Looking for Darwin's footprints in the microbial world

B. Jesse Shapiro[1], Lawrence A. David[1], Jonathan Friedman[1], & Eric J. Alm[1,2,3,4,5]

[1] Program in Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge, MA

[2] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA

[3] Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, MA

[4] The Virtual Institute of Microbial Stress and Survival, http://vimss.lbl.gov

[5] The Broad Institute of MIT and Harvard, Cambridge, MA

Corresponding author: Alm, E.A. (ejalm@mit.edu)

Abstract

As we observe the 200th anniversary of Charles Darwin's birthday, microbiologists interested in the application of Darwin's ideas to the microscopic world have a lot to celebrate: an emerging picture of the (mostly microbial) Tree of Life at ever-increasing resolution, an understanding of horizontal gene transfer as a driving force in the evolution of microbes, and thousands of complete genome sequences to help formulate and refine our theories. At the same time, quantitative models of the microevolutionary processes shaping microbial populations remain just out of reach, a point that is perhaps most dramatically illustrated by the lack of consensus on how (or even whether) to define bacterial species. We summarize progress and prospects in bacterial population genetics, with an emphasis on detecting the footprint of positive Darwinian selection in microbial genomes.

Selection as a window into the microbial world

The microbial world is largely hidden from the naked eye, making it difficult to know the selective pressures acting on a bacterium. Nonetheless, if the genes important to fitness in a particular niche are identified, they can have a dramatic impact on our understanding of that environment. A few recent examples from marine environments help to illustrate this point. The discovery of bacteriorhodopsin in diverse marine bacteria revealed a previously unsuspected evolutionary adaptation that explained the long-standing puzzle of how such a variety of species can thrive in the nutrient poor open ocean [1, 2]. Moreover, spectral tuning of these molecules might help explain the distribution of different species across different regions and depths in the ocean [3]. In another example, phosphorous acquisition genes in *Prochlorococcus* distribute preferentially in strains living in periodically phosphorus-limited waters, suggesting an obvious link between genetics and environmental factors [4]. Many other cases of gene-specific environmental selection, however, probably involve more subtle genetic changes than the gain or loss of an entire gene or pathway, such as amino acid substitutions at specific functional sites [5].

While ongoing advances in genome sequencing technologies have made it possible to obtain complete genome sequences for entire populations of microbes, it is not clear whether genome sequences can be converted directly into evolutionary insight. One appealing and conceptually simple approach comes from the emergent field of population genomics: align the genomes of an entire population of individuals, and use the traditional tools of population genetics to pinpoint loci involved in recent Darwinian selection.

In this paper, we discuss the prospects for uncovering Darwinian selection in microbial genomes, which are becoming more readily available for a broad spectrum of medically, agriculturally and ecologically interesting organisms. We focus on genetic adaptations driven by positive selection, and the challenges involved in detecting them (Box 1). Tests for positive selection will depend on patterns of recombination, which are expected to differ between asexual microbes and sexual eukaryotes, and also among microbes depending on their lifestyles and demography. We summarize the different types of tests (Table 1), highlighting their relative merits under different regimes of recombination, and discuss how the interplay of positive selection and recombination affects the patterns of genetic variation in microbes.

Detecting natural selection in genomic sequence

Genome-wide scans for positively-selected loci in metazoans, especially humans, have yielded substantial insights into the functions of unknown genes, and the genetic basis of phenotypic differences between species. The general approach is to gather genome sequences of related species, compute a sequence-based metric to quantify positive selection on each locus, and take outliers from the genome-wide distribution as candidate positively-selected loci [6, 7]. This approach is exemplified by a recent study of six mammalian genomes, which used the *dN/dS* metric (Table 1) to reveal that genes involved in immunity and sensory perception played a key role in differentiating primates and rodents [8]. Genes acting in the same biochemical pathway were also found to undergo positive selection together, a finding we previously reported in bacteria [9]. Genome-wide scans for positive Darwinian selection have also been performed on finer scales - for example, within human populations, revealing very recent positive selection in

genes involved in malaria resistance [10], hair follicle production [11, 12], and lactose tolerance [13].

But are these approaches, which are being applied in earnest to sexual eukaryotes, theoretically justified in bacteria where recombination might not be frequent enough to provide gene-specific resolution? To investigate this question, we first review the tools available for detecting selection in different types of populations.

Tools for near-clonal populations

In bacteria, patterns of genetic variation depend on the extent to which populations behave clonally. In a perfectly clonal population, every substitution in the genome will have arisen by mutation, never by recombination. Every adaptive allele that arises will therefore be perfectly linked to every other allele in the genome. If the goal is to distinguish adaptive loci from other mutations fixed in the clonal background, one could look for loci with an excess of functional changes (e.g, *dN/dS* and other methods discussed in the 'Detecting selection among species and higher-order groups' section), and/or that show evidence for convergent evolution.

One advantage of a perfectly clonal population, from a practical standpoint, is that genomes are related by a single phylogenetic tree, rather than a more complicated network structure that represents recombination. Alleles that arise independently multiple times in different branches, and are thus incongruous with the tree, stand out as candidate examples of convergent evolution. In a recent study of genetic variation among isolates of *Salmonella enterica* serovar Typhi, convergent evolution was observed at a few loci [14]. Recombination was ruled out as the cause

of the phylogenetic incongruence and convergent mutations resulted in amino acid substitutions, two of which have known adaptive value in conferring antibiotic resistance, further supporting the hypothesis of positive selection.

Sokurenko *et al.* introduced 'zonal analysis' to identify mutations of uropathogenic *Escherichia coli* associated with recent invasion of a new niche, the human urinary tract [15]. By definition, such mutations are recently derived, and are found near the tips of a well-resolved phylogeny. They also tended to involve repeated (convergent) amino acid changes in variable 'hotspots', and these occurred in uropathogenic but not commensal strains. The authors could thus conclude that these mutations conferred a competitive advantage in the uropathogenic niche. This type of analysis - effectively a special case of convergence testing - could be extended and generalized to identify recently selected loci across the genome, even when the population structure and/or selection pressure is obscure.

Tools for sexual populations

A suite of population-genetic tests for non-neutral patterns of evolution have been developed over the past 20 years, and often used to detect positively-selected loci in humans or other sexual eukaryotes. Although many of these tests are sensitive to various deviations from neutral evolution (Table 1), they primarily detect selective sweeps. These tests can be divided into two main classes: (i) Tajima's *D* and related diversity-based tests of deviations from the neutral the allele-frequency spectrum [16, 17] or (ii) long-range haplotype (LRH) and related tests [9, 11]. Diversity-based tests identify alleles that are at unusually high frequency (suggesting a selective sweep of a single beneficial allele) or intermediate frequency (diversifying selection maintaining

multiple alleles in the population). This class of test can be applied to aligned homologous sequences of any length, assuming that sites within the sequence are completely linked (no recombination between them). Thus, diversity-based statistics should be computed within short windows of DNA across the genome. Meanwhile, haplotye-based tests model the decay of linkage disequilibrium (LD) with physical distance in the genome to identify haplotypes that are at unexpectedly high frequency for their age, indicating a recent or ongoing selective sweep.

Both classes of tests should in theory be applicable to populations of bacteria in which homologous recombination among strains is rampant. Diversity-based tests have been applied to a variety of bacterial populations, including cyanobacteria [18], *Buchnera* and related insect endosymbionts [19], *Neisseria* [20], and *Pseudomonas* [21]. In a broad study spanning seven bacterial phyla, Hughes found that Tajima's *D* tends to be lower in nonsynonymous sites than synonymous sites, implying purifying selection on slightly deleterious mutations that lead to amino acid changes [22]. Hughes also observed that this difference between sites implies that recombination must be occurring at some level in order to allow sites to evolve independently. This implies that diversity-based tests have potential to pinpoint selected loci, provided that they are separated from the clonal background by recombination.

Meanwhile, haplotype-based tests have not been applied to bacterial populations - perhaps because population sampling has not been performed at sufficient resolution to capture very recent selective sweeps. Even frequently recombining bacteria differ from sexual eukaryotes in two major ways: (i) in bacteria, recombination occurs by gene conversion rather than crossing-over and (ii) recombination is decoupled from reproduction. As a result of gene conversion,

linkage between nearby loci is expected to be higher than between distant loci in bacteria, making haplotype-based tests valid, in principle, over short genomic distances. But unlike in organisms that recombine by crossing-over, linkage is expected between all loci *not* involved in a gene conversion event regardless of their physical distance on the chromosome [23]. In other words, a non-recombinant locus might be linked to another distant locus, but unlinked to nearby loci that have undergone gene conversion. This type of pattern, called a clonal frame [24], is more likely to occur when gene conversion size fragments are small (e.g. on the order of ~500 bp in *Helicobacter pylori* [25]), when recombination is infrequent, or when only certain combinations of two distant alleles are tolerated, creating linkage between them, with free recombination in the intervening region [26]. Such epistatic interactions among alleles could thus affect patterns of LD, and in principle represent an important determinant of recombination frequency [27].

Can recombination maintain diversity in the face of selective sweeps?

Whether diversity- and haplotype-based tests are able to distinguish adaptive mutations within a population depends on the balance between the opposing forces of positive selection (purging diversity as a new allele approaches fixation) and recombination (maintaining diversity by unlinking distant regions of the genome from a selective sweep). The ratio *r/s* can be used as a shorthand to express this balance between recombination (*r*) and selection (*s*) in a population, and *r/s* is rarely, if ever, measured. Much more commonly measured is the *r/m* ratio, which assesses the relative likelihood that a polymorphic site arises by recombination versus mutation (*m*). *r/m* varies widely among bacteria (e.g. *r/m* is ~ 5-80 in *Neisseria meningitidis*, ~ 50 in *Streptococcus pneumoniae*, ~ 10-50 in *E. coli*, and ~ 1-3 in *Bacillus* [28, 29]), but is generally

larger than 1, suggesting that recombination is quite strong relative to mutation in many species. But the mutation rate is universally quite low in bacteria: on the order of 10^{-10} mutations per site per generation [30]. Even if recombination generates much more diversity than mutation, is it a sufficiently strong diversity generating force to maintain diversity in the face of selection? Selective coefficients for adaptive mutations might be quite high: on the order of 10^{-2} or higher [31]. Thus, adaptive mutations might become fixed before recombination has time to act, leading to genome-wide purges of diversity [32, 33] sometimes referred to as periodic selection. This would render diversity- and haplotype-based tests powerless to discern the selected locus against the background of uniformly low diversity.

So exactly how much recombination is needed to overcome the purging of diversity that can result from periodic selection? Figure 1a illustrates the level of diversity at a single locus in a population after an adaptive mutation at a distant gene locus becomes fixed, as a function of r and s. When recombination rates are low, the resulting diversity at the distant (non-selected) locus is effectively purged resulting in a clonality of 1, which we define as Σp_i^2, where p_i is the frequency of the i^{th} allele at the non-selected locus (also called Simpson's diversity index in ecology). When recombination rates are high, diversity at the non-selected locus is maintained either through the generation of new mutations or retention of initial diversity. Simplifying matters, the trends are only weakly dependent on the population size (Figure 1b) when expressed in the natural variables ρ (=Nr, where N is the population size and r is the per gene or locus per generation recombination rate) and σ (=Ns, where s is the relative fitness advantage conferred by the adaptive mutation).

As illustrated graphically in Figure 1, and as previously modeled by others (*e.g.*, [34, 35]), there exist regimes of recombination and selection that allow an adaptive mutation to purge diversity locally in the genome (at loci near the adaptive site), without substantially reducing diversity elsewhere in the genome. For this panmictic scenario to be viable, *r* must be quite large and/or *s* must be small. High *r/s* or $r > s$ can be used to roughly describe this scenario, but this shorthand does not do justice to the complex relationship between diversity, *r* and *s* (Figure 1a). Panmixis could result from a relatively high population recombination rate, similar to those observed in promiscuous groups such as *Neisseria* or *Helicobacter* [25, 28]. Conversely, if $r << s$, a population will be effectively clonal. This type of clonal population has been observed repeatedly in the context of long-term experimental evolution studies of *E. coli*. In these studies, adaptive point mutations are successively fixed in a clonal background, with little or no contribution of recombination [36-38].

Patterns of recombination and their interplay with selection

Bacterial lineages show substantial variation in population structure - ranging from essentially clonal (*e.g. Salmonella*) to panmictic (*e.g. Neisseria gonorrhoeae*) [39]. In many recombining lineages including *Campylobacter* [40] *Neisseria* [20], *Helicobacter* [25], and *E. coli* [41]. LD decays with distance in the genome. To illustrate this distance-dependent decay of LD, we show LD between pairs of genes located throughout the *E. coli* genome (Figure 2). Genes up to ~20 kilobases apart on the chromosome are often linked, but this linkage drops off around 100 kb. However, linkage never decays to zero in bacteria because some fraction of very distant loci will remain linked as part of the clonal frame [23]. Patterns of LD have a great impact on tests for

selection (Table I, Figure 4), and it is thus important to quantify these patterns in the population of interest.

How can distance-dependent LD be explained? First, some recombination must be occurring although the clonal frame is still evident since a fraction of even very distant loci may be linked [29]. Second, the majority of recombinant fragments introduced by gene conversion are probably small [40-43]. Third, it follows from the simulation results that *r* must be large and/or *s* must be small. This leads to three different scenarios:

1. the strains in question form an effectively sexual population (*r* is large),
2. selection coefficients are low or selection is infrequent (*s* is small), or
3. there are ecological barriers to selective sweeps but not recombination

The first two scenarios are quite straightforward, but the third merits further discussion. In Scenario three, different populations of bacteria might inhabit different micro-niches, making it rare for a single genotype to sweep through all populations. How likely is such a scenario? In the coastal water column, at least 15 ecologically distinct subpopulations of *Vibrio splendidus* co-exist [44] . Because the coastal ocean is relatively well-mixed, there might be even more opportunity for resource partitioning and niche subdivision in other (e.g. terrestrial or host-associated) environments (e.g. [45]). Therefore, a meta-population model, such as the one described by Majewski and Cohan [34], could explain the distance-dependent decay of LD. Their model requires two or more populations, each adapted to a different niche and these niches might only be very slightly different or even transient niches. Each population experiences independent selective sweeps, purging diversity within each sub-population, yet genetic diversity

remains high when summed over all populations. Occasionally, a globally adaptive mutant occurs in one of the populations, but cannot sweep through other populations because its genotype as a whole is relatively unfit in the other micro-niches. But if the globally adaptive mutation is recombined into the 'native' background of another population, it can confer a fitness advantage and rise in frequency.

Eventually, globally adapted mutations can purge diversity locally in the genome without disrupting unlinked genomic diversity. Thus, Majewski and Cohan's model could explain the observed distance-dependent decay of LD without invoking a high rate of recombination. Because niche partitioning in the microbial world likely occurs at a very fine scale [44, 46], it is probable that many bacterial population samples actually encompass multiple subpopulations, connected by rare recombination of globally adaptive alleles.

When is recombination an adaptive event?

Under Scenarios one and two above, neutral recombinational events occur faster than selection, whereas under Scenario three recombinant genotypes are driven to high frequency by selection. So how much of recombination is adaptive? On the one hand, recombination across wide phylogenetic distances (by horizontal gene transfer), followed by conservation of the foreign DNA in the recipient, in itself provides compelling evidence for adaptive evolution. But the picture is not as clear for homologous recombination between closely-related strains.

Recent work by helps clarify the relationship between recombination and positive selection [47]. Recombination and positive selection were both quantified in the *Streptococcus* core genome,

and it was concluded that genes under positive selection are frequently recombined, a result recently supported in a study of *Listeria* genomes [48]. Specifically, 78% of genes under positive selection in the *S. pyogenes* core genome were also inferred to be recombinant [47]. Yet, of the genes identified as recombinant within this species, only a small fraction experienced positive selection (Figure 3). Therefore, although positively selected genes are frequently recombined, a substantial amount of within-species recombination shows no evidence of direct adaptive value. In other words, recombination within a species could be largely neutral. But this is not the case for recombination between species (from *S. agalactiae* to *S. pyogenes*, or vice versa). Comparison of between-species and within-species evolutionary events yields a surprising insight not highlighted in the original work: 81% of all genes recombined between species also experienced positive selection, whereas only 4% of genes recombined within species also experienced positive selection (Figure 3).

This striking result has several implications. First, it provides empirical confirmation of Scenario three above and Milkman's hypothesis that in order for a horizontally transferred gene to be fixed (at least across species), it must enjoy a "considerable selective advantage" [49]. Second, it furnishes evidence that short-distance recombination events are likely to be neutral and unlikely to be under positive selection (Scenarios one and two). Third, it provides inspiration for a new class of tests for positive selection in bacterial populations: identify positively selected genes in bacterial populations as recombinant sequences transferred across population boundaries. Fourth, it suggests a pragmatic solution to the long-standing challenge of defining bacterial species. Sexual eukaryotic species undergo neutral recombination in each generation. By analogy, a group of bacteria that undergo frequent neutral recombination could also constitute

a discrete species. Recombination between species is not precluded in this species definition, but would require positive selection for maintenance of the introduced recombinant sequences (Figure 4). While previous studies have shown that clusters of closely-related strains (or putative species, with more frequent neutral recombination within than between species) can theoretically emerge and be maintained in the absence of selection, these same studies suggest that neutral recombination alone is insufficient to explain fine-scale genetic differentiation actually observed among clusters, supporting the idea that speciation might require population structure (*e.g.* microepidemics) or positive selection [50-52].

Detecting selection among species and higher-order groups

Over millions to billions of years of evolution, populations of bacteria have diverged to form distinct species (although there is considerable controversy over exactly when two populations can be called independent species [53, 54]). This process might occur by restricted gene flow, followed by neutral drift, or by niche partitioning and natural selection [51, 52]. Recombination between distant species is relatively rare and often adaptive (as discussed above), and potentially straightforward to detect using phylogenetic methods (e.g. http://almlab.org/AnGST).

Substitution patterns indicative of positive selection at long time scales (between rather than within populations) could be detected using codon-based (*dN/dS*) and other relative rates-based methods (e.g. selective signatures, M-K test; see Table 1). The ratio of nonsynonymous to synonymous substitution rates has been widely used in genome-wide scans for positive selection in bacteria, often providing evidence for function- or gene-specific selection [47, 55]. Yet *dN/dS* is inappropriate when comparing either very distantly-related strains (*dS* saturated with multiple

substitutions), or very closely-related strains, within which *dN/dS* is inflated by segregating nonsynonymous polymorphism [56, 57].

Metrics that explicitly measure deviations from the expected pattern of amino acid substitution (relative to a within-species near-neutral expectation as in the M-K test, or to a between-species expectation as in selective signatures) are perhaps better suited to detecting sequence-level changes associated with changes in ecological preferences. Recently, such deviations from a protein's expected rate of evolution (based on the genome and protein family to which it belongs) were quantified as its selective signature, identifying substitutions with potential ecological relevance [9]. This approach can yield insights into the cellular functions and pathways that contribute to niche adaptation. For example, selective signatures showed that glycolysis and phenylalanine metabolism genes evolve unusually rapidly in *Idiomarina loihiensis*, mirroring this lineage's shift in carbon source preference from sugars to amino acids.

Nearly every rate-based test (Table 1) has the potential to mistake recombination for positive selection (e.g. [58, 59]), but it is possible to control for recombination by ensuring that the correct gene phylogeny is being used while testing for selection. Certain implementations of the M-K test, for example, assume that all genes in the genomes being compared diverged at the same time [60] – an assumption that is violated when genes have different histories of recombination. Thus, if care is not taken to control for recombination before testing for selection, these two evolutionary events – both potentially interesting and with adaptive merit – may easily be confused.

Finally, there is mounting evidence that taxonomic units broader than individual species indeed have ecological meaning, and thus show similar patterns of selection. For example, clades of bacteria in the same Family or Order tend to have similar habitat preferences [61]. In comparisons of obese (enriched in Firmicutes) and lean (enriched in Bacteroidetes) gut microbiomes, habitat preference was observed at the level of Division [62, 63]. The genetic basis of these higher-order habitat preferences is only just beginning to be elucidated, and likely involves both genome content and sequence-level variation.

Concluding remarks and future directions

Identifying the signature of natural selection in microbial genomes can help to shed light on the hidden world of microbes. Which techniques can be used to identify positive selection depends on the rates and bounds of recombination in microbial populations. The first step in any study of natural selection in bacteria is to quantify the extent of recombination within a population before moving on to sequence-based tests. Once recombinant portions of the genome are identified, they can be tested for evidence of positive selection using diversity-based methods. Meanwhile, the non-recombinant clonal frame can be identified (e.g. using the ClonalFrame program [29]), and tested for convergent evolution or excessive rates of functional substitutions (Figure 4). If possible, all tests should be performed genome-wide to estimate and control for demographic effects (Table 1) that might otherwise provide spurious evidence of positive selection.

In his "Difficulties on Theory" chapter in *On the Origin of Species*, Darwin wrote: "We are profoundly ignorant of the causes producing slight and unimportant variations [...]" [64]. On the sesquicentennial of its original publication, we know that random mutation and recombination

are the causes of heritable fitness variations, yet we remain largely ignorant of the selective pressures that cause advantageous variations to be favored and maintained. Even within our own species, the list of uncontroversial cases of selective pressures leading to genetic adaptations is not long. Yet the list of candidate adaptive variations has grown much longer since genome-wide scans for selection became viable in humans [65], and we are beginning to see the same happen for microbes. Metazoans and microbes will soon be on similar footing; with a list of candidate genes in hand, the challenge will be to translate this list into a meaningful genome-wide map of selection, linking genetic variation to phenotype and ecology. With such genome-wide maps, we are optimistic that evolutionary adaptations will be revealed, even at the finest resolutions - for example, among closely-related, yet ecologically differentiated subpopulations of *Vibrio splendidus* in the coastal ocean [44, 46]. To encourage efforts in this direction, a team of microbial ecologists, population geneticists, and genome sequencing experts have established a database of nearly 100 complete genome sequences of marine *Vibrio* strains encompassing multiple ecologically distinct populations, which will be freely available (at http://micropopgen.org). Darwin was right in saying that many variations are slight (e.g. a single nucleotide mutation that subtly alters protein structure or expression), but cumulatively they leave a trail of footprints, which, given the right set of population genomic tools, will ultimately lead us to a better understanding of the microbial world.

Glossary

Positive/diversifying selection: The evolutionary force causing novel alleles conferring a fitness advantage to rise in frequency in a population. This leads to reduced genetic variation at the selected locus within the population, but increased genetic variation between populations.

Negative/purifying/stabilizing selection: The evolutionary force selecting against deleterious mutations and promoting conservation of the ancestral state.

Neutral drift: The process by which mutations with negligible effects on fitness become stochastically fixed in a population of finite size.

Selective sweep: The process of a positively-selected allele rising in frequency and ultimately becoming fixed in a population. In the absence of recombination, a single clone will sweep through the population, purging genetic diversity genome-wide. In the presence of recombination, diversity may be retained at genomic loci that are unlinked to the selected allele.

Panmictic population: A population undergoing frequent recombination (e.g. a sexual population in which recombination occurs every generation). Qualitatively, this results in random association between loci. More quantitatively, panmixis may be defined as when a single nucleotide change is more likely to have resulted from a recombination event than a mutation event ($r/m > 1$).

Clonal population: A population which never (or extremely rarely) undergoes recombination. All loci in the genome are thus in complete linkage, meaning that a selective sweep will affect diversity in the entire genome, not just at a selected locus.

Globally adaptive mutation: In a meta-population model, a globally adaptive mutation confers a fitness advantage in all of the sub-populations making up the meta-population. If the mutation is recombined into a sub-population, it will purge genetic diversity only in the recombined portion of the genome.

Niche partitioning: The process whereby different organisms co-exist in a community rather than competing for resources. Niches can be partitioned when lineages avoid competition by using different resources, or restricting their activity to different physical spaces, seasons, times of day, etc.

Tajima's *D*: A statistic to measure deviations from allele frequencies expected in a population evolving under a neutral model. Deviations may indicate purifying selection ($D < 0$), diversifying selection or population subdivision ($D > 0$), or a recent selective sweep or population bottleneck ($D < 0$).

Long-range haplotypes: A test to detect positively-selected alleles that have risen to high frequency in a population in a short period of time, so that recombination has not had time to break down linkage to distant hitchhiking mutations. The test exploits the genome-wide distribution of allele frequencies and haplotype lengths to detect outlying haplotypes that are at unusually high frequency for their length.

McDonald-Kreitman (M-K) test: A test for selection on protein-coding nucleotide sequences that measures an unusually high between-species dN/dS, relative to a near-neutral standard of within-population dN/dS.

Selective signatures: A measure of selection that can be applied to nucleotide or protein sequences from distantly related species. Selective signatures quantifies the extent to which a gene deviates from the evolutionary rate (number of substitutions per site) predicted by the gene family and genome it belongs to. Such deviations suggest gene-specific, species-specific changes in the selective pressures on a gene.

Phylogenetic incongruence: If a gene has experienced horizontal transfer, duplication, and/or loss in some lineages, this will often result in the gene's phylogeny (gene tree) having a different topology from the species' phylogeny. Phylogenetic incongruence is often used as evidence for horizontal transfer.

Convergence: The independent fixation of the same mutation in two or more independent (distantly-related) lineages, also called homoplasy, is often used as evidence for positive selection. Because it generates phylogenetic incongruence (see above), it may also be used as evidence for recombination. Recombination may be ruled out if the convergence is restricted to a single mutation, rather than a long stretch of mutations. If the convergence consists of different nucleotide-level mutations that result in convergence to the same amino acid, this also supports positive selection as a more likely explanation that recombination.

Acceptor lineage: In a recombination or horizontal transfer event, the acceptor lineage is the recipient of a stretch of novel DNA.

Restricted gene flow: The reduction or prevention of recombination between bacterial lineages owing to physical or ecological barriers, or DNA sequence divergence.

References

1 Beja, O. *et al*. (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* 289, 1902-1906

2 Sabehi, G. *et al*. (2004) Different SAR86 subgroups harbour divergent proteorhodopsins. *Environ. Microbiol.* 6, 903-910

3 Bielawski, J.P. *et al*. (2004) Darwinian adaptation of proteorhodopsin to different light intensities in the marine environment. *Proc. Natl. Acad. Sci. U. S. A.* 101, 14824-14829

4 Martiny, A.C. *et al*. (2006) Phosphate acquisition genes in *Prochlorococcus* ecotypes: evidence for genome-wide adaptation. *Proc. Natl. Acad. Sci. U. S. A.* 103, 12552-12557

5 Feldgarden, M., Byrd, N. and Cohan, F.M. (2003) Gradual evolution in bacteria: evidence from *Bacillus* systematics. *Microbiol.* 149, 3565-3573.

6 Black, W.C.,4th *et al*. (2001) Population genomics: genome-wide sampling of insect populations. *Annu. Rev. Entomol.* 46, 441-469

7 Luikart, G. *et al*. (2003) The power and promise of population genomics: from genotyping to genome typing. *Nat. Rev. Genet.* 4, 981-994

8 Kosiol, C. *et al*. (2008) Patterns of positive selection in six Mammalian genomes. *PLoS Genet.* 4, e1000144

9 Shapiro, B.J. and Alm, E.J. (2008) Comparing Patterns of Natural Selection across Species Using Selective Signatures. *PLoS Genet.* 4, e23

10 Sabeti, P.C. *et al*. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419, 832-837

11 Kelley, J.L. *et al*. (2006) Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res.* 16, 980-989

12 Sabeti, P.C. *et al*. (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* 449, 913-918

13 Tishkoff, S.A. *et al*. (2007) Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* 39, 31-40

14 Holt, K.E. *et al*. (2008) High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. *Nat. Genet.* 40, 987-993

15 Sokurenko, E.V. *et al*. (2004) Selection footprint in the FimH adhesin shows pathoadaptive niche differentiation in *Escherichia* coli. *Mol. Biol. Evol.* 21, 1373-1383

16 Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585-595

17 Zeng, K. *et al*. (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* 174, 1431-1439

18 Mes, T.H. *et al*. (2006) Selection on protein-coding genes of natural cyanobacterial populations. *Environ. Microbiol.* 8, 1534-1543

19 Herbeck, J.T. *et al*. (2003) A conservative test of genetic drift in the endosymbiotic bacterium *Buchnera*: slightly deleterious mutations in the chaperonin groEL. *Genetics* 165, 1651-1660

20 Jolley, K.A. *et al*. (2005) The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in *Neisseria meningitidis*. *Mol. Biol. Evol.* 22, 562-569

21 Guttman, D.S. *et al*. (2006) Diversifying selection drives the evolution of the type III secretion system pilus of *Pseudomonas syringae*. *Mol. Biol. Evol.* 23, 2342-2354

22 Hughes, A.L. (2005) Evidence for abundant slightly deleterious polymorphisms in bacterial populations. *Genetics* 169, 533-538

23 McVean, G. *et al*. (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 160, 1231-1241

24 Milkman, R. and Bridges, M.M. (1990) Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. *Genetics* 126, 505-517

25 Falush, D. *et al*. (2001) Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock rates, recombination size, and minimal age. *Proc. Natl. Acad. Sci. U. S. A.* 98, 15056-15061

26 Wiehlmann, L. *et al*. (2007) Population structure of *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. U. S. A.* 104, 8101-8106

27 Kondrashov, F.A. and Kondrashov, A.S. (2001) Multidimensional epistasis and the disadvantage of sex. *Proc. Natl. Acad. Sci. U. S. A.* 98, 12089-12092

28 Feil, E.J. *et al*. (1999) The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. *Mol. Biol. Evol.* 16, 1496-1502

29 Didelot, X. and Falush, D. (2007) Inference of bacterial microevolution using multilocus sequence data. *Genetics* 175, 1251-1266

30 Drake, J.W. (1991) A Constant Rate of Spontaneous Mutation in DNA-Based Microbes. *Proc. Natl. Acad. Sci. U. S. A.* 88, 7160-7164

31 Barrett, R.D. *et al*. (2006) Mutations of intermediate effect are responsible for adaptation in evolving *Pseudomonas fluorescens* populations. *Biol. Lett.* 2, 236-238

32 Atwood, K.C. *et al*. (1951) Selective mechanisms in bacteria. *Cold Spring Harb. Symp. Quant. Biol.* 16, 345-355

33 Cohan, F.M. (2001) Bacterial species and speciation. *Syst. Biol.* 50, 513-524

34 Majewski, J. and Cohan, F.M. (1999) Adapt globally, act locally: the effect of selective sweeps on bacterial sequence diversity. *Genetics* 152, 1459-1474

35 Koeppel, A.F. and Cohan, F.M. (2008) The origins of ecological diversity in prokaryotes. *Curr. Biol.* 18, R1024-R1034

36 Rozen, D.E. *et al*. (2005) Long-term experimental evolution in *Escherichia coli*. XIII. Phylogenetic history of a balanced polymorphism. *J. Mol. Evol.* 61, 171-180

37 Hegreness, M. *et al*. (2006) An equivalence principle for the incorporation of favorable mutations in asexual populations. *Science* 311, 1615-1617

38 Blount, Z.D. *et al*. (2008) Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proc. Natl. Acad. Sci. U. S. A.* 105, 7899-7906

39 Smith, J.M. *et al*. (1993) How clonal are bacteria? *Proc. Natl. Acad. Sci. U. S. A.* 90, 4384-4388

40 Fearnhead, P. *et al*. (2005) Analysis of recombination in *Campylobacter jejuni* from MLST population data. *J. Mol. Evol.* 61, 333-340

41 Mau, B. *et al*. (2006) Genome-wide detection and analysis of homologous recombination among sequenced strains of *Escherichia coli*. *Genome Biol.* 7, R44

42 Guttman, D.S. and Dykhuizen, D.E. (1994) Detecting selective sweeps in naturally occurring *Escherichia coli*. *Genetics* 138, 993-1003

43 McKane, M. and Milkman, R. (1995) Transduction, restriction and recombination patterns in *Escherichia coli*. *Genetics* 139, 35-43

44 Hunt, D.E. *et al*. (2008) Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* 320, 1081-1085

45 Koeppel, A. *et al*. (2008) Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. *Proc. Natl. Acad. Sci. U. S. A.* 105, 2504-2509

46 Thompson, J.R. *et al*. (2005) Genotypic diversity within a natural coastal bacterioplankton population. *Science* 307, 1311-1313

47 Lefebure, T. and Stanhope, M.J. (2007) Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome Biol.* 8, R71

48 Orsi, R.H. *et al*. (2008) Genome-wide analyses reveal lineage specific contributions of positive selection and recombination to the evolution of *Listeria monocytogenes*. *BMC Evol. Biol.* 8, 233

49 Milkman, R. *et al*. (2003) Molecular evolution of the *Escherichia coli* chromosome. VI. Two regions of high effective recombination. *Genetics* 163, 475-483

50 Fraser, C. *et al*. (2005) Neutral microepidemic evolution of bacterial pathogens. *Proc. Natl. Acad. Sci. U. S. A.* 102, 1968-1973

51 Hanage, W.P. *et al*. (2006) Modelling bacterial speciation. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 361, 2039-2044

52 Fraser, C. *et al*. (2007) Recombination and the nature of bacterial speciation. *Science* 315, 476-480

53 Gevers, D. *et al*. (2005) Opinion: Re-evaluating prokaryotic species. *Nat. Rev. Microbiol.* 3, 733-739

54 Doolittle, W.F. and Papke, R.T. (2006) Genomics and the bacterial species problem. *Genome Biol.* 7, 116

55 Chen, S.L. *et al*. (2006) Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach. *Proc. Natl. Acad. Sci. U. S. A.* 103, 5977-5982

56 Rocha, E.P. *et al.* (2006) Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J. Theor. Biol.* 239, 226-235

57 Kryazhimskiy, S. and Plotkin, J.B. (2008) The population genetics of dN/dS. *PLoS Genet.* 4 e1000304

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

59 Anisimova, M., Nielsen, R. and Yang, Z. (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* 164, 1229-1236.

60 Bustamante, C.D. *et al.* (2005) Natural selection on protein-coding genes in the human genome. *Nature* 437, 1153-1157

61 von Mering, C. *et al*. (2007) Quantitative Phylogenetic Assessment of Microbial Communities in Diverse Environments. *Science* 315, 1126-1130

62 Ley, R.E. *et al*. (2006) Microbial ecology: human gut microbes associated with obesity. *Nature* 444, 1022-1023

63 Turnbaugh, P.J. *et al*. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444, 1027-1031

64 Darwin, C. (1859) *The Origin of Species,* Penguin Classics

65 Sabeti, P.C. *et al*. (2006) Positive natural selection in the human lineage. *Science* 312, 1614-1620

66 de Queiroz, K. (2005) Different species problems and their resolution. *Bioessays* 27, 1263-1269

67 Hey, J. (2006) Recent advances in assessing gene flow between diverging populations and species. *Curr. Opin. Genet. Dev.* 16, 592-596

68 McDonald, J.H. and Kreitman, M. (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature* 351, 652-654

69 Lewontin, R.C. and Krakauer, J. (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 74, 175-195

70 Kalinowski, S.T. and Hedrick, P.W. (2001) Estimation of linkage disequilibrium for loci with multiple alleles: basic approach and an application using data from bighorn sheep. *Heredity* 87, 698-708

Figure Legends

Figure 1. Population diversity following a selective sweep with varying selection and recombination rate.

(A) We simulated a selective sweep in an initially diverse population of size $N=10^7$, mutation rate $m=10^{-10}$ per bp per generation and a range of selection coefficients (σ) and recombination rates (ρ). After the beneficial allele has swept through the population, we computed the population's Clonality as Σp_i^2, where p_i is the frequency of the i^{th} allele at the background locus. Populations with small ρ and large σ are dominated by few genotypes, while populations with large ρ and small σ consist of many genotypes following fixation of a beneficial allele. (B) Effect of population size on diversity following a selective sweep. $\rho=Nr$ and $\sigma=Ns$ are the 'natural' variables of the system, such that populations of different sizes with same values of ρ and σ (at fixed m) have similar structure following a selective sweep. Eight combinations of ρ and σ are compared using simulations with different population sizes. For each population size, we plot the clonality index vs. that observed for a population of size $N=10^7$. Deviations from the dashed line (y=x), are relatively small indicating that the results hold over a range of population sizes.

Figure 2. Illustration of distance-dependent decay of LD in the *E. coli* genome.

We gathered 1672 core orthologs present in each of 24 *E. coli* strains, as described in [9]. Each unique allele at a given locus was assigned a unique allele number. We then chose pairs of loci separated by increasing distances in the *E. coli* K12 reference genome. Pairs of loci on the same operon and neighboring loci on the same strand were excluded. LD was estimated using the D_A' metric, which provides a summary measure of LD between 2 loci, each containing an arbitrary

number of alleles [70]. When $D_A' = 1$, linkage is at its theoretical maximum. (A) For pairs of loci separated by increasing genetic distance (kb, on a $\log_{10}$ scale), the proportion of pairs in full linkage (# of pairs with $D_A' = 1$ / total # of pairs in that distance bin) is plotted on the y-axis. Points on the x-axis were binned as in Figure 1. Inset: Distances of 0-100 kb shown on a linear scale (red points).

Figure 3. Intersections of sets of recombining and positively selected genes in *Streptococcus* spp.

Overall, Lefebure and Stanhope [47] observed recombination in 753 genes and positive selection in 217 genes; 72 genes experienced both. The overall dataset (top) was segmented into a genus-wide core-genome (between species, bottom left) and a species-specific set of genomes (within species, bottom right). 53 genes were found to recombine between species of which 43 experienced positive selection. By contrast, 700 genes were found to be recombinant within species, but only 29 of these experienced positive selection.

Figure 4. Flow chart of methods to identify positively selected loci in bacteria.

Tajima's *D* and Fay and Wu's *H* tests measure unusually high or low allele frequencies within a population. They require a sample of allele sequences, preferably genome-wide to help quantify recombination and demographics, representing polymorphism within a population. The tests differ in that the *H* test also requires an outgroup species to distinguish derived and ancestral mutations, allowing it to distinguish positive and negative selection [17]. The M-K test also requires a sample of alleles from a population, and at least one outgroup species, but differs from the diversity-based tests in that it is restricted to protein-coding genes. The important assumption

of the M-K 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 outgroup and ingroup) as for segregating polymorphism (within the ingroup). When the ratio of fixed:polymorphic dN/dS exceeds 1, this provides strong evidence that positive selection has played a role in the divergence the of outgroup and ingroup [68]. AnGST is a phylogeny-based approach to detecting recombination ([44], http://almlab.org/AnGST). It identifies ancestral recombinations and specifies donors and acceptor lineages. See the main text and glossary for brief descriptions of other methods.

Box 1. Key challenges in bacterial population genetics

Compared to eukaryotic systems, our limited understanding of microbial population genetics can be summarized by several key challenges in studying environmental and host-associated bacterial communities:

- Limited understanding of gene flow patterns and population boundaries in bacteria. Arguably, a universally accepted species definition is lacking even for animal taxa [66, 67]; however, the problem is exacerbated in bacteria where the rates and bounds (genetic and/or ecological) of gene flow are not known. Complicating matters further, different natural populations likely occupy a continuum of recombinational rates from clonal to panmictic.

- Lack of approaches to detect recent positive selection. A wealth of statistical tools based on allele frequencies or haplotype structure are available for detecting the signature of natural selection in sexual eukaryotes, but which if any of these tests can be adapted for use in bacteria has not been studied.

- The unknown role of the ‘peripheral’ genome. A large fraction of the genetic diversity within microbial lineages is contained within the ‘peripheral’ or ‘flexible’ genome -- strains that are nearly identical in nucleotide sequence at orthologous loci can differ by genomic islands containing megabases of strain-specific DNA. The extent to which this extraordinary diversity contributes to adaptive evolution is not known.

Acknowledgements

This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics Program:GTL through contractDE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.