THEORY-BASED DEMARCATION OF HOT SPRING MICROBIAL MAT SPECIES FROM LARGE DNA SEQUENCE DATASETS

by

Jason Michael Wood

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

of

Doctor of Philosophy

in

Ecology and Environmental Sciences

MONTANA STATE UNIVERSITY
Bozeman, Montana

January, 2018

©COPYRIGHT

by

Jason Michael Wood

2018

All Rights Reserved

DEDICATION

I dedicate my dissertation to my loving family. Without their love and support, this dissertation would have never been written. My grandparents, Euphemia and Harry Wood, and Patricia and Jack Nemmers will always hold a special place in my heart and I appreciate their support and love over the years. My parents, Barbara and Craig Wood instilled in me a love for science and adventure, and I appreciate their encouragement and support of my pursuits. I owe special thanks to my brother Jarrod Wood for being there for me while I finished writing this dissertation. Finally, my wife Janene Wood provided me with endless patience while I dealt with the trials of graduate school, and for that I will forever be grateful.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. David M. Ward for the immense amount of help and support that he has offered during my work on this dissertation. I would also like to thank the members of my graduate committee, Dr. William Inskeep, Dr. Matthew Fields, Dr. Ross Carlson, Dr. Donald Bryant, and Dr. William Cannon for their thoughtful advice and for spending their precious time serving on my committee. I owe special thanks to Dr. Frederick Cohan, and Dr. Danny Krizanc for hosting me at Wesleyan University while we collaborated on developing the new version of Ecotype Simulation and for their help in developing this dissertation.

Financial support for this research was provided by the National Science Foundation Frontiers in Integrative Biological Research (EF-0328698) and the Integrative Graduate Education and Research Traineeship (DGE 0654336) programs. Further support was provided by the U.S. Department of Energy Office of Biological and Environmental Research (GSP 395), originating Foundational Scientific Focus Area at the Pacific Northwest National Laboratory, under contract 112443. Additional support was provided by the Montana Agricultural Experiment Station (project 911352). Additional funding for this research was also provided by the Dean of Graduate Studies, the Department of Land Resources and Environmental Sciences, and the Thermal Biology Institute at Montana State University.

TABLE OF CONTENTS

1. INTRODUCTION 1

1.1 Molecular Methods to Detect Community Composition and Structure 2
1.1.1 16S rRNA Analyses Reveal Representative Populations 3
1.1.2 Patterning Along Gradients 3
1.1.3 Need for Higher Molecular Resolution 5
1.2 Microbial Species Theory 5
1.2.1 Stable Ecotype Model 7
1.3 Ecotype Simulation 8
1.3.1 Canonical Correspondence Analysis 9
1.3.2 Theory Dependent Analyses of Rapidly Evolving Protein-encoding Genes 9
1.4 Thesis Outline 11
1.4.1 Ecotype Simulation 2 12
1.4.2 Island Biogeography of *Synechococcus* in the American Northwest 13
1.4.3 *Synechococcus* Succession and Stability 14
1.4.4 Generality of Ecotype Concept Across Mat Taxa 15

2. ECOTYPE SIMULATION 2: AN IMPROVED ALGORITHM FOR EFFICIENTLY DEMARCATING MICROBIAL SPECIES FROM LARGE SEQUENCE DATASETS 33

2.1 Contribution of Authors and Co-Authors 33
2.2 Manuscript Information Page 34
2.3 Abstract 35
2.4 Background 36
2.5 Implementation 42
2.5.1 Changes to Ecotype Simulation 42
2.5.1.1 Phylogenetic Trees 43
2.5.1.2 Binning 44
2.5.1.3 Estimating *Omega*, *Sigma*, and *Npop* 44
2.5.1.4 The Simulation 45
2.5.1.5 Ecotype Demarcation 46
2.5.2 Evaluation of Ecotype Simulation 46
2.5.2.1 Environmental Sequences 46
2.5.2.2 Tree Building 47
2.5.2.3 Canonical Correspondence Analysis 48
2.5.2.4 Testing for Ecological Interchangeability and Ecological Distinctness 49
2.5.2.5 Analysis of ES2 Runtime 50

TABLE OF CONTENTS – CONTINUED

2.5.2.6 Other Algorithms Tested 50
2.5.2.7 Comparing Demarcated Clusters 51
2.6 Results and Discussion 51
2.6.1 Binning 51
2.6.2 Performance of ES1 and ES2 52
2.6.2.1 Putative Ecotype B′9 53
2.6.2.2 Putative Ecotype A1 53
2.6.2.3 Putative Ecotype A4 54
2.6.2.4 Putative Ecotype A14 55
2.6.3 Comparison of ES1 and ES2 PE Predictions 56
2.6.4 Comparison With Other Algorithms 57
2.6.5 Testing the Runtime of ES2 59
2.7 Conclusions 60

3. BIOGEOGRAPHY OF AMERICAN NORTHWEST HOT SPRING A/B′-LINEAGE *SYNECHOCOCCUS* POPULATIONS 72

3.1 Contribution of Authors and Co-Authors 72
3.2 Manuscript Information Page 73
3.3 Abstract 74
3.4 Introduction 75
3.5 Materials and Methods 77
3.5.1 Samples 77
3.5.2 Molecular and Phylogenetic Analyses 78
3.5.3 Putative Ecotype Demarcation 79
3.5.4 Abundance of HFSs Within Predominant PEs 79
3.5.5 Canonical Correspondence Analyses 80
3.5.6 Shared *psaA* Variants with Distance 81
3.6 Results 81
3.6.1 Phylogenetic Analysis and Abundances of High-Frequncy Sequences Within PEs 83
3.6.2 Regional Endemism 84
3.6.3 Near Endemism Among Greater Yellowstone Ecosystem Basins 84
3.6.4 Within-basin Endemism 85
3.6.5 PEs Reliably Detected in More than One Greater Yellowstone Ecosystem Basin 85
3.6.6 Possible Evidence of Long-distance Dispersal 86
3.6.7 Canonical Correspondence Analyses 86
3.7 Discussion 88

TABLE OF CONTENTS – CONTINUED

4. TEMPORAL STABILITY OF *SYNECHOCOCCUS* ECOLOGICAL POPULATIONS IN MUSHROOM SPRING AND OCTOPUS SPRING MAT COMMUNITIES 102

4.1 Introduction 102
4.2 Methods 104
4.2.1 Environmental Sampling 104
4.2.2 Vertical Dissection of Mat Samples 105
4.2.3 Molecular and Phylogenetic Analyses 105
4.2.4 Putative Ecotype Demarcation 106
4.2.5 Canonical Correspondence Analysis 107
4.3 Results and Discussion 107
4.3.1 Comparison of Winter and Summer Thermal Profiles 107
4.3.2 Comparison of Winter and Summer Vertical Profiles 110
4.3.3 Comparison of Species Composition Over Decadal Time 111
4.4 Conclusions 112

5. ECOTYPES OF OTHER PHOTOTROPHIC TAXA INHABITING THE MUSHROOM SPRING MICROBIAL MAT 121

5.1 Introduction 121
5.1.1 Other Predominant Mushroom Spring Phototrophic Taxa 122
5.1.1.1 *Roseiflexus* spp. 122
5.1.1.2 *Chloroflexus* spp. 123
5.1.1.3 '*Candidatus* Roseilinea' spp. 124
5.1.1.4 *Chloracidobacterium* spp. 124
5.1.1.5 '*Candidatus* Thermochlorobacter' spp. 125
5.1.2 Study Design 125
5.2 Methods 126
5.2.1 Environmental Samples 126
5.2.2 Molecular Methods and Sequence Analysis 126
5.2.3 Putative Ecotype Demarcation 128
5.2.4 Canonical Correspondence Analysis 128
5.3 Results and Discussion 129
5.3.1 *Synechococcus* 129
5.3.2 *Roseiflexus* 130
5.3.3 *Chloroflexus* 131
5.3.4 '*Candidatus* Roseilinea' 131
5.3.5 *Chloracidobacterium* 132
5.3.6 '*Candidatus* Thermochlorobacter' 133

TABLE OF CONTENTS – CONTINUED

5.3.7 Quantitative Distributions of Taxa Along Environmental Gradients 134
5.4 Conclusions 135

6. CONCLUSIONS 144

REFERENCES CITED 149

APPENDICES 162

APPENDIX A : Chapter 2 Supplemental Information 163
A.1 Analysis of B′ *Synechococcus* Lineage 164
A.1.1 Methods 164
A.1.2 Results and Discussion 164
A.2 Analysis of Parameter Estimation Changes 165
A.2.1 Methods 165
A.2.2 Results and Discussion 166
A.3 Canonical Correspondence Analyses With Demarcation From Other Algorithms 166
A.3.1 Methods 166
A.3.2 Results and Discussion 167
APPENDIX B : Chapter 4 Supplemental Information 176
APPENDIX C : Chapter 5 Supplemental Information 188

LIST OF TABLES

Table Page

2.1 Comparison of Clustering Results for Different Trees and Algorithms Using the Variation of Information Statistic 71

3.1 Physical Parameters for Hot Springs Sampled Across the Northwestern United States 99

3.2 Chemical Parameters for Hot Springs Sampled Across the Northwestern United States 100

3.3 Constrained Variability and Significance of Parameters in Canonical Correspondence Analyses 101

A.1 Comparison of Parameter Estimation Methods from Ecotype Simulation 1 and 2 175

B.1 Summary of Samples and *psaA* Sequence Amplicons Collected From Mushroom Spring Summer and Winter Thermal Gradients 183

B.2 Summary of Samples and *psaA* Sequence Amplicons Collected From Mushroom Spring and Octopus Spring Between 1989 and 2011 183

B.3 Summary of Mid-Winter Samples and *psaA* Sequence Amplicons Collected From Mushroom Spring 65 °C Mat 184

B.4 Summary of Mid-Winter Samples and *psaA* Sequence Amplicons Collected From Mushroom Spring 62.5 °C Mat 185

B.5 Summary of Mid-Summer Samples and *psaA* Sequence Amplicons Collected From Mushroom Spring 65 °C Mat 186

B.6 Summary of Mid-Summer Samples and *psaA* Sequence Amplicons Collected From Mushroom Spring 62.5 °C Mat 187

C.1 Summary of Samples and Gene Sequence Amplicons From Predominant Putative Ecotypes of Phototrophic Taxa Collected Along the Mushroom Spring Thermal Gradient 189

C.2 Summary of Samples and Gene Sequence Amplicons From Predominant Putative Ecotypes of Phototrophic Taxa Collected Along the 62.5 °C Mushroom Spring Vertical Gradient 190

LIST OF TABLES – CONTINUED

Table Page

C.3 Summary of Primers Used to Study Phototrophic Taxa of Mushroom Spring..... 191

LIST OF FIGURES

Figure Page

1.1 Octopus Spring and Mushroom Spring Microbial Mats and Microorganisms 17

1.2 Microsensor Measurements of Scalar Irradiance and O_2 Concentration in Hot Spring Microbial Mats 18

1.3 16S rRNA-based Phylogenetic Trees of Cyanobacteria and Chloroflexi Sequences Sampled from Octopus Spring Mat 19

1.4 Denaturing Gradient Gel Electrophoresis of Octopus Spring 16S rRNA Gene Segments Along Summer and Winter Thermal Gradients 20

1.5 Growth Rates of Representative Octopus Spring *Synechococcus* Isolates at Various Temperatures 21

1.6 Denaturing Gradient Gel Electrophoresis of Mushroom Spring 16S rRNA Gene Segments Relative to Depth in 61 °C Mat 22

1.7 Phylogenetic Tree of *Synechococcus* Internal Transcribed Spacer Region Sequences Retrieved from Mushroom Spring 68 °C Mat 23

1.8 The Stable Ecotype Model of Microbial Species and Speciation 24

1.9 Effect of Evolutionary Divergence of Genetic Loci on Putative Ecotype Demarcation by Ecotype Simulation 25

1.10 Vertical Distribution of Predominant *Synechococcus* Ecotypes in Mushroom Spring 60 °C to 63 °C Mat 26

1.11 Canonical Correspondence Analyses of *Synechococcus* Sequence Variants Relative to Depth and Temperature 27

1.12 Growth Rates of Representative Mushroom Spring *Synechococcus* Isolates at Various Light Intensities 28

1.13 Phylogenetic Tree of Cyanobacterial 16S rRNA Sequences Collected from Hot Springs Across the World 29

1.14 Hierarchical Cluster Analysis of Chemistry of Hot Spring Water Collected Across the World 30

LIST OF FIGURES – CONTINUED

Figure Page

1.15 Predominant Taxa Inhabiting the Ocotpus Spring and Mushroom Spring Microbial Mats Based on Metagenomic Assembly Oligonuclotide Usage Frequency 31

1.16 Temperature Distribution of Cyanobacteria, Chloroflexi, and Chloracidobacteria 16S rRNA Sequence Segments Collected from White Creek and Rabbit Creek..... 32

2.1 Dependency Flowchart Comparison of Ecotype Simulation 1 and Ecotype Simulation 2 61

2.2 Sequence Identity Criterion Binning of a Hypothetical Lineage 62

2.3 Comparison of Ecotype Simulation Binning Results of A-like *Synechococcus psaA* Sequence Variants 63

2.4 Screenshot of the Ecotype Simulation 2 Graphical User Interface 64

2.5 Neighbor-joining Phylogeny of A/B′-lineage *Synechococcus psaA* Segments with Comparative Ecotype Demarcations 65

2.6 Maximum-likelihood Phylogeny of A/B′-lineage *Synechococcus psaA* Segments with Comparative Ecotype Demarcations 66

2.7 Canonical Correspondence Analysis Highlighting *psaA* Variants of Predominant *Synechococcus* Putative Ecotypes Occuring >50 Times..... 67

2.8 Canonical Correspondence Analysis Highlighting *psaA* Variants of Predominant *Synechococcus* Putative Ecotypes Occuring >10 Times..... 68

2.9 *psaA* Sequence Abundance of A-like *Synechococcus* Putative Ecotypes Demarcated by Ecotype Simulation 2 69

2.10 Rarefaction and Runtime Curves of A-like *Synechococcus* Putative Ecotypes Demarcated by Ecotype Simulation 2 70

3.1 Geographic Sampling Locations of Hot Springs Studied in the Northwest United States 91

3.2 Percentage of Identical Shared Variants Relative to Separation Among Hot Springs in the American Northwest 92

LIST OF FIGURES – CONTINUED

Figure Page

3.3 A/B′-lineage *Synechococcus psaA* Segment Phylogenies with Ecotype Simulation 2 Putative Ecotype Demarcations 93

3.4 Abundance of A/B′-lineage *Synechococcus* High-Frequency *psaA* Variants of Predominant Putative Ecotypes in Hot Springs of Yellowstone 94

3.5 Abundance of A/B′-lineage *Synechococcus* High-Frequency *psaA* Variants of Predominant Putative Ecotypes in Hot Springs Outside of Yellowstone 95

3.6 Canonical Correspondence Analyses Along Significant Environmental Gradients of Predominant A/B′-lineage *Synechococcus psaA* Variants, Including Oregon Samples 96

3.7 Canonical Correspondence Analyses Along Significant Environmental Gradients of Predominant A/B′-lineage *Synechococcus psaA* Variants, Excluding Oregon Samples 97

3.8 Canonical Correspondence Analyses Along a pH Gradient of Predominant A/B′-lineage *Synechococcus psaA* Variants, Excluding Oregon Samples 98

4.1 Mid-Summer Abundance of A/B′-lineage *Synechococcus* High-Frequency *psaA* Variants of Predominant Putative Ecotypes in Replicated 62.5 °C Samples of Mushroom Spring Mat 113

4.2 Winter Abundance of A/B′-lineage *Synechococcus* High-Frequency *psaA* Variants of Predominant Putative Ecotypes Along the Thermal Gradient of the Mushroom Spring Effluent Channel 114

4.3 Summer Abundance of A/B′-lineage *Synechococcus* High-Frequency *psaA* Variants of Predominant Putative Ecotypes Along the Thermal Gradient of the Mushroom Spring Effluent Channel 115

LIST OF FIGURES – CONTINUED

Figure Page

4.4 Canonical Correspondence Analyses of A/B′-lineage *Synechococcus psaA* Variants from Predominant Putative Ecotypes Along Mushroom Spring Winter and Summer Temperature Gradients 116

4.5 Canonical Correspondence Analyses of A/B′-lineage *Synechococcus psaA* Variants from Predominant Putative Ecotypes from 62.5 °C to 65 °C Mushroom Spring Veritical Profiles 117

4.6 *psaA* Sequence Abundance of A/B′-lineage *Synechococcus* Putative Ecotypes in Winter and Summer Samples Collected at 62.5 °C and 65 °C 118

4.7 Abundance of A/B′-lineage *Synechococcus* High-Frequency *psaA* Variants of Predominant Putative Ecotypes in Mushroom Spring Samples Collected between 1996 and 2014 119

4.8 Abundance of A/B′-lineage *Synechococcus* High-Frequency *psaA* Variants of Predominant Putative Ecotypes in Octopus Spring Samples Collected between 1989 and 2005 120

5.1 *Synechococcus*-Highlighted Canonical Correspondence Analyses of *psaA* Variants from Predominant Putative Ecotypes Along Mushroom Spring Thermal and Vertical Gradients 137

5.2 *Roseiflexus*-Highlighted Canonical Correspondence Analyses of *pufLM* Variants from Predominant Putative Ecotypes Along Mushroom Spring Thermal and Vertical Gradients 138

5.3 *Chloroflexus*-Highlighted Canonical Correspondence Analyses of *pufL* Variants from Predominant Putative Ecotypes Along Mushroom Spring Thermal and Vertical Gradients 139

5.4 '*Candidatus* Roseilinea'-Highlighted Canonical Correspondence Analyses of *pufL* Variants from Predominant Putative Ecotypes Along Mushroom Spring Thermal and Vertical Gradients 140

LIST OF FIGURES – CONTINUED

Figure Page

5.5 *Chloracidobacterium*-Highlighted Canonical Correspondence Analyses of *pscA* Variants from Predominant Putative Ecotypes Along Mushroom Spring Thermal and Vertical Gradients 141

5.6 '*Candidatus* Thermochlorobacter'-Highlighted Canonical Correspondence Analyses of *pscA* Variants from Predominant Putative Ecotypes Along Mushroom Spring Thermal and Vertical Gradients 142

5.7 Canonical Correspondence Analyses Along Thermal and Vertical Gradients of Predominant Putative Ecotypes of Mushroom Spring Mat Phototrophic Taxa 143

A.1 Comparison of Ecotype Simulation Binning Results of B′-like *Synechococcus psaA* Sequence Variants 168

A.2 Neighbor-joining Phylogeny of B′-lineage *Synechococcus psaA* Segments with Comparative Ecotype Demarcations 169

A.3 Maximum-likelihood Phylogeny of B′-lineage *Synechococcus psaA* Segments with Comparative Ecotype Demarcations 170

A.4 *psaA* Sequence Abundance of B′-like *Synechococcus* Putative Ecotypes Demarcated by Ecotype Simulation 2 171

A.5 Rarefaction and Runtime Curves of B′-like *Synechococcus* Putative Ecotypes Demarcated by Ecotype Simulation 2 172

A.6 Canonical Correspondence Analysis Highlighting *psaA* Variants of Predominant *Synechococcus* Putative Ecotypes Demarcated Using Maximum-Likelihood Poisson Tree Processes 173

A.7 Canonical Correspondence Analysis Highlighting *psaA* Variants of Predominant *Synechococcus* Putative Ecotypes Demarcated Using Bayesian Poisson Tree Processes 174

B.1 Photographs of Mushroom Spring Sampling Locations in Winter and Summer 177

B.2 Water Temperature at Winter and Summer Mushroom Spring Sampling Locations 178

LIST OF FIGURES – CONTINUED

Figure Page

B.3 Downwelling Irradiance at Winter and Summer Mushroom Spring Mats 179

B.4 Mid-Winter Abundance of A/B′-lineage *Synechococcus* High-Frequency *psaA* Variants of Predominant Putative Ecotypes in Replicated 65 °C Samples of Mushroom Spring Mat 180

B.5 Mid-Winter Abundance of A/B′-lineage *Synechococcus* High-Frequency *psaA* Variants of Predominant Putative Ecotypes in Replicated 62.5 °C Samples of Mushroom Spring Mat 181

B.6 Mid-Summer Abundance of A/B′-lineage *Synechococcus* High-Frequency *psaA* Variants of Predominant Putative Ecotypes in Replicated 65 °C Samples of Mushroom Spring Mat 182

NOMENCLATURE

CCA	Canonical Correspondence Analysis
DNA / rRNA	Deoxyribonucleic Acid / Ribosomal Ribonucleic Acid
EF / PS / D	Ecotype Formation Event / Periodic Selection Event / Drift Event
ES1	Ecotype Simulation, version 1
ES2	Ecotype Simulation, version 2
HFS	High-Frequency Sequence, includes the minimum sequence count cutoff in subscript, if applicable
ITS	Internal Transcribed Spacer
MS	Mushroom Spring
OS	Octopus Spring
PE	Putative Ecotype
PTP / bPTP	Poisson Tree Processes / Bayesian Poisson Tree Processes
YNP	Yellowstone National Park
npop	The estimated number of ecotype populations
omega	Rate of net ecotype formation
sigma	Rate of periodic selection

ABSTRACT

The identification of closely related, ecologically distinct populations within microbial communities is paramount to understanding the structure and function of these communities. Microbial systematists have long used differences in DNA sequence relatedness to categorize the observed diversity in a community of microbes without including ecological theory to identify whether or not the identified groups are ecologically distinct. Ecotype Simulation, an evolutionary simulation algorithm based on the Stable Ecotype Model of microbial species and speciation, has been used successfully to study the diversification of thermophilic A/B′-lineage *Synechococcus* living in the effluent channels of alkaline-siliceous hot springs in Yellowstone National Park. However, Ecotype Simulation is an extremely slow program that is unable to handle the quantity of data produced by modern DNA sequencing technologies. I introduce a new version of this algorithm, called Ecotype Simulation 2, that permits the rapid analyses of microbial diversity from very large DNA sequence datasets. Results from this new version of the Ecotype Simulation algorithm compare favorably with results from the old version, but with analyses performed much more quickly on a much greater quantity of sequences sampled. The new algorithm was used to analyze three datasets. First, the biogeography of thermophilic A/B′-lineage *Synechococcus* living in hot springs of the American Northwest was analyzed. Results suggested a surprising amount of endemism among springs sampled, as well as implications for adaptations to physical and chemical environmental features not seen before. Second, Ecotype Simulation 2 was used to study the history of change in *Synechococcus* populations, seasonally (winter to summer) and over a twenty-five year period. Results suggested changes in population abundances and distribution seasonally, but stability in population genetic structure over many years. Finally, Ecotype Simulation 2 was used to study the populations of other predominant phototrophic microbes living along temperature and depth gradients in the same microbial mat community. Results suggested that the algorithm and the Stable Ecotype Model can successfully predict ecological diversity within all predominant mat taxa. Ecotype Simulation 2 provides the means for other microbiologists to base their understanding of the communities they study on evolutionary and ecological principals.

CHAPTER ONE

INTRODUCTION

Octopus Spring and Mushroom Spring (Figure 1.1 A and D) are alkaline-siliceous hot springs in Yellowstone National Park, WY, USA that harbor mat-forming microbial communities that have been studied for about 50 years (Brock, 1967; Ward et al., 2012). These hot springs provide relatively simple, stable systems with well-defined environmental gradients (e.g., the water temperature decreases with distance from the source, and light quality changes and availability decreases with depth within the mat) that provide an ideal model for studying fundamental principals of microbial ecology. Thermophilic unicellular cyanobacteria in the genus *Synechococcus* and filamentous Chloroflexi in the genus *Roseiflexus* (see the green arrow pointing to rod-shaped bacteria and the red arrow pointing to filament-shaped bacteria in Figure 1.1E, respectively), in addition to other organisms, form a highly laminated microbial mat (Figure 1.1B) in the effluent channels of these springs. The green-colored upper layer of the mat is illuminated by light during the day while the orange-colored lower layers see little to no light. The makeup of the microbial mat is very amenable to studies using microsensors that are able to demonstrate spatial and temporal changes in light or resource availability (Figure 1.2). Discrete populations of *Synechococcus* spp. can be detected with depth in the upper 1 mm of the mat by observing their autofluorescence (compare the brightness of cells near the top of the mat green layer with those below in Figure 1.1C).

Molecular methods for understanding the composition, structure, and function of these microbial mat communities have undergone rapid development during the past

half-century. Many of the more sophisticated 'omics' technologies have been used in these model communities. I have collaborated in studies using comparative genomics (van der Meer et al., 2010; Olsen et al., 2015), metagenomics (Klatt et al., 2011; Thiel et al., 2016), metatranscriptomics (Liu et al., 2011), meta-metabolomics (Kim et al., 2015), and metaproteomics (unpublished). Since these studies have mainly involved mat community function and this thesis is mainly about species within the mat, I will instead focus on molecular methods for understanding mat community composition and structure.

Molecular Methods to Detect Community Composition and Structure

Understanding the composition and structure of a community requires the identification of closely related, ecologically distinct populations (ecotypes or ecological species) within the community. However, detecting these ecotypes is complicated by the small physical sizes of their microhabitats and the lack of easily identifiable phenotypic differences amongst closely related, but ecologically distinct populations (Becraft et al., 2011). Further complicating the issue of classifying members of a microbial community is the fact that the vast majority of microbes have eluded cultivation in a laboratory environment (Rappé and Giovannoni, 2003).

Originally, different microbes were classified as a species and named by their observable phenotypic differences, including metabolism, cell composition, and shape. Advances in molecular techniques allowed microbiologists to attempt to add a phylogenetic basis to these classifications by utilizing different cutoffs, such as 70% whole-genome DNA-DNA hybridization (Wayne et al., 1987), 95-96% amino acid similarity (Konstantinidis and Tiedje, 2005), or about 1% difference in 16S rRNA (Stackebrandt and Ebers, 2006). These approaches were calibrated to match the traditional phenotype-based classification, so any diversification not evident in the

phenotype of the population could be hidden.

16S rRNA Analyses Reveal Representative Populations

The 16S rRNA locus has proven useful for detecting the composition and structure of microbial communities such as mats in Octopus Spring and Mushroom Spring. Due to its rapid growth in standard media inoculated with mat samples, it was believed that *Synechococcus lividus* was the predominant population of *Synechococcus* in the microbial mats of springs such as these. Ward et al. (1998) used 16S rRNA sequencing to demonstrate that although *S. lividus* can be isolated from the system, the predominant *Synechococcus* populations in the Octopus Spring mat were genetically distinct from *S. lividus* (compare the green arrow pointing at 'OS C1 Isolate' and the green arrow with bar pointing at the A/B′-lineage in Figure 1.3A). Similarly, based on its ability to grow in culture, *Chloroflexus aurantiacus* was thought to be the predominant type of Chloroflexi in the Octopus Spring mat. However, Ward et al. (1998) showed that type-C 16S rRNA sequences, later shown to be contributed by another type of Chloroflexi, *Roseiflexus* spp., dominated (compare the red arrow pointing at *Chloroflexus aurantiacus* and the red arrow with bar pointing at the OS Type C, or *Roseiflexus*, lineage). The results of these 16S rRNA analyses allowed Ward et al. (1998) to realize that those populations that readily grow in the artificial conditions of a laboratory may not be representative of populations predominant *in situ*.

Patterning Along Gradients

Using denaturing gradient gel electrophoresis of 16S rRNA gene segments obtained from mats in the effluent channel of Octopus Spring, Ferris and Ward (1997) demonstrated differences in the distribution of the closely related A/B′-lineage *Synechococcus* variants shown in Figure 1.3 relative to water temperature (Figure

1.4). The *Synechococcus* A″ genotype was only found near the upper temperature limit of the mat (∼72 °C), with the A′ and A genotypes found in somewhat cooler downstream sites. The B′ and then the B genotypes became predominant in water that had cooled to <63 °C. As with the A/B′-lineage of *Synechococcus*, the C-lineage (*Roseiflexus* spp.) displayed a similar pattern, with C″ and C′ genotypes found in warmer water than the C genotype. The E-lineage genotype (later identified as '*Candidatus* Thermochlorobacter aerophilum'-like spp.) was not detected in the highest- or the lowest-temperature water, but was instead found in the approximate range 57 °C to 71 °C.

Allewalt et al. (2006) were able to cultivate isolates representative of A/B′-lineage *Synechococcus*, enabling tests of the effect of water temperature on growth rates. The maximal growth rate, and the maximum survived water temperature, for each isolate mirrored the conditions where the genotype could be found *in situ* (Figure 1.5), with an isolate with an A-like genotype able to survive and grow faster in the warmer water than isolates with B′-like genotypes. These observations suggested different adaptations of these populations with closely related 16S rRNA sequences, which challenged the notion of using molecular cutoffs to define microbial species (Ward, 1998).

The vertical distribution suggested by autofluorescence evidence of *Synechococcus* spp. (Figure 1.1C) led to the hypothesis that different genotypes segregate within the vertical structure of the mat. Ramsing et al. (2000) tested this hypothesis by using denaturing gradient gel electrophoresis of 16S rRNA gene segments recovered from mat samples that had been cryosectioned along 100 µm intervals (Figure 1.6). The *Synechococcus* B′ genotype was found to be more surface-oriented than the *Synechococcus* A, C (*Roseiflexus*-like), and E″ ('*Candidatus* Thermochlorobacter-like) genotypes. However, validation of adaptation to light required more genetic resolution

than that provided by 16S rRNA sequences, as will be presented below.

Need for Higher Molecular Resolution

Phylogenetic analyses of the internal transcribed spacer (ITS) region adjacent to the 16S rRNA gene, which is more variable in sequence, demonstrated that the 16S rRNA gene alone doesn't always provide enough genetic resolution to observe the ecologically distinct species represented by sequences in a dataset (Ferris et al., 2003). By examining an unrooted phylogenetic tree of sampled ITS diversity collected from the upper and lower layers of 68 °C Mushroom Spring mat, separate surface and subsurface clades become evident (compare clades with mostly circles or squares in Figure 1.7). These ITS results pointed towards the need for higher molecular resolution and a theoretical basis for the understanding of microbial species.

Microbial Species Theory

> "*I was much struck how entirely vague and arbitrary is the distinction between species and varieties*" (Darwin, 1859)
>
> "*There is probably no other concept in biology that has remained so consistently controversial as the species concept*" (Mayr, 1982).

As noted by both Darwin and Mayr, the species concept has remained mired in controversy among systematists studying plants and animals, and the situation is no better among systematists studying microbes. As mentioned above, the classical approach used by microbiologists to delimit species has involved the use of unnatural molecular cutoffs, with little regard for speciation theory developed by plant and animal systematists. Although there are disagreements among various species concepts (mainly involving the role of recombination), there are some common

features that most concepts share: i) a species is founded only once, ii) a species is ecologically distinct from other species (minimal sharing of resources with other species), iii) genetic diversity within a species is constrained by a cohesive force, and iv) different species are irreversibly separate from one another (de Queiroz, 2005; Cohan, 2006; Cohan and Perry, 2007).

In highly sexual species like plants and animals, recombination is a powerful cohesive force that serves to constrain genetic diversity within the species. However, between groups that rarely recombine (as in hybrid-forming species), recombination is not a cohesive force. Recombination in bacteria and archaea is different than it is in plants and animals. First, recombination is rare, happening at about the same rate as, or a lower frequency than mutation (Feil et al., 1999, 2000; Smith et al., 1993; Roberts and Cohan, 1995; Selander et al., 1987; Melendrez et al., 2016). Second, unlike plants and animals, homologous recombination can happen even between distant relatives (Duncan et al., 1989; Majewski and Cohan, 1999; Vulić et al., 1997). Third, recombination is highly localized to small segments of the genome (Feil et al., 1999, 2000; Maiden et al., 1998; Smith et al., 1991). Finally, recombination is not limited to homologous transfer, as foreign genes can be introduced into a genome (Hamilton et al., 1989; Majewski and Cohan, 1999; Van Spanning et al., 1995; Ochman et al., 2000).

Even with a low rate of recombination in asexual species, ecologically distinct phylogenetic clusters can still be detected using ecological theory (Melendrez et al., 2016). Van Valen (1976) proposed that species are maintained by their ecological niche and not via recombination. He proposed the ecological species concept, where a species is a lineage of closely related organisms that i) occupy an ecological niche minimally different from other members of the same species, and ii) evolve separately from all other species outside its niche (Van Valen, 1976).

> "*The one [species concept] that seems to fit most situations better than any other is based on the knowledge that a species is characterized not only by reproductive isolation but also by the fact that it occupies a species-specific ecological niche. [...] It is thus customary to combine into species those asexual individuals and clones that fill the same ecological niche or that play the same role in the ecosystem*" (Mayr, 1982).

Stable Ecotype Model

Cohan (2006) has utilized the commonalities in species concepts, and the panoply of ways that bacteria and archaea are thought to speciate, to form a number of models of microbial speciation. In populations like the thermophilic *Synechococcus* of Mushroom Spring and Octopus Spring, where selection pressure exceeds the rate of speciation, the Stable Ecotype Model performs well at predicting ecotypes in a population (Cohan, 2006; Koeppel et al., 2008; Connor et al., 2010; Becraft et al., 2015). The Stable Ecotype Model of microbial speciation posits that extant genetic diversity within a lineage results from two primary forces: ecotype formation and periodic selection (Figure 1.8). Net ecotype formation accounts for ecotype formation as well as extinction, and would have been described by Darwin as the 'divergence of character' between two populations. Periodic selection results in genome-wide diversity within a single ecotype being quashed, and would have been described by Darwin as the 'survival of the fittest' (Darwin, 1859). Periodic selection events can be a result of competition with an especially well-adapted mutant or recombinant, or due to an external force like a phage infection. While these two primary forces shape an ecotype, neutral mutation and recombination cause increased diversity among the members an ecotype. An ecotype (or ecological species) in this model is a cluster of phylogenetically related organisms that are ecologically interchangeable,

but ecologically distinct from members of other ecotypes (Koeppel et al., 2008) and this will be used as the definition of ecotype in future chapters.

Ecotype Simulation

Koeppel et al. (2008) encoded the Stable Ecotype Model into a Monte-Carlo style evolutionary simulation program called Ecotype Simulation that has proven useful for detecting ecologically distinct clades of bacteria from sequence datasets (Koeppel et al., 2008; Connor et al., 2010; Becraft et al., 2015). Each round of the simulation models a lineage of microbes progressing through time as its members diversify through mutation and experience the periodic selection and ecotype formation of the Stable Ecotype Model (Figure 1.8), resulting in the extant diversity seen in a dataset. Rounds of the simulation that match the diversity present in the sequence dataset provide information about the number of putative ecotypes (PEs) and the rates of periodic selection and ecotype formation. Repeated rounds of the simulation provide confidence in these values, with the most likely value used to demarcate PEs in the sequence dataset. Although proven useful detecting PEs from multiple sequence datasets, Ecotype Simulation is extremely slow and is limited in the number of sequences that can be analyzed (<300). See Chapter 2 for more detailed information on how Ecotype Simulation works, and for an introduction to a new faster version of the algorithm that is not limited in the number of sequences that can be analyzed.

The term putative ecotype is used for the simple reason that Ecotype Simulation only provides a prediction of the species sampled by the dataset. It is necessary for the researcher to validate that the PEs meet the expectations of the ecological species (or ecotype) concept. Members of an ecotype should be ecologically interchangeable with other members of the ecotype, but ecologically distinct from other ecotypes.

Canonical Correspondence Analysis

To test ecological distinction and homogeneity, canonical correspondence analysis (CCA) has been employed (ter Braak, 1986; Legendre and Legendre, 1998). CCA is a multivariate method utilized by numerous ecologists (Birks et al., 1996) to elucidate the relationship among the collection of species that utilize an environment and the physical and/or chemical parameters that define the habitat selection (niche) of each species. CCA has also been used to study seasonal and spatial variation in habitat selection (Snoeijs and Prentice, 1989; Bakker et al., 1990; Anderson et al., 1994). Ecologically interchangeable members of a PE are expected to be non-randomly distributed in the CCA ordination space. Thus, each PE is tested for non-random distribution and a p-value is reported by the CCA plotting function. Ecologically distinct PEs are expected to be separated from each other in the CCA ordination space, but unmeasured niche-defining variables or co-existence of geographical species can hide differences among ecotypes (see Chapter 3).

Theory Dependent Analyses of Rapidly Evolving Protein-encoding Genes

In more recent work, the Ward lab has used protein-encoding loci in theory-based analyses. Melendrez et al. (2011) sampled multiple protein-encoding loci and predicted PEs using Ecotype Simulation to demonstrate that the loci studied could provide more resolution for understanding diversity than the 16S rRNA or ITS regions (compare *apcAB*, *aroA*, or *rbsK* to ITS and 16S rRNA in Figure 1.9). Becraft et al. (2011) demonstrated using Ecotype Simulation and cloning that the protein-encoding locus *psaA*, which plays an important role in Photosystem I and is the most transcribed gene in the Mushroom Spring mat (Becraft et al., 2015), also provided more resolution than 16S rRNA, and could be used for understanding the ecological diversity of thermophilic *Synechococcus*. Because these studies were based

on sequences recovered by cloning, which limits the number of sequences obtained, they were limited in their ability to deeply sample the diversity present in the system.

High-throughput sequencing and barcoding were used in the recent study by Becraft et al. (2015), which not only exemplified the need to use more rapidly evolving protein-encoding genes, but also displayed the usefulness of the approach. High-throughput barcoding retains sample-specific information in a DNA barcode annealed to each sequence in a sample, allowing for correlation between sample and environmental sequence in follow-up analyses. The increased amount of sequences generated from high throughput sequencing means that more of the diversity present in the system is sampled. *Synechococcus* PEs studied by Becraft et al. (2015) displayed different distributions with depth in the 60 °C to 63 °C Mushroom Spring mat, with members of an ecotype displaying similar distributions (Figure 1.10). Figure 1.10 shows A-like PEs predicted from *psaA* sequence variation that all share the same type-A 16S rRNA sequence alongside a single B′-like PE. Becraft et al. (2015) also used CCA to show the same pattern in variation of habitat selection by PEs of *Synechococcus* predicted by Ecotype Simulation using a *psaA* gene segment (note the distribution of PEs B′9, A1, A4, and A14 along the depth vector). Importantly, CCA demonstrated the non-random distributions of the members of most PEs (Figure 1.11). Although high-throughput sequencing increased the amount of diversity sampled, only the most abundant genotypes could be studied because the original version of Ecotype Simulation was only able to analyze <300 sequences, a small subset of the data generated.

Isolates representative of *Synechococcus* ecotypes predicted by Becraft et al. (2015) displayed differential growth patterns in an artificial environment setup to test various amounts light, CO_2 availability, and water temperature (Nowack et al., 2015). These experiments demonstrated that isolates representative of PEs found at

different mat depths grew best in light environments more similar to their niches *in situ*. Isolates representative of PEs found deep in the mat (A4 and A14) grew best when the amount of light was low (Figure 1.12A), while an isolate representative of PE A1, found higher in the mat, grew more slowly in low light, but best in high light (Figure 1.12B).

Comparative genomics of these isolates by Olsen et al. (2015) elucidated differences in gene content that could explain the habitat selection by an ecotype. Ecotypes near the surface, where demand for CO_2 is highest during maximum sunlight, have an extra gene that allows for more efficient scavenging of CO_2 when CO_2 is rare. Ecotypes deeper in the mat, which experience a decreased amount of, and different quality of light (Figure 1.2), have an extra suite of genes for harvesting red light that would still be available at the depth to which they are distributed. Comparisons that include multiple isolates from the same ecotype confirm the expectation of the variants within a species being ecologically distinct from other species, but interchangeable with each other (Olsen, 2015).

Thesis Outline

Theory-based analyses of large DNA datasets based on high-resolution genetic markers is needed to understand the structure and function of microbial communities. Although, Ecotype Simulation has provided a means for theory-based analyses of DNA datasets, it is unable to analyze large sequence datasets produced by modern sequencing technologies. The goals of this thesis were to develop a new version of the evolutionary simulation algorithm, Ecotype Simulation, that could analyze these large sequence datasets in a reasonable amount time, and then to use the new algorithm to address fundamental questions in the biology of *Synechococcus* and other predominant mat taxa. In Chapter 2, I introduce the new version of the algorithm, Ecotype

Simulation 2, and in Chapters 3, 4, and 5, I use it to rapidly analyze large sequence datasets that could not be analyzed with the original version of the algorithm.

Ecotype Simulation 2

Although Ecotype Simulation has proven useful for detecting ecotypes from environmental sequence data, the program is notoriously slow, taking days to analyze only 200 sequences on modern computer hardware. Because DNA sequencing technologies can now generate millions of sequences per analysis, Ecotype Simulation is only able to analyze a tiny fraction of the diversity sampled.

In Chapter 2, I introduce a new version of the algorithm called Ecotype Simulation 2 (ES2) that is based on the same ecological theory as the original simulation, but that can detect ecologically distinct populations from very large DNA sequence datasets. ES2 is able to rapidly analyze these very large datasets (as shown in Figure 2.10). I compare the ability of ES2 to predict ecological populations from sequence datasets with the original version of Ecotype Simulation and other similar algorithms in Chapter 2 utilizing a dataset of *Synechococcus psaA* diversity that was incompletely analyzed by Becraft et al. (2015) due to limitations in the original Ecotype Simulation.

Ecotype Simulation 2 enables the analysis of very large sequence datasets to classify the diversity present in a microbial community into PEs. With the knowledge of the PEs (ecological species) in a community, a researcher can begin to understand the ecological parameters that form the niches of, and validate, the PEs. Knowledge of ecotypes in a microbial community also allows for a better understanding of the interactions among all community members.

Island Biogeography of *Synechococcus* in the American Northwest

Hot springs resemble islands in an ocean of land inhospitable to thermophilic microbes, and as such, they present an opportunity to understand the role of allopatric processes (i.e., physical isolation) in the diversification of these populations (Papke et al., 2003). Microbiologist have long thought that the barriers to geographic distribution were non-existent, that "everything is everywhere, and the environment selects" (van Niel and Thayer, 1930; Baas Becking, 1934). Microbes were thought to evolve primarily due to sympatric means, and were distributed globally without barriers to geographic dispersal. Similarities in morphology, which is known to be a low-resolution discriminator among ecologically distinct populations, led to this inference (Finlay and Fenchel, 2004).

Papke et al. (2003) traveled the world sampling thermophilic *Synechococcus* populations from hot springs in Japan, New Zealand, and North America. Analysis of the slowly-evolving 16S rRNA gene in these populations showed that genotypes were specific to each region (Figure 1.13). Analysis of the more rapidly-evolving 16S-23S rRNA ITS region suggested variation in genotypes among hot springs from different geothermal basins in Japan and North America. However, the scope of that study was limited by the lack of theory-based analyses, the use of cloning and sequencing, which limited the degree of sampling, and the use of genetic markers that have been shown to be too slowly evolving to detect the most newly divergent ecotypes (ecological species) of *Synechococcus* (Melendrez et al., 2011; Becraft et al., 2011, 2015). Additionally, this study did not observe the importance of the differences in spring chemistry (Figure 1.14), which may have played a role in the diversification of *Synechococcus*.

Becraft (2014) reanalyzed American Northwest samples collected by Papke et al. (2003) in an attempt to expand the amount of sequence data and increase

molecular resolution by using high-throughput sequencing of a *psaA* gene segment. Although this analysis utilized ecological theory to reexamine the biogeography of *Synechococcus*, it was limited by the use of the original version of Ecotype Simulation and it did not analyze the environmental parameters rigorously using CCA.

In Chapter 3, I reanalyze American Northwest samples collected by Papke et al. (2003) and sequenced by Becraft (2014) using the PE predictions of ES2 to gain a better understanding of the distribution of thermophilic A/B′-lineage *Synechococcus*. I used CCA to test the PEs predicted by ES2 for ecological homogeneity and distinction. I show that everything is not everywhere, and that both allopatric and sympatric processes appear to play an important role in the speciation of A/B′-lineage *Synechococcus*.

Synechococcus Succession and Stability

Populations of microbes have been shown to change seasonally in relation to fluctuating environmental conditions (DuRand et al., 2001; Thompson et al., 2004; Giovannoni and Vergin, 2012). Yet, thermophilic microbial communities inhabiting the effluent channels of Mushroom Spring and Octopus Spring have been mostly treated as static assemblages in studies of these communities over the last fourty years. However, light availability changes over a diel cycle and total light availability changes seasonally, with substantially shorter days in the winter.

Ferris and Ward (1997) showed that there was no change in the community between seasons, based on the the slowly-evolving 16S rRNA gene (Figure 1.4). However, that study did not use ecological theory and was based on a cloning and sequencing approach, which results in few sequences recovered and a low coverage of the diversity present in the system. Furthermore, the low resolution of 16S rRNA sequences may have masked differences by effectively lumping species.

Becraft (2014) expanded the amount of sequence data and increased molecular resolution by using high-throughput sequencing of a *psaA* gene segement from samples collected by various Ward lab members between 1989 and 2011. This analysis utilized ecological theory, but was limited by use of the original version of Ecotype Simulation and it did not analyze replicate samples.

In Chapter 4, I utilize ES2, CCA, and high-throughput sequencing of replicate Mushroom Spring mat samples collected near the winter and summer solstice to gain a better understanding of population differences within A/B′-lineage *Synechococcus* over time. This analysis displayed clear patterns of change in the population structures between seasons. However, I also reexamined the samples sequenced by Becraft (2014), showing the relative stability in population structure over a twenty-five year period.

Generality of Ecotype Concept Across Mat Taxa

Variation in thermophilic A/B′-lineage *Synechococcus* has been well studied, but other mat community members have not been studied to the same extent. Using metagenomic assembly, Klatt et al. (2011) described eight clusters of predominant taxa in these mat systems, six of which were capable of phototrophic metabolisms (Figure 1.15). In addition to A/B′-lineage *Synechococcus* (Cluster 1), which perform oxygenic photosynthesis, the mat community harbors a number of anoxygenic phototrophs. Filamentous anoxygenic phototrophic members of the phylum Chloroflexi make up three clusters, *Roseiflexus* spp. (Cluster 2), *Chloroflexus* spp. (Cluster 3), and the newly described '*Candidatus* Roseilinea gracile' (Cluster 6). Additionally, anoxygenic phototrophic members of the phyla Acidobacteria and Chlorobi were described: *Chloracidobacterium thermophilum* (Cluster 4) and '*Candidatus* Thermochlorobacter aerophilum' (Cluster 5).

Although ecotypes of *Roseiflexus* and '*Ca.* Thermochlorobacter aerophilum' have not previously been studied, diversity in habitat selection was detected by analyzing 16S rRNA variants along the temperature gradient provided by the Octopus Spring effluent channel by Ferris and Ward (1997) (note the C-like and E-like variation with temperature in Figure 1.4). Similarly, Cyanobacteria, Chloroflexi, and Chloracidobacteria lineages have been shown, using 16S rRNA barcoding analyses, to exhibit ecological divergence along the thermal gradients of White Creek and Rabbit Creek in Yellowstone National Park; Figure 1.16) (Miller et al., 2009; Weltzer and Miller, 2013).

In Chapter 5, I utilize high-throughput sequencing of protein-encoding loci similar to and including *psaA*, ES2, and CCA to analyze the distribution of PEs of these other five phototrophic community members. Like *Synechococcus*, PEs of these other phototrophic taxa predicted by ES2 distribute differently with respect to both water temperature and depth within the mat. Clearly, the ecotype concept is useful for describing diversity in more taxa than thermophilic *Synechococcus*.

In the final chapter, Chapter 6, I summarize major findings and discuss uncertanties that need to be resolved in future research.

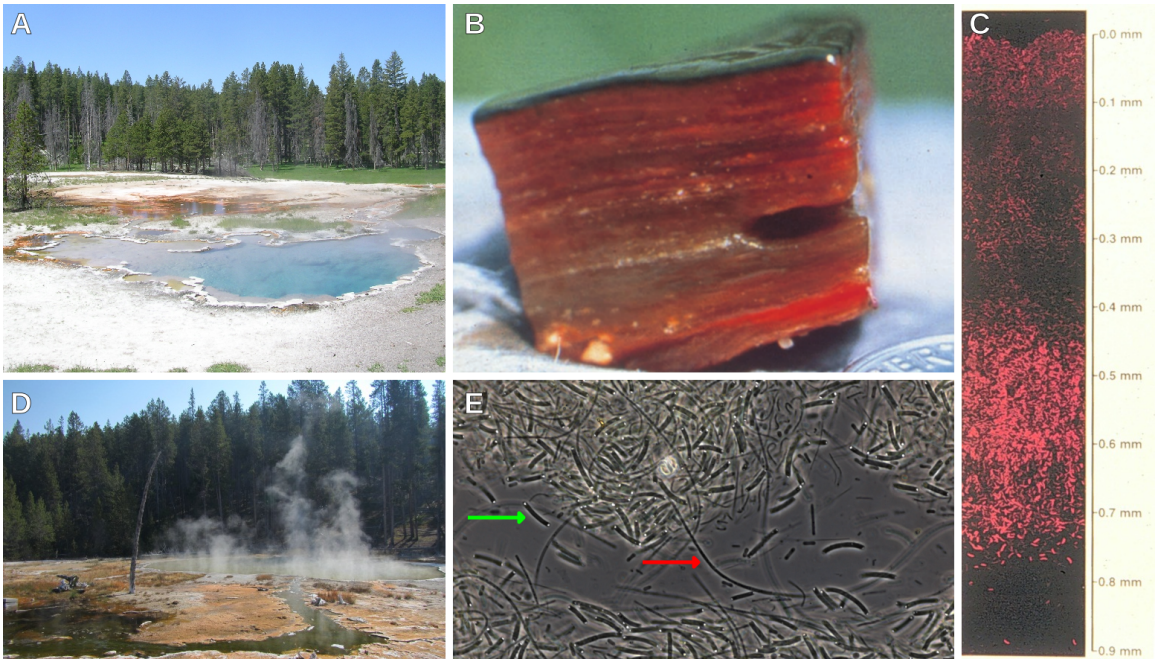


Figure 1.1: Pictures of Octopus Spring (A) and Mushroom Spring (D), a slice of Octopus Spring mat (B), an autofluorescence micrograph of *Synechococcus* spp. inhabiting the top 1 mm green layer, which contain chlorophyll and fluoresce red (C), and a phase-contrast micrograph containing rod-shaped *Synechococcus* spp. (green arrow) and filamentous *Roseiflexus* spp. (red arrow) (E). *Synechococcus* cells are ∼8 µm in length. Adapted from Ward et al. (1998).

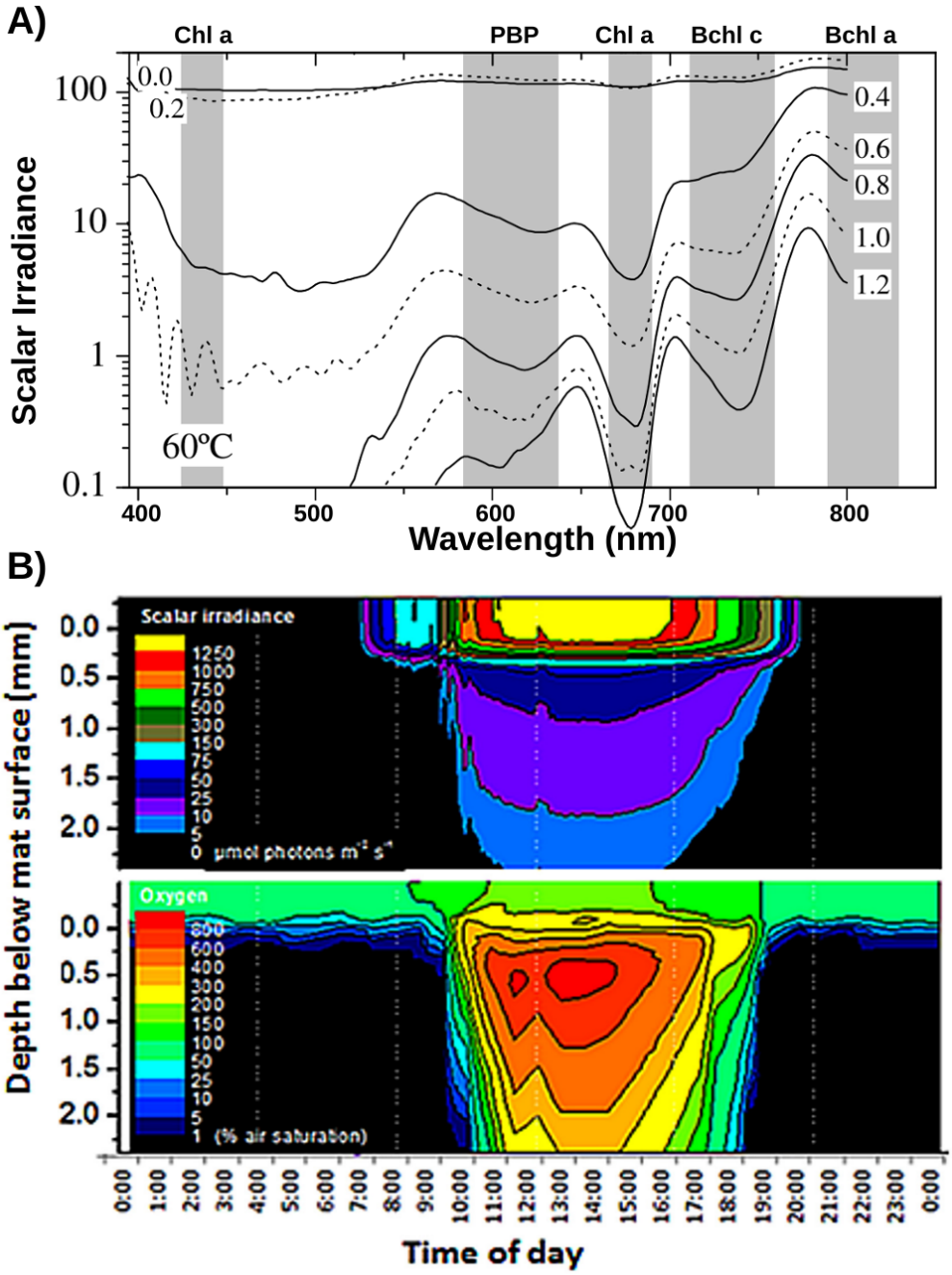


Figure 1.2: Microsensor measurements through the vertical aspect of the Mushroom Spring mat of light intensity versus wavelength (A) and isopleth diagrams showing depth distribution of scalar irradiance and O_2 concentration (% air saturation) over a diel cycle (B). The numbers on the curve in (A) denote depth (mm) below the mat surface, and the shaded regions indicate absorption maxima of photopigments in the mat: chlorophyll a (Chl a), phycobiliproteins (PBP), bacteriochlorophyll c (Bchl c), and bacteriochlorophyll a (Bchl a). Adapted from Ward et al. (2006) (A) and Becraft et al. (2015) (B).

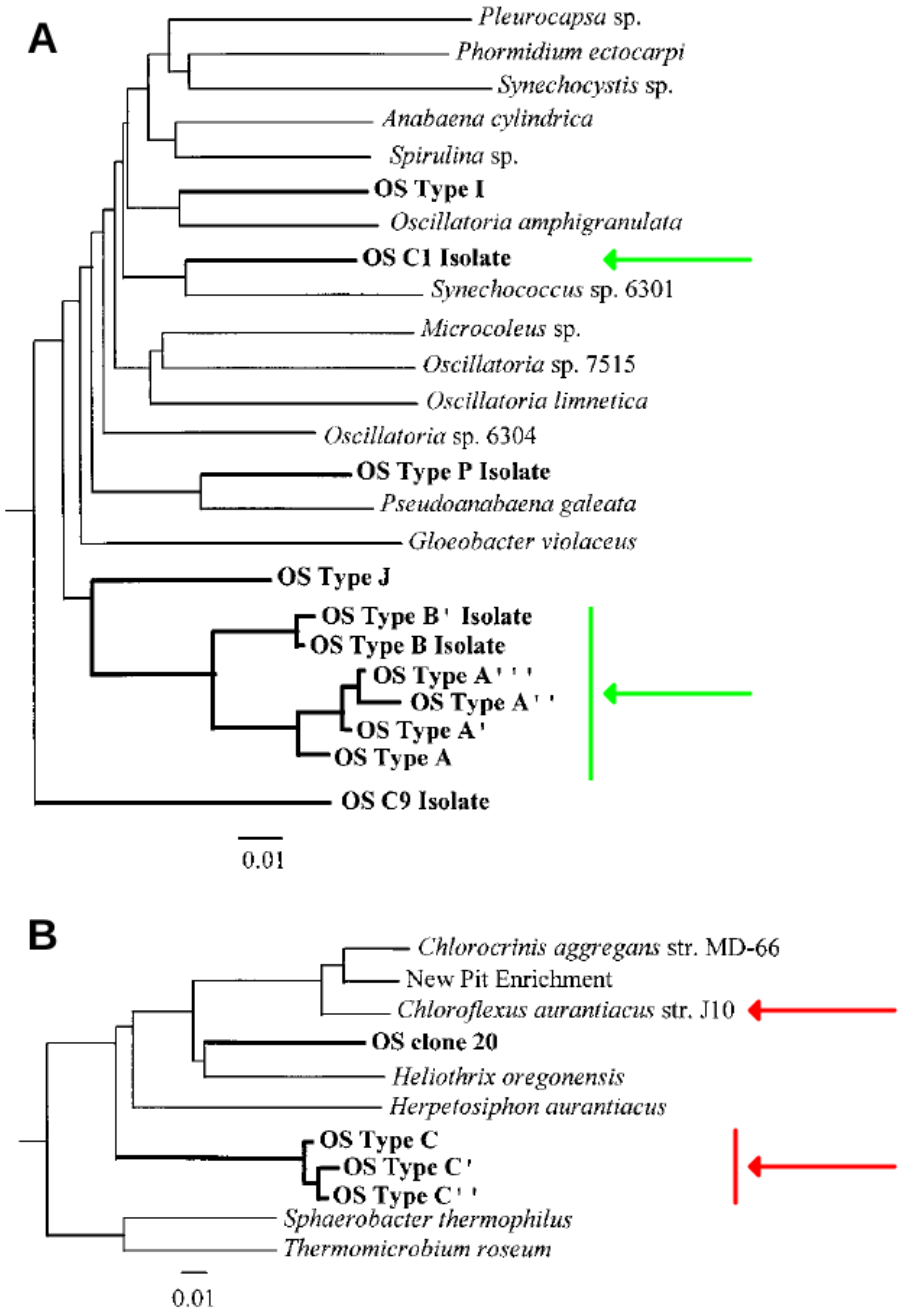


Figure 1.3: 16S rRNA-based phylogenetic trees displaying Cyanobacteria (A) and Chloroflexi (B) lineages. The green arrows point to *Synechococcus lividus* (OS C1 isolate) and the A/B′-lineage clade. The red arrows point to *Chloroflexus aurantiacus* and the *Roseiflexus* (OS Type-C) clade. Bold-highlighted Type A, B, and C variants were retrieved from Octopus Spring microbial mat. Adapted from Ward et al. (1998).

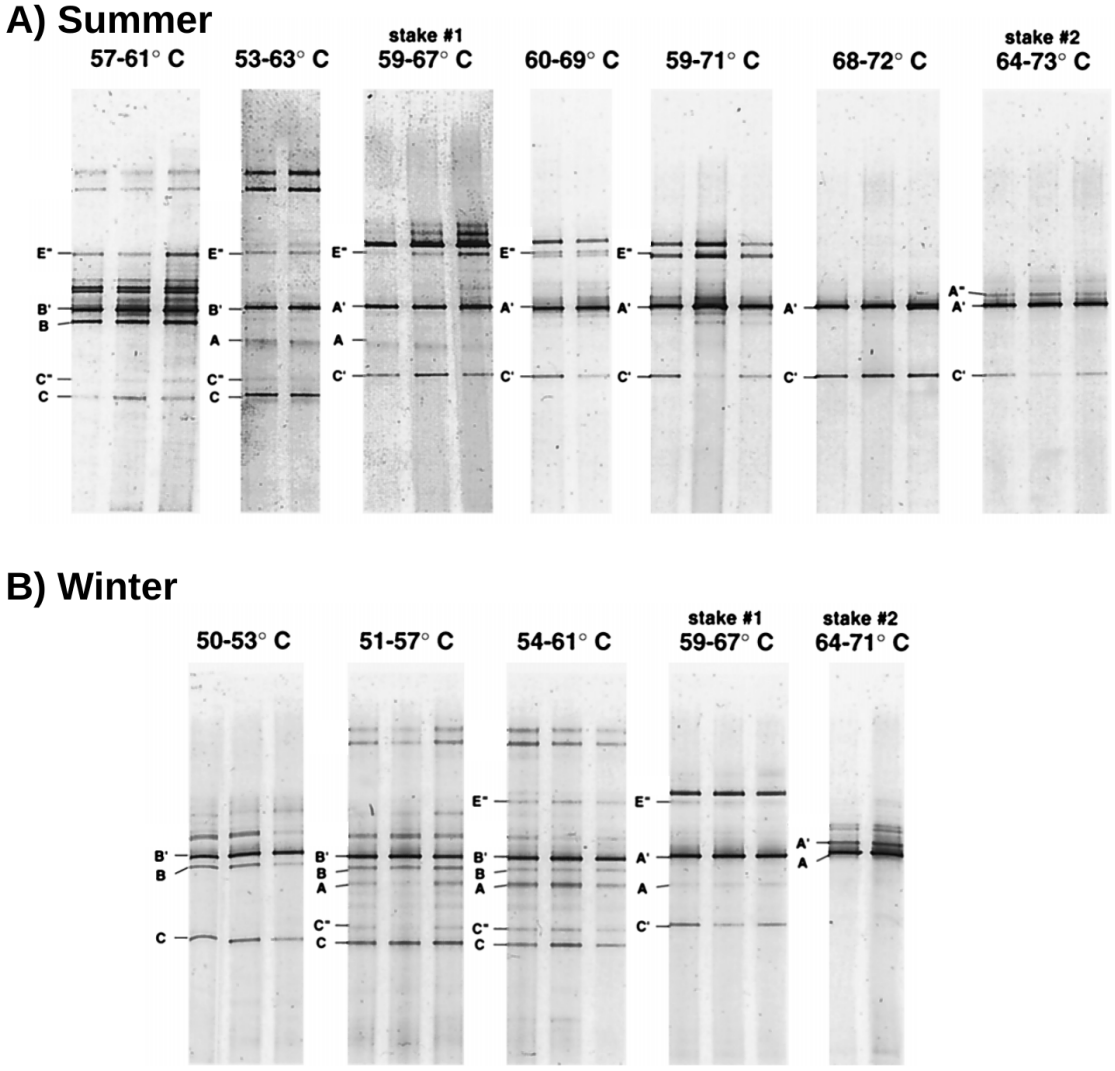


Figure 1.4: Denaturing gradient gel electrophoresis analysis of 16S rRNA gene segments obtained from microbial mat sampled from multiple temperature-defined sites (75 °C to 48 °C) along the Octopus Spring effluent channel during summer (A) and winter (B). Replicate samples were grouped together in adjacent sets of 2 or 3 lanes. Bands labeled A″, A′, A, B′, and B correspond to A/B′-lineage *Synechococcus* diversity. Bands labeled C″, C′, and C correspond to *Roseiflexus*-like diversity. Bands labeled E″ correspond to '*Candidatus* Thermochlorobacter aerophilum'-like diversity. Adapted from Ferris and Ward (1997).

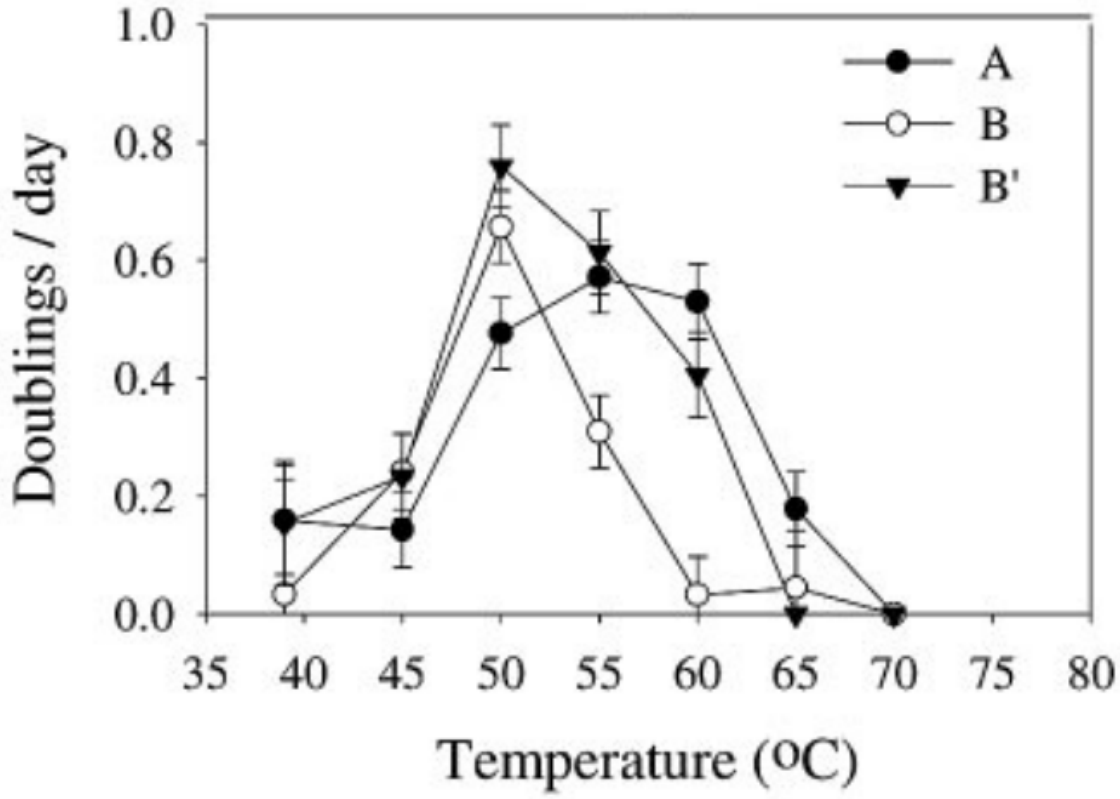


Figure 1.5: Growth rates of Octopus Spring *Synechococcus* isolates with A, B, and B′ genotypes with respect to temperature. Error bars correspond to ±1 standard error. Adapted from Allewalt et al. (2006).

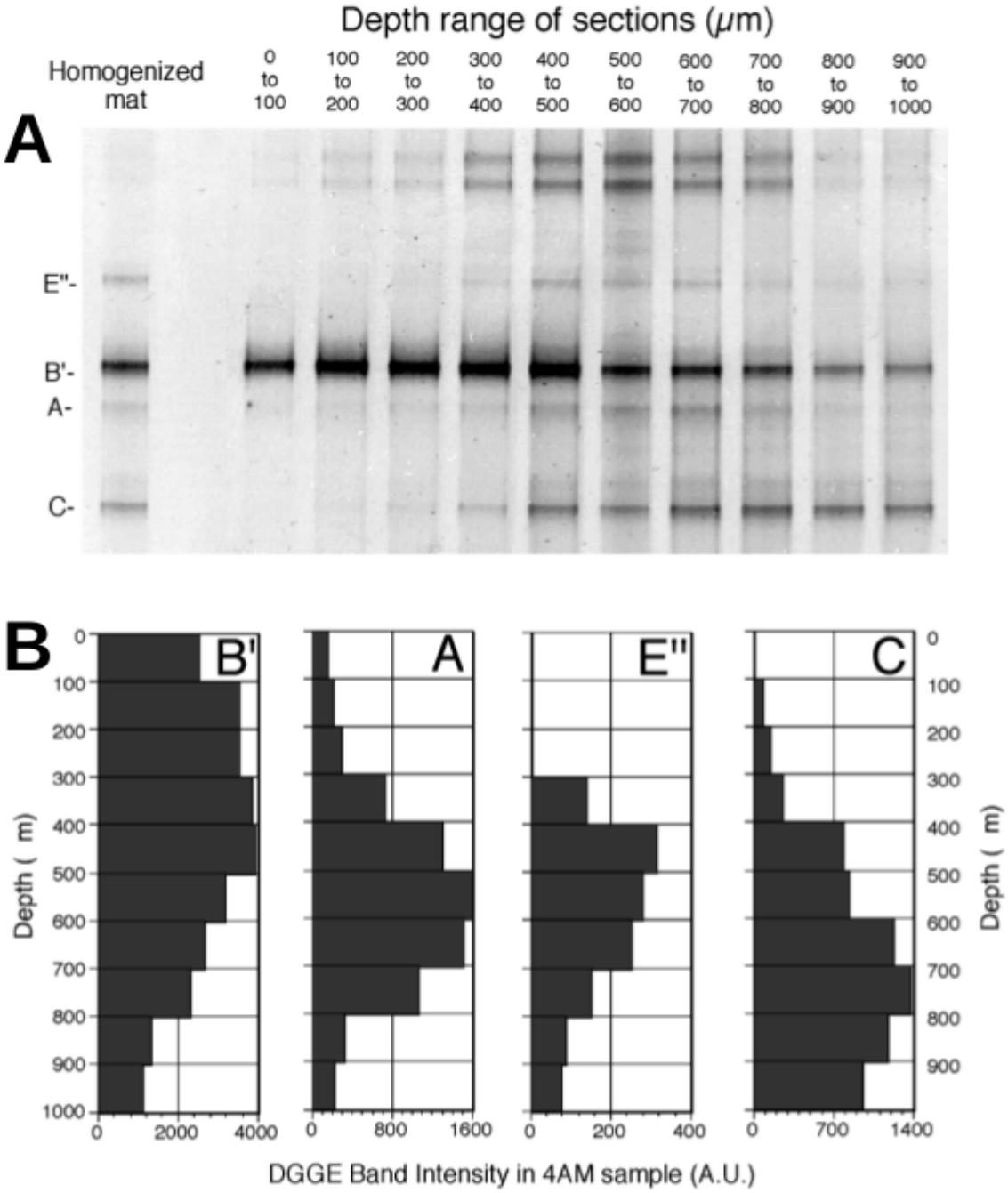


Figure 1.6: Denaturing gradient gel electrophoresis of 16S rRNA gene segments obtained from microbial mat sampled from Mushroom Spring at 61 °C. Bands labeled A and B′ correspond to A/B′-lineage *Synechococcus* diversity. The band labeled C corresponds to *Roseiflexus*-like diversity. The band labeled E″ corresponds to '*Candidatus* Thermochlorobacter aerophilum'-like diversity. Adapted from Ramsing et al. (2000).

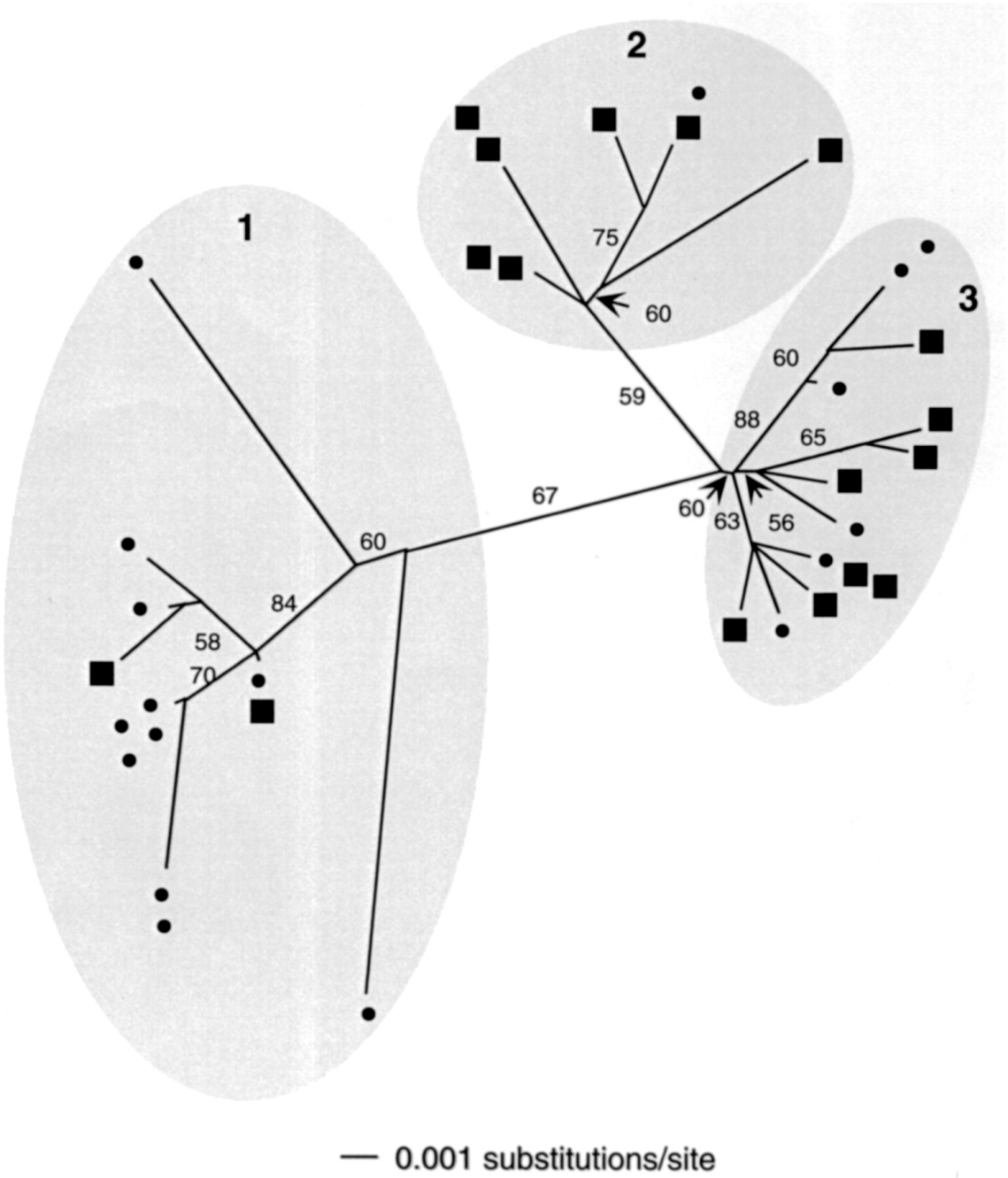


Figure 1.7: Unrooted neighbor-joining phylogenetic tree of the *Synechococcus* internal transcribed spacer region obtained from microbial mat sampled from Mushroom Spring at 68 °C. Circles represent sequences collected from the surface of the mat and squares represent sequences collected from the subsurface (∼2 mm below the surface). Identical sequences are positioned together at the end of the branch. Copied from Ferris et al. (2003).

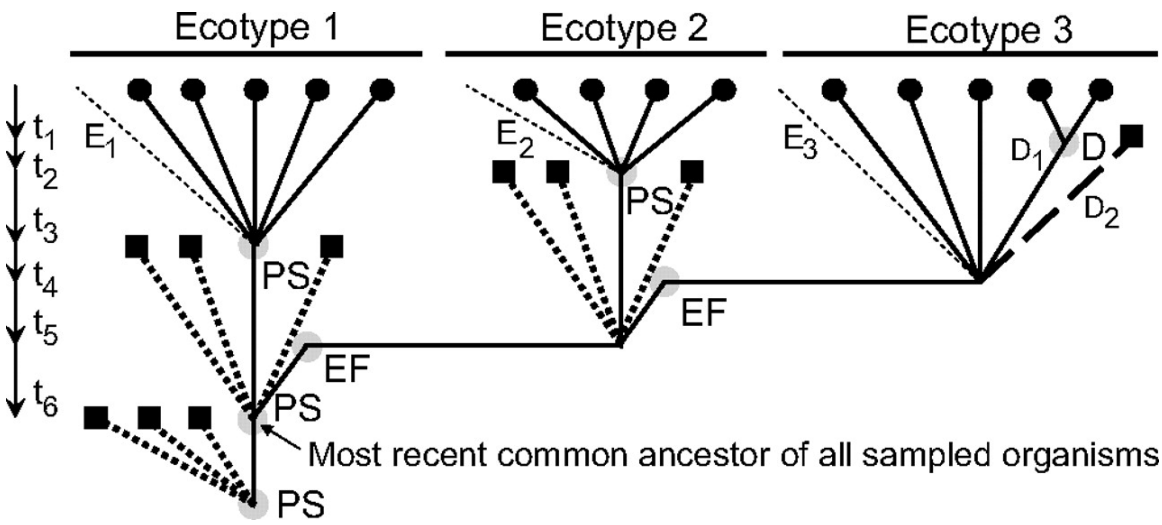


Figure 1.8: The Stable Ecotype Model of microbial species and speciation. The model assumes a lineage of organisms is progressing through time (t_{1-6}, with t_6 being the oldest) experiencing periodic selection events (PS), ecotype formation (EF), and drift (D). Extant lineages sampled are represented by black circles, while those not sampled (E_{1-3}) are displayed with a light dashed line. Extinct lineages caused by periodic selection and drift are displayed with heavy dotted and dashed lines, respectively, with extinction represented by boxes. Copied from Koeppel et al. (2008).

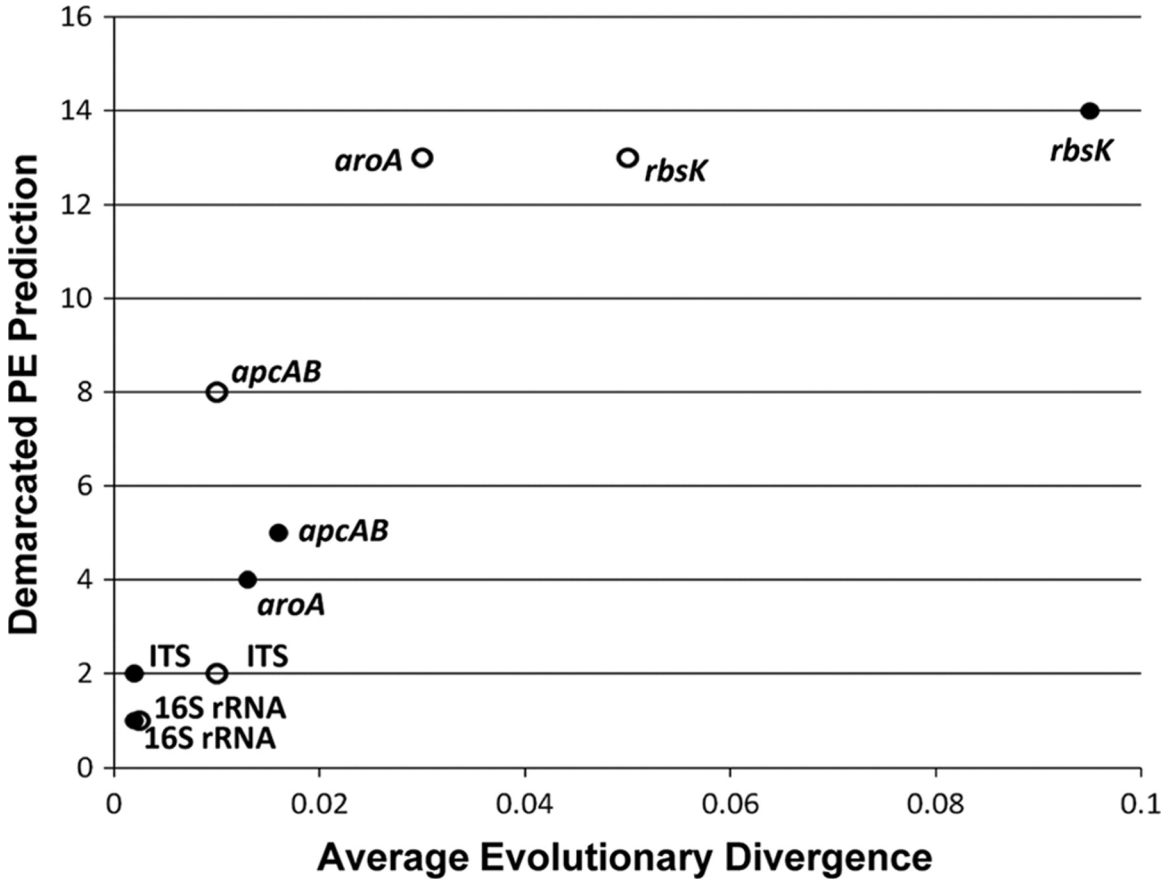


Figure 1.9: Effect of evolutionary divergence on putative ecotype (PE) demarcation by Ecotype Simulation. Genes encoding the allophycocyanin alpha and beta subunits (*apcAB*), 3-phosphoskimate 1-carboxyvinyltransferase (*aroA*), and ribokinase (*rbsK*) were compared with the previously studied 16S rRNA and internal transcribed spacer regions from *Synechococcus* A-like (closed circles) and B′-like (open circles) variants. Copied from Melendrez et al. (2011).

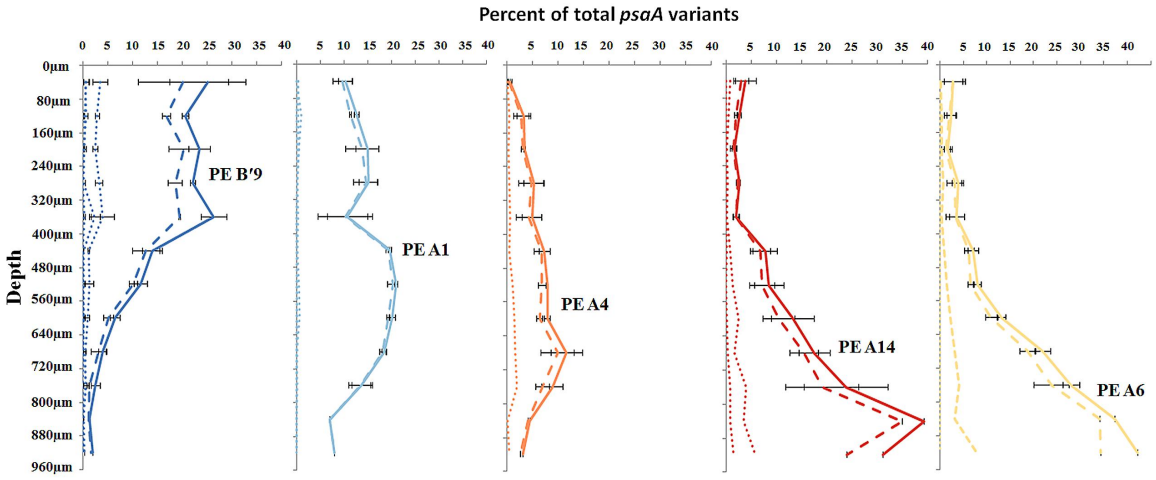


Figure 1.10: Vertical distribution of predominant *Synechococcus* putative ecotypes (PEs) predicted from Ecotype Simulation demarcation of a protein-encoding locus, *psaA*, segment. Mat samples from 60 °C to 63 °C were dissected at 80 µm intervals using a cryotome before analysis. Solid lines represent the relative abundance of all variants within a PE, the dashed line represents the relative abundance of the dominant variant, and the dotted lines represent the relative abundance of other high-abundance variants within the PE. Copied from Becraft et al. (2015).

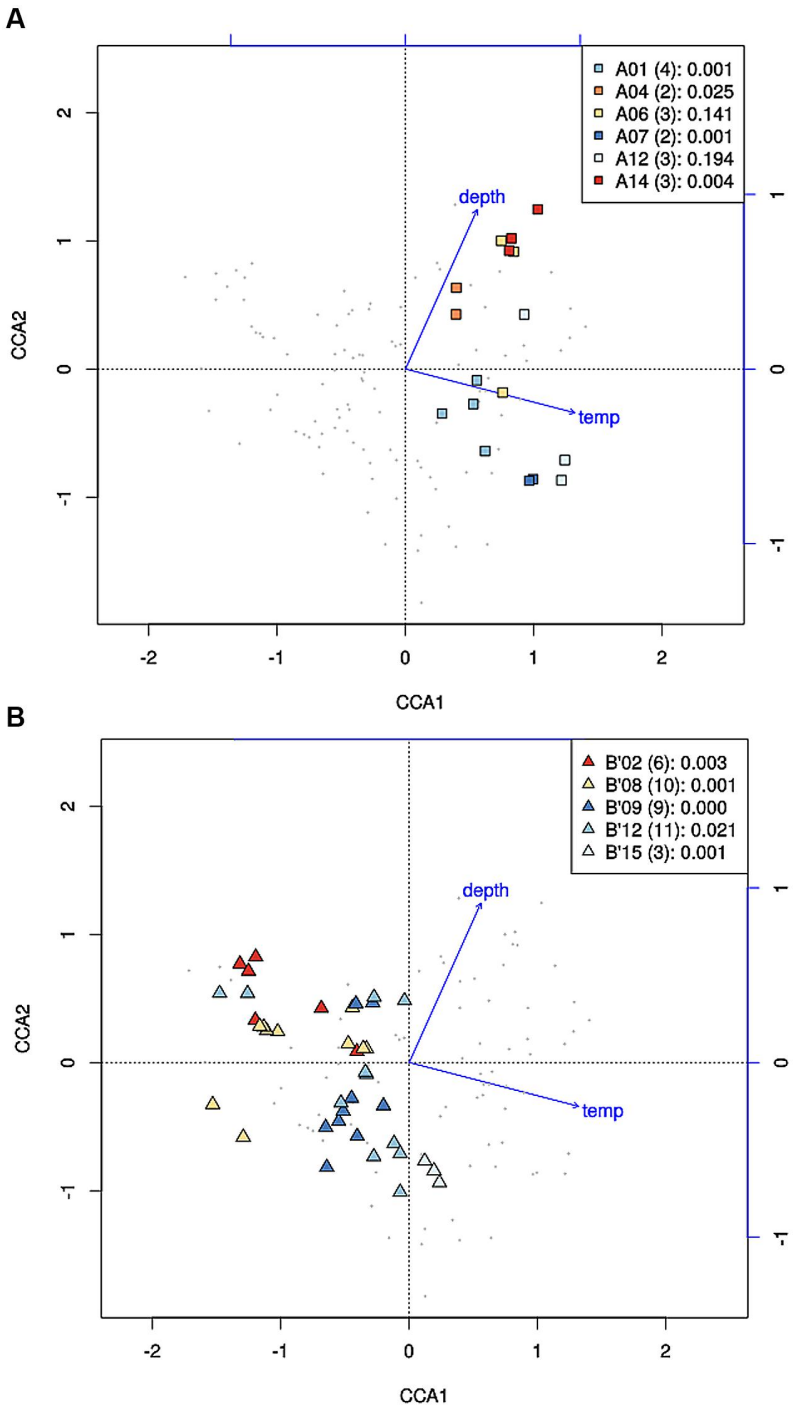


Figure 1.11: Canonical correspondence analyses of *Synechococcus* A-like (A) and B′-like (B) high-frequency *psaA* sequence variants predicted by Ecotype Simulation relative to temperature and depth within the mat. The number of variants is shown in parentheses, and a p-value that represents the probability that the observed ecotype cluster is distributed different from random, are listed in the key after the putative ecotype name. Copied from Becraft et al. (2015).

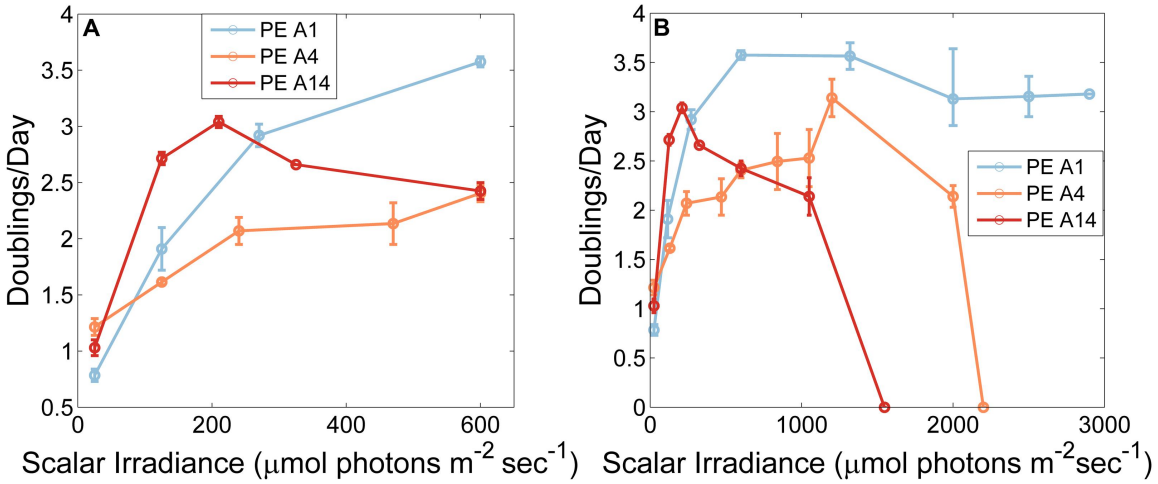


Figure 1.12: Growth rates of *Synechococcus* strains representative of putative ecotypes (PEs) predicted by Ecotype Simulation analyses of *psaA* gene segement sequences as a function of irradiance when grown at 60 °C. Copied from Nowack et al. (2015).

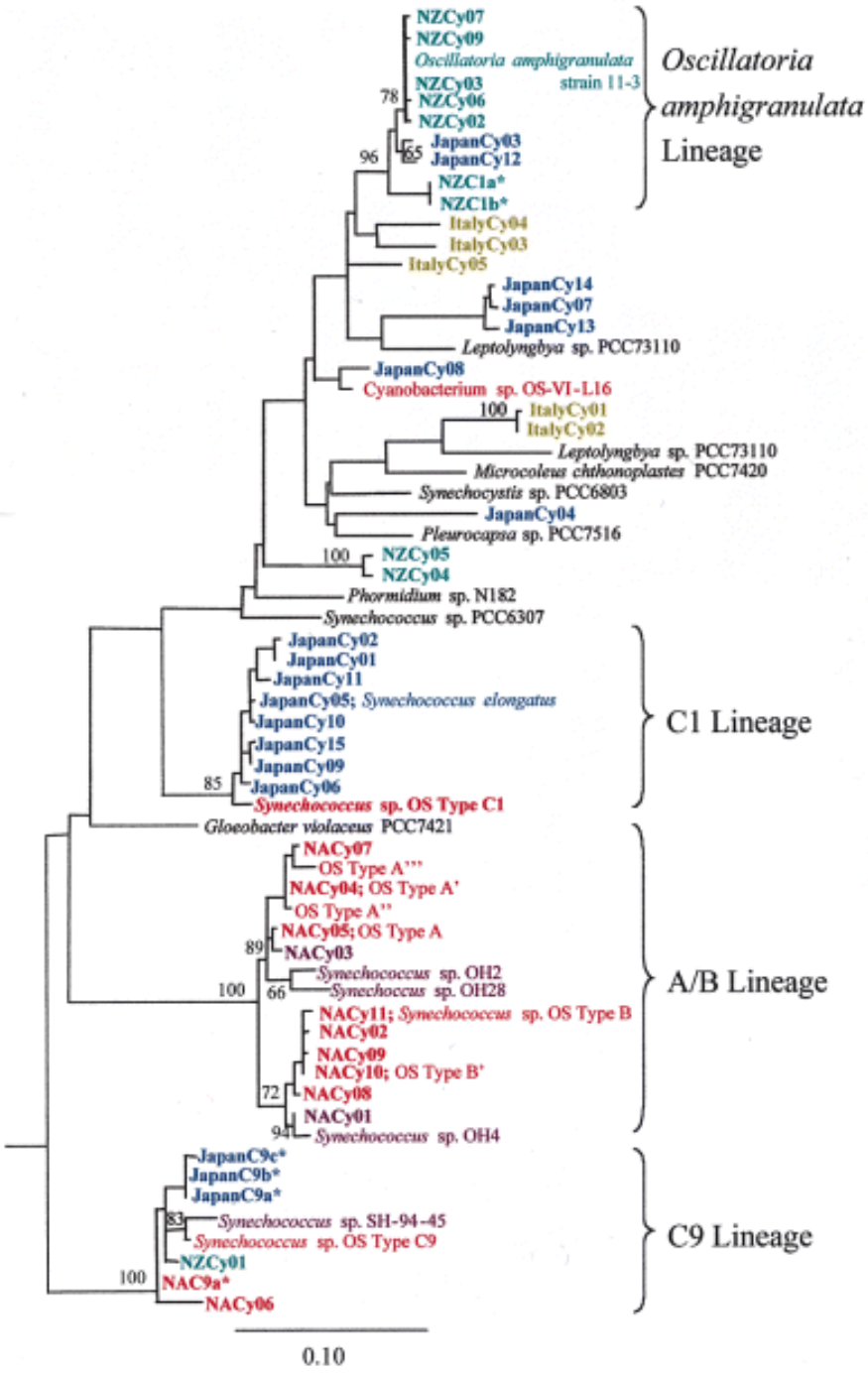


Figure 1.13: Phylogenetic tree of cyanobacterial 16S rRNA sequences collected from hot springs across the world. Colored sequence names correspond to the country of origin, with cyan from New Zealand, blue from Japan, gold from Italy, red from the Greater Yellowstone Ecosystem, and purple from Oregon. Copied from Papke et al. (2003).

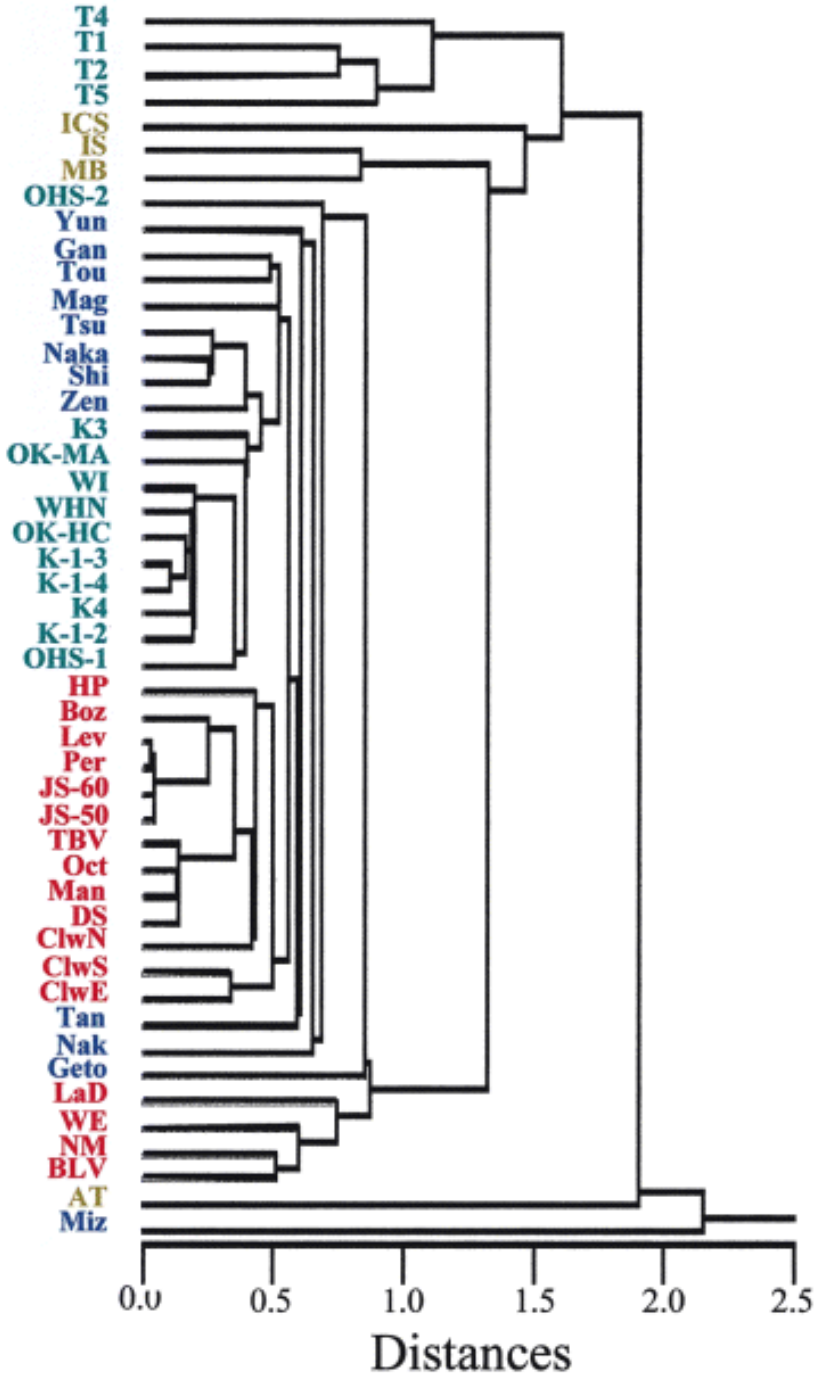


Figure 1.14: Hierarchical cluster analysis of chemistry for springs shown in Figure 1.13. Hot springs are colored by country name, with cyan from New Zealand, blue from Japan, gold from Italy, red from the Greater Yellowstone Ecosystem and Oregon. Copied from Papke et al. (2003).

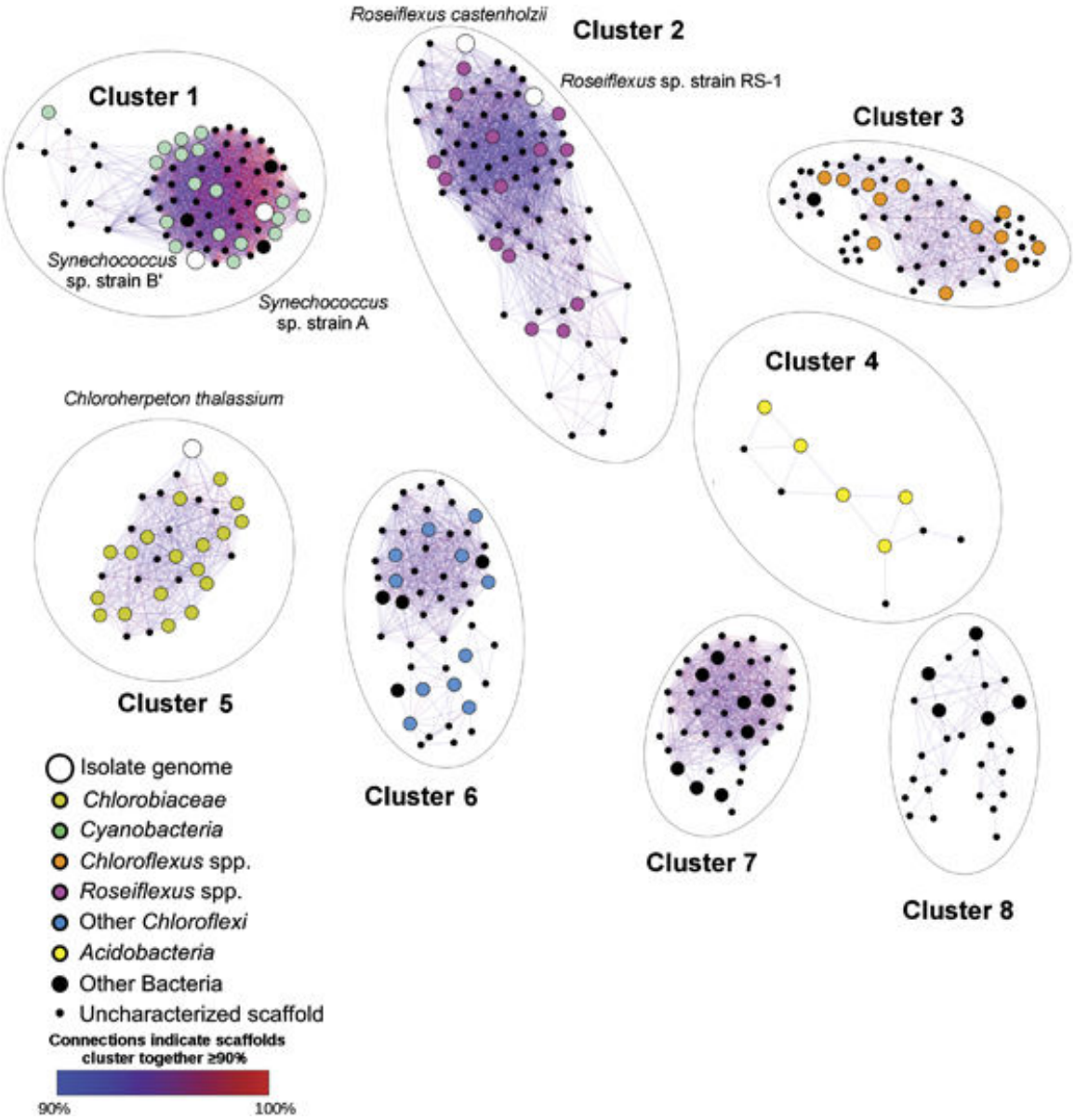


Figure 1.15: Eight predominant taxa inhabiting the Octopus Spring and Mushroom Spring microbial mats detected from metagenomic assembly and clustering based on oligonuclotide usage frequency. Each point represents an assembled sequence scaffold (black points and colored circles) or isolate genome (white circles) used for reference. The colored circles represent scaffolds that could be characterized with phylogenetic marker genes. Colored lines connect points together that cluster in $\geq$ 90% of trials. The size of each ellipse is proportional to the amount of metagenomic sequence data contained within each scaffold cluster. Copied from Klatt et al. (2011).

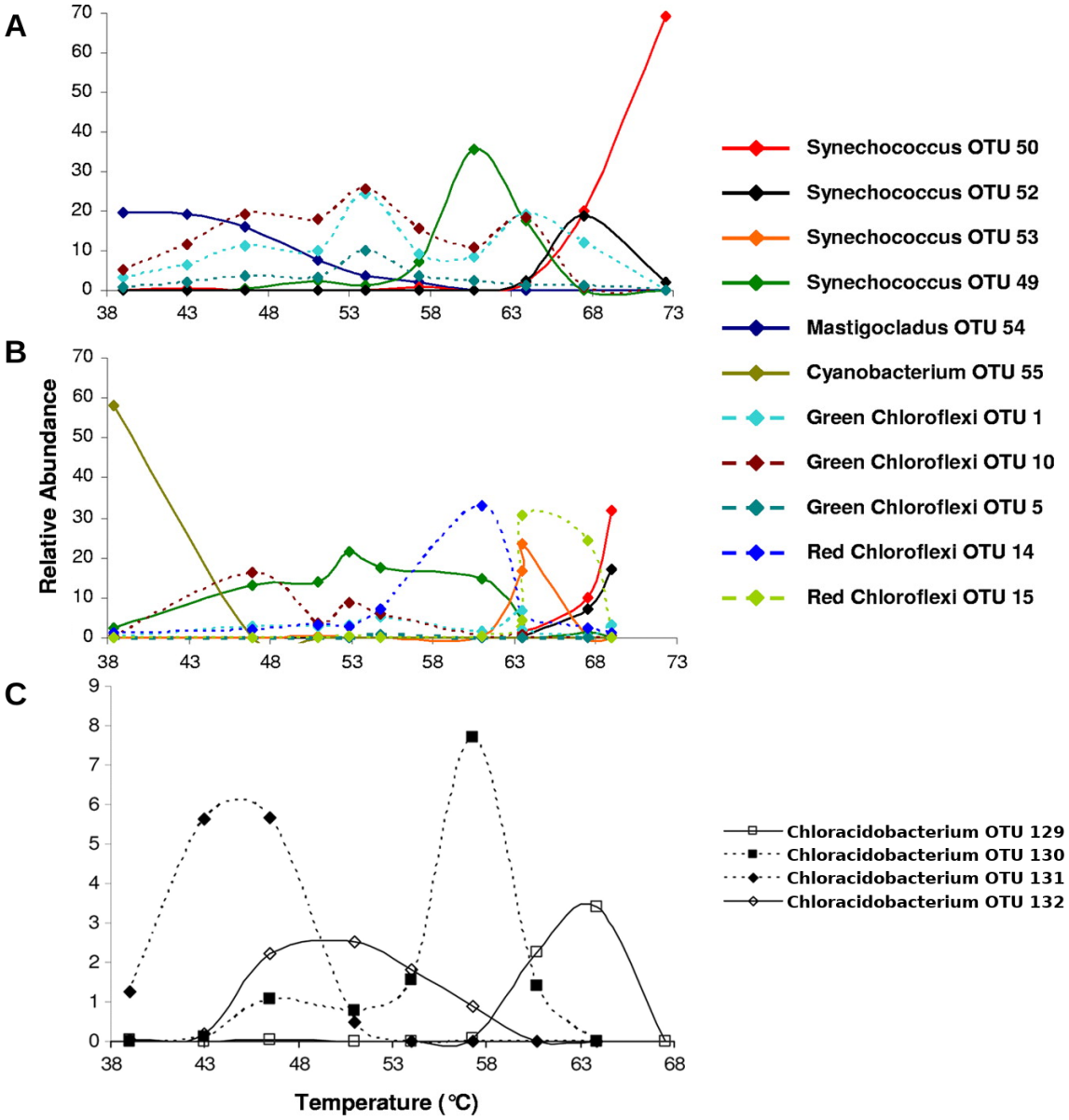


Figure 1.16: Distribution of Cyanobacteria and Chloroflexi along the thermal gradients of White Creek (A) and Rabbit Creek (B), and the distribution of Chloracidobacteria along the White Creek thermal gradient (C). Populations shown are operational taxonomic units (OTUs), based on 16S rRNA sequences that are $\leq 1\%$ divergent. Adapted from Miller et al. (2009).

CHAPTER TWO

ECOTYPE SIMULATION 2: AN IMPROVED ALGORITHM FOR EFFICIENTLY DEMARCATING MICROBIAL SPECIES FROM LARGE SEQUENCE DATASETS

Contribution of Authors and Co-Authors

Manuscript in Chapter 2.

Author: Jason M. Wood

Contributions: helped plan and develop ES2, analyzed barcode sequence data, performed canonical correspondence analyses and comparisons between algorithms, and assisted in preparation of the manuscript.

Co-Author: Dr. Eric D. Becraft

Contributions: tested early versions of ES2 and participated in preparation of the manuscript.

Co-Author: Dr. Danny Krizanc

Contributions: helped plan and develop ES2 and participated in preparation of the manuscript.

Co-Author: Dr. Frederick M. Cohan

Contributions: helped plan and develop ES2, co-supervised the research, and participated in preparation of the manuscript.

Co-Author: Dr. David M. Ward

Contributions: obtained funding for the project, supervised the work, and participated in preparation of the manuscript.

Manuscript Information Page

Jason M. Wood, Eric D. Becraft, Danny Krizanc, Frederick M. Cohan, and David M. Ward

BMC Bioinformatics

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

Published by BioMed Central
Submitted January, 2018

Abstract

Microbial systematists have used molecular cutoffs to categorize the vast diversity present within a natural microbial community without including ecological theory to help identify whether or not demarcated groups are the ecologically distinct, fundamental units (ecotypes) necessary for understanding the system. Ecotype Simulation, a Monte-Carlo approach to modeling the evolutionary dynamics of a microbial population based on the Stable Ecotype Model of microbial speciation, has proven useful for finding these fundamental units. For instance, predicted ecotypes of *Synechococcus* forming microbial mats in Yellowstone National Park hot springs, which were previously considered to be a single species based on phenotype, have been shown to be ecologically distinct, with specialization to different temperature and light availability or quality. Unfortunately, development of high-throughput DNA sequencing methods has outpaced the ability of the program to analyze all of the sequence data produced. Here we developed an improved version of the program called Ecotype Simulation 2, which can rapidly analyze alignments of very large sequence datasets. For instance, while the older version takes days to analyze 200 sequences, the new version can analyze 5000 sequences in about an hour. The faster simulation identified similar ecotypes as found with the slower version, but from larger amounts of sequence data. Algorithms with a similar intent, like the maximum-likelihood and Bayesian implementations of the Poisson Tree Processes model, failed to detect these ecotypes, lumping or splitting the diversity into unnatural groups. Based on ecological theory, Ecotype Simulation 2 provides a much-needed foundation that will help guide microbial ecologists and systematists to the natural, fundamental units.

Keywords: speciation, bacteria, periodic selection, bacterial diversity

Background

The identification of closely related, ecologically distinct populations (ecotypes or ecological species) within a microbial community is paramount to the understanding of the structure and function of the community. Advances in sequencing technology and metagenomic techniques have enabled the identification of the major guild-level constituents (e.g., primary producers like *Synechococcus* spp.) (Dick et al., 2009; Klatt et al., 2011). However, identifying the species-level constituents occupying unique niches within a microbial community can be greatly complicated by the small physical size of the microhabitat and the lack of easily identifiable phenotypic differences amongst closely related but ecologically distinct populations (Becraft et al., 2011).

Early work on microbial species was based on the use of phenotypic differences in cell shape, composition, and metabolism to differentiate between populations. However, systematists were unable to look within natural populations to verify that they were not lumping organisms with similar phenotypes into unnatural groups containing multiple species. With the hope of adding more scientific rigor to species demarcation, molecular techniques such as 70% whole-genome DNA-DNA hybridization (Wayne et al., 1987), 95-96% amino acid identity (Konstantinidis and Tiedje, 2005), or about 1% identity in 16S rRNA (Stackebrandt and Ebers, 2006) were suggested and calibrated to match the historical phenotype-based groupings. However, little attention was given to whether the resulting sequence clusters came from ecologically distinct populations, and indeed, most species taxa were demarcated to include multiple ecologically distinct groups (Aboal et al., 2016; Dugat et al., 2015; Choudhary and Johri, 2011; Oh et al., 2010; Paul et al., 2010; Hunt et al., 2008; Lefébure and Stanhope, 2007; Vernikos et al., 2007; Kettler et al., 2007; Marri et al.,

2006; Jaspers and Overmann, 2004; García-Martínez et al., 2002). Cohan and Kopac (2017) pointed out that lumping species not only hinders a full understanding of ecological diversity within a bacterial community, but also broadens the apparent diversity within and exaggerates the geographical range of clusters, complicates detection of newly emergent species, and may limit biotechnological potential.

Some have claimed that it would be futile to name and describe all ecological diversity. Doolittle and colleagues have argued that, because of the role of horizontal genetic transfer in bacterial diversification, each individual cell might be ecologically unique and so could be demarcated as its own ecological species (Doolittle and Zhaxybayeva, 2009). Indeed, in some bacterial groups, particularly the generalist heterotrophs such as *Bacillus*, the rate of bacterial speciation appears to be quite high (Kopac et al., 2014). Here the discovery of individual ecotypes within a clade may require an extremely highly resolved phylogeny, perhaps based on the entire core genome. On the other hand, speciation appears to be slow in groups with less opportunity for diversification, such as the photoautotrophs, C1-utilizing heterotrophs, and intracellular pathogens (Cohan, 2016). In photoautotrophic, thermophilic, unicellular cyanobacteria, *Synechococcus* spp., speciation is slow enough that even a single gene segment can resolve diversity of individual, ecologically distinct populations (Becraft et al., 2015). The work we describe here is focused on the slowly speciating *Synechococcus* of Yellowstone hot springs which populate well-established environmental gradients. In such systems, phylogenetically distinct clusters of closely related individuals sharing the same adaptations, metabolic requirements, and susceptibility to the same selection regime exist and can be detected by using a theory-based framework for understanding the diversity.

Fortunately, systematists studying plants and animals have created a long list of species concepts that microbial systematists can utilize for inspiration (Ward, 2006).

Although there are disagreements among these species concepts (mainly involving the role of recombination), most share some common features: a species is founded only once, is ecologically distinct (minimal sharing of resources with other species), and is a cohesive group (diversity is limited to ecologically similar individuals), while different species are ecologically distinct (minimal sharing of resources with other species), and are irreversibly separate (de Queiroz, 2005; Cohan, 2006; Cohan and Perry, 2007).

These commonalities in species concepts have been used to formulate multiple models of microbial species, with the Stable Ecotype Model performing well in populations where selection pressures for adaptations within a species exceed the rate of speciation (Cohan, 2006). The Stable Ecotype Model of species and speciation considers the genetic diversity present in a lineage to be the result of two primary variables: net ecotype formation and periodic selection. The rate of net ecotype formation takes into account both extinctions and the formation of new ecotypes. Periodic selection events quash genome-wide diversity within a single ecotype without affecting other ecotypes and can be the result of external forces (e.g., phage infection) or competition within the ecotype with an especially well-adapted mutant or recombinant (Cohan, 2006). In keeping with above-stated criteria, the model defines an ecotype (same as ecological species) to be a cluster of individuals that are ecologically interchangeable, but ecologically distinct from the members of other clusters (Koeppel et al., 2008).

The Stable Ecotype Model was encoded into a Monte-Carlo style evolutionary simulation program called Ecotype Simulation (henceforth termed ES1) by Koeppel et al. (2008) that models a lineage of microbes through time as its members speciate and experience periodic selection events. The first stage of ES1 (see Figure 2.1A for an overview) loads user-supplied, aligned sequence data which can be used prior to the ecotype demarcation stage to generate a neighbor-joining phylogenetic tree using

PHYLIP (Felsenstein, 2005). The second stage runs a series of Fortran programs designed to correct for PCR errors and remove gaps from the sequence data. It then calculates an NxN divergence matrix to quantitatively summarize the diversity present (binning) among the N sequences in the dataset by utilizing complete-linkage clustering to obtain the number of sequence clusters (bins) over a range of sequence identity criteria (see Figure 2.2 for an example lineage). Briefly, the number of bins increases as the lineage diversifies through time from the last common ancestor (the blue dot) of the extant diversity (labeled A-Q). As depicted by the vertical bars to the right of the hypothetical tree shown in Figure 2.2, when the sequence identity criterion is restricted to more closely related sequences, more cluster bins result (see Figure 2.3 and Supplemental Figure A.1 for a graphical presentation of results generated from environmental sequences).

In ES1 (Figure 2.1A), these bins are then used to estimate the rates of periodic selection (termed *sigma*), net ecotype formation (termed *omega*), genetic $drift$, and the estimated number of ecotype populations (termed *npop*) present in the sequence dataset analyzed. These estimated values are found in ES1 by using a brute-force method that repeatedly simulates the evolution of the lineage using a predefined range of *omega*, *sigma*, $drift$, and *npop* values, with the order of events and times at which they occur determined randomly. The estimated values that make up the best model are those that are able to reproduce the binning pattern of the dataset with maximum likelihood. Simulations that are able to reproduce the binning curve are deemed successful, with the most likely *omega*, *sigma*, $drift$, and *npop* value combinations proceeding to the next stage – hill climbing. The hill climbing stage utilizes the Nelder-Mead simplex method (Nelder and Mead, 1965) to optimize the estimated *omega*, *sigma*, $drift$, and *npop* parameter values through repeated simulations. By varying a single parameter at a time and again utilizing the Nelder-Mead method to

optimize the simulation, confidence intervals for each value can be found.

The optimized hill climbing parameter values, along with a phylogenetic tree calculated from the sequence dataset (or provided by the user), are then used in the final stage of ecotype demarcation. The ecotype demarcation stage separately simulates each branch of the tree, starting from the root node (see the blue dot in Figure 2.2) and progressing though the tree recursively (to the right) until an internal node (red dot) is found with an *npop* value of 1 being within the confidence interval, or a single leaf node (labeled A-Q) is left and demarcated as a putative ecotype (PE). This default coarse-scale demarcation method was modified by Becraft et al. (2015) to return fine-scale results by changing the demarcation criterion to require a *npop* value of 1 as the most likely result rather than simply being within the confidence interval. The ecotypes predicted by these analyses are considered PEs until they are proven to have the qualities of ecological species.

ES1 has been tested using microbial communities of Yellowstone National Park and the "Evolution Canyons" of Israel and Death Valley. A microbial mat community containing A/B′-lineage *Synechococcus* living along the effluent channel of Mushroom Spring, Yellowstone National Park, WY, USA has long been known to have phylogenetic clusters of closely related variants differentially distributed along the changing temperature of the flow path. Ferris and Ward (1997) found 16S rRNA-defined A′- and A-clade *Synechococcus* in the higher temperature samples, and B′- or B-clade *Synechococcus* in the lower temperature samples. Isolated strains representative of 16S rRNA-defined linages contain many more closely related clusters that demonstrated different adaptive responses to temperature, with an A-like *Synechococcus* strain, JA-3-3Ab, adapted to higher temperature than a B′-like *Synechococcus* strain, JA-2-3B'a(2-13) (Allewalt et al., 2006; Bhaya et al., 2007). High-throughput 454 Titanium sequencing of a 324nt segment of a more rapidly

evolving gene (*psaA*, which encodes an essential protein subunit of photosystem I) and barcoding, combined with ES1 and statistical methods, have shown that these phylogenetic clusters are ecologically distinct. This approach also showed these clusters to be cohesive groups (i.e., with ecologically interchangeable members). All members of a group were shown to cluster non-randomly in canonical correspondence analyses (CCA) of temperature and depth distribution and reacted differently to environmental perturbation (Becraft et al., 2011, 2015). These PEs were shown to inhabit different locations within the vertical profile of the mat, with PEs B′9, A1 (which shares the same *psaA* sequence as strain JA-3-3Ab), A4, then A14 progressing from the surface to 1 mm below (Becraft et al., 2015). More recently, *Synechococcus* isolates that share the same *psaA* sequence as PEs A1, A4 (strain 65AY6A5), and A14 (strain 60AY4M2) demonstrated different adaptive and acclimative responses to light intensity and quality, with optimal growth *in vitro* similar to conditions present where they can be found in the vertical profile of the mat *in situ* (Nowack et al., 2015). Comparative genomics and transcriptomics have elucidated genetic differences among these PEs that explain their different metabolic requirements (e.g., low-light adapted PEs were found to have an extra cassette of photosynthetic antenna genes that allow spectral fine tuning under low-light conditions) (Olsen et al., 2015). Similarly, PEs of *Bacillus* predicted by ES1 have been shown to inhabit different slope aspects of "Evolution Canyons" in Israel and Death Valley (Koeppel et al., 2008; Connor et al., 2010).

Although ES1 has proven useful for detecting ecologically distinct clades within groups of bacteria that had previously been considered a single species, the program is extremely slow and is greatly limited by the number of sequences that can be analyzed. Analyzing more than 200 sequences with ES1 can take days on modern computer hardware (Intel Core i7-6700) and then quickly becomes impossible as the

number of sequences rises. This is at odds with the large number of DNA sequences that are produced by modern DNA sequencing technologies that have outpaced the ability of ES1 to analyze the data (Illumina MiSeq can now generate 25 million sequences per run). Various portions of ES1 (Figure 2.1A) were not designed to manage the large number of sequences produced by modern sequencing technologies (e.g., the memory usage of the ES1 matrix-based binning algorithm increases with the square of the number of sequences and the CPU usage increases with the cube of the number of sequences). The phylogenetic algorithms provided by PHYLIP and used by ES1 suffer from the same problem, and would offer no help in analyzing the large number of sequences produced with modern sequencing techniques. Other parts of the ES1 program needlessly store excess data in memory (e.g., both the Java object that loads Fasta-formatted sequence data and the Fortran simulation code store all sequences in memory even though the sequence data are unused during the simulation). By addressing these and other issues in this new version of Ecotype Simulation (henceforth termed ES2), we show that it is possible to successfully predict ecological species from very large datasets in a reasonable amount of time.

Implementation

Changes to Ecotype Simulation

ES1 and ES2 work in roughly the same way (see Figures 2.1A and B for an overview comparison) with some fundamental differences that may affect the results. ES2 has a redesigned graphical user interface that will be familiar to ES1 users but is more functional (Figure 2.4). ES2 also has a new command line interface for use when a graphics interface is not available. Like ES1, ES2 utilizes a phylogenetic tree for the ecotype demarcation stage, but in addition, ES2 uses the tree for the binning stage. ES2 incorporates FastTree (Price et al., 2009) rather than treeing algorithms

found in PHYLIP (Felsenstein, 2005) if the user wishes to use ES2 to generate the phylogenetic tree. ES1 and ES2 both calculate the number of sequence clusters (bins) to quantitatively summarize the diversity present in a sequence dataset at various levels of sequence identity criteria, but ES2 utilizes a new tree-based binning method that avoids the need to clean up the sequence data or to spend considerable time calculating a divergence matrix (see Figure 2.2 for an overview and Figure 2.3 for a comparison in resulting curves). Both ES1 and ES2 then use these bins to estimate the rates of periodic selection (*sigma*) and net ecotype formation (*omega*), and estimate the number of ecotype populations (*npop*). However, ES2 uses a custom version of Lloyd's algorithm (Lloyd, 1982) for parameter estimation that converges on a likely solution much faster than the ES1 brute-force method. Since the population sizes of bacteria are incredibly large (Robinson et al., 2011), genetic drift can be safely ignored and has been removed from the simulation in ES2. The ES2 hill climbing stage, the *omega*, *sigma*, and *npop* confidence intervals, and the ecotype demarcation stage are all simulated in the same manner described above for ES1, but with some fundamental differences and numerous optimizations made to the simulation that will be described below.

2.5.1.1 Phylogenetic Trees To generate phylogenetic trees from sequence data for the binning and ecotype demarcation stages of ES2, FastTree (Price et al., 2009) has been incorporated into the program to generate approximate maximum-likelihood trees. FastTree is able to perform single-gene or whole-genome sequence alignments with reasonable memory and CPU usage, and provides higher topological accuracy than distance matrix-based methods similar to the neighbor-joining method provided by PHYLIP (Felsenstein, 2005). Since the ES1 matrix-based binning method is similar to the neighbor-joining method, we expect the higher topological accuracy

of FastTree to benefit the new ES2 tree-based binning method.

2.5.1.2 Binning Because the ecotype demarcation stage of ES1 requires that a phylogenetic tree be provided, we reasoned that a tree-based binning algorithm could utilize the clusters already present in the tree to arrive at a similar binning solution to the CPU- and memory-consuming ES1 matrix-based complete-linkage method but without the need to calculate or store the NxN comparison matrix, where N is the number of sequences analyzed. Whereas the ES1 matrix-based algorithm computes the distance between all members of the sequence dataset (i.e., all possible pairs of sequences, N^2), the new tree-based algorithm of ES2 only computes the maximum distance among leaf-nodes that represent each sequence in the dataset (see the labels A-Q in Figure 2.2)) and the internal nodes that connect them together (red and blue dots). For instance, in Figure 2.2, sequences A and B both exceed an identity criterion of 0.995 and are demarcated as separate bins. At an identity criterion of 0.990, they are combined into a single bin, and at an identity criterion of 0.980, they are grouped with sequences C-I into a much larger bin. But, once defined, each sequence or bin is 'pruned' from the tree. Thus, by using a tree-based algorithm, ES2 is able to avoid calculating distances between nodes that share no direct relation. Utilizing a tree-based binning method thus offers two primary advantages. First, ES2 is able to calculate the number of bins using a method that scales linearly with the number of sequences rather than the ES1 matrix-based binning method that scales cubicly. Second, ES2 can now use binning results for all stages of the program that were calculated using the same phylogeny that ES1 used only in the final ecotype demarcation stage.

2.5.1.3 Estimating *Omega*, *Sigma*, and *Npop* The algorithm that estimates the initial values of *omega*, *sigma*, and *npop* for the hill climbing in ES1 (Figure 2.1) used

a brute-force approach – it searched a large range of predefined *omega*, *sigma*, and *npop* parameter values for likely results. A faster approach was needed to achieve the goals established for ES2. Using a custom variant of Lloyd's k-means algorithm (Lloyd, 1982) designed to partition point data into clusters based on distance from a line rather than a point, the ES2 approach fits two lines to the base-two logarithm of the binning data (the graph in Figure 2.4 shows the two fitted lines). The slope of the green line, closest to the 1.0 sequence identity criterion serves as an estimate of *sigma* – the rate of periodic selection. The slope of the blue line estimates the rate of ecotype formation – *omega*. The number of bins at the intersection of these two lines serves as the estimate for *npop* – the number of predicted ecotypes.

2.5.1.4 The Simulation Multiple changes were made to the simulation code used by hill climbing, demarcation, and the *omega*, *sigma*, and *npop* confidence intervals. Firstly, OpenMP (Dagum and Menon, 1998) was used to thread the main-loop of the simulation that is repeatedly run to allow ES2 to fully utilize modern multi-core computers. To overcome an inability to use the intrinsic Fortran 90 pseudo-random number generator in a threaded environment, the ziggurat algorithm (Marsaglia and Tsang, 2000), was translated from C source code into Fortran 90, and added to ES2. The ziggurat algorithm was modified to use an object-like interface to store the state variables for the pseudo-random number generator, enabling a separate state to be used for each thread.

The simulation itself has also been simplified to reduce both runtime and memory costs. First, and most vital, the simulation now uses an ultrametric approach to estimate the number of mutations, rather than drawing from a Poisson distribution. Second, as mentioned above, the rate of genetic drift is no longer calculated or used in the simulation. Third, simulation style code in ES1 that was run a random number

of times during each round of simulation has been simplified to a mathematical function ($timeWait = -1 * log(x)/rateKey$, with x chosen randomly from a uniform distribution in the range $[0, 1)$) that runs in constant time while retaining the stochasticity of $timeWait$. All three changes make the overall simulation much faster without greatly affecting the results.

One last noteworthy change to the simulation involves the addition of a new Fortran 90 module to store dynamically sized arrays. ES1 used statically sized arrays to store variably sized content, and was usually successful at avoiding stack-overflow errors by using overly large default values for the array size. However, no error checking was done in ES1, and the program would occasionally fail with cryptic error messages. This new module dynamically increases the size of the array as needed, allowing ES2 to store variably sized content and fixing stability issues seen in ES1.

2.5.1.5 Ecotype Demarcation The ecotype demarcation stage has been slightly simplified in order to increase its speed. The ES1 version of the algorithm tested all *npop* values between 1 and the current *npop* estimate for the clade to find the most likely value. Since the point of this stage is to determine whether the clade currently being examined makes up a single or multiple ecotypes, ES2 has been simplified to assume that $npop > 1$ if the average success rate of 1000 trials is found to be zero when $npop = 1$ for the clade. Only when the average success rate of these trials with $npop = 1$ is greater than zero are other values of *npop* tested using a likelihood-ratio test for whether they are significantly more likely ($\alpha = 0.05$).

Evaluation of Ecotype Simulation

2.5.2.1 Environmental Sequences The environmental sequence segments analyzed by Becraft et al. (2015) were used to test the ability of ES2 to predict ecotypes in the Mushroom Spring microbial mat. To demonstrate the enhanced

capabilities of ES2 while retaining the ability to compare results with ES1 (that is only able to analyze ∼300 sequences), the sampled diversity was limited to high-frequency sequences (HFSs) using two different cutoff values. We analyzed the same subset of *Synechococcus psaA* segment sequence diversity that was analyzed by Becraft et al. (2015). That is, we sampled only the HFSs, defined either as those with >50 identical copies across the entire dataset (HFS_{50}) or those with >10 identical copies across the entire dataset (HFS_{10}). In the HFS_{50} analysis, there were 65 A-like *Synechococcus* genotypes and 88 B′-like genotypes. In the HFS_{10} analysis, there were 246 A-like *Synechococcus* genotypes and 298 B′-like genotypes. The script that Becraft et al. (2015) used to find the HFS_{50} was used to find the HFS_{10} from the same environmental sequence dataset (`hfs-finder.pl`; available from https://github.com/sandain/pigeon). In the main text, we focus on predominant PEs in the B′-lineage (PE B′9) and A-lineage (PEs A1, A4, and A14) *Synechococcus*, which were found in ES1 analysis to have distinct vertical positioning in the mat at 60 °C to 63 °C. We use PE B′9 and the A-lineage PEs as examples in the main text, but include similar analyses for the remaining B′-lineage in the Supplemental Information.

2.5.2.2 Tree Building Because of the substitution in ES2 of maximum-likelihood trees generated by FastTree for neighbor-joining trees generated by PHYLIP used in ES1, it was necessary to compare analyses made with both algorithms. The neighbor-joining trees used by Becraft et al. (2015) for analyses of the *Synechococcus* A- and B′-like HFS_{50} subset of sequences were used here (Figure 2.5 and Supplemental Figure A.2). The combination of PHYLIP's `dnadist` and `neighbor` programs (Felsenstein, 2005) were used with default parameters to generate neighbor-joining trees for the HFS_{10} subset of sequences. FastTree (Price et al., 2009) was used to build approximate maximum-likelihood trees for both the HFS_{50} and HFS_{10} subsets of sequences (Figure

2.6 and Supplemental Figure A.3). PEs were separately demarcated for A- or B′-like neighbor-joining or maximum-likelihood phylogenies, generated from HFS_{50} or HFS_{10} datasets.

Since clade structure changes between tree algorithms and datasets used, PE names assigned by ES2 in different analyses do not necessarily correspond. Furthermore, PEs newly demarcated and named by ES2 analyses do not have names that correspond with previously named PEs demarcated with ES1 by Becraft et al. (2015). Newly demarcated PEs presented here from A- or B′-like *Synechococcus* are named using the pattern PEAxx or PEBxx, respectively (with xx being a number between 01 and 36). Correspondence between PEs previously described by Becraft et al. (2015) using ES1 is provided by examining the most dominant variant in each ES2 demarcated PE. Newly demarcated PEs with dominant variants that were members of PEs previously demarcated by Becraft et al. (2015) will hereafter have the prior PE designation enclosed in parentheses after the new demarcation designation presented here. For instance, the ES1 PE A1 demarcation by Becraft et al. (2015) (both fine- and coarse-scale) was split by ES2 into PEs PEA13 (A1), PEA14, and PEA15 in neighbor-joining analysis (Figure 2.5), but only PEA13 (A1) retains the original PE designation, since it contains the same dominant variant (HFS013) as the ES1 PE A1.

2.5.2.3 Canonical Correspondence Analysis Canonical correspondence analysis (CCA) (ter Braak, 1986; Legendre and Legendre, 1998) provided by the `R` package `vegan` (Oksanen et al., 2013) and a new version of the custom plotting software (`cca.R`; available from https://github.com/sandain/R) that was used by Becraft et al. (2015) were used to compare the ecological distribution of the members of PEs with the sampled environmental parameters (temperature of the water and depth

within the mat). The community data matrix analyzed by CCA was created with the `hfs-counter.pl` script (available from https://github.com/sandain/pigeon) that simply counts the abundance of each HFS in each environmental sample. Although PEs were demarcated separately for A- and B′-like *Synechococcus* variants, the CCA community data matrix was comprised of variants from both lineages.

2.5.2.4 Testing for Ecological Interchangeability and Ecological Distinctness

Canonical correspondence analyses provide a method for testing the ecological interchangeability and ecological distinctness of PEs. Ecologically interchangeable members of a PE are expected to form non-randomly distributed clusters within the ordination space of the CCA. The plotting software mentioned above reports a p-value for each PE demarcation that represents the probability that the observed distribution of the PE in ordination space is different from random. Ecologically distinct PEs are expected to form clusters separate from other PEs in the CCA ordination space. Deviation from either of these expectations does not mean that the ecotype demarcation is incorrect. Clusters of PEs that overlap in ordination space could be a result of the niche-defining variable differentiating populations having not been measured. A cluster with a distribution that cannot be differentiated from random could be the result of limited sampling of the PE (i.e., a PE with only two members may appear to be randomly distributed while a PE with the exact same spread in the ordination space but with fifty members may appear distributed differently from random). The researcher must examine each PE predicted by ES2 for these conditions and determine whether they have been satisfied before promoting the putative ecotype to an ecotype. Care must still be taken when interpreting PE distributions with CCA because ES2 may produce PE clusters using a poorly resolving gene segment that could still be differentiated in ordination space. Ecological species,

or ecotypes, should be thought of as the smallest clades meeting the expectations of ecological distinction and ecological interchangeability. Although PEs with only a single member are included in CCA analyses, it is impossible to test for ecological interchangeability in such PE.

2.5.2.5 Analysis of ES2 Runtime To test the capabilities and results of ES2 using variously sized sequence datasets, the entire set of unique sequences sampled by Becraft et al. (2015) was randomly subsampled every 100 sequences from 100 to 5600 sequences (i.e., 100, 200, .., 5600) with the `unique_random_fasta.pl` script (available from https://github.com/sandain/pigeon) and a maximum-likelihood tree was generated with FastTree. Each subset of sequences and its pregenerated tree was run with the ES2 command line interface three times for limited replication on an Intel Core i7-6700 processor and timed using GNU `time` version 1.7.

2.5.2.6 Other Algorithms Tested In addition to the comparison of ES2 ecotype demarcations with those from ES1, the environmental sequence data and associated phylogenetic trees were also used to generate sequence clusters using Bayesian and maximum-likelihood Poisson tree processes (bPTP and PTP) (Zhang et al., 2013) (see the red and orange demaraction lines in Figures 2.5 and 2.6). Default parameters were used for all comparisons.

Attempts were also made to compare ES2 demarcation results to similar algorithms like AdaptML, BAPS, and GMYC, but these attempts were ultimately unsuccessful due to issues analyzing the datasets used in this study. Since Francisco et al. (2014) already demonstrated that ES1 performed significantly better than AdaptML, BAPS, and GMYC, these algorithms were left out of analyses presented here.

2.5.2.7 Comparing Demarcated Clusters The Variation of Information statistic was used to compare clusters demarcated by the various algorithms tested using the same A/B′-lineage *Synechococcus psaA* segment sequence data. Since ES1 has previously been shown to be the most accurate algorithm for species demarcation (Francisco et al., 2014) and Becraft et al. (2015) showed that testing both the coarse- and fine-scale methods was needed to produce the best demarcation, we compare all other demarcations to both the coarse- and fine-scale ES1 methods. Demarcations produced by ES2, PTP, and bPTP from phylogenies produced by PHYLIP and FastTree were compared with ES1 demarcations (both coarse- and fine-scale) produced from the same phylogeny using the `R` package `mcclust` (Fritsch, 2012).

Results and Discussion

Binning

To examine the efficacy of the new tree-based binning method of ES2 compared to the matrix-based binning method of ES1, PHYLIP neighbor-joining and FastTree maximum-likelihood, trees were generated from the same sequence data used for ES1 binning. Because the neighbor-joining method provided by PHYLIP utilizes a matrix-based method that is very similar to the ES1 matrix-based binning method, it is not surprising that the ES2 binning results from neighbor-joining trees more closely follow the ES1 binning results of the environmental sequences (compare the orange and red lines to the blue lines in Figure 2.3 and Supplemental Figure A.1). It is also interesting that the HFS_{10} results appear to be in better agreement than the HFS_{50} results.

Performance of ES1 and ES2

In Figures 2.5 and 2.6, PEs are numbered according to ES2 output, and correspondence with predominant PEs predicted using ES1 by Becraft et al. (2015) is shown within parentheses. In general the ES2 ecotype demarcation results were similar to the results of the ES1 fine-scale and coarse-scale methods. When the ES1 fine-scale and coarse-scale methods were in agreement about the membership of an ecotype, the ES2 demarcation results were also often in agreement (7 PEs and 14 PEs in agreement with the neighbor-joining and maximum-likelihood trees, respectively). When the ES1 fine-scale and coarse-scale methods differed, the ES2 results often agreed with one or the other (4 PEs and 5 PEs in agreement with the fine-scale analysis of the neighbor-joining and maximum-likelihood trees, respectively; and a single agreement with the coarse-scale analysis of the maximum-likelihood tree). However, there were multiple cases where ES2 either splits (24 PEs and 5 PEs split in the neighbor-joining and maximum-likelihood trees, respectively) or lumped (1 PE and 2 PEs in the neighbor-joining and maximum-likelihood trees, respectively) clusters slightly differently than the ES1 fine- and coarse-scale methods.

These differences could result from the use of different treeing algorithms (compare the neighbor-joining tree in Figure 2.5 to the maximum-likelihood tree in Figure 2.6) and/or different demarcation algorithms (compare blue bars of different shade within either figure separately). Because we cannot be certain which tree is closest to reality, we rely on distribution patterns to allow nature to inform us about the existence and composition of PEs. To do so, we focused on the four predominant PEs reported by Becraft et al. (2015), which were differently distributed vertically in 60 °C to 63 °C mat samples. We consider each here in the order of their occurrence from the mat surface to ∼1 mm beneath the surface: PE B′9, PE A1, PE A4 and PE A14.

2.6.2.1 Putative Ecotype B′9 Becraft et al. (2015) initially demarcated PE B′9 based on ES1 coarse-scale analysis using a neighbor-joining phylogeny, while the fine-scale analysis split this clade in two (compare the light-blue and medium-blue lines in Figure 2.5A). The ES2 analysis of the neighbor-joining phylogeny split this clade into five clades, with the four HFS_{50} variants of one clade matching the ES1 fine-scale analysis, PE PEB05 (B′9-2) (see the medium- and dark-blue lines in Figure 2.5A). These four variants of PEB05 (B′9-2) also grouped with a sequence found in prior analyses to be predominant at the mat temperature sampled (DAB9) (Becraft et al., 2015). The HFS_{50} variants in PEB05 (B′9-2) are significantly clustered ($p = 0.001$) in CCA analysis near the surface of the mat (see the blue squares in Figure 2.7A). The ES2 analysis of the maximum-likelihood phylogeny produced the same clade, PEB31 (B′9-2) as both the ES1 coarse- and fine-scale methods (Figure 2.6A). This clade only contains a single HFS_{50} variant, but does clade with the same DAB9 sequence mentioned above. The single HFS_{50} variant is found near the mat surface in CCA analysis, but clustering is impossible to observe with a single variant (see the blue square in Figure 2.7B). Since ES2 permitted the rapid analysis of a larger amounts of sequence data, we analyzed sequences that occurred 10 times or more, thus increasing the number of variants in PE B′9, though different variants were recruited by the two treeing algorithms (PE B′9 inset trees on the right side of Figures 2.5A and 2.6A). The HFS_{10} variants detected in ES2 analyses of both neighbor-joining and maximum-likelihood phylogenies were clustered significantly (PEB076, $p = 0.003$; and PEB104, $p = 0.012$ respectively) near the surface of the mat (see the blue squares in Figure 2.8A and B).

2.6.2.2 Putative Ecotype A1 Becraft et al. (2015) initially demarcated PE A1 based on ES1 fine-scale analysis using a neighbor-joining phylogeny, but the coarse-

scale analysis produced the same demarcation (compare the the light-blue and medium-blue lines in Figure 2.5B). ES2 analysis of the neighbor-joining phylogeny split this clade into three clades, PEA13 (A1), PEA14, and PEA15. The singleton PEA13 (A1) contained HFS013 which shares the same sequence as the PE A1 clade-defining *Synechococcus* strain JA-3-3Ab, and the DAA1 sequence seen previously as a dominant variant of PE A1 (Becraft et al., 2015) (see the dark-blue line in Figure 2.5B). The single HFS_{50} variant is found just beneath the PE B′9 variants and just above the PE A4 variant in CCA space, but clustering is impossible to observe with a single variant (see the red circle in Figure 2.7A). The ES2 PE PEA20 (A1) shares the same demarcation as ES1 fine-scale demarcation of the maximum-likelihood phylogeny (compare the medium-blue and dark-blue lines in Figure 2.6B). The single HFS_{50} variant, HFS013, in PEA20 (A1) is found just beneath the B′9 variants and just above the A4 variant in CCA space, but like above, clustering is impossible to observe (Figure 2.7B). Increasing coverage by analyzing the HFS_{10} variants recruited 8 and 11 additional variants into the neighbor-joining- and maximum-likelihood-defined clades (Figure 2.5B and 2.6B PE A1 inset trees), both of which display significant clustering below PE B′9 variants and above PE A4 variants (PEA013, $p = 0.002$ for neighbor-joining; PEA044, $p = 0.001$ for maximum-likelihood) (compare red circles in Figure 2.8A and B).

2.6.2.3 Putative Ecotype A4 Becraft et al. (2015) initially demarcated PE A4 based on ES1 fine-scale analysis using a neighbor-joining phylogeny, but the coarse-scale analysis produced the same demarcation (compare the the light-blue and medium-blue lines in Figure 2.5B). ES2 analysis of the neighbor-joining phylogeny split this clade into two PEs, only one of which, PEA07 (A4), contained an HFS_{50} variant, HFS015, that shared the same sequence as strain 65AY6A5 and the DAA4

sequence previously observed as a dominant variant of PE A4 by Becraft et al. (2015) (see the dark-blue line in Figure 2.5B). The singleton HFS_{50} variant of PEA07 (A4) could not be analyzed for clustering in CCA of the neighbor-joining phylogeny, though it was positioned beneath the PE A1 variant (see the orange circle in Figure 2.7A). The ES2 analysis of the maximum-likelihood phylogeny produced the same clade, PEA16 (A4) as both the ES1 coarse- and fine-scale methods (Figure 2.6B), and rejoined the same HFS_{50} variants that were grouped in the ES1 analyses of the neighbor-joining phylogeny. These variants were significantly clustered in CCA analysis of the maximum-likelihood phylogeny ($p = 0.029$) beneath the PE A1 variant (see the orange circles in Figure 2.7B). Increasing coverage by analyzing sequences that occurred 10 times or more confirmed that the HFS_{10} variants in the ES2 maximum-likelihood analysis clustered significantly (PEA040, $p = 0.005$), but provided weaker evidence of significant clustering of variants detected in the neighbor-joining tree (PEA030, $p = 0.091$) (compare orange circles in Figure 2.8A and B). In both analyeses, PE A4 variants clustered beneath PE A1 variants.

2.6.2.4 Putative Ecotype A14 Becraft et al. (2015) initially demarcated PE A14 based on ES1 fine-scale analysis using a neighbor-joining phylogeny, while the coarse-scale analysis merged the HFS_{50} variants of PE A13 with PE A14 (compare the light-blue and medium-blue lines in Figure 2.5B). ES2 analysis split the original PE A14 into three singleton PEs, PEA17 (A14), PEA18, and PEA19. PE PEA19 (A14) contained an HFS_{50} variant, HFS019, that shares the same sequence as strain 60AY4M2 and the subA1_1DA sequence used to define the PE A14 clade (Becraft et al., 2015). The singleton HFS_{50} variant of this PE could not be analyzed for clustering in CCA analysis of the neighbor-joining phylogeny, though it was positioned beneath PEs A1 and A4 (see the blue circles in Figure 2.7A). The ES1 coarse-scale

analysis of the maximum-likelihood phylogeny merged the HFS_{50} PEs A1 and A14 from the ES1 fine-scale analysis (see the light-blue and medium-blue bars in Figure 2.6B). The ES2 analysis of the maximum-likelihood phylogeny further divided this clade, splitting the ES1 fine-scale analysis into PEs PEA18 and PEA19 (A14) (see the dark-blue lines in Figure 2.6B). The five HFS_{50} variants of PEA19 (A14) clustered significantly in the CCA of the maximum-likelihood phylogeny beneath PEs B′9, A1, and A4 (see the blue circles in Figure 2.7B). Increased depth of coverage using HFS_{10} added a single new variant to the singleton sequence HFS019 in the neighbor-joining phylogeny (PE A14 inset tree in Figure 2.5B) and these two variants did not significantly cluster in CCA (PEA069, $p = 0.271$). However, the maximum-likelihood tree recruited six new sequences (PE A14 inset tree in Figure 2.6 B), such that the 11 variants now belonging to this PE did cluster significantly (PEA024, $p = 0.006$) below all other predominant PEs analyzed (compare the blue circles in Figure 2.8A and B).

Comparison of ES1 and ES2 PE Predictions

We infer from these observations that ES2 demarcates PEs in a manner similar to that of ES1. While there are multiple differences between versions of the program that mainly affect the variants that are predicted to be within a PE, both analyses of the same dataset yielded similar PE clusters that distributed with depth in the mat similarly to that seen by Becraft et al. (2015). Furthermore, ES2 analysis enabled deeper sequence coverage, which improved the significance of observed clusters. Demarcations based on maximum-likelihood phylogeny appeared to yield the best evidence of the existence and environmental distribution of different closely related *Synechococcus* PEs.

Analysis of the entire HFS_{10} dataset yielded a much larger number of PEs than

those described here, though many are rare contributors to the dataset and are likely to represent rare members of the mat community (note the blue abundance bars in Figure 2.9). Significant environmental clustering in CCA ($p < 0.05$) is observed more frequently in the most abundant PEs (e.g., PEs A1, A4, A14; compare the orange p-value bars and the dashed red confidence interval lines with the corresponding blue bars in Figure 2.9). One exception is PE A6 of Becraft et al. (2015) that is best explained by the observation that the partial *psaA* sequence used to demarcate species is shared by two widely divergent phylogenetic groups (Olsen, 2015). In most cases, however, partial *psaA* sequences permitted the detection of distinct ecological species. If horizontal gene exchange had been rampant, one would expect more well-sampled PE clusters to not be significantly clustered in CCA. The lower degree of significant clustering of variants in rare PEs may indicate the randomization of distributions, possibly due to these members being dispersed randomly in the community, as opposed to occupying a discrete niche within the community.

Comparison With Other Algorithms

The ES2 demarcation results were compared with predictions of other algorithms that share a similar goal of species prediction (Figures 2.5 and 2.6, and Supplemental Figures A.2 and A.3). The maximum-likelihood Poisson Tree Processes (PTP) and the related Bayesian PTP (bPTP) models (Zhang et al., 2013) were compared to the ES1 and ES2 demarcation results. Because each algorithm is dependent upon the phylogenetic tree used, both neighbor-joining trees created with PHYLIP and maximum-likelihood trees created with FastTree were used with all algorithms.

For the most part, the PTP and bPTP demarcation results differed greatly from the ES1 and ES2 results, with PTP generally lumping more and bPTP generally splitting more than ES1 and ES2. There are, however, multiple examples of overlap

in predicted ecotype membership among the five methods tested (see A-like PEA22 (A20) in Figure 2.5B for one example in the neighbor-joining analysis and A-like PEs PEA01 (A′21), PEA02 (A′22), PEA10, PEA11, PEA13, PEA14, PEA15, PEA21 (A20), PEA22, PEA23, and PEA24 in Figure 2.6B for eleven examples in the maximum-likelihood tree). PTP and bPTP detected only one or two of the four PEs focused on in this study, and detected only 7.6 to 46% of all predominant A-like PEs plus PE B′9 detected by ES2.

By comparing the results of all tested algorithms against the results of ES1 coarse- and fine-scale methods with the Variation of Information statistic, one can see that ES2 performs very well at reproducing results similar to ES1 (Table 2.1). ES2 achieved a lower value and, as such, performed better than all other algorithms tested except when comparing results with the ES1 coarse-scale demarcation of the neighbor-joining phylogeny. In this case, the ES1 fine-scale method performed better than all others ($VI_{ES1} = 0.74$ vs. $VI_{ES2} = 1.35$). When using the same neighbor-joining phylogeny for demarcation, but comparing ES2 with the ES1 fine-scale method, ES2 performed better than the ES1 coarse-scale method ($VI_{ES1} = 0.74$ vs. $VI_{ES2} = 0.61$). With a maximum-likelihood phylogeny, ES2 performed better than ES1 fine-scale analysis at reproducing ES1 coarse-scale analysis results ($VI_{ES1} = 1.40$ vs. $VI_{ES2} = 1.26$), and ES2 performed better than ES1 coarse-scale analysis at reproducing ES1 fine-scale analysis results ($VI_{ES1} = 1.37$ vs. $VI_{ES2} = 0.93$). The PTP algorithm only performed better than ES2 when comparing results with the ES1 coarse-scale demarcation of the neighbor-joining phylogeny ($VI_{PTP} = 0.97$ vs. $VI_{ES2} = 1.35$). ES2 performed better than the bPTP algorithm with all comparisons.

Testing the Runtime of ES2

In order to test how likely the changes made to ES2 were to meeting the goal of analyzing a large number of sequences, 5,689 unique A-like *Synechococcus* environmental sequences collected by Becraft et al. (2015) were randomly subsampled at 100-sequence intervals (between 100 and 5600) and ES2 was used on each subsample to predict PEs (Figure 2.10 and Supplemental Figure A.5 for results from B′-like sequences). Although the maximum number of environmental sequences analyzed in this experiment ($n = 5600$) does not fully test the power of ES2, the overall linear nature of the runtime curve (see the blue line in Figure 2.10) and the fact that the analysis didn't exceed 90 minutes, illustrated that analyses of even larger datasets is possible. Indeed, analyses based on much deeper sampling by 454 GS FLX Titanium and Illumina MiSeq sequencing are currently being performed using ES2 (see Chapters 3, 4, and 3). Such analyses would not be possible with ES1.

An interesting side effect of this experiment was that it provided a rarefaction curve for number of PEs that may help offer some insight into how well the diversity present in the environment had been sampled by this study. The total number of PEs predicted by ES2 (see the orange line in Figure 2.10) with a maximum likelihood tree appeared to be leveling off, suggesting that a sample of 5,689 unique sequences revealed nearly the total ecological diversity in A-like *Synechococcus* in this community. The number of PEs with only a single member (see the red line in Figure 2.10) appeared to follow a similar pattern to the total PE count and seemed to be leveling out. Our analyses predicted the existence of about 1000 PEs with more than a single representative. Most of these were rare and it was unclear whether they occupy niches within the system or are merely dispersed into the community.

Conclusions

Ecotype demarcation results produced with ES2 are similar to those produced with ES1 with some cases of differential splitting or lumping that reflect the various changes in the treeing and simulation algorithms. Most importantly, these changes allow for the rapid analysis of large sequence datasets of any sequence length which will permit researchers to explore fine-scale microbial diversity. Analyses based on HFS_{10} provide a clear example of the benefit of deeper coverage of sequence diversity. Since the same snapshot of the phylogeny is used at all stages of ES2, and FastTree is now incorporated for generating that phylogeny, we expect ES2 to provide a higher level of accuracy in ecotype demarcation than ES1 provided.

Although this study only examined the diversity within A/B′-lineage *Synechococcus* populations, we believe that ES2 will be applicable on a large scale because a) ES2 utilizes a theory-based model of speciation that functions on neutral genetic diversity present in the gene segment, gene, or suite of genes analyzed by the program, b) ES1 was shown to predict ecologically distinct ecotypes of *Bacillus* (Koeppel et al., 2008; Connor et al., 2010) and ES2 performs similarly to ES1, and c) because it has been able to predict ecologically distinct ecotypes with ecologically interchangeable members within other phototrophic taxa living in same system studied here (see Chapter 5).

Ecotype Simulation 2 is released under version 2 of the GNU General Public License and can be downloaded for Windows, Linux, and OSX from https://github.com/sandain/ecosim.

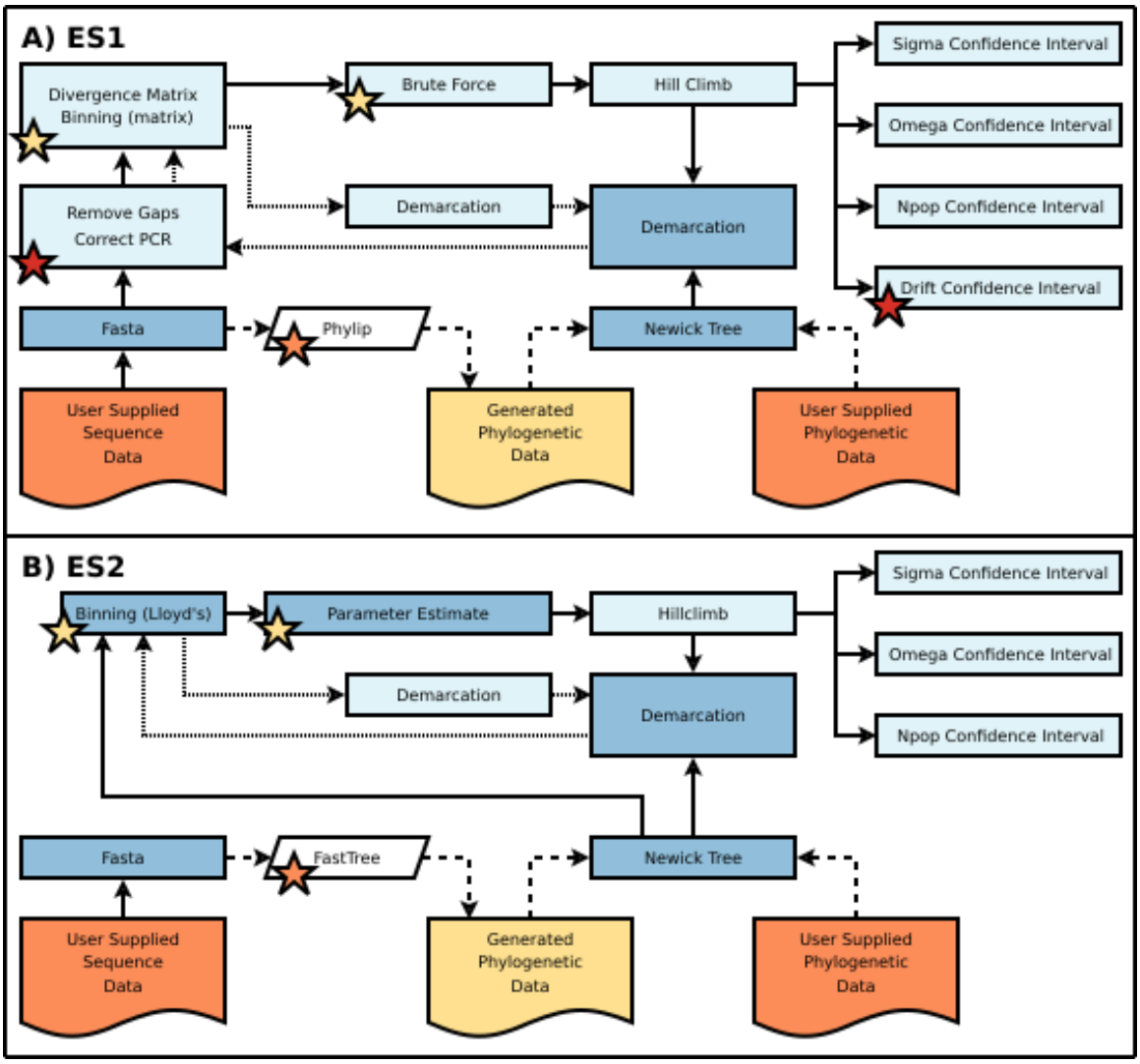


Figure 2.1: Simplified dependency flowchart of (A) Ecotype Simulation 1 (ES1) and (B) Ecotype Simulation 2 (ES2). Orange nodes are user-supplied data, yellow nodes are generated data, white nodes are external programs, dark blue nodes are Java methods, and light blue nodes are Fortran methods. The dashed lines represent an optional route: the user can either supply phylogenetic data in Newick format, or it can be generated using FastTree from the user-supplied sequence data. The dotted lines represent repeated analyses of various branches of the phylogeny during the ecotype demarcation phase. The red stars in A denote Fortran methods no longer used in ES2. The yellow stars denote Fortran methods in ES1 replaced with Java methods in ES2: Parameter Estimate replaces Brute Force, and the tree-based Binning method replaces Divergence Matrix and matrix-based Binning. The orange stars denote the replacement of Phylip in ES1 with FastTree in ES2.

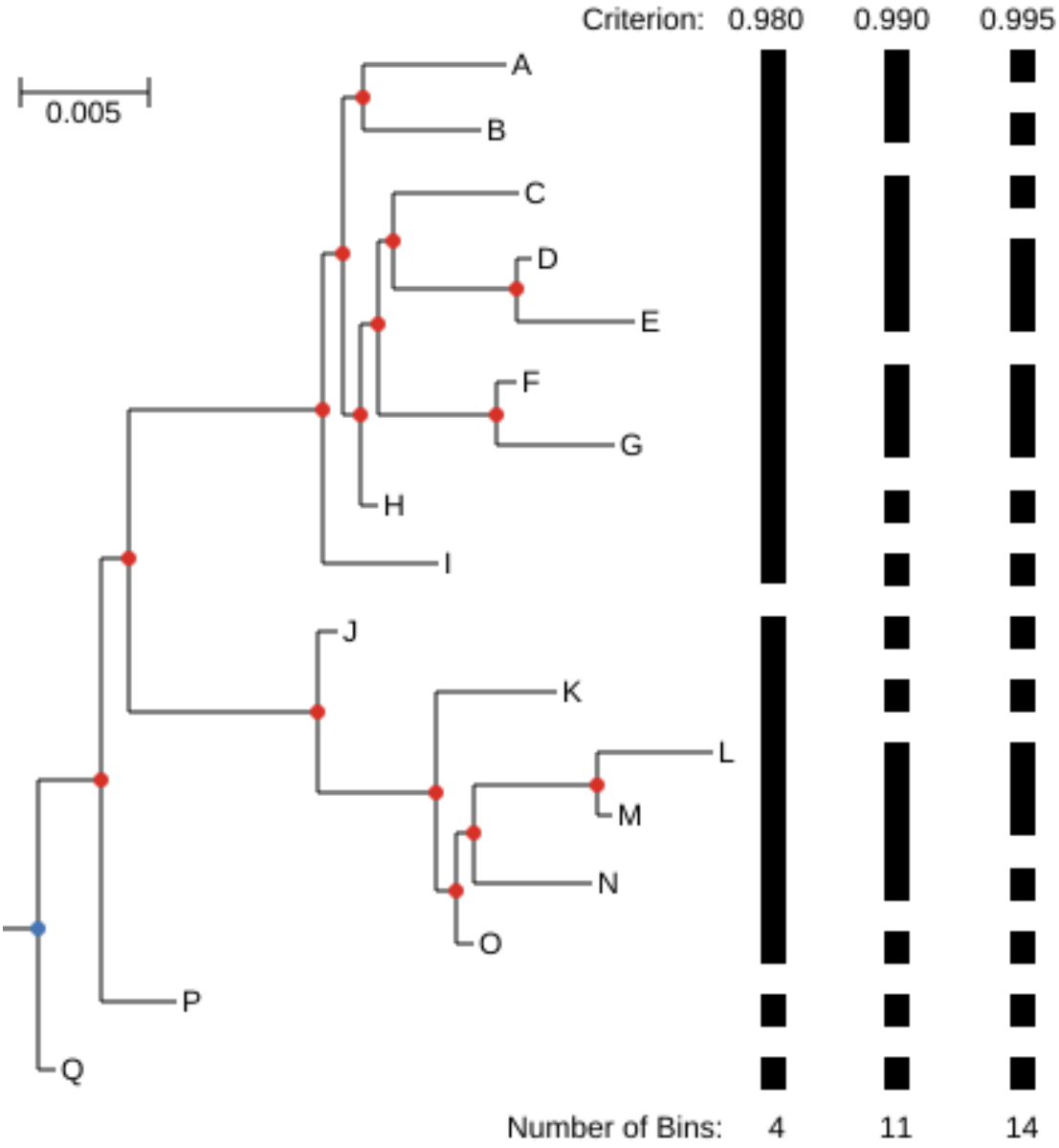


Figure 2.2: Sequence Identity Criterion Binning of a hypothetical lineage. The root node is marked with a blue dot, internal nodes are marked with red dots, and leaf-nodes that represent individual sequences are marked with the letters A-Q. Black bars denote the result of binning the tree at three different sequence identity criterion values with the number of bins listed below. The scale bar represent 0.005 nucleotide substitutions per site.

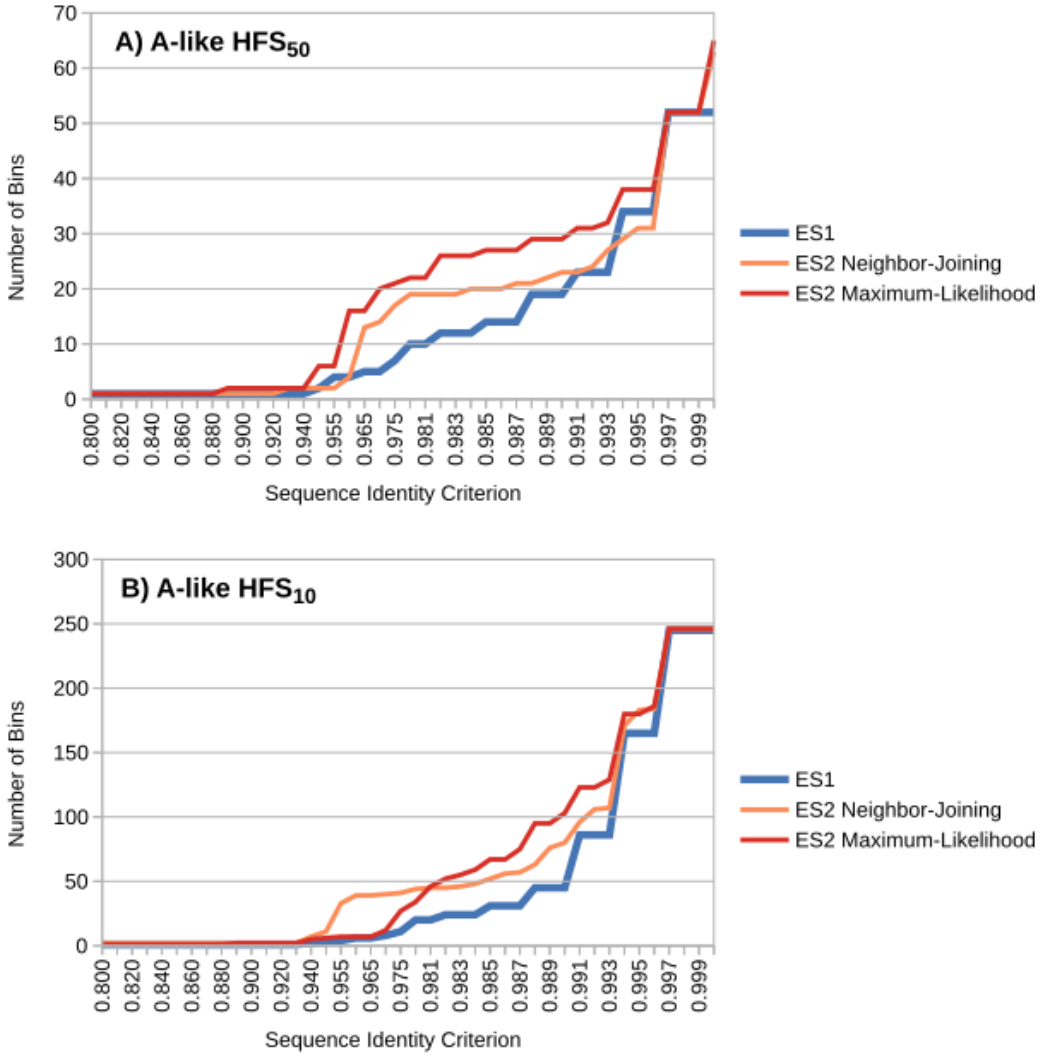


Figure 2.3: Binning results using A-like *Synechococcus psaA* segment high-frequency sequences (HFSs) occurring at least (A) fifty times (HFS_{50}) and (B) ten times (HFS_{10}) in the environmental sequence dataset.

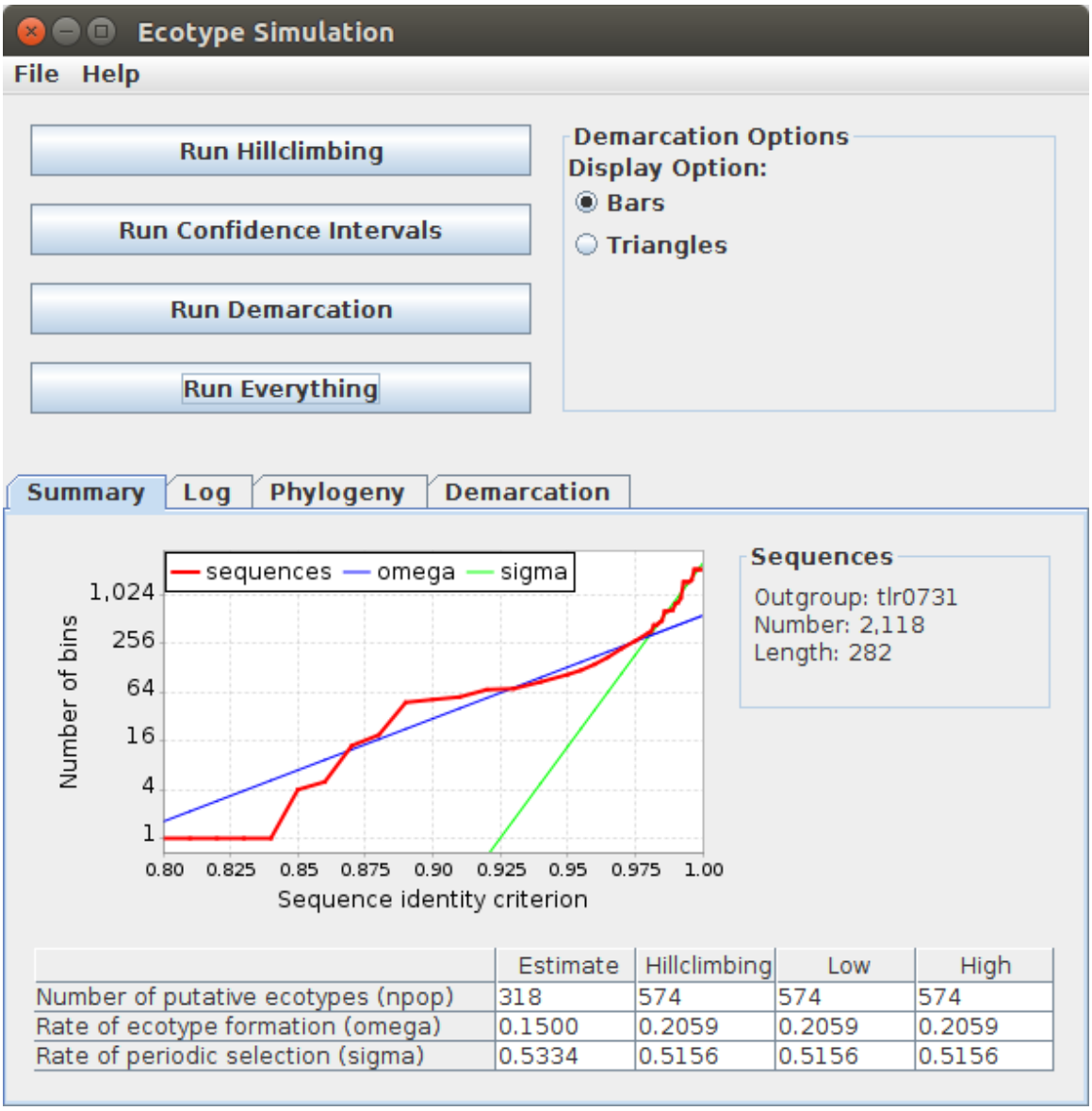


Figure 2.4: Screenshot of the Ecotype Simulation 2 graphical user interface, with the Summary tab displayed.

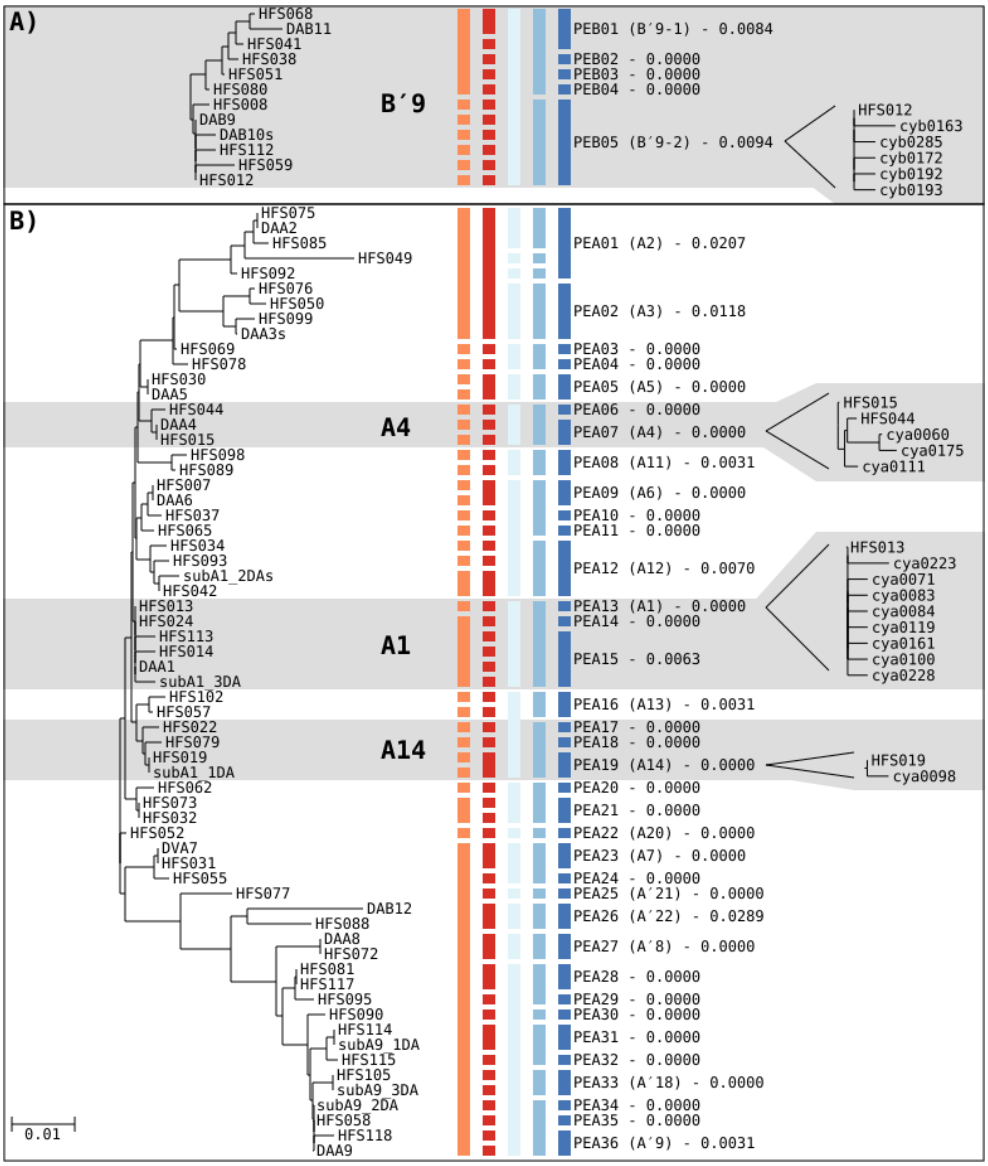


Figure 2.5: Neighbor-joining phylogeny with putative ecotype (PE) demarcation of (A) PE B′9- and (B) A-like *Synechococcus* HFS_{50} environmental *psaA* segments generated using PHYLIP. Gray shading denotes predominant PEs demarcated using Ecotype Simulation 1 (ES1) that are examined in detail in the main text. Colored vertical bars indicate demarcation done by different algorithms, from left to right: orange, maximum-likelihood Poisson tree processes; red, Bayesian Poisson tree processes; light-blue, ES1 coarse-scale; medium-blue, ES1 fine-scale; and dark-blue, Ecotype Simulation 2 (ES2). PEs are labeled with the ES2 demarcation and include the maximum distance among members of the clade. ES2 demarcated PEs that contain the same dominant variant as a PE demarcated by Becraft et al. (2015) using ES1 are indicated in parentheses after the ES2-generated PE names. Portions of the HFS_{10} neighbor-joining tree corresponding to predominant PEs PEB05 (B′9-2), PEA07 (A4), PEA13 (A1), and PEA19 (A14) are included on the right for comparison. The scale bar represent 0.01 nucleotide substitutions per site and is the same in parts A and B.

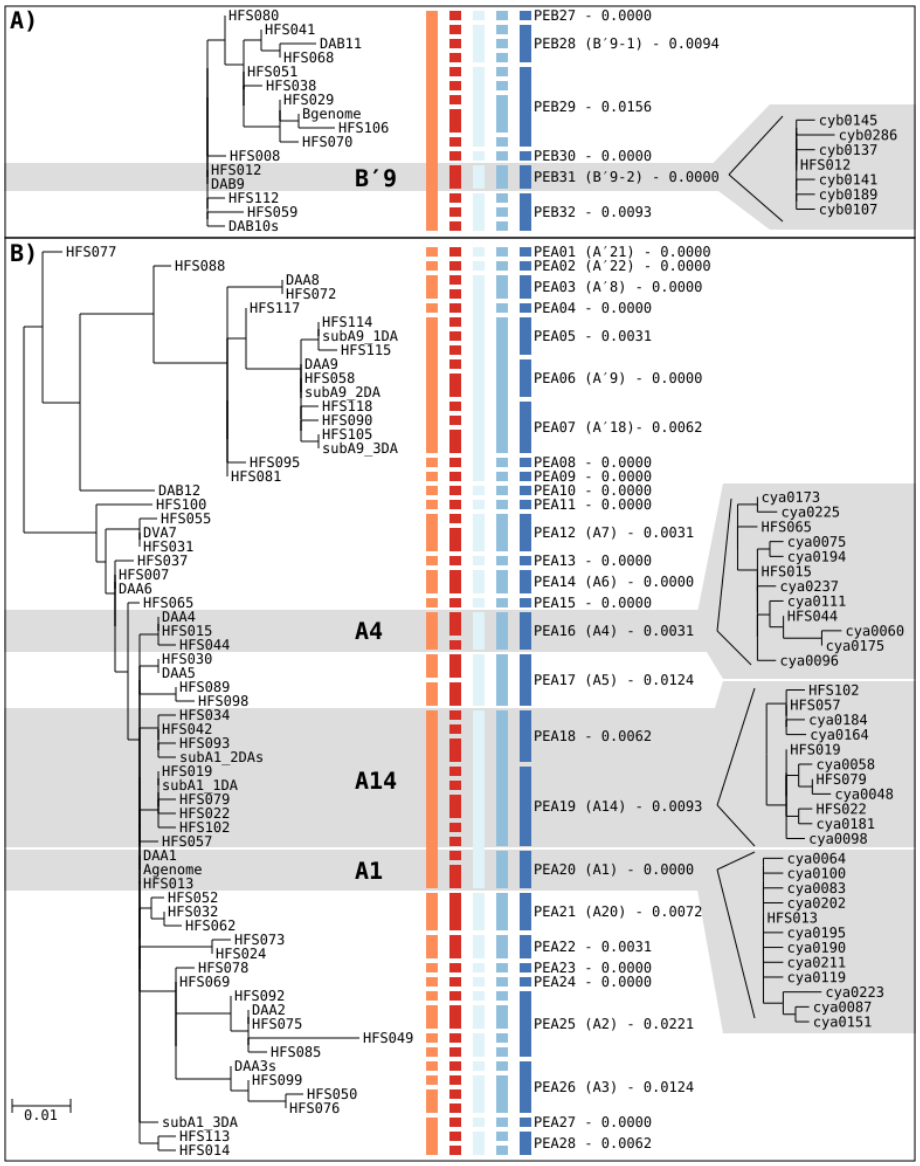


Figure 2.6: Maximum-likelihood phylogeny with putative ecotype (PE) demarcation of (A) PE B′9- and (B) A-like *Synechococcus* HFS_{50} environmental *psaA* segments generated using FastTree. Gray shading denotes predominant PEs demarcated using Ecotype Simulation 1 (ES1) that are examined in detail in the main text. Colored vertical bars indicate demarcation done by different algorithms, from left to right: orange, maximum-likelihood Poisson tree processes; red, Bayesian Poisson tree processes; light-blue, ES1 coarse-scale; medium-blue, ES1 fine-scale; and dark-blue, Ecotype Simulation 2 (ES2). PEs are labeled with the ES2 demarcation and include the maximum distance among members of the clade. ES2 demarcated PEs that contain the same dominant variant as a PE demarcated by Becraft et al. (2015) using ES1 are indicated in parentheses after the ES2-generated PE names. Portions of the HFS_{10} maximum-likelihood tree corresponding to predominant PEs PEB31 (B′9), PEA16 (A4), PEA19 (A14), and PEA20 (A1) are included on the right for comparison. The scale bar represent 0.01 nucleotide substitutions per site and is the same in parts A and B.

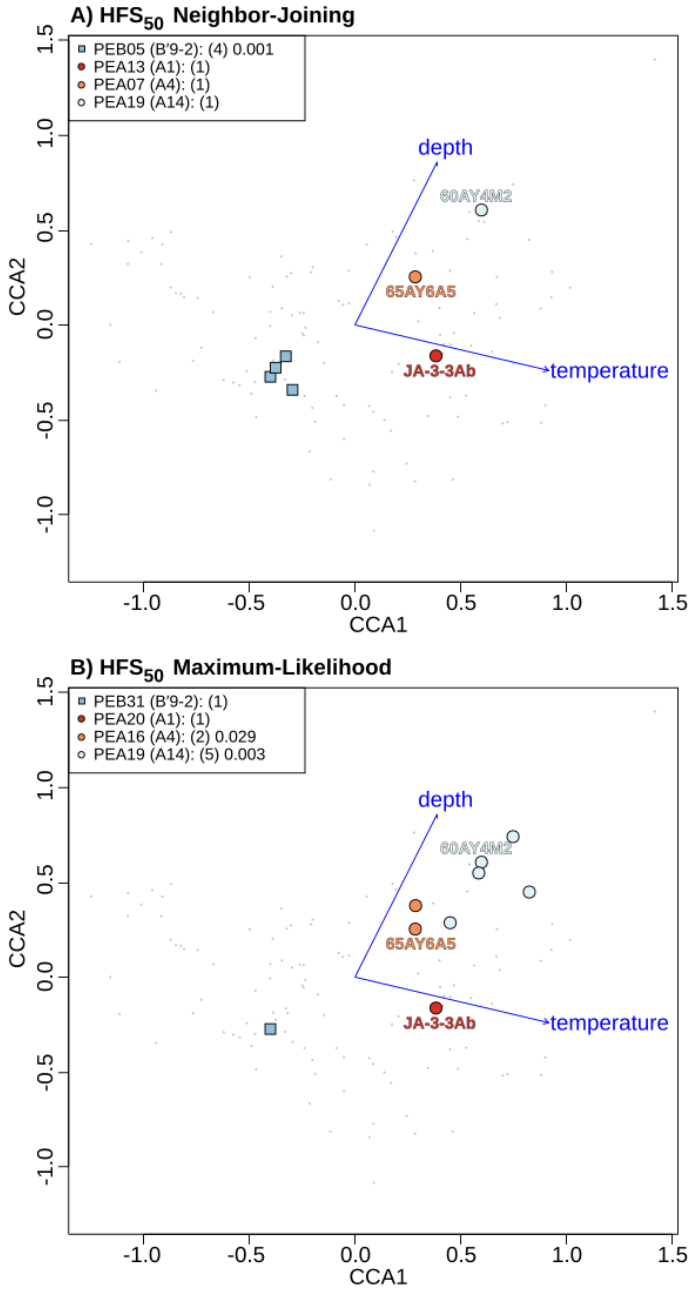


Figure 2.7: Canonical correspondence analysis of putative ecotypes (PEs) demarcated by Ecotype Simualtion 2 using the (A) neighbor-joining and (B) maximum-likelihood trees created from the A- and B′-like *Synechococcus* HFS_{50} environmental *psaA* segments. *Synechococcus* strains JA-3-3Ab, 65AY6A5, and 60AY4M2 share *psaA* sequences with members of PEs PEA13 (A1), PEA07 (A4), and PEA19 (A14), respectively, from the neighbor-joining tree (A) and members of PEs PEA20 (A1), PEA16 (A4), and PEA19 (A14), respectively, from the maximum-likelihood tree (B), and are labeled on each plot.

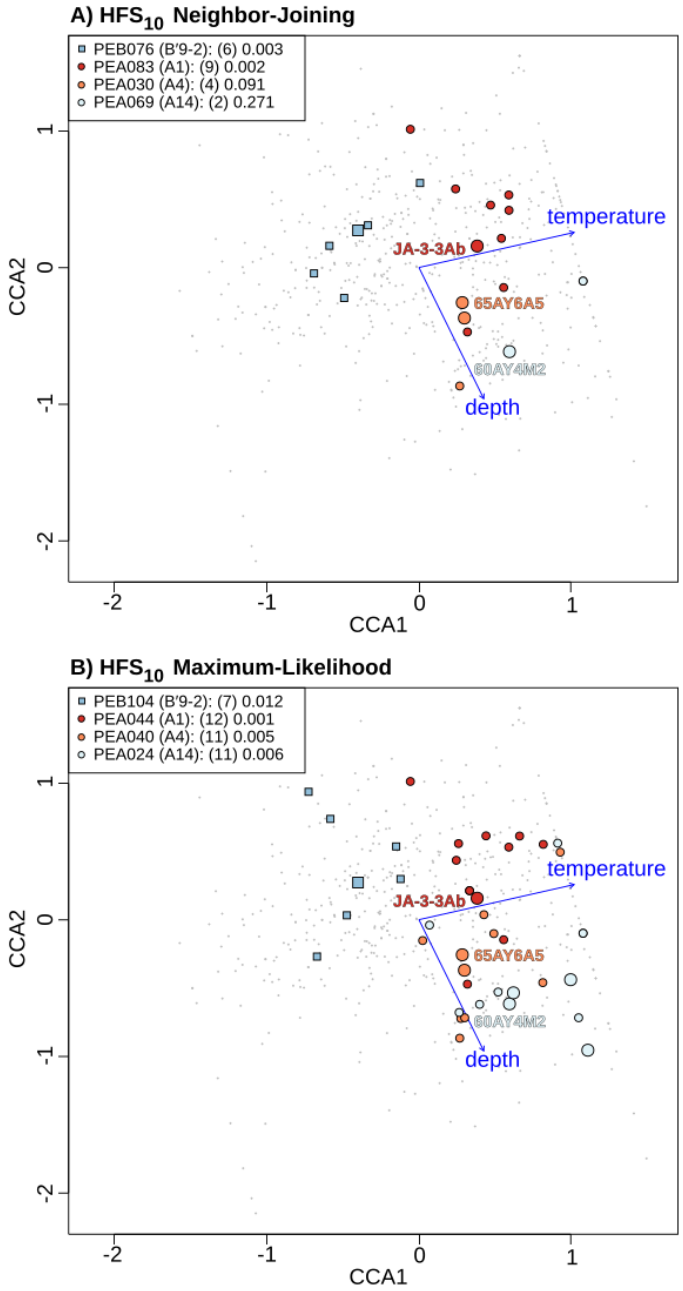


Figure 2.8: Canonical correspondence analysis of putative ecotypes (PEs) demarcated by Ecotype Simulation 2 (ES2) using the (A) neighbor-joining and (B) maximum-likelihood trees created from the A- and B′-like *Synechococcus* HFS_{10} environmental *psaA* seqments. Larger glyphs represent HFS_{50} environmental sequences, while smaller glyphs represent HFS_{10} environmental sequences. *Synechococcus* strains JA-3-3Ab, 65AY6A5, and 60AY4M2 share *psaA* sequences with members of PEs PEA083 (A1), PEA030 (A4), and PEA069 (A14), respectively, from the neighbor-joining tree (A) and members of PEs PEA044 (A1), PEA040 (A4), and PEA024 (A14), respectively, from the maximum-likelihood tree (B), and are labeled on each plot.

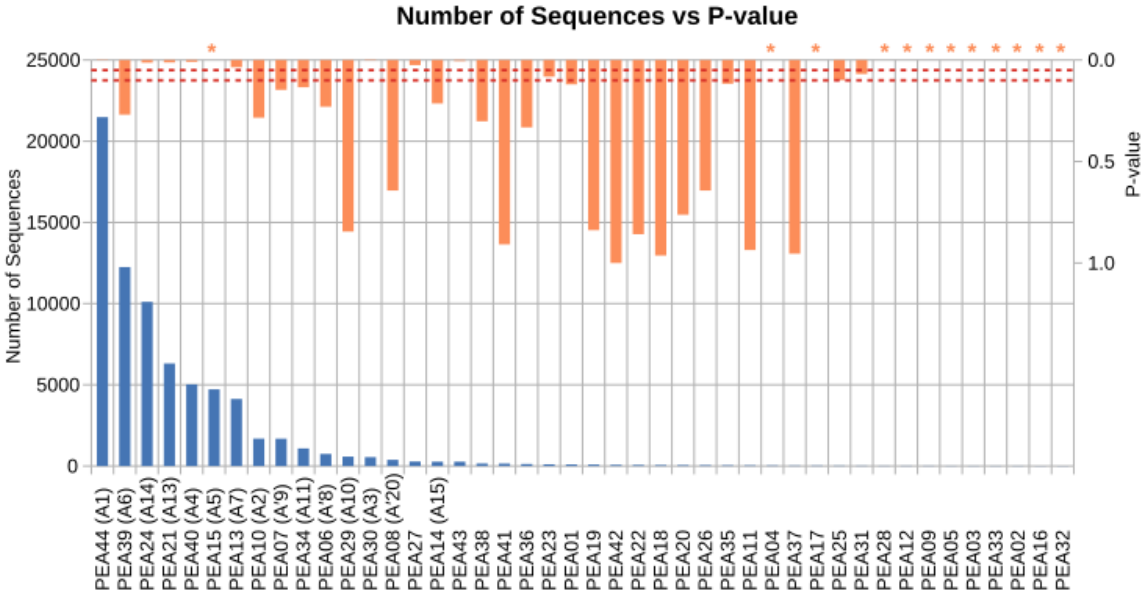


Figure 2.9: *psaA* Sequence abundance of A-like *Synechococcus* HFS_{10} environmental *psaA* segments in putative ecotypes (PEs) demarcated by Ecotype Simulation 2 from the maximum-likelihood tree (blue bars). Orange bars indicate p-values associated with the hypotheses that the members of PEs should not be distributed randomly. Orange asterisks above some columns represent p-values that could not be calculated since the PEs contained only a single member. The red dashed lines represent 0.05 and 0.10 confidence limits.

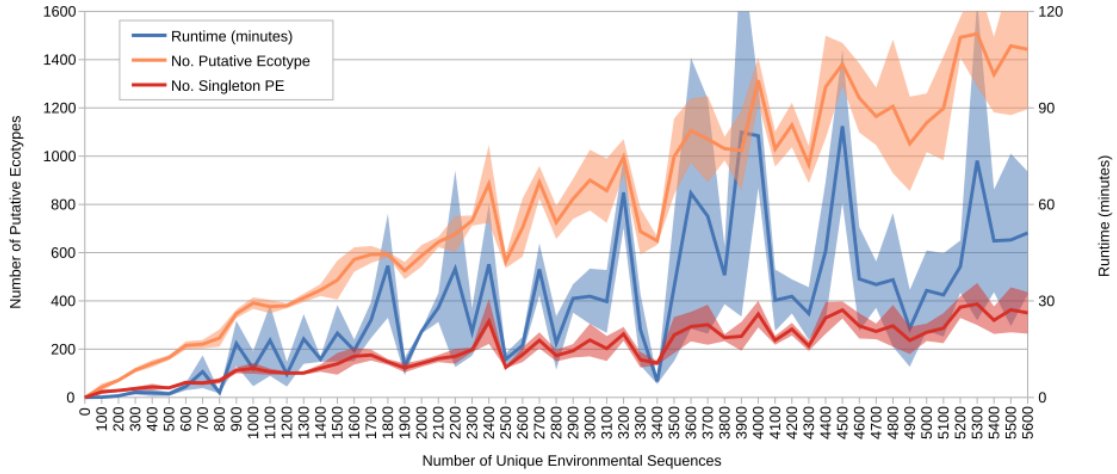


Figure 2.10: Rarefaction and runtime curves produced by repeated subsampling of the 5,689 unique A-like *Synechococcus psaA* segments present in the environmental dataset and subsequent analysis by Ecotype Simulation 2 (ES2). The darker colored lines represent the average and the shaded regions represent the standard error of three trials. The blue line and shaded region demonstrate the amount of time used to run the trial subsample using ES2 (on an Intel Core i7-6700 processor). The orange line and shaded region note the total number of PEs demarcated while the red line and shaded region note the number of PEs with only one member.

Table 2.1: Variation of Information statistic between clustering results provided by the Ecotype Simulation 1 (ES1) coarse- or fine-scale method compared with results provided by the opposite ES1 method, Ecotype Simulation 2 (ES2), maximum-likelihood Poisson tree processes (PTP), and Bayesian Poisson tree processes (bPTP). Lower values indicate less variation between clustering methods compared. Results are provided for both neighbor-joining and maximum-likelihood trees generated from the A- and B′-like *Synechococcus* HFS_{50} environmental *psaA* sequence segments.

	ES1[a]	ES2	PTP	bPTP
Neighbor-Joining				
ES1 coarse-scale	0.74	1.35	0.97	1.83
ES1 fine-scale	0.74	0.61	1.20	1.08
Maximum-Likelihood				
ES1 coarse-scale	1.40	1.26	2.86	1.74
ES1 fine-scale	1.37	0.93	3.36	1.13

[a] Coarse-scale compared to fine-scale or *vice versa*

CHAPTER THREE

BIOGEOGRAPHY OF AMERICAN NORTHWEST HOT SPRING A/B′-LINEAGE *SYNECHOCOCCUS* POPULATIONS

Contribution of Authors and Co-Authors

Manuscript in Chapter 3.

Author: Dr. Eric D. Becraft

Contributions: collected and analyzed barcode sequence data and assisted in preparation of the manuscript.

Co-Author: Jason M. Wood

Contributions: analyzed barcode sequence data using Ecotype Simulation 2 and canonical correspondence analysis and aided in preparation of the manuscript.

Co-Author: Dr. Frederick M. Cohan

Contributions: co-supervised the research and aided in preparation of the manuscript.

Co-Author: Dr. David M. Ward

Contributions: obtained funding for the project, supervised the work, and participated in preparation of the manuscript.

Manuscript Information Page

Eric D. Becraft, Jason M. Wood, Frederick M. Cohan, and David M. Ward

Frontiers in Microbiology

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

Published by Frontiers Media
Submitted January, 2018

Abstract

Island biogeography studies of unicellular cyanobacteria inhabiting hot springs have begun to challenge the notion that "everything is everywhere and nature selects." However, detailed analyses of how diversity is patterned relative to physical separation and physiochemical differences among springs, especially at local to regional scales, have been limited by the low resolution provided by the molecular markers studied. We analyzed large datasets obtained by high-throughput sequencing of a segment of the photosynthesis gene *psaA* from samples collected in hot springs from geothermal basins in Yellowstone National Park, Montana, and Oregon, all known from previous studies to contain populations of A/B′-lineage *Synechococcus*. The fraction of identical sequences was maximal (0-26%) among springs separated by <50 km, while springs separated by 50 km to 110 km rarely shared (0-3%) sequence variants and springs separated by >800 km shared sequence variants extremely rarely (<1%). Phylogenetic analyses provided evidence for endemic lineages that could be related to geographic isolation and/or geochemical differences on regional scales. Ecotype Simulation 2 was used to predict putative ecotypes and their membership and canonical correspondence analysis was used to examine the geographical and geochemical bases for variation in their distribution. Across the range of Oregon and Yellowstone, geographical separation explained the largest percentage of the differences in distribution of ecotypes (9.5% correlated to longitude; 9.4% latitude), with geochemical differences explaining the largest percentage of the remaining differences in distribution (9.3%-7.4% correlated to magnesium, sulfate, and sulfide). However, among samples in the Greater Yellowstone Ecosystem, geochemical differences significantly explained the distribution of ecotypes (6.5%-9.3% correlated to magnesium, boron, sulfate, silicon dioxide, chloride, and pH). Nevertheless, differences in the abundance and membership of ecotypes in Yellowstone springs with similar chemistry suggested that allopatry may be involved even at local scales. *Synechococcus* populations have diverged both by physical isolation and physiochemical differences, and populations on surprisingly local scales have been evolving independently.

Keywords: Biogeography, Microbial Species, Population Genetics, Thermophilic *Synechococcus*

Introduction

Hot spring microbial mats have been used as model systems to demonstrate ecological diversification by sympatric adaptation to parameters that vary along well-established environmental gradients (Ward et al., 2012). For instance, a progression of unicellular cyanobacterial (*Synechococcus*) 16S rRNA genotypes (A″, A′, A, B′, and B, respectively) can be found along a thermal gradient from ∼70 °C to ∼50 °C in alkaline siliceous hot springs in Yellowstone National Park (YNP) (Ferris and Ward, 1997). This patterning led to later studies demonstrating different temperature adaptations of isolates representative of these genotypes (Allewalt et al., 2006). Greater molecular resolution provided by a portion of the *psaA* locus, a major photosystem I reaction center protein, and analyses based on species and speciation theory, were required to demonstrate the existence of more closely related clades distributed to different light environments found at different depths in the upper 1 mm in these mats (Becraft et al., 2015). Isolates representative of *psaA* genotypes uniquely distributed along the vertical aspect of the mat were differently adapted to light intensity and quality (Nowack et al., 2015), and showed differences in gene content that may explain why these adaptations were observed (Olsen et al., 2015).

As hot springs resemble islands, they have also presented an opportunity to understand the role of allopatric processes (i.e., physical isolation) in diversification of indigenous microbial populations (Papke et al., 2003). For many years, it was widely believed that barriers to the distribution of microbial species did not exist, and that "everything is everywhere, and the environment selects" (Baas Becking, 1934). In essence, microorganisms were thought to primarily evolve through sympatric means, and to be distributed globally without geographical barriers limiting their dispersal. Observations leading to such inferences were often based on a low-resolution

taxonomy such as morphology (Finlay and Fenchel, 2004). However, morphology is often a poor indicator of species richness and can mask the genetic and ecological diversity that exists in nature. It can be challenging to study the spatial dynamics of microbial populations at large scale because (i) evolutionarily distinct organisms can share similar morphologies, (ii) cultivated organisms are not always representative of predominant natural populations (Ward et al., 1990), (iii) different taxa can have a variety of evolved dispersal mechanisms (McDougald et al., 2012), and (iv) the immense bacterial diversity that exists in nature can make identifying ecological species in their natural habitats difficult (Dykhuizen, 1998; Ward et al., 2006). Despite these complications, patterns of *Synechococcus* presence or absence in hot springs around the world have suggested a role for dispersal limitations in their biogeography (Castenholz, 1978, 1996).

Molecular technologies have allowed many researchers to infer a role for dispersal limitation in generating biogeographic patterns of microbes (Martiny et al., 2006). Pathogenic and symbiotic microorganisms that are associated with specific eukaryotic hosts have been shown to be restricted to the geographic range of their hosts (Falush et al., 2003; Taylor et al., 2005). Geographic barriers have been shown to isolate populations of hyperthemophilic Archaea (Whitaker et al., 2003; Whitaker, 2006). Thermophilic *Synechococcus* populations have been shown to be differentiated by geography (Papke et al., 2003; Ward and Castenholz, 2002; Ward et al., 2012). These studies provided evidence that geographic isolation has been an important factor in the diversification of microbial populations.

In the case of hot spring *Synechococcus* populations, Papke et al. (2003) showed that *Synechococcus* 16S rRNA genotypes had different distributions in Japanese, New Zealand and North American hot springs (see Chapter 1 Figure 1.13). Analysis of the more rapidly evolving 16S-23S rRNA internal transcribed spacer region suggested

regional variation in community members within Japanese and North American hot springs of different thermal basins, even in springs with very similar chemistry. Papke et al. (2003) characterized diversity using cloning and sequencing methods, so the sampling was limited. Furthermore, these genetic markers have been shown to be too slowly evolving to distinguish the most newly divergent ecological species (or ecotypes) of *Synechococcus* (Melendrez et al., 2011; Becraft et al., 2011, 2015). Additionally, these studies did not take into account how diversity within each hot spring was influenced by differences in the environmental parameters measured. These issues limit our ability to fully understand the relationship between local and regional communities, and thus the roles of dispersal and allopatric and sympatric processes in diversification of *Synechococcus* species.

In this study, we reanalyzed American Northwest samples collected by Papke et al. (2003) by high-throughput sequencing of amplicons of a segment of a more highly-resolving, protein-encoding locus (*psaA*) (Becraft et al., 2011, 2015). We have used Ecotype Simulation 2 (see Chapter 2) to identify the most newly divergent *psaA* segment sequence clusters that can coexist indefinitely either because they are ecologically distinct populations (ecotypes), because they are geographically isolated (geotypes), or both (Cohan and Perry, 2007). This permitted rigorous analysis of the distribution of A- and B′-like *Synechococcus* genetic diversity in hot springs that are separated by <1 to >800 km, with the aim of gaining a better understanding of the role of allopatric processes in the diversification of *Synechococcus*.

Materials and Methods

Samples

Duplicate mat samples and associated water samples for biogeographical analysis were collected from hot springs across the Northwest United States between 30 May

and 8 July 1996 using a #4 cork borer (38.5 mm^2) as reported by Papke et al. (2003) (summarized in Table 3.1 and Figure 3.1). Data from chemical analyses are presented in Table 3.2 (see Papke et al. (2003) for analytical methods). Samples were immediately frozen on dry ice in the field and kept frozen at −80 °C until analysis in 2011. The sampling matrix included replicate springs within four geothermal basins in YNP, two Montana springs near (LaDuke) or ∼150 km (Bozeman Hot Spring) from YNP, and three hot springs in southwest Oregon. Samples used in biogeographical comparisons were collected at ∼50 °C to ∼60 °C and from near-neutral to alkaline pH sites to constrain ecological differentiation among populations, though some local ecological variation was present in the sample set. For instance, (i) samples were collected at different temperatures in Octopus Spring, YNP and Jack's Stream, OR, (ii) samples from Clearwater and Mammoth springs in YNP, and LaDuke Spring had lower pHs than the other springs analyzed (5.2 - 6.9 compared to 8.1 - 8.8), and (iii) samples collected from Mammoth springs and LaDuke Spring had higher concentrations of calcium, magnesium, carbonate and sulfate compared to other Yellowstone springs.

Molecular and Phylogenetic Analyses

DNA was extracted from mat samples, and a segment of the *psaA* locus was PCR-amplified using primers designed to target the *Synechococcus* A/B′-lineage, and sequenced using Ti454-barcoding technology, as described in Becraft et al. (2015). Sequences were trimmed to 302 base pairs to obtain the maximum number of sequences, cleaned and analyzed to identify high-frequency sequences (HFSs; defined here as those with ≥10 identical representatives across all combined samples) using the same process described in Becraft et al. (2015). Sequences similar to those found in publicly available A- and B′-like *Synechococcus* genomes (JA-3-3Ab and JA-2-3B'a(2-

13), respectively) (Bhaya et al., 2007) were split into separate sequence datasets using top-hit BLASTn (Altschul et al., 1990) comparisons with the HFSs found by Becraft et al. (2015). Separate A- and B′-like maximum-likelihood phylogenies were constructed from each sequence dataset with FastTree (Price et al., 2009). Some sample locations (from LaDuke Spring, Bozeman Spring, Bath Lake Vista, Clearwater Spring, Octopus Spring, Perpetual Spring, and Jack's Spring) could not be analyzed in duplicate due to failed sequencing reactions (see Table 3.1 for the number of replicates).

Putative Ecotype Demarcation

To analyze the distribution of HFS variation within ecological populations among chemically similar springs and regions, Ecotype Simulation 2 (ES2) was used to predict putative ecotypes (PEs) from the variation sampled. ES2 uses evolutionary simulation analysis to predict ecologically distinct or geographically isolated clusters in a phylogeny (see Chapter 2). Directions for download and instructions for the use of ES2 are freely available at https://github.com/sandain/ecosim. Because PEs demarcated by ES2 are not guaranteed to have the same HFS membership as previously described by Becraft et al. (2015) from a different dataset, they are named differently here. PEs demarcated in this study that contain members of PEs previously demarcated by Becraft et al. (2015) are indicated by enclosing the previously described PE name in parenthesis after the new PE name assigned by ES2.

Abundance of HFSs Within Predominant PEs

The abundance of HFSs in predominant PEs (defined as those making up >1% of sequences from at least one sample) was compared across environmental samples to measure i) the reproducibility between biological replicates, ii) the differences

between communities in springs of different basins, iii) similarity of communities within different springs within a basin, and iv) the interchangeability of the various HFSs within a PE.

Canonical Correspondence Analyses

The physical and chemical parameters measured at each spring (Tables 3.1 and 3.2) were used as linear predictors in canonical correspondence analyses (CCA) (ter Braak, 1986; Legendre and Legendre, 1998) using software available from the `R` library `vegan` (Oksanen et al., 2013). The `hfs-counter.pl` script (available from https://github.com/sandain/pigeon) was used to count the abundance of HFSs in each environmental sample to provide the data matrix analyzed by CCA. The plotting function used in Chapter 2 was used here to display the distribution of HFSs and predominant PEs in the CCA ordination space. This plotting function reports a p-value for each PE demarcation that represents the probability that the distribution of the PE in ordination space is different from random, and thus provides a test for ecological interchangeability among the members of the PE. PEs with only a single member cannot be tested in this way, so no p-value is provided.

To narrow the list of 18 physical and chemical parameters measured, a customized `R` script adapted from Roberts (2017) was written to find those parameters that significantly ($p < 0.05$) added to the CCA model. This custom script utilizes a forward step-wise approach, starting with the parameter that explained the most variation and stepping through other parameters until no further variation can be explained (`stepCCA.R`; available from https://github.com/sandain/R).

In order to visualize the variation of PEs along a single environmental gradient, a customized `R` script was written to run CCA and perform a weighted-density calculation for each PE in the ordination space. This script utilizes the `R` package

`vegan` (Oksanen et al., 2013) along with the `density` function (`ccadensity.R`; available from https://github.com/sandain/R).

Shared *psaA* Variants with Distance

Geographic distance was calculated from latitude and longitude of all pairwise spring combinations using software available at http://www.movable-type.co.uk/scripts/latlong.html. The number of shared sequences across geographic distance was calculated by identifying all shared genotypes with 100% nucleotide identity present in each pairwise combination of springs across all basins. Sequence counts were summed for individual springs when more than one sample was taken. The percentage of shared sequences for each spring-pair studied was determined by dividing two times the number of all sequences shared between springs by the total number of sequences in both springs.

$$\frac{2 * Shared_{SpringsA\&B}}{Total_{SpringA} + Total_{SpringB}}$$

Springs were arranged by geographic distance from one another (i.e., each data point represents the percentage of sequences in Spring A shared with Spring B and the geographic distance between the two springs, so each of the 15 springs has 14 separate pairwise comparisons for a total of 105 ($\frac{15*14}{2}$) comparative data points).

Results

Samples collected by Papke et al. (2003) from various hot springs in the American Northwest known to contain A/B′-lineage *Synechococcus* (Figure 3.1) were sequenced with Titanium 454 barcoding, resulting in 54,990 *psaA* amplicon sequences (see Table 3.1 for sequence counts from individual springs). The B′-lineage was not detected in Oregon samples, likely due to sequence differences causing inefficient priming. The A-lineage was poorly sampled in Mammoth samples, likely due to the lower

temperatures of the springs sampled (Table 3.2). The degree to which sequence variants were found in more than one spring was compared to the distances separating each pair of springs (Figure 3.2). In general, springs within the same basin (brown, green, pink, red, and purple circles) shared more sequence variants compared to springs from different basins (gray circles). Although the two sites at Clearwater Spring were the closest sampled, they only shared about 9% of sequence variants (pink circle). This may be due to physical and chemical differences between the two sites (Tables 3.1 and 3.2). Oregon springs, which are also close together (brown circles), contained a larger amount of shared sequence variants among springs (about 11-24%). Comparisons among Mammoth hot springs displayed a wide variability (5-25%) in the number of shared sequences, with New Mound and White Elephant Back Springs sharing the highest percentage (green circles). The springs sampled in the Lower Geyser Basin displayed a similar amount of shared sequence variants (about 9-13%). The West Thumb springs displayed the least amount of shared sequence variants (about 4%), possibly owing to the larger distance between springs and/or differences in their chemistry. Comparisons between springs from different basins clustered into two obvious groups: those between springs in YNP (gray circles), and those between Oregon and Montana or YNP springs (open circles). Comparisons between YNP springs displayed some variability in the amount of shared sequences (0-9%), with very little overlap in sequence variants in springs separated by >50 km. Comparisons between Yellowstone and Oregon springs on the other hand displayed almost no overlap in sequence variants. Clearly, there was extreme restriction on the distribution of A/B′-lineage *Synechococcus* among the springs we studied.

Phylogenetic Analysis and Abundances of High-Frequncy Sequences Within PEs

Phylogenetic trees for the *Synechococcus* A- and B′-lineages are shown in Figure 3.3. The shading in the tree compared to the PE demarcation bars to the right of each tree demonstrated that a large number of detected PEs were endemic or nearly endemic ($\geq 90\%$) to a single basin (45 of 66 A-like PEs and 62 of 93 B′-like PEs).

Because each unique sequence was used only once in the phylogeny, it was only when the frequency of HFSs within predominant PEs was taken into account that the degree of endemism of HFSs among different basins and springs could be fully appreciated. Figures 3.4 and 3.5 present the number of occurrences (log scale) of each HFS within predominant PEs detected in each mat sample. With this presentation, it was possible to see the similarity in population and community structure across springs within a basin and the differences in population and community structure among springs in different basins (compare A- and B′' lineages among basins in YNP in Figure 3.4 with those in Montana and Oregon in Figure 3.5).

Comparison of paired samples, especially those with deeper sequence coverage (Mushroom Spring, Twin Butte Vista, Jack Stream and Levee Spring in the A-lineage; White Elephant Back, Clearwater East and Twin Butte Vista in the B′-lineage) suggested the fidelity of the approach to reproducibly sample the populations present. Variants predicted to be within a single PE were often found in all springs within a basin, but some exceptions were noted, including New Mound Spring (differences in variants of PEs PEB85 and PEB88) and Mushroom Spring (differences in PE PEB73 (B′12-2)). In most cases, variability in sampling and the low abundances of HFSs (e.g., <10 sequences of an HFS in a sample) prevented us from making inferences about population genetics within PEs.

Regional Endemism

There were several examples of predominant PEs that were endemic to different regions. For instance, PE PEB49 was found only in Clearwater Springs (YNP), and Oregon PEs PEA43 (A7), PEA55 and PEA57 were endemic to Oregon springs (Figure 3.3). Oregon PE PEA66 would also be endemic, were it not for two HFS variants that were also detected in one of the YNP samples (Heart Pool). As shown by the purple bars in Figure 3.4 (A-lineage, bottom), a single sequence of each of these HFSs was detected in one of the Heart Pool samples, whereas hundreds of sequences of PE PEA66 HFS variants were detected in all Oregon springs. There were several other examples in the dataset, in which HFSs that were abundant in YNP springs were detected rarely (i.e., 1-2 sequences) in Oregon Springs. For instance, rare examples of PEs PEA21 (A1) and PEA30 (A14) HFSs were detected in a Jack Spring High sample, and single sequences of PEs PEB32 (B′9) and PEB80 were recovered from a Levee Spring sample.

Near Endemism Among Greater Yellowstone Ecosystem Basins

Similarly, there were several examples among PEs inhabiting springs in the Greater Yellowstone Ecosystem that were found to be nearly endemic to a single basin. For instance, (i) PE PEA26 appeared endemic to West Thumb springs, except for a single sequence of an HFS variant detected in a New Mound Spring sample from Mammoth Hot Springs, (ii) PE PEB26 appeared endemic to Bozeman Hot Spring, except for a single sequence of an HFS variant detected in a Mantrap Spring sample from West Thumb Basin, (iii) PE PEB32 (B′9) appeared endemic to Lower Geyser Basin springs, except for a single sequence of an HFS variant detected in a West Thumb spring, (iv) PE PEB73 (B′12-2) also appeared endemic to Lower Geyser Basin springs, except for two HFS variants detected at very low quantity in Heart Pool,

Clearwater, and New Mound springs, (v) PE PEB85 appeared endemic to Mammoth springs, except for a single sequence of an HFS variant detected in a Mushroom Spring sample, and (v) PE PEB89 appeared endemic to Mammoth springs, except for the detection of a single HFS variant in Octopus Spring and Clearwater Spring samples. The near endemism of PEs PEA26, PEB32 (B′9), and PEB73 (B′12-2) were especially noteworthy because all other PEs abundant in the Lower Geyser Basin springs were also abundant in the West Thumb springs.

Within-basin Endemism

In general, PEs contained the same HFSs in different springs within the same basin. Exceptions included the absence of (i) PEs PEB66 (B′23) and PEB88 from White Elephant Back Spring in Mammoth Basin, and (ii) PE PEB66 (B′23) from Twin Butte Vista Spring in the Lower Geyser Basin, but these examples suffer from the lack of reproducibility in New Mound Spring paired samples and the relatively poor degree of amplification in Twin Butte Vista samples. Interestingly, the dominant HFS in PE PEB20 (B′12-1) was different in Heart Pool than in Mantrap Spring. Although three of the four samples were low in coverage, the result was reproducible in paired samples.

PEs Reliably Detected in More than One Greater Yellowstone Ecosystem Basin

Some PEs demonstrated a more cosmopolitan distribution and were found at a frequency of >10 per sample in samples from different basins. Notably, A-lineage PEs PEA21 (A1), PEA25, and PEA30 (A14) were found in all Yellowstone basins sampled, and A-lineage PE PEA13 (A′9) was found in all but the samples from Clearwater springs. Similarly, B′-lineage PE PEB20 (B′12-1) could be found in Lower Geyser Basin and West Thumb samples, and B′-lineage PEs PEB88 and PEB93 could be found in samples from Mammoth Hot Springs and Clearwater Springs.

Possible Evidence of Long-distance Dispersal

In a few cases, there was more substantial evidence of possible historical dispersal. For instance, one relatively low-abundance A-like PE, PEA41, was embedded in a clade that was mainly endemic to Oregon (see A-like PEs PEA38-PEA66 in Figure 3.3). This PE contained two HFSs that were found only in the Oregon springs and two HFSs that were found only in Yellowstone springs. Likewise, PE PEB52 shows possible evidence of a potential dispersal event from LaDuke Spring to Bozeman Hot Spring. Note that an HFS which was only found in Bozeman Hot Spring, was embedded within a clade of LaDuke Spring sequence variants (Figure 3.3).

Canonical Correspondence Analyses

Canonical correspondence analyses were run on the data matrix using 3 physical and 15 chemical parameters (Tables 3.1 and 3.2) as potential linear predictors. Spring physical or chemical parameters that added significantly ($p < 0.05$) to the CCA model were: longitude, latitude, magnesium, sulfate, and sulfide (Table 3.3 and Figure 3.6).

When all samples across the American Northwest were analyzed together, we found that longitude and latitude were among the most important parameters (Table 3.3). Given the distance between Oregon and the rest of the environmental samples (>800 km; see Figure 3.1), it was not surprising that longitude and latitude correlated with separation of PEs endemic to Oregon. A-like PEs PEA43 (A7), PEA57, and PEA66 were in very tight clusters with $p < 0.05$ stacked on top of each other on the right side of Figure 3.6. Oregon PEs were well separated from those in and around Yellowstone (see A-like PE PEA26 or B′-like PEs PEB20 (B′12-1), PEB26, PEB32 (B′9), PEB49, PEB73 (B′12-2), PEB85, and PEB93 with $p < 0.05$ that were spread out along the CCA2 axis in Figure 3.6) in the ordination space. Magnesium,

sulfide, and sulfate concentrations separated the PEs in and around Yellowstone, with most PEs with more than one HFS member showing restricted distributions in the ordination space (e.g., A-like PEs PEA25 and PEA30 (A14) with $p < 0.1$, and PE PEA26 with $p < 0.05$; and B′-like PEs PEB20 (B′12-1), PEB26, PEB32 (B′9), PEB49, PEB73 (B′12-2), PEB85, and PEB93 with $p < 0.05$ in Figure 3.6).

To test whether longitude and latitude remained as significant determinants of geographical distribution even within a region, the community data matrix was also analyzed without the Oregon samples. In this case, latitude and longitude were not significant to the CCA model. This resulted in magnesium, boron, sulfate, silicon dioxide, chloride, and pH adding significantly to the CCA model (Table 3.3 and Figure 3.7). Chloride, boron, and pH were associated with the separation of PEs that were shared among the Lower Geyser Basin, Clearwater, and West Thumb basins (e.g., A-like PEs PEA26, PEA30 (A14), and PEA66 $p < 0.01$; and B′-like PEs PEB32 (B′9), PEB49, and PEB73 (B′12-2) $p < 0.05$) from those found in Bozeman and LaDuke hot springs (e.g., B′-like PE PEB26 $p < 0.01$; Figure 3.7). Magnesium, sulfate, chloride, and silicon dioxide separate the predominant B′-like PEs endemic to Mammoth hot springs (e.g., PEs PEB85, PEB88, PEB89, and PEB93) from the Lower Geyser Basin and West Thumb hot springs (Figure 3.7).

Separate CCA analyses of pH provided evidence suggesting that PEs endemic to the most alkaline and acidic springs sampled may be adapted to high and low pH. PEs PEA26 and PEB26, which are endemic to Heart Pool (pH 9.2) and Bozeman Spring (pH 9.8), repectively, form significantly non-random clusters at the high end of the pH range in the ordination space (Figure 3.8A and B). Similarly, PEB49, which is endemic to Clearwater Springs (pH 5.2 to 6.1), forms a significant non-random cluster at the lower end of the pH range. By analyzing the weighted density of these PEs, this pH relationship becomes more apparant (Figure 3.8C and D). PEs PEA26

(pink) and PEB26 (blue) formed distinct density curves on the alkaline end of the pH range, while PE PEB49 (green) formed a distinct density curve on the acidic end of the pH range. Interestingly, PEs PEB88, PEB89, PEB93, and PEB85 formed distinct density curves with minimal overlap and different pH optima near the center of the pH range analyzed.

Discussion

Allopatric processes, like sympatric processes, appear to play a role in the diversification of A/B′-lineage *Synechococcus* inhabiting the hot springs of the American Northwest. Everything does not appear to be everywhere at regional scales – the Yellowstone National Park hot spring "archipelago" has different communities from geographically distinct hot springs in Oregon. Distance between springs serves as a physical barrier that limits dispersal. Many of the predominant PEs found in this study are endemic to a single geothermal basin, with geographical separation explaining the highest degree of differences in populations among basins separated by >50 km.

CCA indicates that magnesium, sulfate, and sulfide serve as good linear predictors of the differences in community composition among all samples. Magnesium, sulfate, boron, pH, chloride, and silicon dioxide serve as good linear predictors when analyzing just the Yellowstone and Montana samples. Although these linear predictors may not all represent metabolic requirements of *Synechococcus* PEs (e.g., SiO_2), they may correlate with important drivers of diversification within springs, such as resource availability or chemical or physical conditions tolerated. For instance, temperature was not identified as a significant parameter in these analyses, though variation along the chloride vector may correlate with temperature, as evaporation along the thermal gradient of the effluent channel causes chloride concentration to

rise (Nordstrom et al., 2005) (note the distribution along the chloride vector of B′-like PEs, which are found further from the source in cooler water, than A-like PEs, which are found closer to the source in warmer water in Figure 3.6). Becraft et al. (2015) previously demonstrated differentiation among PEs based on resource availability and physical conditions within the mat community of Mushroom Spring. Light is quickly attenuated in the microbial mat environment, altering the availability and quality of light available for photosynthesis for subsurface PEs. The concentration of dissolved O_2, H_2, and other ions or chemicals that are important inputs or outputs of *Synechococcus* metabolism also changes with depth over a diel cycle (Revsbech et al., 2016). Though CCA evidence is correlative, the significance of pH in CCA analysis, combined with the endemicity of PEs in the most acid and alkaline springs studied leads to the hypothesis that pH may be another parameter that drives sympatric speciation in these populations.

Overlap in PE structures within and between basins points to these populations serving as metapopulations, with local extinctions possible due to the ephemerality of the hot spring (e.g. springs periodically dry up), interaction with a phage, or any other number of calamities that could plague a bacterium. The ephemerality of some hot springs, particularly those in the Mammoth basin, may provide an excellent resource for testing source-sink dynamics in bacterial communities.

Differences in dominant variants of a PE among nearby springs may allude to independent evolution, or differences in disperal among closely related populations in the springs. B′-lineage PE PEB20 (B′12-1) is found in both Heart Pool and Mantrap Spring in the West Thumb geothermal basin (see Figure 3.4), but the dominant variant of the PE in Mantrap Spring was not detected in Heart Pool and vice versa.

Mechanisms of dispersal may be related to aerosolization of microbes (Bonheyo et al., 2005), transport by one of the Yellowstone brine flies (*Ephydra* spp.) that live

and feed on the mat (Brock et al., 1969), or perhaps by biologists studying hot spring inhabitants as the Geotype plus Boeing model suggests (Cohan and Perry, 2007). Rare dispersal events could b indicative of investigator-mediated contamination or artifacts of crossover between samples during sequencing, or they could simply be natuarl events that happen rarely. The brine fly hypothesis is especially compelling because flies are known vectors for microbial transport (Markus, 1980; Junqueira et al., 2017), and different species of brine flies in YNP are known to distribute differently based on the pH and temperature of the spring (Resh and Barnby, 1984), ensuring that transported microbes are compatible with their new environment. Regardless of the mechanism, evidence is presented here that suggests a history of dispersal among springs within Yellowstone National Park and infrequent recent and historical dispersal events between Yellowstone and Oregon hot springs.

The distribution of ecotypes predicted by Ecotype Simulation fits the expectations of insular biogeography. Springs that are more isolated have communities that are different from springs that are near each other. Our data suggests there are PEs that are endemic to a single basin and PEs that are more cosmopolitan. PEs endemic to a single basin may have specialized to the specific chemical environment provided by the water source for the springs in a basin, or may simply have been unable to emigrate and become established elsewhere. More cosmopolitan PEs may be generalists able to tolerate a variety of conditions, or may simply live in a position, such as the top of the mat, that may allow for easier emigration between springs and subsequently basins.

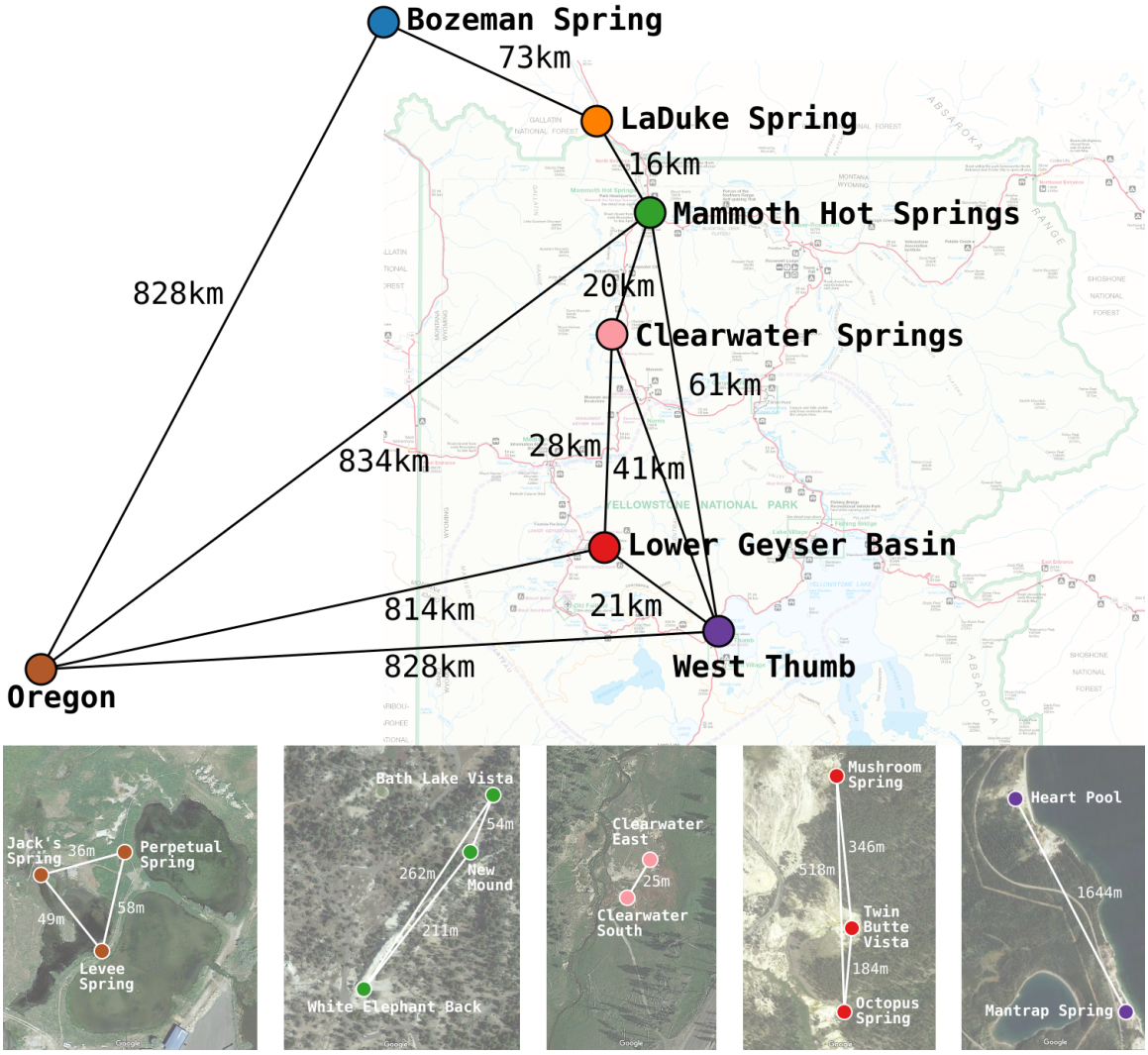


Figure 3.1: Approximate locations and geographic separation (km) among the 7 geographic regions studied across the Northwest United States by Papke et al. (2003). Satellite imagery insets show the approximate locations and geographic separation (m) among springs within Yellowstone National Park and Oregon basins. The satellite imagery was produced using Google Maps. Colors used for basins correspond between the map, satellite imagery, and later figures.

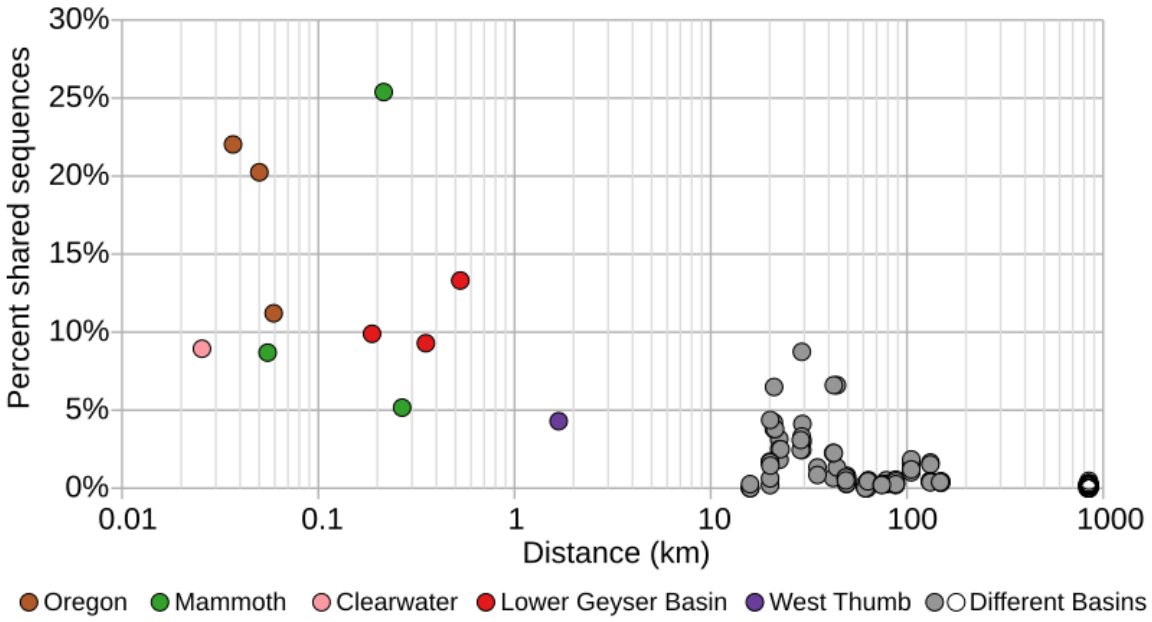


Figure 3.2: Percentage of identical shared *psaA* segment variants relative to separation between pairs of springs sampled. Each circle represents a pair of springs, with colored circles representing pairs from within the same basin. Gray circles represent pairs of springs from different basins within Yellowstone, while open circles represent comparisons between springs in Oregon with springs in basins in and around Yellowstone.

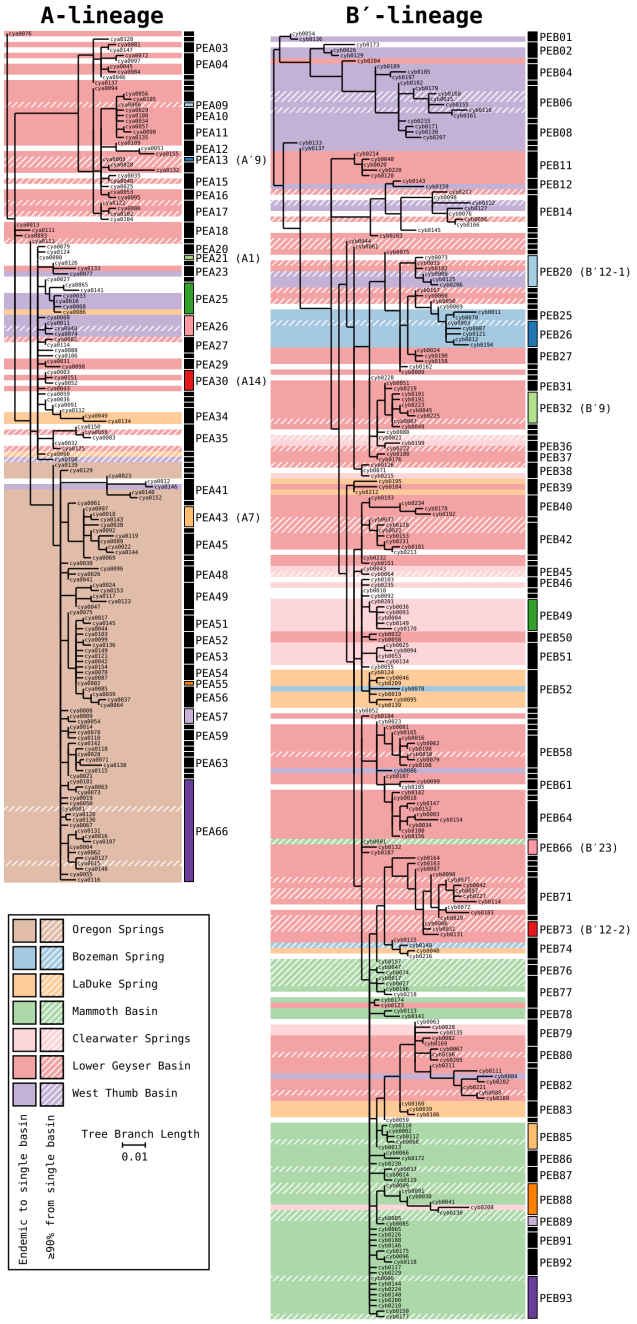


Figure 3.3: Phylogenies of A- and B′-like *Synechococcus* based on partial *psaA* gene segments from Yellowstone National Park, Montana, and Oregon Springs, with Ecotype Simulation 2 (ES2) putative ecotype (PE) demarcations. Shaded regions of the trees mark branches that are found in a single basin. Regions with a diagonal hatch mark branches predominantly found in a single basin (>90%). The ES2 PE demarcations are displayed next to each tree as vertical black and colored bars, with colored bars representing predominant PEs analyzed in detail. Demarcation colors are reused between the A- and B′-like phylogenies and are the same in later figures. Labels are skipped for most PEs with only a single member.

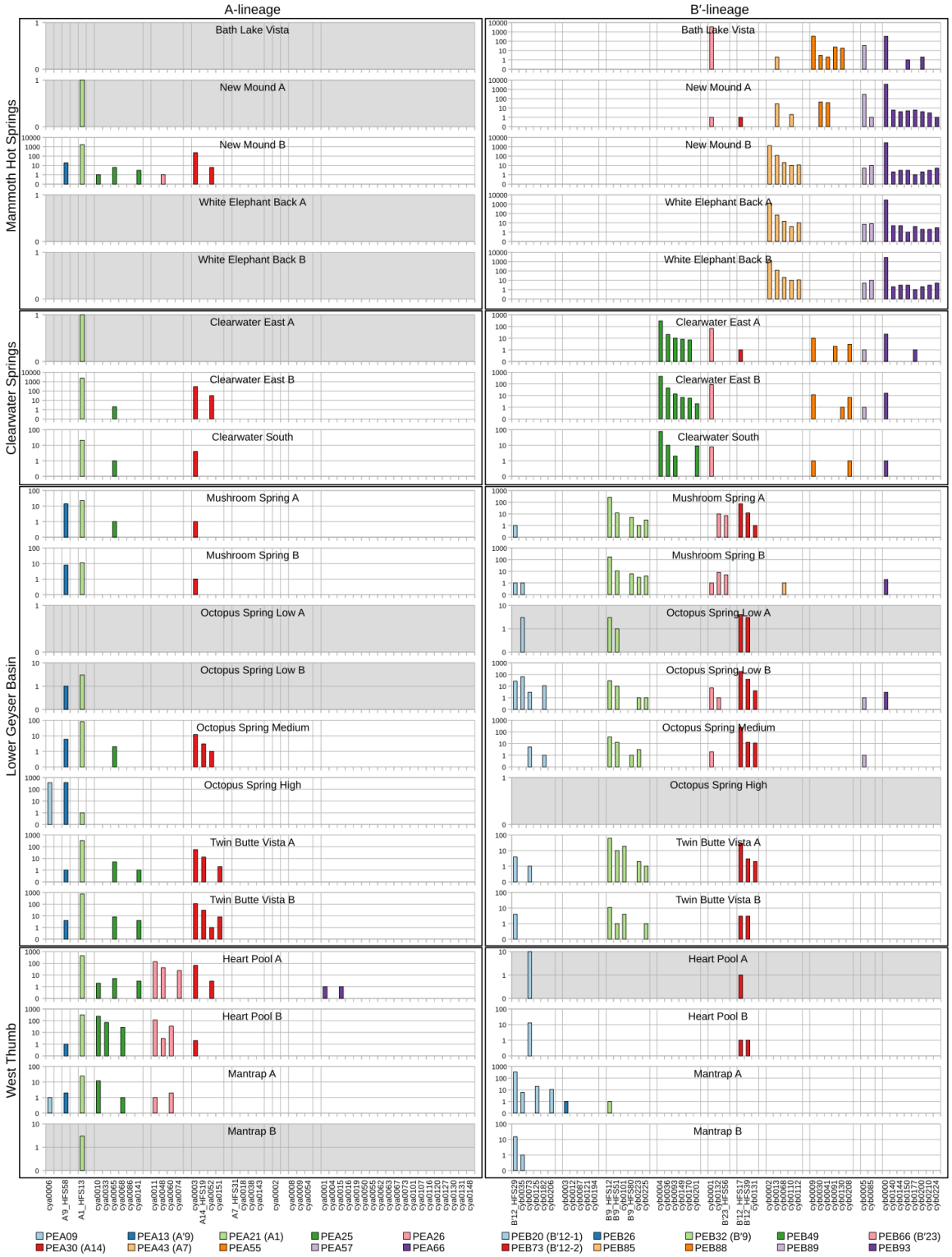


Figure 3.4: Abundance of high-frequency *psaA* sequence segments in predominant *Synechococcus* A- and B′-like putative ecotypes (PEs) for springs within Yellowstone National Park. Poorly sampled springs are shaded gray.

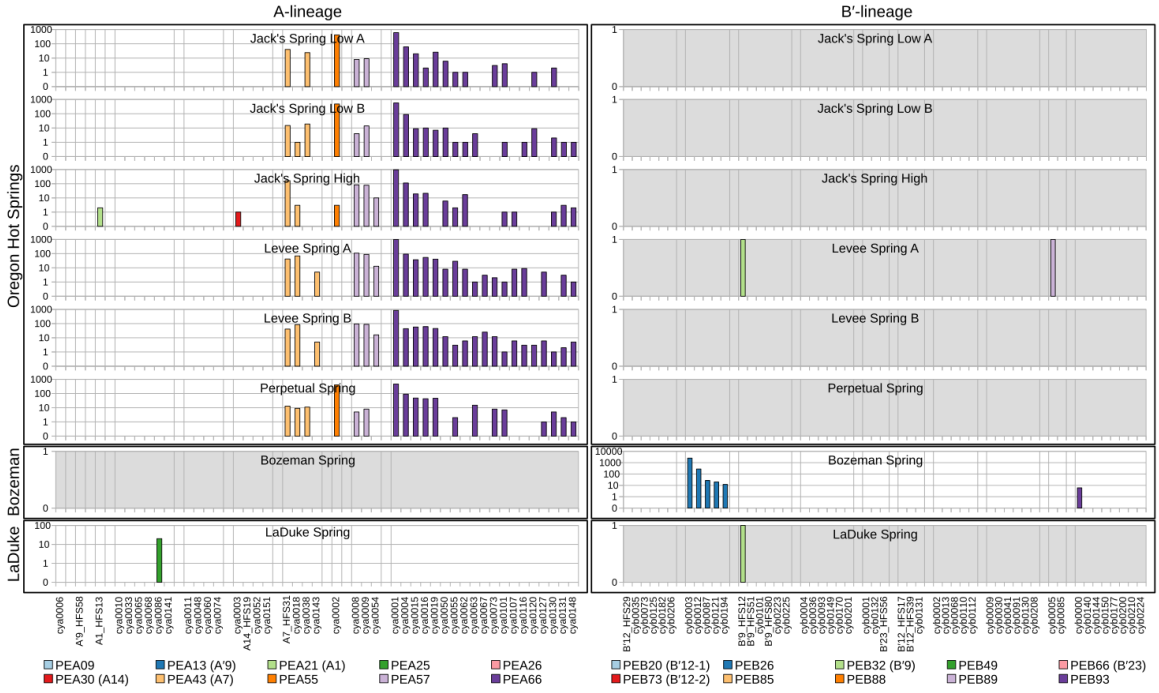


Figure 3.5: Abundance of high-frequency *psaA* sequence segments in predominant *Synechococcus* A- and B′-like putative ecotypes (PEs) for springs outside of Yellowstone National Park. Poorly sampled springs are shaded gray.

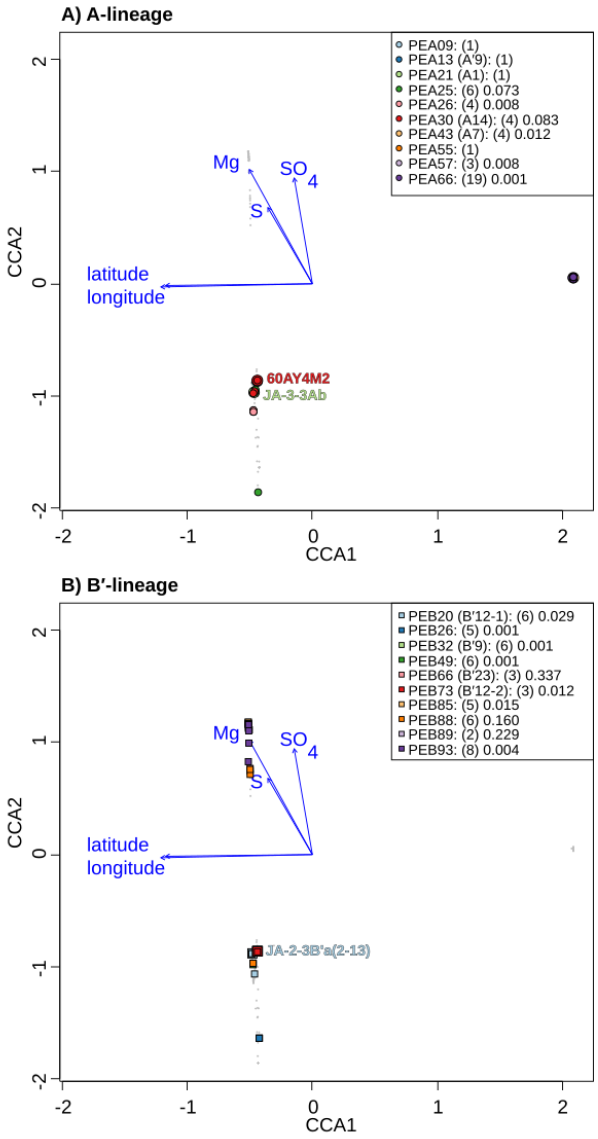


Figure 3.6: Canonical correspondence analyses of *Synechococcus* A- and B′-like *psaA* sequence segment diversity recovered from Yellowstone National Park, Montana, and Oregon hot springs. Larger glyphs represent sequences described previously by Becraft et al. (2015). *Synechococcus* strains JA-3-3Ab, 60AY4M2, and JA-2-3B'a(2-13) share *psaA* sequence segments with high-frequency sequences (HFSs) in putative ecotypes (PEs) PEA21 (A1), PEA30 (A14), and PEB20 (B′12-1), respectively, and are labeled on each plot. Small gray dots represent HFSs from lower-abundance PEs or from the other lineage. Directional arrows represent the vector of influence of each of the significant parameters on the ordination space. PE names in the legend are followed by the number of HFSs making up the PE in parenthesis and a p-value that represents the probability that the observed PE cluster is distributed differently than random.

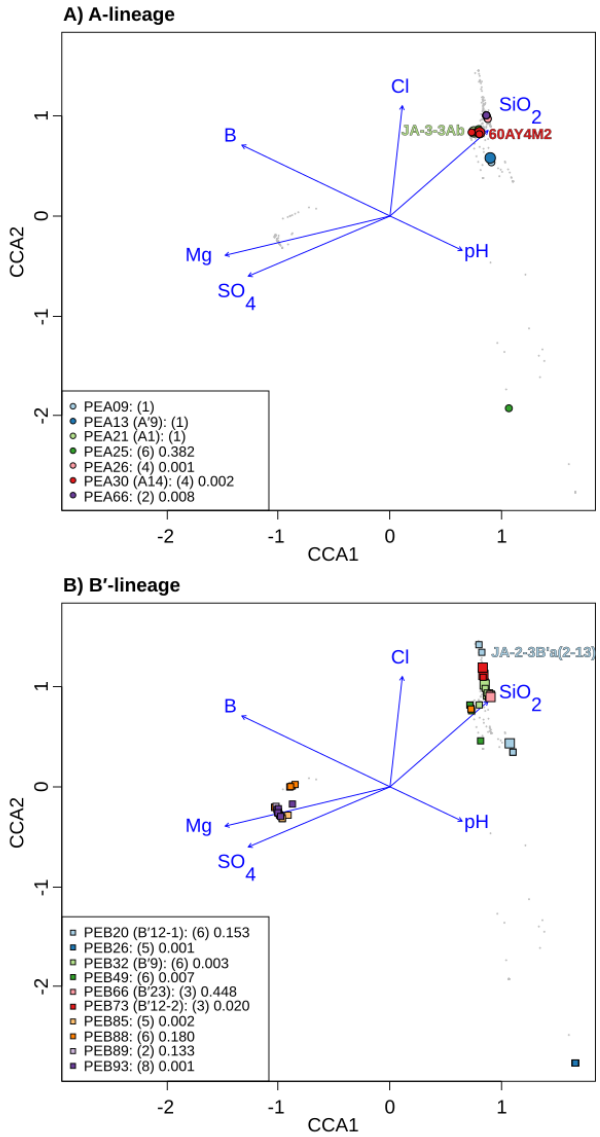


Figure 3.7: Canonical correspondence analyses of *Synechococcus* A- and B′-like *psaA* sequence segment diversity recovered from Yellowstone National Park and Montana hot springs. Larger glyphs represent sequences described previously by Becraft et al. (2015). *Synechococcus* strains JA-3-3Ab, 60AY4M2, and JA-2-3B'a(2-13) share *psaA* sequence segments with high-frequency sequences (HFSs) in putative ecotypes (PEs) PEA21 (A1), PEA30 (A14), and PEB20 (B′12-1), respectively, and are labeled on each plot. Small gray dots represent HFSs from lower-abundance PEs or from the other lineage. Directional arrows represent the vector of influence of each of the significant parameters on the ordination space. PE names in the legend are followed by the number of HFSs making up the PE in parenthesis and a p-value that represents the probability that the observed PE cluster is distributed differently than random.

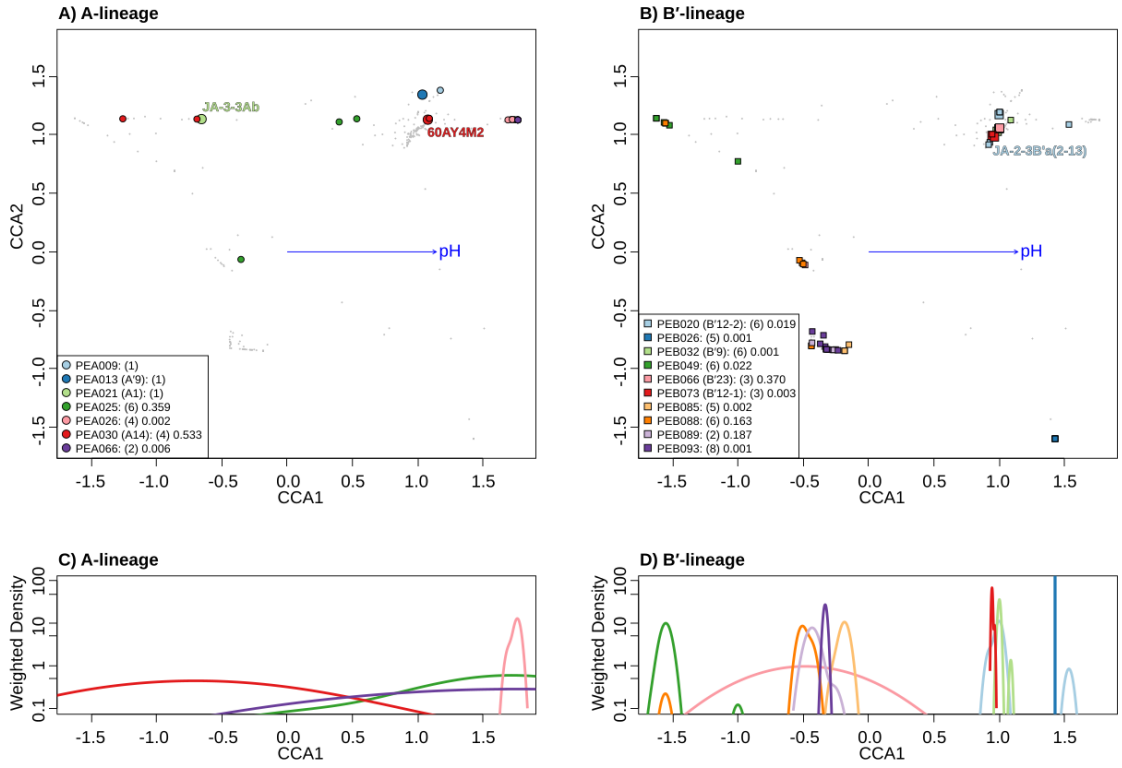


Figure 3.8: Canonical correspondence analyses (CCA) relative to pH of *Synechococcus* A- (A) and B′-like (B) high-frequency *psaA* sequence segments within predominant putative ecotypes (PEs) recovered from Yellowstone National Park and Montana hot springs. Weighted density of predominant A- (C) and B′-like (D) PEs in the ordination space defined by the pH vector. Larger glyphs in A and B represent sequences described previously by Becraft et al. (2015). *Synechococcus* strains JA-3-3Ab, 60AY4M2, and JA-2-3B'a(2-13) share *psaA* sequence segments with high-frequency sequences (HFSs) in putative ecotypes (PEs) PEA21 (A1), PEA30 (A14), and PEB20 (B′12-1), respectively, and are labeled in A and B. Small gray dots in A and B represent HFSs from lower-abundance PEs or from the other lineage. Directional arrows represent the vector of influence of pH on the ordination space in A and B. PE names in the legend are followed by the number of HFSs making up the PE in parenthesis and a p-value that represents the probability that the observed PE cluster is distributed differently than random.

Table 3.1: Physical parameters for hot springs sampled across the Northwestern United States (modified from Papke et al. (2003)). Springs are separated by geothermal basin.

Spring	°C	pH	Latitude	Longitude	Collection Date	Replicates	Sequences
Oregon Springs							
Jack's Spring Low	49.7-50.2	8.4	42.221355	-120.369681	7-Jul-96	2	1,854
Jack's Spring High	59.8-61.2	8.3	42.221355	-120.369681	7-Jul-96	1	3,103
Levee Spring	59.5-61.8	8.1	42.220972	-120.369397	7-Jul-96	2	3,728
Perpetual Spring	60.1-60.9	8.2	42.221387	-120.369242	7-Jul-96	1	1,599
Bozeman Spring	57.2-57.5	8.8	45.660428	-111.187830	8-Jul-96	1	3,492
LaDuke Spring	59.8-60.2	6.7	45.095359	-110.782927	30-May-96	1	687
Mammoth Hot Springs							
Bath Lake Vista	57.0-57.3	6.6	44.964979	-110.711877	30-May-96	1	4,519
New Mound	56.4	6.6	44.964523	-110.712100	30-May-96	2	8,667
White Elephant Back	59.7-60.0	6.9	44.963037	-110.713763	30-May-96	2	8,651
Clearwater Springs							
Clearwater East	57.3-61.9	5.2	44.788868	-110.738900	30-May-96	2	6,468
Clearwater South	56.3-56.5	6.1	44.788670	-110.739061	30-May-96	1	618
Lower Geyser Basin							
Mushroom Spring	58.9-60.7	8.3	44.538751	-110.798180	31-May-96	2	2,146
Octopus Spring Low	51.0-54.0	8.2	44.534103	-110.797846	31-May-96	2	1,070
Octopus Spring Med	54.0-64.0	8.2	44.534103	-110.797846	31-May-96	1	1,090
Octopus Spring High	64.0	8.5	44.534103	-110.797846	31-May-96	1	1,592
Twin Butte Vista	59.3-60.7	8.4	44.535708	-110.797268	31-May-96	2	2,628
West Thumb							
Heart Pool	56.6-58.3	9.2	44.433204	-110.582758	31-May-96	2	2,270
Mantrap Spring	58.2-61.8	8.3	44.420275	-110.572711	31-May-96	2	808

Table 3.2: Chemical parameters for hot springs sampled across the Northwestern United States (modified from Papke et al. (2003)). Springs are separated by geothermal basin. Units for all columns are $mg\,L^{-1}$ unless otherwise noted.

Spring	S^{2-} [a]	CaCO	Ca	Cl	Mg	SiO_2	Na	SO_4	K	As	B	Fe	Mn	Zn
Oregon Springs														
Jack's Spring Low	0.7	65	9	130	0.5	74	200	280	9	0.015	7.5	0.015	0.005	0.015
Jack's Spring High	2.5	64	9	130	0.5	73	190	270	8	0.15	7.4	0.015	0.005	0.015
Levee Spring	5.5	72	10	120	0.5	71	190	260	9	0.13	7.2	0.015	0.005	0.015
Perpetual Spring	4.6	64	10	120	0.5	73	190	260	9	0.15	7.4	0.015	0.005	0.015
Bozeman Spring	7.7	87	1	53	0.5	34	130	130	2	0.11	0.6	0.015	0.005	0.015
LaDuke Spring	0.0	230	340	49	59	23	220	1300	24	0.2	0.93	0.19	0.005	0.015
Mammoth Hot Springs														
Bath Lake Vista	39.3	720	330	170	63	27	120	600	55	0.94	4.7	0.015	0.005	0.015
New Mound	25.4	720	350	160	65	27	120	600	54	0.82	4.3	0.07	0.005	0.015
White Elephant Back	8.20	610	330	170	67	30	120	670	54	0.75	4.4	0.22	0.01	0.07
Clearwater Springs														
Clearwater East	1.3	8	6	130	0.5	40	74	26	16	0.68	2.7	0.53	0.04	0.015
Clearwater South	0.2	31	7	140	0.5	45	94	33	20	0.67	2.8	0.14	0.1	0.015
Lower Geyser Basin														
Mushroom Spring	0.0	250	0.5	270	0.5	110	270	22	19	1.7	3	0.015	0.005	0.015
Octopus Spring Low	0.0	300	0.5	310	0.5	120	300	21	15	1.5	3.5	0.015	0.005	0.015
Octopus Spring Med	0.1	290	0.5	290	0.5	110	300	26	15	1.4	3.3	0.015	0.005	0.015
Octopus Spring High	0.1	280	0.5	270	0.5	100	300	23	15	1.4	3.3	0.015	0.005	0.015
Twin Butte Vista	0.6	270	0.5	320	0.5	110	290	21	13	2.1	3.3	0.015	0.005	0.015
West Thumb														
Heart Pool	19.5	480	0.5	350	0.5	120	410	44	23	2.1	4.3	0.015	0.005	0.015
Mantrap Spring	0.0	350	0.5	180	0.5	94	270	29	12	1	2.5	0.015	0.01	0.015

[a] $\mu mol\,L^{-1}$

Table 3.3: Analyses of constrained variability and significance of parameters ($p < 0.05$) in the canonical correspondence analyses model.

Parameter	Constrained Variability	Significant	Significant w/o Oregon Samples
longitude	9.46%	Yes	No
latitude	9.42%	Yes	No
Mg	9.26%	Yes	Yes
K	9.26%	No	No
Ca	9.24%	No	No
B	9.23%	No	Yes
SO_4	9.10%	Yes	Yes
CaCO	8.61%	No	No
SiO_2	7.84%	No	Yes
As	7.62%	No	No
S^{2-}	7.37%	Yes	No
Cl	7.03%	No	Yes
pH	6.51%	No	Yes
Na	6.34%	No	No
Mn	6.23%	No	No
Fe	5.66%	No	No
Zn	5.13%	No	No
temperature	3.95%	No	No

CHAPTER FOUR

TEMPORAL STABILITY OF *SYNECHOCOCCUS* ECOLOGICAL POPULATIONS IN MUSHROOM SPRING AND OCTOPUS SPRING MAT COMMUNITIES

Introduction

Microbial populations have been shown to change seasonally in relation to fluctuating environmental conditions (DuRand et al., 2001; Thompson et al., 2004; Giovannoni and Vergin, 2012). These seasonal changes could include i) population increase or decrease, ii) population redistribution, or iii) evolutionary changes such as changes in population genetic structure.

In the Ward Lab's work on Mushroom Spring (MS) and Octopus Spring (OS) cyanobacterial mats, which have been extensively studied for over 30 years, the systems have mostly been treated as if they were static. Yet, light availability changes over the diel cycle, and total maximum light intensity, day length, and available light changes seasonally. The day length is considerably shorter in winter months, when the 24 h average irradiance is about 25% of that in the summer. To determine if populations in these mats changed in response to changing environmental conditions, Ferris and Ward (1997) analyzed the population diversity at various temperatures along the thermal gradient in OS seasonally using denaturing gradient gel electrophoresis of partial 16S rRNA gene segments. They found that, despite differences in light availability, *Synechococcus* populations did not change seasonally. In all seasons, a progression of 16S rRNA genotypes A″, A′, A, B′, and B was observed between ∼70 °C and ∼50 °C (see Chapter 1 Figure 1.4). However, the study was complicated by the fact that the 16S rRNA molecule lacks the molecular resolution

needed to accurately demarcate ecological species (Melendrez et al., 2011; Becraft et al., 2011, 2015). Analyses of a segment of the more rapidly evolving *psaA* gene, which encodes an important photosystem I reaction-center protein, combined with demarcation of species using a theory-based algorithm (Ecotype Simulation), revealed the existence of ecological species (or putative ecotypes; PE) of *Synechococcus* that distribute differently within the 1 mm-thick upper green layer of the mat. At 60 °C to 63 °C, a progression of PEs B′9, A1, A4, A6, and A14 was observed from the top of the mat to the lowest part of the ∼1 mm-thick upper green cyanobacterial layer. Isolates representative of these species were obtained and shown to exhibit different light adaptations and acclimation capability (Nowack et al., 2015) and the probable underlying genetic basis for this was discovered in comparative genomic analyses of the same isolates (Olsen et al., 2015). Such observations might suggest that in winter, populations adapted to lower or higher light intensity, might alter their abundances and depth distributions.

The work of Becraft et al. (2015) was limited by the inability of Ecotype Simulation to analyze most of the data collected in a high-throughput sequencing study based on 454 Titanium barcoding. Furthermore, previous studies of temporal datasets (Becraft, 2014) did not include vertical profiles or analysis of a broad range of temperature and often did not include replication in sampling. The development of an improved version of the algorithm called Ecotype Simulation 2 (ES2), described in Chapter 2, which can rapidly analyze very large datasets and recover the same depth distribution patterns as observed by Becraft et al. (2015), enabled me to revisit the issue of temporal stability in a microbial community. Here, much larger sequence datasets produced using Illumina MiSeq technology were analyzed in order to compare winter and summer temperature distributions. Smaller datasets produced with 454 Titanium barcoding were also analyzed to study whether populations changed over

more than two decades.

Methods

Environmental Sampling

The microbial mat in the major effluent channel of Mushroom Spring, an alkaline siliceous hot spring in Yellowstone National Park, WY (44.5387N, 110.7979W) was sampled on January 16 and July 17, 2014. Replicate samples of the microbial mat were collected from the center of the flow channel using a #4 cork borer (38.5 mm^2) at temperature-defined sites corresponding to previous studies of this hot spring mat (Becraft et al., 2011, 2015) (see Supplemental Table B.1 for a summary of sites and Supplemental Figure B.1 for before and after photographs). The highest-temperature site that could be sampled on January 16, 2014 averaged 66.3 °C, while the same site averaged 68.64 °C during the July 17, 2014 sampling trip. The upper ∼5 mm-thick portion of the mat containing the phototrophic members of the microbial community was separated from the rest of the sample and flash-frozen in liquid isopentane cooled by liquid N_2. Frozen samples were placed in 15 mL Falcon tubes cushioned with Kimwipes on both ends, stored on dry ice during transit, and placed in a −80 °C freezer for longer-term storage (Ramsing et al., 2000).

Temperature measurements of the water flowing over the mat were made with a hand-held Eutech Instruments DualLogR thermocouple thermometer for site selection, and temperature was continuously recorded during sampling with Maxim iButton thermistors placed directly on the microbial mat (Supplemental Figures B.1 and B.2). Sites selected at Mushroom Spring for sampling were named by attaching 'MS' to the initial temperature measured by the hand-held thermocouple thermometer (e.g., MS68, MS65, MS62.5, MS60, MS55, and MS50 represent sample sites from ∼68 °C to ∼50 °C). These labels are used since the average temperature

at each site varied throughout the day (Supplemental Figure B.2). Downwelling irradiance (400 nm to 700 nm) was recorded with a LI-COR LI-190R quantum sensor attached to a LI-COR LI-1400 data logger during the January sampling trip and spot measurements were made with a LI-COR LI-250 light meter during the July 2014 trip (Supplemental Figure B.3).

Additional samples that had been collected between 1989 and 2011 by various members of the Ward Lab (see Supplemental Table B.2 for a summary of samples collected from Mushroom Spring and Octopus Springs over this time period) and stored in a similar fashion (though without replication and usually without isopentane preservation), were initially analyzed in 2011 by Becraft (2014).

Vertical Dissection of Mat Samples

Three replicate frozen seasonal mat samples from each of the 65 °C and 62.5 °C sites were separately mounted on a dissection platform of a Leica CM1850 cryotome using Tissue-Tek O.C.T compound and sliced parallel to the mat surface with a sterile diamond knife at −20 °C. Three to five adjacent 40 µm-thick slices of the upper green layer of the mat, which ranged in thickness from 600 µm to 1200 µm, were placed into 1.5 mL microcentrifuge tubes for DNA extraction, resulting in six to eleven 120 µm-200 µm-thick subsamples from each core (see Supplemental Tables B.3, B.4, B.5, and B.6 for an overview of samples). Samples were stored in a cooler packed with dry ice during dissection, and kept in a −80 °C freezer until further analysis.

Molecular and Phylogenetic Analyses

A single unsectioned core comprising the upper green layer of the mat from each sampling location in January and July, in addition to the replicated upper green layer vertically dissected sub-cores, were selected for DNA extraction. Cells were lysed and DNA was extracted and purified using the methods described in Becraft et al. (2011,

2015). Purified DNA was sequenced using MiSeq or Ti454-barcoding sequencing technology using the same primers designed to amplify a segment of the *psaA* gene by Becraft et al. (2011). DNA from seasonal mat samples was sequenced at the Research and Testing Laboratory, Lubbock, TX using the Illumina MiSeq technology. DNA from yearly mat samples was sequenced using 454 Titanium technology at the J. Craig Venter Institute.

As in Becraft et al. (2015), ClustalW (Larkin et al., 2007) controlled by the script `homopolymer-extinguisher.pl` (available from https://github.com/sandain/pigeon), was used to align and trim all sequences to 282 base pairs. A single nucleotide missing from the majority of the 454 sequence data was excised from all other sequences to obtain the maximum number of sequences for analysis. High-frequency sequences (HFSs; defined as those with $\geq$10 identical representatives across all combined samples) were detected using the script `hfs-finder.pl` (available from https://github.com/sandain/pigeon) and separated into A- and B′-like sequence datasets using top-hit BLASTn (Altschul et al., 1990) and comparisons with the HFSs found by Becraft et al. (2015). Separate A- and B′-like phylogenies were created using FastTree (Price et al., 2009).

Putative Ecotype Demarcation

Populations predicted through evolutionary simulation analysis to be ecologically distinct species, or PEs, with ecologically interchangeable members, were demarcated using ES2 (an evolutionary simulation program based on the Stable Ecotype Model of species and speciation), as described in Chapter 2. ES2 can produce different ecotype clusters depending on the dataset used, so correspondence with putative ecotypes detected by Becraft et al. (2015) is provided by identifying PEs in prior and new analyses that share the most-abundnat dominant variant HFS. Previously described

PE names are enclosed in parenthesis after ES2-generated PE names.

Canonical Correspondence Analysis

Seasonal samples were analyzed by canonical correspondence analyses (CCA) (ter Braak, 1986; Legendre and Legendre, 1998) provided by the `R` library `vegan` (Oksanen et al., 2013). The plotting function used in Chapter 2 was used to display the distribution of HFSs with respect to the average temperature recorded at each site and the depth below the surface of the mat from which the samples were originally found (see Supplemental Tables B.1, B.3, B.4, B.5, and B.6 for an overview of samples).

Results and Discussion

By comparing the abundance of HFSs in predominant PEs (those making up >2% of the community sampled from replicate samples; Figure 4.1 and Supplemental Figures B.4, B.5, B.6), it was observed that the *Synechococcus* population structure within the Mushroom Spring mat community was highly similar among different cores collected within an ∼4 cm-diameter sampling site (see Supplemental Figure B.1 for photographs of each sampling site). Although the abundance of HFSs was lower in the unsectioned mat sample than the sum of sequences obtained from each vertically dissected core, the greater coverage of the dissected cores can be explained by the summation of the six to eleven subsamples per core.

Comparison of Winter and Summer Thermal Profiles

The abundance of HFSs from unsectioned samples taken from the temperature-defined sites (Figures 4.2 and 4.3) displayed patterns similar to those seen by Ferris and Ward (1997). The temperature range where the A′-lineage was found was

generally higher than that of the A-lineage, which was generally higher than that of the B′-lineage, and the temperature range for the three lineages fluctuated somewhat between seasons. During the winter, PE PEA0380 (A′9) was most abundant in the two highest temperature samples, MS65 and MS68 (which averaged 65.25 °C to 66.33 °C, respectively), whereas in the summer it was only abundant at the MS68 site. Similarly, the range of the A-lineage extended to cooler waters than that of the A′-lineage during both the winter and summer, with the lineage found in all samples, but most abundant in the MS60, MS62.5, MS65, and MS68 samples during the winter (61.03 °C to 66.33 °C), and the MS62.5, MS65, and MS68 samples during the summer (61.42 °C to 68.64 °C). The upper range of the B′-lineage varied with seasons, with PEs PEB0270 (B′3), PEB0174 (B′9-2), and PEB0641 (B'12-2) detectable in the highest temperature sample, MS68, but not during the summer. During the winter, only two predominant PEs were found in abundance (>100 copies) in the MS65 sample, PE PEB1083 (B′9-2) and PEB0435 (B′9-1), while in the summer all PEs except PEB0270 (B′3) were in abundance in the MS65 sample.

Individual PEs within each lineage displayed different temperature ranges that fluctuated somewhat with the seasons (Figures 4.2 and 4.3). For example, PEA1151 (A1) was the most abundant A-like ecotype in all samples, but was most abundant in the MS62.5, MS65, and MS68 samples in the winter and in the MS65 sample during the summer. A-like PE PEA0181 (A6) was most abundant in the MS62.5, MS65, and MS68 samples during the winter, and the MS65 sample in the summer. A-like PE PEA1175 (A14) was most abundant in the MS62.5 sample during the winter and in the MS65 sample during the summer. B′-like PE PEB0174 (B′9-2) was the most dominant B′-like PE in the MS62.5 and MS60 samples during the winter, but was nearly equal in dominance with PEB0848 (B′2-3) during the summer.

Canonical correspondence analyses of the unsectioned samples taken from the

temperature-defined sites (Figure 4.4) displayed a similar pattern as shown above and by Becraft et al. (2015). A′-like PE PEA0380 (A′9) was found in the highest-temperature region sampled in both winter and summer. A-like PEs PEA0181 (A6), PEA1349 (A12), and PEA1175 (A14) were found next in slightly cooler environments. B′-like PEs PEB0379 (B′8-4), PEB0174 (B′9-2), PEB1083 (B′9-2), PEB0435 (B′9-1), PEB0641 (B′12-2), and PEB0848 (B′2-3) were found in the coolest regions sampled in both seasons. B′-like PE PEB0435 (B′9-1) was positioned closer to the vector line than the other two B′9 PEs in the winter and resided in warmer water than the other two PEs during the summer (Figure 4.4). Since the temperature vector represents only one important parameter, deviation from this vector suggests that one or more other parameters (measured or unmeasured) may also be important to defining the distribution of PEs. However, because it is impossible to know what the deviation from this vector means, we draw inferences only from the distribution of HFSs along the temperature vector.

All predominant PEs except PEA0181 (A6) displayed a distribution significantly ($p < 0.05$) different from random along the thermal gradient during the winter, an important observation, as it suggests that the members of nearly all PEs are ecologically interchangeable. During the summer, PEs PEA0181 (A6), PEB0379 (B′8-4), PEB0174 (B′9-2), and PEB0641 (B′12-2) displayed distributions that were not significantly different from random ($p > 0.05$). The partial *psaA* sequence used to demarcate *Synechococcus* PEs in this study has been observed to be shared between two widely divergent phylogenetic groups in PE A6 (Olsen, 2015), possibly due to recombination between these groups. This may be the reason that the three B′-like PEs also appeared to be randomly distributed in the environment.

Comparison of Winter and Summer Vertical Profiles

Canonical correspondence analyses of replicate dissected-core samples from the MS62.5 and MS65 sites (Figure 4.5) showed a pattern similar to that seen by Becraft et al. (2015) with regard to where in a depth profile PEs were found. B′-like PEs PEB0174 (B′9-2), PEB1083 (B′9-2), and PEB0435 (B′9-1) were found near the top of the mat in both the winter and summer samples, while A-like PEs PEA0181 (A6), PEA1349 (A12), and PEA1175 (A14) were found deeper.

In replicate samples that were vertically dissected (Figure 4.5), several PEs showed differences in distribution between seasons. A-like PEs PEA0181 (A6), PEA1349 (A12), and PEA1175 (A14) were found in warmer water in the summer and cooler water in the winter (compare Figures 4.5 A and C). The three PEs in the B′9 clade, appeared to have distributions that varied differently between winter and summer (compare Figure 4.5 B and D). PEB0435 (B′9-1) displayed seasonal variation not displayed by the other two B′9-like PEs. In the winter samples, PEB0435 (B′9-1) was found at a higher temperature than the rest of the B′9 clade, while in the summer samples it was found at the same temperature, though lower in the mat. The higher amount of solar irradiance that the mat receives in the summer might increase the upper temperature limit of *Synechococcus* species (Allewalt et al., 2006), potentially explaining this difference in distribution between seasons.

By comparing normalized counts of the predominant PEs detected in replicated samples that were vertically dissected (Figure 4.6), it was possible to see differences in the composition of the Mushroom Spring mat between seasons. The A′-like PE PE0380 (A′9-1) displayed a higher abundance in the winter MS65 samples. A-like PE PEA1151 (A1) displayed a two-fold higher abundance in the MS62.5 and MS65 samples during the summer. PE PEA1276 (A5) was more abundant in the summer than it was in the winter. PEs PEA0181 (A6) and PEA1175 (A14) both displayed a

larger range in temperature in the summer when they were found in abundance in both the MS62.5 and MS65 samples. But, in the winter, they were only found in the MS62.5 sample. B′-like PEs PEB0848 (B′2-3) and PEB0270 (B′3) were more abundant in the summer, while B′-like PEs PEB0174 (B′9-2), PEB1083 (B′9-2), PEB0435 (B′9-1), and PEB0641 (B′12-2) were more abundant in the winter. B′-like PE PEB0379 (B′8-4) was more abundant overall in the MS65 samples during the summer, but was more abundant in the MS62.5 sample during the winter.

The drastic changes in abundance of PEs demarcated in the B′9 clade (Figure 4.6) between seasons and temperatures provided further evidence that the subclade structure revealed by ES2 analyses equates to ecological diversity present within this clade. During the winter, PEB0174 (B′9-2) dominated in the MS62.5 samples while PEB0435 (B′9-1) dominated in the MS65 samples. During the summer, PEB0174 (B′9-2) was found in abundance in both MS62.5 and MS65 samples, while PEB0435 (B′9-1) was barely detected. Since the B′9 clade was shown to be surface oriented by Becraft et al. (2015), it is reasonable that its abundance would increase in the winter when solar irradiance is at a minimum. This location would provide maximum access to available light, but the decreased intensity of the light would potentially lead to less photodegradation of important photosystem proteins, and hence to lower maintenance costs, than members of this PE would be subjected to in the summer.

Comparison of Species Composition Over Decadal Time

Although PEs are occasionally missing from individual samples collected over a twenty-five year span from Mushroom Spring and Octopus Spring (Figures 4.7 and 4.8), the overall view of the *Synechococcus* population is one of stability. All predominant A-like PEs analyzed in this study could be detected as far back as 1999, and all predominant B′-like PEs could be detected back to 1989. Usually,

only the dominant variant of a PE was detected in older samples, but the same dominant variant was observed in all PEs in both Mushroom Spring and Octopus Spring through this time period. The deeper-coverage MiSeq sequencing was able to detect sub-dominant variants within PEs, suggesting that if the older samples had been analyzed with this depth of coverage, it might have been possible to witness within-species variation of individual genotypes. However, based on the evidence in hand, it appears that no periodic selection events occurred favoring a more-fit variant over these years. Alternatively, if periodic selection events had occurred, they always favored the same variant.

Conclusions

Putative ecotypes predicted by the evolutionary simulation algorithm Ecotype Simulation 2 displayed clear patterns of seasonal change in population structure. The abundance of PEs changed with season, with some becoming more abundant and others less abundant. The distribution of PEs also changed with the season, with some more abundant in warmer or cooler water, and others more abundant in deeper mat layers, or closer to the mat surface. These changes could be the result of the growth of PEs, or movement, as change in cell orientation (Ramsing et al., 2000) and phototaxis (Kilian et al., 2007) have been reported for A/B′-lineage *Synechococcus*. However, reanalysis of samples sequenced by Becraft (2014), combined with new samples sequenced for this study, showed a relative stability in the structure of individual *Synechococcus* PEs over a twenty-five year period.

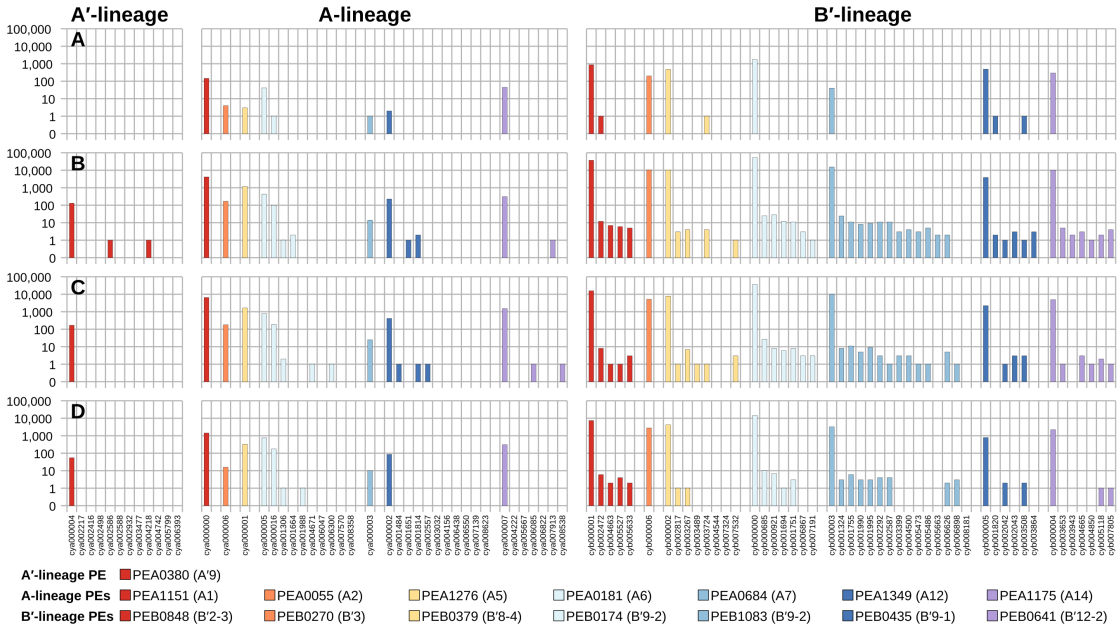


Figure 4.1: Mid-Summer Abundance of high-frequency *psaA* sequence segments that are predicted to be members of predominant A′-, A-, and B′-like putative ecotypes (PEs) from replicate mat samples collected from the 62.5 °C site of the Mushroom Spring effluent channel on July 17, 2014. Abundances are based on sequences obtained from an unsectioned mat sample (A) or from mat subsamples that were obtained by vertical dissection using a cryotome with abundances of all subsamples summed (B-D).

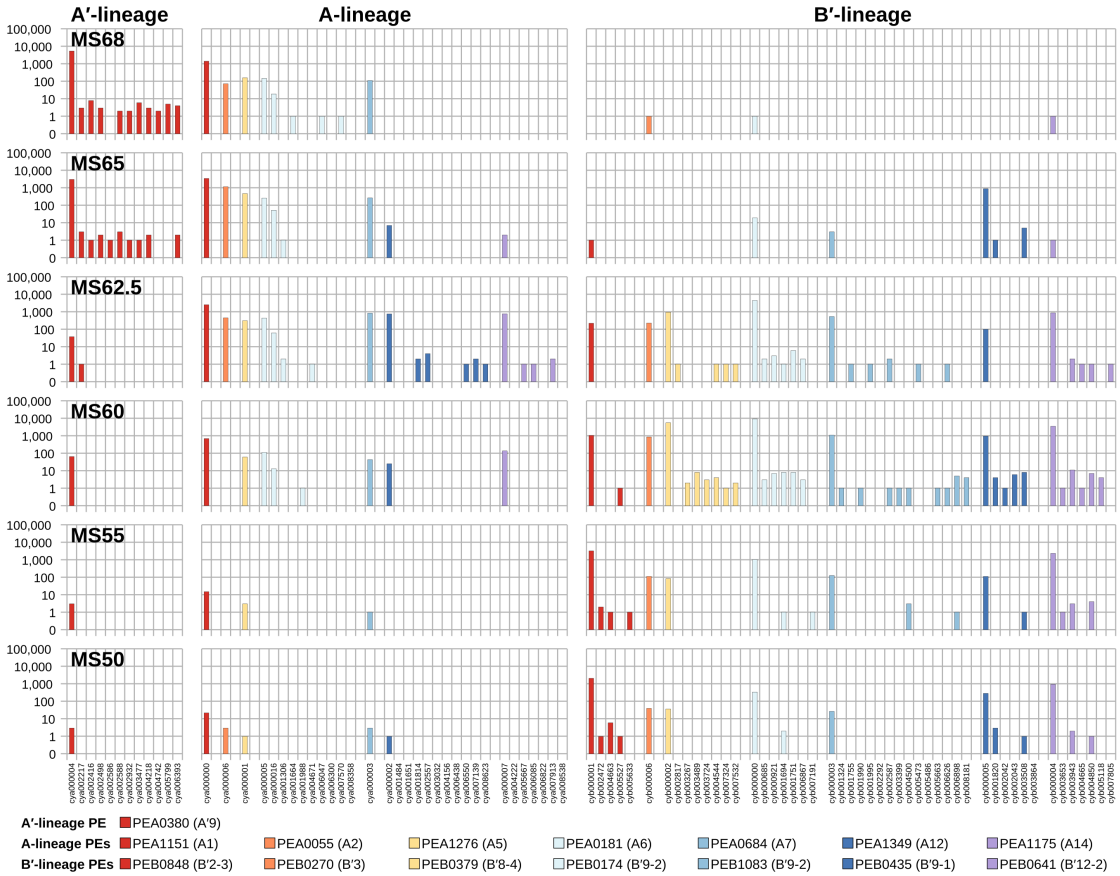


Figure 4.2: Winter abundances of high-frequency *psaA* sequence segments that are predicted to be members of predominant A′-, A-, and B′-like putative ecotypes (PEs) from mat samples taken from temperature-defined sites along the Mushroom Spring effluent channel on January 16, 2014. Abundances are based on sequences obtained from unsectioned mat samples.

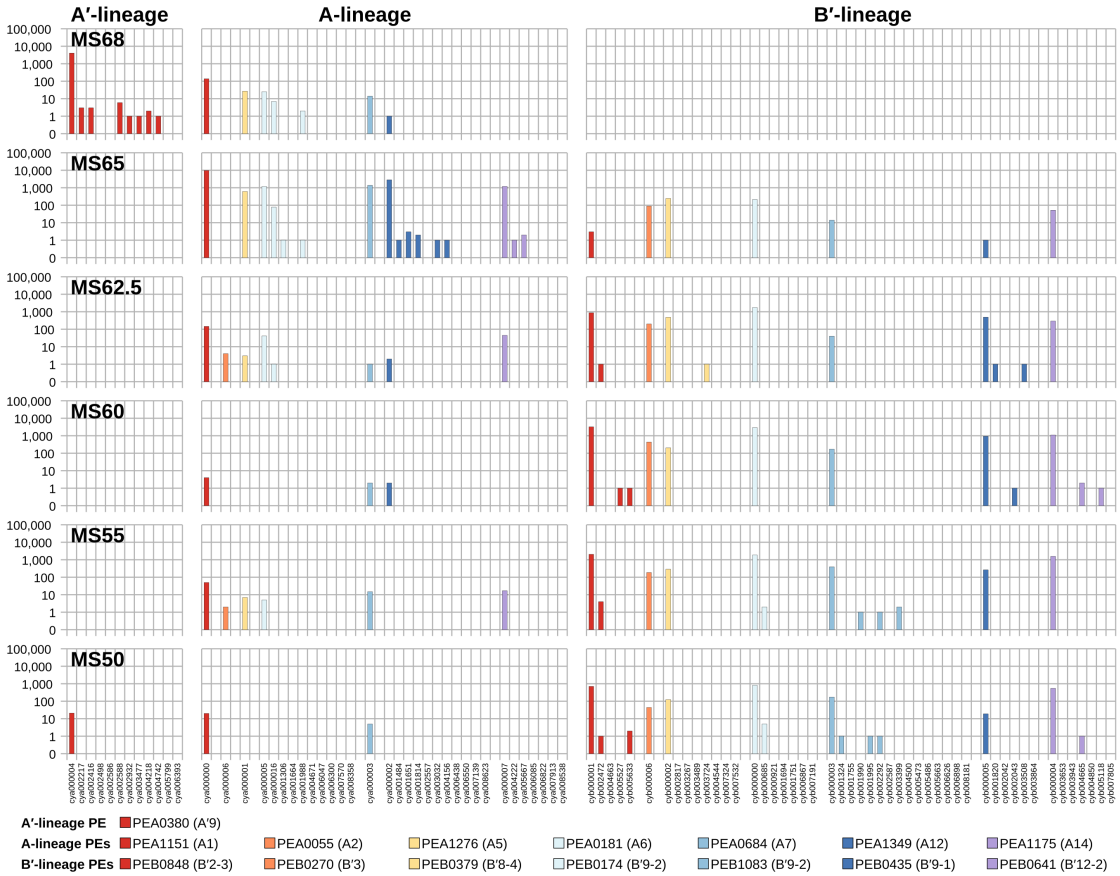


Figure 4.3: Summer abundances of high-frequency *psaA* sequence segments that are predicted to be members of predominant A′-, A-, and B′-like putative ecotypes (PEs) from mat samples taken from temperature-defined sites along the Mushroom Spring effluent channel on July 17, 2014. Abundances are based on sequences obtained from unsectioned mat samples.

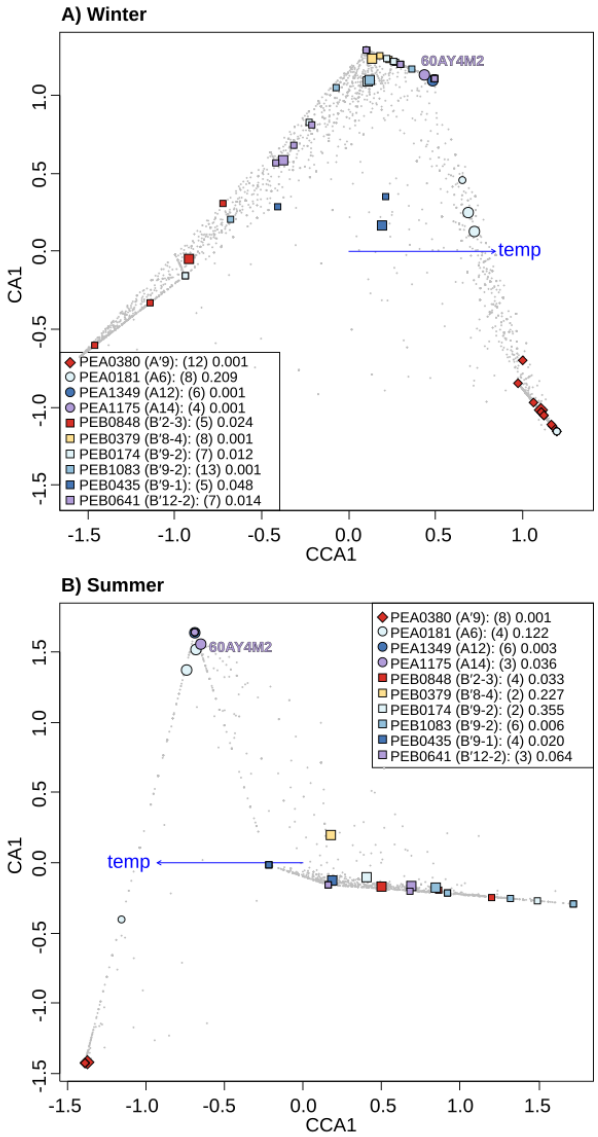


Figure 4.4: Canonical correspondence analyses of *Synechococcus* A′-, A-, and B′-like putative ecotypes (PEs) relative to temperature on January 16, 2014 (A) and July 17, 2014 (B). Predominant PEs described by Becraft et al. (2015) are highlighted with larger glyphs. *Synechococcus* strain 60AY4M2 has an identical *psaA* sequence as PEA1175 (A14), and is labeled on each plot. Small gray dots represent high-frequency sequences from lower-abundance PEs or from the other lineage. The directional arrow represents the vector of influence of the water temperature. Note that the arrows are in opposite directions in the winter and the summer, so high-frequency sequences found in higher temperature samples can be found on the right side of the winter plot, and on the left side of the summer plot. PE names in the legend are followed by the number of high-frequency sequences making up the PE in parenthesis, and a p-value that represents the probability that the observed PE cluster is different from random.

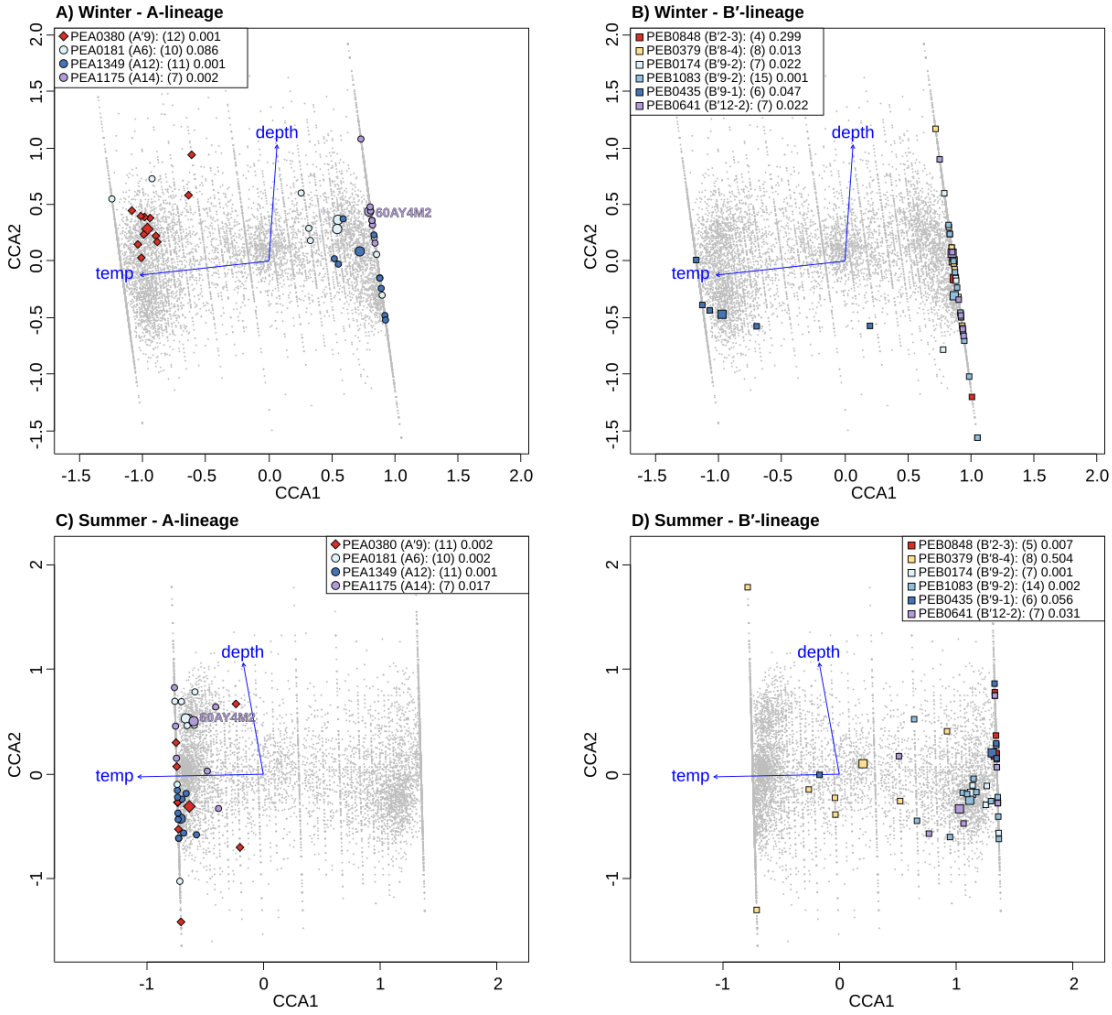


Figure 4.5: Canonical correspondence analyses of *Synechococcus* A′-, A- (A and C), and B′-like (B and D) putative ecotypes (PEs) relative to temperature and depth in samples from 65 °C and 62.5 °C in winter (A and B) and summer (C and D). Predominant PEs described by Becraft et al. (2015) are highlighted with larger glyphs. *Synechococcus* strain 60AY4M2 has an identical *psaA* sequence as PEA1175 (A14), and is labeled on each plot. Small gray dots represent high-frequency sequences from lower-abundance PEs or from the other lineage. The directional arrows represent the vectors of influence of the water temperature and depth within the mat. Note that the depth arrow indicates increasing depth in the mat, so that the top of the mat is found on the bottom of the winter and summer plots. PE names in the legend are followed by the number of high-frequency sequences making up the PE in parenthesis, and a p-value that represents the probability that the observed PE cluster is different from random.

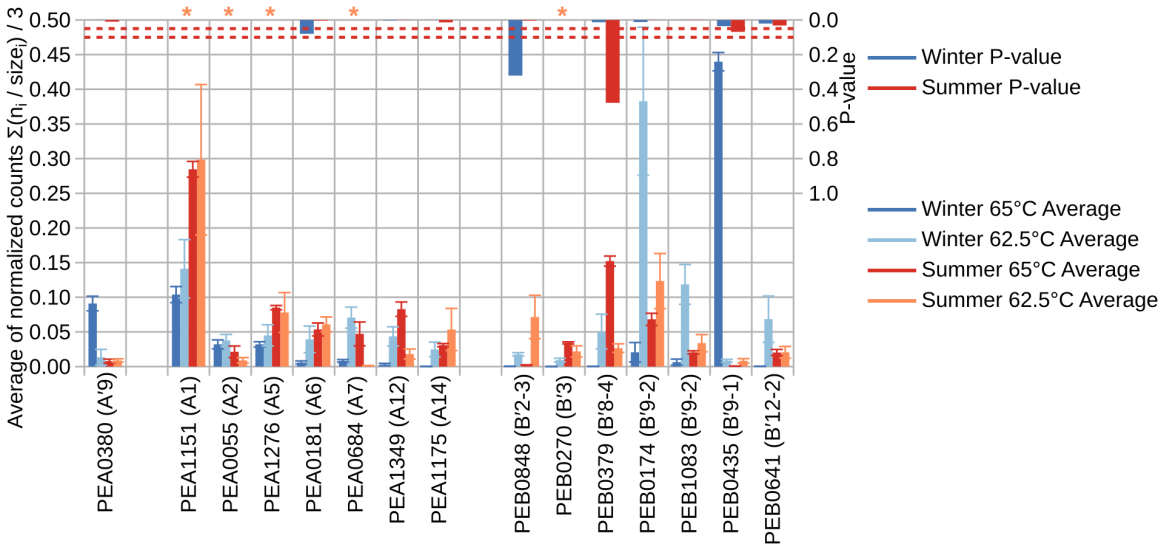


Figure 4.6: Average relative sequence abundance and p-values of predominant putative ecotypes (PEs) found in winter and summer samples collected at 65 °C and 62.5 °C. Orange asterisks above some columns represent p-values that could not be calculated since the PE only contained a single member. The red dashed lines represent 0.05 and 0.10 confidence that the members of the PE are not distributed randomly in the environment.

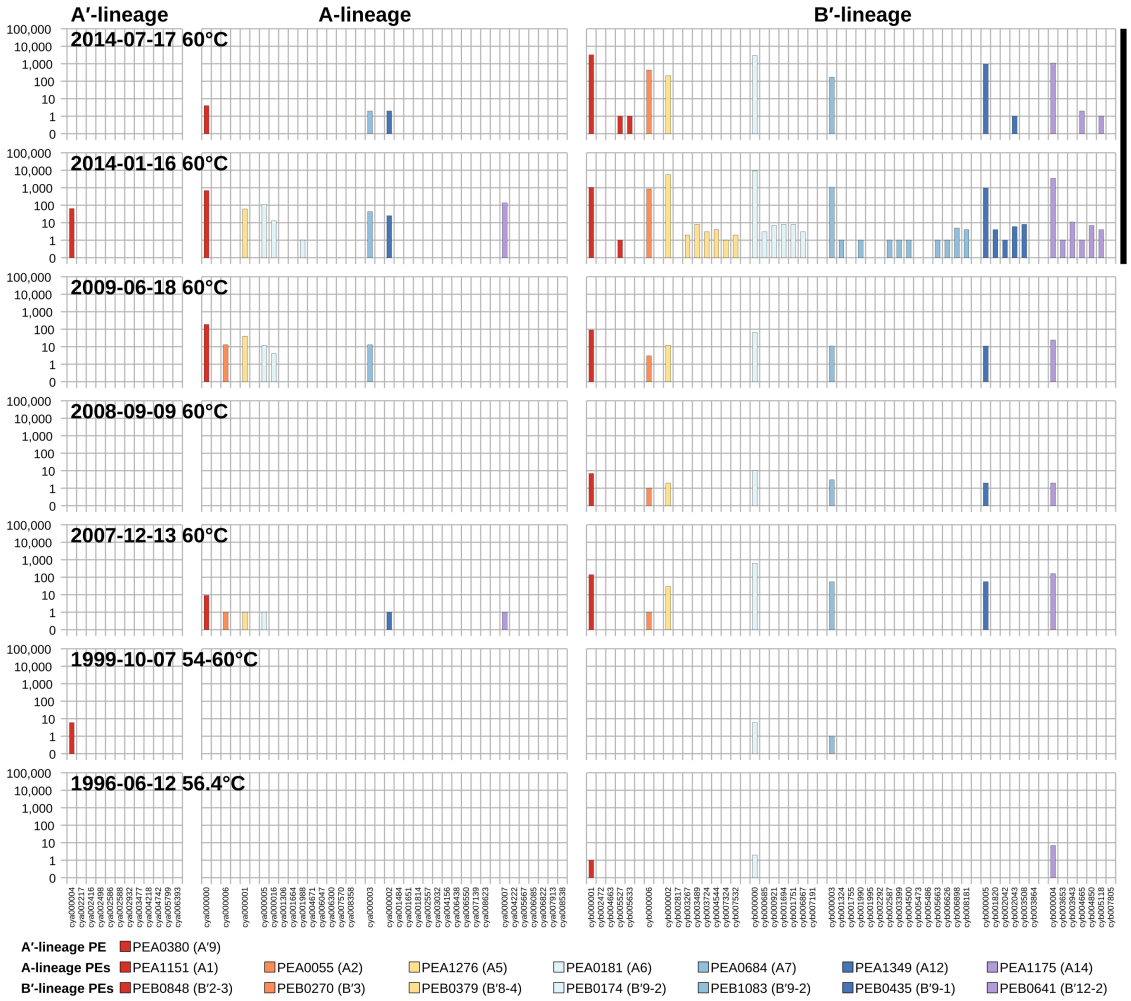


Figure 4.7: Abundance of high-frequency *psaA* sequence segments that are predicted to be members of predominant A′-, A-, and B′-like putative ecotypes (PEs) from mat samples collected from sites measuring 54 °C to 60 °C along the Mushroom Spring effluent channel between June 1996 and July 2014. DNA from samples in January and July, 2014 (black bar on the right) was sequenced using Illumina MiSeq, while DNA from all other samples was sequenced using 454 Titanium, resulting in less sequence coverage in older samples. Abundances are based on sequences obtained from unsectioned mat samples.

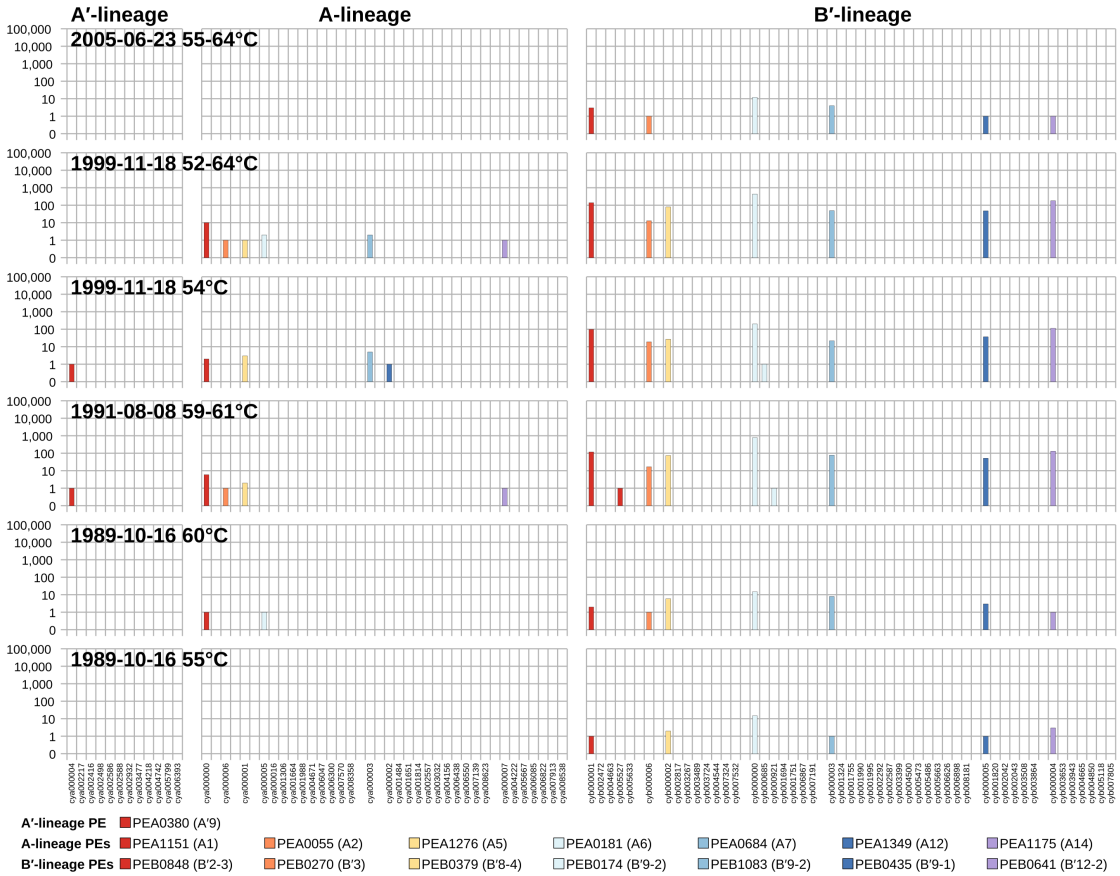


Figure 4.8: Abundance of high-frequency *psaA* sequence segments that are predicted to be members of predominant A′-, A-, and B′-like putative ecotypes (PEs) from mat samples collected from sites measuring 52 °C to 64 °C along the Octopus Spring effluent channel between October 1989 and June 2005. Abundances are based on sequences obtained from unsectioned mat samples.

CHAPTER FIVE

ECOTYPES OF OTHER PHOTOTROPHIC TAXA INHABITING THE MUSHROOM SPRING MICROBIAL MAT

Introduction

The stability of the Mushroom Spring system over time (see Chapter 4), combined with environmental gradients that can be measured at fine-scale, provide an excellent system for studying the genetic and ecological diversity of populations of microbes living in the mat community. Genetic variation among unicellular cyanobacteria (A/B′-lineage *Synechococcus*) living in the upper green layer (1 mm to 2 mm thick) of the microbial mat of the effluent channel has been extensively studied. Using a segment of the 16S rRNA gene, Ferris and Ward (1997) showed that *Synechococcus* populations varied along a well-defined temperature gradient in the system, with A″- and A′-like populations found in the warmest water, followed by A-, B′-, and then B-like populations in the cooler downstream sites (see Chapter 1 Figure 1.4). The *psaA* gene, important to photosynthesis in *Synechococcus*, has been used to show that these A′-, A-, B′-, and B-like populations could be further subdivided into putative ecotype (PE) populations that display even more restricted distributions in the environment than those predicted by using the 16S rRNA gene alone (Becraft et al., 2011, 2015). These PEs were shown to have differing distributions along the temperature gradient provided by the effluent channel and within a depth profile of the mat. Cultivated *Synechococcus* strains that correspond to PEs found at different depths were shown to exhibit different light acclimation and adaptation patterns. Low-light adapted strains predominating ∼1 mm below the surface of the mat grew faster at low light intensity and couldn't tolerate sudden exposure to the high-light

conditions seen by PEs predominating at the top of the mat (Nowack et al., 2015). Genomes of *Synechococcus* isolates representative of these differently adapted PEs have been sequenced and have begun to reveal the genetic bases for these differences in distribution (Olsen et al., 2015).

Although *Synechococcus* spp., have been shown to distribute differently along environmental gradients, other major phototrophic mat taxa have not been studied to the same extent. Using metagenomic assembly, Klatt et al. (2011) described six clusters of predominant phototrophic taxa present in the Mushroom Spring mat (as seen in Chapter 1 Figure 1.15). In addition to the oxygenic phototrophic *Synechococcus*, the mat contains multiple filamentous anoxygenic phototrophic members of the Chloroflexi: *Roseiflexus* spp., *Chloroflexus* spp., and '*Candidatus* Roseilinea' spp. Additional predominant members of the Mushroom Spring mat community include anoxygenic phototrophs from the phyla Acidobacteria and Chlorobi: *Chloracidobacterium* spp. and '*Candidatus* Thermochlorobacter' spp. Each taxon will be described in more detail below.

Other Predominant Mushroom Spring Phototrophic Taxa

5.1.1.1 *Roseiflexus* spp. *Roseiflexus* spp. are the most dominant Chloroflexi in the Mushroom Spring mat system, with their range extending from the green layer at the top of the mat into the orange layer of the 'undermat' (∼2 mm to 3 mm below the surface) (Thiel et al., 2016). *Roseiflexus* spp. utilize a type-2 reaction center and the 3-hydroxypropionate bi-cycle to perform photoautotrophy and photomixotrophy. Depending on the availability of resources, they are hypothesized to fix inorganic CO_2, metabolize organic compounds like glycolate produced by *Synechococcus* spp., break down internal storage molecules such as glycogen, polyhydroxyalkanoates, or wax esters, or by doing a combination of the above metabolisms (Klatt et al., 2013).

These red- (or orange-) colored Chloroflexi synthesize bacteriochlorophyll (BChl) *a* but not Bchl *c*, and they lack chlorosomes (van der Meer et al., 2010). An isolate representative of the population in Mushroom Spring (Klatt et al., 2011), *Roseiflexus* sp. RS-1, was originally sampled from nearby Octopus Spring and described by van der Meer et al. (2010).

Several previous observations indicate the possible existence of ecological species in this taxon. Ferris and Ward (1997) discovered that closely related *Roseiflexus* 16S rRNA variants were distributed differently along the thermal gradient of Octopus Spring (see bands labeled C″, C′, and C in Chapter 1 Figure 1.4), suggesting ecological patterning of diverse *Roseiflexus* populations. Similar results were found by Miller et al. (2009) for populations in White Creek and Rabbit Creek in Yellowstone National Park (see the 'Red Chloroflexi' in Chapter 1 Figure 1.16A and B). Later, a larger diversity of *Roseiflexus* 16S rRNA was identified (with >6000 dereplicated 16S rRNA iTag sequences detected) in the undermat (2 mm to 5 mm below the surface) by Thiel et al. (2016). *Roseiflexus* spp. strains RS-1 and RS-2 were shown to display differences in light absorption maxima (e.g. 900 nm for strain RS-1 and 860 nm for strain RS-2), and differences in carbon substrate utilization (e.g., growth stimulated by malate for strain RS-1 and lactate for strain RS-2) by van der Meer et al. (2010). These observations suggest possible temperature, light, and metabolic specialization.

5.1.1.2 *Chloroflexus* spp. *Chloroflexus* spp. are also abundant Chloroflexi in the mat community, with their numbers at a maximum in the higher-temperature water closer to the source pool (Nübel et al., 2002; Thiel et al., 2016). Like *Roseiflexus* spp., *Chloroflexus* can be found in the surface green layer of the mat as well as the orange 'undermat' (Klatt et al., 2013; Tank et al., 2017). Like the other mat Chloroflexi, *Chloroflexus* spp. utilize a type-2 reaction center to perform photoautotrophy and

photomixotrophy (Klatt et al., 2013). Unlike the other Chloroflexi in the mat community, these green-colored Chloroflexi synthesize both BChl *a* and BChl *c*, along with chlorosome antenna complexes. The *Chloroflexus* sp. MS-G isolate sampled from Mushroom Spring and its genome serve as the best representative of the population in the mat community (Thiel et al., 2014). Miller et al. (2009) demonstrated that various closely related clades of green-Chloroflexi were differentially distributed along the thermal gradients of White Creek and Rabbit Creek in Yellowstone National Park (see the 'Green Chloroflexi' in Chapter 1 Figure 1.16A and B).

5.1.1.3 '*Candidatus* Roseilinea' spp. The remaining predominant population of the mat Chloroflexi studied was first described by Klatt et al. (2011) as an *Anaerolineae*-like metagenomic bin. Later, organisms comprising this bin were given the name '*Candidatus* Roseilinea gracile' by Tank et al. (2017). Like the other Chloroflexi, the members of this cluster utilize a type-2 reaction center to perform photoautotrophy and photomixotrophy (Klatt et al., 2013). Their usage of BChl *a*, but not *c* (Klatt et al., 2011) allowed Tank et al. (2017) to use fluorescence microscopy to identify them as the very-fine filaments within the mat community.

5.1.1.4 *Chloracidobacterium* spp. Evidence of this population was originally found using a 16S rRNA survey of the Octopus Spring mat community, in which the 16S rRNA genotype D was classified as representing 'bacteria of uncertain lineage' by Ward et al. (1998). The Acidobacterium *Chloracidobacterium thermophilum* strain B was later cultivated and characterized as the first discovered chlorophototrophic member of the phylum (Bryant et al., 2007). Metagenomic assembly of DNA from the same mat community was used to describe *Chloracidobacterium* spp. as aerobic photoheterotrophs that utilize a type-1 reaction center, BChl *a* and *c*, chlorophyll (Chl) *a*, and chlorosomes (Bryant et al., 2007). The *Chloracidobacterium*

thermophilum strain B isolate sampled from the mat provides a genome representative of the population in the mat community (Garcia Costas et al., 2012). Miller et al. (2009) demonstrated that various clades of closely related *Chloracidobacterium* were differentially distributed along the thermal gradient of White Creek in Yellowstone National Park.

5.1.1.5 '*Candidatus* Thermochlorobacter' spp. The Chlorobi '*Candidatus* Thermochlorobacter aerophilum' was originally described as green sulfur bacterium-like 16S rRNA Types E, E′, and E″ by Ferris et al. (1996). Later a Chlorobiales-like metagenomics bin was discovered by Klatt et al. (2011). Based on metagenome and metatranscriptome information, Liu et al. (2012) later described organisms comprising this bin. '*Ca.* Thermochlorobacter aerophilum' as the first aerobic chlorophototrophic member described in the phylum *Chlorobi* (Liu et al., 2012). Evidence provided by metagenomics and transcriptomics points to '*Ca.* Thermochlorobacter' utilizing a type-1 reaction center, BChl *a* and *d*, and chlorosomes to perform aerobic photoheterotrophy (Klatt et al., 2011; Liu et al., 2011, 2012).

Study Design

Loci analogous to the *Synechococcus* photosystem I reaction center gene *psaA* in other predominant phototrophic taxa in the Mushroom Spring mat community were used to study population diversity and distribution using the same methods (i.e., protein-encoding loci, high-throughput sequencing, Ecotype Simulation 2, and canonical correspondence analysis) used most recently to study *Synechococcus* in Chapters 2, 3, and 4.

Methods

Environmental Samples

Samples of microbial mat in the main effluent channel of Mushroom Spring, Yellowstone National Park, WY, USA, collected July 17, 2014, and previously described in Chapter 4, were sequenced using primers designed to capture the diversity of major photosynthetic organisms within the mat community. Mat cores collected from the 68, 65, 62.5, 60, 55, and 50 °C sites were used to examine the diversity of each taxon along the temperature gradient provided by the effluent channel (see Supplemental Table C.1 for a summary of samples and sequences collected).

For depth profiling, three replicate mat cores collected from the 62.5 °C site were sliced into 80 µm-thin sections using a cryotome. Adjacent slices of mat core were combined together, resulting in subsamples ranging in thickness from 120 µm to 200 µm.

Molecular Methods and Sequence Analysis

Cells were lysed and DNA was extracted and purified using the same methods described in Becraft et al. (2011, 2015). Prior to sequencing, for all taxa other than *Synechococcus*, DNA from adjacent vertical subsamples of mat cores described in Chapter 4 were combined to reduce the total number of samples to reduce costs. The same *Synechococcus psaA* sequence segments described in Chapter 4 were used for comparison. To account for adjacent slices of mat core being combined prior to sequencing the other taxa, adjacent *Synechococcus psaA* sequence libraries were analyzed together. This resulted in subsamples representing a mat thickness of 240 µm to 400 µm for all taxa (see Supplemental Table C.2 for a summary of samples and sequences collected).

Loci involved in photosynthesis analogous to the photosystem I reaction center *psaA* gene in *Synechococcus* previously described in Chapter 4 were selected for each taxon (see Supplemental Table C.3 for a summary of loci analyzed). The *pufL/pufLM* gene of the *Chloroflexi* that encodes a subunit of the type-2 reaction center was used to capture the diversity of mat *Roseiflexus*, *Chloroflexus*, and '*Ca.* Roseilinea'. The *pscA* genes of *Chloracidobacterium* and '*Ca.* Thermochlorobacter' were used to capture the diversity of mat phototrophs using type-1 reaction centers. Primers for each taxon were designed using PrimerQuest (IDT, Coralville, USA). The primers were chosen based on their location within the gene, and to limit the calculated possibilities for hairpin-loop and prime-dimer formation. The primers were compaired with the NCBI nucleotide database using BLASTn (Altschul et al., 1990) to ensure that there were no other good matches besides the intended taxon. All chosen primers successfully amplified targeted sequences from the Mushroom Spring mat and isolates (for those taxa with isolates available) in PCR trials, and have been cross-tested on isolates from other taxa to ensure specificity (Jane Wiedenbeck and Frederick M. Cohan, unpublished). Sequencing was done by Research and Testing Laboratory, Lubbock, TX using the Illumina MiSeq technology.

ClustalW (Larkin et al., 2007), controlled by the `homopolymer-extinguisher.pl` script, was used to separately align and trim sequences for all taxa. High-frequency sequences for each taxa (HFSs; defined as those with $\geq$50 identical representatives across all combined samples) were detected using the script `hfs-finder.pl` (both scripts are available from https://github.com/sandain/pigeon). *Roseiflexus* and *Chloracidobacterium* sequences obtained using the above primers were diverse (<80% sequence identity), so were split into separate A and B clades similar to the A and B clades of *Synechococcus*, using BLASTn (Altschul et al., 1990) comparisons with the most dominant HFS in each clade. Separate phylogenies were created using FastTree

(Price et al., 2009) for each clade of all taxa.

Putative Ecotype Demarcation

Putative ecotypes (PEs; populations predicted through evolutionary simulation analysis to be ecologically distinct species) were separately demarcated for each taxon using Ecotype Simulation 2 (ES2; an evolutionary simulation program based on the Stable Ecotype Model of species and speciation), as described in Chapter 2. Ecotype demarcation does not guarantee the same membership between runs of the program, so correspondence between PEs newly numbered by ES2 and those described previously by Becraft et al. (2015) are provided by enclosing the previously described PE names in parenthesis.

Canonical Correspondence Analysis

Mat samples were analyzed by canonical correspondence analyses (CCA) (ter Braak, 1986; Legendre and Legendre, 1998) provided by the `R` library `vegan` (Oksanen et al., 2013). The plotting function used in Chapter 2 was employed to display the distribution of HFSs in the most abundant PEs ($> 2\%$ of the lineage) from all taxa with respect to the average temperature recorded at each site or the depth below the surface of the mat from which the samples were obtained (see Supplemental Tables C.1 and C.2 for an overview of samples). The ecological interchangeability of the members of each PE was tested by examining the probability that they were distributed non-randomly. Plots demonstrating quantitatively the distribution of all members of each taxon across temperature- or depth-defined samples were generated using a custom `R` script (`ccadensity.R`; available from https://github.com/sandain/R).

Results and Discussion

Synechococcus

Results produced with CCA for *Synechococcus* (Figure 5.1) replicated patterns seen in previous studies and served as a reference for comparison to the other taxa studied. As Ferris and Ward (1997) found using a 16S rRNA gene segment, and Becraft et al. (2011, 2015) found using the same *psaA* gene segment studied here, *Synechococcus* PEs displayed restricted distributions along the fine-scale environmental gradients measured, water temperature and depth within the mat. Predominant *Synechococcus* PEs from the A-lineage were found in warmer water than PEs from B′-lineage (compare circles and squares in Figure 5.1A). Analysis of the depth profile of the mat (Figure 5.1B), demonstrated differences in distribution previously noticed by Becraft et al. (2015), with PEB553 (B′9-1) and PEB375 (B′9-3) found closer to the surface of the mat than PEA234 (A14) and the *Synechococcus* sp. strain 60AY4M2 isolate, which is representative of the Becraft et al. (2015) PE A14 (compare blue squares and lavender circles in Figure 5.1). Other A-like PEs named by Becraft et al. (2015) found deeper in the mat (e.g., A1, A4, A6) were below the 2% predominance cutoff, or contained a single HFS so were not highlighted in these figures.

Most predominant A-like and all B′-like *Synechococcus* PEs with more than one HFS displayed non-random distributions in the CCA ordination space defined by both the temperature and depth gradients ($p < 0.05$) (Figure 5.1). Although A-like PE PEA438 (A5) obtained poor p-values ($p = 0.881$ in the temperature profile; $p = 0.295$ in the depth profile) in the test for randomness, it appeared to be a high-temperature PE. A-like PE PEA234 (A14) demonstrated mixed p-value results in the test for randomness, with a non-random distribution displayed in the

temperature profile ($p = 0.001$) but a distribution that appeared random in the depth profile ($p = 0.921$). P-values that exceed 0.05 in the test for randomness of a PE's distribution can potentially be explained by: i) the PE being sampled outside of its niche (e.g., a recent immigrant was sampled), ii) the protein-encoding locus used for demarcation having experienced recombination with a distantly related PED, iii) the protein-encoding locus used did not having enough molecular resolution to observe population differences, or iv) a miscall in the generation of the phylogeny or in ecotype demarcation by ES2.

Roseiflexus

Roseiflexus PEs also displayed restricted distributions along the fine-scale environmental gradients measured (Figure 5.2). Most predominant PEs in the A-clade (circle glyphs), including the strain RS-1-like PEA034 (green circles), were localized to the middle of the temperature range sampled. Predominant B′-clade PEs (square glyphs) were found throughout the temperature range (Figure 5.2A). Similarly, PEs in the A-clade (circles) were found in abundance in the deeper samples, while PEs in the B-clade (squares) were found throughout the depth range sampled (Figure 5.2B).

Most predominant A-clade *Roseiflexus* PEs displayed non-random distributions in the CCA ordination space defined by the temperature gradient and all A-clade PEs displayed non-random distributions along the depth gradient ($p < 0.05$; see the circles in Figure 5.2). A-clade PE PEA005 failed the test for random distribution ($p = 0.179$) along the temperature gradient, but appeared to be distributed at depth in the middle-temperature samples. Most B-clade *Roseiflexus* PEs also displayed non-random distributions along the temperature gradient ($p < 0.05$; see the squares in Figure 5.2A), but only one appeared to be non-randomly distributed along the depth profile (PEB004; $p = 0.005$; Figure 5.2B).

Chloroflexus

Chloroflexus PEs were split into three visible clusters of HFSs in the CCA ordination space defined by the temperature gradient sampled (Figure 5.3). HFSs in PE019 were found clustered in the highest-temperature sample as well as the two clusters in the middle- to low-temperature samples. HFSs in PEs PE004, PE017, and the *Chloroflexus* sp. strain MS-G-like PE018 were only found in the two clusters in the middle- to low-temperature samples (compare the light-red circles to the light-blue, light-green, and dark-green circles in Figure 5.3A). HFSs in PE004 and PE019 were found scattered throughout the depth profile with a larger concentration in a cluster near the deepest part of the mat sampled. HFSs in PE017 and PE018 were only found in the cluster in the deepest samples (compare the light-blue and light-red circles with the light-green and dark-green circles in Figure 5.3B).

Although all predominant *Chloroflexus* PEs displayed non-random distributions in the CCA ordination space defined by the temperature gradient ($p < 0.05$), PE019 in particular had a distribution that hinted at a miscall in PE demarcation, with the PE splitting into two thermally separated clusters (compare the light-red circles in Figure 5.3A). Most of the *Chloroflexus* PEs displayed non-random distributions in the ordination space defined by the depth gradient ($p = 0.001$; PE017, PE018, and PE019 in Figure 5.3B).

'*Candidatus* Roseilinea'

'*Candidatus* Roseilinea' PEs displayed restricted distributions similar to *Synechococcus* and *Roseiflexus* PEs (Figure 5.4). HFSs in predominant PE PE006 were localized to the highest- to middle-temperature ranges sampled; in PE PE032 they were localized to the middle-temperature samples; in PEs PE036, PE039, and PE040 they were localized to the middle- to low-temperature ranges sampled; and in PE

PE042 they were localized to the lowest-temperature samples (Figure 5.4A). Most HFSs in predominant '*Ca.* Roseilinea' PEs were localized to the deepest portion of the mat sampled, with HFSs in PEs PE006 and PE042 found throughout the depth profile (Figure 5.4B).

Most predominant '*Ca.* Roseilinea' PEs displayed non-random distributions in the CCA ordination space defined by the temperature and depth gradients ($p < 0.05$). PE PE006, however, displayed a random distribution in both temperature and depth profiles ($p = 0.127$ and $p = 0.996$). Although PE006 is localized to the warmest samples, this poor p-value may allude to population structure hidden by the choice of the *pufL* gene. The fact that PE006 is randomly distributed near the surface of the mat (Figure 5.4B) in 62.5 °C samples, might point to poor p-values simply resulting from populations that have emigrated from elsewhere, in this case upstream in warmer waters. Such populations would be outside of their normal niches, so their distributions would be subject to other forces, such as dispersal from upstream.

Chloracidobacterium

Chloracidobacterium PEs also displayed restricted distributions along the fine-scale temperature and depth gradients sampled (Figure 5.5). HFSs in the A-clade (circle glyphs) were all clustered together in the middle-temperature samples, while HFSs in the B-clade (square glyphs) related to *Chloracidobacterium thermophilum* strain B (labeled in Figure 5.5) were found spread out in the low- to lowest-temperature samples. B-clade PEs PEB003, PEB018, and PEB019 were found in the low-temperature samples, with PEB013 in the lowest-temperature samples (Figure 5.5A). Only the A-clade was found in abundance in the depth profile samples from the 62.5 °C site, matching the observation seen in the temperature profile. Most A-clade HFSs were clustered in the deeper samples of the mat, with a few HFSs from

PEA004 and PEA009 found in samples near the surface of the mat (Figure 5.5B).

All predominant A- and B-clade *Chloracidobacterium* PEs displayed non-random distributions in the CCA ordination space defined by the temperature and depth gradients ($p < 0.05$; Figure 5.5).

'*Candidatus* Thermochlorobacter'

'*Candidatus* Thermochlorobacter' PEs also displayed restricted distributions along the fine-scale temperature and depth gradients sampled (Figure 5.6). Most of the '*Ca.* Thermochlorobacter' PEs (e.g., PE008, PE012, and PE018) were found between the highest- and middle-temperature samples (see the light-green, light-red, and orange circles in Figure 5.6A). HFSs of PE001, on the other hand, were found between the middle- and lowest-temperature samples (see the light-blue circles in Figure 5.6A). HFSs were split into two different clusters along the depth gradient, with PEs PE008 and PE012 found in both (see the light-green and light-red circles in Figure 5.6B). The HFSs in PEs PE018 were restricted to the deeper cluster (see the orange circles in Figure 5.6B). Only two HFSs from low-temperature PE001 were found in the samples from the high-temperature 62.5 °C site, with one on the surface and the other near the deep cluster (see the light-blue circles in Figure 5.6B).

All predominant '*Ca.* Thermochlorobacter' PEs displayed non-random distributions in the CCA ordination space defined by the temperature gradient ($p < 0.05$; Figure 5.6A). Most predominate PEs displayed non-random distributions along the depth gradient ($p < 0.05$) sampled from the 62.5 °C site, with the exception of the poorly sampled low-temperature PE001 that displayed a more varied distribution (Figure 5.6B).

Quantitative Distributions of Taxa Along Environmental Gradients

By comparing quantitatively the distributions of all members of each taxon studied here, it was possible to see a fine-scale partitioning of the environment. Along the thermal gradient (Figure 5.7A), taxa were seen to have maximum densities at different temperatures. '*Candidatus* Thermochlorobacter' was found to have maximum density at the highest temperature, followed by A-like *Synechococcus*, then *Chloracidobacterium*, *Roseiflexus*, '*Ca.* Roseilinea', then B′-like *Synechococcus*, and finally *Chloroflexus*. Along the depth profile of the 62.5 °C site (Figure 5.7B), taxa were seen to have maximum densities at different depths. B′-like *Synechococcus* were found to have maximum density near the surface of the mat, followed by A-like *Synechococcus* a little deeper, then '*Ca.* Thermochlorobacter', *Roseiflexus*, *Chloracidobacterium*, then *Chloroflexus*, and finally '*Ca.* Roseilinea' in the deepest samples. This same placement of '*Ca.* Thermochlorobacter' over *Chloracidobacterium* and *Chloroflexus* was noted by Liu et al. (2012).

Differences in depth distribution may be attributed to varying pigment molecules synthesized by each taxon. '*Candidatus* Thermochlorobacter', which produces Bchl *d*, is found near the surface of the mat but beneath A- and B′-like *Synechococcus*. Bchl *d* absorbs light near 730 nm that is not absorbed by the Chl *a* of *Synechococcus* and *Chloracidobacterium* spp. (note the white bar in Chapter 1 Figure 1.2A at this wavelength). *Chloroflexus* and *Chloracidobacterium*, both of which produce Bchl *c*, were found with maximum density yet deeper in the mat. Bchl *c* absorbs light from 740 nm to 750 nm that is not absorbed by either the Chl *a* of *Synechococcus* and the Bchl *d* of '*Ca.* Thermochlorobacter' that are positioned at depth shallower in the mat (note the white bar in Figure 1.2A at this wavelength).

Chloroflexus, *Chloracidobacterium*, and '*Ca.* Thermochlorobacter', produce chlorosomes (photosynthetic antenna complexes). Chlorosomes are able to harvest

light in low-light conditions, and as such would provide an advantage to organisms that live deeper in the mat. This may explain why taxa with chlorosomes exist in niches deeper in the mat. The location of '*Ca.* Thermochlorobacter' closer to the surface of the mat than *Chloracidobacterium* may relate to differences in the ability of these organisms to tolerate oxygen produced in abundance by *Synechococcus* spp. during the daytime. Bchl *d*-containing chlorosomes have poorer light-harvesting efficiency than chlorosomes containing other bacteriochlorophylls (Liu et al., 2012), potentially explaining why '*Ca.* Thermochlorobacter' is a true aerobe, whereas *Chloracidobacterium* is a microaerophile (Tank et al., 2017). The positioning of *Chloroflexus* and '*Ca.* Roseilinea' deep in the photic zone of the mat may point to chemotrophy rather than photosynthesis performing as the primary source of energy for these organisms.

Lower-density peaks in Figure 5.7 likely represent populations adapted to different water temperature or depth with the mat. For instance, the low-density peak near the high-end of the temperature profile of A-like *Synechococcus* (Figure 5.7A) represents a higher-temperature adapted PE in the A′- or A″-lineages. Similarly, the low-density peak near the top of the mat in the depth profile of '*Candidatus* Thermochlorobacter' (Figure 5.7B) suggest that, in this taxon, there might be different populations adapted to different light qualities available at these depths.

Conclusions

It is clear that the major mat phototrophic taxa, and putative ecotypes predicted by Ecotype Simulation 2 within each, distribute differently in the environment with respect to both water temperature and depth within the microbial mat community. The improved capabilities of the Ecotype Simulation 2 algorithm have allowed for the rapid analysis of diversity in all predominant phototrophic taxa. This analysis

has demonstrated that the ecotype concept is useful for understanding the diversity present in all community members, not just *Synechococcus* diversity.

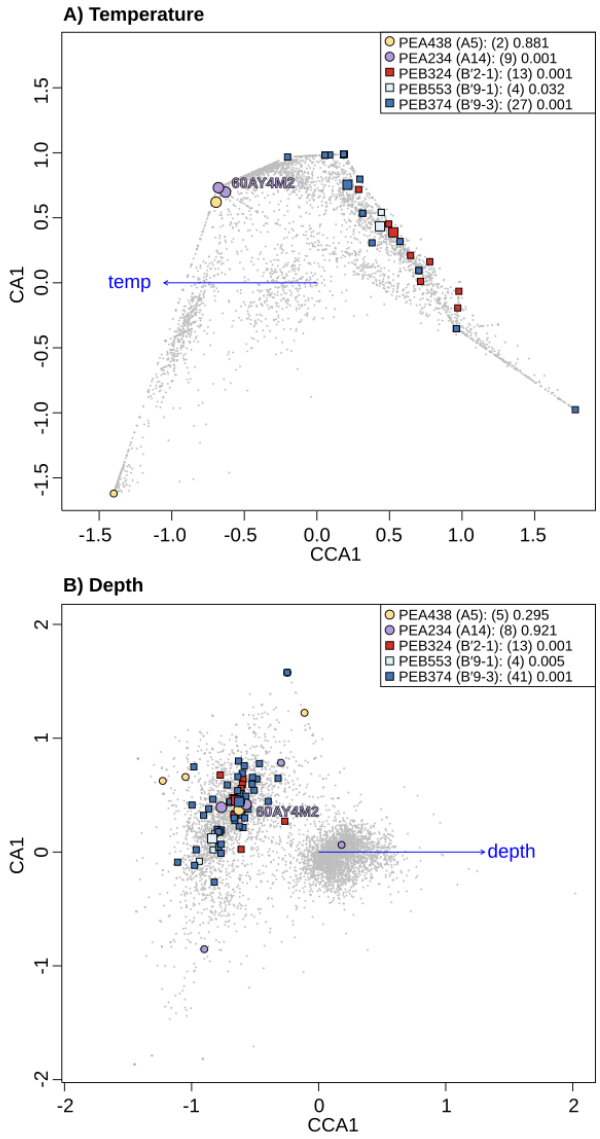


Figure 5.1: *Synechococcus*-highlighted canonical correspondence analyses of *psaA*, *pscA*, and *pufL/pufLM* high-frequency sequence (HFS) segments. HFSs of predominant *Synechococcus* A-, and B′-like putative ecotypes (PEs) are highlighted. The temperature profile (A) contains samples from the 68, 65, 62.5, 60, 55, and 50 °C sites. The depth profile (B) contains three replicate samples of the 62.5 °C site that were disected using a cryotome. The larger labeled glyph represents the HFS that shares the same *psaA* sequence as *Synechococcus* sp. strain 60AY4M2 described by Olsen et al. (2015). Small gray glyphs represent HFSs not highlighted here (<2% of the dataset) or from other taxa. The directional arrows represent vectors of influence of water temperature (A) or depth in the mat core (B). PE names in the legend are followed by the number of HFSs making up the PE in parenthesis and a p-value that represents the probability that the observed PE cluster is distributed differently from random.

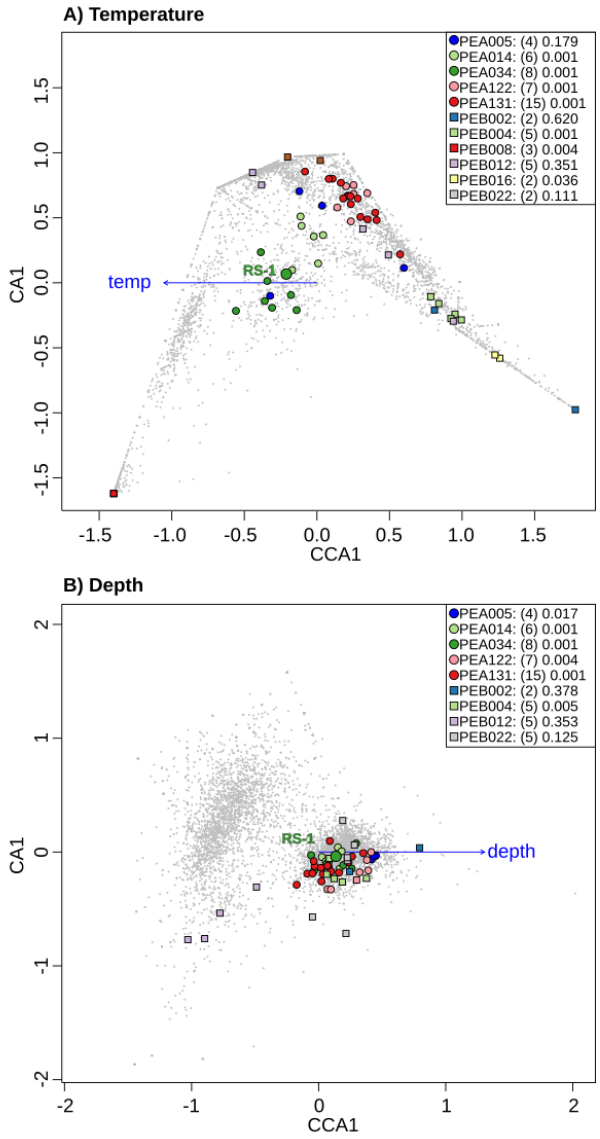


Figure 5.2: *Roseiflexus*-highlighted canonical correspondence analyses of *psaA*, *pscA*, and *pufL/pufLM* high-frequency sequence (HFS) segments. HFSs of predominant *Roseiflexus* putative ecotypes (PEs) are highlighted. The temperature profile (A) contains samples from the 68, 65, 62.5, 60, 55, and 50 °C sites. The depth profile (B) contains three replicate samples of the 62.5 °C site that were disected using a cryotome. The larger labeled glyph represents the HFS that shares the same *pufLM* sequence as *Roseiflexus* sp. strain RS-1 described by Madigan et al. (2005) and by van der Meer et al. (2010). Small gray glyphs represent HFSs not highlighted here (<2% of the dataset) or from other taxa. The directional arrows represent vectors of influence of water temperature (A) or depth in the mat core (B). PE names in the legend are followed by the number of HFSs making up the PE in parenthesis and a p-value that represents the probability that the observed PE cluster is distributed differently from random.

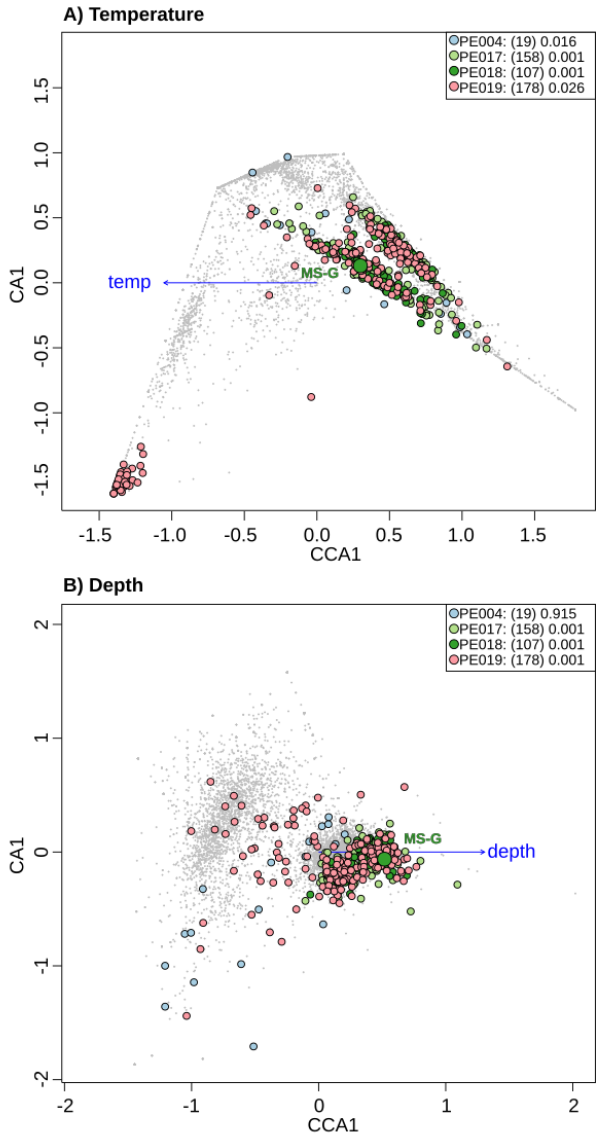


Figure 5.3: *Chloroflexus*-higlighted canonical correspondence analyses of *psaA*, *pscA*, and *pufL/pufLM* high-frequency sequence (HFS) segments. HFSs of predominant *Chloroflexus* putative ecotypes (PEs) are highlighted. The temperature profile (A) contains samples from the 68, 65, 62.5, 60, 55, and 50 °C sites. The depth profile (B) contains three replicate samples of the 62.5 °C site that were disected using a cryotome. The larger labeled glyph represents the HFS that shares the same *pufL* sequence as *Chloroflexus* sp. strain MS-G described by Thiel et al. (2014). Small gray glyphs represent HFSs not highlighted here (<2% of the dataset) or from other taxa. The directional arrows represent vectors of influence of water temperature (A) or depth in the mat core (B). PE names in the legend are followed by the number of HFSs making up the PE in parenthesis and a p-value that represents the probability that the observed PE cluster is distributed differently from random.

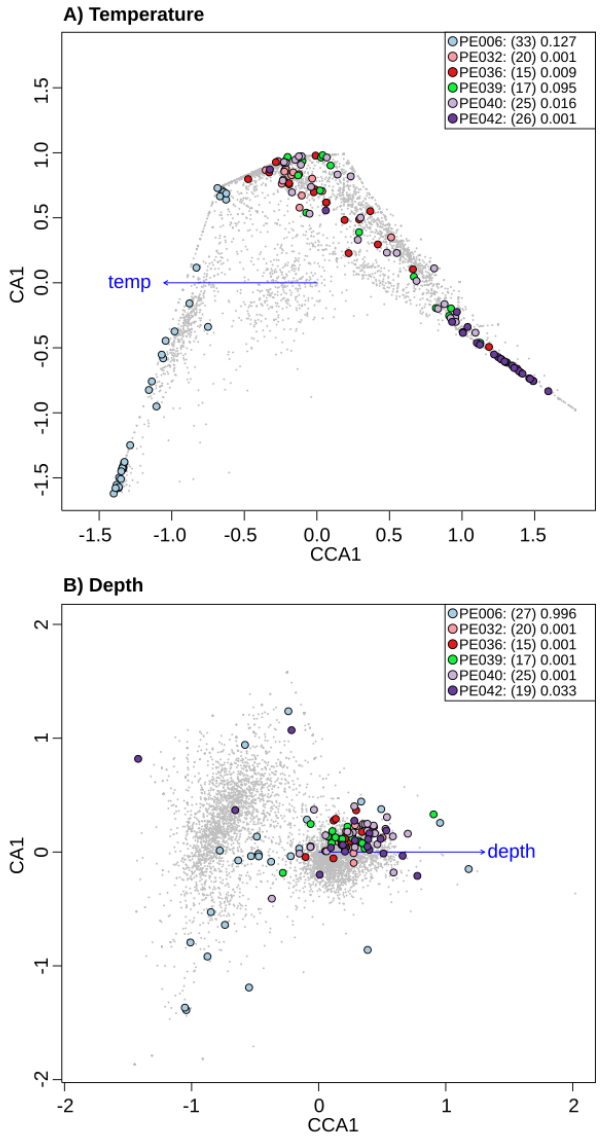


Figure 5.4: '*Candidatus* Roseilinea'-higlighted canonical correspondence analyses of *psaA*, *pscA*, and *pufL/pufLM* high-frequency sequence (HFS) segments. HFSs of predominant '*Ca.* Roseilinea' putative ecotypes (PEs) are highlighted. The temperature profile (A) contains samples from the 68, 65, 62.5, 60, 55, and 50 °C sites. The depth profile (B) contains three replicate samples of the 62.5 °C site that were disected using a cryotome. Small gray glyphs represent HFSs not highlighted here (<2% of the dataset) or from other taxa. The directional arrows represent vectors of influence of water temperature (A) or depth in the mat core (B). PE names in the legend are followed by the number of HFSs making up the PE in parenthesis and a p-value that represents the probability that the observed PE cluster is distributed differently from random.

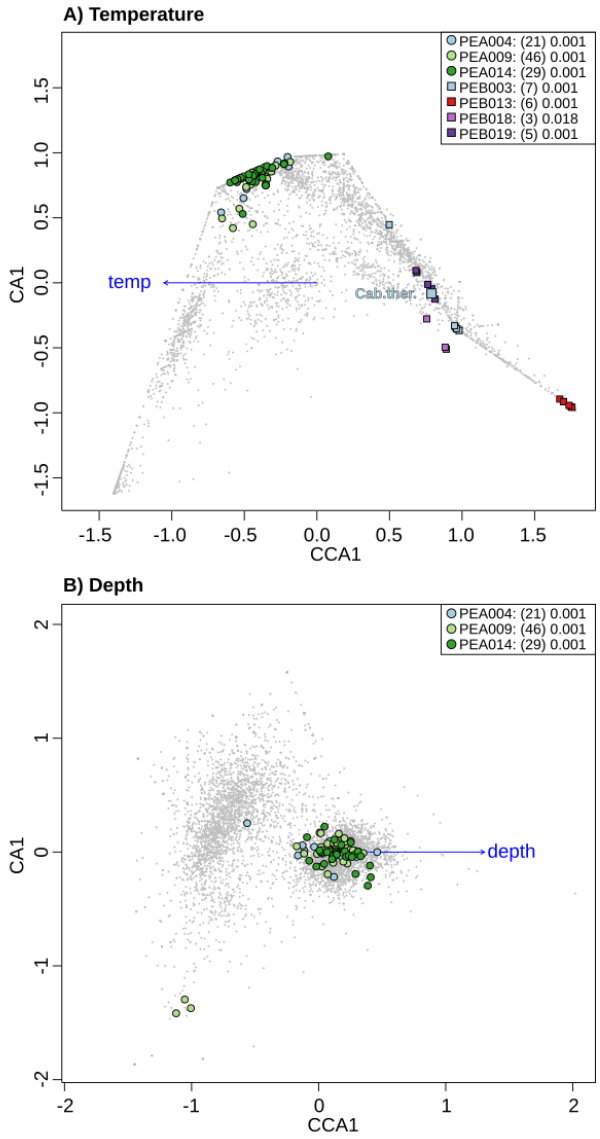


Figure 5.5: *Chloracidobacterium*-highlighted canonical correspondence analyses of *psaA*, *pscA*, and *pufL/pufLM* high-frequency sequence (HFS) segments. HFSs of predominant *Chloracidobacterium* putative ecotypes (PEs) are highlighted. The temperature profile (A) contains samples from the 68, 65, 62.5, 60, 55, and 50 °C sites. The depth profile (B) contains three replicate samples of the 62.5 °C site that were disected using a cryotome. The larger labeled glyph represents the HFS that shares the same *pscA* sequence as *Chloracidobacterium thermophilum* strain B described by Bryant et al. (2007). Small gray glyphs represent HFSs not highlighted here (<2% of the dataset) or from other taxa. The directional arrows represent vectors of influence of water temperature (A) or depth in the mat core (B). PE names in the legend are followed by the number of HFSs making up the PE in parenthesis and a p-value that represents the probability that the observed PE cluster is distributed differently from random.

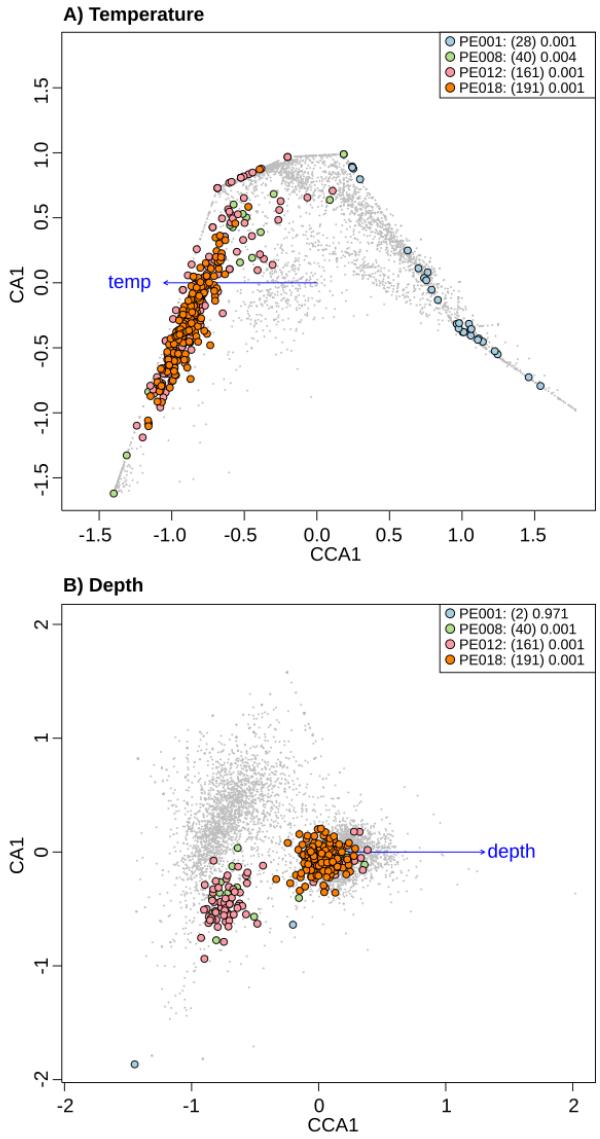


Figure 5.6: '*Candidatus* Thermochlorobacter'-highlighted canonical correspondence analyses of *psaA*, *pscA*, and *pufL/pufLM* high-frequency sequence (HFS) segments. HFSs of predominant '*Ca.* Thermochlorobacter' putative ecotypes (PEs) are highlighted. The temperature profile (A) contains samples from the 68, 65, 62.5, 60, 55, and 50 °C sites. The depth profile (B) contains three replicate samples of the 62.5 °C site that were disected using a cryotome. Small gray glyphs represent HFSs not highlighted here (<2% of the dataset) or from other taxa. The directional arrows represent vectors of influence of water temperature (A) or depth in the mat core (B). PE names in the legend are followed by the number of HFSs making up the PE in parenthesis and a p-value that represents the probability that the observed PE cluster is distributed differently from random.

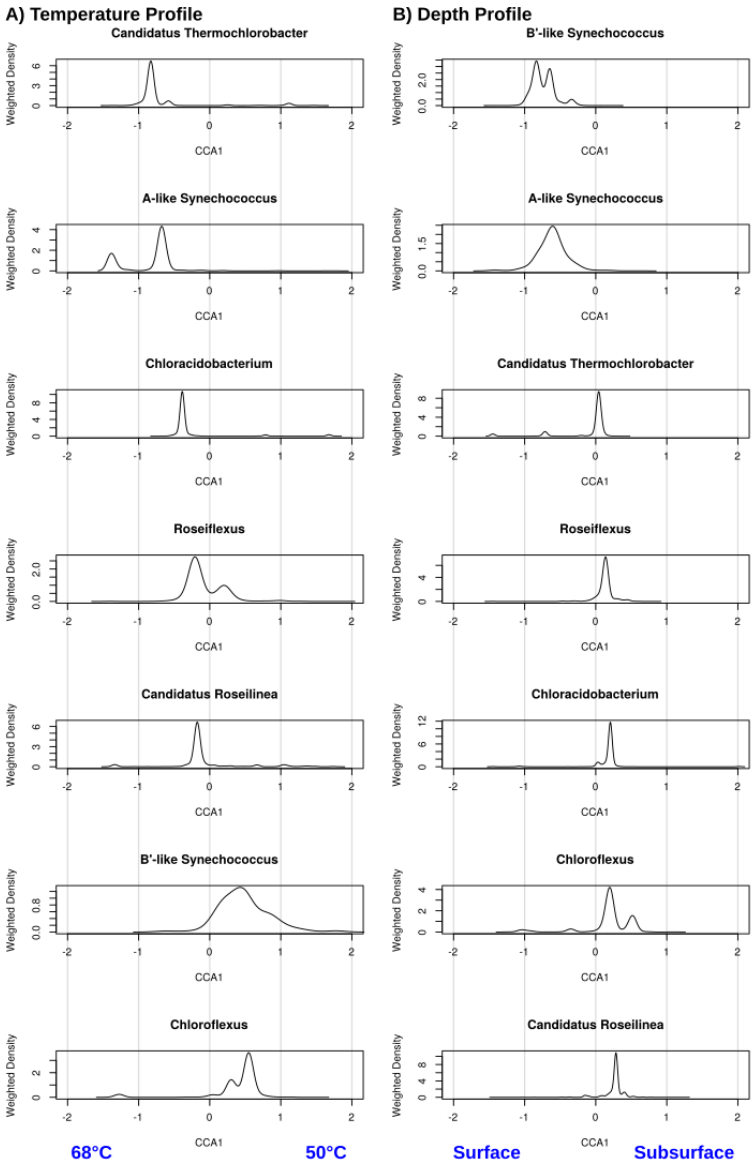


Figure 5.7: Canonical correspondence analyses comparing the distributions of all taxa with respect to water temperature (A) or depth within the mat (B). Taxa are ordered according to their maximum densities along each gradient. The temperature profile (A) is oriented with the warmest temperature water on the left of the plot. The depth profile (B) is oriented with the top of the mat on the left of the plot. The scale of the weighted density axes varies by taxon and analysis, but the scale of the CCA1 axes is held constant for parts A or B to allow for comparison across all taxon studied.

CHAPTER SIX

CONCLUSIONS

The goals of this thesis were to develop a new version of Ecotype Simulation (ES1) that could rapidly analyze a large number of sequences, and to utilize the new algorithm to address fundamental questions in the biology of thermophilic A/B′-lineage *Synechococcus* and other major phototrophic taxa inhabiting the Mushroom Spring microbial mat.

In Chapter 2, I described a new version of the algorithm, Ecotype Simulation 2 that was able to analyze very large sequence datasets that the original ES1 was unable to analyze. Additionally, ES2 simulates the evolution of ecotypes in a dataset much faster than ES1, taking only minutes to predict ecotypes in a dataset that took days with ES1. The putative ecotype (PE) predictions of ES2 compared favorably with those from ES1, but ES2 allowed for analyses of much larger datasets. I also demonstrated that more coverage of sequence diversity provided by newer sequencing technologies is beneficial to the understanding of ecotype diversity. Deep coverage of the genetic diversity in communities has become possible with the most recent sequencing technologies (e.g., Illumina MiSeq), permitting the study of population genetics within species.

In Chapter 3, I utilized ES2 to study the biogeography of American Northwest A/B′-lineage *Synechococcus*. I demonstrated that allopatric processes appear to play a role in the diversification of thermophilic A/B′-lineage *Synechococcus*. A distance of >50 km separating springs was enough to serve as a physical barrier to dispersal between springs, but each spring sampled had its own set of sequence diversity, suggesting that *Synechococcus* populations in each spring are undergoing

independent evolution. The analyses suggested that there is a surprising amount of endemism within springs from the same geothermal basin. I also demonstrated that spring chemistry may be an important driver of diversification within springs, leading to hypotheses about other resources and/or conditions to which A/B′-lineage *Synechococcus* may have adapted. In particular, I discovered low- and high-pH populations that may be adapted to these extremes.

In Chapter 4, I utilized ES2 to study the temporal stability of A/B′-lineage *Synechococcus* populations of Mushroom Spring and Octopus Spring mat communities. Using mat samples collected near the winter and summer solstice, I demonstrated variability in distributions of PEs between seasons, with some PEs found in warmer- or cooler-water, or found deeper or more shallow in the mat during the winter than in the summer. The abundances of PEs also varied with the seasons, with individual PEs increasing or decreasing in abundance. However, mat samples collected over a twenty-five year period demonstrated a surprising stability, in that all predominant A-like *Synechococcus* PEs could be detected back to 1999 and all predominant B′-like PEs could be detected back to 1989. The dominant variant of these PEs could be detected in both Mushroom and Octopus Spring throughout this time period, suggesting that if periodic selection events that favored a more-fit variant occurred during this time the same variant was always selected.

In Chapter 5, I utilized ES2 to study the distribution of PEs in other phototrophic taxa inhabiting the Mushroom Spring mat. Results indicated that PEs predicted in each taxon by ES2 displayed restricted distributions along gradients measured, depth within the mat and the temperature. It was also clear that these other mat phototrophic taxa, and PEs predicted by ES2 within each, distribute differently along these environmental gradients. This analysis revealed that the ecotype concept, the Stable Ecotype Model of species and speciation, and ES2 built

on the model, are useful for understanding the diversity present in all mat community members.

Uncertainties remain from these analyses that will need to be resolved with future research. Putative ecotypes predicted by ES2 that displayed random distributions in the ordination space ($p > 0.05$) need to be examined further to understand the reason for this distribution. We have hypothesized that PEs that displayed this distribution pattern could be the result of i) the PE being sampled outside of its niche (e.g., a recent immigrant not adapted to the niche), ii) recombination in the protein-encoding locus chosen for ecotype demarcation, iii) the protein-encoding locus used did not have enough molecular resolution to observe population differences, or iv) a miscall in the generation of the phylogenetic tree or in the demarcation of ecotypes by ES2.

Interpreting the distributions of PEs in the canonical correspondence analysis ordination space must be done with care. A hierarchy in distributions related to the structure of the clade can be detected. More ancient adaptations to an environmental parameter in a linage can be detected on a coarser scale than more recent adaptations to other environmental parameters. Although a low p-value is indicative of a nonrandom distribution, it may represent an older ecological adaptation. Such large clades are not PEs, but collections of more recently evolved PEs based on some other environmental parameter. The A/B′-lineage *Synechococcus* used in these studies provides a good example of this issue because A-like *Synechococcus* are found at higher temperature than B′-like *Synechococcus* and the split between the two clades is more ancient than adaptations to other environmental parameters studied. Although the maximum-likelihood Poisson tree processes demarcated nearly the entire B′-lineage sampled as a single PE, PTP B02, it was non-randomly distributed ($p = 0.001$; Figure A.6) in the CCA ordination space. Bayesian Poisson tree processes split the B′-lineage into much smaller clades, some of which displayed non-random distributions

in the ordination space ($p < 0.05$; PEs bPTP B13 and bPTP B39 in Supplemental Figure A.7 B). Becraft et al. (2015) also noticed this issue by examining different demarcations based on fine- and coarse-scale analyses by Ecotype Simulation 1, finding that the fine-scale analysis can sometimes break larger low-p-value "PEs" found in coarse-scale analysis into more recently evolved PEs that meet the p-value constraint for species. Thus, p-value alone does not permit one to conclude a PE is an ecological species.

It is therefore imperative that researchers examine the smallest clade that satisfies the requirements of being an ecological species. That is, members of a species must be ecologically interchangeable with other members of the species, but ecologically distinct from other species. A PE that displays a non-random distribution ($p < 0.05$) and that does not overlap with other PEs in the ordination space provides evidence that the PE is an ecological species. While satisfying these requirements may be complicated by inability to measure niche-defining variables in an environment, utilizing algorithms like ES2 that are based on ecological theory can help researchers understand which environmental parameters are important in niche differentiation between ecological species.

This thesis exemplifies an approach that should be helpful in guiding other microbial ecologists to a better understanding of the communities they study. First, genes that offer sufficient molecular resolution to detect species should be targeted. Second, high-throughput sequencing approaches should be used to obtain a deep coverage of the sequence variation in a sample. Third, a theory-based algorithm such as ES2 should be used to predict how variants are grouped into species. Finally, statistical analyses like canonical correspondence analysis should be employed to validate the ecological distinctions and ecological interchangeability of the members of those predicted species.

REFERENCES CITED

Aboal, M., O. Werner, M. E. García-Fernández, J. A. Palazón, J. C. Cristóbal, and W. Williams. 2016. Should ecomorphs be conserved? The case of *Nostoc flagelliforme*, an endangered extremophile cyanobacteria. *Journal for Nature Conservation*, 30:52–64. doi:10.1016/j.jnc.2016.01.001.

Allewalt, J. P., M. M. Bateson, N. P. Revsbech, K. Slack, and D. M. Ward. 2006. Effect 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. *Applied and Environmental Microbiology*, 72:544–550. doi:10.1128/AEM.72.1.544-550.2006.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410. doi:10.1016/S0022-2836(05)80360-2.

Anderson, N. J., T. Korsman, and I. Renberg. 1994. Spatial heterogeneity of diatom stratigraphy in varved and non-varved sediments of a small, boreal-forest lake. *Aquatic Sciences*, 56:40–58. doi:10.1007/BF00877434.

Baas Becking, L. G. M. 1934. *Geobiologie of inleiding tot de milieukunde*. Van Stockum, Den Haag.

Bakker, C., P. M. J. Herman, and M. Vink. 1990. Changes in seasonal succession of phytoplankton induced by the storm-surge barrier in the Oosterschelde (S.W. Netherlands). *Journal of Plankton Research*, 12:947–972. doi:10.1093/plankt/12.5.947.

Becraft, E. D. 2014. *The ecology and evolution of thermophilic Synechococcus species*. PhD thesis, Montana State University, Bozeman, Montana.

Becraft, E. D., F. M. Cohan, M. Kühl, S. I. Jensen, and D. M. Ward. 2011. Fine-scale distribution patterns of *Synechococcus* ecological diversity in microbial mats of Mushroom Spring, Yellowstone National Park. *Applied and Environmental Microbiology*, 77:7689–7697. doi:10.1128/AEM.05927-11.

Becraft, E. D., J. M. Wood, D. B. Rusch, M. Kühl, S. I. Jensen, D. A. Bryant, D. W. Roberts, F. M. Cohan, and D. M. Ward. 2015. The molecular dimension of microbial species: 1. Ecological distinctions among, and homogeneity within, putative ecotypes of *Synechococcus* inhabiting the cyanobacterial mat of Mushroom Spring, Yellowstone National Park. *Frontiers in Microbiology*, 6:590. doi:10.3389/fmicb.2015.00590.

Bhaya, D., A. R. Grossman, A.-S. Steunou, N. Khuri, F. M. Cohan, N. Hamamura, M. C. Melendrez, M. M. Bateson, D. M. Ward, and J. F. Heidelberg.

2007. Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. *The ISME Journal*, 1:703–713. doi:10.1038/ismej.2007.46.

Birks, H. J. B., S. M. Peglar, and H. A. Austin. 1996. An annotated bibliography of canonical correspondence analysis and related constrained ordination methods 1986-1993. *Abstracta Botanica*, 20:17–36.

Bonheyo, G. T., J. Frias-Lopez, and B. W. Fouke. 2005. A test for airborne dispersal of thermophilic bacteria from hot springs. In Inskeep, W. P., and T. R. McDermott, editors, *Geothermal biology and geochemistry in Yellowstone National Park*, pages 327–340. Montana State University Publications, Bozeman, Montana. ISBN 9780963511416.

Brock, M. L., R. G. Wiegert, and T. D. Brock. 1969. Feeding by *Paracoenia* and *Ephydra* (Diptera: Ephydridae) on the microorganisms of hot springs. *Ecology*, 50: 192–200. doi:10.2307/1934846.

Brock, T. D. 1967. Micro-organisms adapted to high temperatures. *Nature*, 214: 882–885. doi:10.1038/214882a0.

Bryant, D. A., A. M. G. Costas, J. A. Maresca, A. G. M. Chew, C. G. Klatt, M. M. Bateson, L. J. Tallon, J. Hostetler, W. C. Nelson, J. F. Heidelberg, and D. M. Ward. 2007. *Candidatus* Chloracidobacterium thermophilum: an aerobic phototrophic Acidobacterium. *Science*, 317:523–526. doi:10.1126/science.1143236.

Castenholz, R. W. 1978. The biogeography of hot springs algae through enrichment cultures. In *Symposium: experimental use of algal culture in limnology 26-28 October 1976, Sandefjord, Norway*, number 21, pages 296–315. Schweizerbart Science Publishers, Stuttgart, Germany. ISBN 9783510520213.

Castenholz, R. W. 1996. Endemism and biodiversity of thermophilic cyanobacteria. In Prasad, A. K., J. A. Nienow, and V. N. Rao, editors, *Nova hedwigia beiheft*, volume 112, pages 33–48. Schweizerbart Science Publishers, Stuttgart, Germany. ISBN 9783443510343.

Choudhary, D. K., and B. N. Johri. 2011. Ecological significance of microdiversity: coexistence among casing soil bacterial strains through allocation of nutritional resource. *Indian Journal of Microbiology*, 51:8–13. doi:10.1007/s12088-011-0068-7.

Cohan, F., and S. Kopac. 2017. A theory-based pragmatism for discovering and classifying newly divergent species of bacterial pathogens. In Tibayrenc, M., editor, *Genetics and evolution of infectious diseases: second edition*, pages 25–49. Elsevier, London. ISBN 9780127999425.

Cohan, F. M. 2006. Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 361:1985–1996. doi:10.1098/rstb.2006.1918.

Cohan, F. M. 2016. Bacterial speciation: genetic sweeps in bacterial species. *Current Biology*, 26:R112–R115. doi:10.1016/j.cub.2015.10.022.

Cohan, F. M., and E. B. Perry. 2007. A systematics for discovering the fundamental units of bacterial diversity. *Current Biology*, 17:R373–R386. doi:10.1016/j.cub.2007.03.032.

Connor, N., J. Sikorski, A. P. Rooney, S. Kopac, A. F. Koeppel, A. Burger, S. G. Cole, E. B. Perry, D. Krizanc, N. C. Field, M. Slaton, and F. M. Cohan. 2010. Ecology of speciation in the genus *Bacillus*. *Applied and Environmental Microbiology*, 76: 1349–1358. doi:10.1128/AEM.01988-09.

Dagum, L., and R. Menon. 1998. OpenMP: an industry standard API for shared-memory programming. *IEEE Computational Science Engineering*, 5:46–55. doi:10.1109/99.660313.

Darwin, C. 1859. *On the origin of species by means of natural selection, or, the preservation of favoured races in the struggle for life*. John Murray, London.

de Queiroz, K. 2005. Ernst Mayr and the modern concept of species. *Proceedings of the National Academy of Sciences*, 102:6600–6607. doi:10.1073/pnas.0502030102.

Dick, G. J., A. F. Andersson, B. J. Baker, S. L. Simmons, B. C. Thomas, A. P. Yelton, and J. F. Banfield. 2009. Community-wide analysis of microbial genome sequence signatures. *Genome Biology*, 10:1–16. doi:10.1186/gb-2009-10-8-r85.

Doolittle, W. F., and O. Zhaxybayeva. 2009. On the origin of prokaryotic species. *Genome Research*, 19:744–756. doi:10.1101/gr.086645.108.

Dugat, T., A.-C. Lagrée, R. Maillard, H.-J. Boulouis, and N. Haddad. 2015. Opening the black box of *Anaplasma phagocytophilum* diversity: current situation and future perspectives. *Frontiers in Cellular and Infection Microbiology*, 5:61. doi:10.3389/fcimb.2015.00061.

Duncan, K. E., C. A. Istock, and J. B. Graham. 1989. Genetic exchange between *Bacillus subtilis* and *Bacillus licheniformis*: variable hybrid stability and the nature of bacterial species. *Evolution*, 43:1585–1609. doi:10.2307/2409377.

DuRand, M. D., R. J. Olson, and S. W. Chisholm. 2001. Phytoplankton population dynamics at the Bermuda Atlantic Time-series station in the Sargasso Sea. *Deep Sea Research Part II: Topical Studies in Oceanography*, 48:1983–2003. doi:10.1016/S0967-0645(00)00166-1.

Dykhuizen, D. E. 1998. Santa Rosalia revisited: why are there so many species of bacteria? *Antonie van Leeuwenhoek*, 73:25–33. doi:10.1023/A:1000665216662.

Falush, D., M. Stephens, and J. K. Pritchard. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164:1567–1587. URL http://www.genetics.org/content/164/4/1567.

Feil, E. J., M. C. Maiden, M. Achtman, and B. G. Spratt. 1999. The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. *Molecular Biology and Evolution*, 16:1496–1502. doi:10.1093/oxfordjournals.molbev.a026061.

Feil, E. J., J. M. Smith, M. C. Enright, and B. G. Spratt. 2000. Estimating recombinational parameters in *Streptococcus pneumoniae* from multilocus sequence typing data. *Genetics*, 154:1439–1450.

Felsenstein, J. 2005. PHYLIP (phylogeny inference package) version 3.6. Distributed by the author. URL http://evolution.genetics.washington.edu/phylip.html.

Ferris, M. J., and D. M. Ward. 1997. Seasonal distributions of dominant 16S rRNA-defined populations in a hot spring microbial mat examined by denaturing gradient gel electrophoresis. *Applied and Environmental Microbiology*, 63:1375–81.

Ferris, M. J., G. Muyzer, and D. M. Ward. 1996. Denaturing gradient gel electrophoresis profiles of 16S rRNA-defined populations inhabiting a hot spring microbial mat community. *Applied and Environmental Microbiology*, 62:340–346.

Ferris, M. J., M. Kühl, A. Wieland, and D. M. Ward. 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. *Applied and Environmental Microbiology*, 69:2893–2898. doi:10.1128/AEM.69.5.2893-2898.2003.

Finlay, B. J., and T. Fenchel. 2004. Cosmopolitan metapopulations of free-living microbial Eukaryotes. *Protist*, 155:237–244. doi:10.1078/143446104774199619.

Francisco, J. C., F. M. Cohan, and D. Krizanc. 2014. Accuracy and efficiency of algorithms for the demarcation of bacterial ecotypes from DNA sequence data. *International Journal of Bioinformatics Research and Applications*, 10:409–425. doi:10.1504/IJBRA.2014.062992.

Fritsch, A. 2012. *mcclust: Process an MCMC Sample of Clusterings*. URL https://CRAN.R-project.org/package=mcclust. R package version 1.0.

Garcia Costas, A. M., Z. Liu, L. P. Tomsho, S. C. Schuster, D. M. Ward, and D. A. Bryant. 2012. Complete genome of *Candidatus* Chloracidobacterium thermophilum, a chlorophyll-based photoheterotroph belonging to the phylum

Acidobacteria. *Environmental Microbiology*, 14:177–190. doi:10.1111/j.1462-2920.2011.02592.x.

García-Martínez, J., S. G. Acinas, R. Massana, and F. Rodríguez-Valera. 2002. Prevalence and microdiversity of *Alteromonas macleodii*-like microorganisms in different oceanic regions. *Environmental Microbiology*, 4:42–50. doi:10.1046/j.1462-2920.2002.00255.x.

Giovannoni, S. J., and K. L. Vergin. 2012. Seasonality in ocean microbial communities. *Science*, 335:671–676. doi:10.1126/science.1198078.

Hamilton, C. M., M. Aldea, B. K. Washburn, P. Babitzke, and S. R. Kushner. 1989. New method for generating deletions and gene replacements in *Escherichia coli*. *Journal of Bacteriology*, 171:4617–4622.

Hunt, D. E., L. A. David, D. Gevers, S. P. Preheim, E. J. Alm, and M. F. Polz. 2008. Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science*, 320:1081–1085. doi:10.1126/science.1157890.

Jaspers, E., and J. Overmann. 2004. Ecological significance of microdiversity: identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies. *Applied and Environmental Microbiology*, 70:4831–4839. doi:10.1128/AEM.70.8.4831-4839.2004.

Junqueira, A. C. M., A. Ratan, E. Acerbi, D. I. Drautz-Moses, B. N. V. Premkrishnan, P. I. Costea, B. Linz, R. W. Purbojati, D. F. Paulo, N. E. Gaultier, P. Subramanian, N. A. Hasan, R. R. Colwell, P. Bork, A. M. L. Azeredo-Espin, D. A. Bryant, and S. C. Schuster. 2017. The microbiomes of blowflies and houseflies as bacterial transmission reservoirs. *Scientific Reports*, 7:1–15. doi:10.1038/s41598-017-16353-x.

Kettler, G. C., A. C. Martiny, K. Huang, J. Zucker, M. L. Coleman, S. Rodrigue, F. Chen, A. Lapidus, S. Ferriera, J. Johnson, C. Steglich, G. M. Church, P. Richardson, and S. W. Chisholm. 2007. Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLOS Genetics*, 3:1–14. doi:10.1371/journal.pgen.0030231.

Kilian, O., A.-S. Steunou, F. Fazeli, S. Bailey, D. Bhaya, and A. R. Grossman. 2007. Responses of a thermophilic *Synechococcus* isolate from the microbial mat of Octopus Spring to light. *Applied and Environmental Microbiology*, 73:4268–4278. doi:10.1128/AEM.00201-07.

Kim, Y.-M., S. Nowack, M. T. Olsen, E. D. Becraft, J. M. Wood, V. Thiel, I. Klapper, M. Kühl, J. K. Fredrickson, D. A. Bryant, D. M. Ward, and T. O. Metz. 2015. Diel metabolomics analysis of a hot spring chlorophototrophic microbial mat leads

to new hypotheses of community member metabolisms. *Frontiers in Microbiology*, 6:209. doi:10.3389/fmicb.2015.00209.

Klatt, C. G., J. M. Wood, D. B. Rusch, M. M. Bateson, N. Hamamura, J. F. Heidelberg, A. R. Grossman, D. Bhaya, F. M. Cohan, M. Kühl, D. A. Bryant, and D. M. Ward. 2011. Community ecology of hot spring cyanobacterial mats: predominant populations and their functional potential. *The ISME Journal*, 5: 1262–1278. doi:10.1038/ismej.2011.73.

Klatt, C. G., Z. Liu, M. Ludwig, M. Kuhl, S. I. Jensen, D. A. Bryant, and D. M. Ward. 2013. Temporal metatranscriptomic patterning in phototrophic Chloroflexi inhabiting a microbial mat in a geothermal spring. *The ISME Journal*, 7:1775–1789. doi:10.1038/ismej.2013.52.

Koeppel, A., E. B. Perry, J. Sikorski, D. Krizanc, A. Warner, D. M. Ward, A. P. Rooney, E. Brambilla, N. Connor, R. M. Ratcliff, E. Nevo, and F. M. Cohan. 2008. Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. *Proceedings of the National Academy of Sciences*, 105:2504–2509. doi:10.1073/pnas.0712205105.

Konstantinidis, K. T., and J. M. Tiedje. 2005. Towards a genome-based taxonomy for prokaryotes. *Journal of Bacteriology*, 187:6258–6264. doi:10.1128/JB.187.18.6258-6264.2005.

Kopac, S., Z. Wang, J. Wiedenbeck, J. Sherry, M. Wu, and F. M. Cohan. 2014. Genomic heterogeneity and ecological speciation within one subspecies of *Bacillus subtilis*. *Applied and Environmental Microbiology*, 80:4842–4853. doi:10.1128/AEM.00576-14.

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23:2947–2948. doi:10.1093/bioinformatics/btm404.

Lefébure, T., and M. J. Stanhope. 2007. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome Biology*, 8:R71. doi:10.1186/gb-2007-8-5-r71.

Legendre, P., and L. Legendre. 1998. *Numerical Ecology*. Elsevier, Amsterdam. ISBN 9780444538697.

Liu, Z., C. G. Klatt, J. M. Wood, D. B. Rusch, M. Ludwig, N. Wittekindt, L. P. Tomsho, S. C. Schuster, D. M. Ward, and D. A. Bryant. 2011. Metatranscriptomic analyses of chlorophototrophs of a hot-spring microbial mat. *The ISME Journal*, 5:1279–1290. doi:10.1038/ismej.2011.37.

Liu, Z., C. G. Klatt, M. Ludwig, D. B. Rusch, S. I. Jensen, M. Kühl, D. M. Ward, and D. A. Bryant. 2012. '*Candidatus* Thermochlorobacter aerophilum:' an aerobic chlorophotoheterotrophic member of the phylum *Chlorobi* defined by metagenomics and metatranscriptomics. *The ISME Journal*, 6:1869–1882. doi:10.1038/ismej.2012.24.

Lloyd, S. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28:129–137. doi:10.1109/TIT.1982.1056489.

Madigan, M. T., D. O. Jung, E. A. Karr, W. M. Sattley, L. A. Achenbach, and M. T. J. van der Meer. 2005. Diversity of anoxygenic phototrophs in contrasting extreme environments. In Inskeep, W. P., and T. R. McDermott, editors, *Geothermal biology and geochemistry in Yellowstone National Park*, pages 203–219. Montana State University Publications, Bozeman, Montana. ISBN 9780963511416.

Maiden, M. C. J., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman, and B. G. Spratt. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences*, 95:3140–3145.

Majewski, J., and F. M. Cohan. 1999. Adapt globally, act locally: the effect of selective sweeps on bacterial sequence diversity. *Genetics*, 152:1459–1474.

Markus, M. B. 1980. Flies as natural transport hosts of *Sarcocystis* and other Coccidia. *The Journal of Parasitology*, 66:361–362. doi:10.2307/3280842.

Marri, P. R., W. Hao, and G. B. Golding. 2006. Gene gain and gene loss in *Streptococcus*: is it driven by habitat? *Molecular Biology and Evolution*, 23:2379–2391. doi:10.1093/molbev/msl115.

Marsaglia, G., and W. W. Tsang. 2000. The ziggurat method for generating random variables. *Journal of Statistical Software*, 5:1–7. doi:10.18637/jss.v005.i08.

Martiny, J. B. H., B. J. M. Bohannan, J. H. Brown, R. K. Colwell, J. A. Fuhrman, J. L. Green, M. C. Horner-Devine, M. Kane, J. A. Krumins, C. R. Kuske, P. J. Morin, S. Naeem, L. Ovreas, A.-L. Reysenbach, V. H. Smith, and J. T. Staley. 2006. Microbial biogeography: putting microorganisms on the map. *Nature Reviews Microbiology*, 4:102–112. doi:10.1038/nrmicro1341.

Mayr, E. 1982. *The growth of biological thought: diversity, evolution and inheritance*. Belknap Press, Cambridge, Massachusetts. ISBN 9780674364455.

McDougald, D., S. A. Rice, N. Barraud, p. D. Steinberg, and S. Kjelleberg. 2012. Should we stay or should we go: mechanisms and ecological consequences for biofilm dispersal. *Nature Reviews Microbiology*, 10:39–50. doi:0.1038/nrmicro2695.

Melendrez, M. C., R. K. Lange, F. M. Cohan, and D. M. Ward. 2011. Influence of molecular resolution on sequence-based discovery of ecological diversity among *Synechococcus* populations in an alkaline siliceous hot spring microbial mat. *Applied and Environmental Microbiology*, 77:1359–1367. doi:10.1128/AEM.02032-10.

Melendrez, M. C., E. D. Becraft, J. M. Wood, M. T. Olsen, D. A. Bryant, J. F. Heidelberg, D. B. Rusch, F. M. Cohan, and D. M. Ward. 2016. Recombination does not hinder formation or detection of ecological species of *Synechococcus* inhabiting a hot spring cyanobacterial mat. *Frontiers in Microbiology*, 6:1540. doi:10.3389/fmicb.2015.01540.

Miller, S. R., A. L. Strong, K. L. Jones, and M. C. Ungerer. 2009. Bar-coded pyrosequencing reveals shared bacterial community properties along the temperature gradients of two alkaline hot springs in Yellowstone National Park. *Applied and Environmental Microbiology*, 75:4565–4572. doi:10.1128/AEM.02792-08.

Nelder, J., and R. Mead. 1965. A simplex method for function minimization. *The Computer Journal*, 7:308–313. doi:10.1093/comjnl/7.4.308.

Nordstrom, D. K., J. W. Ball, and R. B. McCleskey. 2005. Ground water to surface water: chemistry of thermal outflows in Yellowstone National Park. In Inskeep, W. P., and T. R. McDermott, editors, *Geothermal biology and geochemistry in Yellowstone National Park*, pages 73–94. Montana State University Publications, Bozeman, Montana. ISBN 9780963511416.

Nowack, S., M. T. Olsen, G. A. Schaible, E. D. Becraft, G. Shen, I. Klapper, D. A. Bryant, and D. M. Ward. 2015. The molecular dimension of microbial species: 2. *Synechococcus* strains representative of putative ecotypes inhabiting different depths in the Mushroom Spring microbial mat exhibit different adaptive and acclimative responses to light. *Frontiers in Microbiology*, 6:626. doi:10.3389/fmicb.2015.00626.

Nübel, U., M. M. Bateson, V. Vandieken, A. Wieland, M. Kühl, and D. M. Ward. 2002. Microscopic examination of distribution and phenotypic properties of phylogenetically diverse *Chloroflexaceae*-related bacteria in hot spring microbial mats. *Applied and Environmental Microbiology*, 68:4593–4603. doi:10.1128/AEM.68.9.4593-4603.2002.

Ochman, H., J. G. Lawrence, and E. A. Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405:299–304. doi:10.1038/35012500.

Oh, P. L., A. K. Benson, D. A. Peterson, P. B. Patil, E. N. Moriyama, S. Roos, and J. Walter. 2010. Diversification of the gut symbiont *Lactobacillus reuteri* as a result of host-driven evolution. *The ISME Journal*, 4:377–387. doi:10.1038/ismej.2009.123.

Oksanen, J., F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, and H. Wagner. 2013. *vegan: community ecology package*. URL https://CRAN.R-project.org/package=vegan. R package version 2.0-10.

Olsen, M. T. 2015. Comparative genomic analyses of Yellowstone hot spring microbial mat *Synechococcus* spp. M.S. thesis, Montana State University, Bozeman, Montana.

Olsen, M. T., S. Nowack, J. M. Wood, E. D. Becraft, K. LaButti, A. Lipzen, J. Martin, W. S. Schackwitz, D. B. Rusch, F. M. Cohan, D. A. Bryant, and D. M. Ward. 2015. The molecular dimension of microbial species: 3. Comparative genomics of *Synechococcus* strains with different light responses and *in situ* diel transcription patterns of associated ecotypes in the Mushroom Spring microbial mat. *Frontiers in Microbiology*, 6:604. doi:10.3389/fmicb.2015.00604.

Papke, R. T., N. B. Ramsing, M. M. Bateson, and D. M. Ward. 2003. Geographical isolation in hot spring cyanobacteria. *Environmental Microbiology*, 5:650–659. doi:10.1046/j.1462-2920.2003.00460.x.

Paul, S., A. Dutta, S. K. Bag, S. Das, and C. Dutta. 2010. Distinct, ecotype-specific genome and proteome signatures in the marine cyanobacteria *Prochlorococcus*. *BMC Genomics*, 11:103. doi:10.1186/1471-2164-11-103.

Price, M. N., P. S. Dehal, and A. P. Arkin. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, 26:1641–1650. doi:10.1093/molbev/msp077.

Ramsing, N. B., M. J. Ferris, and D. M. Ward. 2000. Highly ordered vertical structure of *Synechococcus* populations within the one-millimeter-thick photic zone of a hot spring cyanobacterial mat. *Applied and Environmental Microbiology*, 66:1038–1049. doi:10.1128/AEM.66.3.1038-1049.2000.

Rappé, M. S., and S. J. Giovannoni. 2003. The uncultured microbial majority. *Annual Review of Microbiology*, 57:369–394. doi:10.1146/annurev.micro.57.030502.090759.

Resh, V. H., and M. A. Barnby. 1984. Distribution of shore bugs and shore flies at Sylvan Springs, Yellowstone National Park. *The Great Basin Naturalist*, 44:99–103. doi:10.2307/41712046.

Revsbech, N. P., E. Trampe, M. Lichtenberg, D. M. Ward, and M. Kühl. 2016. In situ hydrogen dynamics in a hot spring microbial mat during a diel cycle. *Applied and Environmental Microbiology*, 82:4209–4217. doi:10.1128/AEM.00710-16.

Roberts, D. W. 2017. Lab 12 - canonical correspondence analysis. URL http://ecology.msu.montana.edu/labdsv/R/labs/lab12/lab12.html. Accessed: 2017-02-27.

Roberts, M. S., and F. M. Cohan. 1995. Recombination and migration rates in natural populations of *Bacillus subtilis* and *Bacillus mojavensis*. *Evolution*, 49:1081–1094. doi:10.1111/j.1558-5646.1995.tb04435.x.

Robinson, D. A., J. C. Thomas, and W. P. Hanage. 2011. Population structure of pathogenic bacteria. In Tibayrenc, M., editor, *Genetics and evolution of infectious diseases*, pages 43–57. Elsevier, London. ISBN 9780127999425.

Selander, R. K., J. M. Musser, D. A. Caugant, M. N. Gilmour, and T. S. Whittam. 1987. Population genetics of pathogenic bacteria. *Microbial Pathogenesis*, 3:1–7. doi:10.1016/0882-4010(87)90032-5.

Smith, J. M., C. G. Dowson, and B. G. Spratt. 1991. Localized sex in bacteria. *Nature*, 349:29–31. doi:10.1038/349029a0.

Smith, J. M., N. H. Smith, M. O'Rourke, and B. G. Spratt. 1993. How clonal are bacteria? *Proceedings of the National Academy of Sciences*, 90:4384–4388. doi:10.1073/pnas.90.10.4384.

Snoeijs, P. J. M., and I. C. Prentice. 1989. Effects of cooling water discharge on the structure and dynamics of epilithic algal communities in the northern Baltic. *Hydrobiologia*, 184:99–123. doi:10.1007/BF00014306.

Stackebrandt, E., and J. Ebers. 2006. Taxonomic parameters revisited: tarnished gold standards. *Microbiology Today*, 33:152–155.

Tank, M., V. Thiel, D. M. Ward, and D. A. Bryant. 2017. A panoply of phototrophs: an overview of the thermophilic chlorophototrophs of the microbial mats of alkaline siliceous hot springs in Yellowstone National Park, WY, USA. In Hallenbeck, P. C., editor, *Modern topics in the phototrophic prokaryotes: environmental and applied aspects*, pages 87–137. Springer International Publishing, Switzerland. doi:10.1007/978-3-319-46261-5_3.

Taylor, M. W., P. J. Schupp, R. De Nys, S. Kjelleberg, and P. D. Steinberg. 2005. Biogeography of bacteria associated with the marine sponge *Cymbastela concentrica*. *Environmental Microbiology*, 7:419–433. doi:10.1111/j.1462-2920.2004.00711.x.

ter Braak, C. J. F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67:1167–1179. doi:10.2307/1938672.

Thiel, V., T. L. Hamilton, L. P. Tomsho, R. Burhans, S. E. Gay, S. C. Schuster, D. M. Ward, and D. A. Bryant. 2014. Draft genome sequence of a sulfide-oxidizing, autotrophic filamentous anoxygenic phototrophic bacterium, *Chloroflexus* sp. strain MS-G (*Chloroflexi*). *Genome Announcements*, 2. doi:10.1128/genomeA.00872-14.

Thiel, V., J. M. Wood, M. T. Olsen, M. Tank, C. G. Klatt, D. M. Ward, and D. A. Bryant. 2016. The dark side of the Mushroom Spring microbial mat: life in the shadow of chlorophototrophs. I. Microbial diversity based on 16S rRNA amplicons and metagenomic sequencing. *Frontiers in Microbiology*, 7:919. doi:10.3389/fmicb.2016.00919.

Thompson, J. R., M. A. Randa, L. A. Marcelino, A. Tomita-Mitchell, E. Lim, and M. F. Polz. 2004. Diversity and dynamics of a North Atlantic coastal *Vibrio* community. *Applied and Environmental Microbiology*, 70:4103–4110. doi:10.1128/AEM.70.7.4103-4110.2004.

van der Meer, M. T. J., C. G. Klatt, J. M. Wood, D. A. Bryant, M. M. Bateson, L. Lammerts, S. Schouten, J. S. Sinninghe Damsté, M. T. Madigan, and D. M. Ward. 2010. Cultivation and genomic, nutritional, and lipid biomarker characterization of *Roseiflexus* strains closely related to predominant *in situ* populations inhabiting Yellowstone hot spring microbial mats. *Journal of Bacteriology*, 192: 3033–3042. doi:10.1128/JB.01610-09.

van Niel, C. B., and L. A. Thayer. 1930. Report on preliminary observations on the microflora in and near the hot springs in Yellowstone National Park and their importance for the geological formations. Technical Report 731.2, National Park Service.

Van Spanning, R. J. M., W. N. M. Reijnders, and A. H. Stouthamer. 1995. Integration of heterologous DNA into the genome of *Paracoccus denitrificans* is mediated by a family of IS1248-related elements and a second type of integrative recombination event. *Journal of Bacteriology*, 177:4772–4778. doi:10.1128/jb.177.16.4772-4778.1995.

Van Valen, L. 1976. Ecological species, multispecies, and oaks. *Taxon*, 25:233–239. doi:10.2307/1219444.

Vernikos, G. S., N. R. Thomson, and J. Parkhill. 2007. Genetic flux over time in the *Salmonella* lineage. *Genome Biology*, 8:R100. doi:10.1186/gb-2007-8-6-r100.

Vulić, M., F. Dionisio, F. Taddei, and M. Radman. 1997. Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proceedings of the National Academy of Sciences*, 94:9763–9767.

Ward, D. M. 1998. A natural species concept for prokaryotes. *Current Opinion in Microbiology*, 1:271–277. doi:10.1016/S1369-5274(98)80029-5.

Ward, D. M. 2006. A macrobiological perspective on microbial species. *Microbe*, 1: 269–278. doi:10.1128/microbe.1.269.1.

Ward, D. M., and R. W. Castenholz. 2002. Cyanobacteria in geothermal habitats. In Whitton, B. A., and M. Potts, editors, *The ecology of cyanobacteria: their diversity in time and space*, pages 37–59. Springer Netherlands, Dordrecht. doi:10.1007/0-306-46855-7_3.

Ward, D. M., R. Weller, and M. M. Bateson. 1990. 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature*, 345:63–65. doi:10.1038/345063a0.

Ward, D. M., M. J. Ferris, S. C. Nold, and M. M. Bateson. 1998. A natural view of microbial biodiversity within hot spring cyanobacterial mat communities. *Microbiology and Molecular Biology Reviews*, 62:1353–1370.

Ward, D. M., M. M. Bateson, M. J. Ferris, M. Kühl, A. Wieland, A. Koeppel, and F. M. Cohan. 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. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 361:1997–2008. doi:10.1098/rstb.2006.1919.

Ward, D. M., C. G. Klatt, J. M. Wood, F. M. Cohan, and D. A. Bryant. 2012. Functional genomics in an ecological and evolutionary context: maximizing the value of genomes in systems biology. In Burnap, R., and W. Vermaas, editors, *Functional genomics and evolution of photosynthetic systems*, volume 33, pages 1–16. Springer Netherlands, Dordrecht. doi:10.1007/978-94-007-1533-2_1.

Wayne, L. G., D. J. Brenner, R. R. Colwell, P. A. Grimont, O. Kandler, M. I. Krichevsky, W. E. C. Moore, R. G. E. Murray, E. Stackebrandt, M. P. Starr, and H. G. Truper. 1987. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *International Journal of Systematic Bacteriology*, 37:463–464. doi:10.1099/00207713-37-4-463.

Weltzer, M. L., and S. R. Miller. 2013. Ecological divergence of a novel group of *Chloroflexus* strains along a geothermal gradient. *Applied and Environmental Microbiology*, 79:1353–1358. doi:10.1128/AEM.02753-12.

Whitaker, R. J. 2006. Allopatric origins of microbial species. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 361:1975–1984. doi:10.1098/rstb.2006.1927.

Whitaker, R. J., D. W. Grogan, and J. W. Taylor. 2003. Geographic barriers isolate endemic populations of hyperthermophilic Archaea. *Science*, 301:976–978. doi:10.1126/science.1086909.

Zhang, J., P. Kapli, P. Pavlidis, and A. Stamatakis. 2013. A general species delimitation method with applications to phylogenetic placements. *Bioinformatics*, 29:2869–2876. doi:10.1093/bioinformatics/btt499.

APPENDICES

APPENDIX A

CHAPTER 2 SUPPLEMENTAL INFORMATION

Analysis of B′ *Synechococcus* Lineage

Methods

B′-like *Synechococcus* high-frequency sequences (HFSs) were treated exactly like the A-like HFSs presented in the main text.

Results and Discussion

As with the A-lineage presented in the main text, binning results for the B′-lineage are most similar between the Ecotype Simulation 1 (ES1) matrix-based method and the Ecotype Simulation 2 (ES2) result generated with a neighbor-joining tree (Supplemental Figure A.1). As seen in the A-lineage in the main text, overlap in ES1 fine- and coarse-scale ecotype demarcation often means overlap with the ES2 demarcation (compare the dark-blue ES2 demarcation lines to the light-blue coarse-scale ES1 demarcation and the medium-blue fine-scale ES1 lines in Supplemental Figures A.2 and A.3; 15 B′-like PEs in the neighbor-joining tree, and 13 PEs in the maximum-likelihood tree). When the ES1 coarse- and fine-scale methods don't overlap in ecotype demarcation of the B′-lineage, the ES2 demarcation often matches either the ES1 coarse-scale demarcation (7 PEs in the maximum-likelihood tree) or the fine-scale demarcation (7 PEs in the neighbor-joining tree, and 5 PEs in the maximum-likelihood tree).

As with the A-lineage, analysis of the entire B′-lineage HFS_{10} dataset yielded a much larger number of PEs than those described here (Supplementary Figure A.4), though most were rare contributors to the dataset and are likely to be rare members of the mat community. Significant clustering in canonical correspondence analysis (CCA) ordination space was observed in highly abundant PE PEB104 (B′9), and less abundant PEs PEB036 (B′7) and PEB023 (B′8) (Supplemental Figure A.4). Other

highly abundant B′-like PEs either didn't have enough HFS members to analyze with CCA (PEs PEB105 and PEB079), or displayed disjoint distributions (i.e., $p > 0.05$). The B′-lineage has been hypothesized to be older, have more sequence diversity, and exhibits more evidence of recombination than the A-lineage (Melendrez et al., 2016), and as mentioned in the main text, recombination between distantly related PEs might explain the higher frequency of disjointly distributed B′-like PEs.

As with the runtime and rarefaction curves produced for the A-lineage in the main text, ES2 was able to predict ecotype membership for the B′-lineage in less than 90 minutes when analyzing 6400 sequences (see the blue line in Supplemental Figure A.5). The runtime curve was nearly linear with the number of environmental sequences analyzed during this test, illustrating that analyses of even larger datasets is possible. Like the A-lineage, it appeared that the number of PEs detected by ES2 was leveling off, suggesting that a sample of 6,474 unique sequences revealed nearly the total diversity in B′-like *Synechococcus* in this community (see the orange line in Supplemental Figure A.5). Like the A-lineage, our analyses predicted the existence of about 1000 PEs with more than a single representative in the B′-lineage.

Analysis of Parameter Estimation Changes

Methods

The new Lloyd's algorithm-based parameter-estimation algorithm of ES2 was compared with the brute-force approach employed in ES1. Since the brute-force approach is dependent on the underlying simulation method, the brute-force approach was tested with the original ES1 simulation method in addition to the new ES2 simulation method to provide a better comparison to the new Lloyd's algorithm-based approach.

Results and Discussion

ES2 using either the brute-force method or the Lloyd's algorithm-based method provides a lower estimate of the rate of ecotype formation (*omega*) than ES1 in all data tested (Supplemental Table A.1). The rate of periodic selection (*sigma*) estimates were much more varied when the brute-force approach was used rather than the Lloyd's algorithm-based approach, with returned values differing wildly between runs (not displayed). Although the estimated value of *sigma* was highly variable between runs with the brute-force method, the ES2 Lloyd's algorithm-based method produced an estimate approximately within the range produced by the brute-force method (Supplemental Table A.1). As with *sigma*, the estimated number of putative ecotypes (*npop*) was highly varied over multiple runs when using the brute-force approach (not displayed). The Lloyd's algorithm-based method provided a lower *npop* estimate than the ES1 or ES2 brute-force methods (Supplemental Table A.1). Not surprisingly, by increasing the number of sequences analyzed in dataset (by going from HFS_{50} to HFS_{10}) the estimated number of putative ecotypes increased, regardless of the parameter estimation method used (Supplemental Table A.1).

Canonical Correspondence Analyses With Demarcation From Other Algorithms

Methods

Canonical correspondence analyses were run using demarcations produced with both maximum-likelihood Poisson tree processes (PTP) and Bayesian Poisson tree processes (bPTP) on the HFS_{50} set of sequences using the same methods outlined in the main text.

Results and Discussion

Most of the predominant PEs (those that make up >1% of the dataset) demarcated by PTP display non-random distributions with respect to the measured temperature and depth linear predictors ($p < 0.05$; Supplemental Figure A.6). Many of these non-randomly distributed PEs demarcated by PTP are the same as demarcated by either ES1 or ES2 (PTP neighbor-joining B′-like PEs PEB11, PEB16, PEB29, and PEB31; and maximum-likelihood A-like PEs PEA11 and PEA15). Caution must be taken when interpreting these results, however. For instance, the PTP PEB02 clade includes nearly all B′-like HFSs and the low p-value is likely attributed to the adaptation of B′-like species to lower temperatures. Species should be thought of as the smallest clades meeting the expectations of ecological distinction and ecological interchangeability.

The bPTP algorithm was less successful at demarcating predominant PEs with non-random distributions (Supplemental Figure A.7). The neighbor-joining bPTP analysis predicted only 2 predominant PEs that had enough HFS members to analyze with CCA and that displayed non-random distributions (A-like bPTP PE 08 and B′-like 41). Although the maximum-likelihood bPTP analysis predicted 7 predominant PEs, only 2 of them displayed non-random distributions with respect to the depth and temperature linear predictors.

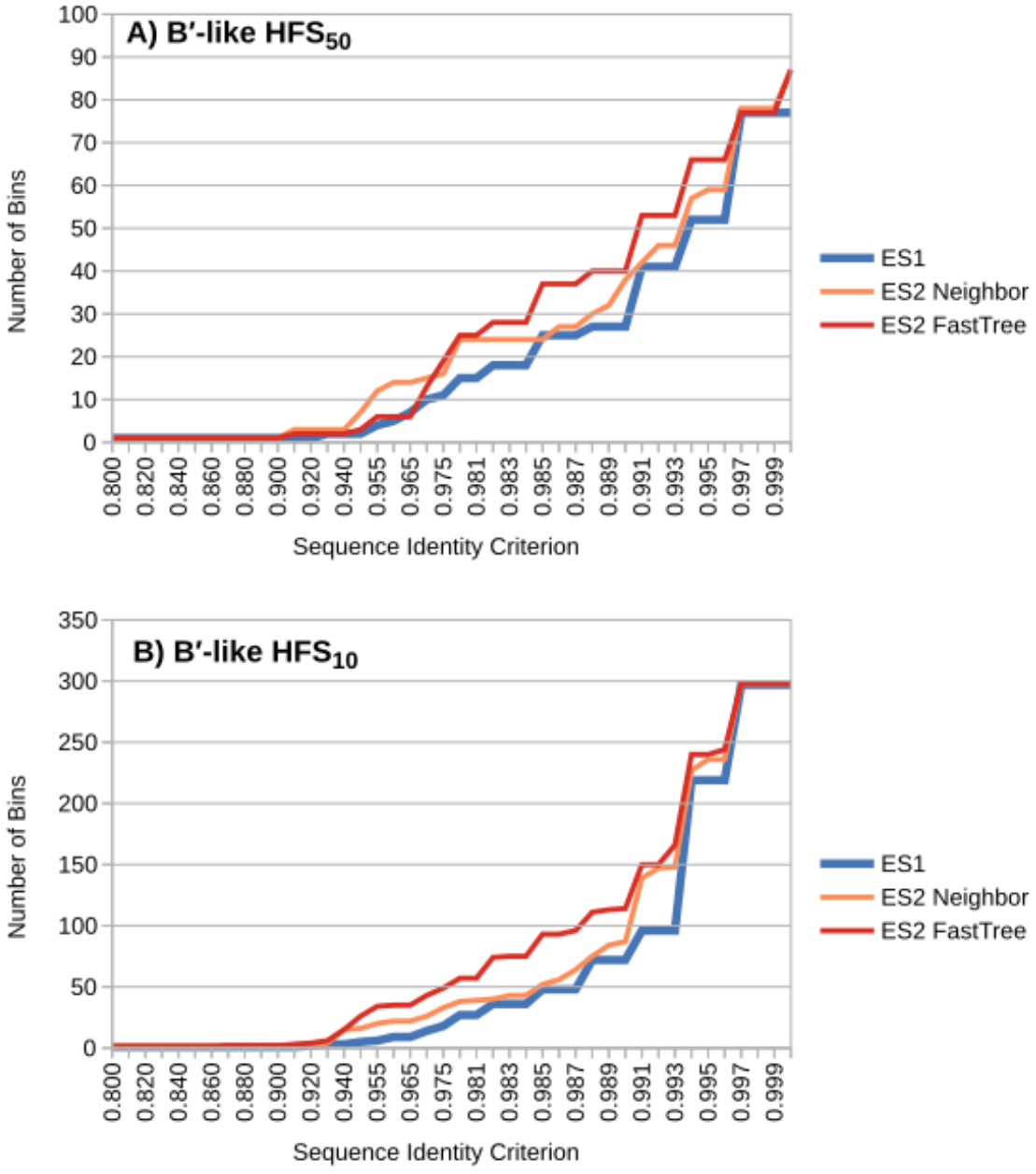


Figure A.1: Binning results using B′-like *Synechococcus psaA* high-frequency sequences (HFS) occuring at least (A) fifty times (HFS_{50}) and (B) ten times (HFS_{10}) in the environmental sequence dataset.

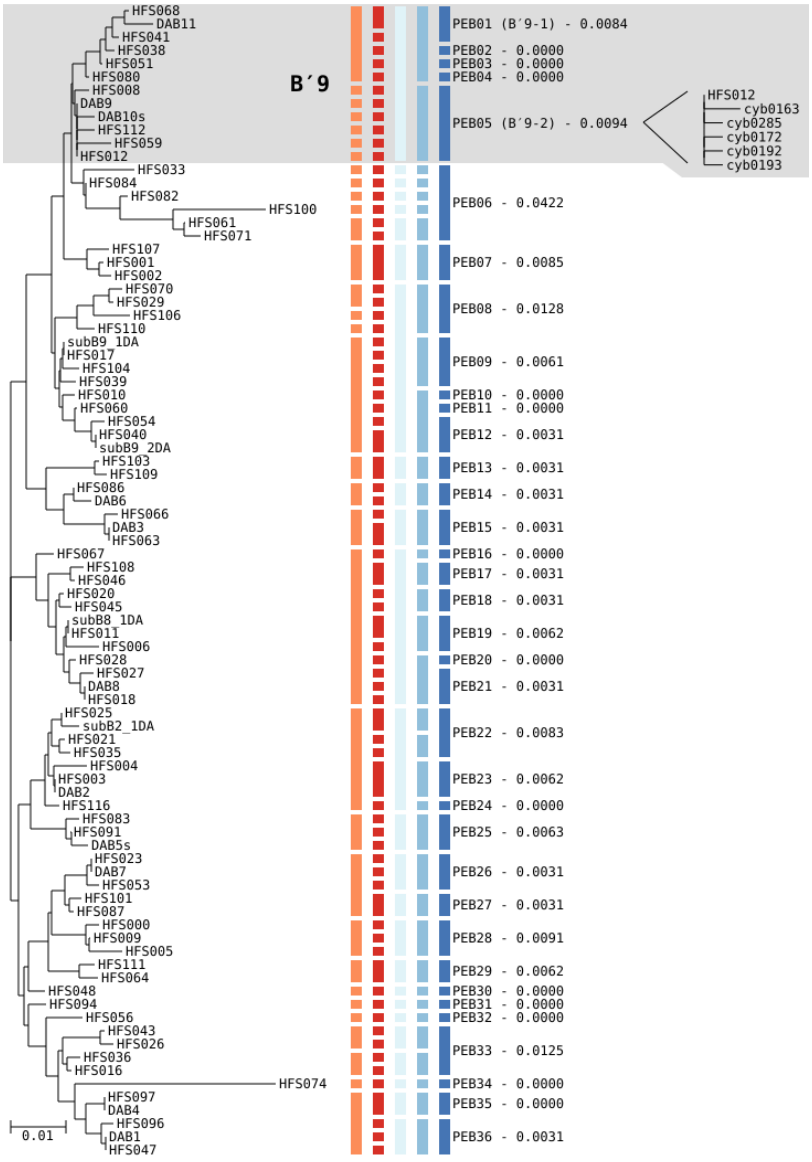


Figure A.2: Neighbor-joining phylogeny with putative ecotype (PE) demarcation of B′-like *Synechococcus* HFS_{50} environmental *psaA* segments generated using PHYLIP. Gray shading denotes predominant PEs demarcated using Ecotype Simulation 1 (ES1) that are examined in detail in the main text. Colored vertical bars indicate demarcation done by different algorithms, from left to right: orange, maximum-likelihood Poisson tree processes; red, Bayesian Poisson tree processes; light-blue, ES1 coarse-scale; medium-blue, ES1 fine-scale; and dark-blue, Ecotype Simulation 2 (ES2). PEs are labeled with the ES2 demarcation and include the maximum distance among members of the clade. ES2 demarcated PEs that contain the same dominant variant as a PE demarcated by Becraft et al. (2015) using ES1 are indicated in parentheses after the ES2-generated PE names. The portion of the HFS_{10} neighbor-joining tree corresponding to predominant B′-like PE PEB05 (B′9-2) is included on the right for comparison. The scale bar represent 0.01 nucleotide substitutions per site.

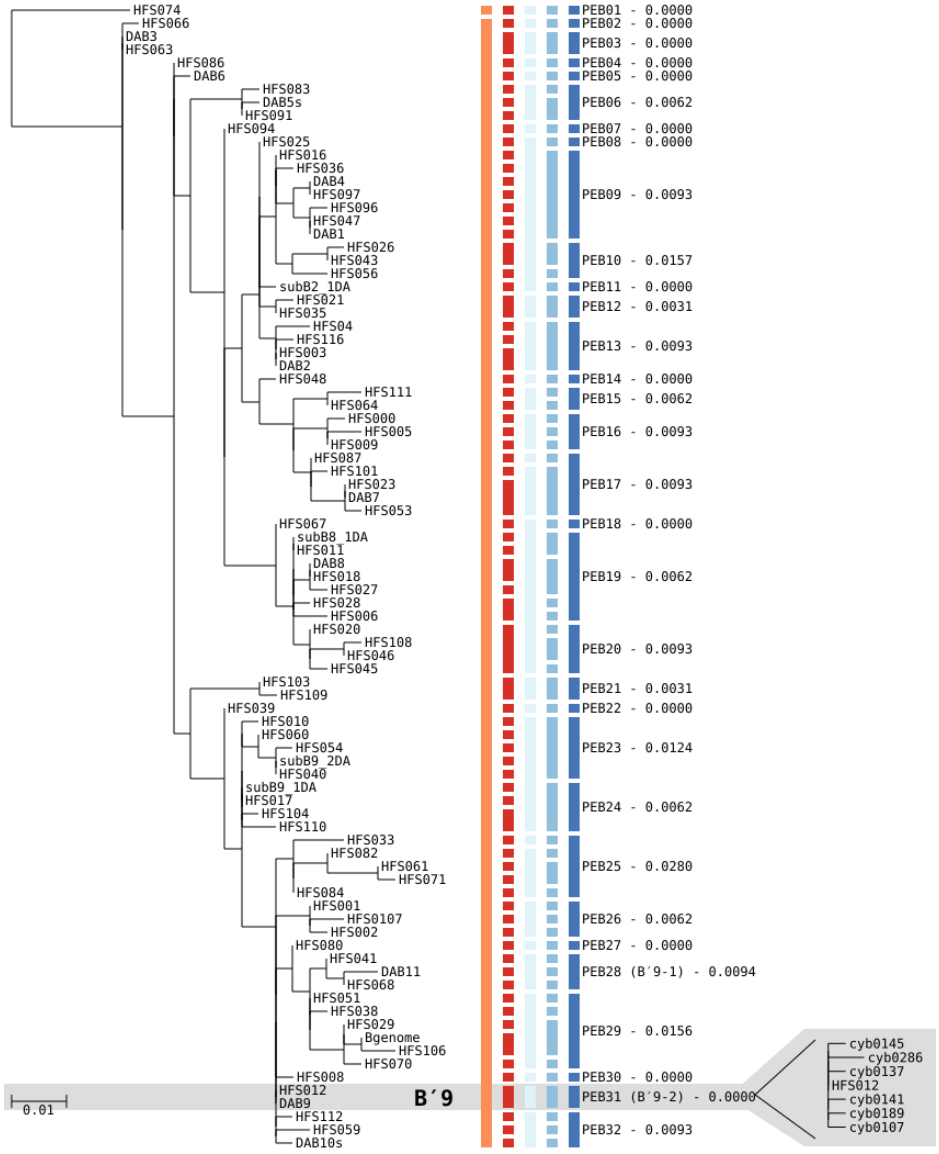


Figure A.3: Maximum-likelihood phylogeny with putative ecotype (PE) demarcation of B′-like *Synechococcus* HFS_{50} environmental *psaA* segments generated using FastTree. Gray shading denotes predominant PEs demarcated using Ecotype Simulation 1 (ES1) that are examined in detail in the main text. Colored bars indicate demarcation done by different algorithms, from left to right: orange, maximum-likelihood Poisson tree processes; red, Bayesian Poisson tree processes; light-blue, ES1 coarse-scale; medium-blue, ES1 fine-scale; and dark-blue, Ecotype Simulation 2 (ES2). PEs are labeled with the ES2 demarcation and include the maximum distance among members of the clade. ES2 demarcated PEs that contain the same dominant variant as a PE demarcated by Becraft et al. (2015) using ES1 are indicated in parentheses after the ES2-generated PE names. The portion of the HFS_{10} maximum-likelihood tree corresponding to predominant B′-like PEB31 (B′9) is included on the right for comparison. The scale bar represent 0.01 nucleotide substitutions per site.

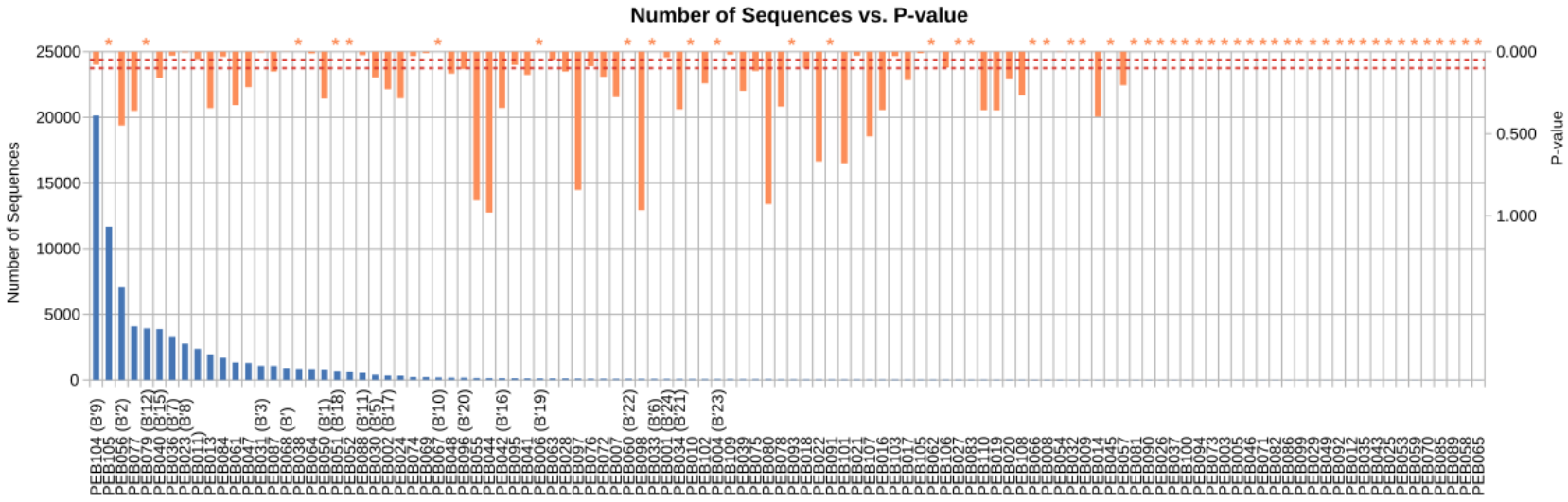


Figure A.4: *psaA* Sequence abundance of B′-like *Synechococcus* HFS_{10} environmental *psaA* segments in putative ecotypes (PEs) demarcated by Ecotype Simulation 2 from the maximum-likelihood tree (blue bars). Orange bars indicate p-values associated with the hypotheses that the members of PEs should not be distributed randomly. Orange asterisks above some columns represent p-values that could not be calculated since the PEs contained only a single member. The red dashed lines represent 0.05 and 0.10 confidence limits.

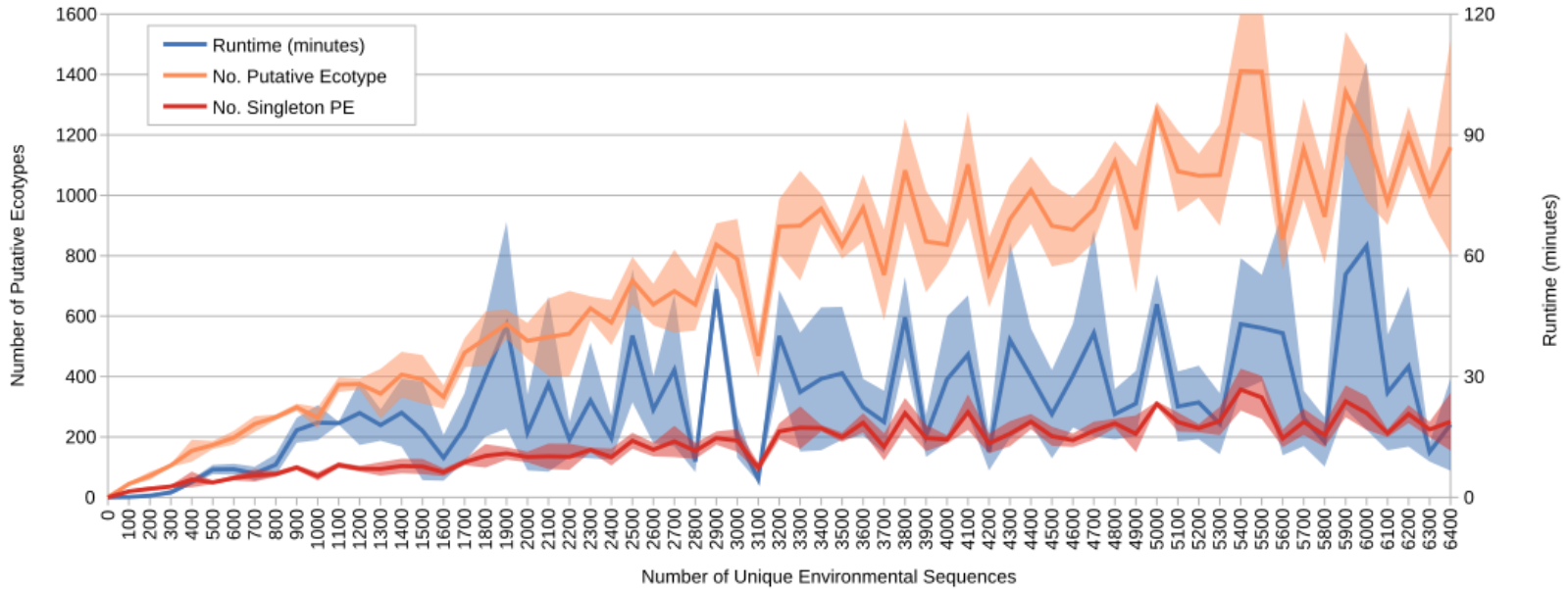


Figure A.5: Rarefaction curve produced by repeated subsampling of the 6,474 unique B′-like *Synechococcus psaA* segments present in the environmental dataset and subsequent analysis by Ecotype Simulation 2 (ES2). The darker colored lines represent the average and the shaded regions represent the standard error of three trials. The blue line and shaded region demonstrate the amount of time used to run the trial subsample using ES2 (on an Intel Core i7-6700 processor). The orange line and shaded region note the total number of PEs demarcated while the red line and shaded region note the number of PEs with only one member.

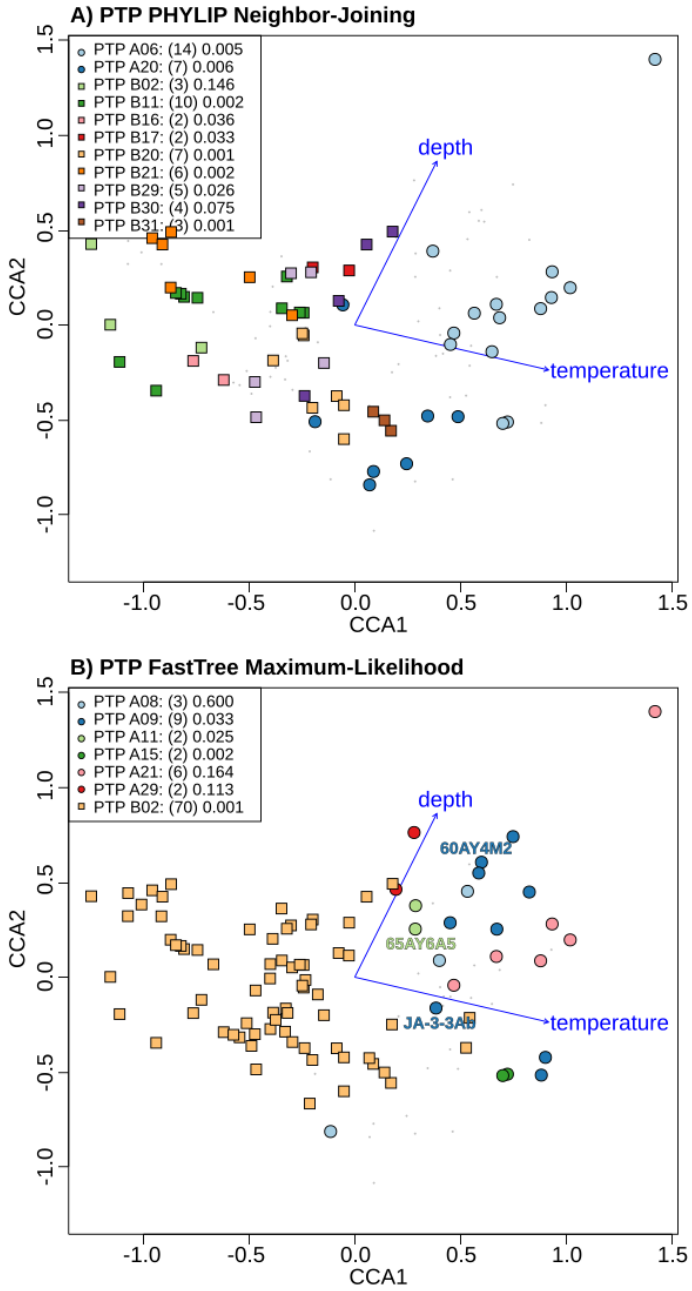


Figure A.6: Canonical correspondence analysis of putative ecotypes (PEs) demarcated by the maximum-likelihood Poisson tree processes (PTP) using the (A) neighbor-joining and (B) maximum-likelihood trees created from the A- and B′-like *Synechococcus* HFS_{50} environmental *psaA* segments. A- and B′-like PTP clades are numbered from top to bottom of their respective trees.

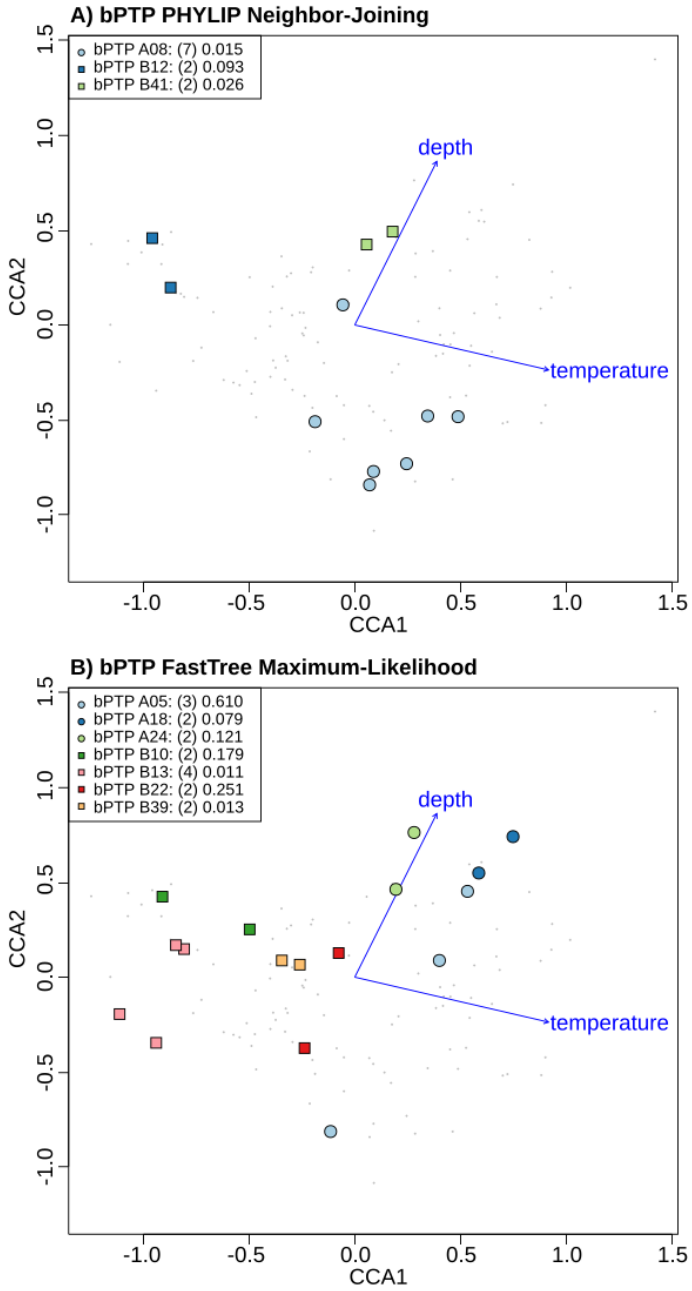


Figure A.7: Canonical correspondence analysis of putative ecotypes (PEs) demarcated by the Bayesian Poisson tree processes (bPTP) using the (A) neighbor-joining and (B) maximum-likelihood trees created from the A- and B′-like *Synechococcus* HFS_{50} environmental *psaA* segments. A- and B′-like bPTP clades are numbered from top to bottom of their respective trees.

Table A.1: Comparison of parameter estimation methods from Ecotype Simulation 1 (ES1) and Ecotype Simulation 2 (ES2), using the same matrix-based binning data in each analysis. The number of sequences, n is provided for each analysis. Optimized results from the hillclimbing algorithm are included in parenthesis next to the estimated values.

Sequence Data	ES1 (Brute-Force)	ES2 (Brute-Force)	ES2 (Lloyd's-Based)
A-like *Synechococcus* HFS$_{50}$			
$n = 65$			
omega	0.993 (0.993)	0.314 (0.330)	0.214 (0.252)
sigma	0.314 (0.324)	55.6 (56.9)	0.467 (0.479)
npop	32 (32)	28 (35)	13 (23)
A-like *Synechococcus* HFS$_{10}$			
$n = 246$			
omega	0.559 (0.838)	0.314 (0.340)	0.271 (0.454)
sigma	0.559 (0.474)	0.559 (0.538)	0.861 (0.845)
npop	39 (49)	62 (70)	27 (51)
B′-like *Synechococcus* HFS$_{50}$			
$n = 87$			
omega	0.559 (0.549)	0.314 (0.323)	0.215 (0.358)
sigma	0.993 (0.992)	17.6 (60.6)	0.405 (0.996)
npop	41 (47)	56 (56)	16 (58)
B′-like *Synechococcus* HFS$_{10}$			
$n = 298$			
omega	0.993 (0.993)	0.314 (0.316)	0.212 (0.223)
sigma	0.314 (0.316)	31.3 (0.990)	0.705 (0.713)
npop	115 (169)	126 (94)	29 (41)

APPENDIX B

CHAPTER 4 SUPPLEMENTAL INFORMATION

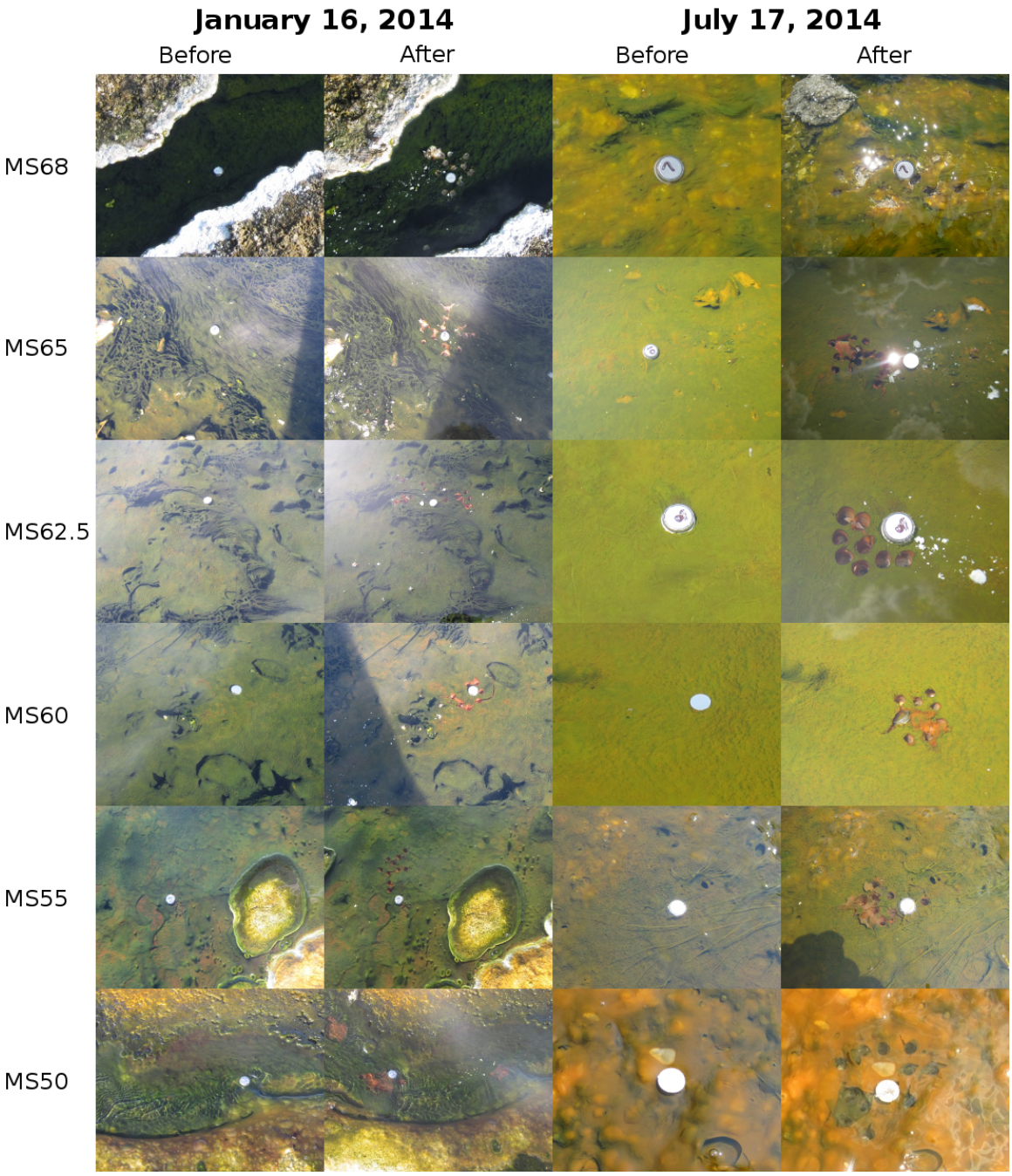


Figure B.1: Before and after photographs of the temperature-defined sampling locations in the *Synechococcus* mat growing in the major effluent channel of Mushroom Spring, Yellowstone National Park. The round metallic objects in each image are the Maxim iButton thermistors used to record the temperature of the water. The diameter of each iButton is ∼1.7 cm for scale. Note that the iButton location for the January 16, 2014 MS50 site was changed to provide better mat to harvest. No visible evidence of mat damaged during the January trip remained when samples were collected on July 17, 2014. Evidence of damage from the July coring was difficult to find during a follow-up visit on August 12, 2014.

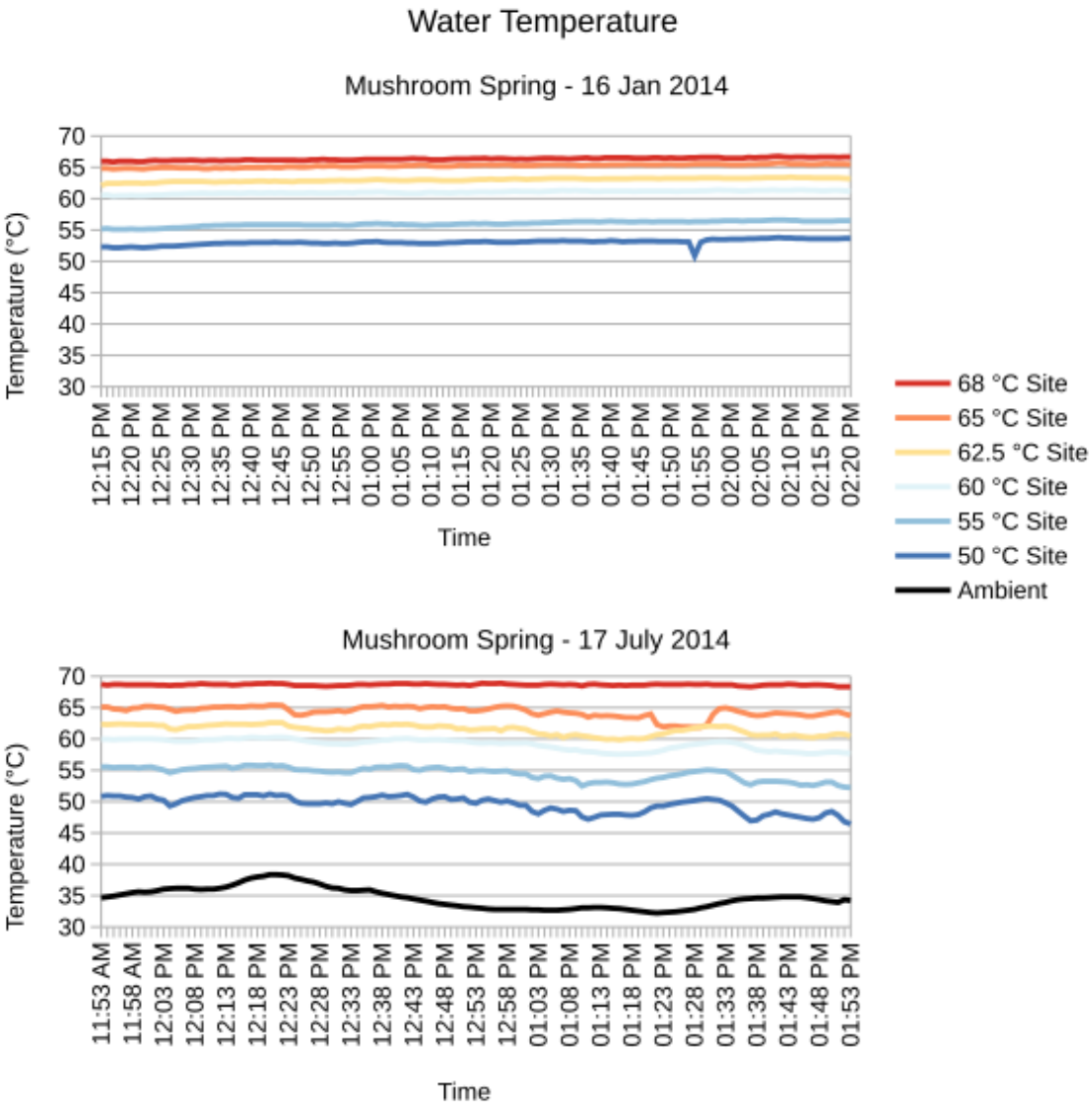


Figure B.2: Water temperature recorded with Maxim iButton thermistors placed directly on the mat surface at each site sampled on January 16, 2014 and July 17, 2014. An extra iButton went unused during the July sampling trip and subsequently recorded the ambient temperature.

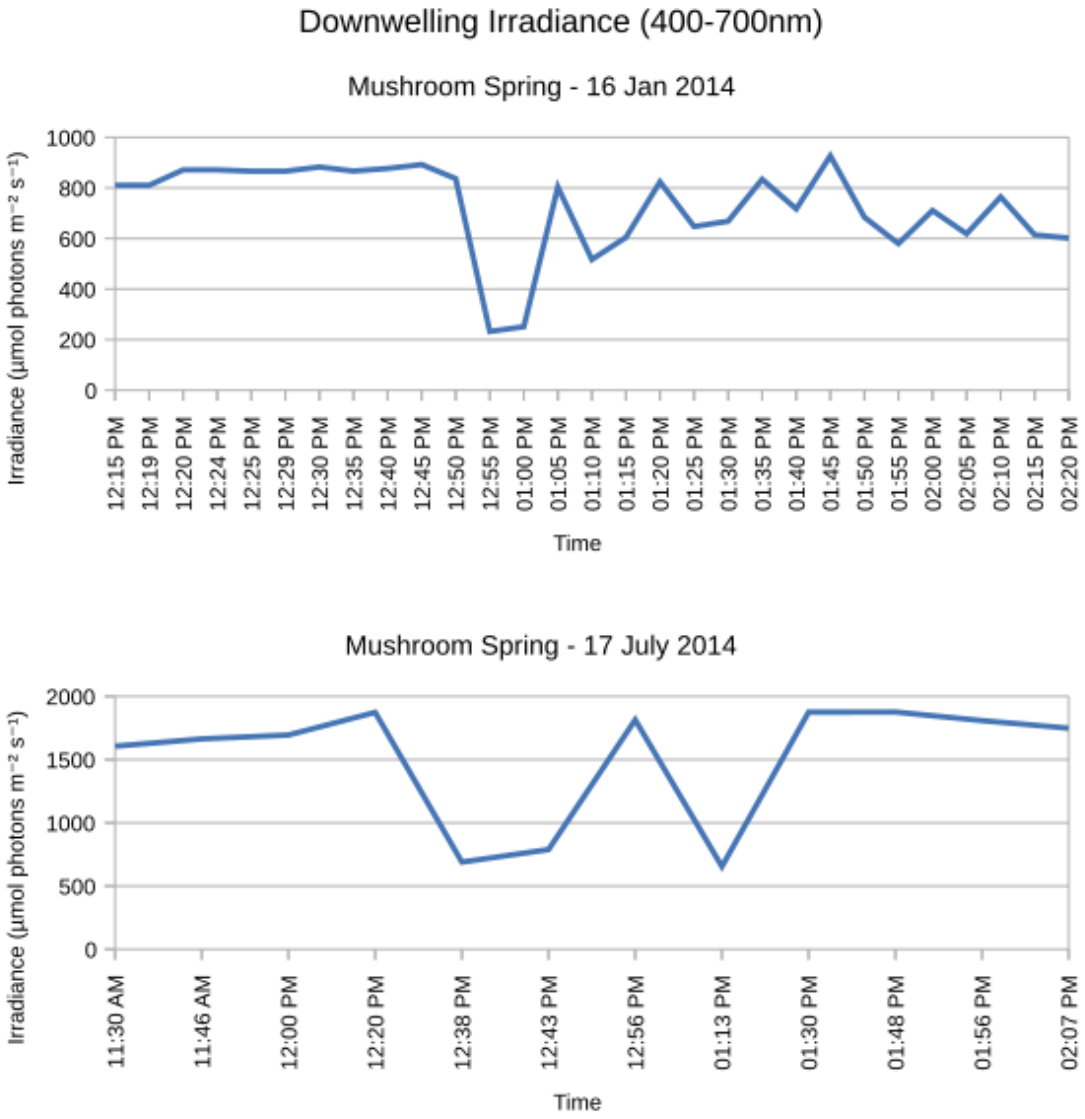


Figure B.3: Downwelling irradiance (400 nm to 700 nm) recorded with a LI-COR LI-1400 data-logger during the sampling conducted on January 16, 2014, and spot measured with a LI-COR LI-250 light meter during the sampling conducted on July 17, 2014.

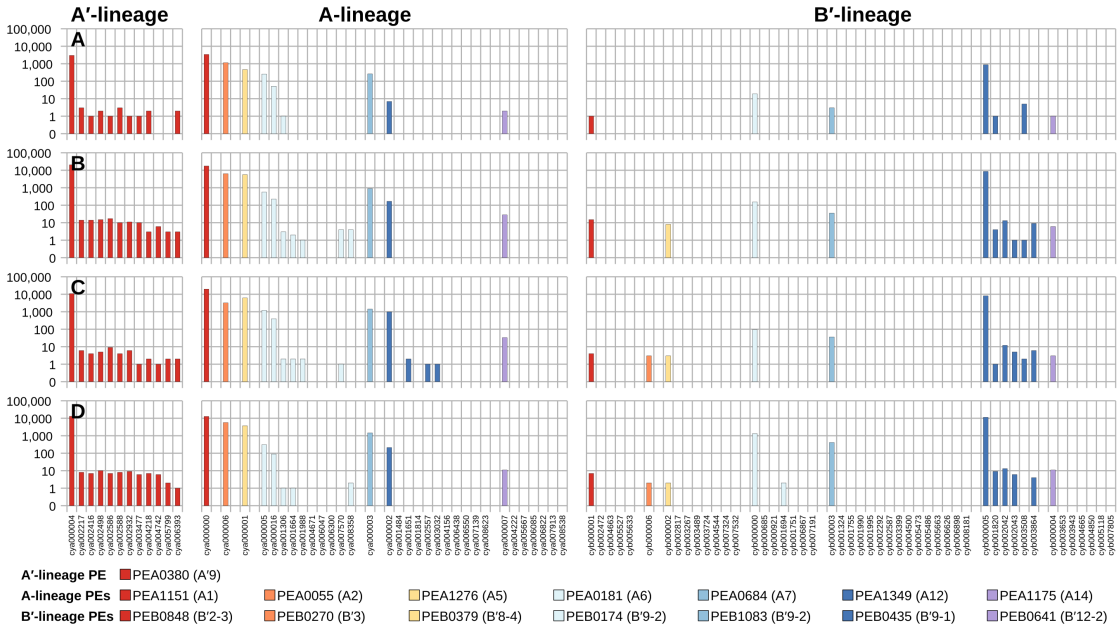


Figure B.4: Abundance of high-frequency *psaA* sequence segments that are predicted to be members of predominant A′-, A-, and B′-like putative ecotypes (PEs) from replicate mat samples collected from the 65 °C site of the Mushroom Spring effluent channel on January 16, 2014. Abundances are based on sequences obtained from an unsectioned mat sample (A) or from mat subsamples that were obtained by vertical dissection using a cryotome with abundances of all subsamples summed (B-D).

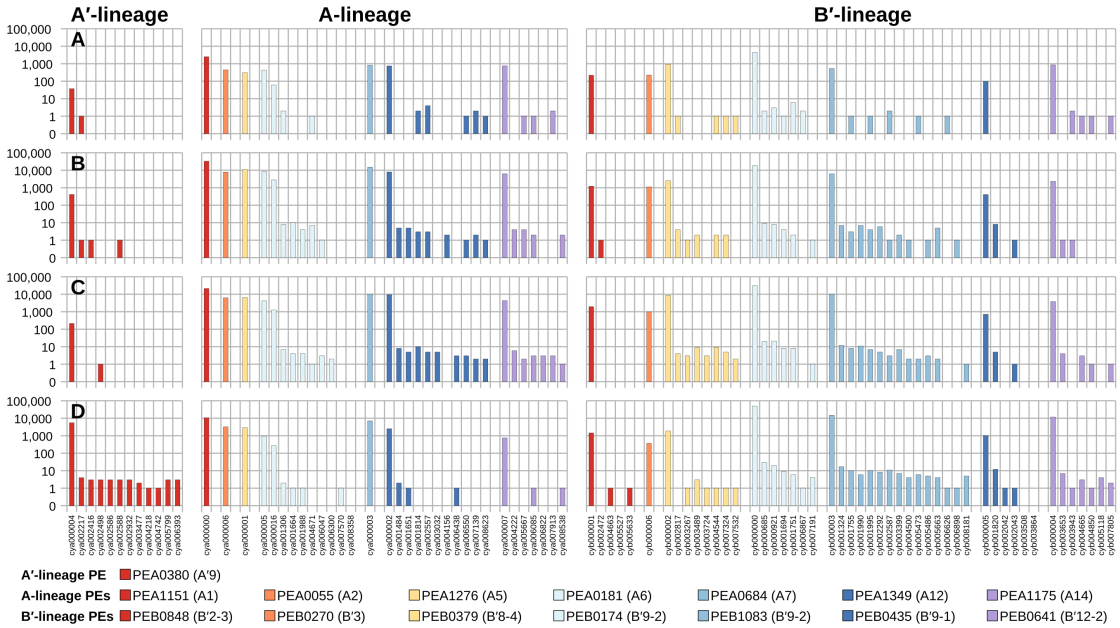


Figure B.5: Abundance of high-frequency *psaA* sequence segments that are predicted to be members of predominant A′-, A-, and B′-like putative ecotypes (PEs) from replicate mat samples collected from the 62.5 °C site of the Mushroom Spring effluent channel on January 16, 2014. Abundances are based on sequences obtained from an unsectioned mat sample (A) or from mat subsamples that were obtained by vertical dissection using a cryotome with abundances of all subsamples summed (B-D).

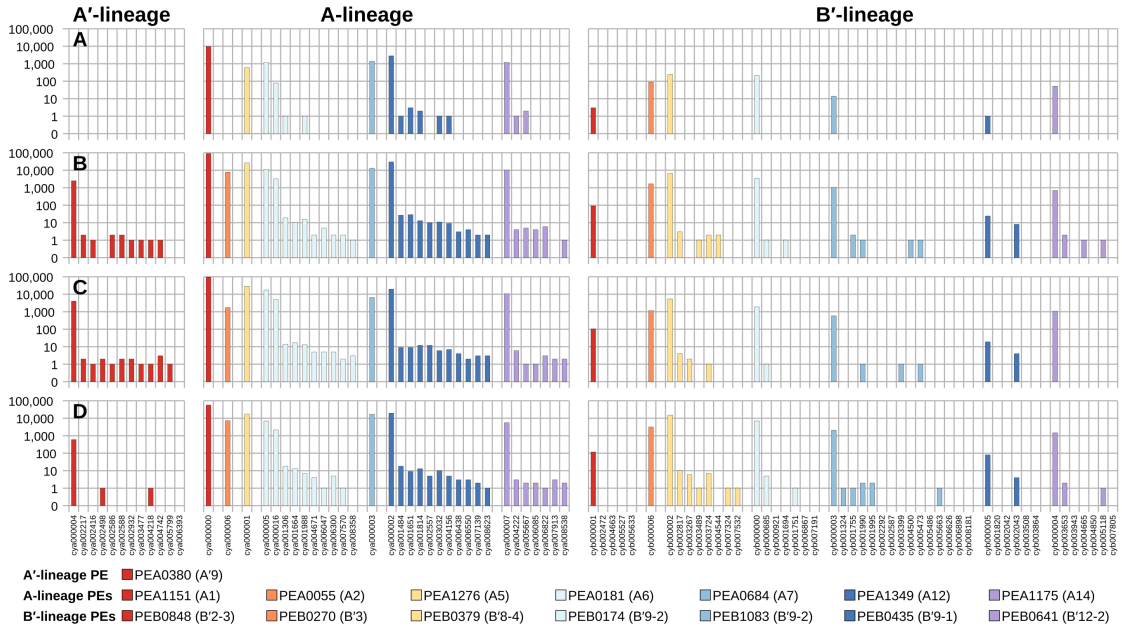


Figure B.6: Abundance of high-frequency *psaA* sequence segments that are predicted to be members of predominant A′-, A-, and B′-like putative ecotypes (PEs) from replicate mat samples collected from the 65 °C site of the Mushroom Spring effluent channel on July 17, 2014. Abundances are based on sequences obtained from an unsectioned mat sample (A) or from mat subsamples that were obtained by vertical dissection using a cryotome with abundances of all subsamples summed (B-D).

Table B.1: Summary of samples and *psaA* sequence amplicons collected from various temperature-defined sites along the effluent channel of Mushroom Spring on January 16 and July 7, 2014. High-frequency sequences (HFSs) are unique sequence types found ≥10 times across all combined samples.

Site Name	Collection Date	Average Temperature	Total Sequences	Total HFSs
MS68	2014-01-16	66.33 °C	62,233	2,681
MS65	2014-01-16	65.25 °C	56,577	3,556
MS62.5	2014-01-16	62.93 °C	42,937	4,176
MS60	2014-01-16	61.03 °C	92,356	4,856
MS55	2014-01-16	56.00 °C	35,502	2,070
MS50	2014-01-16	53.06 °C	82,184	2,149
MS68	2014-07-17	68.64 °C	30,341	1,181
MS65	2014-07-17	64.33 °C	33,893	2,390
MS62.5	2014-07-17	61.42 °C	12,082	1,096
MS60	2014-07-17	59.09 °C	27,927	1,593
MS55	2014-07-17	54.41 °C	34,294	1,868
MS50	2014-07-17	49.47 °C	21,153	830

Table B.2: Summary of additional samples and *psaA* sequence amplicons collected from the effluent channels of Mushroom and Octopus Springs over a 22 year period. High-frequency sequences (HFSs) are unique sequence types found ≥10 times across all combined samples.

Site Name	Collection Date	Average Temperature	Total Sequences	Total HFSs
MS56	1996-06-12	56.4 °C	2,434	131
MS56	1996-06-12	56.4 °C	1,978	130
MS60	1999-10-07	60 °C	2,149	157
MS56	2006-08-11	56.4 °C	3,692	110
MS60	2007-12-13	60 °C	2,809	348
MS60	2008-09-09	60 °C	2,786	145
MS60	2009-06-18	60 °C	2,054	359
MS60	2011-06-02	59.9 °C	15	6
MS63	2011-06-02	63.1 °C	156	20
MS60	2011-12-15	59.9 °C	47	15
MS63	2011-12-15	63.1 °C	32	10
OS60	1989-10-16	60 °C	2,415	156
OS55	1989-10-16	55 °C	2,090	125
OS61	1991-08-08	61 °C	3,320	401
OS60	1996-05-30	60 °C	2,210	109
OS54	1999-11-18	54 °C	2,394	375
OS64	1999-11-18	64 °C	3,225	457
OS64	2005-06-23	64 °C	1,849	150
OS56	2011-06-02	56.5 °C	177	10

Table B.3: Summary of samples and *psaA* sequence amplicons collected from the MS65 site of the effluent channel of Mushroom Spring on January 16, 2014. Mat samples were vertically dissected using a cryotome before extracting and sequencing DNA. The average temperature for each site was measured with a Maxim iButton placed directly on the mat surface. High-frequency sequences (HFSs) are unique sequence types found ≥10 times across all combined samples.

Site Name	Average Temperature	Depth From Mat Surface	Total Sequences	Total HFSs
MS65	65.25 °C	0 µm	34,957	3,022
MS65	65.25 °C	120 µm	17,261	1,890
MS65	65.25 °C	240 µm	26,465	2,843
MS65	65.25 °C	360 µm	25,454	2,854
MS65	65.25 °C	480 µm	18,924	2,319
MS65	65.25 °C	600 µm	22,636	2,583
MS65	65.25 °C	720 µm	25,979	2,668
MS65	65.25 °C	840 µm	21,397	2,344
MS65	65.25 °C	960 µm	32,262	3,054
MS65	65.25 °C	1080 µm	27,777	2,873
MS65	65.25 °C	0 µm	29,242	2,909
MS65	65.25 °C	120 µm	34,357	3,167
MS65	65.25 °C	240 µm	22,719	2,588
MS65	65.25 °C	360 µm	28,969	2,926
MS65	65.25 °C	480 µm	34,181	3,265
MS65	65.25 °C	600 µm	24,100	2,660
MS65	65.25 °C	720 µm	28,829	2,936
MS65	65.25 °C	0 µm	24,383	2,689
MS65	65.25 °C	120 µm	30,285	3,172
MS65	65.25 °C	240 µm	29,890	3,102
MS65	65.25 °C	360 µm	20,302	2,476
MS65	65.25 °C	480 µm	25,212	2,787
MS65	65.25 °C	600 µm	33,703	3,281
MS65	65.25 °C	720 µm	22,018	2,549

Table B.4: Summary of samples and *psaA* sequence amplicons collected from the MS62.5 site of the effluent channel of Mushroom Spring on January 16, 2014. Mat samples were vertically dissected using a cryotome before extracting and sequencing DNA. The average temperature for each site was measured with a Maxim iButton placed directly on the mat surface. High-frequency sequences (HFSs) are unique sequence types found ≥10 times across all combined samples.

Site Name	Average Temperature	Depth From Mat Surface	Total Sequences	Total HFSs
MS62.5	62.93 °C	0 µm	23,281	2,923
MS62.5	62.93 °C	200 µm	24,065	2,968
MS62.5	62.93 °C	320 µm	12,717	1,998
MS62.5	62.93 °C	440 µm	35,060	3,890
MS62.5	62.93 °C	560 µm	23,436	3,045
MS62.5	62.93 °C	680 µm	24,842	3,170
MS62.5	62.93 °C	800 µm	35,581	3,990
MS62.5	62.93 °C	920 µm	29,037	3,612
MS62.5	62.93 °C	1040 µm	34,212	4,110
MS62.5	62.93 °C	1160 µm	28,713	3,650
MS62.5	62.93 °C	0 µm	23,508	2,839
MS62.5	62.93 °C	120 µm	34,067	3,583
MS62.5	62.93 °C	240 µm	28,363	3,344
MS62.5	62.93 °C	360 µm	32,091	3,767
MS62.5	62.93 °C	480 µm	31,485	3,686
MS62.5	62.93 °C	600 µm	29,493	3,560
MS62.5	62.93 °C	720 µm	32,260	3,819
MS62.5	62.93 °C	840 µm	28,252	3,577
MS62.5	62.93 °C	960 µm	29,285	3,641
MS62.5	62.93 °C	0 µm	36,180	2,674
MS62.5	62.93 °C	120 µm	29,215	2,638
MS62.5	62.93 °C	240 µm	30,809	3,161
MS62.5	62.93 °C	360 µm	36,915	3,777
MS62.5	62.93 °C	480 µm	27,976	3,293
MS62.5	62.93 °C	600 µm	35,906	3,910
MS62.5	62.93 °C	720 µm	35,390	3,953
MS62.5	62.93 °C	920 µm	42,713	4,436
MS62.5	62.93 °C	1120 µm	39,040	4,113

Table B.5: Summary of samples and *psaA* sequence amplicons collected from the MS65 site of the effluent channel of Mushroom Spring on July 17, 2014. Mat samples were vertically dissected using a cryotome before extracting and sequencing DNA. The average temperature for each site was measured with a Maxim iButton placed directly on the mat surface. High-frequency sequences (HFSs) are unique sequence types found ≥10 times across all combined samples.

Site Name	Average Temperature	Depth From Mat Surface	Total Sequences	Total HFSs
MS65	64.33 °C	0 µm	28,160	2,786
MS65	64.33 °C	120 µm	39,884	3,896
MS65	64.33 °C	240 µm	46,463	4,139
MS65	64.33 °C	360 µm	34,448	3,445
MS65	64.33 °C	480 µm	39,031	3,576
MS65	64.33 °C	600 µm	29,635	3,068
MS65	64.33 °C	720 µm	40,893	3,824
MS65	64.33 °C	840 µm	45,987	3,694
MS65	64.33 °C	960 µm	28,466	2,980
MS65	64.33 °C	1080 µm	32,772	3,476
MS65	64.33 °C	1200 µm	39,289	3,764
MS65	64.33 °C	0 µm	26,678	2,486
MS65	64.33 °C	120 µm	36,401	3,566
MS65	64.33 °C	240 µm	41,469	3,785
MS65	64.33 °C	360 µm	38,269	3,541
MS65	64.33 °C	480 µm	49,153	4,075
MS65	64.33 °C	600 µm	29,981	2,908
MS65	64.33 °C	720 µm	33,934	3,259
MS65	64.33 °C	840 µm	40,633	3,446
MS65	64.33 °C	960 µm	29,426	2,699
MS65	64.33 °C	1080 µm	34,830	3,000
MS65	64.33 °C	1200 µm	30,445	2,732
MS65	64.33 °C	0 µm	36,979	3,628
MS65	64.33 °C	120 µm	41,561	3,952
MS65	64.33 °C	240 µm	32,603	3,593
MS65	64.33 °C	360 µm	34,091	3,661
MS65	64.33 °C	480 µm	14,024	2,003
MS65	64.33 °C	600 µm	27,665	3,266
MS65	64.33 °C	720 µm	100,688	6,424
MS65	64.33 °C	840 µm	21,319	2,754
MS65	64.33 °C	960 µm	31,247	3,478

Table B.6: Summary of samples and *psaA* sequence amplicons collected from the MS62.5 site of the effluent channel of Mushroom Spring on July 17, 2014. Mat samples were vertically dissected using a cryotome before extracting and sequencing DNA. The average temperature for each site was measured with a Maxim iButton placed directly on the mat surface. High-frequency sequences (HFSs) are unique sequence types found ≥10 times across all combined samples.

Site Name	Average Temperature	Depth From Mat Surface	Total Sequences	Total HFSs
MS62.5	61.42 °C	0 µm	39,094	2,853
MS62.5	61.42 °C	120 µm	26,938	2,465
MS62.5	61.42 °C	240 µm	36,806	3,008
MS62.5	61.42 °C	360 µm	41,521	3,350
MS62.5	61.42 °C	480 µm	23,988	2,407
MS62.5	61.42 °C	600 µm	34,690	3,000
MS62.5	61.42 °C	720 µm	34,513	3,074
MS62.5	61.42 °C	840 µm	43,389	3,533
MS62.5	61.42 °C	960 µm	49,525	3,643
MS62.5	61.42 °C	0 µm	34,170	3,062
MS62.5	61.42 °C	120 µm	40,425	3,545
MS62.5	61.42 °C	240 µm	38,442	3,503
MS62.5	61.42 °C	360 µm	13,736	1,516
MS62.5	61.42 °C	480 µm	13,630	1,486
MS62.5	61.42 °C	600 µm	10,348	1,223
MS62.5	61.42 °C	720 µm	15,757	1,645
MS62.5	61.42 °C	840 µm	7,278	987
MS62.5	61.42 °C	960 µm	11,969	1,462
MS62.5	61.42 °C	1080 µm	20,218	1,981
MS62.5	61.42 °C	1200 µm	8,790	1,017
MS62.5	61.42 °C	0 µm	10,666	1,151
MS62.5	61.42 °C	120 µm	15,485	1,540
MS62.5	61.42 °C	240 µm	8,708	1,040
MS62.5	61.42 °C	360 µm	8,357	926
MS62.5	61.42 °C	480 µm	11,024	1,207
MS62.5	61.42 °C	600 µm	9,464	1,083
MS62.5	61.42 °C	720 µm	9,813	1,077
MS62.5	61.42 °C	840 µm	14,612	1,534

APPENDIX C

CHAPTER 5 SUPPLEMENTAL INFORMATION

Table C.1: Summary of samples and sequences collected from various temperature-defined sites along the effluent channel of Mushroom Spring on July 7, 2014.

Site	Avg. Temp.	*Synechococcus* Seq.	*Synechococcus* HFSs	*Chloroflexus* Seq.	*Chloroflexus* HFSs	'*Ca.* Roseilinea' Seq.	'*Ca.* Roseilinea' HFSs	*Roseiflexus* Seq.	*Roseiflexus* HFSs	*Chloracidobacter* Seq.	*Chloracidobacter* HFSs	'*Ca.* Thermochlorobacter' Seq.	'*Ca.* Thermochlorobacter' HFSs
MS68	68.64 °C	30,341	431	24,731	71	31,359	33	29,573	401	656	97	48,470	522
MS65	64.33 °C	33,893	1,338	24,121	549	24,188	501	38,180	506	22,566	525	37,548	530
MS62.5	61.42 °C	12,082	673	23,256	506	34,106	516	30,368	677	22,122	524	29,291	521
MS60	59.09 °C	27,927	830	28,045	483	14,732	446	26,803	679	11,359	408	17,495	122
MS55	54.41 °C	34,294	936	27,917	587	29,880	283	32,629	670	26,043	60	33,903	149
MS50	49.47 °C	21,153	466	28,173	587	27,112	254	13,153	439	21,728	49	11,042	97

Table C.2: Summary of samples and sequences collected from the MS62.5 site (average temperature 61.42 °C) of the effluent channel of Mushroom Spring on July 7, 2014. *Synechococcus* DNA was not sequenced from samples collected from the bottom of the mat core.

Depth From	*Synechococcus*		*Chloroflexus*		'*Ca.* Roseilinea'		*Roseiflexus*		*Chloracidobacter*		'*Ca.* Thermochlorobacter'	
Mat Surface	Seq.	HFSs	Seq.	HFSs	Seq.	HFSs	Seq.	HFSs	Seq.	HFSs	Seq.	HFSs
0 µm	66,032	2,503	38,637	485	11	5	19,837	633	16,251	503	16,414	513
240 µm	78,327	2,866	14,366	424	11,178	437	20,697	640	18,074	507	28,288	536
480 µm	58,678	2,602	13,815	498	34,472	516	23,505	668	23,279	520	21,964	527
720 µm	77,902	3,080	21,425	541	47,483	518	31,453	690	21,669	523	19,941	491
960 µm	49,525	1,593	22,260	545	22,671	489	26,488	679	37,437	525	32,081	504
1280 µm	-	-	42,112	567	35,535	515	24,777	680	27,168	520	12,046	422
1680 µm	-	-	29,979	544	30,024	514	28,188	692	24,096	516	29,005	494
2080 µm	-	-	22,815	533	38,632	513	22,046	645	26,041	523	6,482	366
0 µm	74,595	3,224	15,401	504	22,448	507	20,726	668	16,933	510	13,994	509
240 µm	52,178	2,689	16,358	570	13,836	473	19,466	665	19,564	516	30,841	535
480 µm	23,978	1,659	18,014	588	39,496	525	24,942	674	24,410	531	19,385	528
720 µm	23,035	1,628	21,553	581	37,125	523	17,272	642	19,846	511	18,850	512
960 µm	32,187	1,921	31,244	588	28,992	520	24,975	682	31,816	531	33,509	531
1200 µm	8,790	576	22,352	567	32,413	521	22,931	674	25,366	520	11,689	422
1520 µm	-	-	31,655	582	37,898	525	28,638	684	18,094	512	29,239	493
1920 µm	-	-	27,749	554	36,770	528	23,544	688	23,823	523	15,503	453
0 µm	26,151	1,628	15,796	507	29,145	512	27,956	693	16,279	501	40,694	544
240 µm	17,065	1,223	15,684	515	17,509	489	21,327	673	17,689	503	44,392	550
480 µm	20,488	1,350	13,179	542	40,661	527	3	0	21,569	520	24,436	533
720 µm	24,425	1,511	22,304	588	39,193	531	30,559	682	22,084	523	28,124	495
960 µm	-	-	18,648	590	34,681	523	30,155	684	31,993	531	53,599	512
1360 µm	-	-	21,040	575	30,230	525	29,729	686	23,619	516	13,826	455
1760 µm	-	-	15,785	564	42,054	524	24,719	681	19,743	517	37,666	531

Table C.3: Summary of primers used to study phototrophic taxa of Mushroom Spring

Taxon Sampled	Loci Sampled	Length Sampled[a]	Diversity Sampled[b]	Forward Primer	Reverse Primer
Synechococcus[c]	*psaA*	282	0.65	CTGAGCGGCATGTACTACCA	CAGGCCACCCTTGAAGGTG
Chloroflexus	*pufL*	273	0.95	ACGAGGGTTTCGTTCTCGGCATTA	AGCCCGACAACAGGATGCACAAAT
'*Ca.* Roseilinea'	*pufL*	316	0.71	GGAGGTAGTCTATGTTGTCGTTCG	ATGTCAATCTGGCGCATCATC
Roseiflexus	*pufLM*	293	0.78	AGCATCCGCAGAAGTCATTCCGTT	GCAGCAACCAGCCAATGAAAGCAA
Chloracidobacter	*pscA*	280	0.73	TTCGCCGTTGATGGTGTTCATC	TGCCAGATCAAGTGGTTTACGC
'*Ca.* Thermochlorobacter'	*pscA*	282	0.83	TGTGGTTTGCTATCCTGTGGTC	TTGACATCGCGTCCAAGTAAGG

[a] Nucleotides

[b] Nucleotide Identity

[c] Originally described by Becraft et al. (2011)