# POPULATION GENETICS OF *SYNEHOCOCCUS* SPECIES INHABITING THE MUSHROOM SPRING MICROBIAL MAT, YELLOWSTONE NATIONAL PARK

by

Melanie Crystal Melendrez

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

of

Doctor of Philosophy

in

Ecology and Environmental Science

MONTANA STATE UNIVERSITY
Bozeman, Montana

March 2010

APPROVAL

of a dissertation submitted by

Melanie Crystal Melendrez

This dissertation has been read by each member of the dissertation committee and has been found to be satisfactory regarding content, English usage, format, citation, bibliographic style, and consistency, and is ready for submission to the Division of Graduate Education.

Dr. David M. Ward

Approved for the Department Land Resources and Environmental Science

Dr. Tracy Sterling

Approved for the Division of Graduate Education

Dr. Carl A. Fox

# STATEMENT OF PERMISSION TO USE

In presenting this dissertation in partial fulfillment of the requirements for a doctoral degree at Montana State University, I agree that the Library shall make it available to borrowers under rules of the Library. I further agree that copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for extensive copying or reproduction of this dissertation should be referred to ProQuest Information and Learning, 300 North Zeeb Road, Ann Arbor, Michigan 48106, to whom I have granted "the exclusive right to reproduce and distribute my dissertation in and from microform along with the non-exclusive right to reproduce and distribute my abstract in any format in whole or in part."

Melanie Crystal Melendrez

March 2010

# DEDICATION

Love is hard to believe, ask any lover.
Life is hard to believe, ask any scientist.
God is hard to believe, ask any believer.
What is your problem with hard to believe?
~ Life of Pi

For those who believed in me.

# ACKNOWLEDGEMENTS

This project was completed with funding from multiple sources; the National Science Foundation Frontiers in Integrative Biology Research Program (EF-0328698) and supplemental support from the NASA Exobiology program (NAG5-8807 and NX09AM87G) and the DOE Pacific Northwest National Laboratory (contract pending). In addition I appreciate the assistance from National Park Service personnel at Yellowstone National Park. I would like to thank the following people for their support during the completion of this project: The Ward lab past and present; E. Becraft, M. Bateson and J. Allewalt; C. Klatt and J. Wood for their assistance and generous sharing of perl scripts and Linux knowledge which made many analyses in this project possible; the administrative staff of LRES – M. Paceley, R. Adams, L. McDonald, M. Schimpf, A. Murphy and D. Barkan; N. Baldridge for organizing video conferencing for the defense; the graduate students and members of the entomology department for their support and nights at Columbos – K. Marske, K. Puliafico, J. Fultz, I. Foley and Dr. R. Hurley. Special thanks to my advisor Dr. Dave Ward for his 30+ years of knowledge, support and enthusiasm for this research; my graduate committee Drs. T. McDermott, M. Franklin, F. Cohan, A. Richman, J. Voyich-Kane, J. Downey and W. Jones for their suggestions and support; and Drs. M. Ivie, W. Hanage and T. Papke for continuous mentorship. A final personal thanks to my family in Hawaii and friends; E. Ivie, T. Vallard, S. Iverson, R. Johnston, J. Johnston, G. Fricke, M. Kumar, D. Coravelli, P. Flair, T. Olson, C. Lease, H. Camper, M. Garre and so many others for believing in me when it was a struggle for me to believe in myself.

TABLE OF CONTENTS

TABLE OF CONTENTS – CONTINUED

TABLE OF CONTENTS – CONTINUED

TABLE OF CONTENTS – CONTINUED

# LIST OF TABLES

LIST OF TABLES—CONTINUED

Table Page

LIST OF TABLES—CONTINUED

Table Page

# LIST OF FIGURES

LIST OF FIGURES—CONTINUED

Figure Page

LIST OF FIGURES—CONTINUED

Figure Page

LIST OF FIGURES—CONTINUED

LIST OF FIGURES—CONTINUED

Figure Page

LIST OF FIGURES—CONTINUED

Figure Page

LIST OF FIGURES—CONTINUED

# MOST COMMONLY USED TERMS AND ACRONYMS

| | |
|---|---|
| BAC: | Bacterial artificial chromosome |
| BLAST: | Algorithm for finding sequences that match a query sequence in the National Biotechnology Information nucleotide collection. |
| DGGE: | Denaturing gradient gel electrophoresis |
| DLV: | Double locus variant |
| dN/dS: | Measure of selection pressure; <1 (purifying), =1 (neutral), >1 (adaptive) |
| DV: | Dominant variant |
| eBURST: | An algorithm which divides an MLST data set of any size into groups of related isolates and clonal complexes and predicts founder(s). |
| Ecotype: | A population of microorganisms that occupy a distinct ecological niche, equivalent to ecological species. |
| EED: | Estimated evolutionary divergence |
| EF: | Ecotype formation |
| ES: | Ecotype Simulation; Simulation of the evolutionary history of organisms sampled from nature according to the Periodic Selection Model under different values of PS, EF and n. for prediction of ecotypes. |
| HMW: | High molecular weight |
| ITS: | Internal transcribed spacer region separating the 16S rRNA and 23S rRNA loci |
| Jointly-recruited: | Sequence data for both ends of the clone was obtained and had a best BLAST match to the same reference genome. |
| M60: | Samples obtained from the 60°C site at Mushroom Spring. |
| M65: | Samples obtained from the 65°C site at Mushroom Spring. |
| MDA: | Multiple displacement amplification |
| MLSA: | Multi-locus sequence analysis |
| MLST: | Multi-locus sequence typing |
| n: | Number of putative ecotypes |
| NCBI: | National Center for Biotechnology Information |
| OTU: | Operational taxonomic unit |
| PE: | Putative ecotype; PEA- ecotype of *Synechococcus* sp. A; PEB´- ecotype of *Synechococcus* sp. B´ |
| PS: | Periodic selection |
| r/m: | Recombination to mutation ratio |
| RDP3: | Software package for detecting recombination signals within genes. |
| sDV: | Subdominant variant |
| SLA: | Single-locus analysis |
| SLV: | Single locus variant |
| SNP: | Single nucleotide polymorphism |
| ST: | Sequence type |

# ABSTRACT

The species concept in microbiology is under considerable debate. Some scientists believe that traditional approaches are adequate, while others search for more natural concepts. The ***ecotype*** concept (ecological species concept) was evaluated in this work. Two temperature sites of a well-studied microbial mat system in Yellowstone National Park were investigated. Previous molecular analyses with 16S rRNA and the adjacent internal transcribed spacer (ITS) suggested the dominance of two putative ecotypes (PEs) of cyanobacteria in these sites, *Synechococcus* genotypes A and B´. Higher resolution molecular approaches were developed to address the hypotheses that (i) there are more *Synechococcus* PEs than those discerned by 16S rRNA/ITS sequence variation, (ii) these PEs exhibit distinct ecological distribution patterns and (iii) recombination has been less important than mutation in shaping the evolution of these *Synechococcus* populations. Analysis of single protein-encoding loci revealed more sample-specific PEs than previously detected, but didn't account for recombination. Bacterial artificial chromosome (BAC) libraries were constructed to sample multiple loci near 16S rRNA genes for multi-locus sequence analyses (MLSA). Analysis of BAC clone end sequences revealed that 16S rRNA regions of the genomes of *Synechococcus* A- and B´-like populations have undergone rearrangement. Multiple BAC loci were analyzed using two population genetics algorithms; Evolutionary Simulation (ES) and eBURST. ES of concatenated MLSA sequences, but not eBURST analysis, suggested a much greater number of PEs than were detected by 16S rRNA and ITS and provided stronger evidence of sample-specificity. Recombination, suggested by phylogenetic incongruency among loci, multiple recombination tests and polymorphism patterns, appears to have been more frequent than mutation, but not to have erased ecotype structure. Many PEs predicted by ES contained a dominant variant surrounded by rare variants. eBURST predicted some clonal complexes with the same dominant variant, but different rare variants. ES appears to miss phylogenetically distant variants that differ at one locus, whereas eBURST appears to miss phylogenetically similar variants that differ at >1 locus. True ecotype populations in nature may contain both types of variants, but this must be evaluated by examining the distribution of all variants relative to environmental gradients.

# CHAPTER 1

# INTRODUCTION

The following work is the product of research into the concept of species based on analysis of the population genetics of cyanobacterial species of the genus *Synechococcus* that inhabit a well-studied hot spring microbial mat community in Yellowstone National Park, Wyoming. In order to come to an understanding of microbial community composition, structure and function it is first necessary to understand and define what the 'fundamental units' (or species-like clusters) of that community are. In this chapter I will examine the species concept and problems associated with the definition of species in microbiology. I will also introduce the hot spring microbial mat system and *Synechococcus* A-like and B´-like populations, which will be the focus of this dissertation. I will discuss the models that have been used to predict species-like clusters based on sequence data that have accumulated historically, and the particular models that were used in this work. Finally, I will introduce the hypotheses to be tested and provide a brief overview of the chapters that follow. This work has increased our understanding of 'species' composition and structure in *Synechococcus* populations and has led to testable hypotheses about species in this microbial mat system that others may conduct in the future.

## The Species Problem: Who's There

The concept of species is still a hotly debated topic in science today. In keeping with human nature (i.e., the need to categorize life into an organized, easily recognizable,

hierarchical structure) several species concepts and definitions have emerged. Mayr (1942) introduced the Biological Species Concept, which states, "A species consists of populations of organisms that can reproduce with one another and that are reproductively isolated from other such populations." The concept was later expanded to further define reproductive isolation to include the ability to produce fertile offspring and specific mate recognition (de Queiroz, 2007). However the Biological Species Concept is difficult to apply to microorganisms that are inherently asexual, though promiscuous in their transfer of DNA. Mayden (1997) highlighted the Evolutionary Species Concept, as introduced by Simpson (1951), which stated that species have to have a "unique evolutionary role, tendencies, and/or historical fate." In a discussion of Mayr and modern concepts of species, de Queiroz suggests that what most can agree on is that species can be considered 'evolving population lineages' (de Queiroz, 2006). However, there are disagreements over the criteria or cutoffs used to demarcate species. Darwin stated in *The Origin of Species*: "…there is no possible test but individual opinion to determine…which shall be considered as species and which as varieties." It is important to realize that this 'individual opinion' should be shaped based on a natural view of organisms (i.e., in their environment) rather than on a view derived from organisms that have been removed from their natural surroundings. The Ecological Species Concept introduced by Van Valen (1976) stated that individual organisms comprising a species need to share the same environmental niche or "adaptive zone." This addition of ecology places the conceptualization of species-like populations or organisms into a natural context.

In microbiology the species problem is further compounded by the vast diversity within the microbial communities in all habitats studied. Traditionally, species of bacteria were demarcated based on morphology and metabolism, phenotypic properties that could be assessed via growth of cultures in laboratories, microscopic examination and laboratory assays. With the advent of molecular techniques, studies of the genetic variation found among strains of a species led to the idea that 70% DNA/DNA hybridization was a "gold standard" in the demarcation of species lineages (Wayne et al., 1987). It was then determined that >70% DNA/DNA hybridization roughly correlated with >97% 16S rRNA homology (Goodfellow et al., 1997; Stackenbrandt and Goebel, 1994). There are two important problems with these approaches. First, we now know that what can be easily cultured in the laboratory may not represent what is actually predominant in the environment (Ward et al., 1990; Ferris and Ward, 1997; Ferris et al., 1996a/b). In the words of Giovannoni et al. (2005) "[Cultured] organisms are rarely observed in 16S rRNA gene diversity surveys, but they frequently dominant experiments in which [conclusions are drawn about] natural microbial communities". Cultivation-independent techniques suggest that we only see a fraction of the millions, perhaps even billions of species of bacteria in any given system (Dykhuizen, 1998; Gans et al., 2005; Venter et al., 2004). Second, although molecular techniques have allowed for further understanding of the relationships among the organisms we study, these molecular cutoffs "lump diverse populations that exhibit distinct species-like traits, leading to an underestimation of functionally important biodiversity" (Ward et al., 2008; Cohan et al., 2002). A more recent cutoff of ~94% average nucleotide identity of genomes has been

introduced and this can be added to the quest for a molecular definition of species (Venter et al., 2004; Konstantinidis and Tiedje, 2005). It is important to note two things in summary: (i) molecular cutoffs have been calibrated based on the assumption that traditionally named species of bacteria are the 'real' species and (ii) studying the population genetics of isolates does not equate to studying the population genetics of organisms in nature using cultivation-independent techniques.

The debate about species that started in the macro-biological world has now made its way to the microbial world. Ward et al. (2008) discuss three principal points of view in microbiology pertaining to species. The first view is that the current classification of microbes using phenotype and phylogeny is adequate for the demarcation of species. In other words, some microbiologists adhere to the traditional approach of demarcating microbial species (Roselló-Mora and Amann, 2001). The second view suggests that the demarcation of species needs to be examined in light of natural distributions such as demonstrated with *Synechococcus* A/B lineage populations and temperature distributions (Ward et al., 2006), *Prochlorococcus* spp. and depth in the oceanic water column (Rocap et al., 2002) or *Bacillus* spp. and niche-defining variations in solar exposure (Koeppel et al., 2008; Cohan, 2006). The third view contends that horizontal gene transfer and recombination are so rampant as to completely erode species, making it impossible to demarcate meaningful units of diversity that would correspond to evolving population lineages of microbes (Doolittle and Papke, 2006).

Mayr (1982) stated, "No ecosystem can be fully understood until it has been dissected into its fundamental parts. The fundamental units need to be identified to

determine their mutual interactions and to understand community function, assembly, and dynamics." In other words, "a species concept is central to achieving a predictive understanding of the composition, structure and function of microbial communities, the population biology of disease outbreaks and the emergence of new diseases" (Ward, 2006). The challenges in identifying the 'fundamental units' of microbial communities are the inherent promiscuity of genetic transfer discussed above and knowing the level of resolution that is needed to demarcate a microbial species. This dissertation will focus on the development and application of empirical techniques to more accurately detect species and on evaluations of analytical models used to predict the ecologically meaningful, 'fundamental units' of diversity in an effort to reevaluate current thinking about microbial species concepts.

## Molecular Evidence of *Synechococcus* Ecological Species in Hot Spring Microbial Mats

In an effort to understand the 'fundamental units' of organization and function within a natural microbial community, Dr. David M. Ward and colleagues have been studying cyanobacterial mat communities in natural hot spring systems in Yellowstone National Park for over 30 years. Two sites have been intensively studied, Octopus Spring (Figure 1.1) and Mushroom Spring (Figure 1.2). Both are alkaline siliceous hot springs with a cyanobacterial mats, which grow below ~72°C. From ~50°C to 72°C (Ward et al., 1998; Ferris et al., 1996a) both mats are dominated by 'sausage-shaped' cyanobacteria of the named genus *Synechococcus*, and filamentous anoxygenic phototrophic bacteria, suggested to be *Roseiflexus*-like or *Chloroflexus*-like (Nübel et al., 2002) (Figure 1.1C).

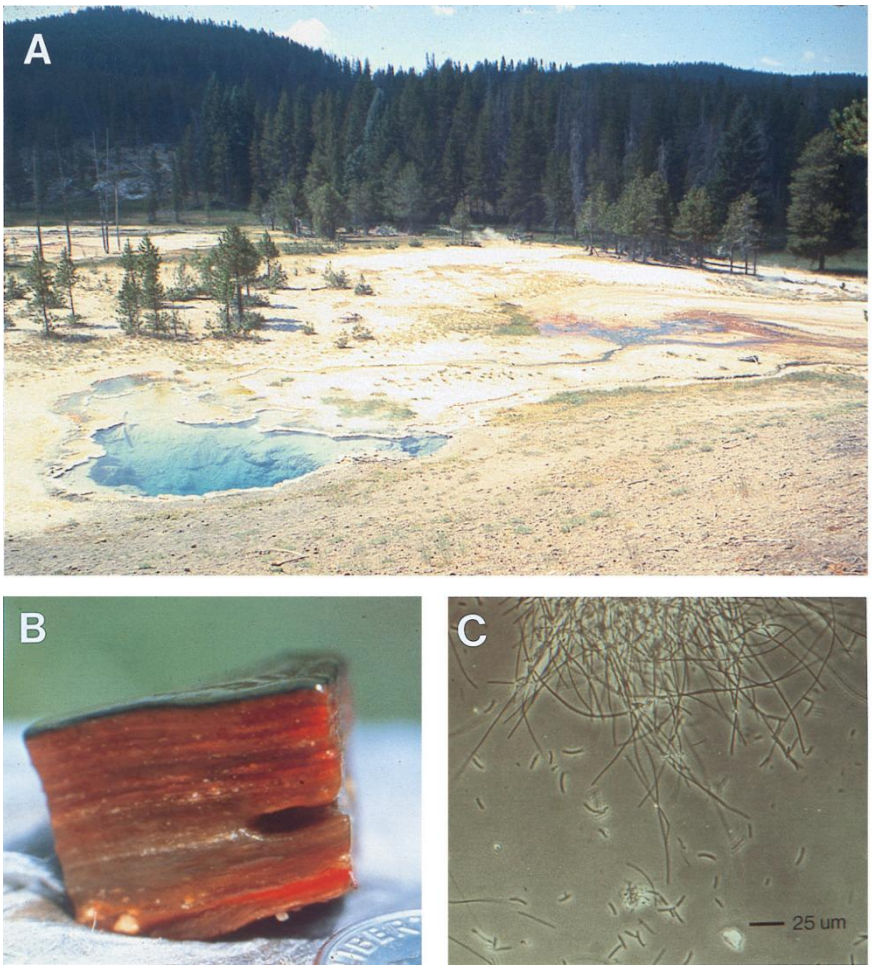



Figure 1.1. Laminated cyanobacterial mat communities of Octopus Spring, Lower Geyser Basin, Yellowstone National Park viewed at different scales. (A) Landscape image showing green-orange mat in the effluent channel. (B) Cross section of a cyanobacterial mat sample from ~60°C. (C) Phase-contrast microscopy image of homogenized 1-mm thick upper green layer of mat showing the predominant cyanobacteria, 'sausage-shaped' *Synechococcus* populations embedded in a matrix of filaments (Reprinted from Ward et al., 1998).

Figure 1.2. Mushroom Spring showing cyanobacterial mat in foreground and spring in background. (Reprinted from Ward et al., 2006).

The focus of this work is Mushroom Spring, located approximately 0.5 km from Octopus Spring (Ramsing et al., 2000) and the temperature sites that will be discussed in this research at Mushroom Spring are 60°C and 65°C (Figure 1.2).

Molecular analyses have shown that the dominant *Synechococcus* populations native to these microbial mats are genetically distinct from readily cultivable *Synechococcus* strains (Ward et al., 1990, 1998; Ferris et al., 1996b). Culture-dependent methods of assaying microbial diversity have been severely biased and the cultivated strains retrieved using these methods are not representative of 'who is there' in nature

(Ward et al., 1990; Ferris and Ward, 1997). With respect to *Synechococcus* populations, Ferris et al. (1996b) showed that the 16S rRNA sequences from readily cultivable *Synechococcus* strains were not representative of the most abundant 16S rRNA sequences in a clone library constructed from mat DNA (Figure 1.3).

The native *Synechococcus* populations, referred to as A/B-type *Synechococcus*, have very similar 16S rRNA sequences that are distributed uniquely along the flow path, and thus, the temperature gradient in the effluent channels (Ferris and Ward, 1997). Specifically, *Synechococcus* 16S rRNA genotypes A′′, A′, A, B′ and B are found at progressively lower temperatures from the upper temperature limit of the mat near 72°C to ~50°C (Ferris and Ward, 1997; Ward et al., 2006; Klatt et al., 2010) and this can be visualized using denaturing gradient gel electrophoresis (DGGE, Figures 1.4 and 1.5). Allewalt et al. (2006) showed that *Synechococcus* isolates with A, A′ and B′ genotypes showed different temperature optima and upper temperature limits (Figure 1.6), as hypothesized from DGGE patterns (Figures 1.4 and 1.5; Ferris et al., 1996a; Ferris and Ward, 1997). The correspondence between these and temperature distribution results (Figures 1.4 and 1.5) led to a reevaluation of the nature of *Synechococcus* species within the microbial mat system and the application of the concept of a species as a population of microorganisms that occupy a distinct ecological niche, ***ecotypes,*** equivalent to Van Valen's ecological species (Ward, 1998).

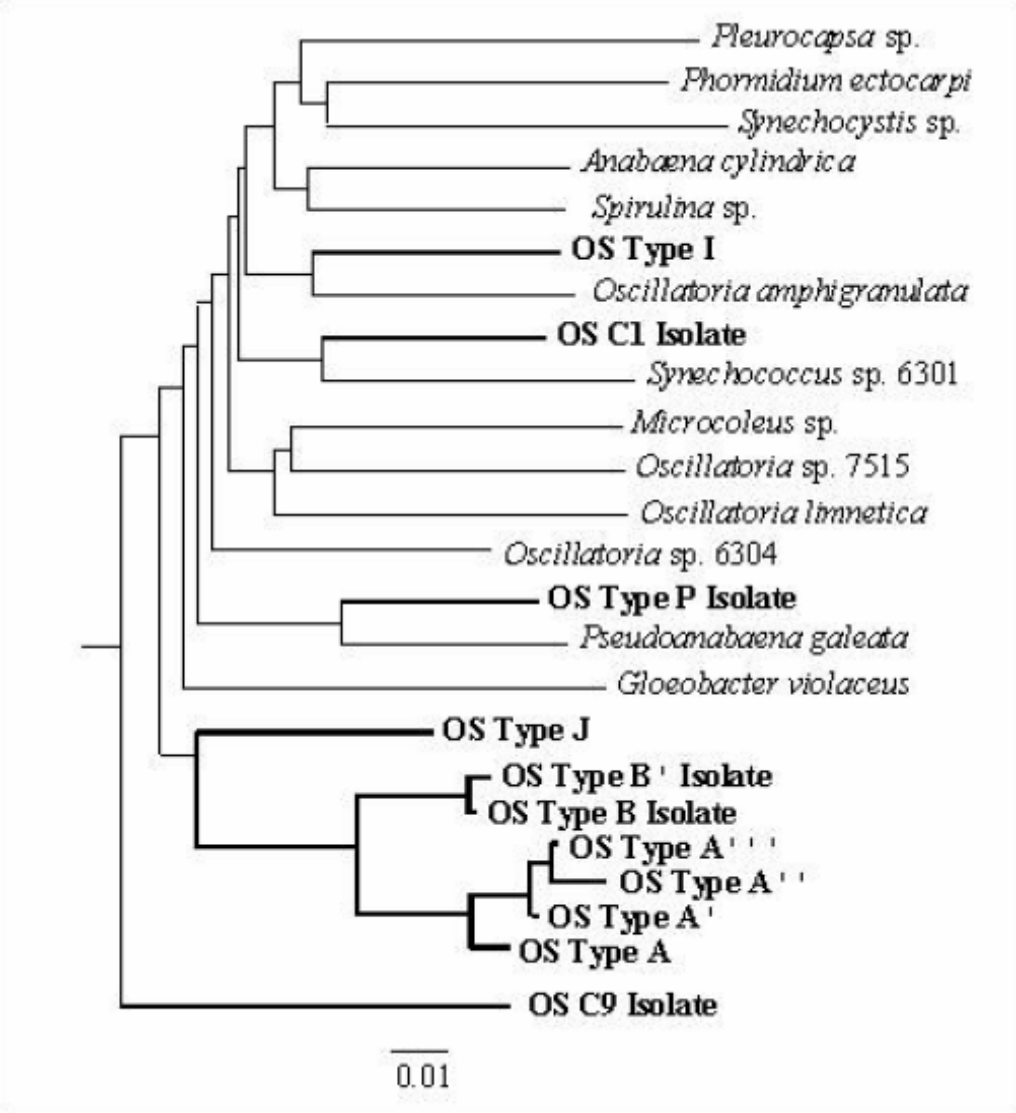



Figure 1.3. Distance matrix phylogenetic tree depicting cyanobacterial 16S rRNA sequences including those from the Octopus Spring mat. The scale bar corresponds to 0.01 nucleotide substitutions per sequence site. *Synechococcus lividus*, which is readily cultivated from undiluted mat samples, is represented by the OS C1 isolate. The cyanobacterial sequences detected in the Octopus Spring mat are highlighted in bold. The OS Type A clade and OS Type B´ clade contain the dominant sequences found in 16S rRNA PCR clone libraries of mat DNA (Reprinted from Ward et al., 1998).

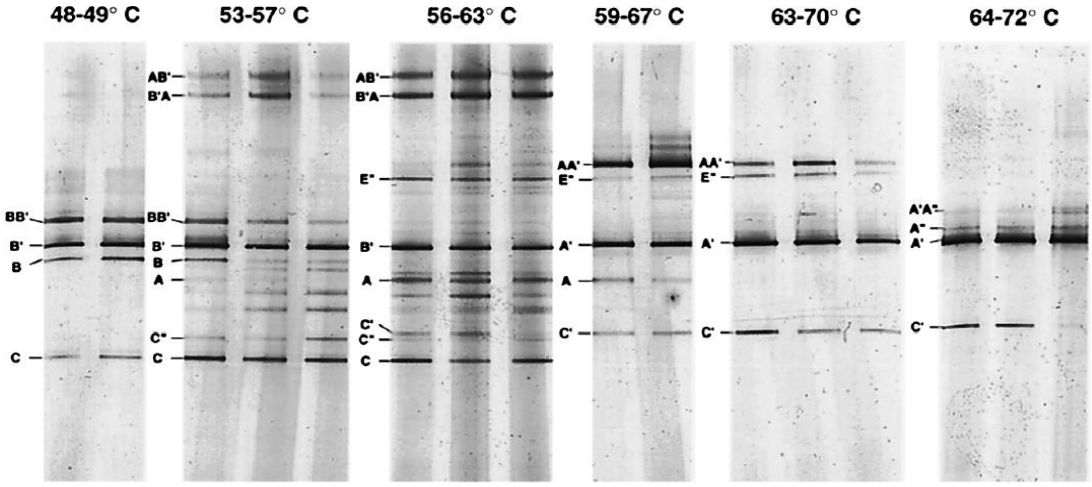



Figure 1.4. Denaturing gradient gel electrophoresis profiles of PCR-amplified 16S rRNA gene segments from DNA extracted from Octopus Spring microbial mat at temperature-defined sites. Letters and primes indicate 16S rRNA genotypes. Genotypes of interest in this dissertation are A, A′ and B′. Double bands (Labeled: AB′, BB′, B′A, and AA′) are heteroduplex artifacts (Reprinted from Ferris and Ward, 1997).

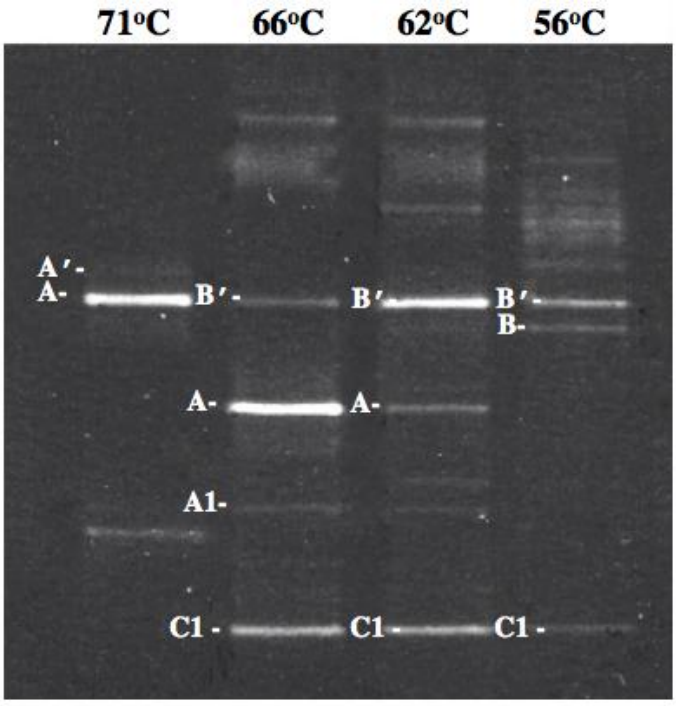



Figure 1.5. Denaturing gradient gel electrophoresis analysis of PCR-amplified 16S rRNA segments for mat samples collected at four temperature-defined sites along the Mushroom Spring effluent channel. Letters and primes indicate 16S rRNA genotypes (Reprinted from Ward et al., 2006).

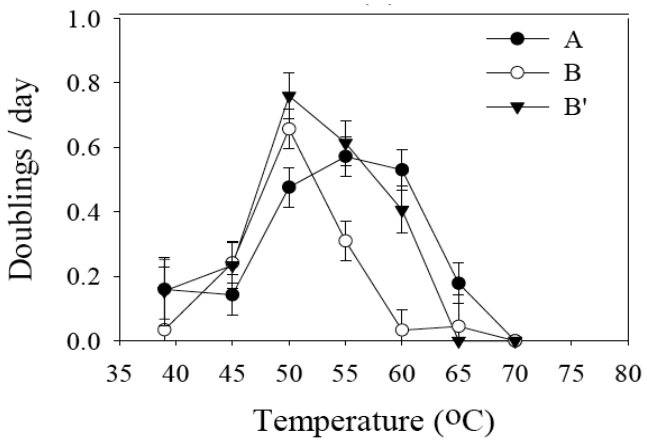



Figure 1.6. Comparison of growth rates of *Synechococcus* isolates with A, B, and B′ 16S rRNA genotypes with respect to temperature. Distribution of *Synechococcus* strains A, B and B′ in the mat according to temperature is 48°C-57°C for B, 48°C-63°C for B′, and 53°C-67°C for A (refer to Figure 1.4) Error bars correspond to +1 standard error (Reprinted from Allewalt et al., 2006).

At all temperatures along the effluent channel *Synechococcus* cells inhabiting the mat surface autoflouresce less intensely than those at greater depths (Figure 1.7). Light intensity and quality (spectral composition) also change dramatically in the lower layers of the mat relative to higher mat layers. Analysis using PCR and DGGE of microbial mat DNA samples from different depth intervals in the green layer, obtained by cryosectioning, provided additional evidence of ecotype populations along the vertical aspect of the mats. At 60°C, the 16S rRNA genotype B′ was the only *Synechococcus* genotype detected in the surface layer, whereas 16S rRNA genotype A was detected only at depths corresponding to the deeper *Synechococcus* population (Ramsing et al., 2000). At 68°C only the A′ *Synechococcus* 16S rRNA genotype occurs, raising the question of whether these phenotypically distinct populations correspond to one differently

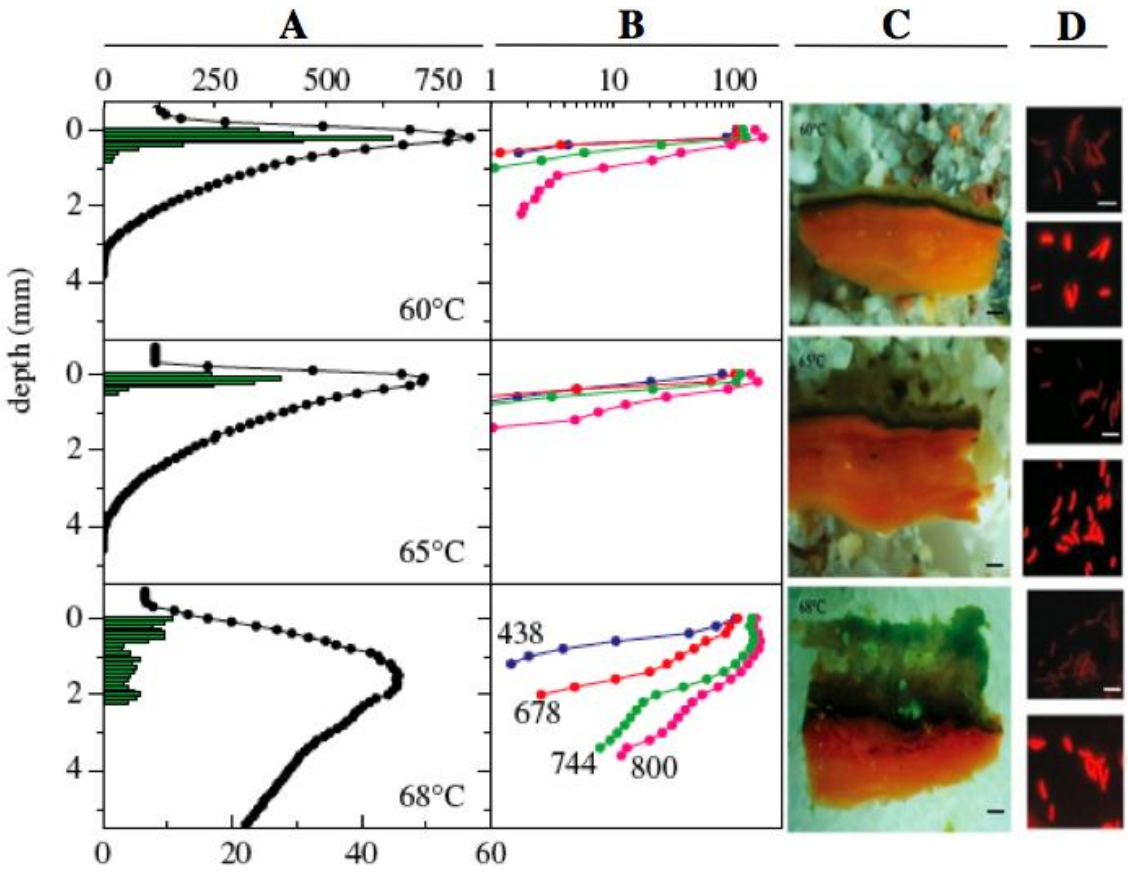



Figure 1.7. Detailed views of vertical profiles within samples from three temperature-sites in the Mushroom Spring mat. Column A, microprofiles of oxygen (black dots, top x-axis, μmol $l^{-1}$) and oxygenic photosynthesis (green bars, bottom x-axis, nmol $O_2$ $cm^{-3}$ $s^{-1}$). Column B, light at photosynthetically useful wavelengths (top x-axis, percentage of incident irradiance, numbers on curves indicate nanometers) measured with microsensors through the vertical aspect of mats at each temperature site. Column C, cross-section view of the mat layers (bar is 1 mm) and Column D, photomicrographs showing autoflourescence of *Synechococcus* populations sampled from the upper yellow-green (top) and deeper dark-green (bottom) parts of the top green layer at the different temperature sites in Column C (bar is 10 μm) (Adapted from Ward et al., 2006).

acclimated (temporarily adjusting to change in the environment) population or two differently adapted (becoming permanently better suited to the habitat) populations that are too closely related (i.e., ecotypes that are too 'young') to be detected using 16S rRNA sequence variation (Ferris et al., 2003). Using the more rapidly evolving 16S-23S rRNA-internal transcribed spacer (ITS) region, surface and subsurface *Synechococcus*

populations at 68°C were observed to be genetically distinct, suggesting they are likely to be differently adapted (based on environmental conditions that vary along the vertical gradient) species-like populations (ecotypes) (Figure 1.8). At 65°C only the A genotype occurs, but no variation is observed in either 16S rRNA or ITS sequences (Ward et al., 2006). ***This underscores the need to use an even more rapidly evolving genetic marker (or markers) to determine if all of the ecotypes have been detected, focusing on those that are closely related yet ecologically distinct.*** Analysis of oxygen and light quality in the mat (see above) suggests that factors other than temperature might define ecotypes. In the chapters of this dissertation I will discuss work that has been done using high-resolution, theory-based population genetics methods to determine if all the ecotypes within the Mushroom Spring microbial mat system have been detected, using single- and multi-locus analyses and two different models to predict ecotypes and 'fundamental populations.'

## Speciation Theory and Modeling Population Structure

### Ecotype Theory

Currently there is no theoretical basis to the practice of bacterial systematists utilizing a single sequence-identity cutoff value to demarcate the fundamental units of bacterial ecology and evolution (Cohan and Perry, 2007; Gevers et al., 2005; Ward et al., 2008). Most bacterial species demarcations to date have been tailored to fit the specific system under study and lack a theory based-guideline in addition to relying heavily of phylogenetic relationships. The problem with species demarcation in phylogenetic

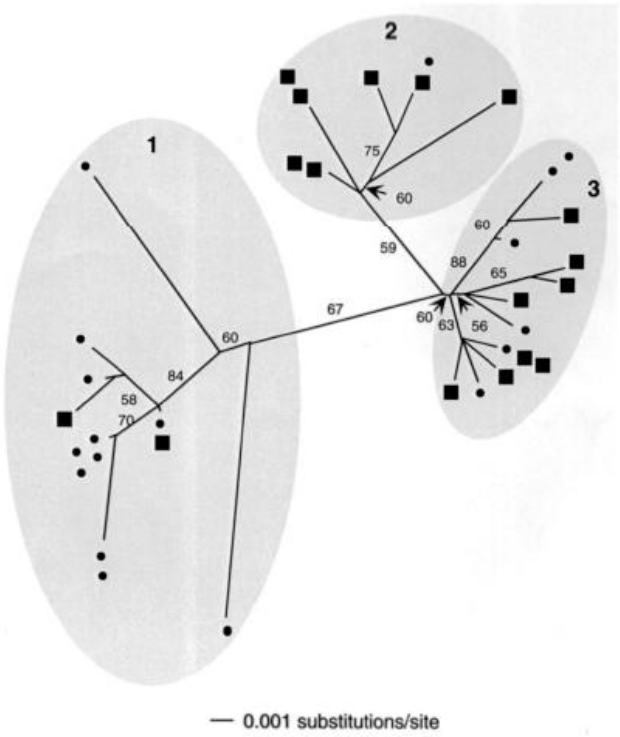



Figure 1.8. Unrooted neighbor-joining phylogenetic tree of *Synechococcus* genotype A´-like sequences based on 396 nucleotides of the ITS region adjacent to the 16S rRNA gene. Circles represent surface sequences; squares represent subsurface sequences. Bootstrap values over 50% are shown (Reprinted from Ferris et al., 2003).

analysis is that it is arbitrary in its interpretation of clustering. "Phylogenies contain a hierarchy of subclusters within clusters and it is never clear what level of clustering corresponds to ecotypes" (Koeppel et al., 2008). Cohan and Perry (2007) discussed 9 models of speciation that may exist for different bacteria. The models focus on the different ways in which populations can evolve considering such factors as the environment (stable ecotype model), geographic isolation (geotype model), and lack of cohesion (species-less model), among others. Because there is evidence of ecological partitioning of *Synechococcus* populations in the microbial mat system, I will focus on the stable ecotype model (Figure 1.9).

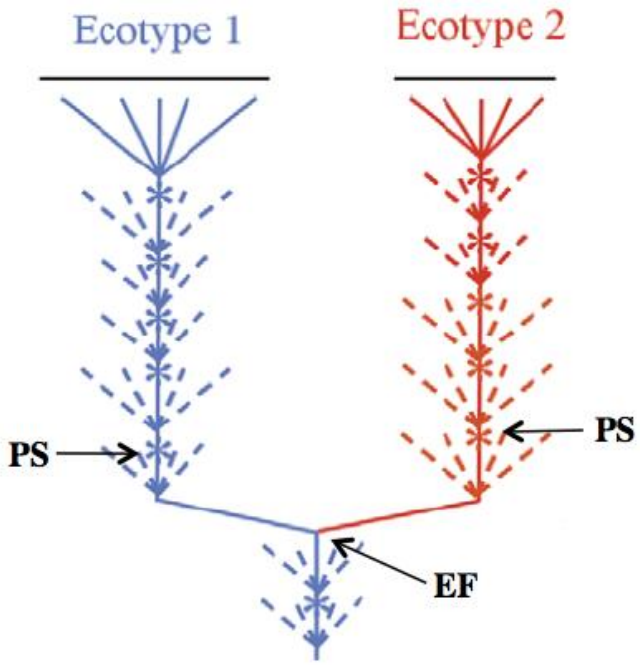



Figure 1.9. The stable ecotype model of speciation depicting the relationship between ecologically distinct populations and DNA sequence clusters. Different ecotypes are represented by different colors. EF shows an ecotype formation event. PS (asterisks) shows examples of periodic selection events that purge population diversity except for one most-fit individual. Dashed lines represent extinct lineages (Reprinted from Cohan and Perry, 2007).

As shown in Figure 1.9, the stable ecotype model assumes that periodic selection (PS) is the cohesive force that holds the individuals of an ecotype population together. An ecotype lineage progresses through time gathering diversity via recombination and mutation until a PS event occurs. PS may result from an environmental disturbance that affects the ecological niche, such as a sudden permanent change in light quality or temperature. Among the diverse variants of an ecotype cluster there are mutants that will be able to survive the environmental change or selection pressure and persist through the PS event while the rest of the population diversity is extinguished. The surviving mutants

will, over time, give rise to new variant diversity until the next PS event. New ecotype formation (EF) occurs when a mutation or set of accumulated mutations allows an individual in the population to exploit a new set of environmental resources in a new niche not used by the parental population. This variant is not affected by PS events that squash diversity in the parent population and diverges from the parental ecotype population founding a new lineage. It gives rise to variants and experiences its own PS events, eventually becoming irreversibly diverged and forming a new ecotype.

Modeling Population Structure

Several approaches have been developed to predict naturally evolved populations from sequence data in an environmental sample. Two will be discussed and implemented in this research: (i) single-locus analysis (SLA) and multi-locus sequence analysis (MLSA) of single and multiple concatenated gene sequences using Ecotype Simulation (ES) (Koeppel et al., 2008) and (ii) MLSA using Multi-Locus Sequence Typing-eBURST (MLST-eBURST) (Feil et al., 2004). Another program worth mentioning but that will not be used in this work is AdaptML (Hunt et al., 2008). AdaptML provides analysis of the presence or absence of correlation between predicted ecotypes and investigator-defined "habitats" (i.e., niche-defining features) using nucleotide sequence data. Use of AdaptML in the current work was limited by the fact that only two 'habitats' could be defined (i.e. high and low temperature sites). Although there are more possible ecotype-defining 'habitats' in the mat system, including (e.g., nutrients that vary along the flow path and features that vary along the vertical dimension, such as light intensity, light quality and

oxygen levels), these were not analyzed in the current study. AdaptML may be evaluated in future work.

Single- and Multi-Locus Analyses Using Ecotype Simulation. In ES, an ecotype is modeled as an ecologically distinct group, whose diversity is limited by a force of cohesion, usually the genome wide purging of diversity known as a PS event but also by genetic drift. In short, it is based on the stable ecotype model. Two ecotypes are considered separate when they are ecologically distinct from each other and belong to genetically cohesive and irreversibly separate evolutionary lineages (Cohan and Koeppel, 2008; Koeppel et al., 2008; Gevers et al., 2005; de Queiroz, 2005).

The ES algorithm simulates the evolutionary history of the organisms sampled from nature under different "values of: PS ($\sigma$), EF ($\Omega$) and the number (n) of putative ecotypes (PE)." Drift is negligible when effective population sizes are large (Cohan and Perry, 2007; Koeppel et al., 2008; Ward and Cohan, 2005; Cohan and Koeppel, 2008). ES uses a "coalescence approach" described in Hudson (1990). Briefly, "a 'backward' simulation is run that stochastically produces a phylogenetic representation of the history of the community comprised of ecotypes, establishing nodes of coalescence of lineages and time between nodes. It ends when all of the branches have coalesced into a single node representing the most recent common ancestor of all the organisms sampled" (Cohan and Koeppel, 2008; Koeppel et al., 2008). The pattern of events generated from the backward simulation is then used as a "scaffold" for the forward simulation. Forward simulation then allows for "[point] mutation nucleotide substitutions throughout the history of the clade according to the established scaffold." ES attempts to model the real

evolutionary history of a gene or concatenation of genes by finding the most likely set of values of the parameters n, EF (Ω) and PS (σ) relative to the observed phylogeny clade sequence diversity (Koeppel et al., 2008). ES does not explicitly account for the introduction of variants into an ecotype by recombination within the analysis. The analysis allows for PS to limit within-ecotype sequence diversity because recombination is never frequent enough to hinder the purging events of PS. Recombinants are removed from the sequence dataset before ES analysis (Koeppel et al., 2008).

Once the analysis has been run, a prediction of the number of PEs and the 95% confidence intervals (95% CIs) is provided. The ecotype prediction then must be overlaid on the phylogeny using either the manual or automatic demarcation program embedded within the ES program. Ecotypes are demarcated conservatively. In the manual demarcation method used in this dissertation, a set of sequences belonging to a clade is uploaded into the ES demarcation program and ES produces a PE prediction (and 95% CI) based on those sequences only. If the 95% CI includes 1, then the clade is considered to be one ecotype, even if the most likely predicted number of PEs is >1. If the 95% CI does not include 1, then sequences are removed from the clade and the remaining subclade is reanalyzed until the 95% CI includes 1. Figure 1.10 illustrates the process of determining a PE through manual demarcation on a phylogenetic tree. All sequences are submitted to the ES demarcation program, if the 95% CI does not include 1, then the sequence (or sequences) lying outside of the terminal clade (most divergent clade), meaning the sequence(s) closest to the common ancestor, are removed. Once the sequence or sequences have been removed the remaining are ‘re-tried’ in demarcation;

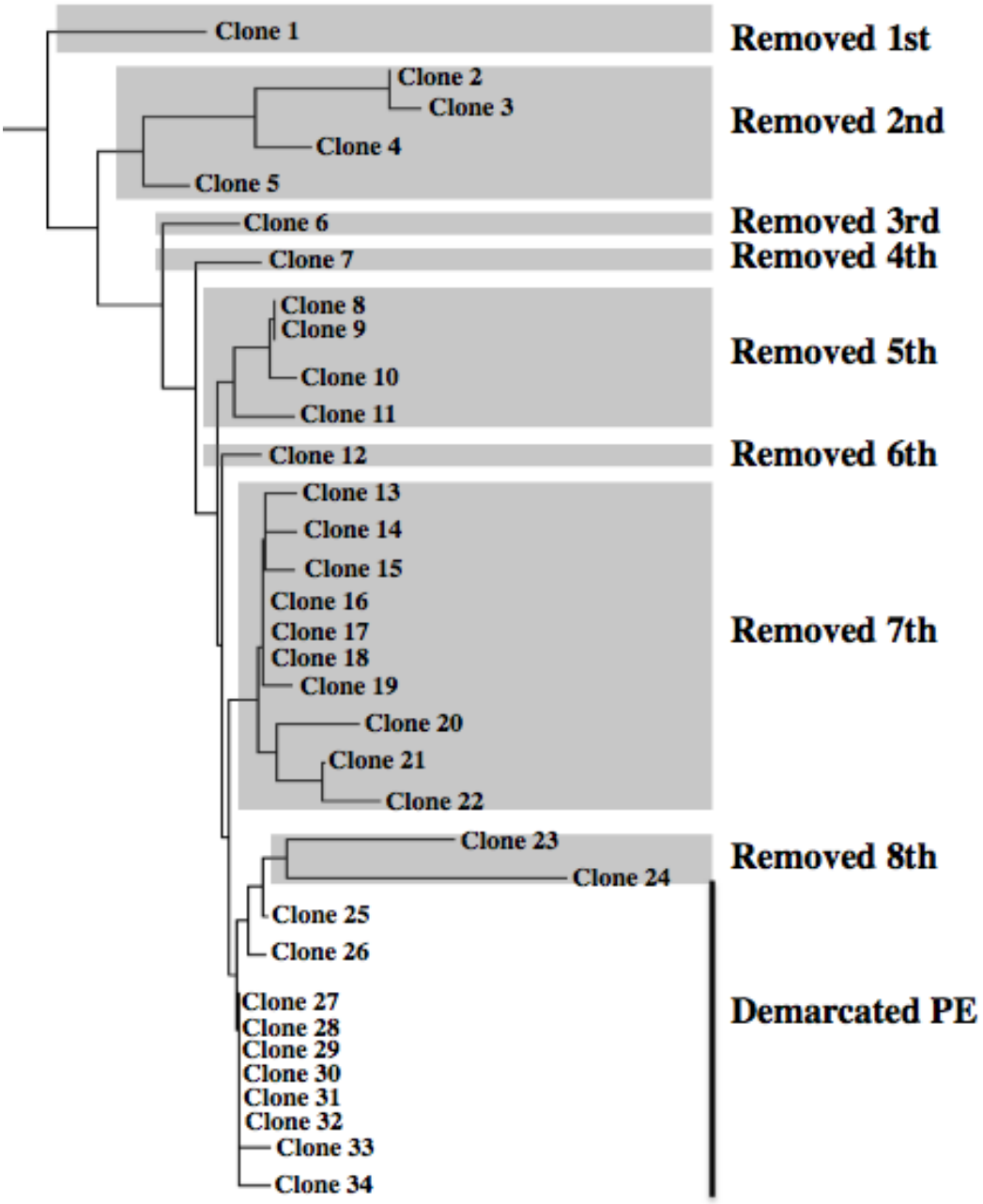



Figure 1.10. An example of the demarcation of ecotypes using manual demarcation in the ecotype simulation program. Sequences are removed progressing from those closest of the common ancestor of a clade toward a terminal clade, until ecotype simulation predicts a putative ecotype clade, where 1 is included within the 95% CI.

again if the 95% CI does not include 1, more outlying sequences are removed and this process is repeated. Notice that the last removal was not a removal of the sequence closest to the common ancestor but rather the 2 sequences that were most divergent from the rest of the clade (removal #8). Once these sequences were removed, the demarcation predicted 2 PEs with a 95% CI that included 1, so the remaining sequences were demarcated as a single ecotype.

ES can be used to analyze single or concatenated multi-locus sequence datasets. For instance, Connor et al. (2010) and Koeppel et al. (2008) have shown using ES and concatenated multiple loci that clades of gene sequence data for *Bacillus simplex* isolates cultivated from the "Evolution Canyons" of Israel are comprised of multiple ecotypes within this named species. Each ecotype was predicted and confirmed to be an ecologically distinct cluster by taking into account the distribution of samples from which the isolates were obtained, which suggested distinct niches on different canyon slopes that have different solar exposures (Koeppel et al., 2008). Further work showed that ecotypes of *Bacillus subtilis* isolates from Death Valley predicted by ES were seen to have distinct habitat associations (as determined by a $\chi^2$ contingency test and with AdaptML analysis) for soil texture and solar exposure (Connor et al., 2010). Preliminary analysis of mat *Synechococcus* A-like and B´-like population 16S rRNA-ITS sequence data (single-locus only) using ES showed ecotypes adapted to temperature and depth confirming the utility of ES in predicting ecotypes (Ward et al., 2006) (Figure 1.11).

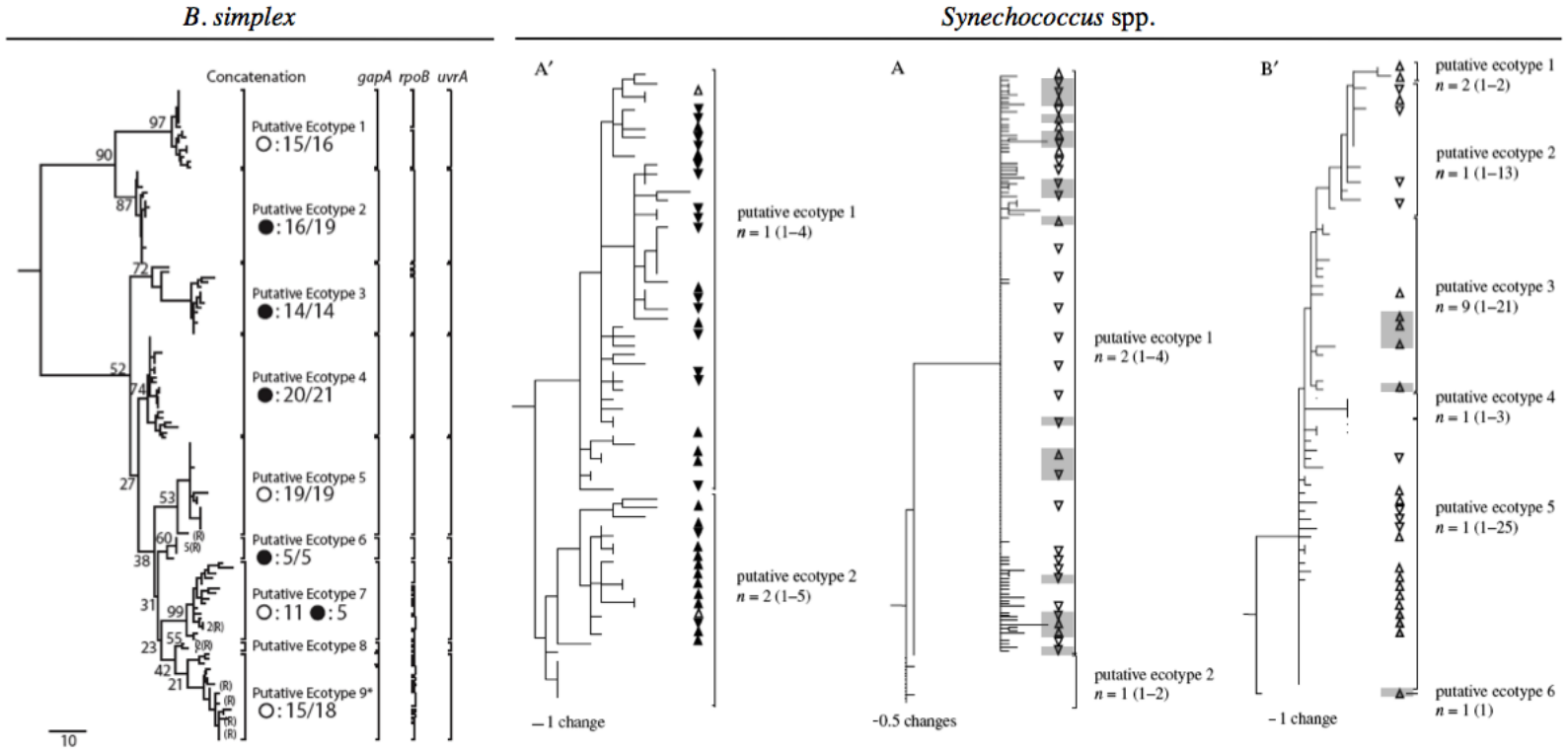



Figure 1.11. Phylogeny and ecotype simulation demarcation of putative ecotypes for *B. simplex* isolates from Evolution Canyon, Israel and *Synechococcus* A´, A and B´ lineages from Mushroom Spring. Ecotype demarcations are indicated by brackets. *B. simplex* results show a 3 gene (*gapA*, *rpoB*, and *uvrA*) concatenated and single-locus parsimony phylogenies with ecotype demarcation. Open circles indicate the south facing-slope microhabitat; closed circles indicate the north-facing slope microhabitat. *Synechococcus* A´, A, and B´ parsimony phylogenetic trees with demarcation of ecotypes by an early version of ES analysis of ITS variants. *n* is the most likely number of ecotypes within a putative ecotype clade and parentheses enclose the 95% confidence intervals. Clones originating from the upper and lower regions of photic zone are indicated by upward and downward-pointing triangles. Shading indicates temperature: white, 60°C; grey-shaded, 65°C; and black, 68°C (Adapted from Koeppel et al., 2008 and Ward et al., 2006).

Multi-Locus Sequence Analysis Using MLST-eBURST. MLST-eBURST, which has been successfully used in medical bacteriology for typing cultivated pathogens, clusters organisms by allelic variation allowing for the independent assortment of alleles, and is not based on the stable ecotype model. MLST-eBURST analyzes allelic profile data compiled by the researcher. However, the advantage of MLSA using MLST-eBURST (referred to from here forward as simply eBURST) is that it can account and 'buffer' for recombination (Hanage et al., 2005; Vos and Didelot, 2009), something ES analysis does not do. The eBURST program essentially divides an allelic dataset into groups of related isolates called clonal complexes and predicts a founding or ancestral genotype for each clonal complex. Allelic datasets are comprised of sequences (typically ~450 nucleotides long) for multiple housekeeping loci (typically 5-7) in the variant being studied. All the sequences for a gene are aligned and sequences that differ are assigned unique allele numbers. The allelic designations for each gene are compiled to make an allelic profile for each clone (or isolate) and given a unique identifier, called a sequence type number (i.e. ST1, ST2 and so on). STs that are identical at all loci form a consensus group, which enucleates a clonal complex in eBURST analysis. Given natural microbial diversification over time, variants can arise that may have different alleles at one or more of the loci compared to the consensus group. Variants that differ at one or two of the 5-7 loci are called single-locus or double-locus variants (SLV or DLV). Clonal complexes are formed from consensus groups plus SLVs and sometimes DLVs depending on the investigator's preference (Feil et al., 2004). The eBURST analysis displays the most likely pattern of evolutionary descent within each clonal complex but does not

reconstruct the evolutionary history between or within clonal complexes. In other words, while the primary "founder" ST is predicted, which SLV it gave rise to first cannot be discerned. The primary founder is predicted on the basis of parsimony, as the ST that has the largest number of SLVs in the group or clonal complex. This assumes that the initial diversification of a clone results in variants of the founder that differ at one of the multiple loci analyzed (Feil et al., 2004).

MLSA using eBURST has allowed the visualization of the complicated population structures of *Staphylococcus aureus* (Figure 1.12)*, Streptococcus pneumoniae,* and *Neisseria meningitidis* (Feil et al., 2004). By taking into account the number of SNPs differentiating SLVs and DLVs from the consensus group, eBURST also provides data on the ratio of recombination to mutation (*r/m*). Allelic variants defined by 1 SNP are hypothesized to have arisen via mutation, whereas variants defined by multiple SNPs are hypothesized to have arisen via recombination (Feil et al., 1999, 2000 and 2001).

## Population Genetics Analysis of *Synechococcus*

### A-like and B´like populations in the Mushroom Spring Mat

This dissertation builds on previous work on ecotype demarcation and phylogenetic analysis using 16S rRNA and ITS sequence data from *Synechococcus* populations, in the Mushroom Spring mat (Ferris et al., 2003; Ward et al., 1998, 2006). Phenotypically distinct *Synechococcus* populations found at different depths within the vertical gradient (Figure 1.7) were distinguishable at 60°C and 68°C using these loci, but indistinguishable at 65°C, underscoring the need to use higher resolution techniques to discern all the ecotypes within the *Synechococcus* populations in the microbial mat.

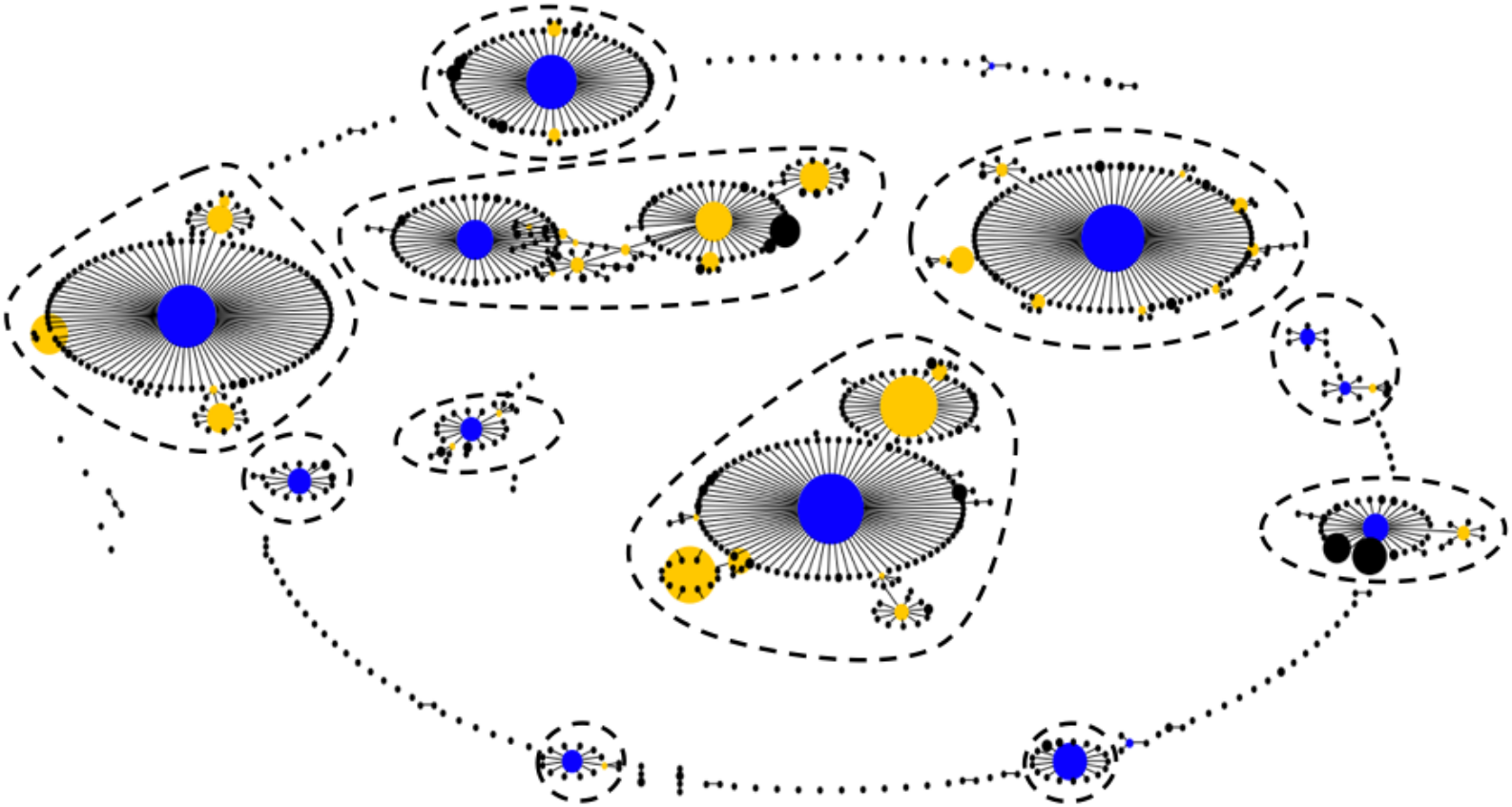


Figure 1.12. An example of an eBURST population snapshot of *Staphylococcus aureus* isolates constructed from allelic profile data from 7 loci available from the MLST database (http://eburst.mlst.net). Examples of clonal complexes (consensus group/founder in blue and at least 3 SLVs; subfounders are in yellow) are enclosed in dashed circles. All other sequence types (single black dots) and clonal complexes are portrayed in a spiral pattern.

The following chapters will address the hypotheses that:

**(i) There are more ecotypes within the Mushroom Spring microbial mat than have been previously identified using cultivation-independent 16S rRNA-ITS PCR, cloning and sequencing methods**

**(ii) These ecotypes exhibit distribution patterns that suggest they are ecologically distinct.**

Chapter 2 considers the use of single protein-encoding loci, genes essential for the metabolic function of the organism, combined with ES and phylogenetic analysis to determine whether the resolution of ecotypes is increased. A concern of SLA analysis is that it does not account explicitly for recombination. I will also introduce evidence to support the need for yet higher resolution techniques using multiple loci (MLSA) to further resolve ecotype structure, which also permits testing the hypothesis that:

**(iii) Recombination has been less important than mutation in shaping the evolution of native *Synechococcus* populations.**

Chapter 3 addresses the creation and characterization of two metagenomic bacterial artificial chromosome (BAC) libraries. Historically, MLSA analysis has been conducted on isolates that have been cultivated due to the need to obtain sequences for many loci from around the genomes of individual organisms. The challenge in a cultivation-independent approach is obtaining multiple loci from an individual genome and, in the case of this study, linking them to the 16S rRNA locus for comparison to previous research. To obtain this linkage it was necessary to consider methods developed in the last decade for cloning large inserts of genomic DNA using BAC, cosmid or

fosmid vectors. (Beja et al., 2000a/b; Rondon et al., 1999 and 2000; Shizuya et al., 1992; Stein et al., 1996). BAC cloning is preferable to other approaches because other cloning vectors, such as yeast artificial chromosomes and cosmids have been found to be unstable or to produce chimeras (Rondon et al., 1999; Shizuya et al., 1992; Monaco and Larin, 1994). BAC clone libraries have been used to link function, phylogeny and, for instance, have led to such discoveries as novel phototrophy in the oceans through the use of proteorhodopsins rather than chlorophylls (Beja et al., 2000b). They have also been used in explorations of bacterial symbionts in sponges and bacterial diversity, antibiotic resistance genes and biologically active small molecules involved in quorum sensing in soil samples (Ouyang et al., 2009; Liles et al., 2003; Riesenfeld et al., 2004; Williamson et al., 2005).

Chapter 4 provides the results of MLSA, using ES and eBURST analyses to demarcate ecotypes and clonal complexes. I will also address the topics of linkage, recombination, and mutation. Finally, I evaluate the programs available for analyzing multi-locus sequence data and the impact of recombination and linkage on MLSA.

Chapter 5 introduces single-cell genomics and multiple-displacement amplification (MDA), a new technique that may prove to be a more efficient method for analysis of microbial population genetics in the future. Chapter 6 provides a synthesis of previous chapters, a discussion of the genetic structure of *Synechococcus* populations in the Mushroom Spring microbial mat and the implications of my results on species concepts as they apply to our system and microbial ecology as a discipline.

References

Allewalt JA, Bateson MM, Revsbech NP, Slack K and Ward DM. (2006). Temperature and light adaptations of *Synechococcus* isolates from the microbial mat community in Octopus Spring, Yellowstone National Park. *Appl Environ Microbiol* **72**: 544-550.

Beja O, Suzuki MT, Koonin EV, Aravind L, Hadd A, Nguyen LP, Villacorta R, Amjadi M, Garrigues C, Jovanovich SB, Feldman RA and DeLong EF. (2000a). Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ Microbiol* **2**: 516-529.

Beja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP, Jovanovich SB, Gates CM, Feldman RA, Spudich JL, Spudich EN and DeLong EF. (2000b). Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289**: 1902-1996.

Cohan FM and Koeppel AF. (2008). The origins of ecological diversity in prokaryotes. *Curr Biol* **18**: R1024-1034.

Cohan FM and Perry EB. (2007). A Systematics for discovering the fundamental units of bacterial diversity. *Curr Biol* **17**: R373-386.

Cohan FM. (2006). Towards a conceptual and operational union of bacterial systematics, ecology and evolution. *Phil Trans Roy Soc B* **361**: 1985-1996.

Cohan FM. (2002). What are bacterial species? *Annu Rev Microbiol* **54**: 457-487.

Connor N, Sikorski J, Rooney AP, Kopac S, Koeppel AF, Burger A, Cole SG, Perry EB, Krizanc D, Field NC, Slaton M and Cohan FM. (2010). The ecology of speciation in *Bacillus*. *Appl Environ Microbiol* Online: doi:10.1128/AEM.01988-09.

Darwin CR. (2001). *The Origin of Species*. Vol 11, Harvard Classics. P.F. Collier & Son: New York.

de Queiroz K. (2007). Species concepts and species delimitation. *Syst Biol* **56**: 879-886.

de Queiroz K. (2005). Ernst Mayr and the modern concept of species. *Proc Natl Acad Sci* 102 **(Suppl 1)**: 6600-6607.

Doolitte WF and Papke RT. (2006). Genomics and the bacterial speciation problem. *Genome Biol* **7**:116.1-116.7

Dykhuizen DE. (1998). Santa Rosalia revisited: why are there so many species of bacteria? *Antonie Van Leeuwenhoek* **73**: 25-33.

Feil EJ, Li BC, Aanensen DM, Hanage WP and Spratt BG. (2004). eBURST: Inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J Bacteriol* **186**: 1518-1530.

Feil EJ, Holmes EC, Bessen DE, Chan M-S, Day NPJ, Enright MC, Goldstein R, Hood DW, Kalia A, Moore CE, Zhou J and Spratt BG. (2001). Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc Natl Acad Sci* **98**:182-187.

Feil EJ, Maynard Smith J, Enright MC and Spratt BG. (2000). Estimating recombination parameters in *Streptococcus pneumoniae* from multi-locus sequence typing data. *Genet* **154**: 1439-1450.

Feil EJ, Maiden MCJ, Achtman M and Spratt BG. (1999). The relative contribution of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. *Mol Biol Evol* **16**: 1496-1502.

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

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

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

Ferris MJ, Ruff-Roberts AL, Kopczynski ED, Bateson MM and Ward DM. (1996b). Enrichment culture and microscopy conceal diverse thermophilic *Synechococcus* populations in a single hot spring microbial mat community. *Appl Environ Microbiol* **62**: 1045-1050.

Gans J, Wolinsky M and Dunbar J. (2005). Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science* **309:**1387-90.

Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackenbrandt E, Van de Peer Y, Vandamme P, Thompson FL and Swings J. (2005). Opinion: Re-evaluating prokaryotic species. *Nat Rev Microbiol* **3**: 733-739.

Giovannoni SJ and Stingl U. (2005). Molecular diversity and ecology of microbial plankton. *Nature* **437**: 343-348.

Goodfellow M, Manfio GP and Chun J. (1997). Towards a practical species concept for cultivable bacteria. In: Claridge MF, Dawah HA and Wilson MR (eds). *Species: the units of biodiversity*. Chapman and Hall: London. Pp. 25-59.

Hanage WP, Fraser C and Spratt BG. (2005). Fuzzy species among recombinogenic bacteria. *BMC Biol* **3**: 6-13.

Hudson RR. (1990). Gene genealogies and the coalescent process. *Oxford Surv Evol Biol* **7**: 1-44.

Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ and Polz MF. (2008). Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* **320**: 1081-1085.

Klatt CG, Wood JM, Rusch DB, Bateson MM, Heidelberg JF, Bryant DA and Ward DM. (2010). Composition of metagenomes from a phototrophic hot spring microbial mat community. In prep.

Koeppel AF, Perry EB, Sikorski J, Kriznac D, Warner WA, Ward DM, Rooney AP, Brambilla E, Connor N, Ratcliff RM, Nevo E and Cohan FM. (2008). Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. *Proc Natl Acad Sci* **105**: 2504-2509.

Konstantinidis KT and Tiedje JM. (2005). Toward a genome-based taxonomy for prokaryotes. *J Bacteriol* **187**: 6258-6264.

Liles MR, Manske BF, Bintrim SB, Handelsman J and Goodman RM. (2003). A census of rRNA genes and linked genomic sequences within a soil metagenomic library. *Appl Environ Microbiol* **69**: 2684-2691.

Mayden RL. (1997). A hierarchy of species concepts: the denouement in the saga of the species problem. In: Claridge MF, Dawah HA and Wilson MR (eds). *Species: the units of biodiversity*. Chapman and Hall: London. Pp. 381–424.

Mayr E. (1982). *The Growth of Biological Thought: Diversity, Evolution, and Inheritance.* Belknap Press of Harvard University Press: Cambridge, MA.

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

Monaco AP and Larin Z. (1994). YACs, BACs, PACs, and MACs: artificial chromosomes as research tools. *Trends Biotechnol* **12**:280-286.

Nübel U, Garcia-Pichel and Muyzer G. (1997). PCR primers to amplify 16S rRNA genes from cyanobacteria. *Appl Environ Microbiol* **63**: 3327-3332.

Oyaung Y, Dai S, Xie L, Ravi Kumar MS, Sun W, Sun H, Tang D and Li X. (2009). Isolation of high molecular weight DNA from marine sponge bacteria for BAC library construction. *Mar Biotechnol* Aug 15.

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

Riesenfeld, CS; RM Goodman and J Handelsman. 2004. Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environmental Microbiology.* **6**: 981-989.

Rocap G, Distel DL, Waterbury JB and Chisholm SW. (2002). Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. *Appl Environ Microbiol* **68**: 1180-1191.

Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR, Loiacono KA, Lynch BA, MacNeil IA, Minor C, Tiong CL, Gilman M, Osburne MS, Clardy J, Handelsman J and Goodman RM. (2000). Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl Environ Microbiol* **66**: 2541-2547

Rondon MR, Raffel SJ, Goodman RM and Handelsman J. (1999). Toward functional genomics in bacteria: analysis of gene expression in *Escherchia coli* from a bacterial artificial chromosome library of *Bacillus cereus. Proc Natl Acad Sci* **96**: 6451-6455.

Roselló-Mora R and Amann R. (2001). The species concept for prokaryotes. *FEMS Microbiol Rev* 25: 39-67.

Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, Tachiri Y and Simon M. (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherchia coli* using an F-factor-based vector. *Proc Natl Acad Sci* **89**: 8794-8797.

Simpson GG. (1961). The species concept. *Evol* **5**: 285-298.

Stackenbrandt E and Geobel BM. (1994). Taxonomic note: a place for DNA: DNA reassociation and the 16S rRNA sequence analysis in the present species definition in bacteriology. *Int J Syst Bacteriol* **44**: 846-849.

Stein JL, Marsh TL, Wu KY, Shizuya H and DeLong EF. (1996). Characterization of uncultivated prokaryotes: isolation and analysis of 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J Bacteriol* **178**: 591-599.

Van Valen L. (1976). Ecological species, multispecies, and oaks. *Taxon* **25**: 233-239.

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

Vos M and Didelot X. (2009). A comparison of homologous recombination rates in bacteria and archaea. *ISME J* **3**: 199-208.

Ward DM, Cohan FM, Bhaya D, Heidelberg JF, Kuhl M and Grossman A. (2008). Genomics, environmental genomics and the issue of microbial species. *Heredity* **100**: 207-219.

Ward DM, Bateson MM, Ferris MJ, Kuhl M, Wieland M, Koeppel A and Cohan FM. (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. *Phil Trans Roy Soc Ser B* **361**: 1997-2008.

Ward DM. (1998). A natural species concept for prokaryotes. *Curr Opin Microbiol* **1**: 271-277.

Ward DM and Cohan FM. (2005). Microbial diversity in hot spring cyanobacterial mats: patter and prediction. In: Inskeep WP and McDermott T (eds) *Geothermal Biology and Geochemistry in Yellowstone National Park*. Thermal Biology Institute: Bozeman, MT. Pp 185-201.

Ward DM, Ferris MJ, Nold SC and Bateson MM. (1998). A natural view of microbial biodiversity within hot spring cyanobacterial mat communities. *Microbiol Mol Biol Rev* **62**: 1353-1370.

Ward DM, Weller R and Bateson MM. (1990). 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature* **344**: 63-65.

Wayne LG, Brenner DJ, Colwell RR, Grimont PAD, Kandler O, Krichevsky MI, Moore WEC, Murray RGE, Stackenbrandt E, Starr MP and Truper HG. (1987). Report of the *ad hoc* committee on reconciliation of approaches to bacterial systematics. *Int J Syst Bacteriol* **37**: 463-464.

Williamson LL, Borlee BR, Schloss PD, Guan C, Allen HK and Handelsman J. (2005). Intracellular screen to identify metagenomic clones that induce or inhibit a quorum-sensing biosensor. *Appl Environ Microbiol* **71**: 6335-6344.

# CHAPTER 2

# ECOLOGICAL DIVERSITY OF *SYNECHOCOCCUS* POPULATIONS INHABITING AN ALKALINE SILICEOUS HOT SPRING MICROBIAL MAT IN YELLOWSTONE NATIONAL PARK, WYOMING, MEASURED USING CULTIVATION-INDEPENDENT ANALYSIS OF PROTEIN-ENCODING GENES AND EVOLUTIONARY SIMULATION[1]

Melanie C. Melendrez[2], Rachel K. Lange[2], Fredrick M. Cohan[3] and David M. Ward[1]

## Abstract

Previous research has shown that sequences of 16S rRNA genes and the 16S-23S rRNA internal transcribed spacer regions may not have enough genetic resolution to define all ecologically distinct *Synechococcus* populations (ecotypes) inhabiting alkaline, siliceous hot spring microbial mats. To achieve higher molecular resolution, we studied sequence variation in three *Synechococcus* protein-encoding loci: *rbsK* (ribokinase), *aroA* (3-phosphoshikimate 1-carboxyvinyltransferase), and *apcAB* (allophycocyanin alpha/beta subunits) using cultivation-independent methods. A segment of each gene was PCR-amplified separately from DNA extracted from samples collected from 60°C and 65°C sites in the Mushroom Spring mat (Yellowstone National Park, WY), cloned and sequenced. Sequences were analyzed using the Ecotype Simulation algorithm, based on the Stable Ecotype Model of speciation, to identify putative ecotypes. Between 6.5 and 14 times more putative ecotypes were predicted from variation in protein-encoding loci

---

[1] This study was conducted in collaboration with MSU undergraduate Rachel K. Lange. We each constructed 3 PCR clone libraries. All analysis, figures and tables were done by me.

[2] Land Resources and Environmental Science, Montana State University, Bozeman, MT.
[3] Biology Department, Wesleyan University, Middletown, CT.

than from variation in 16S rRNA/ITS sequences (1-2 ecotypes). The number of putative ecotypes predicted depended on the number of sequences sampled and the molecular resolution of the locus. The Chao estimation of diversity indicated that not many rare ecotypes were missed and the number of operational taxonomic units predicted based on a 99% similarity cutoff was similar to the number of putative ecotypes demarcated in ecotype simulation analysis. Some putative ecotypes were observed to be sample-specific, suggesting different adaptation to temperature or other parameters that vary along the flow channel. There was evidence, however, that conservative putative ecotype demarcation may underestimate the true number of ecotypes.

## Introduction

Molecular analyses have shown that the dominant *Synechococcus* populations native to microbial mats in alkaline siliceous hot springs, such as Octopus Spring and Mushroom Spring, Yellowstone National Park, Wyoming, are genetically distinct from readily cultivable *Synechococcus* strains (Ward et al., 1990; Ward et al., 1998; Ferris et al., 1996b). The dominant native *Synechococcus* populations, referred to as A/B-type *Synechococcus*, have very similar 16S rRNA sequences that are distributed uniquely along the flow path, and thus, temperature gradient in the effluent channels (Ferris and Ward, 1997). Specifically, *Synechococcus* 16S rRNA genotypes A′′, A′, A, B′′ and B are found at progressively lower temperatures from the upper temperature limit of the mat near 72°C to ~50°C (Ferris and Ward, 1997; Ward et al., 2006; Klatt et al., 2010). This suggested that distinct 16S rRNA sequences might represent populations with

distinct ecologies (ecotypes), corresponding to populations adapted to different temperatures, as had been suggested previously for *Synechococcus* isolates from an Oregon hot spring mat (Peary and Castenholz, 1964; Miller et al., 2000). Allewalt et al. (2006) succeeded in cultivating *Synechococcus* with 16S rRNA sequences corresponding to A-, B´ and B-type *Synechococcus* genotypes, which are found at 56°C to 68°C, 53°C to 60°C and 53°C to 55°C regions, respectively (Ferris and Ward, 1997). Allewalt et al. (2006) also confirmed that these isolates exhibited temperature adaptations corresponding to their *in situ* distributions. Ward (1998) pointed out that ecotypes could be considered species using an ecological species concept, wherein the individual organisms comprising a species share the same environmental niche or "adaptive zone" (Van Valen, 1976).

These ecotypes can be discerned from DNA sequence data provided "a particular model of bacterial evolution applies" (Connor et al., 2010). Ward and Cohan (2005) noted the correspondence between these evolutionary and ecological patterns and the predictions of the Stable Ecotype Model (Cohan and Koeppel, 2008; Cohan and Perry, 2007), which was used in this study. In the Stable Ecotype model, "ecotypes are formed rarely, and there is recurrent purging of diversity within ecotypes, either by periodic selection (PS) or genetic drift" (Connor et al., 2010). After a PS event, ecotypes continue to evolve accumulating sequence diversity along their genomes. Specifically, the model predicts that selection periodically favors a most-fit variant or variants to the exclusion of less-fit variants within the ecotype. Selective sweeps or purging of diversity via a PS event can occur within an ecotype population as environmental pressure acts upon it. A series of PS events provides a cohesive force holding together the individuals of an

ecotype population. Occasionally, a new variant may arise within an ecotype that has the ability to occupy a new niche and founds a new ecotype (ecotype formation, EF) (Cohan and Perry, 2007). The new ecotype population undergoes its own private PS events and is impervious to PS events that affect the parent ecotype population. After diverging, populations at the present time occur as genetically distinct ecotypes adapted to different niches, as ecologically neutral mutations and gene transfer events accumulate uniquely in the separate lineages. A more precise definition of ecotype according to this model is, "an ecologically distinct group whose diversity is limited by a force of cohesion, usually the genome-wide purging of diversity known as periodic selection but also genetic drift" (Koeppel et al., 2008). It assumes that ecotypes are distinct from one another and that there is ecological homogeneity of the members of an ecotype. By this definition, ecotypes are equivalent to ecological species.

A Monte-Carlo evolutionary simulation called Ecotype Simulation (ES) was developed to predict the number of ecotypes within a given clade (a group of organisms sharing a common ancestor) in a given phylogeny and to identify the members of each ecotype population (Koeppel et al., 2008; Cohan and Perry, 2007). Putative ecotypes (PEs) are predicted by ES by first estimating the number of ecotypes (n), omega (EF) and sigma (PS) using the whole set of sequences in an alignment, and then, individual ecotypes are demarcated using an auto- or manual-demarcation tool. Since *Synechococcus* populations in the mats we studied occur in large effective population sizes, drift is negligible. The PEs predicted by ES for *Bacillus* populations inhabiting desert soils (Koeppel et al., 2008; Connor et al. 2010) and hot spring *Synechococcus*

populations based on 16S rRNA/ITS sequence variation (Ward et al., 2006) have been shown to exhibit unique ecological distributions, validating their ecological distinctness.

The existence of phenotypically distinct *Synechococcus* populations at different depths within the upper 1-mm depth interval of these hot spring mats suggested possible adaptations to light and/or other parameters that may vary in relation to the vertical aspect in the mat (Ramsing et. al., 2000; Ward et al., 2006). At all temperatures along the thermal gradient *Synechococcus* cells inhabiting the mat surface autoflouresce less intensely than those at greater depths, where light intensity is much lower and light quality differs dramatically (Ward et al., 2006). At 60°C, the 16S rRNA genotype B´ was the only *Synechococcus* genotype detected in the surface layer, whereas 16S rRNA genotype A was only detected at depths corresponding to the deeper *Synechococcus* population (Ramsing et al., 2000). At 68°C only the A´ *Synechococcus* 16S rRNA genotype occurs at the surface and subsurface, raising the question of whether these phenotypically distinct A´ populations correspond to one differently acclimated (physiologically optimized) population or two differently adapted (evolutionarily optimized) populations that are too closely related to be detected using 16S rRNA sequence variation (Ferris et al., 2003). Using the more rapidly evolving 16S-23S rRNA-internal transcribed spacer (ITS) region it was observed that the A´ surface and subsurface populations at 68°C were genetically distinct (i.e., different ecotypes). At some temperatures between 60°C and 68°C only the A genotype occurs, but no variation was observed in either 16S rRNA or ITS sequences over the depths containing *Synechococcus* cells with distinct autofluorescence intensities (Ward et al., 2006). This

suggested the need for yet higher molecular resolution to distinguish acclimated from adapted populations and thereby to ensure detection of all the ecologically distinct *Synechococcus* populations in the mat.

We hypothesized that there are more *Synechococcus* ecotypes than those revealed by 16S rRNA-ITS variation alone. Previous single and multi-locus studies had suggested that protein-encoding loci are less conserved than the 16S rRNA/ITS region and therefore may yield higher resolution phylogenies for closely related populations of organisms (Koeppel et al., 2008; Papke et al., 2007; Hanage et al., 2006; Whitaker and Banfield, 2005). In this study we investigated three protein-encoding loci, *rbsK* (encoding a ribokinase), *aroA* (encoding the 3-phosphoshikimate 1-carboxyvinyltransferase), and *apcAB* (encoding the allophycocyanin alpha and beta subunits, including the internal transcribed spacer region separating these subunit-encoding genes) for comparison with 16S rRNA and 16S-23S rRNA/ITS variation. Variation at these loci was sampled from mat DNA using PCR amplification, cloning and sequencing. Ecotypes were evaluated using phylogenetic analysis and ES.

## Methodology

### Study Sites

We studied two temperature-defined sites, 60°C and 65°C in the effluent channel of Mushroom Spring (44.5386°N, 110.7979°W) (termed M60 and M65, respectively), an alkaline siliceous hot spring in the Lower Geyser Basin, Yellowstone National Park, Wyoming, USA (Ramsing et al., 2000). This allowed us to exploit genomes for

*Synechococcus* strains A (strain JA-3-3Ab, referred to as the A genome) and B´ (strain JA-2-3B a(2-13), referred to as the B´ genome) both isolated from Octopus Spring, Yellowstone National Park, Wyoming (Allewalt et al., 2006; Bhaya et al., 2007) and metagenomes that were prepared from the same samples (Klatt et al., 2010).

Sample Collection

Samples were collected on 2 October 2003 using a no. 4 cork borer. The top green layer of the mat (~ 1 mm-thick) was separated from core samples using a clean razor blade as previously described (Ramsing et al. 2000). Samples were transported to the laboratory at Montana State University on dry ice and kept frozen at -80°C until DNA extraction and subsequent PCR amplification and cloning.

DNA Extraction and Purification

The top green layers of 3 core samples were combined in a 2 ml tube and cells were lysed using the protocol previously used to obtain DNA from *Synechococcus* strain A and B´ genomes, mat metagenomes and 16S rRNA/ITS sequences (Bhaya et al., 2007; Klatt et al., 2010; Ward et al., 2006). Samples were homogenized by bead beating in a FP120 Fastprep Cell Disrupter (Bio101 Savant Instruments, New York) for 40 seconds at a setting of 6.5 beats per second. Nucleic acids were extracted with buffered phenol and chloroform mixtures followed by ethanol precipitation overnight at -20°C, as previously described (Ferris et al., 1996a; Ferris and Ward, 1997). DNA concentration was estimated using agarose gel electrophoresis analysis with high molecular weight DNA ladders (Invitrogen).

Locus Selection

This study was linked to a cultivation-independent, multi-locus sequence analysis (MLSA) study based on bacterial artificial chromosome cloning of 100-130 kb genome segments containing 16S rRNA genes and surrounding protein-encoding loci (Melendrez et al., 2010a/b; Chapters 3 and 4). Hence, one hundred genes in both directions from the two 16S rRNA loci in the *Synechococcus* strains A and B´ genomes and metagenomic homologs of these genes were analyzed seeking protein-encoding genes with the following preferred characteristics: (i) presence in both genomes, (ii) proximity near but not adjacent to the 16S rRNA locus (iii) high degree of nucleotide divergence between *Synechococcus* strain A and B´ homologs, (iv) high average nucleotide divergence and variance of metagenomic homologs from *Synechococcus* strain A and B´ homologs, (v) not under positive evolutionary selection, (vi) not flanked by transposons and (vii) functionally useful (may be highly expressed or niche defining). Nucleotide divergence between *Synechococcus* strain A and B´ homologs, or between metagenomic sequences and the homolog in the more closely related of the two reference genomes (Figure A2.1) was determined by finding the reciprocal best BLAST matches in the National Center for Biotechnology Information nucleotide collection (NCBI, BLAST-nr/nt) sequence database (default values) (Altschul et. al, 1990). The degree of evolutionary selection pressure was assayed by computing dN/dS values using the method of Nei and Gojobori (1986) implemented in the DnaSP program package (Librado and Rozas, 2009). Genes with dN/dS values greater than 1, indicative of positive selection were not considered. We attempted to avoid loci adjacent to transposons to decrease the probability that they

had recently undergone transfer from some other organism by highly mobile elements and thus might exhibit a different phylogeny than other genes or the organism (Figure 2.1). For this study, *rbsK*, *aroA*, and *apcAB* were selected as genes that satisfied the criteria outlined above (Table 2.1). The *rbsK* and *aroA* are housekeeping genes and *apcAB* is important in photosynthesis and thus potentially useful in expression studies.

PCR Amplification and Cloning

Primers for the separate amplification of *Synechococcus* A-like and B´-like population target genes were obtained from Integrated DNA Technologies. The primers were designed using the primer design tool in SciTools on the Integrated DNA Technologies website (http://www.idtdna.com/Scitools/Applications/Primerquest/) and analyzed by BLAST-nr/nt to assure specificity to *Synechococcus* A-like or B´-like genomic homologs (Table 2.1). All primers were dissolved in water to a final concentration of 50 μM for use in PCR amplification. For the *apcAB*, *aroA* and *rbsK* genes PCR cycling conditions were as follows: initial denaturing step of 94°C for 2 minutes (10 min for *apcAB*), followed by 30 cycles (33 cycles for *apcAB*) of 94°C (1 min), 55°C (1 min) [60°C (1 min, -0.5°C/cycle) for *apcAB*], and 72°C (1 min), with a final extension at 72°C for 10 min followed by storage at 4°C. After verifying sizes of

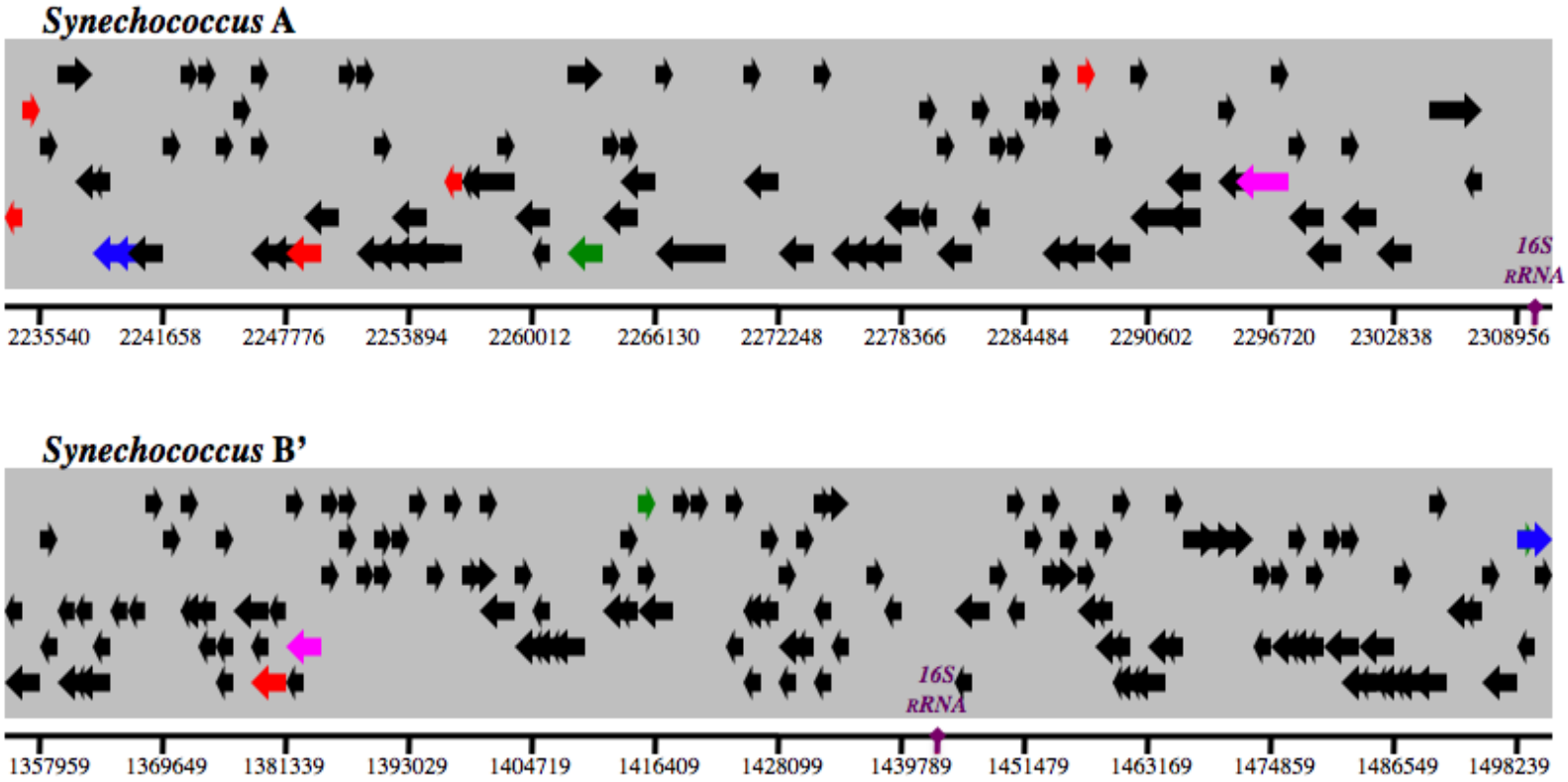



Figure 2.1. Genomic regions of the 16S rRNA genes being studied in *Synechococcus* strain A (top) and B´ (bottom) containing the *apcAB* (blue), *rbsK* (green) and *aroA* (pink) loci. Transposons are color coded in red. Black arrows represent other genes not examined in this study within the 16S rRNA region analyzed.

Table 2.1. Characteristics and primer sequences for protein-encoding loci.

| Locus | Primer | Sequence (5′-3′) | % Divergence of *Synechococcus* A and B′ homologs | Number of metagenomic sequences analyzed A/B′ | % Divergence (Variance) of Metagenomic A-like *Synechococcus* homologs | % Divergence (Variance) of Metagenomic B′-like *Synechococcus* homologs | $d_N/d_S$ for *Synechococcus* A metagenomic homologs | $d_N/d_S$ for *Synechococcus* B′ metagenomic homologs |
|---|---|---|---|---|---|---|---|---|
| *apcAB*[a] | apcABF[d,e] | ttaacttctccatggccgagctgt | 8.5 | 7/11 | 0.04 (0.1) | 1.84 (2.06) | 0.01 | 0.1 |
| | apcABR[d,e] | aagctgccgatcaactgttccaga | | | | | | |
| *aroA*[b] | aroA68F[d] | tgagcattccaggcgataagtcca | 15.1 | 6/10 | 0.3 (0.05) | 3.47 (2.01) | 0.08 | 1.1 |
| | aroA809R[d] | agacccacatcctgtagcagcaat | | | | | | |
| | aroA77F[e] | ccggcgacaaatccatttcccacc | | | | | | |
| | aroA809R[e] | aggcccacatcccttagcagcaat | | | | | | |
| *rbsK*[c] | rbsK34F[d] | aatctggacttggtggtgcaggta | 18.2 | 3/14 | [f] | 3.06 (3.98) | 0.82[g] | 0.7 |
| | rbsK645R[d] | gatgatcactgtcgggatcccctg | | | | | | |
| | rbsK34F[e] | aacatggatttggtggtgcaggtg | | | | | | |
| | rbsK730R[e] | tggtatccaccacttccacaggat | | | | | | |

[a] Role category: Energy metabolism: photosynthesis.
[b] Role category: Amino acid biosynthesis: aromatic amino acid family.
[c] Role category: Energy metabolism: sugars.
[d] Primers used for amplification of A-like *Synechococcus*.
[e] Primers used for amplification of B′-like *Synechococcus*.
[f] Not enough metagenomic sequence data to calculate average divergence and variance.
[g] Based on analysis of 3 metagenomic *rbsK* sequences.

PCR products using agarose gel electrophoresis analysis, products were purified using a MiniElute 96 UF PCR purification kit (Qiagen) or a Qiaquick PCR purification kit (Qiagen) according to the manufacturer's instructions. Cloning of purified PCR products was done using a TOPO-TA Cloning Kit for Sequencing (Invitrogen) according to the manufacturer's instructions. Briefly, DNA was ligated to the cloning vector by adding 2 μl purified PCR product to the TOPO vector, salt solution (1.2 M NaCl and 0.06 M $MgCl_2$), and Sigma water (Sigma) and the suspension was mixed by tapping. This ligation mix was incubated for 30 min. at room temperature, placed on ice overnight, and used to transform OneShot Mach-T1 chemically competent *Escherichia coli* cells (Invitrogen). Transformed cells were spread on LB-Amp plates (Fisher Scientific, 0.1 μg/μl ampicillin) and incubated overnight. Transformant colonies were picked and sizes of the inserts were analyzed using PCR amplification with the M13F/R primers (M13F: 5′-gtaaaacgacggccag-3′, M13R: 5′-caggaaacagctatgac-3′) or T3/T7 primers (T3: 5′-aattaaccctcactaaaggg-3′, T7: 5′-taatacgactcactataggg-3′) specific to the cloning vector, and agarose gel electrophoresis.

Sequencing

Clones with correctly sized inserts were sequenced in both directions using the M13F/R or T3/T7 primer and the BigDye v.3.1 cycle sequencing kit (Applied Biosystems) and analyzed at the University of Nevada-Reno Sequence Center (Reno, NV). The sequences will be submitted to GenBank. Sequence data for 16S rRNA and ITS were previously obtained from the same and other samples by PCR amplification

(Ward et al., 2006) and are available from Genbank under accession numbers DQ979027-DQ979319 (collected October 2, 2003 from Mushroom Spring 60°C and 65°C sites).

Sequence Alignment and Phylogenetic Analysis

Sequence data were analyzed using Sequencher v.4.8 and alignments were made using ClustalX software with *Synechococcus* strain A and B´ genomic homologs as references in the alignments. Sequences were specified to belong to *Synechococcus* A-like or B´-like populations based on a reciprocal best BLAST match to one of the reference genomes using the BLAST-nr/nt database and default parameters. Alignments were manually verified and neighbor-joining trees were constructed using MEGA v4.0 software (Tamura et al., 2007) and viewed using NJPlot (Perriere and Gouy, 1996). Estimates of average evolutionary divergence (EED) were computed over all sequence pairs in the phylogeny using the Maximum Composite Likelihood method in MEGA. All codon positions were included in the analysis, even those positions that were non-coding. All positions containing gaps and missing data were eliminated from the dataset (complete deletion option) for consistency of sequence data input across all analyses.

Samples from the 65°C site had previously been shown to contain both *Synechococcus* A- and A´-like population sequences (Ward et al., 2006). To evaluate whether sequences from these two populations could be separated we examined the percent nucleotide identity distributions of sequences recovered by PCR relative to homologs in the *Synechococcus* strain A genome using a WU-BLAST analysis (default parameters). Since we do not have a reference genome for *Synechococcus* strain A´, we used as a proxy a metagenomic database obtained using Ti454 sequencing from a

Mushroom Spring 68°C sample (unpublished results of Becraft, Klatt, Rusch and Ward), a site that is dominated by *Synechococcus* genotype A′ (Klatt et al., 2010). Metagenomic sequences in the 68°C library that had a top WU-BLAST match to the *Synechococcus* strain A *apcAB*, *aroA* or *rbsK* genes and that matched >50% of the length of these genes were considered homologs in the *Synechococcus* A′-like clade.

Ecotype Simulation and Demarcation

Sequence alignments were analyzed using ES to predict the number of PEs (n), rates of PS, EF and 95% confidence intervals (CI) for all parameters at 1.5x precision match between observed and simulated data (http://fcohan.web.wesleyan.edu/ ecosim/). The n value generated in the ES analysis is a rough estimation of the number of PEs. This value is generally higher than demarcated PE values because PE values obtained from manual demarcation are conservative (see below). Neighbor-joining trees were constructed and uploaded into ES as Newick files for demarcation analysis. PEs were demarcated conservatively by the ES manual demarcation program. Sequences from the largest clades were analyzed and if the 95% CI included 1, then the predicted PE was considered a single ecotype even if the most likely number of ecotypes was greater than 1. If ES predicted the clade to contain more than one ecotype (i.e., 1 was not included in the 95% CI), then sequences were removed systematically starting from the root of the entire phylogeny (closest to the common ancestor) until clades were identified in which the confidence interval included 1 (also see Chapter 1, Figure 1.10).

Chao Estimation of Diversity

Collector's curves and Chao estimates ($S_c$) of species or operational taxonomic unit (OTU) richness were generated for full clone library sequence data and subsets of the data using the software package Mothur v. 1.4.0 (http://schloss.micro.umass.edu/mothur/) (Schloss and Handelsman, 2005 and 2006; Schloss et al., 2004). $S_c$ is a lower-bound estimate of total species richness, including those to rare to be observed in the dataset (Chao, 1984). Two $S_c$ estimates were obtained. The first estimate, $S_{c\text{-}nt}$, was obtained from aligned nucleotide sequence data from all sequences available from each PCR clone library, which were grouped into OTUs that included sequences that are ≥ 99% identical. Collector's curves were generated from the data and an $S_{c\text{-}nt}$ estimate was calculated for 71 sequences (rationale given below) and a binning percent nucleotide identity cutoff of ≥ 99%. The 95% confidence intervals were also calculated from the nucleotide dataset for the $S_{c\text{-}nt}$ value. The second estimate ($S_{c\text{-}pe}$) was calculated using equation 1 and demarcated PE phylogenies generated from ES, where $S_{obs}$ is the observed number of PEs, $n_1$ is the number of PE clades that contain only 1 sequence (singleton), and $n_2$ is the number of PE clades that contain 2 sequences (doubletons) (Chao and Shen, 2003; Chao, 1984).

Equation 1. $$S_{c\text{-}pe} = S_{obs} + n_1(n_1\text{-}1)/[2(n_2+1)]$$

# Results

## Locus Selection and Characteristics

All genes selected for the study exist as one copy in the genomes of *Synechococcus* strains A and B´, occur within 70 kb of the same 16S rRNA gene (Figure 2.1) and are under purifying selection (i.e., $d_N/d_S<1$) (Table 2.1). Percent nucleotide divergence between homologs in the A and B´ genomes ranged between 8.5% and 18.2% (Table 2.1). Average divergences of metagenomic sequences [those bounded by the green and red boxes in Figure A2.1] from homologs in the more closely related reference genome ranged from 0.04% to 0.1% for A-like and 1.8% to 3.5% for B´-like homologs and variance ranged from 0.05 to 0.1 for A-like and 2.01-3.98 for B´-like homologs. Given the low number of metagenomic sequences that this analysis is based on, these were considered rough estimates (Table 2.1, Figure A2.1).

## Clone Library Composition

PCR clone libraries contained 67 to 309 sequenced clones (Table 2.2). *Synechococcus* strain B´-like sequences were recovered only from the M60 sample, whereas *Synechococcus* strain A-like sequences were found in both M60 and M65 samples for all loci, as expected from previous analyses (Ward et al., 2006). Frequency distributions of percent nucleotide divergence comparing *Synechococcus* A-like homologs retrieved from M60 or M65 and homologs in the 68°C metagenome (A´-like proxy) to A genome homologs revealed a low frequency of A´-like sequences in the *aroA* and *apcAB* PCR libraries (Figure A2.2). The *aroA* and *apcAB* genes (Figure A2.2)

showed clear separation of A-like and A´-like homologs. Only one sequence of each locus, which appeared to be more A´-like than A-like was removed from the *apcAB* and *aroA* sequence datasets before further analysis. This was not surprising because a low frequency of A´-like sequences in the M60 and M65 samples had previously been seen with 16S rRNA and ITS sequences from these samples (Ward et al., 2006). In contrast, *rbsK* homologs retreived from M65 by PCR and from the 68°C metagenome showed a similar level of divergence from the *Synechococcus* strain A *rbsK* gene. Thus, no sequences could be confidently assigned as A´-like and none were removed from subsequent analysis. Interestingly, the A-like homologs from the M60 and M65 samples showed different percent nucleotide identity (% nt identity) distributions, suggesting that

Table 2.2. PCR clone library composition

| **Sample** | **Gene** | **No. of Clones in Library** | **% A-like *Synechococcus*** | **% B´-like *Synechococcus*** |
|---|---|---|---|---|
| M60 | *apcAB* | 309 | 34 | 66 |
| M65 | | 117 | 100 | 0 |
| M60 | *aroA* | 423 | 60 | 40 |
| M65 | | 107 | 100 | 0 |
| M60 | *rbsK* | 257 | 35 | 65 |
| M65 | | 67 | 100 | 0 |

these samples contain genetically distinct A-like *rbsK* homologs with those in M65 being more closely related to the *Synechococcus* strain A homolog (Figure A2.2).

Effect of Sampling and
Molecular Resolution on Detection of Diversity

Collector's curve analyses (Figure 2.2, Table 2.3) revealed how sampling affected measurements of genetic diversity (OTUs). Normalization for number of sequences was necessary for comparative analyses of different loci. As a result, we restricted analyses of

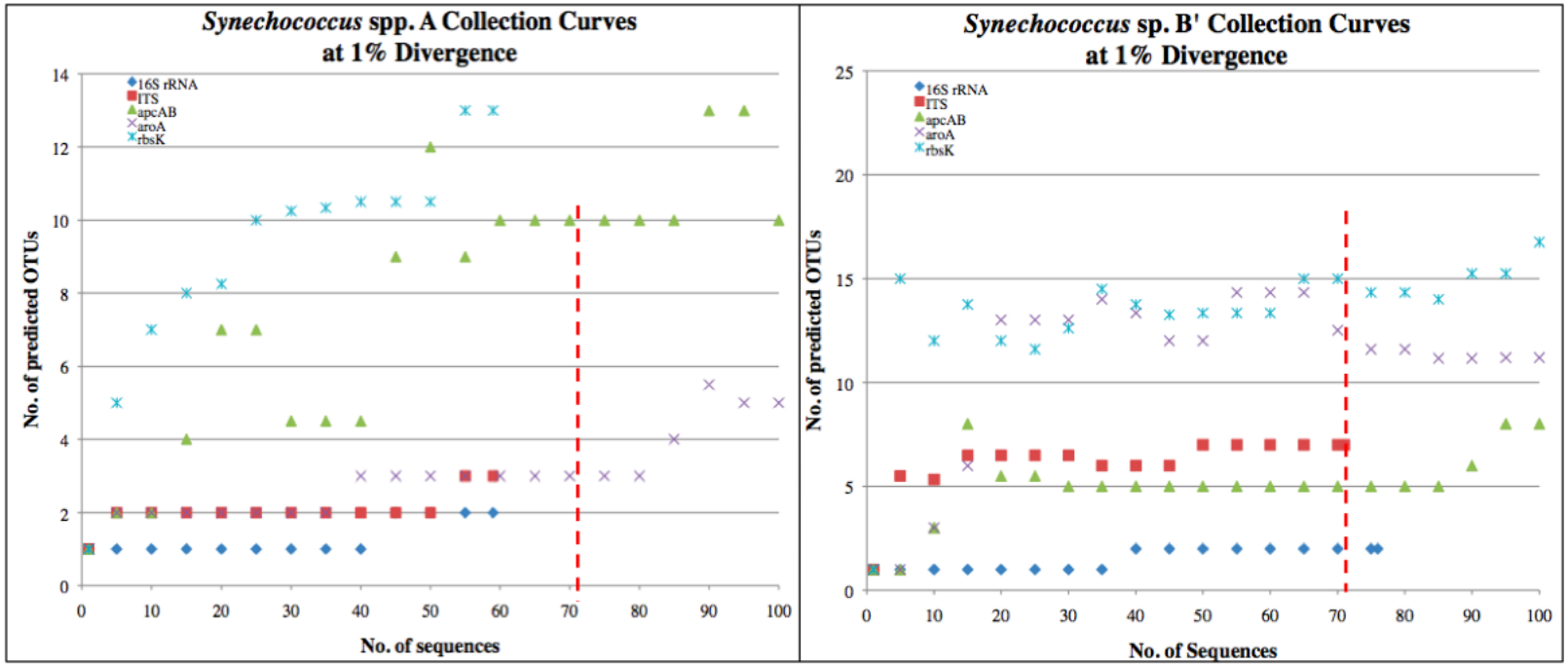



Figure 2.2. Collector's curves for *Synechococcus* A-like and B´-like population sequences for all loci comparing number of OTUs, defined by grouping all sequences with ≥99% nt identity, with the number of sequences sampled. Red dashed lines are set at 71 sequences.

Table 2.3. Comparison of Chao estimates of OTU (99% cutoff) diversity and ecotype diversity predicted from Ecotype Simulation (ES) analysis of ~70 variants and 3 loci of *Synechococcus* A-like and B´-like populations.

| **Sample** | ***Synechococcus* Type** | **Gene** | **No. of Variants** | **EED (%)** | **Chao Richness Estimates** | | **Ecotype Simulation** | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | $S_{c\text{-}pe}$ [a] | $S_{c\text{-}nt}$ (95% CI) | PE prediction –n parameter (95% CI) | Demarcated PEs [b] | Omega (95% CI) | Sigma (95% CI) |
| M60 +M65 | A | 16S rRNA | 70 | 0.002 | 1 | 1 (1-1) | 58 (2-71) | 1 | 5.42 (2.5-18.8) | 3.39 (0.05- >100) |
| | | ITS | 70 | 0.002 | 2 | 3 (0-3) | 3 (2-10) | 2 | 0.136 (0.01-0.45) | 4.04 (0.58-20.1) |
| | | *apcAB* | 73 | 0.016 | 5.5 | 7.8 (7-30) | 11 (3-11) | 5 | 0.035 (0.01-0.12) | 1.82 (0.01->100) |
| | | *aroA* | 73 | 0.013 | 4 | 3.5 (0-3) | 5 (2-30) | 4 | 0.373 (0.17-1.2) | 143 (0.64->100) |
| | | *rbsK* | 72 | 0.095 | 14.3 | 13 (12-23) | 15 (8-30) | 14 | 0.043 (0.02-0.09) | 3.76 (0.25->100) |
| M60 | B´ | 16S rRNA | 73 | 0.002 | 1 | 2 (2-2) | 7 (2-75) | 2 | 0.519 (0.01-13.2) | 5.48 (0.25->100) |
| | | ITS | 70 | 0.01 | 2 | 7 (0-7) | 51 (3-70) | 2 | 1.00 (0.68-2.3) | 1.60 (0.03->100) |
| | | *apcAB* | 67 | 0.01 | 11 | 7 (10-35) | 32 (9-70) | 8 | 0.314 (0.1-0.68) | 1.09 (0.02->100) |
| | | *aroA* | 70 | 0.03 | 15 | 12.5 (11-23) | 32 (3-70) | 13 | 0.242 (0.11-0.81) | 23.8 (0.01->100) |
| | | *rbsK* | 68 | 0.05 | 14.5 | 15 (14-25) | 20 (11-46) | 13 | 0.092 (0.04-0.14) | 1.8 (0.12->100) |

[a] Calculated based on equation 1, no 95% confidence interval, discrete value only.
[b] No 95% confidence interval for demarcated PE value, see n parameter confidence interval values.

each lineage to the largest number of sequences available for all loci, as shown in Table 2.3, which allowed for comparison of an equal number of variants of B´-like and A-like variants (~70 sequences per lineage). Table 2.3 also shows richness estimates ($S_{c\text{-}nt}$), which ranged from 1-13 and 1-15 OTUs for A-like and B´-like populations, respectively for the 16S rRNA locus, ITS region, *apcAB*, *aroA*, and *rbsK* loci. Since the Chao estimate ($S_{c\text{-}nt}$) has been proposed as a lower bound (conservative) richness estimator (Colwell and Coddington, 1994), the value of $S_{c\text{-}nt}$ estimated using ~70 sequences could be considered an underestimation of richness within the environmental sample. In these restricted subsets, the depth of divergence varied with the locus and *Synechococcus* lineage analyzed and the numbers of OTUs increased as the EEDs of the gene from which they were estimated increased (Table 2.3). The relative divergences of the different loci predicted by the divergence of homologs in the A and B´ genomes was greater than the divergence of metagenomic homologs, and metagenomic homolog divergence was greater than that measured by EED within the different lineages (Tables 2.1 and 2.3). Although the *rbsK* gene had the greatest resolution in both populations according to EED (Table 2.3) the divergence of *aroA* in the *Synechococcus* A-like population was lower than expected based on estimates obtained from A/B´ homologs and metagenomic homologs. However, the order of loci from least resolution to greatest resolution was different between the two populations. For A-like *Synechococcus* the *apcAB* and *aroA* loci had the least resolution (EED: 0.016 and 0.013) and the *rbsK* gene had the greatest resolution (0.095). For B´-like *Synechococcus*, the *apcAB* locus had the least resolution (EED: 0.1) followed by *aroA* (EED: 0.3) then *rbsK* (EED: 0.5) (Table 2.3). These values

were generated from nucleotide alignments of ~70 sequences for each gene for each organism and differ from the values obtained in analysis of metagenomic sequences (Table 2.1). Analysis of metagenomic sequences showed the *aroA* locus as having the greatest resolution in B´-like *Synechococcus* population, however, it is important to note the limited number of sequences in the metagenomic analysis and this may explain the discrepancy in gene order of increasing resolution. The A-like population appeared to be more diverse than the B´-like population at the *rbsK* locus (0.095 in A/A´ vs. 0.05 in B´). This may be due to presence of A´-like sequences in the sample because, although one outlier sequence was removed from the analysis (Figure A2.2), we were unable to discern A-like and A´-like sequences.

Putative Ecotype Demarcations Across Loci and Lineages

ES analysis was conducted on the same subsets of sequence data for all loci in both A-like and B´-like lineages (Table 2.3 and Figures 2.3-2.5). In general, well-sampled PEs contained a dominant sequence variant (shaded yellow in Figures 2.4 and 2.5), which we interpret as a possible founder of the PE clade. A-like *Synechococcus aroA* PE-1 (Figure 2.4) and B´-like *Synechococcus apcAB* PE-1 (Figure 2.5) contained two dominant allelic variants (shaded gray in Figures 2.4 and 2.5). The number of demarcated PEs clustered toward the lower end of the 95% CIs for n predicted by ES (Table 2.3). In general, predicted PS rates were higher than predicted EF rates, suggesting that EF events are relatively rare and that putative ecotypes are constrained by PS. This is consistent with the stable ecotype model previously discussed. The average percent nucleotide identity among variants within each PE varied from 0 (all identical) to 2.7%

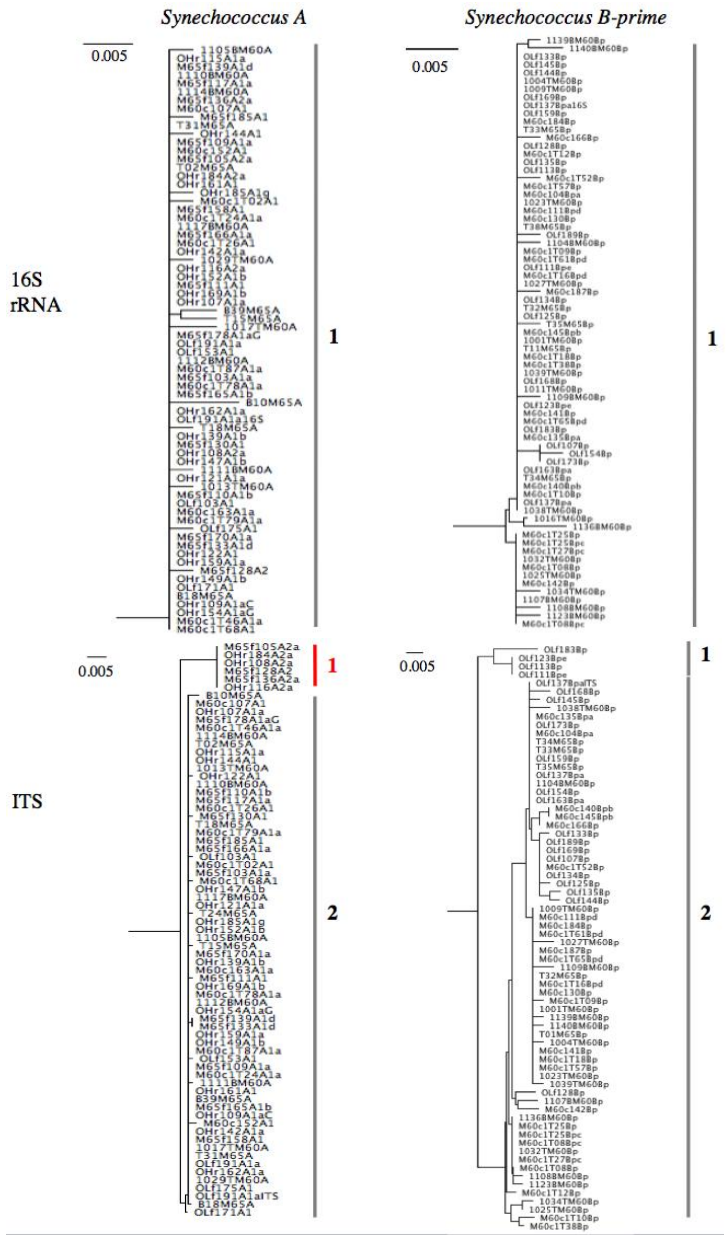



Figure 2.3. Neighbor-joining phylogenetic trees of A-like and B´-like *Synechococcus* diversity in Mushroom Spring at 60°C and 65°C (*Synechococcus* B'-like only observed at 60°C site) for the 16S rRNA gene and internal transcribed spacer region (ITS) with putative ecotypes demarcated by Ecotype Simulation. Numbers represent the names of putative ecotypes. Ecotype highlighted in red contains only clones from Mushroom Spring 65°C or Octopus Spring at ~65°C.

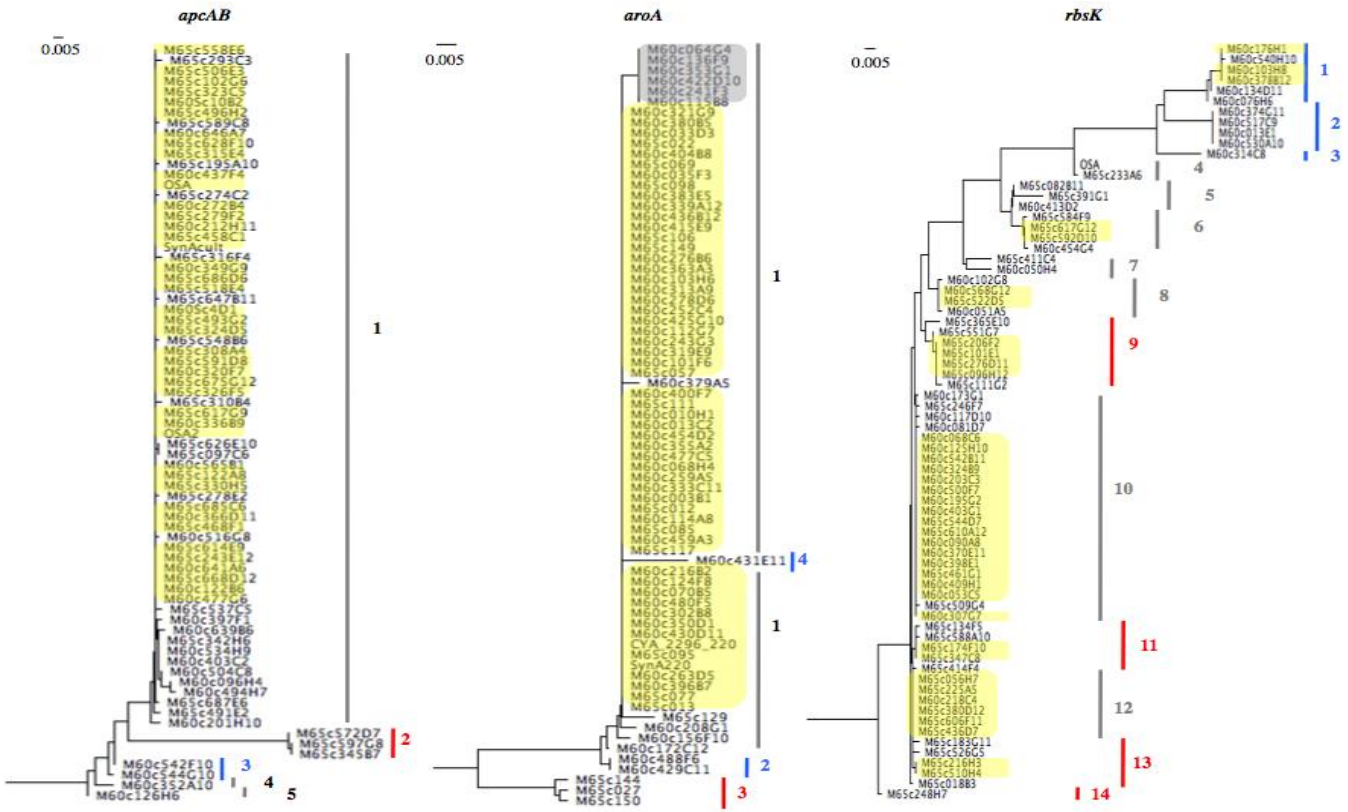



Figure 2.4. Neighbor-joining phylogenies for *Synechococcus* A-like population sequences for *apcAB, aroA* and *rbsK* genes with putative ecotypes demarcated by Ecotype Simulation. Blue and red numbers and bars demarcate putative ecotypes specific to the 60 and 65°C samples, respectively, while those in black demarcate putative ecotypes that contain sequences from both temperature sites. Replicate sequences of the dominant variant within a demarcated putative ecotype are highlighted in yellow boxes. Putative subdominant variants are highlighted in grey boxes.

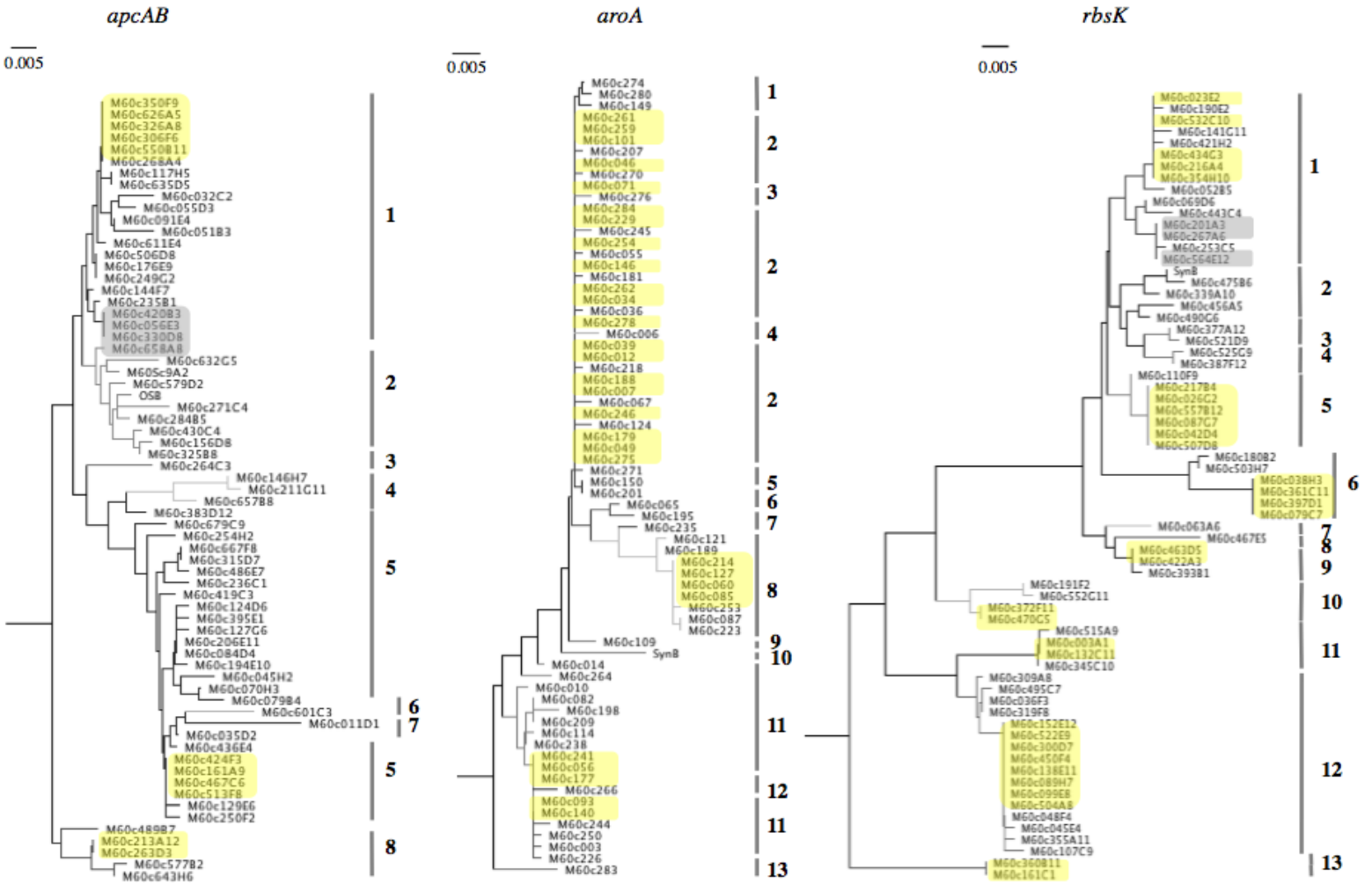



Figure 2.5. Neighbor-joining phylogenies for *Synechococcus* B´-like population sequences for *apcAB, aroA* and *rbsK* genes with putative ecotypes demarcated by Ecotype Simulation. Replicate sequences of the dominant variant are highlighted in yellow boxes for each demarcated ecotype. Putative subdominant variants are highlighted in grey boxes.

Table 2.4. Percent sequence variation within PE clades that contain ≥ 2 sequences.

| Organism | Locus | PE | No. of Sequences in PE | Average % Sequence Variation | Maximum % Sequence Variation |
|---|---|---|---|---|---|
| A-like *Synechococcus* | *apcAB* | 1 | 66 | 0.4 | 2.9 |
| | | 2 | 3 | 0.2 | 1.0 |
| | | 3 | 2 | 1.0 | 1.0 |
| | *aroA* | 1 | 68 | 0.5 | 7.8 |
| | | 2 | 2 | [a] | [a] |
| | | 3 | 3 | 0.7 | 1.0 |
| | *rbsK* | 1 | 6 | 0.4 | 0.9 |
| | | 2 | 4 | [a] | [a] |
| | | 4 | 2 | 0.2 | 0.2 |
| | | 5 | 3 | 1.1 | 1.6 |
| | | 6 | 4 | 0.2 | 0.4 |
| | | 7 | 2 | 2.7 | 2.7 |
| | | 8 | 4 | 0.2 | 0.4 |
| | | 9 | 7 | 0.7 | 1.7 |
| | | 10 | 24 | 0.1 | 0.4 |
| | | 11 | 4 | 0.2 | 0.4 |
| | | 12 | 7 | 0.1 | 0.2 |
| | | 13 | 5 | 0.5 | 0.7 |
| B'-like *Synechococcus* | *apcAB* | 1 | 22 | 1.1 | 2.9 |
| | | 2 | 9 | 1.9 | 4.0 |
| | | 4 | 3 | 2.1 | 3.1 |
| | | 5 | 25 | 1.7 | 3.8 |
| | | 8 | 5 | 1.3 | 2.6 |
| | *aroA* | 1 | 3 | 1.1 | 1.2 |
| | | 2 | 11 | 0.5 | 1.2 |
| | | 5 | 2 | 0.2 | 0.2 |
| | | 6 | 2 | 0.8 | 0.8 |
| | | 8 | 9 | 0.5 | 1.2 |
| | | 11 | 17 | 0.8 | 2.2 |
| | *rbsK* | 1 | 15 | 0.9 | 1.8 |
| | | 2 | 5 | 1.3 | 1.6 |
| | | 3 | 2 | 0.2 | 0.2 |
| | | 4 | 2 | 0.2 | 0.2 |
| | | 5 | 7 | 0.3 | 0.5 |
| | | 6 | 6 | 0.9 | 1.8 |
| | | 9 | 3 | 0.2 | 0.2 |
| | | 10 | 4 | 1.1 | 1.6 |
| | | 11 | 4 | 0.1 | 0.2 |
| | | 12 | 16 | 0.4 | 1.0 |
| | | 13 | 2 | [a] | [a] |

[a] PE clade contained sequences that had 100% nucleotide identity and therefore no sequence variation.

and the maximum percent nucleotide variation within a PE clade ranged from 0.2% to 7.8% (Table 2.4).

ES predicted 1-2 demarcated PEs from 16S rRNA and ITS sequence variation (Figure 2.3) for *Synechococcus* A-like and B´-like sequences, with limited resolution of sample-specific PE clades, similar to the results of Ward et al. (2006). For *Synechococcus* A-like population diversity, ES predicted 4, 5 and 14 demarcated PEs from *apcAB*, *aroA* and *rbsK* variation, respectively (Figure 2.4). For *Synechococcus* B´-like population diversity, ES predicted 8, 13 and 13 demarcated PEs from *apcAB, aroA* and *rbsK*, respectively (Figure 2.5).

Many A-like PEs were specific to either the 60 or 65°C sample, suggesting that they constitute ecologically distinct populations. The Fisher's exact test suggested that there was significant heterogeneity across ecotypes in their habitat associations (Table 2.5) for A-like *Synechococcus* populations. Both the *aroA* and *rbsK* loci showed significant associations. For B´-like *Synechococcus* population diversity, ES predicted fewer demarcated PE's from *apcAB* variation (8) than from *aroA* or *rbsK* variation (both 13) (Figure 2.5). All B´-like sequences were retrieved from the 60°C sample and thus sample-specificity could not be observed. ES predictions of demarcated PEs and OTUs were, in general, similar to Chao estimates of OTUs (99% cutoff) (Table 2.3). Chao analysis of demarcated PEs yielded $S_{c\text{-}pe}$ estimates that were equal or somewhat higher than demarcated PE numbers, probably due to the restriction of these analyses to 71 variants.

## Discussion

The protein-encoding loci we studied offer higher molecular resolution than 16S rRNA and ITS sequences for prediction of genetic and ecological diversity of native *Synechococcus* populations inhabiting these mats. Between 6.5 and 14 times more PEs were predicted from the *rbsK* locus, which offers the greatest molecular resolution. Compared to predictions based on 16S rRNA and ITS sequence variation, ES-demarcated PEs for A-like *Synechococcus* populations are frequently sample-specific (Figure 2.4 and Table 2.5), suggesting that they are, in fact, ecologically distinct populations. Since we only studied samples from two temperature-defined sites along the effluent channel,

Table 2.5. Fishers exact test for habitat associations for PE clades with >1 sequence for *Synechococcus* A-like population sequences.

| Gene | PE | No. of Sequences from M60 | No. of Sequences from M65 | Sample-specificity | Fishers Exact Test p-value |
|---|---|---|---|---|---|
| *apcAB* | 1 | 24 | 35 | none | 0.098 |
| | 2 | 0 | 3 | M65 | |
| | 3 | 2 | 0 | M60 | |
| *aroA* | 1 | 52 | 12 | none | 0.015 |
| | 2 | 2 | 0 | M60 | |
| | 3 | 0 | 3 | M65 | |
| *rbsK* | 1 | 6 | 0 | M60 | <0.001 |
| | 2 | 4 | 0 | M60 | |
| | 4 | 1 | 1 | none | |
| | 5 | 1 | 2 | none | |
| | 6 | 1 | 3 | none | |
| | 8 | 3 | 1 | none | |
| | 9 | 0 | 7 | M65 | |
| | 10 | 17 | 5 | none | |
| | 11 | 0 | 5 | M65 | |
| | 12 | 1 | 5 | none | |
| | 13 | 0 | 5 | M65 | |

further work will be required to understand the adaptations that might explain why different PEs occur in different samples and why more than one PE is present at the same temperature site. For example, some ecological parameters (or dimensions) other than temperature (Allewalt et al., 2006) that may help explain sample specificity and the co-occurrence of many different PEs at one site could include light intensity and quality (Ramsing et al., 2000), nutritional and other chemical differences (Bhaya et al., 2007) and biological associations (Cohan and Koeppel, 2008). The validity of PEs based on protein-encoding sequence variation is being tested through studies of distribution along flow and vertical dimensions, unique gene expression and population dynamics in response to environmental change (Becraft et al., 2010). PE prediction using ES, Chao calculations from PE data, and Chao calculations from direct clone library sequence data (with a 1% divergence cutoff) all correspond closely and fall within each other's 95% CIs (Table 2.3). The predictions however, were obtained from small samples (71 sequences per gene) and PEs were demarcated conservatively. Therefore, the PE predictions likely underestimate the actual number of ecotypes in the environment.

There have been many suggestions of 'cutoffs' for the demarcation of species-like clusters. Originally, a criterion of 70% DNA-DNA hybridization based on phenotypic clustering of bacterial strains was suggested (Wayne et al., 1987), which correlated roughly with 97-98% 16S rRNA gene homology (Goodfellow et al., 1997). This was later revised to 98.7-99%; (Stackenbrandt and Ebers, 2006) and then to ~94% average nucleotide identity of genomes (Venter et al., 2004; Konstantinidis and Tiedje 2005). However, these cutoffs simply reflect the amount of genetic diversity found within the

phenotypic clusters already identified as species in microbiology (Cohan, 2002). They "lump diverse populations that exhibit distinct species-like traits, leading to an underestimation of functionally important biodiversity" (Ward et al., 2008). The current study was an attempt to improve our current understanding of species in light of protein-encoding loci that provide higher resolution of ecologically meaningful population-level units (ecotypes) in nature. The 2-3% cutoff with the 16S rRNA locus would have suggested that *Synechococcus* genotypes A and B´ are from separate species, together with the closely related variants with which they form clades. However, each 16S rRNA variant was found to be uniquely distributed along a temperature gradient (Ferris and Ward, 1997), and corresponding isolates exhibited optimal growth at specific temperatures matching their *in situ* distribution (Allewalt et al., 2006). Based on evidence from 16S rRNA-ITS sequence variation (Ferris et al., 2003;Ward et al., 2006), we hypothesized that there are more *Synechococcus* ecotypes than those revealed by 16S rRNA-ITS variation alone and our results using less conservative protein-encoding loci support this hypothesis.

Konstantinidis and Tiedje (2005) suggested that organisms that share a particular ecology and >99% genome-wide nucleotide identity may be the biologically meaningful units in nature. Interestingly, our Chao estimates of diversity using a 99% sequence identity cutoff to define OTUs were similar to our demarcated PE estimates (Table 2.3). However, we do not favor molecular cutoffs, as different microorganisms evolve at different rates, species may be younger or older and sequence divergence does not equate to a biological notion of species (Ward et al., 2008). Indeed PEs predicted from the same

locus varied in terms of average percent nucleotide identity (Table 2.4). 25% of the PEs analyzed had an average percent nucleotide variation of ≥1% and 55% of the same PEs had a maximum percent nucleotide variation ≥ 1% among the sequences belonging to the same PE. In one instance the maximum percent nucleotide identity within a PE clade was 7.8% (Table 2.4). This illustrates the concern over using molecular cutoffs to demarcate species clusters. The specific percentages are for the loci we studied, which may not be representative of the entire genome, complicating direct comparison with genomic average nucleotide identity values.

One limitation of single-locus analysis of diversity is variation in molecular resolution of the locus being studied; as some loci evolve more slowly than others and it is difficult to discern which locus may best represent the evolutionary trajectory of the organism. The EED values obtained in this study ranged from 0.002 (16S rRNA locus) to 0.095 (*rbsK* locus). The relationship between molecular resolution and PE prediction (Figure 2.6) shows clearly that PE prediction is a function of the resolving power of the locus. Even the most rapidly evolving locus we examined may underestimate ecological diversity. Another major limitation of single-locus analysis is that it does not account for the extent of recombination or linkage between loci. The presence of recombination will cause us to occasionally misidentify a sequence as belonging to a different ecotype than the one it actually belongs to. Linkage would tie together the evolutionary histories of the loci being examined possibly altering the perception of the evolution of the organism.

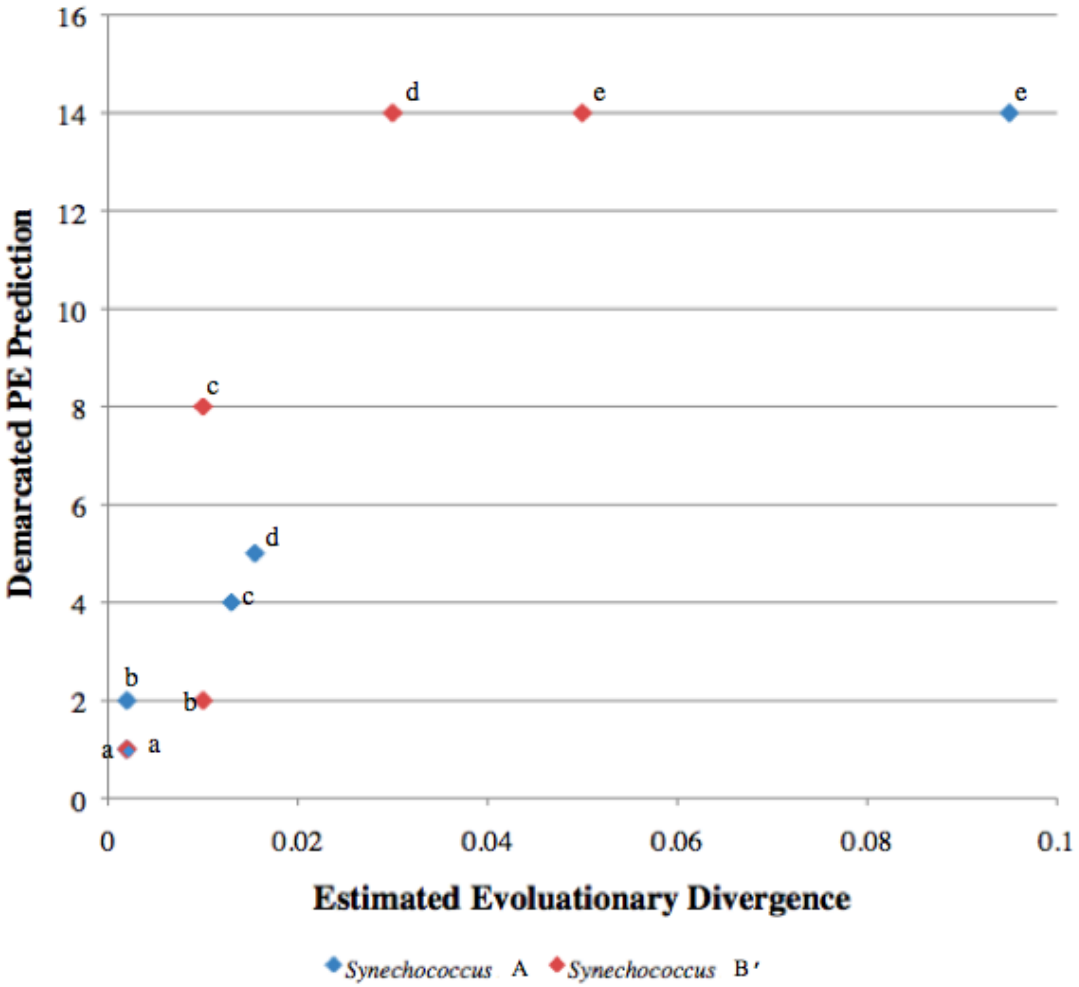



Figure 2.6. Demarcated putative ecotype prediction of Ecotype Simulation as a function of molecular resolution, as measured by estimated evolutionary divergence of 71 sequences for the (a) 16S rRNA, (b) ITS region, (c) *apcAB* for *Synechococcus* B´-like and *aroA* for *Synechococcus* A-like populations, (d) *aroA* for *Synechococcus* B´-like and *apcAB* for *Synechococcus* A-like variants, and (e) *rbsK* loci of A-like and B´-like *Synechococcus* populations.

To account for the different resolving powers of various loci and for recombination, multi-locus analyses have traditionally been implemented providing a more realistic evolutionary history of cultivated microorganisms (Feil et al., 2000, 2004; Salerno et al., 2007; Vitorino et al., 2008). In separate work, we have developed a cultivation-independent, multi-locus sequence analysis (MLSA) approach (Melendrez et al., 2010b; Chapter 4) that further increases molecular resolution and should reveal the importance of recombination and address molecular resolution and linkage within these *Synechococcus* populations. The single-locus data from this study will provide a reference for those studies.

## Acknowledgements

We appreciate support from the National Science Foundation Frontiers in Integrative Biology Research Program (EF-0328698), the National Aeronautics and Space Administration Exobiology Program (NAG5-8807) and the Pacific Northwest National Laboratory (contract pending), as well as the assistance from National Park Service personnel at Yellowstone National Park. We would also like to thank John Heidelberg for his assistance in metagenomic analyses that became important in locus selection, Alex Koeppel for his assistance in answering questions and troubleshooting protocols for Ecotype Simulation, and Jason Wood for metagenomic analysis assistance and optimization of Ecotype Simulation.

## References

Allewalt JP, Bateson MM, Revsbech NP, Slack K and Ward DM. (2006). Effect of temperature and light on growth of and photosynthesis by *Synechococcus* isolates typical of those predominating in Octopus Spring microbial mat community of Yellowstone National Park. *Appl Environ Microbiol* **72**: 544-550.

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403-410. (http://blast.ncbi.nlm.nih.gov).

Becraft ED, Cohan FM, Kühl M, Jensen S and Ward DM. (2010). Identifying and improving the existence of ecologically defined *Synechococcus* sp. in Mushroom Spring, Yellowstone National Park. In prep.

Bhaya D, Grossman AR, Steunou AS, Khuri N, Cohan FM, Hamamura N, Melendrez MC, Bateson MM, Ward DM and Heidelberg JF. (2007). Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analysis. *ISME J* **1**: 703-713.

Chao A and Shen T-J. (2003). Nonparametric estimation of shannon's index of diversity when there are unseen species in sample. *Environ Ecol Stat* 10: 429-443.

Chao A. (1984). Nonparametric estimation of the number of classes in a population. *Scand J Stat* **11**: 265-270.

Cohan FM and Koeppel AF. (2008). The origins of ecological diversity in prokaryotes. *Curr Biol* **18**: R1024-R1034.

Cohan FM and Perry EB. (2007). A systematics for discovering the fundamental units of bacterial diversity. *Curr Biol* **17**: R373-R386.

Cohan FM. (2002). What are bacterial species? *Annu Rev Microbiol* **54**: 457-487.

Colwell RK and Coddington JA. (1994). Estimating terrestrial biodiversity through extrapolation. *Phil Trans Roy Soc B* **345**: 101-118.

Connor N, Sikorski J, Rooney AP, Kopac S, Koeppel AF, Burger A, Cole SG, Perry EB, Krizanc D, Field NC, Slaton M and Cohan FM. (2010). The ecology of speciation in *Bacillus*. *Appl Environ Microbiol* Online: doi:10.1128/AEM.01988-09.

Feil EJ, Li BC, Aanensen DM, Hanage WP and Spratt BG. (2004). eBURST: Inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J Bacteriol* **186**: 1518-1530.

Feil EJ, Maynard Smith J, Enright MC and Spratt BG. (2000). Estimating recombination parameters in *Streptococcus pneumoniae* from multi-locus sequence typing data. *Genet* **154**: 1439-1450.

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

Ferris MJ and Ward DM. (1997). Season distributions of dominant 16S rRNA-defined populations in a hot spring microbial mat examined by denaturing gradient gel electrophoresis. *Appl Environ Microbiol* **63**: 1375-1381.

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

Ferris MJ, Ruff-Roberts AL, Kopczynski ED, Bateson MM and Ward DM. (1996b). Enrichment culture and microscopy conceal diverse thermophilic *Synechococcus* populations in a single hot spring microbial mat habitat. *Appl Environ Microbiol* **62**: 1045-1050.

Goodfellow M, Manfio GP and Chun J. (1997). Towards a practical species concept for cultivable bacteria. In: Claridge MF, Dawah HA and Wilson MR (eds). *Species: the units of biodiversity*. Chapman and Hall: London. Pp. 25-59.

Hanage WP, Fraser W and Spratt BG. (2006). Sequences, sequence clusters and bacterial species. *Phil Trans Roy Soc B* **361**: 1917-1927.

Klatt CG, Wood JM, Rusch DB, Bateson MM, Heidelberg JF, Bryant DA and Ward DM. (2010). Composition of metagenomes from a phototrophic hot spring microbial mat community. In prep.

Koeppel A, Perry EB, Sikorski J, Krizanc D, Warner A, Ward DM, Rooney AP, Brambilla E, Connor N, Ratcliff RM, Nevo E and Cohan FM. (2008). Identifying the fundamental units of bacterial diversity: A paradigm shift to incorporate ecology into bacterial systematics. *Proc Natl Acad Sci* **105**: 2504-2509.

Konstantinidis KT and Tiedje JM. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci* **102**: 2567-2572.

Librado P and Rozas J. (2009). DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**: 1451-1452.

Melendrez MC, Lange RK, Cohan FM and Ward DM. (2010a). Ecological diversity of *Synechococcus* spp. inhabiting an alkaline siliceous hot spring in Yellowstone National Park, WY measured using protein-encoding genes and evolutionary simulation. *ISME J* In prep.

Melendrez MC, Wood JM, Rusch DB, Heidelberg JF and Ward DM. (2010b). Bacterial artificial chromosome (BAC) libraries for Mushroom Spring cyanobacterial mat, Yellowstone National Park, WY. In prep.

Miller SR and Castenholz RW. (2000). Evolution of thermotolerance in hot spring cyanobacteria of the genus *Synechococcus*. *Appl Environ Microbiol* **66**: 4222-4229.

Nei M and Gojobori T. (1986). Simple methods for estimating the number of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**: 418-426.

Papke RT, Zhaxybayeva O, Fiel EJ, Sommerfeld K, Muise D and Doolittle WF. (2007). Searching for species in Haloarchaea. *Proc Natl Acad Sci* **104**: 14092-14097.

Peary JA and Castenholz RW. (1964). Temperature strains of a thermophilic blue-green alga. *Nature* **202**: 720-721.

Perriere G and Gouy M. (1996). WWW-Query: An online retrieval system for biological sequence banks. *Biochimie* **78**: 364-369.

Ramsing NB, Ferris MJ and Ward DM. (2000). Highly ordered vertical structure of *Synechococcus* population within the one-millimeter-thick photic zone of a hot spring cyanobacterial mat. *Appl Environ Microbiol* **66**: 1038-1049.

Salerno A, Deletoile A, Lefevre M, Ciznar I, Krovacek K, Grimont P and Brisse S. (2007). Recombining population structure of *Plesiomonas shigelloides* (*Enterobacteriaceae*) revealed by multilocus sequence typing. *J Bacteriol* **189**: 7808-7818.

Schloss PD and Handelsman J. (2006). Introducing SONS, a tool for OTU-based comparisons of membership and structure between microbial communities. *Appl Environ Microbiol* **72**: 6773-6779.

Schloss PD and Handelsman J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol* **71**: 1501-1506.

Schloss PD, Larget BR and Handelsman J. (2004). Integration of microbial ecology and statistics: a test to compare gene libraries. *Appl Environ Microbiol* **70**: 5485-5492.

Stackenbrandt E and Ebers J. (2006). Taxonomic parameters revisited: tarnished gold standards. *Microbiol Today* **November**: 153-155.

Tamura K, Dudley J, Nei M and Kumar S. (2007). MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**: 1596-1599.

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

Van Valen L. (1976). Ecological species, multispecies, and oaks. *Taxon* **25**: 233-239.

Vitorino LR, Margos G, Feil EJ, Collares-Pereira M, Ze-Ze L and Kurenbach K. (2008). Fine-scale phylogeographic structure of *Borrelia lusitaniae* revealed by multilocus sequence typing. *PLOS ONE* **3**: 1-13.

Ward DM, Cohan FM, Bhaya D, Heidelberg JF, Kuhl M and Grossman A. (2008). Genomes, environmental genomics and the issue of microbial species. *Nat Heredity* 100: 207-219.

Ward DM, Bateson MM, Ferris MJ, Kuhl M, Wieland A, Koeppel A and Cohan FM. (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. *Phil Trans Roy Soc B* **361**:1997-2008.

Ward DM and Cohan FM. (2005). Microbial diversity in hot spring cyanobacterial mats: pattern and prediction. In *Geothermal Biology and Geochemistry in Yellowstone National Park*. Inskeep B and McDermott T (eds.). Thermal Biology Institute: Bozeman, MT. Pp 185-201.

Ward DM; Ferris MJ, Nold SC and Bateson MM. (1998). A natural view of microbial biodiversity within hot spring cyanobacterial mat communities. *Microbiol Mol Biol Rev* **62**: 1353-1370.

Ward DM, Weller R and Bateson MM. (1990). 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature* **345**: 63-65.

Wayne LG, Brenner DJ, Colwell RR, Grimont PAD, Kandler O, Krichevsky MI, Moore WEC, Murray RGE, Stackenbrandt E, Starr MP and Truper HG. (1987). Report of the *ad hoc* committee on reconciliation of approaches to bacterial systematics. *Int J Syst Bacteriol* **37**: 463-464.

Whitaker RJ and Banfield JF. (2005). Population dynamics through the lens of extreme environments. *Rev Mineral Geochem* **59**: 259-277.

# CHAPTER 3

# BACTERIAL ARTIFICIAL CHROMOSOME LIBRARIES FOR MUSHROOM SPRING CYANOBACTERIAL MATS, YELLOWSTONE NATIONAL PARK[4]

Melanie C. Melendrez[2], Jason M. Wood[2], Douglas B. Rusch[5], John F. Heidelberg[6] and David M. Ward[2]

## Abstract

Two bacterial artificial chromosome (BAC) libraries were constructed from Mushroom Spring microbial mat samples in order to obtain multiple linked protein-encoding loci for use in a multi-locus population genetics study of *Synechococcus* populations. Libraries from 60° and 65°C samples contained 304,128 and 64,512 clones with average insert sizes of 120 kb and 100 kb, respectively. 9,216 randomly selected clones from both 60° and 65°C libraries were end sequenced and sequences were compared to 20 reference genomes from isolates relevant to the community. Many sequences of high % nucleotide identity were recruited by *Synechococcus* strains A and B′, *Roseiflexus* sp. RS1 and *Chloroflexus* sp. 396-1 genomes, though the recovery of *Synechococcus* sequences was lower than in metagenomic libraries with shorter inserts (i.e., 2-12 kb). As with these short-insert metagenomes, matches to homologs in *Synechococcus* strain A and *Chloroflexus* sp 396-1 genomes were more numerous in the

---

[4] This was a collaborative study. I conducted all of the sample preparation necessary for the creation of BAC clone libraries by Amplicon Express, which I then screened before sequencing was performed by the J. Craig Venter Institute. Figures involving metagenomic sequences were created with the help of Jason Wood and all subsequent analysis was done by myself under the supervision of Dr. Ward.

[5] J Craig Venter Institute, Rockville, MD

[6] Wrigley Institute, University of Southern California, Avalon, CA

65°C sample than were matches to homologs in the *Synechococcus* strain B´ and *Roseiflexus* sp. RS1 genomes, which recruited a higher proportion of sequences from the 60°C library. Many sequences of lower % nucleotide identity were recruited, mainly from the 60°C library, by the genomes of *Candidatus* Chloroacidobacterium thermophilum and *Chloroherpeton thalassium*. Comparison of BAC clone insert sizes and orientation of end-sequence mate pairs to reference genomes showed that synteny was lower than in metagenomic clones with shorter inserts. Oligonucleotide probe screening identified 1,193 and 236 clones that contained a *Synechococcus* A/B´ lineage-specific 16S rRNA locus from 60°C and 65°C samples, respectively. Comparison of end sequences of these clones with these reference genomes revealed the presence of both syntenous and nonsyntenous populations, which were randomly distributed between the two 16S rRNA loci of the *Synechococcus* strain A and B´ genomes. The nonsyntenous population appears to have undergone rearrangement in the vicinity of the 16S rRNA loci and subsequent mate-pair analysis suggests that this may have been due in part to genomic inversion.

## Introduction

Large-insert metagenomic clone libraries using bacterial artificial chromosome (BAC), cosmid or fosmid vectors have been used for studying environmental microbial diversity for over a decade (Beja et al., 2000a/b; Rondon et al., 1999, 2000; Shizuya et al., 1992; Stein et al., 1996). BAC vectors are based on the well-studied *Escherichia coli* F factor, which is capable of maintaining fragments of DNA as large as 1 Mb (Shizuya et

al., 1992). Replication of the F factor is strictly controlled and the F plasmid is maintained in low copy number (one or two copies per cell), which reduces the potential for recombination between DNA fragments within the plasmids (Shizuya et al., 1992). BAC clone libraries have been used to characterize microbial assemblages from soil, the ocean, bacterial symbionts in sponges and to search for antibiotic resistance genes and biologically active small molecules involved in quorum sensing (Rondon et al., 2000; John et al., 2006; Beja et al., 2000b; Ouyang et al., 2009; Liles et al., 2003; Riesenfeld et al., 2004; Williamson et al., 2005).

We applied BAC cloning to our research on hot spring microbial mats in order to conduct cultivation-independent, high-resolution, multi-locus sequence analysis (MLSA) of the population genetics of its cyanobacterial inhabitants. A progression of studies of single genetic loci, including the 16S rRNA gene (Ferris and Ward, 1997; Ramsing et al., 2000), the internal transcribed spacer separating 16S and 23S rRNA genes (Ferris et al., 2003), and, most recently, protein-encoding loci (Melendrez et al., 2010a; Chapter 2; Becraft et al., 2010), has revealed much about the predominant members of the mat community and their ecology. The predominant cyanobacteria, *Synechococcus* populations of the A/B lineage, exhibit distribution patterns suggesting that they occur as ecological species (Ward et al., 1998; Van Valen, 1976), which are adapted to parameters that vary along the effluent flow path (temperature, nutrients) and vertically (light quantity and quality) within the mat. For instance, only A-like *Synechococcus* populations. occur at 65°C, whereas at 60° B´-like *Synechococcus* populations occur together with A-like *Synechococcus* populations, consistent with temperature adaptations

of isolates of these types (Allewalt et al., 2006). At 60°C, A-like *Synechococcus* populations occur only at depths below 400 μm, suggesting possible adaptation to low and/or spectrally altered light (Ramsing et al., 2000). In addition, comparison of genomes of *Synechococcus* strain A and B´ isolates, to end sequences from clones in small-insert mat metagenomes (Klatt et al., 2010) revealed evidence of possible adaptations in nitrogen, phosphorus and iron nutrition (Bhaya et al., 2007).

Single-locus analyses have two types of limitations. First, molecular resolution varies among loci. Second, they cannot account for the effects of recombination among genes, which has been shown to be important in many bacteria (Hanage et al., 2006; Bilek et al., 2009; Tanabe et al., 2009). MLSA was developed in order to account for recombination in population genetics analyses of cultivated organisms (Feil et al., 1999, 2000; Turner and Feil, 2007; Hanage et al., 2006; Vitorino et al., 2008; Tanabe et al., 2007). BAC cloning provided a means of sampling many genetic loci simultaneously, thus permitting cultivation-independent MLSA analysis of native populations. By focusing on genomic regions that include 16S rRNA genes it is possible to link MLSA analyses with previous single-locus population genetics studies of this mat community.

Here we describe the construction and characterization of BAC libraries for Mushroom Spring microbial mat samples collected at 60°C and 65°C. The genomes of many mat community members, including the predominant cyanobacteria (*Synechococcus* strains A and B´), filamentous anoxygenic phototrophic (FAP) bacteria (*Roseiflexus* strain RS-1, *Chloroflexus* strain 396-1), other newly discovered anoxygenic phototrophs (Bryant et al., 2007) and many heterotrophic mat isolates, such as

*Thermomicrobium roseum* (Wu et al., 2009) have facilitated comparative analyses of mat metagenomes (Klatt et al., 2010) and our analysis of the BAC libraries. Our analyses revealed evidence that BACs containing *Synechococcus* A/B lineage 16S rRNA loci have undergone genomic rearrangements, via inversion events, in the vicinity of 16S rRNA genes.

## Methodology

### Sample Preparation

Samples (approximately 1.5 cm length x 0.5 cm width x 0.5 cm thick) were collected from 60°C and 65°C sites (termed M60 and M65, respectively) in the mat of Mushroom Spring, Yellowstone National Park, WY (44.5386°N, 110.7979°W) on 2 October 2003 as previously described (Melendrez et al., 2010a; see Chapter 2). The top 1-mm thick green layers of three mat samples were separated from the bottom layers using a razor blade, combined and homogenized with 2 ml of 10mM Tris-HCl buffer using a dounce tissue homogenizer. 1 ml aliquots of the mat homogenate were transferred into two 2 ml screw cap tubes and centrifuged at 14,000 rpm (Eppendorf, 5415C) for 5 min to pellet the cells. The supernatant was removed and the cells were resuspended in 1 ml of 10 mM Tris/EDTA buffer (TE: 10 mM Tris-HCl, pH 7.5 and 1 mM EDTA).

### BAC Library Construction

An 850 μl cell suspension from either M60 or M65 samples was transferred to a 2 ml tube. 20 μl of lysozyme (10 mg/ml) was added and the mixture was incubated for 30 min at 37°C. The cell suspension was viewed under the microscope for the appearance of

spheroplasts (Figure B3.1) and then centrifuged at 14,000 rpm (Eppendorf, 5415C) for 5 min to pellet the spheroplasts and remaining cells. The supernatant was removed and the cells were frozen and sent to Amplicon Express (http://www.genomex.com) (Pullman, WA), where they were resuspended in 800 µl of TE buffer and 870 µl of molten 1% Seakem GTG Agarose was added. This suspension was aliquoted into 200-µl plug molds (BioRad) and plugs were solidified at 4°C for 10-15 min. For in-gel lysis, five plugs were incubated in a 50-ml Falcon tube (Eppendorf) containing 40 ml of ESP buffer (ESP: 0.5 M EDTA, pH 9-9.5; 1% Sarkosyl and 50µg/ml proteinase K) and 40 mg of crystaline proteinase K (final concentration, 1 mg/ml proteinase K) overnight at 55$^{o}$C with rotation in a hybridization incubator (Robbins Scientific). The buffer was discarded and replaced with fresh ESP buffer and 40 mg of proteinase K and the mixture was incubated an additional hour at 55°C with rotation. ESP Buffer was discarded and plugs were briefly washed with 10 mM TE Buffer. 40 ml of fresh TE buffer was added and plugs were incubated at room temperature for 1 hour with rotation. TE buffer was discarded and replaced and plugs were stored at 4°C overnight to stabilize high molecular weight (HMW) DNA.

The M60 BAC library was constructed at Amplicon Express from HMW genomic DNA using a method adapted from Liles et al. (2008) and Tao et al. (2002). The HMW DNA was partially digested with HindIII and size-selected (Figure B3.2). Ligation of partially digested HMW DNA to the pECBAC1 vector was carried out in a tube with 100 µl of insert DNA (molar ratio 4:1 vector excess) and 5 units of T4 DNA ligase (GIBCO BRL) incubated at 16°C for 10 hrs. The resultant BAC clones were transformed into

DH10B *Escherchia coli* cells (Invitrogen) and plated on LB agar (Fisher Scientific) with chloramphenicol (12.5 μg/ml), X-gal (40 μg/ml) and IPTG (0.4 mM). Clones were robotically picked with a Genetix QPIX (Genetix) into 792 384-well plates containing Luria Broth (LB) freezing media (Zhang et al., 1996). Plates were incubated for sixteen hours, replicated and then frozen at -80°C. The replicated copy was used as a source plate for nylon filters containing imprinted BAC clones that were sent to Montana State University for probe screening and processing (see below). The M65 library was produced in the same way, except that the cloning vector was pCC1BAC and only 168 384-well plates were made.

To estimate insert sizes, 10 μl aliquots of BAC miniprep DNA (Amplicon Express) were digested with 5 U of Not I enzyme (New England Biolabs) for 3 hrs at 37 °C. The digestion products were separated by pulsed-field gel electrophoresis (CHEF-DRIII system, BioRad) in a 1% agarose gel in 0.5x Tris/borate/EDTA (TBE) electrophoresis buffer (10x TBE: 890 mM Tris, 890 mM Boric acid, 20 mM EDTA, pH 8.0) (Figure B3.3). Insert sizes were compared to those of the Lambda Ladder PFG Marker (New England Biolabs). Electrophoresis was carried out for 18 hrs at 14°C with an initial switch time of 5 sec. and a final switch time of 15 sec. in a voltage gradient of 6 V/cm.

### $^{32}$P-screening for Clones Containing *Synechococcus* A/B Lineage-Specific 16S rRNA Genes

BAC clones imprinted and lysed on Hybond-N+ nylon membranes (Amersham Biosciences) were probed with a $^{32}$P-radiolabeled *Synechococcus* A/B lineage cluster

probe (5´-ctgagacgcggtttttgg-3´) (Papke et al., 2003), which was prepared by mixing 2.7 µl of the A/B cluster probe (50 µM) with 4 µl Sigma water (Sigma), 1.3 µl T4 buffer (Promega), 1.3 µl T4 kinase (Promega) and 4 µl $^{32}$P-ATP (Perkin Elmer). The mixture was incubated at 37°C for 20 min., denatured for 2 min. at 90°C and immediately placed on ice. 133 µl of Sigma water was added to bring the total volume to 146.3 µl.

Membranes were placed into hybridization tubes and rinsed with autoclaved double-distilled water, then with 5x sodium chloride/sodium citrate (SSC) buffer (20x SSC: 3M NaCl, 0.3M $Na_3$citrate-$2H_20$, pH 7.0). 1.34 ml of pre-hybridization buffer [5x SSC, 5x Denhardt's Solution (Fisher Scientific), 0.5% Sodium Dodycel Sulfate (SDS, Fisher Scientific)] was added to each tube and the tubes were placed into a hybridization oven (Robbins Scientific, model 400) for 30 min. with rotation at 51°C. After pre-hybridization, 36.6 µl of probe mixture was added and the probe and filter were hybridized overnight at 51°C with rotation.

To remove unreacted probe following hybridization, membranes were washed twice at low stringency (2x SSC, 0.1% SDS), twice at medium stringency (1x SCC, 0.1% SDS) and four times at high stringency (0.1X SSC, 0.1% SDS) (adapted from the Amersham Biosciences H-bond Nylon+ Membrane standard protocol). Membranes were kept wet, placed in saran wrap and exposed for approximately 30 hrs on a phosphoimaging cassette (Kodak). Positive clones were visualized using a phosphoimager (Phosphoimager) and the program ImageQuant Software (http://www.imsupport.com) (Figure B3.4). Positive clones (see Appendix B) were transferred at Amplicon into 96-well plates containing 1 ml of LB (Fisher Scientific),

grown overnight at 37°C and then glycerol was added to each well for preservation at -80°C. Plates were then frozen until PCR analysis using primers specific for the 16S rRNA-ITS region of cyanobacteria (781cyF and L23cyR), described in Melendrez et al. (2010a; see Chapter 2).

BAC-End Sequencing

A random selection of 9,216 BAC clones from each library (M60 and M65) and BAC clones containing a cyanobacterial 16S rRNA gene were end-sequenced at the J. Craig Venter Institute (JCVI: http://www.jcvi.org) using the Sanger sequencing method and T7 (5′-taatacgactcactatag|gg-3′) and M13R (5′-caggaaacagctatgac-3′) primers.

Metagenomic Analysis of BAC libraries

BAC end sequences were compared to twenty genomes of microorganisms thought to possibly be representative of indigenous mat populations on the basis of their genetic relevance, cultivation history or functional contributions. Custom Perlscripts used in this work were developed for the analysis of end sequences of clones with 2-10 kb inserts from libraries constructed from the same mat samples (see Klatt et al., 2010). Briefly, WU-BLASTN was used to recruit BAC end sequences to the concatenated twenty-genome database and alignments were considered significant if they shared ≥ 50% nucleotide identity with a target of length approximately 100 bp. Sequences that didn't meet these criteria with an e-value as significant as $E = 10^{-10}$, were assigned to a "null" bin, indicating that we have low confidence in their having any relationship with genomes used in this analysis (Klatt et al., 2010). Sequences were considered to be

"jointly recruited" when both end sequences of a particular clone insert had most significant WU-BLASTN matches with the same isolate genome and "disjointly recruited" when matching different genomes. Jointly recruited sequences were considered syntenous, when the end sequences were within 30% of the average insert size determined by gel analysis when tiled to a reference genome and the sequences showed the same directionality and strandedness as observed on the reference genome (see Klatt et al., 2010). For mate pair analysis, categories were identified by fragment recruitment plots generated from tools available via the Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA) website (http://camera.calit2.net).

## Results and Discussion

Characteristics of the BAC libraries are shown in Table 3.1. The M60 BAC library contained 304,128 clones with an average insert size of ~120 kb (Figure B3.3). A total of 10,408 clones were end-sequenced, 9, 216 random clones from the metagenomic BAC library and 1,193 BAC clones that were identified as containing a *Synechococcus* A/B lineage 16S rRNA gene (see below). The M65 BAC library contained 64,512 clones with an average insert size of ~100 kb. A total of 9,452 clones from this library were end-sequenced, 9,216 random clones and 236 that were identified to contain a cyanobacterial 16S rRNA gene.

Table 3.1. BAC library characteristics.

| | **M60** | **M65** |
|---|---|---|
| Number of Clones | 304,128 | 64,512 |
| Average Insert Size | 110-130 kb | 100-110 kb |
| Number of End-Sequenced Clones | 10,408 | 9,452 |
| Number of *Synechococcus* 16S rRNA-bearing BAC clones | 1,193 | 236 |

Composition of BAC Metagenomic libraries

The compositions of these metagenomes are discussed below for different groups of organisms of importance to the mat community on the basis on several kinds of analyses. Genome recruitment results of M60 and M65 random BAC end sequences compared to 20 reference genomes are presented in Figure 3.1, where both the frequency of sequences recruited by each genome and the percent nucleotide identities of the alignments between metagenomic and isolate homologs are displayed. More detailed displays of the % nucleotide identity of alignments between metagenomic sequences and the reference genomes are presented in Figures 3.2 and 3.3, where disjointly recruited, jointly recruited syntenous and jointly recruited nonsyntenous sequences are highlighted. Similar plots for genomes that did not recruit BAC end sequences of high % nucleotide identity are shown in Figures B3.5 and B3.6. Table 3.2 quantifies the percentage of metagenomic sequences that can be confidently associated with a reference genome after normalization for genome length differences that might bias a view of composition toward organisms with longer genomes. Table 3.2 percentage values are used in the text below. Klatt et al. (2010) determined using comparative genome analyses that strains of *named* species typically exhibit up to 70% nucleotide identity, whereas different species

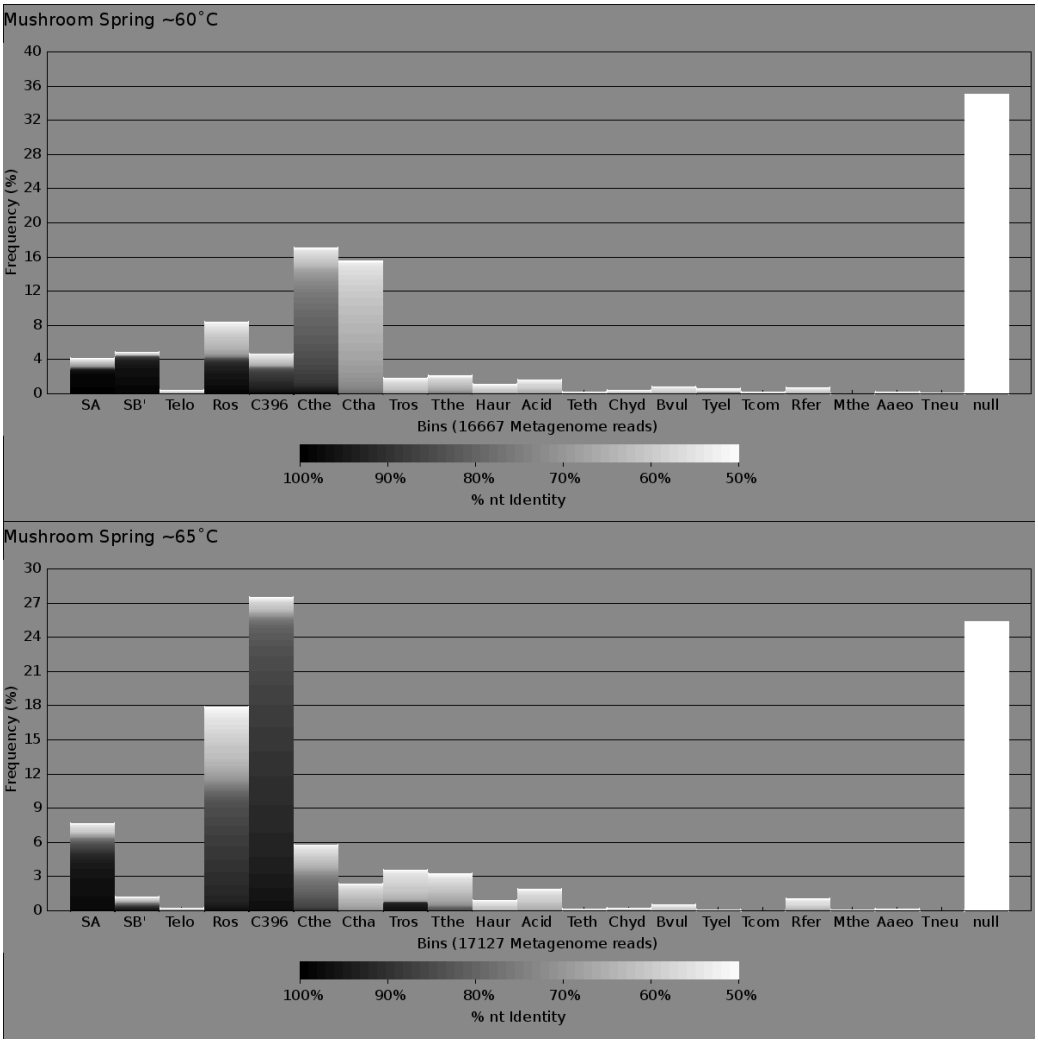



Figure 3.1. BLASTN-based recruitment of metagenomic sequences from BAC libraries prepared from top green (~1 mm) mat layers from M60 (top) and M65 samples (bottom) by genomes of 20 microorganisms of possible relevance to these mats. SA, *Synechococcus* strain A; SB′, *Synechococcus* strain B′; Telo, *Thermosynechococcus elongatus*; Ros, *Roseiflexus* sp. RS-1; C396, *Chloroflexus* sp. 396-1; Cthe, *Candidatus Chloracidobacterium thermophilum*; Ctha, *Chloroherpeton thalassium*; Tros *Thermomicrobium roseum*; The, *Thermus thermophilus*; Haur, *Herpetosiphon aurantiacus*; Acid, *Acidobacterium* sp.; Tpse, *Thermoanaerobacter pseudoethanolicus*; Chyd, *Carboxydothermus hydrogenoformans*; Bvul, *Bacteroides vulgate*; Tyel, *Thermodesulfovibrio yellowstonii*; Tcom, *Thermodesulfobacterium commune*; Rfer *Rhodoferax ferrireducens*; Mthe, *Methanothermobacter thermoautotrophicum*; Aaeo, *Aquifex aeolicus*; and Tneu, *Thermoproteus neutrophilus*. Shading indicates % NT ID of sequences within bins (details regarding selection of genomes in Klatt et al., 2010).

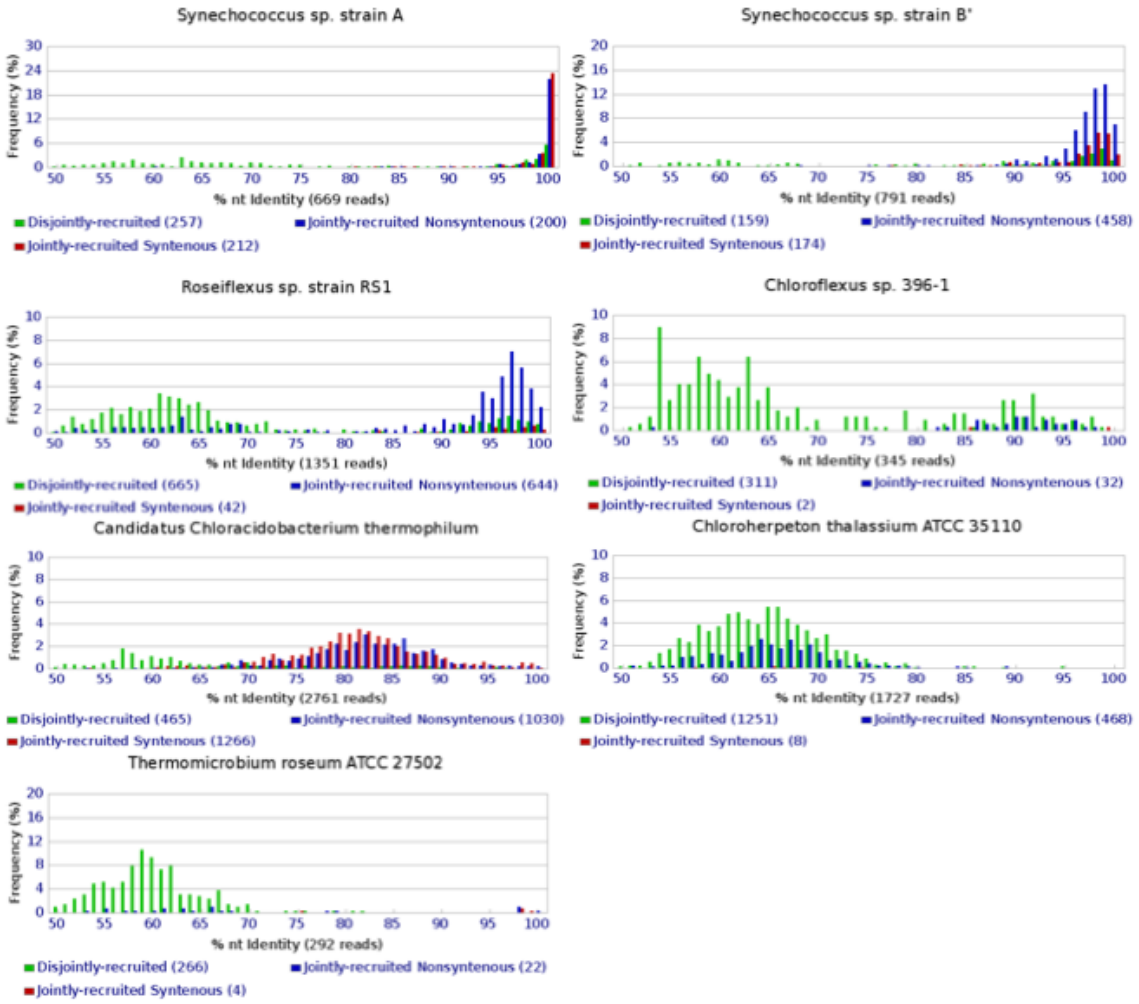



Figure 3.2. Frequency of disjointly recruited (green), jointly recruited syntenous (red) and jointly recruited nonsyntenous (blue) end sequences from randomly selected BACs from the M60 library as a function of their % nt identity relative to homologs in the reference genomes that recruited them.

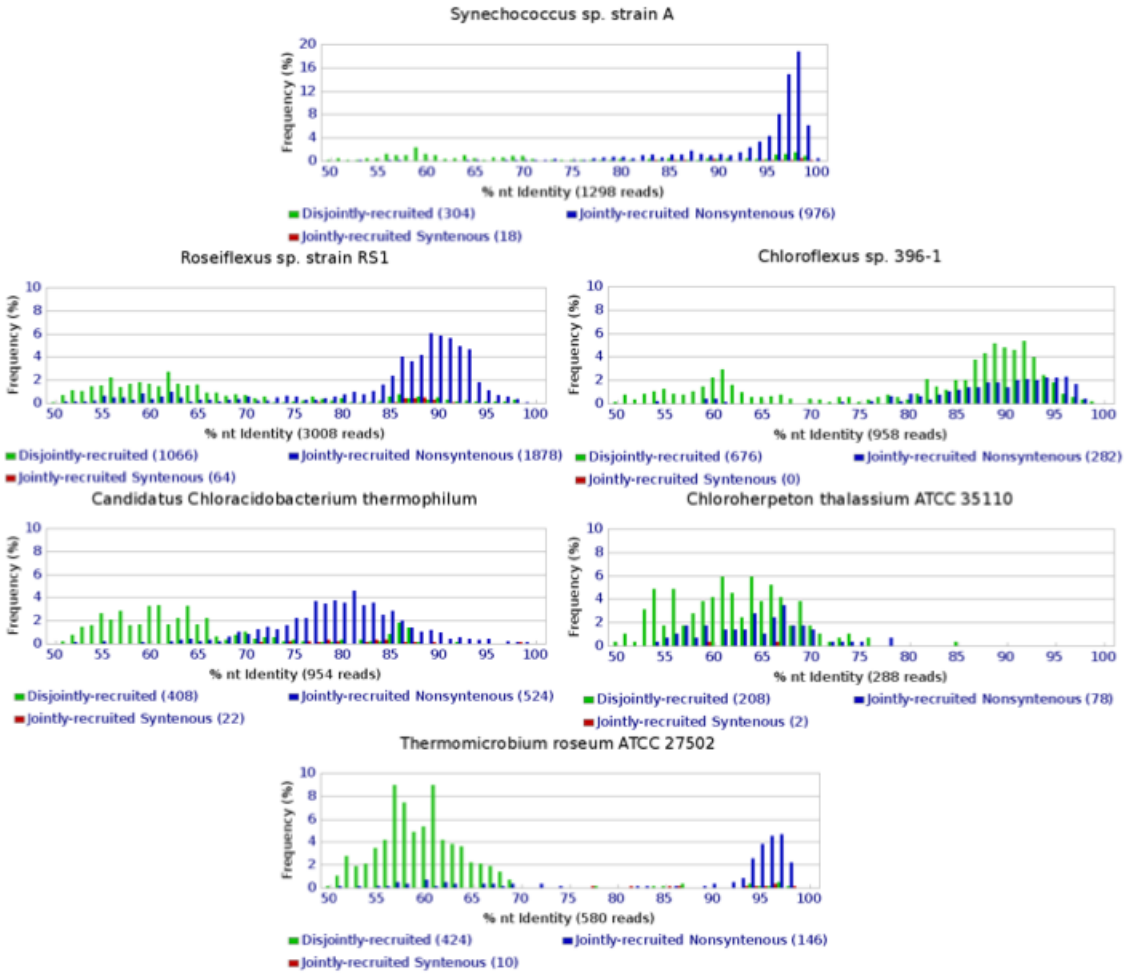



Figure 3.3. Frequency of disjointly recruited (green), jointly recruited syntenous (red) and jointly recruited nonsyntenous (blue) end sequences from randomly selected BACs from the M65 library as a function of their % nt identity relative to homologs in the reference genomes that recruited them.

Table 3.2. Comparison of BAC and small-insert metagenomic library compositions and synteny with reference genomes.

| Reference Genome | Genome Size (Mb) | % nt ID range[a] | % of sequences recruited normalized for genome length[b] | | | | % Syntenous/Nonsyntenous Clones[c] of No. total sequences in analysis | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Small-Insert | | BAC | | Small-Insert | | BAC | |
| | | | MSlow[a,d] | MShigh[a,e] | M60 | M65 | MSlow[a,d] | MShigh[a,e] | M60 | M65 |
| *Synechococcus* strain A | 2.93 | 83-100 | 7.8 | 25.7 | 4.0 | 8.0 | 62/12 (1257) | 69/20 (10172) | 30.8/29.0 (689) | 1.4/74.6 (1308) |
| *Synechococcus* strain B´ | 3.04 | 90-100 | 21.5 | 1.1 | 5.0 | 1.0 | 72/19 (3414) | 7/6 (928) | 21.5/56.7 (808) | n.d. |
| *Roseiflexus* sp. RS-1 | 5.80 | 80-100 | 8.08 | 4.84 | 4.0 | 3.0 | 58/28 (2793) | 41/26 (5616) | 3.0/46.1 (1397) | 2.1/61.5 (3056) |
| *Chloroflexus* sp. 396-1 | 5.20 | 65-100 | 0.43 | 8.96 | 1.0 | 9.0 | 28/2 (201) | 77/9 (5071) | 0.6/8.7 (366) | 0/27.8 (1016) |
| *Candidatus* Chloroacidobacterium thermophilum | 3.70 | 70-100 | 7.64 | 1.69 | 5.0 | 1.3 | 70/15 (1663) | 37/14 (1582) | 45.0/36.6 (2813) | 2.3/54.1 (968) |
| *Chloroherpeton thalassium* | 3.29 | 50-75 | 5.75 | 1.88 | 9.0 | 2.0 | 14/25 (748) | 6/22 (655) | 0.5/26.4 (1770) | 0.7/26.7 (292) |
| *Thermomicrobium roseum* | 2.93 | 75-100 | 0.21 | 0.66 | 2.0 | 3.0 | 15/10 (192) | 18/7 (1141) | 1.3/7.4 (297) | 1.7/24.5 (595) |

[a] Small-insert data obtained from Klatt et al. (2010) was from libraries constructed from the same mat samples used to construct the BAC clone libraries but with different lysing and library construction protocols.

[b] Total number of sequences within range recruited by a genome divided by the total number of sequences in library and multiplied by a normalization factor consisting of the ratio of genome size to the *Synechococcus* strain A genome (Klatt et al., 2010).

[c] Number of syntenous or nonsyntenous sequences divided by the total number of sequences, including disjointly recruited sequences (green bars, Figures 3.2 and 3.3).

[d] MSlow temperature was 60°C (Klatt et al., 2010).

[e] MShigh temperature was 65°C (Klatt et al., 2010).

of a *named* genus exhibit <70% nucleotide identity; % nucleotide identities below ~60% are barely discernable from homologs in genomes belonging to members of different kingdoms. We used the same ranges of % nucleotide identity used by Klatt et al. (2010) to define percentage contributions of taxa. It is important to note that each of these recruitment bins contains many ecological species, challenging the notion that *named* species are true species (Melendrez et al., 2010a; see Chapter 2). Table 3.2 also presents the % syntenous and nonsyntenous sequences and compares all these measures to the composition observed in 2-12 kb-insert metagenomic libraries described by Klatt et al. (2010). Figures 3.4 and 3.5 show the coverage by high % nucleotide homologs of reference genomes that are highly representative of sequences in the M60 and M65 BAC libraries. The sample investigated represents only 3.4% of the M60 clones and 14.7% of the M65 clones, so that the extent of coverage is greatly underestimated in these figures.

Cyanobacteria. The *Synechococcus* strain A and B´ genomes recruited ~4% and ~5% of the BAC-end sequences in the M60 library and ~8% and ~1% of the BAC-end sequences in the M65 library, respectively (Figure 3.2). These percentages were much lower than in 2-12 kb insert libraries, where the *Synechococcus* strain A and B´ genomes recruited 7.8% and 21.5% and 25.7% and 1.1% of the metagenomic sequences obtained from M60 and M65 samples, respectively (Table 3.2). The lower representation of *Synechococcus* in BAC libraries was most likely due to bias resulting from in-gel lysis or to difficulties in obtaining high quality HMW DNA from the mat samples. It was difficult to lyse the cyanobacteria and still maintain the integrity of the HMW DNA (personal communication, Keith Stormo, Amplicon Express). As expected from previous

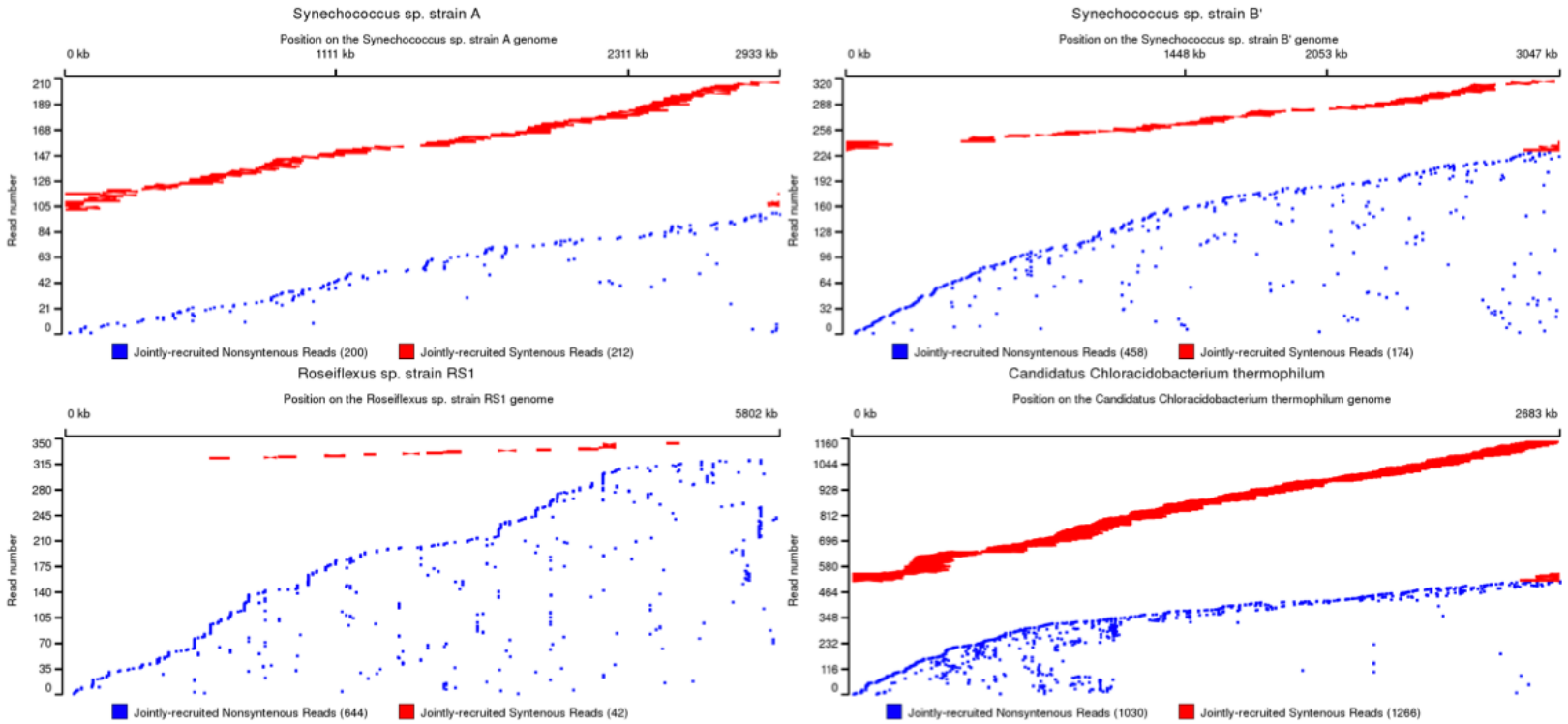



Figure 3.4. Coverage of *Synechococcus* strain A and B´, *Roseiflexus* strain RS-1 and *Cand.* Chloroacidobacterium thermophilum genomes by jointly recruited syntenous (red) and nonsyntenous (blue) end sequences of randomly selected BACs from the M60 library.

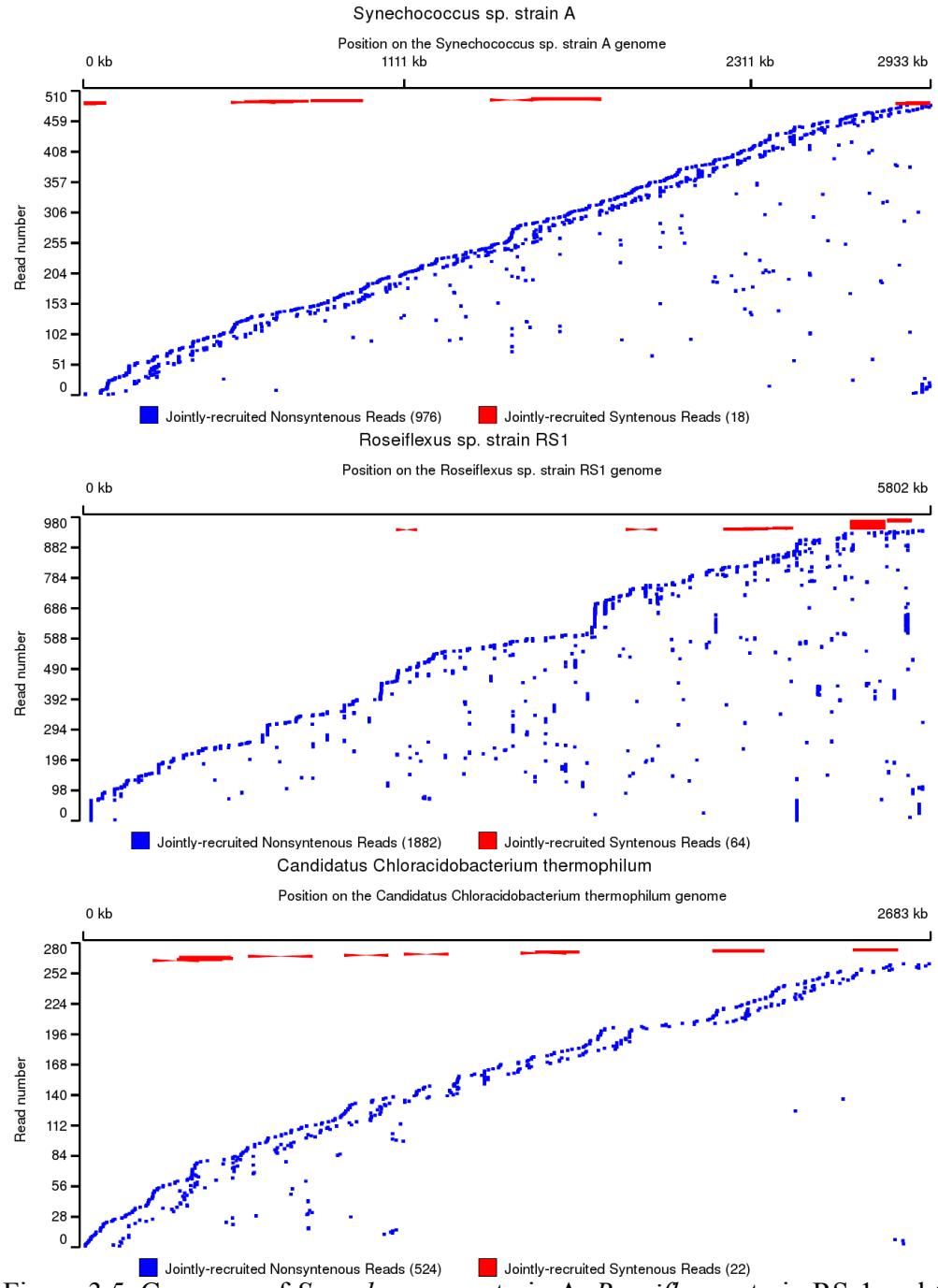



Figure 3.5. Coverage of *Synechococcus* strain A, *Roseiflexus* strain RS-1 and *Cand.* Chloroacidobacterium thermophilum genomes by jointly recruited syntenous (red) and nonsyntenous (blue) end sequences of randomly selected BACs from the M65 library.

distribution analyses (Ferris and Ward, 1997), A-like *Synechococcus* sequences were present in higher abundance than B′-like sequences in the M65 sample (Figure 3.1). Those B′-like sequences that were recruited from the M65 sample had low % nucleotide identity to the reference genome (Figure B3.6) and were mostly disjointly recruited. This suggests that these sequences were not in fact *Synechococcus* B′-like sequences; rather, given the choices of reference genomes in the analysis, they were recruited to the B′ genome rather than by other sequences. The majority of sequences recruited by the *Synechococcus* strain A and B′ genomes were jointly recruited sequences at >90% nt identity compared to homologs in the reference genomes (Figures 3.2 and 3.3). Synteny was lower than that observed in small-insert clones (Table 3.2), likely a result of the greater probability of genomic rearrangements in longer genome segments. Interestingly, synteny in sequences recruited by the *Synechococcus* strain A genome was greater in BACs from M60 than in BACs from M65 (compare Figures 3.2 and 3.3 and see Table 3.2), consistent with the existence of ecologically distinct A-like *Synechococcus* populations at these temperatures (Melendrez et al., 2010a; see Chapter 2; Becraft et al., 2010). Coverage of the reference *Synechococcus* strain A and B′ genome was good considering the relatively low number of BACs analyzed, with few gaps (Figures 3.4 and 3.5). By tiling BAC-end sequences recruited from the M65 library by the *Synechococcus* strain A genome as a function of the % nt identity of homologs (Figure 3.6), it was evident that, as with small-insert libraries (Klatt et al., 2010), there are two populations of A-like *Synechococcus* of different relatedness in this library. The population of clones that are ~10-20% diverged from this reference genome presumably represent the

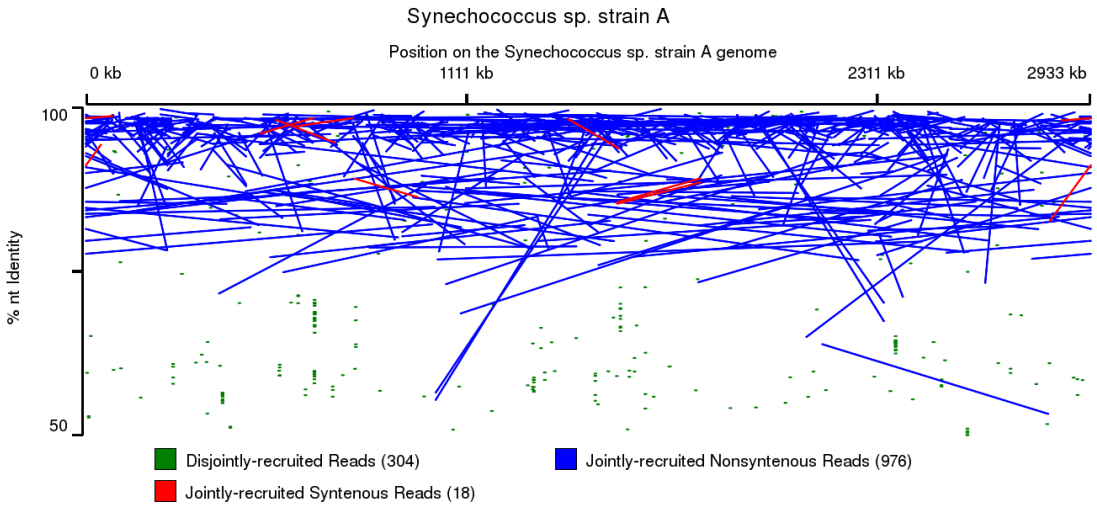



Figure 3.6. Distributions and percent nucleotide identity of jointly recruited syntenous (red) and nonsyntenous (blue) end sequences of BAC clones that contain A-like 16S rRNA genes obtained from the M65 libraries relative to homologs in the *Synechococcus* strain A genome. Lines connect end sequences of the same clone.

presence of A´-like *Synechococcus* in this sample, as had been suggested previously on the basis of recovering A´-like 16S rRNA (Ward et al., 2006), *aroA* and *apcAB* sequences from this sample (Melendrez et al., 2010a; see Chapter 2). Furthermore metagenomic sequences obtained from the 68°C Mushroom Spring mat sample, which is dominated by A´-like *Synechococcus* populations (Becraft, Klatt, Rusch and Ward unpublished), show a similar degree of divergence from homologs in the *Synechococcus* strain A genome, (Figure B3.7).

Filamentous Anoxygenic Phototrophic Bacteria. The *Roseiflexus* sp. RS-1 genome recruited ~4% and ~3% of the BAC-end sequences from the M60 and M65 libraries, respectively (Table 3.2). Many sequences were jointly recruited (mainly nonsyntenous) and the % nt identity to homologs in the reference genome was higher for M60 than M65 sequences (Figure 3.2 and 3.3 and Table 3.2), suggesting that this

*Roseiflexus* strain is more representative of lower-temperature populations. This genome also recruited many disjointly recruited sequences of low % nt identity that cannot be confidently associated with filamentous anoxygenic phototroph genomes (see Klatt et al., 2010). There were more jointly recruited nonsyntenous sequences that covered the genome of *Roseiflexus* sp. RS1 than jointly recruited syntenous sequences at M60 and M65 (644 versus 42 at M60 and 1878 versus 64 at M65) (Figures 3.4 and 3.5).

The *Chloroflexus* sp. 396-1 genome recruited ~1% and ~9% of the BAC-end sequences from the M60 and M65 libraries, respectively (Table 3.2). There were few syntenous sequences, however, and many disjointly recruited sequences of low % nucleotide identity, suggesting that this reference genome is not very representative of native *Chloroflexus* populations. Coverage was also poor in the sample analyzed (data not shown). More sequences were recruited from the M65 metagenome by the *Chloroflexus* sp. 396-1 genome than by the *Roseiflexus* sp. RS1 genome, consistent with findings in the small-insert metagenomes and with past studies of FAPs in these mats (Nübel et al., 2002).

Other Anoxygenic Phototrophs. The genome of *Candidatus* Cab. thermophilum, recently described as the first phototrophic acidobacterium-like organism (Bryant et al., 2007), recruited ~5% and ~1.3% of the BAC-end sequences in the M60 and M65 BAC libraries, respectively (Table 3.2), with equal amounts of syntenous and nonsyntenous jointly recruited clones for M60 and mainly nonsyntenous clones for M65 (Figures 3.2 and 3.3). As noted for small-insert metagenomes (Klatt et al., 2010), a surprisingly high level of synteny was observed with M60 metagenomic sequences (Figure 3.2), given that

homologs in this genome exhibited, on average, only about 80 to 85% nucleotide identity with native populations. Coverage of the Cab. thermophilum genome was excellent (Figures 3.4 and 3.5), especially in the M60 library.

Approximately 9% and 2% of the BAC-end sequences were recruited from the M60 and M65 BAC libraries, respectively, by the genome of the green sulfur bacterium *Chloroherpeton thalassium* (Table 3.2). This reference genome was, however, poorly representative of native populations (Figure 3.2) as evidenced by the low % nucleotide identity and lack of synteny with metagenomic homologs. These results are consistent with evidence of green sulfur bacteria in these mats that are distant relatives of other known green sulfur bacteria, based on 16S rRNA (Ferris and Ward, 1997), reaction center protein sequences (Bryant et al., 2007) and small-insert metagenomic sequence % nucleotide identity results (~50-75%) (Klatt et al., 2010). Coverage of this genome was poor (data not shown).

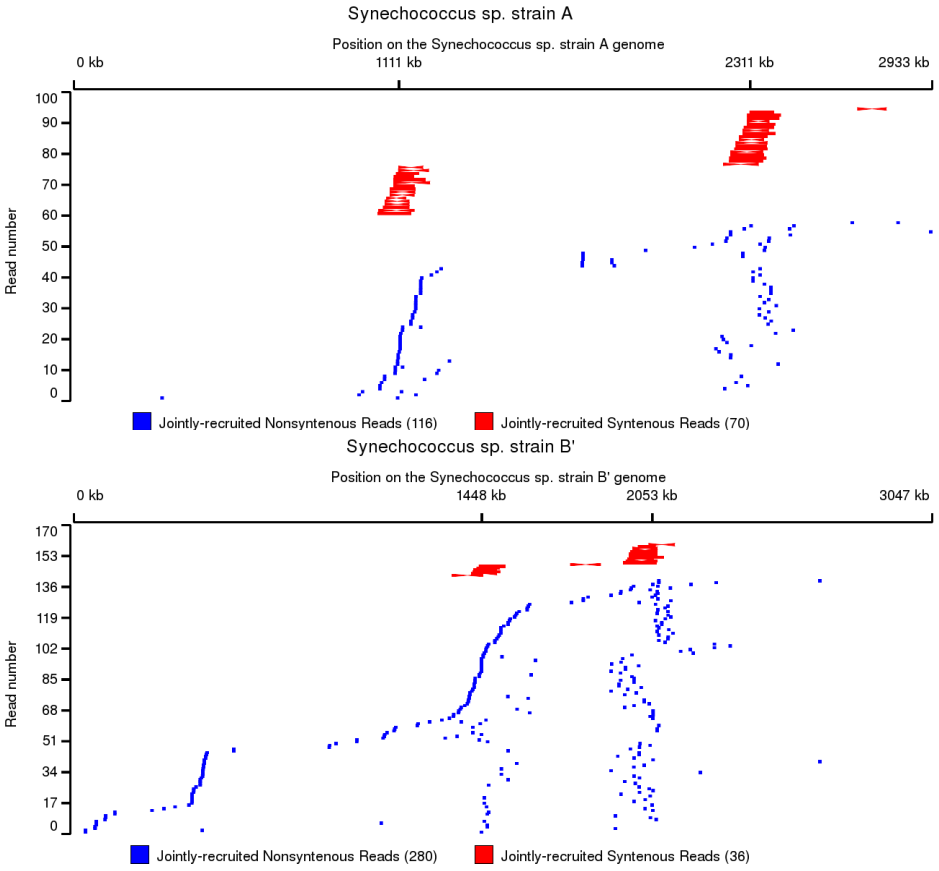
Other Genomes. The *Thermomicrobium roseum* genome recruited less than 4% of the BAC-end sequence sequences from both M60 and M65 BAC libraries (Table 3.2). The % nucleotide identities of the sequences that were recruited from the M60 library were very low (~50-60%); however, this genome recruited two subsets of sequences from the M65 metagenome, one with very high % nucleotide identity (90-100%) and one with very low % nucleotide identity (55-75%) (Figures 3.1 and 3.3). This and previous sequence analysis of small-insert metagenomic libraries (Klatt et al., 2010), suggests that this genome is a good representative for a minority population at 65°C that has high % nucleotide identity to *T. roseum*. Many of the sequences in the M60 and M65 samples

recruited by the *T. roseum* genome were disjointly recruited and there was very little coverage across the genome (data not shown). All other genomes recruited less than 3% of BAC-end sequences. Since the recruited sequences exhibited low % nucleotide identities to these genomes, we suggest that these genomes are not representative of the predominant organisms that inhabit the microbial mat, even though one of these isolates was cultivated from the highly similar Octopus Spring mat and many were cultivated from other Yellowstone hot springs and environments (see Klatt et al., 2010). Both libraries also contained large 'null' bins with ~35% (M60) and ~25% (M65) of metagenomic sequence that did not significantly match any of the reference genomes and thus are of unknown affiliation (Figure 3.1).

BAC Library Clones Bearing
*Synechococcus* A/B Lineage 16S rRNA Genes

Of the 368,640 total BAC clones in both libraries, 1,429 were found to contain a *Synechococcus* A/B´ lineage 16S rRNA gene. This was 0.4% of the BACs recruited by *Synechococcus* strain A and B´ genomes, close to expectations given that these ~3 Mb genomes contain 2 copies of the 16S rRNA gene (Bhaya et al., 2007). These BACs were also end-sequenced and tiled along the genomes of *Synechococcus* strain A and B´ and the following analyses are based on analysis of 634 of the 1,429 M60 and M65 BAC clones (Figures 3.7 and 3.8). Both syntenous and non-syntenous jointly recruited clones were observed, most of which mapped to one or both of the 16S rRNA loci of the reference genomes. It was difficult to discern whether syntenous and/or nonsyntenous sequences differed in % nt identity relative to homologs in the reference *Synechococcus*

A. M60




B. M65

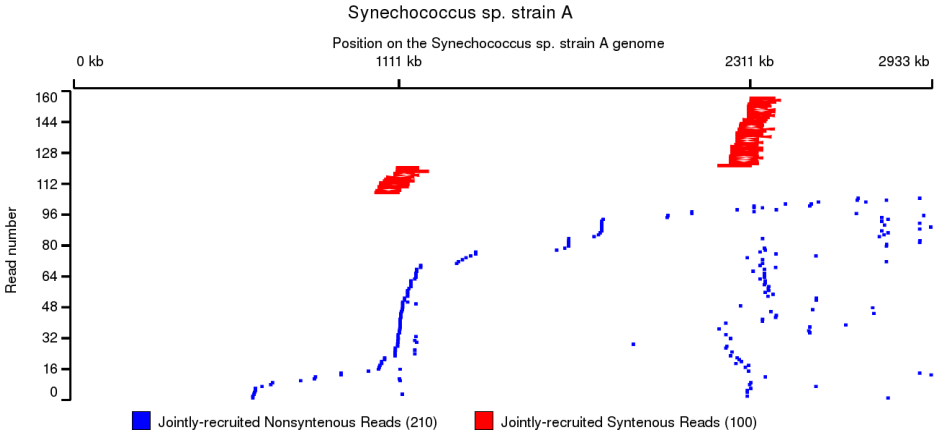
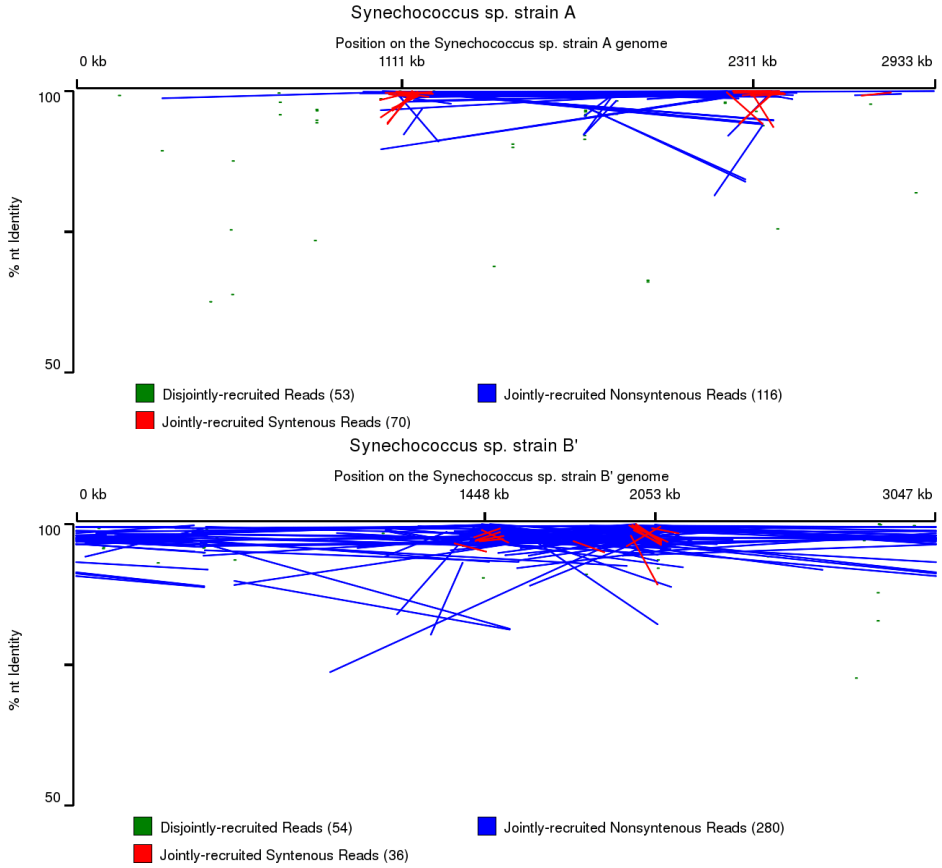



Figure 3.7. Distributions of end sequences of BAC clones that contain A-like and B´-like 16S rRNA genes obtained from the (A) M60 (based on analysis of ~346 M60 BAC clones) and (B) M65 libraries relative to *Synechococcus* strain A and B´ genomes. Jointly recruited syntenous (red lines connect end sequences of the same clone) and nonsyntenous (blue dots) sequences. X-axis ticks represent positions of the 16S rRNA genes.

A. M60




B. M65

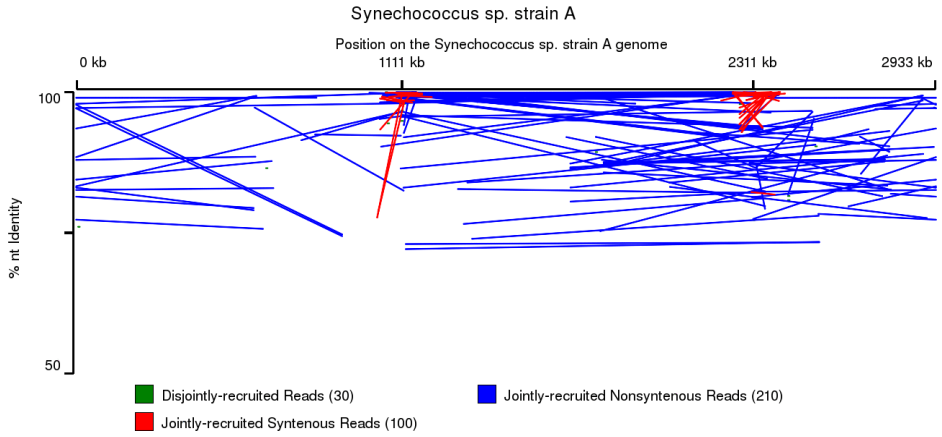



Figure 3.8. Distributions and percent nucleotide identity of jointly recruited syntenous (red) and nonsyntenous (blue) end sequences of BAC clones that contain A-like and B´-like 16S rRNA genes obtained from the (A) M60 (based on analysis of ~346 BAC clones) and (B) M65 libraries relative to homologs in the *Synechococcus* strain A and B´ genomes Lines connect end sequences of the same clone. X-axis ticks represent positions of the 16S rRNA genes.

genomes (Figure 3.8), with the exception that nearly all sequences interpreted as A´-like sequences (i.e., 10-20 % diverged from *Synechococcus* strain A homologs) were nonsyntenous (Figure 3.8B). Many nonsyntenous BACs seem to bridge between the two 16S rRNA loci in the reference genomes. The suspected A´-like nonsyntenous sequences also seemed to bridge between different genomic regions, but the location along the *Synechococcus* strain A genome was displaced, possibly suggesting different locations of the 16S rRNA loci in *Synechococcus* A´-like population genomes.

It has been suggested that the 16S rRNA-23S region may be considered a 'hot spot' for recombination, which may explain why many of the nonsyntenous clones span from one 16S rRNA gene region to the other, nearly 1 Mb away. This has been observed for many genomes, such as those of *Streptococcus anginosus, Ochrobactrum intermedium, Escherchia coli, Bradyrhizobium* sp., *Vibrio* sp. and *Haloarchaea* (Parker, 2001; Schouls et al., 2003; Teyssier et al., 2003; Hashimoto et al., 2003; Boucher et al., 2004; Lan and Reeves, 1998; Jensen et al., 2009; Papke et al., 2004). The frequency of transfer of the 16S rRNA genes is still uncertain, however if it were the case that such genes recombine frequently it would complicate phylogenetic tree construction using the 16S rRNA gene or genes within the 16S rRNA locus region depending on where the 'hot spot' is located. (Eardly et al., 1996; Wang and Zhang, 2000; Han et al., 2009).

Typically there is little or no variation between 16S rRNA sequence(s) within the same genome and it has been hypothesized that this occurs due to some form of homogenization that removes neutral mutations that would otherwise accumulate through a process known as 'concerted evolution' (Yap et al., 1999; Teyssier et al., 2003;

Hashimoto et al., 2003; Rudi et al., 1998 and Liao, 2000). Any recombination events that occur during concerted evolution are considered to occur via gene conversion, a non-reciprocal recombination event in which the sequence of one copy of the gene is converted to the sequence present in another copy. The ability of recombination to act as a homogenizing and diversifying force has been a topic of considerable interest in microbiology (Guttman and Dykhuizen, 1994; Papke et al., 2004; Han et al., 2009; Nystedt et al., 2008).

Mate pair analysis revealed a difference between randomly sampled BACs and BACS containing *Synechococcus* A- and B´-like 16S rRNA genes (Table 3.3). Most nonsyntenous BACs from the random selection had end sequences that were separated by longer or shorter distances than the average estimated clone insert size, suggesting inaccuracy in our +/- 30% criterion or differences from the reference genome due to insertions or deletions (mainly deletions). In contrast, BACs containing A- and B´-like *Synechococcus* 16S rRNA genes show an abundance of normal- and anti-normal long clones, which occur less frequently among randomly sampled BACs. Normal-long and anti-normal long mate pair categories are most associated with inversions (Rusch et al., 2007). As shown in Figure 3.9, most of these clones appear to bridge between the two-16S rRNA regions, suggesting that they may have resulted from genomic inversions involving genes in these genomic neighborhoods (Rusch et al., 2007). Inversions can occur via recombination, apparent 'flipping' of gene order mediated by transposons or mobile elements and recombination events that can occur during replication (Snyder and Champness, 2003; Tillier and Collins, 2000). The reverse gene order of the 16S-23S

rRNA genes themselves between the two 16S rRNA-23S regions seems to support a possible inversion event. In the *Synechococcus* strain A genome, the first 16S rRNA region located approximately at 5′ coordinate 1,110,781 has an order of 23S rRNA (5′ coordinate: 1,108,246) followed by the 16S rRNA. The second 16S rRNA region (located approx. at 5′ coordinate 2,310,964) has the reverse order with the 16S rRNA gene occurring first followed by the 23S rRNA gene (5′ coordinate: 2,313,600). This was also seen in the *Synechococcus* strain B′ genome.

Table 3.3. Distributions of mate-pair types of jointly-recruited end sequences selected from random and cyanobacterial (cyano) BAC clones.

| **Type**[1] | **M60 BAC Library** | | | | **M65 BAC Library** | |
|---|---|---|---|---|---|---|
| | ***Synechococcus* strain A** | | ***Synechococcus* strain B′** | | ***Synechococcus* strain A** | |
| | **Random (%)** | **Cyano (%)** | **Random (%)** | **Cyano (%)** | **Random (%)** | **Cyano (%)** |
| G-G | 50 | 39 | 27 | 11 | 2 | 32 |
| G-L | 7 | 18 | 12 | 16 | 7 | 16 |
| G-S | 34 | 1 | 18 | 0 | 73 | 3 |
| Anti-G | 0.5 | 0 | 2 | 0.6 | 1 | 1 |
| Anti-L | 0 | 27 | 11 | 24 | 5 | 23 |
| Anti-S | 1 | 0 | 1 | 0 | 1 | 2 |
| Nor-G | 0 | 0 | 2 | 1 | 0.4 | 0 |
| Nor-L | 3 | 15 | 14 | 35 | 6 | 15 |
| Nor-S | 0.5 | 0 | 0.6 | 0 | 0.8 | 0 |
| Out-G | 0 | 0 | 0.3 | 0.6 | 0.2 | 0.6 |
| Out-L | 0.5 | 0 | 10.4 | 11 | 4 | 7 |
| Out-S | 0 | 0 | 1 | 0 | 0.4 | 0 |

[1] G: Good, L: Long, S: Short, Anti: Antinormal, Nor: Normal, Out: Outie. For further explanation of mate pair types see Rusch et al. (2007).

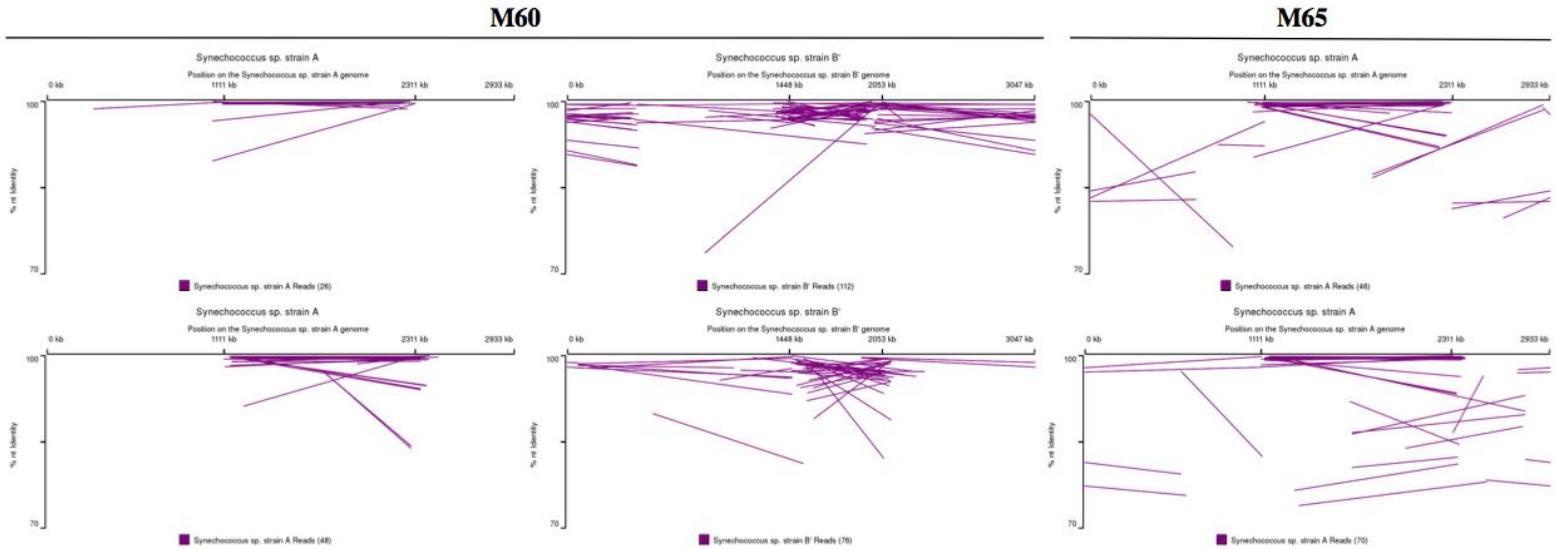



Figure 3.9. Distributions and percent nucleotide identity of jointly recruited normal- and anti-normal, long nonsyntenous sequences in BAC clones containing *Synechococcus* A/B´-lineage16S rRNA genes relative to the *Synechococcus* strain A and B´ genomes. Lines connect end sequences of the same clone. X-axis ticks represent the nucleotide position 0, position of the first and second 16S rRNA genes (1111 kb and 2311 kb for *Synechococcus* strain A and 1448 kb and 2058 kb for *Synechococcus* strain B´), and end of the genome position (2933 kb and 3047 for *Synechococcus* strains A and B´ respectively).

For the purposes of population genetics studies, it is important to note that protein-encoding loci useful for MLSA were found on both syntenous and nonsyntenous clones however; *Synechococcus* B´-like BAC clones (which had a higher percentage of nonsyntenous to syntenous clones) had fewer loci that could be used within an MLSA study (Melendrez et al., 2010b). The BAC libraries described here are a useful resource for cultivation-independent MLSA studies (Melendrez et al., 2010b; see Chapter 4). They have already been useful for linking genes encoding the type I reaction center in photosynthesis (*pscA*) to acidobacterial phylogenetic marker genes (*recA* and the 16S rRNA), leading to the discovery of the first phototrophic acidobacterium (Bryant et al., 2007). They should also be useful in the exploration of other community members, whose roles or capabilities remain unknown (Klatt et al., 2010). Construction of the BAC clone libraries has created a vast metagenomic sequence resource for the Mushroom Spring mat community. The BAC libraries constructed for this work are freely available for further research by contacting Amplicon Express (www.amplicon-express.com) and referencing this paper.

## Acknowledgements

We appreciate the long-term support from the National Science Foundation Frontiers in Integrative Biology Research Program (EF-0328698) and supplemental support from the NASA Exobiology program (NAG5-8807 and NX09AM87G) and the DOE Pacific Northwest National Laboratory (contract pending). In addition we appreciate the assistance from National Park Service personnel at Yellowstone National

Park. We would also like to acknowledge Keith Stormo, Amy Mraz, Robert Bogden and Quanzhou Tao at Amplicon Express, Dr. Mark Liles and Dr. James Elkins for their advice on lysis protocols for the isolation of HMW DNA for construction of the BAC clone libraries. Appreciation is also extended to Christian Klatt at Montana State University for assistance in metagenomic analysis and interpretation.

References

Allewalt JA, Bateson MM, Revsbech NP, Slack K and Ward DM. (2006). Temperature and light adaptations of *Synechococcus* isolates from the microbial mat community in Octopus Spring, Yellowstone National Park. *Appl Environ Microbiol* **72**: 544-550.

Becraft ED, Cohan FM, Kuhl M, Jensen S and Ward DM. (2010). Identifying and proving the existence of ecologically defined *Synechococcus* sp. in Mushroom Spring, Yellowstone National Park. In prep.

Beja O, Suzuki MT, Koonin EV, Aravind L, Hadd A, Nguyen LP, Villacorta R, Amjadi M, Garrigues C, Jovanovich SB, Feldman RA and DeLong EF. (2000a). Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ Microbiol* **2**: 516-529.

Beja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP, Jovanovich SB, Gates CM, Feldman RA, Spudich JL, Spudich EN and DeLong EF. (2000b). Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289**: 1902-1996.

Bhaya D, Grossman AR, Steunou AS, Khuri N, Cohan FM, Hamamura N, Melendrez MC, Bateson MM, Ward DM and Heidelberg JF. (2007). Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. *ISME J* **1**: 703-713.

Bilek N, Ison CA and Spratt BG. (2009). Relative contribution of recombination and mutation to the diversification of the *opa* gene repertoire of *Neisseria gonorrhoeae*. *J Bacteriol* **191**: 1878-1890.

Boucher Y, Douady CJ, Sharma AK, Kamekura M and Doolittle WF. (2004). Intragenomic heterogeneity and intergenomic recombination among Haloarchaeal rRNA genes. *J Bacteriol* **186**: 3980-3990.

Bryant DA, Costas AMG, Maresca JA, Chew AGM, Klatt CG, Bateson MM, Tallon L, Hostetler J, Nelson WC, Heidelberg JF and Ward DM. (2007). *Candidatus Chloracidobacterium thermophilum*: an aerobic phototrophic Acidobacterium. *Science*. **317**: 523-526.

Eardly BD, Wang FS and van Berkum P. (1996). Corresponding 16S rRNA gene segments in Rhizobiaceae and *Aeromonas* yield discordant phylogenies. *Plant Soil* **186**: 69-74.

Feil EJ, Maynard Smith J, Enright MC and Spratt BG. (2000). Estimating recombination parameters in *Streptococcus pneumoniae* from multi-locus sequence typing data. *Genet* **154**: 1439-1450.

Feil EJ, Maiden MCJ, Achtman M and Spratt BG. (1999). The relative contribution of recombination and mutation to the divergence of clones of *Neisseria meningitidis. Mol Biol Evol* **16**: 1496-1502.

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

Ferris MJ and Ward DM. (1997). Season distributions of dominant 16S rRNA-defined populations in a hot spring microbial mat examined by denaturing gradient gel electrophoresis. *Appl Environ Microbiol* **63**: 1375-1381.

Guttman DS and Dykhuizen DE. (1994). Clonal divergence in *Escherchia coli* as a result of recombination not mutation. *Science* **266**: 1380-1383.

Hashimoto JG, Stevenson BS and Schmidt TM. (2003). Rates and consequences of recombination between rRNA operons. *J Bacteriol* **185**: 966-972.

Han D, Fan Y and Hu Z. (2009). An evaluation of four phylogenetic markers in *Nostoc*; Implications for cyanobacterial phylogenetic studies at the intragenic level. *Curr Microbiol* **58**: 170-176.

Hanage WP, Fraser C and Spratt BG. (2006). Sequences, sequence clusters and bacterial species. *Phil Trans Roy Soc B* **361**: 1917-1927.

Jensen S, Frost P and Torsvik VL. (2009). The nonrandom microheterogeneity of 16S rRNA genes in *Vibrio splendidus* may reflect adaptation to versatile lifestyles. *FEMS Microbiol Lett* **294**: 207-215.

John DE, Wawrik B, Tabita FR and Paul JH. (2006). Gene diversity and organization in *rbcL*-containing genome fragments from uncultivated *Synechococcus* in the Gulf of Mexico. *Mar Ecol Ser* **316**: 23-33.

Klatt CG, Wood JM, Rusch DB, Bateson MM, Heidelberg JF, Bryant DA and Ward DM. (2010). Composition of metagenomes from a phototrophic hot spring microbial mat community. In prep.

Lan R and Reeves PR. (1998). Recombination between rRNA operons created most of the ribotypes variation observed in the seventh pandemic clone of *Vibrio cholerae*. *Microbiol* **144**: 1213-1221.

Liao D. (2000). Gene conversion drives within genic sequences: concerted evolution of ribosomal RNA genes in bacteria and archaea. *J Mol Evol* **51**: 305-317.

Liles MR, Williamson LL, Rodbumrer J, Torsvik V, Goodman RM and Handelsman J. (2008). Recovery, purification and cloning of high-molecular-weight DNA from soil microorganisms. *Appl Environ Microbiol* **74**: 3302-3305.

Liles MR, Manske BF, Bintrim SB, Handelsman J and Goodman RM. (2003). A census of rRNA genes and linked genomic sequences within a soil metagenomic library. *Appl Environ Microbiol* **69**: 2684-2691.

Melendrez MC, Lange RK, Cohan FM and Ward DM. (2010a). Ecological diversity of *Synechococcus* spp. inhabiting an alkaline siliceous hot spring in Yellowstone National Park, WY measured using protein-encoding genes and evolutionary simulation. *ISME J* In prep.

Melendrez MC, Cohan FM and Ward DM. (2010). Cultivation-independent multi-locus sequence analysis of *Synechococcus* populations inhabiting a hot spring microbial mat. In prep.

Nübel U, Garcia-Pichel and Muyzer G. (1997). PCR primers to amplify 16S rRNA genes from cyanobacteria. *Appl Environ Microbiol* **63**: 3327-3332.

Nystedt B, Frank AC, Thollesson M and Anderson SGE. (2008). Diversifying selection and concerted evolution of a type IV secretion system in *Bartonella. Mol Biol Evol* **25**: 287-300.

Oyaung Y, Dai S, Xie L, Ravi Kumar MS, Sun W, Sun H, Tang D and Li X. (2009). Isolation of high molecular weight DNA from marine sponge bacteria for BAC library construction. *Mar Biotechnol* Aug 15.

Papke RT, Keonig JE, Rodriguez-Valera F and Doolittle WF. (2004). Frequent recombination in a saltern population of *Halorubrum*. *Science* **306**: 1928-1929.

Papke RT, Ramsing NB, Bateson MM and Ward DM. 2003. Geographical isolation in hot spring cyanobacteria. *Environ Microbiol* **5**: 650-659.

Parker MA. (2001). Case of localized recombination in 23S rRNA genes from divergent *Bradyrhizobium* lineages associated with neotropical legumes. *Appl Environ Microbiol* **67**: 2076-2082.

Ramsing NB, Ferris MJ and Ward DM. (2000). Highly ordered vertical structure of *Synechococcus* population within the one-millimeter-thick photic zone of a hot spring cyanobacterial mat. *Appl Environ Microbiol* **66**: 1038-1049.

Riesenfeld CS, Goodman RM and Handelsman J. (2004). Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environ Microbiol* **6**: 981-989.

Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR, Loiacono KA, Lynch BA, MacNeil IA, Minor C, Tiong CL, Gilman M, Osburne MS, Clardy J, Handelsman J and Goodman RM. (2000). Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl Environ Microbiol* **66**: 2541-2547.

Rondon MR, Raffel SJ, Goodman RM and Handelsman J. (1999). Toward functional genomics in bacteria: analysis of gene expression in *Escherchia coli* from a bacterial artificial chromosome library of *Bacillus cereus. Proc Natl Acad Sci* **96**: 6451-6455.

Rudi K, Skulberg OM and Jakobsen KS. (1998). Evolution of cyanobacteria by exchange of genetic material among phyletically related strains. *J Bacteriol* **180**: 3453-3461.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Lil K, Kravitz S, Heidelberg JF, Utterback T, Rogers Y-H, Falcon LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M and Venter JC. (2007). The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLOS Biol* **5**: 398-431.

Schouls LM, Schot CS and Jacobs JA. (2003). Horizontal gene transfer of segments of the 16S rRNA genes between species of the *Streptococccus anginosus* group. *J Bacteriol* **185**: 7241-7246.

Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, Tachiri Y and Simon M. (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherchia coli* using an F-factor-based vector. *Proc Natl Acad Science* **89**: 8794-8797.

Snyder L and Champness W. (2003). *Molecular Genetics of Bacteria*. 2nd ed. ASM Press: Washington D.C. Pp. 133-135 and 307.

Stein JL, Marsh TL, Wu KY, Shizuya H and DeLong EF. (1996). Characterization of uncultivated prokaryotes: isolation and analysis of 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J Bacteriol* **178**: 591-599.

Tanabe Y, Sano T, Kasai F and Watanabe MM. (2009). Recombination, cryptic clades and neutral molecular divergence of the microcystin synthetase (mcy) genes of the toxic cyanobacterium *Microcystis aeruginosa*. *BMC Evol Biol* **9**: 115.

Tanabe Y, Kasai F and Watanabe MM. (2007). Multilocus sequence typing (MLST) reveals high genetic diversity and clonal population structure of the toxic cyanobacterium *Microcystis aeruginosa. Microbiol* **153**: 3695-3703.

Tao Q, Wang A and Zhang HB. (2002). One large-insert plant-transformation-competent BIBAC library and three BAC libraries of Japonica rice for genome research in rice and other grasses. *Theor Appl Genet* **105**: 1058-1066.

Teyssier C, Marchandin H, De Buochberg MS, Ramuz M and Jumas-Bilak E. (2003). Atypical 16S rRNA gene copies in *Ochrobactrum intermedium* strains reveal large genomic rearrangements by recombination between *rrn* copies. *J Bacteriol* **185**: 2901-2909.

Tillier ERM and Collins RA. (2000). Genome rearrangement by replication-directed translocation. *Nature Genet* **26**: 195-197.

Turner KME and Feil EJ. (2007). The secret life of the multilocus sequence type. *Int J Antimicrob Agents* **29**: 129-135.

Van Valen L. (1976). Ecological species, multispecies, and oaks. *Taxon* **25**: 233-239.

Vitorino LR, Margos G, Feil EJ, Collares-Pereira M, Ze-Ze L and Kurenbach K. (2008). Fine-scale phylogeographic structure of *Borrelia lusitaniae* revealed by multilocus sequence typing. *PLOS ONE* **3**: 1-13.

Wang Y and Zhang Z. (2000). Comparative sequence analyses reveal frequent occurrence of short segments containing an abnormally high number of non-random base variation in bacterial rRNA genes. *Microbiol* **146**: 2845-2854.

Ward DM, Bateson MM, Ferris MJ, Kuhl M, Wieland M, Koeppel A and Cohan FM. (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. *Phil Trans Roy Soc Ser B* **361**: 1997-2008.

Ward DM. (1998). A natural species concept. *Curr Opin Microbiol* **1**: 271-277.

Williamson LL, Borlee BR, Schloss PD, Guan C, Allen HK and Handelsman J. (2005). Intracellular screen to identify metagenomic clones that induce or inhibit a quorum-sensing biosensor. *Appl Environ Microbiol* **71**: 6335-6344.

Wu D, Raymond J, Wu M, Chetterji S, Ren Q, Graham JE, Bryant DA, Robb F, Colman A, Tallon LJ, Badger JH, Madupu R, Ward NL and Eisen JA. (2009). Complete genome sequence of the aerobic CO-oxidizing thermophile *Thermomicrobium roseum. PLOS One* **4**: e4207.

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

Zhang HB, Choi S, Woo SS, Li Z, and Wing RA. (1996). Construction and characterization of two rice bacterial artificial chromosome libraries from the parents of a permanent recombinant inbred mapping population. *Mol Breed* **2**: 11-24.

# CHAPTER 4

# CULTIVATION-INDEPENDENT MULTI-LOCUS SEQUENCE ANALYSIS OF *SYNECHOCOCCUS* POPULATIONS INHABITING A HOT SPRING CYANOBACTERIAL MAT

Melanie C. Melendrez[7], Fredrick M. Cohan[8] and David M. Ward[7]

## Abstract

A high-resolution, theory-based, cultivation-independent, multi-locus population genetics method was developed to test the hypothesis that hot spring microbial mats contain more *Synechococcus* ecotypes than were detected by 16S-23S rRNA internal transcribed spacer sequences or single protein-encoding loci and to provide a means of assaying the impact of recombination on these ecotype populations. The genomes of *Synechococcus* strains A and B´ and mat metagenomes were examined to identify protein-encoding loci near 16S rRNA genes useful for multi-locus sequence analysis (MLSA). Bacterial artificial chromosome (BAC) libraries were constructed to sample MLSA loci simultaneously from Mushroom Spring mats collected at 60° and 65°C. Oligonucleotide probing, PCR and sequencing were used to identify BACs containing a *Synechococcus* A/B lineage-specific 16S rRNA gene and to assay MLSA loci. Co-occurrence of MLSA loci on BACs decreased as a function of their separation in the *Synechococcus* strain A and B´ genomes. Ecotype simulation analysis (ES) of concatenated genes from BACs with A-like or B´-like 16S rRNA sequences revealed putative ecotype clades containing a dominant allelic variant and several closely related

[7] Land Resources and Environmental Science, Montana State University, Bozeman, MT
[8] Department of Biology, Wesleyan University, Middletown, CT

singletons. Separate analyses of a subset of the data permitting equal sampling of clones from the two samples suggested that putative ecotypes predicted by ES were, to a high degree, sample-specific. eBURST, which groups variants based on identity at all but a single locus, without regard for how different the variants are phylogenetically at any locus, predicted fewer clonal complexes, which were less convincingly sample-specific. There were many indications that recombination has molded the evolution of these populations more than mutation, including phylogenetic incongruency, the number and distribution of single-nucleotide polymorphisms and results from numerous algorithms designed to detect recombination. Nevertheless, recombination appears to have been insufficiently intense to erode the detection of species-like populations, although it may have relocated alleles of some variants of an ecotype in terms of the phylogeny of specific genes or limited sets of genes.

## Introduction

We are investigating the species composition of cyanobacterial mats inhabiting alkaline siliceous hot springs in Yellowstone National Park (YNP). Previous molecular analyses have shown that the dominant *Synechococcus* populations native to mats in Octopus Spring and Mushroom Spring, Yellowstone National Park, WY, are genetically distinct from readily cultivable *Synechococcus* strains (Ward et al., 1990; Ward et al., 1998; Ferris et al., 2003). The native *Synechococcus* populations, referred to as A/B-type *Synechococcus*, have very similar 16S rRNA sequences that are distributed uniquely along the flow, and thus, temperature gradient in the effluent channels (Ferris and Ward,

1997). Specifically, genotypes A′′, A′, A, B′ and B occur at progressively lower temperatures, respectively, from the upper temperature limit at ~72°C to ~50°C. This suggested that distinct 16S rRNA sequences might represent populations with distinct ecologies (ecotypes), such as adaptation to different temperatures (demonstrated for relevant isolates by Allewalt et al., (2006)) and/or other parameters that change with flow away from the spring (see Bhaya et al., 2007; Adams et al., 2008). The existence of phenotypically distinct *Synechococcus* populations at different depths within the upper 1-3 mm depth interval of these mats suggested possible adaptations to light and/or other parameters that may vary in relation to depth in the mat (Ramsing et al., 2000; Ward et al., 2006). Ecologically adapted populations can be considered ecological species (Ward et al., 1998, 2006); however, detection of ecological species appears to require analysis of molecular markers offering greater molecular resolution than the slowly evolving 16S rRNA gene (Ferris et al., 2003; Ward et al., 2006; Chapter 2).

When molecular resolution is sufficient to detect individual variants, an evolutionary simulation program called Ecotype Simulation (ES), which is based on the Stable Ecotype Model of species and speciation (Koeppel et al., 2008; Ward et al., 2006), can model ecological species. ES predicts how individual genetic variants are grouped into putative ecotypes (PEs), which we equate with ecological species. ES defines ecotypes more precisely as monophyletic populations whose members are ecologically interchangeable, not just on the basis of unique ecological distribution, as above. ES analysis of more highly resolving protein-encoding loci predicted more species (i.e., 4 to 14 and 8 to13 in A-like and B′-like *Synechococcus* populations, respectively), which

appear to be ecologically distinct based on distribution analyses (Melendrez et al., 2010a; see Chapter 2; Becraft et al., 2010). However, the number of species predicted depends on the molecular resolution of the locus used (Melendrez et al., 2010a; see Chapter 2).

A major shortcoming of single-locus analyses (SLA) is that they cannot account for homologous recombination, which has been shown to be of major importance to the evolution of bacterial diversity (Hanage et al., 2006; Bilek et al., 2009; Tanabe et al., 2009). To date, suspected recombinant sequences have simply been removed from ES analyses, because these recombinant sequences do not reflect the history of the organism at the locus undergoing recombination (Koeppel et al, 2008). Multi-locus sequence analysis (MLSA), which was developed for population genetics studies of cultivated pathogenic isolates (Feil et al., 1999, 2000; Vitorino et al., 2008; Dingle et al., 2005; Salerno et al., 2007; Cesarini et al., 2009) and has been applied to nonpathogenic isolates (Koeppel et al., 2008; Whitaker et al., 2005; Papke et al., 2004) as well, allows for the independent assortment (i.e., recombination) of alleles. MLSA entails the analysis of multiple 'housekeeping' loci that have not been subjected to positive selection pressure. In the established, isolate-based MLSA approach, unique sequences from individual loci (allele types) are combined into an allelic profile and unique allelic profiles are considered sequence types (STs) of isolates. An algorithm, called eBURST, is then used to group the STs into clonal complexes, which are comprised of variants that are identical at all loci (the consensus ST) and variants differing at one (single-locus variants, or SLV) or two (double-locus variants or DLV) of the loci in the allelic profile when compared to the consensus ST (Feil et al., 2004). The degree of sequence difference is not taken into

account, as either mutation or recombination may have caused the difference. Since eBURST allows for independent assortment of alleles in a multi-locus sequence dataset, it also increases resolution by increasing the number of possible combinations of alleles.

An alternative MLSA approach involves phylogenetic analysis of concatenated sequences from multiple loci, which "buffers" against the effect of recombination (Hanage et al., 2005). Concatenated MLSA analysis relies on loci being selected that exhibit little or no signals of recombination. If some loci are experiencing frequent recombination, concatenated sequence analysis will 'buffer' the locus experiencing frequent recombination by keeping it (theoretically) within the phylogenetic clade defined by the cohesive power of the other loci not undergoing recombination. If all loci selected for MLSA are experiencing rampant recombination then this analysis would be problematic. MLSA may also increase resolution as it includes sequence variation in multiple loci, but this is dependent on the molecular resolution of the loci selected. Furthermore, if the majority of the loci selected are conserved, the phylogenetic analysis may be skewed in a more conservative direction and not give an accurate picture of an organism's evolutionary history.

While isolate-based studies can provide valuable population biology insights, insights are limited to these isolates and cannot necessarily be extrapolated to the natural population, particularly in taxa where cultivation of organisms is extremely difficult. The challenge in MLSA of largely uncultivated natural populations is that, without cultivation of isolates, another means is needed to obtain multiple genes that are linked. We developed a cultivation-independent MLSA approach based on the use of bacterial

artificial chromosome (BAC) libraries to sample multiple genes from individual genomes of mat inhabitants (Melendrez et al., 2010b; see Chapter 3). BACs with average insert sizes ranging from 90-120 kb were screened to identify clones containing a *Synechococcus* A/B´ lineage-specific 16S rRNA region. Multiple loci were selected and PCR-amplified from these BAC clones and the sequence data obtained were analyzed using concatenated phylogenies, single gene phylogenies, eBURST and ES analysis.

ES analysis of concatenated loci predicted more PEs than were predicted from single loci, and many PEs were sample-specific, suggesting their unique ecological character. eBURST analysis predicted fewer clonal complexes, which were less convincingly sample-specific. PEs and clonal complexes that were centered on the same dominant variant sequences, contained different sub-dominant variants because of differences in the way the two analytical approaches work. Multiple lines of evidence suggested that recombination has been more important than mutation in generating variation in these lineages, but the impact of recombination has been insufficient to erode the existence and detection of PEs.

## Methodology

Study sites and sample collection were previously described in Melendrez et al. (2010a) (also see Chapter 2). Briefly, samples were collected on 2 October 2003 from the effluent channel of Mushroom Spring (44.5386°N, 110.7979°W) an alkaline siliceous hot spring in the Lower Geyser Basin of Yellowstone National Park, WY, at two different temperatures, 60°C and 65°C (samples designated M60 and M65 respectively). DNA was

extracted using a gentle in-gel lysis protocol previously described in Melendrez et al. (2010b) (also see Chapter 3) to isolate high molecular weight DNA for the construction of metagenomic BAC libraries by Amplicon Express, Pullman, WA). The libraries were screened using radioactive oligonucleotide probing; PCR amplification and sequencing to determine which BACs contained a *Synechococcus* A/B´ lineage specific 16S rRNA gene (Melendrez et al, 2010b; Chapter 3). These BACs were rearrayed by Amplicon Express and sent to Montana State University for MLSA in 96-well format (200 μl of BAC clone suspension per well). The DNA used for construction of these metagenomic BAC libraries was also used for studies on the 16S rRNA and ITS region (Ward et al., 2006), some single protein-encoding loci (*rbsK, aroA* and *apcAB*) (Melendrez et al., 2010a; Chapter 2) and for construction of small-insert (2-12 kb) metagenomic libraries (Bhaya et al., 2007; Klatt et al., 2010).

Locus Selection

From one hundred genes that were upstream and downstream of both 16S rRNA genes in the *Synechococcus* strain A and B´ genomes, loci were selected based on the following preferred characteristics: (i) presence in both genomes, (ii) range of distances from the 16S rRNA locus (iii) high degree of nucleotide divergence between *Synechococcus* strain A and B´ homologs, (iv) high average nucleotide divergence and variance of metagenomic homologs from *Synechococcus* strain A and B´ homologs, (v) not under positive evolutionary selection (dN/dS values < 1) and (vi) functionally useful for gene expression studies (see Melendrez et al., 2010a; Chapter 2). In addition loci

adjacent to transposons or mobile elements were avoided due to the greater possibility of co-migration of adjacent loci. Loci used in this study are presented in Tables 4.1 and 4.2.

PCR Amplification and Sequencing of BAC MLSA Loci

Primers for the separate amplification of *Synechococcus* A-like and B´-like BAC target genes (Tables 4.1 and 4.2) were designed using the primer design tool in SciTools on the Integrated DNA Technologies website (http://www.idtdna.com/Scitools/Applications/Primerquest/), analyzed by BLAST against the NCBI nr database (default parameters) to assure specificity to *Synechococcus* strain A and B´ genomes (Altschul et al., 1990) and obtained from Integrated DNA Technologies (San Diego, CA). Primers were dissolved in water to a final concentration of 50 μM for use in PCR analysis. PCR reaction mixtures contained 2 μl of BAC clone, 10mM dinucleotide triphosphates (Promega), Taq buffer (Promega), Taq Gold Polymerase (Promega), 25mM $MgCl_2$ and primers (50 μM). Cycling conditions for the *dnaG, pcrA, ispE, sufB* and *argD* genes were the same used for *aroA* and *rbsK* genes, and cycling conditions for the protein kinase (*PK*), *lepB* and *accC* genes were the same as for the *apcAB* gene, as described in Melendrez et al. (2010a) (see Chapter 2), except for the use of 40 cycles instead of 30 cycles. Cycling conditions for *hisF* and the conserved hypothetical protein (*CHP*) gene were: an initial denaturing step at 94°C (2 min) followed by 20 cycles of 94°C (1 min), 60°C (1 min) and 72°C (1 min); then 20 cycles of 94°C (1 min), 55°C (1 min) and 72°C (1 min) with a final extension at 72°C for 10 min and storage at 4°C. After verifying sizes

Table 4.1. Characteristics of loci used in analysis of A-like *Synechococcus* mat populations.

| Locus | Genomic annotation[a] | Primer Sequence (5′-3′) | Gene Length | Amplified fragment length | Distance from the 16S rRNA locus (kb) | *Synechococcus* strain A/B′ homolog divergence | No. of Metagenomic Sequences | Average % Divergence of Metagenomic and A-like homologs | Variance of Metagenomic and A-like homologs | dN/dS ratio[f] |
|---|---|---|---|---|---|---|---|---|---|---|
| *apcAB* | Allo-phycocyanin alpha/beta subunits and IS region[b] | Melendrez et al., 2010a | 972 | 500 | 82 | 8.5 | 7 | 0.1 | 0.01 | 0.01 |
| *rbsK* | Ribokinase | Melendrez et al., 2010a | 930 | 583 | 50 | 18.2 | 3 | [d] | [d] | 0.08[e] |
| *PK* | Hypothetical protein kinase | *PK*F:atccatgcccttcgcttggaaac *PK*R:tgacttcagcggtagaaatcggct | 2166 | 666 | 39 | ND | 5 | 1.7 | 1.5 | 0.96 |
| *hisF* | Imidazole-glycerol phosphate synthase, cyclase subunit | hisFF:ccactcacgaagagcgggaaatct hisFR:ggcaattgtcaactggccgtagtg | 762 | 471 | 23 | 11.5 | 7 | 0.4 | 0.01 | 0.71 |
| *lepB* | Signal peptidase I | lepBF:gagaacttgctgacagtggtgctg lepBR:gatactctggcggggtgaaaaagt | 687 | 441 | 28 | 14.5 | 2 | [d] | [d] | 0.66[e] |
| *CHP* | Conserved hypothetical protein | *CHP*F:tcgaggacatgaagggccaaatct *CHP*R:aaatgggtcgtcaaagccgtttcc | 2106 | 680 | 15 | [c] | 9 | 1.8 | 1.9 | 0.09 |
| *aroA* | 3-phospho-shikimate 1-carboxyvinyl-transferase | Melendrez et al., 2010a | 1335 | 654 | 38 | 15.1 | 6 | 0.3 | 0.05 | 0.82 |
| *dnaG* | DNA primase | dnaGF:cacaacgaccacaagcccagcttt dnaGR:ttccgtttccggggagttgaggta | 1896 | 525 | 24 | 12 | 10 | 0.3 | 0.02 | 0.88 |

[a] Links to genomic annotations for *Synechococcus* strain A and B′: http://cmr.jcvi.org/cgi-bin/CMR/GenomePage.cgi?org=gyma and http://cmr.jcvi.org/cgi-bin/CMR/GenomePage.cgi?org=gymb.

[b] The *apcAB* amplified product contained 380 bp from *apcA,* 60 bp of *apcB* and 60 bp of the internal sequence in between.

[c] The conserved hypothetical protein does not contain a homolog in the *Synechococcus* strain B′ genome.

[d] Number of metagenomic sequence < 5 so this statistic was not calculated.

[e] Calculated based on < 5 sequences

Table 4.2. Characteristics of loci used in MSLA analysis of B´-like *Synechococcus* populations.

| Locus | Genomic annotation[a] | Primer Sequence (5´-3´) | Gene Length | Amplified fragment length | Distance from the 16S rRNA locus (kb) | *Synechococcus* strain A/B´ homolog divergence | No. of Metagenomic Sequences | Average % Divergence of Metagenomic and B´-like homologs | Variance of Metagenomic and B´-like homologs | dN/dS ratio |
|---|---|---|---|---|---|---|---|---|---|---|
| *aroA* | 3-phospho-shikimate 1-carboxyvinyl-transferase | Melendrez et al., 2010a | 1305 | 684 | 52 | 15.1 | 10 | 3.5 | 2.0 | 1.1 |
| *rbsK* | Ribokinase | Melendrez et al, 2010a | 918 | 822 | 24 | 18.2 | 14 | 3.1 | 4.0 | 0.7 |
| *pcrA* | ATP-dependent DNA helicase | pcrAF:attgcctacc tggttcgccactat pcrAR:tgacgggtt ggcctgagaacttta | 2364 | 525 | 8 | 13.7 | 19 | 2.5 | 2.1 | 1.0 |
| *ispE* | 4-diphospho-cytidyl-2C-methyl-D-erythritol kinase | ispEF:gcgatggtgt tgcagagcattcat ispER:aaccacctgt tccaagtcattgcg | 936 | 499 | 14 | 19 | 7 | 2.6 | 2.1 | 0.2 |
| *accC* | acetyl-CoA carboxylase, biotin carboxylase | accCF:ggttcttggc ggagaatgccaaat accCR:aatcagatc cagccctgtcaccat | 1356 | 580 | 15 | 8.1 | 11 | 2.5 | 2.8 | 0.04 |
| *sufB* | FeS assembly protein | sufBF:agtcgattgg caaaggcttgaacg sufBR:tcttggagttt gctccctgacaca | 1437 | 679 | 27 | 5.0 | 15 | 2.9 | 2.7 | 0.1 |
| *argD* | Acetyl-ornithine transaminase | argDF:aagccaatg agggagccattaagc argDR:gctgggcat tccaaaccaaactct | 1266 | 551 | 38 | 13.8 | 13 | 0.99 | 1.3 | 0.1 |
| *apcAB* | Allo-phycocyanin alpha/beta subunits and IS region[b] | Melendrez et al., 2010a | 972 | 498 | 61 | 8.5 | 11 | 1.8 | 2.2 | 0.1 |

[a] Links to genomic annotations for *Synechococcus* strain A and B´: http://cmr.jcvi.org/cgi-bin/CMR/GenomePage.cgi?org=gyma and http://cmr.jcvi.org/cgi-bin/CMR/GenomePage.cgi?org=gymb.

[b] The *apcAB* amplified product contained 380 bp from *apcA,* 55 bp of *apcB* and 63 bp of the internal sequence in between.

of PCR products using agarose gel electrophoresis analysis, products were purified using a MiniElute 96 UF PCR purification kit (Qiagen) or a Qiaquick PCR purification kit (Qiagen) according to manufacturers instructions and 3 μl of purified product was used for sequencing reactions. BAC clones exhibiting no PCR product for any gene on first analysis were re-amplified twice more to increase confidence that the gene was absent.

Sequencing

Purified PCR products were sequenced using the forward and reverse primers for each gene described in Tables 4.1 and 4.2 and the BigDye v.3.1 cycle sequencing kit (Applied Biosystems) at the University of Nevada-Reno Sequence Center (Reno, NV). The sequences will be submitted to GenBank.

Sequence Alignment and Phylogenetic Analysis

Bidirectional sequence data for each locus was analyzed using Sequencher v.4.8 to check automated base calls against chromatograms. Clean sequence data were analyzed with NCBI-BLAST and the top BLAST match was determined (with *Synechococcus* strains A, B′ or neither). Alignments were made using the ClustalW algorithm in the MEGA4 software (Tamura et al., 2007) and the *Synechococcus* strain A and B′ genomes as references. Alignments were manually checked to ensure accuracy and ends were trimmed so that all sequences were compared in the same region. Neighbor-joining trees were constructed and bootstrapped (2000 replicates for individual and concatenated loci) using MEGA v4.0 software, viewed and manipulated using TreeView and CTree v1.02 (Page, 1996; Archer and Robertson, 2007). Sequences of

single loci used in MLSA were concatenated using a custom perl script (Jason Wood, personal communication) and uploaded in MEGA4, to ensure that concatenated sequences remained aligned and neighbor-joining trees were constructed. Estimates of average evolutionary divergence (EED) for single and multi-locus phylogenies were computed over all sequence pairs using the Maximum Composite Likelihood method in MEGA4 as previously described (Melendrez et al., 2010a; Chapter 2; Tamura et al., 2004, 2007).

Ecotype Simulation and Demarcation

Concatenated and single-locus sequence alignments were analyzed using ES to predict the number of PEs (n), rates of periodic selection (sigma, PS), ecotype formation (omega, EF) and 95% confidence intervals (CI) for all parameters at 1.5x precision match between observed and simulated data. Ecotypes were manually demarcated conservatively as previously described (Melendrez et al., 2010a; Chapter 2; Cohan and Perry, 2007; Koeppel et al., 2008) (http://fcohan.web.wesleyan.edu/ecosim/).

Multi-Locus and EBurst Analyses

Alignments for each gene used in MLSA were uploaded into Sequencher v 4.8 and were organized into groups that were 100% identical at the nucleotide level. Allele types were assigned for each unique sequence and were used to generate allelic profiles; BACs with identical allelic profiles at all loci were assigned as unique STs, as described in Tables C4.1, C4.2 and C4.3. Allelic profiles and their unique ST designation were uploaded into eBURST (Feil et al., 2004; Spratt et al., 2004) and population snapshots

were generated to view A-like and B´-like *Synechococcus* diversity (Figure C4.1). Clonal complexes were defined as a consensus group of BACs and at least 3 SLVs, as suggested by Feil et al. (2004). Population snapshots to visualize clonal complexes with less stringent criteria were generated by defining clonal complexes as a consensus group with at least 2 SLVs and allowing for DLVs to be included in the clonal complex.

Single Nucleotide Polymorphism Analysis

Single nucleotide polymorphisms were analyzed comparing sequences of PE clade variants (PEVs) and SLVs with that of the dominant variant (DV) of the same PE clade or consensus group of the same clonal complex (often the same as the DV) using the Perl program Pigeon. Pigeon reads in the FASTA file and compares each nucleotide in the DV to the corresponding nucleotide in each PEV (or SLV), locating and reporting the position of SNPs (Jason Wood, personal communication).

Linkage Disequilibrium

The standardized index of association ($I_A^S$), which measures the degree of association between alleles at different loci based on the variance in genetic distance between genotypes, was determined with LIAN version 3.5 (Haubold and Hudson, 2000). $I_A^S$ values were calculated from allelic profiles generated from analysis of MLSA data for *Synechococcus* A-like (7 loci) and B´-like (4 loci; see explanation below) populations. The significance of $I_A^S$ was determined by comparison to the null hypothesis of free recombination simulated by 10,000 randomized reshufflings of alleles for each locus among individuals.

Detection of Recombination Signals

All loci for A-like and B´-like *Synechococcus* populations were tested for putative recombination events using Recombination Detection Program, version 3 (RDP3, Martin et al., 2005). The loci were tested individually and as concatenated sequences. In the RDP suite of programs a number of different methods are implemented. The methods used for recombination detection in this study included the RDP method (Martin et al., 2005), GENECONV (Padidam et al., 1999), Maximum Chi Square (MaxChi, Maynard-Smith, 1992; Posada and Crandall, 2001), Chimaera (Posada and Crandall, 2001), Sister Scanning (Siscan, Gibbs et al., 2000), and 3SEQ (Boni et al., 2007) and Likelihood Assisted Recombination Detection (LARD, Holmes et al., 1999), which constitute the most powerful methods currently available (Vitorino et al., 2008; RDP3 manual)(see Appendix C for descriptions of these methods). The general settings within RDP3 were as follows: the highest acceptable p-value was set to 0.05 with Bonferroni corrections. For the individual methods default parameters. In RDP the window size was set to 30 as recommended (RDP3 Instruction Manual). In MaxChi and Chimaera the 'variable window size' was set. In Siscan the window size was set to 200 bp with a step size of 20. GENECONV was used with default parameters. Recombination signals were considered present if they could be detected by at least 3 methods within the RDP3 package at significant p-values according to the suggestion of the RDP3 manual and Vitorino et al. (2008). The patterning of PEV and SLV SNPs against the DV sequence was also considered in interpreting possible evidence of recombination events.

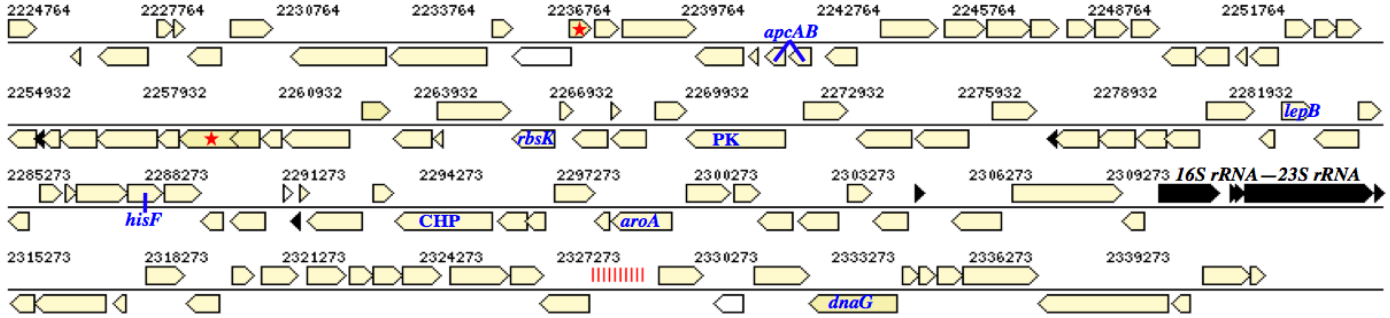
Estimates of the per-locus population recombination parameter ($\rho$) and the average per-site population mutation rate (Waterson's $\theta$) were determined using the recombination rate plot option in the RDP3 package (McVean et al., 2004) with default parameters. A recombination/mutation ratio ($\rho/\theta$) was calculated from the LDHat software (convert, interval and associated lookup tables) embedded within the RDP3 software suite and was run through 1,000,000 updates of Markov Chain Monte Carlo analysis, as recommended in the RDP3 manual. The recombination rate calculated by LDHat assumes a constant recombination rate over the region and a gene conversion model.

## Results

### Characteristics of Loci Selected

Eight hundred protein-encoding loci in the vicinity of the two 16S rRNA loci of the *Synechococcus* strain A and B′ genomes were screened for possible use. Loci within one of these regions in A-like or B′-like populations were selected, as described in Tables 4.1 and 4.2 and their locations are mapped against the references genomes in Figure 4.1. All loci were present in one copy in the genomes of *Synechococcus* strains A and/or B′, and ranged in distance from the 16S rRNA locus from 15-82 kb in the A genome and 8-61 kb in the B′ genome. Average divergence of metagenomic homologs relative to the reference genome ranged from 0.1% to 1.8% for A-like and 0.99% to 3.5% for B′-like variants, with variances ranging from 0.01 to 1.9 for A-like and 1.3 to 4.0 for B′-like variants. All loci were under neutral or purifying selection (dN/dS <1). The *ispE* and

(A)




(B)

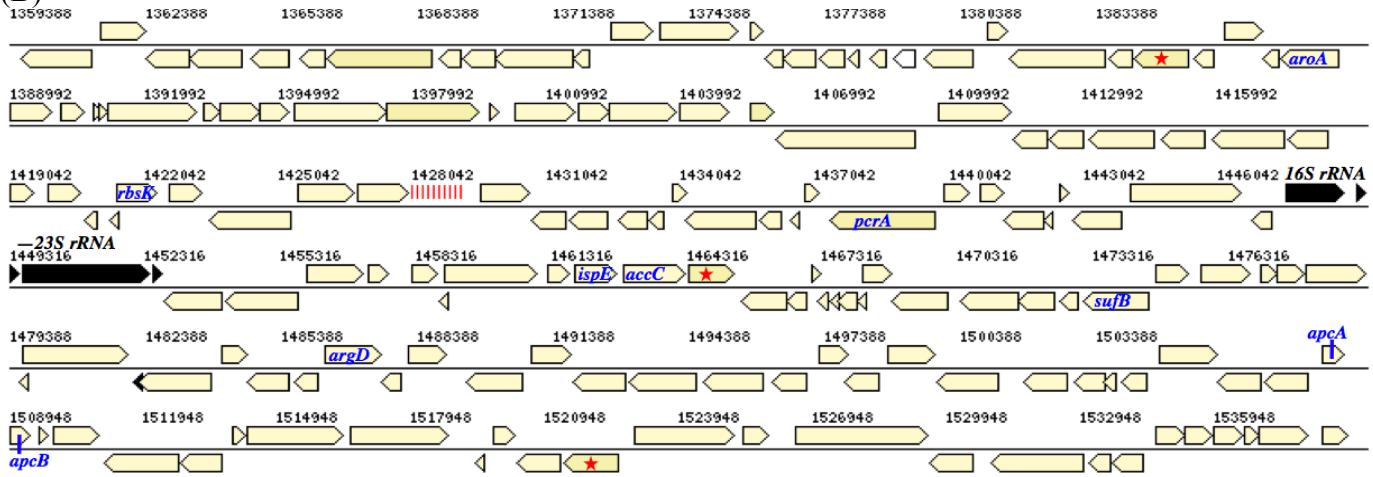



Figure 4.1. Genome region views of MLSA loci (A) *Synechococcus* strain A, coordinates 2,224,764—2,339,273 and (B) *Synechococcus* strain B´, coordinates 1,359,388—1,535,948. Loci used in this study are highlighted in blue. Open reading frames annotated as transposons, resolvases, recombinases and mobile elements are highlighted with red stars. Red vertical bars show presence of CRISPR arrays. (adapted from JGI genome region viewer, http://imgweb.jgi-psf.org/cgi-bin/w/main.cgi, Markowitz et al, 2008, 2009).

*accC* loci in *Synechococcus* strain B′, but not other loci, are adjacent to resolvases, recombinases or other mobile elements (Figure 4.1).

Characterization of BAC Clone Libraries

Of 368,640 BAC clones screened (304,128 from the M60 BAC library and 64,512 from M65), 1429 (1,193 from M60 and 236 from M65) contained a *Synechococcus* A/B′ lineage-specific 16S rRNA gene and these were screened for MLSA loci. 237 *Synechococcus* B′-like BACs and 267 *Synechococcus* A-like BACs contained a 16S rRNA locus and at least one MLSA locus. This was less than the 50% expected given two 16S rRNA loci per genome. The difference was most likely due to insufficient PCR product for sequencing of the 16S rRNA gene because, in preliminary work, 44-47% of 16S rRNA-bearing BACs could be identified as being either *Synechococcus* A- or B′-like.

The distribution of all MLSA loci among BACs containing the 16S rRNA/ITS region and at least 1 protein-encoding locus (i.e., BACs useful for MLSA) is displayed in Tables 4.3 and 4.4. For *Synechococcus* A-like and B′-like populations, of the BACs that contained a 16S rRNA gene and tiled on at least one end to the region being investigated (see Figure 3.7), 159 and 222 *Synechococcus* A- and B′-like BACs exhibited locus combinations expected assuming conservation of gene order relative to the reference genomes (Table 4.3 and 4.4). The number of BACs positive for various loci decreased as the distance from the 16S rRNA locus increased (Figure C4.2). There were 97 *Synechococcus* A-like BACs that contained 7 loci (all included except *apcAB*) in addition

Table 4.3. Locus distribution on *Synechococcus* A-like BACs based on PCR amplification.

| **M60 and M65 *Synechococcus* A-like BACs with Sequence Data for MLSA Loci** | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *apcAB* | *rbsK* | *PK* | *lepB* | *hisF* | *CHP* | *aroA* | 16S rRNA | d*naG* | # of Protein-Encoding Loci | No. M60 BACs | No. M65 BACs | Total |
| + | + | + | + | + | + | + | + | + | 8 | 57 | 7 | 64 |
| | + | + | + | + | + | + | + | + | 7 | 90 | 7 | 97 |
| + | + | + | + | + | + | + | + | | 7 | 83 | 8 | 91 |
| | | + | + | + | + | + | + | + | 6 | 93 | 7 | 100 |
| | + | + | + | + | + | + | + | | 6 | 126 | 10 | 136 |
| | | | + | + | + | + | + | + | 5 | 111 | 9 | 120 |
| | | + | + | + | + | + | + | | 5 | 130 | 10 | 140 |
| | | | | | + | + | + | + | 3 | 127 | 28 | 155 |
| | | | | | + | + | + | | 2 | 181 | 45 | 226 |
| | | | | | | + | + | | 1 | 193 | 57 | 250 |
| | | | | | | | + | + | 1 | 145 | 28 | 173 |
| Number of BACs with Expected Loci | | | | | | | | | | 159 | 17 | 176 |
| ***Synechococcus* A-like BACs with missing MLSA Loci[a]** | | | | | | | | | | | | |
| + | + | + | + | | + | + | + | + | 7 | 6 | 14 | 20 |
| + | + | + | + | + | | + | + | + | 7 | 1 | 0 | 1 |
| + | + | + | + | | + | + | + | | 6 | 4 | 4 | 8 |
| | + | + | + | | + | + | + | + | 6 | 5 | 0 | 5 |
| | + | + | + | + | + | | + | + | 6 | 2 | 0 | 2 |
| + | + | | + | | + | + | + | + | 6 | 1 | 0 | 1 |
| + | + | + | + | + | | + | + | | 6 | 1 | 1 | 2 |
| | | + | + | | | + | + | | 6 | 1 | 0 | 1 |
| + | | + | + | + | + | + | + | | 6 | 1 | 0 | 1 |
| + | | | + | | + | + | + | + | 6 | 0 | 1 | 1 |
| | + | | + | | + | + | + | | 5 | 3 | 0 | 3 |
| + | + | | + | | + | + | + | | 5 | 1 | 0 | 1 |
| + | | + | + | + | | + | + | | 5 | 0 | 1 | 1 |
| + | + | + | + | | | + | + | | 5 | 0 | 2 | 2 |
| + | | + | + | | + | + | + | | 5 | 0 | 3 | 3 |
| | + | + | + | | + | + | + | | 5 | 0 | 2 | 2 |
| | | + | + | | + | + | + | + | 5 | 1 | 0 | 1 |
| | | + | + | | + | + | + | | 4 | 1 | 1 | 2 |
| | | | + | | + | + | + | + | 4 | 6 | 3 | 9 |
| + | + | | + | | | + | + | | 4 | 0 | 1 | 1 |
| | | | + | | + | + | + | | 3 | 0 | 2 | 2 |
| | | | + | + | | | + | + | 3 | 1 | 0 | 1 |
| + | | | + | | | + | + | | 3 | 0 | 5 | 5 |
| | | + | | | | + | + | | 2 | 1 | 0 | 1 |
| | + | | + | | | | + | | 2 | 3 | 0 | 3 |
| | + | | | | + | | + | | 2 | 2 | 0 | 2 |
| | + | | | | | + | + | + | 3 | 1 | 0 | 1 |
| + | + | | | | | | + | | 2 | 1 | 0 | 1 |
| | + | | | | | | + | + | 2 | 7 | 0 | 7 |
| + | | | | | | | + | | 1 | 0 | 1 | 1 |
| Number of BACs with missing 'internal' loci | | | | | | | | | | 50 | 41 | 91 |

[a] BACs that did not yield an amplicon after 3 PCR attempts. These BACs were not used in subsequent analyses of the *Synechococcus* A-like population 7-locus MLSA study but some M65 BACs were included in the 5-locus MLSA study (see below).

Table 4.4. Locus distribution on *Synechococcus* B´-like BACs based on PCR amplification.

| **M60 *Synechococcus* B´-like BACs with Sequence Data for MLSA Loci** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *aroA* | *rbsK* | *pcrA* | 16S rRNA | *ispE* | *accC* | *sufB* | *argD* | *apcAB* | # of Protein-Encoding Loci | No. M60 BACs |
| + | + | + | + | + | + | + | + | | 7 | 2 |
| | + | + | + | + | + | + | + | + | 7 | 1 |
| | + | + | + | + | + | + | + | | 6 | 8 |
| | | + | + | + | + | + | + | + | 6 | 4 |
| + | + | + | + | + | + | | | | 5 | 18 |
| | | | + | + | + | + | + | + | 5 | 35 |
| + | + | + | + | + | | | | | 4 | 28 |
| | + | + | + | + | + | | | | 4 | 47 |
| | | | + | + | + | + | + | | 4 | 48 |
| + | + | + | + | | | | | | 3 | 92 |
| | + | + | + | + | | | | | 3 | 65 |
| | + | + | + | | | | | | 2 | 188 |
| | | | + | + | + | | | | 2 | 83 |
| | | + | + | | | | | | 1 | 197 |
| | | | + | + | | | | | 1 | 102 |
| Number of BACs with Expected Loci | | | | | | | | | | 222 |
| ***Synechococcus* B´-like BACs with missing MLSA Loci[a]** | | | | | | | | | | |
| + | + | | + | + | | + | | | 4 | 2 |
| + | | + | + | + | + | | | | 4 | 1 |
| | + | | + | | | | | | 2 | 8 |
| + | | + | + | | | | | | 2 | 4 |
| Number of BACs with Missing ‘internal’ Loci | | | | | | | | | | 15 |

[a] BACs that were amplified using PCR 3x and still did not yield an amplicon for given gene(s). These BACs were not used in subsequent analyses.

to the 16S rRNA/ITS region and 71 of those BACs contained sequence data clean enough for subsequent analysis. Only two *Synechococcus* B´-like BACs contained 7 MSLA loci. In order to obtain a similar sampling of *Synechococcus* B´-like and A-like BAC clones the number of loci for the *Synechococcus* B´-like population had to be reduced to 3 protein-encoding loci (*rbsK, aroA* and *pcrA*) plus the 16S rRNA-ITS region. This yielded 92 *Synechococcus* B´-like BACs and 71 of those BACs contained sequence data clean enough for further analysis (Tables 4.3 and 4.4).

The average evolutionary divergence (EED) ranged from 0.002 to 22.0 for the loci analyzed in *Synechococcus* A-like BACs and from 0.09 to 0.48 for the loci analyzed in *Synechococcus* B´-like BACs (Table 4.5). These EED values are comparable with the EED values determined from PCR clone library-based phylogenies for each gene (Melendrez et al., 2010a; Chapter 2) and probably provide the best estimate of the potential of the loci sampled for resolving members of each population, since other estimates are either not lineage-specific or are based on metagenomic homologs that are under-sampled in low depth-of-coverage metagenomes (Tables 4.1 and 4.2). A random sampling of BAC and PCR clones amplified for the *rbsK* locus and combined for ES analysis showed that 11 of 13 PEs were sampled by both types of clones. This showed that BACs sampled diversity across ecotypes similarly to PCR clones (Melendrez et al., 2010a; Chapter 2; Figure C4.3).

Ecotype Simulation Analyses and Concatenated PE Clade Structure

Multi-Locus Analysis. Concatenated multi-locus sequence datasets, as well as individual locus sequences sampled by BACs, were analyzed using ES (Figures 4.2 and 4.3 and Tables 4.5 and 4.6). ES predicted 10 A-like and 13 B´-like PEs from concatenated MLSA sequence data and 2 to 6 A-like and 3 to 18 B´-like PEs from individual loci, depending on the locus or loci analyzed. Concatenated PE clade sizes ranged from 1-17 sequences. Three of the 10 PE clades for *Synechococcus* A-like BACs and 4 of the 13 PE clades for *Synechococcus* B´-like BACs contained a DV, which was

Table 4.5. Ecotype Simulation and eBURST output for 7 A-like and 4 B´-like *Synechococcus* BACs.

| ***Synechococcus* A-like** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Locus** | **Estimated Evolutionary Divergence (BACs)** | **ES** | | | | **eBURST[a]** | |
| | | **PEs Demarcated-ES (95% CI)** | **Omega (95% CI)** | **Sigma (95% CI)** | **Sample-Specific PEs** | **No. of Alleles or Sequence Types** | **No. of Clonal Complexes** |
| *rbsK* | 0.03 | 6 (4-30) | 0.03 (0.002-0.14) | 3.5 (0.23->100) | Yes | 24 | [d] |
| *PK* | 3.48 | 2 (2-5) | 0.004 (<2e$^{-7}$-0.01) | 2.26 (0.33-12.2) | No | 14 | [d] |
| *hisF* | 22.0 | 2 (2-3) | 2.e$^{-5}$ (<2e$^{-7}$0.02) | 22.1 (2.2->100) | NA | 4 | [d] |
| *lepB* | 5.11 | 2 (2-71) | 1e$^{-5}$ (<1e$^{-7}$-0.02) | 22.6 (2.3-99.8) | No | 6 | [d] |
| *CHP* | 0.01 | 2 (2-5) | 0.03 (0.001-0.23) | 564 (5.4->100) | Yes | 4 | [d] |
| *aroA* | 0.002 | 2 (2-6) | 0.06 (0.003-0.18) | 15.2 (1.5->100) | No | 7 | [d] |
| *dnaG* | 0.03 | 3 (2-4) | 6.4e$^{-4}$ (<2e$^{-7}$-0.01) | 28.3 (2.7->100) | No | 7 | [d] |
| Concatenation[a] | 0.12 | 10 (4-32) | 0.005 (0.001-0.02) | 0.31 (0.02->100) | [b] | 55 | 4 |
| ***Synechococcus* B´-like** | | | | | | | |
| *aroA* | 0.71 | 9 (5-62) | 0.23 (0.11-0.52) | 15 (0.22->100) | [c] | 37 | [d] |
| *rbsK* | 0.09 | 18 (10-76) | 0.06 (0.04-0.08) | 49 (0.05->100) | [c] | 23 | [d] |
| *pcrA* | 0.48 | 10 (5-77) | 1e$^{-5}$ (1e$^{-7}$-0.01) | 0.75 (0.51-1.7) | [c] | 42 | [d] |
| 16S rRNA /ITS | 0.09 | 3 (3-23) | 0.28 (2e$^{-7}$-0.09) | 1.1 (0.23->100) | [c] | 32 | [d] |
| Concatenation with 16S/ITS[a] | 0.182 | 13 (7-24) | 0.003 (5.3e$^{-4}$-0.01) | 0.05 (0.01-0.16) | [c] | 77 | 2 |
| Concatenation Without 16S/ITS[a] | 0.245 | 19 (5-21) | 0.003 (2e$^{-7}$-0.01) | 0.07 (0.01-0.14) | [c] | 66 | 2 |

[a] Calculated from concatenated sequence datasets only.
[b] Not enough M65 sequences to determine sample specificity of demarcated PEs.
[c] *Synechococcus* B´-like BACs were only retrieved from one temperature site, M60.
[d] Clonal complexes only apply to concatenated sequence datasets.

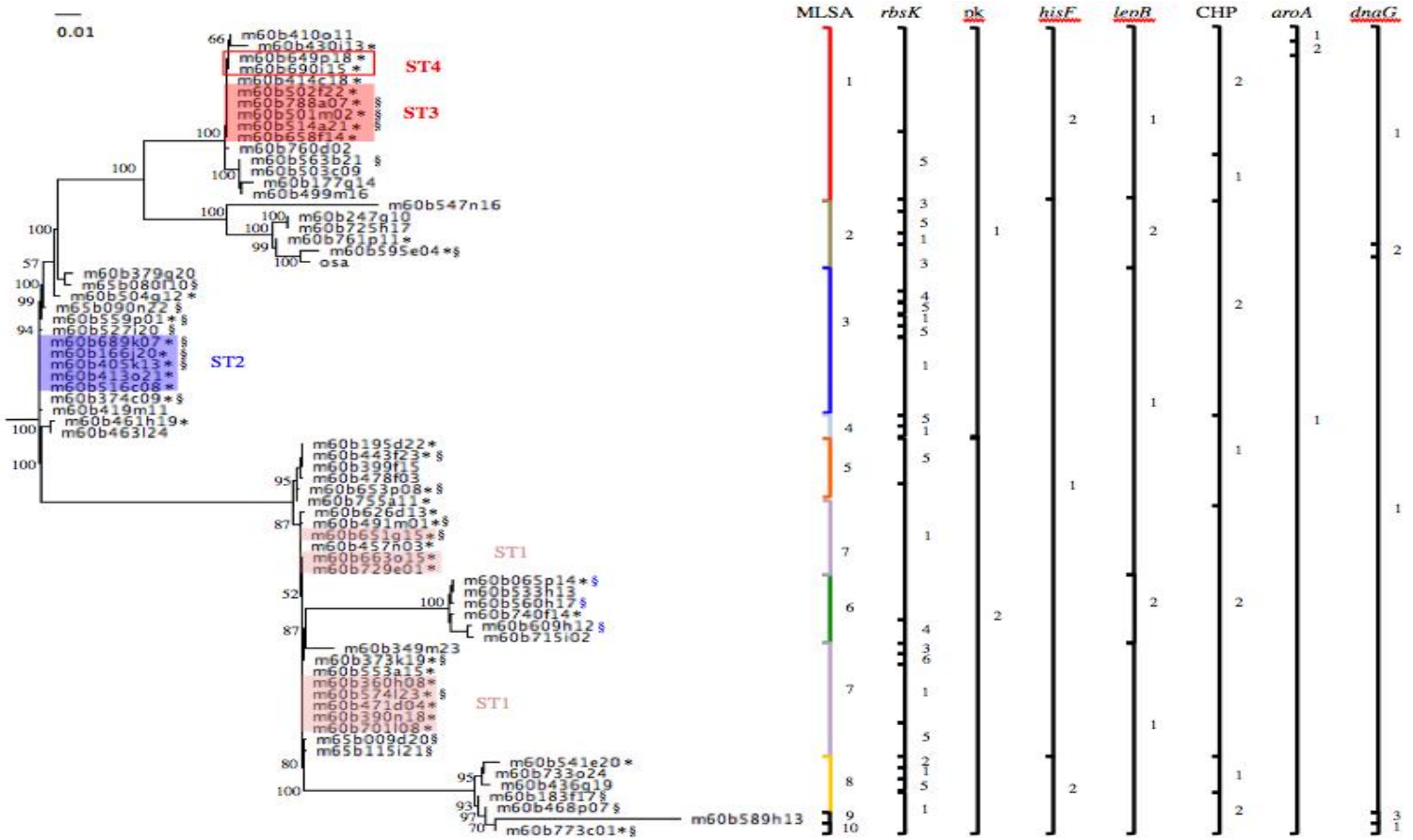



Figure 4.2. Neighbor-joining phylogenetic tree for 7 concatenated loci for *Synechococcus* A-like BACs with PE demarcations. MLSA PEs are color coded to match Figure 4.7. Shaded and boxed variants are DVs and sDVs, respectively representing the adjacent ST. Clones with evidence of recombination as determined by RDP3 analyses in at least one gene, are denoted with asterisks. Clones that also were used in 5-locus analysis are denoted here and in Figure 4.6 with the symbol $ colored according to sample-specific PEs in Figure 4.6 (blue= a part of a low temperature sample-specific PE; black= not a part of a temperature specific PE). Bootstrap values are displayed for major nodes.

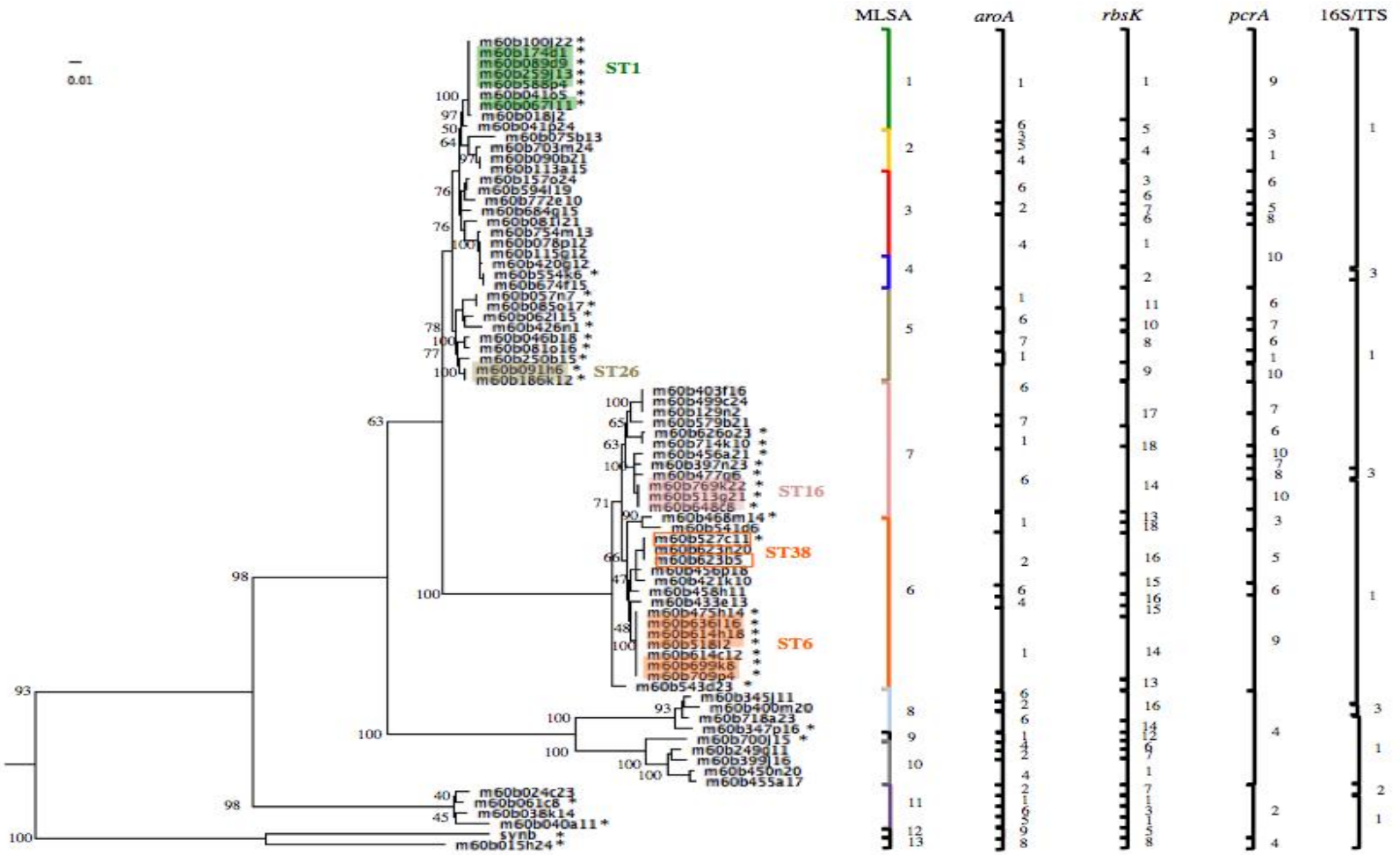



Figure 4.3. Neighbor-joining tree for 4 concatenated loci for *Synechococcus* B´-like BACs with PE demarcations. MLSA PEs are color coded to match Figure 4.8. Shaded and boxed variants are DVs and sDVs, respectively representing the adjacent sequence type. Clones where recombination was detected in at least one gene are denoted with asterisks. Bootstrap values are displayed for major nodes.

Table 4.6. Comparison of singleton STs surrounding DVs of PEs and consensus sequences of clonal complexes observed in 7-locus and 4-locus MLSA of A-like and B´-like *Synechococcus* BACs. STs common to a PE and a corresponding clonal complex are highlighted in grey[a].

| ***Synechococcus* A-like** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **PE** | **DV-ST** | **No. of sequences with DV/sDV-ST** | **Singleton STs within PE** | **Clonal Complex** | **Consensus ST** | **No. of sequences with consensus ST** | **STs (SLVs) within a Clonal Complex** |
| 1 | 3 | 5 | 4, 7, 23, 27, 28, 29, 43, 49, 52 | A-I | 3 | 5 | 2, 4, 7, 14, 43 |
| 1 | 4[b] | 2 | 3, 7, 23, 27, 28, 29, 43, 49, 52 | A-1 | 3 | 5 | 2, 4, 7, 14, 43 |
| 2 | [c] | [c] | | [d] | [d] | [d] | [d] |
| 3 | 2 | 5 | 5, 22, 33, 35, 36, 42, 53, 54 | A-II | 2 | 5 | 1, 3, 5, 6, 36 |
| 4 | [c] | [c] | | [d] | [d] | [d] | [d] |
| 5 | [c] | [c] | 11, 19, 21, 24, 30 | A-IV | 20[e] | 1 | 11, 17, 19, 21, 24 |
| 6 | [c] | [c] | | [d] | [d] | [d] | [d] |
| 7 | 1 | 8 | 9, 15, 26, 31, 38, 40, 41, 48 | A-III | 1 | 8 | 2, 8, 9, 10, 11,15, 38, 41,48 |
| 8 | [c] | [c] | 25, 32, 39, 44, 46 | [d] | [d] | [d] | [d] |
| 9 | [c] | [c] | 12 | [d] | [d] | [d] | [d] |
| 10 | [c] | [c] | 14 | [d] | [d] | [d] | [d] |
| ***Synechococcus* B´-like** | | | | | | | |
| 1 | 1 | 5 | 2, 3, 4, 25 | B-I | 1 | 5 | 2, 3, 5, 6, 14 |
| 2 | [c] | [c] | 50, 51, 56, 65 | [d] | [d] | [d] | [d] |
| 3 | [c] | [c] | 19, 20, 27, 34, 35, 36, 44, 48 | [d] | [d] | [d] | [d] |
| 4 | [c] | [c] | 28, 32, 33 | [d] | [d] | [d] | [d] |
| 5 | 26 | 2 | 9, 10, 46, 47, 55, 57, 58 | [d] | [d] | [d] | [d] |
| 6 | 6 | 5 | 7, 8, 12, 13, 14, 23, 31, 37, 38, 43, 61 | B-II | 6 | 5 | 1, 7, 8 |
| 6 | 38[b] | 2 | 6, 7, 8, 12, 13, 14, 23, 31, 37, 43, 61 | [d] | [d] | [d] | [d] |
| 7 | 16 | 3 | 11, 18, 21, 45, 60, 62, 64 | [d] | [d] | [d] | [d] |
| 8 | [c] | [c] | 17, 22, 24, 39 | [d] | [d] | [d] | [d] |
| 9 | [c] | [c] | 15 | [d] | [d] | [d] | [d] |
| 10 | [c] | [c] | 29, 30, 42, 49 | [d] | [d] | [d] | [d] |
| 11 | [c] | [c] | 5, 41, 53, 54 | [d] | [d] | [d] | [d] |
| 12 | [c] | [c] | 66 | [d] | [d] | [d] | [d] |
| 13 | [c] | [c] | 52 | [d] | [d] | [d] | [d] |

[a] STs designated in the *Synechococcus* A-like population 7 locus MLSA do not correspond to the STs in the 5 locus MLSA.
[b] sDV in PE.
[c] PE did not contain a DV.
[d] PE ST variants do not appear within a clonal complex.
[e] ST20 is represented by only 1 sequence-clone M60B339F15

identical at all loci (color highlighting in Figures 4.2 and 4.3) surrounded by singleton PEVs. There were two cases of subdominant PEV variants (sDVs), surrounded by singletons (colored boxes; Figures 4.2 and 4.3). Table 4.6 shows DVs, sDVs and how STs were distributed among PEs.

Comparison of Multi-locus and Single-locus Analyses. Comparison of ES PE predictions from SLA and concatenated MLSA (Figures 4.2 and 4.3) shows that MLSA sometimes split clades demarcated as single PEs in SLA into many PEs. For instance, the two A-like PEs predicted from ES analysis of *PK* sequences were split into 10 MLSA PEs (Figure 4.2). This presumably occurred because the addition of loci increased resolution. With the exception of B´-like *rbsK*, the number of PEs predicted in MLSA was always greater than that predicted by SLA.

There was a high degree of phylogenetic incongruency between single-locus and concatenated MLSA phylogenies (Figures 4.2 and 4.3), suggesting that recombination may have played a significant role in the evolution *Synechococcus* A- and B´-like populations. This is evidenced by the dispersion of variants within concatenated PEs into many single-locus PEs and the combination of variants in many concatenated PEs into individual single-locus PEs (Figures 4.2 and 4.3). Incongruency is more easily observed by comparing concatenated phylogenies and single-locus phylogenies for the *rbsK* gene, which exhibited the highest degree of phylogenetic incongruency, as shown in Figures 4.4 and 4.5. Despite evidence of recombination, smaller sets of sequences within PE clades (2-9 sequences) remained congruous in both SLA and MLSA. For instance, A-like MLSA PE1 is comprised of 15 sequences, 8 of which (representing 4 STs) remain in the

same PE clade in all single gene phylogenies (ST3-m60b501m02, m60b502f22, m60b514a21, m60b788a07; ST4- m60b690i15, m60b649p18; ST23- m60b410o11 and ST43- m60b414c18 with the exception of clone m60b658f14-ST3) (Figure 4.2).

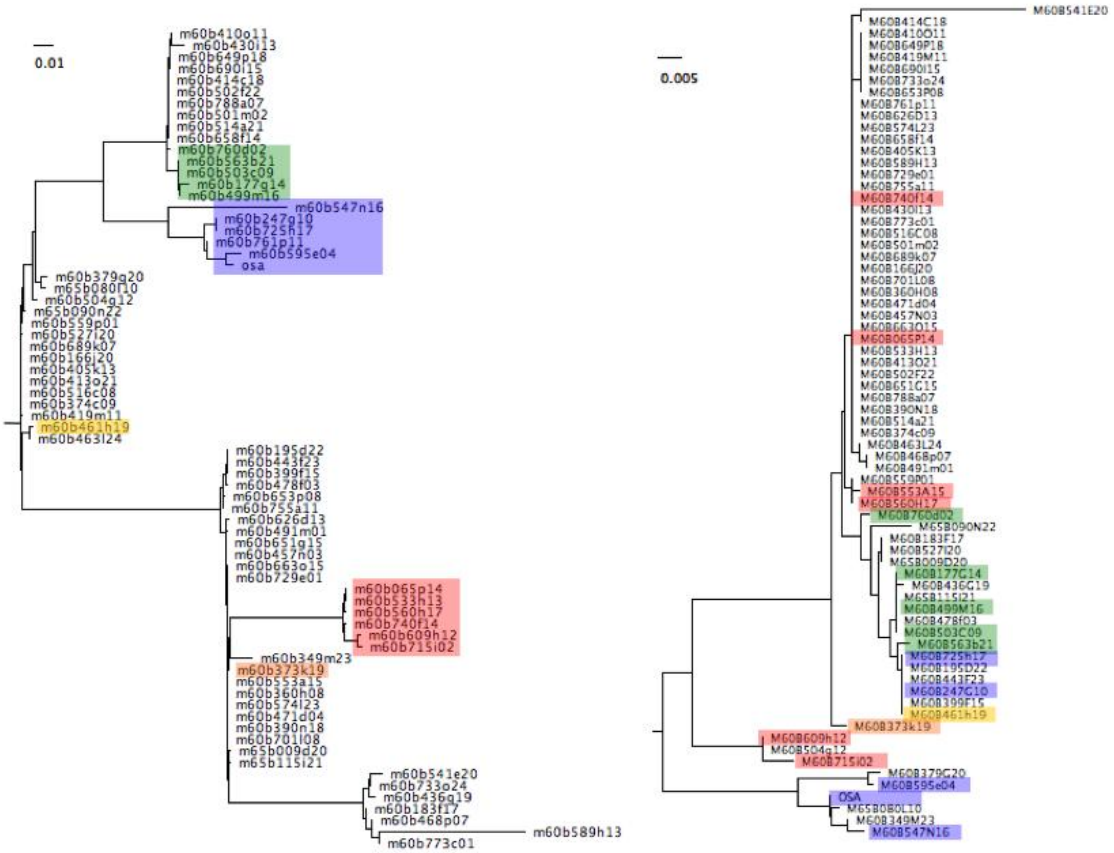



Figure 4.4. *Synechococcus* A-like BAC neighbor-joining phylogenetic trees for concatenated loci (left) and the *rbsK* locus (right) showing phylogenetic incongruence. Colored boxes highlight single and multiple sequences that formed clades (or appeared in a specific location in the tree in the analysis of concatenated sequences) and were separated (or relocated) in the *rbsK* phylogeny.

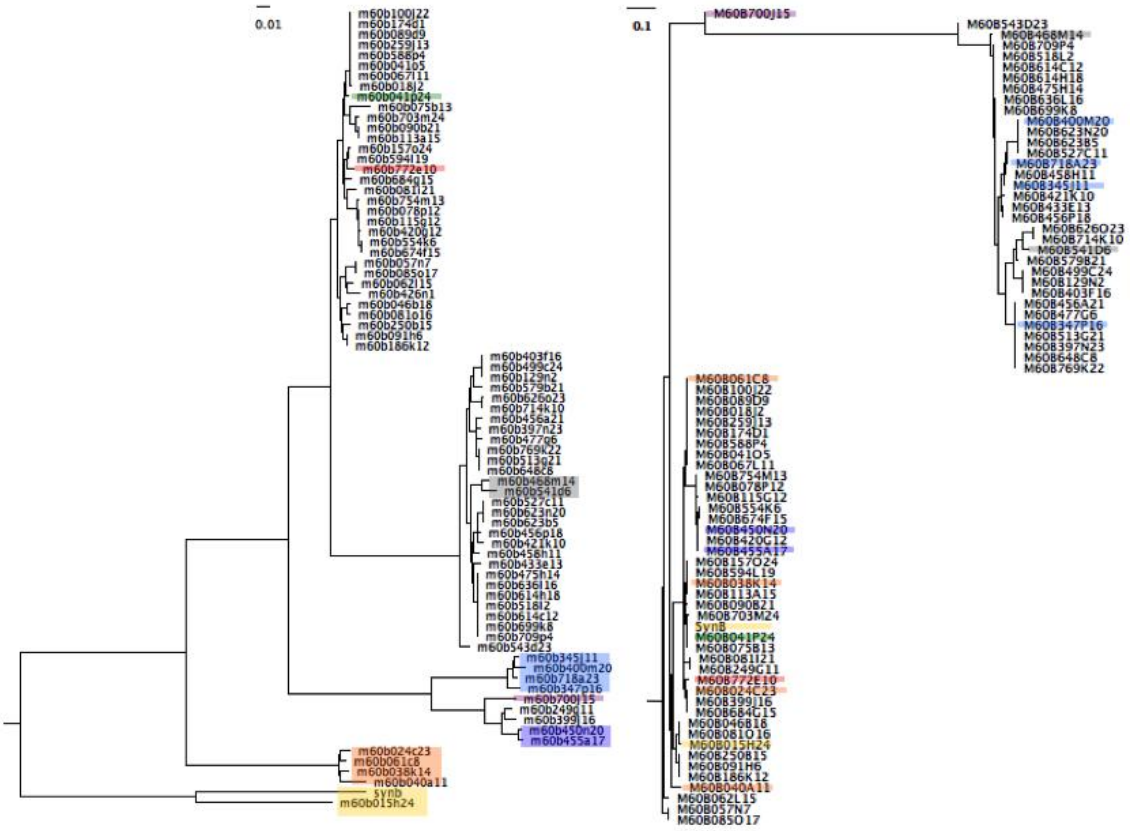



Figure 4.5. *Synechococcus* B´-like BAC neighbor-joining phylogenetic trees for concatenated loci (left) and the *rbsK* locus (right) showing phylogenetic incongruence. Colored boxes highlight single and multiple sequences that formed clades (or appeared in a specific location in the analysis of concatenated sequences) and were separated (or relocated) in the *rbsK* phylogeny.

MLSA PEs A4-7 and B´-like MLSA PEs B´1, 3-7 and 10 also had groups of sequences that remained within the same PE clade in SLA and MLSA (Figures 4.2 and 4.3).

Sample Specificity of PE Clades. In the case of A-like sequences, which were found in both samples, it was possible to evaluate sample-specificity of predicted ecotype clades. Previous research using single-protein encoding genes suggested *Synechococcus* A-like population PE clades that were indeed sample-specific (Melendrez et al., 2010a; Chapter 2). In the 7-locus analysis, only 4 sequences from the M65 BAC library could be included, making it impossible to observe sample specificity. In order to obtain a sufficient and equal number of BACs from the two samples, we reduced the number of loci used for concatenated analysis to 5 (*rbsK, lepB, aroA, CHP* and *PK*), permitting 24 M65 BAC clones to be included in the analysis. A comparable number of A-like BAC clones from M60 containing the same 5 loci were randomly sampled. It is important to note that this dataset is different from the 7-locus dataset described above. [Although allele types correspond, ***the STs do not*** and direct comparison of the 5-locus analysis and the 7-locus analysis described above cannot be made.] ES analysis of a concatenation of these 5 genes in these 48 BACs predicted 8 PEs (Figure 4.6 and Table 4.7), half of which (PEs 2, 6, 7, and 8 in Figure 4.6) contained clones that were recovered from either 60$^{o}$C or 65°C samples. A contingency test using Fisher's exact test suggested significant evidence of heterogeneity across ecotypes in habitat associations in the concatenated analysis (p-value <0.001). The *rbsK* locus showed PEs that were sample-specific, however all of the PEs contained only one sequence. The *CHP* locus also showed sample-specific PEs, including 2 PEs that contained only one sequence and 1 PE that was

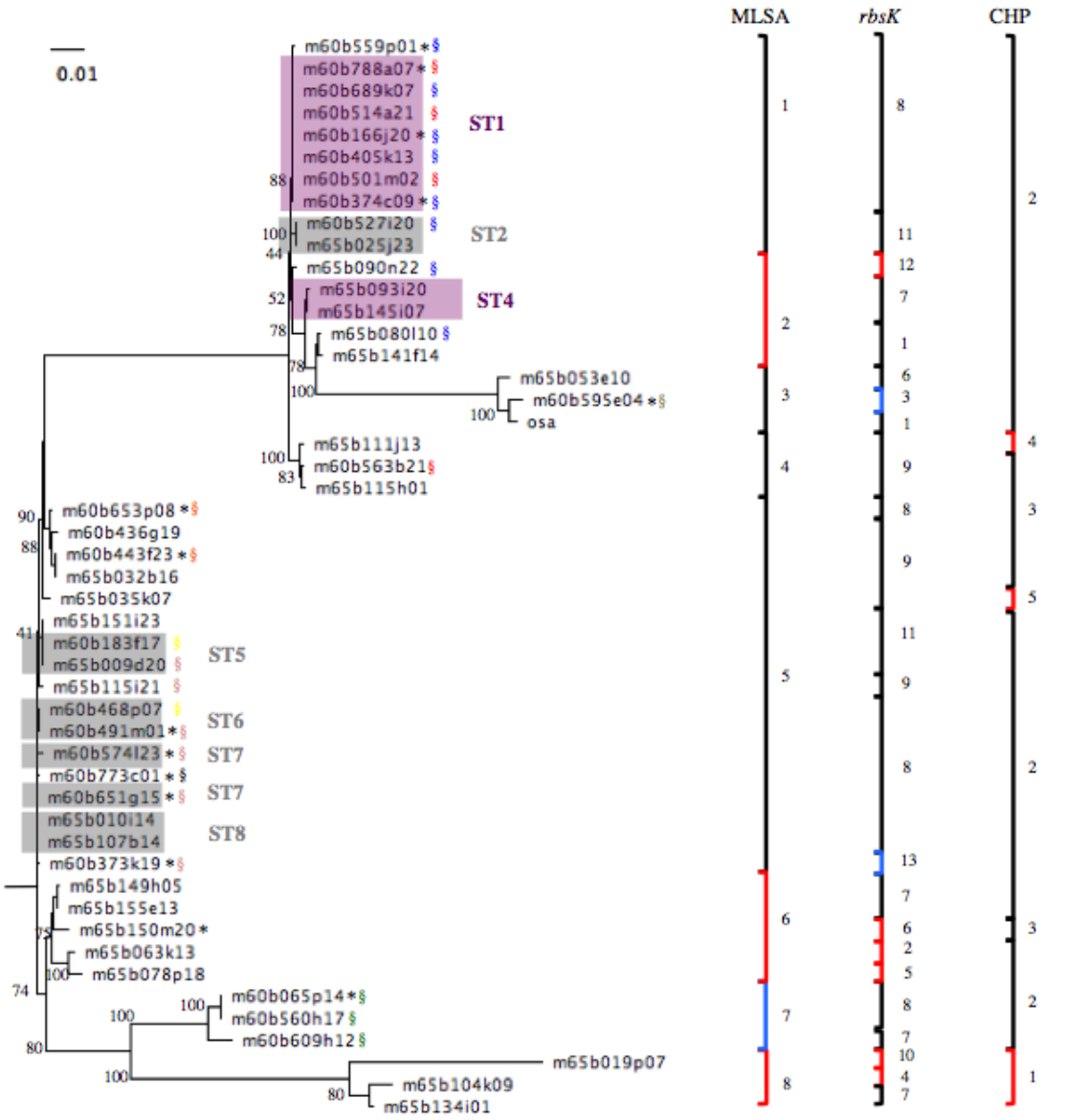



Figure 4.6. Neighbor-joining phylogenetic tree for 5 concatenated loci (*rbsK, PK*, *lepB, CHP* and *aroA*) for *Synechococcus* A-like BACs from 60°C (blue), 65°C (red) or both (black) Mushroom Spring mat DNA samples with PE demarcations. Shaded variants are DVs (purple) and sDVs (grey) representing the adjacent ST. Clones that exhibit evidence of recombination as determined by RDP3 analyses in at least one gene, are denoted with asterisks. Clones that also were used in 7-locus analysis are denoted here and in Figure 4.2 with the symbol § colored according to PEs in Figure 4.2 (blue= PEA3, red=PEA1, brown=PEA2, orange=PEA5, yellow=PEA8, pink=PEA7, green=PEA6 and black=PEA10. Bootstrap values are displayed for major nodes.

Table 4.7. Ecotype Simulation and eBURST output for the 5-locus analysis of *Synechococcus* A-like BACs.

| Locus | Estimated Evolutionary Divergence (BACs) | MLST Fragment Size (bp) | ES | | | | eBURST[a] | |
|---|---|---|---|---|---|---|---|---|
| | | | PEs Demarcated -ES (95% CI) | Omega (95% CI) | Sigma (95% CI) | Sample-Specific PEs | No. of Alleles or Sequence Types | No. of Clonal Complexes |
| *rbsK* | 0.04 | 587 | 13 (5-28) | 0.1 (0.02-0.2) | 8.03 (0.01->100) | Yes | 33 | [b] |
| *PK* | 2.3 | 629 | 2 (2-48) | 0.005 ($2e^{-7}$-44.8) | 3.1 (0.3-31.3) | No | 20 | [b] |
| *lepB* | 3.5 | 439 | 2 (2-3) | $3e^{-5}$ ($2e^{-7}$-0.01) | 56.6 (2.6 ->100) | No | 9 | [b] |
| *CHP* | 0.203 | 653 | 5 (4-5) | 0.003 ($3.2^{e-4}$-0.02) | 13.3 (0.4->100) | Yes | 11 | [b] |
| *aroA* | 0.001 | 609 | 1 (1-49) | 9.2 ($2e^{-7}$-100) | 8.7 (0.1->100) | No | 9 | [b] |
| concatenation[a] | 0.079 | 2917 | 8 | 0.004 ($8.3e^{-4}$-0.01) | 0.32 (0.02->100) | Yes | 37 | 2 |

[a] Calculated from concatenated sequence datasets only.
[b] Clonal complexes only apply to concatenated sequence datasets.

a clade containing three sequences. The *lepB, aroA* and *PK* genes did not show sample-specific PEs; instead they lumped PEs demarcated by higher resolving genes into fewer PEs (*lepB*: 2 PEs, *aroA*: 1 PE and *PK*: 2 PEs; data not shown)(Table 4.7). The lower number of PEs compared to 7-locus analysis may have resulted from limited resolution or also from the use of fewer sequences in the 5-locus analysis. Other demarcated PEs contained sequences retrieved from both samples, suggesting that these populations, if real, have a broader distribution along the effluent channel or represent a population comprised of >1 closely related, younger ecotypes. As with the 7-locus MLSA of the *Synechococcus* A-like population, there was evidence of DVs and sDVs surrounded by singletons (Table 4.8).

EBurst Analyses

Clonal Complexes. The eBURST analysis predicted 4 A-like and 2 B´-like clonal complexes (Table 4.9 and Figures 4.7, 4.8 and C4.1). The lower number of B´-like clonal complexes may be due to the B´-like population being more divergent than the A-like population and to recombination making consensus alleles more rare. Clonal complex A-III contained a consensus group of 8 sequences (ST1) and 9 SLVs, all of which were singleton STs except for ST2, for which there were 5 replicates. ST2 was the consensus group for clonal complex A-II, which included 5 SLVs. The A-II complex had two SLVs that contained more than one sequence, ST1 (9 sequences) and ST 3 (5 sequences). ST3 was the consensus group for clonal complex A-III, which also contained 4 SLVs, all singletons except ST4 (2 sequences). Clonal complex A-IV contained an unreplicated

Table 4.8. Comparison of variant [singleton] STs surrounding dominant sequence types in PEs and clonal complexes observed in 5-locus MLSA analyses of A-like *Synechococcus* populations. Variant STs shared between a PE and clonal complex are highlighted in grey.

| **A-like *Synechococcus* 5 locus MLSA[h]** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **PE** | **DV-ST** | **No. of sequences of DV/sDV-ST** | **Variant STs within PE** | **Clonal Complex** | **Consensus ST** | **No. of sequences with consensus ST** | **STs (SLVs) within a Clonal Complex** |
| 1 | 1 | 7 | 2, 3 | A5-I | 7[i] | 2 | 6, 7[i], 8, 9, 10, 32 |
| 1 | 2[b,d] | 2 | 1, 3 | [g] | [g] | [g] | [g] |
| 2[a] | 4 | 2 | 14, 15 | [g] | [g] | [g] | [g] |
| 3 | [e] | [e] | 16, 17, 18 | [g] | [g] | [g] | [g] |
| 4 | [e] | [e] | 19, 20, 21 | [g] | [g] | [g] | [g] |
| 5[c] | 5 | 3 | 6, 7, 8, 9, 10, 22, 23, 24, 25, 26, 28 | [g] | [g] | [g] | [g] |
| 5[c] | 6 | 2 | 5, 7, 8, 9, 10, 22, 23, 24, 25, 26, 28 | A5-I | 7[i] | 2 | 6, 7[i], 8, 9, 10, 32 |
| 5[c] | 7 | 2 | 5, 6, 8, 9, 10, 22, 23, 24, 25, 26, 28 | A5-I and A5-II | 7 and 8 | 2 and 2 | 1, 6, 8, 9, 10, 32 and 3, 7, 33 |
| 5[c] | 8[f] | 2 | 5, 6, 7, 9, 10, 22, 23, 24, 25, 26, 28 | A5-II | 8 | 2 | 3, 7, 33 |
| 6[a] | [e] | [e] | 11, 12, 29, 30, 31 | [g] | [g] | [g] | [g] |
| 7[b] | [e] | [e] | 32, 33, 34 | [g] | [g] | [g] | [g] |
| 8[a] | [e] | [e] | 35, 36, 37 | [g] | [g] | [g] | [g] |

[a] M65-specific PE.
[b] M60-specific PE.
[c] PE5 contained 1 DV representing 3 sequences and 3 sDVs represented by 2 sequences each.
[d] sDV of PE 1.
[e] PE did not contain a DV.
[f] DV called ST8 (2 sequences) because it enucleates clonal complex A5-II, however the DV is ST5 with 3 sequences; see footnote [c].
[g] PE ST variants do not appear within a clonal complex.
[h] STs from the A-like *Synechococcus* 5-locus analysis ***cannot be equated*** with STs from the 7-locus analysis.
[i] ST7, while the consensus ST of clonal complex A5-I was not the DV. The DV was ST1 with 7 sequences.

Table 4.9. eBURST analysis of clonal complexes for the 7-locus and 4-locus MLSA of *Synechococcus* A-like and B´-like BACs, respectively. The number of BACs within an ST is in parentheses when greater than 1. Superscripts next to the allele number denote number of nucleotide differences compared to the consensus sequence.

| **Clonal Complex A-III** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Locus | Consensus-ST1 (8) | ST2 (5) | ST8 | ST9 | ST10 | ST11 | ST15 | ST38 | ST41 | ST48 |
| *rbsK* | 1 | 1 | 1 | 1 | 1 | 1 | 1 | $8^{2}$ | $10^{5}$ | $17^{5}$ |
| *PK* | 2 | $1^{388}$ | 2 | 2 | 2 | 2 | $14^{11}$ | 2 | 2 | 2 |
| *hisF* | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| *lepB* | 1 | 1 | $2^{233}$ | 1 | $3^{234}$ | 1 | 1 | 1 | 1 | 1 |
| *CHP* | 1 | 1 | 1 | 1 | 1 | $2^{20}$ | 1 | 1 | 1 | 1 |
| *aroA* | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| *dnaG* | 1 | 1 | 1 | $2^{1}$ | 1 | 1 | 1 | 1 | 1 | 1 |

| **Clonal Complex A-II** | | | | | | |
|---|---|---|---|---|---|---|
| Locus | Consensus-ST2 (5) | ST1 (8) | ST3 (5) | ST5 | ST6 | ST36 |
| *rbsK* | 1 | 1 | 1 | 1 | 1 | $7^{2}$ |
| *PK* | 1 | $2^{388}$ | 1 | 1 | 1 | 1 |
| *hisF* | 1 | 1 | $2^{261}$ | 1 | 1 | 1 |
| *lepB* | 1 | 1 | 1 | 1 | $3^{234}$ | 1 |
| *CHP* | 1 | 1 | 1 | 1 | 1 | 1 |
| *aroA* | 1 | 1 | 1 | 1 | 1 | 1 |
| *dnaG* | 1 | 1 | 1 | $3^{2}$ | 1 | 1 |

| **Clonal Complex A-I** | | | | | | |
|---|---|---|---|---|---|---|
| Locus | Consensus-ST3 (5) | ST2 (5) | ST4 (2) | ST7 | ST14 | ST43 |
| *rbsK* | 1 | 1 | $3^{1}$ | 1 | 1 | $12^{1}$ |
| *PK* | 1 | 1 | 1 | 1 | $4^{388}$ | 1 |
| *hisF* | 2 | $1^{261}$ | 2 | 2 | 2 | 2 |
| *lepB* | 1 | 1 | 1 | 1 | 1 | 1 |
| *CHP* | 1 | 1 | 1 | 1 | 1 | 1 |
| *aroA* | 1 | 1 | 1 | $7^{26}$ | 1 | 1 |
| *dnaG* | 1 | 1 | 1 | 1 | 1 | 1 |

| **Clonal Complex A-IV** | | | | | | |
|---|---|---|---|---|---|---|
| Locus | Consensus-ST20 | ST11 | ST17 | ST19 | ST21 | ST24 |
| *rbsK* | 2 | $1^{9}$ | 2 | 2 | 2 | $3^{10}$ |
| *PK* | 2 | 2 | $1^{388}$ | 2 | $4^{1}$ | 2 |
| *hisF* | 1 | 1 | 1 | 1 | 1 | 1 |
| *lepB* | 1 | 1 | 1 | 1 | 1 | 1 |
| *CHP* | 2 | 2 | 2 | $3^{2}$ | 2 | 2 |
| *aroA* | 1 | 1 | 1 | 1 | 1 | 1 |
| *dnaG* | 1 | 1 | 1 | 1 | 1 | 1 |

| | **Clonal Complex B-I** | | | | | | **Clonal Complex B-II** | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Locus | Consensus - ST1 (5) | ST2 | ST3 | ST5 | ST6 (5) | ST14 | Consensus-ST6 (5) | ST1 (5) | ST7 | ST8 |
| aroA | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| rbsK | 1 | 1 | 1 | 1 | $3^{329}$ | $30^{311}$ | 3 | $1^{329}$ | 3 | 3 |
| pcrA | 1 | 1 | 1 | $14^{326}$ | 1 | 1 | 1 | 1 | 1 | 1 |
| 16S rRNA/ITS | 1 | $26^{5}$ | $20^{11}$ | 1 | 1 | 1 | 1 | 1 | $3^{1}$ | $13^{5}$ |

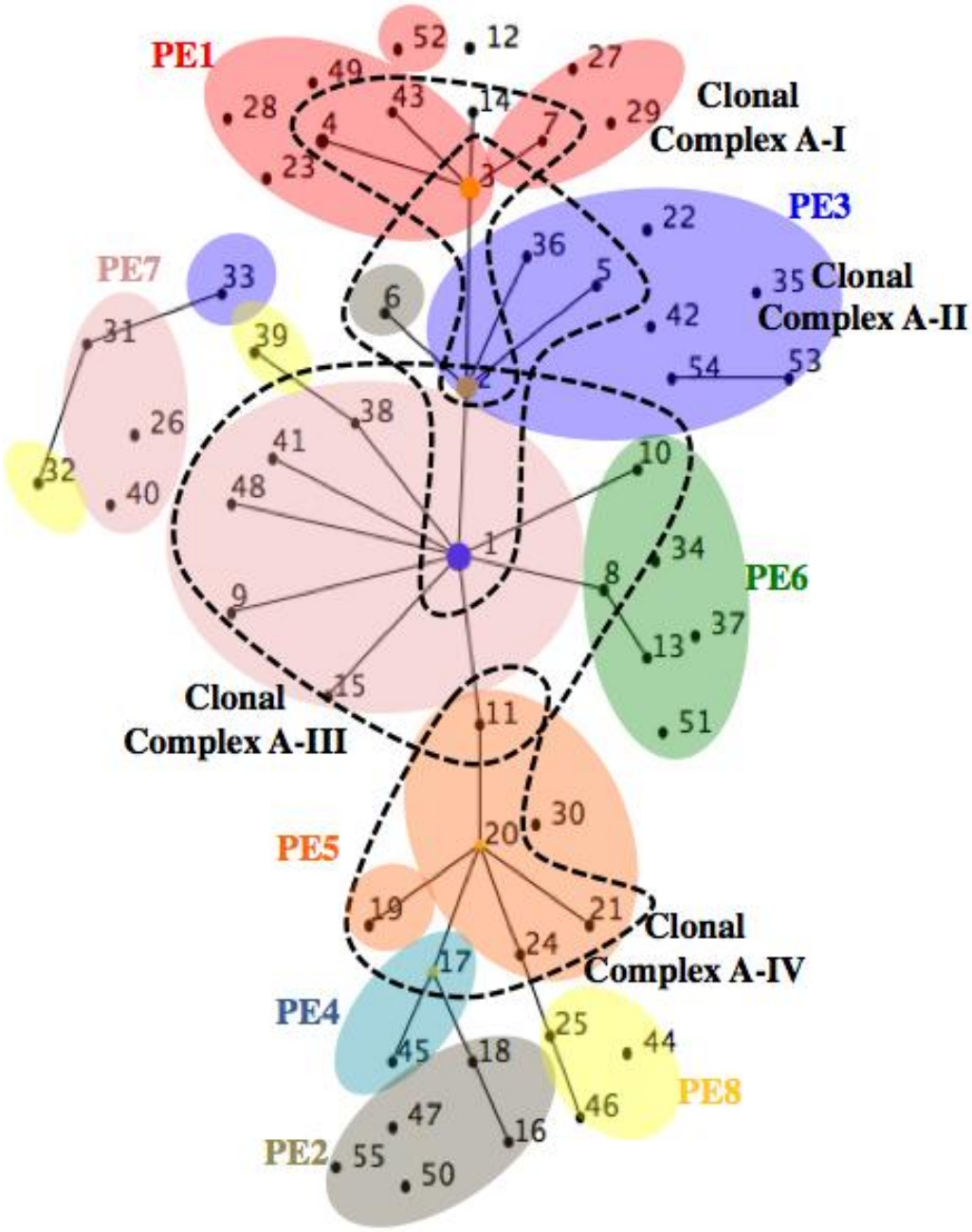



Figure 4.7. eBURST population snapshot of *Synechococcus* A-like BACs showing clonal complexes (enclosed by dashed lines) with PE demarcation from ES analysis overlaid. Different colors represent distinct PEs (corresponding to Figure 4.2). STs are represented by numbers and those not bounded by a colored area belong to PEs demarcated from a single sequence with a unique sequence type. For raw population snapshot from eBURST output refer to Figure C4.1.

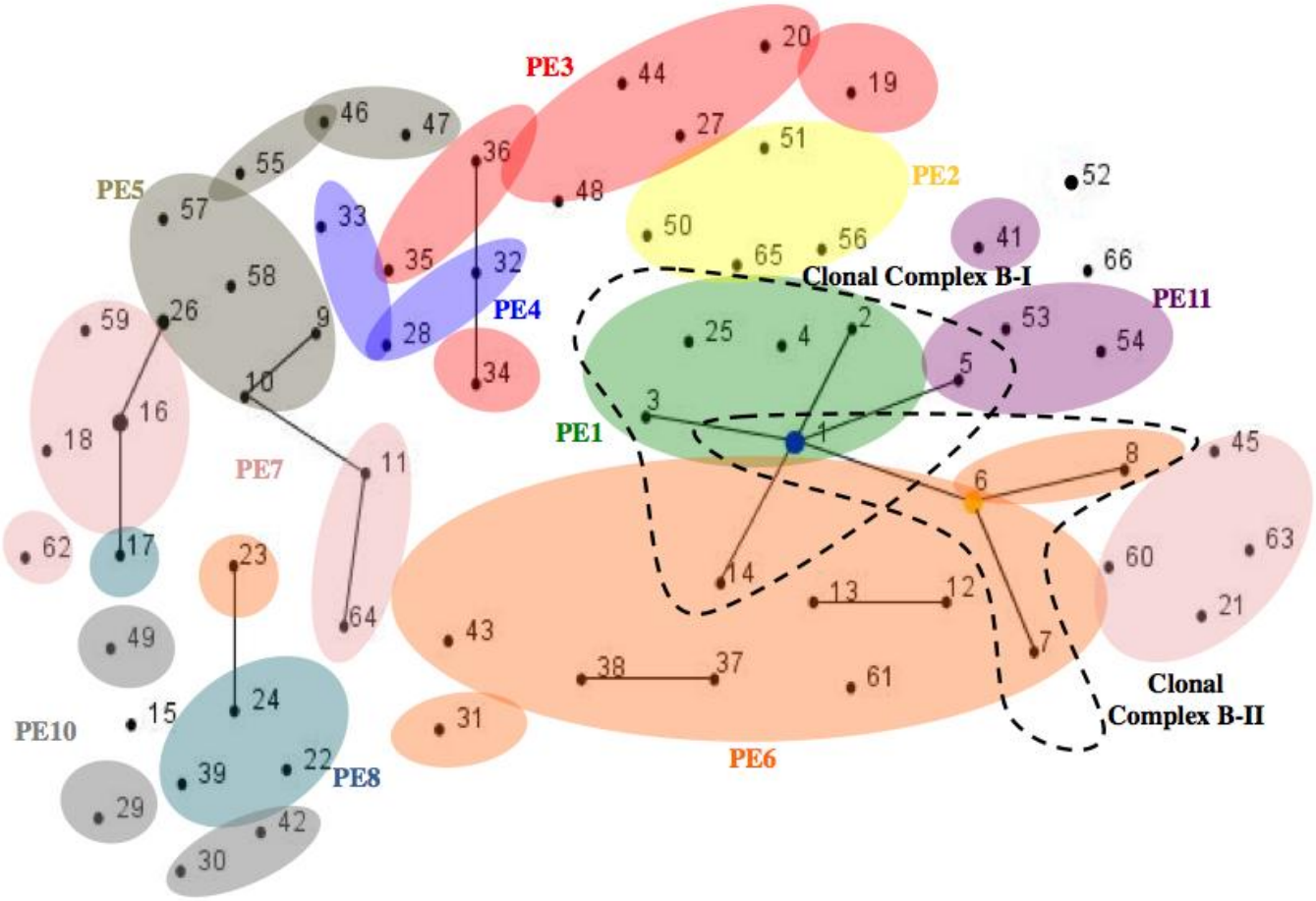



Figure 4.8. eBURST population snapshot of *Synechococcus* B´-like BACs showing clonal complexes (enclosed by dashed lines) with PE demarcation from ES analysis overlaid. Different colors represent distinct PEs (corresponding to Figure 4.3). Sequence types (represented by numbers) not bounded by a colored area belong to PEs demarcated from a single sequence with a unique sequence type. For raw population snapshot from eBURST output refer to Figure C4.1.

sequence (ST20) as the consensus type and 5 singleton SLVs. Clonal complex B′-I contained a consensus group of 5 sequences (ST1) and 5 SLVs. ST6, a replicated SLV of clonal complex B′-I (5 sequences) was the consensus group for clonal complex B′-II, which also contained 3 SLVs. ST1 was a replicated SLV of clonal complex B′-II and all other clonal complex B′-II SLVs were singleton STs.

If the criteria for demarcating clonal complexes were relaxed so that DLVs or fewer SLVs were allowed per clonal complex, more clonal complexes were observed (Figures C4.4 and C4.5). For *Synechococcus* A-like population STs clonal complex composition became more complex with many potential complexes depending on which consensus variant ST was selected as the 'founder'; four potential clonal complexes are highlighted in Figure C4.5. One additional clonal complex (clonal complex Figure C4.5F) could be defined under relaxed conditions. For *Synechococcus* B′-like population STs, 3 additional clonal complexes could be defined (Figure C4.5G, H and J) and clonal complex B′-I and B′-II (Figure 4.8) were combined into one large clonal complex (Figure C4.5I). However, these additional clonal complexes did not improve the correspondence between PEs and clonal complexes.

Sample-Specificity of Clonal Complexes. When the same 48 A-like 5-locus dataset used to test the sample specificity of PEs was analyzed by eBURST two clonal complexes were observed (Tables 4.10 and Figure 4.9). ST8 is the consensus group of clonal complex A5-II and is also an SLV of clonal complex A5-I, whose consensus group is ST7. [Again, STs and clonal complexes here are not the same as those in the 7-locus A-like and 4-locus B′-like analysis, as these experiments were done on different

Table 4.10. eBURST analysis of clonal complexes for the 5-locus MLSA of *Synechococcus* A-like BACs. The number of BACs within STs is in parentheses when greater than 1. Superscripts next to the allele number denote number of nucleotide differences compared to the consensus sequence.

| **Clonal Complex A5-I** | | | | | | | |
|---|---|---|---|---|---|---|---|
| Locus | Consensus-ST7 (2) | ST1 (7) | ST6 (2) | ST8 (2) | ST9 | ST10 | ST32 |
| *rbsK* | 1 | 1 | $8^{8}$ | $7^{7}$ | 1 | $10^{11}$ | 1 |
| *PK* | 2 | $1^{307}$ | 2 | 2 | $4^{1}$ | 2 | $2^{235}$ |
| *lepB* | 1 | 1 | 1 | 1 | 1 | 1 | 2 |
| *CHP* | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| *aroA* | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

| **Clonal Complex A5-II** | | | | |
|---|---|---|---|---|
| Locus | Consensus- ST8 (2) | ST3 | ST7 (2) | ST33 |
| *rbsK* | 7 | 7 | $1^{7}$ | 7 |
| *PK* | 2 | $1^{307}$ | 2 | 2 |
| *lepB* | 1 | 1 | 1 | $2^{235}$ |
| *CHP* | 1 | 1 | 1 | 1 |
| *aroA* | 1 | 1 | 1 | 1 |

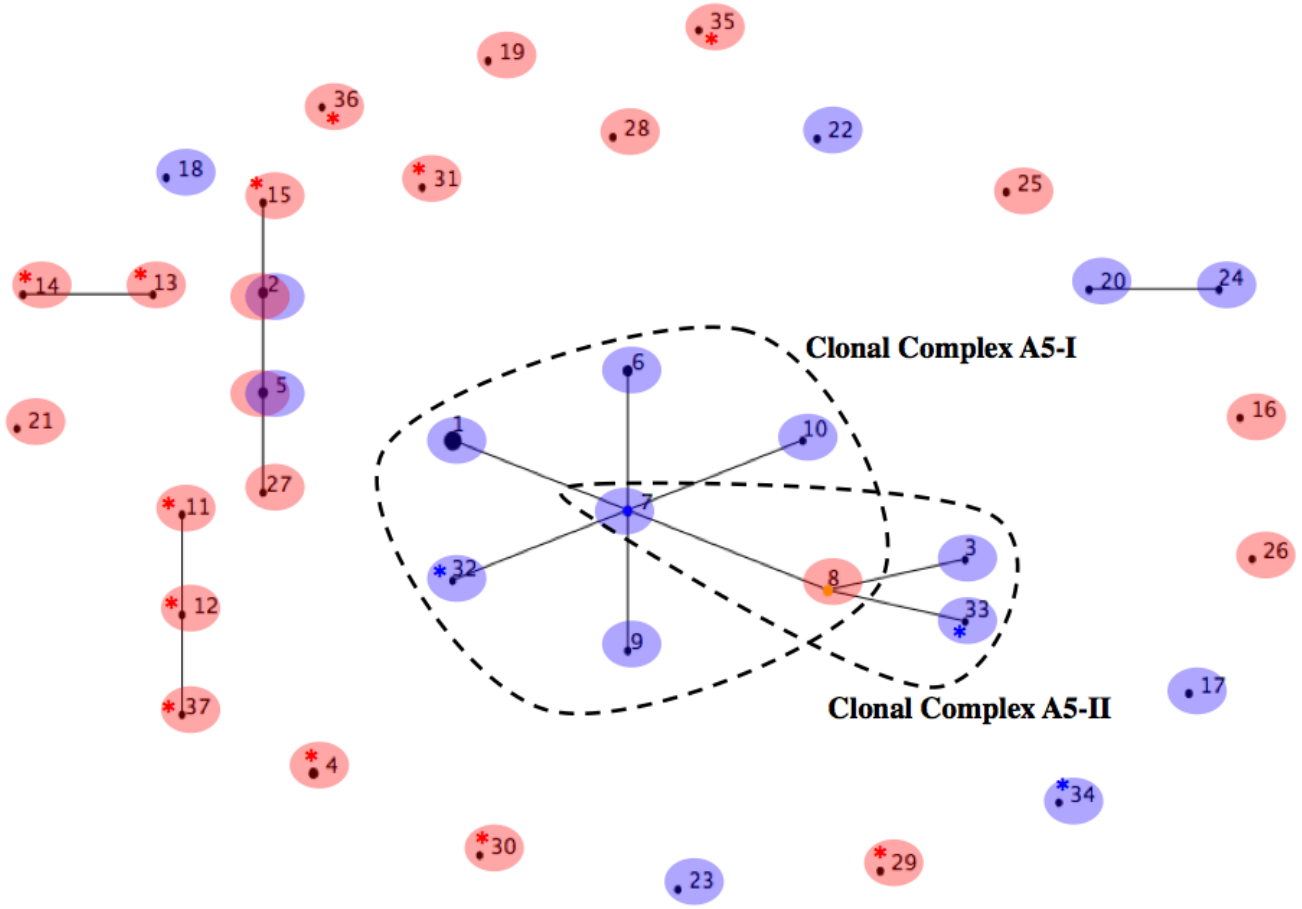



Figure 4.9. eBURST population snapshot of *Synechococcus* A-like BAC clonal complexes with STs (numbers) from M65 (red) and M60 (blue) highlighted  Asterisks indicate sequence types belonging to a sample-specific PE in ES analysis (see Figure 4.6). Dashed lines enclose clonal complexes.

samplings of the BAC MLSA database.] All STs in these two clonal complexes are from the M60 sample, with the exception that ST8 comes from the M65 sample. If the definition of a clonal complex is relaxed to include 2 SLVs rather than 3 SLVs, two more clonal complexes are observed (Figure C4.4A and C). One, which includes ST12 as the consensus group and STs 11 and 37 as SLVs is specific to the M65 sample (Figure C4.4C). The other, which includes STs 2, 5, 15 and 27 contains variants from both M60 and M65 samples. None of the clonal complexes in the more stringent "unrelaxed" analysis corresponded to sample-specific PEs (compare blue and red shading with blue and red asterisks in Figure 4.9) and a Fisher's exact test suggested that the clonal complexes, themselves, did not associate with a particular sample with a p-value =1.0 (two-tailed). This means that there was no evidence of habitat association among the sequence types within a particular clonal complex.

Comparison of Ecotype Simulation and eBURST

When ecotype demarcation by ES from the 7-locus and 4-locus analyses was overlaid on the eBURST population snapshots (Figures 4.7 and 4.8) it was apparent that the clonal complexes tended to group individuals from >1 PE together. For instance, clonal complex A-III contains members of PEs 3, 5, 6 and 7 and clonal complex B´-I contains members of PEs 1, 6 and 11. There were no instances where a PE and a clonal complex were similarly demarcated.

Evidence of Recombination and/or Mutation

SNP Numbers and Patterns. A ratio of recombination to mutation for *Synechococcus* A-like and B´-like BACs of 5.6:1 and 7:1 was suggested based on the number of SNPs differentiating SLVs from consensus sequences in eBURST analysis (Tables 4.9 and 4.10) and the assumption that >1 single nucleotide polymorphism (SNP) is evidence of recombination (Feil et al., 1999). In cases where the concatenated consensus sequences were also the DVs of PEs, it was possible to compare SNP patterns across genes in PEVs grouped around the same dominant variant by either ES or eBURST (Figures 4.10-4.13). The variants most likely to have resulted from recombination, based on hundreds of SNPs in the same genes were detected in eBURST analysis, but not by ES, presumably due to the resulting large phylogenetic difference (blue highlight in figures). eBURST also detected variants with 1-26 SNPs, which were detected by ES as well (purple highlight in figures).

ES detected variants placed within ecotypes that had 1-54 SNPs (red highlight in figures), which were apparently not detected by eBURST because they have SNPs in >1 gene. Many SNPs within some genes suggest recombination, especially in A-like *rbsK*, *PK* and *CHP* and B´-like 16S rRNA/ITS, and to a lesser extent in A-like *lepB* and B´-like *aroA*. However, it is difficult to differentiate whether a limited number ($2<n\leq50$) of SNPs in different loci is the result of recombination or the accumulation of point mutations and, in reality, they are likely to have resulted from a combination of recombination and mutation. LDHat, which assays recombination and mutation among

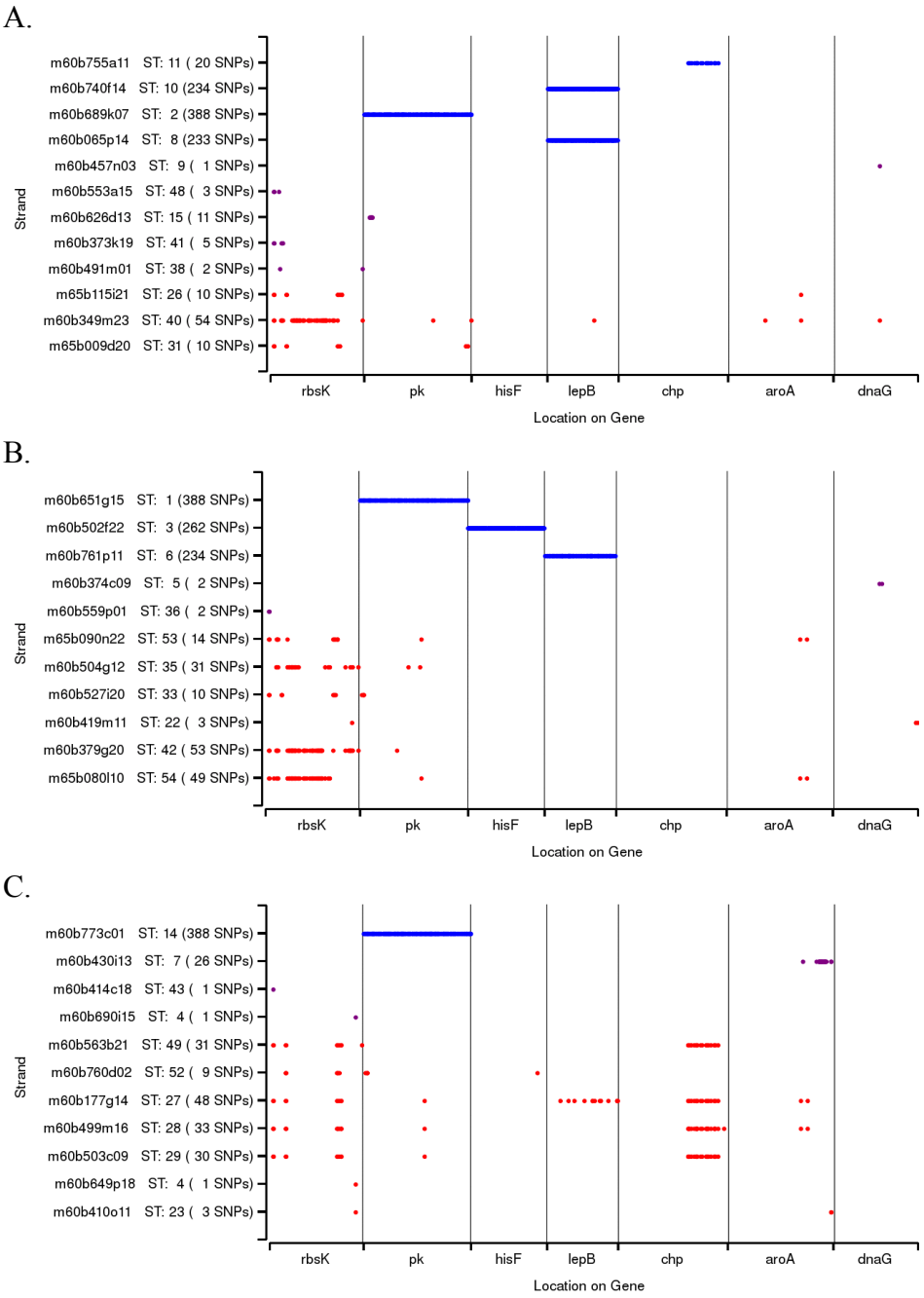



Figure 4.10. Single nucleotide polymorphism patterning in single locus variants (blue), putative ecotype variants (red) or both (purple) surrounding (A) DV-ST1 in PE A7 and clonal complex A-III, (B) DV-ST2 in PE A3 and clonal complex A-II and (C) DV-ST3 in PE A1 and clonal complex A-I defined by ecotype simulation and eBURST analysis of 7 loci in *Synechococcus* A-like BACs (also see Figures 4.2 and 4.7). More detailed SNP maps for individual genes are provided in Appendix D.

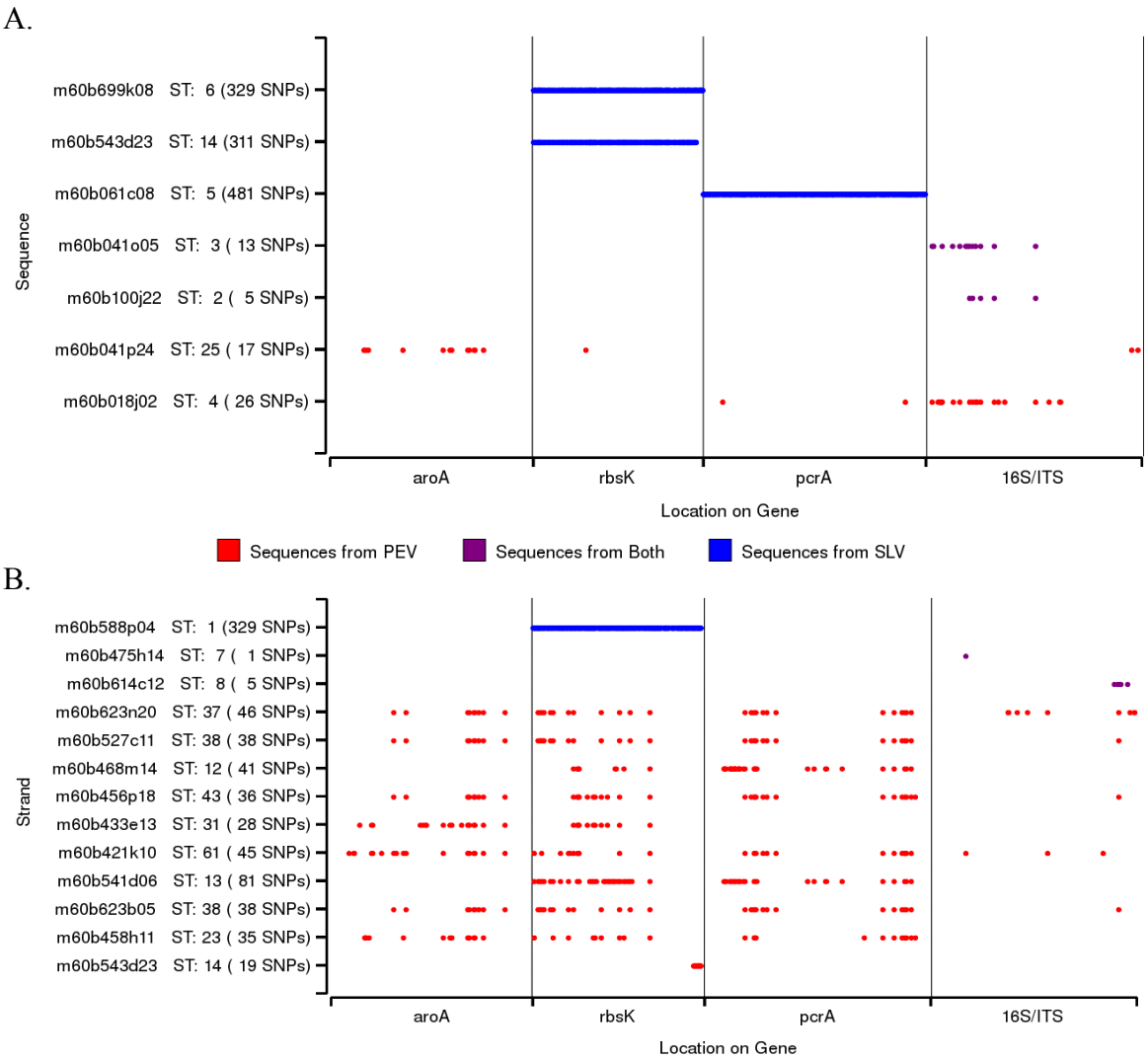



Figure 4.11. Single nucleotide polymorphism patterning in single locus variants (blue), putative ecotype variants (red) or both (purple) surrounding (A) DV-ST 1 in PE B´1 and clonal complex B´-I and (B) DV-ST6 in PE B´6 and clonal complex B´-II defined by ecotype simulation and eBURST in 4-locus analysis of *Synechococcus* B´-like BACs (also see Figures 4.3 and 4.8). More detailed SNP maps for individual genes are provided in Appendix D.

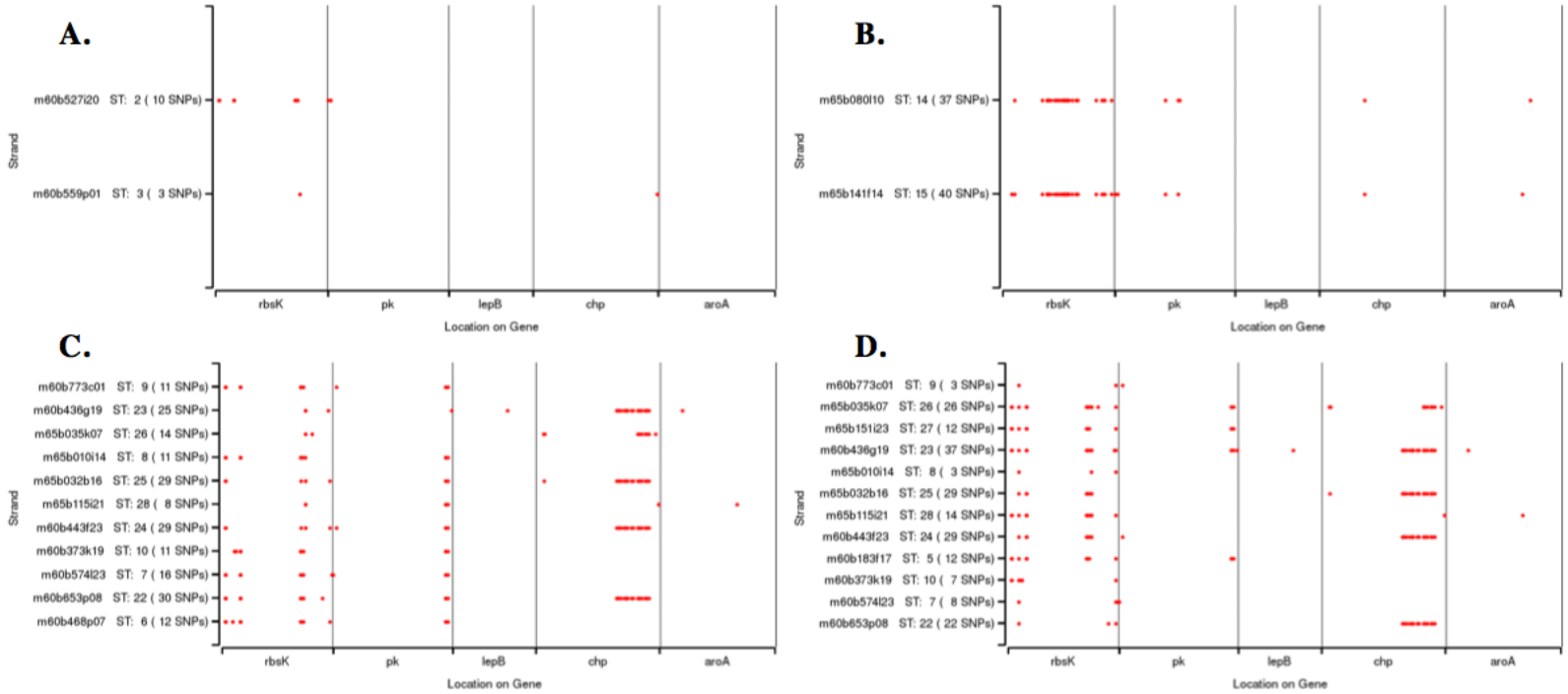
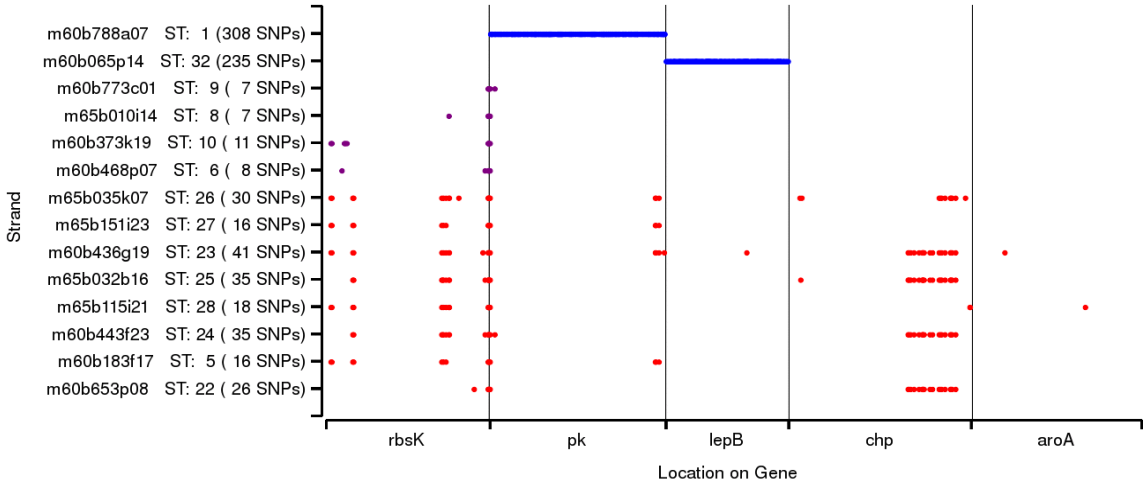



Figure 4.12. Single nucleotide polymorphism patterning in putative ecotype variants (red) surrounding (A) DV-ST1 in PE A5-1, (B) DV-ST4 in PE A5-2, (C) DV-ST5 in PE A5-5 and (D) sDV-ST6 in PE A5-5 defined by ecotype simulation in 5-locus analysis of *Synechococcus* A-like BACs (also see Figure 4.6 and 4.9). More detailed SNP maps for individual genes are provided in Appendix D.

A.




B.

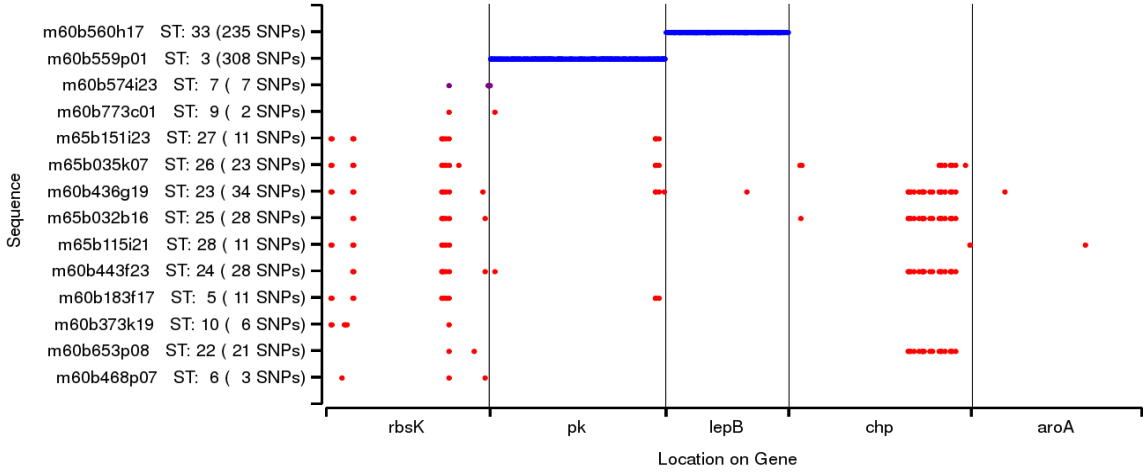



Figure 4.13. Single nucleotide polymorphism patterning in single locus variants (blue), putative ecotype variants (red) or both (purple) surrounding (A) sDV-ST7 in PE A5-5 and clonal complex A5-I and (B) sDV-ST8 in PE A5-5 and clonal complex A5-II defined by ecotype simulation and eBURST in 5-locus analysis of *Synechococcus* A-like BACs (also see figures 4.6 and 4.9).

sets, rather than pairs or triplicates of sequences, predicted recombination to mutation ratios of 2.87 for A-like (7-locus concatenation) and 5.15 for B´-like populations (4-locus concatenation) (Table 4.11). This more global analysis confirmed high relative recombination rates at A-like *CHP* and B´-like 16S rRNA/ITS loci, though recombination rates in other loci, where SNP analysis suggested possible recombination, such as A-like *rbsK*, *PK*, *aroA* and B´-like *aroA*, did not correspond as well. It is noted, however, that the sample size was small. Interestingly, some STs within clonal complexes and PEs show clear relationships, possibly reflecting patterns of evolution of diversity within these populations. For instance, STs 8 and 10 and STs 26 and 31 have similar *lepB* and *rbsK* SNP patterns, respectively (Figure 4.10A) and STs 27, 29 and 49 have similar *rbsK* and *CHP* SNP patterns relative to the DVs of PEA7 and PEA1 (Figure 4.12).

Tests for Recombination Signals. Evidence of recombination was also frequently detected using RDP3 methods (Tables 4.12, 4.13, C4.4, C4.5 and C4.6). More than half of the clones contained signals of recombination (clones denoted with asterisks in Figures 4.2-4.3 and 4.6). The majority of the recombination signals were detected in the *rbsK* locus for both *Synechococcus* A and B´-like BACs. Comparison of clones (STs) determined to be recombinants by RDP3 analysis with those in the SNP analysis showed some lack of correspondence. This may be due to several factors. (i) RDP3 had difficulty in detecting a recombinant that contains ~ 50 SNPs or less. For example, clones m60b349m23 (ST40), m60b379g20 (ST42), m65b80l10 (ST54), and m60b755all (ST11)

Table 4.11. Results from recombination and mutation rate and ratio analyses.

| **Population** | **Locus (No. of Loci)** | **No. of Nucleotides** | **No. of Segregating Sites** | **Mutation rate-θ (per site)** | **Recombination rate-ρ (per site)** | **ρ/θ** | **$r/m^{Feil}$** |
|---|---|---|---|---|---|---|---|
| A-like *Synechococcus* | 7 | 4074 | 1280 | 282.8 (0.0699) | 502.3 (0.201) | 2.87 | 5.6:1 |
| | 5 | 2917 | 990 | 237.2 (0.086) | 331.6 (0.121) | 1.4 | |
| | *aroA* | 614 | 30 | 12.3 (0.011) | 29.3 (0.027) | 2.4 | |
| | *lepB* | 467 | 221 | 90.4 (0.102) | 2.55 (0.003) | 0.028 | |
| | *dnaG* | 638 | 271 | 110.6 (0.105) | 0.013 ($1.3E^{-5}$) | $1.3E^{-4}$ | |
| | *hisF* | 470 | 260 | 124.8 (0.136) | 0.159 ($1.7E^{-4}$) | 0.001 | |
| | *CHP* | 649 | 23 | 12.5 (0.013) | 144.6 (0.15) | 11.5 | |
| | *PK* | 657 | 381 | 122.8 (0.186) | 1.077 (0.002) | 0.009 | |
| | *rbsK* | 579 | 135 | 35.9 (0.031) | 12.4 (0.011) | 0.345 | |
| B´-like *Synechococcus* | 4 | 2878 | 1429 | 301.3 (0.055) | 1553 (0.285) | 5.15 | 7:1 |
| | 16S rRNA | 733 | 417 | 102.9 (0.071) | 562.1 (0.390) | 5.5 | |
| | *aroA* | 615 | 353 | 95.6 (0.069) | 0.021 ($1.5E^{-5}$) | $2.1E^{-4}$ | |
| | *pcrA* | 864 | 429 | 102.9 (0.066) | 8.8 (0.006) | 0.086 | |
| | *rbsK* | 666 | 334 | 80.1 (0.069) | 7.88 (0.007) | 0.098 | |

Table 4.12. Analysis of recombination signals in *Synechococcus* A-like BAC sequences.

| Locus | Recombinant BAC | Sequence Type | Clonal Complex (DV-ST) | No. of SNP differences with consensus group and/or (dominant variant-PE) | Position | Parent | Evidence[m,n] |
|---|---|---|---|---|---|---|---|
| *rbsK* | M60B541E20[a] | 46 | [e] | [e] | 53-581 | M60B065P14 | RDP3 |
| *rbsK* | M60B504G12[a] | 35 | [e] | (29) | 208-359[g] | M60B436G19 | RDP3 |
| *rbsK* | M60B373K19[a,b] | 41 | A-III (1) | 5 (5) | 116-403 [g] | M60B195D22 | RDP3 |
| *aroA* | M60B430I13[a] | 7 | A-I (3) | 26 (26) | 567-616 [g] | OSB | RDP3 |
| *dnaG* | M60B595E04[a] | 50 | [e] | [e] | 1-25[h] | M60B589H13 | RDP3 |
| *PK* | M60B516C08[c] | 2 | A-III (1) | 388 | [i] | [i] | SNPs |
| *rbsK* | | | | 0 | 120-407 | M60B195D22 | RDP3 |
| *PK* | M60B166J20[c] | 2 | A-III (1) | 388 | [i] | [i] | SNPs |
| *rbsK* | | | | 0 | 120-407 | M60B195D22 | RDP3 |
| *PK* | M60B405K13[c] | 2 | A-III (1) | 388 | [i] | [i] | SNPs |
| *rbsK* | | | | 0 | 120-407 | M60B195D22 | RDP3 |
| *PK* | M60B689K07[c] | 2 | A-III (1) | 388 | [i] | [i] | SNPs |
| *rbsK* | | | | 0 | 120-407 | M60B195D22 | RDP3 |
| *PK* | M60B413O21[c] | 2 | A-III (1) | 388 | [i] | [i] | SNPs |
| *rbsK* | | | | 0 | 120-407 | M60B195D22 | RDP3 |
| *lepB* | M60B065P14[c] | 8 | A-III (1) | 233 | [i] | [i] | SNPs |
| *rbsK* | | | | 0 | 120-407 | M60B195D22 | RDP3 |
| *rbsK* | M60B457N03[c] | 9 | A-III (1) | 1 (1) | 120-407[g] | M60B195D22 | RDP3 |
| *lepB* | M60B740F14[c] | 10 | A-III (1) | 234 | [i] | [i] | SNPs |
| *rbsK* | | | | 0 | 120-407 | M60B195D22 | RDP3 |
| *CHP* | M60B755A11[c] | 11 | A-III (1) | 20 | [i] | [i] | SNPs |
| *rbsK* | | | | 0 | 120-407 | M60B195D22 | RDP3 |
| *PK* | M60B626D13[c] | 15 | A-III (1) | 11 (11) | [i] | [i] | SNPs |
| *rbsK* | | | | 0 | 120-407 | M60B195D22 | RDP3 |
| *rbsK* | M60B491M01[c] | 38 | A-III (1) | 2 (2) | 120-407[g] | M60B195D22 | RDP3 |
| *rbsK* | M60B553A15[c] | 48 | A-III (1) | 3 (3) | 120-407[g] | M60B195D22 | RDP3 |
| *PK* | M60B663O15[c] | 1 | A-II (2) | 388 | [i] | [i] | SNPs |
| *rbsK* | | | | 0 | 120-407 | M60B195D22 | RDP3 |
| *PK* | M60B651G15[c] | 1 | A-II (2) | 388 | [i] | [i] | SNPs |
| *rbsK* | | | | 0 | 120-407 | M60B195D22 | RDP3 |

Table 4.12 Continued…

| Locus | Recombinant BAC | Sequence Type | Clonal Complex (DV-ST) | No. of SNP differences with consensus group and/or (dominant variant-PE) | Position | Parent | Evidence[m,n] |
|---|---|---|---|---|---|---|---|
| *PK* | M60B471D04[c] | 1 | A-II (2) | 388 | [i] | [j] | SNPs |
| *rbsK* | | | | 0 | 120-407 | M60B195D22 | RDP3 |
| *PK* | M60B390N18[c] | 1 | A-II (2) | 388 | [i] | [j] | SNPs |
| *rbsK* | | | | 0 | 120-407 | M60B195D22 | RDP3 |
| *PK* | M60B729E01[c] | 1 | A-II (2) | 388 | [i] | [j] | SNPs |
| *PK* | M60B360H08[c] | 1 | A-II (2) | 388 | [i] | [j] | SNPs |
| *rbsK* | | | | 0 | 120-407 | M60B195D22 | RDP3 |
| *PK* | M60B701I08[c] | 1 | A-II (2) | 388 | [i] | [j] | SNPs |
| *rbsK* | | | | 0 | 120-407 | M60B195D22 | RDP3 |
| *PK* | M60B574I23[c] | 1 | A-II (2) | 388 | [i] | [j] | SNPs |
| *rbsK* | | | | 0 | 120-407 | M60B195D22 | RDP3 |
| *hisF* | M60B514A21[c] | 3 | A-II (2) | 262 | [i] | [j] | SNPs |
| *rbsK* | | | | 0 | 120-407 | M60B195D22 | RDP3 |
| *hisF* | M60B658F14[c] | 3 | A-II (2) | 262 | [i] | [j] | SNPs |
| *rbsK* | | | | 0 | 120-407 | M60B195D22 | RDP3 |
| *hisF* | M60B502F22[c] | 3 | A-II (2) | 262 | [i] | [j] | SNPs |
| *rbsK* | | | | 0 | 120-407 | M60B195D22 | RDP3 |
| *hisF* | M60B501M02[c] | 3 | A-II (2) | 262 | [i] | [j] | SNPs |
| *rbsK* | | | | 0 | 120-407 | M60B195D22 | RDP3 |
| *lepB* | M60B761P11[c] | 6 | A-II (2) | 234 | [i] | [j] | SNPs |
| *rbsK* | | | | 0 | 120-407 | M60B195D22 | RDP3 |
| *rbsK* | M60B559P01[c] | 36 | A-II (2) | 2 (2) | 120-407[g] | M60B195D22 | RDP3 |
| *rbsK* | M60B374C09[c] | 5 | A-II (2) | 0 | 120-407 | M60B195D22 | RDP3 |
| *rbsK* | M60B649P18[c] | 4 | A-I (3) | 1 (1) | 120-407[g] | M60B195D22 | RDP3 |
| *rbsK* | M60B690I15[c] | 4 | A-I (3) | 1 (1) | 120-407[g] | M60B195D22 | RDP3 |
| *PK* | M60B773C01[c] | 14 | A-I (3) | 388 | [i] | [j] | SNPs |
| *rbsK* | | | | 0 | 120-407 | M60B195D22 | RDP3 |
| *rbsK* | M60B414C18[c] | 43 | A-I (3) | 1 (1) | 120-407[g] | M60B195D22 | RDP3 |
| *PK* | M60B461H19[c] | 17 | A-IV (20)[f] | 388 | [i] | [j] | SNPs |
| *rbsK* | | | | 0 | 69-409 | M60B595E04[l] | RDP3 |
| *CHP* | M60B195D22[c] | 19 | A-IV (20)[f] | 2 | [i] | [j] | SNPs |
| *rbsK* | | | | 0 | 69-409 | M60B595E04[l] | RDP3 |

Table 4.12 Continued…

| Locus | Recombinant BAC | Sequence Type | Clonal Complex (DV-ST) | No. of SNP differences with consensus group and/or (dominant variant-PE) | Position | Parent | Evidence[m,n] |
|---|---|---|---|---|---|---|---|
| *rbsK* | M60B443F23[c] | 21 | A-IV (20)[f] | 0 | 69-409 | M60B595E04[l] | RDP3 |
| *rbsK* | M60B653P08[c] | 24 | A-IV (20)[f] | 10 | 120-407 | M60B195D22 | RDP3 |
| *rbsK* | M60B788A07[c] | 3 | A-II (2) | 0 | 120-407 | M60B195D22 | RDP3 |
| *hisF* | | | | 262 | [i] | [j] | SNPs |
| *rbsK* | M65B150M20[a,b] | 29[d] | [e] | [e] | 204-355[k] | M60B443F23 | RDP3 |

[a] RDP3-defined recombinant
[b] Sequences defined as recombinants in the *Synechococcus* A-like BAC 5-locus *rbsK* sequence dataset.
[c] Sequence identified as having the same recombination event as an RDP3-defined recombinant as evidenced by phylogenetic incongruency.
[d] ST29 in the M60/65 5-locus MLSA study does not equate to ST29 in the M60/65 7-locus MLSA study.
[e] Not found within a clonal complex or in a PE clade that contains a dominant variant.
[f] Sequence type represented by 1 sequence (clone: M60B399F15).
[g] RDP3 predicted a different position of change than suggested by the SNP map.
[h] RDP3 software unsure of beginning breakpoint site of recombination
[i] <50 SNPs or SNPs extended the length of the gene, therefore RDP3 may not detect as a recombinant.
[j] No detection or not defined.
[k] RDP3 software unsure of ending breakpoint site of recombination.
[l] RDP3 cautions that the parent and recombinant may be reversed.
[m] p-value cut off in RDP3 analysis, $p < 0.05$. P-values can be found in Table C4.4.
[n] Refer to Figures 4.10, 4.12 and 4.13 for SNP maps.

Table 4.13. Analysis of recombination signals in *Synechococcus* B´-like BAC sequences.

| Locus | Recombinant BAC | Sequence Type | Clonal Complex (DV-ST) | No. of SNP differences with consensus group and/or (dominant variant-PE) | Position | Parent | Evidence[j,k] |
|---|---|---|---|---|---|---|---|
| 16S rRNA | M60B100J22[a] | 2 | B-I (1) | 5 (5) | [d] | [h] | SNPs |
| 16S rRNA | M60B041O05[a] | 3 | B-I (1) | 13 (13) | [d] | [h] | SNPs |
| *pcrA* | M60B061C08[a] | 5 | B-I (1) | 481 | [d] | [h] | SNPs |
| *rbsK* | M60B636L06[a] | 6 | B-I (1) | 329 | [d] | [h] | SNPs |
| *rbsK* | M60B614H18[a] | 6 | B-I (1) | 329 | [d] | [h] | SNPs |
| *rbsK* | M60B518L02[a] | 6 | B-I (1) | 329 | [d] | [h] | SNPs |
| *rbsK* | M60B709P04[a] | 6 | B-I (1) | 329 | [d] | [h] | SNPs |
| *rbsK* | M60B699K08[a] | 6 | B-I (1) | 329 | [d] | [h] | SNPs |
| *rbsK* | M60B174D01[a] | 1 | B-II (6) | 329 | [d] | [h] | SNPs |
| *rbsK* | M60B089D09[a] | 1 | B-II (6) | 329 | [d] | [h] | SNPs |
| *rbsK* | M60B259J13[a] | 1 | B-II (6) | 329 | [d] | [h] | SNPs |
| *rbsK* | M60B588P04[a] | 1 | B-II (6) | 329 | [d] | [h] | SNPs |
| *rbsK* | M60B067L11[a] | 1 | B-II (6) | 329 | [d] | [h] | SNPs |
| 16S rRNA | M60B475H14[a] | 7 | B-II (6) | 1 (1) | [d] | [h] | SNPs |
| 16S rRNA | M60B614C12[a] | 8 | B-II (6) | 5 (5) | [d] | [h] | SNPs |
| *aroA* | OSB[a,l] | 66 | [c] | [c] | 70-134 | M60B015H24 | RDP3 |
| *rbsK* | M60B015H24[a] | 52 | [c] | [c] | 125-273 | OSA[m] | RDP3 |
| *pcrA* | M60B541D06[a,b] | 13 | [c] | (33) | 37-198[e] | OSA[i,m] | RDP3 |
| *rbsK* | | | | (48) | 136-344[e] | M60B579B21 | RDP3 |
| *rbsK* | M60B062L15[b] | 46 | [c] | [c] | 181-383 | OSA[m] | RDP3 |
| *rbsK* | M60B700J15[b] | 15 | [c] | [c] | 181-383 & 4[f]-54 | M60B038K14 & M60B347P16 | RDP3 |
| *rbsK* | M60B186K12[b] | 26 | [c] | [c] | 125-273[g] | OSA[m] | RDP3 |
| *rbsK* | M60B543D23[b] | 14 | 1 | 311 | 275-585[g] | M60B090B21 | RDP3 |
| *rbsK* | M60B554K06[b] | 32 | [c] | [c] | 126-499[g] | M60B772E10[i] | RDP3 |
| *rbsK* | M60B347P16[b] | 17 | [c] | [c] | 306[f]-460 | M60B579B21 | RDP3 |
| *rbsK* | M60B397N23[b] | 59 | [c] | [c] | 1[f] -306 | M60B433E13[i] | RDP3 |
| *rbsK* | M60B714K10[a] | 11 | [c] | [c] | 1[f] -306 | M60B433E13 | RDP3 |

Table 4.13 Continued…

| Locus | Recombinant BAC | Sequence Type | Clonal Complex | No. of SNP differences with consensus group and/or (dominant variant-PE) | Position | Parent | Evidence[j,k] |
|---|---|---|---|---|---|---|---|
| *rbsK* | M60B626O23[b] | 64 | [c] | [c] | 203-460[g] | M60B129N02 | RDP3 |
| *rbsK* | M60B085O17[a] | 10 | [c] | [c] | 4-54 | OSA[m] | RDP3 |
| *rbsK* | M60B057N07[a] | 9 | [c] | [c] | 4-54 | OSA[m] | RDP3 |
| *rbsK* | M60B426N01[a] | 47 | [c] | [c] | 4-54 | OSA[m] | RDP3 |
| *rbsK* | M60B046B18[a] | 55 | [c] | [c] | 125-273 | OSA[m] | RDP3 |
| *rbsK* | M60B081O16[a] | 57 | [c] | [c] | 125-273 | OSA[m] | RDP3 |
| *rbsK* | M60B091H06[a] | 26 | [c] | [c] | 125-273 | OSA[m] | RDP3 |
| *rbsK* | M60B250B15[a] | 58 | [c] | [c] | 125-273 | OSA[m] | RDP3 |
| *rbsK* | M60B040A11[a] | 54 | [c] | [c] | 125-273 | OSA[m] | RDP3 |
| *rbsK* | M60B456A21[a] | 18 | [c] | [c] | 1[f] -306 | M60B433E13[i] | RDP3 |
| *rbsK* | M60B477G06[a] | 62 | [c] | [c] | 1[f] -306 | M60B433E13[i] | RDP3 |
| *rbsK* | M60B513G21[a] | 16 | [c] | [c] | 1[f] -306 | M60B433E13[i] | RDP3 |
| *rbsK* | M60B648C08[a] | 16 | [c] | [c] | 1[f] -306 | M60B433E13[i] | RDP3 |
| *rbsK* | M60B769K22[a] | 16 | [c] | [c] | 1[f] -306 | M60B433E13[i] | RDP3 |
| *rbsK* | M60B579B21[a] | 63 | [c] | [c] | 1[f] -306 | M60B433E13[i] | RDP3 |

[a] Sequence identified as having the same recombination event as an RDP3-defined recombinant as evidenced by phylogenetic incongruency.
[b] RDP3-defined recombinant
[c] Not found within a clonal complex or in a PE clade that contains a dominant variant.
[d] <50 SNPs or SNPs extended the length of the gene, therefore RDP3 may not detect as a recombinant.
[e] RDP3 predicted a different position of change than suggested by the SNP map.
[f] RDP3 software unsure of beginning breakpoint site of recombination
[g] RDP3 software unsure of ending breakpoint site of recombination.
[h] No detection or not defined.
[i] RDP3 cautions that the parent and recombinant may be reversed.
[j] p-value cut off in RDP3 analysis, $p < 0.05$. P-values can be found in Table C4.5.
[k] Refer to Figure 4.11 for SNP maps.
[l] Genome annotation available at: http://cmr.jcvi.org/cgi-bin/CMR/GenomePage.cgi?org=gyma.
[m] Genome annotation available at: http://cmr.jcvi.org/cgi-bin/CMR/GenomePage.cgi?org=gymb.

among others all had 20-54 SNPs and were not detected by RDP3. However clustering of these SNPs within a locus (Figure 4.10) would suggest that a recombination event had occurred. (ii) It is possible that a potential recombinant suspected by SNP analysis not detected as a recombinant by at least 3 RDP3 methods as was the case for many clones that shared a recombination event or partial evidence of a recombination event. Clones m60b715i02-ST51, m60bg19-ST44, m60b609h12-ST34 and m60b349m23-ST40, for example, share the same recombination event at the *rbsK* locus as the primary identified recombinant, m60b504g12-ST35 (Figure 4.14). RDP3 analysis identified m60b504g12 as a recombinant, however the other recombinants (named above) did not have the support of significant p-values for 3 analyses within the RDP3 software package and were therefore not included in Table 4.12. (iii) A potential recombinant suspected by SNP analysis contained SNPs that spanned the entire length of the gene, as RDP3 will not detect a recombination signal if it does not have enough sequence to the left or right of the recombined sequence to determine breakpoints. For example, clones m60b740f14-ST10, m60b65p14-ST8 and m60b689k9-ST2 all SLVs were that contained >100 nt differences spanning the entire gene (*lepB* and *PK,* respectively) (Figure 4.10) but were not detected by RDP3 due to lack of enough sequence data on the ends to define breakpoints. A recombination analysis of the 5-locus sequence dataset for *Synechococcus* A-like BACs did not show preferential recombination between or within 60°C and 65°C clades and only one of 24 M65 BAC clones in the analysis (M65B150M20, Table 4.12), exhibited evidence of recombination suggesting that the *Synechococcus* A-like

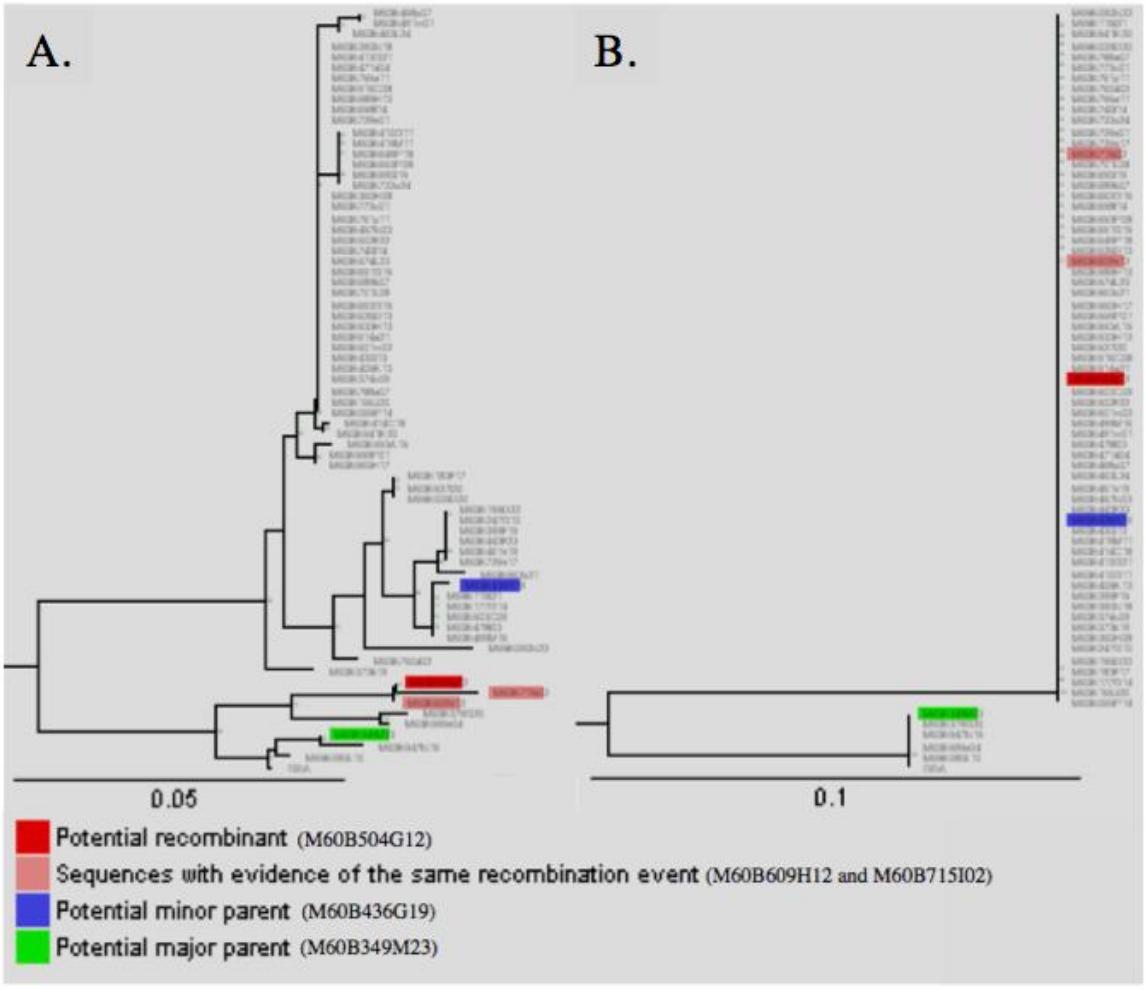



Figure 4.14. Lack of phylogenetic congruence and movement of recombinants among UPGMA phylogenies within the *rbsK* gene of *Synechococcus* A-like M60 BAC clones constructed from (A) non-recombinant regions and (B) recombinant region 208-359 of the sequence data.

population at 65°C may not be undergoing as much recombination as the population at 60°C.

Interestingly, further analysis of recombination determined that many recombinants 'traveled together' between phylogenies. Figures 4.14 and 4.15 show examples of this using the *rbsK* gene for both *Synechococcus* A-like and B´-like BACs. Phylogenies were constructed from recombinant and non-recombinant regions of the sequence data with a specific sequence defined as the 'initial recombinant' by RDP3. The other sequences were identified as carrying the same or partial evidence of the recombination event.

Linkage Disequilibrium

Significant linkage was detected using both the Monte Carlo simulation and parametric methods. Both *Synechococcus* A and B´-like population concatenated BAC sequences showed an $I_A^S > 0$ suggesting that linkage equilibrium does not exist in the dataset (Table 4.14). Testing different combinations of loci at different degrees of kb separation still resulted in linkage disequilibrium. The mean allelic diversity (i.e., proportion of polymorphic sites) was also determined to be lower for *Synechococcus* A-like BACs (0.478) than B´-like BACs (0.894).

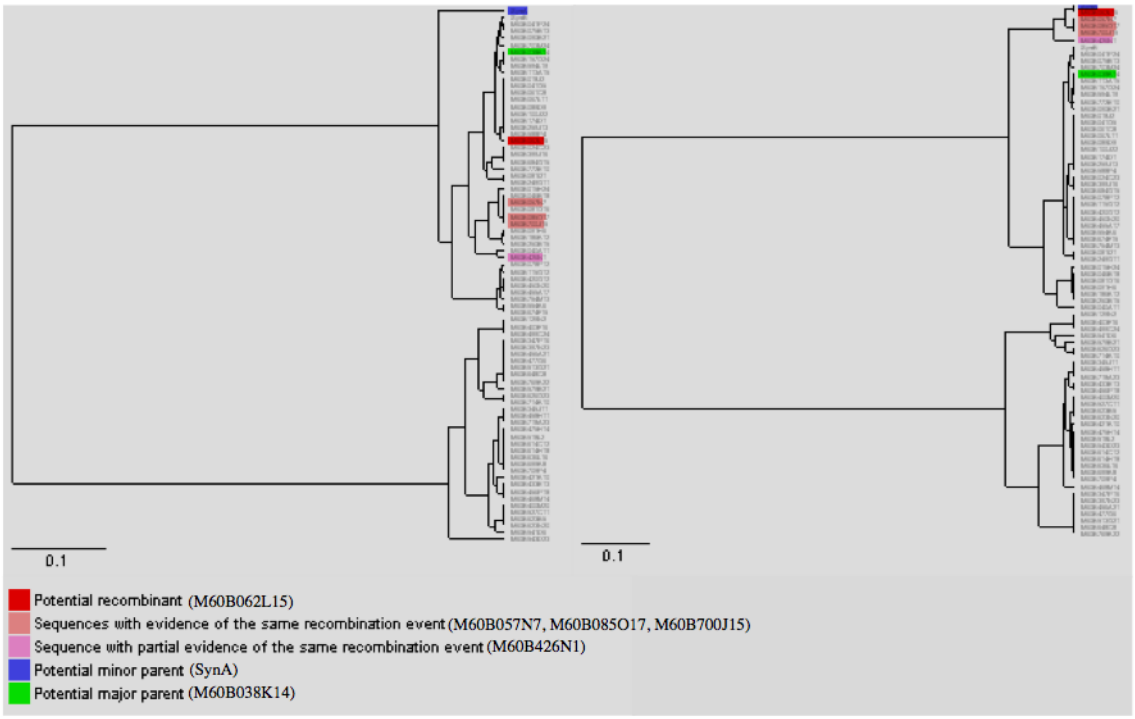



Figure 4.15. Lack of phylogenetic congruence and movement of recombinants among UPGMA phylogenies within the *rbsK* gene of *Synechococcus* B´-like M60 BAC clones constructed from (A) non-recombinant regions and (B) recombinant region 181-393 of the sequence data.

Table 4.14. Linkage disequilibrium results from analysis of concatenated sequence data sets.

| **Organism** | **Sample Size** | **No. of Loci** | **Mean Allelic Diversity*** **+/-** | **$V_D$[a]** | **$V_e$[b]** | **$L_{MC}$[c]** | **$L_{PARA}$[d]** | **$I_A^S$[e]** | **Linkage Detected[f]** |
|---|---|---|---|---|---|---|---|---|---|
| *Synechococcus* A-like | 66 STs | 7 | 0.478 ± 0.084 | 2.27 | 1.45 | 1.69 | 1.67 | 0.094 | Yes |
| *Synechococcus* B´-like | 77 STs | 4 | 0.894 ± 0.033 | 0.63 | 0.36 | 0.39 | 0.39 | 0.239 | Yes |

* Proportion of polymorphic sites.
[a] $V_D$: The observed mismatch variance.
[b] $V_e$: The expected mismatch variance at equilibrium.
[c] $L_{MC}$: Simulated 5% critical value.
[d] $L_{PARA}$: Calculated 5% critical value assuming a normal distribution: $L = V_e + 1.654\sqrt{Var(V_D)}$ (Haubold and Hudson, 2000).
[e] $I_A^S$: Standardized index of association.
[f] Detected at a p-value < 0.01

## Discussion

The use of BAC clone libraries for multi-locus sequence analysis has enabled a detailed exploration of the genetics of natural populations of *Synechococcus* predominating the well-studied hot spring microbial mat in Mushroom Spring without bias due to cultivation. By focusing on BAC clones containing 16S rRNA loci the results could be interpreted in light of the extensive work on this and the adjacent ITS locus (Ward et al., 2006). Analysis of this MLSA database revealed multiple lines of evidence that recombination has been historically significant in the evolution of diversity within the *Synechococcus* A/B lineage, but that it has been insufficient in intensity to erode the phylogenetic coherence of ecologically specialized populations that can be considered ecological species. Analyses of the MLSA database also revealed that a greater number of ecological species are detected than with many single loci, except when a locus is undergoing frequent recombination events.

Evidence for genomic plasticity was first observed as the database was assembled in the form of BAC clones that, despite repeated PCR attempts, could not be shown to contain loci expected, assuming conservation of gene order with the *Synechococcus* strain A and B´ genomes. The poor recovery of B´-like BACs containing a sufficient number of loci for MLSA may indicate that genes had been relocated on the genomes of members of this population. A possible mechanism for the relocation or "erosion" of genes in 16S rRNA genomic neighborhoods may be genomic inversions, which can occur via recombination or transfer mediated by a transposon or mobile element (Melendrez et al., 2010b; Chapter 3). Few A-like, but many B´-like BACs contained an insufficient number

of expected loci (sometimes none), and this caused us to focus on fewer loci in MLSA analysis of the B′ population. It is not known whether BACs with insufficient numbers of loci represent additional B′ PEs or are variants of PEs observed in this study. The greater estimated evolutionary distance of the B′-like population also indicates a longer (or more rapid) evolutionary history than experienced by the A-like populations (Table 4.3; see also Ward et al. (1998), Bhaya et al. (2007), and Melendrez et al. (2010a); Chapter 2).

Evidence of homologous recombination, though not always corresponding across methods of analysis, came from (i) the incongruency of phylogenies created from different individual loci and concatenated MLSA loci sampled from the BAC clones, (ii) SNP number and distribution patterns relative to STs dominating PEs and clonal complexes suggested by ES and eBURST and (iii) multiple methods available for the global analysis of recombination in populations (i.e., methods in RDP3).

The eBURST analysis does not account for the degree of change that defines one allele as different from another, therefore sequences that may be quite far apart in phylogenetic history at one locus may be grouped by eBURST analysis because the sequences share other loci that are identical in nucleotide composition. Conversely, eBURST may not group sequences into the same clonal complexes due to differences in alleles at more than one locus, even though those alleles may differ from the consensus sequence by only a few polymorphisms, making them phylogenetically closely related. It is also important to note, however, that in many cases the consensus groups identified by eBURST analysis were also the DVs in PEs predicted from ES analysis (Tables 4.6 and

4.8). Two exceptions were observed: (i) ST20, the consensus group for clonal complex A-IV of the A-like*Synechococcus* 7-locus MLSA was a singleton, and (ii) ST1 was the dominant variant of PE1, whereas ST7 was the consensus group of the corresponding A5-I clonal complex in the 5-locus MLSA of the *Synechococcus* sp A-like population.

As with the single-locus studies (Chapter 2) the MLSA study was limited with respect to sampling, possibly resulting in an underestimation of number of ecotypes predicted. However, despite low sampling, the limited number of loci studied (for the *Synechococcus* B´-like population) and the influence of recombination did not disrupt the dominant clade structure or the ecological associations.

Both ES and eBURST predicted *Synechococcus* A-like and B´-like populations centered around dominant STs (called consensus sequences in eBURST) (Figures 4.2, 4.3, 4.7 and 4.8), which may represent founders of populations. However, ES and eBURST differed with respect to the variants grouped into populations centered by these dominant variants (Table 4.6 and 4.8, Figures 4.10-4.13). SNP distribution analysis revealed that ES grouped variants that have not diverged enough to lose phylogenetic coherence with the dominant variant from which they presumably arose. Variants that might have diverged due to recombination at any locus or combination of loci, such that they are no longer found within the clade containing the dominant variant from which they arose, were not demarcated as members of the ecotype. eBURST grouped variants that differ at only one locus (or possibly two loci) no matter how different the variant is from the dominant variant. It relies on identity of the alleles at the remaining loci as a measure of clonal complex coherence, but excludes variants found to be phylogenetically

similar because they differ at two or more loci (even if only by a few nucleotide changes).

To determine whether the populations predicted by ES to be PEs or by eBURST to be clonal complexes resemble populations that have a distinct ecology, it was necessary to construct a second dataset in which an equal sampling of BACs from the M60 and M65 samples was possible. It is important to note that the DVs themselves exhibit patterns of distribution and gene expression consistent with their being ecologically distinct (Becraft et al. 2010). If the populations predicted by ES and/or eBURST are ecologically distinct units, both the DVs (consensus groups) and the variants predicted to be associated with them should exhibit the same distribution pattern. ES analysis of this 5-locus dataset revealed 8 PEs, 4 of which were sample-specific (Figure 4.6), suggesting cases where the DVs and their co-occurring variants do have the same distribution. Analysis using eBURST of the same dataset revealed two nearly sample-specific clonal complexes (Figure 4.9), which corresponded to PEs that were not observed to be sample-specific, and, if the stringency of clonal complexes was relaxed to include DLVs, one sample-specific clonal complex emerged (Figures C4.4).

Another analysis that may help clarify habitat associations, or lack thereof, with ecotypes is AdaptML (Hunt et al., 2008). AdaptML provides analysis of the correlation between ecotypes predicted and investigator-defined habitats. Our ability to use AdaptML in the current work was limited, since only two ‘habitats’ could be defined (i.e., two samples that different by only 5°C). We know from other distribution studies (Ramsing et al., 2000; Ward et al., 2006; Becraft et al., 2010; Bhaya et al., 2007; Adams

et al., 2008) that there are more possible niche-defining features in the mat system, including light intensity and quality and differences in nutrient availability, but these were not analyzed in the current study. Thus, we will evaluate AdaptML in future work, where samples are collected with greater spatial resolution along flow and vertical gradients.

Expansion of the techniques available for population genetics studies, such as single-cell whole genome amplification (multiple displacement amplification-MDA), may also clarify the emerging ecological relationships we are seeing in this limited dataset. MDA would enable a more robust analysis of multiple loci or the whole core genome. While multiple displacement amplification remains a promising technique for population genetic studies of environmental diversity, preliminary studies indicate that this technology still has many limitations and would also require thousands of cells be picked by micromanipulation and analyzed to obtain a subset useful for MLSA (Chapter 5).

ES appears to be capable of grouping phylogenetically related variants into sample-specific (i.e., ecologically distinct) clades, but variants that have undergone recombination with phylogenetically more distant individuals at the locus being studied may be excluded. eBURST appears to detect fewer, less well defined sample-specific populations (clonal complexes), excluding close, but capturing more distant phylogenetic relatives, often of the same DV. Population genetics analyses such as these attempt to model the existence of natural evolved populations that are ecologically distinct. Recombination in the taxa and loci we studied appears to have blurred the boundaries of

PE populations and increased the difficulty of detecting all of the members of such populations. However, we can now hypothesize that STs that are phylogenetically distant enough that ES does not include them in a particular PE, which may be suggested by eBURST analysis, are nevertheless associated with individual *organisms* that do belong to this PE population. Since individual organisms of this kind should share the ecological features typical of the PE, despite their phylogenetic distance at particular loci, we can test such hypotheses by studying their co-distribution and gene expression with the DV of the PE and its close phylogenetic relatives.

## Acknowledgements

We appreciate the long-term support from the National Science Foundation Frontiers in Integrative Biology Research Program (EF-0328698), NASA Exobiology program (NAG5-8807 and NX09AM87G) and the DOE Pacific Northwest National Laboratory (contract pending). In addition we appreciate the assistance from National Park Service personnel at Yellowstone National Park. Special thanks are given to Darren Martin in assistance with the RDP3 software suite and Jason Wood for assistance with the creation of SNP graphics.

References

Adams MM, Gómez-García MR, Grossman AR and Bhaya D. (2008). Phosphorus deprivation responses and phosphonate utilization in a thermophilic *Synechococcus* sp. from microbial mats. *J Bacteriol* **190**: 8171-8184.

Allewalt JA, Bateson MM, Revsbech NP, Slack K and Ward DM. (2006). Temperature and light adaptations of *Synechococcus* isolates from the microbial mat community in Octopus Spring, Yellowstone National Park. *Appl Environ Microbiol* **72**: 544-550.

Altschul SF, Gish W., Miller W, Myers EW and Lipman, DJ. (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403-410.

Archer J and Robertson DL. (2007). CTree: comparison of clusters between phylogenetic trees made easy. *Bioinformatics* **23**: 2952-2953.

Becraft ED, Cohan FM, Kuhl M, Jensen S and Ward DM. (2010). Identifying and improving the existence of ecologically defined *Synechococcus* sp. in Mushroom Spring, Yellowstone National Park. In prep.

Bhaya D, Grossman AR, Steunou AS, Khuri N, Cohan FM, Hamamura N, Melendrez MC, Bateson MM, Ward DM, and Heidelberg JF. (2007). Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analysis. *ISME J* **1**: 703-713.

Bilek N, Ison CA and Spratt BG. (2009). Relative contribution of recombination and mutation to the diversification of the *opa* gene repertoire of *Neisseria gonorrhoeae*. *J Bacteriol* **191**: 1878-1890.

Boni MF, Posada D and Feldman MW. (2007). An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genet* **176**: 1035-1047.

Cesarini S, Bevivino A, Tabacchioni S, Chiarini L and Dalmastri C. (2009). *RecA* gene sequence and multilocus sequence typing for species-level resolution of *Burkholderia cepacia* complex isolates. *Lett Appl Microbiol* **49:**580-8

Cohan FM and Perry EB. (2007). A systematics for discovering the fundamental units of bacterial diversity. *Curr Biol* **17**: R373-R386.

Dingle KE, Colles FM, Falush D and Maiden MCJ. (2005). Sequence typing and comparison of population biology of *Campylobacter coli* and *Campylobacter jejuni*. *J Clin Microbiol* **43**: 340-347.

Feil EJ, Maiden MCJ, Achtman M and Spratt BG. (1999). The relative contribution of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. *Mol Biol Evol* **16**: 1496-1502.

Feil EJ, Li BC, Aanensen DM, Hanage WP and Spratt BG. (2004). eBURST: Inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J Bacteriol* **186**: 1518-1530.

Feil EJ, Maynard Smith J, Enright MC and Spratt BG. (2000). Estimating recombination parameters in *Streptococcus pneumoniae* from multi-locus sequence typing data. *Genet* **154**: 1439-1450.

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

Ferris MJ and Ward DM. (1997). Season distributions of dominant 16S rRNA-defined populations in a hot spring microbial mat examined by denaturing gradient gel electrophoresis. *Appl Environ Microbiol* **63**: 1375-1381.

Gibbs MJ, Armstrong JS and Gibbs AJ. (2000). Sister scanning: a monte carlo procedure for assessing signals in recombination sequences. *Bioinformatics* **16**: 573-582.

Hanage WP, Fraser W and Spratt BG. (2006). Sequences, sequence clusters and bacterial species. *Phil Trans Roy Soc B*. **361**: 1917-1927.

Hanage WP, Fraser C and Spratt BG. (2005). Fuzzy species among recombinogenic bacteria. *BMC Biology* **3**: 6.

Haubold B and Hudson RR. (2000). LIAN 3.0: detecting linkage disequilibrium in multilocus data. *Bioinformatics* **16**: 847-848.

Holmes EC, Worobey M and Rambaut A. (1999). Phylogenetic evidence for recombination in Dengue virus. *Mol Biol Evol* **16**: 405.

Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ and Polz MF. (2008). Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* 320: 1081-1085.

Klatt CG, Wood JM, Rusch DB, Bateson MM, Heidelberg JF, Bryant DA and Ward DM. (2010). Composition of metagenomes from a phototrophic hot spring microbial mat community. In prep.

Koeppel A, Perry EB, Sikorski J, Krizanc D, Warner A, Ward DM, Rooney AP, Brambilla E, Connor N, Ratcliff RM, Nevo E and Cohan FM. (2008). Identifying the fundamental units of bacterial diversity: A paradigm shift to incorporate ecology into bacterial systematics. *Proc Natl Acad Sci* **105**: 2504-2509.

Markowitz VM, Mvromatis K, Ivanova N, Chen IA, Chu K and Kyrpides NC. (2009). IMG ER: A system for microbial genome annotation expert review and curation. *Bioinformatics* **25**: 2271-2278.

Markowitz VM, Szeto E, Palaniappan K, Grechkin Y, Chu K, Chen I-M, Dubchak I, Anderson I, Lykidis A, Mavromatis K, Ivanova N, and Kyrpides NC. (2008). The Integrated Microbial Genomes (IMG) system in 2007: data content and analysis tool extensions. *Nuc Acids Res* **36**: D528-533.

Martin DP, Williamson C and Posada D. (2005). RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics* **2**: 260-262.

Maynard Smith J. (1992). Analyzing the mosaic structure of genes. *J Mol Evol* **34**: 126-129.

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR and Donnelly P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science* **304**: 581-584.

Melendrez MC, Lange RK, Cohan FM and Ward DM. (2010a). Ecological diversity of *Synechococcus* spp. inhabiting an alkaline siliceous hot spring in Yellowstone National Park, WY measured using protein-encoding genes and evolutionary simulation. *ISME J* In prep.

Melendrez MC, Wood JM, Rusch DB, Heidelberg JF and Ward DM. (2010b). Bacterial artificial chromosome (BAC) libraries for Mushroom Spring cyanobacterial mat, Yellowstone National Park, WY. In prep.

Padidum M, Sawyer S and Fauquet CM. (1999). Possible emergence of new Geminiviruses by frequent recombination. *Virology* **265**: 218-225.

Page RDM. (1996). TREEVIEW: An application to display phylogenetic trees on personal computers. *Comp Appl Biosci* **12**: 357-358.

Papke TR, Koenig JE, Rodriguez-Valera F and Doolittle WF. (2004). Frequent recombination in a saltern population of *Halorubrum. Science* **306**: 1928-1929.

Posada D and Crandall KA. (2001). Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc Natl Acad Sci* **98**: 13757-13762.

Ramsing NB, Ferris MJ and Ward DM. (2000). Highly ordered vertical structure of *Synechococcus* population within the one-millimeter-thick photic zone of a hot spring cyanobacterial mat. *Appl Environ Microbiol* **66**: 1038-1049.

Salerno A, Deletoile A, Lefevre M, Ciznar I, Krovacek K, Grimont P and Brisse S. (2007). Recombining population structure of *Plesiomonas shigelloides* (*Enterobacteriaceae*) revealed by multilocus sequence typing. *J Bacteriol* **189**: 7808-7818.

Spratt BG, Hanage WP, Li B, Aanensen DM and Feil EJ. (2004). Displaying the relatedness among isolates of bacterial species—the eBURST approach. *FEMS Microbiol Lett* **241**: 129–134.

Tamura K, Dudley J, Nei M and Kumar S. (2007). MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**: 1596-1599.

Tamura K, Nei M and Kumar S. (2004). Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc Natl Acad Sci* **101**:11030-11035.

Tanabe Y, Sano T, Kasai F and Watanabe MM. (2009). Recombination, cryptic clades and neutral molecular divergence of the microcystin synthetase (mcy) genes of the toxic cyanobacterium *Microcystis aeruginosa*. *BMC Evol Biol* **9**:115.

Ward DM, Bateson MM, Ferris MJ, Kuhl M, Wieland A, Koeppel A and Cohan FM. (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. *Phil Trans Roy Soc B* **361**:1997-2008.

Ward DM, Ferris MJ, Nold SC and Bateson MM. (1998). A natural view of microbial biodiversity within hot spring cyanobacterial mat communities. *Microbiol Mol Biol Rev* **62**: 1353-1370.

Ward DM, Weller R and Bateson MM. (1990). 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature* **345**: 63-65.

Whitaker RJ, Grogen DW and Taylor JW. (2005). Recombination shapes the natural population structure of the hyperthermophilic archaeon *Sulfolobus islandicus*. *Mol Biol Evol* **22**: 2354-2361.

Vitorino LR, Margos G, Feil EJ, Collares-Pereira M, Ze-Ze L and Kurenbach K. (2008). Fine-scale phylogeographic structure of *Borrelia lusitaniae* revealed by multilocus sequence typing. *PLOS ONE* **3**: 1-13.

# CHAPTER 5

# MULTIPLE DISCPLACEMENT AMPLIFICATION OF SINGLE-CELL GENOMES FOR CULTIVATION-INDEPENDENT MULTI-LOCUS SEQUENCE ANALYSIS OF *SYNECHOCOCCUS* POPULATION GENETICS: A PILOT STUDY[9]


## Abstract

Multiple displacement amplification (MDA) of genomic DNA of single cells of *Synechococcus* obtained from laboratory cultures and Octopus Spring and Mushroom Spring mats in Yellowstone National Park was evaluated for its utility for cultivation-independent multi-locus population genetics studies. MDAs were analyzed using aggressive PCR protocols to amplify 5 protein-encoding loci and the 16S rRNA gene, followed by sequencing of the PCR products. The percentage of MDAs that contained the 16S rRNA gene plus 1-5 additional specific loci decreased with the number of protein-encoding loci. One thousand to 3,000 MDAs containing a 16S rRNA gene would be needed to permit a multi-locus sequence analysis on 71 MDAs based on the 16S rRNA gene and 5 additional protein-encoding loci.


[9] This study was conducted in collaboration with Dr. Thomas Ishoey and Dr. Robert Laskin from the J. Craig Venter Institute, Rockville, MD. Single cells were picked, MDA reactions, and, in some cases prescreening of MDAs for the presence of a eubacterial 16S rRNA gene by quantitative PCR, were performed at that institute. Amplified, and in some cases screened, genome amplification reaction products were then sent to me at Montana State University, Bozeman, MT where I performed multi-locus sequence analysis and evaluation of MDA as a technique for population genetics studies. PCR primers and protocols for the *psaA* locus were designed and provided by Eric Becraft.

## Introduction

A new technique in molecular biology emerged in 2001, whole genome amplification, that promised to reduce the cost of obtaining genomic sequence data, eliminate biases introduced by cloning methods, and provide data on previously uncultivated microorganisms in many different ecological systems. Multiple displacement amplification (MDA) was the first whole genome amplification method that was isothermal. MDA uses a DNA polymerase from bacteriophage phi29 derived from *Bacillus subtilis* because of its high processivity and proofreading fidelity and strong strand displacing activity. Thus the sequences that are obtained are highly reliable. The reaction produces double-stranded linear DNA, single-stranded DNA and some branched intermediate structures (Binga et al., 2008; Figure 5.1). MDA can yield up to 40 μg of DNA of average length of ~12 kbp per 50 μl reaction (Binga et al., 2008; Lasken 2009; Sorensen et al., 2007).

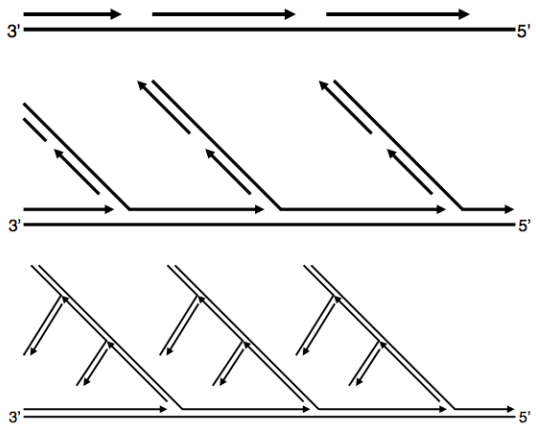



Figure 5.1. Diagrammatic representation of branched structures formed as a multiple displacement amplification reaction progresses (top to bottom). The arrowheads represent the location of the phi29 polymerase synthesizing DNA in the 5' to 3' direction (from Binga et al., 2008, with permission).

MDA holds great potential for the amplification of whole genomes from low-biomass samples and increasing access to the genetics of uncultivated organisms. MDA has been successfully used to amplify ~5 fg of single cell *Escherchia coli* DNA template to ~24 μg of product (Raghunathan et al., 2005), human DNA (Dean et al., 2002) and genomes of uncultured soil microbes collected using fluorescence *in situ* hybridization of the 16S rRNA gene and micromanipulation (Lasken et al., 2005). Research has shown that only 10 bacterial cells are sufficient for generating high-quality DNA that covers an entire genome (Raghunathan et al., 2005).

Limitations of single-cell MDA include (i) nonspecific amplification-derived primer dimer formation, (ii) contaminating DNA template, (iii) formation of chimeras, and (iv) incomplete genomic coverage, causing a 'representational bias' of fragments along the genome. The concentration and copy number of the DNA template also affects the MDA reaction. Reducing the concentration or template copy number (amplifying from a single cell for example instead of several cells) can potentially increase bias and even result in failure to recover of some sequence because some regions of the genome might be more readily amplifiable than others (Wu et al., 2006; Lasken, 2009; Neufeld et al., 2008).

The appeal of MDA for population genetics studies is the ability to conduct multi-locus analyses without culturing isolates, because all loci can be linked to an individual genome, regardless of location on the genome. Previous multi-locus studies that have utilized MDA in place of cloning provided enough amplified genomic product for virtually unlimited downstream PCR analyses, however, locus amplification efficiency

greatly varied (Stepanauskas and Sieracki, 2007; Havryliuk et al., 2008). Those studies that have been successful at amplifying multiple genes from a single MDA reaction used small numbers of cells (e.g., 10-100) as a template, but these studies were aimed at obtaining enough product to sequence full genome(s) rather than assessing the genetics of a population (Fernandez-Ortuno et al., 2007; Aviel-Ronen et al., 2006; Kvist et al., 2007; Sun et al., 2007; Raghunathan et al., 2005). It is important to note, that for the purposes of population genetics it is not necessary to obtain the entire genome but simply the genes required for multi-locus analysis.

In this chapter I present the results of a pilot study in which collaborators at The J. Craig Venter Institute (Rockville, MD) selected single *Synechococcus* cells using micromanipulation, then used MDA to amplify genomic DNA and, in some cases, verified that the MDA reactions had amplified bacterial 16S rRNA genes (Lasken et al., 2005). I used these MDAs as templates for PCR amplification of the 16S rRNA gene and several single protein-encoding loci. Previous work has shown that while Bacterial Artificial Chromosome (BAC) clone library construction is effective at linking loci in cultivation-independent studies of native mat *Synechococcus* populations, it is a costly, time-consuming protocol and hundreds of thousands of BACs need to be prescreened in order to yield a small subset useful for multi-locus analyses giving it a low efficiency (Melendrez et al., 2010b; see Chapter 4). I examined the possibility that single-cell amplification with MDA might decrease the workload associated with obtaining enough individuals with multiple sequenced loci for population genetics studies.

## Methodology

### Picking Individual *Synechococcus* Cells

Single *Synechococcus* cells were picked using capillary micromanipulation, exploiting their red autofluorescence due to the presence of chlorophyll *a* to help avoid contamination with non-cyanobacterial cells (Lasken et al., 2005) (Figure 5.2). Two batches of cells were collected for the study. One batch consisted of 23 cells, eleven of which were from a culture of *Synechococcus* strain A (JA-3-3Ab), and 12 of which were from a culture of *Synechococcus* strain B´ (JA-2-3B´a (2-13)) (Allewalt et al., 2006). This batch of cells was shown to contain a eubacterial 16S rRNA gene based on prescreening with quantitative PCR (qPCR). The second batch, which was not screened for the presence of a eubacterial 16S rRNA gene, contained 96 cells, all of which were picked from a homogenized mat sample collected on 2 October 2003 from Mushroom Spring, YNP at 60$^{o}$C (termed MS60) (44.5386°N, 110.7979°W). Homogenization of mat samples was done with a dounce tissue homogenizer.

### MDA Amplification and Screening for 16S rRNA Genes

Picked cells were lysed with denaturing and neutralizing solutions as described by Lasken et al. (2005) and the lysates were used within one hour for MDA reactions, which were carried out using the Repli-G kit (Qiagen) according to the manufacturers protocols. Taqman qPCR analysis for the presence of a eubacterial 16S rRNA gene has been previously described (Hosono et al., 2003 and Marcy et al., 2007).

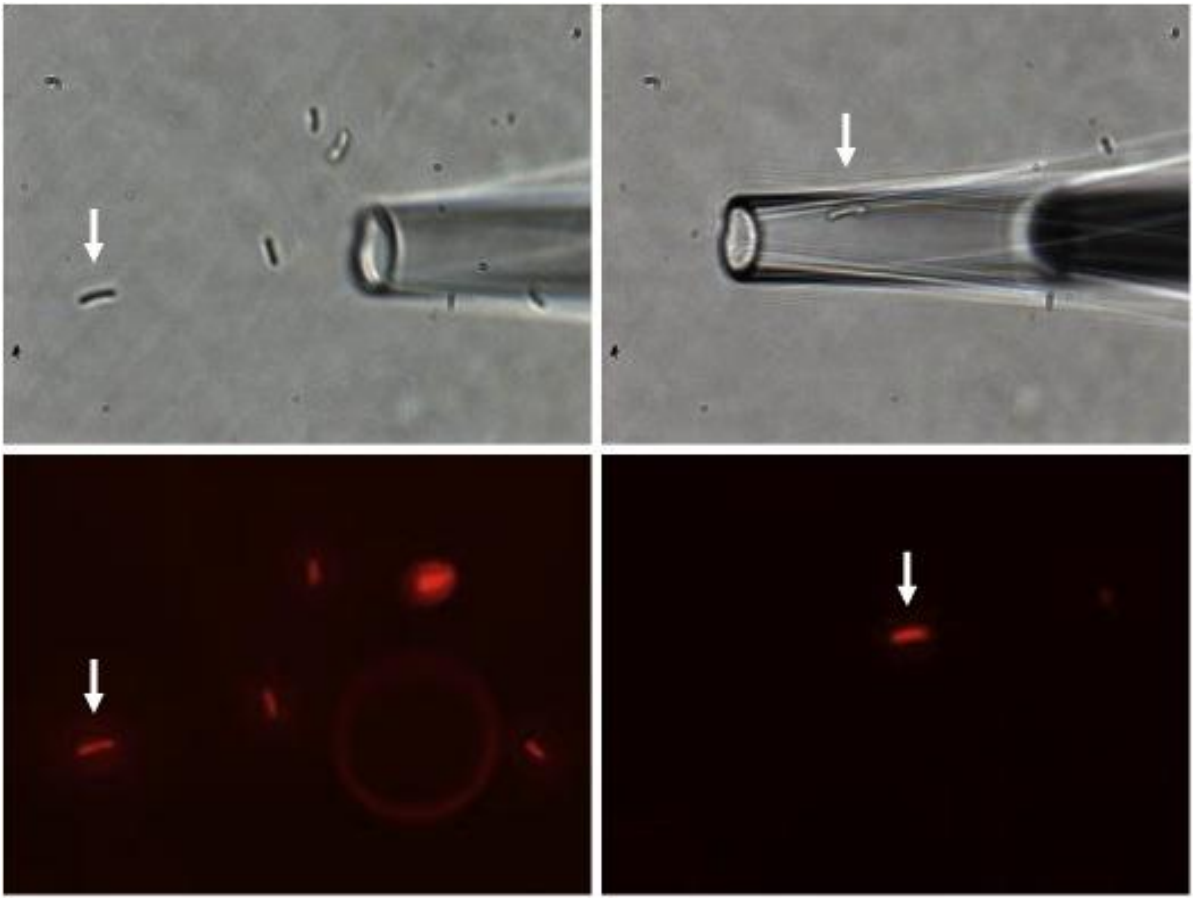


Figure 5.2. Capillary micromanipulation and capture of *Synechococcus* cells under the microscope using autoflourescence to verify the capture of *Synechococcus*. (used with the permission of Dr. Roger Laskin, J. Craig Venter Institute).

PCR Amplification of 16S rRNA and Protein-Encoding Genes

PCR primers and protocols for amplification of eubacterial or cyanobacterial 16S rRNA (Ferris et al., 1996) and *apcAB*, *rbsK, aroA* and *pcrAF* genes (Melendrez et al., 2010a; Chapter 2), which were used in the BAC cloning approach to MLSA analysis targeting a 16S rRNA region of the *Synechococcus* strain A and B´ genomes, were previously described, except that all MDAs were subjected to 40 PCR cycles. The *psaA* locus encoding the core protein for the reaction center of photosystem I, which is under

investigation by Becraft et al., (2010), was amplified using primers specific for the *psaA* gene in members of the *Synechococcus* A/B lineage (*psaA*F259: ctgagcggcatgtactacca and *psaA*R(2)855: caggccacccttgaaggtc). The cycling conditions for *psaA* were different than for all other loci and were as follows: an initial denaturing step at 94$^{o}$C for 5 min was followed by 40 cycles of 94$^{o}$C (45 sec), 53$^{o}$C (45 sec), and 72$^{o}$C (1 min 30 sec); an extension was done at 72$^{o}$C for 7 min followed by storage at 4$^{o}$C. Table 5.1 shows the separation between various pairwise combinations of loci in this study on *Synechococcus* strain A and B´ genomes. All loci were within 24-61 kbp of the 16S rRNA locus, except *psaA*, which was approximately 220 kbp and 1366 kbp away from the 16S rRNA loci being studied in the *Synechococcus* strain A and B´ genomes, respectively. PCR analysis of the cyanobacterial 16S rRNA-ITS region was done with different volumes of MDA product (μl: 0.0001, 0.001, 0.01, 0.1, 0.25, 0.5, 1, and 2) using two individual cells each from *Synechococcus* strain A and B´ cultures and the MS60 sample to evaluate the sensitivity of the MDA amplification products for PCR amplification and to optimize PCR reactions for protein-encoding loci.

Sequence Verification of
PCR-amplified MDAs for Protein-Encoding Loci

Products obtained from PCR amplification of MDA reactions were cleaned and sent to the University of Nevada at Reno for sequencing (see Chapter 2). Sequence data were analyzed using NCBI-BLAST against the nr database (Altschul, 1990).

Table 5.1. The separation among loci in *Synechococcus* strain A (Top) and *Synechococcus* strain B´ genomes (Bottom) and the number of positive MDAs for both loci from the first batch of cells isolated (n = 23).

| ***Synechococcus* strain A** | | |
|---|---|---|
| **Loci** | **Separation (kbp)** | **Number of Positive MDAs of 11 assayed** |
| *apcAB/rbsK* | 32 | 4 |
| 16S rRNA/*aroA* | 38 | 2 |
| *rbsK/aroA* | 39 | 2 |
| 16S rRNA/*rbsK* | 50 | 4 |
| *apcAB/aroA* | 71 | 5 |
| 16S rRNA/*apcAB* | 82 | 5 |
| 16S rRNA/*psaA* | 220 | 4 |
| *aroA/psaA* | 231 | 4 |
| *rbsK/psaA* | 270 | 2 |
| *apcAB/psaA* | 302 | 7 |
| ***Synechococcus* strain B´** | | |
| **Loci** | **Separation (kbp)** | **Number of Positive MDAs of 12 assayed** |
| 16S rRNA/*rbsK* | 24 | 4 |
| *rbsK/aroA* | 28 | 4 |
| 16S rRNA/*aroA* | 52 | 4 |
| 16S rRNA/*apcAB* | 61 | 5 |
| *apcAB/rbsK* | 85 | 6 |
| *apcAB/aroA* | 113 | 4 |
| *aroA/psaA* | 1314 | 4 |
| *rbsK/psaA* | 1342 | 6 |
| 16S rRNA/*psaA* | 1366 | 5 |
| *apcAB/psaA* | 1427 | 6 |

## Results

Prior to amplification of the two batches of cells, two MDAs from *Synechococcus* strain A and B´ cultures and from the M60 sample were amplified to determine the optimal amount of MDA to be used in subsequent PCR amplifications. Amplification of the 16S rRNA-ITS region from samples containing different amounts of MDA reaction products varied among individual cells from different cultures and mat samples (Figure

5.3). Only one of the two *Synechococcus* strain A and B′ cells MDAs yielded PCR products (Figure 5.3A). Amplification of the *Synechococcus* strain B′ cell MDAs was

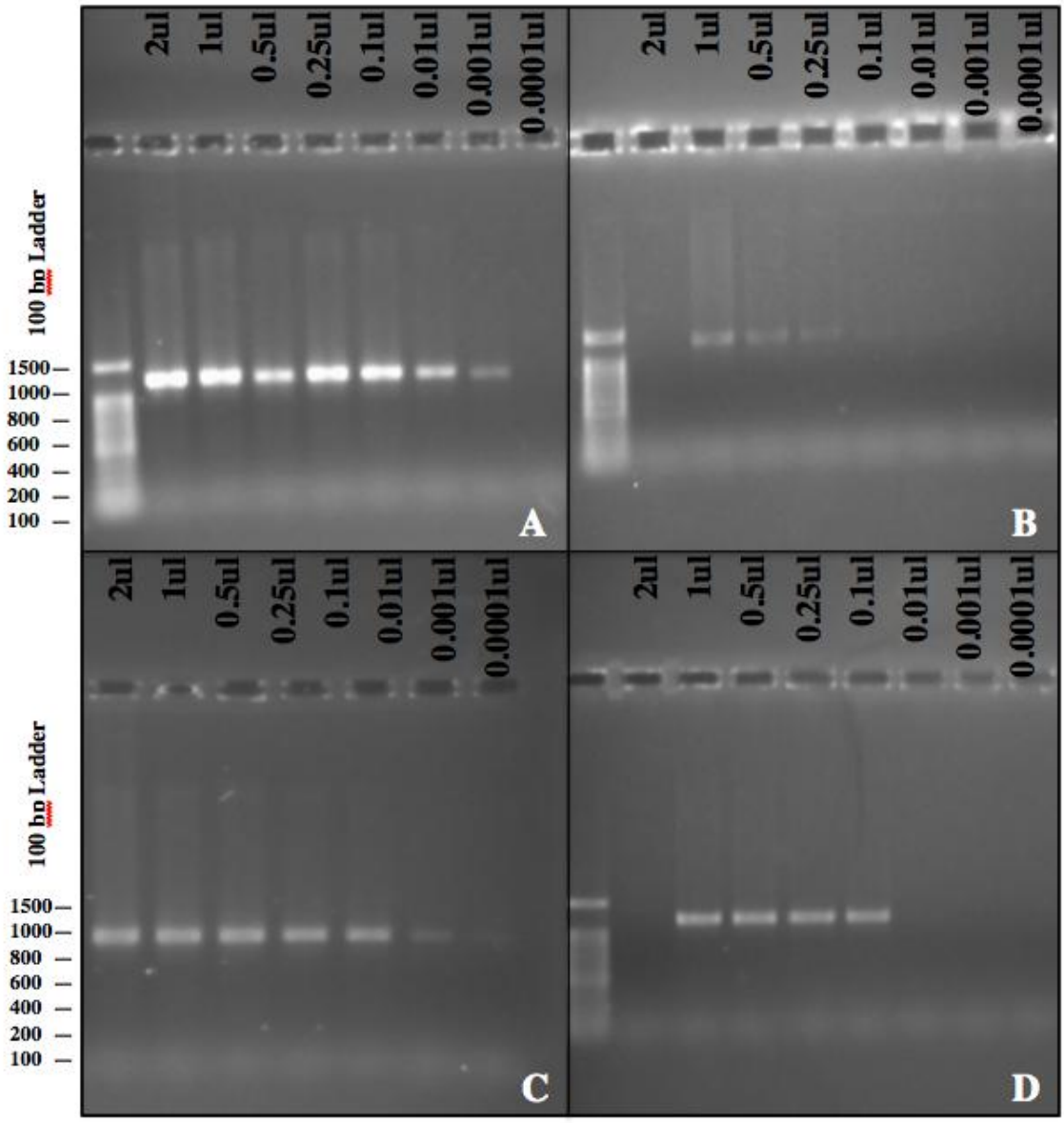



Figure 5.3. 16S rRNA-ITS PCR amplification of MDA reactions of single cells picked from (A) a *Synechococcus* strain A culture, (B) a *Synechococcus* strain B′ culture, (C and D) different cells from Mushroom Spring 60$^{o}$C mat sample. Different concentrations of MDA template were used to optimize the PCR reaction.

weak and it was difficult to distinguish differences in amplification across the gradient of MDA amount (Figure 5.3B). The MDAs from the two MS60 cells amplified well and demonstrated that different cells from the same sample may vary with respect to template yield. This experiment was repeated with similar results. Based on these results, 1-2 μl of MDA template was used for subsequent PCR screening of protein-encoding loci, as these amounts gave consistent and strong results.

All cells were amplified using all sets of primers to determine coverage of loci across the MDAs of cultures and environmental samples. Tables 5.2 and 5.3 summarize PCR and sequencing results for MDAs of A-like and B´-like *Synechococcus* cells obtained from the MS60 sample. All cells that were prescreened by qPCR for eubacterial 16S rRNA genes also contained a cyanobacterial 16S rRNA gene, based on amplification and sequencing (Tables 5.2 and 5.3). However, only 42% of the cells that were not prescreened contained a cyanobacterial 16S rRNA gene (Table 5.4). PCR amplification of other loci varied among the MDA reactions and provided amplicons that sometimes were and sometimes were not of sufficient concentration and quality for sequencing (Tables 5.2 and 5.3). Despite limited sequence data, the MDAs appeared to sample across BAC clone diversity, suggesting this method is comparable to BAC library efforts to sample diversity (Figures E5.1 and E5.2).

Since I wished to conduct an MLSA study that was anchored by the 16S rRNA-ITS region, the MDAs of value to the study are those that contain both the 16S rRNA locus and a series of *specific* additional loci. 56 MDAs contained a 16S rRNA gene as visualized on an agarose gel following PCR, and 42 of these contained a product of

Table 5.2. Summary results for PCR amplification of single-cell MDAs for protein-encoding loci for A-like *Synechococcus*. MDA loci amplifications that contained enough amplified product to confirm the presence of the locus via sequencing are denoted in red (or by a red star).

| **A-like *Synechococcus*** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Cell** | **16S rRNA**[b,c] | ***apcAB*** | ***rbsK*** | ***aroA*** | ***psaA*** | ***pcrA*** | **No. of genes** |
| MDA019 | 1* | 1 | 1 | 1 | 1 | | 5 |
| MDA012 | | 1 | 1 | | 1 | | 3 |
| MDA021 | 1* | 1 | 1 | | | | 3 |
| MDA053 | 1* | 1 | 1 | | | | 3 |
| MDA098p[a] | 1* | 1 | 1 | | | | 3 |
| MDA031 | 1* | | 1 | | 1 | | 3 |
| MDA070 | | 1 | | | 1 | | 2 |
| MDA071 | 1 | | | | 1 | | 2 |
| MDA080 | | 1 | 1 | | | | 2 |
| MDA097p[a] | 1* | | | 1 | | | 2 |
| MDA100p[a] | 1* | 1 | | | | | 2 |
| MDA066 | 1* | | | 1 | | | 2 |
| MDA009 | | | | | 1 | | 1 |
| MDA068 | 1* | | | | | | 1 |
| MDA080 | 1* | | | | | | 1 |
| MDA066 | 1* | | | | | | 1 |
| MDA091 | 1* | | | | | | 1 |

[a] MDAs from the first batch of cell isolations where pre-screening using qPCR was done to verify the presence of a eubacterial 16S rRNA gene.
[b] MDAs positive for a eubacterial 16S rRNA gene are denoted in black.
[c] MDAs positive for a cyanobacterial 16S rRNA gene are denoted in green.

Table 5.3. Summary results for PCR amplification of single-cell MDAs for protein-encoding loci for B´-like *Synechococcus*. MDA loci amplifications that contained enough amplified product to confirm the presence of the locus via sequencing are denoted in red (or by a red star).

| **B´-like *Synechococcus*** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Cell** | **16S rRNA**[b,c] | ***apcAB*** | ***rbsK*** | ***aroA*** | ***psaA*** | ***pcrA*** | **#genes** |
| MDA017 | 1* | 1 | 1 | 1 | 1 | 1 | 6 |
| MDA034 | 1* | 1 | 1 | 1 | 1 | 1 | 6 |
| MDA036 | 1* | 1 | 1 | 1 | 1 | 1 | 6 |
| MDA008 | 1* | 1 | | 1 | 1 | 1 | 5 |
| MDA033 | 1* | | 1 | 1 | 1 | 1 | 5 |
| MDA014 | 1* | 1 | | | 1 | 1 | 4 |
| MDA094 | 1* | 1 | 1 | | 1 | | 4 |
| MDA049 | 1* | 1 | 1 | | 1 | | 4 |
| MDA056 | 1* | | | 1 | 1 | 1 | 4 |
| MDA078 | | 1 | 1 | 1 | | 1 | 4 |
| MDA005 | 1* | | | 1 | | 1 | 3 |
| MDA037 | 1* | | 1 | | | 1 | 3 |
| MDA038 | | | 1 | | 1 | 1 | 3 |
| MDA054 | | 1 | | 1 | | 1 | 3 |
| MDA085 | 1* | | 1 | | | 1 | 3 |
| MDA020 | 1* | 1 | | | | 1 | 3 |
| MDA051 | 1* | 1 | | | | 1 | 3 |
| MDA040 | 1* | | 1 | | 1 | | 3 |
| MDA061 | 1* | 1 | 1 | | | | 3 |
| MDA026 | 1* | | 1 | | | | 2 |
| MDA039 | 1* | 1 | | | | | 2 |
| MDA074 | | 1 | 1 | | | | 2 |
| MDA092 | 1* | | 1 | | | | 2 |
| MDA042 | 1* | | 1 | | | | 2 |
| MDA001 | | 1 | | | 1 | | 2 |
| MDA003 | | 1 | | 1 | | | 2 |
| MDA011 | | | | | 1 | 1 | 2 |
| MDA048 | 1* | | | | 1 | | 2 |
| MDA086 | | | 1 | | 1 | | 2 |
| MDA089 | 1* | | | | | 1 | 2 |
| MDA090 | | | | 1 | 1 | | 2 |
| MDA101p[a] | 1* | 1 | | | | | 2 |

Table 5.3 Continued…

| Cell | 16S rRNA | *apcAB* | *rbsK* | *aroA* | *psaA* | *pcrA* | #genes |
|---|---|---|---|---|---|---|---|
| MDA045 | 1* | | | | | 1 | 2 |
| MDA006 | 1 | | | | 1 | | 2 |
| MDA007 | 1* | | 1 | | | | 2 |
| MDA015 | | | | | 1 | 1 | 1 |
| MDA018 | | | | | 1 | | 1 |
| MDA046 | | | | | | 1 | 1 |
| MDA063 | | | | | | 1 | 1 |
| MDA084 | 1* | | | | | | 1 |
| MDA013 | 1* | | | | | | 1 |
| MDA028 | | | 1 | | | | 1 |
| MDA083 | 1* | | | | | | 1 |
| MDA091 | 1* | | | | | | 1 |
| MDA016 | | | | 1 | | | 1 |
| MDA022 | | | | | 1 | | 1 |
| MDA047 | | | | | 1 | | 1 |
| MDA081 | | | | | 1 | | 1 |
| MDA095 | 1* | | | | | | 1 |

[a] MDAs from the first batch of cell isolations where pre-screening using qPCR was done to verify the presence of a eubacterial 16S rRNA gene.
[b] MDAs positive for a eubacterial 16S rRNA gene are denoted in black.
[c] MDAs positive for a cyanobacterial 16S rRNA gene are denoted in green.

Table 5.4. The number and percent of MDAs that contained positive products for selected genes in order in A-like and B´-like and all *Synechococcus* MDAs that were or were not pre-screened for 16S rRNA by qPCR.

| **A-like *Synechococcus* (n=17)** | | | |
|---|---|---|---|
| **Gene** | **Number of positive MDAs** | **% of 17 Cells (unscreened)** | **% of 12 cells (prescreened)** |
| **16S rRNA** | 11 | 73 | 100 |
| **+apcAB** | 5 | 33 | 45 |
| **+rbsK** | 4 | 27 | 36 |
| **+aroA** | 1 | 7 | 9 |
| **+pcrA** | 1 | 7 | 9 |
| **+psaA** | 0 | 0 | 0 |
| **B´-like *Synechococcus* (n=49)** | | | |
| **Gene** | **Number of positive MDAs** | **% of 49 Cells (unscreened)** | **% 30 cells (prescreened)** |
| **16S rRNA** | 31 | 67 | 100 |
| **+apcAB** | 11 | 24 | 35 |
| **+rbsK** | 5 | 11 | 16 |
| **+aroA** | 3 | 7 | 10 |
| **+pcrA** | 3 | 7 | 10 |
| **+psaA** | 3 | 7 | 10 |
| **All *Synechococcus* (n=101)** | | | |
| **Gene** | **Number of positive MDAs** | **% of 101 Cells (unscreened)** | **% of 42 cells (prescreened)** |
| **16S rRNA** | 42 | 42 | 100 |
| **+apcAB** | 16 | 16 | 38 |
| **+rbsK** | 9 | 9 | 21 |
| **+aroA** | 4 | 4 | 10 |
| **+pcrA** | 4 | 4 | 10 |
| **+psaA** | 3 | 3 | 7 |

sufficient concentration to obtain sequence data. The number of MDAs containing a *Synechococcus* A-like or B´-like 16S rRNA gene plus increasing numbers of protein-encoding genes in the order *apcAB, rbsK*, *aroA, pcrA,* and *psaA* (the combination giving the greatest number of MDAs useful for MLSA) decreased with the number of loci (Table 5.4). In Chapter 4, I discussed the use of 7 and 4 loci for *Synechococcus* A-like and B´-like populations for a MLSA study. Only 2 of the 7 genes used in the MLSA study for A-like *Synechococcus* (*rbsK* and *aroA*) were assayed in this study and only 1 of the 17 MDAs determined to be A-like contained both of these genes. Four of the 49

*Synechococcus* B´-like population MDAs contained all 4 loci used in the MLSA study (Chapter 4); without prescreening, only approximately 4% of the 101 MDAs could be used in a 4-locus MLSA study of B´-like *Synechococcus*. Prescreening for MDAs containing a eubacterial 16S rRNA gene increased recovery somewhat (Table 5.4). The number of MDAs needed to obtain a sample that would be useful for MLSA studies of the same scope as reported in Melendrez et al., (2010) (Chapter 4) is reported as a function of the number of loci in Table 5.5.

Table 5.5. Number of MDAs needed to provide 71 MDAs useful for MLSA analysis as a function of number of loci and pre-screening for all *Synechococcus*.

| **Gene** | **All *Synechococcus* (n=71)[a]** | |
|---|---|---|
| | Unscreened | Screened |
| 16S rRNA | 167 | 71 |
| +apcAB | 439 | 185 |
| +rbsK | 781 | 335 |
| +aroA | 1757 | 703 |
| +pcrA | 1757 | 703 |
| +psaA | 2343 | 1005 |

[a] For comparison with 71 BACs per organism used in the MLSA study (Chapter 4).

## Discussion

As discussed in the introduction of this chapter, the MDA reaction randomly amplifies regions with variable effectiveness around the genome so that successful MDA amplification does not necessarily mean successful amplification of the specific loci of interest. There may be additional causes that contribute to the declining numbers of MDAs useful in an MLSA study. For instance, if primer recognition sites for a targeted gene are not copied during the MDA reaction the locus of interest will not be amplifiable with PCR. I may have observed an example of problems due to recognition site

differences in MDAs that tested positive for a eubacterial 16S rRNA gene but negative for a cyanobacterial 16S rRNA gene (black in Tables 5.1 and 5.2), yet contained *psaA*, which is a gene associated with photosynthesis. This suggests that the cells isolated for MDA were cyanobacteria whose MDAs contained the portion of the eubacterial, but not the cyanobacterial, 16S rRNA locus targeted by PCR primers.

Table 5.5 shows that 439-2343 unscreened or 185-1005 screened cells would have to be picked and amplified using MDA to provide 71 MDAs useful for an MLSA analysis of 16S rRNA plus 1-5 loci. I discussed the need to screen ~360,000 BACs from the environment in order to obtain a subset useful for an MLSA study. Picking and analyzing up to 2,343 cells would seem preferable to screening hundreds of thousands of BACs. However this doesn't take into account the large amount of time and skill needed to pick single cells either by micromanipulation or flow cytometry, to confirm that a single cell was indeed captured and to ensure successful lysis of the cells prior to the MDA reaction. Recent improvements in flow cytometry, such as high-speed droplet-based fluorescence-activated cell sorting (Stephanauskas and Sieracki, 2007 and Laskin et al., 2005) may make high throughput single-cell sorting more feasible if it can be guaranteed that a single cell enters each MDA reaction. While the BAC approach has its own limitations as stated in Chapter 3, there are high-throughput protocols to improve screening efficiency using digoxigenin-labeled overgo probes (non-radioactive) (Hilario et al., 2007), TRFLP (Babcock et al., 2007), and/or radioactive probing (Melendrez et al., 2010a; Chapter 3; Appendix B) that will quickly and efficiently single-out the BACs of interest for a particular study. In addition, creation of a BAC clone library provides a

valuable metagenomic resource with information on all community members in the environmental sample, not just *Synechococcus* populations.

Ensuring that loci are not linked is important in MLSA studies and increases the attractiveness of MDA as a tool in MLSA. All of the genes in this study were separated from any other gene in the study by at least 11 kbp, but amplification of one gene did not ensure that another gene ≥11 kbp away would be amplified as well in the same MDA reaction. It appears that coverage around the genome was random, as expected, given that *psaA* occurs 220 kbp from the 16S rRNA region in *Synechococcus* strain A genome and 1366 kbp from the 16S rRNA region in *Synechococcus* strain B´ genome, yet amplified as well as loci within the 16S rRNA region (Table 5.1) There was no observable trend in the number of positive MDAs for specific locus sets and increased or decreased distances between the two loci (Figure 5.4).

MDA provides the potential advantage of linking any gene in the genome to a specific individual in a population genetics study and thus offers the potential to link phylogeny and function. As an example, observation of metagenomic clones that do not match the *Synechococcus* strain B´ genome on both paired ends (‘illegal clones’) has suggested that some native *Synechococcus* populations may, unlike the *Synechococcus* strain B´ isolate contain genes that may be ecotype-specific such as the *feo* gene, which encodes a protein involved in ferrous iron transport (Bhaya et al., 2007). MDAs might provide a means by which to test whether this property for ferrous iron uptake (for example) is a trait shared by all members of a specific ecotype. Unfortunately, the

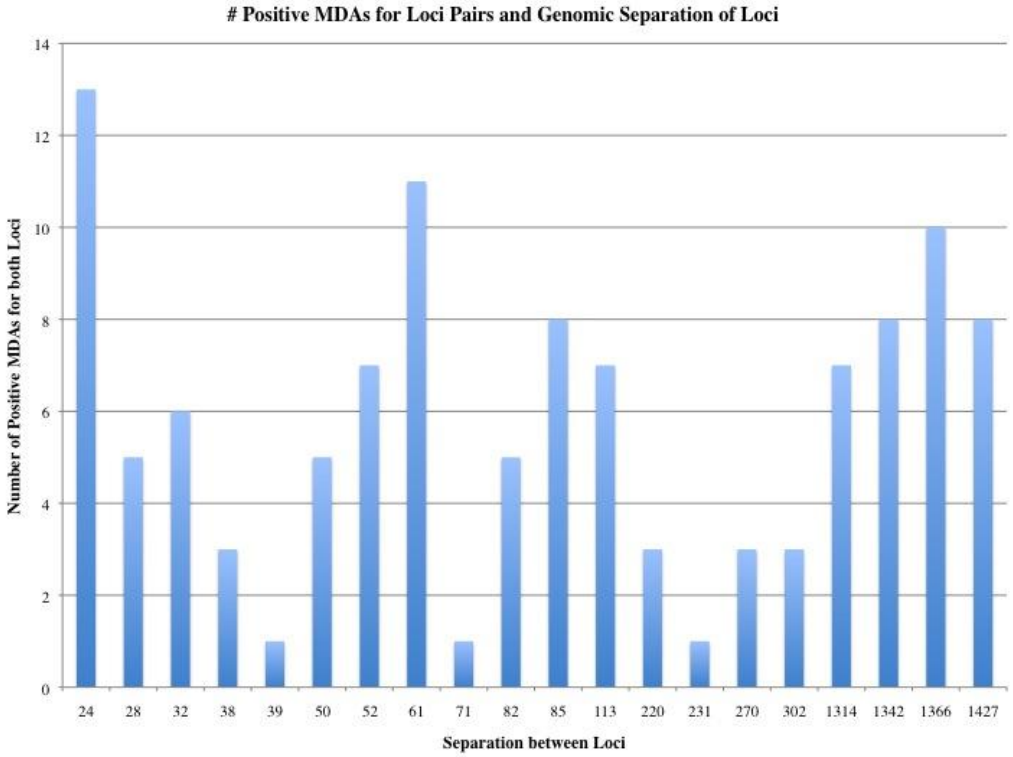



Figure 5.4. The number of positive MDAs for pairs of loci as a function of separation of the loci in the genome.

probability of detecting both ecotype-specifying MLSA loci and such functional genes is reduced by the random amplification currently provided by MDA.

## References

Allewalt JA, Bateson MM, Revsbech NP, Slack K and Ward DM. (2006). Temperature and light adaptations of *Synechococcus* isolates from the microbial mat community in Octopus Spring, Yellowstone National Park. *Appl Environ Microbiol* **72**: 544-550.

Altschul SF, Gish W., Miller W, Myers EW and Lipman, DJ. (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403-410.

Aviel-Ronen S, Zhu Q, Chang C, Bradley P, Liu N, Watson SK, Lam WL and Tsao MS. (2006). Large fragment Bst DNA polymerase for whole genome amplification of DNA from formalin-fixed paraffin-embedded tissues. *BMC Genomics* **7**: 1-10.

Babcock DA, Wawrik B, Paul JH, McGuiness L and Kerkof LJ. (2007). Rapid screening of large insert BAC library for specific 16S rRNA genes using TRFLP. *J Microbiol Method* **71**: 156-161.

Becraft ED, Cohan FM, Kuhl M, Jensen S and Ward DM. (2010). Identifying and improving the existence of ecologically defined *Synechococcus* sp. in Mushroom Spring, Yellowstone National Park. In prep.

Bhaya D, Grossman AR, Steunou AS, Khuri N, Cohan FM, Hamamura N, Melendrez MC, Bateson MM, Ward DM, and Heidelberg JF. (2007). Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analysis. *ISME J* **1**: 703-713.

Binga EK, Lasken RS and Neufeld JD. (2008). Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. *ISME J* **2**: 233-241.

Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, Sun Z, Zong Q, Du Y, Du J, Driscoll M, Song S, Kingsmore SF, Egholm M and Lasken RS. (2002). Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci* **99**: 5261-5266.

Fernández-Ortuño D, Torés JA, de Vicente A and Pérez-García A. (2007). Multiple displacement amplification, a powerful tool for molecular genetic analysis of powdery mildew fungi. *Curr Genet* **51**: 209-219.

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

Havryliuk T, Orjuela-Sanchez P and Ferreira MU. (2008). *Plasmodium vivax:* Microsatellite analysis of multiple-clone infections. *Exp Parasitol* **120**: 330-336.

Hilario E, Bennell TF and Rikkerink E. (2007). Screening a BAC library with nonradioactive overlapping Oligonucleotide (overgo) probes. *Methods Mol Biol* **353**: 79-91.

Hosono S, Faruqi AF, Dean FB, Du Y, Sun Z, Wu X, Du J, Kingsmore SF, Egholm M and Lasken RS. (2003). Unbiased whole-genome amplification directly from clinical samples. *Genome Res* **13**: 954-964.

Kvist T, Ahring BK, Lasken RS and Westermann P. (2007). Specific single-cell isolation and genomic amplification of uncultured microorganisms. *Appl Microbiol Biotechnol* **74**: 926-935.

Lasken RS. (2009). Genomic DNA amplification by the multiple displacement amplification (MDA) method. *Biochem Soc Trans* **37**: 450-453.

Lasken RS, Raghunathan A, Kvist T, Ishoey T, Westermann P, Ahring BK and Boissy R. (2005). Multiple displacement amplification from single bacterial cells. In: *Whole Genome Amplification*. Eds. Hughes S and Lasken R. Scion Publishing: Bloxham. Pp. 119-147.

Marcy Y, Ishoey T, Lasken RS, Stockwell TB, Walenz BP, Halpern AL, Beeson KY, Goldberg SMD and Quake SR. (2009). Nanoliter reactors improve multiple displacement amplification of genomes from single cells. *PLOS Genet* **3**: 1702-1708.

Melendrez MC, Lange RK, Cohan FM and Ward DM. (2010a). Ecological diversity of *Synechococcus* spp. inhabiting an alkaline siliceous hot spring in Yellowstone National Park, WY measured using protein-encoding genes and evolutionary simulation. *ISME J* In prep.

Melendrez MC, Wood JM, Rusch DB, Heidelberg JF and Ward DM. (2010b). Bacterial artificial chromosome (BAC) libraries for Mushroom Spring cyanobacterial mat, Yellowstone National Park, WY. In prep.

Neufeld JD, Chen Y, Dumont MG and Murrell JC. (2008). Marine methylotrophs revealed by stable-isotope probing, multiple displacement amplification and metagenomics. *Environ Microbiol* **10**: 1526-1535.

Raghunathan A, Ferguson HR, Bornath CJ, Song W, Driscoll M and Lasken RS. (2005). Genomic DNA from a single bacterium. *Appl Environ Microbiol* **71**: 3342-3347.

Sorensen KM, Jespersgaard C, Vuust J, Hougaard D, Nørgaard-Pedersen B and Andersen PS. (2007). Whole genome amplification on DNA from filter paper blood spot samples: An evaluation of selected systems. *Genet Testing* **11**: 65-71.

Stepanauskas R and Sieracki ME. (2007). Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proc Natl Acad Sci* **104**: 9052-9057.

Sun YQ, Monstein HJ, Ryberg A and Borch K. (2007). Multiple displacement amplification of DNA isolated from human archival plasma/serum: identification of cytokine polymorphism by pyrosequencing analysis. *Clinica Chimica Acta* **377**: 108-113.

Wu L, Liu X, Schadt CW and Zhou J. (2006). Microarray-based analysis of subnanogram quantities of microbial community DNAs by using whole-community genome amplification. *Appl Environ Microbiol* **72**: 4931-4941.

# CHAPTER 6

# CONCLUSIONS

In this dissertation, together with my collaborators, I addressed the general questions, “do hot spring *Synechococcus* species exist?” and if so, “how might they be detected?” and “does their existence matter in the microbial communities they inhabit?” The challenge in defining a “species” in microbiology stems from the vast genetic diversity, genomic plasticity and promiscuity in bacteria and archaea. At one extreme, concepts have been introduced that imply that some bacterial and archaeal populations experience rampant recombination, such that they do not exist as species (Papke et al., 2007; Doolittle and Papke, 2006; Gogarten et al., 2002; Lawrence, 2002). At the other extreme, concepts have been developed that emphasize the highly organized distribution of bacterial and archaeal diversity both geographically (Papke et al., 2003; Whitaker et al., 2003; Miller et al., 2007) and ecologically (Ward et al., 2006, Ramsing et al., 2000, Ferris et al., 2003; Ferris and Palenik, 1998; Ferris and Ward, 1997; Ferris et al., 1996; Allewalt et al., 2006; Rocap et al., 2002). Without an understanding of the population-level units that make up microbial diversity, it becomes difficult to investigate composition and structure of a community and how that relates to community function, making the issue of microbial species an important one. If every individual within a population is not unique, but is functionally interchangeable, then individuals can be grouped into species and analyzing community function becomes a simpler process. Understanding the ‘individual’ and how it fits into the ‘whole’ affords the opportunity of learning about how evolution shapes microbial species in light of ecological change. This

understanding is of paramount importance in a world that changes rapidly, unleashing new viruses and bacterial lineages in response to human activity or environmental disturbance.

This dissertation has focused on a well-studied hot spring microbial mat community constructed by cyanobacteria. Previous research has suggested that there are more *Synechococcus* putative ecotype (PE) populations than can be detected by the 16S rRNA gene or internal transcribed spacer (ITS) region alone and that these populations may be ecologically distinct in nature (ecotypes). The first hypothesis addressed was that, **there are more ecotypes within the Mushroom Spring microbial mat than have been previously identified using cultivation-independent 16S rRNA-ITS PCR, cloning and sequencing methods.** Ecotype simulation analysis of single- and multi-locus sequence datasets confirmed that more putative ecotypes could be discerned by using protein-encoding loci that increased molecular resolution. Analysis of single protein-encoding loci demarcated 3-14 PEs and 8-13 PEs for *Synechococcus* A-like and B´-like populations respectively, more than the 2-3 PEs predicted from analysis of 16S rRNA and ITS data (Chapter 2). However, the number of putative ecotypes depended on the locus or loci used and the molecular resolution offered. ES analysis of multi-locus data yielded 10 A-like and 13 B´-like PEs, whereas eBURST analysis of the same data yielded fewer clonal complexes (4 A-like and 2 B´-like)(Chapter 4).

Concatenation of multiple protein-encoding genes increased resolution in terms of EED (0.12 and 0.18 for MLSA of *Synechococcus* A- and B´-like BACs) (Figure 6.1), yet fewer B´-like ecotypes were predicted than for the *rbsK* locus in both BACs and PCR

clones (compare red g to e and f in Figure 6.1). The greater PE prediction may be due to recombination in the *rbsK* locus allowing for increased within-PE nucleotide diversity causing ES to predict more ecotypes in single-locus analysis (SLA). Concatenation of a locus (*rbsK*) undergoing recombination with loci that had more conserved genetic histories caused PE prediction to decrease, however this may be more in line with the true organismal phylogeny. That is, one locus did not skew the phylogeny, as would have resulted from using only the *rbsK* locus to assay evolutionary history of *Synechococcus* populations. For *Synechococcus* A-like BACs, ES analysis BAC *rbsK* sequences predicted fewer PEs than from *rbsK* in the SLA (Chapter 2) (compare blue f and e in Figure 6.1). This may be due to the presence of highly divergent A-like or possibly A´-like populations that we might have inadvertently sampled using less specific PCR primers and could not separate from A-like sequences (Figure A2.2). Since the *rbsK* locus was also shown to be undergoing recombination as well, this may also have the same affect of increasing PE prediction from SLA as compared to MLSA (compare blue e and g in Figure 6.1), as discussed for *Synechococcus* B´-like sequences. Overall, concatenation appeared to average the evolutionary histories of the loci so that the high or low molecular resolution and recombination of single loci did not skew the overall phylogeny.

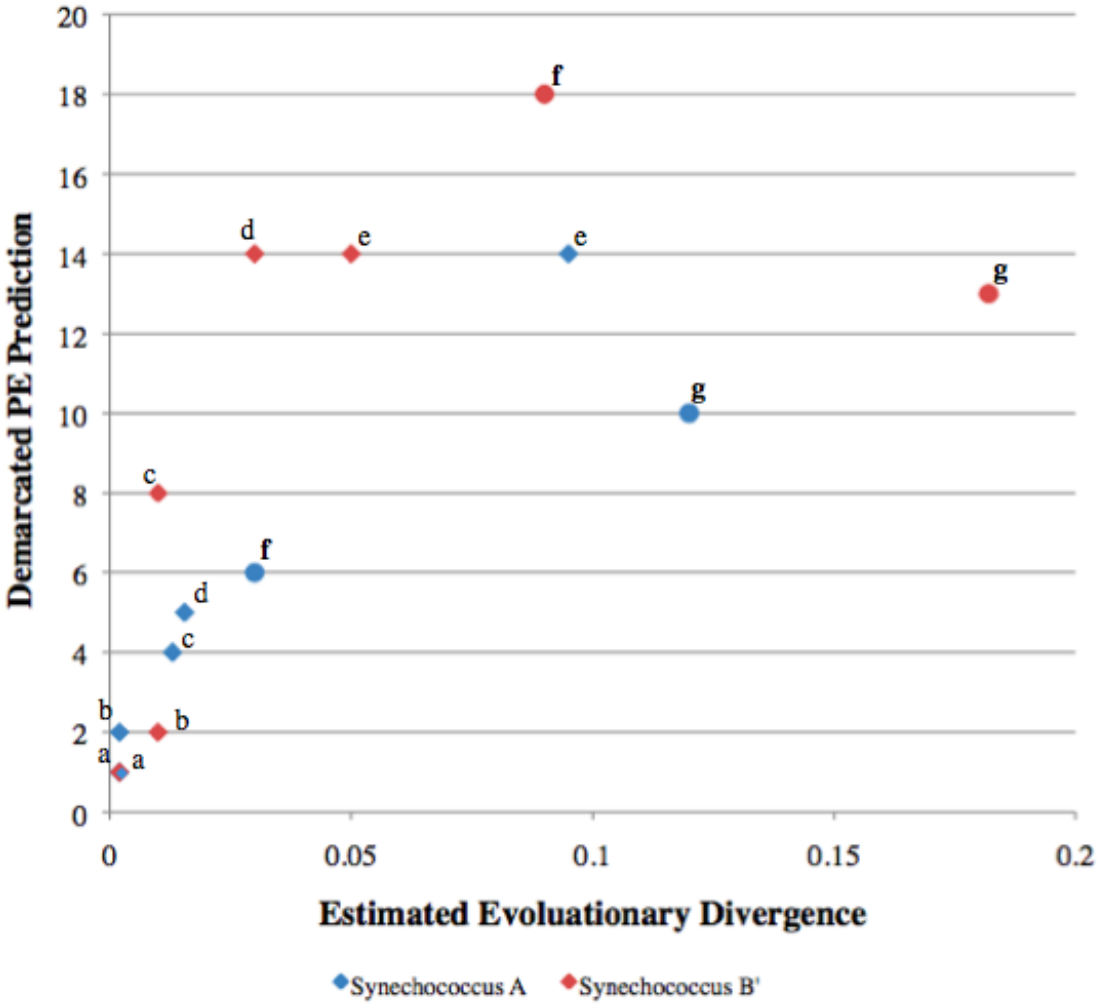



Figure 6.1. Demarcated putative ecotype prediction of Ecotype Simulation as a function of molecular resolution, as measured by estimated evolutionary divergence of ~70 sequences from SLA (a-e) and MLSA (f and g) for the (a) 16S rRNA, (b) ITS region, (c) *apcAB* for B´-like *Synechococcus* and *aroA* for A-like *Synechococcus*, (d) *aroA* for B´-like *Synechococcus* and *apcAB* for A-like *Synechococcus*, (e) *rbsK*, (f) BAC-associated *rbsK* loci and (g) 4- and 7-locus concatenations for A-like and B´-like *Synechococcus* populations.

The second hypothesis was that **these ecotypes exhibit distribution patterns that suggest they are ecologically distinct.** ES analysis of all single protein-encoding loci showed evidence of sample-specific ecotypes for the *Synechococcus* A-like clones with the *rbsK* locus revealing the most (7 of 14 ecotypes). MLSA ES analysis of 5 protein-encoding loci for *Synechococcus* A-like BACs also showed evidence of 8 PEs, 4 of which were sample-specific (PEs A 2, 6, 7 and 8; Figure 4.6). eBURST analysis revealed 2 clonal complexes dominated by sequence types from the low-temperature sample, however these also included a sequence type from the high-temperature sample and there was no statistical support for heterogeneity across ecotypes in habitat associations.

The impact of recombination on the evolution of *Synechococcus* populations was of particular interest in this study, as the strong evidence of ecological adaptation observed in previous studies led naturally to the third hypothesis that **recombination has been less important that mutation in shaping the evolution of native *Synechococcus* populations**, which was tested in Chapter 4. Phylogenetic incongruency, the presence of linkage disequilibrium and evidence from analysis of sequence data using programs available in RDP3 showed that over half of the *Synechococcus* A-like and B´-like clones analyzed were recombinants at one or more loci. Analysis of SNP maps revealed more putative recombinants that were not detected with RDP3 methods. Recombination signals were most commonly seen in the highly resolving loci (i.e. *rbsK, pcrA, PK*). BAC-end sequence data suggested that inversions or rearrangements (which can be mediated by homologous recombination or mobile elements) have occurred within the 16S rRNA

regions (Chapter 3). If the 16S rRNA regions are 'hot spots' of genomic rearrangement, possibly mediated through recombination, it is not surprising that recombination signals were detected in many of the genes in this region. Recombination and mutation rate estimates suggested *r/m* values of 2.87 and 5.15 for *Synechococcus* A- and B´-like BACs. Therefore the above hypothesis must be rejected, as it appears recombination has played a larger role than mutation in shaping the evolution of *Synechococcus* populations.

A major finding was, however, that despite recombination, sample-specific ecotype clades could still be discerned using MLSA. ES appeared to be capable of grouping phylogenetically related variants into sample-specific (i.e., ecologically distinct) clades, but variants that have undergone recombination with phylogenetically more distant individuals at the locus being studied may be excluded. eBURST appears to detect fewer, less well defined sample-specific populations (clonal complexes), excluding close, but capturing more distant phylogenetic relatives. Population genetics analyses such as these, attempt to model the existence of natural evolved organismal populations that are ecologically distinct. Variants resulting from recombination with distant populations, such that they are not detected within PEs demarcated by ES (but may be detected by eBURST), may nevertheless be a member of the organismal PE population. In other words, recombination makes it more difficult to detect the variants within PE populations using individual or multiple loci. It does not to erode the ecological species population itself.

Figure 6.2 illustrates hypothetical recombinants, which fall into different positions in hypothetical phylogenies, depending on the degrees to which they have recombined with closer or more distant relatives. Some recombination events will not have enough impact on the gene sequence to displace the variant from the PE (L in Figure 6.2). For instance, Figure 4.10C shows several PEVs that contained 9-48 SNPs (perhaps indicative of small recombination events) that were still demarcated as members of PEA1. Other recombination events have a greater impact on the gene sequence, however, the buffering resulting from concatenation of multiple loci may "recapture" the variant (M in Figure 6.2). For instance, the *pcrA* locus in *Synechococcus* B´-like BACs has shown evidence of recombination (Table 4.13), however not to the extent of *rbsK*. Figure 4.3 shows that *pcrA* PEs B´5, 6 and 9 contain variants that collapsed into MLSA PE B´6. However, concatenation may be unable to buffer against recombination events with organisms so distantly related that there is a large impact on the gene sequence (H in Figure 6.2). Two examples of this are clones M60B015H24 and M60B773C01 in *Synechococcus* A-like and B´-like BACs (Figures 4.2 and 4.3). Both are recombinants that are distantly related to the other sequences in the MLSA phylogeny. They are also singleton PEs in MLSA-ES. These clone variants may indeed fall within PEs in an organismal phylogeny, however, given the extent of recombination, we are unable to ascertain where they may truly belong from either SLA or MLSA. It should be possible to determine which variants belong to which PEs by studying the co-distribution and gene expression of such variants, which should be identical to the patterns exhibited by the dominant variant (and other variants) of the PE to which they belong.

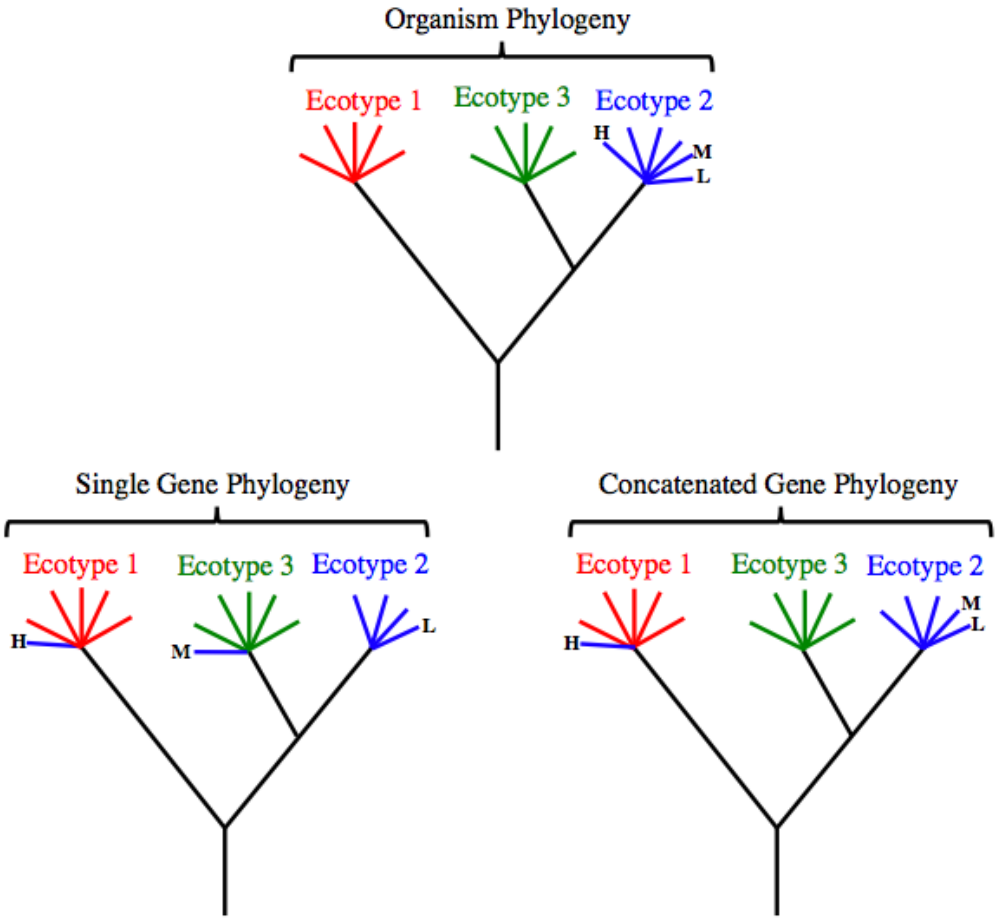



Figure 6.2. Illustration of the hypothesized effect of ecologically neutral recombination events with high, medium and low (H, M, L) impact on the positioning of variants in single gene and multi-locus concatenated phylogenies. All variants belong to the same organismal ecotype (top), but accurate association of the recombinant variant with the correct ecotype clade in single and concatenated MLSA gene phylogenies depends on the degree of impact recombination has.

Perhaps the future of investigating the population genetics of closely related organisms lies in single-cell selection technology and multiple displacement amplification (MDA). The option of MDA was not feasible for this study as discussed in Chapter 5, however improvement of these technologies (e.g., automated cell selection and even amplification across the genome) could provide a quite powerful tool for MLSA

in the future. This would provide more options for selection of loci around the genome and solve problems associated with loci that are linked and may have evolutionary histories that are not independent of each other, possibly skewing multi-locus phylogenetic analysis. If recombination complicates the analysis, loci can be selected that do not exhibit frequent recombination events. However I would argue against the removal of recombinants from any population genetics study as it limits the analysis of true natural diversity created through recombination in the environment.

The environment routinely breaks our preconceived notions of how bacterial evolution should occur or how diversity is organized, underscoring the importance of letting nature guide analysis of bacterial community composition, structure and function. Analysis of single and multiple protein encoding loci using cultivation-independent techniques (BAC cloning) combined with theory-based ES analysis increased detection of PEs, many of which were sample -specific. Furthermore, MLSA revealed that recombination has been important historically (and presumably is presently) in the evolution of these *Synechococcus* populations. Importantly, recombination appears not to have completely disrupted our ability to detect ecotypes. Future studies using distribution analyses will continue to increase our understanding of the relationship between genetic and organismal diversity.

References

Allewalt JA, Bateson MM, Revsbech NP, Slack K and Ward DM. (2006). Temperature and light adaptations of *Synechococcus* isolates from the microbial mat community in Octopus Spring, Yellowstone National Park. *Appl Environ Microbiol* **72**: 544-550.

Cohan FM and Perry EB. (2007). A systematics for discovering the fundamental units of bacterial diversity. *Curr Biol* **17**: R373-386.

Doolitte WF and Papke RT. (2006). Genomics and the bacterial speciation problem. *Genome Biol* **7**:116.1-116.7.

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

Ferris MJ and Palenik B. (1998). Niche adaptation in ocean cyanobacteria. *Nature* **396**: 226-228.

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

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

Gogarten JP, Doolittle WF and Lawrence JP. (2002). Prokaryotic evolution in the light of gene transfer. *Mol Biol Evol* **19**: 2226-2238.

Lawrence JG. (2002). Gene transfer in bacteria: speciation without species? *Theor Pop Biol* **61**: 449-460.

Miller SR, Castenholz RW and Pedersen D. (2007). Phylogeography of thermophilic cyanobacterium *Mastigocladus laminosus*. *Appl Environ Microbiol* **73**: 4751-4759.

Papke RT, Zhaxybayeva O, Feil EJ, Sommerfeld K, Muise D and Doolittle WF. (2007). Searching for species in haloarchaea. *Proc Natl Acad Sci* **104**: 14092-14097.

Papke RT, Ramsing NB, Bateson MM and Ward DM. (2003). Geographical isolation in hot spring cyanobacteria. *Environ Microbiol* **5**: 650-659.

Ramsing NB, Ferris MJ and Ward DM. (2000). Highly ordered vertical structure of *Synechococcus* population within the one-millimeter-thick photic zone of a hot spring cyanobacterial mat. *Appl Environ Microbiol* **66**: 1038-1049.

Rocap G, Distel DL, Waterbury JB and Chisholm SW. (2002). Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. *Appl Environ Microbiol* **68**: 1180-1191.

Ward DM, Bateson MM, Ferris MJ, Kuhl M, Wieland A, Koeppel A and Cohan FM. (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. *Phil Trans Roy Soc B* **361**:1997-2008.

Whitaker RJ, Grogan DW and Taylor JW. (2003). Geographic barriers isolate endemic populations of hyperthermophilic Archaea. *Science* **301**: 976-978.

# APPENDICES

# APPENDIX A

## SUPPLEMENTAL INFORMATION FOR CHAPTER 2: ECOLOGICAL DIVERISTY OF *SYNECHOCOCCUS* POPULATIONS INHABITING AN ALKALINE SILICEOUS HOT SPRING MICROBIAL MAT IN YELLOWSTONE NATIONAL PARK, WYOMING MEASURED USING CULTIVATION-INDEPENDENT ANALYSIS OF PROTEIN-ENCODING GENES AND EVOLUTIONARY SIMULATION

## Estimation of Divergence and Variance of Metagenomic Sequences from Reference Genomes

Metagenomic libraries were generated for mats in Mushroom Spring and Octopus Spring, Yellowstone National Park, Wyoming and are described elsewhere (Klatt et al., 2010). Sequences from the metagenomic libraries were analyzed by reciprocal best BLAST analysis (Altschul, 1990) and the percent nucleotide identities of each sequence relative to homologs in the genomes of *Synechococcus* strain A and B´ were plotted (Figure A2.1). These distributions were used to define cutoffs for A-like and B´-like *Synechococcus* sequences, as suggested by the green and red boxes, respectively. Divergence and variance were calculated using percent nucleotide identities obtained from a reciprocal best BLAST analysis of the sequences recruited by the *Synechococcus* strain A and B´ genomes. The inset diagram illustrates what is meant by divergence and variance. Divergence was calculated by averaging divergences of all metagenomic homologs from the homolog in the most closely related genome. Variance in the divergence from the genomic homolog was taken as a measure of diversity among metagenomic homologs.

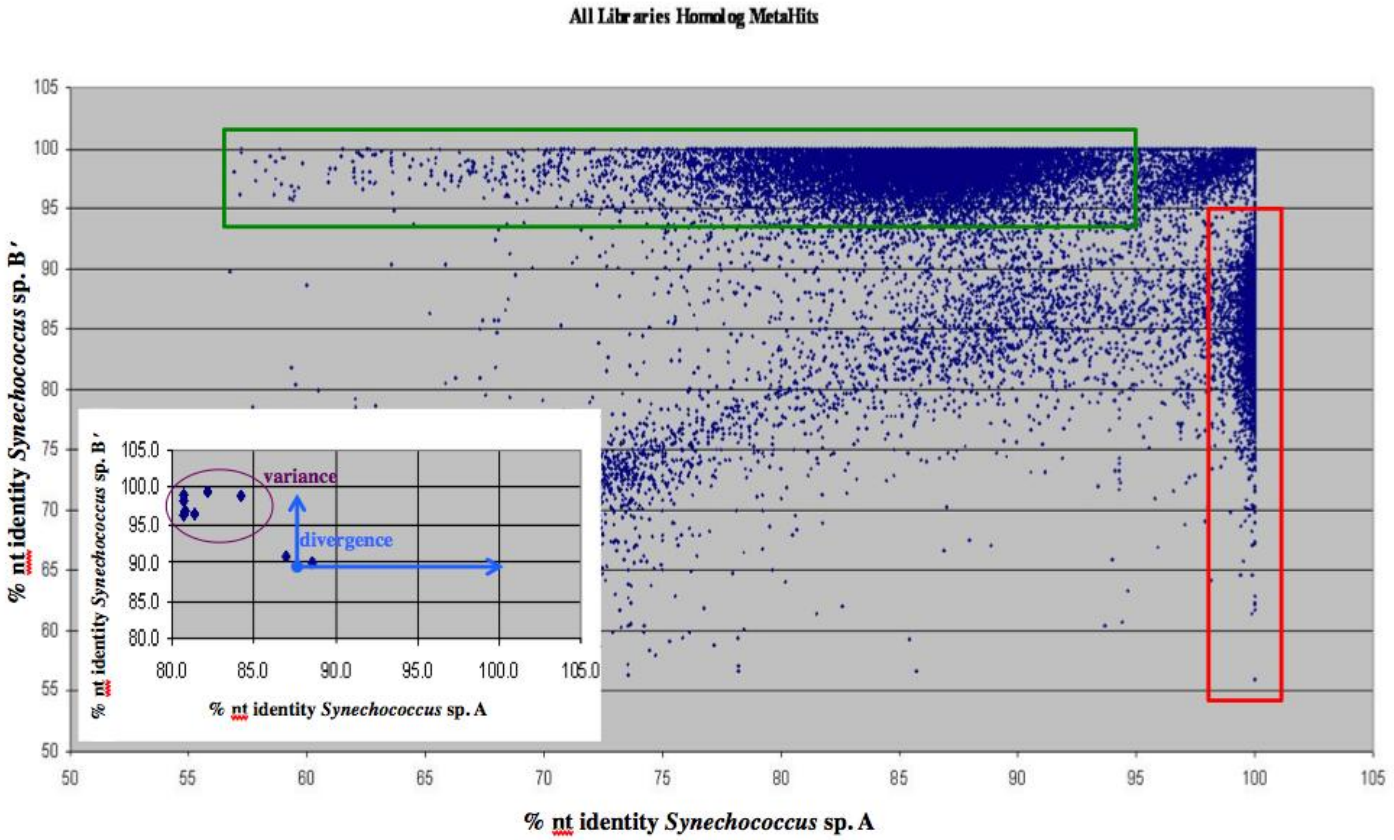



Figure A2.1. Percent nucleotide identities between metagenomic reciprocal best BLAST matches to *Synechococcus* strain A and B′ reference genomes. Matches boxed in green had high nucleotide identity to the *Synechococcus* strain B′ genome and matches boxed in red had high nucleotide identity to the *Synechococcus* strain A genome. Inset diagram shows an example of *rbsK* divergence and variance estimation using metagenomic homologs to *rbsK* with respect to the *Synechococcus* strain B′ reference genome.

Separation of *Synechococcus* A-like from A´-like Sequences

WU-BLAST analyses revealed that *Synechococcus* A-like homologs of *apcAB* and *aroA* genes retrieved by PCR from 60 and 65°C mat samples were well separated from homologs in the 68°C metagenomic sample containing predominantly A´-like populations (Figure A2.2). Nearly all *Synechococcus* A-like sequences PCR amplified from the M60 and M65 samples were typical of *Synechococcus* A-like populations, as opposed to A´-like populations, which are clearly more divergent. Only one *apcAB* and *aroA* sequence each, which were as or more distantly related to the *Synechococcus* strain A homolog as were homologs in the 68°C sample, were removed in further analyses (circled in red, Figure A2.2). In contrast, the percent nucleotide identity distributions of *rbsK* genes retrieved from the M60 and M65 samples were not discretely separated from those in the 68°C metagenome (Figure A2.2). Thus, it was not possible to use this approach to identify A´-like rbsK sequences.

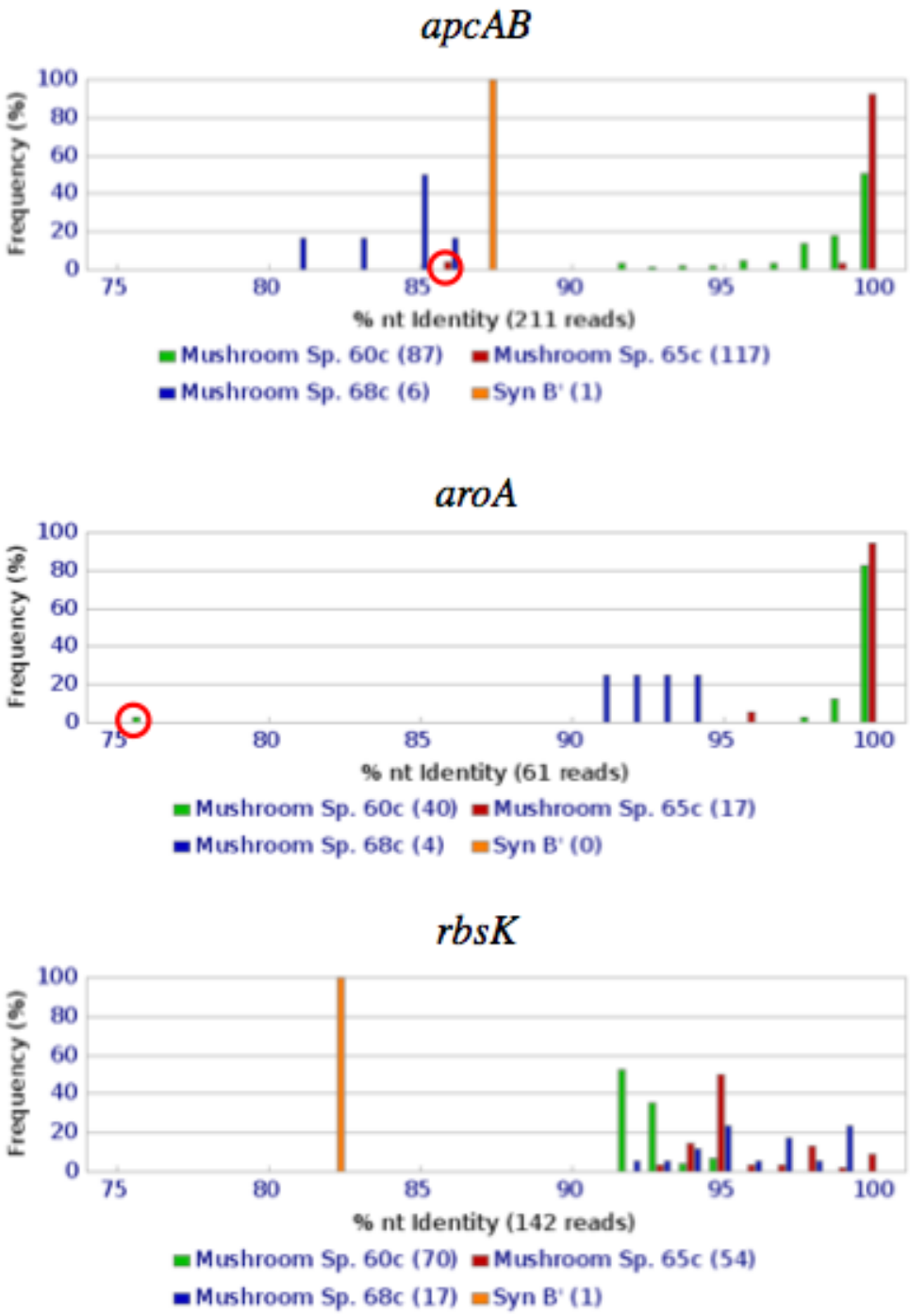



Figure A2.2. Frequency distributions of the percent nucleotide identity between homologous sequences from Mushroom Spring 60° and 65°C PCR clone libraries and the Mushroom Sp. 68°C metagenome and the *Synechococcus* strain A genomic homolog with *Synechococcus* strain B′ (SynB′) as a reference with low nucleotide identity to *Synechococcus* strain A. Sequences that were removed from subsequent analysis are circled in red.

## References

Altschul SF, Gish W., Miller W, Myers EW and Lipman, DJ. (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403-410.

Klatt CG, Wood JM, Rusch DB, Bateson MM, Heidelberg JF, Bryant DA and Ward DM. (2010). Composition of metagenomes from a phototrophic hot spring microbial mat community. In prep.

# APPENDIX B

## SUPPLEMENTAL INFORMATION FOR CHAPTER 3: BACTERIAL ARITIFICAL CHROMOSOME LIBRARIES FOR MUSHROOM SPRING CYANOBACTERIAL MATS, YELLOWSTONE NATIONAL PARK

## Isolation of High Molecular Weight DNA for BAC Clone Library Construction

To assist in obtaining HWM DNA from hot spring mat samples using gentle in-gel lysis samples were pre-treated with lysozyme to form spheroplasts of *Synechococcus* cells (Figure B3.1), which lysed more readily in gels than cells that were not pretreated (data not shown). To determine if HMW DNA had been obtained from the samples following pretreatment and in-gel lysis, gel plugs were analyzed by pulsed-field gel electrophoresis (Figure B3.2).

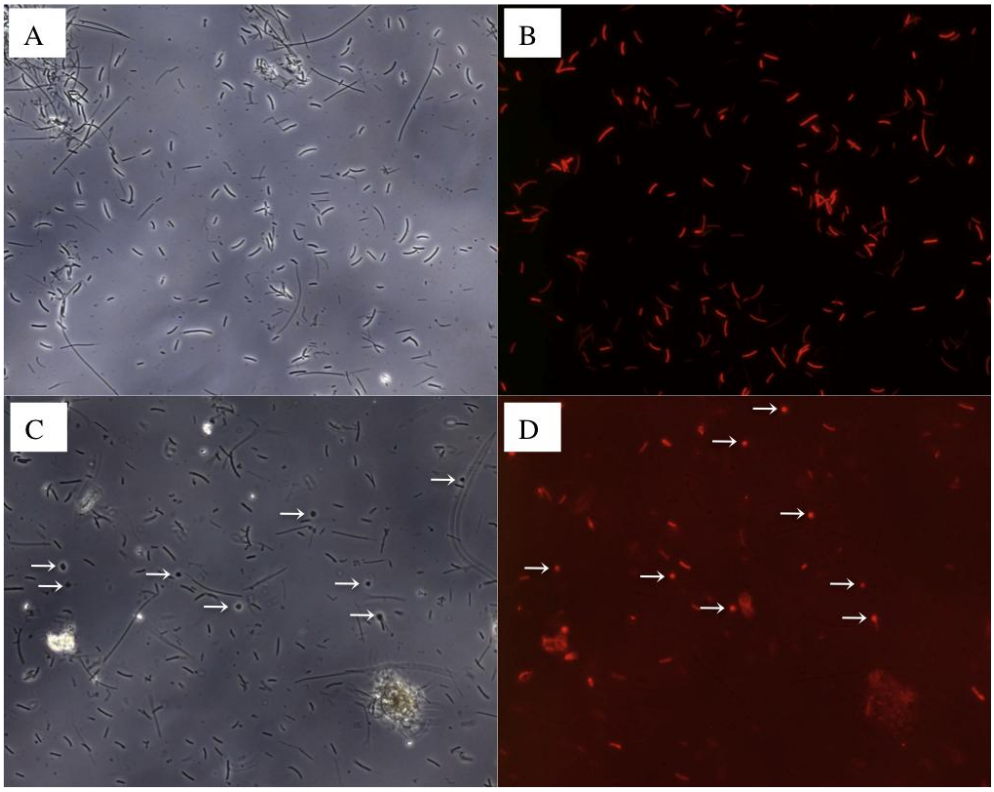



Figure B3.1. Phase contrast (A and C) and autofluorescence (B and D) photomicrographs of mat samples showing *Synechococcus* populations and filamentous cells before (A and B) and after (C and D) treatment with lysozyme prior to in-gel lysis for isolation of HMW DNA. Spheroplasts are denoted with white arrows.

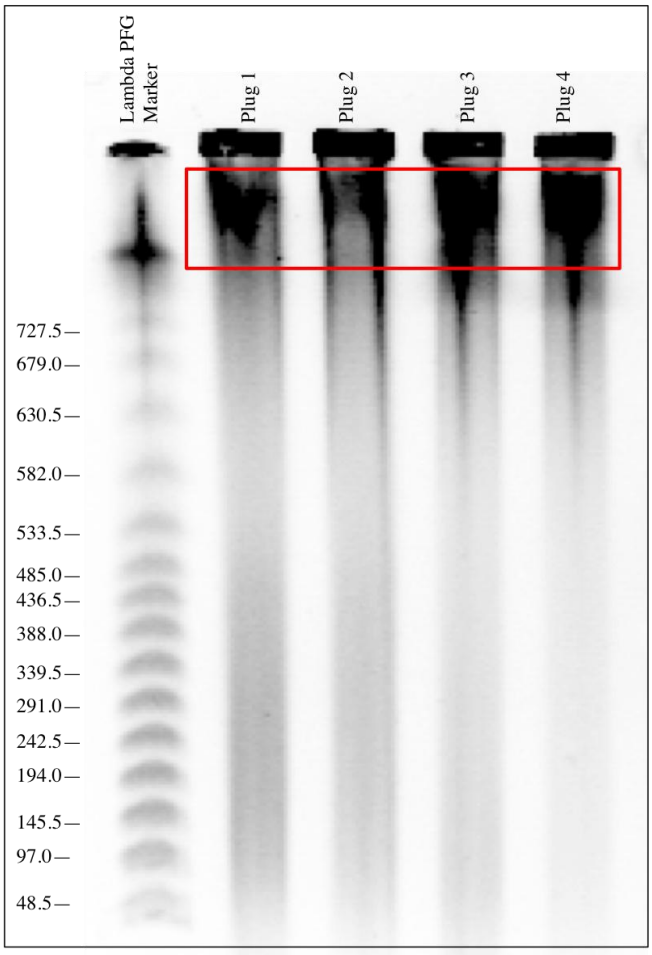



Figure B3.2. Example of pulsed-field gel electrophoresis analysis of HMW DNA obtained after lysozyme pretreatment and in-gel lysis of 4 agarose plugs from the Mushroom Spring 60°C mat sample. HMW DNA used for construction of the M60 BAC library is highlighted in the red box (gel image provided by Q. Tao, Amplicon Express).

BAC Clone Insert Size Analysis

BAC libraries were constructed from HMW DNA and BAC clones were analyzed after digestion with the restriction enzyme NotI to obtain and average insert size for each library (Figure B3.3).

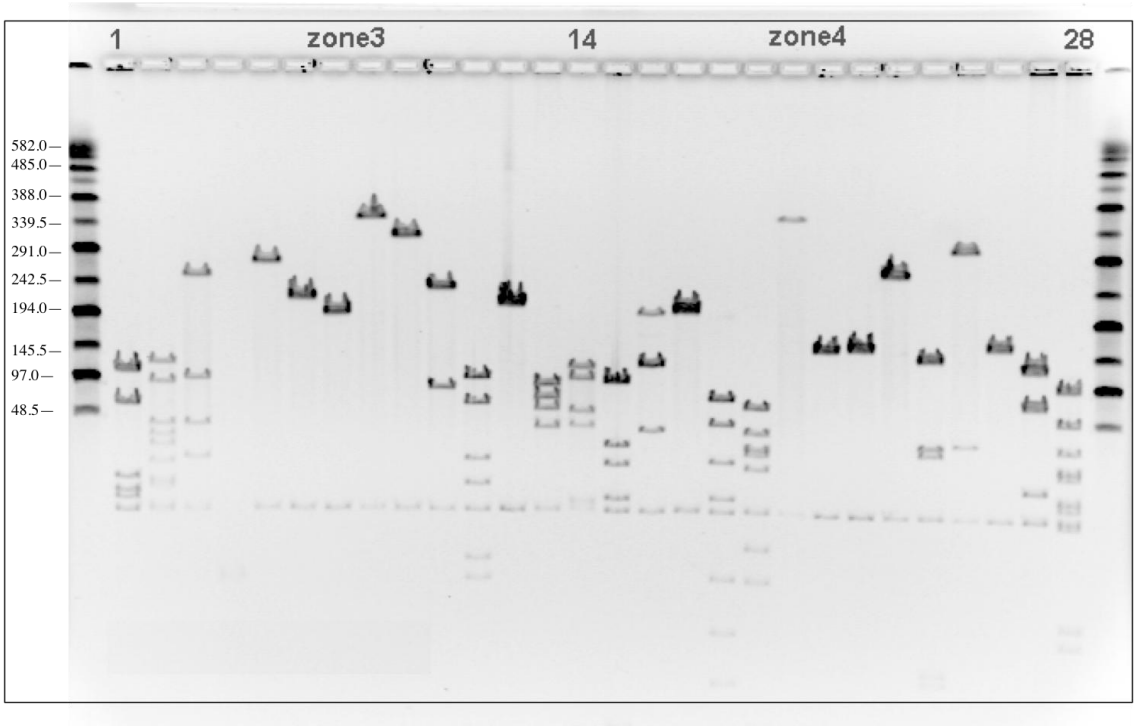



Figure B3.3. Example of a BAC sizing gel from randomly selected M60 BAC clones digested with the restriction enzyme NotI. Sizes of all fragments were added to determine clone insert sizes and insert sizes of all clones sampled from a library were averaged. Lanes 1-14 contain clones from zone 3 (~110 kbp cloned HMW DNA) and lanes 15-28 contain clones from zone 4 (~120 kbp cloned HMW DNA) of the sizing gel (gel image provided by Q. Tao at Amplicon Express).

Probe Screening of BAC Clones

Radioactive, oligonucleotide probes were created to screen 368,640 M60 and M65 BAC clones to identify those that contain a *Synechococcus* A/B lineage-specific 16S rRNA gene (see Methodology). Figure B3.4 shows an example of a nylon filter after exposure on a phosphoimaging cassette. All BAC clones were spotted in duplicate on the nylon filter and positive clones could be visualized as discrete pairs (green arrows).

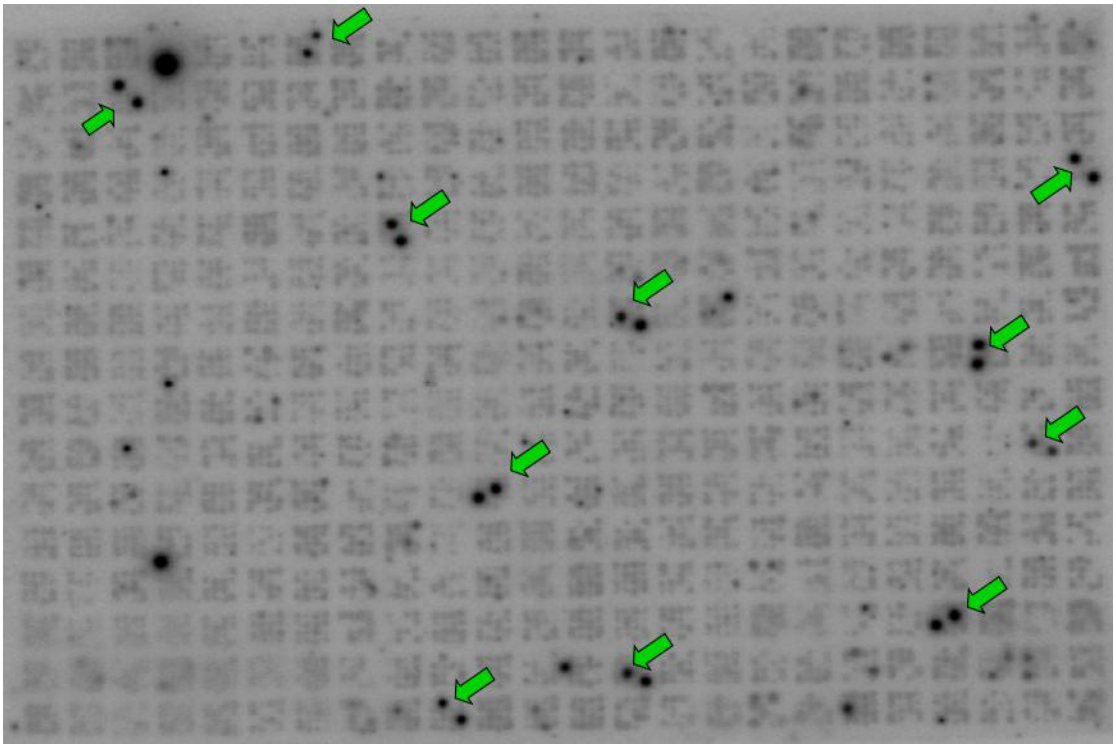


Figure B3.4. Example of 16S rRNA oligonucleotide probing results of the M60 BAC clone library. Here, 3,072 BACs are spotted in duplicate. Hybridization with duplicate spots (green arrows) indicate positive results.

## Histograms of % nt Identities Between Metagenomic Sequences and Reference Genomes that are not Representative of Native Mat Populations

In the end sequence BLAST analysis of 20 genomes, 13 genomes contained BAC end sequences that were mainly disjointly recruited (Figure B3.5 and B3.6 green bars) and had very low % nucleotide identity to the reference genome. This suggests that these reference genomes do not represent the predominant native microbial populations at either site.

## Divergence of *Synechococcus* A´-like Population from the *Synechococcus* strain A Genome

A Ti454 metagenomic library was obtained from samples from Mushroom Spring collected at 68°C, where A´-like *Synechococcus* populations predominate (Becraft, Klatt, Rusch and Ward, unpublished). Sequences obtained from this library that had a best BLAST match to the *Synechococcus* strain A genome were ~10-20% diverged from the *Synechococcus* strain A reference genome (Figure B3.7). This was similar to the divergence found for the sub-population of sequences recruited from M65 by the A genome and interpreted as contributions from *Synechococcus* A´-like community members.

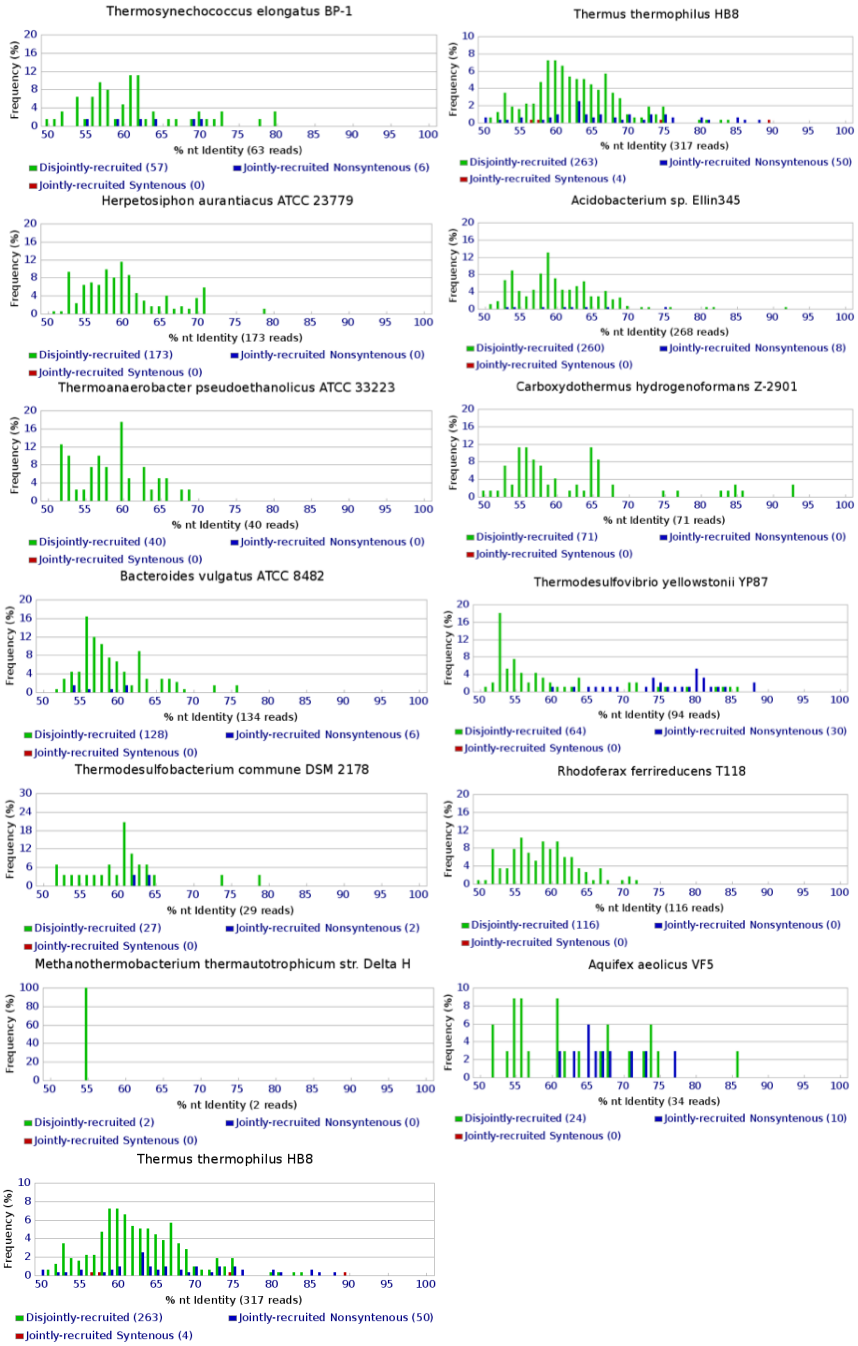



Figure B3.5**.** Frequency distributions of % nucleotide identity of BAC end sequences in the M60 library relative to homologs in recruiting reference genomes that are not representative of native populations. Colors highlight disjointly recruited (green), jointly recruited syntenous (red) and jointly recruited nonsyntenous (blue) metagenomic sequences.

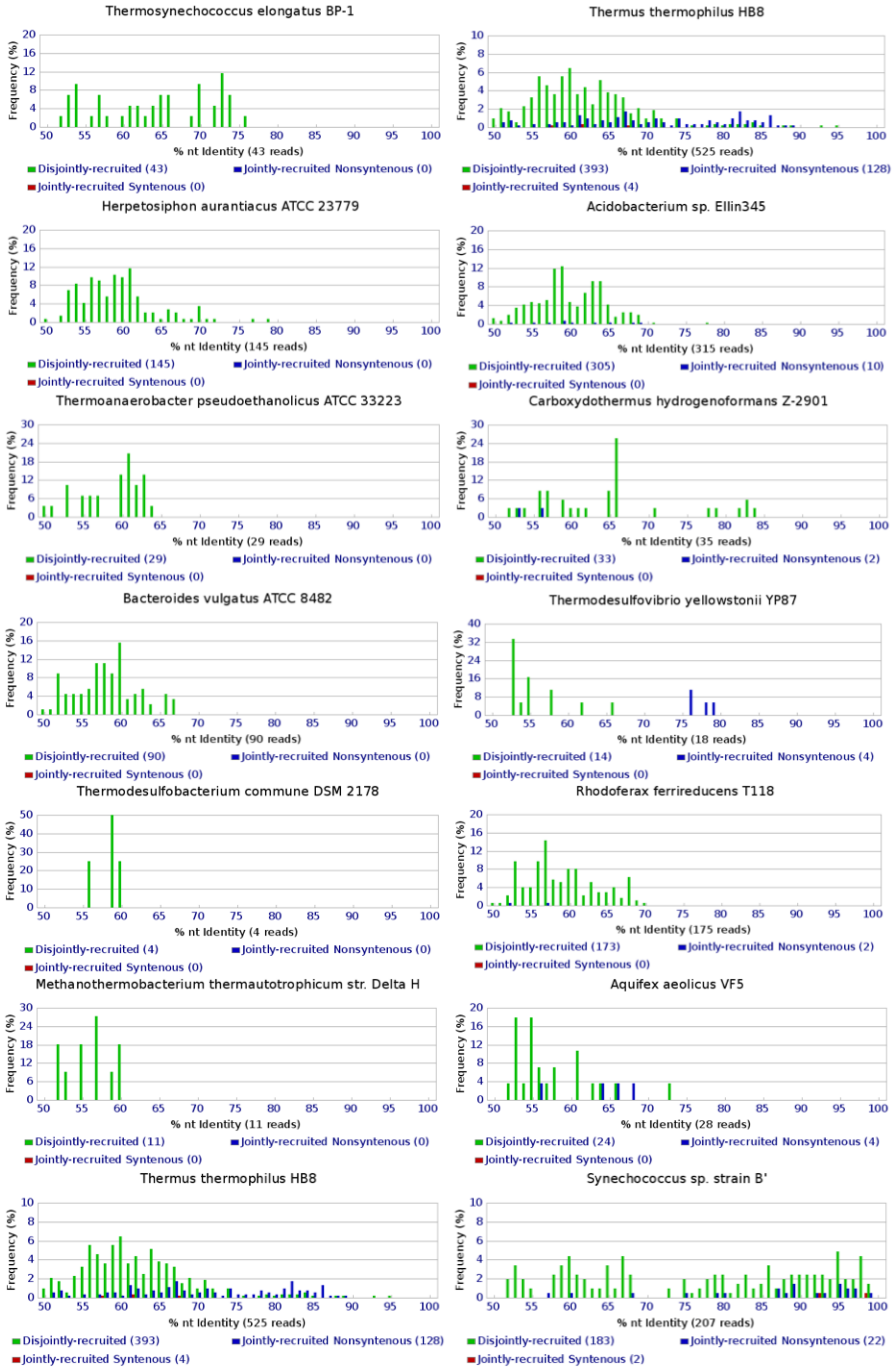



Figure B3.6. Frequency distributions of % nucleotide identity of BAC end sequences in the M65 library relative to homologs in recruiting reference genomes that are not representative of native populations. Colors highlight disjointly recruited (green), jointly recruited syntenous (red) and jointly recruited nonsyntenous (blue) metagenomic sequences

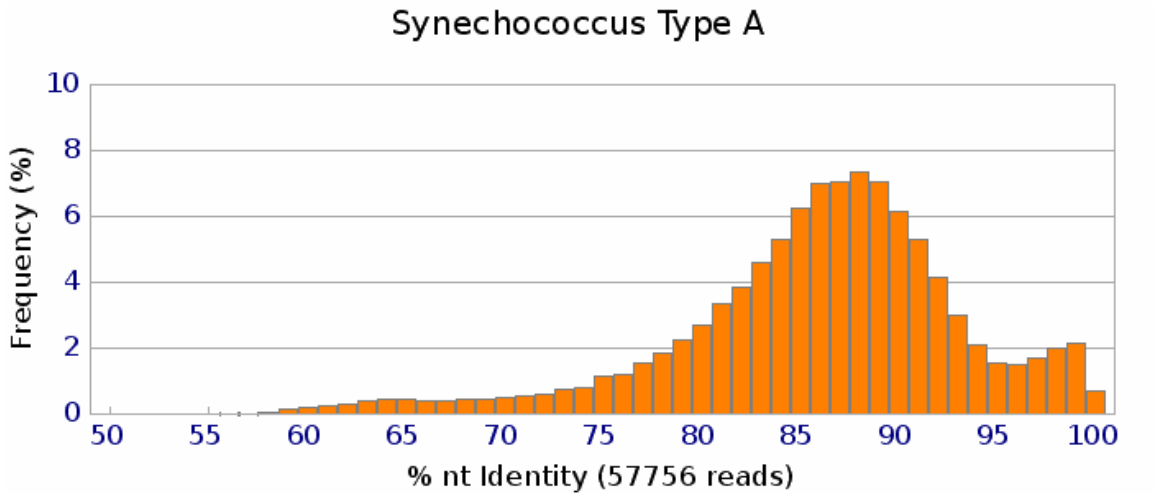



Figure B3.7. Frequency distribution plot of % nucleotide identity between Ti454 metagenomic sequences from a 68°C mat sample from Mushroom Spring, Yellowstone National Park, WY and homologs in the *Synechococcus* strain A genome.

## References

Klatt CG, Wood JM, Rusch DB, Bateson MM, Heidelberg JF, Bryant DA and Ward DM. (2010). Composition of metagenomes from a phototrophic hot spring microbial mat community. In prep.

# APPENDIX C

## SUPPLEMENTAL INFORMATION FOR CHAPTER 4: CULTIVATION-INDEPENDENT MULTI-LOCUS SEQUENCE ANALYSIS OF *SYNECHOCOCCUS* POPULATIONS INHABITING A HOT SPRING CYANOBACTERIAL MAT

## Allelic Profiles

Profiles were generated from a comparison of nucleotide sequence data for each locus and compiled for each clone. Alleles were assigned numbers corresponding to their relative abundance, with allele 1 representing the sequence in highest abundance for that locus. Allelic profiles were then assigned a sequence type (ST) number. ST 1 corresponds to the most abundant ST. All other sequence type designations are arbitrary (Tables C4.1-C4.3). It is important to note that the STs defined in the 7-locus MLSA study (Table 4C.1) do not correspond to the STs defined in the 5-locus MLSA study (Table C4.3), for *Synechococcus* A-like populations.

## eBURST Population Snapshots

Raw population snapshot output from eBURST analysis is shown in Figure C4.1. Figure C4.1A and B are identical to Figures 4.7 and 4.8 except that the designation of clonal complexes and correlation with PEs is not shown.

## Frequency of Recovery of Protein-Encoding Loci in the 16S rRNA Region

Figure C4.2 shows the recovery of BAC clones containing protein-encoding genes as a function of the separation of loci from the 16S rRNA in the *Synechococcus* strain A and B´ genomes.

Table C4.1. Allelic profiles generated from analysis of single nucleotide polymorphisms in the *Synechococcus* A-like BACs for protein-encoding sequence datasets of 7 loci.

| **BAC Clone**[a] | **ST**[b] | ***rbsK*** | ***pk*** | ***hisF*** | ***lepB*** | ***CHP*** | ***aroA*** | ***dnaG*** |
|---|---|---|---|---|---|---|---|---|
| M60B360H08 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 |
| M60B390N18 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 |
| M60B471D04 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 |
| M60B574L23 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 |
| M60B651G15 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 |
| M60B663O15 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 |
| M60B701L08 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 |
| M60B729E01 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 |
| M60B166J20 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| M60B405K13 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| M60B413O21 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| M60B516C08 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| M60B689K07 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| M60B501M02 | 3 | 1 | 1 | 2 | 1 | 1 | 1 | 1 |
| M60B502F22 | 3 | 1 | 1 | 2 | 1 | 1 | 1 | 1 |
| M60B514A21 | 3 | 1 | 1 | 2 | 1 | 1 | 1 | 1 |
| M60B658F14 | 3 | 1 | 1 | 2 | 1 | 1 | 1 | 1 |
| M60B788A07 | 3 | 1 | 1 | 2 | 1 | 1 | 1 | 1 |
| M60B649P18 | 4 | 3 | 1 | 2 | 1 | 1 | 1 | 1 |
| M60B690I15 | 4 | 3 | 1 | 2 | 1 | 1 | 1 | 1 |
| M60B374C09 | 5 | 1 | 1 | 1 | 1 | 1 | 1 | 3 |
| M60B761P11 | 6 | 1 | 1 | 1 | 3 | 1 | 1 | 1 |
| M60B430I13 | 7 | 1 | 1 | 2 | 1 | 1 | 7 | 1 |
| M60B065P14 | 8 | 1 | 2 | 1 | 2 | 1 | 1 | 1 |
| M60B457N03 | 9 | 1 | 2 | 1 | 1 | 1 | 1 | 2 |
| M60B740F14 | 10 | 1 | 2 | 1 | 3 | 1 | 1 | 1 |
| M60B755A11 | 11 | 1 | 2 | 1 | 1 | 2 | 1 | 1 |
| M60B589H13 | 12 | 1 | 2 | 2 | 1 | 1 | 1 | 5 |
| M60B533H13 | 13 | 1 | 4 | 1 | 2 | 1 | 1 | 1 |
| M60B773C01 | 14 | 1 | 4 | 2 | 1 | 1 | 1 | 1 |
| M60B626D13 | 15 | 1 | 14 | 1 | 1 | 1 | 1 | 1 |
| M60B247G10 | 16 | 2 | 1 | 1 | 2 | 2 | 1 | 2 |
| M60B461H19 | 17 | 2 | 1 | 1 | 1 | 2 | 1 | 1 |
| M60B725H17 | 18 | 2 | 1 | 1 | 2 | 2 | 1 | 1 |
| M60B195D22 | 19 | 2 | 2 | 1 | 1 | 3 | 1 | 1 |
| M60B399F15 | 20 | 2 | 2 | 1 | 1 | 2 | 1 | 1 |
| M60B443F23 | 21 | 2 | 4 | 1 | 1 | 2 | 1 | 1 |
| M60B419M11 | 22 | 3 | 1 | 1 | 1 | 1 | 1 | 4 |
| M60B410O11 | 23 | 3 | 2 | 1 | 1 | 1 | 5 | 1 |

Table C4.1 Continued…

| BAC Clone[a] | ST[b] | *rbsK* | *pk* | *hisF* | *lepB* | *CHP* | *aroA* | *dnaG* |
|---|---|---|---|---|---|---|---|---|
| M60B653P08 | 24 | 3 | 2 | 1 | 1 | 2 | 1 | 1 |
| M60B733O24 | 25 | 3 | 2 | 2 | 1 | 2 | 1 | 1 |
| M65B115I21 | 26 | 4 | 2 | 1 | 1 | 1 | 3 | 1 |
| M60B177G14 | 27 | 4 | 3 | 2 | 5 | 2 | 2 | 1 |
| M60B499M16 | 28 | 4 | 3 | 2 | 1 | 4 | 2 | 1 |
| M60B503C09 | 29 | 4 | 3 | 2 | 1 | 2 | 1 | 1 |
| M60B478F03 | 30 | 4 | 11 | 1 | 1 | 2 | 2 | 1 |
| M65B009D20 | 31 | 5 | 5 | 1 | 1 | 1 | 1 | 1 |
| M60B183F17 | 32 | 5 | 5 | 2 | 1 | 1 | 1 | 1 |
| M60B527I20 | 33 | 5 | 7 | 1 | 1 | 1 | 1 | 1 |
| M60B609H12 | 34 | 6 | 8 | 1 | 3 | 1 | 1 | 1 |
| M60B504G12 | 35 | 6 | 12 | 1 | 1 | 1 | 1 | 1 |
| M60B559PO1 | 36 | 7 | 1 | 1 | 1 | 1 | 1 | 1 |
| M60B560H17 | 37 | 7 | 2 | 1 | 2 | 1 | 1 | 2 |
| M60B491M01 | 38 | 8 | 2 | 1 | 1 | 1 | 1 | 1 |
| M60B468P07 | 39 | 8 | 2 | 2 | 1 | 1 | 1 | 1 |
| M60B349M23 | 40 | 9 | 9 | 1 | 4 | 1 | 4 | 2 |
| M60B373K19 | 41 | 10 | 2 | 1 | 1 | 1 | 1 | 1 |
| M60B379G20 | 42 | 11 | 6 | 1 | 1 | 1 | 1 | 1 |
| M60B414C18 | 43 | 12 | 1 | 2 | 1 | 1 | 1 | 1 |
| M60B436G19 | 44 | 13 | 10 | 2 | 4 | 2 | 6 | 2 |
| M60B463L24 | 45 | 14 | 1 | 1 | 1 | 2 | 1 | 1 |
| M60B541E20 | 46 | 15 | 2 | 2 | 1 | 2 | 1 | 1 |
| M60B547N16 | 47 | 16 | 13 | 2 | 6 | 1 | 4 | 1 |
| M60B553A15 | 48 | 17 | 2 | 1 | 1 | 1 | 1 | 1 |
| M60B563B21 | 49 | 18 | 1 | 2 | 1 | 2 | 1 | 1 |
| M60B595E04 | 50 | 19 | 6 | 1 | 3 | 1 | 1 | 6 |
| M60B715I02 | 51 | 20 | 8 | 1 | 2 | 1 | 1 | 1 |
| M60B760D02 | 52 | 21 | 7 | 3 | 1 | 1 | 1 | 1 |
| M65B090N22 | 53 | 22 | 3 | 1 | 1 | 1 | 2 | 1 |
| M65B080L10 | 54 | 23 | 3 | 1 | 1 | 1 | 2 | 1 |
| OSA | 55 | 24 | 1 | 4 | 2 | 1 | 3 | 7 |

[a] Clone names were compiled from spring (M = Mushroom), temperature (60 or 65), BAC library plate (1-792 for M60; 1-168 for M65) and specific well of that plate (A-P and 1-24). Example: M65134I01, Mushroom Spring, 65°C library, plate #134 well I-01.
[b] ST = sequent type.

Table C4.2. Allelic profiles generated from analysis of single nucleotide polymorphisms in the *Synechococcus* B´-like BACs for protein-encoding sequence datasets of 4 loci.

| **Bac Clone**[a] | **ST**[b] | ***aroA*** | ***rbsK*** | ***pcrA*** | **16S/ITS** |
|---|---|---|---|---|---|
| M60B259J13 | 1 | 1 | 1 | 1 | 1 |
| M60B588P04 | 1 | 1 | 1 | 1 | 1 |
| M60B089D09 | 1 | 1 | 1 | 1 | 1 |
| M60B174D01 | 1 | 1 | 1 | 1 | 1 |
| M60B067L11 | 1 | 1 | 1 | 1 | 1 |
| M60B100J22 | 2 | 1 | 1 | 1 | 26 |
| M60B041O5 | 3 | 1 | 1 | 1 | 20 |
| M60B018J02 | 4 | 1 | 1 | 10 | 17 |
| M60B061C08 | 5 | 1 | 1 | 14 | 1 |
| M60B699K08 | 6 | 1 | 3 | 1 | 1 |
| M60B614H18 | 6 | 1 | 3 | 1 | 1 |
| M60B709P04 | 6 | 1 | 3 | 1 | 1 |
| M60B636L16 | 6 | 1 | 3 | 1 | 1 |
| M60B518L02 | 6 | 1 | 3 | 1 | 1 |
| M60B475H14 | 7 | 1 | 3 | 1 | 3 |
| M60B614C12 | 8 | 1 | 3 | 1 | 13 |
| M60B057N07 | 9 | 1 | 12 | 3 | 3 |
| M60B085O17 | 10 | 1 | 12 | 3 | 1 |
| M60B714K10 | 11 | 1 | 16 | 3 | 1 |
| M60B468M14 | 12 | 1 | 28 | 7 | 1 |
| M60B541D06 | 13 | 1 | 29 | 7 | 1 |
| M60B543D23 | 14 | 1 | 30 | 1 | 1 |
| M60B700J15 | 15 | 1 | 33 | 39 | 1 |
| M60B769K22 | 16 | 2 | 2 | 2 | 2 |
| M60B513G21 | 16 | 2 | 2 | 2 | 2 |
| M60B648C08 | 16 | 2 | 2 | 2 | 2 |
| M60B347P16 | 17 | 2 | 2 | 24 | 2 |
| M60B456A21 | 18 | 2 | 2 | 32 | 11 |
| M60B594L19 | 19 | 2 | 6 | 8 | 1 |
| M60B157O24 | 20 | 2 | 6 | 17 | 30 |
| M60B129N02 | 21 | 2 | 7 | 19 | 29 |
| M60B345J11 | 22 | 2 | 8 | 23 | 3 |
| M60B458H11 | 23 | 2 | 8 | 34 | 1 |
| M60B718A23 | 24 | 2 | 8 | 40 | 1 |
| M60B041P24 | 25 | 2 | 10 | 1 | 21 |
| M60B091H06 | 26 | 2 | 14 | 2 | 2 |
| M60B186K12 | 26 | 2 | 14 | 2 | 2 |
| M60B772E10 | 27 | 2 | 36 | 41 | 1 |
| M60B420G12 | 28 | 3 | 5 | 28 | 8 |

Table C4.2 Continued…

| Bac Clone[a] | ST[b] | *aroA* | *rbsK* | *pcrA* | 16S/ITS |
|---|---|---|---|---|---|
| M60B450N20 | 29 | 3 | 5 | 30 | 4 |
| M60B455A17 | 30 | 3 | 5 | 31 | 10 |
| M60B433E13 | 31 | 3 | 15 | 1 | 1 |
| M60B554K06 | 32 | 3 | 17 | 2 | 4 |
| M60B674F15 | 33 | 3 | 17 | 38 | 15 |
| M60B078P12 | 34 | 3 | 21 | 2 | 4 |
| M60B115G12 | 35 | 3 | 24 | 2 | 28 |
| M60B754M13 | 36 | 3 | 35 | 2 | 4 |
| M60B623N20 | 37 | 4 | 4 | 5 | 14 |
| M60B527C11 | 38 | 4 | 4 | 5 | 14 |
| M60B623B05 | 38 | 4 | 4 | 5 | 5 |
| M60B400M20 | 39 | 4 | 4 | 26 | 6 |
| M60B024C23 | 41 | 4 | 9 | 11 | 18 |
| M60B399J16 | 42 | 4 | 9 | 25 | 7 |
| M60B456P18 | 43 | 4 | 15 | 33 | 5 |
| M60B684G15 | 44 | 4 | 32 | 5 | 1 |
| M60B499C24 | 45 | 5 | 7 | 6 | 1 |
| M60B062L15 | 46 | 5 | 20 | 15 | 23 |
| M60B426N01 | 47 | 5 | 27 | 6 | 3 |
| M60B081I21 | 48 | 6 | 13 | 17 | 25 |
| M60B249G11 | 49 | 6 | 13 | 21 | 31 |
| M60B090B21 | 50 | 7 | 22 | 4 | 1 |
| M60B113A15 | 51 | 7 | 23 | 4 | 27 |
| M60B015H24 | 52 | 9 | 18 | 9 | 16 |
| M60B038K14 | 53 | 10 | 6 | 12 | 19 |
| M60B040A11 | 54 | 11 | 19 | 13 | 1 |
| M60B046B18 | 55 | 12 | 11 | 8 | 22 |
| M60B075B13 | 56 | 13 | 10 | 16 | 24 |
| M60B081O16 | 57 | 14 | 11 | 18 | 1 |
| M60B250B15 | 58 | 15 | 25 | 22 | 32 |
| M60B397N23 | 59 | 16 | 2 | 6 | 6 |
| M60B403F16 | 60 | 17 | 7 | 27 | 1 |
| M60B421K10 | 61 | 18 | 26 | 29 | 9 |
| M60B477G06 | 62 | 19 | 2 | 35 | 2 |
| M60B579B21 | 63 | 20 | 31 | 37 | 12 |
| M60B626O23 | 64 | 21 | 16 | 3 | 1 |
| M60B703M24 | 65 | 22 | 34 | 4 | 1 |
| OSB | 66 | 23 | 37 | 42 | 1 |

[a] Clone names were compiled from spring (M = Mushroom), temperature (60), BAC library plate (1-792 for M60) and specific well of that plate (A-P and 1-24). Example: M60703M24, Mushroom Spring, 60$^{o}$C library, plate #703 well M-24.

[b] ST = sequence type.

Table C4.3. Allelic profiles generated from analysis of single nucleotide polymorphisms in the *Synechococcus* A-like BACs for protein-encoding sequence datasets of 5 loci.

| **Clone**[a] | **ST**[b] | ***rbsK*** | ***pk*** | ***lepB*** | ***chp*** | ***aroA*** |
|---|---|---|---|---|---|---|
| M60B788A07 | 1 | 1 | 1 | 1 | 1 | 1 |
| M60B689K07 | 1 | 1 | 1 | 1 | 1 | 1 |
| M60B514A21 | 1 | 1 | 1 | 1 | 1 | 1 |
| M60B166J20 | 1 | 1 | 1 | 1 | 1 | 1 |
| M60B405K13 | 1 | 1 | 1 | 1 | 1 | 1 |
| M60B501M02 | 1 | 1 | 1 | 1 | 1 | 1 |
| M60B374C09 | 1 | 1 | 1 | 1 | 1 | 1 |
| M60B527I20 | 2 | 5 | 7 | 1 | 1 | 1 |
| M65B025J23 | 2 | 5 | 7 | 1 | 1 | 1 |
| M60B559P01 | 3 | 7 | 1 | 1 | 1 | 1 |
| M65B093I20 | 4 | 26 | 16 | 7 | 6 | 3 |
| M65B145I07 | 4 | 26 | 16 | 7 | 6 | 3 |
| M60B183F17 | 5 | 5 | 5 | 1 | 1 | 1 |
| M65B009D20 | 5 | 5 | 5 | 1 | 1 | 1 |
| M65B151I23 | 5 | 5 | 5 | 1 | 1 | 1 |
| M60B468P07 | 6 | 8 | 2 | 1 | 1 | 1 |
| M60B491M01 | 6 | 8 | 2 | 1 | 1 | 1 |
| M60B574L23 | 7 | 1 | 2 | 1 | 1 | 1 |
| M60B651G15 | 7 | 1 | 2 | 1 | 1 | 1 |
| M65B010I14 | 8 | 7 | 2 | 1 | 1 | 1 |
| M65B107B14 | 8 | 7 | 2 | 1 | 1 | 1 |
| M60B773C01 | 9 | 1 | 4 | 1 | 1 | 1 |
| M60B373K19 | 10 | 10 | 2 | 1 | 1 | 1 |
| M65B149H05 | 11 | 26 | 8 | 9 | 6 | 3 |
| M65B155E13 | 12 | 25 | 8 | 9 | 6 | 3 |
| M65B090N22 | 13 | 22 | 3 | 1 | 1 | 2 |
| M65B080L10 | 14 | 23 | 3 | 1 | 1 | 2 |
| M65B141F14 | 15 | 24 | 7 | 1 | 1 | 1 |
| M65B053E10 | 16 | 29 | 17 | 2 | 1 | 1 |
| M60B595E04 | 17 | 19 | 6 | 3 | 1 | 1 |
| OSA | 18 | 24 | 1 | 2 | 1 | 3 |
| M65B111J13 | 19 | 4 | 1 | 4 | 10 | 1 |
| M60B563B21 | 20 | 2 | 1 | 1 | 2 | 1 |
| M65B115H01 | 21 | 4 | 3 | 1 | 5 | 2 |
| M60B653P08 | 22 | 3 | 2 | 1 | 2 | 1 |
| M60B436G19 | 23 | 13 | 18 | 4 | 2 | 8 |
| M60B443F23 | 24 | 2 | 4 | 1 | 2 | 1 |
| M65B032B16 | 25 | 2 | 2 | 1 | 5 | 1 |
| M65B035K07 | 26 | 28 | 5 | 1 | 8 | 1 |

Table C4.3 Continued…

| Clone[a] | ST[b] | *rbsK* | *pk* | *lepB* | *chp* | *aroA* |
|---|---|---|---|---|---|---|
| M65B115I21 | 28 | 4 | 2 | 1 | 1 | 3 |
| M65B150M20 | 29 | 33 | 8 | 9 | 5 | 1 |
| M65B063K13 | 30 | 30 | 19 | 4 | 1 | 3 |
| M65B078P18 | 31 | 31 | 20 | 1 | 1 | 1 |
| M60B065P14 | 32 | 1 | 2 | 2 | 1 | 1 |
| M60B560H17 | 33 | 7 | 2 | 2 | 1 | 1 |
| M60B609H12 | 34 | 6 | 8 | 3 | 1 | 1 |
| M65B019P07 | 35 | 27 | 2 | 2 | 7 | 9 |
| M65B104K09 | 36 | 32 | 2 | 4 | 9 | 3 |
| M65B134I01 | 37 | 25 | 8 | 9 | 11 | 3 |

[a] Clone names were compiled from spring (M = Mushroom), temperature (60 or 65), BAC library plate (1-792 for M60; 1-168 for M65) and specific well of that plate (A-P and 1-24). Example: M65134I01, Mushroom Spring, 65°C library, plate #134 well I-01.
[b] ST = sequent type.

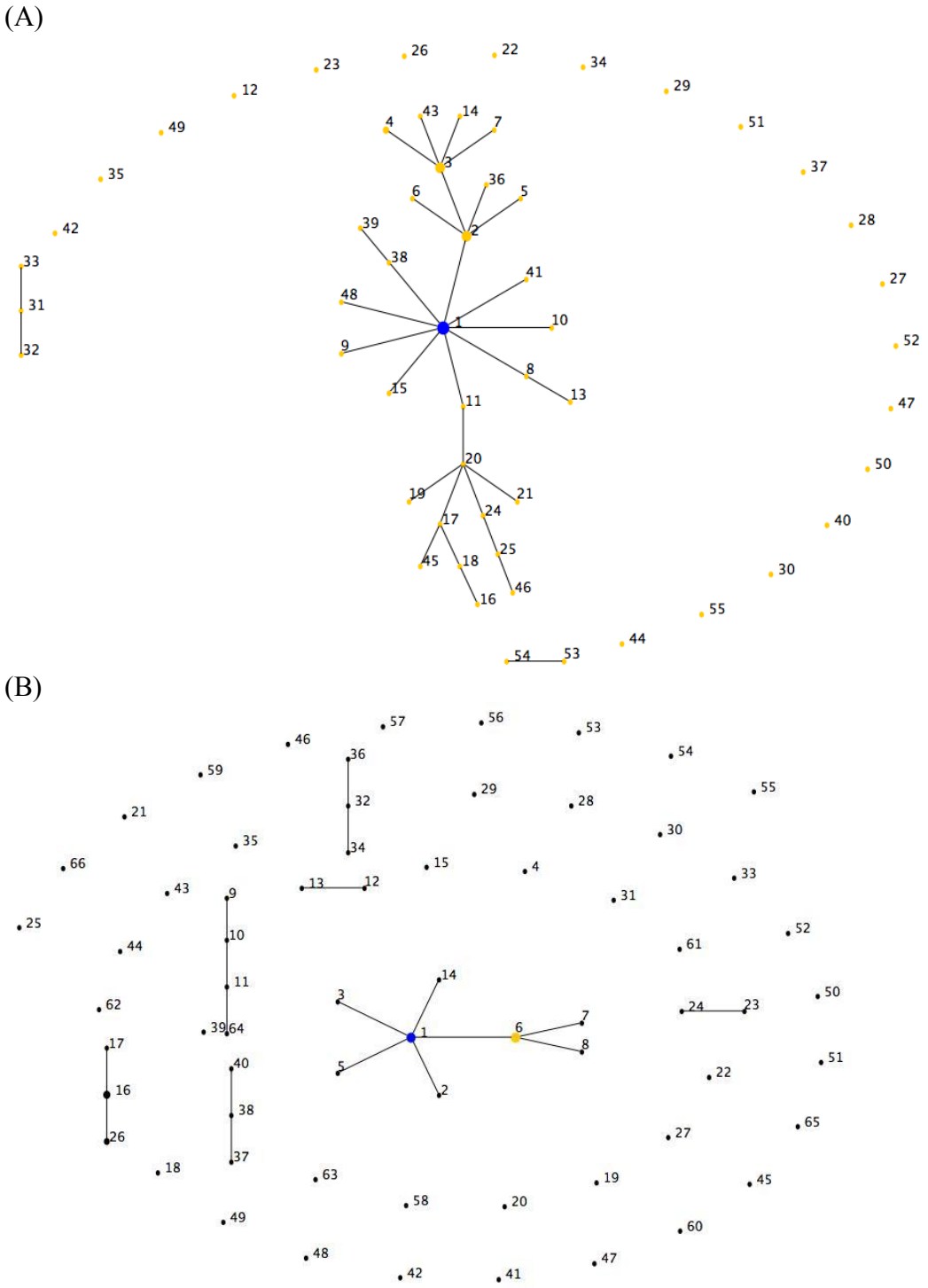



Figure C4.1. Raw output of the population snapshot of (A) *Synechococcus* A-like BACs corresponding to Figure 4.7 and (B) B´-like BACs corresponding to Figure 4.8, from eBURST analysis.

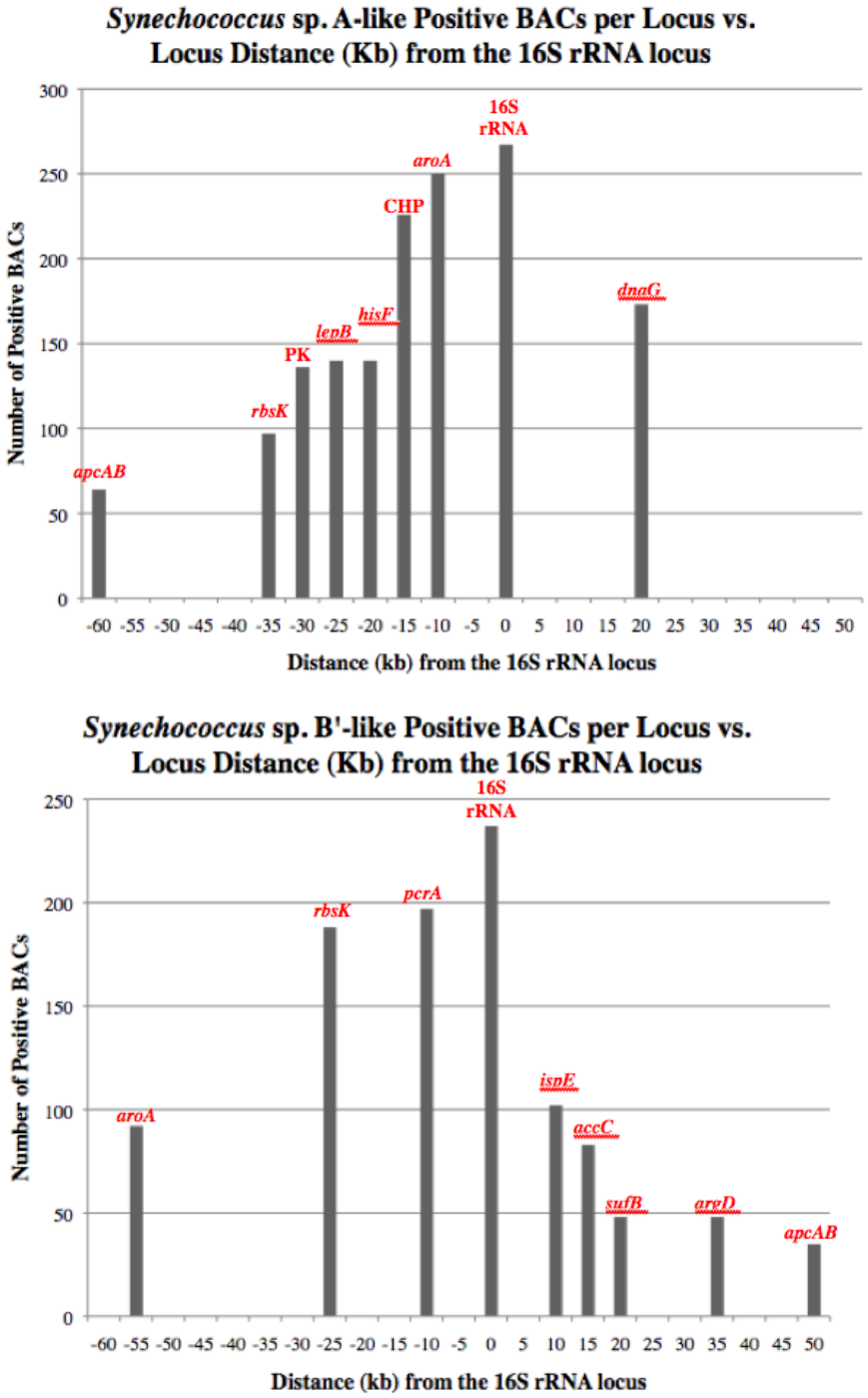



Figure C4.2. Number of positive BACs per locus as a function of separation of the loci from the 16S rRNA locus.

## Correspondence Between BAC and PCR Clone Sampling of Mat *Synechococcus* Diversity

Figure C4.3 shows a phylogeny constructed from *rbsK* sequences of clones retrieved from the same Mushroom Spring mat DNA sample by PCR-based cloning and BAC cloning protocols. BAC clone sequences disperse within the diversity assayed using PCR cloning techniques. Nearly all of the PEs predicted by ES were represented by both PCR and BAC clones, showing that the BAC cloning technique samples across ecotype diversity.

## eBURST Population Snapshots: Relaxation of Criteria

Figures C4.4 and C4.5 A and B show the same population snapshots as Figures 4.7, 4.8 and 4.9, respectively, however the requirements for ≥3 SLVs in a clonal complex has been relaxed to allow for DLVs and/or 2 SLVs. All other sequence types, if not connected to a consensus sequence type, are displayed in a spiral pattern surrounding the clonal complexes. Both Figures C4.4 and C4.5 show an increase in the number of clonal complexes that can be defined. There are 3 clonal complexes visualized for the A-like *Synechococcus* 5-locus MLSA study in Figure C4.4 compared to just 2 (seen by the splitting of clonal complex Figure C4.4B to create the 2 clonal complexes in Figure 4.9 of the main text). There are 4 *Synechococcus* B′-like clonal complexes visualized in Figure C4.5 compared to just 2 clonal complexes in Figure 4.8. Figure C4.5 also shows the temperature samples that the sequence types came

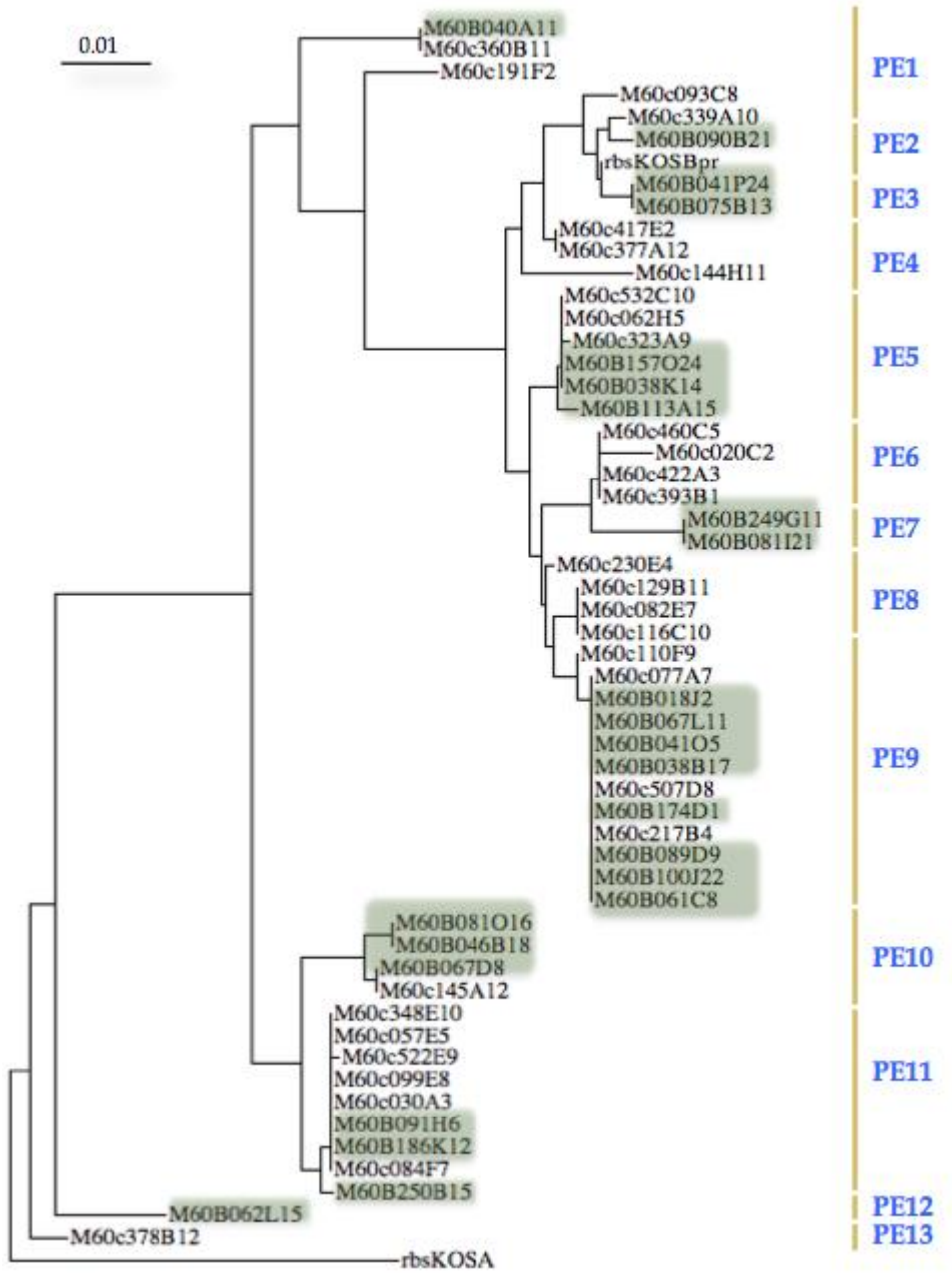



Figure C4.3. Neighbor-joining phylogenetic tree for *Synechococcus* B′-like *rbsK* sequences obtained from the same Mushroom Spring 60°C mat sample using either PCR- or BAC-based cloning approaches. Vertical bars indicate putative ecotype (PE) demarcation from ES analysis. BAC clone sequences are highlighted in green.

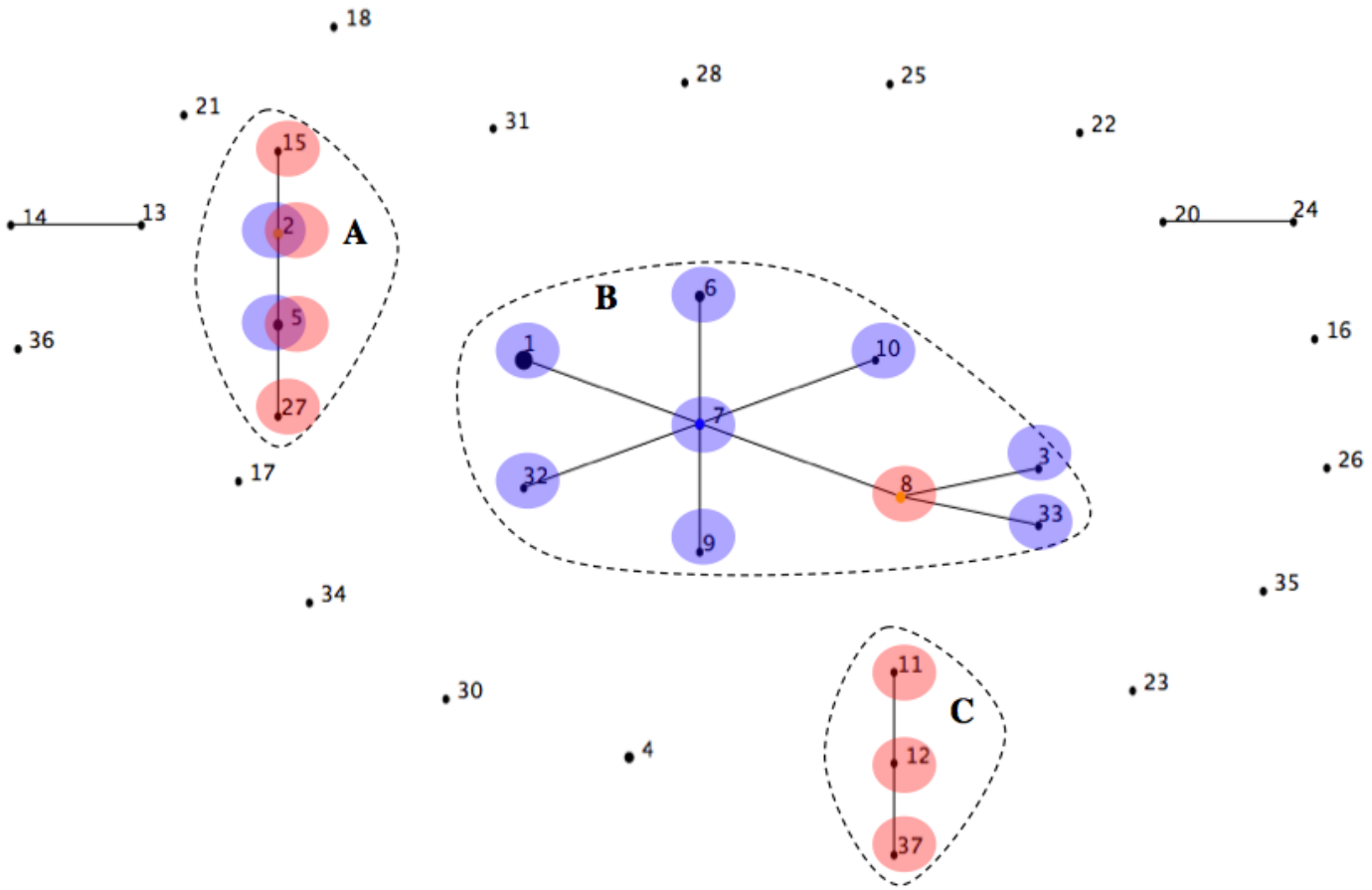
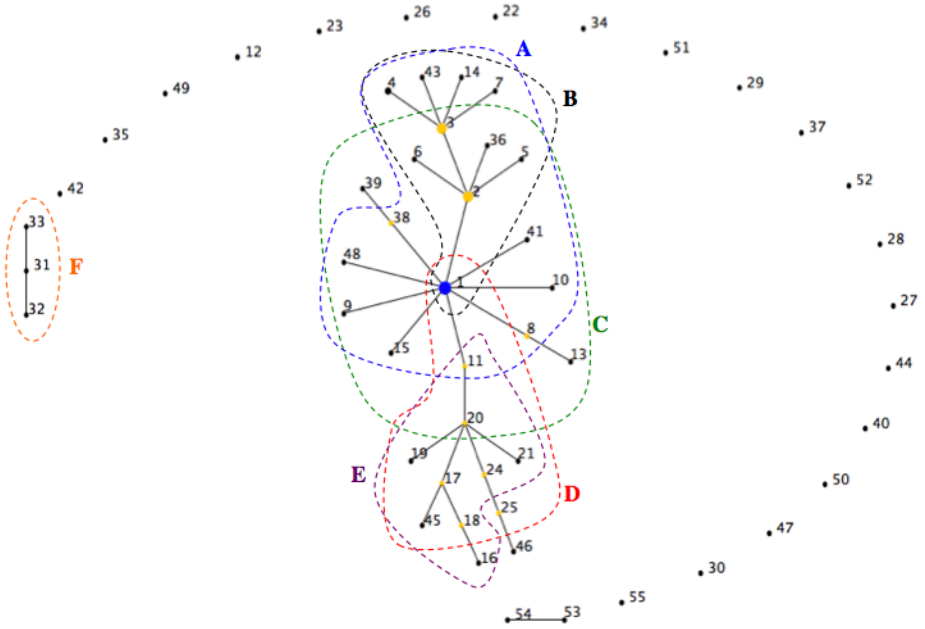



Figure C4.4. Population snapshot of *Synechococcus* A-like BACs from the 5-locus eBURST analysis (corresponding to Figure 4.9) showing 3 clonal complexes (A-C, bounded by dashed lines), one that is temperature defined (C). The criteria have been relaxed to allow for DLVs and 2 SLVs in a clonal complex.

(A)




(B)

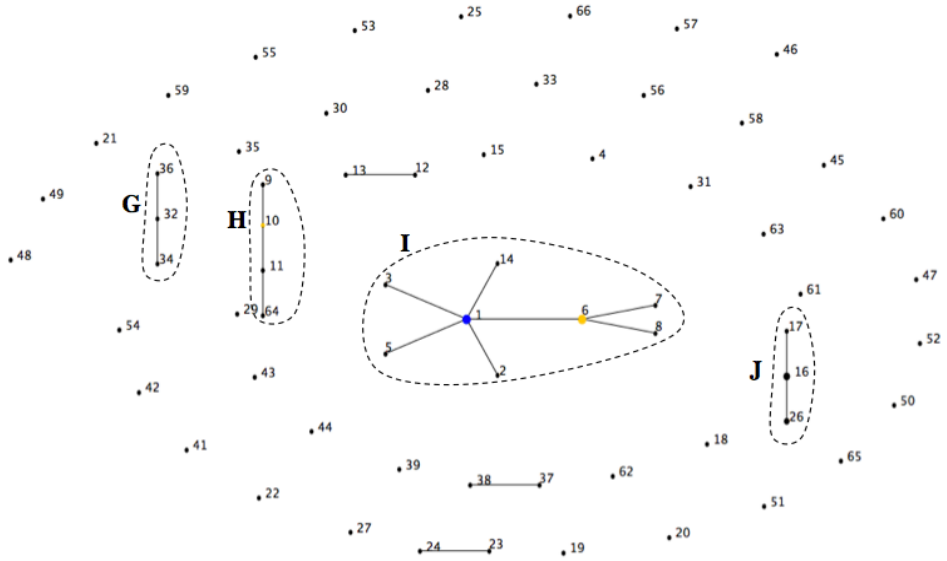



Figure C4.5. Population snapshot of (A) *Synechococcus* A-like BACs (corresponding to Figure 4.7) and (B) B´-like BACs (corresponding to figure 4.8) from eBURST analysis showing 6 potential clonal complexes for A (A-F, bounded by colored dashed lines; these represent some but not all possible clonal complexes in this output) and 4 clonal complexes for B´ (G-J, bounded by black dashed lines). The criteria have been relaxed to allow for DLVs and 2 SLVs a clonal complex.

from and while there is evidence of a temperature defined eBURST clonal complex (Figure C4.4C) it does not correlate to a temperature specific PE clade (compared to Figure 4.6).

## Recombination Signals and Rate Plots

The p-values for the recombination signals detected by the various programs in RDP3 discussed in Tables 4.12 and 4.13 can be found in Tables C4.3 and C4.4. Recombination rate plots were generated as described in the Methodology. Putative 'hot spots' of recombination can be seen in Figure C4.5 (A and B), as spikes (demarcated with stars) along the positional alignment of concatenated sequences. For *Synechococcus* A-like BACs, a small spike occurs in the *rbsK* locus (nucleotides 1-583) and a large spike in the *PK* locus (nucleotides 584-1249). *Synechococcus* B´-like BACs show a predominant spike in the *rbsK* locus (nucleotides 685-1506) and a smaller spike in the 16S rRNA/ITS region (nucleotides 2131-2878).

## RDP3 Analyses

Multiple lines of evidence were used to support whether or not recombination was present within a given gene. Phylogenetic incongruency and the number and pattern of single nucleotide polymorphisms have been used as defined in the main text, however, break points and statistical support for the results are observable in these kinds of analyses. RDP3 relies on the strength of statistically supported analyses that are based on

Table C4.4. P-values for RDP3 analysis of recombinants for *Synechococcus* A-like BACs. Corresponds to Table 4.12 in the text.

| Locus | Recombinant BAC | MaxChi | GENECONV | RDP | Chimaera | Siscan | 3Seq | LARD |
|---|---|---|---|---|---|---|---|---|
| *rbsK* | M60B541E20 | 0.001 | * | 0.01 | * | 0.1 | * | * |
| *rbsK* | M60B504G12 | * | 0.002 | n.d. | * | * | n.d. | * |
| *rbsK* | M65B150M20 | * | * | 0.001 | * | * | * | * |
| *rbsK* | M60B373K19 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *aroA* | M60B430I13 | 0.004 | 0.005 | n.d. | 0.001 | * | * | * |
| *dnaG* | M60B595E04 | 0.002 | 0.002 | n.d. | * | * | * | * |
| *rbsK* | M60B516C08 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B166J20 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B405K13 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B689K07 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B413O21 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B065P14 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B457N03 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B740F14 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B755A11 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B626D13 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B491M01 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B553A15 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B663O15 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B651G15 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B471D04 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B390N18 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B360H08 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B701I08 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B574I23 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B788A07 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B514A21 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B658F14 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B502F22 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B501M02 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B761P11 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B559P01 | 0.001 | 0.03 | n.d. | * | * | * | * |
| */rbsK* | M60B374C09 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B649P18 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B690I15 | 0.001 | 0.03 | n.d. | * | * | * | * |
| */rbsK* | M60B430I13 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B773C01 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B414C18 | 0.001 | 0.03 | n.d. | * | * | * | * |
| *rbsK* | M60B461H19 | 0.017 | 0.046 | n.d. | n.d. | * | * | 0.004 |
| *rbsK* | M60B195D22 | 0.017 | 0.046 | n.d. | n.d. | * | * | 0.004 |
| *rbsK* | M60B443D23 | 0.017 | 0.046 | n.d. | n.d. | * | * | 0.004 |
| *rbsK* | M60B653P08 | 0.001 | 0.03 | n.d. | * | * | * | * |

*P-value < 0.001

n.d.: Recombinants not defined by this method.

Table C4.5. P-values for RDP3 analysis of recombinants for *Synechococcus* B´-like BACs. Corresponds to Table 4.13 in the text.

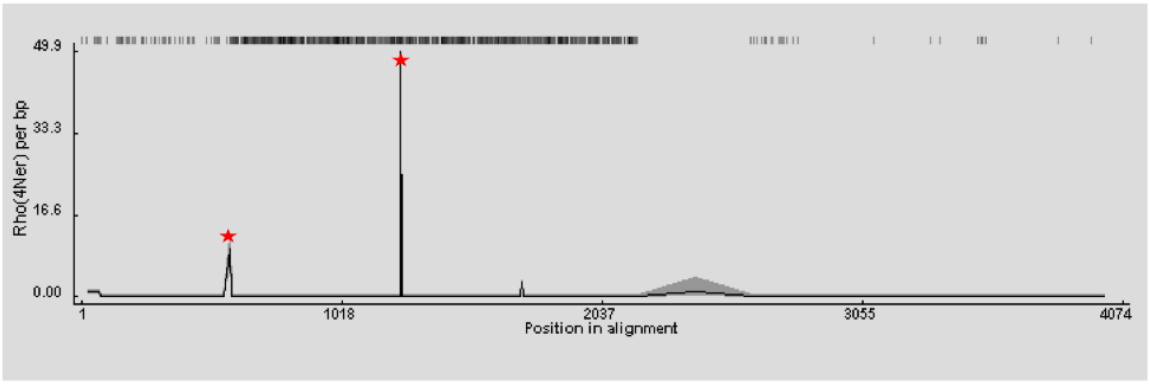
| Locus | Recombinant BAC | MaxChi | GENECONV | RDP | Chimaera | Siscan | 3Seq | LARD |
|---|---|---|---|---|---|---|---|---|
| *aroA* | M60B015H24 | * | * | nd | nd | nd | * | * |
| *pcrA* | M60B541D06 | 0.003 | * | * | * | * | * | * |
| *rbsK* | M60B062L15 | * | * | * | * | * | * | * |
| *rbsK* | M60B700J15 | * | * | * | * | * | * | * |
| *rbsK* | M60B186K12 | * | * | * | * | * | * | * |
| *rbsK* | M60B543D23 | * | * | * | * | * | * | * |
| *rbsK* | M60B541D06 | * | * | * | * | * | * | * |
| *rbsK* | M60B554K06 | * | n.d. | 0.099 | * | * | * | * |
| *rbsK* | M60B347P16 | * | * | * | * | * | * | * |
| *rbsK* | M60B714K10 | * | * | * | 0.03 | * | * | * |
| *aroA* | OSB[++] | * | * | nd | nd | nd | * | * |
| *rbsK* | M60B700J15 | * | * | * | * | * | * | * |
| *rbsK* | M60B397N23 | * | * | * | * | * | * | * |
| *rbsK* | M60B626O23 | * | * | * | * | * | * | * |
| *rbsK* | M60B085O17 | * | * | * | * | * | * | * |
| *rbsK* | M60B057N07 | * | * | * | * | * | * | * |
| *rbsK* | M60B426N01 | * | * | * | * | * | * | * |
| *rbsK* | M60B046B18 | * | * | nd | nd | nd | * | * |
| *rbsK* | M60B081O16 | * | * | nd | nd | nd | * | * |
| *rbsK* | M60B091H06 | * | * | nd | nd | nd | * | * |
| *rbsK* | M60B250B15 | * | * | nd | nd | nd | * | * |
| *rbsK* | M60B040A11 | * | * | nd | nd | nd | * | * |
| *rbsK* | M60B456A21 | * | * | * | * | * | * | * |
| *rbsK* | M60B477G06 | * | * | * | * | * | * | * |
| *rbsK* | M60B513G21 | * | * | * | * | * | * | * |
| *rbsK* | M60B648C08 | * | * | * | * | * | * | * |
| *rbsK* | M60B769K22 | * | * | * | * | * | * | * |
| *rbsK* | M60B579B21 | * | * | * | * | * | * | * |

*P-value < 0.001

n.d.: Recombinants not defined by this method.

[++]Genome annotation available at: http://cmr.jcvi.org/cgibin/CMR/GenomePage.cgi?org=gymb.

(A)




(B)

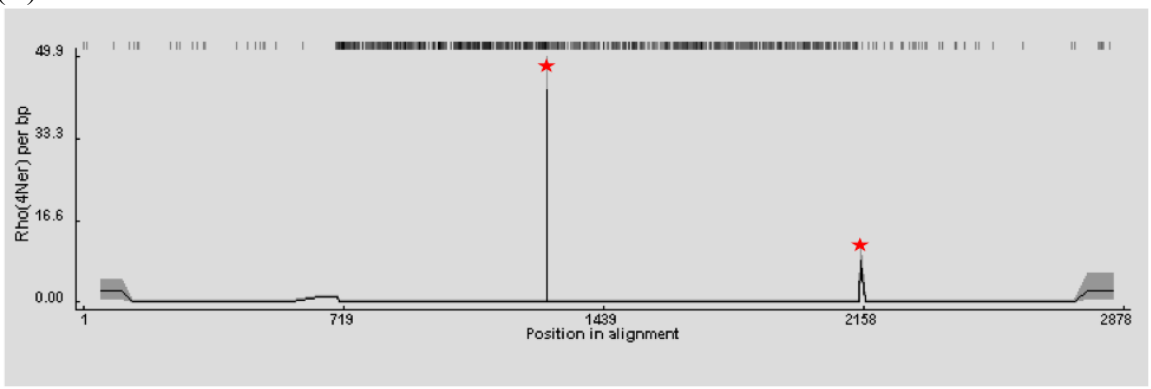



Figure C4.6. Recombination rate plots generated from LDHat interval analysis of concatenated sequence datasets for *Synechococcus* populations (A) A-like BACs and (B) B´-like BACs. Shaded bars above the plot show degree of nucleotide identity ranging from high to low (light to dark coloring) and spikes are demarcated with red stars.

examination of phylogenetic data as well as sequence data. The following summary of descriptions found in the RDP3 manual and respective publications is included to provide an overview; the reader is referred to these resources for more explicit descriptions. Extensive text is quoted from these sources.

RDP Method (Martin and Rybicki, 2000)

The RDP Method scans an alignment of sequences by triplets, discards non-informative sites and uses a sliding window that moves one nucleotide at a time and

calculates a percent nucleotide identity for the three sequence pairs. A UPGMA dendogram is used to infer phylogenetic relationship among the three sequences. A p-value is derived based on the probability that the nucleotide arrangement in the recombinant region arose by chance. This is approximated using a binomial distribution. The method "does not account for rate variation that would lead to non-clocklike trees based on the UPGMA tree constructed during the run". No substitution model is implemented and "this method is reliable for sequences within alignments that have >70% nucleotide identity."

GENECOV (Padidam et al., 1999)

GENECONV looks for regions within a sequence alignment in which sequence pairs are "sufficiently similar" to suspect that they may have arisen through recombination. "Monomorphic sites are excluded from the alignment as a control for constant or highly selected sites." What remains is an alignment of polymorphic sites. For every possible sequence pair in the alignment, regions are found that are either "(1) identical and unusually long for that pair of sequences or (2) have an unusually high degree of similarity" A similarity score is generated and p-values are assigned to high scoring regions through "permutations" and a BLAST method. The GENECONV manual can be obtained from http://www.math.wustl.edu/~sawyer/geneconv/.

MaxChi (Maynard Smith, 1992)

Maynard Smith (1992) proposed a method (called the maximum $\chi^2$ method) for identifying recombination breakpoints that is implemented in the program MaxChi (Posada and Crandall, 2001). Given an alignment, MaxChi scans sequence triplets and seeks to identify recombination breakpoints by looking for "significant differences in the proportions of variable and non-variable polymorphic alignment positions in adjacent regions of sequence." All monomorphic (noninformative) sites in an alignment are discarded. There is also an option to discard sites that contain gaps, which was utilized in this study. The significance (p-value) of $\chi^2$ peaks is determined by a "permutation test". MaxChi provides information on the putative positions of potential breakpoints but does not give information on the extent of recombinant regions. When the "scan triplets" setting is used RDP3 will make an attempt to match potential breakpoints and will "assume that sequences between matched breakpoints are within a single recombinant region." MaxChi will include sites that vary between all three sequences in the analysis, which becomes a problem when the alignment contains very divergent sequences and will lead to false positives– "i.e. results with high *P*-values that cannot be confirmed with any other detection methods."

Chimaera (Posada and Crandall, 2001)

Chimaera is a modification of the maximum $\chi^2$ method. The differences between Chimaera and MaxChi are (1) how polymorphic sites are chosen and (2) that Chimaera can only be used to screen triplets, whereas MaxChi can be used to scan pairs of

sequences or triplets; this option was not used in this study. All monomorphic sites are discarded. As with MaxChi, Chimaera provides information on the positions of potential breakpoints but does not give information on the extent of recombinant regions.

Siscan (Gibbs et al., 2000)

"Sister scanning was developed as a means of analyzing different kinds of signals in nucleotide sequence data." Every possible triplet in an alignment is examined for evidence of recombination using a "fourth" sequence. The fourth sequence is either constructed by "horizontal randomization" of one of the sequences in a triplet or "drawn from the alignment." "Each column of the alignment is sorted into categories and randomized in a process called 'vertical' randomization to produce permuted alignments." A Z-test is used to determine if the columns in the actual alignment significantly differ from the columns generated in the "vertical" randomization. Siscan examines all sites rather than just variable sites.

3Seq (Boni et al., 2007)

3Seq is a nonparametric test for recombination that uses only informative sites, does not use a sliding window and identifies breakpoints and sequences involved in the "mosaic-producing event". 3Seq scans triplet sequences. The method considers two parent sequences that may have recombined, with one or two breakpoints, to form a third sequence the "child sequence". "Excess similarity of the child sequence to a candidate recombinant of the parents" is a sign of recombination and the maximum value of this "excess similarity" is used as a test statistic. 3Seq uses "polynomial time" to compute a

table of p-values that is used as a reference instead of Monte Carlo methods to calculate a p-value. The method considers all possible breakpoints, finding the optimal window size that should be used for inferring recombination in a particular sequence triplet.

LARD (Holmes et al., 1999)

LARD tests the hypothesis of "completely clonal evolution *vs*. the hypothesis of clonal evolution for segments on either side of a breakpoint." LARD detects recombination breakpoints using a method similar to that used by MaxChi. The method scans an alignment of three sequences for the point in the alignment that "optimally" separates regions of "conflicting phylogenetic signal." An alignment is "partitioned into two pieces" and maximum likelihood trees are constructed for each separate piece. The probability that two trees on either side of a partition differ (i.e. that they have different branch lengths due to recombination) is determined by a likelihood ratio test that compares the likelihood scores of the two trees with that of a tree constructed from the full alignment. Every possible "partition" of the alignment is examined as above and "the partition that separates trees with the greatest likelihood (i.e. the greatest difference in branch lengths) is the most probable recombination breakpoint." This method is computationally very slow. It is included in order to confirm recombinant sequences identified by other methods, as suggested by Vitorino et al. (2008). "LARD can account for rate heterogeneity and distinguish recombinants unless the sequence is evolving at a significantly faster or slower rate compared to the other sequences in the triplet. This problem however, is not limited to LARD—all recombination detection programs have this problem."

References

Boni MF, Posada D and Feldman MW. (2007). An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genet* **176**: 1035-1047.

Gibbs MJ, Armstrong JS and Gibbs AJ. (2000). Sister scanning: a Monte Carlo procedure for assessing signals in recombination sequences. *Bioinformatics* **16**: 573-582.

Holmes EC, Worobey M and Rambaut A. (1999). Phylogenetic evidence for recombination in Dengue virus. *Mol Biol Evol* **16**: 405.

Martin D and Rybicki E. (2000). RDP: detection of recombination amongst aligned sequences. *Bioinformatics* 16: 562-563.

Maynard Smith J. (1992). Analyzing the mosaic structure of genes. *J Mol Evol* **34**: 126-129.

Padidum M, Sawyer S and Fauquet CM. (1999). Possible emergence of new Geminiviruses by frequent recombination. *Virol* **265**: 218-225.

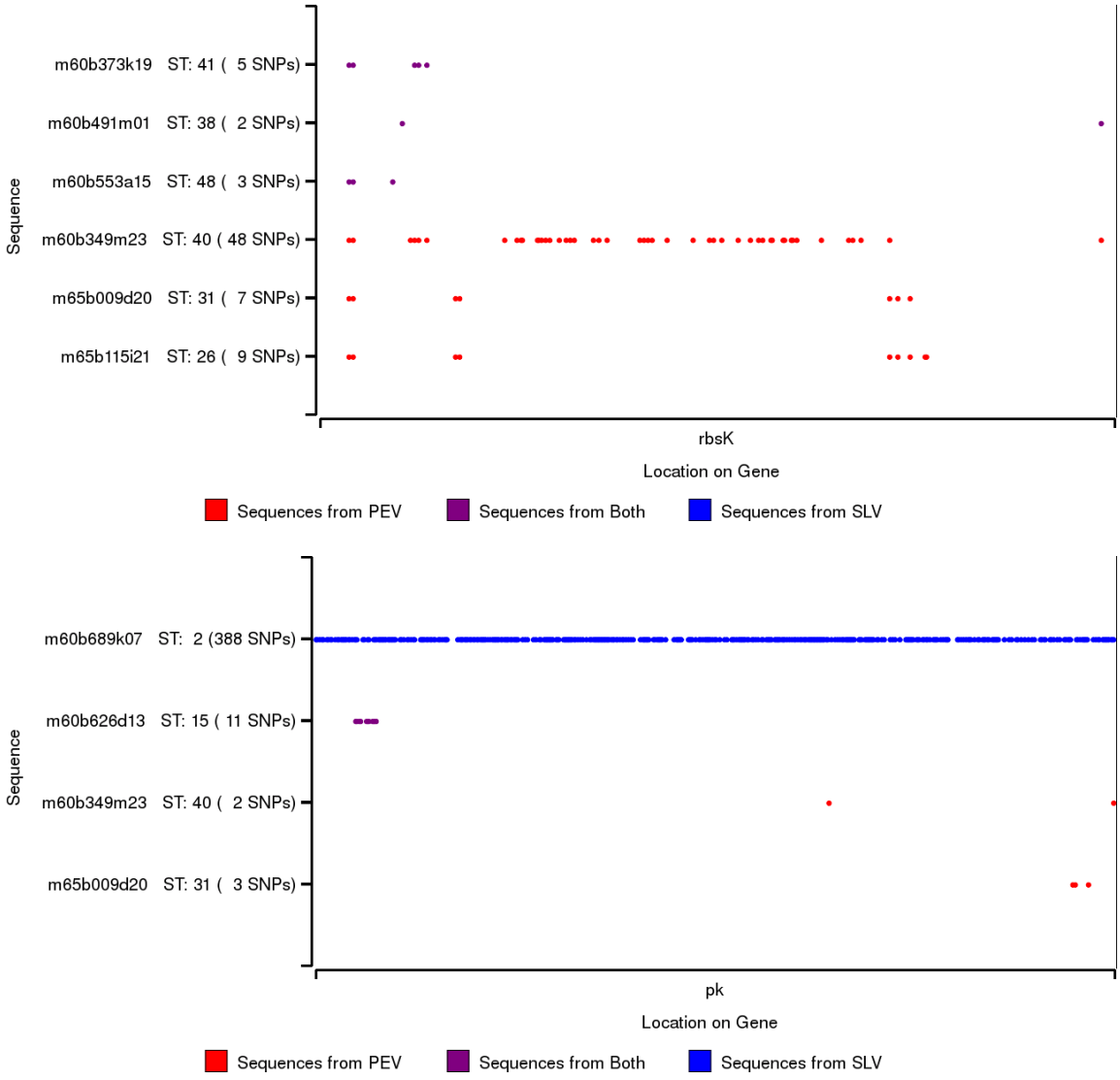
Posada D and Crandall KA. (2001). Evaluation of methods for detecting recombination form DNA sequences: computer simulations. *Proc Natl Acad Sci* **98**: 13757-13762.

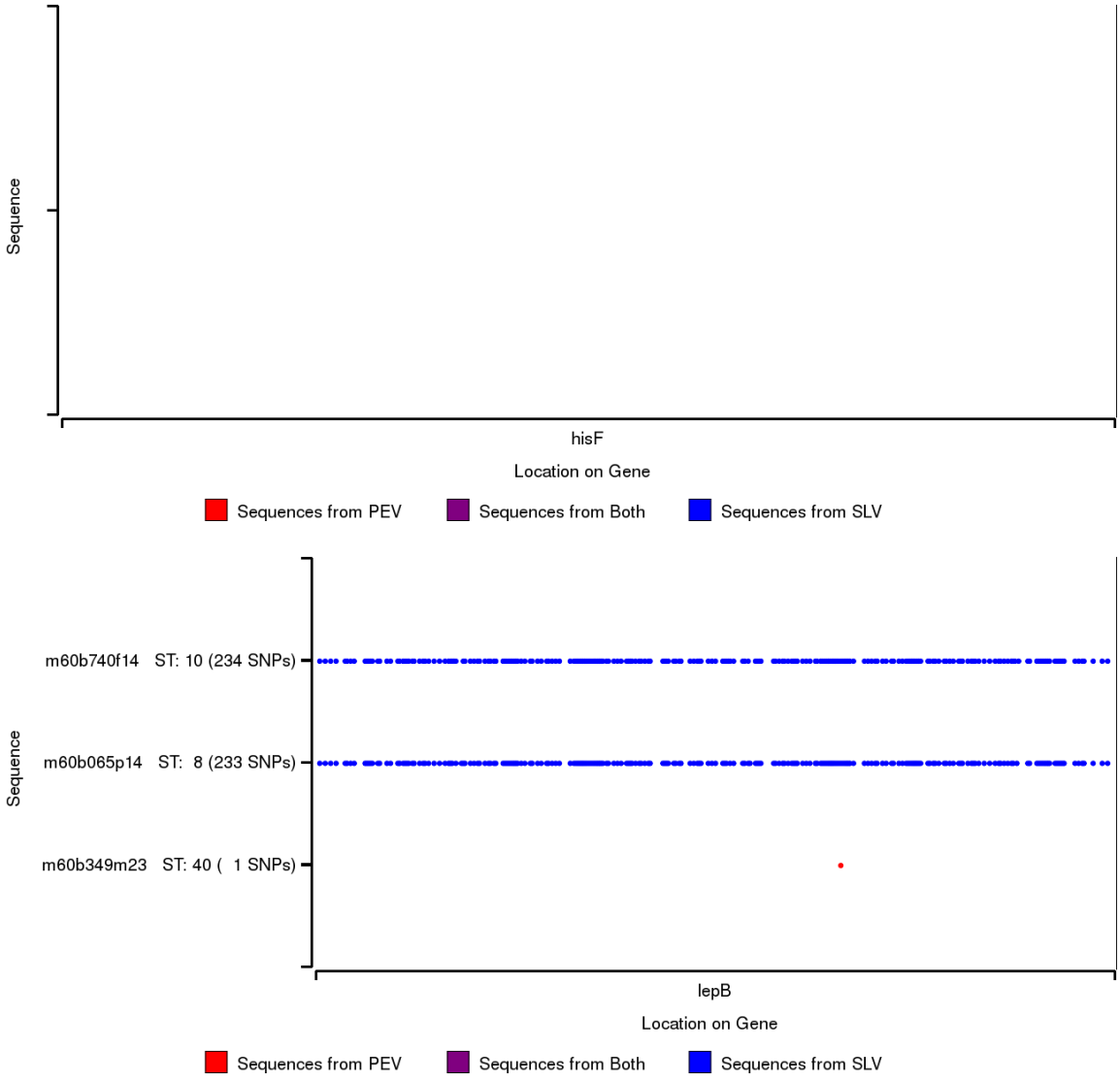
Vitorino LR, Margos G, Feil EJ, Collares-Pereira M, Ze-Ze L and Kurenbach K. (2008). Fine-scale phylogeographic structure of *Borrelia lusitaniae* revealed by multilocus sequence typing. *PLOS ONE* **3**: 1-13.

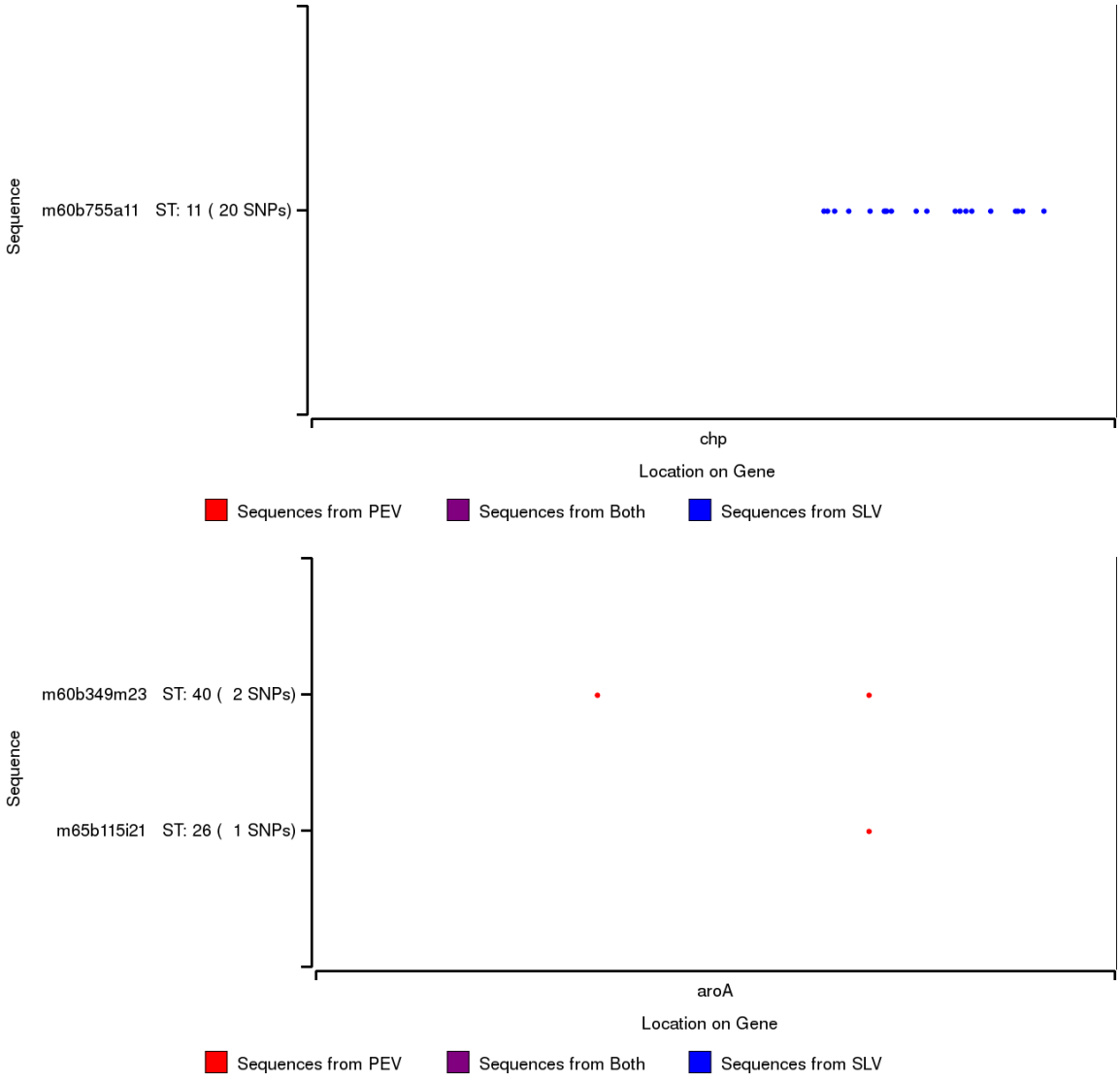
# APPENDIX D

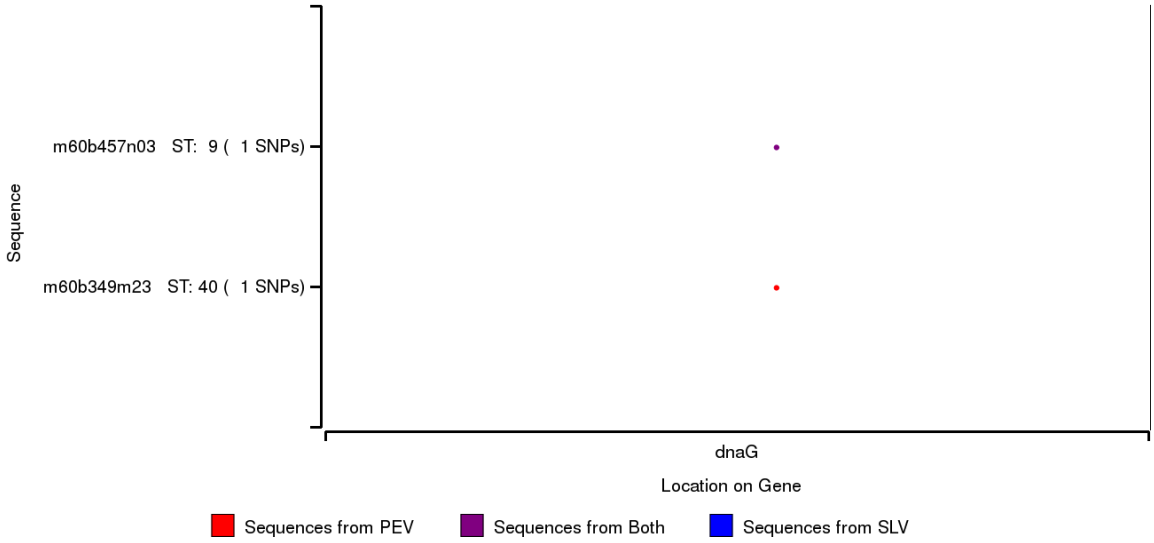
## DETAILED SNP MAPS CORRESPONDING TO FIGURES 4.10 TO 4.13 IN CHAPTER 4

The following pages contain individual SNP maps for each gene corresponding to Figure 4.10A; single nucleotide polymorphism patterning in single locus variants (blue), putative ecotype variants (red) or both (purple) surrounding DV-ST1 in PE A7 and clonal complex A-III defined by ecotype simulation and eBURST analysis of 7 loci in *Synechococcus* A-like BACs. The SNP maps are in the following order, *rbsK, PK, hisF, lepB, CHP, aroA* and *dnaG.*

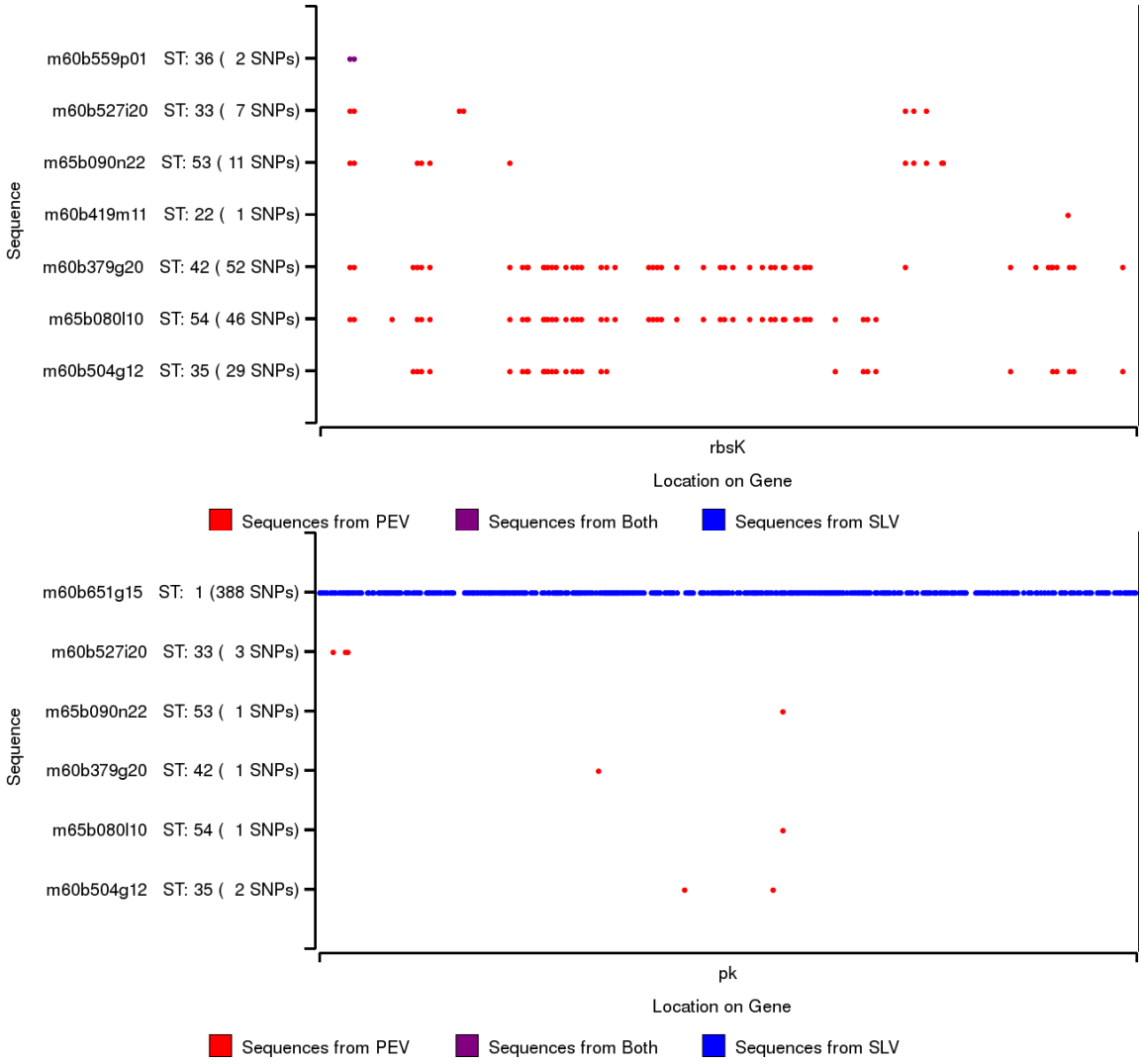


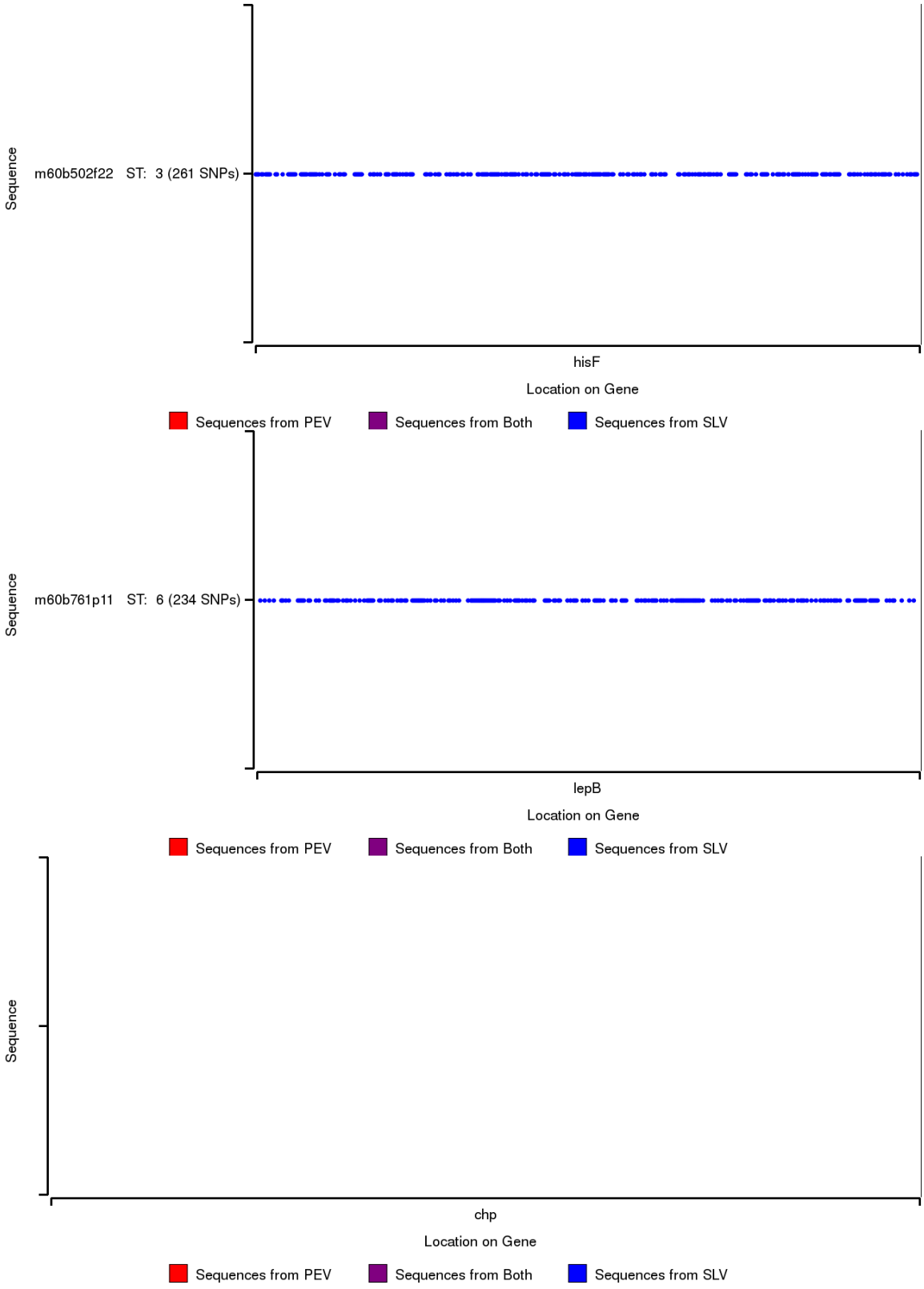

Sequence
hisF
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV
Sequence
m60b740f14 ST: 10 (234 SNPs)
m60b065p14 ST: 8 (233 SNPs)
m60b349m23 ST: 40 ( 1 SNPs)
lepB
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV

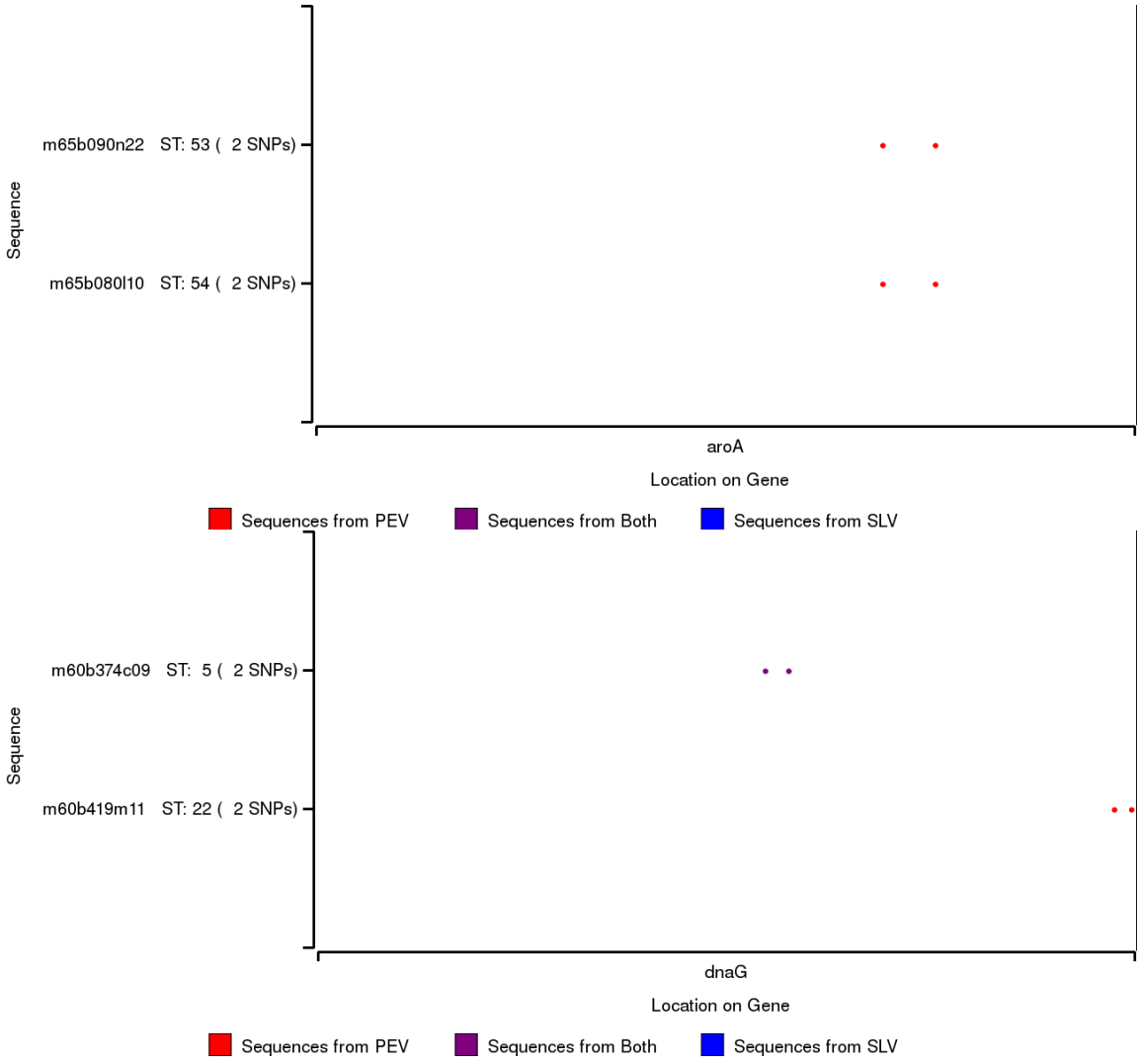

Sequence
m60b755a11 ST: 11 ( 20 SNPs)
chp
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV
Sequence
m60b349m23 ST: 40 ( 2 SNPs)
m65b115i21 ST: 26 ( 1 SNPs)
aroA
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV


m60b457n03 ST: 9 ( 1 SNPs)
m60b349m23 ST: 40 ( 1 SNPs)
Sequence
dnaG
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV

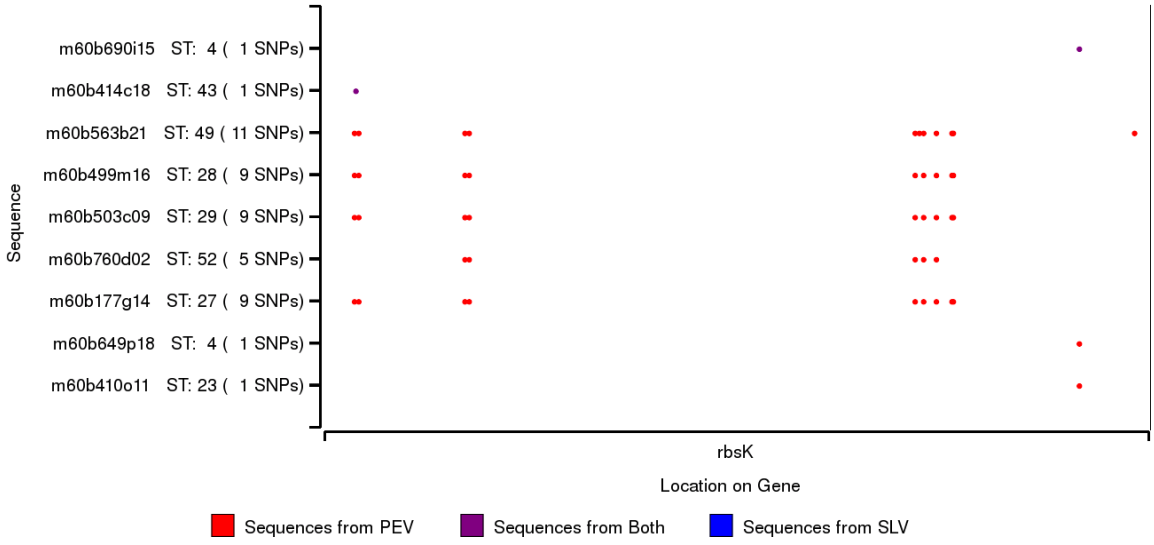
The following pages contain individual SNP maps for each gene corresponding to Figure 4.10B; Single nucleotide polymorphism patterning in single locus variants (blue), putative ecotype variants (red) or both (purple) surrounding DV-ST2 in PE A3 and clonal complex A-II defined by ecotype simulation and eBURST analysis of 7 loci in *Synechococcus* A-like BACs. The SNP maps are in the following order, *rbsK, PK, hisF, lepB, CHP, aroA* and *dnaG.*

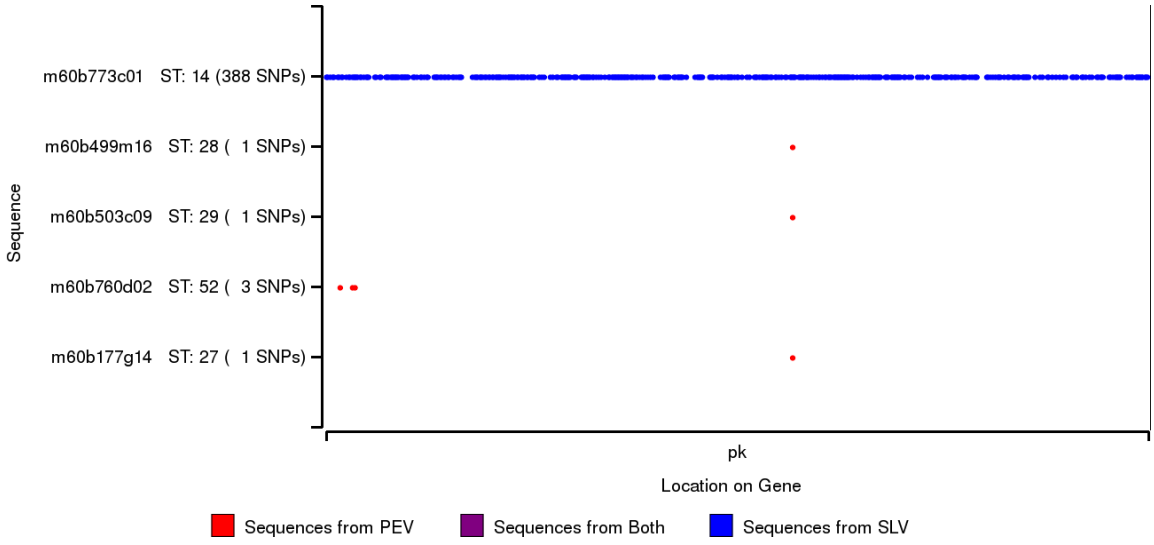


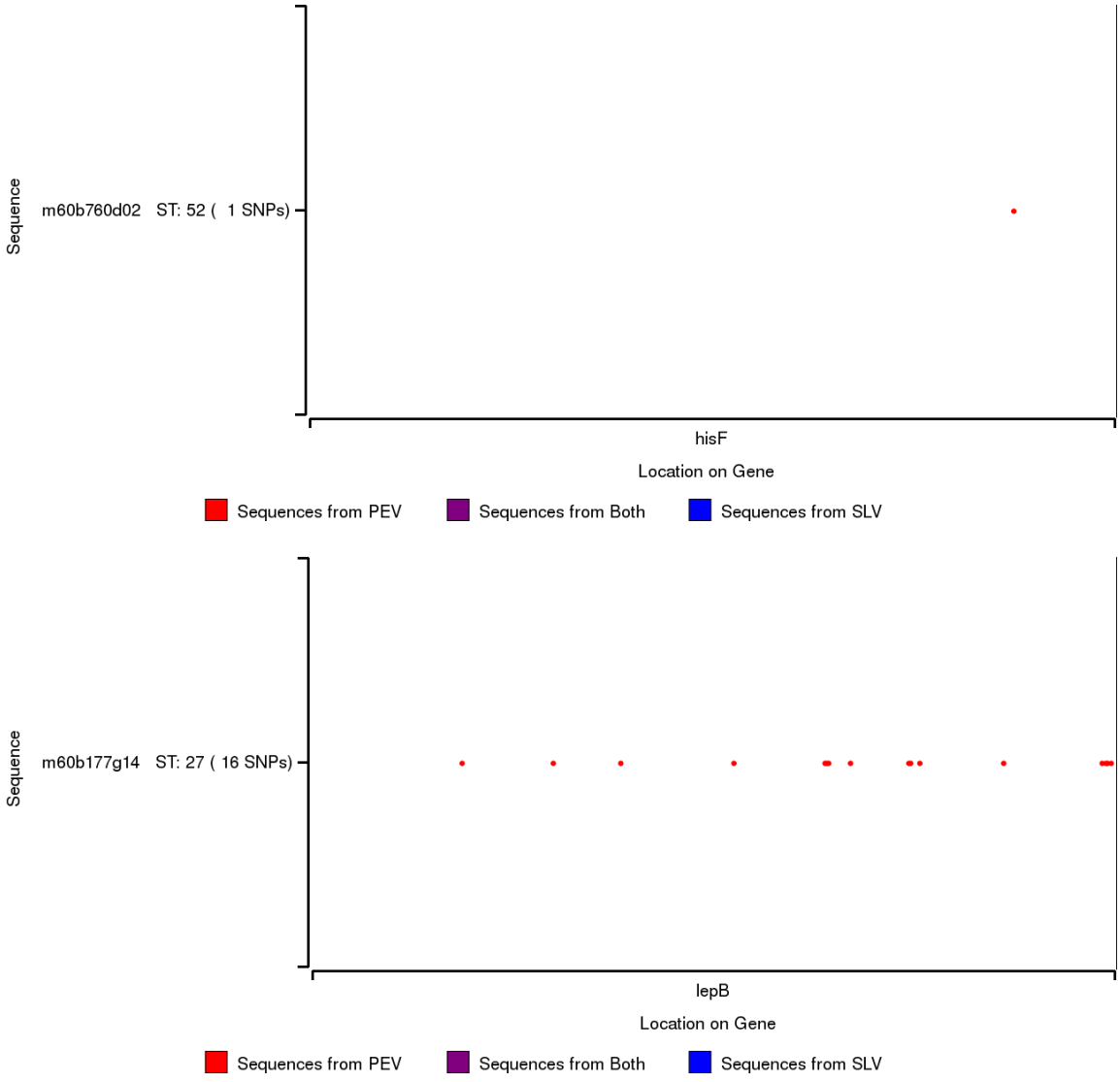

Sequence
m60b502f22 ST: 3 (261 SNPs)
hisF
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV
Sequence
m60b761p11 ST: 6 (234 SNPs)
lepB
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV
Sequence
chp
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV

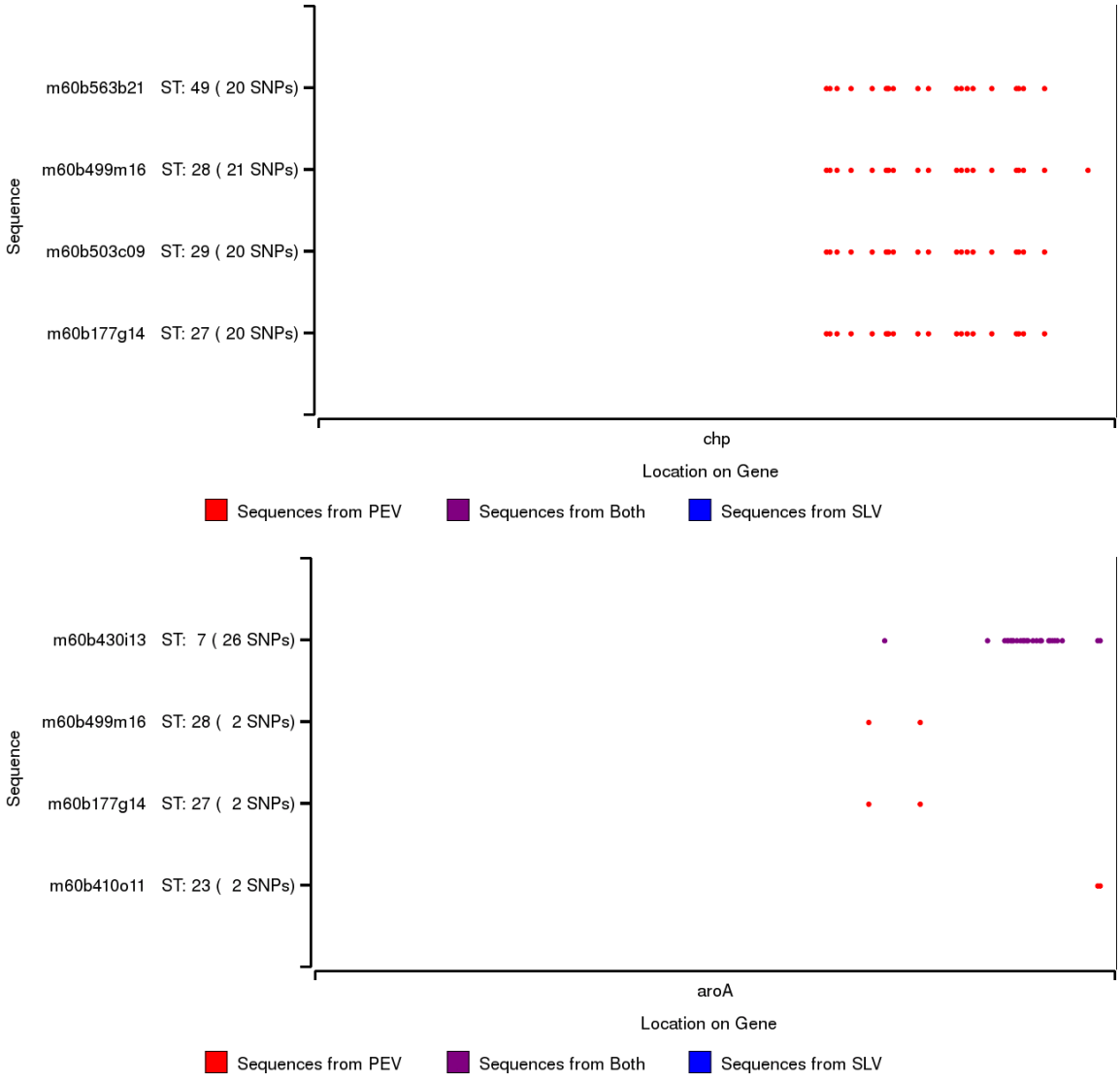

m65b090n22 ST: 53 ( 2 SNPs)
m65b080l10 ST: 54 ( 2 SNPs)
Sequence
aroA
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV
m60b374c09 ST: 5 ( 2 SNPs)
m60b419m11 ST: 22 ( 2 SNPs)
Sequence
dnaG
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV

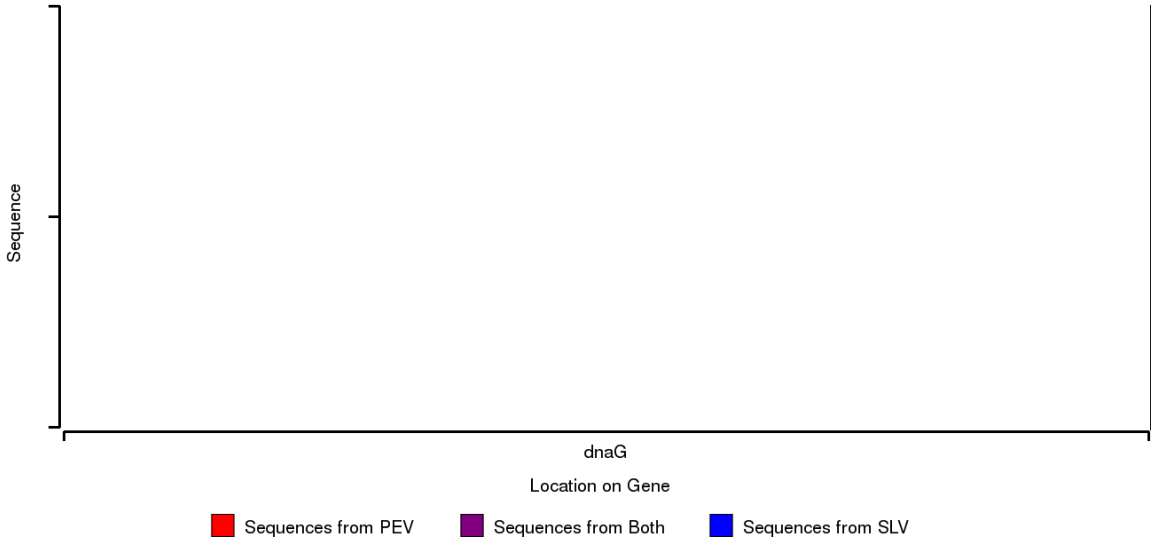
The following pages contain individual SNP maps for each gene corresponding to Figure 4.10C; Single nucleotide polymorphism patterning in single locus variants (blue), putative ecotype variants (red) or both (purple) surrounding DV-ST3 in PE A1 and clonal complex A-I defined by ecotype simulation and eBURST analysis of 7 loci in *Synechococcus* A-like BACs. The SNP maps are in the following order, *rbsK, PK, hisF, lepB, CHP, aroA* and *dnaG.*



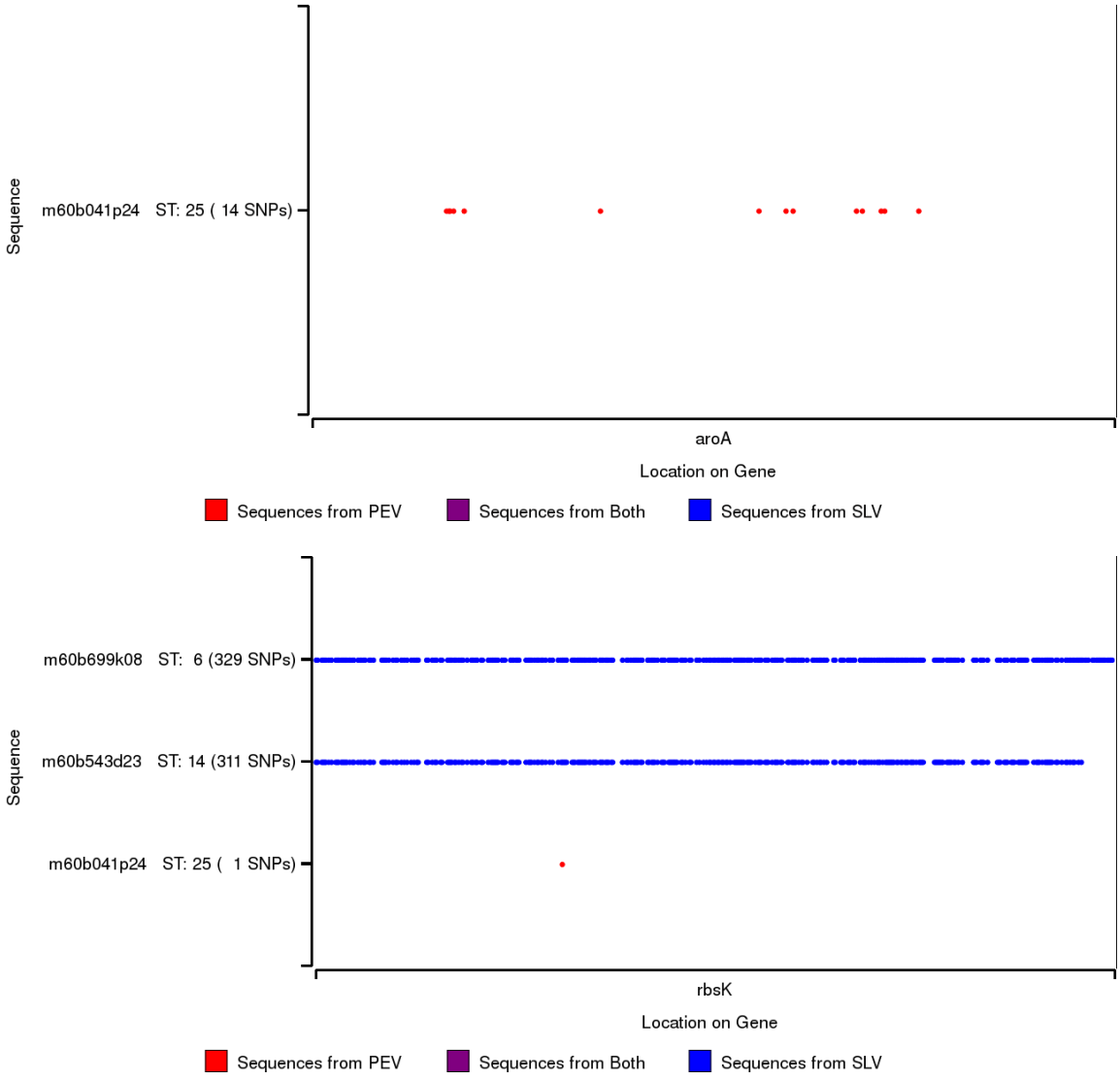



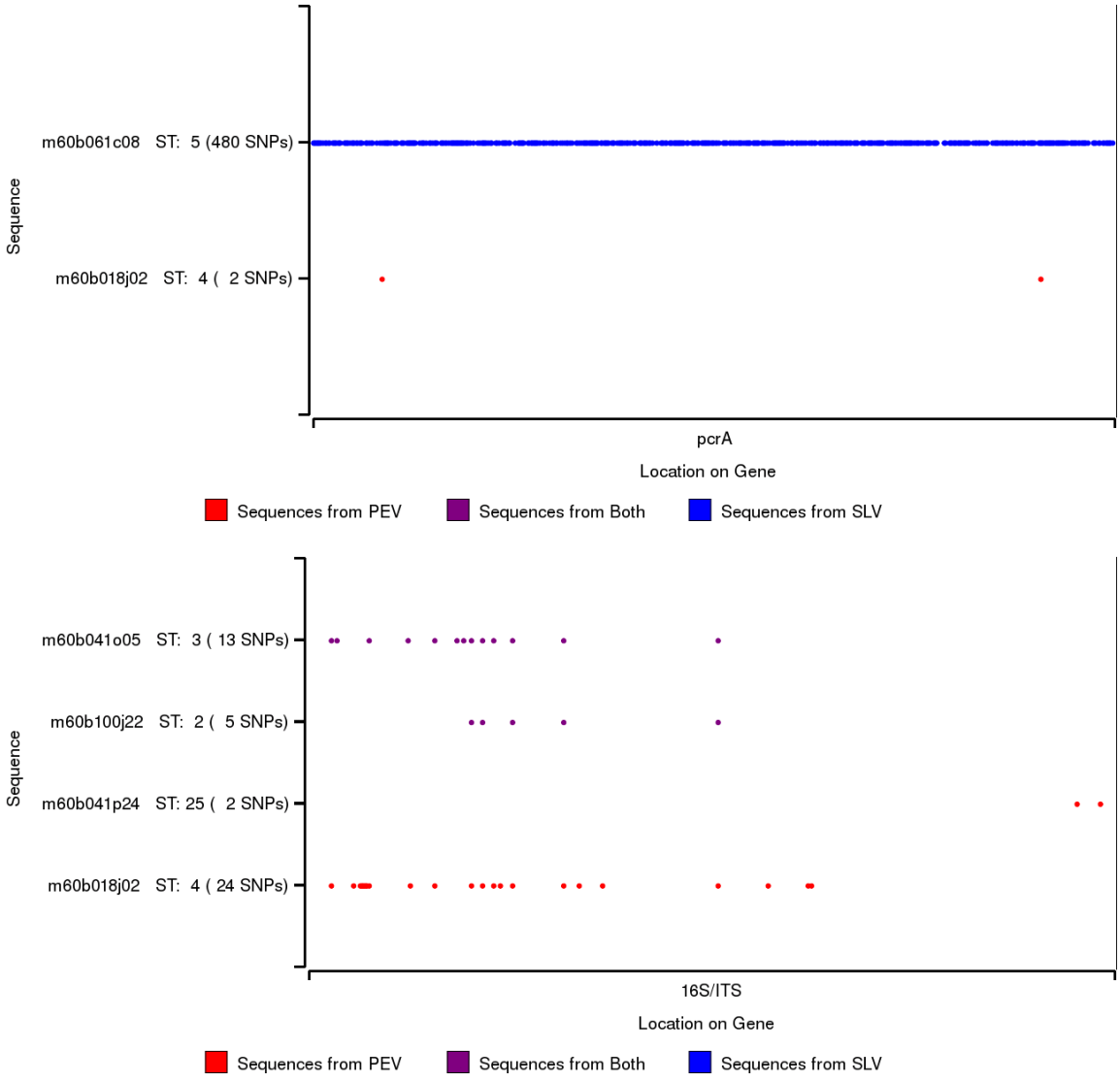

Sequence
m60b760d02 ST: 52 ( 1 SNPs)
hisF
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV
Sequence
m60b177g14 ST: 27 ( 16 SNPs)
lepB
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV


m60b563b21 ST: 49 ( 20 SNPs)
m60b499m16 ST: 28 ( 21 SNPs)
m60b503c09 ST: 29 ( 20 SNPs)
m60b177g14 ST: 27 ( 20 SNPs)
Sequence
chp
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV
m60b430i13 ST: 7 ( 26 SNPs)
m60b499m16 ST: 28 ( 2 SNPs)
m60b177g14 ST: 27 ( 2 SNPs)
m60b410o11 ST: 23 ( 2 SNPs)
Sequence
aroA
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV

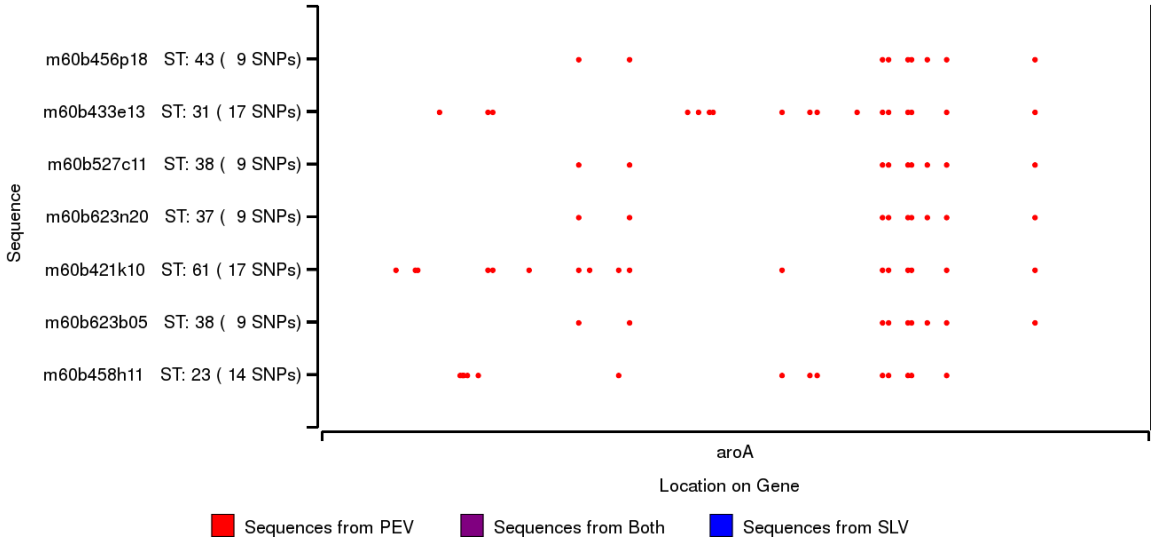

Sequence
dnaG
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV

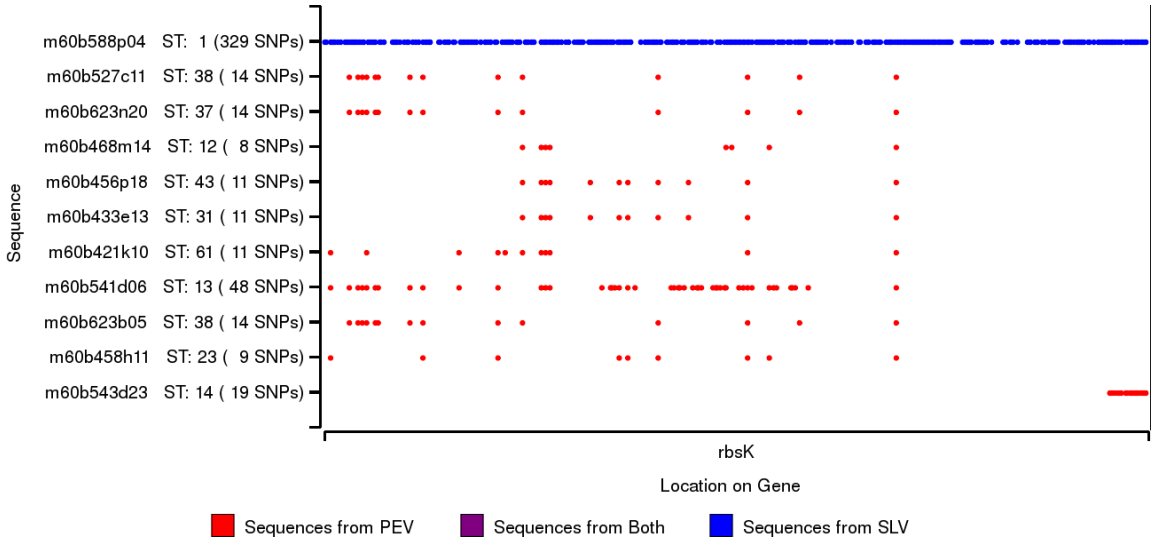
The following pages contain individual SNP maps for each gene corresponding to Figure 4.11A; Single nucleotide polymorphism patterning in single locus variants (blue), putative ecotype variants (red) or both (purple) surrounding DV-ST 1 in PE B´1 and clonal complex B´-I defined by ecotype simulation and eBURST in 4-locus analysis of *Synechococcus* B´-like BACs. The SNP maps are in the following order; *aroA, rbsK, pcrA* and the 16S rRNA/ITS region.

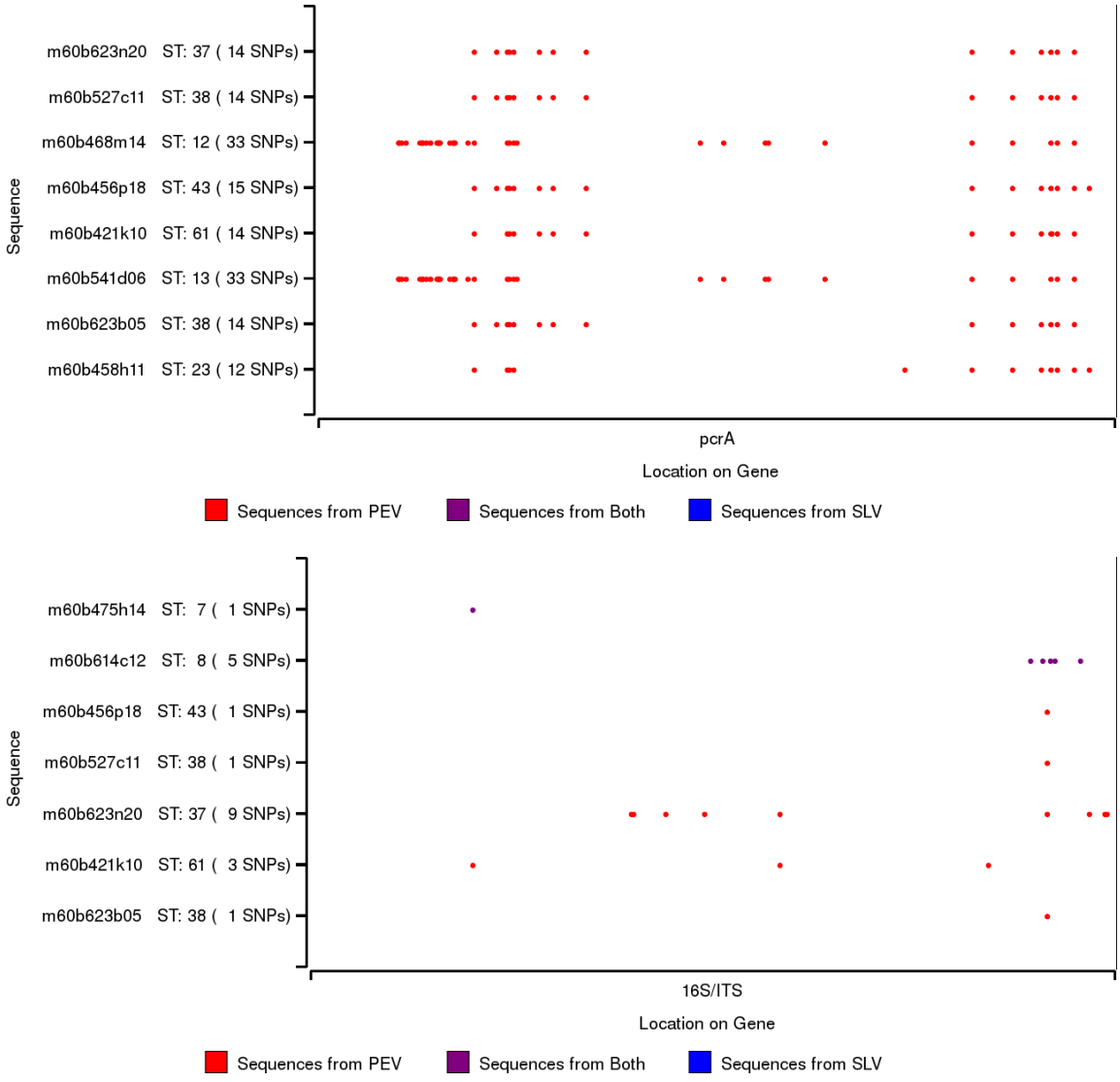



m60b061c08 ST: 5 (480 SNPs)
m60b018j02 ST: 4 ( 2 SNPs)
Sequence
pcrA
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV
m60b041o05 ST: 3 ( 13 SNPs)
m60b100j22 ST: 2 ( 5 SNPs)
m60b041p24 ST: 25 ( 2 SNPs)
m60b018j02 ST: 4 ( 24 SNPs)
Sequence
16S/ITS
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV

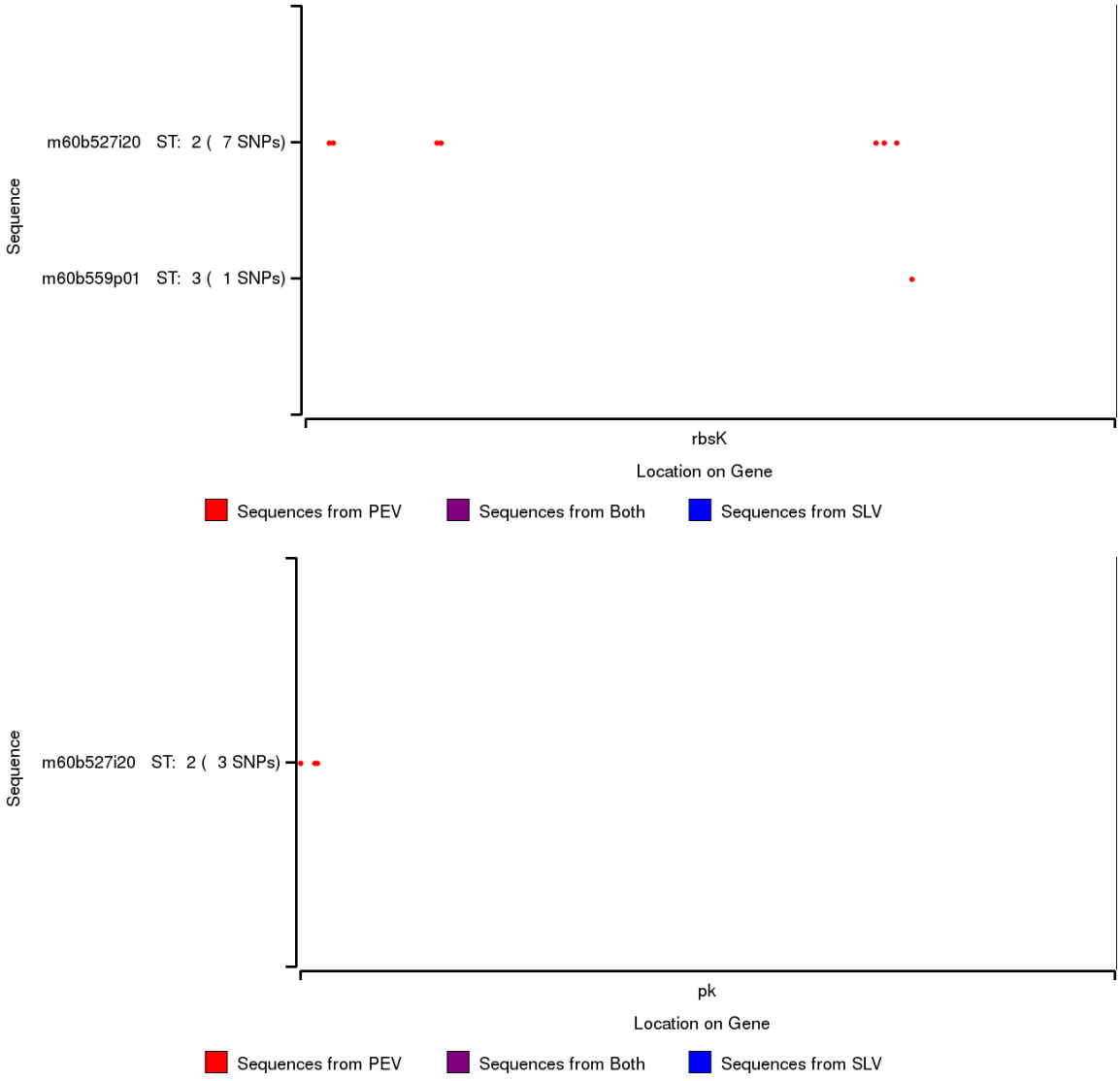
The following pages contain individual SNP maps for each gene corresponding to Figure 4.11B; Single nucleotide polymorphism patterning in single locus variants (blue), putative ecotype variants (red) or both (purple) surrounding DV-ST6 in PE B´6 and clonal complex B´-II defined by ecotype simulation and eBURST in 4-locus analysis of *Synechococcus* B´-like BACs. The SNP maps are in the following order; *aroA, rbsK, pcrA* and the 16S rRNA/ITS region.

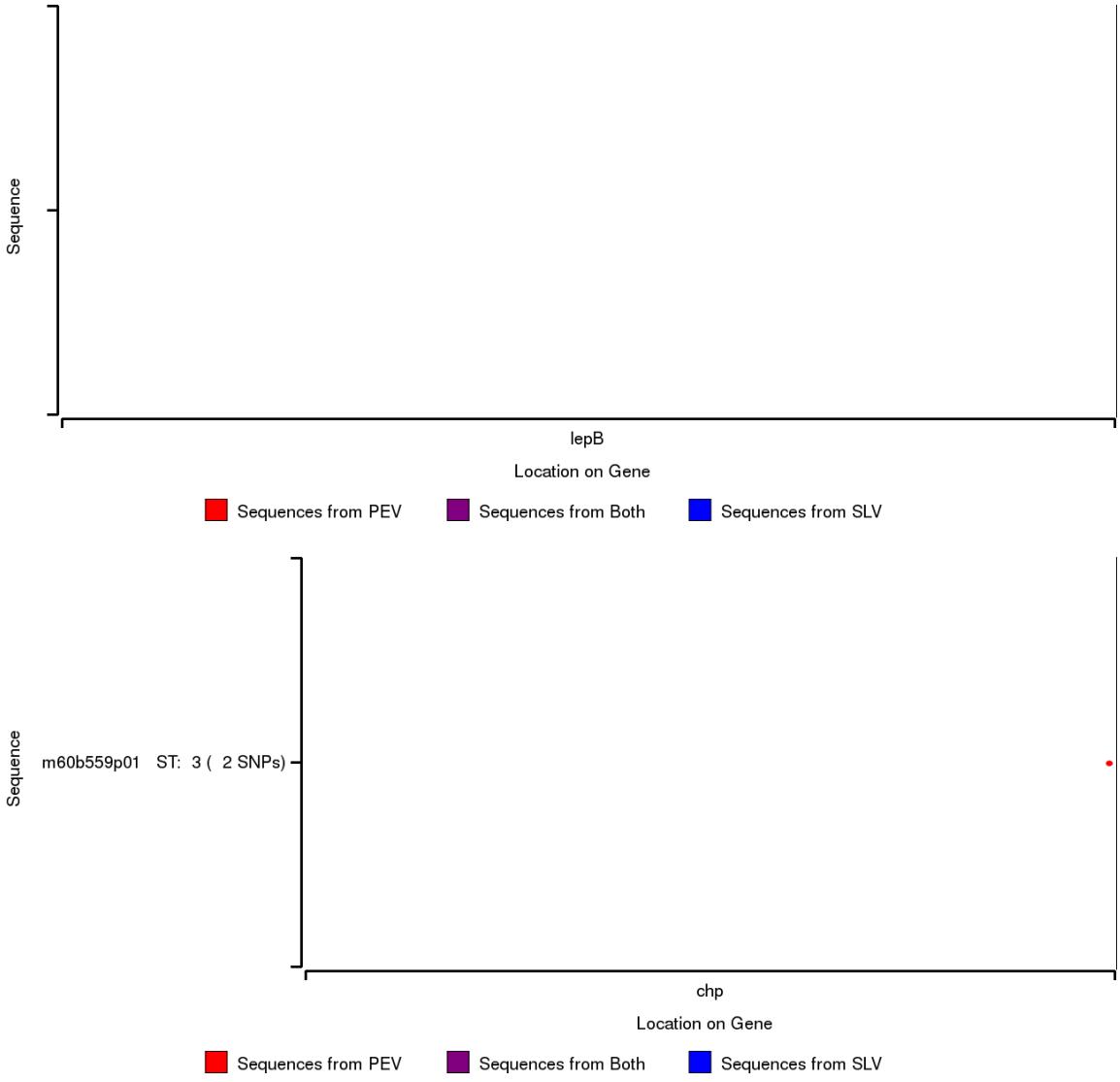


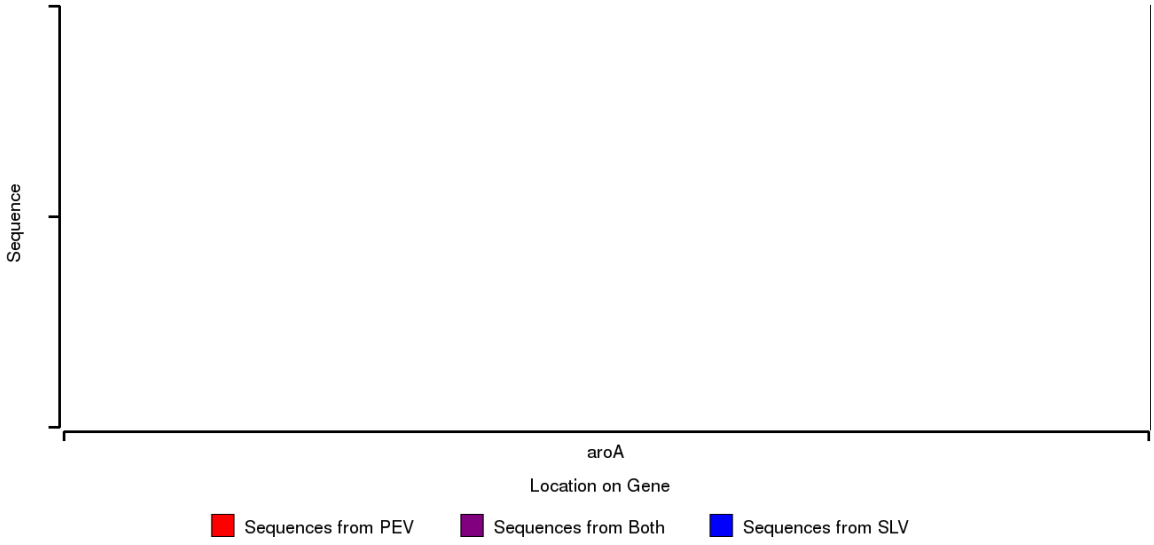




m60b623n20 ST: 37 ( 14 SNPs)
m60b527c11 ST: 38 ( 14 SNPs)
m60b468m14 ST: 12 ( 33 SNPs)
m60b456p18 ST: 43 ( 15 SNPs)
m60b421k10 ST: 61 ( 14 SNPs)
m60b541d06 ST: 13 ( 33 SNPs)
m60b623b05 ST: 38 ( 14 SNPs)
m60b458h11 ST: 23 ( 12 SNPs)
Sequence
pcrA
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV
m60b475h14 ST: 7 ( 1 SNPs)
m60b614c12 ST: 8 ( 5 SNPs)
m60b456p18 ST: 43 ( 1 SNPs)
m60b527c11 ST: 38 ( 1 SNPs)
m60b623n20 ST: 37 ( 9 SNPs)
m60b421k10 ST: 61 ( 3 SNPs)
m60b623b05 ST: 38 ( 1 SNPs)
Sequence
16S/ITS
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV

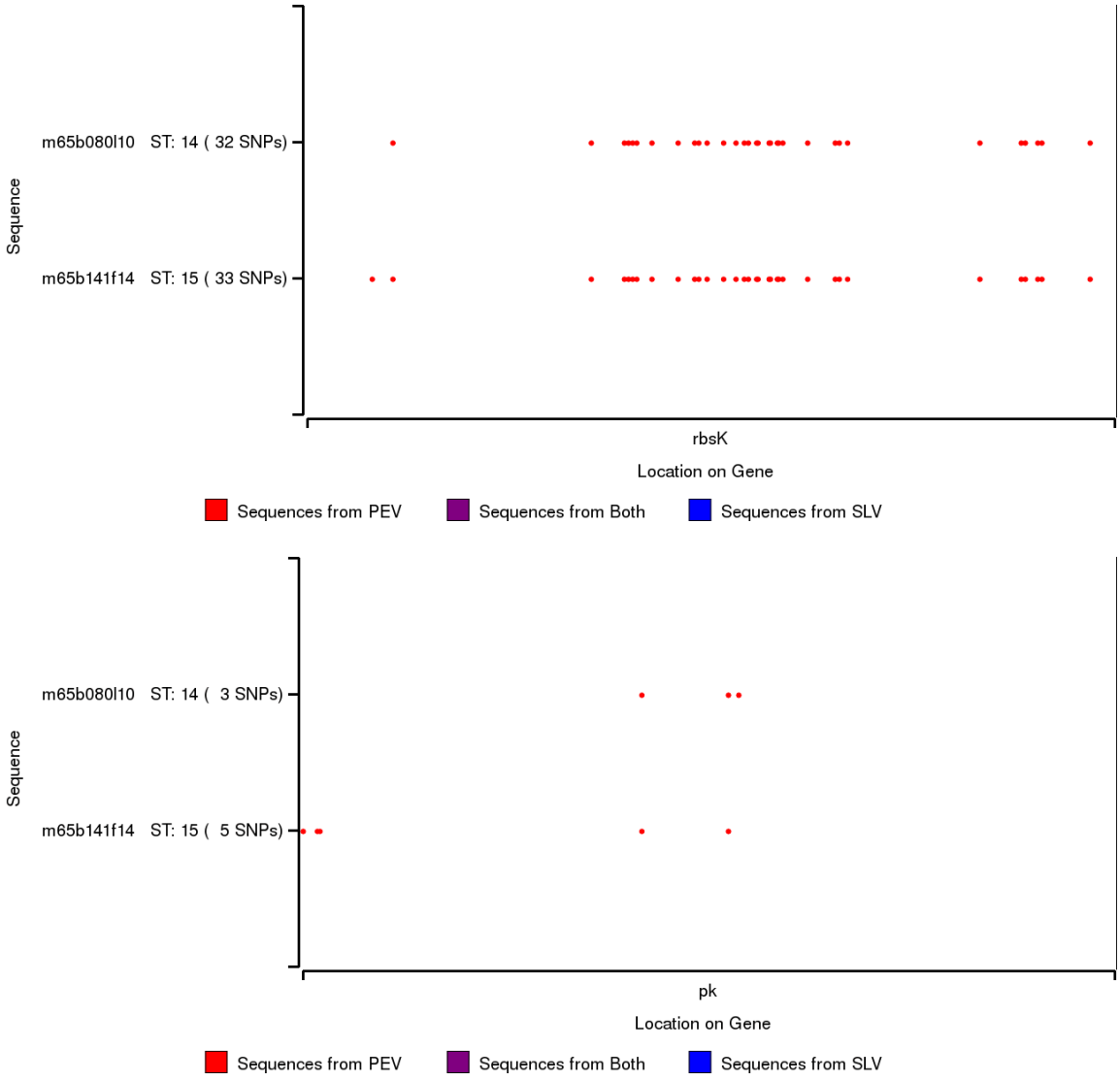
The following pages contain individual SNP maps for each gene corresponding to Figure 4.12A; Single nucleotide polymorphism patterning in single locus variants (blue), putative ecotype variants (red) or both (purple) surrounding DV-ST1 in PE A5-1 defined by ecotype simulation in 5-locus analysis of *Synechococcus* A-like BACs. The SNP maps are in the following order; *rbsK, PK, lepB, CHP* and *aroA*.

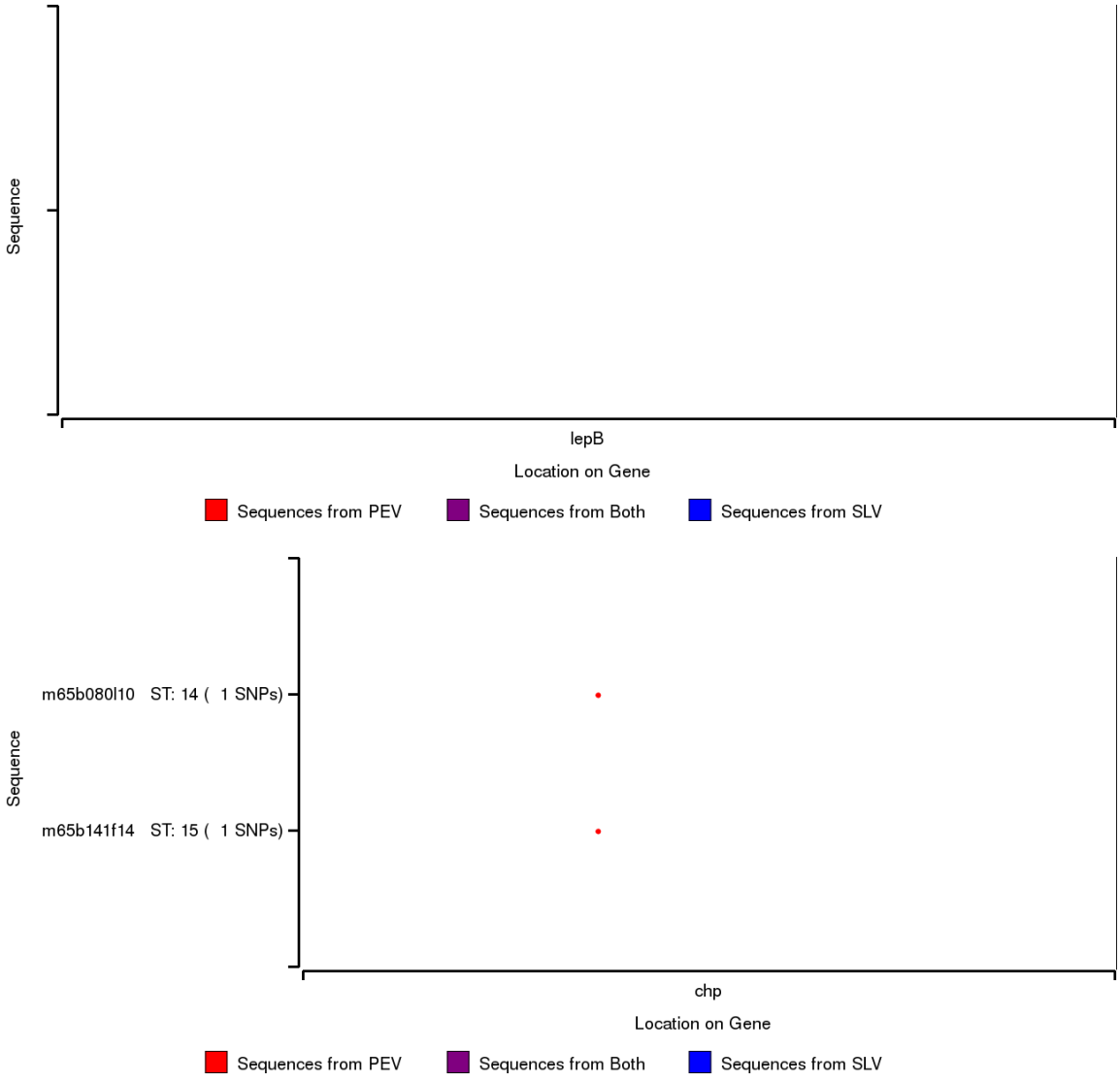


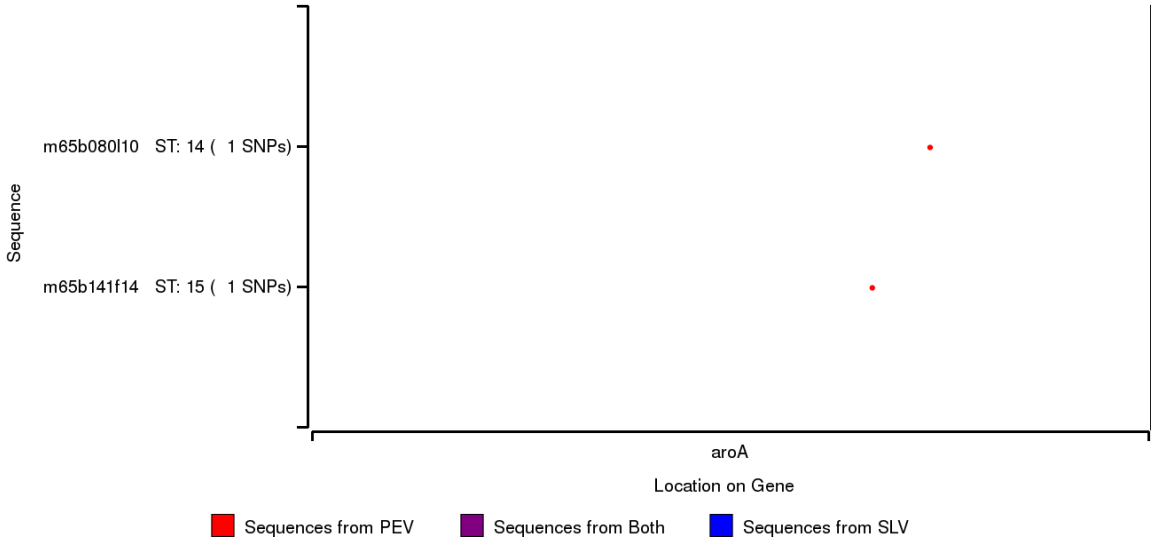

Sequence
lepB
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV
Sequence
m60b559p01 ST: 3 ( 2 SNPs)
chp
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV


Sequence
aroA
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV

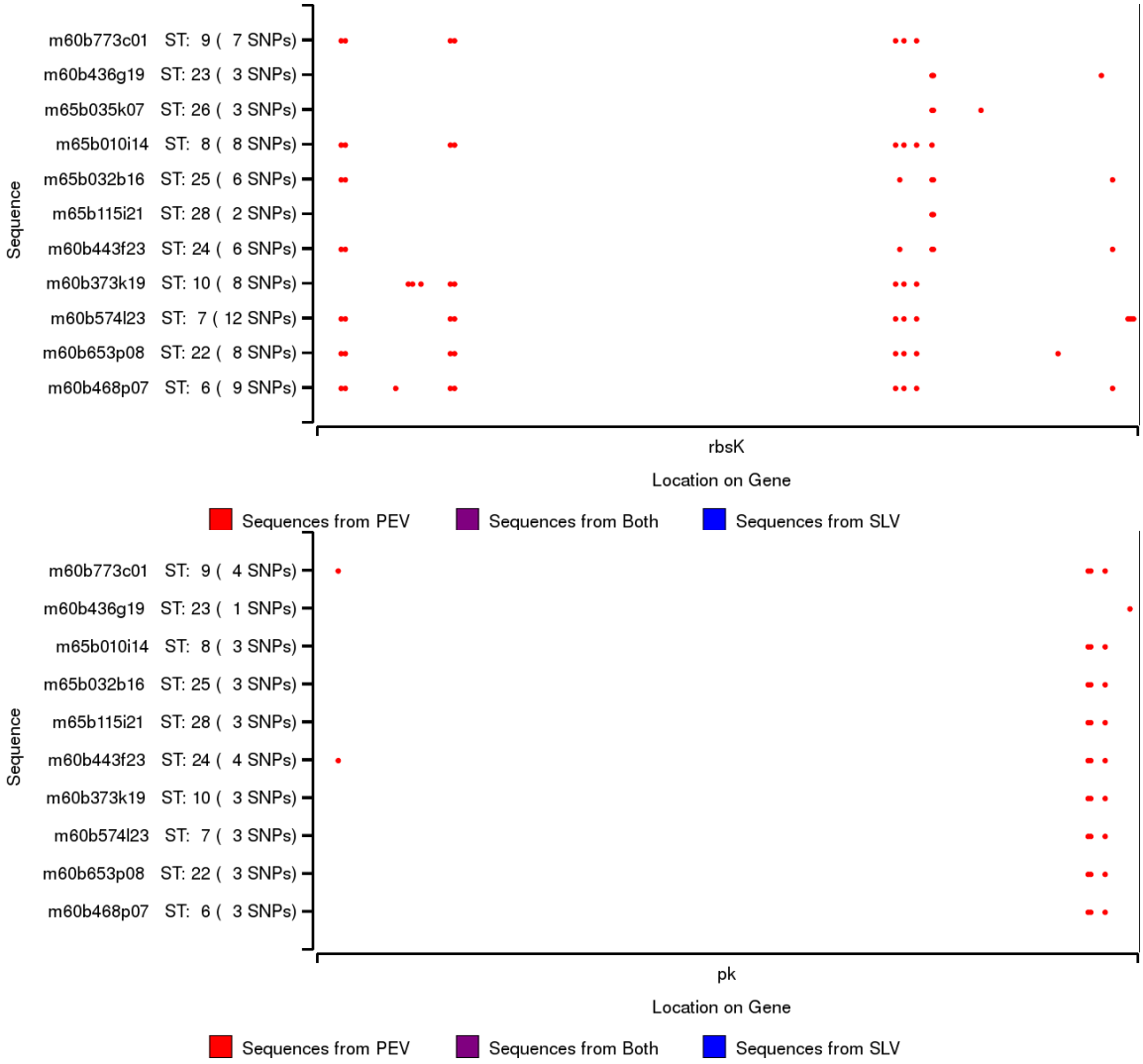
The following pages contain individual SNP maps for each gene corresponding to Figure 4.12B; Single nucleotide polymorphism patterning in single locus variants (blue), putative ecotype variants (red) or both (purple) surrounding DV-ST4 in PE A5-2 defined by ecotype simulation in 5-locus analysis of *Synechococcus* A-like BACs. The SNP maps are in the following order; *rbsK, PK, lepB, CHP* and *aroA*.

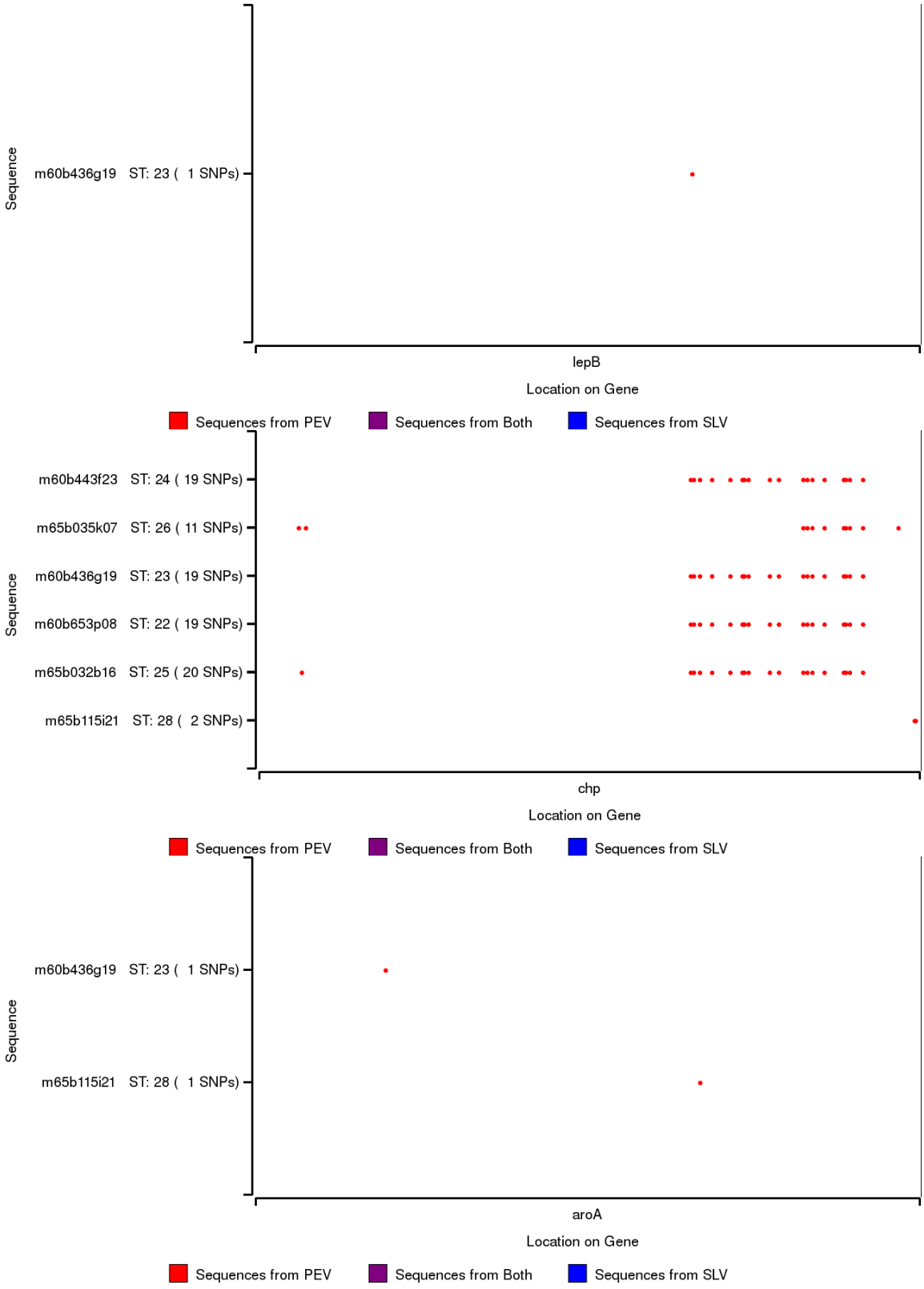



Sequence
lepB
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV
Sequence
m65b080l10 ST: 14 ( 1 SNPs)
m65b141f14 ST: 15 ( 1 SNPs)
chp
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV

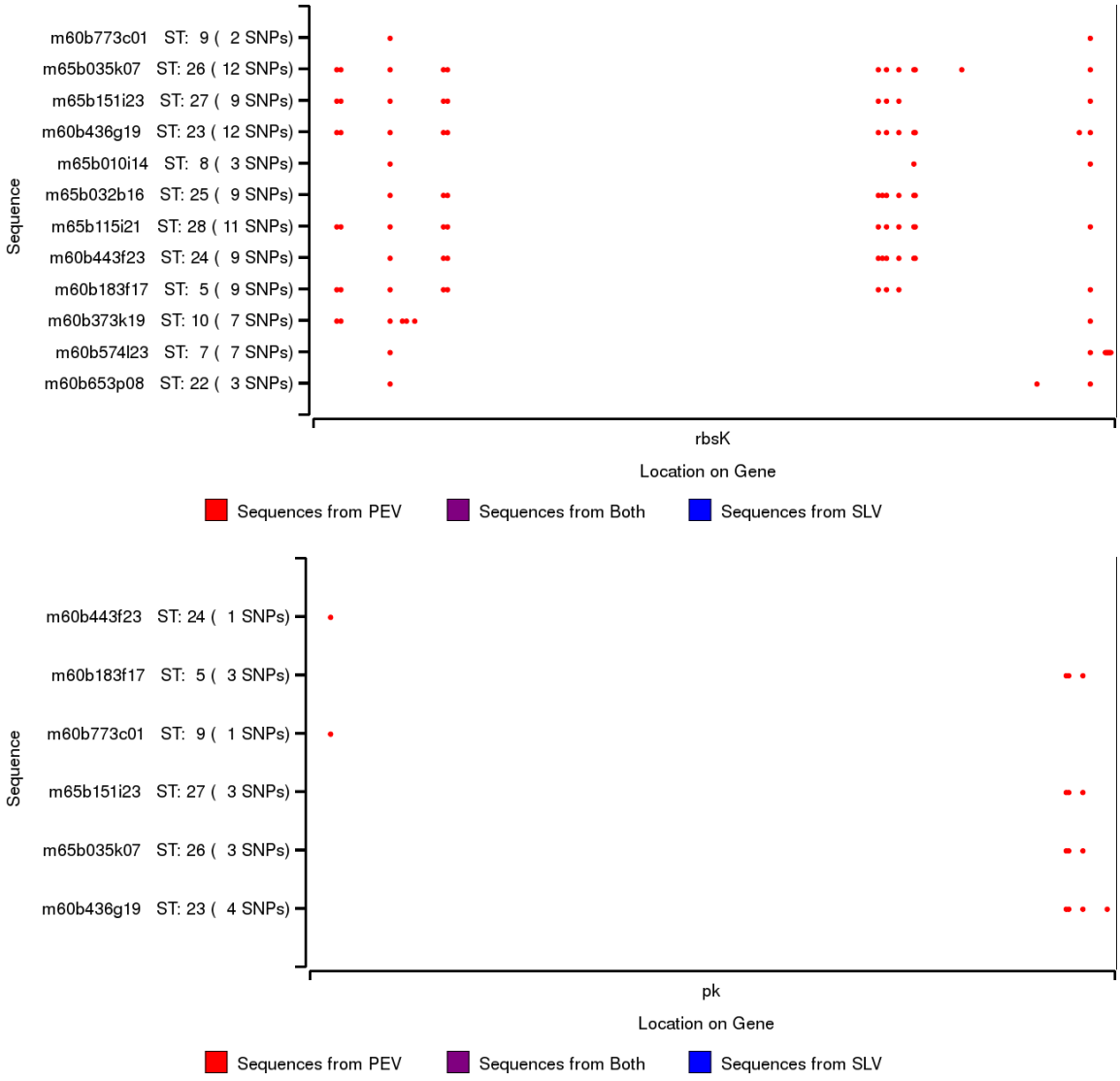

m65b080l10 ST: 14 ( 1 SNPs)
m65b141f14 ST: 15 ( 1 SNPs)
Sequence
aroA
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV

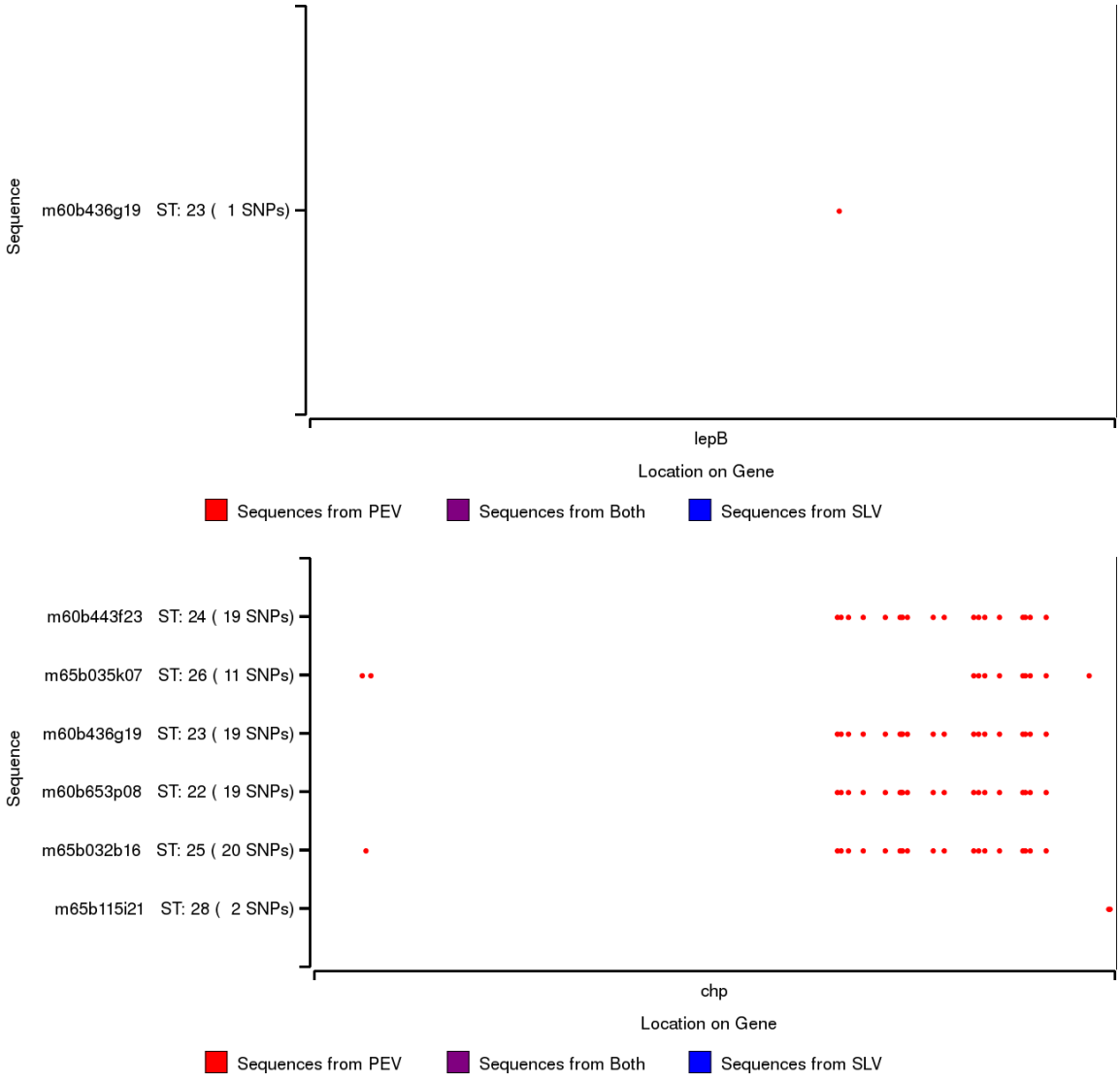
The following pages contain individual SNP maps for each gene corresponding to Figure 4.12C; Single nucleotide polymorphism patterning in single locus variants (blue), putative ecotype variants (red) or both (purple) surrounding DV-ST5 in PE A5-5 defined by ecotype simulation in 5-locus analysis of *Synechococcus* A-like BACs. The SNP maps are in the following order; *rbsK, PK, lepB, CHP* and *aroA*.

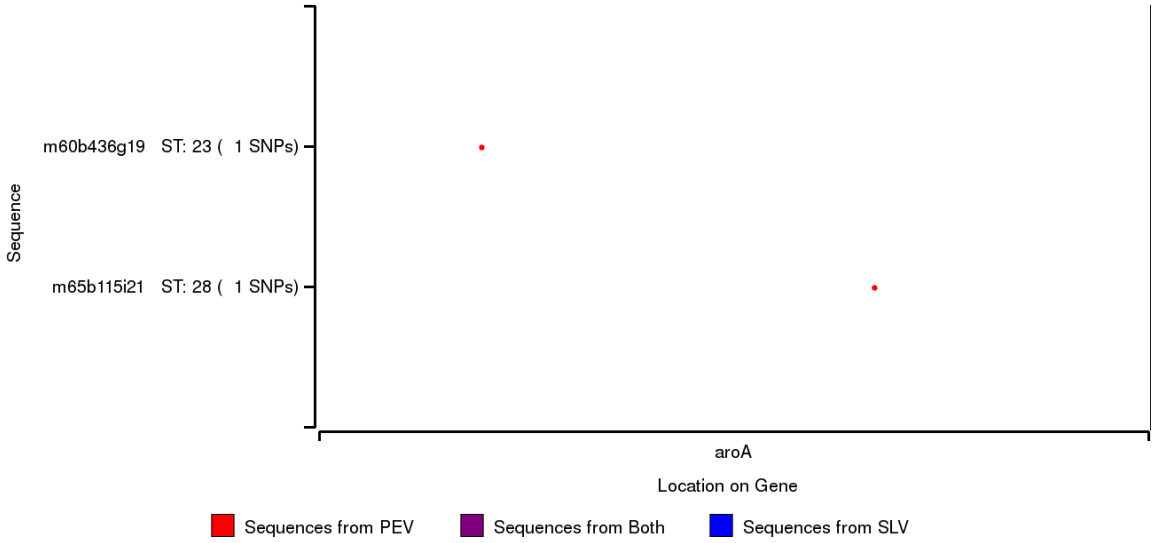



Sequence
m60b436g19 ST: 23 ( 1 SNPs)
lepB
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV
m60b443f23 ST: 24 ( 19 SNPs)
m65b035k07 ST: 26 ( 11 SNPs)
m60b436g19 ST: 23 ( 19 SNPs)
m60b653p08 ST: 22 ( 19 SNPs)
m65b032b16 ST: 25 ( 20 SNPs)
m65b115i21 ST: 28 ( 2 SNPs)
chp
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV
m60b436g19 ST: 23 ( 1 SNPs)
m65b115i21 ST: 28 ( 1 SNPs)
aroA
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV

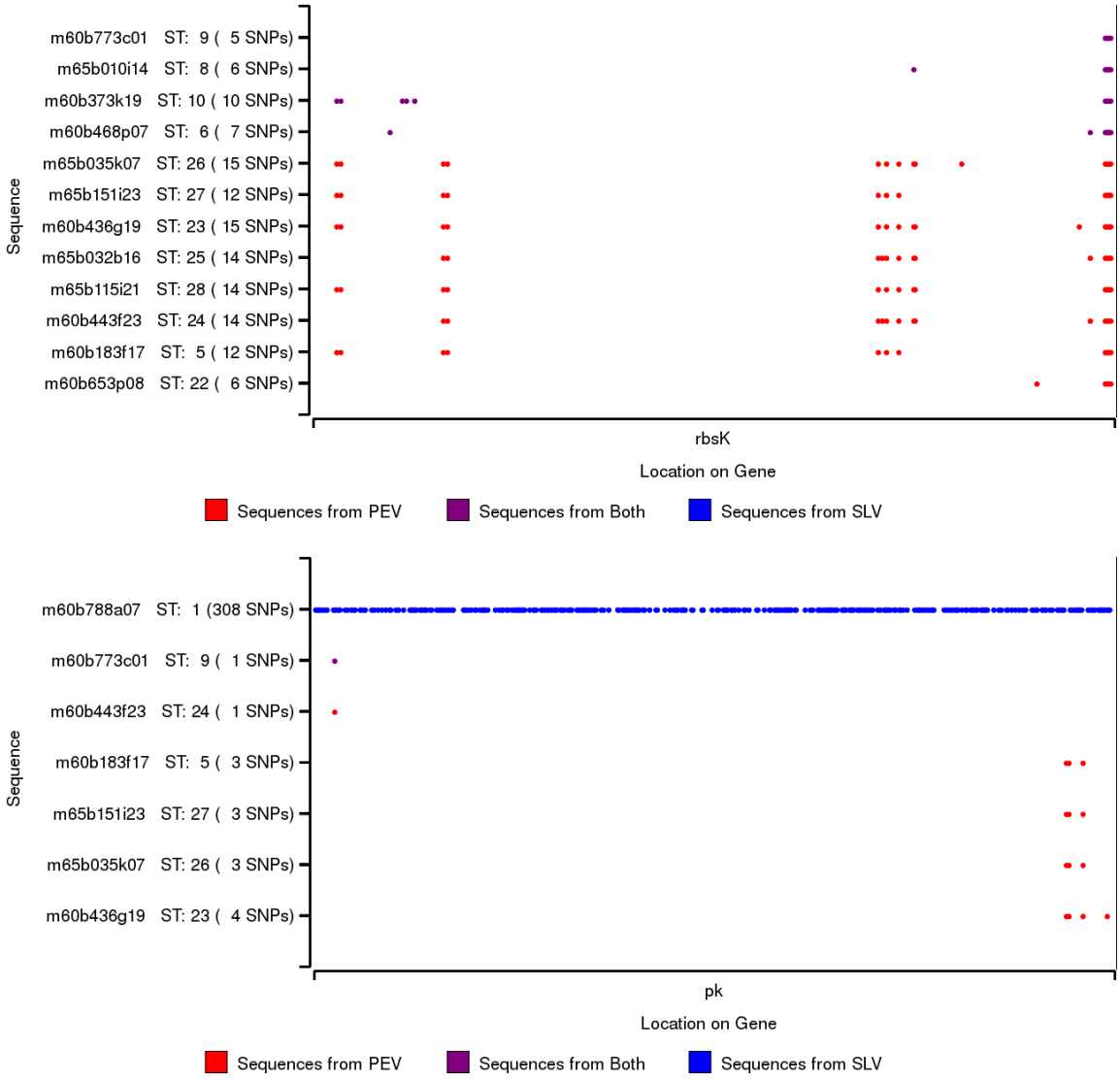
The following pages contain individual SNP maps for each gene corresponding to Figure 4.12D; Single nucleotide polymorphism patterning in single locus variants (blue), putative ecotype variants (red) or both (purple) surrounding sDV-ST6 in PE A5-5 defined by ecotype simulation in 5-locus analysis of *Synechococcus* A-like BACs. The SNP maps are in the following order; *rbsK, PK, lepB, CHP* and *aroA*.

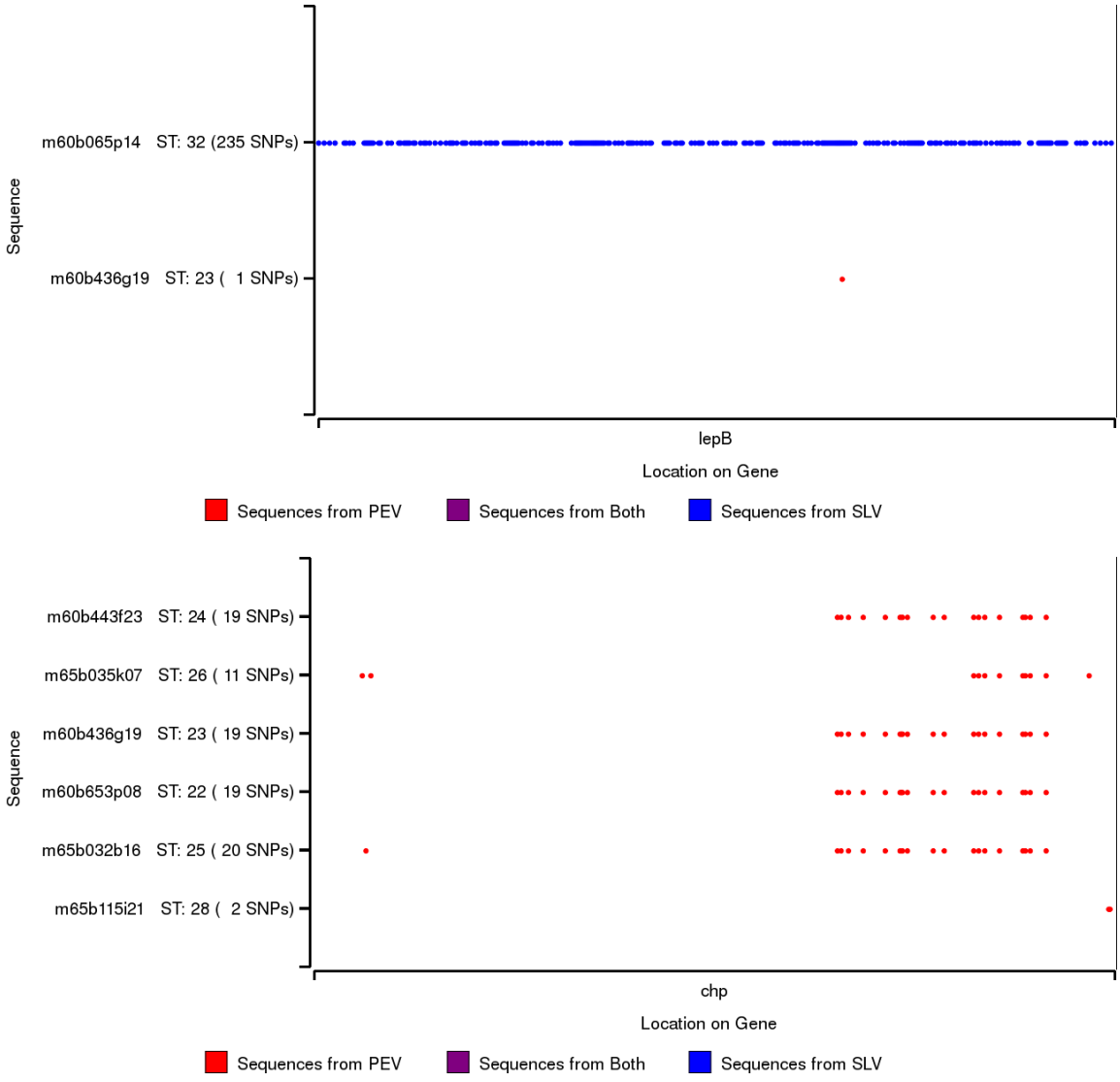


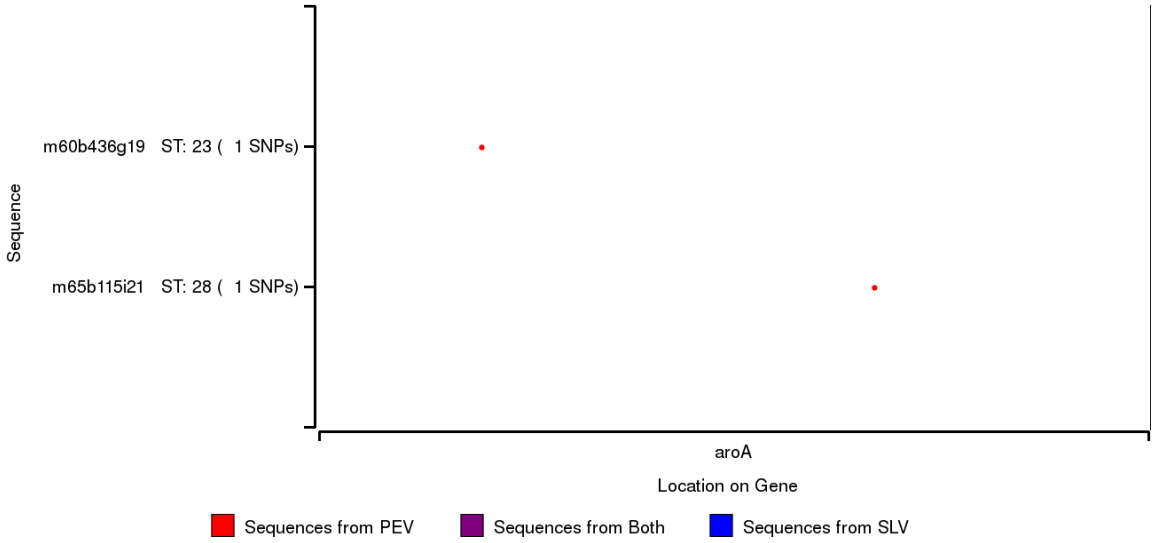

Sequence
m60b436g19 ST: 23 ( 1 SNPs)
lepB
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV
Sequence
m60b443f23 ST: 24 ( 19 SNPs)
m65b035k07 ST: 26 ( 11 SNPs)
m60b436g19 ST: 23 ( 19 SNPs)
m60b653p08 ST: 22 ( 19 SNPs)
m65b032b16 ST: 25 ( 20 SNPs)
m65b115i21 ST: 28 ( 2 SNPs)
chp
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV


m60b436g19 ST: 23 ( 1 SNPs)
m65b115i21 ST: 28 ( 1 SNPs)
Sequence
aroA
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV

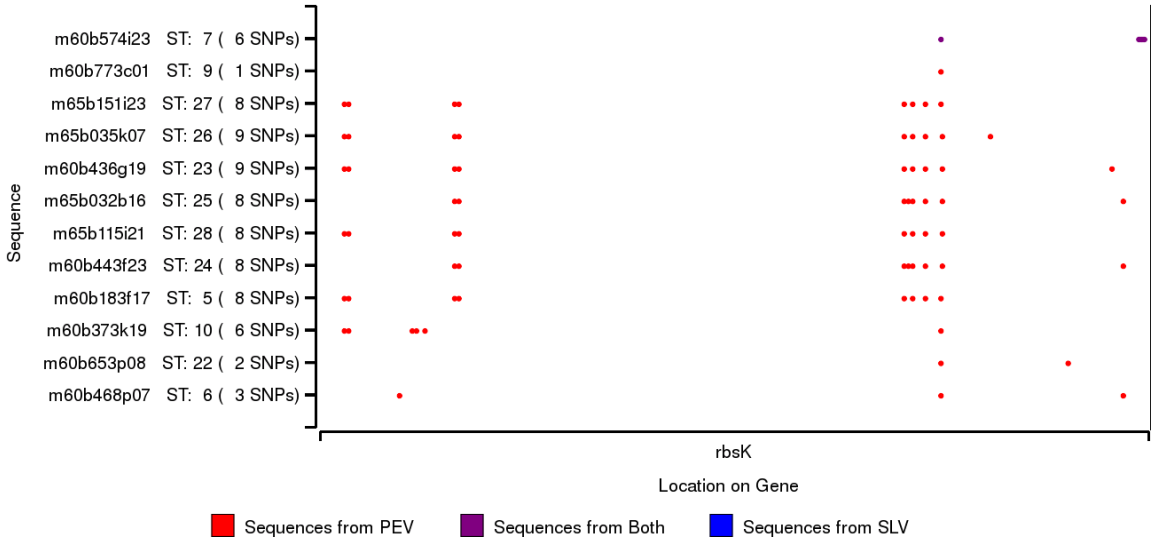
The following pages contain individual SNP maps for each gene corresponding to Figure 4.13A; Single nucleotide polymorphism patterning in single locus variants (blue), putative ecotype variants (red) or both (purple) surrounding sDV-ST7 in PE A5-5 and clonal complex A5-I defined by ecotype simulation in 5-locus analysis of *Synechococcus* A-like BACs. The SNP maps are in the following order; *rbsK, PK, lepB, CHP* and *aroA.*

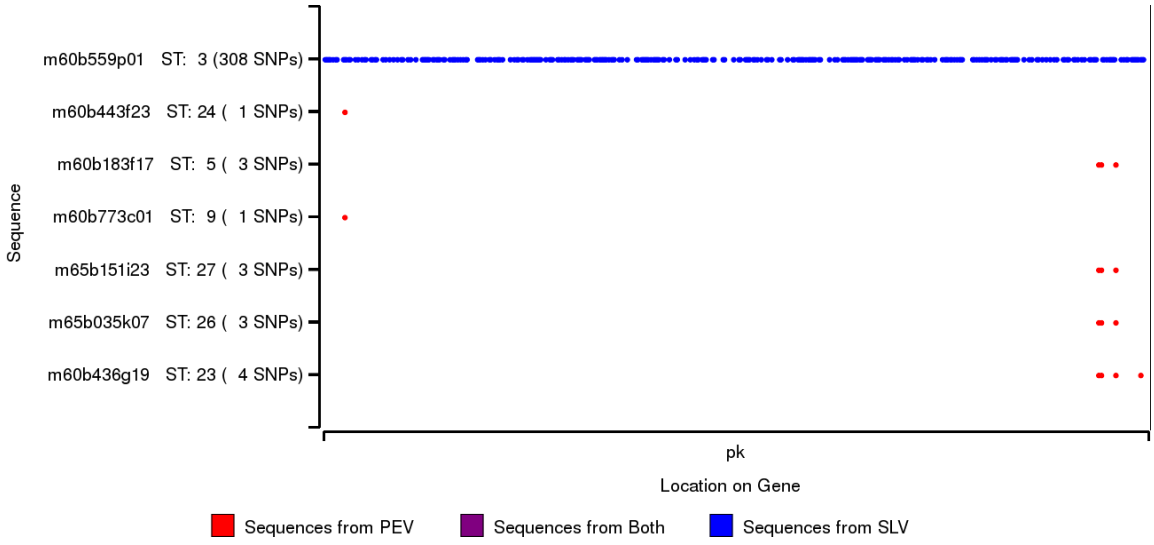


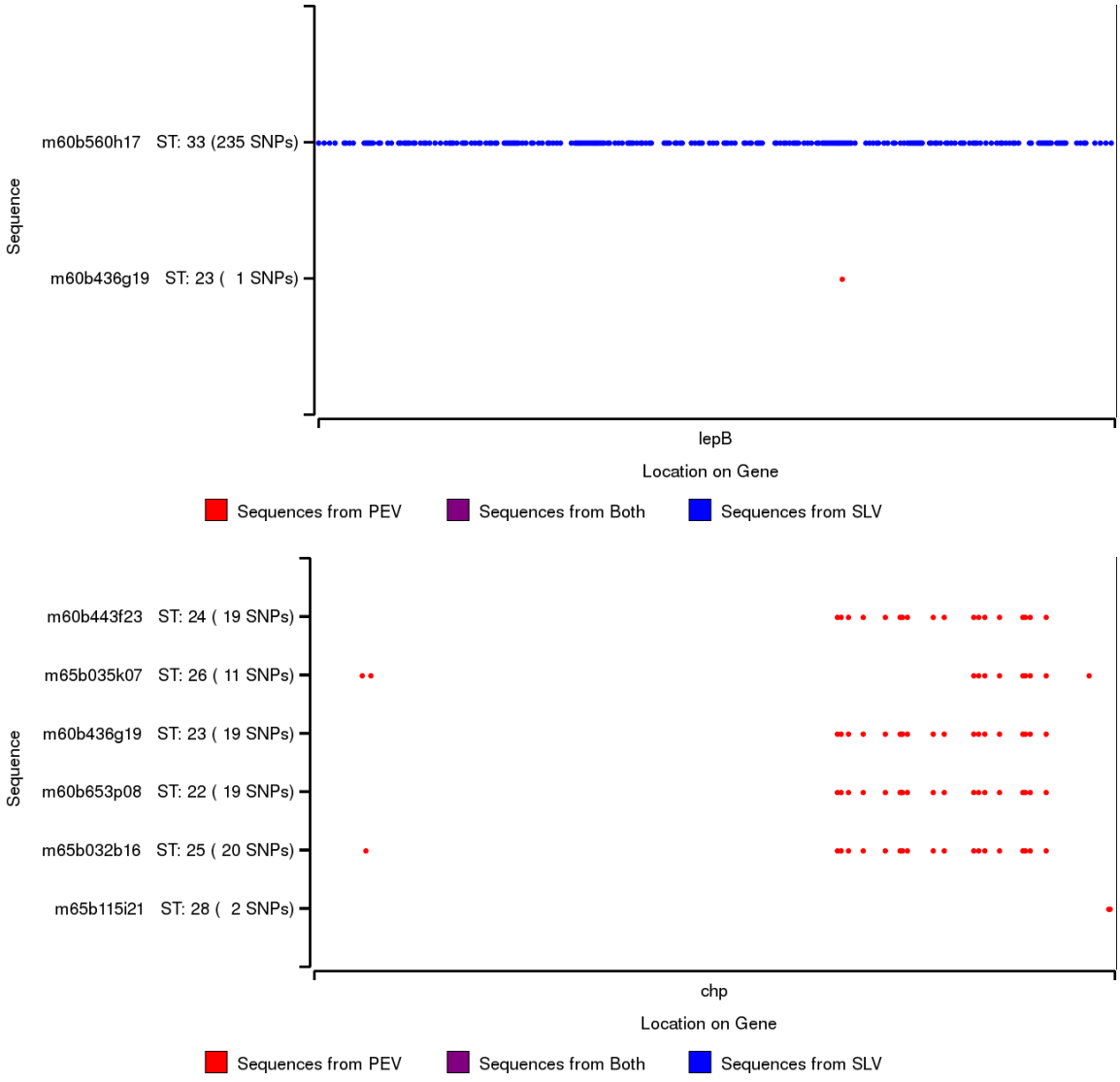

m60b065p14 ST: 32 (235 SNPs)
m60b436g19 ST: 23 ( 1 SNPs)
Sequence
lepB
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV
m60b443f23 ST: 24 ( 19 SNPs)
m65b035k07 ST: 26 ( 11 SNPs)
m60b436g19 ST: 23 ( 19 SNPs)
m60b653p08 ST: 22 ( 19 SNPs)
m65b032b16 ST: 25 ( 20 SNPs)
m65b115i21 ST: 28 ( 2 SNPs)
Sequence
chp
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV

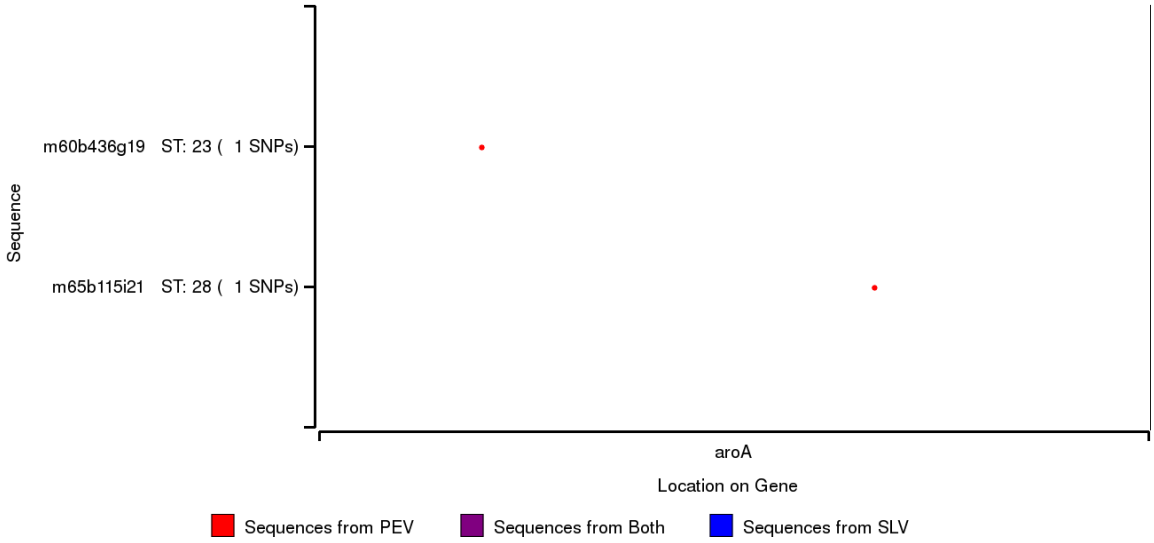

m60b436g19 ST: 23 ( 1 SNPs)
m65b115i21 ST: 28 ( 1 SNPs)
Sequence
aroA
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV

The following pages contain individual SNP maps for each gene corresponding to Figure 4.13B; single nucleotide polymorphism patterning in single locus variants (blue), putative ecotype variants (red) or both (purple) surrounding sDV-ST8 in PE A5-5 and clonal complex A5-II defined by ecotype simulation in 5-locus analysis of *Synechococcus* A-like BACs. The SNP maps are in the following order; *rbsK, PK, lepB, CHP* and *aroA*.




m60b560h17 ST: 33 (235 SNPs)
m60b436g19 ST: 23 ( 1 SNPs)
Sequence
lepB
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV
m60b443f23 ST: 24 ( 19 SNPs)
m65b035k07 ST: 26 ( 11 SNPs)
m60b436g19 ST: 23 ( 19 SNPs)
m60b653p08 ST: 22 ( 19 SNPs)
m65b032b16 ST: 25 ( 20 SNPs)
m65b115i21 ST: 28 ( 2 SNPs)
Sequence
chp
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV


m60b436g19 ST: 23 ( 1 SNPs)
m65b115i21 ST: 28 ( 1 SNPs)
Sequence
aroA
Location on Gene
Sequences from PEV
Sequences from Both
Sequences from SLV

# APPENDIX E

## SUPPLEMENTAL INFORMATION FOR CHAPTER 5: MULTIPLE DISCPLACEMENT AMPLIFICATION OF SINGLE CELL GENOMES FOR CULTIVATION-INDEPENDENT MULTI-LOCUS SEQUENCE ANALYSIS OF POPULATION GENETICS OF *SYNECHOCOCCUS*: A PILOT STUDY

## Correspondence Between Bacterial Artificial Chromosome (BAC) Library Clones and MDAs

To compare how MDA and BACs sample diversity, phylogenetic trees were constructed from MDAs with sufficient sequence data for the various loci combined with a random sample of BAC clones from the M60 BAC library and isolates for *Synechococcus* strain A and strain B´ that were cultured in the lab (highlighted in green; Allewalt et al., 2006) (see Chapter 3) (Figures E5.1 and E5.2).

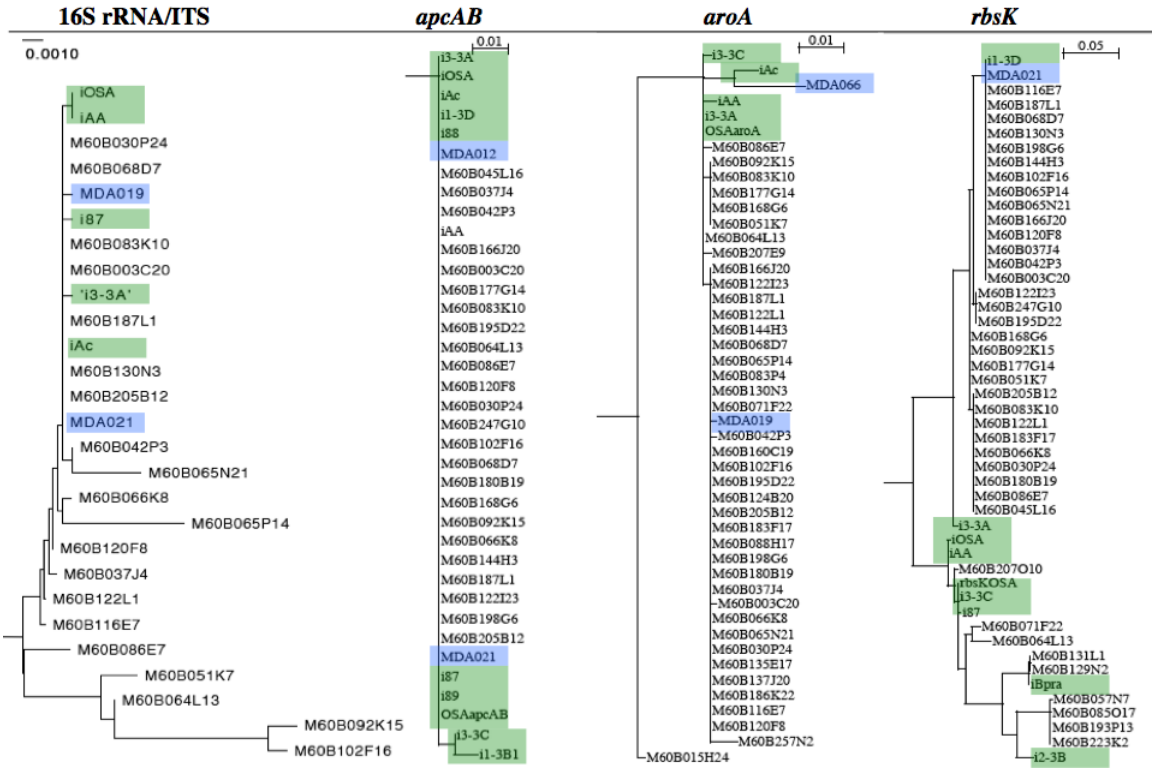



Figure E5.1. Neighbor-joining phylogenetic trees that have a random sample of M60 BAC clones (not highlighted), isolates (green) and MDAs (blue) for *Synechococcus* strain A.

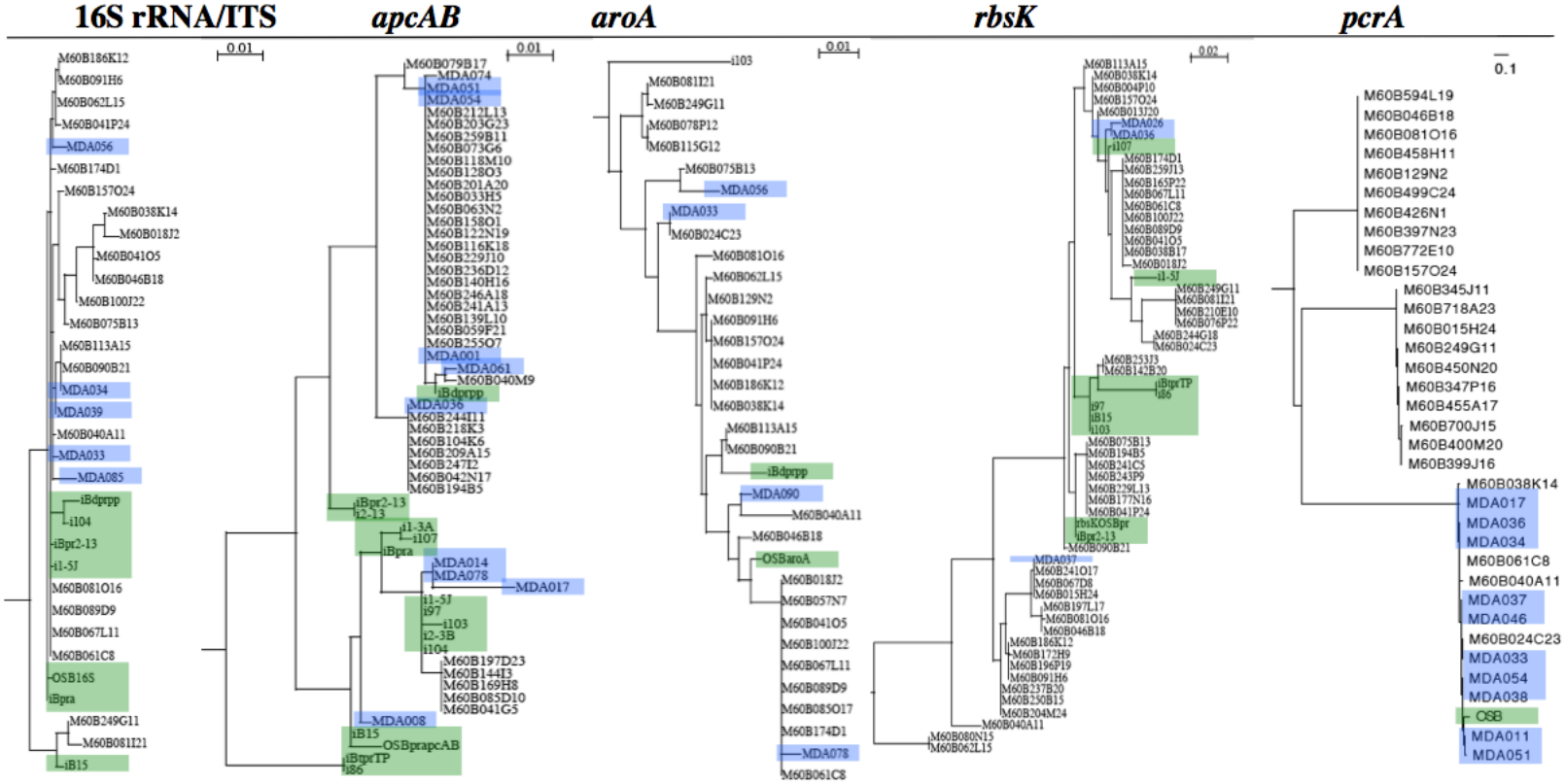



Figure E5.2. Neighbor-joining phylogenetic trees that have a random sample of M60 BAC clones (not highlighted), isolates (green) and MDAs (blue) for *Synechococcus* strain B´.

References

Allewalt JA, Bateson MM, Revsbech NP, Slack K and Ward DM. (2006). Temperature and light adaptations of *Synechococcus* isolates from the microbial mat community in Octopus Spring, Yellowstone National Park. *Appl Environ Microbiol* **72**: 544-550.