Ecotype Simulation 2: An improved algorithm for efficiently demarcating microbial species from large sequence datasets

Jason M. Wood[1], Eric D. Becraft[1,2], Danny Krizanc[3], Frederick M. Cohan[4], and David M. Ward[1]

[1] *Department of Land Resources and Environmental Sciences, Montana State University*
[2] *Bigelow Laboratory for Ocean Sciences*
[3] *Department of Mathematics and Computer Science, Wesleyan University*
[4] *Department of Biology, Wesleyan University*

Abstract

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

Keywords: speciation, bacteria, periodic selection, bacterial diversity

1 Background

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

Early work on microbial species was based on the use of phenotypic differences in cell shape, composition, and metabolism to differentiate between populations. However, systematists were unable to look within natural populations to verify that they were not lumping organisms with similar phenotypes into unnatural groups containing multiple species. With the hope of adding more scientific rigor to species demarcation, molecular divergence cutoffs such as 70% whole-genome DNA-DNA hybridization (Wayne et al., 1987), 95-96% amino acid identity (Konstantinidis and Tiedje, 2005), or about 1% identity in 16S rRNA (Stackebrandt and Ebers, 2006) were suggested and calibrated to match the historical phenotype-based groupings. However, little attention was given to whether the resulting sequence clusters came from ecologically distinct populations, and indeed, most named species have been found to include multiple ecologically distinct groups (Aboal et al., 2016; Dugat et al., 2015; Choudhary and Johri, 2011; Oh et al., 2010; Paul et al., 2010; Hunt et al., 2008; Lefébure and Stanhope, 2007; Vernikos et al., 2007; Kettler et al., 2007; Marri et al., 2006; Jaspers and Overmann, 2004; García-Martínez et al., 2002). Cohan and Kopac (2017) pointed out that lumping species not only hinders a full understanding of ecological diversity within a bacterial community, but also broadens the apparent diversity within and exaggerates the geographical range of clusters, complicates detection of newly emergent species, and may limit full appreciation of biotechnological potential.

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

Fortunately, systematists studying plants and animals have created a long list of species

concepts that microbial systematists can utilize for inspiration (Ward, 2006). Although there are disagreements among these species concepts (mainly involving the role of recombination), most share some common features: a species is ecologically distinct (minimal sharing of resources with other species), and is a cohesive group (diversity is limited among an ecotype's members, and a species is founded only once), while different species are ecologically distinct (minimal sharing of resources with other species), and are irreversibly separate (de Queiroz, 2005; Cohan, 2006; Cohan and Perry, 2007).

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

1.1 Theory-based demarcation of species using Ecotype Simulation

The Stable Ecotype Model was encoded into a Monte-Carlo style evolutionary simulation program called Ecotype Simulation (henceforth termed ES1) by Koeppel et al. (2008) which models a lineage of microbes through time as its members speciate and experience periodic selection events. The first stage of ES1 (see Figure 1A for an overview) loads user-supplied, aligned sequence data which can be used prior to the ecotype demarcation stage to generate a neighbor-joining phylogenetic tree using PHYLIP (Felsenstein, 2005). The second stage runs a series of Fortran programs designed to correct for PCR errors and remove gaps from the sequence data. It then calculates an $N * N$ divergence matrix to quantitatively summarize the diversity present (binning) among the N sequences in the dataset by utilizing complete-linkage clustering to obtain the number of sequence clusters (bins) over a range of sequence identity criteria (see Figure 2 for an example lineage). Briefly, the number of bins increases as the lineage diversifies through time from the last common ancestor (the blue dot) of the extant diversity (labeled A-Q). As depicted by the vertical bars to the right of the hypothetical tree shown in Figure 2, when the sequence identity criterion is restricted to more closely related sequences, more cluster bins result (see Figure 3 and Supplemental Figure 1 for a graphical presentation of results generated from environmental sequences).

In ES1 (Figure 1A), these bins are then used to estimate the rates of periodic selection (termed *sigma*), net ecotype formation (termed *omega*), genetic $drift$, and the estimated number of ecotype populations (termed *npop*) present in the sequence dataset analyzed. These estimated values are found in ES1 by using a brute-force method that repeatedly simulates the evolution of the lineage using a predefined range of *omega*, *sigma*, $drift$, and *npop* values, with the order of events and times at which they occur determined randomly. The estimated values that make up the best model are those that are able to reproduce the binning pattern of the dataset with maximum likelihood. Simulations that are able to reproduce the binning curve are deemed successful, with the most likely *omega*, *sigma*,

drift, and *npop* value combinations proceeding to the next stage – hill-climbing. The hill-climbing stage utilizes the Nelder-Mead simplex method (Nelder and Mead, 1965) to optimize the estimated *omega*, *sigma*, *drift*, and *npop* parameter values through repeated simulations. By varying a single parameter at a time and again utilizing the Nelder-Mead method to optimize the simulation, confidence intervals for each value can be found.

The optimized hill-climbing parameter values, along with a phylogenetic tree calculated from the sequence dataset (or provided by the user), are then used in the final stage of ecotype demarcation. The ecotype demarcation stage separately simulates each branch of the tree, starting from the root node (see the blue dot in Figure 2) and progressing though the tree recursively (to the right) until an internal node (red dot) is found with an *npop* value of 1 being within the confidence interval, or a single leaf node (labeled A-Q) is left, and demarcated as a putative ecotype (PE). This default coarse-scale demarcation method was modified by Becraft et al. (2015) to return fine-scale results by changing the demarcation criterion to require a *npop* value of 1 as the most likely result rather than simply being within the confidence interval. The ecotypes predicted by these analyses are considered PEs until they are proven to have the qualities of ecological species. That is, all PEs are expected to be ecologically distinct from each other, and the members within each PE are expected to be ecologically interchangeable with each other.

1.2 Application of Ecotype Simulation to Predict Species in Natural Systems

ES1 has been tested using microbial communities of Yellowstone National Park. A microbial mat community containing A/B′-lineage *Synechococcus* living along the effluent channel of Mushroom Spring, Yellowstone National Park, WY, USA has long been known to have phylogenetic clusters of closely related variants differentially distributed along the changing temperature of the flow path ranging from ∼72 °C to ∼50 °C. Ferris and Ward (1997) found 16S rRNA-defined A′- and A-clade *Synechococcus* in the higher temperature samples, and B′- or B-clade *Synechococcus* in the lower temperature samples. Isolated strains representative of 16S rRNA-defined linages demonstrated different adaptive responses to temperature, with an A-like *Synechococcus* strain, JA-3-3Ab, adapted to higher temperature than a B′-like *Synechococcus* strain, JA-2-3B'a(2-13) (Allewalt et al., 2006; Bhaya et al., 2007). Analyses based on a 557 nucleotide segment of a more rapidly evolving gene (*psaA*, which encodes an essential protein subunit of photosystem I) revealed that each 16S rRNA variant was comprised of numerous more closely related clusters (Becraft et al., 2011). ES1 analysis of an even shorter 324 nucleotide segment, obtained using high-throughput 454 Titanium sequencing of samples collected along temperature and vertical gradients, combined with barcoding and canonical correspondence analysis (CCA), have shown that these phylogenetic clusters are ecologically distinct (Becraft et al., 2015), and will be demonstrated below. This approach also showed these clusters to be cohesive groups (i.e., with ecologically interchangeable members), in the sense that all members of a group were shown to cluster non-randomly in CCA of temperature and depth distribution. Members of the same cluster reacted similarly, but members of different clusters reacted differently to environmental perturbation (Becraft et al., 2011, 2015). These PEs were shown to inhabit different locations within the vertical profile of the mat, with PEs B′9, A1 (which shares the same *psaA* sequence as strain JA-3-3Ab), A4, then A14 progressing from the surface to 1 mm below (Becraft et al., 2015). More recently, *Synechococcus* isolates that share the same *psaA* sequence as PEs A1, A4 (strain 65AY6A5), and A14 (strain 60AY4M2)

demonstrated different adaptive and acclimative responses to light intensity and quality, with optimal growth *in vitro* similar to conditions present where they can be found in the vertical profile of the mat *in situ* (Nowack et al., 2015). Comparative genomics and transcriptomics have elucidated genetic differences among these PEs that explain their different metabolic requirements (e.g., low-light adapted PEs were found to have an extra cassette of photosynthetic antenna genes that allow spectral fine tuning under low-light conditions) (Olsen et al., 2015). These analyses will be reviewed in detail below, as they provide a test case for comparison of ES1 and the newly developed algorithm described herein.

ES1 has also been used to study ecotypes of *Bacillus* in the "Evolution Canyons" of Israel and Death Valley. PEs detected by ES1 have been shown to differ in their preferences to solar exposure and soil texture, with membrane differences among PEs potentially explaining heat adaptation differences (Koeppel et al., 2008; Connor et al., 2010).

1.3 The need for a more efficient Ecotype Simulation

Although ES1 has proven useful for detecting ecologically distinct clades within groups of bacteria that had previously been considered a single species, the program is extremely slow and is greatly limited by the number of sequences that can be analyzed. Analyzing more than 200 sequences with ES1 can take days on modern computer hardware (Intel Core i7-6700) and then quickly becomes impossible as the number of sequences rises. This is at odds with the large number of DNA sequences that are produced by modern DNA sequencing technologies (Illumina MiSeq can now generate 25 million sequences per run). Various portions of ES1 (Figure 1A) were not designed to manage the large number of sequences produced by modern sequencing technologies (e.g., the memory usage of the ES1 matrix-based binning algorithm increases with the square of the number of sequences and the CPU usage increases with the cube of the number of sequences). The phylogenetic algorithms provided by PHYLIP and used by ES1 suffer from the same problem, and would offer no help in analyzing the large number of sequences produced with modern sequencing techniques. Other parts of the ES1 program needlessly store excess data in memory (e.g., both the Java object that loads Fasta-formatted sequence data and the Fortran simulation code store all sequences in memory even though the sequence data are unused during the simulation). By addressing these and other issues in this new version of Ecotype Simulation (henceforth termed ES2), we show that it is possible to successfully predict ecological species from very large datasets in a reasonable amount of time.

2 Implementation

2.1 Overview of Changes to Ecotype Simulation

ES1 and ES2 work in roughly the same way (see Figures 1A and B for an overview comparison) with some fundamental differences that may affect the results. ES2 has a redesigned graphical user interface that will be familiar to ES1 users but is more functional (Figure 4). ES2 also has a new command line interface for use when a graphics interface is not available. Like ES1, ES2 utilizes a phylogenetic tree for the ecotype demarcation stage, but in addition, ES2 uses the tree for the binning stage. ES2 incorporates FastTree (Price et al., 2009) rather than treeing algorithms found in PHYLIP (Felsenstein, 2005) if the user wishes to use ES2 to generate the phylogenetic tree. ES1 and ES2 both calculate the number of sequence clusters (bins) to quantitatively summarize the diversity present

in a sequence dataset at various levels of sequence identity criteria, but ES2 utilizes a new tree-based binning method that avoids the need to clean up the sequence data or to spend considerable time calculating a divergence matrix (see Figure 2 for an overview and Figure 3 for a comparison in resulting curves). Both ES1 and ES2 then use these bins to estimate the rates of periodic selection (*sigma*) and net ecotype formation (*omega*), and to estimate the number of ecotype populations (*npop*). However, ES2 uses a custom version of Lloyd's algorithm (Lloyd, 1982) for parameter estimation that converges on a likely solution much faster than the ES1 brute-force method. Since the population sizes of bacteria are incredibly large (Robinson et al., 2011), genetic drift can be safely ignored and has been removed from the simulation in ES2. The ES2 hill-climbing stage, the estimation of *omega*, *sigma*, and *npop* confidence intervals, and the ecotype demarcation stage are all simulated in the same manner described above for ES1, but with some fundamental differences and numerous optimizations made to the simulation that will be described below.

2.1.1 Phylogenetic Trees

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

2.1.2 Binning

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

2.1.3 Estimating *Omega*, *Sigma*, and *Npop*

The algorithm that estimates the initial values of *omega*, *sigma*, and *npop* for the hill-climbing in ES1 (Figure 1) used a brute-force approach – it searched a large range of predefined *omega*, *sigma*, and *npop* parameter values for likely results. A faster approach was needed to achieve the goals established for ES2. Using a custom variant of Lloyd's k-means algorithm (Lloyd, 1982) designed to partition point data into clusters based on distance from a line rather than a point, the ES2 approach fits two lines to the base-two logarithm of the binning data (the graph in Figure 4 shows the two fitted lines). The slope of the green line, closest to the 1.0 sequence identity criterion serves as an estimate of *sigma* – the rate of periodic selection. The slope of the blue line estimates the rate of ecotype formation – *omega*. The number of bins at the intersection of these two lines serves as the estimate for *npop* – the number of predicted ecotypes.

2.1.4 The Simulation

Multiple changes were made to the simulation code used by hill-climbing, demarcation, and the *omega*, *sigma*, and *npop* confidence intervals to decrease its runtime and memory usage. First, and most vital, the simulation now uses an ultrametric approach to estimate the number of mutations in each internode, rather than drawing from a Poisson distribution. Second, as mentioned above, the rate of genetic drift is no longer calculated or used in the simulation. Third, the waiting time to an event (either ecotype formation or periodic selection) was previously estimated through simulation-style code in ES1, but in ES2, the waiting time is estimated using a mathematical function ($timeWait = -1 * log(x)/rateKey$, with x chosen randomly from a uniform distribution in the range $[0, 1)$) that runs in constant time while retaining the stochasticity of $timeWait$. All three changes make the overall simulation much faster without greatly affecting the results.

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

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

2.1.5 Ecotype Demarcation

The ecotype demarcation stage has been slightly simplified in order to increase its speed. For a given subclade, the ES1 version of the algorithm tested all *npop* values between 1 and the current *npop* estimate for the entire clade to find the most likely number of ecotypes within the subclade. Since the point of this stage is to determine whether the subclade currently being examined makes up a single or multiple ecotypes, ES2 has been simplified

to assume that $npop > 1$ if the average success rate of 1000 trials is found to be zero when $npop = 1$ for the subclade. Only when the average success rate of these trials with $npop = 1$ is greater than zero are other values of *npop* tested using a likelihood-ratio test for whether they are significantly more likely ($\alpha = 0.05$).

2.2 Evaluation of Ecotype Simulation

2.2.1 Environmental Sequences

The environmental sequence segments analyzed by Becraft et al. (2015) were used to test the ability of ES2 to predict ecotypes inhabiting the Mushroom Spring microbial mat. To demonstrate the enhanced capabilities of ES2 while retaining the ability to compare results with ES1 (which is only able to analyze ∼300 sequences), the sampled diversity was limited to high-frequency sequences (HFSs) using two different cutoff values. We analyzed the same subset of *Synechococcus psaA* segment sequence diversity that was analyzed by Becraft et al. (2015). That is, we sampled only the HFSs, defined either as those with >50 identical copies across the entire dataset (HFS_{50}) or those with >10 identical copies across the entire dataset (HFS_{10}). In the HFS_{50} analysis, there were 65 A-like *Synechococcus* genotypes and 88 B′-like genotypes. In the HFS_{10} analysis, there were 246 A-like *Synechococcus* genotypes and 298 B′-like *Synechococcus* genotypes. The script that Becraft et al. (2015) used to find the HFS_{50} was used to find the HFS_{10} from the same environmental sequence dataset (`hfs-finder.pl`; available from `https://github.com/sandain/pigeon`). Each unique HFS was a single entry in each analysis. In the main text, we focus on predominant PEs in the B′-lineage (PE B′9) and A-lineage (PEs A1, A4, and A14) *Synechococcus*, which were found in ES1 analysis to have distinct vertical positioning in the mat at 60 °C to 63 °C. We use PE B′9 and the A-lineage PEs as examples in the main text, but include similar analyses for the remaining B′-lineage in the Supplemental Information (Supplemental Figures 2, 3, 4, and 5).

2.2.2 Tree Building

Because ES2 substituted the maximum-likelihood phylogenetic trees generated by FastTree for neighbor-joining phylogenetic trees generated by PHYLIP used in ES1, we needed to compare analyses made with both tree building algorithms. The neighbor-joining trees used by Becraft et al. (2015) for analyses of the *Synechococcus* A- and B′-like HFS_{50} subset of sequences were used here (Figure 5 and Supplemental Figure 2). The combination of PHYLIP's `dnadist` and `neighbor` programs (Felsenstein, 2005) were used with default parameters to generate neighbor-joining trees for the HFS_{10} subset of sequences. FastTree (Price et al., 2009) was used to build approximate maximum-likelihood trees for both the HFS_{50} and HFS_{10} subsets of sequences (Figure 6 and Supplemental Figure 3). PEs were separately demarcated for A- or B′-like *Synechococcus* neighbor-joining or maximum-likelihood phylogenies, generated from HFS_{50} or HFS_{10} datasets.

Since clade structure changes between tree algorithms and datasets used, PE names assigned by ES2 in different analyses do not usually correspond. Furthermore, PEs newly demarcated and named by ES2 analyses do not have names that correspond with previously named PEs demarcated with ES1 by Becraft et al. (2015). Newly demarcated PEs presented here from A- or B′-like *Synechococcus* are named using the pattern PEAxx or PEBxx, respectively (with xx being a number between 01 and 36). Correspondence between PEs previously described by Becraft et al. (2015) using ES1 is provided by examining the

most dominant variant in each ES2-demarcated PE. Newly demarcated PEs with dominant variants that were members of PEs previously demarcated by Becraft et al. (2015) will hereafter have the prior PE designation enclosed in parentheses after the new demarcation designation presented here. For instance, the ES1 PE A1 demarcation by Becraft et al. (2015) (both fine- and coarse-scale) was split by ES2 into PEs PEA13 (A1), PEA14, and PEA15 in neighbor-joining analysis (Figure 5), but only PEA13 (A1) retains the original PE designation, since it contains the same dominant variant (HFS013) as the ES1 PE A1.

2.2.3 Canonical Correspondence Analysis

Canonical correspondence analysis (CCA) (ter Braak, 1986; Legendre and Legendre, 1998) provided by the `R` package `vegan` (Oksanen et al., 2013) and a new version of the custom plotting software (`cca.R`; available from `https://github.com/sandain/R`) that was used by Becraft et al. (2015) were used to compare the ecological distribution of the members of PEs with the sampled environmental parameters (temperature of the water and depth within the mat). The community data matrix analyzed by CCA was created with the `hfs-counter.pl` script (available from `https://github.com/sandain/pigeon`) that simply counts the abundance of each HFS in each environmental sample. Although PEs were demarcated separately for A- and B′-like *Synechococcus* variants, the CCA community data matrix was comprised of variants from both lineages.

2.2.4 Testing for Ecological Interchangeability and Ecological Distinctness

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

2.2.5 Analysis of ES2 Runtime

To test the capabilities and results of ES2 using variously sized sequence datasets, the entire set of unique sequences sampled by Becraft et al. (2015) was randomly subsam-

pled every 100 sequences from 100 to 5600 sequences (i.e., 100, 200, .., 5600) with the `unique_random_fasta.pl` script (available from `https://github.com/sandain/pigeon`) and a maximum-likelihood tree was generated with FastTree. Each subset of sequences and its pregenerated tree was run with the ES2 command line interface three times for limited replication on an Intel Core i7-6700 processor and timed using GNU `time` version 1.7.

2.2.6 Other Algorithms Tested

In addition to comparing ES2 ecotype demarcations with those from ES1, the environmental sequence data and associated phylogenetic trees were also used to generate sequence clusters using Bayesian and maximum-likelihood Poisson tree processes (bPTP and PTP) (Zhang et al., 2013) (see the red and orange demaraction lines in Figures 5 and 6). Default parameters were used for all comparisons.

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

2.2.7 Comparing Demarcated Clusters

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

3 Results and Discussion

3.1 Binning

Here we examine the efficacy of the new tree-based binning method of ES2 compared to the distance-matrix-based binning method of ES1. We generated neighbor-joining trees using PHYLIP and maximum-likelihood trees using FastTree from the same sequence data binned previously using ES1. Because the neighbor-joining method provided by PHYLIP utilizes a matrix-based method that is very similar to the ES1 matrix-based binning method, it is not surprising that the ES2 binning results from neighbor-joining trees more closely follow the ES1 binning results of the environmental sequences than the ES2 results from the maximum-likelihood tree (compare the orange and red lines to the blue lines in Figure 3 and Supplemental Figure 1).

3.2 Performance of ES1 and ES2

In Figures 5 and 6, PEs are numbered according to ES2 output, and correspondence with predominant PEs predicted using ES1 by Becraft et al. (2015) is shown in parentheses. In general, the ES2 ecotype demarcation results were similar to the results of the ES1 fine-scale and coarse-scale methods. Differences in demarcation results between ES1 and ES2 could have resulted from the use of different tree algorithms (compare the neighbor-joining tree in Figure 5 to the maximum-likelihood tree in Figure 6) and/or different demarcation methods (compare the blue bars of different shade within either figure separately). Since we cannot be certain which tree is closest to reality, we rely on patterns of distribution in the hot spring mats to allow nature to inform us about the existence and composition of PEs. To do so, we focused on the four predominant PEs reported by Becraft et al. (2015), which showed different vertical distributions in 60 °C to 63 °C mat samples.

3.2.1 Distribution patterns observed in Ecotype Simulation 1 Analyses

We first repeated ES1 analyses based on the HFS_{50} diversity in the neighbor-joining tree performed by Becraft et al. (2015) to allow comparison of ES1 and ES2. Predominant PEs B′9, A1, A4 and A14 were predicted in coarse-scale analysis (light-blue lines in Figure 5). These PEs were previously found to be vertically stratified from top to bottom in the upper ∼1 mm-thick photic zone of the mat in the order listed (Supplemental Figure 6). The fine-scale analysis yielded the same PEs except PE B′9 was split into PEs B′9-2 and B′9-1 (medium-blue lines in Figure 5).

We used CCA analyses to test the predictions of the Stable Ecotype Model with the distribution of PEs in the CCA ordination space. Coarse-scale ES1 analyses of Becraft et al. (2015) are presented in Supplemental Figure 6, but here we focus on the fine-scale ES1 analysis. The graphical presentation of PEs in the ordination space reflect their vertical stratification in the mat environment (Figure 7A), with top to bottom progression of PEs B′9-2, B′9-1, A1, A4, and A14. Note the distribution of centroids for each PE alone the depth vector. The five fine-scale PEs yielded distributions that largely did not overlap, demonstrating ecological distinctness among PEs. We also used CCA to test whether each PE was ecologically homogeneous. The criterion for homogeneity was that the members of a PE should show a non-random distribution in the environment represented by the CCA ordination space. All of these PEs (excluding PE A14) showed non-random distributions in the environment, suggesting that members within a given PE share similar ecological requirements and are thus interchangeable with each other.

Since ES2 uses a maximum-likelihood phylogeny, we next based our ES1 analyses on the HFS_{50} diversity in the maximum-likelihood tree, which presented an alternative view of the A/B′-lineage *Synechococcus* phylogeny than that provided by a neighbor-joining phylogeny. The same predominant PEs detected in the fine-scale neighbor-joining analyses (PEs B′9-2, B′9-1, A1, A4 and A14) were detected in fine-scale maximum-likelihood analysis (light-blue and medium-blue lines in Figure 6B). In CCA analyses, these PEs exhibited the same vertical stratification from top to bottom in the upper ∼1 mm-thick photic zone of the mat, in the order listed above (Figure 7B), demonstrating ecological distinctness. All predominant PEs that contained more than one HFS_{50} member showed a non-random distribution in the environment, demonstrating ecological interchangeability in these PEs. Fine-scale PEs B′9-2, B′9-1, and A1 contained only a single HFS_{50} member, but the members of PEs A4, and A14 showed a non-random distribution in the environment, suggesting ecological

interchangeability among the members of these PEs. All predominant PEs in fine-scale analyses clustered separately in the environment represented by the CCA ordination space, suggesting that these PEs are ecologically distinct from one another (Figure 7B). Similar results were found for coarse-scale analyses based on both treeing algorithms (compare Supplemental Figures 6A and 6B).

3.2.2 Distribution patterns observed in Ecotype Simulation 2 Analyses

In the main text, we report results of ES2 analyses based on maximum-likelihood phylogenies. Results based on neighbor-joining phylogenies, which were highly similar, can be found in the Supplemental Information (Supplemental Figure 7). Analyses based on the HFS_{50} diversity in the maximum-likelihood tree detected predominant PEs PEB31 (B′9-2), PEB28 (B′9-1), PEA20 (A1), PEA16 (A4), and PEA19 (A14) (dark-blue lines Figure 6). In this analysis, there was a single PEB31 (B′9-2) variant and a single PEA20 (A1) variant, but PEs PEB28 (B′9-1), PEA16 (A4) and PEA19 (A14) had multiple variants that clustered non-randomly in the environment analyzed with CCA (Figure 8A). Again, the top to bottom distribution was in the order observed above, PEB31 (B′9-2), PEB28 (B′9-1), PEA20 (A1), PEA16 (A4), and PEA19 (A14).

ES2 analysis of HFS_{10} diversity, which could not be done with ES1 due to its excessive memory and CPU usage, increased the number of variants detected in each predominant PE simply due to the increased number of variants analyzed. ES2 analysis of HFS_{10} diversity in the maximum-likelihood tree resulted in 7-12 variants per PE. Most predominant PEs, with the exception of PEB061 (B′9-1), exhibited non-random distributions in CCA analysis and the same vertical patterning previously observed (Figure 8B).

3.3 Comparison of ES1 and ES2 PE Predictions

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

3.4 Analysis of the full HFS_{10} dataset

Analysis of the entire HFS_{10} dataset yielded a much larger number of PEs than those described here, though many are rare contributors to the dataset, which are likely to represent rare members of the mat community (note the blue abundance bars in Figure 9). Significant environmental clustering in CCA ($p < 0.05$) is observed more frequently in the most abundant PEs (e.g., PEs A1, A4, A14; compare the orange p-value bars and the dashed red confidence interval lines with the corresponding blue bars in Figure 9). One exception is PE A6 of Becraft et al. (2015), which is best explained by the observation that the partial *psaA* sequence used to demarcate species is shared by two widely divergent phylogenetic groups (Olsen, 2015). In most cases, however, partial *psaA* sequences permitted the detection of distinct ecological species. If horizontal gene exchange had been rampant, one would expect more well-sampled PE clusters to not be significantly clustered in CCA. The lower

degree of significant clustering of variants in rare PEs may indicate the randomization of distributions, possibly due to these members being dispersed randomly into the community, as opposed to occupying a discrete niche within the community.

3.5 Comparison With Other Algorithms

The ES2 demarcation results were compared with predictions of other algorithms that share a similar goal of species prediction (Figures 5 and 6, and Supplemental Figures 2 and 3). The maximum-likelihood Poisson Tree Processes (PTP) and the related Bayesian PTP (bPTP) models (Zhang et al., 2013) were compared to the ES1 and ES2 demarcation results. Because each algorithm is dependent upon the phylogenetic tree used, both neighbor-joining trees created with PHYLIP and maximum-likelihood trees created with FastTree were used with all algorithms.

For the most part, the PTP and bPTP demarcation results differed greatly from the ES1 and ES2 results, with PTP generally lumping more and bPTP generally splitting more than ES1 and ES2. There are, however, multiple examples of overlap in predicted ecotype membership among the five methods tested (see A-like PEA22 (A20) in Figure 5B for one example in the neighbor-joining analysis and A-like PEs PEA01 (A′21), PEA02 (A′22), PEA10, PEA11, PEA13, PEA14, PEA15, PEA21 (A20), PEA22, PEA23, and PEA24 in Figure 6B for eleven examples in the maximum-likelihood tree). PTP and bPTP detected only one or two of the four PEs focused on in this study, and detected only 7.6 to 46% of all predominant A-like PEs plus PE B′9 detected by ES2.

By comparing the results of all tested algorithms against the results of ES1 coarse- and fine-scale methods with the Variation of Information statistic, one can see that ES2 performs very well at reproducing results similar to ES1 (Table 1). ES2 achieved a lower value and, as such, performed better than all other algorithms tested except when comparing results with the ES1 coarse-scale demarcation of the neighbor-joining phylogeny. In this case, the ES1 fine-scale method performed better than all others ($VI_{ES1} = 0.74$ vs. $VI_{ES2} = 1.35$). When using the same neighbor-joining phylogeny for demarcation, but comparing ES2 with the ES1 fine-scale method, ES2 performed better than the ES1 coarse-scale method ($VI_{ES1} = 0.74$ vs. $VI_{ES2} = 0.61$). With a maximum-likelihood phylogeny, ES2 performed better than ES1 fine-scale analysis at reproducing ES1 coarse-scale analysis results ($VI_{ES1} = 1.40$ vs. $VI_{ES2} = 1.26$), and ES2 performed better than ES1 coarse-scale analysis at reproducing ES1 fine-scale analysis results ($VI_{ES1} = 1.37$ vs. $VI_{ES2} = 0.93$). The PTP algorithm only performed better than ES2 when comparing results with the ES1 coarse-scale demarcation of the neighbor-joining phylogeny ($VI_{PTP} = 0.97$ vs. $VI_{ES2} = 1.35$). ES2 performed better than the bPTP algorithm with all comparisons.

3.6 Testing the Runtime of ES2

In order to test how well the changes made to ES2 were to meeting the goal of analyzing a large number of sequences, 5,689 unique A-like *Synechococcus* environmental sequences collected by Becraft et al. (2015) were randomly subsampled to test the new algorithm on variously sized datasets. ES2 was used on 56 subsamples, with sample sizes ranging from 100 to 5600 sequences, to predict PEs (Figure 10 and Supplemental Figure 5 for results from B′-like sequences). Although the maximum number of environmental sequences analyzed in this experiment ($n = 5600$) does not fully test the power of ES2, the overall linear nature of the runtime curve (see the blue line in Figure 10) and the fact that the average length

of the analysis didn't exceed 90 minutes, illustrated that analyses of even larger datasets is possible. Indeed, analyses based on much deeper sampling by 454 GS FLX Titanium and Illumina MiSeq sequencing have been performed using ES2 (Wood, 2018). Such analyses would not be possible with ES1.

An interesting side effect of this experiment was that it provided a rarefaction curve for number of PEs that may help offer some insight into how well the diversity present in the environment had been sampled by this study. The total number of PEs predicted by ES2 (see the orange line in Figure 10) with a maximum likelihood tree appeared to be leveling off, suggesting that a sample of 5,689 unique sequences revealed nearly the total ecological diversity in A-like *Synechococcus* in this community. The number of PEs with only a single member (see the red line in Figure 10) appeared to follow a similar pattern to the total PE count and seemed to be leveling out. Our analyses predicted the existence of about 1000 PEs with more than a single representative. Most of these were rare and it was unclear whether they occupy niches within the system or are merely dispersed into the community.

4 Conclusions

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

Although this study only examined the diversity within A/B′-lineage *Synechococcus* populations, we believe that ES2 will be applicable on a large scale because a) ES2 utilizes a theory-based model of speciation that functions on neutral genetic diversity present in the gene segment, gene, or suite of genes analyzed by the program, b) ES1 was shown to predict ecologically distinct ecotypes of *Bacillus* (Koeppel et al., 2008; Connor et al., 2010) and ES2 performs similarly to ES1, and c) because it has been able to predict ecologically distinct ecotypes with ecologically interchangeable members within other phototrophic taxa living in same system studied here (Wood, 2018).

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

5 Availability and requirements

Project name: Ecotype Simulation
Project home page: `https://github.com/sandain/ecosim`
Operating systems: Windows, OSX, Linux
Programming languages: Java, Fortran
Other requirements: Java 8 or higher
License: GNU GPL

6 List of abbreviations

CCA: Canonical correspondence analysis
ES1: Ecotype Simulation version 1.
ES2: Ecotype Simulation version 2.
HFS: High Frequency Sequence (includes the minimum sequence count cutoff in subscript if applicable)
PE: Putative Ecotype.
npop: The estimated number of ecotype populations.
omega: Rate of net ecotype formation.
sigma: Rate of periodic selection.

7 Declarations

7.1 Ethics approval and consent to participate

Not applicable.

7.2 Consent for publication

Not applicable.

7.3 Availability of data and material

The sequence data analyzed for this study are the same analyzed by Becraft et al. (2015) and are available from MG-RAST (`http://metagenomics.anl.gov`; 4613896.3–4614007.3).

7.4 Competing interests

The authors declare that they have no competing interests.

7.5 Funding

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

7.6 Authors' contributions

JW helped plan and develop ES2, analyzed barcode sequence data, performed canonical correspondence analyses and comparisons between algorithms, and assisted in preparation of the manuscript. EB tested early versions of ES2 and participated in preparation of the manuscript. DK helped plan and develop ES2 and participated in preparation of the

manuscript. FC helped plan and develop ES2, co-supervised the research, and participated in preparation of the manuscript. DW obtained funding for the project, supervised the work, and participated in preparation of the manuscript.

7.7 Acknowledgements

This study was conducted under Yellowstone National Park research permits YELL-0129 and 5494 (DW), and we appreciate the assistance from National Park Service personnel. We would also like to thank Lingyuan Ke, a student of DK, with help identifying the issue with using the Fortran 90 intrinsic pseudo-random number generator in a threaded environment.

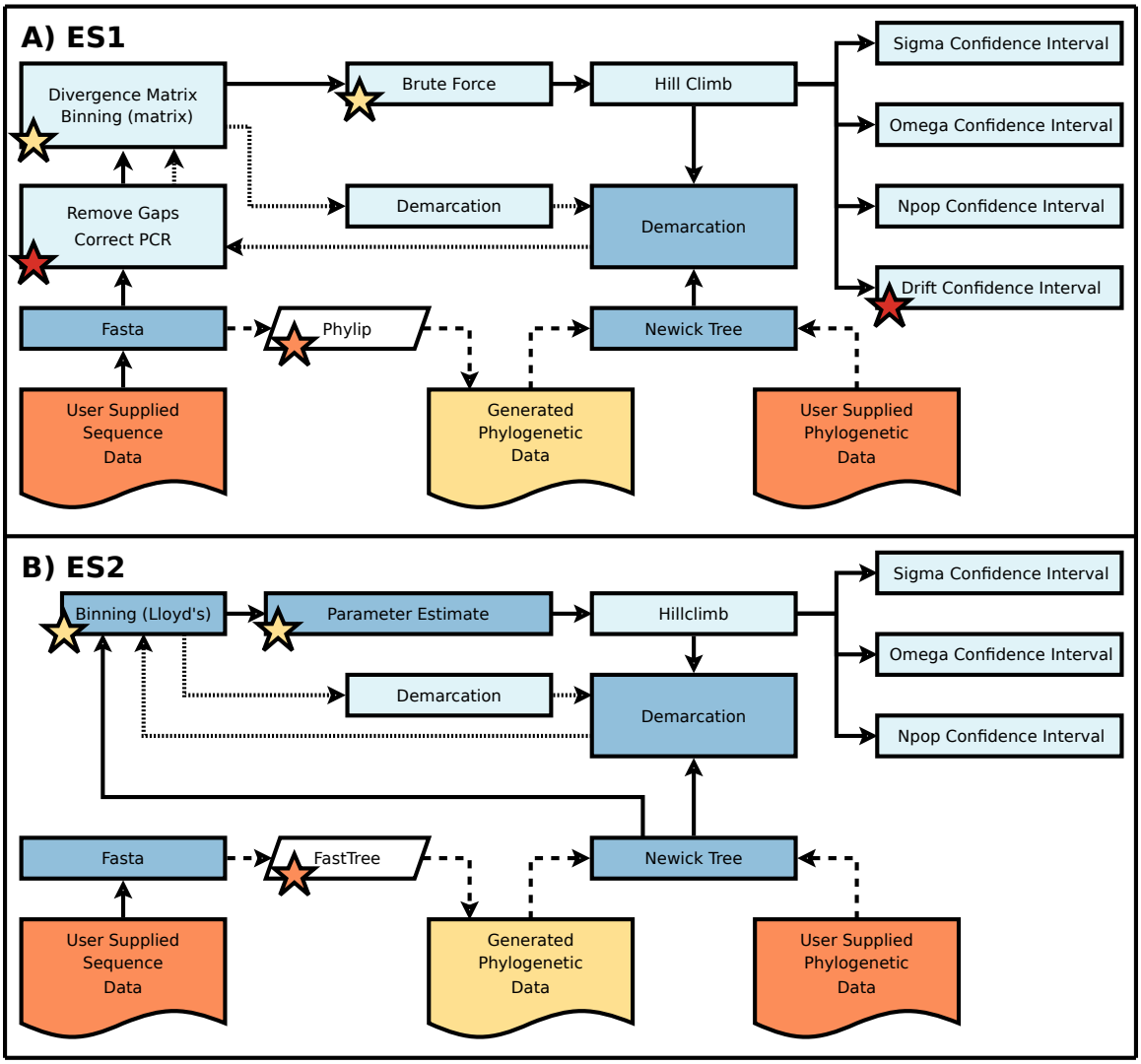


Figure 1: Simplified dependency flowchart of (A) Ecotype Simulation 1 (ES1) and (B) Ecotype Simulation 2 (ES2). Orange nodes are user-supplied data, yellow nodes are generated data, white nodes are external programs, dark blue nodes are Java methods, and light blue nodes are Fortran methods. The dashed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

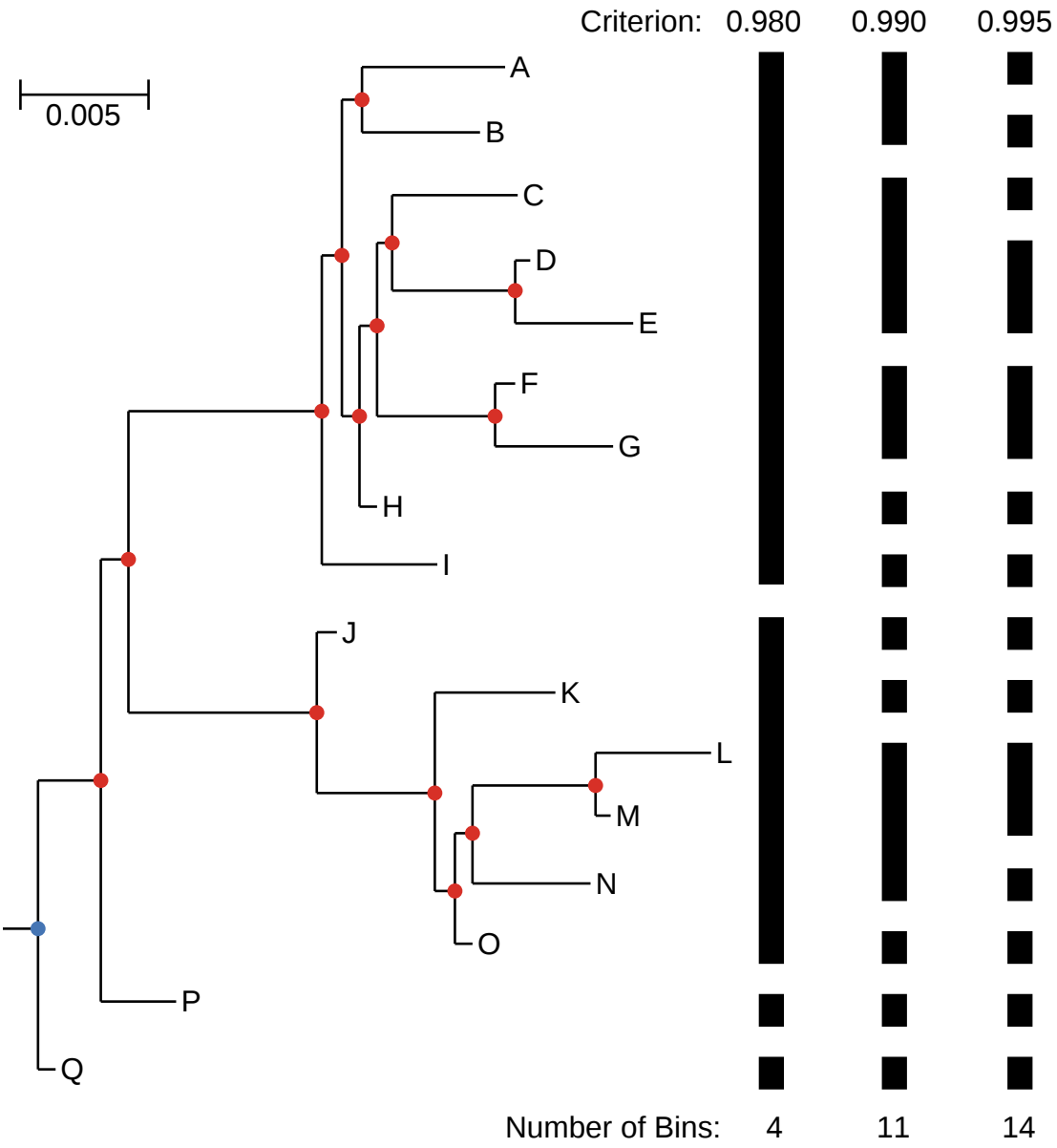


Figure 2: Sequence Identity Criterion Binning of a hypothetical lineage. The root node is marked with a blue dot, internal nodes are marked with red dots, and leaf-nodes that represent individual sequences are marked with the letters A-Q. Black bars denote the result of binning the tree at three different sequence identity criterion values with the number of bins listed below. The scale bar represent 0.005 nucleotide substitutions per site.

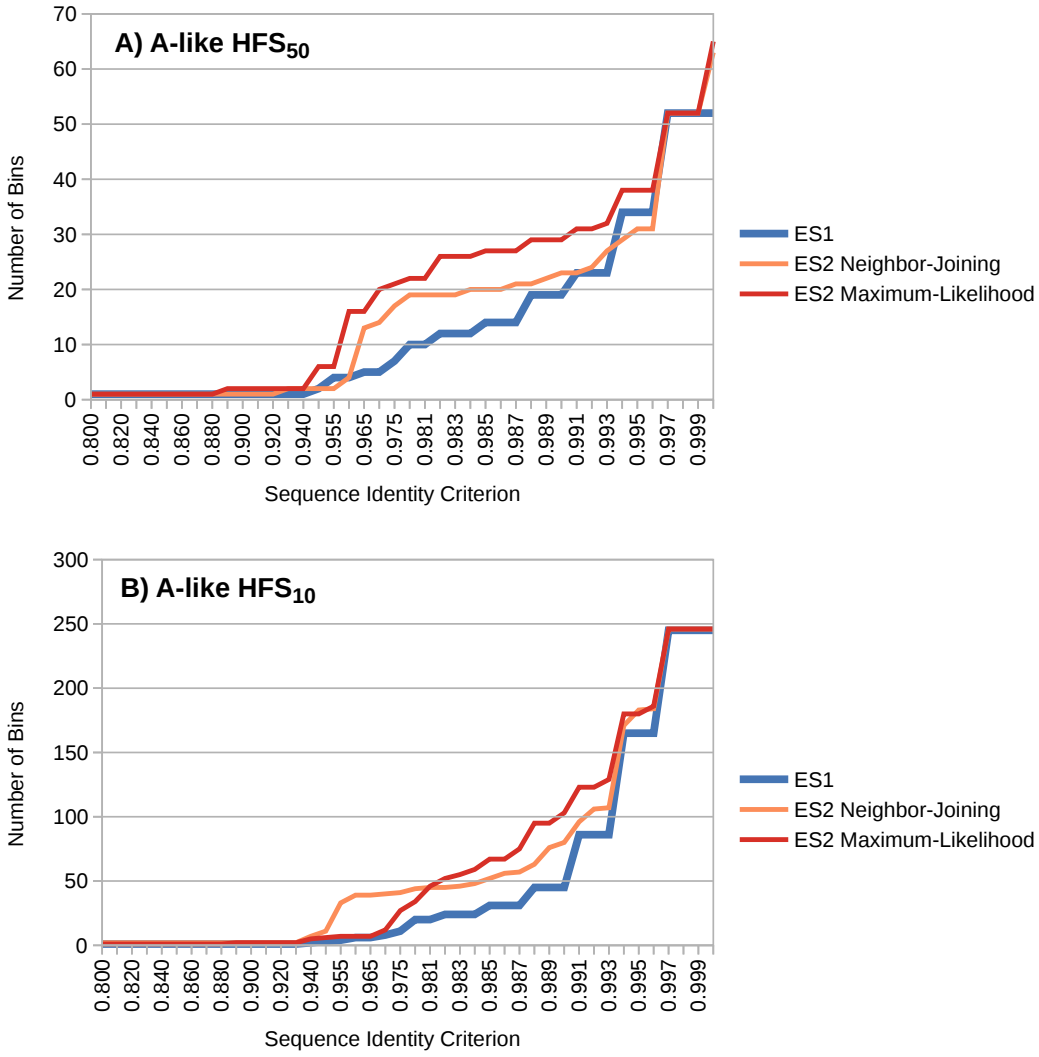
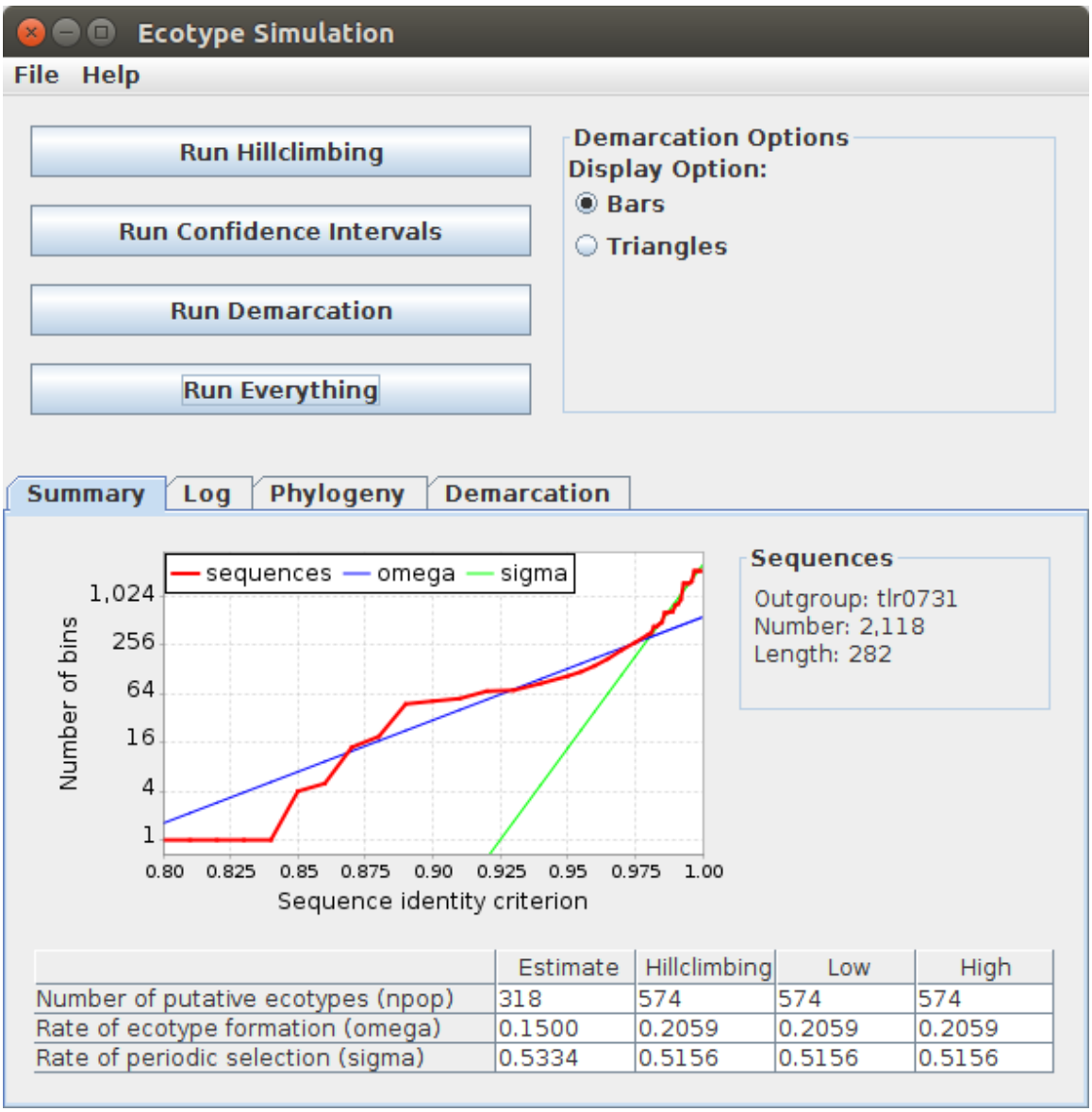


Figure 3: Binning results using A-like *Synechococcus psaA* segment high-frequency sequences (HFSs) occurring at least (A) fifty times (HFS_{50}) and (B) ten times (HFS_{10}) in the environmental sequence dataset.



	Estimate	Hillclimbing	Low	High
Number of putative ecotypes (npop)	318	574	574	574
Rate of ecotype formation (omega)	0.1500	0.2059	0.2059	0.2059
Rate of periodic selection (sigma)	0.5334	0.5156	0.5156	0.5156

Figure 4: Screenshot of the Ecotype Simulation 2 graphical user interface, with the Summary tab displayed.

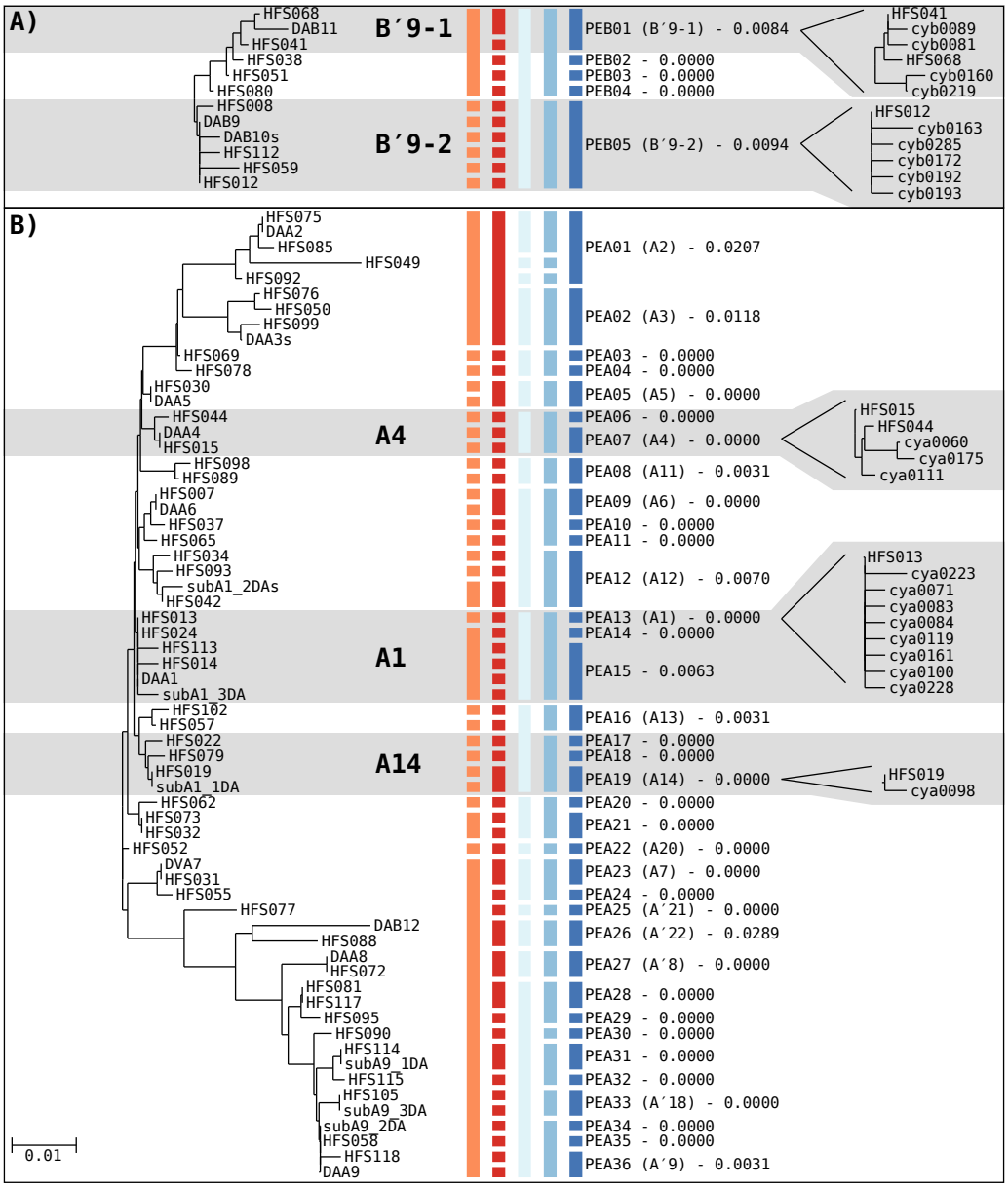


Figure 5: Neighbor-joining phylogeny with putative ecotype (PE) demarcation of (A) PE B′9- and (B) A-like *Synechococcus* HFS_{50} environmental *psaA* segments generated using PHYLIP. Gray shading denotes predominant PEs demarcated using Ecotype Simulation 1 (ES1) that are examined in detail in the main text. Colored vertical bars indicate demarcation done by different algorithms, from left to right: orange, maximum-likelihood Poisson tree processes; red, Bayesian Poisson tree processes; light-blue, ES1 coarse-scale; medium-blue, ES1 fine-scale; and dark-blue, Ecotype Simulation 2 (ES2). PEs are labeled with the ES2 demarcation and include the maximum distance among members of the clade. ES2-demarcated PEs that contain the same dominant variant as a PE demarcated by Becraft et al. (2015) using ES1 are indicated in parentheses after the ES2-generated PE names. Portions of the HFS_{10} neighbor-joining tree corresponding to predominant PEs PEB01 (B′9-1), PEB05 (B′9-2), PEA07 (A4), PEA13 (A1), and PEA19 (A14) are included on the right for comparison. The scale bar represent 0.01 nucleotide substitutions per site and is the same in parts A and B.

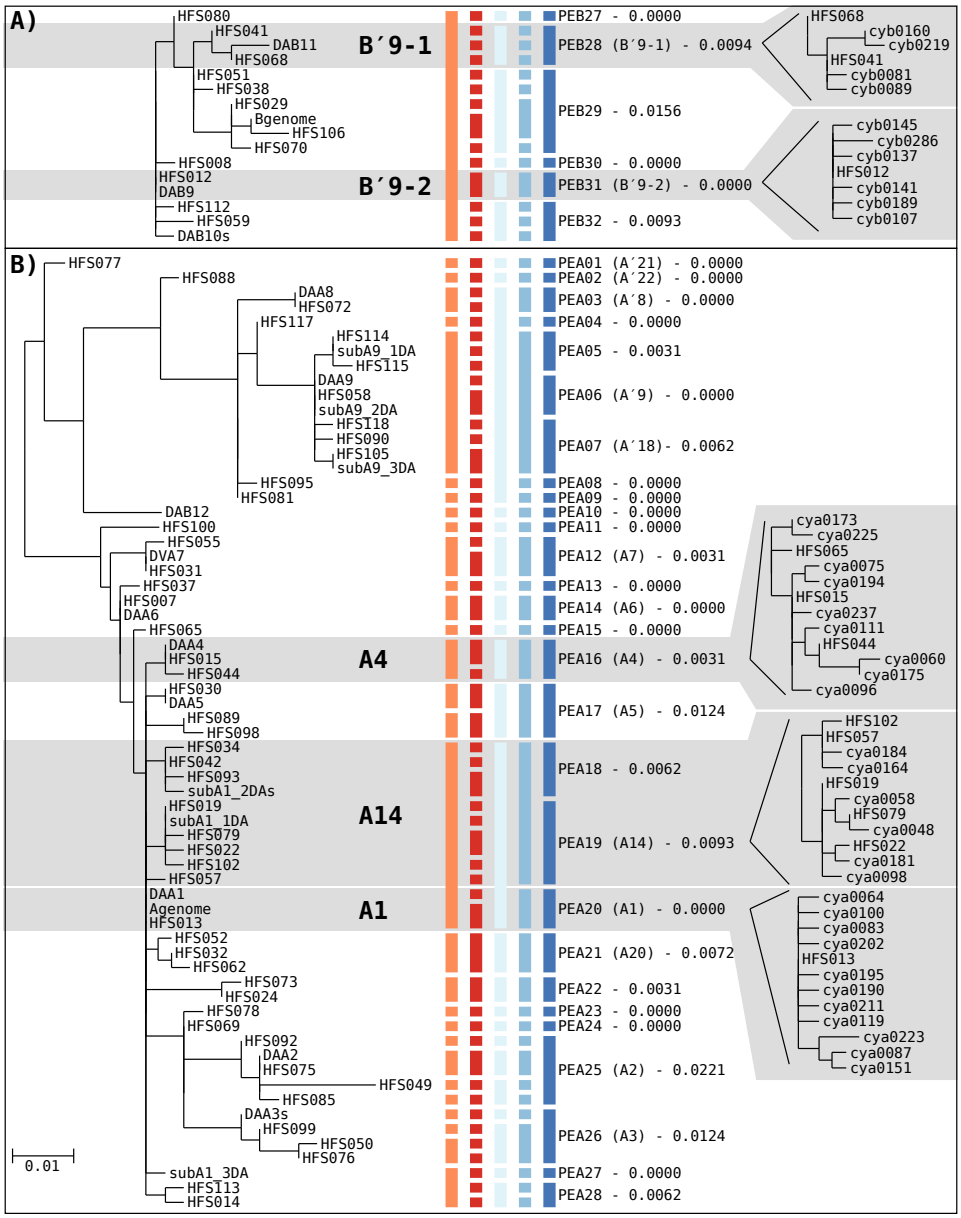


Figure 6: Maximum-likelihood phylogeny with putative ecotype (PE) demarcation of (A) PE B′9- and (B) A-like *Synechococcus* HFS_{50} environmental *psaA* segments generated using FastTree. Gray shading denotes predominant PEs demarcated using Ecotype Simulation 1 (ES1) that are examined in detail in the main text. Colored vertical bars indicate demarcation done by different algorithms, from left to right: orange, maximum-likelihood Poisson tree processes; red, Bayesian Poisson tree processes; light-blue, ES1 coarse-scale; medium-blue, ES1 fine-scale; and dark-blue, Ecotype Simulation 2 (ES2). PEs are labeled with the ES2 demarcation and include the maximum distance among members of the clade. ES2-demarcated PEs that contain the same dominant variant as a PE demarcated by Becraft et al. (2015) using ES1 are indicated in parentheses after the ES2-generated PE names. Portions of the HFS_{10} maximum-likelihood tree corresponding to predominant PEs PEB28 (B′9-1), PEB31 (B′9-2), PEA16 (A4), PEA19 (A14), and PEA20 (A1) are included on the right for comparison. The scale bar represent 0.01 nucleotide substitutions per site and is the same in parts A and B.

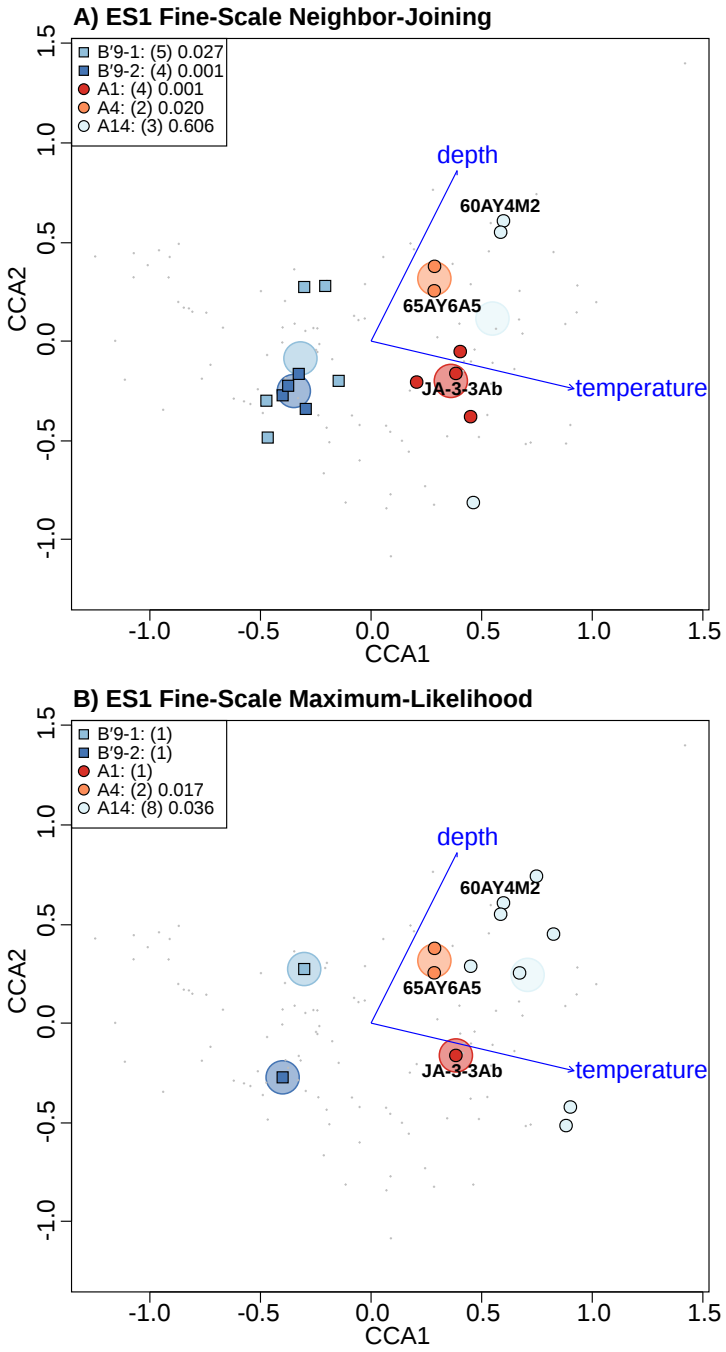


Figure 7: Canonical correspondence analysis highlighting *psaA* variants of predominant *Synechococcus* putative ecotypes (PEs) demarcated by the Ecotype Simulation 1 fine-scale method using (A) neighbor-joining and (B) maximum-likelihood phylogenies. Phylogenies were created from environmental *psaA* segments that numbered >50 (HFS_{50}). Demarcation of PEs in A performed by Becraft et al. (2015). Large, lighter colored circles represent the centroids of highlighted PEs. *Synechococcus* strains JA-3-3Ab, 65AY6A5, and 60AY4M2 share *psaA* sequences with HFS_{50} in these predominant PEs and are labeled on each plot. P-values are associated with the hypothesis that the members of PEs should not be distributed randomly.

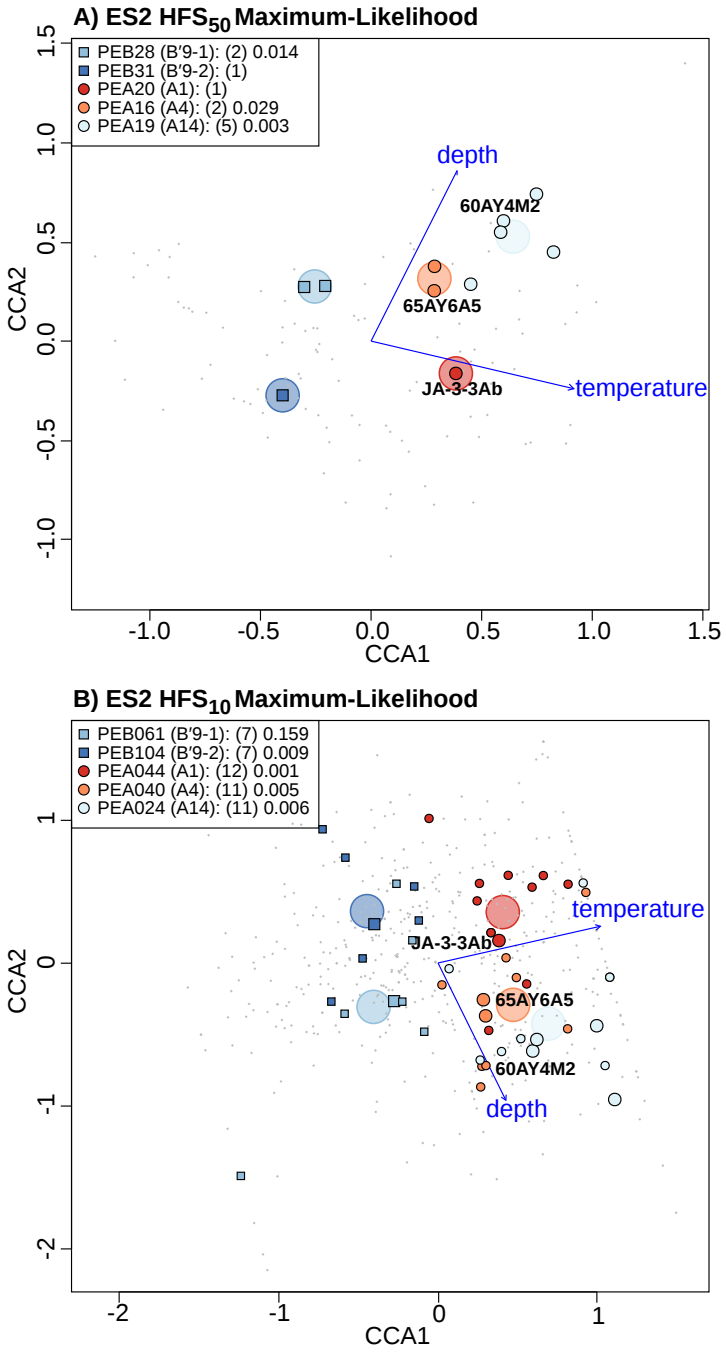


Figure 8: Canonical correspondence analysis highlighting *psaA* variants of predominant *Synechococcus* putative ecotypes (PEs) demarcated by Ecotype Simulation 2 using a maximum-likelihood phylogeny. Phylogenies were created from environmental *psaA* segments that numbered (A) >50 (HFS_{50}), or (B) >10 (HFS_{10}). Larger glyphs in B represent HFS_{50} environmental sequences, while smaller glyphs represent HFS_{10} environmental sequences. Large, lighter colored circles represent the centroids of highlighted PEs. *Synechococcus* strains JA-3-3Ab, 65AY6A5, and 60AY4M2 share *psaA* sequences with HFS_{50} in these predominant PEs and are labeled on each plot. P-values are associated with the hypothesis that the members of PEs should not be distributed randomly. Note the different orientations of the depth vector in A and B.

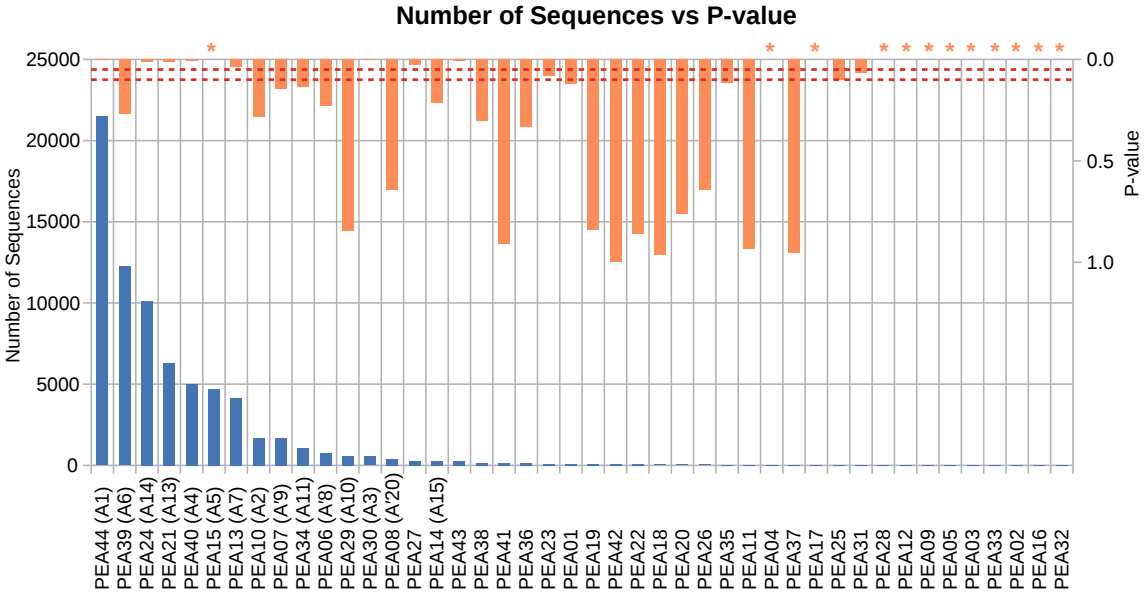


Figure 9: *psaA* Sequence abundance of A-like *Synechococcus* HFS_{10} environmental *psaA* segments in putative ecotypes (PEs) demarcated by Ecotype Simulation 2 from the maximum-likelihood tree (blue bars). Orange bars indicate p-values associated with the hypotheses that the members of PEs should not be distributed randomly. Orange asterisks above some columns represent p-values that could not be calculated since the PEs contained only a single member. The red dashed lines represent 0.05 and 0.10 confidence limits.

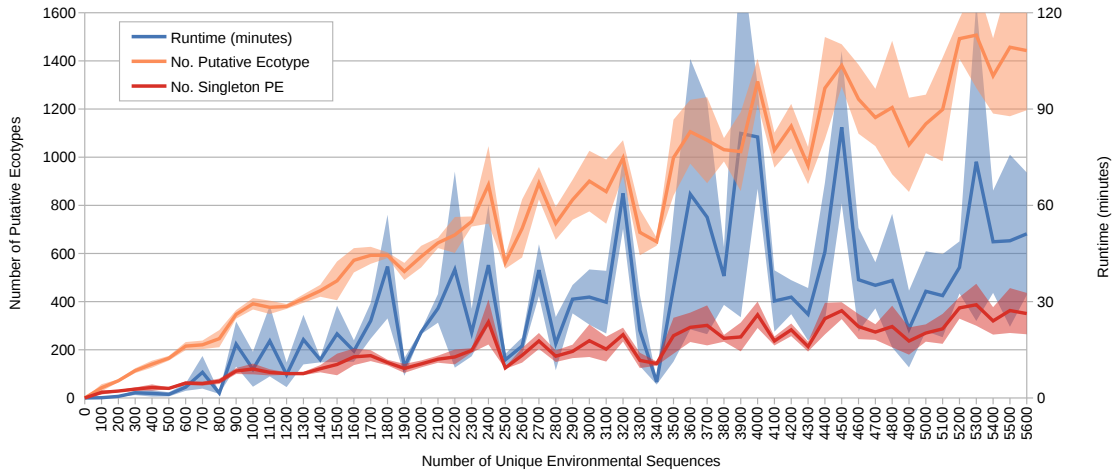


Figure 10: Rarefaction and runtime curves produced by repeated subsampling of the 5,689 unique A-like *Synechococcus psaA* segments present in the environmental dataset and subsequent analysis by Ecotype Simulation 2 (ES2). The darker colored lines represent the average and the shaded regions represent the standard error of three trials. The blue line and shaded region demonstrate the amount of time used to run the trial subsample using ES2 (on an Intel Core i7-6700 processor). The orange line and shaded region note the total number of PEs demarcated while the red line and shaded region note the number of PEs with only one member.

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

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

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

References

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

Allewalt, J. P., M. M. Bateson, N. P. Revsbech, K. Slack, and D. M. Ward. 2006. Effect of temperature and light on growth of and photosynthesis by *Synechococcus* isolates typical of those predominating in the Octopus Spring microbial mat community of Yellowstone National Park. *Appl Environ Microbiol*, 72:544–550. doi:10.1128/AEM.72.1.544-550.2006.

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

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

Bhaya, D., A. R. Grossman, A.-S. Steunou, N. Khuri, F. M. Cohan, N. Hamamura, M. C. Melendrez, M. M. Bateson, D. M. Ward, and J. F. Heidelberg. 2007. Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. *ISME J*, 1:703–713. doi:10.1038/ismej.2007.46.

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

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

Cohan, F. M. 2006. Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. *Philos Trans R Soc London Biol*, 361:1985–1996. doi:10.1098/rstb.2006.1918.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Marsaglia, G., and W. W. Tsang. 2000. The ziggurat method for generating random variables. *J Stat Soft*, 5:1–7. doi:10.18637/jss.v005.i08.

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

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

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

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

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

Olsen, M. T., S. Nowack, J. M. Wood, E. D. Becraft, K. LaButti, A. Lipzen, J. Martin, W. S. Schackwitz, D. B. Rusch, F. M. Cohan, D. A. Bryant, and D. M. Ward. 2015. The molecular dimension of microbial species: 3. Comparative genomics of *Synechococcus* strains with different light responses and *in situ* diel transcription patterns

of associated ecotypes in the Mushroom Spring microbial mat. *Front Microbiol*, 6:604. doi:10.3389/fmicb.2015.00604.

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

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

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

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

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

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

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

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

Wood, J. M. 2018. *Theory-based demarcation of hot spring microbial mat species from large DNA sequence datasets*. PhD thesis, Montana State University, Bozeman, Montana.

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