Neutral Theory, Microbial Practice: Challenges in Bacterial Population Genetics

Eduardo P C Rocha

Neutral Theory, Microbial Practice: Challenges in Bacterial Population Genetics

Eduardo P.C. Rocha*[,1,2]

[1]Microbial Evolutionary Genomics, Institut Pasteur, Paris, France

[2]CNRS, UMR3525, Paris, France

***Corresponding author:** E-mail: erocha@pasteur.fr.

Associate editor: Sudhir Kumar

Abstract

I detail four major open problems in microbial population genetics with direct implications to the study of molecular evolution: the lack of neutral polymorphism, the modeling of promiscuous genetic exchanges, the genetics of ill-defined populations, and the difficulty of untangling selection and demography in the light of these issues. Together with the historical focus on the study of single nucleotide polymorphism and widespread non-random sampling, these problems limit our understanding of the genetic variation in bacterial populations and their adaptive effects. I argue that we need novel theoretical approaches accounting for pervasive selection and strong genetic linkage to better understand microbial evolution.

Key words: **bacteria, horizontal gene transfer, species definition, selection, recombination, demography.**

Introduction

Bacteria have huge census population sizes, organized genomes, and multiple ways of exchanging genetic material. They are perfect subjects of study for researchers looking for complex problems in population genetics. Motoo Kimura was known for his passion for plants and benefitted enormously from decades of work on the population genetics of flies (Crow 1995). Here, I'll argue that the lines of study where he was most influential face particularly interesting challenges when they are applied to microbes. I include the neutral theory among these studies, but I will also dedicate attention to Kimura's work on the use of methods from statistical mechanics to model evolutionary processes. My goal is not to provide an extensive description of Kimura's works, for which his classical book remains an essential reference (Kimura 1983), but to highlight a few of his key results and use them as a basis to understand advances and standstills in microbial evolutionary biology (table 1). Although I will focus on bacteria, this text should be equally pertinent to Archaea, for which population genetics studies remain rare.

Are There Neutral Polymorphisms in Prokaryotes?

"The neutral mutation-random drift hypothesis (or the neutral theory for short) holds that at the molecular level most evolutionary change and most of the variability within species are not caused by Darwinian selection but by random drift of mutant alleles that are selectively neutral or nearly neutral. The essential part of the neutral theory is not so much that molecular mutants are selectively neutral in the strict sense as that their fate is largely determined by random genetic drift" (Kimura 1983).

The second Principle of Molecular Evolution states that "functionally less important molecules or parts of a molecule evolve faster than more important ones" (Kimura and Ohta 1974). It has been the cornerstone of molecular evolution and in particular of studies unraveling the evolution of bacterial genomes, where cellular mechanisms constrain the organization and evolution of the chromosome(s) (Rocha 2008). In bacteria, genes correspond to 85–90% of the genome and intergenic regions are small and packed with regulatory sites (Mira et al. 2001). Genes have few or no introns and are usually organized in polycistronic units (operons) where insertions and deletions of genetic material usually have deleterious effects on gene expression (Price et al. 2006). Nonsynonymous mutations are much more numerous and have stronger fitness effects than synonymous mutations (Nei and Gojobori 1986). Because these fitness effects are usually negative, such mutations are gradually purged from populations by natural selection—purifying selection—leading to low ratios of nonsynonymous (dN) over synonymous (dS) substitution rates ($dN/dS \ll 1$; Kuo et al. 2009). Yet, synonymous substitutions can also be affected by natural selection, for example, for codon usage bias (Gouy and Gautier 1982; Sharp et al. 2005), especially in fast-growing bacteria (Vieira-Silva and Rocha 2010). Intergenic regions rarely correspond to >20% of the genome and are subject to purifying selection at levels intermediate between that of nonsynonymous and synonymous substitutions (Thorpe et al. 2017). Overall, this means that most mutations of the bacterial genome are deleterious (fig. 1A).

Perspective

Table 1. Some Ongoing Debates between Neutralist and Selectionist Arguments in Microbial Evolution.

Topic	Views & Controversies
Huge pan-genomes	Are the genes acquired by HGT neutral, deleterious, or adaptive? Inspired on the infinite site model (Kimura 1969), the IMG model shows that the distribution of accessory genomes in *Prochlorococcus* is consistent with the expectation of the neutral model, whereas in *Synechococcus* it slightly deviates from neutrality (Baumdicker et al. 2012). Others found a positive correlation between a quantity called genome fluidity and genetic diversity (taken as a proxy of effective population size) arguing that most transferred genes are neutral (Andreani et al. 2017) (although in this case diversity may be the consequence of transfer). Still others argue for frequent adaptive HGT (Sela et al. 2016; McInerney et al. 2017). Finally, it has been suggested that many accessory genes are under frequency-dependent selection (Cordero and Polz 2014), but since few gene families are at intermediate frequencies (e.g., fig. 3C), this would suggest that many genes are adaptive only when rare or in very specific environmental conditions.
Genome reduction	Many bacterial lineages have endured genome reduction from ancestors with larger gene repertoires. This has been explained by selection for gene loss or lack of selection for gene maintenance (Batut et al. 2014). Bacteria enduring niche contraction associated with frequent bottlenecks and sexual isolation have shrinking genomes that are thought to result from genetic drift under small effective population sizes (Moran 1996). This is in agreement with the lower efficiency of selection observed in smaller genomes (Kuo et al. 2009). However, some free-living species, with supposedly high effective population sizes, have very small genomes and are thought to select for genome streamlining (Giovannoni et al. 2014). It has also been proposed that bacteria select for the loss of functions that are provided by the community (Black Queen hypothesis; Morris et al. 2012). It is yet unclear if two last mechanisms are very frequent.
Distribution of fitness effects	What is the distribution of fitness effects of mutations segregating in bacterial populations? There is little data available (for a review Gordo et al. 2011). It has been proposed, based on an extension of the McDonald–Kreitman test, that >50% of fixed amino acid substitutions between *E. coli* and *S. enterica* were selected for (Charlesworth and Eyre-Walker 2006). Experimental evolution studies usually reveal initial fixation of many adaptive mutations (see text). But most studies in phylodynamics and phylogenetics intrinsically assume that substitutions are neutral (see references in (Lapierre et al. 2016).
Base composition	GC content varies very widely in bacteria affecting amino acid composition of proteins (Sueoka 1961), which was taken as confirmation of the neutral theory. However, GC content correlates with genome size and is lower in MGEs (Rocha and Danchin 2002; Bentley and Parkhill 2004). The mutation spectra tends to be AT-rich relative to the genome composition (Hershberg and Petrov 2010; Hildebrand et al. 2010), which does not fit the expectation of the neutral theory. There are many theories, and some experimental confirmation (Raghavan et al. 2012), on why there could be selection for nucleotide composition. Alternatively, biased gene conversion (BGC), a proposed bias favoring G and C mismatches during homologous recombination, could explain the gap between mutational patterns and genome composition (Lassalle et al. 2015). Selection and BGC affect polymorphism in very similar ways and are thus hard to distinguish. Selection is favored by recombination. BGC strictly depends on the rate of recombination.
Mutation rate	It has often been stated, with little supporting evidence, that mutation rates were the result of a trade-off between the cost of mutations and the cost of DNA repair. Kimura produced the first rigorous treatment of this question and suggested that mutation rates were under genetic control, that is, could be subject to selection and were a product of past evolution like many other characters (Kimura 1967). Mutation rates vary widely in bacteria: from 10^{-8} to $<10^{-10}$ per site per generation (Lynch et al. 2016). Rates per year are also variable: from 10^{-5} per site per year in *Mycoplasma gallisepticum* (Delaney et al. 2012) to 10^{-7} per site per year in *Mycobacterium tuberculosis* (Ford et al. 2011). These rates are averages, since mutation rates increase dramatically upon loss of repair genes or under deleterious physiological conditions. Furthermore, mutator phenotypes can be transiently selected for (or hitchhike with linked adaptive alleles) when bacteria are poorly adapted (Tenaillon et al. 2001). Mobile elements often evolve to generate diversity under conditions where bacteria are poorly adapted, for example, under stress (Beaber et al. 2004), and some bacteria encode contingency loci that increase local mutation rates in genes under frequent selection for diversification (Moxon et al. 1994). In contrast, it has been proposed that selection favors low mutation rates and the decrease in the latter is limited by the efficiency of selection (Sung et al. 2012). This fits the observation that bacteria, having high Ne, have among the lowest mutation rates (Lynch et al. 2016).

In the absence of strong genetic linkage (an unwise assumption for bacteria, see below), mutations behave neutrally when $|Ne{\cdot}s| \ll 1$ (Kimura 1968), where *Ne* is the effective population size and *s* the selection coefficient. It has often been claimed that bacteria have huge *Ne* (Lynch et al. 2016), and thus even weak fitness effect mutations should be affected by natural selection. As repeatedly pointed out by Kimura, a large load of deleterious mutations is not necessarily evidence against the neutral theory if they are quickly purged from populations (Kimura 1983). This point can be assessed by analyzing the variation in the rates of dN/dS between increasingly divergent genomes. When genomes are very close, the ratio is close to one, indicating little apparent differences between synonymous and nonsynonymous changes. For more divergent genomes, the ratio rapidly decreases to low values, suggesting that many nonsynonymous mutations have weak negative effects in fitness and remain in populations for a long period of time before being purged by natural selection (Rocha et al. 2006). This is not consistent with the neutral theory, but might fit the Nearly Neutral Theory, where molecular evolution is dominated by weak selection of common variants (Ohta 1992). The abundance of weakly deleterious polymorphism has consequences for the evolution of gene repertoires. Bacteria that become obligatory endosymbionts endure sexual isolation and frequent population bottlenecks. This lowers the efficiency with which natural selection can remove deleterious alleles and many genes are then lost (Moran 1996; Andersson and Kurland 1998, table 1).

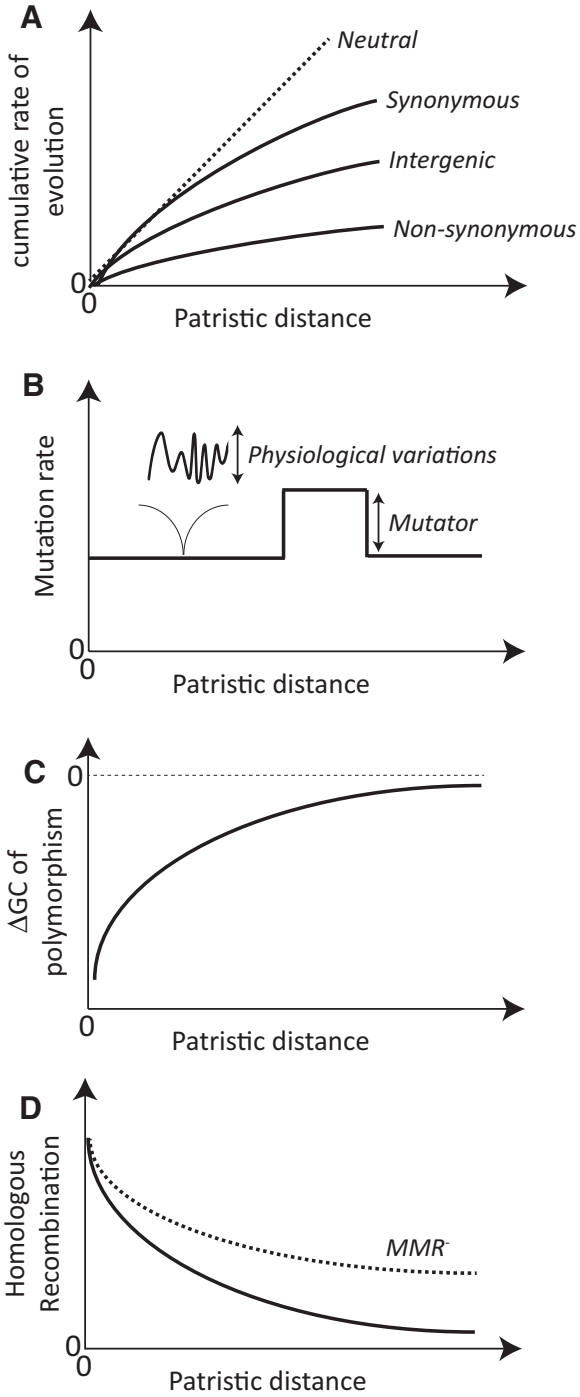

Fig. 1. Association between patristic distance between bacteria (evolutionary distance computed from their phylogenetic tree) and a number of traits. (*A*) As time passes mutations accumulate in the two lineages. If there is no recombination and no major changes in mutation rates, neutral mutations accumulate linearly. In fact, mutations in bacteria tend to accumulate at a slower pace, because the rate of evolution is retarded by purifying selection (purge of deleterious alleles). In general, synonymous substitutions accumulate at higher rates than non-synonymous ones, as expected given the frequent deleterious effects of the latter. (*B*) Mutation rates change in function of physiological processes and by acquisition or loss of DNA repair genes. (*C*) Polymorphism is AT rich relative to the genome composition, suggesting that the effect of the mutational patterns on genome composition is moderated by the action of other forces (table 1). When one compares more distant genomes, a larger fraction of the differences between them has been in populations for some time, during which they endured the effect of selection or biased gene conversion and gradually became less AT rich. (*D*) The rate of homologous recombination decreases quickly with the divergence between sequences.

Adaptation by natural selection depends on the supply of mutations. A lot of work has been devoted to bacterial mutation rates and how they change in response to the cell physiology and environment (table 1, fig. 1B). We know much less about the distribution of fitness effects of these mutations and how they depend on environmental factors. Mutations that actually reach fixation may often be adaptive. For example, one of the few available studies inferred that more than half of the amino acid substitutions were fixed by positive selection in enterobacteria (Charlesworth and Eyre-Walker 2006).

Experimental evolution studies provide simplified but interesting settings where adaptation can be tracked at a very fine timescale. The long-term evolution experiment of *Escherichia coli*, where populations are large, shows that adaptation occurs predominantly by fixation of adaptive nonsynonymous mutations (Barrick and Lenski 2013). The fixation of mutations in this experience is initially almost linear, which could have been misconstrued as a prediction of the neutral theory. However, the mutations are adaptive and the linear trend is presumably caused by clonal interference—the competition between clones carrying different adaptive mutations—where genetic diversity is purged by the rapid population expansion of one clone in the population (hard selective sweep; Barrick et al. 2009). *Escherichia coli* experimental evolution in conditions of intermediate effective population sizes, to minimize the effects of both drift and clonal interference, revealed a rate of beneficial mutations of 10^{-5} per genome per generation with mean selective advantage of 1% (Perfeito et al. 2007). Mutation accumulation studies, which are evolution experiences where effective population sizes are kept very low allowing mutations to drift almost randomly, indicate a mutation rate for this species of $\sim 10^{-3}$ per genome per generation (round of chromosome replication; Foster et al. 2015). These results show that the supply of adaptive mutations is high (1% of all mutations in the experiment above), but many of the adaptive mutations of small effect may be lost by clonal interference.

Kimura presented the cost of adaptation under hard sweeps, where genetic diversity is swept away because of the rapid rise in frequency of the fittest clone, as a key argument in favor of the neutral theory. The cost of adaptation is most dramatic under complete and rapid selective sweeps, because fixation of a variant leads to the extinction of all the others. If many mutations were adaptive and these sweeps frequent, populations should have very low genetic diversity, which is known to be false (Leffler et al. 2012). Studies of *E. coli*

evolving in the mouse gut, a more natural environment that those of the works mentioned above, confirmed the existence of frequent adaptive mutations that were subject to clonal interference. Surprisingly, the analysis of the evolved populations showed rapid complete phenotypic switches without loss of genetic diversity (Barroso-Batista et al. 2014; Lourenco et al. 2016). While the mechanisms operating in these experiments remain to be elucidated, they seem to involve mutation hotspots, facilitating the recurrence of some mutations, and the spread of adaptive changes through recombination or HGT across the population. As mentioned above, the observation that natural populations were highly polymorphic was at the basis of the proposal of the neutral theory because hard sweeps dramatically reduce the efficiency of natural selection. However, soft sweeps—processes where adaptive variants rise in frequency in the population without sweeping away all of its genetic diversity—are much less costly because they don't eliminate the other adaptive variants from the population (Messer and Petrov 2013). They might explain why many bacteria adapt rapidly without dramatic loss of genetic diversity.

It had become clear, before the neutral theory was proposed, that nucleotide composition is extremely variable between microbial genomes (Sueoka 1961), where it determines the average amino acid composition of proteins (Sueoka 1962; King and Jukes 1969). It seems highly unlikely that positive selection would allow the fixation of certain amino acids at the level of the entire genome in function of its nucleotide composition. First, because it's unclear what might be the biological origin of such selection pressure. Second, because fixation of each individual mutation would lead to the removal of genetic diversity and would be slowed-down by intense clonal interference. The proposal that protein composition is a direct consequence of the mutation spectra was thus taken as a demonstration of the neutral theory. Yet, recent works show that the composition of bacterial genomes deviates from the mutational patterns, systematically favoring G and C polymorphism (fig. 1C; Hershberg and Petrov 2010; Hildebrand et al. 2010). These observations revived the interest on the many nonneutral theories proposed to explain genome composition (Rocha and Feil 2010). The current controversy centers around the roles of natural selection and a nonselective process—called biased gene conversion—where allelic recombination favors Gs and Cs over As and Ts when there are mismatches between the two sequences (Raghavan et al. 2012; Lassalle et al. 2015). The controversy is fed by conflicting evidence for an association between allelic recombination rates and bias towards G + C rich polymorphism (Hildebrand et al. 2010; Lassalle et al. 2015; Bobay and Ochman 2017b; see table 1 for details).

In summary, most mutations in bacteria are nonsynonymous and deleterious, many of the others are affected either by selection on codon usage (synonymous changes) or regulatory sequences (intergenic regions). All mutations are affected by neighboring regions under natural selection (strong linkage) and by the mechanism increasing the probability of fixation of Gs and Cs (over As and Ts). One is thus tempted to conclude that there is no single-nucleotide polymorphism (SNP) evolving under random drift in bacteria.

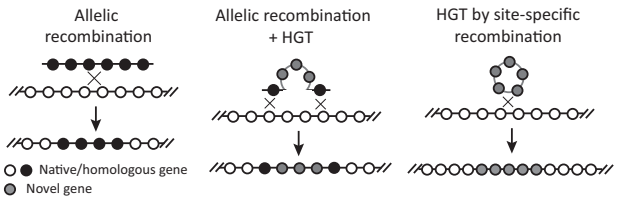

FIG. 2. Processes leading to allelic recombination and HGT in bacteria.

Is the Microbial Pan-Genome (Practically) Infinite?

"A theoretical treatment was presented which enables us to obtain the average number of heterozygous nucleotide sites per individual and related quantities that describe the statistical property of the mutant frequency distribution attained under steady flux of mutations in a finite population. The main assumptions of the model are that (i) a very large (practically infinite) number of sites are available for mutation and (ii) whenever a mutant appears, it represents a mutation at a new (different) site" (Kimura 1969).

Prokaryotes' sexual exchanges are independent of reproduction and can be divided into two processes (fig. 2). Allelic recombination results from gene conversion by homologous recombination between very similar sequences. Horizontal gene transfer (HGT) results from the acquisition of novel genes that can remain extra-chromosomal or integrate the chromosome by homologous or site-specific recombination. The substrate for allelic recombination and HGT, DNA from other individuals, can be acquired by natural transformation or, more frequently, by the action of mobile genetic elements (MGEs, conjugation, or phage-mediated). This novel genetic information can be integrated either by homologous recombination with flanking homologous DNA or by the action of specific recombinases. Natural transformation, ascertained in only half a hundred species, is under the genetic control of the recipient cell, opening the possibility of natural selection on modifiers of its rate and timing. Both are known to vary between species (Johnston et al. 2014). In contrast, most bacterial genomes contain prophages or conjugative elements (Guglielmini et al. 2011; Touchon et al. 2016). Contrary to transformation, MGE-mediated HGT is not under the genetic control of the recipient host. For example, a thorough study in *E. coli* K12 found no single nonessential gene playing an important role in the acquisition of plasmid R388 by conjugation (Perez-Mendoza and de la Cruz 2009). The evolution of the rates of transfer by MGEs takes place predominantly in the element itself. Some of these elements, phages, frequently kill the recipient cell. Hence, acquisition of novel genes by HGT is the result of a complex interaction between bacteria and their MGEs, in which the latter can be symbionts, parasites, or both (depending on the circumstances).

HGT drives intraspecies phenotypic diversification, including antibiotic resistance (Davies and Davies 2010) and virulence

(Wagner and Waldor 2002; Dobrindt et al. 2004). Its impact cannot be ignored. In *Bacillus cereus*, one of the few species for which data is available, each point mutation in the genome is accompanied, on average, by 0.7 SNPs introduced by allelic recombination (Vos and Didelot 2009), and 4.4 genes by HGT (>4,000 nucleotides; Hao and Golding 2006). Many bacterial species have pan-genomes—the set of all nonorthologous gene families in the species—much larger than the average genome size. The description of natural variation and phenotypic adaptation of many bacteria, two key goals of population genetics, must account for HGT because the latter drives their genetic diversification.

The theoretical and practical problems that HGT raises in population genetics have not been explored enough. A key challenge is how to include gain and loss of genetic information in measures of genetic diversity and in neutrality tests that usually focus on SNPs. With this goal in mind, an Infinitely Many Genes model (IMG) has been recently proposed (Baumdicker et al. 2012). Just as the infinitely many sites model of Kimura assumes that the number of sites available for mutation is so large relative to mutation rate that a site only mutates once (Kimura 1969), the IMG model assumes that the supply of novel genes can be considered as (practically) unlimited. A coalescent framework is then used to model gains and losses of accessory genes on the genealogy of a population sample (Baumdicker et al. 2010). This model naturally produces a test for neutrality. Its application to the pan-genomes of two cyanobacterial species with very large census population sizes, and presumably efficient selection, showed marginal rejection of neutrality for one and no rejection for the other (Baumdicker et al. 2012). A development of this approach models the distribution of frequencies of accessory genes using two classes, instead of just one in the original IMG model. This provides a significantly better fit to the distribution of gene frequencies in Firmicutes (Collins and Higgs 2012), suggesting the existence of two classes of accessory genes that differ in terms of turnover rates. At this stage, it is unclear if the genes with fastest turnover—those that tend to be present in very few strains—are neutral, deleterious, or a mixture of both. What is clear is that these results provide ground for another neutralist–selectionist debate (Shapiro 2017), with researchers claiming transfer to be mostly deleterious (Vos et al. 2015), neutral (Baumdicker et al. 2012), or adaptive (Sela et al. 2016; McInerney et al. 2017; table 1). Settling this debate will require extensive experimental work since the fitness effect of these genes depends on the environmental conditions.

The distribution of frequencies of accessory genes is not far from the neutral expectation. This suggests that many genes acquired by HGT are never expressed and thus devoid of fitness effects. However, the compatibility of a distribution with a neutral model does not necessarily implicate that the underlying mechanism is indeed random drift. Natural selection under strong genetic linkage can mimic many of the expectations of neutral models (see below). We do know that at least some of these genes are adaptive whereas others are deleterious. The fundamental question is: How many? The concomitant operational issue is: Can we identify them?

How to Do Population Genetics with Ill-Defined Populations?

"I would like to call the readers' attention to the fact that the effective population size applicable to neutral evolution is that of the entire species (or subspecies if this forms an independent reproductive unit), but not the effective size of a local population which forms a part of it" (Kimura 1983).

It all comes to sex. If there are no sexual exchanges, one can loosely define a species as a cluster of closely related strains with the same ecological niche (ecotype; Cohan and Perry 2007). Whether such ensembles have biological relevance depends on the microbiologist's ability to detect and delimit their niche. In this case, mutations are tightly linked to their genetic background and selection tends to be inefficient (Maruyama and Kimura 1980). If sexual reproduction is delimited to a well-defined type of individuals then the biological definition of species is applicable and species can be defined as the maximal sets of interbreeding and sexually isolated populations (Mayr 1942), or as ecotypes (as for asexuals). The two definitions are not necessarily incompatible. The problem is that many organisms, including most Prokaryotes, exchange genes promiscuously. This renders species fuzzy and complicates the application of a lot of Kimura's most elegant results concerning the vagaries of polymorphisms in natural populations.

The debate on the definition of species in bacteria based on mechanisms of gene exchange has focused on allelic recombination for decades (Dykhuizen and Green 1991). The rate of homologous recombination with foreign DNA decreases exponentially with sequence divergence (fig. 1D; Vulic et al. 1997), and species could be defined by the genetic distances at which homologous recombination becomes unlikely (Gogarten et al. 2002). However, these barriers are fuzzy because there is no discontinuity in the frequency of homologous recombination in function of sequence divergence, and they are unsteady because recombination rates depend on the cell's genetic background (Rayssiguier et al. 1989), physiological state (Matic et al. 1997), and defense mechanisms (Oliveira et al. 2016). Simulations suggest that barriers to recombination cannot arise simply by a cohesive effect of within-species recombination (Cohan 1995; Fraser et al. 2009). As such, allelic recombination defines, at best, fuzzy bacterial species (Lawrence 2002).

The complications brought by HGT to the definition of species are much worse (Doolittle and Zhaxybayeva 2009). I've described above how HGT drives the diversification of bacterial gene repertoires. MGEs, the major vehicles of HGT, transfer between cells with little regard for taxonomic boundaries. For example, conjugative plasmids can transfer across species, genera or even large phyla (Amabile-Cuevas and Chicurel 1992). MGEs are also frequently lost. The high turnover of gene repertoires is in part a consequence of this flux of MGEs and implicates that relatedness between gene repertoires correlates poorly with phylogenetic distance (fig. 3A). It is not uncommon to find pairs of genomes from different species with more orthologs than pairs of genomes from the same species (Touchon et al. 2014). This has the effect of

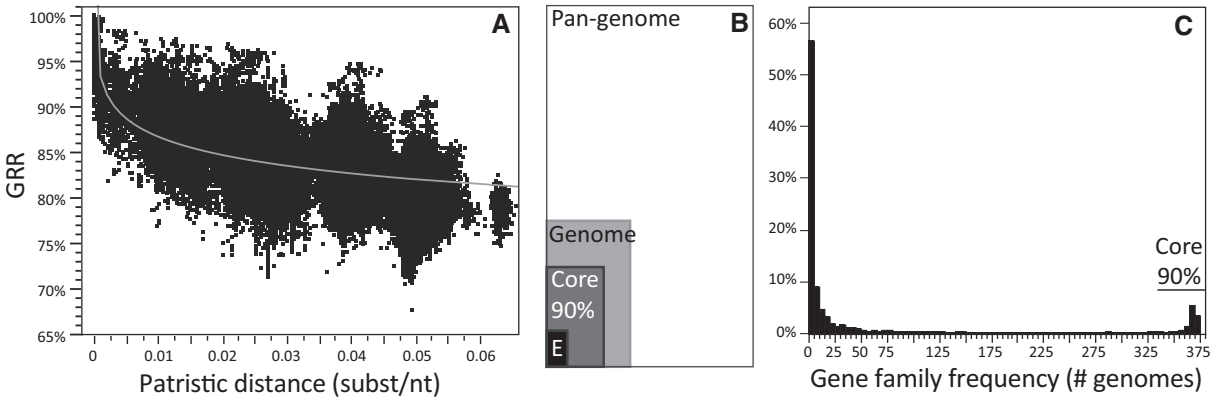

Fig. 3. Analysis of the diversity of *Escherichia coli*'s gene repertoires (update of Touchon et al. 2009). I took the genomes of *E. coli* from RefSeq in March 2018 (Tatusova et al. 2015), removed redundancy (all pairs of genomes are divergent >0.01%), and clustered the gene repertoires of the resulting 370 genomes (protein coding genes only) using MMseqs2 (Steinegger and Söding 2017). Proteins were placed in the same family of the pan-genome when they were >80% identical. (*A*) Gene repertoire relatedness (GRR, defined as the number of gene families present in both genomes divided by the number of gene families in the genome with fewer families) and patristic distance for all pairwise genome comparisons. The latter was computed from a tree built using IQtree v1.5.5 (Nguyen et al. 2015) with automatic choice of model (GTR + F + R8) on the alignment of 664 gene families present in all strains in a single copy (alignment built with MAFFT v7.310, Katoh and Standley 2013). The line is a logarithmic fit (GRR∼log(distance), $R^2 = 0.40$). (*B*) Number of gene families that are essential in rich medium, present in >90% of the genomes (Core 90%), present in the average genome, present in this pan-genome (31,057 families in total). Area is proportional to the number of genes in each group. (*C*) Distribution of frequencies of each gene family. Most families are present in one or very few genomes, some are present in almost all genomes (Core 90%), very few have intermediate frequencies.

blurring the associations between phenotypic and genetic distances (as inferred by SNPs in the core genome).

Most extant species definitions are based on predefined phenotypic markers whose ecological relevance is not always known. For example, the infamous *Bacillus anthracis* is a *B. cereus* clone that acquired a plasmid carrying the anthrax toxin (Okinaka et al. 1999). Other clones of *B. cereus* with other plasmids carrying other toxins were also defined as species because of the relevance of the associated diseases. As a consequence, the *B. cereus* complex is now composed of a dozen named species, many of which are polyphyletic (Bazinet 2017). More generally, a recent study identified intraspecific allelic recombination barriers in a fourth of prokaryotic species (Bobay and Ochman 2017a). The situation has become so confusing that even when ignoring the complications of ecological niches and sexual exchanges, it has been suggested that if one imposes monophyly and a range of sequence divergence to species then 73% of the existing taxonomy needs correction (Parks et al. 2018).

Measures of genetic diversity, estimations of effective population size, and inferences of selective sweeps are extremely sensitive to the definition of species or populations. This means that many traits associated with bacterial species are a direct consequence of the species delimitation. A species that is defined as a very recent clone with a specific trait, like *B. anthracis*, has almost necessarily little genetic diversity and thus a small Ne. This does not mean that we should expect to find in its genome signals of long-term inefficient selection, such as rampant pseudogenization. Conversely, a species defined as a broad collection of phenotypically very diverse strains, like *E. coli*, has high genetic diversity, thus presumably large Ne, even if certain of its lineages contain significant numbers of pseudogenes. Translation of measures of diversity into expected molecular evolution traits must thus be made with care.

Recently, it has become frequent to define species as operational taxonomic units with predefined maximal pairwise diversity (Konstantinidis et al. 2006). This definition has the merits of being clear, precise, and easy to compute. However, it creates a bias towards species with excessively homogeneous values of genetic diversity, obscuring the biological significance of this variable.

It is not yet clear what is the most meaningful way of defining species in bacteria (nor if there is one). The definition of species based on sexual isolation becomes inapplicable when HGT is a relevant mechanism of diversification and adaptation. But the use of ecotypes to define species is extremely complex, and possibly just as fuzzy in many situations. At this stage, the difficulties in defining species, or populations, demands creative thinking from the population geneticist and careful consideration of the underlying hypothesis of existing methodologies.

How to Untangle Demography and Selection?

"If a species consists of a large number of more or less isolated subpopulations, and if extinction and subsequent replacement occur frequently among subpopulations. The effective size of the species is greatly reduced as compared with the situation in which the whole species forms a random mating or panmictic unit. This property is particularly pertinent when we consider genetic variability of a bacterial species" (Kimura 1983).

The efficiency of natural selection is heavily influenced by demography (Wright 1950). Accordingly, the allele frequency

spectrum is affected by changes in population size complicating the identification of alleles under selection (Vitti et al. 2013). It was proposed, following on the footsteps of the neutral theory, that demography affects the entire genome of the species, whereas selection affects individual loci and specific alleles (Lewontin and Krakauer 1973). Hence, many studies disentangle demography from selection by inferring a demographic model from regions thought to evolve neutrally and then using it to identify the alleles under selection as those deviating from the expectations of the demographic model. Whereas such approaches have been faced with significant criticisms (Hahn 2008), they remain popular in population genetics. For the reasons mentioned in the first section of this article, these methods are inapplicable to bacteria. There are probably no locations in the bacterial genome evolving neutrally and even if there were, they would be in linkage disequilibrium with most of the genome (see below). Given the extent and relevance of HGT, the assumption that demography affects similarly all the genome is false.

Many other standard methods to identify regions under positive selection are also poorly applicable to bacteria. In Eukaryotes (diploids), chromosome crossover gradually narrows the region under genetic linkage and measure of linkage-disequilibrium can be used to detect selective sweeps (Sabeti et al. 2006). In bacteria, allelic recombination introduces small patches of homologous sequences leaving distant regions of the genome in the same clonal frame under strong genetic linkage (fig. 4; Milkman and Bridges 1990). High linkage means that mutations often segregate with many others of different, eventually opposed, selective effects. They may hitchhike with adaptive mutations (Maynard Smith and Haig 1974), or be purged because of a background of deleterious mutations (Charlesworth et al. 1993). Identification of changes under positive selection in bacteria has thus focused on techniques derived from the McDonald–Kreitman and dN/dS ratio test, even if such analyses can also be affected by allelic recombination, demography, selection for codon usage bias (Nielsen 2005), and the difficulty of defining fixed substitutions when populations are intrinsically difficult to delimit.

The characterization of demographic changes in bacterial populations would be of great interest in itself. For example, they could inform epidemiological studies as is commonly done in virology (Grenfell et al. 2004). Unfortunately, methods aiming at identifying demographic changes from population polymorphism assume neutrality and are not robust to the presence of selection nor, sometimes, recombination. This is because both processes can distort genealogies in ways resembling demographic changes (see Lapierre et al. 2016 and citations therein). Untangling selection from demography under these conditions is an open, and urgent, problem in bacterial population genomics.

This problem is further amplified by nonrandom sampling, an almost ubiquitous trait in the practice of microbiologists. Most large data sets in bacterial population genetics are associated with bacteria of medical interest. Sampling is usually focused on virulence or antibiotic resistance, which leads to an overrepresentation of a small number of subpopulations. This may have immediate medical interest, for example, pinpoint a virulence factor or a determinant of antibiotic resistance. However, nonrandom sampling affects the inference of most population parameters of interest, notably those associated with demography (Lapierre et al. 2016). Quick fixes, like adding a few other strains chosen uniformly from the species phylogeny can make things worse. It is also important to notice that to really understand the emergence of hypervirulent clones it is important to go beyond the exhaustive study of these strains and move towards the understanding of the evolution of other lineages that have not (yet) evolved in that way. In the long run, excessive focus on strains of medical interest is counterproductive not only to study bacterial population genetics in general but also to understand the emergence of clones of clinical interest.

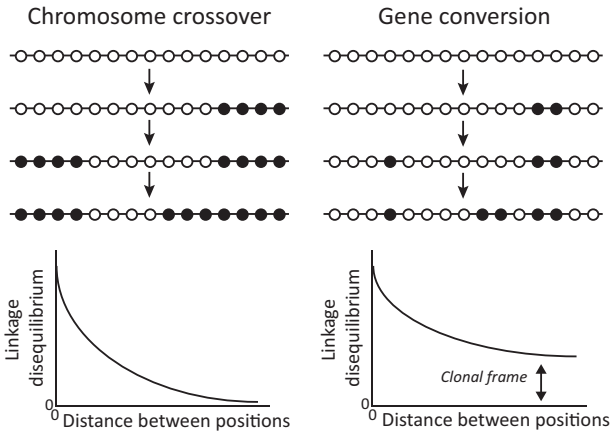

FIG. 4. Effects of recombination on linkage. Chromosome crossover leads to a rapid loss of linkage in Eukaryotes for distant positions in the chromosome. In contrast, gene conversion with exogenous DNA usually involves a small number of contiguous genes and leads to strong linkage between distant regions in the chromosome.

Conclusion

To summarize the complications of bacterial population genetics, let us consider the classical bacterial model: *E. coli*. Its cell machinery has evolved to attain very rapid growth, with doubling times as low as 20 min. Under unlimited resources and in the absence of death, such a cell could produce $> 10^{21}$ descendants in one day (72 generations). In spite of >350,000 scientific articles citing *E. coli* in PubMed, there is very little available data on the average growth rate of this bacteria in nature. Until recently, the reference value was less than one generation per day in the gut (Savageau 1983), but recent data suggest much higher numbers: between 6 and 21 generations per day (Myhrvold et al. 2015; Ghalayini et al. 2018). Replication and survival rates in the environment are poorly known. The gap between potential and effective number of generations per day is explained by suboptimal growth conditions, competition, and predation (Jang et al. 2017). Nevertheless, there are more *E. coli* individuals, a minor component of our gut microbiome, in the average human individual than humans on the planet (Tenaillon et al. 2010). Considering its presence in many other domestic and wild animals, its census population size is certainly several orders of magnitude above 10^{20}. Yet, past estimates of *Ne* ranged from

10^5 to 10^8 (Hartl et al. 1994; Berg 1996; Charlesworth and Eyre-Walker 2006; Sung et al. 2012), leaving selection and demography ample room to shape the evolutionary history of the species.

The *E. coli* species itself is ill-defined. It includes several polyphyletic lineages that together make an entire genus (*Shigella* Pupo et al. 2000), whereas highly divergent strains have "typical" *E. coli* phenotypes (even if more important environmental reservoirs; Walk et al. 2009). *Escherichia coli* core genes have an average diversity of ~0.36 (Walk et al. 2007), which is not very high for the standards of Eukaryotes (Leffler et al. 2012), in spite of *E. coli*'s huge population size. Allelic recombination is moderate (its contribution to SNPs matches that of mutation) but novel alleles often come from outside the species (Didelot et al. 2012). The concomitant strong linkage between mutations in the clonal frame (most of the chromosome) favors genetic draft—a process where changes in allele frequencies are largely due to their association with genetic backgrounds (Gillespie 2001)—caused by acquisitions of genetic material with fitness effects. The pangenome is huge and strains can differ by a fourth of their gene repertoires (fig. 3*A* and *B*). Only around 300 genes are essential in *E. coli* (Baba et al. 2006), which is much less than the number of genes in the core genome (fig. 3B). Furthermore, systematic gene deletion studies show that many genes acquired by HGT provide fitness advantages (Karcagi et al. 2016). The immense diversity of gene repertoires in the species means that no strain is a good representative of the species, and that population genetics is essential to understand the evolution of its numerous pathotypes. To the best of current knowledge, the situation is similar for many other bacteria, which tend to show a U-shaped distribution of gene families frequencies, where most families are either present in most strains or in very few (fig. 3C).

Motoo Kimura was known for his original mathematical treatment of population genetics that solved many outstanding problems in the domain. The contemporary observations that populations were much more polymorphic than previously expected led him naturally to the development of the neutral theory. I would argue that given the current knowledge, it has become unreasonable to assume that most SNPs drift randomly in bacterial populations. The neutral theory may remain relevant as a null hypothesis, especially to describe the patterns of gene acquisition and loss, but evidence suggests that the vast majority of polymorphism has weak fitness effects as proposed by the Nearly Neutral Theory (Ohta and Kimura 1971).

It has been pointed out that if drift is replaced by draft in the Nearly Neutral Theory, then the rate of substitutions no longer depends on population size for large enough populations (Gillespie 2001). This revision of the theory holds great promises for the study of the large bacterial populations where Ne is hard to infer, due to the difficulty in defining meaningful populations, and even harder to interpret. If sweeps are frequent in bacteria, then they should lead to bursts of rapid coalescent between lineages (Neher 2013). Multiple merger coalescent models might thus constitute a promising research avenue to produce more realistic models of microbial populations (Tellier and Lemaire 2014). For this research program to be successful, one also needs further theoretical work on the specifics of bacteria in terms of genetic exchanges, population delimitation, and genome structure, in the context of pervasive selection. In a time of almost unlimited microbial DNA sequence data, and metagenomics data of entire communities, the marriage of novel models in population genetics with molecular evolution is likely to be a very rewarding one.

Acknowledgments

I thank Jorge Moura de Sousa, Louis-Marie Bobay, Guillaume Achaz, and two anonymous reviewers for comments and suggestions on an earlier version of this text. Work in my lab is supported by the CNRS and the Institut Pasteur.

References

Amabile-Cuevas CF, Chicurel ME. 1992. Bacterial plasmids and gene flux. *Cell* 70(2):189–199.

Andersson SGE, Kurland CG. 1998. Reductive evolution of resident genomes. *Trends Microbiol.* 6:263–268.

Andreani NA, Hesse E, Vos M. 2017. Prokaryote genome fluidity is dependent on effective population size. *ISME J.* 11(7):1719.

Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol.* 2:0008.

Barrick JE, Lenski RE. 2013. Genome dynamics during experimental evolution. *Nat Rev Genet.* 14(12):827–839.

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

Barroso-Batista J, Sousa A, Lourenco M, Bergman ML, Sobral D, Demengeot J, Xavier KB, Gordo I. 2014. The first steps of adaptation of *Escherichia coli* to the gut are dominated by soft sweeps. *PLoS Genet.* 10(3):e1004182.

Batut B, Knibbe C, Marais G, Daubin V. 2014. Reductive genome evolution at both ends of the bacterial population size spectrum. *Nat Rev Microbiol.* 12(12):841–850.

Baumdicker F, Hess WR, Pfaffelhuber P. 2010. The diversity of a distributed genome in bacterial populations. *Ann Appl Prob.* 20:1567–1606.

Baumdicker F, Hess WR, Pfaffelhuber P. 2012. The infinitely many genes model for the distributed genome of bacteria. *Genome Biol Evol.* 4(4):443–456.

Bazinet AL. 2017. Pan-genome and phylogeny of *Bacillus cereus sensu lato*. *BMC Evol Biol.* 17(1):176.

Beaber JW, Hochhut B, Waldor MK. 2004. SOS response promotes horizontal dissemination of antibiotic resistance genes. *Nature* 427(6969):72–74.

Bentley SD, Parkhill J. 2004. Comparative genomic structure of prokaryotes. *Annu Rev Genet.* 38:771–791.

Berg OG. 1996. Selection intensity for codon bias and the effective population size of *Escherichia coli*. *Genetics* 142(4):1379–1382.

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

Bobay L-M, Ochman H. 2017b. Impact of recombination on the base composition of Bacteria and Archaea. *Mol Biol Evol.* 34(10):2627–2636.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134(4):1289–1303.

Charlesworth J, Eyre-Walker A. 2006. The rate of adaptive evolution in enteric bacteria. *Mol Biol Evol.* 23(7):1348–1356.

Cohan FM. 1995. Does recombination constrain neutral divergence among bacterial taxa? *Evolution* 164–175.

Cohan FM, Perry EB. 2007. A systematics for discovering the fundamental units of bacterial diversity. *Curr Biol.* 17(10):R373–R386.
Collins RE, Higgs PG. 2012. Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. *Mol Biol Evol.* 29(11):3413–3425.
Cordero OX, Polz MF. 2014. Explaining microbial genomic diversity in light of evolutionary ecology. *Nat Rev Microbiol.* 12(4):263–273.
Crow JF. 1995. Motoo Kimura (1924–1994). *Genetics* 140(1):1–5.
Davies J, Davies D. 2010. Origins and evolution of antibiotic resistance. *Microbiol Mol Biol Rev.* 74(3):417–433.
Delaney NF, Balenger S, Bonneaud C, Marx CJ, Hill GE, Ferguson-Noel N, Tsai P, Rodrigo A, Edwards SV. 2012. Ultrafast evolution and loss of CRISPRs following a host shift in a novel wildlife pathogen, *Mycoplasma gallisepticum*. *PLoS Genet.* 8:e1002511.
Didelot X, Meric G, Falush D, Darling AE. 2012. Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. *BMC Genomics.* 13:256.
Dobrindt U, Hochhut B, Hentschel U, Hacker J. 2004. Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol.* 2(5):414–424.
Doolittle WF, Zhaxybayeva O. 2009. On the origin of prokaryotic species. *Genome Res.* 19(5):744–756.
Dykhuizen DE, Green L. 1991. Recombination in *Escherichia coli* and the definition of biological species. *J Bacteriol.* 173(22):7257–7268.
Ford CB, Lin PL, Chase MR, Shah RR, Iartchouk O, Galagan J, Mohaideen N, Ioerger TR, Sacchettini JC, Lipsitch M. 2011. Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. *Nat Genet.* 43(5):482–486.
Foster PL, Lee H, Popodi E, Townes JP, Tang H. 2015. Determinants of spontaneous mutation in the bacterium *Escherichia coli* as revealed by whole-genome sequencing. *Proc Natl Acad Sci U S A.* 112:E5990–E5999.
Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. 2009. The bacterial species challenge: making sense of genetic and ecological diversity. *Science* 323(5915):741–746.
Ghalayini M, Launay A, Bridier-Nahmias A, Clermont O, Denamur E, Lescat M, Tenaillon O. 2018. Evolution of a dominant natural isolate of *Escherichia coli* in the human gut over a year suggests a neutral evolution with reduced effective population size. *Appl Environ Microbiol* AEM.02377-02317 doi: 10.1128/AEM.02377-17.
Gillespie JH. 2001. Is the population size of a species relevant to its evolution? *Evolution* 55(11):2161–2169.
Giovannoni SJ, Cameron Thrash J, Temperton B. 2014. Implications of streamlining theory for microbial ecology. *ISME J.* 8(8):1553–1565.
Gogarten JP, Doolittle WF, Lawrence JG. 2002. Prokaryotic evolution in light of gene transfer. *Mol Biol Evol.* 19(12):2226–2238.
Gordo I, Perfeito L, Sousa A. 2011. Fitness effects of mutations in bacteria. *J Mol Microbiol Biotechnol.* 21(1–2):20–35.
Gouy M, Gautier C. 1982. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.* 10(22):7055–7074.
Grenfell BT, Pybus OG, Gog JR, Wood JL, Daly JM, Mumford JA, Holmes EC. 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 303(5656):327–332.
Guglielmini J, Quintais L, Pilar Garcillan-Barcia M, de la Cruz F, Rocha EPC. 2011. The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS Genet.* 7:e1002222.
Hahn MW. 2008. Toward a selection theory of molecular evolution. *Evolution* 62(2):255–265.
Hao W, Golding GB. 2006. The fate of laterally transferred genes: life in the fast lane to adaptation or death. *Genome Res.* 16(5):636–643.
Hartl DL, Moriyama EN, Sawyer SA. 1994. Selection intensity for codon bias. *Genetics* 138(1):227–234.
Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet.* 6(9):e1001115.
Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet.* 6(9):e1001107.
Jang J, Hur HG, Sadowsky MJ, Byappanahalli MN, Yan T, Ishii S. 2017. Environmental *Escherichia coli*: ecology and public health implications – a review. *J Appl Microbiol.* 123(3):570–581.
Johnston C, Martin B, Fichant G, Polard P, Claverys JP. 2014. Bacterial transformation: distribution, shared mechanisms and divergent control. *Nat Rev Microbiol.* 12(3):181–196.
Karcagi I, Draskovits G, Umenhoffer K, Fekete G, Kovács K, Méhi O, Balikó G, Szappanos B, Györfy Z, Fehér T, et al. 2016. Indispensability of horizontally transferred genes and its impact on bacterial genome streamlining. *Mol Biol Evol.* 33(5):1257–1269.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30(4):772–780.
Kimura M. 1967. On the evolutionary adjustment of spontaneous mutation rates. *Genet Res.* 9:23–34.
Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* 217(5129):624–626.
Kimura M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* 61:893–903.
Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.
Kimura M, Ohta T. 1974. On some principles governing molecular evolution. *Proc Natl Acad Sci U S A.* 71(7):2848–2852.
King JL, Jukes TH. 1969. Non-Darwinian evolution. *Science* 164(3881):788–798.
Konstantinidis KT, Ramette A, Tiedje JM. 2006. The bacterial species definition in the genomic era. *Philos Trans R Soc Lond B Biol Sci.* 361(1475):1929–1940.
Kuo CH, Moran NA, Ochman H. 2009. The consequences of genetic drift for bacterial genome complexity. *Genome Res.* 19:1450–1454.
Lapierre M, Blin C, Lambert A, Achaz G, Rocha EP. 2016. The impact of selection, gene conversion, and biased sampling on the assessment of microbial demography. *Mol Biol Evol.* 33(7):1711–1725.
Lassalle F, Perian S, Bataillon T, Nesme X, Duret L, Daubin V. 2015. GC-content evolution in bacterial genomes: the biased gene conversion hypothesis expands. *PLoS Genet.* 11(2):e1004941.
Lawrence JG. 2002. Gene transfer in bacteria: speciation without species? *Theor Popul Biol.* 61(4):449–460.
Leffler EM, Bullaughey K, Matute DR, Meyer WK, Segurel L, Venkat A, Andolfatto P, Przeworski M. 2012. Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol.* 10(9):e1001388.
Lewontin R, Krakauer J. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 74(1):175–195.
Lourenco M, Ramiro RS, Guleresi D, Barroso-Batista J, Xavier KB, Gordo I, Sousa A. 2016. A mutational hotspot and strong selection contribute to the order of mutations selected for during *Escherichia coli* adaptation to the gut. *PLoS Genet.* 12(11):e1006420.
Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, Foster PL. 2016. Genetic drift, selection and the evolution of the mutation rate. *Nat Rev Genet.* 17(11):704–714.
Maruyama T, Kimura M. 1980. Genetic variability and effective population size when local extinction and recolonization of subpopulations are frequent. *Proc Natl Acad Sci U S A.* 77(11):6710–6714.
Matic I, Radman M, Taddei F, Picard B, Doit C, Bingen E, Denamur E, Elion J. 1997. Highly variable mutation rates in commensal and pathogenic *Escherichia coli*. *Science* 277(5333):1833–1834.
Maynard Smith J, Haig J. 1974. The hitch-hiking effect of a favourable gene. *Genet Res.* 23:23–35.
Mayr E. 1942. Systematics and the origin of species, from the viewpoint of a zoologist. Cambridge (MA): Harvard University Press.
McInerney JO, McNally A, O'Connell MJ. 2017. Why prokaryotes have pangenomes. *Nat Microbiol.* 2:17040.
Messer PW, Petrov DA. 2013. Population genomics of rapid adaptation by soft selective sweeps. *Tree* 28:659–669.
Milkman R, Bridges MM. 1990. Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. *Genetics* 126(3):505–517.
Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* 17(10):589–596.

Moran NA. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc Natl Acad Sci U S A.* 93(7):2873–2878.
Morris JJ, Lenski RE, Zinser ER. 2012. The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *MBio* 3:e00036–00012.
Moxon ER, Rainey PB, Nowak MA, Lenski RE. 1994. Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr Biol.* 4(1):24–33.
Myhrvold C, Kotula JW, Hicks WM, Conway NJ, Silver PA. 2015. A distributed cell division counter reveals growth dynamics in the gut microbiota. *Nat Commun.* 6:10039.
Neher RA. 2013. Genetic draft, selective interference, and population genetics of rapid adaptation. *Annu Rev Ecol Syst.* 44:195–215.
Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol.* 3(5):418–426.
Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* 32(1):268–274.
Nielsen R. 2005. Molecular signatures of natural selection. *Annu Rev Genet.* 39:197–218.
Ohta T. 1992. The nearly neutral theory of molecular evolution. *Annu Rev Ecol Syst.* 23:263–286.
Ohta T, Kimura M. 1971. On the constancy of the evolutionary rate of cistrons. *J Mol Evol.* 1(1):18–25.
Okinaka R, Cloud K, Hampton O, Hoffmaster A, Hill K, Keim P, Koehler T, Lamke G, Kumano S, Manter D, et al. 1999. Sequence, assembly and analysis of pX01 and pX02. *J Appl Microbiol.* 87(2):261–262.
Oliveira PH, Touchon M, Rocha EP. 2016. Regulation of genetic flux between bacteria by restriction-modification systems. *Proc Natl Acad Sci U S A.* 113(20):5658–5663.
Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P. 2018. A proposal for a standardized bacterial taxonomy based on genome phylogeny. *bioRxiv* 256800. doi: http://dx.doi.org/10.1101/256800.
Perez-Mendoza D, de la Cruz F. 2009. *Escherichia coli* genes affecting recipient ability in plasmid conjugation: are there any? *BMC Genomics.* 10:71.
Perfeito L, Fernandes L, Mota C, Gordo I. 2007. Adaptive mutations in bacteria: high rate and small effects. *Science* 317(5839):813–815.
Price MN, Arkin AP, Alm EJ. 2006. The life-cycle of operons. *PLoS Genet.* 2(6):e96.
Pupo GM, Lan R, Reeves PR. 2000. Multiple independent origins of Shigella clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc Natl Acad Sci U S A.* 97(19):10567–10572.
Raghavan R, Kelkar YD, Ochman H. 2012. A selective force favoring increased G+C content in bacterial genes. *Proc Natl Acad Sci U S A.* 109(36):14504–14507.
Rayssiguier C, Thaler DS, Radman M. 1989. The barrier to recombination between *E. coli* and *S. typhimurium* is disrupted in mismatch-repair mutants. *Nature* 342(6248):396–401.
Rocha E, Danchin A. 2002. Base composition bias might result from competition for metabolic resources. *Trends Genet.* 18(6):291–294.
Rocha E, Smith J, Hurst L, Holden M, Cooper J, Smith N, Feil E. 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J Theor Biol.* 239(2):226–235.
Rocha EPC. 2008. The organization of the bacterial genome. *Annu Rev Genet.* 42:211–233.
Rocha EPC, Feil EJ. 2010. Mutational patterns cannot explain genome composition: are there any neutral sites in the genomes of bacteria? *PLoS Genet.* 6: e1001104.
Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. 2006. Positive natural selection in the human lineage. *Science* 312(5780):1614–1620.
Savageau MA. 1983. *Escherichia coli* habitats, cell types, and molecular mechanisms of gene control. *Am Nat.* 122:732–744.
Sela I, Wolf YI, Koonin EV. 2016. Theory of prokaryotic genome evolution. *Proc Natl Acad Sci U S A.* 113(41):11399–11407.
Shapiro BJ. 2017. The population genetics of pangenomes. *Nat Microbiol.* 2(12):1574.
Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. 2005. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.* 33(4):1141–1153.
Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol.* 35(11):1026.
Sueoka N. 1961. Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. *Proc Natl Acad Sci U S A.* 47(8):1141–1149.
Sueoka N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. *Proc Natl Acad Sci U S A.* 48:582–591.
Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. 2012. Drift-barrier hypothesis and mutation-rate evolution. *Proc Natl Acad Sci U S A.* 109(45):18488–18492.
Tatusova T, Ciufo S, Federhen S, Fedorov B, McVeigh R, O'Neill K, Tolstoy I, Zaslavsky L. 2015. Update on RefSeq microbial genomes resources. *Nucleic Acids Res.* 43:D599–D605.
Tellier A, Lemaire C. 2014. Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. *Mol Ecol.* 23(11):2637–2652.
Tenaillon O, Skurnik D, Picard B, Denamur E. 2010. The population genetics of commensal *Escherichia coli*. *Nat Rev Microbiol.* 8(3):207–217.
Tenaillon O, Taddei F, Radmian M, Matic I. 2001. Second-order selection in bacterial evolution: selection acting on mutation and recombination rates in the course of adaptation. *Res Microbiol.* 152(1):11–16.
Thorpe HA, Bayliss SC, Hurst LD, Feil EJ. 2017. Comparative analyses of selection operating on nontranslated intergenic regions of diverse bacterial species. *Genetics* 206(1):363–376.
Touchon M, Bernheim A, Rocha EP. 2016. Genetic and life-history traits associated with the distribution of prophages in bacteria. *ISME J.* 10(11):2744–2754.
Touchon M, Cury J, Yoon E-J, Krizova L, Cerqueira GC, Murphy C, Feldgarden M, Wortman J, Clermont D, Lambert T, et al. 2014. The genomic diversification of the whole acinetobacter genus: origins, mechanisms, and consequences. *Genome Biol Evol.* 6(10):2866–2882.
Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, et al. 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet.* 5(1):e1000344.
Vieira-Silva S, Rocha EPC. 2010. The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet.* 6:e1000808.
Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting natural selection in genomic data. *Annu Rev Genet.* 47:97–120.
Vos M, Didelot X. 2009. A comparison of homologous recombination rates in bacteria and archaea. *ISME J.* 3(2):199–208.
Vos M, Hesselman MC, te Beek TA, van Passel MW, Eyre-Walker A. 2015. Rates of lateral gene transfer in prokaryotes: high but why? *Trends Microbiol.* 23(10):598–605.
Vulic M, Dionisio F, Taddei F, Radman M. 1997. Molecular keys to speciation: dNA polymorphism and the control of genetic exchange in enterobacteria. *Proc Natl Acad Sci U S A.* 94(18):9763–9767.
Wagner PL, Waldor MK. 2002. Bacteriophage control of bacterial virulence. *Infect Immun.* 70(8):3985–3993.
Walk ST, Alm EW, Calhoun LM, Mladonicky JM, Whittam TS. 2007. Genetic diversity and population structure of *Escherichia coli* isolated from freshwater beaches. *Environ Microbiol.* 9(9):2274–2288.
Walk ST, Alm EW, Gordon DM, Ram JL, Toranzos GA, Tiedje JM, Whittam TS. 2009. Cryptic lineages of the genus *Escherichia*. *Appl Environ Microbiol.* 75(20):6534–6544.
Wright S. 1950. Genetical structure of populations. *Nature* 166(4215):247–249.