Genetics: Early Online, published on July 27, 2017 as 10.1534/genetics.117.300061

Recombination-driven genome evolution and stability of bacterial species

Purushottam D. Dixit[*,1], Tin Yau Pang[†] and Sergei Maslov[‡]

[*]Department of Systems Biology, Columbia University, New York, NY 10032, [†]Institute for Bioinformatics, Heinrich-Heine-Universität Düsseldorf, 40221 Düsseldorf, [‡]Department of Bioengineering and Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana IL 61801

ABSTRACT While bacteria divide clonally, horizontal gene transfer followed by homologous recombination is now recognized as an important contributor to their evolution. However, the details of how the competition between clonality and recombination shapes genome diversity 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 pairs of genomes in the population steadily increases in the course of their evolution. The species 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 temporally stable. Notably, the transition between these two regimes can be affected by relatively small changes in evolutionary parameters. Using the **M**ulti **L**ocus **S**equence **T**yping (MLST) data we classify a number of bacterial species to be either the divergent or the metastable type. Generalizations of our framework to include selection, ecologically structured populations, and horizontal gene transfer of non-homologous regions are discussed as well.

KEYWORDS Bacterial evolution, Recombination, Population genetics

1. 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 (MEDINI *et al.* 2005; TETTELIN *et al.* 2005; HOGG *et al.* 2007; LAPIERRE and GOGARTEN 2009; TOUCHON *et al.* 2009; DIXIT *et al.* 2015; MARTTINEN *et al.* 2015).

Multiple factors shape the diversification of the core genome. For example, point mutations generate single nucleotide polymorphisms (SNPs) within the population that are passed on from mother to daughter. At the same time, stochastic elimination of lineages leads to fixation of polymorphisms which effectively reduces population diversity. The balance between point mutations and fixation determines the average number of genetic differences between pairs of individuals in a population, often denoted by θ.

During the last two decades, exchange of genetic fragments between closely related organisms has also been recognized as a significant factor in bacterial evolution (GUTTMAN and DYKHUIZEN 1994; MILKMAN 1997; FALUSH *et al.* 2001; THOMAS and NIELSEN 2005; TOUCHON *et al.* 2009; VOS and DIDELOT 2009; STUDIER *et al.* 2009; DIXIT *et al.* 2015). Transferred fragments are integrated into the recipient chromosome via homologous recombination. Notably recombination between pairs of strains is limited by the divergence in transferred regions. The probability $p_{\text{success}} \sim e^{-\delta/\delta_{\text{TE}}}$ of successful recombination of foreign DNA into a recipient genome decays exponentially with δ, the local divergence between the donor DNA fragment and the corresponding DNA on the recipient chromosome (VULIĆ *et al.* 1997; MAJEWSKI 2001; THOMAS and NIELSEN 2005; FRASER *et al.* 2007; POLZ *et al.* 2013). Segments with divergence δ greater than divergence δ_{TE} have negligible probability of successful recombination. In this work, we refer to the divergence δ_{TE} as the transfer efficiency. δ_{TE} is shaped at least in part by the restriction modification (RM), the mismatch repair (MMR) systems, and the biophysical mechanisms of homologous recombination (VULIĆ *et al.* 1997; MAJEWSKI 2001). The transfer efficiency δ_{TE} imposes an effective limit on the divergence among subpopulations that can successfully exchange genetic material with each other (VULIĆ *et al.* 1997; MAJEWSKI 2001).

In this work, we develop an evolutionary theoretical framework that allows us to study in broad detail the nature of competition between recombinations and point mutations across a range of evolutionary parameters. We identify two composite parameters that govern how genomes diverge from each other over time. Each of the two parameters corresponds to a competition between vertical inheritance of polymorphisms and their horizontal exchange via homologous recombination.

First is the competition between the recombination rate ρ and the mutation rate μ. Within a co-evolving population, consider a pair of strains diverging from each other. The average time between consecutive recombination events affecting any given small genomic region is $1/(2\rho l_{\text{tr}})$ where l_{tr} is the average length of transferred regions. The total divergence accumulated in this region due to mutations in either of the two genomes is $\delta_{\text{mut}} \sim 2\mu/2\rho l_{\text{tr}}$. If $\delta_{\text{mut}} \gg \delta_{\text{TE}}$, the pair of genomes is likely to become sexually isolated from each other in this region within the time that separates two successive recombination events. In contrast, if $\delta_{\text{mut}} < \delta_{\text{TE}}$, frequent recombination events would delay sexual isolation resulting in a more homogeneous population. Second is the competition between the population diversity θ and δ_{TE}. If $\delta_{\text{TE}} \ll \theta$, 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 $\delta_{\text{TE}} \gg \theta$, unhindered exchange of genetic fragments may result in a single cohesive population.

Using computational models, we show that the two composite parameters identified above, $\theta/\delta_{\text{TE}}$ and $\delta_{\text{mut}}/\delta_{\text{TE}}$, determine qualitative evolutionary dynamics of bacterial species. Furthermore, we identify two principal regimes of this dynamics. In the divergent regime, characterized by a high $\delta_{\text{mut}}/\delta_{\text{TE}}$, 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 $\delta_{\text{mut}}/\delta_{\text{TE}}$ and a low $\theta/\delta_{\text{TE}}$), 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 $\theta/\delta_{\text{TE}}$. 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.

2. Computational models

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 (GILLESPIE 2010). As a result, the population remains constant over time. Strain genomes have $l_G = 5 \times 10^6$. Individual base pairs acquire point mutations at a constant rate μ and recombination events are attempted at a constant rate ρ (see panel a) of Figure. 1). The mutations and recombination events are assumed to have no fitness effects (later, we discuss how this assumption can be relaxed). The probability of a successful integration of a donor gene decays exponentially, $p_{\text{success}} \sim e^{-\delta/\delta_{\text{TE}}}$, with the local divergence δ between the donor and the recipient. Table 1 lists all important parameters in our model.

Unlike point mutations that occur anywhere on the genome, genomic segments involved in recombination events have a well defined starting point and length. In order to understand the effect of these two factors, below we introduce three variants of a model of recombination with increasing complexity illustrated in panel b) of Figure 1. In the first and the only mathematically tractable model we fix both length and start/end points of recombined segments. In the second model, recombined segments have a fixed length but variable starting/ending positions. Finally, in the most realistic third model, recombined segments have variable lengths (drawn from an exponential distribution with an average of 5000 bp (DIXIT *et al.* 2015)) and variable starting/ending positions. *Prima facie*, these three models appear quite distinct from each other, potentially leading to divergent conclusions about the distribution of diversity on the genome. In particular, one might assume that the first model in which different segments recombine (and evolve) completely independently from each other would lead to significantly different evolutionary dynamics than the other two models. This assumption was not confirmed by our numerical simulations. Indeed, later in the manuscript we demonstrate (see Figure 7 below) that all three variants of the model have rather similar evolutionary dynamics. In what follows we first present our mathematical description and simulations of the first model and then compare and contrast it to other two models.

The *effective* population sizes of real bacteria are usually large (TENAILLON *et al.* 2010). This prohibits simulations with realistic parameters wherein genomes of individual bacterial strains are explicitly represented. In what follows (recombination model 1) we overcome this limitation by employing an approach we had proposed earlier DIXIT *et al.* (2015). It allows us to simulate the evolutionary dynamics of only two genomes (labeled X and Y), while representing the rest of the population using evolutionary theory (DIXIT *et al.* 2015). 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 sequence divergence of the i^{th} transferable unit (or gene) between X and Y at time t and by $\Delta(t) = 1/G \sum_i \delta_i(t)$ the genome-wide divergence.

Based on population-genetic and biophysical considerations, we derive the transition 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 (DIXIT *et al.* 2015). 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

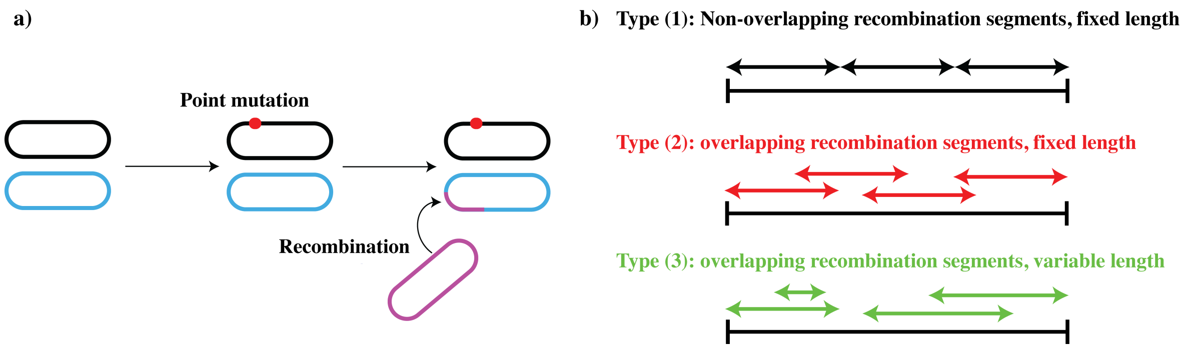
a)
Point mutation
Recombination
b) Type (1): Non-overlapping recombination segments, fixed length
Type (2): overlapping recombination segments, fixed length
Type (3): overlapping recombination segments, variable length

Figure 1 Schematic of the computation models. Panel a) Illustration of the numerical model. N_e bacterial organisms evolve together, we show only one pair of strains. Point mutations (red circles) occur at a fixed rate μ per base pair generation and genetic fragments of length l_{tr} are transferred between organisms at a rate ρ per base pair per generation. Panel b) The schematics of the three models of recombination. In model (1), recombining stretches have fixed end points. As a result, different recombination tracks do not overlap. In models (2) and (3), the recombining stretches have variable end points and as a result different recombination tracks can potentially overlap with each other.

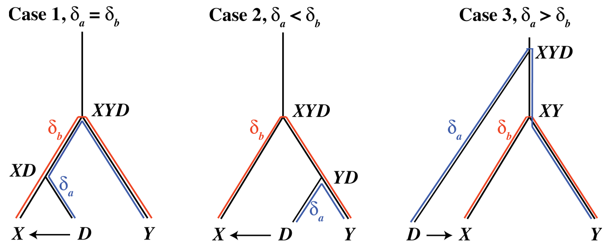
Case 1, $\delta_a = \delta_b$
Case 2, $\delta_a < \delta_b$
Case 3, $\delta_a > \delta_b$
XYD
XD
δ_b
δ_a
X
D
Y
XYD
YD
δ_b
δ_a
X
D
Y
XYD
XY
δ_a
δ_b
D
X
Y

Figure 2 Three possible outcomes of gene transfer that change the divergence δ. XD, YD, XY, and XYD are the most recent common ancestors of the strains. The divergence δ_b before transfer and δ_a after transfer are shown in red and blue respectively.

divergence in a gene by $1/l_{\mathrm{tr}}$. Hence when $\delta_a \neq \delta_b$,

$$\begin{aligned} M(\delta_a|\delta_b) &= 2\mu \text{ if } \delta_a = \delta_b + 1/l_{\mathrm{tr}} \text{ and} && (1) \\ M(\delta_b|\delta_b) &= 1 - 2\mu. && (2) \end{aligned}$$

We assume, without loss of generality, that recombination from donor strain D replaces a gene on strain X. Unlike point mutations, after a recombination, local divergence between X and Y can change suddenly, taking values either larger or smaller than the current divergence (see Figure 2 for an illustration) (Dixit *et al.* 2015). We have the probabilities $R(\delta_a|\delta_b)$ (Dixit *et al.* 2015),

$$\begin{aligned} R(\delta_a|\delta_b) &= \frac{1}{\Omega}\frac{1-e^{-\frac{\delta_b}{\delta_{\mathrm{TE}}}-\frac{2\delta_b}{\theta}}}{2+\theta/\delta_{\mathrm{TE}}} \text{ if } \delta_a = \delta_b \\ R(\delta_a|\delta_b) &= \frac{1}{\Omega}\frac{e^{-\frac{2\delta_a}{\theta}-\frac{\delta_b}{\delta_{TE}}}}{\theta} \text{ if } \delta_a < \delta_b \text{ and} \\ R(\delta_a|\delta_b) &= \frac{1}{\Omega}\frac{e^{-\frac{\delta_a}{\delta_{\mathrm{TE}}}-\frac{\delta_a+\delta_b}{\theta}}}{\theta} \text{ if } \delta_a > \delta_b. && (3) \end{aligned}$$

In Eqs. 3, Ω is the normalization constant.

3. Computational analysis of the simplified model of recombination

A. Evolutionary dynamics of local divergence has large fluctuations

Figure 3 shows a typical stochastic evolutionary trajectory of the local divergence $\delta(t)$ of a single gene in a pair of genomes simulated using $E(\delta_a|\delta_b)$. We have used realistic values of $\theta = 1.5\%$ and $\delta_{\mathrm{TE}} = 1\%$ (Fraser *et al.* 2007; Dixit *et al.* 2015). Mutation and recombination rates (per generation) in real bacteria are extremely small (Dixit *et al.* 2015). In order 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, $\delta_{\mathrm{mut}}/\delta_{\mathrm{TE}} = 0.04$) (Ochman *et al.* 1999; Wielgoss *et al.* 2011) while keeping the ratio of the rates realistic (Touchon *et al.* 2009; Vos and Didelot 2009; Didelot *et al.* 2012; Dixit *et al.* 2015). Alternatively, one may interpret it as one time step in our simulations being considerably longer than a single bacterial generation.

As seen in Figure 3, 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. 3). Eventually, either through the gradual mutational drift or a sudden recombination event, $\delta(t)$ increases beyond the integration barrier set by the transfer efficiency, $\delta(t) \gg \delta_{\mathrm{TE}}$. 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 would be limited to their sub-clades and thus would not further change the divergence within this gene. At the same time, the mutational drift in this region will continue to drive the two strains further apart indefinitely.

In Figure 4, we plot the time evolution of $\Delta(t)$ and its ensemble average $\langle\Delta(t)\rangle$ (as % difference). We have used $\theta = 0.25\%$, $\delta_{\mathrm{TE}} = 1\%$, and $\delta_{\mathrm{mut}}/\delta_{\mathrm{TE}} = 2, 0.2, 0.04$, and 2×10^{-3} respectively. As seen in Figure 4, when $\delta_{\mathrm{mut}}/\delta_{\mathrm{TE}}$ is large, in any local genomic region, multiple mutations are acquired between two successive recombination events. Consequently, individual genes escape

parameter	symbol
population diversity	$\theta = 2\mu N_e$ $(0.1\% - 3.16\%)$
mutation rate	μ $(2 \times 10^{-6}$ per base pair per generation$)$
recombination rate	ρ $(2 \times 10^{-9} - 2 \times 10^{-5}$ per base pair per generation$)$
transfer efficiency	δ_{TE} $(0.5\% - 5\%)$
length of transferred regions	l_{tr} (5000 base pairs)
number of transferable units	G (1000)

Table 1 A list of parameters in the model. The range of values used in this study are indicated in the parentheses.

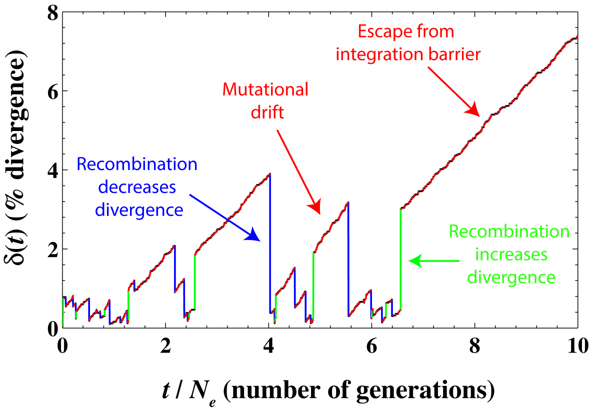

Figure 3 Stochastic evolution of local divergence. A typical evolutionary trajectory of the local divergence $\delta(t)$ within a single gene between a pair of strains. We have used $\mu = 10^{-5}$, $\rho = 5 \times 10^{-6}$ per gase pair per generation, $\theta = 1.5\%$ and $\delta_{\mathrm{TE}} = 1\%$. Red tracks indicate the divergence increasing linearly, at a rate 2μ per base pair 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).

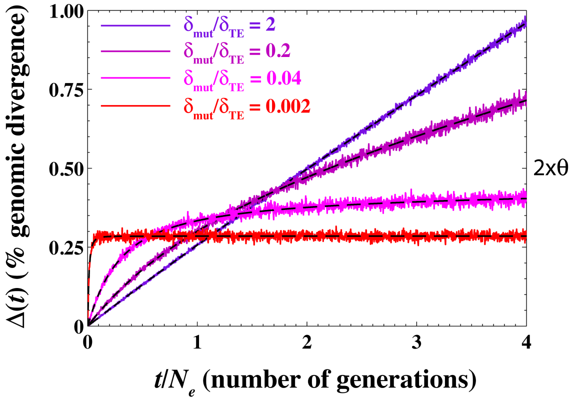

Figure 4 Stochastic evolution of genome-wide divergence. Genome-wide divergence $\Delta(t)$ as a function of time at $\theta/\delta_{\mathrm{TE}} = 0.25$. We have used $\delta_{\mathrm{TE}} = 1\%$, $\theta = 0.25\%$, $\mu = 2 \times 10^{-6}$ per base pair per generation and $\rho = 2 \times 10^{-8}, 2 \times 10^{-7}, 10^{-5}$, and 2×10^{-5} per base pair per generation corresponding to $\delta_{\mathrm{mut}}/\delta_{\mathrm{TE}} = 2, 0.2, 0.04$ and 2×10^{-3} respectively. The dashed black lines represent the ensemble average $\langle\Delta(t)\rangle$. See Figure A1 in the appendix for the evolution of $\Delta(t)$ over a longer time scale.

the pull of recombination rapidly and $\langle\Delta(t)\rangle$ increases roughly linearly with time at a rate 2μ. For smaller values of $\delta_{\mathrm{mut}}/\delta_{\mathrm{TE}}$, the rate of change of $\langle\Delta(t)\rangle$ 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 $\langle\Delta(t)\rangle$ with time albeit with a slope different than 2μ. Thus, the repeated resetting of individual $\delta(t)$s after homologous recombination (see Figure 3) generally results in a $\langle\Delta(t)\rangle$ that increases linearly with time.

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 $\delta_{\mathrm{mut}}/\delta_{\mathrm{TE}}$ 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 Figure 4). 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 faithfully represents the time that has elapsed since their Most Recent Common Ancestor (MRCA). However, Figure 3 and Figure 4 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 Dixit *et al.* (2015) for more details). At the level of genomes, when $\delta_{\mathrm{mut}}/\delta_{\mathrm{TE}}$ is large (e.g. the blue trajectory in Figure 4), the time since MRCA of two strains is directly correlated with the number of mutations that separate their genomes. In contrast, when $\delta_{\mathrm{mut}}/\delta_{\mathrm{TE}}$ is small (see pink and red trajectories in Figure 4), 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.

B. Quantifying metastability

How does one quantify the metastable behavior described above? Figure 4 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 δ_{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.

In Figure 5, we plot the number of generations t_{div} (in units of the effective population size N_e) required for the ensemble average of the genome-wide average divergence $\langle\Delta(t)\rangle$ between a pair of genomes to exceed $2 \times \theta$ (twice the typical intra-population diversity) as a function of $\theta/\delta_{\mathrm{TE}}$ and $\delta_{\mathrm{mut}}/\delta_{\mathrm{TE}}$. Analyzing the ensemble average $\langle\Delta(t)\rangle$ (represented by dashed lines in Figure 4) allows us to avoid the confounding effects of small fluctuations in the stochastic time evolution of $\Delta(t)$ around this average. Note that in the absence of recombination, it takes $t_{\mathrm{div}} = 2N_e$ generations before $\langle\Delta(t)\rangle$ exceeds $2\theta = 4\mu N_e$. The four cases explored in Figure 4 are marked with green diamonds in Figure 5.

We observe two distinct regimes in the behavior of t_{div} as a function of $\theta/\delta_{\mathrm{TE}}$ and $\delta_{\mathrm{mut}}/\delta_{\mathrm{TE}}$. 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, $\langle\Delta(t)\rangle$ increases linearly with time (see e.g. $\delta_{\mathrm{mut}}/\delta_{\mathrm{TE}} = 2$ in Figure 4 and Figure 5) and reaches $\langle\Delta(t)\rangle = 2\theta$ within $\sim 2N_e$ generations. In contrast for smaller values of $\delta_{\mathrm{mut}}/\delta_{\mathrm{TE}}$ in the metastable regime, it takes extremely long time for $\langle\Delta(t)\rangle$ to reach 2θ. In this regime genes get repeatedly exchanged with the rest of the population and $\langle\Delta(t)\rangle$ remains nearly constant over long periods of time (see e.g. $\delta_{\mathrm{mut}}/\delta_{\mathrm{TE}} = 2 \times 10^{-3}$ in Figure 4 and Figure 5). 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.

Do the conclusions about the transition between divergent and metastable dynamics depend on the particular choice of $\delta_{\mathrm{TE}} = 1\%$? In the appendix Figure. A2, we show that in fact the transition is independent of δ_{TE} and is fully determined by the two evolutionary non-dimensional parameters $\theta/\delta_{\mathrm{TE}}$ and $\delta_{\mathrm{mut}}/\delta_{\mathrm{TE}}$ identified in this study.

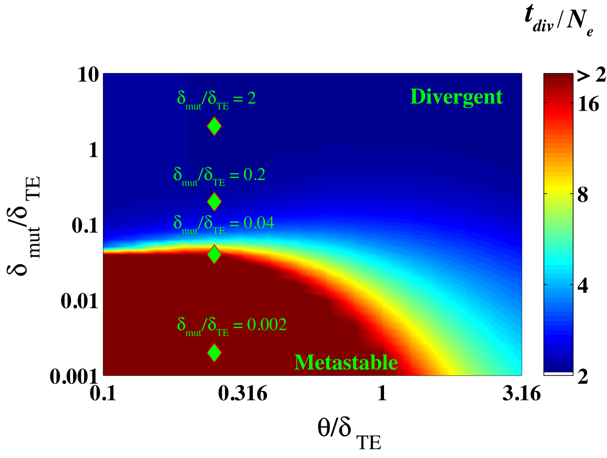

Figure 5 Quantifying metastability in genome evolution. The number of generations t_{div} (in units of the population size N_e) required for a pair of genomes to diverge well beyond the average intra-population diversity (see main text). We calculate the time it takes for the ensemble average $\langle\Delta(t)\rangle$ of the genome-wide average divergence to reach 2θ as a function of $\theta/\delta_{\mathrm{TE}}$ and $\delta_{\mathrm{mut}}/\delta_{\mathrm{TE}}$. We used $\delta_{\mathrm{TE}} = 1\%$, $\mu = 2 \times 10^{-6}$ per base pair generation. In our simulations we varied ρ and θ to scan the $(\theta/\delta_{\mathrm{TE}}, \delta_{\mathrm{mut}}/\delta_{\mathrm{TE}})$ space. The green diamonds represent four populations shown in Figure 4 and Figure 6 (see below).

C. Population structure: the distribution of pairwise divergences of genomes within a population

Can we understand the phylogenetic structure of the entire population by studying the evolutionary dynamics of just a single pair of strains?

Given sufficient amount of time every pair of genomes in our model would diverge indefinitely (see Figure 4). 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, $\overline{p_{\mathrm{c}}(t)} \sim e^{-t/N_e}$ (here and below we use the bar to denote averaging over multiple realizations of the coalescent process, or long-time average over population dynamics) (Kingman 1982; Higgs and Derrida 1992; Serva 2005). Thus, 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

$$\begin{aligned} \pi(\Delta) &= \int_0^\infty p_{\mathrm{c}}(t) \times p(\Delta|t)dt \text{ and} \\ \overline{\pi(\Delta)} &= \int_0^\infty \overline{p_{\mathrm{c}}(t)} \times p(\Delta|t)dt \\ &= \frac{1}{N_e}\int_0^\infty e^{-t/N_e} \times p(\Delta|t)dt \qquad (4) \end{aligned}$$

In Eq. 4, $p_{\mathrm{c}}(t)$ is the probability that a pair of strains in a population snapshot 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

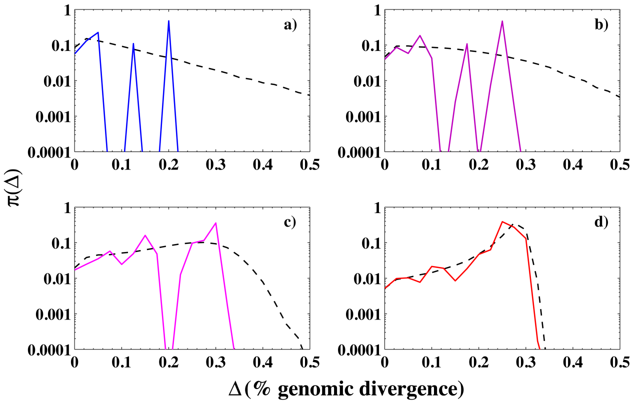

Figure 6 Distribution of genome-wide divergences in a population. Distribution of all pairwise genome-wide divergences δ_{ij} in a co-evolving population for decreasing values of $\delta_{\mathrm{mut}}/\delta_{\mathrm{TE}}$: 2 in a), 0.2 in b), 0.04 in c) and 0.002 in d) In all 4 panels, dashed black 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 Figure 4 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 $p_{\mathrm{c}}(t)$ correspondingly (see text for details).

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. Notably, even for large populations these two distributions could be significantly different from each other. Indeed, $p_{\mathrm{c}}(t)$ in any given population is extremely noisy due to multiple peaks from clonal subpopulations and does not resemble its smooth long-time average profile $\overline{p_{\mathrm{c}}(t)} \sim e^{-t/N_e}$ (Higgs and Derrida 1992; Serva 2005). In panels a) to d) of Figure 6, we show $\pi(\Delta)$ for the four cases shown in Figure 4 (also marked by green diamonds in Figure 5). We fixed the population size to $N_e = 500$. We changed $\delta_{\mathrm{mut}}/\delta_{\mathrm{TE}}$ by changing the recombination rate ρ. The solid lines represent a time snapshot obtained by numerically sampling $p_{\mathrm{c}}(t)$ 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 Figure 5 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, because of a low recombination frequency ρ. In this regime, 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 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 diminishes as well. In this limit, all strains in the population exchange genetic material with each other. Consequently, the population becomes genetically cohesive and temporally stable.

4. Comparison between three models of recombination

So far, we presented results from a simplified model of recombination (model 1, see Figure 1). Employing this model allowed us to develop a mathematical formalism to describe evolutionary dynamics of a pair of bacterial genomes in a co-evolving population. It also allowed us to investigate how genomes diversify across a range of evolutionary parameters in a computationally efficient manner. However, in real bacteria, transfer events have variable lengths and partially overlap with each other (Milkman 1997; Falush *et al.* 2001; Vetsigian and Goldenfeld 2005; Dixit *et al.* 2015).

Here, we systematically study the similarities and differences between the three progressively more realistic models described in section COMPUTATIONAL MODELS and (illustrated in Figure 1 panel b). In order to directly compare results across different types of simulations, we ran each of the three simulations for the four parameter sets used in Figure 4. See appendix for details of the simulations.

The metastability/divergent transition (see Figure 5 above) is based on the dynamics of the ensemble average $\langle\Delta(t)\rangle$. We studied how $\langle\Delta(t)\rangle$ depends on the nature of recombination with an explicit simulation of $N_e = 250$ co-evolving strains each with $L_g = 10^6$ base pairs. Panel a of Figure 7 shows the time evolution of the ensemble average $\langle\Delta(t)\rangle$ estimated from the explicit simulations. The three colors represent three different models of recombination. Notably, $\langle\Delta(t)\rangle$ is insensitive to whether recombination tracks are of variable length or overlapping with each other. Since metastability depends on $\langle\Delta(t)\rangle$, the conclusions about metastability obtained using recombination model (1) can be generalized to more realistic models (2) and (3).

Can the effects of allowing overlapping recombination tracks be seen in population structure? To investigate this, we looked at the stochastic fluctuations in $\Delta(t)$ around $\langle\Delta(t)\rangle$. Intuitively, overlapping recombination events are expected to homogenize highly divergent genetic fragments in the population. As a result, we expect smaller within-population variation i.e. smaller fluctuations in $\Delta(t)$ around $\langle\Delta(t)\rangle$. We tested this by studying the expected distribution $\bar{\pi}(\Delta)$ of pairwise genome-wide divergences within a population (note the above discussion of difference between average $\bar{\pi}(\Delta)$ and the distribution $\pi(\Delta)$ within a sample population) for the three models of recombination.

We only consider the case where $\delta_{\mathrm{mut}}/\delta_{\mathrm{TE}} = 0.002$. As seen in Figure 4 and panel a) of Figure 7, in the metastable state the divergence $\Delta(t)$ virtually does not increase as a function of t at long times (the rate of increase is extremely slow). Thus, the variance in $\bar{\pi}(\Delta)$ largely represents the variance in $\Delta(t)$ around its ensemble average $\langle\Delta(t)\rangle$. In panel b) of Figure 7, we show $\bar{\pi}(\Delta)$ for the three different models of recombination. Indeed, the variance in $\bar{\pi}(\Delta)$ is much smaller when overlapping recombination events are allowed (models (2) and (3) compared to model (1)). The effect of varying the length of recombined segments appears to be minimal.

5. Application to real-life bacterial species

Where are real-life bacterial species located on the divergent-metastable diagram? Instead of $\delta_{\mathrm{mut}}/\delta_{\mathrm{TE}}$, population genetic

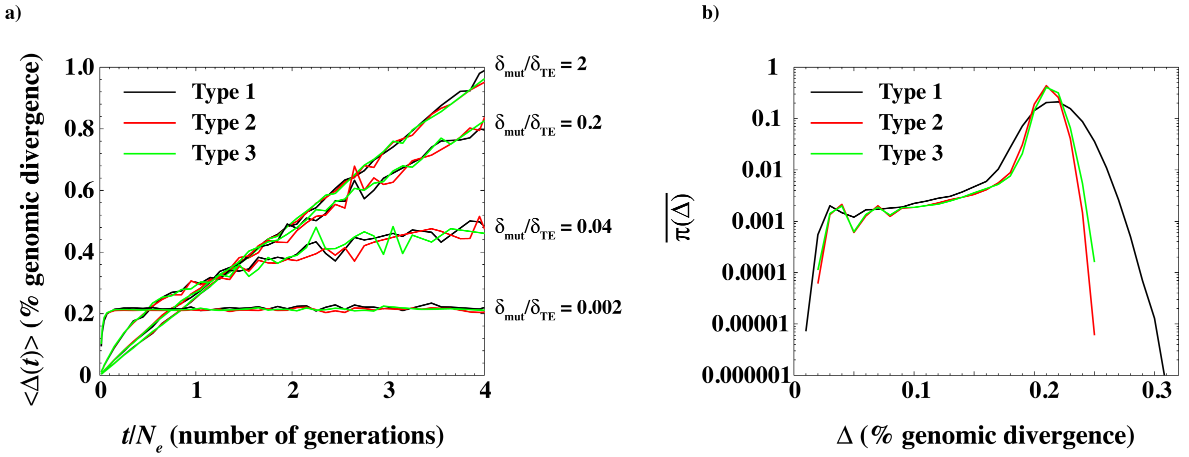

Figure 7 Comparison of different models of recombination. a) The ensemble average $\langle\Delta(t)\rangle$ of pairwise genome-wide divergence $\Delta(t)$ as a function of the pairwise coalescent time t in explicit simulations. Model (1) simulations have non-overlapping transfers of segments of length is 5000 bp. Model (2) simulations have transfers of overlapping 5000 bp segments. Model (3) simulations have overlapping transfer of segnebts if average length 5000 bp. The value of $\delta_{\mathrm{mut}}/\delta_{\mathrm{TE}}$ are on the right side. b) The ensemble average distribution of genome-wide divergence between pairs of strains $\bar{\pi}(\Delta)$ for the three models recombination shown in panel a of Figure 1 when $\delta_{\mathrm{mut}}/\delta_{\mathrm{TE}} = 0.002$.

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 (Guttman and Dykhuizen 1994; Vos and Didelot 2009; Dixit *et al.* 2015). In our framework, r/m is defined as $r/m = \rho_{\mathrm{succ}}/\mu \times l_{\mathrm{tr}} \times \delta_{\mathrm{tr}}$ where $\rho_{\mathrm{succ}} < \rho$ is the rate of successful recombination events and δ_{tr} is the average divergence in transferred regions. Both ρ_{succ} and δ_{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 $\delta_{\mathrm{mut}}/\delta_{\mathrm{TE}}$ used in our previous plots.

In Figure 8, we re-plotted the "phase diagram" shown in Figure 5 in terms of $\theta/\delta_{\mathrm{TE}}$ and r/m and approximately placed several real-life bacterial species on it. To this end we estimated θ from the MLST data (Jolley and Maiden 2010) (see appendix for details) and used r/m values that were determined previously by Vos and Didelot (Vos and Didelot 2009). As a first approximation, we assumed that the transfer efficiency δ_{TE} is the same for all species considered and is given by $\delta_{\mathrm{TE}} \sim 2.26\%$ used in Ref. (Fraser *et al.* 2007). However, as mentioned above, the transfer efficiency δ_{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* (Falush *et al.* 2001) or different combinations of multiple RM systems reported in Ref. (Oliveira *et al.* 2016). We note that *Helicobacter pylori* appears divergent even with minimal barriers to recombination probably because of its ecologically structured population that is dependent on human migration patterns (Thorell *et al.* 2017). 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. Dixit *et al.* (2015).

Figure 8 confirms that both r/m and $\theta/\delta_{\mathrm{TE}}$ are important evolutionary parameters and suggests that each of them alone cannot fully classify a species as either divergent or metastable. Notably, there is a sharp transition between the divergent and the metastable phases implying that a small change in r/m or $\theta/\delta_{\mathrm{TE}}$ can change the evolutionary fate of the species. And finally, one can 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 Figure 5. For example, if the effective population size remains constant, a change in mutation rate changes both $\delta_{\mathrm{mut}}/\delta_{\mathrm{TE}}$ 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 δ_{TE} or the rate of recombination (Vulić *et al.* 1997; Oliveira *et al.* 2016) thus changing the placement of the species on the phase diagram shown in Figure 8.

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 (Krause and Whitaker 2015). 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 (Didelot *et al.* 2012; Dixit *et al.* 2015). Ecological niche separation and/or selective advantages are usually implicated as initiators of such putative speciation events (Polz *et al.* 2013). In our previous analysis of 32 fully sequenced *E. coli* strains, we estimated $\theta/\delta_{\mathrm{TE}} > 3$ and $r/m \sim 8-10$ (Dixit *et al.* 2015) implying that *E. coli* resides deeply in the divergent regime in Figure 8. Thus, based on the analysis presented above one expects

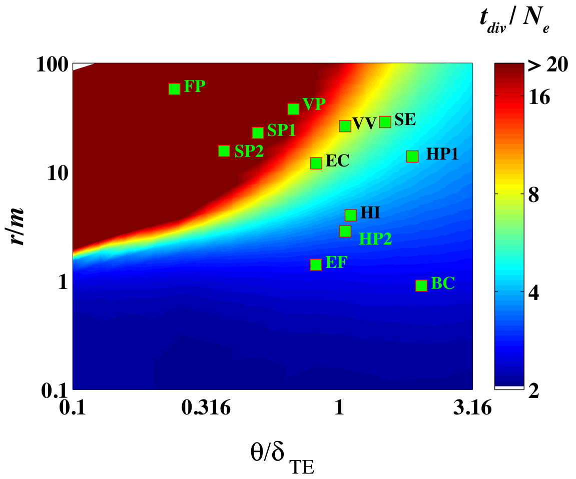

Figure 8 Classifying real bacteria as metastable or divergent. Approximate position of several real-life bacterial spaces on the metastable-divergent phase diagram (see text for details). Abbreviations of species names are as follows: FP: *Flavobacterium psychrophilum*, VP: *Vibrio parahaemolyticus*, SE: *Salmonella enterica*, VV: *Vibrio vulnificus*, SP1: *Streptococcus pneumoniae*, SP2: *Streptococcus pyogenes*, HP1: *Helicobacter pylori*, HP2: *Haemophilus parasuis*, HI: *Haemophilus influenzae*, BC: *Bacillus cereus*, EF: *Enterococcus faecium*, and EC: *Escherichia coli*.

E. coli strains to spontaneously form transient sexually-isolated sub-populations even in the absence of selective pressures or ecological niche separation.

6. Extending the framework to incorporate selection and other factors modulating recombination

Throughout this study we used two assumptions that allowed efficient mathematical analysis: i) exponentially decreasing probability of successful integration of foreign DNA, $p_{\text{success}} \sim e^{-\delta/\delta_{\text{TE}}}$ and ii) exponentially distributed pairwise coalescent time distribution of a neutrally evolving well-mixed population. Here we discuss how to relax these assumptions within our framework.

(i) A wide variety of barriers to foreign DNA entry exist in bacteria (Thomas and Nielsen 2005). For example, bacteria may have multiple RM systems that either act in combination or are turned on and off randomly (Oliveira *et al.* 2016). Moreover, rare non-homologous/illegitimate recombination events can transfer highly diverged segments between genomes (Thomas and Nielsen 2005) potentially leading to homogenization of the population. Such events can be captured by a weaker-than-exponential dependence of the probability of successful integration on local genetic divergence (see Appendix for a calculation with non-exponential dependence of the probability of successful integration p_{success} on the local sequence divergence). One can incorporate these variations within our framework by appropriately modifying p_{success} in the framework.

(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 niches 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 (Takahata 1991; Wakeley 2004). Within our framework, niche-related effects can be incorporated by accounting for pairwise coalescent times of niche-structured populations (Takahata 1991; Wakeley 2004) 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 are thought to have insignificant fitness effect, the evolution of bacterial species may be driven by rare advantageous mutations (Majewski and Cohan 1999). Recombination is thought to be essential for bacterial evolution in order to minimize the fitness loss due to Muller's ratchet (Takeuchi *et al.* 2014) and to minimize the impact of clonal interference (Cooper 2007). Thus, it is likely that both recombination frequency and transfer efficiency are under selection (Takeuchi *et al.* 2014; Lobkovsky *et al.* 2016; Iranzo *et al.* 2016). 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 (Neher and Hallatschek 2013) 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 (Neher and Hallatschek 2013). 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. Dixit *et al.* (2015) 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 (Neher and Hallatschek 2013), and for our main results to remain qualitatively unchanged.

7. Conclusion

While recombination is now recognized as an important contributor to patterns of genome diversity in many bacterial species(Guttman and Dykhuizen 1994; Milkman 1997; Falush *et al.* 2001; Thomas and Nielsen 2005; Touchon *et al.* 2009; Vos and Didelot 2009; Dixit *et al.* 2015), its effect on population structure and stability is still heavily debated (Fraser *et al.* 2007; Wiedenbeck and Cohan 2011; Doolittle 2012; Polz *et al.* 2013; Shapiro *et al.* 2016). In this work, we explored three models of gene transfers in bacteria to study how the competition between mutations and recombinations affects genome evolution. Analysis of each of the three models showed that recombination-driven bacterial genome evolution can be understood as a balance between two competing processes. We identified the two dimensionless parameters $\theta/\delta_{\text{TE}}$ and $\delta_{\text{mut}}/\delta_{\text{TE}}$

that dictate this balance and result in two qualitatively different regimes in bacterial evolution, separated by a sharp transition.

The two competitions give rise to two regimes of genome evolution. In the divergent regime, recombination is insufficient to homogenize 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. 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. Notably, real bacterial species appear to belong to both of these regimes as well as in the cross-over region separating them from each other.

Acknowledgments: We 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. We also thank the three reviewers and the editor for their detailed reading and valuable comments.

Literature Cited

Cooper, T. F., 2007 Recombination speeds adaptation by reducing competition between beneficial mutations in populations of escherichia coli. PLoS Biol **5**: e225.

Didelot, X., G. Méric, D. Falush, and A. E. Darling, 2012 Impact of homologous and non-homologous recombination in the genomic evolution of escherichia coli. BMC genomics **13**: 1.

Dixit, P. D., T. Y. Pang, F. W. Studier, and S. Maslov, 2015 Recombinant transfer in the basic genome of escherichia coli. Proceedings of the National Academy of Sciences **112**: 9070–9075.

Doolittle, W. F., 2012 Population genomics: how bacterial species form and why they don't exist. Current Biology **22**: R451–R453.

Doroghazi, J. R., and D. H. Buckley, 2011 A model for the effect of homologous recombination on microbial diversification. Genome biology and evolution **3**: 1349.

Falush, D., C. Kraft, N. S. Taylor, P. Correa, J. G. Fox, *et al.*, 2001 Recombination and mutation during long-term gastric colonization by helicobacter pylori: estimates of clock rates, recombination size, and minimal age. Proceedings of the National Academy of Sciences **98**: 15056–15061.

Fraser, C., W. P. Hanage, and B. G. Spratt, 2007 Recombination and the nature of bacterial speciation. Science **315**: 476–480.

Gillespie, J. H., 2010 *Population genetics: a concise guide*. JHU Press.

Guttman, D. S., and D. E. Dykhuizen, 1994 Clonal divergence in escherichia coli as a result of recombination, not mutation. Science **266**: 1380.

Higgs, P. G., and B. Derrida, 1992 Genetic distance and species formation in evolving populations. Journal of molecular evolution **35**: 454–465.

Hogg, J. S., F. Z. Hu, B. Janto, R. Boissy, J. Hayes, *et al.*, 2007 Characterization and modeling of the haemophilus influenzae core and supragenomes based on the complete genomic sequences of rd and 12 clinical nontypeable strains. Genome Biol **8**: R103.

Iranzo, J., P. Puigbo, A. E. Lobkovsky, Y. I. Wolf, and E. V. Koonin, 2016 Inevitability of genetic parasites. Genome Biology and Evolution : evw193.

Jolley, K. A., and M. C. Maiden, 2010 Bigsdb: Scalable analysis of bacterial genome variation at the population level. BMC bioinformatics **11**: 595.

Kingman, J. F. C., 1982 The coalescent. Stochastic processes and their applications **13**: 235–248.

Krause, D. J., and R. J. Whitaker, 2015 Inferring speciation processes from patterns of natural variation in microbial genomes. Systematic biology **64**: 926–935.

Lapierre, P., and J. P. Gogarten, 2009 Estimating the size of the bacterial pan-genome. Trends in genetics **25**: 107–110.

Lobkovsky, A. E., Y. I. Wolf, and E. V. Koonin, 2016 Evolvability of an optimal recombination rate. Genome biology and evolution **8**: 70–77.

Majewski, J., 2001 Sexual isolation in bacteria. FEMS microbiology letters **199**: 161–169.

Majewski, J., and F. M. Cohan, 1999 Adapt globally, act locally: the effect of selective sweeps on bacterial sequence diversity. Genetics **152**: 1459–1474.

Marttinen, P., N. J. Croucher, M. U. Gutmann, J. Corander, and W. P. Hanage, 2015 Recombination produces coherent bacterial species clusters in both core and accessory genomes. Microbial Genomics **1**.

Medini, D., C. Donati, H. Tettelin, V. Masignani, and R. Rappuoli, 2005 The microbial pan-genome. Current opinion in genetics & development **15**: 589–594.

Milkman, R., 1997 Recombination and population structure in escherichia coli. Genetics **146**: 745.

Neher, R. A., and O. Hallatschek, 2013 Genealogies of rapidly adapting populations. Proceedings of the National Academy of Sciences **110**: 437–442.

Ochman, H., S. Elwyn, and N. A. Moran, 1999 Calibrating bacterial evolution. Proceedings of the National Academy of Sciences **96**: 12638–12643.

Oliveira, P. H., M. Touchon, and E. P. Rocha, 2016 Regulation of genetic flux between bacteria by restriction–modification systems. Proceedings of the National Academy of Sciences **113**: 5658–5663.

Polz, M. F., E. J. Alm, and W. P. Hanage, 2013 Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends in Genetics **29**: 170–175.

Serva, M., 2005 On the genealogy of populations: trees, branches and offspring. Journal of Statistical Mechanics: Theory and Experiment **2005**: P07011.

Shapiro, B. J., J.-B. Leducq, and J. Mallet, 2016 What is speciation? PLoS Genet **12**: e1005860.

Studier, F. W., P. Daegelen, R. E. Lenski, S. Maslov, and J. F. Kim, 2009 Understanding the differences between genome sequences of escherichia coli b strains rel606 and bl21 (de3) and comparison of the e. coli b and k-12 genomes. Journal of molecular biology **394**: 653–680.

Takahata, N., 1991 Genealogy of neutral genes and spreading of selected mutations in a geographically structured population. Genetics **129**: 585–595.

Takeuchi, N., K. Kaneko, and E. V. Koonin, 2014 Horizontal gene transfer can rescue prokaryotes from muller's ratchet: benefit of dna from dead cells and population subdivision. G3: Genes| Genomes| Genetics **4**: 325–339.

Tenaillon, O., D. Skurnik, B. Picard, and E. Denamur, 2010 The population genetics of commensal escherichia coli.

Nature Reviews Microbiology **8**: 207–217.

Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, *et al.*, 2005 Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: implications for the microbial “pan-genome”. Proceedings of the National Academy of Sciences of the United States of America **102**: 13950–13955.

Thomas, C. M., and K. M. Nielsen, 2005 Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nature reviews microbiology **3**: 711–721.

Thorell, K., K. Yahara, E. Berthenet, D. J. Lawson, J. Mikhail, *et al.*, 2017 Rapid evolution of distinct helicobacter pylori subpopulations in the americas. PLoS genetics **13**: e1006546.

Touchon, M., C. Hoede, O. Tenaillon, V. Barbe, S. Baeriswyl, *et al.*, 2009 Organised genome dynamics in the escherichia coli species results in highly diverse adaptive paths. PLoS genet **5**: e1000344.

Vetsigian, K., and N. Goldenfeld, 2005 Global divergence of microbial genome sequences mediated by propagating fronts. Proceedings of the National Academy of Sciences of the United States of America **102**: 7332–7337.

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

Vulić, M., F. Dionisio, F. Taddei, and M. Radman, 1997 Molecular keys to speciation: Dna polymorphism and the control of genetic exchange in enterobacteria. Proceedings of the National Academy of Sciences **94**: 9763–9767.

Wakeley, J., 2004 Recent trends in population genetics: More data! more math! simple models? Journal of Heredity **95**: 397–405.

Wiedenbeck, J., and F. M. Cohan, 2011 Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. FEMS microbiology reviews **35**: 957–976.

Wielgoss, S., J. E. Barrick, O. Tenaillon, S. Cruveiller, B. Chane-Woon-Ming, *et al.*, 2011 Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with escherichia coli. G3: Genes, Genomes, Genetics **1**: 183–186.

A1. Appendix

A. $\langle \Delta(t) \rangle$ from computer simulations

To compare the three models of recombination, we performed three types of explicit simulations of a Fisher-Wright population of $N_e = 250$ co-evolving strains. The three simulations had different modes of gene transfers as indicated in panel b of Figure 1. Each strain had $L_g = 10^6$ base pairs. Each base pair was represented either by a 0 (wild type) or 1 (mutated). The mutation rate was fixed at $\mu = 5 \times 10^{-6}$ per base pair per generation. We varied the recombination rate $\rho = 2.5 \times 10^{-8}, 2.5 \times 10^{-7}, 1.25 \times 10^{-6}$, and 2.5×10^{-5} per base pair per generation. θ was fixed at $\theta = 0.25\%$ and δ_{TE} was fixed at $\delta_{\text{TE}} = 1\%$. These parameters are identical to the ones used in Figure 4 of the main text. We note that given the low population diversity ($\theta = 0.25\%$), we can safely neglect back mutations. Note that in all three simulations, on an average, a total of 5 kilobase pairs of genome was transferred in a successful transfer event thereby allowing us to directly compare the three simulations.

We strated the simulations with N_e identical genomes. We ran a Fisher-Wright simulation for $5000 = 20 \times N_e$ generations to ensure that the population reached a steady state. In each generation, children chose their parents randomly. This ensured that the population size remained constant over time. Mutation and recombination events were attempted according to the corresponding rates. Note that it is non-trivial to keep track of the divergence between individual pairs over time since one or both of the strains in the pair may either be stochastically eliminated. To study the time evolution of the ensemble average $\langle \Delta(t) \rangle$ of the divergence, at the end of the simulation, we collected the pairwise coalescent times t between all pairs of strains as well as $\Delta(t)$, the genomic divergences between them. Note that due to the stochastic nature of mutations and recombination events, $\Delta(t)$ is a random variable. We estimated the ensemble average $\langle \Delta(t) \rangle$ by binning the pairwise coalescent times in intervals of $dt = 25$ generations (1/10th of the population size) and taking an average over all $\Delta(t)$ in each bin. The ensemble average thus estimated represents the average over multiple realizations of the coalescent process. Mathematically, the ensemble average is given by

$$\langle \Delta(t) \rangle = \int \Delta(t) p(\Delta | t) d\Delta \tag{A1}$$

Here, $p(\Delta|t)$ is the probability that the genomes of two strains whose most recent common ancestor was t generations ago have diverged by Δ. We note that the variance in $\Delta(t)$ is expected to be small since it is an average over a large number of genes. These results are plotted in Figure 7.

B. Behavior of $\langle \Delta(t) \rangle$ in the long time limit

In Figure A1 we show how $\langle \Delta(t) \rangle$ increases with t over a longer range of times. We note that it is exponentially rarer to find a pair of strains in a population that have diverged beyond $t > N_e$ generations where N_e is the population size.

C. Transition between metastable and divergent dynamics doesn't depend on the choice of δ_{TE}

In Figure 5 in the main text, we showed the transition between metastable and divergent evolutionary dynamics. There, we fixed $\delta_{\text{TE}} = 1\%$ and varied θ and ρ to scan the space of non-dimensional parameters $\theta/\delta_{\text{TE}}$ and $\delta_{\text{mut}}/\delta_{\text{TE}}$. However, our results do not depend on this particular value of δ_{TE}. To show this, we recalculated Figure 5 by randomly sampling θ (between 0.5% to 3%), δ_{TE} (between 0.5% to 5%), and ρ (between 2×10^{-7} and 2×10^{-5} per base pair per generation) while keeping the mutation rate constant at $\mu = 2 \times 10^{-5}$ per base pair per generation. In Figure A2 below, we plot the time t_{div} required for the ensemble average genome wide divergence $\langle \Delta(t) \rangle$ to reach an atypical value of 2θ. From Figure A2, it is clear that the time taken to reach 2θ indeed is determined by the two dimensionless constants $\theta/\delta_{\text{TE}}$ and $\delta_{\text{mut}}/\delta_{\text{TE}}$ and not by the particular choice of the value of δ_{TE}.

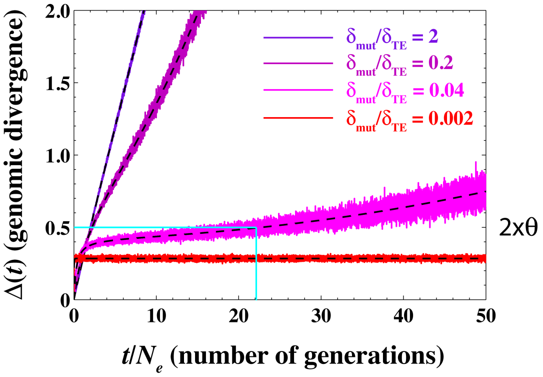

Figure A1 Genome-wide divergence $\Delta(t)$ as a function of time at $\theta/\delta_{\text{TE}} = 0.25$. We have used $\delta_{\text{TE}} = 1\%$, $\theta = 0.25\%$, $\mu = 2 \times 10^{-6}$ per base pair per generation and $\rho = 2 \times 10^{-8}, 2 \times 10^{-7}, 10^{-5}$, and 2×10^{-5} per base pair per generation corresponding to $\delta_{\text{mut}}/\delta_{\text{TE}} = 2, 0.2, 0.04$ and 2×10^{-3} respectively. The dashed black lines represent the ensemble average $\langle \Delta(t) \rangle$. The cyan lines show the time it takes for the ensemble-averaged genomic divergence $\langle \Delta(t) \rangle$ to reach 2θ when $\delta_{\text{mut}}/\delta_{\text{TE}} = 0.04$ (pink line).

D. Estimating r/m from model parameters

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

$$r/m = \rho_{\text{succ}}/\mu \times l_{\text{tr}} \times \delta_{\text{tr}} \tag{A2}$$

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

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

where p_{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_{succ} can be computed from Eq. 3 by integrating over all possible scenarios of successful recombinations. We have

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

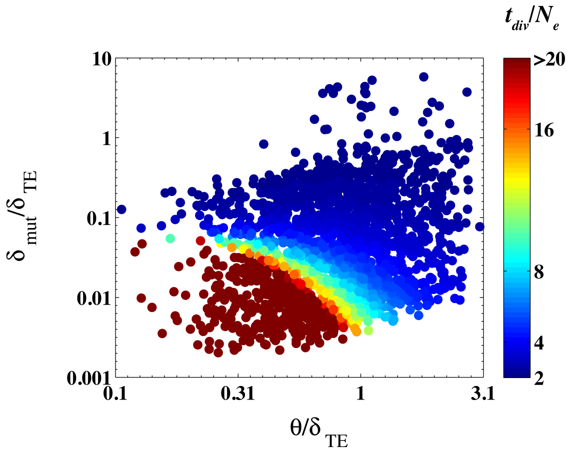

Figure A2 The number of generations t_{div} (in units of the population size N_e) required for a pair of genomes to diverge well beyond the average intra-population diversity. We calculate the time it takes for the ensemble average of the genome-wide average divergence to reach 2θ as a function of $\theta/\delta_{\text{TE}}$ and $\delta_{\text{mut}}/\delta_{\text{TE}}$. We randomly sample θ (between 0.5% to 3%), δ_{TE} (between 0.5% to 5%), and ρ (between 2×10^{-7} and 2×10^{-5} per base pair per generation) while keeping the mutation rate constant at $\mu = 2\times10^{-5}$ per base pair generation.

where $\delta^* = \delta/\delta_{\text{TE}}$ and $\theta^* = \theta/\delta_{\text{TE}}$ 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 (Didelot *et al.* 2012; Dixit *et al.* 2015). Here, we limit the time-integration in Eq. A3 to times $t < \min(N_e = \theta/2\mu, \delta_{\text{TE}}/2\mu)$.

Second, we need to compute the average divergence in transferred segments, δ_{tr}. We have

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

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

E. Computing θ from MLST data

Except for *E. coli* where we used our previous analysis (Dixit *et al.* 2015) (we used $\theta/\delta_{\text{TE}} \sim 3$ and $r/m = 12$), we downloaded MLST sequences of multiple organisms from the MLST database (Jolley and Maiden 2010). For each of the 7 genes present in the MLST database, we performed a pairwise alignment between strains. For a given pair of strains, we evaluated the % nucleotide difference in each gene and estimated the average q over these 7 pairwise differences. The θ for the species was estimated as an average of q over all pairs of strains.

F. Non-exponential dependence of p_{success} on local sequence divergence

In the main text, we showed that when p_{sucess} decays exponentially with the local divergence, the time evolution of local divergence $\delta(t)$ shows metastability. When the recombination rate is low, a few recombination events take place that change $\delta(t)$ to typical values in the population before the local region eventually escapes the integration barrier, leading to a linear increase in $\delta(t)$ (see Figure 3). When the recombination rate is high, the number of recombination events before the eventual escape from the integration barrier increases drastically leading to metastable behavior.

Here, we suggest that weaker-than-exponential dependence of p_{success} can lead to a time evolution of local divergence $\delta(t)$ that never escapes the integration barrier, leading to a genetically homogeneous population independent of the recombination rate ρ.

While it is difficult to carry out analytical calculations for a finite θ and δ_{TE}, following Doroghazi and Buckley (Doroghazi and Buckley 2011), we consider the limit $\theta \to 0$ when μ and ρ are finite. The time evolution of $\delta(t)$ in the limit $\theta \to 0$ when p_{success} decays exponentially with divergence is given by (see Eq. 3)

$$\begin{aligned} p(\delta \to \delta+1) &= 2\mu \text{ and} \\ p(\delta \to 0) &= \rho e^{-\frac{\delta}{\delta_{\text{TE}}}} \end{aligned} \tag{A6}$$

In Eq. A6, $\delta(t)$ is the number of SNPs (as opposed to SNP density used in the main text). As was shown in the main text, the evolution of $\delta(t)$ described by Eq. A6 is a random walk that repeatedly resets to zero before eventually escaping to $\delta \to \infty$. The number of resetting events depends on $\delta_{\text{mut}}/\delta_{\text{TE}}$ as defined in the main text (see low $\theta/\delta_{\text{TE}}$ values in Figure 5).

A generalization to non-exponential dependence of the success probability is straightforward,

$$\begin{aligned} p(\delta \to \delta+1) &= 2\mu \text{ and} \\ p(\delta \to 0) &= \rho f(\delta) \end{aligned} \tag{A7}$$

where $1 \le f(\delta) \ge 0$ is the probability of successful integration. How weak should the integration barrier $f(\delta)$ be so that the time evolution described by Eq. A7 can never escape the pull of recombination? In other words, what are the conditions on $f(\delta)$ that ensure that the time evolution of local divergence described by Eq. A7 results in a random walk that resets to zero infinitely many times?

If the random walk resets infinitely many times, it has a well defined stationary distribution as $t \to \infty$. Note that the random walk described by an exponentially decaying p_{success} does not have a well defined stationary distribution since as $t \to \infty$, $\delta(t) \to \infty$ regardless of the rate of recombination and the transfer efficiency. Let us assume that $f(\delta)$ is such that there exists a well-defined stationary distribution. We define p_i as the probability that $\delta = i$ in the stationary state. We can write balance equations in the stationary state

$$2\mu \times p_0 = \rho \times \sum_{i=1}^{\infty} p_i f(i) \tag{A8}$$

$$2\mu \times p_i + \rho \times p_i f(i) = 2\mu \times p_{i-1} \; \forall\, i > 0 \tag{A9}$$

Rearranging

$$p_i = p_{i-1}\frac{1}{1+\frac{\rho}{2\mu}f(i)} = p_0\prod_{j=1}^{j=i}\frac{1}{1+\frac{\rho}{2\mu}f(j)} \; if\; i > 0 \tag{A10}$$

Since $p_0 \neq 0$, from Eq. A9 and Eq. A10 we have for an arbitrary

$f(\delta)$ (denoting $\rho/2\mu = \tau$)

$$\begin{aligned} s[\tau, f] &= \tau \sum_{i=1}^{\infty} \left(f(i) \prod_{j=1}^{j=i} \frac{1}{1+\tau f(j)} \right) = 1 \\ \Rightarrow m[\tau, f] &= 1 - s[\tau, f] = \prod_{i} \frac{1}{1+\tau f(i)} = 0 \qquad \text{(A11)} \end{aligned}$$

Thus, as long as the *functional* $s[\tau, f]$ in Eq. A11 is equal to 1 (or $m[\tau, f] = 0$), the walk remains localized. Eq. A11 is a surprisingly simple result and is valid for any $0 \leq f(\delta) \leq 1$.

Let us consider a specific case where $f(\delta) = \delta^{-\nu}$. A power-law dependence in p_{success} is weaker than the exponential decay used in the main text, potentially allowing transfers between distant bacteria. Let us examine the self-consistency condition. We have

$$m(\tau, \nu) = 1 - s(\tau, \nu) = \prod_{i=1}^{\infty} \frac{1}{1+\tau i^{-\nu}} \qquad \text{(A12)}$$

Taking logarithms and using the Abel-Plana formula

$$\begin{aligned} \log m(\tau, \nu) &\sim -\int_{1}^{\infty} \log(1+\tau x^{-\nu}) dx \\ &= {}_2F_1(1, \frac{\nu-1}{\nu}; 2 - \frac{1}{\nu}, -\tau) \times \frac{\nu\tau}{\nu-1} - \log(1+\tau) \end{aligned} \qquad \text{(A13)}$$

if $\nu \geq 1$. The integral (and thus the sum) tends to ∞ when $\nu < 1$. Here, ${}_2F_1$ is the hypergeometric function. Thus, when $\nu < 1$, a well defined stationary distribution exists and as long as $\rho > 0$ and $\mu > 0$ regardless of ρ and the population remains genetically cohesive. When $\nu > 1$, we expect behavior similar to the exponential case studied in the main text, viz. a divergent vs metastable transition depending on the competition between forces of recombinations and mutations. We believe that these conclusions will also hold true when θ is finite.