What Microbial Population Genomics Has Taught Us About Speciation

B. Jesse Shapiro

Abstract Population genomics has emerged as a valuable tool to define and delimit species and to understand the mechanisms that drive and maintain speciation. Species and speciation have been notoriously difficult to study in microbes owing to their asexual reproduction, promiscuous horizontal gene transfer, and obscure microscopic niches. Over the past few years, whole-genome sequencing of closely related, locally co-occurring populations of microbes, combined with simulations and modelling, has revealed certain general features of microbial speciation: it is usually driven by divergent natural selection between distinct ecological niches (a form of the ecological species concept), and species distinctness is maintained by barriers to gene flow (a form of the biological species concept). In some cases, gene-flow barriers may come about as a natural consequence of ecological specialization. Although these features appear to be quite general, there are exceptions. Trivially, barriers to gene flow cannot be used to delimit clonal populations where there is negligible gene flow. More interestingly, it is unclear whether other barriers to gene flow, such as genetic incompatibilities or differences in phage-host range, are able to drive speciation in the absence of other selective pressures. Here, I discuss the extent to which speciation is driven by natural selection, gene-flow barriers, or a combination of the two, drawing on recent examples from bacterial and archaeal population genomics, experimental evolution, and modelling. I then describe how population genomic data can be used to define and delimit species boundaries, based upon nucleotide identity cutoffs or upon discontinuities in gene flow. Despite important limitations and caveats, delimitation methods provide a useful starting point for more detailed investigation into the genetic and ecological basis of speciation.

Keywords Archaea · Bacteria · Biological species concept · Ecological species concept · ecoSNP · Gene flow · Mosaic sympatry · Niche · Overlapping Habitat Model · Speciation

B. Jesse Shapiro (✉)
Department of Biological Sciences, University of Montreal, Montreal, QC, Canada
e-mail: jesse.shapiro@umontreal.ca

Martin F. Polz and Om P. Rajora (eds.), *Population Genomics: Microorganisms*,
Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_10, © Springer International Publishing AG 2018

1 Introduction

Over 150 years since Darwin published *On the origin of species*, biologists and philosophers are still debating what species are, how they form, and if they really exist (Doolittle and Zhaxybayeva 2009; Doolittle 2012). I have previously argued that species do exist and their origin (the process of speciation) is generally, if not always, driven by natural selection for adaptation to distinct ecological niches (Shapiro et al. 2016). Here, I will critically re-evaluate this argument and discuss alternatives, drawing on the most recent advances from population genomics. Most of the examples will be from bacteria, with some comparisons across other domains of life. Building on the observation that genetically and ecologically coherent units do exist (Caro-Quintero and Konstantinidis 2011; Shapiro and Polz 2014) even if their boundaries may be "fuzzy" (Hanage et al. 2005; Hanage 2013), I will focus on the mechanisms that give rise to these units and keep them distinct. In other words, this chapter is mainly about speciation, not species. However, I will also discuss methods to define and delimit species, which can provide a practical first step toward better understanding of the mechanisms driving speciation.

2 Species Concepts and Definitions

To begin, let us briefly define population genomics and make the distinction between species concepts and definitions. Species *concepts* require at least some notion of mechanism, whereas species *definitions* can be completely operational and agnostic to mechanism but can also be based on a particular species concept (Gevers et al. 2005). I will focus on two popular types of species concepts. The ecological species concept (ESC), favoured by Darwin, posits that speciation is driven by natural selection, with each species adapted to a unique ecological niche (Schluter 2009). The biological species concept (BSC) posits that speciation is driven by barriers to genetic exchange, which is equivalent to reproductive isolation in sexual species (Dobzhansky 1935; Mayr 1942). Strictly speaking, the BSC will never apply to asexual organisms like bacteria. Moreover, bacteria (and other domains of life, including plants and animals) can exchange genes across species boundaries, so barriers to gene flow will always remain somewhat permeable (Shapiro et al. 2016). Therefore, rather than the strict BSC, I will refer mainly to a *BSC-like* concept in which rates of gene flow are higher within than between species, but cross-species gene transfer can still occur. Other species concepts exist, but most are effectively combinations of the ESC and the BSC. For example, the stable ecotype model is essentially the ESC with relatively low rates of genetic exchange (Wiedenbeck and Cohan 2011). Allopatric speciation is a special case of the BSC in which barriers to genetic exchange are initially due to physical isolation, although they can later be reinforced by genetic incompatibilities. Different species concepts predict different and distinctive patterns of genetic variation within and between species (Krause and Whitaker 2015), which can in principle be harnessed to define species.

Population genomics is a valuable tool—perhaps the most valuable tool available—to both inform our concept of species and to precisely define species. The (relatively new) field of population genomics (see the Chapters in this volume on this topic) uses whole-genome information to answer questions posed by the (more mature) field of population genetics—the study of how mutation, selection, and drift change allele frequencies within a population. Populations are generally defined as sets of locally coexisting members of species. If we do not know what species are in the first place, or how to define them, the task of defining species and populations can become circular. Therefore, the application of population genomics to the study of species and speciation usually requires some a priori notion of species or population boundaries, which can then be critically evaluated based on the fit of observed patterns of genomic variation with the predictions of competing species concepts. In some cases, the prior information can include ecological hypotheses, for example, that speciation in marine vibrios is driven by adaptation to either free-living or particle-associated lifestyles (Shapiro et al. 2012). In other cases, a previously named species or genus might be sampled to test whether genome sequence data fits a particular species concept and whether the sampled genomes constitute one or many species (Cadillo-Quiroz et al. 2012; Bobay and Ochman 2017). In general, population genomics requires complete or near-complete genome sequences from several individuals, be they cultured isolated or single-cell genomes. Metagenomic sequencing of bulk DNA from an environment is usually incapable of linking particular genes or mutations back to a specific individual, making it more difficult to test certain species concepts, particularly versions of the BSC that require testing for differences in recombination rates within and between populations. These shortcomings have not prevented researchers from defining “metagenomic species”, although such definitions are purely operational and not clearly grounded in any particular concept of species other than the prediction that members of the same species should have correlated abundances over time or across samples (Caro-Quintero and Konstantinidis 2011; Alneberg et al. 2014; Nielsen et al. 2014). Nevertheless, metagenomics can help estimate valuable population genetic parameters such as the nucleotide diversity within a species.

3 Selection, Gene-Flow Barriers, or Both?

Both natural selection and barriers to gene flow can be important in the speciation process, but which is usually the driver that initiates speciation? Certain forms of gene flow, namely, homologous recombination, require a certain degree of sequence identity between donor DNA and the recipient genome (although a few dozen base pairs of identity can be sufficient to initiate the transfer of several kilobases of completely nonhomologous DNA; Mell et al. 2011; Croucher et al. 2012). In principle, the accumulation of mutations could gradually create barriers to homologous recombination, driving speciation in the absence of selection and yielding genetically distinct species fitting the BSC. According to computational modelling, this is unlikely to occur, unless recombination rates decline unrealistically rapidly with sequence divergence (Fraser et al. 2007). The model suggests that another force—such as divergent natural selection between two niches—is required to drive speciation. A further theoretical argument why selection is required to initiate speciation is based on the competitive exclusion principle (Gause 1934; Tilman 1982). If two species are ecologically equivalent (meaning they are under identical or near-identical regimes of selection), one will inevitably (after some period of time) drive the other to extinction. Only if species are under divergent selection for adaptation to distinct niches will speciation occur.

Beyond these theoretical considerations, what is the population genomic evidence for selection driving speciation? Perhaps the most direct evidence comes from laboratory evolution experiments, combined with whole-genome sequencing. In a long-term evolution experiment starting with a single clone of *E. coli*, a lineage evolved after ~31,000 generations with the ability to metabolize citrate, a previously unused carbon source present in the growth medium (Blount et al. 2008). Sequencing of ancestral CIT- and derived CIT+ genomes revealed the genetic changes required for citrate utilization (Blount et al. 2012). The two ecologically distinct lineages continue to coexist in the experiment, consistent with the ESC. Despite being a clear example of how ecological selection can drive speciation, it is not really a fair test of whether gene-flow barriers can drive speciation because the *E. coli* in the experiment are not competent and do not recombine DNA.

In another evolution experiment using bacteriophage capable of recombination within host cells, Meyer et al. (2016) showed that speciation readily occurred under both allopatric and sympatric conditions, driven by divergent selection for phage to specialize on one of two available bacterial hosts that differed only in their surface phage receptor. In the allopatric experiment, phage were cultured in media containing only one host, and specialization occurred rapidly. In the sympatric experiment, both bacterial hosts were present in the culture media, but specialization still occurred because of the link between ecological preference (one host or the other) and recombination, which only occurs within a host cell. These barriers to gene flow imposed by host preference are analogous to the barriers imposed by particle preference within the marine water column, which appears to be driving sympatric speciation in natural vibrio populations (discussed below). This subtle spatial structure within seemingly homogeneous sympatric environments has been referred to as “mosaic sympatry” (Mallet 2008; Shapiro and Polz 2014) and explains how ecological selection can initiate speciation, which is later reinforced by gene-flow barriers. By sequencing evolved and ancestral phage genomes, Meyer et al. (2016) further showed that several mutations in a single host-recognition gene in the phage genome (*J*) explained host specialization, with different mutations associated with different hosts. The observation of a single gene apparently responsible for speciation is consistent with theoretical predictions that sympatric speciation proceeds more readily when fewer loci are involved in ecological differentiation or reproductive isolations (Kondrashov and Mina 1986; Friedman et al. 2013). Further reducing gene flow between incipient phage species, recombinant *J* alleles encoding combinations of mutations adapted to different hosts were not viable. Therefore, Meyer et al. (2016) appear to have captured a very early stage of sympatric speciation, driven by ecological differentiation and maintained by gene-flow barriers. A population genomic study of sympatric marine cyanophages suggests the same mechanisms may be at play in natural phage populations, although speciation may be driven by ecological factors other than host identity (Gregory et al. 2016).

Similar patterns have also been observed in recombining natural bacterial populations. For example, we compared the genomes of very closely related *Vibrio cyclitrophicus* isolates (identical 16S and >99% amino acid identity) and concluded that speciation was driven by differential selection for either free-living or particle-associated niches and maintained by the emergence of barriers to gene flow (Shapiro et al. 2012). In other words, the speciation process began with an ESC-like mechanism and was reinforced by a BSC-like mechanism. However, it is difficult to be certain that ecological selection *preceded* the establishment of gene-flow barriers. We found that gene-flow barriers between incipient species are only evident among the most recent detectable recombination events, while older recombination events do not respect species boundaries (Shapiro et al. 2012). I later used an adaptation of the McDonald-Kreitman (MK) test (Vos 2011) to show that the divergence between incipient species involved an unexpected excess of nonsynonymous substitutions, suggesting positive selection driving their divergence (Shapiro 2014). Still, although it is certainly consistent with the "selection first" hypothesis, this does not conclusively prove that ecological selection occurred before the establishment of gene-flow boundaries. Further complicating things, the likely targets of differential selection between free-living and particle-associated habitats—three loci containing >80% of ecoSNPs (the single nucleotide polymorphisms fixed between habitats) and several other genes present in one habitat but not the other—are subject to frequent recombination and were likely acquired from distantly related lineages of *Vibrio*, making it difficult to date their acquisition with certainty. Nevertheless, it is abundantly clear that the two incipient species are ecologically distinct (Yawata et al. 2014) and there is currently no evidence suggesting that gene-flow boundaries emerged before differential selection.

Evidence from several other natural bacterial populations supports the idea that ecological differentiation, due to selection on one or a few "niche-specifying" genes, can occur before any apparent boundaries to gene flow. For example, a population genomic study of *Rhizobium leguminosarum* found that they "form dynamic, diverse populations that are unified by gene flow despite selection acting at one or more loci" (Klinger et al. 2016). Specifically, they found that selection (artificially applied in a 22-year nitrogen fertilization experiment) favoured certain alleles of nitrogen fixation genes, which rose to high frequency in the *R. leguminosarum* population without affecting diversity elsewhere in the genome (Klinger et al. 2016). Such "gene-specific" selective sweeps (Shapiro and Polz 2014) have also been documented in population genomic studies of other bacteria, including *Mesorhizobium* (Porter et al. 2016) and *Streptococcus* (Croucher et al. 2011; Bao et al. 2016). The apparent ease with which natural selection can favour the increase of adaptive genes or alleles in recombining microbial populations suggests that selection could at least plausibly drive speciation, before the establishment of gene-flow boundaries.

Let us now consider the alternative hypothesis that gene-flow barriers directly drive speciation without the need for ecological selection—a version of the BSC without any trace of the ESC. As described above, it is unlikely that gradual mutation accumulation could cause barriers to homologous recombination. But what about other mechanisms of recombination? Phage-mediated transduction requires the donor and recipient cells of a recombination event to be infected by the same phage. Therefore barriers to phage infection could limit gene flow. Consistent with this idea, a comparative analysis of phage and bacterial genome sequences showed that phage-mediated recombination events are mostly limited to closely related bacterial donors and recipients (Popa et al. 2016). This phage-host specificity could limit genetic exchange to close relatives, providing a natural mechanism for the BSC and leading to more genetic exchange within than between species. In principle, a mutation or recombination event changing a phage receptor could instantaneously create a barrier to gene flow (Rodriguez-Valera et al. 2009; López-Pérez and Rodriguez-Valera 2016), but population genomic evidence of such a BSC-like mechanism driving speciation is still lacking. Large-scale chromosomal rearrangements can play an important role in creating reproductive isolation in yeast (Charron et al. 2014; Leducq et al. 2016), but it is unclear which came first—barriers to gene flow or ecological specialization—or whether both occurred more or less simultaneously to initiate speciation.

4 Models to Interpret Population Genomic Data

Population genomic data can be used to operationally define species and, more importantly, to test competing species concepts. An example of an operational species definition based on genome sequence data is the proposed 95% average nucleotide identity (ANI) threshold (Konstantinidis and Tiedje 2005; Konstantinidis et al. 2006). Pairs of genomes that have below 95% ANI always come from distinct species, according to most species concepts or definitions. However, although a 95% threshold may work well for most species, some recently diverged species might still share 97, 98, or 99% ANI (Doolittle and Zhaxybayeva 2009). For example, the nascent phage (Meyer et al. 2016) and *Vibrio* (Shapiro et al. 2012) species described above would be lumped into a single species using a 95% cutoff. ANI may also vary widely across the genome, leading to "fragmented speciation" in which different parts of the genome effectively speciate at different rates (Retchless and Lawrence 2010). Therefore, a universal ANI-based species definition, while appealing in its simplicity, will likely fail to distinguish "good" species, especially at early stages of speciation. ANI, like other sequence-based thresholds (such as 97% identity in the 16S rRNA gene), is still a useful starting point for a more in-depth testing of species concepts. It has been argued that 97% is a much too inclusive cutoff and that 99% 16S identity or unique sequence types better capture ecologically coherent bacterial species (Acinas et al. 2004; Eren et al. 2013; Koeppel and Wu 2014). No one would argue that genomes sharing less than 95% ANI are part of the same species. However, genomes sharing more than 95% ANI might be divided into two, three, or several species, depending on the choice of species concept.

Testing species concepts requires more than population genomic data. It also requires a model describing the mechanism of speciation, which can then be fit to population genomic data. One of the first and most influential such models is the stable ecotype model (SEM), which defines species as ecotypes, each inhabiting a distinct ecological niche, such that selective sweeps and neutral drift affect diversity within but not between species (Wiedenbeck and Cohan 2011). In other words, selective sweeps or population bottlenecks that occur within one species (ecotype) do not affect the genetic diversity of other species. Phylogenies based on marker genes often fit well with the predictions of the SEM, namely, that monophyletic groups of closely related bacteria tend to share the same ecological associations (Hunt et al. 2008; Koeppel et al. 2008). However, applied to marker gene sequences from natural populations of *Bacillus*, the SEM fits slightly worse than a neutral model without ecological niches (Fraser et al. 2009), and patterns that appear consistent with the SEM based on marker genes may be inconsistent when genome-wide information is considered (Shapiro et al. 2012).

In the SEM, sweeps or bottlenecks purge genetic diversity genome-wide, because recombination is not strong enough to decouple the evolutionary fates of loci across the genome. Different versions of the SEM can accept increasing levels of recombination (Majewski and Cohan 1999; Wiedenbeck and Cohan 2011), but the SEM always emphasizes strong selection between ecological niches and relatively low rates of gene flow, such that an adaptive allele will always spread by clonal expansion rather than recombination. Such clonal expansions are expected to result in genome-wide selective sweeps, purging genetic diversity across the genome. Although such clonal expansions and genome-wide sweeps likely occur over relatively short time scales [e.g. pathogen outbreaks; (Shapiro 2016)], they appear to be rare in nature, at least among recombining aquatic and soil bacteria studied with genome-wide surveys (Shapiro et al. 2012; Cui et al. 2015; Rosen et al. 2015; Klinger et al. 2016; Porter et al. 2016). For example, of 30 bacterial populations tracked using metagenomics in a lake over a 9-year period, only one appeared to experience a genome-wide purge of diversity (Bendall et al. 2016), although it remains unclear whether the purge was driven by selection or drift (Shapiro 2016). To explain the apparent rarity of genome-wide sweeps in nature, recent models have shown how combinations of negative frequency-dependent selection [e.g. to avoid phage predation; (Takeuchi et al. 2015)] and migration between habitat patches (Niehus et al. 2015) can allow recombination to outpace natural selection, resulting in gene-specific rather than genome-wide selective sweeps. These models help explain population genomic and metagenomic observations consistent with gene-specific sweeps in nature (Coleman and Chisholm 2010; Shapiro et al. 2012; Shapiro and Polz 2014; Klinger et al. 2016; Porter et al. 2016), but did not specifically investigate the process of speciation.

Fraser et al. (2007) used a computational model to investigate the role of homologous recombination in speciation. They confirmed the prediction of the SEM that, in the absence of distinct ecological niches and in the absence of recombination, genetically distinct clusters of bacteria continuously formed and went extinct. Thus, stable species cannot be maintained in a neutral model with only one niche. They went on to show that recombination homogenized the clusters, resulting in a single, stable "cloud" of genetic diversity. When recombination rates declined with genetic divergence, distinct and stable clusters (reminiscent of species biological species) were maintained—but only using an unrealistically steep rate of decline. In contrast to any parameterization of the Fraser et al. model, real sequence data from the genus *Streptococcus* fall into distinct clusters, despite high rates of recombination. This suggests that a neutral model with or without recombination is not sufficient to explain the formation of stable genetic clusters. For speciation to occur, another ingredient is missing. The missing ingredient could be divergent natural selection between ecological niches or, in special cases of geographic isolation, physical barriers to recombination (Krause and Whitaker 2015).

In the *sym*patric *sim*ulation (*symsim*) model of divergent selection between ecological niches, we found that recombination accelerated the initial rate of niche adaptation but later eroded the distinctness of incipient species, particularly when several (>5) loci are involved in adaptation (Friedman et al. 2013). The model is fully sympatric, meaning that incipient species freely exchange genes despite having completely distinct niches, as might perhaps occur for species inhabiting a well-mixed aquatic environment but specializing on different dissolved nutrients. Qualitatively, the model fit well with the observation of relatively few niche-specifying genes (~3–10) involved in the ecological differentiation of marine vibrios (Shapiro et al. 2012) and suggested that barriers to gene flow (either ecological or physical) might be required to maintain the separateness of species, especially when niche adaptation involves many genes.

Marttinen and Hanage took the next logical step by modelling evolution in two ecological niches with an adjustable level of overlap (Marttinen and Hanage 2017). In this Overlapping Habitat Model (OHM), individuals exchange genes and compete only in their overlapping region of multidimensional niche space (Fig. 1). Unlike *symsim*, which explicitly models the niche-specifying genes, the OHM assumes that niche adaptation is caused by very many loci, such that the recombination of just a few of these loci does not affect niche preference. Using this model, Marttinen and Hanage were able to investigate the rates of genetic divergence under different levels of niche overlap and recombination. Intuitively, with low levels of niche overlap (~20% or less), speciation occurs rapidly due to (implicit) divergent selection between niches and reduced opportunity for genetic exchange (which can only occur in the region of niche overlap). With high niche overlap (~60%), speciation is slow and genetic distances within and between subpopulations (nascent species) continue to overlap significantly, making species difficult to distinguish (as in the case of *V. cyclitrophicus*). Fitting the OHM to real population genomic data, two putative subpopulations of *S. pneumoniae* are predicted to have 41% niche overlap and two putative subpopulations of *C. jejuni* to have 24% overlap. The model further predicts that with fast divergence (no niche overlap), all genes across the genome rapidly accumulate ecoSNPs, similar to the genome-wide divergence predicted by the SEM. With higher niche overlap, ecoSNPs are predicted to accumulate in just a few genes, with most genes containing zero or very few ecoSNPs. This pattern of few dense ecoSNP clusters was observed in both *S. pneumoniae* and *C. jejuni* genomes, suggesting their gradual divergence in the presence of gene flow in partially overlapping niches (Fig. 1). Qualitatively, this also resembles the three dense patches of ecoSNPs in *V. cyclitrophicus* described above, suggesting that the OHM could capture speciation processes in a range

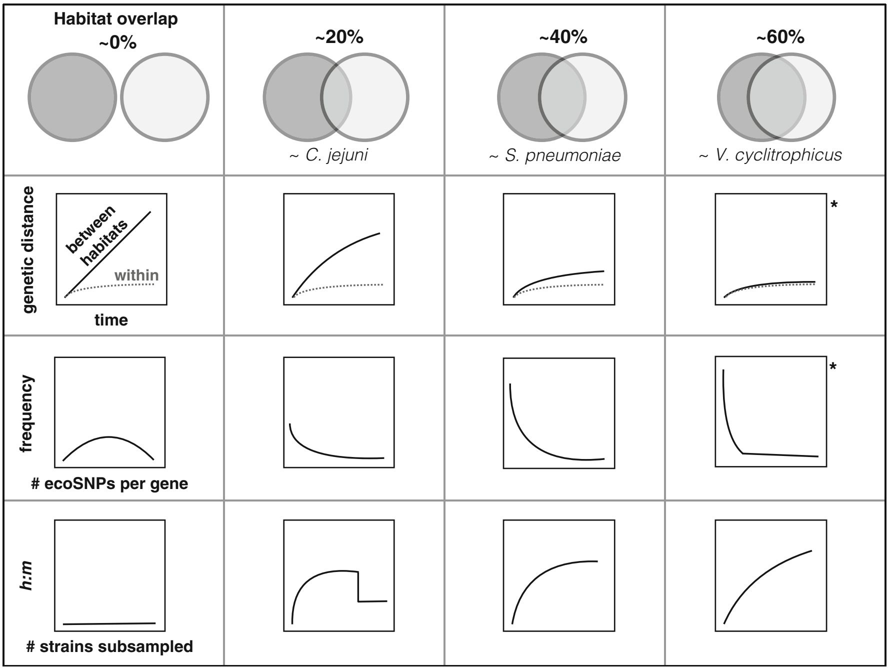

Fig. 1 Population genomic signatures of speciation under the Overlapping Habitat Model (OHM). The first (*top*) row illustrates the extent of habitat overlap between two populations. Populations can live and recombine in their respective habitat or in the region of overlap. Habitats exist in multidimensional niche space. The second row illustrates the genetic distances within and between populations, as predicted by the OHM. When there is little habitat overlap, the two populations diverge rapidly, but as overlap increases, distinct populations become difficult to distinguish from within-species genetic variation. The third row illustrates the predicted distribution of ecoSNPs (fixed single nucleotide differences between the two populations) per gene. The fourth row shows the estimated median homoplasy/mutation (*h:m*) ratio as increasingly large subsamples of genomes are taken from the populations. With ~0% habitat overlap, no recombination is expected between populations; thus the *h:m* ratio will be close to zero, and species are undefinable by the BSC-based method of Bobay and Ochman. In the example of *C. jejuni* (~20% overlap), a discontinuity is observed in the *h:m* ratio, suggesting the existence of two distinct species. The top three panels qualitatively summarize Figs. 2, 4, and 5 from Marttinen and Hanage (2017). Note that the OHM was fit to *C. jejuni* and *S. pneumoniae* datasets, but not *V. cyclitrophicus*. The panels marked with an asterisk are therefore hypothetical, based on the results of Shapiro et al. (2012). The bottom panel qualitatively summarizes portions of Supplementary Fig. 1 from Bobay and Ochman (2017)

of natural bacteria. Because the OHM does not model niche-specifying genes (the genes under divergent selection between niches), it follows that clusters of ecoSNPs in the genome can arise even when these ecoSNPs are not the direct targets of selection. As a consequence, ecoSNP clusters can be either drivers or passengers of the speciation process.

The OHM is appealing for its seamless combination of the ESC and the BSC. Ecology and divergent selection are implicit in the overlap of abstract multidimensional niches.

Barriers to gene flow occur as a consequence of nonoverlapping (or minimally overlapping) niches. The percentage of overlap in multidimensional niche space is a rather abstract concept but provides a point of entry for researchers to determine the main drivers of niche overlap (e.g. physical separation, host preference, nutrient utilization, growth rates, or some combination of these).

5 Species Delimitation Using Population Genomic Data

As discussed above, operational species definitions (such as a 95% ANI threshold) can easily be used to delimit species using population genomic data in the absence of any particular species concept. A more profound use of population genomic data is to detect signals predicted by a specific species concept and define species based on this concept. For example, the BSC predicts higher levels of gene flow within than between species. (Strictly, the BSC predicts zero gene flow between species, a criterion that will never realistically be met in recombining bacteria and archaea; hence only BSC-like concepts are amenable to most microbes and possibly most macrobes; Mallet et al. 2015; Shapiro et al. 2016.) Based on mounting population genomic evidence of higher rates of recombination within than between species or suspected species (Cadillo-Quiroz et al. 2012; Shapiro et al. 2012; Krause and Whitaker 2015; David et al. 2017), a BSC-like concept could plausibly apply to a large variety of microbes. In this BSC-like concept, barriers to gene flow provide a signature of speciation, but the drivers of speciation are not specified.

Bobay and Ochman (2017) recently proposed a way to apply a BSC-like concept to define species based on population genomic data. The method begins with a set of aligned genomes from a putative species (e.g. named species downloaded from NCBI GenBank) and identifies SNPs in the alignment. SNPs are then divided into those that can be placed parsimoniously on the phylogenetic tree, attributed to point mutation, and those that cannot: homoplasies, attributed to recombination. These two classes of SNPs are used to estimate the ratio of recombination to mutation rates (*r:m*) from the ratio of homoplasies to parsimonious mutations (*h:m*). If the alignment includes genomes sampled from just one species, sampling additional genomes will allow the SNP calling procedure to converge on a stable *h:m* ratio. However, if a "contaminant" genome from a second species is added to the alignment, this will cause an abrupt drop in the estimate of *h:m*, because under a BSC-like model, most mutations occurring between species are due to mutation, not recombination. The method therefore accepts "good" species as those that converge on a stable *h:m* estimate and proposes to split species containing "contamination" from other species. Importantly, Bobay and Ochman's method also identifies species that are too clonal (i.e. species with a very low *h:m*) and therefore cannot be classified based on a BSC-like concept.

Studying 105 named species from NCBI GenBank, Bobay and Ochman found that just over half constitute "good" species, about a quarter should be split, and about a quarter are too clonal or lack sufficient numbers of informative SNPs to be defined (e.g. *Mycobacterium tuberculosis* and *Bacillus anthracis*). Encouragingly, the method identifies a species boundary between familiar animal species such as humans and chimpanzees. The two named species analysed with the OHM model, *S. pneumoniae* and *C. jejuni*, were also included in Bobay and Ochman's analysis, providing an opportunity for comparison (although not exactly the same set of genomes were used). In the *C. jejuni* genomes, a clear discontinuity was identified by Bobay and Ochman (Fig. 1), suggesting that this species should be split in two according to the BSC-like concept. In contrast, *S. pneumoniae* behaves as a single cohesive species (Fig. 1). At face value, this contradicts the OHM model, which predicts that *S. pneumoniae* contains two gradually diverging subpopulations that might be considered distinct species. However, the divergent *S. pneumoniae* subpopulation (SC12) identified by the OHM was not represented in Bobay and Ochman's dataset, highlighting the importance of sampling for any population genomic study of speciation or species delimitation. The two nascent species of *V. cyclitrophicus* (Shapiro et al. 2012) were not identified as distinct species based on the BSC-like criterion, likely because divergence was too recent and barriers to gene flow do not yet extend across the genome. Therefore, very early stages of speciation may be difficult to detect based on a genome-wide gene-flow criterion.

Bobay and Ochman's method is attractive for two main reasons. First, it is not based on any arbitrary threshold of genetic similarity, but rather upon a discontinuity in inferred rates of gene flow. As a result, even if some very early stages of speciation may be missed, the method can delimit species across a range of genetic divergences. Second, it is based on genome sequences, meaning it can be readily and reproducibly applied across a range of different species (including bacteria, archaea, eukaryotes, or even viruses) without "expert" knowledge or complicated phenotypic tests. It also comes with some caveats. For practical reasons, the method tests the coherence of an a priori hypothesized species; it does not define species de novo from a database of all sequenced genomes. More importantly, the method depends strongly on sample size and in fact relies on unbalanced sampling between species for discontinuities in gene flow to be identified. As such, the method is optimized to detect single "contaminant" genomes but will fail to distinguish two species sampled in roughly equal proportions. Like any comparative genomic method, it only measures realized (rather than potential) genetic exchange. Under the strict BSC, individuals that *can* exchange genes are members of the same species, even if in practice they do not (e.g. due to geographic separation and the population structure that results). Determining the potential for genetic exchange requires experiments. All that can be reasonably asked of a comparative genomic method is to assess the realized rates and boundaries of recombination. Therefore, the method provides a useful starting point for further investigation. If a species is split, researchers must go on to ask, was the split due to population structure or ecological differentiation? If the latter, what are the relevant ecological niches?

6 Conclusions

Here I have described how speciation can be initiated by ecological differentiation (an ESC-like species concept) and be maintained by barriers to gene flow (a BSC-like species concept). Population genomic evidence from several groups of bacteria support this “ESC + BSC” paradigm, but there are sure to be exceptions. In effectively nonrecombining bacteria, the BSC does not apply. In some groups of bacteria or archaea, speciation could be driven entirely by barriers to gene flow, but strong examples are still lacking. Even in cases where gene-flow barriers appear to maintain species, it is not clear whether these barriers *initiated* speciation (Cadillo-Quiroz et al. 2012; Krause and Whitaker 2015). Moreover, the distinction between ESC and BSC may be somewhat artificial, because ecological differentiation can create barriers to gene flow, for example, when incipient species favour different hosts or particles (Shapiro et al. 2012; Meyer et al. 2016). This combination of the ESC and BSC is elegantly modelled in the Overlapping Habitat Model, in which gene flow occurs only in the region of niche overlap (Marttinen and Hanage 2017). In many instances, ecological specialization and barriers to gene flow may occur effectively simultaneously, which would explain why the two potential drivers of speciation have proven so difficult to disentangle.

Population genomic and, in some cases, metagenomic data have the potential to delimit species in a standard, reproducible way. For example, genomes that differ at more than 5% of nucleotide sites tend to belong to different species (Konstantinidis and Tiedje 2005; Konstantinidis et al. 2006). While this simple cutoff-based species delimitation may work well in many cases, there are exceptions that are better resolved using concept-based delimitation. For example, *Prochlorococcus marinus* includes genomes that share only 72% average nucleotide identity, but this group still behaves as a coherent gene-flow unit according to a BSC-based species delimitation (Bobay and Ochman 2017). On the other hand, it is well established that there are several, if not hundreds, of genetically and ecologically distinct subclusters within *Prochlorococcus* which appear to stably coexist in the ocean (Rocap et al. 2003; Johnson et al. 2006; Kashtan et al. 2014). It may not matter if there are 1000, 100, or only one species of *Prochlorococcus*—but it is useful to note that *Prochlorococcus* appears to be a relatively homogeneous unit of gene flow, which may contain finer-scale units that go undetected by certain methods (Bobay and Ochman 2017). Similarly, *S. pneumoniae* shows finer genetic substructure within the two major subpopulations, suggesting fine-scale niche partitioning (Marttinen et al. 2015; Marttinen and Hanage 2017). Therefore, although species delimitation methods (Bobay and Ochman 2017) and speciation models (Marttinen and Hanage 2017) can provide impressive fits to the major features of population genomic datasets, these methods and models generally provide only a starting point—a very useful starting point—for more detailed investigations into the ecology, phenotypes, and genetics of the organisms in question.

With the possible exception of experimental evolution experiments, it is effectively impossible to follow a speciation event from start to finish in real time. However, if speciation is indeed common—and it must be if all organisms can be placed somewhere along a speciation spectrum (Mallet 2008; Shapiro and Polz 2014)—studying diverse microbes at different stages of speciation will allow us to more fully appreciate the order of events driving and maintaining speciation, the general mechanisms involved, and the inevitable exceptional cases.

Acknowledgements I am grateful to the Canada Research Chairs program for funding and to members of my laboratory for useful discussions and comments that improved the manuscript.

Glossary

Niche 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 between niches. For the purposes of this chapter, "niche" and "habitat" are used more or less interchangeably, although "habitat" has a more spatial connotation, while niches can be temporal, behavioural, physiological, etc.

Ecological species concept (ESC) A species concept in which speciation is driven by adaptation to distinct habitats or ecological niches, with each species inhabiting a distinct niche.

Biological species concept (BSC) A species concept based on reproductive isolation (in the strict sense) or to barriers to gene flow, resulting in more gene flow within than between species, even if some between-species gene flow still occurs.

Allopatric speciation Speciation driven by physical barriers to gene flow between incipient species, such that speciation may occur in the absence of natural selection.

Sympatric speciation Speciation that occurs in the absence of physical barriers to gene flow, such that speciation must be driven by some combination of natural selection and/or genetic barriers to gene flow.

Mosaic sympatry An intermediate between sympatry and allopatric, in which organisms inhabit different niches (e.g. particles or hosts) within an otherwise well-mixed environment.

Gene flow A general term for exchange of DNA between chromosomes, including both homologous and nonhomologous DNA. In sexual organisms, gene flow occurs during meiosis. In microbes, gene flow can occur by phage-mediated transduction, plasmid-mediated conjugation, or natural competence (uptake of free DNA) followed by homologous or nonhomologous recombination.

Gene-specific selective sweep The process in which an adaptive gene or allele 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.

Genome-wide selective sweep The process in which an adaptive gene or allele spreads in a population by clonal expansion of the genome 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.

ecoSNP An ecologically associated single nucleotide polymorphism (SNP) with different nucleotides fixed between two different habitats (e.g. an A allele in habitat 1 and a T allele in habitat 2). Genes under divergent natural selection between niches or habitats ("niche-specifying genes") are expected to contain a large number of ecoSNPs.

References

Acinas SG, Klepac-Ceraj V, Hun DE, Pharino C, Ceraj I, Distel DL, Polz MF. Fine-scale phylogenetic architecture of a complex bacterial community. Nature. 2004;430:551–4.

Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. Binning metagenomic contigs by coverage and composition. Nat Methods. 2014;11:1144–6.

Bao Y-J, Shapiro BJ, Lee SW, Ploplis VA, Castellino FJ, Didelot X, Maiden MCJ, Gevers D, Shapiro BJ, Polz MF, et al. Phenotypic differentiation of streptococcus pyogenes populations is induced by recombination-driven gene-specific sweeps. Sci Rep. 2016;6:36644.

Bendall ML, Stevens SL, Chan L-K, Malfatti S, Schwientek P, Tremblay J, Schackwitz W, Martin J, Pati A, Bushnell B, et al. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. ISME J. 2016;10:1589–601.

Blount ZD, Borland CZ, Lenski RE. Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. Proc Natl Acad Sci U S A. 2008;105:7899–906.

Blount ZD, Barrick JE, Davidson CJ, Lenski RE. Genomic analysis of a key innovation in an experimental Escherichia coli population. Nature. 2012;488:513–8.

Bobay L-M, Ochman H. Biological species are universal across life's domains. Genome Biol Evol. 2017;9:491–501.

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

Caro-Quintero A, Konstantinidis KT. Bacterial species may exist, metagenomics reveal. Environ Microbiol. 2011;14:347–55.

Charron G, Leducq JB, Landry CR. Chromosomal variation segregates within incipient species and correlates with reproductive isolation. Mol Ecol. 2014;23:4362–72.

Coleman ML, Chisholm SW. Ecosystem-specific selection pressures revealed through comparative population genomics. Proc Natl Acad Sci U S A. 2010;107:18634–9.

Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, McGee L, von Gottberg A, Song JH, Ko KS, et al. Rapid pneumococcal evolution in response to clinical interventions. Science. 2011;331:430–4.

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

Cui Y, Yang X, Didelot X, Guo C, Li D, Yan Y, Zhang Y, Yuan Y, Yang H, Wang J, et al. Epidemic clones, oceanic gene pools and epigenotypes in the free living marine pathogen Vibrio parahaemolyticus. Mol Biol Evol. 2015;32:1396–410.

David S, Sánchez-Busó L, Harris SR, Marttinen P, Rusniok C, Buchrieser C, Harrison TG, Parkhill J. Dynamics and impact of homologous recombination on the evolution of Legionella pneumophila. PLoS Genet. 2017;13:e1006855.

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

Doolittle WF. Population genomics: how bacterial species form and why they don't exist. Curr Biol. 2012;22:R451–3.

Doolittle WF, Zhaxybayeva O. On the origin of prokaryotic species. Genome Res. 2009;19:744–56.

Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, Sogin ML. Methods Ecol Evol. 2013;4:1111–9.

Fraser C, Hanage WP, Spratt BG. Recombination and the nature of bacterial speciation. Science. 2007;315:476–80.

Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. The bacterial species challenge: making sense of genetic and ecological diversity. Science. 2009;323:741–6.

Friedman J, Alm EJ, Shapiro BJ. Sympatric speciation: when is it possible in bacteria? PLoS One. 2013;8:e53539.

Gause GF. The struggle for existence. Baltimore: Williams & Williams; 1934.

Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, de Peer YV, Vandamme P, Thompson FL, et al. Opinion: re-evaluating prokaryotic species. Nat Rev Microbiol. 2005;3:733–9.

Gregory AC, Solonenko SA, Ignacio-Espinoza JC, LaButti K, Copeland A, Sudek S, Maitland A, Chittick L, Dos Santos F, Weitz JS, et al. Genomic differentiation among wild cyanophages despite widespread horizontal gene transfer. BMC Genomics. 2016;17:930.

Hanage WP. Fuzzy species revisited. BMC Biol. 2013;11:41.

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

Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF. Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science. 2008;320:1081–5.

Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EM, Chisholm SW. Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients. Science. 2006;311:1737–40.

Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, Ding H, Marttinen P, Malmstrom RR, Stocker R, et al. Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. Science. 2014;344:416–20.

Klinger CR, Lau JA, Heath KD. Ecological genomics of mutualism decline in nitrogen-fixing bacteria. Proc Biol Sci. 2016;283:20152563.

Koeppel AF, Wu M. Species matter: the role of competition in the assembly of congeneric bacteria. ISME J. 2014;8:531–40.

Koeppel AF, Perry EB, Sikorski J, Krizanc D, Warner A, Ward DM, Rooney AP, Brambilla E, Connor N, Ratcliff RM, et al. Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. Proc Natl Acad Sci U S A. 2008;105:2504–9.

Kondrashov AS, Mina MV. Sympatric speciation: when is it possible? Biol J Linn Soc Lond. 1986;27:201–23.

Konstantinidis KT, Tiedje JM. Towards a genome-based taxonomy for prokaryotes. J Bacteriol. 2005;187:6258–64.

Konstantinidis KT, Ramette A, Tiedje JM. The bacterial species definition in the genomic era. Philos Trans R Soc Lond B Biol Sci. 2006;361:1929–40.

Krause DJ, Whitaker RJ. Inferring speciation processes from patterns of natural variation in microbial genomes. Syst Biol. 2015;64:926–35.

Leducq J-B, Nielly-Thibault L, Charron G, Eberlein C, Verta J-P, Samani P, Sylvester K, Hittinger CT, Bell G, Landry CR. Speciation driven by hybridization and chromosomal plasticity in a wild yeast. Nat Microbiol. 2016;1:15003.

López-Pérez M, Rodriguez-Valera F. Pangenome evolution in the marine bacterium Alteromonas. Genome Biol Evol. 2016;8:1556–70.

Majewski J, Cohan FM. Adapt globally, act locally: the effect of selective sweeps on bacterial sequence diversity. Genetics. 1999;152:1459–74.

Mallet J. Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. Philos Trans R Soc Lond B Biol Sci. 2008;363:2971–86.

Mallet J, Besansky N, Hahn MW. How reticulated are species? Bioessays. 2015;38:140–9.

Marttinen P, Hanage WP. Speciation trajectories in recombining bacterial species. PLoS Comput Biol. 2017;13:e1005640.

Marttinen P, Croucher NJ, Gutmann MU, Corander J, Hanage WP. Recombination produces coherent bacterial species clusters in both core and accessory genomes. Microb Genom. 2015;1:e000038.

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

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

Meyer JR, Dobias DT, Medina SJ, Servilio L, Gupta A, Lenski RE. Ecological speciation of bacteriophage lambda in allopatry and sympatry. Science. 2016;354:1301–4.

Niehus R, Mitri S, Fletcher AG, Foster KR. Microbial genomes into multiple niches. Nat Commun. 2015;6:1–9.

Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, Gautier L, Pedersen AG, Le Chatelier E, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol. 2014;32:822–8.

Popa O, Landan G, Dagan T. Phylogenomic networks reveal limited phylogenetic range of lateral gene transfer by transduction. ISME J. 2016;11:543–554.

Porter SS, Chang PL, Conow CA, Dunham JP, Friesen ML. Association mapping reveals novel serpentine adaptation gene clusters in a population of symbiotic Mesorhizobium. ISME J. 2016;11:248–62.

Retchless AC, Lawrence JG. Phylogenetic incongruence arising from fragmented speciation in enteric bacteria. Proc Natl Acad Sci U S A. 2010;107:11453–8.

Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR, et al. Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. Nature. 2003;424:1042–7.

Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pašić L, Thingstad TF, Rohwer F, Mira A. Explaining microbial population genomics through phage predation. Nat Rev Microbiol. 2009;7:828–36.

Rosen MJ, Davison M, Bhaya D, Fisher DS. Microbial diversity. Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. Science. 2015;348:1019–23.

Schluter D. Evidence for ecological speciation and its alternative. Science. 2009;323:737–41.

Shapiro BJ. Signatures of natural selection and ecological differentiation in microbial genomes. Adv Exp Med Biol. 2014;781:339–59.

Shapiro BJ. How clonal are bacteria over time? Curr Opin Microbiol. 2016;31:116–23.

Shapiro BJ, Polz MF. Ordering microbial diversity into ecologically and genetically cohesive units. Trends Microbiol. 2014;22:235–47.

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

Shapiro BJ, Leducq JB, Mallet J. What is speciation ? PLoS Genet. 2016;12:e1005860.

Takeuchi N, Cordero OX, Koonin EV, Kaneko K. Gene-specific selective sweeps in bacteria and archaea caused by negative frequency-dependent selection. BMC Biol. 2015;13:1–11.

Tilman D. Resource competition and community structure. Princeton: Princeton University Press; 1982.

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

Wiedenbeck J, Cohan FM. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. FEMS Microbiol Rev. 2011;35:957–76.

Yawata Y, Cordero OX, Menolascina F, Hehemann J-H, Polz MF, Stocker R. A competition-dispersal tradeoff ecologically differentiates recently speciated marine bacterioplankton populations. Proc Natl Acad Sci U S A. 2014;111:5622–7.