Microbial Speciation

B. Jesse Shapiro[1] and Martin F. Polz[2]

[1]Département de Sciences Biologiques, Université de Montréal, Montréal QC H3C 3J7, Canada

[2]Parsons Laboratory for Environmental Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Correspondence: jesse.shapiro@umontreal.ca; mpolz@mit.edu

Cold Spring Harbor Perspectives in Biology
www.cshperspectives.org

What are species? How do they arise? These questions are not easy to answer and have been particularly controversial in microbiology. Yet, for those microbiologists studying environmental questions or dealing with clinical issues, the ability to name and recognize species, widely considered the fundamental units of ecology, can be practically useful. On a more fundamental level, the speciation problem, the focus here, is more mechanistic and conceptual. What is the origin of microbial species, and what evolutionary and ecological mechanisms keep them separate once they begin to diverge? To what extent are these mechanisms universal across diverse types of microbes, and more broadly across the entire the tree of life? Here, we propose that microbial speciation must be viewed in light of gene flow, which defines units of genetic similarity, and of natural selection, which defines units of phenotype and ecological function. We discuss to what extent ecological and genetic units overlap to form cohesive populations in the wild, based on recent evolutionary modeling and population genomics studies. These studies suggest a continuous "speciation spectrum," which microbial populations traverse in different ways depending on their balance of gene flow and natural selection.

Species, in the vernacular sense, comprise individuals that are phenotypically and, hence, ecologically more similar to each other than to other species (Gevers et al. 2005; Cohan and Koeppel 2008). This notion was extended in the biological species concept of Dobzhansky (1935) and Mayr (1942), which states that species are reproductively isolated units, implying that adaptive mutations can spread within a species leaving other coexisting species unaffected. Although recent evidence has shown that reproductive boundaries can be leaky (Danchin and Rosso 2012; Syvanen 2012; Schönknecht et al. 2013), species are still regarded as congruent genetic and ecological units for sexual eukaryotes, even if hybrids and intermediate forms are common (Mallet 2008). For bacteria and archaea, however, the situation has been marred by several complicating factors that question whether such units can be defined.

In addressing whether we can identify genetically and ecologically congruent units, we need to take into account the peculiarities of bacterial and archaeal evolution, that is, the varying modes and rates of genetic exchange. In these organisms, incorporation of new genetic material is always unidirectional and leads either to gene conversion by homologous

Editor: Howard Ochman
Additional Perspectives on Microbial Evolution available at www.cshperspectives.org

 doi: 10.1101/cshperspect.a018143
Cite this article as *Cold Spring Harb Perspect Biol* 2015;7:a018143

recombination[3] or gene addition by nonhomologous recombination.[4] (In fact, the distinction might not be so clear. There is mounting evidence that homologous recombination is often involved in gene addition and loss [de Vries and Wackernagel 2002; Mell et al. 2011; Cordero et al. 2012a; Croucher et al. 2012].) Importantly, the rates and bounds of this gene transfer can vary considerably. Although some lineages follow a highly clonal mode of evolution, in others, recombination is a much more important evolutionary force than is mutation. Regardless of the overall rate of gene flow, genetic material can, in principle, be incorporated from distantly related organisms. This variation in genetic exchange and its effect on genotypic integrity and ecological adaptation is at the heart of the debate about what constitutes ecological and genetic units for bacteria and archaea.

In particular, horizontal gene transfer[5] (HGT) among distantly related organisms can create genotypes that vary in properties of ecological relevance by acquiring functions, such as antibiotic resistance or nitrogen fixation, that distinguish them from otherwise closely related genotypes (Doolittle and Papke 2006; Syvanen 2012). At the same time, the recipient genotype has also become ecologically similar, in at least one niche[6] dimension, to the organism from which it acquired the novel pathway. In fact, such functional differentiation is observed among closely related environmental isolates (Hahn and Pockl 2005) and, in combination with high gene turnover, has been taken as evidence that gene acquisition and loss is so high as to quickly erode any niche association of lineages (Doolittle and Papke 2006). By extension, the very notion of a lineage has been questioned on the same grounds—with the consequence that nearly each genotype might represent its own, independent ecological unit (Doolittle and Zhaxybayeva 2009) that can only be recognized by the functional genes it carries (Wiedenbeck and Cohan 2011).

In recent years, however, analysis of environmental isolates and metagenomes[7] has shown that microbial communities consist of genotypic clusters of closely related organisms and that these can display cohesive environmental associations and dynamics that clearly distinguish them from other such clusters coexisting in the same samples. Despite also showing evidence for extensive gene flow, genetically distinguishable clusters have been observed among closely related environmental and pathogenic isolates by multilocus sequence analysis and genomics (Gevers et al. 2005; Hanage et al. 2005; Luo et al. 2011), and by metagenomics (Konstantinidis and Delong 2008; Denef et al. 2010a; Oh et al. 2011). Moreover, cohesive ecological dynamics and associations have been shown for a growing number of cases, including for vibrios, sulfate-reducing bacteria, and cyanobacteria, as well for organisms represented in several marine, freshwater and acid-mine drainage community metagenomes. These observations suggest congruence of genotypic and ecological units and are, in principle, consistent with the notion of populations[8] as locally coexisting members of a species. As we will discuss below, selection and recombination are paramount in shaping and maintaining such units, although the effects of biogeography, on both local (Simmons et al. 2008; Denef et al. 2010a) and global (Whitaker et al. 2003; Reno et al. 2009) scales may also come into play.

[3]Homologous recombination is a mechanism of DNA integration requiring at least short tracts of identity between the genome and the foreign DNA, mediated by RecA and mismatch-repair machinery. The integrated DNA can result in single-nucleotide changes and in some cases, addition or loss of relatively long stretch of DNA including entire genes.

[4]Nonhomologous recombination refers to the integration of DNA with no homologous allele already present in the genome, often mediated by phage and integrative elements. This results in the acquisition of entirely new genes.

[5]Horizontal gene transfer (HGT) is the incorporation of foreign DNA into a genome. Incorporation can be mediated by either homologous recombination or nonhomologous recombination of DNA that enters a cell via transformation, transduction, or conjugation. In bacteria and archaea, all gene transfer is horizontal (i.e., always unidirectional).

[6]Here, niche is a specific set of ecological parameters (environments, resources, physical and chemical characteristics, biotic interactions, etc.) to which an organism is adapted. This does not necessarily imply (but does not exclude) physical separation among niches.

[7]Metagenome is the total set of all the genomic DNA in a particular environment or sample.

[8]Herein, population refers to a group of individuals sharing genetic and ecological similarity, and coexisting in a sympatric setting.

The idea that genotypic clusters should be rapidly eroded by HGT might in part be an artifact of early comparative studies of quite anciently diverged genomes. In these, only a fraction of genes showed phylogenetic congruence, whereas the majority seemed to be completely unrelated (Welch et al. 2002; Doolittle and Papke 2006). Moreover, we often call organisms closely related if their 16S rRNA genes, which are commonly used as taxonomic markers, show few percent nucleotide differences, yet such difference may indicate millions of years of separate evolution with associated large genome changes (Kettler et al. 2007). But even as closely related genomes (e.g., identical in 16S rRNA genes) began to be sequenced, these usually were not isolated from the same habitat and, hence, were not part of the same populations of potentially interacting genotypes. This means that the effect of environmental selection might not be easily disentangled from genetic divergence caused by geographic separation (Cordero and Polz 2014). For example, in the marine cyanobacterium *Prochlorococcus* populations in the Atlantic contain genes responsible for efficient phosphorus acquisition that are absent from populations in the Pacific (Coleman and Chisholm 2010). Hence, these genes are part of the core genome[9] (i.e., genes present in all) of Atlantic populations but would be judged flexible genes (i.e., genes present only in a subset) if closely related isolates were compared from both ocean regions. We, therefore, believe that an important step forward will be to emphasize population thinking in microbiology by assembling genomic datasets that represent clusters of close relatives co-occurring in the same environment—as only these will allow interpretation of how environmental selection acts on genomes from within the same population.

The challenge is then to develop an understanding of how genotypic clusters originate and are maintained, and whether they are selectively optimized to occupy sufficiently different niches to coexist with other clusters. Importantly, any such attempt needs to take into account the considerable genotypic diversity encountered in environmental populations, which often consist of genomes differing by a considerable fraction of their gene content and displaying large allelic diversity even if most of their genes suggest close relationships (Cordero and Polz 2014).

In this review, we begin by discussing the extent to which ecological and genetic units overlap, and under what circumstances genetic units can be used as a proxy for ecological units. We argue that, although it is essential to sequence populations of microbial genomes and record ecological metadata, a powerful alternative is represented by a "reverse ecology" approach, in which genomic and gene flow information is used to make predictions about the nature of ecological units (Box 1). What distinguishes reverse ecology from the broader field of ecological genomics is its focus on simultaneously predicting ecological and genetic units, rather than mapping ecological data onto predefined genetic units. These predictions can then be tested using ecological metadata and experimental follow-up. Then we describe in detail two examples of reverse ecology applied to different closely related, sympatric,[10] natural microbial populations. We synthesize conclusions from these examples, along with data from more distantly related genomic comparisons and evolutionary models, and propose a process of speciation that can operate under different regimes of selection and recombination (Fig. 1). Early stages of the speciation process are driven by either gene-specific[11] or genome-wide selective sweeps[12] as microbes adapt to

[9]Core genome is the portion of the genome that is present (or in practice, that can be aligned) in all of a given set of sequenced isolates or metagenomes.

[10]Sympatric means a set of sampled isolates or genomes from the same geographic area, in which barriers to migration and gene flow are low or nonexistent.

[11]The gene-specific selective sweep is the process in which an adaptive gene or allele (possibly a niche-specifying variant) spreads in a population by recombination faster than by clonal expansion. The result is that the adaptive variant is present in more than a single clonal background, and that diversity is not purged genome-wide.

[12]Genome-wide selective sweep is the process in which an adaptive gene or allele (possibly a niche-specifying variant) spreads in a population by clonal expansion of the genome

BOX 1. HOW TO PERFORM A REVERSE ECOLOGY POPULATION GENOMIC STUDY

1. The goal of the reverse ecology approach is to determine whether a sample of closely related, sympatric genome sequences constitute one or more genotypic units, and to test how these units might differ in their ecology either by mapping of these clusters onto environmental gradients or patches, or by laboratory tests. The sampling scheme need not be entirely unbiased. For example, isolates should be intentionally chosen to be closely related. They could also be chosen from two or more hypothesized niches or phenotypic groups to test whether these groups behave as separate genotypic units (Box 2), and to uncover the genes or mutations that might contribute to their ecological differences. Isolates should, however, be sampled from the same geographic location to reduce the effects of allopatric divergence and focus on the effects of local selection and recombination. Some a priori information—perhaps from a previous phylogenetic or metagenomic survey—may also be required to select a subset of closely related populations from the community.
2. Choose a genomic or metagenomic approach. Whole-genome sequencing of cultured isolates or isolated (but uncultured) single cells is preferable because it reveals information about how genes and mutations are linked within genomes, facilitating inferences about recombination events among genomes. Metagenomic sequencing has the advantage of sampling more individuals within an environment than are generally possible to isolate, but linkage information will be limited by the sequencing read length and quality of the assembly. Most importantly, unbiased metagenomic sequencing will only provide an appropriate population genomic dataset for populations that are relatively abundant in the sampled environment. The power of metagenomic data can be boosted significantly if they are gathered as a time series. Although such data sets are currently rare and potentially challenging to collect, they can follow the speciation process (Fig. 1) in real time, and potentially catch selective sweeps and niche-specifying events in action. Fine-grained time series might also follow shifting ecological conditions over time, revealing independent behaviors of different clusters.
3. Assemble genome sequences. Complete genome sequences are more readily assembled from isolates, but assembly can also be attempted on metagenomic data, taking care to guard against or account for different individuals being coassembled into a single genome.
4. Align genome sequences and define core and flexible components. Here, particular care must be taken to only define these categories for organisms that co-occur and, hence, have the potential to be connected by contemporary gene flow and be subject to consistent environmental selection.
5. Evaluate phylogenetic signals in single nucleotide polymorphisms (SNPs) found in the core genome. Standard phylogenetic methods can be used to build a core genome-wide phylogeny, and the average impact of recombination can be measured by assessing linkage disequilibrium among SNPs. Specific recombination events and breakpoints can then be identified using methods, such as BratNextGen (Marttinen et al. 2012), ClonalFrame/ClonalOrigin (Didelot et al. 2010), and STARRInIGHTS (Shapiro et al. 2012). These analyses will reveal the number of major genotypic units (well-supported monophyletic groups), and whether these units are supported genome-wide (consistent with mostly clonal evolution) or in "islands" or "continents" of the genome.
6. If populations were hypothesized a priori based on an ecological axis of interest, assess whether these presumed populations correspond to genotypic clusters or not. If genome-wide diversity is clustered according to ecology, this suggests that stable clusters have formed (Fig. 1, stages 4–5). If there is little or no phylogenetic clustering according to ecology, the hypothesized populations likely constitute a single, phenotypically diverse population. In this case, certain (flexible) genes or (core) mutations that associate with ecology might be identified by GWAS (Fig. 6). If there is a preference for recombination within rather than between ecological groups, the single population might be on a trajectory toward speciation (Fig. 1, stage 3).

Continued

If populations were not hypothesized a priori (a "purer" reverse ecology approach), assess how many phylogenetic groups were identified. If phylogenetic groupings are supported genome-wide, this suggests stable differentiation (Fig. 1, stages 4–5), the ecological basis of which remains unknown but can be tested by phenotypically characterizing representative isolates from each group and/or mapping genotypic clusters onto environmental samples. If groupings are not supported genome-wide, genomic regions containing the bulk of the phylogenetic signal, or signals, of positive selection, frequent recombination, or dense polymorphism, can be functionally annotated to generate hypotheses about their possible ecological roles.

new sympatric niches, with either high or low levels of recombination among niches, respectively. At more advanced stages, if barriers to gene flow among niches emerge, distinct units of microbial diversity come into focus and can potentially be recognized (Box 2). As we have reasoned previously, units can be defined operationally in cases in which both genotypic and phenotypic variance is much greater between than within such units (Polz et al. 2006). We wish to make a strong distinction between this process of speciation—which we define as any stage of the dynamic process of ecological and genetic differentiation—and the concept of species, which we are not attempting to address. Speciation need not proceed to completion, and recognizable units of genotypic and ecological similarity will often contain abundant genetic and phenotypic diversity within them. We conclude by briefly highlighting the potential for reverse ecology to identify natural units of microbial diversity, and for genome-wide association studies[13] (GWAS) to identify mutations and genes underlying ecologically relevant traits.

DEFINING GENETIC AND ECOLOGICAL UNITS

Ecological units, in the most basic sense, denote groups of organisms with common ecological functions. It is immediately obvious that this definition represents an abstraction by the observer and is hence subject to individual preferences of how finely one wishes to demarcate units (Jax 2006). For example, does the acquisition of an antibiotic resistance gene generate a new ecological unit or simply a variant within an existing unit? Do all sulfate-reducing bacteria represent one ecological unit because they all carry out a common, highly relevant environmental function? In other words, is an ecotype (defined here as ecologically completely equivalent genotypes) the right unit, or should we define ecological units more broadly? To understand the genetic basis of ecological preferences, microbiologists will generally make educated guesses about important and measurable dimensions of niche space (e.g., host preference and ability to grow on a particular carbon source) and embark on a population genomics study. Similar to classical, trait-based taxonomy, this approach is potentially subject to arbitrary weighing of phenotypes to define an ecological unit. An alternative approach is to avoid a priori guesses as much as possible, and sample closely related microbes, identify genomic units among them, and make hypotheses about their ecological differences (if any) based on the predicted or experimentally validated effects of these genomic differences or, as we will argue later, based on patterns of gene flow. We refer to this as a reverse ecology approach (Box 1) (Li et al. 2008; Ellison et al. 2011; Levy and Borenstein 2012). If these genomic units correspond to natural populations, this approach also provides the opportunity to test hypotheses about the evolutionary mechanism creating and maintaining diversity within and between such genomic units.

that first acquired it. The result is that diversity is purged genome-wide, and that the adaptive variant is linked in the same clonal frame as the rest of the genome.

[13]Genome-wide association study (GWAS) is a technique commonly used in eukaryotic genetics to identify genomic variants that are associated with a phenotype of interest. In highly structured populations (e.g., clonal microbes), it is essential to correct for false associations owing to phylogenetic structure.

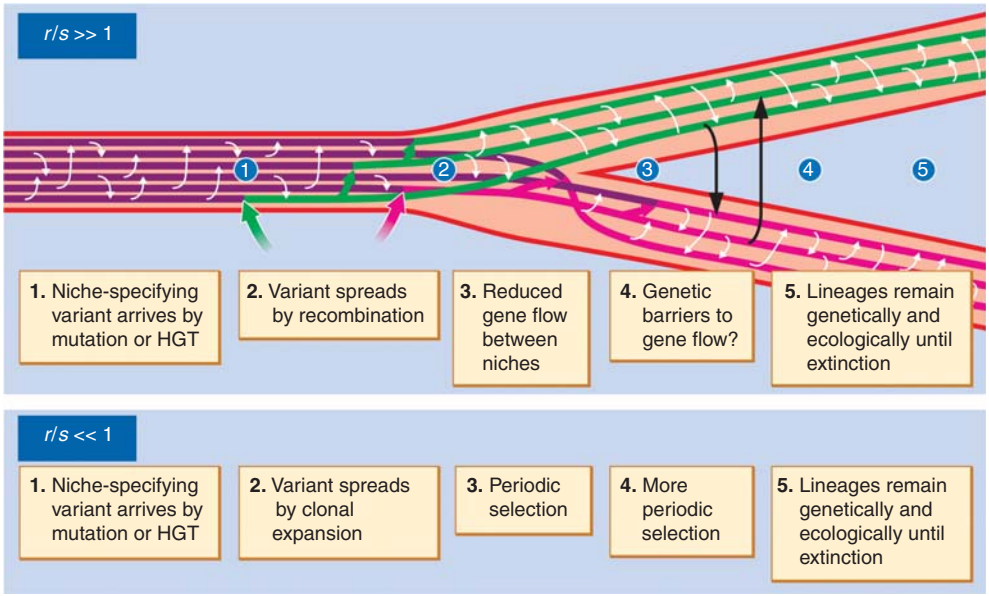

Figure 1. A model for bacterial speciation under different recombination/selection balances. Five stages in the process of speciation are illustrated under the $r/s \gg 1$ regime (*top*). The illustration shows an ancestral population of bacteria (purple), which diverges into two incipient species (red or green) adapted to different habitats following the acquisition of niche-specifying genes (red or green arrows). Thin gray or black arrows show recombination events within and between incipient species, respectively. In the $r/s \ll 1$ regime (*bottom*, not illustrated), stages 2–4 differ, but the start and end points are the same. The $r/s \ll 1$ regime corresponds closely to the stable ecotype model (S.E.M.) (Box 3). Selection (s) is defined here as the average fitness difference experienced by a niche-specifying (adaptive) allele in different niches and recombination (r) is the recombination rate per locus per generation. The stages represent rough, potentially overlapping, and potentially terminal steps (e.g., stage 2 need not lead to stage 3). HGT, Horizontal gene transfer. (From Shapiro et al. 2012; adapted, with permission, from the author.)

Genotypes are in principle easier to delineate than ecological types because a cell's genome can be measured by sequencing, whereas it is not clear how many phenotypic properties have to be measured before a cell's ecology is exhaustively captured. However, defining genetic units suffers from similar problems as for ecological units because it is not clear what measure of similarity to use and where to draw the bounds. Genotypes can be grouped into units based on various measures of genomic similarity: DNA hybridization assays, percent similarity in a marker gene, average nucleotide identity (ANI) across the genome, or the proportion of shared genes (Konstantinidis and Tiedje 2005). These are all convenient measures, but they all require that we decide on a cutoff value to divide units, often referred to as operational taxonomic units (OTUs).

An alternative is to search for natural genetic units based on the mechanisms capable of clustering genetic diversity, namely migration, mutation, recombination, and selection. As detailed in Box 1, this can be performed with a relatively limited sample of genomes from the same environment if one focuses on a defined taxonomic group. We exclude from this article patterns of biogeography that arise because of allopatric[14] speciation, and we refer the reader to excellent recent work on the topic (Denef et al. 2010a; Nemergut et al. 2010; Hanson et al. 2012). Although allopatric speciation is conceptually more straightforward, sympatric speciation is thought to be more common in

[14]Allopatric refers to a set of sampled isolates or genomes from different geographic areas, in which barriers to migration and gene flow are significant.

BOX 2. CHALLENGES IN IDENTIFYING NATURAL UNITS OF MICROBIAL DIVERSITY

To determine whether two hypothesized units are indeed distinct, one must reject the hypothesis that they are both part of the same unit. This means that they must differ in at least one ecological dimension, and must show more genome-wide cohesion within than between units. The cohesion could be caused by higher rates of recombination within than between populations, or caused by independent genome-wide selective sweeps occurring in each population, without significant recombination. Therefore, natural units of genome-wide and ecological similarity can be produced under different regimes of selection and recombination (Fig. 1). Importantly, no absolute cutoff (for either genetic or ecological similarity) is necessary to define these units.

One challenge to overcome in identifying natural units is that some degree of recombination is to be expected between separate units, which might exchange "globally adaptive" genes or alleles, while remaining separate elsewhere in the genome (Majewski and Cohan 1999b). For example, different species of *Campylobacter* exchange genes of certain functions only, while remaining distinct throughout most of the genome (Caro-Quintero et al. 2009; Caro-Quintero and Konstantinidis 2011). In *Streptococcus*, cross-species exchange is often accompanied by positive selection (high dN/dS in the exchanged genes [Shapiro et al. 2009]). This suggests that biased cellular functions and positive selection might be general features of globally adaptive genes, allowing them to be recognized and excluded from phylogenetic or recombination-based tests for separation between units.

A second challenge is that a single cohesive population may still contain significant phenotypic and genotypic variation, but this variation will be restricted to only a relatively small fraction of the core and flexible genome. For example, genes under diversifying or frequency-dependent selection "within" single cohesive populations might be mistaken for niche-specifying genes driving adaptation "between" populations (Fig. 5). With careful application of population-genetic tests for natural selection, combined with phenotypic characterization of these genes, it is possible (if challenging) to distinguish between these two scenarios.

microbial populations (Vos 2011). We therefore focus on recombination and selection in sympatric settings.

MODELING THE INTERPLAY OF SELECTION AND RECOMBINATION

In answering how genotypic clusters originate and are maintained, it is critical to evaluate the interplay of recombination and selection, both of which can vary widely. But, although recombination rates can be measured to some extent, the magnitude of selection is difficult to assess directly in the wild, so that we have to rely on reasonable guesses. Below, we give an overview of current knowledge of recombination rates, and then show how mathematical models that explicitly incorporate recombination have been used to explore (1) the probability that clusters arise in sympatry caused by neutral processes, and (2) the effect of different recombination and selection rates on the spread of adaptive loci or alleles within and across populations.

As noted above, homologous recombination rates can vary tremendously in different lineages of bacteria and archaea, with some evolving in a highly clonal fashion, whereas others are considered sexual, with recombination rates up to 10-fold higher than mutation rates (Smith et al. 1993), resulting in >10-fold more polymorphism from recombination than mutation (Vos and Didelot 2009). Some bacteria, such as *Vibrio*, *Streptococcus*, and *Helicobacter*, tend to recombine frequently, whereas others, such as *Bacillus*, *Staphylococcus*, and *Mycobacterium*, tend to be more clonal. It is, however, likely that most measured recombination rates are underestimates because typical analyses allow inference of recombination only when highly polymorphic segments of DNA are observed. Hence, these measured rates might give a fairly accurate picture of recombination between but not within clusters. Moreover, experimental observations have suggested that the frequency of recombination drops exponential-

BOX 3. THE STABLE ECOTYPE MODEL

The stable ecotype model (SEM) of speciation, as developed by Cohan, invokes a prominent role for natural selection to form and maintain separate genetic clusters. It also provides an appealing mechanistic link between ecological and genetic units. In its basic form, an ecotype can be understood as the domain of competitive superiority of an adaptive mutant (Cohan 2001). When an adaptive mutant arises within an ecotype population, it outcompetes its neighbors, purging diversity in a periodic selection event (Fig. 1, lower panel). Importantly, diversity is not purged in other ecotypes, which compete in independent niches. Ecotypes are also subject to neutral mutation and drift, which, along with periodic selection events, result in separate clusters of ecological and genetic diversity. The model states that observed rates of recombination are not high enough to unlink adaptive and neutral loci in the genome. Therefore, periodic selection is predicted to purge diversity genome-wide.

Although highly plausible, the SEM, in its basic form, has not yet been directly observed in nature. Reasons for this might include recombination rates being underestimated, and niche complementarity maintaining multiple genotypes within a population. Support for the SEM has come from experimental evolution studies, in which diversity can only be generated by mutation within a restricted population, without the possibility for recombination with distant relatives. In these studies, adaptive mutations increase in frequency, eventually reaching fixation on a single genomic background, along with neutral "hitchhiking" mutations (Barrick et al. 2009). Some mutations may found a new ecotype by allowing colonization of a new niche (Koeppel et al. 2013), followed by successive genome-wide sweeps in the new ecotype.

Variants of the SEM that allow a more prominent role for recombination have also been proposed. For example, the "adapt globally" model allows globally adaptive genes to spread by recombination across multiple ecotypes, without affecting their ecological distinctness (Majewski and Cohan 1999b). This model could accommodate, for example, the transfer of an antibiotic resistance gene from *Enterococcus* to *Staphylococcus* without insisting on merging them into the same species.

ly with sequence divergence because of the requirement of a 20-bp stretch of identical DNA sequence for efficient initiation of recombination (Vulić et al. 1997; Majewski and Cohan 1999a). Such a rapid drop in frequency should limit efficient exchange of DNA to closely related genomes, as expected within genotypic clusters, and might play a role in maintaining the cohesion of clusters. Although such relationships have been shown for several, divergent groups of bacteria, in some archaea, the requirement for short, identical DNA stretches seems to be absent (Grogan and Stengel 2008; Naor et al. 2012), even though environmental observations support decreased rates of recombination across clusters (Whitaker et al. 2005; Williams et al. 2012). Moreover, recent comparison of very closely related genomes has also shown that very little sequence similarity appears to be required for integration of long stretches of highly divergent DNA (including single nucleotide changes and structural variants) into the genome (Mell et al. 2011; Cordero et al. 2012a), although the mechanisms remain unclear. These recent results show that much remains to be learned about how recombination proceeds in different groups of bacteria and archaea, making mathematical models an important tool to explore potential outcomes, given reasonable assumptions about the importance of recombination relative to mutation and selection.

Whether genotypic clusters can arise neutrally in sympatry was addressed with a simple computational model starting with a single population that evolves by mutation and varying degrees of recombination, but in the absence of selection (Fraser et al. 2007). Without recombination, clonal clusters emerge by random mutation, but quickly drift to extinction. Because these clusters are short lived, they accumulate very little sequence diversity and would be hard to recognize in samples of microbes. When recombination rates become more frequent than mutation rates, however, clusters

no longer emerge, and the population remains homogenous. A critical parameter in this model is the decline in the rate of homologous recombination with sequence divergence. Separate clusters are only formed if the rate of decline of recombination with mutational divergence is unrealistically high compared with those observed experimentally (Vulić et al. 1997; Majewski and Cohan 1999a; Mell et al. 2011; Croucher et al. 2012). Hence, the model suggests that natural selection should be required to produce stable genotypic clusters, and that neutral cluster formation is extremely unlikely—a prediction that is borne out in long-term microbial experimental evolution studies. These studies have provided evidence that most fixed mutations tend to be adaptive, not neutral (Barrick et al. 2009), and that formation of new genotypic clusters might involve adaptation to using novel resources (Blount et al. 2012).

Building on these results, a model was developed that includes one or more loci under selection, conferring adaptation to either of two sympatric niches, which are completely geographically overlapping, ensuring frequent mixing of all genotypes and an equal probability of sharing genes (Shapiro et al. 2009; Friedman et al. 2013). In this sympatric simulation (symsim) model, niche adaptation is encoded by genes or alleles already segregating within, or recently horizontally transferred into the population, with de novo adaptive mutation assumed to be negligible. The symsim model readily describes the simple case in which niches correspond to two different carbon sources dissolved in a single well-mixed aquatic environment. With rates of recombination much higher than selection ($r/s \gg 1$), diversity at any neutral locus was unaffected by a selective sweep of an adaptive locus (Shapiro et al. 2009). Although at first glance, this scenario seems unlikely because observed rates of recombination are typically much lower than even moderate rates of selection, positive selection[15] might be depressed by high rates of frequency dependent selection (detailed further below) and current data on recombination likely represent underestimates (detailed further above). In the contrasting scenario, when selection coefficients are much higher than recombination rates ($r/s \ll 1$), an adaptive allele will generally sweep to fixation on a single genetic background, homogenizing neutral variation, as in the stable ecotype model (SEM) (Box 3). However, even with $r/s \ll 1$, given enough time before any further selective events, and assuming that the two niches remain sympatric, neutral alleles will eventually become randomly distributed across genotypes, with only adaptive alleles being selectively maintained (Friedman et al. 2013). Moreover, when the selective coefficient is distributed across more than one adaptive locus, this reduces the effective strength of selection and results in even stronger homogenization of neutral loci (and even to some extent, adaptive loci) across niches.

Hence, the model shows that although adaptation spreads differently in the two regimes of recombination versus selection, the eventual outcome is similar, that is, recombination will eventually homogenize perfectly sympatric genotypes even if they carry niche-specific adaptations. The important consequence is that some kind of microgeographic separation between niches, akin to the "mosaic sympatry" described by Mallet (2008), might be required to reduce gene flow between niche-adapted genotypes before clusters of selectively neutral genome-wide diversity may develop. Mosaic sympatry essentially means that niches are distributed patchily, without being completely allopatric (Mallet 2008). This situation might readily describe many microbial environments, such as soil, oceans, and animal hosts, in which resources are distributed in small-scale patches, but patches may be short-lived and colonizing populations may mix frequently because of the need to recolonize new patches (Polz et al. 2006). Barriers to gene flow might also arise because of incompatible restriction modification or competence peptide systems yielding a form of mosaic sympatry, although empirical evidence that either system actually promotes speciation is lacking (Hanage et al. 2005; Cornejo et al. 2010). Overall, there is a growing consensus

[15]Positive selection refers to a type of natural selection that favors variants conferring a fitness advantage, causing them to increase in frequency in a population.

that bacterial speciation generally takes place in sympatric or mosaic sympatric settings (Vos 2011).

Taken together, these models suggest, first, that in the absence of selection, neither clonal nor sexual populations will split into stable, sympatric genotypic clusters because of neutral processes. Second, selection on niche-specifying variants[16] should be accompanied, or followed by, habitat separation for genetic exchange to be reduced across the genome. With $r/s \ll 1$, stable clusters of ecological and genome-wide similarity can develop quickly (as in the SEM, Box 3), and can remain distinct if gene flow is impeded by habitat partitioning. With $r/s \gg 1$, genotypic clusters of distinct ecology would take longer to establish because the gradual accumulation of sequence diversity by the interplay of population-specific mutation and recombination is required for distinct genetic clusters to emerge (Polz et al. 2013).

Important further predictions of these models are, first, that if we observe co-occurring genotypic clusters, these should be ecologically distinct (even if they are closely related). This prediction has largely been supported by surveys of genetic diversity in the wild that identified clusters with overlapping genetic and ecological similarity (Rocap et al. 2003; Johnson et al. 2006; Hunt et al. 2008; Koeppel et al. 2008; Konstantinidis and Delong 2008; Denef et al. 2010a). Second, for a new niche-specifying gene or allele to induce habitat separation, there must be some form of tradeoff that reduces its success in the former habitat while increasing it in the new (Wiedenbeck and Cohan 2011). In the absence of such tradeoffs, an ecological generalist might evolve that is successful in both habitats and remains cohesive by gene flow. As shown below, we have recently detected two nascent populations that appear to have evolved an ecological tradeoff explaining their distribution (Yawata et al. 2014).

[16]Niche-specifying variant is a mutation, gene or allele that allows a cell to be part of a particular niche. These variants are under positive selection within the particular niche, but not outside it.

GENOMICS OF NASCENT CLUSTERS

As suggested by Wiedenbeck and Cohan (2011), detailed investigations of the very early stages of ecological differentiation—whether or not it proceeds to completion—are essential to understand the interplay of recombination and selection in generating ecological and genetic units. We discuss two such snapshots of slightly different stages in this dynamic process.

In the first, 20 *Vibrio cyclitrophicus* genomes with identical 16S rRNA genes were sequenced and found to share >99% amino acid identity genome-wide. Despite being so genetically similar, two separate groups with distinct ecological preferences were recognized: isolates associated with organic particles and those free-living in coastal ocean water (Shapiro et al. 2012). These distinct lifestyles are made possible by the patchy distribution of resources in the ocean, which might promote a form of mosaic sympatry. In the second investigation, Cadillo-Quiroz et al. (2012) sampled thermophilic *Sulfolobus* archaea from a hot spring in Kamchatka, Russia, without any prior knowledge of niche preferences. They then followed a reverse ecology approach to identify two genetically distinct, but closely related groups of *Sulfolobus*, which they later found to differ phenotypically. Both studies used whole-genome sequencing of coexisting, closely related sympatric microbial populations to infer mechanisms of speciation in nature, and the two reports were published in 2012. The studies differed in that the first study had an a priori notion of ecological association for the two *Vibrio* populations because of the sequencing of a gene under potential environmental selection (Hunt et al. 2008; Shapiro et al. 2012), whereas Cadillo-Quiroz et al. (2012) took a purer reverse ecology approach (Box 1), identifying two phylogenetic groups based on overall genomic similarity, then investigating recombination rates within and between groups, and characterizing phenotypic differences between them.

Both studies identified distinct regions of the genome containing single nucleotide polymorphisms (SNPs) clearly dividing the two groups of isolates. In Shapiro et al. (2012), these

were referred to as ecoSNPs, because they were fixed genetic differences between groups with previously known ecological associations. Here, we refer to them more inclusively as "divergent SNPs" (divSNPs)—a term that can also be applied to the *Sulfolobus* populations because in these, association with ecological differentiation is still unclear. In the *Vibrio* genomes, the divSNPs were localized in densely clustered "islands," whereas divSNPs were both more numerous and more broadly dispersed across "continents" of the *Sulfolobus* genomes (Fig. 2), likely reflecting a more advanced stage of differentiation (Fig. 1). (Here we use the terms "islands" and "continents" in a metaphorical sense, not in a biogeographical sense.) Many of the divSNPs in the *Sulfolobus* continents are probably not directly involved in ecological adaptation, and might be hitchhiking with putative adaptive variants. The extent of this divergence hitchhiking (Via 2012) is much smaller in the *Vibrio* islands, which are rich in ecologically relevant genes, such as those involved in stress responses, attachment, and biofilm formation (Shapiro et al. 2012). Outside of these islands or continents, both studies found poorly resolved phylogenetic separation—in fact, a plethora of distinct and conflicting phylogenies across the genome—and shared genetic diversity between groups, indicated by low fixation indices (F_{ST}). This provided evidence for a history of rampant recombination among all sympatric isolates, not just those sharing an ecological preference.

Do these observations support genome-wide or gene-specific selective sweeps (Fig. 3), and what are the implications for ecological/genetic units? One possibility is that genome-wide sweeps did occur in each habitat, but the clonal frames[17] were gradually eroded by recombination of neutral loci between habitats, leaving behind islands or continents as the only traces of the ancient clonal divergence. Modeling suggests that this gradual erosion would require several thousand generations (Friedman et al. 2013), over which time the islands or continents of divSNPs (containing the habitat-specific alleles) would accumulate polymorphism within populations. Yet, in the *Vibrio* genomes, most of the habitat-specific alleles show very low polymorphism and high synonymous divergence between habitats. This suggests their recent acquisition by recombination from more distant relatives, rather than being the remnant of a more ancient genome-wide sweep. Moreover, a genome-wide sweep would not explain the presence of the same habitat-specific allele in different clonal frames, which was observed to be the case at the RpoS/RTX locus (Shapiro et al. 2012). These observations show that niche-specifying genes or alleles may reside in different genotypes that are otherwise homogenized by gene flow (Fig. 3).

Genome-wide sweeps were not as firmly excluded in the *Sulfolobus* populations, although deemed unlikely based on the relatively high inferred recombination rates among populations (Cadillo-Quiroz et al. 2012). As discussed above, archaea, like *Sulfolobus*, which lack mismatch repair machinery, generally show very little reduction in recombination as sequence divergence increases, suggesting weak barriers to recombination between incipient clusters, favoring gene-specific sweeps. Hence, very strong divergent selection would be required for the populations to have diverged before much recombination occurred between them, yielding an effectively genome-wide selective sweep. Even in the absence of selection, however, the populations could have diverged in allopatry before reencountering each other in the same hotspring. Either way, the resulting clonal frame would be observable as large continents of divergence between populations (Fig. 2), interrupted by recombination events following the clonal divergence (a scenario not excluded by Cadillo-Quiroz et al. 2012).

Although the *Sulfolobus* populations appear to behave as two distinct genetic units, at a first glance, the two ecological populations identified in *V. cyclitrophicus* are contained within a single genotypic cluster that appears thoroughly mixed by recombination at all except the

[17]A clonal frame is the portion of the genome transmitted by vertical (clonal) evolution, unimpacted by HGT. Mutations in the clonal frame should all fall parsimoniously on a single phylogenetic tree.

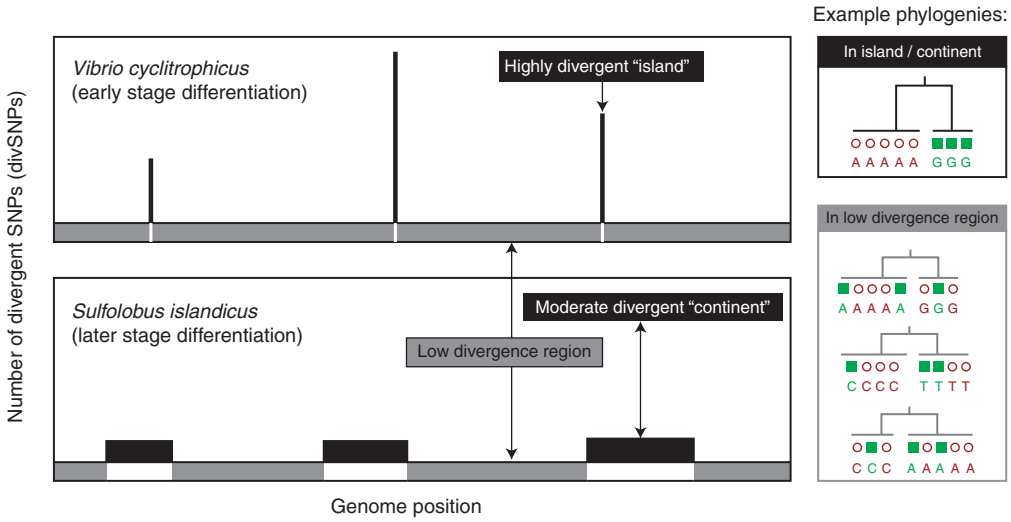

Figure 2. Islands and continents of speciation. Based on data from aligned core genomes of *Vibrio cyclitrophicus* (Shapiro et al. 2012) (*upper* panel) and *Sulfolobus islandicus* (Cadillo-Quiroz et al. 2012) (*lower* panel), the distribution of divergent single nucleotide polymorphisms (SNPs) (divSNPs) fixed between populations is plotted along the genome as black bars. *Vibrio* and *Sulfolobus* contain, respectively, 725 divSNPs distributed more than ~1% of the genome, and 4232 divSNPs distributed more than ~36% of the genome. Scale is approximate for divSNPs (*y*-axis) and genome position (*x*-axis). The *y*-axis is not to scale for SNPs rejecting differentiation between populations (regions of the genome shown in gray).

divSNP-containing loci (Fig. 2). This picture of the vibrios as a single gene-flow unit changes, however, when inferred recombination events are separated into more ancient and more recent ones—those that have presumably occurred before and after the ecological split, respectively. Such analysis shows that the more recent events are biased to occur among genotypes within either of the two habitats, whereas more ancient events connect all genotypes. This suggests an evolutionary trajectory, most likely induced by microhabitat separation, from a single freely recombining population toward two increasingly separate gene pools. The same trend was observed in both the core and flexible components of the population genomes of vibrios and *Sulfolobus* alike, and, if projected into the future, might lead to the evolution of clearly distinct genotypic clusters.

As predicted in the models described above, the habitat separation of the two nascent *Vibrio* populations appears to be associated with an ecological tradeoff. Although one population specializes in organic particle exploitation through strong attachment and growth in biofilms, the other population only rarely attaches, yet is specialized for dispersal by rapidly detecting and swimming toward new particles (Fig. 4), implying that it can better exploit short-lived nutrient patches (Yawata et al. 2014). Based on their genetic distinctness, we would also predict the ecological distinctness of the two *Sulfolobus* populations. Indeed, Cadillo-Quiroz et al. (2012) went on to show that the two populations differed in growth characteristics in the laboratory, suggesting distinct niches. It is not yet clear whether these growth differences are relevant to fitness in the wild, and further research will be needed to exclude the possibility that they evolved as a consequence of neutral divergence in allopatry. However, assuming that sequence divergence does not present a significant barrier to gene flow, the preference for recombination within, rather than between, *Sulfolobus* populations is likely driven at least in part by differences in ecological associations. Hence, these examples of closely related populations suggest that, in

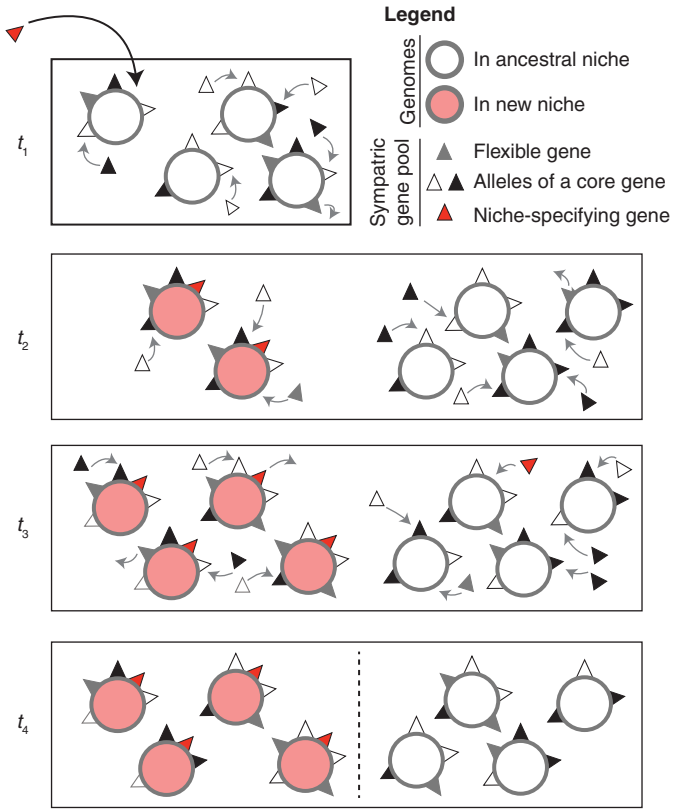

Figure 3. Ecological differentiation via a gene-specific sweep. This illustration follows the basic steps of the symsim model (Friedman et al. 2013). At the first time point (t_1), a niche-specifying gene (red triangle) arrives into a homogeneously recombining population occupying a single niche. Between t_1 and t_2 (as between all time points), recombination (r) occurs at random from a sympatric pool (one to two events per genome, illustrated as arrows), then genomes reproduce clonally, and are culled to a carrying capacity of four genomes per niche. Because the niche-specifying gene confers a selective advantage(s) in the new niche, genomes that contain it grow exponentially until the carrying capacity is reached at t_3. Other genomes are culled at random, because the rest of the gene pool is neutral to fitness. By t_4, the gene-specific sweep is complete. The niche-specifying gene is in perfect association with the new niche, but all other genes are randomly distributed across niches. At this point, barriers to recombination between niches (dashed line) may or may not emerge. (Note that recombination events at t_4 are not shown for purposes of clarity, but this does not mean they do not occur.)

recombining microbes, gene flow barriers may help maintain established units (*Sulfolobus*), and may initiate formation of new units (*Vibrio*). The *Vibrio* example, in particular, further suggests that genes can spread in a population specific manner and, perhaps, initiate microgeographic structure.

This model also helps explain previous findings in genomic and metagenomic surveys that have found location-specific genes or alleles in genotypic clusters that are broadly distributed but seem otherwise phylogenetically "well-mixed" in neutral genes across the genome (Papke et al. 2007; Coleman and Chisholm 2010; Denef et al. 2010b; Boucher et al. 2011; Burke et al. 2011). Although such mixing may be expected in microbes separated by only a few microns in a biofilm (Denef et al. 2010b), it is perhaps more surprising that *Vibrio cholerae* from different continents (Boucher et al. 2011), *Prochloroccous* from different oceans (Coleman and Chisholm 2010), or even haloarchaea separated by a few hundred kilometers (Papke et al. 2007) remain cohesive at neutral

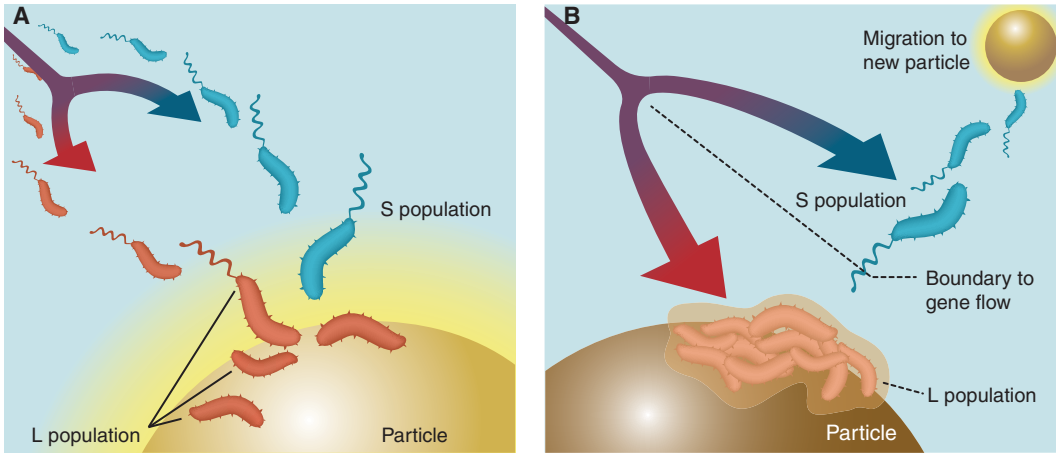

Figure 4. Barriers to recombination emerge via a competition-dispersal tradeoff. (*A*) The particle-associated *Vibrio* population (L) attaches to nutrient-rich particles and forms biofilms, whereas the free-living *Vibrio* population (S) hovers near the surface and scavenges loose nutrients. (*B*) When a new particle becomes available, only the S population is able to rapidly disperse to the new nutrient source. (From Yawata et al. 2014; adapted, with permission, from the author.)

loci. This may be explained by at least a few genotypes being mixed across geographic distances in every generation, allowing allopatric populations to remain homogenous outside of a few environment-specific loci under location-specific selection (e.g., *Prochlorococcus* phosphorus utilization genes discussed above). In other cases, geographic distance has been correlated to significant divergence between populations (Reno et al. 2009), but such divergence need not be permanent once gene flow barriers are removed. For example, clonal divergence was reported between *Leptospirillum* populations that had been separated for ~1000 yr (Simmons et al. 2008; Denef et al. 2010a). However, once the separation ended because of commercial mining activities, it took only ~150 yr (~100,000 bacterial generations) for the incipient populations to become well mixed by recombination of neutral loci across the genome (Simmons et al. 2008).

In summary, these considerations suggest that gene-specific (Fig. 3), rather than genome-wide (clonal) selective sweeps may be more common in nature than previously thought. As we discuss below, such gene-specific mechanisms begin with poor mapping between ecological preference and neutral genetic diversity, but eventually result in tight ecological and genetic units.

STAGES IN THE SPECIATION SPECTRUM

The snapshots described above suggest a gradual process by which a new niche becomes accessible when novel genes or alleles arise by mutation or HGT in an ancestral population (Fig. 1, stage 1). With sufficiently high recombination rates relative to selection, this niche-specifying variant will spread in a gene-specific sweep (stage 2). If the new and ancestral niches remain fully sympatric, with no barriers to recombination between them, the process will stop here, as in the symsim model described above (Friedman et al. 2013). However, if the new niche is also somehow associated with barriers to recombination (perhaps because of a reduced encounter rate with genomes in the ancestral niche), genetic separation will begin to occur at neutral loci throughout the genome (stage 3). These ecological barriers might later be reinforced by genetic barriers, as sequence divergence accumulates between lineages, eventually inhibiting recombination genome-wide (stage 4). Genetic isolation may also develop more quickly if the capacity for recombination

is transiently lost, either genetically (e.g., Katz et al. 2013) or physiologically (e.g., by modulating expression of recombination and mismatch repair machinery).

Whether the early stages of speciation involve gene-specific or genome-wide selective sweeps will depend on the r/s ratio, but both regimes can eventually lead to the same end products of overlapping genotypic and ecological units (Fig. 1). The *Sulfolobus* populations may be in an intermediate regime, with a low enough r/s ratio to allow clonal sweeps at stage 2, but sufficient recombination to generate the patterns of gene flow observed at stages 3 and 4. Based on the number of conflicting phylogenetic signals in the genome, flexible genome[18] diversity, and presence of niche-specific genes in multiple different clonal frames, it appears that the vibrios are firmly in the $r/s \gg 1$ regime. Yet, this seems at odds with experimentally estimated recombination rates, which appear to be generally much lower than even moderate selection coefficients (Wiedenbeck and Cohan 2011).

What factors might keep r/s so high—astoundingly high, in fact, compared with our expectations? One possibility is that genome-wide selective sweeps are slowed by negative frequency-dependent selection (Cordero and Polz 2014), imposed on traits involved in susceptibility to phage predation (e.g., surface structures; Rodriguez-Valera et al. 2009) or in social interactions within microbial populations (e.g., siderophore or antibiotic production; Cordero et al. 2012a,b). Another possibility is that genome-wide sweeps are slowed by clonal interference (e.g., Lieberman et al. 2013), allowing more time for recombination to occur before all diversity is purged. Further research will be needed to distinguish between these possibilities.

We propose that the *Sulfolobus* lineages are approximately at stage 3 or 4, whereas the vibrios are at around stage 2 or 3. As a result, potential niche-specifying genes or alleles are much more readily pinpointed in the *Vibrio* islands than the *Sulfolobus* continents. Whether the nascent *Vibrio* lineages will persist cannot be predicted but we note that 3 yr after the initial sampling, the same populations with the same set of habitat-specific flexible genes were observed once again, suggesting a reasonably stable association between ecological units and selected parts of the genome (Szabó et al. 2013). However, we note that, at stage 2, the two nascent populations cannot be differentiated from a single population with the putative niche-specifying genes under balancing or negative frequency-dependent selection within the population (Cordero and Polz 2014; Shapiro 2014). Only with ecological information that shows poor habitat overlap and/or reduced recombination throughout the genome (stages 3–4) can we reasonably consider the population to be splitting in two.

As this split occurs, lineages might eventually become permanently separate, forming distinct ecological and genome-wide sequence clusters (at neutral loci across the genome, interrupted by occasional exchange between lineages), until one or both go extinct (stage 5). Importantly, this long-term result of formation of congruent ecological and genetic clusters is expected under both high and low r/s ratios. Therefore, comparing microbial genomes that have already reached this stage is not expected to be informative about the relative influence of selection and recombination at early stages. From a practical standpoint, once lineages have diverged to stage 5, they should be easily recognizable as distinct ecological and genetic units. For example, environmental and gut-associated groups of *Escherichia coli* have distinct ecologies and have diverged genetically throughout the genome (Luo et al. 2011). It therefore appears to be justifiable not to group these lineages together into the same species (Box 2).

We stress that, just because these stages of speciation can be defined, it does not mean that all populations that start at stage 1 will make it to stage 5. In fact, the intermediates may be more numerous than the end products. As James Mallet (2008) wrote, "speciation appears to be easy; the intermediate stages are all around us." If this is the case, many more examples

[18]Flexible genome is the set of genes or DNA that is present in only a fraction of a given set of sequenced isolates or metagenomes.

should be forthcoming from across the microbial world. Studies using the framework of population genomics and reverse ecology (Box 1) will test the generality of the speciation process that we propose based on current data. Moreover, as discussed in the next section, this proposed speciation process finds surprising parallels in new models of sympatric animal speciation.

ISLANDS OF SPECIATION IN THE GENOMIC ERA

In 2005, Turner, Hahn, and Nuzhdin (Turner et al. 2005) compared the genomes of the M and S forms of *Anopheles gambiae*, thought to be two incipient species of the malaria mosquito, adapted to different reproductive strategies. They found that the vast majority of the genome contained genetic diversity shared between these two forms—to be expected because they are not geographically separated and there are no physical barriers to genetic exchange. In other words, the M and S forms are sympatric rather than allopatric.

However, they identified three relatively small regions of the genome, which they called islands of speciation (Fig. 5), that were strongly genetically divergent between M and S forms. These islands are thought to contain genes that enable their host to adapt to different ecological niches, and are under divergent natural selection between nascent species. Although it remains controversial to what extent such islands are a cause or consequence of speciation (Turner and Hahn 2010; Pennisi 2014), and to what extent they are really small islands of a few adaptive genes or large "continents" of low recombination, the concept has gained support.

It is worth distinguishing "islands of speciation" from the distinct phenomenon of genomic islands observed in bacterial genomes and metagenomes. In islands of speciation, there is high genetic divergence between incipient species, whereas polymorphisms are shared across the rest of the genome. In bacterial genomes, islands are generally defined as regions of a reference genome, in which genomes from closely related isolates or metagenomic reads align at low frequency or not at all, because of high polymorphism in the island. The same phenomenon is observed with "pathogenicity islands," where inserted virulence factors in these islands can distinguish a pathogenic from a harmless variant.

Therefore, islands of speciation are divergent between species, whereas genomic or pathogenicity islands in bacteria are polymorphic within species (Fig. 5). The two types of islands are related because within-population polymorphism (genomic islands) can be shaped by natural selection and restricted recombination to yield between-species divergence (islands of speciation).

VARIATION WITHIN A COHESIVE POPULATION

Our discussion thus far has focused on how to define and delimit the boundaries between internally cohesive microbial populations. Cohesive populations may nevertheless contain high levels of genotypic (and to some extent, phenotypic) diversity within them (e.g., genomic islands; Fig. 5). How can this be explained?

First, as discussed earlier, niche-specifying variants (genes or alleles) may come with a fitness tradeoff, such that they are adaptive in one niche but not another (indeed, one might even define them as such). In a genetically cohesive population that spans two niches, different niche-specifying variants will be maintained in each niche, leading to variation at the level of the entire population. (In fact, without knowledge of habitat specificity and tendency toward within-habitat recombination, the vibrios could be thought of in this way: as a single cohesive population, with diversity at the level of niche-specifying variants that have failed to sweep through the entire population because of some tradeoff.)

Second, frequency-dependent selection might maintain diversity in a subset of genes involved in niche complementarity, social interactions, and predator–prey interactions (Cordero and Polz 2014). A relatively high proportion of genes in the flexible genome may be involved in such interactions. It has been argued

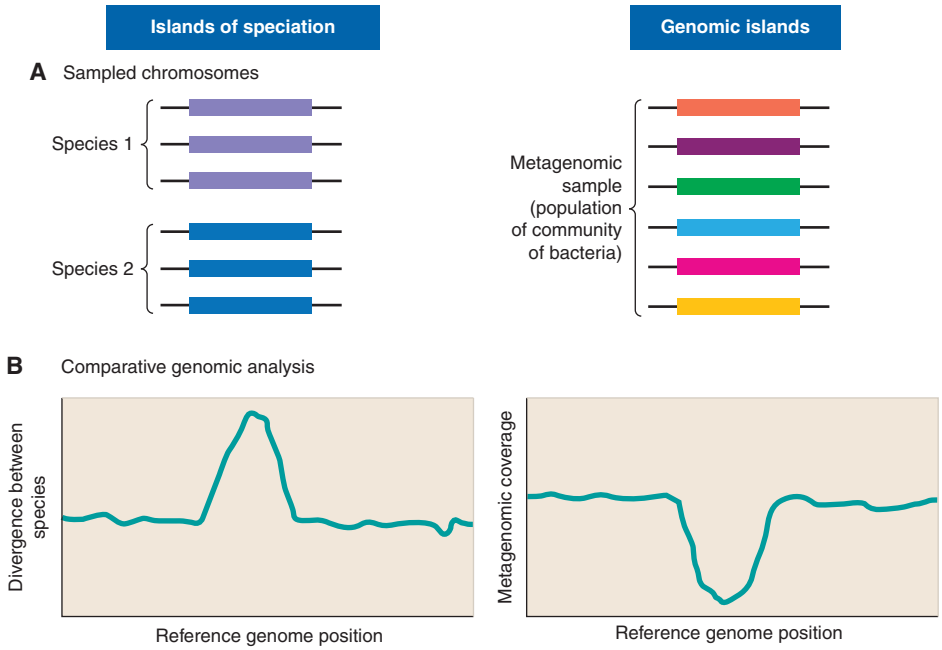

Figure 5. Islands of speciation are distinct from genomic islands. (*A*) Different colors denote different allelic variants at a chromosomal locus. (*B*) Divergence between species (*left y*-axis) is often measured as the number of fixed nucleotide substitutions between species, or a measure, such as the fixation index (F_{ST}); metagenomic coverage (*right y*-axis) is simply the number of metagenomic sequencing reads that align at a given position of the reference genome.

previously that many genes occurring at intermediate frequency within genomes are involved in predation evasion by varying surface antigenicity (Rodriguez-Valera et al. 2009; Cordero and Polz 2014). Moreover, intermediate frequency genes may be involved in frequency dependent interactions, such as public good production and cheating as well as niche-complementation (Cordero and Polz 2014). This may also explain some phenotypic variation frequently observed among closely related genotypes. For example, any secreted compound, such as enzymes, antibiotics, or signaling molecules, can become a public good that may invite cheating given sufficiently stable population structure. If, as observed recently (Cordero et al. 2012b), cheating involves loss of the public good production genes, then gene content variation among closely related isolates can arise.

Perhaps most important, frequency-dependent selection might slow the rates of selective sweeps within a population. For example, phage–bacteria interactions are often modeled with "kill-the-winner" dynamics, analogous to Lotka–Volterra ecological predator–prey models (Rodriguez-Valera et al. 2009), in which the fittest bacterial genotype (the winner) rises to high frequency, only to be targeted by a specific phage, leading to its decline and possible replacement by other genotypes. In this way, selective sweeps by the fittest genotype are prevented, or at least delayed, allowing more recombination events to occur between sweeps. However, it is unclear whether "kill-the-winner" dynamics are sufficient to push a bacterial population into the $r/s > 1$ regime (Cordero and Polz 2014), and other factors might also contribute. For example, social interactions and clonal interference could significantly reduce the rates of selective sweeps, and recombination can accelerate the rate of adaptation (Cooper 2007). Further work will be needed to fully understand the factors that maintain diversity within populations and delay selective sweeps.

Last, we should not forget that many genes, typically localized in genomic islands of high variation, appear to have such high turnover within populations that a high fraction might be (nearly) neutral to bacterial fitness (Berg and Kurland 2002; Thompson et al. 2005; Haegeman and Weitz 2012). Similarly, if genome-wide selective sweeps do not periodically reduce diversity, substantial allelic diversity will be preserved through speciation. In other words, allelic diversity will be much older than the population itself (Castillo-Ramírez et al. 2011). Importantly, interpretation of such microevolutionary changes, in the context of selection and population dynamics, requires that sympatric genomes (i.e., from the same population) are sampled.

Sampling from the same, locally coexisting population is important because another portion of genes that are generally considered as part of the flexible genome may actually be part of the core genome of local populations and hence be under purifying selection. The example of *Prochlorococcus* populations in the Atlantic and Pacific given earlier in this article falls into this category. Another recent example is *Campylobacter jejuni* strains that were isolated from both cattle and chickens, but the genome-wide phylogeny provided little evidence for host preference (Sheppard et al. 2013). In other words, host switching is relatively rapid and long-term host preferences have not been established. However, a gene cluster involved in vitamin B_5 biosynthesis is universally present in cattle isolates, but mostly absent in chicken isolates. This gene cluster appears to provide a selective advantage in B_5-depleted environments, which might include the cattle gut (Sheppard et al. 2013). An ecological trait is therefore associated with variation in a single gene cluster, but not with diversity across the entire genome. The gene cluster can be thought of as a niche-specifying variant, and the cattle- and chicken-associated isolates could be placed at stage 1 of the differentiation spectrum (Fig. 1). This by no means guarantees that stage 1 will proceed to stage 2 and onward to genome-wide divergence. Rather, phenotypic diversity in host preference might be thought of as part of the shared ecology of all *C. jejuni*. Regardless, these examples highlight the importance of considering allele frequencies (in the core and flexible genome) in the context of carefully sampled populations (Box 1).

Whether or not niche-specifying variants trigger further differentiation, they can provide insights into the mechanisms of niche adaptation, and can be identified by properly designed GWAS. An appropriate microbial GWAS should account for the degree of recombination or clonality in the population of interest (Falush and Bowden 2006; Chen and Shapiro 2015). Especially, in highly clonal populations, associations should be based on a convergence criterion (Sokurenko 2004; Chattopadhyay et al. 2013), in which phenotypes of interest are acquired independently in different lineages (Fig. 6). Mutations, alleles, or genes that are repeatedly associated with these phenotypic transitions can then be identified, and the statistical significance of their associations assessed relative to a neutral model (Farhat et al. 2013, 2014). In highly recombining populations, convergence tests are still justified (Shapiro et al. 2009; Shapiro 2014), but might lack power relative to approaches that take into account rapid recombination. For example, in *Vibrio*, the flexible genome turns over very rapidly (Thompson et al. 2005; Boucher et al. 2011; Shapiro et al. 2012), such that associations between habitat preference and flexible genes are unlikely to be maintained by vertical descent, but rather by habitat-specific selective pressures. Similar reasoning was used to identify *E. coli* flexible genes associated with environmental or gut-associated lifestyles (Luo et al. 2011).

CONCLUDING REMARKS AND PROSPECTS FOR REVERSE ECOLOGY

As we have outlined in this review, when genotypic clusters can be detected, they are predicted to be ecologically differentiated from other such clusters. Although this does not preclude some level of ecological diversity within these clusters caused by the acquisition of novel, niche-specifying genes, such diversity should be relatively minor because selection can only maintain a

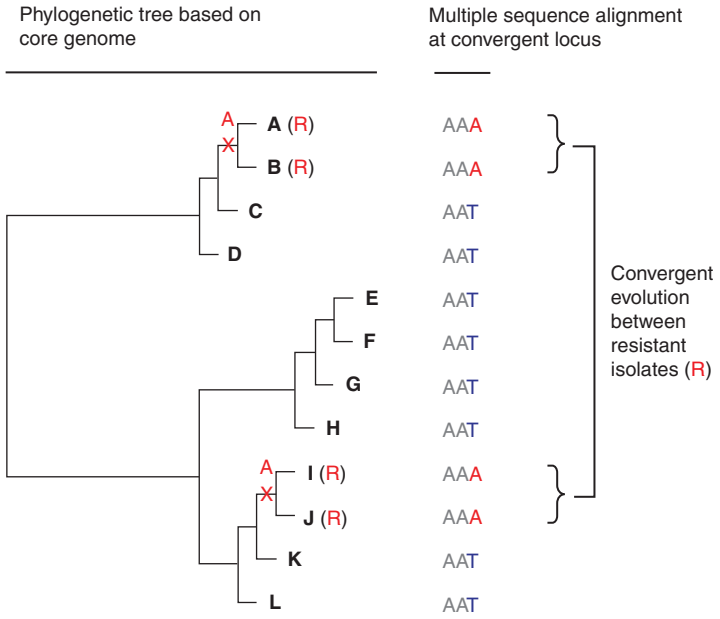

Figure 6. Evolutionary convergence as the basis for genome-wide association studies (GWAS). In this simplified example, the genome-wide core phylogeny has been inferred for a sample of mostly clonal bacterial isolates (A–L). A convergent (homoplasic) single nucleotide polymorphism (SNP) is identified in four genomes. Importantly, this corresponds to only two independent mutation events (T → A, indicated in the phylogenetic tree by the red "X" with an "A" above it), which associate perfectly with two independent transitions from the antibiotic sensitive to resistant (R) state. The significance of the association can be assessed by calculating a *P*-value by resampling from the genome-wide distribution of mutations and phenotypic states (resistant or sensitive) on the phylogeny. By failing to account for population structure (e.g., the phylogenetic information), four events would be counted, thereby overestimating the significance of the association. GWAS can also be performed considering entire genes, instead of individual nucleotide sites, as targets of convergent mutations (for examples, see Sheppard et al. 2013 and Farhat et al. 2013, 2014).

limited number of ecologically divergent loci within the same, genetically mixed population (Friedman et al. 2013). Hence, a reverse ecology strategy, in which genotypic clusters among coexisting microbes are identified as a first step toward identifying ecologically cohesive populations, is potentially easier than the forward approach, which is to map marker genes onto many environmental samples in the hopes of finding significant ecological associations.

As genome sequencing becomes more and more broadly accessible because of decreased cost and increased throughput, it will become feasible to sequence sufficient numbers of closely related genomes from the same environmental samples, either in the form of isolates or single-cell genomes. Moreover, improved coverage and assembly techniques will also allow increased identification of genotypic clusters from metagenomic samples. Once these genomes are available, they can serve two purposes. First, they can be used to delineate clusters, and second, they can help build hypotheses of environmental differentiation by searching for genes of potential ecological relevance. In that way, some guess as to the population's niche can be made before engaging in the exercise of mapping the cluster onto environmental samples and identifying correlations with biotic and abiotic environmental metadata. We stress that this exercise must consider the fine structure of the environment because microbial habitats and interactions often occur at small spatial (micro- to millimeters) and temporal scales (minutes to days) (Polz et al. 2006). Second, given sufficient environmental and genomic sampling, GWAS

can provide valuable further insights as to the causes of allele and gene diversity within and between populations. Even the very early events of speciation, including the acquisition of niche-specifying genes or mutations that will rarely lead to new species, are of interest both for their impacts on health and for exploring how and why subsequent steps toward speciation take place. For example, HGT is rapid and rampant within the human microbiome, allowing bacteria to evolve in response to natural selection imposed by the host immune system, viral predation, antibiotics, and other factors. How much of this HGT triggers speciation and how much remains within-species diversity? The answer will help us identify relevant biomarkers of health and disease, be they mutations, genes, operons, or species, while gaining a deeper understanding of the rates, limitations, and nature of speciation.

ACKNOWLEDGMENTS

We thank Libusha Kelly, Rex Malmstrom, Gabriel Perron, and Rachel Whitaker for their insightful comments. Funding for M.F.P. was provided by National Science Foundation Grant DEB 0821391, National Institute of Environmental Health Sciences Grant P30-ES002109, the Moore Foundation and the Broad Institute's SPARC program. Funding for B.J.S. was provided by the Natural Sciences and Engineering Research Council of Canada and the Canada Research Chairs program.

REFERENCES

Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF. 2009. Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* **461:** 1243–1247.

Berg OG, Kurland CG. 2002. Evolution of microbial genomes: Sequence acquisition and loss. *Mol Biol Evol* **19:** 2265–2276.

Blount ZD, Barrick JE, Davidson CJ, Lenski RE. 2012. Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature* **488:** 513–518.

Boucher Y, Cordero OX, Takemura A, Hunt DE, Schliep K, Bapteste E, Lopez P, Tarr CL, Polz MF. 2011. Local mobile gene pools rapidly cross species boundaries to create endemicity within global *Vibrio cholerae* populations. *mBio* **2:** e00335–10.

Burke C, Steinberg P, Rusch D, Kjelleberg S, Thomas T. 2011. Bacterial community assembly based on functional genes rather than species. *Proc Natl Acad Sci* **108:** 14288–14293.

Cadillo-Quiroz H, Didelot X, Held NL, Herrera A, Darling A, Reno ML, Krause DJ, Whitaker RJ. 2012. Patterns of gene flow define species of thermophilic archaea. *PLoS Biol* **10:** e1001265.

Caro-Quintero A, Konstantinidis KT. 2011. Bacterial species may exist, metagenomics reveal. *Environ Microbiol* **14:** 347–355.

Caro-Quintero A, Rodriguez-Castaño GP, Konstantinidis KT. 2009. Genomic insights into the convergence and pathogenicity factors of *Campylobacter jejuni* and *Campylobacter coli* species. *J Bacteriol* **191:** 5824–5831.

Castillo-Ramírez S, Harris SR, Holden MT, He M, Parkhill J, Bentley SD, Feil EJ. 2011. The impact of recombination on dN/dS within recently emerged bacterial clones. *PLoS Pathog* **7:** e1002129.

Chattopadhyay S, Paul S, Dykhuizen DE, Sokurenko EV. 2013. Tracking recent adaptive evolution in microbial species using TimeZone. *Nat Protoc* **8:** 652–665.

Chen PE, Shapiro BJ. 2015. The advent of genome-wide association studies for bacteria. *Curr Opin Microbiol* **25:** 17–24.

Cohan FM. 2001. Bacterial species and speciation. *Syst Biol* **50:** 513–524.

Cohan FM, Koeppel AF. 2008. The origins of ecological diversity in prokaryotes. *Curr Biol* **18:** 1024–1034.

Coleman ML, Chisholm SW. 2010. Ecosystem-specific selection pressures revealed through comparative population genomics. *Proc Natl Acad Sci* **107:** 18634–18639.

Cooper TF. 2007. Recombination speeds adaptation by reducing competition between beneficial mutations in populations of *Escherichia coli*. *PLoS Biol* **5:** e225.

Cordero OX, Polz MF. 2014. Explaining microbial genomic diversity in light of evolutionary ecology. *Nat Rev Microbiol* **12:** 263–273.

Cordero OX, Ventouras LA, DeLong EF, Polz MF. 2012a. Public good dynamics drive evolution of iron acquisition strategies in natural bacterioplankton populations. *Proc Natl Acad Sci* **109:** 20059–20064.

Cordero OX, Wildschutte H, Kirkup B, Proehl S, Ngo L, Hussain F, Le Roux F, Mincer T, Polz MF. 2012b. Ecological populations of bacteria act as socially cohesive units of antibiotic production and resistance. *Science* **337:** 1228–1231.

Cornejo OE, McGee L, Rozen DE. 2010. Polymorphic competence peptides do not restrict recombination in *Streptococcus pneumoniae*. *Mol Biol Evol* **27:** 694–702.

Croucher NJ, Harris SR, Barquist L, Parkhill J, Bentley SD. 2012. A high-resolution view of genome-wide pneumococcal transformation. *PLoS Pathog* **8:** e1002745.

Danchin EGJ, Rosso MN. 2012. Lateral gene transfers have polished animal genomes: Lessons from nematodes. *Front Cell Infect Microbiol* **2:** 27–27.

Denef VJ, Mueller RS, Banfield JF. 2010a. AMD biofilms: Using model communities to study microbial evolution and ecological complexity in nature. *ISME J* **4:** 599–610.

Denef VJ, Kalnejais LH, Mueller RS, Wilmes P, Baker BJ, Thomas BC, VerBerkmoes NC, Hettich RL, Banfield JF.

2010b. Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. *Proc Natl Acad Sci* **107:** 2383–2390.

de Vries J, Wackernagel W. 2002. Integration of foreign DNA during natural transformation of *Acinetobacter* sp. by homology-facilitated illegitimate recombination. *Proc Natl Acad Sci* **99:** 2094–2099.

Didelot X, Lawson D, Darling A, Falush D. 2010. Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics* **186:** 1435–1449.

Dobzhansky T. 1935. A critique of the species concept in biology. *Philos Sci* **2:** 344–355.

Doolittle WF, Papke RT. 2006. Genomics and the bacterial species problem. *Genome Biol* **7:** 116.

Doolittle WF, Zhaxybayeva O. 2009. On the origin of prokaryotic species. *Genome Res* **19:** 744–756.

Ellison CE, Hall C, Kowbel D, Welch J, Brem RB, Glass NL, Taylor JW. 2011. Population genomics and local adaptation in wild isolates of a model microbial eukaryote. *Proc Natl Acad Sci* **108:** 2831–2836.

Falush D, Bowden R. 2006. Genome-wide association mapping in bacteria? *Trends Microbiol* **14:** 353–355.

Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, Warren RM, Streicher EM, Calver A, Sloutsky A, et al. 2013. Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nat Genet* **45:** 1183–1189.

Farhat MR, Shapiro B, Sheppard SK, Colijn C, Murray M. 2014. A phylogeny-based sampling strategy and power calculator informs genome-wide associations study design for microbial pathogens. *Genome Med* **6:** 101.

Fraser C, Hanage WP, Spratt BG. 2007. Recombination and the nature of bacterial speciation. *Science* **315:** 476–480.

Friedman J, Alm EJ, Shapiro BJ. 2013. Sympatric speciation: When is it possible in bacteria? *PLoS ONE* **8:** e53539.

Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, Van de Peer Y, Vandamme P, Thompson FL, et al. 2005. Opinion: Re-evaluating prokaryotic species. *Nat Rev Microbiol* **3:** 733–739.

Grogan DW, Stengel KR. 2008. Recombination of synthetic oligonucleotides with prokaryotic chromosomes: Substrate requirements of the *Escherichia coli*/λRed and *Sulfolobus acidocaldarius* recombination systems. *Mol Microbiol* **69:** 1255–1265.

Haegeman B, Weitz JS. 2012. A neutral theory of genome evolution and the frequency distribution of genes. *BMC Genomics* **13:** 196–196.

Hahn MW, Pockl M. 2005. Ecotypes of planktonic actinobacteria with identical 16S rRNA genes adapted to thermal niches in temperate, subtropical, and tropical freshwater habitats. *Appl Environ Microbiol* **71:** 766–773.

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

Hanson CA, Fuhrman JA, Horner-Devine MC, Martiny JB. 2012. Beyond biogeographic patterns: Processes shaping the microbial landscape. *Nat Rev Microbiol* **10:** 497–506.

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

Jax K. 2006. Ecological units: Definitions and application. *Q Rev Biol* **81:** 237–258.

Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EM, Chisholm SW. 2006. Niche partitioning among *Prochlorococcus ecotypes* along ocean-scale environmental gradients. *Science* **311:** 1737–1740.

Katz LS, Petkau A, Beaulaurier J, Tyler S, Antonova ES, Turnsek MA, Guo Y, Wang S, Paxinos EE, Orata F, et al. 2013. Evolutionary dynamics of *Vibrio cholerae* O1 following a single-source introduction to Haiti. *mBio* **4:** e00398–13.

Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferriera S, Johnson J, et al. 2007. Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* **3:** 2515–2528.

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

Koeppel AF, Wertheim JO, Barone L, Gentile N, Krizanc D, Cohan FM. 2013. Speedy speciation in a bacterial microcosm: New species can arise as frequently as adaptations within a species. *ISME J* **7:** 1080–1091.

Konstantinidis KT, Delong EF. 2008. Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. *ISME J* **2:** 1052–1065.

Konstantinidis KT, Tiedje JM. 2005. Towards a genome-based taxonomy for prokaryotes. *J Bacteriol* **187:** 6258–6264.

Levy R, Borenstein E. 2012. Reverse ecology: From systems to environments and back. *Adv Exp Med Biol* **751:** 329–345.

Li YF, Costello JC, Holloway AK, Hahn MW. 2008. "Reverse ecology" and the power of population genomics. *Evolution* **62:** 2984–2994.

Lieberman TD, Flett KB, Yelin I, Martin TR, McAdam AJ, Priebe GP, Kishony R. 2013. Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. *Nat Genet* **46:** 82–87.

Luo C, Walk ST, Gordon DM, Feldgarden M, Tiedje JM, Konstantinidis KT. 2011. Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *Proc Natl Acad Sci* **108:** 7200–7205.

Majewski JJ, Cohan FMF. 1999a. DNA sequence similarity requirements for interspecific recombination in *Bacillus*. *Genetics* **153:** 1525–1533.

Majewski J, Cohan FM. 1999b. Adapt globally, act locally: The effect of selective sweeps on bacterial sequence diversity. *Genetics* **152:** 1459–1474.

Mallet J. 2008. Hybridization, ecological races and the nature of species: Empirical evidence for the ease of speciation. *Philos Trans R Soc Lond B Biol Sci* **363:** 2971–2986.

Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, Corander J. 2012. Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res* **40:** e6.

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

Mell JC, Shumilina S, Hall IM, Redfield RJ. 2011. Transformation of natural genetic variation into *Haemophilus influenzae* genomes. *PLoS Pathog* **7:** e1002151.

Naor A, Lapierre P, Mevarech M, Papke RT, Gophna U. 2012. Low species barriers in halophilic archaea and the formation of recombinant hybrids. *Curr Biol* **22:** 1444–1448.

Nemergut DR, Costello EK, Hamady M, Lozupone C, Jiang L, Schmidt SK, Fierer N, Townsend AR, Cleveland CC, Stanish L, et al. 2010. Global patterns in the biogeography of bacterial taxa. *Environ Microbiol* **13:** 135–144.

Oh S, Caro-Quintero A, Tsementzi D, DeLeon-Rodriguez N, Luo C, Poretsky R, Konstantinidis KT. 2011. Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. *Appl Environ Microbiol* **77:** 6000–6011.

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

Pennisi BE. 2014. Disputed islands. *Science* **345:** 611–613.

Polz MF, Hunt DE, Preheim SP, Weinreich DM. 2006. Patterns and mechanisms of genetic and phenotypic differentiation in marine microbes. *Philos Trans R Soc Lond B Biol Sci* **361:** 2009–2021.

Polz MF, Alm EJ, Hanage WP. 2013. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends Genet* **29:** 170–175.

Reno ML, Held NL, Fields CJ, Burke PV, Whitaker RJ. 2009. Biogeography of the *Sulfolobus islandicus* pan-genome. *Proc Natl Acad Sci* **106:** 8605–8610.

Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR, et al. 2003. Genome divergence in two *Prochlorococcus ecotypes* reflects oceanic niche differentiation. *Nature* **424:** 1042–1047.

Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, Pasić L, Thingstad TF, Rohwer F, Mira A. 2009. Explaining microbial population genomics through phage predation. *Nat Rev Microbiol* **7:** 828–836.

Schönknecht G, Weber AP, Lercher MJ. 2013. Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution. *Bioessays* **36:** 9–20.

Shapiro BJ. 2014. Signatures of natural selection and ecological differentiation in microbial genomes. *Adv Exp Med Biol* **781:** 339–359.

Shapiro BJ, David LA, Friedman J, Alm EJ. 2009. Looking for Darwin's footprints in the microbial world. *Trends Microbiol* **17:** 196–204.

Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabó G, Polz MF, Alm EJ. 2012. Population genomics of early events in the ecological differentiation of bacteria. *Science* **336:** 48–51.

Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, Bentley SD, Maiden MC, Parkhill J, Falush D. 2013. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*. *Proc Natl Acad Sci* **110:** 11923–11927.

Simmons SL, Dibartolo G, Denef VJ, Goltsman DS, Thelen MP, Banfield JF. 2008. Population genomic analysis of strain variation in *Leptospirillum* group II bacteria involved in acid mine drainage formation. *PLoS Biol* **6:** 1427–1442.

Smith JM, Smith NH, O'Rourke M, Spratt BG. 1993. How clonal are bacteria? *Proc Natl Acad Sci* **90:** 4384–4388.

Sokurenko EV. 2004. Selection footprint in the FimH adhesin shows pathoadaptive niche differentiation in *Escherichia coli*. *Mol Biol Evol* **21:** 1373–1383.

Syvanen M. 2012. Evolutionary implications of horizontal gene transfer. *Annu Rev Genet* **46:** 341–358.

Szabó G, Preheim SP, Kauffman KM, David LA, Shapiro J, Alm EJ, Polz MF. 2013. Reproducibility of *Vibrionaceae* population structure in coastal bacterioplankton. *ISME J* **7:** 509–519.

Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, Benoit J, Sarma-Rupavtarm R, Distel DL, Polz MF. 2005. Genotypic diversity within a natural coastal bacterioplankton population. *Science* **307:** 1311–1313.

Turner TL, Hahn MW. 2010. Genomic islands of speciation or genomic islands and speciation? *Mol Ecol* **19:** 848–850.

Turner TL, Hahn MW, Nuzhdin SV. 2005. Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biol* **3:** 1572–1578.

Via S. 2012. Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philos Trans R Soc Lond B Biol Sci* **367:** 451–460.

Vos M. 2011. A species concept for bacteria based on adaptive divergence. *Trends Microbiol* **19:** 1–7.

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

Vulić M, Dionisio F, Taddei F, Radman M. 1997. Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proc Natl Acad Sci* **94:** 9763–9767.

Welch RA, Burland V, Plunkett G III, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, et al. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc Natl Acad Sci* **99:** 17020–17024.

Whitaker RJ, Grogan DW, Taylor JW. 2003. Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science* **301:** 976–978.

Whitaker RJ, Grogan DW, Taylor JW. 2005. Recombination shapes the natural population structure of the hyperthermophilic archaeon *Sulfolobus islandicus*. *Mol Biol Evol* **22:** 2354–2361.

Wiedenbeck J, Cohan FM. 2011. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiol Rev* **35:** 957–976.

Williams D, Gogarten JP, Papke RT. 2012. Quantifying homologous replacement of loci between haloarchaeal species. *Genome Biol Evol* **4:** 1223–1244.

Yawata Y, Cordero OX, Menolascina F, Hehemann JH, Polz MF, Stocker R. 2014. Competition-dispersal trade-off ecologically differentiates recently speciated marine bacterioplankton populations. *Proc Natl Acad Sci* **111:** 5622–5627.