2012

Demarcation of bacterial ecotypes from DNA sequence data: A comparative analysis of four algorithms

Juan Carlos Francisco, *Wesleyan University*
Frederick M Cohan, *Wesleyan University*
Danny Krizanc, *Wesleyan University*

Demarcation of Bacterial Ecotypes from DNA Sequence Data: A Comparative Analysis of Four Algorithms

Juan Carlos Francisco[1], Frederick M. Cohan[2], and Danny Krizanc[3]
Department of Mathematics and Computer Science[1, 3] and Department of Biology[2]
Wesleyan University
Middletown, Connecticut, USA
jfrancisco@wesleyan.edu[1], fcohan@wesleyan.edu[2], dkrizanc@wesleyan.edu[3]

Abstract— **Identification of closely related, ecologically distinct populations of bacteria would benefit microbiologists working in many fields including systematics, epidemiology, and biotechnology. Several laboratories have recently developed algorithms aimed at demarcating such "ecotypes." In this paper we examine the ability of four of these algorithms to correctly identify ecotypes from sequence data (along with, in the case of one algorithm, information on the habitats where organisms were isolated). We test the algorithms on synthetic sequences, with known history and habitat associations, generated under the Stable Ecotype model [1], and on data from *Bacillus* strains isolated from Death Valley where previous work [2] has confirmed the existence of multiple ecotypes. We find that one of the algorithms (Ecotype Simulation) performs significantly better than the others (AdaptML, GMYC, BAPS) in both instances. Unfortunately, it is also shown to be the least efficient of the four.**

Keywords - species, speciation, ecotype, ecology, bacteria, demarcation algorithms

I. INTRODUCTION

The taxonomy of bacteria has provided microbiologists a system for routinely identifying species as closely related groups that differ in their disease-causing properties, in their ecological roles in biological communities, and in their physiological capacities [3, 4]. In the last two decades, bacterial taxonomy has adopted various universal molecular criteria for identifying species as clusters, first based on an indirect measure of shared genome content (i.e., DNA-DNA hybridization) [5, 6], and more recently based on measures of sequence identity of shared genes, including 16S rRNA sequence similarity [7], multilocus sequence analysis [8], and most recently, genome-wide average nucleotide identity [9]. Due to these efforts, it is becoming clear that a typical species taxon recognized by systematics is highly heterogeneous in its genome content and physiology, as well as its ecology [10-16]. A more fine-grained taxonomy, which recognizes the closely related, ecologically distinct populations that are now subsumed by bacterial systematics within a single species taxon, would benefit microbiologists from many fields including epidemiology and biotechnology.

Several laboratories have recently developed algorithms aimed to identify bacterial populations with the properties of "ecotypes" [17-20]. An ecotype is defined to constitute a paraphyletic or monophyletic group of close relatives that are ecologically interchangeable, in that the members of an ecotype share genetic adaptations to a particular set of habitats, resources, and conditions, and different ecotypes are distinct in their ecological adaptations [21, 22]. Our goal here is to evaluate four algorithms for their ability to discover the recently divergent ecotypes among close relatives of bacteria. Three of the algorithms we consider assume a force of cohesion within each ecotype, as seen in the Stable Ecotype model [22], and one (AdaptML) is more agnostic about the dynamic forces controlling diversity within ecotypes [18].

A. Models of Bacterial Speciation

The Stable Ecotype model. In the Stable Ecotype model, the homogeneity of ecological adaptations within an ecotype leads to genetic cohesion within the ecotype [1]. That is, sequence diversity within an ecotype is recurrently constrained by one of two cohesive forces, periodic selection and genetic drift. In periodic selection, an adaptive mutant outcompetes all other members of the ecotype, and by virtue of the rare recombination in bacteria [23], natural selection favoring the adaptive mutant causes a nearly genome-wide sweep of diversity [24]. Likewise, genetic drift can purge diversity genome-wide among the adaptively homogeneous membership of an ecotype, although genetic drift will be the primary force of cohesion only for ecotypes with extremely small population sizes [22, 25]. The dynamics of cohesion within each ecotype, coupled with a lack of cohesion between ecotypes, yields a method for recognizing ecotypes through sequence diversity [26]. While each population is recurrently purged of its diversity, each ecotype accumulates its own unique set of neutral mutations at every gene in the genome [22]. In principle, this will cause each ecotype to be recognized as a unique sequence cluster.

Habitat preference model. The algorithm AdaptML does not assume cohesion within ecotypes and works on a different principle. It assumes that each ecotype formation event causes at least a quantitative shift in habitat preferences [18]. That is, newly divergent ecotypes may overlap in the habitats that they utilize, but they are at least quantitatively different in their habitat preferences. This

978-1-4673-1321-6/12/$31.00 ©2012 IEEE

causes the ecotypes to differ in the frequencies at which they are found in each habitat. In this habitat preference model, ecotypes can be discovered as phylogenetic groups (either monophyletic or paraphyletic) that are significantly different in the habitats from which they are isolated.

B. The Algorithms

Ecotype Simulation (ES) [17]. Ecotype Simulation takes as input a lineage-through-time plot of sequence diversity [27], where the number of sequence identity bins is plotted against the sequence identity criteria for binning. This curve represents the history of splitting of lineages that have survived to the present. The algorithm searches for the combination of parameters that yields the observed curve with maximum likelihood. The parameters include: the rate of formation of new ecotypes (Ω), the rate of periodic selection (σ), and the number of ecotypes (*npop*) in the sample. A stochastic simulation, using a backward coalescence approach, generates the phylogenetic tree for the history of the sample, and then a forward simulation adds mutations to the sequences and generates sequence divergences. The algorithm tries many combinations of parameter values over many orders of magnitude, first through brute force and then by a hill-climbing approach.

Next, the algorithm demarcates the individual ecotypes. This is accomplished by recursively analyzing smaller and smaller subsets of the phylogenetic tree to find the most inclusive clades whose optimal value for the number of ecotypes (*npop*) is 1, while applying the rate values Ω and σ estimated for the whole phylogeny. This approach suggests that the analysis should be limited to closely related organisms.

A number of sequence-based analyses have used the ES algorithm. David Ward and colleagues have used it to identify ecologically distinct populations of *Synechococcus* in hot spring microbial mats [28-30], and the hypothesized ecotypes were shown to differ in their associations with temperature and light intensity and quality. The algorithm has also been used for an analysis of *Bacillus* isolates on different slopes of canyons in Death Valley, California [2] and in Israel [17] and on a salinity gradient in Death Valley (S. Kopac, pers. comm.). The hypothesized ecotypes were found to differ in their associations with solar exposure [2, 17], soil texture [2], salinity, and rhizospheres (S. Kopac, pers. comm.). Multiple ecotypes were found within recognized species taxa, and some ecotypes were previously unknown.

Ecotype Simulation is currently available for the Windows operating system, and has a straightforward graphical user interface.

GMYC [20]. The GMYC algorithm was originally designed to delineate species from sequences for sexually reproducing organisms such as insects [31]. It was later shown to be applicable to asexually reproducing organisms such as bdelloid rotifers [32] and more recently to bacterial sequence data (Barraclough *et al.*, 2009)**.** GMYC assumes a Yule model of speciation followed by a neutral coalescent model within species, where drift is the only process yielding coalescence. The algorithm maximizes the likelihood of a transition from the time that new species are being formed to the time when all coalescences are due to drift events within species. The software package that utilizes GMYC needs only a phylogenetic tree to demarcate strains. However, this tree must be ultrametric. The authors recommend the use of an algorithm developed by Sanderson [33], available in the APE package of the R language [34], to produce an ultrametic tree from sequences.

At the time of writing, the GMYC algorithm is available only upon request. GMYC runs on any computer with a version of the R programming language installed, which includes Mac, Linux, and Windows. The algorithm comes in the form of a bundle of R functions that must be imported into one's local R environment.

BAPS [35]**.** The Bayesian Analysis of Population Structure (BAPS) application refines existing Bayesian approaches to determine the structure of populations from genetic data. It assumes a partition-based mixture model and performs classification using a variant of the Metropolis-Hasting algorithm to identify clusters of sequences, with no explicit model of purging of diversity within clusters; the algorithm can take into account recombination within and between populations [35].

BAPS was used to perform a large-scale non-phylogenetic analysis of the population structure of bacteria from the genus *Neisseria* and found multiple clusters within recognized species [36], in spite of recombination frequencies exceeding the rate of mutation [23]. However, the authors did not attempt to determine whether the clusters identified by BAPS corresponded to ecologically distinct populations.

BAPS has Windows, Linux and Mac versions, all with easy to use GUIs. The software package also has a command line equivalent that is designed for batch processing.

AdaptML [18]. AdaptML places strains into ecotypes based on the assumption that the origin of each ecotype is driven by a change in habitat preferences. This algorithm has been used to demarcate ecotypes within the Vibrionaceae in coastal estuaries [18], within *Bacillus* in Death Valley [2], and within *Synechococcus* in Yellowstone hot springs [28, 29].

Unlike the cohesion-based algorithms, AdaptML does not accept as input the nucleotide sequences of the desired data set. Rather, AdaptML uses a phylogenetic tree (usually based on sequences) and data specifying measurements of the habitat of isolation for each strain. The algorithm assumes a hidden Markov model for the evolution of habitat associations and maximizes the likelihood of associations of the strains observed on the tree.

AdaptML runs on any operating system with a version of Python installed, which includes the Mac, Windows, and Linux operating systems. There is also a web app version of the algorithm where one can upload the data set and have AdaptML demarcations emailed.

The two classes of algorithms each have their advantages and disadvantages [37]. The three algorithms

based on the Stable Ecotype model have the advantage that they can demarcate ecotypes even when the investigator has no a priori ideas about either the ecological or physiological dimensions by which the ecotypes have diverged; a disadvantage is that the ecotypes hypothesized by these algorithms must be independently tested for their ecological distinctness. An advantage of AdaptML is that it simultaneously demarcates ecotypes and tests them for differences in their habitat preferences; so no further testing of ecological distinctness is necessary. The disadvantage of AdaptML is that it is limited to identifying ecotypes that are significantly different in preferences to habitat types that are anticipated by the investigators. Interestingly, three recent studies have shown that Ecotype Simulation and AdaptML reach highly similar ecotype demarcations [2, 28, 29], indicating that investigators' a priori guesses about the significant habitat types have been fairly accurate. Finally, a disadvantage of each class of algorithms is that their ability to find the adaptively homogeneous ecotypes is limited by the phylogenetic resolution of the sequences analyzed, and it is difficult to test whether a putative ecotype is homogeneous in its ecological adaptations [4].

In the present study, we perform a comparative evaluation of each of these four algorithms using algorithmically generated bacterial data with *a priori* knowledge of the ecotypes. This data set is generated under the assumptions of the Stable Ecotype model, with periodic selection acting as the force of cohesion within ecotypes. We also present the algorithms' demarcations of *Bacillus* strains in the Radio Facility Wash canyon in Death Valley [2]. Finally, we compare the running times of the algorithms.

II. METHODS

A. Generation of sequences for analysis

Comparing the accuracy of the four algorithms required us to generate data sets where the strain membership of each ecotype was known. This involved generating a history of a clade in silico for a sample of ν organisms, stemming from a single ancestral sequence. The clade history was based on the rate of ecotype formation (Ω), the rate of periodic selection (σ), and the number of ecotypes (*npop*) within the sample. We used the Ecotype Simulation algorithm to generate for each organism its sequence and its ecotype affiliation. Also, each ecotype was assigned a habitat preference. The ancestral organism was labeled as having habitat preference 'A' and each ecotype formation event was associated with a change in habitat preference, either from 'A' to 'B' or from 'B' to 'A'. We generated data sets with ν=20, 30, 50, and 100 simulated sequences, and based the simulations on three values each of Ω and σ. The middle values of Ω and σ were 0.19 and 1.1, respectively, based on prior Ecotype Simulation analysis of *Bacillus* in a Death Valley canyon [2]. Lower and upper values of Ω and σ represented 1/10x and 10x of the *Bacillus* values. *Npop* values of 10, 20 and 30 were used, except when they were greater than or equal to the number of sequences. We ran 100 replications for each combination of input parameters for $\nu < 100$, and 10 replications for each combination of parameters with $\nu = 100$.

Preparing the input. For each run, the maximum likelihood tree construction algorithm PhyML was used to build a phylogenetic tree from the generated sequences, since maximum likelihood trees worked best with the AdaptML algorithm. This tree was converted into an ultrametric chronogram using Sanderson's nonparametric rate smoothing algorithm [33], which is included in the APE package. This ultrametric tree was used as input for GMYC.

AdaptML requires the habitat from which each strain (or sequence) was isolated. While this information was readily available for our real *Bacillus* sequences, the habitat of isolation of each of the simulated sequences needed to be derived from the habitat preference of the strain's ecotype ('A' or 'B'). Ecotypes of habitat preference 'A' preferred habitat '1' and ecotypes of habitat preference 'B' preferred habitat '2'. We took into account different levels of specialization to habitat using the parameter γ, ranging from 0 (absolute specialization) to 50 (no specialization). Members of ecotypes with habitat preference 'A' were isolated from habitat '1' with probability 1-(γ/100) and from habitat '2' with probability γ/100; symmetrically, members of ecotypes with preference 'B' were isolated from habitat '2' with probability 1-(γ/100) and from habitat '1' with probability γ/100. We tested AdaptML with specialization values (γ) of 0, 10, 20, 30, 40 and 50, using the middle (*Bacillus*-based) values of Ω and σ, with *npop*=30.

FASTA sequences were converted into XLS format for compatibility with the BAPS package.

For all of the algorithms, we stored the ecotypes demarcated and the strains in these ecotypes in a MySQL database for easy and quick extraction and analysis.

***Bacillus* sequences.** *Bacillus* strains were isolated from Radio Facility Wash, a west-running canyon in Death Valley, consisting of habitats with three levels of solar exposure, including the canyon's sunny south-facing slope, the shadier and cooler north-facing slope, and the arroyo at the bottom [2]. These solar exposure habitats served as the single dimension of ecology we used as environmental input into AdaptML. DNA extraction, PCR amplification, partial sequencing of three genes, and concatenation of the genes were performed as previously described [2], and we used PhyML to produce a maximum likelihood tree.

B. Variation of Information metric

We used the metric Variation of Information (VI) [38], a criterion for comparing two partitions of the same data set, to determine the closeness of each algorithm's ecotype demarcations to the canonical demarcations generated in silico. The Variation of Information between two clusterings C and C' is given by

$$VI(C,C') = H(C) + H(C') - 2I(C,C')$$

where H(C) is the entropy of a random variable associated with a sequence being in a cluster C and I(C,C') is the mutual information of the two associated variables.

C. *Running time tests*

We tested the running time of each algorithm on a workstation with a dual-core Intel Core i7 processor running at 2.66 GHz and with 8 GB of RAM, running the Windows 7 operating system. We present the mean run times that each algorithm required to analyze synthetic data sets constructed using the parameter values estimated previously for *Bacillus*, with Ω=0.19, σ=1.1, *npop*=2 for γ=20, *npop*=5 for γ=30, and *npop*=10 for γ=50.

III. Results and Discussion

A. *Analysis of in silico-generated sequences*

Over all, ES produced the closest match to the known ecotype demarcations (mean VI over all parameter values=1.37), followed by GMYC (2.25), BAPS (2.38), and AdaptML (2.77), in that order. Even with absolute specialization of ecotypes to different habitats (γ=0), AdaptML was the least accurate (Table 1 – available at wesfiles.wesleyan.edu /home/dkrizanc/web/Table_1.pdf).

ES showed the greatest fidelity to the known ecotype demarcations in every parameter combination (Table 1). Lowering *npop* narrowed the gap in accuracy between ES and the other algorithms and also resulted in AdaptML producing more accurate demarcations than GMYC and BAPS.

Changing the values of Ω (rate of ecotype formation) had little effect on the accuracy of AdaptML and BAPS. GMYC and ES performed worse at the highest level of Ω. GMYC improved when Ω was decreased by a factor of 10, while ES stayed at mostly the same level of accuracy at low Ω.

Increasing or decreasing σ (rate of periodic selection) did not affect the accuracy of each algorithm as much as modifying Ω, but BAPS and GMYC had notable decreases in accuracy with a higher rate of periodic selection.

We found that the value of *npop* had a significant effect on the accuracy of three of the four algorithms: ES was largely unaffected when *npop* was raised or lowered, while AdaptML, BAPS and GMYC suffered a loss of accuracy with a higher number of ecotypes.

Why does ES out-perform the other algorithms? Perhaps it is because the data was generated under a periodic selection model, and ES is the only algorithm that explicitly assumes periodic selection to be the force of cohesion within ecotypes [37]. The other algorithms (especially GYMC) may perform better under a drift-based model. With regard to the lower performance of AdaptML, we note that our testing gave the algorithm the best possible chance to identify ecotypes. In the synthesis of sequence data, every ecotype formation event was coupled with a change in habitat; moreover, there were no ecotype formation events involving habitat changes in environmental dimensions not being analyzed. Also, we allowed complete habitat specialization (γ=0), which should give the algorithm its maximum resolution [18]. We therefore conclude that even when the investigators know the habitats over which ecotypes are specialized, AdaptML offers lower resolution to identify the ecotypes than the other algorithms, especially ES. Nevertheless, we note that AdaptML is extremely valuable in that it simultaneously identifies ecotypes and finds the environmental dimensions by which they have diverged [37].

Finally, we compared the VI scores of the different values of γ among all AdaptML runs (data not shown). As expected, accuracy was negatively correlated with γ, with AdaptML being the most accurate when γ was 0 (with absolute specialization of each ecotype to a different habitat).

B. Bacillus *sequences*

In the analysis of *Bacillus* sequences, the demarcation results of Ecotype Simulation, AdaptML, and BAPS were broadly similar, and AdaptML and BAPS yielded identical ecotypes. At the top of the phylogeny (Figure 1), Putative Ecotypes (P.E.) 1-5 identified by Ecotype Simulation were also identified by AdaptML and BAPS, except that P.E. 2 and 3 of ES were identified as a single ecotype by the other two algorithms. At the bottom of the figure, the clade identified as five ecotypes by ES (P.E. 7-11) was lumped into a single ecotype by AdaptML and BAPS.

GMYC identified many fewer ecotypes than any of the other algorithms. What was identified as six ecotypes or more by the other three algorithms at the top of the figure (P.E. 1-6 plus singleton ecotypes identified by ES), was demarcated as a single ecotype by GMYC.

Overall, ES identified more ecotypes than the other algorithms, and no algorithm split a single ecotype hypothesized by ES (with one exception—that P.E. 7 and 8 of ES were split into a number of single-strain ecotypes by GMYC). The apparent sensitivity of ES raises the question of whether this algorithm is seeing ecotypes that are not really there. Previous data show that the various putative ecotypes hypothesized by ES, but not identified by the other algorithms, are ecologically distinct and merit recognition as ecotypes. For example, at the top of the phylogeny, P.E. 1-6 of ES have been shown to be significantly heterogeneous in their associations with variations in solar exposure and soil texture [2]. While AdaptML and BAPS largely identified these putative ecotypes, GMYC missed them entirely (they are all within GMYC's P.E. 1). At the bottom of the phylogeny, P.E. 7-11 of ES, not discerned by either AdaptML or BAPS, were previously shown also to be significantly heterogeneous in their environmental associations [2]. We conclude Ecotype Simulation can resolve real, closely related ecotypes that the other algorithms cannot.

C. *Running time*

We compared the running times of the four algorithms on synthetic data sets of different sizes, and found that AdaptML, GMYC, and BAPS performed demarcations much more quickly than ES (Table 2). AdaptML, GMYC, and BAPS all completed demarcations on the order of a few seconds, even for data sets of 50 sequences whereas ES required on the order of tens of minutes for the same inputs. GMYC was the fastest algorithm, taking no more than one second on average. AdaptML followed, and BAPS was the slowest of the three faster algorithms.

IV. Conclusion

The algorithm Ecotype Simulation proved the most accurate of the algorithms studied, in analyses of synthetic sequence data and sequences obtained from closely related *Bacillus* strains. AdaptML and BAPS yielded overall similar demarcations as ES, with somewhat less accuracy, but the GMYC algorithm was unable to identify any of the *Bacillus* ecotypes that had previously been shown to be ecologically distinct. While ES is the most accurate, it is by far the slowest of the algorithms tested. If this algorithm is to be adapted to analyzing the huge data sets that are routinely sampled from environmental DNA, it will have to be made much faster. Improvements to the algorithm are currently under investigation.

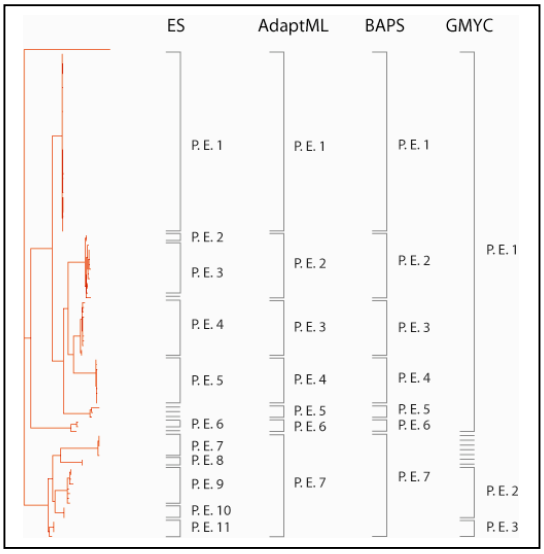

Figure 1. Maximum likelihood phylogeny of *Bacillus* strains, with ecotype demarcations by Ecotype Simulation (ES), AdaptML, BAPS, and GMYC.

Acknowledgment

This work was funded by NSF FIBR grant EF-0328698 and by research funds from Wesleyan University.

References

[1] J. Wiedenbeck and F. M. Cohan, "Origins of bacterial diversity through horizontal gene transfer and adaptation to new ecological niches," *FEMS Microbiology Reviews,* vol. 35, pp. 957-976, 2011.

Table 2. Run times (in seconds) of each algorithm when analyzing synthetic sets of sequences produced under the parameter values estimated for *Bacillus* [2].

Algorithm	20 sequences	30 sequences	50 sequences
ES	69.8	384	2390
AdaptML	1.54	1.57	1.64
GMYC	0.201	0.292	0.549
BAPS	4.80	5.15	6.12

[2] N. Connor, 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, "The ecology of speciation in *Bacillus*," *Applied and Environmental Microbiology,* vol. 76, pp. 1349-1358, 2010.

[3] R. Rosselló-Mora and R. Amann, "The species concept for prokaryotes," *FEMS Microbiol Rev,* vol. 25, pp. 39-67, Jan 2001.

[4] S. Kopac and F. M. Cohan, "A theory-based pragmatism for discovering and classifying newly divergent bacterial species," in *Genetics and Evolution of Infectious Diseases*, M. Tibayrenc, Ed. London: Elsevier, 2011, pp. 21-41.

[5] L. G. Wayne, D. J. Brenner, R. R. Colwell, P. A. D. Grimont, O. Kandler, M. I. Krichevsky, W. E. C. Moore, R. G. E. Murray, E. Stackebrandt, M. P. Starr, and H. G. Trüper, "Report of the ad hoc committee on reconciliation of approaches to bacterial systematics," *Int J Syst Bacteriol,* vol. 37, pp. 463-64, 1987.

[6] R. Lan and P. R. Reeves, "Gene transfer is a major factor in bacterial evolution," *Mol Biol Evol,* vol. 13, pp. 47-55, Jan 1996.

[7] E. Stackebrandt and J. Ebers, "Taxonomic parameters revisited: tarnished gold standards," *Microbiology Today,* vol. 33, pp. 152-155, 2006.

[8] W. P. Hanage, C. Fraser, and B. G. Spratt, "Sequences, sequence clusters and bacterial species," *Phil. Trans. Roy. Soc. Ser. B.,* vol. 361, pp. 1917-1927, 2006.

[9] K. T. Konstantinidis and J. M. Tiedje, "Genomic insights that advance the species definition for prokaryotes," *Proc Natl Acad Sci U S A,* vol. 102, pp. 2567-72, Feb 15 2005.

[10] M. Touchon, C. Hoede, O. Tenaillon, V. Barbe, S. Baeriswyl, P. Bidet, E. Bingen, S. Bonacorsi, C. Bouchier, O. Bouvet, A. Calteau, H. Chiapello, O. Clermont, S. Cruveiller, A. Danchin, M. Diard, C. Dossat, M. E. Karoui, E. Frapy, L. Garry, J. M. Ghigo, A. M. Gilles, J. Johnson, C. Le Bouguenec, M. Lescat, S. Mangenot, V. Martinez-Jéhanne, I. Matic, X. Nassif, S. Oztas, M. A. Petit, C. Pichon, Z. Rouy, C. S. Ruf, D. Schneider, J. Tourret, B. Vacherie, D. Vallenet, C. Médigue, E. P. Rocha, and E. Denamur, "Organised genome dynamics in the *Escherichia coli* species results in

highly diverse adaptive paths," *PLoS Genet,* vol. 5, p. e1000344, Jan 2009.

[11] S. T. Walk, E. W. Alm, D. M. Gordon, J. L. Ram, G. A. Toranzos, J. M. Tiedje, and T. S. Whittam, "Cryptic lineages of the genus *Escherichia*," *Appl Environ Microbiol,* vol. 75, pp. 6534-44, Oct 2009.

[12] G. C. Kettler, 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, "Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*," *PLoS Genet,* vol. 3, p. e231, Dec 2007.

[13] T. Lefebure and M. J. Stanhope, "Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition," *Genome Biol,* vol. 8, p. R71, 2007.

[14] P. R. Marri, W. Hao, and G. B. Golding, "Gene gain and gene loss in *Streptococcus*: is it driven by habitat?," *Mol Biol Evol,* vol. 23, pp. 2379-91, Dec 2006.

[15] S. Paul, A. Dutta, S. K. Bag, S. Das, and C. Dutta, "Distinct, ecotype-specific genome and proteome signatures in the marine cyanobacteria *Prochlorococcus*," *BMC Genomics,* vol. 11, p. 103, 2010.

[16] G. S. Vernikos, N. R. Thomson, and J. Parkhill, "Genetic flux over time in the Salmonella lineage," *Genome Biol,* vol. 8, p. R100, 2007.

[17] A. Koeppel, E. B. Perry, J. Sikorski, D. Krizanc, W. A. Warner, D. M. Ward, A. P. Rooney, E. Brambilla, N. Connor, R. M. Ratcliff, E. Nevo, and F. M. Cohan, "Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics," *Proceedings of the National Academy of Sciences,* vol. 105, pp. 2504-2509, 2008.

[18] D. E. Hunt, L. A. David, D. Gevers, S. P. Preheim, E. J. Alm, and M. F. Polz, "Resource partitioning and sympatric differentiation among closely related bacterioplankton," *Science,* vol. 320, pp. 1081-5, May 23 2008.

[19] J. Corander, P. Marttinen, J. Siren, and J. Tang, "Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations," *BMC Bioinformatics,* vol. 9, p. 539, 2008.

[20] T. G. Barraclough, M. Hughes, N. Ashford-Hodges, and T. Fujisawa, "Inferring evolutionarily significant units of bacterial diversity from broad environmental surveys of single-locus data," *Biol Lett,* vol. 5, pp. 425-8, Jun 23 2009.

[21] D. M. Ward, "A natural species concept for prokaryotes," *Curr Opin Microbiol,* vol. 1, pp. 271-7, Jun 1998.

[22] F. M. Cohan and E. B. Perry, "A systematics for discovering the fundamental units of bacterial diversity," *Current Biology,* vol. 17, pp. R373-R386, 2007.

[23] M. Vos and X. Didelot, "A comparison of homologous recombination rates in bacteria and archaea," *Isme J,* vol. 3, pp. 199-208, Feb 2009.

[24] F. M. Cohan, "Periodic selection and ecological diversity in bacteria," in *Selective Sweep*, D. Nurminsky, Ed. Georgetown, Texas: Landes Bioscience, 2005, pp. 78-93.

[25] C. H. Kuo, N. A. Moran, and H. Ochman, "The consequences of genetic drift for bacterial genome complexity," *Genome Res,* vol. 19, pp. 1450-4, Aug 2009.

[26] F. M. Cohan, "What are bacterial species?," *Annu Rev Microbiol,* vol. 56, pp. 457-87, 2002.

[27] A. P. Martin, "Phylogenetic approaches for describing and comparing the diversity of microbial communities," *Appl Environ Microbiol,* vol. 68, pp. 3673-82, Aug 2002.

[28] E. Becraft, F. M. Cohan, M. Kühl, S. Jensen, and D. M. Ward, "Fine-scale distribution patterns of *Synechococcus* ecological diversity in the microbial mat of Mushroom Spring, Yellowstone National Park," *Applied and Environmental Microbiology,* 2011.

[29] M. C. Melendrez, R. K. Lange, F. M. Cohan, and D. M. Ward, "Influence of molecular resolution on sequence-based discovery of ecological diversity among *Synechococcus* populations in an alkaline siliceous hot spring microbial mat," *Applied and Environmental Microbiology,* vol. 77, pp. 1359-1367, 2011.

[30] D. M. Ward, M. M. Bateson, M. J. Ferris, M. Kühl, A. Wieland, A. Koeppel, and F. M. Cohan, "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.,* vol. 361, pp. 1997-2008, 2006.

[31] J. Pons, T. G. Barraclough, J. Gomez-Zurita, A. Cardoso, D. P. Duran, S. Hazell, S. Kamoun, W. D. Sumlin, and A. P. Vogler, "Sequence-based species delimitation for the DNA taxonomy of undescribed insects," *Syst Biol,* vol. 55, pp. 595-609, Aug 2006.

[32] D. Fontaneto, E. A. Herniou, C. Boschetti, M. Caprioli, G. Melone, C. Ricci, and T. G. Barraclough, "Independently evolving species in asexual bdelloid rotifers," *PLoS Biol,* vol. 5, p. e87, Apr 2007.

[33] M. J. Sanderson, "r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock," *Bioinformatics,* vol. 19, pp. 301-2, Jan 22 2003.

[34] E. Paradis, J. Claude, and K. Strimmer, "APE: Analyses of Phylogenetics and Evolution in R language," *Bioinformatics,* vol. 20, pp. 289-90, Jan 22 2004.

[35] J. Corander and J. Tang, "Bayesian analysis of population structure based on linked molecular information," *Mathematical Biosciences,* vol. 205, pp. 19-31, 2007.

[36] J. Tang, W. P. Hanage, C. Fraser, and J. Corander, "Identifying currents in the gene pool for bacterial populations using an integrative approach," *PLoS Computational Biology,* vol. 5, p. e1000455, 2009.

[37] F. M. Cohan and A. F. Koeppel, "The origins of ecological diversity in prokaryotes," *Current Biology,* vol. 18, pp. R1024-R1034, 2008.

[38] M. Meila, "Comparing clusterings by the variation of information," in *Learning Theory and Kernel Machines*, B. Schölkopf and M. K. Warmuth, Eds. Berlin: Springer, 2003, pp. 173-187.