

# **Title: How clonal are bacteria over time?**

**Author: B. Jesse Shapiro<sup>1</sup>**

## **Affiliation**

1. Département de sciences biologiques, Université de Montréal, Montréal, QC H3C 3J7,  
Canada

\* [jesse.shapiro@umontreal.ca](mailto:jesse.shapiro@umontreal.ca)

## **Abstract**

Bacteria and archaea reproduce clonally (vertical descent), but exchange genes by recombination (horizontal transfer). Recombination allows adaptive mutations or genes to spread rapidly within (or even between) species, and reduces the burden of deleterious mutations. Clonality – defined here as the balance between vertical and horizontal inheritance – is therefore a key microbial trait, determining how quickly a population can adapt and the size of its gene pool. Here, I discuss whether clonality varies over time and if it can be considered a stable trait of a given population. I show that, in some cases, clonality is clearly not static. For example, non-clonal (highly recombining) populations can give rise to clonal expansions, often of pathogens. However, an analysis of time-course metagenomic data from a lake suggests that a bacterial population's past clonality (as measured by its genetic diversity) is a good predictor of its future clonality. Clonality therefore appears to be relatively – but not completely – stable over evolutionary time.

## **Introduction**

Here, I revisit the question posed in the title of a classic paper by John Maynard Smith and colleagues [1]: How clonal are bacteria, and more specifically how does clonality vary among different microbial populations and over time? First, what do we mean by clonality? Perfectly clonal bacteria replicate by cell division (vertical descent) and evolve by random mutations that occur during DNA replication. In this theoretical population, there is negligible horizontal transfer of DNA by recombination across the resulting tree of vertical descent. Very few (if any) natural bacterial populations fit this idealized, theoretical definition of clonality. Or, as discussed below, they might only fit it for a short amount of time. However, knowing where a bacterial

population of interest happens to fall along a spectrum of clonality can help us understand its biology, and even make predictions about its evolution.

The opposite of clonality is panmixis – a situation in which the rate of horizontal transfer is higher than the rate of vertical cell division, resulting in random association (linkage equilibrium) among loci in the genome [1,2]. However, rates of horizontal transfer (recombination) vary widely across the genome, such that a population can be mostly clonal, except for a few loci in the genome [3]. These loci came to be termed genomic islands – a metaphor I will build upon below. Some of the first islands identified were called pathogenicity islands because they contained virulence factors [4]. However, non-pathogenic environmental bacteria also contain islands, conferring adaptation to different ecological niches. For example, genes in *Prochlorococcus* genomic islands confer adaptation to light and nutrient conditions [5,6]. But islands need not confer niche adaptation to their host genome; they can be neutral to host fitness or even detrimental, selfish parasites. Here, I define genomic islands broadly as any piece of DNA that is transferred horizontally (by either homologous or nonhomologous recombination) from cell to cell and therefore evolves independently (*i.e.* is unlinked) from the rest of the genome.

I will begin by extending the use of island analogies to include continents, peninsulas and archipelagos (**Table 1**). I will then use these analogies to discuss to what extent microbial populations are clonal or panmictic, and how often they transition between the two regimes.

### **Are some islands really peninsulas?**

In the classic analogy, an island is totally disconnected from the mainland, meaning that genes in the island evolve independently of the genome (**Table 1**). Examples of islands that fit this strict independence might include integrated phages and other “selfish” elements, or genes that reside in a particular niche but not in a particular genome (*e.g.* a gene ecology model [7]). Peninsulas provide an analogy that might better describe how islands are related to microbial genomes. A peninsula (or “presque-île,” from the French for “almost island”) is a geographic term for a very narrow strip of land connected to (but distinct from) the mainland. In my analogy, an island is evolutionarily independent of the mainland genome, but their fates may become linked, forming a peninsula. For example, a bacterium may acquire a gene from a vast microbial gene pool. This gene allows the bacterium to invade a new ecological niche, triggering a clonal expansion in which the fate of the gene and its new host genome are linked, at least for the duration of the clonal expansion. One such example could be *Yersinia pestis*, which acquired a single gene allowing flea-borne transmission and triggering a clonal expansion in the form of Plague

pandemics [8]. Another peninsula, the prophage-encoded cholera toxin, and its links to the mainland *Vibrio cholerae* genome [9,10], is discussed below.

## **Are some genomes archipelagos?**

The very concept of one or a few islands implies a contrast with the large, clonal genomic mainland or continent. But some microbial genomes may contain so many islands that there is no mainland, only a vast archipelago (**Table 1**). A striking recent example is a population of hotspring cyanobacteria in which virtually every gene in the genome evolved independently due to frequent recombination [11], leading the authors to call the population “quasi-sexual” (in other words, panmictic). Frequent recombination was confirmed by another group, using different methods to study the same cyanobacteria [12]. This group also found that despite a history of panmixis over long time scales, populations are more clonal over shorter time scales. Similarly, the Asian ocean population of *Vibrio parahaemolyticus* also forms a panmictic gene pool, with each recombination block of ~1.8 kbp evolving independently [13]. However, the panmictic gene pool occasionally gives rise to pandemic clones. In another example, we found that almost every gene in a population of *Vibrio cyclitrophicus* genomes showed signs of recombination over relatively recent time scales [14]. Such apparently high rates of recombination in natural populations were mysterious at first, contradicting recombination rates measured in the lab [15,16] and predicted by theory [7,17]. However, theoretical models (discussed below) suggest mechanisms capable of explaining how genes can spread through populations more rapidly by recombination than by clonal expansion [18-20].

## **Clonal expansions from panmictic pools**

I propose that archipelagos are not necessarily static over time, and that archipelagos can sometimes coalesce into continents. Given the right ecological opportunity, a genome from a panmictic gene pool can escape the “gravitational pull” of recombination and take off into a clonal expansion. An example mentioned earlier is *V. cholerae*, a genetically diverse group of coastal marine bacteria, some of which cause cholera. Virulence is mainly determined by two loci in the genome: the cholera toxin and the toxin-coregulated pilus. Both genes are frequently gained and lost by recombination [21,22], but are always found in one lineage of *V. cholerae* – the lineage causing severe disease with pandemic potential, known as the phylocore genome (PG) group [10]. It remains a mystery why the PG lineage evolved once, and only once. If PG *V. cholerae* really did evolve just once, this would be surprising because *V. cholerae* draws on a diverse, global gene pool and can be considered panmictic [23]. Therefore multiple different

lineages would be expected to acquire the two (or perhaps a handful of) genetic elements required for pandemic disease. This leads to the hypothesis that pandemic cholera emergence is *selection limited* rather than *diversity limited*. In other words, benign *V. cholerae* strains constantly acquire virulence genes. However, these strains rarely encounter the right ecological niche to flourish, e.g. a human population consuming brackish water. "The right niche" has appeared a few times in human history: for example in India in the 1800s, when the Classical lineage evolved, and again in Indonesia in the 1950s, when the El Tor lineage evolved [24]. When the right conditions appear, the PG lineage, along with its virulence factors, takes off in a clonal expansions which continue to wreak havoc today (e.g. cholera pandemics from the 1800s to today, all caused by the PG clonal group). The virulence factors, previously islands in an archipelago, became a peninsula connected to the PG mainland. The linkage between virulence factors and PG remains imperfect because different variants of the cholera toxin continue to flow in and out of the PG continent [10,21]; hence the toxin remains a peninsula, not firmly part of the mainland.

*V. cholerae* is a particularly well-characterized example of a panmictic gene pool giving rise to a clonal expansion, but similar evolutionary dynamics are seen in other pathogens as well (e.g. *V. parahaemolyticus* described above [13]). Enterotoxigenic *Escherichia coli* (ETEC) seems to behave similarly, with deep branches of the phylogeny obscured by frequent recombination and plasmid exchange, but more recent branches experiencing mostly clonal descent, with tight linkage between virulence factors and the genomic mainland [25]. These observations are consistent with an ancient, panmictic gene pool giving rise to clonal expansions, which can last for decades or centuries.

## **The balance between recombination and selection**

Let us consider the evolutionary forces that determine clonality: natural selection and recombination. The effect of recombination on clonality is straightforward: more recombination means less clonality. The effect of natural selection is more complex, but is defined here simply as a force which favors clonal expansions of adaptive mutants within an ecological niche. Selection, as defined here, therefore includes ecological effects. When driven by ecological selection, clonal expansions are called selective sweeps, in which one clone outcompetes all others, purging genetic diversity in the population.

Recombination and selection interact to determine the clonality of a population. Recombination rates depend both on the ability of DNA to enter a cell and be incorporated into the genome (the baseline rate) and the ability of that DNA to be retained by a balance of genetic drift and natural selection (the realized rate). Some bacteria, such as *Helicobacter pylori*, have

realized recombination rates that are much higher than point mutation rates, exchanging at least 10% of their genome within a single four-year human infection [26]. Others, such as *Staphylococcus aureus* [27,28] and *Mycobacterium tuberculosis* are decidedly more clonal [29-31]. Recombination rates (both realized and baseline) vary widely across the genome. Of 10 pathogenic bacterial species studied, all had identifiable recombination 'hot' regions, although their length, genomic location and gene content varied [3]. Genes of different functions had different realized recombination rates, implying a role for natural selection on gene function in determining whether newly acquired genes are retained.

### Modeling the recombination-selection balance

When rates of recombination are relatively low compared to rates of natural selection on adaptive genes within niches, entire genomes will sweep to fixation before they can be shuffled by recombination. Following previous modelling work,  $s$  is defined here as the selective coefficient of a niche-adaptive allele, and  $r$  is the recombination rate, per locus per generation [7,32,33]. The  $s \gg r$  regime is well described in the Stable Ecotype Model [17], which predicts that most of the genome will follow a single, clonal phylogeny. Genome-wide sweeps thus increase clonality and can be considered a hallmark of clonal populations (**Table 1, Figure 1A**). In the  $r \gg s$  regime, individual genes (rather than entire genomes) will sweep to fixation (*i.e.* reach 100% frequency) in ecological niches to which they are adapted, without affecting genetic diversity elsewhere in the genome (**Figure 1B**). The first serious theoretical attempt to reconcile the observations of gene-specific sweeps with low recombination rates was the "Adapt Globally, Act Locally" model [18,20], in which globally adaptive genes (adaptive in multiple different niches) trigger genome-wide sweeps within a niche before being transferred to the next niche. When viewed across niches, the result is a gene-specific sweep. A recent model suggested another mechanism by which these gene sweeps can occur at moderate (not unrealistically high) rates of recombination [20]. In this model, a microbial habitat is bombarded with genetically maladapted migrants, allowing gene sweeps to occur, although the adaptive allele never reaches 100% frequency due to the constant input of migrants. In another model, Takeuchi et al. [19] show that gene sweeps can occur when  $r$  is either very high or – counter-intuitively – when  $r$  is very low, but only when negative frequency-dependent selection (NFDS) reduces the rate of genome-wide selective sweeps. NFDS might be commonly imposed on bacteria and archaea by viral (phage) predation, providing a selective advantage to rare alleles of phage receptor genes [34].

One parameter not thoroughly explored in any of these models is the effective population size ( $N_e$ ). Populations with small  $N_e$  are dominated by drift, and natural selection is inefficient.

They may experience genome-wide sweeps independently of  $r$  and  $s$ . As discussed below,  $N_e$  is probably an important determinant of clonality.

### **Genome-wide and gene-specific sweeps in nature**

To date, empirical evidence for gene-specific and genome-wide sweeps has come mostly from cross-sectional studies of a single population of genomes at a single point in time, with recombination and selection inferred backward in time [11,13,14,35]. Sequencing microbial genomes or metagenomes sampled over time – already a typical practice in genomic epidemiology (e.g. [28,36]) – promises to elucidate the rates of gene-specific and genome-wide sweeps in nature (**Figure 1**).

In a pioneering study, Bendall et al. [37] sampled a lake over nine years and followed single-nucleotide polymorphism (SNP) and gene frequencies in 30 bacterial populations by metagenomic sequencing. They inferred that one of the populations (*Chlorobium*-111) had undergone a near-complete genome-wide sweep over the nine-year study, with most SNP diversity purged genome-wide (**Figure 1A**). By identifying regions of the genome with unexpectedly low diversity compared to the genome-wide average, they inferred that gene-specific sweeps had occurred in six of the populations, but these sweeps occurred before the beginning of the nine-year study. During the nine-year study, they observed examples “where a few adjacent SNPs trended toward fixation while genome-wide diversity was maintained” (**Figure 1B**). They took this observation as consistent with gene-specific selective sweeps, but did not attempt to determine whether the sweeps were due to selection or drift. Similarly, the inferred genome-wide sweep could have been caused by selection or drift.

Is the frequency of genome-wide sweeps controlled mainly by the balance of recombination and selection, or could drift (controlled by the effective population size,  $N_e$ ) play an important role as well? In their Figure 2 legend, Bendall et al. observe that “populations with many SNPs were not necessarily sequenced deeper than those with few SNPs.” This statement was simply meant to assure the reader that SNP calling was not biased by sequence coverage, but it also suggests that microbial populations tend to be far from a standard neutral model with constant population size (*i.e.* the Wright-Fisher coalescent [38]). Taking SNP density as a rough measure of  $N_e$  and sequence coverage as a rough measure of the census population size, the data show that census population size is a poor predictor of  $N_e$ , and thus that populations are likely impacted by population bottlenecks and/or selection – although it is difficult to distinguish between them.

As a whole, the study showed that both genome-wide and gene-specific sweeps can occur in different microbial populations from the same environment. Whether microbial populations behaved differently due to differences in their ecology (*i.e.* regime of natural selection) or in their baseline recombination rates remains a question for future study; Cohan recently suggested that ecological differences could be more important [39]. Specifically, he suggests that ecological "generalists" could have more opportunities for diversification, and thus be relatively immune to genome-wide sweeps, compared to ecological specialists. I suggest that these generalists may simply have larger  $N_e$ , such that diversity is rarely purged by drift (*e.g.* bottlenecks), and that diversity is mainly governed by selection and recombination. It is possible that ecological generalists tend to have large  $N_e$ , or that  $N_e$  and ecology exert independent effects (*e.g.* ecological generalists can have low  $N_e$  and still resist genome-wide sweeps). The fact that one genome-wide sweep was observed over a nine year period suggests that such events might be relatively rapid but rare (only observed in one of 30 populations). Meanwhile, gene-sweeps might be more common historically (affecting six of 30 populations), but could take longer to proceed to completion.

### **Is clonality a stable trait?**

As described in the *V. cholerae* example, some pathogenic bacterial populations can switch between panmictic and clonal lifestyles [11,13,14,19,25,34,35]. Therefore clonality can vary over time, but how much and how often? To quantify the stability of clonality over time, not just in pathogens but in free-living environmental bacteria, I re-analyzed the lake time-course of Bendall et al. [37]. Because estimates of selection and recombination rates were not readily available for this dataset, I defined clonality based on the total genetic diversity (SNP density) in a population, which scales with  $N_e$ , and includes the effects of both drift and of genome-wide selective sweeps. Frequent genome-wide selective sweeps ( $s \gg r$ ) and/or small  $N_e$  result in clonality. I identified 20 "old, diverse" populations as those with a high density of SNPs (>1500 SNPs/Mbp) at the beginning of the time-course. These populations are likely "old" because they have gone a relatively long time since the last genome-wide purge of genetic diversity and are relatively non-clonal. They include the six populations inferred to have undergone gene-specific sweeps [37]. The remaining ten populations were defined as "young, low-diversity," having more recently experienced a genome-wide purge of diversity. The "old, diverse" populations have a relatively low ratio of nonsynonymous (N) to synonymous (S) SNPs, suggesting large  $N_e$  and ample time for purifying selection to remove (mostly deleterious) nonsynonymous mutations (**Figure 2**). In contrast, the "young, low-diversity" populations are more likely to have high N:S ratios,



suggesting smaller  $N_e$  and less time for purifying selection to act. The cutoff of 1500 SNPs/Mbp (0.15% SNP density) is somewhat arbitrary, but seems to correspond to an inflection point in **Figure 2**, and also falls squarely within the 0-0.30% SNP density, above which *E. coli* evolution appears to transition from clonal to recombining [40].

With young (presumably clonal, or low  $N_e$ ) and old (less clonal, higher  $N_e$ ) populations thus defined, I asked whether “old, diverse” populations tended to maintain their diversity through the 9-year period of the study. In their paper, Bendall et al. defined two alternative population types: 1) those that maintained stable SNP diversity over nine years, and 2) those that experienced significant fluctuations in diversity due to clonal expansions – defined when one, but not all timepoints are dominated by a single allele ( $\geq 95\%$  frequency) at  $>40\%$  of SNP sites in the genome [37]. By definition, the 19 populations of the first type did not experience genome-wide sweeps during the study, while the 11 populations of the second type did experience genome-wide purges of diversity, which were transient in ten cases and apparently permanent in one case (*Chlorobium*-111). Strikingly, 17 out of 20 “old, diverse” populations maintained their diversity over the nine-year study, compared to only two out of ten “young, low-diversity” populations (Fisher test, Odds Ratio = 19.4,  $P < 0.001$ ). This result suggests that populations with a history of genome-wide sweeps tend to experience subsequent genome-wide sweeps, and those that have maintained genetic diversity in the past tend to maintain their diversity into the future. In other words, clonality can be considered a relatively stable microbial trait. However, we must take care in taking a genome-wide sweep as evidence for clonality ( $s \gg r$ ). In small effective population sizes, genome-wide sweeps can occur due to drift and bottlenecks, independently of  $s$  and  $r$ . Therefore, many of “young, low diversity” populations (**Figure 2**) may have lost diversity over nine years due to low  $N_e$ , not due to low  $r$ .

### History repeats itself

It appears that pathogens are more likely than free-living bacteria to undergo clonal expansions, due in part to their ecology and transmission dynamics [41]. Free-living aquatic bacteria, on the other hand, seem to be more likely to live in large, panmictic populations and behave like archipelagos [11,13,14,37]. If clonality is indeed a stable trait, this implies that history will repeat itself, and that the future behavior of microbial populations can be predicted with some confidence from their past behavior. Diverse populations tend to stay diverse. Clonal populations (that experience frequent genome-wide sweeps) tend to stay clonal. But history is not doomed to repeat itself forever. As we have seen, clonal expansions, such as pandemic *V. cholerae*, may originate when a panmictic gene pool (an archipelago) coalesces into a clonal continent, with



271 virulence factors linked as peninsulas. Many such pathogenic clones have been documented, with  
 272 life spans of decades to thousands of years [25,28,42,43]. Other pathogens, such as *Streptococcus*  
 273 *pneumoniae*, may retain their panmictic population structure throughout an outbreak [44-46].  
 274 Why some pathogens are clonal and others are panmictic is an open question, but surely depends  
 275 on the balance between recombination and selection, on the effective population size, and on the  
 276 time scales considered.

277

278 **Acknowledgments.** I am grateful to Rex Malmstrom, Yan Boucher, and Salvador Almagro-  
 279 Moreno for their thoughtful comments which improved the manuscript. I was supported by a  
 280 Canada Research Chair.

**Table 1. Extended island metaphors of microbial genome evolution.**

Geographic metaphor	Genetic unit to which the metaphor applies	Type of selective sweep experience by the unit	Dominant mode of genetic transmission	Example
Island	Gene	Gene-specific	horizontal	genes in the <i>V. cholerae</i> integron [22,23]
Peninsula	Gene	Genome-wide	vertical (clonal)	the cholera toxin gene, acquired horizontally, then linked to a clonal <i>V. cholerae</i> genome [9,21]
Continent	Genome	Genome-wide	vertical (clonal)	clonal expansions of <i>S. aureus</i> [28], <i>M. tuberculosis</i> [31,43]
Archipelago	Genome	Gene-specific	horizontal	hotspring cyanobacteria [11], ocean vibrios [13,14], pneumococcus [44,46]

## Figures

**Figure 1. Temporal dynamics of genome-wide and gene-specific selective sweeps inferred from metagenomic data.** Genetic diversity can be measured by mapping metagenomic reads to a reference genome, identifying SNPs, and calculating the allele frequencies at each SNP position in the genome over time. The lowest possible genetic diversity occurs when a single allele is present in 100% of metagenomic reads. Alternatively, diversity could be defined in terms of gene presence/absence, based on relative coverage of a gene in the reference genome by metagenomic reads. **(a)** In a hypothetical genome-wide selective sweep, all positions in the genome tend toward low diversity. **(b)** In a hypothetical gene-specific selective sweep, only one or a few positions in the genome tend toward low diversity, while the rest of the genome maintains high or intermediate diversity.

**Figure 2. Past diversity predicts future diversity.** Based on data from Table 2 of Bendall et al. [37], the nonsynonymous to synonymous (N:S) SNP ratio was plotted against the total SNP density (SNPs per megabasepair) for each of 30 bacterial populations. A pseudocount of 1 was added to both N and S counts. These 30 populations were divided into 20 “old, diverse” populations (>1500 SNPs/Mbp) and 10 “young, low-diversity” populations (<1500 SNPs/Mbp). Each point represents one of the 30 populations, colored in black if diversity was maintained over a 9-year metagenomic time-course, or in red if it was not. Seventeen out of 20 “old, diverse” populations maintained their diversity over the 9-year study, compared to only 2 out of 10 “young, low-diversity” populations (Fisher test, Odds Ratio = 19.4,  $P < 0.001$ ). The same result is obtained drawing a cutoff based on N:S rather than SNPs/Mbp: populations with  $N:S < 0.5$  tend to maintain their diversity, whereas those with  $N:S > 0.5$  tend to be purged of diversity over the 9 years (Fisher test, Odds Ratio = 5.94,  $P = 0.042$ ). Consistent with previous observations that N:S depends on the evolutionary time available for purifying selection to act [47,48], N:S is negatively correlated with SNPs/Mbp, a proxy for evolutionary time or the time since the last genome-wide purge of genetic diversity in this dataset (Pearson’s correlation of  $\log_{10}$  transformed data,  $r = -0.81$ ,  $P = 5.6e-8$ ).

## References

1. Smith JM, Smith NH, O'Rourke M, Spratt BG: **How clonal are bacteria?** *Proc Natl Acad Sci USA*. 1993, 90: 4384–4388.
2. Guttman DS, Dykhuizen DE: **Clonal Divergence in Escherichia coli as a Result of Recombination, Not Mutation.** *Science*. 1994, 266: 1380–1383.
3. Yahara K, Didelot X, Jolley KA, Kobayashi I, Maiden MCJ, Sheppard SK, et al.: **The landscape of realized homologous recombination in pathogenic bacteria.** *Mol Biol Evol*. 2016, 33: 456–471. doi:10.1093/molbev/msv237
4. Hacker J, Blum-Oehler G, Mühldorfer I, Tschäpe H: **Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution.** *Mol Microbiol*. 1997, 23: 1089–1097.
5. Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, Delong EF, et al.: **Genomic islands and the ecology and evolution of Prochlorococcus.** *Science*. 2006, 311: 1768–1770.
6. Coleman ML, Chisholm SW: **Ecosystem-specific selection pressures revealed through comparative population genomics.** *Proc Natl Acad Sci USA*. 2010, 107: 18634–18639. doi:10.1073/pnas.1009480107
7. Shapiro BJ, Polz MF: **Ordering microbial diversity into ecologically and genetically cohesive units.** *Trends Microbiol*. 2014, 22: 235–247. doi:10.1016/j.tim.2014.02.006
- \* A comprehensive discussion of the roles of selection and recombination in structuring microbial diversity.
8. Rasmussen S, Allentoft ME, Nielsen K, Orlando L, Sikora M, Sjögren K-G, et al.: **Early Divergent Strains of Yersinia pestis in Eurasia 5,000 Years Ago.** *Cell*. 2015, 163: 571–582. doi:10.1016/j.cell.2015.10.009
9. Waldor MK, Mekalanos JJ: **Lysogenic conversion by a filamentous phage encoding cholera toxin.** *Science*. 1996, 272: 1910–1914.
10. Chun J, Grim CJ, Hasan NA, Lee JH, Choi SY, Haley BJ, et al.: **Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic Vibrio cholerae.** *Proc Natl Acad Sci USA*. 2009, 106: 15442–15447. doi:10.1073/pnas.0907787106
11. Rosen MJ, Davison M, Bhaya D, Fisher DS: **Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche.** *Science*. 2015, 348: 1019–1023. doi:10.1126/science.aaa4456

\* Deep sequencing of 90 cyanobacterial marker genes reveals extensive recombination, with each genome composed of a random mix of alleles from the gene pool.

- 353 12. Melendrez MC, Becraft ED, Wood JM, Olsen MT, Bryant DA, Heidelberg JF, et al.:  
354 **Recombination Does Not Hinder Formation or Detection of Ecological Species of**  
355 ***Synechococcus* Inhabiting a Hot Spring Cyanobacterial Mat.** *Front Microbiol.* 2016,  
356 6: 544. doi:10.1128/AEM.72.1.723-732.2006
- 357 13. Cui Y, Yang X, Didelot X, Guo C, Li D, Yan Y, et al.: **Epidemic Clones, Oceanic Gene**  
358 **Pools, and Eco-LD in the Free Living Marine Pathogen *Vibrio parahaemolyticus*.** *Mol*  
359 *Biol Evol.* 2015, 32: 1396–1410. doi:10.1093/molbev/msv009
- 360 \* In-depth population genomic evidence of a panmictic ocean gene pool containing non-clonal  
361 ecological populations.  
362
- 363 14. Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabo G, et al.:  
364 **Population Genomics of Early Events in the Ecological Differentiation of Bacteria.**  
365 *Science.* 2012, 336: 48–51. doi:10.1126/science.1218198
- 366 15. Vulić M, Dionisio F, Taddei F, Radman M: **Molecular keys to speciation: DNA**  
367 **polymorphism and the control of genetic exchange in enterobacteria.** *Proc Natl Acad*  
368 *Sci USA.* 1997, 94: 9763–9767.
- 369 16. Majewski J, Cohan FM: **DNA sequence similarity requirements for interspecific**  
370 **recombination in *Bacillus*.** *Genetics.* 1999, 153: 1525–1533.
- 371 17. Wiedenbeck J, Cohan FM: **Origins of bacterial diversity through horizontal genetic**  
372 **transfer and adaptation to new ecological niches.** *FEMS Microbiology Reviews.* 2011,  
373 35: 957–976. doi:10.1111/j.1574-6976.2011.00292.x
- 374 18. Majewski J, Cohan FM: **Adapt globally, act locally: the effect of selective sweeps on**  
375 **bacterial sequence diversity.** *Genetics.* 1999, 152: 1459–1474.
- 376 19. Takeuchi N, Cordero OX, Koonin EV, Kaneko K: **Gene-specific selective sweeps in**  
377 **bacteria and archaea caused by negative frequency-dependent selection.** *BMC*  
378 *Biology.* 2015, 13: 20. doi:10.1186/s12915-015-0131-7
- 379 \* Mathematical modeling shows how individual genes can sweep through populations with low  
380 recombination rates in the presence of negative frequency-dependent selection.  
381
- 382 20. Niehus R, Mitri S, Fletcher AG, Foster KR: **Migration and horizontal gene transfer**  
383 **divide microbial genomes into multiple niches.** *Nature Communications.* 2015, 6: 8924.  
384 doi:10.1038/ncomms9924
- 385 21. Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, et al.: **Evidence for**  
386 **several waves of global transmission in the seventh cholera pandemic.** *Nature.* 2011,  
387 477: 462–U111. doi:10.1038/nature10392
- 388 22. Orata FD, Kirchberger PC, Méheust R, Barlow EJ, Tarr CL, Boucher Y: **The Dynamics**  
389 **of Genetic Interactions between *Vibrio metoecus* and *Vibrio cholerae*, Two Close**  
390 **Relatives Co-Occurring in the Environment.** *Genome Biology and Evolution.* 2015, 7:  
391 2941–2954. doi:10.1093/gbe/evv193/-/DC1

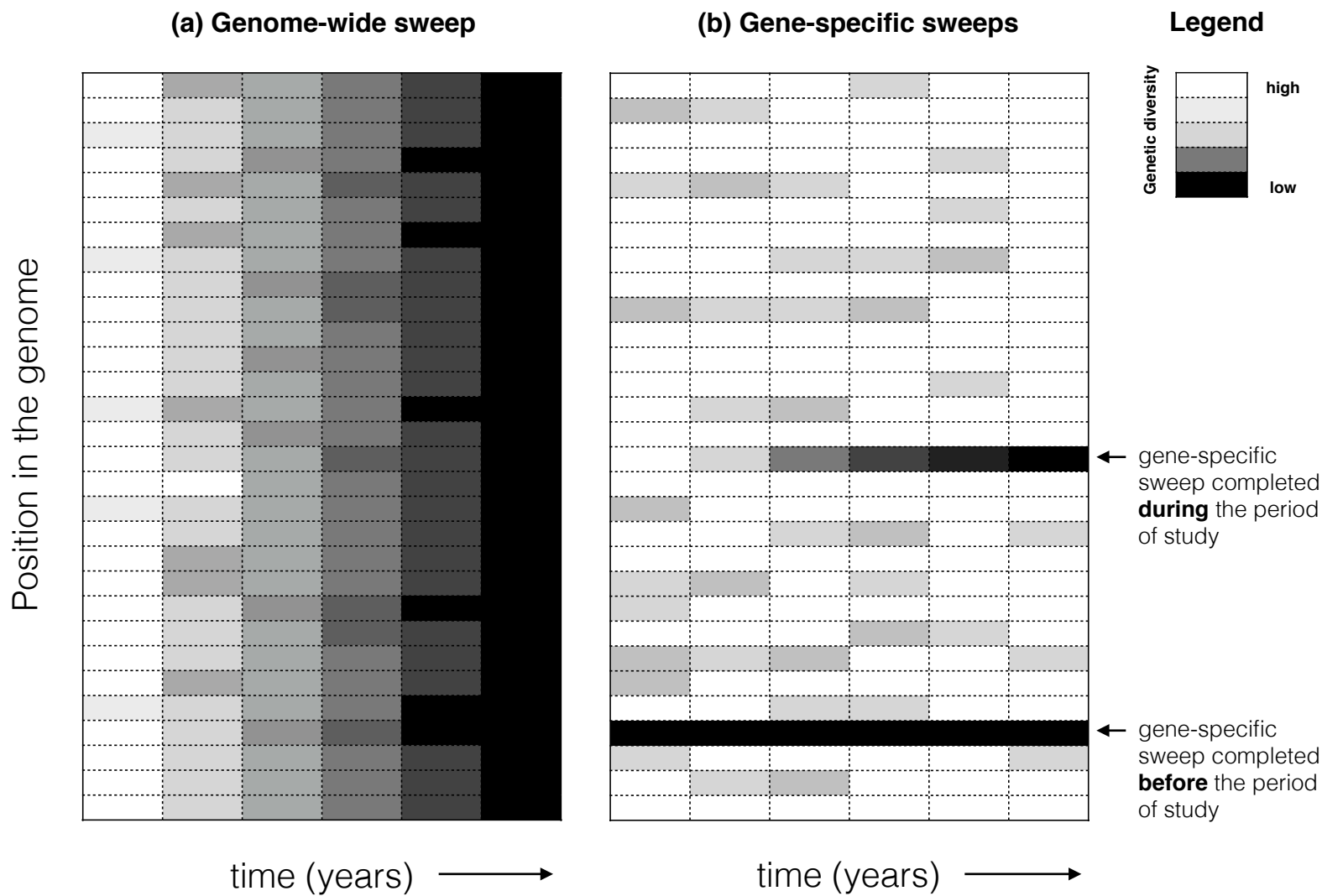
- 392 \* A clear demonstration that genes in the integron are recombined between species just as often  
393 as within species, and that 24% of ‘core’ genes have crossed species boundaries.  
394
- 395 23. Boucher Y, Cordero OX, Takemura A, Hunt DE, Schliep K, Baptiste E, et al.: **Local**  
396 **Mobile Gene Pools Rapidly Cross Species Boundaries To Create Endemicity within**  
397 **Global *Vibrio cholerae* Populations.** *mBio*. 2011, 2: e00335–10.  
398 doi:10.1128/mBio.00335-10
- 399 24. Boucher Y, Orata FD, Alam M: **The out-of-the-delta hypothesis: dense human**  
400 **populations in low-lying river deltas served as agents for the evolution of a deadly**  
401 **pathogen.** *Front Microbiol*. 2015, 6: 1120. doi:10.3389/fmicb.2015.01120
- 402 25. Mentzer von A, Connor TR, Wieler LH, Semmler T, Iguchi A, Thomson NR, et al.:  
403 **Identification of enterotoxigenic *Escherichia coli* (ETEC) clades with long-term**  
404 **global distribution.** *Nature Genetics*. 2014, 46: 1321–1326. doi:10.1038/ng.3145
- 405 26. Cao Q, Didelot X, Wu Z, Li Z, He L, Li Y, et al.: **Progressive genomic convergence of**  
406 **two *Helicobacter pylori* strains during mixed infection of a patient with chronic**  
407 **gastritis.** *Gut*. 2015, 64: 554–561. doi:10.1136/gutjnl-2014-307345
- 408 27. Vos M, Didelot X: **A comparison of homologous recombination rates in bacteria and**  
409 **archaea.** *The ISME journal*. 2009, 3: 199–208. doi:10.1038/ismej.2008.93
- 410 28. Baines SL, Holt KE, Schultz MB, Seemann T, Howden BO, Jensen SO, et al.:  
411 **Convergent adaptation in the dominant global hospital clone ST239 of methicillin-**  
412 **resistant *Staphylococcus aureus*.** *mBio*. 2015, 6: e00080. doi:10.1128/mBio.00080-15
- 413 29. Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, et al.: **High**  
414 **Functional Diversity in *Mycobacterium tuberculosis* Driven by Genetic Drift and**  
415 **Human Demography.** *PLOS Biol*. 2008, 6: e311.  
416 doi:10.1371/journal.pbio.0060311.st006
- 417 30. Namouchi A, Didelot X, Schock U, Gicquel B, Rocha EPC: **After the bottleneck:**  
418 **Genome-wide diversification of the *Mycobacterium tuberculosis* complex by mutation,**  
419 **recombination, and natural selection.** *Genome Research*. 2012, 22: 721–734.  
420 doi:10.1101/gr.129544.111
- 421 31. Casali N, Nikolayevskyy V, Balabanova Y, Harris SR, Ignatyeva O, Kontsevaya I, et al.:  
422 **Evolution and transmission of drug-resistant tuberculosis in a Russian population.**  
423 *Nature*. 2014, 46: 279–286. doi:10.1038/ng.2878
- 424 32. Shapiro BJ, David LA, Friedman J, Alm EJ: **Looking for Darwin's footprints in the**  
425 **microbial world.** *Trends in Microbiology*. 2009, 17: 196–204.  
426 doi:10.1016/j.tim.2009.02.002
- 427 33. Friedman J, Alm EJ, Shapiro BJ: **Sympatric Speciation: When Is It Possible in**  
428 **Bacteria?** *PLOS ONE*. 2013, 8: e53539. doi:10.1371/journal.pone.0053539.g004
- 429 34. Cordero OX, Polz MF: **Explaining microbial genomic diversity in light of evolutionary**  
430 **ecology.** *Nature Reviews Microbiology*. 2014, 12: 263–273. doi:10.1038/nrmicro3218

- 431 35. Cadillo-Quiroz H, Didelot X, Held NL, Herrera A, Darling A, Reno ML, et al.: **Patterns**  
432 **of Gene Flow Define Species of Thermophilic Archaea.** *PLOS Biol.* 2012, 10:  
433 e1001265. doi:10.1371/journal.pbio.1001265.t001
- 434 36. Croucher NJ, Didelot X: **The application of genomics to tracing bacterial pathogen**  
435 **transmission.** *Current Opinion in Microbiology.* 2015, 23: 62–67.  
436 doi:10.1016/j.mib.2014.11.004
- 437 37. Bendall ML, Stevens SL, Chan L-K, Malfatti S, Schwientek P, Tremblay J, et al.:  
438 **Genome-wide selective sweeps and gene-specific sweeps in natural bacterial**  
439 **populations.** *The ISME journal.* 2016, advance online publication, 8 January 2016;  
440 doi:10.1038/ismej.2015.241
- 441 \*\* A pioneering application of time-course metagenomics to infer genome-wide and gene-  
442 specific selective sweeps in natural lake bacterial populations.  
443
- 444 38. Kingman J: **On the genealogy of large populations.** *J Appl Prob.* 1982, 19: 27–43.
- 445 39. Cohan FM: **Bacterial Speciation: Genetic Sweeps in Bacterial Species.** *Curr Biol.* 2016,  
446 26: R112–R115. doi:10.1016/j.cub.2015.10.022
- 447 40. Dixit PD, Pang TY, Studier FW, Maslov S: **Recombinant transfer in the basic genome**  
448 **of *Escherichia coli*.** *Proc Natl Acad Sci USA.* 2015, 112: 9070–9075.  
449 doi:10.1073/pnas.1510839112
- 450 41. Achtman M, Wagner M: **Microbial diversity and the genetic nature of microbial**  
451 **species.** *Nature Reviews Microbiology.* 2008, 6: 431–440. doi:10.1038/nrmicro1872
- 452 42. Snitkin ES, Zelazny AM, Thomas PJ, Stock F, NISC Comparative Sequencing Program,  
453 Henderson DK, et al.: **Tracking a Hospital Outbreak of Carbapenem-Resistant**  
454 ***Klebsiella pneumoniae* with Whole-Genome Sequencing.** *Science Translational*  
455 *Medicine.* 2012, 4: 148ra116–148ra116. doi:10.1126/scitranslmed.3004129
- 456 43. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, et al.: **Out-of-Africa**  
457 **migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern**  
458 **humans.** *Nature Genetics.* 2013, 45: 1176–1182. doi:10.1038/ng.2744
- 459 44. Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J, et al.:  
460 **Population genomics of post-vaccine changes in pneumococcal epidemiology.** *Nature*  
461 *Genetics.* 2013, 45: 656–663. doi:10.1038/ng.2625
- 462 45. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, et al.: **Dense**  
463 **genomic sampling identifies highways of pneumococcal recombination.** *Nature*  
464 *Genetics.* 2014, 46: 305–309. doi:10.1038/ng.2895
- 465 46. Marttinen P, Croucher NJ, Gutmann MU, Corander J, Hanage WP: **Recombination**  
466 **produces coherent bacterial species clusters in both core and accessory genomes.**  
467 *Microbial Genomics.* 2015, 1. doi:10.1099/mgen.0.000038.



- 468 \* The first simulation of bacterial evolution including mutation, allelic exchange (homologous  
469 recombination) and gene gain/loss (non-homologous recombination).  
470
- 471 47. Rocha EP, Smith JM, Hurst LD, Holden MT, Cooper JE, Smith NH, et al.: **Comparisons**  
472 **of dN/dS are time dependent for closely related bacterial genomes.** *J Theor Biol.* 2006,  
473 239: 226–235.
- 474 48. Kryazhimskiy S, Plotkin JB: **The Population Genetics of dN/dS.** *PLOS Genet.* 2008, 4:  
475 e1000304.
- 476

**Figure 1**



**Figure 2**

