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, horizontal gene transfer followed by homologous recombination is now recognized as an important and sometimes even dominant contributor to their evolution. However, the details of how the competition between clonal inheritance and recombination shapes genome diversity, population structure, and species stability remains poorly understood. Using a computational model, we find two principal regimes in bacterial evolution and identify two composite parameters that dictate the evolutionary 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 their evolution. The species as a whole lacks genetic coherence with sexually isolated clonal 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. The transition between these two regimes can be affected by relatively small changes in evolutionary parameters. Using the Multi Locus Sequence Typing (MLST) data we classify a number of well-studied bacterial species to be either the divergent or the metastable type. Generalizations of our framework to include fitness and selection, ecologically structured populations, and horizontal gene transfer of non-homologous regions are discussed.

Significance statement: Homologous recombination is now recognized as an important contributor to evolutionary dynamics and intra-species genetic diversity in many bacterial species. However, the circumstances under which these species can for the most part maintain their clonal population structure in the presence of recombination remains unknown. In this work, we identify and explore two composite evolutionary parameters giving rise to two distinct dynamic evolutionary regimes of bacterial populations and show that bacteria can belong to either of the regimes.

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 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 μ 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, the genetic exchange between closely related organisms integrated into the chromosome via homologous recombination has also been recognized as a significant factor in evolution (5, 6, 9–14). Similar to mutations, recombination events attempted at a rate ρ per base pair per generation, do not happen at every cell division. Notably recombination between genetically distant bacteria is suppressed, so that the probability $p_{\rm success} \sim e^{-\delta/\delta_{\rm TE}}$ of successful recombination of foreign DNA into a recipient genome decays exponentially with δ, the *local* divergence between the donor DNA and the recipient (12, 15–18). The effective barrier $\delta_{\rm TE}$ to successful recombination, referred here as the *transfer efficiency*, is shaped at least in part by the restriction modification (RM), the mismatch repair (MMR) systems, and the biophysical mechanisms of homologous recombination (15, 16).

While vertical inheritance along with point mutations imposes a clonal structure on the population, recombination acts as an homogenizing force, bringing individual strain genomes closer to each other 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 compe-

* Email: maslov@illinois.edu

tition between the diversity within the population θ and the maximal diversity $\delta_{\rm TE}$ within a single sub-population uniformly capable of successful recombinations. If $\delta_{\rm TE} < \theta$, 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 $\delta_{\rm TE} > \theta$, unhindered exchange of genetic fragments may result in a single cohesive population. Second is the competition between the recombination transfer rate ρ and the mutation rate μ. Consider a pair of strains diverging from each other. The average time between consecutive recombination events affecting a given small genomic region in these two strains is $1/(2\rho l_{\rm tr})$ where $l_{\rm tr}$ is the average length of transferred regions. At the same time, the total divergence accumulated in this region due to mutations in either of the two genomes is $\delta_{\rm mut} \sim 2\mu/2\rho l_{\rm tr}$. If $\delta_{\rm mut} \gg \delta_{\rm TE}$, the pair of genomes is likely to become sexually isolated from each other in this region over time separating two successive recombination events. In contrast, if $\delta_{\rm mut} < \delta_{\rm TE}$, frequent recombination events would delay sexual isolation resulting in a more homogeneous population.

What qualitative dynamical regimes in bacterial evolution emerge from the competition between these two factors and which evolutionary parameters dictate the fate of genome diversification and population structure 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 (18–21).

Some aspects of this interplay have been explored before. In their pioneering study Vetsigian and Goldenfeld (22) investigated the effects of a non-recombining segment (for example, an insertion of an auxiliary genomic island via horizontal transfer or a large-scale genomic inversion event) on recombination in its chromosomal neighborhood and how it may result in two propagating waves spreading the divergence 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. (17), working with $\theta = 0.4\%$ (lower than the value typically observed in bacterial species) and the *transfer efficiency* $\delta_{\rm TE} \approx 2.4\%$ concluded that sexual isolation in bacterial species is insufficient to cause speciation with realistic recombination frequencies. Doroghazi and Buckley (24), working with a fixed *transfer efficiency* and a very small population size (limit of $\theta \to 0$ of our study), studied how the competition between mutations and recombination affects the cohesion vs divergence 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 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 genetically 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 generated 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 real-life bacterial species belong to either divergent or metastable regimes. We discuss possible molecular mechanisms and evolutionary forces that decide the role of recombination in a species' evolutionary fate. We also discuss possible extensions of our analysis to include adaptive evolution, effects of ecological niches, and genome modifications such as insertions, deletions, and inversions.

The computational model: We consider a population of N_e co-evolving bacterial strains. The population evolves with non-overlapping generations and in each new generation each of the strains randomly chooses its parent (8). Strain genomes have G indivisible and non-overlapping transferable units. For simplicity, in what follows we refer to these units as *genes* but note that while the average protein-coding gene in bacteria is about $\sim$ 1000 base pairs (bp) long, genomes in our simulations we exchange segments of $l_{\rm tr} = 5000$bp mimicking genetic transfers longer than individual protein-coding genes (6, 10). These genes acquire point mutations at a rate μ per base pair per generation and recombinations into a recipient genome from a randomly selected donor genome in the population are attempted at a rate ρ per base pair per generation. The mutations and recombination events are assumed to have no fitness effects (later on we discuss how this assumption can be relaxed). 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 δ between the donor and the recipient.

The *effective* population sizes for real bacteria are usually too large to allow direct simulation of our model with realistic parameters. In what follows we overcome this constraint by employing an approach we had proposed earlier (6). It allows us to simulate the evolutionary dynamics of only two genomes labeled X and Y, while treating the impact of the rest of the pop-

ulation in a self-consistent manner (reminiscent of the self-consistent Hartree-Fock approximation in physics). 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 average sequence divergence of the i^{th} gene between X and Y at time t and by $\Delta(t) = 1/G \sum_i \delta_i(t)$ the genome-wide mean divergence averaged across all genes. Based on population-genetic and biophysical considerations, we derive the probability $E(\delta_a|\delta_b) = 2\mu M(\delta_a|\delta_b) + 2\rho l_{\text{tr}} R(\delta_a|\delta_b)$ (a for after and b for before) that the divergence in any gene changes from δ_b to δ_a in one generation (see supplementary materials for details) (6). Briefly, there are two components to the probability, M and R. Point mutations in either of two strains, represented by $M(\delta_a|\delta_b)$, occur at a rate 2μ per base pair per generation and increase the divergence in a gene by $1/l_{\text{tr}}$. 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 δ 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: Fig. 1 shows a typical evolutionary trajectory of the *local* divergence $\delta(t)$ of a single gene in a pair of genomes. We have used $\theta = 1.5\%$ and $\delta_{\text{TE}} = 1\%$, similar to values typically observed in bacterial species (6, 17). To keep the simulation times manageable, mutation and recombination rates used in our simulations were 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, $\delta_{\text{mut}}/\delta_{\text{TE}} = 0.04$) (25, 26) while keeping the ratio of the rates realistic (5, 6, 13, 27). Alternatively, one may interpret it as one time step in our simulations being considerably longer than a single bacterial 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, $\delta(t) \gg \delta_{\text{TE}}$. Beyond this point, this particular gene in our two strains splits into two different sexually isolated sub-clades. Any further recombination events in this region in each of two strains would be limited to their sub-clades and thus would not further change the average divergence within this gene. Conversely, the mutational drift in this region will continue to drive them further apart indefinitely.

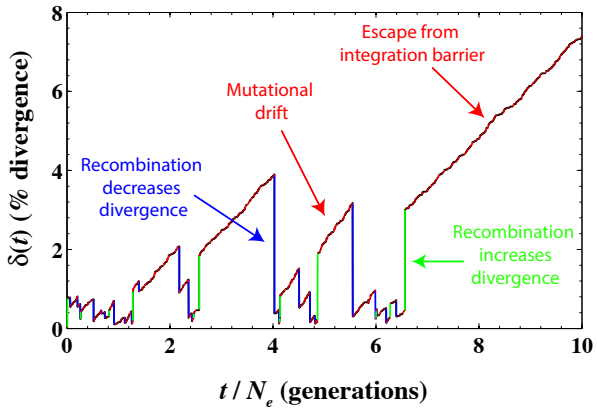

FIG. 1. A typical evolutionary trajectory of the *local* divergence $\delta(t)$ within a single gene between a pair of strains. We have used $\mu = 10^{-5}$, $\rho = 5 \times 10^{-6}$ per base pair per generation, $\theta = 1.5\%$ and $\delta_{\text{TE}} = 1\%$. Red tracks indicate the divergence increasing linearly, at a rate 2μ 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 (red stretch at the right).

Genome-wide divergence: Since genes 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 of the *local* divergences $\delta(t)$. Since the number $G \gg 1$ 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 a more noisy time course of $\delta(t)$ seen in Fig. 1.

In Fig. 2, we plot the time evolution of $\Delta(t)$ between a pair of genomes (as % difference). We have used $\theta = 0.25\%$, $\delta_{\text{TE}} = 1\%$, and $\delta_{\text{mut}}/\delta_{\text{TE}} = 2, 0.5, 0.04$, and 2×10^{-3} respectively. When $\delta_{\text{mut}}/\delta_{\text{TE}}$ is large (either a due to low ρ or a low δ_{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 2μ. For smaller values of $\delta_{\text{mut}}/\delta_{\text{TE}}$, 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, leading to a linear increase in $\Delta(t)$ with time albeit with a slope different than 2μ. Thus, the repeated resetting of individual $\delta(t)$s after homologous recombination (see Fig. 1) generally results in a $\Delta(t)$ that increases linearly (albeit extremely slowly)

with time. The rate of increase of divergence in the long time limit likely somewhat underestimates true rate of escape from the barrier in the metastable regime. Indeed, since individual genes in our model evolve in isolation, they also escape the integration barrier independent from each other. In other words in our current model, there are no correlations, long or short range order in the distribution of local divergences along the chromosome beyond the level of a single gene. As we discuss below, a more realistic model that allows overlaps between transferred regions, for example, the one studied numerically by Vetsigian and Goldenfeld (22), would offer a more accurate estimate of the escape rate.

At the shorter time scale, the trends in genome divergence are opposite to those at the longer time scale. At a fixed θ, 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.

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 recombining 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. One will have to use 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, μ, ρ, θ, and $\delta_{\rm TE}$, the time it takes for a pair of genomes to diverge far beyond the typical population diversity θ can serve as a quantifier for metastability.

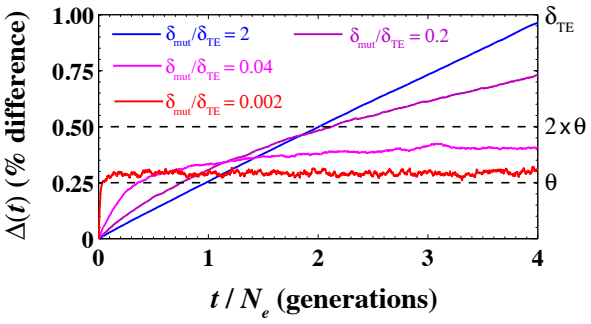

FIG. 2. Genome-wide divergence $\Delta(t)$ as a function of time at $\theta/\delta_{\rm TE} = 0.25$. We have used $\delta_{\rm TE} = 1\%$, $\theta = 0.25\%$, $\mu = 5 \times 10^{-6}$ per base pair per generation and $\rho = 2.5 \times 10^{-8}, 2.5 \times 10^{-7}, 1.25 \times 10^{-6}$, and 2.5×10^{-5} per base pair per generation corresponding to $\delta_{\rm mut}/\delta_{\rm TE} = 2$ (blue), $0.2, 0.04$ (shades of purple), and 2×10^{-3} (red) respectively.

In Fig. 3, we plot the number of generations $t_{\rm div}$ (in units of the effective population size N_e) required for the genome-wide average divergence $\Delta(t)$ between a pair of genomes to exceed $2 \times \theta$ (twice the typical intra-population diversity) 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 $2\theta = 4\mu N_e$. Our results remain qualitatively the same for other thresholds $k \times \theta$ provided that $k > 1$. The four cases explored in Fig. 2 are marked with green diamonds in Fig. 3.

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 e.g. $\delta_{\rm mut}/\delta_{\rm TE} = 2$ in Fig. 2 and Fig. 3) and reaches $\Delta(t) = 2\theta$ within $\sim 2N_e$ generations. In contrast for smaller values of $\delta_{\rm mut}/\delta_{\rm TE}$ in the *metastable* regime, it takes extremely long time for $\Delta(t)$ to reach 2θ. In this regime genes get repeatedly exchanged with the rest of the population and $\Delta(t)$ remains nearly constant over long periods of time (see e.g. $\delta_{\rm mut}/\delta_{\rm TE} = 2 \times 10^{-3}$ in Fig. 2 and Fig. 3). Notably, near the boundary region between the two regimes a small perturbation in the evolutionary parameters could change the evolutionary dynamics from divergent to metastable and vice versa.

Population structure: Can we understand the phylogenetic structure of the entire population by studying the evolutionary dynamics of just a single pair of strains? Or using a quote from William Blake, is it possible "to see a World in a Grain of Sand"?

Given sufficient amount of time every pair of genomes

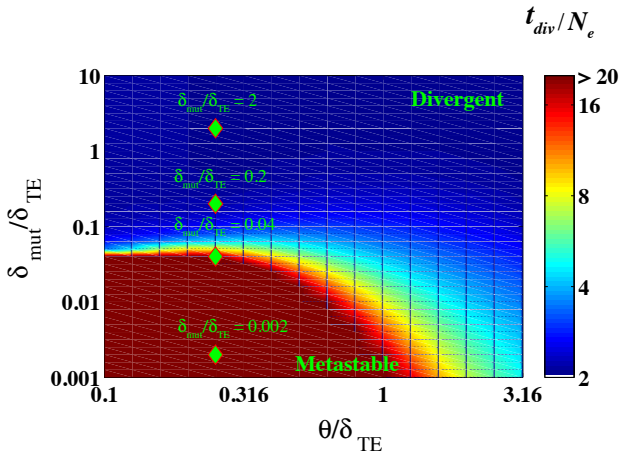

FIG. 3. The number of generations $t_{\rm div}$ (in units of the effective population size N_e) required for a pair of genomes to diverge well beyond the average intra-population diversity (see main text). We calculate the time it takes for the genome-wide average divergence to reach 2θ as a function 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. In our simulations we varied ρ and θ 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).

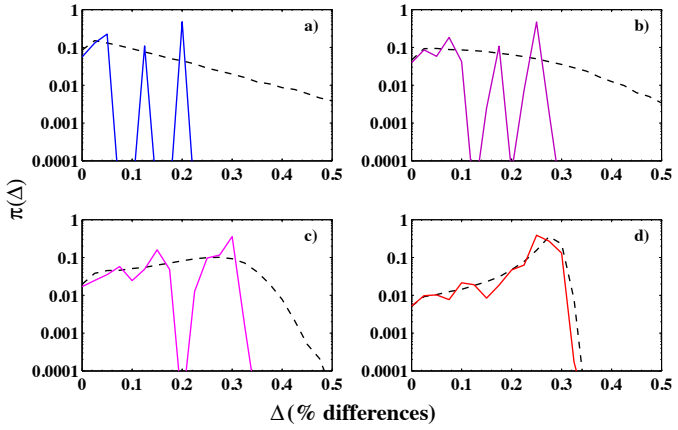

FIG. 4. Distribution of all pairwise genome-wide divergences δ_{ij} in a co-evolving population for decreasing values of $\delta_{\rm mut}/\delta_{\rm TE}$: 2 in a), 0.2 in b), 0.04 in c) and 0.002 in d) In all 4 panels, dashed black lines represent time-averaged distributions $\overline{\pi(\Delta)}$, while solid lines represent typical "snapshot" distributions $\pi(\Delta)$ in a single population. Colors of solid lines match those in Fig. 2 for the same values of parameters. Time-averaged and snapshot distributions were estimated by sampling 5×10^5 pairwise coalescent times from the time-averaged coalescent distribution $p \sim e^{-t/N_e}$ and the instantaneous coalescent distribution $p_{\rm c}(t)$ correspondingly (see text for details).

in our model would diverge indefinitely. However, 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, $\overline{p_{\rm c}(t)} \sim e^{-t/N_e}$ (here and below we use the bar to denote averaging over multiple realizations of the coalescent process, or long-time average over population dynamics) (28–30). Thus, while it may be possible for 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 pair in a finite-sized population.

Let $\pi(\Delta)$ to denote the probability distribution of Δ for all pairs of genomes in a given population, while $\overline{\pi(\Delta)}$ stands for the same distribution averaged over long time or multiple realizations of the population. One has

$$\begin{aligned}\pi(\Delta) &= \int_0^\infty p_{\rm c}(t) \times p(\Delta|t)dt \text{ and} \\ \overline{\pi(\Delta)} &= \int_0^\infty \overline{p_{\rm c}(t)} \times p(\Delta|t)dt \\ &= \frac{1}{N_e}\int_0^\infty e^{-t/N_e} \times p(\Delta|t)dt \qquad (2)\end{aligned}$$

In Eq. 2, $p_{\rm c}(t)$ is the probability that a pair of strains in the current population shared their MRCA t generations ago and $p(\Delta|t)$ is the probability that a pair of strains have diverged by Δ at time t. Given that $\Delta(t)$ is the average of $G \gg 1$ independent realizations of $\delta(t)$, we can approximate $p(\Delta|t)$ as a Gaussian distribution with average $\langle\delta(t)\rangle_G = \int \delta \times p(\delta|t)d\delta$ and variance $\sigma^2 = \frac{1}{G}\left(\langle\delta(t)^2\rangle_G - \langle\delta(t)\rangle_G^2\right)$. Here and below angular brackets and the subscript G denote the average of a quantity over the entire genome.

Unlike the time- or realization- averaged distribution $\overline{\pi(\Delta)}$, only the instantaneous distribution $\pi(\Delta)$ is accessible from genome sequences stored in databases. Indeed, we rarely have the luxury of observing real-life bacterial populations over evolutionary significant stretches of time. Even for large populations these two distributions could be significantly different from each other. Indeed, $p_{\rm c}(t)$ in any given population is extremely noisy due to multiple peaks from clonal subpopulations and does not resemble its smooth long-time average profile $\overline{p_{\rm c}(t)} \sim e^{-t/N_e}$ (29, 30). In panels a) to d) of Fig. 4, we show $\pi(\Delta)$ for the four cases shown in Fig. 2 (also marked by green diamonds in Fig. 3). We fixed the population size to $N_e = 500$. We changed $\delta_{\rm mut}/\delta_{\rm TE}$ by changing the recombination rate ρ. The solid lines represent a time snapshot obtained by numerically sampling $p_{\rm c}(t)$ 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 values of $\delta_{\rm mut}/\delta_{\rm TE} = \mu/(\rho l_{tr}\delta_{TE})$, the *instantaneous* snapshot distribution $\pi(\Delta)$ has multiple peaks corresponding to divergence distances between several spontaneously formed clonal sub-populations *present even in a homogeneous population*. These sub-populations rarely exchange genetic material with each other, either because of a low recombination frequency ρ or due to strict barriers for recombination (small δ_{TE}). In this regime, the time averaged distribution $\overline{\pi(\Delta)}$ has a long exponential tail and,

as expected, does not at all agree with the instantaneous distribution $\pi(\Delta)$.

Notably, in the metastable regime, at lower values of $\delta_{\rm mut}/\delta_{\rm TE}$, 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 dependent on 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 successfully eliminate multiple peaks due to clonal subs-populations thus forming a genetically cohesive and temporally stable population.

Analysis of bacterial species: Where are real-life bacterial species located on the divergent-metastable diagram? Instead of $\delta_{\rm mut}/\delta_{\rm TE}$ as defined here, population genetic studies of bacteria usually quantify the relative strength of recombination over mutations as r/m, the ratio of the number of SNPs brought in by recombination relative to those generated by point mutations in a pair of closely related strains (6, 9, 13). In our framework, r/m is defined as $r/m = \rho_{\rm succ}/\mu \times l_{\rm tr} \times \delta_{\rm tr}$ where $\rho_{\rm succ} < \rho$ is the rate of successful recombination events and $\delta_{\rm tr}$ is the average divergence in transferred regions. Both $\rho_{\rm succ}$ and $\delta_{\rm tr}$ depend on the evolutionary parameters (see appendix for a detailed description of our calculations). r/m is closely related (but not equal) to the inverse of $\delta_{\rm mut}/\delta_{\rm TE}$ used in our previous plots.

In Fig. 5, we re-plotted the "phase diagram" shown in Fig. 3 in terms of $\theta/\delta_{\rm TE}$ and r/m and attempted to place several real-life bacterial species on it. To this end we estimated θ from the MLST data (31) and used r/m values that were determined previously by Vos and Didelot (13). As a first approximation, we assumed that the transfer efficiency $\delta_{\rm TE}$ is the same for all species considered and is given by $\delta_{\rm TE} \sim 2.26\%$ used in Ref. (17). However, as mentioned above, the transfer efficiency $\delta_{\rm TE}$ depends in part on the RM and the MMR systems. Given that these systems vary a great deal across bacterial species including minimal barriers to recombination observed e.g. in *Helicobacter Pylori* (11) or different combinations of multiple RM systems reported in Ref. (32). Thus one expects transfer efficiency δ_{TE} might also vary across bacteria. Further work is needed to collect the extent of this variation in a unified format and location. One possible bioinformatics strategy is to use the slope of the exponential tail in SNP distribution ($p(\delta|\Delta)$ in our notation) to infer the transfer efficiency δ_{TE} as described in Ref. (6).

Fig. 5 allows one to draw the following conclusions. First, it confirms that both r/m and $\theta/\delta_{\rm TE}$ are important evolutionary parameters and suggests that each of them alone cannot fully classify a species as either divergent or metastable. Second, it predicts a sharp transition between the divergent and the metastable phases implying that a small change in r/m or $\theta/\delta_{\rm TE}$ can change the evolutionary fate of the species. And finally, one can

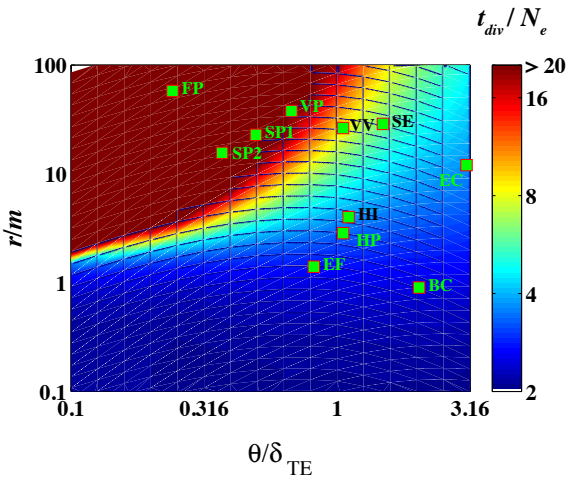

FIG. 5. Approximate position of several real-life bacterial spaces on the metastable-divergent phase diagram (see text for details). Abbreviations of species names are as follows: FP: *Flavobacterium psychrophilum*, VP: *Vibrio parahaemolyticus*, SE: *Salmonella enterica*, VV: *Vibrio vulnificus*, SP1: *Streptococcus pneumoniae*, SP2: *Streptococcus pyogenes*, HP: *Haemophilus parasuis*, HI: *Haemophilus influenzae*, BC: *Bacillus cereus*, EF: *Enterococcus faecium*, and EC: *Escherichia coli*.

see that different bacterial species use diverse evolutionary strategies straddling the divide between these two regimes.

Can bacteria change their evolutionary fate? 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 the effective population size remains constant, a change in mutation rate changes both $\delta_{\rm mut}/\delta_{\rm TE}$ as well as θ. A change in the level of expression of the MMR genes, changes in types or presence of MMR, SOS, or restriction-modification (RM) systems, loss or gain of co-infecting phages, all could change $\delta_{\rm TE}$ or the rate of recombination (15, 32) thus changing the placement of the species on the phase diagram shown in Fig. 5.

Adaptive and ecological events should be inferred from population genomics data only after rejecting the hypothesis of *neutral* evolution. However, the range of behaviors consistent with the neutral model of recombination-driven evolution of bacterial species was not entirely quantified up till now, 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 evolutionary sub-clades (groups A, B1, B2, E1, E2, and D). It is thought that inter-clade sexual exchange is lower compared to intra-clade exchange (6, 27). Ecological niche separation and/or selective advantages are usually implicated as initiators of such putative speciation events (18). 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 deeply in the divergent regime in Fig. 5. Thus, based on the analysis presented above one expects *E. coli* strains to spontaneously form transient sexually-isolated sub-populations even in the absence of selective pressures or ecological niche separation. In conclusion, a more careful analysis is needed to reject neutral models of evolution in the studies of population genetics of bacteria.

Conclusions: While recombination is now recognized as an important and sometimes even dominant contributor to patterns of genome diversity in many bacterial species(5, 6, 9–13), its effect on population structure and 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 competing processes. 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 species. Notably, understanding the time course of divergence between a single pair of genomes allows us to study the structure of the entire population. As shown in Fig. 4, species in the divergent regime are characterized by multi-peaked 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 separating them from each other.

Extending the framework: Throughout this study we used three main assumptions greatly simplifying the problem and allowing for exact mathematical analysis: i) exponentially decreasing probability of successful integration of foreign DNA into a recipient genome as a function of the local sequence divergence, ii) exponentially distributed pairwise coalescent time distribution of a neutrally evolving well-mixed population, and iii) independent evolution of non-overlapping "genes" or larger indivisible units of horizontal genetic transfer. Here we discuss how one can generalize the developed framework to incorporate phenomena violating assumptions.

(i) A wide variety of barriers to foreign DNA entry exist in bacteria (12). For example, *Helicobacter pylori*, is thought to have relatively free import of foreign DNA (11) while other bacteria may have multiple RM systems that either act in combination or are turned on and off randomly (32) leading to potentially non-exponential dependence of the probability of successful integration on local genetic divergence. One can incorporate these variations within our framework by appropriately modifying the functional relationship between the probability of successful integration and local sequence divergence or even by allowing it to change with time (e.g. relax recombination barriers in the presence of stress).

(ii) Bacteria belong to ecological niches defined by environmental factors such as availability of specific nutrient sources, host-bacterial interactions, and geographical characteristics. Bacteria in different environments may rarely compete with each other for resources and consequently may not belong to the same *effective* population and may have lowered frequency of DNA exchange compared to bacteria sharing the same niche. How can one capture the effect of ecological niches on genome evolution? Geographically and/or ecologically structured populations exhibit a coalescent structure (and thus a pairwise coalescence time distribution) that depends on the nature of niche separation (34, 35). Within our framework, niche-related effects can be incorporated by accounting for pairwise coalescent times of niche-structured populations (34, 35) and niche dependent recombination frequencies. For example, one can consider a model with two or more subpopulations with different probabilities for intra- and inter-population DNA exchange describing geographical or phage-related barriers to recombination.

While most point mutations in bacterial genomes are thought to have insignificant fitness effect, the evolutionary dynamics of bacterial species is driven by rare advantageous mutations and thus is far from being neutral. Recombination in bacterial species is thought to be essential for their evolution in order to minimize the fitness loss due to Muller's ratchet (36) and to minimize the impact of clonal interference (37). Thus, it is likely that both recombination frequency and transfer efficiency are under selection (36, 38, 39). How could one include fitness effects in our theoretical framework? Above, we considered the dynamics of *neutrally* evolving bacterial populations. The effective population size is incorporated in our framework only via the coalescent time distribution $\exp(-T/N_e)$ and consequently the intra-species diversity $\exp(-\delta/\theta)$ (see supplementary materials). Neher and Hallatschek (40) recently showed that while pairwise coalescent times in adaptive populations are *not exactly* exponentially distributed, this distribution has a pronounced exponential tail with an *effective* population size N_e weakly related to the actual census population size and largely determined by the variance of mutational fitness effects (40). In order to modify the recombination kernel $R(\delta_a|\delta_b)$ one needs to know the 3-point coalescence distribution for strains X, Y, and the donor strain D (see Supplementary Materials here and in Ref. (6) for details). Once such 3-point coalescence distribution is available in either analytical or even numerical form our results could be straightforwardly generalized for adaptive populations (assuming most genes remain neutral). We expect the phase diagram of thus modified adaptive model to be similar to its neutral predecessor considered here, given that the pairwise coalescent time distribution in adaptive population has an exponential tail as well (40), and for our main results to remain qualitatively unchanged.

(iii) Finally, in this work, we assumed that in a recipient genome, recombinations mix *non-overlapping* segments of length $l_{\rm tr}$ that *always 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, 22). Overlapping transfers can affect the evolutionary dynamics. In particular, when a local region within a genome has diverged above the threshold imposed by the transfer efficiency δ_{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 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 the transfer efficiency threshold can similarly result in two propagating wave fronts of divergence along the chromosome. Our analytical results are rather similar to numerical findings of Ref. (22) indicating that such propagating fronts do not qualitatively change the two regimes of the evolutionary dynamics explored above.

In our future studies we plan to explore these and other extensions on top of the basic mathematically tractable model described here.

Acknowledgments: We would like to thank Kim Sneppen, Erik van Nimwegen, Daniel Falush, Nigel Goldenfeld, Eugene Koonin, and Yuri Wolf for fruitful discussions and comments that lead to an improved manuscript.

[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] C. Fraser, W. P. Hanage, and B. G. Spratt, Science **315**, 476 (2007).
[18] M. F. Polz, E. J. Alm, and W. P. Hanage, Trends in Genetics **29**, 170 (2013).
[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. Oliveira, M. Touchon, and E. P. Rocha, Proceedings of the National Academy of Sciences **113**, 5658 (2016).
[33] D. J. Krause and R. J. Whitaker, Systematic biology **64**, 926 (2015).
[34] N. Takahata, Genetics **129**, 585 (1991).
[35] J. Wakeley, Journal of Heredity **95**, 397 (2004).
[36] N. Takeuchi, K. Kaneko, and E. V. Koonin, G3: Genes—Genomes— Genetics **4**, 325 (2014).
[37] T. F. Cooper, PLoS Biol **5**, e225 (2007).
[38] A. E. Lobkovsky, Y. I. Wolf, and E. V. Koonin, Genome biology and evolution **8**, 70 (2016).
[39] J. Iranzo, P. Puigbo, A. E. Lobkovsky, Y. I. Wolf, and E. V. Koonin, Genome Biology and Evolution , evw193 (2016).
[40] 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 generation, the parent of an offsprings is chosen randomly from the previous generation. 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 μ per base pair per generation and the rate at which recombinations are attempted is ρ 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 2μ, 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 δ_a (a for after) event can either remain the same, decrease, or increase compared to the divergence before recombination δ_b (b for before). The three probabilities are given by (see Fig. A1 for an illustration).

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

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

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

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

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

where $p_{\rm succ}$ is the success probability that a gene that has diverged by δ 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

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

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

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

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

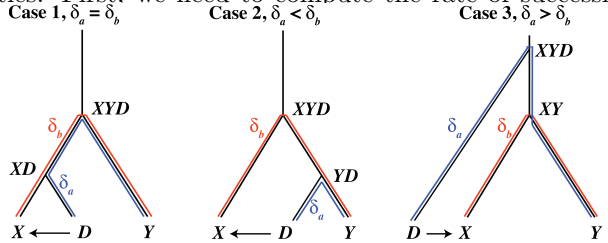

FIG. A1. Three possible outcomes of gene transfer that change the divergence δ. XD, YD, XY, and XYD are the

C. Computing θ from MLST data

Except for *E. coli* where we used our previous analysis (6) (we used $\theta/\delta_{\rm TE} \sim 3$ and $r/m = 12$), 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. θ for each gene was calculated as the average of pairwise SNPs. The θ for the species was estimated as average of the θs calculated for each of the 7 genes.