Correlated Mutations and Homologous Recombination Within Bacterial Populations

Mingzhi Lin* and Edo Kussell*^{†,1}

*Department of Biology and Center for Genomics and Systems Biology and [†]Department of Physics, New York University, New York 10003

ORCID ID: 0000-0002-8435-3780 (M.L.)

ABSTRACT Inferring the rate of homologous recombination within a bacterial population remains a key challenge in quantifying the basic parameters of bacterial evolution. Due to the high sequence similarity within a clonal population, and unique aspects of bacterial DNA transfer processes, detecting recombination events based on phylogenetic reconstruction is often difficult, and estimating recombination rates using coalescent model-based methods is computationally expensive, and often infeasible for large sequencing data sets. Here, we present an efficient solution by introducing a set of mutational correlation functions computed using pairwise sequence comparison, which characterize various facets of bacterial recombination. We provide analytical expressions for these functions, which precisely recapitulate simulation results of neutral and adapting populations under different coalescent models. We used these to fit correlation functions measured at synonymous substitutions using whole-genome data on *Escherichia coli* and *Streptococcus pneumoniae* populations. We calculated and corrected for the effect of sample selection bias, *i.e.*, the uneven sampling of individuals from natural microbial populations that exists in most datasets. Our method is fast and efficient, and does not employ phylogenetic inference or other computationally intensive numerics. By simply fitting analytical forms to measurements from sequence data, we show that recombination rates can be inferred, and the relative ages of different samples can be estimated. Our approach, which is based on population genetic modeling, is broadly applicable to a wide variety of data, and its computational efficiency makes it particularly attractive for use in the analysis of large sequencing datasets.

KEYWORDS bacteria; homologous recombination; population diversity; sample selection bias; sample ages; adapting populations; Bolthausen–Sznitman coalescent

BACTERIA can receive DNA fragments from their environment by different mechanisms, and integrate them into their genome in a set of processes collectively known as horizontal gene transfer (HGT) (Thomas and Nielsen 2005). While the importance of HGT in bacterial evolution is increasingly appreciated (Koonin and Wolf 2008; Shapiro *et al.* 2012; Oren *et al.* 2014; Ravenhall *et al.* 2015; Rosen *et al.* 2015), quantifying its impact across bacterial genomes, and in diverse environmental samples, remains a key challenge (Maynard Smith 1991; Soucy *et al.* 2015). In particular, a prevalent form of HGT involves homologous recombination of fragments that bear a high degree of sequence similarity to the recipient

genome (Andam and Gogarten 2011; Williams *et al.* 2012). Such transfer events are particularly difficult to detect, since they leave no obvious marks, and are indistinguishable based on nucleotide composition, yet they are likely to represent the majority of HGT events in bacteria (Fraser *et al.* 2007).

Three major mechanisms of HGT—transformation, conjugation, and transduction—mediate the passage of external DNA into bacterial cells. Once inside the cell, DNA fragments can be integrated into the genome either by homologous or nonhomologous recombination mechanisms (Thomas and Nielsen 2005). In homologous recombination, the fragment usually recombines into a homologous genomic locus replacing the existing DNA at the recipient locus. DNA transfer by homologous recombination is similar to gene conversion in eukaryotes—it is unidirectional, and changes occur only in the recipient locus. In nonhomologous recombination (also known as illegitimate recombination), a comparatively rare event, the fragment is inserted directly into the genome without replacing DNA. Homologous recombination breaks

Copyright © 2017 by the Genetics Society of America
doi: 10.1534/genetics.116.189621

Manuscript received March 21, 2016; accepted for publication December 15, 2016;
published Early Online December 20, 2016.

¹Corresponding author: Center for Genomics and Systems Biology,
Department of Biology, New York University, 12 Waverly Place, NY 10003.
E-mail: edo.kussell@nyu.edu

genetic linkage within a bacterial population, thereby alleviating clonal interference as well as Muller's ratchet, and is therefore expected to increase the rate at which bacteria evolutionarily adapt to their environment. Theoretical analyses have shown that a principal determinant of the speed of evolutionary adaptation is the recombination rate (Cohen *et al.* 2005; Neher *et al.* 2010; Weissman and Barton 2012; Neher and Hallatschek 2013; Weissman and Hallatschek 2014). Measuring these rates, and other population genetic parameters, is therefore crucial for describing the evolutionary dynamics of different bacterial species.

To infer the parameters and rates of the HGT process, mathematical models based on coalescent theory (Kingman 1982; Hudson 1983) have been used to study the genealogy of a sample of sequences in the presence of gene conversion or homologous recombination (Wiuf 2000; Wiuf and Hein 2000; McVean *et al.* 2002; Didelot and Falush 2007; Touchon *et al.* 2009). Using a coalescent model with gene-conversion, computationally intensive methods have been developed that allow estimation of recombination rates on a whole-genome scale (Didelot *et al.* 2010; Ansari and Didelot 2014). Other methods detect recombination events based on regional differences of single nucleotide polymorphism (SNP) density (Croucher *et al.* 2015) or inferred phylogeny (Marttinen *et al.* 2012), and are often used to establish a clonal phylogeny for a sample based on the vertically inherited regions. Patterns of SNP density combined with Markov Chain Monte Carlo simulations have also been used to infer genome-wide recombination parameters (Dixit *et al.* 2015). One caveat is that phylogenetic methods are particularly sensitive to errors in the reference phylogeny. When recombination occurs as frequently as mutations accumulate, which appears to be the case in many species (Vos and Didelot 2009), it is not clear whether a reference phylogeny can be accurately constructed based on sequence data alone.

In natural populations, the genealogical trees reconstructed from the sampled sequences can exhibit substantial departure from neutrality. Kingman's coalescent (KC) allows only pairwise mergers—at most two ancestral lines can merge at each coalescence event—and the branches of the coalescent tree are distributed evenly (Kingman 1982). In spatially expanding or panmictic adapting populations, multiple mergers of ancestral lines occurring at a single step are not rare in the genealogical trees, and (Brunet *et al.* 2007) showed that a special case of the multiple-merger coalescent, the Bolthausen–Sznitman coalescent (BSC) (Bolthausen and Sznitman 1998), describes their genealogies. More recently, (Desai *et al.* 2013) and (Neher and Hallatschek 2013) showed that, in rapidly adapting populations, exponential amplification of fit lineages leads to multiple mergers, and the resulting genealogies are well described by the BSC.

Here, we present an analytically tractable framework for inferring homologous recombination from sequence data that is applicable for any exchangeable coalescent model, which subsumes KC and BSC models. We introduce new population-genetic quantities based on correlations of substitutions within

a population, and use these to accurately infer recombination parameters and provide consistency tests of the model. We study the impact of selection on these quantities, and demonstrate a smooth transition from KC to BSC predictions with the increase of selection strength. Importantly, our method does not rely on the construction of phylogenetic trees, and, because it is analytically tractable, it yields results as fast as sequence data can be read into memory. We demonstrate the power of this method using large datasets of whole-genome bacterial sequences.

Materials and Methods

Neutral population models

We consider a generalized Moran model that evolves a population of N individuals in continuous time with overlapping generations (Moran 1958), and allows individuals to have multiple offspring at each reproduction event. We choose an arbitrary unit of time to measure all rates in the model. Reproduction events occur with rate G at exponentially distributed times. At each event, exactly one individual reproduces, yielding $U - 1$ new individuals that replace $U - 1$ randomly chosen individuals out of the $N - 1$ existing individuals in the population, not including the parent. The value U is a random variable that can take values from $2, \dots, N$, with probability distribution $P(U)$. To obtain the classic Moran model, one chooses $P(2) = 1$, and $P(U > 2) = 0$, and it is well known that, in this case, the resulting coalescent genealogies are given by KC statistics (Kingman 1982). Alternatively, one can choose $P(U) = \frac{N}{N-1} \frac{1}{U(U-1)}$, which we will call the *Schweinsberg model*, in which case the coalescent genealogies are given by BSC statistics (Schweinsberg 2012).

Mutations, recombination, and fragment size distribution

Mutations and DNA transfer occur with uniform rates throughout the population, and across genomes. We model circular genomes of length L with an “alphabet” of size a letters, where the usual choice is $a = 4$, corresponding to the four bases of DNA. Each site on a genome mutates with rate μ (per generation) to any of the $a - 1$ remaining letters with equal probability. DNA transfer events, in which an external piece of DNA is imported and homologously recombined into the genome, occur at a rate R (per generation) in each genome. We define $\gamma \equiv R/L$ as the transfer rate per site, where L is the length of the genome. Each time a transfer occurs into a given genome (the “recipient”), one of the N individuals is chosen randomly to be the “donor,” and a randomly chosen fragment of size f is copied from the donor, replacing the existing piece at the identical location in the recipient.

The fragment size f is a random variable determined by a probability distribution function, $p(f)$. We define the rate at which a site is affected by recombination, $r \equiv \gamma \bar{f}$, where $\bar{f}$ is

the average fragment size. We call the parameter r , the *recombination coverage rate*, since it measures the fraction of the genome that is “covered” by horizontally transferred pieces of DNA newly acquired in each generation. For two sites separated by distance l along the genome, the coverage rate r at each site can be partitioned into two contributions: the rate at which only the single site is covered by a transfer, $r_1(l)$ (one-site transfer), and the rate at which both sites are covered, $r_2(l)$ (two-site transfer), given by

$$r_1(l) = \gamma \sum_{f=1}^l f p(f) + \gamma l \sum_{f=l}^L p(f) \quad (1)$$

$$r_2(l) = \gamma \sum_{f=l}^L (f-l) p(f), \quad (2)$$

where $r_1(l) + r_2(l) = r$. When l is smaller than the minimal size of transferred fragments, these functions depend only on $\bar{f}$, with $r_1(l) = \gamma l$ and $r_2(l) = \gamma(\bar{f} - l)$. All simulations shown here used a constant fragment size f_0 , setting $p(f) = \delta_{ff_0}$, where δ_{ij} is the Kronecker delta function; all analytical results were derived in terms of $r_1(l)$ and $r_2(l)$, hence, they are generally applicable for any fragment size distribution.

Genome sequences and population diversity

Each genome sequence $\vec{g}$ is specified as g_i , for $i = 0, 1, \dots, L-1$, over an alphabet of size a , with $g_i \in \{1, 2, \dots, a\}$. For a pair of genomes, $\vec{g}$ and $\vec{g}'$, we define the substitution sequence $S_i(\vec{g}, \vec{g}') \equiv 1 - \delta_{g_i, g'_i}$. For notational convenience, we number the possible pairs of sequences $k = 1 \dots N(N-1)/2$, and write the substitution sequence of the k -th pair as S_i^k , which we call the *substitution variable*, and which takes values $\{0, 1\}$. The substitution variable can be thought of as a set of observations indexed by the variables i and k , and we can average over the indices in various ways. For example, we define the average sequence distance between the k -th pair of genomes as $d(k) \equiv \langle S_i^k \rangle$, where we let $\langle \cdot \rangle$ denote the average over sequence positions i . The population average pairwise distance, which we will also call the *population diversity*, is given by

$$d \equiv \overline{d(k)} = \langle \overline{S_i^k} \rangle, \quad (3)$$

where the bar denotes the average over all genome pairs k across the population. The variance of pairwise distances is written as

$$\sigma^2 \equiv \text{Var}[d(k)] = \langle \overline{S_i^k}^2 \rangle - \langle \overline{S_i^k} \rangle^2. \quad (4)$$

Equivalently, and often more conveniently, we consider $S = S_i^k$ to be a random variable that depends on i and k . Since we will be computing averages and correlations of S over the indices, we take i and k to be random variables with a uniform distribution over their respective values. In this notation, we have

$d(k) = E(S|k)$, where $S|k$ is the conditional random variable, and $E(\cdot)$ denotes expectation. The population diversity is given by $d = E[E(S|k)] = E(S)$. The variance of pairwise distances is written as $\sigma^2 = \text{Var}[E(S|k)]$

It is important to emphasize that S , as defined here, depends explicitly on the collection of genomes $\{\vec{g}\}$ that make up the population at any given time. Throughout the paper, any quantity that is defined using S (e.g., d , σ^2 , and the correlation functions below) is therefore a random variable whose distribution is determined by all possible realizations of the stochastic population dynamics process. In simulations, we average over many runs to calculate the expectations of these quantities, while our analytical results are constructed explicitly to compute their expectations over the stochastic dynamics.

Population genetic parameters

We define several additional quantities, which will be useful when analyzing recombination and diversity at the population level. We consider a pair of individuals that coalesced at time t ago. Their *mutational divergence*, $2t\mu$, corresponds to the fraction of the genome that accumulated mutations on the two lineages since coalescence. In the neutral models that we consider here, the mean coalescence time $\bar{t}$ is $N/2$ generations (Appendix B), and the *mean mutational divergence* is given by the well-known population genetic parameter quantity $\theta \equiv N\mu$. Analogously, we define the pairwise *recombinational divergence*, $2t\gamma$, which is the average number of recombination events per site that occurred on two lineages. The *mean recombinational divergence* is then given by the quantity $\phi \equiv N\gamma$. Since each event covers on average $\bar{f}$ base pairs, we define the mean *recombination coverage* as $\rho \equiv \phi\bar{f} = Nr$, which measures the fraction of the genome affected by recombination since divergence. We also define $\rho_1(l) \equiv Nr_1(l)$ and $\rho_2(l) \equiv Nr_2(l)$ as the one-site and two-site recombinational coverage, where $\rho_1 + \rho_2 = \rho$.

Correlation functions

We consider a pair of sites separated by distance l along the genome, denoted i and $i+l$. We define $X \equiv S_i^k$ and $Y \equiv S_{i+l}^k$ to be the substitution variables at any two such sites, and treat X and Y as random variables, as above. This allows us to define three distinct correlation functions, each of which involves all pairs of sites separated by a distance of l along the genome (see Figure 1). The *mutation correlation function*,

$$c_M(l) \equiv \overline{\langle S_i^k S_{i+l}^k \rangle} - \langle \overline{S_i^k} \rangle \langle \overline{S_{i+l}^k} \rangle = E[\text{Cov}(X|k, Y|k)], \quad (5)$$

assesses within each pair of genomes the correlation of mutations across all pairs of sites i and $i+l$; the *structure correlation function*,

$$c_S(l) \equiv \langle \overline{S_i^k S_{i+l}^k} - \overline{S_i^k} \cdot \overline{S_{i+l}^k} \rangle = E[\text{Cov}(X|i, Y|i)], \quad (6)$$

Figure 1 Illustration of population genetic correlation functions. On the left, a population of N genomic sequences each of length L is shown. A pair of sequences, $\vec{g}$ and $\vec{g}'$, is compared at each site i , yielding the substitution sequence S_i^k , where k indexes the pair of genomes among all possible $N(N-1)/2$ pairs. A pair of positions along the sequence separated by a distance l is shown. The population diversity, d , variance of pairwise distances, σ^2 , and correlation functions $c_M(l)$ mutational correlation, $c_S(l)$ structure correlation, and $c_R(l)$ rate correlation, are shown to involve taking averages or covariances in different directions along the substitution sequences.

measures how strongly pairs of sites are correlated across the population; and the *rate correlation function*,

$$c_R(l) \equiv \langle \overline{S_i^k} \cdot \overline{S_{i+l}^k} \rangle - \langle \overline{S_i^k} \rangle \langle \overline{S_{i+l}^k} \rangle = \text{Cov}[E(X|i), E(Y|i)], \quad (7)$$

quantifies the correlation of evolutionary rates at pairs of sites along the genome sequence. Note that, due to the circular genomes, the averaging along the genome sequence goes once around the genome, for $i = 0, \dots, L-1$, using the convention that $i+l$ is taken modulo L .

Adapting population model

To simulate an adapting population (Figure 3), we keep track of individuals' fitness, and allow reproduction events to occur with rates proportional to individuals' fitness, choosing the individual that is replaced at each event randomly and uniformly across the population. To model fitness effects, we consider the genome as a series of L "codons," each of which contains one selective site and one neutral site (mimicking aspects of real codons, in which the first 2 bp are under stronger selective pressures than the third one). Mutations at neutral sites occur with rate μ per site, as before. Mutations at selective sites occur at rate μ_s per site, and change the genome's fitness by $\pm s$, where the parameter s is a constant during the simulation, and the randomly chosen sign of s determines whether it is a beneficial or deleterious mutation. Recombination events are modeled as above.

Data availability

The authors state that all data necessary for confirming the conclusions presented in the article are represented fully within the article.

Results

We initially study populations that evolve neutrally in the presence of recombination, both by simulations and theoretical analysis, and later determine how ongoing adaptation would influence the results. We then show that the theory can be used to explain mutational correlations in whole genome data, and can be applied in a straightforward manner to infer recombination rates, and other key population genetic parameters. Table 1 summarizes the key parameters and variables that are used throughout the text.

Effect of homologous recombination on genetic diversity

We simulated neutral dynamics of a population of sequences that mutate and recombine DNA fragments, and measured their average pairwise distance—or population diversity, d —at different recombination rates. We used two different neutral models that yield either KC or BSC trees for a population of size N , using the mutation rate μ (per site per generation), recombination rate γ (per site per generation), and transferred fragment size f_0 . We define the recombination coverage rate, $r \equiv \gamma f_0$, which measures the fraction of the genome that is "covered" by transferred pieces at each generation. We also define the mutational divergence, $\theta \equiv N\mu$, which, under neutral evolution corresponds to the fraction of the genome that accumulated mutations in a pair of sequences since divergence, and the recombination coverage, $\rho \equiv Nr$, the fraction of the genome covered by recombination since divergence (see *Materials and Methods*).

To study how DNA transfer impacts diversity for different population sizes, we performed comparisons at fixed θ by choosing $\mu = \theta/N$. Figure 2A shows a linear decrease in diversity with increasing recombination, as well as the expected collapse of the values of d obtained across different

Table 1 Table of symbols and key relations

Symbol	Description	Relations
N	Population size	
L	Genome length	
μ	Mutation rate, per site	
γ	Recombination rate, per site	
a	Number of alleles at each locus	$a = 4$ for sequence models; $\tilde{a} \equiv a/(a-1)$
$f, \bar{f}$	Fragment size, mean fragment size	
$p(f)$	Probability distribution of fragment sizes	
l	Physical distance between two loci	
r	Recombination coverage rate	$r = \gamma \bar{f}$
$r_1(l)$	One-site coverage rate	$r_1(l) \approx \gamma l$
$r_2(l)$	Two-site coverage rate	$r_2(l) \approx \gamma(\bar{f} - l)$
$t, \bar{t}$	Coalescence time, mean coalescence time	$\bar{t} = N/2$
$\lambda_{b,k}$	Coalescence rate for subset of k out of b individuals	$\lambda_{b,k} = \lambda_{b+1,k} + \lambda_{b+1,k+1}$ (see Appendix B)
θ	Mean mutational divergence	$\theta = 2\bar{t}\mu = N\mu$
ϕ	Mean recombinational divergence	$\phi = 2\bar{t}\gamma = N\gamma$
ρ	Mean recombination coverage	$\rho = \phi \bar{f} = Nr$
$\rho_1(l)$	One-site coverage	$\rho_1(l) = Nr_1(l)$
$\rho_2(l)$	Two-site coverage	$\rho_2(l) = Nr_2(l)$
S_i^k	Substitution variable	$S_i^k \in \{0, 1\}$, with 0 indicating identity and 1 indicating a difference between the k -th pair of sequences at locus i
d	Population diversity; average pairwise distance	$d = \langle S_i^k \rangle$
σ^2	Variance of pairwise distances	$\sigma^2 = \langle S_i^k \rangle^2 - \langle S_i^k \rangle^2$
$c_M(l)$	Mutational correlation function	$c_M(l) \equiv \langle S_i^k S_{i+l}^k \rangle - \langle S_i^k \rangle \langle S_{i+l}^k \rangle$
$c_S(l)$	Structure correlation function	$c_S(l) \equiv \langle S_i^k S_{i+l}^k \rangle - \overline{S_i^k} \cdot \overline{S_{i+l}^k}$
$c_R(l)$	Rate correlation function	$c_R(l) \equiv \langle S_i^k \cdot S_{i+l}^k \rangle - \langle S_i^k \rangle \langle S_{i+l}^k \rangle$
w	Probability that an external one-site transfer affects t	$w = 2\lambda_{3,2}(\lambda_{3,3} + 3\lambda_{3,2})^{-1}$ (see Appendix E and Figure 7)

All times and rates are measured in units of generation time. The notation $\langle S_i^k \rangle$ indicates averaging over all loci i , and $\overline{S_i^k}$ indicates averaging over all sequence pairs k .

population sizes at constant mutational divergence θ . The decrease in diversity occurs because each recombination event replaces a piece of one genome with a homologous piece of another. DNA transfer within a bacterial population can therefore only decrease the number of alleles segregating at any single site on DNA. However, as seen here and shown analytically below, the slope of the decrease in d is very shallow, hence population diversity is insensitive to large fold changes in r .

While a greater recombination rate has only a weak effect on diversity, it has a stronger effect on the distribution of pairwise distances within the population. To see this, we plotted the variance of pairwise distances σ^2 as a function of r (Figure 2A). When recombination coverage rates are low, the population exhibits clusters of similar sequences (Higgs and Derrida 1992), resulting in high values of σ^2 . Increasing r decreases σ^2 due to transfers between clusters, which reduce the differences between clusters, and create an increasingly isotropic distribution of sequences. While d exhibits only moderate decrease as a function of r , σ^2 decreases much more rapidly with r . For fixed θ , the variance exhibits strong dependence on N , which collapses onto a single curve as a function of the recombination coverage ρ (Figure 2A, inset). These results are consistent with homologous recombination acting as an efficient cohesive force within bacterial populations (Hanage *et al.* 2006; Fraser *et al.* 2007, 2009), since a small increase of r is able to destroy clusters while only marginally reducing population diversity.

Genomic correlations and the mutational covariance identity

Inferring transfer rates from population measurements requires more informative quantities than d and σ^2 . The standard population genetic approach is to explore the pattern of linkage disequilibrium (LD) across genomic regions, and relate it to recombination rates (McVean *et al.* 2002; McVean *et al.* 2004; Ansari and Didelot 2014). However, when more than two alleles can segregate at a locus (e.g., in large, diverse populations), minority and majority alleles cannot be unambiguously defined, and LD must be generalized to account for multiple alleles, which becomes mathematically cumbersome. Moreover, variations in substitution rates across the genome can confound LD measurements, and a framework that accounts for rate variations is particularly useful in bacteria.

To address these issues, we introduce three correlation functions that capture different effects of homologous recombination on the structure of genetic diversity in the population (see Figure 1 and *Materials and Methods*). To avoid arbitrarily assigning a reference sequence, or the related issues of majority and minority alleles mentioned above, we consider all possible pairwise comparisons of sequences, which we denote by the substitution variable S_i^k , where k indexes the pair of sequences, and i is the genomic position; if the two sequences are identical at position i , $S_i^k = 0$, otherwise $S_i^k = 1$. We consider any two loci that are l base pairs apart

Figure 2 Simulation results for population diversity and mutational correlations for KC (top row) and BSC (bottom row). Simulations used parameters $N = 1000$, $L = 1000$, $f_0 = 50$, and $\gamma = 10^{-4}$, except where indicated, and all had identical mutational divergence $N\mu = 0.1$. (A) The population diversity (d) and variance (σ^2) in a population are shown as functions of the recombination coverage rate r for three different population sizes: $\bullet 10^2$, $\blacktriangle 10^3$, and $\blacksquare 10^4$. (B, C) Correlation functions are shown for different recombination rates (γ): $\circ 0$, $\triangle 10^{-4}$, and $\square 10^{-3}$. In (C) inset, structure correlation is shown for different population sizes with shapes corresponding to (A). Full and approximate analytical solutions are shown in solid and dashed lines, respectively. (D) Population variance and correlation functions at adjacent sites ($l = 1$) are shown as functions of γ .

in a randomly chosen sequence pair, and denote the substitution variable at these loci by $X = S_i^k$ and $Y = S_{i+l}^k$. We measure the correlation between X and Y across loci within each sequence pair, and then average over all pairs to obtain the mutational correlation function, $c_M(l)$. We measure the correlation between X and Y across the population, and then average over loci to obtain the structure correlation function, $c_S(l)$. Lastly, we compute the average substitution rates at the two loci across the population, and measure the correlation of substitution rates along the genome using the rate correlation function, $c_R(l)$. Figure 1 indicates pictorially how these functions are obtained by correlating or averaging substitution variables conditionally along the sequence or across the population. The tabular structure suggests that a relation should exist among these functions. Indeed, it is easy to check by substitution that

$$c_M(l) + \sigma^2 = c_R(l) + c_S(l) \quad (8)$$

a relation that we call the mutational covariance identity. In Appendix A, we show how this identity follows from the law of total covariance applied to the variables X and Y . Additionally, the non-negativity of σ^2 yields the inequality, $c_M(l) \leq c_S(l) + c_R(l)$.

The three correlation functions are shown for different recombination rates γ in Figure 2, B and C. An important

aspect of these measurements is that, since we simulate neutrally evolving sequences, the correlations exist only at population level, *i.e.*, in the substitution variables that involve pairwise comparisons. Individual sequences in all cases are purely random, uncorrelated sequences of four letters. Since the process of reproduction builds correlations between pairs of sequences, while mutation breaks them, parameters related to reproduction, mutation, and recombination can all influence the shape of correlation functions.

In the absence of recombination, the mutational correlation, c_M , is identically zero, as expected, since mutations occur randomly throughout each sequence. DNA transfer causes pairs of sequences to become identical over random blocks of size f_0 , thereby inducing local correlations that decay as a function of l at a rate determined by γ ; that is, the higher the recombination rate, the faster the decay of $c_M(l)$. However, the dependence of the magnitude of c_M on γ is nonmonotonic: correlation is low for both low and high values of γ , with a maximum at intermediate values (see also Figure 2D), which we discuss below. The qualitative behavior of the rate correlation, c_R , is similar to that of c_M , both as a function of l (Figure 2B, inset) and γ (Figure 2D), but with smaller overall magnitude. By definition, c_R is constructed to detect correlation of substitution rates, and, indeed, despite the fact that mutation rates in simulations are constant across

all loci, DNA transfer reduces diversity over blocks of size f_0 ; hence, it introduces a correlation length scale for substitution rates.

In contrast, the structure correlation, c_S , is positive and constant in the absence of recombination, where its value is determined by drift-mutation balance, *i.e.*, reproduction events build structure correlation while mutations destroy it (Figure 2C). Recombination causes c_S to decay with l at a rate determined by γ . To see why increasing recombination rate reduces the magnitude of structure correlation, we consider two sites separated by distance l . The recombination coverage rate, r , at a single site can be partitioned into the rate $r_1(l) = \gamma l$ at which transferred fragments overlap only the one site but not the other, and the rate $r_2(l) = \gamma(f_0 - l)$ at which fragments span both sites, where $r_1(l) + r_2(l) = r$ (see *Materials and Methods*). For a given pair of sequences, the rate of one-site transfer is $4r_1$, since four sites are involved, while two-site transfers occur between the pair of sequences with rate $(2/N)r_2$. These two-site transfers will annihilate differences between the pair of individuals at both sites and build correlation, while any one-site transfer will break associations and reduce correlation. Since $N \gg 1$, c_S is determined mainly by one-site transfers that destroy correlation, thus it decays monotonically with both l (Figure 2C) and γ (Figure 2D).

We plotted in Figure 2D the dependence of all three correlation functions as a function of the recombination rate γ using the value $l = 1$, *i.e.*, for proximate sites along the genome. The results indicate that the pair c_M and c_R behave qualitatively similarly, as do the pair c_S and σ^2 . These pair relations could be anticipated from the definitions in Equations 5–7, since both c_M and c_R are calculated by taking covariance across the genome sequence and expectation over the population, whereas both c_S and σ^2 involve the (co)variance over the population and the expectation across the genome. In each case, however, covariance and expectation are taken in different orders, leading to four distinct quantities. Taking expectation before covariance reduces the overall magnitude of the correlation since averaging destroys transient correlations, hence $c_R \leq c_M$ and $\sigma^2 \leq c_S$.

Indeed, the difference $c'_M \equiv c_M - c_R$ can be interpreted as measuring the portion of mutational correlations that result from equilibrium fluctuations within the population. By the mutational covariance identity, $c'_M = c_S - \sigma^2$, and, knowing that c_S and σ^2 are monotonically decreasing, and have different half-maximal positions, explains the pronounced peak of c'_M as a function of γ . Lastly, since σ^2 is the variance in pairwise distances, it is determined by the recombination coverage rate, $r = \gamma f_0$, while c_S as discussed above is determined by $r_1(l) = l\gamma$. For this reason, the correlations measured by $c_S(l = 1)$ persist for higher values of γ by a factor of f_0 than the clusters measured by σ^2 .

We note that all of the dependencies that were measured for correlation functions, genetic diversity d , and variance σ^2 shown in Figure 2 were qualitatively very similar in both KC and BSC models. This opens the question of how these func-

tional forms depend quantitatively on the coalescent process and model parameters, which is the subject of the following section.

Dependence of correlation functions and genetic diversity on the rate of recombination

Mathematical analysis of the models presented above was carried out in the context of a general, exchangeable coalescent model defined by a set of coalescence rates $\lambda_{b,k}$, with which any subset of k out of b ancestral lines coalesce into a single common ancestor (see Appendix B). Exact solution for the correlation functions could be obtained numerically as the solution of a system of linear equations with coefficients that are linear combinations of N , μ , γ , $r_1(l)$, $r_2(l)$, and $\lambda_{b,k}$ (see Appendix D). The exact solution is shown in solid lines in all panels of Figure 2, indicating excellent agreement between simulations and calculations.

An analytically tractable form, however, could be obtained only by approximating the full model dynamics. To this end, we introduced a mean-field approximation for the one-site transfer process, which affects the allelic state of exactly one of the substitution variables X and Y of two sites. Instead of accounting for one-site transfers explicitly, as we do in the full model, we approximate their effect as a mutational event that changes the substitution variable to its expected value d (see Appendix E). This approximation yields tractable expressions given below, which are shown as dashed lines in Figure 2. Generally, the agreement between approximate and full solutions is excellent for the KC model, while deviations are more pronounced under the BSC model, though mainly for mutational and rate correlation functions.

The mean-field expression for average pairwise distance is

$$d \approx \frac{\theta}{1 + r + \theta \tilde{a}}, \quad (9)$$

where $\tilde{a} \equiv a/(a - 1)$, indicating that recombination has a very weak effect on the overall diversity (Figure 2A), since r , like μ , is orders of magnitude smaller than 1. Taking $r \ll 1$, the mutational correlation is

$$c_M(l) = \frac{2wd^2\rho_2(l)}{(1 + 2w\rho_1(l) + 2\theta\tilde{a})(1 + 2w\rho + 2\theta\tilde{a})}. \quad (10)$$

where we define the one-site and two-site recombination coverage, respectively $\rho_1 \equiv Nr_1$ and $\rho_2 \equiv Nr_2$, with $\rho_1 + \rho_2 = \rho$, and where $w = 2/3$ (KC) or $w = 1/2$ (BSC). This expression confirms the basic intuition that two-site transfers build mutational correlation, while one-site transfers destroy correlation. Since $\rho_1, \rho_2, \rho \sim \gamma$, it also predicts that c_M initially increases with increasing γ , goes through a maximum, and eventually decays to zero. Taking $l = 1$ above, we compute the value of γ at which c_M is maximized (see Figure 2B), and obtain $\gamma^* = [\mu\tilde{a} + (2wN)^{-1}]/\sqrt{f}$, which has two regimes: if $\theta \ll 1$ then $\gamma^* \sim 1/N$, while if $\theta \gg 1$ then $\gamma^* \sim \mu$. Calculation of the rate correlation yields

$$c_R(l) \approx q c_M(l), \quad (11)$$

where $q < 1$, which verifies that the overall shape of rate correlation is identical to mutational correlation, while its magnitude is smaller (see Appendix E).

The structure correlation takes the form

$$c_S(l) = \frac{d^2}{1 + 2\theta\bar{a} + 2w\rho_1(l)} \quad (12)$$

indicating that one-site transfer reduces correlation, and that two-site transfer plays a negligible role. The principal difference between c_S and c_M is their sensitivity to one- or two-site transfers: c_M is strongly determined by ρ_2 , while c_S is insensitive to it. Lastly, the variance σ^2 can be found by the large l limit of c_S . In this limit, two loci are sufficiently far that a two-site transfer cannot occur; hence, $\rho_2 = 0$ and $\rho_1 = \rho$. Substituting in (12), we find

$$\sigma^2 = \frac{d^2}{1 + 2\theta\bar{a} + 2w\rho}, \quad (13)$$

which confirms our prediction that the values of γ at which σ^2 or c_S are half-maximal (Figure 2B) differ by a factor of f_0 . Moreover, we recapitulate the collapse of curves shown in Figure 2A (inset) as a function of ρ . Importantly, these results indicate that while the effect of recombination on d is negligible compared to mutation (since $r \ll \theta$, Equation 9), its impact on the fluctuations σ^2 and correlations is substantial, and comparable to that of mutation (since $\rho \gtrsim \theta$, Equation 13).

We generalized the model above to account for recombination barriers, which can limit the efficiency of transfer depending on the divergence of donor and recipient sequences (Fraser *et al.* 2007). Assuming that transfer rates decay exponentially as a function of sequence divergence, we simulated and analytically calculated the impact on transfer rates within the population, and derived the relevant correction to the mean-field approximation (see Appendix H). This analysis demonstrates how recombination barriers quantitatively affect the magnitude of correlations in DNA sequences, and is useful in refining intuition about the basic structure of the theory. When applying our method to bacterial genomic datasets, however, our inference procedure described below accounts for recombination barriers in its basic assumptions, and does not require further corrections.

Mutational correlations in adapting populations

While our analysis can be applied for any exchangeable coalescent, genealogies at neutral sites are often linked to non-neutral sites that can undergo selection, which may violate the condition of exchangeability. In particular, in adapting populations in which fitness changes result from specific mutations that modulate individuals' reproductive rates, exchangeability does not hold in general. To assess the magnitude of this effect on our predictions, we simulated an explicit sequence model with fitness effects in which both neutral and non-neutral sites

exist within each genome (see *Materials and Methods*). Correlation functions were computed using the neutral sites, and, in Figure 3, we show the effects of selection on $c_M(l)$. In the presence of DNA transfer, c_M exhibits the characteristic decay that we observed for neutral models, across all values of the selection strength s (Figure 3A). Selection, however, diminishes the magnitude of correlations, leading to a more shallow decay.

We compared the simulation results of $c_M(l)$ for adapting populations with our predictions based on the exact solution for either KC or BSC models. Since the analytical results require knowledge of $\lambda_{2,2}$, which is model-dependent, we infer $\lambda_{2,2}$ from the measured population diversity, d , according to Equation 26. We find that $c_M(l)$ of adapting populations lies between the two curves defined by the KC and BSC solutions (Figure 3A). As the selection strength increases, the adapting population transitions smoothly from the KC to the BSC curve (Figure 3B). This is consistent with previous results that showed how BSC statistics are obtained in the strong selection limit (Neher *et al.* 2013). These limiting coalescence statistics are model independent (*e.g.*, KC describes both Moran and Wright-Fisher models); hence, the KC and BSC statistics are expected to provide the lower and upper bounds on correlation functions in many different models. We verified this to be the case using a substantially different adaptation model, in which fitness was controlled by a single locus (data not shown).

The effect of changing the recombination rate γ on c_M in adapting populations is shown in Figure 3C, which shows a nonmonotonic behavior similar to that seen in neutral populations (Figure 2D). We also observed a smooth transition from BSC to the KC statistics with increasing γ . For low recombination rates, the population is subject to strong selection, and c_M is correctly predicted using the BSC model. As recombination rates increase, recombination decouples the selective loci, breaking linkage between selective sites and their neutral neighbors, and thus reduces the effects of linked-selection on the statistics at neutral sites. Deviations from BSC statistics are more readily detected using σ^2 (Figure 3, D and E), where, for either increasing selection strength or decreasing recombination rate, σ^2 is noticeably lower than the analytical predictions.

Parameter inference with sample selection bias

We have presented calculations of the correlation functions for randomly sampled individuals from a single large population. In principle, their functional forms could now be fit to measured correlations using bacterial sequencing data, provided that the sampled sequences constitute random individuals from the entire relevant population. In reality, there are strong biases associated with most sampling procedures. First, samples can often consist of very closely related strains, particularly when considering pathogen samples from outbreaks, or strain samples from specific geographic locations, which represent only a fraction of the total diversity within a species (Maynard Smith 1991). Second, sequencing of

Figure 3 Mutational correlation (c_M) and population variance (σ^2) in adapting populations. Simulation results are shown in circles, with error bars indicating SEM. Full analytical solutions either based on KC or BSC models are shown in dashed and solid lines, respectively. (A) shows $c_M(l)$ for different values of the selection strength, s . (B–E) show the dependence of $c_M(l=1)$ and σ^2 on s and γ . The simulations used parameters $N = 10^3$, $L = 10^3$, $\mu = 10^{-4}$, $\gamma = 10^{-4}$, $\mu_s = 10^{-6}$, $s = 10^{-2}$, and $f_0 = 50$, except where indicated.

cultured samples involves selection bias for clones that are able to grow on specific media, which further reduces the sampled diversity. Third, different strains are often identified and grouped based on specific markers (e.g., resistance), or phenotypes (e.g., growth on selective media), which may or may not reflect their actual phylogenetic relationships, especially in the context of extensive recombination. Since all of these effects bias the sample composition in largely unknown ways, and thus most samples may not accurately reflect the genetic composition of the bulk population, accurate measurement of population genetic parameters must account for inherent sampling biases. Here, we present an approach to accurately estimate bulk population parameters from biased samples consisting of closely related sequences.

We have shown that recombination reduces only very slightly the diversity of the bulk population (Equation 9 and Figure 2A). However, in a sample of closely related sequences (where relatedness is measure by the coalescent time of the strictly vertical tree of cell divisions), recombination increases the diversity of the sample by importing DNA fragments from more distant sequences. We denote the average diversity of the imported fragments by d , which we take to be larger than the diversity within the vertically inherited portion of the sampled sequences. To calculate the average diversity d_s within the sample, including both vertically inherited and horizontally acquired sequences, we consider a pair of sampled sequences that coalesce at time t ago. Vertically inherited portions have a mutational divergence $\theta_s = 2\mu t$, while recombined portions have an expected divergence of θ , and corresponding diversity d . The recombination coverage for the given pair is $\rho_s = 2rt$, and, as shown in Appendix F, for closely related sequences (i.e., where $\theta_s, \rho_s \ll 1$) that recombine fragments from a much larger population (i.e., where $\rho \gg 1$), the diversity of the sample is given by

$$d_s \approx \rho_s d. \quad (14)$$

The correlation functions c_M , c_S , and c_R in the mean-field approximation can be expressed in terms of the function

$P_2^{(2)}(l)$, which is analytically derived in Appendix E, and takes the form

$$P_2^{(2)}(l) = d^2 \left(\frac{1}{1 + 2\theta\bar{a} + 2w\phi l} + 1 \right) \quad (15)$$

where $\phi \equiv N\gamma$ is the bulk population's recombinational divergence, i.e., the average number of recombination events per site since coalescence. This function expresses the probability that two sites separated by distance l both have a substitution in a randomly chosen pair of individuals, and the above expression is the analytical result for the bulk population. When the same function is computed within a sample of closely related sequences, denoted by $P_{s,2}^{(2)}(l)$, we obtain

$$P_{s,2}^{(2)}(l) \approx (d_s/d)(1 - l/\bar{f})P_2^{(2)}(l), \quad (16)$$

which is accurate for $l \ll \bar{f}$. Remarkably, we see that the quantity $\tilde{P}_{s,2}^{(2)} \equiv P_{s,2}^{(2)}/d_s$ depends only on the bulk population parameters θ and ϕ , and on the mean fragment size $\bar{f}$, and does not involve the sample-specific quantities ρ_s and θ_s . Fitting this functional form therefore allows us to infer the bulk population parameters, despite the biases inherent in natural population sampling.

As shown from simulation results (Figure 4A), the above expression exhibits a crossover between two regimes. For $l \ll \bar{f}$, a substitution at both sites is most likely the result of an external two-site transfer; hence, we have $\tilde{P}_{s,2}^{(2)} \approx \tilde{P}_2^{(2)} \equiv P_2^{(2)}/d$, which exhibits a hyperbolic decay (blue curve). As l increases, the sites are increasingly likely to have experienced a one-site transfer, which results in the linear decrease of $\tilde{P}_{s,2}^{(2)}(l)$, with a slope of $1/\bar{f}$ (red line). We note that the expression computed for the biased sample (16) involves the fragment size $\bar{f}$, while the same expression for the bulk population (15) does not. Intuitively, in a sample of closely related sequences, recombination coverage is small ($\rho_s \ll 1$ per site) because the coalescence time is short; hence, the boundaries of externally transferred fragments

as a function of distance l , with $\gamma = 10^{-4}$. The blue curve is the analytical form given in Equation 15, the red line is $d(1 - l/\bar{f})$, and the horizontal dashed line indicates the bulk population diversity, d . (B) Inferred values of bulk population parameters θ and ϕ , and fragment size $\bar{f}$, are shown relative to their true values (open circles). The diversity of the biased sample, d_s , is shown relative to the population diversity, d (filled circles). For each biased sample of five sequences, $\tilde{P}_{s,2}^{(2)}(l)$ was calculated and fitting to Equation 16 was used to infer θ , ϕ , and $\bar{f}$. Fitting was performed over the range $1 \leq l \leq 50$ using nonlinear least squares (R Core Team 2016).

are relatively clear, and the mean fragment size can be deduced. In the bulk population, coalescence times are long ($\sim N/2$), and recombination coverage is large ($\rho \gg 1$), which results in overlapping fragments reducing the ability to infer fragment sizes. Sampling bias is therefore not exclusively a nuisance, since it presents the opportunity to infer the mean size of transferred fragments.

To test how well one can infer bulk population parameters based on strongly biased samples, for each simulated population we sampled a small number of closely related sequences. As shown in Figure 4B, in these biased samples, the diversity d_s (filled circles) is a small fraction of the bulk population's diversity $d = 0.088$. Depending on the recombination rate, these samples represent from ~ 1 to 10% of the total diversity, i.e., $d_s \approx 0.001 - 0.01$. We calculated $\tilde{P}_{s,2}^{(2)}(l)$ for the samples, and by fitting Equation 16 to the data, we inferred the bulk population parameters θ and ϕ , and $\bar{f}$, which are shown in Figure 4B. These results indicate that over a realistic range of sample diversity and population transfer rates one can accurately infer bulk population parameters using strongly biased samples.

Using correlated mutations to infer recombination rates and global diversity of bacterial populations

We used whole genome sequences from a recently emerged multidrug-resistant *Escherichia coli* clone, sequence type 131 (ST131) containing 185 isolates (Price *et al.* 2013; Petty *et al.* 2014; Ben Zakour *et al.* 2016), and 1216 *Streptococcus pneumoniae* isolates in a longitudinal pneumococcal carriage study (Chewapreecha *et al.* 2014), two species that are well known for horizontal transfer and recombination (Lorenz and Wackernagel 1994). For each dataset, and several subtypes or clades, we calculated sample diversity and correlation functions, and inferred the bulk population parameters (Table 2) by fitting $\tilde{P}_{s,2}^{(2)}(l)$ (see Appendix G for details of sequence analysis). As shown in Figure 5, all correlation functions of synonymous substitutions exhibit a decay pattern in both species, indicating the presence of recombination. The close agreement between measurements (circles) and predictions of the structure and rate correlation functions (solid green and blue lines) indicates that our

Figure 4 Simulation results on inferring bulk population parameters from biased samples of closely related sequences. Simulations used parameters $N = 1000$, $L = 10,000$, $f_0 = 500$, $\mu = 10^{-4}$ and γ ranging from 10^{-4} to 10^{-3} , and a total of 400 populations were simulated for a total time much longer than the coalescent time. For each population, we constructed a biased sample by selecting the cluster of five sequences having the lowest average pairwise coalescent time. (A) Measurement of the sample's $\tilde{P}_{s,2}^{(2)}(l)$ (circles) and the bulk population's $\tilde{P}_2^{(2)}(l) \equiv P_2^{(2)}(l)/d$ (triangles)

model captures the essential features of the underlying transfer process across different clades of both species.

Parameter fitting yielded values of mean fragment size $\bar{f}$, mutational divergence θ , and recombinational divergence ϕ (Table 2). For each sample, we directly measure the sample diversity, d_s , and using the fitted value of θ we calculate the global diversity, $d \approx \theta(1 + \theta\bar{a})^{-1}$ (Equation 9), and the sample's recombination coverage $\rho_s \approx d_s/d$ (Equation 14). From the inferred parameters, we calculate $\rho = \phi\bar{f}$, and use the identity $\rho_s\theta = \rho\theta_s$ to obtain the value of θ_s . The sample's mutational divergence θ_s provides a measure of the age of the sample since coalescence, $\bar{t}$, and can be converted to generations, if the mutation rate μ is known, i.e., $\bar{t} = \theta_s/(2\mu)$. We also calculate the ratio $\phi/\theta = \gamma/\mu$, which gives the relative rate of recombination to mutation in the population. Another useful quantity is the ratio of the number of SNPs that are brought into a sample by recombination to the number of SNPs that are due to *de novo* mutations. This ratio is given by $\rho_s\theta/\theta_s = \rho$, which is the bulk population's recombinational coverage, and thus it does not depend on the sample itself.

Table 2 lists the inferred values of the model parameters. As shown above, our inference method is applicable for closely related samples of sequences, in which $\rho_s, \theta_s \ll 1$ and $\rho \gg 1$; and, indeed, these conditions hold for the inferred values, indicating that the samples consist of closely related sequences that exchange DNA fragments with a much more diverse population. In *E. coli*, when considering all sequences as a single sample, we infer the bulk population's mutational divergence as $\sim 7\%$, while sequences within the sample differ by $\approx 0.2\%$. Yet the sample's mutational divergence, $\theta_s = 1.2 \times 10^{-5}$, indicates that only $\sim 0.001\%$ of the genome has mutated since the coalescence, such that the vast majority of the sample's diversity has been acquired from external transfer events. Similarly, using the sample of all *S. pneumoniae* sequences, the bulk population exhibits 10% divergence, while sequences within the sample differ by 0.8% and the sample's mutational divergence, $\theta_s = 0.01\%$, indicates again that diversity has been acquired largely from external sequences. The high population diversities also imply

that the recombination barrier within species might not be particularly strong, since recombination evidently can still occur between sequences with divergence as large as 10%.

When analyzing separate clades within each species, we infer bulk population divergence levels, θ , that are relatively similar, with $\theta = 0.07 - 0.09$ in *E. coli*, and $\theta = 0.07 - 0.15$ in *S. pneumoniae*, despite large variations in sample diversity among clades. This result is consistent with the separate isolates having access to a single shared gene pool via recombination. Likewise, the average sizes of recombination fragments $\bar{f}$ are consistent across clades in both species, ranging from 0.5 to 2 kb, which is comparable to the length of a typical gene. One exception is clade BC4-6B of *S. pneumoniae*, which has an inferred size of 23.5 kb. It should be noted that our fitting process did not assume any particular distribution for sizes of recombination fragments, provided that l is much smaller than transferred fragment sizes, a condition that is satisfied here since fitting was performed for $l < 150$ bp.

Notably, the bulk population's recombination coverage, ρ , ranges from 73 to 240 for clades of *E. coli*, and from 48 to 2200 for clades of *S. pneumoniae*, indicating that, in each species, two randomly chosen sequences would have recombined so extensively since divergence that transfers would cover each recombining region many times over. As mentioned above, the value of ρ inferred for each sample is equal to the ratio of the number of polymorphisms due to recombination vs. the number of *de novo* mutations. In both species, $\rho \gg 1$, indicating that the diversity introduced by recombination dominates sample diversity, which is consistent with previous results (Price *et al.* 2013; Chewapreecha *et al.* 2014; Petty *et al.* 2014). Yet, since ρ varies substantially from clade to clade, it suggests that real variability exists between subpopulations as far as the portions of the shared gene pool that each is able to access. Indeed, the inferred recombinational divergence, ϕ , shows substantial variation among clades in both species, with $\phi = 0.048 - 0.347$ for clades of *E. coli* and $\phi = 0.02 - 0.84$ for clades of *S. pneumoniae*. Since mutational divergence is relatively consistent across clades, while γ/μ varies widely from 0.7 to 3.87 in *E. coli*, and from 0.65 to 6.55 in *S. pneumoniae*, it appears the variation in ϕ is not related to differences in population size, and likely arises from differences in access to the shared gene pool.

Lastly, we observe large variation across clades of the sample mutational divergence θ_s , which provides a measure of each sample's age since coalescence. Within *S. pneumoniae*, θ_s varies from 6.0×10^{-7} to 4.6×10^{-5} , i.e., nearly two orders of magnitude, while, in *E. coli*, θ_s varies from 1.3×10^{-6} to 1.4×10^{-5} , or one order of magnitude. The entire collection of *S. pneumoniae* sequences is in fact substantially older than any given clade, with $\theta_s = 1.1 \times 10^{-4}$, while in *E. coli* the sample as a whole has $\theta_s = 1.2 \times 10^{-5}$, which is very similar to the age of its oldest clades. These ages can be converted to generations using known values of the mutation rates in each species. For example, $\mu = 2.2 \times 10^{-10}$ in *E. coli* (Lee *et al.* 2012), and, using this value, the age of the entire sample is found to be 55,000 generations.

While generation times cannot easily be assessed in bacteria, for *E. coli*, we conservatively assume one generation per day, using which we estimate that the sampled strains diverged ~ 150 years ago. We note that, since laboratory measurements of μ may not reflect the mutation rates in natural environments—and variation in mutation rates could exist between strains, conditions, and genomic regions—mapping mutational divergence to generations remains an open problem in the field.

Discussion

We presented a mutational correlation-based methodology for quantifying recombination rates in bacterial populations, which we showed can provide accurate estimates using whole-genome data. Our analytical calculation of the correlation functions enables accurate and rapid fitting to infer DNA recombination parameters, as well as a self-consistency check on the model provided by comparing the predicted correlation functions with measurements (see Figure 5). By decomposing the mutational correlation function $c_M(l)$ into a sum of several distinct terms (Equation 8), we provided an intuitive description of how various types of correlations are related within a population, partitioning observed correlations into meaningful components. We showed by simulations and theory how the correlation functions depend on the population genetic parameters (Figure 2).

We compared our inferred parameters to those of previous work. In *E. coli*, we found that the ratio of recombinational to mutational events $\gamma/\mu = 2.1$ inferred by our models is less than that inferred in Touchon *et al.* (2009) (2.5), more than estimated in Didelot *et al.* (2012) (1.0), and much more than that determined by Dixit *et al.* (2015) (0.31). However, as we noted in our clade-by-clade analysis, these values can vary substantially even between clades of the same species. Since previous studies used different sets of strains, it is not surprising that there exists a range of estimates, and, what is remarkable, given the different methodologies and datasets, is that the various methods are all within an order of magnitude of each other. Our approach has the distinct advantage that it is extremely rapid and computationally efficient, and can therefore handle very large datasets with ease. Since our method is formulated using a population genetics model, it provides meaningful connections between a large body of theoretical work, and measurable quantities such as mutational correlation functions. Importantly, we identified sample selection bias, which is prevalent in most datasets, as a potential source of errors in parameter estimation, and by analyzing the correlation functions for samples of closely related sequences, we showed how to account for its effects.

In our analysis of *S. pneumoniae*, using the genomic sequences from Chewapreecha *et al.* (2014), we found the ratio γ/μ ranges from 0.3 to 6.6 across clades, while measured across the set of all clades $\gamma/\mu = 1.4$. These values are larger than the previous estimates, which range from 0.1 to 0.35, which were performed by determining recombination events by contiguous clusters of SNPs (Chewapreecha *et al.* 2014).

Table 2 Best-fit parameters for natural bacterial populations

Species	Clade	Strains	d_s	Best Fit			Calculated			
				θ	ϕ	$\bar{f}$	γ/μ	ρ	θ_s	ρ_s
<i>E. coli</i>	All	185	0.0018	0.071	0.147	990	2.1	150	1.2×10^{-5}	2.8×10^{-2}
	Clade A	15	0.0005	0.069	0.048	1,500	0.7	73	7.2×10^{-6}	8.4×10^{-3}
	Clade B	43	0.0014	0.072	0.148	710	2.1	100	1.4×10^{-5}	2.2×10^{-2}
<i>S. pneumoniae</i>	Clade C (C1 + C2)	120	0.0003	0.090	0.346	700	3.9	240	1.3×10^{-6}	3.9×10^{-3}
	All	1216	0.0080	0.101	0.138	540	1.4	75	1.1×10^{-4}	8.9×10^{-2}
	BC1-19F	365	0.0011	0.140	0.314	720	2.2	230	5.0×10^{-6}	9.6×10^{-3}
	BC2-23F	213	0.0015	0.129	0.843	710	6.6	600	2.6×10^{-6}	1.4×10^{-2}
	BC3-NT (all)	202	0.0043	0.131	0.284	1,800	2.2	510	8.6×10^{-6}	3.9×10^{-2}
	serotype 14	74	0.0009	0.069	0.020	6,100	0.3	120	7.5×10^{-6}	1.4×10^{-2}
	NT isolates	128	0.0050	0.146	0.366	2,000	2.5	730	6.8×10^{-6}	4.0×10^{-2}
	BC4-6B	126	0.0013	0.103	0.095	23,000	0.9	2200	6.0×10^{-7}	1.5×10^{-2}
	BC5-23A/F	106	0.0068	0.096	0.252	590	2.6	150	4.6×10^{-5}	8.0×10^{-2}
	BC6-15B/C	102	0.0017	0.084	0.161	490	1.9	79	2.1×10^{-5}	2.2×10^{-2}
	BC7-14	102	0.0012	0.067	0.043	1,100	0.7	48	2.6×10^{-5}	2.0×10^{-2}

The values of the three fitted parameters, θ , ϕ , and $\bar{f}$ are given for each dataset, and correspond to the best fitting function for $\tilde{P}_{s,2}^{(2)}$ shown in Figure 5 (dashed black line). The mean fragment size $\bar{f}$ is reported in base pairs.

As noted in previous studies (Croucher *et al.* 2011, 2015; Chewapreecha *et al.* 2014), calling recombination events on a locus-by-locus basis is confounded by overlapping recombination events, which often cannot be detected, and is strongly affected by the age of the sample. For this reason, such methods are typically conservative, and their estimates are in effect lower bounds. Despite the order of magnitude difference in our parameter estimates, our method recapitulates the basic biological finding of the previous work, which showed that encapsulated pneumococci have consistently lower rates than nonencapsulated strains. This result is seen in the BC3-NT clade (Table 2), where serotype 14 isolates are encapsulated ($\gamma/\mu = 0.3$), while the NT isolates are nonencapsulated ($\gamma/\mu = 2.5$). In this regard, our analysis attributes a much bigger effect to encapsulation, which we infer reduces recombination rates by more than eightfold, whereas the previous analysis inferred a less than twofold effect.

The distinct correlation signature of recombination in bacterial genomes enables us to infer several additional quantities that are of broad interest for studies of microbial population structure and phylogeny. First, using closely related genome samples, we are able to infer the bulk population's diversity, d , and divergence, θ . Remarkably, our inferred divergence for the bulk population of *E. coli* ($\theta = 0.07$) closely matches the species' diversity measured across a collection of global samples ($\theta = 0.075$; Dixit *et al.* 2015), which provides additional confirmation of the methodology. Second, our approach provides an estimate of the age of each sample without comparison to an outgroup. Existing methods attempt to partition SNPs into internal to a sample (*de novo* mutations) vs. external SNPs (arising from recombination), and thus encounter difficulties with older samples or overlapping recombination fragments. In contrast, our method compares the decay of mutational correlations with the standing diversity of the sample, and uses the predicted quantitative relations to infer the various rates. Without partitioning SNPs, and thus avoiding the known technical difficulties, we are able to infer

the rate of *de novo* mutations in a sample, *i.e.*, its mutational divergence θ_s , and therefore its relative age.

Within the populations assayed here, we find a wide range of sample ages, *e.g.*, the oldest *S. pneumoniae* clade is ~ 80 times the age of the youngest clade. We do not find any meaningful correlation between a sample's age, and the recombination parameters (γ/μ , ρ) or diversity (θ , d) for the bulk population with which it exchanges DNA, indicating that our method correctly models the overall scaling of divergence with time and appropriately removes such effects. Thus, we believe our method provides a reliable tool for comparing sample ages, although such comparisons inherently assume a relatively constant molecular clock, *i.e.*, a value of μ that does not vary substantially between samples of the same species. At the same time, we expect that comparisons between species may be far less meaningful, since neutral evolutionary rates can differ substantially due to basic molecular differences and environmental factors that influence mutation rates.

Finally, given the prevalence of recombination in bacteria, our analysis can be used to quantify the utility (or futility) of inferring a phylogenetic tree for a given sample of DNA sequences. Construction of a tree based on sequence similarity requires some portion of the sequence to contain vertically inherited (*i.e.*, clonal) information. Each recombination event partially degrades the vertical signal in the data, and the recombination coverage ρ_s determines the proportion of the total sequence that has been degraded. Once ρ_s reaches a value of 1, we expect no remaining vertical signal in the data; hence, such a value indicates that phylogenetic reconstruction of the sample would be futile. The values of ρ_s inferred from the samples we analyzed are all relatively low, ranging from ~ 0.1 to 1.0% coverage (Table 2), indicating that there exists a vertical signal in the data. However, reliable tree construction additionally requires a sufficient number of *de novo* mutations that track the vertical signal, quantified by the value of θ_s times the genome size L . For certain samples,

Figure 5 Whole-genome sequence analysis of natural *E. coli* and *S. pneumoniae* isolates. Measured correlations of synonymous substitutions are shown as circles for $\tilde{P}_{s,2}^{(2)}(l)$ (black), as well as $\tilde{c}_M(l)$ (red), $\tilde{c}_R(l)$ (blue), and $\tilde{c}_S(l)$ (green), where $\tilde{c}_x(l) \equiv c_x(l)/d_s$ for $x = M, R, S$. Dashed black line corresponds to the best fit of $\tilde{P}_{s,2}^{(2)}(l)$ using the form given in Equation 16. Parameter values are given in Table 2. The solid colored lines correspond to the predictions of the three correlation functions based on the fit, using the BSC model value of q (see Appendix E). We note that the excellent fit of $\tilde{c}_M(l)$ is not surprising, since this correlation function is determined entirely by $\tilde{P}_{s,2}^{(2)}(l)$, which was fit, while the predictions of $\tilde{c}_R(l)$ and $\tilde{c}_S(l)$ present an independent test of the theory. The dashed green and blue lines correspond to results of fitting q , indicating that deviations from the predictions are due mainly to the choice of coalescent model, which can be inferred and used to improve the prediction. The fitted values of q range from 0.22 to 0.48, where $q = 0.22$ for KC and $q = 0.33$ for BSC coalecscents, indicating that population tree structures often follow KC or BSC statistics, but certain clades may exhibit more general coalescent statistics.

the number of SNPs may be too low to construct a meaningful tree. Additionally, even when a sufficient number of SNPs is expected to be present, one must reliably identify the SNPs that correspond to the clonal signal, which typically constitute a small fraction of the sample diversity, d_s .

For example, for the largest clade of the *E. coli* sequences $\rho_s = 0.004$ (clade C, Table 2), indicating that, for each pair of sequences, >99% of the genome has been vertically inherited, which seems like good news for tree building. However, $\theta_s = 1.3 \times 10^{-6}$ for this sample, meaning that between any pair of sequences <10 single nucleotide differences correspond to the vertical signal. These few crucial differences must be recognized from among >1000 differ-

ences that exist between any pair of individuals in the sample, the majority of which are due to recombination events. Finding these few informative needles in the haystack of SNPs is the major challenge for sequence-based phylogeny, and, while the problem may be quite difficult, at least in such cases one is relatively certain that a vertical signal exists in the data.

Much more severe, and possibly insurmountable, problems exist in establishing reliably the phylogeny of sequences from the bulk population of a bacterial species, where the relevant recombinational coverage is measured by ρ . Across the populations that we analyzed, $\rho = 48 - 2200$, indicating that, for any pair of sequences sampled from the bulk population, their genomes have been covered many times over

by recombinational transfers since coalescence. We expect no vertical signal would remain in the data. Since the value of ρ represents a genome-wide average, it is of course possible that some genomic regions could exhibit extremely low recombination rates, and thus have effectively much lower values of ρ , which may enable phylogenetic inference in the bulk population, assuming $\rho \ll 1$ for these regions. In the best case scenario, one would again be faced with identifying a tiny fraction of SNPs that capture the vertical signal, only now in a much smaller portion of the genome. Our results therefore indicate that phylogenetic inference in bacteria may, in most cases, be an extremely challenging problem, and our methodology provides the basic measures that quantify the (f)utility of tree building in any given case.

In conclusion, our new approach for inferring within-population recombination rates, based on correlation functions, provides a framework for measurements in a wide range of sequencing datasets. Further generalization of our method could incorporate variability in recombination rates across genomes, as well as modeling explicitly population subdivision and community structures. We expect that this framework, and its generalizations, will find fruitful application in inferring recombination rates within species, as well as providing a useful starting point for analyses of much more complex sets of data such as metagenomic samples.

Acknowledgments

We wish to thank Sergei Maslov, Erik van Nimwegen, Jane Carlton, Alexander Grosberg, Charles Peskin, Matthew Rockman, Wei-Hsiang Lin, and Long Qian for valuable discussions, as well as three anonymous referees for their comments on the manuscript. This work was funded by a Human Frontier Science Program Young Investigators' grant.

Literature Cited

- Andam, C. P., and J. P. Gogarten, 2011 Biased gene transfer in microbial evolution. *Nat. Rev. Microbiol.* 9: 543–555.
- Ansari, M. A., and X. Didelot, 2014 Inference of the properties of the recombination process from whole bacterial genomes. *Genetics* 196: 253–265.
- Ben Zakour, N. L., A. S. Alsheikh-Hussain, M. M. Ashcroft, N. T. Khanh Nhu, L. W. Roberts *et al.*, 2016 Sequential acquisition of virulence and fluoroquinolone resistance has shaped the evolution of *Escherichia coli* ST131. *MBio* 7: e00347–e00416.
- Bolthausen, E., and A. Sznitman, 1998 On Ruelle's probability cascades and an abstract cavity method. *Commun. Math. Phys.* 197: 247–276.
- Brunet, E., B. Derrida, A. Mueller, and S. Munier, 2007 Effect of selection on ancestry: an exactly soluble case and its phenomenological generalization. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 76: 041104.
- Brunet, E., B. Derrida, and D. Simon, 2008 Universal tree structures in directed polymers and models of evolving populations. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 78: 061102.
- Chewapreecha, C., S. R. Harris, N. J. Croucher, C. Turner, P. Marttinen *et al.*, 2014 Dense genomic sampling identifies highways of pneumococcal recombination. *Nat. Genet.* 46: 305–309.
- Cohen, E., D. A. Kessler, and H. Levine, 2005 Recombination dramatically speeds up evolution of finite populations. *Phys. Rev. Lett.* 94: 098102.
- Croucher, N. J., S. R. Harris, C. Fraser, M. A. Quail, J. Burton *et al.*, 2011 Rapid pneumococcal evolution in response to clinical interventions. *Science* 331: 430–434.
- Croucher, N. J., A. J. Page, T. R. Connor, A. J. Delaney, J. A. Keane *et al.*, 2015 Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* 43: e15.
- Desai, M. M., A. M. Walczak, and D. S. Fisher, 2013 Genetic diversity and the structure of genealogies in rapidly adapting populations. *Genetics* 193: 565–585.
- Didelot, X., and D. Falush, 2007 Inference of bacterial microevolution using multilocus sequence data. *Genetics* 175: 1251–1266.
- Didelot, X., D. Lawson, A. Darling, and D. Falush, 2010 Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics* 186: 1435–1449.
- Didelot, X., G. Méric, D. Falush, and A. E. Darling, 2012 Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. *BMC Genomics* 13: 256.
- Dixit, P. D., T. Y. Pang, F. W. Studier, and S. Maslov, 2015 Recombinant transfer in the basic genome of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* 112: 9070–9075.
- Feingold, D. G., and R. S. Varga, 1962 Block diagonally dominant matrices and generalizations of the Gerschgorin circle theorem. *Pac. J. Math.* 12: 1241–1250.
- Fraser, C., W. P. Hanage, and B. G. Spratt, 2007 Recombination and the nature of bacterial speciation. *Science* 315: 476–480.
- Fraser, C., E. J. Alm, M. F. Polz, B. G. Spratt, and W. P. Hanage, 2009 The bacterial species challenge: making sense of genetic and ecological diversity. *Science* 323: 741–746.
- Garrison, E., and G. Marth, 2012 Haplotype-based variant detection from short-read sequencing. *arXiv*: 1207.3907 [q-bio.GN].
- Hanage, W. P., B. G. Spratt, K. M. E. Turner, and C. Fraser, 2006 Modelling bacterial speciation. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 361: 2039–2044.
- Higgs, P. G., and B. Derrida, 1992 Genetic distance and species formation in evolving populations. *J. Mol. Evol.* 35: 454–465.
- Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* 23: 183–201.
- Kingman, J., 1982 On the genealogy of large populations. *J. Appl. Probab.* 19: 27–43.
- Koonin, E. V., and Y. I. Wolf, 2008 Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.* 36: 6688–6719.
- Kuzminov, A., 1999 Recombinational repair of DNA damage in *Escherichia coli* and bacteriophage λ . *Microbiol. Mol. Biol. Rev.* 63: 751–813.
- Lee, H., E. Popodi, H. Tang, and P. L. Foster, 2012 Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. *Proc. Natl. Acad. Sci. USA* 109: E2774–E2783.
- Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The sequence alignment/map format and SAMtools. *Bioinformatics* 25: 2078–2079.
- Lorenz, M. G., and W. Wackernagel, 1994 Bacterial gene transfer by natural genetic transformation in the environment. *Microbiol. Rev.* 58: 563–602.
- Marttinen, P., W. P. Hanage, N. J. Croucher, T. R. Connor, S. R. Harris *et al.*, 2012 Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res.* 40: e6.

- Maynard Smith, J., 1991 The population genetics of bacteria. *Proc. Biol. Sci.* 245: 37–41.
- McVean, G., P. Awadalla, and P. Fearnhead, 2002 A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 160: 1231–1241.
- McVean, G. A., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley *et al.*, 2004 The fine-scale structure of recombination rate variation in the human genome. *Science* 304: 581–584.
- Moran, P. A. P., 1958 Random processes in genetics. *Math. Proc. Camb. Philos. Soc.* 54: 60–71.
- Neher, R. A., and O. Hallatschek, 2013 Genealogies of rapidly adapting populations. *Proc. Natl. Acad. Sci. USA* 110: 437–442.
- Neher, R. A., B. I. Shraiman, and D. S. Fisher, 2010 Rate of adaptation in large sexual populations. *Genetics* 184: 467–481.
- Neher, R. A., T. A. Kessinger, and B. I. Shraiman, 2013 Coalescence and genetic diversity in sexual populations under selection. *Proc. Natl. Acad. Sci. USA* 110: 15836–15841.
- Oren, Y., M. B. Smith, N. I. Johns, M. K. Zeevi, D. Biran *et al.*, 2014 Transfer of noncoding DNA drives regulatory rewiring in bacteria. *Proc. Natl. Acad. Sci. USA* 111: 16112–16117.
- Petty, N. K., N. L. Ben Zakour, M. Stanton-Cook, E. Skippington, M. Totsika *et al.*, 2014 Global dissemination of a multidrug resistant *Escherichia coli* clone. *Proc. Natl. Acad. Sci. USA* 111: 5694–5699.
- Pitman, J., 1999 Coalescents with multiple collisions. *Ann. Probab.* 27: 1870–1902.
- Price, L. B., J. R. Johnson, M. Aziz, C. Clabots, B. Johnston *et al.*, 2013 The epidemic of extended-spectrum- β -lactamase-producing *Escherichia coli* ST131 is driven by a single highly pathogenic subclone, H30-Rx. *MBio* 4: e00377–e00413.
- R Core Team, 2016 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Ran, R.-S., and T.-Z. Huang, 2006 The inverses of block tridiagonal matrices. *Appl. Math. Comput.* 179: 243–247.
- Ravenhall, M., N. Škunca, F. Lassalle, and C. Dessimoz, 2015 Inferring horizontal gene transfer. *PLoS Comput. Biol.* 11: e1004095.
- Rosen, M. J., M. Davison, D. Bhaya, and D. S. Fisher, 2015 Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. *Science* 348: 1019–1023.
- Sagitov, S., 1999 The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Probab.* 36: 1116–1125.
- Schweinsberg, J., 2000 Coalescents with simultaneous multiple collisions. *Electron. J. Probab.* 5: 1–50.
- Schweinsberg, J., 2012 Dynamics of the evolving Bolthausen–Sznitman coalescent. *Electron. J. Probab.* 17: 91.
- Shapiro, B. J., J. Friedman, O. X. Cordero, S. P. Preheim, S. C. Timberlake *et al.*, 2012 Population genomics of early events in the ecological differentiation of bacteria. *Science* 336: 48–51.
- Soucy, S. M., J. Huang, and J. P. Gogarten, 2015 Horizontal gene transfer: building the web of life. *Nat. Rev. Genet.* 16: 472–482.
- Thomas, C. M., and K. M. Nielsen, 2005 Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* 3: 711–721.
- Touchon, M., C. Hoede, O. Tenaillon, V. Barbe, S. Baeriswyl *et al.*, 2009 Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet.* 5: e1000344.
- Vos, M., and X. Didelot, 2009 A comparison of homologous recombination rates in bacteria and archaea. *ISME J.* 3: 199–208.
- Weissman, D., and N. H. Barton, 2012 Limits to the rate of adaptive substitutions in sexual populations. *PLoS Genet.* 8: e1002740.
- Weissman, D., and O. Hallatschek, 2014 The rate of adaptation in large sexual populations with linear chromosomes. *Genetics* 196: 1167–1183.
- Williams, D., J. P. Gogarten, and R. T. Papke, 2012 Quantifying homologous replacement of loci between haloarchaeal species. *Genome Biol. Evol.* 4: 1223–1244.
- Wiuf, C., 2000 A coalescence approach to gene conversion. *Theor. Popul. Biol.* 57: 357–367.
- Wiuf, C., and J. Hein, 2000 The coalescent with gene conversion. *Genetics* 155: 451–462.

Communicating editor: J. Lawrence

Appendix

A. Covariance Decomposition and the Law of Total Covariance

The law of total covariance states that

$$\text{Cov}(X, Y) = E\left(\text{Cov}(X|Z, Y|Z)\right) + \text{Cov}\left(E(X|Z), E(Y|Z)\right) \quad (17)$$

As in *Materials and Methods*, we define random variables for two sites separated by distance l as $X = S_i^k$ and $Y = S_{i+l}^k$, where i is a site, and k is a sequence pair. We note that $E(X|k) = E(Y|k)$, since the expectation is taken over the same set of sites in both cases. The expression $\sigma^2 = \text{Var}(E(S|k))$ can therefore be written equivalently as $\sigma^2 = \text{Cov}(E(X|k), E(Y|k))$. Together with Equation 5 and the law of total covariance (setting $Z = k$), we have

$$c_M + \sigma^2 = \text{Cov}(X, Y). \quad (18)$$

Similarly, using Equations 6 and 7 (setting $Z = i$), we obtain

$$c_S + c_R = \text{Cov}(X, Y). \quad (19)$$

The total covariance, $c_T(l) \equiv \text{Cov}(X, Y)$, is the covariance of the substitution variables at any pair of sites separated by distance l (i.e., over all sites and all sequence pairs). It can be computed either by conditioning on the sequence pair or by conditioning on the site; either way the result is of course the same, which yields the covariance decomposition

$$c_M + \sigma^2 = c_S + c_R. \quad (20)$$

While such a decomposition would be possible for any two-dimensional arrangement of random variables (e.g., magnetic spins on a lattice), one interesting feature in our population genetics application is that a variance appears on the left-hand side rather than a covariance. This happens because the random variables $E(X|k)$ and $E(Y|k)$ turn out to be identical, which is due to the circular genome. The same result also holds to excellent approximation for a very large genome, with corrections of order $1/L$. Therefore we obtain the mutational covariance inequality

$$c_S + c_R \geq c_M. \quad (21)$$

B. Coalescence Rates

The general exchangeable coalescent model is defined by a set of coalescence rates, $\lambda_{b,k}$, with which any subset of k out of b ancestral lines coalesce into a single common ancestor. These rates satisfy a self-consistency relation $\lambda_{b,k} = \lambda_{b+1,k} + \lambda_{b+1,k+1}$ (see Pitman 1999; Sagitov 1999; Schweinsberg 2000; Brunet *et al.* 2008). For the Moran model, we have for all $b \geq 2$, $\lambda_{b,2} = G/\binom{N}{2}$, since $\binom{N}{2}$ is the total number of pairs that could coalesce; and, since only pairwise mergers are possible, $\lambda_{b,k} = 0$ for all $k > 2$ (Moran 1958). For the Schweinsberg model, we obtain

$$\lambda_{2,2} = G \sum_{U=2}^N p(U) \cdot \binom{U}{2} / \binom{N}{2} = G/(N-1), \quad (22)$$

and, since its genealogy follows the BSC, we have

$$\lambda_{b,k}/\lambda_{2,2} = \frac{(k-2)!(b-k)!}{(b-1)!}. \quad (23)$$

For convenience, we take $G = N - 1$ for the Moran model, and $G = 2(N - 1)/N$ for the Schweinsberg model, so that both models have identical $\lambda_{2,2} = 2/N$. Setting G arbitrarily only changes the unit of all times, and does not change any results below. Explicit coalescence computations are not made for the adapting population model; instead, simulation results in the limits of weak and strong selection are compared with theoretical predictions of the KC and BSC models. This requires matching the overall coalescence rates of simulations with theory, which we accomplish by measuring $\lambda_{2,2}$ directly in simulations.

C. Calculation of Population Diversity

We consider substitutions at a single site on a randomly chosen pair of genomic sequences $\vec{g}$ and $\vec{g}'$ as shown in Figure 6A, and analyze the dynamics of $\vec{P}^{(1)} \equiv (p, 1-p)^T$, where p is the probability that these two sequences are identical at that site. We trace their history back one instant Δt , and compute the rates and outcomes of different events that affect the sequences, including coalescence due to reproduction, DNA transfer, and mutation.

First, a coalescence event in which one of $\vec{g}$ or $\vec{g}'$ reproduces and replaces the other yields identity at the site, indicated by a star in Figure 6A corresponding to the state $\vec{P}^{(0)} \equiv (1, 0)^T$, and occurs with rate $\lambda_{2,2}$. The change in the state vector is then $\Delta \vec{P} = \vec{P}^{(0)} - \vec{P}^{(1)}$. Second, an internal DNA transfer, which occurs between the pair of sequences in question (Figure 6A), yields the same result as coalescence, and thus the same $\Delta \vec{P}$. Indeed, such an internal DNA transfer is analogous to a reproduction event in the sense that a piece of DNA “reproduces” from the donor sequence, and replaces its counterpart in the recipient sequence. The rate of such events is $2r/N$. An external DNA transfer, on the other hand, in which the piece of DNA is copied from a third sequence $\vec{g}'' \neq \vec{g}$ or $\vec{g}'$, does not change the probability that the two sequences are identical at the site, and thus does not change the state vector, i.e., $\Delta \vec{P} = 0$. Third, we assume mutations occur independently at each site, and uniformly along the sequence. A mutation causes a change $\Delta \vec{P}$ that can be represented as a linear operator $M\Delta t$ acting on $\vec{P}^{(1)}$. The overall rate for mutations for two sites is 2μ , and the linear operator is given by

$$M = 2\mu \begin{pmatrix} -1 & (a-1)^{-1} \\ 1 & -(a-1)^{-1} \end{pmatrix}. \quad (24)$$

Summing the possible changes ΔP due to each event multiplied by their rates, we find

$$\frac{d\vec{P}^{(1)}}{dt} = M\vec{P}^{(1)} + (\lambda_{2,2} + 2r/N) \left(\vec{P}^{(0)} - \vec{P}^{(1)} \right). \quad (25)$$

Setting the derivative to zero, we obtain the steady-state solution for $d \equiv 1 - p$:

$$d = \frac{2\mu}{\lambda_{2,2} + 2r/N + 2\mu\tilde{a}} = \frac{\theta}{1 + r + \theta\tilde{a}}, \quad (26)$$

where $\tilde{a} \equiv a/(a-1)$. For the case $r = 0$, this yields the well-known expression for heterozygosity in the Moran model.

D. Calculation of Correlation Functions

We consider a pair of sites, a and b , separated by distance l , in a randomly chosen pair of genomic sequences, $\vec{g}$ and $\vec{g}'$. The sequences can differ at 0, 1, or 2 sites, and we write the probability distribution of these three possible states as $\vec{P} = (P_0, P_1, P_2)^T$, where the components are non-negative numbers summing to 1. More generally, a pair of sites can be configured across 2, 3, or 4 sequences, as shown in Figure 6B. Due to the effect of genetic linkage, the distribution of states $\vec{P}$ depends on the configuration. We therefore denote by $\vec{P}^{(i)}$ the distribution of states for $i = 2, 3$, or 4 randomly chosen sequences.

The correlation functions c_M , c_S , and c_R , as well as σ^2 are determined by the values of four different terms (see Equations 4–7):

$$\overline{\langle S_i^k S_{i+l}^k \rangle} = P_2^{(2)}(l); \quad (27)$$

$$\overline{\langle S_i^k \cdot S_{i+l}^k \rangle} = \frac{1}{n_2} P_2^{(2)}(l) + \frac{2 \binom{N-2}{1}}{n_2} P_2^{(3)}(l) + \frac{\binom{N-2}{2}}{n_2} P_2^{(4)}(l); \quad (28)$$

$$\overline{\langle S_i^k \rangle \langle S_{i+l}^k \rangle} = \frac{1}{L} \sum_{l=1}^L \overline{\langle S_i^k S_{i+l}^k \rangle}; \quad (29)$$

Figure 6 Illustration of possible transitions between configurations for one or two pairs of sites. Each horizontal line represents a single sequence, and small vertical lines represent sites. (A) shows the possible transitions and events for one pair of sites. When the pair become identical due to an indicated transition, we denote the site by a star. (B) shows different configurations of two pairs of sites, and the coalescent states in one or both sites, and their possible transitions and events are shown in (C–E). Transitions between configurations are summarized in (F), where solid arrows represent transitions due to reproduction, internal one-site transfer, or two-site transfer, and dashed arrows correspond to an external one-site transfer into a sequence with two sites.

$$\langle \overline{S_i^k} \rangle \langle \overline{S_{i+l}^k} \rangle = \frac{1}{L} \sum_{l=1}^L \langle \overline{S_i^k \cdot S_{i+l}^k} \rangle, \quad (30)$$

where $n_2 = \binom{N}{2}$. For large N , we obtain

$$\langle \overline{S_i^k \cdot S_{i+l}^k} \rangle \approx P_2^{(4)}(l). \quad (31)$$

Given that the maximal size of transferred segments, $f_{\max}$, is far smaller than the genomic length ($f_{\max} \ll L$), $\langle \overline{S_i^k} \rangle \langle \overline{S_{i+l}^k} \rangle$ is determined largely by the flat tail of $P_2^{(2)}(l)$ for $l \geq f_{\max}$, which we denote by $P_2^{(2)}(L)$:

$$\langle \overline{S_i^k} \rangle \langle \overline{S_{i+l}^k} \rangle = \frac{1}{L} \sum_{l=0}^{f_{\max}} P_2^{(2)}(l) + \frac{L - f_{\max}}{L} P_2^{(2)}(L) \approx P_2^{(2)}(L). \quad (32)$$

This corresponds to loci that are sufficiently far such that they are not affected by two-site transfers, which means $r_2 = 0$ and therefore $r_1 = r$. Similar approximation yields

$$\langle \overline{S_i^k} \rangle \langle \overline{S_{i+l}^k} \rangle \approx P_2^{(4)}(L). \quad (33)$$

Based on Equations 4–7, the calculations of c_M , c_S , c_R , and σ^2 are reduced to the analysis of $\vec{P}^{(i)}$:

$$c_M(l) = P_2^{(2)}(l) - P_2^{(2)}(L); \quad (34)$$

$$c_S(l) = P_2^{(2)}(l) - P_2^{(4)}(l); \quad (35)$$

$$c_R(l) = P_2^{(4)}(l) - P_2^{(4)}(L); \quad (36)$$

$$\sigma^2 = P_2^{(2)}(L) - P_2^{(4)}(L) = c_S(L). \quad (37)$$

To analyze the dynamics of $\vec{P}^{(i)}$, we randomly choose i sequences, trace their history back one instant Δt , and compute the rates and outcomes of different events that could affect the sequences. These events include coalescence by reproduction, DNA transfers, and mutations. Mutations have the same effect on $\vec{P}^{(i)}$, for $i = 2, 3, 4$, as follows. Since two pairs of sites are involved, the total rate of mutations is 4μ , and the linear operator of mutations acting on $\vec{P}^{(i)}$ can be written as follows:

$$M = 4\mu M_\mu, \quad (38)$$

where

$$M_\mu \equiv \begin{pmatrix} -1 & \frac{\tilde{a}}{2a} & 0 \\ 1 & -\frac{\tilde{a}}{2} & \frac{\tilde{a}}{a} \\ 0 & \frac{1}{2} & -\frac{\tilde{a}}{a} \end{pmatrix}. \quad (39)$$

Dynamics equation for $\vec{P}^{(2)}$

Possible events that could affect $\vec{P}^{(2)}(t)$ beside mutations are shown in Figure 6C. Coalescence events occur at rate $\lambda_{2,2}$, causing identity at both sites, and the state vector becomes $\vec{P}^{(0)} \equiv (1, 0, 0)^T$; hence, $\Delta \vec{P} = \vec{P}^{(0)} - \vec{P}^{(2)}$. For DNA transfer events, as shown in Figure 6C, we need to consider (i) one-site internal transfers, (ii) one-site external transfers, (iii) two-site internal transfers, and (iv) two-site external transfer. (i) A one-site internal transfer, which occurs at rate $4r_1/N$, results in a coalescence at one site, while leaving the other unchanged. The state vector becomes $\vec{P}^{(1)} \equiv (1-d, d, 0)^T$, where d is the

probability that two sequences differ at the unchanged site; hence, $\Delta \vec{P} = \vec{P}^{(1)} - \vec{P}^{(2)}$. (ii) An external one-site transfer creates a configuration that involves three different sequences, and the state vector becomes $\vec{P}^{(3)}$, thus $\Delta \vec{P} = \vec{P}^{(3)} - \vec{P}^{(2)}$. This occurs with rate $4r_1(N-2)/N$, since each site can receive DNA from any of the $N-2$ sequences not in the given pair. (iii) A two-site internal transfer causes identity at both sites, i.e., $\Delta \vec{P} = \vec{P}^{(0)} - \vec{P}^{(2)}$. Lastly, (iv) the two-site external transfer does not change the state vector, $\Delta \vec{P} = 0$, since sequence labels are exchangeable. Multiplying the possible changes by their rates yields

$$\frac{d\vec{P}^{(2)}}{dt} = M\vec{P}^{(2)} + \left(\lambda_{2,2} + \frac{2r_2}{N}\right)\vec{P}^{(0)} + \frac{4r_1(N-2)}{N}\vec{P}^{(3)} + \frac{4r_1}{N}\vec{P}^{(1)} - \left(\lambda_{2,2} + \frac{4r_1(N-1) + 2r_2}{N}\right)\vec{P}^{(2)}. \quad (40)$$

Dynamics equation for $\vec{P}^{(3)}$

Possible events that could affect $\vec{P}^{(3)}$ are illustrated in Figure 6D. Coalescence can affect the configuration in three different ways: (i) all three sequences could coalesce simultaneously, yielding identity at both sites; hence, $\Delta \vec{P} = \vec{P}^{(0)} - \vec{P}^{(3)}$; (ii) the two-site sequence ($\vec{g}$), and one of the two single-site sequences ($\vec{g}'$ or $\vec{g}''$), could coalesce, causing identity at one site, and leaving the other site unchanged; the state vector becomes $\vec{P}^{(1)}$, and $\Delta \vec{P} = \vec{P}^{(1)} - \vec{P}^{(3)}$; or (iii) the two single-site sequences ($\vec{g}'$ and $\vec{g}''$) could coalesce onto an ancestral sequence that carries both sites; hence, the state vector becomes $\vec{P}^{(2)}$ and $\Delta \vec{P} = \vec{P}^{(2)} - \vec{P}^{(3)}$. The rates for these events are $\lambda_{3,3}$, $2\lambda_{3,2}$, and $\lambda_{3,2}$, respectively.

Since an internal transfer involves choosing two sequences (one as the donor and the other as the recipient), there exist two different types of transfers (Figure 6D): ones between the two-site sequence ($\vec{g}$) and one of the two single-site sequences ($\vec{g}'$ or $\vec{g}''$), and the others between the two single-site sequences ($\vec{g}'$ and $\vec{g}''$). The results of these transfers are exactly the same as those of coalescing two sequences discussed above (ii and iii, see above), such that they change the state vector to $\vec{P}^{(1)}$ and $\vec{P}^{(2)}$, and occur with rates $4r/N$ and $2r/N$, respectively. An external one-site transfer from an external sequence to one of the two single-site sequences changes only the sequence labels, but not the configuration, thus $\Delta \vec{P} = 0$. On the other hand, when an external one-site transfer occurs into the two-site sequence, the configuration becomes that of $\vec{P}^{(4)}$; hence, $\Delta \vec{P} = \vec{P}^{(4)} - \vec{P}^{(3)}$, with rate $2r_1(N-3)/N$. Multiplying rates by changes, we find

$$\frac{d\vec{P}^{(3)}}{dt} = M\vec{P}^{(3)} + \left(\frac{2r}{N} + \lambda_{3,2}\right)\left(2\vec{P}^{(1)} + \vec{P}^{(2)} - 3\vec{P}^{(3)}\right) + \lambda_{3,3}\left(\vec{P}^{(0)} - \vec{P}^{(3)}\right) + \frac{2r_1(N-3)}{N}\left(\vec{P}^{(4)} - \vec{P}^{(3)}\right). \quad (41)$$

Dynamics equation for $\vec{P}^{(4)}$

Possible events that could affect $\vec{P}^{(4)}$ are illustrated in Figure 6E. Since four sequences are involved, each of which carries only one site, four different types of coalescent events are possible: (i) simultaneous coalescence of all four sequences, which yield identity at both sites and $\Delta \vec{P} = \vec{P}^{(0)} - \vec{P}^{(4)}$, (ii) simultaneous coalescence of three sequences, producing identity at one of the two sites; hence, $\Delta \vec{P} = \vec{P}^{(1)} - \vec{P}^{(4)}$, (iii) coalescence of two sequences with sites at the same locus, yielding identity at one of the two sites; hence, $\Delta \vec{P} = \vec{P}^{(1)} - \vec{P}^{(4)}$, and (iv) coalescence of two sequences with two different sites, leading to an ancestral sequence that carries both sites; the state vector then becomes $\vec{P}^{(3)}$, and $\Delta \vec{P} = \vec{P}^{(3)} - \vec{P}^{(4)}$. The rates of these events are $\lambda_{4,4}$, $4\lambda_{4,3}$, $2\lambda_{4,2}$, and $4\lambda_{4,2}$, respectively.

For an internal DNA transfer, the two chosen sequences may carry their single site either at the same locus, or at different loci, and, therefore, there are two types of internal DNA transfers (Figure 6E). The results of these two types of transfer on the state vector are exactly the same as for coalescence of two sequences discussed in types (iii) and (iv) above. The rates are $4r/N$ and $8r/N$, respectively. Since in the configuration of $\vec{P}^{(4)}$, each of the four sequences contains only one site and sequence labels are exchangeable, external transfers do not change the state vector, $\Delta \vec{P} = 0$. Multiplying these rates by changes, we find

$$\frac{d\vec{P}^{(4)}}{dt} = M\vec{P}^{(4)} + \lambda_{4,4}\left(\vec{P}^{(0)} - \vec{P}^{(4)}\right) + 4\lambda_{4,3}\left(\vec{P}^{(1)} - \vec{P}^{(4)}\right) + (4r/N + 2\lambda_{4,2})\left(\vec{P}^{(1)} + 2\vec{P}^{(3)} - 3\vec{P}^{(4)}\right). \quad (42)$$

Exact solution for steady-state values of $\vec{P}^{(i)}$

Combining the above, we express the equations as a system of nonhomogeneous coupled linear differential equations:

$$\frac{d}{dt} \begin{bmatrix} \vec{P}^{(2)} \\ \vec{P}^{(3)} \\ \vec{P}^{(4)} \end{bmatrix} = (S \otimes I + I \otimes M) \begin{bmatrix} \vec{P}^{(2)} \\ \vec{P}^{(3)} \\ \vec{P}^{(4)} \end{bmatrix} + \vec{B}^{(0)} \otimes \vec{P}^{(0)} + \vec{B}^{(1)} \otimes \vec{P}^{(1)} \quad (43)$$

in which I is a 3×3 identity matrix, $\vec{B}^{(0)}$ and $\vec{B}^{(1)}$ describe transitions from $\vec{P}^{(2)}$, $\vec{P}^{(3)}$, and $\vec{P}^{(4)}$ to $\vec{P}^{(0)}$ and $\vec{P}^{(1)}$, respectively, with $\vec{P}^{(0)} = (1, 0, 0)^T$, $\vec{P}^{(1)} = ((1-d), d, 0)^T$, $\vec{B}^{(0)} = (\lambda_{2,2} + 2r_2/N, \lambda_{3,3}, \lambda_{4,4})^T$, and $\vec{B}^{(1)} = (4r_1/N, 2\lambda_{3,2} + 4r/N, 4\lambda_{4,3} + 2\lambda_{4,2} + 4r/N)^T$; and S is the operator for reproduction and transfer, which has the following form:

$$S = \begin{pmatrix} -\frac{2r_2 + 4r_1(N-1)}{N} - \lambda_{2,2} & \frac{4r_1(N-2)}{N} & 0 \\ \frac{2r}{N} + \lambda_{3,2} & -\left(\frac{6r_2}{N} + 2r_1\right) - (3\lambda_{3,2} + \lambda_{3,3}) & 2r_1 \frac{N-3}{N} \\ 0 & \frac{8r}{N} + 4\lambda_{4,2} & -\frac{12r}{N} - (6\lambda_{4,2} + 4\lambda_{4,3} + \lambda_{4,4}) \end{pmatrix}. \quad (44)$$

The tridiagonal form of S leads to a block tridiagonal matrix of the Kronecker product, $S \otimes I + I \otimes M$, which can be written in a general form

$$A = \begin{pmatrix} A_1 & B_1 & 0 \\ D_1 & A_2 & B_2 \\ 0 & D_2 & A_3 \end{pmatrix}. \quad (45)$$

In the block matrix A , the off-diagonal blocks are scalar matrices, and the diagonal blocks are sums of scalar matrices and the mutation operator, M , which is diagonalizable. Importantly, A is a block diagonally dominant matrix in the norm induced by the l_1 -norm, which guarantees that A is nonsingular (Feingold and Varga 1962). The inverse of A can be computed efficiently using an algorithm given by (Ran and Huang 2006), which yields the steady-state solution of Equation 43. Using the steady-state solution for $\vec{P}^{(i)}$ in Equations 27–33 and Equations 4–7 yields the exact solution for correlation functions and variance. The resulting solutions precisely predict the simulation results for both the Moran model and the Schweinsberg model, as shown in Figure 2. Since their exact expression is unwieldy, we seek an approximation that yields a simpler form for further analysis in the following section.

E. Mean-Field Approximation and Solutions

We note that the one-site external transfers make the transitions between $\vec{P}^{(i)}$ a cyclic graph (Figure 6F, dotted arrows), and thus complicate the solutions of the linear equations (Equations 40–42). One possible approximation is thus to remove the transitions $\vec{P}^{(2)} \rightarrow \vec{P}^{(3)}$ and $\vec{P}^{(3)} \rightarrow \vec{P}^{(4)}$, and to account for them implicitly as mutations. As shown in Figure 7, an external transfer involves three sequences with four possible genealogical structures. After the external one-site transfers, the genealogical relationships of the two sequences in question will change (Figure 7), but the coalescent time will only change in two of the four genealogical structures, which happens with probability $w = 2\lambda_{3,2}/(\lambda_{3,3} + 3\lambda_{3,2})$. If the coalescent time does not change [as in cases (i) and (ii), Figure 7], exchangeability implies that there will be no change in the probability distribution for the pair of sequences in question. If the coalescent time changes [as in cases (iii) and (iv), Figure 7], the probability distribution changes, and our mean-field approximation is to assume that the two sites will then differ with probability d , i.e., the average pairwise distance in the population. Accordingly, the operator for these transfers, M_r , can be written as

$$M_r = \begin{pmatrix} -d & \frac{1-d}{2} & 0 \\ d & -\frac{1}{2} & 1-d \\ 0 & \frac{d}{2} & d-1 \end{pmatrix}. \quad (46)$$

Figure 7 Illustration of external one-site transfers and their impacts on the possible genealogical trees. In each tree, X and Y are the pair of sequences under consideration, and D is the donor sequence of the external transfer shown as an arrow from D to X. A red circle represents the MRCA of X and Y before the transfer, and a red star denotes the MRCA after the transfer.

We can combine these external one-site transfers and point mutations to form an effective mutation operator, which is $4\mu M_\mu + 4wr_1(1 - 2/N)M_r$ for $\vec{P}_2^{(2)}$ or $4\mu M_\mu + 2wr_1(1 - 3/N)M_r$ for $\vec{P}_2^{(3)}$. The explicit external one-site transfers appearing in Equation 43 are removed, and replaced by the appropriate mutation operator, yielding the simplified solutions given below.

Solution for the Moran model

We find the solutions of $\vec{P}_2^{(i)}$, which we will use later for the solutions of correlation functions, in the mean-field approximation for the Moran model as follows:

$$Q_2^{(2)} \equiv \frac{P_2^{(2)} - d^2}{d^2} = \frac{1 + r_2}{1 + r + r_1 + 2w(N - 2)r_1 + 2N\mu\tilde{a}}; \quad (47)$$

$$Q_2^{(3)} \equiv \frac{P_2^{(3)} - d^2}{d^2} = \frac{1 + r}{3(1 + r) + w(N - 3)r_1 + 2N\mu\tilde{a}} \cdot Q_2^{(2)}; \quad (48)$$

$$Q_2^{(4)} \equiv \frac{P_2^{(4)} - d^2}{d^2} = \frac{2(1 + r)}{3(1 + r) + N\mu\tilde{a}} \cdot Q_2^{(3)}. \quad (49)$$

Given that $r \ll 1$ and $N \gg 1$, we can simplify the solutions further:

$$Q_2^{(2)} \approx \frac{1}{1 + 2wNr_1 + 2N\mu\tilde{a}}; \quad (50)$$

$$Q_2^{(3)} \approx \frac{1}{3 + wNr_1 + 2N\mu\tilde{a}} \cdot Q_2^{(2)}; \quad (51)$$

$$Q_2^{(4)} \approx \frac{2}{3 + N\mu\tilde{a}} \cdot Q_2^{(3)}. \quad (52)$$

Using the mean-field solution above for $\vec{P}_2^{(i)}$ and Equation 34, we obtain the mean-field solution for c_M :

$$c_M(l) = \frac{2wd^2Nr_2(l)}{(1 + 2wNr_1(l) + 2N\mu\tilde{a})(1 + 2wNr + 2N\mu\tilde{a})}. \quad (53)$$

Similarly, using the mean-field solution for $\vec{P}_2^{(4)}$ and Equation 36, we find the solution for c_R :

$$c_R(l) = q \cdot c_M(l) \quad (54)$$

where $q = \frac{1}{3 + wNr_1(l) + 2N\mu\tilde{a}} \cdot \frac{2}{3 + N\mu\tilde{a}}$.

For c_S , using the solutions for $\vec{P}_2^{(2)}$ and $\vec{P}_2^{(4)}$ and Equation 35, we obtain

$$c_S = \frac{d^2(1-q)}{1 + 2wNr_1(l) + 2N\mu\tilde{a}}. \quad (55)$$

Lastly, we compute the variance

$$\sigma^2 \equiv c_S(L). \quad (56)$$

Solution for the Schweinsberg model

We find the solutions for $\vec{P}_2^{(i)}$ in the mean-field approximation for the Schweinsberg model as follows:

$$Q_2^{(2)} \approx \frac{1}{1 + 2wNr_1 + 2N\mu\tilde{a}}; \quad (57)$$

$$Q_2^{(3)} \approx \frac{1}{2} Q_2^{(2)}; \quad (58)$$

$$Q_2^{(4)} \approx \frac{2}{3} Q_2^{(3)}, \quad (59)$$

for $r \ll 1$.

Given the solutions above, we find

$$c_M(l) = \frac{2wd^2 Nr_2(l)}{\left(1 + 2wNr_1(l) + 2N\mu\tilde{a}\right)\left(1 + 2wNr + 2N\mu\tilde{a}\right)}, \quad (60)$$

$$c_R(l) = q \cdot c_M(l), \quad (61)$$

$$c_S(l) = \frac{d^2(1-q)}{1 + Nr_1(l) + 2N\mu\tilde{a}}, \quad (62)$$

where $q = 1/3$.

Mean-field approximation and coalescent theory

The mean-field results for $\vec{P}^{(i)}$ can equivalently be obtained using coalescent theory, which we illustrate here for the case of $\vec{P}^{(2)}$. Given two pairs of sites on a pair of sequences, a coalescent event could involve both of the two pairs with rate $\lambda_1 \equiv \lambda_{2,2} + 2r_2(l)/N$, where the last term is the rate of coalescence due to two-site transfers between the two sequences, yielding the state $\vec{P}^{(0)}$; or it could involve only one of the two pairs with rate $\lambda_2 \equiv 4r_1(l)/N$, which is the rate of coalescence due to internal one-site transfers between the two sequences, yielding the state $\vec{P}^{(1)}$. The coalescent time t thus follows an exponential distribution with rate $\lambda \equiv \lambda_1 + \lambda_2$. Given the coalescence time, one can compute $\vec{P}^{(2)}$ by propagating forward in time the process of mutations and external one-site transfers for a time t starting from the two ancestral states, $\vec{P}^{(0)}$ and $\vec{P}^{(1)}$. We note that external two-site transfers that occur during this time impact both sites, and are thus equivalent to exchanging one of the sequences for a different individual in the population. Since we study an exchangeable coalescent, these events do not change the distribution of coalescent times, and therefore have no impact on the calculation. We define a combined mutational operator that includes both mutations and external one-site transfers, $M \equiv 4\mu M_\mu + 4w(1 - 2/N)r_1(l)M_r$, and use it to propagate forward in time, while taking the expectation over the coalescent time:

$$\begin{aligned} \vec{P}^{(2)}(l) &= \int_0^\infty e^{-t\lambda} e^{tM} \left(\lambda_1 \vec{P}^{(0)} + \lambda_2 \vec{P}^{(1)} \right) dt \\ &= (\lambda - M)^{-1} \left(\lambda_1 \vec{P}^{(0)} + \lambda_2 \vec{P}^{(1)} \right) \end{aligned} \quad (63)$$

One can check that this is the same equation as the steady-state equation for $\vec{P}^{(2)}$ in the mean-field approximation (see Equation 40). Computing the matrix inverse and applying to $\vec{P}^{(0)} = (1, 0, 0)^T$ and $\vec{P}^{(1)} = ((1-d), d, 0)^T$ therefore yields the same expression for $P_2^{(2)}$ given in Equation 47.

F. Analytical Forms for Parameter Inference in Biased Samples

We consider a pair of individuals with coalescence time $t \ll N$. We assume that t is sufficiently short such that each pair of sites in the genome separated by a distance $l \ll L$ has been affected by, at most, one mutational or recombination event. For portions of the genome that were not affected by recombination, the average per site divergence of the pair of sequences is given by $\theta_s \equiv 2\mu t$. The probability that a single site is affected by recombination is $\rho_s \equiv 2\gamma \bar{f} t$. Since the donor is chosen randomly from the entire population of size N , we can neglect transfers between the given pair of individuals, which have probability $1/N$. After a transfer, the expected divergence in the recombined region is the bulk population diversity, d . Accounting for the rates and effects of mutation and recombination events, we calculate the expected diversity of the pair of sequences as

$$d_s = \rho_s d + (1 - \rho_s) \theta_s. \quad (64)$$

Recalling that $\rho = \gamma \bar{f} N$ and $\theta = \mu N$, where we consider N to represent the size of the bulk population with which sequences can recombine (e.g., that of an entire species), we can assume that $\rho \gg 1$, since typically we have $\gamma N \sim \mu N \sim 0.01 - 0.1$, while $\bar{f}$ is usually on the order of a kilobase or longer. These assumptions are later tested for self-consistency once fitting has been performed. Thus, we have $\rho_s d / \theta_s \approx \rho_s \theta / \theta_s = \rho \gg 1$, so that $(1 - \rho_s) \theta_s < \theta_s \ll \rho_s d$, which allows us to approximate d_s by considering only the contribution of recombination:

$$d_s \approx \rho_s d. \quad (65)$$

We now consider the calculation of $P_2^{(2)}(l)$, i.e., the probability that substitutions have occurred at both sites, where we distinguish the $P_2^{(2)}$ value of the population, which corresponds to the expectation for two sequences chosen at random from the entire population of size N , and that of the sample, which we denote as $P_{s,2}^{(2)}$ and is calculated for the pairs of individuals within the given sample. Since t is sufficiently short, the probability that two or more events occur at the two sites in question is negligible. Thus, the only event that can introduce substitutions at both sites over the timescale t is a two-site transfer from the bulk population into the sample, which occurs with probability $2tr_2(l)$. We obtain

$$P_{s,2}^{(2)}(l) \approx 2t r_2(l) P_2^{(2)}(l). \quad (66)$$

and, substituting the expression for $P_2^{(2)}$, yields

$$P_{s,2}^{(2)}(l) \approx 2r_2(l) t d^2 \left(\frac{1}{1 + 2wNr_1(l) + 2N\mu\bar{a}} + 1 \right). \quad (67)$$

We define

$$\tilde{P}_{s,2}^{(2)}(l) \equiv P_{s,2}^{(2)}(l) / d_s \quad (68)$$

and, given that $d_s \approx \rho_s d = 2\gamma \bar{f} t d$, we obtain

$$\tilde{P}_{s,2}^{(2)}(l) = \frac{dr_2(l)}{\gamma \bar{f}} \left(\frac{1}{1 + 2wNr_1(l) + 2N\mu\bar{a}} + 1 \right). \quad (69)$$

If the distance between sites l is smaller than typical transferred fragment sizes, we have $r_1(l) \approx \gamma l$ and $r_2(l) \approx \gamma(\bar{f} - l)$, using which, we find

$$\tilde{P}_{s,2}^{(2)}(l) = d \left(1 - \frac{l}{\bar{f}} \right) \left(\frac{1}{1 + 2wN\gamma l + 2N\mu\bar{a}} + 1 \right). \quad (70)$$

G. Application to Bacterial Sequence Analysis

We obtained a curated list of 185 *E. coli* sequence type 131 (ST131) isolates from two recent studies (Price *et al.* 2013; Petty *et al.* 2014; Ben Zakour *et al.* 2016), and a total list of 1216 *S. pneumoniae* isolates, which consisted of all isolates from the seven largest clusters in a longitudinal pneumococcal carriage study (Chewapreecha *et al.* 2014).

For each species, we applied a reference-based approach to generate whole-genome alignments from Illumina read data. We used *E. coli* strain EC958 (EMBL accession code HG941718), and *S. pneumoniae* strain ATCC700669 (EMBL accession code FM211187), as the reference genomes, and mapped read pairs against them using SMALT version 0.7.6 (<https://sourceforge.net/projects/smalt/>) with default settings. The resulting alignment was then processed with Samtools (Li *et al.* 2009) and FreeBayes (Garrison and Marth 2012) to generate a consensus sequence. To call bases at each position, we required at least two reads spanning in each direction (*i.e.*, for a total minimum read depth of four), and a $\geq 75\%$ consensus on the major allele. The base quality at the site had to be ≥ 50 , and the mapping quality had to be ≥ 30 , on the Phred scale. When a called base was an SNP, in addition, we required it to be supported by both FreeBayes and Samtools with quality scores ≥ 30 . All other sites that failed to pass these criteria, as well as insertions and deletions, were masked as gaps. Finally, for each clade in a species, the resulting consensus sequences were combined to generate a whole-genome alignment.

To avoid ambiguity in assigning distances along the genome due to the potential for genome rearrangement, the resulting whole-genome alignment was split into multiple gene alignments, and each gene alignment was further split into multiple alignment blocks with a fixed length of 300 bp. In each block, we removed sequences in which gaps constituted $>2\%$ of the total length. After filtering out such sequences, we additionally excluded alignment blocks consisting of less than five sequences. Using the filtered alignment blocks, we compared DNA sequences pairwise. For each pair of sequences k , we obtained a substitution profile S_i^k at the third base of each codon. To reduce the impact of selection, we masked positions containing nonsynonymous substitutions, and did not use them in the analysis, but preserved their genome coordinate so that physical distances remained unchanged.

Using the substitution profiles, we calculated the sample diversity, d_s , and the correlation functions $\tilde{P}_{s,2}^{(2)}(l)$, $c_M(l)$, $c_S(l)$, and $c_R(l)$. To infer population parameters, we fit $\tilde{P}_{s,2}^{(2)}(l)$ using the first 50 data points (codons) in the same manner as we tested using simulation results (Appendix F and Figure 4). To predict $\tilde{c}_M(l)$, $\tilde{c}_S(l)$, $\tilde{c}_R(l)$, we used the relations in Equations 34–36 with the fitted form of $\tilde{P}_{s,2}^{(2)}(l)$, and used $\tilde{P}_{s,2}^{(4)}(l) = q\tilde{P}_{s,2}^{(2)}(l)$, where q was determined by the coalescent structure of the sample (*e.g.*, $q = 1/3$ for BSC statistics). When fitting q (dashed blue and green lines in Figure 5), we performed linear regression using the relation $\tilde{c}_S(l) = (1 - q)\tilde{P}_{s,2}^{(2)}(l)$, with the fitted form $\tilde{P}_{s,2}^{(2)}(l)$ and the measured values of $\tilde{c}_S(l)$.

We note that, in principle, one could infer the full distribution of transferred fragment sizes by using the structure correlation function. One can invert Equation 12 to obtain $r_1(l)$ in terms of $c_S(l)$, and, by differentiating, obtain $p(f) \sim (\partial^2/\partial f^2)(1/c_S(f))$. In practice, however, accurately measuring the curvature of the correlation decay would require a very large amount of data, which we found to be well beyond the sizes of available datasets (data not shown). Nevertheless, as we discussed above, the mean fragment size can be efficiently inferred using currently available data.

H. Recombination Barriers and Population Structure

In the above analysis and simulations, the transfer process was unconstrained, such that each genome can recombine with any other sequence within the population with equal rate. In reality, however, barriers exist that prevent free genetic exchange among individuals within a bacterial population. For example, the mismatch-repair system inhibits interspecies recombination, and reduces recombination frequencies among distantly related individuals (Fraser *et al.* 2007). Moreover, the classical pathway for homologous recombination involves RecA-mediated homology recognition between donor and recipient sequences (Kuzminov 1999). In several bacterial species, laboratory studies have found a log-linear relationship between sequence divergence and the frequency of recombination (Fraser *et al.* 2007).

Recombination barriers in simulation

To assess the magnitude of recombination barriers on our estimation of transfer rates, we modified our simulations as follows. After randomly choosing a pair k of genomes for recombination, we calculate their pairwise distance, $d(k)$, and we allow recombination to occur with probability $\exp[-b d(k)]$, where b is the strength of the recombination barrier that reduces transfer efficiency. During the simulations, we record the total number of successful transfer events across the population. This process limits recombination among distantly related individuals, and thereby reduces the overall rate of successful transfers within the population (Figure 8).

We compared the overall rates of successful transfers, $\tilde{\gamma}$, measured directly from simulations with our estimations based on the mean-field approximation. Due to the recombination barrier, and its effect on the correlation functions, inference of transfer rates based on fitting of measurements from the bulk population will underestimate the rates of successful transfers by a factor of $1 + \theta b/3$, an increasing function of the product of population diversity and the recombination barrier (Figure 8;

Figure 8 Recombination barrier and the rate of successful transfers. The plot shows the population rate of successful transfers, $N\tilde{\gamma}$, as a function of b , which controls transfer efficiency such that a higher b corresponds to a larger recombination barrier. The theoretical prediction, $N\tilde{\gamma} = \phi / (1 + \theta b)$, is shown in solid lines. Simulation results are shown for measured rates (open circles), and inferred rates based on fitting mutational correlations (solid triangles). The solid circles are the inferred rates corrected by the scale factor, $1 + \theta b/3$. Inset shows the data collapsing onto the theoretical prediction when plotted as a function of θb . Simulations used parameters $N = 1000$, $L = 1000$, $f_0 = 50$, and $\gamma = 10^{-4}$, with various values of b and μ as indicated. Each data point corresponds to an average over 10,000 simulations, which were run over a total time much longer than the coalescent time. Fitting mutational correlations was carried out using the mean-field result for $c_M(l)$ in Equation 10.

and see derivation below). Thus, a significant deviation of our estimate requires both high population diversity and a large recombination barrier, and one can correct the estimates if the value of b is known for a given species. However, our inference procedure for fitting genomic samples (Appendix F) implicitly accounts for transfer efficiency in the way it models the sample diversity, d_s , which avoids having to explicitly measure and correct for recombination barriers in different datasets.

Rate of successful transfers in the population

Here, we calculate the average rate of successful transfers, $\tilde{\gamma}$, given the rate of attempted transfers, γ . For any sequence X , transfers occur from donor sequences, D , whose coalescence time with X is a random variable, t , which is exponentially distributed with rate $\lambda_{2,2}$. The average divergence between sequence X and D is then given by $2\mu t$, and the probability that transfer is successful is therefore $\exp(-2\mu t b)$. Integrating over t yields the mean rate of successful transfers,

$$\tilde{\gamma} = \gamma \int_0^\infty \lambda_{2,2} \exp(-\lambda_{2,2} t) \exp(-2\mu t b) dt = \gamma \lambda_{2,2} / (\lambda_{2,2} + 2\mu b) = \gamma / k_b, \quad (71)$$

where $k_b \equiv (\lambda_{2,2} + 2\mu b) / \lambda_{2,2} = 1 + \theta b$ measures the effective strength of the recombination barrier.

The impact of transfer efficiency on the mean-field approximation

As detailed in Appendix E, we obtained the mean-field approximation by studying the distribution of possible phylogenetical trees involved in an external one-site transfer (Figure 7). A recombination barrier changes this distribution—transfers are more likely to occur among closely related sequences rather than among distant ones. It thus overweights the first two types of trees shown in Figure 7, and underweights the last two trees, which changes the value of w , and leads to an underestimate of the rate of successful transfers.

To correct this bias, we study the effect of transfer efficiency on the last two trees in Figure 7. We consider a pair of sequences, X and Y , and an external one-site transfer from a donor sequence D to X . When we trace the three lineages backward in time, there exist two sequential mergers among them, and we let t_1 and t_2 denote the coalescent times of the first and second mergers, respectively. We compute the probability of observing tree (iii) with specified values of t_1 and t_2 . The first merger occurs at t_1 between branches D and Y , while all other possible mergers (D - X , X - Y , or D - X - Y) do not occur; the associated probability is thus $\lambda_{3,2} e^{-(3\lambda_{3,2} + \lambda_{3,3})t_1}$. After the first merger, there remains time $t_2 - t_1$ until the second merger, during which the sample contains two lineages; hence, the associated probability is $\lambda_{2,2} e^{-\lambda_{2,2}(t_2 - t_1)}$. Multiplying the two probabilities, and using the identity $\lambda_{2,2} = \lambda_{3,3} + \lambda_{3,2}$, the probability of observing the given tree is

$$p_{\text{tree}}(t_1, t_2) = \lambda_{2,2} \lambda_{3,2} e^{-2\lambda_{3,2} t_1 - \lambda_{2,2} t_2} \quad (72)$$

Since the two trees (iii) and (iv) have the same topology, their probability is the same. In both cases, the divergence between D and X , which is given by $2\mu t_2$, determines the transfer efficiency. We can therefore calculate the value of w , which is the probability of observing one-site transfers as trees (iii) and (iv), by integrating over the coalescence times:

$$w = 2 \int_0^\infty \int_0^{t_2} p_{\text{tree}}(t_1, t_2) e^{-2b\mu t_2} dt_1 dt_2 \quad (73)$$

$$= \frac{2\lambda_{2,2}\lambda_{3,2}}{(\lambda_{2,2} + 2b\mu)(\lambda_{2,2} + 2\lambda_{3,2} + 2b\mu)}. \quad (74)$$

With no recombination barrier (*i.e.*, $b = 0$), we obtain the original probability $w = w_0 \equiv 2\lambda_{3,2}/(\lambda_{2,2} + 2\lambda_{3,2})$. In the mean-field expressions for the correlation functions (Appendix E), the population recombination rate $N\gamma$ occurs only in the combination of parameters $wN\gamma$. Using the original mean-field solution to fit the population transfer rate (*i.e.*, using $w = w_0$) would thus yield a value $wN\gamma/w_0$, which can be written as $N\tilde{\gamma}wk_b/w_0$, indicating that we would underestimate the rate of successful transfers, $N\tilde{\gamma}$, by a factor $w_0/(wk_b)$. In the Moran model, by noting $\lambda_{2,2} = \lambda_{3,2} = 2/N$, we find the factor to be $1 + \theta b/3$, while in the Schweinsberg model $\lambda_{2,2} = 2\lambda_{3,2} = 2/N$, the factor is $1 + \theta b/2$.