UNIVERSITY OF CALIFORNIA,
IRVINE

Biogeography and the Adaptive Variation of Marine Bacteria in Response to Environmental Change

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biological Sciences

by

Alyssa Giselle Kent

Dissertation Committee:
Professor Adam C. Martiny, Chair
Professor Brandon S. Gaut
Professor Jennifer B. H. Martiny

2017

DEDICATION

To

My family and friends
and to the billions of bacteria sacrificed to accomplish this work

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express the deepest appreciation to my committee chair, Dr. Adam Martiny, who has been more than an advisor in many ways. I appreciated the chance to expand my skills by actually getting into the lab and I am grateful for the consistent constructive feedback over the years. Without his guidance and motivation I would not have had such a successful academic career at UCI.

I thank my committee members, Dr. Brandon Gaut and Dr. Jennifer Martiny. Brandon signed on to have me over the summer before grad school even began and was always excited to hear about the diverse aspects of my projects. Jennifer was always pushing for more out of me, and I am grateful for her incredibly helpful and thoughtful feedback over the years.

I thank the Nature Publishing group for permission to include Chapter 1 of my dissertation, which was originally published in ISME Journal. Financial support was provided by NSF-GRFP (DGE-1321846), the National Institute of Biomedical Imaging and Bioengineering (EB009418), and UCI COR grant for laboratory equipment. I would like to thank Bradford Hawkins for help with niche overlap analysis, Frédéric Partensky and David Scanlan for sharing *rpo*C1 sequences, and Sophie Mazard for sharing multi-locus alignment. I thank my coauthors from chapter 1: Chris Dupont and Shibu Yooseph and chapter 2: Céline Mouginot, Steven Baer, Jeremy Huang, and Mike Lomas.

I thank Shaun Hug for his considerable involvement in the development of my scientific intellect; I appreciated his constant availability to talk out a problem, to bring new ideas to the table, and the pleasure of his company while serving as Graduate Student reps. I thank others in my MCSB cohort, Nancy Drew, Seth Figueroa, Marissa Macchietto, Rabi Murad, Arjun Nair, and Kitt Paraiso for their friendship and the wealth of interactions I had with them and other MCSB members including Edwin Vargas who took the time to respond to my questions even before I arrived at UCI. Many members of the Department of Ecology and Evolutionary Biology became fast friends including Mark Phillips and Jimmy Kezos. I'd like to thank Microbial Group for broadening my view of the microbial world. I thank my lab mates for their assistance and friendship Cecilia Batmalle, Cathy García, Stephen Hatosy, and Alli Moreno. I appreciate the mentorship from the postdoctoral fellows and visiting scientists in our lab, Renaud Berlemont, Pedro Flombaum, Nathan Garcia, Zulema Gómez, Alyse Larkin-Swartout, Junhui Li, and Agathe Talarmin. I especially want to thank Céline Mouginot and Claudia Weihe for their support of my ventures into the lab. I thank Krista Linzner and Jessica Oquist, two undergraduate researchers who enabled me to become a better mentor.

I thank Mary Walton for her antics and her comments on parts of my dissertation. I thank Jessica Greger and Nicolas Canac, two friends I met early in grad school and were valuable resources in their respective disciplines, but also incredible champions in lifting me up no matter the circumstances. I thank Jason Vick for bringing so much laughter, love, support and relaxation to my life. Through the many ups and downs he has been the strength I never knew I needed.

Last I would like to thank my family: my brother Evan and his wife Yolanda for their support from afar and my parents Dr. Nancy Kent and Dr. Robert Kent, who followed me to California. My parents have been my inspiration for pursuing science. Without their love and support I wouldn't have made it past that first quarter.

CURRICULUM VITAE

Alyssa Giselle Kent

EDUCATION

2017	University of California, Irvine Ecology & Evolutionary Biology Doctor of Philosophy
2015	University of California, Irvine Ecology & Evolutionary Biology Master of Science
2011 - 2012	University of California, Irvine Mathematical & Computational Systems Biology Gateway Program
2007 - 2011	Lewis and Clark College Mathematical Sciences Bachelor of Arts – graduated magna cum laude

RESEARCH & TEACHING EXPERIENCE

2017-2018	Postdoctoral Researcher, Cornell University
2012-2017	Research Assistant, UC Irvine
2014-2017	Graduate Teaching Assistant, UC Irvine
2013	Selected Participant, C-MORE Summer Course on Microbial Oceanography, University of Hawaii-Manoa
2013	Algebra tutor, Anaheim, CA
2011	Research Assistant, UC Irvine
2011	Research Fellow, Competitive Edge Summer Research Program, UC Irvine
2010	Undergraduate Research Fellow, NSF-Research Experience for Undergraduates, University of Nebraska, Lincoln

PUBLICATIONS

Kent AG, Baer SE, Mouginot C, Huang J, Lomas MW, Martiny AC. Parallel biogeography across phylogenetic scales of *Prochlorococcus* and *Synechococcus*. In preparation.

Kent AG, Martiny AC. Marine bacterial lifestyle change due to adaptation to high temperature. In revision.

Linzner KA, Kent AG, Martiny AC. Evolutionary pathway determines the stoichiometric response of *Escherichia coli* adapted to high temperature. In review.

Kent AG, Dupont CL, Yooseph, S, Martiny AC. (2016). Global biogeography of *Prochlorococcus* genome diversity in the surface ocean. *The ISME journal*. doi: 10.1038/ismej.2015.265

Berube PM, Biller SJ, Kent AG, Berta-Thompson JW, Roggensack SE, Roache-Johnson KH, … & Chisholm SW. (2015). Physiology and evolution of nitrate acquisition in *Prochlorococcus*. *The ISME journal*. doi:10.1038/ismej.2014.211

Allison S, Lu L, Kent AG, and Martiny AC. (2014). Extracellular enzyme production and cheating in *Pseudomonas fluorescens* depend on diffusion rates. *Front. Microbiol.* 5:169. doi: 10.3389/fmicb.2014.00169

AWARDS AND ACHIEVEMENTS

2013-2017 Graduate Research Fellowship, NSF
2012-2016 Ecology & Evolutionary Biology Department Travel Grants, UC Irvine
2012 Interdisciplinary Research Opportunity Award, Center for Complex Biological Systems, UC Irvine
2011-2013 Predoctoral Training Grant, NIH

PRESENTATIONS

AG Kent and AC Martiny, 2016. Adaptation to rising temperature leads to increased biofilm formation in a marine *Roseobacter.* Lake Arrowhead Microbial Genomics, Lake Arrowhead, CA. [Oral & Poster]

AG Kent and AC Martiny, 2016. Adaptation to rising temperature leads to increased biofilm formation in a marine *Roseobacter.* ASM-Experimental Microbial Evolution, Washington D.C. [Poster]

KL Linzner, AG Kent, AR Moreno, AC Martiny, 2016. Stoichiometric Response of *E. coli* Adapted to High Temperature. Undergraduate Research Opportunity Program Symposium, Irvine, CA. [Poster]

AG Kent, C Mouginot, J Oquist, SE Baer, MW Lomas, AC Martiny, 2016. Biogeography in the microdiversity of *Prochlorococcus*. Southern California Geobiology Symposium, CalTech, Pasadena, CA. [Poster]

KL Linzner, AG Kent, AR Moreno, AC Martiny, 2016. Stoichiometric Response of *E. coli* Adapted to High Temperature. Southern California Geobiology Symposium, CalTech, Pasadena, CA. [Poster]

AC Martiny, AG Kent, C Mouginot, SE Baer, MW Lomas, 2016. Strong latitudinal and vertical biogeography of *Synechococcus* diversity in the equatorial Pacific Ocean. Ocean Sciences Meeting, New Orleans, LA. [Oral]

AG Kent, C Mouginot, J Oquist, SE Baer, MW Lomas, AC Martiny. 2016. Phylogeography of *Prochlorococcus* nitrate reductase gene across ocean biomes. Department of Ecology and Evolutionary Biology Graduate Student Symposium, University of California, Irvine. Irvine, CA. [Oral]

KL Linzner, AG Kent, C Mouginot, AC Martiny, 2015. Physiological and Genetic Response of *E. coli* Adapted to High Temperature. Undergraduate Research Opportunity Program Symposium, Irvine, CA. [Poster]

AG Kent, CL Dupont, S Yooseph, AC Martiny. 2014. Global biogeography of *Prochlorococcus* genome diversity. Southern California Geobiology Symposium, University of Southern California, Los Angeles, CA. [Oral]

AG Kent, CL Dupont, AC Martiny. 2014. Global biogeography of *Prochlorococcus* genome diversity. Center for Complex Biological Systems Retreat, University of California, Irvine, Pasadena, CA. [Poster]

AG Kent, CL Dupont, AC Martiny. 2014. Global biogeography of *Prochlorococcus* genome diversity. Department of Ecology and Evolutionary Biology Graduate Student Symposium, University of California, Irvine. Irvine, CA. [Oral]

AG Kent and AC Martiny. 2013. Geographical differentiation of the genome content of the marine bacteria *Prochlorococcus*. Southern California Regional Conference

in Systems Biology, Irvine CA. [Poster]

AG Kent and AC Martiny. 2012. Geographical differentiation of the genome content of the marine cyanobacterium *Prochlorococcus*. International Society for Microbial Ecology-14, Copenhagen, Denmark. [Poster]

AG Kent and AC Martiny. 2012. Cluster analysis of *Prochlorococcus* gene abundance from widely distributed oceanic samples. NIBIB Training Grantees Meeting, Bethesda, MD. [Poster]

AG Kent, W Zeng, A Mortazavi. 2011. Discovering the evolution of NRSF motifs through comparative genomics. Developmental and Cell Biology Retreat, University of California, Irvine, Dana Point, CA. [Poster]

AG Kent and BS Gaut. 2011. Determining selection on salt tolerance genes in Arabidopsis. Competitive Edge Symposium, University of California, Irvine, Irvine, CA. [Oral]

AG Kent, KK Buschkamp, MS Eickholt, L Ohm, GM Shakan, S Reynolds, and G Ledder. 2011. Asymptotic herbivory and optimal resource allocation: A cause for masting. Lewis & Clark College Spring Math Colloquium Series, Portland, OR. [Oral]

KK Buschkamp, MS Eickholt, AG Kent, L Ohm, GM Shakan, S Reynolds, and G Ledder. 2011. Asymptotic herbivory and optimal resource allocation: A cause for masting. Joint Mathematics Meetings, New Orleans, LA. [Oral & poster]

SERVICE

2016-2017 Organizer, Department Seminar Reception, Ecology and Evolutionary Biology, UC Irvine

2012-2016 Student mentor, Undergraduate Research Opportunity Program, UC Irvine

2015-2016 Graduate Student Representative, Ecology and Evolutionary Biology, UC Irvine

2014-2015 Organizer, Winter Ecology and Evolutionary Biology Graduate Student Symposium, UC Irvine

2013-2014 Organizer, Ecology and Evolutionary Biology Department Recruitment, UC Irvine

2013-2014 Student mentor, Gateway Mentoring Program, UC Irvine

2011-2012 Student Council, Diverse Educational Community and Doctoral Experience, UC Irvine

SOCIETY MEMBERSHIP

2014-2017 American Association for the Advancement of the Sciences

2016 American Society for Microbiology

2012 International Society for Microbial Ecology

2011 Phi Beta Kappa Honors Society

2011 Pi Mu Epsilon Mathematical Honors Society

2011 American Mathematical Society

2010 Association for Women in Mathematics

ABSTRACT OF THE DISSERTATION

Biogeography and the Adaptive Variation of Marine Bacteria
in Response to Environmental Change
By

Alyssa Giselle Kent

Doctor of Philosophy in Biological Sciences

University of California, Irvine, 2017

Professor Adam C. Martiny, Chair

Prochlorococcus, the smallest known photosynthetic bacterium, is abundant in the ocean's surface layer despite large environmental variation. There are several phylogenetically distinct lineages within *Prochlorococcus* and considerable gene gain and loss throughout its evolutionary history. However, the extent to which vertical versus horizontal inheritance shapes its genome diversity across the global oceans is unknown. We observed that *Prochlorococcus* field populations from a global circumnavigation had a significant relationship between phylogenetic and gene content diversity including regional differences in both phylogenetic composition and gene content that were related to environmental factors. Overall we showed that the environment determines the functional capabilities of successful *Prochlorococcus*.

We know *Prochlorococcus* has extensive genetic diversity*,* including the presence of multiple major clades, its sister taxa *Synechococcus* displays similar levels of genetic diversity. *Prochlorococcus* has a clear phylogeography relating to environmental selective pressures, while the biogeography and environmental drivers of *Synechococcus* clades are more difficult to define. To better characterize

Prochlorococcus and *Synechococcus* genetic diversity we used high throughput sequencing of the marker gene *rpo*C1 from 339 samples across the Pacific Ocean and Atlantic Oceans. At multiple taxonomic scales (lineage, clade, and SNP) we observed clear parallel biogeography between these two lineages. Overall, this parallel biogeography suggests similar evolutionary selective pressures for these important marine Cyanobacteria.

Oceans are warming and will continue to increase over the next 100 years due to global climate change. Adaptation will likely play a role, but it is unclear how it will impact microbial distributions and processes. To address this unknown, we experimentally evolved a member of the prevalent marine *Roseobacter* clade to high temperature for 500 generations. We found that this evolved *Roseobacter* shifted from its usual planktonic growth mode to creating more biofilm at the surface of the culture. Furthermore, this altered lifestyle was coupled with a suite of genomic changes linked to biofilm formation and increased growth in low oxygen transfer environments.

INTRODUCTION

Oceans harbor some of the most abundant organisms on the planet. While these organisms are small in size, they account for a vast majority of the global primary production (Flombaum *et al.*, 2013; Partensky *et al.*, 1999). Some of the major primary producers are Cyanobacteria, which are dominated numerically by *Prochlorococcus* and *Synechococcus* (Flombaum *et al.*, 2013; Bouman *et al.*, 2006).

All *Prochlorococcus* lineages vary by less than 3% in their 16S rRNA gene, yet their genome content varies drastically (Kettler *et al.*, 2007; Biller *et al.*, 2015). An individual *Prochlorococcus* has a core genome of approximately 1000 genes shared by all *Prochlorococcus* and around another 1000-2000 genes that they carry but are variable throughout the lineage. This pool of flexible genes has been estimated at ~85,000 genes based on rarefaction of 41 fully sequenced genomes, single cell genomes, and assembled metagenomes (Biller *et al.*, 2015). Indeed, this diversity in flexible accessory genes likely enables the widespread proliferation of *Prochlorococcus* across open ocean waters (Partensky *et al.*, 1999; Kettler *et al.*, 2007) as many of these genes confer environmentally adaptive traits (Martiny *et al.*, 2006; Berube *et al.*, 2015). Yet to what extent this diversity varies across ocean regions has not been elucidated.

To address this, in my first chapter, I use a global ocean circumnavigation coupled with high throughput metagenomic sequencing of surface water to answer the following questions: 1) What is the biogeography of *Prochlorococcus* phylogenetic diversity? 2) What is the biogeography of *Prochlorococcus* genome content diversity? 3) Are these two components of diversity linked to one another? 4) What drives the

biogeographic patterns in each? 5) Is there regional variation in the phylogenetic and genome content diversity?

While *Prochlorococcus* has a clear biogeography that is linked to its evolutionary history, *Synechococcus* has been harder to define (Mazard *et al.*, 2012). Different taxonomic markers and choices in taxonomic resolution have resolved different aspects of its evolutionary history (Tai & Palenik, 2009; Sohm *et al.*, 2016; Farrant *et al.*, 2016) and multi-locus sequencing has strengthened the support for a cohesive structure (Mazard *et al.*, 2012). Yet a clear picture of the biogeography of clades within *Synechococcus* and how they are connected evolutionarily, in particular at what taxonomic level is there biogeography, has also not been fully resolved.

Therefore, in my second chapter, I use the marker gene *rpo*C1 to identify the biogeography at different levels of taxonomic variation across three ocean transects in the eastern Pacific and the northern Atlantic Oceans to answer the following questions: 1) What is the biogeography of *Prochlorococcus* and *Synechococcus* across light, temperature, and nutrient gradients? Does the biogeography of each lineage correlate? Are there patterns of biogeography within the microdiversity of these lineages?

While the first two chapters of my dissertation assess aspects of marine microbial ecology and evolution observationally by surveying *in situ* populations of *Prochlorococcus* and *Synechococcus*, it is difficult to comprehend how organisms will adapt to a changing climate solely from observational data (Collins, 2010). I address the topic of marine microbes in future oceans by experimentally evolving a member of the *Roseobacter* lineage to high temperature. Previous experimental evolution studies have addressed evolution to high temperature in *E. coli* (Tenaillon *et al.*, 2012), yet it is

unclear how generalizable these results are to organisms from different environments and what physiological consequences of adaptation to high temperature may arise. Consequently we ask the following questions: 1) Does growth rate increase in *Roseobacter* adapted to high temperature? 2) Does cellular stoichiometry differ between ancestral and evolved lineages? 3) What are the genomic changes associated with observable phenotypic changes? 4) Are they similar to *E. coli* evolved to high temperature?

References

Berube PM, Biller SJ, Kent AG, Berta-Thompson JW, Roggensack SE, Roache-Johnson KH, *et al.* (2015). Physiology and evolution of nitrate acquisition in *Prochlorococcus*. *ISME J* **9**:1195–1207.

Biller SJ, Berube PM, Lindell D, Chisholm SW. (2015). *Prochlorococcus*:the structure and funciton of collective diversity. *Nat Rev Micro* **13**:13–27. 10.1038/nrmicro3378.

Bouman HA, Ulloa O, Scanlan DJ, Zwirglmaier K, Li WKW, Platt T, *et al.* (2006). Oceanographic Basis of the Global Surface Distribution of *Prochlorococcus* Ecotypes. *Science* **312**:918–921.

Collins S. (2010). Many Possible Worlds: Expanding the Ecological Scenarios in Experimental Evolution. *Evol Biol* **38**:3–14.

Farrant GK, Doré H, Cornejo-Castillo FM, Partensky F, Ratin M, Ostrowski M, *et al.* (2016). Delineating ecologically significant taxonomic units from global patterns of marine picocyanobacteria. *Proc Natl Acad Sci* **113**:E3365–E3374.

Flombaum P, Gallegos JL, Gordillo R a, Rincón J, Zabala LL, Jiao N, *et al.* (2013).

Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proc Natl Acad Sci U S A* **110**:9824–9829.

Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, *et al.* (2007). Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* **3**:2515–2528.

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.

Mazard S, Ostrowski M, Partensky F, Scanlan DJ. (2012). Multi-locus sequence analysis, taxonomic resolution and biogeography of marine *Synechococcus. Environ Microbiol* **14**:372–386.

Partensky F, Hess WR, Vaulot D. (1999). *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *MicrobiolMol BiolRev* **63**:106–127.

Sohm JA, Ahlgren NA, Thomson ZJ, Williams C, Moffett JW, Saito MA, *et al.* (2016). Co-occurring *Synechococcus* ecotypes occupy four major oceanic regimes defined by temperature, macronutrients and iron. *ISME J* **10**:333–345.

Tai V, Palenik B. (2009). Temporal variation of *Synechococcus* clades at a coastal Pacific Ocean monitoring site. *Isme J* **3**:903–915.

Tenaillon O, Rodríguez-Verdugo A, Gaut RL, McDonald P, Bennett AF, Long AD, *et al.* (2012). The molecular diversity of adaptive convergence. *Science* **335**:457–461.

CHAPTER 1

Global biogeography of *Prochlorococcus* genome diversity in the surface ocean

Abstract

Prochlorococcus, the smallest known photosynthetic bacterium, is abundant in the ocean's surface layer despite large variation in environmental conditions. There are several genetically divergent lineages within *Prochlorococcus* and superimposed on this phylogenetic diversity is extensive gene gain and loss. However, it is unknown how vertical and lateral modes of evolution relatively shape the global ocean distribution of genome diversity in *Prochlorococcus*. Here, we show that *Prochlorococcus* field populations from a global circumnavigation harbor extensive genome diversity across the global surface ocean, but this diversity is not randomly distributed. We observed a significant correspondence between phylogenetic and gene content diversity including regional differences in both phylogenetic composition and gene content that were related to environmental factors. Several gene families were strongly associated with specific regions and environmental factors, including the identification of a set of genes related to lower nutrient and temperature regions. In support, metagenomic assemblies of natural *Prochlorococcus* genomes reinforced this association by providing linkage of genes across genomic backbones. Overall our results show, that the environment determines the functional capabilities of *Prochlorococcus* succeeding in that environment, influencing the ecological effects of the most abundant photoautotroph in the world.

Keywords: Comparative metagenomics, phylogeny, marine diversity, Global Ocean Sampling

Introduction

Prochlorococcus is the most abundant marine phytoplankton and an important contributor to global primary production (Flombaum *et al.*, 2013; Bouman *et al.*, 2006). Within the group, there are multiple phylogenetically distinct lineages and the distribution of this diversity has been linked to environmental factors such as light, temperature, and iron availability (Zwirglmaier *et al.*, 2008; Moore *et al.*, 1998; West & Scanlan, 1999; Johnson *et al.*, 2006; Rusch *et al.*, 2010). It has been suggested that pan-genome variability allows *Prochlorococcus* as a genus to proliferate across many environments (Partensky *et al.*, 1999), but it is unclear how the mechanisms of descent contribute to their evolution and the diversity of local populations. Vertical descent, specifically the transfer of genes through cellular division, and horizontal gene transfer, encompassing the several mechanisms of moving genes between organisms without division, are two ways that genes can arrive in an organism. . Based on a 'genomic backbones' concept (Kashtan *et al.*, 2014), where a stable niche partitioning of sub-populations of *Prochlorococcus* is proposed, it should follow that phylogeny will largely correspond to the biogeography of the overall gene content However, the genomic presence of many genes is highly variable as gene gain or loss may be a driving force in the evolution of *Prochlorococcus* (Kettler *et al.*, 2007). Variables affecting simple traits like nutrient acquisition may better explain the flexible gene content than variables associated with complex traits requiring more exchanged genes such as light-acquisition and DNA repair (Martiny *et al.*, 2013; Brown *et al.*, 2014). For example, some of these flexible genes are found in association with phosphate acquisition (Martiny *et al.*, 2006), where the genomes of *Prochlorococcus* cells in environments with

lower phosphate concentrations are enriched in phosphate acquisition genes (Martiny *et al.*, 2009a). While a few sets of flexible genes have well characterized distributions, we know little about the global biogeography of *Prochlorococcus* genome content diversity for the whole of the pan genome. Understanding the distribution of phylogenetic and genomic diversity across environments will provide insight into the selective pressures in the context of vertical inheritance structuring *Prochlorococcus* genome contents and community composition and ultimately lead to identifying drivers of the global biogeography and biogeochemical function of this key marine phytoplankton lineage.

Previous investigations of *Prochlorococcus* biogeography have characterized phylogenetic diversity using specific genetic markers that reflect genome content to an unknown extent (Bouman *et al.*, 2006; Zinser *et al.*, 2006; Zwirglmaier *et al.*, 2008). Metagenomics has also been employed to examine *Prochlorococcus* genomic diversity at smaller regional scales (Hewson *et al.*, 2009; Rusch *et al.*, 2010), but studies have not yet described variation in *Prochlorococcus* genome content itself on a global scale, let alone in comparison to its phylogenetic diversity directly. Here we examine the biogeography of *Prochlorococcus* genome diversity, both in terms of phylogenetic and genome content, across all major ocean basins. We first quantify the biogeography of the phylogenetic diversity based on a set of single-copy core genes. We expect that phylogenetic diversity will largely be driven by factors delineating the major ecotype clades, specifically light followed by temperature (Moore *et al.*, 1998; Johnson *et al.*, 2006; Coleman & Chisholm, 2007; Martiny *et al.*, 2009b). Next, we analyze the distribution of the flexible genome content by asking if there are geographic patterns in gene distributions, if these patterns are driven by environmental factors and if there are

specific genes that are linked to ocean regions. Finally, we determine the extent to which variation in phylogenetic diversity corresponds with difference in the flexible gene content. Although *Prochlorococcus* has been well characterized phylogenetically, identifying the factors controlling its genome content diversity across ocean environments is central to understanding the ecological role of phylotypes within a key lineage in marine ecosystems.

Materials and Methods

We searched for sample sequences highly similar to known *Prochlorococcus* genes (Biller *et al.*, 2014) (see *Prochlorococcus sequence calling* below) and used geographic and environmental data to define the spatial distribution of the phylogenetic lineages and flexible orthologous genes. Subsequent analyses used 87 metagenomes for the phylogenetic, core gene distribution and 56 metagenomes for the non-core or flexible gene content analyses.

Metagenomic Samples

Metagenomic samples from 226 GOS sites were analyzed in this study (Genbank bioproject PRJNA13694 and European Bioinformatics Institute accession numbers ERX913362-ERX913706). Environmental variables were either measured while sampling, (temperature, sample depth, ocean depth, latitude, and longitude) or determined from monthly averages (nitrate and phosphate) for the relevant locations using the World Ocean Atlas 2009 (Garcia *et al.*, 2010) at a 1-degree and 5 m depth resolution. If the sample's coordinates rounded to a land location, we averaged the values from the adjacent oceanic grid points. Temperature ranged from 12.1 to 37.6 °C, sample depth ranged from the surface to 62 m, ocean depth ranged from 0.33 m to

5800 m, nitrate ranged from 0 to 7.54 µmol/L, and phosphate ranged from 0.002 to 0.953 µmol/L. We defined regions first by separating between the oceans, and then by using the 0.3 µmol/L phosphate contours to delineate equatorial regions (Table S1.1).

Reference database

Our reference database was built from 2 *Prochlorococcus* metagenomic assemblies (HLIII and HLIV - nearly complete except for genomic islands), 41 *Prochlorococcus*, and 15 *Synechococcus* fully sequenced genomes (Biller *et al.*, 2014). We used ProPortal v.4 (Kelly *et al.*, 2012) clusters annotated using the RAST server (Aziz *et al.*, 2008) to define our orthologous groups. The non-core genome consists of 8027 genes found in at least one but not all of the *Prochlorococcus* whole genomes. The core genome consists of 504 single-copy genes, which are present in each of the *Prochlorococcus* and *Synechococcus* genomes and metagenomic assemblies. The database had a core genome of 781 genes and a flexible genome of 15 673 genes. *Prochlorococcus*, in particular, had a core genome of 975 genes with 8027 non-core genes (Figure S1.1). Among the full *Prochlorococcus* genomes, 91.5% are single copy.

Prochlorococcus sequence calling

All pyrosequencing and Sanger metagenomic sequences were co-assembled using the CELERA assembler (Miller *et al.*, 2008) at 92% nucleotide identity. The resulting scaffolds encompass 3 Gbp of contiguous DNA sequence, while 85% of the sequence reads could be mapped back to the assembly. Open reading frames (ORFs) on scaffolds were called using MetaGene (Noguchi *et al.*, 2006). To determine the putative phylogenetic origin of the scaffolds, each predicted peptide was phylogenetically annotated using Automated Phylogenetic Inference System (APIS)

(Dupont *et al.*, 2014), which annotates according to the position of the peptide within a phylogenetic tree. Thus a peptide 99% similar to a *Prochlorococcus* protein will be annotated as *Prochlorococcus* (with the associated taxonomic tree), while a peptide that branches basally within the phylogenetic tree next to Cyanobacteria would only be annotated as Cyanobacteria. The scaffolds were taxonomically annotated at the lowest level for which greater than 50% of the ORFs had agreement in the APIS calls. All reads mapped to scaffolds defined as Cyanobacteria were then matched against protein coding regions from our reference database using BLASTn (Camacho *et al.*, 2009), e-value < 10^{-5}. To increase confidence in called sequences, we included *Prochlorococcus*' nearest evolutionary neighbor, *Synechococcus*, as an outgroup in the sequence calling, but all analyses only examine *Prochlorococcus*-related sequences.

Phylogenetic Analysis

We built a reference phylogeny using the 504 core orthologous groups aligned separately in protein space using TranslatorX (Abascal *et al.*, 2010) with ClustalW (Larkin *et al.*, 2007) then concatenated the gene alignments (544 614 bp) to build the maximum likelihood tree using Phylip's DNAML (default parameters, constant rate variation model with WH5701 as outgroup and re-rooted with *Synechococcus* to have a clearer distinction between genera). We analyzed 100 bootstrap resamplings using SEQBOOT (Felsenstein, 2005). For each orthologous group we mapped metagenomic sequences to orthologous groups, used hmmalign from HMMER 3.1b1 (Eddy, 2011) to align sequences to the individual gene reference alignments, PPlacer (Matsen *et al.*, 2010) coupled with RAxML (GTRGAMMA model) (Stamatakis, 2014) to generate the appropriate phylogenetic model statistics and map the sequences onto the reference

phylogeny one core gene at a time, and finally we collapsed single gene branch abundance matrices yielding a phylogenetic distribution across samples (Figure S1.2 and Figure S1.3). We placed a threshold at 500 sequences per sample and rarefied to account for sampling depth biases. After determining dissimilarity (Bray-Curtis) we clustered the samples hierarchically using the 'average' clustering algorithm (R Development Core Team, 2013). We bootstrapped the data to determine the robustness of the cluster signal. We measured the association between sampling regions and phylogenetic clades or deep groups, using a point biserial correlation coefficient (r.g) corrected for differences in sampling sizes (alpha = 0.05) (function 'multipatt') (De Cáceres & Legendre, 2009) (Table S1.2). We used the permutation multivariate analysis of variance test (function adonis) (Oksanen *et al.*, 2013) to test if regions explain clade abundance variation. We used a test for homogeneity of multivariate dispersion (betadisper, TukeyHSD) (Oksanen *et al.*, 2013; R Development Core Team, 2013). There was a significant difference in spread between a few of the pairwise combinations in the phylogenetic analysis (Figure S1.4). This suggested possible differences in both dispersion and location.

Gene Content Analysis

Sequences of high similarity to non-core clusters (best hit, BLASTn, e-value < 10^{-5}) (Camacho *et al.*, 2009) were used in the gene content analysis. We placed a threshold of 1500 sequences per sample and rarefied to account for sampling depth biases. We used the permutation multivariate analysis of variance test (function adonis) (Oksanen *et al.*, 2013) to test if regions explain abundance variation. Using sequence similarity to non-core genes (BLASTp) we identified COG categories from

cyanobacterial eggNOG database (Powell *et al.*, 2014). We measured the association between sampling regions and either COG categories using a point biserial correlation coefficient (r.g) (Table S1.3) or orthologous genes using an indicator value index (indval.g) (Table S1.4), both corrected for differences in sampling sizes (alpha = 0.05) (function 'multipatt')(De Cáceres & Legendre, 2009), but not corrected for multiple testing.

Environmental analysis

We used Canonical Correspondence Analysis to determine the most significant variables with forward selection for both the phylogenetic and the genome content signal (Canoco, v4.5) (ter Braak, 1986) (Table 1.1). Ordination plots were visualized with the 'vegan' package in R (Oksanen *et al.*, 2013).

Genetic distance in phylogeny

Genetic pairwise distances were calculated using Phylip's DNADIST (1). Incorporation of the two HNLC metagenomes and the additional fully sequenced genomes greatly increased the maximum pairwise sequence distance (the proportion of differing nucleotide positions between two sequences) from 0.27 to 0.39.

Transporters

Most transporters that significantly differed among regions were phosphate or nitrogen related as described in Table S1.4. Other transporters associated with a variety of regions: a Zinc transporter negatively associated with the Equatorial Pacific Ocean, a Mn2+ and Fe2+ transporter was found negatively associated with the North Atlantic Ocean, a Na+-dependent transporter was found positively associated with the South Atlantic and Equatorial Pacific Oceans, a glycine betaine transporter was positively

associated with the Equatorial Pacific and North Atlantic Oceans, and a cobalt transporter was positively associated with the Equatorial Atlantic Ocean.

Temperature-related genes

We identified 85 potentially low temperature adaptive genes by first examining orthologous groups with a negative correlation to temperature ($r^2 < -0.3$, $p<0.027$) (Table S1.5). From this set, three of the COGs with the strongest negative correlation temperature and three of the COGs unique or nearly unique to the high-light I (HLI) clade were used in a further analysis. To strengthen the validity of each putatively low temperature gene, we identified scaffolds assembled from the entire GOS dataset with the gene present and identified the nearest two genes closest on each metagenomic assembly (Figure S1.5). We compared the distribution of all reads mapped to assemblies with the lower temperature gene against all reads mapped to assemblies with a neighbor present, but missing the lower temperature gene of interest. We normalized read counts to total *Prochlorococcus* reads in each respective region allowing us to compare across regions and we normalized to the number of reads in each set of assemblies to compare across COGs. We also identified whether a read mapped to a member of the HLI clade or not based on our previous BLASTn calls (Camacho *et al.*, 2009).

Phylogenetic vs. Gene Content Comparison

To compare the phylogenetic variance with the gene content variance we used the Mantel test in 'vegan' package (function mantel) (Oksanen *et al.*, 2013) to test for a correlation between the phylogenetic and gene content sample dissimilarity matrices.

Results

We analyzed a complete global circumnavigation of 226 metagenomic samples (Figure 1.1 and Table S1.1). With the recent addition of 29 strains to the set of fully sequenced *Prochlorococcus* genomes (Biller *et al.*, 2014), we also created a maximum likelihood phylogeny (Figure 1.2A). We first quantified the biogeography of phylogenetic diversity by mapping sequence reads onto this detailed phylogeny and, despite the presence of every clade in every region, we observed unequal lineage representation both in terms of relative abundance and regional distribution (Figure S1.1). As predicted, *Prochlorococcus* phylogenetic community composition varied significantly between regions (permutational multivariate ANOVA, R^2=0.47, p=0.0001)(Figure 1.2B, C).

Prochlorococcus lineages varied in relative frequency among ocean basins. The strongest association was between clade c2, (HLI) and the South Atlantic Ocean and California Current regions (indicator analysis, r.g= 0.80, p=0.001). We also found that sequences mapping to clade c1 (HLIII and HLIV) significantly dominated the Equatorial Pacific Ocean samples (indicator analysis, r.g=0.53, p=0.008) and was consistent with a past analysis of this region, albeit utilizing similar datasets (Rusch *et al.*, 2010). Among HLII clades, c9 (incl. strain MIT9301) was the most frequent subclade and together with c5, c6, and c7 common across all regions except the California Current, South Atlantic Ocean, and Equatorial Pacific Ocean (Table S1.2). Although clade c4 was within the HLII group, it corresponded to the same set of regions but was found at lower frequency in the North Atlantic Ocean (indicator analysis, r.g. = 0.64, p = 0.001). Another HLII group, clade c8, was positively correlated with all regions except the South Atlantic Ocean (indicator analysis, r.g. = 0.55, p = 0.001). Most LL clades (c10, c12, c13)

correlated significantly with the California Current (Table S1.2), a region with a subset of samples deeper than 5 m (Table S1.1) (Dupont *et al.*, 2015). All other clades did not differ in their relative abundance among regions.

In a canonical correspondence analysis, nitrate and temperature accounted for most of the variation in phylogenetic diversity among regions. In contrast, phosphate, sample depth (proxy for light availability), and bottom depth (proxy for coastal influence) had smaller, albeit still significant, effects (Table 1.1). The majority of the Equatorial Pacific Ocean samples grouped with a few from the North Indian and South Pacific Ocean samples along the first canonical axis (CCA1), which predominantly corresponded to higher nitrate (and presumably lower iron) and an elevated frequency of clade c1 (Figure 1.3A). The California Current and the South Atlantic Ocean samples negatively corresponded with temperature and positively with sample depth along the second canonical axis (CCA2) and clade c2 [the lower temperature adapted HLI clade (Johnson *et al.*, 2006)].

We then determined the biogeography of the *Prochlorococcus* gene content using the relative frequency of 1663 recovered orthologous groups of non-core genes between sites not discounted by threshold or rarefaction. Gene content displayed clear biogeographic patterns across regions (permutational multivariate ANOVA, R^2=0.16, p=0.0001). Similar to phylogenetic diversity, gene frequencies were most significantly related to nitrate concentrations and temperature (CCA analysis: p=0.0001 and p=0.0002 respectively; Table 1.1). The additional variables (bottom depth, phosphate, and sample depth) accounted for smaller proportions of variation. Nevertheless, the North Atlantic Ocean samples appeared to cluster together in a negative association

with phosphate concentrations (Figure 1.3B), although phosphate was not a significant environmental factor in our forward selection analysis (CCA analysis: p = 0.1092; Table 1.1). The phylogenetic and gene content diversity of *Prochlorococcus* populations was significantly correlated (R_{Mantel} = 0.72, p=0.001; Figure 1.3C) and thus, sample sites more similar in their phylogeny were more similar in their gene content.

To understand regional differences in *Prochlorococcus* functional potential, we grouped non-core genes into functional categories as defined by the Clusters of Orthologous Groups of proteins (COGs) (Tatusov *et al.*, 1997). The COG categories differentiated significantly with respect to region (permutational multivariate ANOVA, R^2=0.17, p=0.0069). Populations from the North and South Atlantic Oceans were enriched for nucleotide transport and metabolism (COG group F, r.g=0.564, 0.006), and the South Atlantic, North Indian and South Pacific Oceans for amino acid transport and metabolism (COG group E; r.g=0.048, p=0.035) (Table S1.3).

Of the 202 non-core gene clusters significantly differentiating among ocean regions (indicator analysis, p<0.05), 81 (40%) only had hypothetical functions (Table S1.4). Irrespective of any knowledge about gene function, it is notable that significantly differentiated genes were more often found in genomic islands (34.2%) than all recruited flexible gene clusters (17.2%). It is known that genomic islands play a role in *Prochlorococcus* gene gain and loss (Coleman *et al.*, 2006). Moreover, genes in the island regions of *Prochlorococcus* genomes are among the most dynamic in terms of abundance in the Global Ocean Sampling expedition (GOS) dataset.

Next, we wanted to define possible environmental drivers associated with non-core genes. Some of the most significantly differentiating genes (6 out of 29, indicator

analysis, p=0.001) were related to phosphate acquisition (Scanlan *et al.*, 2009; Martiny *et al.*, 2006). Of all significant known phosphate acquisition-related genes, most were positively associated with the North Atlantic Ocean alone (*ptrA*-a transcriptional regulator related to stress response to phosphorus starvation) or the combined regions of the North Atlantic Ocean and the Equatorial Atlantic Ocean (*chrA*-a response regulator*, phoR*-a phosphate regulon sensor histidine kinase, *phoA*-an alkaline phosphatase, *mfs*-a major facilitator superfamily transporter, *arsA*-an arsenate reductase, and a gene expressed in MED4 during phosphate starvation-PMM0720) (Indval.g, $p<0.05$; Table S1.4). In a biogeochemical context, these regions had lower average annual phosphate concentrations than other regions according to the World Ocean Atlas (WOA) (Garcia *et al.*, 2010).

Nitrogen assimilation genes significant in the indicator analysis were almost always associated with the Equatorial Pacific Ocean (Table S1.4) (Martiny, *et al.*, 2009c). Most were negatively associated specifically with the Equatorial Pacific Ocean and South Atlantic Ocean (*napA*-a nitrate/nitrite transporter and *moa*-a molybdenum cofactor biosynthesis protein) or with these two regions and a third (North Atlantic Ocean: *narB*-an assimilatory nitrate reductase; South Pacific Ocean: *moeA*-molydobpterin biosynthesis protein; North Indian Ocean: an agmatinase) (Table S1.4). However, a few nitrogen-related genes were positively associated with the Equatorial Pacific Ocean: an alkyl hydroperoxide reductase subunit C-like protein, a glutamate N-acetyltransferase, and the Equatorial Pacific and South Atlantic Oceans (a leucine dehydrogenase and an ammonium transporter) (Table S1.4).

A set of iron-requiring proteins were frequently underrepresented in the Equatorial Pacific Ocean – perhaps as an adaptation to lower iron availability (Rusch *et al.*, 2010). In our analysis, we recovered a small subset of these orthologous genes (5 out of 17 non-core genes) (Rusch *et al.*, 2010) as significantly differentiating between regions. Of these, all were negatively associated with only the Equatorial Pacific Ocean (a Fe-S oxidoreductase and a transglutaminase) or the Equatorial Pacific and other regions (a transglutaminase, a cytochrome oxidase C subunit, and a metal-binding protein homologous to CopG-a protein involved in copper homeostasis).

Nitrate and temperature were the largest drivers in *Prochlorococcus* community composition, and while the former has tangible physiological underpinnings, the latter is largely unknown. Using a metagenomic assembly approach and the contribution based upon sites to further understand genes underpinning adaptation to lower temperature, we identified a set of orthologous groups negatively correlated to low temperature (Table S1.5, see Materials and Methods). A subset of the most negatively correlated (COGs 6961, 10190, and 6911) and most unique to the cold-adapted c2 (HLI) clade (COGs 29417, 31703, and 31728) were found on assembled scaffolds composed of more sequences from lower temperature sites in the California Current and the South Atlantic Ocean than other warmer regions (Figure 1.4). Further, these sequences mostly mapped to the lower temperature adapted c2 (HLI) lineage using BLASTn (Camacho *et al.*, 2009). In contrast, the syntenic assemblies lacking the lower temperature genes, generally had sequences recruiting from an even distribution across the regions (Figure 1.4, -COG). Of the four genes most negatively correlated with temperature, three were annotated: a possible trypsin 2OG-Fe(II) oxygenase, a dihydroorotase, and a

hydrolase—a pseudouridine-5' phosphatase which dephosphorylates a potential intermediate in rRNA degradation. One of the COGs unique to the HLI clade, COG 29417 annotated as a possible glycoprotein. In several of these cases, the genomic neighborhood is fairly conserved across assemblies while in other cases there is a range of genomic contexts (Figure S1.5). Although the detailed functions are unclear, the geographical distribution of the individual genes and the structure of the genomic neighborhood suggested that specific genes may contribute to the adaption of *Prochlorococcus* to growth at lower ocean temperatures.

Discussion

Questions regarding the relationship between taxonomy and functional gene diversity are critical to understanding the ecological and evolutionary mechanisms structuring microbial populations (Burke *et al.*, 2011; Raes *et al.*, 2011; Martiny *et al.*, 2013; Brown *et al.*, 2014; Dupont *et al.*, 2014) . Increasing amounts of evidence suggest that using the whole genome rather than simply a single-gene classification is often needed to define microbial lineages and more importantly, determine its role in the environment (Brown *et al.*, 2014). Here we observe a positive correlation between phylogenetic and gene content diversity across *Prochlorococcus* populations. This correlation can be driven by at least three mechanisms that may not be exclusive: genome content is phylogenetically structured; the set of genes an individual has depends on its taxonomy, the same environmental selective forces shape the distribution of phylotypes and specific genes, or horizontal gene transfer events could be more prevalent between closely related lineages than observed. In support of a primarily vertically driven evolutionary history, the gene content of 12 *Prochlorococcus*

strains is largely congruent with a molecular phylogeny based on the core genes, suggesting a strong role of phylogenetic descent (Kettler *et al.*, 2007). In addition, core gene alleles share evolutionary history with distinct sets of flexible genes in specific field populations of *Prochlorococcus* (Kashtan *et al.*, 2014). Populations with these core gene alleles can change in abundance depending on environmental conditions due to differential fitness of the allele or the associated flexible genes. In our analysis, phylogeny and gene content is not perfectly correlated. This may be because of sampling error, but can also be due to fine-scale diversity existing within the defined clades or due to particular sets of genes that vary independent of taxonomy. Divergent genes can be associated with variables that are not measured, such as trace metals, vitamins or ocean physical forces.

It is evident here that *Prochlorococcus* has a clear phylogeography, as demonstrated by changes in the lineage distribution of core genes across oceanic regions; consistent with previous observations (Zwirglmaier *et al.*, 2008; Martiny *et al.*, 2009b; Zinser *et al.*, 2007; Johnson *et al.*, 2006; Bouman *et al.*, 2006). In addition, less than 10% of sequences recruited to deeper HL nodes (g14), indicating that the current genomes capture a majority of the core gene variation in the surface ocean. Moreover, deep-node sequences are distributed similarly to sequences recruited to their corresponding tips of the tree. This indicates a reasonable coverage of diversity within the most abundant *Prochlorococcus* clade.

While nitrate and temperature explained the most phylogenetic variation, we did not expect to see the same environmental parameters explaining the genome content variability, as temperature preference separates ecotypes deeper in their evolutionary

history (Martiny *et al.*, 2009b). Nevertheless, temperature remains a strong factor influencing the distribution of the *Prochlorococcus* flexible gene content. The importance of temperature in this dataset for both the phylogenetic and gene content analyses of *Prochlorococcus* echoes that of the SAR11 group (Brown *et al.*, 2012), where temperature has a significant affect both on the abundance of phylotypes and the distribution of temperature-related genes. Several *Prochlorococcus* genes significantly associated with colder regions in our read-based analysis are also disproportionately found on assemblies made up of sequences predominantly from colder climates. It is puzzling that some of these temperature-related genes are found in several genomes rather than only the cold-adapted HLI strains, although the main signal from the distribution of sequences on assemblies with the gene clearly came from sequences mapping to HLI genomes. This suggests multiple levels of adaptation to differences in temperature including ecotype differentiation as well as genome variability within each ecotype. Thus, strains in culture may not fully represent *Prochlorococcus* diversity as it relates to temperature adaptation – as was recently observed for nitrogen assimilation (Martiny *et al.*, 2009c; Berube *et al.*, 2014).

Nutrient acquisition genes follow past descriptions based on fewer samples, whereby the relative abundance is negatively correlated to nutrient concentrations (Martiny *et al.*, 2011). Our analysis also identified novel patterns at the global level, suggesting undiscovered links between environmental conditions and the genetic diversity of *Prochlorococcus.* For example, many genes differentially distributed along the nitrogen gradient. This includes previously identified genes responsible for nitrate uptake present in lower nitrogen waters (Martiny *et al.*, 2009c) as well as a higher

frequency of genes related to reduced nitrogen uptake in Equatorial Pacific and South Atlantic Oceans. In an environmental context, average annual nitrate concentrations of these two regions were higher than other regions (Garcia *et al.*, 2010) and our data suggests that there appears to be a shift in nitrogen assimilation between oxidized and reduced forms of nitrogen. A large proportion of the genes differing between regions are still annotated as hypothetical. These genes define *Prochlorococcus'* biogeography, but we lack the ability to interpret their physiological and ecological role. Of particular interest is developing our understanding of the South Atlantic Ocean, a region that has been limited in its exploration. Eighty-one non-core genes associate with this region alone, and most of these are conserved hypothetical protein (66.6%) with only seven genes found uniquely in a single *Prochlorococcus* genome.

It is clear that *Prochlorococcus* harbors extensive genome diversity across the global surface ocean, and our results demonstrate that this diversity is not randomly distributed. Instead, *Prochlorococcus* genome diversity displays clear regional biogeographic patterns. Recent niche models have provided a quantitative basis for predicting and understanding shifts in overall *Prochlorococcus* distributions and indicates that future changes in ocean environmental conditions can cause broad changes in phytoplankton community structures (Flombaum *et al.*, 2013). However, the large diversity and associated adaptations to local environmental conditions will likely influence such a response to future ocean conditions. Thus, the interaction between changes in the overall distribution of *Prochlorococcus* vs. specific alleles will likely be important for the functional implications of future ocean changes. Considering the high

abundance of *Prochlorococcus*, such interactions have a large impact on future ocean ecosystems and global biogeochemical cycles.

References

Abascal F, Zardoya R, Telford MJ. (2010). TranslatorX: Multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res* **38**:W7–W13.

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, *et al.* (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**.

Berube PM, Biller SJ, Kent AG, Berta-Thompson JW, Roggensack SE, Roache-Johnson KH, *et al.* (2014). Physiology and evolution of nitrate acquisition in *Prochlorococcus*. *ISME J* **9**:1195–1207.

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

Bouman HA, Ulloa O, Scanlan DJ, Zwirglmaier K, Li WKW, Platt T, *et al.* (2006). Oceanographic Basis of the Global Surface Distribution of *Prochlorococcus* Ecotypes. *Science* **312**:918–921.

Ter Braak CJF. (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* **67**:1167–1179.

Brown M V, Lauro FM, DeMaere MZ, Muir L, Wilkins D, Thomas T, *et al.* (2012). Global biogeography of SAR11 marine bacteria. *Mol Syst Biol* **8**:595.

Brown M V, Ostrowski M, Grzymski JJ, Lauro FM. (2014). A trait based perspective on the biogeography of common and abundant marine bacterioplankton clades. *Mar Genomics* **15**:17–28.

Burke C, Steinberg P, Rusch D, Kjelleberg S, Thomas T. (2011). Bacterial community assembly based on functional genes rather than species. *Proc Natl Acad Sci U S A* **108**:14288–14293.

De Cáceres M, Legendre P. (2009). Associations between species and groups of sites: indices and statistical inference. *Ecology* **90**:3566–3574.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, *et al.* (2009). BLAST+: architecture and applications. *BMC Bioinformatics* **10**.

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

Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, Delong EF, *et al.* (2006). Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* **311**:1768–1770.

Dupont CL, Larsson J, Yooseph S, Ininbergs K, Goll J, Asplund-Samuelsson J, *et al.* (2014). Functional tradeoffs underpin salinity-driven divergence in microbial community composition. *PLoS One* **9**:e89549.

Dupont CL, McCrow JP, Valas R, Moustafa A, Walworth N, Goodenough U, *et al.* (2015). Genomes and gene expression across light and productivity gradients in eastern subtropical Pacific microbial communities. *ISME J* **9**:1076–1092.

Eddy SR. (2011). Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**:e1002195.

Felsenstein J. (2005). PHYLIP (Phylogeny Inference Package) version 3.69. Seattle: Department of Genome Sciences, University of Washington.

Flombaum P, Gallegos JL, Gordillo R a, Rincón J, Zabala LL, Jiao N, *et al.* (2013). Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proc Natl Acad Sci U S A* **110**:9824–9829.

Garcia HE, Locarnini RA, Boyer TP, Antonov JI, Zweng MM, Baranova OK, *et al.* (2010). World Ocean Atlas 2009, Volume 4: Nutrients (phosphate, nitrate, and silicate). *NOAA World Ocean Atlas* **71**.

Hewson I, Paerl RW, Tripp HJ, Zehr JP, Karl DM. (2009). Metagenomic potential of microbial assemblages in the surface waters of the central Pacific Ocean tracks variability in oceanic habitat. *Limnol Oceanogr* **54**:1981–1994.

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

Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, *et al.* (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science* **344**:416–420.

Kelly L, Huang KH, Ding H, Chisholm SW. (2012). ProPortal: a resource for integrated systems biology of *Prochlorococcus* and its phage. *Nucleic Acids Res* **40**:D632–D640.

Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, *et al.* (2007). Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* **3**:2515–2528.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, *et al.* (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**:2947–2948.

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.

Martiny AC, Huang Y, Li W. (2011). Adaptation to Nutrient Availability in Marine Microorganisms by Gene Gain and Loss. In:*Handbook of Molecular Microbial Ecology II: Metagenomics in Different Habitats*, Vol. II, pp. 269–276.

Martiny AC, Huang Y, Li W. (2009a). Occurrence of phosphate acquisition genes in *Prochlorococcus* cells from different ocean regions. *Environ Microbiol* **11**:1340–1347.

Martiny AC, Kathuria S, Berube PM. (2009c). Widespread metabolic potential for nitrite and nitrate assimilation among *Prochlorococcus* ecotypes. *Proc Natl Acad Sci U S A* **106**:10787–10792.

Martiny AC, Tai APK, Veneziano D, Primeau F, Chisholm SW. (2009b). Taxonomic resolution, ecotypes and the biogeography of *Prochlorococcus*. *Environ Microbiol* **11**:823–832.

Martiny AC, Treseder K, Pusch G. (2013). Phylogenetic conservatism of functional traits in microorganisms. *ISME J* **7**:830–838.

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

Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, *et al.* (2008). Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24**:2818–2824.

Moore L, Rocap G, Chisholm S. (1998). Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* **576**:220–223.

Noguchi H, Park J, Takagi T. (2006). MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res* **34**:5623–5630.

Oksanen J, Blanchet F, Kindt R, Legendre P, Minchin P, O'Hara R, *et al.* (2013). vegan: Community Ecology Package. R package version 2.0-10. *R Packag version* **1**. http://cran.r-project.org.

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

Powell S, Forslund K, Szklarczyk D, Trachana K, Roth A, Huerta-Cepas J, *et al.* (2014). EggNOG v4.0: Nested orthology inference across 3686 organisms. *Nucleic Acids Res* **42**:231–239.

R Development Core Team. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. *R Found Stat Comput Vienna, Austria*.

Raes J, Letunic I, Yamada T, Jensen LJ, Bork P. (2011). Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Mol Syst Biol* **7**.

Rusch DB, Martiny AC, Dupont CL, Halpern AL, Venter JC. (2010). Characterization of *Prochlorococcus* clades from iron-depleted oceanic regions. *Proc Natl Acad Sci U S A* **107**:16184–16189.

Scanlan DJ, Ostrowski M, Mazard S, Dufresne A, Garczarek L, Hess WR, *et al.* (2009). Ecological genomics of marine picocyanobacteria. *Microbiol Mol Biol Rev* **73**:249–299.

Stamatakis A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**:1312–1313.

Tatusov RL, Koonin E V., Lipman DJ. (1997). A Genomic Perspective on Protein Families. *Science* **278**:631–637.

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.

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

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

Zwirglmaier K, Jardillier L, Ostrowski M, Mazard S, Garczarek L, Vaulot D, *et al.* (2008). Global phylogeography of marine *Synechococcus* and *Prochlorococcus* reveals a distinct partitioning of lineages among oceanic biomes. *Environ Microbiol* **10**:147–161.

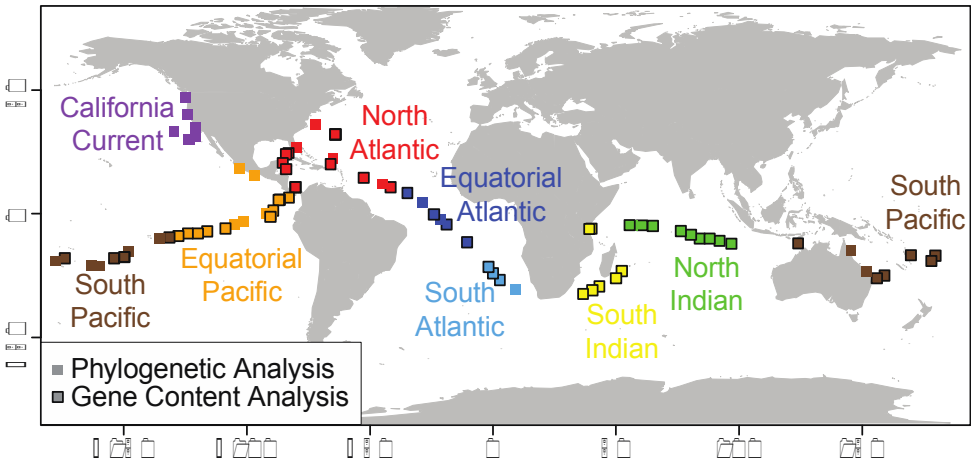


Figure 1.1. Sample map. Samples color coded by region and outlined in black if included in the gene content analysis. See Table S1.1 for metadata for each sample site.

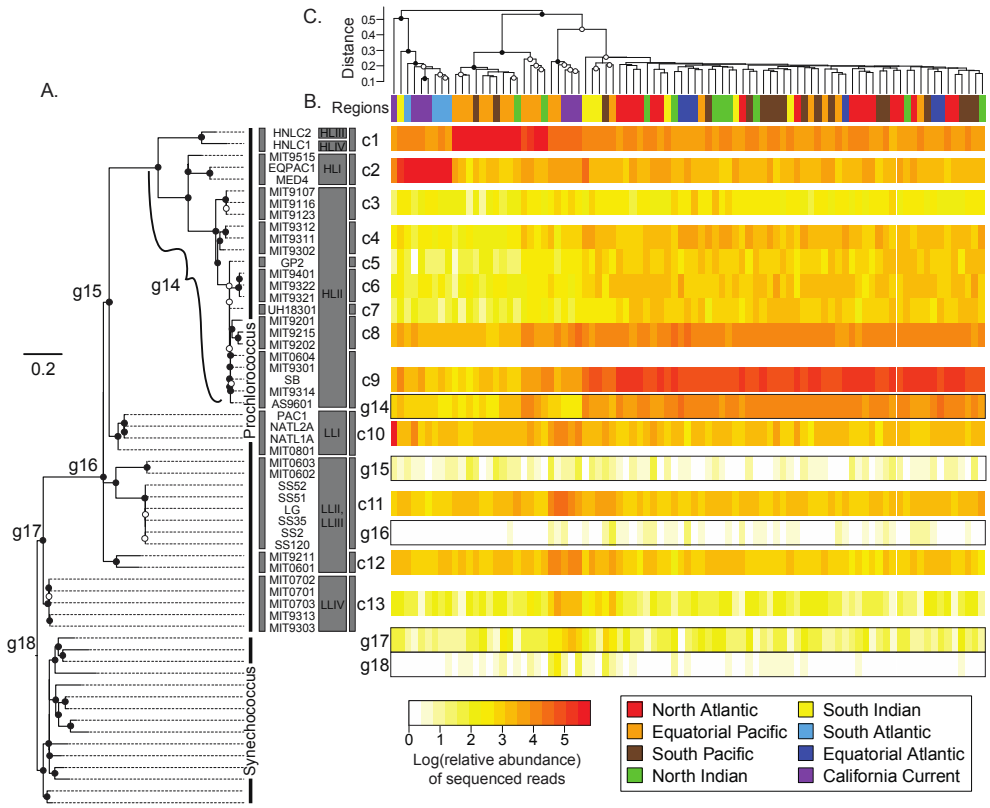


Figure 1.2. Phylogenetic diversity across ocean samples. A) Maximum likelihood phylogeny using core genes from 41 *Prochlorococcus* strains, 2 *Prochlorococcus* metagenomic assemblies and 15 *Synechococcus* strains. Bootstrap values out of 100 resamplings are indicated by circles, filled in have 100% support and empty have at least 50% support. B) Heatmap of samples versus the clades denoted in the phylogeny, c1-c13, and deep phylogenetic group(s), g14-g18. Relative abundances of sequences were log transformed and clustered hierarchically. C) *Prochlorococcus* phylogenetic variation across metagenomic sites clustered using Bray-Curtis dissimilarity of PPlacer placed sequences for samples passing a 500-sequence threshold and rarefied. Node values are bootstrap support for the clade out of 100 resamplings with filled in circles representing >75% support and empty circles representing >25% support for the group. Sample sites are color-coded by region (Figure 1.1).

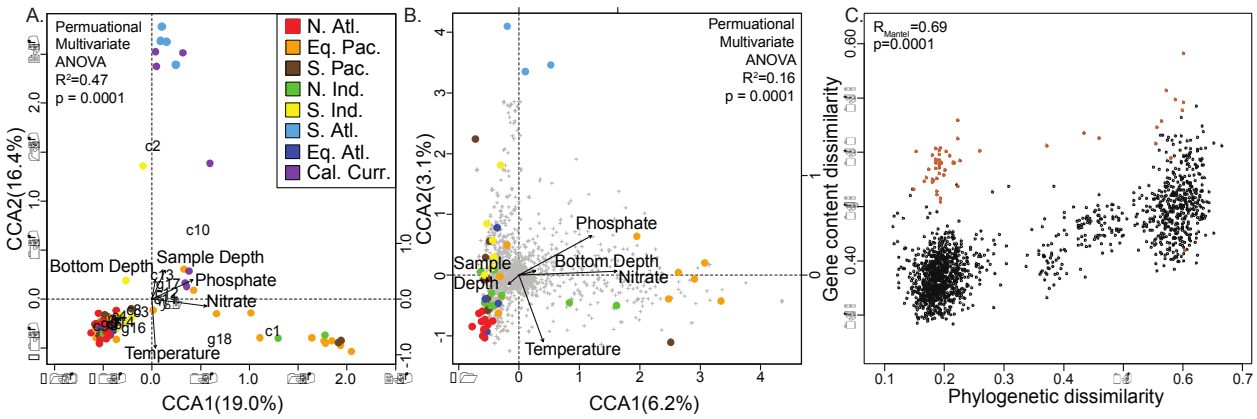


Figure 1.3. Environmental ordination of phylogenetic and genome diversity. Influence of environmental factors on the distribution of *Prochlorococcus* A) phylogenetic and B) gene content diversity. The gene content diversity does not contain any California Current samples due to the increased sample size threshold. The Canonical correspondence analysis triplot showing samples site (colored circles), vectors of environmental parameters: temperature (in situ), nitrate (WOA), phosphate (WOA), depth and bottom depth. The numbers c1-c13 and g14-g18 refers to clades defined in Figure 1.2 (A) and gray pluses refer to orthologous genes. C) Pairwise comparisons between sample sites using phylogenetic and gene content dissimilarity. Correlation estimated based on Mantel test with a Pearson correlation. If GS068 is removed, R_{mantel} increases to 0.82, p=0.001, comparisons involving GS068 are highlighted in gray and contribute to the bimodality of the plot.

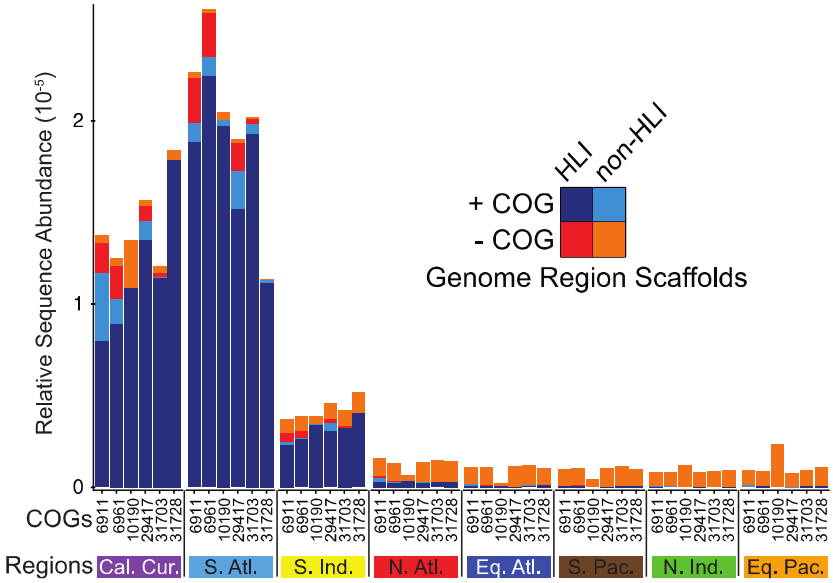


Figure 1.4. Distribution of sequences making up metagenomic assemblies with and without lower temperature genes. The corresponding sequenced reads for each cold temperature gene were mapped to metagenomic assemblies comprised of pooled sequences from the entire dataset (+COG). The two closest neighboring genes were then identified on each metagenomic assembly and used to find assemblies without the lower temperature gene. Sequences on each of these sets of assemblies, with (+COG) and without (-COG) the cold temperature gene, were distributed across regions. Sequences were additionally subdivided based on sequence similarity to high-light I genomes (+HLI) or not (–HLI).

Table 1.1. Partial canonical correspondence analysis of 5 environmental variables.

	Phylogenetic Analysis		Gene Analysis	Content
Variables	Conditional Effects	P-value	Conditional Effects	P-value
Nitrate	16.7%	0.0001	6.2%	0.0001
Temperature	13.3%	0.0001	2.7%	0.0002
Sample Depth	4.4%	0.0038	2.4%	0.0052
Bottom Depth	3.3%	0.0033	1.9%	0.0697
Phosphate	3.3%	0.0035	2.0%	0.1092
Total	41.0%		15.2%	

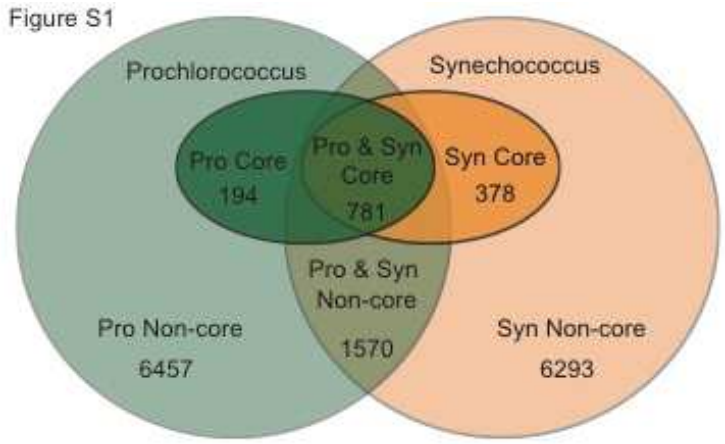


Figure S1.1. Overlap between shared and unshared core and non-core genes. Venn diagram showing the numbers of shared and unshared core and non-core genes using 41 *Prochlorococcus* genomes, 2 metagenomic *Prochlorococcus* assemblies and 15 *Synechococcus* genomes.

Figure S2

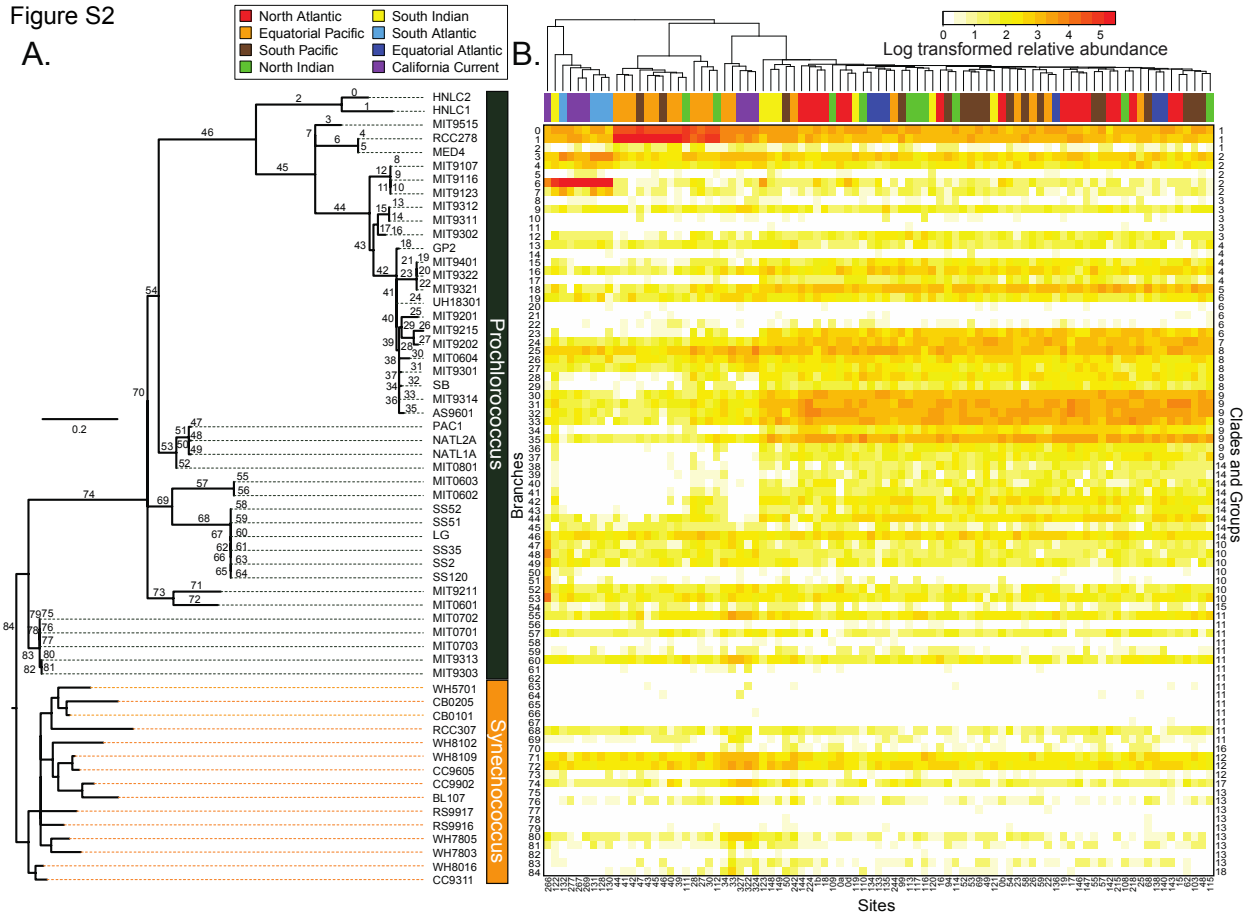


Figure S1.2. Phylogenetic diversity across sampling locations. A. Phylogeny as constructed as in Figure 1.2A with branch labels. B. Heatmap of samples sites versus branch sequence abundance log transformed and clustered hierarchically, clades and groups are labeled on right. Sample sites are colored by region as in Figure 1.1 and clustered hierarchically using the 'average' algorithm on Bray-Curtis dissimilarity between sampling locations.

Figure S3

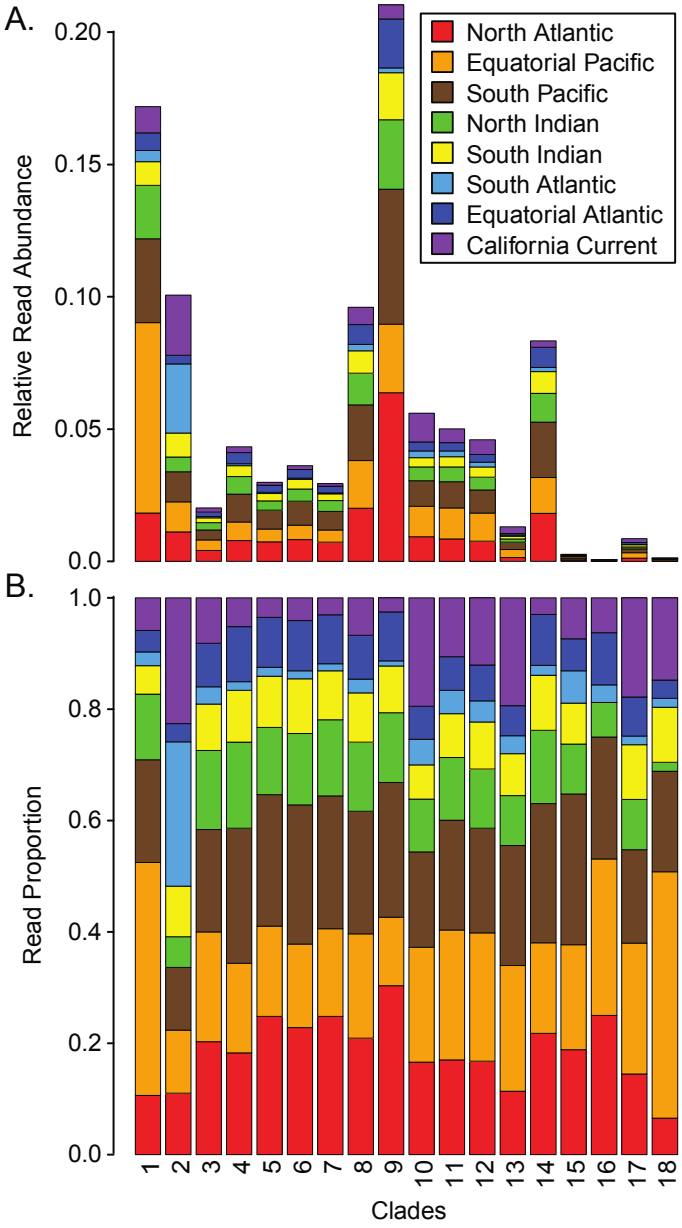


Figure S1.3. Distribution of lineages and deep branches. A) Histogram of the relative sequence abundance from a total of 44,073 sequences falling into the numbered clades and groups from Figure 1.2A and color-coded based on regions as in Figure 1.1. B) Proportional histogram of Figure S1.1A.

Figure S4

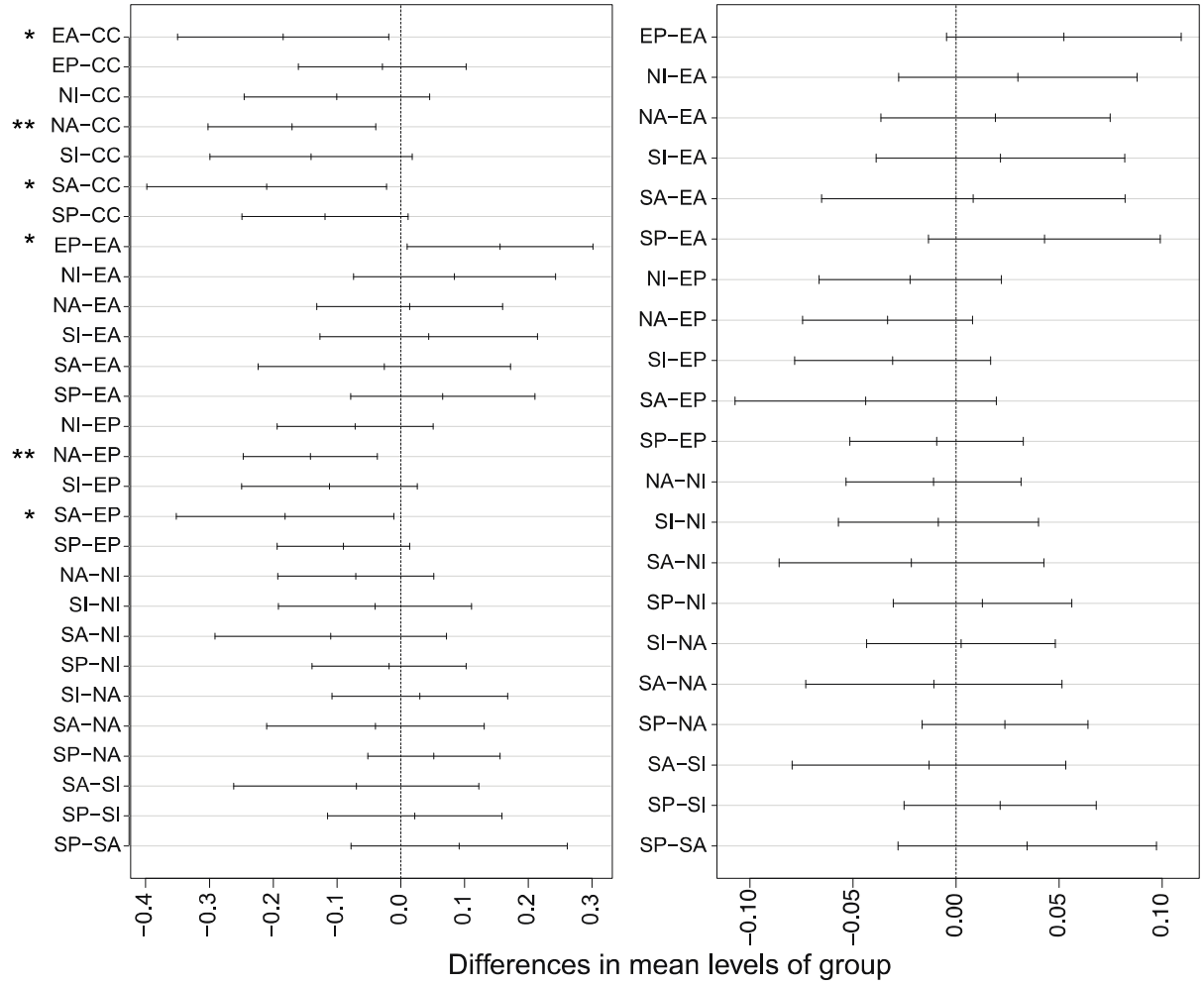


Figure S1.4. Permutational Multivariate ANOVA heteroscedascity. Heteroscedascity in the permutational multivariate analysis of variance in either the A) phylogenetic analysis or the B) gene content analysis. The symbols * and ** represent p-values of <0.05 and <0.01, respectively.

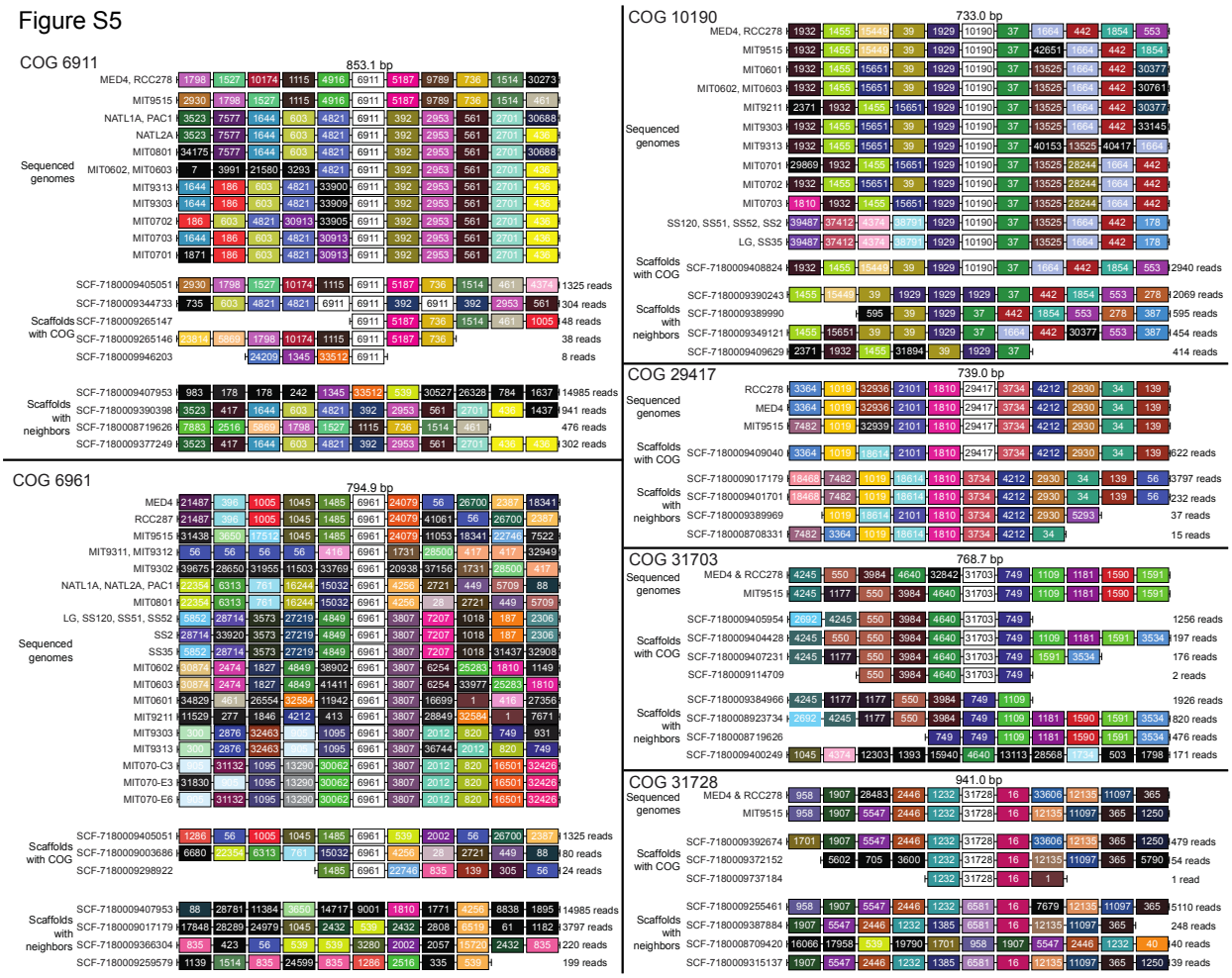


Figure S1.5. Genomic regions of temperature-related genes. Full genomes containing the orthologous groups 6961, 6911, 10190, 29417, 31703, and 31728 are shown with 5 COGs up and downstream of the COG of interest. A subset of scaffolds with or without the lower temperature COG of interest is shown within the genomic region of interest. Genes are color-coded to exemplify shared genes across metagenomic assemblies.

Table S1.1. Sample site metadata. Sample sites included in the gene content analysis are denoted with a *. Regions are abbreviated as follows: North Atlantic (N. Atl.), Equatorial Pacific (Eq. Pac.), South Pacific (S. Pac.), North Indian (N. Ind.), South Indian (S. Ind.), South Atlantic (S. Atl.), Equatorial Atlantic (Eq. Atl.), and California Current (Cal. Cur.) Phosphate and nitrate values are from the World Ocean Atlas 2009 dataset. Temperature measurements were collected at time of sampling. Average sequence lengths are calculated from all picocyanobacterial counts in a sample.

Station	Region	Latitude	Longitude	Phosphate (µmol/L)	Nitrate (µmol/L)	Bottom Depth (m)	Sample Depth (m)	Temp. (°C)	Total Library Counts	Picocyano-bacterial Counts	Average Sequence Length
GS0a*	N. Atl.	32.000	-64.000	0.053	0.369	4 200.0	5.0	20.0	644 551	27 820	836.4
GS0b*	N. Atl.	32.000	-64.000	0.053	0.369	4 200.0	5.0	20.0	317 180	8 958	839.6
GS0d*	N. Atl.	32.000	-64.000	0.053	0.369	4 200.0	5.0	20.0	368 835	20 084	858.6
GS1b	N. Atl.	32.167	-64.500	0.046	0.062	4 200.0	5.0	22.9	90 905	4 471	859.3
GS15*	N. Atl.	24.488	-83.070	0.051	0.941	47.0	1.5	25.3	127 362	20 050	846.2
GS16*	N. Atl.	24.000	-84.000	0.043	0.534	3 300.0	2.0	26.4	127 122	8 897	853.1
GS17*	N. Atl.	20.523	-85.414	0.140	0.110	4 500.0	1.6	27.0	257 581	27 233	842.1
GS18*	N. Atl.	18.000	-84.000	0.125	0.215	4 500.0	2.0	27.4	142 743	17 886	843.6
GS19*	N. Atl.	10.716	-80.254	0.053	0.000	3 300.0	1.6	27.7	135 325	19 103	851.9
GS22*	Eq. Pac.	6.493	-82.904	0.322	0.012	2 400.0	1.9	29.3	121 662	6 980	829.6
GS23*	Eq. Pac.	5.640	-86.565	0.629	0.399	1 100.0	1.9	28.7	133 051	13 971	821.0
GS25*	Eq. Pac.	5.500	-87.067	0.664	0.583	2 400.0	1.0	28.3	120 671	15 891	828.8
GS26*	Ec. Pac.	1.264	-89.295	0.300	0.917	2 400.0	1.8	27.8	102 708	16 584	799.8
GS27*	Ec. Pac.	-1.216	-90.423	0.492	3.068	2.3	2.0	25.5	222 080	17 584	821.3
GS28	Ec. Pac.	-1.219	-90.320	0.492	3.068	160.0	1.8	25.6	189 052	4 628	836.2
GS30	Ec. Pac.	0.200	-92.000	0.541	6.075	19.0	17.0	26.9	359 152	7 807	829.6
GS33	Ec. Pac.	-1.228	-90.429	0.492	3.068	0.3	0.2	37.6	738 651	45 173	816.7
GS34	Ec. Pac.	-0.300	-90.000	0.450	2.066	35.0	2.0	27.5	134 347	24 955	769.1
GS39	Ec. Pac.	-3.343	-101.374	0.659	6.936	2 800.0	1.8	28.7	94 714	5 094	903.9
GS40	Ec. Pac.	-4.499	-105.070	0.720	7.535	3 200.0	2.0	27.8	91 957	7 220	884.9
GS41*	Ec. Pac.	-5.930	-108.687	0.680	6.798	3 000.0	1.8	28.0	93 440	8 997	904.2
GS42*	Ec. Pac.	-7.107	-116.119	0.655	4.974	3 800.0	1.6	27.6	92 684	7 426	905.0
GS43*	Ec. Pac.	-8.000	-120.000	0.684	3.641	4 200.0	2.0	27.6	87 327	6 286	892.0
GS44*	Ec. Pac.	-8.000	-124.000	0.615	3.835	4 000.0	2.0	27.7	86 652	14 004	814.9
GS45*	Ec. Pac.	-9.026	-127.771	0.446	3.281	4 100.0	1.5	28.3	89 014	6 631	820.1
GS46*	S. Pac.	-9.571	-131.492	0.271	2.489	4 100.0	1.7	28.7	76 505	5 773	865.9
GS47	S. Pac.	-10.131	-135.449	0.098	1.003	2 400.0	27.0	28.6	96 876	6 572	884.8
GS48*	S. Pac.	-17.476	-149.812	0.192	0.007	2.1	1.3	28.9	147 203	15 818	840.9
GS49*	S. Pac.	-17.453	-149.799	0.192	0.007	1 000.0	1.2	28.8	91 766	20 708	821.7
GS50	S. Pac.	-15.278	-148.224	0.237	0.042	11.0	1.1	27.8	97 010	17 271	795.1
GS52*	S. Pac.	-18.000	-154.000	0.189	0.044	3 700.0	2.0	28.0	59 264	12 503	861.6
GS53	S. Pac.	-21.183	-159.799	0.206	0.018	650.0	1.6	26.6	10 236	2 619	798.7
GS54	S. Pac.	-20.704	-163.096	0.203	0.073	4 600.0	1.9	26.5	9 712	2 051	793.7
GS55*	S. Pac.	-18.000	-174.000	0.131	0.125	2 000.0	2.0	26.8	58 204	11 488	849.0
GS57	S. Pac.	-19.257	-177.961	0.118	0.077	1 600.0	1.4	26.5	9 769	1 731	796.5
GS58*	S. Pac.	-16.921	179.761	0.173	0.014	760.0	1.6	26.2	180 724	39 840	815.4
GS59*	S. Pac.	-19.042	178.154	0.127	0.060	3.0	1.1	25.7	116 088	7 198	379.6
GS62*	S. Pac.	-16.730	169.787	0.284	0.032	3 100.0	1.5	26.8	60 101	11 197	857.4
GS68*	S. Pac.	-24.972	159.068	0.277	0.385	1 200.0	1.8	22.7	69 908	7 698	373.2
GS69*	S. Pac.	-26.000	156.000	0.233	0.659	4 700.0	2.0	21.9	156 536	26 782	806.3
GS94	S. Pac.	-23.452	151.876	0.068	0.502	18.0	1.5	25.0	62 003	5 922	811.1
GS99	S. Pac.	-14.660	145.453	0.132	0.165	6.1	1.8	26.0	160 226	5 765	804.6
GS103*	S. Pac.	-12.000	124.000	0.150	0.061	67.0	2.0	27.1	58 212	5 900	803.3
GS108*	N. Ind.	-12.093	96.882	0.202	0.010	7.0	1.8	25.8	238 3023	223 632	416.5
GS109*	N. Ind.	-10.944	92.059	0.127	0.034	4 600.0	1.5	27.2	59 812	11 114	886.1
GS110*	N. Ind.	-10.000	88.000	0.115	0.135	1 200.0	2.0	27.0	1 279 359	80 455	457.4
GS111*	N. Ind.	-10.000	84.000	0.091	0.166	3 800.0	2.0	26.4	59 079	5 258	853.9
GS112*	N. Ind.	-8.505	80.376	0.039	0.203	4 600.0	1.8	26.6	1 591 717	132 954	487.7
GS113*	N. Ind.	-7.008	76.332	0.170	0.285	4 600.0	1.8	27.5	109 700	10 100	874.5
GS114*	N. Ind.	-4.990	64.977	0.235	0.099	3 600.0	1.5	28.2	362 739	41 615	796.4
GS115*	N. Ind.	-4.663	60.523	0.267	0.095	3 200.0	1.5	27.9	61 737	7 324	883.5
GS116	N. Ind.	-4.635	56.836	0.128	0.172	2 100.0	1.5	26.2	61 659	5 327	835.3

GS117*	N. Ind.	-4.614	55.509	0.193	0.251	14.0	1.8	26.4	1 589 709	63 702	565.1
GS119*	S. Ind.	-23.216	52.306	0.183	0.144	3 000.0	2.0	23.8	60 987	10 717	881.2
GS120*	S. Ind.	-26.000	50.000	0.210	0.131	5 100.0	3.0	22.5	46 052	8 825	790.6
GS121*	S. Ind.	-29.349	43.216	0.169	0.724	4 300.0	1.5	23.1	110 720	15 121	870.0
GS122*	S. Ind.	-30.898	40.421	0.146	1.022	4 900.0	1.9	20.2	208 301	25 836	748.9
GS123*	S. Ind.	-32.399	36.592	0.193	0.156	1 900.0	2.2	20.4	107 966	7 308	844.3
GS128	S. Atl.	-30.482	9.039	0.514	0.010	5 000.0	1.7	17.9	156 993	3 959	791.1
GS130*	S. Atl.	-26.808	2.832	0.385	0.508	4 700.0	2.1	19.4	101 564	5 981	809.6
GS131*	S. Atl.	-24.000	0.000	0.405	0.682	5 300.0	2.0	20.1	140 910	5 749	357.4
GS132*	S. Atl.	-21.459	-1.778	0.436	0.863	4 000.0	1.8	20.2	154 985	10 831	788.4
GS133*	Eq. Atl.	-11.488	-10.527	0.177	0.302	3 500.0	2.3	23.0	439 611	37 535	364.7
GS134*	Eq. Atl.	-4.290	-18.881	0.124	0.178	4 300.0	2.1	26.4	112 064	10 780	770.8
GS135	Eq. Atl.	-2.379	-21.264	0.164	0.160	4 400.0	1.8	26.5	70 367	3 451	263.8
GS136*	Eq. Atl.	-0.300	-24.000	0.217	0.098	3 400.0	2.0	26.3	61 178	8 418	825.1
GS138	Eq. Atl.	4.711	-28.811	0.052	0.043	2 900.0	1.8	28.2	60 588	3 696	832.0
GS140*	Eq. Atl.	8.390	-34.795	0.080	0.349	4 500.0	1.8	28.2	59 287	9 587	828.0
GS142*	N. Atl.	10.703	-41.572	0.118	0.077	3 000.0	1.8	28.3	10 757	8 206	829.2
GS143	N. Atl.	11.850	-44.995	0.069	0.046	4 400.0	1.8	27.9	58 737	2 344	278.3
GS144*	N. Atl.	14.479	-52.515	0.041	0.041	5 200.0	1.8	27.4	103 164	11 524	840.2
GS146*	N. Atl.	20.000	-66.000	0.080	0.856	5 200.0	2.0	26.7	102 726	5 010	832.5
GS147	N. Atl.	26.698	-79.790	0.021	0.248	3 700.0	1.8	24.1	145 722	5 075	850.6
GS148*	S. Ind.	-6.000	40.000	0.248	0.286	1.0	1.0	25.7	635 662	30 330	506.2
GS149*	S. Ind.	-6.117	39.117	0.247	0.187	5.0	1.5	25.7	710 039	36 044	507.4
GS215	N. Atl.	36.000	-72.000	0.131	0.830	3 800.0	1.0	21.1	46 634	2 928	824.6
GS218	N. Atl.	22.300	-64.800	0.002	0.150	5 800.0	0.9	26.4	28 022	2 408	795.0
GS224*	N. Atl.	10.695	-80.279	0.053	0.000	3 500.0	0.9	26.6	28 010	4 846	789.2
GS242	Ec. Pac.	15.682	-96.789	0.586	0.105	43.0	1.2	26.8	48 149	9 110	753.6
GS244	Ec. Pac.	18.100	-103.099	0.258	0.042	370.0	1.2	25.6	46 374	4 997	800.4
GS266	Cal. Cur.	31.176	-120.913	0.620	3.642	3 900.0	62.0	13.7	45 757	2 346	793.8
GS267	Cal. Cur.	30.168	-122.916	0.350	0.091	3 200.0	5.0	19.6	46 531	2 062	786.5
GS269	Cal. Cur.	30.000	-124.000	0.364	0.021	4 000.0	10.0	19.2	45 458	2 304	751.5
GS277	Cal. Cur.	33.283	-129.417	0.370	0.000	4 500.0	58.0	13.1	47 570	2 266	779.3
GS322	Cal. Cur.	47.111	-124.849	0.520	1.409	140.0	1.1	17.0	669 730	48 459	372.5
GS324	Cal. Cur.	39.896	-124.346	0.953	2.232	1 200.0	1.2	12.1	505 919	19 132	384.1
GS327	Cal. Cur.	34.643	-120.970	0.333	0.184	260.0	0.9	15.0	689 094	15 971	369.2

Table S1.2. Clade correlation analysis. Clades statistically significant, alpha = 0.05, in the analysis of phylogenetic variation. Regions associated with the particular clade or group (Figure 1.2A) are indicated by a 1, the statistic evaluated (r.g) is a point biserial correlation coefficient (Pearson's), which has been corrected for differences in sampling sizes

Clade or group	California Current	Equatorial Atlantic	Equatorial Pacific	North Indian	North Atlantic	South Indian	South Atlantic	South Pacific	Stat: r.g	P-value
c1	0	0	1	0	0	0	0	0	0.526	0.008
c2	1	0	0	0	0	0	1	0	0.804	0.001
c3	1	1	1	1	1	1	0	0	0.300	0.353
c4	0	1	0	1	0	1	0	1	0.642	0.001
c5	0	1	0	1	1	1	0	1	0.646	0.001
c6	0	1	0	1	1	1	0	1	0.740	0.001
c7	0	1	0	1	1	1	0	1	0.700	0.001
c8	1	1	1	1	1	1	0	1	0.553	0.001
c9	0	1	0	1	1	1	0	1	0.772	0.001
c10	1	0	0	0	0	0	0	0	0.497	0.007
c11	1	0	1	0	0	0	0	0	0.327	0.188
c12	1	0	0	0	0	0	0	0	0.445	0.018
c13	1	0	0	0	0	0	0	0	0.481	0.009
g14	0	1	0	1	1	1	0	1	0.760	0.001
g15	0	0	0	0	0	0	1	1	0.212	0.797
g16	0	1	1	0	1	0	0	1	0.240	0.636
g17	1	0	0	0	0	0	0	0	0.448	0.024
g18	1	0	1	0	0	1	0	0	0.306	0.251

Table S1.3. COG categories indicator analysis. COG categories statistically significant, alpha = 0.05, in the indicator analysis of non-core gene content variation across regions. Regions are abbreviated as follows: Equatorial Atlantic-EA, Equatorial Pacific-EP, North Atlantic-NA, North Indian-NI, South Atlantic-SA, South Indian-SI, South Pacific-SP. Regions associated with the particular COG category are indicated by a 1, the statistic evaluated (r.g) is a point biserial correlation coefficient (Pearson's), which has been corrected for differences in sampling sizes.

Definition	COG Letter	EA	EP	NI	NA	SI	SA	SP	Stat: r.g	P-value
Not in cyanobacterial EggNOG database	-	0	1	0	0	0	1	0	0.629	0.003
Energy production and conversion	C	1	0	0	0	0	0	1	0.342	0.375
Cell cycle control, cell division, chromosome partitioning	D	0	1	1	0	0	1	1	0.253	0.715
Amino acid transport and metabolism	E	0	0	1	0	0	1	1	0.488	0.035
Nucleotide transport and metabolism	F	0	0	0	1	0	1	0	0.564	0.005
Carbohydrate transport and metabolism	G	1	1	1	1	1	0	1	0.342	0.379
Coenzyme transport and metabolism	H	1	0	1	0	1	0	0	0.332	0.417
Lipid transport and metabolism	I	0	0	0	0	1	0	0	0.396	0.197
Translation, ribosomal structure and biogenesis	J	1	0	1	0	0	1	1	0.327	0.421
Transcription	K	1	0	1	1	0	1	1	0.223	0.818
Replication, recombination and repair	L	1	1	0	0	1	0	0	0.358	0.312
Cell wall/membrane/envelope biogenesis	M	0	1	0	1	0	0	1	0.325	0.406
Cell motility	N	1	0	0	1	1	0	0	0.370	0.275
Posttranslational modification, protein turnover, chaperones	O	1	0	1	1	0	0	1	0.135	0.982
Inorganic ion transport and metabolism	P	0	0	0	1	0	0	0	0.376	0.252
Secondary metabolites biosynthesis, transport and catabolism	Q	0	1	1	1	0	0	0	0.257	0.71
Function unknown	S	0	1	0	1	1	0	0	0.285	0.622
Signal transduction mechanisms	T	0	1	1	1	0	0	1	0.404	0.161
Intracellular trafficking, secretion, and vesicular transport	U	0	1	1	1	1	1	1	0.270	0.669
Defense mechanisms	V	1	1	1	0	0	0	0	0.315	0.467

Table S1.4. Regionally variable genes. Genes statistically significant, alpha = 0.05, in the gene content variation indicator analysis. Genes are associated with particular regions or sets of regions denoted by having a 1 in the region's column. Region names are abbreviated: Equatorial Atlantic EA, Equatorial Pacific-EP, North Atlantic-NA, North Indian-NI, South Atlantic-SA, South Indian-SI, South Pacific-SP. Sign refers to whether the association between gene and region(s) is positive or negative. The statistic evaluated (indval.g) is an indicator value index, which measures the association between a COG and a region and has been corrected for differences in sampling sizes. CCA1 and CCA2 refer to coordinates for the particular gene on the CCA plot in Figure 1.3b and a 1 under island denotes if members of the COG were present in known genomic islands. The RAST function is given as the set of annotations for the members of a given COG but excludes repetitive elements. Table is organized by COGid.

COGid	EA	EP	NI	NA	SI	SA	SP	Sign	Stat: indval.g	p-value	CCA1	CCA2	Island	RAST function
44	0	1	0	0	0	0	0	-1	0.852	0.023	-0.360	-0.005	0	Nucleoside 2-deoxyribosyltransferase
57	0	1	0	0	0	0	1	-1	0.765	0.042	-0.342	0.204	0	Conserved hypothetical protein FIG00942626: hypothetical protein FIG00941641: hypothetical protein
63	0	0	0	1	1	0	1	-1	0.844	0.001	0.697	0.088	1	Glycosyl transferase| family 2 putative glycosyltransferase family 2
73	0	1	0	0	0	1	0	1	0.643	0.021	1.533	0.247	1	Unknown protein Na+-dependent transporters of the SNF family hypothetical protein
82	0	1	0	0	0	1	0	-1	0.832	0.001	-0.398	-0.141	1	FIG00940844: hypothetical protein
145	1	0	0	0	0	1	0	-1	0.852	0.03	0.139	-0.223	0	Possible SMC domain N terminal domain
155	1	0	0	0	1	1	0	1	0.567	0.033	-0.398	0.687	0	FIG01149297: hypothetical protein FIG00940915: hypothetical protein
255	0	0	0	0	0	1	0	-1	0.910	0.043	0.080	-0.096	0	FIG00940934: hypothetical protein FIG01149311: hypothetical protein
271	0	0	1	0	0	0	1	-1	0.678	0.048	-0.251	-0.045	0	FIG01150634: hypothetical protein FIG00940073: hypothetical protein
305	0	1	0	1	0	0	1	-1	0.764	0.026	-0.231	0.497	1	FIG00944262: hypothetical protein FIG01149883: hypothetical protein FIG00942425: hypothetical protein FIG01152647: hypothetical protein Predicted protein family PM-15 FIG01155851: hypothetical protein FIG01156338: hypothetical protein FIG00940621: hypothetical protein FIG01154924: hypothetical protein FIG01155479: hypothetical protein
404	0	1	0	0	1	1	0	1	0.627	0.02	0.924	0.714	0	Photosystem I subunit IV (PsaE)
474	0	0	0	0	0	1	0	-1	0.890	0.027	-0.080	-0.337	0	Conserved membrane protein TerC family| possibly involved in tellurium resistance Membrane protein TerC| possibly involved in tellurium resistance Conserved membrane protein TerC conserved membrane protein TerC TerC family protein Membrane protein TerC hypothetical protein
475	1	0	0	1	0	1	0	1	0.673	0.004	-0.272	0.119	1	Cytochrome oxidase C subunit VIb-like possible Cytochrome oxidase c subunit VIb
482	0	0	0	0	0	0	1	-1	0.884	0.004	0.094	-0.004	0	Phosphopantetheine adenylyltransferase (EC 2.7.7.3)
484	0	0	0	0	0	1	0	1	0.966	0.001	-0.178	2.672	0	FIG00943123: hypothetical protein Conserved hypothetical protein
493	1	0	0	0	0	0	0	1	0.720	0.003	-0.398	-0.044	0	SSU ribosomal protein S19p (S15e)
503	1	0	1	0	0	1	0	1	0.702	0.006	-0.097	0.611	0	FIG00944489: hypothetical protein FIG00944192: hypothetical protein
623	1	1	0	0	1	0	0	1	0.577	0.038	0.557	0.656	0	FIG00940800: hypothetical protein FIG01150332: hypothetical protein
632	0	1	0	0	0	0	0	-1	0.879	0.034	-0.199	-0.134	0	Hypothetical protein YaeJ with similarity to translation release factor hypothetical protein
659	0	1	0	0	0	0	0	-1	0.893	0.008	-0.312	-0.033	0	Phosphoribosyltransferase Uracil phosphoribosyltransferase (EC 2.4.2.9) / Pyrimidine operon regulatory protein PyrR Phosphoribosyl transferase (EC:2.4.2.9) Xanthine-guanine phosphoribosyltransferase (EC 2.4.2.22) hypothetical protein
747	0	0	0	0	0	1	0	1	0.665	0.013	0.170	0.312	0	Possible Photosystem II reaction center Z protein (PsbZ) Photosystem II protein PsbZ
794	0	1	1	0	0	0	1	-1	0.629	0.033	-0.321	0.445	1	Hypothetical protein FIG00941298: hypothetical protein
817	1	0	0	1	0	0	1	1	0.850	0.002	-0.116	-0.437	0	Permease of the major facilitator superfamily
875	1	0	1	1	0	0	0	1	0.707	0.01	-0.374	-0.387	1	Dienelactone hydrolase Candidate 1: dienelactone hydrolase Dienelactone hydrolase (EC:3.1.1.45) Dienelactone hydrolase and related enzyme possible carboxymethylenebutenolidase (EC:3.1.1.45)

														hypothetical protein
880	1	1	0	0	0	1	0	1	0.576	0.045	1.791	0.428	1	FIG00941388: hypothetical protein
1015	0	0	0	1	0	0	0	-1	0.817	0.02	0.242	-0.072	0	Possible Serine hydroxymethyltransferase
1018	0	0	0	1	0	0	0	-1	0.905	0.021	0.486	0.116	1	Mn2+ and Fe2+ transporter| NRAMP family conserved hypothetical putative integral membrane protein FIG00944807: hypothetical protein
1181	0	1	0	0	0	1	0	-1	0.917	0.001	-0.360	-0.253	1	DUF209 Pirin Protein yhhW
1209	0	1	0	0	0	0	0	-1	0.830	0.016	-0.326	-0.098	0	Uncharacterized HesB family conserved protein Possible AIR synthase related protein| C-termi possible AIR synthase related protein| C-termi FIG00941879: hypothetical protein
1479	0	0	0	0	0	1	0	-1	0.930	0.017	-0.042	-0.039	0	Adenosylcobinamide-phosphate guanylyltransferase (EC 2.7.7.62)
1514	0	0	0	1	0	1	1	-1	0.834	0.002	0.369	0.000	1	FIG00941114: hypothetical protein FIG00942899: hypothetical protein FIG00942684: hypothetical protein FIG00942032: hypothetical protein
1544	0	0	0	0	0	0	1	-1	0.935	0.013	-0.032	0.043	0	Diflavin flavoprotein SYNW2367 DIM6/NTAB family protein
1547	0	1	0	0	0	0	0	-1	0.951	0.005	-0.334	0.087	0	SAM radical protein| elongator protein
1598	1	0	1	0	0	0	0	-1	0.769	0.029	0.034	0.218	1	FIG00944609: hypothetical protein putative protein family PM-19 Predicted protein family PM-19 FIG00943688: hypothetical protein
1629	1	1	0	0	1	0	0	1	0.614	0.034	0.445	0.063	0	Periplasmic divalent cation tolerance protein cutA hypothetical protein cutA1 divalent ion tolerance protein CutA1 divalent ion tolerance protein Periplasmic divalent cation tolerance protein CutA
1646	1	0	0	0	0	1	0	-1	0.830	0.046	0.040	0.030	0	Tryptophan-rich protein DUF2389| Ssr2843 homolog hypothetical protein
1664	1	0	0	0	1	1	0	1	0.732	0.01	-0.270	0.347	0	Arsenate reductase related protein putative arsenate reductase Arsenate reductase (EC 1.20.4.1) Arsenate reductase and related proteins| glutaredoxin family Arsenate reductase related protein (arsC family) putative arsenate reductase (EC:1.20.4.1)
1731	0	0	0	0	0	1	0	1	0.741	0.005	-0.236	2.077	1	Protein family PM-12 FIG00940535: hypothetical protein
1767	0	0	0	0	0	0	1	-1	0.955	0.008	0.086	0.018	0	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase (EC 2.7.1.148)
1772	0	1	0	0	0	1	0	-1	0.843	0.007	-0.242	-0.232	1	Possible DNA gyrase/topoisomerase IV| subunit DNA gyrase/topoisomerase IV| subunit-like DNA gyrase/topoisomerase IV| subunit-like protein FIG00940612: hypothetical protein
1897	0	1	0	0	0	1	0	1	0.660	0.021	0.869	0.970	1	Conserved hypothetical protein FIG00940214: hypothetical protein
1901	0	0	0	0	0	1	0	-1	0.890	0.013	-0.204	-0.142	0	Ribonuclease P protein component (EC 3.1.26.5)
1946	0	0	0	0	0	1	0	-1	0.942	0.009	-0.042	-0.062	0	FIG00940579: hypothetical protein FIG01152141: hypothetical protein
1962	0	0	0	1	1	0	1	-1	0.822	0.001	0.807	0.155	1	N-methyl-L-tryptophan oxidase (EC 1.5.3.-)
1968	0	0	1	0	0	1	0	1	0.681	0.012	-0.229	0.629	0	Protein YidD
2019	0	0	0	0	0	1	0	1	0.913	0.001	-0.223	1.719	0	FIG023675: hypothetical protein
2034	0	1	0	0	0	0	0	-1	0.863	0.006	-0.372	-0.005	0	FIG00942135: hypothetical protein FIG01149229: hypothetical protein
2102	0	0	0	1	0	0	0	-1	0.802	0.008	0.078	0.139	1	FIG00941297: hypothetical protein FIG00941496: hypothetical protein FIG00944881: hypothetical protein FIG00944668: hypothetical protein
2125	0	0	0	0	0	0	1	1	0.522	0.04	-0.563	-0.072	0	Possible photosystem I reaction centre subunit XII (PsaM) photosystem I subunit XII (PsaM)
2292	1	0	0	0	0	0	0	1	0.577	0.037	-0.315	-0.839	1	HupE-UreJ family cobalt transporter
2468	0	1	1	0	1	0	0	1	0.640	0.022	1.805	-0.222	0	DNA ligase (EC 6.5.1.2) putative NAD-dependent DNA ligase (contains BRCT domain type II) hypothetical protein
2516	0	0	1	0	1	1	0	1	0.634	0.024	-0.394	0.425	0	FIG00940553: hypothetical protein FIG00940902: hypothetical protein
2563	0	1	0	0	0	0	0	-1	0.877	0.032	-0.329	-0.048	0	Glutaredoxin Predicted DNA-binding protein with PD1-like DNA-binding motif FIG00942813: hypothetical protein Protein containing glutaredoxin domain and PD1-like DNA-binding domain hypothetical protein
2761	0	0	0	1	0	1	0	-1	0.716	0.011	0.454	-0.141	1	Hypothetical protein FIG00943160: hypothetical protein
2911	0	0	0	0	0	1	0	-1	0.911	0.009	-0.042	-0.077	0	COG0223: Methionyl-tRNA formyltransferase Phosphoribosylglycinamide formyltransferase (EC 2.1.2.2) hypothetical protein
2921	0	0	0	0	0	1	0	1	0.681	0.013	-0.381	0.882	0	FIG00941296: hypothetical protein
3146	0	0	0	0	0	1	0	1	0.935	0.001	-0.069	2.537	0	Conserved hypothetical protein FIG00943604: hypothetical protein Hypothetical protein
3226	0	1	0	0	1	1	0	1	0.676	0.023	1.021	0.558	1	FIG00940823: hypothetical protein FIG00942296: hypothetical protein FIG00940230: hypothetical protein
3405	0	1	0	1	0	1	0	-1	0.786	0.002	-0.370	-0.093	0	Assimilatory nitrate reductase large subunit (EC:1.7.99.4)
3496	0	1	0	1	0	1	0	-1	0.650	0.037	-0.329	0.280	0	FIG00940737: hypothetical protein FIG00940496: hypothetical protein

3554	0	1	0	0	0	0	0	-1	0.940	0.004	-0.358	0.085	0	Ribosomal protein S12p Asp88 (*E. coli*) methylthiotransferase
3599	0	0	0	0	1	1	0	1	0.622	0.03	-0.188	0.487	0	Hypothetical protein FIG00940400: hypothetical protein FIG01150299: hypothetical protein
3799	0	0	0	0	0	1	0	-1	0.942	0.009	-0.227	-0.186	0	Cytochrome b/b6/PetD-like possible lipoprotein hypothetical protein possible Borrelia lipoprotein possible Cytochrome b(C-terminal)/b6/petD
3809	0	1	0	0	0	0	0	-1	0.970	0.001	-0.383	-0.102	1	Protein containing domains DUF404| DUF407 hypothetical protein
4058	1	0	0	0	1	1	0	-1	0.771	0.027	-0.186	-0.383	0	Possible Nucleoside diphosphate kinase
4188	0	1	0	0	0	1	0	1	0.821	0.001	0.928	0.322	1	Gll1939 protein
4330	0	0	0	0	0	0	1	1	0.765	0.005	-0.424	-0.169	0	FIG01150344: hypothetical protein
4818	1	0	1	0	0	1	0	1	0.674	0.011	-0.304	-0.049	0	FIG00940973: hypothetical protein
4849	0	1	1	0	0	0	0	-1	0.811	0.03	-0.187	0.024	0	Possible Influenza RNA-dependent RNA polymerase FIG00940740: hypothetical protein influenza RNA-dependent RNA polymerase-like hypothetical protein
4916	0	0	0	0	0	1	0	1	0.909	0.001	0.353	1.583	0	FIG00941015: hypothetical protein
4958	0	1	0	0	0	0	0	-1	0.920	0.021	-0.324	0.012	1	Transglutaminase-like superfamily domain protein Transglutaminase-like domain FIG001454: Transglutaminase-like enzymes| putative cysteine proteases Protein containing transglutaminase-like domain| putative cysteine protease possible transglutaminase-like enzyme Transglutaminase-like enzymes| putative cysteine proteases NAD-dependent malic enzyme (EC 1.1.1.38)
5062	0	0	0	0	0	1	0	-1	0.918	0.039	0.123	-0.113	0	Possible Paired amphipathic helix repeat
5121	0	1	0	0	0	1	0	-1	0.829	0.002	-0.385	-0.165	0	Nitrate/nitrite transporter
5495	0	0	0	0	0	0	1	1	0.765	0.001	-0.448	-0.107	0	FIG01149450: hypothetical protein
5788	0	0	0	1	0	0	0	-1	0.885	0.005	0.034	-0.063	1	Cytochrome b559 beta chain (PsbF)
5869	0	1	0	0	0	0	0	-1	0.907	0.009	-0.335	-0.002	1	Possible fatty acid desaturase Fatty acid desaturase fatty acid desaturase Fatty acid desaturase| type 2 Possible fatty acid desaturase
6115	0	1	0	0	0	1	0	-1	0.772	0.048	-0.374	-0.215	0	Putative anti-sigma factor antagonist FIG00941060: hypothetical protein anti-sigma F factor antagonist anti-sigma F factor antagonist (spoIIAA-2); anti sigma b factor antagonist RsbV
6195	0	1	0	0	0	1	1	-1	0.704	0.009	-0.371	-0.136	0	Molybdopterin biosynthesis protein MoeA
6203	0	1	1	0	0	1	0	1	0.850	0.001	0.970	0.399	1	Leucine dehydrogenase (EC 1.4.1.9)
6402	0	1	1	0	1	0	0	1	0.711	0.008	0.739	-0.120	1	Predicted ATPase related to phosphate starvation-inducible protein PhoH
6680	0	1	0	0	0	0	0	-1	0.938	0.019	-0.337	0.015	0	Zinc ABC transporter| periplasmic-binding protein ZnuA
6697	1	0	0	0	1	0	1	1	0.629	0.035	-0.386	0.207	1	Asparagine synthetase [glutamine-hydrolyzing] (EC 6.3.5.4)
6872	1	0	0	0	1	0	0	1	0.603	0.025	-0.461	0.296	0	FIG00941014: hypothetical protein
6911	0	0	0	1	0	1	0	1	0.671	0.011	-0.379	0.848	1	Pseudouridine-5' phosphatase (EC 3.1.3.-) HAD-superfamily hydrolase| subfamily IA| variant 3 HAD-superfamily hydrolase subfamily IA| variant 3
6922	1	0	0	0	0	1	0	1	0.699	0.006	0.417	0.766	1	Hypothetical protein
6932	0	0	0	0	0	1	0	1	0.590	0.041	0.262	0.530	0	FIG00940310: hypothetical protein
6961	0	0	0	0	0	1	0	1	0.655	0.014	-0.175	1.715	1	Possible Trypsin 2OG-Fe(II) oxygenase
6968	0	0	0	0	0	1	0	1	0.577	0.047	-0.063	2.707	0	FIG00941313: hypothetical protein
7370	0	0	0	0	0	1	0	-1	0.920	0.047	-0.098	-0.220	1	Metal-dependent membrane protease| abortive infection protein Abortive infection protein
7842	0	0	1	0	1	1	0	1	0.674	0.007	-0.332	0.640	0	Hypothetical protein FIG00940304: hypothetical protein FIG01149538: hypothetical protein
7981	1	0	0	0	0	0	0	1	0.663	0.007	0.210	-0.144	0	Photosystem II protein PsbM
8269	0	1	0	0	0	1	0	-1	0.722	0.042	-0.373	-0.100	0	Tungsten-containing aldehyde ferredoxin oxidoreductase cofactor modifying protein putative molybdenum cofactor biosynthesis protein hypothetical protein Molybdenum cofactor biosynthesis protein MoaA
8356	0	1	1	0	0	1	0	-1	0.743	0.018	-0.405	-0.109	0	Possible Virion host shutoff protein
8463	0	0	0	0	0	1	0	-1	0.858	0.037	0.026	-0.305	0	ATP/GTP-binding site motif A ATP/GTP-binding site motif A (P-loop) (EC:2.7.1.48)
8610	0	0	1	0	0	1	0	1	0.588	0.034	-0.400	0.333	1	FIG00941656: hypothetical protein
8885	0	1	0	0	0	1	0	1	0.689	0.006	1.313	0.324	1	Permeases of the major facilitator superfamily hypothetical protein
9089	0	1	1	0	0	1	0	1	0.708	0.008	1.143	0.262	1	Glucose-methanol-choline (GMC) oxidoreductase:NAD binding site
9323	0	0	0	0	0	1	0	1	0.829	0.001	0.568	0.765	1	Possible Vanadium/alternative nitrogenase delt
10190	0	0	0	0	0	1	0	1	0.853	0.002	-0.088	2.241	0	Dihydroorotase (EC 3.5.2.3)

10290	1	0	0	0	0	0	1	-1	0.742	0.047	-0.241	0.116	0	Putative protein-S-isoprenylcysteine methyltransferase S-isoprenylcysteine O-methyltransferase related enzyme FIG01150808: hypothetical protein FIG01157213: hypothetical protein
10336	0	0	0	0	0	1	0	-1	0.934	0.012	0.108	-0.037	0	Fructose-bisphosphate aldolase class I (EC 4.1.2.13)
10590	1	0	0	0	0	1	0	1	0.672	0.012	-0.022	0.729	0	FIG00943507: hypothetical protein FIG00940267: hypothetical protein
11053	0	1	0	1	0	0	1	-1	0.806	0.004	-0.273	0.379	1	FIG00943760: hypothetical protein FIG00941963: hypothetical protein
11384	0	0	0	1	0	0	0	1	0.645	0.018	-0.366	-0.760	1	Possible Bacterial regulatory proteins| crp fa hypothetical protein
11607	0	0	0	0	0	1	0	1	0.795	0.001	-0.299	1.053	0	Possible Helix-turn-helix protein| copG family Possible Helix-turn-helix protein| copG family
12303	0	0	0	0	0	1	0	-1	0.858	0.043	-0.288	-0.312	1	FIG00940089: hypothetical protein FIG00942931: hypothetical protein
12427	0	0	0	0	1	1	0	1	0.622	0.024	-0.271	0.857	1	FIG00942946: hypothetical protein FIG00941489: hypothetical protein
12919	0	1	1	0	0	1	0	1	0.795	0.001	0.992	0.364	1	3-oxoacyl-[acyl-carrier protein] reductase (EC 1.1.1.100)
13113	0	1	0	0	0	1	0	-1	0.742	0.046	-0.378	-0.222	1	CopG protein
13381	1	0	0	1	0	0	0	1	0.847	0.001	-0.391	-0.500	0	Chromate transporter Chromate transport protein ChrA putative chromate transport protein| CHR family putative chromate transporter| CHR family
13582	1	0	0	1	0	0	0	1	0.919	0.001	-0.367	-0.628	0	Two-component sensor histidine kinase Phosphate regulon sensor protein PhoR (SphS) (EC 2.7.13.3) ATP-binding region| ATPase-like:Histidine kinase| HAMP region:Histidine kinase A| N-terminal
14543	0	0	0	1	0	0	0	-1	0.879	0.002	0.159	-0.065	1	Hypothetical protein FIG00944081: hypothetical protein FIG00942004: hypothetical protein
14614	0	1	0	0	0	1	0	1	0.594	0.04	1.763	1.108	1	FIG01153355: hypothetical protein Uncharacterized conserved secreted protein| specific to cyanobacteria FIG01156744: hypothetical protein possible Small| acid-soluble spore proteins| a FIG01154991: hypothetical protein FIG01155893: hypothetical protein
15427	1	0	0	1	0	0	0	1	0.911	0.001	-0.324	-0.607	0	Alkaline phosphatase (EC 3.1.3.1)
15544	0	0	0	1	0	0	0	1	0.759	0.004	-0.312	-0.335	0	HNH endonuclease:HNH nuclease
16400	0	0	0	1	0	0	0	-1	0.911	0.001	0.059	-0.024	1	Hypothetical protein FIG00941868: hypothetical protein FIG00941742: hypothetical protein
16566	0	0	0	1	0	0	0	1	0.645	0.019	-0.505	-0.493	1	Alkaline phosphatase (EC 3.1.3.1)
16746	0	0	0	0	0	0	1	1	0.655	0.007	-0.439	-0.001	0	FIG00942927: hypothetical protein FIG01155269: hypothetical protein FIG01157942: hypothetical protein
17018	0	1	1	0	0	1	0	-1	0.684	0.028	-0.359	-0.052	0	Hypothetical protein
17755	0	1	0	0	0	0	0	1	0.597	0.035	2.639	-0.108	0	Hypothetical protein
17958	0	1	1	0	1	0	0	1	0.588	0.035	0.544	-0.163	0	FIG00941230: hypothetical protein
18286	1	0	0	1	0	1	0	1	0.583	0.038	-0.393	-0.296	0	FIG01153741: hypothetical protein FIG01153929: hypothetical protein possible M protein repeat possible phage integrase family
20464	0	0	0	1	0	0	1	-1	0.711	0.013	-0.082	0.259	1	FIG00944544: hypothetical protein
21140	0	1	1	0	0	1	0	1	0.738	0.004	0.673	-0.010	1	FIG00942087: hypothetical protein
22394	1	0	0	1	0	0	0	1	0.686	0.007	-0.393	-0.549	1	Arsenate reductase (EC 1.20.4.1) Arsenical-resistance protein ACR3
22608	0	0	0	0	0	1	0	1	0.663	0.026	-0.362	1.192	1	FIG00942811: hypothetical protein FIG00940402: hypothetical protein
22644	0	1	0	1	0	1	0	-1	0.722	0.009	-0.368	-0.130	1	Possible Ribosomal RNA adenine dimethylase
22678	0	1	1	0	0	0	0	1	0.632	0.024	1.545	-0.240	1	Glycine betaine transporter High-affinity choline uptake protein BetT putative glycine betaine transporter| BCCT family
23391	0	0	0	0	0	1	0	1	0.740	0.004	0.709	0.556	0	Protoporphyrinogen IX oxidase|oxygen-independent|HemG (EC 1.3.-.-)
24209	0	0	0	1	0	0	0	1	0.577	0.029	-0.318	-0.676	0	Hypothetical protein
24599	0	1	0	0	0	1	0	1	0.585	0.039	1.346	0.505	0	Hypothetical protein
24841	0	0	0	0	0	1	0	1	0.577	0.047	-0.063	2.707	1	Possible Virion host shutoff protein
24917	0	0	0	0	0	1	0	1	0.816	0.003	-0.028	2.781	0	ATP-dependent protease La (EC 3.4.21.53) Type I FIG00941506: hypothetical protein
24957	0	1	1	0	0	0	1	-1	0.786	0.002	-0.350	-0.225	1	FIG01153374: hypothetical protein FIG00942680: hypothetical protein
25014	0	0	0	0	0	1	0	1	0.516	0.048	-0.410	1.124	0	Predicted protein family PM-12 predicted protein family PM-12 hypothetical protein
25664	0	0	0	0	0	1	0	1	0.862	0.001	-0.169	1.628	1	GP0.7
26024	0	1	0	0	0	0	0	1	0.723	0.005	2.289	-0.065	0	2-octaprenyl-6-methoxyphenol hydroxylase (EC 1.14.13.-)
26278	0	1	0	0	0	1	0	1	0.761	0.001	1.723	0.429	1	Hypothetical protein
26319	0	0	1	0	0	0	0	1	0.577	0.026	-0.393	-0.762	0	FIG00941336: hypothetical protein influenza non-structural protein (NS2)-like
26603	0	0	0	0	0	1	0	1	0.643	0.027	0.079	0.498	0	FIG01150547: hypothetical protein Possible glycosyltransferase conserved hypothetical protein TPR repeat possible glycosyltransferase TPR domain protein TPR repeat-containing protein Translation

														elongation factor P COG0457: FOG: TPR repeat FOG: TPR repeat hypothetical protein
26846	0	1	0	0	0	0	0	1	0.663	0.011	2.077	-0.112	0	Possible ATP synthase protein 8 precursor
26956	1	0	1	0	1	0	0	1	0.639	0.048	-0.275	0.094	0	FIG01150673: hypothetical protein
27101	0	1	1	0	0	1	0	-1	0.778	0.014	-0.344	-0.310	0	Agmatinase (EC 3.5.3.11)
27214	1	0	0	0	1	0	0	1	0.614	0.028	-0.102	0.613	1	Endonuclease
27383	0	0	0	0	0	1	0	1	0.683	0.012	-0.180	2.014	1	GtrA family protein hypothetical protein
27415	0	0	0	1	1	0	0	1	0.595	0.048	-0.425	-0.186	0	N-carbamoyl-L-amino acid amidohydrolase N-carbamoyl-L-amino acid hydrolase (EC 3.5.1.87)
28241	0	0	0	0	0	1	0	-1	0.932	0.006	0.014	-0.174	0	FIG00941603: hypothetical protein
28527	0	0	0	0	0	1	0	1	0.687	0.008	-0.173	1.120	1	COG3558: hypothetical protein
28548	0	1	0	0	0	0	0	1	0.598	0.028	1.791	-0.423	0	DNA polymerase I (EC 2.7.7.7)
28615	1	0	0	1	0	0	0	1	0.866	0.001	-0.361	-0.698	1	Possible Poly A polymerase regulatory subunit possible poly A polymerase regulatory subunit hypothetical protein
28631	0	1	1	0	0	0	1	-1	0.674	0.025	-0.303	-0.123	1	Possible Myosin N-terminal SH3-like domain
28705	0	1	1	0	0	0	0	1	0.616	0.03	1.648	-0.290	0	Zinc metalloprotease hypothetical protein Putative predicted metal-dependent hydrolase Protein of unknown function DUF45
28781	0	0	0	1	0	0	0	1	0.959	0.001	-0.294	-0.708	0	Phosphonate ABC transporter phosphate-binding periplasmic component (TC 3.A.1.9.1)
29378	0	0	0	0	0	1	0	1	0.816	0.003	-0.158	2.777	0	Hypothetical protein
29417	0	0	0	0	0	1	0	1	0.754	0.005	-0.144	2.510	1	Possible Hepatitis C virus envelope glycoprote possible envelope glycoprotein-lik protein
29432	0	1	0	0	0	1	0	1	0.586	0.031	1.586	0.462	0	Hypothetical protein
29501	0	0	0	1	0	0	1	1	0.642	0.028	-0.392	-0.443	0	Biotin synthase-related enzyme
29547	0	1	0	0	0	0	0	1	0.650	0.015	2.299	-0.112	0	Hypothetical protein
29774	0	0	0	0	0	1	0	1	0.944	0.001	-0.223	1.808	1	FIG00942706: hypothetical protein
29836	0	0	0	0	0	1	0	1	0.577	0.047	-0.063	2.707	0	Hypothetical protein
29877	0	1	0	0	0	1	0	1	0.613	0.03	1.263	0.321	0	Transport system permease protein
29990	0	1	0	0	0	1	0	1	0.668	0.005	1.637	0.238	1	Ammonium transporter
30189	0	1	0	0	0	0	0	1	0.599	0.034	2.018	-0.267	0	FIG00940454: hypothetical protein
30324	0	1	0	0	0	1	0	1	0.555	0.034	1.650	1.378	1	FIG00941624: hypothetical protein
30397	0	0	0	1	0	0	0	1	0.943	0.001	-0.355	-0.599	0	Phosphonate ABC transporter permease protein phnE (TC 3.A.1.9.1) Phosphonate ABC transporter permease protein phnE2 (TC 3.A.1.9.1) Phosphonate ABC transporter permease protein phnE1 (TC 3.A.1.9.1)
30630	0	0	0	0	0	1	0	1	0.516	0.048	-0.410	1.124	0	Hypothetical protein
30634	0	0	0	0	0	1	0	1	0.577	0.047	-0.063	2.707	0	Phosphate transport system permease protein PstC (TC 3.A.1.7.1)
30772	0	0	0	0	0	1	0	1	0.838	0.003	0.542	1.512	0	Cell division protein FtsK hypothetical protein
31080	0	0	0	1	0	0	0	1	0.663	0.015	-0.434	-0.482	0	Phosphonate dehydrogenase (EC 1.20.1.1) (NAD-dependent phosphite dehydrogenase) phosphoglycerate dehydrogenase
31613	0	1	0	0	0	0	0	1	0.576	0.041	1.374	-0.261	0	DNA-directed RNA polymerase specialized sigma subunit| sigma24-like protein RNA polymerase sigma-70 factor
31678	0	1	0	0	0	1	0	1	0.589	0.035	2.363	0.039	0	Possible Adenylate cyclase
31703	0	0	0	0	0	1	0	1	0.816	0.003	-0.158	2.777	1	FIG01153305: hypothetical protein FIG00942523: hypothetical protein
31821	0	0	0	0	0	1	0	1	0.764	0.004	-0.086	2.500	1	Possible Trehalase
31946	1	0	0	1	0	0	0	1	0.707	0.002	-0.418	-0.732	0	FIG00942256: hypothetical protein
32080	0	0	0	0	0	1	0	1	0.789	0.001	-0.139	1.016	0	RNA-binding region RNP-1 (RNA recognition motif)
32100	0	1	0	0	0	0	0	1	0.568	0.035	1.904	-0.450	0	Hypothetical protein
32208	0	1	0	0	0	0	0	1	0.582	0.039	2.529	-0.372	0	Hypothetical protein
32272	0	0	0	0	1	1	0	1	0.523	0.049	-0.349	1.309	0	Hypothetical protein
32374	0	0	0	0	0	1	0	1	0.816	0.003	-0.028	2.781	0	Hypothetical protein
32459	0	0	0	0	0	1	0	1	1.000	0.001	-0.140	2.814	1	Possible EPSP synthase (3-phosphoshikimate 1-c
32461	0	0	0	0	0	1	0	1	0.577	0.047	-0.063	2.707	0	FIG00941425: hypothetical protein

32539	1	0	0	1	0	0	0	1	0.612	0.022	-0.276	-0.782	0	Hypothetical protein
32570	0	0	0	0	0	1	0	1	0.776	0.005	-0.158	2.597	0	Hypothetical protein
32681	0	0	0	0	0	1	0	1	0.816	0.002	-0.079	2.853	0	Hypothetical protein
32698	0	1	0	0	0	0	0	1	0.630	0.011	2.742	-0.345	0	Glutamate N-acetyltransferase (EC 2.3.1.35) / N-acetylglutamate synthase (EC 2.3.1.1)
32788	0	0	0	0	0	1	0	1	0.577	0.047	-0.063	2.707	0	Hypothetical protein
32861	0	1	0	0	0	0	0	1	0.662	0.007	1.959	-0.208	0	Exodeoxyribonuclease V beta chain (EC 3.1.11.5)
32970	0	0	0	1	0	0	0	1	0.816	0.001	-0.403	-0.698	0	Possible Herpesvirus UL6 like
33006	0	1	0	0	0	0	0	1	0.572	0.047	1.744	-0.099	0	Possible Adenylate cyclase
33502	0	0	0	0	0	1	0	1	0.816	0.003	-0.189	2.800	0	Possible lipoprotein
33512	0	0	0	1	0	0	0	1	0.577	0.038	-0.476	-0.659	1	Possible PTS system| Lactose/Cellobiose specif
33582	0	1	0	0	0	0	0	1	0.548	0.032	2.680	-0.288	0	Hypothetical protein
34778	0	0	0	0	0	1	0	1	0.577	0.047	-0.063	2.707	0	Cobalt-zinc-cadmium resistance protein CzcD
36738	0	1	0	0	0	0	0	1	0.633	0.026	2.253	-0.190	0	Hypothetical protein
37275	0	1	0	0	0	0	0	1	0.548	0.02	2.641	-0.235	0	Hypothetical protein
39232	0	0	0	0	0	1	0	1	0.577	0.047	-0.063	2.707	0	Hypothetical protein
39426	0	0	0	0	0	1	0	1	0.577	0.047	-0.063	2.707	0	Hypothetical protein
39923	0	0	0	0	0	1	0	1	1.000	0.001	-0.124	2.809	0	Hypothetical protein
40855	0	0	0	0	0	1	0	1	0.761	0.004	0.945	1.909	0	Hypothetical protein
42339	0	0	0	1	0	0	0	1	0.500	0.049	-0.345	-0.495	0	Hypothetical protein
42600	0	1	0	0	0	0	0	1	0.588	0.031	3.111	-0.260	0	Hypothetical protein
42643	0	1	0	0	0	0	0	1	0.707	0.003	2.825	-0.083	0	Alkyl hydroperoxide reductase subunit C-like protein

Table S1.5. Low temperature associated genes. Genes related to low temperature were identified by a negative correlation to temperature (p-value < -0.05). Additionally provided are positive association with the only the South Atlantic Ocean (SA, associated p-values from indicator analysis), axes values from the gene content canonical correspondence analysis (CCA1, CCA2), number of HLI genomes with a gene in the particular COG (HLI), the number of genomes with a gene in the particular COG (Genomes) and the functions associated with members of the COG. An asterisk denotes COGs used in the metagenomic assembly analysis.

COG	Temperature	SA	CCA1	CCA2	HLI	Genomes	Function
3146	-0.56(6.36E-06)	0.001	-0.07	2.54	2	18	Conserved hypothetical protein: FIG00943604
6961*	-0.53(2.19E-05)	0.014	-0.18	1.72	3	39	Possible Trypsin | 2OG-Fe(II) oxygenase
10190*	-0.51(4.89E-05)	0.002	-0.09	2.24	3	33	Dihydroorotase (EC 3.5.2.3)
484	-0.50(7.48E-05)	0.001	-0.18	2.67	3	18	Conserved hypothetical protein: FIG00943123
6911*	-0.49(1.44E-04)		-0.38	0.85	3	29	Pseudouridine-5' phosphatase (EC 3.1.3.-) HAD-superfamily hydrolase| subfamily IA| variant 3 HAD-superfamily hydrolase subfamily IA| variant 3
39923	-0.48(1.54E-04)	0.001	-0.12	2.81	1	1	Hypothetical protein
1288	-0.48(2.16E-04)	0.091	-0.38	0.55	0	34	3-polyprenyl-4-hydroxybenzoate carboxy-lyase (EC 4.1.1.-) | Hydroxyaromatic non-oxidative decarboxylase protein C (EC 4.1.1.-)
27383	-0.47(2.37E-04)	0.012	-0.18	2.01	2	6	GtrA family protein | hypothetical protein
32459	-0.47(2.62E-04)	0.001	-0.14	2.81	2	2	Possible EPSP synthase
29774	-0.46(3.47E-04)	0.001	-0.22	1.81	2	5	Hypothetical protein: FIG00942706
32080	-0.46(3.55E-04)	0.001	-0.14	1.02	1	2	RNA-binding region RNP-1 (RNA recognition motif)
25664	-0.46(3.65E-04)	0.001	-0.17	1.63	2	3	Gp0.7
31821	-0.46(4.17E-04)	0.004	-0.09	2.50	2	2	Possible Trehalase
475	-0.45(4.61E-04)		-0.27	0.12	3	13	Cytochrome oxidase C subunit VIb-like
29417*	-0.44(6.66E-04)	0.005	-0.14	2.51	3	3	Possible envelope glycoprotein-like protein | possible Hepatitis C virus envelope glycoprotein
32570	-0.44(7.54E-04)	0.005	-0.16	2.60	2	2	Hypothetical protein
2019	-0.44(8.07E-04)	0.001	-0.22	1.72	3	51	Hypothetical protein: FIG023675
9846	-0.43(1.09E-03)		-0.31	0.95	3	9	Hypothetical protein
2245	-0.43(1.09E-03)		-0.10	0.17	3	57	L-lactate permease | hypothetical protein: FIG00940131
305	-0.42(1.16E-03)		-0.23	0.50	3	50	Predicted protein family PM-15| hypothetical proteins: FIG00944262; FIG01149883; FIG00942425; FIG01152647; FIG01155851; FIG01156338; FIG00940621; FIG01154924; FIG01155479
1731	-0.42(1.20E-03)	0.005	-0.24	2.08	0	15	Protein family PM-12 | hypothetical protein: FIG00940535
794	-0.42(1.29E-03)		-0.32	0.44	2	19	Hypothetical protein: FIG00941298
25014	-0.41(1.74E-03)	0.048	-0.41	1.12	0	10	Predicted protein family PM-12| hypothetical protein
30630	-0.41(1.74E-03)	0.048	-0.41	1.12	2	3	Hypothetical protein
29378	-0.41(1.92E-03)	0.003	-0.16	2.78	2	7	Hypothetical protein
31703*	-0.41(1.92E-03)	0.003	-0.16	2.78	3	5	Hypothetical protein: FIG00942523; FIG0115330
30772	-0.40(2.01E-03)	0.003	1.31	-0.06	2	6	Cell division protein FtsK | hypothetical protein
1369	-0.40(2.08E-03)		-0.30	0.75	3	51	Phospholipid-lipopolysaccharide ABC transporter| ABC-type multidrug transport system| ATPase and permease components| ATP-binding component| Lipid A export ATP-binding/permease protein | MsbA (EC 3.6.3.25) |ABC transporter transmembrane region: ABC transporter:AAA ATPa... |ATP-binding protein of ABC transporter | ATP-binding/permease protein ABC transporter transmembrane region:ATP/GTP-binding site mot... |ABC-type multidrug transport system |ATPase and permease components ABC-type nitrate/sulfonate/bicarbonate transport system| ATPase component hypothetical protein
11503	-0.40(2.11E-03)	0.13	-0.11	2.29	1	24	Hypothetical protein
31826	-0.40(2.11E-03)	0.13	-0.11	2.29	0	7	Glycosyl transferase| group 1/2 family protein| Exopolysaccharide biosynthesis glycosyltransferase EpsF (EC 2.4.1.-)|

							Glycosyl transferase group 1 |hypothetical protein: FIG01151293| glycosyl transferase| group 1 exopolysaccharide biosynthesis protein | Glycosyl transferase group 1 family protein
33502	-0.39(2.83E-03)	0.003	-0.19	2.80	2	2	Possible lipoprotein
6809	-0.39(3.07E-03)		-0.57	-0.60	0	21	Hypothetical protein: FIG00942137; FIG00940985; FIG00942852
16865	-0.39(3.07E-03)		-0.57	-0.60	0	20	Hypothetical protein: FIG00944261; FIG00942472; FIG01151996
21434	-0.39(3.07E-03)		-0.57	-0.60	0	4	Possible LysM domain
33261	-0.39(3.07E-03)		-0.57	-0.60	2	2	FIG00941908: hypothetical protein
1318	-0.39(3.12E-03)		-0.35	0.73	3	55	LSU ribosomal protein L23p (L23Ae)
21090	-0.39(3.36E-03)	0.056	-0.32	1.05	0	1	Putative site-specific integrase/recombinase
1524	-0.38(3.67E-03)	0.052	-0.28	1.13	0	19	Possible Signal peptidase | hypothetical protein: FIG00941437; FIG00941485
29359	-0.38(3.67E-03)	0.054	-0.28	1.13	0	4	Hypothetical protein
32521	-0.38(3.67E-03)	0.054	-0.28	1.13	0	3	Possible phage integrase family| Mobile element protein hypothetical protein
5168	-0.38(3.67E-03)		-0.27	0.57	0	23	Cyanobacterial hypothetical protein | Hypothetical protein: FIG00942442; FIG01154500
26490	-0.38(3.67E-03)		-0.27	0.57	0	14	Predicted protein family PM-9 | hypothetical protein: FIG00941365
31135	-0.38(3.67E-03)		-0.27	0.57	2	6	Hypothetical protein: FIG00944010
951	-0.38(3.82E-03)		0.06	0.26	0	14	Adhesin-like protein hypothetical protein
24917	-0.38(4.01E-03)	0.003	-0.03	2.78	1	6	ATP-dependent protease La (EC 3.4.21.53) Type I | hypothetical protein: FIG00941506
32374	-0.38(4.01E-03)	0.003	-0.03	2.78	2	2	Hypothetical protein
8911	-0.38(4.38E-03)	0.132	-0.63	0.57	0	32	Hypothetical protein: FIG01154943; FIG00941191; FIG01149531; FIG01150676; FIG01150479
33326	-0.38(4.38E-03)		0.02	2.30	2	2	Hypothetical protein
3554	-0.37(4.47E-03)		-0.36	0.09	3	51	Ribosomal protein S12p Asp88 (*E. coli)* methylthiotransferase
4447	-0.37(4.68E-03)		-0.33	0.04	2	23	Plastoquinol terminal oxidase hypothetical protein
32681	-0.37(4.71E-03)	0.002	-0.08	2.85	2	2	Hypothetical protein
2956	-0.37(5.13E-03)	0.073	-0.06	1.12	2	35	Photosystem II protein PsbX
2824	-0.37(5.29E-03)		-0.29	0.15	3	22	ATP-dependent DNA ligase
1118	-0.37(5.43E-03)		-0.36	0.28	3	51	Hypothetical proteins: FIG01157505; FIG01151667; FIG01149666; FIG01152535; FIG00940112; FIG00943708; FIG00940435; FIG01149201; FIG01155922; FIG00944363; FIG00941013; FIG01152750; FIG01156728
2868	-0.37(5.65E-03)		-0.28	0.37	3	34	Hypothetical protein: FIG00940812
57	-0.36(6.05E-03)		-0.34	0.20	2	25	Hypothetical protein: FIG00942626; FIG00941641
31231	-0.36(6.49E-03)	0.066	-0.03	2.38	2	2	Hypothetical protein
21718	-0.34(9.23E-03)	0.1	-0.23	0.85	3	17	Hypothetical protein FIG00941981
12427	-0.34(9.45E-03)		-0.27	0.86	3	37	Hypothetical protein FIG00942946; FIG00941489
883	-0.34(1.07E-02)		-0.03	0.34	3	57	Cytosine deaminase (EC 3.5.4.1)
26603	-0.34(1.09E-02)		0.08	0.50	0	15	Possible glycosyltransferase | TPR repeat possible glycosyltransferase | TPR domain protein | TPR repeat-containing protein | Translation elongation factor P COG0457: FOG: TPR repeat FOG: TPR repeat hypothetical protein | Hypothetical protein: FIG01150547
191	-0.34(1.15E-02)		-0.20	0.55	3	52	Hypothetical protein: FIG00940803; FIG01150744
24957	-0.34(1.16E-02)		-0.35	-0.22	2	15	Hypothetical protein: FIG01153374; FIG00942680
214	-0.33(1.20E-02)		0.02	0.36	3	54	Pyrroline-5-carboxylate reductase (EC 1.5.1.2)
295	-0.33(1.29E-02)		-0.19	0.32	3	50	Acetylornithine aminotransferase (EC 2.6.1.11)
3566	-0.33(1.29E-02)		-0.37	0.21	3	31	Photosystem II protein PsbY
1664	-0.33(1.35E-02)		-0.27	0.35	3	54	Putative arsenate reductase - Arsenate reductase (EC 1.20.4.1) | Arsenate reductase and related proteins - glutaredoxin family | Arsenate reductase related protein (arsC family)
238	-0.33(1.39E-02)		0.12	-0.26	0	34	Possible beta-N-acetylglucosaminidase | Beta-hexosaminidase (EC 3.2.1.52) | Beta-glucosidase-related glycosidases
11607	-0.33(1.43E-02)	0.001	-0.30	1.05	3	26	Possible Helix-turn-helix protein| copG family Possible Helix-turn-helix protein| copG family
1897	-0.33(1.44E-02)		0.87	0.97	3	14	Conserved hypothetical protein: FIG00940214
19516	-0.32(1.58E-02)		-0.20	0.53	0	11	Aspartate carbamoyltransferase (EC 2.1.3.2)
24726	-0.32(1.67E-02)		-0.33	0.56	2	10	Serine protease precursor MucD/AlgY associated with sigma factor RpoE Serine protease| DegP/HtrA| do-like (EC

							3.4.21.-) | Protease hypothetical protein serine proteinase
28631	-0.31(1.95E-02)		-0.30	-0.12	2	8	Possible Myosin N-terminal SH3-like domain
1371	-0.31(1.97E-02)		-0.23	0.24	2	43	Ferredoxin| 2Fe-2S
1189	-0.31(2.06E-02)		-0.14	0.67	3	57	Hypothetical protein: FIG01149571; FIG00940849; FIG01150457; FIG01153262
27373	-0.30(2.27E-02)		-0.51	-0.04	0	7	Hypothetical protein: FIG00943862
2726	-0.30(2.41E-02)		0.00	1.21	3	28	Hypothetical protein: FIG00943806; FIG00941615
13298	-0.30(2.42E-02)	0.057	-0.25	2.85	0	11	Predicted protein | hypothetical protein
14644	-0.30(2.42E-02)	0.057	-0.25	2.85	0	4	Hypothetical protein: FIG00942127; FIG00943222
18019	-0.30(2.42E-02)	0.057	-0.25	2.85	0	5	Hypothetical protein: FIG00943805
31728*	-0.30(2.42E-02)	0.057	-0.25	2.85	3	3	Hypothetical protein: FIG00944641
32695	-0.30(2.42E-02)	0.057	-0.25	2.85	0	3	Hypothetical protein
35876	-0.30(2.42E-02)	0.057	-0.25	2.85	1	1	Hypothetical protein
41250	-0.30(2.42E-02)	0.057	-0.25	2.85	1	1	Hypothetical protein
16566	-0.30(2.64E-02)		-0.50	-0.49	2	3	Alkaline phosphatase (EC 3.1.3.1)

CHAPTER 2

Parallel phylogeography of *Prochlorococcus* and *Synechococcus*

Abstract

Extensive genetic diversity has been observed within *Prochlorococcus* and *Synechococcus,* including the presence of multiple major clades. *Prochlorococcus* has clear environmental selective pressures controlling its distribution that are mirrored in its evolutionary history. However, the biogeography and underlying environmental drivers of *Synechococcus* clades have been difficult to define. We identified the genetic diversity of *Prochlorococcus* and *Synechococcus* from 339 samples across latitudinal transects of the eastern Pacific Ocean between 3˚S and 19˚N and northern Atlantic Ocean between 19˚N and 55˚N using high throughput sequencing of the marker gene *rpo*C1. We identified the phylogenetic affiliation of each sequence and detected extensive microdiversity within each clade. We observed clear biogeographical domains, with *Synechococcus* Clade CRD1 peaking at the equator, a shift to Clades II+III around 7˚N to 19˚N and dominant from 19˚N to 40˚N in the Atlantic transects, and Clades I+IV in the northern part of the Atlantic transect. This overall biogeography closely matched the distribution of *Prochlorococcus* diversity in this region, suggesting a parallel evolution of ecotypes in these two major lineages of marine Cyanobacteria. The microdiversity within *Prochlorococcus* and *Synechococcus* clades was constrained in part by environmental variation, had distinctive biogeography with clear Atlantic and Pacific lineages in some clades, and also reflected the parallel evolution of the clades within these major lineages. Overall, parallel biogeography at multiple taxonomic

resolutions suggests similar evolutionary selective pressures for these important marine Cyanobacteria.

Keywords: *Prochlorococcus*, *Synechococcus*, phylogenetics, biogeography, microdiversity

Introduction

The marine Cyanobacteria *Prochlorococcus* and *Synechococcus* are globally abundant and together account for approximately a quarter of ocean primary production (Flombaum *et al.*, 2013). The overall distribution of the two lineages is different. *Prochlorococcus* is most abundant at ocean temperatures above 20˚C, whereas *Synechococcus* has a wider range but with a maximum abundance near 10˚C. However, the two lineages co-occur across broad environmental gradients in subtropical and tropical waters (from ~40˚S to 40˚N).

There are many similarities but also a few distinctions in the biology of *Prochlorococcus* and *Synechococcus*. They are both unicellular, derive energy from photosynthesis, have small genomes covering extensive genetic and phylogenetic diversity (Scanlan *et al.*, 2009), use the same nutrient sources (Moore *et al.*, 2002, 2005), and die through a combination of oxidative stress, grazing, and viral infections. Some differences between the lineages include molecular composition of the light-harvesting system (Ting *et al.*, 2001) and cell size where *Synechococcus* is slightly larger than *Prochlorococcus* (Partensky *et al.*, 1999). Thus, it appears that *Prochlorococcus* and *Synechococcus* share many physiological traits in addition to their habitat and therefore may be subject to similar selective pressures in the ocean.

The evolutionary diversification of *Prochlorococcus* has been closely associated with several environmental variables (Figure 2.1). At the basal phylogenetic level, *Prochlorococcus* is broadly divided into two groups: low- and high-light adapted clades (Moore *et al.*, 1998). The low-light adapted clades are found at increasing depth with LLI located near the nutricline, LLII+III right below and LLIV at the bottom of euphotic zone (Ahlgren *et al.*, 2006; Zinser *et al.*, 2007). There are also additional low-light adapted lineages but the phylogenetic position and depth distribution of these are less clear (Martiny, Tai, *et al.*, 2009). At the next-deepest phylogenetic level, the high-light group can be divided into high- and low-iron (named HNLC) adapted clades (Rusch *et al.*, 2010) and the high-iron clade can be further divided into low- (HLI) and high-temperature (HLII) clades. No known mechanisms govern the maintenance of fine-scale diversity (Larkin & Martiny, 2017), but overall *Prochlorococcus* ecotypes have a clear phylogeography.

Despite the shared biology and habitat, the biogeography of *Synechococcus* genetic diversity does not seem to match *Prochlorococcus*. The current notion is that two clades (I and IV) are most abundant in colder, nutrient rich coastal waters, while five clades (Clade II, III, V, VI, and VII) to various degrees are frequent in tropical and subtropical waters (Fuller *et al.*, 2003; Zwirglmaier *et al.*, 2008, 2007; Mella-Flores *et al.*, 2011; Post *et al.*, 2011). Clade VIII may be specifically adapted to hypersaline waters (Dufresne *et al.*, 2008; Huang *et al.*, 2012), while other clades such as IX and X have been rarely seen and only at low abundance (Zwirglmaier *et al.*, 2008, 2007). In *Prochlorococcus*, a single ecotype appears to dominate a specific environment,

whereas the biogeography *Synechococcus* is characterized by overlapping distributions of genotypes with similar biology.

The biogeography of *Synechococcus* has primarily been examined with molecular probes and thus focused on known groups at various taxonomic scales. Thus, the traditional clade designation may not capture important biogeographical patterns – either within clades or as a deeper clade (and thus the sum of multiple defined groups) (Gutiérrez-Rodríguez *et al.*, 2014; Martiny, Tai, *et al.*, 2009; Martiny *et al.*, 2013; Mazard *et al.*, 2012). It could also be that traits related to light, temperature and nutrients are phylogenetically promiscuous and consequently do not lead to a clear phylogeography in *Synechococcus* (Gutiérrez-Rodríguez *et al.*, 2014).

A closer look at the phylogeny of *Synechococcus* suggests that the presumed difference in phylogenetic structure between sister lineages may be a product of the phylogenetic scales applied or the taxonomic marker used. A multi-locus sequencing analysis of strains from the major *Synechococcus* clades shows that Clades II+III actually form a monophyletic clade (Figure 2.1, Mazard *et al.*, 2012). The same is the case for Clades I+IV. There is even evidence that the CRD1 clade diverged before Clades II+III and Clades I+IV as seen for HNLC within the *Prochlorococcus* radiation. Thus, the phylogenetic distribution of *Prochlorococcus* and *Synechococcus* ecotypes appears to occur in a similar hierarchical fashion (Martiny, Tai, *et al.*, 2009).

Here, we want to quantify the phylogeography of *Prochlorococcus* and *Synechococcus* over broad environmental gradients using high-throughput sequencing of a variable phylogenetic marker. The single-copy gene encoding the gamma subunit of RNA polymerase (*rpo*C1) has been effective at delineating between clades of

Prochlorococcus and *Synechococcus* (Gutiérrez-Rodríguez *et al.*, 2014; Palenik, 1994). Based on this dataset, we will test the degree to which the biogeography across phylogenetic scales (lineages, clades, and SNPs) of these two important marine phytoplankton is shared.

Results

To identify the detailed phylogeography of *Prochlorococcus* and *Synechococcus*, we sequenced a phylogenetic marker gene (*rpo*C1) from 339 populations from the Tropical Pacific and North Atlantic Oceans (Figure 2.2, Table S2.1). We sampled the Tropical Pacific Ocean transect (3˚S to 19˚N) and the surface water temperature showed limited variation (27.2 – 29.3˚C) (Figure 2.3A). The water column was highly stratified with a clear shallow thermocline. Macronutrient availability was lower in the northern section; with phosphate concentration between 100 and 200 nM (Figure 2.3B) and nitrate concentration at detection limit (< 10 nM). A sharp transition zone was observed at 5˚N with macronutrient concentrations elevated in the southern section (phosphate > 400 nM). The elevated nutrient concentration near the equator was likely driven by upwelling and indicative of iron stress south of 5˚N (Fitzwater *et al.*, 1996). The environmental conditions in the North Atlantic Ocean between 19.7˚N and 55˚N also displayed a clear transition zone (Figure 2.3A+B). This transition zone was seen in the temperature profile whereby the surface temperature was above 26˚C below 39˚N. North of this transition point, the surface temperature was lower (~19.5˚C) and subsequently dropped to ~10˚C at 55˚N. Similarly, the phosphate concentration was very low (<10 nM) below 39˚N but was elevated above this point (Figure 2.3B). The nutricline was also deep (~ 200 m) below 39˚N and shallower further north (~40 m).

There were two exceptions to this profile. There was an infusion of deep water at 37˚N and slightly elevated nutrient concentrations between 23˚N and 19.7˚N due to horizontal nutrient supply by the Caribbean Current.

The abundances of picophytoplankton lineages also displayed a clear transition, albeit one or two degrees shifted northwards of the nutrient gradient (i.e., 6˚N-7˚N). Both *Synechococcus* and picoeukaryotic phytoplankton abundances were approximately an order of magnitude higher south of 6˚N (Figure 2.3D+E). In contrast, there was little latitudinal variation in *Prochlorococcus* surface abundances (Figure 2.3C). In the Atlantic Ocean, the abundances of picophytoplankton lineages mirrored the environmental transition at 39˚N whereby, a clear shift from *Prochlorococcus* to elevated abundances of *Synechococcus* and picoeukaryotic phytoplankton occurred (Figure 2.3C-E).

Previous *Prochlorococcus* and *Synechococcus* clade designations were confirmed using the marker gene *rpo*C1. After identifying novel references from within the dataset, no major new clades were identified. While there are no strain representatives of clade NC1 this clade was well represented across these transects. Assigning sequences to their respective clades at different sequence identity cutoffs did not significantly change the overall clade distribution until >97% amino acid identity (Mantel test; Table S2.2). Some *Synechococcus* closely related phylogenetic clades overlapped in their biogeography and were analyzed together after ($P<0.01$ for Clades I and IV; $P=0.056$ for Clades II and III; latitudinal randomization test, Figure S2.1 + S2.2).

In parallel with the overall phytoplankton community composition, including picoeukaryotes, the relative frequency of major *Prochlorococcus* and *Synechococcus*

clades also shifted at the 6˚N-7˚N transition point (Figure 2.4). For *Prochlorococcus*, the equatorial section was dominated by the low iron adapted HNLC clade (Figure 2.4C), whereas the northern mixed layer was dominated by the high temperature adapted HLII clade (Figure 2.4B). It is worth noting that the transition between HNLC and HLII matched the shift in abundance of *Synechococcus* and picoeukaryotic phytoplankton. The HLI clade was only observed right above the nutricline in the northern part, but was absent in the equatorial part (Figure 2.4A). The low-light adapted clades showed clear depth partitioning, whereby LLI was detected right below the high-light ecotypes followed by LLII/III, LLIV and NC1 at the bottom of the euphotic zone (Figure S2.3). The latitudinal separation of *Prochlorococcus* ecotypes was also seen for *Synechococcus*. Clade CRD1 was most frequent in the equatorial part (Figure 2.4F) and Clades II+III dominated the surface mixed layer north of 7˚N (Figure 2.4E).

In the North Atlantic Ocean, we again saw a parallel latitudinal ecotype distribution for *Prochlorococcus* and *Synechococcus* (Figure 2.4). There was a clear transition at 39˚N whereby HLII dominated the southern regions. *Prochlorococcus* HLI was most frequent in the northern regions between 43˚N and 49˚N, after which *Prochlorococcus* as a whole nearly disappeared (Figure 2.4A). However, HLI was also present deeper in the water column near the 10 nM phosphate nutricline in the southern part of the transect. As seen in the Pacific samples, *Prochlorococcus* LLI was the dominant low-light clade and became frequent right below the nutricline (Figure S2.3A). LLII/III and LLIV were also detected deeper in the water column (Figure S2.3B+C). *Synechococcus* Clades II+III were most frequent south of 39˚N (Figure 2.4E), whereas Clades I+IV dominated the *Synechococcus* populations between 41 and 55˚N (Figure

2.4D). Thus, both *Prochlorococcus* and *Synechococcus* clades displayed clear latitudinal distributions with transition points at 6˚N for the Pacific Ocean and 39˚N for the Atlantic Ocean transect.

We next quantified the degree to which individual clades had a shared biogeography using Pianka's index (Pianka, 1973). Three pairs of *Prochlorococcus* and *Synechococcus* clades significantly overlapped in their spatial distributions ($P<0.01$; latitudinal randomization test, Figure S2.4). Specifically, *Prochlorococcus* HNLC significantly overlapped with the *Synechococcus* CRD1 with high abundance in the hot, elevated macronutrient waters between 3˚S and 6˚N in the Pacific Ocean and largely absent elsewhere. *Prochlorococcus* HLII overlapped with *Synechococcus* Clades II+III in hot but low macronutrient waters between 6˚N and 20˚N in the Pacific Ocean and between 21˚N and 39˚N in Atlantic Ocean. *Prochlorococcus* HLI overlapped with *Synechococcus* Clades I+IV in cooler but elevated macronutrient waters north of 39˚N in the Atlantic Ocean. Thus, there was significant evidence for a parallel biogeography of *Prochlorococcus* and *Synechococcus* clades.

Within these clades, extensive regional microdiversity co-occurred within closely related members of the two most frequently observed clades of *Prochlorococcus* (HLII and HNLC) and *Synechococcus* (Clades II+III, and CRD1). While sequences within these clades were similar at the amino acid level, there was extensive synonymous variation such that most sequences were entirely unique. Notably, this variation was not random by region, but instead the observed microdiversity was biogeographically and environmentally structured (Figure 2.5). Ocean origin (Pacific vs. Atlantic) significantly explained variation in the single nucleotide polymorphisms (SNPs) profiles across each

clade (PerMANOVA; P<0.005 for each clade; Table S2.3). In addition, microdiversity differed by latitude, temperature and depth (dbRDA; Table S2.4 and Figure S2.5). Within the microdiversity, *Prochlorococcus* HLII and *Synechococcus* Clades II+III had SNP profiles that were unique to the Atlantic and Pacific Oceans suggesting adaptation to environmental differences between the two oceans (Figure 2.5 and Figure S2.6). In addition, there was a major latitudinal transition between profile types of each clade (HLII, HNLC, CRD1, Clades II+III) near where clades shifted in the Pacific Ocean (Figure S2.6). In general, the changes followed the clade transitions with major shifts near 6 to 7˚N, however the Clades II+III shifted between 5 and 6˚N and HLII SNP profiles shifted even further south between 4 and 5˚N. Thus, we found evidence for regional SNP profiles among *Prochlorococcus* and *Synechococcus* populations.

Depth structured the microdiversity of each clade (Figure S2.7). To illustrate this, we sampled several sites at high vertical resolution (every 5 m down to 50m and then every 10 m to 200 m). At 18˚N, the microdiversity for *Prochlorococcus* HLII was uniform down to 70 m, which covered both the mixed layer (depth ~ 20 m) as well as stratified waters below (Figure S2.7A). At 70 m, we observed a subtle shift whereas the population deep in the euphotic zone was very different. Thus, there were distinct vertical populations of the high-light adapted clade HLII in the water column. The population structure of *Synechococcus* Clades II+III showed a parallel vertical distribution (Figure S2.7B). Here, the populations were similar within and below the mixed layer, whereas different populations were found deeper in euphotic zone. At 2˚S in the Pacific Ocean, the microdiversity for *Prochlorococcus* HNLC and *Synechococcus* CRD1 were uniform within the mixed layer (mixed layer depth: 64 m) and below to ~100

m (Figure S2.7C+D). Below this point, we observed slight variations the structure for both HNLC and CRD1. As seen regionally, we observed clear analogous variations in the vertical microdiversity structure for both *Prochlorococcus* and *Synechococcus* clades.

Discussion

By using the same genetic marker and sequencing method, we show that the phylogeography of *Prochlorococcus* and *Synechococcus* are tightly concordant across oceans. Building on a series of recent studies (Farrant *et al.*, 2016; Sohm *et al.*, 2016; Zwirglmaier *et al.*, 2008), we confirm that *Prochlorococcus* and *Synechococcus* each contain three major surface ocean ecotypes and show that these ecotypes are in parallel with one another. The first ecotype (HNLC and CRD1) is adapted to high temperature, elevated macronutrients and presumably low iron availability. In our study, this ecotype is restricted to iron-stressed Eastern Pacific Equatorial Zone but other studies have also detected HNLC/CRD1 in the Indian Ocean upwelling zones (Rusch *et al.*, 2010; Farrant *et al.*, 2016), the Costa Rica Dome water (Ahlgren *et al.*, 2014; Gutiérrez-Rodríguez *et al.*, 2014), and the Benguela Upwelling Zone (Sohm *et al.*, 2016). The second ecotype (HLII and Clades II+III) is present in high temperature, low macronutrient and high iron waters is in our study sandwiched between equatorial upwelling zones and colder, nutrient rich mid-latitude waters. This distribution pattern is also supported by other studies (Zwirglmaier *et al.*, 2008; Johnson *et al.*, 2006). The third major ecotype (HLI and Clades I+IV) is found in higher latitude colder, nutrient rich waters. HLI is restricted to a narrow band as *Prochlorococcus* as a group is restricted to

regions with ocean temperature above ~15˚C, whereas the *Synechococcus* ecotype extend further north (Zwirglmaier *et al.*, 2008).

The parallel phylogeography between *Prochlorococcus* and *Synechococcus* extends even into the fine-scaled genetic diversity of both lineages. Here we observe patterns of biogeography within the microdiversity of both *Prochlorococcus* and *Synechococcus* that correlate with one another between the groups HLII and Clades II+III and the groups CRD1 and HNLC. This variation was also structured by environmental parameters, suggesting the observed diversity represents adaptation to the local environment as opposed to stochastic processes (Larkin & Martiny, 2017). Indeed, others have observed that novel biogeographic patterns were associated with microbial microdiversity unseen at higher taxonomic levels (Needham *et al.*, 2017; Buttigieg & Ramette, 2015). In our case, microdiversity captured shifts within clades between ocean biomes not seen at the clade level (Larkin et al., 2016). A strong difference in the fine-scaled genetic diversity was also observed in a comparison of single cell genomes as well as metagenomes from the Pacific and Atlantic Ocean subtropical gyres (Kashtan *et al.*, 2017; Coleman & Chisholm, 2010; Martiny, Huang, *et al.*, 2009; Martiny, Kathuria, *et al.*, 2009). A difference in nutrient concentrations and consequently nutrient stress between the gyres likely underscores this and our biogeographic pattern (Wu *et al.*, 2000). We also saw that the microdiversity SNP profiles shifted at the same location in the Pacific Ocean as the shifts observed at higher taxonomic levels, suggesting that the diversity at many levels is subject to the same selection (Farrant *et al.*, 2016).

We find that phytoplankton diversity across picophytoplankton lineages, clades, and microdiversity all shift at sharp transition points. These transitions occur at 6˚N in the Tropical Pacific Ocean and at 39˚N in the North Atlantic Ocean and may represent fundamental shifts in the environmental conditions supporting phytoplankton growth. In the equatorial region, picophytoplankton are stressed by iron availability. In the subtropical region, phytoplankton are stressed by macronutrient availability. In the mid-latitude region, phytoplankton are stressed by energy availability driven by a combination of lower irradiance and slower enzyme kinetics (due to temperature). These transitions are biologically amplified as phytoplankton contribute to the availability of resources (i.e., nutrients and to some extent light). The uniformity in distribution of phytoplankton types across these environmental zones suggests that all lineages experience similar stressors. In some cases, adaptation can overcome the stress condition as seen in the leveled distribution of *Prochlorococcus* across the tropical Pacific Ocean. In contrast, the abundance of *Synechococcus and* picoeukaryotic phytoplankton were more sensitive to environmental changes as indicated by the change in overall abundance around 6˚N-7˚N. In sum, our study suggests that *Prochlorococcus* and *Synechococcus* co-occur in regions with specific environmental stress conditions that lead to a parallel evolutionary diversification and biogeography across phylogenetic scales.

Methods

Samples were collected during three cruises over four years: the mid North Atlantic Ocean cruise Bval46 from September 28th to October 12th, 2011, the North Atlantic Ocean cruise AE1319 from August 15th to September 8th, 2013, and the North

Pacific Ocean Cruise NH1418 from September 20th to October 6th, 2014. See Table S2.1 for cruise information and sample data.

Nutrient and cell abundance measurements

Nutrient samples were collected after filtration through 0.8 µm polycarbonate filters (Nucleopore). Soluble reactive phosphorus (SRP) was determined using high-temperature acid persulfate oxidation on a Genesys 10UV spectrophotometer (Thermo Fisher Scientific) after preparation via the magnesium-induced co-precipitation method (Karl & Tien, 1992; Lomas *et al.*, 2010). Nitrate was analyzed on a Seal Analytics AA3 autoanalyzer (Seal Analytics) with a detection limit of 30 nmol N L^{-1}. For cell counts, samples of whole seawater were collected in 2 mL centrifuge tubes. Freshly made 0.2 µm-filtered paraformaldehyde (0.5% v/v final concentration) was added to all samples and allowed to fix for at least 1 hour, after which they were stored at -80°C until analysis. Cell counts were performed on a FACSJazz flow cytometer or a Becton Dickinson Influx flow cytometer utilizing a 200 mW 488 nm laser, with detectors for forward scatter, side scatter, 692 nm, and 530 nm. Instrument alignment was performed with 3.0 µm 6-peak rainbow beads, while hourly checks on forward scatter response were performed with 0.53 µm Nile Red beads (Spherotech). Using 8 g kg^{-1} water for sheath, *Prochlorococcus* populations were discriminated based on forward scatter and red fluorescence, and a gate in orange (530 nm) discriminated for *Synechococcus*. Eukaryotes were all the large red autofluorescing cells that did not fit the cyanobacterial gating scheme.

DNA extraction

4 L of seawater were prefiltered on a GF/D, 2.7μm glass-fiber filter (Whatman, MA) before being collected on a 0.22 μm Sterivex filter (Millipore, MA). 1.62 mL TES buffer (50 mM Tris-HCl pH 7.6, 20 mM EDTA pH 8.0, 400 mM NaCl, 0.75 M sucrose) was added before frozen at -20°C until further processing. DNA was extracted following Bostrom and co-workers (Bostrom *et al.*, 2004). Filters were thawed, 180μl of lysozyme buffer (50 mg/ml) added, and then incubated at 37°C for 30 minutes. After adding 180 μl of proteinase K (1 mg/ml) and 100 μl of 10% sodium dodecyl sulfate, the filters were incubated at 55°C overnight. The filter liquid was removed and combined with 3M sodium acetate (pH 5.2) and cold isopropanol and incubated at -20°C for > 1 h. After centrifugation at 15,000 g for 30 minutes at 4°C, the supernatant was removed and the pellet resuspended in Tris-EDTA buffer (10mM Tris pH 8, 1mM EDTA) in a 37°C water bath for 30 minutes. Finally, DNA was purified using a genomic DNA Clean and Concentrator kit (Zymo Corp., Irvine, CA) and stored at -20°C.

PCR amplification and sequencing

DNA concentration was quantified with Qubit dsDNA HS assay kit (Life Technologies, Invitrogen) and subsequently diluted in Tris (10mM, pH 8) solution to a concentration of 1 ng/μl. We modified primers targeting *rpo*C1 in only marine *Prochlorococcus* and *Synechococcus* lineages to ensure they targeted all known isolates and metagenomics sequences (Tai & Palenik, 2009). We then PCR amplified this region using the primers 5M_newF (5'-GARCARATHGTYTAYTTYA-3') and SACR1039R (5'-CYTGYTTNCCYTCDATDATRT-3'). This region was also chosen such that reference sequences had no indels and would enable merging of non-overlapping regions for analysis. 2 ng of DNA was added to a PCR cocktail of 1 μl each of 6μM

primer, 10 µl Premix F (Epicentre, Madison, WI), and 0.5 µl Taq polymerase (Hotmaster Taq polymerase, 5 PRIME, Hamburg, Germany), up to a final volume of 20 µl. After 35 cycles, 2 µl barcode oligonucleotides (Eurofins MWG Operon, Louisvillle, KY) following Illumina NexteraXT index sequences were added to each sample for another 10 PCR cycles. PCR products were verified on an agarose gel and 1-2 µl of each sample were pooled for cleanup using Agencourt AMPure XP beads (Beckman-Coulter Genomics, Danvers, MA). Final product was sequenced paired end 300bp on a MiSeq from Illumina (Laragen Inc., Culver City, CA).

Data analysis

Sequences were trimmed to the same length to maintain an average quality score >20 using FASTQC (Andrews, 2010). This reduced the second read by 81bp. The reverse complement of the trimmed read 2 was concatenated with read 1. Concatenated sequences were demultiplexed and quality filtered using split_libraries_fastq.py script with default settings from QIIME1.9 (Caporaso *et al.*, 2010). The top 100 references were chosen using pick_open_reference_otus.py from QIIME1.9. All sequences were searched against a custom database of known *Prochlorococcus* and *Synechococcus rpo*C1 sequences using tBlastx (Camacho *et al.*, 2009). Best hits were identified by best e-value. Due to the overabundance of *Prochlorococcus*, we took the best hits to *Synechococcus* references, found additional *Synechococcus* references using pick_open_reference_otus.py and added the top 100 references to the combined reference database to rerun tBlastx on all sequences. Sequences mapping to *Synechococcus* also had to have % GC> 41.22 which was 5 standard deviations below the lowest average % GC from any *Synechococcus* clade

and sequences mapping to *Prochlorococcus* had to have % GC< 49.07 which was 5 standard deviations above the highest average % GC from any *Prochlorococcus* clade. Sequences with less than 90% amino acid identity and 90% coverage were filtered out. Different percent identity thresholds were compared to one another using mantel test (Figure S2.1) and 90% identity was chosen to minimize deep-low sequence sample noise. Sequences that passed these filters were associated with known clades based on phylogeny of all reference sequences using raw sequence distance and clustering (Paradis *et al.*, 2004). This clustering also agreed with a maximum likelihood phylogeny (Felsenstein, 2005; Figure S2.8). Taxonomic abundance and environmental data were interpolated using the 'DIVA' algorithm in Ocean Data View (Schlitzer, 2002). To test the niche overlap, we estimated Pianka's niche overlap index on interpolated data from clades HLI, HLII, HNLC, I+IV, II+III, and CRD1 frequencies in the top 100m (Pianka, 1973). We estimated geographical distribution significance by comparison to a null model of data randomization across latitude.

Microdiversity analysis

Sequences mapping to the *Prochlorococcus* clades HNLC or HLII and the *Synechococcus* clades CRD1 or II+III at greater than 97% amino acid identity and 90% coverage were separated into their respective clades. Single nucleotide polymorphism (SNP) profiles were calculated as the per base pair majority at each nucleotide site for each sample (Biopython, Cock *et al.*, 2009). Sequences were highly unique within each clade (92.5% unique for HNLC, 95.9% for HLII, 99.8% for CRD1 and 99.9% for Clades II+III) potentially due to sequencing errors. Samples were assessed for their most informative nucleotides using minimum entropy decomposition (Eren *et al.*, 2014). The

100 most informative nucleotide positions per clade were identified for each sequence to limit any additional sequencing error noise. We randomly sampled 100 sequences from most samples and included all sequences from samples with fewer than 100 sequences, noted in Table S2.1. We assessed phylogenetic composition of each clade grouped at 100% sequence identity using the weighted Unifrac metric (Lozupone & Knight, 2005) implemented in R (function UniFrac; (Kembel *et al.*, 2010). Using the same top 100 most informative nucleotide positions per sequence, we assessed taxonomic similarity using pick_open_reference_otus.py from QIIME1.9 at an identity threshold of 85% coupled with Bray-Curtis dissimilarity (function vegdist; Oksanen *et al.*, 2013). We used the permutation multivariate analysis of variance test (function adonis; Oksanen *et al.*, 2013) to test if ocean origin explained the Unifrac-based phylogenetic composition variation. We used distance based redundancy analysis with the Unifrac distance to test if other environmental parameters (depth, latitude, and temperature) explained phylogenetic composition variation (function capscale; Oksanen *et al.*, 2013). We used the mantel test with Pearson correlations (function mantel; Oksanen *et al.*, 2013) to test if clades were correlated with one another based on their phylogenetic composition. We also used the mantel test to compare the three sample dissimilarity metrics: phylogenetic composition using Unifrac distance, the sample SNP profiles using Euclidean distance, or the taxonomic diversity using the Bray-Curtis dissimilarity (Table S2.5) yielding significant correlations between metrics for all clades with the weakest between Bray-Curtis dissimilarity of OTUs and Euclidean distance of SNP profiles (r=0.39).

Comparative phylogenies

Multi-locus sequence trees for *Prochlorococcus* and *Synechococcus* were recreated from previous analyses (Berube *et al.*, 2015; Mazard *et al.*, 2012). For *Prochlorococcus*, 503 core genes in *Prochlorococcus* were aligned separately using ClustalW (Larkin *et al.*, 2007; Berube *et al.*, 2015). For *Synechococcus*, 7 core genes were aligned separately with ARB (Ludwig *et al.*, 2004; Mazard *et al.*, 2012). For each lineage, gene sequences were concatenated, then a maximum likelihood tree was constructed using Phylip with neighbor-joining bootstrap support (Felsenstein, 2005), and rooted after with MIT9313 and WH5701 as outgroups for each phylogeny respectively.

References

Ahlgren NA, Noble A, Patton AP, Roache-johnson K, Robinson D, Mckay C, *et al.* (2014). The unique trace metal and mixed layer conditions of the Costa Rica upwelling dome support a distinct and dense community of *Synechococcus*. *Limnol Oceanogr* **59**:2166–2184.

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.

Andrews S. (2010). FastQC: A quality control tool for high throughput sequence data. *Http://WwwBioinformaticsBabrahamAcUk/Projects/Fastqc/*.

Berube PM, Biller SJ, Kent AG, Berta-Thompson JW, Roggensack SE, Roache-Johnson KH, *et al.* (2015). Physiology and evolution of nitrate acquisition in *Prochlorococcus*. *ISME J*.

Bostrom KH, Simu K, Hagstrom A, Riemann L. (2004). Optimization of DNA extraction

for quantitative marine bacterioplankton community analysis. *Limnol Oceanogr Methods* **2**:365–373.

Buttigieg PL, Ramette A. (2015). Biogeographic patterns of bacterial microdiversity in Arctic deep-sea sediments (HAUSGARTEN, Fram Strait). *Front Microbiol* **6**:1–12.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, *et al.* (2009). BLAST+: architecture and applications. *BMC Bioinformatics* **10**.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, *et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Publ Gr* **7**:335–336.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, *et al.* (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**:1422–1423.

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

Dufresne A, Ostrowski M, Scanlan DJ, Garczarek L, Mazard S, Palenik B, *et al.* (2008). Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biol* **9**:R90.

Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML. (2014). Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J* **9**:968–979.

Farrant GK, Doré H, Cornejo-Castillo FM, Partensky F, Ratin M, Ostrowski M, *et al.* (2016). Delineating ecologically significant taxonomic units from global patterns

of marine picocyanobacteria. *Proc Natl Acad Sci* **113**:E3365–E3374.

Felsenstein J. (2005). PHYLIP (Phylogeny Inference Package) version 3.69.

Fitzwater SE, Coale KH, Gordon RM, Johnson KS, Ondrusek ME. (1996). Iron deficiency and phytoplankton growth in the equatorial Pacific. *Deep Sea Res Part II Top Stud Oceanogr* **43**:995–1015.

Flombaum P, Gallegos JL, Gordillo R a, Rincón J, Zabala LL, Jiao N, *et al.* (2013). Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proc Natl Acad Sci U S A* **110**:9824–9829.

Fuller NJ, Marie D, Vaulot D, Post AF, Scanlan DJ. (2003). Clade-Specific 16S Ribosomal DNA Oligonucleotides Reveal the Predominance of a Single Marine. *Appl Environ Microbiol* **69**:2430–2443.

Gutiérrez-Rodríguez A, Slack G, Daniels EF, Selph KE, Palenik B, Landry MR. (2014). Fine spatial structure of genetically distinct picocyanobacterial populations across environmental gradients in the Costa Rica Dome. *Limnol Oceanogr* **59**:705–723.

Huang S, Wilhelm SW, Harvey HR, Taylor K, Jiao N, Chen F. (2012). Novel lineages of *Prochlorococcus* and *Synechococcus* in the global oceans. *ISME J* **6**:285–97.

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

Karl DM, Tien G. (1992). MAGIC: A sensitive and precise method for measuring dissolved phosphorus in aquatic environments. *Limnol Oceanogr* **37**:105–116.

Kashtan N, Roggensack SE, Berta-Thompson JW, Grinberg M, Stepanauskas R, Chisholm SW. (2017). Fundamental differences in diversity and genomic

population structure between Atlantic and Pacific *Prochlorococcus*. *ISME J* 1–15.

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, *et al.* (2010). Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**:1463–1464.

Larkin AA, Martiny AC. (2017). Microdiversity shapes the traits, niche space, and biogeography of microbial taxa. *Environ Microbiol Rep* **9**:55–70.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, *et al.* (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**:2947–2948.

Lomas MW, Burke AL, Lomas DA, Bell DW, Shen C, Dyhrman ST, *et al.* (2010). Sargasso Sea phosphorus biogeochemistry: an important role for dissolved organic phosphorus (DOP). *Biogeosciences* **7**:695–710.

Lozupone C, Knight R. (2005). UniFrac : a New Phylogenetic Method for Comparing Microbial Communities UniFrac : a New Phylogenetic Method for Comparing Microbial Communities. *Appl Environ Microbiol* **71**:8228–8235.

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

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

Martiny AC, Kathuria S, Berube PM. (2009). Widespread metabolic potential for nitrite and nitrate assimilation among *Prochlorococcus* ecotypes. *Proc Natl Acad Sci U S A* **106**:10787–10792.

Martiny AC, Tai APK, Veneziano D, Primeau F, Chisholm SW. (2009). Taxonomic resolution, ecotypes and the biogeography of *Prochlorococcus*. *Environ Microbiol* **11**:823–832.

Martiny AC, Treseder K, Pusch G. (2013). Phylogenetic conservatism of functional traits in microorganisms. *ISME J* **7**:830–838.

Mazard S, Ostrowski M, Partensky F, Scanlan DJ. (2012). Multi-locus sequence analysis, taxonomic resolution and biogeography of marine *Synechococcus*. *Environ Microbiol* **14**:372–386.

Mella-Flores D, Mazard S, Humily F, Partensky F, Mahé F, Bariat L, *et al.* (2011). Is the distribution of *Prochlorococcus* and *Synechococcus* ecotypes in the Mediterranean Sea affected by global warming? *Biogeosciences Discuss* **8**:4281–4330.

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

Moore LR, Post AF, Rocap G, Chisholm SW. (2002). Utilization of different nitrogen sources by the marine cyanobacteria *Prochlorococcus* and *Synechococcus*. *Limnol Oceanogr* **47**:989–996.

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

Needham DM, Sachdeva R, Fuhrman JA. (2017). Ecological dynamics and co-occurrence among marine phytoplankton, bacteria and myoviruses shows microdiversity matters. *ISME J* 1–16.

Oksanen J, Blanchet F, Kindt R, Legendre P, Minchin P, O'Hara R, *et al.* (2013). vegan: Community Ecology Package. R package version 2.0-10. *R Packag version* **1**. http://cran.r-project.org.

Palenik B. (1994). Cyanobacterial Community Structure as Seen from Rna-Polymerase Gene Sequence-Analysis. *Appl Environ Microbiol* **60**:3212–3219.

Paradis E, Claude J, Strimmer K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**:289–290.

Partensky F, Hess WR, Vaulot D. (1999). *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *MicrobiolMol BiolRev* **63**:106–127.

Pianka ER. (1973). The structure of lizard communities. *Annu Rev Ecol Syst* **4**:53–74.

Post AF, Penno S, Zandbank K, Paytan A, Huse SM, Welch DM. (2011). Long term seasonal dynamics of *Synechococcus* population structure in the Gulf of Aqaba, northern Red Sea. *Front Microbiol* **2**:1–12.

Rusch DB, Martiny AC, Dupont CL, Halpern AL, Venter JC. (2010). Characterization of *Prochlorococcus* clades from iron-depleted oceanic regions. *Proc Natl Acad Sci U S A* **107**:16184–16189.

Scanlan DJ, Ostrowski M, Mazard S, Dufresne A, Garczarek L, Hess WR, *et al.* (2009). Ecological genomics of marine picocyanobacteria. *Microbiol Mol Biol Rev* **73**:249–299.

Schlitzer R. (2002). Interactive analysis and visualization of geoscience data with Ocean Data View. *Comput Geosci* **28**:1211–1218.

Sohm JA, Ahlgren NA, Thomson ZJ, Williams C, Moffett JW, Saito MA, *et al.* (2016). Co-occurring *Synechococcus* ecotypes occupy four major oceanic regimes

defined by temperature, macronutrients and iron. *ISME J* **10**:333–345.

Tai V, Palenik B. (2009). Temporal variation of *Synechococcus* clades at a coastal Pacific Ocean monitoring site. *Isme J* **3**:903–915.

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.

Wu J, Sunda W, Boyle E a, Karl DM. (2000). Phosphate depletion in the western North Atlantic Ocean. *Science* **289**:759–762.

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

Zwirglmaier K, Heywood JL, Chamberlain K, Woodward EMS, Zubkov M V., Scanlan DJ. (2007). Basin-scale distribution patterns of picocyanobacterial lineages in the Atlantic Ocean. *Environ Microbiol* **9**:1278–1290.

Zwirglmaier K, Jardillier L, Ostrowski M, Mazard S, Garczarek L, Vaulot D, *et al.* (2008). Global phylogeography of marine *Synechococcus* and *Prochlorococcus* reveals a distinct partitioning of lineages among oceanic biomes. *Environ Microbiol* **10**:147–161.

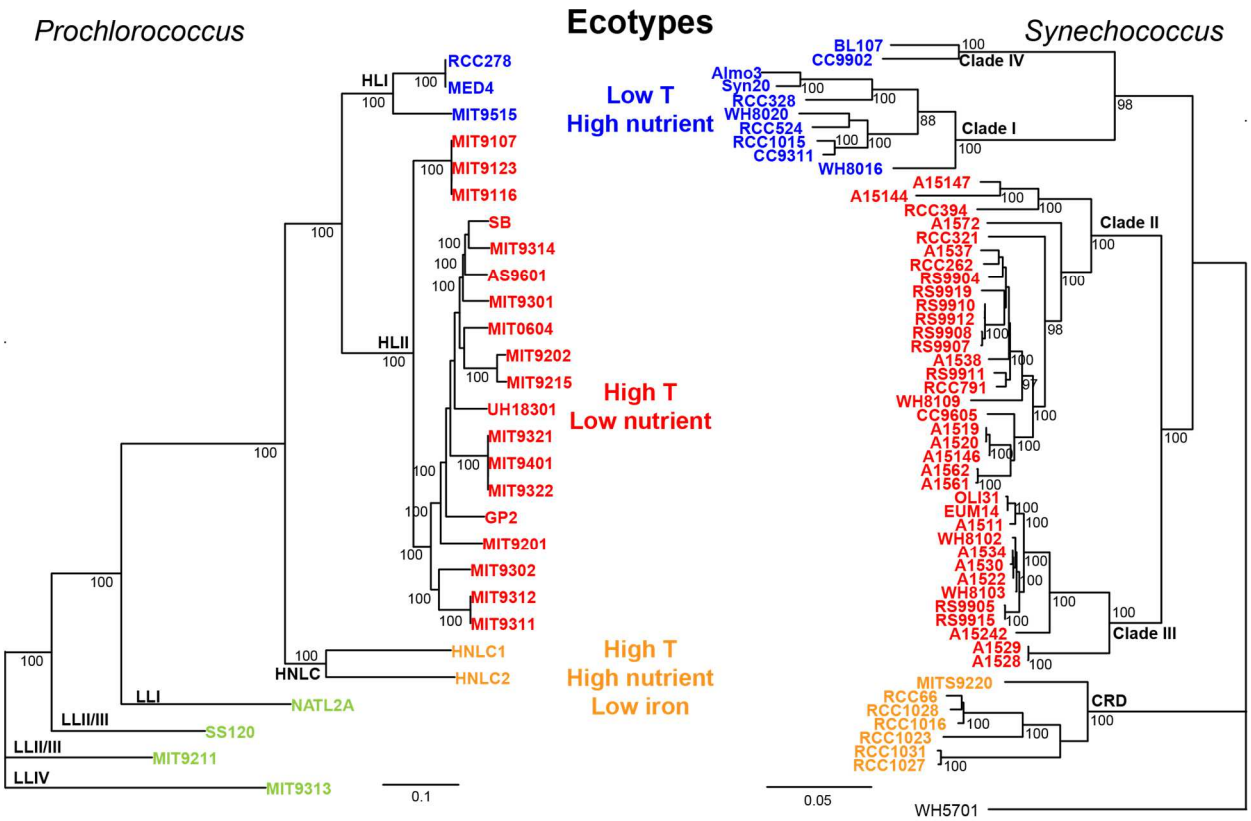


Figure 2.1. Phylogenetic tree comparison of *Prochlorococcus* and *Synechococcus* dominant surface clades. Maximum likelihood phylogenetic trees are supported with neighbor-joining bootstrap values out of 100 replications. Simplified from previous constructions (Mazard *et al.*, 2012; Berube *et al.*, 2015). Strains are colored based on their clade and clade names are at the base of each group. Blue, red and orange clades from each lineage significantly overlap with their lineage counterparts (HLI and Clades I+IV; HLII and Clades II+III; HNLC and CRD1) in niche breadth (Pianka's index, P<0.01, Figure S2.4).

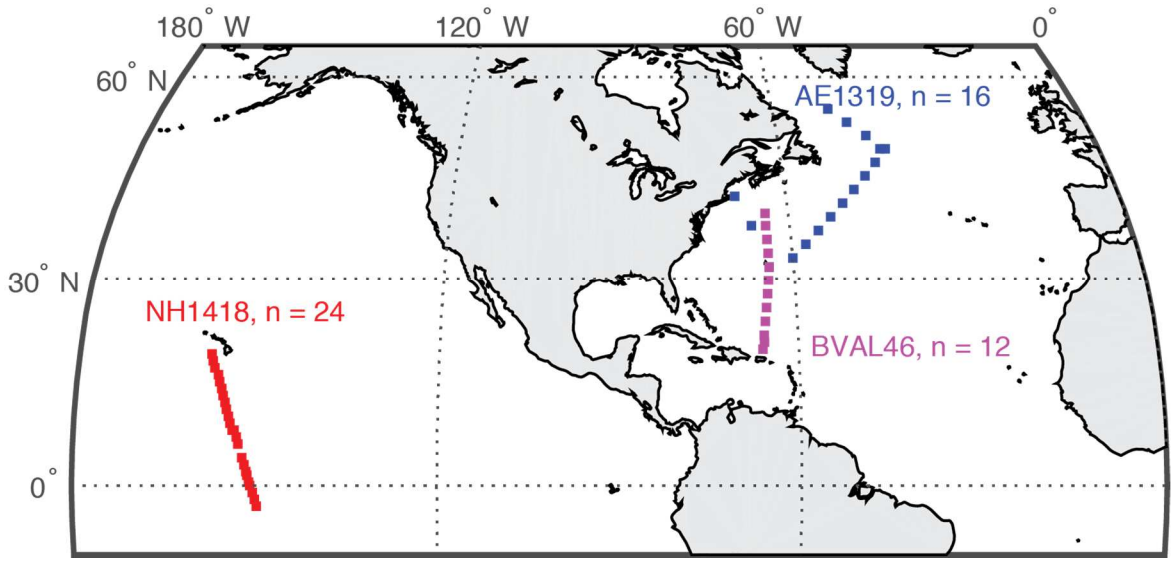


Figure 2.2. Map of Pacific Ocean and North Atlantic Ocean cruise transects. The number of stations in each cruise is shown: n=24, n=12, and n=16.

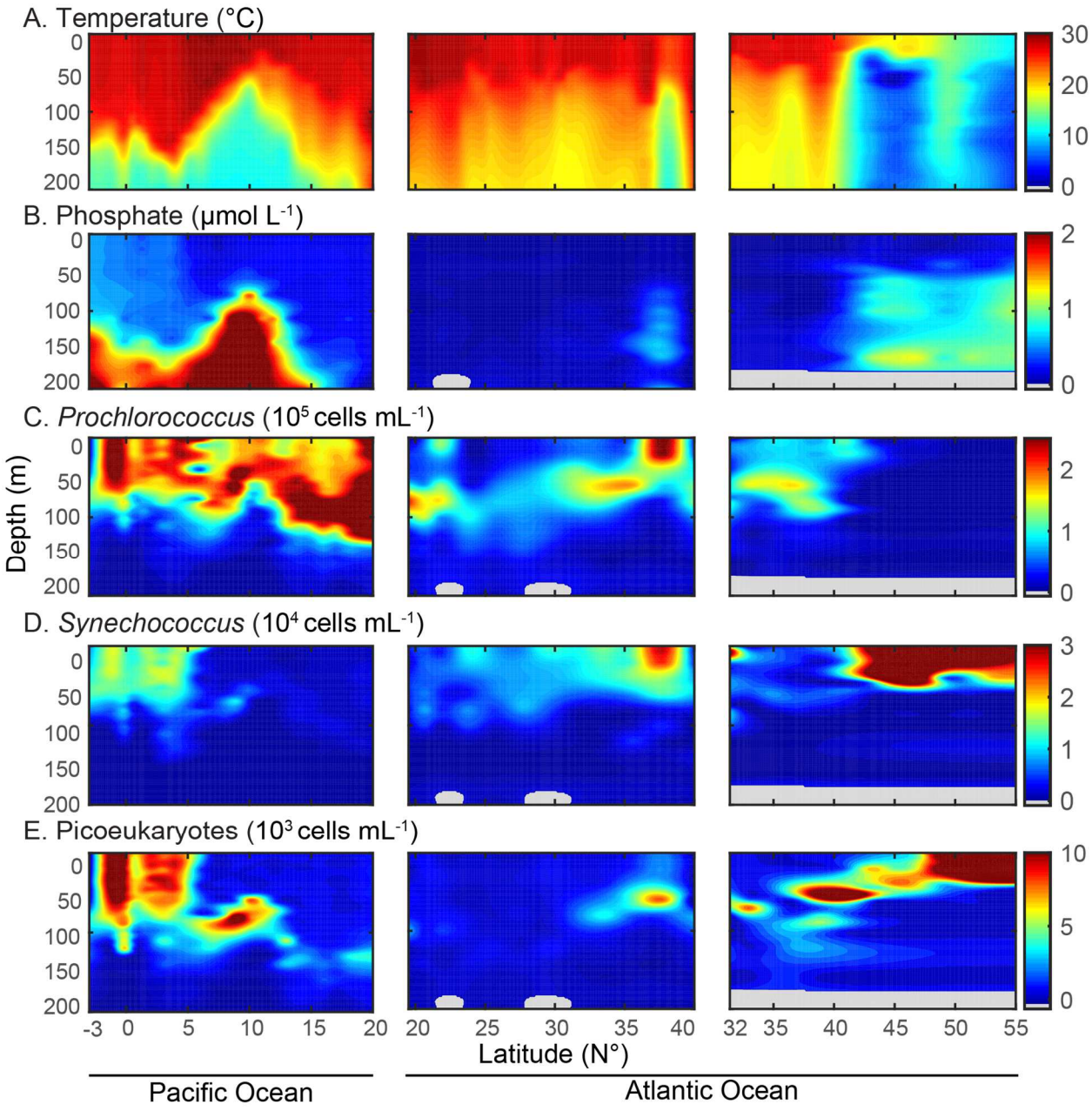


Figure 2.3. Environmental and major lineage variation across Pacific and Atlantic Ocean transects. A) Temperature profiles (˚C) across three cruise transects in the Pacific and Atlantic Oceans. B) Soluble reactive phosphate profiles concentrations in umol/L. C) Flow cytometry counts for *Prochlorococcus.* D) Flow cytometry counts for *Synechococcus.* E) Flow cytometry counts for picoeukaryotes with white points representing sampling locations.

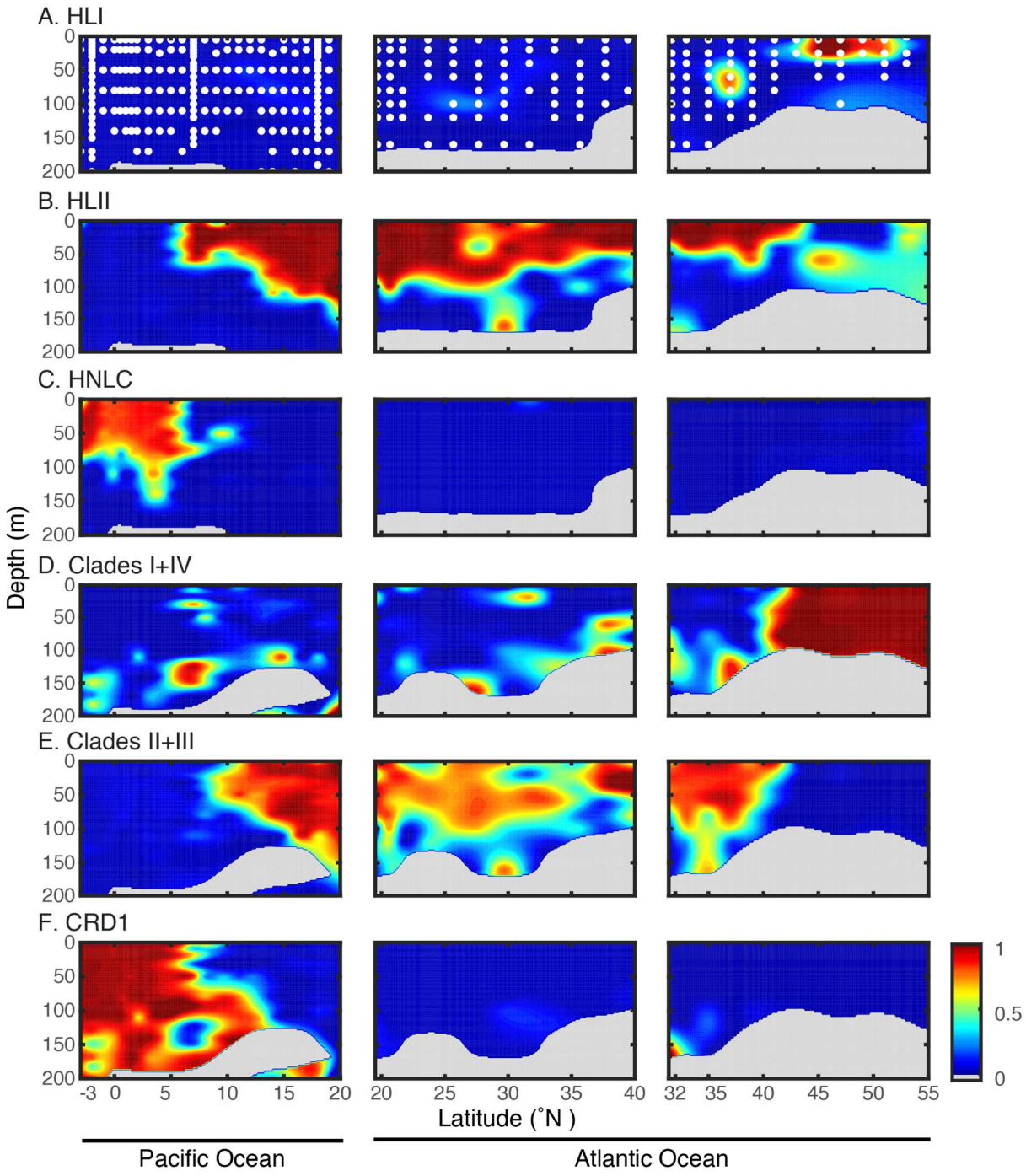


Figure 2.4. Relative clade distribution across ocean transects. *Prochlorococcus* Clades HLI (A), HLII (B), and HNLC (C) with *Synechococcus* Clades I+IV (D), II+III (E), and CRD1 (F) across three cruise transects. Clade abundance normalized to either *Prochlorococcus* or *Synechococcus* total number of sequences per sample.

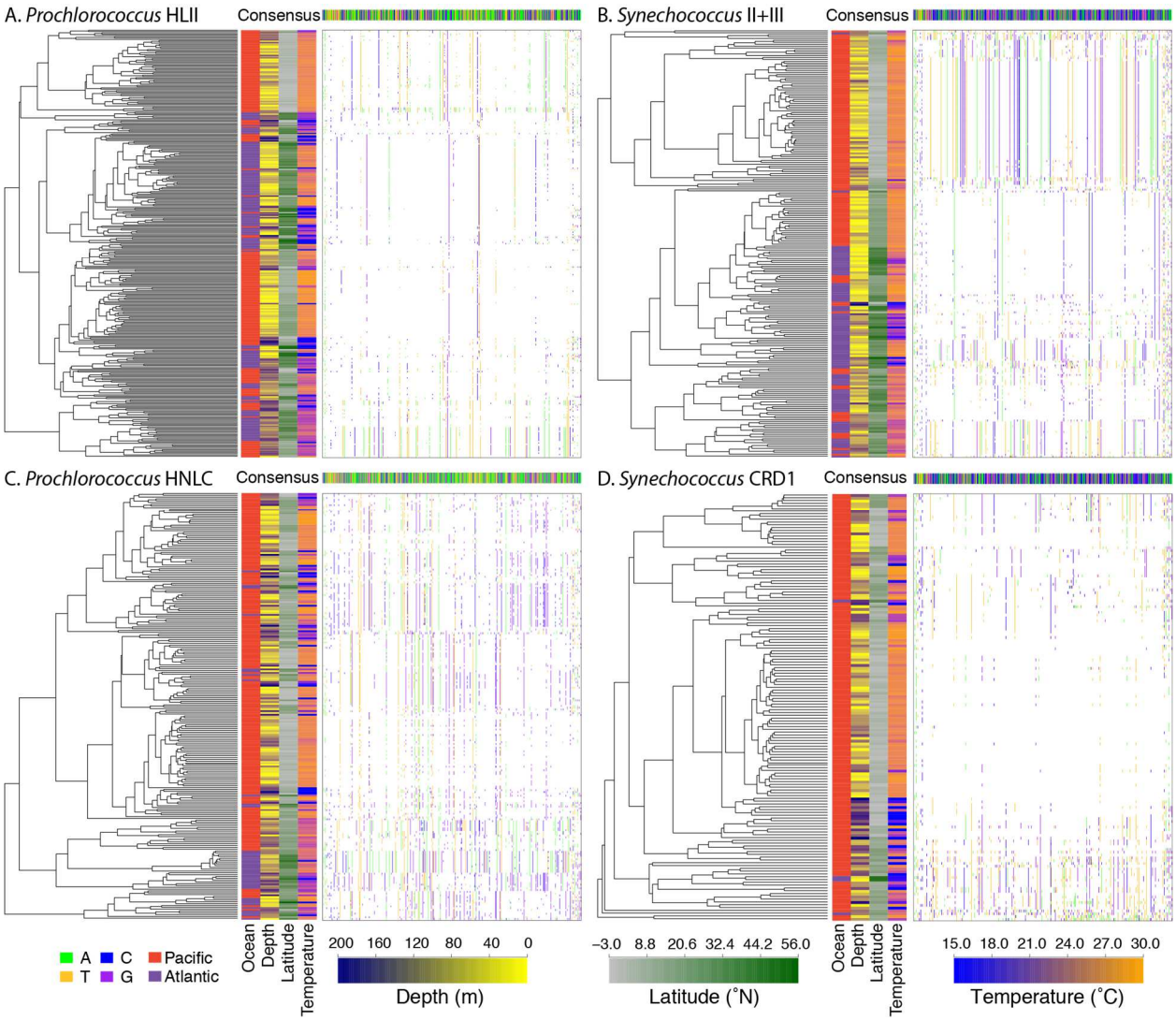
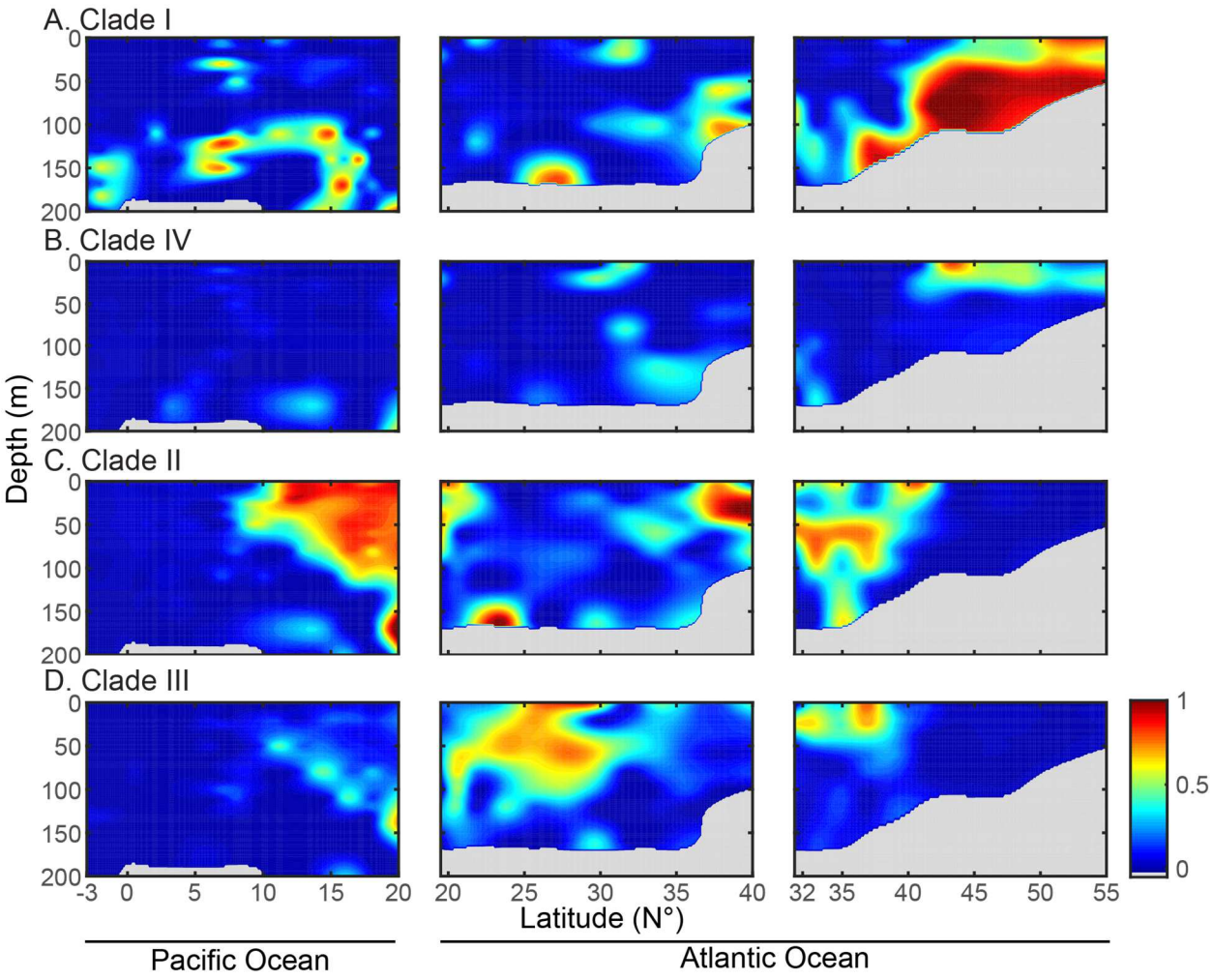


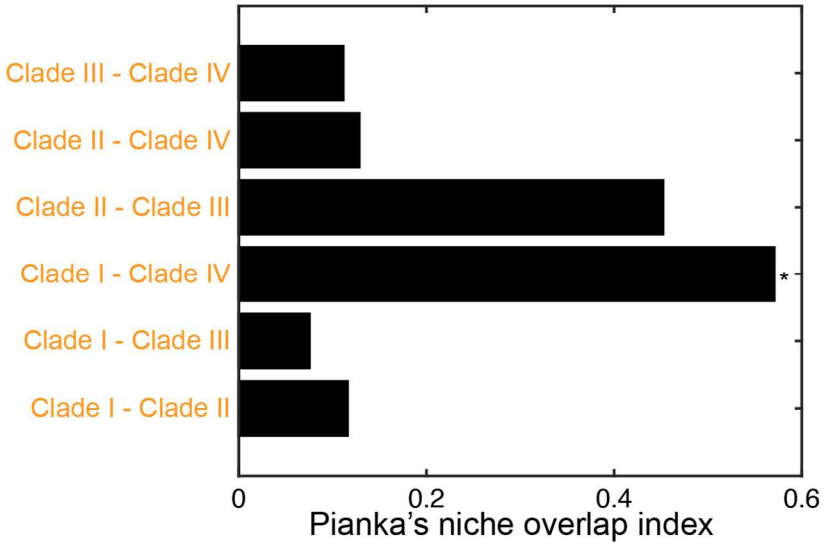
Figure 2.5. Microdiversity profile variation of abundant of clades. Clades A) HLII, B) II+III, C) HNLC, and D) CRD1. Each row represents a sample. Rows were clustered by Unifrac distance of subsampled sequences. Side columns colored by ocean origin, depth, latitude, and temperature. Rows are the SNP profiles for each sample with sequence sites that differ from the consensus of all SNP profiles colored coded by nucleotide.

Table 2.1. Parallel phylogeography of microdiversity in each clade. Unifrac's phylogenetic composition metric computed in each clade and compared across clades. Upper right triangle represents Mantel's statistic r with a Pearson correlation and permutation with P=0.001 for all comparisons.

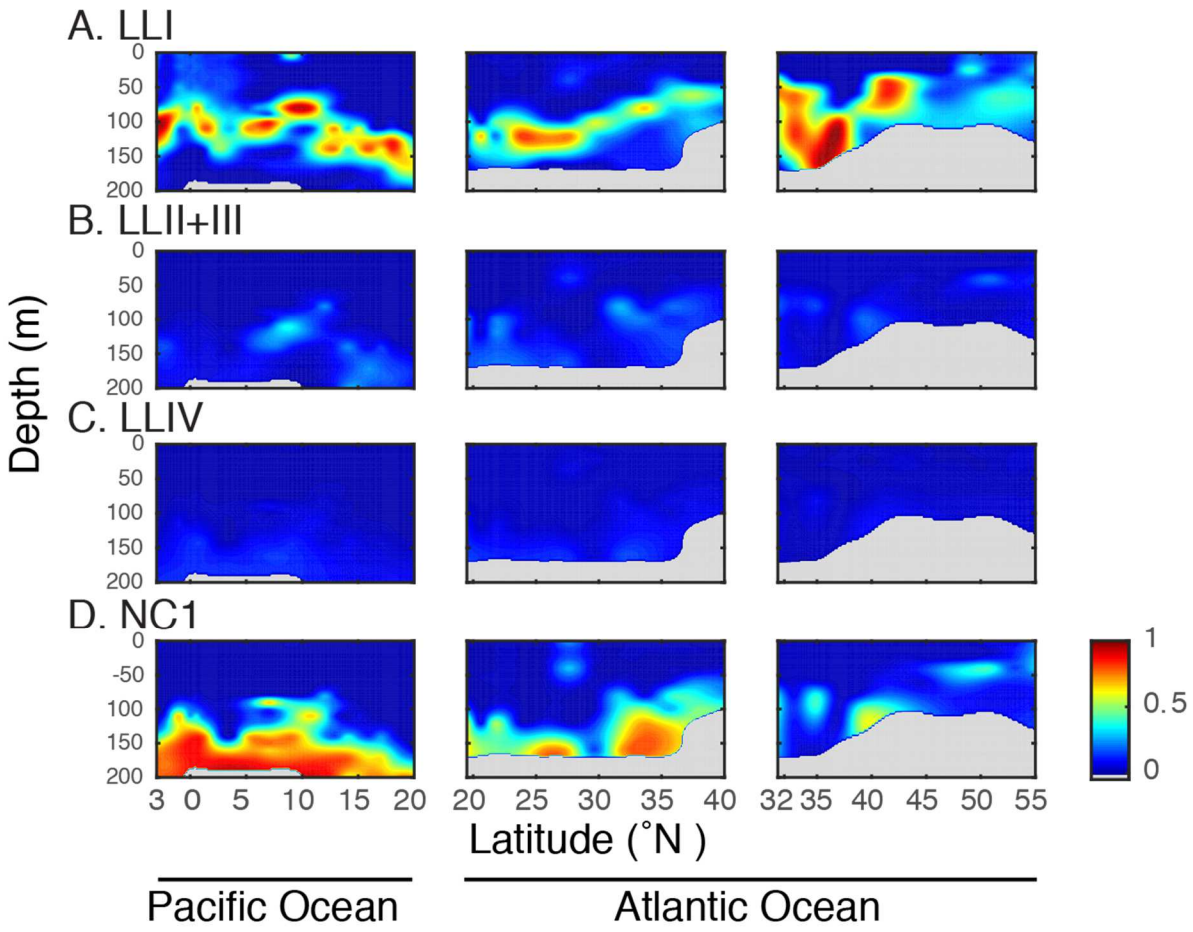
	HNLC	CRD1	Clades II+III
HLII	0.17	0.33	0.61
HNLC	-	0.60	0.27
CRD1	-	-	0.61



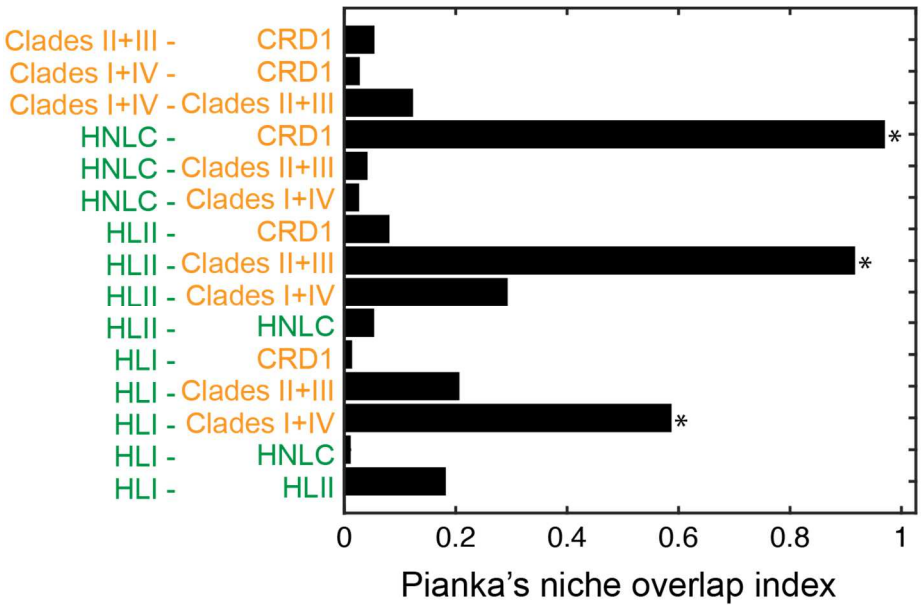
Supplementary Figure S2.1. Relative abundance of *Synechococcus* clades before aggregation. A) Clade I, B) Clade IV, C) Clade II, and D) Clade III across three ocean transects in the Pacific and Atlantic Oceans. Clades normalized by total *Synechococcus* sequences in each sample.



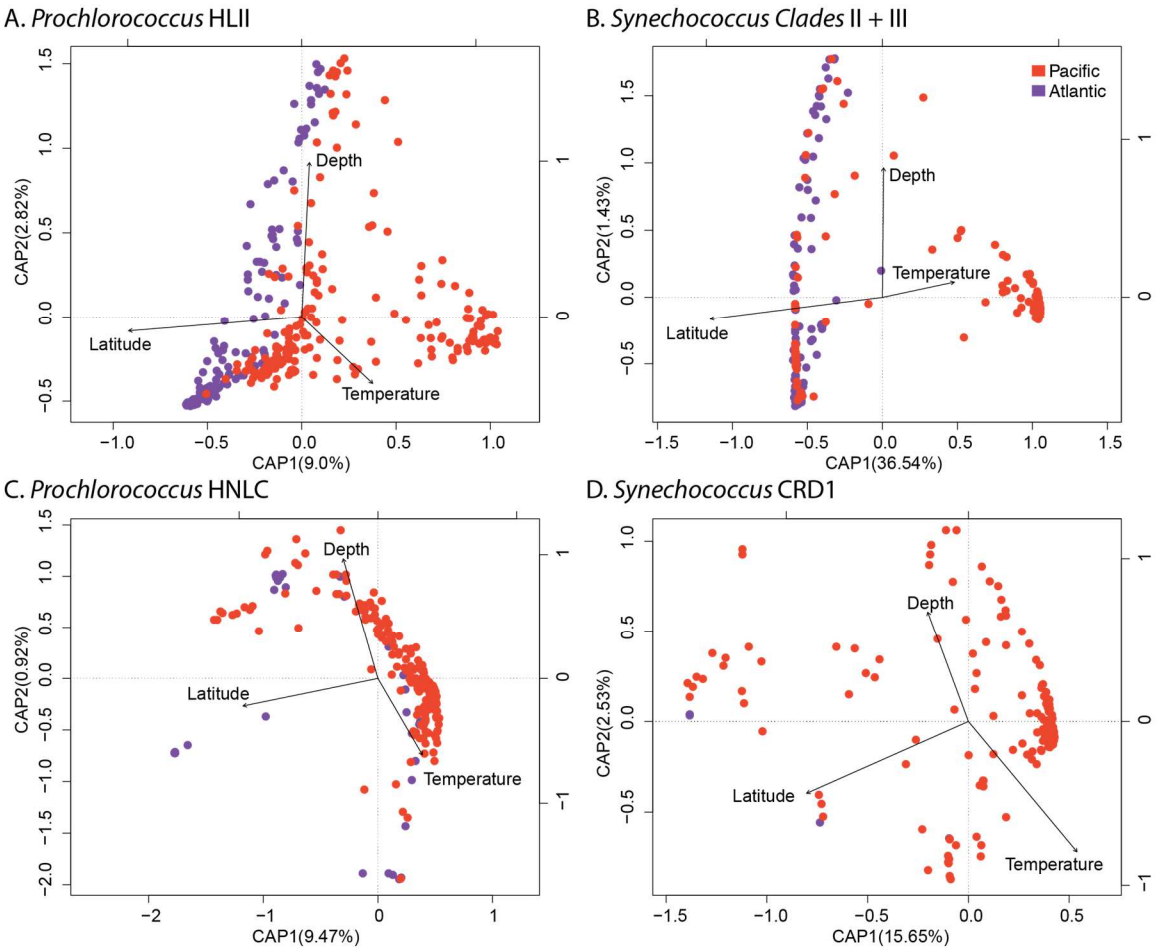
Supplementary Figure S2.2. Niche overlap of *Synechococcus* clades. A value of 0 represents no overlap and a value of 1 signifies complete overlap. Significance values (*: P<0.01) come from testing observations against a null model of latitude randomization.



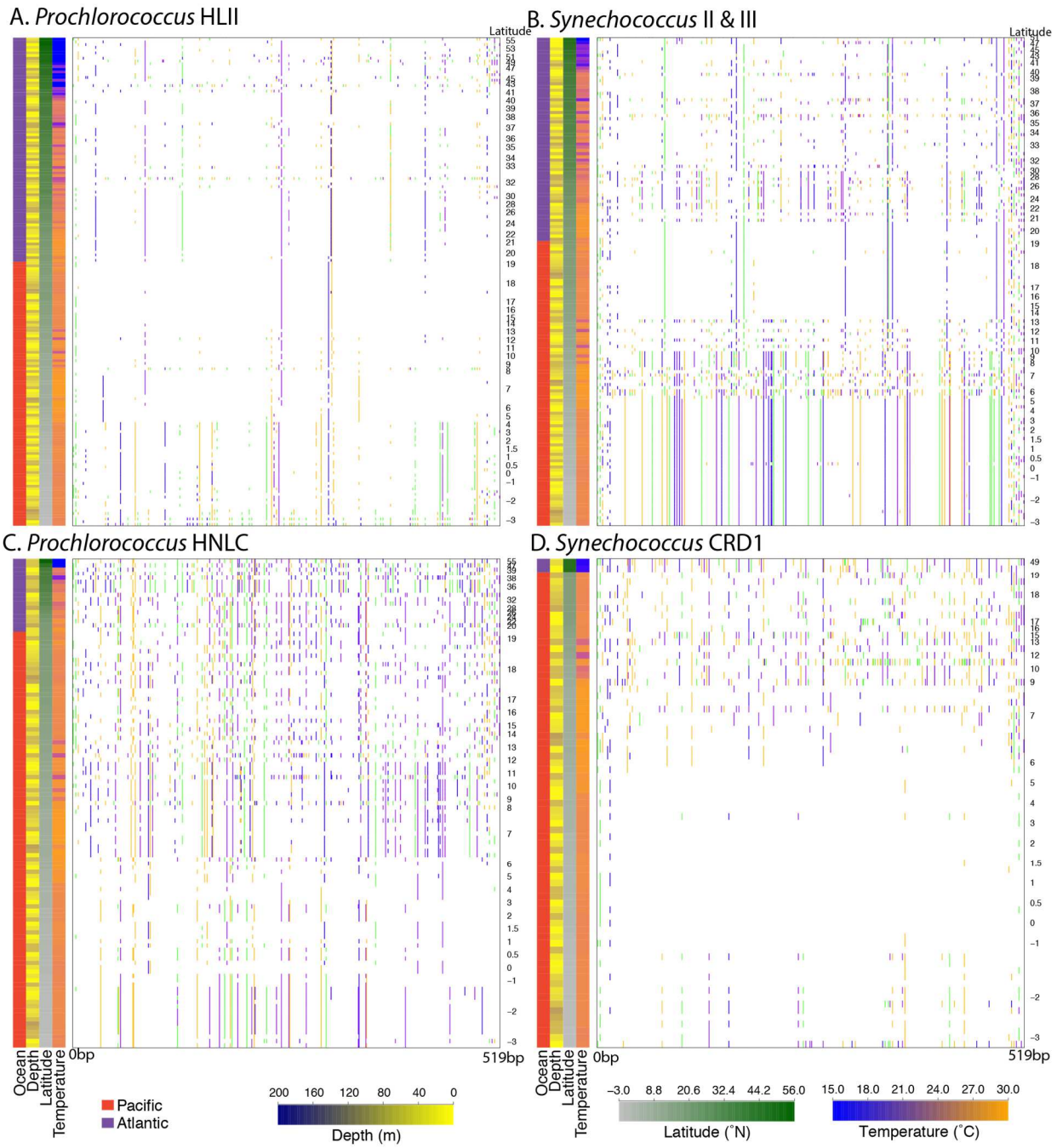
Supplementary Figure S2.3. Relative abundance of low-light *Prochlorococcus* clades. Clades normalized by total *Prochlorococcus* sequences in each sample. A) *Prochlorococcus* LL1, B) LLII+III, C) LLIV, and D) NC1 across three ocean transects in the Pacific and Atlantic Oceans.



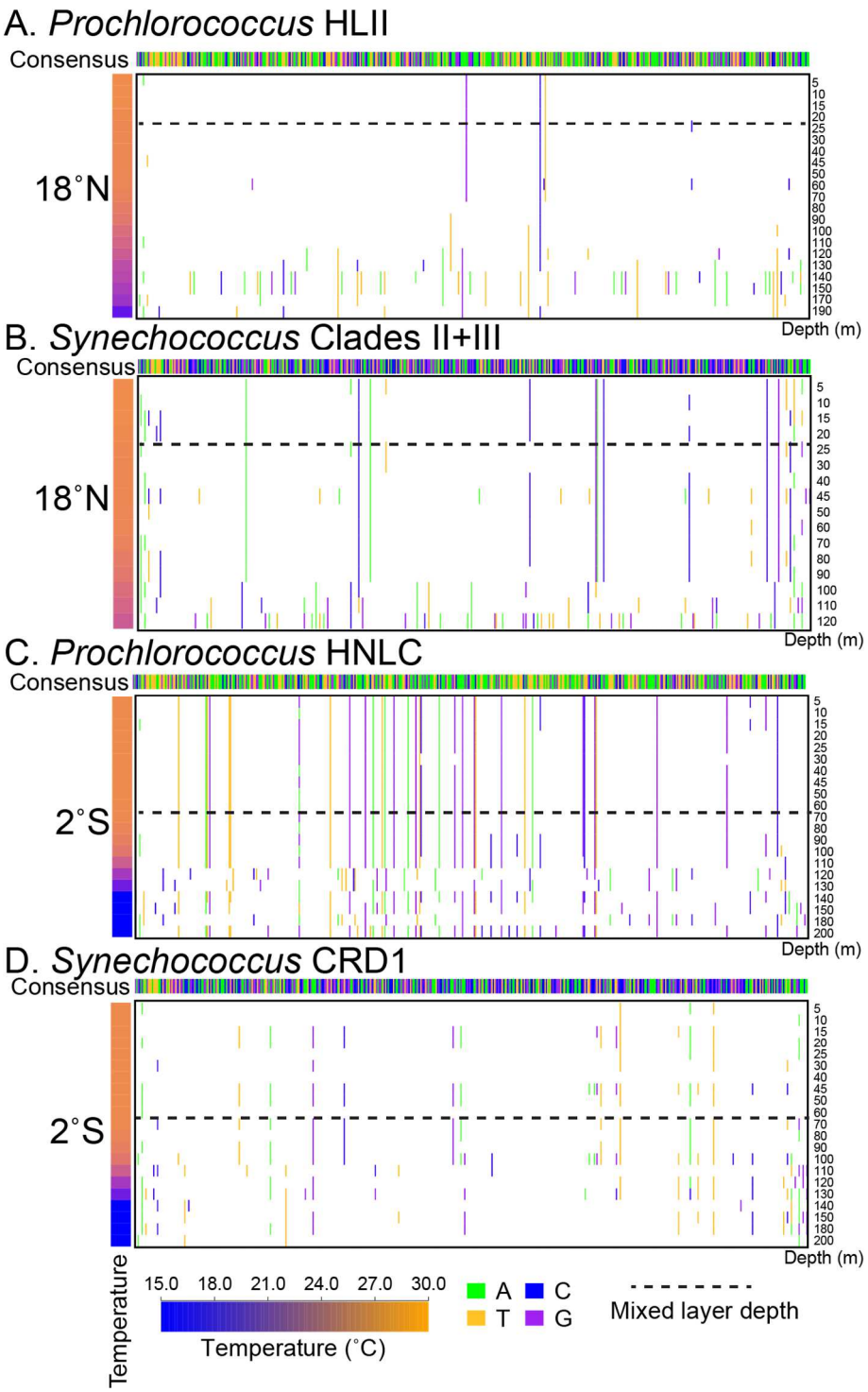
Supplementary Figure S2.4. Niche overlap of *Prochlorococcus* and *Synechococcus* clades. A value of 0 represents no overlap and a value of 1 signifies complete overlap. Significance values (*: P<0.01) come from testing observations against a null model of latitude randomization. *Synechococcus* clades are colored in orange and *Prochlorococcus* clades are colored in green.



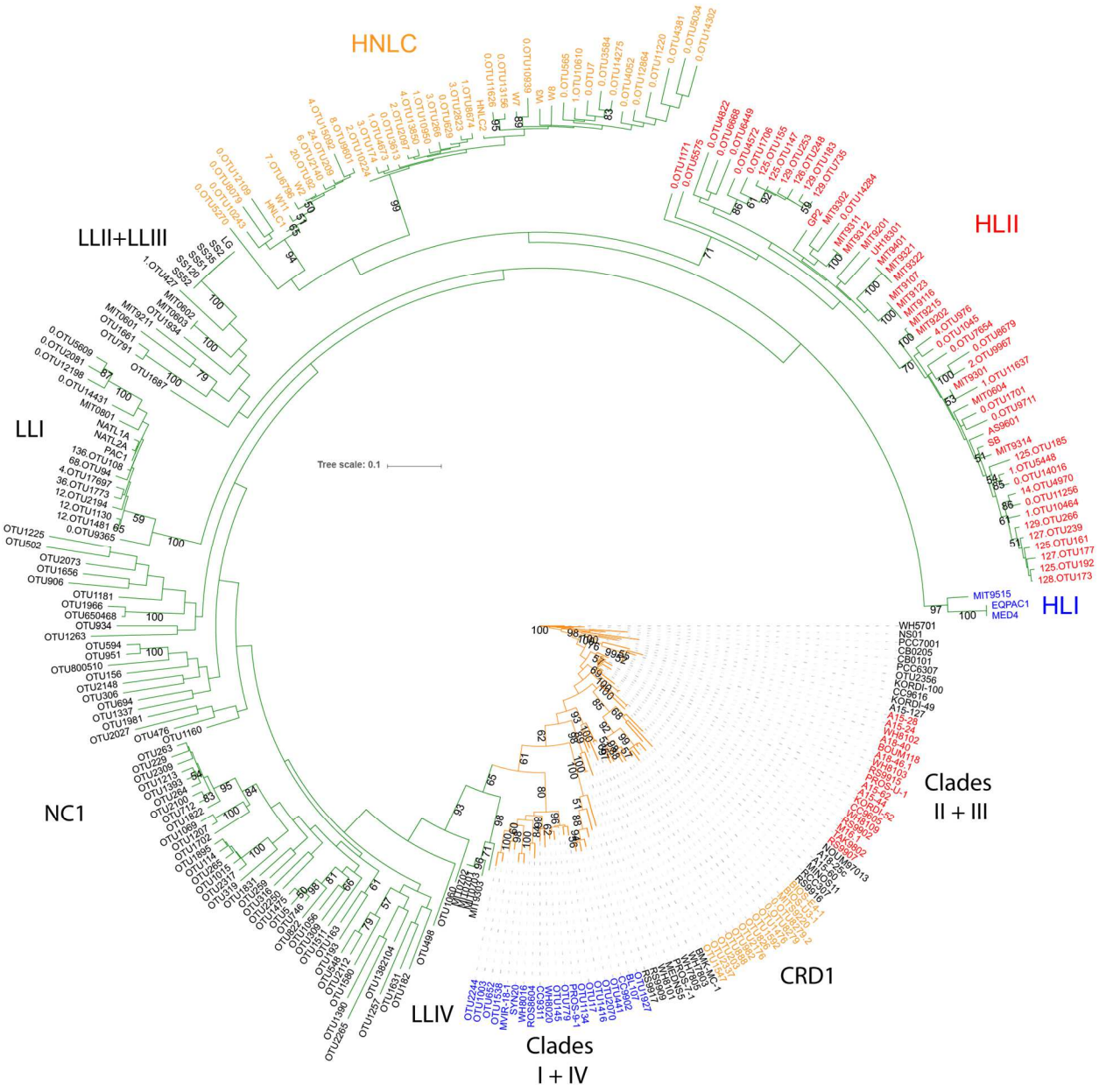
Supplementary Figure S2.5. Environmental variation explains microdiversity in each clade. Two axes shown from a distance based redundancy analysis of phylogenetic population composition constrained by latitude, longitude, and depth. Sample sites are colored by ocean of origin.



Supplementary Figure S2.6. Microdiversity is structured by latitude in the surface ocean. Consensus sequences (0-519bp) from samples in the upper euphotic zone (depth<80m) for each clade ordered by latitude. Color bars represent depth, temperature and latitude for each sample. SNP profiles highlight single nucleotide differences from overall sample consensus (Figure 2.5) for each clade.



Supplementary Figure S2.7. Microdiversity of clades is structured by depth. SNP profiles from each sample for A) HLII and B) Clades II+III at 18˚N and C) HNLC and D) CRD1 at 2˚S in the Pacific Ocean transect. Color bars represent depth and temperature for each sample. SNP profiles highlight single nucleotide differences from overall clade consensus (Figure 2.5) for each clade. Mixed layer depth for each station is marked with a dotted line, approximately 20 meters at 18˚N and 65 meters at 2˚S.



Supplementary Figure S2.8. Phylogeny of *rpo*C1 amplicon region reference sequences. The phylogenetic tree was reconstructed based on DNA sequence similarity using maximum likelihood. Bootstrap values (total 100) are calculated using maximum likelihood. Bootstrap values with >50 support are shown. Clades colored by parallel biogeography from Figure 2.1 and *Prochlorococcus* and *Synechococcus* sections of the tree are colored green and orange, respectively.

Supplementary Table S2.1. Cruise samples and associated environmental data. Depth measured in meters and temperature in ˚C. Sequences retained are the number of sequences in each sample after 90% identity and 90% coverage threshold. HLII, HNLC, Clades II+III, and CRD1 columns with 1=rarefied at 100 sequences, 0=included in analysis but <100 sequences, - represents no sequences at >97% identity.

ID	Cruise	Latitud	Longitud	Dept	Temperatur	Sequence	HLI	HNL	Clade	CRD
A00	AE131	55	-49	1	10.38	16878	0	-	-	-
A00	AE132	55	-49	25	9.99	3208	0	-	-	-
A01	AE132	55	-49	40	7.97	1629	0	0	-	-
A02	AE132	55	-49	120	4.19	1207	0	-	0	-
A04	AE132	53	-46	1	11.77	5181	0	-	-	-
A04	AE132	53	-46	25	11.69	6253	0	-	-	-
A04	AE132	53	-46	40	8.53	5413	0	-	-	-
A08	AE132	51	-43	5	13.35	16835	0	-	0	-
A08	AE132	51	-43	25	13.09	11121	-	-	-	-
A08	AE132	51	-43	40	11.25	5998	0	-	-	-
A08	AE132	51	-43	60	9.1	2077	0	-	-	-
A10	AE133	49	-40	5	15.58	6232	-	-	0	0
A10	AE133	49	-40	25	15.3	6652	0	-	-	0
A13	AE133	47	-42.5	5	18.34	34039	1	-	0	-
A14	AE133	47	-42.5	25	17.77	11236	0	-	0	-
A14	AE133	47	-42.5	40	9.3	99092	0	0	-	-
A14	AE133	47	-42.5	60	7.03	6823	1	-	-	-
A14	AE133	47	-42.5	100	7.23	1004	0	0	-	-
A19	AE133	45	-45	5	18.52	19678	0	-	0	-
A19	AE133	45	-45	25	16.56	12750	-	-	0	-
A19	AE133	45	-45	40	7.99	3182	0	-	-	-
A19	AE134	45	-45	60	0.71	1042	0	-	-	-
A22	AE134	43	-47.5	5	19.36	40724	1	-	0	-
A22	AE134	43	-47.5	25	10.71	7826	0	-	-	-
A24	AE134	41	-50	5	23.74	116852	1	-	1	-
A24	AE134	41	-50	25	21.16	29543	1	-	0	-
A25	AE134	41	-50	40	18.12	16437	1	-	0	-
A25	AE134	41	-50	60	16.35	40799	0	-	0	-
A25	AE134	41	-50	80	14.72	8715	0	0	0	-
A26	AE134	39	-52.5	5	26.29	40399	1	0	0	-
A27	AE134	39	-52.5	25	26.22	10027	1	-	0	-
A27	AE135	39	-52.5	40	26.21	6412	1	-	0	-
A27	AE135	39	-52.5	60	25.78	46926	1	-	0	-
A27	AE135	39	-52.5	80	23.53	82637	1	0	0	-
A27	AE135	39	-52.5	100	22.12	60497	1	0	0	-
A28	AE135	39	-52.5	120	20.81	29845	1	-	-	-
A30	AE135	37	-55	5	26.99	60795	1	-	0	-
A30	AE135	37	-55	25	26.92	16321	1	-	0	-
A30	AE135	37	-55	40	25.92	97609	1	-	0	-
A30	AE135	37	-55	60	21.14	143203	1	-	0	-
A31	AE135	37	-55	80	19.32	15198	1	-	0	-
A31	AE136	37	-55	100	18.28	67754	1	-	0	-
A31	AE136	37	-55	120	17.93	30062	0	-	0	-
A32	AE136	35	-57.5	5	26.63	30123	1	-	0	-
A32	AE136	35	-57.5	25	26.8	54091	1	-	0	-

A32	AE136	35	-57.5	40	24.85	18883	1	-	0	-
A33	AE136	35	-57.5	60	21.67	155578	1	-	0	-
A33	AE136	35	-57.5	80	20.36	43875	1	-	-	-
A33	AE136	35	-57.5	100	19.69	22097	0	-	-	-
A33	AE136	35	-57.5	120	19.29	15017	1	0	-	-
A34	AE136	35	-57.5	160	18.5	13905	0	-	0	-
A35	AE137	33	-60	5	27.51	94530	1	-	0	-
A36	AE137	33	-60	25	27.42	123988	1	-	0	-
A36	AE137	33	-60	40	24.21	34462	1	-	0	-
A36	AE137	33	-60	60	21.64	186286	1	-	0	-
A36	AE137	33	-60	80	20.39	81104	1	-	0	-
A36	AE137	33	-60	100	19.65	32602	1	-	-	-
A37	AE137	33	-60	120	19.1	6321	0	-	0	-
A37	AE137	33	-60	160	18.55	138	0	-	-	-
A42	AE137	31.67	-64.17	5	27.86	46814	0	-	1	-
A43	AE137	31.67	-64.17	25	25.42	26130	1	-	0	-
A43	AE138	31.67	-64.17	40	22.72	56555	1	-	0	-
A44	AE138	31.67	-64.17	60	21.56	75690	1	-	0	-
A44	AE138	31.67	-64.17	80	20.44	99998	1	0	-	-
A45	AE138	31.67	-64.17	100	19.96	147995	1	0	0	-
A46	AE138	31.67	-64.17	120	19.63	27738	1	0	-	-
A46	AE138	31.67	-64.17	160	19.06	6249	1	0	-	0
B02	BVAL4	33.66	-64.17	5	27.73	7095	1	-	-	-
B02	BVAL4	33.66	-64.17	20	27.6	13391	1	-	0	-
B02	BVAL4	33.66	-64.17	60	26.07	16792	1	-	0	-
B03	BVAL4	33.66	-64.17	80	23.58	164725	1	0	0	-
B03	BVAL4	33.66	-64.17	100	22	54646	1	-	-	-
B03	BVAL4	33.66	-64.17	120	21.4	96283	0	-	-	-
B06	BVAL4	35.67	-64.17	5	27.39	2685	1	-	0	-
B06	BVAL4	35.67	-64.17	40	26.47	23139	1	0	0	-
B06	BVAL4	35.67	-64.17	60	25.76	3717	1	0	-	-
B06	BVAL4	35.67	-64.17	80	24.49	10478	1	0	-	-
B07	BVAL4	35.67	-64.17	100	22.87	22532	1	-	0	-
B07	BVAL4	35.67	-64.17	120	21.85	9329	1	-	-	-
B07	BVAL4	35.67	-64.17	160	20.28	18602	0	-	-	-
B09	BVAL4	37.66	-64.16	5	24.97	17493	1	-	0	-
B10	BVAL4	37.66	-64.16	20	24.98	64400	1	0	1	-
B10	BVAL4	37.66	-64.16	40	24.23	26152	1	0	0	-
B10	BVAL4	37.66	-64.16	60	17.69	4532	1	0	0	-
B11	BVAL4	37.66	-64.16	100	15.9	5029	0	-	-	-
B13	BVAL4	39.45	-64.15	80	24.16	24759	1	0	0	-
B15	BVAL4	39.45	-64.15	5	26.49	9838	1	-	0	-
B15	BVAL4	39.45	-64.15	80	24.16	21489	1	0	0	-
B19	BVAL4	31.67	-64.17	5	27.45	63	0	0	-	-
B19	BVAL4	31.67	-64.17	20	27.45	2038	1	0	0	-
B19	BVAL4	31.67	-64.17	40	24.51	14536	1	0	0	-
B20	BVAL4	31.67	-64.17	80	20.75	5344	0	-	-	-
B20	BVAL4	31.67	-64.17	160	18.53	5958	1	0	-	-
B24	BVAL4	29.67	-64.47	5	26.64	4640	1	-	0	-
B25	BVAL4	29.67	-64.47	20	26.66	113	0	-	-	-
B25	BVAL4	29.67	-64.47	40	25.95	5787	1	-	-	-
B25	BVAL4	29.67	-64.47	60	24.92	50551	1	-	0	-

B25	BVAL4	29.67	-64.47	80	22.84	25989	1	-	0	-
B25	BVAL4	29.67	-64.47	100	21.11	9168	1	-	-	-
B26	BVAL4	29.67	-64.47	120	20.14	97	0	-	-	-
B26	BVAL4	29.67	-64.47	160	18.97	23887	1	-	0	-
B30	BVAL4	27.67	-64.77	5	26.77	9946	-	-	-	-
B31	BVAL4	27.67	-64.77	40	26.77	8667	1	-	0	-
B31	BVAL4	27.67	-64.77	60	25.58	11590	1	0	0	-
B31	BVAL4	27.67	-64.77	80	24.84	17160	1	-	0	-
B31	BVAL4	27.67	-64.77	100	22.72	60447	1	0	0	-
B31	BVAL4	27.67	-64.77	120	21.48	87022	1	-	0	-
B32	BVAL4	27.67	-64.77	160	20.4	42404	1	-	-	-
B34	BVAL4	25.67	-65.07	5	27.24	9269	1	-	0	-
B34	BVAL4	25.67	-65.07	20	27.24	40538	1	0	0	-
B34	BVAL4	25.67	-65.07	40	27.25	14609	1	-	0	-
B35	BVAL4	25.67	-65.07	60	26.63	97491	1	-	0	-
B35	BVAL4	25.67	-65.07	100	22.16	118838	1	-	0	-
B35	BVAL4	25.67	-65.07	120	20.79	188407	1	0	0	-
B35	BVAL4	25.67	-65.07	160	19.71	54949	0	0	-	-
B38	BVAL4	23.67	-65.37	5	28.08	1911	1	-	0	-
B38	BVAL4	23.67	-65.37	20	28.1	8409	1	-	0	-
B38	BVAL4	23.67	-65.37	40	26.4	22221	0	0	0	-
B38	BVAL4	23.67	-65.37	60	25.73	40434	1	-	0	-
B39	BVAL4	23.67	-65.37	120	21.75	19290	1	-	-	-
B39	BVAL4	23.67	-65.37	160	20.26	20979	1	0	-	-
B42	BVAL4	21.67	-65.67	5	28.42	6668	1	-	-	-
B42	BVAL4	21.67	-65.67	20	28.42	47662	1	-	0	-
B42	BVAL4	21.67	-65.67	40	28.39	14032	1	0	0	-
B42	BVAL4	21.67	-65.67	80	26	25449	1	-	0	-
B43	BVAL4	21.67	-65.67	100	25.13	43700	1	0	-	-
B43	BVAL4	21.67	-65.67	120	24.37	53906	1	-	-	-
B47	BVAL4	20.67	-65.82	5	28.97	89160	1	-	1	-
B47	BVAL4	20.67	-65.82	20	28.86	161347	1	-	1	-
B47	BVAL4	20.67	-65.82	40	28.83	65010	1	-	0	-
B48	BVAL4	20.67	-65.82	60	28.82	34539	1	-	0	-
B48	BVAL4	20.67	-65.82	80	26.04	17419	1	-	0	-
B48	BVAL4	20.67	-65.82	100	24.38	68604	1	-	0	-
B48	BVAL4	20.67	-65.82	120	23.07	188942	1	-	0	-
B48	BVAL4	20.67	-65.82	160	21.4	184807	1	0	-	-
B51	BVAL4	19.67	-65.97	5	28.71	92648	1	-	0	-
B51	BVAL4	19.67	-65.97	20	28.72	59015	1	0	0	-
B51	BVAL4	19.67	-65.97	40	28.67	102389	1	-	0	-
B51	BVAL4	19.67	-65.97	60	27.13	167914	1	0	0	-
B52	BVAL4	19.67	-65.97	80	26.26	65457	1	1	0	-
B52	BVAL4	19.67	-65.97	100	25.36	93712	1	0	0	0
B52	BVAL4	19.67	-65.97	120	24.24	139464	1	1	-	-
B52	BVAL4	19.67	-65.97	160	21.94	136260	1	0	-	-
N00	NH141	19	-158	5	28.46	29858	1	0	0	-
N00	NH141	19	-158	20	27.58	53491	1	0	0	-
N00	NH141	19	-158	50	27.31	42671	1	0	0	0
N00	NH141	19	-158	80	26.83	24670	1	0	0	-
N00	NH141	19	-158	110	26.48	45115	1	0	0	0
N01	NH141	19	-158	140	23.81	22309	1	0	0	-

N01	NH141	19	-158	170	21.4	27298	1	0	-	-
N01	NH141	19	-158	200	20.3	10678	0	0	-	-
N01	NH141	18	-157.66	5	27.62	18488	1	0	0	0
N01	NH141	18	-157.66	10	27.62	44494	1	0	1	-
N02	NH141	18	-157.66	15	27.62	19156	1	0	0	-
N02	NH141	18	-157.66	20	27.39	97988	1	0	1	0
N02	NH141	18	-157.66	25	27.27	61979	1	0	0	-
N02	NH141	18	-157.66	30	27.23	137062	1	0	1	0
N02	NH141	18	-157.66	40	27.2	45581	1	0	1	0
N03	NH141	18	-157.66	45	27.17	3219	1	0	0	0
N03	NH141	18	-157.66	50	27.13	44028	1	0	1	-
N03	NH141	18	-157.66	60	26.99	42808	1	0	0	0
N03	NH141	18	-157.66	70	26.63	53502	1	0	1	-
N03	NH141	18	-157.66	80	26.11	18473	1	0	0	0
N04	NH141	18	-157.66	90	25.63	56391	1	1	1	0
N04	NH141	18	-157.66	100	24.55	41685	1	1	0	0
N04	NH141	18	-157.66	110	23.82	18477	1	1	0	-
N04	NH141	18	-157.66	120	23.03	30577	1	0	0	0
N04	NH141	18	-157.66	130	21.66	14230	1	0	-	0
N05	NH141	18	-157.66	140	21.06	24158	1	0	-	-
N05	NH141	18	-157.66	150	20.05	35778	1	0	-	0
N05	NH141	18	-157.66	170	19.16	20274	0	0	-	-
N06	NH141	18	-157.66	190	16.52	17708	1	0	-	0
N07	NH141	13	-155.57	5	28.15	40783	1	0	0	0
N07	NH141	13	-155.57	20	28.13	25136	1	0	0	-
N07	NH141	13	-155.57	50	24.32	20344	1	0	0	0
N07	NH141	13	-155.57	80	21.64	42360	1	1	0	0
N08	NH141	13	-155.57	110	18.56	53510	1	0	-	0
N08	NH141	13	-155.57	140	13.39	79777	1	0	-	0
N08	NH141	13	-155.57	200	11.31	8776	0	0	-	0
N09	NH141	12	-155.21	5	28.39	44320	1	0	0	-
N10	NH141	12	-155.21	20	28.39	16684	1	0	0	-
N10	NH141	12	-155.22	50	22.68	43450	1	0	0	0
N10	NH141	12	-155.22	80	20.37	41817	0	0	-	0
N11	NH141	11	-154.87	5	28.34	25718	1	0	0	-
N11	NH141	11	-154.87	20	27.24	53051	1	0	0	-
N11	NH141	11	-154.87	50	22.96	37382	1	0	1	-
N12	NH141	11	-154.87	80	19.8	67614	1	0	-	0
N12	NH141	11	-154.87	110	14.89	27592	0	0	-	-
N13	NH141	10	-154.52	5	28.57	46036	1	0	0	-
N13	NH141	10	-154.52	20	28.48	24009	1	0	0	0
N13	NH141	10	-154.52	50	25.02	126635	1	1	1	1
N13	NH141	10	-154.52	80	14.33	30528	1	0	-	0
N13	NH141	10	-154.52	110	12.25	2980	0	0	-	-
N14	NH141	6	-152.78	5	29	34447	1	1	0	0
N14	NH141	6	-152.78	20	29.03	16083	1	1	0	0
N15	NH141	6	-152.78	50	29.03	11426	1	1	0	0
N15	NH141	6	-152.78	80	28.39	47848	1	1	0	1
N15	NH141	6	-152.78	110	23.19	26378	1	0	-	0
N15	NH141	6	-152.78	170	14.05	12239	0	0	-	0
N16	NH141	7	-153.13	5	29.32	25746	1	0	-	0
N16	NH141	7	-153.13	10	29.33	42040	1	1	0	0

N16	NH141	7	-153.13	15	29.28	58084	1	1	0	0
N16	NH141	7	-153.13	20	29.24	9050	1	0	-	0
N17	NH141	7	-153.13	25	29.23	44092	1	0	0	0
N17	NH141	7	-153.13	30	29.22	65568	1	1	0	-
N17	NH141	7	-153.13	40	29.21	107447	1	1	0	0
N17	NH141	7	-153.13	45	29.21	26286	1	1	-	0
N17	NH141	7	-153.13	50	29.22	14466	1	0	-	0
N18	NH141	7	-153.13	60	29.22	23716	1	1	0	0
N18	NH141	7	-153.13	70	27.92	26507	1	1	0	0
N18	NH141	7	-153.13	80	26.43	32888	1	1	0	0
N18	NH141	7	-153.13	90	22.71	81437	1	0	-	0
N18	NH141	7	-153.13	100	21.08	60330	1	1	0	0
N19	NH141	7	-153.13	110	18.8	144596	1	0	-	0
N19	NH141	7	-153.13	120	16.18	16200	1	0	-	-
N19	NH141	7	-153.13	140	13.19	14545	0	0	-	0
N19	NH141	7	-153.13	150	12.72	31263	1	0	-	0
N20	NH141	7	-153.13	160	11.9	11386	1	1	-	0
N21	NH141	5	-152.43	5	28.87	26412	1	1	0	1
N21	NH141	5	-152.43	20	28.76	15746	1	1	0	1
N21	NH141	5	-152.43	50	28.52	11907	1	1	0	1
N21	NH141	5	-152.43	80	28.28	12358	1	1	0	1
N21	NH141	5	-152.43	110	27.71	42088	1	0	0	1
N22	NH141	5	-152.43	140	20.84	9812	0	0	-	0
N22	NH141	3	-151.74	5	27.49	19639	1	1	0	1
N22	NH141	3	-151.74	20	27.49	8351	1	0	0	1
N23	NH141	3	-151.74	50	27.39	19684	1	1	0	1
N23	NH141	3	-151.74	80	27.35	20524	1	1	0	1
N23	NH141	3	-151.74	110	27.28	10306	1	1	0	1
N23	NH141	3	-151.74	140	27.02	22491	1	1	0	1
N23	NH141	3	-151.74	170	15.13	15202	0	0	-	0
N24	NH141	2	-151.39	5	27.49	34234	1	1	0	1
N24	NH141	2	-151.39	20	27.42	45380	1	1	0	1
N24	NH141	2	-151.39	50	27.35	73953	1	1	0	1
N24	NH141	2	-151.39	80	27.16	20666	1	1	0	1
N25	NH141	2	-151.39	110	26.7	8771	1	1	0	0
N25	NH141	2	-151.39	140	25.23	11451	1	1	-	0
N25	NH141	2	-151.39	170	15.64	22192	0	0	-	0
N25	NH141	0	-150.7	5	27.03	38710	1	1	0	1
N26	NH141	0	-150.7	20	27.03	18813	1	1	0	1
N26	NH141	0	-150.7	50	27.03	28586	1	1	0	1
N26	NH141	0	-150.7	80	26.69	20790	1	1	0	1
N26	NH141	0	-150.7	110	25.97	13817	1	1	0	1
N26	NH141	0	-150.7	140	21.22	4503	0	0	-	0
N27	NH141	-2	-150	5	27.49	14090	0	1	0	1
N27	NH141	-2	-150	10	27.48	29963	1	0	0	1
N27	NH141	-2	-150	15	27.47	33828	0	0	0	1
N28	NH141	-2	-150	20	27.47	16479	0	1	0	1
N28	NH141	-2	-150	25	27.47	60625	1	0	0	1
N28	NH141	-2	-150	30	27.48	17224	0	1	0	1
N28	NH141	-2	-150	40	27.48	30673	1	1	0	1
N28	NH141	-2	-150	45	27.48	14179	0	1	0	1
N29	NH141	-2	-150	50	27.48	15527	0	0	0	1

N29	NH141	-2	-150	60	27.42	18909	0	1	0	1
N29	NH141	-2	-150	70	27.3	21181	1	1	0	0
N29	NH141	-2	-150	80	26.9	14926	0	1	0	1
N29	NH141	-2	-150	90	26.43	20403	1	1	0	1
N30	NH141	-2	-150	100	25.52	23557	0	0	-	0
N30	NH141	-2	-150	110	23.14	53110	0	1	-	0
N30	NH141	-2	-150	120	19.75	21380	0	0	-	0
N30	NH141	-2	-150	130	17.03	41111	0	0	-	0
N30	NH141	-2	-150	140	14.79	17416	0	0	-	0
N31	NH141	-2	-150	150	14.1	10977	1	0	-	0
N31	NH141	-2	-150	170	13.6	11457	0	0	-	0
N31	NH141	-2	-150	180	13.23	7878	0	0	0	0
N32	NH141	-2	-150	200	13.07	3410	0	0	-	1
N32	NH141	-3	-149.67	5	27.22	43	0	0	-	0
N32	NH141	-3	-149.67	20	26.91	13857	0	1	0	1
N32	NH141	-3	-149.67	50	26.88	41367	1	1	0	1
N32	NH141	-3	-149.67	80	26.7	27018	1	1	0	1
N33	NH141	-3	-149.67	110	25.24	90900	1	1	0	1
N33	NH141	-3	-149.67	170	14.73	38616	0	0	-	0
N33	NH141	-1	-150.35	5	27.24	11584	1	1	0	1
N34	NH141	-1	-150.35	20	27.24	46921	1	0	0	1
N34	NH141	-1	-150.35	50	27.13	15140	1	1	0	1
N34	NH141	-1	-150.35	80	26.76	21568	1	1	0	1
N34	NH141	-1	-150.35	110	23.8	38955	1	1	0	0
N35	NH141	0.5	-150.87	5	27.42	13481	1	1	0	1
N35	NH141	0.5	-150.87	20	27.42	22227	1	1	0	1
N35	NH141	0.5	-150.87	50	27.31	30920	1	0	0	1
N36	NH141	0.5	-150.87	80	26.32	34166	1	1	0	1
N36	NH141	0.5	-150.87	110	24.26	16273	0	1	-	0
N37	NH141	1	-151.04	5	27.74	19994	1	1	0	1
N37	NH141	1	-151.04	20	27.75	26774	1	0	0	1
N37	NH141	1	-151.04	50	27.7	63290	1	1	0	1
N37	NH141	1	-151.04	80	27.32	18314	1	1	0	1
N37	NH141	1	-151.04	110	25.88	18333	1	1	-	0
N38	NH141	1	-151.04	140	18.02	21932	0	0	-	1
N38	NH141	1.5	-151.22	5	27.88	43128	1	0	0	1
N38	NH141	1.5	-151.22	20	27.88	32389	1	0	0	1
N39	NH141	1.5	-151.22	50	27.82	38238	1	0	0	1
N39	NH141	1.5	-151.22	80	27.4	43317	1	1	0	1
N39	NH141	1.5	-151.22	110	25.97	15572	1	1	-	0
N39	NH141	1.5	-151.22	140	18.59	3921	0	0	-	0
N40	NH141	4	-152.09	5	27.65	44562	1	1	0	1
N40	NH141	4	-152.09	25	27.69	109121	1	1	1	1
N40	NH141	4	-152.09	50	27.66	17755	1	1	0	0
N40	NH141	4	-152.09	80	27.4	32616	1	1	0	1
N41	NH141	4	-152.09	110	27.21	9159	1	1	0	1
N41	NH141	4	-152.09	140	26.2	8574	1	1	0	1
N41	NH141	8	-153.48	5	29.04	42347	1	-	-	-
N42	NH141	8	-153.48	50	28.71	62198	1	0	0	-
N42	NH141	8	-153.48	80	22.47	32083	1	1	0	0
N42	NH141	8	-153.48	110	15.3	41192	0	0	-	0
N42	NH141	8	-153.48	140	12.36	16698	0	0	-	0

N43	NH141	9	-154.17	5	28.79	54926	1	1	-	0
N43	NH141	9	-154.17	25	28.71	28060	1	0	0	-
N43	NH141	9	-154.17	50	25.85	22842	1	1	0	0
N44	NH141	9	-154.17	80	17.07	78149	1	1	0	0
N44	NH141	9	-154.17	110	12.34	11454	0	0	-	-
N44	NH141	9	-154.17	140	11.48	3904	0	0	-	0
N45	NH141	14	-155.91	25	27.74	130241	1	0	0	-
N45	NH141	14	-155.91	50	27.14	30950	1	0	0	-
N45	NH141	14	-155.91	80	23.91	41526	1	0	1	0
N45	NH141	14	-155.91	110	22.02	23996	1	0	-	0
N46	NH141	14	-155.91	140	19.98	73105	0	0	-	0
N46	NH141	14	-155.91	170	16.88	44775	0	0	-	-
N46	NH141	14	-155.91	200	12.99	42837	0	0	-	-
N46	NH141	15	-156.26	5	27.35	20718	1	0	0	-
N46	NH141	15	-156.26	20	27.35	66709	1	0	1	0
N47	NH141	15	-156.26	50	27.28	24900	1	0	0	-
N47	NH141	15	-156.26	80	24.02	34249	1	1	0	-
N47	NH141	15	-156.26	110	21.93	30688	1	0	-	-
N47	NH141	15	-156.26	140	19.97	16935	1	0	-	0
N48	NH141	16	-156.61	5	27.49	31348	1	0	0	-
N48	NH141	16	-156.61	25	27.46	27259	1	0	0	0
N48	NH141	16	-156.61	50	27.45	70568	1	0	1	-
N48	NH141	16	-156.61	80	26.09	27591	1	0	0	0
N49	NH141	16	-156.61	110	23.35	29022	1	1	0	0
N49	NH141	16	-156.61	140	21.59	25841	0	0	-	-
N49	NH141	16	-156.61	170	19.33	13048	0	0	-	-
N49	NH141	16	-156.61	200	15.47	29126	1	0	0	-
N49	NH141	17	-157.3	5	27.6	90285	1	0	1	0
N49	NH141	17	-157.3	20	27.59	43575	1	0	1	-
N50	NH141	17	-157.3	50	27.56	30369	1	0	0	-
N50	NH141	17	-157.3	80	27.43	29403	1	0	0	0
N50	NH141	17	-157.3	110	24.2	16343	1	1	0	-
N50	NH141	17	-157.3	140	21.63	13718	1	0	-	-
N50	NH141	17	-157.3	170	19.21	21970	0	0	-	-

Supplementary Table S2.2. Clade distribution does not depend on percent identity thresholds. Upper right hand triangle represents mantel's r of Euclidean distance of absolute clade abundances at each tBlastx percent identity threshold for similarity to reference sequences, lower left triangle represents mantel's r of same data using Bray-Curtis distance. All p-values from Mantel test were P=0.001 from 999 permutations, but not corrected for multiple pairwise comparisons.

	70%	**75%**	**80%**	**85%**	**90%**	**95%**	**97%**	**99%**
70%	-	1	0.9998	0.9962	0.9941	0.9743	0.891	0.5235
75%	0.9999	-	0.9999	0.9962	0.9941	0.9748	0.892	0.5256
80%	0.9998	0.9999	-	0.9962	0.9941	0.9743	0.891	0.5235
85%	0.9989	0.9991	0.9994	-	0.9988	0.9783	0.8932	0.5223
90%	0.9974	0.9977	0.9981	0.999	-	0.9801	0.8986	0.5328
95%	0.9869	0.9873	0.988	0.9898	0.9914	-	0.9656	0.6636
97%	0.9596	0.9603	0.9611	0.9629	0.9659	0.9887	-	0.8172
99%	0.8503	0.8511	0.8517	0.8518	0.8549	0.8882	0.9293	-

Supplementary Table S2.3. Differences in clade microdiversity depend on ocean of origin. Permutational multivariate analysis of variance contributed by Ocean (Atlantic or Pacific) using the Unifrac distance metric on sequences derived from four different clades. Each clade was analyzed separately.

	Df	Sum of squares	Mean squares	F	R^2	Pr>F
HLII	1	3.34	3.34	28	0.08	0.001
HNLC	1	2.85	2.85	18.61	0.07	0.001
CRD1	1	0.66	0.66	5.01	0.03	0.005
Clades II+III	1	14.36	14.36	60.34	0.21	0.001

Supplementary Table S2.4. Environmental variation explains within-clade phylogenetic composition. Distance-based redundancy analysis of environmental variables. Latitude, temperature and depth were included in the model. Unifrac distance was used for phylogenetic population composition. Each clade was analyzed separately.

	Variable	Df	Sum Sqs	F	Pr(>F)
HLII	Latitude	1	3.86	33.78	0.001
	Depth	1	1.22	10.69	0.001
	Temperature	1	0.23	2.02	0.027
	Cumulative	3	5.31	15.5	0.001
	Residual	331	37.8	-	-
HNLC	Latitude	1	3.52	23.68	0.001
	Depth	1	0.51	3.46	0.004
	Temperature	1	0.13	0.86	0.468
	Cumulative	3	4.16	9.33	0.001
	Residual	232	34.47	-	-
CRD1	Latitude	1	2.65	23.86	0.001
	Depth	1	0.47	4.21	0.001
	Temperature	1	0.76	6.82	0.001
	Cumulative	3	3.88	11.63	0.001
	Residual	149	16.56	-	-
Clades II+III	Latitude	1	23.79	127.84	0.001
	Depth	1	0.82	4.41	0.01
	Temperature	1	1.69	9.09	0.001
	Cumulative	3	26.31	47.12	0.001
	Residual	221	41.13	-	-

Supplementary Table S2.5. Within-clade microdiversity has similar structure based on different metrics. Correlations between Unifrac distance of phylogenetic composition, Euclidean distance of sample SNP profiles, and Bray-Curtis dissimilarity of OTU-based taxonomic abundances. Mantel test analyzed with Pearson correlations between each distance or dissimilarity metric. All P-values were 0.001.

	HLII	HNLC	CRD1	Clades II+III
Unifrac vs. Euclidean	0.77	0.70	0.89	0.88
Unifrac vs. Bray-Curtis	0.59	0.43	0.62	0.71
Euclidean vs. Bray-Curtis	0.53	0.39	0.61	0.73

CHAPTER 3

Marine bacterial lifestyle change due to adaptation to high temperature

Abstract

Ocean surface temperatures are rising and will continue to increase significantly over the next 100 years due to global climate change (Parry *et al.*, 2007). As temperatures rise and increase beyond current ranges, it is unclear how adaptation will impact microbial distribution and ecological role. To address this major unknown, we imposed a stressful high temperature regime and low temperature control for 500 generations on a strain from the abundant marine *Roseobacter* clade. High temperature adapted lines rapidly improved their growth rates but also adapted to the increased temperature derived selective pressure of a 12.1% decrease in oxygen solubility. We observed a significant departure from the organism's usual growth mode resulting in increased biofilm formation at the air-liquid interface. Furthermore, this altered lifestyle was coupled with a suite of genomic changes linked to biofilm formation. This response is uniquely different from *Escherichia coli* adapted to high temperature (Tenaillon *et al.*, 2012) as only 3% of mutated genes were mutated in both studies, demonstrating the importance for understanding adaptation to elevated temperature in abundant marine bacteria. Temperature had a direct effect on physiology, but we also observed that a small decrease in oxygen solubility could lead to a large difference in microbial lifestyle.

Keywords: Experimental evolution, adaptation, temperature, *Roseobacter*, biofilm, oxygen

Introduction

Climatically driven temperature shifts may extend beyond the confines of acclimation, especially in regions where organisms are already living close to their thermal limits and require adaptations to cope with thermal stress (Franks & Hoffmann, 2012; Bergmann *et al.*, 2010; Thomas *et al.*, 2012). In laboratory experiments, bacteria are capable of adapting rapidly to changes in their environment (Lenski & Bennett, 1993). Due to their short generation times and large population sizes, it is possible that they may do so within the same time frame as ocean temperature changes (Thomas *et al.*, 2012; Collins, 2010). Thus, environmental change can drive selection for different existing lineages but can also lead to selection of *de novo* beneficial mutations.

What we know about short-term temperature adaptation in bacteria is primarily through experimental evolution studies in model organisms (often human associated, like *Escherichia coli)*. Adaptation to high temperature on a short time frame in such organisms has been linked to mutations in genes involved in transcriptional and translational regulation, as small changes at the regulation level can lead to large physiological changes necessary for restoring an organism to a pre-stressed state (Rodríguez-verdugo *et al.*, 2013; Hug & Gaut, 2015). When the model organism *E. coli* evolved to high temperature in an experimental evolution study, all 115 independent lines had at least one mutation either in the RNA polymerase operon or in *rho*, a termination factor, both of which broadly regulate transcription (Tenaillon *et al.*, 2012). Changes in gene expression of heat-inducible genes have also been observed in *E. coli* evolved to high temperature (Riehle *et al.*, 2003). Other researchers have seen a negative correlation between temperature and the concentrations of cellular protein and

ribosomes in bacteria (Nielsen & Jørgensen, 1968; Toseland *et al.*, 2013) leading to a large impact on their cellular allocation and biogeochemical role (Toseland *et al.*, 2013; Galbraith & Martiny, 2015). However, the physiological outcome of adaptation to novel thermal environments is largely unknown in abundant marine bacterial lineages.

The marine *Roseobacter* clade is an abundant heterotrophic bacterium representing up to 25% of cells in marine communities depending on the habitat (Buchan & Moran, 2005; Wagner-Döbler & Biebl, 2006) and plays an important role in marine ecosystem functioning (Moran *et al.*, 2004). They proliferate in coastal and polar waters that range in temperature from 10°C to 35°C (Wagner-Döbler & Biebl, 2006; Giebel *et al.*, 2009). Members of the clade can be planktonic or associated with biofilms and often directly interact with particles or other organisms changing their ecological dynamics (Buchan & Moran, 2005). Due to the widespread range, generalist lifestyle, and importance in carbon and sulfur biogeochemical cycles, it is critical to understand how marine *Roseobacter* will adapt to changes in ocean conditions.

To investigate the capacity and ecological implications of adaptation to warmer ocean environments, we experimentally evolved the strain *Roseovarius sp.* TM1035 to high temperature. We hypothesized that growth rates would rapidly increase due to improved fitness and that adaptation to higher temperature would result in changes in growth and resource allocation strategies and consequently shifts in cellular stoichiometry. We followed this phenotypic characterization with genomic sequencing to identify the genomic basis for adaptation. Finally, to assess the generality of short-term bacterial evolutionary responses to high temperature, we compared the biochemical mechanisms for adaptation to high temperature with those found in *E. coli.*

Results and Discussion

To test our hypotheses, we evolved replicate (2 x 22) lines of *Roseovarius sp.* TM1035 to the optimal (low, 25˚C) and high (33°C) temperature for 500 generations (Figure S3.1 and S3.2). We found *Roseobacter* had the capacity to evolve rapidly to increased temperature as high temperature adapted lines had significantly higher growth rates than low temperature adapted lines (Table S3.1, p=0.006) (Figure 3.1a, 3.1b).

Over the course of the experiment, high temperature lines increased biofilm production at the air-liquid interface. Individual populations attached to the sides of the test tube below the surface (e.g. Line 39, Figure S3.3), formed aggregates and pellicles at the air-liquid interface (Figure S3.4 and Table S3.2), and colony morphology ranged in size, color, and smoothness throughout the experiment with persistent wrinkly types in several lines (Figure S3.1, Table S3.3). The ancestor, low temperature lines, and high temperature lines differed in their biofilm formation (Figure 3.1c, 3.1d, Table S3.1, ANOVA, P=1e-03). Biofilm formation was not observed during the experiment in the low temperature adapted lines but was variable within the high temperature lines (Table S3.1, Nested-ANOVA, P<2e-16). Even when grown at the low temperature of 25˚C, high temperature adapted lines, in general, produced more biofilm than low temperature lines (Figure 3.1c, 3.1d, Table S3.1, P=3e-04). We also assessed the elemental composition of the lines but found no significant difference between groups (Figure S3.5, Table S3.4). Thus, adaptation to high temperature led to increased biofilm formation at the air-liquid interface but did not alter cellular stoichiometry.

We hypothesized that biofilm formation may be a response to oxygen limitation as oxygen solubility in seawater decreases 12.1% between 25˚C and 33˚C (Benson & Krause, 1984). To test this, we manipulated the culture surface-to-volume ratio and associated volume-normalized oxygen transfer rate (OT). There was a significant difference between ancestor, low, and high temperature adapted lines (Table S3.1, ANOVA, P=1.3e-04). The ancestor grew poorly in the low OT environment relative to the high OT environment (mean ratio=0.93, sd=0.1) and continued to grow poorly when adapted to the low temperature environment (mean ratio=0.94, sd=0.04). This disadvantage was overcome in the high temperature adapted lines (mean ratio=1.12, sd=0.16). On average, the high temperature adapted group grew better in the low vs. high OT environment compared to both the ancestor (Table S3.1, P=0.04) and low temperature adapted groups (Figure 3.1e, 3.1f, Table S3.1, P<2e-16).

The physiological changes in growth rate, biofilm formation, and growth in low vs. high OT environments partially co-varied. Biofilm formation and growth rate had the strongest positive correlation (Figure 3.2a, rho=0.53, P=3.5e-04). The correlation between growth rate and growth in low vs. high OT environments was almost as strong (Figure 3.2b, rho=0.51, P=5.2e-04). Biofilm formation and growth in low vs. high OT environments had the weakest correlation (Figure 3.2c, rho=0.26, P=0.097). This relationship had divergent patterns where some high temperature adapted lines were positively correlated while many low biofilm producers grew well under lower OT (Figure 3.2c). Correlations between physiological phenotypes varied in strength but overall were driven by high temperature adaption.

We next identified the genomic variation underlying high temperature adaptation

by sequencing the genomes of the ancestor, the low, and the high temperature adapted lines. A total of 182 mutations were observed among all lines (Figure 3.3a, 3.3b, Table S3.5), with more mutations accrued in high (6.8) vs. low (1.7) temperature lines (Welch t-test, t=-14.27, df=35.49, P=2.9e-16). Lines shared twenty-four mutations at the genic level (Table S3.5). Most were shared among high temperature lines (19 mutations), yet there was convergent evolution to the low temperature environment (1 mutation) and to the shared laboratory environment (4 mutations). To assess the contribution of convergent mutations to phenotypic variation, we created a linear genome-wide model based on genotypic predictors for each phenotypic variable (Figure 3.4). Overall, the models had adjusted-R^2 of 0.82 for biofilm formation, 0.46 for growth in low vs. high OT environments, and 0.3 for growth rate based on 14 common (>2 lines), shared mutations (Table S3.6). Mutations varied in their contribution to each particular phenotypic model but accounted for a significant portion of the observed phenotypic variation.

Mutations grouped broadly into those affecting ‘regulatory’ genes associated with biofilms and those affecting ‘direct’ genes that related to specific functional pathways. Over half (59%) of the high temperature adapted lines had mutations in either the quorum sensing regulators *lux*R + *lux*I or the RNA-binding regulation protein *hfq* and accounted for 35.2% of variance in biofilm formation. However, no line had concurrent mutations in both *hfq* and *lux*RI. Mutations in the adjacent genes encoding the transcriptional activator protein LuxR and the autoinducer synthesis protein LuxI, were, when considered together, the greatest contributors to variance in biofilm formation (Figure 3.4). Furthermore, each line with observed wrinkly morphology had a mutation

in *lux*RI. One mutation in *luxR*, E193K, occurred in a probable DNA-binding location and shifted the amino acid from glutamic acid to lysine, which could be disruptive. All other mutations in *lux*RI were nonsense mutations. The quorum sensing auto-inducing regulation is linked to biofilm formation in many bacteria and may regulate the initial transition from free-living to attached states(Wagner-Döbler & Biebl, 2006). In some *Roseobacter*, the opposite trend has been observed, where inhibition of LuxRI reduces motility and creates thicker biofilms (Zan *et al.*, 2012). This negative relationship between quorum sensing and biofilm formation has also been observed in *Vibrio fisheri* (Yildiz & Visick, 2009) and suggests that suppression of quorum-sensing controls was a component of adaptation to high temperature and subtle declines in oxygen.

The other transcriptional pathway under selection was post-transcriptional small RNA regulation by Hfq, whereby six high temperature lines had changes in or upstream *hfq* (Figure 3.4). One of the six occurred 93 bp upstream of *hfq* and another was a non-synonymous change (R49W). The remaining four were the same non-synonymous mutation (Q54E), which encoded a small shift from glutamine to glutamic acid and was found in a highly conserved region, suggesting a fine-tuning of the gene rather than a loss of function mutation. This small protein is a global regulator of gene expression interacting with many different regulatory sRNAs that target mRNAs for degradation and is involved in the bacterial stress response (Møller *et al.*, 2002; Massé *et al.*, 2003; Glaeser *et al.*, 2007). In *V. cholerae* and *V. harveyii,* Hfq degrades the *lux*R mRNA (Lenz *et al.*, 2004). In other species, it also plays a role in biofilm formation (Zeng *et al.*, 2013; Hammer & Bassler, 2003; Rempe *et al.*, 2012) supporting that *hfq* influences the observed changes in biofilm formation.

Several genes accounting for a significant amount of biofilm variation were directly involved in the secretion system or the secreted biofilm components. A T1-secretion system (T1SS) repeat target protein had deletions in 5 lines ranging from 330 to 1404 bp. This mutation was also a significant predictor for growth rate and growth in low vs. high OT environments (Figure 3.4). The T1SS gene, secreted agglutinin RTX (repeats in toxins), was correlated with mutations in *hfq*. A gene related to exopolysaccharide formation, *epsK,* was mutated in five different lines with three different types of non-synonymous mutations. Many mutations present in only a few high temperature lines appeared to be related to biofilms based on their annotation. These genes were the preprotein translocase subunit *sec*G, the Type I secretion system outer membrane component *lap*E, the Type II/IV secretion system ATPase *tadZ/cpa*E, a large exoprotein involved in adhesion, and another T1SS repeat target protein found elsewhere in the genome. These biofilm-related mutations were found solely in high temperature adapted lines.

A few mutations associated with changes in growth rate and oxygen transfer phenotypes. Mutations in CTP synthase were a significant negative predictor for biofilm formation but a positive predictor for growth rate (Figure 3.4). A large deletion from 2.24 Mbp to 2.28 Mbp, 'Large Deletion 1', was present in four of the lines and was significantly associated with better growth in a low compared to a high OT environment. The deletion within this region varied widely among lines, disrupting between 4 and 51 genes (Figure S6). Among the four genes deleted commonly across the four lines, there was a hypothetical protein, a beta-lactamase, and two transcriptional regulators one of which was in the *lys*R family. Two additional lines had SNPs in *lys*R, making it the most

likely adaptive candidate. Other *lys*R-like transcriptional regulators in the genome were mutated in three lines, including two low temperature adapted lines. As three out of the four lines with 'Large deletion 1' were not large biofilm producers, these deletions were likely unrelated to biofilm formation. The transcriptional regulator containing an amidase domain with an *ara*C□type DNA□binding helix-turn-helix domain had a negative association with growth rate and growth in low vs. high OT environments. This mutation was present in six high temperature adapted lines and affected seventeen low temperature adapted lines. Carboxyl-terminal protease (EC 3.4.21.102), accounted for a lot of variation in growth in low vs. high OT environments, although it was not significant. Lines 27 and 46 had mutations in this protease as well as the C4-dicarboxylate proteins DctBD*,* mutations. As these mutations were found in Line 27, a line that never formed visible biofilm throughout the experiment, they may have contributed to increased growth within a lower oxygen transfer environment rather than increased biofilm formation. Considering the diversity of mutations and their differential contributions to physiological phenotypes, there were multiple divergent genetic pathways for *Roseovarius* to adapt to high temperature.

To test for the generality of evolutionary adaptive response to high temperature, we compared the genetic mutations identified in *Roseovarius sp.* TM1035 with those identified in *E. coli* REL 1206 (Tenaillon *et al.*, 2012) and found an absence of functional convergence. The strains shared some similarities in mutation type (indel, SNP, etc.), but the impacted genes differed between organisms. Most mutations were single nucleotide point mutations, followed by indels, large deletions, duplications, and IS insertions in order of abundance (Figure 3.3a, 3.3b). This structure was similar in the *E.*

coli experiment, except IS insertions were slightly more abundant than duplications. The number of mutations per line was similar to the distribution measured in *E. coli* for duplications and large deletions (Table S3.7), while the rest were more abundant per line in the *E. coli* experiment. In contrast, we saw a limited overlap in the specific genes affected by mutations (Figure 3.3c). *Roseovarius sp.* TM1035 and *E. coli* REL1206 shared 1177 orthologous gene clusters. When considering the mutations found among the 22 high temperature lines in this study and the 115 lines in the *E. coli* study, only 60 of the 1177 shared clusters were involved in a mutation in both organisms (Table S3.8). Only six clusters had mutations in multiple lines in both studies including the prevalent *hfq* and the transcriptional regulator containing an amidase domain with an *ara*C-type DNA-binding HTH domain, but these were found in large duplications in the *E. coli* lines. The NAD-dependent formate dehydrogenase α-subunit was the only gene of the six that had single gene mutations in both studies (i.e. not a large deletion/duplication which have unclear adaptive targets). In *E. coli*, all lines had mutations in either the RNA polymerase operon or *rho* termination factor that led to marked effects on gene expression patterns (Rodŕiguez-Verdugo *et al.*, 2016). However, neither gene contained any mutations among *Roseovarius* evolved lines except within a large duplication found in one line. Thus, adaptation to high temperature differed between organisms whereby *Escherichia coli* experienced modifications in core transcriptional regulatory genes, while this abundant marine *Roseobacter* demonstrated a genetic response supporting a change in lifestyle leading to increased biofilm formation. *Pseudomonas* also develop biofilms at the air-liquid interface when grown under static culture conditions (Rainey & Travisano, 1998). In this setting, oxygen is more severely limited and organisms

outcompete others in the culture by increasing their access to oxygen (Hansen, Haagensen, *et al.*, 2007). As in the Pseudomonas experiments, we also observed wrinkly colony formations in several high temperature adapted lines, which may reflect adaptations to oxygen stress as wrinkles allow greater oxygen penetration into the colony (Rainey & Travisano, 1998; Spiers *et al.*, 2003). However, adaptations in *Pseudomonas* resulted in mutations in cellulose production (*wss* operon or *wap*H) (Spiers *et al.*, 2002; Hansen, Rainey, *et al.*, 2007) which were not seen in *Roseobacter,* demonstrating that there are many different avenues for genetic adaptation to increase biofilm formation (Jefferson, 2004). It is clear that different lineages will take different evolutionary pathways in adaptation to high temperature and we cannot generalize an adaptive response across diverse heterotrophic bacteria.

Abundant marine *Roseobacter* show a clear capacity for rapid adaptation to elevated temperature through a significant lifestyle change. We observed that adaptation led to a rapid increase in growth rate, biofilm formation, and the ability to grow under increased oxygen stress. There was a diversity of genomic changes clearly associated with a variety of biofilm formation pathways, offering a unique response compared to previous experimental tests of adaptation to high temperature (Rodríguez-verdugo *et al.*, 2013; Tenaillon *et al.*, 2012; Riehle *et al.*, 2001; Deatherage *et al.*, 2017) Moreover, an increase in temperature and a derived reduction in oxygen solubility were linked to this major lifestyle change. Ocean oxygen levels are predicted to decrease as a result of ocean heat uptake (Keeling *et al.*, 2010) leading to a strong selective effect on marine microorganisms. While much interest has been focused on the growing oxygen minimum zones and reduced oxygenation to deeper waters due to stratification

(Keeling *et al.*, 2010; Rabalais *et al.*, 2010), our research demonstrates that subtle changes in oxygen solubility in the ocean surface can have substantial physiological consequences for marine bacteria.

Methods

Experimental evolution

Roseovarius sp. TM1035 was isolated from a Chesapeake Bay culture of a dinoflagellate *Pfiesteria piscicida* CCMP1830 (Miller & Belas, 2004; Alavi *et al.*, 2001) and optimally grows at 25°C with a growth rate of 0.15 OD 600 h^{-1} (Figure S3.2). A single colony of ancestral *Roseovarius sp.* TM1035 was grown up in 100 mL of YTSS Media (4 g/L yeast extract, 2.5 g/L tryptone, and 10 g/L sea salts) at its optimal temperature of 25°C. Then, 200 µL of culture were transferred to 44 replicates, (22 control lines held at a constant low temperature of 25°C and 22 experimental lines held at a high temperature of 33°C), which were serially propagated (every 30 or 48 hours respectively to reduce stationary phase dynamics) for 500 generations in 10 mL YTSS media under rigorous shaking. Sample sizes were chosen to be larger than other experimental evolution studies (Lenski & Bennett, 1993; Riehle *et al.*, 2003; Deatherage & Barrick, 2014). At the initiation of the experiment, the ancestor was also preserved from the same starting culture for later comparison. Two low temperature lines were removed from the experiment due to contamination. Plate checks and 16S rRNA PCR assay check confirmed the contamination (Figure S3.7). For the 16S rRNA PCR assay, colonies of interest were diluted in nuclease free water, boiled, and PCR amplified using 16S rRNA primers targeting *Roseovarius sp.* TM1035. Primers were 16S rRNA Forward 5′-ACT AGG GTT TTG GCC CGA TG-3′ and 16S rRNA Reverse 5′-CTT TCC

CCC AAA GGG CGT AT -3′. After 500 generations, populations were frozen and then streaked on an agar plate. A single colony was picked, streaked again on agar and a single colony chosen at random for DNA sequencing. These single colonies were cultured for two days in 10 mL YTSS media at 25°C or 33°C depending on their experimental temperature, and 5 mL was used for DNA extraction, while the remainder was concentrated and frozen for future analyses.

Physiological assays

To begin all post-experimental evolution assays, cultures were taken from frozen stock and incubated for two days at 25°C, transferred via 50-fold dilution into 25°C for two days, and then transferred again to start each experiment. Although the high temperature adapted clones had evolved at 33°C in YTSS, it is common practice in thermal stress studies to allow clones to recover from freezing under less stressful conditions (Lenski & Bennett, 1993; Rodriguez-Verdugo *et al.*, 2014). Biofilm formation was quantified in 96-well microplates using a modified crystal violet method (O'Toole, 2011). This assay likely underestimates biofilm formation because the biofilm must remain attached to the plate after the wash to be quantified. 100 µL of 50-fold diluted cultures were pipetted into 6 or more biological replicates. Plates were kept shaking (120 rpm) at either 25°C or 33°C for 48 hours. After incubation, plates were gently rinsed with milli-Q water six times and 125 µL of 0.1% crystal violet stain was added and incubated at room temperature for 15 minutes. Plates were rinsed six times and left to dry. After drying, 125 µL of 30% acetic acid was added to solubilize the remaining stain and quantified on a spectrophotometer at 550 nm. To remove the confounding variable of temperature, cultures were diluted 50-fold into either 10 mL or 20 mL of media and

incubated shaking at 33°C in 50 mL flasks for the oxygen limitation assay. In a 50 mL, shaking flask, 10 mL of media should be more aerated with a surface to volume ratio of 1.77 cm^2/mL than 20 mL of media with a surface to volume ratio of 0.71 cm^2/mL (Somerville & Proctor, 2013). Optical density (OD) was measured at 600 nm on a spectrophotometer. Growth rates were determined from the 20 mL flasks calculated from the OD measurements at approximately 9 hours and 22 hours. Outliers in all analyses above or below 5 SD range were removed iteratively (2 from oxygen transfer assay).

Sequencing

DNA was extracted using Mini Genomic DNA Kit (Plant), (IBI Scientific, Peosta, IA) and purified with Genomic DNA Clean & Concentrator kit (Zymo, Irvine, CA). DNA concentration was quantified with Qubit dsDNA high sensitivity assay kit (Life Technologies) and subsequently diluted to a concentration of 0.2 ng/µl. Using 1 ng of genomic DNA from each clone, libraries were created and manually pooled using Illumina NexteraXT library preparation kits with NexteraXT barcodes (Illumina, San Diego, CA). The resulting library was checked and quantified on a BioAnalyzer. The pooled libraries were sequenced on two separate runs of the Illumina HiSeq 2500 sequencer producing either paired-end (2 X 100 bp) or single-end (1 X 100 bp) reads with half of the high and low temperature lines on each run (Table S3.9). The sequence reads were reassembled against the draft genome of *Roseovarius sp*. TM1035 (Miller & Belas, 2004) using CLC Genomics Workbench (Version 9.0.1, www.qiagenbioinformatics.com) and CLC Genomics Finishing Module (Version 1.6) with 3 libraries to a coverage of 1180X. The assembly was assessed using REAPR

(Hunt *et al.*, 2013) and possible issues due to repeats were compiled but dismissed when checked against the mate-pair library generated during the original sequencing of *Roseovarius sp.* TM1035 (Table S3.10). After reassembly, this strain appeared to have a plasmid with a length of 156,827 bp and a genomic sequence that was 4,046,715 bp long.

The genome was reannotated using the online RAST server (Aziz *et al.*, 2008). Molecular changes in each line were called against the assembled ancestral sequence as a reference using *breseq* in consensus mode (Deatherage & Barrick, 2014). Many inaccurate polymorphism predictions arise from a strand bias during library preparation and sequencing, we used a cutoff of (α=0.01) with *breseq*'s two-sided Fisher's exact test to identify spurious polymorphisms with this bias. *Breseq* has a harder time assessing rearrangements involving repeat-regions and duplications (Deatherage & Barrick, 2014). Thus rearrangements predicted, but not confirmed by *breseq*, were manually checked in CLC Genomics Workbench 9. As well, a 2-fold increase in read coverage over the ancestor coupled with a novel junction mapping the endpoints of this increase indicated a duplication event.

The mutations found in *Roseovarius sp*. TM1035 were compared to the mutations found in *E. coli* evolved to high temperature (Tenaillon *et al.*, 2012). The REL1206 genome was annotated using the same RAST pipeline as *Roseovarius sp.* TM1035. Genes were included in the genic subset if the molecular change overlapped with a RAST called gene (including multigenic deletions). Intergenic mutations or mutations extending into an intergenic region included the two neighboring genes

surrounding a mutation. Genes were clustered by homology using OrthoMCL V1.4 (Li *et al.*, 2003).

A genome-wide least squares linear regression model was used to associate mutations with changes in growth rate, biofilm formation, and growth in low vs. high OT environments across all lines, including the ancestor, based on shared mutations. Mutations found in several lines provided the best information for associations; accordingly we only included in our linear regression model mutations that were found in three or more lines. Some mutations were correlated with each other; consequently we removed the mutation that contributed less to the models. Mutations in *hfq* were correlated with T1SS secreted agglutinin RTX, mutations in epimerase/dehydratase were correlated with CTP synthase and the T1SS repeat protein, and mutations in the large deletion 1 were correlated with the C4-dicarboxylate proteins *DctBD* (Table S3.11). An ANOVA of the model fit was performed and significant coefficients were noted (Figure 3.4). All statistical analyses were performed in R (Version 3.3) (R Development Core Team, 2016).

Data availability

The data supporting the results of this study are available within the paper and its Supplementary Information and Supplementary Data Files. The assembled genome and sequences from clonal lines used in this study have been deposited in NCBI BioProject database under accession PRJNA386804, SAMN07134944: SAMN07134988 and SAMN07138891 for the genome assembly.

References

Alavi M, Miller T, Erlandson K, Schneider R, Belas R. (2001). Bacterial community associated with *Pfiesteria*-like dinoflagellate cultures. **3**:380–396.

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, *et al.* (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**.

Benson BB, Krause D. (1984). The concentration and isotopic fractionation of oxygen dissolved in freshwater and seawater in equilibrium with the atmosphere1. *Limnol Oceanogr* **29**:620–632.

Bergmann N, Winters G, Rauch G, Eizaguirre C, Gu J, Nelle P, *et al.* (2010). Population-specificity of heat stress gene induction in northern and southern eelgrass Zostera marina populations under simulated global warming. *Mol Ecol* **19**:2870–2883.

Buchan A, Moran MA. (2005). Overview of the Marine *Roseobacter* Lineage †. **71**:5665–5677.

Collins S. (2010). Many Possible Worlds: Expanding the Ecological Scenarios in Experimental Evolution. *Evol Biol* **38**:3–14.

Deatherage DE, Barrick JE. (2014). Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using *breseq*. *Methods Mol Biol* **1151**:165–188.

Deatherage DE, Kepner JL, Bennett AF, Lenski RE, Barrick JE. (2017). Specificity of genome evolution in experimental populations of *Escherichia coli* evolved at different temperatures. *Proc Natl Acad Sci* 201616132.

Franks SJ, Hoffmann A a. (2012). Genetics of climate change adaptation. *Annu Rev*

Genet **46**:185–208.

Galbraith ED, Martiny AC. (2015). A simple nutrient-dependence mechanism for predicting the stoichiometry of marine ecosystems. *Proc Natl Acad Sci* **112**: 8199–8204.

Giebel HA, Brinkhoff T, Zwisler W, Selje N, Simon M. (2009). Distribution of *Roseobacter* RCA and SAR11 lineages and distinct bacterial communities from the subtropics to the Southern Ocean. *Environ Microbiol* **11**:2164–2178.

Glaeser J, Zobawa M, Lottspeich F, Klug G. (2007). Protein synthesis patterns reveal a complex regulatory response to singlet oxygen in Rhodobacter. *J Proteome Res* **6**:2460–2471.

Hammer BK, Bassler BL. (2003). Quorum sensing controls biofilm formation in Vibrio cholerae. **50**:101–114.

Hansen SK, Haagensen JAJ, Gjermansen M, Jørgensen TM, Tolker-Nielsen T, Molin S. (2007). Characterization of a Pseudomonas putida rough variant evolved in a mixed-species biofilm with Acinetobacter sp. strain C6. *J Bacteriol* **189**:4932–4943.

Hansen SK, Rainey PB, Haagensen JAJ, Molin S. (2007). Evolution of species interactions in a biofilm community. *Nature* **445**:533–536.

Hug SM, Gaut BS. (2015). The phenotypic signature of adaptation to thermal stress in *Escherichia coli*. *BMC Evol Biol* 1–12.

Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. (2013). REAPR: a universal tool for genome assembly evaluation. *Genome Biol* **14**:R47.

Jefferson KK. (2004). What drives bacteria to produce a biofilm? *FEMS Microbiol Lett*

236:163–173.

Keeling RE, Körtzinger A, Gruber N. (2010). Ocean deoxygenation in a warming world. *Ann Rev Mar Sci* **2**:199–229.

Lenski R, Bennett A. (1993). Evolutionary Response of *Escherichia coli* to Thermal Stress. *Am Nat*.

Lenz DH, Mok KC, Lilley BN, Kulkarni R V, Wingreen NS, Bassler BL, *et al.* (2004). The Small RNA Chaperone Hfq and Multiple Small RNAs Control Quorum Sensing in Vibrio harveyi and Vibrio cholerae. **118**:69–82.

Li L, Stoeckert CJJ, Roos DS. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes -- Li et al. 13 (9): 2178 -- Genome Research. *Genome Res* **13**:2178–2189.

Lomas MW, Burke AL, Lomas DA, Bell DW, Shen C, Dyhrman ST, *et al.* (2010). Sargasso Sea phosphorus biogeochemistry: an important role for dissolved organic phosphorus (DOP). *Biogeosciences* **7**:695–710.

Massé E, Escorcia FE, Gottesman S. (2003). Coupled degradation of a small regulatory RNA and its mRNA targets in *Escherichia coli*. *Genes Dev* **17**:2374–2383.

Miller TR, Belas R. (2004). Dimethylsulfoniopropionate Metabolism by Dimethylsulfoniopropionate Metabolism by *Pfiesteria*-Associated *Roseobacter spp.*†. **70**.

Møller T, Franch T, Højrup P, Keene DR, Ba HP, Brennan RG, *et al.* (2002). Hfq : A Bacterial Sm-like Protein that Mediates RNA-RNA Interaction. **9**:23–30.

Moran MA, Buchan A, González JM, Heidelberg JF, Whitman WB, Kiene RP, *et al.* (2004). Genome sequence of Silicibacter pomeroyi reveals adaptations to the

marine environment. *Nature* **432**:910–3.

Nielsen ES, Jørgensen EG. (1968). The Adaptation of Plankton Algae: I. General part. *Physiol Plant* **21**:401–413.

O'Toole G a. (2011). Microtiter dish biofilm formation assay. *J Vis Exp* 3–5.

Parry M, Canziani O, Palutikof J, van der Linden P, Hanson C. (2007). Climate Change 2007: Impacts, Adaptation and Vulnerability. *IPPC Clim Chang 2007 Impacts, Adapt Vulnerability* 976.

Rabalais NN, Díaz RJ, Levin LA, Turner RE, Gilbert D, Zhang J. (2010). Dynamics and distribution of natural and human-caused hypoxia. *Biogeosciences* **7**:585–619.

R Development Core Team. (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. *R Found Stat Comput Vienna, Austria*.

Rainey PB, Travisano M. (1998). Adaptive radiation in a heterogeneous environment. *Nature* **394**:69–72.

Rempe KA, Hinz AK, Vadyvaloo V. (2012). Hfq regulates biofilm gut blockage that facilitates flea-borne transmission of Yersinia pestis. *J Bacteriol* **194**:2036–2040.

Riehle MM, Bennett a F, Long a D. (2001). Genetic architecture of thermal adaptation in *Escherichia coli*. *Proc Natl Acad Sci U S A* **98**:525–30.

Riehle MM, Bennett AF, Lenski RE, Long AD. (2003). Evolutionary changes in heat-inducible gene expression in lines of *Escherichia coli* adapted to high temperature. *Physiol Genomics* **14**:47–58.

Rodriguez-Verdugo A, Carrillo-Cisneros D, Gonzalez-Gonzalez A, Gaut BS, Bennett AF. (2014). Different tradeoffs result from alternate genetic adaptations to a

common environment. *Proc Natl Acad Sci* **111**:12121–12126.

Rodríguez-verdugo A, Gaut BS, Tenaillon O. (2013). Evolution of *Escherichia coli* rifampicin resistance in an antibiotic-free environment during thermal stress. *BMC evolutionary biology*, **13**:50.

Rodŕiguez-Verdugo A, Tenaillon O, Gaut BS. (2016). First-Step mutations during adaptation restore the expression of hundreds of genes. *Mol Biol Evol* **33**:25–39.

Sharp JH. (1974). Improved analysis for “particulate” organic carbon and nitrogen from seawater. *Limnol Ocean* **19**:984–989.

Somerville GA, Proctor RA. (2013). Cultivation conditions and the diffusion of oxygen into culture media: The rationale for the flask-to-medium ratio in microbiology. *BMC Microbiol* **13**:1.

Spiers AJ, Bohannon J, Gehrig SM, Rainey PB. (2003). Biofilm formation at the air-liquid interface by the Pseudomonas fluorescens SBW25 wrinkly spreader requires an acetylated form of cellulose. *Mol Microbiol* **50**:15–27.

Spiers AJ, Kahn SG, Bohannon J, Travisano M, Rainey PB. (2002). Adaptive divergence in experimental populations of Pseudomonas fluorescens. I. Genetic and phenotypic bases of wrinkly spreader fitness. *Genetics* **161**:33–46.

Tenaillon O, Rodríguez-Verdugo A, Gaut RL, McDonald P, Bennett AF, Long AD, *et al.* (2012). The molecular diversity of adaptive convergence. *Science* **335**:457–461.

Thomas MK, Kremer CT, Christopher A, Litchman E, Klausmeier C a. (2012). A global pattern of thermal adaptation in marine phytoplankton. *Science* **338**:1085–1088.

Toseland A, Daines SJ, Clark JR, Kirkham A, Strauss J, Uhlig C, *et al.* (2013). The impact of temperature on marine phytoplankton resource allocation and

metabolism. *Nat Clim Chang* **3**:979–984.

Wagner-Döbler I, Biebl H. (2006). Environmental biology of the marine *Roseobacter* lineage. *Annu Rev Microbiol* **60**:255–80.

Yildiz FH, Visick KL. (2009). Vibrio biofilms: so much the same yet so different. *Trends Microbiol* **17**:109–118.

Zan J, Cicirelli EM, Mohamed NM, Sibhatu H, Kroll S, Choi O, *et al.* (2012). A complex LuxR-LuxI type quorum sensing network in a roseobacterial marine sponge symbiont activates flagellar motility and inhibits biofilm formation. *Mol Microbiol* **85**:916–933.

Zeng Q, Mcnally RR, Sundin GW. (2013). Global Small RNA Chaperone Hfq and Regulatory Small RNAs Are Important Virulence Regulators in Erwinia amylovora. **195**:1706–1717.

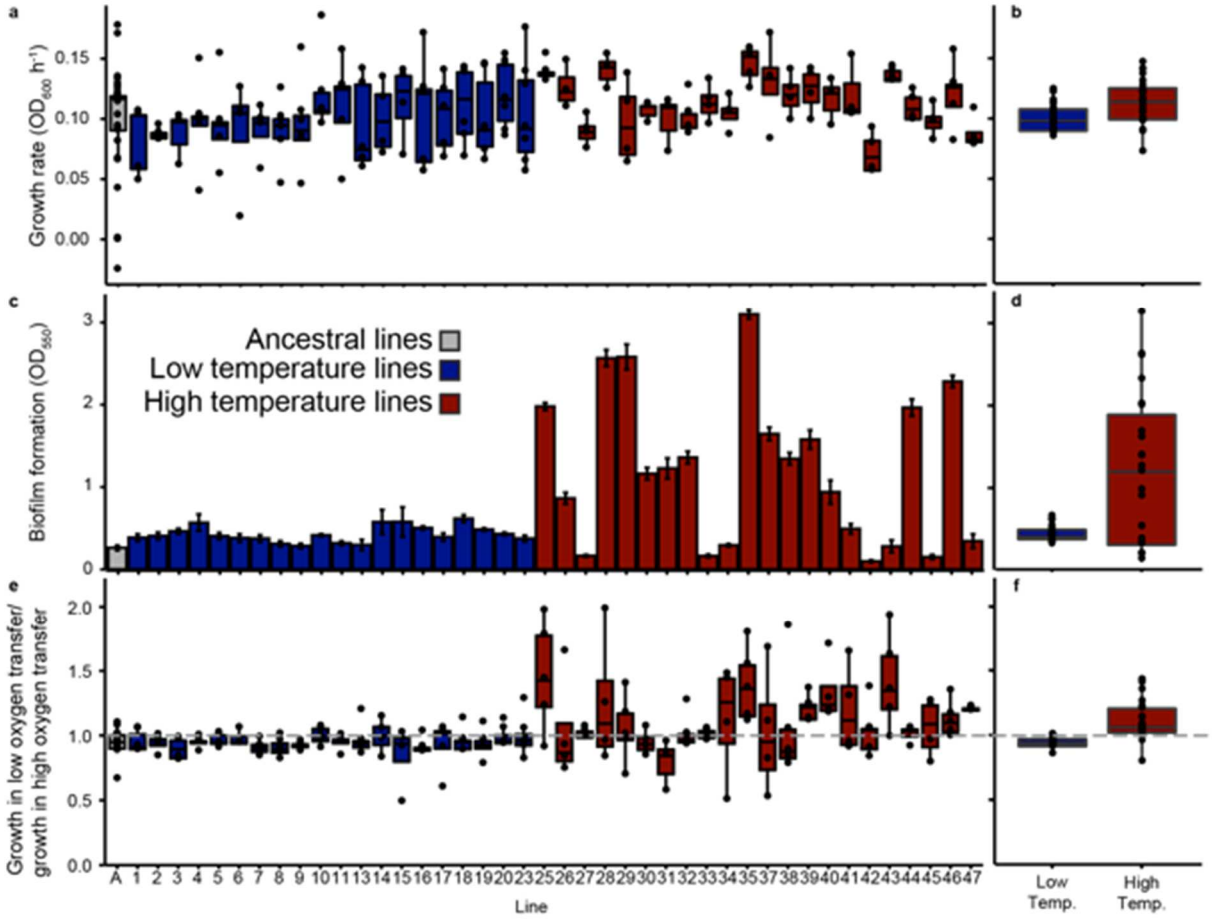


Figure 3.1. Changes in phenotypes due to adaptation. **a**, Changes in growth rate measured as change in optical density (OD) h^{-1} separated by line at 33°C. **b**, as in **a**, but mean by line aggregated by regime. **c**, mean biofilm formation as measured by crystal violet assay grown at 25°C and error bars are standard error of the mean. **d**, as in **c**, but aggregated by regime. **e**, growth ratio of low vs. high oxygen transfer flasks. **f**, as in **e**, but mean of line aggregated by regime. A value at the dashed line of 1 represents clones that grew equally well in their low and high oxygen transfer flasks. In **a, b, d, e,** and **f** the box represents the interquartile range including the median and the ends of the whiskers represent ±1.5*interquartile range. For **b,d,f**, low temperature lines n=20 and high temperature lines n=22.

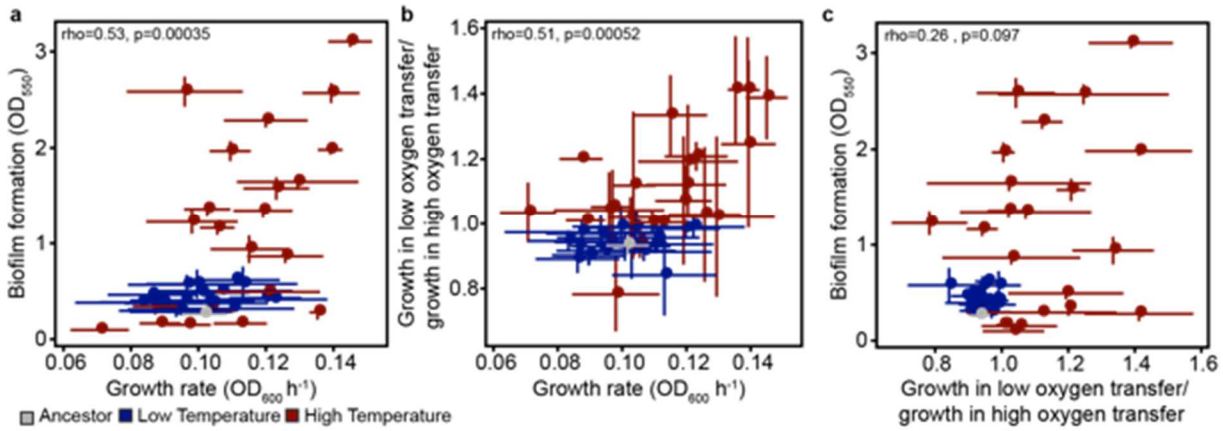


Figure 3.2. Correlation and variation between adaptive phenotypes. **a**, correlation between growth rate and biofilm formation. **b**, correlation between growth rate and the ratio of growth in low oxygen transfer to growth in high oxygen transfer. **c**, correlation between the ratio of growth in low oxygen transfer to growth in high oxygen transfer and biofilm formation. For each, Spearman correlation coefficient, rho, and p-values were included. Points are from Figure 3.1, where values are the mean of biological replicates for each line and error bars indicate standard error.

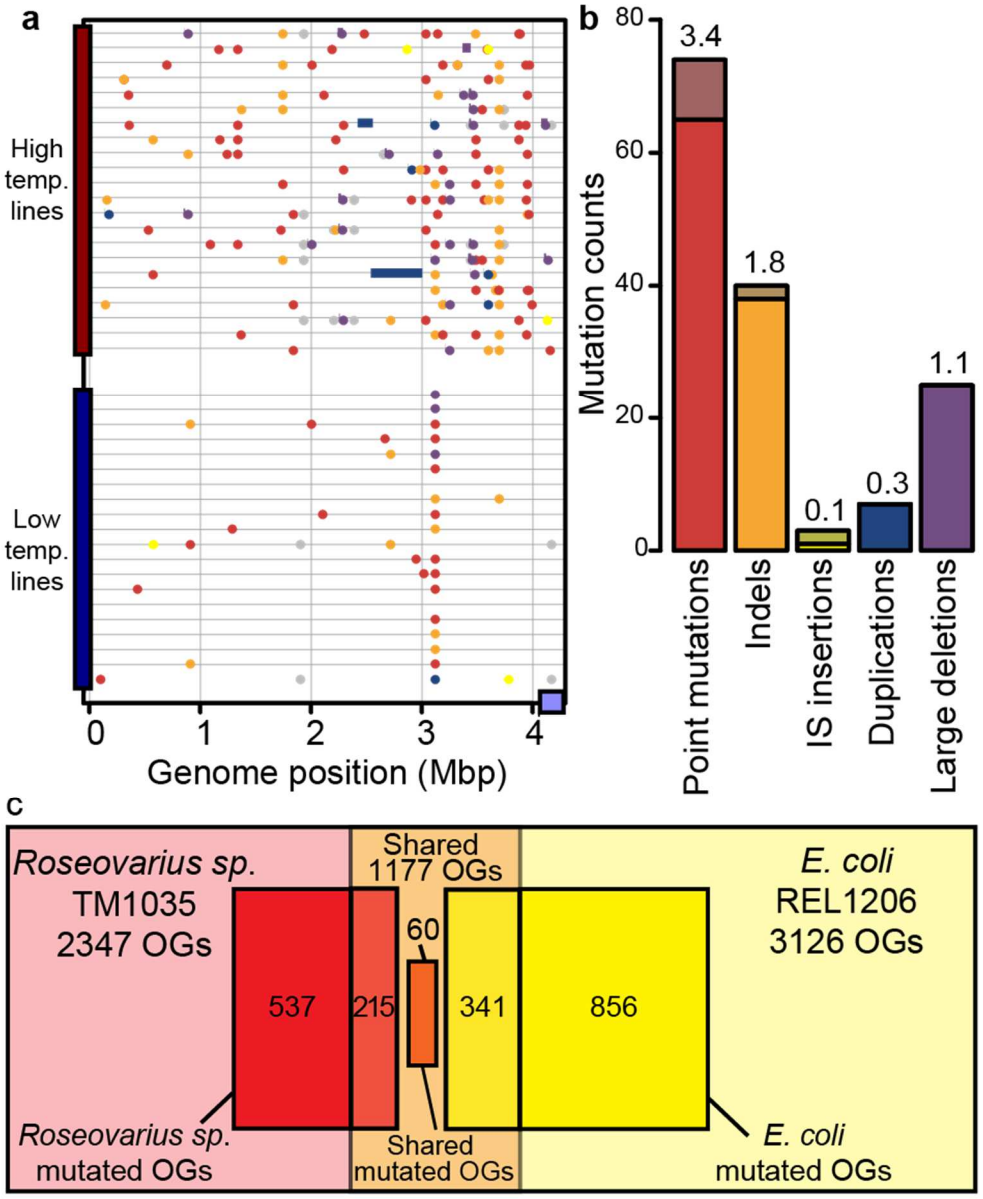


Figure 3.3. Mutation distribution across the genome and compared to *E. coli*. **a**, mutations in 42 independently evolved clones. Mutational types are colored as in **b**. Purple block at end of genome is the hypothetical plasmid sequence. **b**, the distribution of events according to mutational type with mutations split into genic (solid) and intergenic (shaded). **c,** genome and mutation overlap with high temperature adapted *E. coli* lines. Boxes proportionately represent *Roseovarius sp.* TM1035 orthologous groups (OGs) in red, *E. coli* REL1206 OGs in yellow, and shared groups in orange. OGs mutated in either study are represented in bolder colors.

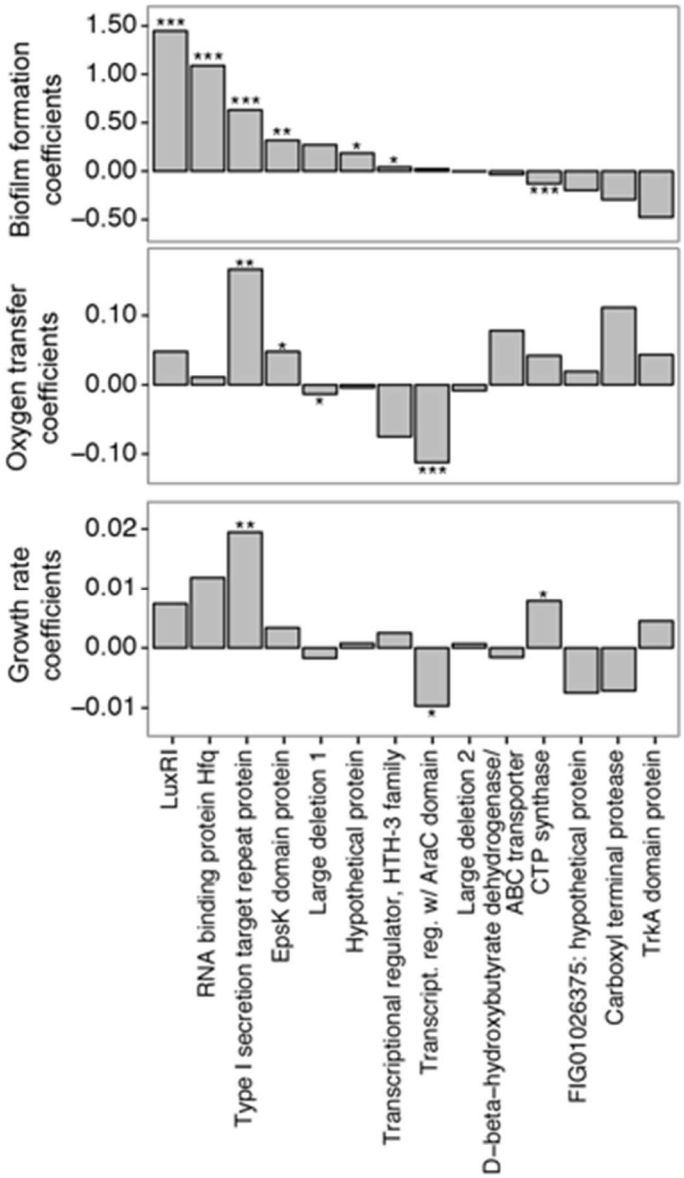
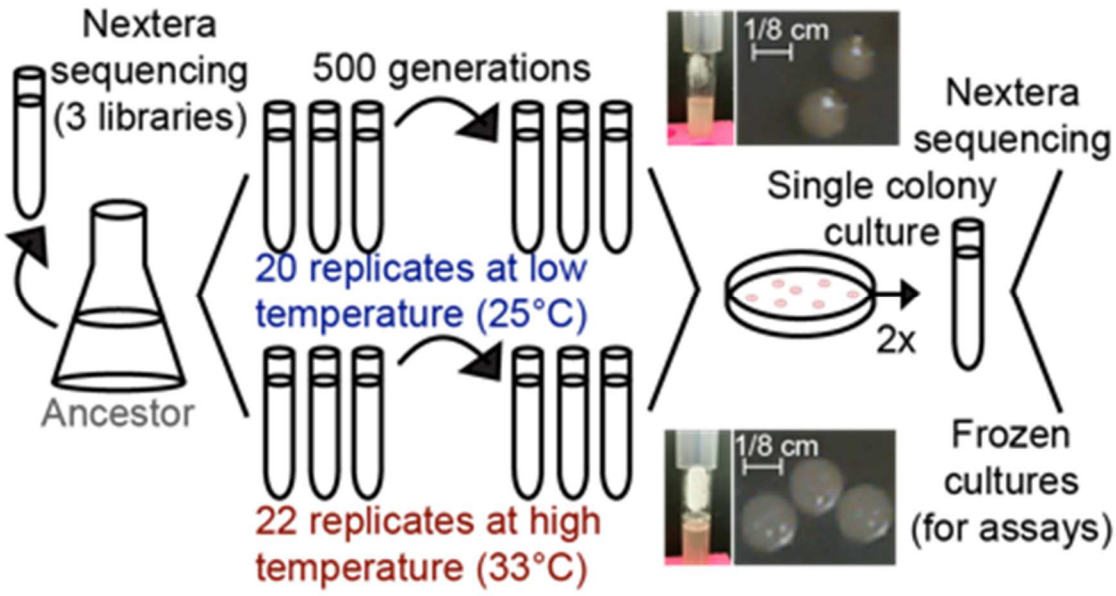
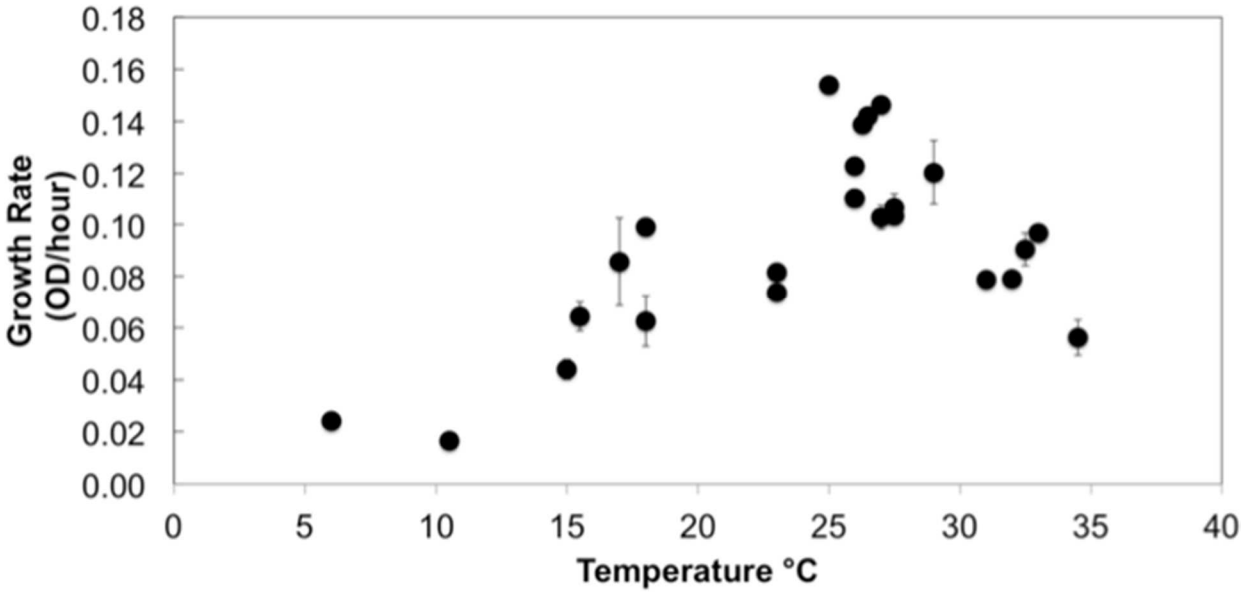


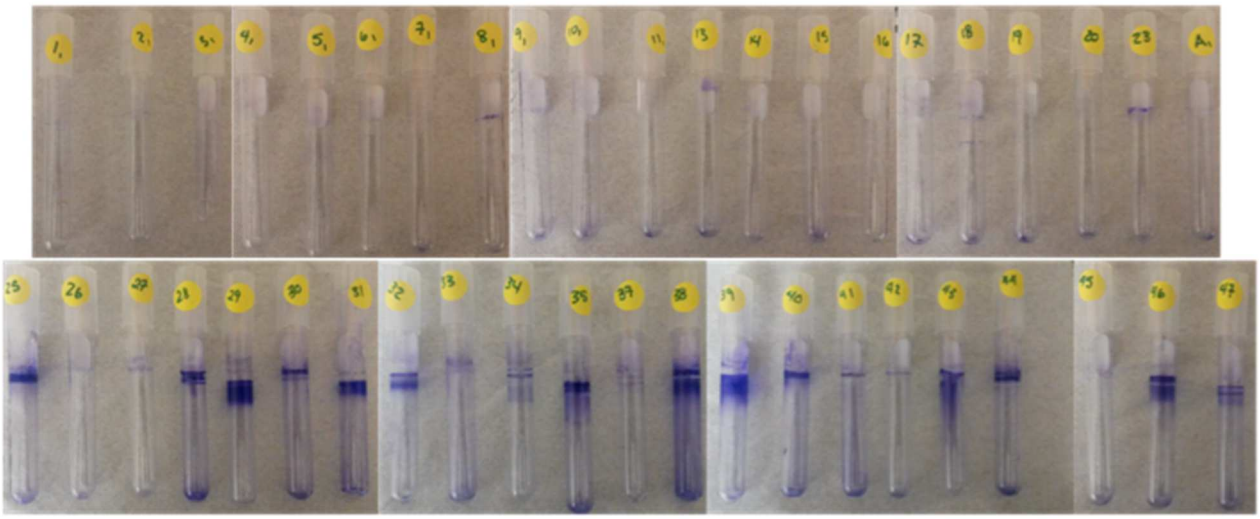
Figure 3.4. Genotypic associations with phenotypic variation**.** Multiple linear regression coefficients are shown for mutations as predictors for each measured phenotype (biofilm formation, growth in low vs. high oxygen transfer, and growth rate). Names of mutations or genes affected noted in Table S3.5 and ordered by biofilm formation coefficients. Asterisk represent *: P<0.05, **: P<0.01, ***: P<0.001 from ANOVA of linear model (Table S3.6).



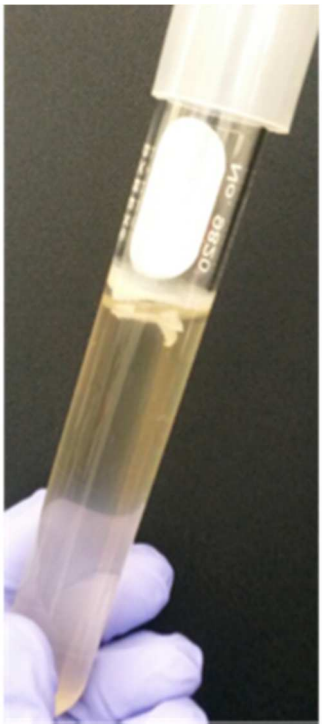
Supplementary Figure S3.1. Experimental setup. Ancestral lineage *Roseovarius sp*. TM1035 was serially propagated for 500 generations at either 25°C (lines 1 to 23) or 33°C (lines 25 to 47). Images are of a representative culture from each treatment. Wrinkly colony morphology was representative of 7 high temperature lines. More images are available in Figure S3.3.



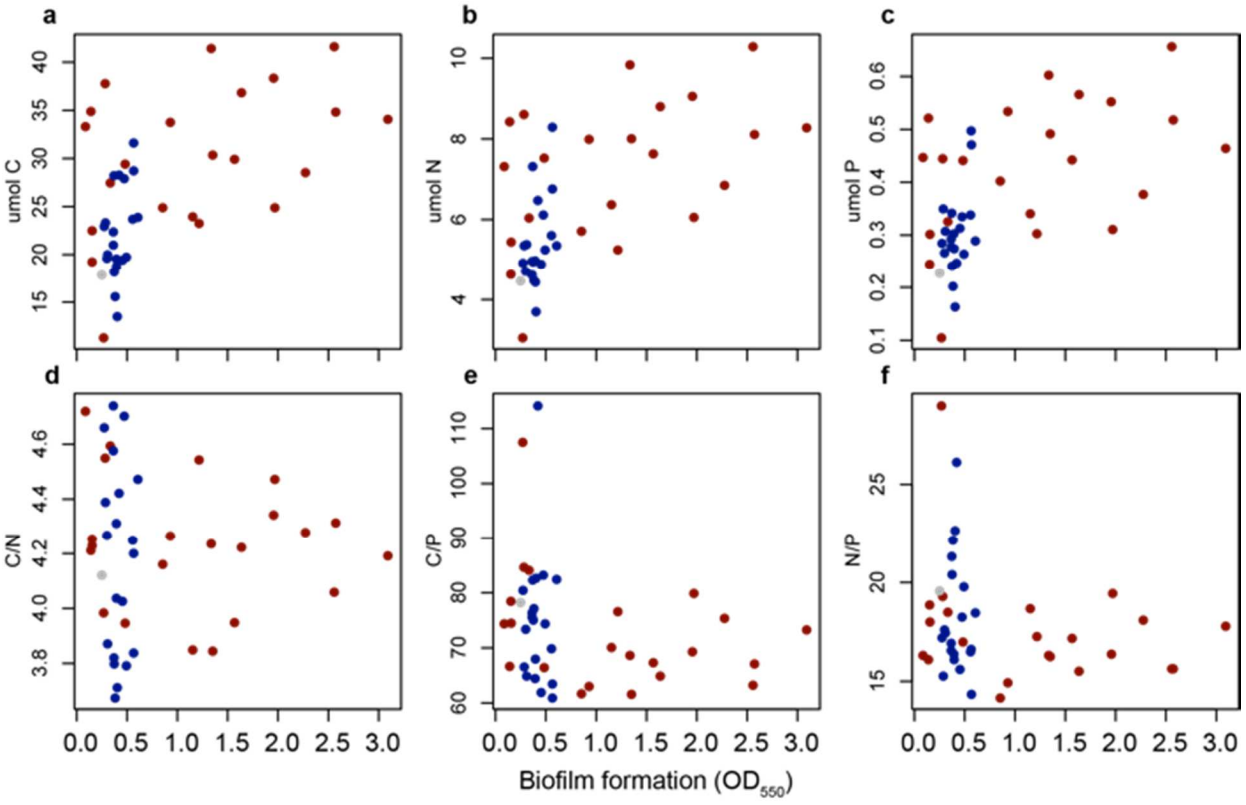
Supplementary Figure S3.2. Ancestral growth rate versus temperature. Growth rates for the ancestral lineage measured prior to experimental evolution were used to determine low and high temperature regimes. Error bars, when present, denote standard error of biological replicates, some temperatures were not repeated.



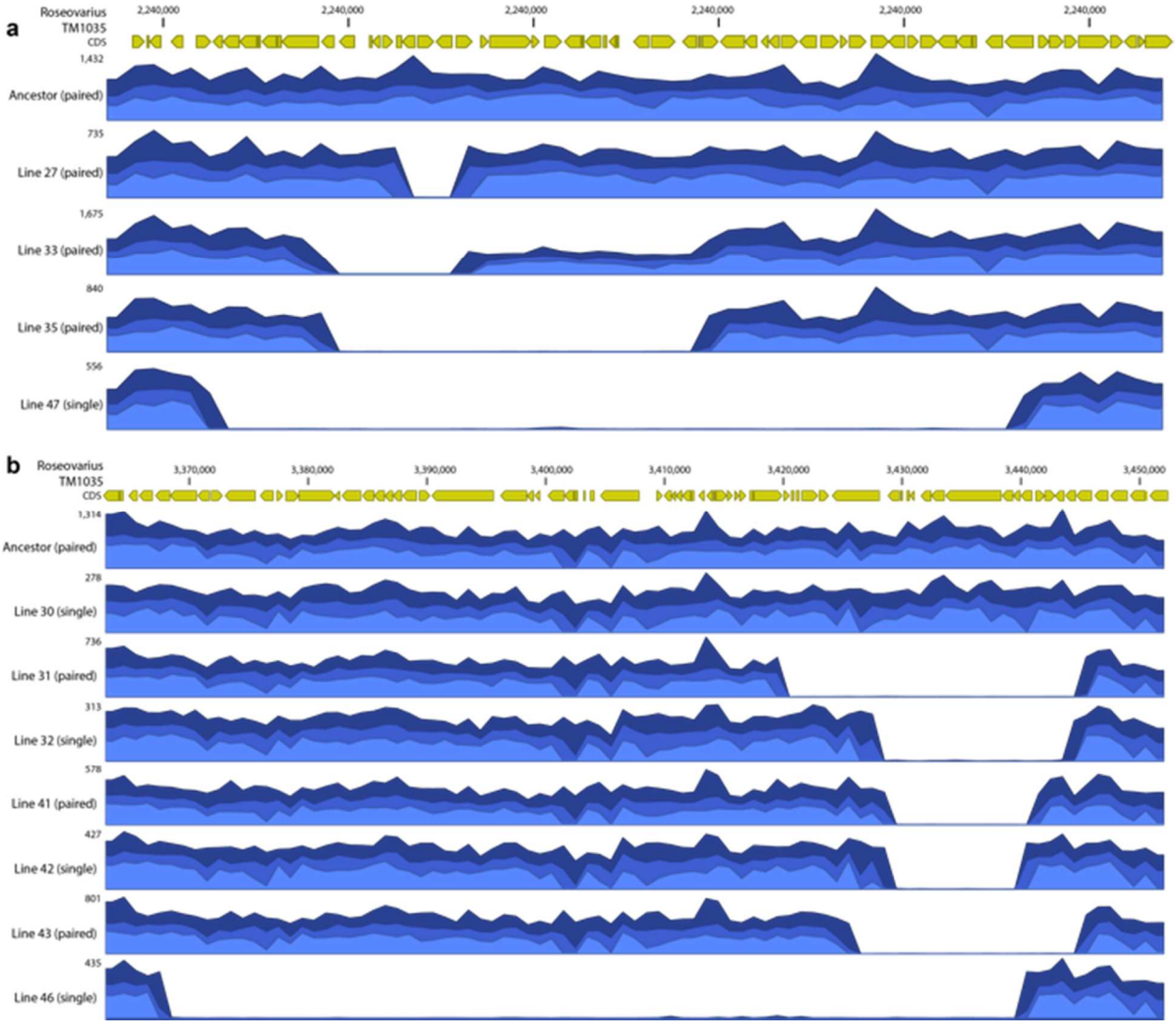
Supplementary Figure S3.3. Crystal violet stained culture tubes. Cultures were grown for two days at their respective temperatures, tubes were rinsed and then stained with crystal violet similar to biofilm assay. Line numbers are written on caps.



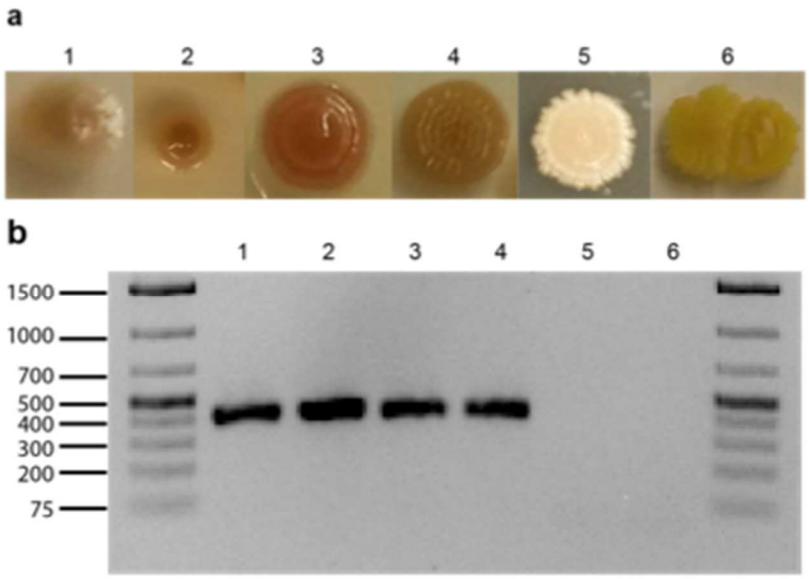
Supplementary Figure S3.4. Plate-like pellicle biofilm formation. Biofilm from line 25 shown here with a plate-like pellicle.



Supplementary Figure S3.5. Cellular stoichiometry across groups. Colored the same as in Figure 3.2. A, B, and C were absolute concentration of carbon, nitrogen and phosphorus measured from 1.5 mL of culture. D, E, F were ratios of carbon to nitrogen, carbon to phosphorus, and nitrogen to phosphorus for each line versus average biofilm formation for each line. Stoichiometric assays quantified particulate organic carbon (POC), nitrogen (PON) and phosphorus (POP). The ancestor, low, and high temperature adapted lines were grown at 33°C and 1.5 ml culture was filtered onto a 0.7 µm glass fiber filter during late exponential phase to maximize biofilm accumulation. For POC and PON, filters were packed into a 30 mm tin capsule (CE Elantech, Lakewood, NJ) and analyzed on a FlashEA 1112 nitrogen and carbon analyzer (Thermo Scientific, Waltham, MA) following Sharp (Sharp, 1974). To quantify POP, filters were analyzed using a modified ash-hydrolysis method (Lomas *et al.*, 2010).



Supplementary Figure S3.6. Large genomic deletions. Genomic deletions **a,** large deletion 1 and **b,** large deletion 2 in Figure 3.3 and Table 3.4. Ancestral library and each line with the deletion were mapped together. Coding sequences (CDS) within the deletion are represented in yellow with the genomic location above.



Supplementary Figure S3.7. Colony rRNA 16S PCR**. a,** images of colonies were 1-ancestor; 2-line 41 (February 18); 3-line 33 redder (March 18); 4-line 30 wrinkly type (February 24); 5-line 21 contamination (February 7); 6-random plate contamination (negative control) not scaled. **b,** gel electrophoresis of PCR product corresponding to colonies in **a**.

Supplementary Table S3.1. Phenotypic differences between experimental groups. Nested analysis of variance of experimental group (ancestral, low, and high temperature) with nested 'line' on growth rate, biofilm formation and ratio of growth in differing oxygen transfer environments. Second table includes Brown-Forsyth test for unequal variances. We tested variance of experimental group (ancestral, low, and high temperature) with Kruskal-Wallis nonparametric test and p-values from pairwise Mann-Whitney U-tests compare groups against one-another, p-value was adjusted with a Bonferroni correction.

Nested ANOVA						
	Sum of Squares	**Mean Square**	**DF**	**Denominator DF**	**F.value**	**Prob(>F)**
Growth Rate	8.4E-03	4.2E-03	2	226	4.14	0.017
Biofilm Formation	1.48	0.74	2	41.9	8.19	1.0E-03
Oxygen Transfer	1.46	0.73	2	16.12	16.29	1.3E-04
Difference of Least Squares-Means						
Growth Rate	**Standard Error**	**DF**	**t-value**	**Lower CI**	**Upper CI**	**P-value**
Ancestor-Low Temperature	0.006	226	0.17	-0.011	0.013	0.86
Ancestor-High Temperature	6.2E-03	226	-1.9	-0.024	0.0005	0.059
Low-High Temperature	4.6E-03	226	-2.75	-0.022	-0.0036	6.0E-03
Biofilm Formation						
Ancestor-Low Temperature	0.65	41.3	-0.26	-1.49	1.15	0.8
Ancestor-High Temperature	0.65	41.3	-1.46	-2.27	0.37	0.2
Low-High Temperature	0.20	42.5	-3.95	-1.18	-0.38	3.0E-04
Oxygen Transfer						
Ancestor-Low Temperature	0.082	10.7	-0.09	-0.19	0.17	0.93
Ancestor-High Temperature	0.083	10.8	-2.39	-0.38	-0.015	0.04
Low-High Temperature	0.034	39	-5.57	-0.26	-0.12	<2.0E-16
Analysis of 'Line' Random Effects:						
	χ^2	**DF**	**P-value**			
Growth Rate	0	1	1			
Biofilm Formation	680	1	<2.0E-16			
Oxygen Transfer	2.06	1	0.2			

Brown-Forsyth test for unequal variances	**F-value**	**DF**	**P-value**
Growth rate	3.43	2	0.042
Biofilm formation	19.76	2	1.08E-06
Oxygen transfer	6.79	2	2.9E-03
Kruskal-Wallis test	χ^2	**DF**	**P-value**
Growth Rate	8.18	2	0.017
Biofilm Formation	107.54	2	< 2.2E-16
Oxygen Transfer	33.37	2	5.7E-08
Pairwise Mann–Whitney U-tests	**Growth rate**	**Biofilm formation**	**Oxygen Transfer**
Ancestor vs. Low Temperature	0.91	3.3E-04	1
Ancestor vs. High Temperature	0.84	1.5E-11	0.023
Low vs. High Temperature	0.012	<2E-16	5.9E-08

Supplementary Table S3.2. Descriptive morphology phenotypes. Phenotypes noted qualitatively from observations. Line 14 had the slightest aggregation so it was included in the table. The plate-like pellicle is as pictured in Figure S3.2. Wrinkly colony formations were observed in the clonal lines. Side film was described from Figure S3.1 as the lines that extended their biofilm deeper than the average depth.

Line	Biofilm	Plate-like pellicle	Wrinkly Colony	Side film
14	1	0	0	0
25	1	1	1	0
26	1	0	0	0
27	0	0	0	0
28	1	1	1	0
29	1	0	0	1
30	1	0	1	0
31	1	0	0	1
32	1	0	0	0
33	1	0	0	0
34	1	0	0	0
35	1	1	1	0
37	1	0	0	1
38	1	1	1	0
39	1	0	0	1
40	1	0	0	1
41	1	0	0	0
42	1	0	0	1
43	1	1	0	0
44	1	1	1	0
45	1	0	0	1
46	1	1	1	0
47	1	0	0	1

Supplementary Table S3.3. Colony morphology observations. Notations are denoted as follows: 'N'-normal, 'W'-wrinkled, 'Ms'-multi-sized, 'S'-small, 'L'-large, 'Wy'-watery, 'D'- smooth dimple, 'A'-over abundant, 'C'-color, '-'- too dry to assess. Some entries have additional notes describing the characteristic.

Line	25	26	27	28	29	30	31	32	33	34	35
31-Jan	Ms	N	N	N	N	N	N	N	N	N	N
8-Feb	Ms	N	N	S mix	N	Ms	Ms	N	N	L mix	N
16-Feb	N	L mix, C-paler	N	S mix	N	W mix	W mix	Wy	N	N	N
2-Mar	-	-	-	-	N	half W	Ms	-	-	-	-
10-Mar	W mix	C-some redder	D	W	N	L mix, W mix	S, Wy	S, Wy	S	S	S
18-Mar	N	mix	D	W	N	mix	D, B-feathery	N	N	mix	W, B-ragged
26-Mar	D	W	N	W	B-fuzzy	W mix, D mix	W mix	D mix	L, W mix	W mix, B	W
3-Apr	W	S, W	N	W	W mix	W	W	N	N	N	N
11-Apr	B-feathery	W mix	D	S, W mix	Ms	W mix	N	S, Wy	W	W mix	D mix
19-Apr	W mix	W mix	Wy	W mix	Ms, B-watery	W mix, B-feathery	N	D	W mix	-	D
27-Apr	L, W	W mix	Ms	W	Ms	W	N	Ms Wy	S	W mix	W mix
5-May	W mix	Ms	Ms	mix	Wy, Ms	W	N	S mix	L	L, C, W	L mix
13-May	S	S, W mix	Ms	D, W	Ms, Wy mix	W	N	S	S, W mix	S, W mix	N
21-May	mix	mix	N	W mix	Ms, B	W	Wy	Ms	N	N	W mix, Wy
29-May	W mix, C	C mix	Ms	W, D mix	Ms, B	W mix, C	N, hard to tell	N, hard to tell	N, hard to tell	N, hard to tell	S mix
4-Jun	L	W mix, C	Ms	W	Ms, Wy	W	S	A	C	N	W mix
14-Jun	A, mix	D, mix	Wy	W, S	Ms, B	W, C	Wy	Wy	D, Wy, A	A	W mix

Line	37	38	39	40	41	42	43	44	45	46	47
31-Jan	L	L	N	N	N	C-dark red ring	N	N	N	N	N
8-Feb	L mix	N	L mix	L mix	Ms	C-red ring	L mix	N	N	one W	S mix
16-Feb	N	N	Ms	N	N	L	W	C-dark red	S mix	W	N
2-Mar	-	-	-	-	-	-	D	N	-	W	N
10-Mar	S, Wy	A	S, Wy	S, Wy	S, Wy	S	S, W mix	S	Wy	W mix	N
18-Mar	S mix	W	N	N	-	W	mix	W mix	D mix	W mix	S mix
26-Mar	N	W mix	-	W	Wy mix	Ms	-	-	Wy, Ms	N	Ms
3-Apr	B-hazy	N	N	N	Ms	N	W mix	Ms	B	Ms	Ms
11-Apr	N	Wy mix	D	N	Ms	S	W, C-some darker red	W mix	Ms	W mix	Ms
19-Apr	Wy mix	-	N mix	Wy mix, S	-	Wy mix	D, C-deep red	W mix	N	N	Ms
27-Apr	N	L	-	L mix	Ms	N	Ms	mix	W mix, Wy mix	W mix	Ms
5-May	N	W	C	Ms, Wy	Ms	N	Ms	-	N	W	Ms
13-May	N	W mix	N	N	Ms	L mix	Ms	Wy	Ms, W mix	W, C-some red	Ms
21-May	A	Wy, Ms	Ms, N	N	Ms	N	Ms	Ms, mix	Bs, L mix, W mix	W, B	Ms
29-May	N, hard to tell	W mix	Ms	Wy	Ms	N	B	N	Ms, W mix	B, C, W mix	N

Supplementary Table S3.4. Stoichiometry variation among experimental groups**.** Particulate organic Carbon, Nitrogen and Phosphorus were assessed, Df-degrees of freedom, Sum Sq- Sum of squares, Mean Sq- Mean Squares, F-value and Pr(>F)-p-value associated with F-value.

	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
C/N	2	0.154	0.07721	0.298	0.744
C/N-residuals	40	10.354	0.25884		
C/P	2	452	226	0.078	0.925
C/P-residuals	40	115423	2886		
N/P	2	57	28.5	0.177	0.838
N/P-residuals	40	6428	160.7		

Supplementary Table S3.5. Genomic changes across lines**.** Genomic changes identified using *Breseq* across low and high temperature adapted lines. Position is given relative to the start of each sequence, shared mutations denote overlapping similar mutations with a running count, short name gives the abbreviated name from Figure 3.5, mutation and bp changed notes the actual change called by *breseq* and information about the mutation position is given in annotation, description gives the RAST gene label or if more than 2 then the number of genes affected, if it is intergenic vs. genic (1 vs 0), and type of change is noted.

Line	Class	Seq.	Position	Final Position	Shared mutation	Short name	Mutation	Bp Change	Description	Inter-genic	Type
L01	Low-Temperature Adapted	Genome	60231	60231	0	NA	C->T	0	Phosphoenolpyruvate_protein phosphotransferase, nitrogen regulation associated	0	SNP
L28	High-Temperature Adapted	Genome	104693	104694	0	NA	(C)7->8	1	Protein export cytoplasm protein SecA ATPase RNA helicase (TC 3.A.5.1.1)	0	Indel
L35	High-Temperature Adapted	Genome	118737	118737	0	NA	+CA	0	Type I secretion target repeat protein	0	Indel
L34	High-Temperature Adapted	Genome	138291	138300	0	NA	(TCGATACCG) 2->3	9	FIG01027979: hypothetical protein	0	Duplication
L44	High-Temperature Adapted	Genome	272092	272093	0	NA	-1bp	1	GTP pyrophosphokinase, (p)ppGpp synthetase II / Guanosine_3',5'_bis(diphosphate) 3'_pyrophosphohydrolase	0	Indel
L44	High-Temperature Adapted	Genome	273928	273928	0	NA	G->A	0	GTP pyrophosphokinase, (p)ppGpp synthetase II / Guanosine_3',5'_bis(diphosphate) 3'_pyrophosphohydrolase	0	SNP
L43	High-Temperature Adapted	Genome	312295	312295	0	NA	G->A	0	Type II/IV secretion system ATPase TadZ/CpaE, associated with Flp pilus assembly	0	SNP
L41	High-Temperature Adapted	Genome	319169	319169	0	NA	C->T	0	Predicted ATPase with chaperone activity, associated with Flp pilus assembly	0	SNP
L07	Low-Temperature Adapted	Genome	396075	396075	0	NA	C->A	0	hypothetical protein	0	SNP
L33	High-Temperature Adapted	Genome	492824	492824	0	NA	G->A	0	Arginine/ornithine antiporter ArcD	0	SNP
L30	High-Temperature Adapted	Genome	535778	535778	1	TrkA domain protein	G->A	0	TrkA domain protein	0	SNP
L40	High-Temperature Adapted	Genome	536224	536238	1	TrkA domain protein	-14bp	14	TrkA domain protein	0	Indel
L10	Low-Temperature Adapted	Genome	537418	537422	1	TrkA domain protein	rearrangement	4	TrkA domain protein/Ser/Thr protein phosphatase family protein	1	IS Insertion
L45	High-Temperature Adapted	Genome	658865	658865	0	NA	G->A	0	Sarcosine dehydrogenase	0	SNP
L34	High-Temperature Adapted	Genome	850113	862126	0	NA	-12013bp	12013	11 genes	0	Large Deletion
L39	High-Temperature Adapted	Genome	850146	850146	0	NA	+AT	0	FIG01027019: hypothetical protein	0	Indel
L47	High-Temperature Adapted	Genome	851335	852594	0	NA	-1259bp	1259	FIG01027019: hypothetical protein/FIG01027843: hypothetical protein	0	Large Deletion
L02	Low-Temperature Adapted	Genome	870894	870895	2	Hypothetical protein	(C)5'-6	1	Hypothetical protein/Identified by similarity to GB:AAO56638.1	1	Intergenic Indel
L19	Low-Temperature Adapted	Genome	870894	870895	2	Hypothetical protein	(C)5->6	1	Hypothetical protein/Identified by similarity to GB:AAO56638.1	1	Intergenic Indel
L10	Low-Temperature Adapted	Genome	870928	870928	2	Hypothetical protein	A->G	0	Hypothetical protein/Identified by similarity to GB:AAO56638.1	1	Intergenic SNP
L32	High-Temperature Adapted	Genome	1051465	1051465	0	NA	A->G	0	FIG01026174: hypothetical protein	0	SNP
L46	High-Temperature Adapted	Genome	1129935	1129935	0	NA	C->T	0	tRNA uridine 5_carboxymethylaminomethyl modification enzyme GidA	0	SNP
L40	High-Temperature Adapted	Genome	1138095	1138095	0	NA	C->T	0	FIG01027707: hypothetical protein	0	SNP
L39	High-Temperature Adapted	Genome	1204523	1204523	0	NA	A->G	0	Na(+) H(+) antiporter subunit A; Na(+) H(+) antiporter subunit B; NADH_ubiquinone oxidoreductase chain L	0	SNP

L11	Low-Temperature Adapted	Genome	1250204	1250204	0	NA	T->G	0	putative Glucose/sorbosone dehydrogenase	0	SNP
L39	High-Temperature Adapted	Genome	1299238	1299238	3	EpsK domain protein	A->C	0	EpsK domain protein	0	SNP
L32	High-Temperature Adapted	Genome	1299752	1299752	3	EpsK domain protein	C->T	0	EpsK domain protein	0	SNP
L46	High-Temperature Adapted	Genome	1299752	1299752	3	EpsK domain protein	C->T	0	EpsK domain protein	0	SNP
L40	High-Temperature Adapted	Genome	1299925	1299925	3	EpsK domain protein	C->T	0	EpsK domain protein	0	SNP
L41	High-Temperature Adapted	Genome	1299925	1299925	3	EpsK domain protein	C->T	0	EpsK domain protein	0	SNP
L26	High-Temperature Adapted	Genome	1329786	1329786	0	NA	G->T	0	FIG140336: TPR domain protein	0	SNP
L42	High-Temperature Adapted	Genome	1333601	1333616	0	NA	-15bp	15	Universal stress protein UspA and related nucleotide_binding proteins	0	Indel
L33	High-Temperature Adapted	Genome	1689989	1689989	0	FIG01023834: hypothetical protein	G->A	0	FIG01023834: hypothetical protein	0	SNP
L42	High-Temperature Adapted	Genome	1706249	1706271	4	FIG01026375: hypothetical protein	-22bp	22	FIG01026375: hypothetical protein	0	Indel
L37	High-Temperature Adapted	Genome	1706412	1706412	4	FIG01026375: hypothetical protein	C->T	0	FIG01026375: hypothetical protein	0	SNP
L31	High-Temperature Adapted	Genome	1706447	1706449	4	FIG01026375: hypothetical protein	(G)5->7	2	FIG01026375: hypothetical protein	0	Indel
L47	High-Temperature Adapted	Genome	1706452	1706458	4	FIG01026375: hypothetical protein	-6bp	6	FIG01026375: hypothetical protein	0	Indel
L43	High-Temperature Adapted	Genome	1706476	1706477	4	FIG01026375: hypothetical protein	-1bp	1	FIG01026375: hypothetical protein	0	Indel
L45	High-Temperature Adapted	Genome	1706541	1706542	4	FIG01026375: hypothetical protein	-1bp	1	FIG01026375: hypothetical protein	0	Indel
L25	High-Temperature Adapted	Genome	1801856	1801856	0	NA	G->A	0	hypothetical protein	0	SNP
L28	High-Temperature Adapted	Genome	1801969	1801969	5	Hypothetical protein/GltI	A->T	0	Hypothetical protein/Glutamate Aspartate periplasmic binding protein precursor GltI	1	Intergenic SNP
L34	High-Temperature Adapted	Genome	1801969	1801969	5	Hypothetical protein/GltI	A->T	0	Hypothetical protein/Glutamate Aspartate periplasmic binding protein precursor GltI	1	Intergenic SNP
L19	Low-Temperature Adapted	Genome	1965758	1965758	0	NA	C->T	0	Periplasmic thiol:disulfide interchange protein DsbA	0	SNP
L32	High-Temperature Adapted	Genome	1969086	1972468	0	NA	-3382bp	3382	6 genes	0	Large Deletion
L45	High-Temperature Adapted	Genome	1969710	1969710	0	NA	C->G	0	Universal stress protein family protein	0	SNP
L13	Low-Temperature Adapted	Genome	2066598	2066598	0	NA	T->G	0	Permeases of the drug/metabolite transporter (DMT) superfamily	0	SNP
L43	High-Temperature Adapted	Genome	2078890	2078890	0	NA	A->G	0	Leucyl_tRNA synthetase	0	SNP
L46	High-Temperature Adapted	Genome	2151392	2151392	0	NA	C->G	0	Signal transduction histidine kinase CheA	0	SNP
L33	High-Temperature Adapted	Genome	2184485	2184486	0	NA	-1bp	1	DNA'-binding response regulator ChvI	0	Indel
L40	High-Temperature Adapted	Genome	2185839	2185839	0	NA	A->G	0	Sensor histidine kinase ChvG	0	SNP
L47	High-Temperature Adapted	Genome	2242058	2286756	6	Large deletion 1	-44,698bp	44698	51 genes	0	Large Deletion
L33	High-Temperature Adapted	Genome	2247091	2256919	6	Large deletion 1	-9828bp	9828	11 genes	0	Large Deletion
L35	High-Temperature Adapted	Genome	2248342	2269831	6	Large deletion 1	-21489bp	21489	26 genes	0	Large Deletion
L27	High-Temperature Adapted	Genome	2252982	2256228	6	Large deletion 1	-3246bp	3246	4 genes	0	Large Deletion
L41	High-Temperature Adapted	Genome	2256111	2256111	7	Transcriptional regulator, LysR family	G->A	0	FIG022886: Transcriptional regulator, LysR family	0	SNP
L38	High-Temperature Adapted	Genome	2256123	2256123	7	Transcriptional regulator, LysR family	C->T	0	FIG022886: Transcriptional regulator, LysR family	0	SNP

L41	High-Temperature Adapted	Genome	2421167	2556479	8	Large duplication	duplication +135312	135312	D_Lactate dehydrogenase, cytochrome c_dependent	0	Duplication
L47	High-Temperature Adapted	Genome	2442994	2442994	0	FIG01027204: hypothetical protein	G->A	0	FIG01027204: hypothetical protein	0	SNP
L30	High-Temperature Adapted	Genome	2539807	3003599	8	Large duplication	duplication +463792	463792	Ferric siderophore transport system, biopolymer transport protein ExbB	0	Duplication
L18	Low-Temperature Adapted	Genome	2628601	2628601	0	Transcriptional regulator, LysR family	A->G	0	Transcriptional regulator, LysR family	0	SNP
L39	High-Temperature Adapted	Genome	2667539	2683879	0	NA	-16340bp	16340	19 genes	0	Large Deletion
L10	Low-Temperature Adapted	Genome	2680031	2680032	9	transcriptional regulator, HTH_3 family	(G)6->7	1	transcriptional regulator, HTH_3 family	0	Indel
L27	High-Temperature Adapted	Genome	2680073	2680074	9	transcriptional regulator, HTH_3 family	(T)7->6	1	transcriptional regulator, HTH_3 family	0	Indel
L17	Low-Temperature Adapted	Genome	2680433	2680434	9	transcriptional regulator, HTH_3 family	(T)6->5	1	transcriptional regulator, HTH_3 family	0	Indel
L46	High-Temperature Adapted	Genome	2828587	2828588	0	NA	rearrangement	1	Sensory histidine protein kinase/ FIG01329746: membrane protein	1	IS Insertion
L35	High-Temperature Adapted	Genome	2868855	2868855	0	NA	G->A	0	NAD_dependent formate dehydrogenase alpha subunit	0	SNP
L38	High-Temperature Adapted	Genome	2873196	2873204	0	NA	(GAGGGGAT)1->2	8	Conserved domain protein	0	Duplication
L09	Low-Temperature Adapted	Genome	2911221	2911221	0	NA	G->A	0	putative facilitator of salicylate uptake	0	SNP
L38	High-Temperature Adapted	Genome	2945791	2945792	0	NA	-1bp	1	Methyltransferase type 12/Mobile element protein	1	Intergenic Indel
L08	Low-Temperature Adapted	Genome	2980376	2980376	0	NA	A->G	0	Ammonium transporter/Hypothetical protein	1	Intergenic SNP
L27	High-Temperature Adapted	Genome	2997522	2997522	10	C4_dicarboxylate transport DctBD	C->T	0	C4_dicarboxylate transport sensor protein DctB	0	SNP
L38	High-Temperature Adapted	Genome	2997522	2997522	10	C4_dicarboxylate transport DctBD	C->T	0	C4_dicarboxylate transport sensor protein DctB	0	SNP
L44	High-Temperature Adapted	Genome	2998324	2998324	10	C4_dicarboxylate transport DctBD	C->T	0	C4_dicarboxylate transport sensor protein DctB	0	SNP
L33	High-Temperature Adapted	Genome	3000184	3000184	10	C4_dicarboxylate transport DctBD	C->T	0	C4_dicarboxylate transport transcriptional regulatory protein DctD	0	SNP
L35	High-Temperature Adapted	Genome	3000184	3000184	10	C4_dicarboxylate transport DctBD	C->T	0	C4_dicarboxylate transport transcriptional regulatory protein DctD	0	SNP
L47	High-Temperature Adapted	Genome	3000184	3000184	10	C4_dicarboxylate transport DctBD	C->T	0	C4_dicarboxylate transport transcriptional regulatory protein DctD	0	SNP
L41	High-Temperature Adapted	Genome	3077582	3079621	0	NA	duplication +2039	2039	FIG01026924: hypothetical protein	0	Duplication
L31	High-Temperature Adapted	Genome	3081659	3082610	11	Transcript. reg. w/ AraC domain	-951bp	951	PQQ-dependent oxidoreductase, gdhB family/Transcriptional regulator containing an amidase domain and an AraC-type DNA-binding HTH domain	0	Large Deletion
L08	Low-Temperature Adapted	Genome	3082129	3082129	11	Transcript. reg. w/ AraC domain	G->A	0	Transcriptional regulator containing an amidase domain and an AraC_type DNA_binding HTH domain	0	SNP
L01	Low-Temperature Adapted	Genome	3082135	3082143	11	Transcript. reg. w/ AraC domain	8bp (ATCAACAGG) 1->2	8	Transcriptional regulator containing an amidase domain and an AraC_type DNA_binding HTH domain	0	Duplication
L13	Low-Temperature Adapted	Genome	3082147	3082147	11	Transcript. reg. w/ AraC domain	C->T	0	Transcriptional regulator containing an amidase domain and an AraC_type DNA_binding HTH domain	0	SNP
L19	Low-Temperature Adapted	Genome	3082147	3082147	11	Transcript. reg. w/ AraC domain	C->T	0	Transcriptional regulator containing an amidase domain and an AraC_type DNA_binding HTH domain	0	SNP
L05	Low-Temperature Adapted	Genome	3082163	3082163	11	Transcript. reg. w/ AraC domain	G->A	0	Transcriptional regulator containing an amidase domain and an AraC_type DNA_binding HTH domain	0	SNP
L07	Low-Temperature Adapted	Genome	3082163	3082163	11	Transcript. reg. w/ AraC domain	G->A	0	Transcriptional regulator containing an amidase domain and an AraC_type DNA_binding HTH domain	0	SNP
L09	Low-Temperature Adapted	Genome	3082163	3082163	11	Transcript. reg. w/ AraC domain	G->A	0	Transcriptional regulator containing an amidase domain and an AraC_type DNA_binding HTH domain	0	SNP
L18	Low-Temperature Adapted	Genome	3082163	3082163	11	Transcript. reg. w/ AraC domain	G->A	0	Transcriptional regulator containing an amidase domain and an AraC_type DNA_binding HTH domain	0	SNP
L16	Low-Temperature Adapted	Genome	3082225	3082225	11	Transcript. reg. w/ AraC domain	T->G	0	Transcriptional regulator containing an amidase domain and an AraC_type DNA_binding HTH domain	0	SNP
L02	Low-Temperature Adapted	Genome	3082575	3082575	11	Transcript. reg. w/ AraC domain	A->T	0	Transcriptional regulator containing an amidase domain and an AraC_type DNA_binding HTH domain	0	SNP

L03	Low-Temperature Adapted	Genome	3082647	3082648	11	Transcript. reg. w/ AraC domain	(C)7->6	1	Transcriptional regulator containing an amidase domain and an AraC_type DNA_binding HTH domain	0	Indel
L04	Low-Temperature Adapted	Genome	3082647	3082648	11	Transcript. reg. w/ AraC domain	(C)7->6	1	Transcriptional regulator containing an amidase domain and an AraC_type DNA_binding HTH domain	0	Indel
L11	Low-Temperature Adapted	Genome	3082647	3082648	11	Transcript. reg. w/ AraC domain	(C)7->6	1	Transcriptional regulator containing an amidase domain and an AraC_type DNA_binding HTH domain	0	Indel
L14	Low-Temperature Adapted	Genome	3082647	3082648	11	Transcript. reg. w/ AraC domain	(C)7->6	1	Transcriptional regulator containing an amidase domain and an AraC_type DNA_binding HTH domain	0	Indel
L26	High-Temperature Adapted	Genome	3082647	3082648	11	Transcript. reg. w/ AraC domain	(C)7->8	1	Transcriptional regulator containing an amidase domain and an AraC_type DNA_binding HTH domain	0	Indel
L29	High-Temperature Adapted	Genome	3082647	3082648	11	Transcript. reg. w/ AraC domain	(C)7->8	1	Transcriptional regulator containing an amidase domain and an AraC_type DNA_binding HTH domain	0	Indel
L30	High-Temperature Adapted	Genome	3082647	3082648	11	Transcript. reg. w/ AraC domain	(C)7->8	1	Transcriptional regulator containing an amidase domain and an AraC_type DNA_binding HTH domain	0	Indel
L37	High-Temperature Adapted	Genome	3082647	3082648	11	Transcript. reg. w/ AraC domain	(C)7->8	1	Transcriptional regulator containing an amidase domain and an AraC_type DNA_binding HTH domain	0	Indel
L32	High-Temperature Adapted	Genome	3082657	3082657	11	Transcript. reg. w/ AraC domain	C->A	0	Transcriptional regulator containing an amidase domain and an AraC_type DNA_binding HTH domain	0	SNP
L17	Low-Temperature Adapted	Genome	3082955	3083017	11	Transcript. reg. w/ AraC domain	-62bp	62	Transcriptional regulator containing an amidase domain and an AraC-type DNA-binding HTH domain	0	Large Deletion
L20	Low-Temperature Adapted	Genome	3082955	3083017	11	Transcript. reg. w/ AraC domain	-62bp	62	Transcriptional regulator containing an amidase domain and an AraC-type DNA-binding HTH domain	0	Large Deletion
L23	Low-Temperature Adapted	Genome	3082955	3083017	11	Transcript. reg. w/ AraC domain	-62bp	62	Transcriptional regulator containing an amidase domain and an AraC-type DNA-binding HTH domain	0	Large Deletion
L39	High-Temperature Adapted	Genome	3105432	3105595	12	D-beta--hydroxybutyrate dehydrogenase/ABC transporter	-163bp	163	D-beta-hydroxybutyrate dehydrogenase/ ABC transporter, periplasmic substrate-binding protein	0	Large Deletion
L34	High-Temperature Adapted	Genome	3105503	3105503	12	D-beta--hydroxybutyrate dehydrogenase/ABC transporter	G->T	0	D-beta-hydroxybutyrate dehydrogenase/ ABC transporter, periplasmic substrate-binding protein	1	Intergenic SNP
L47	High-Temperature Adapted	Genome	3105503	3105503	12	D-beta--hydroxybutyrate dehydrogenase/ABC transporter	G->T	0	D-beta-hydroxybutyrate dehydrogenase/ ABC transporter, periplasmic substrate-binding protein	1	Intergenic SNP
L43	High-Temperature Adapted	Genome	3110006	3110006	0	NA	+T	0	Histidine ABC transporter, ATP_binding protein HisP	0	Indel
L28	High-Temperature Adapted	Genome	3150977	3150978	13	CTP synthase	-1bp	1	Preprotein translocase subunit SecG/ CTP synthase	1	Intergenic Indel
L45	High-Temperature Adapted	Genome	3150984	3150984	13	CTP synthase	C->A	0	Preprotein translocase subunit SecG/ CTP synthase	1	Intergenic SNP
L26	High-Temperature Adapted	Genome	3152232	3152232	13	CTP synthase	C->T	0	CTP synthase	0	SNP
L35	High-Temperature Adapted	Genome	3152232	3152232	13	CTP synthase	C->T	0	CTP synthase	0	SNP
L38	High-Temperature Adapted	Genome	3152583	3152583	13	CTP synthase	C->T	0	CTP synthase	0	SNP
L37	High-Temperature Adapted	Genome	3215320	3216724	14	Type I secretion target repeat protein	-1,404bp	1404	Type I secretion target repeat protein/Type I secretion target repeat protein	0	Large Deletion
L25	High-Temperature Adapted	Genome	3215370	3216396	14	Type I secretion target repeat protein	-1,026bp	1026	Type I secretion target repeat protein/Type I secretion target repeat protein	0	Large Deletion
L32	High-Temperature Adapted	Genome	3215734	3216781	14	Type I secretion target repeat protein	-1,047 bp	1047	Type I secretion target repeat protein/Type I secretion target repeat protein	0	Large Deletion
L35	High-Temperature Adapted	Genome	3215903	3216761	14	Type I secretion target repeat protein	-858bp	858	Type I secretion target repeat protein/Type I secretion target repeat protein	0	Large Deletion
L28	High-Temperature Adapted	Genome	3216335	3216665	14	Type I secretion target repeat protein	-330bp	330	Type I secretion target repeat protein	0	Large Deletion
L45	High-Temperature Adapted	Genome	3281202	3281205	0	NA	-3bp	3	Anhydro_N_acetylmuramic acid kinase	0	Indel
L45	High-Temperature Adapted	Genome	3284738	3284738	0	NA	G->A	0	Shikimate kinase I	0	SNP
L43	High-Temperature Adapted	Genome	3341520	3341671	0	NA	-151bp	151	Transcriptional activator of maltose regulon, MalT	0	Large Deletion
L46	High-Temperature Adapted	Genome	3367298	3440797	15	Large deletion 2	-73,499bp	73499	58 genes	0	Large Deletion
L31	High-Temperature Adapted	Genome	3419931	3445891	15	Large deletion 2	-25960bp	25960	21 genes	0	Large Deletion

L43	High-Temperature Adapted	Genome	3425217	3445345	15	Large deletion 2	-20128bp	20128	16 genes	0	Large Deletion
L32	High-Temperature Adapted	Genome	3427776	3444958	15	Large deletion 2	-17182bp	17182	16 genes	0	Large Deletion
L42	High-Temperature Adapted	Genome	3428067	3440184	15	Large deletion 2	-12117bp	12117	11 genes	0	Large Deletion
L41	High-Temperature Adapted	Genome	3428680	3441438	15	Large deletion 2	-12758bp	12758	11 genes	0	Large Deletion
L30	High-Temperature Adapted	Genome	3440347	3440450	15	Large deletion 2	-103bp	103	FIG01058452: hypothetical protein	0	Large Deletion
L47	High-Temperature Adapted	Genome	3447942	3447943	0	NA	-1bp	1	Mannose_1_phosphate guanylyltransferase (GDP)	0	Indel
L26	High-Temperature Adapted	Genome	3450552	3450552	16	RNA_binding protein Hfq	G->C	0	RNA_binding protein Hfq	0	SNP
L29	High-Temperature Adapted	Genome	3450552	3450552	16	RNA_binding protein Hfq	G->C	0	RNA_binding protein Hfq	0	SNP
L37	High-Temperature Adapted	Genome	3450552	3450552	16	RNA_binding protein Hfq	G->C	0	RNA_binding protein Hfq	0	SNP
L40	High-Temperature Adapted	Genome	3450552	3450552	16	RNA_binding protein Hfq	G->C	0	RNA_binding protein Hfq	0	SNP
L39	High-Temperature Adapted	Genome	3450567	3450567	16	RNA_binding protein Hfq	G->A	0	RNA_binding protein Hfq	0	SNP
L31	High-Temperature Adapted	Genome	3450804	3450804	16	RNA_binding protein Hfq	C->G	0	RNA-binding protein Hfq/Potassium uptake protein TrkH	1	Intergenic SNP
L31	High-Temperature Adapted	Genome	3505901	3505901	17	LysR family/Metallopeptidase	C->A	0	Transcriptional regulator, LysR family/Metallopeptidase, family M24	1	Intergenic SNP
L42	High-Temperature Adapted	Genome	3505907	3505907	17	LysR family/Metallopeptidase	G->A	0	Transcriptional regulator, LysR family/Metallopeptidase, family M24	1	Intergenic SNP
L35	High-Temperature Adapted	Genome	3525269	3525269	0	NA	C->T	0	DNA_binding protein, putative	0	SNP
L46	High-Temperature Adapted	Genome	3553376	3553376	0	NA	A->C	0	glutamine synthetase family protein	0	SNP
L38	High-Temperature Adapted	Genome	3562378	3562378	18	LuxRI	G->A	0	Transcriptional activator protein LuxR	0	SNP
L35	High-Temperature Adapted	Genome	3562513	3562521	18	LuxRI	-8bp	8	Transcriptional activator protein LuxR	0	Indel
L25	High-Temperature Adapted	Genome	3562570	3562571	18	LuxRI	-1bp	1	Transcriptional activator protein LuxR	0	Indel
L30	High-Temperature Adapted	Genome	3562668	3562673	18	LuxRI	(CATAT)1->2	5	Transcriptional activator protein LuxR	0	Duplication
L46	High-Temperature Adapted	Genome	3562769	3562772	18	LuxRI	rearrangement	3	Transcriptional activator protein LuxR	0	Rearrangement
L44	High-Temperature Adapted	Genome	3562820	3562820	18	LuxRI	G->A	0	Transcriptional activator protein LuxR	0	SNP
L28	High-Temperature Adapted	Genome	3563133	3563140	18	LuxRI	(CATCAAT)1->2	7	Autoinducer synthesis protein LuxI	0	Duplication
L30	High-Temperature Adapted	Genome	3596467	3596468	0	NA	(C)8->9	1	Gene Transfer Agent host specificity protein	0	Indel
L29	High-Temperature Adapted	Genome	3629072	3629073	0	NA	(C)8->9	1	Agmatinase	0	Indel
L29	High-Temperature Adapted	Genome	3656801	3656801	0	NA	C->T	0	FIG01027544: hypothetical protein	0	SNP
L25	High-Temperature Adapted	Genome	3661003	3661004	19	Epimerase/ dehydratase	(C)8->7	1	Putative epimerase/dehydratase	0	Indel
L26	High-Temperature Adapted	Genome	3661003	3661004	19	Epimerase/ dehydratase	(C)8->7	1	Putative epimerase/dehydratase	0	Indel
L37	High-Temperature Adapted	Genome	3661003	3661004	19	Epimerase/ dehydratase	(C)8->7	1	Putative epimerase/dehydratase	0	Indel
L14	Low-Temperature Adapted	Genome	3661003	3661004	19	Epimerase/ dehydratase	(C)8->9	1	Putative epimerase/dehydratase	0	Indel
L28	High-Temperature Adapted	Genome	3661003	3661004	19	Epimerase/ dehydratase	(C)8->9	1	Putative epimerase/dehydratase	0	Indel
L29	High-Temperature Adapted	Genome	3661003	3661004	19	Epimerase/ dehydratase	(C)8->9	1	Putative epimerase/dehydratase	0	Indel

L31	High-Temperature Adapted	Genome	3661003	3661004	19	Epimerase/ dehydratase	(C)8->9	1	Putative epimerase/dehydratase	0	Indel
L32	High-Temperature Adapted	Genome	3661003	3661004	19	Epimerase/ dehydratase	(C)8->9	1	Putative epimerase/dehydratase	0	Indel
L33	High-Temperature Adapted	Genome	3661003	3661004	19	Epimerase/ dehydratase	(C)8->9	1	Putative epimerase/dehydratase	0	Indel
L35	High-Temperature Adapted	Genome	3661003	3661004	19	Epimerase/ dehydratase	(C)8->9	1	Putative epimerase/dehydratase	0	Indel
L38	High-Temperature Adapted	Genome	3661003	3661004	19	Epimerase/ dehydratase	(C)8->9	1	Putative epimerase/dehydratase	0	Indel
L42	High-Temperature Adapted	Genome	3661003	3661004	19	Epimerase/ dehydratase	(C)8->9	1	Putative epimerase/dehydratase	0	Indel
L44	High-Temperature Adapted	Genome	3661003	3661004	19	Epimerase/ dehydratase	(C)8->9	1	Putative epimerase/dehydratase	0	Indel
L45	High-Temperature Adapted	Genome	3661003	3661004	19	Epimerase/ dehydratase	(C)8->9	1	Putative epimerase/dehydratase	0	Indel
L01	Low-Temperature Adapted	Genome	3748014	3748018	0	NA	rearrangement	4	Hydrogenase transcriptional regulatory protein hoxA	0	IS Insertion
L27	High-Temperature Adapted	Genome	3838408	3838408	20	Carboxyl_terminal protease	C->T	0	Carboxyl_terminal protease	0	SNP
L40	High-Temperature Adapted	Genome	3838408	3838408	20	Carboxyl_terminal protease	C->T	0	Carboxyl_terminal protease	0	SNP
L41	High-Temperature Adapted	Genome	3838408	3838408	20	Carboxyl_terminal protease	C->T	0	Carboxyl_terminal protease	0	SNP
L47	High-Temperature Adapted	Genome	3838525	3838525	20	Carboxyl_terminal protease	C->T	0	Carboxyl_terminal protease	0	SNP
L47	High-Temperature Adapted	Genome	3849803	3849803	0	NA	G->A	0	4_hydroxybenzoate polyprenyltransferase	0	SNP
L41	High-Temperature Adapted	Genome	3901589	3901589	21	Transporter, Major facilitator superfamily	G->A	0	Transporter, Major facilitator superfamily	0	SNP
L45	High-Temperature Adapted	Genome	3902009	3902009	21	Transporter, Major facilitator superfamily	A->G	0	Transporter, Major facilitator superfamily	0	SNP
L35	High-Temperature Adapted	Genome	3905296	3905296	0	NA	C->T	0	Large exoproteins involved in heme utilization or adhesion	0	SNP
L26	High-Temperature Adapted	Genome	3907571	3907571	0	NA	T->C	0	Thiol peroxidase, Bcp_type	0	SNP
L34	High-Temperature Adapted	Genome	3914865	3914866	0	NA	(C)7->8	1	Type I secretion system, outer membrane component LapE	0	Indel
L43	High-Temperature Adapted	Genome	3916229	3916229	22	T1SS secreted agglutinin RTX	C->T	0	T1SS secreted agglutinin RTX	0	SNP
L29	High-Temperature Adapted	Genome	3916230	3916230	22	T1SS secreted agglutinin RTX	G->A	0	T1SS secreted agglutinin RTX	0	SNP
L37	High-Temperature Adapted	Genome	3916230	3916230	22	T1SS secreted agglutinin RTX	G->A	0	T1SS secreted agglutinin RTX	0	SNP
L39	High-Temperature Adapted	Genome	3916230	3916230	22	T1SS secreted agglutinin RTX	G->A	0	T1SS secreted agglutinin RTX	0	SNP
L29	High-Temperature Adapted	Genome	3930297	3930297	23	FIG01026630: hypothetical protein	T->C	0	FIG01026630: hypothetical protein	0	SNP
L34	High-Temperature Adapted	Genome	3930298	3930298	23	FIG01026630: hypothetical protein	G->C	0	FIG01026630: hypothetical protein	0	SNP
L45	High-Temperature Adapted	Genome	3935145	3935145	0	NA	T->C	0	RNA polymerase sigma factor RpoD	0	SNP
L28	High-Temperature Adapted	Genome	3958178	3958178	0	NA	T->G	0	Scaffold protein for [4Fe_4S] cluster assembly ApbC, MRP_like	0	SNP
L41	High-Temperature Adapted	Plasmid	32440	86974	24	Plasmid deletion	-54534bp	54534	60 genes	0	Large Deletion
L27	High-Temperature Adapted	Plasmid	49802	49805	0	NA	rearrangement	3	FIG01027211: hypothetical protein/FIG01026938: hypothetical protein	1	IS Insertion
L31	High-Temperature Adapted	Plasmid	56600	74435	24	Plasmid deletion	-17835bp	17835	24 genes	0	Large Deletion
L25	High-Temperature Adapted	Plasmid	74589	74589	0	NA	C->T	0	Inner membrane protein/FIG01026951: hypothetical protein	1	Intergenic SNP

Supplementary Table S3.6. Phenotypic trait variation linked to mutations. ANOVA of each multiple regression model of phenotypic variation. Second table has estimation and statistics for each mutation model predictor.

ANOVA on linear model	Biofilm Formation	Oxygen Transfer	Growth Rate
Residual Standard Error	0.33	0.11	0.014
Multiple R-squared	0.88	0.64	0.53
Adjusted R-squared	0.82	0.46	0.3
F-statistic$_{(14,28)}$	14.29	3.43	2.27
P-value	3.3E-09	2.2E-03	0.03

Model estimation								
Biofilm Formation	**Gene ID**	**Estimate**	**Std. Error**	**Df**	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Intercept/Residuals		0.38	0.13	28	3.02	0.11		
TrkA domain protein	peg.541	-0.48	0.25	1	0	0	0.01	0.94
Hypothetical protein	peg.882	0.19	0.22	1	0.54	0.54	5.02	0.03
EpsK domain protein	peg.1293	0.32	0.29	1	1.25	1.25	11.6	2.0E-03
FIG01026375: hypothetical protein	peg.1653	-0.2	0.24	1	0.19	0.19	1.74	0.20
Large deletion 1	-	0.27	0.24	1	0.14	0.14	1.26	0.27
Transcriptional regulator, HTH_3 family	peg.2609	0.04	0.23	1	0.7	0.7	6.53	0.02
Transcriptional regulator w/ AraC domain	peg.2986	0.03	0.14	1	0.43	0.43	3.97	0.06
3-hydroxybutyrate dehydrogenase/ ABC transporter	peg.3009/peg.3010	-0.04	0.24	1	0.42	0.42	3.85	0.06
CTP synthase	peg.3059	-0.13	0.2	1	2.84	2.84	26.29	2.0E-05
Type I secretion target repeat protein	peg.3126,peg.3127	0.63	0.2	1	6.07	6.07	56.24	3.6E-08
Large deletion 2	-	0	0.23	1	0.22	0.22	2	0.17
RNA binding protein Hfq	peg.3345	1.09	0.19	1	3.45	3.45	31.95	4.7E-06
LuxRI	peg.3454, peg.3455	1.45	0.21	1	5.22	5.22	48.37	1.5E-07
Carboxyl terminal protease	peg.3725	-0.29	0.26	1	0.14	0.14	1.29	0.27
Oxygen Transfer	**Gene ID**	**Estimate**	**Std. Error**	**Df**	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Intercept/Residuals		1.04	0.04	28	0.34	0.01		
TrkA domain protein	peg.541	0.04	0.08	1	0.01	0.01	0.86	0.36
Hypothetical protein	peg.882	0	0.07	1	0.03	0.03	2.79	0.11
EpsK domain protein	peg.1293	0.05	0.1	1	0.09	0.09	7.58	0.01
FIG01026375: hypothetical protein	peg.1653	0.02	0.08	1	0.03	0.03	2.74	0.11
Large deletion 1	-	-0.01	0.08	1	0.07	0.07	6.14	0.02
Transcriptional regulator, HTH_3 family	peg.2609	-0.07	0.08	1	0.02	0.02	1.41	0.25
Transcriptional regulator w/ AraC domain	peg.2986	-0.11	0.05	1	0.17	0.17	14.38	0
3-hydroxybutyrate dehydrogenase/ ABC transporter	peg.3009/peg.3010	0.08	0.08	1	7.9E-04	7.9E-04	0.06	0.8
CTP synthase	peg.3059	0.04	0.07	1	0.04	0.04	3.13	0.09
Type I secretion target repeat protein	peg.3126, peg.3127	0.17	0.07	1	0.10	0.10	8.35	0.01
Large deletion 2	-	-0.01	0.08	1	4.8E-04	4.8E-04	0.04	0.84
RNA binding protein Hfq	peg.3345	0.01	0.06	1	4.0E-05	3.5E-05	2.9E-03	0.96
LuxRI	peg.3454, peg.3455	0.05	0.07	1	3.9E-03	4.0E-03	0.32	0.57
Carboxyl terminal protease	peg.3725	0.11	0.09	1	0.02	0.02	1.66	0.21
Growth Rate	**Gene ID**	**Estimate**	**Std. Error**	**Df**	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Intercept/Residuals		0.11	5.8E-03	28	5.6E-03	2.0E-04		
TrkA domain protein	peg.541	4.5E-03	0.01	1	2.0E-04	2.0E-04	0.98	0.33
Hypothetical protein	peg.882	7.5E-04	9.6E-03	1	1.1E-04	1.1E-04	0.53	0.47
EpsK domain protein	peg.1293	3.4E-03	0.01	1	3.8E-04	3.8E-04	1.92	0.18
FIG01026375: hypothetical protein	peg.1653	-7.5E-03	0.01	1	3.5E-05	3.5E-05	0.18	0.68
Large deletion 1	-	-1.7E-03	0.01	1	5.1E-05	5.1E-05	0.26	0.62
Transcriptional regulator, HTH_3 family	peg.2609	2.5E-03	0.01	1	3.3E-05	3.3E-05	0.16	0.69
Transcriptional regulator w/ AraC domain	peg.2986	-9.7E-03	6.2E-03	1	1.1E-03	1.1E-03	5.52	0.03
3-hydroxybutyrate dehydrogenase/ ABC transporter	peg.3009/peg.3010	-1.6E-03	0.01	1	2.1E-04	2.1E-04	1.03	0.32
CTP synthase	peg.3059	7.9E-03	8.5E-03	1	1.5E-03	1.5E-03	7.28	0.01
Type I secretion target repeat protein	peg.3126,peg.3127	0.02	8.5E-03	1	2.1E-03	2.1E-03	10.52	0
Large deletion 2	-	6.9E-04	9.8E-03	1	1.4E-06	1.4E-06	0.01	0.93
RNA binding protein Hfq	peg.3345	0.01	8.0E-03	1	4.5E-04	4.5E-04	2.25	0.15
LuxRI	peg.3454, peg.3455	-7.4E-03	9.1E-03	1	1.5E-04	1.5E-04	0.75	0.39
Carboxyl terminal protease	peg.3725	7.2E-03	0.01	1	8.3E-05	8.3E-05	0.42	0.52

Supplementary Table S3.7. Mutation type differences between *Roseovarius sp.* TM1035 and *E. coli* REL1206. Average number of mutations for each mutation class per high temperature adapted line as described in Table S3.4 and calculated from data provided in Tenaillon *et al.*[2]. Lower and upper confidence intervals and p-value associated with two-tailed Welch's t-test.

Mutation Class	RTM1035 Mean	REL 1206 Mean	Lower CI	Upper CI	P-value
SNPs	3.36	7.31	-4.98	-2.92	1.0E-11
Indels	1.82	2.46	-1.22	-0.07	0.029
IS-elements	0.14	0.56	-0.68	-0.16	2.3E-03
Duplications	0.32	0.22	-0.20	0.40	0.49
Large Deletions	1.14	1.02	-0.33	0.56	0.59

Supplementary Table S3.8. Overlapping mutations between *Roseovarius sp.* TM1035 and *E. coli* REL1206. Mutations in high temperature adapted *Roseovarius sp.* TM1035 lines and *E. coli* REL1206 that overlap in orthologous groups. Whether the mutation was genic, intergenic, or both was noted, then gene ID, gene product description, and the number of high-temperature lines with the gene mutated, followed by the same information for *E. coli* REL1206.

Roseovarius sp. TM1035 genic or intergenic	*Roseovarius sp.* TM1035 gene ID	*Roseovarius sp.* TM1035 gene product	*Roseovarius sp.* TM1035 lines affected	*Roseovarius sp.* TM1035 types	*E. coli* REL1206 genic or intergenic	*E. coli* REL1206 IDs	*E. coli* REL1206 gene product	Lines affected	Types
Genic	peg.3345	RNA_binding protein Hfq	6	SNP	Genic	peg.4161	RNA-binding protein Hfq	3	Duplication
Genic	peg.2896, peg.2986	Transcriptional regulator containing an amidase domain and an AraC-type DNA-binding HTH domain	6	Duplication, Indel, Large Deletion, SNP	Genic	peg.4046	Regulatory protein SoxS	2	Duplication
Genic	peg.3327	GDP-L-fucose synthetase (EC 1.1.1.271)	6	Large Deletion	Genic	peg.1992	GDP-L-fucose synthetase (EC 1.1.1.271)	1	Point mutation
Genic	peg.3725	Carboxyl_terminal protease (EC 3.4.21.102)	4	SNP	Genic	peg.1831	Tail-specific protease precursor (EC 3.4.21.102)	1	Point mutation
Genic	peg.3325	Mobile element protein	2	Large Deletion	Genic	peg.1540	Mobile element protein	8	Duplication, Large deletion, Point mutation
Genic	peg.2783	NAD_dependent formate dehydrogenase alpha subunit	2	Duplication, SNP	Genic	peg.4067	Formate dehydrogenase H (EC 1.2.1.2) @ selenocysteine-containing	6	Point mutation, Duplication
Genic	peg.3986	Cytochrome c-type biogenesis protein DsbD, protein-disulfide reductase (EC 1.8.1.8)	2	Large Deletion	Genic	peg.4128	Cytochrome c-type biogenesis protein DsbD, protein-disulfide reductase (EC 1.8.1.8)	5	Duplication
Genic	peg.2540	Type I restriction-modification system, restriction subunit R (EC 3.1.21.3)/Mobile element protein	2	Duplication, Large Deletion	Genic	peg.4350	Type I restriction-modification system, restriction subunit R (EC 3.1.21.3)	3	Point mutation, Duplication
Genic	peg.2477	Glycyl-tRNA synthetase beta chain (EC 6.1.1.14)	2	Duplication	Intergenic	peg.3501	Glycyl-tRNA synthetase beta chain (EC 6.1.1.14)	2	IS insertion
Intergenic	peg.1756	Hypothetical protein/Glutamate Aspartate periplasmic binding protein precursor GltI (TC 3.A.1.3.4)	2	SNP	Intergenic	peg.3199	Glutamate Aspartate periplasmic binding protein precursor GltI (TC 3.A.1.3.4)	1	Point mutation
Genic	peg.2637	DNA-directed RNA polymerase beta subunit (EC 2.7.7.6)	1	Duplication	Genic	peg.3964	DNA-directed RNA polymerase beta subunit (EC 2.7.7.6)	76	Point mutation, Deletion
Genic	peg.2750	DNA-binding heavy metal response regulator	1	Duplication	Genic	peg.536	Copper-sensing two-component system response regulator CusR	44	Large deletion
Genic	peg.2689	Aspartate aminotransferase (EC 2.6.1.1)	1	Duplication	Genic	peg.572	Methionine aminotransferase, PLP-dependent	38	Large deletion
Genic	peg.3816	RNA polymerase sigma factor RpoD	1	SNP	Genic	peg.3001	RNA polymerase sigma factor RpoD	30	Point mutation, Deletion, Insertion

Genic	peg.2638	DNA-directed RNA polymerase beta' subunit (EC 2.7.7.6)	1	Duplication	Genic	peg.3965	DNA-directed RNA polymerase beta' subunit (EC 2.7.7.6)	21	Point mutation, Deletion
Genic	peg.2904	Methyltransferase corrinoid protein	1	Duplication	Genic	peg.3995	5-methyltetrahydrofolate--homocysteine methyltransferase (EC 2.1.1.13)	15	Point mutation, Deletion, Insertion, Large deletion, Duplication
Genic	peg.871	Capsule polysaccharide export inner-membrane protein	1	Large Deletion	Genic	peg.2878	Capsular polysaccharide export system inner membrane protein KpsE	11	Deletion, Insertion, Point mutation, IS insertion
Genic	peg.2823	Homoserine O-succinyltransferase (EC 2.3.1.46)	1	Duplication	Genic	peg.3988	Homoserine O-succinyltransferase (EC 2.3.1.46)	9	Point mutation, Deletion, Duplication
Genic	peg.873	STRUCTURAL ELEMENTS; Cell Exterior; surface polysaccharides/antigens	1	Large Deletion	Genic	peg.2889	Capsular polysaccharide ABC transporter, permease protein KpsM	7	Point mutation, Deletion
Genic	peg.2769	Aspartate ammonia-lyase (EC 4.3.1.1)	1	Duplication	Genic	peg.4131	Aspartate ammonia-lyase (EC 4.3.1.1)	7	Point mutation, Duplication
Genic	peg.2661	cAMP-binding proteins - catabolite gene activator and regulatory subunit of cAMP-dependent protein kinases	1	Duplication	Genic	peg.3280	Cyclic AMP receptor protein	6	Point mutation, Duplication
Genic	peg.2747	FIG00805073: hypothetical protein	1	Duplication	Genic	peg.4074	FIG00638848: hypothetical protein	5	Duplication
Genic	peg.2624	Translation elongation factor Tu	1	Duplication	Intergenic, Genic	peg.3957, peg.3261	Translation elongation factor Tu	4	Point mutation, Duplication
Genic	peg.4007	Phosphonate ABC transporter permease protein phnE (TC 3.A.1.9.1)	1	Large Deletion	Genic	peg.4094	Phosphonate ABC transporter permease protein phnE (TC 3.A.1.9.1)	4	Duplication
Genic	peg.2558	Alkylphosphonate utilization operon protein PhnA	1	Duplication	Genic	peg.4098	Alkylphosphonate utilization operon protein PhnA	4	Duplication
Genic	peg.3739	4_hydroxybenzoate polyprenyltransferase (EC 2.5.1.39)	1	SNP	Genic	peg.4020	4-hydroxybenzoate polyprenyltransferase (EC 2.5.1.39)	3	Point mutation, Duplication
Genic	peg.2725, peg.2714	Sensory box/GGDEF family protein, Pole remodelling regulatory diguanylate cyclase	1	Duplication	Genic	peg.1478	Putative Heme-regulated two-component response regulator	2	Point mutation
Genic	peg.3985	Transcriptional regulator, MerR family	1	Large Deletion	Genic	peg.3214	HTH-type transcriptional regulator zntR	2	Duplication
Genic	peg.870	Capsular polysaccharide ABC transporter, ATP-binding protein KpsT	1	Large Deletion	Genic	peg.2888	Capsular polysaccharide ABC transporter, ATP-binding protein KpsT	2	Point mutation, Deletion
Genic	peg.2853	Sensory box histidine kinase/response regulator	1	Duplication	Intergenic	peg.3145	Aerobic respiration control sensor protein arcB (EC 2.7.3.-)	2	Duplication
Genic	peg.2850	Outer membrane protein assembly factor YaeT precursor	1	Duplication	Genic	peg.4214	Uncharacterized protein YtfM precursor	2	Duplication
Genic	peg.2849	FIG01027224: hypothetical protein	1	Duplication	Genic	peg.4215	Uncharacterized protein YtfN	2	Duplication

Genic	peg.2797	Replicative DNA helicase (EC 3.6.1.-)	1	Duplication	Genic	peg.4033	Replicative DNA helicase (EC 3.6.1.-)	2	Duplication
Genic	peg.2681	Anthranilate synthase, amidotransferase component (EC 4.1.3.27) @ Para-aminobenzoate synthase, amidotransferase component (EC 2.6.1.85)	1	Duplication	Genic	peg.3283	Para-aminobenzoate synthase, amidotransferase component (EC 2.6.1.85)	2	Duplication
Genic	peg.2539	Type I restriction-modification system, DNA-methyltransferase subunit M (EC 2.1.1.72)	1	Duplication	Genic	peg.4349	Type I restriction-modification system, DNA-methyltransferase subunit M (EC 2.1.1.72)	2	Point mutation, Duplication
Genic	peg.2444	Zinc uptake regulation protein ZUR	1	Duplication	Genic	peg.4026	Zinc uptake regulation protein ZUR	2	Duplication
Genic	peg.2228	tRNA dihydrouridine synthase A (EC 1.-.-.-)	1	Large Deletion	Genic	peg.4030	tRNA dihydrouridine synthase A	2	Duplication
Genic	peg.2226	hypothetical protein	1	Large Deletion	Genic	peg.4239	Hypothetical protein	2	Duplication
Genic	peg.2222	NAD(P)H oxidoreductase YRKL (EC 1.6.99.-) @ Putative NADPH-quinone reductase (modulator of drug activity B) @ Flavodoxin 2	1	Large Deletion	Genic	peg.3274	Glutathione-regulated potassium-efflux system ancillary protein KefG	2	Duplication
Genic	peg.2676	5-formyltetrahydrofolate cyclo-ligase (EC 6.3.3.2)	1	Duplication	Genic	peg.2802	5-formyltetrahydrofolate cyclo-ligase (EC 6.3.3.2)	1	Point mutation
Genic	peg.2763	Branched-chain amino acid ABC transporter, amino acid-binding protein (TC 3.A.1.4.1)	1	Duplication	Genic	peg.3386	High-affinity leucine-specific transport system, periplasmic binding protein LivK (TC 3.A.1.4.1)	1	Duplication
Genic	peg.4012	Regulatory protein, ArsR	1	Large Deletion	Genic	peg.3434	Arsenical resistance operon repressor	1	Duplication
Genic	peg.3523	Agmatinase (EC 3.5.3.11)	1	Indel	Genic	peg.2828	Agmatinase (EC 3.5.3.11)	1	Point mutation
Genic	peg.3282	Transcriptional regulator, LysR family	1	Large Deletion	Genic	peg.2743	Transcriptional activator protein LysR	1	Deletion
Genic	peg.3192	Shikimate kinase I (EC 2.7.1.71)	1	SNP	Genic	peg.3315	Shikimate kinase I (EC 2.7.1.71)	1	Deletion
Genic	peg.2914	Ribonuclease E (EC 3.1.26.12)	1	Duplication	Genic	peg.1101	Ribonuclease E (EC 3.1.26.12)	1	Point mutation
Genic	peg.2879	Branched-chain amino acid transport ATP-binding protein LivF (TC 3.A.1.4.1)	1	Duplication	Genic	peg.3382	Branched-chain amino acid transport ATP-binding protein LivF (TC 3.A.1.4.1)	1	Duplication
Genic	peg.2878	Branched-chain amino acid transport ATP-binding protein LivG (TC 3.A.1.4.1)	1	Duplication	Genic	peg.3383	Branched-chain amino acid transport ATP-binding protein LivG (TC 3.A.1.4.1)	1	Duplication
Genic	peg.2842	ATPase component BioM of energizing module of biotin ECF transporter	1	Duplication	Genic	peg.4254	FIG00638157: hypothetical protein	1	Duplication
Genic	peg.2835	Hydrogen peroxide-inducible genes activator	1	Duplication	Genic	peg.3946	Hydrogen peroxide-inducible genes activator	1	Duplication
Genic	peg.2829	putative facilitator of salicylate uptake	1	Duplication	Genic	peg.2304	Long-chain fatty acid transport protein	1	Point mutation
Genic	peg.2635	LSU ribosomal protein L7/L12 (P1/P2)	1	Duplication	Intergenic	peg.3963	LSU ribosomal protein L7/L12 (P1/P2)	1	Point mutation

Genic	peg.2571	6-carboxytetrahydropterin synthase (EC 4.1.2.50) @ Queuosine biosynthesis QueD, PTPS-I	1	Duplication	Intergenic	peg.2661	Queuosine biosynthesis QueD, PTPS-I	1	Point mutation
Genic	peg.2538	Type I restriction-modification system, specificity subunit S (EC 3.1.21.3)	1	Duplication	Genic	peg.4348	Type I restriction-modification system, specificity subunit S (EC 3.1.21.3)	1	Duplication
Genic	peg.2511	D-alanyl-D-alanine carboxypeptidase (EC 3.4.16.4)	1	Duplication	Intergenic	peg.3117	D-alanyl-D-alanine carboxypeptidase (EC 3.4.16.4)	1	Deletion
Genic	peg.2502	Signal recognition particle receptor protein FtsY (=alpha subunit) (TC 3.A.5.1.1)	1	Duplication	Genic	peg.3393	Signal recognition particle receptor protein FtsY (=alpha subunit) (TC 3.A.5.1.1)	1	Duplication
Genic	peg.2490	Phospho-N-acetylmuramoyl-pentapeptide-transferase (EC 2.7.8.13)	1	Duplication	Genic	peg.84	Phospho-N-acetylmuramoyl-pentapeptide- transferase (EC 2.7.8.13)	1	Point mutation
Genic	peg.2484	Cell division protein MraZ	1	Duplication	Genic	peg.78	Cell division protein MraZ	1	Point mutation
Genic	peg.2447	2-isopropylmalate synthase (EC 2.3.3.13)	1	Duplication	Intergenic	peg.72	2-isopropylmalate synthase (EC 2.3.3.13)	1	Point mutation
Genic	peg.2445	Zinc ABC transporter, periplasmic-binding protein ZnuA	1	Duplication	Genic	peg.1859	Zinc ABC transporter, periplasmic-binding protein ZnuA	1	Point mutation

Supplementary Table S3.9. Sequencing and mapping information. Lines were sequenced either paired-end or single-end then mapped using *Breseq* (Deatherage & Barrick, 2014), percentage of R1-mapped reads and R2-mapped reads using *breseq,* and counts for each library are noted. LA1, LA2, and LA3 are replicate ancestor libraries.

Line	Paired-End (PE) or Single-End (SE)	% R1-mapped	% R2-mapped	R1 Read count	R2 Read count
L01	PE	91	88.8	10,546,674	10,550,566
L02	SE	96.5	-	10,822,129	-
L03	PE	90.3	87.9	5,885,051	5,887,392
L04	SE	95.3	-	13,772,376	-
L05	PE	91.2	89.1	9,754,857	9,758,280
L06	SE	95.5	-	11,782,269	-
L07	PE	85.8	84.2	9,381,640	9,385,140
L08	SE	92	-	9,333,413	-
L09	PE	91.3	87.8	3,583,303	3,584,646
L10	SE	96.1	-	11,188,384	-
L11	PE	85.9	84.6	11,113,144	11,117,251
L13	PE	87	85.3	13,613,559	13,618,511
L14	SE	96.4	-	11,538,852	-
L15	PE	87.7	86	6,755,213	6,757,538
L16	SE	93.4	-	10,002,648	-
L17	PE	85.2	84.1	16,515,277	16,521,546
L18	SE	93	-	8,038,162	-
L19	PE	86.6	85.3	8,201,872	8,205,031
L20	SE	95.3	-	5,812,158	-
L23	PE	88	86.6	8,888,194	8,891,323
L25	PE	97.5	86	13,949,443	13,954,726
L26	SE	93.3	-	9,227,286	-
L27	PE	87.2	85.7	8,583,765	8,586,623
L28	SE	93.8	-	10,541,311	-
L29	PE	88.2	86.4	8,624,410	8,627,611
L30	SE	95.4	-	7,599,254	-
L31	PE	88.2	86.3	8,518,219	8,521,292
L32	SE	95.5	-	10,509,104	-
L33	PE	87	85.3	9,362,871	9,366,342
L34	SE	92.2	-	10,505,506	-
L35	PE	86.3	84.3	8,621,096	8,624,340
L37	PE	85.2	83.3	11,679,228	11,683,490
L38	SE	96.4	-	8,134,582	-
L39	PE	86.1	83.8	10,161,427	10,165,164
L40	SE	96.9	-	10,256,491	-
L41	PE	85	82.7	6,683,775	6,685,994
L42	SE	96.8	-	10,955,734	-
L43	PE	84.7	82.9	9,131,999	9,135,512
L44	SE	95.1	-	9,745,746	-
L45	PE	85	82.9	8,364,889	8,368,202
L46	SE	96.4	-	9,954,523	-
L47	SE	93.8	-	12,636,826	-
LA1	PE	88.1	85.9	15,454,971	14,794,918
LA2	SE	90.8	-	10,132,198	-
LA3	PE	85.5	83.4	7,671,361	7,674,043

Supplementary Table S3.10. Genome break points**.** Genomic break points determined by REAPR(Hunt *et al.*, 2013). Observations of whether or not the break point was in a repetitive region or looked valid—few to no reads overlapping region were determined by manual inspection using CLC Genomics Workbench and assessed with a mate-pair library previously generated during the original *Roseovarius sp.* TM1035 sequencing project(Miller & Belas, 2004) generally had at least 4 reads mapped ±30 bp around proposed REAPR break points with several paired read sets flanking (i.e. junction is in middle of paired read insert).

Plasmid	Repetitive Region	Validity	Mate-pair valid (reads mapped, paired read flank)
73636	0	0	0 (4/7)
Genome			
677305	1	1	0 (12/3)
2309192	0	1	0 (5/5)
2332553	0	1	0 (5/7)
2572924	0	1	0 (6/3)
2643597	0	1	0 (6/6)
2756190	0	0	0 (5/2)
3001121	0	0	0 (4/4)
3405277	0	1	0 (5/4)
3551972	0	1	0 (5/5)
3759693	0	1	0 (6/6)
4012695	1	1	0 (6/7)

Supplementary Table S3.11. Mutation Correlations. Pearson correlations and p-values associated with them using "corr.test" from R package "psych". Correlation coefficients in lower left triangle, p-values in upper right triangle.

	TrkA domain protein	Hyp. protein	EpsK domain protein	FIG01026375 hyp. protein	Large deletion 1	Transcriptional regulator, HTH-3 family	C4-dicarboxylate transport DctBD	Transcriptional regulator w/ AraC domain	3-beta-hydroxybutyrate dehydrogenase/ ABC transporter	CTP synthase	Type I secretion target repeat protein	Large deletion 2	RNA binding protein Hfq	LuxRI	Epimerase/ dehydratase	Carboxyl terminal protease	T1SS secreted agglutinin RTX
TrkA domain protein	-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Hypothetical protein	0.28	-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
EpsK domain protein	0.19	-0.1	-	1	1	1	1	1	1	1	1	0.53	1	1	1	1	1
FIG01026375: hypothetical protein	-0.11	-0.11	-0.15	-	1	1	1	1	1	1	1	1	1	1	1	1	1
Large deletion 1	-0.09	-0.09	-0.12	0.1	-	1	2.6E-08	1	1	1	1	1	1	1	1	0.34	1
Transcriptional regulator, HTH-3 family	0.28	0.28	-0.1	-0.11	0.23	-	1	1	1	1	1	1	1	1	1	1	1
C4-dicarboxylate transport DctBD	-0.11	-0.11	-0.15	0.03	0.8	0.15	-	0.5	1	1	1	1	1	1	1	1	1
Transcript. reg. w/ AraC domain	-0.11	0.07	-0.24	-0.16	-0.34	-0.11	-0.43	-	1	1	1	1	1	1	1	1	1
3-beta-hydroxybutyrate dehydrogenase/ABC transporter	-0.08	-0.08	0.19	0.15	0.23	-0.08	0.15	-0.29	-	1	1	1	1	1	1	1	1
CTP synthase	-0.1	-0.1	-0.13	0.06	0.13	-0.1	0.27	-0.24	-0.1	-	1	1	1	0.53	0.04	1	1
Type I secretion target repeat protein	-0.1	-0.1	0.09	0.06	0.13	-0.1	0.06	-0.1	-0.1	0.32	-	1	1	0.53	0.04	1	1
Large deletion 2	0.13	-0.12	0.43	0.37	-0.14	-0.12	-0.18	-0.09	-0.12	-0.16	0.04	-	1	1	1	1	1
RNA binding protein Hfq	0.15	-0.11	0.27	0.23	-0.13	-0.11	-0.16	0.11	0.15	0.06	0.06	0	-	1	1	1	0.01
LuxRI	0.13	-0.12	0.04	-0.18	0.08	-0.12	0.37	-0.35	-0.12	0.43	0.43	0.15	-0.18	-	1	1	1
Epimerase/dehydratase	-0.19	-0.19	-0.1	0.29	0.12	-0.19	0.29	-0.15	-0.19	0.52	0.52	0.1	0.29	0.37	-	1	1
Carboxyl terminal protease	0.23	-0.09	0.38	0.1	0.45	0.23	0.33	-0.34	0.23	-0.12	-0.12	0.08	0.1	-0.14	-0.22	-	1
T1SS secreted agglutinin RTX	-0.09	-0.09	0.13	0.33	-0.1	-0.09	-0.13	-0.02	0.23	-0.12	0.13	0.08	0.56	-0.14	0.12	-0.1	-