COMPARATIVE GENOMIC ANALYSES OF YELLOWSTONE HOT SPRING MICROBIAL MAT *SYNECHOCOCCUS* SPP.

by

Millie Helen Thornton Olsen

A thesis submitted in partial fulfillment
of the requirements for the degree

of

Master of Science

in

Land Resources and Environmental Sciences

MONTANA STATE UNIVERSITY
Bozeman, Montana

April 2015

©COPYRIGHT

by

Millie Helen Thornton Olsen

2015

All Rights Reserved

DEDICATION

To my parents, who nurtured my love of science and helped me believe I could do anything that I set my mind to. To my husband, who supported me along this journey. And to my grandpa, the ecologist, with whom I would have loved to discuss this research.

ACKNOWLEDGEMENTS

This research was largely supported by the U.S. Department of Energy (DOE) through (i) a Community Sequencing Project from the Joint Genome Institute (proposal #CSP148; user agreement NPUSR006316), and (ii) The Office of Biological and Environmental Research (BER), as part of BER′s Genomic Science Program 395 (GSP). This contribution originates from the GSP Foundational Scientific Focus Area (FSFA) at the Pacific Northwest National Laboratory (PNNL) under contract 112443. We appreciate support from the Montana Agricultural Experiment Station (project 911352). This study was conducted under Yellowstone National Park research permits YELL-0129 and 5494 (D.M.W.) and YELL-02058 (M.K.), and we appreciate the assistance from National Park Service personnel. I would also like to thank the members of the Ward Lab, past and present, who have been an integral part of this research: Eric Becraft, Shane Nowack, Jason Wood, George Schaible, Katrina Jackson, and Chris Klatt. I would especially like to thank my advisor, Dr. Dave Ward, for introducing me to the fascinating world of microbial ecology, hiring me as an undergraduate research assistant, and then as a laboratory technician, convincing me to complete a Master's degree, giving me a tremendous project to work on, and mentoring me along the way.

TABLE OF CONTENTS

1. INTRODUCTION 1

Hot Spring *Synechococcus* Mat Communities as Models for Studying Species 4
Comparative Genomic Analyses 14
Main Aims of this Thesis 19

2. CULTIVATION OF *SYNECHOCOCCUS* SPP. FOR COMPARATIVE GENOMIC AND PHENOTYPIC ANALYSIS 21

Introduction 21
Methods 23
Cultivation of *Synechococcus* Isolates 23
Culture Genotyping 24
DNA Extraction for Culture Purity Testing 25
Culture Purity Testing 26
Culture Preservation 27
Results 27
Cultivation of *Synechococcus* Isolates 27
Culture Purity Testing 28
Discussion 30

3. THE EFFECTS OF RECOMBINATION ON PHYLOGENETIC INFERENCE BASED ON SINGLE-LOCUS, MULTI-LOCUS, AND GENOMIC SEQUENCE ANALYSES 33

Introduction 33
Methods 35
Orthologous-Gene Alignment and Genome Phylogeny Construction 35
Single-Locus Sequence Analysis of Genomic DNA Sequences 36
MLSA of Genomic DNA Sequences 36
Results 38
Orthologous-Gene Alignment and Genome Phylogeny Construction 38
Investigating Incongruence between Single-Gene and Genomic Phylogenies 38
Impacts of Recombining Loci on MLSA Phylogenies 40
Discussion 43

TABLE OF CONTENTS – CONTINUED

4. COMPARATIVE GENOMICS OF *SYNECHOCOCCUS* ISOLATES WITH DIFFERENT LIGHT RESPONSES AND *IN SITU* DIEL TRANSCRIPTION PATTERNS OF ASSOCIATED PUTATIVE ECOTYPES IN THE MUSHROOM SPRING MICROBIAL MAT 46

Contribution of Authors and Co-Authors 46
Manuscript Information Page 48
Abstract 49
Introduction 49
Materials and Methods 54
Synechococcus Isolates 54
Genome Sequencing, Assembly, and Annotation 54
Comparative Analyses 56
Metatranscriptomic Analyses 56
Results 57
Genomic Properties 57
Genes Found Only in Low-Light-Adapted Isolates 61
Genes Found Only in High-Light-Adapted Isolates 66
PsbA Allele Differences between High-Light- and Low-Light-Adapted Isolates 68
Other Gene Content Differences among Isolates 69
Discussion 71
Acknowledgements 79

5. COMPARATIVE GENOMICS OF MULTIPLE *SYNECHOCOCCUS* ISOLATES FOUND IN DIFFERENT TEMPERATURE AND LIGHT ENVIRONMENTS OF HOT SPRINGS MICROBIAL MATS 80

Introduction 80
Methods 81
Synechococcus Isolates 81
Genome Sequencing, Assembly, and Annotation 82
Comparative Analyses 82
PCR Amplification of *nifD* and *nifK* 83
Results 84
Genomic Information 84
Gene Content Differences 89
Differences among Genomes of Lineage B′, A, and A′ Defined by 16S rRNA 89
Differences among Isolates Representative of Different PEs in the B′- and A-Lineages 91

TABLE OF CONTENTS – CONTINUED

Differences among Replicate Isolates Representative of PEs A1 and A14 93
Sites under Positive Selection 94
Differences among Genomes of Lineage B′, A, and A′ Defined by 16S rRNA 94
Differences among Isolates Representative of Different PEs in the B′- and A-Lineages 94
Differences among Replicate Isolates Representative of PEs A1 and A14 96
Genes under Positive Selection 112
Differences among Genomes of Lineage B′, A, and A′ Defined by 16S rRNA 112
Differences among Isolates Representative of Different PEs in the B′- and A-Lineages 113
Differences among Replicate Isolates Representative of PEs A1 and A14 113
Discussion 114

6. CONCLUSIONS 121

Cultivation and Genome Sequencing of *Synechococcus* Isolates 122
Single-Locus, Multi-Locus, and Whole Genome Phylogenies 124
Comparative Analyses of Isolates Representative of PEs with Different Vertical Niches 126
Comparative Analyses of Multiple Isolates Found in Different Temperature and Light Environments 128
Comparative Analyses of Genomes of Lineage B′, A, and A′ Defined by 16S rRNA 128
Comparative Analyses of Isolates Representative of Different PEs in the B′- and A-Lineages 129
Comparative Analyses of Replicate Isolates Representative of PEs A1 and A14 130
Overarching Conclusions 131

REFERENCES CITED 133

APPENDICES 150

APPENDIX A: Supplementary Information for Chapter 4 151
APPENDIX B: Supplementary Information for Chapter 5 157

LIST OF TABLES

Table Page

2.1. Summary of *psaA* Ti-454 or MiSeq analysis of *Synechococcus* isolate purity....29

3.1. Genes chosen from genome-informed analysis for MLSA 42

4.1. Genomic, phenotypic, and environmental information for *Synechococcus* isolates of putative ecotypes with different depth distributions or from different hot spring mats....58

4.2. Average nucleotide identity and genes with identical amino acid sequence in *Synechococcus* isolate genomes....59

4.3. Ortholog table showing discussed gene content differences among isolates with different light adaptations and representative of putative ecotypes with different vertical positioning in the 60-63°C Mushroom Spring mat 60

5.1. Genomic, phenotypic, and environmental information for *Synechococcus* isolates representative of putative ecotypes with different depth distributions or from different hot spring mats....85

5.2. Amino acid and codon usage bias between A'/A-lineage isolates and B'-lineage isolates. 86

5.3. Pairwise percent nucleotide identity comparisons among isolate genomes. 87

5.4. Gene content differences among *Synechococcus* isolates. 90

5.5. Pairwise comparisons among isolate genomes. 97

5.6. Sites under positive selection with the A',A//B' mutation pattern 98

5.7. Sites under positive selection with the A//A',B' mutation pattern 100

5.8. Sites under positive selection with the A',A//B'-OS,B'12//B'5 mutation pattern 101

LIST OF TABLES – CONTINUED

Table Page

5.9. Sites under positive selection with the A',A,B'5 //B'-OS,B'12 mutation pattern 104

5.10. Sites under positive selection within the A-lineage isolate genomes.....106

5.11. Comparative genomic analyses of *Synechococcus* isolates with regards to different ecological parameters.....115

A.1. Orthologous genes among *Synechococcus* isolates .....found at http://scholarworks.montana.edu/xmlui/handle/1/9071

B.1. Comprehensive list of sites under positive selection within the *Synechococcus* isolate genomes 158

B.2. Comprehensive list of sites under positive selection within the *Synechococcus* isolate genomes 168

LIST OF FIGURES

Figure Page

1.1. The Stable Ecotype Model....4

1.2. Mushroom Spring source pool and microbial mat in the effluent channel....5

1.3. Microbial mat communities of Mushroom Spring and Octopus Spring....5

1.4. Distance matrix phylogenetic tree based on cyanobacterial 16S rRNA sequences....7

1.5. Microbial mats collected from 60, 65, and 68°C sites in Mushroom spring displaying different mat phenotype and *Synechococcus* cell fluorescence intensity....8

1.6. *Synechococcus* A-lineage phylogenies based on maximum-likelihood analysis of seven concatenated loci and *rbsK* sequences from environmental BAC clones....11

1.7. Relative abundance of predominant putative ecotypes in ~80 µm-thick subsections along the vertical gradient in 60-63°C mat samples....12

1.8. Diel transcription and change in irradiance and oxygen concentration with depth in the 60°C Mushroom Spring mat....13

1.9. Growth rates of *Synechococcus* isolates representative of predominant PEs as a function of irradiance....14

1.10. The Nano-Niche Model....18

2.1. A'/A-lineage and B'-lineage *Synechococcus* phylogenies based on environmental *psaA* barcode sequences....22

2.2. Floating filter plates showing discrete colony growth....24

3.1. Maximum-likelihood tree of A-lineage isolates based on 1136 aligned and concatenated orthologous genes....38

LIST OF FIGURES – CONTINUED

Figure Page

3.2. Maximum-likelihood trees of A-lineage genome isolate full length and amplified region *psaA* sequences 39

3.3. Alignment of isolate *psaA* sequences 39

3.4. Maximum-likelihood trees of seven concatenated MLSA loci from environmental BAC clones and A-lineage isolates with and without the *rbsK* locus 41

3.5. Maximum-likelihood phylogeny of genome-informed MLSA of A-lineage isolates 42

4.1. Phylogeny based on a concatenation of 460 proteins of A-lineage isolates conserved in the cyanobacterial phylum. 61

4.2. Transcription patterns of transcripts in A-lineage isolates 63

4.3. *rpsbA* diel transcription patterns 69

5.1. Maximum-likelihood tree of *Synechococcus* spp. isolate whole genome alignment using 1136 aligned and concatenated orthologous genes 88

5.2. B′-lineage site selection mutation patterns 95

A.1. Diel transcription patterns of PE A1-specific genes 154

A.2. Diel transcription patterns of assimilatory nitrate reductase duplicate copy in PEs A4 and A14 155

A.3. Growth of the *Synechococcus* isolate of PE A1-MS strain as a function of light intensity and differences in forms of dissolved inorganic carbon supplied 156

ABSTRACT

The question of "What is a microbial species?" has been a highly debated issue in the field of microbiology. Many have accepted a molecular species demarcation approach, that any two organisms with a high enough 16S rRNA sequence similarity are members of the same species. However, the Ward lab has shown that there are many ecologically distinct *Synechococcus* spp. inhabiting hot springs of the Lower Geyser Basin in Yellowstone National Park, WY, that would be defined as members of the same species using the molecular demarcation approach. Using a theory-based species demarcation approach with a conserved photosystem gene (*psaA*), evidence of the existence of putative ecotypes, or predicted ecologically distinct species, has been found in the microbial mat, distributed along both temperature and light gradients. Isolates representative of these ecologically distinct populations have also been shown to have distinct temperature adaptations and light adaptations. I obtained the genomes of these isolates, which include representatives of populations with different temperature distributions and different vertical distributions, and replicate isolates within individual putative ecotypes. Using these genome sequences, I compared the *psaA* gene phylogeny and multi-locus sequence phylogenies with a phylogeny created using genes shared among the genomes to explore the effects of recombination on phylogenies of closely-related organisms. I then explored the underlying genetic mechanisms of the niche adaptations of these ecotypes by (i) comparing the isolate gene content, diel transcription patterns, and positive selection evidence of putative ecotypes with different vertical distributions in the mat and different light adaptations, (ii) comparing the gene content and evidence of positive selection among isolates representative of populations with different temperature distributions, and (iii) comparing the gene content and evidence of positive selection among replicate isolates within individual putative ecotypes. I found that, while recombination may have caused the inaccurate demarcation of genetically distinct isolates into a single PE, there is genomic evidence that species of *Synechococcus* that are ecologically distinct from one another exist, along both temperature and vertical gradients. Members of a species are ecologically homogenous, though there is evidence of some genetic heterogeneity within a species.

CHAPTER ONE

INTRODUCTION

A major question has arisen in modern microbiology: what defines a microbial species? Species demarcation of most eukaryotic organisms can be guided by the Biological Species Concept (Mayr, 1942), which asserts that species are groups of individuals that may reproduce with one another to produce fertile offspring and are reproductively isolated from other groups. Microorganisms reproduce by asexual reproduction, so the criteria of the Biological Species Concept may not be directly applicable for microbial species demarcation, although support for application of the Biological Species Concept to microbial speciation has recently been claimed (e.g. Cadillo-Quiroz et al., 2012; Ellegaard et al., 2013). Traditionally, microbiologists classified microorganisms by observing microbial populations using the only tools at their disposal at the time – laboratory cultivation and phenotypic characterization of strains of the organisms. Molecular analyses challenged these classification methods based on physiology and morphology. Woese and Fox (1977) were among the first to use the highly conserved 16S ribosomal RNA (rRNA) sequence to measure genetic relatedness, or phylogeny, among different organisms. As the use of 16S rRNA sequencing became widespread, microbiologists began to discover the misclassifications caused by phenotyping. For example, strains of bacteria isolated from the rumina of cows and sheep and the ceca of rats and pigs were all classified as *Bacteroides succinogenes* because they produced the same expected fermentation products, acetate and succinate,

and displayed the gram-negative, anaerobic, non-motile, non-spore-forming, rod-shaped phenotype of members of the genus *Bacteroides*. Using 16S rRNA sequencing, Montgomery and colleagues (1988) showed that strains of *B. succinogenes* were only as related to *Bacteroides fragilis*, the type species of the genus *Bacteroides*, as they were to representatives of other bacterial phyla (between 70-78% 16S rRNA sequence similarity). Moreover, they demonstrated that the rumenal strains were phylogenetically and phenotypically distinct from the cecal strains, which were only ~92% similar to the rumenal strains at the 16S rRNA locus and had different growth medium requirements. The phylum Fibrobacter was created based on these results, and the rumenal and cecal strains were classified as *Fibrobacter succinogenes* and *Fibrobacter intestinalis*, respectively, two species phylogenetically distinct from *Bacteroides* and the phylum Bacteroidetes. 16S rRNA sequence analysis influenced classification at all taxonomic levels, as investigators confirmed that the past grouping of cyanobacteria with eukaryotic algae was inaccurate (Doolittle et al., 1975) and revealed a new domain of life, Archaea, whose members were previously combined with Bacteria into the kingdom Monera (Balch et al., 1977). These results eventually lead to a shift in bacterial taxonomy to include phylogenetic relatedness as a criterion for strain description and naming (Wayne et al., 1987).

The use of 16S rRNA sequencing has led to an approach to the molecular demarcation of microbial species that has been accepted and used by most microbiologists. This approach was created by observing the amount of 16S rRNA sequence divergence among strains of classically named species, which exhibit 70%

DNA-DNA hybridization (e.g. Johnson, 1973, 1978; Seki et al., 1978; Mays et al., 1982). This approach requires a >3% divergence of the 16S rRNA locus between two organisms (Stackebrandt and Goebel, 1994) or, more recently, >1% divergence at the 16S rRNA locus (Stackebrandt and Ebers, 2006), to consider two closely related strains as belonging to different species. Since these "molecular cutoffs" are a somewhat arbitrary method of species demarcation, and depend on the assumption that named species are truly species, microbiologists have considered alternatives to the Biological Species Concept to develop theory-based models of microbial species and speciation. The ecologist Simpson (1961) had revised the Biological Species Concept to include evolution, stating "an evolutionary species is a lineage…evolving separately from others and with its own unitary evolutionary role and tendencies". This was further modified by van Valen (1976) to incorporate the ecological distinctness of a lineage, resulting in the Ecological Species Concept. Cohan and Perry (2007) incorporated this concept into the Stable Ecotype Model (Figure 1.1). This model defines an ecotype, essentially an ecological species (van Valen, 1976), as "a group of bacteria that are ecologically similar to one another, so similar that genetic diversity within the ecotype is limited by a cohesive force, either periodic selection or genetic drift, or both". More recently, Kopac et al. (2014) defined an ecotype based on two criteria. First, an ecotype is "a phylogenetic group of close relatives that are ecologically very similar". Second, "different ecotypes are predicted to coexist indefinitely as a result of their ecological differences, while lineages within one ecotype are ecologically too homogeneous to allow indefinite coexistence". Although several models of speciation have been proposed that are based on ecological theory (Cohan and

Perry, 2007), for the hot spring microbial mat system studied in this thesis, the Stable Ecotype Model will be used as suggested in Ward and Cohan (2005; see below) (Figure 1.1). The Stable Ecotype Model describes microbial lineages as ecotypes that are created or extinguished at very low rates and are relatively stable in the environment. Each ecotype is subject to periodic selection events, which quash diversity within the ecotype. A niche invasion event occurs when an individual acquires a mutation that allows it to occupy an ecological niche separate from the other members of an ecotype. This individual and its descendants form a new ecotype, which is subject to its own periodic selection events separate from the parent ecotype.

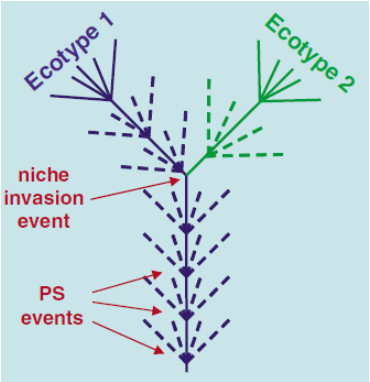


Figure 1.1. The Stable Ecotype Model. The blue line represents a lineage, or ecotype, and other variants within the ecotype that become extinct (dashed lines) as periodic selection (PS) events act on the ecotype. Niche invasion events, caused by acquisition of an ecologically advantageous trait, result in the creation of a second ecotype (green line) (Ward et al., 2008)

Hot Spring *Synechococcus* Mat Communities as Models for Studying Species

The microbial mats of alkaline siliceous hot springs, specifically Mushroom Spring (Figure 1.2) and Octopus Spring, have been studied and their inhabitants



Figure 1.2. Mushroom Spring source pool (background) and microbial mat in the main effluent channel (foreground; red arrow)

characterized for nearly 40 years (Brock, 1978; Ward et al., 2012). The microbial mat communities exclusively contain prokaryotic microorganisms, exist at high temperatures which prohibit predation by eukaryotes, have well established physico-chemical gradients, and are far less complex than microbial communities of other soil and aquatic environments. These qualities make the communities ideal model systems for studying microbial species. The mats are laminated in structure (Figure 1.3), with the upper ~1-

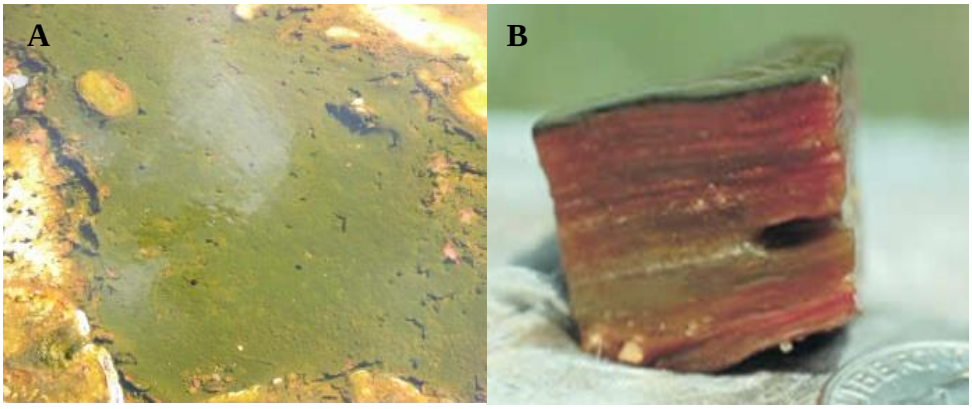


Figure 1.3. Microbial mat communities of Mushroom Spring and Octopus Spring. (A) Photograph of the Mushroom Spring microbial mat in the effluent channel and (B) Cross-section of the Octopus Spring microbial mat, with a dime (lower right-hand corner) for scale (Ward et al., 1998, B only)

mm green oxygenic, phototrophic layer located above a thicker set of underlying orange layers. The mats are constructed by phototrophic microorganisms that begin growing at 72-74°C, which is the upper temperature limit for the cyanobacteria that inhabit the mat, *Synechococcus* spp. Growth of the *Synechococcus* mat continues along the effluent channel, but as the water temperature cools and approaches ambient temperature, multiple genera of other cyanobacteria (*Leptolyngbya* spp. below ~55°C, *Calothrix* spp. below 40°C) are found, along with ephydrid flies that prey on the mat below 42°C (Ward et al., 2012). This study focused on the *Synechococcus* spp. found in the effluent channel between 50 and 70°C.

Based on traditional cultivation and microscopy approaches, *Synechococcus lividus* was once thought to be the only species of cyanobacteria residing in these types of microbial mats (Castenholz, 1973, 1981; Ward et al., 1987). However, molecular analysis of the 16S rRNA locus established the dominance of different populations of *Synechococcus* (designated OS Type A", A', A, B', and B) along the thermal gradient from 70°C to 50°C of the Octopus Spring effluent channel (Ferris and Ward, 1997). Using the >3% 16S rRNA sequence species demarcation approach, the A-type (A", A', and A) and B-type (B' and B) would represent two novel species that were only distantly related to *Synechococcus lividus* isolates (OS C1 Isolate) (see Figure 1.4; Ward et al., 1998). Ferris et al. (1996b) used hybridization probes to prove that, although *S. lividus* was not abundant *in situ* (Ruff-Roberts et al., 1994), it was strongly selected for in enrichment cultures under laboratory conditions. Dilution of mat samples before enrichment culturing yielded *Synechococcus* isolates with OS Type P and OS Type C9

16S rRNA sequences, which were 85.5% and 71.1% similar to *S. lividus* at the 16S rRNA locus, respectively (Figure 1.4). Only the OS Type P 16S rRNA sequence was detected among Octopus Spring environmental 16S rRNA sequences, and neither OS Type P nor OS Type C9 matched the A/B *Synechococcus* sequence types dominating the mat (Ferris et al., 1996a, b). Ward et al. (1997) and Allewalt et al. (2006) were ultimately able to cultivate isolates representative of the predominant 16S rRNA A/B sequence types.

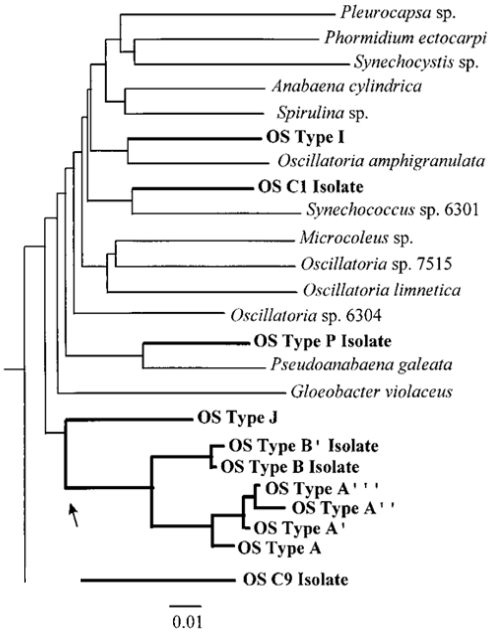


Figure 1.4. Distance matrix phylogenetic tree based on cyanobacterial 16S rRNA sequences. "OS Type" sequences were detected in the Octopus Spring mat or cultures isolated therefrom. Evolutionary distance is indicated by the horizontal line length calibrated by the scale bar, which is equal to 0.01 mutations per sequence position (Ward et al., 1998)

Although the *Synechococcus* 16S rRNA A/B sequence types were only representative of two species according to the 16S rRNA "molecular cutoff", five 16S rRNA genotypes with different temperature distributions were found along the effluent channel (genotype A" at the highest temperatures, followed by genotypes A', A, B, and

finally B' at the lower temperatures) (Ferris et al., 1996a; Ferris and Ward, 1997). Isolates representative of some of these genotypes were shown to have different upper temperature limits for growth: an isolate with the A genotype had a higher upper temperature limit than an isolate with the B' genotype, which had a higher upper temperature limit than an isolate with the B genotype, even though the B and B' genotypes were less than 1% divergent at the 16S rRNA locus and would therefore be considered the same species using the molecular cutoff (Allewalt et al., 2006). Furthermore, there was evidence that physiologically distinct populations were found at different depths along the vertical gradient of the mat. As illustrated by Figure 1.5 (left),

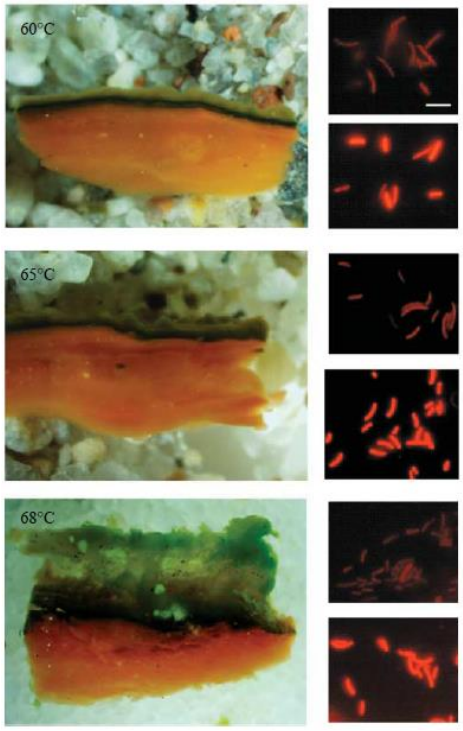


Figure 1.5. Microbial mats collected from 60, 65, and 68°C sites in Mushroom Spring displaying different mat phenotype and *Synechococcus* fluorescence intensity. The photomicrographs to the right of each mat photograph show *Synechococcus* from the upper (top) and lower (bottom) portions of the mat green layer. Mat pictures on the left show difference in mat thickness at different temperatures. Scale bar for photomicrographs is 10 µm (Ward et al., 2006)

the top portion of the mat green layer is yellow green in color while the bottom portion is darker green (Ramsing et al., 2000). Fluorescence microscopy of mat samples suggested that the *Synechococcus* populations associated with these layers were physiologically distinct, since *Synechococcus* cells located in the lower-layer were darker green than the upper layer populations, and were more brightly autofluorescent than those in the upper layers, indicative of a higher concentration of photosynthetic antenna proteins and chlorophyll *a* in the lower layer cells (Figure 1.5, right). Microsensor measurements of oxygen, from which rates of oxygenic photosynthesis can be estimated, also showed difference in net photosynthetic activity between upper and lower populations in the top green layer of the mat. Sequence analysis of the 68°C mat showed that the two populations were identical at the 16S rRNA locus but were genetically distinct at the less highly conserved internal transcribed spacer (ITS) locus that separates the 16S and the 23S rRNA loci (Ferris et al., 2003).

As it became clear that 16S rRNA sequences did not offer enough resolution to demarcate ecologically distinct species, Ward and Cohan (2005) and Ward et al. (2006) began to explore theory-based species prediction models. They determined that studying variation in genes with higher molecular resolution than 16S rRNA and the 16S-23S rRNA ITS region would be essential in the demarcation of these ecological species, and chose to use the Stable Ecotype Model based on the existence of these closely related, but ecologically distinct, populations in the mat, which offers a stable environment with well-defined physico-chemical gradients. An evolutionary simulation algorithm (Ecotype Simulation, ES) was developed to predict the putative ecotypes (PEs), or predicted

ecological species, from sequences within a phylogenetic lineage. ES uses the sequence variation of a lineage and a Monte-Carlo coalescence approach to estimate the number of ecotypes, and the rate and timing of ecotype formation and periodic selection events that best fit the evolutionary history of the lineage (i.e. the observed sequence variation resulting from descent from a single common ancestor). Individual sequences are then demarcated into putative ecotypes based on the most likely parameter values (Koeppel et al., 2008). Melendrez et al. (2011) used ES to analyze the molecular variation of four *Synechococcus* genes, *aroA*, *apcAB*, and *rbsK*, which are conserved protein-encoding loci with differing degrees of molecular resolution. ES predicted 4, 5, and 14 PEs from each gene, respectively, within the 16S rRNA-defined A-lineage, and 8, 13, and 13 PEs within the B'-lineage, respectively. This demonstrated that the number of ecotypes predicted by ES is dependent upon the molecular resolution of the gene used, because analysis of genes with greater sequence diversity resulted in the prediction of more ecotypes within the lineage. Melendrez et al. (in prep) performed multi-locus sequence analysis (MLSA) on environmental bacterial artificial chromosome (BAC) clones, concatenating seven conserved loci to increase molecular resolution compared to the previous single-locus analyses and to buffer against possible effects of recombination (Figure 1.6). They found that ES did predict more PEs within a sample set from MLSA than from single-locus analysis, even when using *rbsK*, the most divergent gene used by Melendrez et al. (2011). However, MLSA can only be accomplished for cultures, BAC clones, or single-amplified genomes used for single-cell genomics, which are not always feasible to obtain for large scale, environmental analyses.

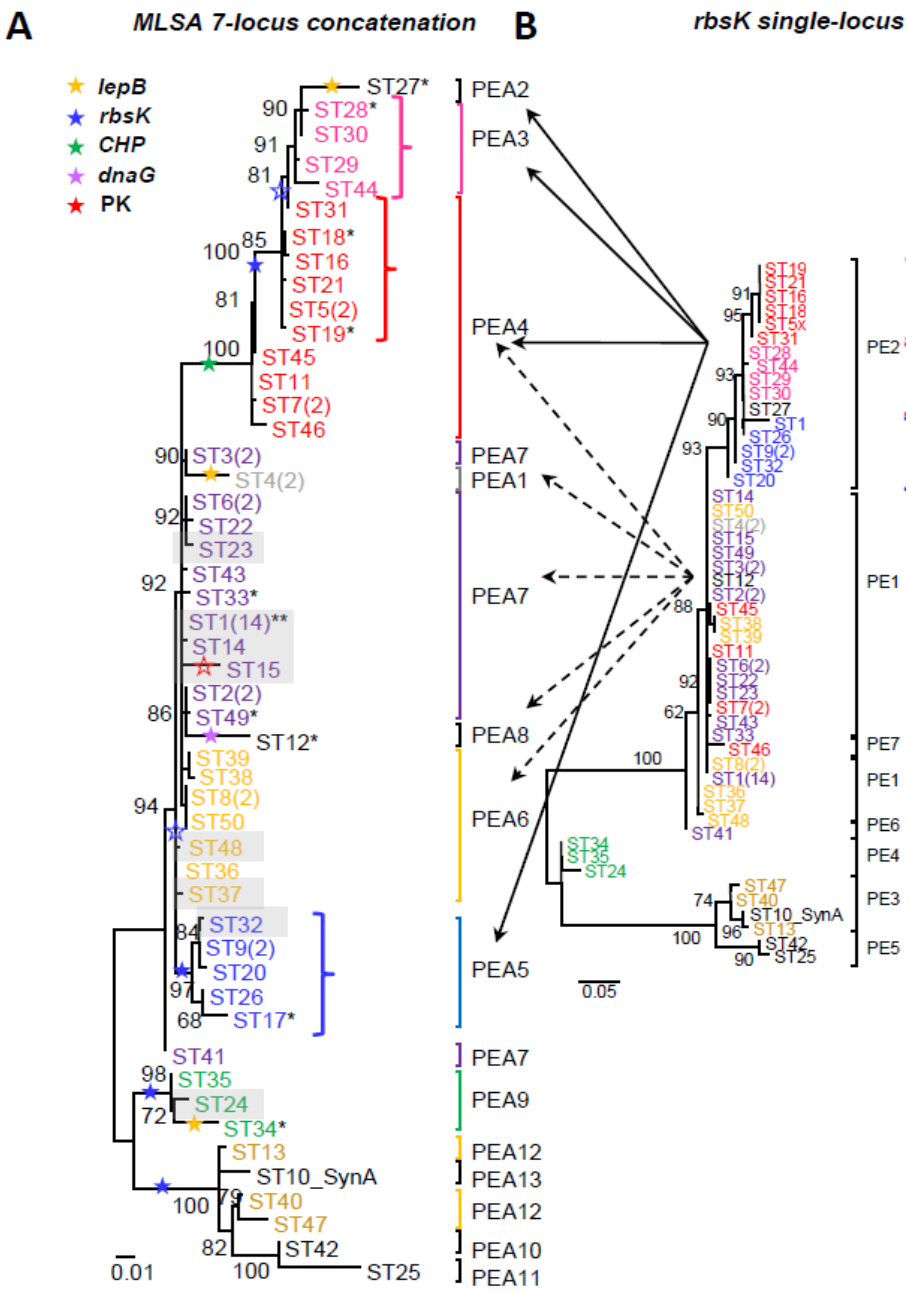


Figure 1.6. *Synechococcus* A-lineage phylogenies based on maximum-likelihood analysis of (A) seven concatenated loci and (B) *rbsK* sequences from environmental bacterial artificial chromosome (BAC) clones. BAC clone sequence types (ST) are color-coded by PE in the multi-locus phylogeny (A) and maintain the same colors in the *rbsK* phylogeny (B). Stars on the tree indicate recombination events in a locus (color coded in the inset). Evolutionary distance is indicated by the horizontal line length calibrated by the scale bar, which is equal to (A) 0.01 mutations per sequence position and (B) 0.05 mutations per sequence position [From Melendrez et al. (in prep)]

Becraft et al. (2011) used ES to analyze sequence variation in the *psaA* gene in the *Synechococcus* spp. inhabiting the Mushroom Spring microbial mat, because *psaA* (which codes for a photosystem I protein subunit) is both highly conserved and one of the most highly transcribed genes *in situ*. ES predicted 9 PEs within the 16S rRNA A'/A-lineage and 12 PEs within the B'-lineage. Denaturing gradient gel electrophoresis of PCR-amplified segments of the *psaA* gene was used to visualize the distribution of separate PEs along both temperature and vertical gradients present in the mat, demonstrating that different PEs predicted by ES had different distributions that could reflect the distinct ecological niche of each PE. Using a deep-coverage, next-generation sequencing method, Becraft et al. (submitted) predicted 22 PEs in the A'/A-lineage and 24 PEs in the B'-lineage, and were able to demonstrate that (i) the distributions of sequences within an ecotype were maximal at distinct temperature and depth (e.g. Figure

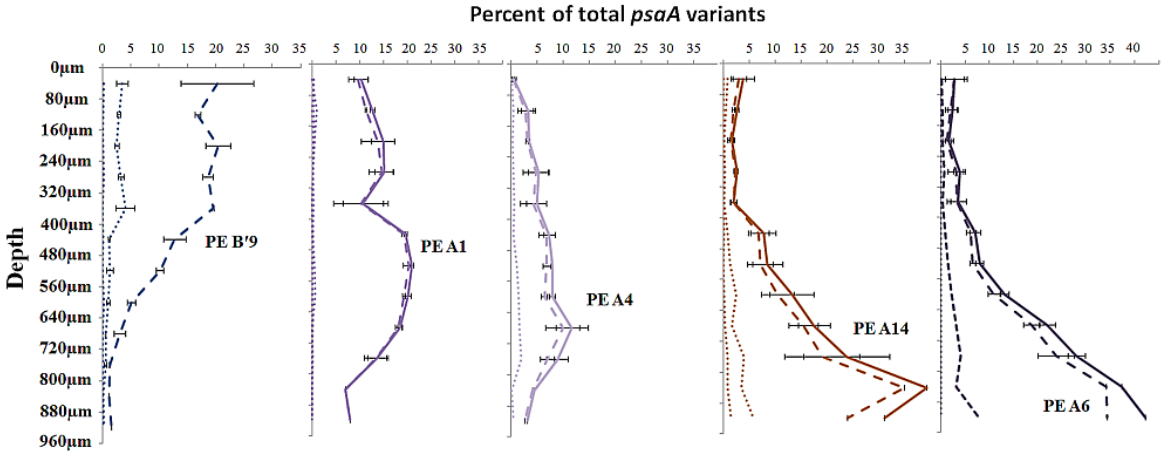


Figure 1.7. Relative abundance of *psaA* sequences representative of predominant putative ecotypes in ~80 µm-thick subsections along the vertical gradient in 60-63°C Mushroom Spring mat samples. Solid line represents all members of a PE, the dashed line represents the number of dominant variant sequences, and dotted lines indicate the number of each additional high-frequency sequence variant (represented >50 times in the sequence data). Error bars indicate range of replicate samples (n=2) [From Becraft et al. (submitted)]

1.7), (ii) members of an ecotype respond similarly to environmental perturbations, and (iii) transcription patterns of genes differ between ecotypes that are located in either the upper (B′-like) or lower (A-like) parts of the 60-63°C mat green layer (Figure 1.8A), with patterns matching the length of the photoperiod experienced by populations in the different layers (Figure 1.8B). Isolates representative of different PEs predicted by

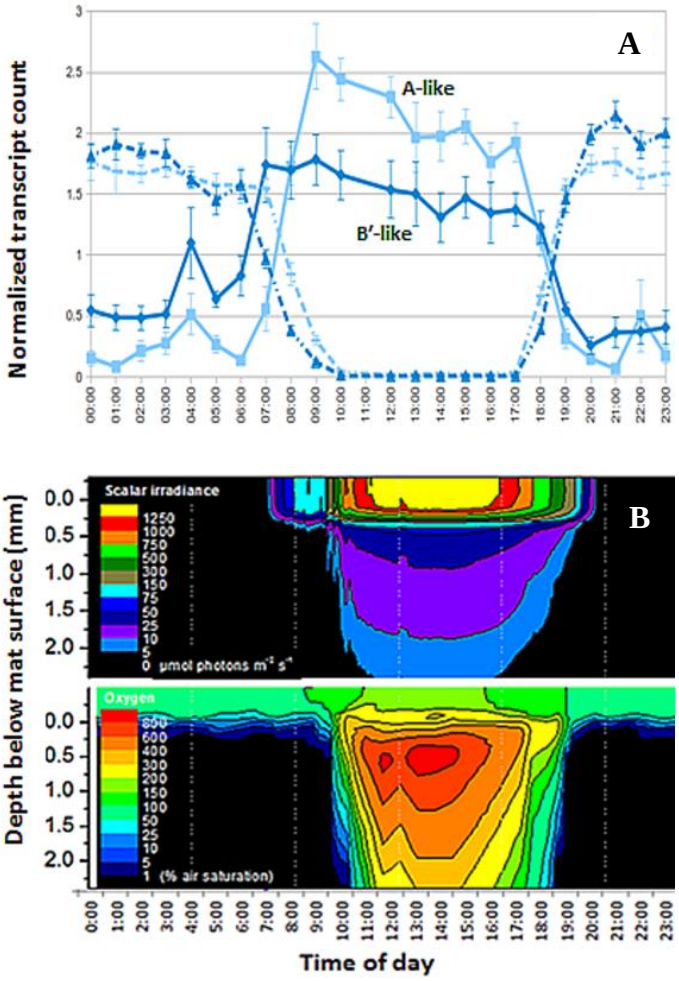


Figure 1.8. Diel transcription and change in irradiance and oxygen concentration with depth in the 60°C Mushroom Spring mat. (A) Normalized count of B′-like (dark blue) and A-like (light blue) photosystem (solid line) and nitrogen fixation (dashed line) transcripts over a diel cycle. (B) Isopleth diagrams showing depth distribution of scalar irradiance (µmol photons m^{-2} s^{-1}) and O_2 concentration (% air saturation) over the same diel cycle [From Becraft et al. (submitted)]

Becraft et al. (2011; submitted) were cultivated and phenotypically characterized by Nowack et al. (submitted), revealing that isolates representative of ecotypes predominant toward the surface of the upper-green layer of the mat (PE A1) had higher light toleranceand faster growth rates at high light intensities (Figure1.9B) than isolates representative of ecotypes located in the lower part of the green layer of the mat (PEs A4 and A14), which had faster growth rates at lower light intensities (Figure 1.9A).

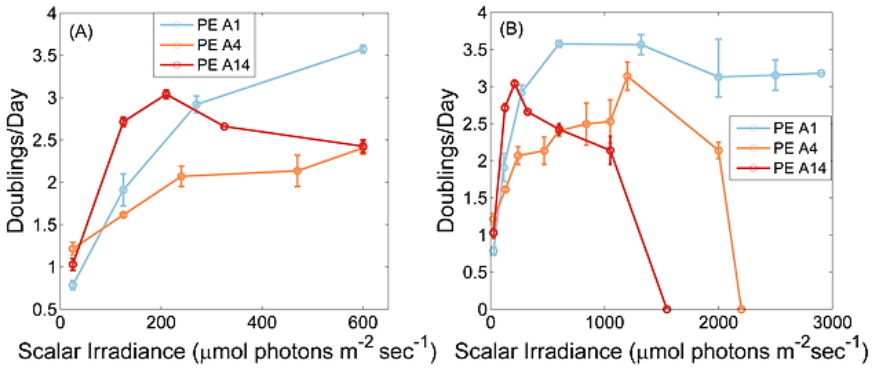


Figure 1.9. Growth rates of *Synechococcus* isolates representative of predominant PEs as a function of irradiance when grown at 60ºC and sparged with 6% CO_2 (v/v) in air. Cultures used in this experiment were pre-grown at 50 µmol photons m^{-2} sec^{-1}. (A) Low irradiances only. (B) All tested irradiances. Each point is the average of 2 measurements and the range is shown by the bars [From Nowack et al. (submitted)]

Comparative Genomic Analyses

Individual genes can be used to predict ecotypes and to infer from their distributions that the ecotypes occupy unique niches within a community, but whole genome comparative analysis can reveal what the mechanisms underlying niche adaptations may be. Since the invention of genome sequencing technologies and methods, especially with newer next-generation sequencing technologies, comparative

genomic analyses have been carried out for many types of microorganisms : (i) pathogens such as *Escherichia coli* (Didelot et al., 2012), *Streptococcus* (Lefebure and Stanhope, 2007), *Haemophilus influenzae* (Hogg et al., 2007), and *Neisseria meningitides* (Schoen et al., 2008), (ii) non-pathogenic bacteria such as environmental *E. coli* (Luo et al., 2011), *Bacillus* (Kopac et al., 2014), *Roseobacter* (Newton et al., 2010), and many members of the phylum Cyanobacteria (Shih et al., 2013), (iii) rhizobia-associated bacteria (Sugawara et al., 2013; Tian et al., 2012), and (iv) thermophiles (Cadillo-Quiroz et al., 2012; Romano et al., 2013)). Recently, there have been novel single-cell genomic analyses (Thrash et al., 2014; Kashtan et al., 2014) that circumvent the need to cultivate and purify strains of microorganisms prior to genome sequencing. Below I will review three examples of comparative genomic analysis that have incorporated information on the environment from which the strains were isolated and phenotypic information on the strain cultures, as those types of comprehensive analysis are most relevant to this project, which incorporates environmental context in the genome analysis.

Perhaps the most comprehensive studies to incorporate an ecological context into comparative genomic analysis have been those for the marine oxygenic phototroph *Prochlorococcus*. Predominating in the surface (top 200 m) waters of oceans between 40°N and 40°S at a population density of 10^4 - 10^5 cells per mL, it is one of the most numerically abundant organisms on Earth (Coleman and Chisholm, 2007). Two major clades of *Prochlorococcus*, designated high-light (HL) and low-light (LL), were discovered to inhabit the surface (HL) and sub-surface (LL) portions of the water column (West and Scanlan, 1999; West et al., 2001), and representative strains from each clade

grew faster at either high (HL) or low (LL) irradiances (Moore and Chisholm, 1999). Comparative genomic analysis of strains of HL and LL *Prochlorococcus* revealed several genetic differences underlying the light adaptations, including chlorophyll-binding proteins unique to the LL strains, as well as high-light-inducible proteins specific to HL strains (Rocap et al., 2003; Kettler et al., 2007). Differences were also discovered among strains of the HL and LL clades. The HL clade in particular is broken into two subclades, HLI and HLII: though the members of the subclades are 99.2% identical at the 16S rRNA locus, 10-15% of the genome of each strain consists of subclade- or strain-specific genomic islands (Coleman et al., 2006), which may contribute to different temperature adaptations or reflect the different distributions of the subclades on a global scale (Johnson et al., 2006). The combination of *in situ* distribution analysis, comparative genomic analysis, and strain isolation, cultivation, and characterization has contributed to a deep understanding of the global *Prochlorococcus* population.

Another comprehensive investigation of closely related strains of bacteria is the analysis of strains of soil *Bacillus*. Koeppel et al. (2008) employed ES in a multi-locus sequence analysis of strains of *Bacillus simplex* isolated from Evolution Canyons I and II and strains of the *Bacillus subtilis-Bacillus licheniformis* clade isolated from Evolution Canyon III, in Israel. There are three different environments within each canyon: the less solar-exposed north-facing slope, the more solar-exposed south-facing slope, and a typically dry streambed at the base. ES predicted multiple ecotypes within the traditionally named species, many of which had distinct environmental distributions in the canyons from which they were isolated. Similarly, strains of *Bacillus* were isolated by

Connor et al. (2010) from a canyon in Death Valley, CA, which hosts the three different environments found in the Evolution Canyons. PE demarcations correlated with difference in solar exposure and difference in soil texture from which the strain was isolated, as well as the heat adaptation index, which is calculated by measuring fatty acid content of the strain. A comparative genomic analysis of four *Bacillus subtilis* strain genomes within one PE (PE 15) and the genome of one strain from a different PE (PE 10) was performed to investigate the ecological heterogeneity among strains within a PE (Kopac et al., 2014). The authors claim that there were no significant gene content differences among isolates of the same PE (no annotated genes that would impart a unique function to the strain coding it). However, each strain was found to have a different set of genes under positive selection, suggesting that there are different selective pressures acting on each strain. Genes that demonstrate positive selection have more non-synonymous nucleotide substitutions than synonymous nucleotide substitutions, which affects the amino acid sequence and possibly the functionality of the protein (Hughes and Nei, 1988). Since any non-synonymous substitutions that decrease the fitness of an organism should be selected against, any genes that show positive selection have most likely increased the fitness of an organism (i.e. Chen et al., 2006). Additionally, one strain that had extra genes for maltose/maltodextrin use displayed significant growth advantages over the other strains when grown on maltose or maltodextrin as the sole carbon source. The strains of *B. subtilis* in PE 15 appeared to be ecologically heterogeneous, and the authors concluded that another model of microbial speciation, the Nano-Niche model (Figure 1.10; Cohan and Perry, 2007; Wiedenbeck and Cohan, 2011),

may better fit their data. This model proposes that subpopulations of an ecotype use the same set of resources, but in different proportions, and therefore may have some ecological heterogeneity.



Figure 1.10. The Nano-Niche Model. An ecotype (the blue line) has ecologically heterogeneous members that may use the same set of resources, but in different proportions. These subpopulations (different colored lines; Abc, aBc, and abC) experience unique periodic selection events (small asterisks), but eventually a mutation will arise that will quash the diversity in all but one subpopulation (large asterisks) since they still use the same set of resources (Wiedenbeck and Cohan, 2011)

The genomes of two *Synechococcus* isolates from Octopus Spring, representative of 16S rRNA-defined genotypes A and B', have been previously sequenced and compared (Bhaya et al., 2007). In addition to different phenotypes exhibited by the A and B' isolates (temperature adaptations corresponding to their location along the effluent channel; Allewalt et al., 2006), their genomes showed gene content differences that may reflect the different ecological niches in which they reside. The B' isolate genome contains genes for phosphonate transport and degradation, and both cyanophycin synthesis and degradation, while the A isolate genome has a urease (urea catabolism)

gene cassette not present in the B' isolate genome (Bhaya et al., 2007). These isolate genomes were used to discover even more differences in gene content among *Synechococcus* spp. metagenomic sequences from the Octopus Spring and Mushroom Spring mats, such as genes for ferrous iron transport (Bhaya et al., 2007) and urea carboxylase, another enzyme for urea catabolism (Klatt et al., 2011). The A and B′ isolates represent different species as demarcated by the 16S rRNA "molecular cutoff" (<97% sequence homology between A and B′), but, as shown by Becraft et al. (2011, submitted), are not as closely-related as the predicted ecological species within the A-lineage and B′-lineage. As explained in further detail below, these genomes alone cannot be used to test hypotheses on ecological species and speciation.

Main Aims of this Thesis

To infer what the specific ecological adaptations of an ecotype might be, as well as to test whether isolates of *Synechococcus* spp. from the Mushroom and Octopus Spring microbial mats fit the Stable Ecotype Model of speciation, in which members of an ecotype are both ecologically distinct from members of another ecotype and ecologically interchangeable with other members of that ecotype, I cultivated *Synechococcus* isolates representative of different PEs, as well as multiple isolates of several PEs, to obtain their genomes for comparative analyses. Chapter 2 describes the cultivation efforts made by myself and other members of the Ward lab to obtain *Synechococcus* isolates for genome sequencing and analysis. In Chapter 3, I compared an orthologous-gene genome phylogeny to single-locus and multi-locus phylogenies of *Synechococcus* spp., and

searched for genes for multi-locus phylogenetic analysis that might offer greater molecular resolution and buffering against the effects of recombination than single-locus phylogenies. In Chapter 4, I compared gene content among isolates representing PEs predominant at 60°C in Mushroom Spring. As in the case with the *Prochlorococcus* HL and LL clades, I expected to find differences in gene content that could explain the different distributions of the PEs and reveal additional ecological differences between the PEs. In Chapter 5, I compared gene content among isolates representing the same PE, as well as isolates with different temperature adaptations. I also compared gene sequences of orthologous genes to find genes and sites in genes that have been under positive selection. Comparisons of multiple strains within a PE (Chapter 5) should reveal evidence of ecological interchangeability within an ecotype. If such evidence was observed (i.e. that the hypothesized ecological species of *Synechoccocus* predicted by ES are both ecologically interchangeable with other individuals of the species, and are ecologically distinct from other species), the Stable Ecotype Model would be supported.

CHAPTER TWO

CULTIVATION OF *SYNECHOCOCCUS* SPP. FOR COMPARATIVE GENOMIC AND PHENOTYPIC ANALYSIS

Introduction

Thermophilic cyanobacteria from the genus *Synechococcus* have been cultivated and studied for over 50 years (Peary and Castenholz, 1964). They have been found in hot springs around the world, including various locations in North America, New Zealand, and Japan (Ward et al., 2012), but perhaps the most extensively studied populations inhabit the microbial mats in hot springs in Yellowstone National Park, WY (Brock, 1978; Ward et al., 1998; Ward et al., 2006). Cultivation of *Synechococcus* from various Yellowstone springs originally yielded one species, *Synechococcus lividus* (Brock, 1978; Ward et al., 1987), but 16S rRNA sequence analysis revealed that all of the strains cultivated from the Yellowstone Springs were identical at the 16S rRNA locus (Ferris et al., 1996), and that cultivated strains and predominant *in situ* populations were phylogenetically unrelated (Figure 1.4; Ward et al., 1990, 1994; Ferris et al., 1996).

Ward et al. (1997), and later Allewalt et al. (2006), succeeded in cultivating isolates representative of the predominant populations of the 16S rRNA-defined A/B-lineage. Allewalt et al. (2006) used floating filter cultivation (which isolates single cells from one another on a filter that is placed atop a pre-filter saturated with culture medium) for growth, preventing competitive exclusion (de Bruyn et al., 1990). However, colonies were not observed on filters inoculated with mat samples diluted more than 10^5-fold.

Because undiluted mat samples had ~10^8-10^9 cells/mL, the low-dilution filters on which colonies were observed should have contained at least 10^3-10^4 cells per filter, which could include representatives of non-predominant populations in the mat, and colonies that developed from multiple cells, potentially from different ecotypes. Becraft et al. (2011; submitted) used a protein-encoding locus, *psaA*, to demarcate species within the A/B-lineage sequence types, revealing more than 40 distinct putative ecotypes (PEs) (Figure 2.1). In order to cultivate isolates representative of these ecological species for genomic and phenotypic characterization more readily, Shane Nowack, myself, and other members of the Ward lab developed an improved culture medium, which increased the

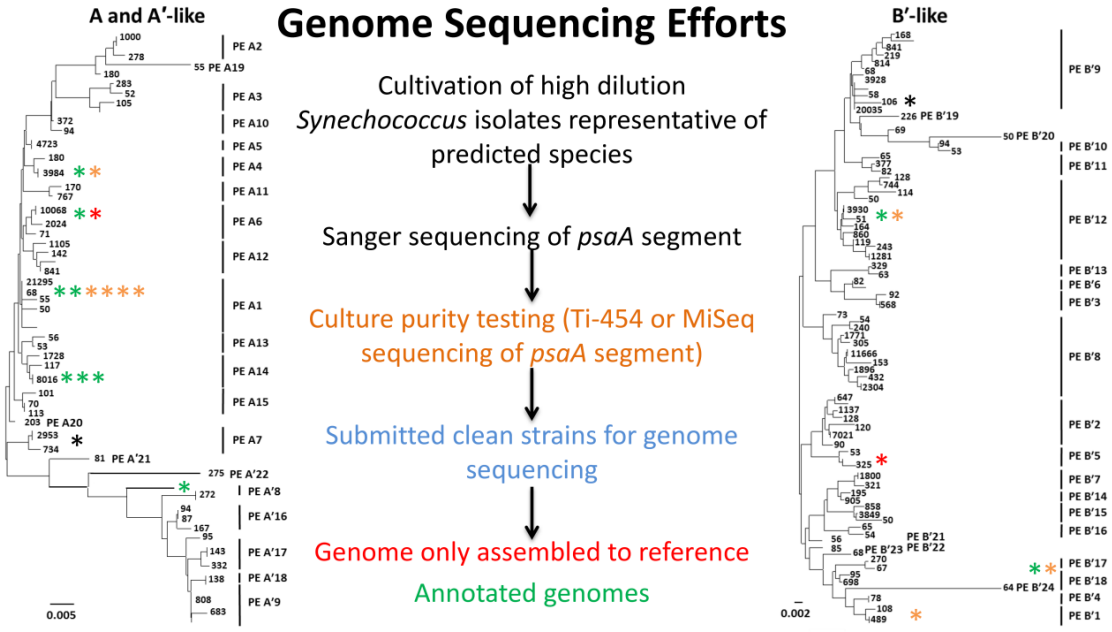


Figure 2.1. A'/A-lineage and B'-lineage *Synechococcus* phylogenies based on environmental *psaA* barcode sequences (Becraft et al., submitted). Predicted putative ecotypes PEs are indicated by bars to the right of each tree. Bar length below the tree represents the number of point mutations per site. Text in the middle of the figure describes different stages in the genome sequencing pipeline. Stars on the phylogenetic trees represent cultures in the various stages of the genome sequencing efforts, with colors representing the progress of the individual culture through the pipeline.

rate of culture growth and yielded growth from more highly-diluted (10^{7}-10^{8}) samples (Nowack, 2014; Nowack et al., submitted). The results presented in Nowack et al. (submitted), as well as the results of the continuation of these cultivation efforts, are reported below. The ultimate goal of these efforts is cultivation of multiple isolates representative of multiple PEs (Figure 2.1) in order to test the expected properties of members of an ecotype, which are hypothesized to be both ecologically interchangeable and ecologically distinct from members of other PEs.

Methods

Cultivation of *Synechococcus* Isolates

Cylindrical core samples 7 mm in diameter were removed from various temperature-defined sites in the Mushroom Spring microbial mat using a #4 cork borer, placed in 15 mL Falcon tubes filled with spring water, then put into a Whirl-Pak® bag. A thermos containing 55°C water from the spring was used to incubate the Whirl-Pak® bags while in transit to the lab (approximately four hours). The top-green layer (1-2 mm thick) was removed from a core sample, placed in a Dounce tissue grinder along with autoclaved Mushroom Spring water, and homogenized. The homogenate was diluted ten-fold from 10^{-1} to 10^{-8} with autoclaved Mushroom Spring water; the diluted cell suspensions were then filtered through sterile polycarbonate filters (47 mm Nucleopore) using a vacuum filtration method. The filtration was performed from highest dilution to lowest dilution (10^{-8} to 10^{-1}) to minimize cross-contamination of the samples. The filters were transferred to 47 mm Millipore Petri dishes prepared with Whatman glass fiber

filters and 1.5 mL Castenholz medium D (Castenholz, 1969) plus HEPES buffer, 0.01% w/v sodium acetate, and 0.01% w/v yeast extract (hereafter referred to as medium DHAY). To prevent excess drying of the dishes, they were placed in plastic bags with wetted paper towels and then incubated at 52°C with 47 µmol photons m^{-2} sec^{-1} of fluorescent light. Medium DHAY was added every 2-3 days, as needed, due to evaporation. Serial dilutions were also performed by preparing Gelrite plates inoculated with the diluted samples. Gelrite was added to medium DHAY at 1.5% w/v, and 20 ml aliquots were transferred into glass, and autoclaved. The bottles were then placed in a 45°C water bath and allowed to cool to just above 45°C. Once the Gelrite was cooled, it was inoculated with 5 mL of 10^{-1} to 10^{-8}-fold diluted mat samples, mixed gently and poured into a 47 mm diameter Petri dish. The Gelrite was allowed to solidify, and the plates were incubated as described above. Discrete colonies were picked from the plates (Figure 2.2) using sterilized toothpicks to inoculate liquid cultures.

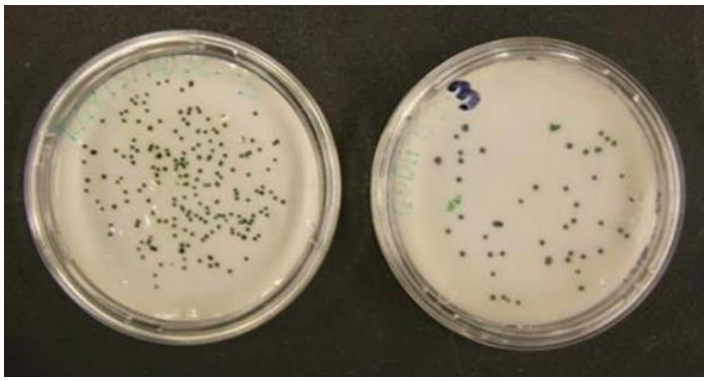


Figure 2.2. Floating filter plates showing discrete colony growth.

Culture Genotyping

After 200-mL liquid cultures were grown to late exponential phase, a 1.5-mL aliquot of cells was pelleted by centrifugation and lysed using the Fastprep Cell Disrupter

(Bio101 Savant Instruments, New York). DNA for Sanger sequencing was extracted using the FastDNA Spin kit (Molecular Biosciences, Boulder, CO), per the instructions of the manufacturer and amplified using PCR primers for the *psaA* locus as described in Becraft et al. (2011). Sequencing of *psaA* amplicons was performed at either the Idaho State University Molecular Biosciences Core Facility (Pocatello, ID) or at Genewiz, Inc. (South Plainfield, NJ). Cultures with *psaA* sequences representing PEs present in environmental samples (Becraft et al. 2011) were purified by serial dilution in Gelrite pour plates, and discrete colonies were picked from high-dilution filters and used to inoculate liquid cultures.

DNA Extraction for Culture Purity Testing

DNA was extracted using an enzymatic extraction procedure. A 60-mL aliquot of culture was pelleted by centrifugation at 1000 RCF for 30 minutes. The supernatant was removed and the pellet was flash-frozen in liquid nitrogen, and stored at -20°C until extraction. Each frozen pellet was thawed and resuspended to 1 mL total volume in medium DHAY. Lysozyme was added to 200 µg/mL, and then the sample was incubated for 45 minutes at 37°C, homogenizing the sample by vortex mixing every 15 minutes. Next, 110 µL of a 20% solution of sodium dodecyl sulfate and 15 µL of 20 mg/mL proteinase K (activity of 30mAU/mg; one milliAnson Unit (mAU) is the activity that releases folin-positive amino acids and peptides corresponding to 1 µmol tyrosine per minute) were added to the sample and the sample was incubated for 50 minutes at 50°C, homogenizing by vortex mixing every 10 minutes. DNA was extracted by adding 950 µL of Tris-EDTA-buffered phenol, gently mixing for 3 minutes, then pelleted by

centrifugation for 5 minutes at 6600 relative centrifugal force (RCF). The aqueous layer was removed, then 450 µL phenol and 450 µL chloroform/isoamyl alcohol (24:1) were added to the sample, which was mixed gently for 3 minutes, then pelleted by centrifugation for 5 minutes at 6600 RCF. The aqueous layer was removed, then 900 µl chloroform/isoamyl alcohol (24:1) added to the sample, which was mixed gently for 3 minutes, then centrifuged for 5 minutes at 6600 RCF. The aqueous layer was removed, 40 µL 3M sodium acetate, pH 5.2, and 1 mL ethanol were added, and the DNA was allowed to precipitate overnight at -20°C. The sample was centrifuged at 4°C (6600 RCF) for 30 minutes, washed in 0.9 ml 70% ethanol twice, and dried using a Savant DNA Speed Vac for 10 minutes at ambient temperature. The sample was resuspended with 200 µL Tris-EDTA buffer, pH 7.5, and the DNA was allowed to hydrate with no mixing for two hours. RNA and other impurities were removed using the RNase I treatment protocol required by the DOE Joint Genome Institute (JGI) and described at http://my.jgi.doe.gov/general/protocols.html.

Culture Purity Testing

The diversity of *Synechococcus* variants in the culture was tested using Ti-454 barcode (Becraft et al., submitted) or Illumina MiSeq sequencing of the *psaA* locus (for details see www.researchandtesting.com/next-generation-sequencing-service.html). Cultures were considered candidates for genome sequencing when they were found to contain a single, predominant cyanobacterial ecotype: when the vast majority of the *psaA* sequences from the culture were identical or exhibited 1-2 single-nucleotide polymorphisms (SNPs) from the dominant allele (allowing for sequencing error and

mutation within the culture). If *psaA* sequences representative of one PE were detected, the culture was again diluted to extinction and DNA from a high-dilution isolate was tested by another round of Ti-454 barcode sequencing of the *psaA* locus. 16S rDNA surveys of the cultures were also carried out using the Ti-454 barcode sequencing of the 16S rRNA gene using the 28F and 519R primers according to the methodology posted on the Research and Testing Laboratory website at www.researchandtesting.com.

Culture Preservation

Cultures were grown into late exponential phase and then two 20-30 mL aliquots were pelleted by centrifugation for 30 minutes at 1000 RCF. The supernatant was discarded and biomass was flash-frozen with liquid nitrogen and stored at -20°C. Three 1.5 mL aliquots were cryopreserved for future cultivation using the protocol for cyanobacterial cryopreservation found at http://www-cyanosite.bio.purdue.edu/protocols/cryo.html. One 1.5-mL aliquot was preserved in 5% gluteraldehyde and frozen for cell counting.

Results

Cultivation of *Synechococcus* Isolates

Several isolates from both the A and B' *Synechococcus* lineages, as well as one isolate from the A'-lineage, have been cultivated from Mushroom Spring samples (Figure 2.2). Currently, we are not able to cultivate axenic strains of *Synechococcus*; all of the cultures contain heterotrophic contaminants (usually *Meiothermus* and *Anoxybacillus* spp. (see Nowack et al., submitted), along with other major taxa found in the Mushroom

Spring mat community (Klatt et al., 2011)). Cultures were screened for *psaA* sequences types representative of dominant sequences found *in situ*; those with dominant alleles representing different PEs were placed in the genome sequencing pipeline (Figure 2.1).

Culture Purity Testing

Results for purity testing of isolates whose genomes have been sequenced are reported in Table 2.1. All isolates sent for genome sequencing (names in black) consisted of a single, predominant cyanobacterial ecotype.

We also tested the purity of previously cultivated strains to evaluate whether they contained a single, predominant cyanobacterial *psaA*-demarcated ecotype. The A-lineage isolate (JA-3-3Ab), previously cultivated by Allewalt et al. (2006), is an example of an isolate we consider to have a single, predominant cyanobacterial ecotype, since 83.4% of the sequences match the dominant variant type for PE A1, and only one sequence matches another dominant variant (the PE A14 sequence type, which only differs by one single nucleotide polymorphism (SNP) different from the PE A1 sequence type in the locus sequenced). We do not, however, consider the B'-lineage isolate (JA-2-3B′a(2-13)), also cultivated by Allewalt et al. (2006), to have a single, predominant cyanobacterial ecotype, because we detected sequence types of three PEs in the culture: PE B′24 (14.6%), which matches the published B′ isolate genome *psaA* sequence type, PE B′19 (67.4%), which differs by 5 SNPs from the PE B′24 sequence type, and PE B′11 (0.2%). Another culture, CIW-10, which was isolated from the B'-lineage culture and described as an axenic strain (Kilian et al., 2007), was also found to contain multiple predominant

Table 2.1. Summary of *psaA* Ti-454 or MiSeq analysis of *Synechococcus* isolate purity. Isolates in red were cultivated using old methods (Allewalt et al., 2006; Kilian et al., 2007) [Expanded from Table 1 in Nowack et al. (submitted)]

				Dominant Variant Sequences[1]			Other Sequences[2]	
Isolate Name	**Temperature[3]**	**Depth[4]**	**Total Sequences**	**Putative Ecotype[5]**	**DVs**	**%**	**Total**	**%**
JA-3-3Ab	58-65°C	400-600	1116	all	931	83.4%	185	16.6%
				A1	930	83.3%		
				A14	1	0.1%		
JA-2-3B'a (2-13)	51-61°C	N/A	2054	all	1690	82.3%	364	17.8%
				B'19	1385	67.4%		
				B'24	300	14.6%		
				B'11	5	0.2%		
CIW-10[6]	51-61°C	N/A	1528	all	1077	70.5%	451	29.5%
				B'24	438	28.7%		
				DV74[7]	639	41.8%		
65AY6Li	65°C	400-600	980	**A1**	780	79.6%	200	20.4%
60AY4M2	60°C	800	443	**A14**	374	84.4%	69	15.6%
63AY4M1	63°C	800	1007	**A14**	816	81.0%	191	19.0%
65AY640	65°C	800	920	**A14**	756	82.2%	164	17.8%
65AY6A5	65°C	640-720	1325	all	1106	83.5%	219	16.5%
				A4	1105	83.4%		
				A1	1	0.1%		
63AY4M2	63°C	400-960	2027	**A6**	1756	86.6%	271	13.4%
M65AY6	65°C	400-960	6730	**A6**	5575	82.8	1155	17.2%
				B-types	6	0.01%		
65AY6A'	65°C	N/A	2041	**A'-like**	920[8]	45.1%	1121	54.9%
R68DH1S1-R03C	68°C	N/A	4992	**B'12**	3825	76.6%	1167	23.4%
55AY5B5	55°C	N/A	1036	**B'5**	746	72.0%	290	28.9%
				A-type	1	0.01%		
A'-3-AG-7	65°C	400-600	26719[9]	**A1**	15025	56.2%	11694	43.8%
A'-4-J-7	65°C	400-600	17665[9]	**A1**	9780	55.4%	7885	44.6%

[1] Number of sequences that are identical to the dominant variant sequence of the putative ecotype (PE) listed in the same row.
[2] Other sequences from the culture that represent sequence variants 1-2 SNPs different from dominant variant sequence listed.
[3] Temperature of the sample from which the isolate was obtained.
[4] Depth (µm) from surface where PE is most abundant at 60°C in Mushroom Spring Microbial mat (see Figure 1.7). N/A means PE was not a dominant population in the 60°C mat.
[5] Putative ecotypes (PEs): in bold are the dominant PEs of the culture.
[6] CIW-10 is reportedly an 'axenic' culture derived from JA-2-3B′a(2-13) (Kilian et al., 2007).
[7] DV74 is a dominant variant sequence type not demarcated as a PE found in Mushroom Spring.
[8] This sample had poor sequence quality, so the DV sequence includes the DV (474 sequences total) and a second sequence variant (446 sequences total), identical except for an ambiguous base call at the same base.
[9] Samples that were sequenced using Illumina MiSeq instead of Ti-454 barcoding. Systematic errors were not removed.

cyanobacterial ecotypes: PE B′24 (28.7%) and DV 74 (41.8%), which differs by 17 SNPs from the PE B′24 sequence type.

Discussion

Synechococcus isolates, representative of multiple PEs demarcated based on *psaA* sequence variation using a theory-based evolutionary simulation analyses, have been cultivated from Mushroom Spring samples using cultivation techniques developed by Shane Nowack, myself, and other members of the Ward lab (Nowack et al., submitted). These isolates are in various stages of a pipeline leading to genome sequencing (e.g. Olsen et al., submitted) and phenotypic characterization (Nowack et al., submitted; Schaible et al., unpublished) of isolates representative of different PEs. We have not been able to cultivate *Synechococcus* cultures without heterotrophic contaminants, but have developed a stringent test for determining whether the cultures have a single, predominant cyanobacterial variant. Deep sequencing of the culture at the *psaA* locus should reveal any contaminating variants prevalent in the culture: Table 2.1 shows cultures with a single, predominant cyanobacterial strain (JA-3-3Ab and all cultures labeled in black), as well as cultures with multiple cyanobacterial strains (JA-2-3B′a(2-13) and CIW-10). Even with next-generation sequencing technologies, sequences cannot represent every cell present in a sample (~10^3 sequences/sample with Ti-454 barcoding, ~10^4 sequences/sample with Illumina MiSeq), or every cell of a 250-mL culture with ~5 x 10^7 cells/mL (~10^8 cells total; Nowack et al., submitted). Care must be taken to ensure a culture maintains the same predominant PE (Shrestha et al., 2013). Although we found multiple strains of *Synechococcus* in our culture of JA-2-3B'a(2-13), the PE B′19 sequence type may or may not have been prevalent in the culture when it was sequenced

(Bhaya et al., 2007). Any analyses using the JA-2-3B′a(2-13) genome should be done with the understanding that the genome sequence may be chimeric. Isolates with the PE B'24 *psaA* sequence type are in the second round of purity testing: obtaining a genome sequence of a PE B'24 isolate should help to resolve questions about ecotypic origin(s) of the published JA-2-3B'a(2-13) genome.

The discrepancy between the number of readily cultivable organisms and the organisms in nature has been known since 1932, when Razumov noted that direct counts of bacterial cells from the environment were orders of magnitude higher than colony counts from plating the same samples. Staley and Konopka (1985) dubbed this phenomenon "The Great Plate Count Anomaly". Molecular sequencing technologies emphasized this "uncultured majority", revealing members of numerous phyla in environmental samples without a single cultured representative, as well as the dearth of cultures that reflect the diversity within each phylum (Rappé and Giovannoni, 2003). Metagenomics, and now single-cell genomics, facilitate investigation of microbial populations without isolation and cultivation (i.e. Simmons et al., 2008; Klatt et al., 2011; Rinke et al., 2013), but phenotypic and genomic analyses of isolated microorganisms offer answers to questions deeper than simply "what is an organism's genomic content and inferred functional potential?". By altering old and developing new cultivation methods, we and others have been able to cultivate and characterize some of the previously "uncultured majority" (Joseph et al., 2003; Tank and Bryant, 2015; Nowack et al., submitted). So long as these isolates are truly representative of predominant PEs *in*

situ (see Chapter 3), we can use them to (i) infer distinct ecological adaptations and their underlying mechanisms from comparative genomic analysis of isolates of different PEs (Chapters 4 and 5) and phenotypic analysis (Nowack et al., submitted) and (ii) test hypotheses about the ecological interchangeability of individuals within the same PE, both genomic (Chapter 5 and ongoing research) and phenotypic (Schaible et al., unpublished).

CHAPTER THREE

THE EFFECTS OF RECOMBINATION ON PHYLOGENETIC INFERENCE BASED ON SINGLE-LOCUS, MULTI-LOCUS, AND GENOMIC SEQUENCE ANALYSES

Introduction

Since the emergence of gene sequencing and sequence-based phylogenetics, investigators have argued over which loci should be used to create the phylogenies (Gee, 2003). Many single-locus phylogenies are incongruent, or have a different branching pattern, from other single-gene phylogenies of the same set of organisms. Although 16S rRNA phylogenies are widely accepted, 16S rRNA is too conserved to distinguish between closely related organisms (as discussed in Chapter 1) and has also proven to be susceptible to recombination (Yap et al., 1999), which has the potential to distort the gene phylogeny away from the true evolutionary history of the organism. To circumvent the problems of single-locus phylogenies, many have turned to multi-locus sequence analysis, or MLSA, which uses several conserved gene sequences to associate related variants (e.g. Urwin and Maiden, 2003; Baldo et al., 2006; Cadillo-Quiroz et al., 2010). MLSA has led to phylogenies that are more highly resolved than 16S rRNA phylogenies (see Chapter 1) but, contrary to previous hypotheses, may also be vulnerable to the effects of recombination with DNA from foreign sources (see below). Phylogenomics, or phylogenies based on whole genome comparison methods, are now considered to show the most accurate relationships among organisms (Delsuc et al., 2005). In particular,

phylogenies created using large numbers of orthologous genes or genomic regions of closely related organisms have correlated with phenotypic characteristics of the organisms (Zhang and Lin, 2012), are highly supported statistically (Dutilh et al., 2007), and have an improved "signal-noise ratio" (referring to vertically inherited sequence as the signal and horizontally acquired, or recombined, sequence as the noise) (Klenk and Göker, 2010).

In this study, I compared the orthologous-gene genome phylogeny of 8 closely-related *Synechococcus* isolates to both single-locus and multi-locus phylogenies, observing how each phylogeny describes the genetic relationships among *Synechococcus* isolates that are identical, or nearly so, at the 16S rRNA locus, but have been shown to have distinct environmental distributions (Becraft et al., submitted), phenotypes (Nowack et al., submitted), and genome content and gene selection patterns (Olsen et al., submitted; Chapters 4 and 5). Based on their *psaA* sequences, these isolates are representative of different PEs that are predominant in the Mushroom Spring microbial mat community at 60°C (Becraft et al., submitted). Three putative ecotype (PE) sequence types are represented by multiple isolates: two isolates representative of PE A1 (*Synechococcus* sp. JA-3-3Ab and 65AY6Li, designated PE A1-OS and PE A1-MS, respectively, in this thesis); two isolates representative of PE A6 (*Synechococcus* sp. 63AY4M2 and M65AY6, designated PE A6-63 and PE A6-65, respectively, in this thesis); three isolates representative of PE A14 (*Synechococcus* sp. 60AY4M2, 63AY4M1, and 65AY640, designated PE A14-60, PE A14-63, and PE A14-65,

respectively, in this thesis); and the single isolate representative of PE A4 (*Synechococcus* sp. 65AY6A5, designated PE A4 in this thesis). I also attempted to find genes suitable for MLSA of the *Synechococcus* spp. studied here, seeking genes that (i) are present in all genomes, (ii) are housekeeping genes (Urwin and Maiden, 2003), (iii) do not show evidence of recombination or selection, and (iv) are divergent enough to differentiate closely related isolates (Urwin and Maiden, 2003). These analyses, along with data from environmental single-locus (Becraft et al., submitted) and multi-locus (Melendrez et al., in prep) sequence phylogenies, test the accuracy of single- and multi-locus sequence analysis and demonstrate the effects that recombination has on the phylogenetic signal of stable ecological lineages. The genomes and gene content of these isolates will be further discussed in Chapters 4 and 5.

Methods

Orthologous-Gene Alignment and Genome Phylogeny Construction

Genomes were obtained for *Synechococcus* isolates as described in Nowack et al. (submitted) and in Chapter 2. Isolate genomes were sequenced and assembled as described in Chapter 4 (Olsen et al., submitted). The *Synechococcus* sp. PE A14-63 genome was designated as heterozygous upon assembly, but assembled to a nearly complete genome (see Chapter 5 for further discussion). *Synechococcus* sp. PE A6-65 was further assembled to reference sequences PE A6-63 and PE A1-OS using the map to reference program in Geneious version 7.1.7 (http://www.geneious.com, Kearse et al.,

2012). Genomes of the isolates mentioned above were aligned using Mauve progressive aligner (Darling et al., 2004). Alignments of the orthologous genes shared among all the genomes were concatenated and a maximum-likelihood tree was constructed using FastTree, with local support values calculated using the Shimodaira-Hasegawa test (Price et al., 2009). 1136 orthologous genes were concatenated for a total sequence length of 1,058,325 bp. The same genes from the genome of a *Synechococcus* isolate representative of PE A' (*Synechococcus* sp. 65AY6A′), which is closely related to the A-lineage *Synechococcus* isolates, but has a different 16S rRNA sequence type, was used as an outgroup.

Single-Locus Sequence Analysis of Genomic DNA Sequences

The *psaA* gene sequence (2286 nt) from the genome of each isolate, as well as the 324-bp PCR-amplified *psaA* region used by Becraft et al. (submitted), were aligned using ClustalW and maximum-likelihood phylogenetic trees were computed using MEGA6 (Tamura et al., 2013), with bootstrapping values calculated with 500 replications. Recombination events were detected with RDP4 (Martin et al., 2010).

MLSA of Genomic DNA Sequences

Sequences of multiple loci (*rbsK, hisF, lepB, aroA, dnaG,* a protein kinase [CYA_2262], and a conserved hypothetical protein [locus tag CYA_2291]) from the *Synechococcus* isolates were concatenated and aligned with existing MLSA concatenations from environmental BAC clones (Melendrez et al., in prep). A maximum-

likelihood tree was created using the seven-gene concatenation, as well as a six-gene concatenation that excluded the *rbsK* locus (Becraft, 2014), with bootstrapping values calculated with 500 replications.

A separate set of seven loci was chosen using a genome-informed approach for an alternative MLSA. Six genomes from the A-lineage (those of the PE A1-OS, PE A1-MS, PE A4, PE A6-63, PE A14-60, PE A14-63, and PE A14-65 isolates) were analyzed using the ODoSE program (Vos et al., 2013), available on the Galaxy platform at https://usegalaxy.org/. All hypothetical genes were discarded. Remaining alignments of orthologous genes with >70% nucleotide homology were scanned for both positive selection using ODoSE (Vos et al., 2013)) and recombination using RDP4 (Martin et al., 2010). Orthologs that did not exhibit either of these traits were scanned for segments of nucleotide divergence small enough to be amplified using high throughput sequencing analyses, and sufficiently divergent to differentiate most, if not all the isolates. Using these criteria, loci from seven genes were chosen for concatenation, alignment, and maximum-likelihood phylogenetic tree construction using MEGA6 (Tamura et al., 2013), with bootstrapping values calculated with 500 replications. Recombination events were detected with RDP4 (Martin et al., 2010).

Results

Orthologous-Gene Alignment and Genome Phylogeny Construction

Figure 3.1 shows the genome phylogeny constructed from an orthologous-gene concatenation. With the exception of the two PE A6 isolates, members of the same PE (i.e. both members of PE A1 and all three members of PE A14) formed clades together, and the high-light-adapted PE A1 isolates formed a different clade (light green) from the low-light-adapted PE A4 and A14 isolates (dark green) (Nowack et al., submitted).

Investigating Incongruence between Single-Gene and Genomic Phylogenies

The phylogeny resulting from the full *psaA* gene sequence (Figure 3.2A) differed from both the phylogeny made using the amplified segment alignment (Figure 3.2B) and the orthologous-gene genome phylogeny (Figure 3.1). RDP4 detected a possible



Figure 3.1. Maximum-likelihood tree of the A-lineage isolates based on 1136 aligned and concatenated orthologous genes. The scale bar describes the branch lengths, and is equal to the number of substitutions per site in the sequence. Tree is rooted by the corresponding concatenation of orthologous *Synechococcus* sp. 65AY6A′ genes (not shown).

recombination event in the PE A1-OS and PE A6-65 isolates in the *psaA* full gene alignment, although the program could not confidently assign a parent sequence. While isolates belonging to the same PE are identical at the amplified *psaA* region used for environmental sequence analysis and culture sequence typing (see Becraft et al., submitted and Nowack et al., submitted), the whole-gene sequences of the two PE A6

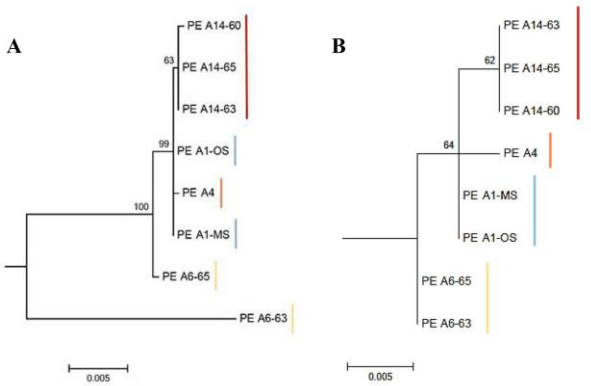


Figure 3.2. Maximum-likelihood trees of A-lineage genome isolate full length (A) and amplified region (B) *psaA* sequences. The scale bars describe the branch lengths, and are equal to the number of substitutions per site in the sequence. Trees are rooted by *Synechococcus* sp. JA-2-3B'a(2-13) (not shown).

isolates are very different, as the isolate PE A6-63 *psaA* sequence has numerous SNPs outside of the amplified region that cause it to differ from the sequence of the other PE A6 isolate and the rest of the A-lineage isolates (Figure 3.3). Most of the SNPs are

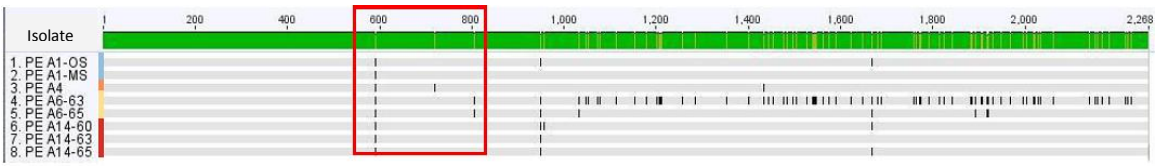


Figure 3.3. Alignment of isolate *psaA* sequences. Base numbers are above the alignment. Black lines indicate mutation with respect to the consensus sequence, while grey indicates that the sequence is identical to the consensus. The PCR-amplified region is between bases 532 and 837 (outlined in red).

synonymous (only 6 amino acid differences in the PE A6-63 isolate sequence compared to the PE A1-OS isolate sequence), so the functionality is expected to be similar.

Impacts of Recombining Loci on MLSA Phylogenies

Figure 3.4 shows the phylogenies constructed using the seven loci used by Melendrez et al. (in prep) for sequence typing environmental BAC clones. Melendrez et al. (in prep) found that the amplified segment in the *rbsK* gene was recombinant, so the phylogeny is displayed both with (Figure 3.4A) and without (Figure 3.4B) the *rbsK* sequence in the concatenated sequence alignment. The isolates representative of PE A1 (highlighted blue) did not form a clade together in the seven-locus phylogeny (Figure 3.4A), but did cluster together in the phylogeny without *rbsK* (Figure 3.4B). The isolates representative of PE A14 (highlighted red) formed a clade together with *rbsK* included in the MLSA (Figure 3.4A), but one isolate (PE A14-63), did not cluster with the other isolates when *rbsK* was omitted. The two isolates representative of PE A6 (highlighted yellow) do not fall together in either phylogeny (Figure 3.4), as observed in the orthologous-gene genome phylogeny (Figure 3.1).

Table 3.1 describes the genes chosen in the genome-informed MLSA. Figure 3.5 shows the genome-informed MLSA maximum-likelihood phylogeny of the A-lineage isolates. Due to poor genome coverage of the PE A6-65 isolate, only 5 of the 7 chosen loci were used for the phylogeny (*hisB*, *folB*, *dapL*, *speA*, and *valS*). Recombination was not detected in any of the genes in the six-genome alignments, but with the addition of the seventh genome (PE A6-65), that was not included in the genome-informed MLSA

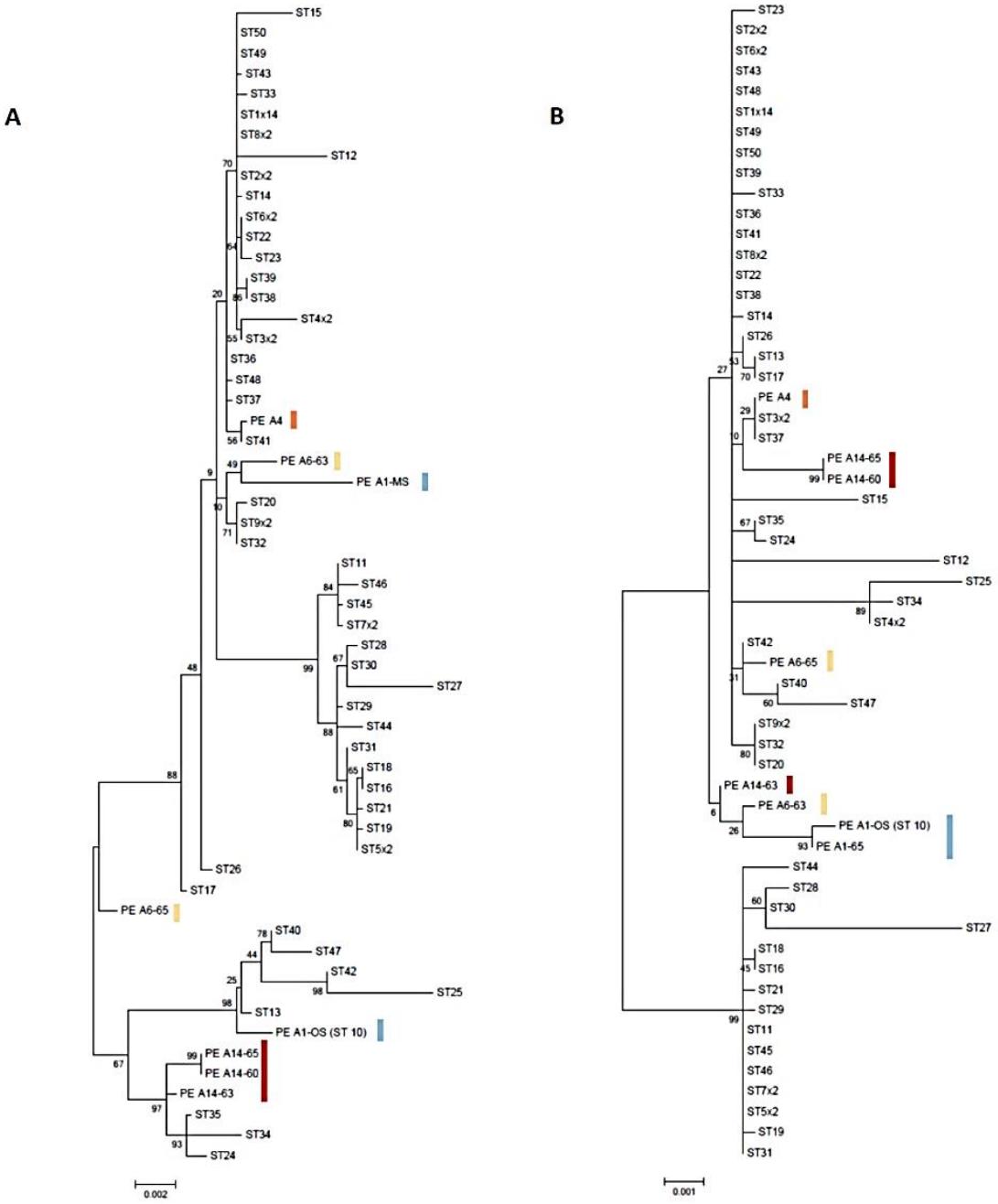


Figure 3.4. Maximum-likelihood trees of seven concatenated MLSA loci from environmental bacterial artificial chromosome (BAC) clones (sequence types designated ST, followed by a sequence type number) (Melendrez et al., in prep) and A-lineage isolates with (A) and without (B) the *rbsK* locus. The scale bars describe the branch lengths, and are equal to the number of substitutions per site in the sequence.

Table 3.1. Genes chosen from genome-informed analysis for MLSA.

Locus tag	Gene description	Protein mnemonic	% Nucleotide (nt) identity	# of nt differences /total gene length	Variable region nt length	MLSA region[1]
CYA_2626	Imidazoleglycerol -phosphate dehydratase (EC 4.2.1.19)	HisB	93.8	72/618	600-618	whole gene
CYA_0021	Dihydroneopterin aldolase (EC 4.1.2.25)	FolB	96.9	24/369	1-396	whole gene
CYA_2460	L,L-diaminopimelate aminotransferase (EC 2.6.1.83)	DapL	98.1	44/1233	490-890	490-1233
CYA_0128	Biosynthetic arginine decarboxylase (EC 4.1.1.19)	SpeA	98.7	58/1923	565-1165/ 1560-1923	1110-1923
CYA_1321	ATP-dependent DNA helicase UvrD/PcrA	PcrA	99.3	30/2328	1-608	1-608
CYA_0512	Valyl-tRNA synthetase (EC 6.1.1.9)	ValS	99.4	41/2736	260-700/ 2000-2200	1-700
CYA_1697	Dihydroxy-acid dehydratase (EC 4.2.1.9)	IlvD	99.5	18/1695	1095-1695	1095-1695

[1] MLSA region describes the region of the gene used for the sequence analysis

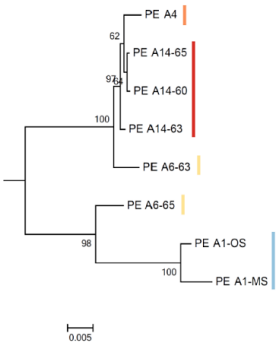


Figure 3.5. Maximum-likelihood phylogeny of genome-informed MLSA of A-lineage isolates. The scale bars describe the branch lengths and are equal to the number of substitutions per site in the sequence. Tree is rooted by the concatenated genes from the *Synechococcus* JA-2-3B'a(2-13) genome (not shown).

gene search, the *dapL* gene did show evidence of recombination. The total number of recombination events detected in the five-gene concatenation also increased from one to four with the addition of the PE A6-65 sequence. The representative PE A6 isolates, as in the orthologous-gene genome phylogeny (Figure 3.1), do not form a clade together in the genome-informed MLSA phylogeny (Figure 3.5). The position of the PE A4 isolate also differs from the orthologous-gene genome phylogeny (Figure 3.1) in this MLSA phylogeny (Figure 3.5), as it falls within, instead of outside of, the PE A14 isolate clade.

Discussion

Single-locus sequence analysis of a region of the *psaA* gene in *Synechococcus* from Mushroom Spring has been used to demarcate PEs in environmental samples (Becraft et al., 2011; submitted), as well as to sequence type cultures (Nowack et al., submitted). Using *Synechococcus* isolate whole-genome sequences, a phylogenetic tree of the PCR-amplified *psaA* region (Figure 3.2B) was compared to a phylogenetic tree constructed using the full *psaA* gene sequence (Figure 3.2A) and the orthologous-gene genome phylogeny (Figure 3.1). The incongruence between the *psaA* amplified-region tree and the orthologous-gene genome tree can be explained by the presence of many SNPs in the un-amplified region of the PE A6-63 isolate gene sequence, which was also detected as a recombinant by RDP4. This, combined with the orthologous-gene genome phylogeny, which separates the two PE A6 isolates, may indicate that the PE A6 *psaA* sequence type has been horizontally transferred between different lineages. This could

also explain the lack of significance for PE A6 high-frequency sequences in canonical correspondence analyses of sequence variation as a function of temperature and vertical distribution in the mat (see Figure 5 in Becraft et al., submitted). Other isolates that formed clades together in both the amplified-region and whole-gene *psaA* sequence (PE A1 and PE A14 representatives) also formed a clade together in the whole genome phylogeny, suggesting that the evolutionary history of *psaA* in these isolates matches the evolutionary history of the organisms.

Melendrez et al. (in prep) used MLSA to create phylogenies of *Synechococcus* environmental BAC clones (see Chapter 1), and found evidence of recombination among the A-lineage sequences, most notably in the *rbsK* locus. When those seven loci were obtained from the sequenced isolate genomes, concatenated, and added into the BAC clone phylogeny, the PE A14 isolates all formed a clade together in the phylogeny, but the two PE A1 isolates did not (Figure 3.4A). When *rbsK* was removed from the analysis, the two PE A1 isolates did form a clade together; however, the three PE A14 isolates were split (Figure 3.4B). Thus, on the one hand, recombination in a gene may create phylogenetic "noise" in one lineage (i.e. separating the PE A1 isolates). On the other hand, recombined genes may act as a strong phylogenetic "signal" in another lineage (i.e. PE A14 isolated clade in seven-gene MLSA phylogeny), concealing "noise" (separation of PE A14-63 isolate from other PE A14 isolates in MLSA phylogeny with *rbsK* omitted) contributed by other genes, as described by Melendrez et al. (in prep). Even when genes are chosen for lack of recombination, the introduction of genes from other isolates to an

analysis may reveal evidence of recombination in a lineage, as in the genome-informed MLSA. Also, incongruence of phylogenies may be difficult to avoid, as comparison of the genome-informed MLSA (Figure 3.5) to the orthologous-gene genome phylogeny (Figure 3.1) shows.

Obtaining genome sequences for multiple isolates within a PE enables examination of the ecological homogeneity of the PEs, through both the comparative analysis of gene-content and gene-selection patterns (see Chapter 5), and through the analysis of single- and multi-locus phylogenies of environmental sequences along with the isolate sequences. Under the assumption that the orthologous-gene genome phylogeny is closest to the "true" phylogeny of the organisms (e.g. Klenk and Göker, 2010), I explored the evolutionary history of loci used in large-scale or environmental phylogenetic analysis of similar organisms. Genes that reflect the "true" evolutionary history of one lineage may blur the evolutionary history of another, as recombination may act as both phylogenetic "signal" and "noise" lineages, and can affect even multi-gene sequence analyses. As with any experimental results, single-gene and multi-gene phylogenies must be treated as hypotheses, and corroborated whenever possible with additional evidence. Based on the phylogenies discussed above, I do not consider the PE A6-63 and PE A6-65 isolates to be replicate isolates of the same PE, but the hypothesized ecotype demarcations still pertain to the PE A1 (PE A1-OS and PE A1-MS) and PE A14 (PE A14-60, PE A14-63, and PE A14-65) isolates. This will be explored in Chapter 5.

CHAPTER FOUR

COMPARATIVE GENOMICS OF *SYNECHOCOCCUS* ISOLATES WITH DIFFERENT LIGHT RESPONSES AND *IN SITU* DIEL TRANSCRIPTION PATTERNS OF ASSOCIATED PUTATIVE ECOTYPES IN THE MUSHROOM SPRING MICROBIAL MAT

Contribution of Authors and Co-Authors

Manuscript in Chapter 4

Author: Millie T. Olsen

Contributions: Cultivated isolates, performed molecular procedures, performed genomic analyses, and wrote the manuscript

Co-Author: Shane Nowack

Contributions: Cultivated isolates and performed growth experiments

Co-Author: Jason M. Wood

Contributions: assisted in analyzing sequence data

Co-Author: Eric D. Becraft

Contributions: Assisted in analyzing sequence data and assisted with molecular procedures

Co-Author: Kurt Labutti

Contributions: Assisted in genome sequencing and assembly

Co-Author: Anna Lipzen

Contributions: Assisted in genome sequencing and assembly

Co-Author: Joel Martin

Contribution of Authors and Co-Authors Continued

Contributions: Assisted in genome sequencing and assembly

Co-Author: Wendy S. Schackwitz

Contributions: Assisted in genome sequencing and assembly

Co-Author: Douglas B. Rusch

Contributions: Acting Co-PI for the Joint Genome Institute Community sequencing project

Co-Author: Frederick M. Cohan

Contributions: Acting Co-PI for the Joint Genome Institute Community sequencing project, discussed results, and edited the manuscript

Co-Author: Donald A. Bryant

Contributions: Acting Co-PI for the Joint Genome Institute Community sequencing project, discussed results, and edited the manuscript

Co-Author: David M. Ward

Contributions: Acting PI for the Joint Genome Institute Community sequencing project, assisted with experimental design, discussed the results, and edited the manuscript at all stages

Manuscript Information Page

Millie T. Olsen, Shane Nowack, Jason M. Wood, Eric D. Becraft, Kurt LaButti, Anna Lipzen, Joel Martin, Wendy S. Schackwitz, Douglas B. Rusch, Frederick M. Cohan, Donald A. Bryant, and David M. Ward

Frontiers in Microbiology

Status of Manuscript:
____ Prepared for submission to a peer-reviewed journal
__x_ Officially submitted to a peer-review journal
____ Accepted by a peer-reviewed journal
____ Published in a peer-reviewed journal

Abstract

Genomes were obtained for three closely related isolates of *Synechococcus* that are representative of putative ecotypes that predominate at different depths in the 1 mm-thick, upper-green layer in the 60°C mat of Mushroom Spring, Yellowstone National Park and exhibit different light adaptation and acclimation responses. Differences in both gene content and orthologous gene alleles between high-light-adapted and low-light-adapted isolates were identified by comparative genomic analyses of these and existing genomes. Evidence of genetic differences that relate to adaptation to light intensity and/or quality, CO_2 uptake, nitrogen metabolism, organic carbon metabolism, and uptake of other nutrients were found between isolates of the different putative ecotypes. *In situ* diel transcription patterns of genes, including genes unique to either low-light-adapted or high-light-adapted isolates and different alleles of an orthologous photosystem gene, revealed that expression is fine-tuned to the different light environments experienced by ecotypes prevalent at various depths in the mat. This study suggests that isolates of closely related putative ecotypes have different genomic adaptations that enable them to inhabit distinct ecological niches while living in close proximity within a microbial community.

Introduction

Thermophilic cyanobacteria of the genus *Synechococcus* predominate in microbial mat communities inhabiting the effluent channels of alkaline, siliceous hot

springs and have been extensively studied and characterized for over 50 years (Peary and Castenholz, 1964; Brock, 1978; Ward et al., 2012). Isolates of *Synechococcus* from mats in Hunters Hot Springs, OR, which were found to exhibit different temperature adaptations, were first observed by Peary and Castenholz (1964). Molecular analyses of Octopus Spring, Yellowstone National Park (YNP) based on 16S rRNA sequences provided further evidence of *Synechococcus* ecotypes--five closely related genotypes were found to be distributed differently from high to low temperatures along the thermal gradient of the effluent channel (Ferris and Ward, 1997). Genotype A″ inhabited the highest temperatures, followed by A′, A, B′, and B as temperatures decreased with distance from the source, though some overlap of adjacent genotypes was observed. Isolates with these 16S rRNA genotypes were shown to have different temperature adaptations that correlated with their distribution along the effluent channel (Allewalt et al., 2006). 16S rRNA genotypes were also found to differ along the vertical aspect of the mat. For example, differential vertical distributions of genotype B′, which occurred above genotype A in the 60°C mat, were observed (Ramsing et al., 2000). However, fluorescence microscopy, combined with estimates of oxygenic photosynthetic rates calculated using microsensor measurements of oxygen in a 68°C mat sample, revealed physiologically distinct populations in the lower and upper parts of the top green mat layer. These populations were identical at the 16S rRNA locus and might have been interpreted as one genotype acclimated differently in response to lower light intensity. However, surface and subsurface populations were genetically distinct at the more

rapidly evolving internal transcribed spacer locus that separates the 16S and the 23S rRNA genes (Ferris et al., 2003), suggesting the possible existence of yet more closely related *Synechococcus* populations with different adaptations to light.

Based on these observations, Ward and Cohan (2005) and Ward et al. (2006) foresaw the need for theory-based models to predict putative ecological species, or putative ecotypes (PEs), from natural variation in sequence data, and for studying genes with even higher molecular resolution than 16S rRNA and the 16S-23S rRNA internal transcribed spacer region. Most recently, *Synechococcus* PEs have been demarcated using highly resolving, protein-encoding loci, from which the evolutionary simulation algorithm Ecotype Simulation (Koeppel et al., 2008) has predicted an even greater number of PEs than 16S-23S rRNA internal transcribed spacer region-based species demarcation (Melendrez et al., 2011; Becraft et al., 2011). As an example, Becraft et al. (submitted) used *psaA* sequence variation and Ecotype Simulation to predict many PEs, including 7 PEs in the A′ lineage, 15 PEs in the A lineage, and 24 PEs in the B′ lineage, several of which were shown to have different vertical distributions in the mat. At 60-63°C, PEs B′9, A1, A4, A14, and A6 were found to be progressively predominant from the mat surface to the bottom of the upper 1 mm-thick green layer (see Figure 3 in Becraft et al., submitted, included in this thesis as Figure 1.7; Table 4.1). Nowack et al. (submitted) were successful in cultivating isolates representative of PEs A1, A4 and A14. They further showed that isolates representative of these PEs have distinctive growth patterns, pigment contents, and low-temperature fluorescence emission spectra when

grown at either high or low irradiance. These differences indicated adaptive and/or acclimative responses to irradiance levels and light qualities that are characteristic of the depth at which each PE predominates *in situ* (see Figures 4 and 8C in Becraft et al., submitted, included in this thesis as Figure 1.8).

While individual genes can be used to predict ecotypes whose unique niches have been inferred from their microhabitat distributions (Becraft et al., submitted), and whose existence can be confirmed by the phenotypes of isolates (Nowack et al., submitted), whole genome comparative analysis can reveal genetic differences among isolates that may be responsible for the adaptive and acclimative mechanisms. For example, the genomes of closely related strains of *Prochlorococcus* spp., which are prevalent phototrophs in marine environments, have been sequenced and compared. Isolates that have different light adaptations maintain differences in gene content related to adaptation to the specific light and nutrient environments of the high-light, surface or low-light, deep-water layers in the ocean (Rocap et al., 2003). *Prochlorococcus* spp. isolates with similar light adaptations also maintained "genomic islands" that may aid in niche differentiation of ecotypes that co-exist within the high-light or the low-light portions of the water column (Coleman et al., 2006; Kettler et al., 2007). Furthermore, genomic analyses are open-ended and unconstrained by the limits of our intuition in that they may reveal unsuspected differences for physiological or metabolic functions that have not yet been tested experimentally. For instance, in previous work the genome of a *Synechococcus* genotype found in downstream regions of these hot spring mats (i.e., B′-

like isolate) was shown to have genes for nitrogen storage and metabolism and for phosphonate utilization that were lacking in an upstream genotype (i.e., A-like), indicating that populations along the flow path differ in adaptations for nutrient metabolism as well as temperature (Bhaya et al., 2007).

In this study, we compared the genomes of three *Synechococcus* isolates within the A lineage that are representative of PEs known to be predominant at different depths in the 60-63°C Mushroom Spring mat (Becraft et al., submitted), and that have different adaptations and acclimative behaviors to low and high irradiance (Nowack et al., submitted). We sought to identify differences in gene content and specific alleles that might underlie these and other adaptations and acclimative responses. We also used these genomes to probe transcript abundances for specific ecotypes using a diel metatranscriptome dataset we had previously obtained (Liu et. al., 2012). We sought evidence of differences in transcription patterns for homologous genes shared among species, which were divergent enough to differentiate PEs, as well as isolate-specific genes, which may be representative of each PE, including genes involved in light harvesting and nutrient uptake. These differences may be indicative of mechanisms underlying adaptive and acclimative responses to light intensity, light quality, and nutrient use that reflect the distinct, ecological niches of these PEs.

Materials and Methods

Synechococcus Isolates

Isolates representative of *Synechococcus* PE A1 (65AY6Li), PE A4 (65AY6A5), and PE A14 (60AY4M2), were selected for comparative genomic analysis. For simplicity we will refer to these isolates by their PE affiliations or by their known adaptions and acclimative responses to low irradiance (isolates of PEs A4 and A14) or high irradiance (isolates of PE A1). The DNA samples used to demonstrate isolate purity by Ti-454 barcode sequencing (Chapter 2; Nowack et al., submitted) were also used for genome sequencing. These genomes were compared to the genome of a second isolate of PE A1 (JA-3-3Ab), previously obtained from the microbial mat of Octopus Spring (Bhaya et al., 2007), which was shown to be adapted to high irradiance (Nowack, 2014). To distinguish between these two PE A1 isolates, we will refer to them here as isolate PE A1-MS (from Mushroom Spring) and PE A1-OS (from Octopus Spring).

Genome Sequencing, Assembly, and Annotation

DNA extracts from each isolate were submitted to the Department of Energy Joint Genome Institute for sequencing and assembly. The DNA from each isolate was randomly sheared into ~270 bp fragments and the resulting fragments were used to create fragment libraries. These libraries were sequenced on Illumina sequencers generating 150-bp paired-end reads. All general aspects of library construction and sequencing are described on the JGI website at http://www.jgi.doe.gov/. Because the variant detection

pipeline requires some non-overlapping paired reads prior to variant detection, the reads were then trimmed to 125 bp. These trimmed reads were then aligned to the reference genome *Synechococcus* sp. JA-3-3Ab using the Burrows-Wheeler Aligner (BWA) (Li and Durban, 2009), and putative single-nucleotide polymorphisms (SNPs) and small indels were identified using samtools and mpileup (Li et al., 2009). Putative structural variants were identified using BreakDancer (Chen et al., 2009), filtering for a confidence score of >90. Genomes were also assembled *de novo* for each strain. Each FASTQ file was QC-filtered for artifact/process contamination and subsequently assembled with AllPathsLG (Gnerre et al., 2011). The resulting contigs contained DNA sequences of heterotrophic contaminants that were found in these isolates (Nowack et al., submitted). These sequences were separated from *Synechococcus* sequences by binning the sequences using NCBI BLASTN and a database of the *Synechococcus* spp. JA-3-3Ab and JA 2-3B′a (2-13) reference genomes (Bhaya et al., 2007), and those of the possible contaminants (*Meiothermus* spp., CP005385; *Anyoxybacillus* spp., CP000922; *Rubrobacter* spp., CP000386) using a method similar to that used in Klatt et al. (2011). The *Synechococcus* DNA assemblies were submitted to the automatic annotation pipelines NCBI PGAAP [NCBI Handbook (Internet) 2nd edition (Tatusova et al., 2013)] and RAST (Aziz et al., 2008) for annotation using the default parameters. The three draft genome sequences have been submitted to Genbank under the following accession numbers: PRJNA209725 (*Synechococcus* sp. 65AY6Li, PE A1), PRJNA210217

(*Synechococcus* sp. 65AY6A5, PE A4), and PRJNA210214 (*Synechococcus* sp. 60AY4M2, PE A14).

Comparative Analyses

The isolate genome phylogeny was obtained using a concatenation of 460 marker proteins identified by Phyla-AMPHORA (Wang and Wu, 2013), which uses phylum specific conserved proteins for metagenomic phylotyping; only conserved cyanobacterial proteins present in all six genomes were selected (highlighted in Supplementary Table A.1, Appendix A). Proteins were aligned with ClustalO (Sievers et al., 2011) and a Newick tree was computed using FastTree, with local support values calculated using the Shimodaira_Hasegawa test (Price et al., 2009). Comparative analyses of genomes were conducted using the RAST SEEDViewer for gene content analyses (Overbeek et al., 2013) and the best-hit and reciprocal best-hit average nucleotide identity (ANI) calculator (Goris et al., 2007).

Metatranscriptomic Analyses

Diel metatranscriptomic datasets described by Liu et al. (2012) were reanalyzed by BWA (Li and Durban, 2009) to locate transcripts for specific genes. Genes targeted in these analyses were chosen either because they are specific to low-irradiance- or high-irradiance-adapted isolates or, in the case of orthologous genes, because they exhibited at least 3% nucleotide sequence difference between/among homologous genes in different isolates. We used the methods described in Liu et al. (2011, 2012), except that (i) we used

genomes instead of metagenomic assemblies to recruit transcripts, and (ii) transcripts associated with the unique alleles of different ecotypes were recruited without allowing any mismatches (i.e., an exact sequence match was required). Recruitment of transcripts associated with B′-like *Synechococcus* was done using the published *Synechococcus* B′ genome (Bhaya et al., 2007), but, as in Liu et al. (2012), up to 5 mismatches were allowed. This was done because this genome is not representative of the predominant B′ PE (B′9) in the 60°C mat (Becraft et al., submitted), which we do not yet have in culture (Nowack et al., this issue). Raw transcript counts were normalized by the total number of mRNA-specific transcripts at each time point and then by the geometric mean of normalized transcript counts across all time points (Liu et al., 2011, 2012).

Results

Genomic Properties

Basic characteristics of the genomes of isolates representative of PEs A1-OS, A1-MS, A4, and A14 are presented in Table 4.1. All of the isolates are >99.9% identical at the 16S rRNA locus and share 2201 orthologous genes as their core genome, including most RAST annotated subsystem genes found in the A1-OS reference genome. The ANI among orthologous genes in the different isolates ranged from 98.35-99.32%, so many of the shared genes predict proteins with 100% sequence identity and may be functionally identical (Table 4.2).

The GC content of the isolates ranged from 60.2-60.4%, while genome sizes varied from 2.93 Mbp for isolates of PEs A1-OS and A1-MS to 3.16 Mbp for the PE A14

Table 4.1. Genomic, phenotypic, and environmental information for *Synechococcus* isolates of putative ecotypes with different depth distributions or from different hot spring mats.

Isolate	**A4**	**A14**	**A1-MS**	**A1-OS**
Strain Name	65AY6A5	60AY4M2	65AY6Li	JA-3-3Ab
Putative Ecotype (PE)	A4	A14	A1	A1
Isolation Source	Mushroom Spring	Mushroom Spring	Mushroom Spring	Octopus Spring
Isolation Date	September 2010	September 2010	September 2010	July 2002
Isolation Temperature (°C)	65	60	65	58-65[1]
PE Relative Abundance ≥5% at 60°C Mushroom Spring	400-800 µm	400-960 µm	0-960 µm	0-960 µm
PE Relative Abundance ≥10% at 60°C Mushroom Spring	640-720 µm	560-960 µm	0-800 µm	0-800 µm
Light Adaptation	Low-light	Low-light	High-light	High-light
Sequencing Method	Illumina	Illumina	Illumina	Sanger
Depth of coverage	22x	35x	16x	2x
Length (Mbp)	2.98	3.16	2.93	2.93
Number of Contigs	9	6	2	1
Largest Contig (bp)	2508234	3142301	2795989	2932766
%GC Content	60.4	60.4	60.3	60.2
CDS	2622	2597	2623	2760
tRNAs	47	47	49[2]	47
rRNA operons	2	2	2	2
Number of Genes Unique to the Isolate with Respect to the Reference A1-OS	173	204	131	N/A

[1] Temperatures in Octopus Spring fluctuate continuously over a 4.5 minute cycle (Miller et al., 1998), therefore the 7°C range of the isolation site is given.
[2] A1-MS contains a duplication of the 23S rRNA locus in one operon.

Table 4.2. Average nucleotide index (ANI) and genes with identical amino acid sequence in *Synechococcus* isolate genomes. The ANI between each pair of genomes, using whole-genome reciprocal best hits is shown below the diagonal. The percentage of genes that encode for proteins with identical amino acid sequences (and may be functionally identical) between each pair of genomes is shown above the diagonal.

	A4	**A14**	**A1-MS**	**A1-OS**
A4	-	42.65%	37.65%	36.63%
A14	99.05%	-	37.08%	43.77%
A1-MS	98.41%	98.72%	-	55.07%
A1-OS	98.42%	98.35%	99.32%	-

isolate (Table 4.1). The isolate genome phylogeny reflects the *psaA* phylogeny (see Figure 1 in Becraft et al. (submitted)), with the high-light-adapted isolates and the low-light-adapted isolates each forming a distinct clade (Figure 4.1). Although all sequenced A-lineage isolates are very closely related, there are several genomic differences among the isolates that may underlie the niche differentiation among the PEs. Though the three isolates were selected because they are representative of PEs that differ in vertical position and exhibit different adaptations to irradiance, the low-light-adapted isolates of PEs A4 and A14 were similar to each other phenotypically (see Nowack et al., submitted) and very different from the two high-light-adapted isolates A1-MS and A1-OS. Hence, this discussion will focus on differences between the two low-light-adapted isolates and the two high-light-adapted isolates. Differences discussed in the main text, including any subsystem genes missing in the newly sequenced isolates, are presented in Table 4.3 and a full ortholog table, which also presents differences in the percentage amino acid identity of all homologous genes, is presented in Supplementary Table A1. Specific differences between the genomes of PE A4 and A14 isolates will also be considered below.

Table 4.3. Ortholog table showing discussed gene content differences among isolates with different light adaptations and representative of putative ecotypes with different vertical positioning in the 60-63°C Mushroom Spring mat.

A4	A14	A1-MS	A1-OS
Chlorophyll a(b) binding protein, photosystem II CP43 protein homolog *isiX*/ Allophycocyanin subunit *apcB*3/ Allophycocyanin subunit *apcD4*/ Hypothetical protein (putative photoreceptor)	Chlorophyll a(b) binding protein, photosystem II CP43 protein homolog *isiX*/ Allophycocyanin subunit *apcB*3/ Allophycocyanin subunit *apcD4*/ Hypothetical protein (putative photoreceptor)		
Ferrous iron transport cassette *feoAB* (2 genes)	Ferrous iron transport cassette *feoAB* (2 genes)		
Maltose/maltodextrin transport cluster *malK* (4 genes)	Maltose/maltodextrin transport cluster *malK* (4 genes)		
Ammonium transporter *amtB2*	Ammonium transporter *amtB2*		
Ammonium transporter *amtB1*	Ammonium transporter *amtB1*	Ammonium transporter *amtB1*	Ammonium transporter *amtB1*
Assimilatory nitrate reductase *narB*	Assimilatory nitrate reductase *narB*	Assimilatory nitrate reductase *narB* (inactivated)	Assimilatory nitrate reductase *narB* (inactivated)
Methyl-accepting chemotaxis protein	Methyl-accepting chemotaxis protein	Methyl-accepting chemotaxis protein (inactivated)	Methyl-accepting chemotaxis protein (inactivated)
		Carbonic Anhydrase	Carbonic Anhydrase
		Peptide/opine/ nickel PepT ABC Transport Cassette (5 genes)	Peptide/opine/ nickel PepT ABC Transport Cassette (5 genes)
		Urease Cluster 1 (5 genes)	Urease Cluster 1 (5 genes)
		Cystine ABC Transporter (2 genes)	Cystine ABC Transporter (2 genes)
		Succinate dehydrogenase flavoprotein subunit *sdhA*/ Omega-amino acid-pyruvate amino-transferase	Succinate dehydrogenase flavoprotein subunit *sdhA*/ Omega-amino acid-pyruvate amino-transferase
		Type III CRISPR/*cas* array	Type III CRISPR/*cas* array
Urea Carboxylase Cassette (7 genes)	Urea Carboxylase Cassette (7 genes)	Urea Carboxylase Cassette (7 genes)	
Polar amino acid ABC Transport (PAAT) (3 genes)		Polar amino acid ABC Transport (PAAT) (3 genes)	Polar amino acid ABC Transport (PAAT) (3 genes)
	Sugar ABC Transport Cassette (3 genes)	Sugar ABC Transport Cassette (3 genes)	Sugar ABC Transport Cassette (3 genes)
Spermidine Putrescine ABC Transporter PotABCD (6 genes)			
Beta-carotene ketolase			
Formamidase *amiF*/Amidase Gene Cluster (8 genes)			
		Type I restriction-modification system/SS exonuclease associated with Rad50/Mre11 complex/Bipolar DNA Helicase	

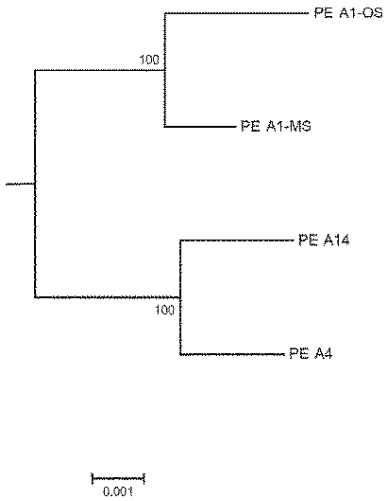


Figure 4.1. Phylogeny based on a concatenation of 460 proteins of A-lineage isolates conserved in the cyanobacterial phylum. The tree was rooted by orthologous proteins from the JA-2-3B′a (2-13) genome and the scale bar is equal to 0.001 amino acid substitutions/site.

Genes Found Only in
Low-Light-Adapted Isolates

The low-light-adapted isolates representative of PEs A4 and A14 have a unique, possibly horizontally acquired gene cluster that includes: a potential photoreceptor predicted to have 4 PAS domains (see Taylor and Zhulin, 1999), 2 GAF domains (see Ho et al., 2000), and a histidine kinase domain, which could act as a light-activated response regulator; *apcD4* and *apcB3* genes, which are predicted to encode a variant allophycocyanin that probably has enhanced far-red absorption (670 to 710 nm absorption maximum) (Gan et al., 2014b); and a gene for an IsiA-like protein, which we have tentatively named IsiX and which probably forms a novel, chlorophyll (Chl) *a*-binding antenna complex distinct from that formed by the iron-starvation induced protein, IsiA (Table 4.3). ApcD4 is approximately 40% identical and 62% similar to ApcD1

(CYA_2790) in these *Synechococcus* strains and is ~63% similar to ApcA (CYA_2227) of the PE A1-OS isolate; ApcB3 is ~80% similar to the ApcB (CYA_2226) of the PE A1-OS isolate. In contrast, the products of the *apcA* and *apcB* genes, which are not located on this cluster, are highly conserved in all four isolate genomes (100% identical). The photosystem (PS) II core subunit PsbC, which is >99% similar among all four isolates, is only 48% similar to IsiX, which has a C-terminal extension of nearly 100 amino acids and probably one additional transmembrane helix relative to other PsbC/IsiA proteins. Moreover, IsiX is only 46% similar to IsiA (CYA_2606), which is likely to be iron-regulated because of its co-localization with *isiB*, encoding flavodoxin (IsiB; CYA_2605), as observed in most other cyanobacteria (Straus, 1994). These observations are consistent with the idea that ApcD4-ApcB3 and IsiX produce specialized antenna proteins that function in PE A4 and A14 isolates under low irradiance or possibly far-red light (or both) conditions. Consistent with this idea, we have noted that isolates corresponding to PEs A4 and A14 have enhanced absorption above 700 nm compared to the PE A1-MS isolate (data not shown, but see Nowack et al., submitted). However, at this point it is not yet clearly established whether these proteins are linked to these differences. Transcript abundances for genes encoding components of the photosynthetic apparatus in the Mushroom Spring mat *Synechococcus* generally rise sharply at sunrise, are maximal during the mid-day, and decline in the late afternoon (see Figure 4D, Liu et al., 2012). This pattern is observed for *psbC*, which encodes a core subunit of the PS II reaction center (Figure 4.2A). However, the transcript abundance for *isiX* has a different

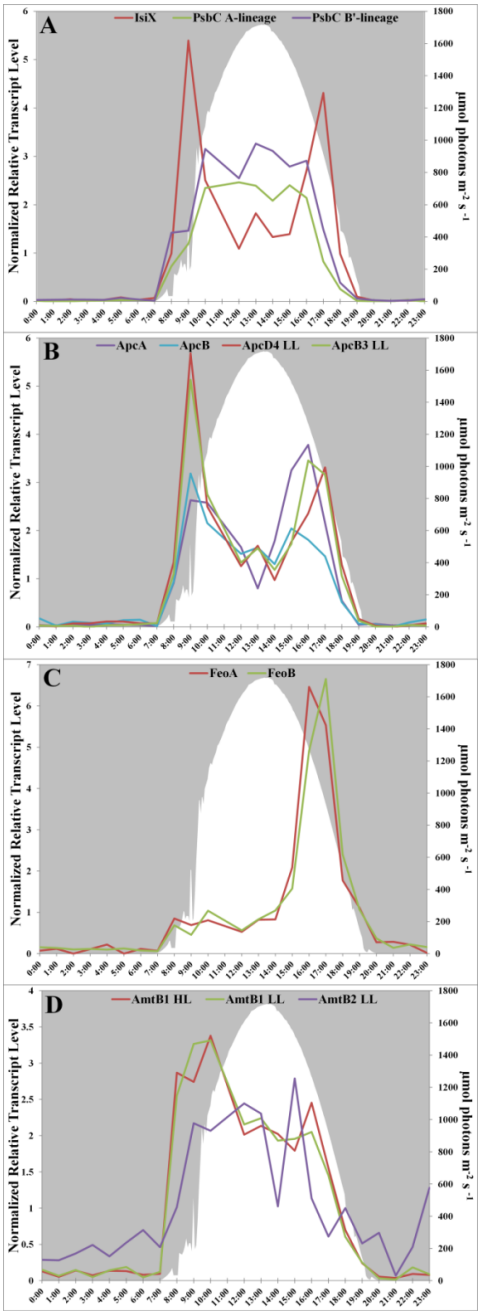


Figure 4.2. Transcription patterns of transcripts encoding (A) PsbC in all A-lineage isolates and IsiX, present in PE A4 and A14 isolates, (B) ApcA/ApcB, present in all A-lineage isolates, and ApcD4 LL/ApcB3 LL, present in PE A4 and A14 isolates, (C) FeoA and FeoB, present in PE A4 and A14 isolates, and (D) ammonium transporter genes in A1 (*amtB1*), PE A4/A14 (*amtB1* LL), and duplication (*amtB2* LL) in A4/A14. All panels show downwelling irradiance (µmol photons $m^{-2}s^{-1}$) measured at Mushroom Spring from September 11-12, 2009.

pattern and is most abundant during the low-irradiance periods in the early morning (07:00 to 10:00) and late afternoon (15:00 to 19:00) (Figure 4.2A). Transcripts for the *apcD4* and *apcB3* genes have a similar overall abundance pattern to *isiX* (Figure 4.2B). Although transcripts for *apcA*, *apcB*, *apcD4*, and *apcB3* were all maximal at the same time in the morning (09:00), transcripts for *apcD4* and *apcB3* were maximal about an hour later than transcripts for *apcA* and *apcB* in the late afternoon period.

Another gene cassette unique to the low-light-adapted isolates representative of PEs A4 and A14 contains the *feoAB* genes, which encode subunits of a ferrous iron transporter. The transcript abundances for these genes are maximal in the late afternoon, when the mat is becoming anoxic (Figure 4.2C; and see Figure 8C in Becraft et al., submitted, included as Figure 1.8 in this thesis). The most closely related FeoAB protein sequences are found in other cyanobacteria, but it is not clear if the genes were acquired by horizontal gene transfer, or lost in the high-light-adapted isolates, which are representative of a PE that predominates in the more oxic portions of the mat. Isolates of PEs A14 and A4 also share an ABC transporter cassette for sugar (possibly maltose/maltodextrin) transport and a methyl-accepting chemotaxis protein, one of many copies found in all genomes. The latter gene has two identical in-frame stop codons in genomes of the high-light-adapted PE A1-OS and A1-MS isolates.

In addition to horizontal gene transfer, gene duplication and subsequent nucleotide divergence can provide novel functionality to an organism, even though the resulting variant protein retains homology to the original product. All four isolates

contain a gene (*amtB1*) encoding a putative ammonium transporter and the predicted AmtB1 proteins are ~90% identical. However, the isolates of PEs A4 and A14 additionally contain a paralogous gene that has apparently arisen by duplication and divergence: AmtB2 is ~70% identical to AmtB1. The transcription patterns of the *amtB1* gene in the PE A1 isolates and in the PE A4 and A14 isolates are comparable, but the transcription pattern of the *amtB2* gene in the PE A4 and A14 isolates differ from the *amtB1* pattern. This gene has a transcription pattern similar to many genes for components of the photosynthetic apparatus (Liu et al., 2012) and largely reflects the light period except for a late-afternoon decline (Figure 4.2D). Functional studies have shown that AmtB can transport both NH_3 and CO_2 (Musa-Aziz et al., 2009), and it is possible that these variants are functionally differentiated with respect to substrate. Because the transcript abundance pattern mirrors photosynthetic activity in the mat, this pattern is consistent with the possibility that AmtB2 could be a CO_2 transporter. Alternatively, AmtB2 could transport ammonium but have a high affinity for the substrate. In addition to the duplicated *amtB* genes, isolates of PEs A4 and A14 contain a second copy of *narB*, encoding assimilatory nitrate reductase, which is transcribed diurnally (Supplementary Figure A.2). This copy is unlinked and divergent (~68% amino acid identity) from the nitrate reductase of the *nirA-narB* gene cluster found in all of the isolates. Although this gene is also found in the genomes of the PE A1-MS and A1-OS isolates, it is disrupted by a mobile element gene and thus is not likely to be active.

Genes Found Only in High-Light-Adapted Isolates

The genomes of high-light-adapted isolates PE A1-MS and PE A1-OS possess a copy of a carbonic anhydrase gene that has 68% amino acid identity to the zinc-dependent, gamma-class carbonic anhydrase found in *Thermosynechococcus* sp. NK55a. Carbonic anhydrase catalyzes the interconversion of carbon dioxide and bicarbonate by a reversible hydration reaction, and while the carbon-concentrating mechanism (CCM, present in all four isolates) also includes a different carbonic anhydrase, cyanobacteria (Cannon et al., 2010) and other prokaryotes (Smith and Ferry, 2000) can have multiple copies of the genes and multiple classes of the enzyme that may play different functional roles in photosynthesis. The expression of this gene was too low to confidently ascertain its transcription pattern.

The urease cassette (Cluster 1 urease in Bhaya et al., 2007) found in the genome of the PE A1-OS isolate is also found in the PE A1-MS isolate genome, but these genes are not present in the genomes of the low-light adapted PE A4 and A14 isolates. This urease cassette includes the genes that encode the larger alpha subunit UreC, smaller beta and gamma subunits UreB and UreA, which form the heterotrimeric urease enzyme, and UreDEFG accessory proteins that aid in assembly of the nickel metallocenter of the enzyme (Farrugia et al., 2013). All of the genes in the urease cassette have >90% identity to the urease genes found in *Thermus islandicus*, which indicates a possible horizontal gene transfer of this cassette to an ancestor common to the high-light-adapted isolates but not the low-light-adapted isolates.

The PE A1-OS and A1-MS isolate genomes have a five-gene cluster annotated as a peptide/opine/nickel ABC transporter (PepT family), which includes a periplasmic substrate-binding protein, two permease subunits, and two ATP-binding protein genes. Additionally, the PE A1-OS and A1-MS isolate genomes possess two components of a cystine ABC transporter, genes encoding the periplasmic cystine binding protein and the permease protein, as well as two genes, flanked by genes for mobile element proteins, that are annotated as succinate dehydrogenase flavoprotein subunit *sdhA* and omega-amino acid-pyruvate aminotransferase. Transcript abundances for all of these genes are higher during the light period and lower at night, similar to other genes that are expressed during the day (Supplementary Figure A.1). Finally, along with the Type I and Type II CRISPR/*cas* arrays that are conserved among all four isolates, the PE A1-MS genome contains a Type III CRISPR/*cas* array previously found in the PE A1-OS isolate genome by Heidelberg et al. (2009). This is a unique CRISPR/*cas* array that is shared by *Roseiflexus* sp. RS-1, an anoxygenic photosynthetic organism that is also abundant in these microbial mat communities (Klatt et al., 2011). Although the amino acid similarities of the homologous genes are only 40-66% between the two organisms, there are transposons flanking the array in the PE A1 isolate genomes, which suggests a possible, if not recent, lateral gene transfer event in the mat (Heidelberg et al., 2009).

PsbA Allele Differences between High-Light- and Low-Light-Adapted Isolates

The genome of the high-light-adapted PE A1-OS isolate encodes four *psbA* genes, CYA_1274, CYA_1748, CYA_1811, and CYA_1894, while the B′ genome (JA-2-3B′a (2-13)) of *Synechococcus* has three *psbA* genes, designated CYB_0216, CYB_0371, and CYB_0433 (Bhaya et al., 2007). CYA_1274, CYA_1811, CYA_1849, CYB_0371, and CYB_0433 are nearly identical and differ by only 1 or 2 conserved amino acids. CYA_1748 and CYB_0216 are very similar to one another (94% identity, 96% similarity) but are only about 73% identical and 85% similar to the other PsbA sequences. These latter sequences have been called "rogue PsbA" sequences (rPsbA) by Murray (2012). Rogue PsbA sequences lack key functional residues and thus are not expected to support oxygen evolution by PS II complexes that might contain them. The new isolate genomes also possess multiple copies of the *psbA* gene: the PE A1-MS isolate genome has four copies, which appear to be orthologous to those in the PE A1-OS isolate genome, while the PE A4 and A14 isolate genomes each have three *psbA* genes, two of which are identical to CYA_1274 and CYA_1849, as well as a copy of the *rpsbA* gene. The gene encoding rPsbA, CYA_1748, is sufficiently divergent to differentiate between the high-light- and low-light-adapted isolates (the PE A4 and A14 alleles are 90% and 91% identical to the *rpsbA* gene in the PE A1-MS isolate, respectively). Interestingly, the low-light- and high-light-adapted isolates exhibited similar, but clearly temporally offset transcript abundance patterns (Figure 4.3), with the transcripts of low-light-adapted isolates declining later in the morning and increasing earlier in the late

afternoon. The transcript abundance for *rpsbA* (CYB_0216) of the B′-lineage *Synechococcus* also declines earlier in the morning than those of the low-light adapted isolates, but they increase even later than transcripts of PE A1 isolates.

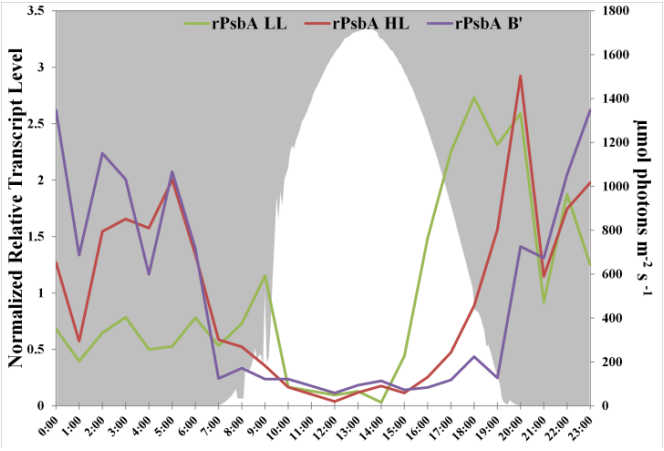


Figure 4.3. *rpsbA* diel transcription patterns of alleles from the (i) B′ lineage, (ii) populations represented by high-light adapted isolates (HL) of PE A1-OS and A1-MS, and (iii) populations represented by low-light adapted isolates (LL) of PEs A4 and A14. Downwelling irradiance (µmol photons $m^{-2}s^{-1}$) measured at Mushroom Spring on September 11-12, 2009 is also shown.

Other Gene Content

Differences among Isolates

The genomes of isolates PEs A1-MS, A4, and A14 encode the genes necessary to synthesize urea carboxylase, including urea carboxylase/allophanate hydrolase, two urea carboxylase-related aminomethyltransferases, three genes for a urea carboxylase-related ABC transporter, and a biotin-protein ligase gene (Table 4.3). The urea carboxylase cluster proteins have between 60-85% identity to proteins found in other cyanobacteria, so it is not clear if the cluster was acquired horizontally or vertically. Isolates of PEs A1-OS, A1-MS, and A4 share a three-gene cassette for an ABC transporter for polar amino

acids that is not found in the PE A14 isolate (Table 4.3). This is a potentially interesting difference because aspartate and glutamate are the only two amino acids that are not taken up and metabolized by *Chloracidobacterium thermophilum*, which is co-localized with the low-irradiance-adapted *Synechococcus* that occur deeper in the mat (see Tank and Bryant, 2015). The PE A1-MS isolate also has a gene cluster that is not found in the PE A1-OS isolate and that consists of genes predicted to encode a bipolar DNA helicase, a Type I restriction-modification system DNA-methyltransferase subunit M, and the single-stranded exonuclease associated with Rad50/Mre11 complex (Table 4.3). The PE A14 isolate possesses a PotABCD cassette for spermidine/putrescine transport that is not found in any of the other isolates, which could provide an alternative nitrogen source for this isolate. The proteins are most similar to PotABCD proteins in alpha- and gamma-proteobacteria, which may be indicative of a lateral gene transfer event. Additionally, the genome of the PE A14 isolate encodes beta-carotene ketolase (CrtO), which is also encoded in the genome of the B′ isolate JA-2-3B′a(2-13) (Bhaya et al., 2007). Keto-carotenoids provide better protection from reactive oxygen species than hydroxylated xanthophyll derivatives and are differently localized in membranes than other xanthophyll derivatives (Zhu et al., 2010). Further suggesting functional differences among these ecological species, transcripts were found for all of the isolate-specific genes *in situ*.

Discussion

In this study we compared the genomes of isolates from extremely closely related yet ecologically distinct PEs, each of which has a unique distribution along the vertical gradient at 60-63°C (Becraft et al., submitted) and differing light adaptations and acclimation responses corresponding to their vertical distributions (Nowack et al., submitted). Our aim was to discover the genetic bases for the physiological differences that cause these organisms to occupy different niches along the vertical gradient. Some of the most conspicuous differences are found between the high-light-adapted isolates of PE A1 and the low-light-adapted isolates of PEs A4 and A14. Becraft et al. (submitted) showed that PE A1 predominates in the upper to middle part (0-760 µm deep) of the upper green layer of the mat, while PEs A4 and A14 are most abundant in deeper layers of the mat (640-720 µm and 640-960 µm, respectively; Figure 1.7). The difference in scalar irradiance received by the different populations is striking (see Figure 8C in Becraft et al, submitted; Figure 1.8); while members of PE A1 may experience up to 1250 µmol photons $m^{-2}s^{-1}$ scalar irradiance, PE A4 and A14 populations may only experience 50-75 µmol photons $m^{-2}s^{-1}$ at the peak irradiance level during a diel cycle. These ecophysiological differences are reflected in the gene contents of these organisms.

Gene content differences suggest different adaptations for the high-light- and low-light-adapted organisms. The low-light-adapted isolates of PEs A4 and A14 possess a gene cluster with xenologous copies of *apcD4*, *apcB3*, and *isiX*, all of which are highly expressed *in situ*. Genes in this cassette are most likely responsible for the long-wave

absorption and fluorescence emission features observed in those isolates when grown at low irradiance but missing in high-light-adapted organisms (Nowack et al., submitted). This would be consistent with selection pressure to improve light harvesting when the ambient light is strongly filtered by Chl *a* and phycobiliproteins by organisms in the upper regions of the mat and by the greater relative abundance of far-red light at increasing depth in the mat (see Figure 4 in Becraft et al., submitted). Some of this shift to the far-red would simply be due to greater penetration by light of longer wavelengths, which is less readily scattered. This gene cassette is also found in several other cyanobacteria (see Shih et al., 2013, Figure S5 CP43 phylogeny, members of clade CBPVI; and Gan et al., 2014b), including *Chlorogloeopsis* spp. PCC 6912 and 9212, *Fischerella* sp. PCC 9605, *Chroococcidiopsis thermalis* PCC 7203, *Gloeocapsa* sp. PCC 7428, *Xenococcus* sp. 7305, and *Leptolyngbya* sp. PCC 6406. These genes may encode a common adaptive mechanism among low-light-adapted cyanobacteria that are primarily found in benthic or terrestrial environments, by enabling them to acclimate to low irradiance conditions and/or to far-red light. Only three genes, *apcD4*, *apcB3*, and *isiX*, are required, which is far simpler than the FaRLiP response recently described by Gan et al. (2014a) that involves 17 genes and changes to all three major photosynthetic complexes. Interestingly, several organisms that can perform FaRLiP (Gan et al., 2014a, b) also have this simpler system, which strongly suggests that the systems, at least in those cyanobacteria that have both capabilities, respond to different light cues.

The PE A4 and A14 isolates additionally contain the *feoAB* genes for the ferrous iron transport system, which were initially described to be present in the metagenomes of the Mushroom Spring and Octopus Spring mats. The transcription pattern of the genes in the metatranscriptome (Liu et al., 2012, and Figure 4.2C) match the transcription pattern of *feoB* measured with q-RT-PCR over a diel cycle by Bhaya et al. (2007). Under alkaline conditions, ferrous iron is only present in the absence of oxygen, which may occur more often in the deeper parts of the mat, away from the higher levels of oxygen in the upper part of the mat that are produced by *Synechococcus* populations experiencing higher irradiance levels and longer periods of exposure to light (Jensen et al., 2010). Interestingly, the *feoAB* genes discovered by Bhaya et al. (2007) were found on metagenomic clones that were most closely related to the B′-lineage, which may indicate the existence of low-light adapted B′-lineage ecotypes as well as low-light-adapted A-lineage PEs. This might explain the inability of the B′-like isolate studied by Kilian et al. (2007) to grow at high irradiance, if it contained only low-light-adapted, B′-lineage ecotypes.

In contrast to the low-light-adapted isolates, the high-light-adapted isolates PE A1-OS and A1-MS both contain an extra carbonic anhydrase gene, which may enhance growth under CO_2-limiting conditions when bicarbonate is present. The extra carbonic anhydrase may enhance conversion of bicarbonate to CO_2. CO_2 limitation caused by high rates of photosynthesis during peak irradiance has been indicated by an increase in pH when rates of oxygenic photosynthesis are high (Jensen et al., 2010). This observation led

us to demonstrate that the growth rate of the PE A1-MS isolate, but not an isolate without the extra carbonic anhydrase gene, was increased by the addition of bicarbonate under carbon-limiting conditions (Supplementary Figure A.3 and Supplementary Data in Appendix A), which implies that this gene may provide increased fitness under such conditions. The high-light-adapted PE A1-OS and A1-MS isolates also have unique genes involved in the TCA cycle (*sdhA*) and virus infection (Type III CRISPR/*cas* array), which may be indicative of uncharacterized environmental realities of the high-light-adapted isolates compared to the low-light-adapted isolates.

Transcription patterns differ for genes associated with isolates representative of different PEs. We were able to exploit the relatively high sequence divergence of the *rpsbA* gene to show that the transcription timing of this gene by low-light-adapted PE A4 and A14 populations found deepest in the mat green layer differed from that of the high-light-adapted PE A1 population residing above them. Specifically, transcription in PEs found deeper in the mat started earlier in the afternoon and ended later in the morning. Jensen et al. (2010) reported a similar transcription pattern for this gene in B′-like *Synechococcus*. Furthermore, by recruiting B′-like transcripts from the metatranscriptome, we were able to show that B′-like populations in the 60°C mat, which have been shown to predominate in the uppermost part of the mat green layer (see Figure 3 of Becraft et al., submitted; Ramsing et al., 2000), express *rpsbA* genes even later in the afternoon and have declining transcript abundances for these genes even earlier in the

morning. Similarly, Becraft et al. (this issue) reported offsets in the timing of B′-like and A-like expression of photosynthesis and nitrogen fixation genes.

The function of rogue-PsbA in PS II has not yet been established, but because this subunit is missing essential amino acid residues for the Mn_4CaO_5 cluster of the water oxidation center and has key differences in the binding pocket for quinone Q_B, it seems unlikely that PS II complexes containing this protein can oxidize water (Murray, 2012). Considering that transcript abundance pattern for this gene is similar to those for nitrogen fixation genes (Figure 8B in Becraft et al., submitted, included in this thesis as Figure 1.8), and that transcripts for "typical" *psbA* alleles increase rapidly as nitrogen fixation wanes and photosynthesis increases, we hypothesize that rPsbA subunits are involved in the oxidation of sulfide, which is present in the mats due to sulfate reduction during periods of anoxia (van der Meer et al., 2005; Dillon et al., 2007). Although *Synechococcus* lacks sulfide quinone reductase, which occurs in some cyanobacteria that oxidize sulfide to polysulfide (e.g., *Oscillatoria limnetica*; Arieli et al., 1991, 1994), most cyanobacteria that oxidize sulfide actually produce thiosulfate as the sole product in a reaction that has never been fully characterized biochemically (de Wit and van Gemerden, 1987; Rabbenstein et al., 1995). We hypothesize that rPsbA is involved in the oxidation of sulfide to thiosulfate, and that this process could provide electrons for nitrogen fixation by nitrogenase, which would otherwise be inactivated by oxygen production if PS II contained "typical" PsbA subunits. This scenario is further supported and is completely consistent with previous results suggesting that sulfide stimulated early

morning incorporation of CO_2 into cyanobacterial lipids (van der Meer et al., 2005). Such a process would be expected to occur under anoxic conditions, which occur earlier in deeper portions of the mat (see Figure 8C in Becraft et al., submitted, included in this thesis as Figure 1.8).

Besides parameters linked to photosynthesis (light intensity and quality, CO_2 acquisition) several other parameters appear to be involved in defining niches that different *Synechococcus* putative ecotypes may occupy. Nitrogen metabolism appears to differ among the isolates, even between the two isolates representative of PE A1. Both PE A1 isolates are capable of urea degradation with urease, while isolates of PEs A1-MS, A4, and A14 have urea carboxylase. This may reflect a difference in the chemistry of Octopus Spring and Mushroom Spring; urea degradation with urea carboxylase involves two separate reactions and is ATP-dependent, while urease involves only one reaction and is not ATP-dependent, but requires nickel for the enzyme metallocenter (Sakamoto and Bryant, 2001; Solomon et al., 2010; Farrugia et al., 2013). Rates of urea uptake are usually higher than for nitrate or nitrite, even when the concentration of these oxidized nitrogen sources is higher, and urea is preferable in CO_2-limited environments because CO_2 is a useful by-product of urea assimilation (Solomon et al., 2010). Different genes for nutrient uptake may also reflect different strategies for nitrogen metabolism. The peptide/opine/nickel transport cassette in PE A1-OS and A1-MS isolates may provide the nickel for the urease enzyme when available, or it might be involved in scavenging of environmental peptides or opines as a source of both nitrogen and organic carbon.

Similarly, the cystine transport genes in PE A1-OS and A1-MS isolates, the polar amino acid transport cassette in PE A1-OS, A1-MS, and A4 isolates, and the PotABCD spermidine/putrescine transporter in the PE A14 isolate are all transcribed *in situ*, and all transport possible sources of nitrogen into the cells. Other gene content differences among isolates may indicate differences in organic carbon use (the putative MalK transport cassette in isolates of PEs A4 and A14) and DNA protection and repair (bipolar DNA helicase and single-strand exonuclease in the PE A1-MS isolate and beta-carotene ketolase in the PE A14 isolate). Some of these gene content differences may help to explain the niche differentiation between the two low-light-adapted isolates of PEs A4 and A14. Although both grow faster at lower irradiances than the PE A1 isolates and are thus characterized as low-light-adapted, they do have different patterns of growth relative to light intensity (Nowack et al., submitted) and different vertical distributions in the mat (Becraft et al., submitted). The PE A4 distribution is maximal in the lower-middle part of the mat upper green layer, while PE A14 is maximal at the greatest depths where irradiance is most attenuated.

Woese and Fox (1977) used the highly conserved 16S ribosomal RNA sequence to estimate phylogenetic relatedness among organisms. 16S rRNA sequencing revealed several inaccuracies of traditional classification methods, among which was the complete oversight of the domain Archaea (Balch et al., 1977). However, the extensive use of 16S rRNA sequencing has led to a somewhat arbitrary molecular demarcation of microbial species that is widely accepted and used by many microbiologists. Molecular cutoffs

were created by observing the sequence divergence among isolates of classically named species (e.g. Seki et al. (1978) within the genus *Bacillus*): that a >3% divergence of the 16S rRNA locus between two organisms (Stackebrandt and Goebel, 1994) or, more recently, >1% divergence at the 16S rRNA locus (Stackebrandt and Ebers, 2006) is required to consider that the two isolates belong to different species. Using the highly resolving locus *psaA*, we have (i) predicted the existence of different putative ecological species within traditional 16S rRNA-defined species using a theory-based model (Becraft et al., 2011), (ii) shown that they are ecologically distinct through differences in distribution, (iii) shown that many contain ecologically homogeneous members (Becraft et al., submitted), (iv) shown that isolates representative of different PEs have different adaptations and acclimation responses to light (Nowack et al., submitted), and (v) through comparative genomic analysis of these isolates, shown that isolates of the different PEs contain differences in gene content and gene alleles which appear to underlie the adaptations and acclimation responses of each PE to their distinct ecological niches. Comparative analyses of multiple isolates of different PEs, including analyses of genes under selection within and between ecotypes, will provide further evidence for the ecological differentiation among PEs and will demonstrate whether the adaptations and acclimative responses of these isolates are typical of members of a PE. These results, along with the differences in the timing of gene expression by different PEs located in distinct niches, demonstrate that ecologically adapted populations, which may in fact be true ecological species, matter in microbial communities.

Acknowledgements

This research was largely supported by the U.S. Department of Energy (DOE) through (i) a Community Sequencing Project from the Joint Genome Institute (proposal #CSP148; user agreement NPUSR006316), and (ii) The Office of Biological and Environmental Research (BER), as part of BER′s Genomic Science Program 395 (GSP). This contribution originates from the GSP Foundational Scientific Focus Area (FSFA) at the Pacific Northwest National Laboratory (PNNL) under contract 112443. We appreciate support from the Montana Agricultural Experiment Station (project 911352). This study was conducted under Yellowstone National Park research permits YELL-0129 and 5494 (D.M.W.) and YELL-02058 (M.K.), and we appreciate the assistance from National Park Service personnel.

CHAPTER FIVE

COMPARATIVE GENOMIC ANALYSES OF MULTIPLE *SYNECHOCOCCUS* ISOLATES FOUND IN DIFFERENT TEMPERATURE AND LIGHT ENVIRONMENTS OF HOT SPRING MICROBIAL MATS

Introduction

Chapter 4 demonstrated the differences that can be found among genomes of isolates representative of different putative ecotypes (PEs) of A-lineage *Synechococcus* spp. based on gene content comparison. This chapter includes comparative genomic analyses involving detection of both sites and genes under positive selection within an isolate or lineage, in addition to gene content comparisons. Evidence of positive selection of a protein, or a site within a protein, can indicate mutations that have had an advantageous effect on an organism, such as the evidence of selection in uropathogenic *E. coli* relative to non-uropathogenic *E. coli* strains in genes that are associated with urinary tract infection (Chen et al., 2006), or the increase in hydrophobicity of amino acid residues in thermostable proteins (Miller, 2003).

In this chapter, I used comparative genomic analyses to make three different types of genome comparisons. First, I compared isolates from different 16S rRNA sequence types predominant at different temperatures along the Mushroom Spring effluent channel (from high temperature to low temperature: A′, A, and B′; Ferris and Ward, 1997), which may lead to information on genomic changes associated with temperature adaptation.

Populations that have distinct temperature adaptations, such as the 16S rRNA-defined A′-, A-, and B′-lineages, should exhibit even more evidence of genomic difference because they are more distantly related than different PEs within a 16S rRNA-defined lineage. Second, I compared isolates representative of different PEs, especially PEs with different light adaptations (high-light-adapted PE A1 and low-light-adapted PEs A4 and A14; Chapter 4; Olsen et al., submitted), which may further illuminate the ecological distinctions of PEs based on light and other ecological parameters within a restricted temperature range. Third, I compared multiple isolate genomes from members of the same PE (PEs A1 and A14; Becraft et al., 2011; submitted), which will allow me to test the hypothesis of ecological interchangeability of members of a PE. If members of a PE are both ecologically interchangeable and ecologically distinct from members of other PEs (Cohan and Perry, 2007), they should have different adaptations from members of another PE, while maintaining many, if not all, of the same adaptations within a PE.

Methods

Synechococcus Isolates

Synechococcus isolates were obtained as described in Nowack et al. (submitted) and in Chapter 2, with the exception of *Synechococcus* sp. PE B′12 R68DH1S1R03C, which was isolated using medium DH (Allewalt et al., 2006), instead of medium DHAY (Nowack et al., submitted). Isolates used in this study are as follows, with names in parentheses used hereafter in this thesis: *Synechococcus* sp. JA-3-3Ab (PE A1-OS) and

JA-2-3B′a(2-13) (PE B′-OS) (Bhaya et al., 2007); *Synechococcus* sp. 65AY6Li (PE A1-MS), 65AY6A5 (PE A4), and 60AY4M2 (PE A14-60) (Chapter 4; Olsen et al., submitted); and, newly isolated and sequenced *Synechococcus* sp. 63AY4M2 (PE A6-63), M65AY6 (PE A6-65), 63AY4M1 (PE A14-63), and 65AY640 (PE A14-65) (Chapter 3), as well as *Synechococcus* sp. 65AY6A′ (PE A′), 55AY5B5 (PE B′5), and R68DH1S1R03C (PE B′12).

Genome Sequencing, Assembly, and Annotation

Isolate genomes were sequenced and assembled as described in Chapter 4 (Olsen et al., submitted). The genomes of *Synechococcus* spp. PE A6-65 and PE B′5 isolates were further assembled to reference sequences PE A6-63 and PE A1-OS, and PE B'-OS isolate genomes, respectively, using the map to reference program in Geneious version 7.1.7 (http://www.geneious.com, Kearse et al., 2012). Genomes were annotated as described in Chapter 4 (Olsen et al., submitted).

Comparative Analyses

Genomes were aligned using the Mauve progressive aligner (Darling et al., 2004). Alignments of the orthologous genes shared among all the genomes were concatenated and a maximum-likelihood tree was constructed as described in Chapter 3. 1136 orthologous genes were concatenated for a total sequence length of 1,058,325 bp. Any orthologous gene sets are identified herein using the locus tag from the PE A1-OS isolate, which is termed CYA_[gene number]. Percent nucleotide identity was also calculated

using this alignment with the compute pairwise distance feature in MEGA version 6 (Tamura et al., 2013), and using whole-genome reciprocal best hits as described in Chapter 4. GC content, codon use, and amino acid use were computed using Geneious version 7.1.7 (http://www.geneious.com, Kearse et al., 2012). Comparative gene content analyses of genomes were conducted using the RAST SEEDViewer (Overbeek et al., 2013). Positive selection was tested for in all orthologous genes using the M1a and M2a CodeML tests available in PAML: whole gene selection was designated by dN/dS>1 (the rate of non-synonymous mutations, or dN, divided by the rate of synonymous mutations, or dS, greater than one is indicative of positive selection) and site selection was determined using the Bayes Empirical Bayes method ($\omega > 1$; $P > 90\%$) (Yang, 2007).

PCR Amplification of *nifD* and *nifK*

DNA isolated for genome sequencing of the PE A14-60 and PE A14-65 isolates (see Chapter 4, Olsen et al., submitted), as well as DNA from four *Synechococcus* cultures extracted using the FastDNA spin kit (Molecular Biosciences, Boulder, CO) (see Nowack et al., submitted, for protocol), was amplified by PCR using *nifD* and *nifK* primers and thermocycler conditions described in Steunou et al. (2006) to test for the presence of *nif* genes. An aliquot of TE buffer was amplified as a negative control. PCR products were stained with ethidium bromide and bands were separated using agarose gel electrophoresis, before visualization of the DNA using UV illumination.

Results

Genomic Information

Table 5.1 describes the genomic and phenotypic characteristics of the 12 isolates used in this comparative analysis, as well as the environmental distributions of the PEs of the genomes studied, where available. Differences between the *Synechococcus* A'/A-lineage isolates and the B'-lineage isolates are apparent, especially with regards to GC content. A'/A-lineage genomes have a GC content range of 60.2-60.7%, while the B'-lineage genome GC content range is lower at 58.5-58.7%. This difference is reflected in the codon usage bias between the A'/A-lineage and the B'-lineage (Table 5.2), which shows that the A'/A-lineage isolates tend to use more high-GC codons compared to the B'-lineage isolates. For example, alanine has four codons: GCC, GCG, GCA, and GCT. The A′/A-lineage isolates use GCC, on average, 54.34% of the total alanine codons in the isolate genomes, while less frequently using GCG, GCT, and GCA for 19.38%, 16.74%, and 9.53% of the total alanine codons, respectively. The B′-lineage isolates use GCC and GCG for a lower overall percentage of the total alanine codons than the A'/A-lineage isolates (50.80% and 17.37%, respectively), and use GCT and GCA for a higher overall percentage than the A′/A-lineage isolates (20.47% and 11.37%, respectively). The average amino acid usage of the A′/A-lineage and B′-lineage isolates are also reported in Table 5.2.

Table 5.3 describes the pairwise percent nucleotide identity between the genomes. The number of orthologous genes shared between PE A6-65 and the other genomes is

Table 5.1. Genomic, phenotypic, and environmental information for *Synechococcus* isolates representative of putative ecotypes with different depth distributions or from different hot spring mats.

Isolate	A′	A1-OS	A1-MS	A4	A6-63	A6-65	A14-60	A14-63	A14-65	B′-OS	B′5	B′12
PE	A′	A1	A1	A4	A6	A6	A14	A14	A14	B′24	B′5	B′12
Isolation Source	Mushroom Spring	Octopus Spring	Mushroom Spring	Mushroom Spring	Mushroom Spring	Mushroom Spring	Mushroom Spring	Mushroom Spring	Mushroom Spring	Octopus Spring	Mushroom Spring	Mushroom Spring
Isolation Date	September 2010	July 2002	September 2010	September 2010	September 2010	September 2010	September 2010	September 2010	September 2010	July 2002	October 2011	October 2008
Isolation Temp (°C)	65	58-65[1]	65	65	63	65	60	63	65	58-65[1]	55	68[2]
PE Relative Abundance >5% at 60°C Mushroom Spring (MS) (depth in µm)	N/A	0-960	0-960	440-760	440-960	440-960	440-960	440-960	440-960	N/A	N/A	N/A
PE Relative Abundance >10% at 60°C MS (depth in µm)	N/A	0-760	0-760	680	560-960	560-960	600-960	600-960	600-960	N/A	N/A	N/A
PE Light Distribution at 60°C MS[3]	N/A	High light	High light	Low light	Low light	Low light	Low light	Low light	Low light	N/A	N/A	N/A
Sequencing Method	Illumina	Sanger	Illumina	Illumina	Illumina	Illumina	Illumina	Illumina	Illumina	Sanger	Illumina	Illumina
Length (Mbp)	2.74	2.93	2.93	2.98	3.09	3.10	3.16	3.16	3.16	3.05	3.05	2.85
# Contigs	140	1	2	9	3	3[4]	6	5	6	1	1[4]	230
Largest Contig (bp)	102466	2932766	2795989	2508234	3083974	3084139	3142301	3150163	2867041	3046682	3046481	69680
% GC content	60.5	60.2	60.3	60.4	60.5	60.7	60.4	60.5	60.4	58.5	58.6	58.7
CDS	2787	2760	2623	2597	2636	1762[5]	2622	2634	2631	2862	2948	2854
tRNAs	41	47	49	47	45	28	47	41	44	45	45	40
rRNA operons	≥1	2	2[6]	2	≥1	≥1	2	≥1	≥1[6]	2	2	≥0

[1] Temperatures in Octopus Spring fluctuate continuously over a 4.5 minute cycle (Miller et al., 1998), therefore the 7°C range of the isolation site is given.
[2] Temperature taken at site and sample site location differ, sample temperature was most likely ~63°C.
[3] See Becraft et al. (submitted) and Figure 1.7 for PE vertical distributions and Figure 1.8 for vertical gradient light intensity.
[4] Genomes are assembled to a reference and therefore have the same number of contigs as the reference.
[5] Due to poor genome coverage (see Chapter 3).
[6] Duplication of the 23S rRNA locus in one operon.

Table 5.2. Amino acid and codon usage bias between A'/A-lineage isolates and B'-lineage isolates. Values shown are average values in each lineage for percent of total amino acid use (left) and percent codon use per total use for an amino acid (right; larger of the two values for each codon in yellow). Amino acids are color coded: green=hydrophobic; blue=polar, charged; and purple=polar,uncharged. *represents the stop codon.

Amino Acid (AA) Usage			Codon Usage		
AA	Average A'/As	Average B's	Codon and AA	Average A'/As	Average B's
A	10.21%	9.80%	GCA A	9.53%	11.37%
			GCC A	54.34%	50.80%
			GCG A	19.38%	17.37%
			GCT A	16.74%	20.47%
C	1.01%	1.07%	TGC C	66.96%	61.47%
			TGT C	33.04%	38.53%
D	4.30%	4.33%	GAC D	45.59%	38.37%
			GAT D	54.41%	61.63%
E	6.40%	6.23%	GAA E	33.53%	39.43%
			GAG E	66.47%	60.57%
F	3.40%	3.40%	TTC F	41.67%	40.17%
			TTT F	58.33%	59.83%
G	7.77%	7.80%	GGA G	14.00%	16.83%
			GGC G	46.37%	39.93%
			GGG G	32.24%	33.10%
			GGT G	7.41%	10.13%
H	1.90%	2.00%	CAC H	71.94%	64.17%
			CAT H	28.06%	35.83%
I	4.70%	4.80%	ATA I	2.73%	2.87%
			ATC I	66.57%	61.77%
			ATT I	30.71%	35.37%
K	2.81%	2.87%	AAA K	36.88%	42.27%
			AAG K	63.12%	57.73%
L	12.62%	12.43%	CTA L	5.02%	5.03%
			CTC L	17.47%	15.43%
			CTG L	47.14%	44.53%
			CTT L	6.01%	6.40%
			TTA L	1.91%	2.23%
			TTG L	22.44%	26.30%
M	1.63%	1.70%	ATG M	97.63%	96.97%
			CTG M	0.00%	0.00%
			TTG M	2.36%	3.03%
N	2.40%	2.50%	AAC N	76.69%	70.10%
			AAT N	23.31%	29.90%
P	6.30%	6.23%	CCA P	9.33%	11.03%
			CCC P	44.88%	44.37%
			CCG P	28.87%	25.73%
			CCT P	16.89%	18.87%
Q	5.69%	5.77%	CAA Q	34.03%	40.07%
			CAG Q	65.97%	59.93%
R	7.20%	7.07%	AGA R	3.07%	3.47%
			AGG R	5.19%	5.10%
			CGA R	6.86%	9.03%
			CGC R	46.51%	41.93%
			CGG R	31.20%	30.70%
			CGT R	7.18%	9.77%
S	5.50%	5.70%	AGC S	36.04%	32.27%
			AGT S	5.87%	7.80%
			TCA S	3.69%	4.67%
			TCC S	27.29%	27.40%
			TCG S	14.19%	12.97%
			TCT S	12.91%	14.80%
T	4.49%	4.60%	ACA T	7.78%	8.90%
			ACC T	57.94%	57.60%
			ACG T	21.60%	19.47%
			ACT T	12.67%	14.00%
V	6.94%	7.00%	GTA V	7.23%	7.73%
			GTC V	21.07%	18.83%
			GTG V	57.48%	57.73%
			GTT V	14.26%	15.70%
W	1.80%	1.80%	TGG W	100.00%	100.00%
Y	2.60%	2.60%	TAC Y	70.56%	65.50%
			TAT Y	29.44%	34.50%
*	0.30%	0.30%	TAA *	21.34%	21.77%
			TAG *	43.56%	40.57%
			TGA *	35.12%	37.70%

Table 5.3. Pairwise percent nucleotide identity comparisons among isolate genomes. Above the diagonal is the percent nucleotide identity for each pair of genomes calculated from an 1136 orthologous gene concatenation. Below the diagonal is the percent nucleotide identity for each pair of genomes calculated from the whole genome using reciprocal best hits. Bold text indicates within-PE comparisons. PE A6-63/PE A6-65 isolate orthologous gene comparison is are in red. Blue text indicates within-16S rRNA-defined lineage comparisons. Gold text is PE A6-65 whole-genome reciprocal best hit value, which is low due to a high number of ambiguous bases (see Chapter 3).

	PE	**A1**		**A4**	**A6**		**A14**			**A' like**	**B'24**	**B'5**	**B'12**
PE	**Strain**	**A1-OS**	**A1-MS**	**A4**	**A6-63**	**A6-65**	**A14-60**	**A14-63**	**A14-65**	**A'**	**B'-OS**	**B'5**	**B'12**
A1	**A1-OS**	-	**99.32**	98.96	99.02	98.97	98.87	98.94	98.88	98.96	85.28	85.22	85.28
	A1-MS	**99.32**	-	98.92	98.99	98.96	98.84	98.90	98.85	99.01	85.31	85.24	85.31
A4	**A4**	98.46	98.41	-	99.53	99.12	99.50	99.51	99.50	98.88	85.37	85.31	85.37
A6	**A6-63**	98.40	98.33	99.27	-	**99.12**	99.57	99.61	99.57	98.87	85.34	85.28	85.34
	A6-65	92.59	92.68	90.34	**89.48**	-	99.14	99.14	99.14	99.04	85.37	85.32	85.37
A14	**A14-60**	98.35	98.72	99.05	99.13	90.36	-	**99.92**	**99.99**	98.89	85.36	85.30	85.36
	A14-63	98.49	98.39	99.11	99.19	90.47	**99.72**	-	**99.92**	98.90	85.35	85.30	85.36
	A14-65	98.42	98.32	99.11	99.20	90.36	**99.82**	**99.86**	-	98.89	85.35	85.30	85.36
A' like	**A'**	98.62	98.61	98.42	98.36	93.95	98.54	98.49	98.48	-	85.43	85.38	85.45
B'24	**B'-OS**	87.77	87.86	87.59	87.75	89.11	87.66	87.61	87.43	87.92	-	97.78	98.96
B'5	**B'5**	87.88	87.86	87.65	87.77	89.47	87.77	87.73	87.54	88.00	97.80	-	97.70
B'12	**B'12**	87.78	87.86	87.63	87.81	88.70	87.78	87.73	87.56	87.98	98.75	97.62	-

low because of poor coverage and a high number of ambiguous regions in the PE A6-65 genome: for this reason, pairwise percent nucleotide identity was calculated only using genes shared among all 12 genomes, with a requirement of >70% nucleotide identity for genes to be considered as orthologous. It is apparent that inter-PE nucleotide identity is higher than intra-PE nucleotide identity, and that isolates of the same PE share more genes than isolates of different PEs share, with the exception of the PE A6 isolates. This corroborates the whole genome phylogeny (Figure 5.1), which shows that members of PE A14 and PE A1 form distinct clades, while the two PE A6 isolates do not (see Chapter 3).

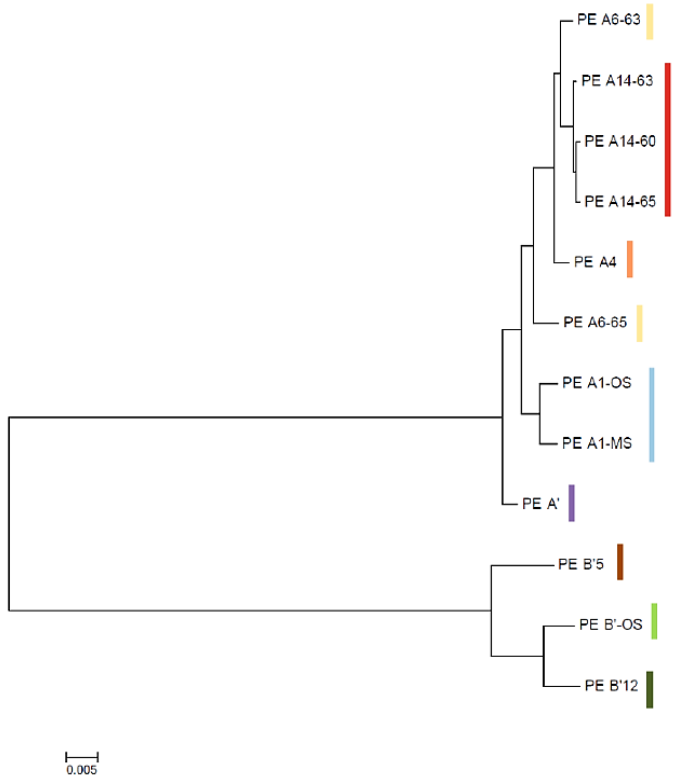


Figure 5.1. Maximum-likelihood tree of *Synechococcus* spp. isolate whole genome alignment using 1136 aligned and concatenated orthologous genes. The scale bar describes the branch lengths, and is equal to the number of substitutions per site in the sequence.

Gene Content Differences

Differences among Genomes of Lineages B′, A, and A′ Defined by 16S rRNA. Several gene content differences are observed among the 12 isolate genomes, summarized in Table 5.4. The B′-lineage isolate genomes contained several genes not found in any of the A'/A-lineage isolate genomes. The PE B'12 and PE B'5 isolate genomes both contained the phosphonate transport and metabolism gene clusters found in PE B'-OS (CYB_0159 to CYB_0168) that are missing in the PE A1-OS isolate (Bhaya et al., 2007), and are not present in any of the A'/A-lineage isolates. All 12 isolates possess orthologs of the putative phosphonatase genes (CYA_1475 and CYA_2058) described by Gomez-Garcia et al. (2011) (with the exception of the CYA_1475 ortholog in A6-65, which may be missing due to the poor coverage of that genome). All three B′-lineage isolates also have cyanophycin synthetase (CYB_0911) and cyanophycinase genes (CYB_2043), which appear to be specific to the B'-lineage. The B'-lineage isolates also have a gene coding for the succinate dehydrogenase iron-sulfur protein subunit, along with the succinate dehydrogenase flavoprotein subunit, which is found in the PE A1 isolate genomes (Chapter 3; Olsen et al., submitted). The PE A' isolate genome has a single lineage-specific gene that codes for an extra monoamine oxidase: PSI-BLAST top hit has 62% amino acid identity to a gene annotated as tyramine oxidase in *Nostoc punctiforme*. No genes are found in all of the A-lineage isolate genomes that are absent in the B′-lineage and PE A′ isolates.

Table 5.4. Gene content differences among *Synechococcus* isolates. Black cell indicates gene presence in the isolate genome, blank cell indicates probable gene absence.

PE	A′	A1		A4	A6		A14			B′24	B′12	B′5
Isolate	**PE A'**	**PE A1-OS**	**PE A1-MS**	**PE A4**	**PE A6-63**	**PE A6-65**	**PE A14-60**	**PE A14-63**	**PE A14-65**	**PE B'-OS**	**PE B'12**	**PE B'5**
Phosphonate metabolism/degradation *phnCDE/phnGHIJKLM*										■	■	■
Cyanophycin synthetase/cyanophycinase										■	■	■
Succinate dehydrogenase iron-sulfur subunit										■	■	■
Polar amino acid transport	■	■	■	■		■				■	■	■
Carbonic anhydrase	■	■	■			■				■	■	■
Cystine ABC transport	■	■	■			■				■	■	■
Succinate dehydrogenase flavoprotein subunit	■	■	■			■				■	■	■
Peptide/opine/nickel transport *pepT*	■	■	■			■				■	■	■
Beta-carotene ketolase					■		■	■	■	■	■	■
Hydroxymethyl pyrimidine ABC transport	■				■		■	■	■	■	■	■
Ferrous Iron Transport *feoAB*	■			■	■	■	■	■	■			
Nitrate reductase paralog	■			■	■	■	■	■	■			
Urea carboxylase			■	■	■		■	■	■			
Ammonium transport duplication *amtB2*	■			■	■		■	■	■			
Predicted photosynthetic antenna genes *isiX/apcA3/apcD4*				■	■		■	■	■			
Sugar (possibly maltose/maltodextrin) transport				■	■		■	■	■			
Formamidase/amidase cluster	■						■	■	■			
Spermidine/putrescine transport *potABCD*							■	■	■			
Type III CRISPR/cas array		■	■			■						
Urease	■	■	■			■						
Bipolar DNA helicase/Type I restriction modification subunit M/single-stranded endonuclease	■		■									
Monoamine oxidase (orthologous copy)	■											
Nitrogen fixation cassette *nif*	■	■	■	■	■	■	■	■		■	■	■

Differences among Isolates Representative of Different PEs in the B′- and A-Lineages. The genomes of B'- lineage isolates PE B'-OS, PE B'12, and PE B'5, did not have any notable content differences, although the reference-based assembly of the PE B'5 genome would not have included genes unique to that genome. Isolates from the B'-lineage also share several A-lineage, isolate-specific genes that were reported in initial genome comparison studies in Chapter 4 (Olsen et al., submitted). The B' isolate genomes have genes encoding carbonic anhydrase, two subunits of a cystine ABC transporter (encoding the periplasmic cystine binding protein and the permease protein), succinate dehydrogenase flavoprotein subunit SdhA, omega-amino acid-pyruvate aminotransferase cluster, and PepT-like peptide/opine/nickel transport cassette that were previously found only in PE A1 isolates (Chapter 4; Olsen et al., submitted). B'-lineage isolates also have the genes encoding beta-carotene ketolase and a hydroxymethylpyrimidine ABC transport system. These hydroxymethylpyrimidine transport genes are also found in the three PE A14 isolates, the PE A' isolate, and the PE A6-63 isolate.

The PE A' isolate also shares genes described in Chapter 4 (Olsen et al., submitted) that are not shared by every isolate in the A-lineage. These include (i) the ferrous iron transport *feoAB* genes, which are also found in the three PE A14 isolates, the two PE A6 isolates, and the PE A4 isolate; (ii) the cystine ABC transport genes, succinate dehydrogenase flavoprotein subunit, omega-amino acid-pyruvate aminotransferase, and peptide/opine/nickel *pepT* ABC transport cassette, all found in the two PE A1 isolates

and the three B'-lineage isolates; (iii) a polar amino acid ABC transport cassette found in the PE A1 isolates, the PE A4 isolate, the PE A6-65 isolate, and also the B'-lineage isolates; (iv) a six gene cluster that includes putative amidase and formamidase (*amiF*) genes, and is shared by the three PE A14 isolates; (v) the urease 1 genes found in PE A1-OS (Bhaya et al., 2007) and subsequently in PE A1-MS (Chapter 4; Olsen et al., submitted); and (vi) a three gene cluster found in PE A1-MS that contains genes annotated as a bipolar DNA helicase, a Type I restriction-modification system DNA-methyltransferase subunit M, and the single-stranded exonuclease associated with Rad50/Mre11 complex. The PE A' isolate also shares the ammonium transport gene duplication *amtB1/2* with the three PE A14, PE A4, and PE A6-63 isolates, and a nitrate reductase gene duplicated in all PE A14, both PE A6, and the PE A4 isolates.

The A-lineage isolates have several gene content differences, many of which were identified in Chapter 4 (Olsen et al., submitted), or are shared with either the B′-lineage or PE A′ isolates, as described above. The two PE A6 isolates, A6-65 and A6-63, differ with regard to gene content, which matches the orthologous-gene genome phylogeny and the conclusion reached in Chapter 4, that the two isolates are not replicate isolates of the same ecotype. The PE A6-65 isolate shares several genes with the PE A1 and PE A' isolates, including urease 1, carbonic anhydrase, cystine ABC transport genes, succinate dehydrogenase flavoprotein subunit, omega-amino acid-pyruvate aminotransferase, and peptide/opine/nickel *pepT* ABC transport cassette, as well as a partial (2 gene) match to the Type III CRISPR/cas array found in the A1-OS and A1-MS isolates (Heidelberg et

al., 2009; Olsen et al., submitted). The PE A6-65 isolate shares some of the low-light-isolate specific genes as well, such as those coding for *feoAB* and the paralogous nitrate reductase gene, which are also found in the PE A6-63 isolate genome. The PE A6-63 isolate is closer to the other A-lineage low-light-adapted strains in gene content: it shares the *apcA*/*apcD*/*isiX* gene cassette hypothesized to be associated with low-light adaptation and a sugar (possibly maltose/maltodextrin) ABC transport cluster with the three PE A14 and PE A4 isolates, as well as the beta-carotene ketolase genes found in the three PE A14 isolates and in the B'-lineage isolates, and a urea-carboxylase gene cluster, which is found in PE A1-MS, the three PE A14 isolates, and the PE A4 isolates.

Differences among Replicate Isolates Representative of PEs A1 and A14. The PE with the most replicate isolates, PE A14, has the most similar inter-PE gene content. All three isolates share PE A14 genes listed previously, as well as a PE-specific gene cluster *potABCD*, which codes for a spermidine/putrescine transport system (Chapter 4; Olsen et al., submitted). One notable difference among the PE A14 genomes is the absence of the *nif* gene cluster in the PE A14-65 isolate. The *nif* gene cluster is highly conserved, not only among the other genomes representative of PE A14, but among the genomes of all other isolates. This gene absence has been further evaluated by PCR amplification of the *nifD* and *nifK* genes, which were successfully amplified from the PE A14-60 and PE A14-63 isolates, but not from the PE A14-65 isolate. The PE A1 isolates do not have any PE-specific genes, and have two gene-content differences (PE A1-MS urea carboxylase cassette and the bipolar DNA helicase cluster described above).

Sites under Positive Selection

Out of 1136 orthologous genes in the set of 12 genomes, CodeML analysis detected 297 site-specific mutations with a signal of positive selection in a total of 148 genes (Table 5.5, below diagonal). The genes with sites under selection have a variety of annotated functions, including cell signaling and motility, transport of metabolites into the cell, biosynthesis of various coenzyme and cofactor, DNA binding and repair, peptidoglycan metabolism, glycogen and glucose metabolism, proteolysis, stress response, transcription and translation, and even photosynthesis, that could impact the fitness and environmental niche of an organism if the protein function were altered (Tables 5.6-5.10 below; Supplementary Table B.1, found in Appendix B).

Differences among Genomes of Lineages B′, A, and A′ Defined by 16S rRNA. Several patterns of lineage-specific site mutations were apparent in these analyses (Figure 5.2), two of which were associated with divergences among the three 16S rRNA lineages. One pattern, termed the A',A//B' mutation pattern (Figure 5.2A), involved 36 sites in 21 genes (Table 5.6), which indicate selection in either the B′-lineage or the A′/A-lineage. Another pattern, termed the A//A′,B′ pattern (Figure 5.2B; Table 5.7) , involved 7 sites in 4 genes that showed selection only within the A-lineage.

Differences among Isolates Representative of Different PEs in the B′- and A-Lineages. The most highly repeated pattern of selection involved a mutation resulting in a site with one amino acid shared by the A'/A-lineage isolate, another amino acid in the PE

B'-OS and PE B'12, and a third amino acid in PE B'5, thus termed the A',A//B'-OS,B'12//B'5 pattern (Figure 5.2C). This pattern was found in 33 genes, with a total of 62 sites (Table 5.8). There were also 19 genes (25 sites) that showed a fourth pattern (Table 5.9), termed the A',A,B'5//B'-OS,B'12 mutation pattern (Figure 5.2D).

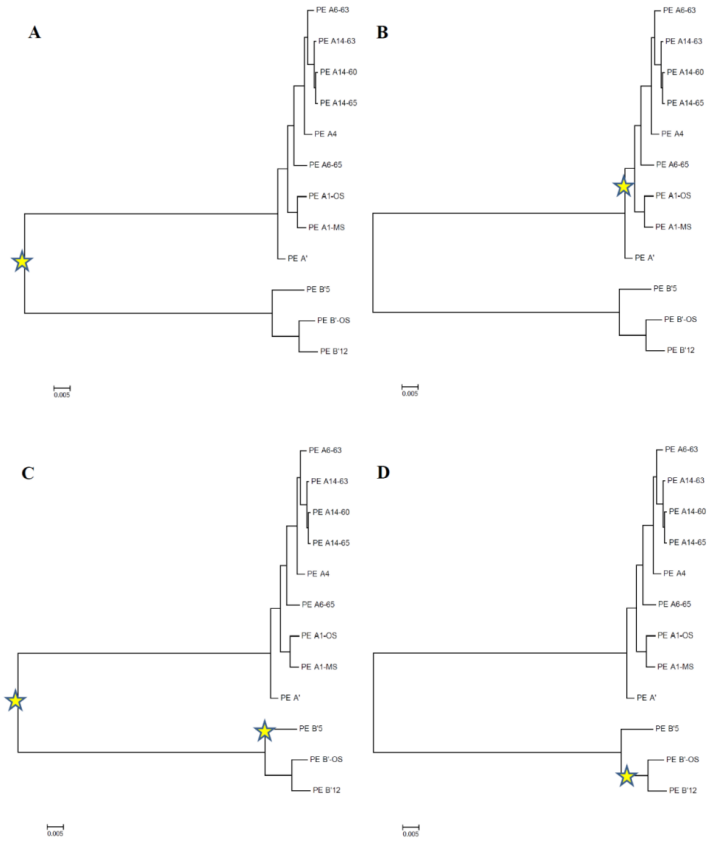


Figure 5.2. B′-lineage site selection mutation patterns. Stars on the tree represent hypothesized mutation events in the evolutionary history of the A/B′ lineage, with the A) A',A//B' pattern, and B) A//A′,B′ pattern, C) A',A//B'-OS,B'12//B'5 pattern, and D) A',A,B'5//B'-OS,B'12 pattern.

Some genes have multiple sites with different mutation patterns, like dephospho-CoA kinase (CYA_0565), which is involved in Coenzyme A biosynthesis, and has 11

positively selected sites: 6 sites with the A',A//B'-OS,B'12//B'5 pattern (Figure 5.2C), three sites with an A',A,B'5//B'-OS,B'12 pattern (Figure 5.2D), and two with an A',A//B' pattern (Figure 5.2B). The patterns of selection in this gene may indicate a selection between the A′/A- and B′-lineages, with further selection within the B′-lineage.

There were also site mutation patterns within the A'/A-lineage isolate genome sequences (Table 5.5, below diagonal), as well as intra- and inter-PE mutation patterns (Table 5.10). All of the A-lineage isolates shared the same amino acid residue for 179 of the positively selected sites, and only 9 sites (in 5 genes) of those 179 sites did not have the amino acid residue conserved between the PE A' sequence and the A-lineage sequences (pattern in Figure 5.2B). The low-light-adapted PE isolates (PE A4, A6-63, and three PE A14s) shared the same site mutation 204 out of the total 297 selected sites, but there was some variation between the low-light-adapted PE isolates. PE A4 and PE A6-63 isolates had different amino acids at 47 positively selected sites, PE A14 and PE A4 isolates had different amino acids at 62 positively selected sites, and PE A14 and PE A6-63 isolates had different amino acids at 64 positively selected sites. The PE A6 isolates shared 232, which was less than the 253 sites shared between PE B′-OS and PE B′12 isolates.

Differences among Replicate Isolates Representative of PEs A1 and A14. PE A1 isolates shared 291 of the 297 total site mutation patterns, while the three PE A14 isolates shared 265 site mutations (Table 5.5, below diagonal). PE A14-63 had 32 mutations that differentiated it from PE A14-60 and PE A14-65 (which only differed in 2 selected sites

Table 5.5. Pairwise comparisons among isolate genomes. Below the diagonal is the number of positively selected sites with different mutations in each pair of genomes. Above the diagonal displays the number of genes under selection (dN/dS > 1) between each pair of genomes. Bold text indicates within-PE comparisons. PE A6-63/PE A6-65 isolate comparisons are in red. Blue text indicates within-16S rRNA-defined lineage comparisons.

	PE	A1		A4	A6		A14			A' like	B'24	B'5	B'12
PE	Strain	A1-OS	A1-MS	A4	A6-63	A6-65	A14-60	A14-63	A14-65	A'	B'-OS	B'5	B'12
A1	A1-OS	-	**0**	2	1	2	0	0	0	3	0	0	0
	A1-MS	**6**	-	2	0	2	0	0	0	3	0	0	0
A4	A4	65	64	-	2	1	2	2	2	3	0	0	0
A6	A6-63	47	46	47	-	**3**	3	2	2	3	1	0	0
	A6-65	94	89	55	**65**	-	3	2	4	1	0	0	0
A14	A14-60	108	103	44	64	41	-	**0**	**0**	3	0	0	0
	A14-63	76	72	40	37	48	**32**	-	**0**	3	0	0	0
	A14-65	106	101	44	62	43	**2**	**32**	-	4	0	0	0
A' like	A'	83	80	74	83	37	60	65	62	-	0	0	0
B'24	B'-OS	251	253	258	273	265	256	261	258	248	-	5	3
B'5	B'5	228	228	243	248	242	233	244	235	223	128	-	6
B'12	B'12	250	252	268	276	268	255	269	257	253	44	131	-

Table 5.6. Sites under positive selection with the A',A//B' mutation pattern (Figure 5.2A). The site under selection and the probability it is under selection are given in, as well as the mutation pattern (right-most column). Mutation patterns highlighted in red do not fit the A',A//B' pattern.

Gene	Annotated Function	Gene Subsystem	Alignment Length (nt)	Positively selected sites (*: P>95%; **: P>99%)[1]			A',A//B's (Other Pattern)[2]
CYA_0016	gas vesicle protein W	Cell motility/Buoyancy	747	**7 P**	**0.902**	**8.010 +- 2.888**	**P (As and A')-> G (B's)**
				214 M	**0.960***	**8.470 +- 2.318**	**M(A1s/A6-63)->T(A14s/A4/A6-65/A')/A(B's)**
CYA_0109	Cell division protein FtsZ (EC 3.4.24.-)	Bacterial Cytoskeleton	1119	**326 V**	**0.944**	**4.094 +- 2.469**	**V(As and A')->A(B's)**
CYA_0164	2-octaprenyl-6-methoxyphenol hydroxylase (EC 1.14.13.-)	Ubiquinone Biosynthesis	1242	**136 H**	**0.984***	**9.975 +- 1.341**	**H(A1s, A14-63,A6-63)->Y(REST)**
				165 A	**0.967***	**9.813 +- 1.807**	**A(As and A')-> R(B's)**
				385 R	**0.901**	**9.205 +- 2.859**	**R(As and A')-> A(B's)**
				406 -	**1.000****	**10.122 +- 0.678**	**STOP CODON IN B's**
CYA_0182	Phosphate transport system regulatory protein PhoU	High affinity phosphate transporter and control of PHO regulon; Phosphate metabolism	639	**118 G**	**0.962***	**6.789 +- 2.364**	**G(A1s,B's,A6-63)->S(REST)**
				119 R	**0.963***	**6.795 +- 2.358**	**R(A1s,B's,A6-63)->C(REST)**
				174 K	**0.950**	**6.719 +- 2.456**	**K(As and A') -> R(B's)**
CYA_0291	23S rRNA (guanosine-2'-O-) - methyltransferase rlmB (EC 2.1.1.-) ## LSU rRNA Gm2251	RNA methylation	1062	**50 -**	**1.000****	**10.431 +- 0.266**	**STOP CODON B's**
				67 -	**1.000****	**10.431 +- 0.266**	**STOP CODON B's**
				341 -	**1.000****	**10.431 +- 0.266**	**STOP CODON B's**
CYA_0331	Plastid and cyanobacterial ribosome-associated protein PSRP-3	Translation	297	**62 T**	**0.956***	**9.670 +- 1.996**	**T(As and A')->E(B's)**
				89 -	**1.000****	**10.067 +- 0.732**	**STOP CODON B'12**
CYA_0336	Chorismate synthase (EC 4.2.3.5) # AroG	Chorismate Synthesis; Common Pathway For Synthesis of Aromatic Compounds (DAHP synthase to chorismate)	1134	**97 S**	**1.000****	**8.239 +- 1.881**	**S(As and A')->Q(B's)**

[1] Amino acid sites indicated to be under positive selection by the BEB method are listed, along with the probability (P) that the sites are in a positive selection site class and the post mean +/- SE for positive selection.

[2] Amino acids (and the genomes of the isolates and lineages in which they occur) are listed. (REST) indicates the amino acid is present in the remainder of the isolate genomes not listed in the entry. STOP CODON indicates a stop codon insertion in the isolate genomes listed.

Table 5.6 (continued)

				189 A	0.967*	8.021 +- 2.247	A(As and A')->Q(B's)
				218 N	0.940	7.822 +- 2.507	N(As and A')->R(B's)
CYA_0406	pyrimidine regulatory protein PyrR	Nucleotide synthesis	546	177 E	0.985*	8.330 +- 2.122	E(As and A')->I(B's)
				179 S	0.976*	8.265 +- 2.224	S(As and A')->P(B'-OS, B'12) (no seq B'5)
CYA_0416	hyp		393	3 F	1.000**	10.394 +- 0.361	F(A1s, A4, A6-63)->L(A14s, A6-65, A')/H(B's)
				16 L	0.960*	10.020 +- 1.882	L(As and A')->N(B's)
				81 -	1.000**	10.397 +- 0.329	STOP CODON B's
				86 -	1.000**	10.397 +- 0.329	STOP CODON B'-OS and B'12
CYA_0423	Rrf2 family transcriptional regulator	Transcription	450	94 Q	0.969*	6.057 +- 2.344	Q(As and A')->T(B'-OS and B'12)/S(B'5)
				95 Q	0.915	5.747 +- 2.549	Q(As and A')->A(B's)
CYA_0552	nucleic acid binding, OB-fold, tRNA/helicase-type	DNA binding; DNA repair	366	11 G	0.989*	6.707 +- 2.307	G(As, A', B'5)->K(B'-OS and B'12)
				20 A	0.956*	6.539 +- 2.484	A(As and A')->L(B's)
CYA_0565	Dephospho-CoA kinase (EC 2.7.1.24)	Coenzyme A Biosynthesis	690	111 A	0.959*	9.622 +- 1.943	A(As and A')->E(B'-OS and B'12)/L(B'5)
				113 E	0.965*	9.669 +- 1.844	E(As, A', B'5)->A(B'-OS and B'12)
				114 A	0.994**	9.939 +- 1.046	A(As and A')->S(B'-OS and B'12)/E(B'5)
				116 P	1.000**	9.985 +- 0.823	P(As, A', B'5)->K(B'-OS and B'12)
				118 C	0.953*	9.566 +- 2.057	C(As and A')->S(B's)
				119 Q	0.974*	9.751 +- 1.643	Q(As, A', B'5)->G(B'-OS and B'12)
				120 C	0.994**	9.937 +- 1.050	C(As and A')->F(B'-OS and B'12)/K(B'5)
				121 P	0.925	9.311 +- 2.496	P(As and A')->S(B'-OS and B'12)/N(B'5)
				122 H	0.990*	9.899 +- 1.201	H(As and A')->S(B's)
				127 P	0.996**	9.952 +- 0.985	P(As and A')->Q(B'-OS and B'12)/S(B'5)
				213 G	0.992**	9.917 +- 1.131	G(As and A')->Q(B'-OS and B'12)/E(B'5)
CYA_0569	hyp		1824	28 -	1.000**	9.347 +- 1.332	STOP CODON B's
CYA_1778	Transcription antitermination protein NusG	LSU ribosomal proteins cluster; Transcription factors bacterial	669	3 -	1.000**	9.754 +- 1.033	STOP CODON B's
CYA_1892	two-component response regulator		753	238 Q	0.975*	7.867 +- 2.387	Q(As and A')->V(B's)
				250 P	0.935	7.603 +- 2.710	P(As and A')->K(B's)
CYA_1904	CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase (EC 2.7.8.5)	Glycerolipid and Glycerophospholipid Metabolism in Bacteria	558	178 T	0.959*	8.331 +- 2.338	T(As and A')->H(B's)
				181 A	0.959*	8.327 +- 2.344	A(As and A')->Q(B's)
CYA_1988	UDP-N-acetylglucosamine 1-carboxyvinyltransferase (EC 2.5.1.7)	Peptidoglycan Biosynthesis; UDP-N-acetylmuramate from Fructose-6-phosphate Biosynthesis	1443	91 S	0.901	8.459 +- 2.799	T(As)->A(A' and B's)
				461 A	1.000**	9.278 +- 1.347	A(As, A', B'5)-> Q(B'-OS and B'12)

Table 5.6 (continued)

				462 T	0.995**	9.240 +- 1.460	T(As and A')-> L(B'-OS and B'12)/V(B'5)
				463 P	0.999**	9.272 +- 1.367	P(As and A')-> G(B'-OS and B'12)/S(B'5)
				464 Q	0.982*	9.138 +- 1.723	Q(As and A')-> K(B'-OS and B'12)/P(B'5)
CYA_2269	Peptidase family M23/M37	Proteolysis	990	68 A	0.958*	2.948 +- 0.794	A(As and A')->P(B's)
				71 L	0.904	2.815 +- 0.874	L(As and A')->E(B's)
				81 S	0.961*	2.954 +- 0.788	S(As and A')->T(B'-OS)/P(B'5 and B'12)
				85 P	0.934	2.888 +- 0.834	P(As and A')->M(B'-OS and B'5)/B(B'12)
				99 A	0.935	2.894 +- 0.833	A(As and A')->R(B'-OS)/L(B'5 and B'12)
				144 P	0.981*	3.000 +- 0.745	P(As and A')->G(B'-OS and B'12)/Q(B'5)
CYA_2292	Acetyltransferase (GNAT) family, Syn7942_2240 homolog	Transfer of CoA from donor to receiver	585	3 P	0.991**	7.972 +- 2.035	P(As and A')->G(B'-OS and B'12)/C(B'5)
				9 L	0.986*	7.938 +- 2.091	L(As and A')->Q(B'-OS and B'12)/R(B'5)
				13 F	0.913	7.393 +- 2.711	F(As and A')->Q(B's)
CYA_2447	bacterial pre-peptidase C-terminal domain protein	Transcription regulator	900	4 K	0.933	7.854 +- 2.592	K(As and A')->S(B's)
				8 L	0.920	7.752 +- 2.700	L(As and A')->N(B's)
				20 V	0.917	7.732 +- 2.722	V(As and A')->Q(B's)
				30 D	0.914	7.709 +- 2.747	D(As and A')->P(B's)
				188 R	0.911	7.687 +- 2.775	R(As and A')->I(B's)
				275 A	0.977*	8.204 +- 2.158	A(A1s, A14-63/65, A6-63)->T(REST)
CYA_2803	hyp		738	49 C	0.922	8.240 +- 2.647	C(As and A')->V(B's)
				71 P	0.999**	8.826 +- 1.679	P(A1s, A14-63, A6-63)->H(B's)/T(REST)

Table 5.7. Sites under positive selection with the A//A',B' mutation pattern (Figure 5.2B). The site under selection and the probability it is under selection are given in, as well as the mutation pattern (right-most column). Mutation patterns highlighted in red do not fit the A//A',B' pattern.

Gen	Annotated Function	Gene Subsystem	Alignment Length (nt)	Positively selected sites (*: P>95%; **: P>99%)[1]			A//A'/B's (Other Pattern)[2]
CYA_0005	short-chain dehydrogenase/reductase	Fatty acid biosynthesis	873	274 Q	1.000**	8.890 +- 1.624	Q(As) -> S(A' and B's)
				284 A	0.989*	8.815 +- 1.801	A(As) -> G(A' and B's)/R(B'5)

[1] Amino acid sites indicated to be under positive selection by the BEB method are listed, along with the probability (P) that the sites are in a positive selection site class and the post mean +/- SE for positive selection.

[2] Amino acids (and the genomes of the isolates and lineages in which they occur) are listed. (REST) indicates the amino acid is present in the remainder of the isolate genomes not listed in the entry. STOP CODON indicates a stop codon insertion in the isolate genomes listed.

Table 5.7 (continued)

				289 K	0.999**	8.887 +- 1.630	K(As) -> G (A' and B's)
CYA_0319	ABC-2 transporter, permease protein	ABC transport	813	4 Q	0.944	6.518 +- 2.471	Q(As)->R(A' and B's)
				7 H	0.993**	6.778 +- 2.184	H(As)->T(A' and B's)
CYA_1988	UDP-N-acetylglucosamine 1-carboxyvinyltransferase (EC 2.5.1.7)	Peptidoglycan Biosynthesis; UDP-N-acetylmuramate from F-6-P Biosynthesis	1443	91 S	0.901	8.459 +- 2.799	T(As)->A(A' and B's)
				461 A	1.000**	9.278 +- 1.347	A(As, A', B'5)-> Q(B'-OS and B'12)
				462 T	0.995**	9.240 +- 1.460	T(As and A')-> L(B'-OS and B'12)/V(B'5)
				463 P	0.999**	9.272 +- 1.367	P(As and A')-> G(B'-OS and B'12)/S(B'5)
				464 Q	0.982*	9.138 +- 1.723	Q(As and A')-> K(B'-OS and B'12)/P(B'5)
CYA_2229	peptidase, S1C family(EC:3.4.21.-)	Proteolysis	1260	23 F	0.993**	9.417 +- 1.451	F(As)->I(A' and B's)
				24 V	0.995**	9.431 +- 1.410	V(rest)->I(A6-65))/T(A' and B's)
				25 L	1.000**	9.471 +- 1.284	F(As)->I(A' and B's)

Table 5.8. Sites under positive selection with the A',A//B'-OS,B'12//B'5 mutation pattern (Figure 5.2C). The site under selection and the probability it is under selection are given in, as well as the mutation pattern (right-most column). Mutation patterns highlighted in red do not fit the A',A//B'-OS,B'12//B'5 pattern.

Gene	Annotated Function	Gene Subsystem	Alignment Length (nt)	Positively selected sites (*: P>95%; **: P>99%)[1]			(A',A//B'-OS,B'12//B'5) (Other Pattern)[2]
CYA_0014	hyp		462	152 L	0.912	5.077 +- 2.961	L(As and A') -> A(B'-OS and B'12)/V(B'5)
CYA_0026	O-antigen polymerase family protein		723	4 R	0.914	6.871 +- 3.005	R(As and A')->T(B'-OS and B'12)/E(B'5)
CYA_0051	peptide transport system permease protein	ABC transport	1101	111 W	0.989*	8.250 +- 2.069	W(As, A', B'5) -> S(B'-OS and B'12)
				123 T	0.980*	8.191 +- 2.167	T(As and A')->I(B'-OS and B'12)/A(B'5)
				124 E	0.966*	8.095 +- 2.310	E(As, A', B'5)->G(B'-OS and B'12)
CYA_0111	hyp		468	96 R	0.934	6.355 +- 2.949	R(As and A')->F(B'-OS and B'12)/R(B'5)
CYA_0233	Fibronectin/fibrinogen-binding protein	Biofilm formation	1806	228 S	0.998**	9.497 +- 1.192	S(A1s, A14-63, A6s, A', B'5)->T(REST)
				590 -	1.000**	9.513 +- 1.133	G(As, A')/A(B'-OS,B'12)/STOP CODON B'5
				599 E	0.993**	9.457 +- 1.323	E(As and A')->F(B'-OS,B'12)/STOP B'5
				600 L	0.916	8.797 +- 2.619	L(As and A')->F(B'-OS,B'12)/STOP B'5
CYA_0323	truncated cyclic nucleotide-binding protein		513	4 S	0.913	5.880 +- 2.736	S(As and A')->P(B'-OS and B'12)/T(B'5)
				139 V	0.970*	6.213 +- 2.530	V(A1s, A14-63, A4, A6-63)->L(rest)

[1] Amino acid sites indicated to be under positive selection by the BEB method are listed, along with the probability (P) that the sites are in a positive selection site class and the post mean +/- SE for positive selection.

[2] Amino acids (and the genomes of the isolates and lineages in which they occur) are listed. (REST) indicates the amino acid is present in the remainder of the isolate genomes not listed in the entry. STOP CODON indicates a stop codon insertion in the isolate genomes listed.

Table 5.8 (continued)

CYA_0423	Rrf2 family transcriptional regulator	Transcription	450	**94 Q**	**0.969***	**6.057 +- 2.344**	**Q(As and A')->T(B'-OS and B'12)/S(B'5)**
				95 Q	**0.915**	**5.747 +- 2.549**	**Q(As and A')->A(B's)**
CYA_0454	hyp		513	**138 A**	**0.996****	**8.272 +- 2.011**	**A(As and A')->R(B'-OS and B'12)/V(B'5)**
CYA_0565	Dephospho-CoA kinase (EC 2.7.1.24)	Coenzyme A Biosynthesis	690	**111 A**	**0.959***	**9.622 +- 1.943**	**A(As and A')->E(B'-OS and B'12)/L(B'5)**
				113 E	**0.965***	**9.669 +- 1.844**	**E(As, A', B'5)->A(B'-OS and B'12)**
				114 A	**0.994****	**9.939 +- 1.046**	**A(As and A')->S(B'-OS and B'12)/E(B'5)**
				116 P	**1.000****	**9.985 +- 0.823**	**P(As, A', B'5)->K(B'-OS and B'12)**
				118 C	**0.953***	**9.566 +- 2.057**	**C(As and A')->S(B's)**
				119 Q	**0.974***	**9.751 +- 1.643**	**Q(As, A', B'5)->G(B'-OS and B'12)**
				120 C	**0.994****	**9.937 +- 1.050**	**C(As and A')->F(B'-OS and B'12)/K(B'5)**
				121 P	**0.925**	**9.311 +- 2.496**	**P(As and A')->S(B'-OS and B'12)/N(B'5)**
				122 H	**0.990***	**9.899 +- 1.201**	**H(As and A')->S(B's)**
				127 P	**0.996****	**9.952 +- 0.985**	**P(As and A')->Q(B'-OS and B'12)/S(B'5)**
				213 G	**0.992****	**9.917 +- 1.131**	**G(As and A')->Q(B'-OS and B'12)/E(B'5)**
CYA_0576	hyp		909	**72 E**	**0.946**	**6.229 +- 2.660**	**E(A1s, A14-63, A6-63)->T(B's)/A(rest)**
				195 E	**0.957***	**6.269 +- 2.616**	**E(As and A')->R(B'-OS and B'12)/Q(B'5)**
CYA_0635	hyp		180	**1 P**	**0.945**	**7.849 +- 2.626**	**P(As, A', B'5)->Q(B'-OS and B'12)**
				3 V	**0.966***	**8.003 +- 2.441**	**V(As and A')->G(B'-OS and B'12)/L(B'5)**
CYA_0636	Sensory subunit of low CO_2-induced protein complex, putative		498	**151 E**	**0.952***	**5.748 +- 2.259**	**E(As and A')->R(B'-OS and B'12)/D(B'5)**
CYA_1860	hyp		795	**260 E**	**0.993****	**7.470 +- 2.355**	**E(As and A')->P(B'-OS and B'12)/D(B'5)**
				263 E	**0.983***	**7.401 +- 2.435**	**E(As and A')->I(B'-OS and B'12)/A(B'5)**
CYA_1988	UDP-N-acetylglucosamine 1-carboxyvinyltransferase (EC 2.5.1.7)	Peptidoglycan Biosynthesis; UDP-N-acetylmuramate from Fructose-6-phosphate Biosynthesis	1443	**91 S**	**0.901**	**8.459 +- 2.799**	**T(As)->A(A' and B's)**
				461 A	**1.000****	**9.278 +- 1.347**	**A(As, A', B'5)-> Q(B'-OS and B'12)**
				462 T	**0.995****	**9.240 +- 1.460**	**T(As and A')-> L(B'-OS and B'12)/V(B'5)**
				463 P	**0.999****	**9.272 +- 1.367**	**P(As and A')-> G(B'-OS and B'12)/S(B'5)**
				464 Q	**0.982***	**9.138 +- 1.723**	**Q(As and A')-> K(B'-OS and B'12)/P(B'5)**
CYA_2022	DNA recombination and repair protein RecF	DNA repair, bacterial RecFOR pathway; DNA replication cluster 1	1134	**134 T**	**0.961***	**5.970 +- 2.394**	**T(As)->H(A' and B'5)/N(B'-OS and B'12)**
				373 S	**0.947**	**5.878 +- 2.446**	**S(A's and A')->G(B'-OS and B'12)/W(B'5)**
CYA_2023	hyp		690	**86 A**	**0.965***	**9.031 +- 2.064**	**A(A1s, A4, A6-63)->E(REST)**
				93 L	**0.964***	**9.009 +- 2.095**	**L(As and A')->H(B'-OS and B'12)/R(B'5)**
				182 Q	**0.982***	**9.156 +- 1.791**	**Q(As and A')->N(B'-OS and B'12)/E(B'5)**
CYA_2078	hyp		1506	**260 K**	**0.988***	**8.375 +- 2.094**	**K(As and A') -> P(B'-OS and B'12)/E(B'5)**
				422 P	**0.998****	**8.443 +- 1.971**	**P(As and A') -> E(B'-OS and B'12)/L(B'5)**
				423 W	**0.928**	**7.941 +- 2.689**	**W(As, A', B'5)->S(B'-OS and B'12)**

Table 5.8 (continued)

				424 T	0.981*	8.321 +- 2.181	T(As and A') -> L(B'-OS and B'12)/G(B'5)
CYA_2104	Acetyl-coenzyme A carboxyl transferase beta chain (EC 6.4.1.2)	Fatty acid biosynthesis	942	123 L	0.907	7.974 +- 2.850	L(As, A', B'-OS, B'12)->A(B'5)
				303 L	0.993**	8.608 +- 1.932	L(As and A') -> I(B'-OS and B'12)/Q(B'5)
				304 E	0.977*	8.490 +- 2.142	E(As and A') -> D(B'-OS and B'12)/R(B'5)
CYA_2128	hyp		1059	200 E	0.997**	8.801 +- 1.676	E(As, A', B'-OS)-> T(B'5 and B'12)
				288 G	0.925	8.240 +- 2.592	G(As, A')-> I(B'-OS and B'5)/T(B'12)
				351 A	0.999**	8.816 +- 1.641	A(As, A')->Q(B'-OS and B'12)/I(B'5)
CYA_2172	peptidase M16B family, nonpeptidase-like protein	Proteolysis	1311	344 G	0.937	6.651 +- 2.615	G(As, A')-> L(B'-OS, B'12)/P(B'5)
				410 N	0.974*	6.844 +- 2.417	N(As, A', B'5)->P(B'-OS and B'12)
CYA_2184	sensor histidine kinase(EC:2.7.3.-)		1332	307 A	0.911	6.080 +- 2.987	A(As, A')-> G(B'-OS, B'12)/T(B'5)
				309 G	0.948	6.294 +- 2.858	G(As, A')-> K(B'-OS, B'12)/E(B'5)
CYA_2206	Pyruvate kinase (EC 2.7.1.40)	Fermentations: Mixed acid; Glycolysis and Gluconeogenesis; Pyruvate metabolism I: anaplerotic reactions, PEP	1839	374 S	0.938	3.497 +- 1.973	S(As and A')->G(B'-OS and B'12)/M(B'5)
CYA_2269	Peptidase family M23/M37	Proteolysis	990	68 A	0.958*	2.948 +- 0.794	A(As and A')->P(B's)
				71 L	0.904	2.815 +- 0.874	L(As and A')->E(B's)
				81 S	0.961*	2.954 +- 0.788	S(As and A')->T(B'-OS)/P(B'5 and B'12)
				85 P	0.934	2.888 +- 0.834	P(As and A')->M(B'-OS and B'5)/B(B'12)
				99 A	0.935	2.894 +- 0.833	A(As and A')->R(B'-OS)/L(B'5 and B'12)
				144 P	0.981*	3.000 +- 0.745	P(As and A')->G(B'-OS and B'12)/Q(B'5)
CYA_2292	Acetyltransferase (GNAT) family, Syn7942_2240 homolog	Transfer of CoA from donor to receiver	585	3 P	0.991**	7.972 +- 2.035	P(As and A')->G(B'-OS and B'12)/C(B'5)
				9 L	0.986*	7.938 +- 2.091	L(As and A')->Q(B'-OS and B'12)/R(B'5)
				13 F	0.913	7.393 +- 2.711	F(As and A')->Q(B's)
CYA_2361	NAD-dependent epimerase/dehydratase family protein		954	264 I	0.944	7.256 +- 2.669	I(As and A')->L(B'-OS and B'12)/P(B'5)
				269 T	0.965*	7.380 +- 2.530	T(rest)->A(A6-65, A')/G(B's)
				280 A	0.938	7.236 +- 2.702	A(As, A', B'5)->E(B'-OS and B'12)
CYA_2405	adenylate/guanylate cyclase	cAMP/cGMP synthesis	1014	205 Q	0.917	5.770 +- 2.924	Q(As and A')->A(B'-OS and B'12)/V(B'5)
CYA_2473	3,4-dihydroxy-2-butanone 4-phosphate synthase (EC 4.1.99.12) / GTP cyclohydrolase II (EC 3.5.4.25)	Molybdenum cofactor biosynthesis; Riboflavin, FMN and FAD metabolism	1245	1 F	0.997**	9.776 +- 1.048	F(As and A')->Q(B'-OS and B'12)/M(B'5)
				2 L	0.986*	9.680 +- 1.390	L(As and A')->V(B'-OS and B'12)/R(B'5)
				3 S	0.989*	9.699 +- 1.330	S(As and A')->F(B'-OS and B'12)/Q(B'5)
				4 L	0.944	9.310 +- 2.221	L(As and A')->P(B'-OS and B'12)/I(B'5)
				6 R	0.939	9.264 +- 2.300	R(As and A')->C(B'-OS and B'12)/P(B'5)

Table 5.8 (continued)

				265 T	0.934	9.213 +- 2.377	T(As and A')->D(B'-OS and B'12)/A(B'5)
				410 G	0.993**	9.737 +- 1.199	G(As and A')->C(B'-OS and B'12)/N(B'5)
				411 K	0.987*	9.688 +- 1.371	K(As and A')->T(B'-OS and B'12)/R(B'5)
				412 S	0.997**	9.775 +- 1.054	S(As and A')->T(B'-OS and B'12)/F(B'5)
				413 E	0.978*	9.608 +- 1.590	E(As and A')->L(B'-OS and B'12)/R(B'5)
CYA_2518	hyp		1191	64 Q	0.953*	5.899 +- 2.307	Q(As and A')->A(B'-OS and B'12)/E(B'5)
				73 S	0.956*	5.916 +- 2.293	S(As and A')->R(B'-OS and B'12)/K(B'5)
				132 Q	0.942	5.839 +- 2.353	Q(As and A')->T(B'-OS and B'12)/A(B'5)
				292 S	0.978*	6.020 +- 2.194	S(As, A', B'5)->R(B'-OS and B'12)
CYA_2560	Membrane protein PxcA, involved in light-induced proton extrusion # located in cytoplasmic membrane	photosynthetic electron transport	1380	149 A	0.909	5.500 +- 2.666	A(As and A')->R(B'-OS and B'12)/T(B'5)
				152 R	0.944	5.659 +- 2.559	R(As and A')->Q(B'-OS and B'12)/V(B'5)
CYA_2675	Inositol-1-monophosphatase (EC 3.1.3.25)	Streptomycin biosynthesis	819	75 T	0.926	5.816 +- 2.687	T(A1s, A14-63, A6-63)->A(REST)
				258 S	0.953*	5.904 +- 2.596	S(As and A')->Q(B'-OS and B'12)/R(B'5)
CYA_2727	response regulator		1200	13 E	0.920	6.048 +- 2.682	E(As and A')->A(B'-OS and B'12)/L(B'5)
				19 L	0.908	5.976 +- 2.731	L(As, A', B'-OS, B'12)->S(B'5)
CYA_2730	hyp		1137	378 P	0.960*	5.701 +- 2.672	P(As and A')->F(B'-OS and B'12)/A(B'5)
CYA_2886	response regulator receiver sensor signal transduction histidine kinase		3066	77 Q	0.917	6.114 +- 2.496	Q(As and A')->G(B'-OS and B'12)/R(B'5)
				115 G	0.996**	6.495 +- 2.086	G(A1s, A14-63, A6-63, A')->C(B's)/E(REST)

Table 5.9. Sites under positive selection with the A',A,B'5 //B'-OS,B'12 mutation pattern (Figure 5.2D). The site under selection and the probability it is under selection are given in, as well as the mutation pattern (right-most column). Mutation patterns highlighted in red do not fit the A',A,B'5 //B'-OS,B'12 pattern.

Gene	Annotated Function	Gene Subsystem	Alignment Length (nt)	Positively selected sites (*: P>95%; **: P>99%)[1]			A',A,B'5//B'-OS,B'12 (Other Pattern)[2]
CYA_0051	peptide transport system permease protein	ABC transport	1101	111 W	0.989*	8.250 +- 2.069	W(As, A', B'5) -> S(B'-OS and B'12)
				123 T	0.980*	8.191 +- 2.167	T(As and A')->I(B'-OS and B'12)/A(B'5)
				124 E	0.966*	8.095 +- 2.310	E(As, A', B'5)->G(B'-OS and B'12)
CYA_0160	hyp		756	233 V	0.979*	6.420 +- 2.685	V(As, A',B'5)->A(B'-OS and B'12)

[1] Amino acid sites indicated to be under positive selection by the BEB method are listed, along with the probability (P) that the sites are in a positive selection site class and the post mean +/- SE for positive selection.

[2] Amino acids (and the genomes of the isolates and lineages in which they occur) are listed. (REST) indicates the amino acid is present in the remainder of the isolate genomes not listed in the entry. STOP CODON indicates a stop codon insertion in the isolate genomes listed.

Table 5.9 (continued)

CYA_0378	hyp		543	**5 D**	**0.961***	**7.933 +- 2.564**	**D(As, A', B'5)->K(B'-OS and B'12)**
CYA_0552	nucleic acid binding, OB-fold, tRNA/helicase-type	DNA binding; DNA repair	366	**11 G**	**0.989***	**6.707 +- 2.307**	**G(As, A', B'5)->K(B'-OS and B'12)**
				20 A	**0.956***	**6.539 +- 2.484**	**A(As and A')->L(B's)**
CYA_0565	Dephospho-CoA kinase (EC 2.7.1.24)	Coenzyme A Biosynthesis	690	**111 A**	**0.959***	**9.622 +- 1.943**	**A(As and A')->E(B'-OS and B'12)/L(B'5)**
				113 E	**0.965***	**9.669 +- 1.844**	**E(As, A', B'5)->A(B'-OS and B'12)**
				114 A	**0.994****	**9.939 +- 1.046**	**A(As and A')->S(B'-OS and B'12)/E(B'5)**
				116 P	**1.000****	**9.985 +- 0.823**	**P(As, A', B'5)->K(B'-OS and B'12)**
				118 C	**0.953***	**9.566 +- 2.057**	**C(As and A')->S(B's)**
				119 Q	**0.974***	**9.751 +- 1.643**	**Q(As, A', B'5)->G(B'-OS and B'12)**
				120 C	**0.994****	**9.937 +- 1.050**	**C(As and A')->F(B'-OS and B'12)/K(B'5)**
				121 P	**0.925**	**9.311 +- 2.496**	**P(As and A')->S(B'-OS and B'12)/N(B'5)**
				122 H	**0.990***	**9.899 +- 1.201**	**H(As and A')->S(B's)**
				127 P	**0.996****	**9.952 +- 0.985**	**P(As and A')->Q(B'-OS and B'12)/S(B'5)**
				213 G	**0.992****	**9.917 +- 1.131**	**G(As and A')->Q(B'-OS and B'12)/E(B'5)**
CYA_0635	hyp		180	**1 P**	**0.945**	**7.849 +- 2.626**	**P(As, A', B'5)->Q(B'-OS and B'12)**
				3 V	**0.966***	**8.003 +- 2.441**	**V(As and A')->G(B'-OS and B'12)/L(B'5)**
CYA_1805	hyp		567	**174 E**	**0.920**	**6.286 +- 3.004**	**E(As, A', B'5)->T(B'-OS and B'12)**
CYA_1988	UDP-N-acetylglucosamine 1-carboxyvinyltransferase (EC 2.5.1.7)	Peptidoglycan Biosynthesis; UDP-N-acetylmuramate from Fructose-6-phosphate Biosynthesis	1443	**91 S**	**0.901**	**8.459 +- 2.799**	**T(As)->A(A' and B's)**
				461 A	**1.000****	**9.278 +- 1.347**	**A(As, A', B'5)-> Q(B'-OS and B'12)**
				462 T	**0.995****	**9.240 +- 1.460**	**T(As and A')-> L(B'-OS and B'12)/V(B'5)**
				463 P	**0.999****	**9.272 +- 1.367**	**P(As and A')-> G(B'-OS and B'12)/S(B'5)**
				464 Q	**0.982***	**9.138 +- 1.723**	**Q(As and A')-> K(B'-OS and B'12)/P(B'5)**
CYA_2078	hyp		1506	**260 K**	**0.988***	**8.375 +- 2.094**	**K(As and A') -> P(B'-OS and B'12)/E(B'5)**
				422 P	**0.998****	**8.443 +- 1.971**	**P(As and A') -> E(B'-OS and B'12)/L(B'5)**
				423 W	**0.928**	**7.941 +- 2.689**	**W(As, A', B'5)->S(B'-OS and B'12)**
				424 T	**0.981***	**8.321 +- 2.181**	**T(As and A') -> L(B'-OS and B'12)/G(B'5)**
CYA_2172	peptidase M16B family, nonpeptidase-like protein	Proteolysis	1311	**344 G**	**0.937**	**6.651 +- 2.615**	**G(As, A')-> L(B'-OS, B'12)/P(B'5)**
				410 N	**0.974***	**6.844 +- 2.417**	**N(As, A', B'5)->P(B'-OS and B'12)**
CYA_2178	transporter, major facilitator family		1284	**257 S**	**0.976***	**4.420 +- 2.328**	**S(As, A', B'5)->E(B'-OS and B'12)**
CYA_2199	LSU rRNA 2'-O-methyl-C2498 methyltransferase RlmM	RNA methylation	1053	**85 I**	**0.949**	**5.660 +- 2.891**	**I(As, A', B'5)->G(B'-OS and B'12)**
CYA_2320	hyp		477	**2 A**	**0.922**	**7.901 +- 2.751**	**A(As, A', B'5)-> L(B'-OS and B'12)**
				7 L	**0.970***	**8.254 +- 2.316**	**L(As, A', B'5)-> E(B'-OS and B'12)**
				20 W	**0.989***	**8.393 +- 2.100**	**W(As, A', B'5)-> L(B'-OS and B'12)**
CYA_2394	Nitrogen-responsive response regulator NrrA	Transcription	777	**35 N**	**0.941**	**7.809 +- 2.652**	**N(A1s, A14-63, A6-63)->T(REST)**

Table 5.9 (continued)

				236 A	0.956*	7.889 +- 2.527	A(As and A')->S(B's)
				237 A	0.971*	7.990 +- 2.388	A(As, A', and B'5)->V(B'-OS and B'12)
CYA_2466	glycosyl transferase, family 2	Polysaccharide synthesis	996	93 R	0.965*	5.524 +- 2.603	R(As, A', B'5)->N(B'-OS and B'12)
CYA_2515	hyp		537	173 S	0.993**	8.093 +- 2.305	S(As, A', B'5)->G(B'-OS and B'12)
CYA_2518	hyp		1191	64 Q	0.953*	5.899 +- 2.307	Q(As and A')->A(B'-OS and B'12)/E(B'5)
				73 S	0.956*	5.916 +- 2.293	S(As and A')->R(B'-OS and B'12)/K(B'5)
				132 Q	0.942	5.839 +- 2.353	Q(As and A')->T(B'-OS and B'12)/A(B'5)
				292 S	0.978*	6.020 +- 2.194	S(As, A', B'5)->R(B'-OS and B'12)
CYA_2563	MscS Mechanosensitive ion channel	Transport	1854	151 A	0.950*	4.623 +- 2.385	A(As, A', B'-OS)->R(B'5 and B'12)
CYA_2882	ABC transporter	ABC transport	1008	331 A	0.980*	6.207 +- 2.642	A(As, A', B'5)->L(B'-OS and B'12)

Table 5.10. Sites under positive selection within the A-lineage isolate genomes. The site under selection and the probability it is under selection are given in, as well as the mutation pattern (right-most column). Mutation patterns highlighted in red are not within the A-lineage.

Gene	Annotated Function	Gene Subsystem	Alignment Length (nt)	Positively selected sites (*: P>95%; **: P>99%)[1]			Mutation Within As (Other Pattern)[2]
CYA_0016	gas vesicle protein W	Buoyancy	747	7 P	0.902	8.010 +- 2.888	P (As and A')-> G (B's)
				214 M	0.960*	8.470 +- 2.318	M(A1s/A6-63)->T(A14s/A4/A6-65/A')/A(B's)
CYA_0061	RNA polymerase sigma-70 factor	Transcription initiation, sigma factors	579	128 G	0.918	2.980 +- 1.485	G(A1s, A4, A6s, A')->Q(A14s, B's)
				145 H	0.911	2.952 +- 1.472	H(A1s, A4, A6s, A')->S(A14s, B's)
CYA_0072	Geranylgeranyl hydrogenase ChlP	Chlorophyll Biosynthesis	1218	369 M	0.978*	5.968 +- 2.918	M(A1s, A14s, A4, A6-63, B's)->L(A6-65, A')
CYA_0074	Protein-export membrane protein SecD (TC 3.A.5.1.1)	Protein transport	1344	79 D	0.917	4.144 +- 2.802	D(A1-OS,A14-63,A4,A6-63)->N(A1-MS,A14-60/65,A6-65, A',B's)
CYA_0097	transporter, major facilitator family	Transport	1377	1 A	0.946	6.085 +- 2.214	A(rest)->L(A6-65)/S(B's)
				4 T	0.984*	6.272 +- 1.999	T(A1s,A14-63,A6-63,A')->I(A14-60/65,A4)/N(A6-65)/D(B's)
				76 W	0.948	6.085 +- 2.201	W(As, A', B'-OS)->Q(B'5,B'12)
CYA_0126	DNA double-strand break repair protein Mre11	Rad50-Mre11 DNA repair cluster	1302	425 E	0.907	6.758 +- 2.946	E(rest)->D(A1-MS)/S(B's)

[1] Amino acid sites indicated to be under positive selection by the BEB method are listed, along with the probability (P) that the sites are in a positive selection site class and the post mean +/- SE for positive selection.

[2] Amino acids (and the genomes of the isolates and lineages in which they occur) are listed. (REST) indicates the amino acid is present in the remainder of the isolate genomes not listed in the entry. STOP CODON indicates a stop codon insertion in the isolate genomes listed.

Table 5.10 (continued)

CYA_0164	2-octaprenyl-6-methoxyphenol hydroxylase (EC 1.14.13.-)	Plastoquinone Biosynthesis	1242	**136 H**	**0.984***	**9.975 +- 1.341**	**H(A1s, A14-63,A6-63)->Y(REST)**
				165 A	**0.967***	**9.813 +- 1.807**	**A(As and A')-> R(B's)**
				385 R	**0.901**	**9.205 +- 2.859**	**R(As and A')-> A(B's)**
				406 -	**1.000****	**10.122 +- 0.678**	**STOP CODON IN B's**
CYA_0167	5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase (EC 2.1.1.14)	Methionine Biosynthesis	1662	**226 I**	**0.989***	**10.052 +- 1.137**	**I(A1s,A4, A6-63)->V(REST)**
				534 G	**0.999****	**10.140 +- 0.700**	**G(A1s, A', B's)->P(A4, A6s, A14s)**
				535 L	**0.997****	**10.125 +- 0.794**	**L->G (SIM)**
				536 K	**1.000****	**10.147 +- 0.652**	**K->T (SIM)**
				537 E	**0.987***	**10.034 +- 1.203**	**E->P(SIM)**
				538 R	**0.940**	**9.598 +- 2.275**	**R->T (SIM)**
				544 S	**0.999****	**10.145 +- 0.666**	**STOP CODON (A4, A6s, A14s)**
				545 W	**0.978***	**9.944 +- 1.497**	**STOP CODON (A4, A6s, A14s)**
				548 D	**0.968***	**9.859 +- 1.725**	**STOP CODON (A4, A6s, A14s)**
				550 F	**0.980***	**9.970 +- 1.420**	**STOP CODON (A4, A6s, A14s)**
CYA_0177	pilin domain protein	cell motility	339	**98 -**	**1.000****	**9.783 +- 0.997**	**STOP (A4, A6s, A14s)/W(A1s)/R(B's)**
CYA_0182	Phosphate transport system regulatory protein PhoU	Phosphate metabolism	639	**118 G**	**0.962***	**6.789 +- 2.364**	**G(A1s,B's,A6-63)->S(REST)**
				119 R	**0.963***	**6.795 +- 2.358**	**R(A1s,B's,A6-63)->C(REST)**
				174 K	**0.950**	**6.719 +- 2.456**	**K(As and A') -> R(B's)**
CYA_0233	Fibronectin/fibrinogen-binding protein	Biofilm formation	1806	**228 S**	**0.998****	**9.497 +- 1.192**	**S(A1s, A14-63, A6s, A', B'5)->T(REST)**
				590 -	**1.000****	**9.513 +- 1.133**	**G(As, A')/A(B'-OS,B'12)/STOP B'5**
				599 E	**0.993****	**9.457 +- 1.323**	**E(As and A')->F(B'-OS,B'12)/STOP B'5**
				600 L	**0.916**	**8.797 +- 2.619**	**L(As and A')->F(B'-OS,B'12)/STOP B'5**
CYA_0234	hyp		567	**4 F**	**0.950**	**3.059 +- 1.185**	**F(A1s and B's)-> R(REST)**
				10 K	**0.925**	**3.003 +- 1.224**	**K(A1s and B's)-> R(REST)**
				11 G	**0.972***	**3.134 +- 1.190**	**G(A1s and B's)-> S(REST)**
				18 L	**0.935**	**3.011 +- 1.184**	**L(A1s and B's)-> G(REST)**
				20 L	**0.929**	**3.006 +- 1.209**	**L(A1s and B's)-> L(REST)**
				22 L	**0.932**	**2.996 +- 1.169**	**L(A1s and B's)-> G(REST)**
				26 G	**0.968***	**3.124 +- 1.192**	**G(A1s and B's)-> L(REST)**
				46 P	**0.976***	**3.141 +- 1.181**	**P(A1s and B's)-> N(REST)**
CYA_0235	oxidoreductase, short chain dehydrogenase/reductase family	Fatty acid biosynthesis	1266	**409 R**	**0.984***	**3.124 +- 1.101**	**R(A1s, A4, A')->P(A14s, A6s)/S(B'-OS)/Q(B'5,B'12)**
				412 L	**0.959***	**3.070 +- 1.136**	**L(A1s, A4, A')->V(A14s, A6s)/P(B'-OS)/I(B'5,B'12)**
CYA_0266	hyp		267	**85 L**	**0.990****	**7.231 +- 2.112**	**L(A1s, A14s, A6s, A')->C(A4, B's)**
				88 Y	**0.951***	**6.972 +- 2.395**	**Y(A1s, A14s, A6s, A')->Q(A4, B's)**
				89 P	**0.982***	**7.178 +- 2.174**	**P(A1s, A14s, A6s, A')->D(A4, B's)**
CYA_0312	hyp		369	**8 L**	**0.917**	**6.954 +- 3.068**	**L(A1s, A14-63, A4)->V (REST)**
CYA_0317	Phytoene synthase (EC 2.5.1.32)	Carotenoids	933	**3 S**	**0.997****	**3.868 +- 1.508**	**S(A1s, A14-63, A4, A6s)->S(REST)**

Table 5.10 (continued)

CYA_0323	truncated cyclic nucleotide-binding protein		513	**4 S**	**0.913**	**5.880 +- 2.736**	**S(As and A')->P(B'-OS and B'12)/T(B'5)**
				139 V	**0.970***	**6.213 +- 2.530**	**V(A1s, A14-63, A4, A6-63)->L(REST)**
CYA_0337	Pantothenate kinase type III, CoaX-like (EC 2.7.1.33)	Coenzyme A Biosynthesis cluster	732	**80 S**	**0.999****	**8.496 +- 1.902**	**S(A1s, A4, A6-63)->G(REST)**
CYA_0350	Transcription termination protein NusB	Transcription factors	864	**267 L**	**0.968***	**6.526 +- 2.735**	**L(A1s, A4, A6-63)->V(REST)**
CYA_0416	hyp		393	**3 F**	**1.000****	**10.394 +- 0.361**	**F(A1s, A4, A6-63)->L(A14s, A6-65, A')/H(B's)**
				16 L	**0.960***	**10.020 +- 1.882**	**L(As and A')->N(B's)**
				81 -	**1.000****	**10.397 +- 0.329**	**STOP CODON B's**
				86 -	**1.000****	**10.397 +- 0.329**	**STOP CODON B'-OS and B'12**
CYA_0418	hyp		492	**155 V**	**0.988***	**8.040 +- 2.223**	**V(A1s, A4, A6-63)->A(A14s, A6-65, A') /T(B's)**
CYA_0426	hyp		558	**12 W**	**0.984***	**7.973 +- 2.405**	**W(A1s, A6-63)->C(A14s, A4, A6-65, A')/L(B's)**
CYA_0467	Peptide methionine sulfoxide reductase MsrA (EC 1.8.4.11)	Peptide methionine sulfoxide reductase	672	**220 P**	**1.000****	**10.307 +- 0.475**	**P(A1-OS,B'12)->A(B'5)/E(REST)**
				221 D	**1.000****	**10.307 +- 0.479**	**D(A1-OS)>C(B'5)/P(B'-OS)/V(B'12)/S(rest)**
				222 -	**1.000****	**10.309 +- 0.458**	**STOP CODON B'12**
				223 -	**1.000****	**10.309 +- 0.458**	**STOP CODON B'-OS**
CYA_0555	hyp		345	**96 K**	**0.990****	**8.291 +- 2.202**	**K(A1s, A14-65, A4, A6-63)->E(REST)**
CYA_0567	hyp		369	**15 V**	**0.988***	**9.192 +- 1.690**	**V(A1s, A4, B's)->L(A14s, A6-63)/I(A6-65, A')**
				87 V	**0.961***	**8.975 +- 2.140**	**V(A1s, A4, A6-63)->A(REST)**
				98 L	**0.979***	**9.123 +- 1.849**	**L(A1-OS, A6-65, A')->V(REST)**
CYA_0576	hyp		909	**72 E**	**0.946**	**6.229 +- 2.660**	**E(A1s, A14-63, A6-63)->T(B's)/A(REST)**
				195 E	**0.957***	**6.269 +- 2.616**	**E(As and A')->R(B'-OS and B'12)/Q(B'5)**
CYA_0580	Iojap protein	NAD and NADP cofactor biosynthesis global; tRNA modification	336	**83 V**	**0.943**	**6.944 +- 2.959**	**V(A1s, A14-63, A4, A6-63, A', B's)->I(REST)**
CYA_0614	Nitrate ABC transporter, permease protein	Nitrate and nitrite ammonification	957	**45 G**	**0.991****	**5.190 +- 1.261**	**G(rest)->H(A1-MS, B'5)**
				86 A	**0.987***	**5.175 +- 1.284**	**I(rest)->A(A1s, B'5)**
CYA_0628	Nitroreductase	Nitrogen reduction	624	**104 A**	**0.945**	**7.061 +- 2.897**	**A(A1s, A14-63, A6-63)->P(REST)**
CYA_0644	Photosystem II CP47 protein (PsbB)	Photosystem II	1530	**351 S**	**0.982***	**7.270 +- 2.667**	**S(A1s, A14-63, A4, A6s)->G(REST)**
CYA_0646	Phosphoribosylformylglycinamidine synthase, glutamine amidotransferase subunit (EC 6.3.5.3)	De Novo Purine Biosynthesis	675	**224 A**	**0.945**	**6.165 +- 2.348**	**A(A1s, A6s, B's)->S(REST)**
				225 G	**0.988***	**6.367 +- 2.119**	**G(A1s, A6s)->S(A14s, A4, A')/A(B's)**
CYA_0649	hyp		771	**115 V**	**0.982***	**7.134 +- 2.610**	**V(A1s, A6-63, A')->M(A14s, A4, A6-65)/L(B's)**
				116 I	**0.945**	**6.950 +- 2.800**	**I(A1s, A6-63, A')->M(A14s, A4, A6-65)/L(B's)**
CYA_0650	hyp		282	**88 D**	**0.992****	**7.322 +- 2.149**	**D(A1s, A14-63, A6s, A')->A(A14-60/65,A4)/P(B's)**
				89 K	**0.949**	**7.033 +- 2.463**	**K(A1s, A14-63, A6s, A')->E(A14-60/65,A4)/D(B's)**
CYA_0651	2-phosphosulfolactate phosphatase(EC:3.1.3.71)	Coenzyme M biosynthesis; methane	726	**215 S**	**0.983***	**7.695 +- 2.224**	**S(A1s, A6-65, A')->Q(B',B'12)/K(B'5)/R(REST)**

Table 5.10 (continued)

		metabolism					
				222 K	0.987*	7.725 +- 2.186	K(A1s, A6-65, A')->D(B',B'12)/Q(B'5)/E(REST)
CYA_1776	hyp		351	97 C	0.995**	8.258 +- 2.146	C(A1s, A14-63, A4, A6-63)->Q(B's), R(REST)
CYA_1905	Crossover junction endodeoxyribonuclease RuvC	RuvABC plus a hypothetical	579	1 E	0.998**	10.336 +- 0.545	E(As and A')->V(B'-OS and B'5)/C(B'12)
				2 S	0.998**	10.336 +- 0.545	S(As and A')->M(B'-OS and B'5)/Q(B'12)
				3 S	0.902	9.418 +- 2.860	S(As and A')->F(B'-OS and B'5)/I(B'12)
				4 R	0.992**	10.278 +- 0.912	R(As and A')->S(B'-OS and B'5)/P(B'12)
				5 A	1.000**	10.349 +- 0.409	A(A1s)->S(B'-OS and B'5)/P(B'12)/V(REST)
				6 T	1.000**	10.348 +- 0.427	T(A1s)->V(B'-OS and B'5)/P(B'12)/S(REST)
				7 R	0.999**	10.342 +- 0.483	R(A1s)->Q(B'-OS, B'5)/L(B'12)/G(REST)
				8 T	0.999**	10.342 +- 0.486	T(As and A')->R(B'-OS and B'5)/A(B'12)
				9 E	0.990*	10.255 +- 1.022	E(As and A')->K(B'-OS and B'5)/D(B'12)
				10 S	0.994**	10.295 +- 0.819	S(As and A')->C(B'-OS and B'5)/R(B'12)
				11 S	0.969*	10.060 +- 1.681	S(REST)->G(B'12)
CYA_1969	phosphoglucomutase	Glycogen; Glycolysis/ gluconeogenesis	1629	248 L	0.988*	4.711 +- 2.353	L(A1s, A14-63, A6-63)->G(REST)
				249 H	0.915	4.481 +- 2.465	H(A1s, A14-63, A6-63)->R(REST)
CYA_2054	hyp		279	34 L	0.994**	7.367 +- 2.403	L(A1s, A14-60/65, A6-65, A')>A(B's)/G(A14-63,A6-63, A4)
CYA_2055	DnaJ-class molecular chaperone CbpA	Protein chaperones	930	21 L	0.969*	3.609 +- 1.544	L(A1s, A14s, A4, A')->A(A6s and B's)
CYA_2116	ATP synthase F0 sector subunit b' (EC 3.6.3.14)	ATP synthesis; Oxidative phosphorylation	471	4 S	0.998**	7.993 +- 2.057	S(A1s, A6s, A')->P(A14s, A4, B')/Q(B'12, B'5)
CYA_2123	2,3-bisphosphoglycerate-independent phosphoglycerate mutase (EC 5.4.2.1)	Glycolysis and Gluconeogenesis	1590	526 A	0.934	2.747 +- 1.464	A(A1s, A6-65, A')->S(A14s, A4, A6-63, B's)
CYA_2132	hyp		744	214 R	0.979*	7.761 +- 2.387	R(A1-OS, B's)->V(REST)
				241 G	0.983*	7.791 +- 2.353	G(A1-OS, B's)->L(REST)
CYA_2151	competence protein/comEA-related protein	DNA binding; DNA repair	1761	143 D	0.929	4.675 +- 2.676	D(A1s, A14-63, A4, A6-63, B', B'12)->A(REST)
				144 Q	0.938	4.692 +- 2.660	Q(A1s, A14-63, A4, A6-63, B'-OS, B'12)->E(REST)
CYA_2153	hyp		480	115 T	0.997**	7.751 +- 2.128	T(A1s, A14-63, A4, A6-63)->P (REST)
CYA_2161	FOG: CheY-like receiver		372	122 A	0.946	7.248 +- 2.738	A(A1s, A6-63)->P(REST)
CYA_2162	two-component response regulator		1242	3 C	0.996**	9.740 +- 1.170	C(A1s, A6-63, B'5)->R(REST)
				126 G	0.958*	9.403 +- 2.049	G(rest)->S(A6-65, A')
				280 E	0.999**	9.773 +- 1.037	E(A1s, A6s, A')->G(REST)
				308 E	0.958*	9.403 +- 2.049	E(rest)->G(A6-65, A')
				322 A	0.996**	9.740 +- 1.170	A(A1s, A6-63, B'12)->T(REST)
CYA_2229	peptidase, S1C (protease Do) family(EC:3.4.21.-)	Proteolysis	1260	23 F	0.993**	9.417 +- 1.451	F(As)->I(A' and B's)
				24 V	0.995**	9.431 +- 1.410	V(As)->I(A6-65))/T(A' and B's)
				25 L	1.000**	9.471 +- 1.284	F(As)->I(A' and B's)

Table 5.10 (continued)

CYA_2305	alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen	Stress response	465	**8 A**	**0.942**	**7.170 +- 2.896**	**P(A4)->V(B'-OS and B'12)/A(REST)**
CYA_2329	DNA primase (EC 2.7.7.-)	Macromolecular synthesis operon	1893	**99 E**	**0.996****	**7.573 +- 2.175**	**E(A1s, A4, A6-63)->Q(B's)/G(REST)**
				189 A	**0.968***	**7.425 +- 2.391**	**A(A1s, A4, A6-63)->V(REST)**
				415 T	**0.978***	**7.465 +- 2.321**	**T(A1s, A6-63)->L(B's)/I(REST)**
				551 V	**0.951***	**7.297 +- 2.517**	**V(A1s, A', A6-63, B's)->L(REST)**
CYA_2332	Conserved membrane protein in copper uptake, YcnI	Metal binding	564	**70 F**	**0.941**	**3.673 +- 1.574**	**F(A1s, B's)->A(REST)**
				127 A	**0.989***	**3.834 +- 1.521**	**A(A1s, B'-OS, B'12)->L(REST)**
				163 K	**0.932**	**3.649 +- 1.594**	**K(A1s, B'-OS, B'12)->N(B'5)/H(REST)**
				180 G	**0.940**	**3.676 +- 1.583**	**G(A1s, B's)->T(REST)**
CYA_2361	NAD-dependent epimerase/dehydratase family protein		954	**264 I**	**0.944**	**7.256 +- 2.669**	**I(As and A')->L(B'-OS and B'12)/P(B'5)**
				269 T	**0.965***	**7.380 +- 2.530**	**T(rest)->A(A6-65, A')/G(B's)**
				280 A	**0.938**	**7.236 +- 2.702**	**A(As, A', B'5)->E(B'-OS and B'12)**
CYA_2374	fibrillin	Extracellular matrix	615	**176 T**	**0.982***	**6.304 +- 2.524**	**T(A1s, A14-63, A4, A6-63, B's)->R (REST)**
				190 L	**0.980***	**6.301 +- 2.532**	**L(A1s, A6-63, B's)->I(REST)**
CYA_2390	Cobalt-precorrin-6x reductase (EC 1.3.1.54)	Coenzyme B12 biosynthesis	747	**248 R**	**0.989***	**6.890 +- 2.646**	**R(A1s, A14-63, A4, A6-63, B'-OS, B'12)->G(REST)**
CYA_2394	N-responsive response regulator NrrA	Transcription	777	**35 N**	**0.941**	**7.809 +- 2.652**	**N(A1s, A14-63, A6-63)->T(REST)**
				236 A	**0.956***	**7.889 +- 2.527**	**A(As and A')->S(B's)**
				237 A	**0.971***	**7.990 +- 2.388**	**A(As, A', and B'5)->V(B'-OS and B'12)**
CYA_2408	Cobalamin synthase (EC 2.7.8.26)	Coenzyme B12 biosynthesis	777	**166 D**	**0.965***	**7.235 +- 2.744**	**D(A1s, A14-63, A6s, A')->N(A14-60/65,A4)/E(B's)**
CYA_2447	bacterial pre-peptidase C-terminal domain protein	Poss. transcription regulator	900	**4 K**	**0.933**	**7.854 +- 2.592**	**K(As and A')->S(B's)**
				8 L	**0.920**	**7.752 +- 2.700**	**L(As and A')->N(B's)**
				20 V	**0.917**	**7.732 +- 2.722**	**V(As and A')->Q(B's)**
				30 D	**0.914**	**7.709 +- 2.747**	**D(As and A')->P(B's)**
				188 R	**0.911**	**7.687 +- 2.775**	**R(As and A')->I(B's)**
				275 A	**0.977***	**8.204 +- 2.158**	**A(A1s, 63/A14-65, A6-63)->T(REST)**
CYA_2495	Histidyl-tRNA synthetase (EC 6.1.1.21)	Histidine Biosynthesis	1257	**193 L**	**0.980***	**5.542 +- 2.415**	**L(A1s, A14-63, A4, A6-63)->P(REST)**
				419 R	**0.922**	**5.314 +- 2.581**	**R(As and A')->G(B'-OS and B'5)/S(B'12)**
CYA_2511	hyp		183	**37 G**	**0.983***	**7.037 +- 2.583**	**G(A1s, A6-65, A', B'5)->L(REST)**
CYA_2605	Flavodoxin 1	Flavodoxin	507	**40 T**	**0.933**	**3.870 +- 1.585**	**T(A1s, A4, A')->S(REST)**
CYA_2606	IsiA [Chlorophyll a(b) binding protein, photosystem II CP43 protein (PsbC) homolog]	Chl *a*-binding protein; light harvesting	1032	**223 A**	**0.971***	**4.853 +- 2.589**	**A(A1s, A4, A6-63, A', B'5)->S(A14s, A6-65, B'-OS, B'12)**
CYA_2608	putative iron-regulated protein A		1230	**74 G**	**0.960***	**7.159 +- 2.773**	**G(A1s, A14-63, A6-63)->R(REST)**
CYA_2614	hyp		774	**6 L**	**0.986***	**8.640 +- 2.025**	**L(A1s, A6-63, B's)->S(REST)**
				57 A	**0.986***	**8.639 +- 2.027**	**A(A1s, A4, A6-63, B's)->T(A14s, A6-65, A')**
CYA_2615	hyp		660	**139 T**	**0.936**	**6.988 +- 2.828**	**T(REST)->K(A6-65, A')**

Table 5.10 (continued)

CYA_2618	response regulator		996	**329 D**	**0.921**	**8.584 +- 2.595**	**D(A1s, A14s)->Y(A4, A6s, A')/L(B's)**
				331 S	**0.999****	**9.217 +- 1.420**	**S(A1s, A14s, B's)->P(A4, A6s, A')**
				332 Q	**1.000****	**9.219 +- 1.414**	**Q(A1s, A14s)->S(A4, A6s, A')/G(B's)**
CYA_2675	Inositol-1-monophosphatase (EC 3.1.3.25)		819	**75 T**	**0.926**	**5.816 +- 2.687**	**T(A1s, A14-63, A6-63)->A(REST)**
				258 S	**0.953***	**5.904 +- 2.596**	**S(As and A')->Q(B'-OS and B'12)/R(B'5)**
CYA_2678	Cobyric acid synthase (EC 6.3.5.10)	Coenzyme B12 biosynthesis	1485	**234 V**	**0.922**	**3.393 +- 2.344**	**V(REST)->T(A14-60/65, A')**
CYA_2769	Histidinol-phosphatase [alternative form] (EC 3.1.3.15)	Histidine metabolism	792	**261 S**	**0.997****	**9.871 +- 0.999**	**S(A1s, A6-65, A')/G(A14s, A6-63)/E(A4)/H(B's)**
				262 G	**1.000****	**9.891 +- 0.908**	**G(A1s, A6-65, A')/I(A14s, A6-63)/T(A4)/F(B's)**
				263 D	**1.000****	**9.893 +- 0.896**	**D(A1s, A6-65, A')/E(A14s, A6-63)/S(A4 and B's)**
				264 R	**1.000****	**9.891 +- 0.908**	**R(A1s, A6-65, A')/K(A14s, A6-63)/G(A4)/P(B's)**
CYA_2802	hyp		306	**4 G**	**0.997****	**8.255 +- 2.045**	**G(A1-OS, A6-63, A')->N(B's)/A(REST)**
CYA_2803	hyp		738	**49 C**	**0.922**	**8.240 +- 2.647**	**C(As and A')->V(B's)**
				71 P	**0.999****	**8.826 +- 1.679**	**P(A1s, A14-63, A6-63)->H(B's)/T(REST)**
CYA_2877	peptidase M14, carboxypeptidase A	Proteolysis	627	**148 N**	**0.957***	**5.971 +- 2.791**	**N(rest)->D(A14-60/65)/A(B's)**
CYA_2886	response regulator receiver sensor signal transduction histidine kinase		3066	**77 Q**	**0.917**	**6.114 +- 2.496**	**Q(As and A')->G(B'-OS and B'12)/R(B'5)**
				115 G	**0.996****	**6.495 +- 2.086**	**G(A1s, A14-63, A6-63, A')->C(B's)/E(REST)**

in two genes); at those 32 sites, PE A14-63 often had the same mutation as the PE A1 isolates (21 sites).

Genes under Positive Selection

CodeML also detected genes with dN/dS ratio >1, indicating positive selection of a gene, in 18 orthologous genes (Table 5.5, above diagonal; Supplementary Table B.2).

Differences among Genomes of Lineages B′, A, and A′ Defined by 16S rRNA. A two-component response regulator (CYA_2162) is the only gene to display selection between A′/A- and B′-lineage isolates, showing selection in (i) PE A6-63 relative to PE B′-OS, PE A′, PE A4, PE A6-65, PE A14-60, and PE A14-63; and (ii) PE A14-65 relative to PE A6-65 and PE A′. Many genes showed positive selection between the PE A′ isolate and A-lineage isolates. DNA primase (CYA_2329) exhibited evidence of positive selection in A′ relative to the A-lineage isolates. Geranylgeranyl hydrogenase *chlP* (CYA_0072) exhibited selection evidence in PE A′ and PE A6-65 relative to the rest of the A-lineage isolates). A response regulator (CYA_2618) exhibited evidence of selection in PE A4, the PE A6 isolates, and PE A′ relative to PE A1-OS and the three PE A14s, along with selection in a ComEA-related competence protein (CYA_2151; selection in (i) PE A6 isolates relative to PE A14-60 and PE A14-65, and (ii) PE A′ relative to PE A4) and a hypothetical protein (CYA_0567; selection in PE A1-MS relative to PE A6-65 and PE A′).

Differences among Isolates Representative of Different PEs in the B′ and A Lineages. Two genes, the glutamine amidotransferase subunit of phosphoribosyl-formylglycinamidine synthase (CYA_0646) and phosphoglucomutase (CYA_1969) exhibited evidence of selection between PE A4 and the PE A1 isolates. The other genes include a histidinol-phosphatase (CYA_2769; selection in PE A4 relative to the three PE A14s and PE A6-63) and a molecular chaperone gene *cbpA* (CYA_2055; selection in PE A6-63 relative to PE A6-65).

The remaining genes under selection are within the B′-lineage. Four genes showed selection in the PE B′5 isolate relative to PE B′-OS and PE B′12 (which could also be interpreted as selection in PE B′-OS and PE B′12 relative to the PE B′5): two hypothetical proteins (CYA_0416 and CYA_2132), a gene annotated as gas vesicle protein W (CYA_0016), and a protein with a bacterial pre-peptidase C-terminal domain (CYA_2447). Two genes had dN/dS ration >1 in the PE B′12 isolate genome relative to PE B′-OS and PE B′5, peptide methionine sulfoxide reductase *msrA* (CYA_0467) and crossover junction endodeoxyribonuclease *ruvC* (CYA_1905). Two other genes exhibited evidence of selection indicated within B′-lineage isolates: histidyl trRNA synthetase (CYA_2495) in PE B′-OS relative to PE B′5 and a pilin domain protein (CYA_0177) in PE B′-OS relative to PE B′12.

Differences among Replicate Isolates Representative of PEs A1 and A14. There was no evidence of genes under selection between members of an ecotype (i.e. the PE A1 and PE A14 isolate genomes).

Discussion

Numerous differences were found among the genomes of the 12 *Synechococcus* sp. isolates compared in this study. The genomes from the A'/A and B' lineages defined by 16S rRNA differ in GC content; the A'/A-lineage genomes have an average of 60.4% GC content and the B'-lineage genomes have an average GC content of 58.6%. This is associated with a codon-usage bias evident in the A'/A-lineage isolates, which tend to use codons with higher GC content (Table 5.2). This is most likely due to the higher thermostability of a G-C pair, which has three hydrogen bonds, compared to an A-T pair, which only has two hydrogen bonds; this phenomenon has been observed in thermophiles (e.g. in *Thermus thermophilus*, Kagawa et al., 1984) and thermotolerant strains (*Corynebacterium efficiens*, Nishio et al., 2003). Differences have also been found in gene content, sites under positive selection, and genes under positive selection for isolates representative of (i) populations with different temperature distributions, (ii) PEs within these different temperature-distributed populations, some of which have distinct vertical distributions in the mat, and (iii) replicate members of PE A1 and A14. These differences are summarized in Table 5.11.

There are several lineage-specific genes separating the B'-lineage isolate genomes from the A′/A-lineage isolate genomes, including genes for cyanophycin metabolism and phosphonate transportation and degradation. There is only one gene specific to the PE A′ isolate (an extra copy of a putative monoamine oxidase similar to tyramine oxidase, which produces an aldehyde and ammonia from decarboxylated

Table 5.11. Comparative genomic analyses of *Synechococcus* isolates with regards to different ecological parameters. HL refers to high-light-adapted isolates (representatives of PE A1) from Chapter 4 (Olsen et al., submitted). LL refers to low-light-adapted isolates (representatives of PEs A4 and A14) from Chapter 4 (Olsen et al., submitted), as well as the PE A6-63 isolate.

	Gene Content Differences	**Sites under Positive Selection**	**Genes under Positive Selection**
Among Genomes of Lineages A', A, and B' Defined by 16S rRNA	B'-lineage specific genes • Cyanophycin metabolism • Phosphonate transport/metabolism • Succinate dehydrogenase, iron-sulfur subunit A'-lineage specific gene • Monoamine oxidase (extra copy) No A-lineage specific genes	B's vs A': 223-258 sites B's vs As: 228-276 sites A' vs As: 37-83 sites	B's vs A': 0 genes under selection B's vs As: 1 gene under selection A' vs As: 1-4 genes under selection
Among Isolates Representative of Different PEs in the B′- and A-lineages	No within B'-lineage differences Within A-lineage differences • HL (e.g. carbonic anhydrase) • LL (e.g. putative antenna proteins, ferrous iron transport) Other Intra-PE differences • Nutrient transport • Nitrogen metabolism	Within B'-lineage: 44-131 sites Within A-lineage: 37-108 sites HL vs LL isolates: 64-108 sites Within LL isolates: 37-64 sites	Within B'-lineage: 3-6 genes Within A-lineage: 1-4 genes HL vs LL isolates: 0-2 genes Within LL isolates: 1-3 genes
Among Replicate Isolates Representative of PEs A1 and A14	PE A1 isolates • No PE-specific genes • Differences (urea carboxylase, bipolar DNA helicase cluster) PE A14 isolates • PE specific (PotABCD cassette) • Differences (no *nif* genes in one isolate)	Between PE A1 isolates: 6 sites Among PE A14 isolates: 2-32 sites	Between PE A1 isolates: 0 genes Among PE A14 isolates: 0 genes

tyrosine), and there are no lineage-specific genes shared by the A-lineage isolates. Most of the site-specific positive selection patterns differentiated one or more B'-lineage isolates from the A'/A-lineage as a whole. Between 233 and 276 different sites are under selection between B′-lineage and A′/A-lineage isolates compared to 37-83 sites that differed between the A-lineage and the PE A′ isolate.

There are no genes that exhibit evidence of positive selection between the PE A′ isolate and the B′-lineage isolates, however, and there is only one gene that exhibits evidence of positive selection between A-lineage isolates and B′-lineage isolates (specifically between the PE A6-63 and PE B′-OS isolates; see Table 5.5, above diagonal). This may be due to a phenomenon observed by Rocha and colleagues (2006), who observed a variable dN/dS rate among closely related strains of bacteria. The authors hypothesized that the positive selection of an allele to fixation in a population could carry along several "hitchhiking" non-synonymous mutations that, although not deadly, are slightly deleterious. A period of time would pass before these slightly deleterious, non-synonymous mutations could be selected against, which would be not be observed in closely-related populations, but would be observed in more distantly-related populations. Therefore, even though a site may be under positive selection, the purifying selection of the non-synonymous sites surrounding it would eventually dampen the dN/dS signal of positive selection.

The hypothesis presented by Rocha et al. (2006) could also explain the relatively high (1-4) number of genes under selection in the PE A' isolate compared to the A-

lineage isolates. Because the PE A' isolate is more closely related to the A-lineage than the B'-lineage isolates (Figure 5.1), there would have been less time since the divergence of the A′ and A lineages for slightly deleterious, non-synonymous mutations in the PE A' and A-lineage isolate sequences to be removed by purifying selection. Though the PE A' isolate genome only has one novel gene, there are several genes and sites that exhibit positive selection. This could be indicative of the evolutionary changes needed for an ecotype to occupy a niche with a higher temperature than that of the parent ecotype. Expanding into a higher temperature niche requires protein thermostability for continuing cell function (e.g. Miller, 2003), not a novel function encoded by a gene. The sites and genes under selection may be changes providing such as thermostability.

As discussed in Chapter 4 (Olsen et al., submitted), many isolate-specific genes reflect different strategies for nitrogen metabolism and nitrogen acquisition, such as the presence and absence of the urease/urea carboxylase genes, or the ability to store and degrade cyanophycin. Although a *de novo* assembly of the PE B'5 genome could reveal isolate-specific genes, the B'-lineage isolates have similar gene content, with no notable differences within the lineage. Several site mutations were found within the B'-lineage, especially between the PE B'5 isolate and the other two B′-lineage isolates, PE B'-OS and PE B'12, that are indicative of the greater evolutionary divergence between the B'-lineage PEs and the A-lineage PEs, which had fewer mutations that differentiated PEs.

The PE A6 isolates do not appear to be ecologically interchangeable, as hypothesized in Chapter 3, because they have several gene content and positive selection

pattern differences. The absence of a gene cassette specific to the other low-light-adapted isolates in PE A6-65, but not in PE A6-63, that is thought to confer distinct fitness advantages to organisms living in low light (Shen, unpublished data) and the presence of carbonic anhydrase in PE A6-65, which is found only in A-lineage isolates predominating in the upper portion of the mat, suggest that PE A6-63 and PE A6-65 do not occupy the same niche. This is also reflected in the site-selection pattern differences (PE B' and PE B'12 have more sites under selection in common than PE A6-63 and PE A6-65) and the genes under selection, for which three genes (geranylgeranyl hydrogenase *chlP*, molecular chaperone *cbpA*, and histidinol-phosphatase) show selection between the PE A6-63 and the PE A6-65 sequences. I hypothesize that the PE A6-63 isolate, which has similar growth patterns (Nowack, unpublished) and gene content to the other low-light-adapted isolates, reflects the observed PE A6 *psaA* environmental distribution (Becraft et al., submitted; Figure 1.7) and is representative of a low-light-adapted PE.

The differences among the PE A14, PE A4, and PE A6-63 isolates in (i) gene content, such as the presence of beta-carotene ketolase in the PE A14 and PE A6-63 isolates, but not the PE A4 isolate; (ii) sites exhibiting positive selection (37-64 sites under selection between any two low-light-adapted isolates); and (iii) genes under selection could reflect the evolution of these closely-related PEs and the distinct, ecological niches they occupy which allow them to co-exist in the bottom portion of the microbial mat. The higher number of genes under selection within the low-light-adapted isolates (1-3 genes) compared to the number under selection between the high-light-

adapted and low-light-adapted isolates (0-2 genes) is consistent with the variable dN/dS rate hypothesis (Rocha et al., 2006).

Comparative genomics of multiple strains of a PE have revealed that members of a PE are mostly similar in their gene content and selection patterns, and are more similar to each other than to members of other closely related PEs. PE A1 and PE A14 isolates have identical within-PE gene content, with a few exceptions. The most prominent, the absence of *nif* genes in the PE A14-65 isolate, might reflect a recent evolutionary strategy of the PE A14 population at higher temperatures. PE A14-65 was isolated from a 65°C sample, and might have different nitrogen acquisition needs than down-channel populations. Alternatively, this PE A14 variant might utilize a "cheating" strategy, by feeding off the fixed nitrogen produced by surrounding populations and gaining a fitness advantage by conserving the energy otherwise spent on nitrogen fixation (Stefanic et al., 2012). PE A1 and PE A14 isolates also have very similar within-PE selection patterns, although the PE A14-63 isolate seems to be more different from the PE A14-60 and PE A14-65 isolates than they are from each other. This could be a sequencing artifact, as the PE A14-63 isolate genome was found to be heterozygous (two different sequence types existing at a locus) (Chapter 3). Alternatively, this could be indicative of ecological heterogeneity within a PE, similar to that found in multiple strains of a PE in *Bacillus* sp. (Kopac et al., 2014), or genome plasticity within strains of an ecotype, in which differences may accumulate that have a negligible effect on the niche-defining characteristics of an organism.

The genes identified to be under selection in the genomes are involved in many vital cell functions, including DNA repair, motility, cell signaling, and chlorophyll biosynthesis. These differences, along with the differences indicated in the site-specific positive selection and gene content analyses, may not always suggest a clear niche for each PE, but do strongly suggest differences between members of different PEs. Using these differences, experiments can be designed that can qualitatively differentiate the response of PEs and solidify our understanding of their ecological niche.

CHAPTER SIX

CONCLUSIONS

The ability to obtain and analyze the genome sequences of microorganisms has revolutionized the field of microbiology. No longer limited to phenotypic characterization, genome sequencing and analysis can be used to uncover mechanisms of ecology, pathogenicity, and secondary metabolite biosynthesis, among other microbial capabilities, and to characterize strains of microorganisms based on their genetic makeup (Fraser-Liggett, 2005). This knowledge has led to many breakthroughs in medicine, bioengineering, and bioremediation that are beneficial to humanity, such as novel diagnostic tool and vaccine development for pathogenic bacteria, but also to a deeper understanding of the microbial world. As sequencing technologies became less expensive and time consuming, investigators began to sequence multiple strains of closely related bacteria for comparative analysis. One of the first comparative genomic analyses of this kind, the comparison of *E. coli* strains K-12 and O157:H7, revealed that almost 1/3 of the genes in the two genomes were strain-specific (Perna et al., 2001). As more strains of named species were sequenced, more within-species genome sequence diversity was discovered. Genomospecies and pan-genome species concepts (e.g. Boucher et al., 2001; Medini et al., 2005) emerged, and some workers were even prompted to declare that, due to rampant gene exchange, microorganisms may not form groups that could be defined as species (Doolittle and Papke, 2006). Others called to question the validity of the initial,

phenotype-based species demarcations, and proposed alternative models of bacterial speciation based in ecological theory (Cohan, 2002; Cohan and Perry, 2007; Koeppel et al., 2008), including the Stable Ecotype Model (Figure 1.1), upon which the Ward lab has based prediction of ecological species, or putative ecotypes (PEs), of *Synechococcus* in hot spring microbial mats (Ward and Cohan, 2005; Ward et al., 2006; Melendrez et al., 2011; Becraft et al., 2011; submitted). My thesis project has involved the cultivation and comparative genomic analysis of multiple isolates representing PEs predominant in the environment to test the hypothesis that members of a PE are both ecologically interchangeable with other members of the same PE, and ecologically distinct from members of other PEs, and to begin to explore genetic differences that underlie adaptation to temperature, light, and other physico-chemical parameters. Although these efforts are ongoing, I will summarize what I have learned to date.

Cultivation and Genome Sequencing of *Synechococcus* Isolates

Together with Shane Nowack and other members of the Ward lab, I have cultivated isolates of *Synechococcus* representing predominant PEs of the Mushroom Spring mat community (Chapter 2; Nowack et al., submitted; Schaible et al., unpublished data). Because we have not yet been able to grow these organisms axenically, a protocol was developed to verify that each culture is predominated by a *Synechococcus* sequence type representative of a single PE. Genome sequencing and assembly of *Synechococcus* sequences from these enrichment cultures is possible due to the existence of reference

genomes for many of the heterotrophic contaminants, to which most contaminating sequences may be recruited and removed from the *Synechococcus* sequence pool. One of the most important aspects of these cultivation efforts was the development of a specialized growth medium, which allowed the growth of *Synechococcus* colonies from highly diluted mat samples and increased the growth rate of the cyanobacteria in these cultures. Single-cell genome sequencing has become more accessible, and has already revealed considerable information about uncultivated organisms in the environment (e.g. Rinke et al., 2013; Kashtan et al., 2014; Thrash et al., 2014), but microbiologists should still strive for cultivation of microorganisms, which can often be accomplished through changes to growth medium, growth conditions, or by exploiting co-culture and enrichment cultures (Pham and Kim, 2012; Tank and Bryant, 2015). Cultivation of *Synechococcus* isolates representing different PEs has not only allowed us to obtain genome sequences. The representative cultures have also allowed us to test growth rates under different light intensities and temperatures, observe differences in cell morphology under different growth conditions, and measure differences in absorbance spectra related to photosynthetic apparatus variations among different PEs, none of which would have been possible without obtaining representative isolates (Nowack et al., submitted; Schaible et al., unpublished data). To date, we have isolated and obtained genome sequences for *Synechococcus* isolates representative of (i) four of the five *psaA*-defined PEs predominant at 60°C in the Mushroom Spring microbial mat (Figure 1.6; PEs A1, A4, A6, and A14), (ii) multiple individuals of three PEs (PEs A1, A14, and A6), and (iii)

three 16S rRNA lineages (A′, A, and B′-lineage isolates), which are differently distributed along a temperature gradient. In addition to the ten new genomes sequenced, two more genomes from two new PE A1 isolates are forthcoming (Figure 2.1), and new cultures, including isolates representative of PE B′24 (PE B′-OS published genome *psaA* variant) and DV 74 (‘axenic’ strain *psaA* variant) (see Chapter 2 culture purity discussion), and a second isolate representative of PE A4, are in the early stages of purification for next-generation sequencing and culture purity testing.

Single-Locus, Multi-Locus, and Whole Genome Phylogenies

Obtaining whole genome sequences also facilitated construction of orthologous-gene genome phylogenies, which should be the best representation of the evolutionary history of a lineage (Klenk and Göker, 2010). The orthologous-gene genome phylogeny (Figure 3.1) parallels the clades formed by the isolate PEs using the entire *psaA* sequence alignment, in which the two PE A6 isolates are very different, while the multiple isolates of PE A14 and PE A1 are very similar. The phylogeny based on the PCR-amplified region of the *psaA* gene is incongruous with the whole-gene phylogeny (Figure 3.2), and RDP4 detected recombination in the PE A6-63 and PE A1-OS sequences of the *psaA* whole gene alignment. These observations may mean that the PE A6 *psaA* sequence type is a product of recombination that lumps together different ecotypes found in the mat community. I conclude that analysis using this segment of *psaA* groups the PE A6-63 and PE A6-65 isolates into one ecotype. The orthologous-gene genome phylogeny shows the

PE A6 isolates are not as closely related as the isolates of PEs A1 and A14 are to the other isolates in the same PE, and thus are likely not the same ecotype.

The effects of recombination on phylogenies are also apparent in the MLSA phylogenies of Melendrez et al. (in prep). When the sequences from the isolate genomes were added to the phylogeny, the two PE A1 isolate MLSA locus concatenations did not form a clade together (Figure 3.4A), although they did in both *psaA* and orthologous-gene genome phylogenies (Figures 3.1 and 3.2). When the *rbsK* locus, detected as recombinant in the MLSA phylogeny, was removed from the analysis, the PE A1 isolate concatenations formed a clade together (Figure 3.4B). However, the PE A14 isolate concatenations did not form a clade together when *rbsK* was omitted, even though they did form a clade in the MLSA phylogeny (Figure 3.4). This highlights the difference in the evolutionary history of a gene in different lineages. The inclusion of the *rbsK* locus in the MLSA phylogeny masked other differences between the PE A14 isolates that became apparent when it was removed from the analysis; *rbsK* acts as both a phylogenetic "signal" (vertically inherited by the PE A14 isolates) and "noise" (horizontally acquired by the PE A1 isolates).

Bearing this point in mind, it is perhaps not surprising that a genome-informed MLSA study did not produce a recombination-free phylogeny. Six genomes were analyzed to obtain the orthologous gene alignments. Genes that did not exhibit evidence of recombination or selection (Urwin and Maiden, 2003) were considered for a genome-driven MLSA of microbial mat *Synechococcus*. However, the addition of one genome not

used in the search for these genes resulted in detection of recombination within the phylogeny, and the MLSA phylogeny was incongruous with the orthologous-gene genome phylogeny. Because, as we see above, wholesale removal of recombination signals from a phylogeny removes "signal" as well as "noise", whole-genome phylogenies should be more accurate than single or even multi-locus phylogenies, as the "noise" effects of recombination are more strongly buffered by the substantial amount of phylogenetic data per organism (Klenk and Göker, 2010).

Comparative Analyses of Isolates Representative of PEs with Different Vertical Niches

Gene content comparisons of isolates of *Synechococcus* spp. have revealed many genes that could potentially cause niche-defining differences among the isolates. In Chapter 4 (Olsen et al. submitted), several differences between genomes of isolates representative of PEs that are high-light-adapted and which inhabit the middle to upper part of the mat green layer (PE A1) and PEs that are low-light-adapted, which inhabit the lower part of the mat green layer (PE A4 and A14) were found. For instance, the low-light-adapted isolates contain a photosynthetic antenna gene cassette that may enable them to use far-red light (Nowack et al., submitted; Olsen et al., submitted; Shen and Bryant, unpublished data) and transport genes specific for ferrous iron, which is only found under anoxic conditions that are more common in the lower part of the mat. These genes were shown to be expressed in the mat over the diel cycle, with the antenna gene cassette transcription peaking in the low-light periods in the early-morning and late-

afternoon, and *feoAB* transcription peaking in the late-afternoon (Figure 4.2). The high-light-adapted isolates contain a carbonic anhydrase gene that, by catalyzing the formation of CO_2 from bicarbonate, could increase fitness in peak photosynthesis periods. The light intensity in the upper part of the mat can be over 20x the intensity in the lower part of the mat (see Figure 1.8), and this high irradiance can lead to CO_2 consumption that causes pH levels to reach 9.5-10 (Jensen et al., 2011). Under these conditions, bicarbonate will be more prevalent than CO_2.

Isolates representative of populations that inhabit different parts of the upper-green layer of the microbial mat were also shown to have different transcription patterns over a diel cycle. For example, the *rpsbA* transcription pattern differed between B′-lineage, high-light-adapted A-lineage (PE A1), and low-light-adapted (PE A4 and A14) populations in the 60°C Mushroom Spring mat over a diel cycle (Figure 4.3). The transcription of this gene ends earlier in the morning and starts later in the afternoon for the upper layer B′-lineage populations (representative of populations in the very top of the mat; Figure 1.7; Chapter 4) than the high-light-adapted A-lineage populations (PE A1, predominant in the upper-middle part of the mat; Figures 1.7 and 1.9). The transcription of the gene in low-light-adapted A-lineage populations (PEs A4 and A14; Figure 1.7 and 1.9; Nowack et al., submitted) begins earliest of the three populations in the afternoon, and ends the latest in the morning. This corresponds to a difference in light exposure over the diel cycle, as populations in the upper layer (0-0.5 μm depth)

experience a longer exposure than populations in the lower layer (0.5-1 μm depth) (Figure 1.8).

Comparative Analyses of Multiple Isolates Found in Different Temperature and Light Environments

Genomes were obtained for isolates representative of (i) three 16S rRNA-defined clades, A', A, and B', which are found at different temperatures along the effluent channel (Ferris and Ward, 1997) and isolates of which have been shown to have different temperature adaptations (Allewalt et al., 2006); (ii) multiple different PEs, including PEs with distinct vertical distributions in the mat (Figure 1.7), isolates of which have been shown to have faster growth rates at different levels of light irradiance (Figure 1.9); and (iii) replicate isolates within PE A1 (two isolates) and PE A14 (three isolates). I compared the gene content, sites under positive selection, and genes under positive selection among the *Synechococcus* isolates with regards to the different ecological parameters (temperature, light, intra-, and inter-PE), summarized in Table 5.11.

Comparative Analyses of A′, A, and B′ Lineages Defined by 16S rRNA

The B′-lineage isolates have a few lineage-specific genes, such as the genes for cyanophycin synthesis and degradation found by Bhaya et al. (2007). The PE A′ isolate had one lineage-specific gene, but the A-lineage isolate genomes did not share any lineage-specific genes. There was evidence of site-specific positive selection between the different lineages defined by 16S rRNA, with more sites selected between the and A′-

lineage isolates (223-258 sites) or B′-lineage and A-lineage isolates (228-276 sites) than between the A′-lineage isolate and the A-lineage isolates (37-38 sites). There are also several instances of positive selection of a gene in the PE A′ isolate compared to the A-lineage isolates (1-4 genes), but only one instance of positive selection of a gene between the B′-lineage and the A′/A-lineage. As discussed in Chapter 5, this may be due to greater divergence between the B′-lineage and the A-lineage than between the A′-lineage and A-lineage, which would allow any slightly deleterious non-synonymous mutations, which could have "hitchhiked" to fixation alongside an advantageous mutation, to be removed by purifying selection over time (Rocha et al., 2006).

Comparative Analyses of Isolates Representative of Different PEs in the B′- and A-Lineages

There are not any gene content differences found in the present assemblies of the B′-lineage isolate genomes. However, the B'-lineage isolates do contain several genes that are not shared among all of the A'/A-lineage isolates, such as carbonic anhydrase found in the PE A1 and PE A' isolates, and beta carotene ketolase, which is found in the three PE A14 isolates, the PE A4 isolate, and the PE A' isolate. The PE A' isolate is more closely related to the A-lineage isolates than to the B'-lineage isolates (see Chapter 5, Figure 5.1 and Table 5.3), and, similar to the B'-lineage isolates, shares many genes, previously thought to be clade-specific, with both the high-light-adapted and low-light-adapted A-lineage clades. Other gene content variation among the *Synechococcus* isolates seems to center around nitrogen metabolism: genes for ammonium transporters, amino

acid transporters, and urea metabolism, among others, differ among the isolates. This variation in gene content may be similar to the local mobile gene pools observed by Boucher et al. (2011), or could indicate gene loss events happening as new ecotypes are formed and move into distinct ecological niches, such as the different niches occupied by low-light-adapted PEs as they co-exist in the lower layer of the microbial mat.

Site selection patterns within the B′-lineage revealed that the PE B'5 isolate sequences are often divergent from the PE B'-OS and PE B'12 isolates (128-131 sites, compared to 44 sites between the PE B'-OS and PE B'12 isolates) which matches the orthologous-gene genome phylogeny (Figure 5.1), and is higher than within-A-lineage site selection (37-108 sites). The number of genes under selection within the B′-lineage (3-6) is also higher than within the A-lineage (0-4). Selection patterns are similar within the A-lineage in that more closely related isolates (such as the low-light-adapted isolates, 37-64 sites) have fewer selected site differences than do more distantly related isolates (high-light-adapted isolates compared to the low-light-adapted isolates, 64-108 sites).

Comparative Analyses of Replicate Isolates Representative of PEs A1 and A14

Gene content differences were found within isolates representative of both PE A1 and PE A14, although the gene content among the isolates within a PE is more similar than the gene content shared between the two PEs (Table 5.4). There were no genes under selection between the PE A1 isolates or among the PE A14 isolates, but there was evidence of sites under selection in members of both PEs (6 sites between the PE A1

isolates, and 2-32 sites between PE A14 isolates). As discussed in Chapter 5, the high number of sites under selection in the PE A14-63 isolate compared to the other two PE A14 isolates may be the result of a sequencing artifact, although the number of sites is still lower than that for any inter-PE comparisons.

Overarching Conclusions

I assert that the hypothesis, that the *Synechococcus* spp. isolates of the same PE are both ecologically interchangeable and ecologically distinct from isolates of other PEs, cannot be rejected. While members of PE A1 or PE A14 seem to be similar in terms of gene content and patterns of selection within the genome, there are also differences between isolates of the same PE. The question is: do these differences matter ecologically? Isolates PE A14-60 and PE A14-65 are nearly identical in patterns of selection within their genomes (two site mutations differentiate PE A14-60 and PE A14-65 out of 298 positively selected sites detected), in orthologous gene sequence identity (99.99%), and in gene content, except for one small difference. The PE A14-65 isolate does not have the capability to fix nitrogen, which potentially changes the fundamental ecological niche of the organism. Perhaps members of PE A14 do not need to fix nitrogen, and this chance deletion, which removed unneeded genes from one organism, has not yet had a chance to sweep through the population. Alternatively, perhaps this isolate is representative of a “cheating” subpopulation that can obtain fixed nitrogen from surrounding individuals and gain the fitness advantage of not having to fix nitrogen.

Alternatively, perhaps our definition of an ecotype is too broad, and we need to categorize isolates from different temperatures as fundamentally different ecotypes, as they have different environmental realities, such as enzyme-substrate affinity differences or differences in the solubility of available nutrients. Continuing the sequencing efforts to obtain multiple (>3) isolates in multiple PEs may show that this kind of occurrence is rare, and that most isolates as closely related as PE A14-60 and PE A14-65 are truly ecologically interchangeable. Additional sequencing may reveal more differences between nearly identical isolates. Hopefully, as more replicate isolates of PEs such as PE A1 and PE A4 are added to the analyses, these questions will be answered.

REFERENCES CITED

REFERENCES CITED

Allewalt, J. P., Bateson, M. M., Revsbech, N. P., Slack, K., and Ward, D. M. (2006) Effects of temperature and light on growth of and photosynthesis by *Synechococcus* isolates typical of those predominating in the Octopus Spring microbial mat community of Yellowstone National Park. *Appl. Environ. Microbiol.* 72, 544-550.

Arieli, B., Padan, E., and Shahak, Y. (1991). Sulfide-induced sulfide-quinone reductase activity in thylakoids of *Oscillatoria limnetica*. *J. Biol. Chem.* 266, 104-111.

Arieli, B., Shahak, Y., Taglicht, D., Hauska, G. and Padan, E. (1994). Purification and characterization of sulfide-quinone reductase, a novel enzyme driving anoxygenic photosynthesis in *Oscillatoria limnetica*. *J. Biol. Chem.* 269, 5705-5711.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G. D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., and Zagnitko, O. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*. 9, 75

Balch, W. E., Magrum, L. J., Fox, G. E., Wolfe, R. S., and Woese, C. R. (1977) An ancient divergence among the Bacteria. *J. Mol. Evol.* 9, 305-311.

Baldo, L., Hotopp, J. C. D., Jolley, K. A., Bordenstein, S. R., Biber, S. A., Choudhury, R. R., Hayashi, C., Maiden, M. C., Tettelin, H., and Werren, J. H. (2006). Multi-locus sequence typing system for the endosymbiont *Wolbachia pipientis*. *Appl. Environ. Microbiol.* 72, 7098-7110.

Becraft, E. D. (2014). "The ecology and evolution of thermophilic *Synechococcus* species". Doctoral dissertation, Montana State University, Bozeman.

Becraft, E. D., Cohan, F. M., Kühl, M., Jensen, S. I., and Ward, D. M. (2011) Fine-scale distribution patterns of *Synechococcus* ecological diversity in microbial mats of Mushroom Spring, Yellowstone National Park. *Appl. Environ. Microbiol.* 77, 7689-7697.

Becraft, E. D., Wood, J. M., Rusch, D. B., Kühl, M., Jensen, S. I., Bryant, D. A., Roberts, D. W., Cohan, F. M., and Ward, D. M. Ecological distinctions among, and homogeneity within, putative ecotypes of *Synechococcus* inhabiting the cyanobacterial mat of Mushroom Spring, Yellowstone National Park. Submitted.

Bhaya, D., Grossman, A. R., Steunou, A.-S., Khuri, N., Cohan, F. M., Hamamura, N., Melendrez, M. C., Bateson, M. M., Ward, D. M., and Heidelberg, J. F. (2007) Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. *ISME J.* 1, 703-713.

Brock, T. D. (1978). *Thermophilic Microorganisms and Life at High Temperatures,* (New York: Springer-Verlag).

Boucher, Y., Cordero, O. X., Takemura, A., Hunt, D. E., Schliep, K., Bapteste, E., Lopez, P., Tarr, C. L., and Polz, M. F. (2011). Local mobile gene pools rapidly cross species boundaries to create endemicity within global *Vibrio cholerae* populations. *M. Bio*. 2, e00335-10.

Boucher, Y., Nesbo, C. L., and Doolittle, W. F. (2001). Microbial genomes: dealing with diversity. *Curr. Opin. Microbiol.* 4, 285-289.

Cadillo-Quiroz, H., Didelot, X., Held, N. L., Herrera, A., Darling, A., Reno, M. L., Krause, D. J., and Whitaker, R. J. (2012). Patterns of gene flow define species of thermophilic Archaea. *PLoS Biol.* 10, e1001265.

Cannon, G. C., Heinhorst, S., and Kerfeld, C. A. (2010). Carboxysomal carbonic anhydrases: Structure and role in microbial CO_2 fixation. *Biochim. Biophys. Acta.* 1804, 382-392.

Castenholz, R. W. (1969). Thermophilic blue-green algae and the thermal environment. *Bacteriol. Rev.* 33, 476.

Castenholz, R. W. (1973). Ecology of blue-green algae in hot springs. In *Botanical monographs*. eds. N. G. Carr, B. A. Whitton (Los Angeles, Calif: University of California Press), 379–414.

Castenholz, R. W. (1981). Isolation and cultivation of thermophilic cyanobacteria. In *The Prokaryotes.* (Berlin, Heidelberg: Springer), 236-246.

Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., McGrath, S. D., Wendl, M. C., Zhang, Q., Locke, D. P., Shi, X., Fulton, R. S., Ley, T. J., Wilson, R. K., Ding, L., and Mardis, E. R. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods.* 6, 677-681.

Chen, S. L., Hung, C. S., Xu, J., Reigstad, C. S., Magrini, V., Sabo, A., Blasiar, D., Bieri, T., Meyer, R. R., Ozersky, P., Armstrong, J. R., Fulton, R. S., Latreille, J. P., Spieth, J., Hooton, T. M., Mardis, E. R., Hultgren, S. J., and Gordon, J. I. (2006). Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach. *Proc. Natl. Acad. Sci. USA.* 103, 5977-5982.

Cohan, F. M. (2002). What are bacterial species?. *Ann. Rev. Microbiol.* 56, 457-487.

Cohan, F. M. and Perry, E. B. (2007) A systematics for discovering the fundamental units of bacterial diversity. *Curr. Biol.* 17, R373-R386.

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

Coleman, M. L., Sullivan, M. B., Martiny, A. C., Steglich, C., Barry, K., DeLong, E. F., and Chisholm, S. W. (2006). Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science.* 311, 1768-1770.

Connor, N., Sikorski, J., Rooney, A. P., Kopac, S., Koeppel, A. F., Burger, A., Cole, S. G., Perry, E. B., Krizanc, D., Field, N. C., Slaton, M., and Cohan, F. M. (2010). Ecology of speciation in the genus *Bacillus*. *Appl. Environ. Microbiol.* 76, 1349-1358.

Darling, A. C., Mau, B., Blattner, F. R., and Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14, 1394-1403.

de Bruyn, J. C., Boogerd, F. C., Bos, P., and Kuenen, J. G. (1990). Floating filters, a novel technique for isolation and enumeration of fastidious, acidophilic, iron-oxidizing, autotrophic bacteria. *Appl. Environ. Microbiol.* 56, 2891-2894.

Delsuc, F., Brinkmann, H., and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* 6, 361-375.

de Wit, R. and van Gemerden, H. (1987). Oxidation of sulfide to thiosulfate by *Microcoleus chtonoplastes. FEMS Microbiol. Lett.* 45, 7-13.

Didelot, X., Méric, G., Falush, D., and Darling, A. E. (2012). Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. *BMC Genomics.* 13, 256.

Dillon, J. G., Fishbain, S., Miller, S. R., Bebout, B. M., Habicht, K. S., Webb, S. M., and Stahl, D. A. (2007). High rates of sulfate reduction in a low-sulfate hot spring microbial mat are driven by a low level of diversity of sulfate-respiring microorganisms. *Appl. Environ. Microbiol*. 73, 5218-5226.

Doolittle, W. F. and Papke, R. T. (2006). Genomics and the bacterial species problem. *Genome Biol.* 7, 116.

Doolittle, W. F., Woese, C. R., Sogin, M. L., Bonen, L., and Stahl, D. (1975) Sequence studies on 16S ribosomal RNA from a blue-green alga. *J. Mol. Evol.* 4, 307-315.

Dutilh, B. E., van Noort, V., van der Heijden, R. T., Boekhout, T., Snel, B., and Huynen, M. A. (2007). Assessment of phylogenomic and orthology approaches for phylogenetic inference. *Bioinformatics*. 23, 815-824.

Ellegaard KM, Klasson L, Naslund K, Bourtzis K, and Andersson SG (2013) Comparative genomics of *Wolbachia* and the bacterial species concept. *PLoS Genet*. 9, e1003381.

Farrugia, M. A., Macomber, L., and Hausinger, R. P. (2013). Biosynthesis of the urease metallocenter. *J. Biol. Chem.* 288, 13178-13185.

Ferris, M. J., Kühl, M., Wieland, A., and Ward, D. M. (2003) Cyanobacterial ecotypes in different optical microenvironments of a 68°C hot spring mat community revealed by 16S-23S rRNA internal transcribed spacer region variation. *Appl. Environ. Microbiol.* 69, 2893-2898.

Ferris, M. J., Muyzer, G., and Ward, D. M. (1996a) Denaturing gel gradient electrophoresis profiles of 16S rRNA-defined populations inhabiting a hot spring microbial mat community. *Appl. Environ. Microbiol.* 62, 340-346.

Ferris, M. J., Ruff-Roberts, A. L., Kopczynski, E. D., Bateson, M. M., and Ward, D. M. (1996b). Enrichment culture and microscopy conceal diverse thermophilic *Synechococcus* populations in a single hot spring microbial mat habitat. *Appl. Environ. Microbiol.* 62, 1045-1050.

Ferris, M. J. and Ward, D. M. (1997) Seasonal distributions of dominant 16S rRNA-defined populations in a hot spring microbial mat examined by denaturing gel gradient electrophoresis. *Appl. Environ. Microbiol.* 63, 1375-1381.

Fraser-Liggett, C. M. (2005). Insights on biology and evolution from microbial genome sequencing. *Genome Res.* 15, 1603-1610.

Gan, F., Zhang, S., Rockwell, N. C., Martin, S. S., Lagarias, J. C., and Bryant, D. A. (2014a). Extensive remodeling of a cyanobacterial photosynthetic apparatus in far-red light. *Science.* 345, 1312-1317.

Gan, F., Shen, G., and Bryant, D. A. (2014b). Occurrence of far-red light photoacclimation (FaRLiP) in diverse cyanobacteria. *Life*. 5, 2-24.

Gee, H. (2003). Evolution: ending incongruence. *Nature*. 425, 782-782.

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., Sharpe, T., Hall, G., Shea, T. P., Sykes, S., Berlin, A. M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E. S., and Jaffe, D. B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. USA.* 108, 1513-1518.

Gomez-Garcia, M. R., Davison, M., Blain-Hartnung, M., Grossman, A. R., and Bhaya, D. (2011). Alternative pathways for phosphonate metabolism in thermophilic cyanobacteria from microbial mats. *ISME J.* 5, 141-149.

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

Heidelberg, J. F., Nelson, W. C., Schoenfeld, T., and Bhaya, D. (2009). Germ warfare in a microbial mat community: CRISPRs provide insights into the co-evolution of host and viral genomes. *PLoS ONE.* 4, 1.

Ho, Y. S. J., Burden, L. M., and Hurley, J. H. (2000). Structure of the GAF domain, a ubiquitous signaling motif and a new class of cyclic GMP receptor. *EMBO J.* 19, 5288-5299.

Hogg, J. S., Hu, F. Z., Janto, B., Boissy, R., Hayes, J., Keefe, R., Post, C., and Ehrlich, G. D. (2007). Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biol.* 8, R103.

Hughes, A. L. and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature.* 335, 167-170.

Jensen, S. I., Steunou, A. S., Bhaya, D., Kühl, M., and Grossman, A. R. (2010). *In situ* dynamics of O_2, pH and cyanobacterial transcripts associated with CCM, photosynthesis and detoxification of ROS. *ISME J*. 5, 317-328.

Johnson, J. L. (1973). Use of nucleic-acid homologies in the taxonomy of anaerobic bacteria. *Int. J. Syst. Bacteriol.* 23, 308-315.

Johnson, J. L. (1978). Taxonomy of the *Bacteroides* I. Deoxyribonucleic acid homologies among *Bacteroides fragilis* and other saccharolytic *Bacteroides* species. *Int. J. Syst. Bacteriol.* 28, 245-256.

Johnson, Z. I., Zinser, E. R., Coe, A., McNulty, N. P., Woodward, E. M. S., and Chisholm, S. W. (2006). Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science.* 311, 1737-1740.

Joseph, S. J., Hugenholtz, P., Sangwan, P., Osborne, C. A., and Janssen, P. H. (2003). Laboratory cultivation of widespread and previously uncultured soil bacteria. *Appl. Environ. Microbiol.* 69, 7210-7215.

Kagawa, Y., Nojima, H., Nukiwa, N., Ishizuka, M., Nakajima, T., Yasuhara, T., Tanaka, T., and Oshima, T. (1984). High guanine plus cytosine content in the third letter of codons of an extreme thermophile. DNA sequence of the isopropylmalate dehydrogenase of *Thermus thermophilus*. *J. Biol. Chem.* 259, 2956-2960.

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

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Mentjies, P., and Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. 28, 1647-1649.

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

Kilian, O., Steunou, A. S., Fazeli, F., Bailey, S., Bhaya, D., and Grossman, A. R. (2007). Responses of a thermophilic *Synechococcus* isolate from the microbial mat of Octopus Spring to light. *Appl. Environ. Microbiol.* 73, 4268-4278.

Klatt, C. G., Wood, J. M., Rusch, D. B., Bateson, M. M., Hamamura, N., Heidelberg, J. F., Grossman, A. R., Bhaya, D., Cohan, F. M., Kühl, M., Bryant, D. A., and Ward, D. M. (2011). Community ecology of hot spring cyanobacterial mats: predominant populations and their functional potential. *ISME J.* 5, 1262-1278.

Klenk, H. P. and Göker, M. (2010). En route to a genome-based classification of Archaea and Bacteria?. *Syst. Appl. Microbiol.* 33, 175-182.

Koeppel, A., Perry, E. B., Sikorski, J., Krizanc, D., Warner, A., Ward, D. M., Rooney, A. P., Brambilla, E., Conner, N., Ratcliff, R. M., Nevo, E., and Cohan, F. M. (2008) Identifying the fundamental units of bacterial diversity: A paradigm shift to incorporate ecology into bacterial systematics. *Proc. Natl. Acad. Sci. USA.* 105, 2504-2509.

Kopac, S., Wang, Z., Wiedenbeck, J., Sherry, J., Wu, M., and Cohan, F. M. (2014). Genomic heterogeneity and ecological speciation within one subspecies of *Bacillus subtilis*. *Appl. Environ. Microbiol.* 80, 4842-4853.

Lefébure, T. and Stanhope, M. J. (2007). Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome Biol.* 8, R71.

Li H. and Durbin R. (2009). Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics.* 25, 1754-1760.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*. 25, 2078-9.

Liu, Z., Klatt, C. G., Wood, J. M., Rusch, D. B., Ludwig, M., Wittekindt, N., Tomsho, L. P., Schuster, S. C., Ward, D. M., and Bryant, D. A. (2011). Metatranscriptomic analyses of chlorophototrophs of a hot-spring microbial mat. *ISME J.* 5, 1279-1290.

Liu, Z., Klatt, C. G., Ludwig, M., Rusch, D. B., Jensen, S. I., Kühl, M., Ward, D. M., and Bryant, D. A. (2012). '*Candidatus* Thermochlorobacter aerophilum:'an aerobic chlorophotoheterotrophic member of the phylum Chlorobi defined by metagenomics and metatranscriptomics. *ISME J.* 6, 1869-1882.

Luo, C., Walk, S. T., Gordon, D. M., Feldgarden, M., Tiedje, J. M., and Konstantinidis, K. T. (2011). Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *Proc. Natl. Acad. Sci. USA.* 108, 7200-7205.

Martin, D. P., Lemey, P., Lott, M., Moulton, V., Posada, D., and Lefeuvre, P. (2010). RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics*. 26, 2462-2463.

Mayr, E. (1942). *Systematics and the origin of species*. (New York: Columbia University Press.)

Mays, T. D., Holdeman, L. V., Moore, W. E. C., Rogosa, M., and Johnson, J. L. (1982). Taxonomy of the genus *Veillonella* Prévot. *Int. J. Syst. Bacteriol.* 32, 28-36.

Medini, D., Donati, C., Tettelin, H., Masignani, V., and Rappuoli, R. (2005). The microbial pan-genome. *Curr. Opin. Genet. Dev.* 15, 589-594.

Melendrez, M. C., Becraft, E. D., Wood, J. M., Heidelberg, J. F., Rusch, D., Cohan, F. M., and Ward, D. M. The relative importance of recombination and mutation on the evolution of *Synechococcus* species inhabiting a hot spring cyanobacterial mat. In prep.

Melendrez, M. C., Lange, R. K., Cohan, F. M., and Ward, D. M. (2011) Influence of molecular resolution on sequence-based discovery of ecological diversity among *Synechococcus* populations in an alkaline siliceous hot spring microbial mat. *Appl. Environ. Microbiol.* 77, 1359-1367.

Miller, S.R. (2003) Evidence for the adaptive evolution of the carbon fixation gene *rbcL* during diversification in temperature tolerance of a clade of hot spring cyanobacteria. *Mol. Ecol.* 12, 1237-1246.

Miller, S. R., Wingard, C. E., and Castenholz, R. W. (1998). Effects of visible light and UV radiation on photosynthesis in a population of a hot spring cyanobacterium, a *Synechococcus* sp., subjected to high-temperature stress. *Appl. Environ. Microbiol.* 64, 3893-3899.

Montogomery, L., Flesher, B., and Stahl, D. (1988). Transfer of *Bacteroides succinogenes* (Hungate) to *Fibrobacter succinogenes* gen. nov. as *Fibrobacter succinogenes* comb. nov. and description of *Fibrobacter intestinalis* sp. nov. *Int. J. Syst. Bacteriol.* 38, 430-435.

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

Murray, J. W. (2012). Sequence variation at the oxygen-evolving centre of photosystem II: a new class of 'rogue' cyanobacterial D1 proteins. *Photosyn. Res.* 110, 177-184.

Musa-Aziz, R., Chen, L. M., Pelletier, M. F., and Boron, W. F. (2009). Relative CO_2/NH_3 selectivities of AQP1, AQP4, AQP5, AmtB, and RhAG. *Proc. Natl. Acad. Sci. USA.* 106, 5406-5411.

Newton, R. J., Griffin, L. E., Bowles, K. M., Meile, C., Gifford, S., Givens, C. E., Howard, E. C., King, E., Oakley, C. A., Reisch, C. R., Rinta-Kanto, J. M., Sharma, S., Sun, S., Varaljay, V., Vila-Costa, M., Westrich, J. R., and Moran, M. A. (2010). Genome characteristics of a generalist marine bacterial lineage. *ISME J.* 4, 784-798.

Nishio, Y., Nakamura, Y., Kawarabayasi, Y., Usuda, Y., Kimura, E., Sugimoto, S., Matsui, K., Yamagishi, A., Kikuchi, H., Ikeo, K, and Gojobori, T. (2003). Comparative complete genome sequence analysis of the amino acid replacements responsible for the thermostability of *Corynebacterium efficiens*. *Genome Res.* 13, 1572-1579.

Nowack, S. (2014). "Niche character in a temporally varying environment." Doctoral dissertation, Montana State University: Bozeman, MT.

Nowack, S., Olsen, M. T., Schaible, G. A., Becraft, E. D., Shen, G., Klapper, I., Bryant, D. A., and Ward, D. M. *Synechococcus* isolates representative of putative ecotypes inhabiting different depths in the Mushroom Spring microbial mat exhibit different adaptive and acclimative responses to light. Submitted.

Olsen, M. T., Nowack, S., Wood, J. M., Becraft, E. D., LaButti, K., Lipzen, A., Martin, J., Schackwitz, W. S., Rusch, D. B., Cohan, F. M., Bryant, D. A., and Ward, D. M. Comparative genomics of *Synechococcus* isolates with different light responses and *in situ* diel transcription patterns of associated putative ecotypes in the Mushroom Spring microbial mat. Submitted.

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A. R., Xia, F., and Stevens, R. (2013). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* gkt1226.

Papke, R. T., Ramsing, N. B., Bateson, M. M., and Ward, D. M. (2003). Geographical isolation in hot spring cyanobacteria. *Environ. Microbiol.* 5, 650-659.

Peary J. A. and Castenholz, R. W. (1964). Temperature strains of a thermophilic blue-green alga. *Nature*. 202, 720-721.

Perna, N. T., Plunkett, G., Burland, V., Mau, B., Glasner, J. D., Rose, D. J., Mayhew, G. F., Evans, P. S., Gregor, J., Kirkpatrick, H. A., Pósfai, G., Hackett, J., Klink, S., Boutin, A., Shao, Y., Miller, L., Grotbeck, E. J., Davis, N. W., Lim, A., Dimalanta, E. T., Potamousis, K. D., Apodaca, J., Anatharaman, T. S., Lin, J., Yen, G., Schwartz, D. C., Welch, R. A., and Blattner, F. R. (2001). Genome sequence of enterohaemorrhagic *Escherichia coli* O157: H7. *Nature*. 409, 529-533.

Pham, V. H. and Kim, J. (2012). Cultivation of unculturable soil bacteria. *Trends Biotech.* 30, 475-484.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* 26, 1641-1650.

Rabenstein, A., Rethmeier, J. and Fischer, U. (1995). Sulphite as intermediate sulfur compound in anaerobic sulphide oxidation to thiosulphate by marine cyanobacteria. *Z. Naturforsch.* 50c, 769-774.

Ramsing, N. B., Ferris, M. J., and Ward, D. M. (2000). Highly ordered vertical structure of *Synechococcus* populations within the one-millimeter-thick photic zone of a hot spring cyanobacterial mat. *Appl. Environ. Microbiol.* 66, 1038-1049.

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

Razumov, A. S. (1932). Direct count method for bacteria in water. *Mikrobiologiya.* 1, 131-146.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J. F., Darling, A., Malfatti, S., Swan, B. K., Gies, E. A., Dodsworth, J. A., Hedlund, B. P., Tsiamis, G., Sievert, S. M., Liu, W., Eisen, J. A., Hallam, S. J., Kyrpides, N. C., Stepanauskas, R., Rubin, E. M., Hugenholtz, P., and Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature.* 499, 431-437.

Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N. A., Arellano, A., Coleman, M., Hauser, L., Hess, W. R., Johnson, Z. I., Land, M., Lindell, D., Post, A. F., Regala, W., Shah, M., Shaw, S. L., Steglich, C., Sullivan, M. B., Ting, C. S., Tolonen, A., Webb, E. A., Zinser, E. R., and Chisholm, S. W. (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature.* 424, 1042-1047.

Rocha, E. P., Smith, J. M., Hurst, L. D., Holden, M. T., Cooper, J. E., Smith, N. H., and Feil, E. J. (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J. Theor. Biol.* 239, 226-235.

Romano, C., D'Imperio, S., Woyke, T., Mavromatis, K., Lasken, R., Shock, E. L., and McDermott, T. R. (2013). Comparative genomic analysis of phylogenetically closely related *Hydrogenobaculum* sp. isolates from Yellowstone National Park. *Appl. Environ. Microbiol.* 79, 2932-2943.

Ruff-Roberts, A. L., Kuenen, J. G., and Ward, D. M. (1994). Distribution of cultivated and uncultivated cyanobacteria and *Chloroflexus*-like bacteria in hot spring microbial mats. *Appl. Environ. Microbiol.* 60, 697-704.

Sakamoto, T. and Bryant, D. A. (2001). Requirement of nickel as an essential micronutrient for the utilization of urea in the marine cyanobacterium *Synechococcus* sp. PCC 7002. *Microbes Environ*. 16, 177-184.

Schoen, C., Blom, J., Claus, H., Schramm-Glück, A., Brandt, P., Müller, T., Goesmann, A., Joseph, B., Konietzny, S., Kurzai, O., Schmitt, C., Friedrich, T., Linke, B., Vogel, U., and Frosch, M. (2008). Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in *Neisseria meningitidis*. *Proc. Natl. Acad. Sci. USA.* 105, 3473-3478.

Seki, T., Chung, C., Mikami, H., and Oshima, Y. (1978) Deoxyribonucleic acid homology and taxonomy of the genus *Bacillus*. *Int. J. of Syst. Bacteriol.* 28, 182-189.

Shih, P. M., Wu, D., Latifi, A., Axen, S. D., Fewer, D. P., Talla, E., Calteau, A., Cai, F., Tandeau de Marsac, N., Rippka, R., Herdman, M., Sivonen, K., Coursin, T., Laurent, T., Goodwin, L., Nolan, M., Davenport, K. W., Han, C. S., Rubin, E. M., Eisen, J. A., Woyke, T., Guggar, M., and Kerfeld, C. A. (2013). Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc. Natl. Acad. Sci. USA.* 110, 1053-1058.

Shrestha, P. M., Nevin, K. P., Shrestha, M., and Lovley, D. R. (2013). When is a microbial culture "pure"? Persistent cryptic contaminant escapes detection even with deep genome sequencing. *MBio*. 4, e00591-12.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol*. **7**, 1.

Simmons, S. L., DiBartolo, G., Denef, V. J., Goltsman, D. S. A., Thelen, M. P., and Banfield, J. F. (2008). Population genomic analysis of strain variation in *Leptospirillum* group II bacteria involved in acid mine drainage formation. *PLoS Biol.* 6, e177.

Simpson, G. G. (1961). *Principles of animal taxonomy*, No. 20. (New York: Columbia University Press).

Smith, K. S. and Ferry, J. G. (2000). Prokaryotic carbonic anhydrases. *FEMS Microbiol. Rev.* 24, 335-366.

Solomon, C. M., Collier, J. L., Berg, G. M., and Glibert, P. M. (2010). Role of urea in microbial metabolism in aquatic systems: a biochemical and molecular review. *Aquat. Microb. Ecol.* 59, 67-88.

Stackebrandt, E. and Ebers, J. (2006) Taxonomic parameters revisited: tarnished gold standards. *Microbiol. Today.* 33, 152-155.

Stackebrandt, E. and Goebel, B.M. (1994) Taxonomic note: a place for DNA:DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int. J. of Syst. Bacteriol.* 44, 846-849.

Staley, J. T. and Konopka, A. (1985). Measurement of *in situ* activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu. Rev. Microbiol.* 39, 321-346.

Stefanic, P., Decorosi, F., Viti, C., Petito, J., Cohan, F. M., and Mandic-Mulec, I. (2012). The quorum sensing diversity within and between ecotypes of *Bacillus subtilis*. *Environ. Microbiol.* 14, 1378-1389.

Steunou, A. S., Bhaya, D., Bateson, M. M., Melendrez, M. C., Ward, D. M., Brecht, E., Peters, J. W., Kühl, M., and Grossman, A. R. (2006). *In situ* analysis of nitrogen fixation and metabolic switching in unicellular thermophilic cyanobacteria inhabiting hot spring microbial mats. *Proc. Natl. Acad. Sci. USA.* 103, 2398-2403.

Straus, N. A. (1994). "Iron deprivation: physiology and gene regulation," In *The Molecular Biology of Cyanobacteria*, ed. D. A. Bryant (Dordrecht, The Netherlands: Kluwer Academic), 731-750.

Sugawara, M., Epstein, B., Badgley, B. D., Unno, T., Xu, L., Reese, J., Gyaneshwar, P., Denny, R., Mudge, J., Bharti, A. K., Farmer, A. D., May, G. D., Woodward, J. E., Médigue, C., Vallenet, D., Lajus, A., Rouy, Z., Martinez-Vaz, B., Tiffin, P., Young, N. D., and Sadowsky, M. J. (2013). Comparative genomics of the core and accessory genomes of 48 *Sinorhizobium* strains comprising five genospecies. *Genome Biol.* 14, R17.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725-2729.

Tank, M. and Bryant, D. A. (2015). Nutrient requirements and growth physiology of the photoheterotrophic Acidobacterium, *Chloracidobacterium thermophilum*. *Front. Microbiol.* 6, 226.

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Ciufo, S., and Li, W. (2013). Prokaryotic Genome Annotation Pipeline. In *The NCBI handbook* (Internet), 2nd ed. (Bethesda, MD: NCBI).

Taylor, B. L. and Zhulin, I. B. (1999). PAS domains: internal sensors of oxygen, redox potential, and light. *Microbiol. Molec. Biol. Rev.* 63, 479-506.

Thrash, J. C., Temperton, B., Swan, B. K., Landry, Z. C., Woyke, T., DeLong, E. F., Stepanauskas, R., and Giovannoni, S. J. (2014). Single-cell enabled comparative genomics of a deep ocean SAR11 bathytype. *ISME J.* 8, 1440-1451.

Tian, C. F., Zhou, Y. J., Zhang, Y. M., Li, Q. Q., Zhang, Y. Z., Li, D. F., Zhang, Y. Z., Li, D. F., Wang, S., Gilbert, L. B., Li, Y. R., and Chen, W. X. (2012). Comparative genomics of rhizobia nodulating soybean suggests extensive recruitment of lineage-specific genes in adaptations. *Proc. Natl. Acad. Sci. USA.* 109, 8629-8634.

Urwin, R. and Maiden, M. C. (2003). Multi-locus sequence typing: a tool for global epidemiology. *Trends Microbiol.* 11, 479-487.

van der Meer, M. T. J., Schouten, S., Bateson, M. M., Nübel, U., Wieland, A., Kühl, M., de Leeuw, J. W., Sinninghe-Damsté, J. S., and Ward. D. M. (2005). Diel variations in carbon metabolism by green nonsulfur-like bacteria in alkaline silicious hot spring microbial mats from Yellowstone National Park, U.S.A. *Appl. Environ. Microbiol*. 71, 3978-3986.

van Valen, L. (1976) Ecological species, multispecies, and oaks. *Taxon.* 25, 233-9.

Vos, M., te Beek, T. A., van Driel, M. A., Huynen, M. A., Eyre-Walker, A., and van Passel, M. W. (2013). ODoSE: a webserver for genome-wide calculation of adaptive divergence in prokaryotes. *PloS ONE*. 8, e62447.

Wang, Z., and Wu, M. (2013). A phylum-level bacterial phylogenetic marker database. *Mol. Biol. Evol*. 30, 1258-1262.

Ward, D.M., Bateson, M.M., Ferris, M.J., Kühl, M., Wieland, A., Koeppel, A., and Cohan, F.M. (2006) Cyanobacterial ecotypes in the microbial mat community of Mushroom Spring (Yellowstone National Park, Wyoming) as species-like units, linking microbial community composition, structure and function. *Philos. Trans. R. Soc. B., Biol. Sci*. 361, 1997-2008.

Ward, D. M., Castenholz, R. W., and Miller, S. R. (2012) "Cyanobacteria in geothermal habitats". In *Ecology of Cyanobacteria, 2nd Ed*, ed. B.A. Whitton (Heidleberg, GE: Springer Science+Business Media), 39-63.

Ward, D.M. and Cohan, F.M. (2005) Microbial diversity in hot spring cyanobacterial mats: pattern and prediction. In *Geothermal Biology and Geochemistry in Yellowstone National Park,* eds. W.P. Inskeep and T. McDermott (Bozeman: Thermal Biology Institute), pp. 185-202.

Ward, D. M., Cohan, F. M., Bhaya, D., Heidelberg, J. F., Kühl, M., and Grossman, A. (2008). Genomics, environmental genomics and the issue of microbial species. *Heredity.* 100, 207-219.

Ward, D. M., Ferris, M. J., Nold, S. C., and Bateson, M. M. (1998). A natural view of microbial biodiversity within hot spring cyanobacterial mat communities. *Microbiol. Mol. Biol. Rev.* 62, 1353-1370.

Ward, D. M., Ferris, M. J., Nold, S. C., Bateson, M. M., Kopczynski, E. D., and Ruff-Roberts, A. L. (1994). Species diversity in hot spring microbial mats as revealed by both molecular and enrichment culture approaches—relationship between biodiversity and community structure. In *Microbial Mats.* (Berlin, Heidelberg: Springer), 33-44.

Ward, D. M., Santegoeds, C. M., Nold, S. C., Ramsing, N. B., Ferris, M. J., and Bateson, M. M. (1997). Biodiversity within hot spring microbial mat communities: molecular monitoring of enrichment cultures. *Antonie Van Leeuwenhoek*. 71, 143-150.

Ward, D. M., Tayne, T. A., Anderson, K. L., and Bateson, M. M. (1987). Community structure and interactions among community members in hot spring cyanobacterial mats. *Symp. Soc. Gen. Microbiol.* 41, 179-210.

Ward, D. M., Weller, R., and Bateson, M. M. (1990). 16S rRNA sequences reveal uncultured inhabitants of a well-studied thermal community. *FEMS Microbiol. Rev.* 6, 105-115.

Wayne, L. G., Brenner, D. J., Colwell, R. R., Grimont, P. A. D., Kandler, O., Krichevsky, M. I., Moore, L. H., Moore, W. E. C., Murray, R. G. E., Stackbrandt, E., Starr, M. P., and Trüper, H. G. (1987). Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int. J. Syst. Bacteriol.* 4, 463-464.

West, N. J. and Scanlan, D. J. (1999). Niche-partitioning of *Prochlorococcus* populations in a stratified water column in the eastern North Atlantic Ocean. *Appl. Environ. Microbiol.* 65, 2585-2591.

West, N. J., Schönhuber, W. A., Fuller, N. J., Amann, R. I., Rippka, R., Post, A. F., and Scanlan, D. J. (2001). Closely related *Prochlorococcus* genotypes show remarkably different depth distributions in two oceanic regions as revealed by *in situ* hybridization using 16S rRNA-targeted oligonucleotides. *Microbiol.* 147, 1731-1744.

Wiedenbeck, J. and Cohan, F. M. (2011). Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiol. Rev.* 35, 957-976.

Woese, C. F. and Fox, G. E. (1977) Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci. USA.* 74, 5088-5090.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586-1591.

Yap, W. H., Zhang, Z., and Wang, Y. (1999). Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon. *J. Bacteriol.* 181, 5201-5209.

Zhang, Y. and Lin, K. (2012). A phylogenomic analysis of *Escherichia coli*/*Shigella* group: implications of genomic features associated with pathogenicity and ecological adaptation. *BMC Evol. Biol.* 12, 174.

Zhu, Y., Graham, J. E., Ludwig, M., Xiong, W., Alvey, R. M., Shen, G., and Bryant, D. A. (2010). Roles of xanthophyll carotenoids in protection against photoinhibition and oxidative stress in the cyanobacterium *Synechococcus* sp. strain PCC 7002. *Arch. Biochem. Biophys*. 504, 86-99.

APPENDICES

APPENDIX A

SUPPLEMENTARY INFORMATION FOR CHAPTER 4

Supplementary Methods

Stimulation of *Synechococcus* Isolate PE A1-MS by Bicarbonate

To assess the interconnectivity between light responses and the form of available dissolved inorganic carbon (DIC), the growth rates with respect to light intensity of the *Synechococcus* PE A1-MS isolate were measured when three different forms of DIC were provided. The genome sequence of this isolate contains an extra putative carbonic anhydrase gene that the low-light-adapted isolates do not possess. This gene is known to produce an enzyme that catalyzes the bi-directional interconversion of CO_2 and water to bicarbonate. Hence, we hypothesized that the PE A1-MS isolate would be stimulated by addition of bicarbonate under conditions of CO_2 limitation. The protocol to measure growth rates described in Nowack et al. (submitted) was followed with these additional specifications. The cultures were pre-grown at 52°C and at a scalar irradiance of 50 μmol photons $m^{-2}s^{-1}$, and the temperature during the growth rate experiments was 52°C. To alleviate CO_2 limitation in one set of conditions, 6% CO_2 in air was sparged at a rate of ~ 1 bubble/second. To achieve conditions of CO_2 limitation, the only form of DIC was diffusion from the air above the culture. Filter-sterilized $NaHCO_3$ was added (to 12 mM final concentration) to provide conditions in which bicarbonate was the primary DIC source. Bicarbonate stimulated the rate of growth of this isolate at irradiances typically observed at the depth in the mat at which PE A1 is the predominant *Synechococcus* population (Figure A3 and see Figures 3 and 8C in Becraft et al., submitted; Figure 1.7).

Supplementary Table

Supplementary Table A.1. Orthologous genes among *Synechococcus* isolates. Columns represent gene locus number, percent amino acid identity to the A1-OS reference protein (when applicable), and RAST annotated gene name for each isolate, as well as the designated Phyla-AMPHORA marker peptide name. Genes not present in all four isolates are highlighted in green, while gene absence in an isolate is highlighted in yellow. Genes highlighted in orange have >90% amino acid identity to the PE A1-OS isolate reference protein sequence. Phyla-AMPHORA marker proteins are highlighted in blue. Any manual verification of gene absence is noted in red text within the yellow highlighted cell.

This table is available at http://scholarworks.montana.edu/xmlui/handle/1/9071

Supplementary Figures

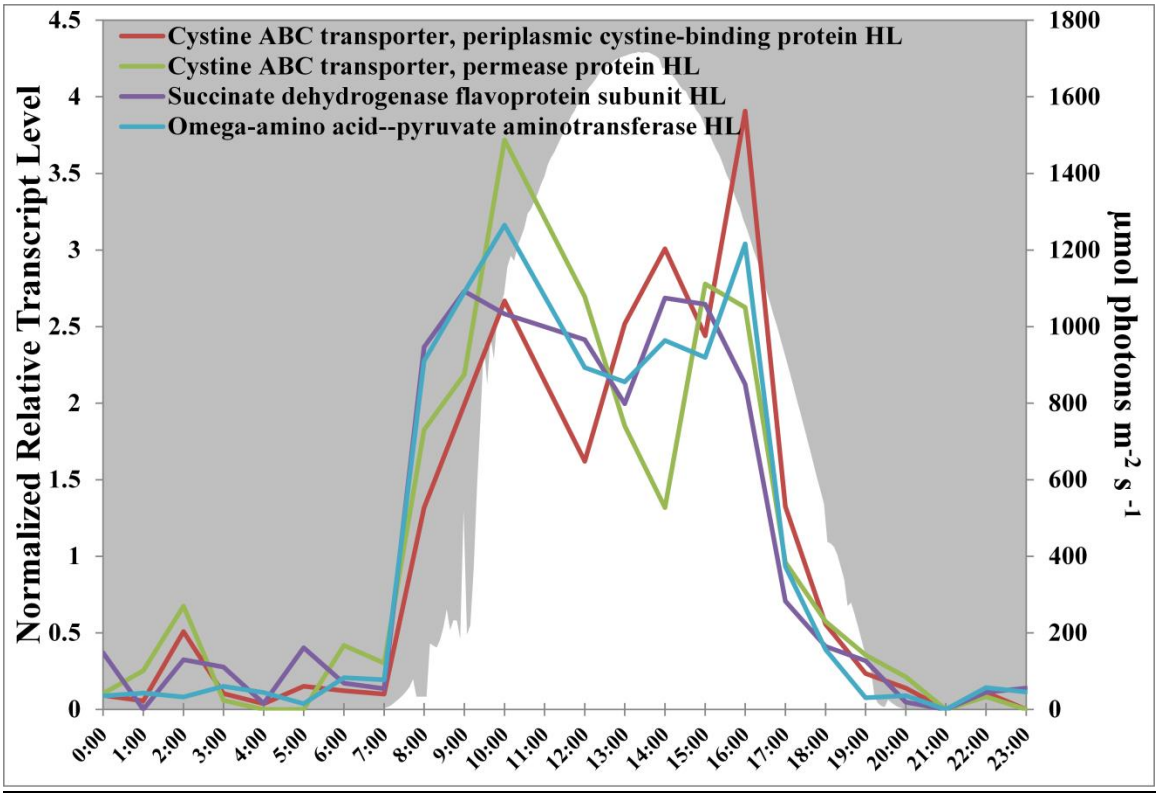


Figure A.1. PE A1-specific genes: Diel transcription patterns of PE A1-specific cystine ABC transporter periplasmic binding protein and permease protein, succinate dehydrogenase flavoprotein subunit *sdhA*, and omega-amino-acid-pyruvate transferase with downwelling irradiance (µmol photons $m^{-2}s^{-1}$) measured at Mushroom Spring from September 11-12, 2009.

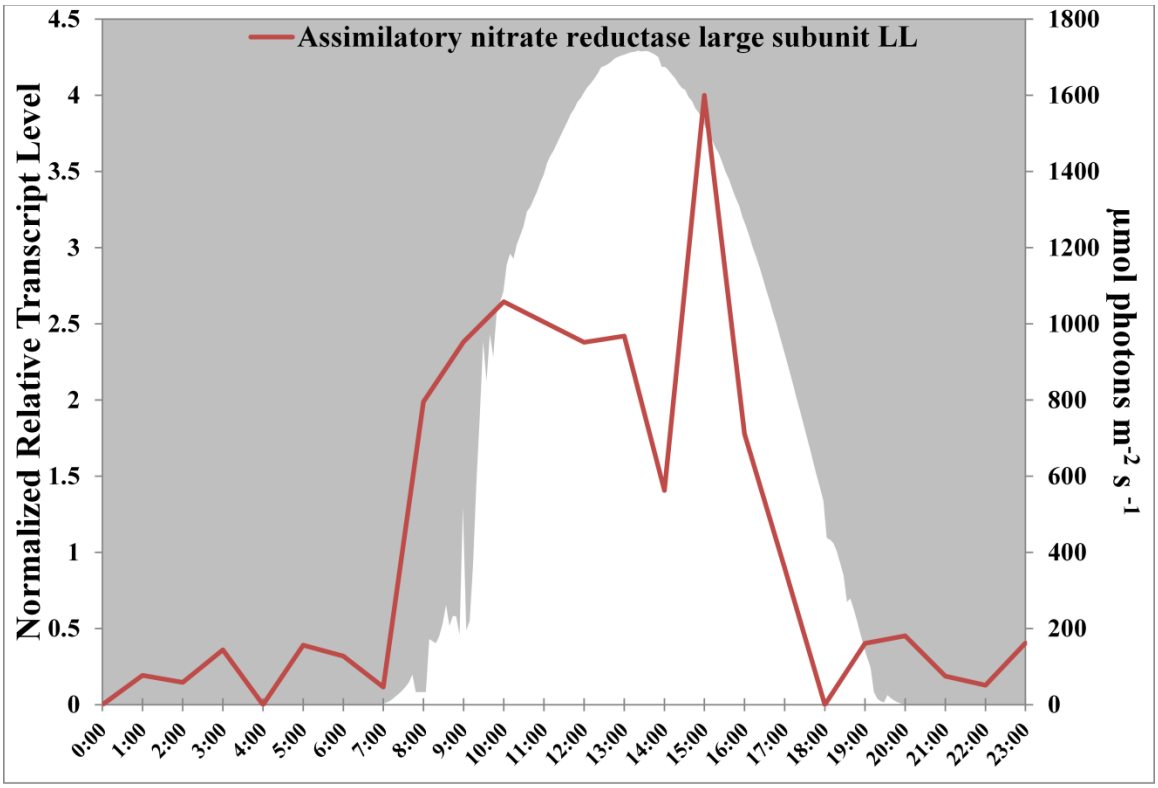


Figure A.2. Diel transcription patterns of assimilatory nitrate reductase duplicate copy in PEs A4 and A14 with downwelling irradiance (µmol photons $m^{-2}s^{-1}$) measured at Mushroom Spring from September 11-12, 2009.

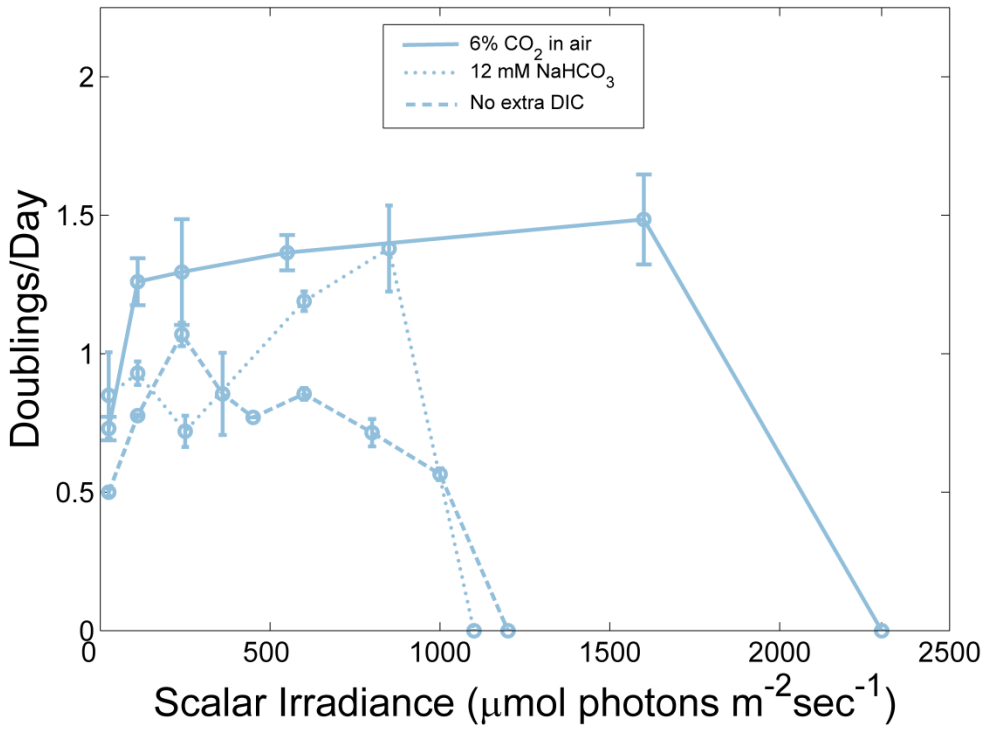


Figure A.3. Growth of the *Synechococcus* isolate of PE A1-MS strain as a function of light intensity and differences in forms of dissolved inorganic carbon supplied.

APPENDIX B

SUPPLEMENTARY INFORMATION FOR CHAPTER 5

Table B.1. Comprehensive list of sites under positive selection within the *Synechococcus* isolate genomes

Gene	Annotated Function	Gene Subsystem	Length (nt)	Selection Evidence			
				Positively selected sites (*: P>95%; **: P>99%)[1]			Mutation for each isolate[2]
CYA_0005	short-chain dehydrogenase/reductase SDR(EC:1.1.1.100)	fatty acid biosynthesis	873	**274 Q**	**1.000****	**8.890 +- 1.624**	**Q(As) -> S(A' and B's)**
				284 A	**0.989***	**8.815 +- 1.801**	**A(As) -> G/R(A' and B's/B'5)**
				289 K	**0.999****	**8.887 +- 1.630**	**K(As) -> G (A' and B's)**
CYA_0014	hyp		462	**152 L**	**0.912**	**5.077 +- 2.961**	**L(As and A') -> A(B' and B'12)/V(B'5)**
CYA_0016	gas vesicle protein W	cell buoyancy	747	**7 P**	**0.902**	**8.010 +- 2.888**	**P -> G (As and A' -> B's)**
				214 M	**0.960***	**8.470 +- 2.318**	**M(A1s/63A6)->T(A14s/A4/MA6/A')/A(B's)**
CYA_0026	O-antigen polymerase family protein	cell surface	723	**4 R**	**0.914**	**6.871 +- 3.005**	**R(As and A')->T(B' and B'12)/E(B'5)**
CYA_0051	peptide transport system permease protein	ABC transport	1101	**111 W**	**0.989***	**8.250 +- 2.069**	**W(As, A', B'5) -> S(B' and B'12)**
				123 T	**0.980***	**8.191 +- 2.167**	**T(As and A')->I(B' and B'12)/A(B'5)**
				124 E	**0.966***	**8.095 +- 2.310**	**E(As, A', B'5)->G(B' and B'12)**
CYA_0061	RNA polymerase sigma-70 factor	Transcription initiation, bacterial sigma factors	579	**128 G**	**0.918**	**2.980 +- 1.485**	**G(A1s, A4, A6s, A')->Q(A14s, B's)**
				145 H	**0.911**	**2.952 +- 1.472**	**H(A1s, A4, A6s, A')->S(A14s, B's)**
CYA_0072	Geranylgeranyl hydrogenase ChlP	Chlorophyll Biosynthesis	1218	**369 M**	**0.978***	**5.968 +- 2.918**	**M(A1s, A14s, A4, 63A6, B's)->L(MA6, A')**
CYA_0074	Protein-export membrane protein SecD (TC 3.A.5.1.1)	Protein transport	1344	**79 D**	**0.917**	**4.144 +- 2.802**	**D(3-3A1,63A14,A4,63A6)->N(65A1,60/65A14,MA6, A',B's)**
CYA_0097	transporter, major facilitator family	Transport	1377	**1 A**	**0.946**	**6.085 +- 2.214**	**A(As/A')->L(MA6)/S(B's)**
				4 T	**0.984***	**6.272 +- 1.999**	**T(A1s,63A14,63A6,A')->I(60/65A14,A4)/N(MA6)/D(B's)**
				76 W	**0.948**	**6.085 +- 2.201**	**W(As, A', B')->Q(B'5,B'12)**
CYA_0109	Cell division protein FtsZ (EC 3.4.24.-)	Bacterial Cytoskeleton	1119	**326 V**	**0.944**	**4.094 +- 2.469**	**V(As and A')->A(B's)**
CYA_0111	hyp		468	**96 R**	**0.934**	**6.355 +- 2.949**	**R(As and A')->F(B' and B'12)/R(B'5)**
CYA_0126	DNA double-strand break repair protein Mre11	Rad50-Mre11 DNA repair cluster	1302	**425 E**	**0.907**	**6.758 +- 2.946**	**E(As,A')->D(65A1)/S(B's)**
CYA_0134	hyp		318	**92 T**	**0.985***	**6.171 +- 2.486**	**T(As, A', B')->Q(B'5,B'12)**
CYA_0160	hyp		756	**233 V**	**0.979***	**6.420 +- 2.685**	**V(As, A',B'5)->A(B' and B'12)**
CYA_0164	2-octaprenyl-6-methoxyphenol hydroxylase (EC 1.14.13.-)	Plastoquinone Biosynthesis	1242	**136 H**	**0.984***	**9.975 +- 1.341**	**H(A1s, 63A14,63A6)->Y(REST)**

[1] Amino acid sites indicated to be under positive selection by the BEB method are listed, along with the probability (P) that the sites are in a positive selection site class and the post mean +/- SE for positive selection.

[2] Amino acids (and the genomes of the isolates and lineages in which they occur) are listed. (REST) indicates the amino acid is present in the remainder of the isolate genomes not listed in the entry. STOP CODON indicates a stop codon insertion in the isolate genomes listed.

Table B.1 (continued)

				165 A	**0.967***	**9.813 +- 1.807**	**A(As and A')-> R(B's)**
				385 R	**0.901**	**9.205 +- 2.859**	**R(As and A')-> A(B's)**
				406 -	**1.000****	**10.122 +- 0.678**	**STOP CODON IN B'**
CYA_0167	5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase (EC 2.1.1.14)	Methionine Biosynthesis	1662	**226 I**	**0.989***	**10.052 +- 1.137**	**I(A1s,A4, 63A6)->V(REST)**
				534 G	**0.999****	**10.140 +- 0.700**	**G(A1s, A', B's)->P(REST)**
				535 L	**0.997****	**10.125 +- 0.794**	**L->G (SIM)**
				536 K	**1.000****	**10.147 +- 0.652**	**K->T (SIM)**
				537 E	**0.987***	**10.034 +- 1.203**	**E->P(SIM)**
				538 R	**0.940**	**9.598 +- 2.275**	**R->T (SIM)**
				544 S	**0.999****	**10.145 +- 0.666**	**STOP CODON IN REST**
				545 W	**0.978***	**9.944 +- 1.497**	**STOP CODON IN REST**
				548 D	**0.968***	**9.859 +- 1.725**	**STOP CODON IN REST**
				550 F	**0.980***	**9.970 +- 1.420**	**STOP CODON IN REST**
CYA_0177	pilin domain protein	cell motility	339	**98 -**	**1.000****	**9.783 +- 0.997**	**STOP CODON(LL As)/W(A1s)/R(B's)**
CYA_0182	Phosphate transport system regulatory protein PhoU	High affinity phosphate transporter and control of PHO regulon; Phosphate metabolism	639	**118 G**	**0.962***	**6.789 +- 2.364**	**G(A1s,B's,63A6)->S(REST)**
				119 R	**0.963***	**6.795 +- 2.358**	**R(A1s,B's,63A6)->C(REST)**
				174 K	**0.950**	**6.719 +- 2.456**	**K(As and A') -> R(B's)**
CYA_0230	gas vesicle protein	cell motility	243	**74 G**	**0.978***	**8.626 +- 2.070**	**G(As and A') ->E(B' and B'5), Q(B'12)**
				76 A	**0.913**	**8.153 +- 2.781**	**A(As, A', B'12) ->E(B' and B'5)**
				78 E	**0.977***	**8.622 +- 2.084**	**E(As and A') ->STOP(B' and B'5), G(B'12)**
CYA_0233	Fibronectin/fibrinogen-binding protein	biofilm formation	1806	**228 S**	**0.998****	**9.497 +- 1.192**	**S(A1s, 63A14, A6s, A', B'5)->T(REST)**
				590 -	**1.000****	**9.513 +- 1.133**	**STOP CODON B'5/G(As, A')/A(B',B'12)**
				599 E	**0.993****	**9.457 +- 1.323**	**E(As and A')->F(B',B'12)/STOP B'5**
				600 L	**0.916**	**8.797 +- 2.619**	**L(As and A')->F(B',B'12)/STOP B'5**
CYA_0234	hyp		567	**4 F**	**0.950**	**3.059 +- 1.185**	**F(A1s and B's)-> R(rest)**
				10 K	**0.925**	**3.003 +- 1.224**	**K(A1s and B's)-> R(rest)**
				11 G	**0.972***	**3.134 +- 1.190**	**G(A1s and B's)-> S(rest)**
				18 L	**0.935**	**3.011 +- 1.184**	**L(A1s and B's)-> G(REST)**
				20 L	**0.929**	**3.006 +- 1.209**	**L(A1s and B's)-> L(REST)**
				22 L	**0.932**	**2.996 +- 1.169**	**L(A1s and B's)-> G(REST)**
				26 G	**0.968***	**3.124 +- 1.192**	**G(A1s and B's)-> L(REST)**
				46 P	**0.976***	**3.141 +- 1.181**	**P(A1s and B's)-> N(REST)**
CYA_0235	oxidoreductase, short chain dehydrogenase/reductase family	fatty acid biosynthesis	1266	**409 R**	**0.984***	**3.124 +- 1.101**	**R(A1s, A4, A')->P(A14s, A6s)/S(B')/Q(B'5,B'12)**
				412 L	**0.959***	**3.070 +- 1.136**	**L(A1s, A4, A')->V(A14s, A6s)/P(B')/I(B'5,B'12)**
CYA_0266	hyp		267	**85 L**	**0.990****	**7.231 +- 2.112**	**L(A1s, A14s, A6s, A')->C(A4, B's)**

Table B.1 (continued)

				88 Y	**0.951***	**6.972 +- 2.395**	**Y(A1s, A14s, A6s, A')->Q(A4, B's)**
				89 P	**0.982***	**7.178 +- 2.174**	**P(A1s, A14s, A6s, A')->D(A4, B's)**
CYA_0291	23S rRNA (guanosine-2'-O-) - methyltransferase rlmB (EC 2.1.1.-) ## LSU rRNA Gm2251	RNA methylation	1062	**50 -**	**1.000****	**10.431 +- 0.266**	**STOP CODON B's**
				67 -	**1.000****	**10.431 +- 0.266**	**STOP CODON B's**
				341 -	**1.000****	**10.431 +- 0.266**	**STOP CODON B's**
CYA_0312	hyp		369	**8 L**	**0.917**	**6.954 +- 3.068**	**L(A1s, 63A14, A4)->V (REST)**
CYA_0317	Phytoene synthase (EC 2.5.1.32)	Carotenoids	933	**3 S**	**0.997****	**3.868 +- 1.508**	**S(A1s, 63A14, A4, A6s)->S(REST)**
CYA_0319	ABC-2 transporter, permease protein	ABC transport	813	**4 Q**	**0.944**	**6.518 +- 2.471**	**Q(As)->R(A' and B's)**
				7 H	**0.993****	**6.778 +- 2.184**	**H(As)->T(A' and B's)**
CYA_0323	truncated cyclic nucleotide-binding protein	cell signaling	513	**4 S**	**0.913**	**5.880 +- 2.736**	**S(As and A')->P(B' and B'12)/T(B'5)**
				139 V	**0.970***	**6.213 +- 2.530**	**V(A1s, 63A14, A4, 63A6)->L(REST)**
CYA_0331	Plastid and cyanobacterial ribosome-associated protein PSRP-3	translation	297	**62 T**	**0.956***	**9.670 +- 1.996**	**T(As and A')->E(B's)**
				89 -	**1.000****	**10.067 +- 0.732**	**STOP CODON B'12**
CYA_0336	Chorismate synthase (EC 4.2.3.5) # AroG	Chorismate Synthesis; Common Pathway For Synthesis of Aromatic Compounds (DAHP synthase to chorismate)	1134	**97 S**	**1.000****	**8.239 +- 1.881**	**S(As and A')->Q(B's)**
				189 A	**0.967***	**8.021 +- 2.247**	**A(As and A')->Q(B's)**
				218 N	**0.940**	**7.822 +- 2.507**	**N(As and A')->R(B's)**
CYA_0337	Pantothenate kinase type III, CoaX-like (EC 2.7.1.33)	Coenzyme A Biosynthesis cluster	732	**80 S**	**0.999****	**8.496 +- 1.902**	**S(A1s, A4, 63A6)->G(REST)**
CYA_0350	Transcription termination protein NusB	Riboflavin synthesis cluster; Transcription factors bacterial	864	**267 L**	**0.968***	**6.526 +- 2.735**	**L(A1s, A4, 63A6)->V(REST)**
CYA_0378	hyp		543	**5 D**	**0.961***	**7.933 +- 2.564**	**D(As, A', B'5)->K(B' and B'12)**
CYA_0406	pyrimidine regulatory protein PyrR	Nucleotide synthesis	546	**177 E**	**0.985***	**8.330 +- 2.122**	**E(As and A')->I(B's)**
				179 S	**0.976***	**8.265 +- 2.224**	**S(As and A')->P(B', B'12) (no seq B'5)**
CYA_0416	hyp		393	**3 F**	**1.000****	**10.394 +- 0.361**	**F(A1s, A4, 63A6)->L(A14s, MA6, A')/H(B's)**
				16 L	**0.960***	**10.020 +- 1.882**	**L(As and A')->N(B's)**
				81 -	**1.000****	**10.397 +- 0.329**	**STOP CODON B's**
				86 -	**1.000****	**10.397 +- 0.329**	**STOP CODON B' and B'12**
CYA_0418	hyp		492	**155 V**	**0.988***	**8.040 +- 2.223**	**V(A1s, A4, 63A6)->A(A14s, MA6, A') /T(B's)**
CYA_0423	Rrf2 family transcriptional regulator	transcription	450	**94 Q**	**0.969***	**6.057 +- 2.344**	**Q(As and A')->T(B' and B'12)/S(B'5)**
				95 Q	**0.915**	**5.747 +- 2.549**	**Q(As and A')->A(B's)**
CYA_0426	hyp		558	**12 W**	**0.984***	**7.973 +- 2.405**	**W(A1s, 63A6)->C(A14s, A4, MA6, A')/L(B's)**
CYA_0429	hyp		588	**5 P**	**0.947**	**6.783 +- 2.634**	**P(As and A')->A(B' and B'5)/I(B'12)**
				6 A	**0.905**	**6.512 +- 2.844**	**A(As and A')->Q(B' and B'5)/K(B'12)**
CYA_0431	hyp		732	**201 V**	**0.999****	**8.827 +- 1.605**	**V(As and A')->D(B')/R(B'5 and B'12)**

Table B.1 (continued)

				202 P	**0.995****	**8.800 +- 1.670**	**P(As and A')->S(B')/I(B'5 and B'12)**
				204 Q	**0.973***	**8.629 +- 2.014**	**Q(As and A')->G(B')/K(B'5 and B'12)**
CYA_0454	hyp		513	**138 A**	**0.996****	**8.272 +- 2.011**	**A(As and A')->R(B' and B'12)/V(B'5)**
CYA_0467	Peptide methionine sulfoxide reductase MsrA (EC 1.8.4.11)	Peptide methionine sulfoxide reductase	672	**220 P**	**1.000****	**10.307 +- 0.475**	**P(3-3A1,B'12)->A(B'5)/E(REST)**
				221 D	**1.000****	**10.307 +- 0.479**	**D(3-3A1)->C(B'5)/P(B')/V(B'12)/S(REST)**
				222 -	**1.000****	**10.309 +- 0.458**	**STOP CODON B'12**
				223 -	**1.000****	**10.309 +- 0.458**	**STOP CODON B'**
CYA_0540	hyp		255	**77 D**	**0.988***	**8.324 +- 2.095**	**D(As and A')->N(B')/G(B'5 and B'12)**
				78 R	**0.992****	**8.353 +- 2.048**	**R(As and A')->G(B')/T(B'5 and B'12)**
CYA_0547	Molybdopterin binding motif, CinA N-terminal domain / C-terminal domain of CinA type S	NAD and NADP cofactor biosynthesis global; Riboflavin, FMN and FAD metabolism in plants; Universal stress protein family	1314	**424 H**	**0.977***	**9.505 +- 1.673**	**H(As, A', B')->C(B'5 and B'12)**
				428 -	**1.000****	**9.702 +- 1.066**	**STOP CODON B'5 and B'12**
CYA_0552	nucleic acid binding, OB-fold, tRNA/helicase-type	DNA binding; DNA repair	366	**11 G**	**0.989***	**6.707 +- 2.307**	**G(As, A', B'5)->K(B' and B'12)**
				20 A	**0.956***	**6.539 +- 2.484**	**A(As and A')->L(B's)**
CYA_0555	hyp		345	**96 K**	**0.990****	**8.291 +- 2.202**	**K(A1s, 65A14, A4, 63A6)->E(REST)**
CYA_0565	Dephospho-CoA kinase (EC 2.7.1.24)	Coenzyme A Biosynthesis	690	**111 A**	**0.959***	**9.622 +- 1.943**	**A(As and A')->E(B' and B'12)/L(B'5)**
				113 E	**0.965***	**9.669 +- 1.844**	**E(As, A', B'5)->A(B' and B'12)**
				114 A	**0.994****	**9.939 +- 1.046**	**A(As and A')->S(B' and B'12)/E(B'5)**
				116 P	**1.000****	**9.985 +- 0.823**	**P(As, A', B'5)->K(B' and B'12)**
				118 C	**0.953***	**9.566 +- 2.057**	**C(As and A')->S(B's)**
				119 Q	**0.974***	**9.751 +- 1.643**	**Q(As, A', B'5)->G(B' and B'12)**
				120 C	**0.994****	**9.937 +- 1.050**	**C(As and A')->F(B' and B'12)/K(B'5)**
				121 P	**0.925**	**9.311 +- 2.496**	**P(As and A')->S(B' and B'12)/N(B'5)**
				122 H	**0.990***	**9.899 +- 1.201**	**H(As and A')->S(B's)**
				127 P	**0.996****	**9.952 +- 0.985**	**P(As and A')->Q(B' and B'12)/S(B'5)**
				213 G	**0.992****	**9.917 +- 1.131**	**G(As and A')->Q(B' and B'12)/E(B'5)**
CYA_0567	hyp		369	**15 V**	**0.988***	**9.192 +- 1.690**	**V(A1s, A4, B's)->L(A14s, 63A6)/I(MA6, A')**
				87 V	**0.961***	**8.975 +- 2.140**	**V(A1s, A4, 63A6)->A(REST)**
				98 L	**0.979***	**9.123 +- 1.849**	**L(3-3A1, MA6, A')->V(REST)**
CYA_0569	hyp		1824	**28 -**	**1.000****	**9.347 +- 1.332**	**STOP CODON B's**
CYA_0576	hyp		909	**72 E**	**0.946**	**6.229 +- 2.660**	**E(A1s, 63A14, 63A6)->T(B's)/A(REST)**
				195 E	**0.957***	**6.269 +- 2.616**	**E(As and A')->R(B' and B'12)/Q(B'5)**
CYA_0580	Iojap protein	NAD and NADP cofactor biosynthesis global; tRNA modification Bacteria	336	**83 V**	**0.943**	**6.944 +- 2.959**	**V(A1s, 63A14, A4, 63A6, A', B's)->I(REST)**
CYA_0614	Nitrate ABC transporter, permease	Nitrate and nitrite	957	**45 G**	**0.991****	**5.190 +- 1.261**	**G(REST)->H(65A1, B'5)**

Table B.1 (continued)

	protein	assimilation					
				86 A	**0.987***	**5.175 +- 1.284**	**I(REST)->A(A1s, B'5)**
CYA_0628	Nitroreductase	Nitrogen reduction	624	**104 A**	**0.945**	**7.061 +- 2.897**	**A(A1s, 63A14, 63A6)->P(REST)**
CYA_0635	hyp		180	**1 P**	**0.945**	**7.849 +- 2.626**	**P(As, A', B'5)->Q(B' and B'12)**
				3 V	**0.966***	**8.003 +- 2.441**	**V(As and A')->G(B' and B'12)/L(B'5)**
CYA_0636	Sensory subunit of low CO_2-induced protein complex, putative	cell signaling	498	**151 E**	**0.952***	**5.748 +- 2.259**	**E(As and A')->R(B' and B'12)/D(B'5)**
CYA_0644	Photosystem II CP47 protein (PsbB)	Photosystem II	1530	**351 S**	**0.982***	**7.270 +- 2.667**	**S(A1s, 63A14, A4, A6)->G(REST)**
CYA_0646	Phosphoribosylformylglycinamidine synthase, glutamine amidotransferase subunit (EC 6.3.5.3)	De Novo Purine Biosynthesis	675	**224 A**	**0.945**	**6.165 +- 2.348**	**A(A1s, A6s, B's)->S(REST)**
				225 G	**0.988***	**6.367 +- 2.119**	**G(A1s, A6s)->S(A14s, A4, A')/A(B's)**
CYA_0649	hyp		771	**115 V**	**0.982***	**7.134 +- 2.610**	**V(A1s, 63A6, A')->M(A14s, A4, MA6)/L(B's)**
				116 I	**0.945**	**6.950 +- 2.800**	**I(A1s, 63A6, A')->M(A14s, A4, MA6)/L(B's)**
CYA_0650	hyp		282	**88 D**	**0.992****	**7.322 +- 2.149**	**D(A1s, 63A14, A6s, A')->A(60/65A14,A4)/P(B's)**
				89 K	**0.949**	**7.033 +- 2.463**	**K(A1s, 63A14, A6s, A')->E(60/65A14,A4)/D(B's)**
CYA_0651	2-phosphosulfolactate phosphatase(EC:3.1.3.71)	Coenzyme M biosynthesis; methane metabolism	726	**215 S**	**0.983***	**7.695 +- 2.224**	**S(A1s, M65, A')->Q(B',B'12)/K(B'5)/R(REST)**
				222 K	**0.987***	**7.725 +- 2.186**	**K(A1s, M65, A')->D(B',B'12)/Q(B'5)/E(REST)**
CYA_1776	hyp		351	**97 C**	**0.995****	**8.258 +- 2.146**	**C(A1s, 63A14, A4, 63A6)->Q(B's), R(REST)**
CYA_1778	Transcription antitermination protein NusG	LSU ribosomal proteins cluster; Transcription factors bacterial	669	**3 -**	**1.000****	**9.754 +- 1.033**	**STOP CODON B's**
CYA_1805	hyp		567	**174 E**	**0.920**	**6.286 +- 3.004**	**E(As, A', B'5)->T(B' and B'12)**
CYA_1860	hyp		795	**260 E**	**0.993****	**7.470 +- 2.355**	**E(As and A')->P(B' and B'12)/D(B'5)**
				263 E	**0.983***	**7.401 +- 2.435**	**E(As and A')->I(B' and B'12)/A(B'5)**
CYA_1892	two-component response regulator	cell signaling	753	**238 Q**	**0.975***	**7.867 +- 2.387**	**Q(As and A')->V(B's)**
				250 P	**0.935**	**7.603 +- 2.710**	**P(As and A')->K(B's)**
CYA_1902	Circadian clock protein KaiA	Cyanobacterial Circadian Clock;	912	**152 P**	**0.900**	**2.476 +- 1.009**	**P(As and A')->S(B')/T(B'5 and B'12)**
				(many more <0.900 prob)			**(As and A' vs B's)**
CYA_1904	CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase (EC 2.7.8.5)	Glycerolipid and Glycerophospholipid Metabolism in Bacteria	558	**178 T**	**0.959***	**8.331 +- 2.338**	**T(As and A')->H(B's)**
				181 A	**0.959***	**8.327 +- 2.344**	**A(As and A')->Q(B's)**
CYA_1905	Crossover junction endodeoxyribonuclease RuvC (EC 3.1.22.4)	RuvABC plus a hypothetical	579	**1 E**	**0.998****	**10.336 +- 0.545**	**E(As and A')->V(B' and B'5)/C(B'12)**
				2 S	**0.998****	**10.336 +- 0.545**	**S(As and A')->M(B' and B'5)/Q(B'12)**
				3 S	**0.902**	**9.418 +- 2.860**	**S(As and A')->F(B' and B'5)/I(B'12)**

Table B.1 (continued)

				4 R	**0.992****	**10.278 +- 0.912**	**R(As and A')->S(B' and B'5)/P(B'12)**
				5 A	**1.000****	**10.349 +- 0.409**	**A(A1s)->S(B' and B'5)/P(B'12)/V(REST)**
				6 T	**1.000****	**10.348 +- 0.427**	**T(A1s)->V(B' and B'5)/P(B'12)/S(REST)**
				7 R	**0.999****	**10.342 +- 0.483**	**R(A1s)->Q(B' and B'5)/L(B'12)/G(REST)**
				8 T	**0.999****	**10.342 +- 0.486**	**T(As and A')->R(B' and B'5)/A(B'12)**
				9 E	**0.990***	**10.255 +- 1.022**	**E(As and A')->K(B' and B'5)/D(B'12)**
				10 S	**0.994****	**10.295 +- 0.819**	**S(As and A')->C(B' and B'5)/R(B'12)**
				11 S	**0.969***	**10.060 +- 1.681**	**G(B'12)->S(REST)**
CYA_1969	phosphoglucomutase	Glycogen metabolism; Glycolysis/gluconeogenesis; PPP	1629	**248 L**	**0.988***	**4.711 +- 2.353**	**L(A1s, 63A14, 63A6)->G(REST)**
				249 H	**0.915**	**4.481 +- 2.465**	**H(A1s, 63A14, 63A6)->R(REST)**
CYA_1988	UDP-N-acetylglucosamine 1-carboxyvinyltransferase (EC 2.5.1.7)	Peptidoglycan Biosynthesis; UDP-N-acetylmuramate from Fructose-6-phosphate Biosynthesis	1443	**91 S**	**0.901**	**8.459 +- 2.799**	**T(As)->A(A' and B's)**
				461 A	**1.000****	**9.278 +- 1.347**	**A(As, A', B'5)-> Q(B' and B'12)**
				462 T	**0.995****	**9.240 +- 1.460**	**T(As and A')-> L(B' and B'12)/V(B'5)**
				463 P	**0.999****	**9.272 +- 1.367**	**P(As and A')-> G(B' and B'12)/S(B'5)**
				464 Q	**0.982***	**9.138 +- 1.723**	**Q(As and A')-> K(B' and B'12)/P(B'5)**
CYA_2022	DNA recombination and repair protein RecF	DNA repair, bacterial RecFOR pathway; DNA replication cluster 1	1134	**134 T**	**0.961***	**5.970 +- 2.394**	**T(As)->H(A' and B'5)/N(B' and B'12)**
				373 S	**0.947**	**5.878 +- 2.446**	**S(A's and A')->G(B' and B'12)/W(B'5)**
CYA_2023	hyp		690	**86 A**	**0.965***	**9.031 +- 2.064**	**A(A1s, A4, 63A6)->E(REST)**
				93 L	**0.964***	**9.009 +- 2.095**	**L(As and A')->H(B' and B'12)/R(B'5)**
				182 Q	**0.982***	**9.156 +- 1.791**	**Q(As and A')->N(B' and B'12)/E(B'5)**
CYA_2054	hyp		279	**34 L**	**0.994****	**7.367 +- 2.403**	**L(A1s, 60/65A14, MA6, A')>A(B's)/G(63A14,63A6)**
CYA_2055	DnaJ-class molecular chaperone CbpA	Protein chaperones	930	**21 L**	**0.969***	**3.609 +- 1.544**	**L(A1s, A14s, A4, A')->A(A6s and B's)**
CYA_2072	Glycogen branching enzyme, GH-57-type, archaeal (EC 2.4.1.18)	Glycogen metabolism	1605	**131 K**	**0.940**	**6.142 +- 3.284**	**K(As, A', B'5)->R(B' and B'12)**
CYA_2078	hyp		1506	**260 K**	**0.988***	**8.375 +- 2.094**	**K(As and A') -> P(B' and B'12)/E(B'5)**
				422 P	**0.998****	**8.443 +- 1.971**	**P(As and A') -> E(B' and B'12)/L(B'5)**
				423 W	**0.928**	**7.941 +- 2.689**	**W(As, A', B'5)->S(B' and B'12)**
				424 T	**0.981***	**8.321 +- 2.181**	**T(As and A') -> L(B' and B'12)/G(B'5)**
CYA_2104	Acetyl-coenzyme A carboxyl transferase beta chain (EC 6.4.1.2)	fatty acid biosynthesis	942	**123 L**	**0.907**	**7.974 +- 2.850**	**L(As, A', B', B'12)->A(B'5)**
				303 L	**0.993****	**8.608 +- 1.932**	**L(As and A') -> I(B' and B'12)/Q(B'5)**
				304 E	**0.977***	**8.490 +- 2.142**	**E(As and A') -> D(B' and B'12)/R(B'5)**
CYA_2116	ATP synthase F0 sector subunit b' (EC 3.6.3.14)	ATP synthesis; Oxidative phosphorylation	471	**4 S**	**0.998****	**7.993 +- 2.057**	**S(A1s, A6s, A')->P(A14s, A4, B')/Q(B'12, B'5)**

Table B.1 (continued)

CYA_2123	2,3-bisphosphoglycerate-independent phosphoglycerate mutase (EC 5.4.2.1)	Glycolysis and Gluconeogenesis	1590	**526 A**	**0.934**	**2.747 +- 1.464**	**A(A1s, MA6, A')->S(A14s, A4, 63A6, B's)**
CYA_2128	hyp		1059	**200 E**	**0.997****	**8.801 +- 1.676**	**E(As, A', B')-> T(B'5, B'12)**
				288 G	**0.925**	**8.240 +- 2.592**	**G(As, A')-> I(B', B'5)/T(B'12)**
				351 A	**0.999****	**8.816 +- 1.641**	**A(As, A')->Q(B' and B'12)/I(B'5)**
CYA_2132	hyp		744	**214 R**	**0.979***	**7.761 +- 2.387**	**R(3-3A1, B's)->V(REST)**
				241 G	**0.983***	**7.791 +- 2.353**	**G(3-3A1, B's)->L(REST)**
CYA_2151	competence protein/comEA-related protein	DNA binding; DNA repair	1761	**143 D**	**0.929**	**4.675 +- 2.676**	**D(A1s, 63A14, A4, 63A6, B', B'12)->A(REST)**
				144 Q	**0.938**	**4.692 +- 2.660**	**Q(A1s, 63A14, A4, 63A6, B', B'12)->E(REST)**
CYA_2153	hyp		480	**115 T**	**0.997****	**7.751 +- 2.128**	**T(A1s, 63A14, A4, 63A6)->P (REST)**
CYA_2161	FOG: CheY-like receiver	cell signaling	372	**122 A**	**0.946**	**7.248 +- 2.738**	**A(A1s, 63A6)->P(REST)**
CYA_2162	two-component response regulator	cell signaling	1242	**3 C**	**0.996****	**9.740 +- 1.170**	**C(A1s, 63A6, B'5)->R(REST)**
				126 G	**0.958***	**9.403 +- 2.049**	**G(REST)->S(MA6, A')**
				280 E	**0.999****	**9.773 +- 1.037**	**E(A1s, A6s, A')->G(REST)**
				308 E	**0.958***	**9.403 +- 2.049**	**E(REST)->G(MA6, A')**
				322 A	**0.996****	**9.740 +- 1.170**	**A(A1s, 63A6, B'12)->T(REST)**
CYA_2172	peptidase M16B family, nonpeptidase-like protein	Proteolysis	1311	**344 G**	**0.937**	**6.651 +- 2.615**	**G(As, A')-> L(B', B'12)/P(B'5)**
				410 N	**0.974***	**6.844 +- 2.417**	**N(As, A', B'5)->P(B' and B'12)**
CYA_2178	transporter, major facilitator family	transport	1284	**257 S**	**0.976***	**4.420 +- 2.328**	**S(As, A', B'5)->E(B' and B'12)**
CYA_2181	N-acetylmuramoyl-L-alanine amidase (EC 3.5.1.28)	Murein Hydrolases; Recycling of Peptidoglycan Amino Acids	1869	**34 L**	**0.909**	**5.166 +- 2.946**	**L(As and A')->E(B')/P(B'5 and B'12)**
CYA_2184	sensor histidine kinase(EC:2.7.3.-)	cell signaling	1332	**307 A**	**0.911**	**6.080 +- 2.987**	**A(As, A')-> G(B', B'12)/T(B'5)**
				309 G	**0.948**	**6.294 +- 2.858**	**G(As, A')-> K(B', B'12)/E(B'5)**
CYA_2199	LSU rRNA 2'-O-methyl-C2498 methyltransferase RlmM	RNA methylation	1053	**85 I**	**0.949**	**5.660 +- 2.891**	**I(As, A', B'5)->G(B' and B'12)**
CYA_2206	Pyruvate kinase (EC 2.7.1.40)	Fermentations: Mixed acid; Glycolysis and Gluconeogenesis; Pyruvate metabolism I: anaplerotic reactions, PEP	1839	**374 S**	**0.938**	**3.497 +- 1.973**	**S(As and A')->G(B' and B'12)/M(B'5)**
CYA_2229	peptidase, S1C (protease Do) family(EC:3.4.21.-)	Proteolysis	1260	**23 F**	**0.993****	**9.417 +- 1.451**	**F(As)->I(A' and B's)**
				24 V	**0.995****	**9.431 +- 1.410**	**V(As)->I(MA6))/T(A' and B's)**
				25 L	**1.000****	**9.471 +- 1.284**	**F(As)->I(A' and B's)**
CYA_2269	Peptidase family M23/M37	Proteolysis	990	**68 A**	**0.958***	**2.948 +- 0.794**	**A(As and A')->P(B's)**
				71 L	**0.904**	**2.815 +- 0.874**	**L(As and A')->E(B's)**
				81 S	**0.961***	**2.954 +- 0.788**	**S(As and A')->T(B')/P(B'5 and B'12)**
				85 P	**0.934**	**2.888 +- 0.834**	**P(As and A')->M(B' and B'5)/B(B'12)**
				99 A	**0.935**	**2.894 +- 0.833**	**A(As and A')->R(B')/L(B'5 and B'12)**

Table B.1 (continued)

				144 P	**0.981***	**3.000 +- 0.745**	**P(As and A')->G(B' and B'12)/Q(B'5)**
CYA_2292	Acetyltransferase (GNAT) family, Syn7942_2240 homolog	Transfer of CoA from donor to receiver	585	**3 P**	**0.991****	**7.972 +- 2.035**	**P(As and A')->G(B' and B'12)/C(B'5)**
				9 L	**0.986***	**7.938 +- 2.091**	**L(As and A')->Q(B' and B'12)/R(B'5)**
				13 F	**0.913**	**7.393 +- 2.711**	**F(As and A')->Q(B's)**
CYA_2305	alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen	Stress response	465	**8 A**	**0.942**	**7.170 +- 2.896**	**P(A4)->V(B' and B'12)/A(REST)**
CYA_2320	hyp		477	**2 A**	**0.922**	**7.901 +- 2.751**	**A(As, A', B'5)-> L(B' and B'12)**
				7 L	**0.970***	**8.254 +- 2.316**	**L(As, A', B'5)-> E(B' and B'12)**
				20 W	**0.989***	**8.393 +- 2.100**	**W(As, A', B'5)-> L(B' and B'12)**
CYA_2329	DNA primase (EC 2.7.7.-)	Macromolecular synthesis operon; dNTP triphosphohydrolase protein family	1893	**99 E**	**0.996****	**7.573 +- 2.175**	**E(A1s, A4, 63A6)->Q(B's)/G(REST)**
				189 A	**0.968***	**7.425 +- 2.391**	**A(A1s, A4, 63A6)->V(REST)**
				415 T	**0.978***	**7.465 +- 2.321**	**T(A1s, 63A6)->L(B's)/I(REST)**
				551 V	**0.951***	**7.297 +- 2.517**	**V(A1s, A', 63A6, B's)->L(REST)**
CYA_2332	Conserved membrane protein in copper uptake, YcnI	Metal binding	564	**70 F**	**0.941**	**3.673 +- 1.574**	**F(A1s, B's)->A(REST)**
				127 A	**0.989***	**3.834 +- 1.521**	**A(A1s, B', B'12)->L(REST)**
				163 K	**0.932**	**3.649 +- 1.594**	**K(A1s, B', B'12)->N(B'5)/H(REST)**
				180 G	**0.940**	**3.676 +- 1.583**	**G(A1s, B's)->T(REST)**
CYA_2361	NAD-dependent epimerase/dehydratase family protein		954	**264 I**	**0.944**	**7.256 +- 2.669**	**I(As and A')->L(B' and B'12)/P(B'5)**
				269 T	**0.965***	**7.380 +- 2.530**	**T(REST)->A(MA6, A')/G(B's)**
				280 A	**0.938**	**7.236 +- 2.702**	**A(As, A', B'5)->E(B' and B'12)**
CYA_2374	fibrillin	Extracellular matrix	615	**176 T**	**0.982***	**6.304 +- 2.524**	**T(A1s, 63A14, A4, 63A6, B's)->R (REST)**
				190 L	**0.980***	**6.301 +- 2.532**	**L(A1s, 63A6, B's)->I(REST)**
CYA_2378	GTP-binding protein Obg	Universal GTPases	1155	**335 A**	**0.967***	**6.761 +- 2.478**	**A(As and A')->G(B')/L(B'12)/V(B'5)**
				336 A	**0.993****	**6.886 +- 2.330**	**A(As and A')->S(B')/W(B'12)/G(B'5)**
CYA_2390	Cobalt-precorrin-6x reductase (EC 1.3.1.54)	Coenzyme B12 biosynthesis	747	**248 R**	**0.989***	**6.890 +- 2.646**	**R(A1s, 63A14, A4, 63A6, B', B'12)->G(REST)**
CYA_2394	Nitrogen-responsive response regulator NrrA	transcription	777	**35 N**	**0.941**	**7.809 +- 2.652**	**N(A1s, 63A14, 63A6)->T(REST)**
				236 A	**0.956***	**7.889 +- 2.527**	**A(As and A')->S(B's)**
				237 A	**0.971***	**7.990 +- 2.388**	**A(As, A', and B'5)->V(B' and B'12)**
CYA_2405	adenylate/guanylate cyclase	cAMP/cGMP synthesis	1014	**205 Q**	**0.917**	**5.770 +- 2.924**	**Q(As and A')->A(B' and B'12)/V(B'5)**
CYA_2408	Cobalamin synthase (EC 2.7.8.26)	Porphyrin and Chlorophyll metabolism	777	**166 D**	**0.965***	**7.235 +- 2.744**	**D(A1s, 63A14, A6s, A')->N(60/65A14,A4)/E(B's)**
CYA_2447	bacterial pre-peptidase C-terminal domain protein	transcription regulator	900	**4 K**	**0.933**	**7.854 +- 2.592**	**K(As and A')->S(B's)**
				8 L	**0.920**	**7.752 +- 2.700**	**L(As and A')->N(B's)**

Table B.1 (continued)

				20 V	**0.917**	**7.732 +- 2.722**	**V(As and A')->Q(B's)**
				30 D	**0.914**	**7.709 +- 2.747**	**D(As and A')->P(B's)**
				188 R	**0.911**	**7.687 +- 2.775**	**R(As and A')->I(B's)**
				275 A	**0.977***	**8.204 +- 2.158**	**A(A1s, 63/65A14, 63A6)->T(REST)**
CYA_2466	glycosyl transferase, family 2	Polysaccharide synthesis	996	**93 R**	**0.965***	**5.524 +- 2.603**	**R(As, A', B'5)->N(B'12 and B')**
CYA_2473	3,4-dihydroxy-2-butanone 4-phosphate synthase (EC 4.1.99.12) / GTP cyclohydrolase II (EC 3.5.4.25)	Molybdenum cofactor biosynthesis; Riboflavin, FMN and FAD metabolism	1245	**1 F**	**0.997****	**9.776 +- 1.048**	**F(As and A')->Q(B' and B'12)/M(B'5)**
				2 L	**0.986***	**9.680 +- 1.390**	**L(As and A')->V(B' and B'12)/R(B'5)**
				3 S	**0.989***	**9.699 +- 1.330**	**S(As and A')->F(B' and B'12)/Q(B'5)**
				4 L	**0.944**	**9.310 +- 2.221**	**L(As and A')->P(B' and B'12)/I(B'5)**
				6 R	**0.939**	**9.264 +- 2.300**	**R(As and A')->C(B' and B'12)/P(B'5)**
				265 T	**0.934**	**9.213 +- 2.377**	**T(As and A')->D(B' and B'12)/A(B'5)**
				410 G	**0.993****	**9.737 +- 1.199**	**G(As and A')->C(B' and B'12)/N(B'5)**
				411 K	**0.987***	**9.688 +- 1.371**	**K(As and A')->T(B' and B'12)/R(B'5)**
				412 S	**0.997****	**9.775 +- 1.054**	**S(As and A')->T(B' and B'12)/F(B'5)**
				413 E	**0.978***	**9.608 +- 1.590**	**E(As and A')->L(B' and B'12)/R(B'5)**
CYA_2495	Histidyl-tRNA synthetase (EC 6.1.1.21)	Histidine Biosynthesis	1257	**193 L**	**0.980***	**5.542 +- 2.415**	**L(A1s, 63A14, A4, 63A6)->P(REST)**
				419 R	**0.922**	**5.314 +- 2.581**	**R(As and A')->G(B' and B'5)/S(B'12)**
CYA_2511	hyp		183	**37 G**	**0.983***	**7.037 +- 2.583**	**G(A1s, MA6, A', B'5)->L(REST)**
CYA_2515	hyp		537	**173 S**	**0.993****	**8.093 +- 2.305**	**S(As, A', B'5)->G(B' and B'12)**
CYA_2518	hyp		1191	**64 Q**	**0.953***	**5.899 +- 2.307**	**Q(As and A')->A(B' and B'12)/E(B'5)**
				73 S	**0.956***	**5.916 +- 2.293**	**S(As and A')->R(B' and B'12)/K(B'5)**
				132 Q	**0.942**	**5.839 +- 2.353**	**Q(As and A')->T(B' and B'12)/A(B'5)**
				292 S	**0.978***	**6.020 +- 2.194**	**S(As, A', B'5)->R(B' and B'12)**
CYA_2560	Membrane protein PxcA, involved in light-induced proton extrusion # located in cytoplasmic membrane	photosynthetic electron transport	1380	**149 A**	**0.909**	**5.500 +- 2.666**	**A(As and A')->R(B' and B'12)/T(B'5)**
				152 R	**0.944**	**5.659 +- 2.559**	**R(As and A')->Q(B' and B'12)/V(B'5)**
CYA_2563	MscS Mechanosensitive ion channel	transport	1854	**151 A**	**0.950***	**4.623 +- 2.385**	**A(As, A', B')->R(B'5 and B'12)**
CYA_2605	Flavodoxin 1	Flavodoxin	507	**40 T**	**0.933**	**3.870 +- 1.585**	**T(A1s, A4, A')->S(REST)**
CYA_2606	IsiA [Chlorophyll *a*(b) binding protein, photosystem II CP43 protein (PsbC) homolog]	Chl *a*-binding protein; light harvesting	1032	**223 A**	**0.971***	**4.853 +- 2.589**	**A(A1s, A4, 63A6, A', B'5)->S(A14s, MA6, B', B'12)**
CYA_2608	putative iron-regulated protein A		1230	**74 G**	**0.960***	**7.159 +- 2.773**	**G(A1s, 63A14, 63A6)->R(REST)**
CYA_2614	hyp		774	**6 L**	**0.986***	**8.640 +- 2.025**	**L(A1s, 63A6, B's)->S(REST)**
				57 A	**0.986***	**8.639 +- 2.027**	**A(A1s, A4, 63A6, B's)->T(A14s, MA6, A')**
CYA_2615	hyp		660	**139 T**	**0.936**	**6.988 +- 2.828**	**T(REST)->K(MA6, A')**
CYA_2618	response regulator	cell signaling	996	**329 D**	**0.921**	**8.584 +- 2.595**	**D(A1s, A14s)->Y(A4, A6s, A')/L(B's)**
				331 S	**0.999****	**9.217 +- 1.420**	**S(A1s, A14s, B's)->P(A4, A6s, A')**
				332 Q	**1.000****	**9.219 +- 1.414**	**Q(A1s, A14s)->S(A4, A6s, A')/G(B's)**
CYA_2675	Inositol-1-monophosphatase (EC		819	**75 T**	**0.926**	**5.816 +- 2.687**	**T(A1s, 63A14, 63A6)->A(REST)**

Table B.1 (continued)

	3.1.3.25)						
				258 S	**0.953***	**5.904 +- 2.596**	**S(As and A')->Q(B' and B'12)/R(B'5)**
CYA_2678	Cobyric acid synthase (EC 6.3.5.10)	Coenzyme B12 biosynthesis	1485	**234 V**	**0.922**	**3.393 +- 2.344**	**V(REST)->T(60/65A14, A')**
CYA_2727	response regulator	cell signaling	1200	**13 E**	**0.920**	**6.048 +- 2.682**	**E(As and A')->A(B' and B'12)/L(B'5)**
				19 L	**0.908**	**5.976 +- 2.731**	**L(As, A', B', B'12)->S(B'5)**
CYA_2730	hyp		1137	**378 P**	**0.960***	**5.701 +- 2.672**	**P(As and A')->F(B' and B'12)/A(B'5)**
CYA_2769	Histidinol-phosphatase [alternative form] (EC 3.1.3.15)	Histidine metabolism	792	**261 S**	**0.997****	**9.871 +- 0.999**	**S(A1s, MA6, A')/G(A14s, 63A6)/E(A4)/H(B's)**
				262 G	**1.000****	**9.891 +- 0.908**	**G(A1s, MA6, A')/I(A14s, 63A6)/T(A4)/F(B's)**
				263 D	**1.000****	**9.893 +- 0.896**	**D(A1s, MA6, A')/E(A14s, 63A6)/S(A4 and B's)**
				264 R	**1.000****	**9.891 +- 0.908**	**R(A1s, MA6, A')/K(A14s, 63A6)/G(A4)/P(B's)**
CYA_2802	hyp		306	**4 G**	**0.997****	**8.255 +- 2.045**	**G(3-3A1, 63A6, A')->N(B's)/A(REST)**
CYA_2803	hyp		738	**49 C**	**0.922**	**8.240 +- 2.647**	**C(As and A')->V(B's)**
				71 P	**0.999****	**8.826 +- 1.679**	**P(A1s, 63A14, 63A6)->H(B's)/T(REST)**
CYA_2877	peptidase M14, carboxypeptidase A	Proteolysis	627	**148 N**	**0.957***	**5.971 +- 2.791**	**N(REST)->D(60/65A14)/A(B's)**
CYA_2882	ABC transporter	ABC transport	1008	**331 A**	**0.980***	**6.207 +- 2.642**	**A(As, A', B'5)->L(B', B'12)**
CYA_2886	response regulator receiver sensor signal transduction histidine kinase	cell signaling	3066	**77 Q**	**0.917**	**6.114 +- 2.496**	**Q(As and A')->G(B' and B'12)/R(B'5)**
				115 G	**0.996****	**6.495 +- 2.086**	**G(A1s, 63A14, 63A6, A')->C(B's)/E(REST)**

Table B.2. Genes under selection in *Synechococcus* isolates

Gene	Annotated Function	Gene Subsystem	Length (nt)	Site Selection Evidence			Gene Selection
				Positively selected sites (*: P>95%; **: P>99%)[31]			Nei & Gojobori dN/dS
CYA_0016	gas vesicle protein W	cell motility	747	**7 P**	**0.902**	**8.010 +- 2.888**	**B'5 vs 2-3B' and B'12**
				214 M	**0.960***	**8.470 +- 2.318**	
CYA_2132	hyp		744	**214 R**	**0.979***	**7.761 +- 2.387**	**B'5 vs 2-3B' and B'12**
CYA_2447	bacterial pre-peptidase C-terminal domain protein	transcription regulator	900	**4 K**	**0.933**	**7.854 +- 2.592**	**B'5 vs 2-3B' and B'12**
				8 L	**0.920**	**7.752 +- 2.700**	
				20 V	**0.917**	**7.732 +- 2.722**	
				30 D	**0.914**	**7.709 +- 2.747**	
				188 R	**0.911**	**7.687 +- 2.775**	
				275 A	**0.977***	**8.204 +- 2.158**	
CYA_0416	hyp		393	**3 F**	**1.000****	**10.394 +- 0.361**	**B'5 vs 2-3B' and B'12**
				16 L	**0.960***	**10.020 +- 1.882**	
				81 -	**1.000****	**10.397 +- 0.329**	
				86 -	**1.000****	**10.397 +- 0.329**	
CYA_0467	Peptide methionine sulfoxide reductase MsrA (EC 1.8.4.11)	Peptide methionine sulfoxide reductase	672	**220 P**	**1.000****	**10.307 +- 0.475**	**B'12 vs 2-3B' and B'5**
				221 D	**1.000****	**10.307 +- 0.479**	
				222 -	**1.000****	**10.309 +- 0.458**	
				223 -	**1.000****	**10.309 +- 0.458**	
CYA_1905	Crossover junction endodeoxyribonuclease RuvC (EC 3.1.22.4)	Homologous Recombination; Holliday junction	579	**1 E**	**0.998****	**10.336 +- 0.545**	**B'12 vs 2-3B' and B'5**
				2 S	**0.998****	**10.336 +- 0.545**	
				3 S	**0.902**	**9.418 +- 2.860**	
				4 R	**0.992****	**10.278 +- 0.912**	
				5 A	**1.000****	**10.349 +- 0.409**	
				6 T	**1.000****	**10.348 +- 0.427**	
				7 R	**0.999****	**10.342 +- 0.483**	
				8 T	**0.999****	**10.342 +- 0.486**	
				9 E	**0.990***	**10.255 +- 1.022**	
				10 S	**0.994****	**10.295 +- 0.819**	
				11 S	**0.969***	**10.060 +- 1.681**	
CYA_2495	Histidyl-tRNA synthetase (EC 6.1.1.21)	Histidine Biosynthesis	1257	**193 L**	**0.980***	**5.542 +- 2.415**	**2-3B' vs B'5**

[31] Amino acid sites indicated to be under positive selection by the BEB method are listed, along with the probability (P) that the sites are in a positive selection site class and the post mean +/- SE for positive selection.

Table B.2 (continued)

				419 R	**0.922**	**5.314 +- 2.581**	
CYA_0177	pilin domain protein	cell motility	339	**98 -**	**1.000****	**9.783 +- 0.997**	**2-3B'a vs B'12**
CYA_2162	two-component response regulator	cell signaling	1242	**3 C**	**0.996****	**9.740 +- 1.170**	**63A6 vs 60/63A14, A4, MA6, A', and 2-3B'; MA6 and A' vs 65A14**
				126 G	**0.958***	**9.403 +- 2.049**	
				280 E	**0.999****	**9.773 +- 1.037**	
				308 E	**0.958***	**9.403 +- 2.049**	
				322 A	**0.996****	**9.740 +- 1.170**	
CYA_0646	Phosphoribosylformylglycinamidine synthase, glutamine amidotransferase subunit	De Novo Purine Biosynthesis	675	**224 A**	**0.945**	**6.165 +- 2.348**	**A4 vs A1s**
				225 G	**0.988***	**6.367 +- 2.119**	
CYA_1969	phosphoglucomutase	Glycogen metabolism; Glycolysis/gluconeogenesis; PPP	1629	**248 L**	**0.988***	**4.711 +- 2.353**	**A4 vs A1s**
				249 H	**0.915**	**4.481 +- 2.465**	
CYA_2769	Histidinol-phosphatase [alternative form] (EC 3.1.3.15)	Histidine metabolism	792	**261 S**	**0.997****	**9.871 +- 0.999**	**A4 vs A14s and 63A6 (0.9 A1s vs A14s and A4/63A6; MA6 and A' vs A14s, A4, and 63A6)**
				262 G	**1.000****	**9.891 +- 0.908**	
				263 D	**1.000****	**9.893 +- 0.896**	
				264 R	**1.000****	**9.891 +- 0.908**	
CYA_0072	Geranylgeranyl hydrogenase BchP	Chlorophyll Biosynthesis	1218	**369 M**	**0.978***	**5.968 +- 2.918**	**MA6 and A' vs As**
CYA_0567	hyp		369	**15 V**	**0.988***	**9.192 +- 1.690**	**65A1 vs MA6 and A'**
				87 V	**0.961***	**8.975 +- 2.140**	
				98 L	**0.979***	**9.123 +- 1.849**	
CYA_2055	DnaJ-class molecular chaperone CbpA	Protein chaperones	930	**21 L**	**0.969***	**3.609 +- 1.544**	**63A6 vs MA6**
				241 G	**0.983***	**7.791 +- 2.353**	
CYA_2151	competence protein/comEA-related protein	DNA binding; DNA repair	1761	**143 D**	**0.929**	**4.675 +- 2.676**	**A6s vs 60/65A14; A' vs A4**
				144 Q	**0.938**	**4.692 +- 2.660**	
CYA_2329	DNA primase (EC 2.7.7.-)	Macromolecular synthesis operon; dNTP triphosphohydrolase family	1893	**99 E**	**0.996****	**7.573 +- 2.175**	**A' vs As**
				189 A	**0.968***	**7.425 +- 2.391**	
				415 T	**0.978***	**7.465 +- 2.321**	
				551 V	**0.951***	**7.297 +- 2.517**	
CYA_2618	response regulator	cell signaling	996	**329 D**	**0.921**	**8.584 +- 2.595**	**3-3A1, A14s vs A4, A6s, A'**
				331 S	**0.999****	**9.217 +- 1.420**	
				332 Q	**1.000****	**9.219 +- 1.414**	