Signatures of Natural Selection and Ecological Differentiation in Microbial Genomes

17

B. Jesse Shapiro

Abstract

We live in a microbial world. Most of the genetic and metabolic diversity that exists on earth – and has existed for billions of years – is microbial. Making sense of this vast diversity is a daunting task, but one that can be approached systematically by analyzing microbial genome sequences. This chapter explores how the evolutionary forces of recombination and selection act to shape microbial genome sequences, leaving signatures that can be detected using comparative genomics and population-genetic tests for selection. I describe the major classes of tests, paying special attention to their relative strengths and weaknesses when applied to microbes. Specifically, I apply a suite of tests for selection to a set of closely-related bacterial genomes with different microhabitat preferences within the marine water column, shedding light on the genomic mechanisms of ecological differentiation in the wild. I will focus on the joint problem of simultaneously inferring the boundaries between microbial populations, and the selective forces operating within and between populations.

Keywords

Microbial genomics • Natural selection • Recombination • Reverse ecology • Evolution • Bacteria • Convergent evolution • McDonald-Kreitman test • Speciation • Ecological differentiation • Adaptive divergence • Vibrio • Long-range haplotype test

17.1 Introduction

Microbes are key players in global biogeochemical cycles, human health and disease; yet the microbial world is largely hidden from view. Even with the best microscopes and experimental techniques, it is exceedingly difficult to know the predominant selective

B.J. Shapiro (✉)
Département de sciences biologiques, Université de Montréal, Montréal, QC, Canada
e-mail: jesse.shapiro@umontreal.ca

C.R. Landry and N. Aubin-Horth (eds.), *Ecological Genomics: Ecology and the Evolution of Genes and Genomes*, Advances in Experimental Medicine and Biology 781,
DOI 10.1007/978-94-007-7347-9_17, © Springer Science+Business Media Dordrecht 2013

pressures and ecological interactions at play in the wild. Microbial genome sequences provide a comprehensive and accessible record of the forces that drive microbial evolution. Using a *reverse ecology* approach (Li et al. 2008; Whitaker and Banfield 2006), we can analyze genome sequences – for example, by deploying sequence-based statistical tests to identify genes under positive selection – in order to discover ecologically distinct populations and how they adapt to different niches. Our motivation for this line of research could be driven by basic curiosity about the microbial world, but could also have practical goals in both environmental (e.g. linking microbial populations to nutrient cycles) and clinical spheres (e.g. understanding mechanisms of pathogenesis). The long-term goal of reverse ecology is to gain a mechanistic understanding of ecological processes, but short-term benefits can be expected along the way. For example, genes or mutations that are consistently associated with niches or phenotypes of interest (e.g. antibiotic resistance) can serve as diagnostic biomarkers, helping to predict environmental or clinical outcomes and suggesting effective interventions.

Gaining biological insight from microbial genome sequences and tests for selection poses several challenges. First, there are challenges arising from the enormous range of microbial evolutionary time scales: we may be interested in comparing species that diverged hundreds of millions to billions of years ago, or that diverged so recently that it is unclear if they constitute separate species or not. Second, while it was once thought that microbes do not form species in the classical sense because they reproduce clonally and do not recombine their DNA through sex, the idea is now gaining popularity that they do not form proper species because they have *too much* promiscuous sex, due to their ability to exchange genes by horizontal transfer spanning great genetic distances (Doolittle and Papke 2006).

In this chapter, I will begin by explaining how the problem of defining bacterial species is inextricably bound to the process of natural selection. I will then describe how genomic sequence data, analyzed with appropriate statistical and computational methods, can distinguish among evolutionary hypotheses, and ultimately provide insight into the structure and function microbes in their natural environments. The chapter will focus mainly on relatively closely-related (same genera or species) populations of 'wild' bacteria (i.e. outside of lab or microcosm settings). My goal is to provide an introduction for readers new to microbial evolutionary genomics, while raising outstanding questions in the field and synthesizing knowledge in a way that is useful to more experienced readers.

The chapter begins by asking the question, how do we define and identify ecologically distinct species of bacteria (Sect. 17.2)? It then describes different models of speciation, and the importance of natural selection in these models (Sect. 17.3). The major classes of tests for natural selection in genome sequences are briefly described (Sect. 17.4), and applied to a population of natural *Vibrio* genomes (Sect. 17.5), focusing on the McDonald-Kreitman test (Sect. 17.5.2) and the long-range haplotype test (Sect. 17.5.3). Other methods that can be applied to detect selection in rarely recombining bacteria, including time course studies and convergent evolution, and in the 'flexible' (horizontally transferred) component of the genome, are discussed briefly (Sect. 17.6). The chapter closes with an outlook (Sect. 17.7) on how new datasets and populations models are beginning to be incorporated into a better understanding of microbial evolution and ecology.

17.2 Recombination and the Bacterial Species Problem

Partitioning biological diversity into discrete units is challenging in general, and perhaps most challenging in microbes. To begin with, microbes are by definition microscopic and we can only categorize them into a limited number of morphological classes based on cell wall characteristics, shape and size, presence or absence of flagella, etc. They are much more diverse in terms of physiology and metabolic capability, leading to the problem of how to properly weight an abundance of traits into a

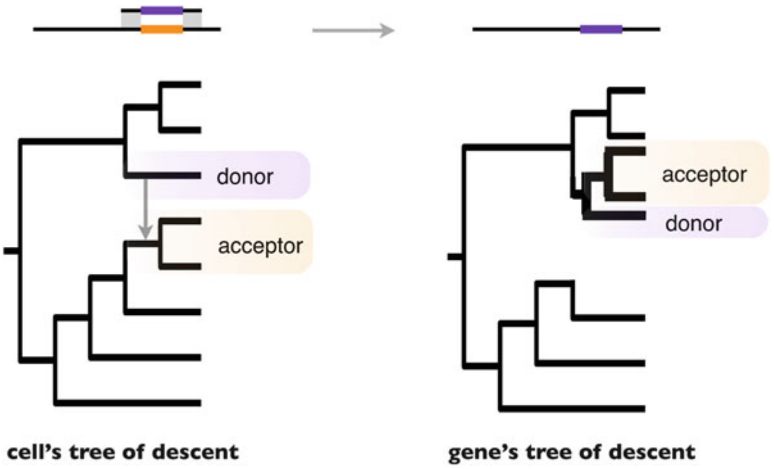


Fig. 17.1 Recombination can result in incongruence between gene-trees and cell-trees. A piece of DNA with flanking homology (*grey shading in upper panel*) is recombined from a donor into an acceptor genome, replacing the original allele (*orange*) with a new one (*purple*). The result is that the acceptor genome now has an identical allele as the donor, so the acceptor and donors branch closely together on the gene tree (*right*; branch lengths not to scale), whereas the acceptor and donor cells share a much more distant common ancestor

meaningful species classification scheme. The same problem arises when trying to classify multicellular organisms into species based on shared traits.

One solution to this problem is to privilege genetic data over other measurable traits. The reasoning behind this solution is that genetic similarity provides the best evidence that two individuals are similar *by descent*, as opposed to *by chance* or *by convergence*. I will discuss the concept of convergence in Sect. 17.6, but for now let's explore the idea of descent. When we talk about descent in bacteria, we could mean at least two different things: one is the bifurcating tree of cellular descent by clonal cell division; the other is the tree describing the replication of DNA molecules. When a cell divides in two, its genome replicates into two copies as well. The tree of cellular descent is identical to the tree of genomic descent. Now imagine that after the first cell division, one of the daughter cells encounters a molecule of DNA in its environment. The DNA – let's say it encodes an allele of a gene already present in the genome – enters the cell and replaces the original version of the gene by the mechanism of homologous recombination. The history of this gene is now different from the history of the cell. They are described by different trees. In the cell's tree, the two daughter cells branch together. In the gene's tree, the daughter that accepted the foreign DNA branches with the source of that DNA rather than with the other daughter (Fig. 17.1). I am intentionally using the word 'tree' instead of 'phylogeny' because the latter usually implies relationships between species, whereas my intention is to more generally describe patterns of descent. So which tree do we care about, the gene's tree of descent or the cell's?

Let's begin by examining the cell's tree. This tree describes the exponential process of binary cell division. The tree topology remains the same, no matter how many genes have been swapped for different alleles. In what I will call the purely *clonal* scenario, absolutely no genes have been swapped by recombination. In the extreme opposite of the clonal scenario, genes are exchanged at a rate that far outpaces cell division, so the tree for any given gene will have nothing to do with the cell's tree. In the clonal scenario, the gene's tree and the cell's tree are identical, so DNA sequence data from any (or every) gene in the genome can be used to infer, using a model of sequence evolution, the correct tree of cellular descent with reasonable statistical certainty. Now we have a trustworthy tree, but we are still left with the problem of defining species: where should we make a cut in the tree to divide one species from another? Just like species definitions based on morphology or physiology, we are faced with a decision. Should we make an arbitrary cut in the tree, perhaps dividing branches with greater than 95 % DNA sequence similarity across the genome? A given threshold is generally chosen because it provides a good empirical match to other species definitions (Konstantinidis and Tiedje 2005), but this type of reasoning quickly becomes circular.

17.3 Natural Selection and Speciation

17.3.1 Models of Bacterial Speciation

Up until this point, I have focused on using genetic similarity to infer patterns of descent. But is this really what we want from a bacterial species concept? I argue that we should care more about the *process* that generates genetic similarity than the exact level of genetic similarity itself. The process is an *evolutionary process*, involving natural selection of the fittest within a diverse, replicating population. A good example of a process-based species concept is the Ecotype Model, developed by Cohan and others (see (Cohan and Perry 2007) for a comprehensive overview). In its simplest form, the Ecotype Model defines species as independent evolutionary units. If a mutation occurs in the genome of a member of one species, it only competes with members of the same species, all sharing the same ecological niche. If the mutation is adaptive, genomes containing the mutation will multiply more rapidly, or escape predation more effectively, than those without the mutation, eventually dominating the population in a *selective sweep*. Importantly, the selective sweep will have absolutely no effect on other populations that compete in different ecological niches. New species emerge when a member of an existing species gains a function (by mutation or recombination) that allows it to exploit a new ecological niche, founding a new evolutionarily independent population. The process described by the Ecotype Model generates *clusters of ecological and genetic similarity*. Although it has been suggested that clusters of genetic similarity could arise through neutral processes (by mutation and genetic drift alone, without natural selection), theory suggests that selection is required for microbes to differentiate into genotypic clusters (Polz et al. 2013).

In the Ecotype Model, the evolutionary process of natural selection is paramount. But what about the process of recombination? Taken to an extreme, recombination will obscure clusters of genetic similarity because different genes will have different trees, leading us to infer different clusters. Adding selection to this scenario of extreme recombination yields a model that I will call Gene Ecology. In this model, genes, not species, inhabit ecological niches and are the targets of natural selection. Species only exist insofar as genes have to work together in order to reproduce themselves within genomes. This model may be extreme – ignoring epistasis and gradual coevolution among genes – but it can be a useful tool for understanding the distribution of different genes in different environments (Coleman and Chisholm 2010; Delong 2006; Mandel et al. 2009).

One way of moderating the extreme gene-centrism of Gene Ecology is to introduce elements of Mayr's Biological Species concept (de Queiroz 2005; Mayr 1942). This concept defines species based on reproductive isolation, so strictly speaking, it does not apply to asexually-reproducing bacteria. In sexual reproduction, genes are recombined every generation. Reproductive isolation therefore results in separate gene pools. In bacteria, reproduction is decoupled from recombination, and genes from very distantly related bacteria can be exchanged by recombination (Koonin et al. 2002). Therefore, we can never expect bacterial species to have completely isolated gene pools. But we needn't discard the Biological Species concept entirely. In bacteria that recombine frequently, different genes could be selected in different niches, independently of the cell or genome that they (transiently) inhabit. This begins to resemble Gene Ecology. But if there is preferential recombination among cells in the same niche (due to physical proximity, or

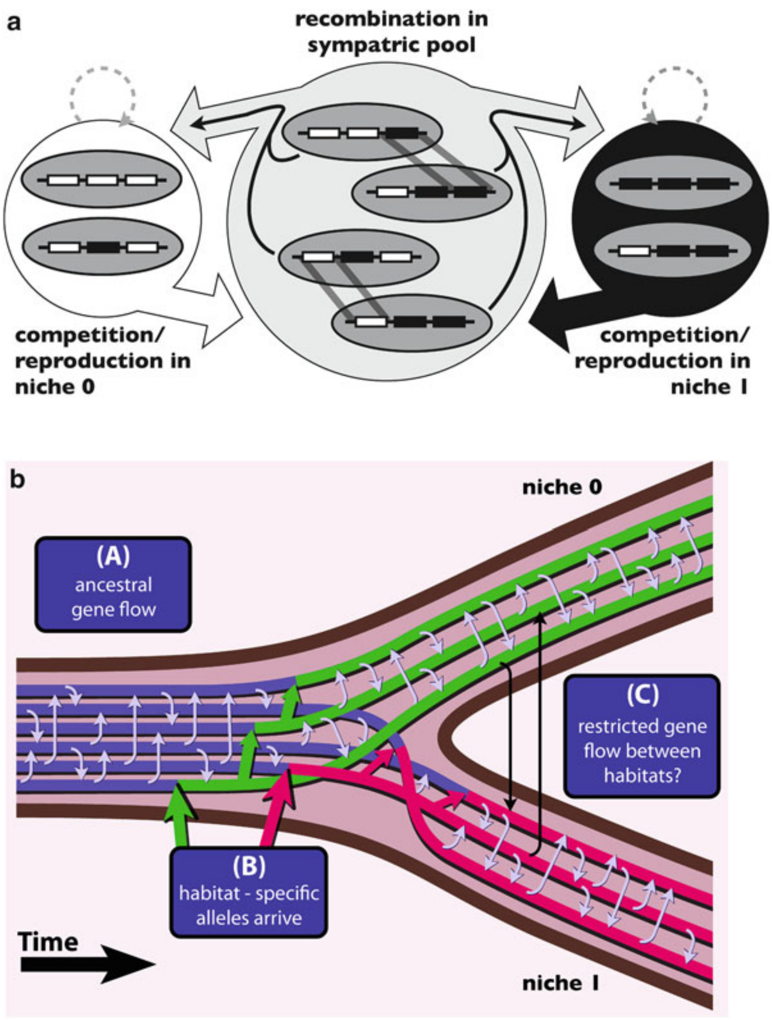

a
recombination in sympatric pool
competition/ reproduction in niche 0
competition/ reproduction in niche I
b
niche 0
(A) ancestral gene flow
(C) restricted gene flow between habitats?
(B) habitat - specific alleles arrive
Time
niche I

Fig. 17.2 A model of ecological differentiation for sympatric recombining bacteria. (**a**) A sympatric model (Modified from Friedman et al. 2013) in which microbial cells (*dark grey ovals*) compete in either of two niches. Cells containing mostly *black* alleles are best adapted to niche 1; *white* alleles to niche 0. Gene conversion of homologous loci (*diagonal lines*) take place in a sympatric, mixed pool of genotypes from both niches, and cells return to the niche to which their genotype is best adapted (e.g. in this 3-locus example, genotypes with mostly white alleles go to niche 0; those with mostly black alleles go to niche 1). Some degree of allopatry could be added to the model by increasing the rate of recombination within niches (*dashed circular arrows*). (**b**) The resulting temporal dynamics of such a model, supported by data from *V. cyclitrophicus* populations adapted to large- or small-particle niches in the marine water column (Modified from Shapiro et al. 2012. Reprinted with permission from John Kaufmann and from © AAAS 2012. All Rights Reserved)). *Thin gray* or *black arrows* represent recombination within or between ecologically associated populations. *Thick red* or *green colored arrows* represent acquisition of adaptive alleles for different habitats/niches

increased efficiency of recombination due to DNA sequence similarity, or both), a hybrid of Gene Ecology and Biological Species might apply. This is the sort of hybrid model that I proposed for a pair of closely-related, recombining populations of marine *Vibrio* bacteria, described in Sect. 17.5.1 and illustrated in Fig. 17.2.

I hope that you now have some appreciation for the connections between speciation, recombination and selection. From here on, I won't focus any further on the bacterial species problem

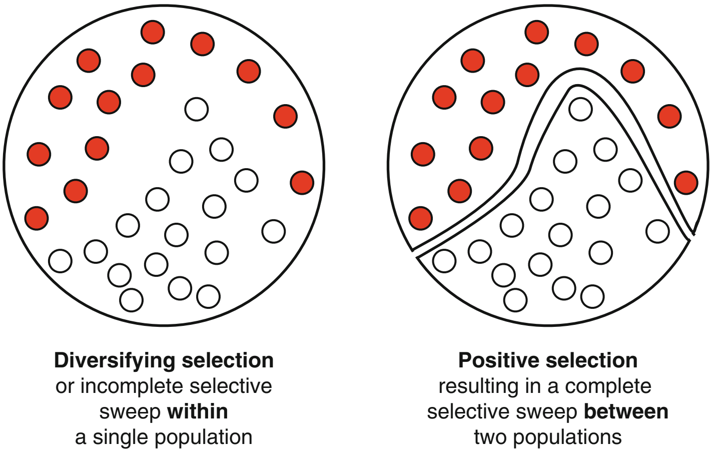


Fig. 17.3 The concepts of positive and diversifying selection depend on population boundaries. The *right* and *left* panels differ only in the definition of boundaries between populations. *Small circles* represent individual sampled bacteria with different alleles (*red filled or empty circles*) at a polymorphic locus

per se, but instead on the process of ecological differentiation, which by definition is driven by selection for adaptation to different niches. I take this 'adaptationist' perspective because it is likely to describe the behavior of many microbial populations on earth. With some exceptions, Baas-Becking's theory that "everything is everywhere; the environment selects" (Baas-Becking 1934) has been largely supported by observations of natural microbial populations. This means that microbial populations are generally *sympatric* (part of the "same country," without geographic structure, e.g. freshwater cyanobacteria described in van Gremberghe et al. 2011) and are rarely separated by physical separation, as occurs in *allopatric* speciation (a rare microbial instance of which is presented by hotspring archaea; see Whitaker et al. 2003). In reality, most microbes probably fall on the spectrum between absolute sympatry and allopatry. The point is that physical separation is much less important for microbes than for most species of plants and animals. As a result, natural selection should be the most important contributor to ecological differentiation and speciation (Fig. 17.2).

17.3.2 Forces of Natural Selection

Natural selection can operate in a variety of ways, but I find it useful to consider three major forms of selection: negative, positive, and diversifying selection. Negative (sometimes called purifying) selection is the tendency of unfit individuals to reproduce less and therefore to be eliminated from the population. This results in traits or genes remaining *conserved* over time, because deleterious genetic variants are weeded out. Positive selection favors the survival and reproduction of variants conferring a competitive advantage over the rest of the population. During a *selective sweep*, positively selected variants replace unselected variants. Diversifying selection can be thought of as favoring incomplete selective sweeps. For example, in a special case of diversifying selection called negative frequency-dependent selection, a mutation is favored by positive selection when it is at low frequency, but becomes deleterious at high frequency. The mutation never sweeps the entire population, but fluctuates around an intermediate frequency. Depending on how boundaries between populations are drawn, diversifying and positive selection can be hard to distinguish (Fig. 17.3).

17.4 Signatures of Selection and Adaptive Divergence

The goal of microbial ecological and evolutionary genomics is to use genetic sequences sampled from microbial populations to learn how these populations adapt to different niches. To solve this reverse ecology problem, we need to identify signatures of selection and niche adaptation in microbial genomes. A whole battery of sequence-based statistical tests for selection have been developed, but because most of them were designed for sexual populations we must be careful which tests we choose to apply to asexual microbes (Shapiro et al. 2009). The basic premise of these tests is to define patterns of genetic variation that are shaped by selection, and distinguish them from the *neutral* patterns expected by random mutation and genetic drift.

One of the most popular tests for selection involves comparing the relative rates of nonsynonymous (amino acid-changing) to synonymous mutations, often called the *dN/dS* ratio. The key assumption is that nonsynonymous changes (measured by *dN*) affect protein structure, change the phenotype, and are thus subject to natural selection. Synonymous changes have no effect on protein structure, and are thus subject to less natural selection and reflect mostly random mutation and genetic drift. In fact, synonymous mutations may also be under selection for RNA stability, translational efficiency, etc., e.g. (Gingold and Pilpel 2011; Raghavan et al. 2012), but the *dN/dS* test assumes that selection is generally stronger on nonsynonymous mutations. Imagine that we have sequenced orthologous protein-coding genes from two species, aligned the two sequences and counted nucleotide differences between them. We can then count the differences as synonymous or nonsynonymous according to the genetic code, and normalize the counts by the number of synonymous or nonsynonymous sites, respectively, to obtain *dN* and *dS*. Averaged across the entire gene, $dN/dS \approx 1$ suggests very little selection at the protein level (characteristic of pseudogenes), $dN/dS > 1$ suggests very strong positive selection to fix different amino acids between species, and $dN/dS < 1$ suggests negative selection for a conserved protein structure in both species. I have deliberately chosen one of the simplest possible applications of *dN/dS* in order to illustrate the principle, but applications of *dN/dS* can be tailored to consider more than two species, or to consider separately individual codons within a gene (Yang 2008; Yang and Nielsen 2002).

A powerful extension of *dN/dS* called the McDonald-Kreitman (MK) test (McDonald and Kreitman 1991) employs protein-coding sequences sampled both within and between species. We have already discussed the difficulties in defining species boundaries in bacteria, but the MK test can still be useful. If we are satisfied with a species concept based on adaptive divergence (i.e. ecological differentiation driven by positive selection), the MK test can be flipped on its head: rather than testing for positive selection between a priori defined species, we are instead testing whether a sample of gene sequences come from the same or different species (Simmons et al. 2008; Vos 2011). I will return to the 'flipped' MK in Sect. 17.5.2, but will first describe the original version of the test.

The key assumption of the MK test is that in the absence of selection, the *dN/dS* ratio should remain constant over time and, thus, be the same for fixed substitutions (between species) as for segregating polymorphism (within species). The MK test normalizes *DN/DS* (measured between a pair of species) by *PN/PS*, the equivalent measure *within* one or both of the species. In this case *D* and *P* refer to divergence and polymorphism, and *DN, DS, PN,* and *PS* are the absolute counts (rather than rates per nonsynonymous or synonymous site) of each category of mutation in the gene of interest. The Fixation Index (FI) is defined as (*DN/DS*)/(*PN/PS*), with FI >1 suggesting positive selection between species and FI <1 suggesting negative selection. The MK test is preferable to simply computing *dN/dS* because an average $dN/dS > 1$ across an entire gene is very unlikely, even if a few individual amino acids are genuinely under positive selection. By normalizing by *PN/PS*, the MK test is more sensitive. Second, $dN/dS > 1$ may occur due to a *relaxation of negative selection* rather than positive selection, whereas FI >1 is much more likely to indicate positive selection only.

Tests for selection need not be based on the genetic code, like *dN/dS* and the MK test. Another group of tests that I will collectively call *allele-frequency* and *haplotype-frequency* tests look for mutations (or clusters of mutations linked together as alleles or haplotypes) that have risen to an unexpectedly high frequency in the population, suggesting positive or diversifying selection. Allele-frequency tests, such as Tajima's *D* (Tajima 1989) or Fay and Wu's *H* (Fay and Wu 2000), calculate mutation frequencies within a gene or region of interest, under the assumption of no recombination within it. Haplotype-frequency tests, including the long-range haplotype (LRH) test and its variations, explicitly consider recombination as a sort of 'clock' (Sabeti et al. 2002; Voight et al. 2006). When a new mutation occurs in the genome, it is necessarily linked to other mutations on the chromosome. This haplotype of mutations is initially long, spanning the entire chromosome. In sexual population, recombination occurs with some frequency every generation by crossing-over of homologous chromosomes. This results in the slow erosion of the haplotype, from the edges of the chromosome toward the new mutation. As a result, older mutations will be part of shorter haplotypes than newer mutations. If they are neutral to fitness, new mutations should not rise very quickly, or at all, to high frequency in the population. But if they are subject to positive selection, they are more likely to increase in frequency. If the increase in frequency is fast relative to the recombination 'clock,' selected mutations will tend to be observed at high frequency on unexpectedly long haplotype backgrounds.

The LRH test was designed with sexual populations in mind, and is not expected to work in bacteria – at least not in its original form. While many bacteria are capable of homologous recombination, they do so by gene conversion rather than crossing over. Instead of eroding linear haplotypes from the edges inward, gene conversion generates a characteristic 'patchwork' pattern known as the *clonal frame* (Milkman and Bridges 1990). The clonal frame refers to the chromosome background, with its own clonal ancestry and tree topology (presumably congruent with the tree of cellular descent), which is interrupted by short recombinant blocks, usually a few kilobases (kb) that have been introduced by gene conversion. These recombinant blocks have different evolutionary histories than the clonal frame. In the clonal frame model, because gene conversion events are of fairly uniform size, there should be little association between haplotype length and frequency in positively selected regions of the genome. Therefore, the original formulation of the LRH test is not strictly applicable to bacteria.

17.5 Testing for Selection in Bacterial Genomes

In the last section, I touched on some of the issues involved in applying tests for selection to recombining bacterial genomes. On the one hand, if bacteria are perfectly clonal (no recombination), every gene in the genome will be linked in the same clonal frame. When an adaptive mutation occurs in a particular genome in the population, the resulting selective sweep will bring to high frequency not only the adaptive mutation, but any other neutral or slightly deleterious mutations that happen to be 'hitchhiking' in the same genome (Hanage et al. 2006; Shapiro et al. 2009). Selective sweeps therefore purge population diversity genomewide, and it becomes difficult, based on any of the tests described above, to distinguish adaptive mutations from hitchhikers (Fig. 17.4). On the other hand, if bacteria recombine frequently or promiscuously, care must be taken to ensure that recombination does not obscure or lead to false signals of selection. Recombination has the potential to introduce adaptive alleles (by homologous recombination), or entirely new genes or operons (by illegitimate recombination, often mediated by phage, plasmids or integrative conjugative elements). We could in principle design tests to look for recombination events that are adaptive, based on a consistent association with a niche or phenotype of interest, unexpectedly high population frequency, or recombination across species boundaries. For example, in an analysis of *Streptococcus* genomes, we found that genes recombined between recognized species tended to have higher *dN/dS* than genes recombined within species, suggesting that recombination across species boundaries requires positive selection (Shapiro et al. 2009).

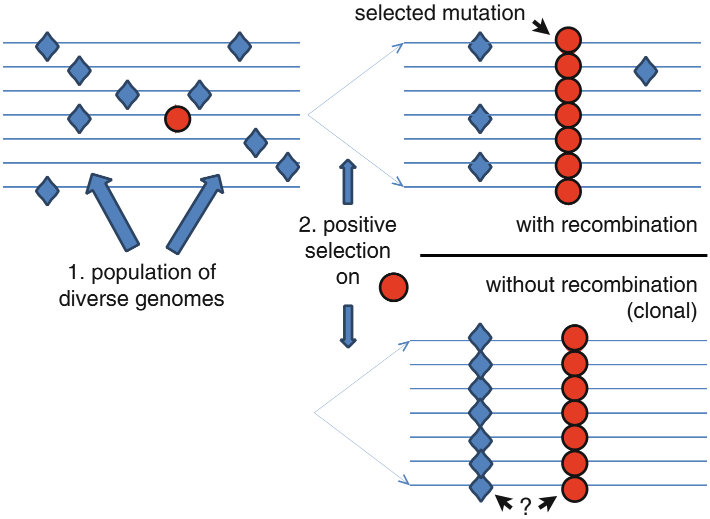

selected mutation
1. population of diverse genomes
2. positive selection on
with recombination
without recombination (clonal)
?

Fig. 17.4 In clonal populations, selected mutations (*red circle*) can be confused with neutral mutations (*blue diamonds*) in the genome (*horizontal line*)

I will now walk through a workflow for detecting regions of the genome under selection in populations of bacteria. I will use the example of marine *Vibrio cyclitrophicus*, which my colleagues and I have studied extensively (Hunt et al. 2008; Shapiro et al. 2012; Szabó et al. 2013), highlighting aspects of the analyses that can be generalized to other data, and focusing on the interplay between selection and ecological differentiation.

17.5.1 Ecological Differentiation Among Marine *Vibrio*

In 2006, we sampled coastal seawater off the coast of Massachusetts, and ran it through a series of progressively finer filters. We then isolated *Vibrio* from each of the filters on *Vibrio*-selective media. I will focus on two groups of isolates: those from the largest filter, which catches mainly large particles (>63 μm, consisting primarily of zooplankton) and those from an intermediate filter, which catches small particles that are still larger than a typical *Vibrio* cell (∼1 μm in diameter). The large and small particles are proxies for different microhabitats in the water column, and thus constitute a potential axis of ecological differentiation.

We sequenced 20 whole genomes from two closely-related clusters of *V. cyclitrophicus* that appeared to have undergone a recent habitat switch, finding that just a few loci in the genome appear to have driven the switch (Fig. 17.2b). We inferred that the ecological switch had been relatively recent because all the isolates had identical 16S sequences, and only differed by about ten mutations in the faster-evolving *hsp60* gene. Genomewide, 725 single nucleotide polymorphisms (SNPs) clearly partitioned the large – and small-particle isolates into two distinct groups. Surprisingly, these 725 'ecoSNPs' were not distributed evenly across the genome, but were clustered in only 11 regions, the three densest of which contained >80 % of ecoSNPs (Shapiro et al. 2012). Outside of these regions, SNPs tended to conflict with the partitioning of large- and small-particle isolates. The extent of recombination and conflicting phylogenetic signal is displayed in a STARRInIGHTS plot (Fig. 17.5a and b), showing many small, sometimes barely visible 'constellations' of support (in white) for many different phylogenetic partitions, none of which accounts for much of the genome. These conflicting phylogenetic signals suggested high rates of homologous recombination since the divergence of these isolates, with recombination breakpoints inferred to have occurred about once per kilobase. Together, this suggests that large- and small-particle populations actually constituted a single homogeneously recombining

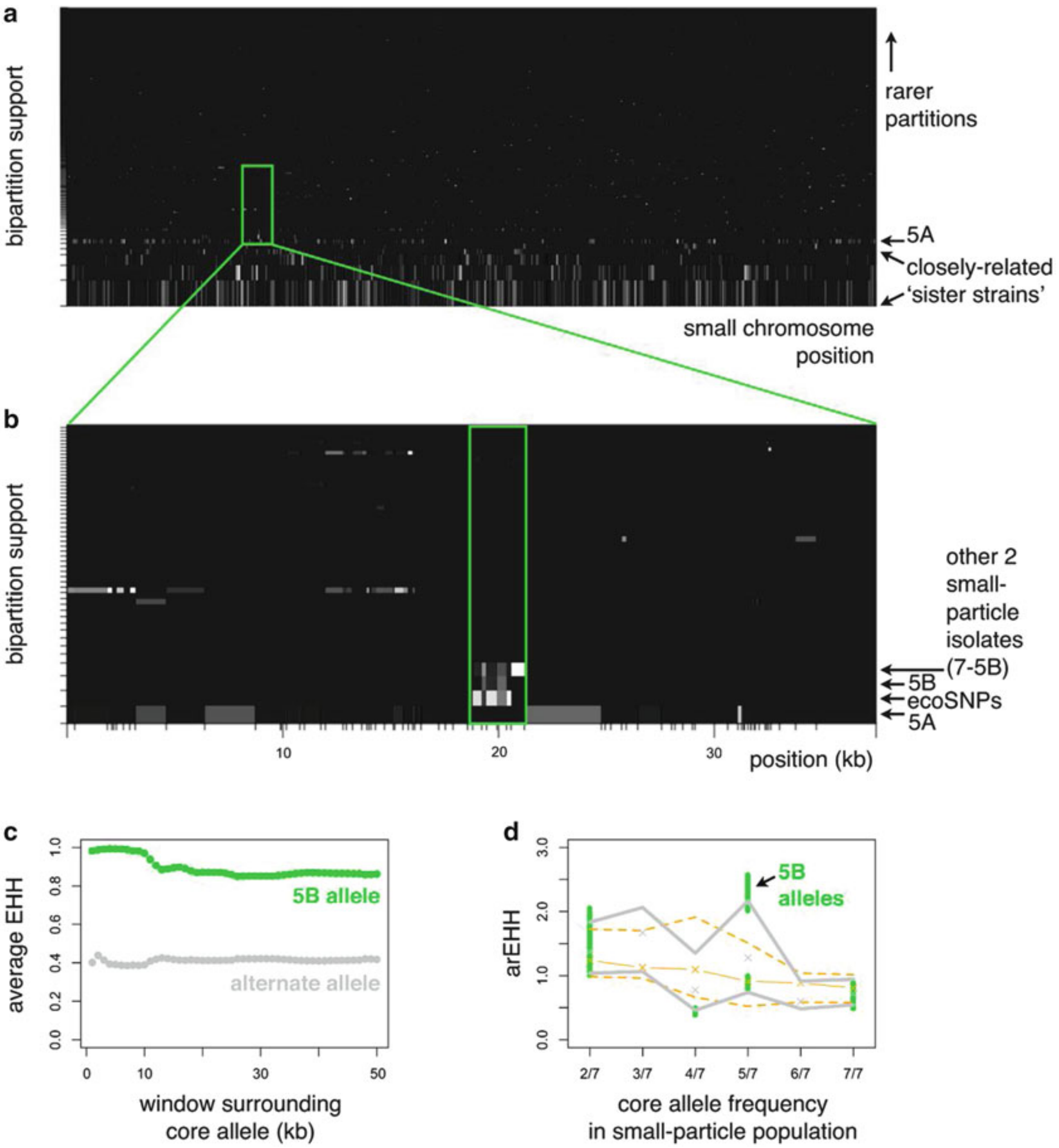


Fig. 17.5 Recombination and selection at the *RpoS2/RTX* locus in *V. cyclitrophicus* genomes. (**a**) Recombination blocks supporting different phylogenetic partitions across the small chromosome. Strain-based Tree Analysis and Recombinant Region Inference In Genomes from High-Throughput Sequencing projects (STARRInIGHTS; http://almlab.mit.edu/starrinights.html) was used to infer breakpoints between recombination blocks across the chromosome (x-axis). *Brighter white* indicates higher numbers of SNPs within a block supporting a particular partition of the 20 genomes. Partitions are ranked on the y-axis in increasing order of their prevalence genomewide. The row width is also proportional to genomewide prevalence of each partition. (**b**) Detail of 37.5 kb surrounding the *RpoS2/RTX* region (*green box*). *Small tick marks* on the x-axis indicate recombination breakpoints. (**c**) Decay of linkage (average extended haplotype homozygosity) with distance around a representative SNP in the *RpoS2/RTX* region. The 5B-supporting variant (*green*) is surrounded by a longer linked haplotype than the alternate allele (*grey*; present in the other two small-particle genomes and the large-particle outgroup). (**d**) SNPs within the ∼2 kb *RpoS2/RTX* region (*green points*) at frequency 5/7, supporting the 5B partition, have high average relative EHH (arEHH) compared to neutral simulations (*dashed orange lines*) and other sites on the small chromosome (*grey lines*). Lines denote upper and lower 95 % confidence bounds and x denotes the median arEHH

population for most of their history (Fig. 17.2b). The ecoSNP regions are the exception, and we reasoned that they might contain alleles conferring adaptation to different microhabitats, driving ecological differentiation. Certain genes in the 'flexible' genome, differing in their pattern of presence and absence across the 20 sampled genomes, are probably also involved in ecological differentiation, and I will discuss them briefly in Sect. 17.6.

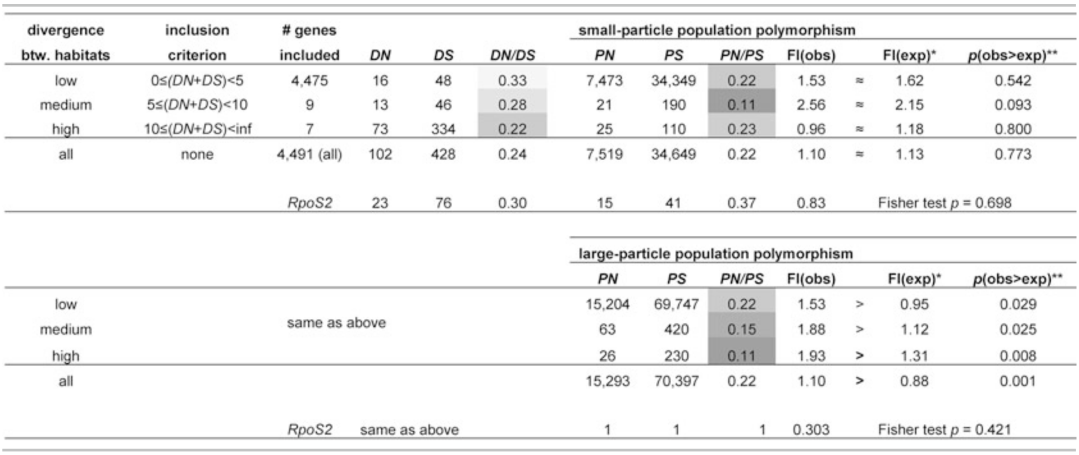
Before formally testing the ecoSNP regions for evidence of divergent positive selection between habitats, I will briefly discuss the implications of these regions having been acquired by recombination from very distant relatives of the 20 sequenced isolates – which could be considered evidence for selection in and of itself (Shapiro et al. 2009). Acquisition by recombination is by no means a necessary characteristic of positively selected loci, but it certainly adds a layer of evidence. There are two main reasons why we suspect the ecoSNP regions to have been acquired by recombination. First, they constitute only a small fraction of a genome that mostly rejects the ecoSNP phylogeny, making it highly unlikely that they are part of the clonal frame. Second, most genes in the ecoSNP regions have very high rates of synonymous substitutions (*dS*) between habitats; several times higher than anywhere else in the genome. Such high *dS* is best explained by recombination with relatives beyond the 20 genomes considered here. One consequence of such high *dS* is that, despite relatively high nonsynonymous divergence (*dN*), traditional tests for positive selection at the protein level (such as *dN/dS* and the MK test) suffer a substantial loss of power. Potentially due to this power loss, none of the genes in the three densest ecoSNP regions have FI significantly greater than 1 (Shapiro et al. 2012).

17.5.2 Insights from the MK Test

As alluded to in Sect. 17.4, the MK test can be 'flipped' in order to test whether two populations constitute distinct species (Shapiro et al. 2009; Vos 2011). Using a species concept based on adaptive divergence, Vos proposed that if the genomewide FI is significantly greater than 1 between populations, then these populations can be considered separate species (Vos 2011). Computing FI genomewide can be done by pooling genes, but combining genes with different histories of recombination, and different levels of polymorphism and divergence, can lead to biased estimates of FI. To control for this, the observed genomewide FI can be compared to the expected neutral distribution of FI, obtained by summing *DN*, *DS*, *PN* and *PS* across a set of bootstrapped contingency tables with marginal sums equal to those at each individual gene (Shapiro et al. 2007). By repeating this bootstrap resampling procedure 1,000 times, I was able to obtain an empirical *p*-value for the deviation of the observed from the expected FI.

Using all 4,491 aligned core *V. cyclitrophicus* genes, the genomewide FI is significantly greater than expected – but only when *PN* and *PS* are estimated from the large-particle population, not from the small-particle population (Table 17.1). Even though *PN/PS* is similar in both populations, the overall level of polymorphism is much lower (about half) in the small-particle population, which might explain the ambiguous results. In general, both *DN/DS* and *PN/PS* decrease as genes with progressively higher divergence (*DN* + *DS*) between habitats are included. If we accept that divergence measures evolutionary time, then this is consistent with purifying selection purging deleterious nonsynonymous mutations over time, both within and between populations. However, there is an exception to this trend: the highest *PN/PS* is actually observed in the seven most highly divergent genes, in the small-particle population (Table 17.1). This suggests diversifying selection among small-particle strains might be acting to increase *PN/PS*, specifically among the most divergent genes. Meanwhile, in the large-particle population, *PN/PS* is low among the most divergent genes, resulting in much stronger evidence for speciation in these genes ($\text{FI} = 1.93$, $p = 0.008$) than elsewhere in the genome. Overall, this reinforces that a few highly divergent genes seem to be driving ecological differentiation. However, it also reinforces

Table 17.1 MK test applied to core genes in *V. cyclitrophicus* ecological populations



divergence	inclusion	# genes				small-particle population polymorphism						
btw. habitats	criterion	included	*DN*	*DS*	*DN/DS*	*PN*	*PS*	*PN/PS*	FI(obs)		FI(exp)*	*p*(obs>exp)**
low	0≤(*DN*+*DS*)<5	4,475	16	48	0.33	7,473	34,349	0.22	1.53	≈	1.62	0.542
medium	5≤(*DN*+*DS*)<10	9	13	46	0.28	21	190	0.11	2.56	≈	2.15	0.093
high	10≤(*DN*+*DS*)<inf	7	73	334	0.22	25	110	0.23	0.96	≈	1.18	0.800
all	none	4,491 (all)	102	428	0.24	7,519	34,649	0.22	1.10	≈	1.13	0.773
		RpoS2	23	76	0.30	15	41	0.37	0.83		Fisher test *p* = 0.698	
						large-particle population polymorphism						
						PN	*PS*	*PN/PS*	FI(obs)		FI(exp)*	*p*(obs>exp)**
low						15,204	69,747	0.22	1.53	>	0.95	0.029
medium	same as above					63	420	0.15	1.88	>	1.12	0.025
high						26	230	0.11	1.93	>	1.31	0.008
all						15,293	70,397	0.22	1.10	>	0.88	0.001
		RpoS2	same as above			1	1	1	0.303		Fisher test *p* = 0.421	

[a] Mean FI in 1,000 bootstrap resamplings
[b] Based on 1,000 bootstrap resamplings

how different genes in the genome speciate at different rates (Retchless and Lawrence 2010), making it difficult to decide on a firm threshold between species.

Let's consider one of the ecoSNP regions as a candidate driver of ecological differentiation between large- and small-particle habitats: the single densest ecoSNP region, located on the smaller of the two chromosomes, encodes an RTX toxin and RpoS2, an RNA polymerase sigma factor involved in stress response. The *RTX* gene has sequence similarity to an excreted cytotoxic protein in *V. cholerae* (Lin et al. 1999). *RTX* is highly divergent between habitats, with ten fixed nonsynonymous changes and significant domain reorganization: the gene is split into three aligned coding regions, with other domains uniquely present in either small- or large-particle genomes only. The sigma factor appears to be a *Vibrio*-specific second copy of RpoS (hence the "RpoS2" designation), the first copy of which is located on the large chromosome. The *RpoS2* gene contains 23 fixed nonsynonymous differences (*DN*) between small- and large-particle isolates – the highest *DN* in the genome – three of which occur in predicted DNA binding domains (Lee and Gralla 2002). An additional two DNA binding residues differ between RpoS2 and the canonical RpoS, but are identical in large- and small-particle genomes. These observations suggest, first, that RpoS2 may target different DNA binding sites than the canonical stress-response sigma factor, and second, that RpoS2 may have experienced functional modifications between small- and large-particle isolates – potentially modulating major differences in gene expression between habitats. And yet, the MK test does not support this evidence of positive selection between habitats.

For the moment, let's assume that selection is real, and the MK test simply lacked power to detect it. There are at least two reasons why this could happen. First, in addition to the 23 fixed nonsynonymous differences, *RpoS2* also contains $DS = 76$ fixed synonymous differences (Table 17.1). This is consistent with *RpoS2* having diversified for a long time outside the populations considered here, and different alleles being acquired recently in different habitats (Fig. 17.2b). If all 99 substitutions were acquired simultaneously by recombination, even if some of the nonsynonymous substitutions were adaptive, the signal of positive selection might be obscured by high *DS*. Second, *RpoS2* contains a lot of polymorphism, particularly within the small-particle population, with a *PN/PS* ratio slightly higher than the *DN/DS* ratio (Table 17.1). This suggests that *RpoS2* might be under diversifying selection within the small-particle population, resulting in FI <1.

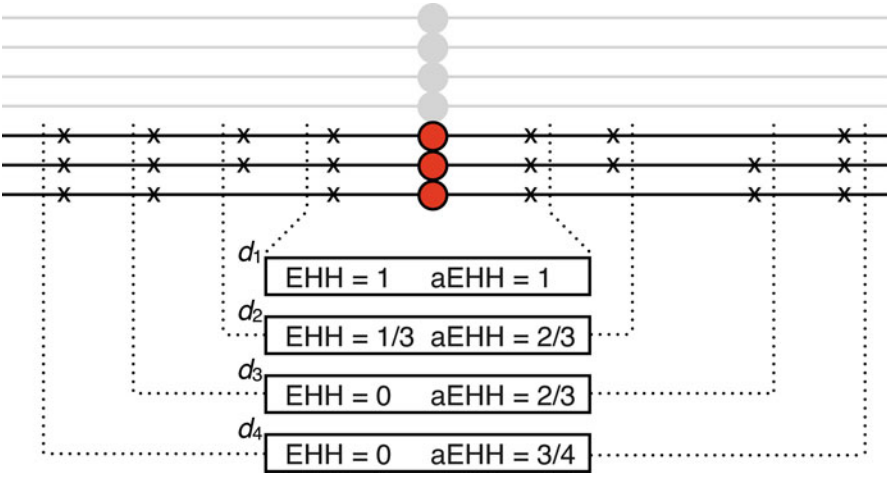


Fig. 17.6 Illustration of 'classical' and 'average' long-range haplotype (LRH) tests. The extended haplotype homozygosity (EHH) is defined in a window *d* around a core site (*circles*) as the probability that two randomly chosen chromosomes with the same allele at the core site (*black horizontal lines*) are also identical at *all* polymorphic sites within *d*. For example, in window d_1, all polymorphic sites have identical alleles (all have the 'x' allele), yielding EHH = 1. The average EHH (aEHH) is also 1. In window d_2, EHH and aEHH differ. This is because there are two identical haplotypes and one unique haplotype in d_2, yielding EHH = 1/3. However, aEHH is averaged over four polymorphic sites, with homozygosities of 1/3, 1, 1, and 1/3 (from *left to right*), for an average equal to 2/3. Note that EHH eventually decays to 0 with increasing distance from the core site, whereas aEHH can fluctuate up and down between 0 and 1. To compute the average *relative* EHH (arEHH), the aEHH values for the first allele at the core site (shown) would be normalized by aEHH in haplotypes around the other allele (*grey circles*; surrounding polymorphism not shown)

Because *DN/DS* = 0.30 is quite high (compared to *DN/DS* = 0.22 across other highly-divergent genes; see Table 17.1), within-population diversifying selection might be more important than elevated *DS* in explaining why the MK test does not detect significant positive selection between habitats.

17.5.3 Long-Range Haplotype Testing in Bacteria

I do not mean to push the hypothesis of positive selection on *RpoS2* too hard, merely to highlight some of the difficulties in applying and interpreting tests for selection. In the same spirit, let's investigate another test for selection: a variant of the LRH test. I explained in Sect. 17.4 that we don't expect the classical LRH test to work in microbes because of its assumptions about recombination. Yet a simple modification of the LRH test might be used to detect selection in bacterial genomes. In its original form (Sabeti et al. 2002), the LRH was implemented as follows. First, a 'core' nucleotide site is defined, with at least two polymorphic variants at the site. The extended haplotype homozygosity (EHH) is defined in a window *d* around the core site as the probability that two randomly chosen chromosomes from the population are identical (homozygous) at *all* single nucleotide polymorphisms (SNPs) in *d* (Fig. 17.6). EHH ranges from 0 (all extended haplotypes are different; rapid breakdown of linkage due to recombination) to 1 (all extended haplotypes are identical; perfect linkage, uninterrupted by recombination). Relative EHH is defined separately for each nucleotide variant at a core site as the EHH for the variant divided by the average EHH of other variants at the same locus, in order to control for the local recombination rate. Relative EHH ranges from 0 to infinity, with higher values indicating a variant on an unusually long haplotype, relative to other variants at the same locus. The LRH test then identifies variants with a high relative EHH that are also at high frequency in the population (relative to a modeled or genomewide empirical distribution of relative EHH), suggesting recent positive selection.

In bacteria, our goal is to identify regions of the genome that have been acquired by homologous gene conversion *and* are under positive selection. Several methods already exist to identify recombinant regions (Didelot and Falush 2007; Didelot et al. 2010; Marttinen et al. 2012; Shapiro et al. 2012), but they do not detect positive selection; this is where a version of the LRH test could be useful. Imagine that an operon of size 10 kb arrives in a bacterial genome by gene conversion, replacing the existing allele. The new allele comes from a distantly related lineage, but confers a selective advantage to its recipient. The allele can now spread in the populations by two mechanisms: gene conversion into other genomes, or clonal expansion of the genome that originally acquired the new allele. The predominant mechanism will depend on the strength of selection relative to the rate of recombination, and we can reasonably expect a combination of mechanisms to operate in bacteria with moderate to high recombination rates, like *Vibrio* and *E. coli* (Schubert et al. 2009; Touchon et al. 2009).

First consider a clonal expansion following acquisition of the new allele. The new allele will increase in frequency on a haplotype background consisting of the entire chromosome (the clonal frame). If this selected genome sweeps to fixation without any further recombination in either the selected genome or other genomes in the population, both selected and unselected alleles will have identical haplotype lengths, and any version of the LRH test will be uninformative (and the selected variant will be indistinguishable from other mutations linked in the clonal frame; see Fig. 17.4). If, however, some (neutral) recombination events occur within the population before and during the clonal expansion, we expect the selected allele to be in better *average* linkage with the clonal frame than the unselected allele at the same locus. It will be in better *average* linkage to sites anywhere on the chromosome, because gene conversion events will interrupt linkage for a few kb, at which point linkage to the clonal frame will resume. This contrasts with sexual crossing over, which interrupts linkage from the recombination point to the end of the chromosome. An LRH test modified for bacteria should therefore measure *average* linkage (homozygosity) in a window *d* around a core site, to allow for interruption and resumption of linkage (Fig. 17.6). Under certain circumstances, this 'average' LRH (aLRH) test could also be sensitive to selected alleles that spread primarily by multiple recombination events, rather than by clonal expansion. For example, in a population in which linkage only extends for ~1 kb (a realistic figure for *V. cyclitrophicus* and other populations such as *Leptospirillum* (Simmons et al. 2008) and *E. coli* (Mau et al. 2006) due to frequent, potentially overlapping gene conversions), an adaptive allele of ~10 kb that spreads rapidly by recombination of roughly the same block of DNA, might stand out as a long haplotype at high frequency.

I will now demonstrate how the aLRH test might work in practice, using the small chromosome of *V. cyclitrophicus* as an example. I do not intend this as a rigorous benchmarking of the test, but rather as an illustration of how such a test could be useful, and of its shortcomings.

The STARRInIGHTS plot, an illustration of recombination 'blocks' of the genome containing SNPs supporting different binary partitions of the isolates, shows that the *RpoS2/RTX* locus contains a dense cluster of ecoSNPs, supporting the partition between the 7 small- and 13 large-particle isolates (Fig. 17.5a and b). The partitions are ranked on the *y*-axis according to the prevalence of SNPs supporting them genomewide. The partition shown just below the ecoSNP partition, for example, is supported by 796 sites genomewide, slightly more than the 725 ecoSNPs. This partition – let's call it 5A – groups together a clade of 5 small-particle isolates, and is almost entirely restricted to the small chromosome (790/796 SNPs), strongly suggesting that the 5A strains share a closely related copy of the small chromosome. However, the *RpoS2/RTX* locus supports a grouping of a *different* five small-particle isolates – let's call it 5B – that is phylogenetically inconsistent with 5A. SNPs supporting 5B are common on the large chromosome (630 SNPs), and rare on the small chromosome. The only 52 5B-supporting SNPs on the small chromosome occur in the *RpoS2/RTX* region. This suggests that a new *RpoS2/RTX* allele spread by recombination through the 5B isolates, which all share a chromosomal background supporting the 5A phylogeny. Does this suggest positive selection acting to favor a new allele in a cryptic, ecologically distinct 5B population, or perhaps diversifying selection acting within the small-particle population?

To help evaluate these scenarios, I applied the aLRH test to the small chromosome of seven small-particle isolates and two large-particle isolates (as an outgroup). For each phylogenetically informative SNP (with a minor allele present in two or more isolates), I calculated the average homozygosity (average EHH) within windows of up to 50 kb centered around the SNP. The choice of window size is somewhat arbitrary, but was chosen to be sensitive to relatively long stretches of recombined DNA (a few genes or an operon), without picking up linkage across the clonal frame (spanning the whole chromosome, on the order of hundreds to thousands of kb). The average EHH surrounding a representative 5B-supporting allele in the *RpoS2/RTX* region stays near the maximum value of 1 within a window of ~10 kb, whereas the alternative allele (present in the two non-5B small-particle isolates and the large-particle isolates) has a stable average EHH of ~0.4 (Fig. 17.5c). We can then calculate the average relative EHH (arEHH) for the 5B variant as arEHH $\approx 1/0.4 \approx 2.5$ within a ~10 kb window. It turns out that this 5B-SNP, and others in the ~2 kb *RpoS/RTX* region, have exceptionally high values of arEHH, meaning that they have unusually long linked haplotypes for their frequency (Fig. 17.5d). Remarkably, it is mostly 5B-SNPs and *not* ecoSNPs (fixed in all seven small-particle isolates) that are responsible for the high arEHH in the region. This is a clear shortcoming of the aLRH test: the ecoSNPs are likely under selection, but do not occur on very long haplotypes, so they are missed.

In parallel to the 'real' small chromosome data, I analyzed a 'matched' dataset (with the same rate of polymorphism) from a coalescent model (Hudson 2002) simulating a population evolving neutrally and recombining by gene conversion (with population recombination rate $\rho = 1$ gene conversion of tract length 500 bp per generation; ρ as high as 10 and tract length of 5 kb also fit the real data reasonably well). Instead of the characteristic exponential decay of relative EHH with SNP frequency observed in humans and other sexual populations (Sabeti et al. 2002), both simulated and real microbial populations show a relatively uniform distribution of arEHH across SNP frequencies (Fig. 17.5d). The slight drop in arEHH of high-frequency SNPs might be due to small clusters of ecoSNPs recombined into different clonal frames, resulting in rapid decay of linkage across the recombination breakpoint. An alternative, but not exclusive explanation is that near-identical pairs of 'sister strains,' almost certainly sharing a clonal frame, are responsible for the slightly higher arEHH of SNPs at frequency 2/7. Whatever the explanation, factors such as population subdivision and uneven gene flow, are likely contributors to the difference between the real and simulated data. One clear example of this is the spike in arEHH at frequency 5/7 in the real data, corresponding to support for the 5A partition across the entire small chromosome (likely a feature of this chromosome's clonal frame phylogeny). And yet the 5B-supporting SNPs in the *RpoS2/RTX* region stand out as having an *even higher* arEHH than 5A-SNPs (shown as green points at frequency 5/7 in Fig. 17.5d). This certainly suggests that a relatively large segment of DNA has been recombined into the 5A isolates. Whether this is due to recent positive selection on a cryptic sub-population or diversifying selection within the small-particle population is unknown, but the aLRH analysis does show that the *RpoS2/RTX* region stands out from other loci in the genome, and from neutral simulations. It also demonstrates how the aLRH test can be used to explain the results of the MK test, and to more fully explore the complex layers of selective pressures on this locus.

17.5.4 Conclusions from *Vibrio* Ecological Genomics

What conclusions can we draw about natural selection acting on these *V. cyclitrophicus* populations, and in particular on ecologically differentiated regions of the genome such as *RpoS2/RTX*? First, these regions were probably acquired by recombination from more distantly related populations. The stable maintenance (of different allelic versions in different ecological populations) of foreign pieces of DNA – which we would generally expect to be maladaptive – immediately suggests divergent positive selection between microhabitats. However, because the evolutionary history of ecoSNP regions such as *RpoS2/RTX* is so different from the history of the ecological populations themselves, standard tests for selection give ambiguous results. Second, natural selection may be operating on different levels of organization: both between and within populations. The *RpoS2* gene, for example, is highly differentiated between large- and small-particle populations, but is also highly polymorphic within the small-particle population. This could suggest diversifying selection within the small-particle population, or divergent positive selection between two small-particle cryptic sub-populations adapted to unobserved niches. Without further knowledge of these unobserved niches, or evidence of reduced gene flow between putative sub-populations, it is difficult to distinguish between diversifying and positive selection scenarios (Fig. 17.3). Third – and related to the second point – selective processes are inextricably linked to ecological differentiation and speciation: the processes that define and bound populations. In the *V. cyclitrophicus* study, we were able to define populations based on both ecological (particle size preference) and biological (preference for recombination within rather than between populations) criteria. This allowed us to identify as candidates for selection genes and alleles that are highly divergent between populations. However, this does not guarantee that the populations we defined are by any means 'the most important,' and other important axes of ecological differentiation and selection may well exist. Fourth, 'flexible' genes, acquired by illegitimate recombination, may also be subject to habitat-specific selection and are likely involved in ecological differentiation. In most comparative studies of microbial genomes, it is almost implicitly assumed that this is the case (Haegeman and Weitz 2012). However, beyond looking for stable and significant associations between gene presence/absence and a particular niche or phenotype of interest, formal tests for selection on the flexible genome are not well developed. However, neutral models of genome evolution are starting to be explored (Haegeman and Weitz 2012), and tests for selection based on deviation from such models will hopefully follow.

Finally, the population genomic tests for selection I have described will never constitute unassailable proof of positive selection. They can, however, generate strong and specific hypotheses about the mechanisms of ecological differentiation. The adaptive value of putative selected loci can be tested in competition experiments. For example, isogenic *Vibrios* with different *RpoS2* alleles could be competed in microcosms containing different mixtures of large- and small-particles obtained from filtered, sterilized seawater. Data can also be compared across studies to discern trends. In *E. coli* experimental evolution studies, *RpoS* was frequently mutated or duplicated in replicate populations subjected to heat stress (Riehle et al. 2001; Tenaillon et al. 2012). This reinforces that *RpoS* genes may be frequent targets of selection in novel environments.

How general are these conclusions, and to what extent can the analyses I've described in *Vibrio* be applied to other populations of microbes? One study of hotspring *Sulfolobus* archaea showed a similar pattern of ecological differentiation with gradual reduction in gene flow between populations (Cadillo-Quiroz et al. 2012). However, the population-specific regions (equivalent to ecoSNP regions) were much larger (several hundred kb) and widely distributed across the genome, making it difficult to pinpoint genes under divergent positive selection between habitats.

The *Sulfolobus* study is also a truer example of reverse ecology than the *Vibrio*. This is because the two *Sulfolobus* populations were not predefined based on known ecology, but were inferred *de novo* based on comparative genomics. They were subsequently found to have distinct growth characteristics, but more work will be required to understand the nature of their niche partitioning, and the genetic changes that drive it.

The MK test failed to identify any genes under selection between *Sulfolobus* populations, either because many genes are involved in ecological differentiation, each under only weak selection, or because of lack of recombination within the extended 'continents' of population-specific differentiation. While the MK test lacked power to pinpoint selected genes within large linked continents, a variant of the LRH test might be successful in genomes with longer stretches of linkage, as appears to be more the case in the *Sulfolobus* than the *Vibrio* genomes.

In a study of vaccine evasion by serotype-switching in highly recombining *Streptococcus pneumoniae*, the recombination events leading to serotype switches were generally much larger (∼20–40 kb) than other recombinations in the genome (Croucher et al. 2011). This suggests that serotype-switch alleles may have spread rapidly in the population due to selection (probably some form of frequency-dependent selection) before the long stretches of linkage could be eroded by recombination. *Streptococcus* could therefore be another good candidate to apply a variant of the LRH test for selection. Generally speaking, the balance between selection and recombination within and between populations will dictate the sorts of tests for selection that can be applied (Fraser et al. 2009; Shapiro et al. 2009). In the next section, I will describe tests that perform well regardless of the recombination rate.

17.6 Convergence and Evolution in Real Time

So far, I have been considering 'cross-sectional' samples of genomes at a single moment in time. It is remarkable how much can be inferred about the evolutionary history of a population based on this kind of data, but microbes give us the opportunity to go further. In many natural microbial populations, large population sizes and high genetic diversity mean that the exact same mutation is likely to arise independently in different genomes. This allows a form of pseudo-replication, even in natural settings, that provides the basis for a class of tests for selection based on convergent evolution. Moreover, many bacteria replicate very rapidly, allowing us to track them in real time, and effectively watch evolutionary processes in action. The best examples of positive or diversifying selection in action come from viruses (HIV-1 in particular), and some of the techniques developed in their study (e.g. Pybus and Rambaut 2009) will prove very useful as time-course sequencing of natural bacterial populations becomes more feasible. For example, *dN* and *dS* can be measured precisely over time in replicate populations in order to estimate the strength of selection on nonsynonymous mutations (Neher and Leitner 2010).

In some instances, evolution in 'stasis' is just as interesting as evolution in action. In 2009, we isolated *Vibrio* from the same coastal location, using the same sampling strategy as 2006 in order to test whether the ecological associations we initially observed were stable over time (Szabó et al. 2013). Although some of the populations observed in 2006 were not observed at all in 2009 (perhaps due to association with cryptic niches, or stochastic extinctions), most of those that were observed again also had the same habitat association as in 2006. This suggests that particle-associated microniches are, to a large extent, stable over time. In particular, the small- and large-particle associated *V. cyclitrophicus* populations were re-identified in 2009, again based on shared habitat association and similarity in the *hsp60* phylogenetic marker gene. We did not sequence whole genomes of the 2009 isolates, but instead asked specifically whether the habitat-specific flexible genes were still associated with the same habitats in 2009. This type of stable association of flexible genes is not necessarily expected, because flexible genes are exchanged frequently by recombination in these populations, resulting in even the most closely-related pairs of 'sister' strains (almost certainly sharing a recent clonal ancestor) differing by several flexible genes (Shapiro et al. 2012). In fact, all five flexible genes tested in a PCR-based screen were consistently habitat-associated in 2009.

These genes, involved in mannose-sensitive hemagglutinin (MSH) pilus biosynthesis, intercellular adhesion, and surface carbohydrate biosynthesis, are generally present in large- but not small-particle associated isolates. Given the expected high flexible genome turnover, it is difficult to explain the stable habitat-association of these genes without invoking selection. The involvement of the MSH genes in attachment to chitinous surfaces (Frischkorn et al. 2013; Meibom et al. 2005) also suggests they may be positively selected in large-particle strains, which likely rely on attachment to chitinous copepods or diatoms. The stable habitat association of these genes strongly suggests, but of course does not prove, that they are under habitat-specific positive selection. Alternatively, the association could be maintained by a preference for habitat-specific recombination. Even though we did observe such a preference, it was only discernible in a small fraction of the genome (Shapiro et al. 2012). Given the high turnover of flexible genes, it is hard to imagine how recombination alone could result in the stable gene-habitat association without invoking some form of selection. Overall, this is consistent with findings from other bacteria showing that flexible gene content tends to be structure by horizontal transfer within ecological niches with similar selective pressures (Boucher et al. 2011; Smillie et al. 2011).

As discussed extensively throughout this chapter, *dN/dS*, the MK test, allele-frequency- and haplotype-based tests for selection will often lack power or fail in highly clonal (rarely recombining) populations. An attractive alternative is a class of tests for selection based on convergent evolution, the independent fixation of the same mutation in different clonal backgrounds. Given a sufficient sample of related genomes, convergence has the potential to pinpoint positive selection to a single mutation. Alternatively, if several different mutations in the same gene confer the same selected phenotype, then convergence at the gene level would provide a sensitive test. For example, in a pioneering application of convergence to study pathogenic *E. coli*, Sokurenko and colleagues found a significant excess of nonsynonymous mutations in the adhesin gene *FimH*, specifically in lineages (tips of the phylogenetic tree) that independently transitioned from a gut commensal to a uropathogenic phenotype (Sokurenko 2004). Both *dN/dS* and allele-frequency tests lacked power to detect selection in *FimH*, whereas convergence provided strong evidence for selection at the level of the entire protein, and at individual amino acid sites. Sokurenko and colleagues began with a specific hypothesis about selection on *FimH*, but convergence can also be applied genomewide. For example, a comparative genomic study of *Salmonella* Typhi identified convergent mutations in *gyrA* as likely targets of selection for antibiotic resistance (Holt et al. 2008). Depending on the mutation rate and evolutionary time scales considered, a certain level of 'background' convergence is expected even in the absence of selection (Rokas and Carroll 2008). Provided that care is taken to control for this, convergence will provide a powerful framework for detecting selection in a wide variety of microbial genomes, ranging from highly clonal to highly recombining.

17.7 Outlook

In this chapter, I have drawn from selected examples to illustrate how one should go about assessing evidence for selection and ecological differentiation among microbial genomes. I introduced the major classes of tests for selection, and attempted to highlight their strengths and weaknesses, and cases in which special care is warranted in interpreting them. In particular, haplotype-based tests and tests for 'unexpected' recombination across great phylogenetic distances, although not yet fully developed for use in bacteria, have great potential for inferring selection in recombining populations (Shapiro et al. 2009). The MK test provides a useful framework for assessing adaptive divergence genomewide, but often lacks power when applied to loci recently acquired by recombination with distance relatives. Convergence tests and time-course studies provide powerful tools for both recombining and clonal populations, provided that sufficient evolutionary time has transpired for convergent mutations to occur, or to observe changes in allele frequencies over time. There is currently no 'standardized' way to detect selection in microbial genomes; the current 'best practice' is to apply the tests that are justified given the recombination rate, and to interpret test results with a critical eye.

I focused on comparisons among closely-related whole genome sequences, but of course there are other rich sources of data that can be used to test for selection. The most obvious of these is metagenomic shotgun sequencing data, which yields allele-frequency information, but very limited information on how sequences of DNA are linked together into genomes. This accentuates the joint problem of detecting selection within or between species, while simultaneously defining the boundaries between species. However, when combined with ecological metadata and/or time course sampling, metagenomics can be a powerful tool for identifying genes, alleles and populations associated with particular phenotypes or ecological niches (Delong 2006; Denef et al. 2010b). Associations between protein expression and environmental metadata can be revealing as well. For example, proteins involved in cobalamin biosynthesis, phosphate uptake and motility were more highly expressed in late- than early-colonizing strains of *Leptospirillum* in an acid mine community, suggesting particular adaptations to the nutrient-limited environment experienced by late-colonizers (Denef et al. 2010a). Many of these differentially expressed proteins also had high *dN/dS* between early- and late-colonizers, further suggesting niche partitioning by positive selection.

As well as developing new and more refined tests for selection in microbial genomes and exploiting new datasets, we are also improving our basic understanding of the boundaries between microbial populations. In our current understanding, 'globally adaptive' genes may sweep through multiple sub-populations at once (Majewski and Cohan 1999), while sub-populations are maintained by local adaptation and preferences for within-population recombination. However, new findings are complicating models based on recombination and selection alone. For example, microbial populations might also be defined based on cooperation within but not between populations (Cordero et al. 2012). In such cases, it might be appropriate to consider selection at the level of the population, not just the individual or the gene. While genes are generally exchanged freely as 'public goods' (McInerney et al. 2011), they can sometimes become private to specific populations due to ecological boundaries, or barriers to recombination. An emerging challenge in microbial evolutionary genomics will therefore be to devise tests for selection that account for, and test the generality of, these new population models. Two papers have recently applied phylogenetic convergence tests for selection, discussed in Sect. 17.6, to identify genotype-phenotype associations. The tests were applied genomewide to identify variants associated with host-specificity (Sheppard et al. 2013) and antibiotic resistance (Farhat et al. 2013), respectively.

References

Baas-Becking LGM (1934) Geobiologie of Inleiding Tot de Milieukunde. W.P. Van Stockum & Zoon, The Hague

Boucher Y, Cordero OX, Takemura A et al (2011) Local mobile gene pools rapidly cross species boundaries to create endemicity within global *Vibrio cholerae* populations. mBio 2:e00335–10

Cadillo-Quiroz H, Didelot X, Held NL et al (2012) Patterns of gene flow define species of thermophilic archaea. PLoS Biol 10:e1001265

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

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

Cordero OX, Wildschutte H, Kirkup B et al (2012) Ecological populations of bacteria act as socially cohesive units of antibiotic production and resistance. Science 337:1228–1231

Croucher NJ, Harris SR, Fraser C et al (2011) Rapid pneumococcal evolution in response to clinical interventions. Science 331:430–434

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

Delong EF (2006) Community genomics among stratified microbial assemblages in the ocean's interior. Science 311:496–503

Denef VJ, Kalnejais LH, Mueller RS et al (2010a) Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. Proc Natl Acad Sci USA 107:2383–2390

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

Didelot X, Falush D (2007) Inference of bacterial microevolution using multilocus sequence data. Genetics 175:1251–1266

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

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

Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, Warren RM, Streicher EM, Calver A, Sloutsky A, Kaur D, Posey JE, Plikaytis B, Oggioni MR, Gardy JL, Johnston JC, Rodrigues M, Tang PK, Kato-Maeda M, Borowsky ML, Muddukrishna B, Kreiswirth BN, Kurepina N, Galagan J, Gagneux S, Birren B, Rubin EJ, Lander ES, Sabeti PC, Murray M (2013) Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. Nat Genet 45(10):1183–1189

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413

Fraser C, Alm EJ, Polz MF et al (2009) The bacterial species challenge: making sense of genetic and ecological diversity. Science 323:741–746

Friedman J, Alm EJ, Shapiro BJ (2013) Sympatric speciation: when is it possible in bacteria? PLoS ONE 8:e53539

Frischkorn KR, Stojanovski A, Paranjpye R (2013) *Vibrio parahaemolyticus* type IV pili mediate interactions with diatom-derived chitin and point to an unexplored mechanism of environmental persistence. Environ Microbiol 15:1416–1427

Gingold H, Pilpel Y (2011) Determinants of translation efficiency and accuracy. Mol Syst Biol 7:1–13

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

Hanage WP, Spratt BG, Turner KM, Fraser C (2006) Modelling bacterial speciation. Philos Trans R Soc Lond B Biol Sci 361:2039–2044

Holt KE, Parkhill J, Mazzoni CJ et al (2008) High-throughput sequencing provides insights into genome variation and evolution in *Salmonella typhi.* Nat Genet 40:987–993

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338

Hunt DE, David LA, Gevers D et al (2008) Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science 320:1081–1085

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

Koonin EV, Makarova KS, Wolf YI, Aravind L (2002) Horizontal gene transfer and its role in the evolution of prokaryotes. In: Syvanen M, Kado CI (eds) Horizontal gene transfer, 2nd edn. Academic, London, pp 277–304

Lee SJ, Gralla JD (2002) Promoter use by sigma 38 (rpoS) RNA polymerase. Amino acid clusters for DNA binding and isomerization. J Biol Chem 277: 47420–47427

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

Lin W, Fullner KJ, Clayton R et al (1999) Identification of a *Vibrio cholerae* RTX toxin gene cluster that is tightly linked to the cholera toxin prophage. Proc Natl Acad Sci USA 96:1071–1076

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

Mandel MJ, Wollenberg MS, Stabb EV et al (2009) A single regulatory gene is sufficient to alter bacterial host range. Nature 457:215–218

Marttinen P, Hanage WP, Croucher NJ et al (2012) Detection of recombination events in bacterial genomes from large population samples. Nucleic Acids Res 40:e6

Mau B, Glasner JD, Darling AE, Perna NT (2006) Genome-wide detection and analysis of homologous recombination among sequenced strains of *Escherichia coli.* Genome Biol 7:R44

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

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652–654

McInerney JO, Pisani D, Bapteste E, O'Connell MJ (2011) The public goods hypothesis for the evolution of life on earth. Biol Direct 6:41

Meibom KL, Blokesch M, Dolganov NA et al (2005) Chitin induces natural competence in *Vibrio cholerae*. Science 310:1824–1827

Milkman R, Bridges MM (1990) Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. Genetics 126:505–517

Neher RA, Leitner T (2010) Recombination rate and selection strength in HIV intra-patient evolution. PLoS Comput Biol 6:e1000660

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

Pybus OG, Rambaut A (2009) Evolutionary analysis of the dynamics of viral infectious disease. Nat Rev Genet 10:540–550

Raghavan R, Kelkar YD, Ochman H (2012) A selective force favoring increased G+C content in bacterial genes. Proc Natl Acad Sci 109:14504–14507

Retchless AC, Lawrence JG (2010) Phylogenetic incongruence arising from fragmented speciation in enteric bacteria. Proc Natl Acad Sci USA 107:11453–11458

Riehle MM, Bennett AF, Long AD (2001) Genetic architecture of thermal adaptation in *Escherichia coli*. Proc Natl Acad Sci USA 98:525–530

Rokas A, Carroll SB (2008) Frequent and widespread parallel evolution of protein sequences. Mol Biol Evol 25:1943–1953

Sabeti PC, Reich DE, Higgins JM et al (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837

Schubert S, Darlu P, Clermont O et al (2009) Role of intraspecies recombination in the spread of pathogenicity islands within the *Escherichia coli* species. PLoS Pathog 5:e1000257

Shapiro JA, Huang W, Zhang C et al (2007) Adaptive genic evolution in the Drosophila genomes. Proc Natl Acad Sci USA 104:2271–2276

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 et al (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 et al (2013) Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. Proc Natl Acad Sci USA 110: 11923–11927

Simmons SL, DiBartolo G, Denef VJ et al (2008) Population genomic analysis of strain variation in *Leptospirillum* group II bacteria involved in acid mine drainage formation. Plos Biol 6:1427–1442

Smillie CS, Smith MB, Friedman J et al (2011) Ecology drives a global network of gene exchange connecting the human microbiome. Nature 480: 241–244

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

Szabó G, Preheim SP, Kauffman KM et al (2013) Reproducibility of Vibrionaceae population structure in coastal bacterioplankton. ISME J 7:509–519

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Tenaillon O, Rodriguez-Verdugo A, Gaut RL et al (2012) The molecular diversity of adaptive convergence. Science 335:457–461

Touchon M, Hoede C, Tenaillon O et al (2009) Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet 5:e1000344

van Gremberghe I, Leliaert F, Mergeay J et al (2011) Lack of phylogeographic structure in the freshwater cyanobacterium *Microcystis aeruginosa* suggests global dispersal. PLoS ONE 6:e19561

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. Plos Biol 4:e72

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

Whitaker RJ, Banfield JF (2006) Population genomics in natural microbial communities. Trends Ecol Evol 21:508–516

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

Yang Z (2008) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15(5):568–573

Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19: 908–917