# Recombination-driven genome evolution and stability of bacterial species

Purushottam D. Dixit
*Department of Systems Biology, Columbia University, New York, NY 10032*

Tin Yau Pang
*Institute for Bioinformatics, Heinrich-Heine-Universität Düsseldorf, 40221 Düsseldorf, Germany*

Sergei Maslov*
*Department of Bioengineering and Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana IL 61801, USA*

While bacteria divide clonally, occasional homologous recombination is known to be an important contributor to their evolution. However, the details of how the competition between clonal inheritance and recombination shapes genome diversification, population structure, and species stability remains poorly understood. Using a computational model, we propose two evolutionary regimes and identify two composite parameters that dictate the fate of bacterial species. In the *divergent* regime, characterized by either a low recombination frequency or strict barriers to recombination, cohesion due to recombination is not sufficient to overcome the mutational drift. As a consequence, the divergence between any pair of genomes in the population steadily increases in the course of evolution. The species as a whole lacks coherence at the population level with sub-populations continuously formed and dissolved. In contrast, in the *metastable* regime, characterized by a high recombination frequency combined with low barriers to recombination, genomes continuously recombine with the rest of the population. The population remains genetically cohesive and stable over time. We demonstrate that the transition between these two regimes can be affected by relatively small changes in evolutionary parameters. Using the data from Multi Locus Sequence Typing (MLST) analysis we classify a number of well-studied bacterial species to be either the divergent or the metastable type. Mechanisms that allow bacterial species to transition from one regime to another are discussed. Generalizations of the framework to understand adaptive populations, horizontal gene transfer of non-homologous regions, and spatial correlations in diversity along the chromosome are also discussed.

**Introduction:** Bacterial genomes are extremely variable, comprising both a consensus 'core' genome which is present in the majority of strains in a population, and an 'auxiliary' genome, comprising genes that are shared by some but not all strains (1–7).

Multiple factors shape the diversification of the core bacterial genome. Bacteria divide clonally thereby inheriting the entirety of their mother's genome. The balance between this vertical inheritance and random fixation of single nucleotide polymorphisms (SNPs), generated at a rate $\mu$ per base pair per generation, limits the typical population diversity to $\theta = 2\mu N_e$ where $N_e$ is the *effective* population size of the species (8). During the last two decades, genetic exchange between closely related organisms through homologous recombination, attempted at a rate $\rho$ per base pair per generation, has also been recognized as a significant factor in evolution (5, 6, 9–14). Notably recombination between genetically distant bacteria is suppressed, the probability $p_{\text{success}} \sim e^{-\delta/\delta_{\text{TE}}}$ of successful recombination of foreign DNA into a recipient genome decays exponentially with $\delta$, the *local* divergence between the donor DNA and the recipient (12, 15–17). The effective barrier $\delta_{\text{TE}}$ to successful recombination, referred here as the *transfer efficiency*, is shaped at least in part by the biophysical mechanisms of homologous recombination (15, 16).

While clonal inheritance with mutations imposes a clonal structure on the population, recombination acts as an homogenizing force, keeping populations homogeneous and potentially destroying the genetic signatures of clonality (6, 17, 18). There are two principal components to the interplay between mutations and recombinations. First is the competition between the diversity within the population $\theta$ and the maximal diversity within one subpopulation $\delta_{\text{TE}}$ uniformly capable of successful recombination. If $\delta_{\text{TE}} < \theta$, one expects spontaneous fragmentation of the entire population into several transient sexually isolated sub-populations that rarely exchange genetic material between each other. In contrast, if $\delta_{\text{TE}} > \theta$, unhindered exchange of genetic fragments may result in a single cohesive population. Second is the competition between the recombination transfer rate $\rho$ and the mutation rate $\mu$. The typical time between consecutive recombination events in any local genomic region is $1/2\rho \times l_{\text{tr}}$ where $l_{\text{tr}}$ is the typical length of transferred regions. In the same time, the total divergence accumulated in this

* Email: maslov@illinois.edu

region is $\delta_{\rm mut} \sim 2\mu/2\rho \times l_{\rm tr}$). If $\delta_{\rm mut} \gg \delta_{\rm TE}$, the pair of genomes may become sexually isolated from each other between successive recombination events. In contrast, if $\delta_{\rm mut} < \delta_{\rm TE}$, frequent recombination events may disallow sexual isolation resulting in a homogeneous population.

What qualitative dynamical regimes in bacterial evolution emerge from the competition and balance between these two factors and which evolutionary parameters dictate the evolutionary fate of bacterial genome diversification remains poorly understood. Importantly, even the question of whether bacteria can retain their clonal inheritance in the presence of recombination and whether signatures of clonal structure and recombination can be inferred from population genetic data is still heavily debated (17, 19–21).

Nonetheless, some aspects of this interplay have been explored before. In their pioneering study Vetsigian and Goldenfeld (22) investigated the effects of non-recombining segments (for example, a large scale inversion or insertion) on recombination events in their chromosomal neighborhoods' vicinity and how it may result in divergence spreading along the chromosome. Falush et al. (23) suggested that a low *transfer efficiency* $\delta_{\rm TE}$ leads to sexual isolation in *Salmonella enterica*. Fraser et al. (18), working with a $\theta = 0.4\%$ (lower than those observed in typical bacterial species) and a *transfer efficiency* $\delta_{\rm TE} \approx 2.4\%$ concluded that realistic values of sexual isolation in bacterial species is insufficient to cause speciation with realistic recombination frequencies. Doroghazi and Buckley (24), working with a fixed *transfer efficiency* but a very small population size (limit of $\theta \to 0$), studied how the competition between mutations and recombination affects the cohesion of two isolated subpopulations.

In this work, using a computational model and mathematical calculations, we show that the two composite parameters identified above, $\theta/\delta_{\rm TE}$ and $\delta_{\rm mut}/\delta_{\rm TE}$, determine qualitative evolutionary dynamics of a given bacterial species. Furthermore, we identify two principal regimes of this dynamics. In the *divergent regime*, characterized by a high $\delta_{\rm mut}/\delta_{\rm TE}$, local genomic regions acquire multiple mutations between successive recombination events and rapidly isolate themselves from the rest of the population. The population remains mostly clonal where transient sexually isolated sub-populations are continuously formed and dissolved. In contrast, in the *metastable regime*, characterized by a low $\delta_{\rm mut}/\delta_{\rm TE}$ and a low $\theta/\delta_{\rm TE}$), local genomic regions recombine repeatedly before ultimately escaping the pull of recombination (hence the name "metastable"). At the population level, in this regime all genomes can exchange genes with each other resulting in a sexually cohesive and temporally stable population. Notably, our analysis suggests that only a small change in evolutionary parameters can have a substantial effect on evolutionary fate of bacterial genomes and populations.

We also show how to classify bacterial species using the conventional measure of the relative strength of recombination over mutations, $r/m$ (defined as the ratio of the number of single nucleotide polymorphisms (SNPs) brought by recombinations and those brought by point mutations in a pair of closely related strains), and our second composite parameter $\theta/\delta_{\rm TE}$. Based on our analysis of the existing MLST data, we find that different bacterial species bacteria belong to either divergent and metastable regimes. We discuss possible molecular mechanisms and evolutionary forces that decide the role of recombination in a species's evolutionary fate. We also discuss possible extensions of our analysis to include adaptive evolution and genome modifications such as insertions, deletions, and inversions.

**The computational model**: We consider a population of $N_e$ strains. The population evolves with non-overlapping generations and in each generation the strains choose their parents randomly (8). The genome of each strain has $G$ indivisible and non-overlapping transferable units. For simplicity, in what follows we refer to these units as *genes* but note that while protein coding genes in a typical bacteria are $\sim$ 1000 base pair long, we use $l_{\rm tr} = 5000$ base pairs mimicking genetic transfers longer than individual protein coding genes (6, 10). These genes acquire point mutations at a rate $\mu$ per base pair per generation and recombinations are attempted on a recipient genome from a randomly selected donor in the population at a rate $\rho$ per base pair per generation. The mutations and recombination events are assumed to have no fitness effects. Finally, the probability of a successful integration of a donor gene decays exponentially, $p_{\rm success} \sim e^{-\delta/\delta_{\rm TE}}$, with the *local* divergence $\delta$ between the donor and the recipient.

In order to avoid simulating extremely large bacterial population sizes, we focus on the evolution of divergence between two randomly chosen genomes labeled $X$ and $Y$ in a co-evolving population. $X$ and $Y$ start diverging from each other as identical twins at time $t = 0$ (when their mother divides). We denote by $\delta_i(t)$, the divergence in the $i^{\rm th}$ gene between $X$ and $Y$ at time $t$ and by $\Delta(t) = 1/G \sum_i \delta_i(t)$ the average genome-wide divergence. Based on population genetic and biophysical considerations, we derive the probability $E(\delta_a|\delta_b) = 2\mu M(\delta_a|\delta_b) + 2\rho l_{\rm tr} R(\delta_a|\delta_b)$ ($a$ for after and $b$ for before) that the divergence in any gene changes from $\delta_b$ to $\delta_a$ in one generation (see supplementary materials for details) (6). Briefly, there are two components to the probability, $M$ and $R$. Point mutations, represented by $M(\delta_a|\delta_b)$, occur at a rate $2 \times \mu$ per base pair per generation and increase the divergence in a gene by $1/l_{\rm tr}$. Unlike point mutations, after a recombination event (represented by $R(\delta_a|\delta_b)$), the divergence can change suddenly, taking values either larger or smaller than the current divergence (6). Note that recombinations from highly diverged members in the population are suppressed exponentially and consequently not all recombination attempts are successful. Intuitively, the time evolution of $p(\delta|t)$ of the probability of observing a divergence $\delta$ in a

gene at time $t$ can be written as

$$\frac{\partial p(\delta|t)}{\partial t} = -2\mu\frac{\partial p(\delta|t)}{\partial \delta} + 2\rho l_{tr}\underbrace{\int R(\delta|\delta_b)\times p(\delta_b|t)d\delta_b}_{\text{entry into }\delta}$$
$$- 2\rho l_{tr}\underbrace{\int R(\delta_a|\delta)\times p(\delta|t)d\delta_a}_{\text{exit from }\delta}. \quad (1)$$

**Evolution of *local* divergence has large fluctuations:** In Fig. 1 we show a color-coded typical trajectory of evolution of the *local* divergence $\delta(t)$ of a single gene in a pair of genomes. We have used $\theta = 1.5\%$ and $\delta_{\rm TE} = 1\%$, values typically observed in bacterial species (6, 18). To keep the simulation times manageable, the mutation and the recombination rates used here are 4-5 orders of magnitude higher compared to those observed in real bacteria ($\mu = 10^{-5}$ per base pair per generation and $\rho = 5\times 10^{-6}$ per base pair per generation, $\delta_{\rm mut}/\delta_{\rm TE} = 0.04$) (25, 26) while keeping the ratio of the rates realistic (5, 6, 13, 27). Alternatively, the unit of time in these simulations is considerably longer than one single generation.

The time evolution of $\delta(t)$ is noisy; mutational drift events that gradually increase the divergence linearly with time (red) are frequently interspersed with homologous recombination events (green if they increase $\delta(t)$ and blue if they decrease it) that suddenly change the divergence to typical values seen in the population (see Eq. A1 in the appendix). Eventually, either through the gradual mutational drift or a sudden recombination event, $\delta(t)$ increases beyond the integration barrier set by the transfer efficiency, $\delta(t) \gg \delta_{\rm TE}$. Beyond this point, the two strains belong to two different sexually isolated sub-clades. Any further recombination events on any one of the strains are limited to their own sub-clades and do not change the divergence between the two strains. Consequently, the mutational drift keeps on driving them further apart indefinitely.

**Genome-wide divergence:** Because genomic regions in our model evolve independently of each other, the genome-wide average divergence $\Delta(t)$ can be calculated as the mean of $G$ independent realizations *local* divergences $\delta(t)$. Consequently, since the number $G$ of genes in the genome is large, the law of large numbers implies that the fluctuations in the dynamics of $\Delta(t)$ are substantially suppressed compared to more noisy $\delta(t)$ seen in Fig. 1.

In Fig. 2, we plot the time evolution of $\Delta(t)$ between a pair of strains (as % difference). We have used $\theta = 0.25\%$, $\delta_{\rm TE} = 1\%$, and $\delta_{\rm mut}/\delta_{\rm TE} = 2, 0.5, 0.04$, and $2\times 10^{-3}$ respectively. When $\delta_{\rm mut}/\delta_{\rm TE}$ is large (either a due to low $\rho$ or a low $\delta_{\rm TE}$), in any local genomic region, multiple mutations are acquired between two successive recombination events. Consequently, individual genes escape the pull of recombination rapidly and $\Delta(t)$ increases roughly linearly with time at a rate $2\mu$. For smaller values of $\delta_{\rm mut}/\delta_{\rm TE}$, the rate of change of $\Delta(t)$ in the long term decreases as many of the individual genes repeatedly recombine with the population. However, even then the fraction of genes that have escaped the integration barrier slowly increases over time, eventually leading to a linear increase in $\Delta(t)$ with time albeit with a slope different than $2\mu$. The repeated resetting of $\delta(t)$ after homologous recombination (see Fig. 1) generally results in a $\Delta(t)$ that increases extremely slowly with time.

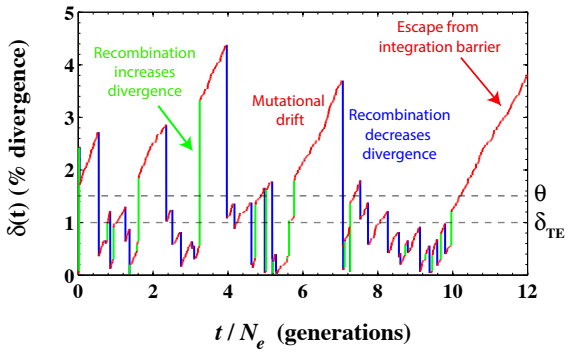


FIG. 1. A typical evolutionary trajectory of the *local* divergence $\delta(t)$ of a single gene between a pair of strains. We have used $\mu = 10^{-5}$, $\rho = 5\times 10^{-6}$ per base pair per generation, $\theta = 1.5\%$ and $\delta_{\rm TE} = 1\%$. Red tracks indicate divergence increasing linearly, at a rate $2\mu$ per base pair per generation, with time due to mutational drift. Green tracks indicate recombination events that suddenly increase the divergence and blue tracks indicate recombination events that suddenly decrease the divergence. Eventually, the divergence increases sufficiently and the local genomic region escapes the pull of recombination.

At the shorter time scale, the trends in genome divergence are opposite of those at the longer time scale. At a fixed $\theta$, a low value of $\delta_{\rm mut}/\delta_{\rm TE}$ implies faster divergence and vice versa. When recombination rate is high, genomes of strains quickly 'equilibrate' with the population and the genome-wide average divergence between a pair of strains reaches the population average diversity $\sim \theta$ (see the red trajectory in Fig. 2). From here, any new mutations that increase the divergence are constantly wiped out through repeated recombination events with the population. Over time, a rare event, wherein mutations accumulate and no recombinations take place for a sufficiently long time, allows individual 'genes' to escape the pull of recombination. Subsequently, genes on any pair of genomes escape the pull of recombination one after the other and the genomes can diverge indefinitely.

Computational algorithms that build phylogenetic trees from multiple sequence alignments often rely on the assumption that the sequence divergence, for example between a pair of strains (at the level of individual genes or at the level of genomes), faithfully represents the time that has elapsed since their most recent common ancestor (MRCA). However, Fig. 1 and Fig. 2 serve as a cautionary tale. Notably, after just a single recombination event the *local* divergence at the level of individual genes does not at all reflect time elapsed since divergence but rather depends on statistics of divergence within a recombin-

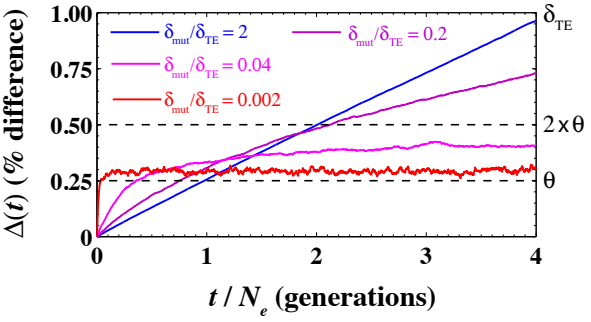


FIG. 2. Genome-wide divergence $\Delta(t)$ as a function of time at $\theta/\delta_{\rm TE} = 0.25$. We have used $\delta_{\rm TE} = 1\%$, $\mu = 5\times10^{-6}$ per base pair per generation and $\rho = 2.5\times10^{-8}, 2.5\times10^{-7}, 1.25\times10^{-6}$, and $2.5\times10^{-5}$ per base pair per generation corresponding to $\delta_{\rm mut}/\delta_{\rm TE} = 2, 0.2, 0.04$, and $2\times10^{-3}$ respectively. The dashed black lines at $\Delta = 0.25\%$ and $\Delta = 0.5\%$ show $\theta$ and $2\times\theta$.

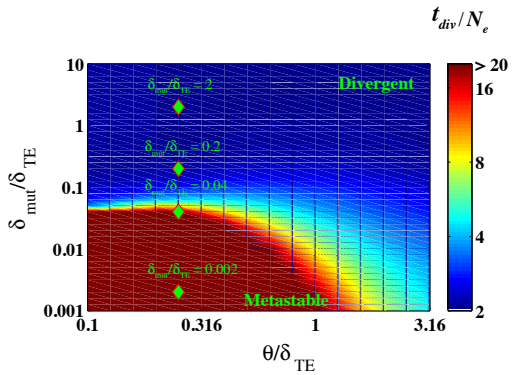


FIG. 3. The time $t_{\rm div}$ required for a pair of genomes to diverge beyond the typical population diversity $\sim\theta$ (see main text). We calculate the time taken for the genome-wide average divergence to reach $2\times\theta$ as a function of $\theta/\delta_{\rm TE}$ and $\delta_{\rm mut}/\delta_{\rm TE}$. We used $\delta_{\rm TE} = 1\%$, $\mu = 10^{-5}$ per base pair per generation. We changed $\rho$ and $\theta$ to scan the $(\theta/\delta_{\rm TE}, \delta_{\rm mut}/\delta_{\rm TE})$ space. The green diamonds represent four populations shown in Fig. 2 and Fig. 4 (see below).

ing population (see (6) for more details). At the level of genomes, when $\delta_{\rm mut}/\delta_{\rm TE}$ is large (e.g. the blue trajectory in Fig. 2), the time since MRCA of two strains is directly correlated with the number of mutations that separate their genomes. In contrast, when $\delta_{\rm mut}/\delta_{\rm TE}$ is small (see pink and red trajectories in Fig. 2), frequent recombination events repeatedly erase the memory of the clonal ancestry. Nonetheless, individual genomic regions slowly escape the pull of recombination at a fixed rate. Thus, the time since MRCA is reflected not in the total divergence between the two genomes but in the fraction of the length of the total genomes that has escaped the pull of recombination. As described above one uses a very different rate of accumulation of divergence to estimate evolutionary time from genome-wide average divergence.

**Quantifying metastability:** How does one quantify the *metastable* behavior described above? At the level of individual genes it is manifested through constant resetting of $\delta(t)$ to typical population values and at the level of entire genomes through a very slow increase in $\Delta(t)$ when $\delta_{\rm mut}/\delta_{\rm TE}$ is small. Fig. 2 suggests that high rates of recombination prevent pairwise divergence from increasing beyond the typical population divergence $\sim\theta$ at the whole-genome level. Thus, for any set of evolutionary parameters, $\mu$, $\rho$, $\theta$, and $\delta_{\rm TE}$, the time it takes for a pair of genomes to diverge far beyond the typical population diversity $\theta$ can serve as a quantifier for metastability.

In Fig. 3, we plot the time $t_{\rm div}$ required for the genome-wide average divergence $\Delta(t)$ between a pair of genomes to exceed twice the typical population diversity $2\times\theta$ as a function of $\theta/\delta_{\rm TE}$ and $\delta_{\rm mut}/\delta_{\rm TE}$. Note that in the absence of recombination, it takes $t_{\rm div} = 2N_e$ generations before $\Delta(t)$ exceeds $2\theta$. While we work with an arbitrary threshold of twice the average diversity, we do not expect that our results will change for any other non-typical divergences $> \theta$.

We observe two distinct regimes in the behavior of $t_{\rm div}$ as a function of $\theta/\delta_{\rm TE}$ and $\delta_{\rm mut}/\delta_{\rm TE}$. In the *divergent* regime, after a few recombination events, the divergence $\delta(t)$ at the level of individual genes quickly escapes the integration barrier and increases indefinitely. Consequently, $\Delta(t)$ increases linearly with time (see $\delta_{\rm mut}/\delta_{\rm TE} = 2$ in Fig. 2) and reaches $\Delta(t) = 2\theta$ within $\sim 2N_e$ generations. In contrast, it takes extremely long for $\Delta(t)$ to reach $2\theta$ as $\delta_{\rm mut}/\delta_{\rm TE}$ decreases in the *metastable* regime. Here, genes get repeatedly exchanged with the population and $\Delta(t)$ between a pair of strains appears to remain constant over large periods of time (see $\delta_{\rm mut}/\delta_{\rm TE} = 2\times10^{-3}$ in Fig. 2). Notably, a small perturbation in evolutionary parameters can change evolutionary dynamics from divergent to metastable and vice versa near the boundary region between the two regimes.

**Population structure:** Can we understand the phylogenetic structure of the population by studying the dynamics of evolution of divergence between a pair of strains in the population? Every pair of genomes diverges indefinitely when sufficient amount of time has elapsed since their MRCA. But, in a finite population of size $N_e$, the *average* probability of observing a pair of strains whose MRCA existed $t$ generations ago is exponentially distributed, $\overline{p_c(t)} \sim e^{-t/N_e}$ (the line indicates averaging over multiple realizations of the coalescent process) (28–30). Thus, while it may be possible a pair of genomes considered above to diverge indefinitely from each other (see Fig. 2), it becomes more and more unlikely to find such a a pair in a finite sized population.

Let us define $\pi(\Delta)$ as the probability that the genomes of two randomly picked strains in a population have diverged from each other by $\Delta$. Let $\overline{\pi(\Delta)}$ be the time

average of $\pi(\Delta)$. We have

$$\pi(\Delta) = \int_0^\infty p_c(t) \times p(\Delta|t)dt \text{ and}$$
$$\overline{\pi(\Delta)} = \int_0^\infty \overline{p_c(t)} \times p(\Delta|t)dt$$
$$= \frac{1}{N_e}\int_0^\infty e^{-t/N_e} \times p(\Delta|t)dt \qquad (2)$$

In Eq. 2, $p_c(t)$ is the probability of that a pair of strains share their MRCA $t$ generations ago in any time snapshot of the population and $p(\Delta|t)$ is the probability that a pair of strains have diverged by $\Delta$ at time $t$. Given that $\Delta(t)$ is the average of $G \gg 1$ independent realizations of $\delta(t)$, we can approximate it as a Gaussian distribution with average $\langle\delta(t)\rangle_G = \int \delta \times p(\delta|t)d\delta$ and variance $\sigma^2 = \frac{1}{G}\left(\langle\delta(t)^2\rangle_G - \langle\delta(t)\rangle_G^2\right)$. The angular brackets and the subscript $G$ indicate an average over the entire genome.

Unlike the time averaged distribution $\overline{\pi(\Delta)}$, the instantaneous distribution $\pi(\Delta)$ is accessible from genome sequences. However, we expect substantial differences between the two distributions even in the large population limit given that $p_c(t)$ is extremely noisy and does not resemble its long-time average $\overline{p_c(t)} \sim e^{-t/N_e}$ (29, 30). In panels a) to d) of Fig. 4, we show $\pi(\Delta)$ for the four cases shown in Fig. 2. We fixed the population size to $N_e = 500$. We changed $\delta_{\text{mut}}/\delta_{\text{TE}}$ by changing the recombination rate $\rho$. The solid lines represent a time snapshot obtained by numerically sampling $p_c(t)$ in a Fisher-Wright population of size $N_e = 500$. The dashed black line represents the time average $\overline{\pi(\Delta)}$.

In the divergent regime of Fig. 3, at high $\delta_{\text{mut}}/\delta_{\text{TE}}$ values, the *instantaneous* snapshot distribution $\pi(\Delta)$ has multiple peaks indicating the spontaneous formation of clonal sub-populations *even in a homogeneous population* that exchange genetic material within the clade but not outside of it, either because of a low recombination frequency or because of a low transfer efficiency. In this case, the time averaged distribution $\overline{\pi(\Delta)}$ has a long exponential tail and, as expected, does not agree with the instantaneous distribution $\pi(\Delta)$.

Notably, in the metastable regime, at lower values of $\delta_{\text{mut}}/\delta_{\text{TE}}$, the exponential tail shrinks into a Gaussian-like peak. The width of this peak relates to fluctuations in $\Delta(t)$ around its mean value which in turn are related to the total number of genes $G$. Moreover, the difference between the instantaneous and the time averaged distributions decreases as well. In this limit, all strains in the population exchange genetic material with each other. Thus, in the metastable regime, frequent recombination events can successfully eiliminate the clonal structure of the population leading to a sexually cohesive and temporally stable population.

**Analysis of bacterial species:** How are bacterial species placed on the divergent-metastable diagram? Instead of $\delta_{\text{mut}}/\delta_{\text{TE}}$ as defined here, population genetic studies of bacteria quantify the relative 'strength' of recombination over mutations as $r/m$. In our framework, $r/m$ can be estimated as $r/m = \rho_{\text{succ}}/\mu \times l_{\text{tr}} \times \delta_{\text{tr}}$ where $\rho_{\text{succ}} < \rho$ is the rate of successful recombination events and $\delta_{\text{tr}}$ is the average divergence in transferred regions. Both $\rho_{\text{succ}}$ and $\delta_{\text{tr}}$ depend on the evolutionary parameters (see appendix for a detailed description of the calculations).

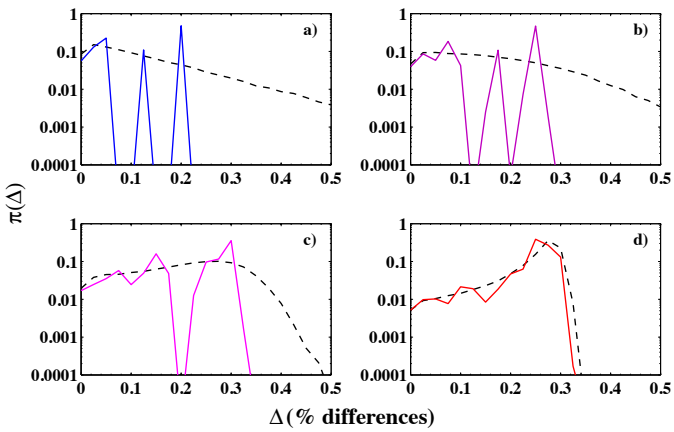


FIG. 4. Distribution of genome-wide divergence in a co-evolving population for increasing values of $\delta_{\text{mut}}/\delta_{\text{TE}}$. In all 4 panels, dashed black lines represent $\overline{\pi(\Delta)}$, the time averaged distribution of genome-wide diversity in the population. The solid lines represent an instance of the distribution $\pi(\Delta)$ at one snap shot. The time averaged and the snapshot distributions were estimated by sampling $5 \times 10^5$ pairwise coalescent times from the time averaged coalescent distribution $p \sim e^{-t/N_e}$ and the instantaneous coalescent distribution $p_c(t)$.

In Fig. 5, we re-plot the 'phase diagram' in Fig. 3 in terms of $\theta/\delta_{\text{TE}}$ and $r/m$ and place multiple bacterial species on it. We estimated $\theta$ from MLST data (31) and used $r/m$ values that were determined previously by Vos and Didelot (13). We assumed that the transfer efficiency $\delta_{\text{TE}}$ was approximately equal to $\delta_{\text{TE}} \sim 2.26\%$ (18). There are three striking features. First, our analysis identifies both $r/m$ and $\theta/\delta_{\text{TE}}$ as important evolutionary parameters and suggests that individually $r/m$ or $\theta/\delta_{\text{TE}}$ alone cannot determine population structure. Second, the sharp transition between the divergent and the metastable phase is also observed in the modified phase diagram as well, implying that a small change in either $r/m$ or $\theta/\delta_{\text{TE}}$ can change the evolutionary fate of a bacterial species. And finally, we observe that real bacterial species can indeed be both divergent as well as metastable.

Can bacteria change their evolutionary fates? There are multiple biophysical and ecological processes by which bacterial species may move from the metastable to the divergent regime and vice versa in Fig. 3. For example, if population size remains constant, a change in mutation rate changes $\delta_{\text{mut}}/\delta_{\text{TE}}$ as well as $\theta$. A change in the expression of the mismatch repair (MMR) system or

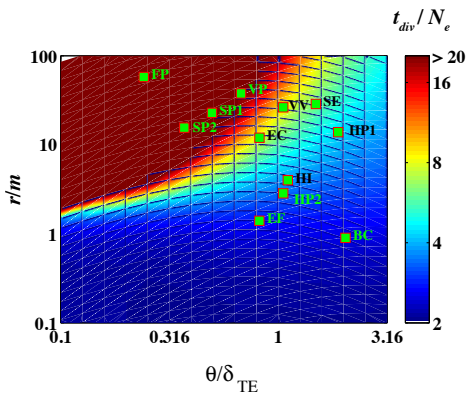


FIG. 5. Placement of real bacteria on the metastable-divergent phase diagram. Abbreviation of species as follows: FP: *Flavobacterium psychrophilum*, VP: *Vibrio parahaemolyticus*, SE: *Salmonella enterica*, VV: *Vibrio vulnificus*, SP1: *Streptococcus pneumoniae*, SP2: *Streptococcus pyogenes*, HP1: *Helicobacter pylori*, HP2: *Haemophilus parasuis*, HI: *Haemophilus influenzae*, BC: *Bacillus cereus*, EF: *Enterococcus faecium*, and EC: *Escherichia coli*.

the type of the MMR system can change $\delta_{\rm TE}$ (15). Loss of co-infecting phages or the SOS system and the number of restriction-modification (RM) systems (32) can change the rate of recombination.

Adaptive and ecological events are regularly inferred from population genomics data only after rejecting models of *neutral* evolution. However, the range of qualitative behaviors explained by the neutral models of recombination-driven evolution of bacterial species was not entirely quantified leading to potentially unwarranted conclusions, as illustrated in (33).

Consider *E. coli* as an example. Known strains of *E. coli* are usually grouped into 5-6 different sub-clades. It is thought that inter-clade sexual exchange is lower compared to intra-clade exchange (6, 27). Ecological niche separation and/or selective advantage are usually implicated as initiators of such putative speciation events (17). In our previous analysis of 32 fully sequenced *E. coli* strains, we estimated $\theta/\delta_{\rm TE} > 3$ and $r/m \sim 8-10$ (6) implying that *E. coli* resides in the divergent regime in Fig. 5. Thus, the analysis presented here indicates that *E. coli* strains should spontaneously form sexually isolated sub-populations even in the absence of selective pressures or ecological niche separation. Consequently, careful analysis is needed to reject neutral models of evolution in the study of population genetics of bacteria.

**Conclusions:** While recombination is now recognized as the chief contributor to the observed genome diversity in many bacterial species(5, 6, 9–13), its effect on population structure and species stability is still heavily debated (17–21).

In this work, we have shown that recombination-driven bacterial genome evolution can be understood as a balance between two important competitions. We identified the two dimensionless parameters $\theta/\delta_{\rm TE}$ and $\delta_{\rm mut}/\delta_{\rm TE}$ that dictate this balance and result in two qualitatively different regimes in bacterial evolution, separated by a sharp transition.

As seen in Fig. 3 and Fig. 5, in the divergent regime, the pull of recombination is insufficient to homogenize individual genes and entire genomes leading to a temporally unstable and sexually fragmented population. Notably, understanding divergence between a pair of genomes allows us to study the structure of the population as well. As shown in Fig. 4, genomes of members of divergent population form a clonal population structure. On the other hand, in the metastable regime, individual genomes repeatedly recombine genetic fragments with each other leading to a sexually cohesive and temporally stable population. As seen in Fig. 5, real bacterial species appear to belong to both of these regimes as well as in the cross-over region between the regimes.

**Extending the framework:** Recombination in bacterial species is thought to be essential in order to minimize the fitness loss due to Muller's ratchet (34) and to minimize clonal interference (35). Thus, it is likely that both recombination frequency and transfer efficiency are under selection (34). Could one include fitness effects in our theoretical framework? Here, we considered the dynamics of *neutrally* evolving bacterial population of effective population size $N_e$. The effective population size is incorporated in our framework via the coalescent time distribution of a neutral population, $\exp(-T/N_e)$ (see supplementary materials). Neher and Hallatschek (36) recently showed that while pairwise coalescent times in adaptive populations are *not* exponentially distributed, this distribution has a pronounced exponential tail with an *effective* population size $N_e$ weakly related to the actual population size and largely determined by the variance of mutational fitness effects (36). Our results could be generalized for adaptive populations by incorporating the effects of such non-exponential coalescent time distribution.

In this work, we assumed that in a recipient genome, recombinations bring *non-overlapping* segments of length $l_{\rm tr}$ that recombine in their entirety. In real bacteria, different recombined segments have variable lengths and partially overlap with each other thereby creating a *mosaic* of clonal and transferred regions along a chromosome (6, 10, 11). Overlapping transfers can affect evolutionary dynamics. In particular, when a local region within a genome has diverged above the threshold imposed by transfer efficiency $\delta_{TE}$ it reduces the likelihood of successful homologous recombination near both of its boundaries leading to a gradual expansion of the highly diverged region along the chromosome (22). Vetsigian and Goldenfeld proposed (22) that non-core genome segments e.g. horizontally acquired pathogenicity genomic islands could nucleate such propagating fronts of diversity and ultimately give rise to a new species. Genome rearrangements such as large-scale inversions are also expected to reduce the local rate of recombination in their vicinity. One expects that a long stretch of a genome

that diverged above transfer efficiency threshold can also result in two propagating wave fronts of divergence along the chromosome. In our future studies we plan to explore these and other extensions on top of the basic mathematically tractable model described here.


[1] D. Medini, C. Donati, H. Tettelin, V. Masignani, and R. Rappuoli, Current opinion in genetics & development **15**, 589 (2005).
[2] H. Tettelin et al., Proceedings of the National Academy of Sciences of the United States of America **102**, 13950 (2005).
[3] J. S. Hogg et al., Genome Biol **8**, R103 (2007).
[4] P. Lapierre and J. P. Gogarten, Trends in genetics **25**, 107 (2009).
[5] M. Touchon et al., PLoS genet **5**, e1000344 (2009).
[6] P. D. Dixit, T. Y. Pang, F. W. Studier, and S. Maslov, Proceedings of the National Academy of Sciences **112**, 9070 (2015).
[7] P. Marttinen, N. J. Croucher, M. U. Gutmann, J. Corander, and W. P. Hanage, Microbial Genomics **1** (2015).
[8] D. L. Hartl, A. G. Clark, and A. G. ClarkPrinciples of population genetics Vol. 116 (Sinauer associates Sunderland, 1997).
[9] D. S. Guttman and D. E. Dykhuizen, Science **266**, 1380 (1994).
[10] R. Milkman, Genetics **146**, 745 (1997).
[11] D. Falush et al., Proceedings of the National Academy of Sciences **98**, 15056 (2001).
[12] C. M. Thomas and K. M. Nielsen, Nature reviews microbiology **3**, 711 (2005).
[13] M. Vos and X. Didelot, The ISME journal **3**, 199 (2009).
[14] F. W. Studier, P. Daegelen, R. E. Lenski, S. Maslov, and J. F. Kim, Journal of molecular biology **394**, 653 (2009).
[15] M. Vulić, F. Dionisio, F. Taddei, and M. Radman, Proceedings of the National Academy of Sciences **94**, 9763 (1997).
[16] J. Majewski, FEMS microbiology letters **199**, 161 (2001).
[17] M. F. Polz, E. J. Alm, and W. P. Hanage, Trends in Genetics **29**, 170 (2013).
[18] C. Fraser, W. P. Hanage, and B. G. Spratt, Science **315**, 476 (2007).
[19] J. Wiedenbeck and F. M. Cohan, FEMS microbiology reviews **35**, 957 (2011).
[20] W. F. Doolittle, Current Biology **22**, R451 (2012).
[21] B. J. Shapiro, J.-B. Leducq, and J. Mallet, PLoS Genet **12**, e1005860 (2016).
[22] K. Vetsigian and N. Goldenfeld, Proceedings of the National Academy of Sciences of the United States of America **102**, 7332 (2005).
[23] D. Falush et al., Philosophical Transactions of the Royal Society B: Biological Sciences **361**, 2045 (2006).
[24] J. R. Doroghazi and D. H. Buckley, Genome biology and evolution **3**, 1349 (2011).
[25] H. Ochman, S. Elwyn, and N. A. Moran, Proceedings of the National Academy of Sciences **96**, 12638 (1999).
[26] S. Wielgoss et al., G3: Genes, Genomes, Genetics **1**, 183 (2011).
[27] X. Didelot, G. Méric, D. Falush, and A. E. Darling, BMC genomics **13**, 1 (2012).
[28] J. F. C. Kingman, Stochastic processes and their applications **13**, 235 (1982).
[29] P. G. Higgs and B. Derrida, Journal of molecular evolution **35**, 454 (1992).
[30] M. Serva, Journal of Statistical Mechanics: Theory and Experiment **2005**, P07011 (2005).
[31] K. A. Jolley and M. C. Maiden, BMC bioinformatics **11**, 595 (2010).
[32] P. H. Oliviera, T. Marie, and R. E. P. C, Proceedings of the National Academy of Sciences **0**, 0 (2016).
[33] D. J. Krause and R. J. Whitaker, Systematic biology **64**, 926 (2015).
[34] N. Takeuchi, K. Kaneko, and E. V. Koonin, G3: Genes—Genomes— Genetics **4**, 325 (2014).
[35] T. F. Cooper, PLoS Biol **5**, e225 (2007).
[36] R. A. Neher and O. Hallatschek, Proceedings of the National Academy of Sciences **110**, 437 (2013).

## A1. APPENDIX

### A. The computational model

We consider a model population of $N_e$ bacteria. The population evolves with non-overlapping generations. In each generations, offsprings choose their parent randomly. Genome of every bacteria consists of $G$ genes. The genes can mutate and can be transferred *one at a time* in their entirety. The genes in this model are in fact indivisible units of homologous recombination. We denote by $l_{\rm tr}$ the length of each gene. We use $l_{\rm tr} = 5000$ base pairs reflecting transfers larger than individual genes (6, 10). The mutation rate is $\mu$ per base pair per generation and the rate at which recombinations are attempted is $\rho$ per base pair per generation. We assume that recombinations always start at the first base pair of each gene.

In this co-evolving population, we focus on the divergence between a pair of strains $X$ and $Y$ that at time $t = 0$ start as identical twins. The divergence $\delta(t)$ on any one of the genes between these two pairs evolves stochastically as a function of time. With probability $2\mu$, the divergence increases by $1/l_{\rm tr}$. Recombinations are attempted from the population into one of the genomes (say X) at a rate $2 \times \rho$. The divergence after a recombination $\delta_a$ ($a$ for after) event can either remain the same, decrease, or increase compared to the divergence before recombination $\delta_b$ ($b$ for before). The three probabilities are given by (see Fig. A1 for an illustration).

$$p_{=}(\delta_a|\delta_b) = \frac{1 - e^{-\frac{\delta_b}{\delta_{\rm TE}} - \frac{2\delta_b}{\theta}}}{2 + \theta/\delta_{\rm TE}} \times Di(\delta_a - \delta_b),$$
$$p_{<}(\delta_a|\delta_b) = \frac{e^{-\frac{2\delta_a}{\theta} - \frac{\delta_b}{\delta_{TE}}}}{\theta} \times H(\delta_b - \delta_a) \text{ and,}$$
$$p_{>}(\delta_a|\delta_b) = \frac{e^{-\frac{\delta_a}{\delta_{\rm TE}} - \frac{\delta_a + \delta_b}{\theta}}}{\theta} \times H(\delta_a - \delta_b). \quad \text{(A1)}$$

Here, $Di(x)$ is the Dirac Delta function and $H(x)$ is the Heaviside theta function. The full evolutionary kernel $E(\delta_a|\delta_b)$ is the combination of mutational events and recombination events.

### B. Estimating $r/m$

As mentioned in the main text, $r/m$ is defined in a pair of strains as the ratio of SNPs brought in by recombination events and the SNPs brought in by point mutations. Clearly, $r/m$ will depend on a strain-to-strain comparison however, usually it is reported as an average over all pairs of strains. How do we compute $r/m$ in our framework? We have

$$r/m = \rho_{\rm succ}/\mu \times l_{\rm tr} \times \delta_{\rm tr} \quad \text{(A2)}$$

Thus, in order to compute $r/m$, we need two quantities. First, we need to compute the rate of successful recombinations $\rho_{\rm succ} < \rho$. We can calculate $\rho_{\rm succ}$ as

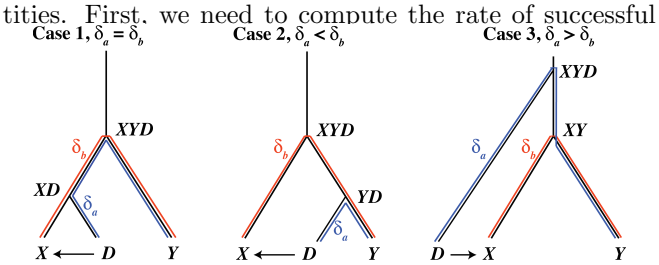


FIG. A1. Three possible outcomes of gene transfer that change the divergence $\delta$. $XD$, $YD$, $XY$, and $XYD$ are the most recent common ancestors of the strains. The divergence $\delta_b$ before transfer and $\delta_a$ after transfer are shown in red and blue respectively.

$$\rho_{\rm succ} = \int\int \frac{1}{N_e} \rho e^{-t/N_e} \times p_{\rm succ}(\delta) p(\delta|t) d\delta dt \quad \text{(A3)}$$

where $p_{\rm succ}$ is the success probability that a gene that has diverged by $\delta$ will have a successful recombination event. The integration over exponentially distributed pairwise coalescent times averages over the population. $p_{\rm succ}$ can be computed from Eq. A1 by integrating over all possible scenarios of successful recombinations. We have

$$p_{\rm succ}(\delta) = e^{-\frac{\delta^*(2+\theta^*)}{\theta^*}} \times \left( \frac{1}{1 + 3\theta^* + \theta^* \times \theta^*} - \frac{1}{2} \right) + \frac{e^{-\delta^*}}{2} + \frac{1}{2 + \theta^*} \quad \text{(A4)}$$

where $\delta^* = \delta/\delta_{\rm TE}$ and $\theta^* = \theta/\delta_{\rm TE}$ are normalized divergences and $p(\delta|t)$ is the distribution of *local* divergences at time $t$. In practice, $r/m$ can only estimated by analyzing statistics of distribution of SNPs on the genomes of closely related strain pairs where both clonally inherited and recombined parts of the genome can be identified (6, 27). Here, we limit the time-integration in Eq. A3 to times $t < \min(N_e = \theta/2\mu, \delta_{\rm TE}/2\mu)$.

Second, we need to compute the average divergence in transferred segments, $\delta_{\rm tr}$. We have

$$\delta_{\rm tr} = \frac{1}{N_e} \int\int e^{-t/N_e} \times \delta_t(\delta) p(\delta|t) dt d\delta \quad \text{(A5)}$$

where $\delta_t(\delta)$ is the average divergence *after* a recombination event if the divergence before transfer was $\delta$.

### C. Computing $\theta$ from MLST data

Except for *E. coli* where we used our previous analysis (6), we downloaded MLST sequences of multiple organisms from the MLST database (31). For each of the 7 genes present in the MLST database, we performed a pairwise alignment between strains. $\theta$ for each gene was calculated as the average of pairwise SNPs. The $\theta$ for the species was estimated as average of the $\theta$s calculated for each of the 7 genes.