The Impact of Selection, Gene Conversion, and Biased Sampling on the Assessment of Microbial Demography

Marguerite Lapierre,[1,2] Camille Blin,[3,4,5] Amaury Lambert,[2,6] Guillaume Achaz,[1,2] and Eduardo P. C. Rocha*[,4,5]

[1]Atelier de Bioinformatique, UMR7205 ISYEB, MNHN-UPMC-CNRS-EPHE, Muséum National d'Histoire Naturelle, Paris, France

[2]Collège de France, Center for Interdisciplinary Research in Biology (CIRB), CNRS UMR 7241, Paris, France

[3]Sorbonne Universités, UPMC Univ Paris06, IFD, 4 Place Jussieu, Paris Cedex05, France

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

[5]CNRS, UMR3525, Paris, France

[6]UPMC Univ Paris 06, Laboratoire de Probabilités et Modèles Aléatoires (LPMA), CNRS UMR 7599, Paris, France

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

Associate editor: Helen Piontkivska

Abstract

Recent studies have linked demographic changes and epidemiological patterns in bacterial populations using coalescent-based approaches. We identified 26 studies using skyline plots and found that 21 inferred overall population expansion. This surprising result led us to analyze the impact of natural selection, recombination (gene conversion), and sampling biases on demographic inference using skyline plots and site frequency spectra (SFS). Forward simulations based on biologically relevant parameters from *Escherichia coli* populations showed that theoretical arguments on the detrimental impact of recombination and especially natural selection on the reconstructed genealogies cannot be ignored in practice. In fact, both processes systematically lead to spurious interpretations of population expansion in skyline plots (and in SFS for selection). Weak purifying selection, and especially positive selection, had important effects on skyline plots, showing patterns akin to those of population expansions. State-of-the-art techniques to remove recombination further amplified these biases. We simulated three common sampling biases in microbiological research: uniform, clustered, and mixed sampling. Alone, or together with recombination and selection, they further mislead demographic inferences producing almost any possible skyline shape or SFS. Interestingly, sampling sub-populations also affected skyline plots and SFS, because the coalescent rates of populations and their sub-populations had different distributions. This study suggests that extreme caution is needed to infer demographic changes solely based on reconstructed genealogies. We suggest that the development of novel sampling strategies and the joint analyzes of diverse population genetic methods are strictly necessary to estimate demographic changes in populations where selection, recombination, and biased sampling are present.

Key words: **bacteria, population size, natural selection, gene conversion, *Escherichia coli*, population genomics.**

Article

Introduction

Bacterial populations show extensive demographic variations across space and time (Martiny et al. 2006), such as frequent expansions and bottlenecks. The characterization of these demographic changes among populations of infectious agents provides epidemiological information that can guide public health interventions. A recent field of research, phylodynamics, aims at understanding the association between ecological processes and epidemiological patterns in an evolutionary framework (Grenfell et al. 2004). It integrates phylogenic inference and population genetics to study variations in demography through time (Grad and Lipsitch 2014; Li et al. 2014). Phylodynamics has been particularly useful to characterize transmission dynamics from sequence data, and could facilitate the evaluation of public health policies for diseases with low reporting rates (Volz et al. 2013).

Demographic changes imprint the reconstructed genealogies of the population, the so-called coalescent tree, by affecting the intervals of time between successive splits in the tree (Tajima 1989a). These values (coalescent rates) are proportional to the inverse of the effective population size (N_e) in the standard neutral model. If one takes two idealized populations with the same contemporary population size, then the one with a history of population expansion will have (on average) shorter branches throughout, including at the tips. However, the relative length of the tips compared with the internal branches will be longer than in a nonexpanding population. Since nodes in the reconstructed genealogy of the expanding population are more concentrated closer to the root of the tree, the site frequency spectrum (SFS), that is, the distribution of the frequencies of all nucleotide polymorphisms, shows an excess of alleles shared by few individuals (rare alleles) (Adams and Hudson 2004). Conversely,

populations with a history of population size contraction exhibit an excess of polymorphism shared by many individuals when compared with stable populations with the same contemporary population size. Their reconstructed genealogies have longer branches overall, but the average length of the tips compared with the internal branches are shorter than in a noncontracting population (coalescence rates are higher than expected closer to the present).

Under the assumptions of the standard neutral model (no population structure, random sampling, no recombination, no selection), it is often implicitly assumed that variations in N_e (or equivalently, variations in the coalescence rate) are indications of demographic changes. Parametric approaches were developed to infer these demographic changes under explicit models, such as the Approximate Bayesian Computation method (Beaumont et al. 2002) or the likelihood-based method (*e.g.*, Nielsen and Wakeley 2001; Drummond et al. 2002). In this context, skyline plots were introduced to quantify the relationship between the coalescence rate of the population and the genealogy of the sequences in a non-parametric approach, that is, without an explicit model to test. Coalescent rates can then be used to produce detailed demographic histories from sequence data assuming that all other assumptions of the neutral coalescent are met (Pybus et al. 2000; Drummond et al. 2005). Demographic trends can also be inferred using SFS-based neutrality tests (Fu 1997; Fu and Voordouw 1997; Ramos-Onsins and Rozas 2002; Achaz 2009). For example, Tajima's *D* measures the difference between the mean number of pairwise differences and the number of segregating sites, and is skewed to negative values in case of population expansion (Tajima 1989b). SFS-based model-flexible methods (*i.e.*, exploring the space of possible demographic models) have also been recently proposed (Liu and Fu 2015). They approximate the demography using piecewise constant population sizes.

Violations of the assumptions of the neutral coalescent, such as presence of recombination or selection, may affect reconstructed genealogies and SFS in ways resembling demography (*e.g.*, Schierup and Hein 2000; Nielsen and Beaumont 2009; Mazet et al. 2015). Recombination by gene conversion has a very moderate effect on the topology of phylogenetic trees (Touchon et al. 2009), but affects skyline models (Hedge and Wilson 2014). Removing sites incompatible with the tree topology, that is, homoplasies, actually aggravates the effect of recombination in skyline models, presumably because it preferentially removes polymorphisms in deeper branches of the tree (Hedge and Wilson 2014). Recombination in the absence of selection has actually little effect on the expected SFS, apart from decreasing its variance (Wall 1999). The effect of selection on skyline plots has been less studied. Strong purifying selection is not expected to affect drastically the SFS because the deleterious mutations are quickly purged (Kimura 1983). On the other hand, mild purifying selection or recent selective sweeps lead to an excess of recent polymorphism, creating the impression of recent population expansion (Braverman et al. 1995). Diversifying or balancing selection can produce more complex patterns (Navarro and Barton 2002). Some studies have found that deleterious mutations of mild effect have a negligible effect on the time back to the most recent common ancestor (TMRCA) (Neuhauser and Krone 1997), and very little effect on the shape of the reconstructed genealogies (Przeworski et al. 1999) even though linkage between sites may affect the distribution of mutations (Williamson and Orive 2002). Mutations of mild deleterious effect are abundant in some bacteria (Hughes 2005; Balbi et al. 2009). If bacterial evolution is dominated by these mutations then selection might not strongly affect demographic inference using skyline plots. However, recent studies have suggested that weak purifying selection, when occurring at multiple sites, could affect the shape of the coalescent tree (O'Fallon et al. 2010). The effect of selection on skyline plots remains unclear.

The possibility of producing large sequence datasets for microbial populations has spurred interest on the use of these methods to study microbial demography. The skyline plot has been particularly popular because it allows precisely detailing demographic changes (Ho and Shapiro 2011). This method was initially used to study RNA viruses, which exhibit low recombination rates between individuals in different hosts and small effective population sizes (Holmes 2007). These viruses also have very high mutation rates, which increases mutational load and decreases the efficiency of selection (especially under no recombination) (Kimura 1983). Skyline plots have been increasingly used to study cellular microbes, most notably pathogenic bacteria. Yet, it is unclear if violations to the neutral coalescent model (biased sampling, selection, or recombination) can be safely ignored in these cases. Many bacterial populations are extremely large, show a very strong imprint of natural selection, endure rapid population fluctuations, exhibit low mutation rates, and recombine at high rates (Rocha et al. 2006; Vos and Didelot 2009; Tellier and Lemaire 2014). In fact, abundant evidence suggests that there are few, if any, positions evolving according to the neutral model in bacterial genomes (reviewed in Rocha and Feil 2010).

Most demographic analyses assume random sampling. However, sampling is usually not random in microbial studies, either on purpose or by the intrinsic difficulties of defining appropriate sampling strategies in microbiology, and this may severely affect the conclusions taken from the analysis of reconstructed genealogies. There are three major sampling biases in microbiology. *Clustered* sampling occurs when all samples are taken from a single sub-population, for example, a particularly virulent lineage. *Uniform* sampling of all major lineages is frequently found in studies aiming at maximizing the genetic diversity of samples. This bias may also result from sampling different environments (or patients) while analyzing a single isolate per site (thus disregarding differences in population sizes in each site). Finally, a very common type of *mixed* sampling bias is found in studies extensively sampling a sub-population and a small number of very diverse individuals from other sub-populations. This gives a broad view of

the genetic diversity in the species, while focusing in a subpopulation of interest. Analyses using sequences available in databanks are prone to combine the sampling biases of the different underlying studies.

We surveyed the available literature on the use of skyline plots to describe bacterial population demography and found that nearly all studies showed skyline plots suggestive of population expansion. We then decided to test if the violations of the assumptions of the neutral coalescent could be reasonably ignored when studying bacterial populations. For this, we simulated the evolution of bacterial populations of constant size using biologically realistic parameters for natural selection, recombination, and sampling bias. These sequences were then used to build skyline plots and make SFS-based inference of demographic changes. We did not use time calibration in the inference of the skyline plots. Therefore, the *Y*-axis in the skyline plots represents the inferred product of N_e by the mutation rate u ($N_e.u$) and the *X*-axis represents the expected number of mutations per site, which is an estimate of the distance from the present (Ho and Shapiro 2011). By convention, we represent zero mutations per site at the left of the skyline plots. Hence, the *X*-axes of the skyline plots are ordered from the present (left) to the past (right). In the last section, we present the analysis of data from *Escherichia coli* in the light of the results of simulations.

Results

The Puzzling Expansion of Most Bacterial Populations

We found 26 recent studies using skyline plots to analyze bacterial demography. We analyzed their characteristics in terms of TMRCA, demographic changes, and their presumed justifications (table 1). The TMRCA of these populations was extremely variable, from 3 years to over 100 million years. Many of these studies proposed some type of justification for the observed demographic changes. For example, demographic expansion in *Bordetella pertussis* was associated with the introduction of vaccination and expansion of escape variants (Bart et al. 2014). Demographic expansion in *Clostridium difficile* was associated with the date when the bacterium became a recognized nosocomial pathogen (He et al. 2010), and in *Salmonella enterica* serovar Typhi with the introduction of antibiotics (Roumagnac et al. 2006). Skyline plots suggested that the effective population size of *Neisseria gonorrhoeae* in Baltimore increased during most of the twentieth century and then decreased, presumably as the result of urban planning and changes in patterns of drug addiction (Perez-Losada et al. 2007). Some works suggested associations between the increase in effective population sizes and environmental changes, for example, glacial cycles in *Thiomonas* spp. (Liao and Huang 2012), and human population growth in *Mycobacterium tuberculosis* (Comas et al. 2013). However, a careful analysis of table 1 revealed a most puzzling trend: the vast majority of studies (21 out of 26) concluded that effective population sizes have increased.

Are all bacterial populations expanding? Researchers might focus preferentially on expanding bacterial populations, for example, recent epidemic clones, thus producing an ascertainment bias towards population expansion. Also, human populations have been growing exponentially and human-specific pathogens might have followed similar trends. However, a number of arguments cast doubt on these results. (1) The prevalence of bacterial pathogens (the majority of species in table 1) has decreased in the last century as the result of hygiene and the use of antibiotics (Cohen 2000). (2) Most of the remaining species in table 1 are commensals associated with multiple hosts (eventually including some nosocomials), or free-living bacteria for which human population growth might be of little relevance (especially since it is associated with decrease in the population of closely related animals that are often within the commensal host range). For example, *E. coli* is associated with most warm-blooded and some cold-blooded animals (Tenaillon et al. 2010), *Moraxella* was until recently regarded exclusively as a commensal of animals (Brenner et al. 2005), and *Thiomonas* spp. are free-living bacteria inhabiting extreme environments (Liao and Huang 2012). (3) The majority of the studies in table 1 have not checked for the assumptions of the standard neutral model, and those that did, only checked for the presence of recombination. Very few studies have used SFS to infer demographic changes in bacterial populations. While several of these works obtained SFS compatible with recent demographic expansions, they also showed that distortions in the SFS were partially caused by purifying selection (Cornejo et al. 2013; Pepperell et al. 2013; Touchon et al. 2014). These arguments led us to study the effects of violations of the assumptions of the standard neutral model in the inference of bacterial demography.

The Effect of Recombination

We made forward population genetics simulations of a locus of 20 kb with gene conversion and constant population size (see section "Methods"). Hence, deviations from the expectations of the neutral coalescent in the simulations were necessarily caused by recombination, not demography. The parameters for the simulations were taken from the literature for the model bacterium *E. coli* (table 2). Several studies estimated the rate of recombination over mutation in *E. coli* (reviewed in Bobay et al. 2015). We used an estimate based on the analysis of complete genomes (Touchon et al. 2009), which is among the lowest proposed and might therefore be conservative. The sequences resulting from our simulations were used to obtain skyline plots with BEAST (Drummond and Rambaut 2007). Our results show that even the moderate recombination rate observed in *E. coli*, leads to skyline plots with increasing values of $N_e.u$ for recent dates (fig. 1). This could be spuriously interpreted as an indication of population expansion. Simulations using ten times larger recombination rates (as observed in highly recombining bacteria), showed even stronger distortions in the skyline plots. Expectedly, recombination had no effect on the number of segregating sites (see *Recombination* in fig. 2), and lowered the variance, but did not affect the average, of the genome-wide average SFS (fig. 1). Consequently, recombination had no effect on the average estimate of Tajima *D* (although for a single locus see Thornton 2005).

Table 1. Published Works Using Skyline Plots to Estimate Demographic Changes in Bacteria.

Species	Conclusion	TMRCA	Authors' Comments
Bordetella pertussis	Expansion	200 Y	Surprisingly, vaccination was followed by increase not decrease in $N_e.u$, suggesting diversification of lineages escaping the vaccine (Bart et al. 2014)
Clostridium difficile	Expansion	35 Y	Population expansion coincides with the first reports of hospital outbreaks (He et al. 2010). Recombination tracts removed
Escherichia coli	Expansion	140 MY	A population bottleneck had a founding effect by purging diversity and leading to the formation of the extant major groups of *E. coli* (Wirth et al. 2006). 50-fold population expansion in the last 5 MY. Mentions the caveat of recombination
Legionella pneumophila	Expansion	20 Y	Correlation between population and reported number of clinical cases (Sanchez-Buso et al. 2014). Recombination tracts removed
Moraxella catarrhalis	Expansion	50 MY	The populations of antibiotic resistant isolates expand faster than those of sensitive bacteria (Wirth et al. 2007). Recombination tracts removed
Mycobacterium tuberculosis	All expansion	70 KY, 6.6 KY, 40Y	(1) Concludes about a parallel evolution between human (mitochondria) and this clade's N_e caused by a tight host-parasite association (Comas et al. 2013). (2) One expansion is associated with the industrial revolution, another with the first world war, and a recent contraction is associated with the introduction of antibiotherapy (Merker et al. 2015). (3) Expansion is associated with acquisition of multi-drug resistance (Eldholm et al. 2015)
Mycoplasma gallisepticum	Expansion	17 Y	Population expansion (Delaney et al. 2012)
Neisseria gonorrhoeae	Expansion, contraction	40 Y[a], 120 Y	(1) Population expansion measured in housekeeping functions parallels the number of clinical cases, but not when measured in an antibiotic resistance gene, suggesting it has been subject to positive selection. Results could be used in managing resistance (Tazi et al. 2010). Found no recombination events in the set. (2) Suggests that demographic changes are associated with selective sweeps caused by antibiotic resistance, crack epidemics and urban-planning. N_e decrease associated with 5× decrease in the prevalence of this obligatory human pathogen (Perez-Losada et al. 2007). Recombination tracts were removed
Pseudomonas aeruginosa	Expansion	0.005/nt[b]	Assigns the presence of a recent selective sweep (Guttman et al. 2008)
Pseudomonas fluorescens	Stable	0.07/nt[b]	Suggests ancient rapid growth followed by stabilization, but very close strains are absent (Guttman et al. 2008)
Pseudomonas syringae	Stable	0.1/nt[b]	Suggests it is an endemic pathogen (Sarkar and Guttman 2004)
Salmonella enterica serovar Paratyphi A	Expansion	450 Y	Population contraction associated with the introduction of antibiotics, followed by expansion that would be associated with environmental changes (Zhou et al. 2014). Recombination tracts removed
Salmonella enterica serovar Typhi	All expansion	10–71 KY, 25 Y	(1) Steady increase in population size in the last 3,000 years. Recombinant SNPs removed and strong selection checked (Roumagnac et al. 2006). (2) Expansion is consistent with epidemiological data reporting drug-resistant isolates. Recombinant regions removed (Wong et al. 2015)
Shigella sonnei	Stable	500 Y	The population size was found to be constant through time (Holt et al. 2012)
Staphylococcus aureus	Expansion	20 Y, 50 Y, 30 Y	(1) Rampant expansion might have followed trans-Atlantic spread (Nubel et al. 2010). (2) Phylodynamics analysis used to estimate epidemiological parameters such as the potential reproductive number. No signs of recombination identified (Prosperi et al. 2013). (3) Fit between demographic expansion and the epidemiology of the CC80 clone (Stegger et al. 2014)
Streptococcus pneumoniae	Contraction	15 Y	Population expansion and then contraction fits the observed number of clinical cases (Croucher et al. 2014). Recombination tracts removed
Streptococcus pyogenes	Expansion	80 Y	Associates population expansion with the acquisition of super-antigens (Davies et al. 2015). Recombination tracts removed
Streptococcus suis	Expansion	90 Y	Correlates population expansion with the introduction of new methods used for improved pig genetics (Weinert et al. 2015). Recombination tracts removed
Thiomonas spp	Expansion	7 MY	The demographic history matches the glacial cycles (Liao and Huang 2012)
Vibrio cholerae	Expansion	3 Y	Association with the history of the progression of an epidemic (Azarian et al. 2014). Found no evidence for recombination

Note—We show the TMRCA, the conclusion of the work, and the authors' justifications of the results. Multiple studies published for a given species are indicated as multiple lines in the column TMRCA and by the respective numbers in the last column.

[a]TMRCA not indicated. The value indicates the span of the *X*-axis on the skyline plot.

[b]Studies did not perform time calibration and present only the number of mutations per site.

Table 2. Parameters for *E. coli* Populations Used in the Simulations.

Parameter	Value	Reference
Effective population size (N_e)	1.8×10^8	Hartl et al. (1994)
Genomic adaptive mutation rate	1×10^{-5}	Perfeito et al. (2007)
Genomic deleterious mutation rate	2×10^{-4}	Kibota and Lynch (1996)
Average value of s[a]	$\pm 7 \times 10^{-3}$	Perfeito et al. (2007) and Gallet et al. (2012)
Mutation rate per generation (u)	8.9×10^{-11}	Wielgoss et al. (2011)
Genome size (nt)	5×10^6	Touchon et al. (2009)
Recombination/mutation rate	1	Touchon et al. (2009)
Size of recombination tracts	542	Didelot et al. (2012)
SNPs recombination/mutation	2.5	Touchon et al. (2009)
Weak selection ($N_e.s$)	5	
Strong recombination/mutation rate	10	

[a]The absolute values of s for adaptive and deleterious mutations being in the same order of magnitude we used an average for both.

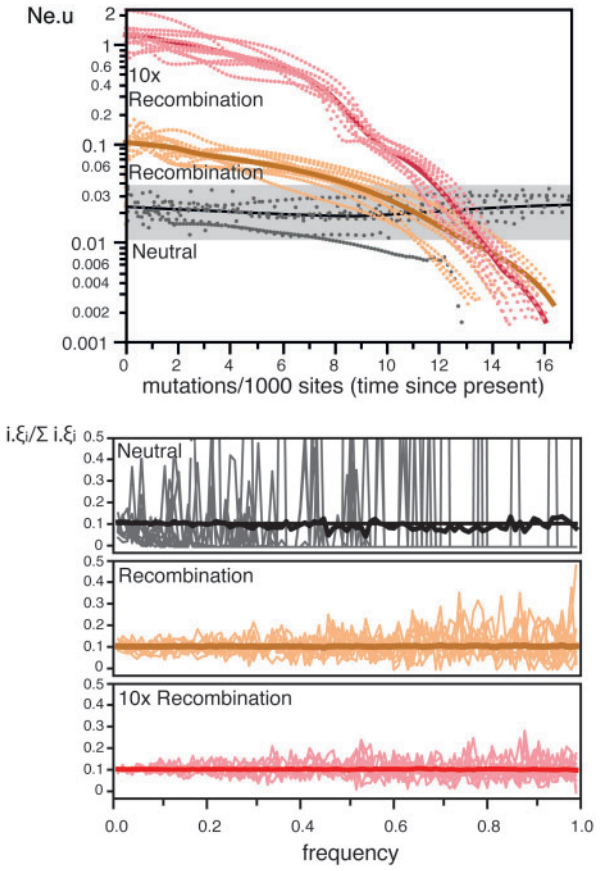

Fig. 1. The effect of recombination on skyline plots and SFS. The simulations used the *E. coli* population parameters (*Recombination*), ten times higher recombination rates (*10× Recombination*), or no recombination (*Neutral*). *Top* The simulations in the skyline plots are represented as dotted lines. The thick lines represent the smooth kernel fit (resp. $R^2 = 0.81$, $R^2 = 0.87$, and $R^2 = 0.38$). *Bottom*. SFS (distribution of the frequencies of all nucleotide polymorphisms in the sample) for each condition. The thick line indicates the average SFS over 1,000 replicates whereas the thin shaded lines are the observed SFS for ten random replicates. All SFS were transformed and normalized (see section "Methods"). Colors match the same datasets in both plots.

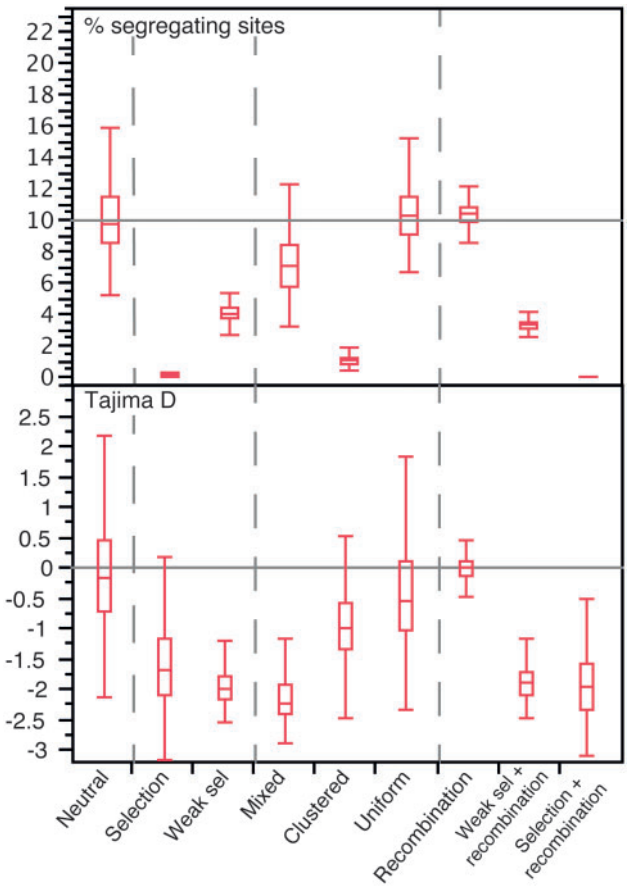

Fig. 2. Distribution of the number of segregating sites and Tajima D values in each set of 1,000 simulations. The gray line in the top panel corresponds to the expected number of segregating sites under the standard neutral model: $\pi = \theta.L.a_n$ where $a_n = \sum_{1}^{n-1} \frac{1}{i}$. Here, $\theta = 0.02$, $L = 20,000$, and $n = 100$. The gray line in the bottom panel corresponds to the expected Tajima D under the neutral model ($D = 0$).

We then tested if state-of-the-art methods aiming at producing "recombination free" phylogenetic trees could produce unbiased skyline plots. We analyzed ten simulations with ClonalFrame to obtain a matrix of distances between individuals purged from recombination (Didelot and Falush 2007). We used these matrices to infer phylogenies and these phylogenies to compute skyline plots. The latter showed very clear and systematic increase in the values of $N_e.u$ for recent times (supplementary fig. S1, Supplementary Material online). The average amplitude in $N_e.u$ (measured as the ratio between the maximal and the minimal value) was three times higher than the one obtained without the use of ClonalFrame, that is, with the primary data (see *After ClonalFrame* in fig. 3). This suggests that ClonalFrame distance matrices are skewed so that the trees inferred from them have internal branches more affected by the removal of recombination than the external branches. These results are in line with a previous study showing that removing homoplasies in recombined sequences worsens the distortions in skyline plots (Hedge and Wilson 2014). Hence, trying to remove polymorphism caused by recombination may aggravate the biases of demographic studies using skyline plots.

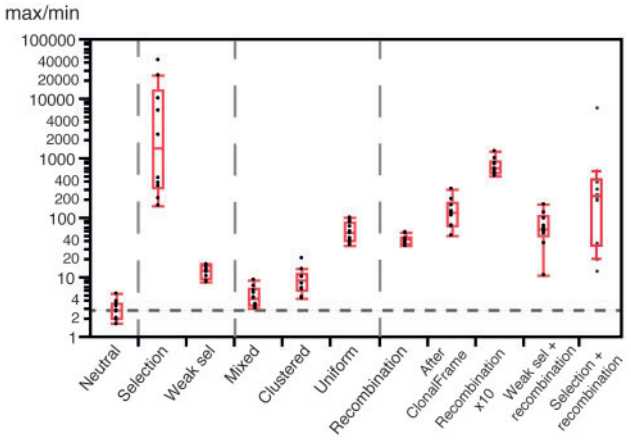

FIG. 3. Boxplots of the ratios between the maximal and minimal $N_e.u$ values for skyline plots (ten simulations each), across the different types of simulations. All other categories were significantly different from *Neutral* (all $P < 0.01$ Wilcoxon tests, except the comparison between *Neutral* and *Mixed*, $P = 0.0102$, same test).

The Effect of Selection

Experimental works indicate that >45% of the mutations are deleterious (Kibota and Lynch 1996), and >2% are adaptive (Perfeito et al. 2007) in *E. coli*. The effective population size of the species is estimated at $>10^8$ (Hartl et al. 1994; Lynch 2006). The average selective effects of mutations in *E. coli* are much larger than the inverse of the effective population size (table 2), which implies that their fate is mostly driven by selection (Kimura 1983). Our simulations using these parameters resulted in very strong distortions in the skyline plots, showing higher $N_e.u$ values for recent dates (see *Selection* in fig. 3). These patterns might have been interpreted as population expansions if the effect of selection had been ignored. Under strong selection, diversity is constantly being purged and swept away by recurrent selective sweeps. Accordingly, the fraction of segregating sites in these simulations was only ~0.16%, to be compared with ~10% for the neutral simulations (see *Selection* in fig. 2). The effect of strong selection was also apparent in the SFS, where extremely rare and frequent alleles were in large excess (fig. 4), presumably due to the selective sweeps caused by beneficial mutations. This resulted in negative values of Tajima *D* (fig. 2).

Some of the species listed in table 1 have narrow host ranges and might have much smaller N_e than *E. coli*. We therefore made simulations using parameters corresponding to populations with $N_e.s = \pm 5$ (*s* being the average selection coefficient on sites under selection) and a distribution of the frequency of sites under selection similar to *E. coli*. If these species have similar distributions of selective effects as those used *for E. coli* (*i.e.*, similar *s*), this value corresponds to N_e close to 1,000 (five orders of magnitude lower than *E. coli*). One should note that even bacteria obligatorily associated with humans are thought to have higher absolute values of N_e or $N_e.s$, for example, the N_e of *Neisseria meningitidis* was estimated at 10^5 (Treangen et al. 2008), and the average nonsynonymous values of $N_e.s$ were estimated at −5 for *M. tuberculosis* (Pepperell et al. 2013) and at −17 for

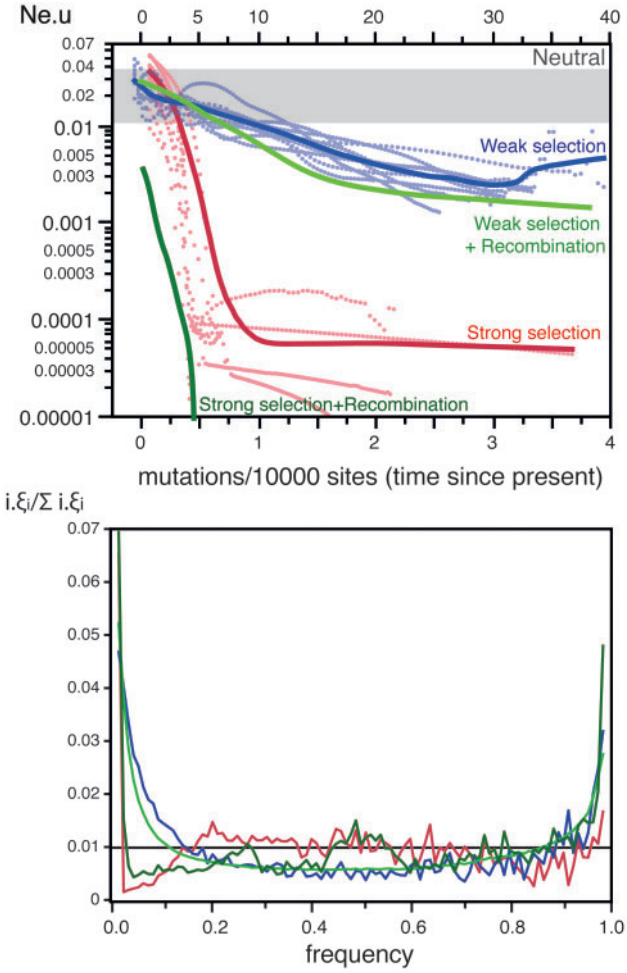

FIG. 4. The effect of selection on ten skyline plots (top) and 1,000 SFS (bottom). *Top* The simulations were represented as dotted lines. The thick lines represent the smooth kernel fit for strong and weak selection (resp. $R^2 = 0.78$, $R^2 = 0.79$). For the analysis of selection and recombination only the kernel fits are indicated ($R^2 = 0.80$). The grey box indicates the range of variation of the *Neutral* simulations in figure 1. *Bottom* The thick lines represent the average SFS over 1,000 simulations. In all SFS plots, the horizontal black line indicates the neutral expectation. Colors match the same datasets in both plots.

Streptococcus mutans (Cornejo et al. 2013). As expected, simulations incorporating such weak selection showed patterns much less extreme than those obtained under strong selection. For example, the average fraction of segregating sites in the former was ~4%, less than half of the neutral expectation but over two orders of magnitude more than under strong selection (see *Weak sel* in fig. 2). The skyline plots and the SFS under weak selection also showed less striking distortions (see *Weak sel* in figs. 3 and 4). Nevertheless, deviations from the expectation under neutral evolution were still very important in both analyses (negative Tajima *D*, fig. 2). These are likely to be caused by low-frequency segregating mildly deleterious mutations and by the selective sweeps caused by beneficial mutations. Hence, selection affects the inference of demography even when the values of N_e are uncharacteristically low for bacterial populations.

In our previous simulations, we have included positive and purifying selection. We therefore assessed the separate impact of each of these components of the evolutionary process on the skyline plots and on the SFS. For this we made simulations with just either positive or purifying selection. The effect of strong selection on skyline plots and SFS was caused exclusively by positive selection (supplementary fig. S2, Supplementary Material online). Accordingly, the SFS for strong purifying selection shows no excess of rare or frequent variants. This is because of the extremely rapid purge of deleterious mutations of strong effect. On the other hand, the significant effect of weak selection on the skyline plots and SFS is caused by both purifying and positive selection (supplementary fig. S3, Supplementary Material online). The SFS and skyline plots of populations evolving under weak purifying selection show an excess of rare variants and an increase in $N_e.u$ for recent times (supplementary fig. S4, Supplementary Material online). This shows that when selection is very strong, only positive selection affects the reconstructed genealogies, whereas when selection is weaker, both positive and purifying selection affect the reconstructed genealogies (and thus the skyline plot).

We then simulated the joint effects of selection and recombination on the reconstructed genealogies to check if recombination might moderate the effects of selection (fig. 4). The joint effect of recombination and selection (weak or strong) on the skyline plots was noticeable, that is, led to even stronger distortions in the plots, than the independent effects of each taken separately ($P < 0.0001$, Wilcoxon test). The SFS with selection and recombination were not appreciably different from the ones with selection under no recombination (compare the pairs of lines in the SFS of fig. 4). As a result, Tajima D is negative whenever there is selection, that is, with or without recombination (fig. 2). These results show that one cannot ignore the effect of selection on the analyses of bacterial demography.

The Effect of Sampling Bias

We simulated three types of typical sampling biases in the study of microbial population genetics. In these simulations, there were no changes in population size, no selection, and no recombination. We simulated sampling biases by clustering the final individuals evolved in the simulations in groups using sequence similarity and then sampling these groups in different ways (see section "Methods"). The results showed that different types of sampling bias affect in very diverse ways the shape of the tree and of the SFS, and thus the inference of demographic changes (fig. 5).

The sampling of a single group (clustered sampling), resulted in skyline plots with lower average values of $N_e.u$, as expected, and a peak of high $N_e.u$ for times very close to the present (see supplementary fig. S5, Supplementary Material online, for the values close to 0). The amplitudes of $N_e.u$ values were on average three times larger than those of neutral populations (*Clustered* in fig. 3). The simulations also showed slight over-representation of rare and frequent variants in the SFS. Clustered sampling produced alignments with far fewer (approximately ten times) segregating sites than the neutral simulations (*Clustered* in fig. 2). Hence, sampling a sub-population produces patterns akin to very recent population size expansions.

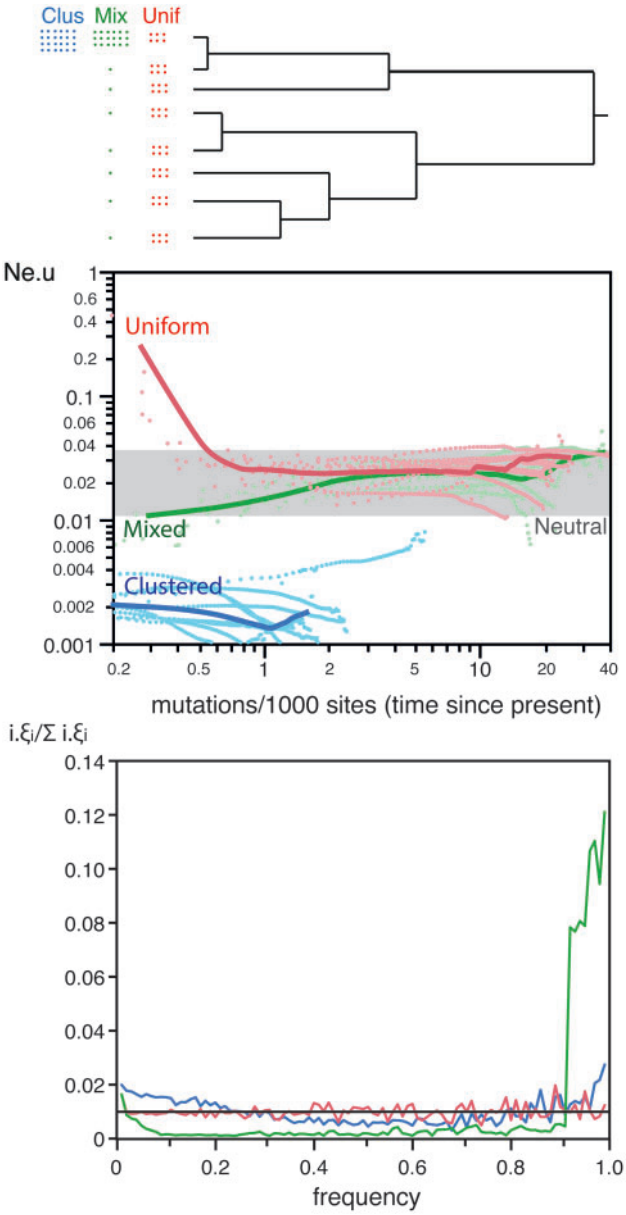

Fig. 5. Analysis of three types of sampling biases. *Top* Schematic representation of the different types of sampling biases in a species tree (see section "Methods" for a precise definition). *Center* Skyline plots for each set of ten simulations. The dotted lines represent the simulations. The thick line represents the smooth kernel fit (resp. Clustered $R^2 = 0.63$, Uniform $R^2 = 0.86$, Mixed $R^2 = 0.40$). The grey box indicates the range of variation of the *Neutral* simulations in figure 1. See supplementary figure S5, Supplementary Material online for a zoom for values of clustered bias close to zero. *Bottom* Average SFS for the three datasets (1,000 simulations for each). Colors match the same datasets in both plots.

We simulated uniform sampling by re-sampling the same number of individuals in each group. This led to skyline plots with increasing values of $N_e.u$ for recent dates (fig. 5). In fact, this sampling bias resulted in reconstructed genealogies with fewer than expected short terminal branches, which is akin to

the effect produced by strong population expansion. The consequent distortion of the reconstructed genealogies can be extremely important since these skyline plots had $N_e.u$ amplitudes >100 times higher than those found on neutral populations (*Uniform* in fig. 3). On the other hand, uniform sampling had essentially no effect on the SFS (fig. 5).

Mixed sampling bias was simulated by retrieving 91 individuals from one group and one from each of the remaining nine groups. These samples showed complex skyline plots, with initially increasing $N_e.u$ values followed by a sharp decrease for very recent dates (fig. 5). The SFS showed striking over-abundance of very frequent variants, some over-representation of rare variants and nearly no variants of intermediate frequency. This was associated with a negative Tajima *D* (*Mixed* in fig. 2). This pattern is the joint effect of the excess of very small external branches in the highly sampled group and the long internal branches linking the remaining groups in the reconstructed genealogy.

Joint Effects of Selection, Recombination, and Sampling Bias

We then studied the joint effect of sampling biases, recombination, and weak selection on skyline plots and SFS (as shown before, strong selection rapidly erases genetic diversity in the simulations). The increase in $N_e.u$ values in skyline plots inferred under uniform sampling bias was highly amplified when weak selection and recombination were also present, rising by almost four orders of magnitude (fig. 6). The SFS of these simulations showed a large excess of rare variants and a small excess of very frequent ones.

Clustered sampling of populations enduring recombination and weak selection resulted in skyline plots with a rapid increase in $N_e.u$, which then rapidly dropped to values very close to the initial ones. This process mimics initial strong population expansion, followed by very recent strong population contraction. The SFS showed a slight excess of rare variants and a large excess of frequent ones.

Finally, the skyline plots of simulations with mixed sampling, recombination, and weak selection showed a steady increase in $N_e.u$ and then a sharp decrease near the present. These patterns are also akin to the effects caused by ancient population expansions and recent population contractions. The SFS of these simulations showed an excess of both rare and frequent variants, with few intermediate values.

Analysis of the *E. coli* Core Genome

The parameters of fitness effects used in the simulations were measured on *E. coli* in the laboratory. It might be argued that these parameters are not representative of the effects observed in structured locally adapted natural populations. To assess the imprint of natural selection in *E. coli* we built its core genome (see section "Methods"). The analysis of the polymorphism in the ~1.3 million positions of the alignment of *E. coli* core genes, showed a pervasive pattern of purifying selection as expected from the simulations (fig. 7). Indeed, the ratio between the rates of nonsynonymous and synonymous substitutions (*dN*/*dS*) was significantly lower than one for all pairwise comparisons with sufficient polymorphism.

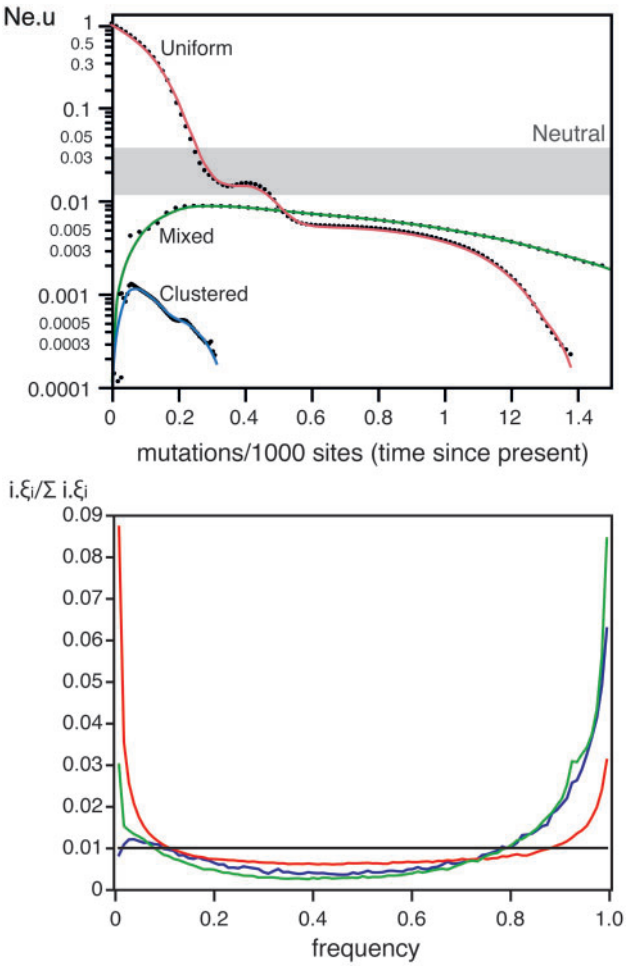

FIG. 6. *Top* Skyline plots for clustered, uniform and mixed sampling on simulations with weak selection and recombination (each point is an average of the ten simulations). The grey box indicates the range of variation of the *Neutral* simulations in figure 1. *Bottom* Average SFS for the same three datasets (1,000 simulations). Colors match the same datasets in both plots.

Importantly, when *dS* was higher than 1/5,000 the value of *dN*/*dS* was always smaller than 0.5. Multi-locus sequence typing (MLST) analyses use ~5 kb of sequenced data and thus only start becoming informative when there is more than one SNP per 5 kb. At this level of divergence, the values of *dN*/*dS* show that the distribution of polymorphism is already imprinted by natural selection, precluding the use of MLST to make demographic inferences using skyline plots.

We then made ten random samples of 10% of the core genome positions to produce ten skyline plots for *E. coli*. The results were highly concordant between samples, showing a pattern of increase in $N_e.u$ followed by a sudden drop for times closer to the present (fig. 7). The SFS of the *E. coli* core genome showed a strong over-representation of very frequent variants (fig. 7). We then restricted our analysis to genes of the core genome with individual phylogenies not significantly different from those of the concatenate of the core genes. We found that the topologies of the reconstructed trees of 1,146 of the 1,371 core genes were significantly different from the one of the core genome ($P < 0.01$,

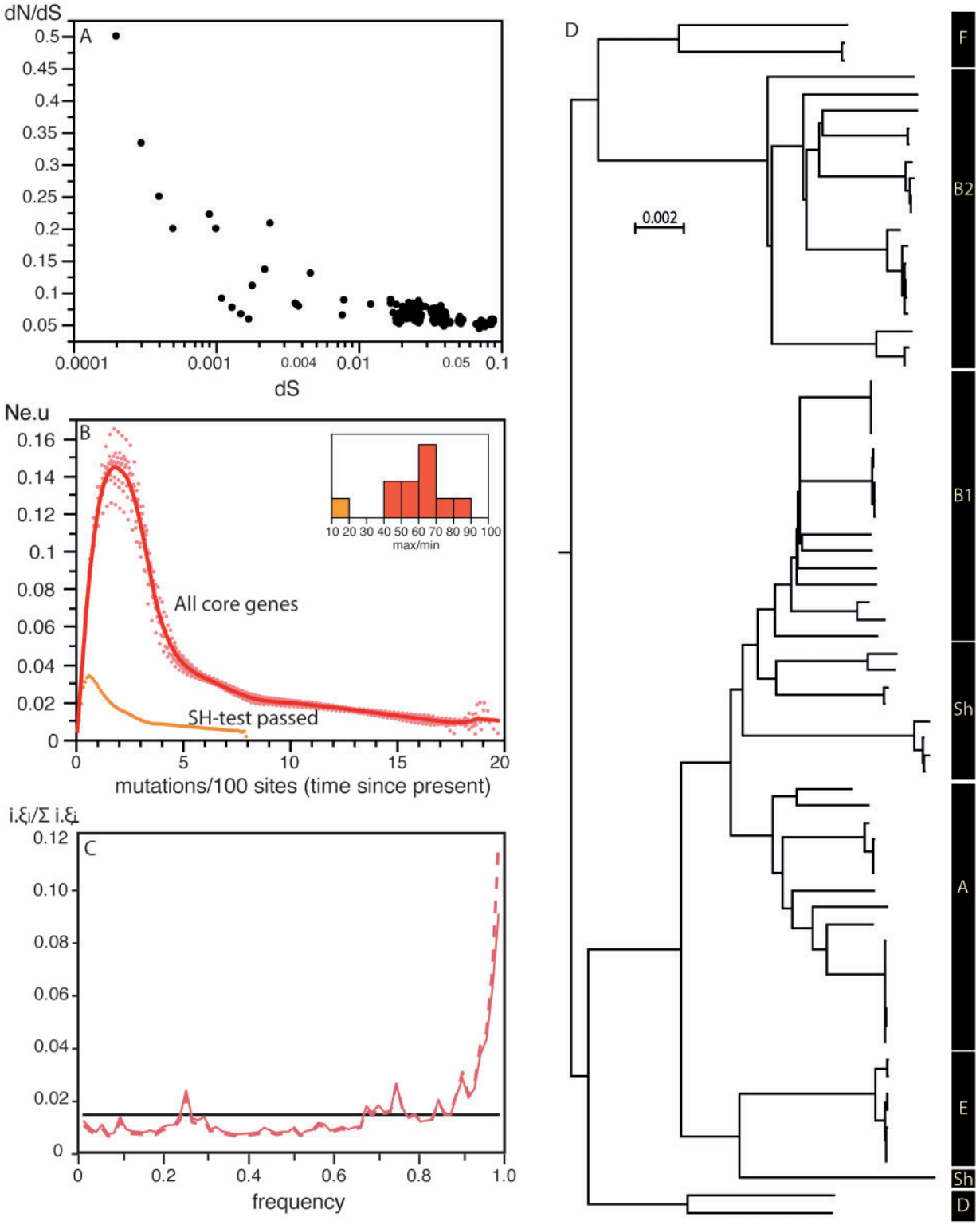

Fig. 7. Analysis of the core genome of *E. coli*. (*A*) Values of *dN*/*dS* versus *dS*. Each point represents a comparison between two strains using the concatenate of alignments of genes of the core genome. (*B*) Skyline plot. We made ten analyses of the dataset by randomly sampling each time a tenth of the core genome. The orange line represents the skyline of the concatenate of genes with reconstructed genealogies not significantly different from those of the core genome (passed the SH test at $P < 0.01$). The inset represents the ratio between the maximum and minimum values of $N_e.u$ for the 11 skyline plots (10 with the 1/10th samples of the core genome and one with the analysis of the concatenate of genes passing the SH-test). (*C*) The observed SFS is indicated in dashed red line, the corrected SFS (with Kimura's two-parameter model) is indicated in solid red line. The horizontal black line indicates the neutral expectation. The corrected SFS with the JC69 model (not shown here) is similar to the SFS corrected with Kimura's two-parameter model except for the last point, which is slightly higher. (*D*) *E. coli* distance-based phenetic tree with the major clades indicated on the right. A similar tree indicating all strains used in the analysis is in supplementary figure S6, Supplementary Material online.

Shimodaira–Hasegawa [SH] test). This analysis confirmed that the vast majority of genes in the genome are significantly affected by recombination, in spite of the low estimated rate of recombination in *E. coli*. We used the remaining 225 genes to build a skyline plot. This showed qualitatively identical trends, but less striking variations (fig. 7B).

Together, these results are consistent with a mixture of strong purifying selection and recombination producing patterns akin to demographic expansion in *E. coli* skyline plots. The excess of high-frequency variants observed in the unfolded SFS might be due to hitchhiking effects, appearing under strong selection and recombination. However, one cannot exclude the possibility that part of this excess might result from misoriented polymorphisms (polymorphisms for which the ancestral allele was wrongly assigned) (Baudry and Depaulis 2003), since corrections tend to lower this excess (see section "Methods" and fig. 7C). Alternatively, a mixed sampling bias could produce a drop in $N_e.u$ for the most recent times in skyline plots and a large excess of high-frequency variants in SFS. To test this hypothesis we built a phenetic tree for *E. coli* using a distance-based method (to minimize reconstruction artifacts associated with recombination). The analysis of this tree does not support the existence of a very strong mixed sampling bias (fig. 7).

Discussion

Recent advances in the analysis of genetic data using coalescent theory have the potential to unravel many novel aspects of microbial population genetics. The limitations of the underlying models are well known from the theoretical point of view (Frost et al. 2014). However, at the beginning of this work it was unclear if these limitations could compromise the use of such approaches to analyze bacterial data. Our study suggests that neglecting the effect of natural selection, recombination, and sampling biases may severely affect conclusions from phylodynamics analyses. These results are likely to be applicable to other phyla where these effects are important. An important effect that we have not quantified in this study concerns population structure, which tends to produce patterns akin to population contraction (Pannell 2003). Unfortunately, we could not study them due to current lack of modeling frameworks for simulating bacterial population structure. Previous studies have confirmed that animal population structure leads to distortions in skyline plots (Heller et al. 2013).

Some of the studies in table 1 tried to eliminate the effect of recombination by removing detectable recombination tracts from the analysis. Using ClonalFrame, we obtained even worst distortions in skyline plots. Similar results were previously found for the removal of homoplasies (Hedge and Wilson 2014). While we cannot offer a clear explanation for this observation, we presume it is caused by the removal of only certain specific types of recombination events (or polymorphism) from the data. Interestingly, the analysis of *E. coli* genomes suggests that removing all genes whose trees are incongruent with that of the core genome (SH test) attenuates the effect of recombination. The reasons for this, and the consequences of removing these sequences, will require further study. Yet, the relative apparent success of this method might just derive from the bias of the SH test toward removing the recombining genes producing genealogies incompatible with the average genealogy of the core genome (while leaving for further analysis those that are compatible with this genealogy). This is expected to decrease the bias toward higher coalescent rates closer to the TMRCA. Importantly, the expectation of the SFS is insensitive to the presence of recombination and can be used to analyze genomic data deeply imprinted by recombination.

Previous theoretical studies suggested that selection on mutations of mild deleterious effect might not distort genealogies. This might explain why none of the studies in table 1 assessed the effect of natural selection on demographic inference. Yet, using population genetics parameters of *E. coli*, and even using much smaller values for $N_e.s$, we found striking distortions in skyline plots.

We observed very frequent selective sweeps in the simulations with the selection parameters from *E. coli*. It must be emphasized that the high genetic diversity of the *E. coli* core genome is not fully consistent with such a succession of sweeps. However, it could be compatible with frequent soft sweeps, as recently described in *E. coli* adaptation to the mouse gut (Barroso-Batista et al. 2014). It would also be compatible with sweeps associated with local adaptation of certain lineages (Cohan and Perry 2007), or negative-frequency-dependent selection (Takeuchi et al. 2015). Finally, the existence of abundant strongly adaptive mutations in *E. coli* is consistent with previous results showing that a large fraction of amino acid substitutions between the *E. coli* and *Salmonella* lineages have been fixed by positive selection (Charlesworth and Eyre-Walker 2006).

To benefit from the power of coalescent-based approaches, one must find ways of controlling the distortions produced by selection on reconstructed genealogies. Unfortunately, practical and efficient ways of using the coalescent with selection are not yet available. Meanwhile, some simple controls might allow to identify or even estimate the effect of selection on demographic inference. For example, synonymous and nonsynonymous changes are very differently affected by selection, in spite of codon usage (Sharp et al. 2010), and partitioning the data in these two categories could shed light on the effect of selection on skyline plots and SFS. Comparisons between highly expressed and weakly expressed genes may also be informative since the former endure more intense selection for both synonymous and nonsynonymous substitutions (Rocha and Danchin 2004). Very recent polymorphism is relatively less imprinted by selection (Ho et al. 2005; Rocha et al. 2006), and might produce less biased patterns in skyline plots. Interestingly, the only published skyline plots in table 1 showing population contractions were based on samples with very short TMRCA (table 1). Unfortunately, the analysis of dN/dS in *E. coli* shows that even the very recent polymorphism was

affected by purifying selection (fig. 7). Skyline plots on larger time spans are even more imprinted by natural selection and interpretation purely in terms of demographic changes should not be made in the absence of control for natural selection.

Random sampling is a key underlying hypothesis of most statistical methods for the inference of demographic changes. However, funding agencies often stimulate researchers to focus on particular bacterial sub-populations of societal interest. This renders random sampling effectively impossible and might explain why surveys of microbial populations rarely explicit the statistical design of the sampling. As an example, despite the fact that *E. coli* is a commensal present in most warm-blooded animals, the vast majority of complete genomes available for this species are from strains pathogenic to humans. Since host-association, virulence, and antibiotic resistance vary between lineages of a species, over-sampling isolates of direct interest in terms of public health almost inevitably leads to statistical biases. Our results show that three common sampling strategies can severely bias the inference of demographic changes, especially in the presence of selection and recombination. Skyline plots studies of populations where these factors are important can exhibit almost any possible pattern of change.

The sampling of sub-groups of a population led to reconstructed genealogies suggesting recent population expansion. These results show that sub-trees of coalescent trees have distributions of coalescent rates different from those of the population tree. Hence, sampling a sub-population inevitably produces biased skyline plots. This brings to the fore the importance of precisely defining bacterial populations when inferring demographic changes using coalescent rates. The study of past demographic changes in microbial populations requires the use of adapted sampling techniques. Many such techniques have been developed in ecology (Young and Young 2013), even if their implementation poses technical challenges in microbiological research.

Many approaches alternative to skyline plots allow the inference of demographic changes. They all have specific advantages and disadvantages and their combination might facilitate the use of the available sequence data to make demographic inference. Lack of obvious neutral sites in bacteria renders difficult the establishment of demographic models independent of selection. Nevertheless, *dN*/*dS*-based approaches can be used to assess if natural selection has imprinted sequence data (although care must be taken to check if absence of evidence of selection is not due to lack of statistical power). Furthermore, the expectations of the SFS are insensitive to recombination and to uniform sampling when there is no selection or recombination. They are also less affected by differences in the intensity of natural selection, although in case of pervasive selection with recombination, the SFS shape will correspond to the predictions of multiple merger coalescent models (Tellier and Lemaire 2014). Therefore, joint analyses of skyline plots, detection of recombination, SFS (and derived statistics), *dN*/*dS*, and other population genetics methods are necessary to accurately infer changes in microbial demography.

Methods

Simulations

We made 1,000 simulations for each set of parameters. Simulations were done using SFS_code, which implements a generalized version of the Wright–Fisher forward population genetic model allowing finite-site mutation models with selection, recombination, and demography (Hernandez 2008). The typical simulation was done using a population of haploids with $N_e = 1{,}000$ individuals and one single genetic locus of 20,000 nucleotides. The length of the locus was chosen in order to be much larger than the average recombination tract in *E. coli* (~542 nt) (Didelot et al. 2012). In simulations under selection and recombination, we increased the length of the locus to 200,000 nucleotides, to obtain a sufficient number of polymorphic sites for further analyses. For simplicity, all nucleotides were included at similar frequencies and the substitution model was set to JC69 (equal mutation rates between all pairs of nucleotides) (Jukes and Cantor 1969). We used a 3-point mass model for selection (including negative, positive, and null values for the selection coefficient) (table 2). Modeling positive and purifying selection as two exponential distributions provides qualitatively similar results (but often produced numerical instabilities). Recombination was introduced exclusively as gene conversion (no crossovers allowed) in populations simulated as diploids (due to the constraints of the software). In this case, only half of the loci were used (1,000). The simulations were done using population scaled parameters accounting for the N_e of *E. coli* (table 2). Under these conditions, the size of the population effectively simulated does not affect the outcome of the analysis (Hernandez et al. 2007). In all cases, except those concerning sampling biases, we took 100 individuals from each final simulated population for further analysis.

Simulations of Biased Sampling

When analyzing biased sampling we took all 1,000 individuals from the final simulated populations. These sequences were used to build a distance matrix with FastTree v 2.1.7 using default parameters and the option-makematrix (Price et al. 2009). This distance matrix was then partitioned into clusters around medoids, a more robust version of *K*-means (Reynolds et al. 2006), using R. We simulated biased sampling of 100 individuals from the population in three ways. We simulated uniform distribution by picking one individual per cluster in an analysis where the population was clustered in 100 groups. We simulated mixed sampling bias by picking one individual per cluster for a total of ten individuals and then picking the remaining 90 individuals from one single cluster (analysis where the population was clustered in ten groups). We simulated clustered distribution by selecting all 100 individuals from a single cluster (analysis where the population was clustered in ten groups). It is important to note that a cluster obtained with this method may not exactly correspond to a monophyletic group as described in figure 5. The goal of our approach was to mimic the typical identification of clusters of bacterial groups used to select strains for sequencing, which are based on relatively imprecise methods (MLST or PFGE).

Analyses of Reconstructed Genealogies

We analyzed sequences using the generalized skyline plot model in BEAST with piecewise-linear modeling of the population size (skyline.popSize priors: initial = 3.2×10^{-4}, upper = 100, lower = 0), using the HKY model (the mutation model was parameterized so that its stationary frequencies were the empirical frequencies) (Hasegawa et al. 1985), setting a tight prior for k (lognormal, initial = 1, logMean = 0, Logstdev = 0.25), a strict molecular clock (as used in the simulations), and 30,000,000 iterations (sampling every 3,000 iterations). For simulations involving selection we made 300,000,000 iterations. The effective sample size (ESS) values were checked using Tracer and the runs were accepted when the ESS was higher than 200 for all parameters with eventual exception for some skyline.population parameters (as suggested by the manual of BEAST—[Drummond and Rambaut 2007]). Analyses resulting in poor ESS values were discarded and re-run. Tracer was used to compute all skyline plots except those made after the ClonalFrame analysis (see below). Given the computational cost of these analyses we only analyzed ten simulations per condition. However, the results were very consistent between simulations resulting in kernel fits with high R^2 (see text).

Analysis of the SFS

SFS were generated from random samples of 100 individuals. The mean SFS was calculated using 1,000 simulations. The exact ancestral state of each SNP was obtained using SFS_code. The SFS of the simulations were thus unfolded. For a better representation of the results, the SFS were transformed as follows. Let ξ_i denote the number of polymorphic sites at frequency $\frac{i}{n}$ in the sample of size n. We plot $i.\xi_i$ for $i \in [1, n-1]$, normalized by its sum, which is an unbiased estimator of the (supposedly unknown) mutation rate, often noted θ under the standard neutral model. Thus, the transformed SFS has a flat expectation under the standard neutral model, due to the well-known fact that $E[\xi_i] = \frac{\theta}{i}$.

For the analysis of *E. coli* data, the ancestral state is unknown and we used outgroup sequences. To correct for potential ancestral misorientations (i.e., when the nucleotide of the outgroup is erroneously inferred as the ancestral state), we calculated the probability of misorientations, using sites for which the outgroup nucleotide is different from the two nucleotides of the SNP (see Baudry and Depaulis 2003; Hernandez et al. 2007).

If q is the probability that the outgroup nucleotide is identical to the ancestral nucleotide, we have in expectation:

$$\xi_k^{obs} = \xi_k q + \xi_{n-k}(1-q) \text{ for } k \in [1, n-1],$$

where ξ_k^{obs} is the number of polymorphic sites at frequency $\frac{k}{n}$ before correction and ξ_k the real value.

We denoted by S the event that a given site is segregating, and by U the event that it is segregating *and* the outgroup nucleotide is different from the two nucleotides of the SNP. On one hand, $P(U \mid S)$ is easily estimated by the proportion x of sites that are segregating and yet have a different outgroup nucleotide. On the other hand, under the JC69 model of mutation, $P(U \cap S) = 2q\ P(S)$, neglecting the case when the ancestral nucleotide is different from the other three. Combining these two arguments we can estimate q by $x/2$.

Once q is estimated from the data, we can calculate the corrected values of the SFS:

$$\xi_k = \frac{\xi_k^{obs} - \xi_{n-k}^{obs}(1-q)}{2q-1} \text{ for } k \in [1, n-1].$$

We estimated q with two corrections, depending on the mutation model. Under the JC69 model of mutation, $q = 0.960$. Under Kimura's two parameters model (Kimura 1980), taking into account the transition and transversion rates, $q = 0.947$ (Baudry and Depaulis 2003).

ClonalFrame Analysis and Subsequent Skyline Plot

ClonalFrame was used with default parameters on the results of ten simulations with recombination, no selection and no sampling bias. All ClonalFrame outputs were imported in the ClonalFrame GUI (Didelot and Falush 2007). The convergence of MCMC traces was visually assessed. ClonalFrame outputs ultra-metric trees with multifurcations, but bifurcating trees are necessary to compute skyline plots. Hence, for each simulation, we exported the recombination-free distance matrix and used the R package *phangorn* to construct the UPGMA trees (Schliep 2011). We computed generalized skyline plots using the skyline function of the *ape* package (Paradis et al. 2004). The AIC criterion was applied to find the optimal ϵ spline parameter.

Analysis of *E. coli* Genome

We downloaded from RefSeq in November 2013 (Tatusova et al. 2015) the 62 genomes of *E. coli*, the nine genomes of *Shigella* spp. (in fact *E. coli* strains—Ochman et al. 1983) and the genome of *E. fergusonii* (the outgroup). Pairs of orthologous genes between two genomes were defined as bi-directional best hits, with >80% similarity in protein sequence, <20% difference in gene size, present within similar genetic neighborhoods (see Touchon et al. 2009 for details). The list of the core genome was defined as the intersection of all lists of pairwise analyses and included 1,371 genes. Genes from the same family of the core genome were aligned in protein sequence using MUSCLE v3.8 (default parameters, Edgar 2004) and back translated to DNA. These alignments were concatenated, making a total of 1,349,016 positions. They were used to compute the pairwise values of *dS*, *dN* and *dN*/*dS* between *E. coli* genomes using codeML from PAML v4 (parameters: runmode = −2; CodonFreq = 2; clock = 0; model = 2) (Yang 2007). Comparisons between very closely related isolates (i.e., with no single synonymous or nonsynonymous substitution in the core genome) were discarded.

SH Tests and Phenetic Tree

We built a phylogenetic tree of the core genome of *E. coli* using IQ-Tree (Nguyen et al. 2015) with the option to search for the best substitution model. The best model based on the BIC criterion was GTR + I + G4. For each gene we used IQ-Tree to make the SH test (1,000 replicates) using as a

reference tree the core genome tree. The phenetic tree in figure 7 was built using BIONJ (Gascuel 1997) from a distance matrix computed using TreePuzzle with the model GTR + I+G4 (Schmidt et al. 2002).

Supplementary Material

Supplementary figures S1–S6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Sylvain Brisse, Adam Eyre-Walker, and Guillaume Laval for comments on an earlier version of this manuscript. This project was financed by the Centre National de la Recherche Scientifique (CNRS) and the Institut Pasteur. G.A. and M.L. acknowledge support from the grant ANR-12-BSV7-0012 Demochips from the Agence Nationale de la Recherche (France). M.L. is funded by the PhD program Interfaces for Life of the University Pierre and Marie Curie (Paris). C.B. is funded by the PhD program Complexité du Vivant of the University Pierre and Marie Curie (Paris).

References

Achaz G. 2009. Frequency spectrum neutrality tests: one for all and all for one. *Genetics* 183:249–258.

Adams AM, Hudson RR. 2004. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics* 168:1699–1712.

Azarian T, Ali A, Johnson JA, Mohr D, Prosperi M, Veras NM, Jubair M, Strickland SL, Rashid MH, Alam MT, et al. 2014. Phylodynamic analysis of clinical and environmental *Vibrio cholerae* isolates from Haiti reveals diversification driven by positive selection. *MBio* 5:e01824–14.

Balbi KJ, Rocha EP, Feil EJ. 2009. The temporal dynamics of slightly deleterious mutations in *Escherichia coli* and Shigella spp. *Mol Biol Evol*. 26:345–355.

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:e1004182.

Bart MJ, Harris SR, Advani A, Arakawa Y, Bottero D, Bouchez V, Cassiday PK, Chiang CS, Dalby T, Fry NK, et al. 2014. Global population structure and evolution of *Bordetella pertussis* and their relationship with vaccination. *MBio* 5:e01074.

Baudry E, Depaulis F. 2003. Effect of misoriented sites on neutrality tests with outgroup. *Genetics* 165:1619–1622.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. *Genetics* 162:2025–2035.

Bobay LM, Traverse CC, Ochman H. 2015. Impermanence of bacterial clones. *Proc Natl Acad Sci U S A*. 112:8893–8900.

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* 140:783–796.

Brenner DJ, Krieg NR, Staley JT. 2005. Bergey's manual of systematic bacteriology. New York: Springer.

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

Cohen ML. 2000. Changing patterns of infectious disease. *Nature* 406:762–767.

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

Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, Parkhill J, Malla B, Berg S, Thwaites G, et al. 2013. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nat Genet*. 45:1176–1182.

Cornejo OE, Lefebure T, Bitar PD, Lang P, Richards VP, Eilertson K, Do T, Beighton D, Zeng L, Ahn SJ, et al. 2013. Evolutionary and population genomics of the cavity causing bacteria *Streptococcus mutans*. *Mol Biol Evol*. 30:881–893.

Croucher NJ, Hanage WP, Harris SR, McGee L, van der Linden M, de Lencastre H, Sa-Leao R, Song JH, Ko KS, Beall B, et al. 2014. Variable recombination dynamics during the emergence, transmission and 'disarming' of a multidrug-resistant pneumococcal clone. *BMC Biol*. 12:49.

Davies MR, Holden MT, Coupland P, Chen JH, Venturini C, Barnett TC, Zakour NL, Tse H, Dougan G, Yuen KY, et al. 2015. Emergence of scarlet fever *Streptococcus pyogenes* emm12 clones in Hong Kong is associated with toxin acquisition and multidrug resistance. *Nat Genet*. 47:84–87.

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, Falush D. 2007. Inference of bacterial microevolution using multilocus sequence data. *Genetics* 175:1251–1266.

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.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*. 7:214.

Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* 161:1307–1320.

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol*. 22:1185–1192.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32:1792–1797.

Eldholm V, Monteserin J, Rieux A, Lopez B, Sobkowiak B, Ritacco V, Balloux F. 2015. Four decades of transmission of a multidrug-resistant *Mycobacterium tuberculosis* outbreak strain. *Nat Commun*. 6:7119.

Frost SDW, Pybus OG, Gog JR, Viboud C, Bonhoeffer S, Bedford T. 2014. Eight challenges in phylodynamic inference. *Epidemics* 10:88–92.

Fu YX. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147:915–925.

Fu R, Voordouw G. 1997. Targeted gene-replacement mutagenesis of dcrA, encoding an oxygen sensor of the sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough. *Microbiology* 143:1815–1826.

Gallet R, Cooper TF, Elena SF, Lenormand T. 2012. Measuring selection coefficients below 10(-3): method, questions, and prospects. *Genetics* 190:175–186.

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol*. 14:685–695.

Grad YH, Lipsitch M. 2014. Epidemiologic data and pathogen genome sequences: a powerful synergy for public health. *Genome Biol*. 15:538.

Grenfell BT, Pybus OG, Gog JR, Wood JL, Daly JM, Mumford JA, Holmes EC. 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 303:327–332.

Guttman DS, Morgan RL, Wang PW. 2008. The evolution of the pseudomonads. In: Fatmi M, Collmer A, Iacobellis MS, Mansfield JW, Murillo J, Schaad NW, Ullrich M, editors. *Pseudomonas syringae* pathovars and related pathogens—identification, epidemiology and genomics. Dordrecht: Springer. p. 307–319.

Hartl DL, Moriyama EN, Sawyer SA. 1994. Selection intensity for codon bias. *Genetics* 138:227–234.

Hasegawa M, Kishino H, Yano T, Stabler RA, Dawson LF, Martin MJ, Holt KE, Seth-Smith HM, Quail MA, Rance R, 1985. Dating of the human-

ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*. 22:160–174.
He M, Sebaihia M, Lawley TD, Stabler RA, Dawson LF, Martin MJ, Holt KE, Seth-Smith HM, Quail MA, Rance R, et al. 2010. Evolutionary dynamics of *Clostridium difficile* over short and long time scales. *Proc Natl Acad Sci U S A*. 107:7527–7532.
Hedge J, Wilson DJ. 2014. Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not. *MBio* 5:e02158.
Heller R, Chikhi L, Siegismund HR. 2013. The confounding effect of population structure on Bayesian skyline plot inferences of demographic history. *PLoS One* 8:e62992.
Hernandez RD. 2008. A flexible forward simulator for populations subject to selection and demography. *Bioinformatics* 24:2786–2787.
Hernandez RD, Williamson SH, Bustamante CD. 2007. Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Mol Biol Evol*. 24:1792–1800.
Ho SY, Phillips MJ, Cooper A, Drummond AJ. 2005. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Mol Biol Evol*. 22:1561–1568.
Ho SY, Shapiro B. 2011. Skyline-plot methods for estimating demographic history from nucleotide sequences. *Mol Ecol Resour*. 11:423–434.
Holmes EC. 2007. Viral evolution in the genomic age. *PLoS Biol*. 5:e278.
Holt KE, Baker S, Weill FX, Holmes EC, Kitchen A, Yu J, Sangal V, Brown DJ, Coia JE, Kim DW, et al. 2012. *Shigella sonnei* genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nat Genet*. 44:1056–1059.
Hughes AL. 2005. Evidence for abundant slightly deleterious polymorphisms in bacterial populations. *Genetics* 169:533–538.
Jukes TH, Cantor CR. 1969. Evolution of protein molecules. Mammalian protein metabolism. New York: Academic Press. p. 21–132.
Kibota TT, Lynch M. 1996. Estimate of the genomic mutation rate deleterious to overall fitness in *E. coli*. *Nature* 381:694–696.
Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*. 16:111–120.
Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.
Li LM, Grassly NC, Fraser C. 2014. Genomic analysis of emerging pathogens: methods, application and future trends. *Genome Biol*. 15:541.
Liao P-C, Huang S. 2012. Patterns of microbial genetic diversity and the correlation between bacterial demographic history and geohistory. In: Caliskan M, editor. Genetic diversity in microorganisms. Shangai: INTECH Open Access Publisher. p. 123–148.
Liu X, Fu YX. 2015. Exploring population size changes using SNP frequency spectra. *Nat Genet*. 47:555–559.
Lynch M. 2006. The origins of eukaryotic gene structure. *Mol Biol Evol*. 23:450–468.
Martiny JB, Bohannan BJ, Brown JH, Colwell RK, Fuhrman JA, Green JL, Horner-Devine MC, Kane M, Krumins JA, Kuske CR, et al. 2006. Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol*. 4:102–112.
Mazet O, Rodriguez W, Chikhi L. 2015. Demographic inference using genetic data from a single individual: Separating population size variation from population structure. *Theor Popul Biol*. 104:46–58.
Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, Blum MG, Rusch-Gerdes S, Mokrousov I, Aleksic E, et al. 2015. Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. *Nat Genet*. 47:242–249.
Navarro A, Barton NH. 2002. The effects of multilocus balancing selection on neutral variability. *Genetics* 161:849–863.
Neuhauser C, Krone SM. 1997. The genealogy of samples in models with selection. *Genetics* 145:519–534.
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*. 32:268–274.
Nielsen R, Beaumont MA. 2009. Statistical inferences in phylogeography. *Mol Ecol*. 18:1034–1047.
Nielsen R, Wakeley J. 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* 158:885–896.
Nubel U, Dordel J, Kurt K, Strommenger B, Westh H, Shukla SK, Zemlickova H, Leblois R, Wirth T, Jombart T, et al. 2010. A timescale for evolution, population expansion, and spatial spread of an emerging clone of methicillin-resistant *Staphylococcus aureus*. *PLoS Pathog*. 6:e1000855.
O'Fallon BD, Seger J, Adler FR. 2010. A continuous-state coalescent and the impact of weak selection on the structure of gene genealogies. *Mol Biol Evol*. 27:1162–1172.
Ochman H, Whittam TS, Caugant DA, Selander RK. 1983. Enzyme polymorphism and genetic population structure in *Escherichia coli* and Shigella. *J Gen Microbiol*. 129:2715–2726.
Pannell JR. 2003. Coalescence in a metapopulation with recurrent local extinction and recolonization. *Evolution* 57:949–961.
Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290.
Pepperell CS, Casto AM, Kitchen A, Granka JM, Cornejo OE, Holmes EC, Birren B, Galagan J, Feldman MW. 2013. The role of selection in shaping diversity of natural *M. tuberculosis* populations. *PLoS Pathog*. 9:e1003543.
Perez-Losada M, Crandall KA, Zenilman J, Viscidi RP. 2007. Temporal trends in gonococcal population genetics in a high prevalence urban community. *Infect Genet Evol*. 7:271–278.
Perfeito L, Fernandes L, Mota C, Gordo I. 2007. Adaptive mutations in bacteria: high rate and small effects. *Science* 317:813–815.
Price MN, Dehal PS, Arkin AP. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol*. 26:1641–1650.
Prosperi M, Veras N, Azarian T, Rathore M, Nolan D, Rand K, Cook RL, Johnson J, Morris JG Jr, Salemi M. 2013. Molecular epidemiology of community-associated methicillin-resistant *Staphylococcus aureus* in the genomic era: a cross-sectional study. *Sci Rep*. 3:1902.
Przeworski M, Charlesworth B, Wall JD. 1999. Genealogies and weak purifying selection. *Mol Biol Evol*. 16:246–252.
Pybus OG, Rambaut A, Harvey PH. 2000. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* 155:1429–1437.
Ramos-Onsins SE, Rozas J. 2002. Statistical properties of new neutrality tests against population growth. *Mol Biol Evol*. 19:2092–2100.
Reynolds AP, Richards G, de la Iglesia B, Rayward-Smith VJ. 2006. Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *J Math Model Algorithms*. 5:475–504.
Rocha E, Danchin A. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol Biol Evol*. 21:108–116.
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*. 239:226–235.
Rocha EPC, Feil EJ. 2010. Mutational patterns cannot explain genome composition: are there any neutral sites in the genomes of bacteria? *Plos Genetics* 6:e1001104.
Roumagnac P, Weill FX, Dolecek C, Baker S, Brisse S, Chinh NT, Le TA, Acosta CJ, Farrar J, Dougan G, et al. 2006. Evolutionary history of Salmonella typhi. *Science* 314:1301–1304.
Sanchez-Buso L, Comas I, Jorques G, Gonzalez-Candelas F. 2014. Recombination drives genome evolution in outbreak-related *Legionella pneumophila* isolates. *Nat Genet*. 46:1205–1211.
Sarkar SF, Guttman DS. 2004. Evolution of the core genome of *Pseudomonas syringae*, a highly clonal, endemic plant pathogen. *Appl Environ Microbiol*. 70:1999–2012.
Schierup MH, Hein J. 2000. Consequences of recombination on traditional phylogenetic analysis. *Genetics* 156:879–891.
Schliep KP. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* 27:592–593.
Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18:502–504.
Sharp PM, Emery LR, Zeng K. 2010. Forces that influence the evolution of codon bias. *Philos Trans R Soc Lond B Biol*. 365:1203–1212.

Stegger M, Wirth T, Andersen PS, Skov RL, De Grassi A, Simoes PM, Tristan A, Petersen A, Aziz M, Kiil K, et al. 2014. Origin and evolution of European community-acquired methicillin-resistant *Staphylococcus aureus*. *MBio* 5:e01044–e01014.

Tajima F. 1989a. The effect of change in population size on DNA polymorphism. *Genetics* 123:597–601.

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

Takeuchi N, Cordero OX, Koonin EV, Kaneko K. 2015. Gene-specific selective sweeps in bacteria and archaea caused by negative frequency-dependent selection. *BMC Biol*. 13:20.

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.

Tazi L, Perez-Losada M, Gu W, Yang Y, Xue L, Crandall KA, Viscidi RP. 2010. Population dynamics of *Neisseria gonorrhoeae* in Shanghai, China: a comparative study. *BMC Infect. Dis*. 10:13.

Tellier A, Lemaire C. 2014. Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. *Mol Ecol*. 23:2637–2652.

Tenaillon O, Skurnik D, Picard B, Denamur E. 2010. The population genetics of commensal *Escherichia coli*. *Nat Rev Microbiol*. 8:207–217.

Thornton K. 2005. Recombination and the properties of Tajima's D in the context of approximate-likelihood calculation. *Genetics* 171:2143–2148.

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*. 6:2866–2882.

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:e1000344.

Treangen TJ, Ambur OH, Tonjum T, Rocha EPC. 2008. The impact of the neisserial DNA uptake sequences on genome evolution and stability. *Genome Biol*. 9:R60.

Volz EM, Koelle K, Bedford T. 2013. Viral phylodynamics. *PLoS Comput Biol*. 9:e1002947.

Vos M, Didelot X. 2009. A comparison of homologous recombination rates in bacteria and archaea. *ISME J*. 3:199–208.

Wall JD. 1999. Recombination and the power of statistical tests of neutrality. *Genet Res*. 74:65–79.

Weinert LA, Chaudhuri RR, Wang J, Peters SE, Corander J, Jombart T, Baig A, Howell KJ, Vehkala M, Valimaki N, et al. 2015. Genomic signatures of human and animal disease in the zoonotic pathogen *Streptococcus suis*. *Nat Commun*. 6:6740.

Wielgoss S, Barrick JE, Tenaillon O, Cruveiller S, Chane-Woon-Ming B, Medigue C, Lenski RE, Schneider D. 2011. Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with *Escherichia coli*. *G3* 1:183–186.

Williamson S, Orive ME. 2002. The genealogy of a sequence subject to purifying selection at multiple sites. *Mol Biol Evol*. 19:1376–1384.

Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MC, Ochman H, et al. 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol Microbiol*. 60:1136–1151.

Wirth T, Morelli G, Kusecek B, van Belkum A, van der Schee C, Meyer A, Achtman M. 2007. The rise and spread of a new pathogen: seroresistant *Moraxella catarrhalis*. *Genome Res*. 17:1647–1656.

Wong VK, Baker S, Pickard DJ, Parkhill J, Page AJ, Feasey NA, Kingsley RA, Thomson NR, Keane JA, Weill FX, et al. 2015. Phylogeographical analysis of the dominant multidrug-resistant H58 clade of Salmonella Typhi identifies inter- and intracontinental transmission events. *Nat Genet*. 47:632–639.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24:1586–1591.

Young LJ, Young J. 2013. Statistical ecology. New York: Springer Science & Business Media.

Zhou Z, McCann A, Weill FX, Blin C, Nair S, Wain J, Dougan G, Achtman M. 2014. Transient Darwinian selection in *Salmonella enterica* serovar Paratyphi A during 450 years of global spread of enteric fever. *Proc Natl Acad Sci U S A*. 111:12199–12204.