STRUCTURE AND DYNAMICS OF GENOME-WIDE DIVERSITY IN *PROCHLOROCOCCUS*

by

Maureen Lynn Coleman

A.B. Biology
Dartmouth College, 2002

Submitted to the Department of Civil and Environmental Engineering in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy
at the
Massachusetts Institute of Technology

June 2008

Signature of Author:_____
Department of Civil & Environmental Engineering
May 15, 2008

Certified by:_____
Sallie W. Chisholm
Lee and Geraldine Martin Professor of Environmental Studies
Professor of Civil and Environmental Engineering
Thesis Supervisor

Accepted by:_____
Daniele Veneziano
Chairman, Departmental Committee for Graduate Students

Structure and dynamics of genome-wide diversity in *Prochlorococcus*

by

Maureen Lynn Coleman

Submitted to the Department of Civil and Environmental Engineering in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the field of Environmental Biology

Abstract

The capability of microbes to thrive in myriad environments has its foundation in the diversity of microbial genomes. Here we explore adaptation and diversification through the lens of the marine cyanobacterium *Prochlorococcus*, which comprises a group of closely-related ecotypes that together perform most of the primary production in low-nutrient regions of the world oceans. *Prochlorococcus* was one of the first microbes in which a genomic basis for ecological differentiation was characterized, in the distinction between high- and low-light adapted ecotypes. It is clear, however, that other axes of differentiation are important, including temperature, nutrient availability, and biotic interactions. This thesis seeks to characterize salient aspects of genomic diversity in *Prochlorococcus* and to advance understanding of the ecological and evolutionary forces that shape this variation. We show that closely related isolates harbor remarkably dissimilar gene complements, and much of this variation is concentrated in specific genome regions, termed islands, that appear to have arisen through phage-mediated gene transfer. Several island-encoded genes likely play important metabolic roles, as inferred from their strong and specific upregulation under stress conditions. A region of the genome involved in phosphate assimilation has highly variable gene content that appears to reflect oceanic phosphate availability. Accordingly, we find extreme differences between strains in the transcriptional response to phosphate starvation. Using metagenomics approaches, we describe high coexisting diversity in natural *Prochlorococcus* populations. Nevertheless, this diversity is structured: a core genome of universal single-copy genes is augmented by a flexible genome. The population genome changes with water depth, reflecting genotypic variation among ecotypes and within the dominant ecotype. Finally, we show that the transcriptomes of wild *Prochlorococcus* correlate strongly with transcriptomes in culture as measured by microarrays. Genes of unknown function are among the most highly expressed in the wild. Several highly expressed genes show signatures of intragenic recombination, a process that likely influences their diversity and function. Overall, this work demonstrates that environmental factors such as light, temperature, nutrient availability, and interspecies interactions each leave different marks in the genome over different scales of time and space. Understanding microbial evolution requires that we dissect diversity over these multiple scales.

Thesis supervisor: Sallie W. Chisholm
Title: Lee and Geraldine Martin Professor of Environmental Studies
Professor of Civil and Environmental Engineering

ACKNOWLEDGMENTS

This work is a product of the creative, challenging, and dynamic environment in the Chisholm lab – the "context" in which I have been embedded over the past few years. This environment has been driven by my advisor, Penny Chisholm. She has taught me to see the big picture and to ask "so what?", without sacrificing the important details. She has been a constant source of wisdom, encouragement, and inspiration, and I am forever grateful.

The people in the Chisholm lab have brought diverse talents, expertise, and personalities to the lab and have made it a fun and scientifically interesting place to be. In particular, I am grateful to my early mentors in the lab, Debbie Lindell, Zackary Johnson, and Erik Zinser, who taught me how to do three different kinds of science. Matt Sullivan has been not only one of my closest collaborators but also one of my closest friends, and I am glad to have shared so many good times with him, both inside and outside the lab (and especially on boats). He remains an unfailing source of advice and Ohioan humor. I thank those colleagues I have worked closely with – Claudia Steglich, Adam Martiny, Anne Thompson, Jorge Frias-Lopez, Greg Kettler, Sébastien Rodrigue, Rex Malmstrom, Katherine Huang, Allison Coe, Marcia Osburne, Scott Chilton – for their enthusiasm and patience. I am sincerely grateful to everyone in the lab for their friendship and willingness to share their ideas and expertise.

I thank my committee members, Martin Polz and Ed DeLong, for their direct and indirect influence on my thinking and on my projects. Martin has taught me to think critically about the significance of microbial diversity and about natural populations. Ed has helped launch the field of metagenomics, and his presence at MIT really shaped the direction of my thesis. I thank him for generously sharing clones, resources, and ideas.

I thank my friends and officemates for sharing laughter, beer, Red Sox, wine, hikes, coffee, movies, and conversation, especially Virginia Rich, Vanja Klepac-Ceraj (and Ivan and Zara), Frédéric Chagnon, Janelle Thompson, Dana Hunt, Ramahi Sarma-Rupavtarm, Noreen Tuross, and Molly Redmond. Vicki Murphy has provided two necessities, chocolate and humor, on a daily basis. I am glad I made a rash decision in 2003 – agreeing to go on a roadtrip with a mere acquaintance, Alison Cohen – because since then she has been a rock-solid friend (and Nurit too). I am still amazed by her energy and her never-ending athletic and professional pursuits.

Thanks to my entire family for keeping me grounded and supporting my sometimes bizarre pursuits – Becky & Mike, Mark & Aileen, Erin & Dan, Nink & PeeWee, Peg & Tim, and especially Mom and Dad.

To Jake, thanks for all you do.

TABLE OF CONTENTS

LIST OF FIGURES

Appendix A

Appendix B

Appendix C

Appendix D

Appendix E

Appendix F

List of Tables

Appendix B

Appendix C

Appendix D

Appendix E

Appendix F

CHAPTER ONE

Introduction

INTRODUCTION

Microbial genomes represent an enormous and dynamic pool of genetic diversity. This diversity enables bacteria and archaea to inhabit nearly every environment on earth and to carry out the metabolic processes that drive global biogeochemical cycles. Understanding how this diversity arises and persists is therefore fundamental to understanding the functioning of the Earth system.

Microbial diversity is shaped by processes occurring at a huge range of temporal and spatial scales — from the nanoseconds and angstroms of molecular DNA and protein interactions, through micron-scale cellular motility and competition with neighboring cells for nutrients during a cell cycle of several hours, to global climate change over geologic time. Consequently the effects of these processes must be studied at different scales as well. This is hardly a unique problem in science; in medical research, both epidemiology and patient-centered approaches tell us something about the causes of disease and the efficacy of various treatments, but neither is sufficient alone. On a case by case basis, the progression of a condition can be traced and the pathology characterized in detail. But individuals are genetically variable and subject to idiosyncratic environmental influences. When an entire population is considered, trends emerge that can suggest cause-and-effect relationships with observable environmental factors. Similarly, understanding microbial diversity benefits from both intensive study of individual isolates and their genomes, and broad population and community surveys in the environment.

The nature and extent of microbial diversity: an evolving landscape

Our comprehension of the nature and extent of microbial diversity has been intimately tied to the available methods for measuring it. The development of molecular-phylogenetic approaches in the 1980's and 1990's represented the most fundamental leap: microorganisms could now be recognized by their 16S rRNA barcode *in situ* without the need for cultivation (Pace 1997). Molecular surveys have since uncovered whole new bacterial divisions or phyla, both in extreme environments and in our own proverbial backyard, dramatically changing our view of bacterial phylogeny and of the entire tree of life (Hugenholtz et al. 1998; Rappe and Giovannoni 2003). At one level, then, microbial diversity refers to this taxonomic diversity as measured by the 16S rRNA metric.

The 16S rRNA and other single-gene markers have also been used to capture the depth, rather than breadth, of microbial diversity within a particular natural population. Deeply sequenced clone libraries have shown that bacterial populations contain many similar, but not identical, gene variants (Thompson et al. 2005; Klepac-Ceraj et al. 2004; Acinas & Klepac-Ceraj et al. 2004). Through these studies, we have come to recognize intrapopulation microheterogeneity as another component of microbial diversity. Thompson et al. (2005) went beyond single-gene markers and observed large variation in genome size

among coexisting *Vibrio* isolates, suggesting that microdiversity in a single locus might be associated with massive changes in gene content throughout the genome.

Parallel efforts to sequence entire microbial genomes have indeed documented large-scale changes in gene content among closely related isolates. This intraspecies diversity has been well documented among pathogens with different virulence or targets of infection. Strikingly, three *E. coli* strains share only 39% of their collective genes, and 10-12% of any *Salmonella* genome is unique (Welch et al. 2002; Edwards et al. 2002). These strain-specific genes are concentrated in pathogenicity islands and often encode functions such as adhesion, secretion systems, toxin production, and iron acquisition – functions important during infection (Hacker and Carniel 2001). This genome-wide intraspecies variation complicates our understanding of microbial taxa delineated by 16S rRNA. A complete picture of microbial diversity must, as a result, consider the whole genome in addition to marker loci.

Because these first genome comparisons focused on specific host-associated bacteria, they are not necessarily generalizable to other lifestyles or taxa. The *E. coli* and *Salmonella* genomes are about 5 Mb in size, but we know that many environmentally important microorganisms have much smaller genomes (Giovannoni et al. 2005; Mira et al. 2001). Do bacteria with small genomes have the same flexibility in gene content? The well-studied pathogens generally belong to the gamma-Proteobacteria, are heterotrophic, and often live attached to surfaces. Accordingly, many of the genes gained and lost in these organisms relate to catabolism of various carbon substrates and adhesion. Are genes gained and lost in autotrophs to the same extent, and what pathways are affected? How diverse, in terms of both gene content and sequence divergence, are groups like the Cyanobacteria, Acidobacteria, and Planctomycetes, which all lead very different lifestyles from *E. coli*? What selective pressures shape their genomes? We might expect genome evolution to follow distinct trajectories in different bacterial phyla or in different environments, but until recently such comparisons were not available.

The combination of molecular-phylogenetic approaches in natural communities and whole genome sequencing of cultured isolates has certainly helped delineate the scope of microbial diversity. At the same time, it has raised many new questions about the generation, maintenance, and significance of this diversity. How do we reconcile 16S rRNA diversity with whole genome comparisons? What is the extent of genome-wide diversity within a particular species or genus? In a single coexisting population, how variable are genomes of a particular species? What are the relative impacts of sequence substitutions, horizontal gene transfer, and recombination on microbial evolution? Answering these questions is crucial for developing an explicit model of microbial evolution.

The *Prochlorococcus* system

To integrate knowledge obtained at different scales and with different approaches, it is useful to have a well-studied model system. The marine cyanobacterium *Prochlorococcus* is a valuable model for studying patterns of diversity for a variety of reasons. Physiologically distinct isolates can be grown in the lab under controlled conditions, enabling hypotheses about phenotypic variation to be tested. Its relatively simple metabolism, planktonic lifestyle, and small genome size reduce the number of variables in the system. *Prochlorococcus* is also abundant and widespread throughout the open oceans from about 40°N to 40°S latitude (Partensky et al. 1999), making it both easily observable and ecologically important.

Early studies of *Prochlorococcus* diversity revealed niche differentiation with respect to light intensity: some isolates grew best at high light intensity, while others grew best at low light intensity and were inhibited in high light (Moore et al. 1995; Moore et al. 1998; Moore and Chisholm 1999). These different physiological types were also found to be genetically distinct and were termed ecotypes (Moore et al. 1998; Rocap et al. 2002). Field studies using genetic markers, primarily the 16S rRNA and 16S-23S internal transcribed spacer (ITS), have documented the abundance of these two major ecotypes, demonstrating that high-light adapted (HL) types dominate the surface waters, while low-light adapted (LL) types tend to increase in abundance at depth (West and Scanlan et al. 1999; West et al. 2001; Ahlgren et al. 2006; Zinser et al. 2006; Johnson et al. 2006).

Whole genome sequences of HL isolate MED4 and LL isolates MIT9313 and SS120 revealed signatures of this niche differentiation (Rocap et al. 2003, Appendix F; Dufresne et al. 2003). MED4 and MIT9313 were chosen for comparative sequencing because they span the largest evolutionary distance among *Prochlorococcus* isolates, and their genomes show evidence of extensive rearrangements and gene losses and/or gains (Rocap et al. 2003, Appendix F). Genome size and G+C content vary considerably among these strains: MED4 is the smallest and has the lowest percent G+C (1,657,990 bp, 31%), followed by SS120 (1,751,080 bp, 36%) and MIT9313 (2,410,873 bp, 51%). The picture emerging from these first *Prochlorococcus* genomes, together with the genome of *Synechococcus* WH8102 (Palenik et al. 2003), suggested gradual genome reduction and sequential decay of particular pathways during the course of evolution from *Synechococcus* to LL *Prochlorococcus* to HL *Prochlorococcus* (Dufresne et al. 2005).

Many of the genes gained and lost in these isolates make sense in light of vertical niche partitioning in the water column. The two LL isolates, for example, encode multiple *pcb* genes for light harvesting (2 in MIT9313 and 8 in SS120), while the HL isolate carries only one. Conversely, the HL isolate carries several genes for DNA repair that are absent from the LL isolates (Rocap et al. 2003, Appendix F; Dufresne et al. 2003). Nutrients also vary with depth, with nitrate and phosphate increasing below the

mixed layer, and nitrite exhibiting a subsurface maximum. Genes involved in nitrogen acquisition appear to have been lost sequentially: *Synechococcus* WH8102 carries genes for nitrate and nitrite utilization, while LL *Prochlorococcus* strain MIT9313 has lost the capacity to use nitrate and HL strain MED4 has further lost the capacity to use nitrite (Rocap et al. 2003, Appendix F). This scenario is complicated, however, by the fact that MIT9313's nitrite transporter appears to have been horizontally acquired (Rocap et al. 2003, Appendix F). Thus other mechanisms besides stepwise genome reduction are required to explain the evolution of *Prochlorococcus*.

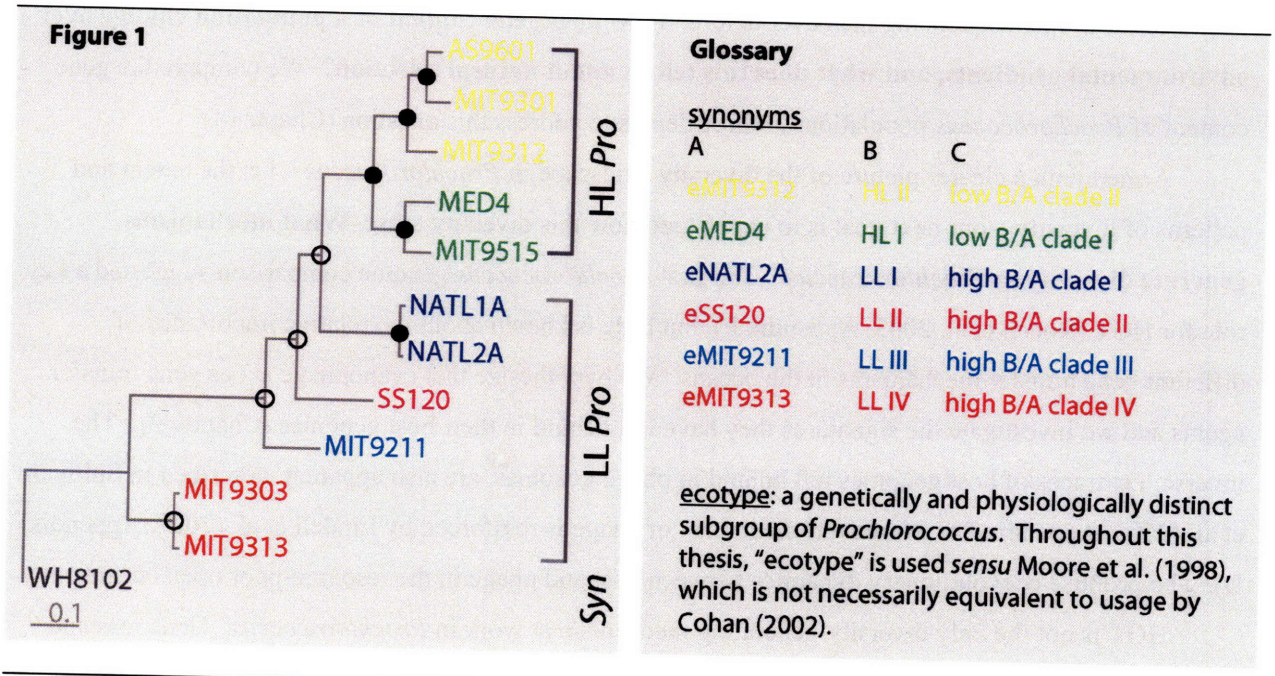
Diversity beyond the HL/LL paradigm

A combination of approaches – field surveys using the ITS and 16S rRNA markers, physiology studies using cultured isolates, and whole genome sequences – have clearly demonstrated the importance of light as a selective force shaping the ecology and evolution of *Prochlorococcus*. It has become increasingly clear, though, that other forces are also at work, and their effects are more subtle and likely occur *within* the HL and LL groups. **One major goal of this thesis is to investigate environmental factors beyond light that shape the ecology of *Prochlorococcus* and drive its genome evolution.**

Within the HL ecotype, two distinct clades exist based on the ITS phylogeny (Rocap et al. 2002) (see Figure 1 and Glossary). Isolates from both clades are adapted for growth at high light intensities (Moore and Chisholm 1999) and their type strains share 99.2% 16S rRNA identity. What is surprising, then, is that these two HL clades exhibit strikingly different geographic and depth distributions. Early studies using 16S rRNA probes showed that the North Atlantic and the Red Sea were dominated by eMED4 and eMIT9312, respectively (West and Scanlan 1999; West et al. 2001). Since then, a qPCR assay has been developed (Ahlgren et al. 2006; Zinser et al. 2006) and employed along a meridional transect (Johnson et al. 2006). The results show eMED4 dominates at higher latitudes near the surface while eMIT9312 dominates at lower latitudes and deeper in the water column (Johnson et al. 2006). These complementary ecological distributions likely result from physiological adaptations and ultimately genomic differences. This observed niche partitioning within the HL ecotype raises many new questions that serve as the motivation for this work.

Besides light, what other axes of differentiation are important for *Prochlorococcus*? What environmental factors have driven the more recent divergence of clades within the HL and LL ecotypes? Here we address these questions using several approaches. First, we compare the genomes of two isolates representing the two distinct HL clades (MED4 and MIT9312), to explore genomic clues to their observed ecological differentiation (Chapter 2). This work complements the work of Johnson et al. (2006) and Zinser et al. (2007), who show through field studies and physiology experiments that temperature adaptation distinguishes the eMED4 and eMIT9312 clades. In addition to light and

temperature, nutrient acquisition capabilities have been shown to vary among isolates, and this variation does not neatly fit the major HL/LL split (Moore et al. 2002; Moore et al. 2005). We hypothesized that phosphate availability in particular selects for different *Prochlorococcus* genotypes in the environment, based on their ability to respond to phosphate limitation. Here we investigate the role of phosphate availability as a selective force and examine its effects on gene expression, genome evolution, and population genetics (Chapter 3).



Phylogenetic relationships and nomenclature of *Prochlorococcus* ecotypes. Figure 1 shows relationships among cultured isolates of *Prochlorococcus* based on the *rpoB* gene (adapted from Coleman and Chisholm 2007). The glossary lists all synonymous names for each clade, represented by the same color in the tree. Column A is from Ahlgren et al. (2007); column B, West and Scanlan (1999); column C, Rocap et al. (2002).

The divergence between HL and LL *Prochlorococcus* left obvious imprints, at least in the first sequenced genomes (MED4, MIT9313, and SS120), in terms of genome size, G+C content, gene content, and sequence divergence, as described above. **How much genome-wide diversity exists within a single ecotype? Between ecotypes?** Answering these questions requires multiple genome representatives from a single ecotype or clade. To this end, we compare genome fragments from wild eMIT9312 cells with their cultured representative (MIT9312) and with each other to assess within-ecotype diversity (Chapter 2). We then examine the diversity of a particular set of genes – those involved in phosphate acquisition – across eleven *Prochlorococcus* genomes including multiple representatives from each major clade (Chapter 3). This work has since been expanded genome-wide by Kettler et al. (2007, Appendix C).

New metagenomics tools enable us to explore genome-wide diversity not just in cultured isolates, but also in natural populations. **What is the extent of diversity in coexisting natural populations of *Prochlorococcus*?** This question has been addressed using individual loci, but the goal of the work presented in Chapter 4 is to expand beyond marker genes to the entire genome of a particular organism. We use metagenomics, specifically short sequence reads obtained by pyrosequencing, to quantify the coexisting diversity in natural *Prochlorococcus* populations in the Pacific. Because the eMIT9312 clade dominates these natural populations, we also obtain a much clearer picture of within-clade diversity, thanks to far deeper sequencing than ever before. **How does gene content of a population change over environmental gradients, and what does this tell us about natural selection?** We compare the gene content of *Prochorococcus* populations at three depths to address this question (Chapter 4).

Armed with a clearer picture of the diversity landscape in *Prochlorococcus* – i.e. the extent and patterns of diversity – our next goal is to understand how this diversity arose. **What mechanisms generate diversity in *Prochlorococcus*?** The first *Prochlorococcus* genome comparison suggested a key role for HGT (Rocap et al. 2003, Appendix F), but little is known about the relative importance of different gene transfer mechanisms in the oceans. We hypothesize that cyanophage act as gene transfer agents and we investigate the signatures they have left behind in their host genomes (Chapter 2). The inverse, i.e. traces of host genomes left behind in phage genomes, are also apparent, described in Sullivan et al. (2005, Appendix E). Moreover, the impact of phage is reinforced by Lindell et al. (2007, Appendix D), who propose coevolutionary dynamics between host and phage in the resource-poor open oceans.

HGT is not the only diversity-generating mechanism at work in *Prochlorococcus*. Gene loss and gene duplication have also played roles in generating the diversity of phosphate acquisition systems observed in *Prochlorococcus* (Chapter 3). Homologous recombination has not been studied at all in *Prochlorococcus*, and thus the rates or extent are unknown. We use sequence data from natural populations to look for instances of recombination in key metabolic genes (Chapter 4). An important next step will be to gauge the relative importance of HGT, point mutations, and recombination in *Prochlorococcus* evolution.

Measuring diversity is useful for understanding the evolutionary history of *Prochlorococcus*, but from an ecological perspective, we know little about the consequences of this diversity. **What is the functional significance of genome-wide diversity in *Prochlorococcus*?** Given the vast differences in genome architecture and gene content between MED4 and MIT9313 (Rocap et al. 2003, Appendix F), we expect important physiological consequences. We examine their physiological and transcriptional responses to one stress, phosphate starvation, and attempt to understand its genomic basis (Chapter 3). We hypothesize that finer scale genomic diversity, for instance within an ecotype or clade, also has important functional consequences. We test this hypothesis by measuring expression of so-called

"flexible" genes – genes that are found in some, but not all, *Prochlorococcus* isolates (Kettler et al. 2007, Appendix C). Often these flexible genes are clustered in the chromosome, and we examine the expression of these clusters in cultured isolates (Chapter 2). We then extend our analysis to natural populations and employ metatranscriptomic methods to measure *Prochlorococcus* gene expression (Chapter 4). **How important are recently acquired "flexible" genes for *Prochlorococcus* function?** The transcriptomic approach is first applied to a 75m population in the subtropical Pacific by Frias-Lopez et al. (2008, Appendix B). We then extend this analysis to three depths and explore the functional role of genes in the flexible genome (Chapter 4). The core genome – genes shared by all *Prochlorococcus* isolates – clearly encodes the central functions of the cell (Kettler et al. 2007, Appendix B). But the flexible genome is dynamic and, we hypothesize, important for adapting to local environmental conditions.

As a whole, this thesis investigates genome-wide diversity and adaptation in *Prochlorococcus* at a finer phylogenetic resolution than the HL/LL groups, and along axes of differentiation besides light. We employ not only traditional comparative genomics, but also physiological studies, *in situ* population genomics, and gene expression studies in both the lab and the field. This combination leads to, we believe, a more complete picture of the origins, nature, and consequences of genome diversity in *Prochlorococcus*. By merging this picture and knowledge from other diverse phylogenetic groups and environments, we can begin to understand the fundamental principles underlying microbial evolution.

References

Acinas, S, V Klepac-Ceraj, DE Hunt, C Pharino, I Ceraj, DL Distel, and MF Polz. 2004. Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* 430: 551-554.

Ahlgren, NA, G Rocap, and SW Chisholm. 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.

Cohan, FM. 2002. What are bacterial species? *Ann. Rev. Microbiol.* 56: 457-487.

Coleman, ML and SW Chisholm. 2007. Code and context: *Prochlorococcus* as a model for cross-scale biology. *Trends Microbiol.* 15: 398-407.

Dufresne A, L Garczarek, and F Partensky. 2005. Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biology* 6: R14.

Dufresne, A, M Salanoubat, F Partensky, F Artiguenave, IM Axmann, V Barbe, S Duprat, MY Galperin, EV Koonin, F Le Gall, KS Makarova, M Ostrowski, S Oztas, C Robert, IB Rogozin, DJ Scanlan, N Tandeau de Marsac, J Weissenbach, P Wincker, YI Wolf, and WR Hess. 2003. Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *PNAS* 100: 10020-10025.

Edwards, RA, G Olsen, and S Maloy. 2002. Comparative genomics of closely related salmonellae. *Trends Microbiol.* 10: 94-99.

Frias-Lopez, F, Y Shi, GW Tyson, ML Coleman, SC Schuster, SW Chisholm, and EF DeLong. 2008. Microbial community gene expression in ocean surface waters. *PNAS* 105: 3805-10.

Giovannoni, SJ, H Tripp, S Givan, M Podar, KL Vergin, D Baptista, L Bibbs, J Eads, T Richardson, M

Noordewier, Michael S Rappe, J Short, J Carrington, and E Mathur. 2005. Genome streamlining in a cosmopolitan oceanic bacterium. *Science* 309: 1242-1245.

Hacker, J, and E Carniel. 2001. Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Reports* 2: 376-381.

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

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

Kettler, G, AC Martiny, K Huang, J Zucker, ML Coleman, S Rodrigue, F Chen, A Lapidus, S Ferriera, J Johnson, C Steglich, GM Church, PM Richardson, and SW Chisholm. 2007. Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genetics* 3: e231.

Klepac-Ceraj, V, M Bahr, B Crump, A Teske, J Hobbie, and MF Polz. 2004. High overall diversity and dominance of microdiverse relationships in salt marsh sulphate-reducing bacteria. *Environ. Microbiol.* 6: 686-698.

Lindell D, JD Jaffe, ML Coleman, ME Futschik, IM Axmann, T Rector, G Kettler, MB Sullivan, R Steen, WR Hess, GM Church, and SW Chisholm. 2007. Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature* 449: 83-6.

Mira, A, H Ochman, and NA Moran. 2001. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* 17: 589-596.

Moore, LR, and SW Chisholm. 1999. Photophysiology of the marine cyanobacterium *Prochlorococcus*: Ecotypic differences among cultured isolates. *Limnol. Oceanogr.* 44: 628-638.

Moore, LR, R Goericke, and SW Chisholm. 1995. Comparative physiology of *Synechococcus* and *Prochlorococcus* - Influence of light and temperature on growth, pigments, fluorescence and absorptive properties. *Marine Ecology-Progress Series* 116: 259-275.

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

Moore, LR, A Post, G Rocap, and SW Chisholm. 2002. Utilization of different nitrogen sources by the marine cyanobacteria *Prochlorococcus* and *Synechococcus*. *Limnol. Oceanogr.* 49: 989-996.

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

Pace, N. 1997. A molecular view of microbial diversity and the biosphere. *Science* 276: 734-740.

Palenik, B, B Brahamsha, FW Larimer, M Land, L Hauser, P Chain, J Lamerdin, W Regala, EE Allen, J McCarren, I Paulsen, A Dufresne, F Partensky, EA Webb, and JB Waterbury. 2003. The genome of a motile marine *Synechococcus*. *Nature* 424: 1037-1042.

Partensky, F, WR Hess, and D Vaulot. 1999. *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol. Mol. Biol. Rev. : MMBR* 63: 106-127.

Rappe, MS, and SJ Giovannoni. 2003. The uncultured microbial majority. *Ann. Rev. Microbiol.* 57: 369-394.

Rocap, G, DL Distel, JB Waterbury, and SW Chisholm. 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, FW Larimer, J Lamerdin, S Malfatti, P Chain, NA Ahlgren, A Arellano, ML Coleman, L Hauser, WR Hess, ZI Johnson, M Land, D Lindell, AF Post, W Regala, M Shah, SL Shaw, C Steglich, MB Sullivan, CS Ting, AC Tolonen, EA Webb, ER Zinser, and SW Chisholm. 2003. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* 424: 1042-1047.

Sullivan, MB, ML Coleman, P Weigele, F Rohwer, and SW Chisholm. 2005. Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. *PLoS Biology* 3: e144.

Thompson, JR, SE Pacocha, C Pharino, V Klepac-Ceraj, DE Hunt, J Benoit, R Sarma-Rupavtarm, DL Distel, and MF Polz. 2005. Genotypic diversity within a natural coastal bacterioplankton

population. *Science* 307: 1311-1313.

Welch, RA, V Burland, G Plunkett, P Redford, P Roesch, D Rasko, E Buckles, S Liou, A Boutin, J Hackett, D Stroud, G Mayhew, D Rose, S Zhou, D Schwartz, N Perna, H Mobley, M Donnenberg, and FR Blattner. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *PNAS* 99: 17020-17024.

West, NJ, and DJ Scanlan. 1999. Niche-partitioning of *Prochlorococcus* populations in a stratified water column in the eastern North Atlantic Ocean. *Appl. Environ. Microbiol.* 65: 2585-2591.

West, NJ, W Schonhuber, NJ Fuller, R Amann, R Rippka, A Post, and DJ Scanlan. 2001. Closely related *Prochlorococcus* genotypes show remarkably different depth distributions in two oceanic regions as revealed by in situ hybridization using 16S rRNA-targeted oligonucleotides. *Microbiology-Sgm* 147: 1731-1744.

Zinser, ER, ZI Johnson, A Coe, E Karaca, D Veneziano, and SW Chisholm. 2007. Influence of light and temperature on *Prochlorococcus* ecotype distribution in the Atlantic Ocean. *Limnol. Oceanogr.* 52: 2205-2220.

Zinser, ER, A Coe, ZI Johnson, AC Martiny, NJ Fuller, DJ Scanlan, and SW Chisholm. 2006. *Prochlorococcus* ecotype abundances in the North Atlantic Ocean as revealed by an improved quantitative PCR method. *Appl. Environ. Microbiol.* 72: 723-732.

CHAPTER TWO

Genomic islands and the ecology and evolution of *Prochlorococcus*

Maureen L. Coleman, Matthew B. Sullivan, Adam C. Martiny, Claudia Steglich, Kerrie Barry, Edward F. DeLong, and Sallie W. Chisholm

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: 1768-1770.

Genomic Islands and the Ecology and Evolution of *Prochlorococcus*

Maureen L. Coleman,[1] Matthew B. Sullivan,[1] Adam C. Martiny,[1] Claudia Steglich,[1]* Kerrie Barry,[2] Edward F. DeLong,[1] Sallie W. Chisholm[1]†

Prochlorococcus ecotypes are a useful system for exploring the origin and function of diversity among closely related microbes. The genetic variability between phenotypically distinct strains that differ by less that 1% in 16*S* ribosomal RNA sequences occurs mostly in genomic islands. Island genes appear to have been acquired in part by phage-mediated lateral gene transfer, and some are differentially expressed under light and nutrient stress. Furthermore, genome fragments directly recovered from ocean ecosystems indicate that these islands are variable among co-occurring *Prochlorococcus* cells. Genomic islands in this free-living photoautotroph share features with pathogenicity islands of parasitic bacteria, suggesting a general mechanism for niche differentiation in microbial species.

Closely related bacterial isolates often contain remarkable genomic diversity (*1*, *2*). Although its functional consequences have been described in a few model heterotrophic microbes (*3*), little is known about genomic microdiversity in the microbial phototrophs that dominate aquatic ecosystems. The marine cyanobacterium *Prochlorococcus* offers a useful system for studying this issue, because they are globally abundant, have very simple growth requirements, have a very compact genome [1.7 to 2.4 megabases (Mb)], and live in a well-mixed habitat. Although the latter appears to offer few opportunities for niche differentiation, *Prochlorococcus* populations consist of multiple coexisting ecotypes (*4*), whose relative abundances vary markedly along gradients of light, temperature, and nutrients (*5*–*9*). Even two high-light adapted (HL) ecotypes, whose type strains (MED4 and MIT9312) differ by only 0.8% in 16*S* ribosomal RNA (rRNA) sequence, have substantially different distributions in the wild (*5*–*9*).

Although whole-genome comparisons between the most distantly related *Prochlorococcus* isolates (97.9% 16*S* rRNA identity) have revealed the gross signatures of this niche differentiation (*10*), important insights into the evolution of diversity in this group likely lie in comparisons between very closely related strains, and between coexisting genomes from wild populations. Thus, we compared the complete genomes of the type strains, MED4 and MIT9312, that represent the two HL clades, and we analyzed genome fragments from wild cells belonging to these clades from the Atlantic and Pacific oceans.

The 1574 shared genes of MED4 and MIT9312 have conserved order and orientation, except for a large inversion around the replication terminus (Fig. 1). The average G + C content is similar in both genomes (31%), and the median sequence identity of the shared genes is 78%, surprisingly low for strains so similar at the rRNA locus (*11*). For most genes, synonymous sites are saturated and protein sequence identity is low (median 80%); this is likely a function of high mutation rates, given that HL *Prochlorococcus* lack several important DNA-repair enzymes (*10*, *12*).

The strain-specific genes between MED4 and MIT9312 (236 in MIT9312 and 139 in MED4) occur primarily (80 and 74%, respectively) in five major islands (Fig. 1). Thus, these genomes have a mosaic structure similar to that of *Escherichia coli* genomes (*1*), though on a smaller scale. The islands are located in the same position in both genomes, implying that they are hotspots for recombination, and the length of island genes is similar to the whole-genome average, suggesting that they are not degraded. We hypothesize that these islands arose via lateral gene transfer and continually undergo rearrangement, on the basis of a number of characteristics. First, three islands are associated with tRNA genes (fig. S1), which are common integration sites for mobile elements (*13*). Second, the 3′ end of tRNA-proline, which flanks ISL3 in both genomes, is repeated 13 times in MIT9312-ISL3 (Fig. 2A) and three times in MED4-ISL3 (fig. S2), suggesting repeated remodeling of this island. Third, some of the genes found in a particular island in MED4 are found in a different island in MIT9312 (Fig. 1), a rearrangement that may have been mediated by a 48–base pair sequence element we call PRE1 (*Prochlorococcus* repeat element 1; fig. S3); portions of PRE1 are repeated, almost exclusively in islands, 13 times in MED4 (fig. S2), and 9 times in MIT9312 (Fig. 2A). Finally, up to 80% of the genes in any given MIT9312 island are most similar to the genes of noncyanobacterial organisms including phage, Eukarya, and Archaea, consistent with the recent observation that horizontally acquired genomic islands reflect a gene pool that differs from that of the core genome (*14*).

It is likely that phage, which often carry host genes (*15*, *16*), mediate some of the island-associated lateral gene transfer, and the *hli* gene family in particular appears to have undergone repeated phage-host gene exchange (*16*). Of the 24 *hli* genes in MIT9312, 18 are found in the five major islands or their flanking regions. All 18 belong to the multicopy and sporadically distributed group that includes phage copies (Fig. 2A) and is well differentiated from widespread single-copy *hli* genes found in cyanobacteria (*16*). Other phagelike genes in islands include an integrase, DNA methylases, a second *phoH*, a MarR-family transcriptional regulator, a putative hemagglutinin neuraminidase, and an endonuclease (*15*), further supporting a link between phage and island dynamics.

Many island genes in the two strains appear to encode functions related to physiological stress and nutrient uptake and thus may be important in the high-light, low-nutrient surface waters dominated by HL *Prochlorococcus*. ISL2 and ISL5 in MIT9312, for example, encode 12 of the 24 *hli* genes, known to be important under a variety of stress conditions (*17*); they also encode two outer-membrane transport

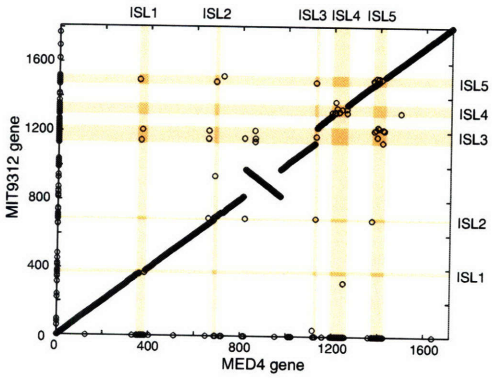


Fig. 1. Whole-genome alignment showing the positions of orthologous genes in MED4 and MIT9312. Strain-specific genes appear on the axes. The locations of five major islands defined by whole-genome alignment (*25*) are shaded.

[1]Massachusetts Institute of Technology, Department of Civil and Environmental Engineering, 15 Vassar Street, Cambridge, MA 02139, USA. [2]U.S. Department of Energy Joint Genome Institute, Production Genomics Facility, Walnut Creek, CA 94598, USA.

*Present address: University Freiburg, Department of Biology II/Experimental Bioinformatics, Schänzlestrasse 1, D-79104 Freiburg, Germany.

†To whom correspondence should be addressed. E-mail: chisholm@mit.edu

proteins; and a cyanophage-like homolog of *phoH* thought to be involved in the phosphate stress response (*15*). ISL3 in this strain contains a paralog of *psbF*, which encodes part of cytochrome b559, thought to protect against photoinhibition (*18*). Islands also contain genes involved in nutrient assimilation, including a cyanate transporter and lyase in MED4 and two transporters, for manganese/iron and amino acids, in MIT9312 (fig. S1).

In addition to genes involved in potentially growth-limiting processes, islands also contain genes that could play a role in selective mortality. ISL4 in both MED4 and MIT9312 encodes proteins involved in cell surface modification, including biosynthesis of lipopolysaccharide, a common phage receptor (*19*) (fig. S1). Phages are important agents of mortality in the oceans (*20*), and thus cell surface properties are likely under strong selection.

Clearly, for island genes to influence a cell's fitness, they must be expressed. When MED4 cells are starved for phosphorus, nine ISL5 genes are differentially expressed, nearly all of unknown function (table S1). When cells are shifted to high light, 38 island genes are differentially expressed, including seven *hli* genes (table S1) that in *Synechocystis* encode proteins that accumulate when cells absorb excess excitation energy (e.g., under high light, nutrient limitation, and low temperatures) (*17*). Thus, 26% of all MED4 island genes are differentially expressed under P starvation or high-light stress; only one of these is differentially expressed under both conditions (conserved hypothetical gene PMM1416), suggesting that island genes contribute to specific stress responses.

The genome variation within the eMIT9312 clade [sensu (*7*)] was examined in wild populations of *Prochlorococcus* by aligning short genome fragments from the Sargasso Sea (*21*), where this clade dominates (*7*), against the MIT9312 genome (Fig. 2B). Nearly constant coverage was observed, confirming a stable core genome, except for notable gaps at ISL1, ISL3, and ISL4. This finding indicates that very few wild sequences match genes in these islands, and it supports the hypothesis that these regions are hypervariable in HL *Prochlorococcus* genomes. In contrast, genes belonging to ISL2 and ISL5 are relatively well represented in the Sargasso Sea data set (Fig. 2B, fig. S2). In MED4 and MIT9312, these islands contain about half of the *hli* genes, lack the tRNA genes implicated in integration of mobile elements, and contain a smaller fraction of noncyanobacterial genes than do the other islands. This finding suggests that the genes in these islands have become fixed in this wild population.

Examination of 36 large genome fragments (1.1 Mb total sequence; median size 34 kb) (table S2) from the Hawaii Ocean Time-Series Station (*22*) further confirms that a stable core genome surrounds islands of variability, because most fragments showed remarkable conservation of gene content and order with respect to the MED4 and MIT9312 genomes. Thirty-four of the 36 fragments were more similar to MIT9312 than to MED4; two contained rRNA operons, confirming their phylogenetic affiliation with the eMIT9312 clade (fig. S4). The eMIT9312 fragments have about 90% identity with the MIT9312 genome and about 80% with MED4 (Table 1). Collectively, these results suggest that the wild eMIT9312 population is a coherent group identifiable by sequence similarity in the absence of an rRNA operon (*11*). eMIT9312 genome fragments from this wild population are more similar to each other than to the genome of the type strain MIT9312 (isolated from the Atlantic Ocean), but still share only 93% average sequence identity (Table 1), indicating high coexisting diversity in core genes.

Five eMIT9312 genome fragments from the Hawaii sample border the major islands defined above. About 60% of the genes in these islands have no ortholog in either MED4 or MIT9312, and two fragments border ISL1, yet their gene content is largely different from each other and from the MIT9312 and MED4 genomes (fig. S5). Indeed, a third of the island genes in these two fragments are novel, i.e., have no detectable homologs, implying that cells have access to a large novel gene pool in the oceans (*14*). Like the islands in the MED4 and MIT9312 genomes, these two fragments contain signatures of mobility, including duplicated tRNA genes, copies of the repeat PRE1, and an integrase gene. This reveals that islands are dynamic even within a single ecotype clade as we have defined it.

One observation that stimulated this work is the dramatic difference in distribution and abundance of the two HL *Prochlorococcus* ecotype clusters (*5–9*), as defined by their rRNA internal transcribed spacer (ITS) sequence similarity. Although strains belonging to these two clusters have different island gene content, so do cells from field populations that belong to a single cluster. Therefore, other genomic features are likely to be important in explaining niche differentiation between eMED4 and eMIT9312 cells in the wild. Differential temperature adaptation, for example, which is thought to be an important determinant of ecotype distribu-

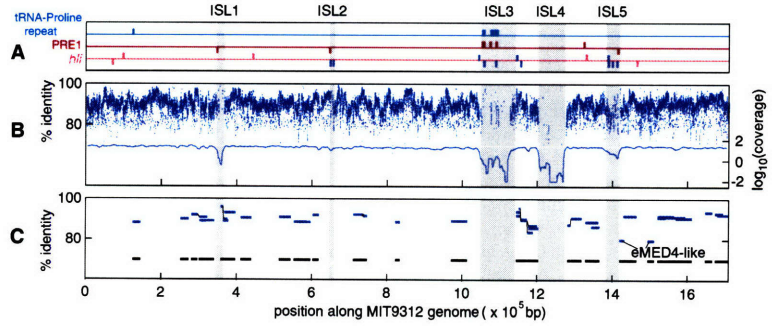


Fig. 2. Features of genomic islands (shaded) in the *Prochlorococcus* strain MIT9312 genome compared with wild sequences from the Atlantic and Pacific Oceans. **(A)** Locations of repetitive elements and *hli* genes in MIT9312, shown above or below the horizontal line for the forward or reverse strand, respectively. *hli* genes shown in pink belong to the single-copy conserved group and those shown in blue belong to the multicopy phage-encoded group (*16*). **(B)** Percent identity of Sargasso Sea shotgun database sequences (*21*) aligned to MIT9312 (top, left axis) and average coverage in the database of a given position in the MIT9312 genome (bottom, right axis). Log_{10}(coverage) is set to −2 when coverage equals 0. **(C)** Genomic locations and percent identity of wild genome fragments (eMIT9312-like unless noted) aligned to MIT9312. Where the alignment is interrupted, a black line connects aligned segments of a single fragment. Fragments are projected down to 70% horizontal to visualize total coverage.

Table 1. Median pairwise percent identities, for all orthologous gene pairs and for large aligned regions >4 kb (*25*). Numbers in parentheses indicate the number of orthologous gene pairs from which the median was calculated.

	MED4-MIT9312	MED4-eMIT9312 fragments	MIT9312-eMIT9312 fragments	Overlapping fragments
Orthologs (nucleotides)	78.4 (1574)	79.5 (1063)	90.6 (1092)	93.2 (434)
Orthologs (amino acids)	80.0 (1574)	82.4 (1063)	92.9 (1092)	95.2 (434)
Large aligned regions	79.0	79.9	90.7	92.6

tions (*5*), can be achieved through sequence (*23*) or regulatory (*24*) changes in the core genome. Nonetheless, given their prevalence, mobility, and expression under relevant conditions, islands likely play a role in adaptation, but on shorter time scales, or more local spatial scales, in the context of large populations that harbor substantial genomic variability.

Thus, although streamlined for life in the oligotrophic oceans, the genomes of HL *Prochlorococcus* are not static. Cell-to-cell genome variability is concentrated in islands containing genes that are differentially expressed under stresses typical of oceanic environments. Just as pathogenicity islands alter the host specificity and virulence of pathogenic bacteria (*3*), genomic islands in *Prochlorococcus* may contribute to niche differentiation in the surface oceans. Although other factors, such as small insertions and deletions, substitutions in homologous proteins, and differential regulation are important contributors to diversity, the prevalence of genomic islands and their features argue that these also play an influential role. We postulate that lateral gene transfer in genomic islands is an important mechanism for local specialization in the oceans. If true, genomic islands of natural taxa should contain genes that are ecologically important in a given environment, regardless of the core genome phylogeny. Testing this hypothesis will not only advance our understanding of microbial diversity in the ocean, but also contribute to a unified understanding of genomic evolutionary mechanisms and their impact on microbial ecology.

References and Notes

1. R. A. Welch *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 17020 (2002).
2. J. R. Thompson *et al.*, *Science* **307**, 1311 (2005).
3. J. Hacker, J. B. Kaper, *Annu. Rev. Microbiol.* **54**, 641 (2000).
4. L. R. Moore, G. Rocap, S. W. Chisholm, *Nature* **393**, 464 (1998).
5. Z. I. Johnson *et al.*, *Science* **311**, 1737 (2006).
6. E. R. Zinser *et al.*, *Appl. Environ. Microbiol.* **72**, 723 (2006).
7. N. Ahlgren, G. Rocap, S. W. Chisholm, *Environ. Microbiol.* **8**, 441 (2006).
8. N. J. West, D. J. Scanlan, *Appl. Environ. Microbiol.* **65**, 2585 (1999).
9. N. J. West *et al.*, *Microbiology* **147**, 1731 (2001).
10. G. Rocap *et al.*, *Nature* **424**, 1042 (2003).
11. K. T. Konstantinidis, J. M. Tiedje, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 2567 (2005).
12. A. Dufresne, L. Garczarek, F. Partensky, *Genome Biol.* **6**, R14 (2005).
13. W. D. Reiter, P. Palm, S. Yeats, *Nucleic Acids Res.* **17**, 1907 (1989).
14. W. W. Hsiao *et al.*, *PLoS Genet.* **1**, e62 (2005).
15. M. B. Sullivan, M. L. Coleman, P. Weigele, F. Rohwer, S. W. Chisholm, *PLoS Biol.* **3**, e144 (2005).
16. D. Lindell *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 11013 (2004).
17. Q. He, N. Dolganov, O. Bjorkman, A. R. Grossman, *J. Biol. Chem.* **276**, 306 (2001).
18. D. H. Stewart, G. W. Brudvig, *Biochim. Biophys. Acta* **1367**, 63 (1998).
19. A. Wright, M. McConnell, S. Kanegasaki, in *Virus Receptors*, L. L. Randall, L. Philipson, Eds. (Chapman and Hall, New York, 1980), pp. 27–57.
20. J. A. Fuhrman, *Nature* **399**, 541 (1999).
21. J. C. Venter *et al.*, *Science* **304**, 66 (2004).
22. E. F. DeLong *et al.*, *Science* **311**, 496 (2006).
23. G. N. Somero, *Annu. Rev. Physiol.* **57**, 43 (1995).
24. M. M. Riehle, A. F. Bennett, R. E. Lenski, A. D. Long, *Physiol. Genomics* **14**, 47 (2003).
25. Materials and methods are available as supporting material on *Science* Online.

26. We thank T. Rector, N. Hausman, and R. Steen for Affymetrix microarray processing; M. Polz for helpful discussions; and D. Lindell and A. Tolonen for comments on the manuscript. This work was supported by grants from NSF Biological Oceanography (S.W.C.) and Microbial Observatory (E.F.D.) Programs, the U.S. Department of Energy (DOE) GTL Program (to S.W.C. and G. Church), and the Gordon and Betty Moore Foundation (S.W.C. and E.F.D.). Sequencing support came from the DOE Microbial Genomics Program (E.F.D.) and DOE GTL and Community Sequencing Program (S.W.C.), conducted at the DOE Joint Genome Institute. Sequences are available in GenBank: BX548174 (MED4 genome), CP000111 (MIT9312 genome), and DQ366711 to DQ366746 (environmental genome fragments).

Supporting Online Material

www.sciencemag.org/cgi/content/full/311/5768/1768/DC1
Materials and Methods
Figs. S1 to S5
Tables S1 to S3
References

31 October 2005; accepted 17 February 2006
10.1126/science.1122050

Supporting Online Material

Materials and Methods

Genome sequencing, assembly, and annotation

Prochlorococcus strain MIT9312 was isolated from the Gulf Stream by flow cytometric sorting (*1*). DNA was isolated using a modified phenol/chloroform extraction (*2*). Library construction, sequencing, assembly, and automated annotation were performed by the Department of Energy Joint Genome Institute (JGI). The annotation was manually edited to reflect the human-curated MED4 genome annotation.

Large-insert environmental genomic libraries

Seawater samples were collected from the Hawaii Ocean Time-series Station ALOHA as described (*3*). Clones with a "HOT0M-#" identifier are BAC clones from the surface (0m) collected in December 2001, and BAC libraries were constructed as previously described (*4*). Clones with "HF10-#" or "ASNC#" identifiers are fosmid clones from 10m depth sampled 7 October 2002 (*3*). Fosmid library construction and end-sequencing were performed as described (*3*). Putative *Prochlorococcus* BAC and fosmid clones were identified based on end-sequence similarity and inserts were fully sequenced by the JGI.

Annotation of environmental genome fragments

Protein coding genes were predicted using a combination of GeneMarkS (*5*) and Glimmer 2.0 (*6*). tRNA genes were predicted using tRNAscan-SE (*7*). Predicted protein-coding genes were compared to the sequenced *Prochlorococcus* genomes using standalone BLAST (*8*) and annotations were assigned to match the human-curated MED4 genome annotation. Predicted genes with no similarity to *Prochlorococcus* were searched against the nonredundant GenBank database using blastp.

Sequence comparisons

Whole genome alignments and fosmid-genome alignments were performed with NUCmer, part of the MUMmer 3.10 package (*9*), with the following parameters: minimum match length, 10; break length, 1200; maximum gap between clusters, 1000; minimum cluster length, 220. Major islands were defined as unaligned regions greater than 10kb. Major islands and smaller unaligned regions ("islets") are listed in Table S2. Orthologs were defined as bidirectional best blastp hits with an e-value less than 1e-10. Reciprocal best hits that did not meet the e-value cutoff but that showed conserved synteny were also inferred to be orthologs. Protein sequences were aligned using ClustalW 1.7 (*10*).

Putative *Prochlorococcus* sequences were identified in the Sargasso Sea dataset, binned by sample (SAR1-7) (*11*), using the MIT9312 and MED4 proteomes as queries. SAR clones with an e-value less than 1e-10 were extracted from the database and aligned to the MIT9312 and MED4 genomes using NUCmer with minimum match length 10. Samples SAR1-4, all collected in February 2003, are combined and shown in Figures 2B and S2. Coverage was calculated for each position in MIT9312 and MED4 by summing the number of clones covering that position in the NUCmer alignment. Coverage is plotted in Figures 2A and S2A for a 10kb sliding window (step = 500bp).

Phylogeny

ITS sequences from fosmid clones were aligned using the ARB software package (*12*) against a custom database containing all known *Prochlorococcus* ITS sequences. As in Rocap et al. (*13*), the two tRNA sequences within the ITS region were excluded. A total of 476 basepairs were used for the analysis. Selected sequences were exported to PAUP* v.4b10 (*14*) for phylogenetic analysis using distance (minimum evolution as criterion and paralinear (logdet) distance correction), maximum parsimony, and maximum likelihood (HKY model parameters, gamma shape optimized by iterative likelihood searches starting from a maximum parsimony tree). Bootstrap analysis (100 resamplings) was done with heuristic

searches utilizing random addition and tree-bisection reconnection branch swapping methods. The neighbor-joining topology is shown in Figure S4.

hli gene clustering

MIT9312 hli genes were aligned to other known cyanobacterial *hli* genes and clustered using GENERAGE as described (*15*). These clusterings were examined and MIT9312 *hli* genes were designated as "multicopy" or "single copy" based upon whether they grouped with one or the other gene type (as designated in ref. 14).

Experimental conditions

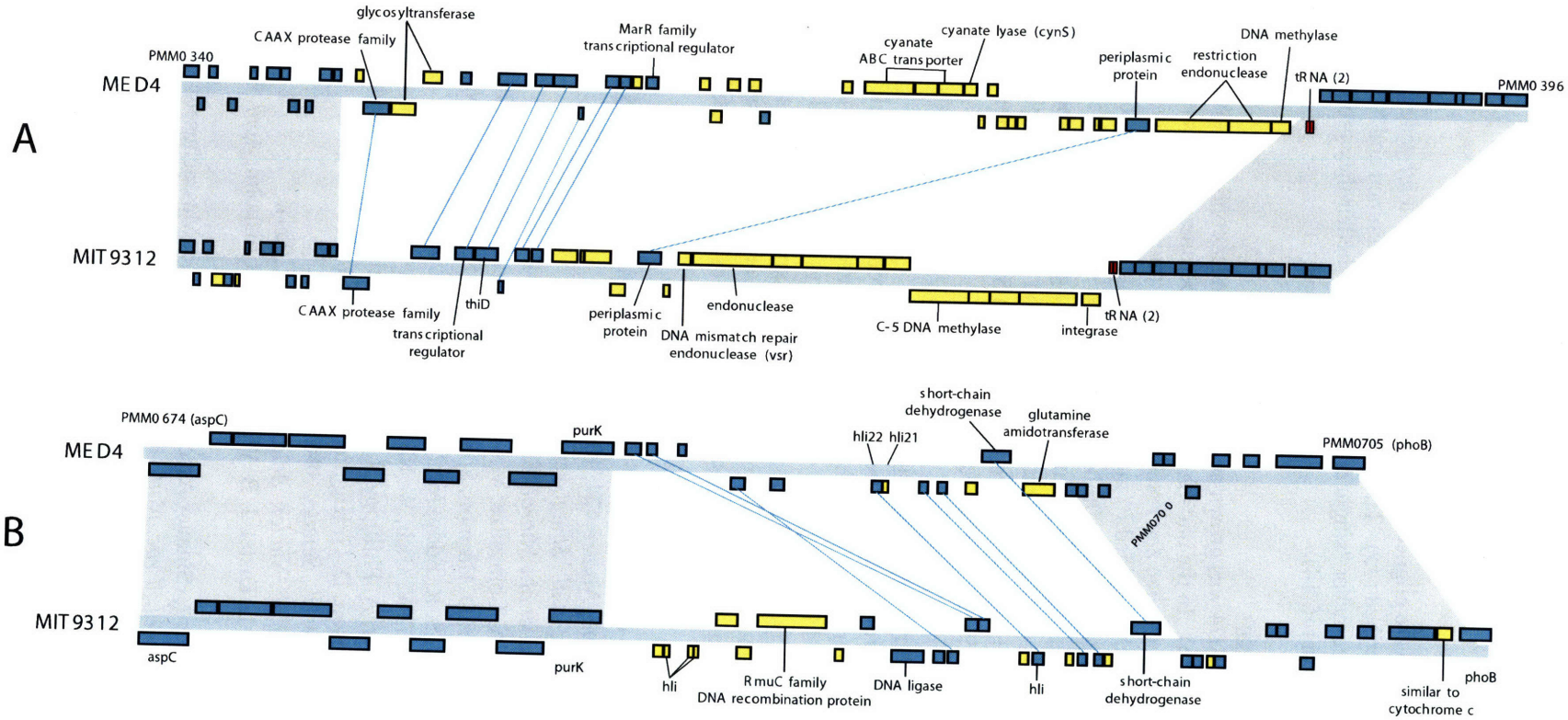
Prochlorococcus MED4 was grown at 21°C or 24°C under continuous white light (30 μmol photon m^{-2} s^{-1}) in Pro99 media (*16*). In the phosphate starvation experiment, triplicate cultures were harvested by centrifugation, washed twice and resuspended in either phosphate replete (standard Pro99) or minus-phosphate (Pro99 with no added P) media. After 48 hours, cells were collected by centrifugation, resuspended in storage buffer (200mM sucrose, 10mM sodium acetate, 5mM EDTA; pH 5.2), snap-frozen in liquid nitrogen, and stored at –80°C. In the light shift experiment, cells were dark-acclimated for 5 h then shifted to high white light (55 μmol photon m^{-2} s^{-1}). Control cells were kept in darkness. After 45 minutes, cells were harvested by filtration on Supor-450 membranes, which were immersed in storage buffer, snap frozen, and stored at –80°C.

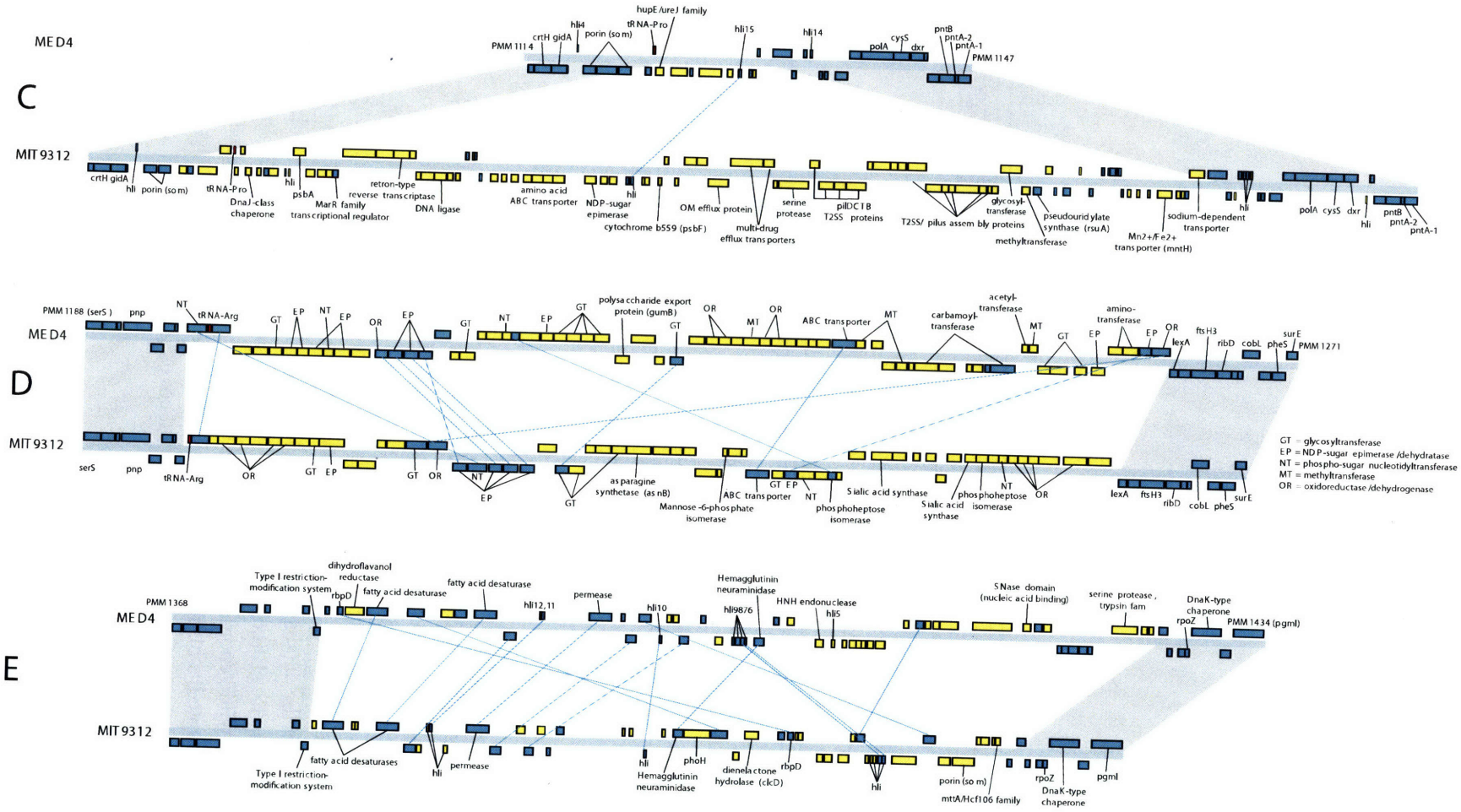
RNA extraction and transcript analysis

RNA was extracted from cells resuspended in buffer according to Lindell et al. (*17*). RNA was extracted from cells on the filters using a hot phenol method as described (*18*). 2 μg total RNA was labeled and hybridized to custom MD4-9313 Affymetrix microarrays using standard protocols (http://www.affymetrix.com/technology/index.affx). Expression summaries were computed using RMA normalization (*19*) and differentially expressed genes were identified using Bayesian methods implemented in Cyber-T (*20*) and false discovery rates estimated using QVALUE (*21*).

1. L. R. Moore, G. Rocap, S. W. Chisholm, *Nature* **393**, 464 (1998).
2. G. Rocap *et al.*, *Nature* **424**, 1042 (2003).
3. E. F. DeLong *et al.*, *Science* (in press).
4. M. T. Suzuki *et al.*, *Microb Ecol* **48**, 473 (2004).
5. J. Besemer, A. Lomsadze, M. Borodovsky, *Nucleic Acids Res.* **29**, 2607 (2001).
6. A. L. Delcher, D. Harmon, S. Kasif, O. White, S. L. Salzberg, *Nucleic Acids Res* **27**, 4636 (1999).
7. T. M. Lowe, S. R. Eddy, *Nucleic Acids Res.* **25**, 955 (1997).
8. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).
9. A. L. Delcher, A. Phillippy, J. Carlton, S. L. Salzberg, *Nucleic Acids Res.* **30**, 2478 (2002).
10. J. D. Thompson, D. G. Higgins, T. J. Gibson, *Nucleic Acids Res* **22**, 4673 (1994).
11. J. C. Venter *et al.*, *Science* **304**, 66 (2004).
12. W. Ludwig *et al.*, *Nucleic Acids Res* **32**, 1363 (2004).
13. G. Rocap, D. L. Distel, J. B. Waterbury, S. W. Chisholm, *Appl. Environ. Microbiol.* **68**, 1180 (2002).
14. D. L. Swofford. (Sinauer Associates, Sunderland, Massachusetts, 2000).
15. D. Lindell *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 11013 (2004).
16. L. R. Moore, A. F. Post, G. Rocap, S. W. Chisholm, *Limnol. Oceanogr.* **47**, 989 (2002).
17. D. Lindell, J. D. Jaffe, Z. I. Johnson, G. M. Church, S. W. Chisholm, *Nature* **438**, 86 (2005).
18. D. Lindell, A. F. Post, *Appl. Environ. Microbiol.* **67**, 3340 (2001).
19. B. M. Bolstad, R. A. Irizarry, M. Astrand, T. P. Speed, *Bioinformatics* **19**, 185 (2003).
20. P. Baldi, A. D. Long, *Bioinformatics* **17**, 509 (2001).
21. J. D. Storey, R. Tibshirani, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 9440 (2003).

Figure S1. Individual island comparisons between MED4 and MIT9312. Genes with an ortholog in the other genome are blue; unique genes are yellow. Core genome segments are connected by gray shading; segments not connected by shading are islands (as defined in Materials and Methods). Blue lines connect orthologs within islands.




C
MED4
PMM 1114
crtH gidA
hli4
porin (som)
hupE/ureJ family
tRNA-Pro
hli15
hli14
cysS
polA
dxr
pntB
pntA-2
pntA-1
PMM 1147
MIT 9312
crtH gidA
hli
porin (som)
tRNA-Pro
DnaJ-class chaperone
hli
psbA
MarR family transcriptional regulator
retron-type reverse transcriptase
DNA ligase
amino acid ABC transporter
NDP-sugar epimerase
hli
cytochrome b559 (psbF)
OM efflux protein
multi-drug efflux transporters
serine protease
pilDCTB T2SS proteins
T2SS/ pilus assembly proteins
glycosyl-transferase
pseudouridylate synthase (rsuA)
methyltransferase
Mn2+/Fe2+ transporter (mntH)
sodium-dependent transporter
hli
polA cysS dxr
hli
pntB
pntA-2
pntA-1
D
MED4
PMM 1188 (serS)
pnp
NT
tRNA-Arg
GT
EP
NT
EP
OR
EP
GT
NT
EP
GT
polysaccharide export protein (gumB)
GT
OR
MT
OR
ABC transporter
MT
carbamoyl-transferase
acetyl-transferase
MT
GT
EP
amino-transferase
EP
OR
lexA
ftsH3
ribD
cobL
pheS
surE
PMM 1271
MIT 9312
serS
pnp
tRNA-Arg
OR
GT
EP
GT
OR
NT
EP
GT
asparagine synthetase (asnB)
Mannose-6-phosphate isomerase
ABC transporter
GT
EP
NT
phosphoheptose isomerase
Sialic acid synthase
Sialic acid synthase
phosphoheptose isomerase
NT
OR
lexA
ftsH3
ribD
cobL
pheS
surE
GT = glycosyltransferase
EP = NDP-sugar epimerase /dehydratase
NT = phospho-sugar nucleotidyltransferase
MT = methyltransferase
OR = oxidoreductase /dehydrogenase
E
MED4
PMM 1368
Type I restriction-modification system
dihydroflavanol reductase
rbpD
fatty acid desaturase
fatty acid desaturase
hli12,11
permease
hli10
Hemagglutinin neuraminidase
hli9876
HNH endonuclease
hliS
SNase domain (nucleic acid binding)
serine protease, trypsin fam
DnaK-type chaperone
rpoZ
PMM 1434 (pgml)
MIT 9312
fatty acid desaturases
Type I restriction-modification system
hli
permease
hli
Hemagglutinin neuraminidase
phoH
dienelactone hydrolase (clcD)
rbpD
hli
porin (som)
mttA/Hcf106 family
rpoZ
DnaK-type chaperone
pgml

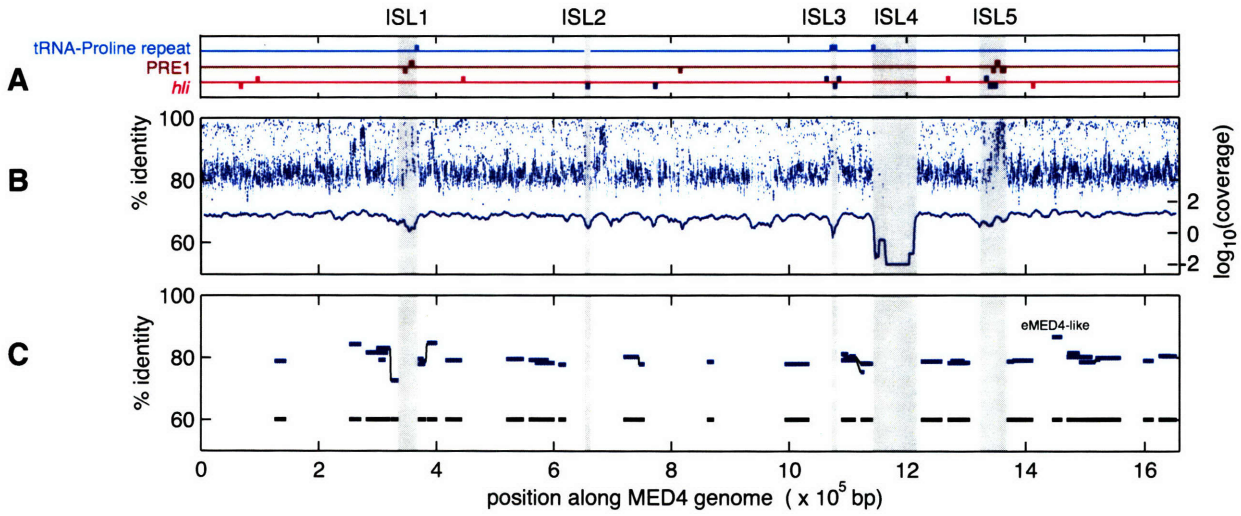
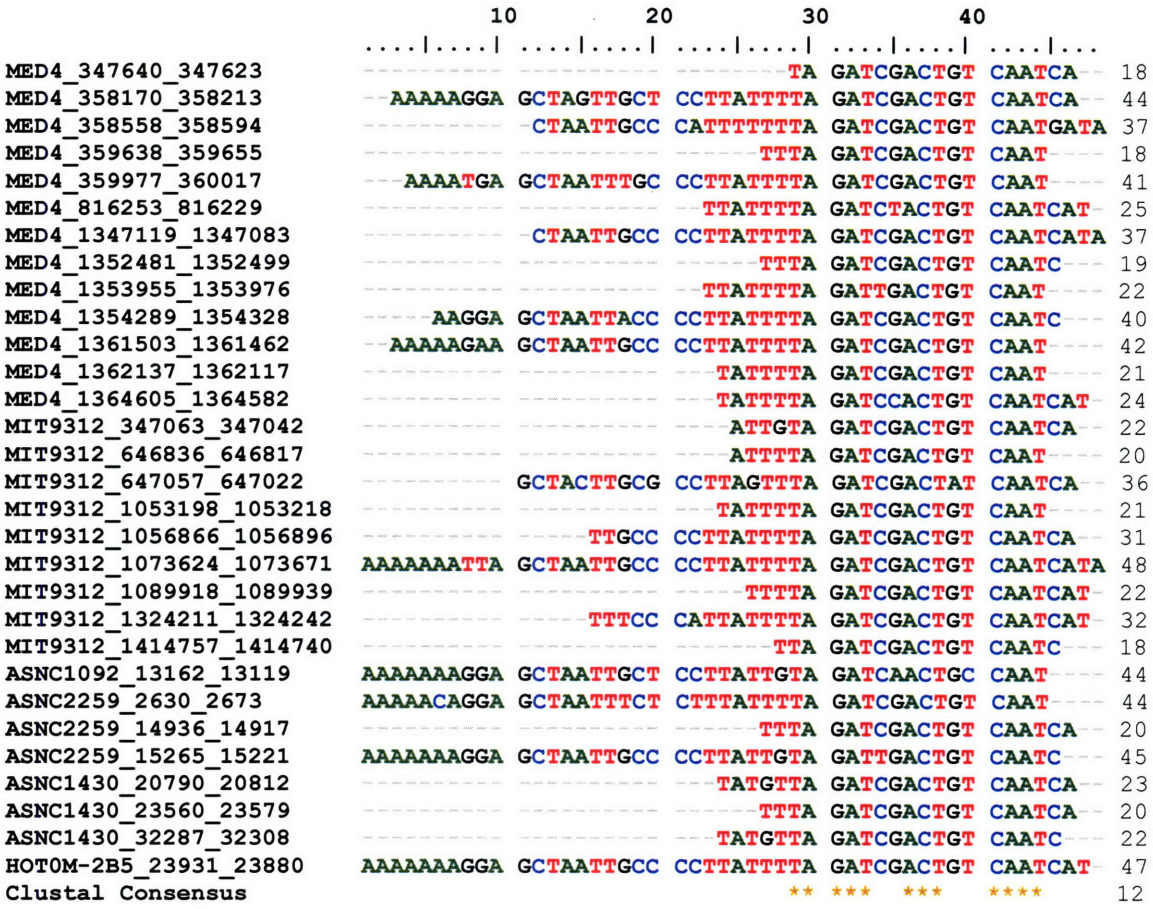


Figure S2. Features of genomic islands (shaded) in the Prochlorococcus strain MED4 genome compared with wild sequences from the Atlantic and Pacific Oceans. A Locations of repetitive elements and hli genes in MED4, shown above/below the horizontal line for forward/reverse strand. Magenta, hli genes belong to the single-copy conserved group and blue, to the multi-copy/phage-encoded group. B Percent identity of Sargasso Sea shotgun database sequences aligned to MED4 (top, left axis) and average coverage in the database of a given position in the MED4 genome, calculated for a 10kb window (bottom, right axis). Log10(coverage) set to –2 when coverage equals 0. C Genomic locations and percent identity of wild genome fragments (eMIT9312-like unless noted) aligned to MED4. Where the alignment is interrupted, a black line connects aligned segments of a single fragment. Fragments are projected down to 60% horizontal to visualize total coverage.

Figure S3. Sequence alignment of the PRE1 repeat element found in *Prochlorococcus* genomic islands. The source DNA name and genomic location are listed on the left. MED4 = *Prochlorococcus* strain MED4, MIT9312 = *Prochlorococcus* strain MIT9312; ASNC# and HOT0M sequences are environmental fosmid or BAC clones. The positions conserved and present in every sequence are starred in the bottom line, and the number of aligned nucleotides is given on the right.



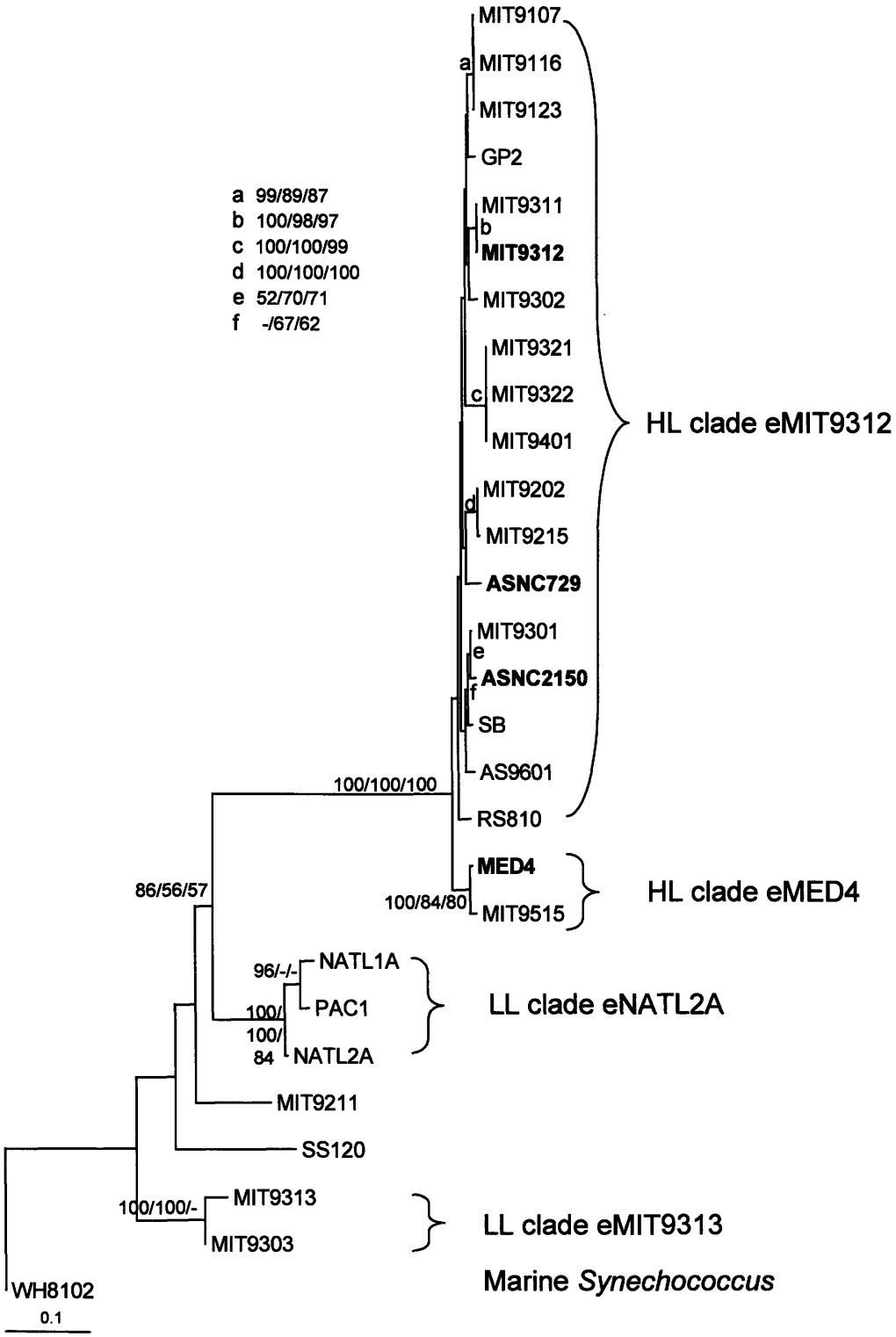


Figure S4. Phylogenetic tree based on the ITS sequence (see Materials and Methods for details). Sequences discussed in the text are shown in bold. ASNC# identifiers indicate environmental fosmid clones from HOT. Bootstrap support values greater than 50% are shown in the following order: neighbor joining/maximum parsimony/maximum likelihood.

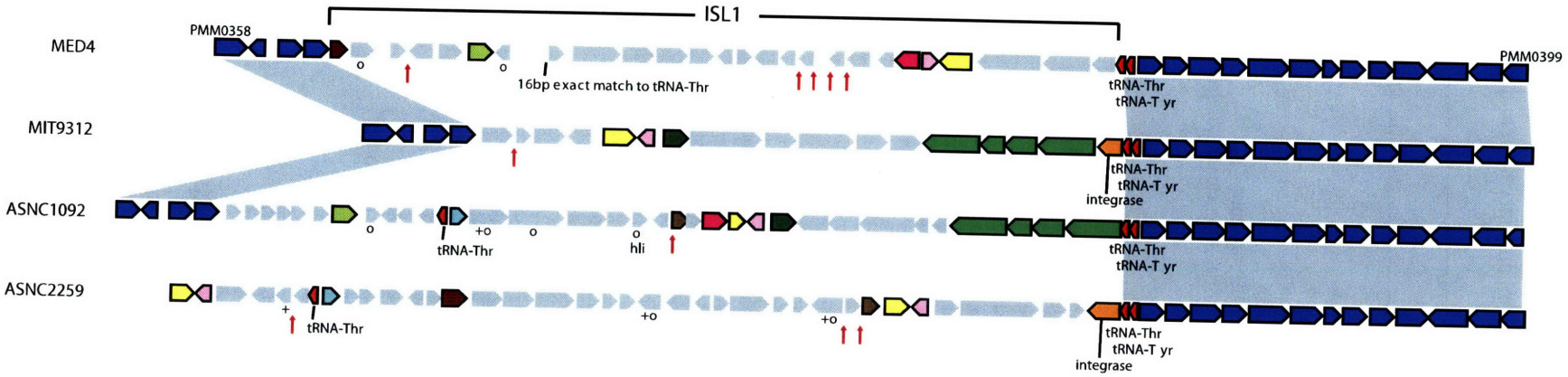


Figure S5. Diversity of ISL1 in wild and cultured *Prochlorococcus*. MED4 = *Prochlorococcus* strain MED4, MIT9312 = *Prochlorococcus* strain MIT9312, ASNC# are environmental fosmid clones. Homologous genes in this region are indicated by the same color. Core genes shared by all are dark blue and connected by gray shading. Gray genes are unique to a given genome/genome fragment, or have homologs outside this region as indicated by the legend.

Table S1. Island genes differentially expressed under phosphate starvation and high light shift.

Gene	Fold change	q-value	Description
Phosphate starvation			
PMM1416	44.52	2.04E-12	conserved hypothetical
PMM1414	16.33	1.72E-07	conserved hypothetical
PMM1409	21.03	6.37E-07	conserved hypothetical
PMM1408	2.94	1.71E-03	conserved hypothetical
PMM1403	6.50	1.82E-03	HNH endonuclease:HNH nuclease
PMM1415	3.68	1.98E-03	predicted hydrolase of the alpha/beta superfamily
PMM1407	3.12	3.46E-03	hypothetical protein
PMM1411	3.63	1.14E-02	hypothetical protein
PMM1405	-2.09	4.78E-02	hypothetical protein
High Light			
PMM0690	32.96	1.31E-12	hli21 high light inducible protein
PMM1404	40.02	1.02E-11	hli5 high light inducible protein
PMM1397	57.42	1.50E-11	hli8 high light inducible protein
PMM1396	50.34	2.59E-11	hli9 high light inducible protein
PMM0689	12.88	2.93E-10	hli22 high light inducible protein
PMM1384	14.17	6.74E-10	hli12 high light inducible protein
PMM1385	10.00	6.16E-09	hli11 high light inducible protein
PMM1387	7.22	3.66E-07	hypothetical protein
PMM0366	4.86	2.60E-06	Type-1 copper (blue) domain
PMM1379	-4.91	1.72E-05	putative dape gene and orf2
PMM1422	4.28	2.66E-05	conserved hypothetical protein
PMM0380	4.01	3.26E-05	conserved hypothetical protein
PMM1128	3.93	3.35E-05	hli15 high light inducible protein
PMM0356	-3.53	1.76E-04	Alpha/beta hydrolasefold:Esterase/lipase/thioesterase family
PMM0365	2.98	4.88E-04	possible DsrE-like protein
PMM1421	3.29	5.61E-04	possible Gibberellin regulated protein
PMM1400	-4.01	6.55E-04	possible Hemagglutinin-neuraminidase
PMM1416	2.50	8.32E-04	conserved hypothetical protein
PMM1395	3.84	1.65E-03	hypothetical protein
PMM0367	2.21	3.52E-03	conserved hypothetical protein
PMM0691	-2.50	3.58E-03	conserved hypothetical protein
PMM1206	-1.65	3.67E-03	hypothetical protein
PMM1402	-2.12	1.08E-02	hypothetical protein
PMM0364	2.44	1.15E-02	conserved hypothetical protein
PMM1229	2.07	1.33E-02	Dehydrogenase, E1 component
PMM0374	-1.42	1.42E-02	mttA/Hcf106 family
PMM0355	-2.00	1.68E-02	conserved hypothetical protein
PMM1124	2.30	1.81E-02	conserved hypothetical protein
PMM1420	3.35	2.00E-02	possible Fumarate reductase subunit D
PMM1412	-1.79	2.54E-02	conserved hypothetical protein
PMM0363	2.26	2.58E-02	possible MarR family
PMM1242	1.95	3.24E-02	hypothetical protein
PMM0378	-1.73	3.26E-02	conserved hypothetical protein
PMM1131	-1.95	3.32E-02	conserved hypothetical protein
PMM1243	2.17	3.37E-02	possible methyltransferase
PMM1259	1.95	3.38E-02	pyridoxal-phosphate-dependent aminotransferase
PMM1248	1.66	3.42E-02	hypothetical protein
PMM1126	2.26	4.77E-02	Domain of unknown function DUF33

Table S2. Sequence information for genomes and environmental genome fragments (fosmid/BAC clones) presented here.

Genome	Size (bp)	# CDS	%GC	Percent shared with other HL genome
MIT9312	1709204	1811	31.2	84%[1]/87%[2]
MED4[3]	1657990	1713	30.8	86%[1]/92%[2]

Fosmid/BAC clone	Size (bp)	# CDS	%GC	Region of MED4 covered[4]
ASNC1092	35772	50	33	PMM0358-0399
ASNC1363	37759	43	31	PMM1165-1197+
ASNC1430	38037	63	33	PMM1428-1443+
ASNC2150	37884	38	35	PMM0290-0327
ASNC2259	35161	51	31	PMM0388-0401+
ASNC2388	37730	49	32	PMM1150-1190
ASNC3046	33332	55	31	PMM1132-1146+
ASNC612	41637	48	32	PMM1140-1178
ASNC729	38521	43	34	PMM0311-0352
HF10-110B11	42294	40	32	PMM1049-1088
HF10-11A3_c2	33701	30	32	PMM0589-0618
HF10-11D6	34867	42	32	PMM1439-1480
HF10-11H11	37311	38	31	PMM1560-1596
HF10-11H7	34000	35	32	PMM0598-0632
HF10-88D1	37878	39	32	PMM0756-0793
HF10-88F10	32924	37	32	PMM0391-0436
HF10-88G4	36389	37	33	PMM1269-1305
HF10-88H10	42262	44	33	PMM1588-1630
HF10-88H9	34302	40	32	PMM1315-1353
HOT0M-10B5	30968	30	32	PMM1583-1610
HOT0M-10D2	28741	32	32	PMM0436-0466
HOT0M-10E12	18233	19	33	PMM1665-1682
HOT0M-10G7	23045	32	30	PMM0639-0656+
HOT0M-1A11	23461	23	32	PMM1322-1344
HOT0M-2B5	35254	46	32	PMM1135-1165
HOT0M-3E5	19926	18	32	PMM1688-1704
HOT0M-5C8	20051	20	33	PMM1504-1524
HOT0M-6C1	21832	22	32	PMM1334-1355
HOT0M-7B6	11735	14	31	PMM0897-0910
HOT0M-7C8	12007	13	34	PMM0315-0327
HOT0M-8C8	20952	35	35	PMM1527-1561
HOT0M-8E2	30260	35	33	PMM0549-0583
HOT0M-8F9	27928	21	32	PMM1694-1714
HOT0M-8G12	20915	22	32	PMM0261-0282
HOT0M-9F4	43714	57	33	PMM1525-1581
HOT0M-9H9	24039	25	31	PMM0127-0148+
median for all fosmid clones	*33851*	*37*	*32.0*	

[1] percent shared based on MUMmer whole genome alignment.

[2] percent of predicted protein coding genes shared with other HL genome.

[3] Rocap et al. 2003

[4] Indicates the first and last gene of each fosmid/BAC, given as MED4 orthologs.

+ indicates that fosmid/BAC genes continue past this range but do not match the corresponding MED4 genes.

Table S3. Comparison of islands and small variable regions between MED4 and MIT9312.

Name Approximate Location	Length (kb)	Total number of genes	Number of non-shared genes	notes	Name Approximate Location	Length (kb)	Total number of genes	Number of non-shared genes	notes
Major Islands									
		MED4					**MIT9312**		
MED4-ISL1 334900-368300	33.4	36	25	flanked by 2 tRNA; cyanate transporter and lyase; restriction-modification genes; Range: PMM0351-0386.	MIT9312-ISL1 346000-365400	19.4	17	16	flanked by 2 tRNA; DNA modification genes (nucleases, methylase, integrase); Range: Pmt9312_0366-Pmt9312_0382.
MED4-ISL2 652800-663100	10.3	12	3	No tRNA; 2 *hli* genes; Range: PMM0684-0695.	MIT9312-ISL2 646400-659700	13.3	21	10	No tRNA; 4 *hli* genes; DNA recombination protein and ligase; Range: Pmt9312_0685-Pmt9312_0694.
MED4-ISL3 1071500-1082200	10.7	9	6	flanked by tRNA-Pro2; 1 *hli* inside, 1 *hli* in flanking 5' region, 1 *hli* in flanking 3' region;repeats; 6 proteins with predicted transmembrane domains. Range: PMM1123-1131.	MIT9312-ISL3 1049100-1141300	92.2	90	71	flanked by tRNA-Pro2; 5 transporters; 2 *hli* genes inside region, 1 *hli* gene in flanking 5' region, 3 *hli* genes in flanking 3' region; many repeats; 2nd copy of *psbA;* retron-type RT; T4pilus/T2SS genes. Range: Pmt9312_1134-Pmt9312_1223.
MED4-ISL4 1141700-1216200	74.5	66	52	Flanked by tRNA-Arg at 5' end, also contains tRNA-Ala near 3' end; LPS genes; Range: PMM1196-1261	MIT9312-ISL4 1203200-1274000	70.8	60	48	Flanked by tRNA-Arg at 5'end, also contains tRNA-Ala and pseudo-tRNA; cell surface modification genes; Range: Pmt9312_1297-Pmt9312_1355.
MED4-ISL5 1323000-1367600	44.6	53	25	No tRNA; 8 hli; Range: PMM1375-1427	MIT9312-ISL5 1382100-1419800	37.7	49	30	No tRNA; 8 *hli* , *phoH* ; Range: Pmt9312_1472 Pmt9312_1520.
Smaller Variable Regions and Indels									
		MED4					**MIT9312**		
(PMM0309-tRNA-Phe1)	na	na	na	na	296700-300800	4.1	4	2	flanked by tRNA-Phe; Range: Pmt9312_0312-Pmt9312_0315.
681200-685700	4.5	8	7	Transporter, transcriptional regulator; Range: PMM0715-0722.	(Pmt9312_0727-Pmt9312_0728)	na	na	na	na
767200-776400	9.2	12	6	Flanked by tRNA-Met; 4 *hli* genes; flanks inversion; Range: PMM0809-0820.	759200-763000	3.8	6	2	flanks inversion; Range: Pmt9312_0818-Pmt9312_0823.
814400-820300	5.9	4	0	Flanked by tRNA-Met, also includes tRNA-Thr; all 4 genes are elsewhere in MIT9312; Range: PMM0858-0861.	874000-876500	2.5	3	1	flanked by tRNA-Met; Range: Pmt9312_0940-Pmt9312_0942.
962100-966500	4.4	6	3	Low similarity region; Range: PMM1014-1019.	943300-947100	3.8	3	0	Low similarity region; Range: Pmt9312_1027-Pmt9312_1029.
1112500-1117200	4.7	3	2	tRNA-Ser; *idiA* iron transport protein (shared with MIT9312); two cons. hyp.; Range: PMM1162-1164.	1172800-1177500	4.7	4	3	tRNA-Ser; *idiA* (shared with MED4), *piuC,* cons. hyp., porin; Range: Pmt9312_1261-Pmt9312_1264.

CHAPTER THREE

Phosphate acquisition genes in *Prochlorococcus* ecotypes: Evidence for genome-wide adaptation

Adam C. Martiny*, Maureen L. Coleman*, and Sallie W. Chisholm

*Co-first authors.

Reprinted with permission from *Proceedings of the National Academy of Sciences*

Martiny, A.C.*, Coleman, M.L.* and Chisholm, S.W. (2006) Phosphate acquisition genes in *Prochlorococcus* ecotypes: evidence for genome-wide adaptation. *Proc. Nat. Acad. Sci.* 103:12552-12557.

Phosphate acquisition genes in *Prochlorococcus* ecotypes: Evidence for genome-wide adaptation

Adam C. Martiny*, Maureen L. Coleman*, and Sallie W. Chisholm†

Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139

Edited by Rita R. Colwell, University of Maryland, College Park, MD, and approved June 28, 2006 (received for review February 20, 2006)

The cyanobacterium *Prochlorococcus* is the numerically dominant phototroph in the oligotrophic oceans. This group consists of multiple ecotypes that are physiologically and phylogenetically distinct and occur in different abundances along environmental gradients. Here we examine adaptations to phosphate (P) limitation among ecotypes. First, we used DNA microarrays to identify genes involved in the P-starvation response in two strains belonging to different ecotypes, MED4 (high-light-adapted) and MIT9313 (low-light-adapted). Most of the up-regulated genes under P starvation were unique to one strain. In MIT9313, many ribosomal genes were down-regulated, suggesting a general stress response in this strain. We also observed major differences in regulation. The P-starvation-induced genes comprise two clusters on the chromosome, the first containing the P master regulator *phoB* and most known P-acquisition genes and the second, absent in MIT9313, containing genes of unknown function. We examined the organization of the *phoB* gene cluster in 11 *Prochlorococcus* strains belonging to diverse ecotypes and found high variability in gene content that was not congruent with rRNA phylogeny. We hypothesize that this genome variability is related to differences in P availability in the oceans from which the strains were isolated. Analysis of a metagenomic library from the Sargasso Sea supports this hypothesis; most *Prochlorococcus* cells in this low-P environment contain the P-acquisition genes seen in MED4, although a number of previously undescribed gene combinations were observed.

genome evolution | microarrays | phoB

The oceans play a key role in global nutrient cycling and climate regulation. The unicellular cyanobacterium *Prochlorococcus* is a significant contributor to these processes, because it accounts for ≈30% of primary productivity in midlatitude oceans (1). *Prochlorococcus* is composed of closely related physiologically distinct cells, enabling proliferation of the group as a whole over a broad range of environmental conditions (2). Early observations revealed that there are two genetically and physiologically distinct types of *Prochlorococcus*, high-light (HL) and low-light (LL)-adapted (2), which are distributed differently in the water column (3, 4). Cells belonging to these two groups differ not only in light optima and pigmentation (5) but also in nitrogen (6) and phosphorus (7) utilization capabilities, presumably adaptations that are related to depth-dependent nutrient concentrations.

The HL and LL groups can be further divided into at least six clades (two HL- and four LL-adapted) based on the phylogeny of the 16S/23S rRNA internal transcribed spacer region (8). The relative abundance of cells belonging to these clades has been measured in several ocean regions, revealing patterns that agree, for the most part, with their HL/LL phenotype: HL-adapted cells dominate the surface mixed layer, and LL-adapted cells most often dominate in deeper waters (3, 9–12). By combining physiological studies of isolates and clade abundance in the ocean, it was recently shown that temperature, in addition to light, is an important determinant of the ocean-scale abundance of these six phylogenetic clades (12). Based on the observed correlations between phylogenetic origin, physiological properties, and environmental distributions, these six clades are considered ecotypes, i.e., distinct phylogenetic clades with ecologically relevant physiological differences (2, 13).

A closer examination of physiological properties among cultured isolates reveals variability that is not consistent with their phylogenetic relationships. For example, some LL-adapted strains can use nitrite as sole nitrogen source, whereas others require ammonium (6). Moreover, one HL-adapted strain (MED4) can grow on organic phosphates as a sole phosphorus source, whereas another (MIT9312) and a LL-adapted strain (MIT9313) cannot (7). Thus strains with similar temperature and light optima for growth can vary in nutrient assimilation capabilities. This implies that nutrient adaptation has occurred more recently than adaptation to light and temperature gradients. One mechanism for rapid adaptation to a specific environment is the acquisition of genes by lateral transfer. Indeed, several key genes involved in nutrient assimilation in *Prochlorococcus* are thought to be of foreign origin (13), and we have recently identified variable genomic islands in *Prochlorococcus*, thought to have arisen by lateral gene transfer (14), that contain a number of genes involved in nutrient assimilation.

To better understand the relationship between variability in nutrient acquisition mechanisms, phylogeny, and light adaptation, we undertook a detailed analysis of phosphate (P) acquisition in *Prochlorococcus*. We first identified P-starvation-induced genes in HL- and LL-adapted isolates using DNA microarrays. Having identified these genes, we then analyzed their distribution among the genomes of 11 phylogenetically diverse *Prochlorococcus* strains. Finally, we compared these findings with the collective P-acquisition gene content of a natural *Prochlorococcus* population from the surface waters of the Sargasso Sea, which is periodically P-limited.

Results and Discussion

Identification of Differentially Expressed Genes Under P Starvation. To determine genes involved in the P-starvation response in *Prochlorococcus*, we subjected strains MED4 (HL-adapted) and MIT9313 (LL-adapted) to abrupt P limitation and monitored changes in gene expression. To initially map the time course of the response, we used quantitative RT-PCR to measure expression levels of *pstS*, which encodes a periplasmic P-binding protein known to be induced under P-limiting conditions in many cyanobacteria, including MED4 (15). The temporal profile of the P-starvation response differed significantly between the two strains. In MED4, the transcript level of *pstS* began to increase 12 h after cells were resuspended in P-free medium (Fig. 1*A*) and

Conflict of interest statement: No conflicts declared.

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: HL, high-light; LL, low-light; P, phosphate.

Data deposition: Orthologs to genes in the MED4 *phoB* region reported in this paper have been deposited in the GenBank database (accession nos. DQ786954–DQ787011 and DQ856305–DQ856313).

*A.C.M. and M.L.C. contributed equally to this work.

†To whom correspondence should be addressed. E-mail: chisholm@mit.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0601301103

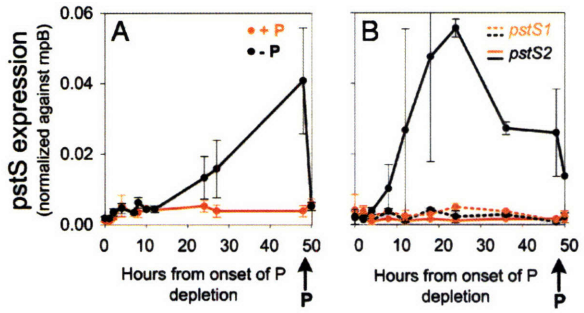


Fig. 1. Time course of expression of *pstS* in *Prochlorococcus* cells resuspended in medium with no added P at 0 h (black lines), compared to cells resuspended in P-replete medium (orange lines). Arrows indicate P addition after 48 h. (*A*) MED4: *pstS* is ORF PMM0710. (*B*) MIT9313: *pstS1* is ORF PMT0508 (dashed lines), and *pstS2* is ORF PMT0993 (solid lines).

increased steadily until P was added at 48 h. This release from P starvation caused a rapid decline in transcript level, which reached the control value within 2 h. In MIT9313, which has two copies of *pstS*, the expression of one (*pstS1*) was unresponsive to P starvation, whereas that of the other (*pstS2*) was elevated 50-fold by 24 h (Fig. 1*B*), followed by a decline. The addition of P to the medium after 48 h appeared to accelerate this decrease. Despite 94% amino acid sequence identity between the two copies of *pstS* in MIT9313, the genes responded very differently to P starvation. The function of *pstS1* is unknown.

We next examined genome-wide differences in gene expression in response to P starvation between the two strains. In MED4, a progressive induction of genes was observed over 48 h after the cells were resuspended in P-free medium. Thirty genes were significantly up-regulated, and four were down-regulated, by 48 h (Fig. 2*A*; Table 1, which is published as supporting information on the PNAS web site). The general response was different in MIT9313, where 176 genes were differentially expressed after 24 h, but most (143) were down-regulated (Fig. 2*B* and Table 2, which is published as supporting information on the

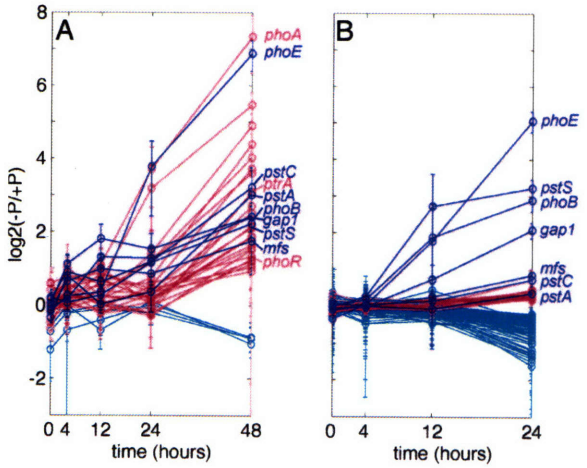


Fig. 2. Time course of gene expression in P-starved *Prochlorococcus* cultures. Differentially expressed genes ($q < 0.05$) in MED4 (*A*) and MIT9313 (*B*). Dark-blue lines indicate genes that were up-regulated in both strains, magenta lines are genes up-regulated in only one strain, and light-blue lines are genes down-regulated in only one strain. Error bars represent one standard deviation of fold change.

PNAS web site). The high fraction of down-regulated genes, including many ribosomal proteins, could indicate a general reduction in the metabolic rate of MIT9313 cells (16).

Only seven up-regulated genes were common to both strains (blue lines with gene names in Fig. 2). Most are orthologs to *Escherichia coli* genes implicated in P scavenging, such as the response regulator (*phoB*) and the transport system for orthophosphate (*pstABCS*). A porin gene located just downstream from *phoB* (PMM0709 in MED4 and PMT0998 in MIT9313) was also induced in both strains, and we propose that this gene encodes *phoE*, which is known to facilitate transport of orthophosphate across the outer membrane in other organisms. In addition to known P-starvation genes, genes previously unassociated with P starvation were up-regulated in both strains (Fig. 2 and Tables 1 and 2). Only two of these genes were common to both MED4 and MIT9313: *gap1*, which encodes glyceraldehyde-3-phosphate dehydrogenase, and *mfs*, which encodes a major facilitator superfamily transporter. Both genes are located just downstream from *phoB*, suggesting they play an important but unknown role in the P-starvation response, as has been suggested (17).

A number of orthologs to genes involved in the P-starvation response in other bacteria (18) were not induced in either *Prochlorococcus* strain, including *phoH* (whose function is unknown) and phosphonate transport genes (*phnCDE*). The lack of an identifiable phosphonatase or C-P lyase gene suggests that *phnCDE* encode a transport system for a different substrate in *Prochlorococcus* or may be nonfunctional. Also, genes encoding polyphosphate utilization (*ppK* and *ppX*) did not respond to P starvation in either strain of *Prochlorococcus*, although they are known to respond in some bacteria (19).

Despite similarities between the responses of MED4 and MIT9313, there were also important differences. MIT9313 lacks an ortholog to the most highly up-regulated gene in MED4, *phoA*, encoding alkaline phosphatase, which cleaves P from organic compounds. *ptrA*, which encodes a transcription factor thought to be involved in the P-starvation response (17), is up-regulated 8-fold in MED4 (PMM0718), whereas MIT9313 carries only a remnant of this gene (between PMT0998 and -999) that is not expressed. Similarly, MIT9313 carries a pseudogene of the sensor kinase *phoR* (17), which was not up-regulated, whereas the intact version of this gene was up-regulated in MED4. Despite the absence of *phoR* expression, both *phoB* and *pstABCS*, which normally depend on *phoR*, were induced under P starvation in MIT9313. Several regulatory genes that do not have orthologs in MED4 (PMT0265, PMT1357, and PMT2151) were differentially expressed in MIT9313 (Table 2), and these may be involved in activating *phoB* and in turn *pstABCS*. The remaining differentially expressed genes are unique to either strain and are primarily of unknown function. They should be further examined as potentially important for shaping the ecotype-specific response to P starvation.

The genes that are differentially expressed under P starvation are not distributed randomly along the chromosomes of the two strains (Fig. 3, $P < 0.0001$). Fifteen are located in a 21-gene stretch of the genome in MED4 (PMM0705–PMM0725), which includes *phoB*, most of the known P-acquisition genes, and several transporters. MIT9313 lacks intact orthologs to eight of these 15 genes, but most of the remaining seven are similarly located in the "*phoB* region." In addition, MED4 contains a second cluster of up-regulated genes located between PMM1403 and PMM1416, which is part of a variable genomic island (14). This organization suggests that the gene cluster around *phoB* is involved in the uptake of various forms of P, whereas the second cluster encodes an unknown component of the P-starvation response.

MICROBIOLOGY

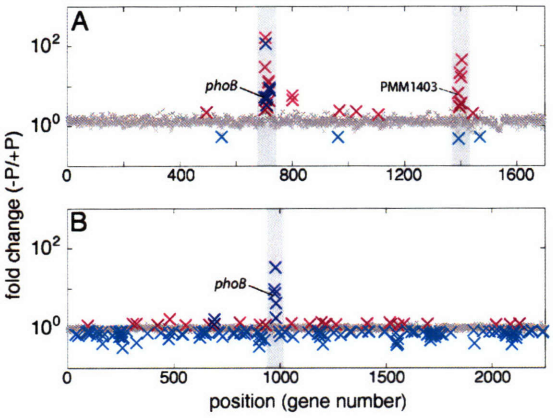


Fig. 3. Genome position of genes that were differentially expressed under P starvation in MED4 (*A*) and MIT9313 (*B*). The color code is the same as for Fig. 2 for differentially expressed genes; gray indicates genes with no significant ($q < 0.05$) change. The data plotted are from the 48-h time point in MED4 and the 24-h time point in MIT9313, the time of maximal *pstS* expression in each strain.

Genome Content and Organization of P-Acquisition Genes. Genes that are differentially expressed in response to P starvation in MED4 and MIT9313 were more likely to be lost or gained than randomly selected genes ($P < 0.0001$) in the genomes of 11 *Prochlorococcus* strains. In particular, genes found in the *phoB* region in MED4 are often missing or rearranged in the other genomes (Fig. 4*A*). Some strains (MED4, NATL1A, NATL2A, MIT9312, and MIT9301) share many orthologs with MED4, similarly grouped in a large cluster. In contrast, MIT9303, MIT9313, SS120, MIT9211, MIT9515, and AS9601 harbor fewer than half the *phoB* region genes found in MED4, and many of these are scattered throughout the genome.

This variability in genome content and architecture of P-acquisition genes is not related to phylogeny, as defined by rRNA sequence divergence (Fig. 4 *A* and *B*). Two HL-adapted strains belonging to the eMED4 clade (MIT9515 and MED4) share 99.9% 16S rRNA sequence identity, yet MIT9515 lacks orthologs to 15 MED4 genes from the *phoB* region. Similarly, three strains belonging to the eMIT9312 clade (MIT9312, MIT9301, and AS9601; 99.9% 16S rRNA identity) differ in gene content and organization relative to the MED4 *phoB* region. In fact, MIT9312 is more similar to MED4 and AS9601 to MIT9515 in terms of P-acquisition gene content (Fig. 4*A*), which is the inverse of their rRNA similarity. Thus it is reasonably clear, even from this limited data set, that the organization of P-acquisition genes in *Prochlorococcus* strains is not dictated by phylogenetic origin.

Ordering the genomes by gene content and organization relative to the MED4 *phoB* region, as depicted in Fig. 4*A*, reveals patterns that suggest that P availability in the waters from which these strains were isolated could influence genome content. MED4, the strain with the most-expansive *phoB* region, was isolated from surface waters in the northwest Mediterranean Sea, where the P concentration is typically <100 nM and has been shown to limit growth of cyanobacteria (20, 21). NATL1A and NATL2A, which possess orthologs to most of the MED4 *phoB* region genes, came from surface waters in the central North Atlantic Ocean, where surface P levels were between 50 and 150 nM (22) at the time these strains were isolated. Conversely, the strains with the fewest orthologs to the *phoB* region in MED4 (AS9601, MIT9515, and MIT9211) were isolated from ocean regions with high surface P levels (>600 nM; refs. 23 and 24). The remaining five strains in Fig. 4*A* contain an intermediate number of orthologs relative to the *phoB* region in MED4. Although they were isolated from regions where P concentrations are either low (<100 nM throughout the euphotic zone in Sargasso Sea) or variable (Gulf Stream; refs. 25 and 26), all came from deep in the euphotic zone (between 90 and 135 m). Light is likely the primary limiting factor for growth at this depth, perhaps relaxing selective pressure on the P-acquisition system. Thus, we predict that in P-limited environments, cells will contain many P-acquisition genes, primarily in a cluster around *phoB*.

Frequency of *Prochlorococcus* P-Acquisition Genes in the Sargasso Sea. To test this hypothesis, we examined gene stoichiometries in surface waters of the Sargasso Sea (27), where the P concentration is extremely low (25, 26). Indeed, all genes from the MED4 *phoB* region were present at roughly one copy per *Prochlorococcus* genome in this population (Fig. 4*C*). This includes genes between PMM0717 and PMM0722, which are largely absent from the other genomes, including ones affiliated with eMIT9312, the ecotype dominating this wild *Prochlorococcus* population (based on internal transcribed spacer sequence analysis from this data set). The abundance of P-acquisition genes similar to those found in MED4, in a population dominated by eMIT9312 cells, further supports our hypothesis that the regional environment influences the P-acquisition gene content of *Prochlorococcus* cells.

We also analyzed the frequency of occurrence of orthologs to the second up-regulated cluster in MED4 (spanning PMM1403 to -1416; Fig. 3*A*) in the Sargasso Sea population. As mentioned previously, the cluster is present only in MED4 and is located in a variable genomic island. In the Sargasso Sea, most genes from this cluster were present in a ratio close to 0.5 compared to core genes (data not shown), indicating that some, but not all, *Prochlorococcus* genomes contained these genes (see also ref 14). We discovered genome fragments containing genes from this island in proximity to known P-acquisition genes commonly found around *phoB* (Fig. 4*D*). These fragments demonstrated physical linkage between PMM1406 and *phoBR*, PMM1416 and *phoA* and several other combinations. This association of genes from two separate P-starvation-induced clusters in the MED4 genome supports the importance of these genes in responding to P limitation.

In MIT9301 and in several genome fragments from the Sargasso Sea, we also saw an intriguing linkage between genes found in the *phoB* region of MED4 and phosphonate uptake genes (*phnCDE*; Fig. 4 *A* and *D*). It has been proposed that phosphonates are an important phosphorus resource in marine ecosystems (28), but efforts to grow *Prochlorococcus* on phosphonates as a sole P source have been unsuccessful thus far. The clustering of phosphonate uptake genes and genes up-regulated under P starvation suggests that some *Prochlorococcus* lineages may be capable of using this organic phosphorus source.

Adaptation to P Limitation in *Prochlorococcus*. Our analysis revealed genomic variation among *Prochlorococcus* isolates that is not consistent with their rRNA-based phylogenetic relationships. We propose that these differences are related, in part, to the nutrient regime from which the cells were isolated. However, other forces are likely shaping genome content as well, such as phages using outer membrane proteins (e.g., PhoE) as receptors (29), crosstalk between regulatory circuits (e.g., PhoBR; ref. 30), and limitation by other factors (e.g., light). Stochastic variation may also play a role.

Lateral gene transfer may explain the lack of correspondence between the gene complements of the strains and their phylogenetic relationships. The *pstS* gene is encoded in the genomes

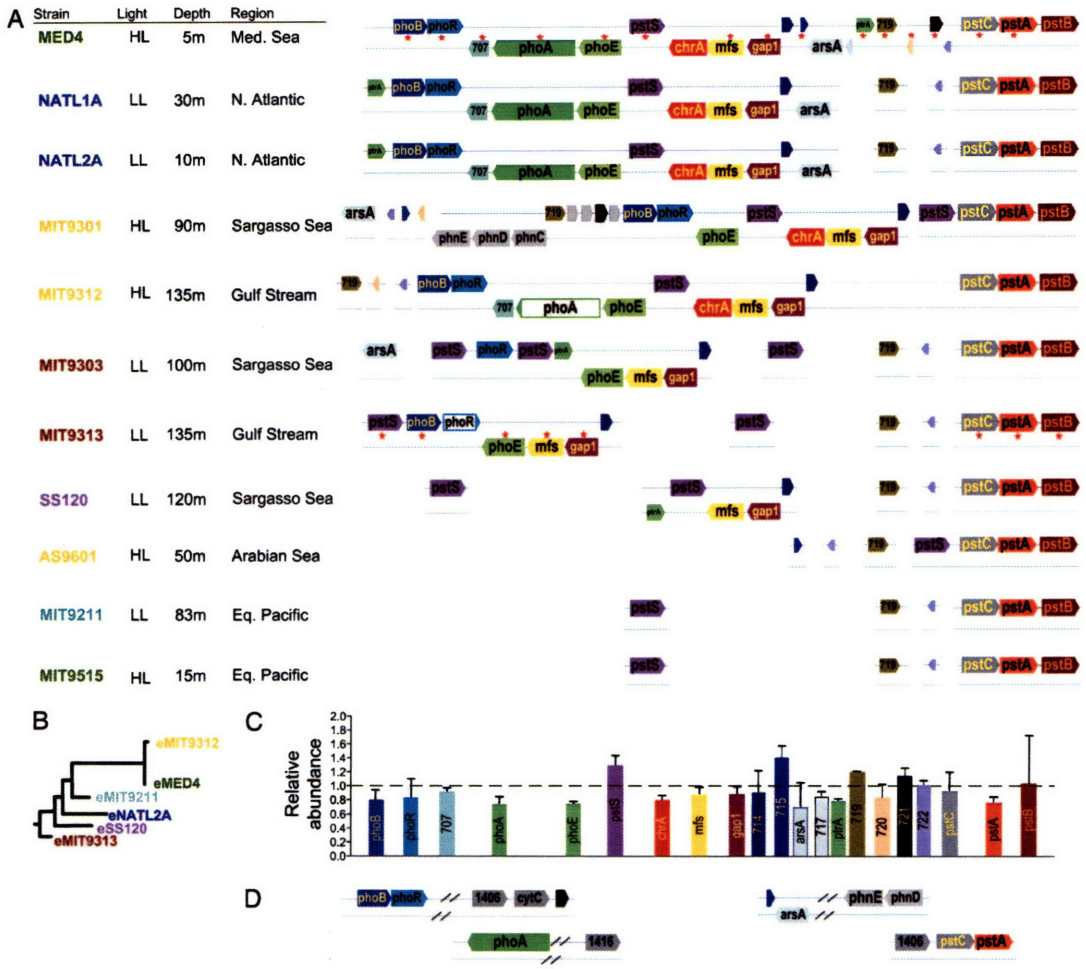


Fig. 4. P-acquisition genes in *Prochlorococcus*. (*A*) Genes located in proximity to *phoB* in MED4 (at the top) and the presence of their orthologs in the genomes of 11 *Prochlorococcus* strains. A red star indicates a gene that was significantly up-regulated in MED4 or MIT9313 from the microarray analyses. Gene numbers refer to PMM0XXX in MED4. Unfilled genes are likely pseudogenes. Color coding of strain names reflects ecotype affiliation shown in *B* (2). (*B*) Schematic of the phylogenetic relationship among different *Prochlorococcus* ecotypes (9). (*C*) Gene frequency in small insert libraries from the surface waters of the Sargasso Sea (27). Error bars indicate standard deviation of abundance based on all 150-bp fragments covering a gene. (*D*) Examples of genomic variants in the Sargasso Sea, showing linkage between genes found in the *phoB* region of MED4 and genes found elsewhere in the MED4 genome. Diagonal lines represent unknown sequence between two end reads of a clone in the data set.

of cyanophages that infect *Prochlorococcus* (31), suggesting a mechanism for moving genes across phylogenetic clades, and there is evidence that *phoA* and other genes involved in nutrient assimilation have been acquired laterally in some *Prochlorococcus* lineages (13). Furthermore, we observed that genes clustered in a variable genomic island in MED4 are up-regulated during P starvation (14). We were unable to detect any other obvious events of lateral gene transfer in the *phoB* region using phylogenetic analysis, but we anticipate that these events will become apparent as the sequences of more genomes from marine environments become available.

Unlike the P-starvation response, some traits, such as adaptations to light and temperature, are consistent with the phylogeny of *Prochlorococcus* (2, 12). One explanation for this difference is that photosynthesis requires a large protein complex that does not readily incorporate whole genes from foreign organisms (32, 33), and temperature adaptation can occur through genome-wide changes in amino acid and membrane lipid composition (34, 35). In contrast, the acquisition of a few key genes can rapidly change the spectrum of nutrient sources for a cell (e.g., nitrite reductase and alkaline phosphatase). A simplified calculation (see *Materials and Methods*) shows that if a *Prochlorococcus* cell acquires genes that improve growth rate by 1%, its progeny will dominate the entire population in an ocean basin in a few decades. This time scale is comparable to the observed domain shift in the North Pacific Ocean gyre from

a nitrogen- to a P-controlled state, purportedly fueled by increased nitrogen fixation in this region (36). Considering the strong feedback between the metabolic activity of *Prochlorococcus* (and all phytoplankton) and the local nutrient regime (37), understanding this type of genomic adaptation may be crucial for understanding shifts in biogeochemical processes in the oceans.

Materials and Methods

Culture Conditions. *Prochlorococcus* strains were grown at 22°C in Pro99 medium (6). Before the experiment, cultures were maintained in continuous light in log-phase growth at an irradiance of 12 μE m^{-2}·s^{-1} [E, einstein (1 mol of photons)] for MIT9313 (growth rate = 0.18 d^{-1}), and 30 μE·m^{-2}·s^{-1} for MED4 (growth rate = 0.27 d^{-1}) for >30 generations. Chlorophyll fluorescence was monitored on a Synergy HT fluorometer (BioTek, Burlington, VT).

P-Starvation Time Series. To induce P starvation, triplicate 4-liter cultures were harvested by centrifugation (10,000 × *g*), split in two, and washed twice in either P-replete (Pro99 with 50 μM PO_4) or -depleted (Pro99 with no added PO_4) medium and resuspended in 2 liters of the same medium. Samples were taken for RNA extraction, microarray hybridization, and quantitative RT-PCR (qRT-PCR) analysis at 0, 4, 12, 24, and 48 h after resuspension. Additional samples were taken for qRT-PCR at selected time points. After 48 h, 50 μM P was added to the P-depleted cultures to monitor the recovery response.

RNA Extraction. RNA was isolated according to ref. 38. In brief, cells were harvested by centrifugation (10,000 × *g*), resuspended in storage buffer (200 mM sucrose/10 mM NaOAc, pH 5.2/5 mM EDTA) and stored at −80°C. Before RNA extraction, MIT9313 cells were treated with 10 μg/μl lysozyme (Sigma, St. Louis, MO) for 1 h at 37°C (39). Total RNA was extracted by using the mirVana miRNA kit (Ambion, Austin, TX). DNA was removed by using Turbo DNase (Ambion). RNA was concentrated by ethanol precipitation and resuspended in milli-Q water.

Quantitative RT-PCR. RNA (2–10 ng of total RNA) was reverse-transcribed by using 100 units of SuperScript II (Invitrogen, Carlsbad, CA) in the presence of 200 units of SuperaseIN (Ambion). Primers are described in Table 3, which is published as supporting information on the PNAS web site. The resulting cDNA was diluted 5-fold in 10 mM Tris, pH 8. Triplicate real-time PCRs were performed by using the Qiagen (Valencia, CA) SYBR green kit and the diluted cDNA as template. The following program was run on an MJ Research (Cambridge, MA) Opticon DNA engine: 15 min at 95°C, followed by 40 cycles of denaturation (95°C, 15 s), annealing (56°C, 30 s), and extension (72°C, 30 s), followed by 5 min at 72°C. cDNA for *pstS* was quantified relative to *rnpB* by using the Δ-Δ C_T method (40).

Array Analysis. cDNA synthesis, labeling, and hybridization onto custom MD4–9313 Affymetrix (Santa Clara, CA) microarrays was done following the standard Affymetrix protocol. The probe arrays were scanned, and data visualization was done with GeneSpring software (Version 7.1; Silicon Genetics, Palo Alto, CA). Normalization was done by using the Robust Multichip Average algorithm (41) implemented in GeneSpring. Bayesian statistical analysis was applied to identify differentially expressed genes using Cyber-T (42). The Bayesian estimate of variance, which incorporates both the experimental variance for a given gene and variance of genes with similar expression levels (42), was calculated by using window sizes of 81 for MED4 and 101 for MIT9313 and a confidence value of 10 for both strains. A *t* test was then performed on log-transformed expression values by using the Bayesian variance estimate. To account for the multiple *t* tests performed, we used the program QVALUE, which measures significance in terms of the false discovery rate (43). A gene was identified as differentially expressed if the *q* value was <0.05. Signal intensities of individual probes targeting intergenic regions and potential miscalled ORFs were extracted by using Intensity Mapper (Affymetrix).

Tests for Clustering and Selective Loss/Gain of Induced Genes. We tested whether differentially expressed genes were distributed randomly along the genome by comparing the gene distance (in base pairs) against a simulated random distribution of genes. The weighted gene distance (*d*) was calculated by using the following decay function (adjusted for a circular genome):

$$d = \sum_i \text{sort} \sum_j \frac{1}{j} (n_i - n_{j+l}), \qquad [1]$$

where $i, j = 1, 2, \ldots$, number of expressed genes, and n = position in genome. The second summation is based on a sorted array to nearest neighbor of n_i (i.e., $n_i - n_1 = 0$). The physical distance between differentially expressed genes was then compared to the *d* value of *i* randomly selected genes (10,000 permutations). We also tried other decay functions (e.g., different log bases of $n_i - n_j$) as well as using gene order as a measure for distance instead of actual base-pair difference, but all summations yielded the same result.

We also tested whether differentially expressed genes in MED4 (34 genes) and MIT9313 (176 genes) were more commonly lost or gained compared to randomly selected genes in the other *Prochlorococcus* genomes. We randomly chose 34 genes in the MED4 genome, counted the total number of orthologs to these 34 genes in the other 10 genomes, and repeated this process 10,000 times to generate a distribution. We then tested whether the total number of orthologs of the 34 differentially expressed genes in MED4 fell significantly outside this distribution. We repeated the test using the 176 differentially expressed MIT9313 genes. Orthologs were identified as pairwise best blastp hits. To further support the ortholog assignments, we constructed phylogenetic trees (maximum parsimony) for each gene in the MED4 *phoB* region and its putative orthologs.

Blast Analysis of Sargasso Sea Shotgun Library. We examined the occurrence of genes found in the *phoB* region of MED4 (between PMM0705 and PMM0725), in the Sargasso Sea environmental sequence data set sampled in February 2003 (excluding samples 5, 6, and 7; ref. 27). We used MED4 as the template for PMM0715 to PMM0722 and MIT9312 for the remaining genes. A sliding window of 150-bp fragments (step length = 50 bp) from the *phoB* region was first searched (blastn or tblastx; ref. 44) against the environmental sequence data set. A positive hit was scored if the environmental sequence and the paired end recovered *Prochlorococcus* as best hit when searched against a database consisting of *Prochlorococcus*, marine *Synechococcus* (WH8102, CC9905, and CC9902), *Pelagibacter ubique*, and *Silicibacter pomeroyi*. The number of copies of a particular *phoB*-region gene in the Sargasso Sea data set was estimated by averaging the number of hits for 150-bp segments comprising that gene and normalized against the average occurrence of known single-copy genes in all sequenced Cyanobacteria: *cpeA*, *glnA*, *gyrB*, *hemA*, 16S/23S internal transcribed spacer region (single copy in HL *Prochlorococcus* clades), *recA*, *rpl10*, *rpoB*, *rpsD*, and *tyrS*.

Changes in Genotype Frequency as a Function of Relative Fitness. To calculate how long it might take a new genotype with slightly improved fitness to overtake a population of *Prochlorococcus* cells in an ocean, we used equation **11** from ref. 45:

$$\ln[x_1(t)/x_2(t)] = \ln[x_1(0)/x_2(0)] + st, \qquad [2]$$

where $x_1(t)$ is the fraction of the new genotype, and $x_2(t)$ is the fraction of the ancestral genotype at time t (days). At $t = 0$, x_1 was set to 10^{-24}, and x_2 was set at 1, assuming 10^{24} cells in an ocean basin such as the Sargasso Sea (46). We assumed a growth rate of 0.5 per day^{-1} (47) for the ancestral genotype and an increase in growth rate (or relative fitness) of new genotype (s) of 1%, so $s = 0.005$ d^{-1}.

We thank Debbie Lindell for many helpful discussions and Robert Steen and Trent Rector at Harvard Biopolymer Facility for labeling RNA and hybridizing the microarrays. We also thank numerous members of the Chisholm and DeLong labs for helpful comments on the manuscript. This work was supported in part by a fellowship from the Danish National Science Foundation (to A.C.M.); a National Science Foundation Graduate Fellowship (to M.L.C.); and grants from the National Science Foundation, the Gordon and Betty Moore Foundation, and the U.S. Department of Energy GTL Program (to S.W.C.).

1. Goericke, R. & Welschmeyer, N. A. (1993) *Deep-Sea Res.* **40,** 2283–2294.
2. Moore, L. R., Rocap, G. & Chisholm, S. W. (1998) *Nature* **393,** 464–467.
3. West, N. J. & Scanlan, D. J. (1999) *Appl. Environ. Microbiol.* **65,** 2585–2591.
4. Urbach, E. & Chisholm, S. W. (1998) *Limnol. Oceanogr.* **43,** 1615–1630.
5. Moore, L. R. & Chisholm, S. W. (1999) *Limnol. Oceanogr.* **44,** 628–638.
6. Moore, L. R., Post, A. F., Rocap, G. & Chisholm, S. W. (2002) *Limnol. Oceanogr.* **47,** 989–996.
7. Moore, L. R., Ostrowski, M., Scanlan, D. J., Feren, K. & Sweetsir, T. (2005) *Aquat. Microbial Ecol.* **39,** 257–269.
8. Rocap, G., Distel, D. L., Waterbury, J. B. & Chisholm, S. W. (2002) *Appl. Environ. Microbiol.* **68,** 1180–1191.
9. Ahlgren, N. A., Rocap, G. & Chisholm, S. W. (2006) *Environ. Microbiol.* **8,** 441–454.
10. Zinser, E. R., Coe, A., Johnson, Z. I., Martiny, A. C., Fuller, N. J., Scanlan, D. J. & Chisholm, S. W. (2006) *Appl. Environ. Microbiol.* **72,** 723–732.
11. West, N. J., Schonhuber, W. A., Fuller, N. J., Amann, R. I., Rippka, R., Post, A. F. & Scanlan, D. J. (2001) *Microbiology* **147,** 1731–1744.
12. Johnson, Z. I., Zinser, E. R., Coe, A., McNulty, N. P., Woodward, E. M. & Chisholm, S. W. (2006) *Science* **311,** 1737–1740.
13. Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N. A., Arellano, A., Coleman, M., Hauser, L., Hess, W. R., *et al.* (2003) *Nature* **424,** 1042–1047.
14. Coleman, M. L., Sullivan, M. B., Martiny, A. C., Steglich, C., Barry, K., Delong, E. F. & Chisholm, S. W. (2006) *Science* **311,** 1768–1770.
15. Scanlan, D. J., Silman, N. J., Donald, K. M., Wilson, W. H., Carr, N. G., Joint, I. & Mann, N. H. (1997) *Appl. Environ. Microbiol.* **63,** 2411–2420.
16. Nomura, M. (1999) *J. Bacteriol.* **181,** 6857–6864.
17. Scanlan, D. J. & West, N. J. (2002) *FEMS Microbiol. Ecol.* **40,** 1–12.
18. Wanner, B. L. (1996) in *Escherichia coli and Salmonella: Cellular and Molecular Biology*, ed. Neidhardt, F. C. (Am. Soc. Microbiol., Washington, DC), pp. 1357–1381.
19. Kornberg, A., Rao, N. N. & Ault-Riche, D. (1999) *Annu. Rev. Biochem.* **68,** 89–125.
20. Marty, J. C., Chiaverini, J., Pizay, M. D. & Avril, B. (2002) *Deep-Sea Res.* **49,** 1965–1985.
21. Vaulot, D., LeBot, N., Marie, D. & Fukai, E. (1996) *Appl. Environ. Microbiol.* **62,** 2527–2533.
22. Irwin, B. (2000) *Nutrient Data from the Atlantic, JGOFS Canada Data Sets 1989–1998* (Marine Enviromental Data Service, Department of Fisheries and Oceans, Canada). CD-ROM Version 1.0.
23. Coale, K. H., Johnson, K. S., Fitzwater, S. E., Gordon, R. M., Tanner, S., Chavez, F. P., Ferioli, L., Sakamoto, C., Rogers, P., Millero, F., *et al.* (1996) *Nature* **383,** 495–501.
24. Morrison, J. M., Codispoti, L. A., Gaurin, S., Jones, B., Manghani, V. & Zheng, Z. (1998) *Deep-Sea Res.* **45,** 2053–2101.
25. Cavender-Bares, K. K., Karl, D. M. & Chisholm, S. W. (2001) *Deep-Sea Res.* **48,** 2373–2395.
26. Wu, J., Sunda, W., Boyle, E. A. & Karl, D. M. (2000) *Science* **289,** 759–762.
27. Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., *et al.* (2004) *Science* **304,** 66–74.
28. Karl, D. M. & Björkman, K. M. (2002) in *Biogeochemistry of Marine Dissolved Organic Matter*, eds. Hansell, D. A. & Carlson, C. A. (Academic, London), pp. 250–366.
29. Ho, T. D. & Slauch, J. M. (2001) *J. Bacteriol.* **183,** 1495–1498.
30. Fisher, S. L., Jiang, W., Wanner, B. L. & Walsh, C. T. (1995) *J. Biol. Chem.* **270,** 23143–23149.
31. Sullivan, M. B., Coleman, M. L., Weigele, P., Rohwer, F. & Chisholm, S. W. (2005) *PLoS Biol.* **3,** e144.
32. Blankenship, R. E. (1992) *Photosynth. Res.* **33,** 91–111.
33. Shi, T., Bibby, T. S., Jiang, L., Irwin, A. J. & Falkowski, P. G. (2005) *Mol. Biol. Evol.* **22,** 2179–2189.
34. Morgan-Kiss, R. M., Priscu, J. C., Pocock, T., Gudynaite-Savitch, L. & Huner, N. P. A. (2006) *Microbiol. Mol. Biol. Rev.* **70,** 222–252.
35. Tekaia, F., Yeramian, E. & Dujon, B. (2002) *Gene* **297,** 51–60.
36. Karl, D. M., Letelier, R., Tupas, L., Dore, J. E., Christian, J. & Hebel, D. V. (1997) *Nature* **388,** 533–538.
37. Redfield, A. C. (1934) in *James Johnstone Memorial Volume*, ed. Daniel, R. J. (Univ. Press of Liverpool, Liverpool, U.K.).
38. Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M. & Chisholm, S. W. (2005) *Nature* **438,** 86–89.
39. Tolonen, A. C., Aach, J., Lindell, D., Johnson, Z. I., Rector, T., Steen, R., Church, G. M. & Chisholm, S. W. (2006) *Mol. Syst. Biol.*, in press.
40. Livak, K. J. & Schmittgen, T. D. (2001) *Methods* **25,** 402–408.
41. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. (2003) *Bioinformatics* **19,** 185–193.
42. Baldi, P. & Long, A. D. (2001) *Bioinformatics* **17,** 509–519.
43. Storey, J. D. & Tibshirani, R. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 9440–9445.
44. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
45. Dykhuizen, D. E. & Hartl, D. L. (1983) *Microbiol. Rev.* **47,** 150–168.
46. Partensky, F., Hess, W. R. & Vaulot, D. (1999) *Microbiol. Mol. Biol. Rev.* **63,** 106–127.
47. Liu, H., Landry, M. R., Vaulot, D. & Campbell, L. (1999) *J. Geophys. Res.* **104,** 3391–3399.

Table 1. Gene expression summaries for MED4 genes identified as differentially expressed (q<0.05 at t=48). FC, fold change (-P/+P).

ORF	Description	FC t=0	q-value t=0	FC t=4	q-value t=4	FC t=12	q-value t=12	FC t=24	q-value t=24	FC t=48	q-value t=48
PMM0709	som possible porin	-1.133	0.5211	1.175	0.5285	1.274	0.5349	13.670	0.0000	117.337	0.0000
PMM1416	conserved hypothetical	-1.017	0.6962	1.020	0.6530	1.537	0.1640	9.086	0.0000	44.524	0.0000
PMM0708	phoA alkaline phosphatase	-1.055	0.6412	1.336	0.4679	2.142	0.0080	13.059	0.0000	162.295	0.0000
PMM0707	possible Lipoprotein	-1.381	0.2181	1.161	0.5223	1.094	0.7691	2.677	0.0053	29.907	0.0000
PMM0720	possible Poly A polymerase regulatory subunit	1.019	0.6806	-1.038	0.6097	-1.047	0.8067	1.431	0.1031	13.245	0.0000
PMM1414	unnamed	-1.070	0.6371	1.726	0.2221	1.506	0.1640	2.299	0.0004	16.334	0.0000
PMM0710	ABC transporter, substrate binding protein,phosphate	1.142	0.4626	1.041	0.6121	1.202	0.3275	2.232	0.0000	4.670	0.0000
PMM1409	possible Rubredoxin	-1.059	0.6305	-1.045	0.6328	1.130	0.7051	1.679	0.1864	21.031	0.0000
PMM0719	hypothetical	-1.289	0.3027	-1.034	0.6578	1.313	0.4148	2.257	0.0042	12.183	0.0000
PMM0723	pstC,phoW putative phosphate ABC transporter	1.104	0.5570	1.866	0.2065	1.536	0.1464	2.264	0.0051	9.222	0.0000
PMM0724	pstA,phoT putative phosphate ABC transporter	-1.275	0.2839	1.728	0.2691	1.012	0.8901	1.280	0.6342	8.034	0.0000
PMM0705	phoB two-component response regulator, phosphate	1.010	0.7034	2.158	0.1601	3.490	0.0000	2.882	0.0001	5.298	0.0000
PMM0713	gap1 Putative glyceraldehyde 3-phosphatedehydrogenase	-1.048	0.6416	1.109	0.5695	2.467	0.0011	2.424	0.0000	5.273	0.0001
PMM0806	Bacterial regulatory proteins, Crp family	1.343	0.3219	1.384	0.3502	1.638	0.1464	1.109	0.7591	4.461	0.0005
PMM1408	hypothetical	-1.125	0.5183	1.154	0.5204	1.173	0.5916	1.172	0.6342	2.941	0.0017
PMM1403	HNH endonuclease:HNH nuclease	-1.265	0.2876	1.315	0.3998	-1.159	0.5916	-1.038	0.8580	6.497	0.0018
PMM1415	possible UBX domain	-1.034	0.6805	2.042	0.2383	1.175	0.6484	1.109	0.7575	3.684	0.0020
PMM0712	multidrug efflux transporter, MFS family	-1.094	0.5830	1.555	0.3013	1.974	0.0171	1.805	0.0151	3.361	0.0022
PMM1407	possible SRP19 protein	1.004	0.7145	1.008	0.6778	1.312	0.3209	1.246	0.6215	3.122	0.0035
PMM0805	hypothetical	1.041	0.6281	1.506	0.3436	1.039	0.7812	1.316	0.6342	5.710	0.0036
PMM0706	phoR two-component sensor histidine kinase, phosphatesensing	-1.083	0.5734	1.241	0.4769	1.455	0.1464	1.513	0.4544	2.522	0.0084
PMM1411	hypothetical	-1.126	0.5183	1.041	0.6433	-1.139	0.7493	1.073	0.8482	3.629	0.0114
PMM0715	possible chorismate binding enzyme	-1.533	0.1344	1.337	0.3840	1.001	0.8901	-1.216	0.6342	2.871	0.0119
PMM0496	sigA, rpoD Putative principal RNA polymerase sigma factor	-1.065	0.6166	1.274	0.4769	2.031	0.0180	1.002	0.8786	2.194	0.0131
PMM1457	atp1 possible ATP synthase subunit 1	1.490	0.1084	1.056	0.6301	1.822	0.0134	1.369	0.3724	2.060	0.0131
PMM0718	possible Bacterial regulatory proteins, crp fa	-1.091	0.4626	1.265	0.2691	1.020	0.8655	-1.164	0.6342	8.517	0.0220
PMM0970	urtA putative urea ABC transporter, substrate bindingprotein	-1.657	0.0285	-1.160	0.3536	1.003	0.8832	-1.065	0.7814	-1.883	0.0256
PMM0975	conserved hypothetical protein	-1.027	0.7145	1.172	0.5625	1.076	0.7494	-1.153	0.7041	2.392	0.0348
PMM0721	possible Myosin N-terminal SH3-like domain	1.011	0.6962	1.052	0.5986	-1.143	0.6188	1.245	0.6215	4.356	0.0367
PMM1113	two-component response regulator	1.514	0.1381	1.693	0.0818	1.817	0.0180	1.276	0.4692	1.910	0.0374
PMM0551	rbcS, cbbS Ribulose bisphosphate carboxylase, small chain	-2.345	0.0068	-1.616	0.0733	-1.335	0.2073	1.015	0.8724	-1.842	0.0378
PMM1035	possible DNA gyrase/topoisomerase IV, subunit	1.056	0.6922	1.627	0.2707	1.586	0.0638	1.147	0.7422	2.307	0.0478
PMM1405	hypothetical	-1.524	0.2839	-1.012	0.4769	-1.617	0.1294	1.076	0.7814	-2.086	0.0478
PMM1482	hli3 possible high light inducible protein	-1.036	0.6500	-1.170	0.4646	-1.084	0.7578	-1.159	0.6342	-1.864	0.0486

Table 2. Gene expression summaries for MIT9313 genes identified as differentially expressed (q<0.05 at t=24). FC, fold-change (-P/+P).

ORF	Description	FC t=0	q-value t=0	FC t=4	q-value t=4	FC t=12	q-value t=12	FC t=24	q-value t=24	notes
PMT0994	phoB two-component response regulator, phosphate	1.013	0.9971	1.129	0.4487	3.661	0.0490	7.385	0.0000	
PMT0998	som possible porin	1.035	0.9971	1.017	0.9369	3.422	0.0789	33.238	0.0000	
PMT1000	Putative glyceraldehyde 3-phosphatedehydrogenase	-1.071	0.9971	-1.011	0.9591	1.661	0.0098	4.210	0.0000	
PMT0993	ABC transporter, substrate binding protein,phosphate	-1.009	0.9971	1.191	0.4553	6.577	0.0003	9.279	0.0000	
PMT0508	ABC transporter, substrate binding protein,phosphate	-1.060	0.9971	1.022	0.9329	2.349	0.0003	5.364	0.0000	1
PMT0999	multidrug efflux transporter, MFS family	1.067	0.9971	-1.012	0.9591	1.192	0.1749	1.808	0.0000	
PMT1572	conserved hypothetical protein	1.014	0.9971	-1.393	0.9221	-1.320	0.0174	-2.662	0.0000	
PMT1576	ABC transporter, ATP binding protein	1.111	0.9971	-1.105	0.9105	-1.129	0.8464	-2.458	0.0000	
PMT0169	possible Chitin synthase	1.288	0.9971	-1.259	0.5027	-1.190	0.0765	-1.536	0.0000	
PMT0702	pstC,phoW putative phosphate ABC transporter	-1.275	0.9971	1.033	0.9221	1.102	0.7149	1.656	0.0000	
PMT0915	Bacterial outer membrane protein	1.171	0.9971	-1.076	0.8135	-1.168	0.1484	-1.971	0.0000	
PMT0916	hypothetical	1.182	0.9971	1.008	0.9591	-1.170	0.7086	-2.846	0.0000	
PMT0995	two component sensor histidine kinase fragment,likely pseudogene	-1.083	0.9971	-1.021	0.9329	1.035	0.9187	1.386	0.0000	2
PMT2117	possible D12 class N6 adenine-specific DNA met	1.084	0.9971	1.054	0.9221	-1.159	0.8466	-2.553	0.0000	
PMT0923	hypothetical	1.074	0.9971	-1.115	0.6700	-1.085	0.8466	-1.858	0.0000	
PMT1940	possible Peptidase family C9	-1.003	0.9971	-1.157	0.9329	-1.081	0.9520	-2.426	0.0000	
PMT0265	two component sensor histidine kinase	-1.075	0.9971	-1.020	0.9329	1.009	0.9604	-1.645	0.0001	
PMT1571	hypothetical	-1.091	0.9984	-1.392	0.8409	-1.237	0.6797	-2.556	0.0001	
PMT0262	possible pilin	-1.052	0.9971	1.120	0.4568	1.002	0.9604	-3.064	0.0001	
PMT0258	hypothetical protein	-1.247	0.9971	1.130	0.5027	1.019	0.9604	-1.573	0.0001	
PMT0924	possible Carbamoyl-phosphate synthase L chain,	-1.034	0.9994	-1.075	0.9221	-1.040	0.9520	-1.425	0.0002	
PMT1223	hypothetical	1.100	0.9971	-1.064	0.9657	-1.037	0.9604	-2.553	0.0002	
PMT0168	possible Negative factor, (F-Protein) or Nef.	1.012	0.9971	-1.118	0.9591	-1.148	0.9520	-2.399	0.0004	
PMT0488	possible Integrase Zinc binding domain	1.128	0.9971	1.064	0.8681	1.030	0.9587	1.663	0.0004	
PMT1732	rpl3, rplC 50S ribosomal protein L3	1.132	0.9971	-1.178	0.8409	-1.321	0.2887	-2.020	0.0005	
PMT0794	possible Spectrin repeat	1.080	0.9971	1.053	0.8907	-1.023	0.9604	-1.409	0.0005	
PMT1789	S1 RNA binding domain:Ribonuclease E and G	1.073	0.9971	-1.025	0.9329	-1.065	0.7977	-1.475	0.0006	
PMT1753	rps13, rpsM 30S ribosomal protein S13	1.113	0.9971	-1.100	0.9161	-1.369	0.2149	-1.658	0.0008	
PMT2138	conserved hypothetical protein	-1.102	0.9971	-1.339	0.7608	-1.184	0.5650	-1.455	0.0008	
PMT0038	possible Penicillin amidase	1.002	0.9994	-1.080	0.8409	-1.274	0.1320	-1.416	0.0008	
PMT0546	probable periplasmic protein	1.024	0.9971	-1.049	0.8409	-1.063	0.8438	-1.412	0.0010	
PMT1733	rpl4, rplD 50S ribosomal protein L4	1.069	0.9971	-1.119	0.8409	-1.102	0.9187	-1.631	0.0010	
PMT0740	rps18, rpsR 30S Ribosomal protein S18	-1.074	0.9971	1.060	0.8907	-1.211	0.2190	-1.621	0.0010	
PMT0847	possible Photosystem I reaction centre subunitVI	-1.125	0.9971	1.033	0.9329	1.047	0.8830	-1.404	0.0010	
PMT0216	conserved hypothetical protein	-1.056	0.9971	-1.027	0.9329	-1.204	0.2149	-1.523	0.0014	
PMT0701	pstA,phoT putative phosphate ABC transporter	-1.038	0.9971	-1.001	0.9632	-1.049	0.9520	1.285	0.0014	
PMT1741	rps17, rpsQ 30S Ribosomal protein S17	1.010	0.9971	-1.091	0.9221	-1.172	0.7977	-1.518	0.0014	
PMT0251	possible Replication protein	1.030	0.9971	1.046	0.8409	-1.008	0.9604	-1.259	0.0018	
PMT0824	Domain of unknown function DUF143:Iojap-relatedprotein	1.051	0.9971	1.021	0.9329	1.043	0.9490	1.356	0.0018	
PMT0263	possible pilin	1.000	0.9971	1.029	0.9329	1.046	0.9520	-1.771	0.0020	
PMT0142	mreB Rod shape determining protein	1.165	0.9971	-1.041	0.9329	-1.125	0.7977	-1.483	0.0021	
PMT0328	possible bromodomain adjacent to zinc fingerdomain, 2B...	1.246	0.9971	1.002	0.9400	-1.119	0.9520	-2.284	0.0022	
PMT0102	Surface polysaccharide biosynthesis protein,possible cytidylyltransferase	1.053	0.9971	1.008	0.9591	-1.034	0.9520	-1.290	0.0025	
PMT0659	possible Sodium:sulfate symporter transmembrane	1.247	0.9971	1.007	0.9613	-1.054	0.8438	-1.255	0.0028	
PMT1716	DUF206	1.034	0.9971	-1.036	0.9329	-1.192	0.2887	-1.465	0.0029	
PMT1210	hypothetical	-1.034	0.9971	1.098	0.6398	-1.051	0.9520	-1.375	0.0033	
PMT0095	capsular polysaccharide biosynthesis protein	1.251	0.9971	1.082	0.6727	-1.054	0.8551	-1.254	0.0033	
PMT0943	hypothetical	1.101	0.9971	-1.028	0.9632	-1.015	0.9604	-1.974	0.0033	
PMT1754	rps11, rpsK 30S ribosomal protein S11	1.166	0.9971	-1.047	0.9221	-1.067	0.9520	-1.446	0.0033	
PMT0630	possible Gonadotropin-releasing hormone	-1.046	0.9971	1.016	0.9591	-1.079	0.7128	-1.290	0.0041	
PMT0110	conserved hypothetical protein	1.101	0.9971	1.156	0.2234	-1.119	0.5107	-1.369	0.0041	
PMT1514	possible Neuromedin U	1.009	0.9971	1.062	0.8338	-1.028	0.9520	-1.255	0.0042	
PMT2130	possible (M20568) ORF 11 [Azotobactervinelandii]	-1.023	0.9971	1.030	0.9221	-1.006	0.9604	-1.333	0.0043	
PMT1570	conserved hypothetical protein	-1.106	0.9971	-1.193	0.9329	-1.055	0.9604	-2.246	0.0048	
PMT0657	conserved hypothetical protein	-1.008	0.9971	1.092	0.8388	-1.086	0.7798	-1.413	0.0049	

ORF	Description	FC t=0	q-value t=0	FC t=4	q-value t=4	FC t=12	q-value t=12	FC t=24	q-value t=24	notes
PMT0483	possible Malic enzyme	-1.034	0.9971	-1.031	0.9613	1.065	0.9520	-1.860	0.0056	
PMT1273	hemC Porphobilinogen deaminase	1.011	0.9971	1.003	0.9632	-1.075	0.7798	-1.296	0.0056	
PMT1209	conserved hypothetical	1.197	0.9971	-1.015	0.9591	-1.205	0.7086	-1.658	0.0070	
PMT2109	rluD putative pseudouridylate synthase specific toribosomal large subunit	-1.002	0.9975	1.027	0.9221	-1.009	0.9604	1.208	0.0070	
PMT0329	conserved hypothetical protein	-1.089	0.9971	1.001	0.9632	-1.020	0.9604	1.282	0.0074	
PMT1155	probable GTP-binding protein	-1.053	0.9971	-1.065	0.8388	-1.026	0.9587	-1.288	0.0075	
PMT1577	hypothetical protein	1.043	0.9971	1.119	0.8409	-1.002	0.9604	-1.881	0.0086	
PMT1357	two-component response regulator	-1.232	0.9971	-1.042	0.8907	-1.003	0.9604	-1.275	0.0090	
PMT2262	possible NADH-Ubiquinone/plastoquinone (complexI	1.189	0.9971	-1.035	0.9221	-1.072	0.7977	-1.287	0.0092	
PMT0855	Isochorismatase hydrolase family	1.093	0.9971	1.044	0.9007	-1.095	0.7086	-1.288	0.0093	
PMT0484	possible Domain of unknown function DUF33	1.085	0.9971	1.025	0.9329	-1.013	0.9604	-1.323	0.0094	
PMT1780	rps7, rpsG 30S ribosomal protein S7	1.101	0.9971	1.004	0.9632	-1.188	0.6797	-1.423	0.0094	
PMT1231	conserved hypothetical protein	1.091	0.9971	-1.022	0.9632	-1.178	0.6797	-1.550	0.0097	
PMT2137	hypothetical	-1.018	0.9971	1.082	0.8755	1.173	0.7977	-1.774	0.0097	
PMT2153	unnamed product	-1.142	0.9971	1.157	0.1794	1.121	0.5755	1.261	0.0122	
PMT1199	csoS1 carboxysome structural protein CsoS1	-1.160	0.9971	-1.043	0.9221	-1.107	0.7977	-1.391	0.0123	
PMT1212	conserved hypothetical protein	1.061	0.9971	-1.113	0.8052	-1.056	0.8425	-1.284	0.0126	
PMT1742	rpl14, rplN 50S Ribosomal protein L14	1.066	0.9971	-1.058	0.9329	-1.124	0.8521	-1.383	0.0141	
PMT2090	rpl10, rplJ 50S ribosomal protein L10	1.066	0.9971	-1.086	0.9221	-1.097	0.9399	-1.539	0.0148	
PMT2077	small mechanosensitive ion channel, MscS family	1.069	0.9971	1.035	0.9329	-1.010	0.9604	-1.263	0.0149	
PMT0092	hypothetical protein	-1.060	0.9971	1.037	0.8907	-1.092	0.5218	-1.219	0.0152	
PMT1737	rpl22, rplV 50S ribosomal protein L22	1.017	0.9971	-1.080	0.9221	-1.123	0.8480	-1.542	0.0152	
PMT0929	Hemolysin-type calcium-binding region:RTXN-terminal domain	1.177	0.9971	-1.048	0.9667	-1.045	0.9604	-1.729	0.0163	
PMT1567	conserved hypothetical protein	-1.019	0.9971	1.031	0.9221	-1.104	0.6797	-1.333	0.0172	
PMT0631	hypothetical	1.126	0.9971	1.101	0.8409	1.174	0.7977	-1.693	0.0179	
PMT0656	conserved hypothetical protein	1.035	0.9971	-1.065	0.8354	1.036	0.9520	-1.193	0.0179	
PMT0947	hypothetical	-1.114	0.9971	-1.039	0.9007	1.065	0.8029	1.175	0.0179	
PMT0852	possible HAMP domain	1.064	0.9971	-1.022	0.9329	-1.038	0.9244	-1.164	0.0195	
PMT1022	rps21, rpsU 30S Ribosomal protein S21	1.163	0.9971	-1.152	0.8956	-1.356	0.2275	-1.541	0.0215	
PMT1539	F3H9.20 conserved hypothetical membrane protein	-1.045	0.9971	-1.026	0.9369	1.029	0.9604	1.343	0.0215	
PMT2125	hypothetical	1.156	0.9971	-1.063	0.8388	-1.051	0.9270	-1.212	0.0215	
PMT0311	galactosyl-1-phosphate transferase	1.003	0.9971	1.053	0.9007	1.126	0.6793	1.284	0.0216	
PMT1746	rpl6, rplF 50S ribosomal protein L6	1.067	0.9971	-1.083	0.9221	-1.212	0.6718	-1.331	0.0216	
PMT0099	HisH imidazoleglycerol-phosphate synthase, glutamineamidotransferase subunit	1.004	0.9971	1.042	0.8936	1.101	0.6797	1.168	0.0218	
PMT0551	sdmt putative sarcosine-dimethylglycinemethyltransferase	-1.006	0.9971	1.066	0.8409	-1.025	0.9604	-1.417	0.0218	
PMT2037	conserved hypothetical protein	1.058	0.9971	1.029	0.9329	-1.136	0.4629	-1.238	0.0226	
PMT0920	conserved hypothetical	-1.038	0.9971	-1.039	0.9221	-1.073	0.8029	1.202	0.0226	
PMT1545	possible ABC transporter, ATP binding component	-1.079	0.9971	1.071	0.8338	1.078	0.7977	1.233	0.0226	
PMT0925	hypothetical	1.113	0.9971	1.033	0.9329	1.080	0.9520	-1.688	0.0227	
PMT0382	possible Signal peptidase I	-1.160	0.9971	-1.035	0.9369	1.352	0.2806	-1.341	0.0228	
PMT0700	putative phosphate ABC transporter, ATP bindingsubunit	1.051	0.9971	-1.001	0.9632	-1.038	0.9587	1.211	0.0230	
PMT0857	hypothetical	-1.076	0.9971	-1.045	0.8409	-1.056	0.7977	-1.165	0.0232	
PMT1606	ABC transporter, membrane component	1.266	0.9971	1.074	0.8124	1.104	0.8029	1.269	0.0238	
PMT1718	ATP:corrinoid adenosyltransferase BtuR/CobO/CobP	-1.064	0.9971	1.032	0.9221	1.121	0.4766	1.247	0.0238	
PMT1758	rpl13, rplM 50S ribosomal protein L13	1.147	0.9971	-1.163	0.8388	-1.247	0.5744	-1.361	0.0238	
PMT0642	conserved hypothetical protein	1.130	0.9971	1.114	0.6015	-1.136	0.4059	-1.249	0.0254	
PMT1235	DUF152	-1.051	0.9971	1.033	0.9221	-1.015	0.9604	1.239	0.0255	
PMT1738	rps3, rpsC 30S ribosomal protein S3	1.010	0.9971	-1.009	0.9632	-1.052	0.9587	-1.494	0.0258	
PMT1198	conserved hypothetical protein	-1.025	0.9971	-1.065	0.8409	1.039	0.9520	-1.184	0.0263	
PMT0106	neuB N-acetylneuraminate synthase	1.075	0.9971	1.034	0.9221	-1.058	0.8029	-1.179	0.0272	
PMT0256	Hemolysin-type calcium-binding region:RTXN-terminal domain	-1.233	0.9971	1.070	0.8388	1.046	0.9520	-1.276	0.0272	
PMT1734	rpl23, rplW 50S ribosomal protein L23	1.060	0.9971	-1.125	0.8388	-1.140	0.7977	-1.378	0.0272	
PMT2151	crp1 Bacterial regulatory proteins, GntR family:Cyclicnucleotide-...	1.068	0.9971	-1.082	0.7512	-1.173	0.1320	-1.293	0.0272	
PMT0445	hemE Uroporphyrinogen decarboxylase (URO-D)	1.062	0.9971	-1.094	0.8388	1.030	0.9604	-1.291	0.0284	
PMT0678	possible ABC transporter	1.024	0.9971	1.053	0.8493	1.122	0.5555	1.200	0.0288	
PMT0725	conserved hypothetical protein	1.011	0.9971	-1.066	0.8409	-1.110	0.6797	-1.290	0.0288	
PMT1408	Peptide methionine sulfoxide reductase	-1.104	0.9971	1.001	0.9674	-1.005	0.9604	-1.215	0.0288	
PMT1643	conserved hypothetical protein	-1.038	0.9971	1.083	0.6727	-1.034	0.9587	-1.252	0.0288	
PMT2011	conserved hypothetical protein	-1.105	0.9971	1.074	0.8338	-1.085	0.7086	-1.231	0.0288	
PMT2091	rpl12, rplL, rpl7 50S ribosomal protein L7/L12	1.114	0.9971	-1.101	0.8409	-1.128	0.8029	-1.465	0.0288	
PMT0167	hypothetical	-1.200	0.9971	1.023	0.9329	-1.019	0.9604	-1.207	0.0296	
PMT1216	chlB Oxidoreductase, nitrogenase component 1	1.049	0.9971	-1.057	0.9221	-1.052	0.9587	1.377	0.0319	
PMT0055	rpl20, rplT 50S ribosomal protein L20	1.009	0.9971	1.010	0.9632	-1.196	0.6797	-1.514	0.0321	
PMT0827	hypothetical	1.111	0.9971	1.055	0.9221	-1.035	0.9604	-1.306	0.0321	
PMT1981	hypothetical	-1.057	0.9971	1.030	0.9252	-1.008	0.9604	-1.203	0.0321	
PMT2129	hypothetical	1.092	0.9971	1.136	0.6172	-1.098	0.7977	-1.315	0.0321	
PMT2195	petF, fdx ferredoxin, PetF like protein	-1.026	0.9971	1.003	0.9632	-1.026	0.9587	-1.263	0.0321	
PMT0246	possible Profilin	1.238	0.9971	1.121	0.8388	1.037	0.9587	-1.623	0.0327	
PMT1441	conserved hypothetical protein	1.030	0.9971	1.032	0.9221	-1.100	0.6797	-1.223	0.0327	
PMT1827	hypothetical	1.126	0.9971	1.056	0.8591	-1.024	0.9604	-1.309	0.0327	
PMT2122	possible REJ domain	1.078	0.9971	1.030	0.9329	-1.076	0.7977	-1.289	0.0327	
PMT0282	hypothetical	-1.057	0.9971	1.043	0.8409	1.027	0.9520	-1.162	0.0328	
PMT1015	conserved hypothetical	1.156	0.9971	1.046	0.8936	-1.142	0.4059	-1.235	0.0328	
PMT0247	possible (M20568):ORF 11 [Azotobactervinelandii]	1.064	0.9971	1.129	0.6172	-1.012	0.9604	-1.265	0.0329	

ORF	Description	FC t=0	q-value t=0	FC t=4	q-value t=4	FC t=12	q-value t=12	FC t=24	q-value t=24	notes
PMT1438	possible Mannitol dehydrogenase	1.025	0.9971	-1.007	0.9594	-1.021	0.9604	-1.206	0.0329	
PMT1743	rpl24, rplX 50S ribosomal protein L24	1.110	0.9971	-1.106	0.9161	-1.233	0.6797	-1.331	0.0329	
PMT1042	hypothetical	1.007	0.9971	-1.015	0.9586	-1.105	0.6280	-1.208	0.0371	
PMT1161	dxr 1-deoxy-D-xylulose 5-phosphate reductoisomerase	-1.133	0.9971	1.002	0.9632	1.156	0.6155	1.227	0.0371	
PMT0107	conserved hypothetical protein	-1.062	0.9971	1.202	0.1794	1.026	0.9604	-1.192	0.0373	
PMT1081	rbpD RNA-binding protein RbpD	1.063	0.9971	-1.000	0.9632	-1.383	0.2428	-1.637	0.0373	
PMT1657	hypothetical	1.148	0.9971	1.079	0.8124	1.012	0.9604	-1.232	0.0373	
PMT1043	possible Domain of unknown function DUF38	-1.037	0.9971	1.049	0.8907	-1.075	0.7977	-1.266	0.0374	
PMT2054	hypothetical	1.048	0.9971	-1.076	0.7379	-1.069	0.8047	-1.188	0.0374	
PMT0503	possible ATLS1-like light-inducible protein	-1.051	0.9971	1.071	0.8124	-1.041	0.9520	-1.217	0.0375	
PMT0253	possible Galanin	1.083	0.9971	1.093	0.8052	-1.018	0.9604	-1.206	0.0377	
PMT1736	rps19, rpsS 30S Ribosomal protein S19	1.040	0.9971	-1.242	0.6501	-1.170	0.7638	-1.418	0.0377	
PMT0232	hupE putative hydrogenase accessory protein	1.074	0.9971	1.059	0.8734	-1.080	0.7909	-1.206	0.0380	
PMT2087	nusG transcription antitermination protein, NusG	-1.045	0.9971	-1.006	0.9632	-1.081	0.7977	-1.357	0.0380	
PMT0795	conserved hypothetical protein	1.197	0.9971	-1.033	0.9221	-1.089	0.7086	-1.259	0.0380	
PMT1739	rpl16, rplP 50S ribosomal protein L16	-1.020	0.9971	-1.140	0.8907	-1.160	0.7977	-1.425	0.0388	
PMT2118	possible tRNA synthetases class I (C)	-1.003	0.9971	1.202	0.8338	-1.014	0.9604	-1.643	0.0399	
PMT2212	Class I peptide chain release factor	-1.120	0.9971	-1.052	0.8907	1.019	0.9604	-1.161	0.0399	
PMT0145	Bacterial extracellular solute-binding protein,family 1	-1.020	0.9971	-1.013	0.9454	-1.007	0.9604	-1.181	0.0410	
PMT1074	kprS Ribose-phosphate pyrophosphokinase	-1.078	0.9971	-1.014	0.9591	1.025	0.9604	1.303	0.0410	
PMT1776	conserved hypothetical protein	1.008	0.9975	-1.038	0.9161	-1.047	0.9187	-1.237	0.0416	
PMT0249	fatty acid desaturase, type 2	1.046	0.9971	1.094	0.6700	-1.074	0.7977	-1.201	0.0420	
PMT0856	possible large-conductance mechanosensitivechannel mscL	1.049	0.9971	1.007	0.9632	-1.121	0.6718	-1.287	0.0420	
PMT2089	rpl1, rplA 50S ribosomal protein L1	1.077	0.9971	-1.038	0.9329	-1.166	0.6718	-1.273	0.0425	
PMT0140	DedA family; putative alkaline phosphatase-likeprotein	1.189	0.9971	1.004	0.9632	-1.073	0.8039	-1.208	0.0435	
PMT0718	conserved hypothetical protein	-1.081	0.9971	-1.058	0.8388	-1.013	0.9604	-1.225	0.0435	
PMT0907	possible Octicosapeptide repeat	-1.025	0.9971	1.159	0.6700	1.024	0.9604	-1.285	0.0435	
PMT1936	possible DNA polymerase family B	-1.059	0.9971	1.151	0.4553	1.008	0.9604	-1.152	0.0438	
PMT1759	rps9, rpsI 30S ribosomal protein S9	1.163	0.9971	-1.033	0.9409	-1.180	0.7909	-1.306	0.0438	
PMT2042	Pseudouridine synthase	-1.003	0.9975	-1.032	0.9329	-1.041	0.9520	1.196	0.0438	
PMT0430	conserved hypothetical protein	-1.038	0.9971	-1.017	0.9444	-1.022	0.9604	1.191	0.0446	
PMT2264	mraY Putativephospho-N-acetylmuramoyl-pentapeptide-transferase	1.008	0.9971	1.016	0.9400	-1.062	0.8140	-1.211	0.0446	
PMT1579	possible Flavivirus glycoprotein, immunoglobul	-1.059	0.9971	-1.072	0.8124	1.025	0.9520	1.150	0.0449	
PMT0101	acetyltransferase	1.031	0.9971	1.064	0.7372	1.031	0.9520	-1.173	0.0451	
PMT0652	STAS domain:Anti-sigma factor antagonist	1.048	0.9971	1.006	0.9615	-1.054	0.9430	-1.260	0.0451	
PMT1307	Conserved hypothetical protein	1.170	0.9971	1.024	0.9591	-1.167	0.7977	-1.480	0.0451	
PMT0103	aminotransferase, Class III pyridoxal-phosphatedependent	1.084	0.9971	-1.031	0.9221	-1.004	0.9604	-1.152	0.0455	
PMT1646	similar to serum resistance locus BrkB	1.024	0.9971	1.032	0.9221	-1.082	0.7798	-1.204	0.0459	
PMT1432	conserved hypothetical protein	-1.033	0.9971	-1.008	0.9632	1.032	0.9604	1.288	0.0459	
PMT1239	ilvB acetolactate synthase	-1.033	0.9971	-1.000	0.9667	1.104	0.6949	1.207	0.0460	
PMT1621	possible 4'-phosphopantetheinyl transferasefamily protein	-1.116	0.9971	-1.075	0.7608	-1.055	0.8464	-1.181	0.0461	
PMT0564	Putative 6-phosphogluconolactonase (DevB, Pgl)	-1.011	0.9971	-1.034	0.9221	-1.074	0.8425	1.180	0.0469	
PMT1211	conserved hypothetical protein	1.159	0.9971	1.020	0.9401	-1.078	0.8464	-1.208	0.0477	
PMT1284	conserved hypothetical protein	-1.098	0.9971	-1.055	0.9007	1.047	0.9520	1.218	0.0477	
PMT1697	possible NADH-Ubiquinone/plastoquinone (comple	1.025	0.9971	1.080	0.8338	-1.014	0.9604	-1.242	0.0477	
PMT1515	Sulfatase	1.009	0.9971	1.022	0.9329	-1.158	0.3954	-1.204	0.0485	
PMT2209	possible AIR synthase related protein, C-termi	1.055	0.9971	-1.001	0.9632	1.002	0.9604	-1.187	0.0486	
PMT1351	conserved hypothetical protein	1.075	0.9971	-1.061	0.9007	-1.078	0.8010	-1.223	0.0487	
PMT0259	hypothetical protein	-1.063	0.9971	1.077	0.7512	1.009	0.9604	-1.187	0.0494	

notes

1 signal is likely due to cross-hybridization with PMT0993 probes. Using qPCR with specific primers, we do not see upregulation of PMT0508.

2 part of *phoR* pseudogene. See text.

Table 3. List of primer sequences used in RT-PCR

Strain	Gene	Orientation	Position	Sequence (5'-3')
MED4	*pstS*	F	386	TGGGTATGGTTAAAAACTGG
MED4	*pstS*	R	545	GGCCACTTAACTGATTTACC
MED4*	*rnpB*	F	1	TTGAGGAAAGTCCGGGCTC
MED4*	*rnpB*	R	91	GCGGTATGTTTCTGTGGCACT
MIT9313	*pstS1*	F	11	CAAAGAAGGCCCTTTTGCTC
MIT9313	*pstS1*	R	87	GGATGACCCTGAGGTGCTGC
MIT9313	*pstS2*	F	9	TCTCAAGAAGGGCTTTCTGC
MIT9313	*pstS2*	R	88	CTGAGGTGCCTGAGGTGCT
MIT9313	*rnpB*	F	284	TCTGCCACGTTCCACATAAA
MIT9313	*rnpB*	R	361	AGAGCAGTGGGTGCTCATCT

F, forward; R, reverse.

*From Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M. & Chisholm, S. W. (2005) *Nature* **438,** 86-89.

CHAPTER FOUR

Structure and dynamics of genomes and gene expression in natural *Prochlorococcus* populations

Maureen L. Coleman, Yanmei Shi, Gene W. Tyson, Jorge Frias-Lopez, Edward F. DeLong, and Sallie W. Chisholm

Chapter 4

Structure and dynamics of genomes and gene expression in natural *Prochlorococcus* populations

Abstract

Metagenomic sequencing, especially when placed in the context of whole-genome sequences of physiologically characterized isolates, promises to yield a wealth of new insights into the structure and functioning of microbial populations. In this study, pyrosequencing data sets, prepared from genomic DNA and cDNA at three depths in the water column at Hawaii Ocean Time Series station ALOHA, were used to characterize *Prochlorococcus* populations. These sequence databases afford a detailed and quantitative picture of genomic diversity and expression in a natural microbial population, defined here as coexisting *Prochlorococcus* cells at a given depth. Principal aspects of *Prochlorococcus* diversity observed in cultured isolates, including the presence of core and flexible genomes and ecotypic variation in gene content, were manifested in the population metagenome. Both operon and genome-wide expression patterns in the metatranscriptomic data mirrored those observed in laboratory microarray experiments, reinforcing the validity of using cultured isolates to help understand wild populations. Depth-dependent variations in gene abundance suggest that *Prochlorococcus* cells found towards the base of the euphotic zone are distinct from those near the surface, even when they belong to the same ecotype, confirming that they are indeed two separate populations. A set of genes showing differential abundance along the depth profile also implicate *Prochlorococcus* in an active water column cycle of thiamin. Finally, we detected signals of intragenic recombination and we hypothesize this recombination may be linked to high cellular expression levels of these particular genes. These results demonstrate the power of quantitative metagenomics and metatranscriptomics to extend inferences from laboratory studies to natural habitats and to reveal unexpected features of the microbial populations that live in them.

Bacteria inhabit nearly every corner of the earth, a testament to their genomic plasticity and adaptive potential. Phenotypic change can arise from gene gain and loss, substitutions in coding regions, and regulatory modification. Whole genome sequences have revealed, for instance, that even closely related taxa differ widely in gene content (Welch et al., 2002; Kettler et al., 2007, Appendix C; Konstantinidis and Tiedje, 2005). It is difficult to say, however, how much of this observed gene content variability is the result of natural selection and how much is neutral. Even less is known about the significance of allelic variants and about the scope of regulatory evolution in bacteria. Thus a major challenge is to characterize patterns in gene content, sequence variation, and regulation in a microbial population across environmental gradients, and to understand the contribution of these components to phenotypic change.

Metagenomics offers an approach for studying these questions without the need for cultured isolates. Yet for understanding the processes of evolutionary change, whole genome sequences of reference taxa, and some knowledge of their physiology, are essential. The marine cyanobacterium *Prochlorococcus* is uniquely suited to this challenge: as one of the most abundant taxa in the open oceans, it is readily captured in marine metagenomic databases (Venter et al., 2004; DeLong et al., 2006; Rusch et al., 2007), and a dozen isolate genomes have been sequenced to date. Further, physiological diversity has been characterized with respect to light (Moore and Chisholm, 1999), temperature (Moore et al., 1995;

Johnson et al., 2006; Zinser et al., 2007), nitrogen (Moore et al., 2002), phosphorus (Moore et al., 2005; Martiny et al., 2006), and metals (Mann et al., 2002; Saito et al., 2002). By further exploring how genome variability in *Prochlorococcus* populations is structured over environmental gradients, we can begin to infer which features are invisible to selection and which contribute to ecological differentiation.

The relatively simple metabolism, well-characterized habitat, and small genome size (1.7-2.4 Mb) of *Prochlorococcus* make it even more amenable as a model organism. *Prochlorococcus* is a small (0.6um) unicellular cyanobacterium, distinguished from its close relative *Synechococcus* by its unique pigment composition. The *Prochlorococcus* group consists of genetically and physiologically distinct ecotypes, recognizable by their internal transcribed spacer (ITS) sequence and by their light physiology (Moore et al., 1998; Rocap et al., 2002). Throughout this work, we use the term "ecotype" as used historically for *Prochlorococcus* (Moore et al., 1998; Rocap et al., 2002), which is not necessarily congruent with usage by others (see Coleman and Chisholm, 2007 for a discussion). High light-adapted (HL) ecotypes (eMIT9312 and eMED4; ecotypes are identified by "e" and their type strain) dominate near the surface, while low light-adapted (LL) ecotypes (eNATL2A, eMIT9313) can become more abundant deeper in the water column. The two HL ecotypes are thought to differ in their temperature physiology (Johnson et al., 2006), but the traits distinguishing different LL ecotypes remain unclear (Zinser et al., 2007). The abundance of these four major ecotypes strongly depends on light and temperature, but there remains unexplained variability in their distributions (Bouman et al., 2006; Johnson et al., 2006; Zinser et al., 2007; Zwirglmaier et al., 2007).

Complicating this ecotype picture, gene content and cellular physiology vary significantly even within a single ITS-defined ecotype. Phosphate acquisition genes, for example, are abundant in some strains isolated in extremely low-phosphate waters, while these genes are missing entirely in other strains of the same ecotype (Martiny et al., 2006). This relationship between genome content and phosphate availability has also been observed in metagenomic samples from different oceans (Rusch et al., 2007). Similarly, the gene complement for nitrogen assimilation is very different between strains MED4 and MIT9515, both members of the eMED4 ecotype (Kettler et al., 2007, Appendix C). In general, much of the gene content variation among *Prochlorococcus* isolates exists in the "leaves of the tree", i.e. between closely related isolates within an ecotype, particularly for genes involved in cell surface biosynthesis and transport (Kettler et al., 2007, Appendix C). It is unclear, however, whether this variation is adaptive or whether it represents random gene gains and losses. Although one layer of ecological differentiation is captured by the ITS and ecotype paradigm, other layers potentially exist at finer phylogenetic resolution.

The complementary *Prochlorococcus* datasets currently available — whole genome sequences, the Global Ocean Survey (GOS) and similar metagenomic datasets (Venter et al., 2004; DeLong et al., 2006; Rusch et al., 2007), and global ecotype profiles based on the ITS sequence (Ahlgren et al., 2006; Bouman et al., 2006; Johnson et al., 2006; Zinser et al., 2006; Garczarek et al., 2007; Zwirglmaier et al., 2007) — have defined key patterns of diversity within this microbial group, yet each has specific limitations when we seek to understand natural populations at the genome level. Genome sequences from isolates have begun to define the pan-genome space (Kettler et al., 2007, Appendix C) and are crucial for studying gene interactions within a cell, yet reflect single isolates from a few points in space and time. ITS-based

ecotype profiles reveal patterns along environmental gradients and through time, but are only indicative of a single locus. The available metagenomic datasets, while offering genome-wide sequences along environmental gradients, provide relatively low coverage of *Prochlorococcus* and are subject to cloning biases (Sorek et al., 2007). These factors make it difficult to draw quantitative inferences about the distribution and function of genes within populations.

What is needed to leverage the knowledge gained from these datasets is much deeper genome-wide sequence coverage of a single taxon along environmental gradients, free from cloning biases. Recent advances in sequencing technology (Margulies et al., 2005) have the power to dramatically sharpen the picture of microbial ecology and evolution at the genome level. With high throughput and reasonable template requirements, these tools enable observation of microbes at depth, breadth, and resolution commensurate with their capacities for adaptation. In combination with methods for amplification of bacterial mRNA (Van Gelder et al., 1990; Wendisch et al., 2001; Poretsky et al., 2005; Moreno-Paz and Parro, 2006; Rachman et al., 2006; Frias-Lopez et al., 2008, Appendix B), sequencing has become a tool for functional analysis. Information about gene expression will provide insight into which gene content differences might be functionally important, and into regulatory differences among genes shared in different environments.

To advance our goal of understanding the origins and functional significance of coexisting diversity in natural microbial populations, we analyzed metagenomic and metatranscriptomic sequences of microbial communities along a depth gradient at Station ALOHA, near Hawaii. The sequences were obtained as part of a methods development project (Frias-Lopez et al. 2008, Appendix B) and metagenomic surveys of depth stratified microbial communities (Shi et al. *in prep.*). Here we characterize the *Prochlorococcus* populations embedded in this community at three depths by both their gene content and gene expression. We address the following questions: what is the extent and organization of diversity within and between samples, and how does this compare to our understanding from cultured isolates? Which genes are likely to confer a fitness advantage at different depths? How does gene expression change with depth and time? And what are the contributions of strain-specific genes and genes of unknown function to this expression profile? The sequences were generated using a modified RNA amplification protocol (Frias-Lopez et al. 2008, Appendix B) and pyrosequencing (Margulies et al., 2005), and this approach, combined with the availability of whole genome scaffolds and ecotype characterization, allows unprecedented population-level quantitative inferences about gene content and expression in *Prochlorococcus*.

MATERIALS AND METHODS

Sampling

Seawater was collected from 25m (22:00 local time), 75m (03:30), and 125m (08:00) at the Hawaii Ocean Time-series (HOT) Station ALOHA (22 44' N, 158 2' W) in March 2006. Hydrocasts were conducted aboard the R/V *Kilo Moana* using a conductivity-temperature-depth (CTD) rosette water

sampler equipped with 24 Scripps 12L sampling bottles. The mixed layer depth was about 50m during sampling, so 25m is within the mixed layer while 75m is below. The 125m sample corresponded to the chlorophyll maximum. Samples for RNA and DNA were collected as in Frias-Lopez et al. (2008, Appendix B) and Shi et al. (*in prep.*).

Extraction, amplification, and sequencing of nucleic acids

DNA and cDNA were prepared, and methods development carried out, as described in Frias-Lopez et al. (2008, Appendix B) and Shi et al. (*in prep.*).

Removal of low-quality cDNA reads and rRNA

cDNA sequences were filtered to remove low-quality and rRNA sequences and trimmed as described (Frias-Lopez et al., 2008, Appendix B; Shi et al. *in prep.*).

Identification of putative *Prochlorococcus* fragments

Prochlorococcus protein-coding sequences were defined as reads with a top BLASTX hit against the NCBI-nr database to *Prochlorococcus*, with a bit score greater than 40. The number of *Prochlorococcus* coding sequences identified are listed in Table 1. Reads assigned to *Prochlorococcus* protein coding sequences by this method constituted 21, 31, and 10% of the total community genomic DNA reads and 6.2, 6.2, and 6.0% of the community cDNA reads at 25, 75, and 125m, respectively.

Assigning sequences to orthologous gene clusters

Prochlorococcus fragments were mapped to orthologous gene clusters (as defined in Kettler et al., 2007, Appendix C) as follows. Each fragment was searched using BLASTN against a database of predicted coding sequences from fully sequenced *Prochlorococcus* genomes, with a bit score cutoff of 40 and minimum alignment length 40. If the top three gene hits all belonged to the same orthologous cluster, then the query fragment was assigned to that cluster. If the top three hits represented more than one cluster, then the fragment was classified as a "conflict". If only two gene hits were found, and they belonged to the same cluster, then the fragment was assigned to that cluster; if the two genes belonged to different clusters, the fragment was classified "conflict". If only one gene hit was retrieved, the fragment was assigned to that cluster. Using this conservative procedure, 92-95% of the *Prochlorococcus* protein-coding genomic DNA fragments and 93-96% of the cDNA fragments were assigned to orthologous clusters (Table 1). Fewer assignments were possible from the 125m samples compared with the 25m and 75m samples, reflecting the higher diversity among low-light adapted *Prochlorococcus* and the relative lack of representative whole genome sequences.

Assigning fragments to *Prochlorococcus* ecotypes

To understand the contributions of different ecotypes to the bulk *Prochlorococcus* genomic DNA and mRNA, we examined the ecotype breakdown of blast hits. Fragments that could be assigned to gene clusters were examined for hits to the following *Prochlorococcus* genomes: AS9601, MIT9312, MIT9301 (all eMIT9312 ecotype); MED4, MIT9515 (both eMED4 ecotype); NATL1A, NATL2A (both eNATL2A ecotype); MIT9303, MIT9313 (both eMIT9313 ecotype); SS120 and MIT9211 (not included in the ecotype assignments because they are very rare). A fragment was assigned to an ecotype if the top two (for eMED4, eMIT9313, eNATL2A) or three hits (for eMIT9312) were to the two or three representative genomes of that ecotype. For example, for a fragment to be assigned *geneA* from eMIT9312, the top

three hits must include *geneA* from AS9601, MIT9301, and MIT9312, in any order. Thus an ecotype-assigned fragment clusters (by BLASTN) with the orthologous genes of its own ecotype, to the exclusion of all other ecotypes. By this conservative procedure, 82-85% of the genomic DNA fragments and 72-77% of the cDNA fragments were assigned unambiguously to an ecotype (Table 1).

Statistical analyses of gene frequencies in the genomic DNA

Within a single sample, we wished to identify genes that were detected by pyrosequencing more or less often than would be expected if they were single copy in every cell. The number of DNA fragments n_i observed for a gene i of length L_i and average multiplicity per genome m_i in a sample of N_{Pro} sequences is expected to be binomially distributed:

$$n_i(L_i, m_i) \sim binomial\left(N_{Pro}, \frac{L_i m_i}{\sum L_i m_i}\right)$$

Thus each sequencing read we sample can be classified as either a "success" or "failure": either it belongs to gene i or it does not. The probability of success (i.e. that the read came from a given gene i) depends on the length of the gene, as longer genes will be detected more frequently. This distribution was used to estimate 99.9% confidence intervals for the number of fragments expected for each gene length L_i assuming a multiplicity m_i of 1, as we expect for core genes (plotted in Figure 1). Genes that were detected more or less frequently than predicted by this interval presumably have an average multiplicity different from 1 copy per cell, and were therefore classified as "abundant" or "rare" in the population (relative to core genes).

Between samples, we wished to compare the frequency of each gene to identify those with significantly higher or lower abundance between depths. To do so, we estimated the multiplicity of each gene per genome, taking a normal approximation to the binomial:

$$m_i = \frac{n_i}{bL_i} \sim normal\left(m_i, \frac{m_i}{bL_i}\right)$$

A gene was classified as having significantly different multiplicity at two depths if the 99.9% confidence intervals for m_i did not overlap.

Statistical analyses of transcript abundance in the cDNA

The RNA amplification and cDNA sequencing approach has been shown to be unbiased for protein coding genes (Frias-Lopez et al., 2008, Appendix B), and therefore we assumed that the proportion of reads from gene i in the cDNA library reflects the proportion of mRNA molecules from gene i in the average cell. This proportion p_i is calculated as n_i/N_{Pro}, where n_i is the number of reads detected from gene i and N_{Pro} is the total number of *Prochlorococcus* cDNA reads identified by BLASTX as above. Using a normal approximation to the binomial with a mean equal to p_i and variance equal to $p_i(1-p_i)$, 95% confidence intervals were estimated. To compare the proportion of a gene i in two samples (*1* and *2*), Z scores were calculated as follows:

$$Z_i = \frac{(p_{i,1} - p_{i,2})}{\sqrt{\frac{p_{i,0}(1-p_{i,0})}{N_{Pro,1}} + \frac{p_{i,0}(1-p_{i,0})}{N_{Pro,2}}}}$$

where

$$p_{i,0} = \frac{(n_{i,1} + n_{i,2})}{(N_{Pro,1} + N_{Pro,2})}$$

Genes with a Z score (absolute value) greater than 1.96, corresponding to 95% confidence intervals, had significantly different expression in the two samples.

QPCR analysis of ecotype abundance

As another measure of the ecotype composition of the *Prochlorococcus* assemblage, we used quantitative PCR targeting the ITS region to quantify each of four major ecotypes: eMED4, eMIT9312, eNATL2A, and eMIT9313, using the primers and protocol developed by Ahlgren (2006) and Zinser (2006).

RESULTS AND DISCUSSION

Structure of the *Prochlorococcus* assemblage genome

Although the average *Prochlorococcus* genome contains just 2000 genes, the *Prochlorococcus* pan-genome based on 12 isolates includes nearly 6000 genes and continues to grow by roughly 50 genes with each newly sequenced isolate (Kettler et al., 2007, Appendix C). This pan-genome encompasses the core genome, consisting of 1273 genes – 1221 of which are present in a single copy per genome – found in all 12 sequenced isolates, and the flexible genome, consisting of genes found in some but not all isolates. We wondered how well these core/flexible designations, based on the genomes of 12 isolates, applied to natural *Prochlorococcus* populations at our study site in the subtropical Pacific. We detected every core gene at 25m and 75m (i.e. at least one read mapped to every core gene), and all but two core genes at 125m, where there were far fewer *Prochlorococcus* sequences and thus less depth of coverage (Table 1). We quantified the sequence coverage of each gene as the number of reads of that gene normalized to 1kb. The coverage of core genes was normally distributed around a distinct mean that scaled with the total number of *Prochlorococcus* sequence reads at each depth, consistent with these genes being single-copy in each cell (Figure 1A). Nearly all core genes were observed at frequencies predicted by their gene length (Figure 1B), which reflects the minimally biased representation of the template DNA afforded by pyrosequencing. Four core genes were observed at significantly lower frequencies, outside the 99.9% confidence interval, at all three depths (labeled w-z in Figure 1B). All four are located in the genomic island encoding lipolysaccharide biosynthesis (ISL4 of Coleman et al., 2006), and these same genes (plus three others) are also underrepresented in *Prochlorococcus* from the GOS dataset (Kettler et al., 2007, Appendix C). Their underrepresentation in both datasets is likely due to their history of gene transfer and their high sequence divergence, which makes their sequences unrecognizable in some cases (Kettler et al.,

2007, Appendix C). Overall, these observations suggest that the core genome identified on the basis of a limited number of whole genomes — 12 isolates totaling 22.4 Mb of genome sequence — is indeed a set of genes common to *Prochlorococcus* cells in natural populations, both in the subtropical Pacific and along the GOS transect (Kettler et al., 2007, Appendix C).

By contrast, genes belonging to the so-called "flexible genome" showed a greater dispersion in coverage (Figure 1A) and could be divided into two main groups: genes that occurred very rarely in the population (points below the 99.9% confidence interval, Figure 1B), and genes that occurred at roughly the same frequency as core genes, i.e. in roughly every cell (points within the interval, Figure 1B; Suppl. Table 1). The latter group essentially constitutes an extension of the core genome in this particular environment. Moreover, these genes are also relatively abundant in the genomes of HL *Prochlorococcus* isolates sequenced to date (see ecotype discussion below) and therefore might represent an extension of the core in these lineages.

Patterns and processes in the flexible genome

The two classes of flexible genes – those rare in the population and those as abundant as core genes – tend to be located in different regions of the chromosome in cultured isolates. Many of the genes located in genomic islands ISL1, ISL2, and ISL5 (identified in Coleman et al., 2006) in reference strain MIT9301, are close to one copy per cell in these samples; in contrast, genes located in MIT9301 ISL3 and ISL4, and in a new island not described by Coleman (Coleman et al., 2006), are rare in these samples. This pattern bears strong similarity to data from the Sargasso Sea metagenome (Venter et al., 2004): in the Sargasso, genes from ISL2 and ISL5 (using reference strain MED4) were abundant in the population, while genes in ISL1, ISL3, and ISL4 were rare (Coleman et al., 2006). It appears that not all islands are created equal: unlike the other major islands, ISL2 and ISL5 contain genes that are abundant in natural populations, lack tRNA genes, and contain a large number of genes including *hli* genes that are expressed under various stress conditions (Coleman et al., 2006). We hypothesize that these two islands have essentially become core in some HL populations and have lost their mobility or propensity for gene acquisition. These patterns suggest a possible evolutionary progression: genes are acquired in genomic islands, such as the newly discovered island in MIT9301 adjacent to tRNA-Ser; most acquired genes are lost, but adaptive loci are maintained by selection and gradually become an extension of the core, and eventually the island loses its mobility features. A similar role for islands in bacterial genome evolution was proposed by Lawrence and Hendrickson (2005). Thus genomic islands could be important for adaptation in large populations over many generations.

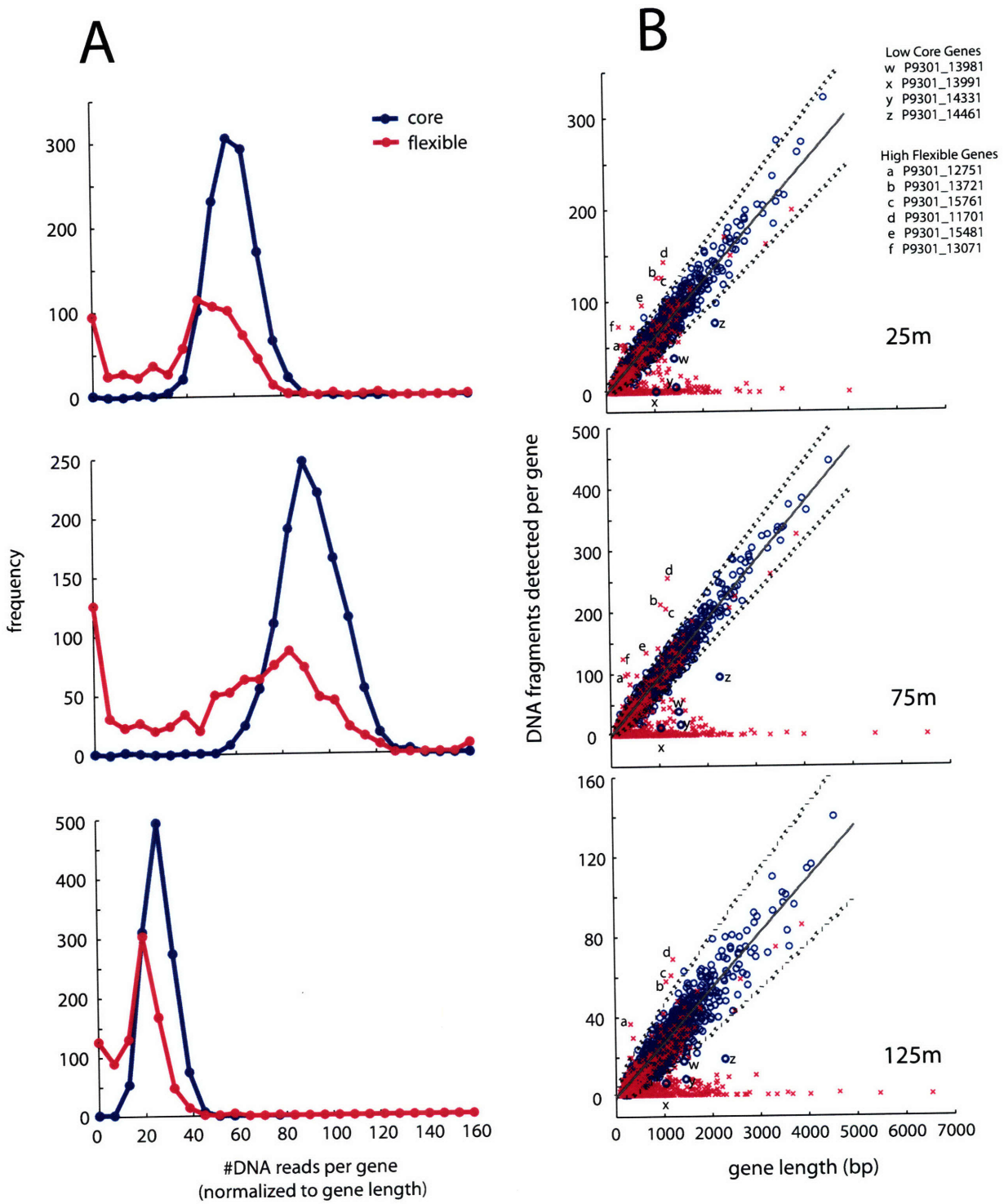
Environmental factors likely drive selection for subsets of the flexible genome, and thus the abundance of specific flexible genes can inform our understanding of the environment experienced by the cells. Phosphate availability is one such factor shaping genome content in *Prochlorococcus* (Martiny et al., 2006) and indeed the frequency of phosphate uptake genes is higher in low-P waters than in higher P waters (Rusch et al., 2007). In this dataset, the phosphate acquisition genes *phoA*, which encodes alkaline phosphatase, and *phoBR*, which encodes a two-component regulatory system, are rare relative to core genes at all three depths, reflecting, we hypothesize, that phosphate is non-limiting for *Prochlorococcus*

Table 1. Summary of database sizes, listed as the number of pyrosequencing reads. In each database, over 90% of reads assigned to *Prochlorococcus* (top hit by blastx) could also be assigned to a particular orthologous gene cluster. Between 70 and 84% of *Prochlorococcus* reads could also be assigned to an ecotype, using blastn and conservative criteria described in the Methods. For cDNA, the total number of community reads is presented both including and excluding rRNA reads.

	Total # of community reads (non-rRNA)*	Total # Prochl. CDS Reads	# / % of CDS reads assigned to a gene cluster		# / % of CDS reads assigned to an ecotype	
25m DNA	385193	84671	81492	96	71051	84
75m DNA	414323	131402	126405	96	109833	84
125m DNA	348394	37885	35106	93	30135	80
25m cDNA	85283 (51507)	3219	3054	95	2429	75
75m cDNA	120427 (58710)	3654	3482	95	2588	71
125m cDNA	112686 (66974)	4031	3712	92	2822	70

* from Frias-Lopez et al. 2008 (Appendix B) and Shi et al. *in prep* .

Figure 1 (opposite page). Two views of the quantitative sampling of the *Prochlorococcus* metagenome. (A) Frequency histograms showing the number of genomic DNA fragments assigned to each *Prochlorococcus* gene in samples from 3 depths. Values are normalized to a gene of length 1000bp. Blue, frequencies for core genes; pink, flexible genes. Core genes show distributions around defined means that scale with the database size (125m<25m<75m), a result of their single-copy occurrence in nearly all cells. The distributions of flexible genes are broader, with peaks lower and less well-defined than those of core genes, reflecting the more variable occurrence of flexible genes. (B) Relationship between gene length (in a reference genome) and the number of times it was detected in the genomic DNA fragments, plotted for each sample. Blue, core genes; pink, flexible genes. Lines are the mean (solid) and 99.9% confidence intervals (dashed) of a binomial distribution used to describe the results of random DNA sampling by pyrosequencing (see Methods). Nearly all core genes and many flexible genes are detected according to this theoretical distribution, confirming the representative, quantitative nature of the sequence sampling. The low-abundance group of flexible genes likely reflects contributions from minor ecotypes as well as a few genes that are infrequently present in eMIT9312 cells. Letters a-f indicate flexible genes present at higher than expected abundance (above the 99.9% confidence interval), while letters w-z indicate core genes present at lower than expected abundance (below the confidence interval).


A
core
flexible
frequency
#DNA reads per gene
(normalized to gene length)
B
Low Core Genes
w P9301_13981
x P9301_13991
y P9301_14331
z P9301_14461
High Flexible Genes
a P9301_12751
b P9301_13721
c P9301_15761
d P9301_11701
e P9301_15481
f P9301_13071
25m
75m
125m
DNA fragments detected per gene
gene length (bp)

at station ALOHA. This relatively high phosphate availability (compared to more P-limited oceanic regions such as the Sargasso Sea and Southwest Pacific) may be a transient phenomenon with decadal-scale periodicity (Karl, 2002; Moutin et al., 2008). Notably, however, *pstS* shows a different pattern: its abundance is similar to core genes at 25m and 125m, consistent with its presence in all sequenced isolates of *Prochlorococcus*, but is higher than the average core gene at 75m. This spike at 75m may result from its presence in cyanophage genomes (Sullivan et al., 2005; Appendix E), whose representation in metagenomic samples is greatest at 75m at Station ALOHA (DeLong et al., 2006). Similarly, amino acids (Zubkov et al., 2003) and cyanate (Palenik et al., 2003) are thought to be utilized by some populations as nitrogen sources, but the rarity of amino acid and cyanate transport genes, relative to core genes, at all three depths suggests that they are not widely utilized nutrient sources here. Urease and urea transport genes, on the other hand, are relatively abundant at all depths, suggesting that urea or similar small organic nitrogen compounds might be frequently assimilated by much of the *Prochlorococcus* population.

Five flexible genes were detected at even higher frequency than core genes at all three depths, implying more than one copy per cell (a-f in Figure 1B). These genes are all members of multi-gene families that are also present in more than one copy per cell in most *Prochlorococcus* isolates. Two of these gene families encode conserved hypothetical proteins (represented by P9301_07651/10791/12751 and P9301_07881/08771/10041) which are present in at least three copies per cell in HL *Prochlorococcus*, suggesting critical biological functions. Each copy may serve a different functional role, as supported by the fact that these genes were specifically upregulated in response to nitrogen starvation (Tolonen et al., 2006), phosphate starvation (Martiny et al., 2006), high light shift (Kettler et al., 2007), and phage infection (Lindell et al., 2007; Appendix D) (Table 2). Similarly, gene families encoding lycopene cyclases (P9301_06601/11701), Pcb light harvesting antenna proteins (P9301_06541/13721), and fatty acid desaturases (P9301_15721/15761) were detected more frequently than the average core gene at all depths, and they are, with few exceptions, multi-copy in *Prochlorococcus* isolate genomes as well. The preservation of these gene stoichiometries in our dataset reinforces the quantitative, representative nature of this dataset, and points to a few specific gene families for which having multiple copies is beneficial in *Prochlorococcus* cells.

Undoubtedly, numerous flexible genes that have never been seen in *Prochlorococcus* isolates are also present in the samples but invisible to our analyses. Among the sequenced HL *Prochlorococcus* isolates, 75.4% of the total length of protein coding genes comes from core genes and 24.6% from flexible genes. Therefore we would expect these same percentages of the *Prochlorococcus* protein-coding reads in the wild population to come from core and flexible genes, respectively. We observe, however, 84% of the reads assignable to a gene cluster come from core genes and 16% from flexible. Thus we are missing roughly 10% of the protein coding reads from the average genome (assuming similar genome sizes and core/flexible gene ratios as in cultured isolates). Flexible genes with higher sequence divergence and shorter length, and those with paralogs in different gene clusters, are more difficult to assign unambiguously to a gene cluster by the conservative method we used. With this in mind, we can account for about a third of the missing flexible gene reads: they were assigned as *Prochlorococcus*, but

Table 2. Paralagous genes of unknown function: are they functional? These genes were detected at higher frequency than core genes in natural populations and appear multicopy in genomes of cultured isolates. In several cases these genes are upregulated in response to various stressors, and different gene copies from the same gene cluster appear to respond specifically to different stressors, suggesting unique functions for each copy. All experiments were done with strain MED4. A "+" symbol indicates significantly higher expression in the experimental treatment than in the control. Cluster designations are from Kettler et al. 2007 (Appendix C); phosphate starvation data from Martiny et al. 2006 (Chapter 3); nitrogen starvation from Tolonen et al. 2006; phage infection from Lindell et al. 2007 (Appendix D); high light shift from Steglich et al. 2006.

clusterID	MIT9301 locus	MED4 locus	phosphate starvation	nitrogen starvation	phage infection	high light shift
16	P9301_07651	PMM0701				+
16	P9301_10791	PMM0996		+		
16	P9301_12751	PMM0368		+	+	
16		PMM0379				
16		PMM1409	+	+		
240	P9301_08771	PMM0811				
240	P9301_10041	PMM0858		+		
240	P9301_07881	PMM0734				

could not be unambiguously mapped to one gene cluster (Table 1). The rest, however, must not be recognizable as *Prochlorococcus* at all. To fully capture this "foreign" diversity within *Prochlorococcus* cells requires taxon-specific cell sorting followed by sequencing — a task which is feasible for *Prochlorococcus* given its unique flow cytometric signature. This work is underway.

We next looked for significant differences in gene frequencies between depths, hypothesizing that the environment would select for different gene content in the flexible genome at each depth. The water column at Station ALOHA is often stably stratified and it is therefore plausible that cells at 25m in the mixed layer have been separated from cells at 75m and 125m, below the mixed layer, for some time. For each gene at each depth, we estimated the multiplicity per genome (i.e. copy number). Genes with multiplicity less than one are present in only a subset of the population, while genes with multiplicity greater than one are multi-copy in at least some cells, or are present in unrelated (i.e. not *Prochlorococcus*) genomes. We found 9 flexible genes with significantly different multiplicity ($p<0.001$) between 25m and 75m, 22 genes between 25m and 125m, and 35 genes between 75m and 125m (Suppl. Table 2). These genes with differential distributions across depth could arise two ways: genes that are found exclusively in one ecotype will track with the abundance of that ecotype, while genes that are differentially distributed within an ecotype may have variable distributions. Therefore we sought to differentiate sequence fragments by ecotype as well as by gene cluster.

Ecotypic structure and variability derived from the assemblage genome

To understand the contribution of different ecotypes to these overall meta-population genome signatures, we further discriminated DNA reads by ecotype, as defined according to their ITS phylogeny. Reads were assigned to an ecotype if and only if they shared highest sequence similarity with the cultured, fully sequenced representatives of their ecotype, to the exclusion of other strains. Over 80% of the genomic DNA reads were unambiguously assigned by this method (Table 1), suggesting that most genes have been vertically inherited or exchanged only within the ecotype clusters defined by the ITS sequence (Rocap et al., 2002). The relative abundance of each ecotype based on core genome fragments correlated with that measured by quantitative PCR for the ITS region, but the core genome reads give higher estimates of abundance for the three lower-abundance ecotypes (Figure 2). This discrepancy suggests that genome-wide sequence similarity is apparent within each ecotype even when ITS primers fail to detect some genotypes. Moreover, these data indicate that short (~100bp) sequencing reads are generally sufficient for characterizing the ecotype composition of *Prochlorococcus* in natural communities, largely because of the availability of multiple reference genomes (Edwards et al., 2006; Huson et al., 2007; Wommack et al., 2008).

<u>"Missing" core genes and potential recombination</u>

Because core genes are found in every cell and tend to have phylogenetic signal consistent with the ITS tree (Kettler et al., 2007; Appendix C), we expect that, for any given core gene, the number of eMIT9312-assignable reads will closely mirror the total number of *Prochlorococcus* reads, and their ratio will equal the relative abundance of eMIT9312 cells. Unexpectedly, we find a number of core genes that

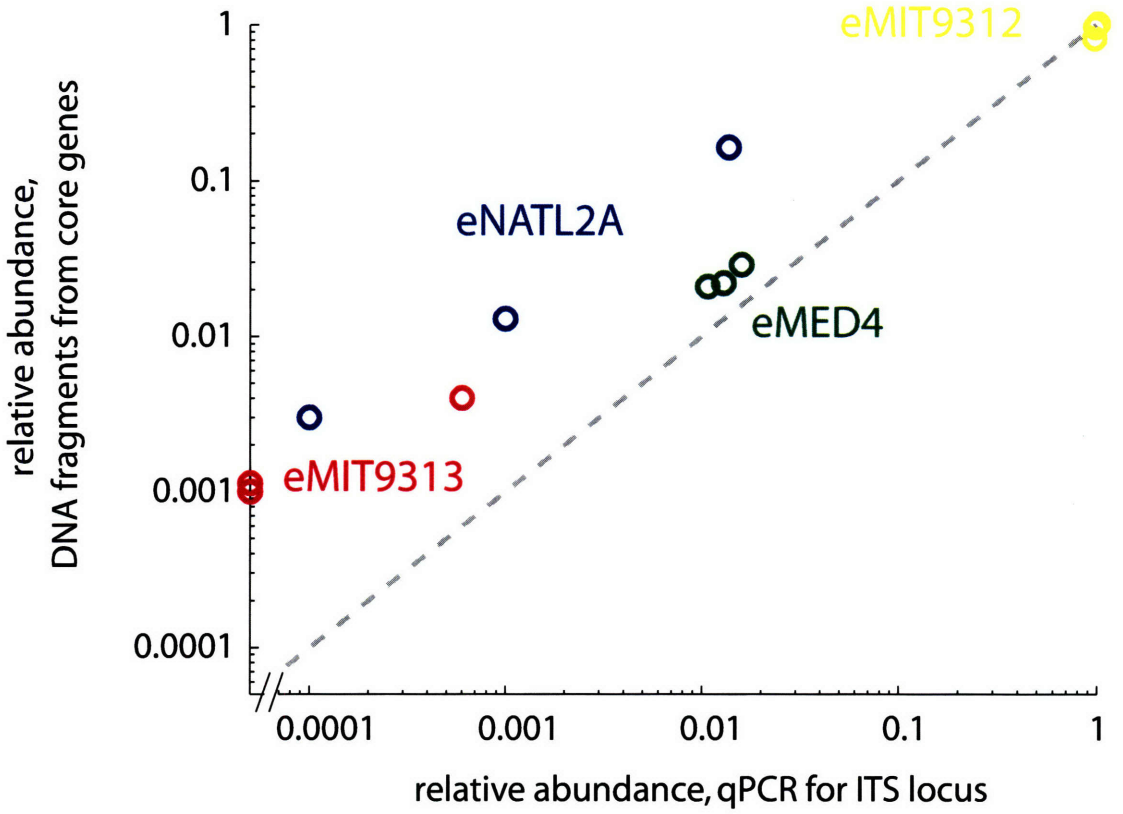


Figure 2. Relative abundance of *Prochlorococcus* ecotypes measured by qPCR for the ITS locus (x-axis) and by ecotype-assignable core gene fragments (y-axis). The gray line represents a 1:1 correspondence. For the less abundant ecotypes, core genome fragments give higher estimates of ecotype abundance than qPCR. This suggests that ecotypes are recognizable as sequence clusters genome-wide, even when qPCR primers specific for an ecotype fail to detect the ITS locus.

are observed less frequently than expected in the eMIT9312-assigned reads (Figure 3; Suppl. Table 3). These underrepresented core genes are likely present in every eMIT9312 cell, based on the gene cluster analysis described above, but could not be assigned unambiguously to one ecotype. This might be expected for highly conserved genes with nearly identical gene sequences across ecotypes. To test this, we compared the percent identity (from BLASTN) for each read aligned to its target gene, for reads that could or could not be assigned to one ecotype ("ecotype-assignable" and "ecotype conflicts", respectively). If the underrepresentation of core genes in the eMIT9312-assignable reads were due to high sequence identity, we would expect the conflict reads to be enriched in high identities, but in fact we see the opposite (Figure 4). Another explanation could be that the underrepresented core genes have relatively low sequence identity, and that this divergence prevents unambiguous ecotype assignments. This may be true for some genes or some variable regions within an otherwise conserved gene but in general, reads shared at least 80% identity with their target gene (Figure 4), which should be sufficient for ecotype assignment. It is possible that different evolutionary rates in particular genes and particular lineages could account for some of this signal; recently developed methods have begun to explore this scenario (Shapiro and Alm, 2008).

Another explanation is that recombination events have blurred the phylogenetic signal for a particular gene. Recombination will result in a hybrid sequence that has portions derived from distinct sources, so that different regions of the gene will show conflicting phylogenetic histories. This hypothesis is supported by the fact that the underrepresented core genes are significantly clustered on the chromosome in cultured isolates, suggesting regions larger than a single gene are involved in recombination (Suppl. Table 3). For example, the eight genes P9301_03041-03111, encoding several enzymes including histidine biosynthesis and DNA repair, are underrepresented in eMIT9312-assignable reads. Sequence similarity among cultured isolates reveals a probable recombination event: five HL isolates share high similarity with each other in this region, but strain MIT9515 is distinctly different from its HL counterparts. Beyond the edges of this region, similarity among the HL isolates returns to its usual state: the HLII strains are most similar to each other, as are the HLI strains (including MIT9515). The high sequence coverage obtained via pyrosequencing enables detection of such patterns and anomalies that might otherwise be overlooked, and in turn points us back to whole genomes where we can understand the behavior of contiguous genome regions.

The list of underrepresented core genes includes essential genes for *Prochlorococcus* metabolism and we hypothesize that these undergo recombination as well. The phylogeny of *amtB*, for instance, encoding an ammonium transporter, is not congruent with the ITS phylogeny because strain MIT9312 does not cluster with its eMIT9312 sister taxa (Figure 5a). Upon closer inspection, we find that trees constructed from different parts of the gene have different topologies, another indicator of recombination (Figure 5b-e). We used the program RDP3 (Recombination Detection Program; Martin et al., 2005) to look for further evidence of recombination using a variety of methods and algorithms. An example of such evidence is shown in Figure 5g-h for the *amtB* gene. An *amtB* sequence recovered from Station ALOHA in a fosmid clone shares high similarity with MIT9312 on its outer edges, while in the middle it is more similar to MED4 and other HL isolates. Thus we can infer a recombination event that swapped

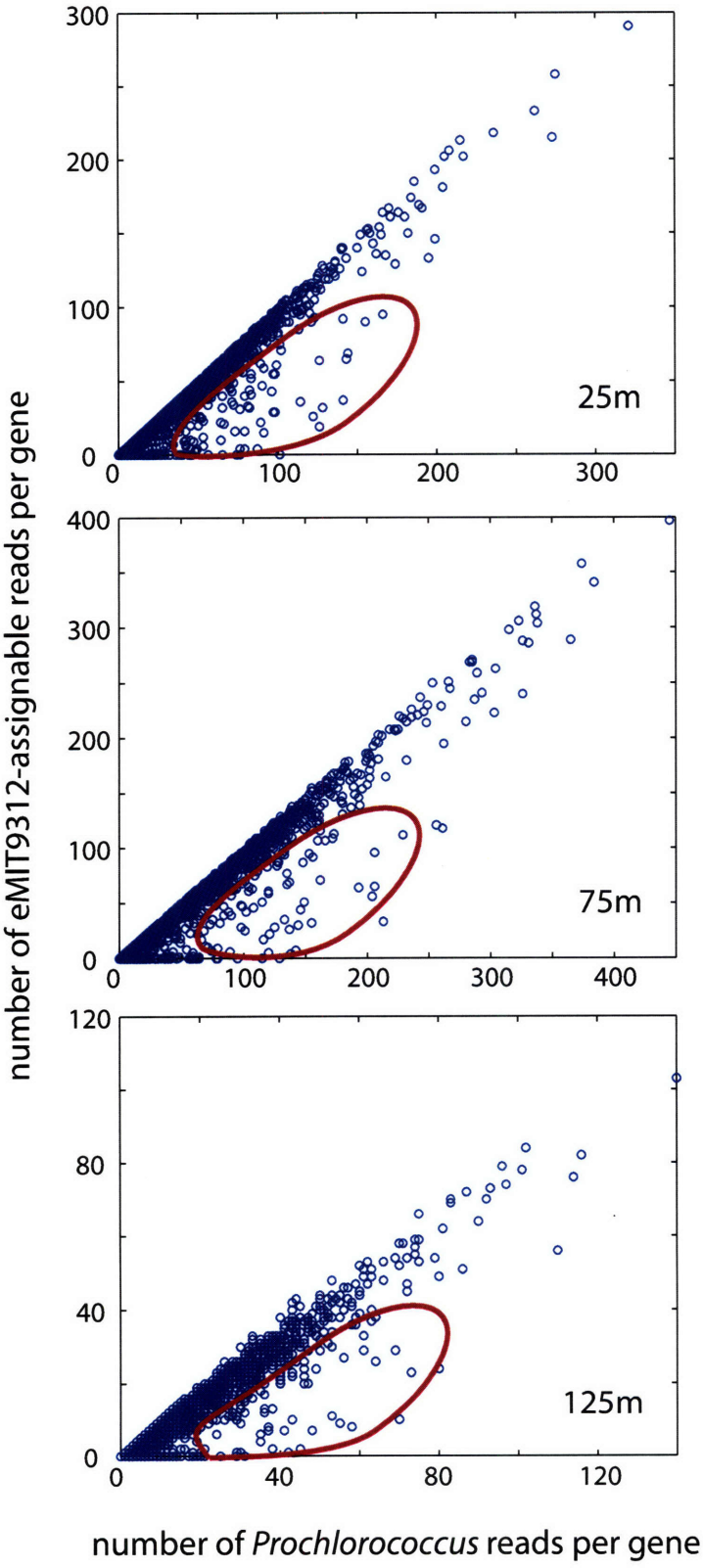


Figure 3. Underrepresentation of numerous core genes in eMIT9312, relative to their abundance in all *Prochlorococcus*. Each point represents a single-copy core gene, with the number of eMIT9312-assignable reads for that gene vs. the total number of *Prochlorococcus* reads for that gene. We expect to see a linear relationship between the two, reflecting the contribution of eMIT9312 cells to the *Prochlorococcus* population. However we see many points below the trend (circled in red) indicating genes that are underrepresented in eMIT9312 relative to the entire population. This is likely due to several factors, including high sequence conservation and intragenic recombination, that blur the ecotype signal for certain genes.

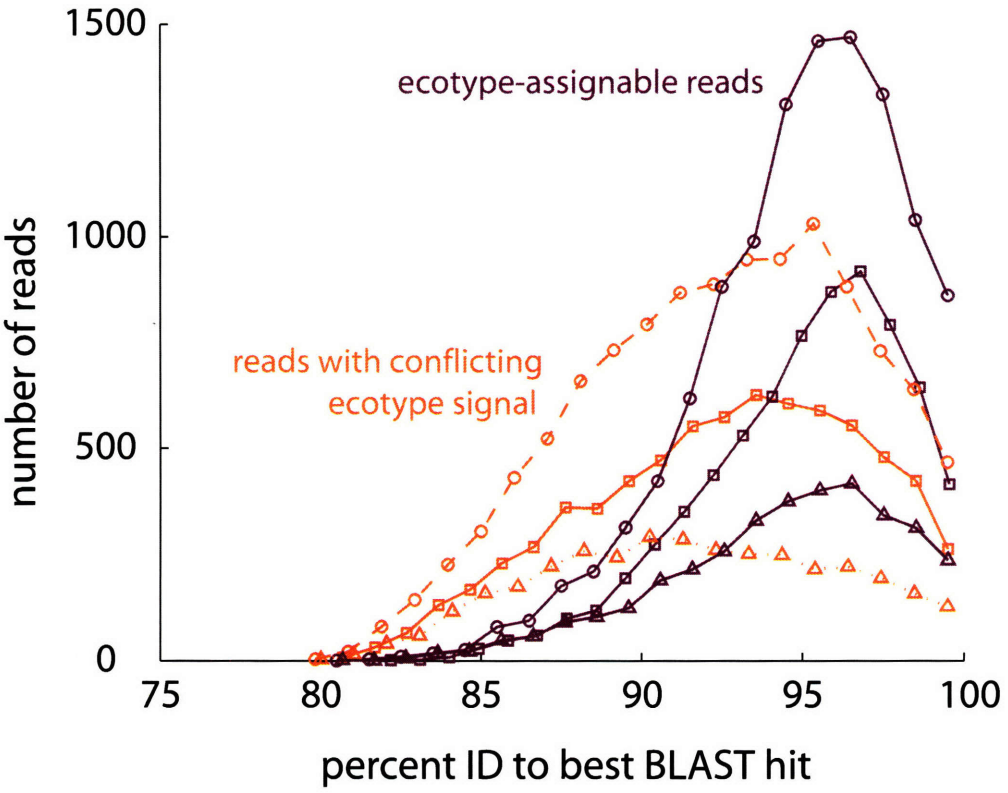


Figure 4. Histogram of ecotype-assignable reads (purple) and reads that could not be assigned to an ecotype (orange), binned by percent identity between each read and its best blastn hit in *Prochlorococcus* genomes. Distributions for samples from 25m (squares), 75m (circles) and 125m (triangles) are shown. Reads that could not be assigned to an ecotype have generally lower percent identity values than ecotype-assignable reads. This indicates that the inability to assign those reads to an ecotype is not a result of very high sequence conservation between ecotype representatives.

the middle portion of the gene, resulting in a mosaic. If these breakpoints are hotspots, i.e. if recombination tends to occur in the same part of this gene over and over, then we might expect the regions around these breakpoints to have blurred phylogenetic signal, and therefore sequence reads from these regions would tend not to be ecotype-assignable. This is indeed what we see: the ecotype-assignable reads for *amtB* tend to match the beginning, end, and middle of the gene, but not the intermediate portions, perhaps due to recombinant phylogenetic signal (Figure 5h). Using similar methods, we found significant evidence for recombination in *csoS2*, *psaA*, and *rbcL* as well (data not shown). Recent work has documented intragenic recombination in the *psbA* gene, both within the cyanophage gene pool and between phage and hosts (Sullivan et al., 2006). Together, these findings suggest that recombination is far more important for genome evolution in *Prochlorococcus* than previously thought. Despite their short length, pyrosequencing reads can be useful for detecting recombination and other anomalous sequence properties at multi-gene and intragenic scales, especially when used as a screening tool for downstream analyses with longer sequences.

Ecotypic and environmental signals in the flexible genome

Reads from the flexible genome were less likely to be assignable to an ecotype, reflecting both their sequence divergence and their often sporadic distribution among strains (Kettler et al., 2007, Appendix C). Moreover, ecotype assignments, particularly for flexible genes, are only inferences; genes acquired by horizontal gene transfer from another ecotype would be indistinguishable by this method from their 'native' counterparts. With these limitations in mind, we explored the ecotypic breakdown of the flexible genes. For abundant flexible genes (with average multiplicity near 1 copy per cell) we can infer that they came from eMIT9312 cells even if an ecotype assignment was not possible, since most cells in these samples are of the eMIT9312 ecotype. Rare flexible genes, on the other hand, could come from one of the minor ecotypes exclusively, or could be rare in eMIT9312 cells. To gauge the contribution of rare genes by the minor ecotypes, we examined whether the prevalence of a given flexible gene in six whole genome sequences from HL isolates could predict its prevalence in these samples. Greater than 98% of cells at 25m and 75m, and over 83% at 125m, belong to HL ecotypes, so we expect the prevalence of a gene among HL isolates to be a predictor of its abundance here. In accord with this, we observed that genes that are abundant in the six HL isolate genomes (2 from the Equatorial Pacific, 2 Atlantic, 1 Mediterranean, 1 Arabian Sea) also tend to be abundant in these subtropical Pacific samples, while genes that are rare in the isolates tend to be rare in these waters (Figure 6). Genes that are absent from all six HL isolate genomes but are present in these samples (gray circles in Figure 6) are likely attributable to LL ecotypes; accordingly these genes are more common at 125m where the eNATL2A ecotype is abundant. Thus we can conclude that genes rare among total *Prochlorococcus* represent both rare genes within the dominant eMIT9312 ecotype and relatively abundant genes exclusive to other ecotypes.

There are a handful of counterexamples (nearly all hypothetical proteins) which are present in all or nearly all HL isolate genomes and nonetheless are quite rare in the metagenome samples (red circles near the horizontal in Figure 6), or, conversely, which are seen in only one or two whole genomes but appear nearly single-copy per cell in the metagenome samples (blue circles along the diagonal trend). One set of

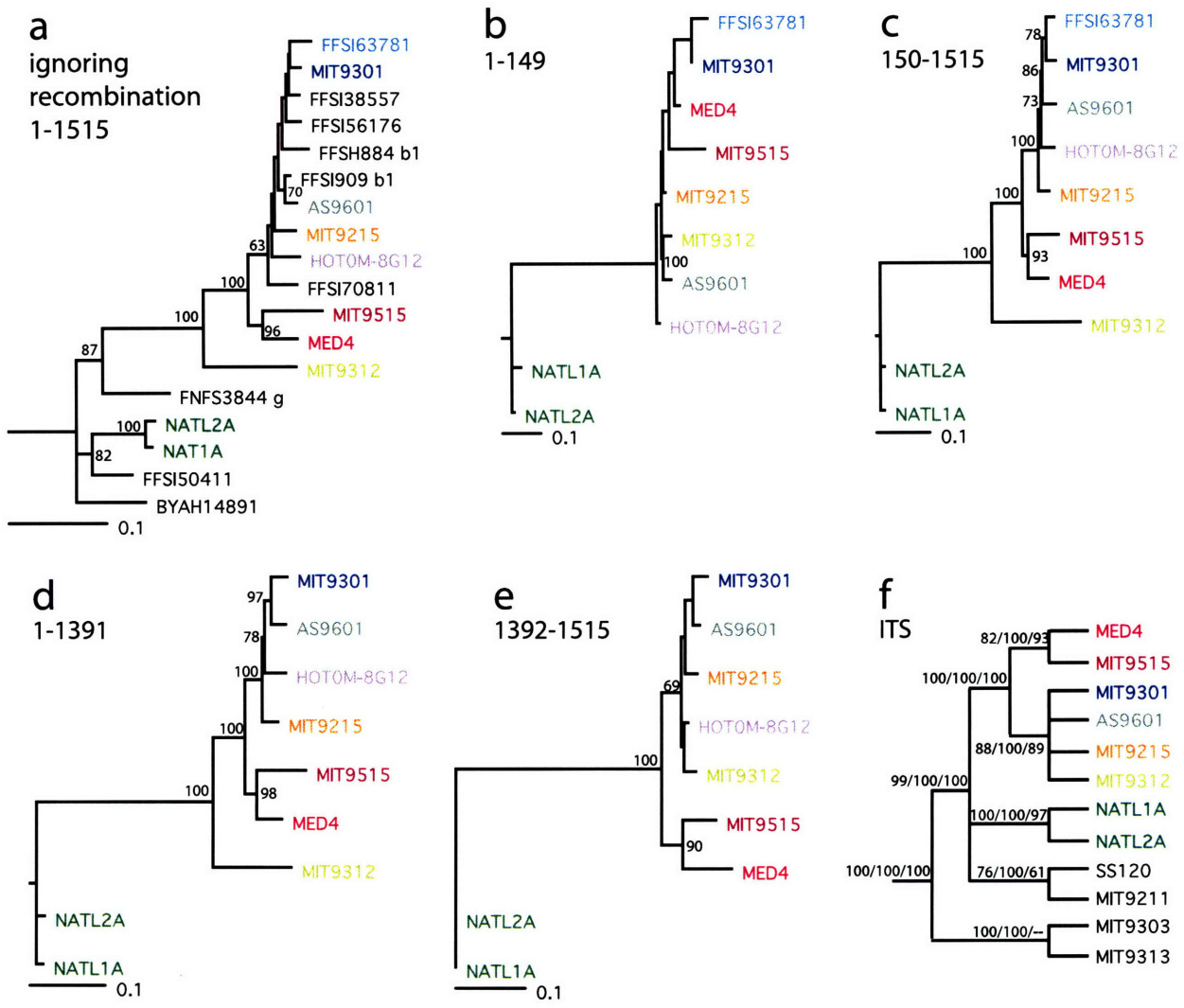


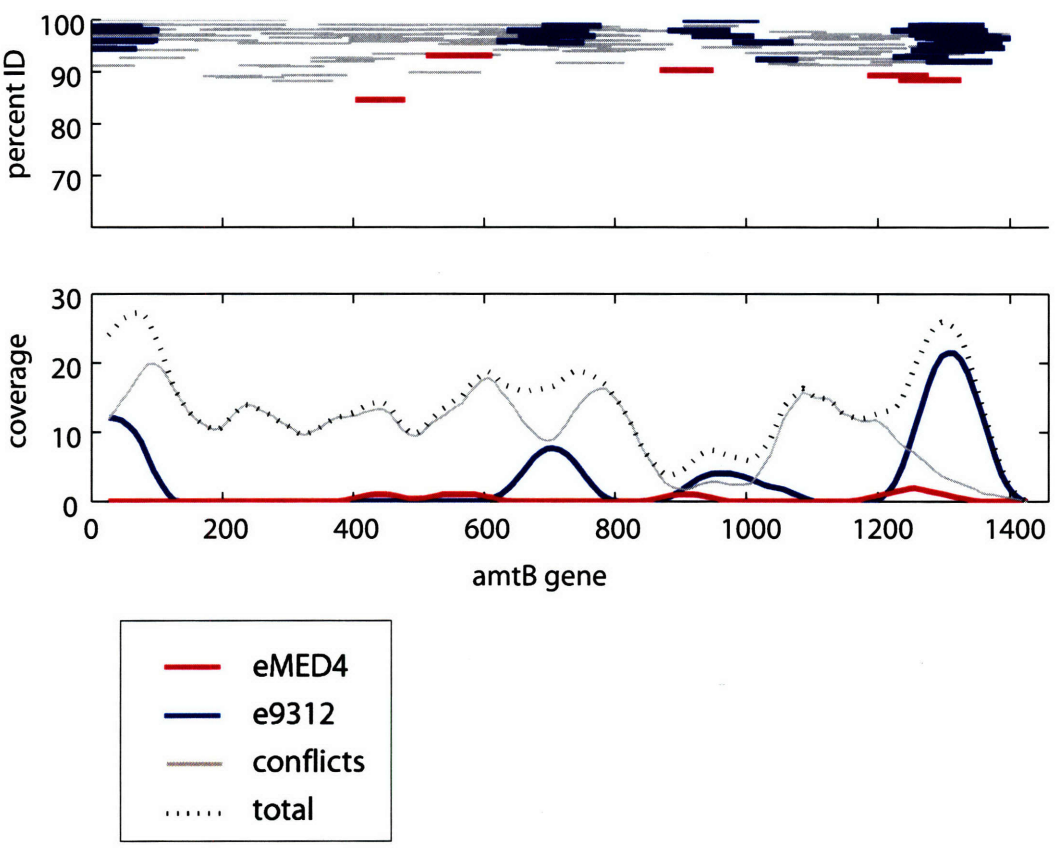
Figure 5. Evidence for recombination in *amtB*, an ammonium transporter gene. (A) Neighbor-joining phylogenetic tree for the full length (1515 positions) of the *amtB* alignment. Note that this topology is incongruent with the ITS phylogeny (shown in (F)): MIT9312 does not group with its sister taxa. Trees for discrete portions of the alignment are shown in (B) positions 1-149, (C) positions 150-1515, (D) positions 1-1391 and (E) positions 1392-1515; and for the ITS locus in (F) (Kettler et al. 2007, Appendix C). Each shows a different topology (though many nodes are unsupported) and only the last recovers the ITS phylogeny. These distinct phylogenetic signals from different portions of the gene are suggestive of recombination. (G) Likelihood of recombination breakpoints occurring at positions indicated, as calculated by a variety of methods in RDP. (H) Recruitment and coverage plot for HOT metagenomic reads along the length of *amtB*. Blue, e9312-assigned reads; red, eMED4-assigned reads; gray, reads unassignable to an ecotype. Ecotype-assignable reads cluster in a few sections of the gene, while the remainder has a blurred phylogenetic signal.

g

			Detection Method					
event	**begin**	**end**	**Bootscan**	**Maxchi**	**Chimaera**	**SiSscan**	**LARD**	**3Seq**
1	1392	1515*	1.57E-04	NS	NS	1.07E-08	NS	NS
2	1*	149	2.70E-04	1.58E-02	6.68E-03	7.20E-09	4.55E-09	8.55E-03

* = The actual breakpoint position is undetermined
NS = not significant

h



genes found in ISL4 (Coleman et al., 2006) in 5 or 6 HL isolate genomes is very rare at all 3 depths, demonstrating the especially high gene content variability and/or sequence divergence of the genomic islands. Population metagenomics is particularly valuable for these hypervariable regions, which are more likely to present a biased picture on the basis of sequencing a smaller number of cells, particularly with traditional cloning-based sequencing. Another anomalous set of genes, P9301_13111 through 13161, are only found in the genome of MIT9301, but appeared at much higher frequencies than expected, especially at 25 and 75m (indicated on Figure 6). These genes seem to form a cassette and transfer together, and their observed abundance may be related to water-column cycling of thiamin; these are discussed in detail below. Finally, two genes found in 3 or 4 of the HL isolate genomes appeared so frequently in the DNA samples as to be clearly single-copy per cell (Figure 6): a pyrimidine photolyase (P9301_03921, present in 4 genomes) and a fructose bisphosphate aldolase (FBA) (A9601_08451, 3 genomes). Overall, we can conclude that our isolate genomes are fairly representative, to a first approximation, of natural populations.

Using these ecotype assignments, we can now reexamine genes that are present in different frequencies between depths. In particular, we would like to deconvolve the effects of ecotype and depth on gene abundance – that is, to understand to what exent a gene's prevalence is a function of the ecotype structure of the population and to what extent it specifically reflects adaptation to prevailing physicochemical conditions at a position in the water column. Since the genomic factors that drive differential success of *Prochlorococcus* ecotypes remain incompletely understood, and likely include allelic and regulatory variations as well as gene content, it is important to understand which genes may be 'tagging along' with changes in ecotype abundance (while not necessarily driving those changes themselves) and which genes have purely depth-related trends in frequency. As an example, a gene that is common to HL but not LL strains is expected to become less abundant with depth, whether or not it affords a fitness advantage in and of itself. But were that gene to reverse the ecotype trend and become *more* abudant with depth, we might hypothesize that it is involved with enabling growth of HL-adapted cells under lower-light conditions. Examining the depth distributions of gene abundance with the ecotype framework in mind will help in understanding both the genomic basis of ecotypic differentiation and spatial patterns of selection in natural environments.

A total of 50 flexible genes show significant multiplicity differences between depths, as listed in Supplemental Table 2. They are not a functionally well-characterized group: 24 are hypotheticals and the annotations of most others are uncertain or general. While we do not see signals of large-scale depth differences in identifiable cellular biochemistry, some abundance trends are apparent. First, the genes appear on the list because they are more abundant at 25 and/or 75m than at 125m. The reverse pattern (more abundant at 125m than the upper two depths) is not observed, with the exception of the single LL-specific gene on the list, NATL2_14561, which shows a clearly ecotype-related rise rise in multiplicity from 0.0 to 0.21. This gene is similar to membrane-associated hydrolases, but its role in the deep euphotic zone is unknown. This suggests that the issue of 'invisible' flexible genes described above becomes more acute at 125m, a result not unexpected from the phylogenetically sparser representation of LL strain genomes in our collection.

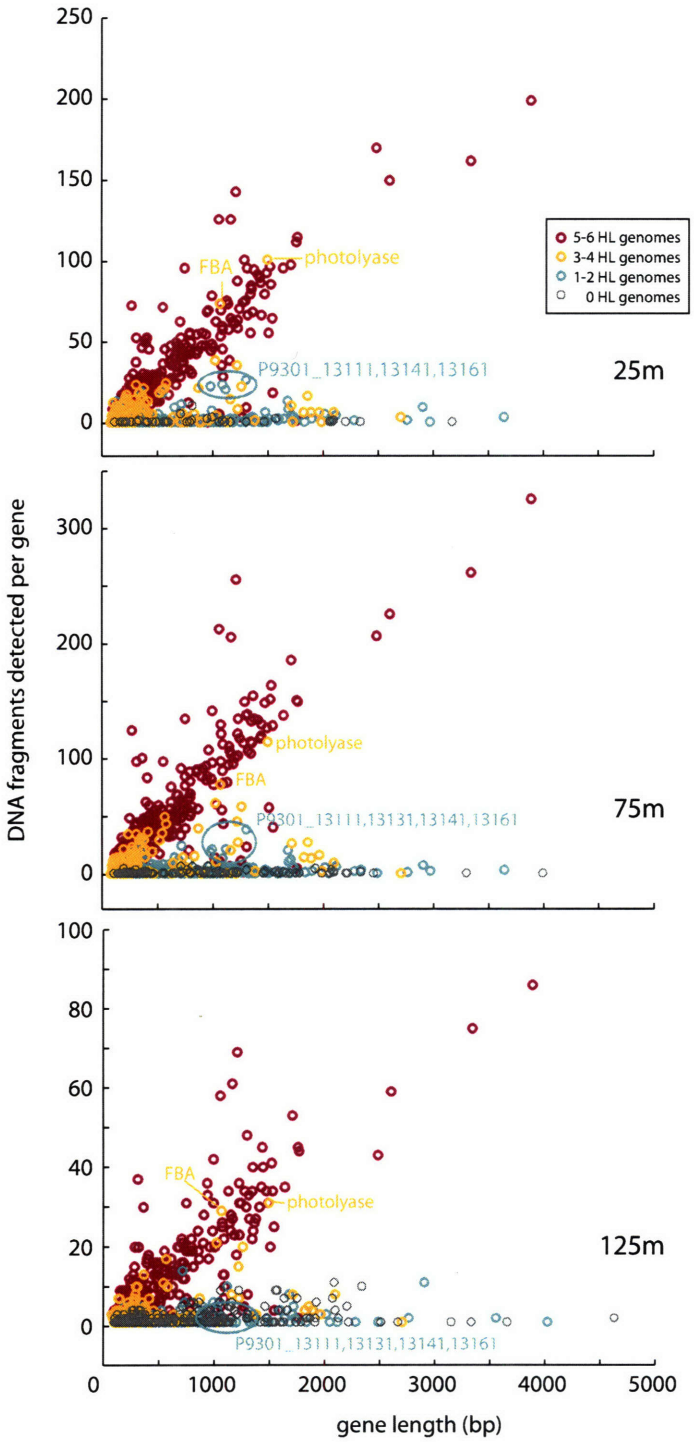


Figure 6. Occurrence of flexible genes in sequenced *Prochlorococcus* isolates and in natural populations. All points are the same as the pink points (i.e., flexible genes) in Figure 1B, but here colored by the number of HL genomes that carry the gene. Genes found in nearly all HL genomes are also found, on average, in nearly all individuals in these samples. Genes found in few or no HL genomes tend to be less abundant in these samples as well. A cassette of syntenous genes (indicated by P9301_13111-13161) found only in MIT9301 was unusually abundant, especially at 25 and 75m; some of these genes were hypothesized to be involved in thiamin metabolism. Two genes (pyrimidine photolyase and fructose bisphosphate aldolase (FBA)), were present at single-copy per genome frequency, despite being present in 4 and 3 HL isolate genomes, respectively.

It appears that HL cells at depth are different from those at the surface, as evinced by a second predominant pattern in gene abundance with depth: 35 genes show a larger drop in abundance between the shallower samples and 125m than the ~20% drop that would be expected from the shift in ecotype composition alone. The presence of ~20% LL cells (mostly eNATL2A) at 125m means that single-copy, HL-specific genes would be expected to drop about 20% in multiplicity, yet many of these fall by more than 50% or disappear altogether. These genes must be less prevalent among the HL (and mostly eMIT9312) cells living at 125m than those in the upper 75m. This is the first evidence of a genetically distinct population of HL-ecotype cells living at depth. A preliminary hypothesis is that at least some of these deep-dwelling HL cells have specialized for that environment, and are not simply surface-adapted cells that have been mixed downward. Testing of this hypothesis will require further metagenomic characterization of the deep euphotic zone; this twilight region is a promising area to prospect for new *Prochlorococcus* strains and genes.

Variability with depth and ecotype: the case of thiamin metabolism

One gene that shows this greater-than-expected dropoff at 125m is *thiD*, a kinase involved in the biosynthesis of thiamin. Thiamin is an important cofactor for enzymes in central carbon metabolism, participating in glycolysis and the citrate, pentose phosphate and Calvin cycles. Most bacteria, including *Prochlorococcus*, can synthesize thiamin, but it is an essential vitamin for a number of eukaryotic algae (Croft et al., 2006) and is maintained at picomolar concentrations in seawater (He et al., 2005; Okbamichael and Sañudo-Wilhelmy, 2005), suggesting active cycling in the water column. Thiamin has a short half-life in seawater — about 6hr in warm sunlit waters (Gold 1964; Okbamichael and Sañudo-Wilhelmy, 2005) — so salvage pathways might be expected to be a part of this cycling. The biosynthesis of thiamin is shown in Figure 7a. Dedicated thiamin biosynthesis begins with the conversion of aminoimidazole ribonucleotide (AIR) (an intermediate of purine nucleotide biosynthesis) to hydroxymethylpyrimidine (HMP) by the enzyme ThiC. HMP is then phosphorylated by the kinase ThiD to HMP-PP. The enzyme ThiE then condenses HMP-PP with a thiazole phosphate moiety, which derives from a separate branch of the pathway, to yield thiamin monophosphate. The biologically active cofactor form, thiamin pyrophosphate, is generated by one further phosphorylation catalyzed by ThiL.

These genes have an intriguing distribution in both *Prochlorococcus* isolates and the water-column metagenome samples, as shown in Figure 7b. All *Prochlorococcus* isolate genomes contain genes for ThiC, ThiE and ThiL, as well as the PurM protein that synthesizes AIR. LL *Prochlorococcus*, however, lack ThiD, and may use an alternative (and as yet unidentified) kinase to produce HMP-PP. There are a number of candidate kinases with unknown substrates in *Prochlorococcus* genomes, and both *E. coli* and *B. subtilis* have been found to have alternative HMP kinases (Park et al., 2004). As noted by Kettler et al. (Kettler et al., 2007, Appendix C), all HL *Prochlorococcus* do have *thiD*, and it resides in a predicted operon with a gene called *tenA* (Figure 7b). TenA has recently been found to catalyze the hydrolytic deamination of aminomethylpyrimidine, regenerating HMP in a novel thiamin salvage pathway (Jenkins et al., 2007; red arrows in Figure 7a). The presence and synteny of these two genes in all 6 HL isolates suggests that HL *Prochlorococcus* may both synthesize and scavenge thiamin.

Moreover, the operon structure of *thiD* with *tenA* might mean that ThiD in HL cells is more involved in scavenging than biosynthesis *per se*, and that HL *Prochlorococcus*, like their LL counterparts, might possess a different kinase to produce HMP-PP in primary biosynthesis.

These six thamin genes occur at multiplicities near one (i.e., approximately single-copy per *Prochlorococcus* genome) at 25 and 75m, in accord with their uniform presence in HL isolate genomes (Figure 7b). As noted above, *thiD* decreases in abundance significantly at 125m, falling from a multiplicity of 0.9 to 0.3. This is a functionally significant example of HL cells being genetically distinct at depth. We can also detect this result by looking at the eMIT9312-assignable reads: the multiplicity of *thiD* in eMIT9312 cells is 1.0 at both 25m and 75m, but drops to 0.4 at 125m. Its predicted operon partner *tenA*, though, does not appear to be lost to the same extent at 125m, suggesting differential retention of the two genes. It may be the case that HL cells at 125m rely more on the unidentified HMP kinase that they share with LL cells, and *thiD* is not maintained. In the upper water column, however, it is apparently advantageous to maintain *thiD* in the genome, possibly as a dedicated kinase for thiamin salvage.

A second group of potentially thiamin-related genes includes the seven-gene cassette noted above for its unexpectedly high abundance (Figure 6). These genes have been found only in isolate MIT9301, but occur with identical synteny in five strains of *Synechococcus* (Figure 7b) as well as *Synechocystis* and a number of proteobacteria. This group of genes appears to be horizontally transferred as a module, an idea reinforced by the presence of high-light inducible genes — a common feature of hypervariable regions in *Prochlorococcus* genomes (Coleman et al., 2006) — flanking both sides of the cassette in MIT9301. A large-insert *Prochlorococcus* clone from Station ALOHA also contains this cassette, confirming its presence in *Prochlorococcus* cells in this particular environment (Figure 7b). Conserved domains within a few of these proteins hint at their potential biochemical role. P9301_13161 contains two domains found in AIR synthase, which supplies the substrate for ThiC (Figure 7a). P9301_13131 contains an amidohydrolase domain and bears similarity to YlmB, another enzyme of the recently-discovered thiamin salvage pathway that enables cells to use abiotic degradation products of thiamin (Jenkins et al., 2007). The simultaneous presence of TenA and a YlmB-like amidohydrolase suggests that some HL *Prochlorococcus* might make use of this pathway as well.

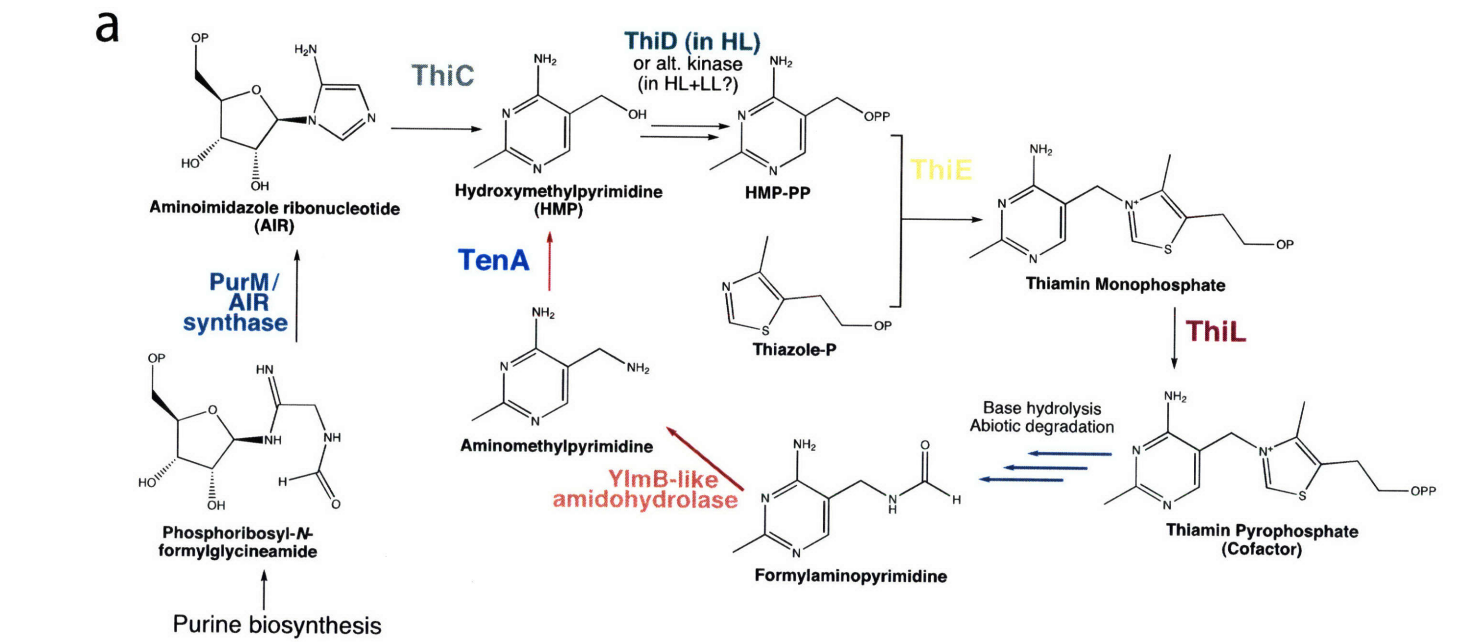
Though they have been observed in only a single *Prochlorococcus* isolate, the genes of this cassette have multiplicities of up to 0.4 in the 25m sample, implying that they are fairly widespread in the upper water column (Figure 7b). While the overall abundances of the putative cassette genes are roughly similar, they do show somewhat different distribution patterns with depth. This may be because some of the genes also occur outside of the cassette in some genomes, or that some genes gained in older cassette-integration events have been lost. Significantly lower abundances at 125m as compared to 25m are observed for the putative FAD-binding protein, radical SAM protein, and AIR synthase, which is consistent with the hypothesis that the maintenance of these genes is related to an upper water column process.

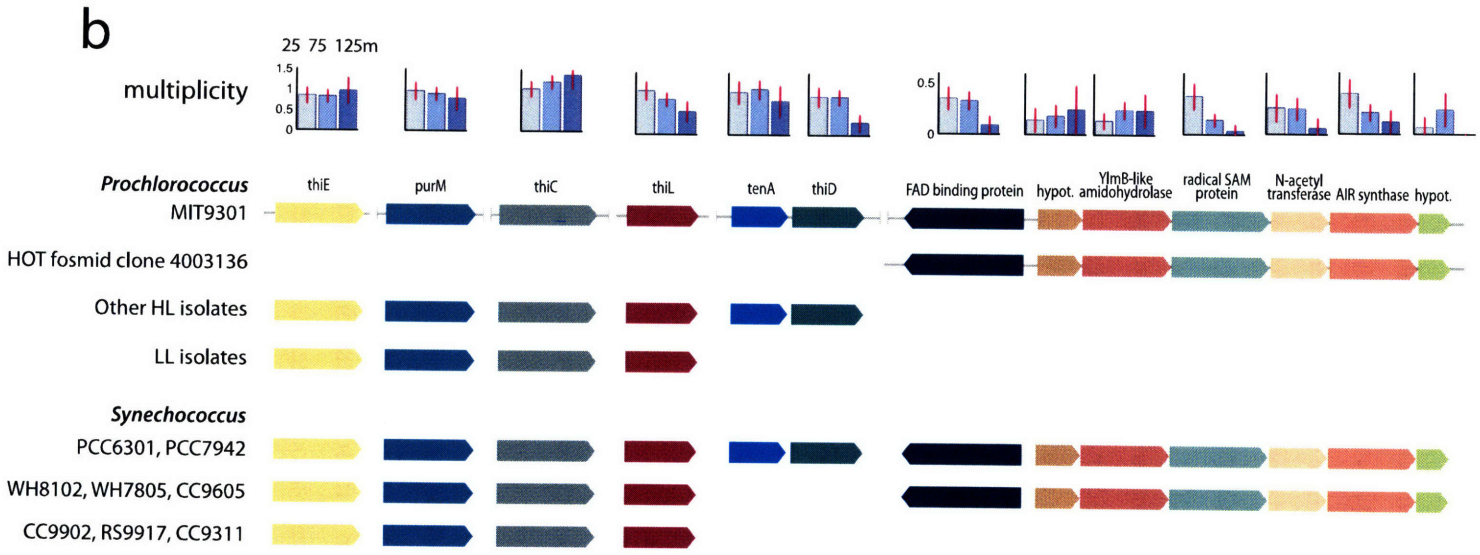
Taken together, the distribution of these thiamin-related genes suggests that *Prochlorococcus* is a participant in an active water-column cycle of thiamin in the subtropical Pacific. Elements of this cycle

Table 3. Number of *Prochlorococcus* gene clusters detected in the genomic DNA and cDNA datasets. Only a very few gene clusters at each depth were detected in the cDNA but not in the DNA, a result of the much greater database sizes in the genomic DNA samples.

Number of gene clusters detected in:					
	cDNA	DNA	cDNA only	DNA only	Both cDNA & DNA
25m	999	1981	12	994	987
75m	898	2141	4	1247	894
125m	942	2109	19	1186	923

Figure 7 (opposite page). Thiamin cycling in the upper water column. (A) Metabolism of thiamin in bacteria. Dedicated thiamin biosynthesis begins with an intermediate of purine nucleotide biosynthesis, amidoimidazole ribonucleotide (AIR), and eventually yields the biologically active cofactor thiamin pyrophosphate. If released into the water column, thiamin undergoes rapid, photochemically mediated breakdown (blue arrows) to degradation products that include formylaminopyrimidine. A newly discovered salvage pathway, indicated by red arrows, allows use of this breakdown product. The formyl group is first cleaved by an amidohydrolase such as YlmB, and the resulting amidomethylpyrimidine is hydrolytically deaminated by TenA, regenerating HMP. (B) Occurrence of thiamin-related genes in *Prochlorococcus/Synechococcus* genomes and the ALOHA metagenome samples. Genes are arranged according to synteny in MIT9301, and their multiplicities in the metagenome samples (25m, 75m, and 125m) are shown in the small bar graphs above each gene. Note the different multiplicity scale between the *thi/pur/ten* genes at left and the seven-gene cassette at right. All sequenced *Prochlorococcus* have a basic set of genes for thiamin biosynthesis, including *thiCEL* and *purM*, but lack *thiD* and so must use an alternative kinase for HMP phosphorylation. HL (but not LL) *Prochlorococcus* and two strains of *Synechococcus* have *thiD* and *tenA*. MIT9301 is the only sequenced strain to contain the indicated seven-gene cassette, which it shares – with perfect synteny – with five *Synechococcus* strains. This cassette has also been found in a *Prochlorococcus* fosmid clone from HOT. All annotations of the cassette genes are putative and based on conserved domain similarity.


a
ThiC
ThiD (in HL)
or alt. kinase
(in HL+LL?)
Aminoimidazole ribonucleotide
(AIR)
Hydroxymethylpyrimidine
(HMP)
HMP-PP
ThiE
Thiamin Monophosphate
ThiL
PurM/
AIR
synthase
TenA
Thiazole-P
Aminomethylpyrimidine
YlmB-like
amidohydrolase
Base hydrolysis
Abiotic degradation
Thiamin Pyrophosphate
(Cofactor)
Phosphoribosyl-N-
formylglycineamide
Formylaminopyrimidine
Purine biosynthesis


b
multiplicity
25 75 125m
Prochlorococcus
MIT9301
HOT fosmid clone 4003136
Other HL isolates
LL isolates
Synechococcus
PCC6301, PCC7942
WH8102, WH7805, CC9605
CC9902, RS9917, CC9311
thiE
purM
thiC
thiL
tenA
thiD
FAD binding protein
hypot.
YlmB-like amidohydrolase
radical SAM protein
N-acetyl transferase
AIR synthase
hypot.

include biosynthesis by bacteria; release by viral lysis, sloppy grazing or exudation; uptake by eukaryotes; abiotic degradation of dissolved thiamin; and scavenging of degradation products. The greater abundance of thiamin-related genes at 25 and 75m suggests that thiamin cycling is most active in the upper water column. This may be related to water column photochemistry: thiamin is known to rapidly photodegrade to stable analogs (Carlucci et al., 1969) so salvaging partially degraded molecules may greatly increase the ability to make use of this nutrient. At the slightly alakaline pH of seawater, thiamin undergoes base hydrolysis to products such as formylaminopyrimidine (Jenkins et al., 2007). *Prochlorococcus*, along with other marine cyano- and proteobacteria, have apparently acquired the ability to recycle those degradation products through a YlmB-TenA pathway. This 'dual fuel' strategy, combining biosynthesis and salvage, may be a way for microbes to sustain cofactor supplies under variable nutrient-stress conditions. The scenario described here, where a central metabolic pathway is elaborated and extended with modules acquired horizontally, is very much like that described by Pal et al (Pál et al., 2005) in a detailed analysis of *E. coli* metabolism. That study found an inverse correlation between the centrality of a gene in a metabolic pathway and the likelihood that it had been involved in a horizontal transfer event. Should the same pattern hold true in marine microbial populations — as our data here suggest — then the fringes of metabolic networks will be prime places to look for adaptive variation.

Three *Prochlorococcus* transcriptomes: the big picture

Just as gene content is a fundamental determinant of microbial community function, so too is the corresponding regulation and expression of these genes. Homologous genes that evolve different regulatory control can result in important phenotypic differences (Winfield and Groisman, 2004). In collaboration with the work of others (Frias-Lopez et al. 2008, Appendix B; Shi et al. *in prep.*), we obtained cDNA sequences from the same seawater samples described above (25m collected at 22:00, 75m collected at 03:30, and 125m collected at 08:00) to explore *Prochlorococcus* metabolism and function. We detected transcripts from roughly 900-1000 gene clusters at each of the three depths, or 40-50% of the gene clusters for which we detected genomic DNA (Table 3). These transcripts, however, were unevenly distributed: in the three samples combined, 149 transcripts accounted for over 50% of the reads, while 25% of the transcripts were represented by a single read. This reflects the unevenness of gene expression in the cell (Zinser et al., *in prep.*) and the fact that the coverage obtained from a single pyrosequencing run is far from saturating the complex microbial community transcriptome. Despite this low coverage, clear biological signals emerge from the data and inform our understanding of the functioning of *Prochlorococcus* populations.

Expression of the core genome

Of the 1221 single-copy core genes, 88% were detected in the cDNA from at least one sample. The core genes not observed in the cDNA do not reside in particular genomic regions, nor do they constitute specific biochemical modules, suggesting that their absence is simply a result of sampling depth. With deeper sequencing and broader sampling over the diel cycle — a strong driver of *Prochlorococcus* metabolism — we suspect that nearly all of the core genes would be detected in cDNA, supporting the

idea that the set of core genes encodes the essential *Prochlorococcus* metabolic machinery.

Among the most abundant transcripts in all three samples are core genes involved in transcription (RNA polymerase and sigma factors), translation (EF-Tu, EF-G, IF-2), photosynthesis (*psaAB*, *psbA*), and nutrient transport (*amtB*). These are central activities of the *Prochlorococcus* metabolism and are also highly expressed in microarray experiments using cultured isolates (Zinser et al., *in prep.*; Martiny et al., 2006; Steglich et al., 2006; Tolonen et al., 2006), so the abundance of these transcripts suggests that cDNA sequencing has provided at least a first-order quantitative picture of *Prochlorococcus* gene expression. Operons provide a further test of the fidelity of environmental cDNA sequencing as a measure of gene expression. Several pairs of genes that lie in predicted operons and that appear to be coexpressed in microarray experiments (Zinser et al., *in prep.*), are also detected in similar ratios in the cDNA (Figure 8a, open circles). In two operons, however, this stoichiometry is not observed in the cDNA fragments (filled red and purple circles in Figure 8a). These two operons each contain one very short gene (*psbT* and *psaI*; Figure 8b). With very short genes on the edges of a transcript, there is lower probability that enough sequence will be obtained to unambiguously map the read to a gene. This effect does not scale linearly with gene length, but rather seems to disappear beyond a length threshold of about 500bp (Figure 8b). Thus short genes may be underrepresented in the cDNA, but for longer genes our results are consistent with predictions from whole genome sequences and microarray experiments.

Among the top 50 most highly expressed core genes in the field data, we found that 16 of them were also underrepresented in the ecotype-assignable reads, suggesting recombination or other unusual sequence features. These include key genes in nutrient uptake (*amtB*), photosynthesis (*psaABFL*, *psbC*, *chlN*), carbon fixation (*rbcLS*, *csoS2*), and transcription and translation (*pnp*, *rpoB-C2*, EF-G). This overlap between potential recombination hotspots and highly expressed genes could have important functional consequences for the cell. An allele with slightly improved catalytic efficiency in such crucial processes as photosynthesis and carbon fixation might spread rapidly in the population via homologous recombination, thereby leading to faster adaptation than through single mutations alone (Lawrence, 2002). Functional differences between variants of these important proteins remain to be explored through biochemical methods.

The correspondence between highly expressed genes and potential regions of recombination may in fact have a mechanistic basis. Transcription has been shown to significantly alter the supercoil structure of DNA, with highly expressed genes having a much stronger effect (Deng et al., 2004). One hypothesis, then, is that highly expressed genes alter the chromosome structure in such a way that homologous recombination is facilitated, perhaps by making certain regions more accessible to recombination proteins. This effect of transcription is complicated by DNA replication, which also changes chromosome organization (Sherratt, 2003). In *Prochlorococcus*, these effects are likely intimately tied to the diel cycle as well, since transcription of most genes exhibits strong daily cycles (Zinser et al., *in prep.*) and replication happens synchronously in natural populations in the evening (Vaulot et al., 1995). Such temporal interactions and their effects on recombination and gene transfer remain to be explored.

Given the critical influence of light and the diel cycle on *Prochlorococcus* physiology, it is not surprising that we observed strong differential expression between the three samples, which were

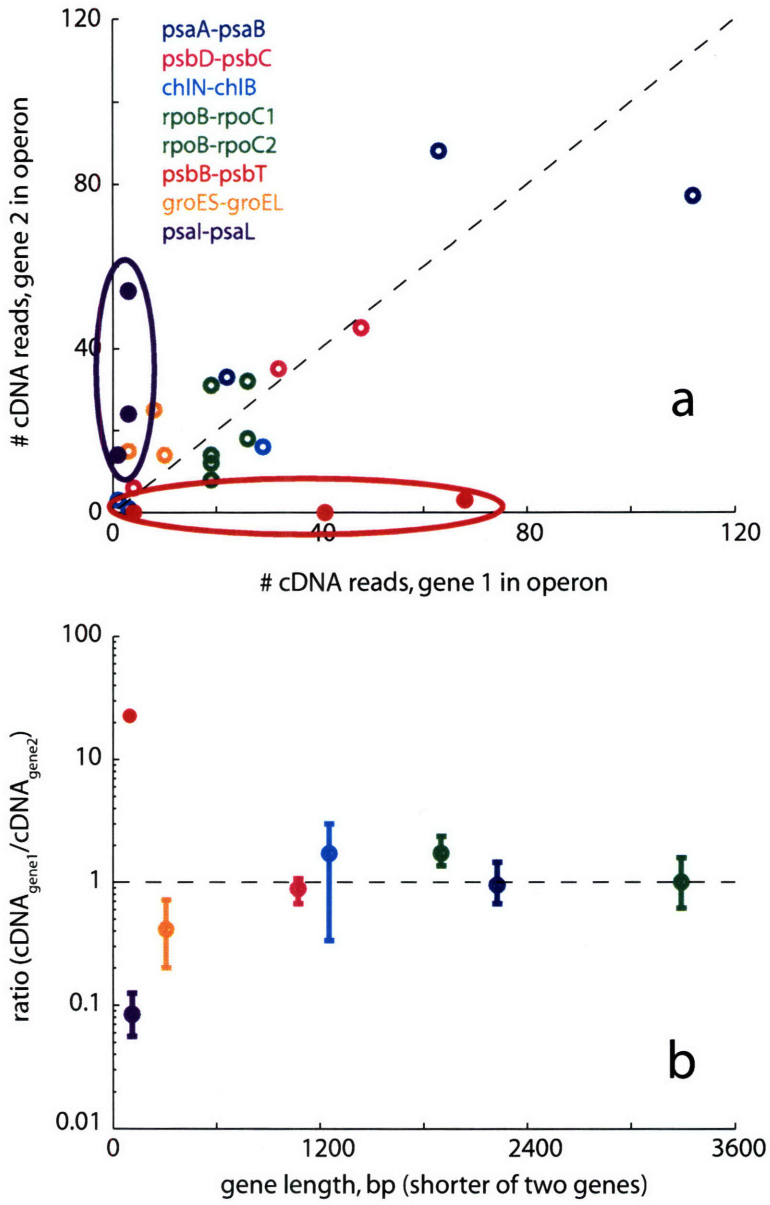


Figure 8. Transcript abundance of pairs of genes in predicted operons. Each color represents a different operon, and three points representing the three samples are plotted for each pair of genes. Gene 1 is the upstream gene. For all operons shown, the genes in the operon had similar transcript abundance in laboratory microarray experiments (Zinser et al., *in prep.*). Two pairwise comparisons are plotted for the *rpoB-C1-C2* operon (*rpoB-C1* and *rpoB-C2*). (A) Number of cDNA reads detected for the upstream gene, plotted against the number of reads detected for the downstream gene. The dashed line represents a 1:1 correlation. In most cases, the two genes are detected at very similar frequencies (open circles). For two operons, however, the frequencies are very different (filled circles). (B) The length of the shorter of two genes, plotted against the ratio of cDNA reads detected for the two genes (mean and range for 3 samples). Only one point is plotted for *psbB-T* because the other two points require division by zero (i.e. *psbT* was not detected in two samples). In the *psbB-T* and *psaI-L* operons, where the two genes were detected with different frequencies, one gene is very short, which could explain its underrepresentation relative to its operon partner.

collected at different times of day as well as different depths. Over 100 genes had significantly different transcript abundances ($p<0.01$) between depths (Suppl. Table 4). *Prochlorococcus* cells divide synchronously in the evening (Vaulot et al., 1995), and thus transcripts of *ftsZ*, for example, were more abundant in the 25m sample collected at 22:00, right after cell division. Carbon fixation gene transcripts including Rubisco (*rbcL, rbcS, csoS2*) were more abundant in the 75m and 125m samples, taken at 03:30 and 08:00, respectively, consistent with maximal carbon fixation in the early morning hours (Bruyant et al., 2005). Thus some of the observed differential gene expression is most readily explained by temporal dynamics, but more work needs to be done to deconvolute the spatial and temporal signals.

To further resolve patterns of diel gene expression, we compared transcript abundance measured by cDNA sequencing to that from microarray analysis, sampled every two hours over a diel cycle in laboratory cultures (Zinser et al., *in prep.*). We found a positive correlation between abundance of a transcript in the lab, measured by microarrays, and in the field, measured by cDNA sequencing (Figure 9a). We then calculated the pairwise correlation coefficients for each time point in the microarray experiment compared to the cDNA frequencies from the three field samples. We predicted that the strongest correlations would arise from the 22:00, 03:00, and 08:00 time points in the microarray experiment — the same time points sampled by cDNA sequencing and with a similar photoperiod in both cases. Remarkably, the strongest correlations were indeed observed near the analogous time point (Figure 9b). Even with limited cDNA sequencing of natural *Prochlorococcus* populations, we were able to detect similar temporal fluctuations in expression as detected by high resolution microarray methods using axenic cultures in a controlled laboratory setting.

Expression of the flexible genome

In contrast to the core genes, the vast majority of which were detected in the cDNA, only 31% (362 of 1188 genes) of the flexible genes observed in the combined DNA from all three samples were also observed in the cDNA. The flexible genome likely contains a number of genes that are nonfunctional and not expressed, such as genes in genomic islands that have not been integrated into *Prochlorococcus* regulatory networks or genes that are decaying and will eventually be lost. On the other hand, flexible genes that are abundant in the population and are essentially "core" in this environment (Figure 1; Suppl. Table 1) are likely to be functional and expressed. Therefore we predicted that the flexible genes we detected in the cDNA would be enriched in these "abundant" (more than 1 copy per cell) or "average" (1 copy per cell) flexible genes. Indeed this was the case: 81% of the 362 flexible genes detected in both cDNA and DNA were as abundant as core genes at all depths and thus behave like core genes in both their DNA abundance and their detectable expression.

A few flexible genes, however, were rare in a particular sample's DNA but were detected in the cDNA from that same sample (53 genes; Suppl. Table 5), suggesting relatively high cellular expression in a subset of the population. One dramatic example is a putative DNA repair gene, rare in the DNA at 25m but detected 12 times in the cDNA (P9301_10691), indicating high expression perhaps due to increased DNA damage in surface waters. A number of these highly expressed rare flexible genes are located in genomic islands in cultured *Prochlorococcus* isolates (Coleman et al., 2006) and may represent recent

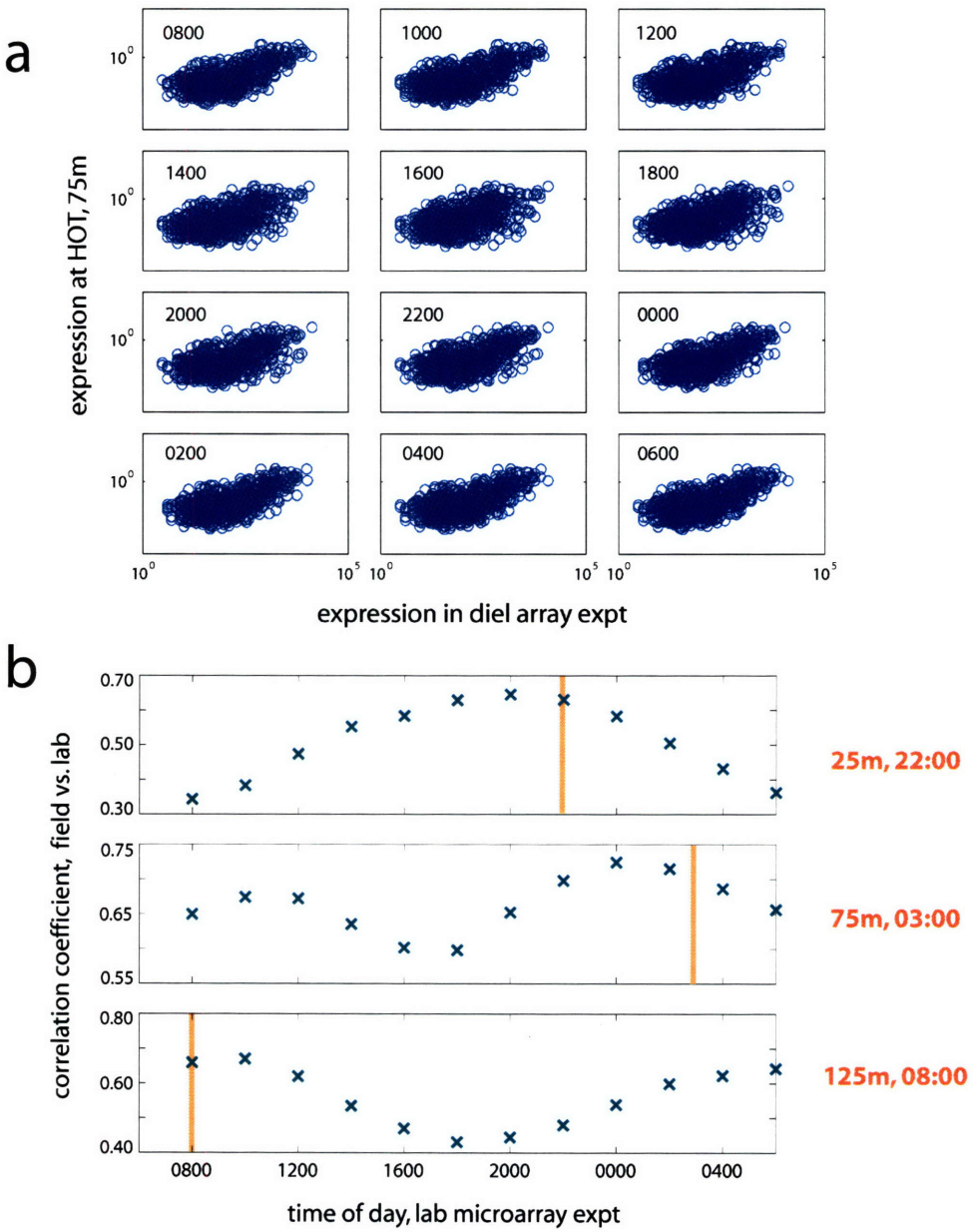


Figure 9. Comparing the transcriptome of wild and cultured *Prochlorococcus*. (A) Relationship between expression inferred from two methods: microarray analysis of cultured isolate MED4 and cDNA sequencing of natural populations at HOT. Each panel shows the relationship between the 75m cDNA sequencing (y-axis) versus one time point from the diel microarray experiment using strain MED4 (Zinser et al., *in prep.*). Each point represents a single core gene: on the x-axis, its normalized expression from microarrays; on the y-axis, its frequency in the cDNA from 75m. (B) Correlation coefficients for all pairwise comparisons of a microarray time point with the cDNA dataset, such as those depicted in (a) for 75m. The vertical line indicates the time of sampling for the cDNA at each depth, and this is where we would predict the best correlation. Indeed, the strongest correlation is observed near this time point.

horizontal transfer events in these natural populations. Examples include a putative cyanate transport gene (PMED4_04031) and an outer membrane secretion system gene (*pulD*, P9301_06961). One notable gene, detected in the cDNA at 75m, encodes a 453 amino acid protein found only in strain AS9601 among *Prochlorococcus*, adjacent to a tRNA gene. The next best hit, with an e-value of e-78, is found in *Mariprofundus ferrooxydans* (SPV1_04378), an iron-oxidizing proteobacterium isolated from the Loihi Seamount (512km from Station ALOHA) at 1100m depth. This gene's occurrence in unrelated taxa and its apparent expression at 75m warrant further study. Rare flexible genes that are detected in the cDNA may also be confined to one of the minor ecotypes due to metabolic constraints or gene transfer limitations and may be highly expressed in these ecotypes. A light harvesting Pcb protein (NATL2_14951), for example, appears limited to LL *Prochlorococcus* based on whole genome sequences (though more distant paralogs are found in HL isolates). It is rare in the DNA at all three depths but is detected in the cDNA at 75m and 125m (Suppl. Table 5).

This gene expression data suggests that some rare flexible genes, including genes found in genomic islands and genes confined to particular ecotypes, play a role in *Prochlorococcus* physiology. It is important to keep in mind, however, that despite closest sequence similarity to *Prochlorococcus* isolates, these genes may actually belong to other organisms in these communities. Sequencing large-insert clones (fosmids and BACs) or flow-sorted cell populations will allow us to match such flexible genes with informative phylogenetic markers.

Expression of hypothetical proteins: do they matter and where?

Making sense of the expression of genes of unknown function is a challenging task. Of all the "conserved hypothetical", "hypothetical", and unannotated genes detected in the combined DNA from all three depths (either core or flexible), 55% were also detected in the cDNA from at least one depth (282 out of 513 gene clusters). This fraction is only slightly lower than the 61% observed for functionally annotated genes (1152 out of 1896 gene clusters). In contrast, hypothetical and conserved hypothetical proteins were strongly underrepresented in the proteome relative to the meta-genome in an acid mine drainage community, and the authors inferred that many hypothetical proteins might be nonfunctional, required in low abundance, or expressed only in certain conditions (Ram et al., 2005). This might also be the case in our samples if, for instance, many transcribed genes do not get translated into proteins, which has been suggested for horizontally acquired genes in *E. coli* (Taoka et al., 2004). In addition, we are only capturing genes that have been seen in other *Prochlorococcus* genomes, and this recurrence supports the functional significance of these particular genes. Genes that have been recently acquired by a *Prochlorococcus* cell in these waters are invisible to our analyses, and these might be enriched in hypotheticals that are nonfunctional and not expressed. Alternatively, it may be the case that a greater fraction of hypothetical genes are expressed in these open ocean samples than in the acid mine drainage biofilm, for instance if we have captured a greater diversity of microhabitats or cell states. Moreover, the categories "hypothetical" and "conserved hypothetical" are likely not congruent in the two systems, given the discrepancy in available genome annotations (1 for *Leptospirillum* (Ram et al., 2005) vs 12 for *Prochlorococcus*), and the variable fraction of genes falling into these two categories (20% in HL

Prochlorococcus genomes vs 40% in *Leptospirillum*). The difference does not seem to be attributable simply to deeper sequencing in one system compared to the other, as the proteomic study achieved similar coverage of predicted proteins (48% of the dominant *Leptospirillum* Group II predicted proteins were detected) as we did with *Prochlorococcus* transcripts.

CONCLUSIONS

One potentially powerful use of metagenomics in microbial ecology is testing hypotheses generated in studies of cultured isolates. The reverse is also true: cultured isolates can be used to test specific hypotheses generated by metagenomics. It is unlikely that all variant strains of each organism of interest - or even representatives of every relevant ecotype - can routinely be isolated and physiologically characterized. Model systems, such as *Prochlorococcus*, are therefore especially important for integrating findings from various approaches. Investigations that leverage knowledge of model organisms can reveal structures in metagenomic data that might enable greater interpretation of sequences from uncultivated taxa. The greatest benefit will likely be realized by a combined approach that sees laboratory studies and meta-analysis of natural communities as synergistic, and builds on the respective strengths of each technique. Studies such as this one provide a framework for the interpretation of metagenomic data to understand microbial populations - collections of individual cells that occupy a biogeochemically similar habitat, are subject to related environmental stresses, and exchange genetic information.

This study has confirmed several hypotheses deriving from analysis of the collection of *Prochlorococcus* isolate genomes. Nearly the entire single-copy core genome of 1221 genes proposed by Kettler et al. (Kettler et al., 2007, Appendix C) was found to be present at close to one copy per cell. Most of these genes (88%) were found to be expressed, and the absence of some is likely a result only of sequencing coverage. This reinforces the idea that the core genome encodes a set of metabolic processes central to the functioning of *Prochlorococcus* cells and is broadly conserved in natural populations. Whether such sets of core genes are a common feature of microbial populations remains to be seen, but this study suggests one method of identifying them: core genes will appear in a pyrosequencing dataset with a frequency that is linearly proportional to their length.

Flexible genes, as expected, were more variably distributed than core genes. About a quarter of the flexible genes known from isolate genomes were observed in our samples, yet there are almost certainly more *Prochlorococcus* flexible genes in the genomic DNA samples that were invisible to our analysis. Nevertheless, some principal features of the flexible genome as defined by isolates were observed in the population metagenome. Notably, the observed abundance of flexible genes tracked their prevalence in HL isolate genomes, affirming that the genomes of the cultured strains are broadly representative those found in of natural populations. And some genes showed clear abundance changes with depth that suggest they are being differentially spread through or maintained by distinct parts of the population. The extent to which spatial variation in gene abundance is driven by selective adaptation to

immediate, local conditions versus being controlled by barriers to host dispersal and gene transfer remains a key unresolved question in microbial population genomics.

The metatranscriptomic analysis demonstrated that at least some of this genomic diversity is functionally significant. Genes in hypervariable genomic islands and rare flexible genes were found to be expressed, sometimes at high levels. And gene expression has now been measured in *Prochlorococcus* using a suite of tools, including qRTPCR, microarrays and pyrosequencing, in both laboratory and field settings, and remarkably congruent patterns have emerged. Genes in predicted operons showed expression at consistent stoichiometries, demonstrating the utility of environmental sequencing for quantitative functional analysis. Whole-cell patterns of genome-wide expression over a diel cell cycle were found to be very similar between metatranscriptomic sequencing of a natural population and microarray analysis of controlled growth in the laboratory. When there is a clear, strong environmental driver of gene expression – as the light/dark cycle is for *Prochlorococcus* – independent measures of expression yield convergent results.

Several previously unknown features of the ecology and evolution of *Prochlorococcus* are also suggested by this data. We found the first genomic evidence, in the form of depth-related multiplicity trends, that cells belonging to HL ecotypes but dwelling nearer the base of the euphotic zone have genetic complements distinct from their relatives towards the surface. Whether these cells constitute entirely new HL ecotypes or bear specific adaptations to life at greater depth remains to be seen. We also identified, through anomalous sequence coverage patterns, potential genes and regions involved in recombination, which has not been extensively documented in *Prochlorococcus*. Finally, the greater abundance of a special set of thiamin-related genes in the 25 and 75m samples implicates *Prochlorococcus* as an active player in a water-column cycle of that vitamin, about which very little is presently known. Thiamin metabolism appears to have been expanded in some cells through the horizontal acquisition of accessory salvage pathways, a pattern that may be replicated in other areas of metabolism in response to different environmental stresses.

The picture of the population metagenome emerging from this study places *Prochlorococcus* at an intermediate position along a spectrum of genomic variability. The diversity and dynamism of the flexible genome is unlike populations thought to be nearly clonal, such as *Crocosphaera* (Zehr et al., 2007). The enormous population size and wide geographic range of *Prochlorococcus* are likely factors promoting a large pan-genome, and gene transfer mediated by phages or other modes of exchage can communicate genetic variation through the population. The presence of an identifiable core genome, however, means that there are genetic commonalities, perhaps even universals, among *Prochlorococcus* cells. Certain loci are likely much more resistant to transfer and loss, and these constitute a steadily functioning scaffold around which flexible genomic content is arranged and incorporated into cellular metabolism to varying degrees. By employing metagenomic and metatranscriptomic studies similar to this one in a wide range of environments, we can further elucidate the adaptive significance of this flexible genome for *Prochlorococcus* evolution.

ACKNOWLEDGEMENTS & CONTRIBUTIONS

Samples were collected on HOT179 by Ed DeLong, Tracy Mincer, Matt Sullivan, Anne Thompson, Maureen Coleman, and Jake Waldbauer. We thank the captain and crew of the R/V Kilo Moana and the entire HOT program for enabling our sample collection. Genomic DNA was extracted by Tracy Mincer and Jay McCarren. Yanmei Shi extracted and amplified the RNA, using a protocol developed by Jorge Frias-Lopez and Yanmei Shi (Frias-Lopez et al., 2008). Sequencing was done by Stephan Schuster at Penn State University, through funds obtained by Ed DeLong and Stephan Schuster. Gene Tyson and Yanmei Shi processed the raw sequences including removal of low-quality and rRNA reads from the cDNA. We thank Allison Coe for performing the qPCR quantitation shown in Figure 2. We thank Jake Waldbauer and Vanja Klepac-Ceraj for discussion and comments on the manuscript.

REFERENCES

Ahlgren, NA, G Rocap, and SW Chisholm. 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.

Bouman, HA, O Ulloa, DJ Scanlan, K Zwirglmaier, WK Li, T Platt, V Stuart, R Barlow, O Leth, L Clementson, V Lutz, M Fukasawa, S Watanabe, and S Sathyendranath. 2006. Oceanographic basis of the global surface distribution of *Prochlorococcus* ecotypes. *Science* 312: 918-921.

Bruyant, F, M Babin, B Genty, and O Prasil. 2005. Diel variations in the photosynthetic parameters of *Prochlorococcus* strain PCC 9511: Combined effects of light and cell cycle. *Limnol. Oceanogr.* 50: 850-863.

Carlucci, AF, SB Silbernagel and PM McNally. 1969. Influence of temperature and solar radiation on persistence of vitamin B12, thiamine, and biotin in seawater. *J. Phycology* 5: 302-305.

Coleman, ML and SW Chisholm. 2007. Code and context: : *Prochlorococcus* as a model for cross-scale biology. *Trends in Microbiology* 15: 398-407.

Coleman, ML, MB Sullivan, AC Martiny, C Steglich, K Barry, EF DeLong, and SW Chisholm. 2006. Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* 311: 1768-1770.

Croft, MT, MJ Warren, and AG Smith. 2006. Algae need their vitamins. *Eukaryotic Cell* 5: 1175-1183.

DeLong, EF, CM Preston, T Mincer, V Rich, SJ Hallam, NU Frigaard, A Martinez, MB Sullivan, R Edwards, BR Brito, SW Chisholm, and DM Karl. 2006. Community genomics among stratified microbial assemblages in the ocean's interior. *Science* 311: 496-503.

Deng, S, RA Stein, and NP Higgins. 2004. Transcription-induced barriers to supercoil diffusion in the *Salmonella typhimurium* chromosome. *PNAS* 101: 3398-3403.

Edwards, RA, B Rodriguez-Brito, L Wegley, M Haynes, M Breitbart, DM Peterson, MO Saar, S Alexander, EC Alexander, and F Rohwer. 2006. Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* 7: 57.

Frias-Lopez, J, Y Shi, G Tyson, ML Coleman, SC Schuster, SW Chisholm, EF DeLong. 2007. Microbial community gene expression in ocean surface waters. *PNAS* 105: 3805-3810.

Garczarek, L, A Dufresne, S Rousvoal, NJ West, S Mazard, D Marie, H Claustre, P Raimbault, AF Post, DJ Scanlan, and F Partensky. 2007. High vertical and low horizontal diversity of *Prochlorococcus* ecotypes in the Mediterranean Sea in summer. *FEMS Microbiol. Ecol.* 60: 189-206.

He, HZ, HB Li, and F Chen. 2005. Determination of vitamin B1 in seawater and microalgal fermentation media by high-performance liquid chromatography with fluorescence detection. *Analytical and Bioanalytical Chemistry* 383: 875-879.

Huson, DH, AF Auch, J Qi, and SC Schuster. 2007. MEGAN analysis of metagenomic data. *Genome Res.* 17: 377-386.

Jenkins, AH, G Schyns, S Potot, G Sun, and TP Begley. 2007. A new thiamin salvage pathway. *Nature Chemical Biology* 3: 492-497.

Johnson, ZI, ER Zinser, A Coe, and NP McNulty. 2006. Niche Partitioning Among *Prochlorococcus* Ecotypes Along Ocean-Scale Environmental Gradients. *Science* 311: 1737-1740.

Karl, DM. 2002. Nutrient dynamics in the deep blue sea. *Trends in Microbiology* 10: 391-434.
Kettler, G, AC Martiny, K Huang, J Zucker, ML Coleman, S Rodrigue, F Chen, A Lapidus, S Ferriera, J Johnson, C Steglich, GM Church, PM Richardson, and SW Chisholm. 2007. Patterns and Implications of Gene Gain and Loss in the Evolution of *Prochlorococcus*. *PLoS Genetics* 3: e231.
Konstantinidis, KT, and JM Tiedje. 2005. Genomic insights that advance the species definition for prokaryotes. *PNAS* 102: 2567-2572.
Lawrence, JG. 2002. Gene transfer in bacteria: speciation without species? *Theor. Pop. Biol.* 61: 449-460.
Lawrence, JG, and H Hendrickson. 2005. Genome evolution in bacteria: order beneath chaos. *Current Opinion in Microbiology* 8: 572-578.
Lindell, D, JD Jaffe, ML Coleman, ME Futschik, IM Axmann, T Rector, G Kettler, MB Sullivan, R Steen, WR Hess, GM Church, and SW Chisholm. 2007. Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature* 449: 83-86.
Mann, EL, NA Ahlgren, JW Moffett, and SW Chisholm. 2002. Copper Toxicity and Cyanobacteria Ecology in the Sargasso Sea. *Limnol. Oceanogr.* 47: 976-988.
Margulies, M, M Egholm, WE Altman, S Attiya, JS Bader, LA Bemben, J Berka, MS Braverman, YJ Chen, Z Chen, SB Dewell, L Du, JM Fierro, XV Gomes, BC Godwin, W He, S Helgesen, CH Ho, CH Ho, GP Irzyk, SC Jando, ML Alenquer, TP Jarvie, KB Jirage, JB Kim, JR Knight, JR Lanza, JH Leamon, SM Lefkowitz, M Lei, J Li, KL Lohman, H Lu, VB Makhijani, KE McDade, MP McKenna, EW Myers, E Nickerson, JR Nobile, R Plant, BP Puc, MT Ronan, GT Roth, GJ Sarkis, JF Simons, JW Simpson, M Srinivasan, KR Tartaro, A Tomasz, KA Vogt, GA Volkmer, SH Wang, Y Wang, MP Weiner, P Yu, RF Begley, and JM Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376-380.
Martin, DP, C Williamson, and D Posada. 2005. RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics* 21: 260-262.
Martiny, AC, ML Coleman, and SW Chisholm. 2006. Phosphate acquisition genes in *Prochlorococcus* ecotypes: evidence for genome-wide adaptation. *PNAS* 103: 12552-12557.
Moore, LR, and SW Chisholm. 1999. Photophysiology of the Marine Cyanobacterium *Prochlorococcus*: Ecotypic Differences among Cultured Isolates. *Limnol. Oceanogr.* 44: 628-638
Moore, LR, R Goericke, and SW Chisholm. 1995. Comparative physiology of *Synechococcus* and *Prochlorococcus* - Influence of light and temperature on growth, pigments, fluorescence and absorptive properties. *Marine Ecology Progress Series* 116: 259-275.
Moore, LR, M Ostrowski, DJ Scanlan, and K Feren. 2005. Ecotypic variation in phosphorus-acquisition mechanisms within marine picocyanobacteria. *Aquatic Microb. Ecol.* 39: 257-269.
Moore, LR, AF Post, G Rocap, and SW Chisholm. 2002. Utilization of Different Nitrogen Sources by the Marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. *Limnol. Oceanogr.* 47: 989-996.
Moore, LR, G Rocap, and SW Chisholm. 1998. Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* 393: 464-467.
Moreno-Paz, M, and V Parro. 2006. Amplification of low quantity bacterial RNA for microarray studies: time-course analysis of *Leptospirillum ferrooxidans* under nitrogen-fixing conditions. *Environ. Microbiol.* 8: 1064-1073.
Moutin, T, DM Karl, S Duhamel, P Rimmelin, P Raimbault, BAS Van Mooy, H Claustre. 2008. Phosphate availability and the ultimate control of new nitrogen input by nitrogen fixation in the tropical Pacific Ocean. *Biogeosciences* 5: 95–109.
Okbamichael, M, and SA Sañudo-Wilhelmy. 2005. Direct determination of vitamin B12 in seawater by solid-phase extraction and high-performance liquid chromatography quantification. *Limnol. Oceanogr.: Methods* 3: 241-246.
Pál C, B Papp, and MJ Lercher. 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat. Genet.* 37: 1372-1375.
Palenik, B, B Brahamsha, FW Larimer, M Land, L Hauser, P Chain, J Lamerdin, W Regala, EE Allen, J McCarren, I Paulsen, A Dufresne, F Partensky, EA Webb, and J Waterbury. 2003. The genome of a motile marine *Synechococcus*. *Nature* 424: 1037-1042.
Park, JH, K Burns, C Kinsland, and TP Begley. 2004. Characterization of two kinases involved in thiamine pyrophosphate and pyridoxal phosphate biosynthesis in *Bacillus subtilis*: 4-amino-5-hydroxymethyl-2methylpyrimidine kinase and pyridoxal kinase. *J. Bacteriol.* 186: 1571-1573.
Poretsky, RS, N Bano, A Buchan, G LeCleir, J Kleikemper, M Pickering, WM Pate, MA Moran, and JT Hollibaugh. 2005. Analysis of microbial gene transcripts in environmental samples. *Appl. Environ. Microbiol.* 71: 4121-4126.
Rachman, H, JS Lee, J Angermann, J Kowall, and SH Kaufmann. 2006. Reliable amplification method

for bacterial RNA. *J. Biotechnol.* 126: 61-68.
Ram, RJ, NC VerBerkmoes, MP Thelen, GW Tyson, BJ Baker, RC Blake, M Shah, RL Hettich, and JF Banfield. 2005. Community proteomics of a natural microbial biofilm. *Science* 308: 1915-1920.
Rocap, G, DL Distel, JB Waterbury, and SW Chisholm. 2002. Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. *Appl. Environ. Microbiol.* 68: 1180-1191.
Rusch, DB, AL Halpern, G Sutton, KB Heidelberg, S Williamson, S Yooseph, D Wu, JA Eisen, JM Hoffman, K Remington, K Beeson, B Tran, H Smith, H Baden-Tillson, C Stewart, J Thorpe, J Freeman, C Andrews-Pfannkoch, JE Venter, K Li, S Kravitz, JF Heidelberg, T Utterback, YH Rogers, LI Falcón, V Souza, G Bonilla-Rosso, LE Eguiarte, DM Karl, S Sathyendranath, T Platt, E Bermingham, V Gallardo, G Tamayo-Castillo, MR Ferrari, RL Strausberg, K Nealson, R Friedman, M Frazier, and JC Venter. 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biology* 5: e77.
Saito, MA, JW Moffett, SW Chisholm, and JB Waterbury. 2002. Cobalt limitation and uptake in *Prochlorococcus*. *Limnol. Oceanogr.* 47: 1629-1636.
Shapiro, BJ, and EJ Alm. 2008. Comparing Patterns of Natural Selection across Species Using Selective Signatures. *PLoS Genetics* 4: e23.
Sherratt, DJ. 2003. Bacterial chromosome dynamics. *Science* 301: 780-785.
Sorek, R, Y Zhu, CJ Creevey, MP Francino, P Bork, and EM Rubin. 2007. Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* 318: 1449-1452.
Steglich, C, M Futschik, T Rector, R Steen, and SW Chisholm. 2006. Genome-wide analysis of light sensing in Prochlorococcus. *J. Bacteriol.* 188: 7796-7806.
Sullivan, MB, ML Coleman, P Weigele, F Rohwer, and SW Chisholm. 2005. Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. *PLoS Biology* 3: e144.
Sullivan, MB, D Lindell, JA Lee, LR Thompson, JP Bielawski, and SW Chisholm. 2006. Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biology* 4: e234.
Taoka, M, Y Yamauchi, T Shinkawa, H Kaji, W Motohashi, H Nakayama, N Takahashi, and T Isobe. 2004. Only a small subset of the horizontally transferred chromosomal genes in *Escherichia coli* are translated into proteins. *Mol. Cell. Proteomics* 3: 780-787.
Tolonen, AC, J Aach, D Lindell, ZI Johnson, T Rector, R Steen, GM Church, and SW Chisholm. 2006. Global gene expression of *Prochlorococcus* ecotypes in response to changes in nitrogen availability. *Mol. Systems Biol.* 2: 53.
Van Gelder, RN, ME von Zastrow, A Yool, WC Dement, JD Barchas, and JH Eberwine. 1990. Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *PNAS* 87: 1663-1667.
Vaulot, D, D Marie, RJ Olson, and SW Chisholm. 1995. Growth of *Prochlorococcus*, a Photosynthetic Prokaryote, in the Equatorial Pacific Ocean. *Science* 268: 1480-1482.
Venter, JC, K Remington, JF Heidelberg, AL Halpern, D Rusch, JA Eisen, DY Wu, I Paulsen, KE Nelson, W Nelson, DE Fouts, S Levy, AH Knap, MW Lomas, K Nealson, O White, J Peterson, J Hoffman, R Parsons, H Baden-Tillson, C Pfannkoch, YH Rogers, and HO Smith. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304: 66-74.
Welch, RA, V Burland, G Plunkett, P Redford, P Roesch, D Rasko, EL Buckles, SR Liou, A Boutin, J Hackett, D Stroud, GF Mayhew, DJ Rose, S Zhou, DC Schwartz, NT Perna, HLT Mobley, MS Donnenberg, and FR Blattner. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *PNAS* 99: 17020-17024.
Wendisch, VF, DP Zimmer, A Khodursky, B Peter, N Cozzarelli, and S Kustu. 2001. Isolation of *Escherichia coli* mRNA and comparison of expression using mRNA and total RNA on DNA microarrays. *Analytical Biochemistry* 290: 205-213.
Winfield, MD, and EA Groisman. 2004. Phenotypic differences between *Salmonella* and *Escherichia coli* resulting from the disparate regulation of homologous genes. *PNAS* 101: 17162-17167.
Wommack, KE, J Bhavsar, and J Ravel. 2008. Metagenomics: Read length matters. *Appl. Environ. Microbiol.* 74: 1453-1463.
Zehr, JP, SR Bench, EA Mondragon, J McCarren, and EF DeLong. 2007. Low genomic diversity in tropical oceanic N_2-fixing cyanobacteria. *PNAS* 104: 17807-17812.
Zinser, ER, A Coe, ZI Johnson, AC Martiny, NJ Fuller, DJ Scanlan and SW Chisholm. 2006. *Prochlorococcus* Ecotype Abundances in the North Atlantic Ocean As Revealed by an Improved Quantitative PCR Method. *Appl. Environ. Microbiol.* 72: 723-732.
Zinser, ER, ZI Johnson, A Coe, E Karaca, D Veneziano and SW Chisholm. 2007. Influence of light and

temperature on *Prochlorococcus* ecotype distribution in the Atlantic Ocean. *Limnol. Oceanogr.* 52: 2205-2220.

Zubkov, M, B Fuchs, G Tarran, P Burkhill, and R Amann. 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, JL Heywood, K Chamberlain, EMS Woodward, MV Zubkov and DJ Scanlan. 2007. Basin-scale distribution patterns of picocyanobacterial lineages in the Atlantic Ocean. *Environ. Microbiol.* 9: 1278-1290.

Suppl. Table 1. Relative abundance of flexible genes detected at each depth. Genes detected less frequently than core genes ($p<0.01$) are labeled 'rare'; genes detected more frequently than core genes are labeled 'high'; and the rest are labeled 'average'. ND indicates that a gene was not detected at that depth. Roughly half the flexible genes were rare or undetected at most or all depths, while the other half were of average abundance (close to that of single-copy core genes) at most depths. Only a handful of genes (also indicated by letters a-f in Figure 1B) were consistently present at high abundance.

			number of DNA fragments observed			relative abundance				
cluster	core	protLength	25DNA	75DNA	125DNA	25m	75m	125m	locus	description
1954	0	59	18	32	5	average	high	average	P9301_13211	Conserved hypothetical protein
493	0	331	79	142	42	average	high	high	P9301_12391	ABC transporter, substrate binding protein, phosphate
523	0	249	96	135	31	high	high	average	P9301_15481	NAD-dependent DNA ligase N-terminus
13	0	89	73	125	13	high	high	average	P9301_13071	possible Phosphatidylinositol-specific phospho
1012	0	183	72	98	19	high	high	average	P9301_15461	Predicted membrane protein
145	0	135	53	84	18	high	high	average	P9301_11261	Uncharacterized protein conserved in bacteria
423	0	131	50	63	17	high	high	average	P9301_08441	conserved hypothetical protein
142	0	82	28	49	10	high	high	average	P9301_11291	conserved hypothetical
102	0	403	143	256	69	high	high	high	P9301_11701	putative lycopene beta cyclase
591	0	352	126	213	58	high	high	high	P9301_13721	*chlorophyll a/b binding light harvesting protein PcbD*
1672	0	388	126	206	61	high	high	high	P9301_15761	fatty acid desaturase, type 2
192	0	103	53	98	37	high	high	high	P9301_12751	
266	0	120	47	101	30	high	high	high	P9301_10041	hypothetical
78	0	363	58	113	7	average	average	rare	P9301_11941	putative glycerol dehydrogenase
384	0	384	39	107	28	rare	average	average	P9301_08851	putative urea ABC transporter
901	0	259	43	66	6	average	average	rare	P9301_03941	Phosphomethylpyrimidine kinase
2149	0	205	27	52	3	average	average	rare	P9301_15781	conserved hypothetical protein
2068	0	150	6	20	4	rare	average	average	P9301_12821	conserved hypothetical protein
906	0	105	4	9	0	rare	average	ND	A9601_12391	
630	0	1296	199	326	86	average	average	average	P9301_06151	conserved hypothetical
1527	0	1114	162	262	75	average	average	average	P9301_16851	Translation initiation factor IF-2
182	0	868	150	226	59	average	average	average	P9301_10901	probable aminopeptidase N
485	0	828	170	207	43	average	average	average	P9301_07821	DEAD/DEAH box helicase:Helicase C-terminal domain
376	0	569	98	186	53	average	average	average	P9301_08931	Urease alpha subunit
795	0	589	115	150	44	average	average	average	P9301_04511	putative penicillin binding protein
374	0	586	112	151	45	average	average	average	P9301_08951	Glycoside hydrolase family 13
787	0	506	97	152	41	average	average	average	P9301_04591	putative NADH dehydrogenase (complex I) subunit (chain 2)
931	0	509	86	164	34	average	average	average	P9301_03701	Bacterial-type phytoene dehydrogenase
1211	0	429	101	150	30	average	average	average	P9301_00891	putative ATPase, AAA family
484	0	546	96	138	35	average	average	average	P9301_07831	possible ATP-dependent DNA ligase
1337	0	437	96	139	33	average	average	average	P9301_18701	ATP-dependent DNA ligase
98	0	490	80	149	34	average	average	average	P9301_11741	Predicted membrane-associated HD superfamily hydrolase
44	0	432	76	138	48	average	average	average	P9301_13641	porin-like
136	0	449	85	133	40	average	average	average	P9301_11351	Cobalamin synthesis protein/P47K
1935	0	458	95	135	27	average	average	average	P9301_13401	putative thioredoxin reductase
1307	0	469	92	134	30	average	average	average	P9301_19011	Sucrose phosphate synthase
991	0	478	87	120	45	average	average	average	P9301_03091	putative DNA photolyase
797	0	496	93	127	31	average	average	average	P9301_04471	Uncharacterized protein related to deoxyribodipyrimidine photolyase
1183	0	454	69	155	26	average	average	average	P9301_01171	possible Fe-S oxidoreductase
905	0	498	101	115	31	average	average	average	P9301_03921	putative deoxyribodipyrimidine photolyase
375	0	408	88	122	36	average	average	average	P9301_08941	possible Vng0271c
1188	0	475	90	118	35	average	average	average	P9301_01121	possible RND family outer membrane efflux protein
129	0	480	67	130	40	average	average	average	P9301_11431	Uncharacterized conserved protein
1291	0	447	80	115	34	average	average	average	P9301_00101	Protein phosphatase 2C domain
383	0	425	75	114	34	average	average	average	P9301_08861	putative urea ABC transporter, substrate binding protein
966	0	409	64	135	23	average	average	average	P9301_03351	Carbohydrate kinase, FGGY family
1037	0	450	83	114	24	average	average	average	P9301_02641	Phosphotransferase superclass
1689	0	414	71	118	31	average	average	average	P9301_15581	Predicted flavoproteins
137	0	513	65	129	25	average	average	average	P9301_11341	ABC transporter, substrate binding protein, possibly Mn.
1898	0	358	70	122	25	average	average	average	P9301_13791	Glycosyl transferases group 1
135	0	357	54	130	25	average	average	average	P9301_11361	G-protein beta WD-40 repeats
480	0	439	67	107	27	average	average	average	P9301_07871	Uncharacterized protein conserved in bacteria
1239	0	320	70	108	20	average	average	average	P9301_00601	Aldo/keto reductase family
2034	0	357	74	98	26	average	average	average	P9301_08421	Fructose-bisphosphate/sedoheptulose-1, 7-bisphosphate aldolase
200	0	388	65	104	27	average	average	average	P9301_10711	conserved hypothetical protein
781	0	380	74	95	26	average	average	average	P9301_04651	Aldo/keto reductase family
120	0	409	59	103	31	average	average	average	P9301_11521	possible multidrug efflux transporter, MFS family
1700	0	382	58	111	24	average	average	average	P9301_15471	similar to DNA photolyase
385	0	376	76	80	34	average	average	average	P9301_08841	putative membrane protein of urea ABC transport system
1961	0	447	56	105	22	average	average	average	P9301_13021	Na+-dependent transporter of the SNF family
128	0	313	55	94	33	average	average	average	P9301_11441	Uncharacterized protein conserved in bacteria
276	0	394	63	96	23	average	average	average	P9301_09921	probable ribonuclease II
425	0	355	74	78	29	average	average	average	A9601_08451	Fructose-1,6-bisphosphate aldolase class I
1165	0	400	62	101	18	average	average	average	P9301_01351	Uncharacterized protein conserved in bacteria
217	0	331	49	97	31	average	average	average	P9301_10511	putative multidrug efflux ABC transporter
551	0	302	56	92	28	average	average	average	P9301_07181	Hsp33 protein
729	0	362	60	90	22	average	average	average	P9301_05171	Putative GTPases (G3E family)
908	0	316	69	85	17	average	average	average	P9301_03901	proline iminopeptidase

882	0	237	53	92	21	average	average	average	P9301_00401	hypothetical
1345	0	312	48	82	36	average	average	average	P9301_18621	Fatty acid desaturase, type 1
127	0	285	55	91	19	average	average	average	P9301_11451	Transglutaminase-like superfamily
809	0	264	56	81	20	average	average	average	P9301_04351	putative methyltransferase
1315	0	284	43	89	24	average	average	average	P9301_18931	formyltetrahydrofolate deformylase
379	0	300	48	85	21	average	average	average	P9301_08901	urease accessory protein UreD
716	0	263	45	89	19	average	average	average	P9301_05301	possible precorrin-6X reductase
483	0	328	45	83	24	average	average	average	P9301_07841	Predicted exonuclease of the beta-lactamase fold involved in RNA processing
1175	0	235	63	65	22	average	average	average	P9301_01251	possible POLO box duplicated region.
386	0	249	43	85	19	average	average	average	P9301_08831	putative ATP binding subunit of urea ABC transport system
174	0	335	47	78	20	average	average	average	P9301_10981	putative nitrogen regulation protein NifR3 family homolog
583	0	343	63	61	21	average	average	average	P9301_06621	small mechanosensitive ion channel, MscS family
215	0	264	55	67	22	average	average	average	P9301_10541	conserved membrane protein, multidrug efflux associated
1690	0	249	49	77	15	average	average	average	P9301_15571	GAF domain
40	0	363	46	79	13	average	average	average	P9301_16881	Glycosyl transferase, family 2
1507	0	207	48	76	14	average	average	average	P9301_17071	Site-specific recombinase
450	0	227	48	69	20	average	average	average	P9301_08171	D-ala-D-ala dipeptidase
373	0	265	43	70	19	average	average	average	P9301_08961	Predicted hydrolase (HAD superfamily)
1715	0	224	44	69	15	average	average	average	P9301_15321	conserved hypothetical protein
1240	0	223	50	57	20	average	average	average	P9301_00591	ATPases involved in chromosome partitioning
153	0	256	47	57	19	average	average	average	P9301_11181	possible Gram-negative pili assembly chaperone
199	0	267	50	54	19	average	average	average	P9301_10721	possible Paired amphipathic helix repeat
2226	0	340	39	62	21	average	average	average	P9301_12361	Putative glyceraldehyde 3-phosphate dehydrogenase
1652	0	288	45	57	17	average	average	average	A9601_16211	Predicted dehydrogenase
1083	0	239	45	61	12	average	average	average	P9301_02201	possible ribonuclease HI
387	0	236	47	49	21	average	average	average	P9301_08821	Putative ATP-binding subunit of urea ABC transport system
1098	0	241	38	63	16	average	average	average	P9301_02051	Predicted Zn-dependent hydrolases of the beta-lactamase fold
1694	0	243	41	56	20	average	average	average	P9301_15531	DUF209
811	0	234	42	57	16	average	average	average	P9301_04331	putative short chain dehydrogenase
902	0	207	38	63	13	average	average	average	P9301_03931	TENA/THI-4 protein
692	0	189	46	51	16	average	average	average	P9301_05541	
1704	0	205	43	57	13	average	average	average	P9301_15431	Steroid 5-alpha reductase, C-terminal domain
1609	0	245	38	61	12	average	average	average	P9301_13091	
937	0	169	31	60	16	average	average	average	P9301_03641	conserved hypothetical
1703	0	183	37	55	14	average	average	average	P9301_15441	Uncharacterized conserved protein
213	0	270	31	59	15	average	average	average	P9301_10571	Predicted permeases
216	0	264	33	51	20	average	average	average	P9301_10531	putative multidrug efflux ABC transporter
486	0	216	41	49	13	average	average	average	P9301_07811	Serine/threonine specific protein phosphatase
382	0	203	38	45	17	average	average	average	P9301_08871	urease accessory protein UreG
126	0	157	26	56	16	average	average	average	P9301_11461	conserved hypothetical protein
834	0	161	29	56	12	average	average	average	P9301_04101	conserved hypothetical protein
196	0	226	35	47	14	average	average	average	P9301_10751	Peptidase E
1054	0	208	38	48	10	average	average	average	P9301_02471	possible 2-keto-3-deoxy-6-phosphogluconate aldolase
204	0	152	30	51	13	average	average	average	P9301_10671	conserved hypothetical
381	0	228	32	52	8	average	average	average	P9301_08881	urease accessory protein UreF
587	0	180	28	48	15	average	average	average	P9301_06581	possible NADH-Ubiquinone/plastoquinone (comple
89	0	178	25	57	8	average	average	average	P9301_11831	conserved hypothetical protein
2040	0	188	22	50	17	average	average	average	P9301_09781	Predicted membrane protein
435	0	182	22	51	14	average	average	average	P9301_08321	conserved hypothetical protein
929	0	161	19	52	16	average	average	average	P9301_03721	conserved hypothetical protein
1590	0	178	29	47	11	average	average	average	P9301_16231	Phosphoribosyl transferase
417	0	164	27	46	13	average	average	average	P9301_08501	Predicted hydrolase of the HAD superfamily
1375	0	143	27	52	6	average	average	average	P9301_18321	possible transcription regulator
87	0	179	32	40	12	average	average	average	P9301_11851	conserved hypothetical
889	0	211	36	34	14	average	average	average	P9301_03331	Uncharacterized conserved protein
921	0	149	18	49	16	average	average	average	P9301_03801	putative bacterioferritin comigratory protein
566	0	129	29	46	7	average	average	average	P9301_06791	Glyoxalase I
1654	0	135	28	36	18	average	average	average	P9301_15991	
256	0	250	34	40	7	average	average	average	P9301_10131	conserved hypothetical
274	0	207	28	39	14	average	average	average	P9301_09941	possible SMC domain N terminal domain
1422	0	143	27	39	14	average	average	average	P9301_17871	conserved hypothetical protein
147	0	135	30	42	7	average	average	average	P9301_11241	putative stress-induced protein OsmC
563	0	194	25	42	12	average	average	average	P9301_06821	Conserved hypothetical protein
1916	0	222	24	42	13	average	average	average	P9301_13621	Prolyl 4-hydroxylase, alpha subunit
2041	0	192	25	44	10	average	average	average	P9301_10231	conserved hypothetical protein
897	0	125	24	49	5	average	average	average	P9301_03981	possible Phosphoenolpyruvate carboxykinase
567	0	165	19	48	10	average	average	average	P9301_06781	possible VHS domain
51	0	168	24	43	9	average	average	average	P9301_12211	conserved hypothetical protein
349	0	155	23	39	13	average	average	average	P9301_09201	possible Serine hydroxymethyltransferase
1611	0	99	27	44	3	average	average	average	P9301_13001	protein family PM-11
218	0	129	20	47	6	average	average	average	P9301_10501	conserved hypothetical protein
393	0	98	24	42	5	average	average	average	P9301_08751	hypothetical
926	0	130	16	39	16	average	average	average	P9301_03751	Conserved hypothetical protein
207	0	167	17	43	9	average	average	average	P9301_10641	possible Josephin
2023	0	102	24	35	10	average	average	average	P9301_05011	conserved hypothetical protein
220	0	164	16	40	12	average	average	average	P9301_10481	putative protein
2050	0	180	21	36	11	average	average	average	P9301_11091	conserved hypothetical
150	0	124	31	28	8	average	average	average	P9301_11211	Conserved hypothetical protein
201	0	198	27	36	4	average	average	average	P9301_10701	ATP/GTP-binding site motif A (P-loop)
20	0	170	22	38	6	average	average	average	P9301_00321	conserved hypothetical protein
597	0	136	26	35	5	average	average	average	P9301_06481	conserved hypothetical protein
2013	0	155	21	39	6	average	average	average	P9301_03971	Uncharacterized conserved protein
771	0	185	22	35	8	average	average	average	P9301_04751	possible Arenavirus glycoprotein

960	0	120	23	39	3	average	average	average	P9301_03411	possible ferredoxin
1415	0	151	19	36	10	average	average	average	P9301_17941	Conserved hypothetical protein
1065	0	128	13	44	7	average	average	average	P9301_02371	Glyoxalase/Bleomycin resistance protein/Dioxygenase superfamily
28	0	116	25	33	5	average	average	average	P9301_03831	
56	0	171	22	31	10	average	average	average	P9301_12161	conserved hypothetical protein
175	0	130	16	38	9	average	average	average	P9301_10971	hypothetical
380	0	149	19	36	8	average	average	average	P9301_08891	urease accessory protein UreE
185	0	101	24	25	12	average	average	average	P9301_10871	conserved hypothetical protein
250	0	115	19	31	11	average	average	average	P9301_10181	Predicted Fe-S-cluster oxidoreductase
472	0	124	20	32	9	average	average	average	P9301_07951	possible Alpha-2-macroglobulin family N-termin
1042	0	149	21	34	6	average	average	average	P9301_02591	conserved hypothetical protein
1157	0	119	18	34	9	average	average	average	P9301_01461	possible Signal peptide binding domain
2038	0	104	24	31	6	average	average	average	P9301_08781	hypothetical
19	0	128	18	33	9	average	average	average	P9301_18741	conserved hypothetical
763	0	164	14	32	14	average	average	average	P9301_04831	conserved hypothetical protein
1921	0	200	18	33	8	average	average	average	P9301_13571	
1963	0	124	21	32	6	average	average	average	P9301_12991	Conserved hypothetical protein
352	0	120	21	25	12	average	average	average	P9301_09171	possible cAMP phosphodiesterases class-II
1969	0	137	14	37	7	average	average	average	A9601_12531	
378	0	100	24	28	5	average	average	average	P9301_08911	Urease gamma subunit
527	0	99	23	29	5	average	average	average	P9301_07371	hypothetical
860	0	125	22	30	5	average	average	average	A9601_12481	
1697	0	88	24	29	4	average	average	average	P9301_15501	hypothetical
1911	0	110	27	28	2	average	average	average	P9301_13671	Macrophage migration inhibitory factor family
2161	0	108	25	27	5	average	average	average	P9301_16391	ferredoxin
756	0	86	19	28	9	average	average	average	P9301_04901	hypothetical
2209	0	117	20	31	5	average	average	average	P9301_07721	cytochrome c
1679	0	111	26	26	3	average	average	average	P9301_15681	Predicted membrane protein
165	0	134	16	31	7	average	average	average	P9301_11061	Predicted membrane protein
1744	0	81	20	28	6	average	average	average	P9301_15031	possible high light inducible protein
109	0	100	25	20	8	average	average	average	P9301_11631	cytochrome cM
2079	0	116	15	32	6	average	average	average	P9301_13331	Ferredoxin
2069	0	172	19	30	3	average	average	average	P9301_12831	Predicted metal-binding protein
132	0	91	10	36	5	average	average	average	P9301_11391	DNA gyrase/topoisomerase IV, subunit-like
499	0	127	16	30	5	average	average	average	P9301_07661	possible DUP family
568	0	130	21	23	7	average	average	average	P9301_06771	possible LEM domain
589	0	85	15	35	1	average	average	average	P9301_06561	conserved hypothetical protein
1549	0	85	18	24	9	average	average	average	P9301_16631	conserved hypothetical protein
1966	0	121	17	21	13	average	average	average	A9601_12561	
124	0	117	20	25	5	average	average	average	P9301_11481	possible Borrelia lipoprotein
646	0	118	12	30	8	average	average	average	P9301_05991	Transcriptional regulator AbrB
1040	0	111	19	30	1	average	average	average	P9301_02611	
1663	0	158	24	22	4	average	average	average	P9301_15871	carbon storage regulator-like
205	0	131	23	17	9	average	average	average	P9301_10661	putative PURINE PHOSPHORIBOSYLTRANSFERASE related protein
520	0	78	8	35	5	average	average	average	P9301_07441	Conserved hypothetical protein
1049	0	73	21	21	6	average	average	average	P9301_02521	conserved hypothetical protein
1215	0	117	16	27	5	average	average	average	P9301_00851	conserved hypothetical
1908	0	81	19	18	11	average	average	average	P9301_13701	hypothetical
476	0	65	13	29	5	average	average	average	P9301_07911	Conserved hypothetical protein
514	0	70	18	24	5	average	average	average	P9301_07511	conserved hypothetical protein
933	0	96	15	25	7	average	average	average	P9301_03681	conserved hypothetical
272	0	103	15	26	5	average	average	average	P9301_09961	possible Fusion glycoprotein F0.
470	0	145	12	28	6	average	average	average	P9301_07971	hypothetical membrane protein
513	0	78	11	23	12	average	average	average	P9301_07521	possible DDT domain
1018	0	83	24	16	6	average	average	average	P9301_02821	conserved hypothetical protein
1384	0	109	15	23	8	average	average	average	P9301_18231	conserved hypothetical protein
59	0	105	14	21	10	average	average	average	P9301_12131	conserved hypothetical protein
482	0	137	23	18	4	average	average	average	P9301_07851	possible COMC family
556	0	85	16	25	4	average	average	average	P9301_07131	possible RNA recognition motif. (a.k.a. RRM, R
594	0	99	20	20	5	average	average	average	P9301_06511	hypothetical
938	0	144	17	24	4	average	average	average	P9301_03631	NADH-plastoquinone oxidoreductase chain 5-like
1061	0	72	16	21	8	average	average	average	P9301_02411	Conserved hypothetical protein
1144	0	83	15	24	6	average	average	average	P9301_01591	putative Ycf34
1661	0	67	14	23	8	average	average	average	P9301_15891	regulatory proteins, DeoR-like
1678	0	108	16	24	5	average	average	average	P9301_15691	possible Type I restriction modification DNA s
1936	0	74	11	28	6	average	average	average	P9301_13391	conserved hypothetical
528	0	60	16	23	5	average	average	average	P9301_02401	conserved hypothetical
925	0	88	20	22	2	average	average	average	P9301_03761	mttA/Hcf106 family
1695	0	104	14	25	5	average	average	average	P9301_15521	possible MATH domain
15	0	86	12	24	7	average	average	average	P9301_13191	possible high light inducible protein
505	0	83	13	26	4	average	average	average	P9301_07601	possible DnaJ central domain (4 repeats)
942	0	89	13	21	9	average	average	average	P9301_03601	conserved hypothetical protein
1900	0	83	15	20	8	average	average	average	P9301_13771	possible Helix-turn-helix protein, copG family
163	0	97	16	20	5	average	average	average	P9301_11081	possible Virion host shutoff protein
194	0	106	17	23	1	average	average	average	P9301_10771	hypothetical
593	0	77	16	20	5	average	average	average	P9301_06521	hypothetical
1506	0	70	14	22	5	average	average	average	P9301_17081	hypothetical
1594	0	83	13	23	5	average	average	average	P9301_16191	conserved hypothetical protein
1688	0	78	15	19	7	average	average	average	P9301_15591	Conserved hypothetical protein
1798	0	83	15	20	6	average	average	average	P9301_14521	hypothetical
16	0	68	14	19	7	average	average	average	P9301_15821	high light inducible protein-like
368	0	90	12	24	4	average	average	average	P9301_09011	possible GRAM domain
799	0	78	11	20	9	average	average	average	P9301_04451	conserved hypothetical protein
801	0	67	15	21	4	average	average	average	P9301_04431	Conserved hypothetical protein
807	0	85	13	22	5	average	average	average	P9301_04371	hypothetical
927	0	105	10	19	11	average	average	average	P9301_03741	possible Helper component proteinase
1610	0	98	11	19	10	average	average	average	P9301_13081	possible Heat-labile enterotoxin alpha chain
475	0	64	9	26	4	average	average	average	P9301_07921	Conserved hypothetical protein

907	0	57	14	16	9 average	average	average	P9301_03911	Conserved hypothetical protein
1951	0	94	12	23	4 average	average	average	P9301_13241	possible Cytochrome b(C-terminal)/b6/petD
1981	0	66	13	22	4 average	average	average	P9301_12771	
8	0	126	13	18	7 average	average	average	P9301_15881	Conserved hypothetical protein
212	0	113	13	19	6 average	average	average	P9301_10581	HNH endonuclease:HNH nuclease
234	0	65	8	21	9 average	average	average	P9301_10351	conserved hypothetical protein
237	0	122	17	21	0 average	average	ND	P9301_10321	possible Nucleoside diphosphate kinase
397	0	64	13	22	3 average	average	average	P9301_08711	Conserved hypothetical protein
487	0	100	14	17	7 average	average	average	P9301_07801	hypothetical
269	0	95	10	21	6 average	average	average	P9301_09991	Conserved hypothetical protein
377	0	106	14	21	2 average	average	average	P9301_08921	Urease beta subunit
932	0	67	8	27	2 average	average	average	P9301_03691	Conserved hypothetical protein
1973	0	91	9	21	7 average	average	average	P9301_12901	Conserved hypothetical protein
401	0	84	19	12	5 average	average	average	P9301_08661	hypothetical
428	0	54	10	23	3 average	average	average	P9301_08391	Conserved hypothetical protein
503	0	78	15	16	5 average	average	average	P9301_07621	conserved hypothetical
517	0	84	11	20	5 average	average	average	P9301_07481	possible high light inducible protein
557	0	64	9	23	4 average	average	average	P9301_07121	Conserved hypothetical protein
1741	0	103	9	24	3 average	average	average	P9301_15061	hypothetical
131	0	77	5	29	1 average	average	average	P9301_11401	hypothetical
154	0	60	10	20	5 average	average	average	P9301_11171	conserved hypothetical protein
481	0	93	11	17	7 average	average	average	P9301_07861	possible Major surface glycoprotein
1742	0	80	13	17	5 average	average	average	P9301_15051	possible Beta-lactamase
1902	0	77	13	16	6 average	average	average	P9301_13751	hypothetical
1975	0	96	15	12	8 average	average	average	P9301_12881	Conserved hypothetical protein
2067	0	76	15	16	4 average	average	average	P9301_12811	possible Uncharacterized protein family UPF003
178	0	80	11	18	5 average	average	average	P9301_10941	possible Legume lectins alpha domain
225	0	72	3	28	3 average	average	average	P9301_10441	hypothetical
268	0	129	9	22	3 average	average	average	P9301_10031	
506	0	51	6	22	6 average	average	average	P9301_07591	protein family PM-16
507	0	75	14	16	4 average	average	average	P9301_07581	possible D12 class N6 adenine-specific DNA met
709	0	84	11	20	3 average	average	average	P9301_05371	possible Reverse transcriptase (RNA-dependent
896	0	60	18	11	5 average	average	average	P9301_03991	Conserved hypothetical protein
1581	0	52	11	19	4 average	average	average	P9301_16311	hypothetical
1952	0	83	11	21	2 average	average	average	P9301_13231	possible Uncharacterized protein family UPF005
130	0	78	13	15	5 average	average	average	P9301_11411	conserved hypothetical
214	0	75	10	20	3 average	average	average	P9301_10561	hypothetical
1164	0	89	14	18	1 average	average	average	P9301_01361	conserved hypothetical protein
1953	0	71	14	13	6 average	average	average	P9301_13221	Conserved hypothetical protein
1958	0	60	10	22	1 average	average	average	P9301_13061	Conserved hypothetical protein
1971	0	80	11	18	4 average	average	average	P9301_12931	Conserved hypothetical protein
2170	0	50	12	19	2 average	average	average	P9301_18711	conserved hypothetical protein
2180	0	117	9	22	2 average	average	average	P9301_04001	
2181	0	115	12	16	5 average	average	average	P9301_04011	
148	0	80	8	22	2 average	average	average	P9301_11231	possible lactate/malate dehydrogenase, alpha/b
156	0	73	18	8	6 average	average	average	P9301_11151	
171	0	78	12	17	3 average	average	average	P9301_11011	hypothetical
501	0	63	13	14	5 average	average	average	P9301_07641	Conserved hypothetical protein
530	0	128	13	15	4 average	average	average	A9601_16081	
808	0	63	7	18	7 average	average	average	P9301_04361	Conserved hypothetical protein
1613	0	71	12	16	4 average	average	average	P9301_15941	Conserved hypothetical protein
1683	0	56	9	20	3 average	average	average	P9301_15641	Conserved hypothetical protein
1988	0	84	11	20	1 average	average	average	P9301_12661	
209	0	73	14	10	7 average	average	average	P9301_10611	Conserved hypothetical protein
371	0	75	12	13	6 average	average	average	P9301_08981	conserved hypothetical
1526	0	78	13	17	1 average	average	average	P9301_16861	hypothetical
1980	0	38	13	18	0 average	average	ND	P9301_12781	
2006	0	43	14	14	3 average	average	average	A9601_03611	Conserved hypothetical protein
2062	0	98	5	22	4 average	average	average	A9601_12551	
969	0	59	8	20	2 average	average	average	P9301_03321	hypothetical
1671	0	53	10	18	2 average	average	average	P9301_15771	Conserved hypothetical protein
1978	0	50	9	17	4 average	average	average	P9301_12791	Conserved hypothetical protein
2071	0	86	10	16	4 average	average	average	P9301_12851	possible Fumarate reductase subunit D
2182	0	44	11	15	4 average	average	average	P9301_04931	Photosystem I PsaJ protein (subunit IX)
9	0	90	6	17	6 average	average	average	A9601_12581	
155	0	59	12	16	1 average	average	average	P9301_11161	Conserved hypothetical protein
186	0	61	7	19	3 average	average	average	P9301_10861	hypothetical
260	0	73	11	14	4 average	average	average	P9301_10101	influenza RNA-dependent RNA polymerase-like
262	0	61	9	17	3 average	average	average	P9301_10081	Conserved hypothetical protein
395	0	60	15	10	4 average	average	average	P9301_08731	Conserved hypothetical protein
595	0	97	11	13	5 average	average	average	P9301_06501	possible Hantavirus glycoprotein G2
1430	0	59	5	21	3 average	average	average	P9301_17791	conserved hypothetical protein
1621	0	40	6	19	4 average	average	average	P9301_15901	Conserved hypothetical protein
1684	0	91	14	12	3 average	average	average	P9301_15631	conserved hypothetical protein
2063	0	71	15	13	1 average	average	average	P9301_12611	
2461	0	129	7	18	4 average	average	average	PMED4_03941	conserved hypothetical protein
168	0	65	4	21	3 average	average	average	P9301_11031	Conserved hypothetical protein
516	0	39	10	16	2 average	average	average	P9301_07491	Conserved hypothetical protein
1227	0	61	9	18	1 average	average	average	P9301_00731	Photosystem II protein X PsbX
1972	0	44	8	19	1 average	average	average	P9301_12921	Conserved hypothetical protein
2014	0	100	11	13	4 average	average	average	P9301_04031	
2026	0	77	9	13	6 average	average	average	P9301_05961	NADH dehydrogenase subunit NdhL (ndhL)
149	0	74	13	7	7 average	average	average	P9301_11221	Conserved hypothetical protein
202	0	109	5	19	3 average	average	average	P9301_10691	Helix-hairpin-helix DNA-binding motif class 1
508	0	84	8	15	4 average	average	average	P9301_07571	hypothetical
849	0	67	5	19	3 average	average	average	P9301_04061	
978	0	48	8	15	4 average	average	average	P9301_03221	Cytochrome b559 beta-subunit
1269	0	65	8	16	3 average	average	average	P9301_00331	conserved hypothetical protein
2030	0	51	8	15	4 average	average	average	P9301_07781	Conserved hypothetical protein

2044	0	88	8	19	0	average	average	ND	P9301_10551	hypothetical
2219	0	97	10	10	7	average	average	average	P9301_12071	hypothetical
48	0	42	8	15	3	average	average	average	P9301_12241	possible Photosystem II reaction center Y protein (PsbY)
141	0	48	13	12	1	average	average	average	P9301_11301	Conserved hypothetical protein
180	0	66	12	10	4	average	average	average	P9301_10921	Conserved hypothetical protein
518	0	65	5	16	5	average	average	average	P9301_07471	Conserved hypothetical protein
1567	0	81	11	11	4	average	average	average	P9301_16451	ATP synthase subunit c
1677	0	51	12	12	2	average	average	average	P9301_15711	Conserved hypothetical protein
1977	0	65	14	10	2	average	average	average	P9301_12861	Conserved hypothetical protein
1990	0	76	11	15	0	average	average	ND	P9301_00291	possible Transcription factor TFIID (or TATA-b
2031	0	33	13	12	1	average	average	average	P9301_08001	Cytochrome b6-f complex subunit VIII
157	0	100	9	15	1	average	average	average	P9301_11141	possible Integrin alpha cytoplasmic region
198	0	77	10	14	1	average	average	average	P9301_10731	Conserved hypothetical protein
509	0	46	5	13	7	average	average	average	P9301_07561	Conserved hypothetical protein
525	0	50	4	12	9	average	average	average	P9301_07391	Conserved hypothetical protein
526	0	46	8	15	2	average	average	average	P9301_07381	Conserved hypothetical protein
880	0	61	11	13	1	average	average	average	P9301_04041	Conserved hypothetical protein
909	0	47	9	12	4	average	average	average	P9301_03891	Conserved hypothetical protein
934	0	68	11	14	0	average	average	ND	P9301_03671	Conserved hypothetical protein
1950	0	47	9	14	2	average	average	average	P9301_13251	Conserved hypothetical protein
2058	0	101	9	14	2	average	average	average	P9301_12601	
12	0	105	9	14	1	average	average	average	P9301_00371	
1058	0	42	7	13	4	average	average	average	P9301_02441	Conserved hypothetical protein
1743	0	70	16	5	3	average	average	average	A9601_15171	Predicted membrane protein
2078	0	42	6	14	4	average	average	average	P9301_13101	possible high light inducible protein
14	0	72	6	14	3	average	average	average	P9301_15701	possible M protein repeat
159	0	59	10	11	2	average	average	average	P9301_11121	Conserved hypothetical protein
1656	0	50	11	9	3	average	average	average	A9601_16181	
1673	0	47	9	11	3	average	average	average	P9301_15751	Conserved hypothetical protein
143	0	47	7	14	1	average	average	average	P9301_11281	Conserved hypothetical protein
181	0	53	8	13	1	average	average	average	P9301_10911	possible Photosystem II reaction centre N prot
265	0	50	5	15	2	average	average	average	P9301_10051	Conserved hypothetical protein
369	0	44	6	13	3	average	average	average	P9301_09001	Conserved hypothetical protein
511	0	68	8	12	2	average	average	average	P9301_07541	Conserved hypothetical protein
512	0	56	7	13	2	average	average	average	P9301_07531	Conserved hypothetical protein
899	0	47	10	7	5	average	average	average	P9301_03961	Conserved hypothetical protein
1675	0	50	7	13	2	average	average	average	P9301_15731	Conserved hypothetical protein
1702	0	57	10	11	1	average	average	average	P9301_15451	Conserved hypothetical protein
1974	0	56	5	13	4	average	average	average	P9301_12891	Conserved hypothetical protein
172	0	54	8	8	5	average	average	average	P9301_11001	Conserved hypothetical protein
259	0	52	5	13	3	average	average	average	P9301_10111	Conserved hypothetical protein
327	0	45	6	13	2	average	average	average	P9301_09431	Conserved hypothetical protein
345	0	110	9	12	0	average	average	ND	P9301_09241	conserved hypothetical protein
1197	0	47	8	9	4	average	average	average	P9301_01031	Conserved hypothetical protein
1665	0	54	6	12	3	average	average	average	P9301_15851	Conserved hypothetical protein
1788	0	43	3	14	4	average	average	average	P9301_14621	conserved hypothetical protein
1983	0	54	8	10	3	average	average	average	P9301_12731	Conserved hypothetical protein
2007	0	36	7	14	0	average	average	ND	P9301_03711	Conserved hypothetical protein
2022	0	47	7	12	2	average	average	average	P9301_04491	Conserved hypothetical protein
2446	0	103	7	12	2	average	average	average	P9312_17691	hypothetical protein
2470	0	87	6	10	5	average	average	average	PMED4_04101	hypothetical
11	0	60	8	10	2	average	average	average	P9301_07461	Conserved hypothetical protein
151	0	53	4	14	2	average	average	average	P9301_11201	Conserved hypothetical protein
152	0	75	9	4	7	average	average	average	P9301_13321	
389	0	57	5	12	3	average	average	average	P9301_08791	Conserved hypothetical protein
515	0	60	8	12	0	average	average	ND	P9301_07501	Conserved hypothetical protein
1129	0	62	5	8	7	average	average	average	P9301_01741	Conserved hypothetical protein
1655	0	48	5	8	7	average	average	average	A9601_16191	Conserved hypothetical protein
1910	0	57	6	11	3	average	average	average	P9301_13681	Conserved hypothetical protein
2303	0	106	6	13	1	average	average	average	P9301_16871	
160	0	59	5	14	0	average	average	ND	P9301_11111	Conserved hypothetical protein
193	0	49	4	12	3	average	average	average	P9301_10781	Conserved hypothetical protein
264	0	59	5	11	3	average	average	average	P9301_10061	Conserved hypothetical protein
267	0	50	5	11	3	average	average	average	A9601_10051	
394	0	49	5	10	4	average	average	average	P9301_08741	Conserved hypothetical protein
495	0	43	7	9	3	average	average	average	P9301_07701	Conserved hypothetical protein
680	0	44	5	12	2	average	average	average	P9301_05661	possible photosystem I reaction centre subunit XII (PsaM)
936	0	46	8	7	4	average	average	average	P9301_03651	Conserved hypothetical protein
1682	0	60	5	13	1	average	average	average	P9301_15651	Conserved hypothetical protein
1968	0	62	11	8	0	average	average	ND	A9601_12541	
2074	0	76	4	13	2	average	average	average	P9301_12941	
2080	0	32	9	6	4	average	average	average	P9301_13441	cytochrome b6-F complex subunit VII
488	0	48	8	8	2	average	average	average	P9301_07791	Conserved hypothetical protein
1651	0	46	9	8	1	average	average	average	P9301_15801	high light inducible protein-like
1680	0	54	6	11	1	average	average	average	P9301_15671	Conserved hypothetical protein
1915	0	79	6	12	0	average	average	ND	P9301_13631	peptidase family M20/M25/M40-like
2001	0	64	6	9	3	average	average	average	A9601_02341	
2073	0	43	7	11	0	average	average	ND	P9301_12911	Conserved hypothetical protein
497	0	45	9	7	1	average	average	average	P9301_07681	Conserved hypothetical protein
1199	0	58	4	10	3	average	average	average	P9301_01011	Conserved hypothetical protein
1238	0	45	7	10	0	average	average	ND	P9301_00611	Conserved hypothetical protein
1612	0	51	10	4	3	average	average	average	P9301_15951	Conserved hypothetical protein
2151	0	46	4	10	3	average	average	average	P9301_15811	Conserved hypothetical protein
2179	0	55	2	14	1	average	average	average	P9301_03451	Conserved hypothetical protein
2225	0	90	5	8	4	average	average	average	P9301_12351	bacterial regulatory proteins, ArsR family
123	0	63	8	8	0	average	average	ND	P9301_11491	Conserved hypothetical protein
392	0	49	4	10	2	average	average	average	P9301_08761	Conserved hypothetical protein
558	0	49	5	8	3	average	average	average	P9301_07101	Conserved hypothetical protein

939	0	37	4	10	2	average	average	average	A9601_03631	Conserved hypothetical protein
959	0	50	8	6	2	average	average	average	P9301_03421	possible Photosystem II reaction center M protein (PsbM)
962	0	32	4	10	2	average	average	average	P9301_03391	Photosystem II PsbT protein
1025	0	50	5	11	0	average	average	ND	P9301_02761	Photosystem II reaction centre N protein (psbN)
1955	0	42	8	6	2	average	average	average	P9301_13201	possible high light inducible protein
2037	0	47	7	8	1	average	average	average	P9301_08701	
2048	0	38	8	7	1	average	average	average	P9301_10811	Conserved hypothetical protein
2325	0	63	6	5	5	average	average	average	P9312_08681	Zn-ribbon protein
2432	0	49	4	11	1	average	average	average	P9312_16431	
133	0	32	3	10	2	average	average	average	P9301_11381	Conserved hypothetical protein
524	0	45	2	10	3	average	average	average	P9301_07401	Conserved hypothetical protein
1502	0	37	7	4	4	average	average	average	P9301_17121	photosystem I subunit VIII (PsaI)
1903	0	85	6	6	3	average	average	average	A9601_13661	Uncharacterized conserved protein
1920	0	52	4	7	4	average	average	average	P9301_13581	
1959	0	53	7	6	2	average	average	average	P9301_13051	Conserved hypothetical protein
2027	0	41	8	5	2	average	average	average	A9601_06441	
2039	0	29	5	8	2	average	average	average	P9301_09081	
2066	0	38	6	5	4	average	average	average	P9301_12801	Conserved hypothetical protein
144	0	58	7	5	2	average	average	average	P9301_11271	Conserved hypothetical protein
211	0	64	4	7	3	average	average	average	P9301_10591	Conserved hypothetical protein
43	0	55	5	7	1	average	average	average	P9301_12281	Conserved hypothetical protein
158	0	44	3	8	2	average	average	average	P9301_11131	Conserved hypothetical protein
188	0	68	4	9	0	average	average	ND	P9301_10841	
477	0	65	5	6	2	average	average	average	P9301_07901	Conserved hypothetical protein
900	0	52	8	2	3	average	average	average	P9301_03951	hypothetical
912	0	65	4	9	0	average	average	ND	P9301_03881	Conserved hypothetical protein
2003	0	68	4	8	1	average	average	average	P9301_03291	Conserved hypothetical protein
2060	0	40	5	7	1	average	average	average	A9601_12461	
2088	0	61	3	9	1	average	average	average	P9301_13591	
2447	0	38	4	9	0	average	average	ND	P9312_17811	50S Ribosomal protein L36
504	0	41	4	6	2	average	average	average	P9301_07611	Conserved hypothetical protein
532	0	49	6	5	1	average	average	average	P9301_10021	
1681	0	60	4	6	2	average	average	average	P9301_15661	Conserved hypothetical protein
1693	0	42	6	4	2	average	average	average	P9301_15541	Conserved hypothetical protein
2043	0	47	5	5	2	average	average	average	A9601_10501	Conserved hypothetical protein
2157	0	46	4	5	3	average	average	average	P9301_15971	Conserved hypothetical protein
2158	0	40	3	9	0	average	average	ND	A9601_16171	
208	0	75	4	5	2	average	average	average	P9301_10621	Conserved hypothetical protein
1254	0	47	7	3	1	average	average	average	P9301_00421	Conserved hypothetical protein
1657	0	37	4	6	1	average	average	average	A9601_16161	
2046	0	37	5	6	0	average	average	ND	P9301_10631	
7	0	56	5	5	0	average	average	ND	P9301_16011	
370	0	41	3	5	2	average	average	average	P9301_08991	Conserved hypothetical protein
1949	0	52	3	7	0	average	average	ND	P9301_13261	Conserved hypothetical protein
1173	0	60	5	3	1	average	average	average	P9301_01271	Conserved hypothetical protein
1646	0	51	1	5	3	average	average	average	A9601_12451	
1686	0	40	2	7	0	average	average	ND	P9301_15611	Conserved hypothetical protein
2152	0	49	2	6	1	average	average	average	A9601_15991	Conserved hypothetical protein
2208	0	33	2	5	2	average	average	average	P9301_07111	
2261	0	76	3	4	2	average	average	average	P9301_13601	
498	0	30	1	7	0	average	average	ND	P9301_07671	
911	0	31	3	5	0	average	average	ND	P9312_03791	
1234	0	51	2	4	2	average	average	average	P9301_00661	Conserved hypothetical protein
1685	0	44	4	4	0	average	average	ND	P9301_15621	Conserved hypothetical protein
2029	0	85	0	6	2	ND	average	average	A9601_07451	
2210	0	50	1	4	3	average	average	average	P9301_07731	
5	0	42	2	5	0	average	average	ND	A9601_12381	
388	0	36	2	3	2	average	average	average	P9301_08801	Conserved hypothetical protein
474	0	60	5	2	0	average	average	ND	P9301_07931	Conserved hypothetical protein
887	0	64	5	2	0	average	average	ND	P9301_13551	Conserved hypothetical protein
910	0	46	3	2	2	average	average	average	P9312_03801	
2365	0	84	0	6	1	ND	average	average	P9312_12551	cytochrome oxidase C subunit VIb-like
17	0	48	1	3	2	average	average	average	P9301_12711	Conserved hypothetical protein
261	0	56	0	4	2	ND	average	average	P9301_10091	Conserved hypothetical protein
873	0	53	2	4	0	average	average	ND	P9312_03221	Conserved hypothetical protein
1761	0	35	0	3	3	ND	average	average	P9301_14861	Conserved hypothetical protein
2148	0	55	3	3	0	average	average	ND	A9601_15771	
2327	0	59	2	3	1	average	average	average	P9312_09851	Conserved hypothetical protein
2389	0	45	3	2	1	average	average	average	P9312_13551	
251	0	25	1	1	3	average	average	average	PMED4_09461	Conserved hypothetical protein
878	0	49	0	5	0	ND	average	ND	P9301_10011	
940	0	45	2	3	0	average	average	ND	P9301_03621	Conserved hypothetical protein
1637	0	35	2	2	1	average	average	average	P9301_16001	
1639	0	38	0	5	0	ND	average	ND	A9601_16221	
2005	0	54	0	2	3	ND	average	average	A9601_03531	
2017	0	51	2	3	0	average	average	ND	A9601_04091	
2072	0	73	1	4	0	average	average	ND	P9301_12871	
2075	0	65	1	3	1	average	average	average	A9601_12881	
2165	0	36	1	4	0	average	average	ND	P9301_17031	
2458	0	42	3	2	0	average	average	ND	PMED4_03891	
3293	0	64	0	5	0	ND	average	ND	NATL1_00401	
6	0	52	0	3	1	ND	average	average	PMED4_16221	
164	0	41	0	4	0	ND	average	ND	P9301_11071	Conserved hypothetical protein
468	0	60	0	4	0	ND	average	ND	A9601_07991	Conserved hypothetical protein
879	0	45	0	2	2	ND	average	average	P9301_04051	Conserved hypothetical protein
914	0	54	2	2	0	average	average	ND	P9301_03861	Conserved hypothetical protein
922	0	43	2	1	1	average	average	average	P9301_03791	Conserved hypothetical protein
941	0	30	0	2	2	ND	average	average	P9301_03611	Conserved hypothetical protein

2064	0	55	3	1	0	average	average	ND	P9301_12741	possible chorismate binding enzyme
2316	0	52	1	3	0	average	average	ND	P9312_07161	
238	0	53	0	3	0	ND	average	ND	P9301_10311	Conserved hypothetical protein
258	0	37	2	1	0	average	average	ND	P9312_10071	
473	0	42	1	1	1	average	average	average	P9301_07941	Conserved hypothetical protein
886	0	41	0	3	0	ND	average	ND	P9301_13561	Conserved hypothetical protein
1003	0	37	0	3	0	ND	average	ND	P9301_02971	
1026	0	55	0	3	0	ND	average	ND	P9301_02751	
1658	0	38	0	3	0	ND	average	ND	P9301_15961	Conserved hypothetical protein
1979	0	34	1	2	0	average	average	ND	A9601_12711	
2061	0	35	0	2	1	ND	average	average	A9601_12501	
2070	0	45	1	2	0	average	average	ND	P9301_12841	gibberellin regulated protein-like
3173	0	53	0	1	2	ND	average	average	NATL2_15771	
108	0	39	0	1	1	ND	average	average	P9301_11641	Cytochrome b6/f complex, subunit V
1664	0	51	0	2	0	ND	average	ND	P9301_15861	
1964	0	38	0	2	0	ND	average	ND	P9301_12981	Conserved hypothetical protein
2052	0	32	1	1	0	average	average	ND	P9301_11421	Conserved hypothetical protein
2059	0	55	1	1	0	average	average	ND	A9601_12441	
2076	0	38	1	1	0	average	average	ND	P9301_12971	Conserved hypothetical protein
2232	0	55	0	1	1	ND	average	average	P9301_12441	
2517	0	42	0	1	1	ND	average	average	PMED4_13231	
2580	0	38	0	1	1	ND	average	average	PMED4_17481	50S Ribosomal protein L36
3047	0	54	1	1	0	average	average	ND	NATL2_12941	
93	0	37	0	1	0	ND	average	ND	P9301_11791	Conserved hypothetical protein
1638	0	56	0	1	0	ND	average	ND	A9601_16231	
2145	0	53	0	1	0	ND	average	ND	A9601_15211	
2230	0	37	0	1	0	ND	average	ND	P9301_12421	Conserved hypothetical protein
2443	0	42	0	1	0	ND	average	ND	P9312_16661	twin-arginine translocation pathway signal sequence domain protein
2484	0	43	0	1	0	ND	average	ND	PMED4_07451	
2705	0	52	0	1	0	ND	average	ND	P9515_16021	
2762	0	54	0	1	0	ND	average	ND	NATL2_02351	
2803	0	49	0	1	0	ND	average	ND	NATL2_05361	
2840	0	55	0	1	0	ND	average	ND	NATL2_08531	
2905	0	46	0	1	0	ND	average	ND	NATL2_10711	
2926	0	47	0	1	0	ND	average	ND	NATL2_11061	
3011	0	47	0	1	0	ND	average	ND	NATL2_12471	
3085	0	35	0	1	0	ND	average	ND	NATL2_13441	
3090	0	57	0	1	0	ND	average	ND	NATL2_13531	
3441	0	55	0	1	0	ND	average	ND	NATL1_18561	
1635	0	141	46	60	12	high	average	average	P9301_16041	conserved hypothetical protein
210	0	133	52	39	17	high	average	average	P9301_10601	possible ATP synthase protein 8
935	0	102	38	45	10	high	average	average	P9301_16021	Conserved hypothetical protein
10	0	93	31	39	15	high	average	average	P9301_13011	Conserved hypothetical protein
18	0	94	22	38	20	average	average	high	P9301_11101	hypothetical
758	0	105	23	33	20	average	average	high	P9301_04881	conserved hypothetical protein
1837	0	364	29	44	10	rare	rare	rare	P9301_14441	GDPmannose 4,6-dehydratase
166	0	289	22	40	7	rare	rare	rare	P9301_11051	conserved hypothetical protein
2252	0	433	27	39	3	rare	rare	rare	P9301_13111	
41	0	514	19	41	4	rare	rare	rare	P9301_14321	
36	0	372	21	22	10	rare	rare	rare	P9312_12781	adhesin-like protein
875	0	619	17	28	5	rare	rare	rare	P9312_03241	carbamoyltransferase
2122	0	4723	11	29	8	rare	rare	rare	A9601_14361	
29	0	570	11	27	8	rare	rare	rare	P9301_12651	Succinate dehydrogenase/fumarate reductase, flavoprotein subunit
2257	0	326	23	19	3	rare	rare	rare	P9301_13161	
2228	0	408	9	28	7	rare	rare	rare	P9301_12381	putative chromate transporter, CHR family
2229	0	386	15	21	8	rare	rare	rare	P9301_12411	two-component sensor histidine kinase, phosphate sensing
2338	0	559	14	21	8	rare	rare	rare	P9312_12231	hypothetical protein
1965	0	434	10	24	8	rare	rare	rare	P9301_12961	Mn2+ and Fe2+ transporter, NRAMP family
2255	0	362	24	14	1	rare	rare	rare	P9301_13141	
850	0	260	13	21	4	rare	rare	rare	A9601_04251	probable periplasmic protein
2254	0	330	8	22	6	rare	rare	rare	P9301_13131	Predicted amidohydrolase
2117	0	968	10	8	11	rare	rare	rare	A9601_14311	
839	0	654	7	17	3	rare	rare	rare	A9601_04361	
1922	0	607	7	15	4	rare	rare	rare	A9601_13391	conserved hypothetical protein
2256	0	217	10	15	1	rare	rare	rare	P9301_13151	
842	0	698	7	10	8	rare	rare	rare	A9601_04331	DNA-cytosine methyltransferase
2102	0	629	7	14	4	rare	rare	rare	P9301_14281	Nucleotide-diphosphate-sugar epimerase, membrane associated
3414	0	271	11	7	6	rare	rare	rare	P9313_11921	conserved hypothetical protein
1859	0	337	4	15	4	rare	rare	rare	P9301_14451	Putative fucose synthetase
1877	0	352	6	12	5	rare	rare	rare	P9301_14241	UDP-glucose 4-epimerase
2231	0	242	6	15	2	rare	rare	rare	P9301_12431	two-component response regulator, phosphate
2189	0	568	5	13	3	rare	rare	rare	P9301_06911	Type II secretory pathway ATPase PulE/Tfp pilus assembly pathway ATPase PilB-like
2314	0	229	4	13	4	rare	rare	rare	P9312_12211	hypothetical protein
2464	0	561	9	11	1	rare	rare	rare	PMED4_04031	putative cyanate ABC transporter, substrate binding protein
1783	0	417	4	12	4	rare	rare	rare	A9601_17951	
2465	0	260	4	11	3	rare	rare	rare	PMED4_04041	putative cyanate ABC transporter
840	0	343	5	11	1	rare	rare	rare	A9601_04351	
2194	0	572	5	7	5	rare	rare	rare	P9301_06961	Type II secretory pathway component PulD-like
2171	0	213	4	10	1	rare	rare	rare	A9601_19081	Dolichol kinase
2515	0	671	6	6	3	rare	rare	rare	PMED4_12741	Domain of unknown function DUF33
2759	0	523	3	3	9	rare	rare	rare	NATL2_01771	
2188	0	357	3	8	2	rare	rare	rare	P9301_06901	twitching motility protein
2198	0	705	5	7	1	rare	rare	rare	P9301_07001	Conserved hypothetical protein
2315	0	248	3	7	3	rare	rare	rare	P9312_06881	hypothetical protein

2514	0	491	7	4	2	rare	rare	rare	PMED4_12721	possible Natural resistance-associated macroph
3171	0	779	1	2	10	rare	rare	rare	NATL2_15731	
3415	0	316	3	7	3	rare	rare	rare	P9313_11931	conserved hypothetical protein
2159	0	577	4	6	2	rare	rare	rare	A9601_16271	
3274	0	690	1	4	7	rare	rare	rare	NATL2_21051	
1880	0	422	4	2	5	rare	rare	rare	P9301_13971	Conserved hypothetical protein
2187	0	453	2	6	3	rare	rare	rare	P9301_06891	Type II secretory pathway component PulF-like
1	0	632	3	3	4	rare	rare	rare	P9312_13011	TPR repeat
1831	0	384	1	7	2	rare	rare	rare	P9301_14251	
1999	0	440	3	6	1	rare	rare	rare	A9601_00601	
2012	0	255	3	6	1	rare	rare	rare	A9601_03901	possible Glycosyl transferase
841	0	237	1	6	2	rare	rare	rare	A9601_04341	
1822	0	586	2	5	2	rare	rare	rare	P9301_14231	
2118	0	387	3	5	1	rare	rare	rare	A9601_14321	
2360	0	364	1	7	1	rare	rare	rare	P9312_12461	ATP-dependent DNA ligase
2370	0	297	2	2	5	rare	rare	rare	P9312_12611	Amino acid ABC transporter, permease protein, 3-TM region, His/Glu/Gln/Arg/opine
2371	0	349	1	2	6	rare	rare	rare	P9312_12621	extracellular solute-binding protein, family 3
2466	0	284	3	4	2	rare	rare	rare	PMED4_04051	putative cyanate ABC transporter
2777	0	263	1	2	6	rare	rare	rare	NATL2_03721	
30	0	459	2	3	3	rare	rare	rare	P9301_12621	
2237	0	261	2	3	3	rare	rare	rare	P9301_12501	
2321	0	532	3	1	4	rare	rare	rare	P9312_07241	hypothetical protein
2369	0	343	3	1	4	rare	rare	rare	P9312_12601	Amino acid ABC transporter, permease protein, 3-TM region, His/Glu/Gln/Arg/opine
2775	0	434	1	4	3	rare	rare	rare	NATL2_03691	
2780	0	331	4	1	3	rare	rare	rare	NATL2_03791	
2814	0	698	2	2	4	rare	rare	rare	NATL2_06951	
2355	0	546	3	2	2	rare	rare	rare	P9312_12411	uncharacterized conserved protein containing SWIM-like Zn-finger
2525	0	399	3	3	1	rare	rare	rare	PMED4_13701	NDP-hexose 3,4-dehydratase
2748	0	734	1	1	5	rare	rare	rare	NATL2_01231	
847	0	901	4	1	1	rare	rare	rare	A9601_04281	
2137	0	288	3	2	1	rare	rare	rare	A9601_14541	possible N-terminal fragment of transketolase
2169	0	453	4	1	1	rare	rare	rare	A9601_17341	
2173	0	921	2	2	2	rare	rare	rare	P9301_00541	
2323	0	760	2	3	1	rare	rare	rare	P9312_07661	putative secreted protein
2384	0	201	2	2	2	rare	rare	rare	P9312_13041	hypothetical protein
2554	0	225	2	3	1	rare	rare	rare	PMED4_15661	unknown
2644	0	236	2	3	1	rare	rare	rare	P9515_12821	
2657	0	456	3	1	2	rare	rare	rare	P9515_13781	
2960	0	489	1	3	2	rare	rare	rare	NATL2_11661	
325	0	151	1	3	1	rare	rare	rare	A9601_09461	putative lipoprotein signal peptidase
2094	0	318	1	2	2	rare	rare	rare	P9301_14271	UDP-N-acetylmuramyl pentapeptide phosphotransferase/UDP-N- acetylglucosamine-1-phosphate transferase
2116	0	245	3	1	1	rare	rare	rare	A9601_14301	
2195	0	497	1	3	1	rare	rare	rare	P9301_06971	Conserved hypothetical protein
2660	0	442	2	1	2	rare	rare	rare	P9515_13821	
2119	0	291	1	2	1	rare	rare	rare	A9601_14331	
2538	0	610	1	2	1	rare	rare	rare	PMED4_14111	Carbamoyltransferase
2679	0	682	2	1	1	rare	rare	rare	P9515_14051	
2761	0	386	1	1	2	rare	rare	rare	NATL2_02221	
3240	0	187	1	1	2	rare	rare	rare	NATL2_19081	
2196	0	201	1	1	1	rare	rare	rare	P9301_06981	Conserved hypothetical protein
3065	0	152	1	1	1	rare	rare	rare	NATL2_13151	
3138	0	370	1	1	1	rare	rare	rare	NATL2_14951	
45	0	502	56	58	20	rare	rare	average	P9301_00391	TPR repeat
1989	0	360	36	56	13	rare	rare	average	P9301_02451	Photosystem II PsbA protein (D1)
2227	0	419	23	59	20	rare	rare	average	P9301_12371	multidrug efflux transporter, MFS family
2144	0	406	36	46	15	rare	rare	average	P9301_14651	putative bifuntional enzyme: tRNA methyltransferase; 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase
2289	0	238	12	25	14	rare	rare	average	P9301_16031	dienelactone hydrolase
2260	0	182	19	9	2	average	rare	rare	P9301_13541	
2011	0	277	6	16	0	rare	rare	ND	A9601_03891	Glycosyl transferase family 11
3413	0	236	6	7	5	rare	rare	average	P9313_11911	conserved hypothetical protein
2253	0	160	4	8	3	rare	rare	average	P9301_13121	
1602	0	168	6	8	0	rare	rare	ND	PMED4_04201	possible Ribosomal RNA adenine dimethylase
2429	0	149	5	9	0	rare	rare	ND	P9312_16201	hypothetical protein
3128	0	695	0	2	11	ND	rare	rare	NATL2_14561	
2467	0	147	5	4	3	rare	rare	average	PMED4_04061	Cyanate lyase
848	0	137	7	3	1	rare	rare	average	A9601_04271	
2396	0	243	2	3	6	rare	rare	average	P9312_14241	Sugar transferases involved in lipopolysaccharide synthesis-like
2832	0	518	0	2	9	ND	rare	rare	NATL2_07821	
3195	0	407	0	2	9	ND	rare	rare	NATL2_16811	
3215	0	641	0	5	6	ND	rare	rare	NATL2_17571	
3283	0	352	0	1	9	ND	rare	rare	NATL2_21291	
838	0	321	2	7	0	rare	rare	ND	A9601_04371	
903	0	119	2	7	0	rare	rare	ND	P9515_04011	
2558	0	130	3	4	2	rare	rare	average	PMED4_15911	hypothetical
2757	0	544	0	2	7	ND	rare	rare	NATL2_01661	
2783	0	349	0	3	6	ND	rare	rare	NATL2_04031	
2839	0	287	0	3	6	ND	rare	rare	NATL2_08451	
866	0	1214	4	4	0	rare	rare	ND	P9301_12591	
1615	0	95	1	6	1	rare	rare	average	P9515_12371	
2172	0	437	2	6	0	rare	rare	ND	P9301_00381	
2688	0	432	2	6	0	rare	rare	ND	P9515_14171	
890	0	89	1	6	0	rare	rare	ND	A9601_12371	

2002	0	135	1	6	0	rare	rare	ND	P9301_02561	possible Methylpurine-DNA glycosylase (MPG)
2258	0	91	1	6	0	rare	rare	ND	P9301_13171	
2259	0	103	3	4	0	rare	rare	ND	P9301_13181	
2459	0	106	1	5	1	rare	rare	average	PMED4_03911	possible Malic enzyme
2460	0	134	2	5	0	rare	rare	ND	PMED4_03931	Type-1 copper (blue) domain
2766	0	311	0	3	4	ND	rare	rare	NATL2_02721	
561	0	125	2	3	1	rare	rare	average	PMED4_07021	possible 5'-3' exonuclease, C-terminal SAM fol
1992	0	248	4	2	0	rare	rare	ND	P9301_00551	
2047	0	89	1	2	3	rare	rare	average	P9301_10651	
2246	0	209	1	5	0	rare	rare	ND	P9301_12631	
2291	0	150	4	1	1	rare	rare	average	P9301_16061	
2359	0	395	1	5	0	rare	rare	ND	P9312_12451	hypothetical protein
3127	0	486	0	1	5	ND	rare	rare	NATL2_14521	
3256	0	384	0	1	5	ND	rare	rare	NATL2_20271	
3281	0	303	0	2	4	ND	rare	rare	NATL2_21271	
865	0	84	1	4	0	rare	rare	ND	A9601_04181	
1645	0	142	2	2	1	rare	rare	average	A9601_04071	
1993	0	146	3	1	1	rare	rare	average	P9301_00571	
2020	0	166	2	1	2	rare	rare	average	A9601_04231	
2233	0	133	1	4	0	rare	rare	ND	P9301_12451	possible myosin N-terminal SH3-like domain
2363	0	97	1	4	0	rare	rare	ND	P9312_12501	vanadium/alternative nitrogenase delta-like
2457	0	133	1	3	1	rare	rare	average	PMED4_03881	possible MarR family
2574	0	135	1	2	2	rare	rare	average	PMED4_16231	Staphylococcus nuclease (SNase) homologues
2676	0	529	2	3	0	rare	rare	ND	P9515_14021	
3292	0	255	1	4	0	rare	rare	ND	NATL1_00391	Glutathione S-transferase
2393	0	347	0	4	1	ND	rare	rare	P9312_14211	Pyruvate dehydrogenase (lipoamide)
2716	0	402	0	2	3	ND	rare	rare	NATL2_00381	
2855	0	310	0	2	3	ND	rare	rare	NATL2_09341	
2871	0	230	0	2	3	ND	rare	rare	NATL2_09971	
3280	0	553	0	1	4	ND	rare	rare	NATL2_21261	
3294	0	287	0	4	1	ND	rare	rare	NATL1_00411	
845	0	661	1	3	0	rare	rare	ND	A9601_04301	
883	0	254	1	3	0	rare	rare	ND	A9601_04161	Putative dehydrogenase
1782	0	245	1	3	0	rare	rare	ND	P9312_14971	tRNA (guanine-N1)-methyltransferase
2053	0	330	2	2	0	rare	rare	ND	A9601_12311	
2127	0	501	3	1	0	rare	rare	ND	A9601_14421	putative ADP-heptose synthase
2309	0	303	3	1	0	rare	rare	ND	P9312_03921	hypothetical protein
2326	0	154	1	1	2	rare	rare	average	P9312_09491	putative transcriptional regulator
2354	0	990	1	3	0	rare	rare	ND	P9312_12401	protein with signal peptide
2699	0	581	2	2	0	rare	rare	ND	P9515_15731	
3097	0	131	1	3	0	rare	rare	ND	NATL2_13641	
1849	0	296	0	3	1	ND	rare	rare	P9515_14101	glucose-1-phosphate thymidylyltransferase
2442	0	210	0	1	3	ND	rare	rare	P9312_16611	trypsin-like
2717	0	242	0	2	2	ND	rare	rare	NATL2_00391	
633	0	88	1	1	1	rare	rare	average	P9301_06121	conserved hypothetical
2092	0	314	1	2	0	rare	rare	ND	P9301_14161	
2104	0	370	1	2	0	rare	rare	ND	P9301_14311	
2112	0	283	2	1	0	rare	rare	ND	A9601_14231	
2130	0	223	1	2	0	rare	rare	ND	A9601_14451	SAM-dependent methyltransferase
2192	0	241	1	2	0	rare	rare	ND	P9301_06941	Conserved hypothetical protein
2197	0	191	2	1	0	rare	rare	ND	P9301_06991	Conserved hypothetical protein
2203	0	273	1	2	0	rare	rare	ND	P9301_07051	
2293	0	132	2	1	0	rare	rare	ND	P9301_16081	
2356	0	482	1	2	0	rare	rare	ND	P9312_12421	retron-type reverse transcriptase
2378	0	311	1	2	0	rare	rare	ND	P9312_12791	hypothetical protein
2472	0	155	1	2	0	rare	rare	ND	PMED4_04171	conserved hypothetical protein
2480	0	417	1	2	0	rare	rare	ND	PMED4_05631	putative dihydroorotase
2577	0	122	1	2	0	rare	rare	ND	PMED4_16281	possible Gibberellin regulated protein
2613	0	432	1	2	0	rare	rare	ND	P9515_07291	
2652	0	322	2	1	0	rare	rare	ND	P9515_13291	
2689	0	377	2	1	0	rare	rare	ND	P9515_14181	
2779	0	362	2	1	0	rare	rare	ND	NATL2_03781	
1821	0	441	0	1	2	ND	rare	rare	P9312_14181	UDP-glucose 6-dehydrogenase
1995	0	166	0	2	1	ND	rare	rare	A9601_00561	
2240	0	318	0	2	1	ND	rare	rare	P9301_12531	Putative dehydrogenase
2718	0	264	0	1	2	ND	rare	rare	NATL2_00401	
2805	0	182	0	1	2	ND	rare	rare	NATL2_05921	
3132	0	357	0	1	2	ND	rare	rare	NATL2_14721	
3158	0	213	0	1	2	ND	rare	rare	NATL2_15521	
1860	0	375	1	1	0	rare	rare	ND	A9601_13921	putative glycosyl transferase, group 1
2095	0	297	1	1	0	rare	rare	ND	P9301_14261	
2193	0	288	1	1	0	rare	rare	ND	P9301_06951	Conserved hypothetical protein
2296	0	170	1	1	0	rare	rare	ND	P9301_16111	
2300	0	328	1	1	0	rare	rare	ND	P9301_16151	
2320	0	110	1	1	0	rare	rare	ND	P9312_07231	hypothetical protein
2394	0	329	1	1	0	rare	rare	ND	P9312_14221	Pyruvate dehydrogenase (lipoamide)
2421	0	271	1	1	0	rare	rare	ND	P9312_14681	hypothetical protein
2675	0	334	1	1	0	rare	rare	ND	P9515_14011	
2680	0	248	1	1	0	rare	rare	ND	P9515_14061	
2682	0	310	1	1	0	rare	rare	ND	P9515_14091	
2867	0	136	1	1	0	rare	rare	ND	NATL2_09791	
3515	0	687	1	1	0	rare	rare	ND	SS120_06341	DNA helicase, predicted restriction/modification system component, ortholog of BS_yeeB
2138	0	304	0	1	1	ND	rare	rare	A9601_14551	Transketolase, C-terminal subunit
2222	0	234	0	1	1	ND	rare	rare	P9301_12311	putative hydrogenase accessory protein
2357	0	187	0	1	1	ND	rare	rare	P9312_12431	hypothetical protein
2373	0	336	0	1	1	ND	rare	rare	P9312_12651	NADPH-dependent reductase
2474	0	829	0	1	1	ND	rare	rare	PMED4_04231	possible uncharacterized restriction enzyme, interrupted-C
2767	0	220	0	1	1	ND	rare	rare	NATL2_02791	

2823	0	151	0	1	1	ND	rare	rare	NATL2_07311	
2896	0	210	0	1	1	ND	rare	rare	NATL2_10551	
3008	0	152	0	1	1	ND	rare	rare	NATL2_12441	
3145	0	1821	0	1	1	ND	rare	rare	NATL2_15121	
3184	0	739	0	1	1	ND	rare	rare	NATL2_16151	
3257	0	265	0	1	1	ND	rare	rare	NATL2_20401	
3451	0	2178	0	1	1	ND	rare	rare	NATL1_21051	
240	0	236	29	34	16	average	rare	average	P9301_10281	Membrane protein TerC, possibly involved in tellurium resistance
42	0	172	14	18	7	average	rare	average	P9301_14411	Adenine/guanine phosphoribosyltransferase or related PRPP-binding protein
2290	0	102	4	4	4	average	rare	average	P9301_16051	conserved hypothetical
562	0	66	5	1	3	average	rare	average	P9301_06831	Conserved hypothetical protein
2065	0	87	5	4	0	average	rare	ND	A9601_12681	
2160	0	173	0	5	4	ND	rare	average	A9601_16281	
2374	0	201	0	9	0	ND	rare	ND	P9312_12661	hypothetical protein
2571	0	112	0	7	2	ND	rare	average	PMED4_16191	Hypothetical protein
2445	0	74	4	3	1	average	rare	average	P9312_17621	photosystem I subunit VIII (PsaI)
2433	0	100	0	3	3	ND	rare	average	P9312_16461	RNA-binding region RNP-1
2849	0	99	0	4	2	ND	rare	average	NATL2_08761	
529	0	70	4	1	0	average	rare	ND	A9601_12471	
2107	0	400	0	5	0	ND	rare	ND	A9601_14181	Predicted dehydrogenase
2395	0	367	0	5	0	ND	rare	ND	P9312_14231	pyridoxal phosphate-dependent enzyme apparently involved in regulation of cell wall biogenesis
2498	0	93	0	1	4	ND	rare	average	PMED4_09021	hypothetical
2601	0	134	0	3	2	ND	rare	average	P9515_04321	
2837	0	111	0	1	4	ND	rare	average	NATL2_08431	
191	0	59	3	1	0	average	rare	ND	P9301_10801	Conserved hypothetical protein
2224	0	70	2	1	1	average	rare	average	P9301_12341	Conserved hypothetical protein
2449	0	81	0	1	3	ND	rare	average	P9312_19551	hypothetical protein
2641	0	390	0	4	0	ND	rare	ND	P9515_12791	
2662	0	780	0	4	0	ND	rare	ND	P9515_13861	
3039	0	180	0	1	3	ND	rare	average	NATL2_12851	
3040	0	124	0	1	3	ND	rare	average	NATL2_12861	
3122	0	118	0	2	2	ND	rare	average	NATL2_14041	
2015	0	83	0	1	2	ND	rare	average	A9601_04061	
2183	0	58	1	1	1	average	rare	average	P9301_05081	
2202	0	189	0	3	0	ND	rare	ND	P9301_07041	
2285	0	85	0	2	1	ND	rare	average	P9301_15041	conserved hypothetical
2468	0	107	0	3	0	ND	rare	ND	PMED4_04081	possible Type II intron maturase
2524	0	360	0	3	0	ND	rare	ND	PMED4_13681	CDP-glucose 4,6-dehydratase
2532	0	309	0	3	0	ND	rare	ND	PMED4_13951	dehydrogenase E1 component beta subunit
2612	0	252	0	3	0	ND	rare	ND	P9515_07281	
2649	0	322	0	3	0	ND	rare	ND	P9515_13001	putative GDP-D-mannose dehydratase
2653	0	632	0	3	0	ND	rare	ND	P9515_13301	
2719	0	124	0	2	1	ND	rare	average	NATL2_00441	
2904	0	145	0	1	2	ND	rare	average	NATL2_10701	
2938	0	81	0	1	2	ND	rare	average	NATL2_11241	
2950	0	73	0	2	1	ND	rare	average	NATL2_11481	
2958	0	87	0	1	2	ND	rare	average	NATL2_11631	
2969	0	85	0	2	1	ND	rare	average	NATL2_11811	
190	0	89	0	2	0	ND	rare	ND	A9601_10821	possible Protein of unknown function DUF67
851	0	64	1	1	0	average	rare	ND	A9601_04241	
904	0	127	0	1	1	ND	rare	average	P9515_04001	
1404	0	71	1	1	0	average	rare	ND	A9601_18221	conserved hypothetical
1607	0	64	0	1	1	ND	rare	average	PMED4_09041	Conserved hypothetical protein
1634	0	77	0	1	1	ND	rare	average	P9515_12551	
1643	0	69	1	1	0	average	rare	ND	A9601_04111	
2098	0	257	0	2	0	ND	rare	ND	A9601_14071	glucose-1-phosphate cytidylyltransferase
2135	0	287	0	2	0	ND	rare	ND	A9601_14511	SAM-dependent methyltransferase
2174	0	144	0	2	0	ND	rare	ND	P9301_00561	
2249	0	61	1	1	0	average	rare	ND	P9301_12721	
2334	0	70	0	2	0	ND	rare	ND	P9312_11541	cytochrome b6/f complex, subunit V
2361	0	140	0	2	0	ND	rare	ND	P9312_12481	hypothetical protein
2364	0	62	1	1	0	average	rare	ND	P9312_12531	virion host shutoff protein-like
2516	0	163	0	2	0	ND	rare	ND	PMED4_12751	possible NAD(P) transhydrogenase beta subunit
2527	0	230	0	2	0	ND	rare	ND	PMED4_13771	conserved hypothetical protein
2581	0	61	0	1	1	ND	rare	average	PMED4_19191	Conserved hypothetical protein
2607	0	101	0	1	1	ND	rare	average	P9515_04431	
2615	0	402	0	2	0	ND	rare	ND	P9515_07581	
2656	0	359	0	2	0	ND	rare	ND	P9515_13771	
2658	0	389	0	2	0	ND	rare	ND	P9515_13791	
2661	0	314	0	2	0	ND	rare	ND	P9515_13841	
2838	0	72	0	1	1	ND	rare	average	NATL2_08441	
2864	0	68	0	1	1	ND	rare	average	NATL2_09751	
2874	0	75	0	1	1	ND	rare	average	NATL2_10151	
2906	0	97	0	1	1	ND	rare	average	NATL2_10721	
2913	0	86	0	1	1	ND	rare	average	NATL2_10821	
2916	0	81	0	2	0	ND	rare	ND	NATL2_10851	
2954	0	65	0	1	1	ND	rare	average	NATL2_11531	
2962	0	87	0	1	1	ND	rare	average	NATL2_11701	
2978	0	81	0	1	1	ND	rare	average	NATL2_11961	
2984	0	126	0	1	1	ND	rare	average	NATL2_12061	
3098	0	97	0	1	1	ND	rare	average	NATL2_13651	
3142	0	68	0	2	0	ND	rare	ND	NATL2_14991	
3218	0	137	0	2	0	ND	rare	ND	NATL2_17611	
3229	0	88	0	2	0	ND	rare	ND	NATL2_18061	
3698	0	395	0	2	0	ND	rare	ND	SS120_16021	Beta-lactamase class C and other penicillin binding proteins
39	0	150	0	1	0	ND	rare	ND	PMED4_18511	hypothetical protein

146	0	86	0	1	0	ND	rare	ND	P9301_11251	Conserved hypothetical protein
1493	0	131	0	1	0	ND	rare	ND	P9312_03231	hypothetical protein
1626	0	563	0	1	0	ND	rare	ND	P9515_12641	cytochrome c oxidase subunit I
1644	0	83	0	1	0	ND	rare	ND	A9601_04101	
1785	0	167	0	1	0	ND	rare	ND	P9312_14951	MECDP-synthase
1841	0	398	0	1	0	ND	rare	ND	P9301_14101	Predicted pyridoxal phosphate-dependent enzyme apparently involved in regulation of cell wall biogenesis
1852	0	402	0	1	0	ND	rare	ND	A9601_14051	
2028	0	68	0	1	0	ND	rare	ND	P9301_07421	
2051	0	387	0	1	0	ND	rare	ND	A9601_11341	ABC transporter, substrate binding protein, possibly Mn
2056	0	184	0	1	0	ND	rare	ND	A9601_12341	
2096	0	268	0	1	0	ND	rare	ND	A9601_14031	
2128	0	315	0	1	0	ND	rare	ND	A9601_14431	putative nucleoside-diphosphate sugar epimerase
2136	0	355	0	1	0	ND	rare	ND	A9601_14521	Zinc-containing alcohol dehydrogenase superfamily
2139	0	198	0	1	0	ND	rare	ND	A9601_14561	conserved hypothetical protein
2168	0	74	0	1	0	ND	rare	ND	A9601_17331	
2185	0	127	0	1	0	ND	rare	ND	P9301_06851	
2191	0	461	0	1	0	ND	rare	ND	P9301_06931	Conserved hypothetical protein
2199	0	208	0	1	0	ND	rare	ND	P9301_07011	Conserved hypothetical protein
2200	0	145	0	1	0	ND	rare	ND	P9301_07021	
2205	0	78	0	1	0	ND	rare	ND	P9301_07071	
2207	0	263	0	1	0	ND	rare	ND	P9301_07091	
2279	0	401	0	1	0	ND	rare	ND	P9301_14361	
2282	0	235	0	1	0	ND	rare	ND	P9301_14391	
2292	0	213	0	1	0	ND	rare	ND	P9301_16071	
2328	0	79	0	1	0	ND	rare	ND	P9312_10021	hypothetical protein
2408	0	526	0	1	0	ND	rare	ND	P9312_14471	hypothetical protein
2437	0	132	0	1	0	ND	rare	ND	P9312_16501	hypothetical protein
2439	0	72	0	1	0	ND	rare	ND	P9312_16531	hypothetical protein
2444	0	84	0	1	0	ND	rare	ND	P9312_16691	poly A polymerase regulatory subunit-like
2469	0	122	0	1	0	ND	rare	ND	PMED4_04091	hypothetical
2518	0	170	0	1	0	ND	rare	ND	PMED4_13241	possible Trehalase
2519	0	67	0	1	0	ND	rare	ND	PMED4_13251	Conserved hypothetical protein
2523	0	317	0	1	0	ND	rare	ND	PMED4_13651	Glycosyl transferase, family 2
2544	0	324	0	1	0	ND	rare	ND	PMED4_14201	putative glycosyl transferase, family 2
2547	0	388	0	1	0	ND	rare	ND	PMED4_14251	pyridoxal-phosphate-dependent aminotransferase
2640	0	326	0	1	0	ND	rare	ND	P9515_12781	
2648	0	151	0	1	0	ND	rare	ND	P9515_12991	
2659	0	363	0	1	0	ND	rare	ND	P9515_13801	
2725	0	564	0	1	0	ND	rare	ND	NATL2_00621	
2736	0	70	0	1	0	ND	rare	ND	NATL2_00751	
2791	0	583	0	1	0	ND	rare	ND	NATL2_04191	
2799	0	59	0	1	0	ND	rare	ND	NATL2_04591	
2809	0	95	0	1	0	ND	rare	ND	NATL2_06631	
2848	0	81	0	1	0	ND	rare	ND	NATL2_08711	
2892	0	180	0	1	0	ND	rare	ND	NATL2_10471	
2911	0	127	0	1	0	ND	rare	ND	NATL2_10771	
2917	0	58	0	1	0	ND	rare	ND	NATL2_10861	
2951	0	66	0	1	0	ND	rare	ND	NATL2_11501	
2972	0	58	0	1	0	ND	rare	ND	NATL2_11861	
2980	0	134	0	1	0	ND	rare	ND	NATL2_11981	
2992	0	73	0	1	0	ND	rare	ND	NATL2_12171	
2997	0	60	0	1	0	ND	rare	ND	NATL2_12271	
3121	0	107	0	1	0	ND	rare	ND	NATL2_14031	
3162	0	106	0	1	0	ND	rare	ND	NATL2_15611	
3163	0	113	0	1	0	ND	rare	ND	NATL2_15621	
3165	0	64	0	1	0	ND	rare	ND	NATL2_15641	
3166	0	71	0	1	0	ND	rare	ND	NATL2_15651	
3175	0	235	0	1	0	ND	rare	ND	NATL2_16031	
3177	0	661	0	1	0	ND	rare	ND	NATL2_16051	
3205	0	64	0	1	0	ND	rare	ND	NATL2_17401	
3208	0	92	0	1	0	ND	rare	ND	NATL2_17451	
3269	0	322	0	1	0	ND	rare	ND	NATL2_20591	
3275	0	151	0	1	0	ND	rare	ND	NATL2_21091	
3290	0	305	0	1	0	ND	rare	ND	NATL1_00371	Site-specific DNA methylase
3295	0	484	0	1	0	ND	rare	ND	NATL1_00421	
3460	0	1099	0	1	0	ND	rare	ND	P9313_27071	SNF2 related domain:DEAD/DEAH box helicase:Helicase C-termina...
3554	0	404	0	1	0	ND	rare	ND	P9313_10221	Serine/threonine specific protein phosphatase:DNA repair exon...
3615	0	818	0	1	0	ND	rare	ND	SS120_12861	Predicted protein
3668	0	593	0	1	0	ND	rare	ND	SS120_15501	Gamma-glutamyltransferase
3695	0	212	0	1	0	ND	rare	ND	SS120_15951	Predicted membrane protein
3712	0	532	0	1	0	ND	rare	ND	P9313_28191	Fe-S oxidoreductase
3743	0	93	0	1	0	ND	rare	ND	P9211_05921	
3921	0	387	0	1	0	ND	rare	ND	P9303_01061	
5155	0	293	0	1	0	ND	rare	ND	P9303_27001	
5255	0	334	0	1	0	ND	rare	ND	P9303_28711	
5403	0	1330	0	1	0	ND	rare	ND	P9313_03631	conserved hypothetical protein
2166	0	328	2	0	1	rare	ND	rare	A9601_17311	
3459	0	269	2	0	1	rare	ND	rare	P9313_27081	Uncharacterized conserved protein
2582	0	202	1	0	1	rare	ND	rare	P9515_00381	
2583	0	155	1	0	1	rare	ND	rare	P9515_00401	
2655	0	183	1	0	1	rare	ND	rare	P9515_13761	
3230	0	583	0	0	7	ND	ND	rare	NATL2_18231	
2752	0	386	0	0	6	ND	ND	rare	NATL2_01431	
2750	0	378	0	0	5	ND	ND	rare	NATL2_01331	
2774	0	287	0	0	4	ND	ND	rare	NATL2_03681	

2808	0	525	0	0	4	ND	ND	rare	NATL2_06531	
3191	0	321	0	0	4	ND	ND	rare	NATL2_16371	
3255	0	394	0	0	4	ND	ND	rare	NATL2_19941	
2221	0	186	3	0	0	rare	ND	ND	P9301_12301	
2322	0	108	1	0	2	rare	ND	average	P9312_07651	lipoprotein-like
2831	0	104	1	0	2	rare	ND	average	NATL2_07461	
3033	0	162	1	0	2	rare	ND	average	NATL2_12751	
2186	0	267	0	0	3	ND	ND	rare	P9301_06881	leader peptidase (prepilin peptidase) / N-methyltransferase
2813	0	273	0	0	3	ND	ND	rare	NATL2_06861	
3137	0	351	0	0	3	ND	ND	rare	NATL2_14941	
3252	0	251	0	0	3	ND	ND	rare	NATL2_19761	
1845	0	322	2	0	0	rare	ND	ND	P9301_14141	
1994	0	185	2	0	0	rare	ND	ND	A9601_00551	
2120	0	145	2	0	0	rare	ND	ND	A9601_14341	
2596	0	155	2	0	0	rare	ND	ND	P9515_04121	
2643	0	170	2	0	0	rare	ND	ND	P9515_12811	
2885	0	110	2	0	0	rare	ND	ND	NATL2_10351	
853	0	1185	0	0	2	ND	ND	rare	SS120_06121	Superfamily I DNA/RNA helicase
2383	0	239	0	0	2	ND	ND	rare	P9312_13021	hypothetical protein
2398	0	439	0	0	2	ND	ND	rare	P9312_14271	hypothetical protein
2724	0	196	0	0	2	ND	ND	rare	NATL2_00611	
2743	0	179	0	0	2	ND	ND	rare	NATL2_00861	
2773	0	198	0	0	2	ND	ND	rare	NATL2_03671	
2781	0	262	0	0	2	ND	ND	rare	NATL2_03831	
2868	0	622	0	0	2	ND	ND	rare	NATL2_09851	
2870	0	202	0	0	2	ND	ND	rare	NATL2_09901	
2948	0	195	0	0	2	ND	ND	rare	NATL2_11601	
3143	0	362	0	0	2	ND	ND	rare	NATL2_15001	
3146	0	1543	0	0	2	ND	ND	rare	NATL2_15141	
3159	0	206	0	0	2	ND	ND	rare	NATL2_15581	
3242	0	612	0	0	2	ND	ND	rare	NATL2_19101	
3258	0	459	0	0	2	ND	ND	rare	NATL2_20411	
3288	0	708	0	0	2	ND	ND	rare	NATL2_21521	
3435	0	430	0	0	2	ND	ND	rare	NATL1_18201	
3551	0	638	0	0	2	ND	ND	rare	P9313_01411	FAD binding site:Fumarate reductase/succinate dehydrogenase f...
3954	0	434	0	0	2	ND	ND	rare	P9313_01541	Bacterial extracellular solute-binding protein, family 1
4555	0	365	0	0	2	ND	ND	rare	P9313_10051	Serine/threonine specific protein phosphatase:Purple acid pho...
170	0	78	1	0	0	rare	ND	ND	A9601_11021	Conserved hypothetical protein
650	0	198	1	0	0	rare	ND	ND	P9301_06861	
844	0	294	1	0	0	rare	ND	ND	A9601_04311	
1834	0	574	1	0	0	rare	ND	ND	A9601_14011	
1840	0	370	1	0	0	rare	ND	ND	P9301_14071	UDP-N-acetylglucosamine 2-epimerase
1864	0	339	1	0	0	rare	ND	ND	P9515_18311	
2055	0	86	1	0	0	rare	ND	ND	A9601_12331	
2126	0	204	1	0	0	rare	ND	ND	A9601_14411	putative phosphoheptose isomerase
2132	0	229	1	0	0	rare	ND	ND	A9601_14471	2OG-Fe(II) dioxygenase superfamily protein
2142	0	394	1	0	0	rare	ND	ND	A9601_14591	UDP-galactopyranose mutase
2167	0	117	1	0	0	rare	ND	ND	A9601_17321	
2244	0	133	1	0	0	rare	ND	ND	P9301_12571	
2263	0	263	1	0	0	rare	ND	ND	P9301_14021	2,4-dihydroxyhept-2-ene-1,7-dioic acid aldolase
2298	0	170	1	0	0	rare	ND	ND	P9301_16131	
2343	0	119	1	0	0	rare	ND	ND	P9312_12281	hypothetical protein
2358	0	140	1	0	0	rare	ND	ND	P9312_12441	hypothetical protein
2368	0	210	1	0	0	rare	ND	ND	P9312_12581	sulfotransferase family protein
2379	0	109	1	0	0	rare	ND	ND	P9312_12831	hypothetical protein
2401	0	456	1	0	0	rare	ND	ND	P9312_14371	hypothetical protein
2528	0	386	1	0	0	rare	ND	ND	PMED4_13781	Glycosyl transferases group 1
2563	0	133	1	0	0	rare	ND	ND	PMED4_16041	HNH endonuclease:HNH nuclease
2663	0	148	1	0	0	rare	ND	ND	P9515_13871	
2672	0	475	1	0	0	rare	ND	ND	P9515_13981	
2673	0	391	1	0	0	rare	ND	ND	P9515_13991	
2677	0	234	1	0	0	rare	ND	ND	P9515_14031	
2681	0	354	1	0	0	rare	ND	ND	P9515_14071	
3061	0	88	1	0	0	rare	ND	ND	NATL2_13111	
3105	0	86	1	0	0	rare	ND	ND	NATL2_13771	
3176	0	199	1	0	0	rare	ND	ND	NATL2_16041	
3828	0	352	1	0	0	rare	ND	ND	P9211_12861	
4015	0	1676	1	0	0	rare	ND	ND	P9303_02941	
5286	0	693	1	0	0	rare	ND	ND	P9303_29441	
5396	0	1057	1	0	0	rare	ND	ND	P9313_03551	TPR repeat
5417	0	194	1	0	0	rare	ND	ND	P9313_04541	possible M protein repeat
1249	0	169	0	0	1	ND	ND	rare	P9211_12491	
1625	0	198	0	0	1	ND	ND	rare	P9515_12651	possible cytochrome c oxidase subunit III
1942	0	183	0	0	1	ND	ND	rare	P9515_13071	
1991	0	1341	0	0	1	ND	ND	rare	A9601_00521	
2206	0	283	0	0	1	ND	ND	rare	P9301_07081	
2245	0	336	0	0	1	ND	ND	rare	P9301_12581	Arsenite efflux pump ACR3 and related permeases
2438	0	185	0	0	1	ND	ND	rare	P9312_16521	hypothetical protein
2631	0	202	0	0	1	ND	ND	rare	P9515_11761	
2714	0	580	0	0	1	ND	ND	rare	NATL2_00361	
2739	0	595	0	0	1	ND	ND	rare	NATL2_00821	
2744	0	169	0	0	1	ND	ND	rare	NATL2_00911	
2772	0	155	0	0	1	ND	ND	rare	NATL2_03661	
2800	0	258	0	0	1	ND	ND	rare	NATL2_04691	
3016	0	184	0	0	1	ND	ND	rare	NATL2_12531	
3139	0	349	0	0	1	ND	ND	rare	NATL2_14961	
3144	0	354	0	0	1	ND	ND	rare	NATL2_15011	

3152	0	189	0	0	1	ND	ND	rare	NATL2_15201	
3220	0	637	0	0	1	ND	ND	rare	NATL2_17631	
3251	0	170	0	0	1	ND	ND	rare	NATL2_19611	
3264	0	1219	0	0	1	ND	ND	rare	NATL2_20521	
3278	0	270	0	0	1	ND	ND	rare	NATL2_21241	
3304	0	205	0	0	1	ND	ND	rare	P9313_21521	TPR repeat
3410	0	302	0	0	1	ND	ND	rare	NATL1_15801	
3423	0	190	0	0	1	ND	ND	rare	NATL1_16331	
3472	0	166	0	0	1	ND	ND	rare	P9211_02301	
3473	0	262	0	0	1	ND	ND	rare	P9313_22931	Creatininase
3475	0	472	0	0	1	ND	ND	rare	P9211_03411	
3507	0	1110	0	0	1	ND	ND	rare	SS120_06211	AbrB family trancriptional regulator fused to LRR containing domain
3625	0	355	0	0	1	ND	ND	rare	P9211_14531	
3671	0	333	0	0	1	ND	ND	rare	P9313_22891	conserved hypothetical
3687	0	306	0	0	1	ND	ND	rare	P9211_15411	
3703	0	379	0	0	1	ND	ND	rare	P9313_18461	Alanine dehydrogenase
3706	0	888	0	0	1	ND	ND	rare	P9313_22641	conserved hypothetical protein
3752	0	185	0	0	1	ND	ND	rare	P9211_06071	
3826	0	515	0	0	1	ND	ND	rare	P9313_24901	Dolichyl-phosphate-mannose-proteinmannosyltransferase
3830	0	317	0	0	1	ND	ND	rare	P9211_12891	
3880	0	307	0	0	1	ND	ND	rare	P9211_15351	
3899	0	542	0	0	1	ND	ND	rare	P9313_00371	SNF2 related domain:DEAD/DEAH box helicase:Helicase C-termina...
3944	0	1049	0	0	1	ND	ND	rare	P9303_01471	
3956	0	317	0	0	1	ND	ND	rare	P9313_01601	Uncharacterized conserved protein
4091	0	507	0	0	1	ND	ND	rare	P9313_19171	Uncharacterized conserved protein
4192	0	242	0	0	1	ND	ND	rare	P9313_16251	conserved hypothetical
4476	0	252	0	0	1	ND	ND	rare	P9303_12471	
4634	0	486	0	0	1	ND	ND	rare	P9313_08001	Conserved hypothetical protein
4729	0	177	0	0	1	ND	ND	rare	P9313_05881	Ferritin
4772	0	533	0	0	1	ND	ND	rare	P9313_05091	Permeases
4805	0	837	0	0	1	ND	ND	rare	P9313_04441	Predicted bile acid beta-glucosidase
4869	0	294	0	0	1	ND	ND	rare	P9303_20531	
5047	0	320	0	0	1	ND	ND	rare	P9313_23031	Predicted N-acetylglucosamine kinase
5102	0	345	0	0	1	ND	ND	rare	P9313_24291	conserved hypothetical protein
5147	0	509	0	0	1	ND	ND	rare	P9313_26691	Phage integrase
5165	0	338	0	0	1	ND	ND	rare	P9313_25591	Pseudouridine synthase
5226	0	158	0	0	1	ND	ND	rare	P9313_27091	conserved hypothetical protein
5229	0	374	0	0	1	ND	ND	rare	P9313_27121	possible Tripartite transporter component (TRAP-T family), substrate binding protein
1648	0	71	6	0	1	average	ND	average	A9601_12431	
531	0	85	0	0	5	ND	ND	average	P9312_07351	hypothetical protein
3282	0	236	0	0	5	ND	ND	average	NATL2_21281	
1541	0	48	3	0	1	average	ND	average	P9301_16711	Conserved hypothetical protein
2311	0	41	2	0	2	average	ND	average	P9312_04471	hypothetical protein
2956	0	86	0	0	4	ND	ND	average	NATL2_11571	
3119	0	91	0	0	4	ND	ND	average	NATL2_13981	
3232	0	105	0	0	4	ND	ND	average	NATL2_18481	
1608	0	34	2	0	1	average	ND	average	P9301_12291	Conserved hypothetical protein
2734	0	181	0	0	3	ND	ND	average	NATL2_00731	
2764	0	161	0	0	3	ND	ND	average	NATL2_02561	
2819	0	116	0	0	3	ND	ND	average	NATL2_07251	
2842	0	60	0	0	3	ND	ND	average	NATL2_08561	
2915	0	98	0	0	3	ND	ND	average	NATL2_10841	
2942	0	86	0	0	3	ND	ND	average	NATL2_11321	
3013	0	110	0	0	3	ND	ND	average	NATL2_12491	
3193	0	42	0	0	3	ND	ND	average	NATL2_16611	
3663	0	160	0	0	3	ND	ND	average	SS120_15391	Predicted protein
1641	0	78	2	0	0	average	ND	ND	A9601_04141	
2156	0	69	2	0	0	average	ND	ND	P9301_15931	
2164	0	56	1	0	1	average	ND	average	A9601_17001	
2236	0	82	0	0	2	ND	ND	average	P9301_12481	
2297	0	61	0	0	2	ND	ND	average	P9301_16121	
2729	0	54	0	0	2	ND	ND	average	NATL2_00681	
2811	0	84	0	0	2	ND	ND	average	NATL2_06751	
2833	0	161	0	0	2	ND	ND	average	NATL2_08051	
2843	0	125	0	0	2	ND	ND	average	NATL2_08631	
2857	0	61	1	0	1	average	ND	average	NATL2_09671	
2873	0	102	0	0	2	ND	ND	average	NATL2_10001	
2876	0	73	0	0	2	ND	ND	average	NATL2_10181	
2889	0	83	0	0	2	ND	ND	average	NATL2_10421	
2895	0	71	0	0	2	ND	ND	average	NATL2_10521	
2898	0	84	0	0	2	ND	ND	average	NATL2_10571	
2914	0	34	0	0	2	ND	ND	average	NATL2_10831	
2963	0	68	0	0	2	ND	ND	average	NATL2_11711	
2993	0	87	0	0	2	ND	ND	average	NATL2_12191	
3037	0	108	0	0	2	ND	ND	average	NATL2_12831	
3064	0	64	0	0	2	ND	ND	average	NATL2_13141	
3067	0	80	0	0	2	ND	ND	average	NATL2_13181	
3074	0	48	0	0	2	ND	ND	average	NATL2_13271	
3168	0	118	0	0	2	ND	ND	average	NATL2_15681	
3200	0	63	0	0	2	ND	ND	average	NATL2_17211	
3202	0	107	0	0	2	ND	ND	average	NATL2_17311	
3203	0	81	0	0	2	ND	ND	average	NATL2_17321	
3212	0	50	0	0	2	ND	ND	average	NATL2_17531	
3238	0	90	0	0	2	ND	ND	average	NATL2_18741	
3277	0	71	0	0	2	ND	ND	average	NATL2_21151	
47	0	47	0	0	1	ND	ND	average	NATL2_18551	

1650	0	38	0	0	1	ND	ND	average	A9601_12411	
1670	0	35	0	0	1	ND	ND	average	P9301_15791	
1997	0	51	1	0	0	average	ND	ND	P9301_00581	
2016	0	44	0	0	1	ND	ND	average	A9601_04081	
2211	0	59	0	0	1	ND	ND	average	P9301_08811	Conserved hypothetical protein
2234	0	62	0	0	1	ND	ND	average	P9301_12461	possible poly A polymerase regulatory subunit
2454	0	86	0	0	1	ND	ND	average	PMED4_03701	possible Small, acid-soluble spore proteins, a
2455	0	57	1	0	0	average	ND	ND	PMED4_03741	
2549	0	87	0	0	1	ND	ND	average	PMED4_15331	hypothetical
2610	0	63	1	0	0	average	ND	ND	P9515_06341	
2723	0	110	0	0	1	ND	ND	average	NATL2_00601	
2728	0	68	1	0	0	average	ND	ND	NATL2_00671	
2730	0	58	0	0	1	ND	ND	average	NATL2_00691	
2731	0	75	0	0	1	ND	ND	average	NATL2_00701	
2733	0	111	0	0	1	ND	ND	average	NATL2_00721	
2738	0	90	0	0	1	ND	ND	average	NATL2_00791	
2742	0	81	0	0	1	ND	ND	average	NATL2_00851	
2769	0	87	0	0	1	ND	ND	average	NATL2_03181	
2786	0	93	0	0	1	ND	ND	average	NATL2_04111	
2796	0	67	0	0	1	ND	ND	average	NATL2_04511	
2815	0	56	0	0	1	ND	ND	average	NATL2_06961	
2817	0	137	0	0	1	ND	ND	average	NATL2_07131	
2824	0	90	0	0	1	ND	ND	average	NATL2_07341	
2828	0	111	0	0	1	ND	ND	average	NATL2_07391	
2830	0	74	0	0	1	ND	ND	average	NATL2_07431	
2835	0	56	0	0	1	ND	ND	average	NATL2_08331	
2846	0	117	0	0	1	ND	ND	average	NATL2_08681	
2847	0	54	0	0	1	ND	ND	average	NATL2_08691	
2858	0	59	0	0	1	ND	ND	average	NATL2_09681	
2861	0	87	0	0	1	ND	ND	average	NATL2_09711	
2878	0	68	0	0	1	ND	ND	average	NATL2_10231	
2883	0	68	0	0	1	ND	ND	average	NATL2_10301	
2884	0	60	0	0	1	ND	ND	average	NATL2_10341	
2886	0	78	0	0	1	ND	ND	average	NATL2_10371	
2887	0	65	0	0	1	ND	ND	average	NATL2_10381	
2888	0	73	0	0	1	ND	ND	average	NATL2_10401	
2890	0	51	0	0	1	ND	ND	average	NATL2_10431	
2899	0	66	0	0	1	ND	ND	average	NATL2_10591	
2910	0	59	0	0	1	ND	ND	average	NATL2_10761	
2936	0	58	0	0	1	ND	ND	average	NATL2_11221	
2940	0	54	0	0	1	ND	ND	average	NATL2_11301	
2952	0	64	0	0	1	ND	ND	average	NATL2_11511	
2959	0	57	0	0	1	ND	ND	average	NATL2_11641	
2965	0	53	0	0	1	ND	ND	average	NATL2_11731	
2968	0	89	0	0	1	ND	ND	average	NATL2_11791	
2971	0	71	0	0	1	ND	ND	average	NATL2_11841	
2988	0	94	0	0	1	ND	ND	average	NATL2_12111	
2995	0	67	0	0	1	ND	ND	average	NATL2_12231	
2998	0	68	0	0	1	ND	ND	average	NATL2_12281	
2999	0	70	0	0	1	ND	ND	average	NATL2_12291	
3001	0	44	0	0	1	ND	ND	average	NATL2_12361	
3002	0	56	0	0	1	ND	ND	average	NATL2_12371	
3005	0	60	0	0	1	ND	ND	average	NATL2_12411	
3006	0	115	0	0	1	ND	ND	average	NATL2_12421	
3014	0	33	0	0	1	ND	ND	average	NATL2_12501	
3020	0	72	1	0	0	average	ND	ND	NATL2_12571	
3032	0	121	0	0	1	ND	ND	average	NATL2_12721	
3042	0	38	0	0	1	ND	ND	average	NATL2_12881	
3073	0	38	1	0	0	average	ND	ND	NATL2_13261	
3081	0	62	0	0	1	ND	ND	average	NATL2_13401	
3087	0	59	0	0	1	ND	ND	average	NATL2_13471	
3088	0	53	0	0	1	ND	ND	average	NATL2_13481	
3101	0	55	0	0	1	ND	ND	average	NATL2_13681	
3106	0	68	0	0	1	ND	ND	average	NATL2_13801	
3114	0	65	0	0	1	ND	ND	average	NATL2_13921	
3135	0	62	0	0	1	ND	ND	average	NATL2_14901	
3153	0	45	0	0	1	ND	ND	average	NATL2_15211	
3209	0	59	0	0	1	ND	ND	average	NATL2_17481	
3213	0	70	0	0	1	ND	ND	average	NATL2_17541	
3216	0	74	0	0	1	ND	ND	average	NATL2_17591	
3222	0	75	0	0	1	ND	ND	average	NATL2_17651	
3223	0	94	0	0	1	ND	ND	average	NATL2_17661	
3224	0	74	0	0	1	ND	ND	average	NATL2_17671	
3233	0	47	0	0	1	ND	ND	average	NATL2_18511	
3243	0	56	0	0	1	ND	ND	average	NATL2_19111	
3244	0	76	0	0	1	ND	ND	average	NATL2_19121	
3254	0	143	0	0	1	ND	ND	average	NATL2_19821	
3273	0	55	0	0	1	ND	ND	average	NATL2_20891	
3279	0	65	0	0	1	ND	ND	average	NATL2_21251	
3306	0	54	0	0	1	ND	ND	average	NATL1_00791	
3307	0	39	0	0	1	ND	ND	average	NATL1_00811	
3454	0	73	0	0	1	ND	ND	average	SS120_00311	Predicted protein
3520	0	41	0	0	1	ND	ND	average	P9211_06051	
3539	0	65	0	0	1	ND	ND	average	SS120_06971	Predicted protein
4454	0	78	0	0	1	ND	ND	average	P9313_10621	Conserved hypothetical protein
4918	0	119	0	0	1	ND	ND	average	P9313_02471	possible Tropomyosin

Suppl. Table 2. Genes with significantly different multiplicity per genome at different depths. Hypothesis test results ($p<0.001$) are listed for each pairwise depth comparison; 1 indicates we can reject the null hypothesis of equal multiplicities, while 0 indicates a failure to reject. The majority of these genes are hypothetical or of unknown function; the few with annotated functions code for single steps in disparate biochemical pathways.

			Multiplicity estimate			Significantly different?				
cluster	core	protLength	25mult	75mult	125mult	25v75	25v125	75v125	locus	description
13	0	89	4.56	5.05	1.89	0	1	1	P9301_13071	possible Phosphatidylinositol-specific phosphotase
41	0	514	0.21	0.29	0.10	0	0	1	P9301_14321	
78	0	363	0.89	1.12	0.25	0	1	1	P9301_11941	putative glycerol dehydrogenase
131	0	77	0.36	1.35	0.17	1	0	1	P9301_11401	hypothetical
132	0	91	0.61	1.42	0.71	1	0	0	P9301_11391	DNA gyrase/topoisomerase IV, subunit-like
135	0	357	0.84	1.31	0.91	1	0	0	P9301_11361	G-protein beta WD-40 repeats
160	0	59	0.47	0.85	0.00	0	0	1	P9301_11111	Conserved hypothetical protein
194	0	106	0.89	0.78	0.12	0	1	1	P9301_10771	hypothetical
201	0	198	0.76	0.65	0.26	0	1	0	P9301_10701	ATP/GTP-binding site motif A (P-loop)
210	0	133	2.17	1.05	1.66	1	0	0	P9301_10601	possible ATP synthase protein 8
225	0	72	0.23	1.40	0.54	1	0	0	P9301_10441	hypothetical
237	0	122	0.77	0.62	0.00	0	1	1	P9301_10321	possible Nucleoside diphosphate kinase
345	0	110	0.46	0.39	0.00	0	0	1	P9301_09241	conserved hypothetical protein
384	0	384	0.56	1.00	0.94	1	0	0	P9301_08851	putative urea ABC transporter
485	0	828	1.14	0.90	0.67	0	1	0	P9301_07821	DEAD/DEAH box helicase:Helicase C-terminal domain
515	0	60	0.74	0.72	0.00	0	0	1	P9301_07501	Conserved hypothetical protein
520	0	78	0.57	1.61	0.83	1	0	0	P9301_07441	Conserved hypothetical protein
589	0	85	0.98	1.48	0.15	0	0	1	P9301_06561	conserved hypothetical protein
897	0	125	1.07	1.41	0.52	0	0	1	P9301_03981	possible Phosphoenolpyruvate carboxykinase
901	0	259	0.92	0.92	0.30	0	1	1	P9301_03941	Phosphomethylpyrimidine kinase
925	0	88	1.26	0.90	0.29	0	1	0	P9301_03761	mttA/Hcf106 family
932	0	67	0.66	1.45	0.39	0	0	1	P9301_03691	Conserved hypothetical protein
934	0	68	0.90	0.74	0.00	0	1	1	P9301_03671	Conserved hypothetical protein
960	0	120	1.07	1.17	0.32	0	1	1	P9301_03411	possible ferredoxin
1025	0	50	0.56	0.79	0.00	0	0	1	P9301_02761	Photosystem II reaction centre N protein (psbN)
1040	0	111	0.95	0.97	0.12	0	1	1	P9301_02611	
1065	0	128	0.56	1.24	0.71	1	0	0	P9301_02371	Glyoxalase/Bleomycin resistance protein/Dioxygenase superfamily
1164	0	89	0.87	0.73	0.15	0	0	1	P9301_01361	conserved hypothetical protein
1183	0	454	0.85	1.23	0.74	1	0	1	P9301_01171	possible Fe-S oxidoreductase
1227	0	61	0.82	1.06	0.21	0	0	1	P9301_00731	Photosystem II protein X PsbX
1375	0	143	1.05	1.31	0.54	0	0	1	P9301_18321	possible transcription regulator
1611	0	99	1.52	1.60	0.39	0	1	1	P9301_13001	protein family PM-11
1679	0	111	1.30	0.84	0.35	0	1	0	P9301_15681	Predicted membrane protein
1911	0	110	1.37	0.92	0.24	0	1	1	P9301_13671	Macrophage migration inhibitory factor family
1915	0	79	0.42	0.55	0.00	0	0	1	P9301_13631	peptidase family M20/M25/M40-like
1958	0	60	0.93	1.32	0.22	0	0	1	P9301_13061	Conserved hypothetical protein
1968	0	62	0.99	0.46	0.00	0	1	0	A9601_12541	
1972	0	44	1.01	1.55	0.29	0	0	1	P9301_12921	Conserved hypothetical protein
1980	0	38	1.90	1.70	0.00	0	1	1	P9301_12781	
1988	0	84	0.73	0.86	0.15	0	0	1	P9301_12661	
1990	0	76	0.80	0.71	0.00	0	1	1	P9301_00291	possible Transcription factor TFIID (or TATA-binding)
2007	0	36	1.08	1.40	0.00	0	0	1	P9301_03711	Conserved hypothetical protein
2011	0	277	0.12	0.21	0.00	0	0	1	A9601_03891	Glycosyl transferase family 11
2044	0	88	0.51	0.78	0.00	0	0	1	P9301_10551	hypothetical
2073	0	43	0.91	0.92	0.00	0	0	1	P9301_12911	Conserved hypothetical protein
2149	0	205	0.73	0.91	0.19	0	1	1	P9301_15781	conserved hypothetical protein
2252	0	433	0.35	0.32	0.09	0	1	1	P9301_13111	
2255	0	362	0.37	0.14	0.04	0	1	0	P9301_13141	
2257	0	326	0.39	0.21	0.12	0	1	0	P9301_13161	
3128	0	695	0.00	0.01	0.21	0	1	0	NATL2_14561	

Suppl. Table 3. Core genes that are observed less frequently than expected in the eMIT9312-assigned pool of reads. This apparent discrepancy is likely due to factors such as very high sequence conservation or intragenic recombination that make the ecotype signal more difficult to discern.

cluster	core	protLength	25DNA	75DNA	125DNA	25	75	125	locus	desc
1297	1	813	124	179	28	0	0	-1	P9301_00041	DNA gyrase/topoisomerase IV, subunit A
1294	1	244	27	39	9	0	-1	0	P9301_00071	Uncharacterized conserved protein
1292	1	439	62	84	17	0	-1	0	P9301_00091	signal recognition particle docking protein FtsY
1289	1	217	11	25	9	-1	-1	0	P9301_00121	RNA-binding region RNP-1 (RNA recognition motif)
1280	1	448	47	120	23	-1	0	0	P9301_00211	Probable UDP-N-acetylmuramate-alanine ligase
1279	1	340	36	78	16	-1	0	0	P9301_00221	Glyceraldehyde 3-phosphate dehydrogenase(NADP+)(phosphorylating)
1273	1	292	59	47	15	0	-1	0	P9301_00281	Nucleoside-diphosphate-sugar epimerases
1271	1	150	12	13	4	0	-1	0	P9301_00311	type II secretion system protein-like
1191	1	113	9	8	4	0	-1	0	P9301_01091	possible Zinc finger, C3HC4 type (RING finger)
1172	1	69	2	6	4	-1	0	0	P9301_01281	conserved hypothetical protein
1149	1	472	57	68	15	0	-1	0	P9301_01541	fatty acid/phospholipid synthesis protein PlsX
1132	1	392	31	77	10	-1	0	0	P9301_01711	Putative sugar-phosphate nucleotidyl transferase
1082	1	131	15	11	5	0	-1	0	P9301_02211	50S ribosomal protein L7/L12
1067	1	550	61	109	21	-1	-1	0	P9301_02351	putative sulfate transporter
1043	1	241	29	38	6	0	-1	0	P9301_02581	Beta-carotene hydroxylase
1031	1	321	32	35	7	-1	-1	0	P9301_02701	conserved hypothetical protein
1030	1	145	1	2	0	-1	-1	0	P9301_02711	conserved hypothetical protein
1022	1	212	6	6	1	-1	-1	-1	P9301_02781	3-isopropylmalate dehydratase small subunit
1021	1	467	55	107	12	0	0	-1	P9301_02791	3-isopropylmalate dehydratase large subunit
1014	1	486	21	30	7	-1	-1	-1	P9301_02861	Ammonium transporter family
1011	1	517	71	95	23	0	-1	0	P9301_02891	AICARFT/IMPCHase bienzyme:Methylglyoxal synthase-like domain
1008	1	378	33	51	10	-1	-1	0	P9301_02921	two-component sensor histidine kinase
1007	1	170	7	11	1	-1	-1	-1	P9301_02931	Cobalamin-5-phosphate synthase
1006	1	372	40	47	16	-1	-1	0	P9301_02941	tRNA-guanine transglycosylase
996	1	494	22	44	9	-1	-1	-1	P9301_03041	Retinal pigment epithelial membrane protein
995	1	201	7	7	2	-1	-1	-1	P9301_03051	Imidazoleglycerol-phosphate dehydratase
994	1	260	3	6	1	-1	-1	-1	P9301_03061	enoyl-[acyl-carrier-protein] reductase
993	1	197	4	9	1	-1	-1	-1	P9301_03071	conserved hypothetical protein
992	1	401	15	17	4	-1	-1	-1	P9301_03081	putative pleiotropic regulatory protein
990	1	187	1	1	0	-1	-1	0	P9301_03101	NUDIX hydrolase
989	1	188	6	5	3	-1	-1	0	P9301_03111	possible 2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase
988	1	711	92	137	29	0	-1	0	P9301_03121	Protoporphyrin IX Magnesium chelatase, ChlD subunit
945	1	253	21	47	14	-1	0	0	P9301_03571	possible LysM domain
831	1	88	3	9	3	-1	-1	0	P9301_04131	conserved hypothetical protein
813	1	100	5	9	1	-1	-1	0	P9301_04311	Thioredoxin family protein
812	1	511	29	35	6	-1	-1	-1	P9301_04321	UDP-N-acetylmuramyl-tripeptide synthetase
810	1	390	41	61	10	-1	-1	0	P9301_04341	putative L-cysteine/cystine lyase
785	1	200	9	17	0	-1	-1	0	P9301_04611	hypothetical
752	1	184	11	19	8	-1	-1	0	P9301_04941	Photosystem I PsaF protein (subunit III)
726	1	501	49	94	14	-1	-1	-1	P9301_05201	Carboxypeptidase Taq (M32) metallopeptidase
676	1	523	36	51	20	-1	-1	0	P9301_05701	Light-independent protochlorophyllide reductase subunit B
675	1	418	10	22	10	-1	-1	-1	P9301_05711	Light-independent protochlorophyllide reductase subunit N
670	1	471	32	71	13	-1	-1	-1	P9301_05761	Ribulose bisphosphate carboxylase, large chain
669	1	113	5	8	2	-1	-1	0	P9301_05771	Ribulose bisphosphate carboxylase, small chain
668	1	764	29	64	11	-1	-1	-1	P9301_05781	carboxysome shell protein CsoS2
667	1	509	55	107	24	-1	0	0	P9301_05791	carboxysome shell protein CsoS3
625	1	513	55	95	16	-1	-1	0	P9301_06201	putative NADH Dehydrogenase (complex I) subunit (chain 4)
619	1	444	16	26	2	-1	-1	-1	P9301_06261	conserved hypothetical protein
618	1	302	22	42	13	-1	-1	0	P9301_06271	ABC transporter, substrate binding protein, possibly Mn
599	1	240	37	32	11	0	-1	0	P9301_06461	conserved hypothetical protein
590	1	462	77	89	20	0	-1	0	P9301_06551	possible sodium:solute symporter, ESS family
545	1	376	40	86	26	-1	0	0	P9301_07241	putative phosphonate ABC transporter
494	1	351	36	60	20	-1	-1	0	P9301_07711	putative potassium channel, VIC family
491	1	315	26	63	9	-1	0	0	P9301_07751	putative phosphate ABC transporter
490	1	297	10	32	6	-1	-1	0	P9301_07761	putative phosphate ABC transporter

489	1	269	20	39	10	-1	-1	0	P9301_07771	putative phosphate ABC transporter, ATP binding subunit
464	1	637	59	139	36	-1	0	0	P9301_08031	FtsH ATP-dependent protease homolog
454	1	234	43	32	14	0	-1	0	P9301_08131	30S ribosomal protein S2
452	1	383	39	56	7	-1	-1	-1	P9301_08151	possible Adenylate cyclase
447	1	446	36	59	13	-1	-1	0	P9301_08201	Aromatic-ring hydroxylase (flavoprotein monooxygenase)
412	1	182	10	18	7	-1	-1	0	P9301_08551	conserved hypothetical protein
367	1	266	26	29	8	0	-1	0	P9301_09021	Bacitracin resistance protein BacA
361	1	379	49	86	6	0	0	-1	P9301_09071	carbamoyl-phosphate synthase small chain
359	1	334	23	44	7	-1	-1	-1	P9301_09101	RNA methyltransferase TrmH, group 3
355	1	149	9	30	6	-1	0	0	P9301_09141	conserved hypothetical protein
328	1	461	52	113	24	-1	0	0	P9301_09421	Pyridoxal-dependent decarboxylase family protein
314	1	111	10	12	1	0	-1	0	P9301_09541	conserved hypothetical protein
313	1	100	1	11	1	-1	-1	0	P9301_09551	conserved hypothetical protein
246	1	148	12	15	5	0	-1	0	P9301_10221	possible Uncharacterized secreted proteins, Ya
197	1	134	15	17	12	0	-1	0	P9301_10741	conserved hypotheticcal protein
138	1	256	10	25	2	-1	-1	-1	P9301_11331	ABC transporter, ATP binding domain, possibly Mn transport
103	1	865	135	180	49	0	-1	0	P9301_11691	DNA gyrase/topoisomerase IV, subunit A
83	1	164	18	16	7	0	-1	0	P9301_11891	possible Helix-turn-helix
82	1	551	93	116	16	0	0	-1	P9301_11901	possible kinase
77	1	841	95	112	26	-1	-1	-1	P9301_11951	ClpC
75	1	457	61	87	20	0	-1	0	P9301_11971	Diaminopimelate decarboxylase
73	1	267	26	64	15	-1	0	0	P9301_11991	Undecaprenyl pyrophosphate synthetase (UPPS)
63	1	567	97	157	14	0	0	-1	P9301_12091	possible exodeoxyribonuclease V 67 kD polypeptide
61	1	157	22	22	3	0	-1	0	P9301_12111	Putative protein-S-isoprenylcysteine methyltransferase
57	1	204	22	32	8	0	-1	0	P9301_12151	Phospholipid and glycerol acyltransferase (from 'motifs_6.msf')
49	1	470	40	61	13	-1	-1	-1	P9301_12231	NAD binding site:Glucose inhibited division protein A family
1986	1	240	2	2	0	-1	-1	0	P9301_12681	Pseudouridine synthase, Rsu
1984	1	212	13	39	11	-1	0	0	P9301_12701	Conserved hypothetical protein
1926	1	460	44	69	20	-1	-1	0	P9301_13501	Photosystem II PsbC protein (CP43)
1924	1	509	52	92	22	-1	-1	0	P9301_13521	Cobyric acid synthase CobB
1923	1	84	1	2	1	-1	-1	0	P9301_13531	conserved hypothetical protein
1917	1	340	25	46	12	-1	-1	0	P9301_13611	putative iron ABC transporter, substrate binding protein
1913	1	268	45	42	11	0	-1	0	P9301_13651	Glycyl-tRNA synthetase alpha subunit
1912	1	128	1	4	0	-1	-1	0	P9301_13661	conserved hypothetical protein
1909	1	212	4	8	1	-1	-1	-1	P9301_13691	conserved hypothetical protein
1894	1	192	30	27	7	0	-1	0	P9301_13831	conserved hypothetical protein
1888	1	425	57	70	32	0	-1	0	P9301_13891	Seryl-tRNA synthetase
1887	1	359	40	65	13	0	-1	0	P9301_13901	Predicted membrane-associated Zn-dependent proteases 1
1885	1	721	26	56	10	-1	-1	-1	P9301_13921	polyribonucleotide nucleotidyltransferase
1884	1	300	0	1	0	0	-1	0	P9301_13931	CysQ protein homolog
1883	1	288	1	1	0	-1	-1	0	P9301_13941	putative tetrapyrrole methylase family protein
1879	1	471	0	1	0	0	-1	0	P9301_13981	UDP-glucose 6-dehydrogenase
1820	1	344	0	1	0	0	-1	0	P9301_13991	putative nucleotide sugar epimerase
1803	1	205	0	2	0	0	-1	0	P9301_14471	SOS function regulatory protein, LexA repressor
1802	1	308	4	1	1	-1	-1	-1	P9301_14481	Ornithine carbamoyltransferase
1801	1	620	32	32	7	-1	-1	-1	P9301_14491	cell division protein FtsH3
1800	1	364	43	67	15	0	-1	0	P9301_14501	putative Diaminohydroxyphosphoribosylaminopyrimidine deaminase and 5-amino-6-(5-phosphoribosylamino)uracil reductase
1799	1	166	9	26	5	-1	0	0	P9301_14511	conserved hypothetical protein
1796	1	302	49	48	14	0	-1	0	P9301_14541	predicted sugar kinase
1793	1	177	24	24	12	0	-1	0	P9301_14571	conserved hypothetical protein
1791	1	351	42	60	20	0	-1	0	P9301_14591	Thiamine monophosphate synthase (TMP)
1776	1	492	66	93	12	0	-1	-1	P9301_14711	signal recognition particle protein (SRP54)
1773	1	313	32	61	12	-1	0	0	P9301_14741	Type II alternative RNA polymerase sigma factor, sigma-70 family
1762	1	456	62	88	22	0	-1	0	P9301_14851	UDP-N-glucosamine 1-carboxyvinyltransferase
1753	1	241	30	29	5	0	-1	0	P9301_14941	Cell division septal protein
1723	1	712	85	154	24	-1	0	-1	P9301_15241	chloroplast outer envelope membrane protein homolog
1722	1	245	12	29	5	-1	-1	0	P9301_15251	SAICAR synthetase
1720	1	688	90	113	33	0	-1	0	P9301_15271	two-component sensor histidine kinase
1719	1	509	48	122	27	-1	0	0	P9301_15281	possible circadian clock protein KaiC

1713	1	391	71	57	15	0	-1	0	P9301_15341	possible sporulation protein SpoIID
1711	1	186	16	18	5	0	-1	0	P9301_15361	Pentapeptide repeats
1706	1	162	19	21	8	0	-1	0	P9301_15411	conserved hypothetical protein
1667	1	401	6	10	2	-1	-1	-1	P9301_15831	possible permease
2153	1	194	8	13	3	-1	-1	0	P9301_15841	conserved hypothetical protein
1591	1	522	61	111	20	-1	0	0	P9301_16221	putative DnaK-type molecular chaperone (HSP70 family)
1589	1	540	97	94	31	0	-1	0	P9301_16241	Phosphoglycerate mutase, co-factor-independent (iPGM)
1582	1	158	24	16	5	0	-1	0	P9301_16301	conserved hypothetical protein
1576	1	565	88	109	31	0	-1	0	P9301_16361	Carbon-nitrogen hydrolase:NAD+ synthase
1572	1	316	38	31	11	0	-1	0	P9301_16401	ATP synthase gamma subunit
1571	1	505	55	87	24	-1	-1	0	P9301_16411	ATP synthase F1, alpha subunit
1569	1	170	24	24	8	0	-1	0	P9301_16431	ATP synthase B/B' CF(0)
1559	1	112	17	8	2	0	-1	0	P9301_16531	Nitrogen regulatory protein P-II
1538	1	1366	215	289	82	0	-1	0	P9301_16741	RNA polymerase beta prime subunit
1536	1	1097	133	223	56	-1	-1	0	P9301_16761	RNA polymerase beta subunit
1529	1	467	56	88	20	0	-1	0	P9301_16831	N utilization substance protein A
1517	1	310	46	54	12	0	-1	0	P9301_16961	SAM (and some other nucleotide) binding motif
1513	1	691	124	146	22	0	0	-1	P9301_17001	Elongation factor G
1503	1	199	9	17	5	-1	-1	0	P9301_17111	Photosystem I PsaL protein (subunit XI)
1501	1	126	6	7	0	-1	-1	0	P9301_17131	possible Annexin
1499	1	742	69	118	23	-1	-1	-1	P9301_17151	Photosystem I PsaB protein
1498	1	767	37	65	24	-1	-1	-1	P9301_17161	Photosystem I PsaA protein
1482	1	130	9	14	3	0	-1	0	P9301_17281	30S ribosomal protein S11
1477	1	206	20	30	6	0	-1	0	P9301_17331	30S ribosomal protein S5
1450	1	121	13	8	2	0	-1	0	P9301_17591	conserved hypothetical protein
1449	1	381	60	64	19	0	-1	0	P9301_17601	molybdopterin biosynthesis protein
1417	1	455	50	87	33	-1	-1	0	P9301_17921	putative Na+/H+ antiporter, CPA2 family
1391	1	472	47	88	18	-1	-1	0	P9301_18171	putative adenosylhomocysteinase
1387	1	344	49	61	16	0	-1	0	P9301_18211	Type II alternative RNA polymerase sigma factor, sigma-70 family
1380	1	190	15	42	7	-1	0	0	P9301_18271	Translation Initiation factor 3
1373	1	65	3	1	0	0	-1	0	P9301_18341	possible Photosystem II reaction center Z protein (PsbZ)
1308	1	358	42	61	13	0	-1	0	P9301_19001	Putativephospho-N-acetylmuramoyl-pentapeptide-transferase
1305	1	559	96	122	15	0	0	-1	P9301_19031	DNA REPAIR PROTEIN RECN, ABC transporter

Suppl. Table 4. Differentially expressed genes between samples ($p<0.01$). The three samples were sampled from different depths and at different times of day: 25m at 22:00, 75m at 03:30, 125m at 08:00. Shown are the number of cDNA reads detected for each gene cluster and the Z score for each pairwise depth comparison, if the Z score was significant at the 99% confidence level.

			# of reads			Z score				
clusterID	core?	protLength	25m	75m	125m	25v75	25v125	75v125	locus	description
1289	1	217	5	10	2			2.588	P9301_00121	RNA-binding region RNP-1 (RNA recognition motif)
1279	1	340	1	7	12		-2.908		P9301_00221	Glyceraldehyde 3-phosphate dehydrogenase(NADP+)(phosphorylating)
1266	1	528	9	0	1	3.372	3.053		P9301_00361	Glutamine amidotransferase class-I:GMP synthase
1226	1	309	11	8	2		3.064		P9301_00741	conserved hypothetical protein
1201	1	112	5	0	0		2.639		P9301_00991	conserved hypothetical protein
1183	0	454	1	6	0			2.702	P9301_01171	possible Fe-S oxidoreductase
1142	1	139	1	9	33		-5.307	-3.748	P9301_01611	RNA-binding region RNP-1 (RNA recognition motif)
1141	1	302	1	3	12		-2.908		P9301_01621	Squalene and phytoene synthases
1135	1	673	6	0	1	2.753			P9301_01681	putative NADH Dehydrogenase (complex I) subunit (chain 5)
1080	1	175	8	19	1		2.824	4.460	P9301_02221	50S ribosomal protein L10
1061	0	72	5	1	0		2.639		P9301_02411	Conserved hypothetical protein
1051	1	279	6	10	28		-3.458	-2.918	P9301_02501	Photosystem II manganese-stabilizing protein
1050	1	418	1	6	0			2.702	P9301_02511	putative p-pantothenate cysteine ligase and p-pantothenenoylcysteine decarboxylase
1019	1	423	7	0	1	2.974	2.578		P9301_02811	Serine hydroxymethyltransferase (SHMT)
1014	1	486	71	95	6		8.930	9.872	P9301_02861	Ammonium transporter family
1010	1	205	13	7	1		3.844		P9301_02901	probable esterase
1009	1	122	21	24	7		3.404	3.460	P9301_02911	conserved hypothetical protein
1004	1	334	5	9	20		-2.710		P9301_02961	probable oxidoreductase
963	1	159	9	2	2		2.618		P9301_03381	Predicted transcriptional regulator, consists of a Zn-ribbon and ATP-cone domains
961	1	507	4	41	68	-5.371	-7.263		P9301_03401	Photosystem II PsbB protein (CP47)
952	1	428	8	1	2	2.669			P9301_03501	carboxyl-terminal protease
829	1	191	16	0	1	4.499	4.348		P9301_04151	Uncharacterized protein conserved in bacteria
828	1	270	7	1	0		3.123		P9301_04161	Delta 1-pyrroline-5-carboxylate reductase
820	1	455	14	11	3		3.304		P9301_04241	Dihydrolipoamide acetyltransferase
818	1	328	13	4	1	2.589	3.844		P9301_04261	O-acetylserine (thiol)-lyase A
815	1	202	5	28	8	-3.835		3.776	P9301_04291	30S ribosomal protein S4
808	0	63	7	16	4			3.024	P9301_04361	Conserved hypothetical protein
777	1	541	6	0	2	2.753			P9301_04691	Cytochrome c oxidase, subunit I
2182	0	44	0	5	10		-3.073		P9301_04931	Photosystem I PsaJ protein (subunit IX)
723	1	394	27	13	4	2.778	5.035		P9301_05231	Putative principal RNA polymerase sigma factor
701	1	333	10	1	0	3.092	3.734		P9301_05451	Transaldolase
694	1	587	14	4	2	2.781	3.648		P9301_05521	Acetolactate synthase large subunit
684	1	246	1	11	2	-2.808		2.793	P9301_05621	putative GTP cyclohydrolase I
676	1	523	16	1	3	4.125	3.667		P9301_05701	Light-independent protochlorophyllide reductase subunit B
675	1	418	29	3	1	5.249	6.082		P9301_05711	Light-independent protochlorophyllide reductase subunit N
670	1	471	3	54	65	-6.630	-7.260		P9301_05761	Ribulose bisphosphate carboxylase, large chain
669	1	113	1	15	24	-3.419	-4.433		P9301_05771	Ribulose bisphosphate carboxylase, small chain
668	1	764	1	19	17	-3.942	-3.618		P9301_05781	carboxysome shell protein CsoS2
642	1	307	7	1	0		3.123		P9301_06031	Putative type II alternative sigma factor, sigma70 family
638	1	116	3	13	23		-3.693		P9301_06071	plastocyanin
614	1	111	0	4	9		-2.915		P9301_06311	conserved hypothetical protein
2208	0	33	1	19	6	-3.942		2.954	P9301_07111	
555	1	777	11	4	0		3.916		P9301_07141	ribonucleotide reductase (Class II)
488	0	48	5	0	0		2.639		P9301_07791	Conserved hypothetical protein
487	0	100	7	0	0	2.974	3.123		P9301_07801	hypothetical
454	1	234	5	11	1			3.206	P9301_08131	30S ribosomal protein S2

449	1	595	7	2	0		3.123		P9301_08181	Ferredoxin-sulfite reductase
440	1	333	5	14	4			2.667	P9301_08271	Fructose-1,6-bisphosphatase/sedoheptulose-1, 7-bis phosphatase
432	1	96	5	1	0		2.639		P9301_08351	conserved hypothetical
421	1	298	0	12	6	-3.417			P9301_08461	phosphoribulokinase
411	1	538	9	1	2	2.888	2.618		P9301_08561	conserved hypothetical protein
315	1	86	2	4	13		-2.653		P9301_09531	Photosystem I PsaK protein (subunit X)
311	1	78	2	13	8	-2.731			P9301_09571	50S ribosomal protein L28
257	1	194	9	23	7			3.315	P9301_10121	thioredoxin peroxidase
224	1	1098	8	3	0		3.339		P9301_10451	carbamoyl-phosphate synthase, large subunit
202	0	109	12	5	0		4.091		P9301_10691	Helix-hairpin-helix DNA-binding motif class 1
142	0	82	3	7	16		-2.757		P9301_11291	conserved hypothetical
133	0	32	2	45	117	-6.162	-10.285	-5.673	P9301_11381	Conserved hypothetical protein
104	1	387	5	1	0		2.639		P9301_11681	putative IMP dehydrogenase
100	1	546	9	1	2	2.888	2.618		P9301_11721	2-isopropylmalate synthase
91	1	507	7	3	1		2.578		P9301_11811	Glucose-6-phosphate dehydrogenase
90	1	321	5	11	22		-2.982		P9301_11821	ferredoxin-NADP oxidoreductase (FNR)
85	1	331	19	5	1	3.357	4.802		P9301_11871	Ribose-phosphate pyrophosphokinase
77	1	841	15	4	9	2.966			P9301_11951	ClpC
1937	1	192	6	0	1	2.753			P9301_13381	possible EF-1 guanine nucleotide exchange doma
1934	1	88	5	2	0		2.639		P9301_13411	translation initiation factor IF-1
1927	1	358	4	32	48	-4.519	-5.831		P9301_13491	Photosystem II PsbD protein (D2)
1926	1	460	6	35	45	-4.347	-5.147		P9301_13501	Photosystem II PsbC protein (CP43)
44	0	432	40	23	15	2.818	4.416		P9301_13641	porin-like
1908	0	81	5	5	0		2.639		P9301_13701	hypothetical
591	0	352	14	49	88	-4.122	-6.875	-3.191	P9301_13721	chlorophyll a/b binding light harvesting protein PcbD
1777	1	121	1	10	5	-2.635			P9301_14701	30S Ribosomal protein S16
1774	1	357	1	11	8	-2.808			P9301_14731	Pyruvate dehydrogenase E1 alpha subunit
1772	1	521	7	2	1		2.578		P9301_14751	Predicted sugar kinase fused to uncharacterized domain
1768	1	157	7	14	4			2.667	P9301_14791	putative nickel-containing superoxide dismutase precursor (NISOD)
1752	1	371	13	1	0	3.643	4.258		P9301_14951	Cell division protein FtsZ:Tubulin/FtsZ family
1745	1	337	0	7	1	-2.609			P9301_15021	putative cobalamin biosynthetic protein
1732	1	412	7	2	0		3.123		P9301_15151	Tyrosyl-tRNA synthetase
1723	1	712	7	0	2	2.974			P9301_15241	chloroplast outer envelope membrane protein homolog
1719	1	509	9	2	1		3.053		P9301_15281	possible circadian clock protein KaiC
1675	0	50	2	1	12				P9301_15731	Conserved hypothetical protein
1672	0	388	4	8	20		-3.007		P9301_15761	fatty acid desaturase, type 2
1589	1	540	17	4	4	3.312	3.534		P9301_16241	Phosphoglycerate mutase, co-factor-independent (iPGM)
1584	1	486	5	19	36	-2.680	-4.552		P9301_16281	ATP synthase F1, beta subunit
1572	1	316	0	9	7	-2.959			P9301_16401	ATP synthase gamma subunit
1556	1	461	5	0	0		2.639		P9301_16561	Fumarate lyase
1538	1	1366	32	31	12		3.947	3.329	P9301_16741	RNA polymerase beta prime subunit
1513	1	691	13	33	29	-2.651			P9301_17001	Elongation factor G
1511	1	124	4	17	6	-2.680		2.619	P9301_17021	30S ribosomal protein S12
1510	1	1523	13	7	1		3.844		P9301_17041	Ferredoxin-dependent glutamate synthase, Fd-GOGAT
1503	1	199	14	24	54		-4.377	-3.347	P9301_17111	Photosystem I PsaL protein (subunit XI)
1499	1	742	33	88	77	-4.554	-3.444		P9301_17151	Photosystem I PsaB protein
1498	1	767	22	63	112	-4.079	-7.221	-3.549	P9301_17161	Photosystem I PsaA protein
1477	1	206	1	10	3	-2.635			P9301_17331	30S ribosomal protein S5
1476	1	122	2	8	0			3.121	P9301_17341	50S ribosomal protein L18
1468	1	160	1	10	3	-2.635			P9301_17421	50S ribosomal protein L16
1466	1	243	4	9	17		-2.577		P9301_17431	30S ribosomal protein S3
1456	1	365	10	0	2	3.555	2.848		P9301_17531	RecA bacterial DNA recombination protein
1450	1	121	5	1	0		2.639		P9301_17591	conserved hypothetical protein
1440	1	140	1	10	22	-2.635	-4.216		P9301_17691	Photosystem I protein PsaD
1416	1	848	14	1	4	3.810	2.982		P9301_17931	phosphorylase
1406	1	668	7	25	15	-2.972			P9301_18031	Transketolase
1405	1	456	17	2	6	3.934	2.972		P9301_18041	ThiC family
1380	1	190	7	9	0		3.123	3.311	P9301_18271	Translation Initiation factor 3
1373	1	65	0	3	10		-3.073		P9301_18341	possible Photosystem II reaction center Z protein (PsbZ)
1350	1	82	9	0	3	3.372			P9301_18571	conserved hypothetical protein
1345	0	312	2	5	15		-2.964		P9301_18621	Fatty acid desaturase, type 1
1309	1	104	3	11	2			2.793	P9301_18991	conserved hypothetical protein

Suppl. Table 5. Flexible genes that were rare in the genomic DNA relative to core genes, but were detected in the cDNA from the same sample. These genes are present in only a subset of the population, perhaps in genomic islands or confined to one of the minor ecotypes, and their detection in the cDNA suggests relatively high per-cell expression.

			cDNA			DNA							
cluster	core	protLength	25	75	125	25	75	125	25m RelAbun	75m RelAbun	125m RelAbun	locus	description
29	0	570	1	0	0	11	27	8	rare	rare	rare	P9301_12651	Succinate dehydrogenase/fumarate reductase, flavoprotein subunit
36	0	372	0	0	1	21	22	10	rare	rare	rare	P9312_12781	adhesin-like protein
78	0	363	0	2	1	58	113	7	average	average	rare	P9301_11941	putative glycerol dehydrogenase
201	0	198	0	0	1	27	36	4	average	average	rare	P9301_10701	ATP/GTP-binding site motif A (P-loop)
202	0	109	12	5	0	5	19	3	rare	average	average	P9301_10691	Helix-hairpin-helix DNA-binding motif class 1
205	0	131	1	1	0	23	17	9	average	rare	average	P9301_10661	putative PURINE PHOSPHORIBOSYLTRANSFERASE related protein
256	0	250	1	1	0	34	40	7	average	rare	average	P9301_10131	conserved hypothetical
384	0	384	1	0	0	39	107	28	rare	average	average	P9301_08851	putative urea ABC transporter
583	0	343	3	1	0	63	61	21	average	rare	average	P9301_06621	small mechanosensitive ion channel, MscS family
875	0	619	1	0	0	17	28	5	rare	rare	rare	P9312_03241	carbamoyltransferase
890	0	89	0	1	0	1	6	0	rare	rare	average	A9601_12371	
906	0	105	0	1	0	4	9	0	rare	rare	average	A9601_12391	
1173	0	60	0	2	0	5	3	1	average	rare	average	P9301_01271	Conserved hypothetical protein
1743	0	70	0	1	0	16	5	3	average	rare	average	A9601_15171	Predicted membrane protein
1837	0	364	4	0	1	29	44	10	rare	rare	rare	P9301_14441	GDPmannose 4,6-dehydratase
1877	0	352	0	1	0	6	12	5	rare	rare	rare	P9301_14241	UDP-glucose 4-epimerase
1989	0	360	6	10	7	36	56	13	rare	rare	average	P9301_02451	Photosystem II PsbA protein (D1)
2005	0	54	3	4	7	0	2	3	average	rare	average	A9601_03531	
2012	0	255	0	0	1	3	6	1	rare	rare	rare	A9601_03901	possible Glycosyl transferase
2047	0	89	1	0	0	1	2	3	rare	rare	average	P9301_10651	
2144	0	406	2	1	0	36	46	15	rare	rare	rare	P9301_14651	putative bifuntional enzyme: tRNA methyltransferase; 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase
2160	0	173	0	2	0	0	5	4	average	rare	average	A9601_16281	
2169	0	453	0	1	0	4	1	1	rare	rare	rare	A9601_17341	
2194	0	572	0	0	2	5	7	5	rare	rare	rare	P9301_06961	Type II secretory pathway component PulD-like
2197	0	191	1	0	0	2	1	0	rare	rare	average	P9301_06991	Conserved hypothetical protein
2198	0	705	0	0	1	5	7	1	rare	rare	rare	P9301_07001	Conserved hypothetical protein
2237	0	261	1	0	0	2	3	3	rare	rare	rare	P9301_12501	
2253	0	160	0	1	0	4	8	3	rare	rare	average	P9301_13121	
2254	0	330	0	1	0	8	22	6	rare	rare	rare	P9301_13131	Predicted amidohydrolase
2258	0	91	0	1	0	1	6	0	rare	rare	average	P9301_13171	
2320	0	110	1	0	2	1	1	0	rare	rare	average	P9312_07231	hypothetical protein
2326	0	154	0	0	1	1	1	2	rare	rare	rare	P9312_09491	putative transcriptional regulator
2355	0	546	0	1	0	3	2	2	rare	rare	rare	P9312_12411	uncharacterized conserved protein containing SWIM-like Zn-finger
2393	0	347	0	1	0	0	4	1	average	rare	rare	P9312_14211	Pyruvate dehydrogenase (lipoamide)
2464	0	561	0	1	0	9	11	1	rare	rare	rare	PMED4_04031	putative cyanate ABC transporter, substrate binding protein
2558	0	130	0	1	0	3	4	2	rare	rare	average	PMED4_15911	hypothetical
2580	0	38	0	1	0	0	1	1	average	rare	average	PMED4_17481	50S Ribosomal protein L36
2761	0	386	0	0	1	1	1	2	rare	rare	rare	NATL2_02221	
2777	0	263	0	0	1	1	2	6	rare	rare	rare	NATL2_03721	
2837	0	111	0	1	0	0	1	4	average	rare	average	NATL2_08431	
2868	0	622	0	0	1	0	0	2	average	average	rare	NATL2_09851	
2938	0	81	0	1	0	0	1	2	average	rare	average	NATL2_11241	
2948	0	195	0	1	1	0	0	2	average	average	rare	NATL2_11601	
2960	0	489	0	0	1	1	3	2	rare	rare	rare	NATL2_11661	
3127	0	486	0	0	1	0	1	5	average	rare	rare	NATL2_14521	
3137	0	351	0	0	1	0	0	3	average	average	rare	NATL2_14941	
3138	0	370	0	1	2	1	1	1	rare	rare	rare	NATL2_14951	
3139	0	349	0	0	1	0	0	1	average	average	rare	NATL2_14961	
3143	0	362	0	0	1	0	0	2	average	average	rare	NATL2_15001	
3229	0	88	0	1	0	0	2	0	average	rare	average	NATL2_18061	
3230	0	583	0	0	1	0	0	7	average	average	rare	NATL2_18231	
3264	0	1219	0	0	1	0	0	1	average	average	rare	NATL2_20521	
3278	0	270	0	0	1	0	0	1	average	average	rare	NATL2_21241	

CHAPTER FIVE

Code and context: *Prochlorococcus* as a model for cross-scale biology

Maureen L. Coleman and Sallie W. Chisholm

Reprinted with permission from *Trends in Microbiology*

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

Review TRENDS in Microbiology Vol.15 No.9

Code and context: *Prochlorococcus* as a model for cross-scale biology

Maureen L. Coleman and Sallie W. Chisholm

Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

***Prochlorococcus* is a simple cyanobacterium that is abundant throughout large regions of the oceans, and has become a useful model for studying the nature and regulation of biological diversity across all scales of complexity. Recent work has revealed that environmental factors such as light, nutrients and predation influence diversity in different ways, changing our image of the structure and dynamics of the global *Prochlorococcus* population. Advances in metagenomics, transcription profiling and global ecosystem modeling promise to deliver an even greater understanding of this system and further demonstrate the power of cross-scale systems biology.**

The organism in context

Since the pioneering work on *Escherichia coli*, model organisms have been a cornerstone of molecular biology. Research on these model systems, however, rarely extends beyond the organism itself into the realm of the ecosystem in which it is embedded. Yet we know that the properties of all organisms are shaped by context, through adaptations to the heterogeneous, diverse and dynamic biotic and abiotic environments in which the organisms live. To understand living systems, we must understand their properties at all scales of organization, from the underlying pool of genetic information to the complex ecosystems that emerge from it.

New technologies are now enabling a cross-scale integrative biology, whereby a model organism can be investigated at scales from the molecular to the ecosystem, in the laboratory and *in situ*. This cross-scale biology seeks to elucidate reciprocal links between genes, metabolism, ecological interactions, environmental change and genome evolution. The requirements for such a system are demanding: it must be observable over time and space in nature and amenable to controlled growth under relevant conditions in the laboratory, and its natural habitat must be well characterized. The marine cyanobacterium *Prochlorococcus* meets these conditions (Box 1) and as such has advanced our understanding of microbial ecology and evolution.

Although the first evidence for the existence of *Prochlorococcus* dates back three decades [1,2], the isolation of *Prochlorococcus* into culture and the recognition of its significance as a globally abundant phototroph are more recent [3–6]. It is the numerically dominant photoautotroph at latitudes from 40°S to beyond 40°N in surface waters (upper 200 m) of the open ocean, making it one of the most abundant organisms on the planet [3] and an important object of study for oceanographers. The oligotrophic waters in which it thrives are characterized by steep gradients of light, temperature and nutrients, which vary not only with depth but also geographically and seasonally. The challenge is to understand how these environmental gradients dictate the distributions of ecological variants of *Prochlorococcus*, and how they drive genome evolution and diversification. Here we first describe how our understanding of light adaptation in *Prochlorococcus* has developed through studies at the cellular, genomic and population levels, both in cultured isolates and wild populations. We then apply a similar framework to understanding how nutrient availability and predators influence *Prochlorococcus* across scales.

The *Prochlorococcus* cell

Prochlorococcus is the smallest known oxygenic phototroph (0.5–0.7 μm diameter) and contains a unique photosynthetic apparatus. It is the only organism known to use divinyl chlorophyll *a* and *b* as the major light-harvesting pigments [7]. Furthermore, it harvests light with chlorophyll-binding antenna proteins (Pcb proteins) instead of the phycobilisomes used by most cyanobacteria, including its close relative, the marine *Synechococcus* [8].

Diversity viewed through the lens of light adaptation

Physiological differentiation

Prochlorococcus cells are found in abundance – typically 10^4–10^5 per ml – throughout the euphotic zone of the oceans, thriving at light intensities spanning four orders of magnitude. Isolates fall into two broad ecotypes (see Glossary) (Box 2) that are differentially adapted to high- and low-light conditions (HL and LL). HL cells can grow at

Glossary

Ecotype: a genetically and physiologically distinct population (see Box 2).
HLI: a phylogenetically distinct clade of high light-adapted *Prochlorococcus*, represented by the type strain MED4. This clade is also called 'Low B/A I' and 'eMED4'.
HLII: a phylogenetically distinct clade of high light-adapted *Prochlorococcus*, represented by the type strain MIT9312. This clade is also called 'Low B/A II' and 'eMIT9312'.
ITS: internal transcribed spacer sequence, located between the 16S and 23S rRNA genes.
LL: the low light-adapted *Prochlorococcus* clades. These clades are also referred to as 'High B/A' I–IV, and by their type strains: eNATL2A, eMIT9211, eSS120 and eMIT9313.

Corresponding author: Chisholm, S.W. (chisholm@mit.edu).
Available online 10 August 2007.

www.sciencedirect.com 0966-842X/$ – see front matter doi:10.1016/j.tim.2007.07.001

Box 1. *Prochlorococcus* as a model organism for cross-scale systems biology

Life is a system of systems, and thus can only be understood by studying it at all scales of biological organization. Each scale has properties that emerge from the interactions of its component parts, and those properties, in turn, influence the behavior of the component parts, and their evolution (Figure I).

Prochlorococcus has several features that make it a useful model system for elucidating the properties of life across scales, thereby connecting genomic information to global processes.

Genome features

- As an autotroph *Prochlorococcus* creates biomass from sunlight, CO_2 and inorganic nutrients, and thus has minimal requirements for growth.
- It has a small genome (roughly 2000 genes) and a simple regulatory system (as few as 28 predicted transcription regulators) [28,30].
- Many genome variants exist (12 cultured strains have been sequenced to date), which creates raw material for exploring evolutionary and functional diversity (genome sequences available in GenBank).

Cellular machinery and physiology

- *Prochlorococcus* is easily isolated into culture for studies of physiology, biochemistry and cellular systems biology [9,14], and its simple lifestyle – passively floating in a relatively well-mixed fluid medium – enables its natural environment to be recreated easily in the laboratory.
- Its limited requirements for growth greatly simplify physiological experiments and cellular modeling.

Population and community dynamics

- The natural habitat of *Prochlorococcus* is well studied, particularly at two long-term ocean time series sites near Hawaii and Bermuda (HOT and BATS) [68,69].
- Cells are abundant (up to 10^5 cells per ml), and can be easily identified and counted *in situ* [6].
- Wild cells can be easily sorted away from the rest of the microbial community using flow cytometry [14].
- Diverse strains of *Prochlorococcus*-infecting phage, which serve as a source of mortality and gene transfer, have been isolated [62].
- *Prochlorococcus* genes are abundant in metagenomic databases, enabling comparative genomics in the wild [42,63,64].

Global processes

- *Prochlorococcus* carries out a measurable fraction of global ocean production [3].
- Global models of its habitat are well developed, facilitating the simulation of its global dynamics and the relative fitness of different ecotypes in past, present and future oceans [65].

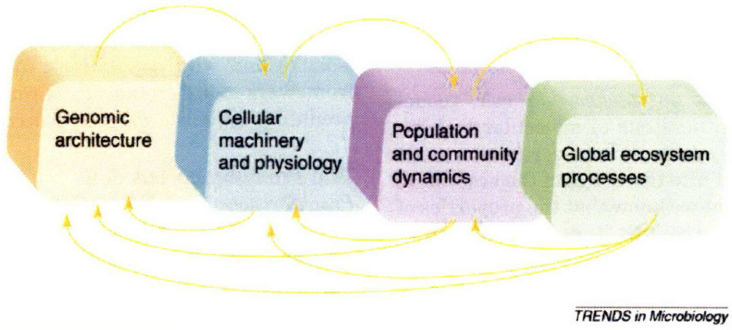


Figure I. Scales of biological organization. A comprehensive understanding of biological systems requires studying all scales of organization, and the reciprocal interactions across scales.

light intensities that are great enough to inhibit LL cells, whereas LL cells can grow at intensities that are too low to support the growth of HL strains [9]. Underlying these patterns are differences in light absorption efficiency resulting from different pigment ratios [8–10]. HL and LL cells also encode different forms of the phycobiliprotein phycoerythrin, although the role of this protein in light sensing or light harvesting is unknown [11,12]. This ecotype diversity helps to explain the broad habitat range of *Prochlorococcus*, and the ability of *Prochlorococcus* to dominate the base of the euphotic zone in the oceans. In fact, LL *Prochlorococcus* cells are arguably more efficient light harvesters than any other photosynthetic cell [13].

HL- and LL-adapted cells are distinguishable genetically at several loci [14–18], with HL cells forming a monophyletic clade divided into at least two subclades, HLI and HLII. By contrast, the LL isolates form at least four distinct lineages [16] (Figure 1a). These evolutionary reconstructions suggest that each clade might represent an ecologically distinct population, resulting from a selective sweep [16] (Box 2).

Distributions in the wild

Because these clades represent physiologically distinct cell types with different light optima for growth, we might expect their members to have different distributions along light (depth) gradients in the oceans. Indeed, initial field studies using the 16S rRNA and 16S–23S internal transcribed spacer (ITS) loci to differentiate between HL- and LL-adapted cells showed that although both coexist throughout the water column, HL cells outnumber LL cells in the surface by orders of magnitude. LL cells, although they usually have their abundance maximum in deeper waters, can be outnumbered by HL cells throughout the water column (Figure 1b) [19–24]. This asymmetry in distributions is probably the result of the asymmetry in physiological adaptation: HL cells have a greater fitness advantage, relative to LL cells, at high irradiances than do LL cells, relative to HL cells, at low irradiances [9].

As more data have accumulated we have learned that the situation is more complex. This classic pattern of HL

Box 2. What are ecotypes?

The term ecotype has been adopted in many different contexts to describe genetically and ecologically distinct units of diversity. We applied the term in describing physiological and genetic diversity in *Prochlorococcus* [14], and Cohan and others [70] have used it more formally in theoretical models of bacterial evolution. These two usages, however, are not (necessarily) equivalent.

In Cohan's stable ecotype model [70], a bacterial ecotype is a population of cells having the same ecological niche and whose divergence is constrained by periodic selection events. When one cell in a population acquires an adaptive mutation, the mutant and its descendants outcompete the rest of the population. Assuming that recombination is rare, the entire genome associated with the adaptive mutation will sweep through the population. Cohan proposes an approach for discovering ecotypes using sequence data, and under this stable ecotype model, sequence clusters will correspond one-to-one with ecotypes. One can then test whether these sequence-based ecotypes are indeed ecologically distinct.

Prochlorococcus ecotypes have been defined empirically using a combination of phenotypic and genetic data. High-light-adapted and low-light-adapted ecotypes were defined as coexisting populations with distinct photophysiology and phylogenetically distinct rRNA markers [14]. Thus, in contrast to the Cohan approach, the sequence clusters called ecotypes were chosen to reflect their known physiology. Within the HL ecotype, two ecologically distinct subclusters have been identified (see main text). Within the more divergent LL ecotype, several lineages are apparent. Although the term 'ecotype' has been applied to these LL lineages (i.e. eNATL2A, eSS120, eMIT9211 and eMIT9313 [19,20,26]), they are not well-resolved phylogenetically and their ecological differentiation is unclear.

Can we reconcile these two usages of the term 'ecotype'? If Cohan's stable ecotype model applies to *Prochlorococcus*, we should be able to delineate ecologically distinct populations as sequence clusters using a multilocus sequence analysis (MLSA) approach. In the absence of MLSA data, recent evidence hints that ecologically distinct lineages, distinguished by nutrient physiology, might exist within the named HL and LL *Prochlorococcus* clusters [34]. Several alternatives to the stable ecotype model predict multiple ecotypes within a sequence cluster, for instance when the recombination rate is high, when new traits are easily gained and lost by horizontal gene transfer, or when many ephemeral 'nano-niches' exist in the environment [70]. Testing these models in the *Prochlorococcus* system is an important step towards understanding the ecology and evolution of open-ocean bacterioplankton and developing a robust vocabulary to describe it.

cells dominating surface waters, and LL cells having an abundance maximum in deep waters, emerges when the water column has been stratified for some time. This physically isolates the surface and deep waters, enabling the HL and LL cells to photoacclimate and growth differences to manifest themselves in the population structure. When the water column is physically well mixed, however, LL cells can be as abundant as HL cells at the surface [25].

The two subclades of HL cells, HLI and HLII, display distinct distributions along oceanic gradients. HLII cells, represented by the type strain MIT9312, tend to be most abundant in warmer, highly stratified waters [19,20,23–27] and seem to be the most abundant cells on a global scale, whereas HLI cells (type strain MED4) dominate in cooler, weakly stratified waters [25,26] including the high-latitude cold-water limits of the *Prochlorococcus* range [26]. Consistent with these observations, laboratory studies have shown that HLI cells can grow at lower temperatures than HLII cells [26,27].

Among the LL-adapted clades, the forces shaping patterns of global distribution are less clear, partly because the diversity of LL lineages is not well represented in culture or in sequence databases [19,20]. When LL cells are found near the surface, they typically belong to one clade, represented by the cultured NATL2A strain, whereas another clade, represented by strain MIT9313, tends to localize deeper in the euphotic zone [19,20,23,26,27]. Although cultured isolates representing these two LL clades have similar light optima for growth, they might also possess different photoprotective mechanisms that would enable survival at elevated light intensity during mixing events [27]. Interestingly, this 'moderate' position of the NATL2A clade in the water column corresponds to its phylogenetic position as the most closely related LL clade to the HL clades [16].

Insights from whole genomes

The differentiation of HL and LL ecotypes is further revealed by whole-genome comparisons [28–30]. HL strain MED4, for example, has only one gene encoding the light-harvesting antenna protein Pcb, whereas LL strain MIT9313 has two, and LL strain SS120 – which grows at the lowest light intensity of these three [9] – has eight [31] (Figure 2). By contrast, MED4 has more genes encoding high-light-inducible proteins (HLIPs), thought to protect the cell from excess excitation energy [32], than do most LL strains (Figure 2). MED4 and other HL strains also encode photolyase, which repairs UV-induced DNA lesions, whereas the photolyase gene is absent from most LL strains [28–30]. Notably, however, the LL strains NATL1A and NATL2A possess a photolyase gene and 41 HLIPs (Figure 2), which might provide photoprotection and help explain the ability of this clade to thrive in high-light surface waters.

More surprising are the differences between the closely related HLI and HLII ecotypes. A genome comparison of two strains representing these two HL clades (MED4 and MIT9312) revealed that 10–15% of each genome is not found in the other strain, despite 99.2% 16S rRNA identity [33]. These differences are concentrated in a few major genomic islands, which contain signatures of horizontal gene transfer, including repeats, tRNA genes and genes that apparently came from phage genomes. Although the functional significance of island-encoded genes is unclear, many of them are expressed under certain conditions, suggesting that they are integrated into *Prochlorococcus* metabolic networks [33–35].

This analysis of light adaptation in *Prochlorococcus* serves as a template for beginning to understand the drivers of diversity in microorganisms. Through the simultaneous analysis of physiological properties of cultures, the distributions of their relatives in the wild, and their genome content, we can begin to tease apart the environmental factors that are most important in driving evolution and shaping contemporary populations.

Phosphorus economy and ecology

Cross-scale investigations are pointing to phosphorus (P) availability, which varies with both depth and geography, as a key ecological factor and a potential driver of genome

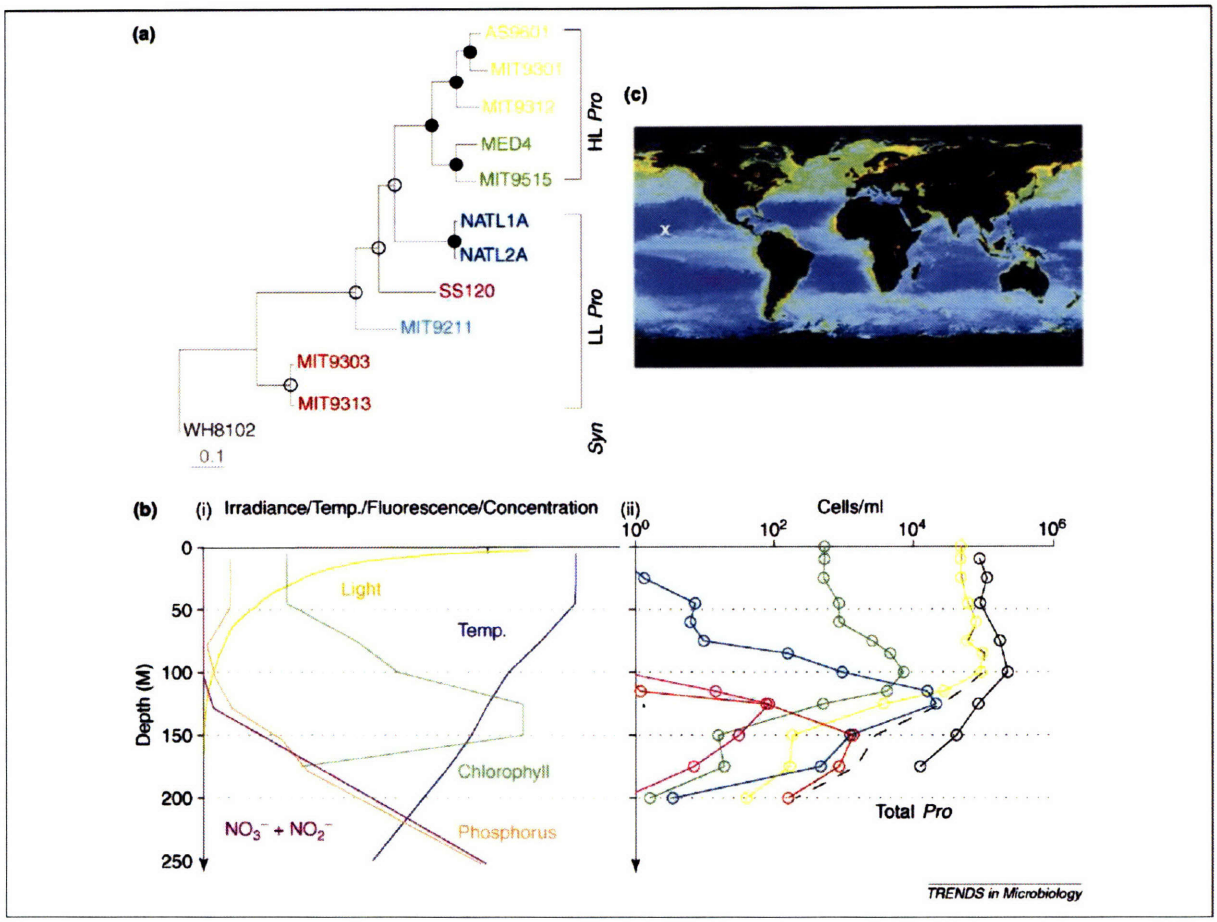


Figure 1. *Prochlorococcus* ecotypes: Evolutionary relationships and distributions along environmental gradients. **(a)** Phylogenetic tree showing the high-light-adapted and low-light-adapted clades (HL and LL), as defined by their physiological light-adaptation properties. The tree was generated from the *rpoB* gene by maximum likelihood (ML) with model parameters estimated from the data; nodes with >70% bootstrap support by two (unfilled circles) or three (filled circles) methods (distance, parsimony, ML) are shown. **(b)** (i) Depth profile (in meters) of water column characteristics [data from Hawaii Ocean Time-series (HOT): http://hahana.soest.hawaii.edu/hot/hot-dogs/interface.html] and (ii) abundance of *Prochlorococcus* clades (data from HOT; E. Zinser *et al.*, unpublished); both (i) and (ii) are from April 2003. Colors for each clade measured by quantitative PCR (qPCR). The black broken line is the sum of qPCR abundances, whereas the black unbroken line is total *Prochlorococcus* counted by flow cytometry. The discrepancy between the two totals indicates that current qPCR methods fail to detect some *Prochlorococcus* in deep waters. **(c)** Inset map shows surface water chlorophyll concentrations from NASA SeaWiFS (http://oceancolor.gsfc.nasa.gov/SeaWiFS/), seasonally averaged for spring 2003. HOT station shown by 'x'.

evolution in *Prochlorococcus*. This is not surprising, because P limits production in some regions of the oceans, with phosphate concentrations drawn down to the nanomolar range in some areas [36,37]. This extreme P depletion has selected for biochemical efficiency in *Prochlorococcus*, which has an unusually low cellular P:N ratio of 1:16–24 compared with other phytoplankton [38,39]. One biochemical adaptation enabling this P economy is the replacement of phospholipids with sulfolipids [40]. The small genome size (1.7 Mb for MED4) also reduces the P requirement relative to other essential elements, because the greatest demand for P in the cell is for nucleotide synthesis. Nevertheless, the MED4 chromosome contains more than half of the total cellular P [38].

Beyond this overall economization, different *Prochlorococcus* lineages possess further physiological adaptations for dealing with low P availability. Strain MED4 (HLI), for instance, can use a variety of organic P compounds for growth, whereas MIT9313 (LL) grows on a more limited range of P sources [41]. Furthermore, MED4 upregulates dozens of genes in response to P starvation, including known P-assimilation genes and novel genes of unknown function, which apparently enable it to endure prolonged P starvation [34]. By contrast, MIT9313 lacks many of these genes and is less capable of recovering from prolonged starvation [34]. These differences are in keeping with the preferred habitats of HL- and LL-adapted cells: surface waters often contain vanishingly small concentrations of inorganic phosphate, but regenerated organic P is available because of the intense biological activity in these waters. By contrast, LL *Prochlorococcus* cells are more abundant in deep waters, in closer proximity to the large reservoir of inorganic phosphate below the nutricline (Figure 1).

This simplistic HL–LL view of P acquisition quickly falls apart, however, as more strains are examined. MIT9312

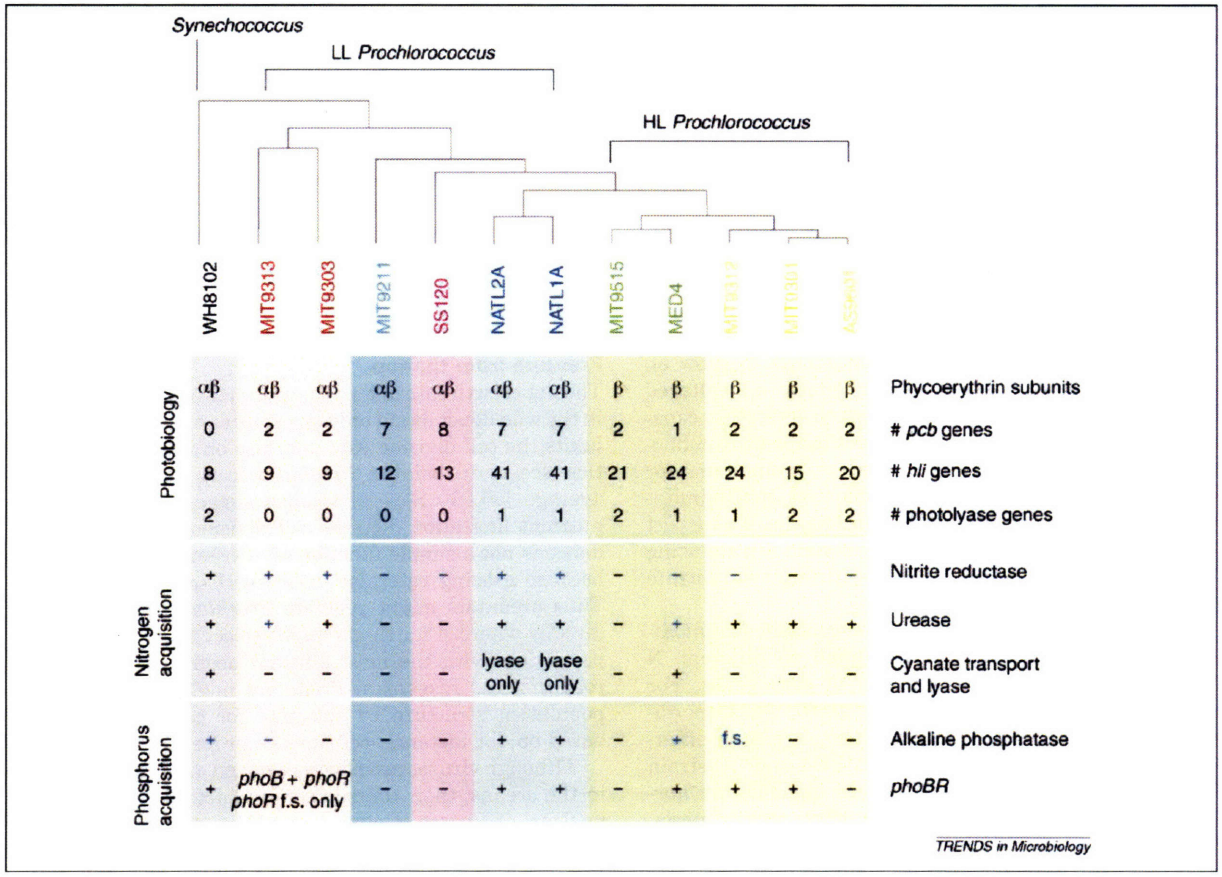


Figure 2. Presence or absence of genes involved in photosynthesis and nutrient acquisition in *Prochlorococcus* isolates. For +, – and f.s. entries in blue type, the activity of the gene, or lack thereof, has been confirmed by physiology experiments [41,45]. The presence of some genes, especially genes involved in photobiology, appears to map onto the phylogeny of the isolates, whereas the presence of nutrient acquisition genes varies even among the most closely related isolates. Cladogram shows the branching order of the isolates (from Figure 1). F.s. indicates frameshift mutation; + indicates presence of genes; – indicates absence of genes; and # indicates number of copies of a gene. Genomes from refs [28,30] and G. Kettler *et al.* (unpublished).

(HLII) lacks the broad P-utilization capabilities of MED4 [41]. In fact, among all HL isolates there is considerable variability in gene content related to P acquisition (Figure 2) [34]. In MED4, one large cluster of genes encodes regulation of P uptake (*phoBR*, *ptrA*), transport of phosphate (*phoE*, *pstSCAB*), and cleavage of organic P compounds (*phoA*), but nearly all of these genes are missing from HLI strain MIT9515, which shares 99.9% 16S rRNA identity with MED4. Similarly, HLII strains AS9601, MIT9301 and MIT9312 share 99.9% 16S rRNA identity, but they possess different gene complements for P acquisition. This variability in gene content is thought to reflect variation in selection imposed by P availability in the oceans: isolates from P-limited regions such as the Mediterranean and Sargasso Sea (e.g. MED4, MIT9301) have more genes for P acquisition, whereas isolates from iron-limited regions like the equatorial Pacific (e.g. MIT9515), or isolates from deep in the water column (e.g. MIT9312, MIT9313), have fewer genes for P acquisition [34]. Thus in addition to the vertical (depth) component there seems to be a horizontal or geographic component to P adaptation.

This hypothesis is supported by recent metagenomic data showing that the abundance of phosphate transport genes is correlated with P availability [42]. In the low-P waters of the Caribbean, for instance, the gene encoding the phosphate-binding protein PstS is seven times more abundant compared with the higher-P waters of the Pacific. Several P-related genes, including *pstS* and *phoH*, are also found in cyanophage genomes, suggesting that P might limit phage replication and offering a potential mechanism for lateral gene transfer of P genes among hosts [43]. An important next step is to combine these genome surveys with *in situ* functional measurements of gene expression, protein abundance and activity [24], and phosphate uptake rates, to understand how these gene content changes translate to physiology.

Obtaining nitrogen: more than one solution

Prochlorococcus does not fix dinitrogen gas (N_2), and surprisingly none of the cultured isolates can use nitrate as a nitrogen (N) source. Nevertheless, the upper waters of the open ocean contain several other N sources, and *Prochlorococcus* seems to use several of them. As expected, all strains can use ammonium, the form incorporated biosynthetically by glutamine synthetase and generated through recycling. Some, but not all, LL strains possess

nitrite reductase (Figure 2) and can use nitrite as their sole N source, consistent with the fact that their optimal light intensity for growth often occurs near the depth of the nitrite maximum in the ocean water column. Interestingly, the nitrite permease gene in MIT9313 seems to have been horizontally transferred [28]. Likewise horizontal transfer might have introduced a cyanate transporter and lyase in MED4 and an amino acid transporter in MIT9312, both in genomic islands [33]. Field studies suggest that *Prochlorococcus* can take up amino acids at an elevated rate [44], suggesting that these genes might be widespread in the wild. Finally, several HL and LL strains can grow on urea [45] (Figure 2).

The inability of cultured *Prochlorococcus* to grow on nitrate is surprising, given that virtually all cultured *Synechococcus*, a close relative, seem to have this capability and that nitrate can be abundant in the deep euphotic zone. Nitrate must be reduced to ammonium for biosynthesis, however, and it is possible that energy limitation precludes this pathway at extreme depths. We suspect that directed isolation approaches will yield nitrate-using *Prochlorococcus* from environments with sufficient nitrate and light.

In addition to having altered gene content, *Prochlorococcus* lineages seem to have adapted to different N regimes by altering the regulation of core genes. For example, the P_{II} protein (encoded by *glnB)*, which coordinates carbon and nitrogen metabolism, responds differently to N starvation in HL strain MED4 and LL strain MIT9313 [35]. The regulation of N metabolism in *Prochlorococcus* seems to be simpler than in other cyanobacteria, probably as a result of the unique selective pressures in oligotrophic environments [46].

Behind the scenes: essential micronutrients

In the open oceans, vanishingly low trace metal concentrations are maintained by, and select for, efficient uptake mechanisms in microorganisms [47]. These trace metals are essential cofactors for the metalloenzymes underlying major metabolic processes such as photosynthesis, carbon fixation and nutrient assimilation. Cobalt [48] and nickel (for superoxide dismutase [49]) are required for *Prochlorococcus* growth, and copper is probably required for plastocyanin. But copper can also be toxic to *Prochlorococcus*, even at the low concentrations found in the surface waters of oligotrophic oceans. HL cells are more resistant to copper toxicity than LL cells, consistent with the observation that free copper concentrations are greater near the surface [50]. Given this concentration gradient with depth, and the role of metals in both causing and curing (e.g. through superoxide dismutase) oxidative stress, we suspect that requirements for and sensitivities to trace metals will frequently be related to the light adaptation of the strain.

Iron has a key role in photophysiology and thus we also expect to find different iron physiologies in HL and LL ecotypes of *Prochlorococcus*. Iron is required in large amounts for photosystems but occurs at low bioavailable concentrations throughout much of the open oceans, and as a result limits primary production in these regions [51]. Indeed, the cell division rate, cell size and cellular chlorophyll content of *Prochlorococcus* have been shown to increase in response to iron addition in the equatorial Pacific, indicating cellular iron limitation [52,53]. The physiological iron–light interaction is exemplified by the light-harvesting antenna proteins (Pcb proteins). When iron is limiting, the LL strains SS120 and MIT9313, but not the HL strain MED4, upregulate specific Pcb antenna proteins that might be necessary for light harvesting or photoprotection (or both) [54]. The demand for iron, particularly by cells at low light intensity that need more photosystems, has undoubtedly influenced the genomes and physiology of *Prochlorococcus* in other ways that remain to be elucidated.

Pressure from the top

The net growth rate and population size of *Prochlorococcus* in the wild depends not only on resource availability, which limits the cell division rate, but also on mortality. Predation seems to balance *Prochlorococcus* growth rate on average [55,56]. Several studies suggest that although nutrient limitation dictates cell division rate, cell abundance is not a simple function of nutrient concentrations, because grazing rate also varies with nutrients [52,55]. Thus predators might regulate *Prochlorococcus* biomass directly whereas nutrients might select for subsets of the population with the most efficient acquisition and most robust stress responses. Predators might also influence population structure by selecting for certain cell types based on, for instance, cell surface properties [57].

Although viruses are now recognized as being abundant in the oceans, their contribution to bacterial mortality is unclear; estimates of virus-induced mortality range from 1% to 50% [58]. It is clear, however, that phages are important for *Prochlorococcus* evolution and diversity. Photosynthesis genes are found in phage genomes and probably recombine with host versions, thereby increasing diversity [59–61]. Phage genomes also contain genes involved in phosphate and carbon metabolism, suggesting further coevolution of host and phage [43]. Furthermore, *Prochlorococcus* genomes harbor strain-specific genomic islands encoding biosynthesis of lipopolysaccharides and other cell surface features [28,33]; these islands might help explain the different phage susceptibility [62] observed for cultured isolates, and might contribute to differential mortality and relative fitness of cells in the wild.

Revealing organization and complexity through metagenomics

Recently there has been a flood of metagenomics data from *Prochlorococcus*-rich regions of the oceans [42,63,64], and these data offer an unprecedented look at evolutionary processes and patterns of diversity. For example, large-insert fosmid sequences can reveal insertions and deletions of genes in wild cells compared with cultured isolates, whereas small-insert shotgun libraries can reveal large-scale patterns of gene presence and absence in the whole community (Figure 3). These datasets also reveal that the extensive sequence diversity is clearly structured, and that regions of the genome are hypervariable even in a single population [42]. Moreover, abundant cyanophage sequences in the cellular size fraction, probably derived from replicating intracellular phages, reinforce the importance

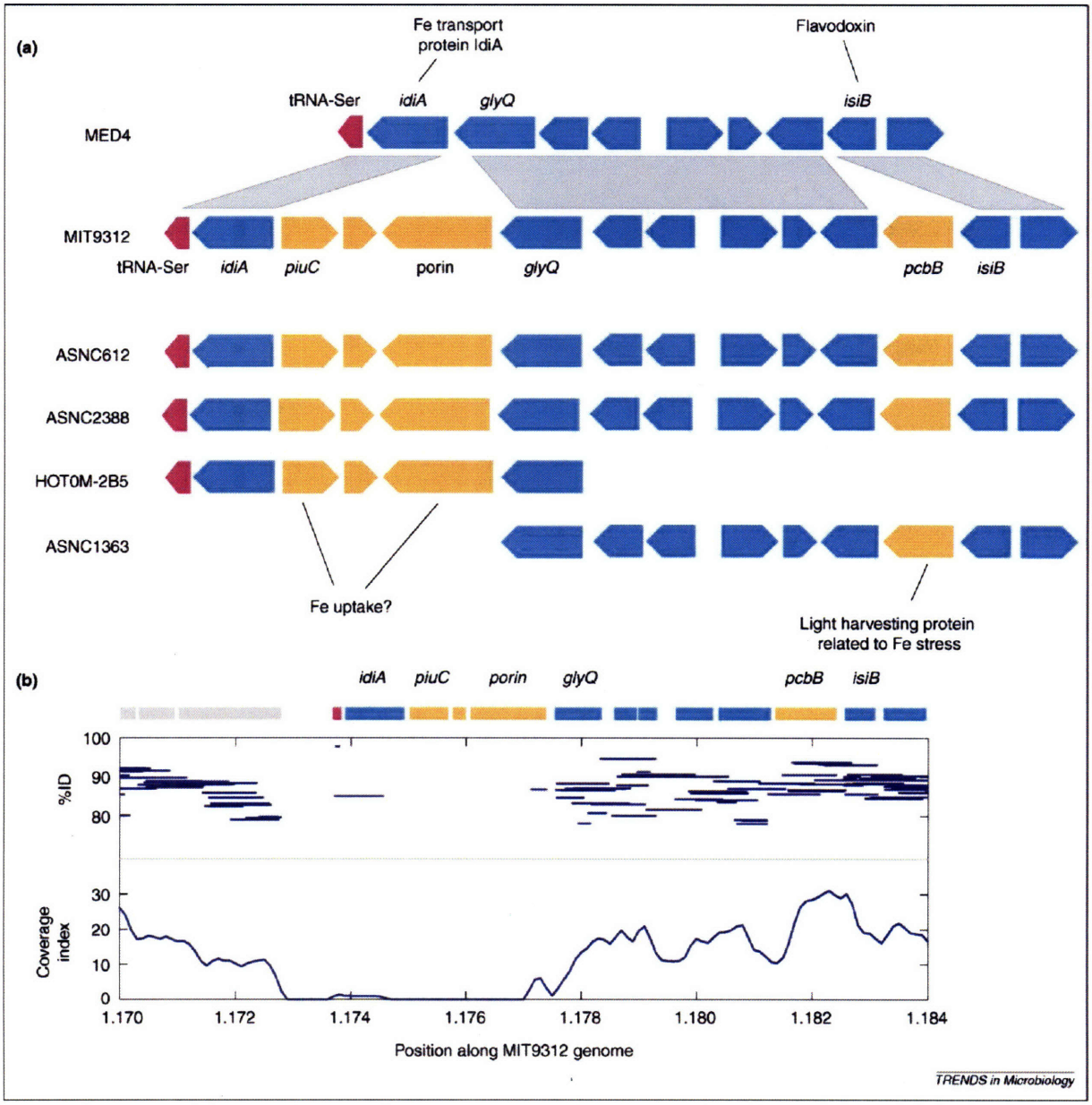


Figure 3. Patterns of diversity in wild *Prochlorococcus* revealed through metagenomics. Here we use metagenomics data to test whether genome variation observed in cultured strains of *Prochlorococcus* is observed in wild populations from two different (Atlantic and Pacific) oceanic regimes. **(a)** A genome fragment from two cultured strains, MED4 and MIT9312, compared with the homologous section of four large pieces of *Prochlorococcus* DNA collected from the Hawaii Ocean Time-series station (HOT) [33,64]. This region of the genome contains several genes thought to be involved in iron stress [28]. MED4 lacks the genes identified with orange color, whereas these particular wild cells from HOT all have them. **(b)** Smaller genome fragments from cells from the Sargasso Sea [63] aligned to the MIT9312 genome in this same region. The top panel shows the percent nucleotide identity (%ID) between the fragments and homologous regions in the MIT9312 genome, revealing ~80–90% identity in shared genes. The bottom panel shows an estimate of 'coverage', or how many sequences in this sample were homologous to a particular section of the genome. The Sargasso population seems to lack several putative iron stress genes found in the Pacific.

of these viruses for mortality and host evolution [64]. Metagenomic data have been so informative for *Prochlorococcus* and its viruses in part because of the availability of whole-genome sequences of cultured isolates, which serve as scaffolds for interpreting the environmental data. Thus it is crucial that future sequencing efforts include both wild populations and whole genomes of cultured isolates.

Conclusions and future perspectives

A picture of *Prochlorococcus* ecology and evolution is emerging thanks to the combination of approaches and scales of interrogation employed (Box 1; Figure 4). Clearly this picture is far more complex than can be captured by the HL–LL ecotype paradigm that first emerged from studies of light physiology and molecular phylogenies. For nutrients like N and P, gene loss and gene gain have occurred

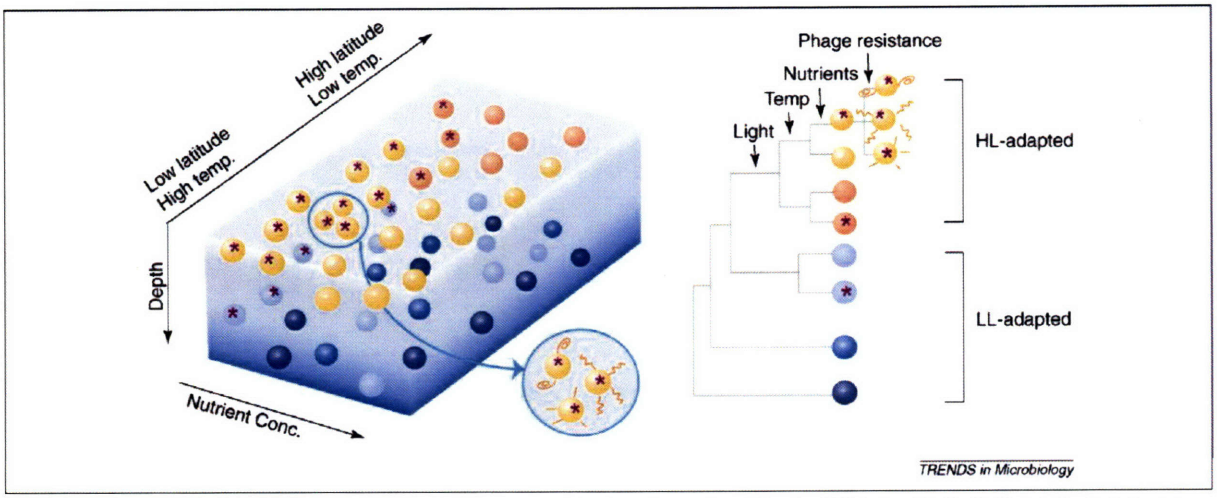


Figure 4. Hypothetical patterns of diversity along different *Prochlorococcus* niche dimensions and spatial scales. Within the globally distributed HL-adapted group, we might see differentiation based on temperature, and within each temperature-adapted clade we might see multiple 'nutrient ecotypes' at more local scales, and multiple phage-resistance types within a single nutrient ecotype. A robust understanding of these patterns will require all of the tools at our disposal: metagenomics, single cell genomics and models of cellular metabolism. Key: spheres represent *Prochlorococcus* cells. Shades of orange represent HL-adapted clades; shades of blue, LL-adapted clades. Purple stars represent additional adaptations to a specific nutrient-depleted environment (e.g. low P, low Fe). Exterior decoration represents cell surface diversity, such as different lipopolysaccharide structures or surface proteins, which might confer resistance to certain phages or grazers.

within the 'ecotypes' delineated by rRNA ITS sequence clusters, and as a result nutrient physiology does not map neatly onto the ITS phylogeny. At an even finer resolution, extremely closely related cells have different phage susceptibilities (Figure 4). Thus, the recognition of clades and clusters, and their interpretation in light of ecological factors, depends on the scale of observation.

Prochlorococcus is but one of many model organisms (e.g. *Pelagibacter*, *Roseobacter*, *Vibrio* and *Synechococcus*) that are proving useful for understanding marine microbial processes, and biological systems in general. As a microbially dominated ecosystem, where the physics and chemistry are fairly well understood and modeled, the oceans provide us with an opportunity to connect, for the first time, the information in genomes to the dynamics of whole ecosystems [65]. One of the new frontiers for marine microbiology is moving from the 'parts list', that is, the genetic information carried by marine microbes, to understanding the function of the metabolic machinery of the oceans. This machinery is an integral component of global biogeochemical cycles, and we must strive to understand how it responds to perturbations at local and global scales. Techniques for measuring gene expression *in situ* at both the RNA and protein levels are on the horizon, and these can be validated using model systems like *Prochlorococcus*. Furthermore, single cell analyses [66,67] combined with revolutionary advances in DNA sequencing technologies will change the way we study marine microbial systems.

Even in these relatively simple living systems, we are humbled and challenged by the complexity and robustness of nature. Our hope is that *Prochlorococcus* and other model microbial systems from the oceans will not only advance our understanding of microbial processes, but also yield fundamental insights into the structure, function and evolution of life in general.

www.sciencedirect.com

Acknowledgements

We thank members of the Chisholm Laboratory, especially J. Waldbauer and M. Sullivan, for comments on the manuscript: Ed DeLong for providing fosmid clones; and *Prochlorococcus* enthusiasts, past and present, for their contributions to this story. We thank Jim Tiedje, Dan Drell, Frank Larimer, Elbert Branscomb and the DOE Joint Genome Institute for guiding the entry of *Prochlorococcus* into the genomic era. This research on *Prochlorococcus* has been supported over the years by grants from the DOE, NSF, ONR and the Seaver Foundation (to S.W.C.). M.L.C. is supported by an NSF Graduate Fellowship. Particular thanks are due to the Gordon and Betty Moore Foundation for their support of some of this research, and their recent transformative role in the field of marine microbiology.

References

1 Johnson, P.W. and Sieburth, J.M. (1979) Chroococcoid cyanobacteria in the sea – ubiquitous and diverse phototropic biomass. *Limnol. Oceanogr.* 24, 928–935
2 Gieskes, W.W. and Kraay, G.W. (1983) Unknown chlorophyll-a derivatives in the North-Sea and the tropical Atlantic-Ocean revealed by HPLC analysis. *Limnol. Oceanogr.* 28, 757–766
3 Partensky, F. *et al.* (1999) *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol. Mol. Biol. Rev.* 63, 106–127
4 Chisholm, S.W. *et al.* (1992) *Prochlorococcus-marinus* nov gen-nov sp – an oxyphototrophic marine prokaryote containing divinyl chlorophyll-a and chlorophyll-B. *Arch. Microbiol.* 157, 297–300
5 Chisholm, S.W. *et al.* (1988) A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature* 334, 340–343
6 Olson, R.J. *et al.* (1990) Spatial and temporal distributions of prochlorophyte picoplankton in the North-Atlantic Ocean. *Deep-Sea Res.* 37, 1033–1051
7 Goericke, R. and Repeta, D.J. (1992) The pigments of *Prochlorococcus marinus* – the presence of divinyl chlorophyll-a and chlorophyll-B in a marine prokaryote. *Limnol. Oceanogr.* 37, 425–433
8 Partensky, F. *et al.* (1997) The divinyl-chlorophyll a/b-protein complexes of two strains of the oxyphototrophic marine prokaryote *Prochlorococcus* – Characterization and response to changes in growth irradiance. *Photosynth. Res.* 51, 209–222
9 Moore, L.R. and Chisholm, S.W. (1999) Photophysiology of the marine cyanobacterium *Prochlorococcus*: Ecotypic differences among cultured isolates. *Limnol. Oceanogr.* 44, 628–638

10 Moore, L.R. *et al.* (1995) Comparative physiology of *Synechococcus* and *Prochlorococcus* – Influence of light and temperature on growth, pigments, fluorescence and absorptive properties. *Mar. Ecol. Prog. Ser.* 116, 259–275
11 Steglich, C. *et al.* (2005) A green light-absorbing phycoerythrin is present in the high-light-adapted marine cyanobacterium *Prochlorococcus* sp MED4. *Environ. Microbiol.* 7, 1611–1618
12 Steglich, C. *et al.* (2003) Photophysical properties of *Prochlorococcus marinus* SS120 divinyl chlorophylls and phycoerythrin *in vitro* and *in vivo*. *FEBS Lett.* 553, 79–84
13 Morel, A. *et al.* (1993) *Prochlorococcus* and *Synechococcus* – a comparative study of their optical properties in relation to their size and pigmentation. *J. Mar. Res.* 51, 617–649
14 Moore, L.R. *et al.* (1998) Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* 393, 464–467
15 Urbach, E. *et al.* (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
16 Rocap, G. *et al.* (2002) Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. *Appl. Environ. Microbiol.* 68, 1180–1191
17 Ferris, M.J. and Palenik, B. (1998) Niche adaptation in ocean cyanobacteria. *Nature* 396, 226–228
18 Steglich, C. *et al.* (2003) Analysis of natural populations of *Prochlorococcus* spp. in the northern Red Sea using phycoerythrin gene sequences. *Environ. Microbiol.* 5, 681–690
19 Ahlgren, N.A. *et al.* (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
20 Zinser, E.R. *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
21 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, 2585–2591
22 West, N.J. *et al.* (2001) Closely related *Prochlorococcus* genotypes show remarkably different depth distributions in two oceanic regions as revealed by *in situ* hybridization using 16S rRNA-targeted oligonucleotides. *Microbiology* 147, 1731–1744
23 Fuller, N.J. *et al.* (2006) Molecular analysis of picocyanobacterial community structure along an Arabian Sea transect reveals distinct spatial separation of lineages. *Limnol. Oceanogr.* 51, 2515–2526
24 Fuller, N.J. *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
25 Bouman, H.A. *et al.* (2006) Oceanographic basis of the global surface distribution of *Prochlorococcus* ecotypes. *Science* 312, 918–921
26 Johnson, Z.I. *et al.* (2006) Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* 311, 1737–1740
27 Zinser, E.R. *et al.* Influence of light and temperature on *Prochlorococcus* ecotype distributions in the Atlantic Ocean. *Limnol. Oceanogr.* (in press)
28 Rocap, G. *et al.* (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* 424, 1042–1047
29 Hess, W.R. *et al.* (2001) The photosynthetic apparatus of *Prochlorococcus*: insights through comparative genomics. *Photosynth. Res.* 70, 53–71
30 Dufresne, A. *et al.* (2003) Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *Proc. Natl. Acad. Sci. U. S. A.* 100, 10020–10025
31 Garczarek, L. *et al.* (2000) Multiplication of antenna genes as a major adaptation to low light in a marine prokaryote. *Proc. Natl. Acad. Sci. U. S. A.* 97, 4098–4101
32 He, Q. *et al.* (2001) The high light-inducible polypeptides in *Synechocystis* PCC6803. Expression and function in high light. *J. Biol. Chem.* 276, 306–314
33 Coleman, M.L. *et al.* (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* 311, 1768–1770
34 Martiny, A.C. *et al.* (2006) Phosphate acquisition genes in *Prochlorococcus* ecotpes: evidence for genome-wide adaptation. *Proc. Natl. Acad. Sci. U. S. A.* 103, 12552–12557
35 Tolonen, A.C. *et al.* (2006) Global gene expression of *Prochlorococcus* ecotypes in response to changes in nitrogen availability. *Mol. Syst. Biol.* 2, 53
36 Ammerman, J.W. *et al.* (2003) Phosphorus deficiency in the Atlantic: an emerging paradigm in oceanography. *Eos. Trans. A.G.U.* 84, 165
37 Thingstad, T.F. *et al.* (2005) Nature of phosphorus limitation in the ultraoligotrophic eastern Mediterranean. *Science* 309, 1068–1071
38 Bertilsson, S. *et al.* (2003) Elemental composition of marine *Prochlorococcus* and *Synechococcus*: Implications for the ecological stoichiometry of the sea. *Limnol. Oceanogr.* 48, 1721–1731
39 Heldal, M. *et al.* (2003) Elemental composition of single cells of various strains of marine *Prochlorococcus* and *Synechococcus* using X-ray microanalysis. *Limnol. Oceanogr.* 48, 1732–1743
40 Van Mooy, B.A. *et al.* (2006) Sulfolipids dramatically decrease phosphorus demand by picocyanobacteria in oligotrophic marine environments. *Proc. Natl. Acad. Sci. U. S. A.* 103, 8607–8612
41 Moore, L.R. *et al.* (2005) Ecotypic variation in phosphorus-acquisition mechanisms within marine picocyanobacteria. *Aquat. Microb. Ecol.* 39, 257–269
42 Rusch, D.B. *et al.* (2007) The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.* 5, e77
43 Sullivan, M.B. *et al.* (2005) Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. *PLoS Biol.* 3, e144
44 Zubkov, M.V. *et al.* (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
45 Moore, L.R. *et al.* (2002) Utilization of different nitrogen sources by the marine cyanobacteria *Prochlorococcus* and *Synechococcus*. *Limnol. Oceanogr.* 47, 989–996
46 Garcia-Fernandez, J.M. *et al.* (2004) Streamlined regulation and gene loss as adaptive mechanisms in *Prochlorococcus* for optimized nitrogen utilization in oligotrophic environments. *Microbiol. Mol. Biol. Rev.* 68, 630
47 Morel, F.M.M. and Price, N.M. (2003) The biogeochemical cycles of trace metals in the oceans. *Science* 300, 944–947
48 Saito, M.A. *et al.* (2002) Cobalt limitation and uptake in *Prochlorococcus*. *Limnol. Oceanogr.* 47, 1629–1636
49 Eitinger, T. (2004) *In vivo* production of active nickel superoxide dismutase from *Prochlorococcus marinus* MIT9313 is dependent on its cognate peptidase. *J. Bacteriol.* 186, 7821–7825
50 Mann, E.L. *et al.* (2002) Copper toxicity and cyanobacteria ecology in the Sargasso Sea. *Limnol. Oceanogr.* 47, 976–988
51 Boyd, P.W. *et al.* (2007) Mesoscale iron enrichment experiments 1993-2005: Synthesis and future directions. *Science* 315, 612–617
52 Mann, E.L. and Chisholm, S.W. (2000) Iron limits the cell division rate of *Prochlorococcus* in the eastern equatorial Pacific. *Limnol. Oceanogr.* 45, 1067–1076
53 Cavender-Bares, K.K. *et al.* (1999) Differential response of equatorial Pacific phytoplankton to iron fertilization. *Limnol. Oceanogr.* 44, 237–246
54 Bibby, T.S. *et al.* (2003) Low-light-adapted *Prochlorococcus* species possess specific antennae for each photosystem. *Nature* 424, 1051–1054
55 Worden, A.Z. and Binder, B.J. (2003) Application of dilution experiments for measuring growth and mortality rates among *Prochlorococcus* and *Synechococcus* populations in oligotrophic environments. *Aquat. Microb. Ecol.* 30, 159–174
56 Landry, M.R. *et al.* (2003) Phytoplankton growth and microzooplankton grazing in high-nutrient, low-chlorophyll waters of the equatorial Pacific: Community and taxon-specific rate assessments from pigment and flow cytometric analyses. *J. Geophys. Res.* 108, 8142
57 Monger, B.C. *et al.* (1999) Feeding selection of heterotrophic marine nanoflagellates based on the surface hydrophobicity of their picoplankton prey. *Limnol. Oceanogr.* 44, 1917–1927
58 Fuhrman, J.A. (1999) Marine viruses and their biogeochemical and ecological effects. *Nature* 399, 541–548

59 Lindell, D. *et al.* (2004) Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc. Natl. Acad. Sci. U. S. A.* 101, 11013–11018
60 Sullivan, M.B. *et al.* (2006) Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol.* 4, 1344–1357
61 Zeidner, G. *et al.* (2005) Potential photosynthesis gene recombination between *Prochlorococcus* and *Synechococcus* via viral intermediates. *Environ. Microbiol.* 7, 1505–1513
62 Sullivan, M.B. *et al.* (2003) Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. *Nature* 424, 1047–1051
63 Venter, J.C. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304, 66–74
64 DeLong, E.F. *et al.* (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science* 311, 496–503
65 Follows, M. *et al.* (2007) Emergent biogeography of microbial communities in a model ocean. *Science* 315, 1843–1846
66 Zhang, K. *et al.* (2006) Sequencing genomes from single cells by polymerase cloning. *Nat. Biotechnol.* 24, 680–686
67 Stepanauskas, R. and Sieracki, M.E. (2007) Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proc. Natl. Acad. Sci. U. S. A.* 104, 9052–9057
68 DuRand, M.D. *et al.* (2001) Phytoplankton population dynamics at the Bermuda Atlantic Time-series station in the Sargasso Sea. *Deep-Sea Res. Pt II* 48, 1983–2003
69 Campbell, L. *et al.* (1994) The importance of *Prochlorococcus* to community structure in the central North Pacific Ocean. *Limnol. Oceanogr.* 39, 954–961
70 Cohan, F.M. (2006) Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. *Philos. Trans. R. Soc.Lond. B Biol. Sci.* 361, 1985–1996

CHAPTER SIX

Conclusions and Future Directions

SUMMARY AND FUTURE DIRECTIONS

The work presented here has helped improve our grasp of the nature and extent of genome-wide diversity in *Prochlorococcus*, both in cultured isolates and natural populations, and has identified several ecological and evolutionary factors that influence this diversity. Building upon the foundational discovery of high- and low-light adapted ecotypes, we have extended this understanding of adaptation and diversification to finer phylogenetic resolution and to other environmental drivers.

Summary

Comparing the genomes of two closely related high-light adapted isolates reveals substantial gene content diversity: 10-15% of the genes in each strain are not found in the other (Chapter 2). These variable genes are not spread randomly throughout the genome, but are concentrated in a few genomic islands, reminiscent of pathogenicity islands that have been widely observed in host-associated bacteria (Hacker and Carniel, 2001). These islands appear to be hotspots for horizontal gene transfer, likely via phages, based on their anomalous GC content, proximity to tRNA loci, and presence of genes such as *hli* genes that are also found in virus genomes (Lindell et al., 2004). These islands are also present in natural populations as inferred from large DNA fragments from the subtropical Pacific, and in fact their gene content varies even within a single ecotype and a single population. Such high variability could be interpreted to mean that these genes are mostly nonadaptive and may be transient in the genome, eventually to be lost. However, many of these island-encoded genes do appear to be functional, as they are expressed in response to phosphate starvation (Martiny et al., 2006, Chapter 3), nitrogen starvation (Tolonen et al., 2006), phage infection (Lindell et al., 2007, Appendix D), and high light shift (Steglich et al., 2006), and many exhibit regulated, cyclic expression over the diel period (Zinser et al., *in prep.*). The proportion of island-encoded genes that are functional vs. those that are "junk" remains unknown and deserves further study, in order to fully appreciate the role of genomic islands in adaptive evolution.

The work presented in Chapter 3 represents a case study for understanding how an environmental driver, in this case phosphate availability, can shape gene content and physiology over evolutionary time. Two isolates from different geographic locations, different depths in the water column, with different light physiology, and with vastly different genome architecture and content, were compared with respect to their physiology and gene expression in response to phosphate starvation. As predicted, MED4, isolated from the low-phosphate waters of the Mediterranean Sea, appears better-adapted to low phosphate conditions than MIT9313. But given so many confounding differences between these two strains, it is difficult to say which factor contributed most to their differentiation. Therefore we explored the organization of phosphate-related genes in the genomes of a wider range of isolates. Strikingly, the

most closely related isolates based on 16S rRNA identity had the most dissimilar gene complements for phosphate assimilation. These gene complements instead appeared correlated to the environment where the strain was isolated: isolates from low phosphate regions of the oceans have a greater number of genes for phosphate assimilation. Another important result from this work is that several genes located in a genomic island in strain MED4 and absent from the other cultured strains are strongly induced under phosphate limitation. These genes are also found to be physically linked to known phosphate assimilation genes in clones from the Sargasso Sea (Venter et al., 2004), an environment thought to be phosphate limited. Taken together, this evidence strongly suggests that these island genes are adaptive in low phosphate environments and that horizontal gene transfer plays an important role in adaptive genome evolution. The biochemical or physiological functions of these novel phosphate-related genes remain to be elucidated.

Using metagenomics, we have extended our understanding of genome-wide diversity and gene expression from cultured isolates to natural populations (Chapter 4). The core and flexible genomes, as delineated by Kettler et al. (2007, Appendix C) are clearly observable in natural populations as well. Within the flexible genome, a subset of genes behaves like core genes, being present in about one copy per cell. Another subset of the flexible genome, in contrast, is found only rarely in individual cells. Moreover, some genes within the flexible genome show trends with depth, such as genes involved in thiamin metabolism. Such a quantitative picture of gene stoichiometry in *Prochlorococcus*, extended to other environments, will enable us to identify genes potentially involved in local adaptation.

The deep coverage and short read length of this dataset reveal anomalous regions of the chromosome that, upon closer inspection, appear to have undergone recombination. These regions contain a number of highly expressed genes involved in key metabolic pathways, and therefore recombination could have an enormous effect on *Prochlorococcus* phenotypic evolution. Among the most highly expressed genes at each depth are conserved hypothetical proteins with no known function, several of which are found in genomic islands. We also found that gene expression in natural populations correlated strongly with that in cultured isolates as measured by microarrays over a diel cycle (Zinser et al., *in prep.*), and that 88% of single-copy core genes were detected as mRNA, reinforcing the critical role of the core for *Prochlorococcus* metabolism.

Emergent themes

Some genes found in genomic islands and acquired by gene transfer, likely play important metabolic roles. This is evidenced by their strong and specific upregulation under various environmental stressors, and also by their recurrence in natural populations. Some of these island genes have essentially become "core" in certain natural populations, suggesting they confer a fitness advantage in specific

habitats. Alternatively, these genes may be hitchhiking along with other adaptive loci, or they may be in the process of slowly being lost but were once advantageous. These genes represent important targets for biochemical characterization. Moreover, they support our hypothesis that islands serve as dynamic reservoirs of gene gain and loss that enable adaptation to local environmental conditions.

One of the most gratifying conclusions emerging from this work is that our cultured *Prochlorococcus* isolates paint a fairly representative picture of natural diversity. Genes identified as single-copy core genes based on genome sequences from just 12 isolates are, in fact, single-copy in natural populations as well, with very few exceptions. Flexible genes that occur in most of the cultured HL genome representatives also tend to be abundant in natural populations, while flexible genes that occur in only one or two HL genomes tend to be rare in natural populations. Thus the 12 sequenced genomes currently available are, overall, a good indicator of genome diversity in the wild, at least for the dominant eMIT9312 clade which makes up 80-99% of the *Prochlorococcus* cells in our samples. Moreover, the temporal patterns of expression we observed in natural populations correlate strongly with those observed in a cultured isolate. This correspondence between isolates and natural populations could be considered surprising given the small fraction of *Prochlorococcus*' natural ecological space represented by the isolate collection, but it demonstrates that genomic sequencing of even a few strains can be a valuable step towards understanding a large and widely-distributed microbial population.

Together, this thesis documents that *Prochlorococcus* genomic diversity has a clear structure and that this structure is scale-dependent. An emergent picture of this cross-scale diversity is presented in Chapter 5. Previous work documented genome evolution along the most conspicuous environmental axis, light/depth. We can now add more and more axes including nutrient availability, temperature, phage infection and predation, and each of these drivers is reflected differently in the genome. Much work remains to be done to fully understand these genomic imprints, but there is now a framework for thinking about these layers of diversity.

Future Directions

This thesis both advances understanding and raises further questions. For instance, the genomic basis of temperature adaptation in HL *Prochlorococcus* remains unexplained and may lie in the sequences of a few proteins. While temperature physiology explains the distinct distributions of two HL ecotypes, the traits distinguishing different LL lineages are unknown. Moreover, the extent of diversity in LL *Prochlorococcus* is much less certain, and is certainly far greater than in HL ecotypes. The relative importance of mutation, recombination, and gene transfer for genome evolution in LL cells deep in the water column may be different as well, since *Prochlorococcus* cell densities are orders of magnitude lower than at the surface, which may affect the rates of horizontal gene transfer and homologous

recombination, and mutation frequencies may differ due to UV exposure and different DNA repair systems.

Although we can conclude that both horizontal gene transfer and homologous recombination are important processes in *Prochlorococcus* genome evolution, we know essentially nothing about the mechanisms underlying them. We suspect that phage are important vehicles for moving genes around, but so far no lysogens have been documented nor have rates of "sloppy packaging" of phage DNA been bounded. Nothing is known about conjugation or direct DNA uptake in natural *Prochlorococcus* populations either. Identifying these mechanisms is key to understanding rates of these processes.

Finally, this thesis reinforces the ecological importance of many genes of unknown function in the open oceans. For example, genes induced under phosphate starvation and located in a genomic island are annotated as conserved hypotheticals and bear little similarity to any described proteins. The recent Global Ocean Survey uncovered thousands of novel protein families that have no similarity to sequences in GenBank (Yooseph et al., 2007). Thus it is possible we will yet discover some of the most important and interesting biological features of these organisms.

References

Hacker, J, and E Carniel. 2001. Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Reports* 2: 376-381.

Kettler, G, AC Martiny, K Huang, J Zucker, ML Coleman, S Rodrigue, F Chen, A Lapidus, S Ferriera, J Johnson, C Steglich, GM Church, PM Richardson, and SW Chisholm. 2007. Patterns and Implications of Gene Gain and Loss in the Evolution of *Prochlorococcus*. *PLoS Genetics* 3: e231.

Lindell D, JD Jaffe, ML Coleman, ME Futschik, IM Axmann, T Rector, G Kettler, MB Sullivan, R Steen, WR Hess, GM Church, and SW Chisholm. 2007. Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature* 449: 83-6.

Lindell D, MB Sullivan, ZI Johnson, AC Tolonen, F Rohwer, and SW Chisholm. 2004. Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *PNAS* 101: 11013-8.

Martiny, AC, ML Coleman, and SW Chisholm. 2006. Phosphate acquisition genes in *Prochlorococcus* ecotypes: evidence for genome-wide adaptation. *PNAS* 103: 12552-12557.

Steglich, C, M Futschik, T Rector, R Steen, and SW Chisholm. 2006. Genome-wide analysis of light sensing in Prochlorococcus. *Journal of Bacteriology* 188: 7796-7806.

Tolonen, AC, J Aach, D Lindell, ZI Johnson, T Rector, R Steen, GM Church, and SW Chisholm. 2006. Global gene expression of *Prochlorococcus* ecotypes in response to changes in nitrogen availability. *Molecular Systems Biology* 2: 53.

Venter, JC, K Remington, JF Heidelberg, AL Halpern, D Rusch, JA Eisen, DY Wu, I Paulsen, KE Nelson, W Nelson, DE Fouts, S Levy, AH Knap, MW Lomas, K Nealson, O White, J Peterson, J Hoffman, R Parsons, H Baden-Tillson, C Pfannkoch, YH Rogers, and HO Smith. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304: 66-74.

Yooseph, S, G Sutton, DB Rusch, AL Halpern, SJ Williamson, K Remington, JA Eisen, KB Heidelberg, G Manning, W Li, L Jaroszewski, P Cieplak, CS Miller, H Li, ST Mashiyama, MP Joachimiak, C van Belle, JM Chandonia, DA Soergel, Y Zhai, K Natarajan, S Lee, BJ Raphael, V Bafna, R Friedman, SE Brenner, A Godzik, D Eisenberg, JE Dixon, SS Taylor, RL Strausberg, M Frazier, JC Venter. 2007. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol* 5: e16.

Appendix A

Portal protein diversity and phage ecology

Matthew B. Sullivan, Maureen L. Coleman, Vanessa Quinlivan, Jessica E. Rosenkrantz, Alicia S. DeFrancesco, Grace Tan, Ross Fu, Jessica Lee, John B. Waterbury, Joseph P. Bielawski, and Sallie W. Chisholm

Sullivan, M.B., **Coleman, M.L.**, Quinlivan, V., Rosenkrantz, J.E., DeFrancesco, A.S., Tan, G., Fu, R., Lee, J., Waterbury, J.B., Bielawski, J.P. and Chisholm, S.W. Portal protein diversity and phage ecology. *Environmental Microbiology, in press.*

Abstract

Oceanic phages are critical components of the global ecosystem, where they play a role in microbial mortality and evolution. Our understanding of phage diversity is greatly limited by the lack of useful genetic diversity measures. Previous studies, focused on myophages that infect the marine cyanobacterium *Synechococcus*, have used the coliphage T4 portal-protein-encoding homolog, gene 20 (g20), as a diversity marker. These studies revealed ten sequence clusters, 9 oceanic and 1 freshwater, where only three contained cultured representatives. We sequenced g20 from 38 marine myophages isolated using a diversity of *Synechococcus* and *Prochlorococcus* hosts to see if any would fall into the clusters that lacked cultured representatives. On the contrary, all fell into the three clusters that already contained sequences from cultured phages. Further, there was no obvious relationship between host of isolation, or host range, and g20 sequence similarity. We next expanded our analyses to all available g20 sequences (769 sequences), which includes PCR amplicons from wild uncultured phages, non-PCR-amplified sequences identified in the Global Ocean Survey (GOS) metagenomic database, as well as sequences from cultured phages, to evaluate the relationship between g20 sequence clusters and habitat features from which the phage sequences were isolated. Even in this meta-dataset, very few sequences fell into the sequence clusters without cultured representatives, suggesting that the latter are very rare, or sequencing artifacts. In contrast, sequences most similar to the culture-containing clusters, the freshwater cluster and two novel clusters were more highly represented, with one particular culture-containing cluster representing the dominant g20 genotype in the unamplified GOS sequence data. Finally, while some g20 sequences were non-randomly distributed with respect to habitat, there were always numerous exceptions to general patterns, indicating that phage portal proteins are not good predictors of a phage's host or the habitat in which a particular phage may thrive.

Virus-like particles occur in high abundance (to 10^8 ml^{-1}) in the oceans (Bergh, 1989; Bratbak et al., 1990; Proctor and Fuhrman, 1990). One of the most well studied phage-host systems in this habitat is the phages that infect the marine cyanobacteria *Prochlorococcus* and *Synechococcus*, which are globally important marine primary producers (Waterbury et al., 1986; Partensky et al., 1999). These 'cyanophages' are abundant (Waterbury and Valois, 1993; Suttle and Chan, 1994; Suttle, 2000; Lu et al., 2001; Frederickson et al., 2003; Marston and Sallee, 2003; Sullivan et al., 2003), contribute to host mortality (Waterbury and Valois, 1993; Suttle and Chan, 1994; Suttle, 2000), and are thought to play a role in maintaining the extensive microdiversity of their hosts (Waterbury and Valois, 1993; Suttle and Chan, 1994; Marston and Sallee, 2003; Sullivan et al., 2003) likely through killing the winner (*sensu* Thingstad, 2000) and through the movement of genes throughout the host population (Lindell et al., 2004; Coleman et al., 2006; Sullivan et al., 2006).

Studying the diversity of phages has proven difficult because no universal gene, analogous to the 16S rRNA gene used for microbes, exists throughout all phage families (Paul et al., 2002). Thus family-specific genes have been proposed for use as taxonomic tools in phage ecology (Rohwer and Edwards, 2002). One such marker, a homolog to the coliphage T4 portal protein gene 20 (g20), has been developed to study the diversity of *Myoviridae* – one of the most common phage types observed in metagenomics surveys (Breitbart et al., 2002; Breitbart et al., 2004c; DeLong et al., 2006) and among *Synechococcus* cyanophage isolates (Suttle and Chan, 1993; Waterbury and Valois, 1993; Wilson et al., 1993; Sullivan et al., 2003). The g20 homolog is ubiquitous among T4-like myoviruses (see T4-like phages genome website http://phage.bioc.tulane.edu/) with hosts ranging from proteobacteria to cyanobacteria (Fuller et al., 1998; Hambly et al., 2001; Mann et al., 2005; Sullivan et al., 2005). The evolution of g20 is likely constrained because its protein product initiates capsid assembly (at least in T4), a process which involves geometric precision (Coombs and Eiserling, 1977; Hsiao and Black, 1978; van Driel and Couture, 1978) through the formation of a proximal vertex (van Driel and Couture, 1978) used for DNA packaging (Hsiao and Black, 1978) and binding the capsid to the tail junction (Coombs and Eiserling, 1977).

The availability of cultured cyanomyophage (Waterbury & Valois 1993, Wilson et al. 1993, Suttle & Chan 1993, Marston & Sallee 2003, Sullivan et al. 2003) has allowed the design of cyanomyophage-specific g20 sequence PCR primers that have been used to study this component of viral populations in the wild. Early studies using non-degenerate PCR primers and DNA 'fingerprinting' techniques (e.g., denaturing gradient gel electrophoresis and terminal-restriction fragment length polymorphism banding patterns) revealed variability in g20 diversity across gradients in space and time from a variety of different environments (Wilson et al., 1999, 2000; Frederickson et al., 2003; Dorigo et al., 2004; Wang and Chen, 2004; Muhling et al. 2005; Sandaa and Larsen, 2006). These studies concluded that g20 diversity was as great within a sample as between oceans (Wilson et al., 1999), that phage g20 diversity increased as *Synechococcus* abundance increased (Wilson et al., 1999, 2000; Frederickson et al., 2003; Wang and Chen, 2004; Sandaa and Larsen, 2006), that some g20 types were ubiquitous in the habitats examined (Wilson et al., 1999, 2000; Frederickson et al., 2003; Dorigo et al., 2004), as well as a temporal study by Muhling et al. (2005) that correlated 'cyanophage' diversity (inferred from g20 sequence types) with *Synechococcus* diversity (inferred from *rpoC1* sequence types).

Subsequent cloning and sequencing of g20 PCR amplicons from both cultured isolates and wild populations have allowed phylogenetic analyses of cyanomyophage diversity. Although initial studies (Zhong et al., 2002) suggested some correlation between ocean habitat and g20 phylogeny (e.g., phylogenetic cluster II represents "open ocean" g20 sequences), further sampling revealed this was not the case, as seven g20 sequences from coastal *Synechococcus* myophages isolated from Rhode Island waters clustered with the putative "open ocean" sequences (Marston and Sallee, 2003). As more g20 sequence data have accumulated from diverse environments (Zhong et al., 2002; Marston and Sallee, 2003; Dorigo et al., 2004; Short and Suttle, 2005; Sandaa and Larsen, 2006; Wilhelm et al., 2006), it has become clear that marine g20 sequences form nine phylogenetic clusters (first described by Zhong et al. (2002)), and g20 sequences originating from freshwater environments form a separate, tenth cluster (Dorigo et al., 2004; Short and Suttle, 2005; Wilhelm et al., 2006). Three of the 9 marine clusters (clusters I, II, III in Zhong et al. 2002) contain cultured representatives (hereafter called "culture-containing clusters"), whereas the remaining six marine clusters (clusters A-F) and the "freshwater" cluster do not (hereafter called "environmental-sequence-only clusters"). The cultured representatives were isolated using only *Synechococcus* hosts (7 strains = WH7803, WH7805, WH8007, WH8012, WH8018, WH8101, WH8113), which undoubtedly limits the diversity represented considering the larger diversity of *Synechococcus* strains (Rocap et al., 2002; Fuller et al., 2003; Ahlgren and Rocap, 2006) and that the sister genus *Prochlorococcus* is also abundant in open ocean waters. This raises the question: Could these 7 environmental-sequence-only clusters represent novel cyanomyophages that infect this broader diversity of *Synechococcus* host strains, *Prochlorococcus*, or other cyanobacteria?

To address this question, we isolated phages on a broad diversity of *Prochlorococcus* and *Synechococcus* hosts (Table 1), sequenced their g20 homologs, and analyzed their diversity in the context of published PCR-generated sequences from natural populations. We then combined the g20 sequences from these new cultured isolates with all environmental g20 sequences available (including all PCR-generated environmental sequences, as well as primer-independent sequences available in the Global Ocean Survey metagenomic dataset), to examine the broad diversity of g20 observed in the wild. This allowed us to ask: Do any of the new environmental sequences cluster with the previously observed environmental-sequence-only clusters? Furthermore, are g20 sequence clustering patterns ecologically meaningful? Do they reflect the habitat —and by inference the microbial community — of the site from which they were isolated?,

RESULTS AND DISCUSSION

Analysis of g20 diversity captured by several g20 primer sets

As our understanding of marine myoviruses has grown over the years, multiple primer sets have been developed and used to specifically amplify cyanomyophage g20 sequences from field samples (Fuller et al., 1998; Wilson et al., 1999, , 2000; Zhong et al., 2002; Frederickson et al., 2003; Marston and Sallee, 2003; Dorigo et al., 2004; Wang and Chen, 2004; Sandaa and Larsen, 2006; Wilhelm et al., 2006).

Each of these primer sets was designed based on a limited number of sequences from cultured isolates. Thus we wondered how well these primer sets would capture the diversity of g20 sequences in our relatively extensive *Prochlorococcus* and *Synechococcus* cyanophage collection (Table 1).

We found that the CPS4GC/5 primer set (Wilson et al., 1999) amplified g20 sequences from 80% of the cyanomyophages screened (bold entries in Table 1). This primer set, however, amplifies only a small region of this gene (~165 bp), thus its utility for subsequent phylogenetic analyses is limited. In contrast, the CPS1/8 primer set (Zhong et al., 2002), which captures a larger segment of the gene (~594bp), amplified the g20 sequence of only 56% of the cyanomyophages screened (Table 1). Using genome sequence data from two *Prochlorococcus* cyanomyophages (Sullivan et al., 2005) that became available after these primer sets were designed, we modified the CPS1/8 primer set with the hope of amplifying g20 from all of our isolates for use in subsequent phylogenetic analyses. Indeed, the redesigned set (CPS1.1/8.1) captured g20 homologs from all cyanomyophage isolates screened (Table 1). Despite their degeneracy, the redesigned primer set remained specific only for cyanomyophage isolates as inferred from repeatedly negative PCR results against the sipho- and podo- cyanophage, as well as the non-cyanomyophages we examined (Table 1).

Phylogenetic relationships of g20 sequences

We next analyzed how these new g20 sequences from cultured isolates compared to selected sequences (see methods) from the databases (Fig. 1). Randomly paired g20 sequence identities from this dataset ranged from 59-100% amino acid identity, notably with some identical g20 protein sequences observed multiple times (alphanumeric clusters #1-13 in Fig. 1). This is not unprecedented: even at the level of the gene, identical viral sequences have been previously reported from vastly different aquatic environments using two separate gene markers including g20 (Zhong et al., 2002; Marston and Sallee, 2003; Short and Suttle, 2005) and DNA polymerase (Breitbart et al., 2004a; Breitbart and Rohwer, 2005).

In phylogenetic analyses, 40 of 45 g20 sequences from cyanomyophages (38 new, 7 previously published) grouped within the clusters that contain cultured representatives (I, II and III), four fell into a new monophyletic cluster (indicated by 'PSSM9/11/12 new cluster' on Fig. 1), and one (P-ShM1) fell onto a long branch. None fell into the previously defined (by Zhong et al. 2002) environmental-sequence-only clusters A-F, which were thought to be from marine cyanomyophages because of the use of isolate-designed and -tested 'cyanophage-specific primers'. Thus either our phage culture collection is still not diverse enough to represent the g20 diversity of phages that infect marine cyanobacteria, or the sequences in the environmental-sequence-only clusters A-F represent myophages that infect other hosts. Observations made by Short & Suttle (2005) lend support to the latter. They found three g20 sequences in waters 3,246m deep in the Arctic Chukchi Sea, waters unlikely to contain cyanobacteria and their phages, that grouped with Cluster A.

Given our extensive host range information for these cyanobacteria phage-host systems, we examined g20 clustering patterns for relationships with respect to the host strains upon which the phage were isolated or could cross infect. None of the three culture-containing clusters (I, II, III) were comprised solely of g20 sequences from phages with similar hosts (Fig. 1), and no clear-cut patterns emerged when subclusters within these clusters were evaluated. This is consistent with the observations of Stoddard et al. (2007), who recently reported that g20 sequences could not predict the pattern of cross-resistance observed when selecting for cyanophage resistance in *Synechococcus*. Conversely, they also found that *Synechococcus* DNA-dependent RNA polymerase genotypes were not related to phage sensitivities (Stoddard et al. 2007). Thus for the *Prochlorococcus*/*Synechococcus*/myophage system in Fig. 1, it appears that commonly used phage and host genetic markers lack the ability to predict either the range of hosts that a phage can infect, or the range of phages to which a host is susceptible.

We next added more recently published g20 sequences to this analysis, including those from the non-PCR-based Global Ocean Survey (GOS) metagenomics database (Rusch et al., 2007) and all published PCR-based environmental sequences ((Fig. 2, Table 2). Only sequences of sufficient length for phylogenetic analysis were used. The majority (464 of 769) of these environmental sequences, including 401 GOS sequences, grouped in culture-containing clusters I, II and III, First we found that thirteen of

the thirty-eight GOS sample sites included in our analysis lack *Prochlorococcus* and *Synechococcus* (as determined by dot-blots in Rusch et al., 2007), yet 75 g20 sequences from these sites fell into clusters I, II and III (Fig. 2), thought, from earlier studies, to represent myophages that infect marine pico-cyanobacteria. Thus it appears that clusters I, II, and III likely represent phages that infect a diversity of hosts and are not limited to pico-cyanobacteria dominated environments. Second, these analyses revealed that cluster II contains ~10-fold more GOS sequences than clusters I and III (336 vs. 32 and 33, respectively). If we ignore possible cloning bias, this suggests that cluster II sequences are by far the most abundant type in the environments sampled. Third, we note that a relatively tiny number of the GOS sequences fell into the the environmental-sequence-only clusters — clusters A-F in Fig. 1 — that were defined by Zhong et al. (2002) (Fig. 2). The 12 that fell into cluster A originated from 7 sites with different physicochemical characteristics (see color rings, Fig. 2). Even fewer sequences fell into environmental-sequence-only clusters B-F, suggesting that these types of g20 sequences are either extremely rare in the environments sampled to date, or are sequencing artifacts.

This expanded dataset lends support for three additional g20 lineages (Fig. 2). These include 93 sequences that group with the previously identified 'freshwater' cluster (Dorigo et al. 2004, Short & Suttle 2005, Wilhelm et al. 2006, labeled as 'new cluster #1' in Fig. 2), 25 sequences that group with the new culture-containing P-SSM9/11/12 cluster (named after the original phage isolates forming this cluster in Fig. 1, labeled as 'new cluster #2' in Fig. 2), and 84 environmental sequences (74 GOS + 10 non-GOS environmental sequences, labeled as 'new cluster #3' in Fig. 2) of mixed biogeographic and habitat origin that form a new environmental-sequence-only cluster.

Relationship between g20 clusters and habitat

Using Unifrac distance metric statistical tools (Lozupone et al. 2006), we examined the meta-g20 dataset for correlates between sequence clustering and habitat descriptors, such as the microbial community type, temperature and salinity of the original sample. As a first approximation of the microbial community type, we used previously defined environmental categories originally inferred from ribotype dot-blots and metagenomic sequence data (Figs. 9 and 10 in Rusch et al. 2007) for the GOS g20 sequences, then assigned such categories where reasonable assumptions could be made for non-GOS sequences (details in Table 3 legend). We found that the g20 sequence clusters were non-randomly distributed with respect to sequences that originated from freshwater, tropical freshwater, arctic/polar, estuarine, Sargasso, and hypersaline environments, while eight other environments lacked statistically significant clustering (Table 3). Beyond habitat-related properties, we also observed non-random g20 sequence distributions relative to abiotic factors such as salinity (4 of 5 categories significant, Table 4) and temperature (3 of 5 categories significant, Table 5). In both cases the outermost categories (e.g., "cold" and "hot", but not "medium" for temperature) were significantly structured, but median categories were not. Qualitatively, some of these clustering patterns are also evident in the color-coded rings in Figure 2.

Notably, however, clustered sequences, when significantly correlated with a habitat characteristic, always contained exceptions. For example, the 'freshwater' category was one of the most significantly non-random sequence categories (Fig. 2, Table 3-5). In spite of this, the 'freshwater' cluster also contained 6 sequences from brackish waters, while 68 additional freshwater sequences were distributed elsewhere in the tree (light blue in the outer circle in Fig. 2). Similarly, while sequences in the 'tropical freshwater' category were found to be non-randomly distributed (Table 3), this is likely driven by the 24 sequences that form a well defined sub-cluster within cluster II (GOS site 20 sub-cluster in Fig 2). However, another 18 sequences from this same sample are scattered throughout the rest of the tree (11 in cluster II, 4 in cluster I, and 3 in other clusters).

In other words, while some patterns emerge, exceptions are so frequent that one must conclude that the g20 sequence is not a good predictor of the habitat from which the phage originated. This is perhaps not surprising given the sheer abundance of phages on the planet (10^{31} phages) and the apparent promiscuity of viral-host interactions allow a lot of 'rule breakers' to persist. For example, not only can viral particles survive the physical challenges of extreme environmental shifts (Breitbart et al., 2004b),

but also viruses from one environment (e.g., freshwater Great Lakes) are readily capable of infecting hosts from another environment (e.g., oceanic *Synechococcus*; (Wilhelm et al., 2006). Further, in coliphage T4, the g20 gene encodes a portal protein (Marusich and Mesyanzhinov, 1989) involved in functions quite removed from the direct interaction between phage and host. In contrast, the distal tail fiber gene is known to be the direct determinant of host range in T-even coliphages (Henning and Hashemolhosseini, 1994; Tetart et al., 1998). Thus, g20 sequence patterns might no longer correlate to host range at the fine scales (e.g., cyanobacteria and their phages) where host range 'jumps' could more commonly occur (e.g., by simple tail-fiber-switching *sensu* Tetart et al. 1998) that would de-couple host properties from vertically evolved g20 sequence lineages.

Concluding remarks

Taken together, these data reveal that oceanic phage g20 sequence clustering patterns are, at a fine level (e.g., cyanobacteria-cyanophages), largely uncorrelated to host factors. As one zooms out to more generally consider the relationship between g20 sequences from the wild and the habitat characteristics from which they were collected, we find that they are non-randomly distributed, reflecting in some cases a connection between habitat properties, microbial community structure, and phage community composition as defined by the g20 gene. We posit that the latter patterns, when evident, reflect host-range-limited vertical evolution of g20 sequences, while the former reflects highly specific 'tip-of-the-tree' phage-host interactions that are evolutionarily disconnected from that of the g20 protein product.

ACKNOWLEDGEMENTS

This research was supported in part by funding from NSF (CMORE contribution # ...), DOE, The Seaver Foundation and the Gordon and Betty Moore Foundation Marine Microbiology Program to SWC an NIH Bioinformatics Training Grant supported MBS; MIT Undergraduate Research Opportunities Program supported VQ, JAL, GT, RF and JER; Howard Hughes Medical Institute funded MIT Biology Department Undergraduate Research Opportunities Program supported ASD; NSERC (Canada) Discovery Grant (DG 298394) and a Grant from the Canadian Foundation for Innovation (NOF10394) to JPB; NSF Graduate Fellowship funding supported MLC. F. Chen and M. Marston kindly provided phage isolates used in testing of PCR primer sets. MBS thanks C. Lozupone and M. Hamady for interpretive and technical support using Unifrac, as well as F. Chen, M. Marston, U. Dorigo, J. Waterbury and S. Wilhelm for providing unpublished meta-data for published g20 sequences to make the meta-g20 dataset analyses as comprehensive as possible. The comments of V. Rich greatly improved the manuscript.

MATERIALS AND METHODS

Phage isolates. Forty five cyanomyophages were isolated (Table 1) as described previously (Waterbury and Valois, 1993; Wilson et al., 1993; Marston and Sallee, 2003; Sullivan et al., 2003). S-PM2 and S-WHM1 were provided by W. Wilson and all S-RIM phages were provided by M. Marston. The specificity of cyanomyophage g20 primers was tested using five marine *Pseudoalteromonas* spp. bacteriophages (HER320, HER321, HER322, HER327, HER328; (Wichels et al., 1998) that were purchased from the Felix d'Herelle Reference Center for Bacterial Viruses (contact H. Ackermann) as well as 7 heterotrophic bacteriophages (IH6-ϕ1, IH6-ϕ7, IH11-ϕ2, IH11-ϕ5, CB8-ϕ2, CB8-ϕ6, CB-ϕ8; (Zhong et al., 2002) kindly provided by F. Chen.

Primer redesign. To obtain g20 PCR amplicons from myophage that would not amplify using published primers, we added degeneracies to both CPS1 and CPS8, and shifted the CPS8 primer based upon genomic sequence data from two *Prochlorococcus* myophage isolates, P-SSM2 and P-SSM4 (Sullivan et al. 2005) to design CPS1.1 5'-GTAGWATWTTYTAYATTGAYGTWGG-3' and CPS8.1 5'-ARTAYTTDCCDAYRWAWGGWTC-3'.

PCR amplification and sequencing. Previous g20 PCR primer sets (non-degenerate CPS4GC/CPS5 (Wilson et al., 1999) and degenerate CPS1/CPS8 (Fuller et al., 1998; Zhong et al., 2002) were designed to amplify ~200bp and ~592 bp fragments, respectively, of the T4 g20 homologue in myophages.

PCR reactions for CPS4GC/CPS5 and CPS1/CPS8 were conducted as described previously (Wilson et al., 1999; Zhong et al., 2002). Briefly, 2 μl of cyanophage lysate was added as DNA template to a PCR reaction mixture (total volume 50 μl) containing the following: 20 pmol each of a forward and reverse primer, 1x PCR buffer (50mM Tris-HCl, 100 mM NaCl, 1.5 mM $MgCl_2$), 250 μM of each dNTP, and 0.75 U of Expand High Fidelity DNA polymerase (Roche, Indianapolis, IN). PCR amplification was carried out with a PTC-100 DNA Engine Thermocycler (MJ Research, San Francisco, CA). Optimized thermal cycling conditions varied slightly from those reported as follows: CPS4GC/CPS5 required an initial denaturation step of 94°C for 3 minutes, followed by 35 cycles of denaturation at 94°C for 1 minute, annealing at 50°C for 1 min, ramping at 0.3°C/s, and elongation at 73°C for 1 minute with a final elongation step at 73°C for 4 minutes, whereas both primer sets CPS1/CPS8 and CPS1.1/CPS8.1 required an initial denaturation step of 94°C for 3 minutes, followed by 35 cycles of denaturation at 94°C for 15s, annealing at 35°C for 1 min, ramping at 0.3°C/s, and elongation at 73°C for 1 minute with a final elongation step at 73°C for 4 minutes. Systematic PCR screening using various primer sets was conducted using the same PCR reaction conditions and amplification protocol, but replacing the High Fidelity DNA polymerase with the less expensive Taq DNA polymerase (Invitrogen, Carlsbad, CA) and only using 20 μl reactions since replicate (range 3-8) PCR reactions were pooled before sequencing to decrease PCR bias (Polz and Cavanaugh, 1998). In all cases, a 5-10 μl aliquot of PCR product was analyzed in a 1.5% TAE gel stained with EtBr. The gel image was captured and analyzed with an Eagle Eye II gel documentation system (Stratagene, La Jolla, CA). For purification and sequencing, replicate PCR reactions were combined, run out on a 1.5% TAE gel and purified using the QIAGEN QIAquick gel extraction kit (Qiagen, Valencia, CA). The purified PCR products were sequenced directly on both strands using the degenerate PCR primers used to obtain the product (CPS1, CPS8, CPS1.1, CPS8.1) with best results at primer concentrations ~10-fold those suggested by the sequencing facility (40 pmol per reaction). To have greater confidence in negative PCR results, templates that did not produce amplified product were tested against optimized primer sets multiple times (data not shown). To confirm that our correctly-sized amplicons from "positive" PCR reactions were in fact g20 sequences, we sequenced the products. In all cases, the amplicon sequences were from g20 homologues

Where identical g20 sequences were observed in our study, we confirmed the match was real and not the result of PCR contamination by re-amplifying and sequencing directly from fresh phage isolates (e.g., for P-SSM4, P-RSM3, S-SSM2, and "Syn" phages Syn2, Syn9, Syn10, Syn26, Syn30, Syn33, Syn1, Syn19) – many of which were obtained from stocks kept at a separate institution.

Phylogenetic analysis. For the new sequences presented in Fig. 1 of this study, paired sequence data were aligned using ClustalW (Thompson et al., 1997) and corrected manually using the sequence chromatograms. Consensus sequences for each cyanophage isolate were then translated in-frame into amino acids. Published g20 sequences from PCR-amplified environmental clone libraries and phage isolates were screened by building preliminary neighbor-joining trees to select representative sequences that spanned the known g20 diversity and added to this dataset. Multiple sequence alignments of translated amino acid consensus sequences were done with ClustalW using the Gonnet protein weight matrix, a gap opening penalty of 15 and gap extension penalty of 0.30 (although changing these penalties did not significantly alter the alignments). Phylogenetic reconstruction was done using PAUP 4.0 (Swofford, 2002) for parsimony and distance trees and Tree-Puzzle 5.0 (Schmidt et al., 2002) for maximum likelihood trees. Evolutionary distances for neighbor-joining trees were calculated based on mean character distances, while evolutionary distances for maximum likelihood trees were calculated using the JTT model of substitution assuming a gamma-distributed model of rate heterogeneities with 16 gamma-rate categories empirically estimated from the data. A heuristic search with 10 random addition replicates using the tree-bisection-reconnection branch swapping algorithm was used for parsimony trees.

Bootstrap analysis was used to estimate node reproducibility and tree topology for neighbor-joining (1,000 replicates) and parsimony (100 replicates) trees, while quartet puzzling (10,000 replicates) indicates support for the maximum likelihood tree. The g20 sequence from coliphage T4 was used as the outgroup taxon for all analyses.

Phylogenetic analyses of 183 amino acids from viral g20 sequence from 79 taxa yielded robust, similar trees using both algorithmic (neighbor-joining) and tree-searching (parsimony and maximum likelihood) methods. The translated g20 sequences contained phylogenetically informative regions (e.g., for parsimony analyses, 41 positions were constant, 25 were parsimony uninformative and 117 were parsimony informative). Differences between the parsimony, distance and maximum likelihood trees were limited to the branching order of the terminal nodes in a given cluster. To evaluate whether g20 sequence diversity correlated to the host-related properties presented in Figure 1, we empirically defined a "well supported node" as one where the average support across all three phylogenetic methods was 80% or greater.

GOS g20 identification, filtering and phylogenetic analyses: Using the 549 bp g20 fragment from all available cultured isolates as queries (Table 1), we retrieved 553 sequence reads with similarity (bit score > 100) to this region of the g20 gene from the Global Ocean Survey (GOS) databases (downloaded from http://camera.calit2.net/ 1 May 07), then combined these GOS sequences with available published g20 sequences. The combined sequences were aligned using Clustal X and filtered to remove short, phylogenetically uninformative sequences, as well as sequences with poor quality at the ends. This manual curation left 769 total sequences (512 GOS sequences, details in Table 2) with 554 aligned nucleotide positions. 11 maximum likelihood trees were generated using GARLI (Zwickl, 2006), starting from a neighbor joining topology calculated in PAUP v4b10 (Swofford, 2002). Tree searching was terminated after 100,000 generations with no significantly better scoring topology, and a score improvement threshold for termination of 0.05. Topology mutation proportions were 0.1-0.2 NNI (nearest neighbor interchange) and 0.8-0.9 limited SPR (subtree pruning-regrafting), with the maximum SPR range of 8-10 branches. From the 11 resulting trees, a majority-rule consensus tree (threshold 50% agreement) was generated in PAUP and is presented in Figure 2.

Statistical analyses to evaluate whether g20 clustering patterns uncovered in the phylogenetic reconstructions were related to the habitat features of the original sample (e.g., microbial community type, temperature and salinity) were carried out using the Unifrac distance metric statistical tools available at http://bmf2.colorado.edu/unifrac/index.psp (Lozupone and Knight, 2005). The database and the treefile used for the analysis is provided in Supplementary Information (Suppl. Files 1-2). Briefly, all g20 sequences were assigned to environmental categories using metadata for each sequence, with some assumptions made as described in Table 3 legend. Missing metadata for published g20 sequences were obtained where possible from the authors of the original work, as indicated in Tables 4 and 5. The patterns of these metadata were evaluated for 'each environment separately' in the context of a single neighbor-joining tree that included branch lengths (Suppl. File 2) using Unifrac; all statistical results were similar using the P-test (also available at the Unifrac site, data not shown).

Nucleotide sequence accession numbers. The nucleotide sequences determined in this study were submitted to GenBank and assigned accession numbers AYXXXXXX to AYXXXXXX.

REFERENCES

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

Bergh, O. (1989) High abundance of viruses found in aquatic environments. *Nature* **340**: 467-468.

Bratbak, G., Heldal, M., Norland, S., and Thingstad, T.F. (1990) Viruses as partners in spring bloom microbial trophodynamics. *Appl. Environ. Microbiol.* **56**: 1400-1405.

Breitbart, M., and Rohwer, F. (2005) Here a virus, there a virus, everywhere the same virus? *Trends in Microbiology* **13**: 278-284.

Breitbart, M., Miyake, J.H., and Rohwer, F. (2004a) Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiology Letters* **236**: 249-256.

Breitbart, M., Wegley, L., Leeds, S., Schoenfeld, T., and Rohwer, F. (2004b) Phage community dynamics in hot springs. *Appl. Environ. Microbiol.* **70**: 1633-1640.

Breitbart, M., Felts, B., Kelley, S., Mahaffy, J.M., Nulton, J., Salamon, P., and Rohwer, F. (2004c) Diversity and population structure of a near-shore marine-sediment viral community. *Proc R Soc Lond B Biol Sci* **271**: 565-574.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D. et al. (2002) Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A* **99**: 14250-14255.

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**: 1768-1770.

Coombs, D., and Eiserling, F.A. (1977) Studies on the structure, protein composition and assembly of the neck of bacteriophage T4. *J. Molecular Biology* **116**: 375-407.

DeLong, E.F., Preston, C.M., Mincer, T., Rich, V., Hallam, S.J., Frigaard, N.U. et al. (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**: 496-503.

Dorigo, U., Jacquet, S., and Humbert, J.-F. (2004) Cyanophage diversity, inferred from g20 gene analyses, in the Largest Natural Lake in France, Lake Bourget. *Appl. Environ. Microbiol.* **70**: 1017-1022.

Frederickson, C.M., Short, S.M., and Suttle, C.A. (2003) The physical environment affects cyanophage communities in British Columbia Inlets. *Microbial Ecology* **46**: 348-357.

Fuller, N.J., Wilson, W.H., Joint, I.R., and Mann, N.H. (1998) Occurrence of a sequence in marine cyanophages similar to that of T4 g20 and its application to PCR-based detection and quantification techniques. *Appl. Environ. Microbiol.* **64**: 2051-2060.

Fuller, N.J., Marie, D., Partensky, F., Vaulot, D., Post, A.F., and Scanlan, D.J. (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.

Hambly, E., Tetart, F., Desplats, C., Wilson, W.H., Krisch, H.M., and Mann, N.H. (2001) A conserved genetic module that encodes the major virion components in both the coliphage T4 and the marine cyanophage S-PM2. *Proc Natl Acad Sci U S A* **98**: 11411-11416.

Hendrix, R.W., Smith, M.C., Burns, R.N., Ford, M.E., and Hatfull, G.F. (1999) Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc Natl Acad Sci U S A* **96**: 2192-2197.

Henning, U., and Hashemolhosseini, S. (1994) Recepor recognition by T-even-type coliphages. In *Molecular biology of bacteriophage T4*. Karam, J. (ed). Washington D.C.: ASM Press, pp. 291-298.

Hsiao, C.L., and Black, L.W. (1978) Head morphogenesis of bacteriophage T4. III. The role of g20 in DNA packaging. *Virology* **91**: 26-38.

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**: 11013-11018.

Lozupone, C., and Knight, R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71**: 8228-8235.

Lu, J., Chen, F., and Hodson, R.E. (2001) Distribution, isolation, host specificity, and diversity of cyanophages infecting marine *Synechococcus* spp. in river estuaries. *Appl. Environ. Microbiol.* **67**: 3285-3290.

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

Marston, M.F., and Sallee, J.L. (2003) Genetic diversity and temporal variation in the cyanophage community infecting marine *Synechococcus* species in Rhode Island's coastal waters. *Appl. Environ. Microbiol.* **69**: 4639-4647.

Marusich, E.I., and Mesyanzhinov, V.V. (1989) Nucleotide and deduced amino acid sequence of bacteriophage T4 gene 20. *Nucleic Acids Research* **17**: 7514.

Mühling et al., (2005) Genetic diversity of marine Synechococcus and co-occurring cyanophage communities: evidence for viral control of phytoplankton. *Environ. Microbiol.* **7:** 499-508.

Partensky, F., Hess, W.R., and Vaulot, D. (1999) *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiology and Molecular Biology Reviews* **63**: 106-127.

Paul, J.H., Sullivan, M.B., Segall, A.M., and Rohwer, F. (2002) Marine phage genomics. *Comp Biochem Physiol B Biochem Mol Biol* **133**: 463-476.

Polz, M.F., and Cavanaugh, C.M. (1998) Bias in template-to-product ratios in multitemplate PCR. *Appl. Environ. Microbiol.* **64**: 3724-3730.

Proctor, L.M., and Fuhrman, J.A. (1990) Viral mortality of marine bacteria and cyanobacteria. *Nature* **343**: 60-62.

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**: 1180-1191.

Rohwer, F., and Edwards, R. (2002) The Phage Proteomic Tree: a genome-based taxonomy for phage. *J. Bacteriology* **184**: 4529-4535.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S. et al. (2007) The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biology* **5**: e77.

Sandaa, R.A., and Larsen, A. (2006) Seasonal variations in virus-host populations in Norwegian coastal waters: Focusing on the cyanophage community infecting marine *Synechococcus* spp. *Appl. Environ. Microbiol.* **72**: 4610-4618.

Schmidt, H.A., Strimmer, K., Vingron, M., and von Haeseler, A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**: 502-504.

Short, C.M., and Suttle, C.A. (2005) Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. *Appl. Environ. Microbiol.* **71**: 480-486.

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**:5516-22.

Sullivan, M.B., Waterbury, J.B., and Chisholm, S.W. (2003) Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. *Nature* **424**: 1047-1051.

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

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 Biology* **4**: e234.

Suttle, C.A. (2000) Cyanophages and their role in the ecology of cyanobacteria. In *The Ecology of Cyanobacteria*. Whitton, B.A., and Potts, M. (eds). Netherlands: Kluwer Academic Publishers, pp. 563-589.

Suttle, C.A., and Chan, A.M. (1993) Marine cyanophages infecting oceanic and coastal strains of *Synechococcus*: abundance, morphology, cross-infectivity and growth characteristics. *Marine Ecological Progress Series* **92**: 99-109.

Suttle, C.A., and Chan, A.M. (1994) Dynamics and distribution of cyanophages and their effects on marine *Synechococcus* spp. *Appl. Environ. Microbiol.* **60**: 3167-3174.

Swofford, D.L. (2002) *PAUP*. Phylogenetic analysis using parsimony (*and other methods), version 4.* Sunderland, MA: Sinauer Associates.

Tetart, F., Desplats, C., and Krisch, H.M. (1998) Genome plasticity in the distal tail fiber locus of the T-even bacteriophage: recombination between conserved motifs swaps adhesin specificity. *J. Molecular Biology* **282**: 543-556.

Thingstad, T.F. (2000) Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic ecosystems. *Limnol. Oceanogr.* **45**: 1320-1328.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* **25**: 4876-4882.

van Driel, R., and Couture, E. (1978) Assembly of the scaffolding core of bacteriophage T4 proheads. *J. Molecular Biology* **123**: 713-719.

Wang, K., and Chen, F. (2004) Genetic diversity and population dynamics of cyanophage communities in the Chesapeake Bay. *Aquatic Microbial Ecology* **34**: 105-116.

Waterbury, J.B., and Valois, F.W. (1993) Resistance to co-occurring phages enables marine *Synechococcus* communities to coexist with cyanophage abundant in seawater. *Appl. Environ. Microbiol.* **59**: 3393-3399.

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.

Wichels, A., Biel, S.S., Gelderblom, H.R., Brinkhoff, T., Muyzer, G., and Schutt, C. (1998) Bacteriophage diversity in the North Sea. *Appl. Environ. Microbiol.* **64**: 4128-4133.

Wiggins, B.A., and Alexander, M. (1985) Minimum bacterial density for bacteriophage replication: implications for significance of bacteriophages in natural ecosystems. *Appl. Environ. Microbiol.* **49**: 19-23.

Wilhelm, S.W., Carberry, M.J., Eldridge, M.L., Poorvin, L., Saxton, M.A., and Doblin, M.A. (2006) Marine and freshwater cyanophages in a Laurentian Great Lake: Evidence from infectivity assays and molecular analyses of g20 genes. *Appl. Environ. Microbiol.* **72**: 4957-4963.

Wilson, W.H., Joint, I.R., Carr, N.G., and Mann, N.H. (1993) Isolation and molecular characterization of five marine cyanophages propogated on *Synechococcus* sp. strain WH 7803. *Appl. Environ. Microbiol.* **59**: 3736-3743.

Wilson, W.H., Fuller, N.J., Joint, I.R., and Mann, N.H. (1999) Analysis of cyanophage diversity and population structure in a south-north transect of the Atlantic Ocean. In *Marine Cyanobacteria.* Sharpy, L., and Larkum, A.W.D. (eds). Monaco: Bulletin de l'Institut oceanographique, pp. 209-216.

Wilson, W.H., Fuller, N.J., Joint, I.R., and Mann, N.H. (2000) Analysis of cyanophage diversity in the marine environment using denaturing gradient gel electrophoresis. In *Microbial Biosystems: New Fronteirs. Proceedings of the 8th International Symposium on Microbial Ecology.* Bell, C.R., Brylinksky, M., and Johnson-Green, P. (eds). Halifax, Nova Scotia, Canada: Atlantic Canada Society for Microbial Ecology.

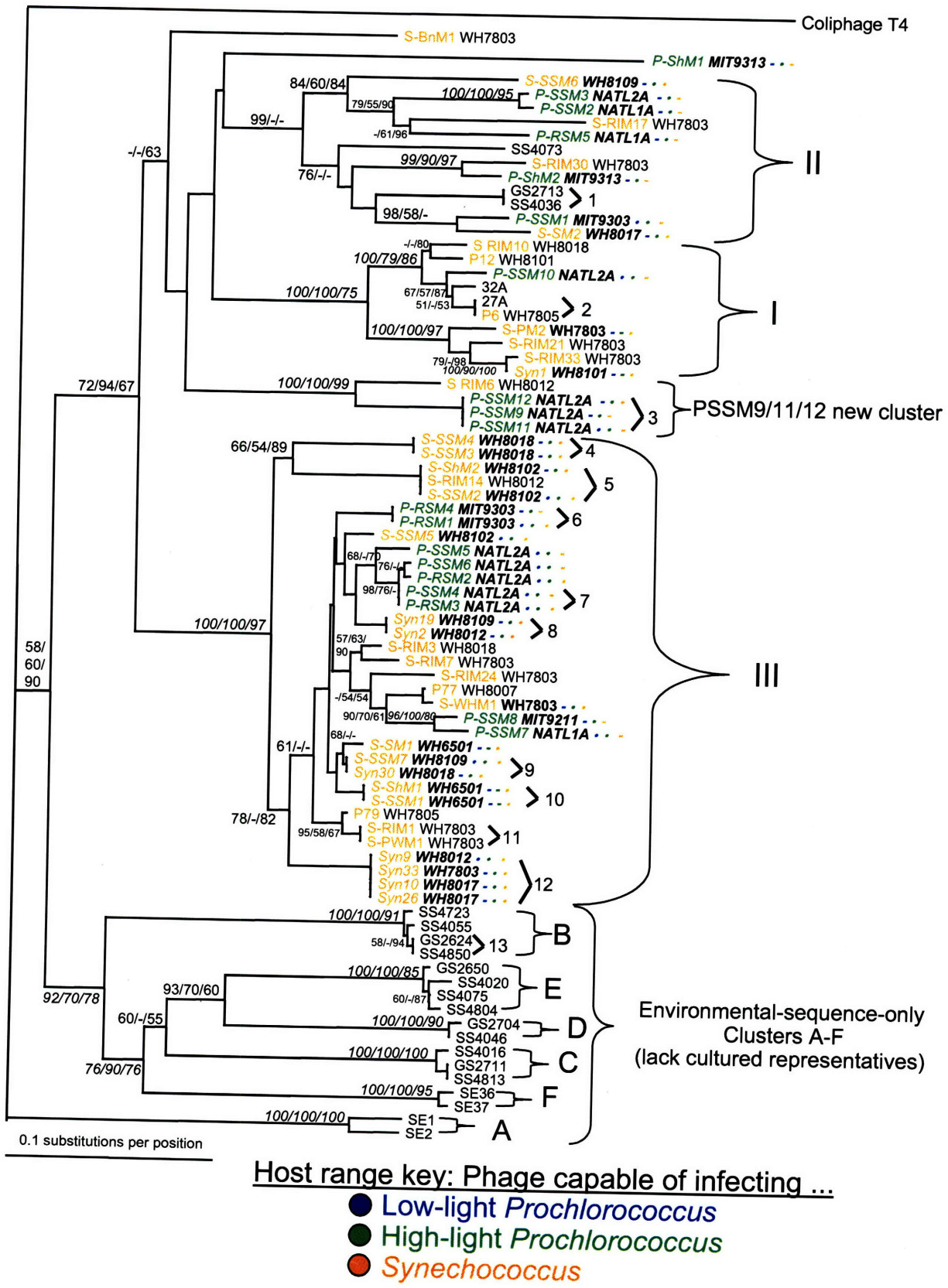
Zhong, Y., Chen, F., Wilhelm, S.W., Poorvin, L., and Hodson, R.E. (2002) Phylogenetic diversity of marine cyanophage isolates and natural virus communities as revealed by sequences of viral capsid assembly protein gene g20. *Appl. Environ. Microbiol.* **68**: 1576-1584.

Zwickl, D.J. (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. PhD Thesis, University of Texas at Austin.

Figure 1. Evolutionary relationships determined using 183 amino acids of the portal protein gene (g20) amplified from cultured phage isolates (names begin with "S-" or "P-" and are colored orange or green for *Synechococcus* or *Prochlorococcus* phages respectively) from this study (*italicized*), as well as previous studies (non-italicized), and environmental g20 sequences (names in black)(Zhong et al., 2002; Marston and Sallee, 2003). Clusters defined by Zhong et al. (2002) are as follows: clusters I, II and III contain g20 sequences from cultured phage isolates, while clusters A-F represent only environmental g20 sequences. Clusters containing identical g20 protein sequences are numbered with alphanumeric numbers (1-13). For cultured phages, the phage isolate names are followed by black lettering that indicates the original host strain used for isolation, while the phage host range is indicated as high-light adapted *Prochlorococcus* (green circle or dash), low-light adapted *Prochlorococcus* (blue circle or dash) or *Synechococcus* (orange circle or dash). The circles represent cross-infection was observed within this group of hosts tested, whereas a dash indicates that no cross-infection was observed. Isolates not available for host range testing have no indication of their host range. The tree shown was inferred by neighbor-joining as described in the methods. Support values shown at the nodes are neighbor-joining bootstrap / maximum parsimony bootstrap / maximum likelihood quartet puzzling support (only values >50 are shown). Well-supported nodes (as defined in methods) are designated by italicized support values, including 6 nodes that represent sub-clusters within the culture-containing clusters I, II and III. The g20 sequence from the non-cyanomyophage isolate T4 was used as an outgroup to root this tree.

Figure 2. Evolutionary relationships determined using 554 base pairs of the portal protein gene (g20) from 769 available g20 sequences. Clusters defined by Zhong et al. (2002) are identified as culture-based clusters I,II,II and environmental-sequence-only clusters A-F. New clusters defined since Zhong et al. (2002) are indicated with the preface 'new cluster', a number and a brief description. The tree shown is the consensus (majority rules) tree from 11 GARLI iterations inferred using the maximum likelihood criterion (see methods), with the Aeromonas phage Aeh1 g20 sequence used as an outgroup to root the tree. Three color rings reflect the habitat type from which the g20 sequence originated. For most of these sequences (Global Ocean Survey sequences) there is ribotype dot-blot and metagenomic information about the microbial community structure at the site, while for non-GOS sequences such information was assumed where reasonable to do so (see Table 3 legend). The inner ring is this microbial community structure information listed as Rusch et al. (2007) defined environmental categories, while the other two rings reflect the temperature and salinity of the original sampling site.

Fig. 1



Host range key: Phage capable of infecting ...

- Low-light *Prochlorococcus*
- High-light *Prochlorococcus*
- *Synechococcus*

re 2

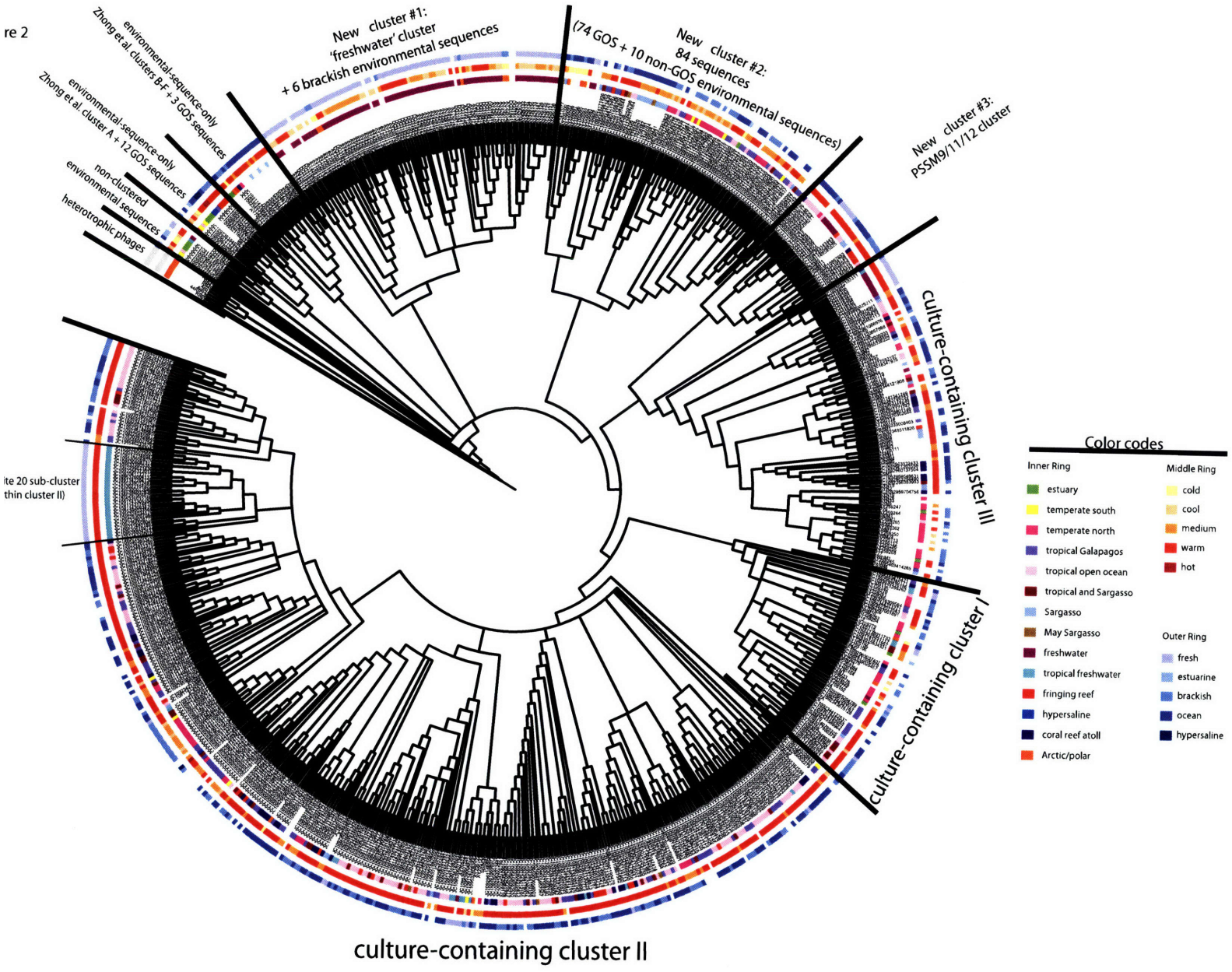

New cluster #1:
'freshwater' cluster
+ 6 brackish environmental sequences
(74 GOS + 10 non-GOS environmental sequences)
New cluster #2:
84 sequences
New cluster #3:
PSSM9/11/12 cluster
environmental-sequence-only
Zhong et al. clusters B-F + 3 GOS sequences
environmental-sequence-only
Zhong et al. cluster A + 12 GOS sequences
non-clustered
environmental sequences
heterotrophic phages
culture-containing cluster III
culture-containing cluster I
culture-containing cluster II
ite 20 sub-cluster
thin cluster II)
Color codes
Inner Ring
estuary
temperate south
temperate north
tropical Galapagos
tropical open ocean
tropical and Sargasso
Sargasso
May Sargasso
freshwater
tropical freshwater
fringing reef
hypersaline
coral reef atoll
Arctic/polar
Middle Ring
cold
cool
medium
warm
hot
Outer Ring
fresh
estuarine
brackish
ocean
hypersaline

Table 1. Efficacy of three different primer sets at amplifying the *g20* gene from cultured cyanophage. "+" indicates positive PCR amplification; "–" indicates that there was no PCR product of the expected size. The new g20 sequences contributed in this study are shown in bold letters. CPS1.1/8.1 is the new primer set designed for this study, while CPS4GC/5 and CPS1/8 were published previously.

Phage Strain	Original host strain isolated on	Site of Isolation	Depth (m)	Date isolated	Family[a]	g20 primer set			Ref[b]
						CPS 4GC/5	CPS 1/8	CPS 1.1/8.1	
Prochlorococcus cyanophage									
P-SSP1	MIT 9215	BATS / 31°48'N, 64°16'W	100	6 June 2000	P	–	–	–	1
P-RSP1	MIT 9215	Red Sea / 29°28'N, 34°53'E	0	15 July 2000	P	–	–	–	1
P-RSP2	MIT 9302	Red Sea / 29°28'N, 34°53'E	0	15 July 2000	P	–	–	–	1
P-SSP2	MIT 9312	BATS / 31°48'N, 64°16'W	120	29 Sep 1999	P	–	–	–	1
P-SSP3	MIT 9312	BATS / 31°48'N, 64°16'W	100	29 Sep 1999	P	–	–	–	1
P-SSP4	MIT 9312	BATS / 31°48'N, 64°16'W	70	26 Sep 1999	P	–	–	–	1
P-SSP5	MIT 9515	BATS / 31°48'N, 64°16'W	120	29 Sep 1999	P	–	–	–	1
P-SSP6	MIT 9515	BATS / 31°48'N, 64°16'W	100	26 Sep 1999	P	–	–	–	1
P-SSP7	MED4	BATS / 31°48'N, 64°16'W	100	26 Sep 1999	P	–	–	–	1
P-GSP1	MED4	Gulf Stream / 38°21'N, 66°49'W	40	6 Oct 1999	P	–	–	–	1
P-SSP8	NATL2A	BATS / 31°48'N, 64°16'W	100	26 Sep 1999	P	–	–	–	1
P-RSP3	NATL2A	Red Sea / 29°28'N, 34°55'E	50	13 Sep 2000	P	–	–	–	1
P-SP1	SS120	Slope / 38°10'N, 73°09'W	83	17 Sep 2001	P	–	–	–	1
P-SSM8	**MIT 9211**	**W Sargasso Sea / 34°24'N, 72°03'W**	**30**	**22 Sept 2001**	**M**	+	+	+	**2**
P-SSM1	**MIT 9303**	**BATS / 31°48'N, 64°16'W**	**100**	**6 June 2000**	**M**	+	–	+	**1**
P-RSM1	**MIT 9303**	**Red Sea / 29°28'N, 34°53'E**	**0**	**15 July 2000**	**M**	+	–	+	**1**
P-RSM4	**MIT 9303**	**Red Sea / 29°28'N, 34°55'E**	**130**	**13 Sep 2000**	**M**	+	+	+	**2**
P-ShM1	**MIT 9313**	**Shelf / 39°60'N, 71°48'W**	**40**	**16 Sep 2001**	**M**	+	–	+	**1**
P-ShM2	**MIT 9313**	**Shelf / 39°60'N, 71°48'W**	**0**	**16 Sep 2001**	**M**	–	–	+	**1**
P-SSM2	**NATL1A**	**BATS / 31°48'N, 64°16'W**	**100**	**6 June 2000**	**M**	+	+	+	**1**
P-RSM5	**NATL1A**	**Red Sea / 29°28'N, 34°55'E**	**130**	**13 Sep 2000**	**M**	+	+	+	**2**
P-SSM7	**NATL1A**	**BATS / 31°48'N, 64°16'W**	**120**	**29 Sep 1999**	**M**	–	–	+	**2**
P-SSM3	**NATL2A**	**BATS / 31°48'N, 64°16'W**	**100**	**6 Jun 2000**	**M**	–	–	+	**1**
P-SSM4	**NATL2A**	**BATS / 31°48'N, 64°16'W**	**10**	**6 June 2000**	**M**	–	–	+	**1**
P-SSM5	**NATL2A**	**BATS / 31°48'N, 64°16'W**	**15**	**26 Sep 1999**	**M**	+	–	+	**1**
P-SSM6	**NATL2A**	**BATS / 31°48'N, 64°16'W**	**40**	**29 Sep 1999**	**M**	–	–	+	**1**
P-RSM2	**NATL2A**	**Red Sea / 29°28'N, 34°55'E**	**50**	**13 Sep 2000**	**M**	+	–	+	**1**
P-RSM3	**NATL2A**	**Red Sea / 29°28'N, 34°55'E**	**50**	**13 Sep 2000**	**M**	–	–	+	**1**
P-SSM9	**NATL2A**	**W Sargasso Sea / 34°24'N, 72°03'W**	**0**	**22 Sep 2001**	**M?**	+	–	+	**2**
P-SSM10	**NATL2A**	**W Sargasso Sea / 34°24'N, 72°03'W**	**0**	**22 Sep 2001**	**M?**	+	–	+	**2**
P-SSM11	**NATL2A**	**W Sargasso Sea / 34°24'N, 72°03'W**	**0**	**22 Sep 2001**	**M?**	+	–	+	**2**
P-SSM12	**NATL2A**	**W Sargasso Sea / 34°24'N, 72°03'W**	**95**	**22 Sep 2001**	**M?**	+	–	+	**2**
Synechococcus cyanophage									
Syn5	WH 8109	Sargasso Sea / 36°58'N, 73°42'W	0	Dec 1990	P	–	–	–	1
Syn12	WH 8017	Gulf Stream / 34°06'N, 61°01'W	0	July 1990	P	–	–	–	1
S-SM1	**WH 6501**	**Slope / 38°10'N, 73°09'W**	**0**	**17 Sep 2001**	**M**	–	–	+	**1**
S-ShM1	**WH 6501**	**Shelf / 39°60'N, 71°48'W**	**0**	**16 Sep 2001**	**M**	+	+	+	**1**
S-SSM1	**WH 6501**	**W Sargasso Sea / 34°24'N, 72°03'W**	**70**	**22 Sep 2001**	**M**	+	+	+	**1**
Syn 2	**WH 8012**	**Sargasso Sea / 34°06'N, 61°01'W**	**0**	**July 1990**	**M**	–	+	+	**3**
Syn 9	**WH 8012**	**Woods Hole / 41°31'N, 71°40'W**	**0**	**Oct 1990**	**M**	+	+	+	**3**
Syn 10	**WH 8017**	**Gulf Stream / 36°58'N, 73°42'W**	**0**	**Dec 1990**	**M**	+	+	+	**3**
Syn 26	**WH 8017**	**NE Providence Channel / 25°53'N, 77°34'W**	**0**	**Jan 1992**	**M**	+	+	+	**3**
S-SM2	**WH 8017**	**Slope / 38°10'N, 73°09'W**	**15**	**17 Sep 2001**	**M**	+	–	+	**2**
Syn30	**WH 8018**	**NE Providence Channel / 25°53'N, 77°34'W**	**0**	**Jan 1992**	**M**	+	–	+	**3**
S-SSM3	**WH 8018**	**W Sargasso Sea / 34°24'N, 72°03'W**	**0**	**22 Sep 2001**	**M**	+	+	+	**2**
S-SSM4	**WH 8018**	**W Sargasso Sea / 34°24'N, 72°03'W**	**110**	**22 Sep 2001**	**M**	+	+	+	**2**
S-RIM3	WH 8018	Mt. Hope Bay, RI / 41°39'N, 71°15'W	0	Sept. 1999	M?	+	–	+	4
Syn 33	**WH 7803**	**Gulf Stream / 25°51'N, 79°26'W**	**0**	**Jan 1995**	**M**	+	+	+	**3**
S-PM2	WH 7803	English Channel / 50°18'N, 4°12'W	0	23 Sep 1992	M	+	+	+	5
S-WHM1	WH 7803	Woods Hole / 41°31'N, 71°40'W	0	11 Aug 1992	M	+	+	+	5
S-RIM9	WH 7803	Mt. Hope Bay, RI / 41°39'N, 71°15'W	0	May 2000	M?	+	–	+	4

Phage Strain	Original host strain isolated on	Site of Isolation	Depth (m)	Date isolated	Family[a]	g20 primer set			Ref[b]
						CPS 4GC/5	CPS 1/8	CPS 1.1/8.1	
S-RIM17	WH 7803	Mt. Hope Bay, RI / 41°39'N, 71°15'W	0	July 2001	M?	+	–	+	4
S-RIM24	WH 7803	Mt. Hope Bay, RI / 41°39'N, 71°15'W	0	Dec 2001	M?	+	–	+	4
S-RIM30	WH 7803	Mt. Hope Bay, RI / 41°39'N, 71°15'W	0	June 2002	M?	+	–	+	4
Syn 1	**WH 8101**	**Woods Hole / 41°31'N, 71°40'W**	**0**	**Aug 1990**	**M**	**+**	**–**	**+**	**3**
S-ShM2	**WH 8102**	**Shelf / 39°60'N, 71°48'W**	**0**	**16 Sep 2001**	**M**	**+**	**+**	**+**	**1**
S-SSM2	**WH 8102**	**W Sargasso Sea / 34°24'N, 72°03'W**	**0**	**22 Sep 2001**	**M**	**+**	**+**	**+**	**1**
S-SSM5	**WH 8102**	**W Sargasso Sea / 34°24'N, 72°03'W**	**95**	**22 Sep 2001**	**M**	**+**	**+**	**+**	**2**
Syn 19	**WH 8109**	**Sargasso Sea / 34°06'N, 61°01'W**	**0**	**July 1990**	**M**	**–**	**–**	**+**	**3**
S-SSM6	**WH 8109**	**W Sargasso Sea / 34°24'N, 72°03'W**	**70**	**22 Sep 2001**	**M**	**+**	**+**	**+**	**2**
S-SSM7	**WH 8109**	**W Sargasso Sea / 34°24'N, 72°03'W**	**95**	**22 Sep 2001**	**M**	**+**	**+**	**+**	**2**
Other phages									
IH6-ϕ1	IH6	Inner Harbor, Baltimore, MD	0	17 Nov 2000	M	–	–	–	6
IH6-ϕ7	IH6	Inner Harbor, Baltimore, MD	0	17 Nov 2000	P	–	–	–	6
IH11-ϕ2	*Alteromonas*	Inner Harbor, Baltimore, MD	0	17 Nov 2000	M	–	–	–	6
IH11-ϕ5	*Alteromonas*	Inner Harbor, Baltimore, MD	0	17 Nov 2000	P	–	–	–	6
CB8-ϕ2	CB8	Chesapeake Bay, MD	0	17 Nov 2000	M	–	–	–	6
CB8-ϕ6	CB8	Chesapeake Bay, MD	0	17 Nov 2000	M	–	–	–	6
CB-ϕ8	*Vibrio alginolyticus*	Chesapeake Bay, MD	0	17 Nov 2000	M	–	–	–	6
HER320	H7	Helgoland, North Sea	0	1976-1978	M	–	–	–	7
HER321	H100	Helgoland, North Sea	0	1976-1978	P	–	–	–	7
HER322	H100	Helgoland, North Sea	0	1976-1978	M	–	–	–	7
HER327	11-68	Helgoland, North Sea	0	1976-1978	S	–	–	–	7
HER328	H105	Helgoland, North Sea	0	1976-1978	S	–	–	–	7

[a] M, P and S represent the virus families *Myoviridae, Podoviridae* and *Siphoviridae* respectively, as determined by morphology. "M?" indicates that the assignment is based solely on amplification and sequencing of a g20 PCR product and has not been confirmed with electron microscopy.

[b] Reference where cultured isolate was originally described: 1 = Sullivan et al, 2003; 2 = This study; 3 = Waterbury & Valois; 1993; 4 = Marston & Salee, 2003; 5 = Wilson et al., 1993; 6 = Zhong et al., 2002; 7 = Wichels et al., 1998

Table 2. Origins of the g20 sequences used in 'meta' phylogenetic analyses shown in Fig. 2. The "PCR-based" column indicates whether the environmental sequence was obtained by PCR or metagenomic approaches (N/A indicates that this is not applicable for sequences from cultured phage isolates).

# sequences	Description	PCR-based?	Sequence label in Figure 2	Refs
512	Environmental sequences from 42 oceanic sample sites from the Global Ocean Survey	N	JC#	1
56	Environmental sequences from 19 globally distributed freshwater and marine sites	Y	AY705#	2
25	Environmental sequences from Rhode Island coastal waters, USA	Y	AY259#	3
43	Environmental sequences from Lake Erie, USA	Y	DQ318#	4
47	Environmental sequences from Lake Bourget, France	Y	AY426#	5
27	Environmental sequences and mixed lysates from coastal northwestern Atlantic Ocean	Y	variable	6
51	Cultured marine cyanomyophages of variable coastal and open ocean origins	N/A	variable	3, 7
8	Cultured non-cyanomyophages from sewage	N/A	variable	8

References code: 1 = Rusch et al, 2007; 2 = Short & Suttle 2005; 3 = Marston & Salee, 2003; 4 = Wilhelm et al. 2006; 5 = Dorigo et al. 2004; 6 = Zhong et al., 2002; 7 = this study; 8 = T4-like phage genomes website http://phage.bioc.tulane.edu/

Table 3. Relationship between g20 sequence clusters and the microbial community types of the original habitats from which they were collected. Unifrac distance metric (Lozupone and Knight, 2005) was used for the analysis. A P-value <0.05 (italicized) indicates that sequences from that category are non-randomly distributed with respect to habitat in the phylogenetic analysis. In the Unifrac analysis presented here, we used the environmental categories given to the GOS sample g20 sequences by Rusch et al. (inferred using ribotype dot-blots and shared metagenomic content; Figs. 9 and 10 in Rusch et al. 2007), whereas we assumed which environmental category non-GOS sequences belonged to as follows: (a) Woods Hole, Plymouth, NE Providence Channel, Rhode Island waters were considered 'temperate ocean - north' (akin to GOS sample 8, Newport Harbor, RI), (b) freshwater, the Sargasso Sea or estuaries were considered 'freshwater', 'Sargasso Sea' or 'estuary', respectively, (c) arctic or polar water sequences were given their own category. We did not assume an environmental category for non-GOS samples originating from the Red Sea, Atlantic Ocean continental shelf and slope waters, Dauphin Island and Gulf Stream so they were not used in this analysis (temperature and salinity data were available for many of these samples, so they were used in subsequent analyses). A total of 698 categorized sequences were used in the Unifrac analysis. To provide an overall picture of the microbial community for each environmental category, we provide qualitative relative abundance microbial community data for each environmental category inferred from the ribotype data published for the GOS samples in Rusch et al. (2007) as follows: dark squares = highly dominant ribotypes, lighter squares = ribotypes that are present but not dominant, white squares = ribotype was not detected.

Environmental category	Unifrac P-value	# sequences	Qualitative relative abundance of dominant ribotypes inferred for the GOS samples in these categories[%]
Temperate ocean – North	0.2736	84	
Temperate ocean – South	0.0532	13	
Tropical and Sargasso	0.5098	44	
Tropical – Open Ocean	0.8969	139	
Tropical – Near Galapagos	0.1411	116	
May Sargasso Sea	0.0812	6	
Coral reef atoll	0.0714	48	
Fringing Reef	0.4238	34	
Tropical freshwater	*0.0001*	*47*	
Estuary	*0.0020*	*14*	
Sargasso Sea	*0.0029*	*32*	
Hypersaline	*0.0474*	*3*	
Freshwater	*0.0009*	*99*	NO GOS SAMPLE DATA
Arctic / Polar	*<=0.0001*	*19*	NO GOS SAMPLE DATA

SAR11 – surface 1; SAR11 – surface 2; SAR11 – surface 3; *Prochlorococcus*; *Synechococcus*; Roseobacter; SAR86; SAR116; Archaea; Cytophaga; Rhodospirillaceae; Alphaproteobacteria; Gammaproteobacteria; Acidomicrobidae; Cellulomonadaceae; Chlorobi; Acidobacteria; Frankineae; Bdellovibrionales; Comamonadaceae

[%] Qualitative characterization of the relative abundance of dominant ribotypes using published data from the GOS (detailed data available in Rusch et al. 2007). These data represent only those microbes captured in 0.1-0.8 _m size fraction samples, except for the Fringing Reef sample which is the 0.8-3.0 _m size fraction. No data are available for freshwater and arctic/polar samples because these were not part of the Global Ocean Survey sampling expedition.

Table 4. Probability that g20 sequence clusters are non-random with respect to the salinity at the site from which they were collected. The Unifrac distance metric (Lozupone et al. 2006) was used for the analysis. Salinity values, when not available from the published work, were obtained from the communicating author of the paper in which the g20 sequence was first reported. All freshwater samples were assumed to have a salinity of < 0.50 ppt. All but the sequences from brackish waters clustered non-randomly ($p<0.05$) with respect to the habitat type as defined by salinity.

Environmental category	salinity (ppt)	# sequences	Unifrac P-value
sewage	*N/A*	*6*	*<=0.0001*
fresh	*<0.50*	*149*	*<=0.0001*
estuarine	*0.5-17.99*	*6*	*0.0096*
brackish	18-32.99	183	0.1456
ocean	*33-38*	*286*	*0.0006*
hypersaline	*>38*	*8*	*0.0474*

Table 5. Probability that g20 sequence clusters are non-random with respect to the temperature at the site from which they were collected. The Unifrac distance metric (Lozupone et al. 2006) was used for the analysis. Temperature values, when not available from the published work, were obtained from the communicating author of the paper in which the g20 sequence was first reported. All but the sequences from moderate temperatures clustered non-randomly ($p<0.05$) with respect to the habitat type as defined by temperature.

Environment	Temperature (°C)	# sequences	Unifrac P-value
sewage	*N/A*	*6*	*<=0.0001*
cold	*<4.99*	*20*	*<=0.0001*
cool	5-14.99	57	0.2209
medium	15-21.99	141	0.2296
warm	*22-29.99*	*467*	*0.0003*
hot	*>30*	*3*	*0.0394*

Appendix B

Microbial community gene expression in ocean surface waters

Jorge Frias-Lopez*, Yanmei Shi*, Gene W. Tyson, Maureen L. Coleman, Stephan C. Schuster, Sallie W. Chisholm, and Edward F. DeLong

* Co-first authors

Reprinted with permission from *Proceedings of the National Academy of Sciences*

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. Nat. Acad. Sci.* 105(10): 3805-10.

Microbial community gene expression in ocean surface waters

Jorge Frias-Lopez*, Yanmei Shi*, Gene W. Tyson*, Maureen L. Coleman*, Stephan C. Schuster†, Sallie W. Chisholm*‡, and Edward F. DeLong*‡§

Departments of *Civil and Environmental Engineering and §Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139; †Center for Comparative Genomics and Bioinformatics, Pennsylvania State University, University Park, PA 16802

Edited by David M. Karl, University of Hawaii, Honolulu, HI, and approved January 22, 2008 (received for review September 19, 2007)

Metagenomics is expanding our knowledge of the gene content, functional significance, and genetic variability in natural microbial communities. Still, there exists limited information concerning the regulation and dynamics of genes in the environment. We report here global analysis of expressed genes in a naturally occurring microbial community. We first adapted RNA amplification technologies to produce large amounts of cDNA from small quantities of total microbial community RNA. The fidelity of the RNA amplification procedure was validated with *Prochlorococcus* cultures and then applied to a microbial assemblage collected in the oligotrophic Pacific Ocean. Microbial community cDNAs were analyzed by pyrosequencing and compared with microbial community genomic DNA sequences determined from the same sample. Pyrosequencing-based estimates of microbial community gene expression compared favorably to independent assessments of individual gene expression using quantitative PCR. Genes associated with key metabolic pathways in open ocean microbial species—including genes involved in photosynthesis, carbon fixation, and nitrogen acquisition—and a number of genes encoding hypothetical proteins were highly represented in the cDNA pool. Genes present in the variable regions of *Prochlorococcus* genomes were among the most highly expressed, suggesting these encode proteins central to cellular processes in specific genotypes. Although many transcripts detected were highly similar to genes previously detected in ocean metagenomic surveys, a significant fraction (≈50%) were unique. Thus, microbial community transcriptomic analyses revealed not only indigenous gene- and taxon-specific expression patterns but also gene categories undetected in previous DNA-based metagenomic surveys.

bacterial communities | metagenomics | metatranscriptomics | marine | cDNA

Cultivation-independent genomic approaches have greatly advanced our understanding of the ecology and diversity of microbial communities in the oceans (1, 2). Metagenomic methods applied in a variety of microbial habitats have led to the discovery and characterization of new genes and gene products from uncultivated microorganisms (3), assembly of whole genomes from community DNA sequence data (4), and comparisons of community gene content among diverse microbial assemblages (4–9). Recently, a very large metagenomic sampling survey was conducted in ocean surface waters, doubling the number of predicted protein sequences in public databases (10). All currently available data suggest that gene and protein "sequence space" still remain largely under sampled.

At the same time, studies of cultured members of the microbial community, such as *Prochlorococcus*, are helping to further link the ecology of genes and the ecology of organisms (11). From the considerable *Prochlorococcus* diversity observed in metagenomic datasets, clear structure has emerged, including clusters of sequence similarity and chromosomal hot spots for rearrangements (6, 8, 10). Meanwhile, laboratory studies have described physiological differentiation among isolates (12), and field surveys have documented the distribution of ecotypes in the oceans (13). These cross-scale comparisons provide a useful approach in which taxon-specific metagenomic information can be embedded and understood in the context of ecological and physiological data.

Given current research trends, it seems likely that metagenomic datasets will continue to grow rapidly and soon will dwarf whole-genome sequence datasets derived from cultivated microorganisms. The nature, size, and complexity of this information present formidable challenges to analyses and interpretation. In addition, although these data provide information about genome content, there is no clear indication of gene expression or expression dynamics. Although techniques like quantitative PCR (qPCR) can be used to quantify gene expression in natural samples, these are limited usually to measurement of a small number of known genes. Many questions remain to be answered. What fraction of the many new genes discovered in metagenomic datasets are actually expressed? Of the many hypothetical genes present, which are significantly expressed, and what is their function? What are the dynamics and time scales for gene expression in different microbial species, gene suites, and environments?

Measuring bacterial and archaeal gene expression in the wild has been challenging. The half-life of mRNA is short (14, 15), and mRNA in bacteria and archaea usually comprises only a small fraction of total RNA. Several approaches for overcoming these challenges recently have been developed. In one approach, ribosomal RNA (rRNA) subtraction was used in combination with randomly primed RT-PCR to generate microbial community cDNA for cloning and downstream sequence analysis (16). Although preliminary results were encouraging, relatively large sample volumes (≈10 liters) and long sample collecting times were required. Linear RNA amplification methods have been widely used to study gene expression in eukaryotic tissues (17, 18) but are not generally applicable to bacterial and archaeal mRNA because of the requirement of a poly(A) tail. Wendisch *et al.* (19) developed a method for the polyadenylation of bacterial messenger RNA using *Escherichia coli* poly(A) polymerase, which facilitates preferential isolation of bacterial mRNA from rRNA in crude extracts. This approach has been adapted in a commercially available kit (MessageAmp

Author contributions: J.F.-L. and Y.S. contributed equally to this work. J.F.-L., Y.S., G.W.T., S.W.C., and E.F.D. designed research; Y.S. performed research; S.C.S. contributed new reagents/analytic tools; J.F.-L., Y.S., G.W.T., M.L.C., S.W.C., and E.F.D. analyzed data; and J.F.-L., Y.S., G.W.T., M.L.C., S.W.C., and E.F.D. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: DNA and cDNA sequences reported in this paper have been deposited in the GenBank database (accession nos. SRA000262 and SRA000263).

‡To whom correspondence may be addressed. E-mail: delong@mit.edu or chisholm@mit.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0708897105/DC1.

Table 1. Characterization of the pyrosequenced DNA and cDNA libraries from the microbial community analyzed in this study

	DNA library	cDNA library
Total number of reads	414,323	128,324
Average length, bp	110	114
Number of rRNA reads	5,877	67,859
Total base pairs, Mb	45.4	14.7
Number of NCBI-nr hits	205,747 (50% of reads)	7,275 (13% of reads)
Number of GOS peptide hits	290,741 (70% of reads)	23,203 (43% of reads)

Only sequences with bit score cutoffs ≥40 were considered hits.

II-Bacteria Kit, Ambion), which couples microbial RNA polyadenylation with a linear amplification step using T7 RNA polymerase (20). Polyadenylation-dependent RNA amplification approaches have been used in studies of cultured microbes using single-genome microarrays (21, 22). We adapted this approach to enable the synthesis of microbial community cDNA from small amounts of mixed population microbial RNA. Specifically, after polyadenylation of nanogram quantities of RNA (19), the RNA was linearly amplified with T7 RNA polymerase (20) and then converted to cDNA. The cDNA was directly sequenced by pyrosequencing, avoiding the need to prepare clone libraries and their associated biases (23, 24). By pyrosequencing both genomic DNA and cDNA from the same sample, the abundance of cDNA copies can be normalized to corresponding gene copy numbers in the community DNA pool.

We report here the application, validation, and field testing in the North Pacific Subtropical Gyre (25) of these methodologies for studying microbial community gene expression. We used the technique to analyze the expression of genes across the entire microbial community, to assess the taxonomic origins of the expressed genes, and to examine gene expression in *Prochlorococcus*, the dominant phototroph in the surface waters at this site. Genes from *Prochlorococcus* are highly represented in metagenomic databases (5, 8, 10), and extensive genomic and transcriptomic data exists from culture studies (6, 26–28) and so were useful in guiding the interpretation of field observations.

Results and Discussion

Assessing the Fidelity of Bacterial mRNA Amplification. We tested the fidelity of the RNA amplification technique using *Prochlorococcus* cultures and custom-designed Affymetrix arrays [see supporting information (SI) *Methods*] (27). Levels of gene expression measured from the amplified *Prochlorococcus* RNA compared favorably with those of unamplified RNA for protein coding genes (r^2 between 0.85 and 0.92; SI Fig. 4), and the results were highly reproducible (r^2 between 0.94 and 0.99 for biological replicates; SI Fig. 5). Linearly amplified RNA also revealed the same physiologically relevant changes in gene expression, as did unamplified RNA in an experiment designed to examine the response of strain MIT9313 to phosphate starvation (SI Fig. 6) (27). Both amplified and unamplified RNA identified the same four genes, all involved in phosphate acquisition, as highly up-regulated under phosphate starvation. In contrast to this high fidelity for mRNA, rRNA transcripts were consistently underrepresented in amplified versus unamplified RNAs (SI Fig. 7), reflecting a preferential polyadenylation of mRNA, consistent with previous reports of this polyadenylation bias in crude extracts (19) and with the known inefficiency of amplification of molecules with a high degree of secondary structure (29).

Field Testing Microbial Gene Expression Profiling in the Open Ocean. As a field test, we analyzed a picoplanktonic sample collected from 75-m depth at the well characterized Hawaii Ocean Time Series (HOT) station ALOHA in the North Pacific Subtropical Gyre (25). Because metagenomic analyses already have been performed at this site (5), and the cyanobacterium *Prochlorococcus* comprises a large fraction of its microbial communities (30), significant databases exist to facilitate the interpretation of our field results. The detection frequency for any given transcript in the community depends on the abundance of transcript-bearing cells and the average number of transcripts per cell. We recovered sequence data from both cDNA and genomic DNA in the same sample, which facilitated representation of specific cDNA classes relative to their occurrence in the genomic DNA pool.

The diversity of sequences captured in the cDNA and DNA reads (Table 1) was determined by comparing all sequences to the National Center for Biotechnology Information nonredundant protein database (NCBI-nr; as of March 28, 2007) and to predicted peptides from the recent Global Ocean Sampling (GOS) metagenomic dataset (31). The number of cDNA and DNA reads with significant database matches (bit score >40; SI Fig. 8) was higher with GOS peptides than with the NCBI-nr database. A large number of GOS matches was expected because the GOS data are derived from similar microbial communities and contain a larger number of total protein sequences. The enrichment in GOS matches over NCBI-nr matches was much greater for the cDNA library (≈3-fold) compared with the DNA library (≈1.4-fold) (Table 1). The fraction of reads matched in the cDNA, however, still was relatively low (43% of total reads) compared with the DNA library (70% of reads). The large proportion of unmatched cDNA reads in part may reflect the presence of novel, rare genes, not detected in the GOS metagenomic survey, that nevertheless contribute significantly to the microbial community expression profile.

To corroborate the results, we selected a suite of genes and performed RT-qPCR and qPCR on the same RNA and DNA samples analyzed by pyrosequencing (*SI Methods*, SI Table 2, and SI Fig. 9). Three different gene expression classes were investigated: (*i*) genes shared in both genomic DNA and cDNA sequence datasets but with higher relative frequency in the cDNA pool, (*ii*) genes present in both genomic DNA and cDNA datasets but with lower relative frequency in the cDNA pool, and (*iii*) genes detected in the cDNA but not in the genomic DNA sequence dataset. The calculated RT-qPCR/qPCR ratios followed the same trends as gene expression patterns inferred from cDNA/DNA pyrosequencing analyses (SI Fig. 9). In some cases, the RT-qPCR/qPCR analysis appeared more sensitive for detecting a broader range of gene expression patterns. For example, genes found only in the cDNA sequence dataset were detected by qPCR in both RNA and DNA samples, which likely reflects the limited extent of sampling depth of the DNA pyrosequencing relative to indigenous genetic complexity.

To evaluate the protein family representation in our dataset and to functionally categorize genes, reads from both cDNA and DNA libraries were assigned to GOS protein clusters. DNA reads were assigned to 35,178 different GOS protein clusters, and cDNA reads were assigned to 4,376 clusters. There were

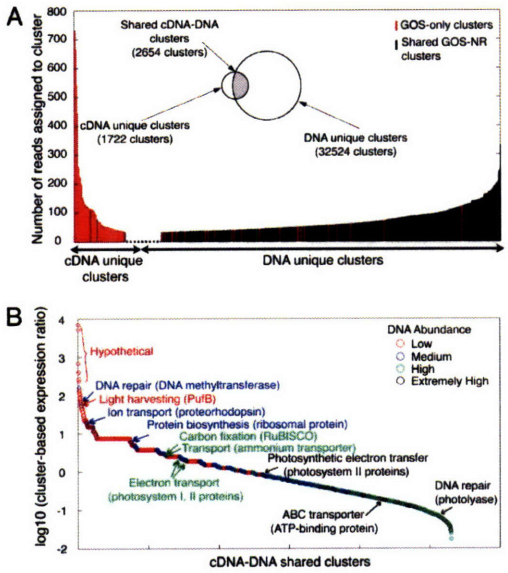


Fig. 1. Community-level gene expression profile based on GOS peptide database. (*A*) GOS protein clusters with DNA or cDNA matches at bit scores ≥40 are shown in the Venn diagram. Numbers of reads assigned to GOS protein clusters, when >70, are plotted for both cDNA-unique protein clusters and DNA-unique protein clusters. GOS protein clusters shared by DNA and cDNA libraries (shaded in gray) were further illustrated in *B*. (*B*) GOS protein clusters shared by cDNA and DNA libraries were ranked by their cluster-based expression ratio (representation of each cluster in the cDNA library normalized by its representation in the DNA library). Furthermore, each protein cluster was categorized (and color-coded) according to its abundance in the DNA library. Representative protein clusters were highlighted from each category and discussed in the text.

2,654 clusters that had both DNA and cDNA reads (Fig. 1). The smaller number of cDNA assignments is in part because the total number of cDNA reads was only one-eighth the number of DNA reads after removing rRNA sequences. Another factor likely responsible for the decreased number of high-quality sequence reads in the cDNA relative to genomic DNA includes the inefficient enzymatic removal of the poly(A) tail produced during the amplification of the mRNA. These homopolymers cause a significant number of sequences to be filtered out during processing because of lower-quality scores, low flow counts, and carry forward (premature incorporation of bases caused by incomplete flushing) (see *Materials and Methods* and ref. 24). Nevertheless, 40% of the cDNA-containing GOS clusters (referred to as "cDNA-unique clusters" hereafter) did not overlap with those in the DNA library, suggesting that the full diversity of sequences was undersampled in both the DNA and cDNA pools. This difference in representation is supported by rarefaction analysis, showing a near linear increase in the rate of recovery of GOS protein clusters with increasing number of sequence reads for both cDNA and DNA (SI Fig. 10). This finding is consistent with other large-scale metagenomic surveys that showed no sign of sequencing saturation for similar marine microbial communities (31, 32).

To maximize functional genomic information drawn from the data, the 2,654 GOS protein clusters (protein families) that were represented in both the DNA and cDNA libraries were analyzed further, calculating the number of cDNA reads matching a given GOS protein cluster, divided by the number of corresponding DNA reads in the same cluster (see *Materials and Methods*)—the "cluster-based expression ratio." This approach allowed us to bypass the difficulties associated with traditional annotation of short pyrosequencing reads (average trimmed length of ≈96 bp), which would have segmented the reads into many apparently unrelated, nonoverlapping clusters, even though they were potentially derived from the same gene. This level of analysis allows us to look at the expression profile of the microbial community at the level of protein family, without losing the resolution inherent in the data.

The 2,654 shared GOS protein clusters were categorized based on their abundance in the DNA library (low, medium, high, and extremely high; SI Fig. 11). Protein clusters with the highest cluster-based expression ratios (up to 10^3 higher than the average ratio) tended to fall into the low DNA abundance category (Fig. 1*B*). This observation, together with apparent high expression levels in cDNA-unique clusters, suggested the presence of actively transcribed genes that are relatively low in abundance in the total community. Interestingly, these highly expressed protein clusters consist mostly of hypothetical proteins that are found only in the GOS peptide database (Fig. 1 and SI Table 3). The high degree of sequence similarity (up to 100%; average 89.5%) between these GOS-only hypothetical protein matches and the cDNA reads supports the GOS gene predictions and confirms that these genes are actively expressed *in situ*. Conversely, the DNA-unique clusters are composed of protein families that are well represented in current protein databases (e.g., NCBI-nr and fully sequenced microbial genomes; Fig. 1 and SI Table 4). This finding indicates that cDNA analysis captures novel genes, with potentially important functions, that have escaped detection even in the largest metagenomic DNA survey conducted to date.

Highly Expressed Gene Categories in Known Metabolic Pathways. Expression patterns of environmentally diagnostic genes can provide significant insight into microbial processes active in the environment. For example, genes involved in microbial phototrophy—e.g., oxygenic and anoxygenic photosynthesis and phototheterotrophy—were among the most highly expressed classes in cluster-based expression ratios (Fig. 1*B*), even though the sample was collected 3 h before sunrise.

In the case of genes related to oxygenic photosynthesis, ribulose bisphosphate carboxylase (RuBisCo) large subunit (*rbc*L) homologs, encoding subunits of the key enzyme in the Calvin Cycle carbon fixation enzyme, were among the highly expressed genes in the sample (Fig. 1*B*). Expression levels of this gene were on a par with those of glutamine synthase (GS), suggesting high expression levels of this key enzyme in nitrogen metabolism that is found in all microorganisms. RuBisCo and GS gene copies were present in comparable numbers in the microbial genomic DNA of our sample, in contrast to the recently reported GOS datasets, where relatively low numbers of the *rbc*L gene were identified relative to GS (31). With respect to alternative forms of phototrophy, several protein clusters associated with aerobic, anoxygenic phototrophy showed extremely high cluster-based expression ratios (Fig. 1*B*). These proteins include light-harvesting protein β-chain (PufB), photosynthetic reaction center cytochrome *C* subunit (PufC), and chlorophyllide reductase subunit Y (BchY), which all appear to be derived from Alphaproteobacteria closely related to *Roseobacter* species (33). Although these correspond to relatively low abundances in the DNA libraries, their high expression levels support the potential ecological importance of aerobic anoxygenic phototrophy to microbial species in the open ocean.

Another important family of proteins involved in phototrophy are the proteorhodopsins (PRs), a group of membrane proteins that function as a light-driven proton pump (3). PR genes were not only abundant in community genomic DNA but also were

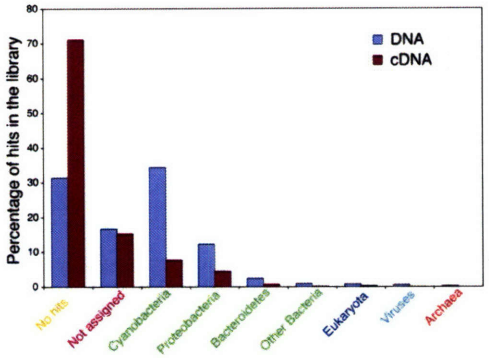


Fig. 2. Distribution of different phylogenetic groups in DNA and cDNA libraries. Percentages of the different phylogenetic groups were calculated from the MEGAN analysis results at the phylum level cutoff (SI Table 5 shows a detailed list of the distribution of number of hits and percentages for all phyla). Not assigned reads are sequences with an NR hit but a bit score <40.

among the most highly expressed genes in the cDNA pool (Fig. 1*B*). Preliminary taxonomic assignments suggest that the expressed PR genes were derived from diverse microbial taxa, supporting their general ecological significance in planktonic microbial communities (3, 34). Heterologous expression experiments have confirmed the ability of PR to function as a proton pump and enable photophosphorylation in *E. coli* (3, 35). Moreover, some (but not all) PR-containing bacteria display enhanced growth rates and cell yields in the presence of light (36, 37).

Putative Taxonomic Origins of Expressed Genes. Metatranscriptomic analyses, in principle, can be used to associate specific microbial taxa with *in situ* expression dynamics. However, phylogenetic inference based on protein-coding genes is highly dependent on a given gene's conservation across taxa, the depth of taxonomic sampling, taxon richness and evenness in the sample, and sequence read length. Further, taxonomic inferences also have the potential to be confused by horizontal gene transfer events (38). With these caveats in mind, we performed a preliminary taxonomic assessment of DNA and cDNA reads using MEGAN (39), software that assigns putative taxonomic origins based on BLAST outputs and NCBI taxonomic hierarchy. Not surprisingly, based on their known abundance in the wild and their abundance in the genomic databases, the genus *Prochlorococcus* and Alphaproteobacteria (genus *Pelagibacter*) were the two most highly represented taxonomic groups in both DNA and cDNA libraries (Fig. 2 and SI Table 5). Another noteworthy observation was the detection of expressed genes of viral origin, suggesting there was active viral infection occurring in cells *in situ* in the sample we analyzed (SI Table 5). The most common viral transcripts were related to the major capsid protein of myoviridae. Previous metagenomic analyses reported a high viral abundance in the cellular fraction from the same depth and site (5). For the most abundant groups, there was general agreement between the taxonomic origins of sequence reads in the DNA and cDNA datasets.

Evaluating Gene Expression in a Naturally Occurring *Prochlorococcus* Assemblage. The vast majority (>90%) of putative *Prochlorococcus* reads shared highest sequence similarity with strains MIT9301, AS9601, and MIT9312, all representatives of the high light-adapted eMIT9312 ecotype (40). This result (data not shown) is consistent with depth-specific ecotype abundance data based on qPCR analysis of the rRNA internally transcribed

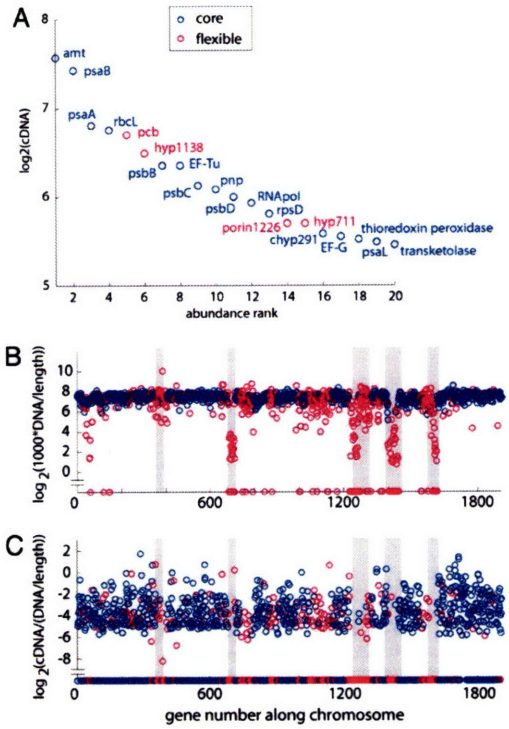


Fig. 3. *Prochlorococcus* gene and transcript abundance using strain MIT9301 as a reference genome. (*A*) Rank abundance of the 20 genes with highest frequency in the raw cDNA, reflecting transcription of the entire *Prochlorococcus* population. (*B*) Frequency of DNA hits from the natural sample along the genome of MIT9301 normalized to gene length. (*C*) Frequency of cDNA hits from the natural sample normalized to the DNA values in *B*. Gray bars indicate the location of genomic islands identified through whole-genome analysis of cultured isolates (6). Core genes, genes present in all genomes of *Prochlorococcus* sequenced, are shown in blue. Flexible genes, genes not present in all genomes of *Prochlorococcus* sequenced, are shown in pink.

spacer (ITS) region (13). Our current analysis using short pyrosequencing sequence reads from both DNA and cDNA therefore supports ecotype distributions inferred from independent analyses using a single taxonomic marker, the ITS.

Observed frequencies of the putative *Prochlorococcus* cDNA sequences reflect which genes are the most highly expressed in the *Prochlorococcus* assemblage sampled. These highly expressed genes include ammonium uptake (*amt*), photosynthesis (*psa*AB), and carbon fixation (*rbc*L) genes, pointing to key biogeochemical processes being driven, in part, by *Prochlorococcus* (Fig. 3*A* and SI Table 6). Two of the top 20 most highly expressed *Prochlorococcus* genes were hypothetical proteins: P9301_11381, which has orthologs only in the other MIT9312-like genomes (AS9601, MIT9312, and MIT9215), and P9301_07111, which has no orthologs in other *Prochlorococcus* genomes (but is paralogous to P9301_04361) (SI Table 6). High-level expression of hypothetical proteins has previously been observed in *Prochlorococcus* under nutrient limitation in laboratory experiments (27, 28). The current data indicate the potential relevance of these proteins to *Prochlorococcus* in its native environment. When a gene-length correction is applied (see *Materials and Methods*, SI Fig. 12, and SI Table 6), additional hypothetical proteins (P9301_03541 and P9301_02451) with high

per-copy transcript abundance appear to be rare in the population but are highly expressed.

The *Prochlorococcus* core genome (i.e., those genes shared by all sequenced *Prochlorococcus* isolates) consists of ≈1,250 genes (41). The "flexible" genome represents the remaining genes found in one or more genomes, and many of these variable genes are concentrated in genomic islands (6). Using strain MIT9301 as a reference, we calculated the abundance of genes belonging to the core and flexible genomes in both the DNA and cDNA libraries. In the DNA library, all *Prochlorococcus* core genes were represented with roughly equal abundance, supporting the idea that these genes are conserved and present in single copy in virtually every *Prochlorococcus* cell (Fig. 3*B*). In contrast, genes belonging to the MIT9301 flexible genome had highly variable occurrence in the DNA library, suggesting that the natural population likely harbors a different suite of such genes. In the cDNA library, core genes involved in photosynthesis and carbon fixation, for instance, were highly represented, but, surprisingly, a number of genes belonging to the flexible genome, some of which are located in genomic islands in MIT9301, also were highly represented (Fig. 3 *A* and *C*). Thus, some of these island genes appear to be highly expressed, corroborating laboratory evidence and suggesting that they are likely functionally important to naturally occurring *Prochlorococcus*. Furthermore, the majority of "flexible" genes, and hypothetical genes, were found in the cDNA pool and expressed at levels comparable to most other core genes, further indicating their significance in the biology and ecology of *Prochlorococcus*.

Microbial Community Transcriptomics: Prospects and Challenges. Many challenges are associated with the interpretation of microbial gene expression patterns at the community level. These arise in part from the remarkable diversity and complexity of microbial communities in the ocean environment, logistics associated with field sampling, and the lack of comprehensive representation in metagenomic databases. Rapid collection and processing of samples for gene expression studies, for example, still presents significant challenges. Although our approach used relatively small volumes (1 liter) and short filtration times (<15 min), there still remains significant room for improvement. Other factors that will influence community transcriptomic analyses include the specifics of mRNA synthesis and degradation rates, environmental conditions at the time of sampling (time of day, for example), sequence read size and target gene size, and the specific method used for gene identification and annotation. Some of these variables can be controlled or improved, and others are inherent to the specific environment or community being sampled.

It is well accepted that longer sequence reads are generally more informative, allowing more robust annotation. Side-by-side comparisons of Sanger dideoxy sequences versus pyrosequences derived from the same metagenomic samples, however, have been generally consistent and comparable (9, 42). The sequence reads in our dataset have an average size of ≈96 bp, sufficient for general functional annotation and, in the case of *Prochlorococcus*, for assignment of reads to specific genes and ecotypes. For as-yet-uncultivated microorganisms, 100 bp is not always sufficient for specific gene assignment. Improvements in pyrosequencing, however, now produce >230-bp length reads. Further improvements in pyrosequencing read lengths are anticipated, which will improve accuracy and precision in future microbial community transcriptomics studies.

Despite the caveats and potential improvements, we have shown metatranscriptomic sequencing and characterization is sufficient to identify many expressed biological signatures in complex biological samples such as seawater. Whole-community analysis relying on gene family clustering for analyses of pyrosequencing reads revealed clear patterns in community gene expression for individual taxa, specific genes, and within protein families. Taxon-specific analyses focusing on *Prochlorococcus* provided deep insight into the most highly expressed genes among these populations. Interestingly, both in the case of the whole community as well as in the case of *Prochlorococcus*, hypothetical genes were among the most highly expressed, underlining the potential importance of these unidentified proteins. The fact that a large fraction of cDNA reads were not present in the available databases, including the GOS database, indicates that we have just scratched the surface of the microbial metabolic diversity present in the ocean.

Metatranscriptomics (ref. 16 and this report) and proteomics (43, 44) represent two approaches in microbial ecology that have potential to significantly leverage, apply, and extend existing microbial metagenomic datasets. The two approaches each measure a different component and dynamic of the macromolecular pool, reflecting the different regulatory controls, expression rates, and turnover kinetics of mRNAs and proteins. Although transcriptomics has potential to reveal the near-instantaneous responses to environmental fluctuation, proteomics more directly reflects the immediate catalytic potential of the microbial community. In conjunction with metagenomic data, these approaches offer significant promise to advance measurement and prediction of *in situ* microbial responses and activities in complex, naturally occurring, or engineered microbial communities.

Materials and Methods

Sampling. Seawater was collected at the HOT station ALOHA (22°44′N, 158°2′W), 75-m depth, on March 9, 2006, 03:30 a.m. local time. Hydrocasts for sampling and hydrographic profiling were conducted by using a conductivity-temperature-depth (CTD) rosette water sampler equipped with 24 Scripps 12-liter sampling bottles aboard the R/V Kilo Moana. Continuous vertical profiles of physical and chemical parameters thus were recorded. DNA and RNA extraction, processing, and sequencing are detailed in the *SI Methods*.

RNA Amplification and cDNA Synthesis. Approximately 5 μl of RNA (≈100 ng total) was amplified by using MessageAmp II-Bacteria Kit (Ambion) following the manufacturer's instructions. Briefly, the method is based on polyadenylation of the 3′ end of total RNA. The A-tailed RNA is reverse-transcribed primed with an oligo(dT) primer containing a T7 promoter sequence and a restriction enzyme (BpmI) recognition site sequence [T7-BpmI-$(dT)_{16}$VN], then double-stranded cDNA is synthesized. Finally, the cDNA templates are transcribed *in vitro* (37°C for 6 h), yielding large amounts of antisense RNA (aRNA; ≈1,000-fold amplification). The aRNA is polyadenylated and further reverse-transcribed to cDNA with the SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen). Finally, ≈2 μg of cDNA is digested with BpmI, purified, and used for pyrosequencing.

Pyrosequencing. DNA and cDNA libraries were constructed as previously described (23, 45) and sequenced with a Roche GS20 DNA sequencer. A full run of the sequencer yielded 45,380,301 bp from 414,323 reads (110-bp average length) from the DNA library, and 14,675,424 bp from 128,324 reads (114-bp average length) from the cDNA library (Table 1). The lower number of cDNA library reads may be because of shorter cDNA fragments and highly polymeric sequences resulting from inefficient removal of poly(A) tails introduced during mRNA amplification. To pass GS20 quality filters, flow-grams for each read require at least 84 flows (21 cycles or ≈50 bp) and <5% of flows with ambiguous bases (N) and <3% of flow-gram values between 0.5 and 0.7.

Analysis of Metagenomic GS20 DNA and cDNA Data. DNA and trimmed non-RNA cDNA reads were compared with the NCBI-nr (as of March 28, 2007) and GOS peptide databases using BLASTX (46). Top BLASTX hits with bit score >40 were used to assign DNA and cDNA reads to GOS peptides and NCBI-nr proteins (Table 1). Reads assigned to GOS peptides were linked to GOS protein clusters and associated GO, Pfam, and TIGRfam annotations (if available). Additional details are provided in *SI Methods*.

ACKNOWLEDGMENTS. We thank the HOT team, the captain and crew of the R/V Kilo Moana for the expert assistance at sea, and Chon Martinez for preparing the sample DNA. This work was supported by the Gordon and Betty Moore Foundation (E.F.D. and S.W.C.), the National Science Foundation (S.W.C.), National Science Foundation Microbial Observatory Award MCB-0348001 (to E.F.D.), the Department of Energy Genomics GTL Program (E.F.D. and S.W.C.), and the Department of Energy Microbial Genomics Program (E.F.D. and S.W.C.). This article is a contribution from the National Science Foundation Science and Technology Center for Microbial Oceanography: Research and Education (C-MORE).

1. Giovannoni SJ, Stingl U (2005) Molecular diversity and ecology of microbial plankton. *Nature* 437:343–348.
2. DeLong EF, Karl DM (2005) Genomic perspectives in microbial oceanography. *Nature* 437:336–342.
3. Beja O, Spudich EN, Spudich JL, Leclerc M, DeLong EF (2001) Proteorhodopsin phototrophy in the ocean. *Nature* 411:786–789.
4. Tyson GW, *et al.* (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428:37–43.
5. DeLong EF, *et al.* (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science* 311:496–503.
6. Coleman ML, *et al.* (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* 311:1768–1770.
7. Tringe SG, Rubin EM (2005) Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet* 6:805–814.
8. Venter JC, *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304:66–74.
9. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444:1027–1031.
10. Rusch DB, *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PloS Biol* 5:398–431.
11. Coleman ML, Chisholm SW (2007) Code and context: *Prochlorococcus* as a model for cross-scale biology. *Trends Microbiol* 15:398–407.
12. Moore LR, Ostrowski M, Scanlan DJ, Feren K, Sweetsir T (2005) Ecotypic variation in phosphorus acquisition mechanisms within marine picocyanobacteria. *Aquatic Microb Ecol* 39:257–269.
13. 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.
14. Selinger DW, Saxena RM, Cheung KJ, Church GM, Rosenow C (2003) Global RNA half-life analysis in *Escherichia coli* reveals positional patterns of transcript degradation. *Genome Res* 13:216–223.
15. Andersson AF, Lundgren M, Eriksson S, Rosenlund M, Bernander R, Nilsson P (2006) Global analysis of mRNA stability in the archaeon Sulfolobus. *Genome Biol* 7:R99.
16. Poretsky RS, *et al.* (2005) Analysis of microbial gene transcripts in environmental samples. *Appl Environ Microbiol* 71:4121–4126.
17. Feldman AL, *et al.* (2002) Advantages of mRNA amplification for microarray analysis. *Biotechniques* 33:906–912, 914.
18. Moll PR, Duschl J, Richter K (2004) Optimized RNA amplification using T7-RNA-polymerase based *in vitro* transcription. *Anal Biochem* 334:164–174.
19. Wendisch VF, Zimmer DP, Khodursky A, Peter B, Cozzarelli N, Kustu S (2001) Isolation of *Escherichia coli* mRNA and comparison of expression using mRNA and total RNA on DNA microarrays. *Anal Biochem* 290:205–213.
20. Vangelder RN, Vonzastrow ME, Yool A, Dement WC, Barchas JD, Eberwine JH (1990) Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proc Natl Acad Sci USA* 87:1663–1667.
21. Moreno-Paz M, Parro V (2006) Amplification of low quantity bacterial RNA for microarray studies: time-course analysis of *Leptospirillum ferrooxidans* under nitrogen-fixing conditions. *Environ Microbiol* 8:1064–1073.
22. Rachman H, Lee JS, Angermann J, Kowall J, Kaufmann SH (2006) Reliable amplification method for bacterial RNA. *J Biotechnol* 126:61–68.
23. Margulies M, *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380.
24. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* 8:R143.
25. Karl DM, *et al.* (1996) Seasonal and interannual variability in primary production and particle flux at Station ALOHA. *Deep Sea Res Part II Topical Stud Oceanogr* 43:539–568.
26. Rocap G, *et al.* (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* 424:1042–1047.
27. 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.
28. Tolonen AC, *et al.* (2006) Global gene expression of *Prochlorococcus* ecotypes in response to changes in nitrogen availability. *Mol Syst Biol* 2:53.
29. von Wintzingerode F, Gobel UB, Stackebrandt E (1997) Determination of microbial diversity in environmental samples: Pitfalls of PCR-based rRNA analysis. *FEMS Microbiol Rev* 21:213–229.
30. 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.
31. Yooseph S, *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: Expanding the universe of protein families. *PloS Biol* 5:432–466.
32. Sogin ML, *et al.* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere." *Proc Natl Acad Sci USA* 103:12115–12120.
33. Oz A, Sabehi G, Koblizek M, Massana R, Beja O (2005) Roseobacter-like bacteria in Red and Mediterranean Sea aerobic anoxygenic photosynthetic populations. *Appl Environ Microbiol* 71:344–353.
34. Sabehi G, *et al.* (2005) New insights into metabolic properties of marine bacteria encoding proteorhodopsins. *PloS Biol* 3:1409–1417.
35. Martinez A, Bradley AS, Waldbauer JR, Summons RE, DeLong EF (2007) Proteorhodopsin photosystem gene expression enables photophosphorylation in a heterologous host. *Proc Natl Acad Sci USA* 104:5590–5595.
36. Gomez-Consarnau L, *et al.* (2007) Light stimulates growth of proteorhodopsin-containing marine Flavobacteria. *Nature* 445:210–213.
37. Giovannoni SJ, *et al.* (2005) Proteorhodopsin in the ubiquitous marine bacterium SAR11. *Nature* 438:82–85.
38. Boucher Y, *et al.* Lateral gene transfer and the origins of prokaryotic groups. *Ann Rev Genet* 37:283–328.
39. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Res* 17:377–386.
40. 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.
41. Kettler GC, *et al.* (2007) Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* 3:e231.
42. Gill SR, *et al.* (2006) Metagenomic analysis of the human distal gut microbiome. *Science* 312:1355–1359.
43. Ram RJ, *et al.* (2005) Community proteomics of a natural microbial biofilm. *Science* 308:1915–1920.
44. Lo I, *et al.* (2007) Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* 446:537–541.
45. Poinar HN, *et al.* (2006) Metagenomics to paleogenomics: Large-scale sequencing of mammoth DNA. *Science* 311:392–394.
46. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.

Supporting information.

Sample collection for DNA extraction. Bacterioplankton samples for DNA extraction were collected as previously described with minor modifications (1). Briefly, the seawater was pre-filtered in line through 125 mm Whatman® GFA filter (Whatman, Maidstone, UK) before the final collection of bacterioplankton cells onto 0.22 μm Steripak-GP20 filter (Millipore, Bedford, MA) using a Masterflex® peristaltic pump (Cole Parmer Instrument Company, Vernon Hills, IL). After a total of 260 l of seawater was filtered, the Steripak filter was covered with lysis buffer (50 mM Tris•HCl, 40 mM EDTA and 0.75 M sucrose) and frozen in -80°C aboard before shipped frozen to the laboratory where they were stored at –80°C until DNA extraction.

DNA extraction. DNA was extracted using slightly modified lysis and purification methods (2). Briefly, a solution of 5 mg/ml of lysozyme in 3 ml of lysis buffer was added to the Steripak-GP20 filter cartridge (Fisher, Fairlawn, NJ) after thawing, and incubated at 37°C for 30 min. Proteinase K (Sigma, St Louis, MO) in sterile water was added (at a final concentration of 0.5 $mg \cdot ml^{-1}$) into the Steripak-GP20 filter cartridge, followed by addition of SDS (Sigma, St Louis, MO) to a final concentration of 1 %. The filter cartridges were sealed and incubated at 55°C for 20 minutes, followed by further incubation at 70°C for 5 minutes to further promote cell lysis. The lysate was remove from the filter cartridge, and nucleic acids were extracted twice with phenol:chloroform:IAA (25:24:1, Sigma, St Louis, MO) and once with chloroform:isoamyl alcohol (24:1, Sigma). The purified aqueous phase was concentrated by spin dialysis using a Centricon 100 filter. An aliquot (~2 μg) of the extracted DNA was used for GS20 pyrosequencing.

Sample collection for RNA extraction. Bacterioplankton cells for total RNA extraction were collected filtering seawater from the same water sample that was used in DNA sample collection. We modified the collection process to shorten sampling time and improve sample preservation, which is critical in transcriptomics studies. The Niskin bottle transportation time in the water column is entirely dependent on the depth the CTD reaches, however, immediately upon shipboard retrieval of the CTD, a smaller volume of seawater (~ 1 L) was filtered as rapidly as possible. The time from the start of filtration to storage in RNA later was 12 minutes. Briefly, the seawater was pre-filtered through 1.6 um GF/A filters (Whatman, Maidstone, UK) and then filtered through 25 mm 0.22 um Durapore® filters (Millipore, Bedford, MA) using a 4-head peristaltic pump system. The pre-filtering step was used to remove most eukaryotic cells although picoeukaryote cells (eukaryotes < 2.0 um in diameter) were present in the sample. The four Durapore® filters (identica replicates) were immediately transferred to a screw cap tube containing 1 ml RNAlater (Ambion Inc., Austin, TX) after filtration, and frozen and kept at –80°C aboard

the R/V Kilo Moana. Samples were transported frozen to the laboratory in a dry shipper and stored at –80°C until RNA extraction procedures.

RNA extraction. Total RNA was extracted using a mirVana® RNA isolation kit (Ambion, Austin, TX), with several modifications to recover RNA possibly released to the 1 ml RNAlater due to the sample freeze and thaw. Samples were thawed on ice, and the 1 ml RNAlater was gently pipetted out and loaded onto two Microcon YM-50 columns (Millipore, Bedford, MA) for desalting and concentrating by centrifugal filtration. The resulting 50 µl RNAlater was added back to the sample tubes, and total RNA extraction was proceeded following the mirVana® manual. Genomic DNA was removed using a Turbo® DNA-free kit (Ambion, Austin, TX). Finally, extracted RNA (DNase treated) from four replicate filters were combined, purified, and concentrated using the MinElute PCR Purification Kit (Qiagen, Valencia, CA).

Microarray Analysis of *Prochlorococcus* gene expression. For the experiments with *Prochlorococcus* MED4, cells were grown in the Pro99 seawater based medium (3) at 21 °C under continuous white light at 16 mol photon·m^{-1}·s^{-1}. Cells were harvested by centrifugation (10,000x*g*) in log phase growth. Growth conditions and cell collection under phosphorus starvation of *Prochlorococcus* MIT 9313 were as described by Martiny et al. (4). Samples of *Prochlorococcus* MIT 9313 were taken after 12 hours under phosphorus starvation.

Before microarray analysis and RNA amplification, DNA was removed using the Turbo DNA-*free* kit (Ambion, Austin, TX). Synthesis, labeling, and hybridization of cDNA onto customized MD4–9313 Affymetrix (Santa Clara, CA) microarrays was done following the standard Affymetrix protocol and scanning was carried out according to Affymetrix protocols for *Escherichia coli* (http://www.affymetrix.com/support/technical/manual/expression_manual.affx). Data visualization was carried out using GeneSpring software (Version 7.3.1; Silicon Genetics, Palo Alto, CA). An initial normalization was applied using the Robust Multichip Average algorithm (5) implemented in GeneSpring. Those values were later normalized using the lowess correction performed using the software R (http://www.R-project.org) (6).

RT-qPCR analysis. Possible traces of DNA were removed using Ambion's Turbo DNA-*free* kit (Ambion, Austin, TX) following manufacturers instructions with minor modifications. The volume of Turbo DNase I was increased to 3µL Turbo DNase I (Ambion's Turbo DNA-*free*, Ambion) and the reaction mixture was incubated at 37 °C for 60 min. RNA (1 ng) was reverse transcribed using random hexamer primers and Superscript II reverse transcriptase (Invitrogen, Carlsbad, CA) following

manufacturers instructions. RT was performed at 42°C for 2 hours, after an initial incubation step of 10 minutes at 25°C. The synthesized cDNA and purified environmental DNA (1 ng) were used in SYBR Green quantitative PCR (qPCR) using the specific primers for the genes of interest (SI Table 6). To compare the relative expression of genes we modified the $2^{-\Delta\Delta C_T}$ method (7) and used the formula $cDNA/DNA = (1+E_{DNA})^{C_{T,DNA}} / (1+E_{cDNA})^{C_{T,cDNA}}$ to take into consideration the different amplification efficiencies in separate qPCR runs.

Sequence analyses of cDNA and DNA reads. The defined bitscore cutoff for assigning reads to GOS peptides and NCBI-nr protein was based on in silico tests using BLASTX comparisons against non-marine microbial genomes (SI Fig. 8) where a bitscore of > 40 were shown to result in a low false positive frequencies (<2%). Furthermore, a breakdown of amino acid identity and length values for bitscores greater than 40 observed in DNA library (SI Fig. 8) highlights the stringency of this cutoff.

Assignment of reads to GOS protein clusters enabled the calculation of cluster based expression ratio, a normalized comparison of the number of reads found for each protein cluster in the cDNA library relative to that found in the DNA library. In order to normalize this ratio for the difference in DNA and cDNA library size, the number of reads assigned to any given protein cluster were divided by the total number of reads in the respective library. The resulting cluster fraction for the cDNA library was then expressed as a function of the representation in DNA library. The cluster based expression ratios were ranked from highest to lowest (Fig. 1) in order to look at clusters being expressed at elevated levels.

The relative abundance of detected clusters was taken into consideration by dividing cluster based expression ratios into categories based on their abundance in the DNA library. Using an empirical cumulative density function (SI Fig. 11) clusters were categorized as either low (< 9 read members), medium (9 – 161 read members), high (161- 461 read members) or extremely high abundance (> 461 read members). This abundance measure also reflects the conservation of protein clusters, as more conserved proteins clusters are likely to have more members (e.g. RNA polymerase). Rarefaction analysis for each sample was based on best matches against the GOS database. The frequency of observed best matches to GOS protein clusters for each library was used to calculate rarefaction curves with the program Analytic Rarefaction 1.3.

Putative *Prochlorococcus* reads were identified as reads with top BLASTX hit (against NCBI-nr) to *Prochlorococcus*, and with a bitscore > 40. Each of these putative *Prochlorococcus* reads was then searched against a database of 11 whole genome sequences using BLASTN and assigned to the best hit gene. For comparison with a single reference genome, MIT9301, the assigned genes from 11 strains were all translated to their MIT9301 ortholog (8), where one exists. The number of raw cDNA reads per gene was used to indicate the most transcribed genes in the entire *Prochlorococcus* population. To normalize

cDNA reads per gene copy, the number of DNA reads per gene was first divided by the gene length (x1000 to give reads/kb) to account for a clear direct relationship between gene length and its representation in the DNA reads (SI Fig 12). A clear, direct relationship with gene length does not exist for cDNA reads. The number of cDNA reads per gene was then divided by this normalized DNA (DNA reads/kb), to give an indication of per-copy cDNA abundance. This additional normalization to gene length, which is not possible for the whole community without good reference genomes, is generally consistent with the expression ratio (cDNA/DNA) – analogous to the cluster-based expression ratio used for whole community analyses – except, for example, in cases of very short genes (SI Fig. 12).

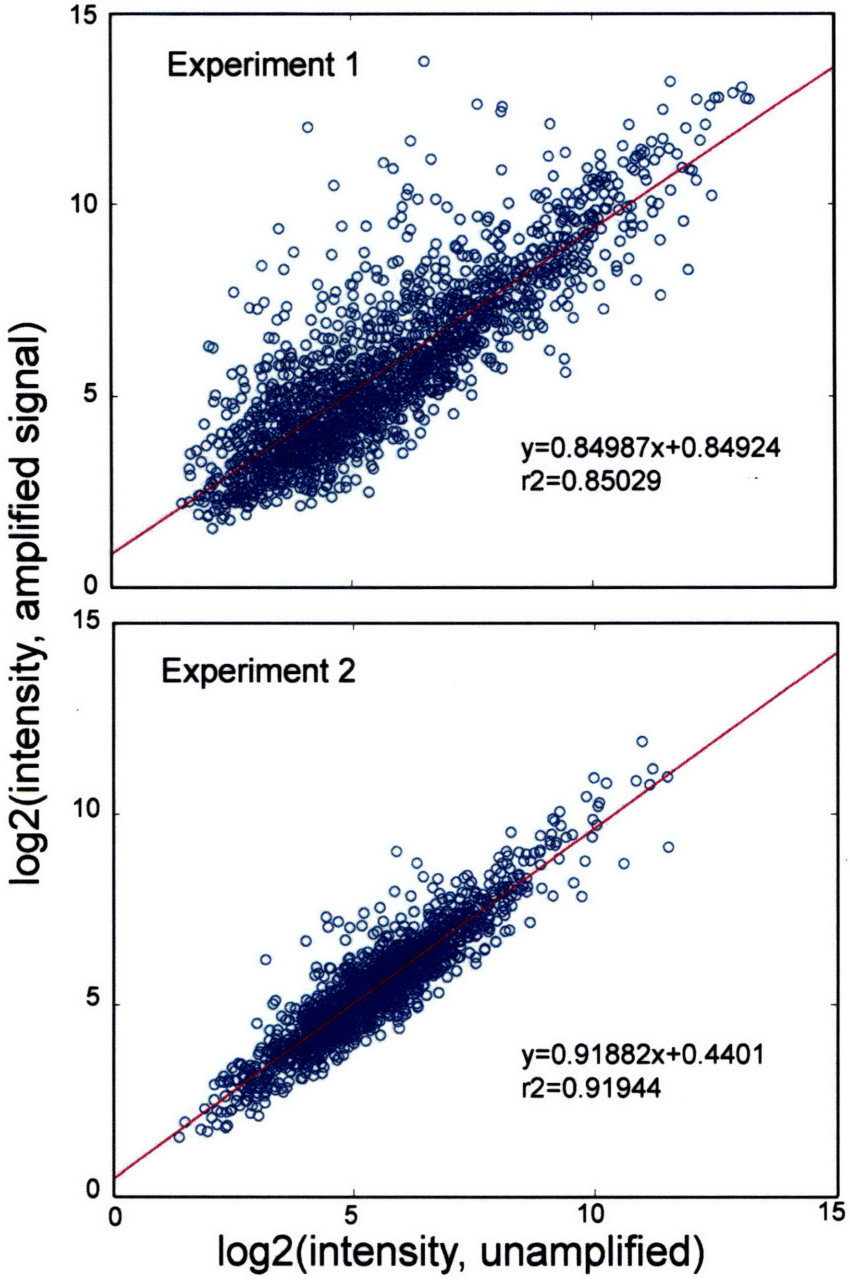
Removal of Low-Quality and rRNA GS20 cDNA sequences. Polymeric sequences inadvertently introduced into the cDNA library during cDNA synthesis (via polyadenylation of mRNA/aRNA and subsequent amplification step)were trimmed from reads based on the observed frequency of polymeric sequences in the DNA library (SI Fig.13). A noticeable peak in polyA/T sequences in the cDNA library around 16 bp (SI Fig.13) is attributable to polyadenylation of the mRNA and subsequent amplification with a T7-BmpI-$(dT)_{16}$VN primer. To remove residual T7 promoter and priming sites not cleaved by BmpI, reads were initially screened using crossmatch (-minmatch 10, -minscore 10; found in 32,246 reads). Reads containing a polyA/T sequence >10 bp (cutoff based on SI Fig.13), or multiple polyA/T runs in a single read (4 x 6 bp) were trimmed unless a significant BLASTN match across the polymeric sequence in the cDNA read was identified in a read from the DNA library (39,444 reads remained untrimmed). Using these criteria, bases flanking the ends of each cDNA read were trimmed and reads with polymeric sequences located in the middle of reads were deemed putative chimeras and removed from the dataset (5232 chimeric reads).

Ribosomal RNAs (rRNAs) were removed from the cDNA library using a combined 5S, 16S, 18S, 23S, and 28S rRNA database derived from available microbial genomes and sequences from the ARB SILVA LSU and SSU databases (www.arb-silva.de). BLASTN matches with bit score > 40 were considered significant and deemed rRNA sequences (65,859 reads; 51.3% of reads). This bit score cutoff resulted in <1.7% false positives against a database of all non-rRNA microbial genes from available microbial genomes. After trimming and removal of rRNAs, 54,568 reads (average length 95 bps) totaling 5,194,332 bps remained in the cDNA sample. Raw metagenomic GS20 DNA and cDNA reads have been deposited in GenBank, accession numbers XXX-XXX.

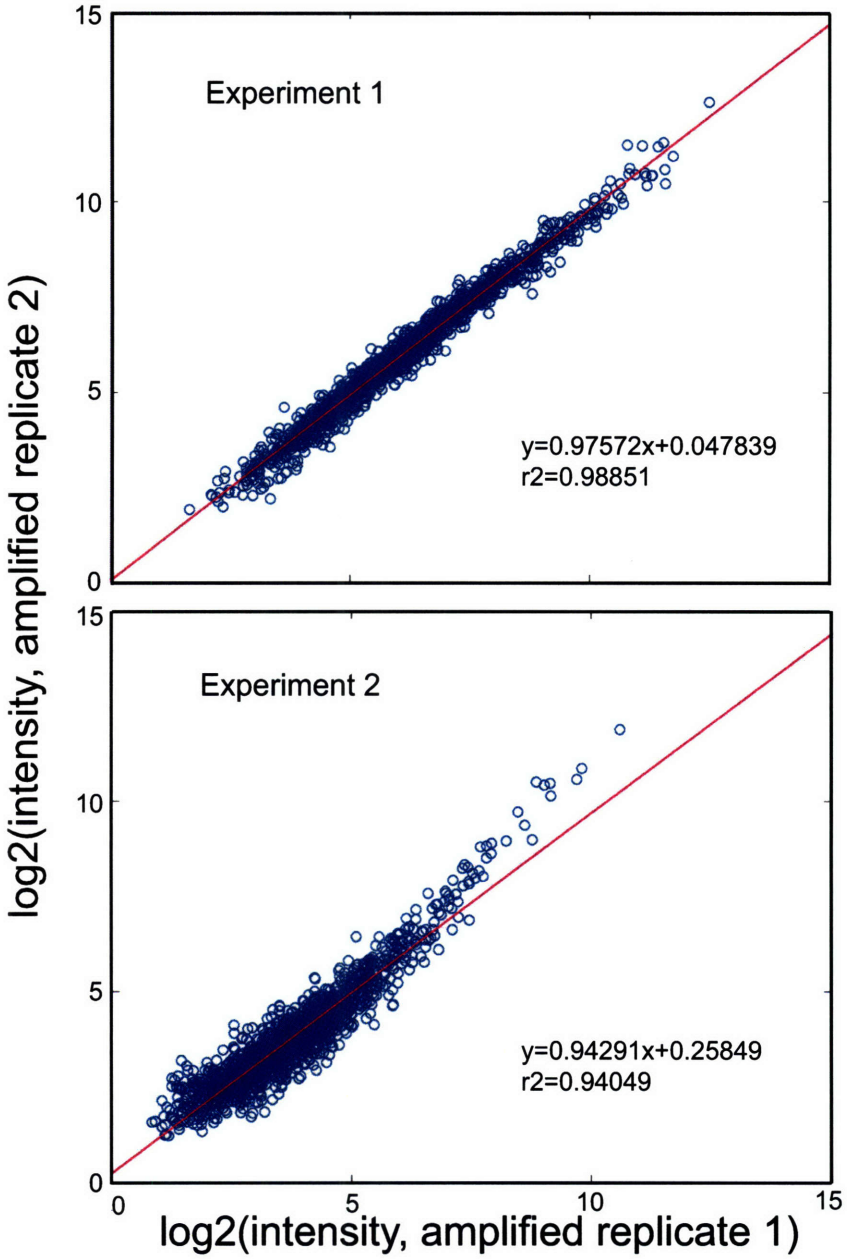
MEGAN and statistical analysis. We performed sequence comparisons of DNA and cDNA pyrosequencing results against the NCBI-nr database. Only the best hit of the top BLASTX hits with a bitscore greater than 40 was used for MEGAN analysis (version 2beta3 Aug. 2007). MEGAN is a new

software program (9) used to explore the taxonomical content of the data set, employing the NCBI taxonomy to summarize and order the results. Moreover, MEGAN gives the number of hits obtained for the different taxonomic groups, which allows for statistical comparison of the distribution of those groups on the phylogenetic trees. Statistical differences between taxonomic groups on the DNA and cDNA trees obtained in MEGAN was assessed using the software R (http://www.R-project.org)(6). Chi-squared test was used to estimate differences at the level of Kingdom. In this case we used the Pearson's Chi-squared test with simulated p-value (based on 10000 replicates) and the log likelihood ratio (G-test) test with Williams' correction (g.test.r code in R, from Peter L. Hurd http://www.psych.ualberta.ca/~phurd/cruft/).

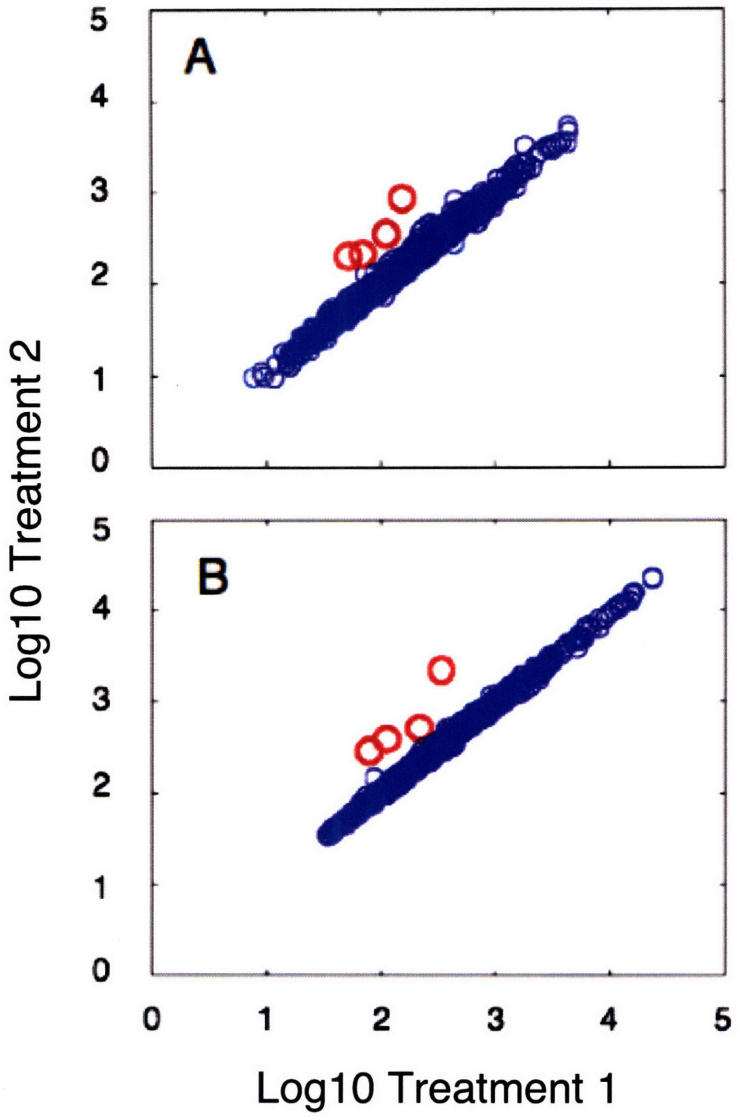
1. M. L. Coleman, M. B. Sullivan, A. C. Martiny, C. Steglich, K. Barry, E. F. Delong and S. W. Chisholm (2006). Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* **311,** 1768-70.
2. M. T. Suzuki, C. M. Preston, O. Beja, J. R. de la Torre, G. F. Steward and E. F. DeLong (2004). Phylogenetic screening of ribosomal RNA gene-containing clones in Bacterial Artificial Chromosome (BAC) libraries from different depths in Monterey Bay. *Microb Ecol* **48,** 473-88.
3. R. Rippka, T. Coursin, W. Hess, C. Lichtle, D. J. Scanlan, K. A. Palinska, I. Iteman, F. Partensky, J. Houmard and M. Herdman (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,** 1833-47.
4. A. C. Martiny, M. L. Coleman and S. W. Chisholm (2006). Phosphate acquisition genes in *Prochlorococcus* ecotypes: Evidence for genome-wide adaptation. *Proceedings of the National Academy of Sciences of the United States of America* **103,** 12552-12557.
5. B. M. Bolstad, R. A. Irizarry, M. Astrand and T. P. Speed (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19,** 185-193.
6. R. D. C. Team (2007). R: A Language and Environment for Statistical Computing.
7. K. J. Livak and T. D. Schmittgen (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(T)(-Delta Delta C) method. *Methods* **25,** 402-408.
8. G. C. Kettler, A. C. Martiny, K. Huang, J. Zucker, M. Coleman, S. Rodrigue, F. Chen, A. Lapidus, S. Ferriera, J. Johnson, C. Steglich, G. Church, P. M. Richardson and S. W. Chisholm (2007). Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genetics* **preprint,** e231.eor.
9. D. H. Huson, A. F. Auch, J. Qi and S. C. Schuster (2007). MEGAN analysis of metagenomic data. *Genome Research* **17,** 377-386.



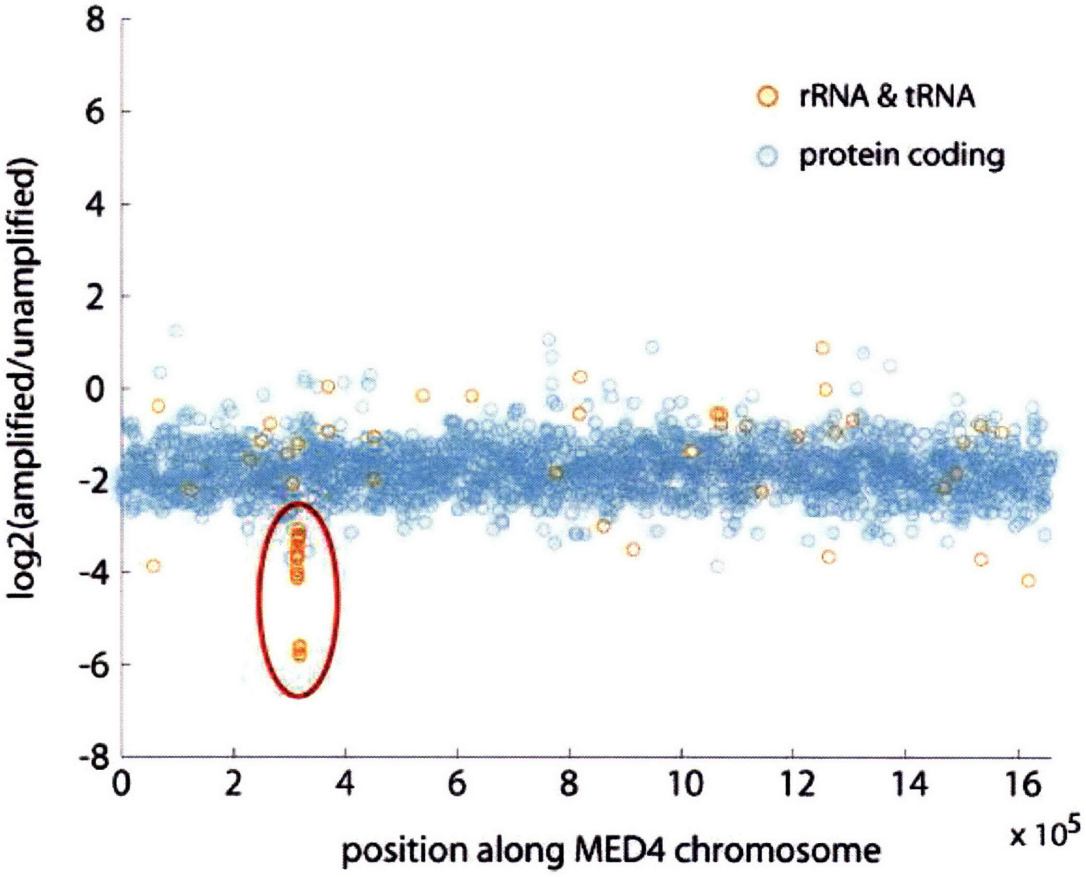
Supporting information. Fig. 4. Comparison of linearly amplified and un-amplified mRNA from cultures of *Prochlorococcus* (MED4) cells using custom Affymetrix arrays. Expression values for protein-coding genes of *Prochlorococcus* MED4 for unamplified RNA vs. the amplified RNA obtained from a 100 ng aliquot from the former. Results from two independent experiments are shown.



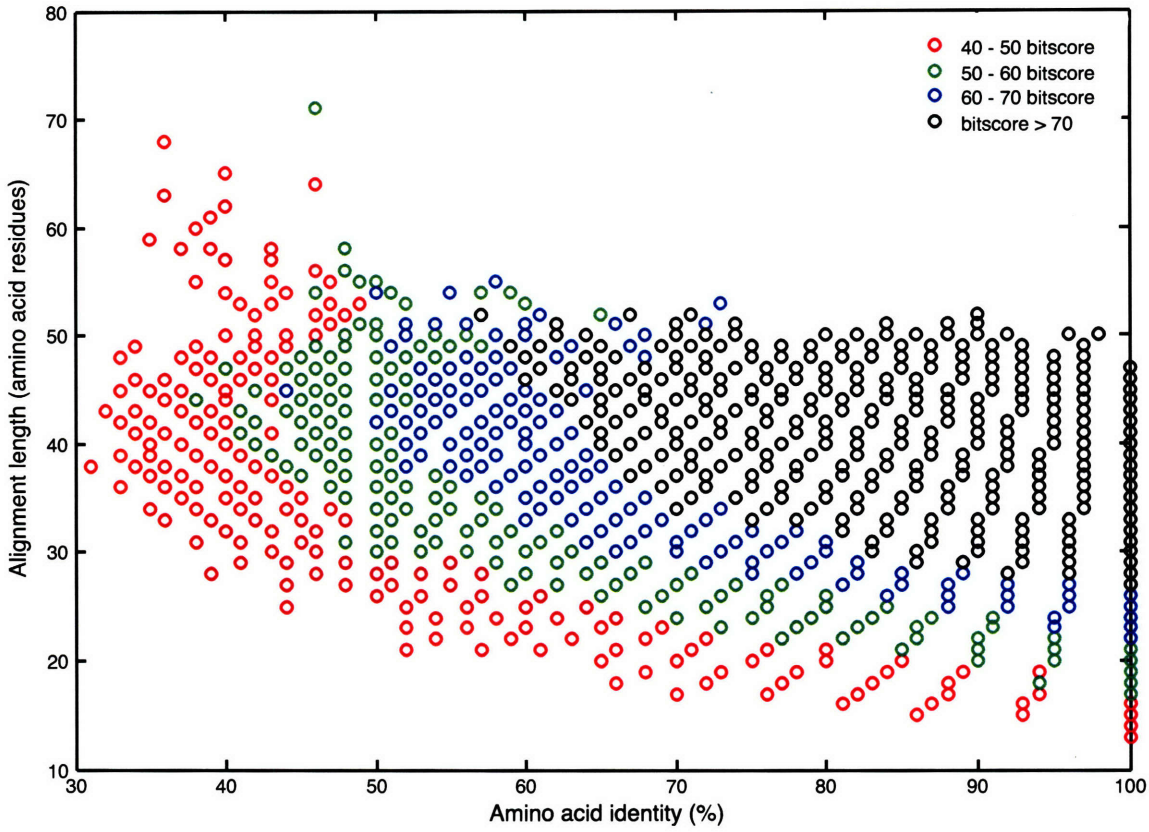
Supporting information. Fig. 5. Comparison of linearly amplified mRNA from duplicate cultures of *Prochlorococcus* (MED4) cells using custom Affymetrix arrays. Expression values for protein-coding genes of *Prochlorococcus* MED4 of replicate amplified samples plotted against each other showing the reproducibility of the amplification. Results are from 2 independent experiments.



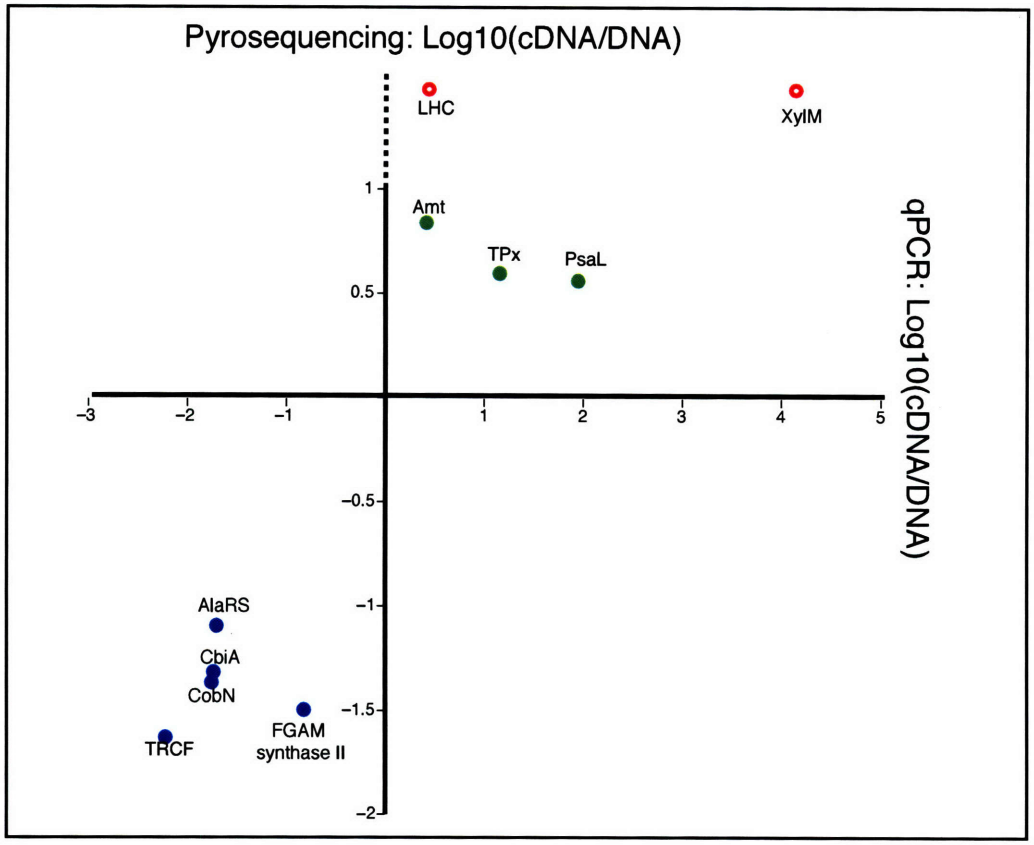
Supporting information. Fig. 6. Comparison of the results of an experiment designed to reveal upregulated genes in *Prochlorococcus* (MIT9313) under P-starvation, using unamplified (A) and amplified (B) RNA using custom Affymetrix arrays. Treatment 1: Control culture in P-replete media. Treatment 2: P-starved cultures. The same four genes appear as differentially expressed in both amplified and un-amplified treatments: a *phoB* two component response regulator, a Som like protein (P-limitation inducible outer membrane porins) and two ABC transporter substrate (phosphate) binding protein.



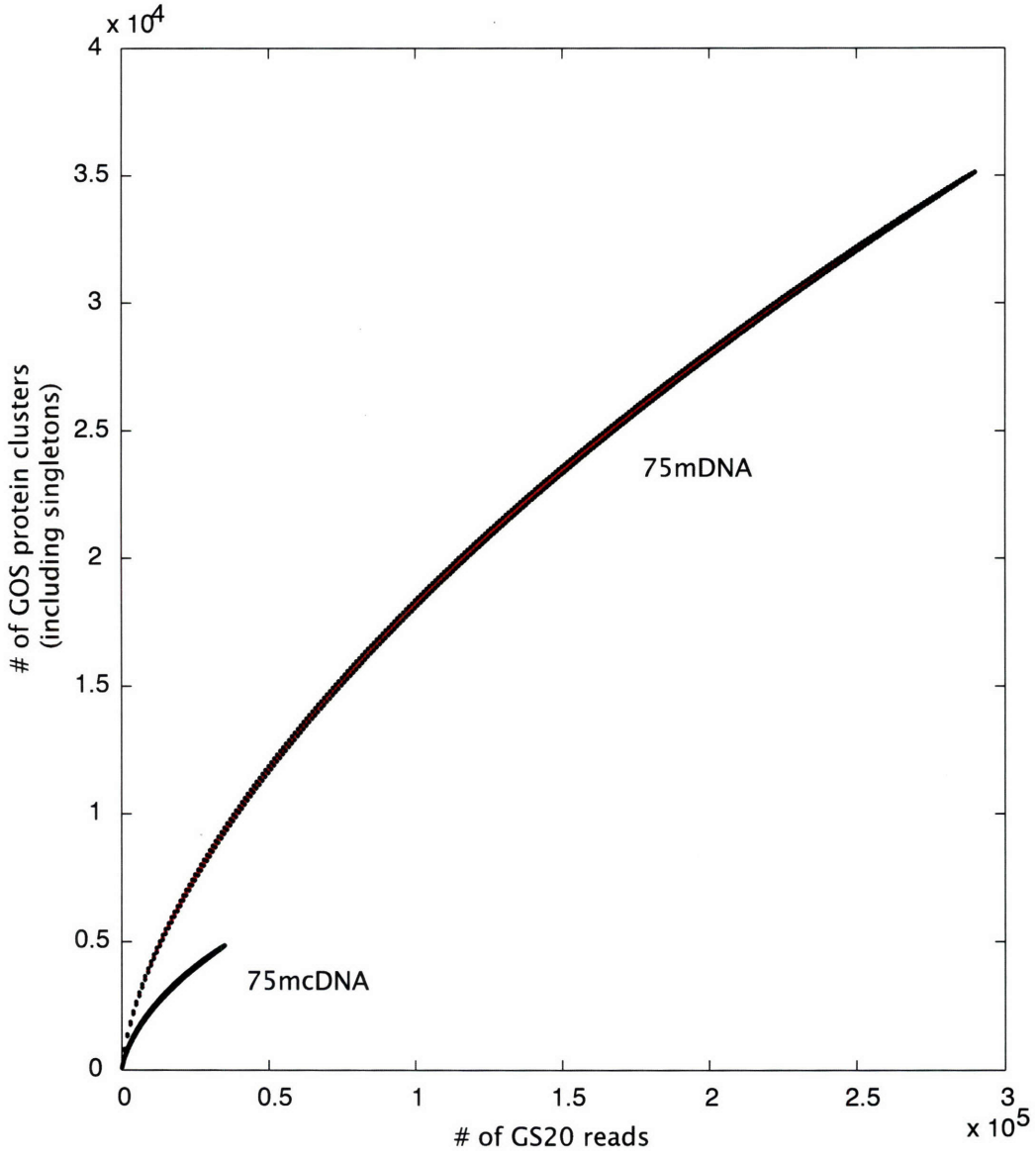
Supporting information. Fig. 7. Analysis of accuracy of RNA amplification as a function of position along the *Prochlorococcus* MED4 chromosome using custom Affymetrix arrays. The ratio of the expression values yielded from amplified and unamplified RNA for protein coding genes (blue) and ribosomal RNAs and tRNAs (red dots). The circled red dots are rRNAs.



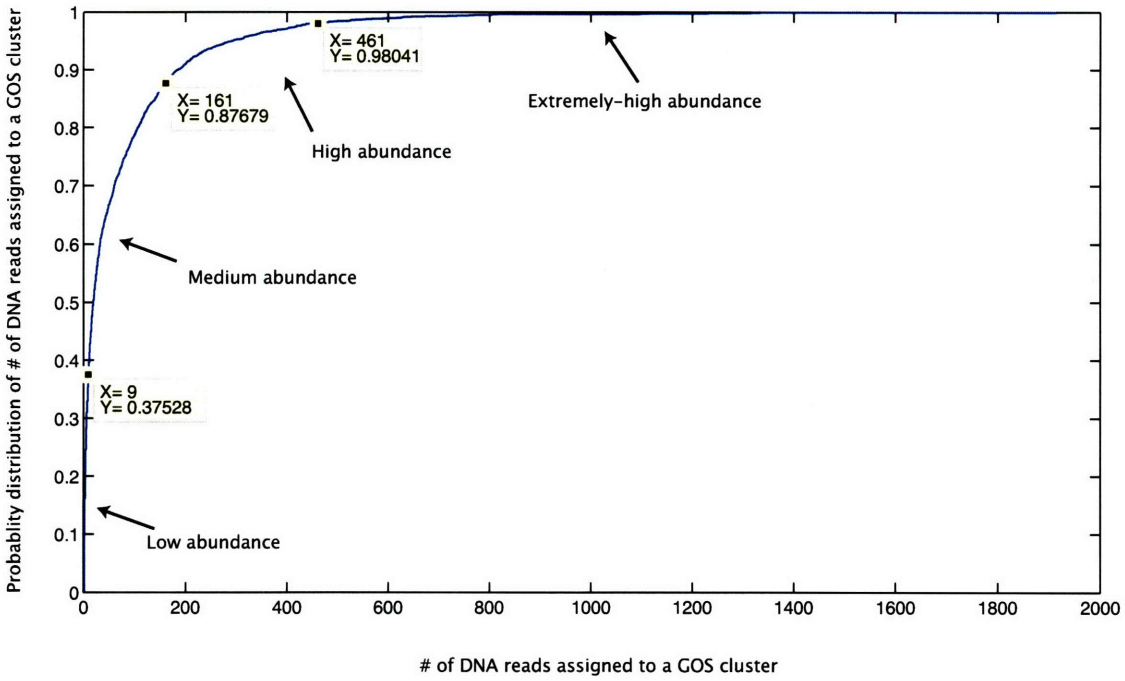
Supporting information. Fig. 8. Stringency of the BLASTX bitscore cutoff, in terms of alignment length and amino acid identity. Each circle represents an alignment between a cDNA pyrosequencing read and an NCBI-NR database sequence. Alignments with a bitscore >40 were considered significant in our analyses.



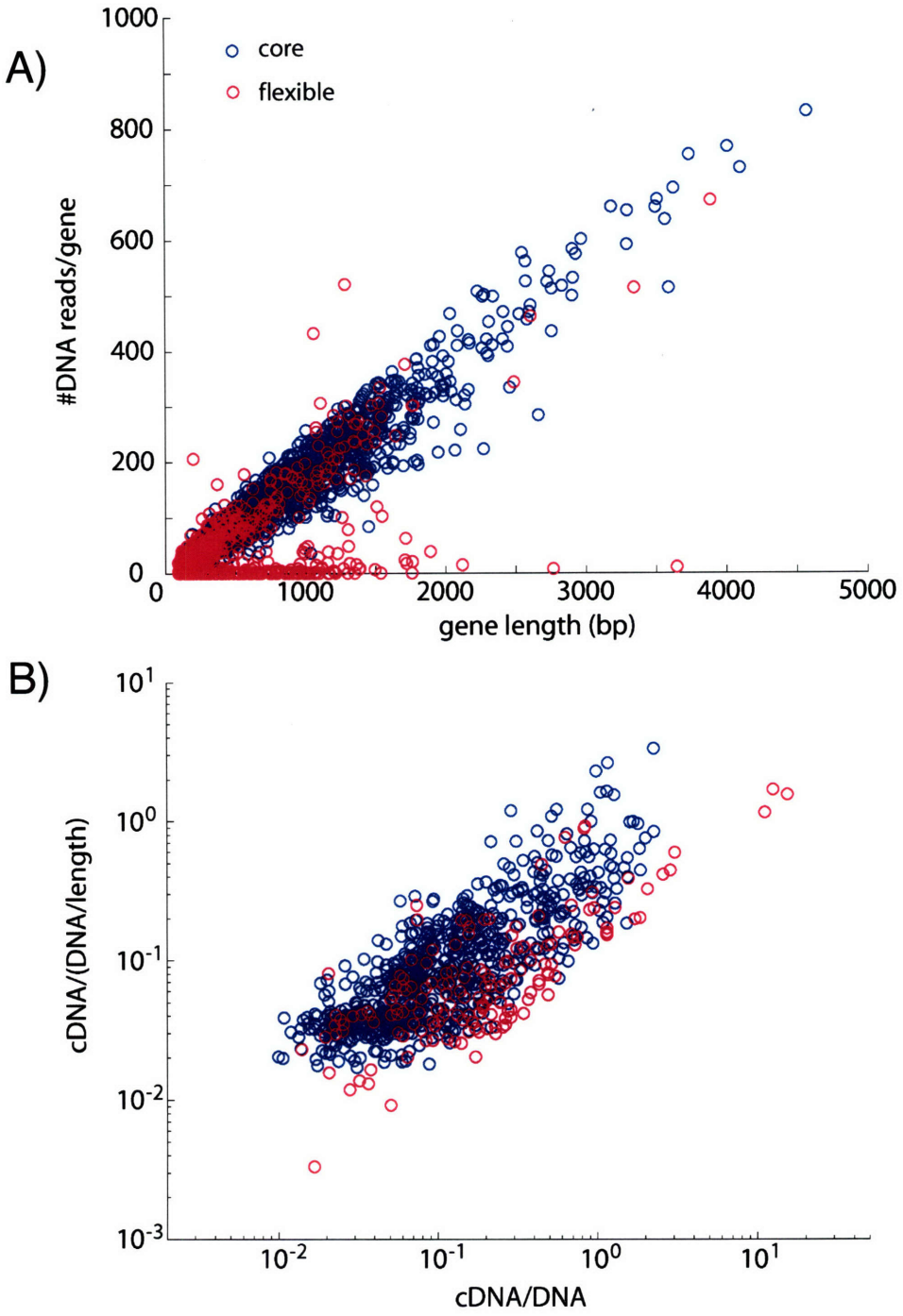
Supporting information. Fig. 9. Comparison of transcriptional levels of selected genes using pyrosequencing and RT-qPCR/qPCR. The unamplified environmental RNA and DNA samples were used for quantitative PCR. The cDNA to DNA ratio in qPCR analysis (x-axis) was calculated based on the modified method (see SI Methods). The cDNA to DNA ratio in pyrosequence analysis (y-axis) was normalized to the size of the respective libraries. More specifically the ratio was calculated as the fraction of reads assigned to the targeted gene in the cDNA library divided by that in the DNA library. Three sets of genes were selected based on their enrichment in the cDNA pyrosequence library. Green solid circle: genes with normalized cDNA/DNA ratio>1. Blue solid circle: genes with normalized cDNA/DNA ratio <1. Red open circle: Gene only detected in the cDNA library but not in the DNA library, and thus the cDNA/DNA ratio could not be calculated for pyrosequencing data (dotted part of y axis). The full names of the 10 selected genes are listed in SI Table 6.



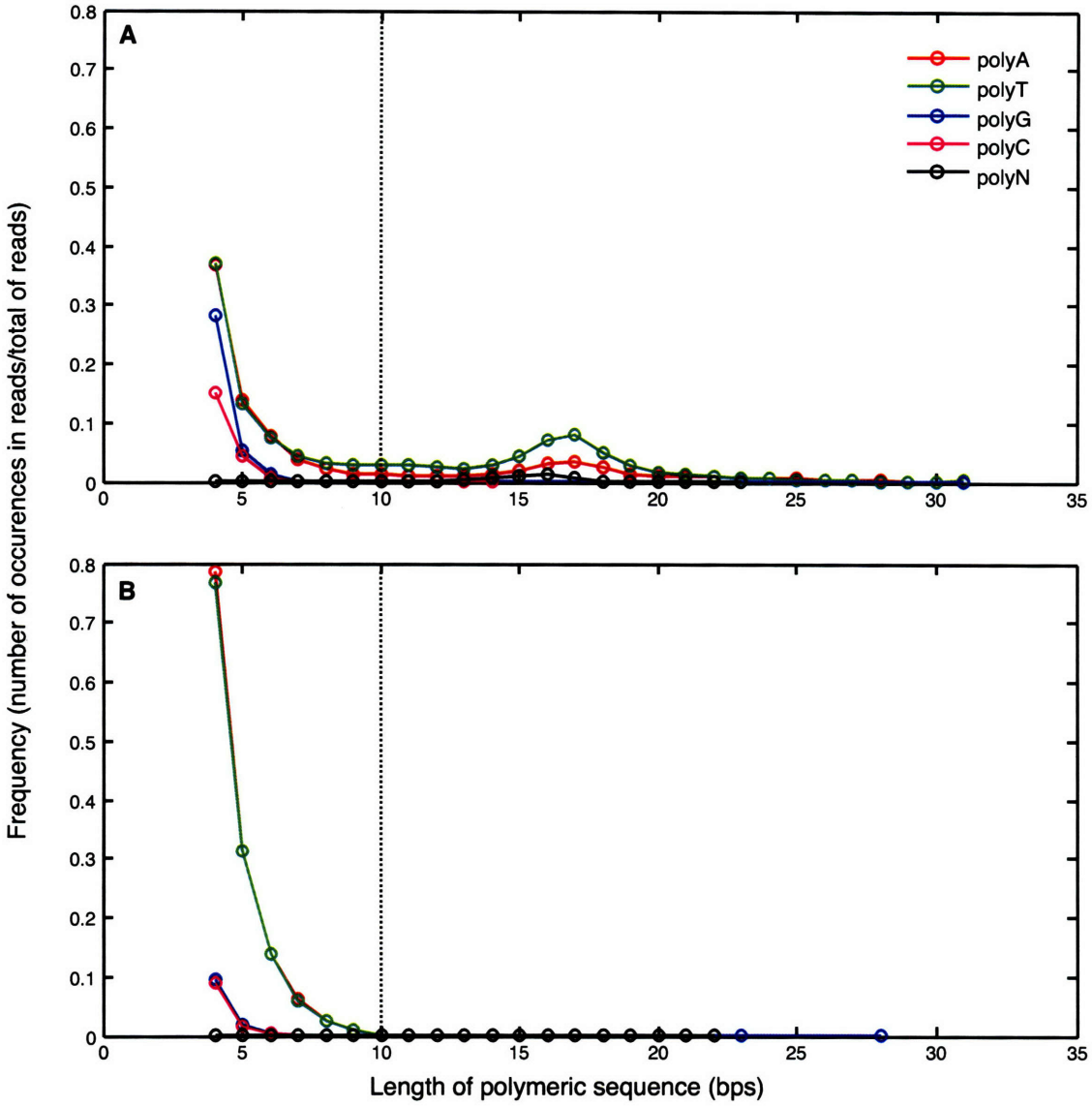
Supporting information. Fig. 10. Rarefaction analyses for cDNA and DNA libraries. The rarefaction analysis was based on the frequency of significant BLASTX matches in the GOS peptide database, with increasing number of Roche GS20 DNA pyrosequencing reads. Red dots represent the average values and the black dots represents the 95% confidence interval values.



Supporting information. Fig. 11. Empirical cumulative probability density function of the number of DNA reads assigned to a GOS protein cluster. The GOS protein clusters were arbitrarily binned to low, medium, high, and extremely high categories. Boundary values for each category, e.g., the number of DNA reads assigned to the cluster and its probability, are also shown.



Supporting information. Fig. 12. Effect of gene length on the number of hits in the DNA library, assessed using *Prochlorococcus* MIT9301. A) Linear relationship between the number of hits in the DNA database and gene length in the genome of MIT9301. B) Relationship between the normalized cDNA against DNA hits, and the normalized cDNA already normalized against gene length. In blue, core genes, i.e. genes present in all genomes of *Prochlorococcus* sequenced to date. In pink, flexible genes, i.e.genes not present in all genomes of *Prochlorococcus* sequenced.



Supporting information. Fig. 13. Distribution of the frequency of polymeric nucleotide sequence (A, T, G, C and N) lengths found in the A) 75m cDNA and B) 75m DNA pyrosequencing libraries. The peak in polymeric sequence length at 15-16 bps in the cDNA reads is a result of the polyadenylation in library preparation. The dashed line at 10bp indicates the cutoff used in the trimming of the cDNA data.

Table 2. Oligonucleotides used for qPCR analysis of genes identified by pyrosequencing. Sequences were compared against the NCBI-NTdatabase of nucleotide sequences using BLASTN.

Best hit in nr database Common, highly expressed	Oligonucleotide sequences 5'-3'	Comments
Thioredoxin peroxidase (Tpx)	TAT TAA GTG CTG AGA AAT CTT GA TGG GTT GTT CTA TTC TTT TAC CC	Specific only for Prochlorococcus MIT9312
Ammonium transporter (Amt)	ATTGGATTTGGAATTATGTATTAC AGTATTCCAGGAATTATTTCC	Specific only for Prochlorococcus MIT9312
Photosystem I PsaL protein (subunit XI) (PsaL)	TTG TTA ATC CGC CAA AGG AC AAG CAA AAA CAG CTC CTC CA	Amplifies Prochlorococcus MIT9301 and AS9601
Common, low expressed		
Alanyl-tRNA synthetase (AlaRS)	CAG ACA TGG GAG ATT TGT TAG G TCA GGA TAA TTA TTT TGC ATT AAA	Amplifies Prochlorococcus MIT9312 and MIT9301
Transcription-repair coupling factor (TRCF)	AAG GTT GAA ATC TAT TAT TTA TTG TTC TTA CAT CAG GCA AAC AGG TAA	Amplifies Prochlorococcus MIT9312, MIT9301 and AS9601
Phosphoribosylformylglycin amidine synthase II (FGAM synthaseII)	GCAGCAATAGTTCCTCTAAAAGGG TTC TGG TGT TGC TGC TTC TG	Amplifies Prochlorococcus MIT9312 and MIT9515
Cobaltochelatase, CobN subunit (CobN)	TTTTAATGCGAATGCTATTTGCC CCT ATA GAT TTG CCA GGT AAC CA	Amplifies Prochlorococcus MIT9301,MIT 9515, AS9601, MIT 9312 and MED4
Cobyrinic acid a,c-diamide synthase (CbiA)	AAG AGA ATT CAT ATT TCA AAG AAT GTT CCA ACC TAT TTG CAG GAA TTT	Amplifies Prochlorococcus 9301, 9515, AS9601, 9312 and MED4
Only in cDNA library		
Putative light-harvesting protein alpha chain (LHC)	AGCAATGATACATCTTGTTCTGC AGT TGC TGC TGC CTC AAA C	Specific for uncultured proteobacterium eBACred25D05
Predicted xylene monooxygenase hydroxylase component (XylM)	TTTGCAGTGTGATAACTCAT TGTGCTATCAACAGGTATATTTGCCGG	Specific for uncultured bacterium BAC13K9BAC

Table 3. Representatives of the GOS protein clusters that are unique to 75-m cDNA library

Cluster ID	Abundance	GO term	Pfam	TIGRfam	NR
14275698	667	–	–	–	–
11297554	28	–	–	–	ZP_01470602.1| hypothetical protein RS9916_32857 [Synechococcus sp. RS9916]
14230436	19	–	–	–	AAT90307.1| putative light-harvesting protein alpha chain [uncultured proteobacterium eBACred25D05]
12073604	14	photosynthesis\light reaction	–	–	ZP_01583951.1| antenna complex, alpha/beta subunit [Dinoroseobacter shibae DFL 12]
12023158	8	–	–	–	ZP_01470602.1| hypothetical protein RS9916_32857 [Synechococcus sp. RS9916]
11699146	6	–	–	–	YP_001008748.1| hypothetical protein A9601_03531 [Prochlorococcus marinus str. AS9601]
7478	4	translational initiation	–	–	–
11393514	4	photosynthesis\light reaction	–	–	AAT90308.1|putative light-harvesting protein beta chain [uncultured proteobacterium eBACred25D05]
19661	3	–	–	–	putative proteorhodopsin [uncultured bacterium]
11054015	3	–	–	–	CAL01029.1| chlorophyll a/b binding light harvesting protein pcbA [uncultured Prochlorococcus sp.]
16914	3	–	–	–	ZP_01255953.1| Substrate-binding region of ABC-type glycine betaine transport system [Psychroflexus torquis ATCC 700755]
17232	2	transcription	–	–	EAZ99485.1|DNA-directed RNA polymerase subunit beta [Marinobacter sp. ELB17]
14025838	2	transport	–	–	ZP_00949339.1|putative outer membrane protein [Croceibacter atlanticus HTCC2559]
14212924	1	–	TonB-dependent receptor	–	–

Table 4. Representatives of the GOS protein clusters that are unique to 75-m DNA library

Cluster ID	Abundance	GO term	Pfam	TIGRfam	NR
174	333	de novo' IMP biosynthesis	–	–	GAR transformylase 2 [Prochlorococcus marinus str. MIT 9301]
260	245	–	–	–	Glycosyl transferase, family 2 [Prochlorococcus marinus str. MIT 9301]
5431	241	lipopolysaccharide biosynthesis	–	–	Glycosyl transferase, family 2 [Prochlorococcus marinus str. MIT 9301]
700	209	mismatch repair	–	–	putative DNA mismatch repair protein MutS family [Prochlorococcus marinus str. AS9601]
442	200	–	–	small_GTP: small GTP-binding protein domain	Small GTP-binding protein domain [Prochlorococcus marinus str. MIT 9312]
3200	198	urea metabolism	Amidohydrolase family	urease_alph: urease, alpha subunit	Urease alpha subunit [Prochlorococcus marinus str. MIT 9301]
3868	196	lipopolysaccharide biosynthesis	–	–	UDP-N-acetylglucosamine pyrophosphorylase [Prochlorococcus marinus str. MIT 9301]
152	193	cobalamin biosynthesis	–	–	precorrin-2 C20-methyltransferase [uncultured Prochlorococcus marinus clone ASNC2259]
428	190	tryptophanyl-tRNA aminoacylation	–	trpS: tryptophanyl-tRNA synthetase	Tryptophanyl-tRNA synthetase [Prochlorococcus marinus str. AS9601]
1225	190	coenzyme A biosynthesis	–	–	ATP/GTP-binding site motif A (P-loop) [Prochlorococcus marinus str. AS9601]
4133	184	amino acid biosynthesis	Homoserine dehydrogenase	–	YP_001009547.1| Homoserine dehydrogenase:ACT domain-containing protein [Prochlorococcus marinus str. AS9601]
2731	180	intracellular protein transport	–	chloroplast envelope protein translocase, IAP75 family	outer envelope membrane protein-like protein [Prochlorococcus marinus str. AS9601]
3940	176	GTP biosynthesis	Radical SAM superfamily		Fe-S oxidoreductase [Prochlorococcus marinus str. MIT 9301]
288	173	–	–	–	DEAD/DEAH box helicase:Helicase C-terminal domain-containing protein [Prochlorococcus marinus str.AS9601]
133	172	pentose-phosphate shunt	Transaldolase	transaldolase	Transaldolase [Prochlorococcus marinus str. AS9601]
2871	171	electron transport	Pyridine nucleotide-disulphide oxidoreductase		Selenide,water dikinase [Prochlorococcus marinus str. MIT 9301]

Table 5. Phylogenetic diversity of DNA and cDNA libraries computed by MEGAN after removal of rRNA sequences from the databases. BLASTX results with a score (bits) cutoff of 40 was used to construct the trees. The cutoff level used to assign the different hits was phylum as described in MEGAN. Color-coding corresponds to that in Fig.2. Bacteria: green. Archaea: red. Eukaryota: blue. Viruses: lightblue. Taxa within each kingdom have been ordered by rank abundance based on the total number of hits in the DNA library.

Phylum	Number of hits in the DNA library	Number of hits in the cDNA library	Percentage (%) of hits in the DNA library	Percentage (%) of hits in the cDNA library
Cyanobacteria	142,084	4,167	34.313	7.636
Proteobacteria	50,506	2,413	12.197	4.422
Bacteroidetes	9,943	375	2.401	0.687
Firmicutes	2,477	243	0.598	0.445
Actinobacteria	1,507	26	0.364	0.048
Planctomycetes	561	8	0.135	0.015
Chlorobi	517	9	0.125	0.016
Chloroflexi	335	13	0.081	0.024
Spirochaetes	251	6	0.061	0.011
Acidobacteria	219	5	0.053	0.009
Thermotogae	191	0	0.046	0
Deinococcus-Thermus	113	0	0.027	0
Verrucomicrobia	112	0	0.027	0
Fusobacteria	83	0	0.020	0
Aquificae	63	0	0.015	0
Chlamidiae	47	2	0.011	0.004
Nitrospirae	41	0	0.010	0
candidate division WS3	7	0	0.002	0
Unclassified bacteria	4	0	0.001	0
Candidate division 0P8	4	0	0.001	0
Candidatus Poribacteria	4	0	0.001	0
Dictyoglomi	2	0	0.0005	0
Euryarchaeota	708	10	0.171	0.018
Crenarchaeota	168	0	0.041	0.000
Nanoarchaeota	3	0	0.001	0.000
Streptophyta	509	18	0.123	0.033
Chordata	495	21	0.120	0.038
Ascomycota	468	4	0.113	0.007
Chlorophita	307	9	0.074	0.016
Arthropoda	257	15	0.062	0.027
Ciliophora	167	16	0.040	0.029
Apicomplexa	166	6	0.040	0.011
Cnidaria	157	0	0.038	0.000
Mycetozoa	140	13	0.034	0.024
Echinodermata	110	3	0.027	0.005

Table 5, continued.

Phylum	Number of hits in the DNA library	Number of hits in the cDNA library	Percentage (%) of hits in the DNA library	Percentage (%) of hits in the cDNA library
Basidiomycota	93	3	0.022	0.005
Kinetoplastida	74	2	0.018	0.004
Nematoda	65	2	0.016	0.004
Parabasalidea	54	2	0.013	0.004
Haptophyceae	39	8	0.009	0.015
Bacillariophyta	29	5	0.007	0.009
Rhodophyta	28	0	0.007	0.000
Cryptophyta	22	0	0.005	0.000
Entamoebidae	19	0	0.005	0.000
Platyhelminthes	14	6	0.003	0.011
Dinophyceae	10	2	0.002	0.004
Cercozoa	9	2	0.002	0.004
Oomycetes	8	3	0.002	0.005
Synurophyceae	7	3	0.002	0.005
Jakobidae	7	0	0.002	0
Microsporidia	5	0	0.001	0
Mollusca	4	0	0.001	0
Diplomonadina group	4	0	0.001	0
Glaucocystophyceae	3	0	0.001	0
Phaeophyceae	3	2	0.001	0.004
Euglenida	3	0	0.001	0
Raphidophyceae	2	0	0.0005	0
Bicosoecida	2	0	0.0005	0
Porifera	2	0	0.0005	0
Heterolobosea	2	0	0.0005	0
Malawimonadidae	0	2	0	0.004
Viruses	2,102	24	0.508	0.044
Not assigned	68,961	8,304	16.654	15.218
No hits	129,780	38,816	31.342	71.133
Total	414,077	54,568	100.000	100.000

Table 6. Top 20 *Prochlorococcus* highly expressed genes in the cDNA library depending on the kind of normalization applied on the dataset.

Raw cDNA		cDNA/DNA		cDNA/(DNA/length)	
P9301_02861	Ammonium transporter family	P9301_11381	no description	P9301_02861	Ammonium transporter family
P9301_17151	Photosystem I PsaB protein	P9301_03541	no description	P9301_17151	Photosystem I PsaB protein
P9301_17161	Photosystem I PsaA protein	P9301_07111	no description	P9301_17161	Photosystem I PsaA protein
P9301_05761	Ribulose bisphosphate carboxylase, large chain	P9301_04361	Predicted protein	P9301_03541	no description
P9301_06541	Chlorophyll a/b binding light harvesting protein PcbD	P9301_03421	Photosystem II reaction center M protein (PsbM)	P9301_05761	Ribulose bisphosphate carboxylase, large chain
P9301_11381	no description	P9301_13581	no description	P9301_03401	Photosystem II PsbB protein (CP47)
P9301_03401	Photosystem II PsbB protein (CP47)	P9301_02911	conserved hypothetical protein	P9301_11381	no description
P9301_16991	Elongation factor Tu	P9301_02861	Ammonium transporter family	P9301_16991	Elongation factor Tu
P9301_13501	Photosystem II PsbC protein (CP43)	P9301_00661	Predicted protein	P9301_13501	Photosystem II PsbC protein (CP43)
P9301_13921	polyribonucleotide nucleotidyltransferase	P9301_17021	30S ribosomal protein S12	P9301_13921	polyribonucleotide nucleotidyltransferase (pnp)
P9301_13491	Photosystem II PsbD protein (D2)	P9301_09571	50S ribosomal protein L28	P9301_16741	RNA polymerase beta prime subunit
P9301_16741	RNA polymerase beta prime subunit	P9301_12251	Possible high light inducible protein	P9301_07111	no description
P9301_04291	30S ribosomal protein S4	P9301_05771	Ribulose bisphosphate carboxylase, small chain	P9301_17001	Elongation factor G
P9301_12261	Porin homolog	P9301_02221	50S ribosomal protein L10	P9301_13491	Photosystem II PsbD protein (D2)
P9301_07111	no description	P9301_17121	photosystem I subunit VIII (PsaI)	P9301_10121	thioredoxin peroxidase
P9301_02911	conserved hypothetical protein	P9301_10121	thioredoxin peroxidase	P9301_04291	30S ribosomal protein S4
P9301_17001	Elongation factor G	P9301_04291	30S ribosomal protein S4	P9301_02221	50S ribosomal protein L10
P9301_10121	thioredoxin peroxidase	P9301_16451	ATP synthase subunit c	P9301_02451	no description
P9301_17111	Photosystem I PsaL protein (subunit XI)	P9301_03211	Cytochrome b559 alpha-subunit	P9301_06541	Chlorophyll a/b binding light harvesting protein PcbD
P9301_18031	Transketolase	P9301_06071	plastocyanin	P9301_18031	Transketolase

Appendix C

Patterns and Implications of Gene Gain and Loss in the Evolution of *Prochlorococcus*

Gregory C. Kettler, Adam C. Martiny, Katherine Huang, Jeremy Zucker, Maureen L. Coleman, Sebastien Rodrigue, Feng Chen, Alla Lapidus, Steven Ferriera, Justin Johnson, Claudia Steglich, George M. Church, Paul Richardson, and Sallie W. Chisholm

Kettler, G.C., Martiny, A.C., Huang, K., Zucker, J., **Coleman, M.L.**, Rodrigue, S., Chen, F., Lapidus, A., Ferriera, S., Johnson, J., Steglich, C., Church, G.M., Richardson, P. and Chisholm, S.W. (2007) Patterns and Implications of Gene Gain and Loss in the Evolution of *Prochlorococcus*. *PLoS Genetics* 3: e231.

OPEN ACCESS Freely available online PLoS GENETICS

Patterns and Implications of Gene Gain and Loss in the Evolution of *Prochlorococcus*

Gregory C. Kettler[1,2☯], Adam C. Martiny[2☯¤], Katherine Huang[2], Jeremy Zucker[3], Maureen L. Coleman[2], Sebastien Rodrigue[2], Feng Chen[4], Alla Lapidus[4], Steven Ferriera[5], Justin Johnson[5], Claudia Steglich[6], George M. Church[3], Paul Richardson[4], Sallie W. Chisholm[1,2*]

1 Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **2** Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **3** Department of Genetics, Harvard Medical School, Boston, Massachusetts, United States of America, **4** Joint Genome Institute, United States Department of Energy, Walnut Creek, California, United States of America, **5** J. Craig Venter Institute, Rockville, Maryland, United States of America, **6** Department of Biology II/Experimental Bioinformatics, University Freiburg, Freiburg, Germany

Prochlorococcus **is a marine cyanobacterium that numerically dominates the mid-latitude oceans and is the smallest known oxygenic phototroph. Numerous isolates from diverse areas of the world's oceans have been studied and shown to be physiologically and genetically distinct. All isolates described thus far can be assigned to either a tightly clustered high-light (HL)-adapted clade, or a more divergent low-light (LL)-adapted group. The 16S rRNA sequences of the entire *Prochlorococcus* group differ by at most 3%, and the four initially published genomes revealed patterns of genetic differentiation that help explain physiological differences among the isolates. Here we describe the genomes of eight newly sequenced isolates and combine them with the first four genomes for a comprehensive analysis of the core (shared by all isolates) and flexible genes of the *Prochlorococcus* group, and the patterns of loss and gain of the flexible genes over the course of evolution. There are 1,273 genes that represent the core shared by all 12 genomes. They are apparently sufficient, according to metabolic reconstruction, to encode a functional cell. We describe a phylogeny for all 12 isolates by subjecting their complete proteomes to three different phylogenetic analyses. For each non-core gene, we used a maximum parsimony method to estimate which ancestor likely first acquired or lost each gene. Many of the genetic differences among isolates, especially for genes involved in outer membrane synthesis and nutrient transport, are found within the same clade. Nevertheless, we identified some genes defining HL and LL ecotypes, and clades within these broad ecotypes, helping to demonstrate the basis of HL and LL adaptations in *Prochlorococcus*. Furthermore, our estimates of gene gain events allow us to identify highly variable genomic islands that are not apparent through simple pairwise comparisons. These results emphasize the functional roles, especially those connected to outer membrane synthesis and transport that dominate the flexible genome and set it apart from the core. Besides identifying islands and demonstrating their role throughout the history of *Prochlorococcus*, reconstruction of past gene gains and losses shows that much of the variability exists at the "leaves of the tree," between the most closely related strains. Finally, the identification of core and flexible genes from this 12-genome comparison is largely consistent with the relative frequency of *Prochlorococcus* genes found in global ocean metagenomic databases, further closing the gap between our understanding of these organisms in the lab and the wild.**

Citation: Kettler CG, Martiny AC, Huang K, Zucker J, Coleman ML, et al. (2007) Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. PLoS Genet 3(12): e231. doi:10.1371/journal.pgen.0030231

Introduction

The oceans play a key role in global nutrient cycling and climate regulation. The unicellular cyanobacterium *Prochlorococcus* is an important contributor to these processes, as it accounts for a significant fraction of primary productivity in low- to mid-latitude oceans [1]. *Prochlorococcus* and its close relative, *Synechococcus* [2], are distinguished by their photosynthetic machinery: *Prochlorococcus* uses chlorophyll-binding proteins instead of phycobilisomes for light harvesting and divinyl instead of monovinyl chlorophyll pigments. Although *Prochlorococcus* and *Synechococcus* coexist throughout much of the world's oceans, *Synechococcus* extends into more polar regions and is more abundant in nutrient-rich waters, while *Prochlorococcus* dominates relatively warm, oligotrophic regions and can be found at greater depths [3]. The

Editor: David Guttman, University of Toronto, Canada

Received July 30, 2007; **Accepted** November 13, 2007; **Published** December 21, 2007

A previous version of this article appeared as an Early Online Release on November 13, 2007 (doi:10.1371/journal.pgen.0030231.eor).

Abbreviations: HL, high-light; LGT, lateral gene transfer; LL, low-light

* To whom correspondence should be addressed. E-mail: chisholm@mit.edu

☯ These authors contributed equally to this work.

¤ Current address: Department of Earth System Science and Department of Ecology and Evolutionary Biology, University of California, Irvine, California, United States of America

Author Summary

Prochlorococcus—the most abundant photosynthetic microbe living in the vast, nutrient-poor areas of the ocean—is a major contributor to the global carbon cycle. *Prochlorococcus* is composed of closely related, physiologically distinct lineages whose differences enable the group as a whole to proliferate over a broad range of environmental conditions. We compare the genomes of 12 strains of *Prochlorococcus* representing its major lineages in order to identify genetic differences affecting the ecology of different lineages and their evolutionary origin. First, we identify the core genome: the 1,273 genes shared among all strains. This core set of genes encodes the essentials of a functional cell, enabling it to make living matter out of sunlight and carbon dioxide. We then create a genomic tree that maps the gain and loss of non-core genes in individual strains, showing that a striking number of genes are gained or lost even among the most closely related strains. We find that lost and gained genes commonly cluster in highly variable regions called genomic islands. The level of diversity among the non-core genes, and the number of new genes added with each new genome sequenced, suggest far more diversity to be discovered.

Prochlorococcus group consists of two major ecotypes, high-light (HL)-adapted and low-light (LL)-adapted, that are genetically and physiologically distinct [4] and are distributed differently in the water column [5,6]. Given their relatively simple metabolism, well-characterized marine environment, and global abundance, these marine cyanobacteria represent an excellent system for understanding how genetic differences translate to physiological and ecological variation in natural populations.

The first marine cyanobacterial genome sequences suggested progressive genome decay from *Synechococcus* to LL *Prochlorococcus* to HL *Prochlorococcus*, characterized by a reduction in genome size (from 2.4 to 1.7 Mb) and a drop in G + C content from ~59% to ~30% [7–9]. Notably, genes involved in light acclimation and nutrient assimilation appeared to have been sequentially lost, consistent with the niche differentiation observed for these three groups [7]. This comparison suggested that the major clades of marine cyanobacteria differentiated in a stepwise fashion, leading to patterns of gene content that corresponded to the isolates' 16S rRNA phylogeny.

Recently, however, molecular sequence data and physiology studies have revealed complexity beyond the HL/LL paradigms. Within the LL ecotype, for instance, some but not all isolates can use nitrite as a sole nitrogen source [10], and the LL genomes range widely in size [7,8]. Moreover, the distribution of phosphate acquisition genes among *Prochlorococcus* genomes does not correlate to their rRNA phylogeny but instead appears related to phosphate availability: strains isolated from low-phosphate environments are genetically better equipped to deal with phosphate limitation than those from high-phosphate environments, regardless of their 16S rRNA phylogeny [11]. Thus, while the HL/LL distinction has held up both phenotypically and genotypically, there are other differences among isolates that are not consistent with their rRNA phylogeny. Thus, to understand diversification and adaptation in this globally important group, we must characterize the underlying patterns of genome-wide diversity.

Lateral gene transfer (LGT) is one mechanism that creates complex gene distributions and phylogenies incongruent with the rRNA tree. The question of whether a robust organismal phylogeny can be inferred despite extensive LGT is still hotly debated [12,13]. If a core set of genes exists that is resistant to LGT, then gene trees based on these core genes should reflect cell division and vertical descent, as has been argued for the gamma *Proteobacteria* [13]. Others argue that genes in a shared taxon core do not necessarily have the same evolutionary histories, making inference of an organismal phylogeny difficult [14]. In spite of this debate, the core genome remains a useful concept for understanding biological similarity within a taxonomic group. Recent compar-

Table 1. General Characteristics of the *Prochlorococcus* and *Synechococcus* Isolates Used in This Study

Cyanobacterium	Isolate	Light Adaptation	Length (bp)	GC %	Number of Genes[a]	Isolation Depth	Region	Date	Accession Number	Reference
Prochlorococcus	MED4	HL(I)	1,657,990	30.8	1,929	5m	Med. Sea	Jan. 1989	BX548174	[7,38]
	MIT9515[b]	HL(I)	1,704,176	30.8	1,908	15m	Eq. Pacific	Jun. 1995	CP000552	[18]
	MIT9301[b]	HL(II)	1,642,773	31.4	1,907	90m	Sargasso Sea	Jul. 1993	CP000576	[18]
	AS9601[b]	HL(II)	1,669,886	31.3	1,926	50m	Arabian Sea	Nov. 1995	CP000551	[21]
	MIT9215[b]	HL(II)	1,738,790	31.1	1,989	5m	Eq. Pacific	Oct. 1992	CP000825	[19]
	MIT9312	HL(II)	1,709,204	31.2	1,962	135m	Gulf Stream	Jul. 1993	CP000111	[4,60]
	NATL1A[b]	LL(I)	1,864,731	35.1	2,201	30m	N. Atlantic	Apr. 1990	CP000553	[20]
	NATL2A[b]	LL(I)	1,842,899	35	2,158	10m	N. Atlantic	Apr. 1990	CP000095	[22]
	SS120	LL(II)	1,751,080	36.4	1,925	120m	Sargasso Sea	May 1988	AE017126	[8,26]
	MIT9211[b]	LL(III)	1,688,963	38	1,855	83m	Eq. Pacific	Apr. 1992	CP000878	[19]
	MIT9303[b]	LL(IV)	2,682,807	50.1	3,022	100m	Sargasso Sea	Jul. 1992	CP000554	[4]
	MIT9313	LL(IV)	2,410,873	50.7	2,843	135m	Gulf Stream	Jul. 1992	BX548175	[4,27]
Synechococcus	CC9311	Syn.	2,606,748	52.5	3017	95m	Calif. Current	1993	CP000435	[9]
	CC9902	Syn.	2,234,828	54.2	2504	5m	Calif. Current	1999	CP000097	Palenik, unpublished data
	WH8102	Syn.	2,434,428	59.4	2787		Sargasso Sea	Mar. 1981	BX548020	[2,36]
	CC9605	Syn.	2,510,659	59.2	2991	51m	Calif. Current	1996	CP000110	Palenik, unpublished data

[a]Number of protein coding genes excluding pseudogenes

[b]Genomes of isolates are being reported for the first time here. The gene counts of previously published genomes are slightly different from those of earlier reports [7,8,60] as new annotation pipelines have identified more genes. References refer to either the paper in which the genome was first reported, or the first paper describing the particular isolate.

doi:10.1371/journal.pgen.0030231.t001

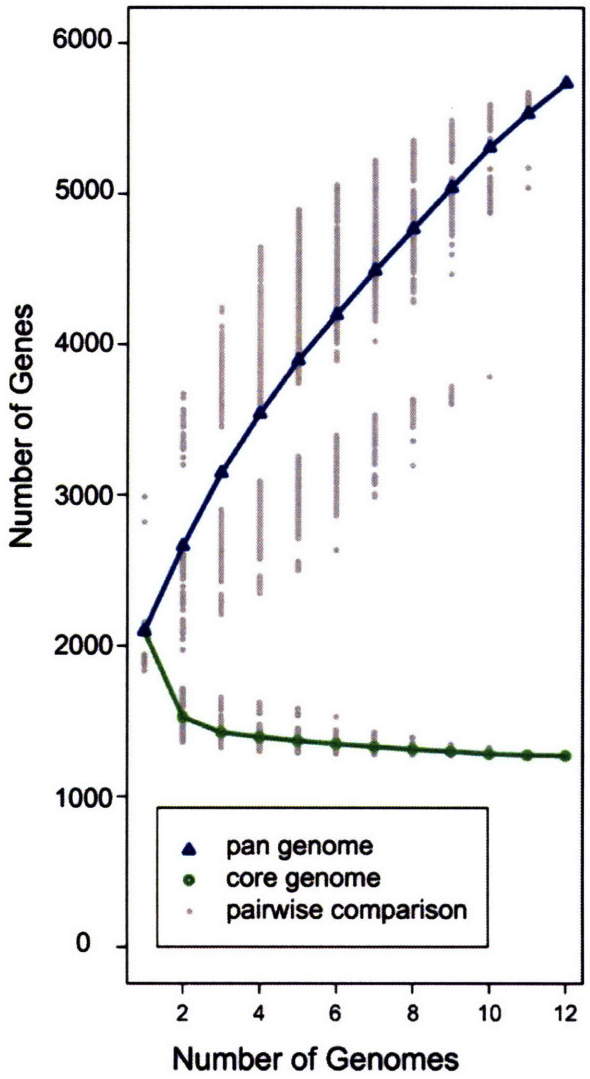


Figure 1. The Sizes of the Core and Pan-Genomes of *Prochlorococcus*

The calculated sizes depend on the number of genomes used in the analysis. If k genomes are selected from 12, there are 12!/(k!(12 − k)!) possible selections from which to calculate the core and pan-genomes. Each possible selection is plotted as a grey point, and the line is drawn through the average. This analysis is based on a similar one in [15].

doi:10.1371/journal.pgen.0030231.g001

isons within the lactic acid bacteria, cyanobacteria, and *Streptococcus agalacticae* groups, for instance, have each revealed a core set of genes shared by all members of the group, on top of which is layered the flexible genome [15–17]. The vast majority of genes in the core genome encode housekeeping functions, while genes in the flexible genome reflect adaptation to specific environments [16] and are often acquired by LGT. Thus the core and flexible genomes are informative not only in a phylogenetic context, for understanding the mechanisms and tempo of genome evolution, but also in an ecological context, for understanding the selective pressures experienced in different environments.

To further understand diversification and adaptation in *Prochlorococcus*, we obtained sequences of eight additional genomes representing diverse lineages, both LL- and HL-adapted, spanning the complete 16S rRNA diversity (97% to 99.93% similarity) of cultured representatives of this group [18–22] (Table 1). Comparing these genomes with available genomes for *Prochlorococcus* and marine *Synechococcus*, our goal was to reconstruct the history of vertical transmission, gene acquisition, and gene loss for these marine cyanobacteria. In particular we identified functions associated with the core and flexible genomes and analyzed the metabolic pathways encoded in each. This analysis reveals not only what differentiates *Synechococcus* from LL *Prochlorococcus* from HL *Prochlorococcus*, but also informs our understanding of how adaptation occurs in the oceans along gradients of light, nutrients, and other environmental factors, providing essential biological context for interpreting rapidly expanding metagenomic datasets.

Results/Discussion

Core Genome

The genomes of 12 *Prochlorococcus* isolates, representing all known major phylogenetic clades, range in size from 1.6 Mbp (MIT9301) to 2.7 Mbp (MIT9303) (Table 1). As more genomes are compared, we observe an asymptotic decline in the number of shared (core) genes (Figure 1), similar to observations for *Streptococcus* genomes [15]. This suggests a finite size of the core genome of approximately 1,250 genes, or 40% to 67% of the genes of any particular isolate. In contrast, the pan-genome [15,23] of these isolates, encompassing the core genes, plus the total of all additional genes found in any of the isolates (the "flexible genes"), contains 5,736 genes (Table S1). The gene accumulation curve as more genomes are added to the analysis is clearly far from saturated (Figure 1), indicating a far larger gene pool within the *Prochlorococcus* clade than is captured by our sequenced isolates, and suggesting the presence of *Prochlorococcus* lineages in the wild, with yet-to-be discovered traits.

Although the closely related marine cyanobacterium *Synechococcus* commonly coexists with *Prochlorococcus*, it is considered more of a generalist, and, collectively, is capable of growth over a broader range of nutrient concentrations and temperatures than is *Prochlorococcus*. To understand the divergence of marine *Synechococcus* and *Prochlorococcus* since their last common ancestor, we looked for genes present in all *Prochlorococcus* but absent from some or all *Synechococcus*. We found 33 such genes, 13 of which are not found in any sequenced marine *Synechococcus* (Table S2). Eight of these *Prochlorococcus*-only genes have been assigned putative functions including one HL inducible protein (MED4's *hli11*, which responds only slightly to light stress [24]), a possible sodium-solute symporter, an iron-sulfur protein, and a *deoR*-like transcription factor, but it is unclear what role these genes have in distinguishing *Prochlorococcus* from *Synechococcus*. Perhaps more importantly, the differentiation between these two groups is defined by the absence in *Prochlorococcus* of 140 genes that are present in all four sequenced marine *Synechococcus* (Table S3). All *Prochlorococcus* isolates sequenced to date lack, for example, divinyl protochlorophyllide a reductase (*dvr*) [25], resulting in one of the defining phenotypic properties of *Prochlorococcus*: divinyl chlorophyll *a* as the primary light harvesting pigment [26]. Other light harvesting genes absent in *Prochlorococcus* include allophyco-

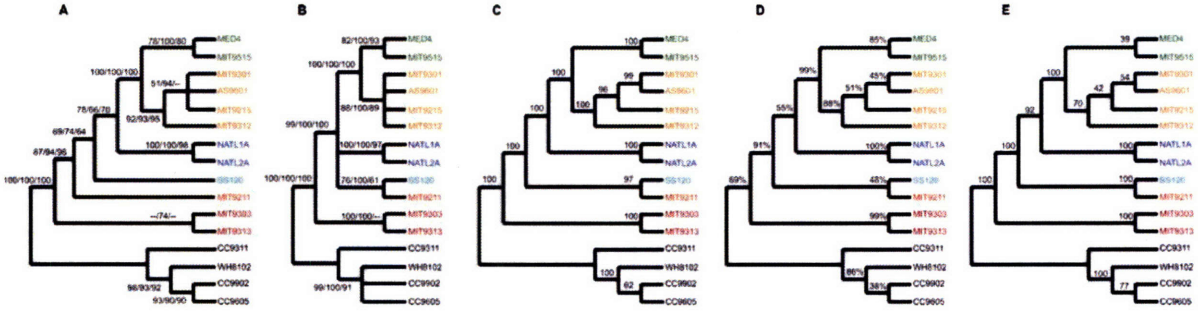


Figure 2. Phylogenetic Relationship of *Prochlorococcus* and *Synechococcus* Reconstructed by Multiple Methods

(A) 16S rRNA and (B) 16S-23S rRNA ITS region reconstructed with maximum parsimony, neighbor-joining, and maximum likelihood. Numbers represent bootstrap values (100 resamplings).

(C) Maximum parsimony reconstruction of random concatenation of 100 protein sequences sampled from core genome. Values represent average bootstrap values (100 resamplings) from 100 random concatenation runs.

(D) Consensus tree of all core genes using maximum parsimony on protein sequence alignments. Values represent fraction of genes supporting each node.

(E) Genome phylogeny based on gene content using the approach of [34]. Values represent bootstrap values from 100 resamplings.

doi:10.1371/journal.pgen.0030231.g002

cyanin (*apcABCDE*), some phycoerythrins, and phycobilisome linkers. *Synechococcus* also possess several molybdopterin biosynthesis enzymes not found in *Prochlorococcus* (*mobA*, *moaABCDE*), which may be necessary for the function of nitrate reductase [27,28]. Although all 12 *Prochlorococcus* isolates also lack the gene for nitrate reductase, this might be a result of the isolation conditions, and further study may reveal nitrate-utilizing isolates [29].

The underpinnings of *Prochlorococcus* diversity should be reflected in the respective roles of the core and flexible genomes. If the core genome provides for central metabolic needs shared by all isolates, it should be possible to reconstruct those pathways with the core genes alone. Therefore we asked whether the core genome encodes all the biochemical pathways needed for growth from the nutrients available to *Prochlorococcus* using Pathway Tools [30] and compared the resulting map with the manually curated, but less detailed, metabolic map for *Prochlorococcus* SS120 [8]. The automated approach is more detailed (Figures S1–S4 and see http://procyc.mit.edu), but the results recapitulate the previous manual effort.

We have identified core genes responsible for nearly all the reactions in the central metabolism, from the Calvin Cycle to the incomplete TCA cycle, including pathways to synthesize all 20 amino acids, several cofactors, and chlorophylls (Figures S1–S4). Among the genes that were assigned functions in the *Prochlorococcus* SS120 core metabolic model, all but seven are found to be part of the core genome in this study. Five of these seven additional genes in SS120 are transporters: SS120_12271, an iron or manganese transporter; SS120_15671, a sodum/alanine symporter; and SS120_06831–06851, three genes encoding an ABC-type amino acid transporter. The other two, *sdhA* and *sdhB*, are putatively responsible for the conversion of fumarate to succinate in the incomplete TCA cycle, but they have no apparent orthologs in many *Prochlorococcus* isolates. Importantly, *sdhAB* in the TCA cycle and *pdxH* in pyridoxal phosphate synthesis are the only cases in which one of the pathways examined could be reconstructed in some strains, but not in the core genome. An additional case, the phosphorylation of pantothenate in coenzyme A synthesis, is incomplete in the core and pan reconstructions, indicating that we have most likely failed to identify the gene or an alternate pathway (Figure S4). This observation supports the view in which essential life functions are unchanging across all *Prochlorococcus*, while nonessential or environment-specific functions are found in the flexible genome (see below). The functions of the latter, then, may relate to niche-specific adaptations that are not required for growth under optimal conditions, but that provide a fitness advantage in particular habitats. The pattern of their gain and loss in phylogenetic space could therefore help us understand when and how *Prochlorococcus* lineages evolved adaptations to particular environments. However, a close examination of their gain and loss requires a robust phylogenetic tree as a scaffold for analysis.

Phylogeny of *Prochlorococcus* Isolates Using the Core Genomes

Identification of the core genome shared by all *Prochlorococcus* isolates provides a new opportunity for determining the phylogenetic relationship among isolates. Our current understanding of the branching order among isolates is based on single gene phylogenies including 16S rRNA [10], 16S-23S rRNA internal transcribed spacer sequence (ITS) [18], *rpoC1* [31], *psbA* [32], and *petBD* [6]. Although trees based on these genes generally agree on the phylogenetic position of most isolates, they disagree, or lack bootstrap support, for the branching order of internal nodes among LL isolates (see Figure 2A and 2B for 16S rRNA and ITS trees). To reconstruct a robust phylogeny, we randomly concatenated 100 protein sequences from a pool of all core genes and compared the topology of the resulting trees (Figure 2C), analogous to the approach described by Rokas and coworkers [33]. This random concatenation was repeated 100 times and the same highly supported topology emerged every time. This tree is very similar to the 16S rRNA tree (Figure 2A) except for the position of LL isolates MIT9211 and SS120. We attribute this discrepancy to the limited information in any single gene (including 16S rRNA), and our analysis suggests that MIT9211 and SS120 form a separate clade. Each node in the concatenated protein tree is also supported by a

plurality of individual core genes (as defined above) (Figure 2D). Based on these results, we postulate that this tree represents the most probable evolutionary relationship among *Prochlorococcus* isolates. However, it is unclear if the physiology of SS120 and MIT9211 warrants considering them as one or separate ecotypes. Furthermore, many single gene phylogenies supported alternative topologies for this node, and future analyses with more genomes or alternative phylogenetic approaches may result in different topologies for this node.

The history of *Prochlorococcus* is marked not only by sequence divergence among the core genes, but also by the gain and loss of genes. We constructed a dendrogram based on the presence or absence of individual orthologous groups (Figure 2E) [34]. Again, the topology of this tree is identical to that of Figure 2C. This suggests that shared gene content among *Prochlorococcus* isolates is significantly influenced by the isolates' phylogenetic relationship despite the occurrence of lateral gene gain and loss.

Flexible Genome

Patterns of gene gain and loss in the evolutionary tree. We used our most probable phylogenetic tree (Figure 2C) as a map for the evolution of each isolate and superimposed the gain and loss of flexible genes (i.e., non-core) upon it (Figure 3A). By assigning costs to gain and loss events (see Methods) and then minimizing the total cost (maximum parsimony criterion), we estimated for each gene in each node of the tree whether it was more likely to have been inherited from a common ancestor or acquired at that node [35].

As mentioned above, 140 genes found in all *Synechococcus* are absent in all *Prochlorococcus* (Table S3). This is consistent with our earlier image, based on only four genomes, of progressive gene loss from *Synechococcus* to LL *Prochlorococcus* to HL *Prochlorococcus* [7,8,36]. However, our analysis suggests an alternative to this view, in that the MIT9313 lineage (i.e., the MIT9313/MIT9303 "cluster" or eMIT9313 clade, *sensu* [37]) is not simply an intermediate step in this gene loss process. Although the genome sizes within eMIT9313 are similar to those of *Synechococcus*, the eMIT9313 clade appears to have gained a large number of genes, including many unique to each isolate. These genes are not found in any other sequenced *Prochlorococcus* or *Synechococcus* strain, and the eMIT9313 strains may therefore have acquired them after their divergence from the other *Prochlorococcus*. The large difference between strains MIT9313 and MIT9303 is then most likely the result of further gene gains after they diverged from each other. After the divergence of eMIT9313, all *Prochlorococcus* genomes have a roughly constant size (1.66 to 1.84 Mbp). However, we still observe significant gene gain and loss. A few particular examples are discussed below, but additional work remains to show how these dynamics contribute to the distribution patterns we observe in the oceans for specific lineages.

Ecotypic differences: Genes underlying the HL/LL ecotypes. As described in many previous studies, *Prochlorococcus* can be classified into two broad groups based on their growth adaptation to specific light intensity (and corresponding phylogeny) [4]. In addition to the core genome shared by all 12 *Prochlorococcus* examined in this study, HL isolates all share an additional 257 genes, 95 of which are not found in any of the LL isolates (Table S6). This HL core provides further clues to the genetic bases for the HL/LL physiological and ecological differentiation that has been observed in previous works [4,5,19,20,37–42]. All HL isolates carry an operon containing a DNA ligase, exonuclease, and helicase, which might be involved in DNA repair or other nucleic acid processing. HL isolates also possess large numbers of HLIPs (although NATL1A and NATL2A have more), which are thought to protect photosystems from oxidative damage [39] and are upregulated in stress conditions such as high light [24], nitrogen starvation [43], and phage infection [44]. In particular, they share at least three additional genes for HL inducible proteins not found in any other strain. In addition to HL stress, one (hli8/18 in MED4) is upregulated in response to phage infection, and the other two (hli15 and hli22) by nitrogen starvation [43,44]. The HL isolates also share some genes with no clear connection to photobiology, such as a uridine kinase that may provide an alternative pathway for uracil recycling to UMP. In all *Prochlorococcus*, UMP can be generated by core pathways involving the core *upp* or *pyrBCDEF* genes [45]. All HL isolates also share the operon *tenA-thiD*, which may be involved in thiamine salvage and/or degradation [46,47]. In addition, the HL core contains dozens of hypothetical and conserved hypothetical genes not found in any LL isolate, and these might be critical for survival in the commonly nutrient-poor, HL environment of the surface oceans. Finally, all HL and eNATL2A isolates (which are LL, but closest to the HL clade) include at least one photolyase (orthologs of P9301_3091) and a second possible (P9301_03091), and some HL strains have a third (P9301_03921), the function of which is to repair UV-induced DNA lesions (Table 2).

Likewise, LL isolates share an additional 92 genes beyond the *Prochlorococcus* core, 48 of which are not found in any HL isolates (Table S7). All *Prochlorococcus* have lost the majority of genes involved in phycobilisome synthesis but LL isolates retain several phycoerythrin genes (*cpeABSTYZ*), whereas HL isolates have lost all but *cpeB* and *cpeS*, consistent with previous observations based on fewer genomes [48]. The role of phycoerythrin in *Prochlorococcus* remains uncertain, but may be related to signal transduction rather than light harvesting [49,50]. Individual *Prochlorococcus* strains possess different complements of amino acid transporters. But all LL isolates, and only some HL isolates, contain the tandemly arranged amino acid transporter components *glnQ* and *hisM*, suggesting some variation among *Prochlorococcus* ecotypes in the ability to take up amino acids [51].

Several exonucleases that repair UV-induced lesions, encoded by *recJ* and *xseA*, are exclusive to LL isolates, which is surprising given their reduced exposure to UV radiation. These genes might be necessary to protect against UV exposure during mixing events, and their absence from HL isolates suggests the HL isolates have different strategies to limit DNA damage. Moreover, LL isolates exclusively encode *mutY*, whose product prevents mutations arising from oxidatively damaged guanine residues [52]. The absence of the *mutY* gene in HL *Prochlorococcus* has been hypothesized to underlie their extremely low %G + C content, by increasing the frequency of G-C to A-T mutations [7]. However, this gene is present in LL isolates with %G + C as low as 35%, suggesting that *mutY* alone is not responsible for genomic A + T enrichment [53].

Ecotypic differences: Clades within the HL and LL

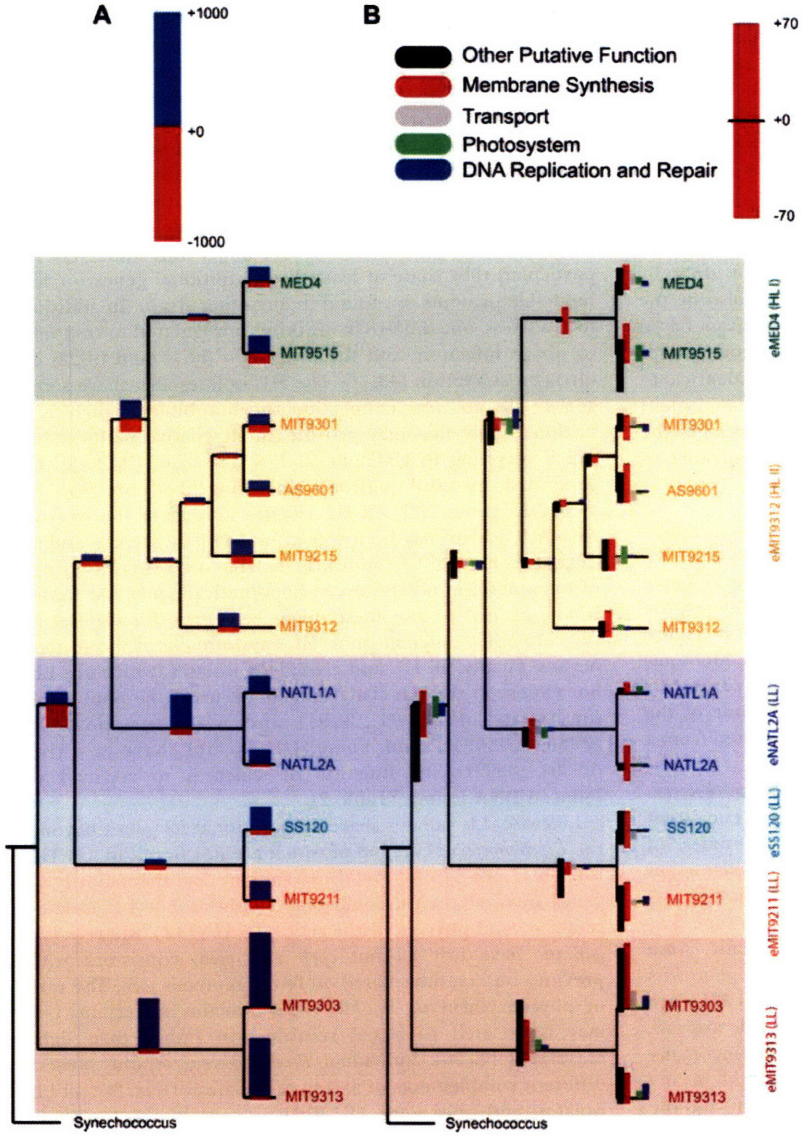


Figure 3. The Loss and Gain of Genes through the Evolution of *Prochlorococcus*

The ancestor node in which a gain or loss event took place was estimated by maximum parsimony. Four marine *Synechococcus* genomes (not shown) were included in the calculation, and the phylogenetic tree from Figure 2C was rooted between the *Synechococcus* and *Prochlorococcus* lineages.
(A) The total number of genes gained and lost at each node.
(B) The loss and gain of genes in that could be assigned functional roles through homology. Note that (B) focuses on the small minority of genes that do have an assigned function. Genes were assigned to one of five categories on the basis of keyword matches against the gene name or COG description. "Other Putative Function" refers to genes with assigned function but not belonging to the four major categories. Note the difference in scale for (A and B).

doi:10.1371/journal.pgen.0030231.g003

ecotypes. Going beyond the HL and LL ecotypes, two distinct subclades have been identified within the HL ecotype (eMED4 and eMIT9312), and several lineages within the LL ecotype (eNATL2A, eMIT9313, and eSS120 + eMIT9211) [18] (Figure 3). The distribution of cells belonging to these subclades has been measured along extensive environmental gradients in the oceans, and the two HL subclades have distinct distributions most strongly correlated with surface temperature [39,40]. Moreover, two LL clades (eNATL2A and eMIT9313) have distinct distributions as well: cells related to eNATL2A can be abundant at the surface, while cells related to eMIT9313 are generally found at the base of the euphotic zone in stratified waters and never at the surface [40]. This is in spite of the two clades' similar optimum light intensity for growth [19,40]. Given these ecological distinctions, we looked for genes distinguishing these subclades (Table 2).

The eMIT9313 clade has many features that distinguish it

Table 2. Non-core Genes Referred to in the Discussion

Location	Gene	Function	Locus	MIT9313	MIT9303	MIT9211	SS120	NATL1A	NATL2A	MIT9301	AS9601	MIT9215	MIT9312	MIT9515	MED4
Underlying the HL/LL ecotypes	*lhr*	Helicase	PMED4_08051							x	x	x	x	x	x
		DNA ligase	PMED4_08061							x	x	x	x	x	x
		Beta-lactamase fold exonuclease	PMED4_08071							x	x	x	x	x	x
	hli8/18	Photosystem protection	PMED4_15941							x	x	x	x	x	x
	hli15	Photosystem protection	PMED4_12761							x	x	x	x	x	x
	hli22	Photosystem protection	PMED4_07541							x	x	x	x	x	x
	udk	Uridine kinase	PMED4_11091					x	x	x	x	x	x	x	x
	tenA-thiD	Thiamine salvage	PMED4_03811–21							x	x	x	x	x	x
	phrB	Photolyase	PMED4_02901					x	x	x	x	x	x	x	x
	phrB	Photolyase	P9301_03921							x	x	x		x	
		Possible photolyase	P9301_04471					x	x	x	x	x	x	x	x
	cpeADR-TYZ	Phycorethryin	non-adjacent	x	x	x	x	x	x						
	cpeBS	Phycorethryin	non-adjacent	x	x	x	x	x	x	x	x	x	x	x	x
	hisM	Amino acid transport	P9313_10601	x	x	x	x	x	x				x	x	
	glnQ	Amino acid transport	P9313_10611	x	x	x	x	x	x	x			x	x	
	recJ	Exonuclease	P9313_08931	x	x	x	x	x	x						
	xseA	Exonuclease	P9313_20731	x	x	x	x	x	x						
	mutY	Mismatch repair	P9313_01441	x	x	x	x	x	x						
Within the HL/LL ecotypes		Sigma factor	P9313_13631	x	x										
		Sigma factor	P9313_27801	x	x										
		Sigma factor	A9601_12341								x				
		Sigma factor	P9313_09171	x	x										
	gdhA (1)	Amino acid synthesis	P9313_07431	x	x										
	gdhA (2)	Amino acid synthesis	P9515_04091									x		x	
	cytA	Electron transporter	P9313_06071	x	x										
	cypX	Electron transporter	P9313_19741	x	x										
	melB	Disaccharide transport	P9211_03411			x	x								
	glcD	Dehydrogenase	P9211_13031			x	x								
	citB-baeS	Signal Transduction	P9211_15001–11			x	x								
		Disulfide bond formation	P9211_15411			x	x								
	nirA	Nitrite reductase	P9313_28061	x	x			x	x						
	sdhA	Possibly TCA cycle	A9601_12591							x	x	x	x		
	sdhAB	Possibly TCA cycle	P9313_01411–21	x	x		x								
	phoBR		PMED4_07791–801	x[a]	x			x	x	x			x		x
	phoE		PMED4_07831	x	x			x	x	x			x		x
	cynS	Cyanate lyase	PMED4_04061					x	x						x
	amtB (1)	Ammonia permease	PMED4_02681	x	x	x	x	x	x	x	x	x	x	x	x
	amtB (2)	Ammonia permease	P9515_04231											x	
	urtBCD	Urea transport	PMED4_10831–51	x	x			x	x	x	x	x	x		x

Each line is an orthologous group, for which the gene name and putative function are given, if available. The locus given is that of an arbitrarily selected gene in the group; the complete list for any orthologous group is available in Table S3. The presence or absence in each *Prochlorococcus* isolate is given.

[a]*phoR* is not functional in MIT9313.

doi:10.1371/journal.pgen.0030231.t002

from other *Prochlorococcus* (Table S8). Acquired genes include multiple sigma factors and kinases, likely involved in signal transduction, outer membrane synthesis enzymes, and transporters. Their possession of transporters not found in other *Prochlorococcus* or in *Synechococcus* may imply that they are exploiting nutrient resources unique to their environment, or they may simply have experienced weaker selection for reducing genome size. Likewise, the two isolates in this clade (MIT9313 and MIT9303) share three sigma factors (MIT9303 has a fourth) and several other transcriptional regulators not found in any other isolate, suggesting they have more complexity in their ability to respond to various stimuli. The eMIT9313 isolates also share a glutamate dehydrogenase gene (*gdhA*), absent from most other *Prochlorococcus* (two HL isolates share a distantly related allele), which provides an alternative pathway for ammonium incorporation besides the standard GS-GOGAT pathway. This enzyme has been shown in *Synechocystis* to be important during the late stages of growth when energy is limiting, and for ammonia detoxification [54]. We also observe that photosystem II genes *psbU* and *psbV* are exclusively found in eMIT9313 (as well as most other cyanobacteria) along with possible electron transporters (*cytA*, *cypX*). The eMIT9313 isolates carry only three *pcb* genes, encoding light harvesting antenna proteins, compared to six or seven in the other LL isolates. This relative lack of *pcb* genes, however, does not seem to prevent growth at very low irradiances, as eMIT9313 cells are often found at the base of the euphotic zone. The eMIT9313 isolates also have relatively few genes for HLIPs (nine in eMIT9313, compared to 12–13 in SS120/MIT9211 and 41 in eNATL2A), which might help explain why this clade is not found in surface waters.

Five genes with assigned functions were unique to eSS120/eMIT9211 (P9211_03411, P9211_13031, P9211_15001, P9211_15011, P9211_15411), but there were no clear linkages between these genes and the distribution pattern of this group in the ocean.

In contrast, the eNATL2A isolates (NATL1A and NATL2A), whose low optimum light intensity for growth marks them as LL [19,40] have some notable HL-like properties. The eNATL2A isolates possess photolyase genes, like HL isolates, and they harbor more genes for HLIPs than any other HL or LL isolate. Together these genes may help explain the abundance of eNATL2A at the surface relative to other LL clades [40]. They also share the uridine kinase found in HL isolates.

All isolates in the eMIT9313 and eNATL2A clades possess a nitrite reductase gene, *nirA*, whereas no other *Prochlorococcus* lineages (HL or LL) have this gene, a difference that has been confirmed through physiology studies [10]. The availability of nitrite may therefore influence the distribution of these two clades, although this pattern has not emerged in the field studies to date [39,41].

In spite of their different distributions in the ocean, we could identify only one gene with a described function that distinguishes the two HL clades eMIT9312 and eMED4. All isolates in eMIT9312 possess a gene similar to *sdhA* which encodes succinate dehydrogenase. Unlike the proteobacteria-like *sdhA* found in SS120, MIT9313, and MIT9303 and previously assigned to the incomplete TCA cycle [8], the HL gene is actinobacteria-like and is not accompanied by *sdhB*, raising the possibility that this dehydrogenase/reductase acts on a different substrate. Temperature variability is most strongly correlated with differences in the abundances of eMED4 and eMIT9312 along a longitudinal gradient in the oceans, and this is consistent with the temperature limits for growth for strains representing these ecotypes in culture [39]. These properties could emerge from differences within orthologous proteins, yielding different enzymatic reaction temperature optima, rather than from the presence or absence of entire genes. This complicates the search for ecotype-defining genes in their case.

Isolate-specific genes. We found that a large fraction of variability was in the "leaves of the tree," that is, genes gained by one isolate but not necessarily by others in the same clade (Figure 3B and Table S4). The greatest differentiator between the most closely related isolates are genes related to outer membrane synthesis (Table S5). For example, while MIT9515 and MED4 each have several genes in COG438 and COG451 (both COGs described as acyltransferases connected to outer membrane synthesis), these genes are only distantly related [55]. Six genes matching COG438 are found in MIT9515 but not MED4, and these six all have best matches to genes in lineages outside *Prochlorococcus*. The rapid turnover of genomic content contrasts with the broader similarity of their roles: even though the genes found in different isolates are not orthologs and have little to no sequence similarity, they share the same biological role. Such membrane synthesis genes were probably lost or gained continuously throughout the evolution of *Prochlorococcus*, as every ancestor node is estimated to have lost or gained some in that category (Figure 3B).

Certain cell surface proteins are potentially under strongly diversifying selection if they serve as attachment or recognition sites for predators or phages. The observed variation among genomes in relation to this category supports this idea and suggests that the predatory environment could be different in each of the locations where these isolates originated. However, it is deceptive to consider these the most recent changes, as there are innumerable undiscovered *Prochlorococcus* genotypes in the wild, some of which could fill the gap between MIT9515 and MED4, for example. Such variation, some of which may be adaptive, is below the resolution of current methods for measuring ecotype abundance in the oceans [39,42,56].

After cell surface synthesis, the next largest fraction of the flexible genome is transporters (Figure 3B). As discussed above, the larger genomes of MIT9303 and MIT9313 have a significant number of transporters not shared with other *Prochlorococcus*, although some are shared with *Synechococcus*. Among their predicted substrates are toxins, sugars, and metal ions. Relatively few transporters are specific to the other LL isolates. In addition, each HL isolate possesses a different set of transporters, but there is no set both universal among HL isolates and absent from LL isolates. Furthermore, the presence of specific transporters does not follow the phylogeny of the HL ecotype. Transport genes must therefore be subject to rapid gain and loss, such that their presence is not conserved within the subclades. Transport reactions are peripheral to metabolic pathways, and such peripheral reactions are predicted to be subject to the most rapid turnover [57].

Individual *Prochlorococcus* isolates also contain multiple copies of specific light-related genes but in different

numbers. MED4, the first HL genome to be studied, has only one *pcb* light harvesting antenna gene whereas the first LL genomes had two (MIT9313) or eight (SS120) [58]. Our new data identify MED4 as the exception, since the other five HL isolates share a second copy in the same well-conserved neighborhood. Surprisingly, there is huge variation in the number of genes encoding HLIPs, ranging from nine in eMIT9313 to 41 in eNATL2A. Even at the leaves of the tree, within the HL clades, HLIPs range in copy number from 15 to 24.

A second copy of the core photosystem II gene *psbA* also appears in more than half the genomes. This gene is especially interesting because it is also found in all *Prochlorococcus*-infecting myoviruses and podoviruses sequenced to date [59]. While it is possible that *psbA* might have been inserted into the genome by those viruses, much as the genes in genomic islands are thought to have been [60], the similarity between *psbA* copies in the same genome suggests they are the result of intragenomic duplication events, not transduction. Indeed, in all of these strains the two copies are identical or nearly identical in nucleotide sequence, suggesting that they result from a very recent duplication event. Furthermore, while extra *psbA* copies sometimes appear in islands, they do not always. In MIT9515, for example, the two copies lie in tandem but not in an island. It is not clear why *psbA* is subject to such duplication events while other photosystem genes are not. The most likely reason is that the PsbA protein (D1) has an exceptionally brief half-life due to light-induced damage [61], and therefore two gene copies help ensure sufficient product via a gene dosage effect and/or by promoter differences leading to expression under different conditions.

The complement of nutrient assimilation genes also varies among the most closely related isolates, suggesting frequent gain and loss events. Such variability was recently described for genes involved in phosphorus assimilation [11]. Within the eMIT9312 clade, for instance, the isolates AS9601 and MIT9215 are lacking the *phoBR* two-component system, the *phoE* porin, and several related genes that are present in MIT9312 and MIT9301. Now equipped with whole genomes for 12 isolates, we see a similar situation for nitrogen assimilation genes. MED4 is the only HL isolate with cyanate lyase, and likewise MIT9515 exclusively carries a second ammonia permease gene. In contrast, MIT9515 is the only HL isolate lacking urea transport and metabolism genes. This variability may reflect the available nitrogen sources in the local environment where these isolates originated, as has been hypothesized for phosphorus [11].

Chromosomal Location of the Flexible Genome

Previous work comparing the genomes of two closely related *Prochlorococcus* isolates has highlighted the importance of highly variable island regions in genomes as the sites of genomic variation [60]. These variable genome segments appear to contain genes that could be important for adaptation to local conditions, and include many of the functions encoded in the flexible genome analyzed here, such as outer membrane synthesis. Thus, we analyzed the chromosomal geography of the flexible genome. Are flexible genes preferentially located in island regions, and if so are the most recently acquired genes more likely to be island genes?

To answer these questions, we plotted the timing of gene gain events against their chromosome positions (Figures 4 and S5 and S6). In HL isolates, the islands contain the majority of gained genes. Furthermore, the islands include not only recent acquisitions but also genes that were gained long ago, based on their presence in divergent modern isolates. However, particular islands show different levels of gain or loss events throughout the evolution of *Prochlorococcus*. Apparently, these sites have been important for adaptation throughout the history of most *Prochlorococcus* lineages.

In the earlier comparison of two genomes at a time, islands were identified as breaks in syntenic regions [60]. Among LL isolates, this approach is difficult because the genomes are more divergent, and numerous rearrangements have disrupted synteny, even for core genes. Plotting gene gain events along the chromosome, however, reveals island structure in several LL genomes. MIT9211 and SS120 have clearly defined islands much like the HL isolates, while NATL1A and NATL2A have one large potential island and several much smaller sites (Figures 4 and S5 and S6).

Surprisingly, this approach is less helpful in the two large genomes, MIT9313 and MIT9303, which have apparently gained a large number of genes throughout the chromosome (Figure 4). In their organization and content, the large genomes are exceptional among *Prochlorococcus* in three ways: they share a large number of genes with *Synechococcus* that the other isolates do not, they gain additional genes not shared with any *Prochlorococcus* or with marine *Synechococcus*, and those genes do not cluster in discernible islands. The first two differences mean that their genome sizes are much greater than those of the other isolates. The lack of islands together with the larger genome size could indicate that these isolates have acquired genes through a different mechanism that does not direct them toward islands. The relative lack of pressure towards genome reduction in the evolution of eMIT9313 may also play a role. However, additional sequenced genomes may provide better coverage of the eMIT9313 clade and clarify the timing of gene gain events.

The Frequency of Core and Flexible Genes in Wild Populations

Because *Prochlorococcus* is very abundant in many regions of the oceans that have recently been sampled and subjected to metagenomic analysis [62–64], we have an opportunity to test the robustness of our distinction between core and flexible genes in *Prochlorococcus*. If the core genome we have defined, based on the genomes of 12 isolates, is reasonably universal and core genes are generally single copy per genome, we would expect to find core genes represented with equal frequency in the ocean; the occurrence of non-core genes, in contrast, would be more variable. To test this hypothesis, we used the MIT9301 core and flexible genomes as queries against the Global Ocean Survey dataset [64], as MIT9301 often shares the highest sequence similarity with GOS sequences. As expected, the core genes, after normalization to gene size, are represented in roughly equal abundance in the database, with only a few exceptions (Figure 5A). In the case of non-core, or flexible genes, many had few or no hits, and a few were even more abundant than the average core gene, suggesting more than one copy per genome (Figure 5A). Seven core genes are underrepresented in the GOS dataset relative to other core genes, and all seven are located in a genomic island in MIT9301 largely related to cell surface

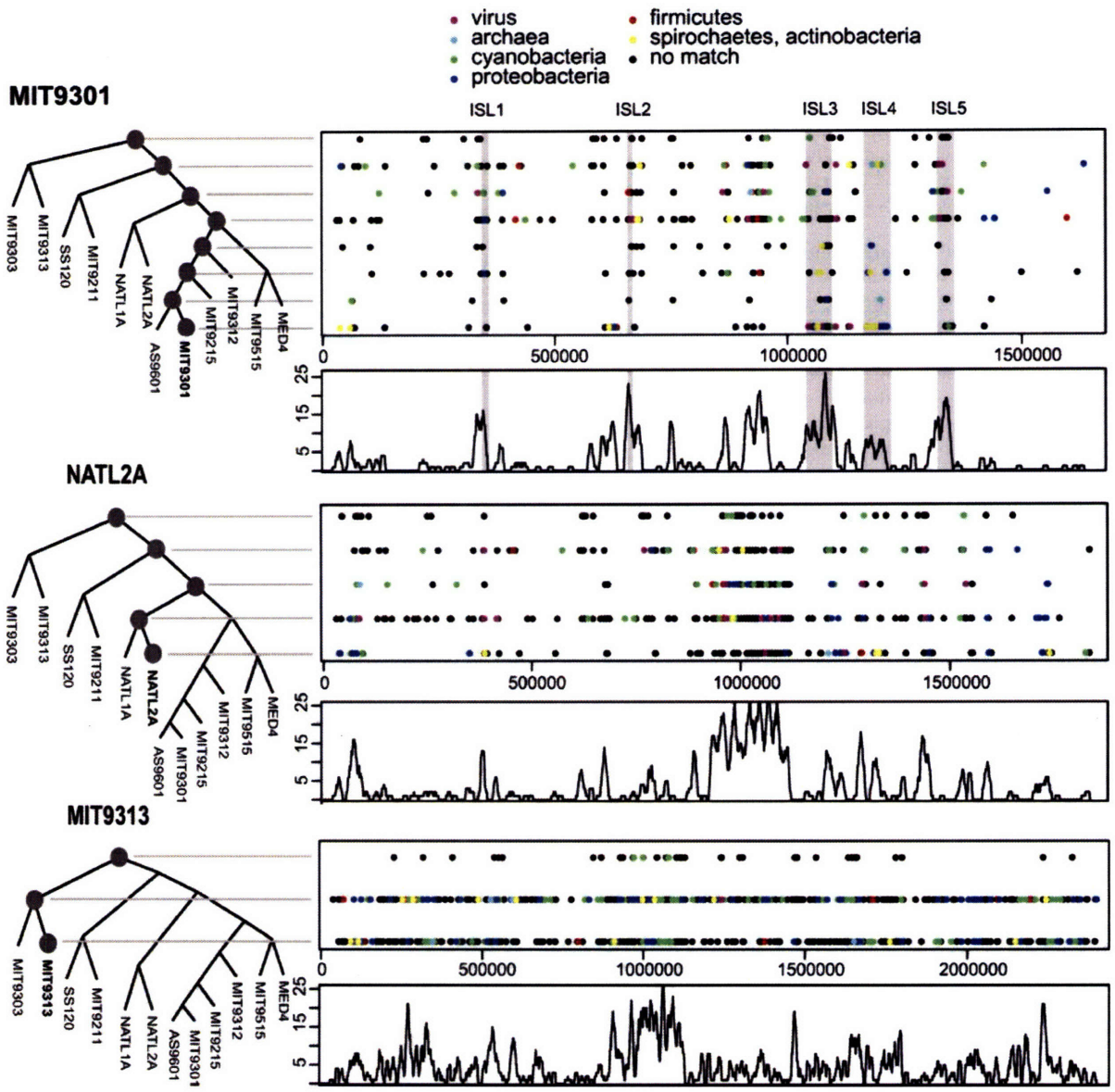


Figure 4. Gene Acquisitions Confirm Known, and Identify Novel, Genomic Islands in *Prochlorococcus*

The dot plots indicate the location on the chromosome and the ancestor node in which the gene is estimated to be gained. The color indicates where the best match was found. In MIT9301, The shaded regions are islands as defined by [60]. Gained genes are defined for each node as in Figure 3. The lower plot is the number of genes gained in a sliding window (size 10,000 bp, interval 1,000 bp) along the chromosome.

doi:10.1371/journal.pgen.0030231.g004

biosynthesis (Figure 5B). The most abundant flexible genes encode HLIPs and hypothetical proteins and are also found in islands in MIT9301 (Figure 5B). This supports the hypothesis that islands are dynamic reservoirs for recent and local adaptation.

Conclusion

In this study we have attempted to advance our understanding of the evolutionary origins of diversity in *Prochlorococcus* by defining the core and flexible genomes and examining the patterns of gain and loss of non-core genes over the course of evolution. We have learned, for example, that many genes involved in adaptation to different light intensities and DNA repair were apparently fixed before the modern clades diverged, and as a result, the HL-/LL-adapted dichotomy has persisted both genetically and phenotypically. The eNATL2A clade appears to be a refinement on the HL/LL paradigm, as its isolates grow optimally at light intensities typical of the LL ecotype, but have the photoprotective abilities of the HL ecotype. More recent changes in genome content, i.e., those occurring at the tips of the phylogenetic tree, involve cell surface features that are likely under

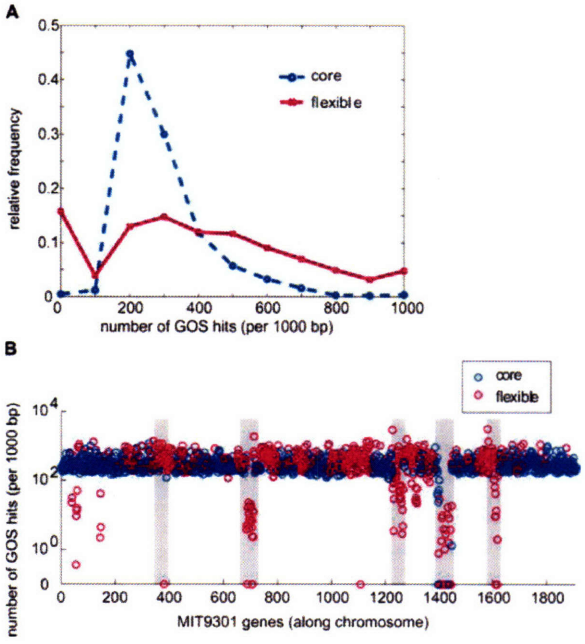


Figure 5. *Prochlorococcus* Core and Flexible Genes in the Global Ocean Survey (GOS) Dataset [64]

(A) Frequency distribution of GOS hits per gene, using genes in the *Prochlorococcus* MIT9301 genome as queries. Most core genes retrieve a similar number of GOS hits, as one would expect from single copy genes shared by all *Prochlorococcus*, resulting in a relatively tight frequency distribution. In contrast, flexible genes retrieve a broad range of GOS hits per gene, consistent with their scattered distribution among genomes. (B) The number of GOS hits per gene, again using MIT9301 genes as queries, plotted against position along the chromosome. Shaded regions represent genomic islands, after [60]. Flexible genes with low representation in the GOS dataset tend to be located in genomic islands. In both (A) and (B), the number of GOS hits per gene is normalized to gene length and plotted as hits per gene, per 1,000 bp.

doi:10.1371/journal.pgen.0030231.g005

selection pressure via predators and phage and transporter composition, which likely plays a role in both defense from toxins and differences in nutrient availability. The latter is consistent with our earlier observation that genes involved in phosphorus acquisition are distributed among *Prochlorococcus* isolates not according to phylogeny, but rather the P concentrations in their ocean of origin [11]. However, despite the clear evidence for common gene gains and losses throughout the evolution of *Prochlorococcus*, we still observed a significant correlation between genome content and phylogeny. This suggests an important contribution of vertically inherited genes to the overall genome content that cannot be easily substituted through lateral gene transfer or lost altogether.

The core genome of *Prochlorococcus*, with 81% of the 1,273 genes having an inferred function, is now reasonably well understood and appears to encode a viable cell. That this could be circumscribed through the analysis of only 12 genomes is encouraging, and likely emerges from the reasonably small evolutionary distance between these isolates. The close agreement between manually curated core pathway reconstruction for one isolate [8], and the automatic reconstruction of the core metabolism shared by all 12 isolates in our study, promises to help streamline the analysis of new genomes. To date, discussions of minimal genomes to support life have focused on the set of genes that enable heterotrophic cells to replicate on rich organic media, where they benefit from nutrients that must have been synthesized by other organisms [65]. Here, however, we are approximating the minimum number of genes necessary to convert solar energy, carbon dioxide, and inorganic nutrients to living biomass.

The *Prochlorococcus* flexible genome is still only loosely defined, as over 70% of the orthologous groups in this category have no known homolog in MicrobesOnline and no inferred function. Moreover, as the last genomes are added to the analysis, they each add roughly 150 new genes to the *Prochlorococcus* pan-genome (Figure 1); thus it appears that the global pool of genes that are residing, at this moment, in a *Prochlorococcus* cell cannot even be approximated from this dataset. Therefore, one of the most daunting unanswered questions is: How many *Prochlorococcus* genotypes truly exist in the ocean, and what fraction of these has differential fitness at any point in time?

The level of diversity found in the flexible genes, and the steady increment of genes added to the *Prochlorococcus* pan genome with each new genome, suggests that we have barely begun to observe the extent of micro-diversity among *Prochlorococcus* in the ocean. Although the sequencing of 12 genomes represents one of the larger sequencing projects of closely related isolates to date, each isolate undoubtedly represents a subclade of a very large number of cells—especially considering the approximately 10^{25} *Prochlorococcus* cells in the ocean [3]. Additional sequencing, especially metagenomic [63] and single-cell sequencing [66], will help us understand more about on what scale, and where in the genomes, the flexible genes vary. In particular, it will be enlightening to understand the complete genome diversity of the 10^5 cells in a milliliter of ocean water, and conversely, how widely separated in space two cells with identical genomes might be.

Materials and Methods

DNA sequencing and assembly. The genome sequences of eight of the isolates used in our analysis are reported for the first time here. The genomes of MIT9211, MIT9515, NATL1A, MIT9303, MIT9301, and AS9601 were sequenced by the J. Craig Venter Institute as follows: Two genomic libraries with insert sizes of 4 and 40 kb were made as described in [67]. The prepared plasmid and fosmid clones were sequenced from both ends to provide paired-end reads at the J. Craig Venter Institute Joint Technology Center on ABI3730XL DNA sequencers (Applied Biosystems). Successful reads for each organism were used as input for the Celera Assembler. WGS sequence produced by the assembler was then annotated using the PGAAP at NCBI. Accession numbers for all genomes are provided in Table 1.

NATL2A was sequenced at the DOE Joint Genome Institute by methods described previously (http://www.jgi.doe.gov/sequencing/protocols/prots_production.html). Briefly, three whole genome shotgun libraries were constructed containing inserts of approximately 3 kb, 8 kb, or 40 kb and sequenced to a depth of 9X using BigDye Terminators on ABI3730 sequencersb (Applied Biosystems). Shotgun reads were assembled with parallel PHRAP (http://www.phrap.org).

The MIT9215 genome was sequenced with a combination of approximately 20X coverage of 454 pysoquencing (454 Life Sciences) and standard Sanger sequencing of 3-kb insert libraries. All genomes were completed to finished quality with no gaps, except MIT9211, with one gap of less than 1 kb and an estimated error rate of less than 1 in 50,000 bases.

Genome annotation. We re-annotated 12 sequenced *Prochlorococcus*

and four finished marine *Synechococcus* genomes by a uniform method for the purpose of this study. We used the gene prediction programs CRITICA [68] and GLIMMER [69]. The results from both programs were combined into a preliminary set of unique ORFs. Overlapping gene models from the two programs are considered the same gene if sharing the same stop position and in the same reading frame, in which case the gene start site of the CRITICA model is preferred. Coding genes that are shorter than 50 aa long are excluded unless they are conserved in more than one genome. Orthologous genes between two given genomes are assigned automatically using MicrobesOnline's [70] (http://www.microbesonline.org) genome annotation pipeline. The new annotations are also available at that site.

Two genes are considered orthologs if they are reciprocal best BLASTp hits and the alignment covers at least 75% of the length of each gene. An orthologous group includes all genes that are orthologous to any other gene in the group. The most common challenge of clustering orthologous genes is the risk of merging paralogous genes into one group. However, our method yields only 127 paralog-containing groups. In those cases, gene neighborhoods were also compared. Because a single missing ortholog effectively removes a gene from the core genome, the clusters that are absent in only one or two genomes were verified by BLAST search.

While the COG categories alone provide enough information to draw these conclusions about the membrane synthesis enzymes, there are some shortcomings. Some *Prochlorococcus* orthologous groups can be annotated with a gene name but not a COG (for example the LPS synthesis gene wcaK, or many photosystem genes like psbA), where literature searches show that they are likely involved in LPS synthesis. Other categories are hampered by the arrangement of the COG categories, which were not chosen with any particular focus on this system. For example, the category "Amino acid transport and metabolism" includes transporters and intracellular enzymes. When we found that transporters are among the most recently gained genes, we desired a way to group all of them by themselves. We decided the best approach was to group genes into five broad categories on the basis of keyword searches: membrane or cell wall synthesis, transporters, photosynthesis, DNA repair or modification, and other. HLI proteins were identified by their possession of six out of ten conserved residues in the motif AExxNGRxAMIGF, and lengths under 120 amino acids [32].

Phylogenetic analysis. 16S rRNA and 16S-23S rRNA ITS region sequences were manually aligned in ARB and phylogenetic reconstruction using maximum parsimony, neighbor-joining, and maximum likelihood was done in PAUP [71]. Following the approach described in [33] to identify the phylogenetic relationship between the sequenced isolates, we aligned all core genes using clustalw using the protein sequence as reference. We randomly concatenated 100 alignments and constructed a phylogenetic tree using maximum parsimony and bootstrap resampled 100 times. The random concatenation was repeated 100 times and the average bootstrap values for concatenated alignments are reported in Figure 2. In addition, we also constructed a phylogenetic tree using maximum parsimony on each individual alignment and the most likely tree for each gene (plurality consensus tree based on 100 bootstraps) was identified. We also calculated the phylogenetic relationship based on the presence and absence of orthologous groups as previously described [34]. However, we used bootstrap instead of jack-knife resampling to test how well individual nodes were supported to ensure easy comparison with other phylogenetic trees.

Estimation of the timing of gene loss and gain events was as described using a maximum parsimony approach [35]. We used the phylogenetic tree in Figure 2C rooted between the *Prochlorococcus* and *Synechococcus* last common ancestors as guide. We included the cost of a "gain" event in the tree's common ancestor node. We assigned a gene gain event twice the cost of a loss event, and in cases where two scenarios had equal scores we chose the one with fewer gains. We also tested a ratio of three to one, which changes the behavior of 117 genes.

Metabolic reconstruction. To predict the metabolic pathways present in the sequenced isolates, we ran Pathway Tools software [30] to generate a Pathway/Genome database (PGDB). This software creates gene, protein, reaction, small-molecule, and pathway objects based on Enzyme Commission (E.C.) numbers and enzyme names assigned in the genome annotation. We hand-curated the PGDB to eliminate unlikely pathways, and from it we created a pathway model of the central carbon metabolism [72]. To aid in the analysis of the core and flexible genes, we created a pseudogenome, Pan, which includes all genes from all isolates. We created another pseudogenome for the core genome. The database is available in flat file, BioPAX, and SBML format.

Supporting Information

Figure S1. The Core Genome Includes Enzymes for Central Carbon Metabolism, Including the Calvin Cycle, Glycolysis, and an Incomplete TCA Cycle Producing Fumarate and 2-Oxoglutarate

Some genomes, but not the core genome, also include *sdhAB*, encoding an enzyme for the reaction 1.3.99.1, the conversion of fumarate to succinate (Table 2). The pathway diagram includes the stuctures of intermediate metabolites, the locus name, in MED4, of the gene encoding each enzyme, the enzyme name, and the E.C. number.

Found at doi:10.1371/journal.pgen.0030231.sg001 (1.4 MB EPS).

Figure S2. The Core Genome Includes Enzymes for the Synthesis of All 20 Amino Acids

The pathway diagram is annotated as in Figure S1.

Found at doi:10.1371/journal.pgen.0030231.sg002 (2.2 MB EPS).

Figure S3. The Core Genome Includes Enzymes for the Synthesis of Divinyl Chlorophyll

The pathway diagram is annotated as in Figure S1.

Found at doi:10.1371/journal.pgen.0030231.sg003 (1.5 MB EPS).

Figure S4. The Core Genome Includes Enzymes for the Synthesis of the Cofactors NAD (A), Coenzyme A (B and C), and FAD (D)

The pathway diagrams are annotated as in Figure S1. One reaction (2.7.1.33) in coenzyme A synthesis is highlighted; its enzyme (pantothenate kinase) has not been identified in the core or pangenomes.

Found at doi:10.1371/journal.pgen.0030231.sg004 (1.3 MB EPS).

Figure S5. Islands of LL Genomes Not Represented in Figure 4

The dot plot shows the location of each gene, the ancestor in which it is estimated to be acquired, and when possible, the best match outside *Prochlorococcus*. The lower plot is the number of genes gained in a sliding window (size 10,000 bp, interval 1,000 bp) along the chromosome.

Found at doi:10.1371/journal.pgen.0030231.sg005 (4.2 MB EPS).

Figure S6. Islands of HL Genomes Not Represented in Figure 4

The dot plot shows the location of each gene, the ancestor in which it is estimated to be acquired, and when possible, the best match outside *Prochlorococcus*. The lower plot is the number of genes gained in a sliding window (size 10,000 bp, interval 1,000 bp) along the chromosome. When available, the locations of islands previously defined by hand are represented by shaded regions.

Found at doi:10.1371/journal.pgen.0030231.sg006 (6.1 MB EPS).

Table S1. All *Prochlorococcus* Orthologous Groups in This Study

For each group, its locus names are given for those genomes in which it is found. Also given are the COG match [55], gene name, and description as assigned by MicrobesOnline (http://www.microbesonline.org).

Found at doi:10.1371/journal.pgen.0030231.st001 (1.8 MB XLS).

Table S2. Prochlorococcus Core Genes Absent in Synechococcus

33 orthologous groups are shared by all *Prochlorococcus* but absent in some *Synechococcus*, and only 13 of those are absent in all *Synechococcus*. For each such orthologous group, its presence or absence in each of the four *Synechococcus* genomes in this analysis is given. Also given is the locus name for the gene in MED4, its COG match, and its gene name, if available.

Found at doi:10.1371/journal.pgen.0030231.st002 (68 KB DOC).

Table S3. Genes Found in All Synechococcus but No Prochlorococcus

The locus name for *Synechococcus* is given, in addition to the COG and gene name, if available.

Found at doi:10.1371/journal.pgen.0030231.st003 (45 KB XLS).

Table S4. Genes Lost or Gained at Each Ancestor

For each gene, the name and COG are given, in addition to a locus name. The role assigned is one of "nomatch," "shortnomatch," "conserved_unknown," "hli," "photosynthesis," "DNA," "membrane," "transport," or "other," on the basis of keyword matches in the gene name, COG, or description. The latter five categories are

reported individually in Figure 3B; the totals are reported in Figure 3A.

Found at doi:10.1371/journal.pgen.0030231.st004 (1.5 MB XLS).

Table S5. The Most Common COGs in the Core and Flexible Genomes

We used matches against the COG database as a first impression of the differences between the core and flexible genomes. The number of *Prochlorococcus* orthologous groups and the total number of genes in those groups, matching each COG is given. The top ten COGs matching the core and flexible genomes are shown.

Found at doi:10.1371/journal.pgen.0030231.st005 (43 KB DOC).

Table S6. Orthologous Groups Found in All HL Isolates

These include those exclusive to HL isolates and those shared with some, but not all, LL isolates, as indicated. Also given are the gene name, description, and COG assignments as in Table S1.

Found at doi:10.1371/journal.pgen.0030231.st006 (114 KB XLS).

Table S7. Orthologous Groups Found in All LL Isolates

As Table S6, but those found in all LL isolates.

Found at doi:10.1371/journal.pgen.0030231.st007 (60 KB XLS).

Table S8. Notable Genes Exclusive to eMIT9313 Isolates

These are orthologous groups from Table S1, each found only in MIT9303, MIT9313, and in some cases marine *Synechococcus*. This list includes only those genes with hypothetical functions and with no BLAST alignment against the other genomes. Note that some belong to COGs shared with other *Prochlorococcus* isolates, but their extreme sequence divergence suggests their precise roles differ.

Found at doi:10.1371/journal.pgen.0030231.st008 (118 KB XLS).

Acknowledgments

Sequencing, assembly, and annotation efforts of MIT9211, MIT9515, NATL1A, MIT9303, MIT9301, and AS9601 were supported by Marine Microbial Sequencing Project of the Gordon and Betty Moore Foundation, and carried out at the J. Craig Venter Institute (JCVI) Joint Technology Center, under the leadership of Robert Friedman and Yu-Hui Rogers. We thank Granger Sutton, Aaron Halpern, and Saul A. Kravitz for their contributions to completion of these genomes. Sequencing, assembly, and annotation of the genomes NATL2A and MIT9215 were produced by the DOE Joint Genome Institute JGI. We thank Brian Palenik for allowing us to use unpublished *Synechococcus* genomes as out-group reference genomes in our analyses

Author contributions. ACM, GMC, and SWC conceived and designed the experiments. ACM, SR, FC, AL, SF, JJ, CS, and PR performed the experiments. GCK, ACM, KH, JZ, and MLC analyzed the data. GCK, ACM, MLC, and SWC wrote the paper.

Funding. This work was supported in part by grants from the National Science Foundation (NSF), Department of Energy (DOE), and the Gordon and Betty Moore Foundation (SWC), and a DOE-GTL Grant (SWC and GC). It is C-MORE Contribution #43. GCK was supported in part by a National Institutes of Health Training Grant through the MIT Biology Department, MLC by an NSF Graduate Fellowship, and ACM by a fellowship from the Danish National Science Foundation.

Competing interests. The authors have declared that no competing interests exist.

References

1. Goericke RE, Welschmeyer NA (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: 2283–2294.
2. Waterbury JB, Watson SW, Valois FW, Franks DG (1986) Biological and ecological characterization of the marine unicellular bacterium *Synechococcus*. Can Bull Fish Aquat Sci 214: 71–120.
3. Partensky F, Hess WR, Vaulot D (1999) *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. Microbiol Mol Biol Rev 63: 106–127.
4. Moore LR, Rocap G, Chisholm SW (1998) Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. Nature 393: 464–467.
5. West NJ, Scanlan DJ (1999) Niche-partitioning of *Prochlorococcus* populations in a stratified water column in the eastern North Atlantic Ocean. Appl Environ Microbiol 65: 2585–2591.
6. Urbach E, Scanlan DJ, Distel DL, Waterbury JB, Chisholm SW (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.
7. Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, et al. (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424: 1042–1047.
8. Dufresne A, Salanoubat M, Partensky F, Artiguenave F, Axmann IM, et al. (2003) Genome sequence of the cyanobacterium *Prochlorococcus* marinus SS120, a nearly minimal oxyphototrophic genome. Proc Natl Acad Sci U S A 100: 10020–10025.
9. Palenik B, Ren Q, Dupont CL, Myers GS, Heidelberg JF, et al. (2006) Genome sequence of *Synechococcus* CC9311: insights into adaptation to a coastal environment. Proc Natl Acad Sci U S A 103: 13555–13559.
10. Moore LR, Goericke RE, Chisholm SW (2002) Utilization of different nitrogen sources by the marine cyanobacteria, *Prochlorococcus* and *Synechococcus*. Limnol Oceanogr 47: 989–996.
11. Martiny AC, Coleman ML, Chisholm SW (2006) Phosphate acquisition genes in *Prochlorococcus* ecotypes: evidence for genome-wide adaptation. Proc Natl Acad Sci U S A 103: 12552–12557.
12. Bapteste E, Boucher Y, Leigh J, Doolittle WF (2004) Phylogenetic reconstruction and lateral gene transfer. Trends Microbiol 12: 406–411.
13. Lerat E, Daubin V, Moran NA (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-proteobacteria. PLoS Biol 1: e19. doi:10.1371/journal.pbio.0000019
14. Doolittle WF (1999) Phylogenetic classification and the universal tree. Science 284: 2124–2129.
15. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, 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: 13950–13955.
16. Makarova K, Slesarev A, Wolf Y, Sorokin A, Mirkin B, et al. (2006) Comparative genomics of the lactic acid bacteria. Proc Natl Acad Sci U S A 103: 15611–15616.
17. Mulkidjanian AY, Koonin EV, Makarova KS, Mekhedov SL, Sorokin A, et al. (2006) The cyanobacterial genome core and the origin of photosynthesis. Proc Natl Acad Sci U S A 103: 13126–13131.
18. 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.
19. Moore LR, Chisholm SW (1999) Photophysiology of the marine Cyanobacterium *Prochlorococcus*: ecotypic differences among cultured isolates. Limnol Oceanogr 44: 628–638.
20. Partensky F, Hoepffner N, Li W, Ulloa O, Vaulot D (1993) Photoacclimation of *Prochlorococcus* sp. (Prochlorophyta) strains Isolated from the North Atlantic and the Mediterranean Sea. Plant Physiol 101: 285–296.
21. Shalapyonok A, Olson RJ, Shalapyonok LS (1998) Ultradian Growth in *Prochlorococcus* spp. Appl Environ Microbiol 64: 1066–1069.
22. Scanlan DJ, Hess WR, Partensky F, Vaulot D (1996) High degree of genetic variation in *Prochlorococcus* (Prochlorophyta) revealed by RFLP analysis. European Journal of Phycology 31: 1–9.
23. Lawrence JG, Hendrickson H (2005) Genome evolution in bacteria: order beneath chaos. Curr Opin Microbiol 8: 572–578.
24. Steglich C, Futschik M, Rector T, Steen R, Chisholm SW (2006) Genome-wide analysis of light sensing in *Prochlorococcus*. J Bacteriol 188: 7796–7806.
25. Nagata N, Tanaka R, Satoh S, Tanaka A (2005) Identification of a vinyl reductase gene for chlorophyll synthesis in *Arabidopsis thaliana* and implications for the evolution of Prochlorococcus species. Plant Cell 17: 233–240.
26. Chisholm SW, Frankel SL, Goericke RE, Olson RJ, Palenik B, et al. (1992) *Prochlorococcus marinus* nov. gen. nov. sp.: an oxyphototrophic marine prokaryote containing divinyl chlorophyll *a* and *b*. Archives of Microbiology 157: 297–300.
27. Rubio LM, Flores E, Herrero A (1999) Molybdopterin guanine dinucleotide cofactor in *Synechococcus* sp. nitrate reductase: identification of *mobA* and isolation of a putative *moeB* gene. FEBS Lett 462: 358–362.
28. Rubio LM, Flores E, Herrero A (2002) Purification, cofactor analysis, and site-directed mutagenesis of *Synechococcus* ferredoxin-nitrate reductase. Photosynth Res 72: 13–26.
29. Lomas MW, Lipschultz F (2006) Forming the primary nitrite maximum: nitrifiers or phytoplankton? Limnol Oceanogr 51: 2453–2467.
30. Paley SM, Karp PD (2006) The pathway tools cellular overview diagram and omics viewer. Nucleic Acids Res 34: 3771–3778.
31. Ferris MJ, Palenik B (1998) Niche adaptation in ocean cyanobacteria. Nature 396: 226–228.
32. Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, et al. (2004) Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. Proc Natl Acad Sci U S A 101: 11013–11018.
33. Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425: 798–804.

34. Snel B, Bork P, Huynen MA (1999) Genome phylogeny based on gene content. Nat Genet 21: 108–110.
35. 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.
36. Palenik B, Brahamsha B, Larimer FW, Land M, Hauser L, et al. (2003) The genome of a motile marine *Synechococcus*. Nature 424: 1037–1042.
37. Ahlgren NA, Rocap G, Chisholm SW (2006) Measurement of *Prochlorococcus* ecotypes using real-time polymerase chain reaction reveals different abundances of genotypes with similar light physiologies. Environ Microbiol 8: 441–454.
38. Moore LR, R. G, S.W. C (1995) Comparative physiology of *Synechococcus* and *Prochlorococcus*: influence of light and temperature on growth, pigments, fluorescence and absorptive properties. Marine Ecology Progress Series 116: 259–275.
39. Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EM, et al. (2006) Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. Science 311: 1737–1740.
40. Zinser ER, Johnson ZI, Coe A, Karaca E, Veneziano D, et al. (2007) Influence of light and temperature on *Prochlorococcus* ecotype distribution in the Atlantic Ocean. Limnol Oceanogr 52: 2205–2220.
41. Bouman HA, Ulloa O, Scanlan DJ, Zwirglmaier K, Li WK, et al. (2006) Oceanographic basis of the global surface distribution of *Prochlorococcus* ecotypes. Science 312: 918–921.
42. West NJ, Schonhuber WA, Fuller NJ, Amann RI, Rippka R, et al. (2001) Closely related *Prochlorococcus* genotypes show remarkably different depth distributions in two oceanic regions as revealed by in situ hybridization using 16S rRNA-targeted oligonucleotides. Microbiology 147: 1731–1744.
43. Tolonen AC, Aach J, Lindell D, Johnson ZI, Rector T, et al. (2006) Global gene expression of *Prochlorococcus* ecotypes in response to changes in nitrogen availability. Mol Syst Biol 2: 53.
44. Lindell D, Jaffe JD, Coleman ML, Futschik ME, Axmann IM, et al. (2007) Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. Nature 449: 83–86.
45. Sasaki Y, Ishikawa J, Yamashita A, Oshima K, Kenri T, et al. (2002) The complete genomic sequence of *Mycoplasma penetrans*, an intracellular bacterial pathogen in humans. Nucleic Acids Res 30: 5293–5300.
46. Park JH, Burns K, Kinsland C, Begley TP (2004) Characterization of two kinases involved in thiamine pyrophosphate and pyridoxal phosphate biosynthesis in Bacillus subtilis: 4-amino-5-hydroxymethyl-2methylpyrimidine kinase and pyridoxal kinase. J Bacteriol 186: 1571–1573.
47. Toms AV, Haas AL, Park JH, Begley TP, Ealick SE (2005) Structural characterization of the regulatory proteins TenA and TenI from *Bacillus subtilis* and identification of TenA as a thiaminase II. Biochemistry 44: 2319–2329.
48. Hess WR, Rocap G, Ting CS, Larimer F, Stilwagen S, et al. (2001) The photosynthetic apparatus of *Prochlorococcus*: insights through comparative genomics. Photosynth Res 70: 53–71.
49. Steglich C, Mullineaux CW, Teuchner K, Hess WR, Lokstein H (2003) Photophysical properties of *Prochlorococcus marinus* SS120 divinyl chlorophylls and phycoerythrin in vitro and in vivo. FEBS Lett 553: 79–84.
50. Steglich C, Frankenberg-Dinkel N, Penno S, Hess WR (2005) A green light-absorbing phycoerythrin is present in the high-light-adapted marine cyanobacterium *Prochlorococcus* sp. MED4. Environ Microbiol 7: 1611–1618.
51. Zubkov MV, Fuchs BM, Tarran GA, Burkill PH, 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.
52. Lu AL, Li X, Gu Y, Wright PM, Chang DY (2001) Repair of oxidative DNA damage: mechanisms and functions. Cell Biochem Biophys 35: 141–170.
53. Dufresne A, Garczarek L, Partensky F (2005) Accelerated evolution associated with genome reduction in a free-living prokaryote. Genome Biol 6: R14.
54. Chavez S, Lucena JM, Reyes JC, Florencio FJ, Candau P (1999) The presence of glutamate dehydrogenase is a selective advantage for the Cyanobacterium *Synechocystis* sp. strain PCC 6803 under nonexponential growth conditions. J Bacteriol 181: 808–813.
55. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, et al. (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4: 41.
56. Zinser ER, Coe A, Johnson ZI, Martiny AC, Fuller NJ, 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.
57. Pal C, Papp B, Lercher MJ (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet 37: 1372–1375.
58. Bibby TS, Mary I, Nield J, Partensky F, Barber J (2003) Low-light-adapted *Prochlorococcus* species possess specific antennae for each photosystem. Nature 424: 1051–1054.
59. Sullivan MB, Lindell D, Lee JA, Thompson LR, Bielawski JP, et al. (2006) Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLoS Biol 4: e234. doi:10.1371/journal.pbio.0040234
60. Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, et al. (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. Science 311: 1768–1770.
61. Adir N, Zer H, Shochat S, Ohad I (2003) Photoinhibition: a historical perspective. Photosynth Res 76: 343–370.
62. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304: 66–74.
63. DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, et al. (2006) Community genomics among stratified microbial assemblages in the ocean's interior. Science 311: 496–503.
64. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. (2007) The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. PLoS Biol 5: e77. doi:10.1371/journal.pbio.0050016
65. Gil R, Silva FJ, Pereto J, Moya A (2004) Determination of the core of a minimal bacterial gene set. Microbiol Mol Biol Rev 68: 518–537.
66. Zhang K, Martiny AC, Reppas NB, Barry KW, Malek J, et al. (2006) Sequencing genomes from single cells by polymerase cloning. Nat Biotechnol 24: 680–686.
67. Goldberg SM, Johnson J, Busam D, Feldblyum T, Ferriera S, et al. (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proc Natl Acad Sci U S A 103: 11240–11245.
68. Badger JH, Olsen GJ (1999) CRITICA: coding region identification tool invoking comparative analysis. Mol Biol Evol 16: 512–524.
69. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Improved microbial gene identification with GLIMMER. Nucleic Acids Res 27: 4636–4641.
70. Alm EJ, Huang KH, Price MN, Koche RP, Keller K, et al. (2005) The MicrobesOnline Web site for comparative genomics. Genome Res 15: 1015–1022.
71. Swofford DL (2003) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. 4 ed. Sunderland, Massachusetts: Sinauer Associates.
72. Segre D, Zucker J, Katz J, Lin X, D'Haeseleer P, et al. (2003) From annotated genomes to metabolic flux models and kinetic parameter fitting. Omics 7: 301–316.

Figure S1
Available online at doi:10.1371/journal.pgen.0030231.sg001.

Figure S2
Available online at doi:10.1371/journal.pgen.0030231.sg002.

Figure S3
Available online at doi:10.1371/journal.pgen.0030231.sg003.

Figure S4
Available online at doi:10.1371/journal.pgen.0030231.sg004.

Table S1
Available online at doi:10.1371/journal.pgen.0030231.st001.

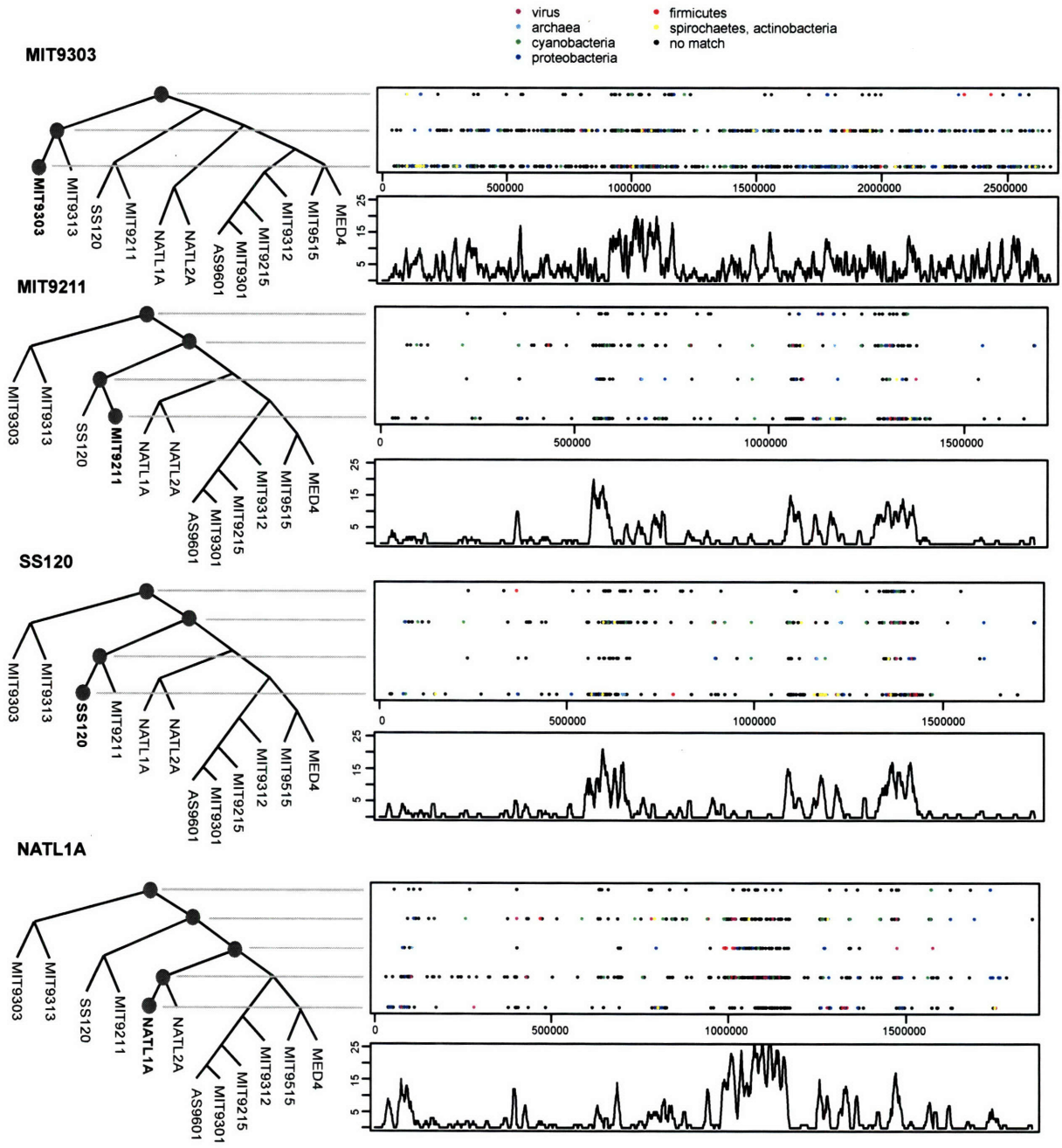
Table S3
Available online at doi:10.1371/journal.pgen.0030231.st003.

Table S4
Available online at doi:10.1371/journal.pgen.0030231.st004.

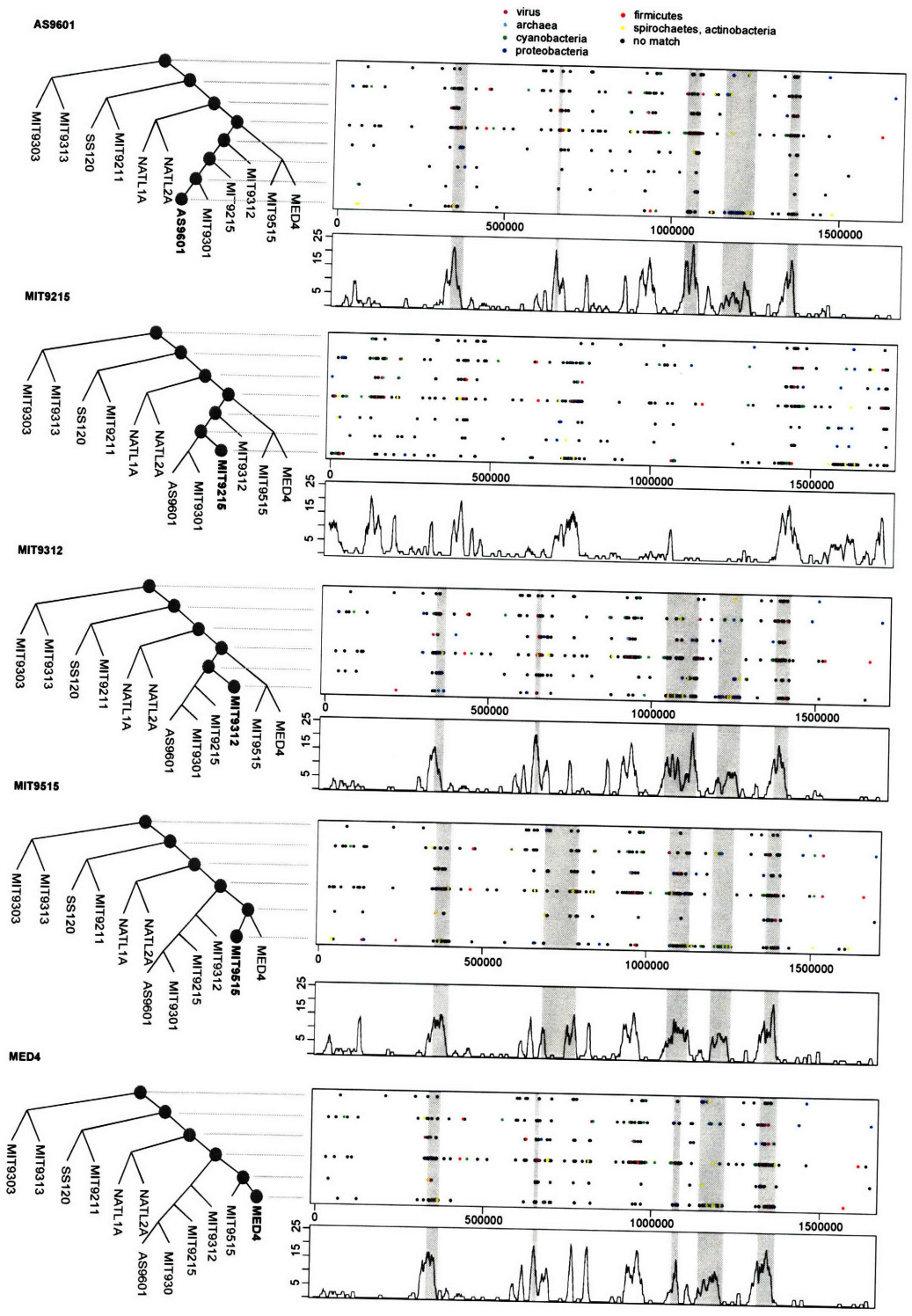
Table S6
Available online at doi:10.1371/journal.pgen.0030231.st006.

Table S7
Available online at doi:10.1371/journal.pgen.0030231.st007.

Table S8
Available online at doi:10.1371/journal.pgen.0030231.st008.



FigureS5. Islands of LL Genomes Not Represented in Figure 4. The dot plot shows the location of each gene, the ancestor in which it is estimated to be acquired, and when possible, the best match outside *Prochlorococcus*. The lower plot is the number of genes gained in a sliding window (size 10,000bp, interval 1,000bp) along the chromosome.



FigureS6. Islands of HL Genomes Not Represented in Figure 4. The dot plot shows the location of each gene, the ancestor in which it is estimated to be acquired, and when possible, the best match outside *Prochlorococcus*. The lower plot is the number of genes gained in a sliding window (size 10,000bp, interval 1,000bp) along the chromosome. When available, the locations of islands previously defined by hand are represented by shaded regions.

Table S2: *Prochlorococcus* core genes absent in *Synechococcus*. 33 orthologous groups are shared by all *Prochlorococcus* but absent in some *Synechococcus*, and only 13 of those are absent in all *Synechococcus*. For each such orthologous group, its presence or absence in each of the four *Synechococcus* genomes in this analysis is given. Also given is the locus name for the gene in MED4, its COG match, and its gene name, if available.

SynW H8102	SynCC 9605	SynCC 9902	SynCC 9311	MED4 locus	COG	gene name
Absent from *Synechococcus*						
0	0	0	0	PMED4_00681		
0	0	0	0	PMED4_02181		
0	0	0	0	PMED4_06781	COG786:Na+/glutamate symporter [Amino acid transport and metabolism]	GltS
0	0	0	0	PMED4_06861	COG1535:Isochorismate hydrolase [Secondary metabolites biosynthesis, transport, and catabolism]	EntB/pncA
0	0	0	0	PMED4_07061		
0	0	0	0	PMED4_08191		
0	0	0	0	PMED4_08941	COG2146:Ferredoxin subunits of nitrite reductase and ring-hydroxylating dioxygenases [Inorganic ion transport and metabolism / General function prediction only]	NirD / hcaI
0	0	0	0	PMED4_11001		
0	0	0	0	PMED4_11581		
0	0	0	0	PMED4_11671		
0	0	0	0	PMED4_12731		
0	0	0	0	PMED4_15181		
0	0	0	0	PMED4_15741		hli11
In some *Synechococcus*						
1	0	0	1	PMED4_00811	COG2091:Phosphopantetheinyl transferase [Coenzyme metabolism]	Sfp
0	1	1	1	PMED4_02171	COG492:Thioredoxin reductase [Posttranslational modification, protein turnover, chaperones]	TrxB
0	0	0	1	PMED4_02191	COG3329:Predicted permease [General function prediction only]	sbtA
1	1	0	0	PMED4_02251		
1	1	0	0	PMED4_03481		
0	0	0	1	PMED4_03761	COG5470:Uncharacterized conserved protein [Function unknown]	
1	1	1	0	PMED4_05531	COG1324:Uncharacterized protein involved in tolerance to divalent cations [Inorganic ion transport and metabolism]	cutA
0	1	1	1	PMED4_06771		pcbA
1	0	0	0	PMED4_07161		
0	0	1	0	PMED4_07701		
0	1	1	1	PMED4_08901	COG1528:Ferritin-like protein [Inorganic ion transport and metabolism]	/ Ftn
0	1	1	1	PMED4_08921	COG664:cAMP-binding proteins - catabolite gene activator and regulatory subunit of cAMP-dependent protein kinases [Signal transduction mechanisms]	Crp
1	1	1	0	PMED4_11181		
0	1	1	1	PMED4_11231		
0	0	0	1	PMED4_12131		
1	1	0	1	PMED4_12251		
0	1	1	1	PMED4_13361	COG716:Flavodoxins [Energy production and conversion]	FldA
0	1	0	1	PMED4_14561	COG63:Predicted sugar kinase [Carbohydrate transport and metabolism] / COG62:Uncharacterized conserved protein [Function unknown]	
1	1	1	0	PMED4_15631		
0	0	0	1	PMED4_18811		

Table S5: The most common COGs in the core and flexible genomes. We used matches against the COG database as a first impression of the differences between the core and flexible genomes. The number of *Prochlorococcus* orthologous groups, and the total number of genes in those groups, matching each COG is given. The top 10 COGs matching the core and flexible genomes are shown.

COG Name	COG Description	*Prochlorococcus* Orthologous Groups	genes
Core			
COG524	Sugar kinases, ribokinase family [Carbohydrate transport and metabolism]	3	42
COG456	Acetyltransferases [General function prediction only]	3	42
COG1132	ABC-type multidrug transport system, ATPase and permease components [Defense mechanisms]	4	56
COG740	Protease subunit of ATP-dependent Clp proteases [Posttranslational modification, protein turnover, chaperones / Intracellular trafficking and secretion]	4	56
COG745	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain [Signal transduction mechanisms / Transcription]	4	56
COG596	Predicted hydrolases or acyltransferases (alpha/beta hydrolase superfamily) [General function prediction only]	4	56
COG1028	Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases) [Secondary metabolites biosynthesis, transport, and catabolism / General function prediction only]	4	56
COG465	ATP-dependent Zn proteases [Posttranslational modification, protein turnover, chaperones]	4	56
COG568	DNA-directed RNA polymerase, sigma subunit (sigma70/sigma32) [Transcription]	5	70
COG451	Nucleoside-diphosphate-sugar epimerases [Cell envelope biogenesis, outer membrane / Carbohydrate transport and metabolism]	6	112
Flexible			
COG457	FOG: TPR repeat [General function prediction only]	7	8
COG1028	Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases) [Secondary metabolites biosynthesis, transport, and catabolism / General function prediction only]	7	20
COG2274	ABC-type bacteriocin/lantibiotic exporters, contain an N-terminal double-glycine peptidase domain [Defense mechanisms]	7	20
COG463	Glycosyltransferases involved in cell wall biogenesis [Cell envelope biogenesis, outer membrane]	8	25
COG1943	Transposase and inactivated derivatives [DNA replication, recombination, and repair]	10	11
COG399	Predicted pyridoxal phosphate-dependent enzyme apparently involved in regulation of cell wall biogenesis [Cell envelope biogenesis, outer membrane]	11	16
COG5010	Flp pilus assembly protein TadD, contains TPR repeats [Intracellular trafficking and secretion]	12	20
COG451	Nucleoside-diphosphate-sugar epimerases [Cell envelope biogenesis, outer membrane / Carbohydrate transport and metabolism]	13	33
COG3063	Tfp pilus assembly protein PilF [Cell motility and secretion / Intracellular trafficking and secretion]	17	24
COG438	Glycosyltransferase [Cell envelope biogenesis, outer membrane]	25	89

Appendix D

Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution

Debbie Lindell, Jacob D. Jaffe, Maureen L. Coleman, Matthias E. Futschik, Ilka M. Axmann, Trent Rector, Gregory Kettler, Matthew B. Sullivan, Robert Steen, Wolfgang R. Hess, George M. Church, and Sallie W. Chisholm

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

Vol 449|6 September 2007|doi:10.1038/nature06130 nature

LETTERS

Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution

Debbie Lindell[1]†, Jacob D. Jaffe[3]†, Maureen L. Coleman[1], Matthias E. Futschik[5], Ilka M. Axmann[5], Trent Rector[4], Gregory Kettler[1], Matthew B. Sullivan[1], Robert Steen[4], Wolfgang R. Hess[6], George M. Church[3] & Sallie W. Chisholm[1,2]

Interactions between bacterial hosts and their viruses (phages) lead to reciprocal genome evolution through a dynamic co-evolutionary process[1–5]. Phage-mediated transfer of host genes—often located in genome islands—has had a major impact on microbial evolution[1,4,6]. Furthermore, phage genomes have clearly been shaped by the acquisition of genes from their hosts[2,3,5]. Here we investigate whole-genome expression of a host and phage, the marine cyanobacterium *Prochlorococcus* MED4 and the T7-like cyanophage P-SSP7, during lytic infection, to gain insight into these co-evolutionary processes. Although most of the phage genome was linearly transcribed over the course of infection, four phage-encoded bacterial metabolism genes formed part of the same expression cluster, even though they are physically separated on the genome. These genes—encoding photosystem II D1 (*psbA*), high-light inducible protein (*hli*), transaldolase (*talC*) and ribonucleotide reductase (*nrd*)—are transcribed together with phage DNA replication genes and seem to make up a functional unit involved in energy and deoxynucleotide production for phage replication in resource-poor oceans. Also unique to this system was the upregulation of numerous genes in the host during infection. These may be host stress response genes and/or genes induced by the phage. Many of these host genes are located in genome islands and have homologues in cyanophage genomes. We hypothesize that phage have evolved to use upregulated host genes, leading to their stable incorporation into phage genomes and their subsequent transfer back to hosts in genome islands. Thus activation of host genes during infection may be directing the co-evolution of gene content in both host and phage genomes.

Prochlorococcus is the dominant photosynthetic organism in vast regions of the world's oceans[7], where T7-like podoviruses are also abundant[8]. Therefore this phage–host system is likely to be of great relevance for bacterial and phage global evolution, for modelling their population dynamics, and for understanding the role of phage in the oceanic carbon cycle.

Phages infecting marine cyanobacteria encode a number of host-like genes including photosynthesis and stress-response genes[5,9–11]. Phage photosynthesis genes are expressed during infection while transcripts of homologous genes in the host decline[12,13], and are hypothesized to facilitate production of carbon and energy through cell photosynthesis for optimal phage production[5,10,12–14]. This physiological interdependence between host and phage is likely to have led to the observed prevalence of photosynthesis genes in cyanophage[10,15], providing a reservoir for genetic exchange, and influencing the co-evolutionary process of both host and phage[14,15].

Although the analysis of single genes has provided insight into this dynamic, a systems approach is essential for a broader understanding of this co-evolutionary process. Here we investigate genome-wide transcriptome dynamics of *Prochlorococcus* MED4 and the T7-like podovirus P-SSP7 over the course of infection—the first such detailed view of infection for any lytic host–phage system.

We first characterized the gross features of the lytic cycle (Fig. 1). Phage genomic DNA (gDNA) began to increase, and host gDNA to decrease, 4 h after infection, and phage progeny were first released into the extracellular medium 8 h post infection (Fig. 1a). Phage

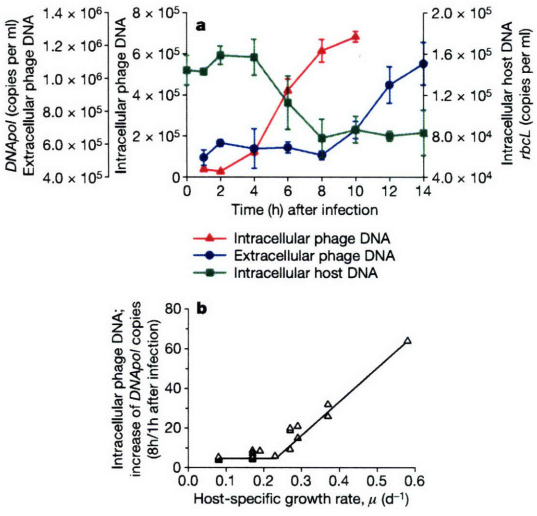


Figure 1 | Infection dynamics of *Prochlorococcus* MED4 by podovirus P-SSP7. a, Timing of phage gDNA replication (intracellular phage DNA) and length of the lytic cycle (extracellular phage DNA) was determined by quantifying the phage DNA polymerase gene (*gene 5*/*DNApol* gene copy number). Host gDNA degradation (intracellular host DNA) was determined by disappearance of the host *rbcL* gene. Average and s.d. of three biological replicates. **b**, Dependence of phage gDNA replication on host growth rate. Phage *DNApol* intracellular copy number was measured 8 h after infection and normalized to that at 1 h after infection as a measure for phage gDNA replication. $n = 24$.

[1]Department of Civil and Environmental Engineering, [2]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [3]Department of Genetics, and [4]BioPolymers Facility, Department of Genetics Harvard Medical School, Boston, Massachusetts 02115, USA. [5]Institute for Theoretical Biology, Humboldt University, Berlin 10115, Germany. [6]Institute of Biology, University of Freiburg, Freiburg 79104, Germany. †Present addresses: Department of Biology, Technion—Israel Institute of Technology, Haifa 32000, Israel (D.L.); The Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02141, USA (J.D.J).

replication is a function of host growth rate (Fig. 1b), which, together with a dependence on photosynthesis[12], suggests an intimate link between phage fitness and host physiology. We wanted to know: what are the dynamics of the phage and host transcriptome during infection, and where do phage-encoded 'bacterial-like' genes fit into this transcriptional program?

The genome content and architecture of the cyanophage P-SSP7 are similar to that of the *Escherichia coli*-infecting T7 podovirus[11]. As in T7 (ref. 16), the P-SSP7 genome was transcribed linearly from the left to the right of the genome map over the course of infection (with important exceptions, see below) (Fig. 2), and three expression clusters were discerned (Fig. 2a, and Supplementary Fig. 1). The first cluster contains a putative *marR* transcriptional regulator gene, a T7 homologue (g0.7) suggesting a role in redirecting transcription from the host towards the phage. The second cluster contains genes involved in DNA metabolism and replication (Fig. 2a, and Supplementary Fig. 1) as well as RNA polymerase (RNAP), which may be involved in RNA transcription and/or DNA replication[16]. The third cluster consists of genes involved in phage particle formation and DNA maturation. Proteins encoded by this latter cluster were detected in the mature phage particle (Fig. 2b, and Supplementary Table 1), further supporting that many are phage structural genes. Thus the three expression clusters in this cyanophage are analogous to T7-coliphage class I, II and III genes in both gene content[11] and the timing of genome expression (Fig. 2). That these fundamental operational properties are conserved across cyanophage and enteric phage, the hosts of which are drastically different with respect to energy source (autotroph versus heterotroph), habitat (nutrient-poor oceanic waters versus the nutrient-rich human gut), and growth rate (generation time of a day versus less than an hour), is remarkable.

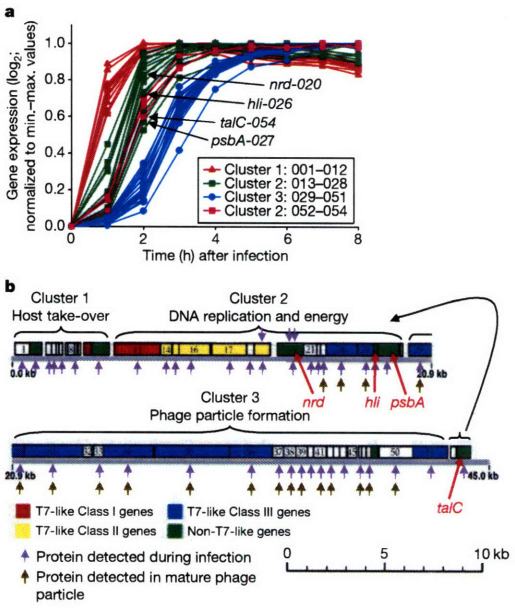


Figure 2 | Temporal expression dynamics of P-SSP7 phage genes during infection of *Prochlorococcus* MED4. a, Transcript levels with time after infection reveal three transcription clusters (see Supplementary Fig. 1). Profiles determined from microarrays, were normalized to minimum–maximum values for each gene. Average of three biological replicates; Supplementary Fig. 6 shows RT–PCR verification of results. **b**, Genome map[11] highlighting the position of *talC* at the end of the genome, even though it is transcribed in cluster 2. Protein detection during infection (purple arrows) and in mature phage particle (brown arrows) showing that 74% of phage genes produced proteins, including three overlapping genes that escaped previous annotation (Supplementary Table 6). Supplementary Table 1 encompasses gene identification and peptide detection.

Despite the similar overall infection strategies of P-SSP7 and T7, transcription cluster 2 in the cyanophage displays novel features in both gene content and regulation and bears signatures of host–phage co-evolution unique to the marine ecosystem. This cluster contains four 'bacterial-like' genes: the ribonucleotide reductase gene *nrd* (ORF 020), the high-light-inducible stress response gene *hli* (ORF 026), the photosystem II gene *psbA* (ORF 027), and the transaldolase gene *talC* (ORF 054). Although *nrd*, *hli* and *psbA* are in the middle of the genome, *talC* is at the end[11] (Fig. 2b). The co-transcription of these four genes, despite their physical separation (Fig. 2, and Supplementary Fig. 2), suggests that they are functionally linked[3].

Clues as to the function of the 'bacterial-like' genes are given by their position in the transcriptional and translational program of the entire host–phage system. First, the proteins encoded by these genes are present during infection but absent from the mature phage particle (Fig. 2b, and Supplementary Table 1), indicating that they function intracellularly. Second, these genes are transcribed together with DNA replication genes, and include ribonucleotide reductase, which converts host ribonucleotides, recycled from degraded RNA (see below), to deoxynucleotides. The photosynthesis genes found in this cluster are thought to be involved in the production of energy[5,10,12–14] and transaldolase may function in the host's pentose phosphate pathway to produce reducing power (NADPH) and/or ribose substrates for nucleotide synthesis[11]. Together, these findings suggest that these genes form a functional unit to produce energy and deoxynucleotide carbon substrates necessary for cyanophage DNA replication in the resource-poor oceans.

The bacterial-like metabolism genes found in P-SSP7 are also commonly found in myoviruses that infect marine cyanobacteria, despite drastic differences in their core genome content[9,11]. In some myoviruses, however, the genes are situated together on the genome[10,11]. Therefore we may be seeing a snapshot of evolution in progress, from spatial separation with cotranscription in P-SSP7, to physically linked genes in other cyanophage genomes.

It is not at all clear how the transcription of this cyanophage genome is regulated, and, in particular, how the last three genes are co-regulated with cluster 2 genes. Although we bioinformatically detected host-like RNAP promoters upstream of each phage expression cluster, and ORF 052, no clear phage-like RNAP promoters were detected[17] (Supplementary Table 2). A transcription initiation site consistent with bacterial-like promoters was experimentally mapped upstream of cluster 2 genes, whereas 5′ ends consistent with RNA processing sites and with weak similarity to T7-like promoters, were found upstream of cluster 1 and cluster 3 genes (Supplementary Fig. 3). However, it remains unclear whether these sequences serve as phage promoters and/or RNA processing sites for transcripts generated by either host or phage RNAP.

Given the reliance of phage replication on host physiology (Fig. 1b), the behaviour of the host transcriptome during infection is of interest. Whereas the transcript levels for approximately 75% of the 1,716 host genes declined during infection (Fig. 3), 41 genes were significantly upregulated. This is distinctly different from other lytic host–phage systems where few, if any, host genes become activated[16,18]. The upregulated genes fall into two groups (Fig. 3, and Supplementary Fig. 4 and Supplementary Table 3). The first was transiently upregulated immediately after infection and consists of high-light-inducible stress response (*hli*), carbon metabolism (*rbcLS*), transcription (*rpoC2*, *rpoD*) and ribosome (*rpl5*, *rpl6*, *rps8*, *rps11*, *rps17*) genes. Transcripts of the second group appeared 2 h after infection and included genes involved in RNA degradation and modification (*rne*, *rnhB*, *dus* and *sun*), protein turnover (*clpS*, and an AAA ATPase family gene), stress responses (*umuD* and *phoH*), and those of unknown function. Two of the latter were

transcribed from bacterial-like promoters (Supplementary Fig. 3), suggesting involvement of host RNAP. Upregulated host gene expression may constitute a direct stress response to phage infection, or may have been facilitated by phage factors[16] injected into the cell or expressed from phage expression cluster 1.

Regardless of the mechanism of upregulation, we hypothesize that phages may have evolved to make use of the products of some upregulated host genes as part of the 'arms-race' between host and phage[19]. Certainly, phages are known to exploit host stress-response proteins during infection in other systems[20–22]. The T4 and T7 phages infecting *E. coli*, for example, have evolved to modify host RNase E (involved in RNA degradation) leading to the degradation of host RNA[22]. It is perhaps not coincidental that *rne* (encoding RNase E) is one of the upregulated genes in *Prochlorococcus* during infection. This may have initially served as a host defence mechanism for degrading phage RNA, but could also be exploited by phage to degrade host RNA for use as substrates for phage deoxynucleotide synthesis.

Perhaps the most compelling evidence that upregulated host genes are part of the co-evolutionary process in this system is that 34% of them (more than would occur by chance $P < 0.001$) are found in hypervariable host genome islands (Supplementary Table 3), which are thought to be mobilized by phages[6]. Furthermore, homologues of a number of these host genes are found in phage genomes, including *hli*, *phoH*, and HNH endonuclease and sigma factor genes, as well as RNase H and heat-shock genes[9,11].

Thus there seems to be a connection between genes upregulated during infection, their position in the host genome, and the presence of homologues in phage. Although there are a number of possible explanations for this connection, the most parsimonious evolutionary scenario is as follows: Host stress response genes are upregulated in response to phage infection. Phages that have evolved to use these gene products gain a fitness advantage. Random incorporation[3] of these genes into their own genomes would enable phages to more tightly regulate their expression, conferring a fitness advantage, and leading to preferential retention. This retention would increase the probability of transfer back to the host in genome islands, by lysogeny or unsuccessful infection, and those genes beneficial to the host would remain in the host genome. Analysis of the *hli* gene family provides an interesting illustration in support of this scenario. *hli* genes are upregulated in the host in response to phage infection (Fig. 3, and Supplementary Table 3), are common in *Prochlorococcus*-infecting phage genomes[5,11], and multiple phage-like copies[5] are found in *Prochlorococcus* genome islands[6] (Supplementary Table 4). Furthermore, their differential expression in *Prochlorococcus* in response to various environmental stressors (Supplementary Table 4) and the presence of a binding site for the nitrogen transcriptional regulator NtcA upstream of the nitrogen-regulated *hli10* gene[23], suggests that copies acquired from phage[5] have undergone specialization of function in the host. It remains to be seen whether host fitness has been enhanced by the acquisition of these *hli* genes from cyanophages.

This system-wide analysis of the infection of a cyanobacterium by a phage has led to new insights and hypotheses regarding co-evolutionary interactions between host and phage. These interactions clearly shape the gene content of both host and phage, and probably play a role in shaping the distribution and abundance of cyanobacterial ecotypes in the oceans.

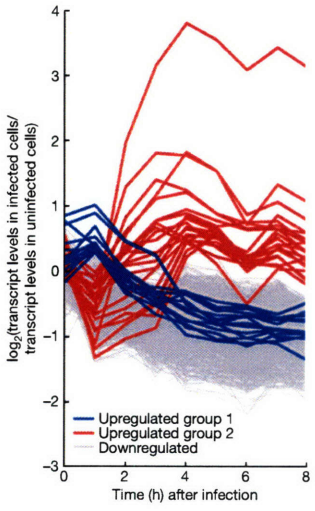


Figure 3 | Transcriptional profiles of *Prochlorococcus* MED4 genes with time after infection by P-SSP7. Transcript levels, determined from microarrays, are presented as $\log_2$-fold change in infected cells relative to uninfected cells over the 8 h latent period of infection. Only genes whose expression levels were significant at a false-discovery rate of $q < 0.05$ are shown. Blue and red indicate significantly upregulated genes in transcription groups 1 and 2, respectively (see Supplementary Table 3). Grey indicates genes significantly downregulated at 8 h after infection. Average of three biological replicates. Supplementary Fig. 7 shows RT–PCR verification of results.

METHODS SUMMARY

Prochlorococcus MED4 was grown at 21 °C under 10–25 μmol photon m^{-1} s^{-1} continuous white light in Pro99 seawater medium with HEPES and sodium bicarbonate. The length of the lytic cycle was determined by quantifying phage DNA in the extracellular medium using a real-time quantitative PCR (qPCR) assay (see Supplementary Fig. 5 for a comparison with standard methods). The timing of phage DNA replication and host DNA degradation were determined intracellularly using qPCR assays for the phage *DNApol* and host *rbcL* genes, respectively. For expression analysis triplicate cultures (10^8 cells ml^{-1}) were infected with the P-SSP7 podovirus (3×10^8 infective phage particles ml^{-1}) and the paired control cultures were amended with filter-sterilized spent medium. Samples were collected by centrifugation, resuspended in storage buffer and snap frozen in liquid nitrogen. RNA was extracted using Ambion's *mir*Vana RNA isolation kit and DNA was removed by DNaseI digestion using Ambion's Turbo DNA-*free* kit. Transcriptional analyses were carried out using a custom-made high-density antisense Affymetrix array—MD4-9313. Two micrograms of total RNA were subjected to Affymetrix protocols for *E. coli*. Array analyses were carried out using R and Bioconductor, and array data were normalized and probe set summaries calculated using the robust multi-array average (RMA) procedure[24]. Array results were validated by RT–PCR (Supplementary Figs 6, 7) and the appropriate normalization method was determined by comparing normalized transcription profiles to RT–PCR results (Supplementary Table 5 and Supplementary Figs 8, 9, 10). Promoters were computationally predicted and experimentally assessed using the 5′ RACE technique. For the detection of phage proteins, *Prochlorococcus* cells were harvested 3 and 7 h after infection with phage, and 10^{10} caesium-chloride-purified phage particles were subjected to mass spectrometry proteomic analysis as in ref. 25. See Supplementary Methods for details of all experimental procedures.

Received 5 June; accepted 26 July 2007.

1. Canchaya, C., Proux, C., Fournous, G., Bruttin, A. & Brussow, H. Prophage genomics. *Microbiol. Mol. Biol. Rev.* **67**, 238–276 (2003).
2. Filee, J., Forterre, P. & Laurent, J. The role played by viruses in the evolution of their hosts: a view based on informational protein phylogenies. *Res. Microbiol.* **154**, 237–243 (2003).
3. Hendrix, R. W., Lawrence, J. G., Hatfull, G. F. & Casjens, S. The origins and ongoing evolution of viruses. *Trends Microbiol.* **8**, 504–508 (2000).
4. Hsiao, W. W. L. *et al.* Evidence of a large novel gene pool associated with prokaryotic genome islands. *PLoS Genet.* **1**, e62 (2006).
5. Lindell, D. *et al.* Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc. Natl Acad. Sci. USA* **101**, 11013–11018 (2004).
6. Coleman, M. L. *et al.* Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* **311**, 1768–1770 (2006).
7. Partensky, F., Hess, W. R. & Vaulot, D. *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol. Mol. Biol. Rev.* **63**, 106–127 (1999).
8. Breitbart, M., Miyake, J. H. & Rohwer, F. Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiol. Lett.* **236**, 249–256 (2004).

9. Mann, N. H. *et al.* The genome of S-PM2, a "photosynthetic" T4-type bacteriophage that infects marine *Synechococcus*. *J. Bacteriol.* **187,** 3188–3200 (2005).
10. Millard, A., Clokie, M. R., Shub, D. A. & Mann, N. H. Genetic organization of the *psbAD* region in phages infecting marine *Synechococcus* strains. *Proc. Natl Acad. Sci. USA* **101,** 11007–11012 (2004).
11. Sullivan, M. B., Coleman, M., Weigele, P., Rohwer, F. & Chisholm, S. W. Three *Prochlorococcus* cyanophage genomes: Signature features and ecological interpretations. *PLoS Biol.* **3,** e144 (2005).
12. Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M. & Chisholm, S. W. Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* **438,** 86–89 (2005).
13. Clokie, M. R. J., Shan, J., Bailey, S., Jia, Y. & Krisch, H. M. Transcription of a 'photosynthetic' T4-type phage during infection of a marine cyanobacterium. *Environ. Microbiol.* **8,** 827–835 (2006).
14. Zeidner, G. *et al.* Potential photosynthesis gene recombination between *Prochlorococcus* and *Synechococcus* via viral intermediates. *Environ. Microbiol.* **7,** 1505–1513 (2005).
15. Sullivan, M. B. *et al.* Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol.* **4,** e234 (2006).
16. Molineux, I. in The Bacteriophages (ed. Calendar, R.) 277–301 (Oxford University Press, New York, 2005).
17. Chen, Z. & Schneider, T. D. Information theory based T7-like promoter models: classification of bacteriophages and differential evolution of promoters and their polymerases. *Nucleic Acids Res.* **33,** 6172–6187 (2005).
18. Miller, E. S. *et al.* Bacteriophage T4 genome. *Microbiol. Mol. Biol. Rev.* **67,** 86–156 (2003).
19. Lenski, R. E. & Levin, B. R. Constraints on the coevolution of bacteria and virulent phage: A model, some experiments, and predictions for natural communities. *Am. Nat.* **124,** 585–602 (1985).
20. Tabor, S., Huber, H. E. & Richardson, C. C. *Escherichia coli* thioredoxin confers processivity on the DNA polymerase activity of the gene 5 protein of bacteriophage T7. *J. Biol. Chem.* **262,** 16212–16223 (1987).
21. Tilly, K., Murialdo, H. & Georgopoulos, C. P. Identification of a second *Escherichia coli groE* gene whose product is necessary for bacteriophage morphogenesis. *Proc. Natl Acad. Sci. USA* **78,** 1629–1633 (1981).
22. Ueno, H. & Yonesaki, T. Phage-induced change in the stability of mRNAs. *Virology* **329,** 134–141 (2004).
23. Tolonen, A. C. *et al.* Global gene expression of *Prochlorococcus* ecotypes in response to changes in nitrogen availability. *Mol. Systems Biol.* **2,** 53 (2006).
24. Irizarry, R. A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31,** e15 (2003).
25. Jaffe, J. D. *et al.* The complete genome and proteome of *Mycoplasma mobile*. *Genome Res.* **14,** 1447–1461 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Steglich, S. Bhattacharya, H. Keller, L. Thompson, P. Weigele, S. Choe, D. Endy, S. Kosuri, M. Shmoish and J. Aach for discussions, and J. Waldbauer and M. Osburne for comments on the manuscript, and the MIT Center for Environmental Health Sciences. This work was funded by the DOE Genomes to Life System Biology Center Grant (G.M.C. and S.W.C.), the Gordon and Betty Moore Foundation's Marine Microbiology Program (S.W.C.), and the National Science Foundation (S.W.C).

Author Information The microarray data have been deposited in the GEO database under the accession number GSE8382. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.W.C. (chisholm@mit.edu).

SUPPLEMENTARY INFORMATION

Supplemental Methods

Culture Growth and Experimental Design

Prochlorococcus MED4 was grown in the Pro99 seawater based medium amended with 10 mM HEPES (pH7.5) and 12 mM sodium bicarbonate at 21 °C under continuous white light at 10-25 μmol photon·m^{-1}·s^{-1} as described in Lindell et al.[12]. For experiments in which host gDNA was quantified and expression analyses carried out, the ratio of infective phage to host (the MOI) was 3.0 to maximize levels of infection. Phage at $3x10^8$ infective particles·ml^{-1} were added to 10^8 cells·ml^{-1} and samples were collected each hour during the course of infection. Prior to phage addition for the expression experiment, cells were concentrated to 10^8 cells·ml^{-1} by centrifugation. Experiments were carried out with triplicate independent cultures in paired experiments and control treatments were amended with filter-sterilized spent medium. For experiments in which the length of the lytic cycle and the timing of phage gDNA replication were determined, the ratio of infective phage to host was 0.1. Phage at 10^7 infective particles·ml^{-1} were added to 10^8 cells·ml^{-1} and allowed to adsorb for 1 h. The phage-cell mix was then diluted 100 fold and samples collected at different times after phage addition.

Quantification of phage particles and phage and host genomic DNA

Extracellular Phage Quantification

Phage particles from the extracellular medium were quantified using a quantitative PCR (qPCR) method. Samples were filtered over 0.2 μm sterile syringe filters (Tuffryn HT), and the filtrate, containing phage particles, was collected. To prevent PCR inhibition by the seawater based growth medium, the filtrate was diluted 20-100 fold in 10 mM Tris pH8, 10 μl of which was used in triplicate qPCR assays with the P-SSP7 specific DNA polymerase primers (see Suppl. Table 7 for primer sequences). Quantification was achieved using a standard curve of phage particles in 10 mM Tris pH8 that had been enumerated by epifluorescence microscopy after SYBR staining (see below).

This qPCR method was compared to standard methodology for determining phage numbers in the extracellular medium (Suppl. Fig. 5). The number of infective phages was determined by the Most Probable Number (MPN) assay[26]. Briefly, phage samples were serially diluted and added to exponentially growing MED4 cells in 96-well plates. The clearing of wells, as compared to control wells, was monitored using a Synergy HT Biotek fluorescence plate reader. The number of cleared wells at the appropriate phage dilutions was used to calculate the most probable number of infective phage in the undiluted sample. Total phage particles were enumerated using epifluoresence microscopy after DNA staining of the phage[27]. Briefly phage samples were filtered onto a 0.02 μm Anodisc (Whatman) filter with a vacuum of 7 in of Hg and allowed to dry. The filter was then stained with SYBR Green I (Molecular Probes), and the sample enumerated by epifluorescence microscopy after addition of anti-fade solution containing 0.1% p-phenylenediamine.

Intracellular phage and Prochlorococcus DNA quantification

Prochlorococcus cells were collected onto 0.2 μm pore-sized polycarbonate filters (Osmonics) by filtration at 8-10 in of Hg. The filters were washed 3 times with sterilized

seawater to reduce the presence of extracellular phage, once with 3 ml preservation solution (10 mM Tris, 100 mM EDTA, 0.5 M NaCl; pH8) and then frozen at -80 °C. A heat lysis method was used to extract DNA from *Prochlorococcus* cells[28]. Briefly, the polycarbonate filter with *Prochlorococcus* cells was immersed in 650 µl of 10 mM Tris pH8, and agitated in a mini-bead beater for 2 min at 5000 rpm without beads. Five hundred µl of the sample was removed from the shards of filter and heated at 95 °C for 15 min. Ten µl was used in triplicate qPCR reactions. Phage DNA was amplified with P-SSP7 specific DNA polymerase primers, and *Prochlorococcus* DNA with *rbcL* primers (see Suppl. Table 7 for primer sequences).

Quantitative PCR protocol
Triplicate real-time PCR assays were carried out for each sample using Qiagen's QuantiTect SYBR Green PCR kit, primers at 0.3-1.0 µM and 10 µl samples (in 10 mM Tris pH8) in 25 µl volume reactions run on an DNA Engine Opticon (MJ Research). After 15 min denaturation at 95 °C, 40 amplification cycles were carried out as follows: Denaturation (95 °C for 15 sec); annealing (56 or 58 °C for 30 sec); elongation (72 °C for 30 sec); and fluorescence plate read (for quantification of SYBR green incorporation into double stranded DNA), and were followed by 5 min at 72 °C and melt curve analysis (read every degree from 50-90°C). Quantification of template was determined from standard curves produced with dilution series of P-SSP7 phage particles or *Prochlorococcus* MED4 genomic DNA. Melt curve analysis was used to verify that a single product was amplified.

RNA extraction
Samples were collected by centrifugation (12,400 Xg for 15 min at 20 °C), resuspended in storage buffer (200 mM sucrose, 10 mM sodium acetate pH5.2, 5 mM EDTA), snap frozen in liquid nitrogen and stored at -80 °C. Prior to RNA extraction the storage buffer was removed after spinning the cells for 2 min at 20,000 Xg at 20 °C. RNA was extracted using Ambion's *mir*Vana RNA isolation kit. DNA was removed by DNase I digestion using the Turbo DNA-*free* kit (Ambion). For microarray analysis 8 µg of the nucleic acid extract was digested with 6 U of Turbo DNase during a 60 min incubation at 37 °C followed by DNase I inactivation with inactivation slurry. The RNA was purified and concentrated by sodium acetate/ethanol precipitation. DNA removal was verified by gel electrophoresis. For RT-PCR analysis DNA was removed from 0.1-0.5 µg of the nucleic acid extract using the above procedure but without the precipitation step. DNA removal was verified by running no RT controls followed by qPCR for each sample (see RT-PCR validation of array results).

Array experimentation
Transcriptional analysis was carried out using a custom-made high density antisense Affymetrix array – MD4-9313. Synthesis of complementary DNA (cDNA), labeling, hybridization, staining and scanning was carried out according to Affymetrix protocols for *E.coli* (http://www.affymetrix.com/support/technical/manual/expression_manual.affx) with minor changes. Total RNA (2 µg) was denatured at 70 °C and annealed to random hexamer primers (25 ng/µl) at 25 °C for 10 min. The RNA was reverse transcribed to produce cDNA with Superscript II (25 U/µl – Invitrogen Life Technologies) and 0.5 mM

dNTPs in the presence of 1 U/μl RNase Out RNase Inhibitor (Invitrogen). The mix was incubated at 25 °C for 10 min followed by 60 min incubations at 37 °C and 42 °C respectively. Superscript II was inactivated with a 10 min incubation at 70 °C. Sodium hydroxide (0.25 N) was used to remove RNA during a 30 min incubation at 65 °C, followed by neutralization with HCl. The cDNA was purified with MinElute PCR purification columns (Qiagen). Fragments of cDNA, 50-200 nt long, were produced from a 10 min incubation at 37 °C with DNase I (0.6 U per μg cDNA), followed by heat inactivation of the DNase I enzyme (10 min at 98 °C). The cDNA fragments were end-labeled with biotin using the BioArray Terminal Labeling Kit (Enzo) during a 60 min incubation at 37 °C. The reaction was stopped by freezing at –20 °C. The quality of biotin end-labeling was verified by gel-shift assays with NeutrAvidin (Pierce Chemicals) on 1% TBE agarose gels.

The cDNA was hybridized to the MD4-9313 custom Affymetrix array (see below for array description) in aqueous hybridization solution (100 mM MES, 1 M NaCl, 20 mM EDTA, 0.01% Tween-20, 0.1 mg/mL Herring Sperm DNA, 0.5 mg/mL BSA, 7.8 % DMSO and 3 nM prelabeled Affymetrix hybridization B2 oligo control probe mix) during a 16 h incubation at 45 °C in a GeneChip Hybridization Oven 320 rotating at 60 rpm. Washes and stains were carried out on a GeneChip Fluidics Station 450 (Affymetrix) following the ProkGE_WS2v3 Affymetrix protocol. Briefly, following two stringency washes the array was sequentially incubated with 10 μg/mL streptavidin (Pierce Chemical), 5 μg/mL biotinylated anti-streptavidin goat antibody (Vector Laboratories) and 0.1 mg/mL goat IgG (Sigma) and 10 μg/mL streptavidin-phycoerythrin conjugate (Mol. Probes) each for 10 min at 25 °C. After a final wash the arrays were scanned with the GeneChip Scanner (Affymetrix) at a 2.5 μm resolution with excitation set for 570 nm.

Array Design

The MD4-9313 (MD4-9313a520062) array is a custom-made high density antisense Affymetrix array. This array detects labeled cDNA that is antisense to the original RNA. It contains probes for the genomes of two cyanobacterial strains, *Prochlorococcus* MED4 and *Prochlorococcus* MIT9313[29], as well as two dsDNA phages that infect *Prochlorococcus* MED4 – the podovirus P-SSP7 and the myovirus P-SSM4[11]. For the *Prochlorococcus* genomes, the array contains probe sets to detect all predicted open reading frames with probe pairs approximately every 80 bases. For short open reading frames (ORF), the length of the gap was reduced to ensure a minimum of 11 probe pairs per ORF where possible. Probes were designed for intergenic regions longer than 35 bases and were spaced every ca 45 bases on both strands. For short intergenic regions the gap between probe pairs was reduced to ensure a minimum of 4 probe pairs where possible. For some short sequences, where insufficient high performance probes were designed using this approach, the best probes possible were designed with no regard for their spacing along the genome feature. For the phage isolates, probe pairs were designed across the genomes for both strands at an approximate interval of 90 bases. Probes are 25 bases long and each probe pair consists of a perfect match probe (identical to the sequence) and a mismatch probe (containing a single base change at the center of the probe).

Array Data Analyses

Normalization and Statistical Analyses

Data analyses were carried out in the statistical language R using several Bioconductor packages[30]. The array data were normalized and probe set summaries calculated from perfect match probe intensities in Affymetrix CEL files using quantile robust multi-array average (RMA) analysis[24] as implemented in the Bioconductor package *affy*[31]. See below for determination of appropriate normalization method for this experiment. Statistical significance of differentially expressed genes between infected and control cells at each time point was determined using the Bayesian t-test function implemented in the GoldenSpike[32] package (originally derived from the *hdarray* package) with the confidence level set to 9[32, 33]. The results are comparable to those obtained when RMAExpress Version 0.3 beta 1[24], Cyber-T[34] and Q-value[35] were used as stand-alone programs (data not shown). Control arrays at 4 and 8 h after infection gave the same expression profiles (data not shown) and were used for comparison with infected cells at 5, 6, 7 h after infection.

Clustering Analyses

Hierarchical clustering of phage genes was carried out using Pearson correlation and average linkage in the *stats* package in R. Input data was the average logged expression values of 3 biological replicates, standardized so that mean expression values for each gene equal zero and standard deviation equals one. The dendogram, visualized with Java TreeView[36], suggests the presence of several distinct expression clusters (Suppl. Fig. 1a). To determine the number of reliable clusters in the data, a resampling approach was applied[37] using the Bioconductor package *clusterStab*[38] whereby randomly selected sub-sets of genes are repeatedly clustered and the extent of similarity between the resulting clusters are examined. Reliable (or stable) clusters are those which repeatedly occur for the random sub-sets of genes. The similarity between clusters of different repeats was measured by the Jaccard coefficient ranging from zero (no similarity) to one (identical clustering). This resampling strategy was used for a range of number of clusters (k=2 to k=5) and the resulting distribution of Jaccard coefficients compared. If an adequate number of clusters is chosen, the distribution of coefficients will show an enrichment of values equal or close to one. A comparison of the histograms for the Jaccard coefficients strongly indicated the existence of three stable expression clusters for P-SSP7 genes (Suppl. Fig. 1b). Their temporal profiles are shown in Suppl. Fig. 1c. While this normalization methodology is the most appropriate for cluster analysis, we show temporal phage gene expression profiles in Fig. 2a using minimum-maximum normalized data as these more appropriately describe the dynamics of phage genome expression from a biological perspective.

The same strategy was used to determine the number of stable clusters for upregulated MED4 genes. Here the histograms indicate that two stable expression clusters exist (Suppl. Fig. 4).

Significance of co-expression of 'bacterial-like' phage genes

Clustering analysis indicated that the 'bacterial-like' phage genes *nrd, hli, psbA* and *talC* are temporally coexpressed (Suppl. Fig. 1, Fig. 2). To stringently assess the validity of

this co-expression, two approaches were used. First, we assessed the reliability that the four genes are assigned to the same expression cluster (namely cluster 2) by a bootstrap approach using the Bioconductor package *hopach*[39] whereby genes are assigned to a particular cluster when only partial time series are used. Reliable cluster assignments should not depend on single data points and should therefore be found using only partial data. Thus, reliability can be examined by repeated bootstrap sampling and re-clustering of genes with subsequent calculation of cluster memberships. Cluster membership is defined here as the percentage of bootstrap samples that were assigned to the same original cluster. A cluster membership close to one indicates reliable assignment of a gene to the cluster. Applying this bootstrap approach, we detected high membership values for the phage genes *nrd, hli, psbA* and *talC* (1.000, 0.998, 0.9984 and 0.9999) for cluster 2 (Suppl. Fig. 2a). This strongly indicates that the 4 'bacterial-like' phage genes are co-transcribed together within cluster 2 despite their spatial separation on the genome.

In addition, a regression approach was applied to determine the degree of certainty of co-expression of the four 'bacterial-like' genes (Suppl. Fig. 2b, 2c). In this analysis we wished to estimate the time points at which expression of each P-SSP7 gene changed from being non-expressed to expressed – termed here the switch time t*. If genes are co-regulated, we expect them to have the same time t* within acceptable confidence intervals. Here we defined the switch time t* as the time at which expression values reach half maximum values. We used the following procedure to estimate t*: After averaging expression values for the 3 biological replicates, values for each gene across the time series were normalized to a minimum value of zero and maximum value of 1. All P-SSP7 genes displayed sigmoidal expression patterns. To improve the fitting by sigmoidal curves, expression values across the time series were truncated to values ranging from 0.01 to 0.95 .The truncated expression values y were fitted to the sigmoidal function; $y(t) = \exp(a.t+b)/(\exp(a.t+b) + 1)$, where a and b are fitting parameters and t is the time after infection. To allow fitting of the data by linear regression (which simplifies the calculation of confidence intervals), the data were transformed such that $y' = \log(y/(1-y))$. Subsequently, linear regression given by $y' = a.t + b$, was performed and confidence intervals were calculated for each gene. Seeing as $y=0.5$ (half maximal expression) corresponds to $y'=0$, we can use the confidence intervals for y' to assess whether the induction of genes occurred at the same time t*. The confidence intervals for *nrd* (020), *hli* (026), *psbA* (027) and *talC* (054) all overlap indicating that they are turned on simultaneously, together with the remaining genes in cluster 2 (Suppl. Fig. 2b). An example of the fitting procedure is illustrated in Suppl. Fig. 2c for the *nrd* gene.

Determination of Appropriate Normalization Method

Analysis of microarray data after implementation of various normalization methods showed differences in putative expression patterns, in particular for down-regulated genes (Suppl. Fig. 8). Therefore to ascertain which normalization method should be used for this dataset, normalized expression patterns for select genes were compared to those determined empirically with RT-PCR (see below for RT-PCR methodology). Normalization procedures tested were: RMA with quantile normalization at the probe level; RMA with normalization based on positive hybridization control spikes (AFFX-Bio* and AFFX-Cre*); Goldenspike which computes an expression summary based on 8

different normalization methods[32]; and Goldenspike without the second loess normalization at the summary level – as the assumption for this summary level normalization, that the majority of genes is not differentially expressed, may not hold for this experiment.

Comparisons were carried out on representative genes with the following expression patterns: (a) 1 unchanged, internal control gene; (b) 4 down-regulated genes and (c) 7 up-regulated genes. See Suppl. Table 5 for a list of the genes tested. (Note that PMM0550 and PMM1629 are considered both up- and down-regulated based on RMA quantile normalization.) See the "RT-PCR validation of array results" section below for more details of the genes chosen for RT-PCR validation. We compared the performance of the different normalization schemes to detect differential expression as validated by RT-PCR. These analyses show that RMA and the two versions of Goldenspike performed similarly for up-regulated expression, whereas differential expression patterns for down-regulated genes were best represented by RMA with quantile normalization (Suppl. Table 5, and see Suppl. Figs. 8, 9).

Initially, the superior performance of RMA with quantile normalization was somewhat surprising, seeing as it assumes similar overall distribution of probe intensities in different arrays, and we observed downregulation of a large number of genes. However it is important to note that a considerable percentage (25%) of the host MED4 genes were not significantly down-regulated at 8 h after infection. More importantly, however, is that the array used in this study includes a large number of probe sets other than for the host genes. It contains probe sets for intergenic regions, for the P-SSP7 phage genome and for an additional *Prochlorococcus* and phage strain. In fact, most probes on the microarray are not assigned to the organisms examined in our study. Therefore, most expression signals on the array are not expected to change and the underlying assumption of quantile normalization may hold. To assess this issue further, we compared the frequency of probe intensities for the whole array to the subset of intensities for MED4 genes after normalization. The density plots show that the signal distributions for the subset of MED4 probe sets are distinct for different arrays even though the overall distributions are similar for all arrays due to quantile normalization (Suppl. Fig. 10). Thus, quantile normalization can still be applied to our study as it did not erase differences in expression for the host genes.

RT-PCR validation of array results

Total RNA (2-5 ng) was reverse transcribed with Superscript II (Invitrogen) using 2 pmol gene-specific reverse primers in 20 µl reactions following the manufacturer's instructions. The resultant cDNA was diluted with 80 µl 10 mM Tris pH8, and 10 µl was used in each of 3 triplicate quantitative PCR reactions using Quantitect Sybr Green 2x kit (Qiagen) and 0.5-1.0 µM primers (see Suppl. Table 7 for primer sequences) in 25 µl reactions, such that cDNA resulting from 0.2-0.5 ng total RNA was used in each qPCR reaction (see above for quantitative PCR protocol). Results were further normalized to *rnpB* transcript levels which served as an internal control. No RT controls, carried out under identical conditions but without the reverse transcriptase enzyme, indicated that gDNA contamination was less than 1 % of the RT-PCR signal in all samples. Standard

curves were carried out with MED4 gDNA or P-SSP7 phage particles. Expression levels from infected cells at different times after infection was compared to that for control cells. Significance of differential expression was determined from two-tailed t-tests.

Phage genes chosen for RT-PCR validation included the first gene of each expression cluster as well as the last gene in the genome (*talC*) as a representative of the 3 genes transcribed out of order on the genome. RT-PCR validation of host genes included downregulated and upregulated genes from both up-regulated expression clusters and were chosen to span low, medium and high array signal intensities as well as to include genes of potential biological interest where possible.

Promoter analysis

Bioinformatic analysis of phage P-SSP7 transcriptional signals

The computational prediction of bacterial promoters was based on a position specific weight matrix established for the -10 box of *Prochlorococcus* MED4 promoters[40]. Bacterial terminators were found by using the TransTerm algorithm[41], which detects rho-independent transcription terminators by searching for stem-loop-structures (inverted repeats) followed by a row of T's in the genome. Putative recognition sites for the phage RNA polymerase were searched *in silico* with a consensus sequence for T7 RNA polymerase allowing substitutions at positions, which are not common among all 47 natural phage promoters[42].

Experimental detection of 5' transcript ends

The 5' ends of mRNA transcripts from P-SSP7 and MED4 were mapped using the 5' Rapid Amplification of cDNA ends (RACE) technique, described previously by Bensing et al.[43] and modified for *Prochlorococcus* by Vogel et al.[40]. Briefly, 0.7-1.5 μg total RNA was used to cleave the 5' triphosphate, found in primary transcripts, with tobacco acid pyrophosphatase (TAP) (Epicentre, Madison, Wisconsin USA). The resulting 5' monophosphate was subsequently ligated, using T4 RNA ligase (Epicentre, Madison, Wisconsin USA), to the 3' hydroxyl group of an RNA oligonucleotide (5' adaptor: GAU AUG CGC GAA UUC CUG UAG AAC GAA CAC UAG AAG AAA). A gene-specific DNA primer (see Suppl. Table 8 for gene specific primer sequences) was used for reverse transcription followed by PCR amplification with a nested gene-specific primer and the 5' adaptor primer (ATA TGC GCG AAT TCC TGT AGA ACG AAC ACT AG). The amplification products were cloned and sequenced, and the first nucleotide downstream of the 5' adaptor RNA was assigned as the 5' end. This method enables the differentiation between transcription initiation sites of primary transcripts and RNA processed sites. For primary transcripts (carrying a 5' triphosphate) the TAP treated samples (TAP+) yield a specific or strongly enhanced amplification product relative to untreated samples (TAP-), whereas amplification products of equal intensity found for both TAP treated and untreated RNA samples are indicative of processed 5' ends that already carried a monophosphate at the 5' end.

Protein Analyses

Protein Extraction and Digestion to Peptides

Cells were collected by centrifugation and stored as described for RNA work, prior to lysis in 3 M urea, 0.05% SDS, and 50 mM Tris-HCl pH 8.0. Protein levels were quantified using the bicinchoninic acid method (Pierce). Samples were digested with sequencing grade trypsin (Promega) (protein:trypsin = 137.5:1) overnight at 37 °C, reduced with 10 mM DTT, alkylated with 50 mM iodoacetamide, and acidified to < pH 3.0 with HCl.

Peptide Fractionation and identification by Ion Trap Mass Spectrometry (MS)

Two phage infection time points (3 h and 7 h post infection) were subjected to comprehensive MS/MS sequencing experiments. For this purpose, each sample was adjusted to 25% acetonitrile and centrifuged to remove particulates. The entire sample was subjected to two-dimensional chromatographic fractionation (strong cation exchange followed by reversed phase) as in Jaffe et al. [44]. The eluate of the nano-flow reversed phase column was coupled directly to a LTQ linear ion trap mass spectrometer (ThermoElectron, Waltham, MA) where the top 10 most abundant MS ions were sampled for MS/MS sequencing in each scan cycle. Dynamic exclusion was employed to increase depth of coverage. In all, 60 Strong Cation Exchange (SCX) fractions were analyzed for each sample. The accumulated spectra were analyzed with SEQUEST[45], searching against a database of all predicted proteins from *Prochlorococcus* MED4 as well as the entire genomic sequence of cyanophage P-SSP7 prepared for proteogenomic mapping as in Jaffe et al.[44]. Criteria for valid spectra assignments and creation of proteogenomic maps for the phage were as in Jaffe et al.[44]. All validated peptides were considered to be potentially 'present' in subsequent analyses. It should be noted that detection of peptides is dependent on its ionization properties and on it being of suitable length, therefore the inability to detect a particular peptide using this methodology is not a definitive indication of the lack of its presence.

Phage Particle Purification for protein analysis

Prochlorococcus MED4 was infected with P-SSP7 and harvested once the culture had cleared. The cell lysate was centrifuged at 12,000 Xg for 30 min to remove unlysed cells and cellular debris. The supernatant was incubated for 30 min at 25 °C with 1 μg DNase I (Sigma) to degrade host gDNA from lysed cells. The salt concentration of the solution was brought up to 2 M with NaCl and incubated at 25 °C for an additional 30 min and then spun at 15,000 Xg for 30 min and the pellet discarded. Triton X-100 (0.1 % v/v final concentration) and PEG 8000 (10 % w/v final concentration) was added to the supernatant and stirred gently until fully dissolved and incubated overnight at 4 °C. Phages were collected by centrifugation at 12,400 Xg for 30 min at 4 °C and resuspended in Pro99 medium. Phage particles were purified on a cesium chloride step gradient (rho=1.4/1.6) prepared in 0.2 μm filtered seawater amended with 50 mM $MgCl_2$, 50 mM Tris pH8 and 0.1 % Triton X-100, and spun at 150,000 Xg for 2 hours. Purified phage particles were dialyzed in a step-wise fashion against 1 l of 1M NaCl, 50 mM $MgCl_2$, 50 mM Tris pH8 for 1 hour and twice for an hour each against 1 l of 100 mM NaCl, 50 mM $MgCl_2$, 50 mM Tris pH8.

Determination of proteins in purified phage particles
The equivalent of 10^{10} purified phage particles was subjected to proteomic analysis. The sample was digested for 18 hours at 37 °C with 0.2 μg of trypsin in a buffer consisting of 3M urea, 25 mM Tris pH 8.0, 25 mM $MgCl_2$, and 50 mM NaCl. The sample was reduced and alkylated as above, except that 5 mM DTT and 12.6 mM iodoacetamide were used. The sample was desalted using an Oasis HLB solid phase extraction column (Waters, 10 mg resin) according to the manufacturer's directions, reduced to dryness by vacuum centrifugation, and resuspended in 10 μl of 5% acetonitrile/5% formic acid. The sample was analyzed with 3 x 3μl injection to LCMS as above using a top 10 MS/MS method on an LTQ-FT mass spectrometer. Spectra were extracted and searched using SpectrumMill (Agilent, Palo Alto, CA) against the same hybrid database of phage and host proteins described above. Standard SpectrumMill autovalidation parameters were used to select confidently identified proteins and peptides.

Supplemental References

26. Taylor, J. The estimation of numbers of bacteria by tenfold dilution series. Journal of Applied Bacteriology 25, 54-61 (1962).
27. Noble, R. T. & Fuhrman, J. A. Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. Aq. Microb. Ecol. 14, 113-118 (1998).
28. Zinser, E. R. et al. *Prochlorococcus* ecotype abundances in the North Atlantic Ocean as revealed by an improved quantitative PCR method. Applied and Environmental Microbiology 72, 723-32 (2006).
29. Rocap, G. et al. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424, 1042-1047 (2003).
30. Gentleman, R. C. et al. Bioconductor: Open software development for computational biology and bioninformatics. Genome Biology 5, R80 (2004).
31. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. Affy - analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20, 307-315 (2004).
32. Choe, S. E., Boutros, M., Michelson, A. M., Church, G. M. & Halfon, M. S. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. Genome Biology 6, R16 (2005).
33. Baldi, P. & Long, A. D. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. Bioinformatics 17, 509-519 (2001).
34. Long, A. D. et al. Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Journal of Biological Chemistry 276, 19937-19944 (2001).
35. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences U S A 100, 9439-9445 (2003).
36. Saldanha, A. J. Java Treeview - extensible visualization of microarray data. Bioinformatics 20, 3246-3248 (2004).
37. Ben-Hur, A., Elisseeff, A. & Guyon, I. A stability based method for discovering structure in clustered data. Pacific Symposium on Biocomputing (2002).
38. Smolkin, M. & Ghosh, D. Cluster stability scores for microarray data in cancer studies. BMC Bioinformatics 4, 36-42 (2003).

39. Pollard, K. S. & van der Laan, M. J. Cluster Analysis of Genomic Data. *In:* Bioinformatics and Computational Biology Solutions Using R and Bioconductor; R. Gentleman, V. Carey, W. Huber, R. Irizarry, S. Dudoit (eds.) Springer, 209-229 (2005).
40. Vogel, J., Axmann, I. M., Herzel, H. & Hess, W. R. Experimental and computational analysis of transcriptional start sites in the cyanobacterium *Prochlorococcus* MED4. Nucleic Acids Research 31, 2890-9 (2003).
41. Ermolaeva, M. D., Khalak, H. G., White, O., Smith, H. O. & Salzberg, S. L. Prediction of transcription terminators in bacterial genomes. J. Mol. Biol. 301, 27-33 (2000).
42. Imburgio, D., Rong, M., Ma, K. & McAllister, W. T. Studies of promoter recognition and start site selection by T7 RNA polymerase using a comprehensive collection of promoter variants. Biochemistry 39, 10419-10430 (2000).
43. Bensing, B. A., Meyer, B. J. & Dunny, G. M. Sensitive detection of bacterial transcription initiation sites and differentiation from RNA processing sites in the pheromone-induced plasmid transfer system of *Enterococcus faecalis*. Proceedings of the National Academy of Sciences U S A 93, 7794-7799 (1996).
44. Jaffe, J. D., Berg, H. C. & Church, G. M. Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics 45, 59-77 (2004).
45. Eng, J. K., McCormack, A. L. & Yates, J. R. r. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. Journal of the American Society for Mass Spectromotry (1994).
46. Steglich, C., Futschik, M., Rector, T., Steen, R. & Chisholm, S. W. Genome-wide analysis of light sensing in *Prochlorococcus*. Journal of Bacteriology 188, 7796-7806 (2006).
47. Dunn, J. J. & Studier, F. W. Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. J. Mol. Biol. 166, 477-535 (1983).

Supplementary Table 1: Detection of phage proteins inside the host cell during infection (infection) or in the purified phage particle (virion). Number of unique peptides detected: 1 = +; 2-4 = ++; 5-10 = +++; more than 10 = ++++. Transcription cluster designations as per Fig.2 and Suppl. Fig. 1.

ORF ID	Gene Name – Product	Infection	Virion	Transcription cluster
PSSP7_001	unknown	++		Cluster 1
PSSP7_002	unknown	++		
PSSP7_003	unknown	++		
PSSP7_004	unknown	++		
PSSP7_005	unknown			
PSSP7_006	unknown	+++		
PSSP7_007	unknown			
PSSP7_008	unknown			
PSSP7_009	unknown	+		
PSSP7_010	unknown			
PSSP7_011	*gene 0.7* – MarR family of transcriptional regulators			
PSSP7_012	*int* – integrase	+		
PSSP7_013	*gene 1* – RNA polymerase	++		Cluster 2
PSSP7_014	*gene 2.5* – ssDNA binding protein	+++		
PSSP7_015	*gene 3* – endonuclease	+		
PSSP7_016	*gene 4* – primase/helicase	++++		
PSSP7_017	*gene 5* – DNA polymerase	+++		
PSSP7_018	unknown	++		
PSSP7_019	*gene 6* – exonuclease	+++		
PSSP7_019A	unknown	++		
PSSP7_020	*nrd* – ribonucleotide reductase	+++		
PSSP7_020A	unknown	+		
PSSP7_020B	unknown	+		
PSSP7_021	unknown	+		
PSSP7_022	unknown	++		
PSSP7_023	unknown	++	+++	
PSSP7_024	*gene 8* – head-to-tail connector	+++	++++	
PSSP7_025	*gene 9* – capsid assembly protein (scaffolding protein)	+++	+	
PSSP7_026	*hli* – high-light inducible protein	+		
PSSP7_027	*psbA* – D1 photosystem II reaction center protein	++		
PSSP7_028	Unknown			Cluster 3
PSSP7_029	*gene 10* – capsid protein	++++	++++	
PSSP7_030	*gene 11* – tail tubular protein A	+	++++	
PSSP7_031	*gene 12* – tail tubular protein B	++++	++++	
PSSP7_032	unknown (putative *gene 13?)*			
PSSP7_033	unknown (putative *gene 14?* – internal core protein)	++	++++	
PSSP7_034	*gene 15* – internal core protein	+++	++++	
PSSP7_035	*gene 16* – internal core protein	++++	++++	
PSSP7_036	*gene 17* – tail fiber	++++	++++	
PSSP7_037	unknown	++	+	
PSSP7_038	unknown	+	+++	
PSSP7_039	unknown	+	++++	
PSSP7_040	unknown	+		
PSSP7_041	unknown		+	
PSSP7_042	unknown		+	
PSSP7_043	unknown			
PSSP7_044	unknown	+		
PSSP7_045	unknown			
PSSP7_046	unknown		+++	
PSSP7_047	unknown			
PSSP7_048	unknown		++	
PSSP7_049	possible endonuclease			
PSSP7_050	unknown	++	++++	
PSSP7_051	*gene 19* – DNA maturase	+		
PSSP7_052	unknown			Cluster 2
PSSP7_053	unknown			
PSSP7_054	*talC* – transaldolase family protein	+++		

Supplementary Table 2: Promoter analyses: predictions and experimental detection of 5' ends upstream of the denoted ORF. Promoter analysis: nd – not determined. None = tested but no 5' end found. Processed 5' end = product found in both TAP- and TAP+ treatments. Motif = conserved motif found in proximity of processed 5' end.

ORF ID	Product	Bioinformatic Predictions	Experimental detection of 5' ends
PSSP7_001	unknown	Bacterial -10 box: 52..57	Processed 5' end=113 motif=91..116; Processed 5' end=44686; motif=44664..44689
PSSP7_002	unknown		nd
PSSP7_003	unknown	Terminator: 1675..1691	Processed 5' end=1672; motif=1650..1675
PSSP7_004	unknown		Same site as per PSSP7_003
PSSP7_005	unknown		nd
PSSP7_006	unknown		nd
PSSP7_007	unknown	Bacterial -10 box (x2): 2591..2596; 2629..2634	nd
PSSP7_008	unknown		Nothing conserved; Processed 5' ends (5x)=2444; 2446; 2450; 2451; 2459; Nothing conserved; Processed 5' ends (3x)= 2724; 2725; 2726
PSSP7_009	unknown		nd
PSSP7_010	unknown		nd
PSSP7_011	*gene 0.7* - MarR transcriptional regulator	Bacterial -10 box (x3): 3566..3571; 3577..3582, 3585..3590	none
PSSP7_012	*int* - phage related integrase		nd
PSSP7_013	*gene 1* - RNA polymerase	Terminator: 5005..5028 Bacterial -10 box: 5034..5039	Bacterial tis=5045 Non-Processed (x2) 5' ends=5032; 5060
PSSP7_014	*gene 2.5* - ssDNA binding protein		none
PSSP7_015	*gene 3* - endonuclease		nd
PSSP7_016	*gene 4* - primase/helicase		nd
PSSP7_017_018	*gene 5* - DNA polymerase		Nothing conserved; ; Processed 5' ends (3x)=9690; 9692; 9695
PSSP7_019	*gene 6* - exonuclease		none
PSSP7_019a	unknown		nd
PSSP7_020	*nrd* - ribonucleotide reductase domain	Possible -10 box (13160..13165), identical to that found experimentally for rpl21[40].	Nothing conserved; Processed 5' end=12818; Non-Processed 5' end=12928
PSSP7_020a	unknown		nd
PSSP7_021	unknown		nd
PSSP7_022	unknown		nd
PSSP7_023	unknown		nd
PSSP7_024	*gene 8* - head-to-tail connector		none
PSSP7_025	*gene 9* capsid assembly protein		nd
PSSP7_026	*hli* - high-light inducible protein		none
PSSP7_027	*psbA* - D1 photosystem II reaction center protein		none
PSSP7_028	unknown	Bacterial -10 box: 19430..19435	none (no signal found further upstream of Processed 5' end: 19749)
PSSP7_029	*gene 10* - capsid protein		Processed 5' end: 19749 motif: 19725..19750
PSSP7_030	*gene 11* - tail tubular protein A	Terminator: 21031..21046	none
PSSP7_031	*gene 12* - tail tubular protein B		nd
PSSP7_032	Unknown (*gene 13??*)	Terminator: 24626..24650	none
PSSP7_033	unknown (*gene*14??)		nd
PSSP7_034	*gene 15* - internal core protein		nd
PSSP7_035	*gene 16* - internal core protein		nd
PSSP7_036	*gene 17* - tail fiber		none
PSSP7_037	unknown		nd
PSSP7_038	unknown		nd
PSSP7_039	unknown		nd
PSSP7_040	unknown		nd
PSSP7_041	unknown		nd
PSSP7_042	unknown		nd
PSSP7_043	unknown		nd
PSSP7_044	unknown		nd

PSSP7_045	unknown		nd
PSSP7_046	unknown		nd
PSSP7_047	unknown		nd
PSSP7_048	unknown		nd
PSSP7_049	possible endonuclease		nd
PSSP7_050	unknown	Terminator: 39402..39417	none
PSSP7_051	*gene 19* - DNA maturase	Terminator: 41165..41176	none
PSSP7_052	unknown	Bacterial -10 box (x2): 42978..42983; 42988..42993	none
PSSP7_053	unknown		nd
PSSP7_054	*talC* - transaldolase		nd
		Terminator: 44063..44075	

Supplementary Table 3: Upregulated *Prochlorococcus* MED4 genes determined from microarray analysis. Fold change (infected/control) with time (h) after infection. Positive and negative values indicate an increase and decline in transcript levels respectively. Significant increases in fold change are shown in blue and the level of significance is shown: * for q<0.05; ** q<0.01; ***q<0.001.

ORF – gene name, possible product and function	Fold Change (inf/ctrl)								
TRANSCRIPTION GROUP 1	**0 h**	**1 h**	**2 h**	**3 h**	**4 h**	**5 h**	**6 h**	**7 h**	**8 h**
PMM0549 – *csoS1* carboxysome shell protein 1, carbon fixation	1.70	**1.31****	-1.22	-1.59	-1.76	-2.07	-2.03	-2.16	-2.58
PMM0550 – *rbcL* rubisco large subunit, carbon fixation	1.79	**2.01*****	**1.38****	**1.18***	-1.41	-1.69	-1.62	-2.00	-2.05
PMM0551 – *rbcS* rubisco small subunit, carbon fixation	1.63	**1.85*****	**1.36****	**1.17***	-1.43	-1.67	-1.62	-2.04	-2.01
#PMM0815/PMM1396 – *hli19/09* high-light inducible stress response protein	1.15	**1.23***	-1.27	-1.45	-1.25	-1.38	-1.45	-1.58	-1.39
#PMM0816/PMM1397 – *hli18/08* high-light inducible stress response protein	1.09	**1.29****	-1.12	-1.26	-1.26	-1.50	-1.56	-1.64	-1.56
#PMM0817/PMM1398 – *hli17/07* high-light inducible stress response protein	1.16	**1.35*****	-1.17	-1.19	-1.24	-1.47	-1.46	-1.64	-1.60
#PMM0818/PMM1399 – *hli16/06* high-light inducible stress response protein	1.14	**1.33****	-1.18	-1.24	-1.28	-1.53	-1.49	-1.66	-1.70
PMM0970 – *urtA* urea ABC transporter periplasmic binding protein	1.34	**1.43*****	1.01	-1.24	-1.51	-1.77	-1.59	-1.75	-1.52
PMM1135 – *hli14* high-light inducible stress response protein	1.24	**1.26****	-1.17	-1.26	-1.54	-1.76	-1.73	-1.88	-1.94
PMM1483 – *rpoC2* RNA polymerase subunit, transcription	1.01	**1.26***	-1.02	-1.28	-1.49	-1.59	-1.61	-1.67	-1.67
PMM1536 – *rps11* ribosome small subunit protein 11, translation	1.35	**1.19***	-1.09	-1.35	-1.80	-1.94	-1.95	-2.05	-2.09
PMM1544 – *rpl6* ribosome large subunit protein 6, translation	1.02	**1.23***	-1.05	-1.44	-1.70	-1.74	-2.04	-1.86	-1.97
PMM1545 – *rps8* ribosome small subunit protein 8, translation	1.07	**1.21***	-1.19	-1.39	-1.68	-1.86	-1.93	-1.85	-1.90
PMM1546 – *rpl5* ribosome large subunit protein 5, translation	-1.06	**1.33****	-1.09	-1.33	-1.82	-1.94	-2.22	-1.94	-1.94
PMM1549 – *rps17* ribosome small subunit protein 17, translation	-1.13	**1.22***	-1.17	-1.40	-1.97	-2.09	-2.15	-2.10	-1.98
PMM1629 – *rpoD type II* alternative sigma factor, transcription	1.10	**1.61*****	-1.09	-1.34	-1.38	-1.74	-1.80	-1.84	-1.92
TRANSCRIPTION GROUP 2	**0 h**	**1 h**	**2 h**	**3 h**	**4 h**	**5 h**	**6 h**	**7 h**	**8 h**
PMM0014 – *dus* tRNA dihydrouridine synthase, RNA modification	1.20	-1.28	**1.23***	**1.21***	**1.44***	**1.48****	1.06	**1.49****	1.21
PMM0030 – unknown	-1.10	-1.15	1.09	**1.66****	**1.41***	1.08	-1.41	1.00	1.00
PMM0334 – unknown	1.11	-1.49	-1.07	1.05	1.20	**1.26***	-1.11	1.04	-1.17
#PMM0368 – unknown	1.02	-1.05	**1.21***	**1.70*****	**1.85****	**1.61*****	**1.68*****	**1.70*****	1.32
PMM0426 – *sun* tRNA and rRNA methyltransferase, RNA modification	1.09	-1.66	-1.48	-1.20	**1.48***	**1.41****	1.23	**1.81*****	**1.49***
#PMM0684 – unknown (homologous to PMM0819 and PMM1134)	1.01	-1.06	**1.27***	**1.58*****	1.35	1.16	1.08	1.07	-1.16
#PMM0685 – unknown (homologous to PMM1427)	-1.05	-1.36	**1.77*****	**2.65*****	**2.37****	**2.01*****	**1.44***	**1.49***	**1.39***
#PMM0686 – *clpS-like* protease adaptor, protease inhibition and redirection	-1.07	-1.48	**3.91*****	**8.95*****	**14.00*****	**11.71*****	**8.55*****	**10.82*****	**8.75*****
#PMM0819 – unknown (homologous to PMM0684 and PMM1134)	1.28	-1.01	**1.52*****	**2.16*****	**2.32****	**1.84*****	**1.66*****	**1.57*****	1.21
PMM0830 – *DHPS-like* folate biosynthesis, nucleotide & amino acid synthesis	1.16	-1.84	-1.69	-1.38	**1.38***	**1.36****	1.10	1.18	-1.12

PMM0936 – *umuD* SOS response to DNA damage	1.01	-1.55	**1.31***	**1.39*****	**1.75****	**1.53*****	1.15	**1.43****	1.10
PMM1114 – unknown	1.26	-1.81	-1.05	-1.18	1.22	**1.29***	1.05	1.04	1.02
PMM1115 – *crtH*, phytoene dehydrogenase, secondary metabolite biosynthesis	1.01	-1.42	-1.15	-1.07	1.26	**1.17***	-1.08	1.00	-1.14
PMM1187 – AAA ATPase family, protein turnover, stress response	1.01	-1.39	1.09	**1.21***	**1.56***	**1.45****	1.00	**1.27***	1.15
#PMM1201 – dTDP-D-glucose 4,6-dehydratase, cell envelope biogenesis	1.27	-1.63	-1.43	-1.16	1.22	**1.27****	1.02	1.17	-1.08
#PMM1248 – unknown	1.08	-2.27	-1.76	-1.43	**1.37***	**1.60*****	1.12	**1.33***	1.25
PMM1284 – *phoH-like* phosphate stress ATPase	1.13	-1.34	1.06	1.11	**1.80****	**1.56*****	1.12	**1.26***	1.07
#PMM1403 – HNH nuclease domain, site-specific endonuclease	1.11	-1.88	-1.56	-1.60	1.24	**1.45****	1.11	**1.56****	**1.51***
#PMM1426 – unknown	1.47	-2.50	-2.05	-1.82	1.16	**1.49****	1.18	**1.31***	1.36
#PMM1427 – unknown (homologous to PMM0685)	1.00	-1.20	**1.23***	**1.71*****	**1.93****	**1.74****	**1.52****	**1.54****	1.35
PMM1428 – unknown	1.02	-1.27	-1.03	1.03	**1.60****	**1.50****	1.16	**1.32***	1.04
PMM1501 – *rne* RNase E, mRNA degradation	1.01	-1.16	**2.44*****	**3.51*****	**3.43*****	**2.91*****	**1.83*****	**2.09*****	**1.72***
PMM1502 – *rnhB* RNase HII, DNA replication and repair	1.13	-2.16	**1.47***	**2.25*****	**3.55*****	**2.91*****	**1.64****	**2.52*****	**2.09****
PMM1517 – unknown	-1.06	-1.58	-1.32	1.03	**1.39***	**1.31***	1.10	1.06	1.15
PMM1529 – *prfA* peptide release factor, translation	1.31	-1.11	1.10	**1.52*****	**1.67****	**1.29****	-1.04	1.17	-1.10

#Genes found in genome islands as per Coleman et al.[6].

228

Supplementary Table 4: High-light inducible genes (*hli*) in *Prochlorococcus* MED4 and their expression patterns during exposure to environmental stressors.

MED4 Gene IDs	High Light Stress	Nitrogen Stress	Phage Infection
PMM0093 - *hli01*			
PMM0064 - *hli02*			
PMM1482 - *hli03*			
*PMM1118 - *hli04*	+		
*#PMM1404 - *hli05*	+		
*#PMM0818/PMM1399 - *hli06/hli16*	+		+
*#PMM0817/PMM1398 - *hli07/hli17*	+		+
*#PMM0816/PMM1397 - *hli08/hli18*	+		+
*#PMM0815/PMM1396 - *hli09/hli19*	+		+
*#PMM1390 - *hli10*		+	
*#PMM1385 - *hli11*	+		
*#PMM1384 - *hli12*	+		
PMM1317 - *hli13*			
*PMM1135 - *hli14*	+		+
*#PMM1128 - *hli15*	+	+	
PMM0471 - *hli20*			
*#PMM0690 - *hli21*	+	+	
*#PMM0689 - *hli22*	+	+	

*Clusters with phage *hli* genes as per Lindell & Sullivan et al.[5]; #Found in genome islands as per Coleman et al.[6]. High-light stress[46], Nitrogen stress[23], Phage infection (this study).

Supplementary Table 5: Comparison of array normalization methods to RT-PCR. Significance is assigned to differentially expressed genes determined by RT-PCR (p-values) normalized to *rnpB* and microarray analysis (q-values) normalized using different methods (Quantile RMA; Hyb Ctrl;.Golden Spike (GS) regular[32]; GS without 2nd Loess)

Expression Pattern	ORF *gene*	Time	RT-PCR	RMA	Hyb Ctrl	GS w/ 2nd loess (regular)	GS w/o 2nd Loess
Unchanged	PMM_rnpB *rnpB*	0	0.528	0.912	0.759	0.991	0.978
		1	0.530	0.818	0.674	0.419	0.795
		3	0.938	0.805	0.158	0.614	0.550
		4	0.878	0.489	0.712	0.884	0.986
		8	0.970	0.756	0.281	0.627	0.903
DownRegulated	PMM0496 *rpoD*	0	0.171	0.529	0.231	0.995	0.998
		4	0.000	0.003	0.470	0.539	0.123
		8	0.001	0.003	0.043	0.966	0.100
	PMM0627 *pcb*	0	0.422	0.337	0.150	0.655	0.769
		4	0.004	0.001	0.449	0.572	0.208
		8	0.000	0.000	0.007	0.152	0.048
	PMM1309 *ftsZ*	0	0.034	0.405	0.072	0.813	0.803
		4	0.000	0.008	0.602	0.703	0.113
		8	0.000	0.000	0.003	0.453	0.011
	PMM1629 *rpoD type II* (up- and down-regulated)	0	0.195	0.231	0.123	0.889	0.895
		1	0.388	0.000 up	0.636	0.000 up	0.044
		3	0.000 dn	0.001 dn	0.005	0.457	0.057
		8	0.000 dn	0.001 dn	0.054	0.991	0.042
UpRegulated	PMM0550 *rbcL* (up- and down-regulated)	0	0.041 up	0.974	0.000 up	0.022 up	0.016 up
		1	0.066	0.000 up	0.283	0.000 up	0.000 up
		3	0.317	0.010 up	0.021	0.000 up	0.170
		8	0.269	0.000 dn	0.051 dn	0.809	0.050
	PMM0684 unknown	0	0.095	0.999	0.200	0.992	0.989
		3	0.049	0.000	0.000	0.000	0.005
		4	0.001	0.088	0.223	0.000	0.171
		8	0.075	0.269	0.154	0.000	0.702
	PMM0686 *clpS-like*	0	0.467	0.337	0.231	0.801	0.624
		4	0.006	0.000	0.000	0.000	0.000
		8	0.000	0.000	0.000	0.000	0.000
	PMM0819 unknown	0	0.599	0.719	0.101	0.373	0.476
		4	0.000	0.002	0.035	0.000	0.003
		8	0.013	0.164	0.014	0.000	0.230
	PMM0936 *umuD*	0	0.439	0.193	0.416	0.889	0.921
		1	0.276	0.000 dn	0.041	0.251	0.038 dn
		3	0.001	0.000	0.005	0.006	0.059
		4	0.005	0.005	0.101	0.003	0.075
		8	0.001	0.497	0.350	0.459	0.843
	PMM1284 *phoH-like*	0	0.002	0.486	0.105	0.707	0.765
		4	0.002	0.002	0.037	0.003	0.111
		8	0.738	0.569	0.336	0.256	0.754
	PMM1501 *rne*	0	0.451	0.297	0.403	0.993	0.988
		3	0.025	0.000	0.000	0.000	0.000
		8	0.026	0.019	0.036	0.027	0.160

p determined from 2-tailed t-test for RT-PCR results and q determined using Cyber-T and Q-value for microarray results.

Supplementary Table 6: Mass spectrometric detection of previously unannotated proteins from the P-SSP7 phage genome. Detected tryptic peptides are in bold and underlined. The entire ORF inferred from these peptide detections are shown. For PSSP7_033, detection of the peptide suggests an N-terminal extension of the previously annotated protein (shown in italics).

GENE ID	PROTEIN SEQUENCE
PSSP7_19A	MTTRKK**NQSFGPPPPITK**LTTEQDFKLRQLEILLSKPETRKEDIAIVMIALQE QAFVLSNCIKNLIEKWPKPPTTTDPR**TTNEVPLMFGILLETK**DSDFTSET
PSSP7_20A	MKYLGEVVRTVTVPAFYTLILITPILLTSCKSKDKNSHGLNDVWTSPENSGLI QKLEQR**KQLYKELLGETSGSTK**
PSSP7_20B	**METESIQTSVLR**FTCPHAERASYSTLICQPVVSATYEKVCVKVCQICASSIV GQGLKNLESILHQISTGKLDSDS
PSSP7_033	MCLGAAAKAANENARRRYKYENERRER***NWMQTMSIYNAQK****VKYDEDVQ NAGLAQAQVKTDQQEAMDLARGEAQIKYAELFRKLLNDSTYGKLVASGQT GQSTRRRATMDYAKYGRDVSDIARRLTLNDRELARKSSEQISKYKQFKDE AFAKVAFQPIPDVAPPQPVMRNVGAEAFMGALSIASNVATMGGQSGFGW WGG

Supplementary Table 7: Primers used for RT-PCR verification of microarray results and normalization methods for representative phage and host genes.

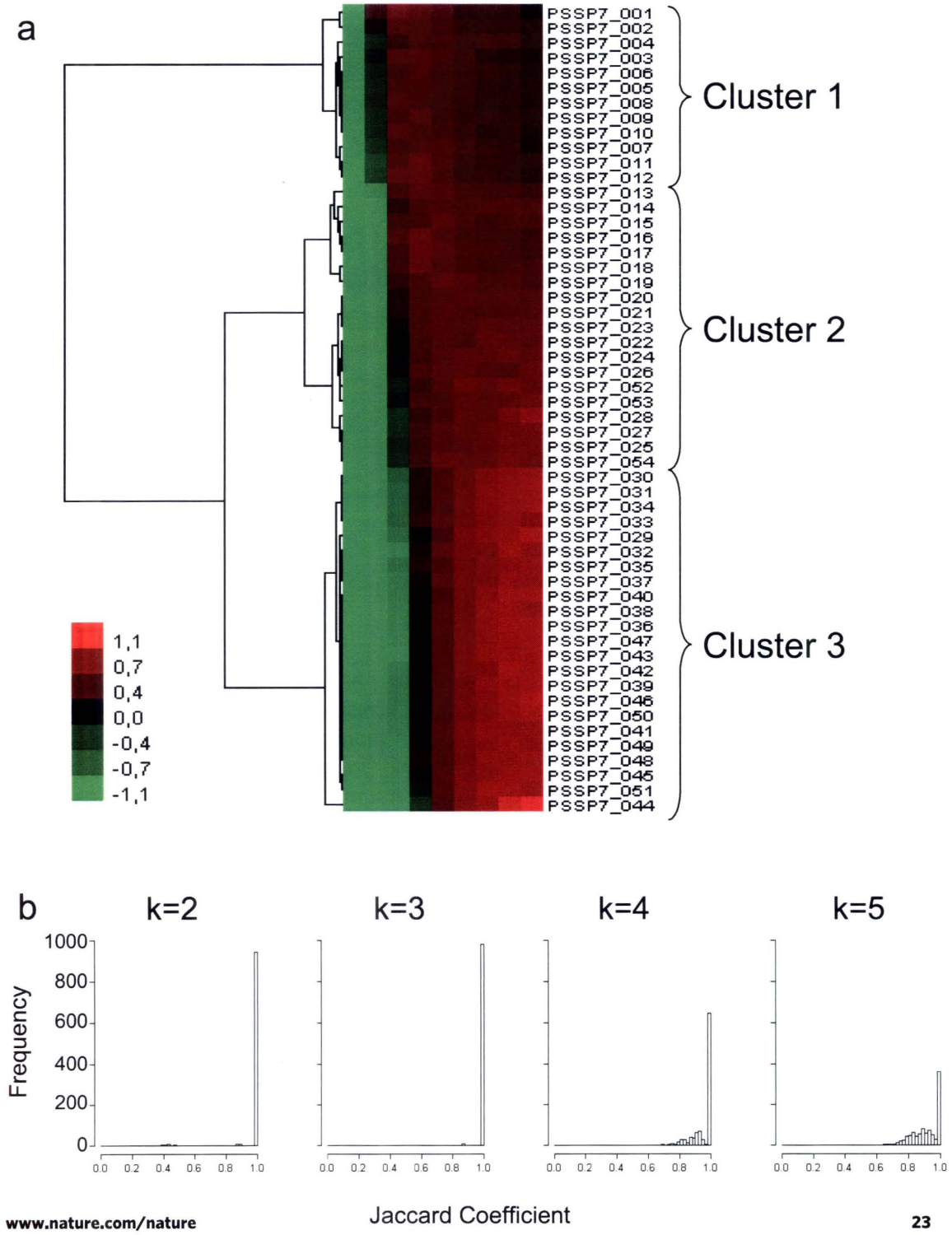
Target ORF ***gene* – product**	**Primer Direction**	**Primer Sequence 5' – 3'**
Phage P-SSP7 genes		
PSSP7_001 Unknown	F R	CCAAGCCAAAGGCTACACAT GCATCCCTTGATTCATTGCT
PSSP7_003 Unknown	F R	ATGGTTCACTTCCTAACCAAGC CCCCCTTACCCATAGGTGTT
PSSP7_013 *gene 1* – RNA polymerase	F R	CGACTATGGAGGAGCGGTTA GTCTGCTGCTTCCCAATCTC
PSSP7_017 *gene 5* – DNA polymerase	F R	AAACACTTCCGCCCTTACCT CTGCAACGAAAGGGAATTGT
PSSP7_020 *nrd* – ribonucleotide reductase	F R	TTGTGCAAGCTCCATAGTCG GCCTTACCAAACTCGGCATA
PSSP7_027 *psbA* – D1 photosystem II protein	F R	CTCTGCTATGCACGGAAGTT GCAGATTCCCATGGAGGTAA
PSSP7_029 *gene 10* – capsid protein	F R	GGCTTCCAGCATGAAACAAT TGGTCTTCTCTGCAACTGGA
PSSP7_054 *talC* – transaldolase family protein	F R	TGGTCGAAAATACGGAGAGG TACGTAGCACCAGCATGAGC
Host *Prochlorococcus* MED4 genes		
PMM_rnpB *rnpB* – RNA of RNase P	F R	TTGAGGAAAGTCCGGGCTC GCGGTATGTTTCTGTGGCACT
PMM0496 *rpoD* – principle RNA polymerase sigma factor	F R	AATCAGAGCTGCCGAAAATA TGATCTGCTATCGCTCGTGT
PMM0550 *rbcL* – rubisco large subunit	F R	CCTGAATATGTCCCCCTCGA CCGCTGCTGCAACTTCTTCT
PMM0627 *pcb* – chlorophyll a/b binding protein	F R	TCATGTCGCTCATGCAGGG GACCCATTGGGACACTGGG
PMM0684 Unknown	F R	CGCAAGGCAGCTTTTTAATC TCCATGTTTCAAACGCAGAG
PMM0686 *clpS-like* – protease adaptor	F R	CAGTTGTAGATCCAAAGACAACG CAAGACAATTTGCTACGTGTTCA
PMM0819 Unknown	F R	CCCAAGTGGTTGGCTTCTTA ATCCCAGGCTTTTTCCAAAT
PMM0936 *umuD* – SOS response to DNA damage	F R	GTGATTCGGTCTCAGCAGGT TTCTCCATCTATCATCGCAATA
PMM1284 *phoH-like* – phosphate stress induced ATPase	F R	GTTTGTGCCGCCAGATTATT TGCTAATGGTGCGACTTCAA
PMM1309 *ftsZ* – cell division protein	F R	AATGACTGAAGCTGGCACTGC ACTATTCATTGCGGCTTGAGC
PMM1501 *rne* – RNase E	F R	AACCGCCTAGCACAGGATTA TGCTTTTTCGAGAGCGATTT
PMM1629 *rpoD type II* – alternative sigma factor	F R	GAGTTGCCCGAAGATGATGT ACATTGGCTCATCTCCATCC

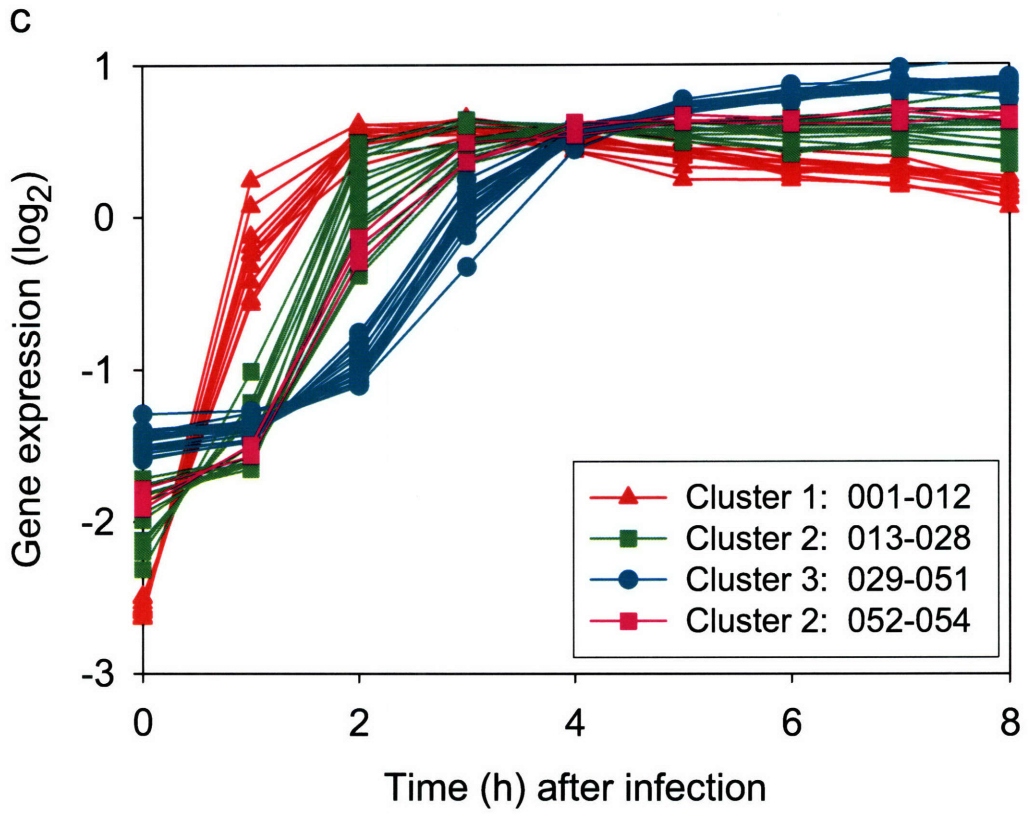
Supplementary Table. 8. Primers used in 5' RACE analysis for phage P-SSP7 and host *Prochlorococcus* MED4 genes. Primers used for reverse transcription are designated by "rt", whereas those used for nested or second nested PCR are designated by "nest" and "nest2" respectively in the oligonucleotide name. "up" – the primer was designed upstream of the gene.

Gene	Oligo. Name	Oligonucleotide Sequence (5'–3')	Length
Phage P-SSP7 genes			
PSSP7_001	P001rtREV	TCTCCATAATTGACCCGCTT	20
	P001nestREV	CTTAATAAAGTCAGTCAGTCCATCCCAGTCAGT	33
	P001nest2rev	TGATGGGAGAGGAATTGAACCTCTCGAT	28
PSSP7_003	P003nestREV	CCTTCTTACCGAATTTTCCTATGGTGTACTTAG	33
	P003rtREV	TCTCCCCCTTACCCATAG	18
PSSP7_004	P004nestREV	GGAATCGTTTACAGGGAATAACTCAGGCTCG	31
	P004rtREV	TTGTGCTTGGGTTTCTCTCA	20
PSSP7_008	P008nestREV	CGAAGGCTTCATAAGGCATCGAAGGAAAG	29
	P008rtREV	GGTATCTGATAGCCATAAATCTT	23
PSSP7_011	P011nestREV	CAGTCAGTATTACGGCTACCACTTGCGC	28
	P011rtREV	TCAATCTATGAAACTCACTGAG	22
PSSP7_013	P013rtREV	TCTTTACGTGGGGAGAATAG	20
	P013nestREV	GGTTATCTTTGCAGTTATAGCTCCTTGCGATTCTG	35
PSSP7_014	P014nestREV	CTATAAACTTACCTTCTTCTACCTCCTCCCATG	33
	P014rtREV	CTGGAGGACGCTTATCTTC	19
PSSP7_017	P017nestREV	CAGCATGGGTGAGCCAATGCAGAGC	25
	P017rtREV	ACAGGTAAATCGTAGCCAATAAT	23
PSSP7_019	P019nestREV	CTTAAGTTCCTGTATGACACGTTTGTATCCAC	32
	P019rtREV	CATCTGCTTCTAGAGTATCTC	21
PSSP7_020	P020rtREV	GTCCTGATGCAACGAGAG	18
	P020nestREV	CCCTTATTTGTTCTTGTTCCCGCCGGTC	28
PSSP7_020 up	P020nest2REV	CCGAGGTGAAATCCGAGTCCTTGGTC	26
	P020rtREV	ACCCTGCTCTGCATGTGT	18
PSSP7_024	P024nestREV	GGAGGTAGAACTGCGAGCATGAGCTTC	27
	P024rtREV	CCTAACTTGTCATCACGTAC	20
PSSP7_026	P027nest2REV	CAGCCATTAAATCTTTCTGCTTCTGGTGACATTAG	35
PSSP7_027	P027nestREV	CCTATTGCATTGGAGCTTGGAACTACTGC	29
	P027rtREV	CGGCTTCCCAGATCGG	16
PSSP7_029 up	P029nest2REV	GGACTCGCACCTCCCTGATCGG	22
PSSP7_029	P029rtREV	GCCTTGTCAGCATTACCC	18
	P029nestREV	GGAAGGAACTTGTCATACGACCTGTGTAG	29
PSSP7_030	P030nestREV	GCCATCCTTCACCCTGTACATCTTTGTTTG	30
	P030rtREV	TCTGGGGTAACAAGTACATG	20
PSSP7_032	P032nestREV	CATCTTCCTGTACGCCGGCTACTCC	25
	P032rtREV	CGGGTGTACATAGCATCC	18
	P032nest2rev	CTCCATAAGCGGCACATAGTGGGATTACC	29
PSSP7_036	P036nestREV	GTCGAGTTCAACTTTGACATCGACATTCGC	30
	P036rtREV	GGTGTAGTCATTATTTGTTTGAC	23
PSSP7_050	P050nestREV	GTCCATTATTTCAGGGTTAGCTTTTTGTGCAGG	33
	P050rtREV	ACCTTTCCATCTCATTTTAGAGA	23
PSSP7_051	P051nestREV	GGCTTGAATCTGGAGTCTCTTGGGTCC	27
	P051rtREV	CCAAGATTTACCAACACCTC	20
PSSP7_052	P052rtREV	CCTTTGCGTTAAAGATGCTG	20
	P052nestREV	CTTCCATCTCATCCAGGTAGTCCTCAATAGC	31
Host MED4 genes			
PMM0368	PMM0368nestREV	CTATTGCCCAACCATTAGCTCCCTGAGG	28
	PMM0368rtREV	GCTGACACCACTGCCAAA	18
PMM0684	PMM0684nestREV	CTGCCTTGCGGGTTAAGTATCCAGCC	26
	PMM0684rtREV	TTGTAAATAAGAGATGGGTATAAAC	25
PMM0819	PMM0819nestREV	TGACTTGATGACTTGATTGCTGAGGGCTC	29
	PMM0819rtREV	CATTTATCCCAGGCTTTTTCC	21
PMM1500	PMM1500nestREV	CCATCCATATCCTTGCTCTGGGCCG	25
	PMM1500rtREV	GATGAGATTTGACACCCTC	19

PMM1501	PMM1501nestREV	CTATAAAGGCAGCATCAATACCTGGTAGGAC	31
	PMM1501rtREV	GGACCTAGATCTGATACATG	20

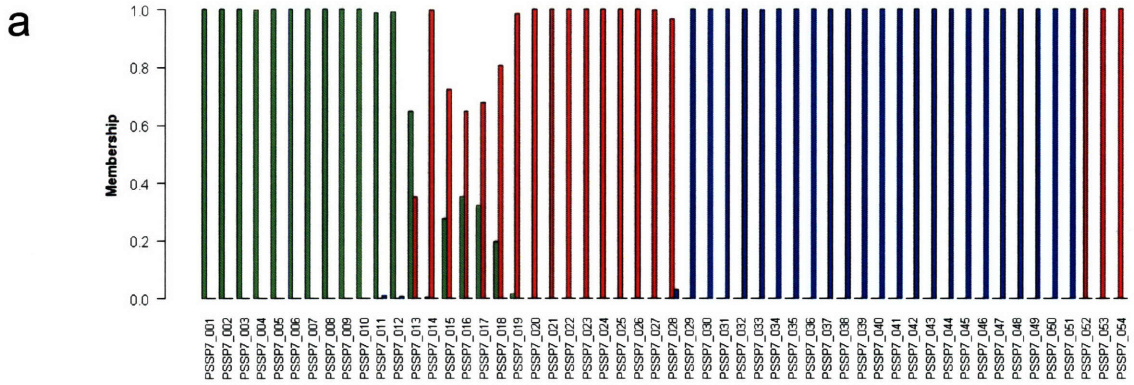
Supplementary Figure 1

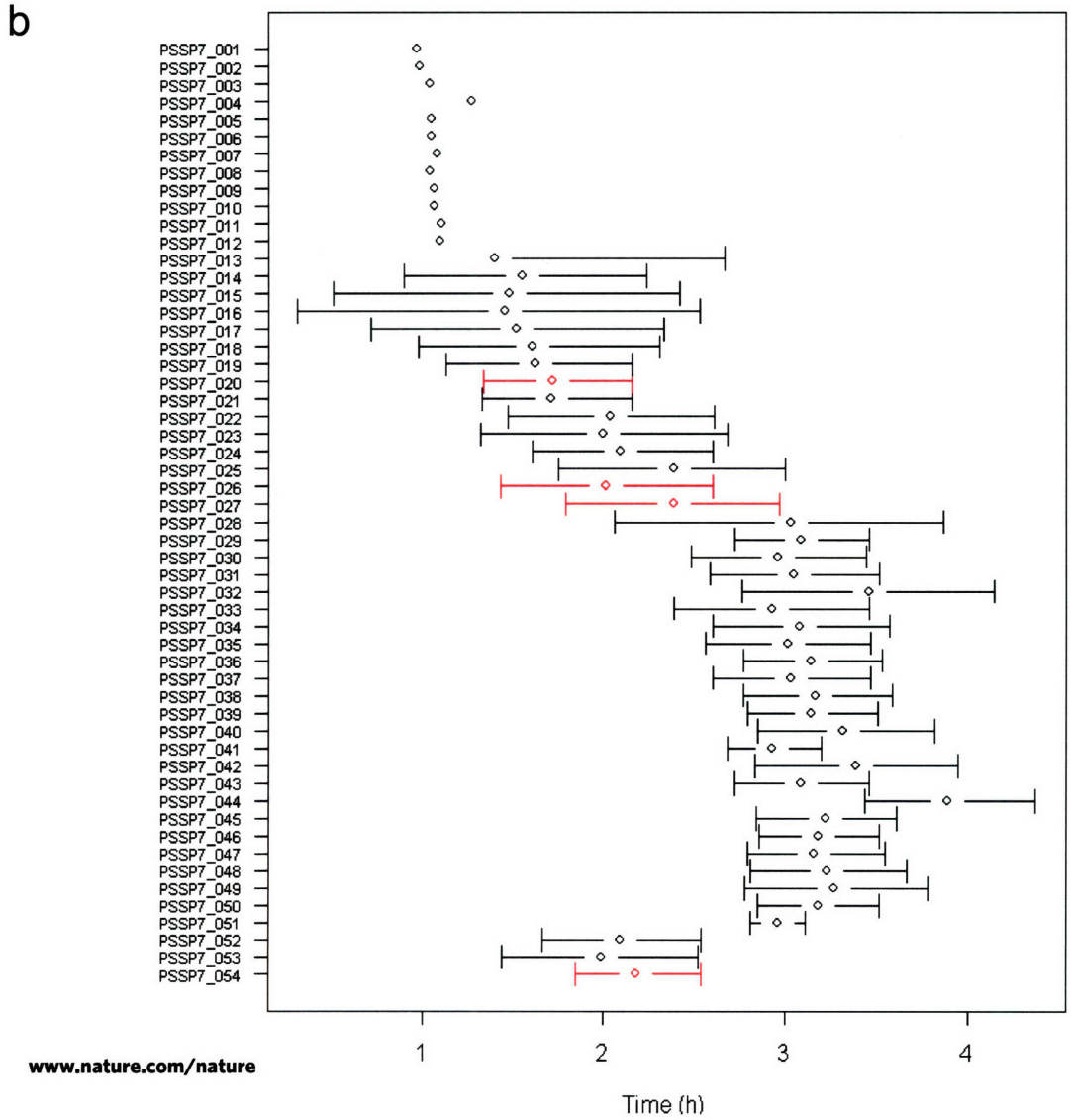




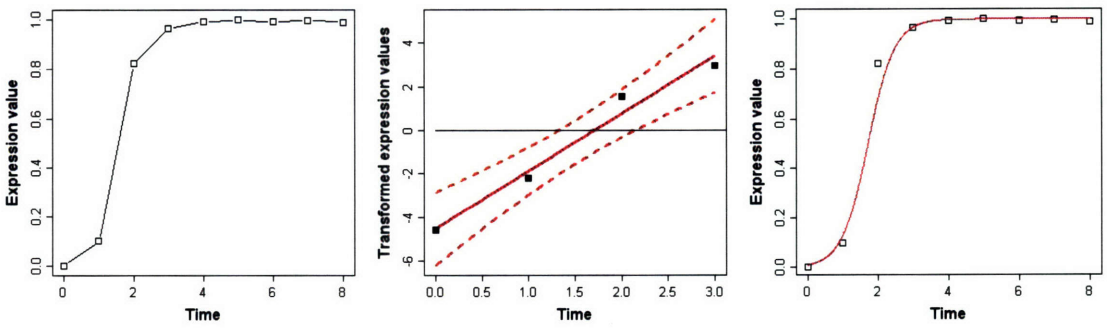
Supplementary Figure 1. Cluster analysis of phage gene expression profiles. (a) Hierarchical clustering of phage gene expression profiles, after standardization of logged data (mean expression equal to zero and standard deviation equal to one) was performed with average linkage and Pearson correlation. (b) Distribution of Jaccard coefficients derived from 1000 random independent resamplings of phage genes. The proportion of genes used for resampling was 0.7. The number of clusters (k) tested ranged from 2 to 5. Average linkage and Pearson correlation was used for the hierarchical re-clustering. For k=3 clusters, 982 out of 1000 Jaccard coefficients equaled 1 indicating that phage genes form three stable clusters. (c) Temporal profiles of the 3 clusters detected by hierarchical clustering. See Figure 2a for a representation of these temporal profiles after minimum-maximum normalization. Note that the last 3 genes in the genome cluster together with genes from cluster 2 and are transcribed prior to genes in cluster 3. See Suppl. Fig. 2 for statistical analysis of the significance of the clustering of all 4 'bacterial-like' genes in cluster 2. See Suppl. Table 1 for gene name and function for each ORF.

Supplementary Figure 2

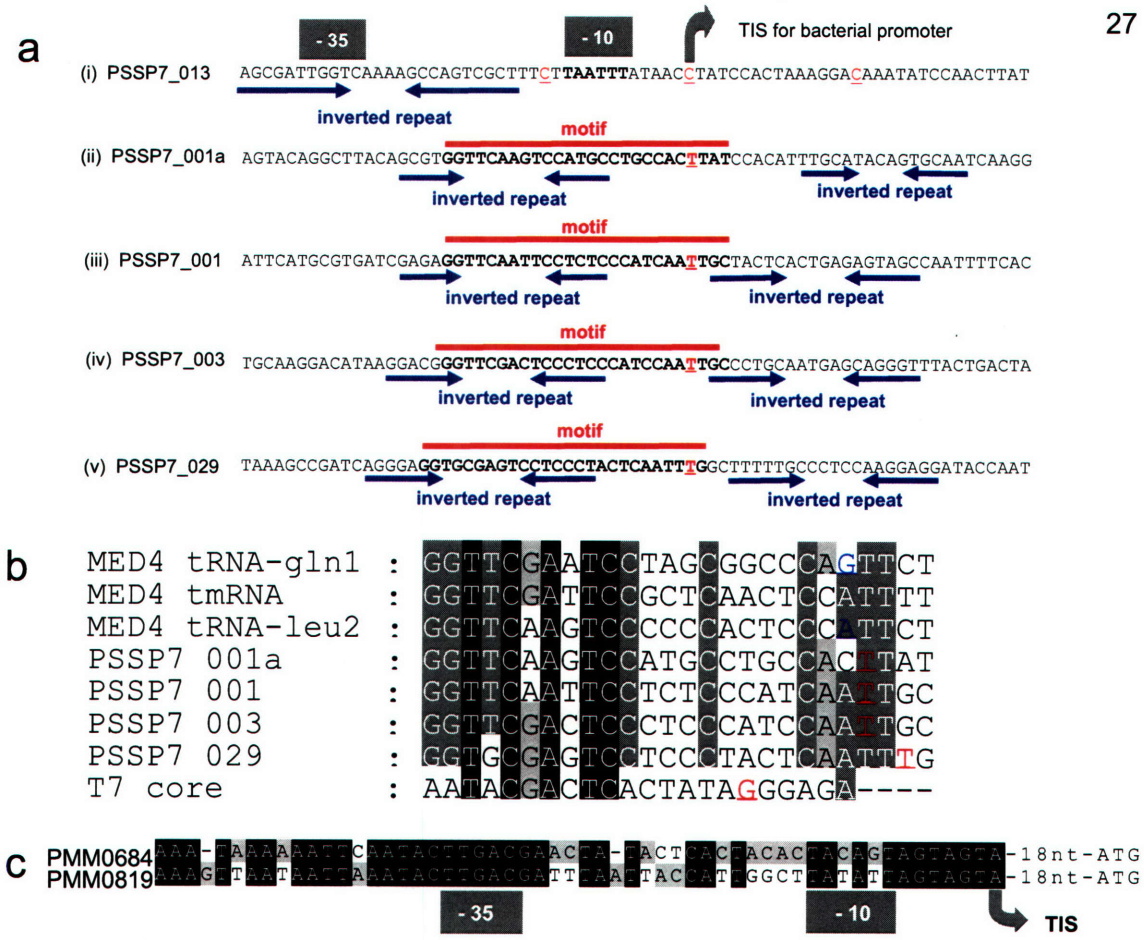




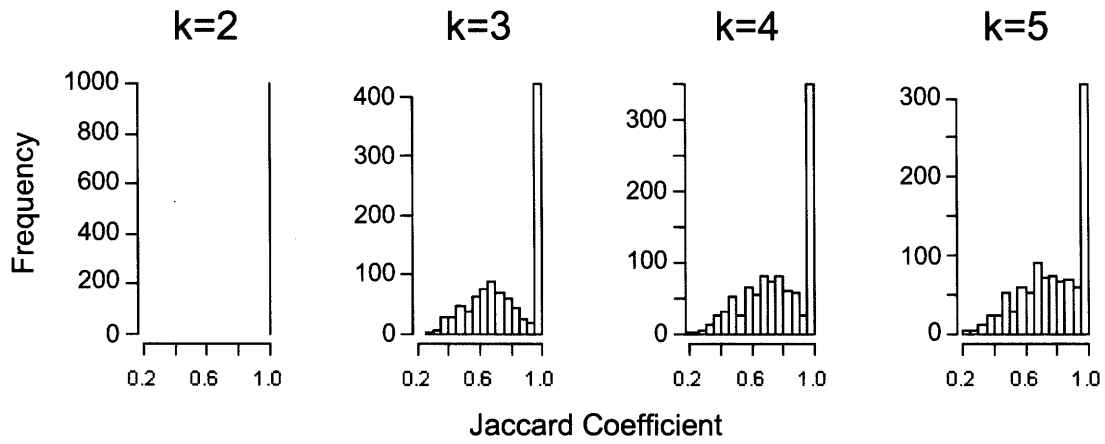
C



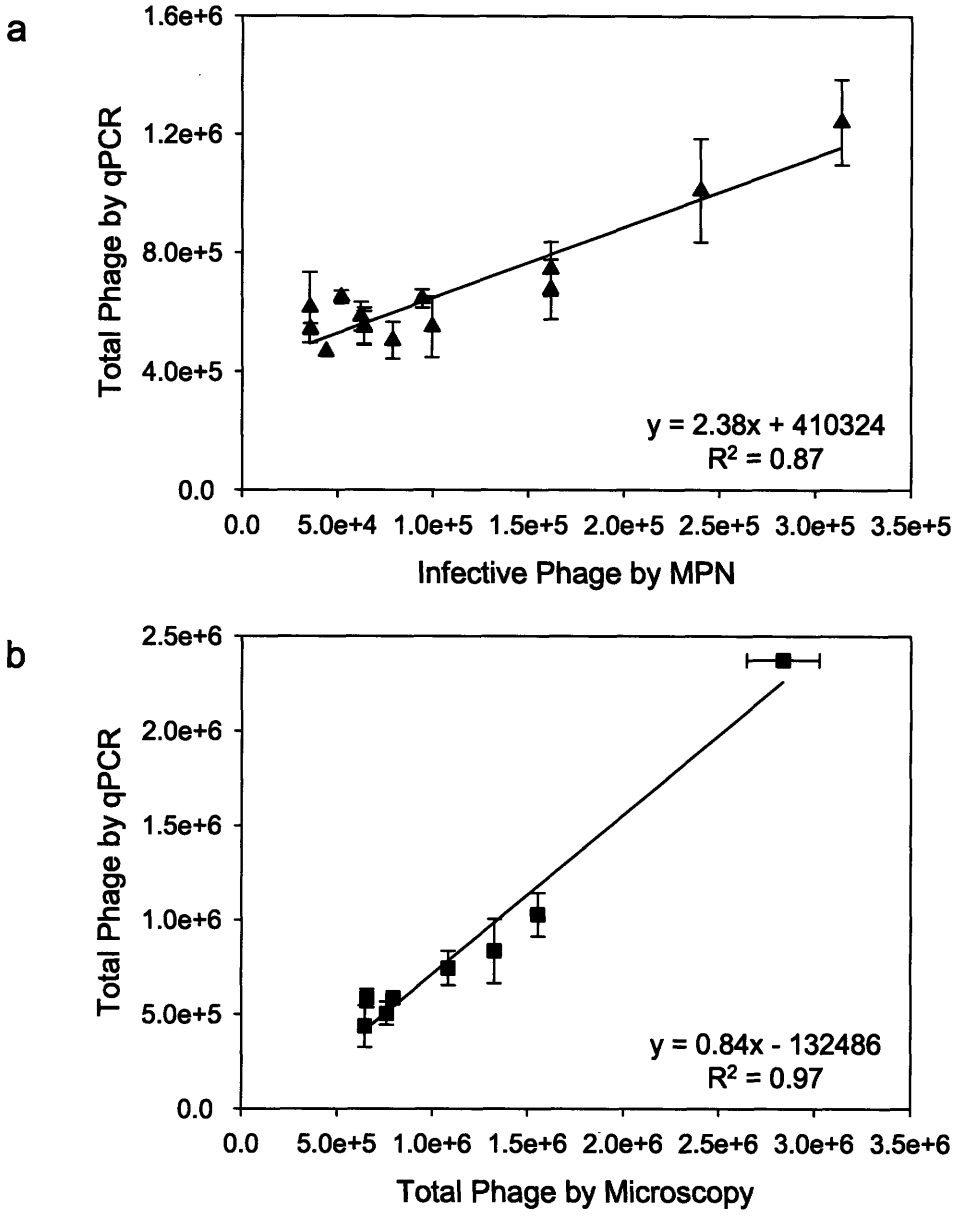
Supplementary Figure 2. Significance of the temporal coexpression of the last 3 genes of the genome with genes in cluster 2 and therefore of the 4 'bacterial-like' genes *nrd* (020), *hli* (026), *psbA* (027) and *talC* (054). (a) Cluster membership for genes in cluster 1 (green), cluster 2 (red) and cluster 3 (blue) were based on 10000 independent bootstrap samplings with replacement of expression values for the same gene. (b) 90% confidence intervals are shown for the switch time (t*) at which transcription of the phage genes went from being non-expressed to expressed – defined here as the time point at which 50% of maximal expression was reached. Confidence intervals for the 4 'bacterial-like' genes are shown in red. Note that no intervals could be derived for ORFs 1-12 due to immediate initiation of expression. (c) An example of the fitting procedure carried out (for *nrd*) to determine the confidence intervals shown in (b). The left panel shows the expression data (y) after normalization so that minimum expression equals zero and maximal expression equals 1. The middle panel shows the linear regression (solid red line) and the confidence intervals (dashed red lines) determined after the transformation y'=log(y/(1-y)). Note time t* is derived as the time point for which the regression line crosses y'=0 (set at half maximum expression). The regressed sigmoidal curve is shown in the right panel. The reliable assignment of the last 3 genes on the genome with genes in expression cluster 2 as well as the overlap of their confidence intervals provides strong evidence for the temporal coexpression of the 4 'bacterial-like' genes in cluster 2 despite their spatial separation on the genome.



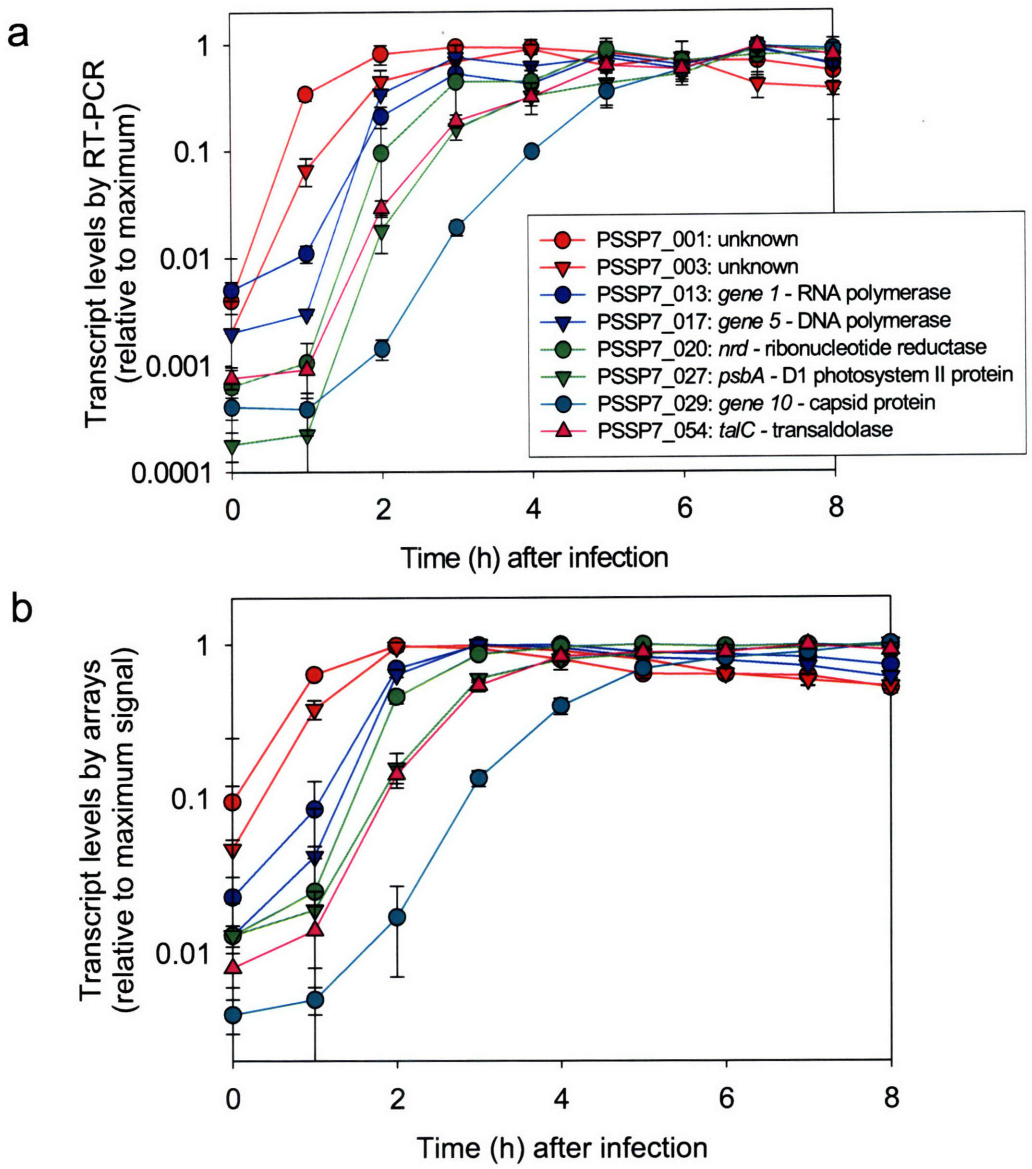
Supplementary Figure 3. RACE mapping of 5' transcript ends for phage and host genes and associated possible regulatory elements. (a) Experimentally mapped 5' ends for phage genes. A typical bacterial promoter was found upstream of the 5'ends of the cluster 2 gene PSSP7_013 coding the RNA polymerase (i). The 5' ends upstream of cluster 1 and 3 genes – ORF PSSP7_001 (2 transcripts), PSSP7_003 and PSSP7_029 – were found in both TAP+ and TAP- treatments (data not shown) suggesting that they are mature transcripts that have undergone post-transcriptional processing (ii to v). The nt at the 5' ends are shown in red and underlined. A 26 nt conserved motif (bold type face and marked with a red line above the sequence) was found in the region of these processed 5' ends, as were inverted repeats (blue arrows below the sequence). (b) Comparison of the conserved motif found upstream of processed transcripts in (a) shows that they have weak similarity to the T7 core promoter. A bioinformatic search for sequence similarity between this motif and the MED4 host genome revealed similarity to putative 3' cleavage sites of 2 tRNA genes and tmRNA, suggesting that the inverted repeats associated with this motif may serve as an RNase III recognition site in a similar fashion to that known for T7 with linked promoter and processing sites[47]. The red underlined nucleotide indicates the 5' end determined by RACE and for the T7 core promoter it indicates the known transcript start. The annotated 3' end of the tRNAs is blue and underlined. (c) Transcription initiation sites for two up-regulated MED4 homologous genes of unknown function (PMM0684 and PMM0819). The consensus bacterial regulatory elements (the -10 and -35 box) are shown upstream of the transcription initiation site.



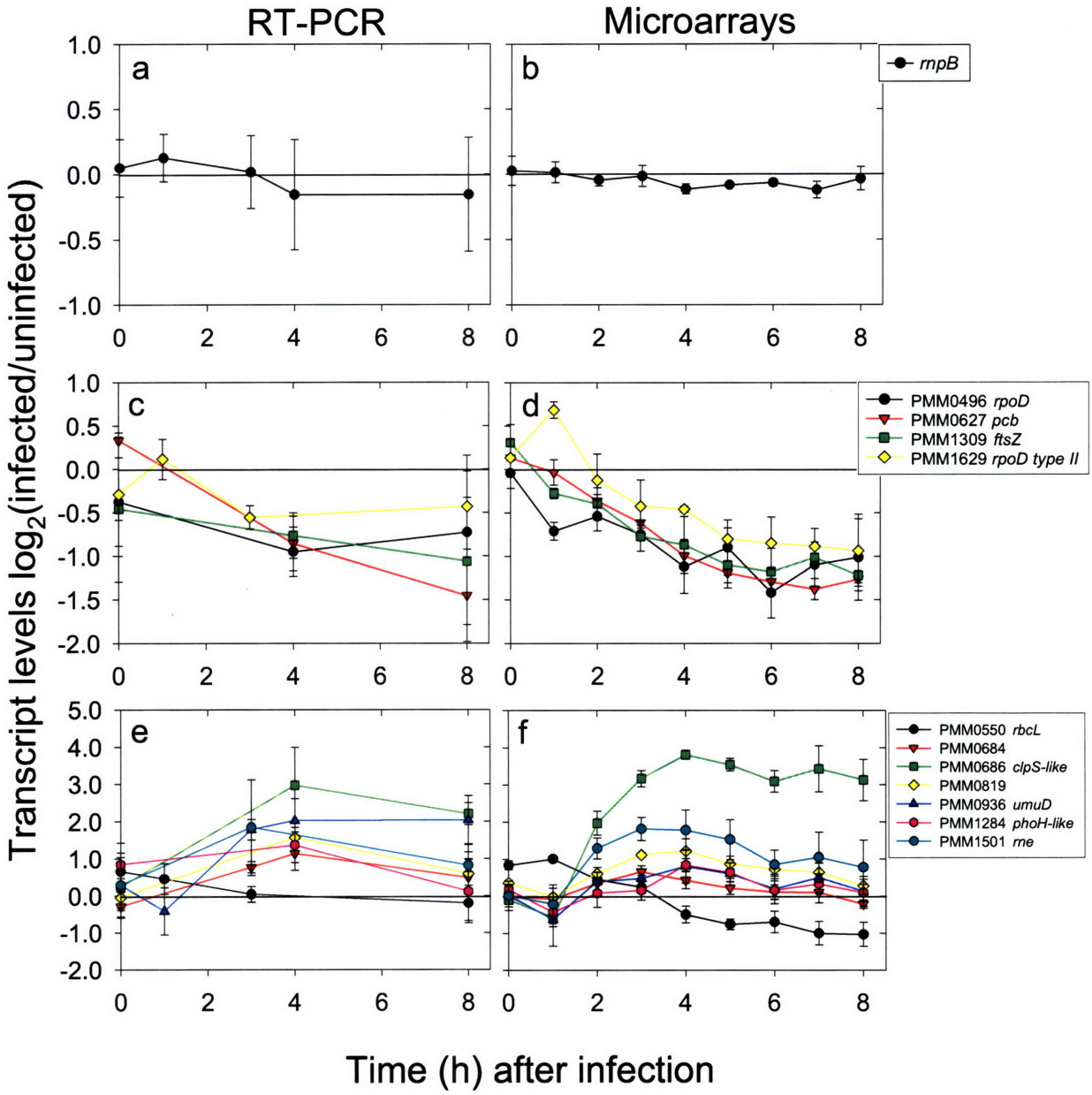
Supplementary Figure 4. Analysis of the number of stable clusters of upregulated host (MED4) genes. The distribution of Jaccard coefficents was derived from 1000 independent random samplings of upregulated host genes. The number of clusters (k) tested ranged from 2 to 5. The proportion of genes used for resampling was 0.7. For hierarchical re-clustering average linkage and Pearson correlation was used. For k=2 clusters the coefficients are concentrated at 1 indicating that upregulated host genes form two stable clusters.



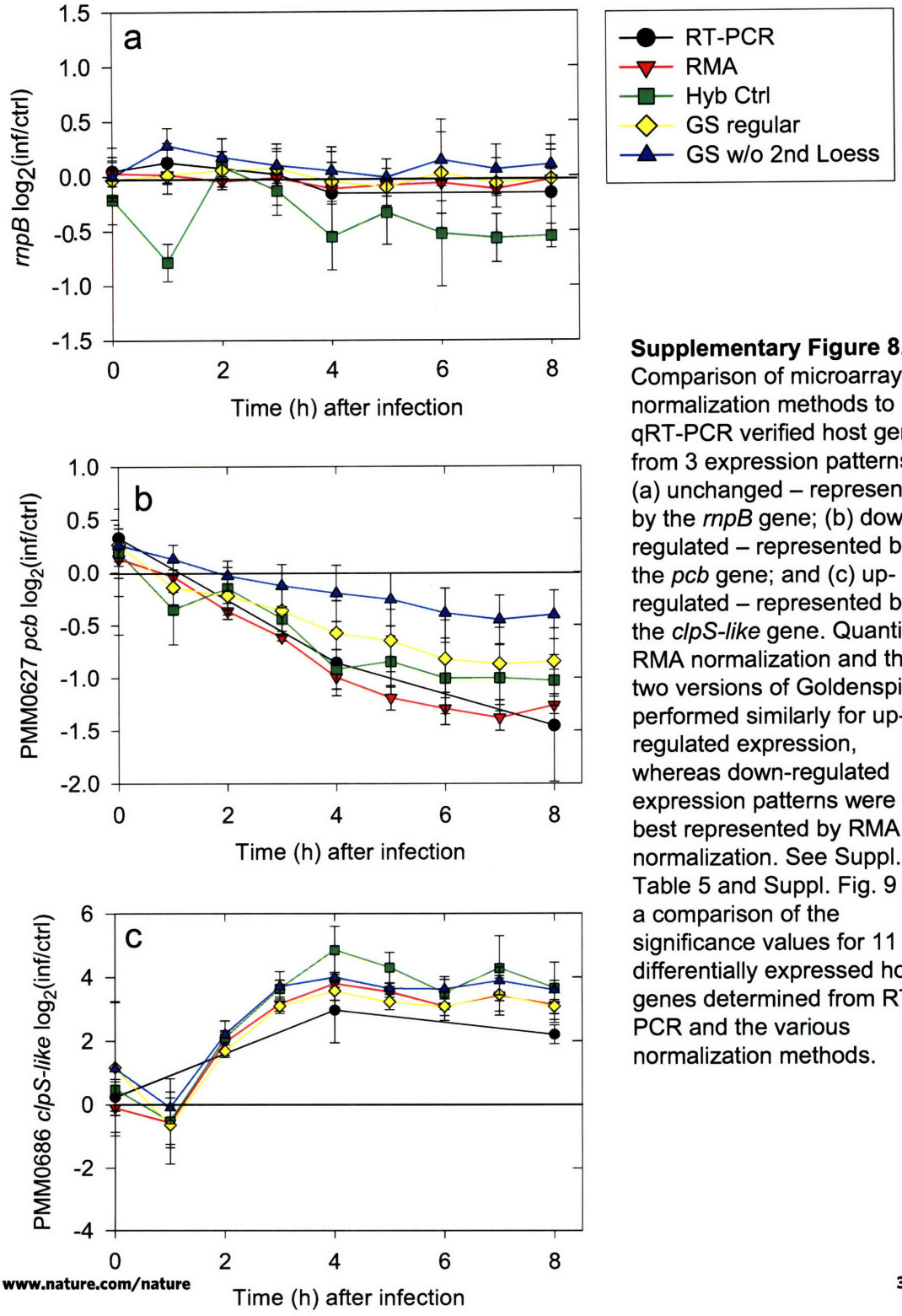
Supplementary Figure 5. Comparison of extracellular P-SSP7 quantification using a quantitative PCR (qPCR) assay for the phage DNA polymerase gene with (a) infective titer determined by the most probable number (MPN) assay and (b) total phage particles after staining with the DNA SYBR Green I stain and enumerated by epifluorescence microscopy. The linear regression for (a) is y = 2.38x + 410324, R^2 = 0.87; and (b) is y = 0.84x – 132486, R2 = 0.97. Note that qPCR quantification provides close to a 1:1 ratio with the SYBR stained particles, but was 2.5 fold higher than infective phage.



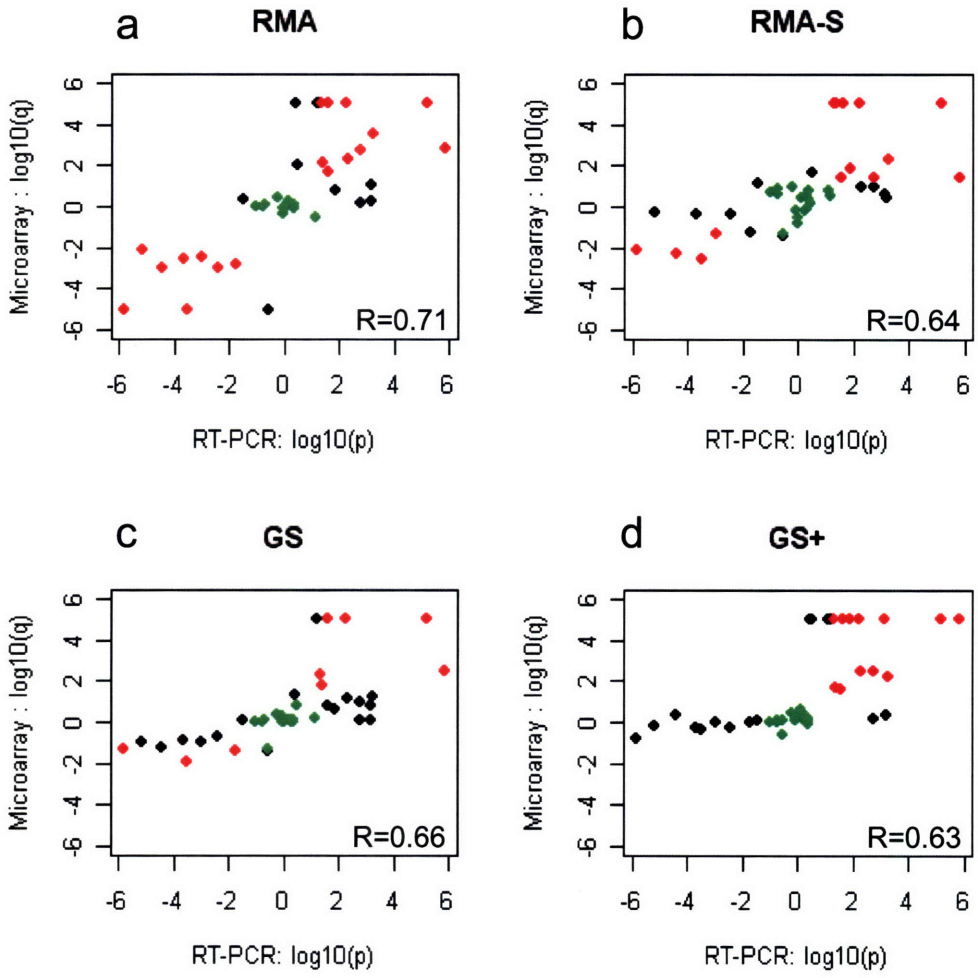
Supplementary Figure 6. qRT-PCR verification of phage gene expression patterns determined from microarray results. (a) Expression profiles for representative genes from each transcription cluster were analyzed by RT-PCR using gene specific primers. The results were normalized to *rnpB* (an internal control gene) to correct for potential differences in input RNA, and are presented relative to maximum levels for each gene. The results are shown on a logarithmic scale to better discern differences in expression patterns at the early time points which are low relative to maximal transcript levels. (b) Microarray results for the same representative genes shown to facilitate direct comparison to the RT-PCR results. Note that, as is commonly found, changes in expression determined by RT-PCR were orders of magnitude greater than by microarray analysis.



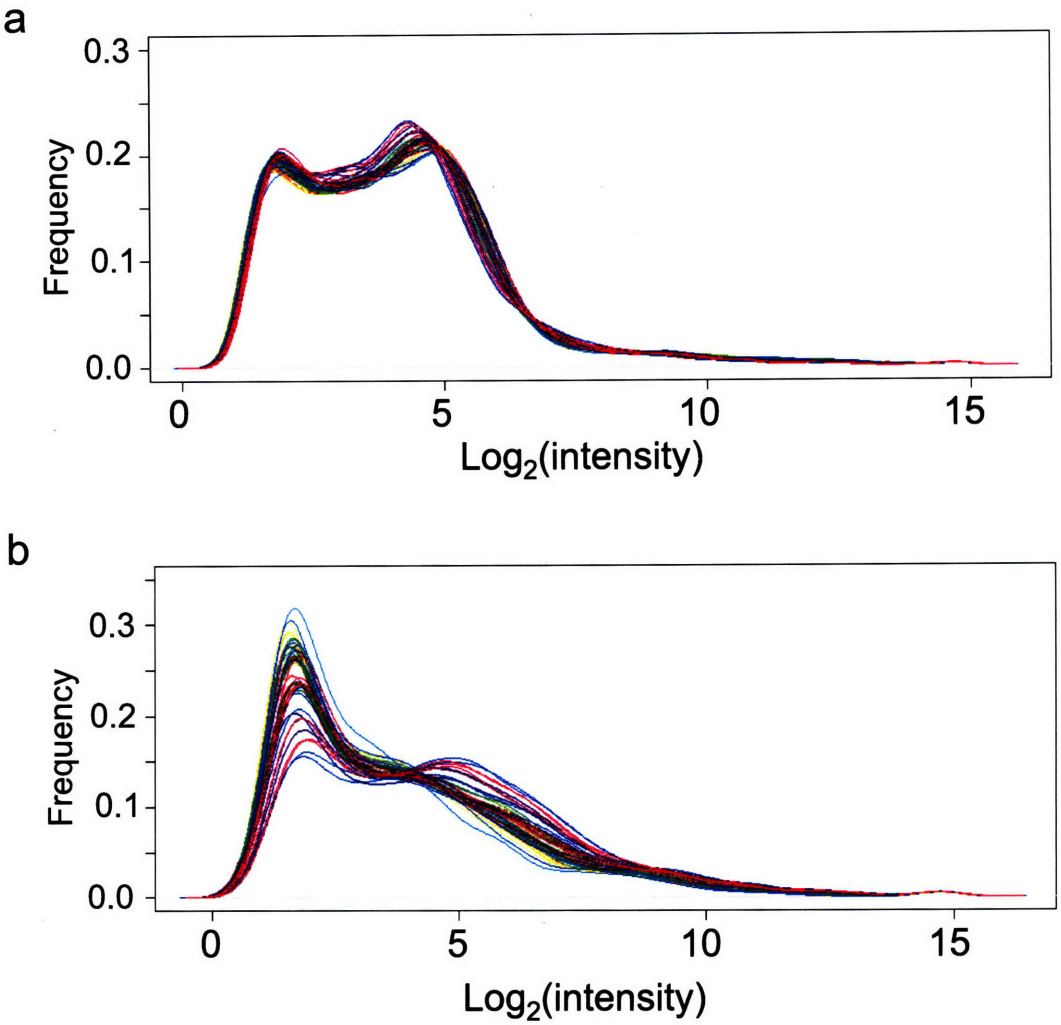
Supplementary Figure 7. qRT-PCR verification of host gene expression patterns determined from microarray analysis. Left panel (a, c, e) shows RT-PCR results and right panel (b, d, f) shows microarray results for representative genes displaying: (a, b) unchanged expression profile; (c, d) down-regulated genes; and (e, f) up-regulated genes. PMM1629 is the only gene whose up-regulated expression was not verified by RT-PCR (compare c and d at T=1 h). Suppl. Table 5 provides a direct comparison of the significance of differentially expressed from qRT-PCR and microarrays after quantile RMA normalization.



Supplementary Figure 8. Comparison of microarray normalization methods to qRT-PCR verified host genes from 3 expression patterns: (a) unchanged – represented by the *rnpB* gene; (b) down-regulated – represented by the *pcb* gene; and (c) up-regulated – represented by the *clpS-like* gene. Quantile RMA normalization and the two versions of Goldenspike performed similarly for up-regulated expression, whereas down-regulated expression patterns were best represented by RMA normalization. See Suppl. Table 5 and Suppl. Fig. 9 for a comparison of the significance values for 11 differentially expressed host genes determined from RT-PCR and the various normalization methods.



Supplementary Figure 9. Comparison of the performance of different microarray normalization methods for detection of significant differences in gene expression as compared to RT-PCR analysis. (a) Quantile RMA normalization at the probe level; (b) RMA normalization based on spiked in hybridization controls; (c) Goldenspike (GS) without summary level normalization; and (d) Goldenspike with summary level normalization. R = Pearson correlation between RT-PCR and microarray analysis. Red and green symbols denote significant and insignificant differences respectively in gene expression called for both RT-PCR and microarray analysis, whereas black symbols denote discrepancies in significance calls between the RT-PCR and microarray analyses. The significance of differential expression for microarray analyses (q-values) was calculated using the Bayes t-test and significance of differential expression for RT-PCR was determined from a standard two-tailed t-test (p-values). Q-values <0.00001 were set to 0.00001 to ensure non-zero values. These findings show that quantile RMA normalization gave the highest correlation and the largest number of correctly identified differentially expressed genes, especially for downregulated genes.



Supplementary Figure 10. Density distribution of signal intensities for probe sets from each microarray after quantile RMA normalization: (a) All probe sets are displayed. The overall distribution for all arrays is similar due to the quantile normalization. (b) MED4 probe sets only are displayed. Note that different arrays displayed various distributions for the MED4 probe sets despite the similar distribution for all probe sets. Therefore quantile normalization can be used for this experiment without erasing differences in MED4 gene expression. Each line represents a different array.

Appendix E

Three *Prochlorococcus* Cyanophage Genomes: Signature Features and Ecological Interpretations

Matthew B. Sullivan, Maureen L. Coleman, Peter Weigele, Forest Rohwer, and Sallie W. Chisholm

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 Biology* 3:e144.

Three *Prochlorococcus* Cyanophage Genomes: Signature Features and Ecological Interpretations

Matthew B. Sullivan[1], Maureen L. Coleman[2], Peter Weigele[3], Forest Rohwer[4], Sallie W. Chisholm[2,3*]

1 Joint Program in Biological Oceanography, Woods Hole Oceanographic Institution and Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, 2 Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, 3 Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, 4 Department of Biology, San Diego State University, San Diego, California, United States of America

The oceanic cyanobacteria *Prochlorococcus* are globally important, ecologically diverse primary producers. It is thought that their viruses (phages) mediate population sizes and affect the evolutionary trajectories of their hosts. Here we present an analysis of genomes from three *Prochlorococcus* phages: a podovirus and two myoviruses. The morphology, overall genome features, and gene content of these phages suggest that they are quite similar to T7-like (P-SSP7) and T4-like (P-SSM2 and P-SSM4) phages. Using the existing phage taxonomic framework as a guideline, we examined genome sequences to establish "core" genes for each phage group. We found the podovirus contained 15 of 26 core T7-like genes and the two myoviruses contained 43 and 42 of 75 core T4-like genes. In addition to these core genes, each genome contains a significant number of "cyanobacterial" genes, i.e., genes with significant best BLAST hits to genes found in cyanobacteria. Some of these, we speculate, represent "signature" cyanophage genes. For example, all three phage genomes contain photosynthetic genes (*psbA, hliP*) that are thought to help maintain host photosynthetic activity during infection, as well as an aldolase family gene (*talC*) that could facilitate alternative routes of carbon metabolism during infection. The podovirus genome also contains an integrase gene (*int*) and other features that suggest it is capable of integrating into its host. If indeed it is, this would be unprecedented among cultured T7-like phages or marine cyanophages and would have significant evolutionary and ecological implications for phage and host. Further, both myoviruses contain phosphate-inducible genes (*phoH* and *pstS*) that are likely to be important for phage and host responses to phosphate stress, a commonly limiting nutrient in marine systems. Thus, these marine cyanophages appear to be variations of two well-known phages—T7 and T4—but contain genes that, if functional, reflect adaptations for infection of photosynthetic hosts in low-nutrient oceanic environments.

Citation: Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW (2005) Three *Prochlorococcus* cyanophage genomes: Signature features and ecological interpretations. PLoS Biol 3(5): e144.

Introduction

Prochlorococcus is the numerically dominant primary producer in the temperate and tropical surface oceans [1]. These cyanobacteria are the smallest known photosynthetic organisms (less than a micron in diameter), yet are significant contributors to global photosynthesis [2,3] because they occur in high abundance (as many as 10^5 cells/ml) throughout much of the world's oceans. They are adapted to living in low-nutrient oceanic regions [4] and are physiologically and genetically diverse with at least two "ecotypes" that have distinctive light physiology [5], nitrogen [6] and phosphorus (L. R. Moore, personal communication) utilization, and copper [7] and virus (phage) [8] sensitivity. Cyanobacterial phages are also abundant in these environments [8,9,10,11,12] and have a small, but significant, role in mediating population sizes [9,10]. Further, cyanophages likely play a role in maintaining the extensive microdiversity within marine cyanobacteria [9,10] through keeping "competitive dominants" (sensu [13]) in check, as well as by carrying photosynthetic "host" genes [14,15,16] and mediating horizontal transfer of genetic material between cyanobacterial hosts [14].

Although there are more than 430 completed double-stranded DNA phage genomes in GenBank, only nine phages with published genomes infect marine hosts (cyanophage P60; vibriophages VpV262, KVP40, VP16T, VP16C, K139, and VHML; roseophage SIO1; and *Pseudoalteromonas* phage PM2). Of those nine, only one infects cyanobacteria (cyanophage P60, a member of the Podoviridae). P60 was isolated from estuarine waters using *Synechococcus* WH7803 as a host and appears most closely related to the T7-like phages [17]. It contains 11 T7-like phage genes and has no genes with homology to non-T7-like phages. However, it lacks the conserved T7-like genome architecture. Thus, P60 is thought to be only distantly related to the T7-like phages, but still part of a T7 supergroup [18] proposed by Hardies et al. [19]. The

Received June 19, 2004; Accepted February 23, 2005; Published April 19, 2005
DOI: 10.1371/journal.pbio.0030144

Abbreviations: DNAP, deoxyribonucleic acid polymerase; HL, high-light-adapted; HN, hemagglutinin neuraminidase; LL, low-light-adapted; LPS, lipopolysaccharide; ORF, open reading frame; PSII, photosystem II; RNAP, ribonucleic acid polymerase; RNR, ribonucleotide reductase; ssRNA, single-stranded ribonucleic acid

Academic Editor: Nancy A. Moran, University of Arizona, United States of America

*To whom correspondence should be addressed. E-mail: chisholm@mit.edu

T7 supergroup also contains two other marine phages (roseophage SIO1 and vibriophage VpV262) that show similarity to some (three) T7-like genes. However, these phages lack many T7-like genes including the hallmark T7-like RNA polymerase (RNAP) gene [18]. Thus, there is clearly a gradient in relatedness among the T7 supergroup, with these newer marine phage genomes at the distant, less-similar end of the group.

Marine phages are subject to different selection pressures (e.g., dispersal strategies, encounter rates, limiting nutrients, and environmental variability) than their relatively well-studied terrestrial counterparts. Thus, beyond informing phage taxonomy, the analysis of their genomes should unveil "signatures" of these selective agents. For example, genomic analysis of two marine phages, roseophage SIO1 [20] and vibriophage KVP40 [21], has revealed phosphate-inducible genes. It is thought that these genes play an important regulatory role in the phosphorus-limited waters from which they were isolated. Similarly, some *Prochlorococcus* and *Synechococcus* phages (including the three cyanophage genomes presented here) contain core photosynthetic genes that are full-length, conserved, and cyanobacterial in origin [14,15,16]. They are hypothesized to be important for maintaining active photosynthetic reaction centers—and hence the flow of energy—during phage infection [14,15,16].

With a large collection of phages from which to choose [8], we used host range and phage morphology to select strains for sequencing. The selected podovirus (P-SSP7) is very host-specific, infecting a single high-light-adapted (HL) *Prochlorococcus* strain of 21 *Prochlorococcus* and *Synechococcus* strains tested. In contrast, the two myoviruses that were selected cross-infect between *Prochlorococcus* (but not *Synechococcus*) hosts: P-SSM2 can infect three low-light-adapted (LL) host strains, and P-SSM4 can infect two HL and two LL hosts [8]. We had no prior knowledge of the gene content of these phages; thus, with regard to their genomes, these phages were selected randomly.

As mentioned earlier, our first survey of these phage genomes led to the surprising discovery of photosynthetic genes in all three *Prochlorococcus* phages [14], similar to the findings in *Synechococcus* cyanophages [15,16,22]. In this report, we present a more thorough analysis of these three cyanophage genomes, which, we argue, appear to be T7-like (P-SSP7) and T4-like (P-SSM2 and P-SSM4) phages.

Results/Discussion

General Features of the Podovirus P-SSP7

P-SSP7 is morphologically similar to the Podoviridae (tails are short and noncontractile; Figure 1A). It also includes a rectangular region of electron transparency (Figure 1A) that is similar to the gp14/gp15/gp16 core located at the unique portal vertex found in coliphage T7 [23]. Its genome contains 44,970 bp (54 open reading frames [ORFs]; 38.7% G+C content; Figure 1B), including a T7-like RNAP and a phage-related integrase gene (a more detailed analysis of this feature is discussed later). Thus, the P-SSP7 genome is more T7-like or P22-like than ϕ29-like among the Podoviridae (Table 1). Thirty-five percent of the translated ORFs have best hits to phage proteins; nearly all of these are T7-like, whereas none are P22-like (Figure 1C). Together, these data suggest that P-SSP7 is most closely related to the T7-like phages. Surprisingly, 11% of the translated ORFs have best hits to bacterial proteins, with well over half of these being cyanobacterial (see later discussion). Roughly half (54%) of the translated ORFs could not be assigned a function (Figure 1C).

An examination of the genomes of coliphage T7 and its closest coliphage relatives (T3, gh-1, ϕYe03–12, ϕA1122) revealed that they share 26 genes, which we define as core genes (Table 2). P-SSP7 has 15 of these 26 core genes and an additional gene *(0.7)* that is common, but not universal, among T7-like phages (Table 2). Further, only two non-T7-like phage genes were identified in this genome: hypothetical gene *12* from a *Burkholderia* phage, *Bcep1*, of the Myoviridae family, and the phage-related integrase gene discussed later. Strikingly, the T7-like genes found in P-SSP7 are arranged in exactly the same order as in other T7-like phages (Figure 1B). The gene content and genome architecture of P-SSP7 contrast with those from the three other sequenced marine podovirus genomes in the T7 supergroup [17,19,20]. SIO1 and VpV262 lack the hallmark T7-like RNAP and contain only three T7-like core genes (Table 2), whereas cyanophage P60 contains 11 core genes (Table 2) but clearly lacks the conserved T7-like genome architecture [17].

The putative functions of the 16 T7-like genes in P-SSP7 would allow for the majority of host interactions and phage production as follows (T7-like gene designations are shown in parentheses): shutdown of host transcription *(0.7)*, phage gene transcription *(1)*, degradation of host DNA *(3, 6)*, DNA replication *(1, 2.5, 4, 5)*, formation of a channel across the cell envelope via an extensible tail *(15, 16)* [24], DNA packaging *(19)*, and virion formation *(8, 9, 10, 11, 12, 17)*. We found two stretches of DNA (frame +1 from nucleotides 9994–10525, then frame +3 from nucleotides 10485–11759) with matches to T7 gp5 (DNA polymerase [DNAP]): one corresponding to the 3′-exonuclease and one to the polymerase (nucleotidyl transferase) segments of the T7 enzyme. This region may encode a split variant of T7 family DNAP (V. Petrov and J. Karam, personal communication), an arrangement that has been shown to be functional in archaea [25] and some T4-like phages (V. Petrov and J. Karam, personal communication).

As described earlier, we identified only 15 of the 26 core T7-like genes in P-SSP7. What are the functions of the absent gene set? It includes genes that in T7 are involved in ligation of DNA fragments *(1.3)*, inhibition of host RNAP *(2)*, interactions that are specific to the host cell envelope during virion formation *(6.7, 13, 14)*, lysis events *(3.5, 17.5)*, small-subunit terminase activity *(18)*, and unknown functions *(5.7, 6.5, 18.5)* [23]. These same genes are also absent in the marine podovirus genomes in the T7 supergroup (cyanophage P60, vibriophage VpV262, and roseophage SIO1; Table 3). If we assume a conserved genomic architecture among the T7-like phages, we find hypothetical ORFs in homologous positions to these T7 core genes in P-SSP7 (Figure 1B) that may fulfill these core (e.g., *5.7, 6.5, 6.7, 13, 14, 17.5, 18, 18.5*) and common (e.g., antirestriction gene *0.3*) T7-like gene functions. Alternatively, their functions may be unnecessary for this phage.

The P-SSP7 genome assembled as a circular chromosome, suggesting that it is circularly permuted, thus lacking the terminal repeats that are common among T7-like phages [26]. Confirmation of this hypothesis would require direct sequencing of the genome ends (I. Molineux, personal communication), which was not possible in this study because

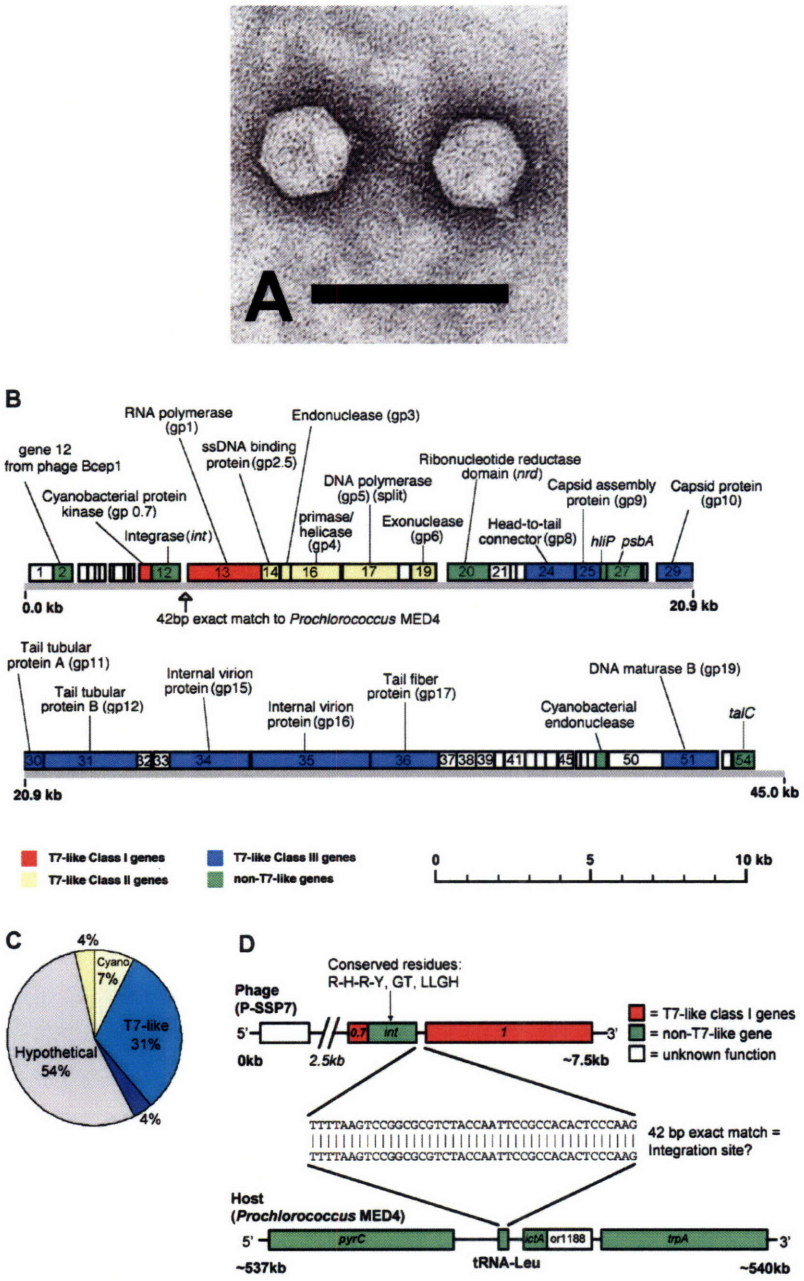


Figure 1. Features of the *Prochlorococcus* Podovirus P-SSP7

(A) Electron micrograph of negative-stained podovirus P-SSP7. Note the distinct T7-like capsid and tail structure. Scale bar indicates 100 nm.
(B) Genome arrangement of *Prochlorococcus* podovirus P-SSP7. The ORFs are sequentially numbered within the boxes, and gene names are designated above the boxes. Gene designations use T7 nomenclature for T7-like genes [24] or microbial nomenclature for non-phage genes. Class I, II, and III genes refer to those in T7 [66] that belong to gene regions primarily involved in host transcription of phage genes (class I), DNA replication (class II), and the formation of the virion structure (class III). The ORFs are designated by boxes, and in this genome, all ORFs are oriented in the same direction. Although the phage genome is one molecule of DNA, the representation is broken to fit on a single page. Note that the P-SSP7 genome is most similar to genomes of the T7-like phages.
(C) Taxonomy of best BLASTp hits for P-SSP7. Each predicted coding sequence from the phage genomes was used as a query against the nonredundant database to identify the taxon of the best hit (details in Materials and Methods). Blue slices indicate phage hits, while yellow slices indicate cellular hits.
(D) Diagrammatic representation of the genomic regions surrounding a putative phage and host integration site. This site consists of a 42-bp exact match between the podovirus P-SSP7 and its host *Prochlorococcus* MED4 located directly downstream of the phage integrase gene and the noncoding strand of a host tRNA gene.
DOI: 10.1371/journal.pbio.0030144.g001

Table 1 Genome-Wide Characteristics of the *Prochlorococcus* Cyanophage P-SSP7 Relative to the Other Recognized Phage Groups within the Podoviridae [105]

Phage	Hosts	Size (kb)	Number of ORFs	Terminal Repeats	RNAP	Integrase Gene
P-SSP7	Cyanobacteria	45	53	?	Y	Y
T7-like	Gram negatives	38–43	43–56	Y	Y	N
P22-like	Gram negatives	38–50	60–65	N	N	Y
ϕ29-like	Gram positives	18–22	17–35	N	N	N

Y indicates that the feature is present, N indicates that the feature is absent, and a question mark indicates that the presence or absence of the feature is unknown.

DOI: 10.1371/journal.pbio.0030144.t001

of the difficulty of obtaining significant quantities of purified DNA [27].

Hypothesized Lysogeny in P-SSP7

One of the more interesting discoveries in the podovirus genome is the presence of a tyrosine site-specific recombinase *(int)* gene (Figure 1B), which in temperate phages encodes a protein that enables the phage to integrate its genome into the host genome [28]. T7 is a classically lytic phage, and there has been only one other report of *int* genes in a T7-like phage: in an integrated prophage in the *Pseudomonas putida* KT2440 genome [29]. The P-SSP7 *int* contains conserved amino acid motifs previously identified for site-specific recombinases (Arg-His-Arg-Tyr, Leu-Leu-Gly-His, and Gly-Thr [30]) suggesting it is functional. Downstream of *int*, we find a 42-bp sequence that is identical to part of the noncoding strand of the leucine tRNA gene in the phage's host genome (*Prochlorococcus* MED4) (Figure 1D). tRNA genes are a common integration site for phages and other mobile elements [31], adding support to the hypothesis that this *int* gene is functional.

P-SSP7 was isolated from surface ocean waters at the end of summer stratification [8], when nutrients are extremely limiting. We have hypothesized [8] that the integrating phase

Table 2. Shared Genes in T7-Like Phages

Class	Gene	P-SSP7	e-Value	T7 Supergroup Phages								Functions
				T7	T3	gh-1	ϕYe03–12	ϕA1122	P60	VpV262	SIO1	
Class I	0.7	125	10^{-18}[a]	359	370	—	370	—	—	—	—	Protein kinase
	1	779	10^{-91}	883	885	886	885	884	574	—	—	RNA polymerase
	1.3	—	—	359	347	355	347	341	—	—	—	DNA ligase
Class II	2	—	—	64	55	56	79	65	—	—	—	Host RNA polymerase inhibitor
	2.5	190	0.004	232	233	234	233	233	—	—	—	Single-stranded binding protein
	3	117	10^{-19}	149	153	148	154	152	116[b]	—	135	Endonuclease
	3.5	—	—	151	152	147	152	152	—	—	—	Amidase (lysozyme)
	4	521	10^{-132}	566	567	563	567	567	531	287/408	523	Primase-helicase
	5	589[c]	10^{-124}	704	705	710	705	705	587	661	581	DNA polymerase
	5.7	—	—	69	69	70	70	70	—	—	—	
	6	260	10^{-44}	300	303	315	304	301	243	—	—	Exonuclease
Class III	6.5	—	—	84	81	81	82	85	—	—	—	
	6.7	—	—	88	83	91	84	89	—	—	—	Internal virion protein
	8	523	10^{-171}	536	536	544	536	537	555	—	—	Head–tail connector protein
	9	266	10^{-50}	307	311	292	311	305	246[b]	—	—	Head assembly protein
	10	376	10^{-40}	345	348	348	348	345	221	—	—	Major capsid protein
	11	205	10^{-23}	196	197	196	197	197	192[b]	—	—	Tail protein
	12	977	10^{-58}	794	802	809	802	795	680	—	—	Tail protein
	13	—	—	138	137	145	139	139	—	—	—	
	14	—	—	196	198	194	198	197	—	—	—	Internal core protein
	15	838	—	747	749	739	748	748	—	—	—	Internal core protein
	16	1,246	—	1,318	1,319	1,393	1,321	1,319	—	—	—	Internal core protein
	17	716	10^{-10}	553	559	619	646	559	—	—	—	Tail fiber protein
	17.5	—	—	67	67	72	68	68	—	—	—	Putative lysis protein
	18	—	—	89	89	86	89	90	—	—	—	Small terminase subunit
	18.5	—	—	143	148	150	151	148	—	—	—	
	19	578	10^{-121}	586	587	583	588	587	566 [b]	535	—	Large terminase subunit

The T7 supergroup contains phages with close similarity to T7 (the T7-like phages T3, gh-1, ϕYe03-12, and ϕA1122), as well as more distant relatives (e.g., P60, VpV262, ϕ-KMV, and SIO1) [19]. All T7-like phages are represented as well as the marine phages belonging to the T7 supergroup for comparison. The size (amino acids) of each predicted coding region is presented using gene numbers and function assignments according to T7 terminology [24]. For P-SSP7, No e-value is given for ORFs that were assigned using size, domain homology, and synteny. A long dash indicates the lack of a particular gene using standard searches.

[a]The best e-value was microbe-related rather than related to the T7-like phages.

[b]Putative split genes in cyanophage P60.

[c]A putative frameshifted gene in cyanophage P-SSP7.

DOI: 10.1371/journal.pbio.0030144.t002

Table 3. Genome-Wide Characteristics of the *Prochlorococcus* Cyanomyophages P-SSM2 and P-SSM4 Relative to the Other Recognized Phage Groups within the Myoviridae [105]

Phage	Hosts	Size (kb)	Number of ORFs	Integrating Phage Element	DNA Conformation
P-SSM2	Cyanobacteria	252	327	N	Circularly permuted
P-SSM4	Cyanobacteria	178	198	N	Circularly permuted
T4-like	Gram negatives	164–255	252–384	N	Circularly permuted
P1-like	Gram negatives	<50	40	?	Circularly permuted
P2-like	Gram negatives	30–34	40–44	N	Circularly permuted
Mu-like	Gram negatives	37	55	Y[a]	Linear
Spo1-like	Gram positives	<50	40	?	Linear
ϕH-like	Archaea	58–78	98–121	Y	Circularly permuted

Y indicates that the feature is present, N indicates that the feature is absent, and a question mark indicates that no representative phage genomes have been completely sequenced, so the presence or absence of the character is unknown.
[a] Phage integrates using a transposase rather than a site-specific integrase.
DOI: 10.1371/journal.pbio.0030144.t003

of the temperate-phage life cycle may be selected for under these conditions; thus, finding the *int* gene in this particular phage is consistent with this hypothesis. None of the complete genome sequences of cyanobacterial hosts reported to date have intact prophages [4,32,33,34]. Moreover, temperate phages have not been induced from unicellular freshwater or marine cyanobacterial cultures [9,35,36]. Although some field experiments suggest that temperate cyanophages can be induced from *Synechococcus* [37,38], prophage integration has not been demonstrated. Thus, experimental validation that P-SSP7 is capable of integration would confirm indirect evidence and establish a valuable experimental system.

General Features of the Myoviruses P-SSM2 and P-SSM4

P-SSM2 and P-SSM4 are morphologically similar to the Myoviridae (tails are long and contractile; Figure 2). Both have an isometric head, contractile tail, baseplate, and tail fiber structures (Figure 2) that are most consistent (but see isometric head discussion later) with the morphological characteristics of the T4-like phages [39]. Their genomes also have general characteristics that are fully consistent with T4-like status within the Myoviridae (Table 3). Both genomes are relatively large: P-SSM2 has 252,401 bp (327 ORFs; 35.5% G+C content; Figure 3) and P-SSM4 has 178,249 bp (198 ORFs; 36.7% G+C content; Figure 4). An apparent strand bias is noteworthy because only 12 (of 327) and six (of 198) ORFs are predicted on the minus strand in the P-SSM2 and P-SSM4 genomes, respectively. Similar to the lytic T4-like phages, integrase genes were absent. Both genomes assembled and closed, suggesting the circularly permuted chromosome common among the T4-like phages (Table 3). A large portion of the nonhypothetical ORFs have best hits to phage proteins (14% and 21%, respectively) and bacterial proteins (26% and 21%, respectively; Figure 5). The phage hits were most similar

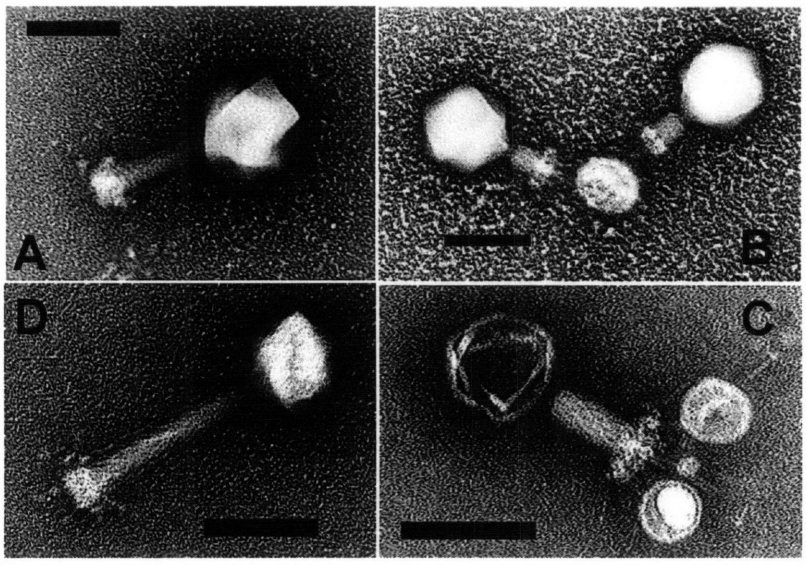


Figure 2. Electron Micrograph of Negative-Stained *Prochlorococcus* Myoviruses P-SSM2 and P-SSM4
Myovirus P-SSM2 with (A) non-contracted tail and (B) contracted tail, and myovirus P-SSM4 with (C) contracted tail and (D) non-contracted tail. Note the T4-like capsid, baseplate, and tail structure in both myoviruses. Scale bars indicate 100 nm.
DOI: 10.1371/journal.pbio.0030144.g002

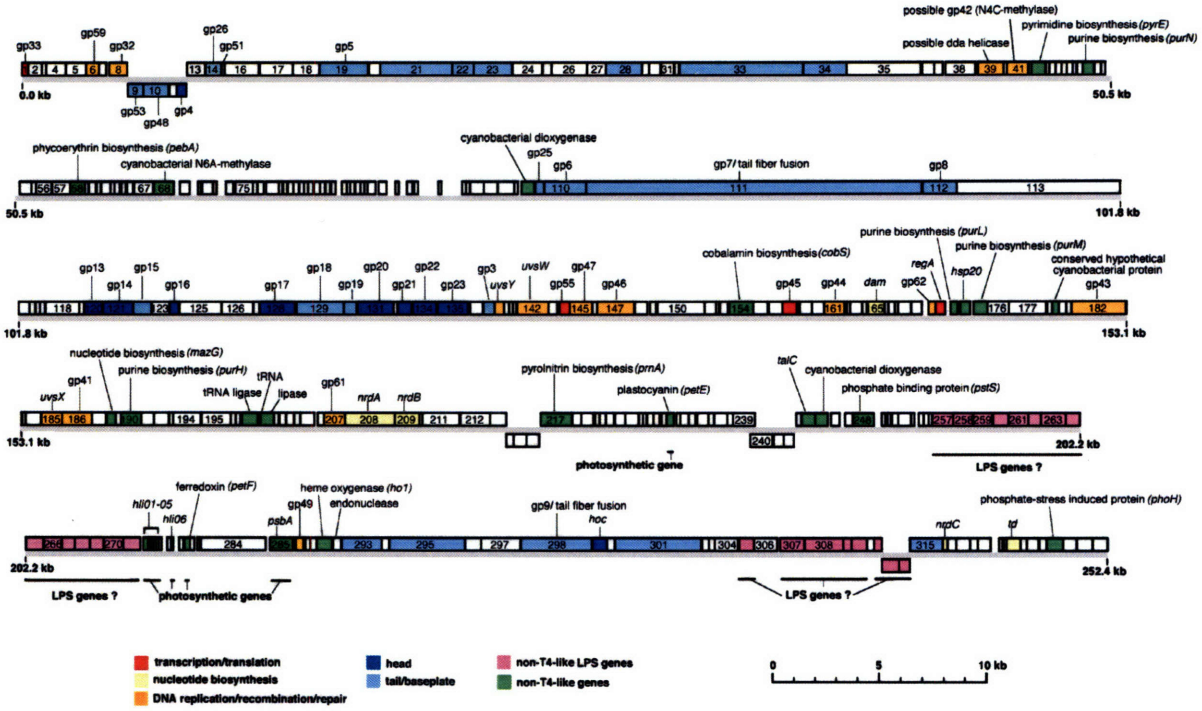


Figure 3. Genome Arrangement of the *Prochlorococcus* Myovirus P-SSM2

Gene names are designated above the box representing the ORF where genes were identified; descriptions of genes are in Table 4. The ORFs located above the centering line are on the forward DNA strand, whereas those below the line are on the reverse strand. Although the genome is one molecule, the representation is broken to fit the page. Colors indicate the putative role for the identified genes as inferred from T4 phage. Gene designations use T4 nomenclature for T4-like genes [104] or microbial nomenclature for non-phage genes.

DOI: 10.1371/journal.pbio.0030144.g003

to T4-like phage proteins, and about half of the bacterial ORFs were most similar to those from cyanobacteria. As with P-SSP7, most of the translated ORFs from P-SSM2 and P-SSM4 could not be assigned a function (60% and 58%, respectively). The majority of the differences between these two phages are due to the presence of two large clusters of genes (24 total) in P-SSM2 (see Figure 3) that are absent from P-SSM4. These clusters contain many sugar epimerase, transferase, and synthase genes that we hypothesize to be involved in lipopolysaccharide (LPS) biosynthesis. The large genome size, collective gene complement, and morphology suggest both P-SSM2 and P-SSM4 are most closely related to T4-like phages.

The six sequenced T4-like phage genomes (T4, RB69, RB49, 44RR2.8t, KVP40, and Aeh1; available as of 15 May 2004 at http://phage.bioc.tulane.edu/) share 75 genes (Table 4), which suggests a core gene complement required for T4-like phage infection. This core contains 18 genes involved in DNA replication, recombination, and repair, seven regulatory genes, ten nucleotide metabolism genes, 34 virion structure and assembly genes, and six genes involved in chaperonin, lysis exclusion, and other activities. Again, despite cyanobacterial hosts being quite divergent from the hosts of these other T4-like phages, our myoviruses contained 43 and 42 of the 75 T4-like core genes, as well as other noncore T4-like genes in each phage (*uvsX, uvsY,* and possibly *dam, 42,* and *hoc* in P-SSM2; *uvsX, uvsY,* and possibly *dam, 42,* and *denV* in P-SSM4; Table 4). Furthermore, aside from the low-complexity tail fiber related genes (see "Tail-Fiber-Related Genes in the Myoviruses" below), we found no genes with sequence similarity to any phage type other than T4-like phages.

Slightly fewer than half of the core T4-like genes were absent in both myoviruses P-SSM2 and P-SSM4. P-SSM2 and P-SSM4 lack the genes required for anaerobic nucleotide biosynthesis (*nrdD, nrdG,* and *nrdH),* which is perhaps not surprising because these phages were isolated from the well-mixed, oxygenated surface oceans. Both myoviruses also lack homologs to the prohead core-encoding genes *(67* and *68)* of the T4-like phages (Table 4). However, we note that the capsids of both *Prochlorococcus* myoviruses are isometric (see Figure 2), rather than prolate as is often observed for other T4-like phage capsids [39]. In T4, mutations in the prohead core proteins (gp67 and gp68) are known to cause a capsid structural defect whereby isometric heads are observed [40,41,42]. Thus, functional homologs of prohead core proteins may not be required for the formation of isometric heads in these *Prochlorococcus* myoviruses.

Other T4-like phage gene functions may be represented by divergent homologs filling the T4-like phage role in these cyanomyophages. P-SSM2 and P-SSM4 lack core T4-like chaperonin genes *(rnlA, 31,* and *57A;* Table 4) and nucleotide metabolism genes (T4-like pyrimidine biosynthesis: *cd, frd, 1,* and *tk*; Table 4). However, both P-SSM2 and P-SSM4 contain non-T4-like hsp20-family chaperonins, as well as a non-T4-

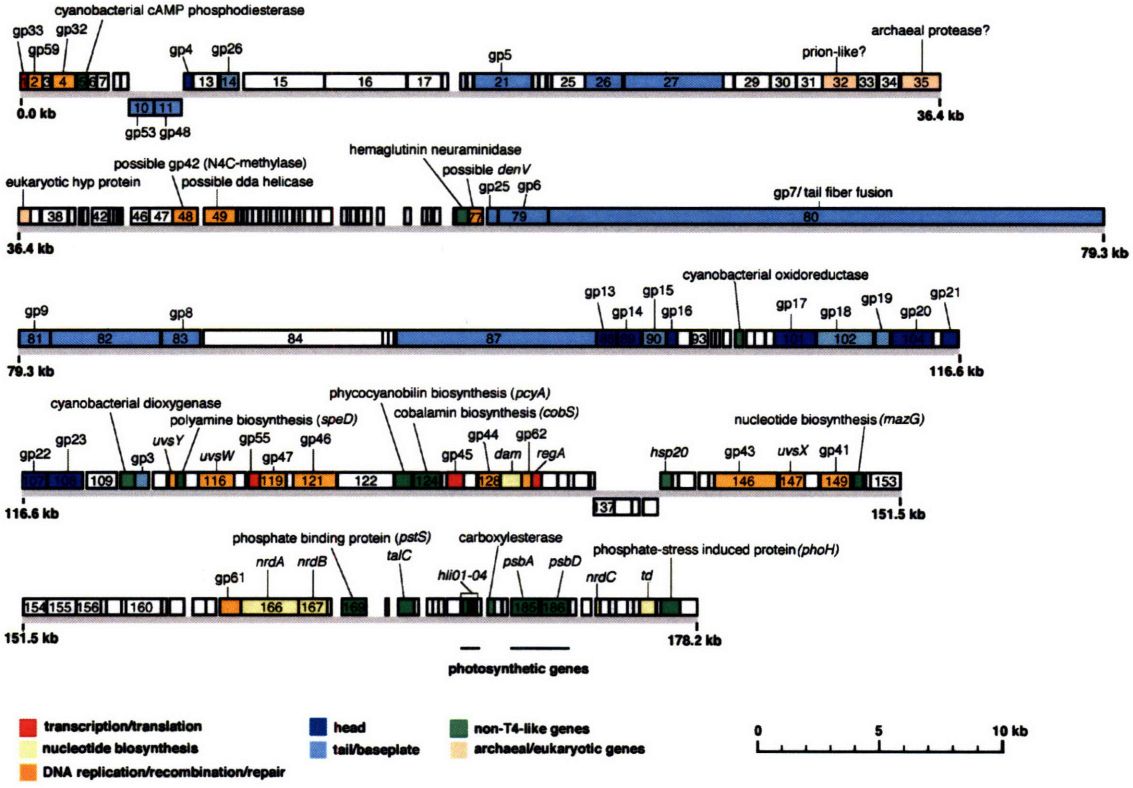


Figure 4. Genome Arrangement of the *Prochlorococcus* Myovirus P-SSM4.
Gene nomenclature is as in Figure 3.
DOI: 10.1371/journal.pbio.0030144.g004

like gene *(mazG)* that in bacteria is involved in degradation of DNA (Table 5) [43,44]. Furthermore, P-SSM2 contains ORFs with high sequence similarity to host-encoded homologs of five genes involved in pyrimidine *(pyrE)* and purine *(purH, purL, purM,* and *purN)* biosynthesis (Table 5). These non-T4-like genes might compensate for T4-like nucleotide metabolism and/or chaperone genes that are absent. Despite the structural similarities between our myophages (see Figure 2) and the T4-like phages, some core virion structural genes (e.g., head genes, *2, 24, 67, 68,* and *inh*; tail/tail fiber genes, *10, 11, 12, 34, 35, 37,* and *wac*) have yet to be identified in these myophage genomes (see Table 4). Similarly, genes involved in transcriptional regulation *(dsbA, rnlA,* and *pseT),* lysis events *(rIIa* and *rIIb),* and replication, recombination, and repair (DNA ligase, *30;* topoisomerases, *39* and *52;* RNase H, *rnh;* and an exonuclease, *dexA*) also have yet to be identified.

Tail-Fiber-Related Genes in the Myoviruses

Sequence analysis of phage tail fiber genes has revealed extensive swapping of gene fragments between loci [45,46]. Such exchanges yield phages with altered host ranges [47]. Although this mosaic gene construction makes computational identification of tail fiber genes by sequence homology difficult, we have attempted to do so in the two *Prochlorococcus* T4-like genomes. The analysis is motivated by the belief that understanding mechanisms of attachment and host range is critical for developing assays for studying phage–host interactions in wild populations—one of the underlying motivations of our work with this system.

We identified ORFs as potential tail fiber genes by a three-tiered bioinformatics approach using sequence similarity, repeat analysis, and paralogy (details in Materials and Methods). First, sequence similarity to known tail fiber genes was used to add ORFs to the pool of possible tail fiber genes (Figure 6). Seven ORFs in P-SSM2 and three ORFs in P-SSM4 had similarity to known tail fiber genes. In T4, the long tail fiber of T4 is composed of four protein subunits including a

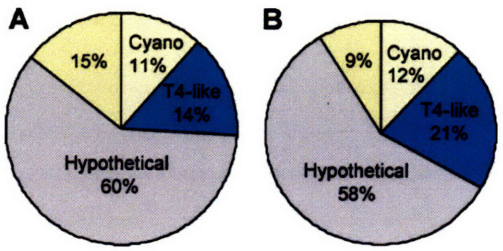


Figure 5. Taxonomy of Best BLASTp Hits for P-SSM2 and P-SSM4
Each predicted coding sequence from both phage genomes was used as a query against the nonredundant database to identify the taxon of the best hit (details in Materials and Methods). Blue slices indicate phage hits, while yellow slices indicate cellular hits.
DOI: 10.1371/journal.pbio.0030144.g005

Table 4. Shared Genes in T4-like Phages

Category	Gene	P-SSM2	e-Value	P-SSM4	e-Value	T4 Supergroup Phages						Description of Protein Product
						T4	RB69	RB49	44RR2.8t	KVP40	Aeh1	
Transcription	*dsbA*	—	—	—	—	89	95	91	95	90	95	DNA binding
	33	87	10^{-7}	87	10^{-4}	112	112	89	85	98	76	Late activator
	55	160	10^{-17}	159	10^{-17}	185	185	177	172	170	173	Late sigma
	45	222	10^{-26}	223	10^{-28}	228	228	228	224	221	225	Late activator
Translation, RNA	*regA*	148	10^{-34}	140	10^{-34}	122	125	120	118	132	122	Repressor
	rnlA	—	—	—	—	374	374	389	383	381	389	RNA ligase 1
	pseT	—	—	—	—	301	299	292	295	305	305	Polynucleotide kinase
	tRNAs	×1	—	—	—	×8	×2	—	×17	×30	×24	tRNA
Nucleotide metabolism	*frd*	—	—	—	—	193	195	189	184	181	172	Dihydrofolate reductase; product used by *td*
	nrdD	—	—	—	—	606	605	620	608	611	704	Anaerobic NTP reductase
	td	212	10^{-84a}	211	10^{-78a}	286	238	410	279	300	277	Thymidylate synthase
	nrdH	—	—	—	—	102	90	89	91	79	94	Anaerobic glutaredoxin
	cd	—	—	—	—	193	169	168	172	150	182	dCMP deaminase
	tk	—	—	—	—	193	193	198	191	194	199	Thymidine kinase
	nrdA	776	0	769	0	754	751	747	570	742	796[b]	Aerobic NDP reductase (alpha)
	nrdB	385	10^{-104}	388	10^{-104}	388	390	386	323	374	376	Aerobic NDP reductase (beta)
	nrdC	83	10^{-5a}	59	0.002[a]	87	92	93	75	99	90	Aerobic thioredoxin
	1	—	—	—	—	241	244	218	224	212	229	dNMP kinase
	dam	265	10^{-24a}	276	10^{-43a}	260	—	270	?	?	?	N6-methyladenine
	42	323	10^{-75a}	327	10^{-79a}	246	—	—	231	?	?	N4-methylcytosine; produces hydroxymethyl cytosine in T4
Replication, recombination, repair	*rnh*	—	—	—	—	305	290	315	307	335	306	RNase H
	59	199	10^{-16}	192	10^{-15}	217	175	215	219	216	215	Helicase loader
	32	273	10^{-42}	308	10^{-45}	301	299	322	295	304	302	Single-stranded DNA binding protein
	uvsX	336	10^{-29}	337	10^{-26}	390	390	356	?	366	411	RecA-like protein
	41	461	10^{-93}	393	10^{-88}	475	458	470	469	466	485	DNA helicase
	61	324	10^{-48}	278	10^{-36}	342	340	342	334	352	344	Primase
	dexA	—	—	—	—	227	225	222	221	230	226	Exonuclease A
	39	—	—	—	—	516	606	607	607	601	613	Topoisomerase II
	30	—	—	—	—	487	497	498	501	447	495	DNA ligase
	47	347	10^{-41}	344	10^{-41}	339	339	341	355	346	342	Recombination nuclease
	46	571	10^{-84}	573	10^{-84}	560	562	560	570	745	772	Recombination nuclease
	45	222	10^{-26}	223	10^{-28}	228	228	228	223	221	225	Clamp
	44	314	10^{-51}	293	10^{-37}	319	320	324	319	318	321	Clamp loader
	62	111	10^{-8}	135	10^{-43}	187	187	192	192	163	193	Clamp loader
	43	832	10^{-142}	827	10^{-149}	898	903	892	889[b]	850	919	DNA polymerase
	49	122	10^{-4}	—	—	157	157	157	157	151	161	Recombination endonuclease VII
	dda	397	10^{-8}	424	10^{-9}	439	437	463	439	421	454	DNA helicase
	denV	—	—	185	0.46[a]	138	137	—	152	138	?	N-glycosidase; pyrimidine dimer repair
	52	—	—	—	—	442	441	454	555	428	438	Topoisomerase II
	uvsW	488	10^{-82}	488	10^{-75}	587	504	500	493	507	503	RNA-DNA helicase
	uvsY	143	10^{-9}	99	10^{-4}	137	164	135	?	144	136	UvsX assistant; SSB
Capsid genes	*24*	—	—	—	—	427	427	413	410	298	392	Vertex precursor
	2	—	—	—	—	274	273	273	268	198	332	DNA end binding
	4	146	10^{-33}	109	10^{-23}	150	149	157	150	151	156	Head completion
	13	282	10^{-10}	269	10^{-14}	309	308	310	307	307	306	Head completion
	14	471	10^{-23}	309	10^{-26}	256	254	246	252	278	262	Head completion
	16	144	10^{-5}	145	10^{-5}	165	164	165	154	219	172	Terminase subunit
	17	548	10^{-96}	551	10^{-97}	610	611	607	613	601	633	Terminase subunit
	20	559	0	538	0	624	523	521	516	515	521	Portal vertex protein
	67	—	—	—	—	80	78	76	69	55	53	Prohead core protein
	68	—	—	—	—	141	141	135	140	163	157	Prohead core protein
	21	217	10^{-68}	215	10^{-77}	212	213	231	210	213	209	Prohead protease
	22	367	10^{-54}	334	10^{-46}	269	270	264	274	283	264	Prohead core protein
	23	471	10^{-139}	463	10^{-149}	521	522	528	529	514	534	Head protein
	inh	—	—	—	—	226	222	242	244	163	232	Head protease inhibitor
	hoc	222	10^{-4}	—	—	376	471	404	180	?	?	Highly immunogenic capsid protein
Tail, tail fiber genes	*37*	—	—	—	—	1,026	1,103	979	1,420[c]	1,085/1,094	1,358/1,303	Distal long tail fiber
	3	178	10^{-8}	184	10^{-5}	176	194	196	175	177	189	Tail sheath stabilizer
	26	240	10^{-13}	237	10^{-13}	208	208	209	204	282	258	Baseplate hub
	48	387	2.2	357	P-SSM2	364	369	352	342	379	358	Baseplate tube cap

Table 4. Continued

Category	Gene	P-SSM2	e-Value	P-SSM4	e-Value	T4 Supergroup Phages						Description of Protein Product
						T4	RB69	RB49	44RR2.8t	KVP40	Aeh1	
	53	242	10^{-4}	324	0.003	196	191	184	179	192	188	Baseplate wedge
	5	753	0.025	770	10^{-8}	575	577	600	600	421	604	Baseplate hub
	25	134	10^{-8}	140	10^{-8}	132	132	128	127	139	140	Baseplate wedge
	6	648	10^{-46}	663	10^{-34}	660	656	634	627	646	651	Baseplate wedge
	7	5,196	7.9	7,313	0.005	1,032	1,032	3,087	1,019	1,165	1,163	Baseplate wedge
	8	534	e^{-6}	511	10^{-6}	334	334	331	328	343	328	Baseplate wedge
	9	1,095	Fig. 6	410	10^{-4}	288	287	284	285	327	308	Baseplate; socket
	10	—	—	—	—	602	604	600	604	748	718	Baseplate; pin
	11	—	—	—	—	219	219	214	220	234	324	Baseplate; pin
	12	—	—	—	—	527	516	466	466	512/473	436	Short tail fiber
	wac	—	—	—	—	487	480	589	587	559	1,035	Whiskers
	15	281	10^{-26}	311	10^{-23}	272	258	277	272	450	275	Tail sheath stabilizer
	18	730	10^{-95}	750	10^{-101}	659	660	666	663	671	679	Tail sheath
	19	196	10^{-32}	197	10^{-35}	163	163	164	162	166	162	Tail tube
	35	—	—	—	—	372	374	379	377	894	1,312	Fiber hinge
	34	—	—	—	—	1,289	1,277	1,246	1,222	1,290	1,236	Proximal long tail fiber
Chaperonins, catalysis genes	*rnlA*	—	—	—	—	374	374	389	383	381	389	Fiber attachment
	31	—	—	—	—	111	110	107	136	112	136	Cochaperonin for head assembly
	57A	—	—	—	—	80	76	86	78	89	77	Chaperone for long and short tail fibers
Lysis exclusion	*rIIA*	—	—	—	—	725	737	702	705	689	732	Lysis inhibition
	rIIB	—	—	—	—	312	311	330	377	345	438	Lysis inhibition
Other	*57B*	—	—	—	—	152	151	154	142	151	164	Unknown function, conserved

Table modified from [22,104]. The T4 supergroup is divided into T-evens (e.g., T4 and RB69), pseudo T-evens (e.g., RB49 and 44RR2.8t), Schizo T-evens (e.g., Aeh1), and the Exo T-evens (e.g., S-PM2) [106,107]. For previously published T4 supergroup phages, only the size (amino acids) of selected predicted coding regions are presented using gene names according to T4 terminology. For P-SSM2 and P-SSM4, the size of each translated gene and the e-value of the best phage–T4-like (or microbe-related see below) e-value is presented; Where no e-value is given, these ORFs were assigned based upon size, domain homology, and synteny except where "Fig.6" is listed, which refers to designations made using tail fiber analyses summarized in Figure 6, and P-SSM2 or P-SSM4 indicates designation made through paralogy. A long dash indicates the lack of a particular gene.

[a]The best e-value was microbe-related rather than related to T4-like phages.

[b]The gene is split into two segments, often by an intron or homing endonuclease.

[c]The gene is fused.

DOI: 10.1371/journal.pbio.0030144.t004

proximal-end subunit (gp34) anchoring the fiber to the phage baseplate and a distal-end subunit (gp37) responsible for host recognition and attachment (reviewed in [48]). Thus P-SSM2 and P-SSM4 ORFs contained regions similar to T4-like phage distal tail fiber genes (gp37; P-SSM2 orf023, orf033, orf295, and orf298; P-SSM4 orf087) and proximal tail fiber genes (gp34; P-SSM2 orf295 and orf315; P-SSM4 orf026 and orf087). Further, two P-SSM2 ORFs (orf034 and orf315) and a P-SSM4 ORF (orf027) are similar to other known tail fiber genes, albeit with low sequence similarity, and for only a small portion of the ORF.

Second, ORFs containing repeat sequences were added to the pool of possible tail fiber genes. Both simple (amino acid triplets) and complex (longer amino acid motifs) repeats are associated with phage tail fiber genes [49,50]. Simple repeats are found in two P-SSM2 ORFs (orf23 and orf28; Figure 6), with nearly 49% of orf028 encoding the simple triplet repeat Gly-X-Y (where X and Y are often proline, serine, or threonine). Proteins with extended runs of these collagen-like amino acid motifs are thought to fold into trimeric coiled coils, consistent with a tail-fiber-like structure [50]. Complex repeat motifs of 15 to 51 amino acids in length are found in P-SSM2 (orf111 and orf298) and P-SSM4 (orf087; Figure 6). Some of these motifs are similar to those found in the long distal tail fiber (gp37) and short tail fiber (gp12) genes in T4, where they encode tandem, beta-strand-rich, supersecondary structural elements that are correlated with the beaded or knobbed shaft structure of these tail fibers [49,51].

Third, possible tail-fiber-encoding ORFs were identified through paralogy to other *Prochlorococcus* phage tail fiber ORFs already identified (Figure 6). This approach follows the observation of homology between three T4 tail fiber genes (gp12, gp34, and gp37) [49], which are thought to have arisen via gene duplication events [52]. These analyses added four ORFs to the pool of possible tail fiber genes for P-SSM2 (orf021, orf022, orf293, and orf301) and two for P-SSM4 (orf080 and orf082).

After identification of a pool of putative tail fiber genes, we used sequence similarity to known tail fiber and/or baseplate genes as a guideline to annotate ORFs according to the known T4 phage architecture. Three tail-fiber-like ORFs of P-SSM2 (orf111, orf295, and orf298) have N-terminal domains that are similar to T4 baseplate proteins (Figure 6). In T4, the N-terminus of the proximal long tail fiber (gp34) is bound to the baseplate via the baseplate protein gp9 and possibly gp10 [53,54,55]. The N-terminus of P-SSM2 orf298 is similar to the P-SSM4 orf081 (a gp9 homolog by sequence), suggesting that P-SSM2 orf298 could be analogous to a T4 proximal long tail fiber subunit (gp34), albeit fused to the baseplate socket in P-SSM2. Although such a fused protein does not appear to exist for the other myophage, P-SSM4, the adjacent reading frame to orf081 encodes a possible tail fiber ORF with significant

Table 5. Summary Table of Unique Features of *Prochlorococcus* Cyanophage Genomes That Are Uncommon among Known Phages

Functional Category	Genes	Putative Function	P-SSP7	e-Value	Marine T7-Likes	P-SSM2	e-Value	P-SSM4	e-Value	Marine T4-Likes
Phosphate	*pstS*	Phosphate uptake				322	e^{-136}	322	e^{-137}	
	phoH	Phosphate-stress-induced			+	251	e^{-24}	258	e^{-20}	+
Carbon mobilization	*talC*	MipB/TalC family transaldolase	215	e^{-43}		216	e^{-47}	218	e^{-53}	+
Lysogeny	*int*	Phage integration	291	e^{-13}						
Nucleotide metabolism	*mazG*	pyrophosphohydrolase/ pyrophosphatase				139	e^{-11}	134	e^{-27}	
	pyrE	Orotate phosphoribosyltransferase				215	e^{-44}			
	purH	Phosphoribosylformyl glycinamide synthase				108	e–7			
	purL	Phosphoribosyl formyl glycinamide cyclo-ligase				223	e^{-80}			
	purM	AICARFT/IMPCHase bienzyme				314	e^{-97}			
	purN	phosphoribosyl glycinamide formyltransferase				175	e^{-33}			
	nrd	RNR domain	469	e^{-11}	+	universal among T4-like phages				
Photosynthesis-related genes	*psbA*	D1 protein, PSII	360	e=0		361	e = 0	366	e = 0	+
	hli	Thylakoid-associated proteins	×1 *hli* gene			× 6 *hli* genes		×4 *hli* genes		+
	petE	Plastocyanin, PET				115	e^{-21}			
	petF	Ferredoxin, PET				98	e^{-29}			
	pebA	Phycoerythrobilin biosynthesis				234	e^{-12}			
	ho1	Heme biosynthesis				234	e^{-63}			
	psbD	D2 protein, PSII						359	e = 0	+
	speD	Polyamine biosynthesis						102	e^{-17}	
	pcyA	Phycocyanobilin biosynthesis						230	e^{-26}	
Other functions	*cobS*	Vitamin B12 biosynthesis				365	e^{-21}	365	e^{-23}	
	prnA	Bacterial tryptophan halogenase				486	e^{-52}			
	nol	Carbomoyltransferase				572	e^{-52}			
	hn	HN						158	e^{-33}	
	LPS	Epimerases, transferases, phospholipases				×24 genes				

Non-marine T7-like/T4-like phages completely lack these genes. The size (amino acids) and best BLASTp e-value of each predicted coding region are presented using gene names and function assignments according to their function in cellular organisms. The *hli* genes were assigned using e-value and a signature sequence as reported in Lindell et al. [14]. A plus sign indicates that the feature is present in the phage group, otherwise the feature is absent or is yet to be identified. PET, photosynthetic electron transport; PSII, photosystem II reaction center.

DOI: 10.1371/journal.pbio.0030144.t005

similarity to C-terminal stretches of P-SSM2 orf298. Thus, it appears that P-SSM4 orf081 and orf082 are orthologous with the PSSM2 orf298 N- and C-terminal regions, respectively. P-SSM2 orf295 also appears to be a tail fiber fused to a baseplate protein, gp10, which, in T4, may also play a role in binding tail fiber proteins, although this role is less clear. Similarly, the very large homologous genes (>15,000 nt) P-SSM2 orf113 and P-SSM4 orf080 appear fused to baseplate wedge initiator (gp7) homologs, which are not known to bind tail fiber in T4 [53]. Regardless of their precise assignments relative to T4 tail fiber genes, these putative fusions likely encode tail fiber subunits that bind directly to the baseplate through incorporation of their N-termini into the baseplate complex. Assuming that the long tail fibers of P-SSM2 or P-SSM4 are composed of more than one kind of protein subunit, as in T4 [48], we hypothesize that these baseplate-domain-containing tail fibers are unlikely to determine host specificity, but rather are analogous to the proximal long tail fiber (gp34) or short tail fiber (gp12) of T4.

Thus we identify a pool of 12 and five putative tail-fiber-related genes (awaiting experimental confirmation) in the P-SSM2 and P-SSM4 genomes, respectively. Some are quite large relative to those in T4, whereas others appear fused to baseplate genes, which has not been observed for the T4-like phages.

Metabolic Genes Uncommon among Phages

All three cyanophages contained genes that are not commonly found in phages. We have selected the following cyanobacterial genes for discussion because we hypothesize that they could play defining functional roles in the marine cyanophage–cyanobacterium phage–host system.

Photosynthesis-related genes in cyanophages. We previously reported photosynthesis-related genes (*psbA* and *hli*) in all three of these *Prochlorococcus* phages, as well as other photosynthesis genes (*petE, petF,* and *psbD*) in one of the two *Prochlorococcus* myovirus genomes [14]. In addition, genomic analyses have revealed that P-SSM2 contains *pebA* and *ho1*, whereas P-SSM4 contains *pcyA* and *speD* (see Table 5). In cyanobacteria these genes are involved in phycobilin biosynthesis (*ho1*, *pebA*, and *pcyA*) [56,57] and polyamine biosynthesis (*speD*). Although the phycobilin biosynthesis genes are found in *Prochlorococcus* [4,34], their function is unclear because *Prochlorococcus* does not have the intact phycobilisomes characteristic of most cyanobacteria. These genes are thought to be a remnant of the evolutionary reduction of the phycobilisome-based antenna to a chlorophyll-b-based antenna [4,58,59,60]. Although low levels of phycoerythrin occur in some LL *Prochlorococcus* strains [61], they have, as yet, no known function in the host.

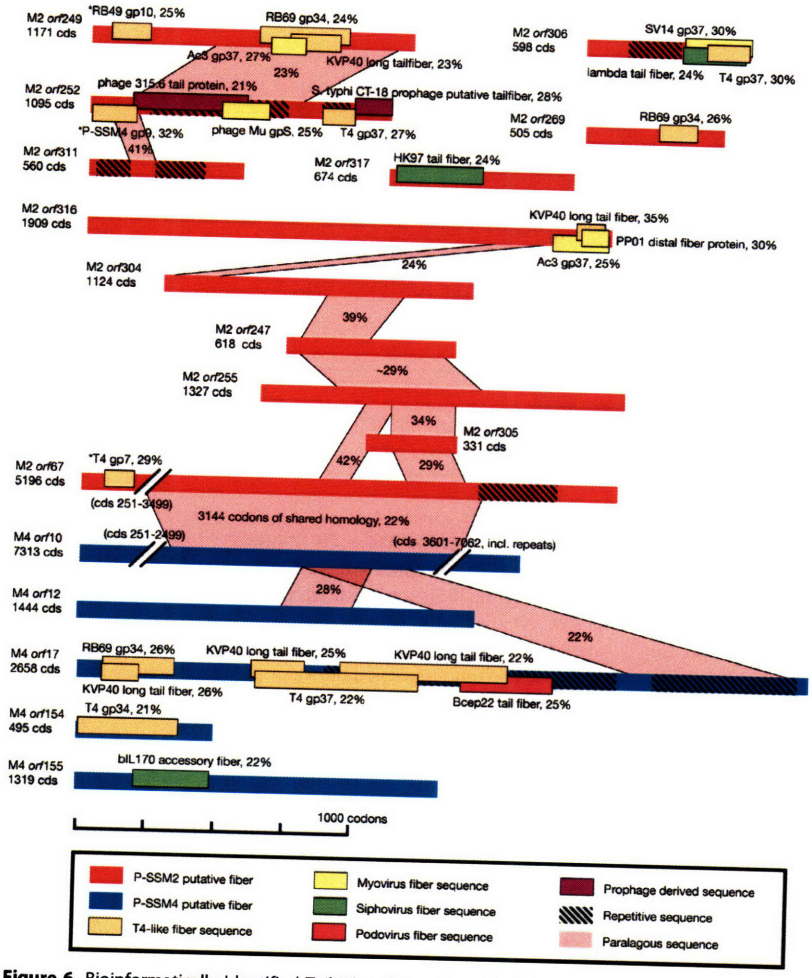


Figure 6. Bioinformatically Identified Tail Fiber Genes from *Prochlorococcus* Myoviruses

Red bars indicate P-SSM2 ORFs (labeled as M2); blue bars indicate P-SSM4 ORFs (labeled as M4). Due to space constraints, P-SSM2 orf67 and P-SSM4 orf10 are broken as indicated.

DOI: 10.1371/journal.pbio.0030144.g006

The polyamine biosynthesis gene *speD* found in the phage has a homolog in all of the marine cyanobacteria with complete genome sequences. Although its function has not been confirmed in these organisms, SpeD is known to catalyze the terminal step in polyamine synthesis in other prokaryotes, and polyamines affect the structure and oxygen evolution rate of the photosystem II (PSII) reaction center in higher plants [62]. Therefore, SpeD, if expressed, may play a role in maintaining the host PSII reaction center during phage infection.

Nucleotide metabolism genes. The podovirus P-SSP7 contains an ORF (orf20) with a putative ribonucleotide reductase (RNR) domain (see Table 5). In prokaryotes and T4-like phages, RNRs provide the building blocks for DNA synthesis through catalyzing a thioredoxin-mediated reduction of diphosphates (e.g., rNDP → dNDP) during nucleotide metabolism [63]. Among T7-like genomes, these domains have been observed only in marine phages (see Table 5) including cyanophage P60 and roseophage SIO1 [17,20]. An examination of the two genes (*nrdA* and *nrdB*) in P60 that contain homology to RNRs suggests that they represent a split RNR (as described earlier for DNAP): *nrdA* is similar to the 5′-end and *nrdB* is similar to the 3′-end of cyanobacterial class II RNRs (data not shown). When analyzed for the presence of a class II RNR diagnostic motif [64], all three marine T7-like phage putative RNRs were found to contain homology to this motif (seven of nine residues in SIO1, P-SSP7; eight of nine residues in P60; as compared to eight of nine residues in the marine cyanobacteria) (Figure S1). Furthermore, the putative RNRs are located in the genomes at the distal end of a region homologous to the nucleotide metabolism region in T7 [65]. It is plausible that T7-like phage infection in phosphorus-limited environments requires extra nucleotide-scavenging genes.

Both *Prochlorococcus* myoviruses contain the alpha and beta RNR subunits that are found in all known T4-like phages (see Table 4). The genes have closer sequence homology to those in T4-like phages than cyanobacterial hosts (Figure S2). Interestingly, our myoviruses also contain a noncyanobacterial *cobS* gene, which has never been found in phages. This

Table 6. Signature Cyanophage Genes?

Phage Family	Host	Cyanophage	Genes							Reference
			psbA	*hli*	*talC*	*int*	*phoH*	*pstS*	*cobS*	
Podoviridae	*Synechococcus*	P60	-	-	-	-	-	-	-	16
	Prochlorococcus	P-SSP7	+	+	+	+	-	-	-	This study
Myoviridae	*Prochlorococcus*	P-SSM2	+	+	+	-	+	+	+	This study
		P-SSM4	+	+	+	-	+	+	+	This study
	Synechococcus	S-PM2	+	+	-	-	+	-	+	108
		S-RSM2[a]	+	?	+	?	?	?	?	22
		S-BM4 [a]	+	+	?	?	?	?	?	22
		S-WHM1 [a]	+	?	?	?	?	?	?	22
		S-RSM88 [a]	+	?	?	?	?	?	?	22

There are genes that are not commonly found in phages, but are commonly found among the limited cyanophage sequences available.

[a] These phage genomes were not completely sequenced, but were part of a study that did targeted analyses of ~5kb regions surrounding the *psbA* gene. A question mark indicates that the presence or absence of the feature is unknown.

DOI: 10.1371/journal.pbio.0030144.t006

gene encodes a protein that catalyzes the final step in cobalamin (vitamin B12) biosynthesis in bacteria [66,67], and cobalamin is an RNR cofactor during nucleotide metabolism in cyanobacteria [68]. Both physiological assays [69,70] and genomic evidence [4,34] indicate that *Prochlorococcus* synthesizes its own cobalamin. It is tempting to speculate that the phage *cobS* gene serves to boost cobalamin production in the host during infection, thus improving the activity of RNRs. However, these phage RNRs clearly contain the α2 and β2 subunits (typical of class I RNRs) and lack the class II motif described earlier. Thus, if the phage *cobS* does increase cobalamin production and if this production increase is important, then either the phage class I RNRs are cobalamin dependent (which is unprecedented) or cobalamin must be useful for some other process.

Carbon metabolism genes. In cyanobacteria, the pentose phosphate pathway oxidizes glucose to produce NADPH for biosynthetic reactions (oxidative branch) and ribulose-5-phosphate for nucleotides and amino acids (non-oxidative branch). This pathway (both branches) is particularly important in cyanobacteria for metabolizing the products of photosynthesis during dark metabolism [71]. Long ago, it was hypothesized that cyanophages utilize this pathway as a source of energy and carbon when the host is not photosynthesizing [72]. Interestingly, genomic sequencing has recently revealed that *Synechococcus* cyanophage S-RSM2 [16] and the *Prochlorococcus* cyanophages P-SSM2 and P-SSM4 [14] contain a transaldolase gene *(talC)*. In *Escherichia coli*, transaldolase is a key enzyme in the non-oxidative branch of the pentose phosphate pathway [73]. It has been suggested that the product of the phage *talC* gene may facilitate phage access to stored carbon pools during the dark period [16].

Recent work in *E. coli* has revealed two genes *(mipB/fsa* and *talC)* that are divergent from the bona fide transaldolases (*talA* and *talB*) [74], but encode a structurally similar enzyme [75]. Members of this new subfamily (MipB/TalC) of aldolases, which have a striking sequence similarity to each other, can have distinctly different functions, acting either as a transaldolase or fructose-6-phosphate aldolase, but not both [74]. All three of the genes previously reported as "transaldolase" genes in cyanophages [14,16], as well as an ORF in the podovirus P-SSP7, are most similar to these MipB/TalC aldolase genes (see Table 5; Figure S3). The translated cyanophage genes contain 26 (P-SSM2), 28 (P-SSP7 and S-RSM2), and 29 (P-SSM4) of 32 diagnostic (as designated by Thorell et al. [75]) amino acid residues (Figure S4). In the active site of this enzyme, as inferred from the crystal structure of *E. coli* fructose-6-phosphate aldolase, eight of 14 residues are not conserved between the MipB/TalC subfamily, varying depending on enzyme specificity (fructose-6-phosphate aldolase versus transaldolase) [75]. When aligned with MipB/TalC members of known substrate specificity, the cyanophage putative active site residues match all eight of those enzyme sequences with transaldolase activity (Figure S4). Thus, it appears that each of the four cyanophage *talC* genes encodes an enzyme with transaldolase activity. If functional, these genes are likely to be important for metabolizing carbon substrates—which is central to biosynthesis and energy production—during phage infection of cyanobacterial hosts.

Phosphate stress genes in the myoviruses. Phosphorus is a scarce resource in the oligotrophic oceans [76,77]. It is often growth limiting for cyanobacteria [78] and is required in significant amounts for phage replication. Thus it is perhaps not surprising that the phosphate-inducible *phoH* gene, which has been found in two marine phage genomes [20,21], is also found in both *Prochlorococcus* myoviruses (see Table 5; see Figures 3 and 4). Although the *phoH* gene is found widely distributed among both eubacteria and archaea [79], including all cyanobacteria, and is known to be induced under phosphate stress in *E. coli* [80], its function has not been experimentally determined. Bioinformatic analyses suggest that these *phoH* genes are part of a multi-gene family with divergent functions from phospholipid metabolism and RNA modification (COG1702 *phoH* genes) to fatty acid beta-oxidation (COG1875 *phoH* genes) [79].

Both P-SSM2 and P-SSM4 also contain a phosphate-inducible *pstS* gene—which is also widespread among the archaea and eubacteria, including all known cyanobacteria—that has not been reported in phages. In bacteria, the *pstS* gene encodes a periplasmic phosphate-binding protein involved in phosphate uptake [81]. If expressed by the phage, it might serve to enhance phosphorus acquisition during infection of phosphate-stressed cells.

LPS biosynthesis genes in P-SSM2. The myovirus P-SSM2 contains 24 LPS genes that form two major clusters in the genome (see Figure 3). Reports of phage-encoded LPS genes have previously been limited to temperate phages [82]. Such temperate phage LPS genes are thought to be used during infection and establishment of the prophage state to alter the cell-surface composition of the host, preventing other phages from attaching to the host cell. Although T4-like phages are commonly thought of as lytic phages, the lytic process can be stalled upon infection (sometimes termed "pseudolysogeny") during suboptimal host growth [83]. If this phenomenon occurs in marine phages, as has been suggested [22,84,85], then a phage-encoded LPS gene cluster, even in a lytic phage, might maintain a similar functional role.

Signature genes for oceanic cyanophages? Although data are too limited to be conclusive (Table 6), some of the host genes that appear common in oceanic cyanophages may ultimately represent signature genes for these phages. For example, the genomes of all three cyanophages presented here and five partial genomes (<5 kb) of *Synechococcus* cyanomyophages presented by Millard et al. [16] all contain a *psbA* gene. Further, all three cyanophages presented here contain at least one *hli* and a *talC* gene, and both myoviruses presented here are unique among the phages in that they contain *pstS* and *cobS* (Table 6). As more phages are sequenced, will we find that these genes are specifically characteristic of oceanic cyanophages? If true, this would provide us with a powerful tool for studying these phages in the wild because quantitative PCR could be used to differentiate between cyanophages and other phages in environmental samples.

Hypothesized Transient Genes

There are genes of interest, found in only one of the myoviruses, that we hypothesize are not functional, but rather were obtained by cyanomyophages through packaging random DNA, probably by illegitimate recombination [86,87] with DNA from a common phage genome pool [88].

Trytophan halogenase. P-SSM2 contains a gene *(prnA)* that is known to exist in only nine species of bacteria, in which it encodes a tryptophan halogenase that catalyzes the NADH-consuming first step of four that are involved in converting tryptophan to the antibiotic pyrrolnitrin [89,90,91]. Although this gene is full length (Figure S5), *prnA* is part of a unique metabolic pathway missing in most bacteria, including cyanobacteria.

Archaeal and eukaryotic genes. The other myovirus, P-SSM4, contains three grouped genes with homology only to eukaryotic prion-like proteins (orf32), an archaeal protease (orf35), and a hypothetical protein from a eukaryotic slime mold (orf36) (see Figure 4). Other eukaryotic and prion-like genes have been predicted in the genomes of mycobacteriophages that infect actinobacterial hosts [92], although they have no similarity to those found in P-SSM4.

Hemagglutinin neuraminidase. P-SSM4 contains a possible hemagglutinin neuraminidase (HN), which has only been observed in single-stranded RNA (ssRNA) viruses and *Prochlorococcus* MED4 (orf1400). In ssRNA viruses, HN cleaves sialic acid from glycolipids on the host cell surface, which enables these viruses to attach. Protein alignments show, however, that both the MED4 and P-SSM4 HN genes are only partial genes—they are missing the N- and C-termini (approximately 200 amino acids)—relative to other ssRNA HNs (Figure S6). It is noteworthy that the HN gene occurs nowhere else in the prokaryotic world except for MED4. Could this gene have been obtained by P-SSP7 through the phage genome pool (sensu Hendrix et al. [88]), then transferred to MED4? This postulate is buttressed by the observation that the HN gene in MED4 is found next to three *hli* genes (which encode high-light-inducible proteins)—genes which we have argued earlier are susceptible to horizontal gene transfer in this phage–host system [14].

Ecological and Evolutionary Implications of Phages Carrying Host Genes

Prochlorococcus cells are slow-growing (doubling times range from 1 to 10 d), oxygenic phototrophs that thrive in nutrient-poor, aerobic surface waters [1]—conditions that are fundamentally different from those of most of the host cells of the phages sequenced to date. Thus, oceanic cyanophages are subject to substantially different selective pressures than most other sequenced phages in the database. The presence in these phages of host genes that are likely involved in the maintenance of photosynthesis, response to phosphate stress, and mobilization of carbon stores during infection may be interpreted as evidence of such unique pressures (see Table 5).

If phage genomes interact as "local neighborhoods" (sensu Hendrix et al. [88]) within a "global phage metagenome" (sensu Rohwer [93]), one would expect to find biologically cohesive units akin to species, defined by local gene transfers as proposed for "microbial species" [94]. Such cohesive units would be characterized by core genes that determine a general phage infection lifestyle (e.g., T4-like or T7-like), as well as host-specific genes within phages that infect similar hosts. Indeed, 26 and 75 such core genes exist among the T7-like and T4-like phages, respectively (see Tables 3 and 4), and host-specific genes abound among these cyanophages (see Figures 1C, 5A, and 5B). That these core genes represent mostly morphological and DNA replication genes suggests a T7-like or T4-like lifestyle that would involve a specific means of delivering DNA from host to host (in a tailed, capsid structure) as well as converting the host into a phage factory. Based upon the presence of many such core genes in our *Prochlorococcus* phages, one would predict they would behave as T7-like (P-SSP7; although probably with the ability to integrate into its host) and T4-like phages (P-SSM2 and P-SSM4) during cyanobacterial infection.

Beyond these core genes, our *Prochlorococcus* phages contain many "nonphage" genes that are of greatest sequence similarity to cyanobacterial genes (see Figures 1C, 5A, and 5B). We speculate that the acquisition and use of some host genes by phages plays an important role in phage ecology, even *shaping* the evolution of the phage host range. The initial host range alterations are likely to occur by phage tail fiber switching [47], but beyond that, these co-opted host genes could either shift or expand the phage's host range depending upon whether they affect fitness of the phage in the original hosts. Understanding this dynamic fitness landscape will require modeling efforts directed by a thorough knowledge of the mechanisms and relative rates for this complex genetic shuffling—factors that likely underpin the complexity of phage–host interactions in the environment.

Materials and Methods

Electron microscopy. *Prochlorococcus* phages were concentrated using ultracentrifugation. Concentrates were prepared for microscopy by spotting phage lysates onto freshly glow-discharged carbon/formvar-coated copper grids. Grids were negatively stained with 1% uranyl acetate, dried, and viewed in a JEOL (Peabody, Massachusetts, United States) 1200 EXII transmission electron microscope operated at 80 kV.

Preparation of cyanophages for genome sequencing. Three *Prochlorococcus* phages were chosen for sequencing based upon their host ranges, which were restricted to *Prochlorococcus* hosts (see Introduction).

Phages were prepared for genomic sequencing as previously described [14,95]. Briefly, phage particles were concentrated from phage lysates using polyethylene glycol. Concentrated DNA-containing phage particles were purified from other material in phage lysates using a density cesium chloride gradient. Purified phage particles were broken open (SDS/proteinase K), and DNA was extracted (phenol:chloroform) and precipitated (ethanol) yielding small amounts of DNA (<1 μg). A custom 1- to 2-kb insert linker-amplified shotgun library was constructed by Lucigen (Middletown, Wisconsin, United States) as described previously [95]. Additional larger insert (3–8 kb) clone libraries were constructed from genomic DNA by the Department of Energy (Joint Genome Institute, Walnut Creek, California, United States) using a similar protocol to provide larger scaffolds during assembly. Inserts were sequenced by the Department of Energy Joint Genome Institute from all of these clone libraries and used for initial assembly of these phage genomes. The Stanford Human Genome Center Finishing Group (Palo Alto, California, United States) closed the genomes using primer walking.

Gene identification and characterization. Protein coding genes were predicted using GeneMark [96] and manual curation. Translated ORFs were compared to known proteins in the nonredundant GenBank database (http://www.ncbi.nlm.nih.gov/BLAST/) and in the KEGG database (http://www.genome.ad.jp/kegg/kegg2.html) using the BLASTp program (ftp://ftp.ncbi.nih.gov/blast). Translated ORFs were also analyzed for signal sequences and transmembrane regions using the Web-based software SignalP and TMHMM, respectively (available at the CBS prediction servers; http://www.cbs.dtu.dk/services/). Where BLASTp e-values were high (>0.001) or no sequence similarity was observed, ORF annotation was aided by the use of PSI-BLAST, gene size, domain conservation, and/or synteny (gene order), the last as suggested for highly divergent genes encountered during phage genome annotation [97]. Identification of tRNA genes was done using tRNAscan-SE [98].

Taxonomy of best hits. For global genome comparison, we used BLASTp (e-values < 0.001) or manual annotation to classify to which group of organisms or phages each predicted coding sequence was most similar. In most cases this was obvious. However, approximately 2% of the coding sequences were less obvious, so we established an operational definition of "most similar" as the query sequence having e-values within four orders of magnitude of the top cluster of organismal types. For example, if a query sequence was similar to noncyanobacterial sequences with e-values of 10^{-29} to 10^{-25} and to cyanobacterial sequences with e-values of 10^{-20} or greater, then, despite sequence similarity to cyanobacterial sequences, the query would be considered noncyanobacterial.

Tail fiber gene identification. Tail fiber genes were identified by generating alignments (stand-alone Basic Local Alignment Search Tool, BLAST [99], 2.2.8 release) of conceptually translated, computationally identified ORFs from the P-SSM2 and P-SSM4 genomes against a database consisting of 33,270 sequences encompassing all known phage sequences obtained from the NCBI NR database in April 2004. Only ORFs whose alignments to known tail fiber genes were longer than 100 residues and had e-values less than 0.001 were designated as tail-fiber-like. Sequences close to this cutoff were re-aligned using the bl2seq command of BLAST, which computes e-values independently of database size. Tail-fiber-like paralogs were identified by individually aligning the set of tail-fiber-like ORFs with all other ORFs in the genomes. All ORFs with alignments greater than 100 residues and e-values less than 0.001, were designated as tail fiber paralogs. All BLAST searches and alignments were performed with the low-complexity sequence filter and default parameters. Amino acid sequence repeats were identified by self-alignment matrices using the program Dotter [100].

Sequence manipulation and phylogenetic analyses. Alignments were generated using Clustal X [101] and edited manually as necessary. PAUP V4.0b10 [102] was used for the construction of distance and maximum parsimony trees. Amino acid distance trees were inferred using minimum evolution as the objective function, and mean distances. Heuristic searches were performed with 100 random addition sequence replicates and the tree bisection and reconnection branch-swapping algorithm. Starting trees were obtained by stepwise addition of sequences. Bootstrap analyses of 1,000 resamplings were carried out. Maximum likelihood trees were constructed using TREE-PUZZLE 5.0 [103]. Evolutionary distances were calculated using the JTT model of substitution assuming a gamma-distributed model of rate heterogeneities with 16 gamma-rate categories empirically estimated from the data. Quartet puzzling support was estimated from 10,000 replicates.

Supporting Information

Figure S1. Class II RNR Motif Compared Against Cyanobacterial and Non-T4-Like Phage RNRs

A question mark indicates this sequence data is not known; a period indicates identical residue to the reference sequence; and a dash indicates a gap in the alignment. *Anab, Anabaena; Pro, Prochlorococcus; Syn, Synechococcus; Syncy, Synechocystis.*

Found at DOI: 10.1371/journal.pbio.0030144.sg001 (10 KB PDF).

Figure S2. Distance Tree of RNR Family Proteins, Including Phage Sequences from P-SSM2, P-SSM4, and P-SSP7

Sequences from P-SSM2, P-SSM4, and P-SSP7 are shown in bold. Trees were generated from 900 amino acids. Bootstrap values for distance and maximum parsimony analyses and quartet puzzling values for maximum likelihood analysis, greater than 50%, are shown at the nodes (distance/maximum likelihood/maximum parsimony). Trees were unrooted; abbreviations as in Figure S1.

Found at DOI: 10.1371/journal.pbio.0030144.sg002 (14 KB PDF).

Figure S3. Distance Tree of Tal Proteins, Including Phage Sequences from P-SSM2, P-SSM4, and P-SSP7

Sequences from P-SSM2, P-SSM4, and P-SSP7 are shown in bold. Trees were generated from 566 amino acids. Bootstrap values for distance and maximum parsimony analyses and quartet puzzling values for maximum likelihood analysis, greater than 50%, are shown at the nodes (distance/maximum likelihood/maximum parsimony). Trees were unrooted; abbreviations as in Figure S1.

Found at DOI: 10.1371/journal.pbio.0030144.sg003 (14 KB PDF).

Figure S4. Alignment of TalC Subfamily Aldolases, Including Phage Sequences from P-SSM2, P-SSM4, P-SSP7, and S-RSM2

The 32 amino acid residues suggested to be diagnostic by Thorell et al. [75] are labeled with an asterisk and shaded where identical to bona fide TalC proteins, whereas the active site residues are labeled with an "at" symbol. Note the active site residues in the cyanophage TalC sequences exclusively match those from enzymes known to have transaldolase activity rather than fructose-6 phosphate aldolase activity.

Found at DOI: 10.1371/journal.pbio.0030144.sg004 (14 KB PDF).

Figure S5. Alignment of Tryptophan Halogenase Amino Acid Sequences Deduced from Phage and Cellular Encoded *prnA* Gene Sequences

Note the phage gene appears full-length relative to the other cellular genes. *Bdellovibrio, Bdellovibrio bacteriovorus; Bordtella, Bordetella pertussis; Burkpyrro, Burkholderia pyrrocinia; Caulobacter, Caulobacter crescentus; Myxfulvus, Myxococcus fulvus; Pschloro, Pseudomonas chlororaphis; Pseud_fl, Pseudomonas fluorescens; Shewanella, Shewanella oneidensis* MR-1; *Xanaxon, Xanthomonas axonopodis; Xancamp, Xanthomonas campestris.*

Found at DOI: 10.1371/journal.pbio.0030144.sg005 (35 KB PDF).

Figure S6. Alignment of HN Amino Acid Sequences Deduced from Phage and ssRNA Viral Gene Sequences

Note the *Prochlorococcus* phage and host gene appears to contain only the central region of the gene relative to the other ssRNA viral genes.APMV6, avian paramyxovirus 6; BPIV3, bovine parainfluenza virus 3; Gparamyxovirus, goose paramyxovirus; HPIV1,2,3, human parainfluenza virus 1,2,3; ProMED4, *Prochlorococcus* MED4.

Found at DOI: 10.1371/journal.pbio.0030144.sg006 (36 KB PDF).

Accession Numbers

The GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) accession numbers for the genomes discussed in this paper are MED4 (BX548174), P-SSM2 (AY939844), P-SSM4 (AY940168), and P-SSP7 (AY939843).

Acknowledgments

We thank David Mead (Lucigen) and Chris Detter (Department of Energy Joint Genome Institute [DOE JGI]) for clone library construction from minimal DNA. The sequencing and assembly of the phage genomes was performed by the production sequencing group at the DOE JGI through the Sequence-for-Others Program under the auspices of the US DOE's Office of Science, Biological, and Environmental Research Program and the University of California, Lawrence Livermore National Laboratory, under contract number W-7405-ENG-48; Lawrence Berkeley National Laboratory under contract number DE-AC03-76SF00098; Los Alamos National Laboratory under contract number W-7405-ENG-36; and Stanford University under contract number DE-FC02-99ER62873. This research was supported by the US DOE under grant numbers DE-FG02-99ER62814 and DE-FG02-02ER63445, and the National Science Foundation under grant number OCE-9820035 (to SWC). We thank Sherwood Casjens, Drew Endy, Hector Hernandez, and Roger Hendrix for discussions about phage biology, evolution, and RNRs, as well as Virginia Rich, Debbie Lindell, and Erik Zinser for valuable comments on the manuscript.

Particular thanks go to Ian Molineux for providing access to his unpublished T7 Group review chapter and extensive suggestions on the manuscript; the teams of Henry Krisch and Jim Karam for providing data at the T4-like Genome Web site (http://phage.bioc.tulane.edu); Jim Karam and Vasiliy Petrov for their analysis of the gp5 DNAP split in P-SSP7; Luke Thompson for analytical assistance with the cyanophage transaldolase family genes; and Anca Segall for finding the 42-bp exact match sequence in P-SSP7 and *Prochlorococcus* MED4 that supported our hypothesis that the P-SSP7 integrase gene might be functional.

Competing interests. The authors have declared that no competing interests exist.

Author contributions. MBS grew, purified, and extracted DNA from the phages. Non-authors (see Acknowledgments) prepared clone libraries, sequenced the inserts, and assembled the genomes. MBS, MLC, and FR did the majority of the genome annotation, while PW evaluated tail-fiber-related genes and provided electron micrographs of the particles. MBS and SWC wrote the majority of the paper with significant contributions from all authors, as well as non-authors (detailed in the Acknowledgments). ■

References

1. Partensky F, Hess WR, Vaulot D (1999) *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. Microbiol Mol Biol Rev 63: 106–127.
2. Liu H, Nolla HA, Campbell L (1997) *Prochlorococcus* growth rate and contribution to primary production in the equatorial and subtropical North Pacific Ocean. Aquatic Microb Ecol 12: 39–47.
3. Liu H, Campbell L, Landry MR, Nolla HA, Brown SL, et al. (1998) *Prochlorococcus* and *Synechococcus* growth rates and contributions to production in the Arabian Sea during the 1995 Southwest and Northeast Monsoons. Deep-Sea Res II 45: 2327–2352.
4. Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, et al. (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424: 1042–1047.
5. Moore LR, Rocap G, Chisholm SW (1998) Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. Nature 393: 464–467.
6. 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.
7. Mann EL, Ahlgren N, Moffett JW, Chisholm SW (2002) Copper toxicity and cyanobacteria ecology in the Sargasso Sea. Limnol Oceanogr 47: 976–988.
8. Sullivan MB, Waterbury JB, Chisholm SW (2003) Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. Nature 424: 1047–1051.
9. Waterbury JB, Valois FW (1993) Resistance to co-occurring phages enables marine *Synechococcus* communities to coexist with cyanophage abundant in seawater. Appl Environ Microbiol 59: 3393–3399.
10. Suttle CA, Chan AM (1994) Dynamics and distribution of cyanophages and their effects on marine *Synechococcus* spp. Appl Environ Microbiol 60: 3167–3174.
11. Marston MF, Sallee JL (2003) Genetic diversity and temporal variation in the cyanophage community infecting marine *Synechococcus* species in Rhode Island's coastal waters. Appl Environ Microbiol 69: 4639–4647.
12. Lu J, Chen F, Hodson RE (2001) Distribution, isolation, host specificity, and diversity of cyanophages infecting marine *Synechococcus* spp. in river estuaries. Appl Environ Microbiol 67: 3285–3290.
13. Thingstad TF (2000) Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic ecosystems. Limnol Oceanogr 45: 1320–1328.
14. Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, et al. (2004) Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. Proc Natl Acad Sci U S A 101: 11013–11018.
15. Mann NH, Cook A, Millard A, Bailey S, Clokie M (2003) Marine ecosystems: Bacterial photosynthesis genes in a virus. Nature 424: 741.
16. Millard A, Clokie MR, Shub DA, Mann NH (2004) Genetic organization of the *psbAD* region in phages infecting marine *Synechococcus* strains. Proc Natl Acad Sci U S A 101: 11007–11012.
17. Chen F, Lu J (2002) Genomic sequence and evolution of marine cyanophage P60: A new insight on lytic and lysogenic phages. Appl Environ Microbiol 68: 2589–2594.
18. Scholl D, Kieleczawa J, Kemp P, Rush J, Richardson CC, et al. (2004) Genomic analysis of bacteriophages SP6 and K1–5, an estranged subgroup of the T7 supergroup. J Mol Biol 335: 1151–1171.
19. Hardies SC, Comeau AM, Serwer P, Suttle CA (2003) The complete sequence of marine bacteriophage VpV262 infecting *Vibrio parahaemolyticus* indicates that an ancestral component of a T7 viral supergroup is widespread in the marine environment. Virology 310: 359–371.
20. Rohwer F, Segall A, Steward G, Seguritan V, Breitbart M, et al. (2000) The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with nonmarine phages. Limnol Oceanogr 45: 408–418.
21. Miller ES, Heidelberg JF, Eisen JA, Nelson WC, Durkin AS, et al. (2003) Complete genome sequence of the broad-host-range vibriophage KVP40: Comparative genomics of a T4-related bacteriophage. J Bacteriol 185: 5220–5233.
22. Mann NH (2003) Phages of the marine cyanobacterial picophytoplankton. FEMS Microbiol Rev 27: 17–34.
23. Molineux I (2005) The T7 group. In: Calendar R, editor. The bacteriophages. New York: Oxford University Press. In press.
24. Molineux IJ (2001) No syringes please, ejection of phage T7 DNA from the virion is enzyme driven. Mol Microbiol 40: 1–8.
25. Kelman Z, Pietrokovski S, Hurwitz J (1999) Isolation and characterization of a split B-type DNA polymerase from the archaeon *Methanobacterium thermoautotrophicum* deltaH. J Biol Chem 274: 28751–28761.
26. Lavigne R, Burkal'tseva MV, Robben J, Sykilinda NN, Kurochkina LP, et al. (2003) The genome of bacteriophage phiKMV, a T7-like virus infecting *Pseudomonas aeruginosa*. Virology 312: 49–59.
27. Paul JH, Sullivan MB, Segall AM, Rohwer F (2002) Marine phage genomics. Comp Biochem Physiol B Biochem Mol Biol 133: 463–476.
28. Groth AC, Calos MP (2004) Phage integrases: Biology and applications. J Mol Biol 335: 667–678.
29. Nelson KE, Weinel C, Paulsen IT, Dodson RJ, Hilbert H, et al. (2002) Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. Environ Microbiol 4: 799–808.
30. Nunes-Duby SE, Kwon HJ, Tirumalai RS, Ellenberger T, Landy A (1998) Similarities and differences among 105 members of the Int family of site-specific recombinases. Nucleic Acids Res 26: 391–406.
31. 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.
32. Canchaya C, Proux C, Fournous G, Bruttin A, Brussow H (2003) Prophage genomics. Microbiol Mol Biol Rev 67: 238–276.
33. Casjens S (2003) Prophages and bacterial genomics: What have we learned so far? Mol Microbiol 49: 277–300.
34. Dufresne A, Salanoubat M, Partensky F, Artiguenave F, Axmann IM, et al. (2003) Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. Proc Natl Acad Sci U S A 100: 10020–10025.
35. Adolph KW, Haselkorn RH (1973) Isolation and characterization of a virus infecting a blue-green alga of the genus *Synechococcus*. Virology 54: 230–236.
36. Sherman LA, Connelley M (1976) Isolation and characterization of a cyanophage infecting the unicellular blue-green algae *A. nidulans* and *S. cedorum*. Virology 72: 540–544.
37. Ortmann AC, Lawrence JE, Suttle CA (2002) Lysogeny and lytic viral production during a bloom of the cyanobacterium *Synechococcus* spp. Microb Ecol 43: 225–231.
38. McDaniel L, Houchin LA, Williamson SJ, Paul JH (2002) Lysogeny in marine *Synechococcus*. Nature 415: 496.
39. Ackermann HW, Krisch HM (1997) A catalogue of T4-type bacteriophages. Arch Virol 142: 2329–2345.

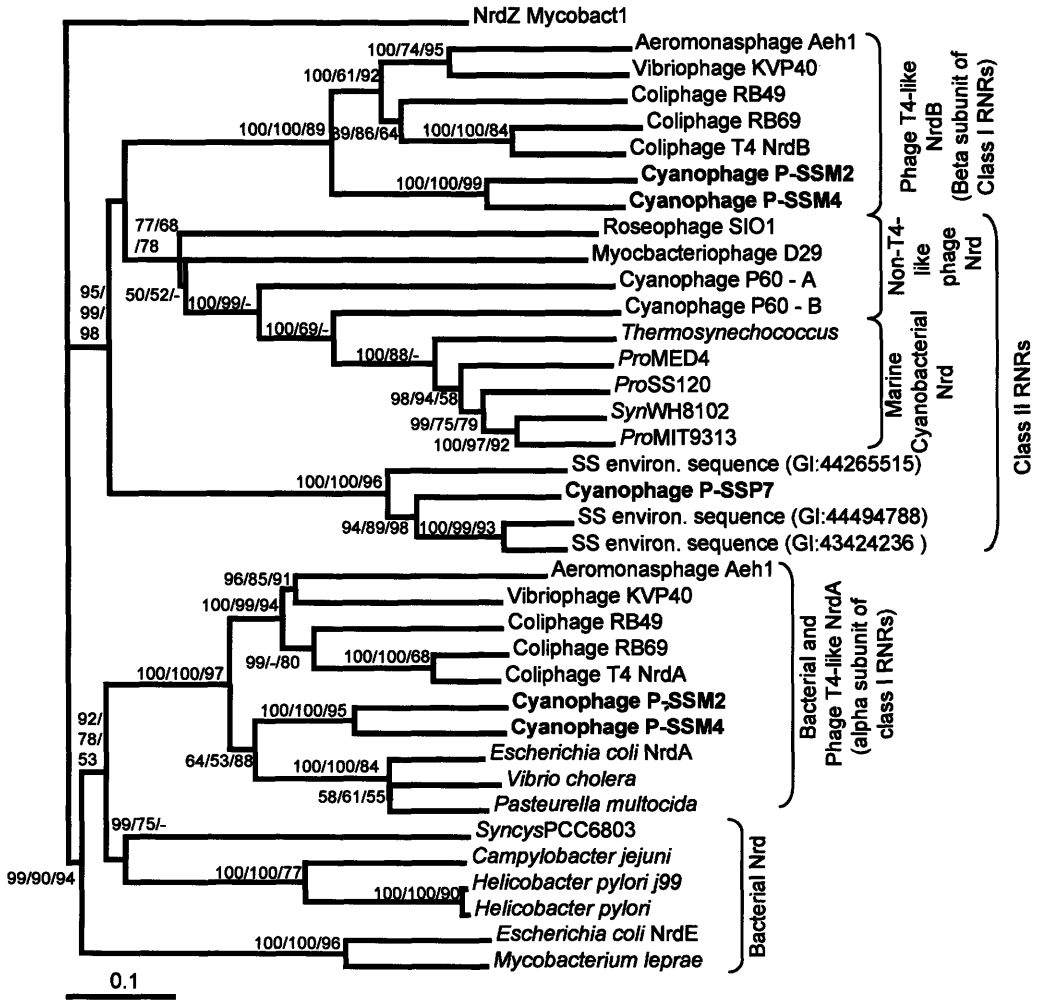
40. Keller B, Dubochet J, Adrian M, Maeder M, Wurtz M, et al. (1988) Length and shape variants of the bacteriophage T4 head: Mutations in the scaffolding core genes 68 and 22. J Virol 62: 2960–2969.
41. Volker TA, Gafner J, Bickle TA, Showe MK (1982) Gene 67, a new, essential bacteriophage T4 head gene codes for a prehead core component, PIP. I. Genetic mapping and DNA sequence. J Mol Biol 161: 479–489.
42. Volker TA, Kuhn A, Showe MK, Bickle TA (1982) Gene 67, a new, essential bacteriophage T4 head gene codes for a prehead core component, PIP. II. The construction in vitro of unconditionally lethal mutants and their maintenance. J Mol Biol 161: 491–504.
43. Zhang J, Inouye M (2002) MazG, a nucleoside triphosphate pyrophosphohydrolase, interacts with Era, an essential GTPase in *Escherichia coli*. J Bacteriol 184: 5323–5329.
44. Zhang J, Zhang Y, Inouye M (2003) *Thermotoga maritima* MazG protein has both nucleoside triphosphate pyrophosphohydrolase and pyrophosphatase activities. J Biol Chem 278: 21408–21414.
45. Haggard-Ljungquist E, Halling C, Calendar R (1992) DNA sequences of the tail fiber genes of bacteriophage P2: Evidence for horizontal transfer of tail fiber genes among unrelated bacteriophages. J Bacteriol 174: 1462–1477.
46. Xue Q, Egan JB (1995) Tail sheath and tail tube genes of the temperate coliphage 186. Virology 212: 218–221.
47. Tetart F, Desplats C, Krisch HM (1998) Genome plasticity in the distal tail fiber locus of the T-even bacteriophage: Recombination between conserved motifs swaps adhesin specificity. J Mol Biol 282: 543–556.
48. Leiman PG, Kostyuchenko VA, Shneider MM, Kurochkina LP, Mesyanzhinov VV, et al. (2000) Structure of bacteriophage T4 gene product 11, the interface between the baseplate and short tail fibers. J Mol Biol 301: 975–985.
49. Cerritelli ME, Wall JS, Simon MN, Conway JF, Steven AC (1996) Stoichiometry and domainal organization of the long tail-fiber of bacteriophage T4: A hinged viral adhesin. J Mol Biol 260: 767–780.
50. Smith MC, Burns N, Sayers JR, Sorrell JA, Casjens SR, et al. (1998) Bacteriophage collagen. Science 279: 1834.
51. van Raaij MJ, Schoehn G, Jaquinod M, Ashman K, Burda MR, et al. (2001) Identification and crystallisation of a heat- and protease-stable fragment of the bacteriophage T4 short tail fibre. Biol Chem 382: 1049–1055.
52. Kutter E, Gachechiladze K, Poglazov A, Marusich E, Shneider M, et al. (1995) Evolution of T4-related phages. Virus Genes 11: 285–297.
53. Kostyuchenko VA, Leiman PG, Chipman PR, Kanamaru S, van Raaij MJ, et al. (2003) Three-dimensional structure of bacteriophage T4 baseplate. Nat Struct Biol 10: 688–693.
54. Kostyuchenko VA, Navruzbekov GA, Kurochkina LP, Strelkov SV, Mesyanzhinov VV, et al. (1999) The structure of bacteriophage T4 gene product 9: The trigger for tail contraction. Structure Fold Des 7: 1213–1222.
55. King J (1968) Assembly of the tail of bacteriophage T4. J Mol Biol 32: 231–262.
56. Frankenberg N, Lagarias JC (2003) Phycocyanobilin:ferredoxin oxidoreductase of *Anabaena* sp. PCC 7120. Biochemical and spectroscopic. J Biol Chem 278: 9219–9226.
57. Frankenberg N, Mukougawa K, Kohchi T, Lagarias JC (2001) Functional genomic analysis of the HY2 family of ferredoxin-dependent bilin reductases from oxygenic photosynthetic organisms. Plant Cell 13: 965–978.
58. Ting CS, Rocap G, King J, Chisholm SW (2002) Cyanobacterial photosynthesis in the oceans: The origins and significance of divergent light-harvesting strategies. Trends Microbiol 10: 134–142.
59. Ting CS, Rocap G, King J, Chisholm SW (2001) Phycobiliprotein genes of the marine photosynthetic prokaryote *Prochlorococcus*: Evidence for rapid evolution of genetic heterogeneity. Microbiology 147: 3171–3182.
60. Hess WR, Rocap G, Ting CS, Larimer FW, Stilwagen S, et al. (2001) The photosynthetic apparatus of *Prochlorococcus*: Insights through comparative genomics. Photosynth Res 70: 53–71.
61. 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.
62. Bograh A, Gingras Y, Tajmir-Riahi HA, Carpentier R (1997) The effects of spermine and spermidine on the structure of photosystem II proteins in relation to inhibition of electron transport. FEBS Lett 402: 41–44.
63. Madigan MT, Martinko JM, Parker J (2003) Brock biology of microorganisms. Upper Saddle River: Prentice Hall. 1,019 p.
64. Borovok I, Kreisberg-Zakarin R, Yanko M, Schreiber R, Myslovati M, et al. (2002) *Streptomyces* spp. contain class Ia and class II ribonucleotide reductases: Expression analysis of the genes in vegetative growth. Microbiology 148: 391–404.
65. Dunn JJ, Studier FW (1983) Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. J Mol Biol 166: 477–535.
66. Maggio-Hall LA, Escalante-Semerena JC (1999) In vitro synthesis of the nucleotide loop of cobalamin by *Salmonella typhimurium* enzymes. Proc Natl Acad Sci U S A 96: 11798–11803.
67. Lawrence JG, Roth JR (1995) The cobalamin (coenzyme B12) biosynthetic genes of *Escherichia coli*. J Bacteriol 177: 6371–6380.
68. Gleason FK, Olszewski NE (2002) Isolation of the gene for the B12-dependent ribonucleotide reductase from *Anabaena* sp. strain PCC 7120 and expression in *Escherichia coli*. J Bacteriol 184: 6544–6550.
69. Moore LR, Chisholm SW (1999) Photophysiology of the marine cyanobacterium *Prochlorococcus:* Ecotypic differences among cultured isolates. Limnol Oceanog 44: 628–638.
70. Waterbury JB, Watson SW, Valois FW, Franks DG (1986) Biological and ecological characterization of the marine unicellular cyanobacterium *Synechococcus*. Can Bull Fish Aquat Sci 214: 71–120.
71. Stanier RY (1973) Autotrophy and heterotrophy in unicellular blue-green algae. In: Carr NG, Whitton BA, editors. The biology of blue-green algae. Berkeley: University of California Press. pp. 501–518.
72. Sherman LA (1976) Infection of *Synechococcus cedrorum* by the cyanophage AS-1M. III. Cellular metabolism and phage development. Virology 71: 199–206.
73. Sprenger GA (1995) Genetics of pentose-phosphate pathway enzymes of *Escherichia coli* K-12. Arch Microbiol 164: 324–330.
74. Schurmann M, Sprenger GA (2001) Fructose-6-phosphate aldolase is a novel class I aldolase from *Escherichia coli* and is related to a novel group of bacterial transaldolases. J Biol Chem 276: 11055–11061.
75. Thorell S, Schurmann M, Sprenger GA, Schneider G (2002) Crystal structure of decameric fructose-6-phosphate aldolase from *Escherichia coli* reveals inter-subunit helix swapping as a structural basis for assembly differences in the transaldolase family. J Mol Biol 319: 161–171.
76. Karl DM (1999) A sea of change: Biogeochemical variability in the North Pacific Subtropical Gyre. Ecosystems 2: 181–214.
77. Wu J, Sunda W, Boyle EA, Karl DM (2000) Phosphate depletion in the western North Atlantic Ocean. Science 289: 759–762.
78. Scanlan DJ, Silman NJ, Donald KM, Wilson WH, Carr NG, et al. (1997) An immunological approach to detect phosphate stress in populations and single cells of photosynthetic picoplankton. Appl Environ Microbiol 63: 2411–2420.
79. Kazakov AE, Vassieva O, Gelfand MS, Osterman A, Overbeek R (2003) Bioinformatics classification and functional analysis of PhoH homologs. In Silico Biol 3: 3–15.
80. Kim SK, Makino K, Amemura M, Shinagawa H, Nakata A (1993) Molecular analysis of the *phoH* gene, belonging to the phosphate regulon in *Escherichia coli*. J Bacteriol 175: 1316–1324.
81. Wanner BL (1996) Phosphorus assimilation and control of the phosphate regulon. In: Neidhardt FC, editor. *Escherichia coli* and *Salmonella:* Cellular and molecular biology, 2nd ed. Washington (DC): ASM Press. pp. 1357–1381.
82. Calendar R, editor (1988) The bacteriophages. New York: Plenum Press..
83. Los M, Wegrzyn G, Neubauer P (2003) A role for bacteriophage T4 *rI* gene function in the control of phage development during pseudolysogeny and in slowly growing host cells. Res Microbiol 154: 547–552.
84. Moebus K (1996) Marine bacteriophage reproduction under nutrient-limited growth of host bacteria. II. Investigations with phage-host system [H3:H3/1]. Mar Ecol Prog Ser 144: 13–22.
85. Williamson SJ, McLaughlin MR, Paul JH (2001) Interaction of the PhiHSIC virus with its host: Lysogeny or pseudolysogeny? Appl Environ Microbiol 67: 1682–1688.
86. Mosig G (1998) Recombination and recombination-dependent DNA replication in bacteriophage T4. Annu Rev Genet 32: 379–413.
87. Mosig G, Gewin J, Luder A, Colowick N, Vo D (2001) Two recombination-dependent DNA replication pathways of bacteriophage T4, and their roles in mutagenesis and horizontal gene transfer. Proc Natl Acad Sci U S A 98: 8306–8311.
88. Hendrix RW, Smith MC, Burns RN, Ford ME, Hatfull GF (1999) Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. Proc Natl Acad Sci U S A 96: 2192–2197.
89. Hammer PE, Burd W, Hill DS, Ligon JM, van Pee K (1999) Conservation of the pyrrolnitrin biosynthetic gene cluster among six pyrrolnitrin-producing strains. FEMS Microbiol Lett 180: 39–44.
90. Kirner S, Hammer PE, Hill DS, Altmann A, Fischer I, et al. (1998) Functions encoded by pyrrolnitrin biosynthetic genes from *Pseudomonas fluorescens*. J Bacteriol 180: 1939–1943.
91. Hammer PE, Hill DS, Lam ST, Van Pee KH, Ligon JM (1997) Four genes from *Pseudomonas fluorescens* that encode the biosynthesis of pyrrolnitrin. Appl Environ Microbiol 63: 2147–2154.
92. Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, et al. (2003) Origins of highly mosaic mycobacteriophage genomes. Cell 113: 171–182.
93. Rohwer F (2003) Global phage diversity. Cell 113: 141.
94. Lawrence JG, Hendrickson H (2003) Lateral gene transfer: When will adolescence end? Mol Microbiol 50: 739–749.
95. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, et al. (2002) Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci U S A 99: 14250–14255.
96. Besemer J, Lomsadze A, Borodovsky M (2001) GeneMarkS: A self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res 29: 2607–2618.
97. Brussow H, Hendrix RW (2002) Phage genomics: Small is beautiful. Cell 108: 13–16.
98. Lowe TM, Eddy SR (1997) tRNAscan-SE: A program for improved

detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25: 955–964.
99. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.
100. Sonnhammer EL, Durbin R (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167: GC1–GC10.
101. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25: 4876–4882.
102. Swofford DL (2002) PAUP: Phylogenetic analysis using parsimony (and other methods), version 4 [computer program]. Sunderland (Massachusetts): Sinauer.
103. Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18: 502–504.
104. Miller ES, Kutter E, Mosig G, Arisaka F, Kunisawa T, et al. (2003) Bacteriophage T4 genome. Microb Mol Biol Rev 67: 86–156.
105. van Regenmortel MHV, Fauquet CM, Bishop DHL, Carstens EB, Estes MK, et al. (2000) Virus taxonomy: The classification and nomenclature of viruses. San Deigo: Academic Press. 1,167 p.
106. Desplats C, Krisch HM (2003) The diversity and evolution of the T4-type bacteriophages. Res Microbiol 154: 259–267.
107. Tetart F, Desplats C, Kutateladze M, Monod C, Ackermann HW, et al. (2001) Phylogeny of the major head and tail genes of the wide-ranging T4-type bacteriophages. J Bacteriol 183: 358–366.
108. Mann NH, Clokie MR, Millard A, Cook A, Wilson WH, et al. The genome of S-PM2, a "photosynthetic" T4-type bacteriophage that infects marine *Synechococcus*. J Bacteriol. In press.

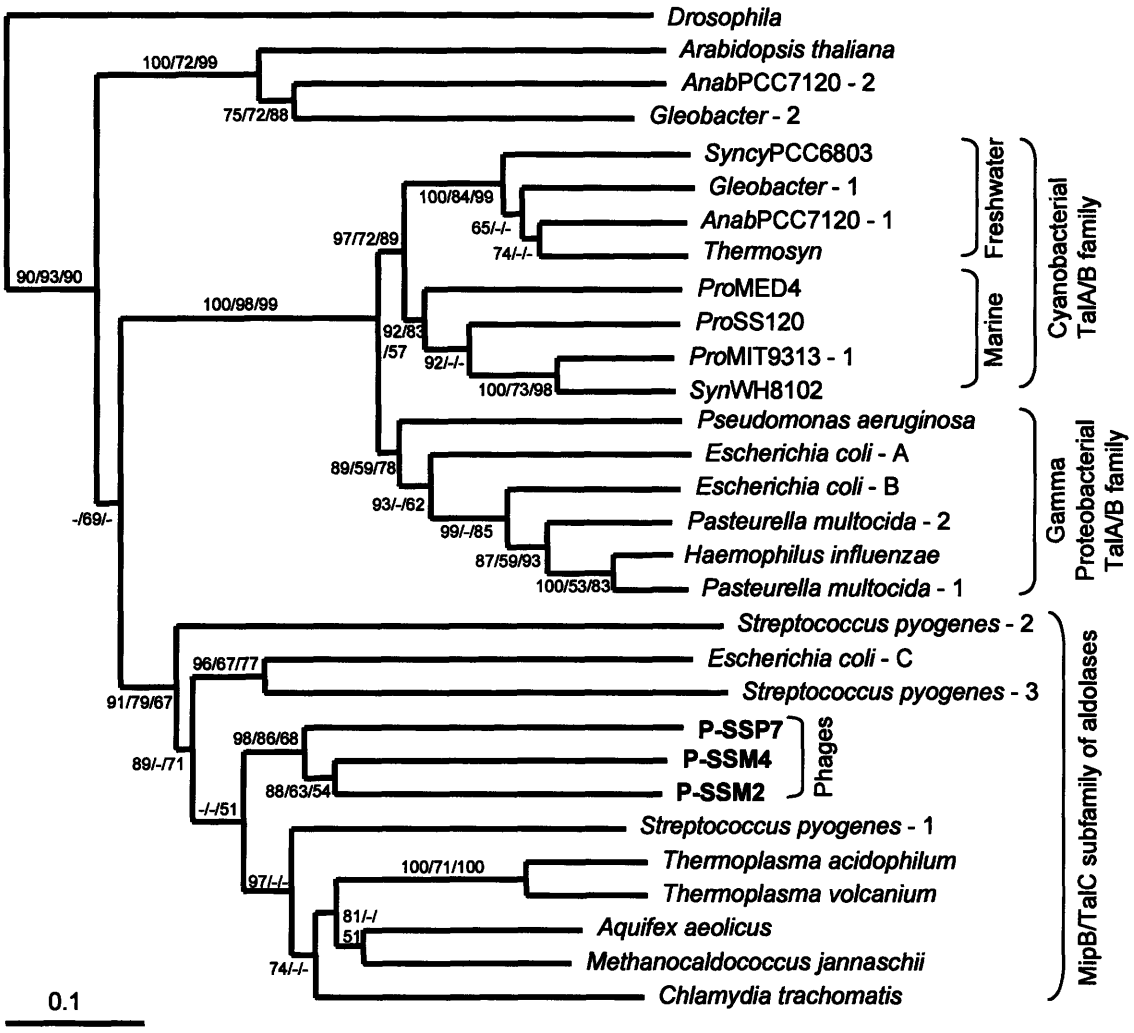
Supplementary Figure 1: Comparison of Cyanobacterial and non-T4-like phages against class II RNR motif.

Class II RNR MOTIF	T/S	N	P	C	G/S	E	X	X	X	X	X	X	X	C	N/L	L
Thermosynechococcus	L	.	.	.	G	.	I	I	-	G	A	D	F	H	.	N	.
ProMED4	L	.	.	.	G	.	I	L	-	G	N	D	F	H	.	N	.
ProMIT9313	L	.	.	.	G	.	I	L	-	G	A	D	F	H	.	N	.
ProSS120	L	.	.	.	G	.	I	L	-	G	A	D	F	H	.	N	.
SynWH8102	L	.	.	.	G	.	I	L	-	G	A	D	F	H	.	N	.
Cyanophage P60_nrdB	L	.	.	.	G	.	I	L	-	G	K	D	F	H	.	N	.
Mycobacteriophage D29	T	.	.	.	G	.	I	T	L	E	P	W	E	P	.	N	.
Roseophage SIO1	V	.	.	.	A	.	I	L	L	G	N	K	S	F	.	N	.
Cyanophage P-SSP7	S	.	V	.	L	.	V	Y	L	P	S	R	G	T	.	L	.
SS (GI:44265515)	G	.	V	.	L	.	V	Y	L	P	S	R	G	T	.	L	.
SS (GI:44494788)	G	.	V	.	L	.	V	Y	L	P	S	R	G	T	.	L	.
SS (GI:43424236)	?	?	?	.	L	.	V	Y	L	P	S	R	G	T	.	L	.

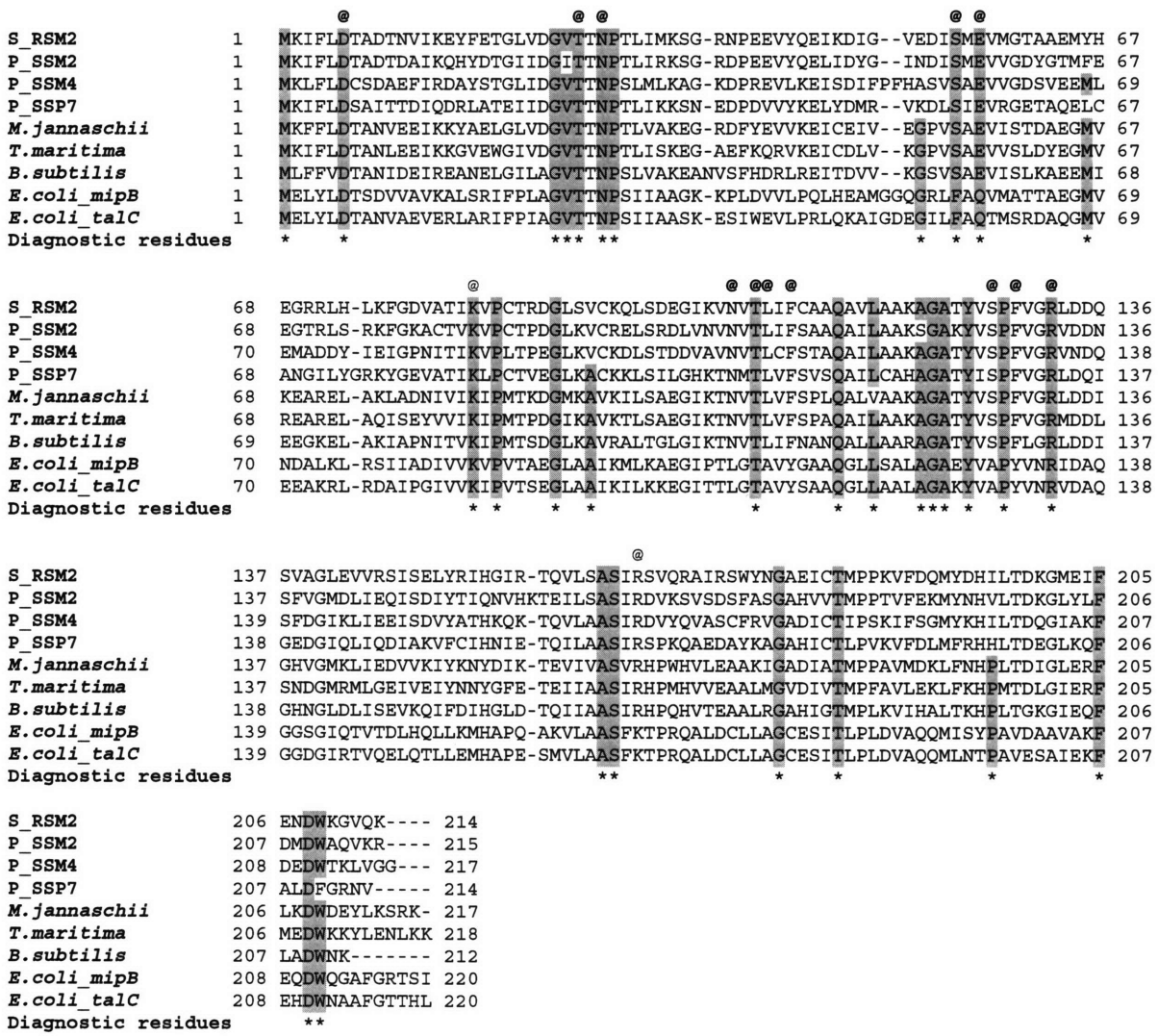
Supplementary Figure 2: Distance tree of ribonucleotide reductase proteins



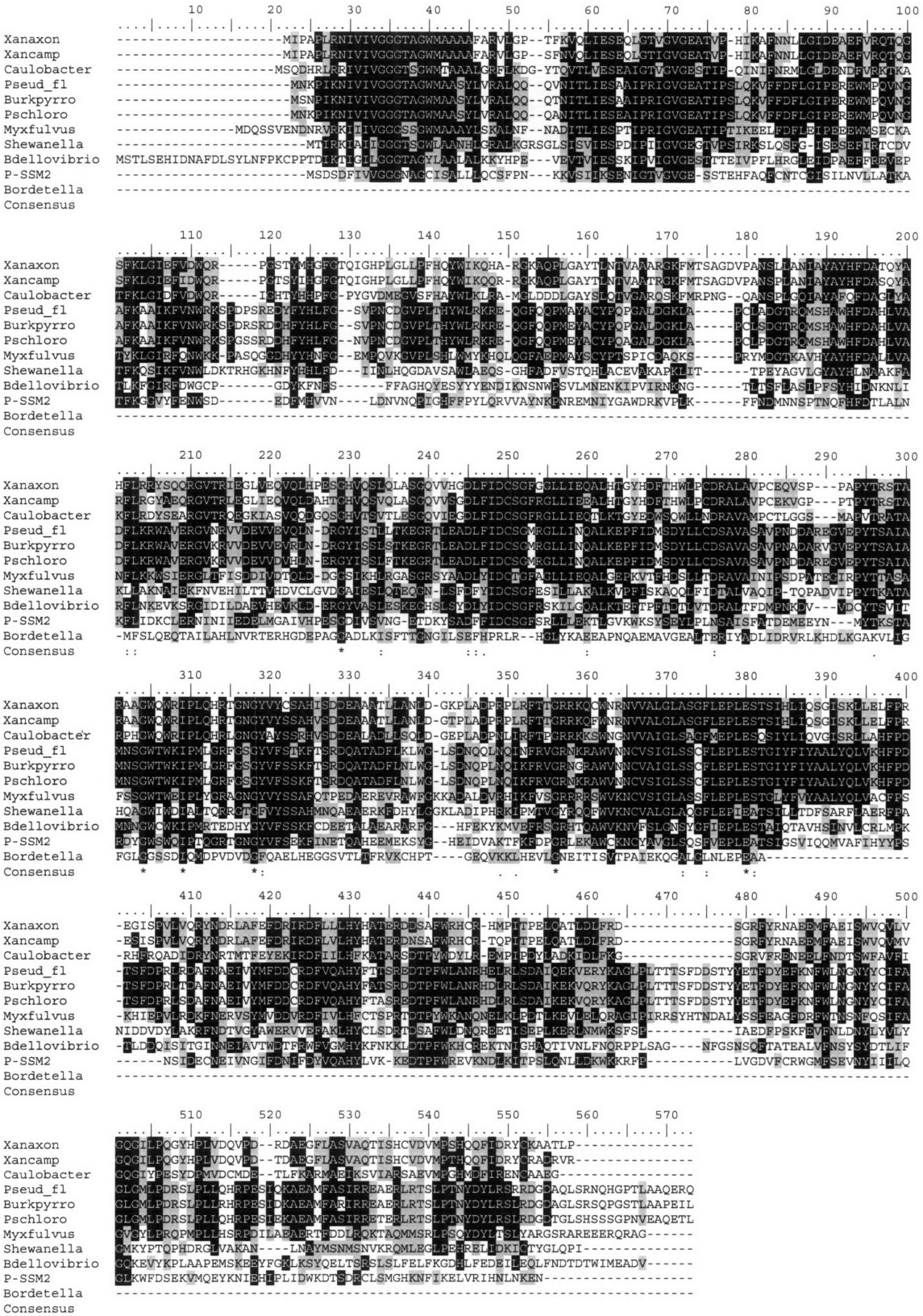
Supplementary Figure 3: Distance tree of transaldolase proteins



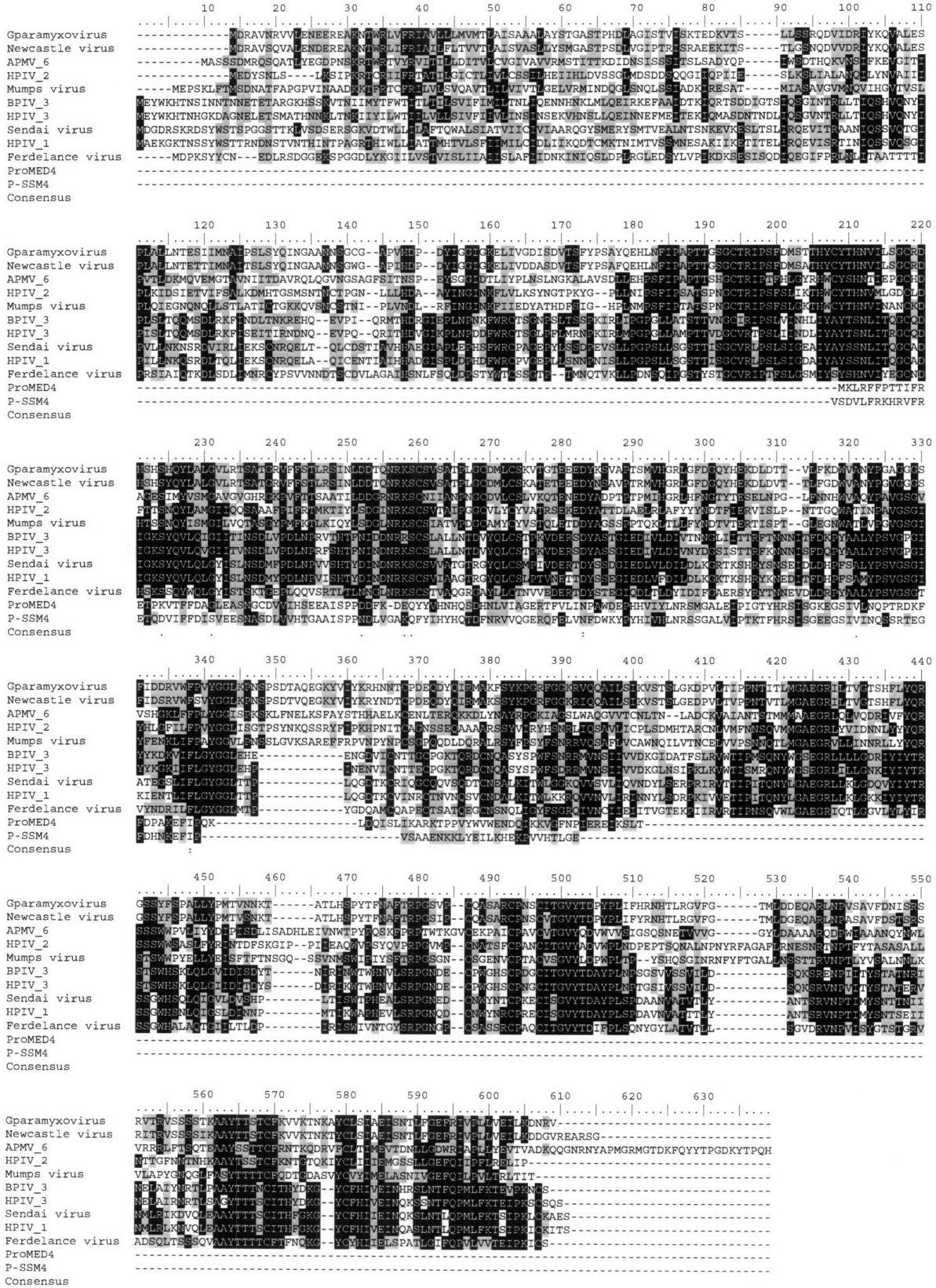
Supplementary Figure 4: Alignment of TalC Subfamily aldolases, including phage sequences from P-SSM2, P-SSM4, P-SSP7 and P-RSM2



Supplementary Figure 5: PrnA protein alignments



Supplementary Figure 6: Hemagglutinin neuraminidase alignment



Appendix F

Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation

Gabrielle Rocap, Frank W. Larimer, Jane Lamerdin, Stephanie Malfatti, Patrick Chain, Nathan A. Ahlgren, Andrae Arellano, Maureen Coleman, Loren Hauser, Wolfgang R. Hess, Zackary I. Johnson, Miriam Land, Debbie Lindell, Anton F. Post, Warren Regala, Manesh Shah, Stephanie L. Shaw, Claudia Steglich, Matthew B. Sullivan, Claire S. Ting, Andrew Tolonen, Eric A. Webb, Erik R. Zinser, and Sallie W. Chisholm

Rocap, G., Larimer, F.W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N.A., Arellano, A., **Coleman, M.L.**, Hauser, L., Hess, W.R., Johnson, Z.I., Land, M., Lindell, D., Post, A.F., Regala, W., Shah, M., Shaw, S.L., Steglich, C., Sullivan, M.B., Ting, C.S., Tolonen, A., Webb, E.A., Zinser, E.R. and Chisholm, S.W. (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* 424:1042-1047.

Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation

Gabrielle Rocap[1], Frank W. Larimer[2,3], Jane Lamerdin[3], Stephanie Malfatti[3], Patrick Chain[3,4], Nathan A. Ahlgren[1], Andrae Arellano[3], Maureen Coleman[5], Loren Hauser[2,3], Wolfgang R. Hess[9]*, Zackary I. Johnson[5], Miriam Land[2,3], Debbie Lindell[5], Anton F. Post[10], Warren Regala[3], Manesh Shah[2,3], Stephanie L. Shaw[6]*, Claudia Steglich[9], Matthew B. Sullivan[7], Claire S. Ting[8], Andrew Tolonen[7], Eric A. Webb[11], Erik R. Zinser[5] & Sallie W. Chisholm[5,8]

[1]*School of Oceanography, University Of Washington, Seattle, Washington 98195, USA*
[2]*Computational Biology, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA*
[3]*Joint Genome Institute, Walnut Creek, California 94598, USA*
[4]*Lawrence Livermore National Laboratory, Livermore, California 94550, USA*
[5]*Department of Civil and Environmental Engineering,* [6]*Department of Earth, Atmospheric and Planetary Sciences,* [7]*Joint Program in Biological Oceanography, Woods Hole Oceanographic Institution, and* [8]*Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
[9]*Institute of Biology, Humboldt-University, D-10115 Berlin, Germany*
[10]*Interuniversity Institute of Marine Science, 88103 Eilat, Israel*
[11]*Biology Department, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA*

* Present addresses: Ocean Genome Legacy, Beverly, Massachusetts 01915, USA (W.R.H.); Department of Environmental Science Policy and Management, University of California, Berkeley, California 94720, USA (S.L.S.)

The marine unicellular cyanobacterium *Prochlorococcus* is the smallest-known oxygen-evolving autotroph[1]. It numerically dominates the phytoplankton in the tropical and subtropical oceans[2,3], and is responsible for a significant fraction of global photosynthesis. Here we compare the genomes of two *Prochlorococcus* strains that span the largest evolutionary distance within the *Prochlorococcus* lineage[4] and that have different minimum, maximum and optimal light intensities for growth[5]. The high-light-adapted ecotype has the smallest genome (1,657,990 base pairs, 1,716 genes) of any known oxygenic phototroph, whereas the genome of its low-light-adapted counterpart is significantly larger, at 2,410,873 base pairs (2,275 genes). The comparative architectures of these two strains reveal dynamic genomes that are constantly changing in response to myriad selection pressures. Although the two strains have 1,350 genes in common, a significant number are not shared, and these have been differentially retained from the common ancestor, or acquired through duplication or lateral transfer. Some of these genes have obvious roles in determining the relative fitness of the ecotypes in response to key environmental variables, and hence in regulating their distribution and abundance in the oceans.

As an oxyphototroph, *Prochlorococcus* requires only light, CO_2 and inorganic nutrients, thus the opportunities for extensive niche differentiation are not immediately obvious—particularly in view of the high mixing potential in the marine environment (Fig. 1a). Yet co-occurring *Prochlorococcus* cells that differ in their ribosomal DNA sequence by less than 3% have different optimal light intensities for growth[6], pigment contents[7], light-harvesting efficiencies[5], sensitivities to trace metals[8], nitrogen usage abilities[9] and cyanophage specificities[10] (Fig. 1b, c). These 'ecotypes'—distinct genetic lineages with ecologically relevant physiological differences—would be lumped together as a single species on the basis of their rDNA similarity[11], yet they have markedly different distributions within a stratified oceanic water column, with high-

Table 1 **General features of two *Prochlorococcus* genomes**

Genome feature	MED4	MIT9313
Length (bp)	1,657,990	2,410,873
G+C content (%)	30.8	50.7
Protein coding (%)	88	82
Protein coding genes	1,716	2,275
With assigned function	1,134	1,366
Conserved hypothetical	502	709
Hypothetical	80	197
Genes with orthologue in:		
Prochlorococcus MED4	–	1,352
Prochlorococcus MIT9313	1,352	–
Synechococcus WH8102	1,394	1,710
Genes without orthologue in:		
MED4 and WH8102	–	527
MIT9313 and WH8102	284	–
Transfer RNA	37	43
Ribosomal RNA operons	1	2
Other structural RNAs	3	3

light-adapted ecotypes most abundant in surface waters, and their low-light-adapted counterparts dominating deeper waters[12] (Fig. 1a). The detailed comparison between the genomes of two *Prochlorococcus* ecotypes we report here reveals many of the genetic foundations for the observed differences in their physiologies and vertical niche partitioning, and together with the genome of their close relative *Synechococcus*[13], helps to elucidate the key factors that regulate species diversity, and the resulting biogeochemical cycles, in today's oceans.

The genome of *Prochlorococcus* MED4, a high-light-adapted strain, is 1,657,990 base pairs (bp). This is the smallest of any oxygenic phototroph—significantly smaller than that of the low-light-adapted strain MIT9313 (2,410,873 bp; Table 1). The genomes of MED4 and MIT9313 consist of a single circular chromosome (Supplementary Fig. 1), and encode 1,716 and 2,275 genes respectively, roughly 65% of which can be assigned a functional category (Supplementary Fig. 2). Both genomes have undergone numerous large and small-scale rearrangements but they retain conservation of local gene order (Fig. 2). Break points between the orthologous gene clusters are commonly flanked by transfer RNAs, suggesting that these genes serve as loci for rearrangements caused by internal homologous recombination or phage integration events.

The strains have 1,352 genes in common, all but 38 of which are also shared with *Synechococcus* WH8102 (ref. 13). Many of the 38 '*Prochlorococcus* -specific' genes encode proteins involved in the atypical light-harvesting complex of *Prochlorococcus*, which contains divinyl chlorophylls *a* and *b* rather than the phycobilisomes that characterize most cyanobacteria. They include genes encoding the chlorophyll *a*/*b*-binding proteins (*pcb*)[14], a putative chlorophyll *a* oxygenase, which could synthesize (divinyl) chlorophyll *b* from (divinyl) chlorophyll *a*[15], and a lycopene epsilon cyclase involved in the synthesis of alpha carotene[16]. This remarkably low number of 'genera defining' genes illustrates how differences in a few gene families can translate into significant niche differentiation among closely related microbes.

MED4 has 364 genes without an orthologue in MIT9313, whereas MIT9313 has 923 that are not present in MED4. These strain-specific genes, which are dispersed throughout the chromosome (Fig. 2), clearly hold clues about the relative fitness of the two strains under different environmental conditions. Almost half of the 923 MIT9313-specific genes are in fact present in *Synechococcus* WH8102, suggesting that they have been lost from MED4 in the course of genome reduction. Lateral transfer events, perhaps

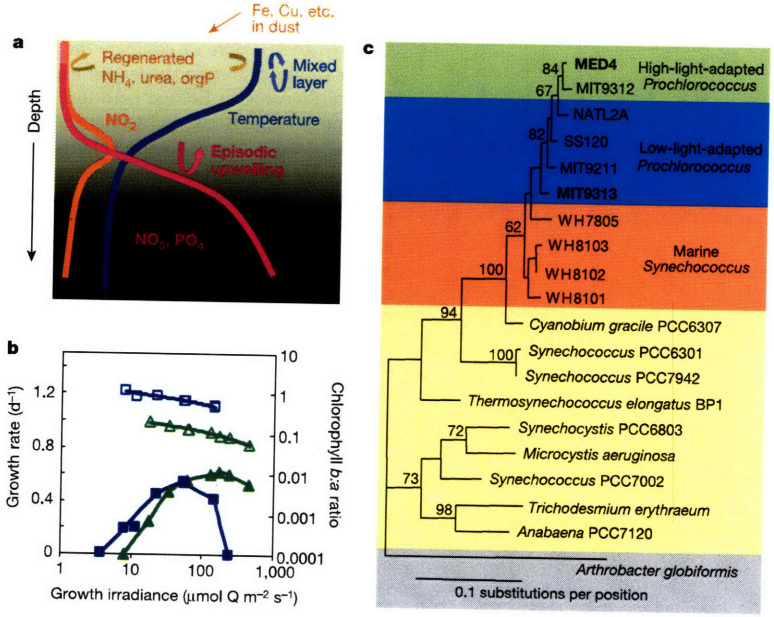


Figure 1 Ecology, physiology and phylogeny of *Prochlorococcus* ecotypes. **a**, Schematic stratified open-ocean water column illustrating vertical gradients allowing niche differentiation. Shading represents degree of light penetration. Temperature and salinity gradients provide a mixing barrier, isolating the low-nutrient/high-light surface layer from the high-nutrient/low-light deep waters. Photosynthesis in surface waters is driven primarily by rapidly regenerated nutrients, punctuated by episodic upwelling. **b**, Growth rate (filled symbols) and chlorophyll *b*:*a* ratio (open symbols) as a function of growth irradiance for MED4 (ref. 7) (green) and MIT9313 (ref. 6) (blue). **c**, Relationships between *Prochlorococcus* and other cyanobacteria inferred using 16S rDNA.

mediated by phage[10], may also be a source of some of the strain-specific genes (Supplementary Figs 3–6).

Gene loss has played a major role in defining the *Prochlorococcus* photosynthetic apparatus. MED4 and MIT9313 are missing many of the genes encoding phycobilisome structural proteins and enzymes involved in phycobilin biosynthesis[15]. Although some of these genes remain, and are functional[17], others seem to be evolving rapidly within the *Prochlorococcus* lineage[18]. Selective genome reduction can also be seen in the photosynthetic reaction centre of *Prochlorococcus*. Light acclimation in cyanobacteria often involves differential expression of multiple, but distinct, copies of genes encoding photosystem II D1 and D2 reaction centre proteins (*psbA* and *psbD* respectively)[19]. However, MED4 has a single *psbA* gene, MIT9313 has two that encode identical photosystem II D1 polypeptides, and both possess only one *psbD* gene, suggesting a diminished ability to photoacclimate. MED4 has also lost the gene encoding cytochrome *c*550 (*psbV*), which has a crucial role in the oxygen-evolving complex in *Synechocystis* PCC6803 (ref. 20).

There are several differences between the genomes that help account for the different light optima of the two strains. For example, the smaller MED4 genome has more than twice as many genes (22 compared with 9) encoding putative high-light-inducible proteins, which seem to have arisen at least in part through duplication events[15]. MED4 also possesses a photolyase gene that has been lost in MIT9313, probably because there is little selective pressure to retain ultraviolet damage repair in low light habitats. Regarding differences in light-harvesting efficiencies, it is noteworthy that MED4 contains only a single gene encoding the chlorophyll *a*/*b*-binding antenna protein Pcb, whereas MIT9313 possesses two copies. The second type has been found exclusively in low-light-adapted strains[21], and may form an antenna capable of binding more chlorophyll pigments.

Both strains have a low proportion of genes involved in regulatory functions. Compared with the freshwater cyanobacterium *Thermosynechococcus elongatus* (genome size <2.6 megabases)[22], MIT9313 has fewer sigma factors, transcriptional regulators and two-component sensor-kinase systems, and MED4 is even more reduced (Supplementary Table 1). The circadian clock genes provide an example of this reduction as both genomes lack several components (*pex*, *kaiA*) found in the model *Synechococcus* PCC7942 (ref. 23). However, genes for the core clock proteins (*kaiB*, *kaiC*) remain in both genomes, and *Prochlorococcus* cell division is tightly synchronized to the diel light/dark cycle[24]. Thus, loss of some circadian components may imply an alternative signalling pathway for circadian control.

Gene loss may also have a role in the lower percentage of G+C content of MED4 (30.8%) compared with that of MIT9313 (50.74%), which is more typical of marine *Synechococcus*. MED4 lacks genes for several DNA repair pathways including recombinational repair (*recJ*, *recQ*) and damage reversal (*mutT*). Particularly, the loss of the base excision repair gene *mut*Y, which removes adenosines incorrectly paired with oxidatively damaged guanine residues, may imply an increased rate of G•C to T•A transversions[25]. The tRNA complement of MED4 is largely identical to MIT9313 and is not optimized for a low percentage G+C genome, suggesting that it is not evolving as fast as codon usage.

Analysis of the nitrogen acquisition capabilities of the two strains points to a sequential decay in the capacity to use nitrate and nitrite during the evolution of the *Prochlorococcus* lineage (Fig. 3a). In *Synechococcus* WH8102—representing the presumed ancestral state—many nitrogen acquisition and assimilation genes are grouped together (Fig. 3a). MIT9313 has lost a 25-gene cluster, which includes genes encoding the nitrate/nitrite transporter and nitrate reductase. The nitrite reductase gene has been retained in MIT9313, but it is flanked by a proteobacterial-like nitrite transporter rather than a typical cyanobacterial nitrate/nitrite permease (Supplementary Fig. 4), suggesting acquisition by lateral gene transfer. An additional deletion event occurred in MED4, in which the nitrite reductase gene was also lost (Fig. 3a). As a result of these serial deletion events MIT9313 cannot use nitrate, and MED4 cannot use nitrate or nitrite[9]. Thus each *Prochlorococcus* ecotype uses the N species that is most prevalent at the light levels to which they are best adapted: ammonium in the surface waters and nitrite at depth (Fig. 1a). *Synechococcus*, which is the only one of the three that has nitrate reductase, is able to bloom when nitrate is upwelled (Fig. 1a), as occurs in the spring in the North Atlantic[3] and the north Red Sea[26].

The two *Prochlorococcus* strains are also less versatile in their organic N usage capabilities than *Synechococcus* WH8102 (ref. 13). MED4 contains the genes necessary for usage of urea, cyanate and oligopeptides, but no monomeric amino acid transporters have been identified. In contrast, MIT9313 contains transporters for urea, amino acids and oligopeptides but lacks the genes necessary for cyanate usage (cyanate transporter and cyanate lyase) (Fig. 3a). As expected, both genomes contain the high-affinity ammonium transporter *amt1* and both lack the nitrogenase genes essential for nitrogen fixation. Finally, both contain the nitrogen transcriptional regulator encoded by *ntcA* and there are numerous genes in both genomes, including *ntcA*, *amt1*, the urea transport and GS/GOGAT genes (glutamine synthetase and glutamate synthase, both involved in ammonia assimilation), with an upstream NtcA-binding-site consensus sequence.

The genomes also have differences in genes involved in phosphorus usage that have obvious ecological implications. MED4, but not MIT9313, is capable of growth on organic P sources (L. R. Moore and S.W.C., unpublished data), and organic P can be the prevalent form of P in high-light surface waters[27]. This difference may be due to the acquisition of an alkaline phosphatase-like gene in MED4 (Supplementary Fig. 5). Both genomes contain the high-affinity phosphate transport system encoded by *pstS* and *pstABC*[28], but MIT9313 contains an additional copy of the phosphate-binding component *pstS*, perhaps reflecting an increased reliance on orthophosphate in deeper waters. MED4 contains

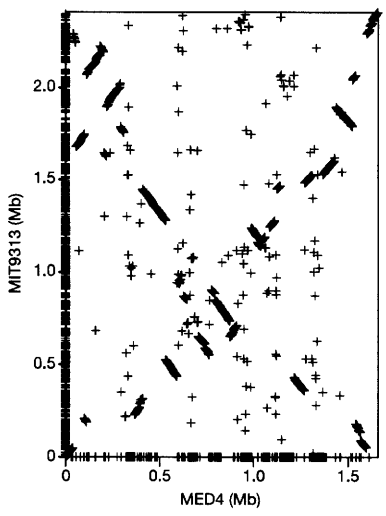


Figure 2 Global genome alignment as seen from start positions of orthologous genes. Genes present in one genome but not the other are shown on the axes. The 'broken X' pattern has been noted before for closely related bacterial genomes, and is probably due to multiple inversions centred around the origin of replication. Alternating slopes of many adjacent gene clusters indicate that multiple smaller-scale inversions have also occurred.

several P-related regulatory genes including the *phoB*, *phoR* two-component system and the transcriptional activator *ptrA*. In MIT9313, however, *phoR* is interrupted by two frameshifts and *ptrA* is further degenerated, suggesting that this strain has lost the ability to regulate gene expression in response to changing P levels.

Both *Prochlorococcus* strains have iron-related genes that are missing in *Synechococcus* WH8102, which may explain its dominance in the iron-limited equatorial Pacific[2]. These genes include flavodoxin (*isiB*), an Fe-free electron transfer protein capable of replacing ferredoxin, and ferritin (located with the ATPase component of an iron ABC transporter), an iron-binding molecule implicated in iron storage. Additional characteristics of the iron acquisition system in these genomes include: an Fe-induced transcriptional regulator (Fur) that represses iron uptake genes; numerous genes with an upstream putative *fur* box motif that are candidates for a high-affinity iron scavenging system; and absence of genes involved in Fe–siderophore complexes.

Prochlorococcus does not use typical cyanobacterial genes for inorganic carbon concentration or fixation. Both genomes contain a sodium/bicarbonate symporter but lack homologues to known families of carbonic anhydrases, suggesting that an as yet unidentified gene is fulfilling this function. One of the two carbonic anhydrases in *Synechococcus* WH8102 was lost in the deletion event that led to the loss of the nitrate reductase (Fig. 3a); the other is located next to a tRNA and seems to have been lost during a genome rearrangement event. Similar to other *Prochlorococcus* and marine *Synechococcus*, MED4 and MIT9313 possess a form IA ribulose-1,5-bisphosphate carboxylase/oxygenase, rather than the typical cyanobacterial form IB. The ribulose-1,5-bisphosphate carboxylase/oxygenase genes are adjacent to genes encoding structural carboxysome shell proteins and all have phylogenetic affinity to genes in the γ-proteobacterium *Acidithiobacillus ferroxidans*[15], suggesting lateral transfer of the extended operon.

Prochlorococcus has been identified in deep suboxic zones where it is unlikely that they can sustain themselves by photosynthesis alone[29], thus we looked for genomic evidence of heterotrophic capability. Indeed, the presence of oligopeptide transporters in both genomes, and the larger proportion of transporters (including some sugar transporters) in the MIT9313 strain-specific genes (Supplementary Fig. 2), suggests the potential for partial hetero-

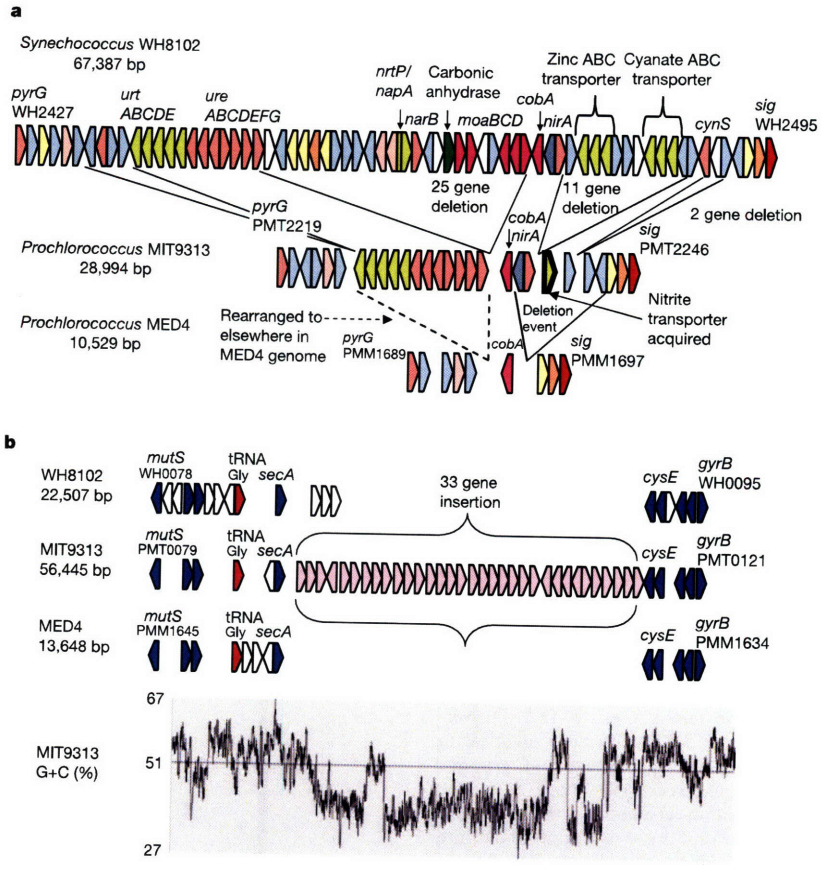


Figure 3 Dynamic architecture of marine cyanobacterial genomes. **a**, Deletion, acquisition and rearrangement of nitrogen usage genes. In MIT9313, 25 genes including the nitrate/nitrite transporter (*nrtP*/*napA*), nitrate reductase (*narB*) and carbonic anhydrase have been deleted. The cyanate transporter and cyanate lyase (*cynS*) were probably lost after the divergence of MIT9313 from the rest of the *Prochlorococcus* lineage, as MED4 possesses these genes. MIT9313 has retained nitrite reductase (*nirA*) and acquired a nitrite transporter. In MED4 *nirA* has been lost and the urea transporter (*urt* cluster) and urease (*ure* cluster) genes have been rearranged (dotted line). Genes in different functional categories are colour-coded to guide the eye. **b**, Lateral transfer of genes involved in lipopolysaccharide biosynthesis including sugar transferases, sugar epimerases, modifying enzymes and two pairs of ABC-type transporters. Blue, genes in all three genomes; pink, genes hypothesized to have been laterally transferred; red, tRNAs; white, other genes. The percentage of G + C content in MIT9313 along this segment is lower (42%) than the whole-genome average (horizontal line).

trophy. However, neither genome contains known pathways that would allow for complete heterotrophy. They are both missing genes for steps in the tricarboxylic acid cycle, including 2-oxoglutarate dehydrogenase, succinyl-CoA synthetase and succinyl-CoA-acetoacetate-CoA transferase.

Cell surface chemistry has a major role in phage recognition and grazing by protists and thus is probably under intense selective pressure in nature. The two *Prochlorococcus* genomes and the *Synechococcus* WH8102 genome show evidence of extensive lateral gene transfer and deletion events of genes involved in lipopolysaccharide and/or surface polysaccharide biosynthesis, reinforcing the role of predation pressures in the creation and maintenance of microdiversity. For example, MIT9313 has a 41.8-kilobase (kb) cluster of surface polysaccharide genes (Fig. 3b), which has a lower percentage G+C composition (42%) than the genome as a whole, implicating acquisition by lateral gene transfer. MED4 has acquired a 74.5-kb cluster consisting of 67 potential surface polysaccharide genes (Supplementary Fig. 6a) and has lost another cluster of surface polysaccharide biosynthesis genes shared between MIT9313 and *Synechococcus* WH8102 (Supplementary Fig. 6b).

The approach we have taken in describing these genomes highlights the known drivers of niche partitioning of these closely related organisms (Fig. 1). Detailed comparisons with the genomes of additional strains, such as *Prochlorococcus* SS120 (ref. 30), will enrich this story, and the analysis of whole genomes from *in situ* populations will be necessary to understand the full expanse of genomic diversity in this group. The genes of unknown function in all of these genomes hold important clues for undiscovered niche dimensions in the marine pelagic zone. As we unveil their function we will undoubtedly learn that the suite of selective pressures that shape these communities is much larger than we have imagined. Finally, it may be useful to view *Prochlorococcus* and *Synechococcus* as important 'minimal life units', as the information in their roughly 2,000 genes is sufficient to create globally abundant biomass from solar energy and inorganic compounds. □

Methods

Genome sequencing and assembly

DNA was isolated from the clonal, axenic strain MED4 and the clonal strain MIT9313 essentially as described previously[4]. The two whole-genome shotgun libraries were obtained by fragmenting genomic DNA using mechanical shearing and cloning 2–3-kb fragments into pUC18. Double-ended plasmid sequencing reactions were carried out using PE BigDye Terminator chemistry (Perkin Elmer) and sequencing ladders were resolved on PE 377 Automated DNA Sequencers (Perkin Elmer). The whole-genome sequence of *Prochlorococcus* MED4 was obtained from 27,065 end sequences (7.3-fold redundancy), whereas *Prochlorococcus* MIT9313 was sequenced to ×6.2 coverage (33,383 end sequences). For *Prochlorococcus* MIT9313, supplemental sequencing (×0.05 sequence coverage) of a pFos1 fosmid library was used as a scaffold. Sequence assembly was accomplished using PHRAP (P. Green). All gaps were closed by primer walking on gap-spanning library clones or PCR products. The final assembly of *Prochlorococcus* MED4 was verified by long-range genomic PCR reactions, whereas the assembly of *Prochlorococcus* MIT9313 was confirmed by comparison to the fosmid clones, which were fingerprinted with *Eco*RI. No plasmids were detected in the course of genome sequencing, and insertion sequences, repeated elements, transposons and prophages are notably absent from both genomes. The likely origin of replication in each genome was identified based on G+C skew, and base pair 1 was designated adjacent to the *dnaN* gene.

Genome annotation

The combination of three gene-modelling programs, Critica, Glimmer and Generation, were used in the determination of potential open reading frames and were checked manually. A revised gene/protein set was searched against the KEGG GENES, Pfam, PROSITE, PRINTS, ProDom, COGs and CyanoBase databases, in addition to BLASTP against the non-redundant peptide sequence database from GenBank. From these results, categorizations were developed using the KEGG and COGs hierarchies, as modified in CyanoBase. Manual annotation of open reading frames was done in conjunction with the *Synechococcus* team. The three-way genome comparison was used to refine predicted start sites, add additional open reading frames and standardize the annotation across the three genomes.

Genome comparisons

The comparative genome architecture of MED4 and MIT9313 was visualized using the Artemis Comparison Tool (http://www.sanger.ac.uk/Software/ACT/). Orthologues were determined by aligning the predicted coding sequences of each gene with the coding sequences of the other genome using BLASTP. Genes were considered orthologues if each was the best hit of the other one and both *e*-values were less than e^{-10}. In addition, bidirectional best hits with *e*-values less than e^{-6} and small proteins of conserved function were manually examined and added to the orthologue lists.

Phylogenetic analyses used PAUP*, logdet distances and minimum evolution as the objective function. The degree of support at each node was evaluated using 1,000 bootstrap resamplings. Ribosomal DNA analyses used 1,160 positions. The Gram-positive bacterium *Arthrobacter globiformis* was used to root the tree.

Received 11 May; accepted 25 July 2003; doi:10.1038/nature01947.
Published online 13 August 2003.

1. Chisholm, S. W. *et al.* A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature* **334,** 340–343 (1988).
2. Campbell, L., Liu, H., Nolla, H. & Vaulot, D. Annual variability of phytoplankton and bacteria in the subtropical North Pacific Ocean at Station ALOHA during the 1991–1994 ENSO event. *Deep Sea Res. I* **44,** 167–192 (1997).
3. DuRand, M. D., Olson, R. J. & Chisholm, S. W. Phytoplankton population dynamics at the Bermuda Atlantic Time-series Station in the Sargasso Sea. *Deep Sea Res. II* **48,** 1983–2003 (2001).
4. Rocap, G., Distel, D. L., Waterbury, J. B. & Chisholm, S. W. Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S–23S rDNA internal transcribed spacer (ITS) sequences. *Appl. Environ. Microbiol.* **68,** 1180–1191 (2002).
5. Moore, L. R. & Chisholm, S. W. Photophysiology of the marine cyanobacterium *Prochlorococcus*: Ecotypic differences among cultured isolates. *Limnol. Oceanogr.* **44,** 628–638 (1999).
6. Moore, L. R., Rocap, G. & Chisholm, S. W. Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* **393,** 464–467 (1998).
7. Moore, L. R., Goericke, R. E. & Chisholm, S. W. Comparative physiology of *Synechococcus* and *Prochlorococcus*: influence of light and temperature on growth, pigments, fluorescence and absorptive properties. *Mar. Ecol. Prog. Ser.* **116,** 259–275 (1995).
8. Mann, E. L., Ahlgren, N., Moffett, J. W. & Chisholm, S. W. Copper toxicity and cyanobacteria ecology in the Sargasso Sea. *Limnol. Oceanogr.* **47,** 976–988 (2002).
9. Moore, L. R., Post, A. F., Rocap, G. & Chisholm, S. W. Utilization of different nitrogen sources by the marine cyanobacteria, *Prochlorococcus* and *Synechococcus*. *Limnol. Oceanogr.* **47,** 989–996 (2002).
10. Sullivan, M. B., Waterbury, J. B. & Chisholm, S. W. Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. *Nature* **424,** 1047–1051 (2003).
11. Hagström, Å. *et al.* Use of 16S ribosomal DNA for delineation of marine bacterioplankton species. *Appl. Environ. Microbiol.* **68,** 3628–3633 (2002).
12. West, N. J. *et al.* Closely related *Prochlorococcus* genotypes show remarkably different depth distributions in two oceanic regions as revealed by *in situ* hybridization using 16S rRNA-targeted oligonucleotides. *Microbiology* **147,** 1731–1744 (2001).
13. Palenik, B. *et al.* The genome of a motile marine *Synechococcus*. *Nature* **424,** 1037–1042 (2003).
14. La Roche, J. *et al.* Independent evolution of the prochlorophyte and green plant chlorophyll a/b light harvesting proteins. *Proc. Natl Acad. Sci. USA* **93,** 15244–15248 (1996).
15. Hess, W. *et al.* The photosynthetic apparatus of *Prochlorococcus*: Insights through comparative genomics. *Photosynth. Res.* **70,** 53–71 (2001).
16. Stickforth, P., Steiger, S., Hess, W. R. & Sandmann, G. A novel type of lycopene ε-cyclase in the marine cyanobacterium *Prochlorococcus marinus* MED4. *Arch. Microbiol.* **179,** 407–415 (2003).
17. Frankenberg, N., Mukougawa, K., Kohchi, T. & Lagarias, J. C. Functional genomic analysis of the HY2 family of ferredoxin-dependent bilin reductases from oxygenic photosynthetic organisms. *Plant Cell* **13,** 965–978 (2001).
18. Ting, C., Rocap, G., King, J. & Chisholm, S. W. Phycobiliprotein genes of the marine prokaryote *Prochlorococcus*: Evidence for rapid evolution of genetic heterogeneity. *Microbiology* **147,** 3171–3182 (2001).
19. Golden, S. S., Brusslan, J. & Haselkorn, R. Expression of a family of *psbA* genes encoding a photosystem II polypeptide in the cyanobacterium *Anacystis nidulans* R2. *EMBO J.* **5,** 2789–2798 (1986).
20. Shen, J. R., Qian, M., Inoue, Y. L. & Burnap, R. L. Functional characterization of *Synechocystis* sp. PCC 6803 Δ*psbU* and Δ*psbV* mutants reveals important roles of cytochrome c-550 in cyanobacterial oxygen evolution. *Biochemistry* **37,** 1551–1558 (1998).
21. Garczarek, L., van der Staay, G. W. M., Hess, W. R., Le Gall, F. & Partensky, F. Expression and phylogeny of the multiple antenna genes of the low-light-adapted strain *Prochlorococcus marinus* SS120 (Oxyphotobacteria). *Plant Mol. Biol.* **46,** 683–693 (2001).
22. Nakamura, Y. *et al.* Complete genome structure of the thermophilic Cyanobacterium *Thermosynechococcus elongatus* BP-1. *DNA Res.* **9,** 123–130 (2002).
23. Ishiura, M. *et al.* Expression of a gene cluster *kaiABC* as a circadian feedback process in cyanobateria. *Science* **281,** 1519–1523 (1998).
24. Vaulot, D., Marie, D., Olson, R. J. & Chisholm, S. W. Growth of *Prochlorococcus*, a photosynthetic prokaryote, in the Equatorial Pacific Ocean. *Science* **268,** 1480–1482 (1995).
25. Michaels, M. L. & Miller, J. H. The GO system protects organisms from the mutagenic effect of the spontaneous lesion 8-Hydroxyguanine (7,8-Dihydro-8-Oxoguanine). *J. Bacteriol.* **174,** 6321–6325 (1992).
26. Lindell, D. & Post, A. F. Ultraphytoplankton succession is triggered by deep winter mixing in the Gulf of Aqaba (Eilat). *Red Sea. Limnol. Oceanogr.* **40,** 1130–1141 (1995).
27. Karl, D. M., Bidigare, R. R. & Letelier, R. M. Long-term changes in plankton community structure and productivity in the North Pacific Subtropical Gyre: The domain shift hypothesis. *Deep Sea Res. II* **48,** 1449–1470 (2001).
28. Scanlan, D. J., Mann, N. H. & Carr, N. G. The response of the picoplanktonic marine cyanobacterium *Synechococcus* species WH7803 to phosphate starvation involves a protein homologous to the periplasmic phosphate binding protein of *Escherichia coli*. *Mol. Microbiol.* **10,** 181–191 (1993).
29. Johnson, Z. *et al.* Energetics and growth kinetics of a deep *Prochlorococcus* spp. population in the Arabian Sea. *Deep Sea Res. II* **46,** 1719–1743 (1999).
30. Dufresne, A. *et al.* Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a near minimal oxyphototrophic genome. *Proc. Natl Acad. Sci. USA* (in the press).

Supplementary Information accompanies the paper on **www.nature.com/nature.**

Acknowledgements This research was funded by the Biological and Environmental Research Program of the US Department of Energy's Office of Science. The Joint Genome Institute managed the overall sequencing effort. Genome finishing was carried out under the auspices of the US Department of Energy by the University of California, Lawrence Livermore National Laboratory. Computational annotation was carried out at the Oak Ridge National Laboratory, managed by UT-BATTELLE for the US Department of Energy. Additional support was provided by the DOE, NSF and the Seaver Foundation to S.W.C., the Israel–US Binational Science Foundation to A.F.P. and S.W.C., and FP5-Margenes to W.R.H. and A.F.P. We thank the *Synechococcus* WH8102 annotators (B. Palenik, B. Brahamsha, J. McCarren, E. Allen, F. Partensky, A. Dufresne and I. Paulsen) for their help with curating the *Prochlorococcus* genomes and E. V. Armbrust and L. Moore for critical reading of the manuscript.

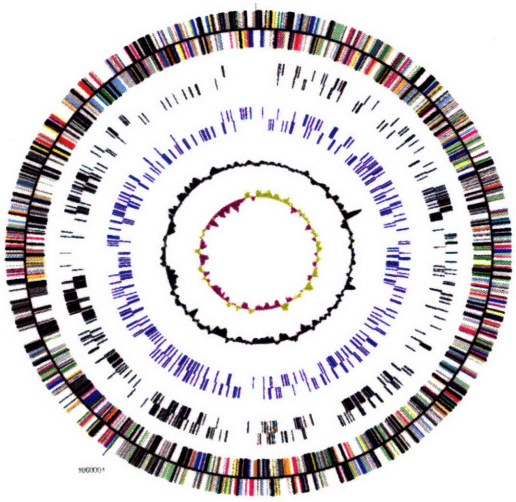
Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.W.C. (chisholm@mit.edu). The complete nucleotide sequences and sequences of predicted open reading frames have been deposited in the EMBL/GenBank/DDBJ databases under accession numbers BX548174 (MED4) and BX548175 (MIT9313).

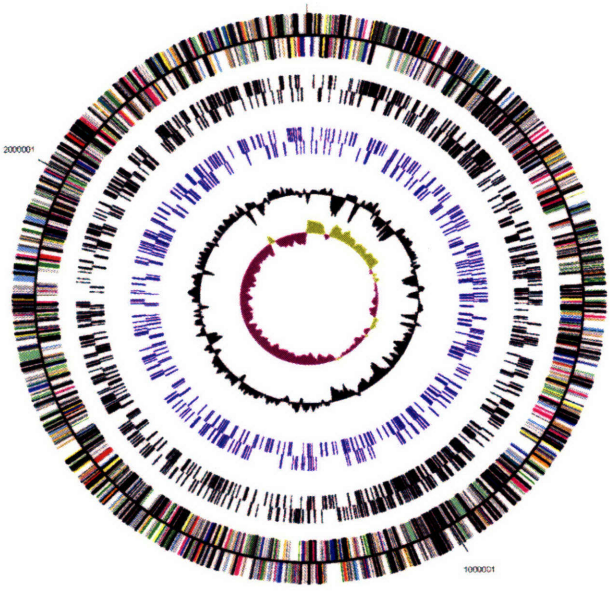
Supp. Table 1. Number of predicted signal transduction and transcription factors suggests reduced regulatory capacity in *Prochlorococcus*

	MED4	MIT 9313	*T. elongatus*
Sigma Factors	5	8	8
Two Component systems			
Histidine Kinases	4	5	17
Response regulators	6	8	27
Ser/Thr protein Kinases	0	1	11
Transcription Factors			
LuxR family	2	5	4
LysR family	1	1	3
CRP family	3	4	3
ArsR family	1	2	2
FUR family	2	3	3
Other	2	3	3
Light sensors/transducers			
Cryptochrome	2	0	2
Bacteriophytochrome	0	0	5
Phototropin	0	0	1

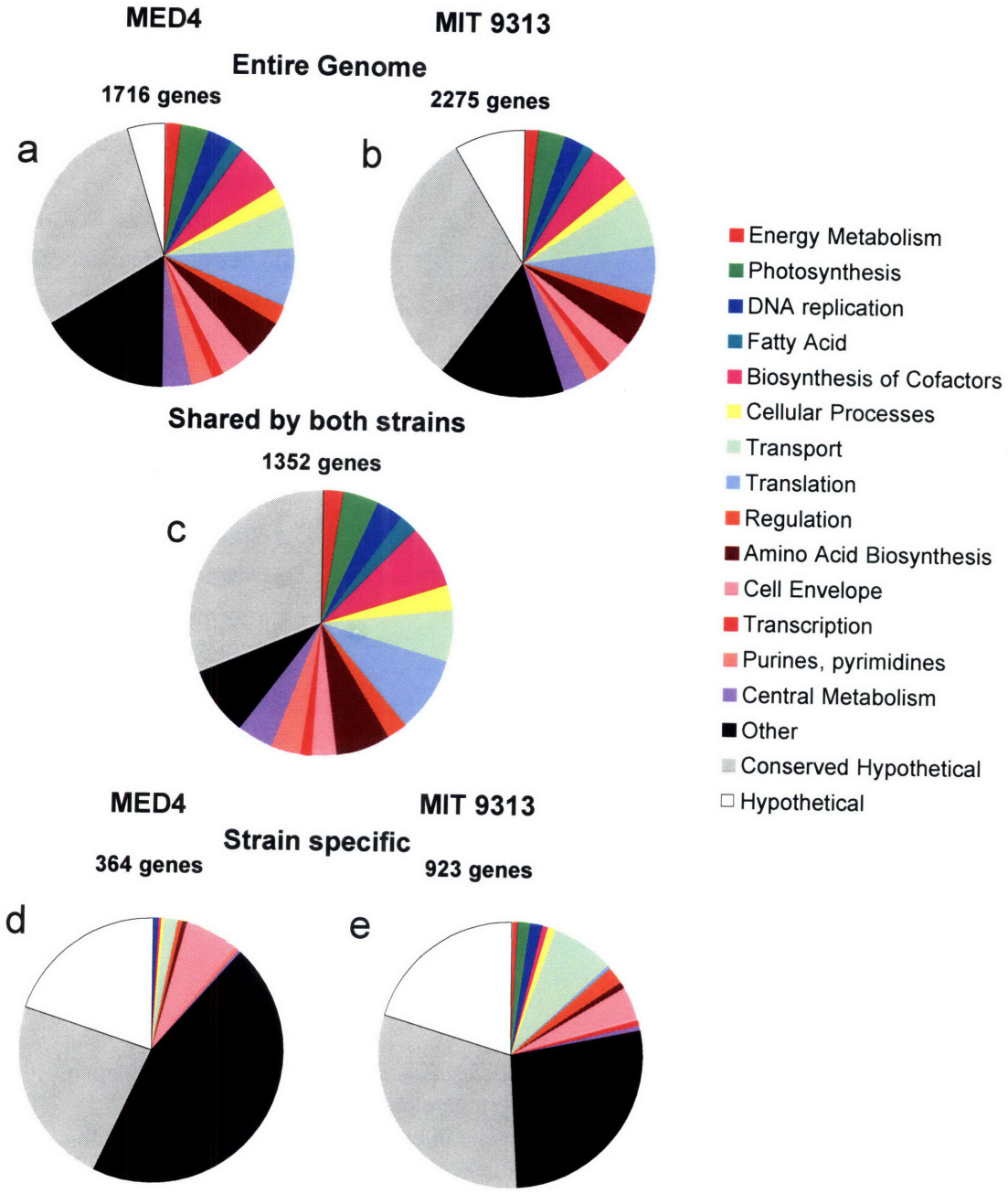
a



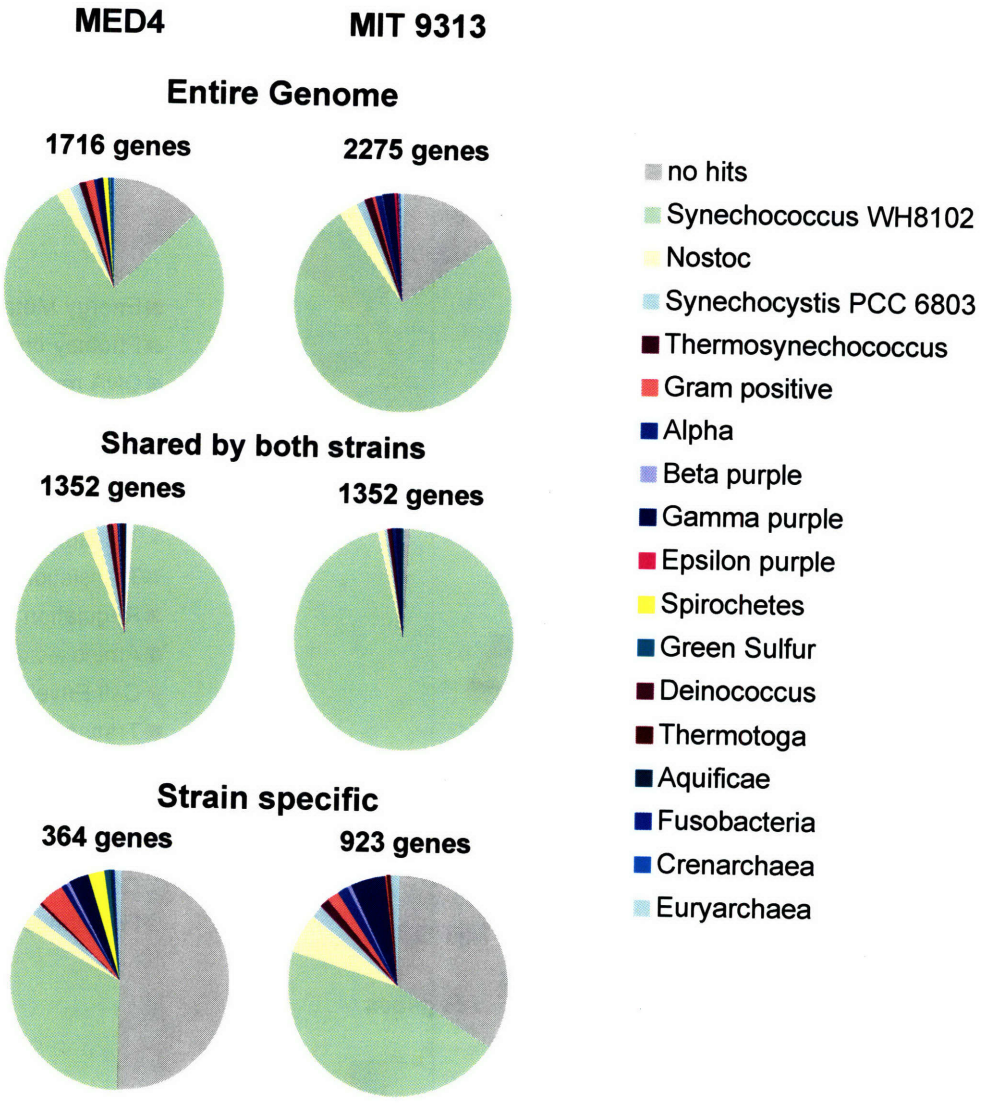
b



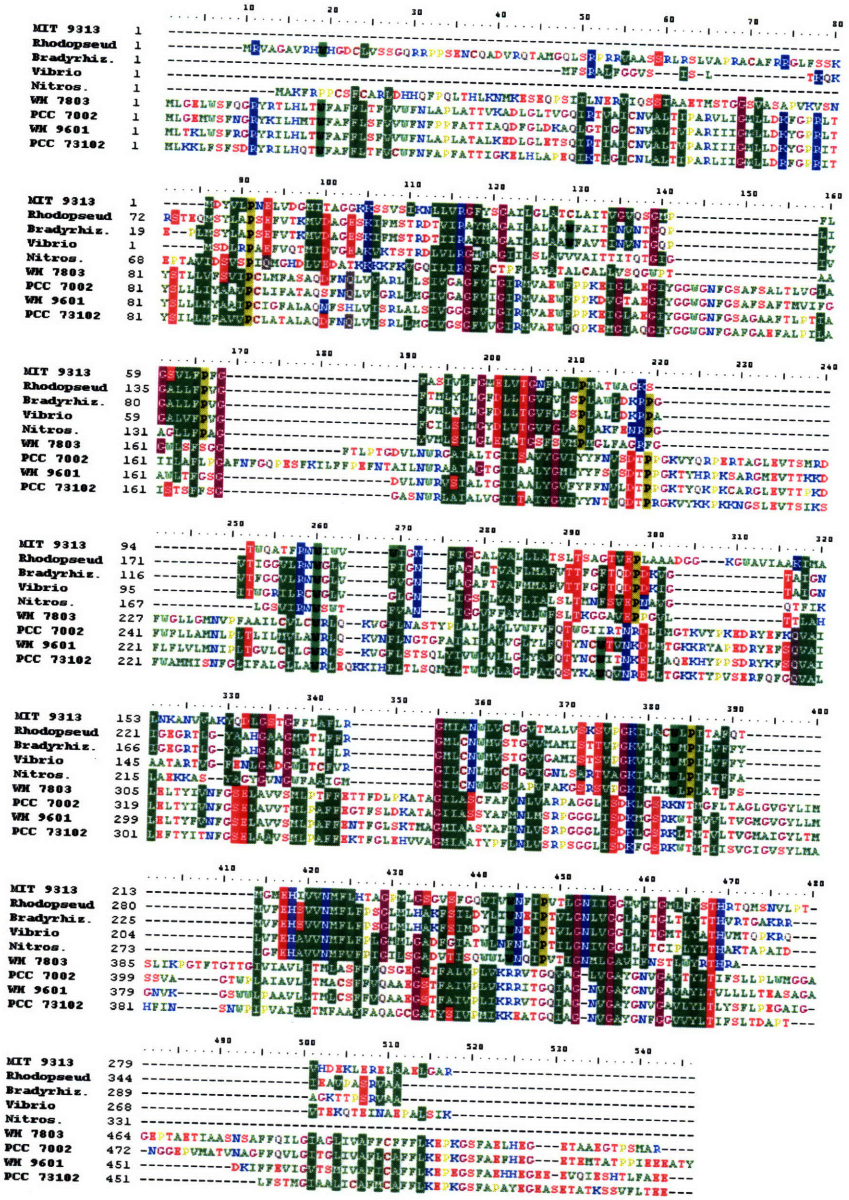
Supp. Figure 1. Circular representation of the *Prochlorococcus* genomes. a, MED4. b, MIT 9313. For both genomes outermost circles (1 and 2) are predicted protein coding regions on the plus and minus strands, respectively. Color coding is as in Supplementary Figure 2. The next two circles show genes not present in the other *Prochlorococcus* genome on the plus (circle 3) and minus (circle 4) strands. Circles 5 and 6 show genes on the plus and minus strands, respectively that contain transmembrane domains. Circle 7 is % G+C content (deviation from average). Innermost circle (8) represents the GC skew curve.



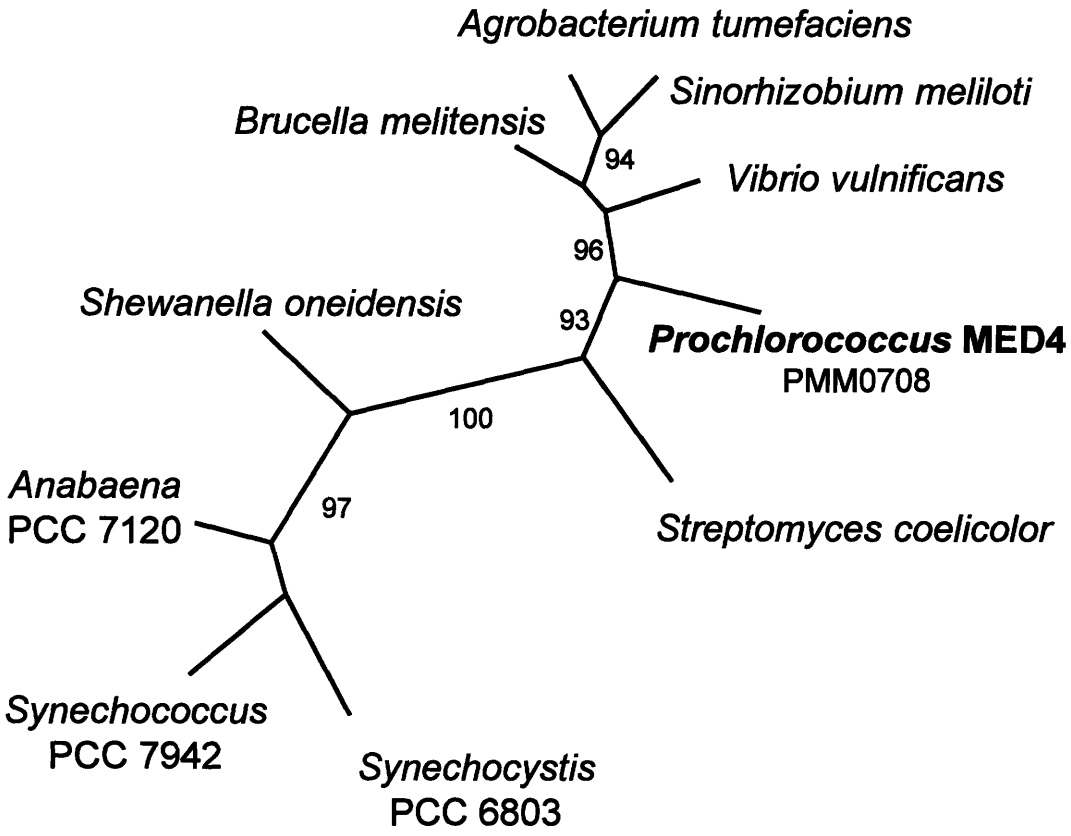
Supp. Figure 2. Functional categorization of predicted open reading frames in the *Prochlorococcus* genomes, following the classification scheme used by CyanoBase. a, MED4, entire genome. b, MIT 9313, entire genome. C, Genes present in both MED4 and MIT 9313. d, Genes in MED4 not present in MIT 9313. e, Genes in MIT 9313 not present in MED4.



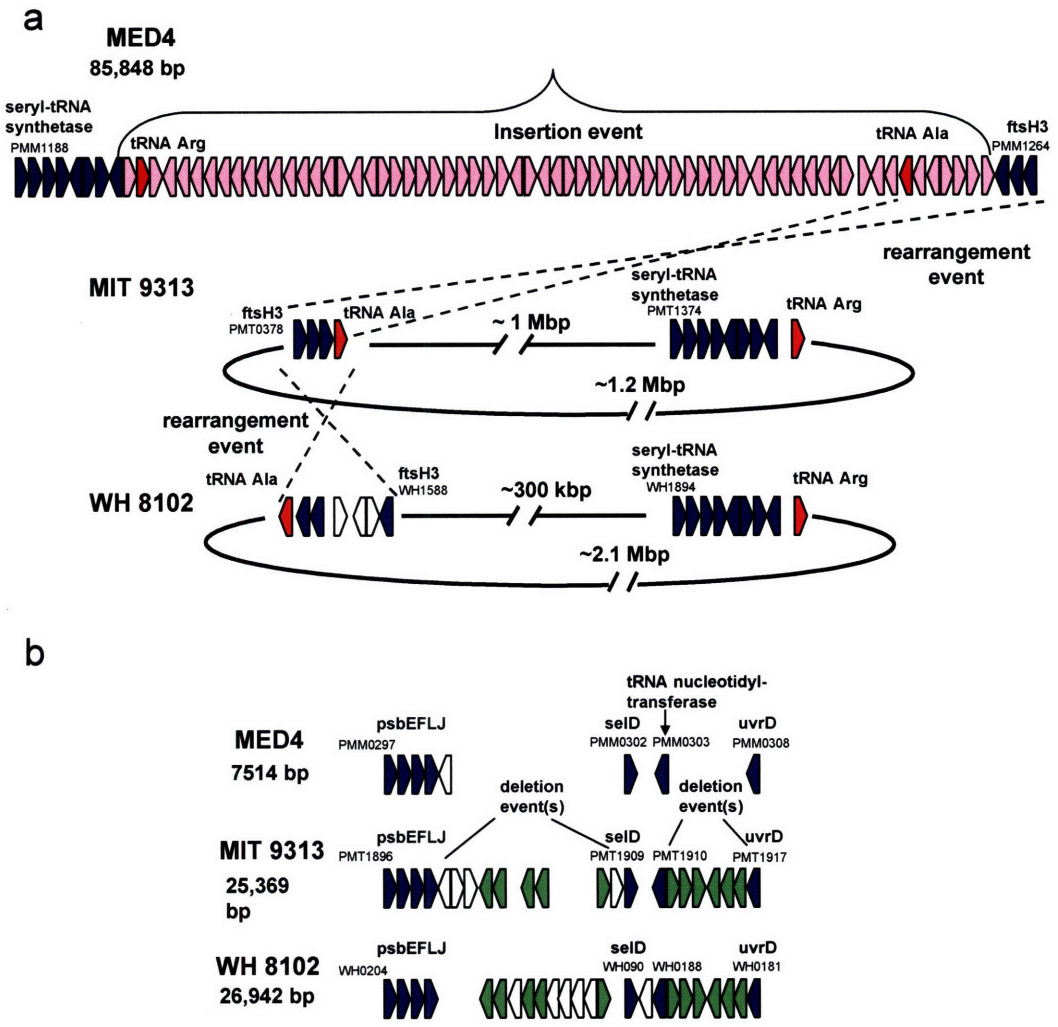
Supp. Figure 3. Comparison of *Prochlorococcus* MED4 and MIT 9313 open reading frames with those of other complete prokaryotic genomes. The predicted coding sequences of each gene in both genomes were aligned with the coding sequences of 90 bacterial genomes using BLASTP. Significant alignments were defined as having an e-value less than 10-6. The bacterial genomes comprised the 89 completed bacterial genomes available from ftp.ncbi.nih.gov/genbank/genomes/Bacteria on 30 October 2002 and *Synechococcus* WH 8102[8]. a, MED4, entire genome. B, MIT 9313, entire genome. c, MED4 genes present in MIT 9313 c, MIT 9313 genes present in MED4 e, Genes in MED4 not present in MIT 9313 f, Genes in MIT 9313 not present in MED4.



Supp. Figure 4 Alignment of the putative nitrite transporter in *Prochlorococcus* MIT9313 (PMT2240) with its most significant matches in the NR database (all proteobacteria) and with cyanobacterial nitrate/nitrate transporters. The MIT9313 gene has a formate/nitrite transporter domain (Pfam PF01226) in contrast to the cyanobacterial nitrate transporters which are permeases of the major facilitator superfamily (Pfam PF00083). Furthermore, the MIT 9313 gene has no significant matches (BLASTP evalue < e-2) in the genomes of *Prochlorococcus* MED4, *Synechococcus* WH8102, *Synechocystis* sp. PCC 6803, *Thermosynechococcus elongatus* BP-1, or *Anabaena* sp. PCC 7120 suggesting it may have been acquired via lateral gene transfer. Alignment generated using ClustalW. Shaded residues indicate >50% similarity. Abbreviations and accession numbers as follows: Rhodopseud., *Rhodopseudomonas palustris* (ZP_00012718.1); Bradyrhiz., *Bradyrhizobium japonicum* (NP_769441); Vibrio, *Vibrio vulnificus* (NP_762336.1); Nitros., *Nitrosomonas europaea* (NP_840759); WH 7803, *Synechococcus* WH 7803 *napA* (AAG45172); PCC 7002, *Synechococcus* PCC 7002 *nrtP* (AAD45941); WH9601, *Trichodesmium* WH9601 *napA* (AAF00917); PCC 73102, *Nostoc punctiforme* PCC 73102 (ZP_00107423).



Supp. Figure 5. Phylogenetic tree showing the relationship of a possible alkaline phosphatase like gene in *Prochlorococcus* MED4 (PMM0708) with the most significant matches in the NR database, which include several proteobacterial sequences, and with the atypical alkaline phosphatase of *Synechococcus* PCC 7942 and related cyanobacterial genes. Accession numbers as follows: *Brucella melitensis* (NP_541633.1), *Agrobacterium tumefaciens* str. C58 (NP_531956.1); *Sinorhizobium meliloti*, (NP_385365.1); *Vibrio vulnificus* (NP_762849.1), *Streptomyces coelicolor* A3(2) (NP_624650.1), *Shewanella oneidensis* MR-1 (NP_717877.1) *Anabaena* PCC 7102 (NP_489331.1), *Synechocystis* sp. PCC 6803 (NP_440276); *Synechococcus* sp. PCC 7942 (A47026).



Supp. Figure 6. Insertions, deletions and rearrangements of genes involved in lipopolysaccharide biosynthesis (LPS clusters) in MED4. Color coding is as follows: blue, orthologous genes present in all three genomes; pink, genes hypothesized to be part of lateral transfer events, many have roles in LPS biosynthesis; red, tRNAs; green, orthologous genes present in two genomes, many have roles in LPS biosynthesis; white, other genes. Length in bp represents the size of the region shown for each genome. a, Insertion of a 74.5 kbp cluster of LPS genes in MED4, roughly between two tRNAs. The 67 potential surface polysaccharide genes in this cluster include sugar transferases, sugar epimerases, and modifying enzymes such as aminotransferases, methyltransferases, carbamoyltransferases, and acetyltransferases. In MIT 9313 and WH 8102 the genes that flank this insertion are rearranged to other parts of the genome. b, Deletion of LPS biosynthesis genes in MED4. LPS related genes present in MIT 9313 and WH 8102, several of which have homologs in the acquired genes shown in part a, have been deleted. In this region a selenophosphate synthase (selD) and a tRNA nucleotidyl-transferase in the center of the cluster have been retained suggesting that they are essential genes and separate deletion events have occurred on either side of them.