Scaling approaches for modeling the fine-scale diversity of microbes

Benjamin H Good, Stanford University

Overview. Large-scale sequencing efforts have uncovered extensive genetic diversity in natural populations of viruses and bacteria. Understanding the causes and consequences of this fine-scale variation – and how it shifts and changes over time – is crucial for efforts to understand the evolution of antibiotic resistance or the long-term resilience of microbial communities. At present, however, the dynamics of this fine-scale diversity are still very poorly understood. A central challenge is that these population-level processes emerge from a mixture of stochastic evolutionary forces, each of which acts across a broad range of length and time scales. While the underlying ingredients – mutation, selection, recombination, and genetic drift – are each well-understood in isolation, it is extremely difficult to predict how they combine to determine the emergent patterns of diversity within a population. This is particularly true for natural populations of bacteria, where the interactions between natural selection and horizontal gene transfer are known to play an important role. The current project aims to bridge this gap in our understanding by developing new theoretical methods to predict the statistics of microbial diversity under the joint action of selection, recombination, and genetic drift. Our central goal is to develop a new class of scaling approaches to systematically probe the evolutionary forces that operate on different time scales in large microbial datasets, and to use these approaches to quantify the dynamics of horizontal gene transfer in a diverse range of human gut bacteria. We will pursue these objectives through the following aims:

Aim 1: Develop a theoretical framework to predict how selection and recombination influence correlations between mutations across a broad range of length and time scales.

Aim 2: Extend the analytical framework in Aim 1 to account for non-equilibrium demography.

Aim 3: Develop model-independent approaches to directly measure recombination events among closely related strains in a large cohort of human gut bacteria.

Intellectual Merit. This project will utilize concepts of scaling and coarse-graining from statistical physics to address fundamental questions in evolutionary dynamics. Our forward-time framework will provide the first analytical predictions for the distributions of mutation frequencies and their correlations ("linkage disequilibria") under the joint action of selection, recombination, and non-equilibrium demography. These analytical results will help resolve longstanding controversies about the relative contributions of selection and demography in shaping the fine-scale diversity of natural populations, and will enable a variety of new approaches for quantifying the dynamics of bacterial recombination on different time scales. Our empirical applications to human gut bacteria will provide one of the highest resolution views of the rates and lengths of successful recombination events to date, and will allow us to identify universal trends that are shared across many closely related species. These analyses will help address fundamental questions about the maintenance of bacterial diversity across a broad range of length and time scales, and will allow us to better exploit the unique dimensions of modern genomic datasets.

Broader impacts. Fine-scale microbial diversity is the fundamental fuel for evolution, and underlies some of the greatest challenges facing humanity today. The novel theoretical methods and data analysis techniques developed in this project will therefore serve as a valuable community resource for understanding the evolution of other microbial systems (e.g. bacterial pathogens or certain cancers), as well as higher organisms (e.g. humans) where large numbers of sequenced genomes are currently available. In addition, we anticipate that the specific rates and timescales of horizontal gene transfer we will infer in our empirical applications to human gut bacteria will have broader relevance for efforts to understand and manipulate these medically important ecosystems. This interdisciplinary nature of this project, which integrates methods from statistical physics, population genetics, microbial ecology, and genomics, will enable diverse educational, mentorship, and dissemination activities that magnify the broader impact of the project. These activities include interdisciplinary graduate, undergraduate, and postdoctoral training, a new course on quantitative evolutionary modeling for physicists and engineers, and a local conference on population genetics and genomics.

Scaling approaches for modeling the fine-scale diversity of microbes

1. Introduction and Summary of Project Objectives

The commoditization of DNA sequencing has fueled tremendous growth in the number of sequenced genomes over the last decade. The numbers are particularly striking for microbial organisms: millions of SARS-CoV2 genomes have been sequenced during the COVID-19 pandemic [1], and hundreds of thousands of human gut bacteria have been sequenced from different fecal samples [2]. The fine-scale variation within these large cohorts provides unprecedented opportunities to quantify the evolutionary forces that operate within these rapidly evolving populations. These strain-level dynamics have important practical consequences, from the spread of antibiotic resistance mutations [3–5], to the success of fecal microbiome transplants and other personalized therapies [6, 7].

Yet despite their central importance, the evolutionary processes that shape microbial diversity remain poorly characterized. Even basic questions about the relevant parameter regimes are still hotly debated: How important are stochastic vs selective effects in generating the fine-scale variation we observe today? When does horizontal gene transfer enable the transition between selection on entire genomes and selection on individual genes or mutations? Our inability to answer these questions – even at an order of magnitude level – stems in part from our limited theoretical understanding of their underlying evolutionary dynamics. The basic rules of evolution are simple: mutation and recombination generate variation, while natural selection and random genetic drift alter the frequencies of the variants. Each of these forces is well understood in isolation [8, 9]. However, it is extremely difficult to predict how they combine in empirically relevant settings where all four forces operate simultaneously. A further challenge is that the forces of natural selection, recombination, and genetic drift each often act across a broad range of length and time scales within a population. This makes it difficult to predict how these competing stochastic processes contribute to the emergent patterns of diversity that we observe at the population level.

Classical approaches from population genetics tend to neglect one or more of these forces [8–10], or else assume that they can be captured by rescaling the remaining parameters [11–17]. Recent work has highlighted the limitations of this approach [17–20], particularly in microbial populations where selection and recombination can both play a major role [21–25]. A second body of work has sought to overcome these difficulties using computer simulations, ranging from numerical integration and computer algebra techniques [26–30] to agent-based simulations and machine learning methods [31–37]. While these computational approaches have greatly increased the range of evolutionary scenarios one can consider, the large number of inputs and outputs can make it difficult to draw connections between competing models [17, 38], or to understand when they can – or can't – explain the observable data [39–41]. This greatly limits our ability to answer the fundamental questions above.

The proposed project aims to bridge this gap in our understanding by developing new theoretical methods to predict how natural selection, recombination, and random genetic drift combine to shape the statistics of microbial diversity. We plan to do this by exploiting an additional degree of freedom inherent in the large sizes of modern genomic datasets. The genetic variants observed within these large cohorts are distributed across a huge range of sample frequencies, which reflect the broad range of time scales over which these mutations accumulated. By examining how the statistics of microbial diversity scale across these different frequency ranges, it should be possible to systematically probe the evolutionary forces that operate across these corresponding time scales. Similar scaling approaches have been enormously useful for understanding the collective behavior of physical systems [42–44], as well as recent extensions to ecology [45] and neuroscience [46]. At present, however, few methods exist for predicting how the statistics of genetic diversity scale across different frequency ranges, or for inverting these scaling laws to make inferences about the underlying evolutionary forces. ***The central goal of this proposal is to extend these scaling approaches to the evolutionary domain, with a specific focus on the dynamics of horizontal gene transfer.*** We will pursue this objective using a combination of mathematical modeling and analysis of genomic data from natural populations of human gut bacteria.

Aim 1: Predict how selection and recombination influence correlations between mutations across a broad range of length and time scales. We will use new theoretical approaches we have recently developed to predict how correlations between pairs of neutral and deleterious mutations depend on their present-day frequencies. We will use these approaches to analyze data from a large collection of

human gut bacteria, to systematically probe how rates of recombination vary across a broad range of length and time scales in a variety of different species.

Aim 2: Extend the analytical framework in Aim 1 to account for non-equilibrium demography. We will leverage new theoretical results we have recently derived for predicting the frequencies of neutral and deleterious mutations under arbitrary time-varying population sizes. We will use these results to identify scaling properties of genetic diversity that cannot be produced by any simple demographic model.

Aim 3. Develop model-independent approaches to directly measure recombination events among closely related strains in large cohorts. We will leverage the local nature of bacterial recombination to develop a new approach for directly resolving recombination events that accumulate between closely related bacterial strains. We will apply our method to a large panel of human gut bacteria in order to systematically quantify how the rates and lengths of successful intraspecific recombination events vary among strains with different degrees of genetic relatedness.

2. Background and Significance

Microbial organisms are incredibly diverse, occupying almost every possible ecological niche [47]. Environmental surveys have identified tens of thousands of different species [48] and more than a million more are thought to remain undiscovered [49]. The widespread deployment of DNA sequencing has also revealed extensive genetic variation within these species as well [23, 50–52]. The phenotypic consequences of this "*fine-scale" variation* are less well understood, but a growing body of evidence suggests that they can have a major impact on the structure and function of different microbial ecosystems. In the human gut microbiota, for example, fine-scale genetic variants are known to influence the activity of certain pharmaceuticals [53], as well as the success of microbiome transplants designed to eliminate resistant infections [6]. Fine-scale genetic variation also plays a role in the spread of antibiotic resistance mutations [3] and the emergence of variants-of-concern in the ongoing COVID-19 pandemic [54]. Understanding the causes and consequences of this genetic variation, and how it shifts and changes over time, is one of the most pressing open questions in the field.

Fine-scale variation must ultimately arise through the basic laws of population genetics [9]. The underlying rules of this process are simple: mutations create genetic variation, and natural selection and random genetic drift alter the frequencies of the variants. In microbes, the horizontal exchange of DNA can also enable recombination between both distantly and closely related strains [55]. The qualitative effects of these forces are well understood in isolation [8, 9]. However, it is often surprisingly difficult to predict how they interact – even in simple and well-defined settings. The basic problem is that too many things are going on at once. Unlike the classical picture, where mutations arise one-by-one, large microbial populations typically harbor many different mutations at the same time. Natural selection, recombination, and genetic drift cannot act on these mutations individually, but only on *combinations of mutations* that happen to be inherited together on the same physical DNA molecules. These effects, collectively known as *genetic linkage*, create complicated correlations along the genome that are difficult to disentangle from each other. This makes it very difficult to predict how microbial populations will evolve.

In addition to genetic linkage, a second major challenge is that the forces of natural selection, recombination, and genetic drift will each often act over a large range of length and time scales within a population. For example, mutations in different genes can have a broad range of effects on cellular fitness, and will be amplified or purged at different rates [56, 57]. Similarly, recombination breaks up genetic linkage at vastly different rates depending on the distance between the mutations [24, 58–60]. The overall magnitude of genetic drift can also vary over time due to bottlenecks and other demographic changes [61, 62]. This large number of internal time scales can make it difficult to predict how these basic evolutionary forces contribute to the emergent patterns of diversity within a population.

Existing approaches from population genetics tend to neglect one or more of these forces [8–10], or assume that they can be captured by an appropriate rescaling of the other parameters [11–17]. The most well-known example is the *neutral theory of molecular evolution* [63], which neglects the action of natural selection. Dramatic simplifications occur in this limit, enabling detailed inferences of demographic history and other evolutionary parameters using tools derived from coalescent theory [64–68]. However, while these "backward-time" approaches are particularly well-suited for statistical inference, a growing number of studies have shown that they often fail to capture key features of empirical data [17–20], particularly in

microbial populations where the interactions between natural selection and genetic linkage are thought to play an important role [21–25].

More recent work has started to account for these effects by focusing on the opposite extreme, and studying the collective behavior of many linked mutations in the absence of recombination [10, 69–80]. Significant analytical progress has been possible in this case by exploiting connections to "traveling wave" models of front propagation in other areas of statistical physics [81–84]. However, while these traveling wave models have provided significant insights into the dynamics and diversity of *asexual* populations, from laboratory evolution experiments [21, 85–87] and certain cancers [88, 89], to the nonrecombining gene segments in influenza [90, 91], they have been difficult to extend to natural populations of bacteria, where recombination is known to play a significant role [92].

Genetic exchange in bacteria is commonly associated with the acquisition of new genes or pathways, which are frequently observed to transmit across traditional species boundaries [93–96]. However, genetic material can also be acquired from more closely related strains, and can overwrite existing regions of the chromosome through homologous recombination [93]. This more subtle form of horizontal transfer acts to reshuffle genetic variants within species, similar to meiotic recombination in sexually reproducing organisms [55]. Bacterial recombination plays a critical role in many areas of microbial evolution, from the definition of bacterial species [97–100] to transition between genome-level selection and selection on individual genes or mutations [101–104]. The existence of this process is no longer controversial: a variety of studies have shown that bacterial recombination can occur so frequently that the genomes of many species are best viewed as "mosaics" of acquired genetic material [105–110, 23–25], similar to sexually reproducing organisms like humans [111]. However, despite this broad agreement about the pervasiveness of bacterial recombination, many of the *quantitative* aspects of this process are still poorly understood [92].

For example, little is currently known about the dominant mechanisms of horizontal transfer in most species, though phage infection, bacterial conjugation, and uptake of environmental DNA have all been shown to take place in certain cases [93]. The typical rates and lengths of transferred fragments are also controversial. Several approaches have been developed to infer these parameters from the fine-scale variation among sequenced genomes [112–119, 23–25, 60]. However, most of these existing methods are based on neutral models of molecular evolution, and therefore suffer from some of the inherent limitations discussed above. In addition, many of these methods make restrictive assumptions about population demography, or the distribution of recombination rates within and between different genomes. Previous studies have shown that these simplified models often fail to capture key features of the observed data [23–25, 119], which suggests that any parameter estimates should be treated with a degree of caution. This makes it difficult to answer many fundamental questions about the dynamics of bacterial recombination: How rapidly do bacterial genomes lose their clonal backbone and acquire a mosaic structure? How does this compare to the characteristic timescales of natural selection or genetic drift? Do the rates and lengths of transferred fragments vary over time, or as a function of genetic relatedness? What can these patterns tell us about the interactions between circulating genetic variants?

The present proposal aims to bridge this gap in our understanding by developing new theoretical methods to predict how the processes of mutation, selection, recombination, and genetic drift combine to shape the diversity of microbial genomes. Our theoretical approach is designed to exploit an additional degree of freedom inherent in modern bacterial genomics datasets. These studies increasingly contain genomes from thousands (or tens of thousands) of different individuals per species [120–123], with even larger sizes likely in the near future [124]. This makes it possible to estimate the frequencies of mutations (or larger combinations of mutations) across several orders of magnitude of frequency space. This additional dimension of the data provides new opportunities to leverage the *scaling properties* of genetic diversity across a wide range of frequency scales to start to disentangle the population genetic processes that act on different time scales. Frequency-stratified measurements have a long history of application in population genetics, and underlie several common methods for inferring selection and demographic history from the shape of the mutation frequency distribution [125–128, 28]. However, most of these existing approaches rely on complex numerical calculations or computer simulations of specific evolutionary models. ***There is currently no analytical theory that allows us to predict how simple summaries of genetic diversity scale across different frequency ranges under the joint action of natural selection, recombination, and genetic drift. The central goal of this proposal is to develop***

theoretical methods to predict these scaling behaviors, and to use these methods to study the dynamics of bacterial recombination in diverse species of human gut bacteria. This analysis will help address fundamental questions about the maintenance of bacterial diversity across a broad range of time scales, and will allow us to better exploit the unique dimensions of modern genomic datasets.

3. Research Plan

Aim 1: Predict how selection and recombination influence correlations between mutations across a broad range of length and time scales.

Rationale: The statistical associations between mutations, collectively known as *linkage disequilibrium* (LD), are a critical component of fine-scale diversity. Correlations between mutations enable the genome-wide association studies that are used to map the genetic basis of complex traits [129, 130]. Genetic correlations also play a critical role in evolution, since combinations of linked mutations constitute the raw material on which natural selection and other evolutionary forces can act. Yet while extensive theory has been developed for predicting the *marginal distributions* of mutation frequencies [131–134, 76, 79], *higher-order correlations* like LD remain poorly understood in comparison.

Classical measures of LD typically focus on the pairwise correlation coefficient,

$$r^2 = \frac{(f_{AB} - f_A f_B)^2}{f_A(1-f_A)f_B(1-f_B)}\ , \qquad \textit{Eq. (1)}$$

where f_A and f_B denote the marginal frequencies of the two mutations, and f_{AB} denotes the fraction of individuals that possess both mutations; r^2 measures how the joint distribution of a pair of mutations differs from a null model where the mutations are independently distributed across individuals. Recombination tends to break down these correlations on average, while selection and drift tend to enhance them. A celebrated result by Ohta and Kimura [135] shows that, in a neutral population with constant size N, the frequency-weighted expectation of r^2 is given by

$$\sigma_d^2 = \frac{\mathbb{E}[r^2 \cdot f_A(1-f_A)f_B(1-f_B)]}{\mathbb{E}[f_A(1-f_A)f_B(1-f_B)]} = \frac{5+NR}{11+13NR+2(NR)^2}\ , \qquad \textit{Eq. (2)}$$

where R is the recombination rate between the two mutations; this expression approaches a constant value when $NR \ll 1$ and decays as $\sim 1/NR$ when $NR \gg 1$. The shape of this "LD curve" is frequently used to estimate recombination rates in empirical settings, e.g., by examining how genome-wide averages of σ_d^2 (or related quantities) decay as a function of the distance between sites [23, 24, 60, 117].

While this classical result has been enormously influential for building intuition about linkage disequilibrium, it suffers from several limitations that are increasingly important in modern genomic datasets. Chief among these is the absence of natural selection. While there has been some progress in predicting LD under specific selection scenarios (e.g. a single selective sweep [136]), we currently lack theoretical predictions for the empirically relevant case where a subset of the observed mutations are deleterious. This is a crucial limitation: most mutations are either neutral or deleterious, and numerous studies have documented differences in the genome-wide patterns of LD between synonymous and nonsynonymous mutations [137–140], where negative selection is thought to play an important role. Several recent studies have begun to explore these effects in computer simulations [29, 139–142]. Yet without a corresponding analytical theory, it can be difficult to understand how these patterns depend on the underlying parameters of the model, or to determine when more exotic forces like positive selection, genetic interactions, or ecological structure are necessary to explain the observed data.

A second and related limitation arises from the averaging scheme in σ_d^2, which effectively weights each pair of mutations by their joint heterozygosity, $f_A(1-f_A)f_B(1-f_B)$. This tends to favor mutations with intermediate frequencies ($10\% \leq f \leq 90\%$); these are usually older variants that have been circulating in the population for a long time. Understanding how LD varies across other frequency scales could provide new information about the evolutionary forces that operate on different time scales. Such an approach could be particularly useful for probing the dynamics of recombination, which are difficult to observe from single-site statistics alone. Yet at present, little is known about how different frequency and time scales contribute to the expected patterns of linkage disequilibrium within a population. This limited frequency resolution is particularly problematic in the presence of natural selection, which is known to strongly influence the distribution of mutation frequencies. This makes it difficult to interpret the varying

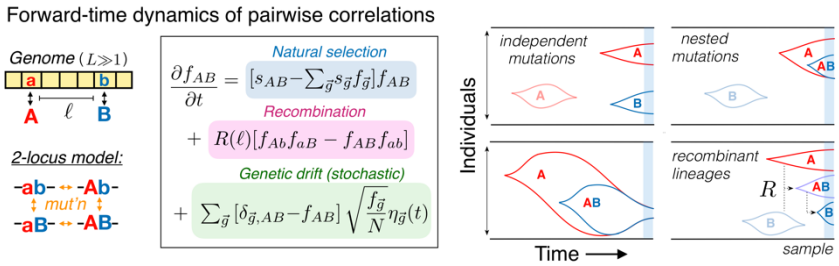

Fig 1: Schematic showing how correlations between mutations emerge from the forward-time dynamics of their underlying lineages (right). The frequencies of these lineages can be modeled by a set of nonlinear stochastic differential equations (left) under the joint action of natural selection, recombination, and random genetic drift.

LD patterns that are observed between different classes of selected sites in natural populations, or to use these observations to make inferences about the underlying recombination process.

In **Aim 1**, we plan to address this challenge by developing a new theoretical approach to predict how correlations between neutral and deleterious mutations scale as a function of their present-day frequencies. Our approach will leverage asymptotic methods we and others have recently developed for modeling the stochastic trajectories of rare mutations in related population genetic contexts [79, 143–146]. We will use these results to analyze new LD measurements from a large collection of human gut bacteria, in order to systematically probe the rates of recombination across a broad range of time scales.

Theoretical Methods and Preliminary Results: To quantity the dynamics of LD on different frequency scales, we will focus on a generalization of Ohta and Kimura's σ_d^2 statistic,

$$\sigma_d^2(f_0) = \frac{\mathbb{E}[r^2 \cdot f_A(1-f_A) f_B(1-f_B) \cdot e^{-f_A/f_0 - f_B/f_0}]}{\mathbb{E}[f_A(1-f_A) f_B(1-f_B) \cdot e^{-f_A/f_0 - f_B/f_0}]}, \qquad \textit{Eq. (3)}$$

where f_0 is a characteristic frequency scale. The exponential terms in this definition act like a "soft" version of a step function, preferentially excluding mutations with frequencies much larger than f_0. By scanning over a large range of f_0 values, this weighted version of Ohta and Kimura's σ_d^2 statistic allows us to systematically probe how linkage disequilibrium varies over different frequency scales.

We will characterize the asymptotic behavior of these *frequency-resolved LD measures* for pairs of genetic loci that evolve under the joint action of mutation, selection, recombination, and genetic drift. While these *two-locus models* are simple to write down (**Fig 1**), an analytical understanding of their behavior has so far remained elusive. We plan to overcome this challenge using novel "forward-time" approaches that we have recently developed for modeling the stochastic trajectories of rare mutations. Existing methods for predicting LD using coalescent theory [60, 147] or moment hierarchies [29, 135, 148] rapidly become unwieldy in larger samples, or when higher-order moments are required. Forward-time approaches offer a powerful alternative: rather than tracking unobserved genealogies or moments, these methods directly model the trajectories of observable lineages using a lower-dimensional system of coupled Langevin equations (**Fig 1**). The key simplification is that at small frequency scales ($f_0 \lesssim 10\%$), these coupled Langevin equations reduce to a nearly linear form that can be analyzed perturbatively using branching process methods. Since most mutations reside at these low frequencies, this asymptotic approach is particularly well-suited for characterizing the scaling properties of LD, since it exploits – rather than suffers from – the large sample sizes of modern genomic datasets.

In preliminary work (currently available in preprint [149]), we have shown that we can use this approach to derive analytical predictions for $\sigma_d^2(f_0)$ for pairs of neutral and deleterious mutations in a population of constant size N. Our results show how the scaling behavior at low frequencies collapses onto a lower dimensional manifold that depends on compound parameters like NRf_0 (**Fig 2**). These scaling laws suggest a novel approach for measuring recombination rates in empirical data, by varying the

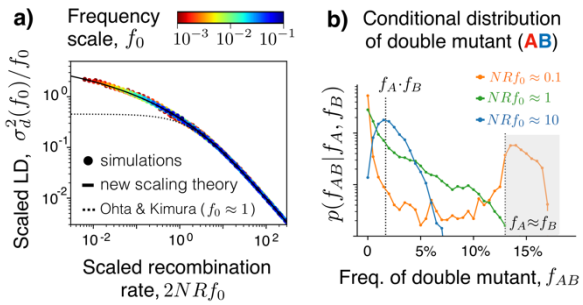

Fig 2: (a) Our frequency-resolved LD statistic, $\sigma_d^2(f_0)$, collapses onto a universal curve across a range of recombination rates and frequency scales, which can be predicted with our new theory. **(b)** Additional information is contained in the shape of the conditional distribution of f_{AB}, which undergoes a transition at different values of NRf_0. This suggests a new approach for measuring recombination rates.

frequency scale f_0 at a fixed genomic distance. By identifying the critical frequency where the effects of recombination are "frozen out" ($NRf_0^* \sim 1$; **Fig 2**), we can obtain a corresponding estimate for NR.

Our next steps will focus on extending these calculations to other LD statistics that are better suited for measuring recombination rates. Traditional metrics like $\sigma_d^2(f_0)$ and its cousins are poorly suited for this task, since they quantify the *enhancement* of genetic linkage relative to an infinite recombination limit. This makes it difficult to isolate the effects of recombination from other factors (e.g. correlated mutation rates [150] or epistasis [139]) that can enhance these correlations even in the absence of recombination. We plan to address this challenge by exploring a new family of *linkage equilibrium* (LE) metrics,

$$\Lambda(f_0) = \frac{\mathbb{E}[f_{AB} f_{Ab} f_{aB} f_{ab} \cdot e^{-f_A/f_0 - f_B/f_0}]}{\mathbb{E}[f_A^2(1-f_A)^2 f_B^2(1-f_B)^2 \cdot e^{-f_A/f_0 - f_B/f_0}]} \,. \qquad \textit{Eq. (4)}$$

This metric can be viewed as a quantitative generalization of the classical 4-gamete test [151], which vanishes in the absence of recombination, and approaches $\Lambda(f_0) \approx 1$ in the infinite recombination limit. We will also analyze the full conditional distribution of double mutant frequencies, $p(f_{AB}|f_A, f_B)$, since our preliminary results suggest that this distribution displays qualitatively distinct shapes at different values of NR (**Fig 2B**). Understanding the scaling behavior of these new statistics will enable more robust approaches for quantifying rates of bacterial recombination in genomic data.

Applications to genomic data from a large panel of human gut bacteria: We plan to compare our theoretical predictions to fine-scale diversity data from a large collection of human gut bacteria assembled by the Unified Human Gastrointestinal Genome (UHGG) project [123]. Human gut bacteria provide several unique advantages for our purposes. Recent advances in culturing and metagenomics have enabled the recovery of >100,000 bacterial genomes from >10,000 human subjects [2, 123]. These reconstructed genomes are distributed across a range of phylogenetically diverse species, many of which have been sampled in >1,000 different individuals. These large sample sizes allow us to focus on general trends that are shared by many closely related species [24], rather than the idiosyncratic features of any single population. In addition, the underlying biology of commensal gut bacteria – along with the randomized nature of many microbiome sampling efforts – makes them less susceptible to the "epidemic" outbreaks that are known to bias the genetic diversity of many pathogens [106, 112]. These features provide an ideal setting in which to study the dynamics of bacterial recombination.

Previous work has shown that the human gut is a hot spot for horizontal gene transfer *between species* [94–96], but the dynamics of recombination *within species* are still poorly understood. In a recent study [24], we found evidence for widespread recombination in the core genomes of many gut species, by measuring how σ_d^2 decays with the distance between mutations (ℓ) on genomes sampled from unrelated hosts (**Fig 3A**); this reflects the long-term action of recombination over multiple host colonization cycles. At present, however, it is difficult to use these LD curves to obtain quantitative estimates of the underlying recombination rates, since the mapping between physical distance and R depends on many unknown details of the relevant DNA transfer mechanisms (e.g. the length distribution of transferred fragments).

Our preliminary theoretical results suggest a new way to overcome this problem, by examining how frequency-resolved statistics like $\sigma_d^2(f_0)$ and $\Lambda(f_0)$ scale with f_0 at a fixed physical distance. We aim to explore this idea by constructing an analogous set of LD curves for ~40 prevalent bacterial species in the UHGG collection. We will then compare these measurements with our theoretical predictions derived above. Preliminary results obtained using data from a smaller cohort already show some qualitative agreement with our predictions (**Fig 3**), most notably in the shape of the conditional distribution $p(f_{AB}|f_A, f_B)$ at different physical distances (e.g. compare **Figs 3C & 2B**).

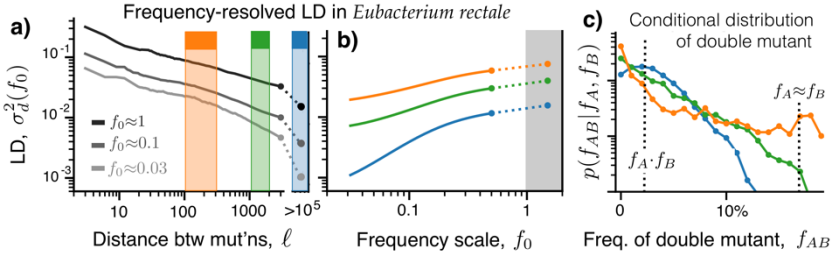

Fig 3: Scaling behavior of linkage disequilibrium (LD) in a global population of human gut bacteria. (a,b) Preliminary data showing frequency-resolved LD, $\sigma_d^2(f_0)$, between pairs of synonymous mutations in the core genomes of ~100 *Eubacterium rectale* strains sampled from different hosts. **(c)** Conditional distribution of the double mutant frequency f_{AB} when $f_A \approx f_B \approx 15\%$.

This suggests that there is at least some power to resolve individual recombination rates from the scaling behavior of frequency-resolved LD metrics. We plan to explore this idea by developing a pipeline to scale up these measurements to the approximately ~40 species in the UHGG collection with at least 1000 sampled genomes, and to extend them to new LD metrics like $\Lambda(f_0)$ in Eq. (4) above. This will allow us to quantify how LD varies across almost three decades of frequency space, both within and between genes, and between synonymous and nonsynonymous mutations, providing a rich source of empirical "scaling laws" to guide the additional theoretical work in **Aims 2 & 3** below.

Expected Outcomes and Significance: The theoretical methods developed in **Aim 1** will allow us to predict for the first time how correlations between mutations vary as a function of their frequency scale, recombination rate, and their additive and epistatic fitness costs. These scaling results will enable new methods for inferring bacterial recombination rates across a broad range of length and time scales, which require fewer assumptions about the mechanisms of the DNA transfer process (and could therefore be useful for inferring them). Our empirical applications to the UHGG dataset will allow us to test these predictions across a phylogenetically diverse range of human gut bacteria, and will help identify deviations from the simple model that require additional theoretical explanation.

Aim 2. Extend the analytical framework in Aim 1 to account for non-equilibrium demography.

Rationale: While the results of **Aim 1** will provide important intuition about the scaling properties of LD, they make a crucial assumption that the rates of genetic drift are constant over time. This is an important limitation: natural population sizes are never truly constant, and the magnitude of genetic drift will generally vary – sometimes dramatically – in response to these population size changes [152]. A major focus of modern population genetics has been to use this connection to reconstruct detailed demographic histories of specific populations (e.g. humans [153]). In our case, however, these historical population sizes are not the primary object of interest. We will often be more interested in inferring recombination rates or selection strengths in ways that are robust to changing population sizes. In other cases, we would like to know whether a given observation is incompatible with *any* time-varying population size $N(t)$, or whether more interesting scenarios like pervasive genetic hitchhiking [134] or ecological diversification [154] are required to explain the observed data. In these settings, $N(t)$ behaves more like a nuisance parameter, which we would like to "integrate out" if possible. However, the high-dimensional nature of this parameter makes this a challenging task.

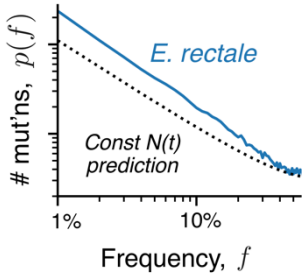

Fig 4: Frequency distribution of synonymous mutations in the *E. rectale* population from Fig 3. The observed distribution (blue) deviates from the equilibrium prediction (dashed line).

Much of existing our intuition about non-equilibrium demography comes from coalescent theory [64]. The behavior is particularly simple in this picture: variable population sizes lead to a time-varying rate of coalescence, $p_c(t) = 1/N(t)$, between each pair of individuals. This implies that an average pair of individuals will share a common ancestor roughly

$$\langle T_{\mathrm{MRCA}} \rangle = \int_0^\infty e^{-\int_0^T \frac{dt}{N(-t)}} dT \qquad \textit{Eq. (5)}$$

generations ago; $\langle T_{\mathrm{MRCA}} \rangle$ defines a characteristic timescale over which fine-scale genetic diversity can accumulate. This same idea can be extended to understand how time-varying population sizes influence the frequencies of segregating variants as well [64]. In a constant population, neutral variants follow a universal scaling law $p(f) \propto 1/f$, with an overall magnitude set by $\langle T_{\mathrm{MRCA}} \rangle$. Compared to this baseline, population expansions lead to fewer coalescence events in the recent past – and a corresponding enrichment of rare mutations – while population crashes tend to produce the opposite effect. These effects lead to well-known deviations from the classical $p(f) \propto 1/f$ result above. Such deviations are routinely observed in natural populations [152], including the gut bacteria in **Aim 1** (**Fig 4**). This could potentially explain the widespread deviations we have observed in their LD curves as well (**Fig 3A**; [24]).

However, while this coalescent picture has been enormously influential in shaping our intuition about non-equilibrium demography, our quantitative understanding of these effects remains limited – particularly in larger samples. Even the simplest observables, e.g. the neutral frequency distribution $p(f)$, still lack a

compact analytical description for a general $N(t)$. Coalescent methods can be used to derive formal series solutions for $p(f)$ in finite samples [133]. However, the number of terms in these series scales with the size of the sample, which makes it difficult to extrapolate their predictions to larger samples. This makes it hard to determine which aspects of $N(t)$ will "matter" at a given frequency scale [155].

The situation becomes even more complicated for multi-site statistics, or when natural selection is present. Many studies have attempted to approximate the effects of non-equilibrium demography in an ad-hoc manner, by extrapolating the equilibrium predictions with an effective population size $N_e \approx \langle T_{\mathrm{MRCA}} \rangle$ [60, 117, 119, 127, 156]. While this can be a reasonable approximation in some settings [157], it is known to break down in many others [28, 29, 158–160], particularly when the important dynamics occur on timescales much shorter than $\langle T_{\mathrm{MRCA}} \rangle$. More recent work has sought to overcome these challenges using numerical methods. These vary in complexity, from agent-based simulations [31–34, 37] to numerical integration of the Fokker-Planck PDEs [27, 28] or their corresponding moment hierarchies [29, 30]. While these computational approaches provide a principled way to predict the effects of a ***given*** $N(t)$, it can be difficult to search over this vast parameter space to make inferences about the underlying evolutionary forces. Existing methods typically try to infer lower-dimensional representations $N(t)$ from finite collections of summary statistics. However, since we rarely know which aspects of $N(t)$ will be relevant for different statistics *a priori*, it can be hard to choose these inputs and outputs in a principled way. For similar reasons, it is difficult to use these numerical approaches to determine whether an observation is fundamentally incompatible with ***any*** simple demographic model. This is an important limitation: many of the qualitative effects of non-equilibrium demography can also be produced by other scenarios, including pervasive genetic hitchhiking [134] or "ecotype" structure within species [154]. Understanding which of these scenarios is more relevant in natural populations is one of the major open questions in population genetics [39]. Identifying scaling patterns that can rigorously distinguish between these scenarios would therefore be a crucial step forward – one that will critically rely on additional analytical progress.

In **Aim 2**, we plan to address this challenge by extending the forward-time framework in **Aim 1** to account for non-equilibrium demography. We will use these results to explore the scaling properties of both single- and multi-site diversity statistics in the presence of selection and recombination, and we will compare these predictions to data from the human gut bacteria in **Aim 1**. We will use these results to determine which features of $N(t)$ "matter" at different frequency or time scales, and to identify general scaling patterns that cannot be produced by variable population sizes alone.

Theoretical Methods and Preliminary Results: Our theoretical approach will utilize the same forward-time picture we employed in **Aim 1**. As above, the basic idea is to sum over the possible mutation trajectories that could contribute to a present-day frequency f (a type of *path integral* [161]; **Fig 5**). The key difference is that we will now assume a general time-varying population size, $N(t)$. At low frequencies ($f \lesssim 10\%$), the dynamics of these variants can still be reduced to a nearly linear form (similar to **Aim 1**) except that the strength of genetic drift will now vary over time. This will alter the simple mapping between frequency and time ($t \leftrightarrow Nf$) that was implicit in our analysis above.

In unpublished preliminary work, we have found that we can overcome this problem by exploiting a new decomposition for the trajectories of rare mutations based on their age T (**Fig 5**). By integrating over these stochastic trajectories, we can obtain a novel analytical formula for the frequency distribution of neutral variants,

$$p(f) \propto \int_0^\infty \frac{dT}{\left(\int_0^T \frac{dt}{2N(-t)}\right)^2} \cdot e^{-f/\int_0^T \frac{dt}{2N(-t)}} \quad , \qquad \textit{Eq. (6)}$$

which is valid when $f \ll 1$. This compact analytical formula allows us to directly observe how different timescales contribute to the marginal frequency distribution in non-equilibrium populations. For example, Eq. (6) shows that the most likely age of a mutation with present-day frequency f is given by

$$f = \int_0^{\hat{T}(f)} \frac{dt}{2N(-t)} \, , \qquad \textit{Eq. (7)}$$

which generalizes the simple mapping between frequency and time above [162]. Roughly speaking, this result tells us that the only timescales of $N(t)$ that contribute to the frequency distribution near f are those

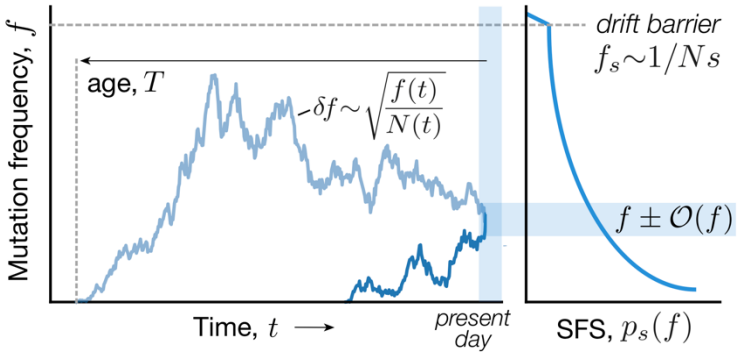

Fig 5: Forward-time picture of mutation trajectories. The observed distribution of mutation frequencies $p_s(f)$ (right) can be expressed as a sum (or path integral) over random mutation trajectories (left) with different ages T, and the same final frequency f. Genetic drift produces fluctuations that depend on the instantaneous frequency $f(t)$ and the population size $N(t)$. In constant populations, genetic drift prevents natural selection from purging deleterious mutations below a characteristic frequency $f_s \sim 1/Ns$ (the "drift barrier").

with $t \lesssim \hat{T}(f)$. This provides a principled way to "invert" an observed frequency distribution (e.g. **Fig 4**) to infer the underlying population size $N(t)$ – which is easier to interpret than existing methods based on formal series expansions [128]. The neutral frequency spectrum is only the simplest possible application of this approach. We now aim to use this basic framework to answer two fundamental questions about how selection and recombination interact with genetic drift in non-equilibrium settings:

The "drift barrier" and the efficiency of natural selection. Natural selection purges deleterious variants from the population, while genetic drift allows them to proliferate. The competition between these two forces enables deleterious mutations to grow to intermediate frequencies. In constant populations, the ratio between the frequency distributions of neutral and deleterious variants scales as

$$\frac{p_s(f)}{p(f)} = e^{-2Nsf} \ , \qquad \textit{Eq. (8)}$$

where s is the fitness cost of the mutation [131]. Eq. 8 shows that natural selection is only effective above a characteristic *"drift barrier"* $f_s \sim 1/Ns$. This celebrated result places strong constraints on evolution's ability to maintain a desired trait [163–165]. It is also frequently used to infer the prevalence and strength of negative selection in empirical settings, e.g. by comparing the frequencies of synonymous and non-synonymous variants and identifying the frequency scales at which selection is "frozen out" [127]. Yet despite its central importance, we still lack an analytical understanding of how the drift barrier generalizes to non-equilibrium settings. This makes it difficult to apply this powerful principle in practice.

We aim to address this challenge using our forward-time framework above. In preliminary work, we have found that we can extend our earlier result for $p(f)$ to handle deleterious mutations:

$$p_s(f) \propto \int_0^\infty \frac{e^{-sT}\, dT}{\left(\int_0^T \frac{e^{-s\cdot t}}{2N(-t)}\right)^2} \cdot e^{-f/\int_0^T \frac{e^{-s\cdot t}}{2N(-t)}} \ . \qquad \textit{Eq. (9)}$$

This result allows us to directly observe how selection and genetic drift interact on different frequency and time scales, and how this relates to the underlying features of $N(t)$. By comparing this expression to the formula for $p(f)$ above, we will derive asymptotic expressions for the characteristic frequency where $p_s(f)$ becomes an order-of-magnitude smaller than $p(f)$; this will constitute a non-equilibrium generalization of the classical drift barrier $f_s \sim 1/Ns$. We will then compare these results to the measured frequency distributions of synonymous and nonsynonymous variants from the human gut bacteria in **Aim 1**. This will allow us to estimate the typical fitness costs of protein-coding mutations in a range of phylogenetically diverse species.

Scaling properties of linkage disequilibrium (LD). We will also extend this approach to derive non-equilibrium versions of our frequency-resolved LD statistics from **Aim 1**. By combining these results with our predictions for $p(f)$ above, we will develop scaling approaches for inferring recombination rates that are robust to time-varying population sizes. We will also focus on identifying scaling patterns that cannot be reproduced by demography alone. Our initial steps will focus on a particular pattern motivated by the preliminary data from the gut bacteria in **Fig 3**. When compared with our equilibrium predictions (**Fig 2A)**, we see that data follow the same qualitative trend with increasing distance ℓ (i.e., greater $R(\ell)$ → lower LD). However, the scaling with f_0 follows the opposite of the predicted trend, with larger distances

experiencing *greater* relative reductions in LD at lower frequency scales (**Fig 6**) – as if they were experiencing *higher* effective recombination rates. While non-equilibrium demography will generally alter the shape of the scaling curve in **Fig 2B,** we suspect that this particular trend is fundamentally inconsistent with a time-varying population size: Eq. (7) suggests that the relationship between frequency and time – though warped by $N(t)$ – is still monotonic, so there should always be *fewer* opportunities for recombination at lower frequencies. We plan to test this idea using our analytical framework, by deriving bounds on the monotonicity of ratios like $\sigma_d^2(f_0, \ell_1)/\sigma_d^2(f_0, \ell_2)$. We will also measure the prevalence of this signal across the larger panel of gut species in **Aim 1**. This will allow us to obtain robust evidence that linked selection or other evolutionary scenarios are required to explain the data.

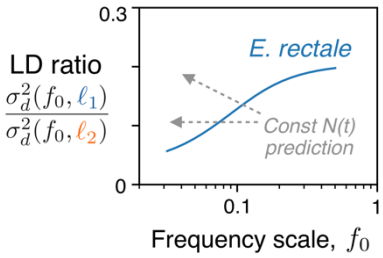

Fig 6: Identifying scaling properties of LD that cannot be produced by time-varying population sizes. The ratio between the blue and orange lines from the *E. rectale* population in Fig 3B. The observed data decline at lower frequencies, while the theory in Fig 2A predicts a constant or upward trend (dashed lines).

Expected Outcomes and Significance: The theoretical results in **Aim 2** will provide the first analytical predictions for the distributions of mutation frequencies and linkage disequilibria in non-equilibrium populations. These analytical results will allow us to predict which features of $N(t)$ interact with selection and recombination at different frequency and time scales, and will suggest new ways of measuring these quantities directly from the empirical scaling patterns observed in data. Our analytical framework will also allow us to identify scaling patterns that cannot be produced by demography alone. These "no-go theorems" will help resolve key open questions about the relative contributions of selection and demography in shaping the fine-scale diversity of natural populations – including the human gut bacteria examined in this project. Our basic framework will be useful for extending these scaling ideas to other evolutionary scenarios, e.g. population structure [29] or recurrent genetic hitchhiking (**Aim 3**), that may also be important for microbes.

Aim 3. Develop model-independent approaches to directly measure recombination events among closely related bacterial strains in large cohorts.

Rationale: The previous sections used the frequencies of mutations as a type of *coarse-graining* scheme to probe the evolutionary dynamics on different time scales. However, these site-based decompositions are not the only possible way to coarse-grain microbial genomes: are there other coarse-graining schemes that could take advantage of the higher-order correlations along the genome as well? These "collective modes" would be particularly useful for probing the dynamics of bacterial recombination: transferred DNA fragments often contain multiple linked mutations, and the correlations between these variants encode information about the lengths and sources of the acquired DNA. In addition, since many individual transfers are required to tile a bacterial genome, these populations also possess an emergent timescale, T_{mosaic}, below which their genomes "crystallize" into clonal backbones with global correlations across their entire length. This *mosaic timescale* plays a crucial role in bacterial evolution, since it controls the transition between genome-level selection and selection on individual genes or mutations.

Despite their importance, these collective dynamics of bacterial recombination are still poorly characterized. Existing efforts to infer these key parameters have mostly relied on *parametric approaches*, which attempt to fit the fine-scale diversity of microbial genomes using simple null models from population genetics [25, 60, 110, 112–119]. However, most of these existing methods are based on neutral models of molecular evolution with a constant rate of genetic drift. They also tend to utilize relatively simple models of recombination, e.g. with exponentially distributed transfer lengths and uniform rates of recombination among strains. These are important limitations, as a number of recent studies – including our preliminary data in **Figs 3, 4 & 5** above – have shown that these simple models often fail to capture the global patterns of diversity in many bacterial populations [23–25, 119]. Little is currently known about how these and other model misspecification errors might bias our existing parameter estimates. While the theoretical methods in **Aims 1 & 2** offer some additional prospects for extending this approach to non-equilibrium settings, these pairwise models will still break down if linked selection or ecological structure are more common. This highlights the central challenge with any "global" parametric

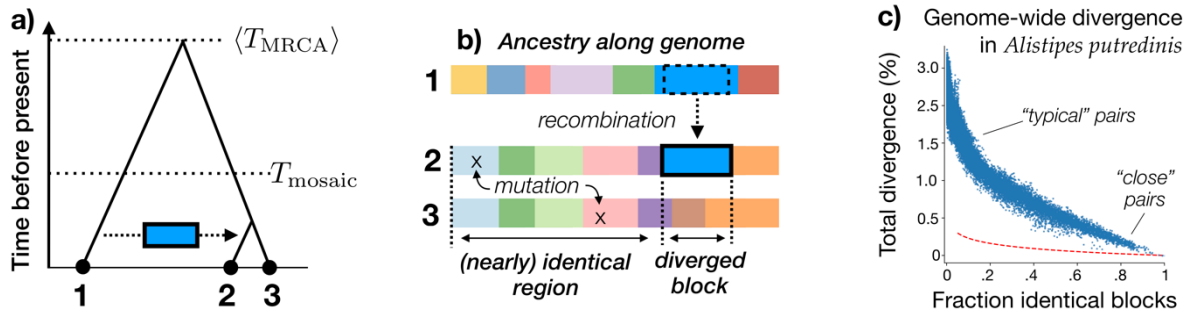

Fig 7: Accumulation of homologous recombination events in quasi-sexual bacterial populations. (a) Most pairs of strains share a common ancestor $\sim\langle T_{MRCA}\rangle$ generations ago (e.g. strains 1 & 2), and their ancestral genomes are completely overwritten by subsequent recombination events **(b)**. However, if a pair of strains happens to share a common ancestor more recently than the *clonal crystallization timescale* T_{mosiac} (e.g. strains 2 & 3), their accumulated recombination events will be visible as "blocks" of typical divergence against a backdrop of nearly identical DNA sequence. **(c)** Preliminary data suggest that this pattern can explain the broad range of genetic divergence observed in many species of human gut bacteria. Blue points show pairwise comparisons between ~250 *Alistipes putredinis* strains sampled from different hosts; red line shows the expectation for clonal mutations.

inference method: when fine-scale diversity emerges from a mixture of evolutionary forces, it is often necessary to model all of them accurately to make inferences about any single one. This is a major barrier in our ability to infer the dynamics of bacterial recombination from genomic data.

In **Aim 3**, we plan to overcome this challenge by developing a "non-parametric" approach to directly resolve large numbers of recombination events in microbial genomes in a model-independent fashion. Our approach aims to exploit the local nature of bacterial recombination, along with a common empirical feature of large microbial datasets. Previous work by our group [24] and others [25, 60, 118, 119] has shown that the genetic divergence between different pairs of genomes can often vary over several orders of magnitude – even within the same species. This broad range of genetic divergences between strains (which is a *collective property* of many individual sites) reflects the broad range of timescales in their shared genealogical history. Most pairs of strains share a common ancestor $\langle T_{\mathrm{MRCA}} \rangle \gg T_{\mathrm{mosiac}}$ generations ago, so that their present-day genomes now comprise a mosaic of overlapping recombination events (**Fig 7**). However, in sufficiently large samples, some pairs of strains will inevitably share a common ancestor on timescales much shorter than $\langle T_{\mathrm{MRCA}} \rangle$; in extreme cases, these may even fall below the crystallization timescale T_{mosiac} (**Fig 7A**). Among these "*closely related strains*", recombination will not have had enough time to completely cover the ancestral genome with DNA from other, more typically diverged strains. Rather, individual recombination events will be visible as "blocks" of typical genetic divergence ($\propto \langle T_{\mathrm{MRCA}} \rangle$) against a backdrop of nearly identical DNA sequence (**Fig 7B**) [25, 118]. We note that this basic signature is independent of most assumptions about natural selection or demography, and therefore provides a powerful approach for detecting recombination events directly from empirical data without the need for any sophisticated phylogenetic inference.

Previous studies have observed these partially recombined genomes in a handful of bacterial species – most notably in *E. coli* [25, 118]. Our preliminary results suggest that these dynamics could explain a large portion of the genetic divergence within many species of human gut bacteria as well (**Fig 7C**). While the forces responsible for these closely related strains are still poorly understood (see below), their empirical existence presents a valuable opportunity for disentangling and comparing the dynamics of recombination in a broad range of different species. In **Aim 3**, we will use this basic idea to develop a systematic approach for resolving the individual recombination events that accumulate between closely related strains in the large panel of human gut bacteria from **Aim 1**. These data will allow us to obtain high-resolution measurements of rates and lengths of transferred fragments in a variety of phylogenetically diverse strains. Guided by these results, we will also develop theoretical models to explore how the large number of closely related pairs might emerge from different evolutionary scenarios, including the hitchhiking of locally adaptive mutations.

Theoretical Approach and Preliminary Data: We will resolve individual recombination events by analyzing the spatial distribution of pairwise genetic differences along the "core genomes" of closely related strains. These pairwise differences reflect a mixture of all the genetic changes that accumulated

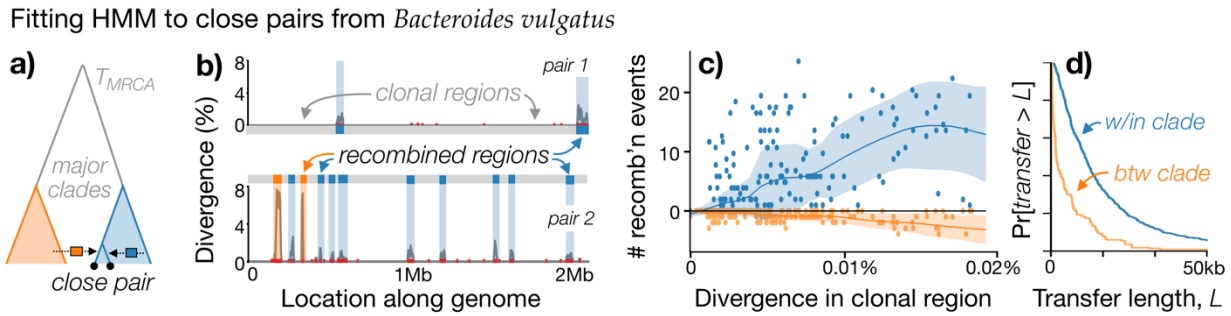

Fig 8: Hidden Markov model (HMM) for detecting recombination events between closely related strains. (a) Preliminary data for *Bacteroides vulgatus*, which has a strong population structure with two major clades. **(b)** The HMM partitions each closely related pair into clonal regions (grey) and recombined regions (blue, orange) based on their local sequence divergence; data from two example pairs are shown. **(c)** Preliminary data showing the # of detected transfers and # of mutations in clonal regions for ~250 pairs of closely related *B. vulgatus* strains from different hosts. **(d)** Distribution of transfer lengths for the recombination events in (c). These data show that the rates and lengths of successful transfers strongly depend on the divergence of the donated fragment.

within these two lineages since they last shared a common ancestor (**Fig 7**). We will assume that these changes arise through a mixture of two different processes: (i) point mutations (which alter individual sites) and (ii) homologous recombination events (which replace longer stretches of DNA with a corresponding fragment sampled from another strain in the population). The joint dynamics of these two processes can be quite complex in the general case (even in the absence of selection) due to the complexities of the underlying *ancestral recombination graph* [166–168]. Our approach will avoid these difficulties by exploiting on a key simplification that occurs at short timescales: when $t \ll T_{\text{mosaic}} \ll \langle T_{\text{MRCA}} \rangle$, mutation and recombination events can both be approximated by independent Poisson processes, with a negligible chance of overlapping. This is a crucial simplification, since it implies that the spatial patterns of genetic divergence can be captured by a *hidden Markov model* (HMM) that transitions between *clonal regions* (dominated by de novo mutations) and *recombined regions* (dominated by imported mutations) at different locations along the genome (**Fig 8B**). The underlying parameters of this HMM will vary between different pairs of strains, due to the differences in their time-aggregated rates of mutation and recombination. In this way, our HMM approximation resembles a lightweight version of existing parametric approaches like ClonalFrameML [114], except that it is robust to various forms of selection, non-equilibrium demography, and other deviations from the simple neutral models assumed in previous work.

We will use this HMM approximation to analyze closely related strains from the large panel of human gut bacteria in **Aim 1**. For simplicity, we will focus on the subset of samples from *unrelated* hosts in which the *dominant strains* can be confidently identified [24]; the recombination events among these strains will therefore represent *successful transfers* that have risen to appreciable frequencies within their host community. The basic signatures of these transfers are the "clusters" of genetic differences that are introduced together during a single recombination event [25, 118]. Modeling this imported variation is still a major challenge, however: the divergence of these transferred fragments can vary over several orders of magnitude, depending on the location along the genome and the identities of the donor and recipient strains. Existing approaches that attempt to capture this imported variation with a single parameter [25, 60, 113–115, 118] can therefore be severely biased in their ability to detect true recombination events.

We plan to overcome this challenge by exploiting the additional information contained in the *typically diverged* pairs of strains from the same species; these constitute the vast majority of the genomes in any given sample. We will use the empirical distribution of divergences at a given location to parameterize a mixture model for the divergence of each imported fragment – this will allow us to capture the broad variation in the divergence of different transfers in a way that is directly informed by the available data. In species with strong population structure (**Fig 8A**; [24]), we will also introduce additional HMM states to differentiate between within- vs between-clade transfers; this will allow us to model systematically different transfer lengths from donors with different degrees of genetic relatedness (as observed in some previous *in vitro* experiments [169]). We can then use standard dynamic programming techniques [170] to fit these HMMs to pairs of closely related strains in our dataset. This will allow us to obtain a list of the inferred recombination events for each pair of strains, along with the corresponding divergence in their non-recombined regions (a proxy for their T_{MRCA}).

To demonstrate the feasibility of this approach, we have recently used this method to identify ~1700 recombination events that have accumulated between ~250 pairs of closely related *Bacteroides vulgatus* strains that were sampled from different hosts (**Fig 8**). These preliminary data allow us to directly observe the accumulation of successful recombination events with time, as well as systematic differences in the rates and lengths of successful transfers within vs between the major *B. vulgatus* clades. Our next steps will focus on extending this approach to the larger panel of species (and the significantly larger sample sizes) contained in the UHGG collection from **Aim 1**. This will allow us to obtain a high-resolution view of the rates and lengths of successful recombination events across ~40 phylogenetically diverse species, and to determine how they vary between strains with different degrees of genetic relatedness. These data will allow us to address a number of open questions about the basic mechanisms of homologous recombination in these species [171], and can be directly contrasted with our population-level estimates from **Aims 1 & 2** above. Furthermore, by extrapolating our observations to longer timescales, we will be able to obtain an independent estimate of the crucial timescale T_{mosiac} that controls the transition between clonal vs quasi-sexual evolution.

These new estimates will also allow us to address a fundamental puzzle that underlies our entire approach: why do quasi-sexual bacterial populations have so many closely related strains in the first place? The existence of these strains was crucial for the analysis in **Fig 8**, but their underlying causes are still unclear. We previously showed [24] that specific features of the human gut system allow us to rule out common sampling biases (e.g. microepidemics / transmission chains / clonal blooms) that have been conjectured to play a role in other species (e.g. pathogens) [106, 112]. This suggests that the closely related strains in **Figs 7 & 8** emerge from evolutionary – rather than epidemiological – processes. In principle, these partially recombined pairs could emerge in a neutral population with sufficiently deep sampling: some cells will inevitably share a common ancestor $\ll T_{\text{mosaic}}$ generations ago and will not have had time to fully recombine yet. We plan to test this hypothesis by calculating the total probability of these early common ancestors as a function of the sample size n, the mosaic timescale T_{mosaic}, and the population demography, $N(t)$. By comparing these predictions across species, we will be able to assess whether the overall frequency of close pairs is consistent with this simple baseline expectation.

An alternative explanation is that these partially recombined strains arise through frequent genetic "hitchhiking" with sweeping beneficial mutations [134]. These sweeps can produce genome-wide correlations due to the inherent asymmetry of bacterial recombination: beneficial mutations will drag their entire genomic background to fixation unless they are successfully transferred onto another genetic background during the sweep (**Fig 9**). Conversely, recombination events in other regions of the genome will introduce blocks of diverged sequence within this nearly clonal background – similar to our earlier picture in **Fig 7**. While this hitchhiking scenario provides a plausible mechanism for producing the partially recombined strains we observe in data, the statistical properties of these HGT-mediated hitchhiking events are still poorly understood. We plan to explore this scenario by developing new theory to predict the frequency distributions of clonal backbones and recombined fragments that emerge from pervasive, HGT-mediated hitchhiking. We will also consider scenarios where the sweeping mutations are beneficial in only a subset of local environments [172, 173] – this scenario is likely to be particularly relevant for the global populations of human gut bacteria we consider here, where the local conditions can vary dramatically in different host communities. We will attack these problems using a heuristic approach we have recently developed for modeling the forward-time dynamics of recurrent genetic hitchhiking (similar to **Aims 1 & 2**) that incorporates the fictitious selection forces produced by genetic draft [174]. By comparing these theoretical predictions with the empirical data above, we will be able to test whether this common alternative scenario is broadly consistent with the data.

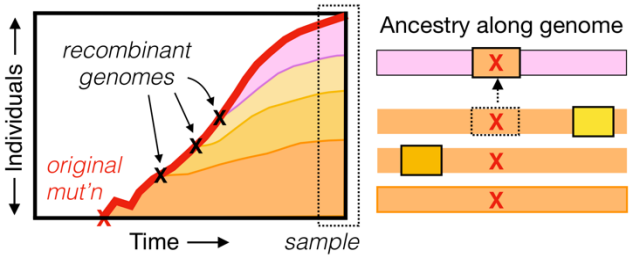

Fig 9: Genetic hitchhiking in quasi-sexual bacterial populations. A beneficial mutation (red) arises on an initial genetic background (orange) and sweeps through the population. The original genetic background will "hitchhike" in frequency until the beneficial mutation is transferred onto a different background (purple). Recombination elsewhere along the genome (yellow) will introduce blocks of local divergence on the original clonal backbone.

Expected Outcomes and Significance: The "non-parametric" inference methods developed in **Aim 3** will provide a high-resolution view of the rates and lengths of successful recombination events within ~40 different species of human gut bacteria. This will constitute one of the largest surveys of bacterial recombination to date. These data will allow us to address fundamental open questions about the dynamics of homologous recombination in these species, and how they vary between strains with different degrees of genetic relatedness. The theoretical methods developed in this Aim will provide the first analytical predictions for how these patterns can emerge from the "crystallization" of microbial genomes on short timescales. These results will help us resolve a longstanding puzzle about the origins of the broad range of genetic diversity observed within many bacterial species, and will provide a starting point for understanding the scaling behavior of linkage correlations under recurrent selective sweeps.

Timeline

This project will take 3 years to complete. In **Aim 1**, we will develop theory for predicting the scaling properties of different LD statistics in year 1, and we will compare these predictions to data from the large panel of human gut bacteria in year 2. We will then extend these approaches to non-equilibrium populations in **Aim 2** in years 2 and 3. In **Aim 3**, we will implement and apply our new inference method in year 1, and we will analyze the data and develop theory in years 2-3.

4. Broader Impacts

Training Opportunities: This project integrates methods from statistical physics, population genetics, microbial ecology, and computational biology. This will provide excellent interdisciplinary training opportunities for the postdoctoral fellow, graduate student, and undergraduate researcher who will be supported or partially supported by this proposal. The PI's group provides an ideal environment for this interdisciplinary work: the PI was trained in theoretical physics and evolutionary biology, and the group currently contains a mixture of graduate and undergraduate students with backgrounds in physics, evolutionary biology, and bioengineering. The PI's lab is located within Stanford's Bio-X institute, which is the hub for interdisciplinary research in biology and medicine at Stanford. The PI has active collaborations with several experimentalists on campus, including David Relman (Microbiology & Immunology), KC Huang (Bioengineering), and Dmitri Petrov (Biology), and students and postdocs regularly collaborate and interact across these groups. The Bio-X institute also hosts an undergraduate research internship program each summer. Underrepresented students are actively recruited, and the PI has hosted two freshman interns from this program over the last two years. The PI will continue to host one student from this program each summer.

Education and Outreach: The proposed work is similar to much of modern evolutionary genetics, in that it increasingly demands training that spans both the biological and mathematical sciences [175]. Physicists are uniquely poised to contribute at this interface, given their extensive training in both theoretical and applied problems. However, while substantial efforts have been devoted to improving quantitative training for biologists in recent decades [176–179], comparatively few resources exist for physics students seeking training in modern quantitative biology. This is particularly true for evolutionary biology and population genetics, where the underlying mathematical models are sufficiently complicated that they are rarely covered – even in graduate-level courses – in the traditional biology curriculum. Existing courses are most often geared toward "consumers" of population genetic methods [180], and assume that students have little familiarity with the mathematical tools (e.g. PDEs, series expansions, probability distributions) that are a core part of the undergraduate physics curriculum. Conversely, coverage of these topics in mathematics and statistics courses tends to emphasize formal properties (e.g. convergence proofs), with few connections to modern experimental data [181]. This makes it difficult for physical scientists to get a mathematically rigorous but biologically naive introduction to the field.

To fill this gap, the PI has developed a new mezzanine-level course that aims to provide an introduction to quantitative evolutionary modeling through the lens of statistical physics. The course covers topics ranging from the foundations of theoretical population genetics to experimental evolution in laboratory microbes, while emphasizing techniques like order-of-magnitude estimation and the method of successive approximations. Crucially for physics students, this course often provides their first formal exposure to non-equilibrium approaches in statistical physics (e.g. stochastic differential equations and continuous-time branching processes), which challenges them to move beyond the familiar mathematical

frameworks that have learned in their standard physics courses. The last iteration of the course was delivered online and was attended by ~30 graduate and undergraduate students, including virtual attendees from Harvard, Berkeley, Caltech, UCLA, and WashU. Handwritten lecture notes and homework problems are publicly available on the PI's website; during the present project, the PI will continue to develop this material and expand the handwritten notes into a fully typeset pedagogical document, which will be disseminated to the broader community. The PI will also work with graduate students in the group to adapt some of the order-of-magnitude estimation problems in the course to use for high-school outreach events that they have (and will continue to) participate in through the Stanford Splash program.

Finally, in collaboration with Dmitri Petrov, the PI will co-organize the Bay Area Population Genomics meeting in Fall 2022. This day-long meeting brings together population geneticists and evolutionary biologists in the bay area, and features student talks and posters from researchers at Stanford, UC Berkeley, UC Santa Cruz, UC Davis, UCSF, and San Francisco State University. Though traditionally attended by evolutionary biologists, the PI will actively solicit participation from students and postdocs in local biophysics groups (including Daniel Fisher, Shamit Kachru, Stephen Quake, KC Huang, and Oskar Hallatschek) in order to foster additional interactions between these fields.

Broader Research Impacts: Fine-scale microbial diversity is the fundamental fuel for evolution, and underlies some of the greatest challenges facing humanity today. A quantitative understanding of this genetic variation – and how it shifts and changes over time – will therefore have broader relevance across multiple subfields of biology. In particular, our empirical applications in this proposal focus on commensal human gut bacteria, which have become a central focus of modern biomedical research. Thus, in addition to providing general insights into the dynamics of bacterial recombination, the specific rates and mechanisms that we will infer for these species will also have broader relevance for existing efforts in the microbiome field, from mapping the genetic basis of adaptation to different host environments [182–184] to designing synthetic consortia of gut bacteria that are robust to short-term evolution [185]. We will facilitate these connections by releasing our inference methods and recombination estimates in publicly available repositories, and by disseminating our results at microbiome meetings (e.g. CSHL Microbiome). More broadly, the theoretical methods developed in this proposal will provide concrete examples of how concepts like scaling and coarse-graining – which play a central role in modern physics – can be extended to population genetic settings. These theoretical approaches will be relevant beyond bacteria, and will be useful for understanding the evolution of other natural populations (e.g. humans [186], SARS-CoV2 [1], cancer tumors [187]) where large numbers of sequenced genomes are now available.

5. Results from Prior NSF Support

DISSERTATION RESEARCH: Evolutionary Dynamics in Rapidly Evolving Populations
(PI Desai, Co-PI Good; 6/1/15-5/31/16, $21,936, NSF DEB-1501580)

Intellectual Merit. The goal of this project was to directly measure the dynamics of molecular evolution in the longest running laboratory evolution experiment that had been performed to date. Using a multiplexed DNA sequencing approach, along with novel statistical methods for tracking the dynamics of new mutations, we generated a series of high-resolution "molecular fossil records" covering 60,000 generations of evolution in 12 laboratory populations of *E. coli*. This project resulted in a major publication [86], in which we used these molecular fossil records to show that long-term adaptation to a fixed environment can involve a rich and dynamic set of population genetic processes – in stark contrast to the evolutionary desert expected near a fitness optimum.

Broader impacts. This project was a Doctoral Dissertation Improvement Grant for Dr. Good's PhD thesis. It provided a valuable interdisciplinary training opportunity to supplement Dr. Good's formal background in theoretical physics with hands-on training in experimental evolution and genomics. Dr. Good presented results from this work in numerous conferences and invited seminars, including the Society for Molecular Biology and Evolution's Walter M. Fitch Prize Symposium (2015) and the Microbiology Society meeting (2018). In addition, the large dataset produced by this study now serves as an important community resource for other researchers to investigate questions in experimental evolution and bacterial genetics. In keeping with our Data Management Plan, the raw sequencing reads have been deposited at NCBI, and a user-friendly plaintext database, as well as the source code for the sequencing pipeline, downstream analyses, and figure generation, have been posted on Github.

References Cited

1. Maxmen A (2021) One million coronavirus sequences: popular genome site hits mega milestone. *Nature*, 593(7857):21–21. https://doi.org/10.1038/d41586-021-01069-w

2. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, Collado MC, Rice BL, DuLong C, Morgan XC, Golden CD, Quince C, Huttenhower C, Segata N (2019) Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell*, 176(3):649-662.e20. https://doi.org/10.1016/j.cell.2019.01.001

3. Collins J, Robinson C, Danhof H, Knetsch C, Van Leeuwen H, Lawley T, Auchtung J, Britton R (2018) Dietary trehalose enhances virulence of epidemic Clostridium difficile. *Nature*, 553(7688):291–294.

4. Zlitni S, Bishara A, Moss EL, Tkachenko E, Kang JB, Culver RN, Andermann TM, Weng Z, Wood C, Handy C, Ji HP, Batzoglou S, Bhatt AS (2020) Strain-resolved microbiome sequencing reveals mobile elements that drive bacterial competition on a clinical timescale. *Genome Medicine*, 12(1):50. https://doi.org/10.1186/s13073-020-00747-0

5. Vasquez KS, Willis L, Cira NJ, Ng KM, Pedro MF, Aranda-Díaz A, Rajendram M, Yu FB, Higginbottom SK, Neff N, Sherlock G, Xavier KB, Quake SR, Sonnenburg JL, Good BH, Huang KC (2021) Quantifying rapid bacterial evolution and transmission within the mouse intestine. *Cell Host & Microbe*, 29(9):1454-1468.e4. https://doi.org/10.1016/j.chom.2021.08.003

6. Kim SG, Becattini S, Moody TU, Shliaha PV, Littmann ER, Seok R, Gjonbalaj M, Eaton V, Fontana E, Amoretti L, Wright R, Caballero S, Wang Z-MX, Jung H-J, Morjaria SM, Leiner IM, Qin W, Ramos RJJF, Cross JR, Narushima S, Honda K, Peled JU, Hendrickson RC, Taur Y, Brink MRM van den, Pamer EG (2019) Microbiota-derived lantibiotic restores resistance against vancomycin-resistant Enterococcus. *Nature*, 572(7771):665–669. https://doi.org/10.1038/s41586-019-1501-z

7. Schmidt TS, Li SS, Maistrenko OM, Akanni W, Coelho LP, Dolai S, Fullam A, Glazek AM, Hercog R, Herrema H, Jung F, Kandels S, Orakov A, Rossum TV, Benes V, Borody TJ, Vos WM de, Ponsioen CY, Nieuwdorp M, Bork P (2021) Drivers and Determinants of Strain Dynamics Following Faecal Microbiota Transplantation. :2021.09.30.462010. https://doi.org/10.1101/2021.09.30.462010

8. Ewens WJ (2004) Mathematical population genetics: theoretical introduction. 1

9. Walsh B, Lynch M (2018) Evolution and Selection of Quantitative Traits.

10. Fisher DS (2013) Asexual evolution waves: fluctuations and universality. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(01):P01011.

11. Charlesworth B (2009) Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, 10(3):195–205.

12. Santiago E, Caballero A (1998) Effective Size and Polymorphism of Linked Neutral Loci in Populations Under Directional Selection. *Genetics*, 149(4):2105–2117.

13. Nicolaisen LE, Desai MM (2013) Distortions in Genealogies due to Purifying Selection and Recombination. *Genetics*, 195(1):221–230. https://doi.org/10.1534/genetics.113.152983

14. Neher RA, Shraiman BI (2011) Statistical genetics and evolution of quantitative traits. *Reviews of Modern Physics*, 83(4):1283–1300. https://doi.org/10.1103/RevModPhys.83.1283

15. Weissman DB, Hallatschek O (2014) The Rate of Adaptation in Large Sexual Populations with Linear Chromosomes. *Genetics*, 196(4):1167–1183. https://doi.org/10.1534/genetics.113.160705

16. Neher RA, Kessinger TA, Shraiman BI (2013) Coalescence and genetic diversity in sexual populations under selection. *Proceedings of the National Academy of Sciences*, 110(39):15836–15841. https://doi.org/10.1073/pnas.1309697110

17. Good BH, Walczak AM, Neher RA, Desai MM (2014) Genetic Diversity in the Interference Selection Limit. *PLOS Genetics*, 10(3):e1004222. https://doi.org/10.1371/journal.pgen.1004222

18. Messer PW, Petrov DA (2013) Frequent adaptation and the McDonald–Kreitman test. *Proceedings of the National Academy of Sciences*, 110(21):8615–8620. https://doi.org/10.1073/pnas.1220835110

19. Ewing GB, Jensen JD (2016) The consequences of not accounting for background selection in demographic inference. *Molecular Ecology*, 25(1):135–141. https://doi.org/10.1111/mec.13390

20. Pouyet F, Aeschbacher S, Thiéry A, Excoffier L (2018) Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *eLife*, 7:e36317. https://doi.org/10.7554/eLife.36317

21. Lang GI, Rice DP, Hickman MJ, Sodergren E, Weinstock GM, Botstein D, Desai MM (2013) Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature*, 500(7464):571–574.

22. Zanini F, Brodin J, Thebo L, Lanz C, Bratt G, Albert J, Neher RA (2015) Population genomics of intrapatient HIV-1 evolution. *eLife*, 4:e11282. https://doi.org/10.7554/eLife.11282

23. Rosen MJ, Davison M, Bhaya D, Fisher DS (2015) Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. *Science*, 348(6238):1019–1023. https://doi.org/10.1126/science.aaa4456

24. Garud NR, Good BH, Hallatschek O, Pollard KS (2019) Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLoS biology*, 17(1):e3000102.

25. Sakoparnig T, Field C, Nimwegen E van (2021) Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species. *eLife*, 10:e65366. https://doi.org/10.7554/eLife.65366

26. Song YS, Steinrücken M (2012) A Simple Method for Finding Explicit Analytic Transition Densities of Diffusion Processes with General Diploid Selection. *Genetics*, 190(3):1117–1129. https://doi.org/10.1534/genetics.111.136929

27. Ragsdale AP, Gutenkunst RN (2017) Inferring Demographic History Using Two-Locus Statistics. *Genetics*, 206(2):1037–1048. https://doi.org/10.1534/genetics.117.201251

28. Kim BY, Huber CD, Lohmueller KE (2017) Inference of the Distribution of Selection Coefficients for New Nonsynonymous Mutations Using Large Samples. *Genetics*, 206(1):345–361. https://doi.org/10.1534/genetics.116.197145

29. Ragsdale AP, Gravel S (2019) Models of archaic admixture and recent history from two-locus statistics. *PLOS Genetics*, 15(6):e1008204. https://doi.org/10.1371/journal.pgen.1008204

30. Friedlander E, Steinrücken M (2021) A numerical framework for genetic hitchhiking in populations of variable size. :2021.03.25.437048. https://doi.org/10.1101/2021.03.25.437048

31. Garud NR, Messer PW, Buzbas EO, Messer PW (2015) Recent selective sweeps in North American ıt Drosophila melanogaster show signatures of soft sweeps. *PLoS Biology*, 11:e1005004.

32. Haller BC, Messer PW (2019) SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. *Molecular Biology and Evolution*, 36(3):632–637. https://doi.org/10.1093/molbev/msy228

33. Adrion JR, Cole CB, Dukler N, Galloway JG, Gladstein AL, Gower G, Kyriazis CC, Ragsdale AP, Tsambos G, Baumdicker F, Carlson J, Cartwright RA, Durvasula A, Gronau I, Kim BY, McKenzie P, Messer PW, Noskova E, Ortega-Del Vecchyo D, Racimo F, Struck TJ, Gravel S, Gutenkunst RN, Lohmueller KE, Ralph PL, Schrider DR, Siepel A, Kelleher J, Kern AD (2020) A community-maintained standard library of population genetic models. *eLife*, 9:e54967. https://doi.org/10.7554/eLife.54967

34. Schrider DR, Kern AD (2016) S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning. *PLOS Genetics*, 12(3):e1005928. https://doi.org/10.1371/journal.pgen.1005928

35. Sheehan S, Song YS (2016) Deep Learning for Population Genetic Inference. *PLOS Computational Biology*, 12(3):e1004845. https://doi.org/10.1371/journal.pcbi.1004845

36. Schrider DR, Kern AD (2018) Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends in genetics: TIG*, 34(4):301–312. https://doi.org/10.1016/j.tig.2017.12.005

37. Johri P, Charlesworth B, Jensen JD (2020) Toward an Evolutionarily Appropriate Null Model: Jointly Inferring Demography and Purifying Selection. *Genetics*, 215(1):173–192. https://doi.org/10.1534/genetics.119.303002

38. Schrider DR (2020) Background Selection Does Not Mimic the Patterns of Genetic Diversity Produced by Selective Sweeps. *Genetics*, 216(2):499–519. https://doi.org/10.1534/genetics.120.303469

39. Johri P, Aquadro CF, Beaumont M, Charlesworth B, Excoffier L, Eyre-Walker A, Keightley PD, Lynch M, McVean G, Payseur BA, Pfeifer SP, Stephan W, Jensen JD (2021) Statistical inference in population genomics. :2021.10.27.466171. https://doi.org/10.1101/2021.10.27.466171

40. Feder AF, Pennings PS, Petrov DA (2021) The clarifying role of time series data in the population genetics of HIV. *PLOS Genetics*, 17(1):e1009050. https://doi.org/10.1371/journal.pgen.1009050

41. Garud NR, Messer PW, Petrov DA (2021) Detection of hard and soft selective sweeps from Drosophila melanogaster population genomic data. *PLOS Genetics*, 17(2):e1009373. https://doi.org/10.1371/journal.pgen.1009373

42. Cardy J (1996) Scaling and Renormalization in Statistical Physics. https://doi.org/10.1017/CBO9781316036440

43. Stanley HE (1999) Scaling, universality, and renormalization: Three pillars of modern critical phenomena. *Reviews of Modern Physics*, 71(2):S358–S366. https://doi.org/10.1103/RevModPhys.71.S358

44. Nightingale P (1982) Finite-size scaling and phenomenological renormalization (invited). *Journal of Applied Physics*, 53(11):7927–7932. https://doi.org/10.1063/1.330232

45. Banavar JR, Green JL, Harte J, Maritan A (1999) Finite Size Scaling in Ecology. *Physical Review Letters*, 83(20):4212–4214. https://doi.org/10.1103/PhysRevLett.83.4212

46. Meshulam L, Gauthier JL, Brody CD, Tank DW, Bialek W (2019) Coarse Graining, Fixed Points, and Scaling in a Large Population of Neurons. *Physical Review Letters*, 123(17):178103. https://doi.org/10.1103/PhysRevLett.123.178103

47. Brown JW (2014) Principles of Microbial Diversity.

48. Dykhuizen D (2005) Species Numbers in Bacteria. *Proceedings. California Academy of Sciences*, 56(6 Suppl 1):62–71.

49. Curtis TP, Sloan WT, Scannell JW (2002) Estimating prokaryotic diversity and its limits. *Proceedings of the National Academy of Sciences*, 99(16):10494–10499. https://doi.org/10.1073/pnas.142680199

50. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, Waller A, Mende DR, Kultima JR, Martin J, Kota K, Sunyaev SR, Weinstock GM, Bork P (2013) Genomic variation landscape of the human gut microbiome. *Nature*, 493(7430):45–50.

51. Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, Ding H, Marttinen P, Malmstrom RR, Stocker R, Follows MJ, Stepanauskas R, Chisholm SW (2014) Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild Prochlorococcus. *Science*, 344(6182):416–420. https://doi.org/10.1126/science.1248575

52. Dorp L van, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, Owen CJ, Pang J, Tan CCS, Boshier FAT, Ortiz AT, Balloux F (2020) Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infection, Genetics and Evolution*, 83:104351. https://doi.org/10.1016/j.meegid.2020.104351

53. Maini Rekdal V, Bess EN, Bisanz JE, Turnbaugh PJ, Balskus EP (2019) Discovery and inhibition of an interspecies gut bacterial pathway for Levodopa metabolism. *Science*, 364(6445)https://doi.org/10.1126/science.aau6323

54. Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, Doolabh D, Pillay S, San EJ, Msomi N, Mlisana K, Gottberg A von, Walaza S, Allam M, Ismail A, Mohale T, Glass AJ, Engelbrecht S, Van Zyl G, Preiser W, Petruccione F, Sigal A, Hardie D, Marais G, Hsiao N, Korsman S, Davies M-A, Tyers L, Mudau I, York D, Maslo C, Goedhals D, Abrahams S, Laguda-Akingba O, Alisoltani-Dehkordi A, Godzik A, Wibmer CK, Sewell BT, Lourenço J, Alcantara LCJ, Kosakovsky Pond SL, Weaver S, Martin D, Lessells RJ, Bhiman JN, Williamson C, Oliveira T de (2021) Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature*, 592(7854):438–443. https://doi.org/10.1038/s41586-021-03402-9

55. Arnold BJ, Huang I-T, Hanage WP (2021) Horizontal gene transfer and adaptive evolution in bacteria. *Nature Reviews Microbiology*, :1–13. https://doi.org/10.1038/s41579-021-00650-4

56. Eyre-Walker A, Keightley PD (2007) The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, 8(8):610–618. https://doi.org/10.1038/nrg2146

57. Levy SF, Blundell JR, Venkataram S, Petrov DA, Fisher DS, Sherlock G (2015) Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature*, 519(7542):181–186.

58. Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Walters GB, Jonasdottir A, Gylfason A, Kristinsson KT, Gudjonsson SA, Frigge ML, Helgason A, Thorsteinsdottir U, Stefansson K (2010) Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 467(7319):1099–1103. https://doi.org/10.1038/nature09525

59. Yahara K, Didelot X, Ansari MA, Sheppard SK, Falush D (2014) Efficient inference of recombination hot regions in bacterial genomes. *Molecular Biology and Evolution*, 31(6):1593–1605. https://doi.org/10.1093/molbev/msu082

60. Lin M, Kussell E (2019) Inferring bacterial recombination rates from large-scale sequencing datasets. *Nature Methods*, 16(2):199–204. https://doi.org/10.1038/s41592-018-0293-7

61. Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. *Molecular Biology and Evolution*, 22(5):1185–1192. https://doi.org/10.1093/molbev/msi103

62. Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. *Nature*, 475:493–496.

63. Kimura M (1983) The neutral theory of molecular evolution.

64. Wakeley J (2009) Coalescent theory: an introduction.

65. Spence JP, Steinrücken M, Terhorst J, Song YS (2018) Inference of population history using coalescent HMMs: review and outlook. *Current Opinion in Genetics and Development*, 53:70–76.

66. Schiffels S, Durbin R (2014) Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46(8):919–925. https://doi.org/10.1038/ng.3015

67. Barroso GV, Puzović N, Dutheil JY (2019) Inference of recombination maps from a single pair of genomes and its application to ancient samples. *PLOS Genetics*, 15(11):e1008449. https://doi.org/10.1371/journal.pgen.1008449

68. DeWitt WS, Harris KD, Ragsdale AP, Harris K (2021) Nonparametric coalescent inference of mutation spectrum history and demography. *Proceedings of the National Academy of Sciences*, 118(21)https://doi.org/10.1073/pnas.2013798118

69. Tsimring LS, Levine H, Kessler DA (1996) RNA virus evolution via a fitness-space model. *Physical review letters*, 76(23):4440.

70. Rouzine IM, Wakeley J, Coffin JM (2003) The solitary wave of asexual evolution. *Proceedings of the National Academy of Sciences*, 100(2):587–592.

71. Desai MM, Fisher DS (2007) Beneficial mutation–selection balance and the effect of linkage on positive selection. *Genetics*, 176(3):1759–1798.

72. Park S-C, Simon D, Krug J (2010) The Speed of Evolution in Large Asexual Populations. *Journal of Statistical Physics*, 138(1):381–410. https://doi.org/10.1007/s10955-009-9915-x

73. Hallatschek O (2011) The noisy edge of traveling waves. *Proceedings of the National Academy of Sciences*, 108(5):1783–1787. https://doi.org/10.1073/pnas.1013529108

74. Schiffels S, Szöllősi GJ, Mustonen V, Lässig M (2011) Emergent Neutrality in Adaptive Asexual Evolution. *Genetics*, 189(4):1361–1375. https://doi.org/10.1534/genetics.111.132027

75. Good BH, Rouzine IM, Balick DJ, Hallatschek O, Desai MM (2012) Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. *Proceedings of the National Academy of Sciences*, 109(13):4950–4955.

76. Neher RA, Hallatschek O (2013) Genealogies of rapidly adapting populations. *Proceedings of the National Academy of Sciences*, 110(2):437–442. https://doi.org/10.1073/pnas.1213113110

77. Desai MM, Walczak AM, Fisher DS (2013) Genetic Diversity and the Structure of Genealogies in Rapidly Adapting Populations. *Genetics*, 193(2):565–585. https://doi.org/10.1534/genetics.112.147157

78. Walczak AM, Nicolaisen LE, Plotkin JB, Desai MM (2012) The Structure of Genealogies in the Presence of Purifying Selection: A Fitness-Class Coalescent. *Genetics*, 190(2):753–779. https://doi.org/10.1534/genetics.111.134544

79. Cvijović I, Good BH, Desai MM (2018) The effect of strong purifying selection on genetic diversity. *Genetics*, 209(4):1235–1278.

80. Melissa MJ, Good BH, Fisher DS, Desai MM (2021) Population genetics of polymorphism and divergence in rapidly evolving populations. :2021.06.28.450258. https://doi.org/10.1101/2021.06.28.450258

81. Kolmogorov A, Petrovsky I, Piskunov N (1937) Investigation of the equation of diffusion combined with increasing of the substance and its application to a biology problem. *Bull. Moscow State Univ. Ser. A: Math. Mech*, 1(6):1–25.

82. Fisher RA (1937) The wave of advance of advantageous genes. *Annals of eugenics*, 7(4):355–369.

83. Brunet E, Derrida B, Mueller A, Munier S (2006) Phenomenological theory giving the full statistics of the position of fluctuating pulled fronts. *Physical Review E*, 73(5):056126.

84. Hallatschek O, Nelson DR (2008) Gene surfing in expanding populations. *Theoretical population biology*, 73(1):158–170.

85. Frenkel EM, Good BH, Desai MM (2014) The fates of mutant lineages and the distribution of fitness effects of beneficial mutations in laboratory budding yeast populations. *Genetics*, 196(4):1217–1226.

86. Good BH, McDonald MJ, Barrick JE, Lenski RE, Desai MM (2017) The dynamics of molecular evolution over 60,000 generations. *Nature*, 551(7678):45–50.

87. Nguyen Ba AN, Cvijović I, Rojas Echenique JoséI, Lawrence KR, Rego-Costa A, Liu X, Levy SF, Desai MM (2019) High-resolution lineage tracking reveals travelling wave of adaptation in laboratory yeast. *Nature*, 575(7783):494–499.

88. McFarland CD, Mirny LA, Korolev KS (2014) Tug-of-war between driver and passenger mutations in cancer and other adaptive processes. *Proceedings of the National Academy of Sciences*, 111(42):15138–15143. https://doi.org/10.1073/pnas.1404341111

89. Tilk S, Curtis C, Petrov DA, McFarland CD (2021) Most cancers carry a substantial deleterious load due to Hill-Robertson interference. :764340. https://doi.org/10.1101/764340

90. Strelkowa N, Lässig M (2012) Clonal interference in the evolution of influenza. *Genetics*, 192(2):671–682. https://doi.org/10.1534/genetics.112.143396

91. Neher RA, Russell CA, Shraiman BI (2014) Predicting evolution from the shape of genealogical trees. *eLife*, 3:e03568. https://doi.org/10.7554/eLife.03568

92. Hanage WP (2016) Not So Simple After All: Bacteria, Their Population Genetics, and Recombination. *Cold Spring Harbor Perspectives in Biology*, 8(7):a018069. https://doi.org/10.1101/cshperspect.a018069

93. Thomas CM, Nielsen KM (2005) Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria. *Nature Reviews Microbiology*, 3(9):711–721. https://doi.org/10.1038/nrmicro1234

94. Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ (2011) Ecology drives a global network of gene exchange connecting the human microbiome. *Nature*, 480(7376):241–244. https://doi.org/10.1038/nature10571

95. Kent AG, Vill AC, Shi Q, Satlin MJ, Brito IL (2020) Widespread transfer of mobile antibiotic resistance genes within individual gut microbiomes revealed through bacterial Hi-C. *Nature Communications*, 11(1):4379. https://doi.org/10.1038/s41467-020-18164-7

96. Groussin M, Poyet M, Sistiaga A, Kearney SM, Moniz K, Noel M, Hooker J, Gibbons SM, Segurel L, Froment A, Mohamed RS, Fezeu A, Juimo VA, Lafosse S, Tabe FE, Girard C, Iqaluk D, Nguyen LTT, Shapiro BJ, Lehtimäki J, Ruokolainen L, Kettunen PP, Vatanen T, Sigwazi S, Mabulla A, Domínguez-Rodrigo M, Nartey YA, Agyei-Nkansah A, Duah A, Awuku YA, Valles KA, Asibey SO, Afihene MY, Roberts LR, Plymoth A, Onyekwere CA, Summons RE, Xavier RJ, Alm EJ (2021) Elevated rates of horizontal gene transfer in the industrialized human microbiome. *Cell*, 184(8):2053-2067.e18. https://doi.org/10.1016/j.cell.2021.02.052

97. Fraser C, Hanage WP, Spratt BG (2007) Recombination and the nature of bacterial speciation. *Science (New York, N.Y.)*, 315(5811):476–480. https://doi.org/10.1126/science.1127573

98. Bobay L-M, Ochman H (2017) Biological Species Are Universal across Life's Domains. *Genome Biology and Evolution*, 9(3):491–501. https://doi.org/10.1093/gbe/evx026

99. Arevalo P, VanInsberghe D, Elsherbini J, Gore J, Polz MF (2019) A Reverse Ecology Approach Based on a Biological Definition of Microbial Populations. *Cell*, 178(4):820-834.e14. https://doi.org/10.1016/j.cell.2019.06.033

100. Olm MR, Crits-Christoph A, Diamond S, Lavy A, Carnevali PBM, Banfield JF (2020) Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries. *mSystems*, 5(1)

101. Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabó G, Polz MF, Alm EJ (2012) Population Genomics of Early Events in the Ecological Differentiation of Bacteria. *Science*, 336(6077):48–51. https://doi.org/10.1126/science.1218198

102. Neher RA, Shraiman BI (2009) Competition between recombination and epistasis can cause a transition from allele to genotype selection. *Proceedings of the National Academy of Sciences*, 106(16):6866–6871. https://doi.org/10.1073/pnas.0812560106

103. Niehus R, Mitri S, Fletcher AG, Foster KR (2015) Migration and horizontal gene transfer divide microbial genomes into multiple niches. *Nature Communications*, 6(1):8924. https://doi.org/10.1038/ncomms9924

104. Bendall ML, Stevens SL, Chan L-K, Malfatti S, Schwientek P, Tremblay J, Schackwitz W, Martin J, Pati A, Bushnell B, Froula J, Kang D, Tringe SG, Bertilsson S, Moran MA, Shade A, Newton RJ, McMahon KD, Malmstrom RR (2016) Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *The ISME Journal*, 10(7):1589–1601. https://doi.org/10.1038/ismej.2015.241

105. Smith JM, Dowson CG, Spratt BG (1991) Localized sex in bacteria. *Nature*, 349(6304):29–31. https://doi.org/10.1038/349029a0

106. Smith JM, Smith NH, O'Rourke M, Spratt BG (1993) How clonal are bacteria? *Proceedings of the National Academy of Sciences*, 90(10):4384–4388. https://doi.org/10.1073/pnas.90.10.4384

107. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43. https://doi.org/10.1038/nature02340

108. Falush D, Kraft C, Taylor NS, Correa P, Fox JG, Achtman M, Suerbaum S (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(26):15056–15061. https://doi.org/10.1073/pnas.251396098

109. Hanage WP, Fraser C, Spratt BG (2006) The impact of homologous recombination on the generation of diversity in bacteria. *Journal of Theoretical Biology*, 239(2):210–219. https://doi.org/10.1016/j.jtbi.2005.08.035

110. Vos M, Didelot X (2009) A comparison of homologous recombination rates in bacteria and archaea. *The ISME Journal*, 3(2):199–208. https://doi.org/10.1038/ismej.2008.93

111. Pääbo S (2003) The mosaic that is our genome. *Nature*, 421(6921):409–412. https://doi.org/10.1038/nature01400

112. Fraser C, Hanage WP, Spratt BG (2005) Neutral microepidemic evolution of bacterial pathogens. *Proceedings of the National Academy of Sciences*, 102(6):1968–1973. https://doi.org/10.1073/pnas.0406993102

113. Didelot X, Falush D (2007) Inference of Bacterial Microevolution Using Multilocus Sequence Data. *Genetics*, 175(3):1251–1266. https://doi.org/10.1534/genetics.106.063305

114. Didelot X, Wilson DJ (2015) ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLOS Computational Biology*, 11(2):e1004041. https://doi.org/10.1371/journal.pcbi.1004041

115. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR (2015) Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research*, 43(3):e15. https://doi.org/10.1093/nar/gku1196

116. Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, Corander J (2012) Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Research*, 40(1):e6. https://doi.org/10.1093/nar/gkr928

117. Ansari MA, Didelot X (2014) Inference of the Properties of the Recombination Process from Whole Bacterial Genomes. *Genetics*, 196(1):253–265. https://doi.org/10.1534/genetics.113.157172

118. Dixit PD, Pang TY, Studier FW, Maslov S (2015) Recombinant transfer in the basic genome of Escherichia coli. *Proceedings of the National Academy of Sciences*, 112(29):9070–9075. https://doi.org/10.1073/pnas.1510839112

119. Cui Y, Yang X, Didelot X, Guo C, Li D, Yan Y, Zhang Y, Yuan Y, Yang H, Wang J, Wang J, Song Y, Zhou D, Falush D, Yang R (2015) Epidemic Clones, Oceanic Gene Pools, and Eco-LD in the Free Living Marine Pathogen Vibrio parahaemolyticus. *Molecular Biology and Evolution*, 32(6):1396–1410. https://doi.org/10.1093/molbev/msv009

120. Berube PM, Biller SJ, Hackl T, Hogle SL, Satinsky BM, Becker JW, Braakman R, Collins SB, Kelly L, Berta-Thompson J, Coe A, Bergauer K, Bouman HA, Browning TJ, De Corte D, Hassler C, Hulata Y, Jacquot JE, Maas EW, Reinthaler T, Sintes E, Yokokawa T, Lindell D, Stepanauskas R, Chisholm SW (2018) Single cell genomes of Prochlorococcus, Synechococcus, and sympatric microbes from diverse marine environments. *Scientific Data*, 5(1):180154. https://doi.org/10.1038/sdata.2018.154

121. Petit RA, Read TD (2018) Staphylococcus aureus viewed from the perspective of 40,000+ genomes. *PeerJ*, 6:e5261. https://doi.org/10.7717/peerj.5261

122. Frentrup M, Zhou Z, Steglich M, Meier-Kolthoff JP, Göker M, Riedel T, Bunk B, Spröer C, Overmann J, Blaschitz M, Indra A, Müller L von, Kohl TA, Niemann S, Seyboldt C, Klawonn F, Kumar N, Lawley TD, García-Fernández S, Cantón R, Campo R del, Zimmermann O, Groß U, Achtman M, Nübel U 2020 A publicly accessible database for Clostridioides difficile genome sequences supports tracing of transmission chains and epidemics. *Microbial Genomics*, 6(8):e000410. https://doi.org/10.1099/mgen.0.000410

123. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, Pollard KS, Sakharova E, Parks DH, Hugenholtz P, Segata N, Kyrpides NC, Finn RD (2021) A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology*, 39(1):105–114. https://doi.org/10.1038/s41587-020-0603-3

124. Million Microbiomes from Humans Project. https://db.cngb.org/mmhp/

125. Myers S, Fefferman C, Patterson N (2008) Can one learn history from the allelic spectrum. *Theoretical population biology*, 73(3):342–348. https://doi.org/10.1016/j.tpb.2008.01.001

126. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS Genetics*, 5(10):e1000695. https://doi.org/10.1371/journal.pgen.1000695

127. Lawrie DS, Petrov DA (2014) Comparative population genomics: power and principles for the inference of functionality. *Trends in Genetics*, 30(4):133–139. https://doi.org/10.1016/j.tig.2014.02.002

128. Bhaskar A, Wang YXR, Song YS (2015) Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Research*, :gr.178756.114. https://doi.org/10.1101/gr.178756.114

129. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J (2017) 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*, 101(1):5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

130. Power RA, Parkhill J, Oliveira T de (2017) Microbial genome-wide association studies: lessons from human GWAS. *Nature Reviews Genetics*, 18(1):41–50. https://doi.org/10.1038/nrg.2016.132

131. Sawyer SA, Hartl DL (1992) Population genetics of polymorphism and divergence. *Genetics*, 132(4):1161–1176.

132. Griffiths RC, Tavaré S (1998) The age of a mutation in a general coalescent tree. *Communications in Statistics. Stochastic Models*, 14(1–2):273–295. https://doi.org/10.1080/15326349808807471

133. Polanski A, Kimmel M (2003) New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics*, 165(1):427–436.

134. Coop G, Ralph P (2012) Patterns of Neutral Diversity Under General Models of Selective Sweeps. *Genetics*, 192(1):205–224. https://doi.org/10.1534/genetics.112.141861

135. Ohta T, Kimura M (1971) Linkage Disequilibrium between Two Segregating Nucleotide Sites under the Steady Flux of Mutations in a Finite Population. *Genetics*, 68(4):571–580.

136. McVean G (2007) The structure of linkage disequilibrium around a selective sweep. *Genetics*, 175(3):1395–1406. https://doi.org/10.1534/genetics.106.062828

137. Sohail M, Vakhrusheva OA, Sul JH, Pulit SL, Francioli LC, GENOME OF THE NETHERLANDS CONSORTIUM, ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE, Berg LH van den, Veldink JH, Bakker PIW de, Bazykin GA, Kondrashov AS, Sunyaev SR (2017) Negative selection in humans and fruit flies involves synergistic epistasis. *Science*, 356(6337):539–542. https://doi.org/10.1126/science.aah5238

138. Rosen MJ, Davison M, Fisher DS, Bhaya D (2018) Probing the ecological and evolutionary history of a thermophilic cyanobacterial population via statistical properties of its microdiversity. *PloS one*, 13(11)

139. Arnold B, Sohail M, Wadsworth C, Corander J, Hanage WP, Sunyaev S, Grad YH (2020) Fine-Scale Haplotype Structure Reveals Strong Signatures of Positive Selection in a Recombining Bacterial Pathogen. *Molecular Biology and Evolution*, 37(2):417–428. https://doi.org/10.1093/molbev/msz225

140. Garcia JA, Lohmueller KE (2021) Negative linkage disequilibrium between amino acid changing variants reveals interference among deleterious mutations in the human genome. *PLOS Genetics*, 17(7):e1009676. https://doi.org/10.1371/journal.pgen.1009676

141. Sandler G, Wright SI, Agrawal AF (2021) Patterns and Causes of Signed Linkage Disequilibria in Flies and Plants. *Molecular Biology and Evolution*, 38(10):4310–4321. https://doi.org/10.1093/molbev/msab169

142. Ragsdale AP (2021) Can we distinguish modes of selective interactions using linkage disequilibrium? :2021.03.25.437004. https://doi.org/10.1101/2021.03.25.437004

143. Ghosh OM, Good BH (2021) Emergent evolutionary forces in spatial models of luminal growth in the human gut microbiota. :2021.07.15.452569. https://doi.org/10.1101/2021.07.15.452569

144. Fisher DS (2007) Course 11 Evolutionary dynamics. *Les Houches*, 85:395–446. https://doi.org/10.1016/S0924-8099(07)80018-7

145. Weissman DB, Desai MM, Fisher DS, Feldman MW (2009) The rate at which asexual populations cross fitness valleys. *Theoretical Population Biology*, 75(4):286–300. https://doi.org/10.1016/j.tpb.2009.02.006

146. Weissman DB, Feldman MW, Fisher DS (2010) The Rate of Fitness-Valley Crossing in Sexual Populations. *Genetics*, 186(4):1389–1410. https://doi.org/10.1534/genetics.110.123240

147. McVean GAT (2002) A genealogical interpretation of linkage disequilibrium. *Genetics*, 162(2):987–991. https://doi.org/10.1093/genetics/162.2.987

148. Song YS, Song JS (2007) Analytic computation of the expectation of the linkage disequilibrium coefficient r2. *Theoretical Population Biology*, 71(1):49–60. https://doi.org/10.1016/j.tpb.2006.09.001

149. Good BH (2020) Linkage disequilibrium between rare mutations. :2020.12.10.420042. https://doi.org/10.1101/2020.12.10.420042

150. Harris K, Nielsen R (2014) Error-prone polymerase activity causes multinucleotide mutations in humans. *Genome Research*, 24(9):1445–1454. https://doi.org/10.1101/gr.170696.113

151. Hudson RR, Kaplan NL (1985) Statistical Properties of the Number of Recombination Events in the History of a Sample of Dna Sequences. *Genetics*, 111(1):147–164.

152. Beichman AC, Huerta-Sanchez E, Lohmueller KE (2018) Using genomic data to infer historic population dynamics of nonmodel organisms. *Annual Review of Ecology, Evolution, and Systematics*, 49:433–456.

153. Schraiber JG, Akey JM (2015) Methods and models for unravelling human evolutionary history. *Nature Reviews. Genetics*, 16(12):727–740. https://doi.org/10.1038/nrg4005

154. Wiedenbeck J, Cohan FM (2011) Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS microbiology reviews*, 35(5):957–976. https://doi.org/10.1111/j.1574-6976.2011.00292.x

155. Rosen Z, Bhaskar A, Roch S, Song YS (2018) Geometry of the Sample Frequency Spectrum and the Perils of Demographic Inference. *Genetics*, 210(2):665–682. https://doi.org/10.1534/genetics.118.300733

156. Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M (2012) Drift-barrier hypothesis and mutation-rate evolution. *Proceedings of the National Academy of Sciences*, 109(45):18488–18492. https://doi.org/10.1073/pnas.1216223109

157. Wahl LM, Gerrish PJ, Saika-Voivod I (2002) Evaluating the impact of population bottlenecks in experimental evolution. *Genetics*, 162(2):961–971. https://doi.org/10.1093/genetics/162.2.961

158. Messer PW, Petrov DA (2013) Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology & Evolution*, 28(11):659–669. https://doi.org/10.1016/j.tree.2013.08.003

159. Torres R, Stetter MG, Hernandez RD, Ross-Ibarra J (2020) The Temporal Dynamics of Background Selection in Nonequilibrium Populations. *Genetics*, 214(4):1019–1030. https://doi.org/10.1534/genetics.119.302892

160. Otto SP, Whitlock MC (1997) The Probability of Fixation in Populations of Changing Size. *Genetics*, 146(2):723–733.

161. Schraiber JG (2014) A path integral formulation of the Wright–Fisher process with genic selection. *Theoretical Population Biology*, 92:30–35. https://doi.org/10.1016/j.tpb.2013.11.002

162. Slatkin M (2000) Allele age and a test for selection on rare alleles. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 355(1403):1663–1668.

163. Serohijos AWR, Shakhnovich EI (2014) Merging molecular mechanism and evolution: theory and computation at the interface of biophysics and evolutionary population genetics. *Current Opinion in Structural Biology*, 26:84–91. https://doi.org/10.1016/j.sbi.2014.05.005

164. Lynch M, Ackerman MS, Gout J-F, Long H, Sung W, Thomas WK, Foster PL (2016) Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, 17(11):704–714. https://doi.org/10.1038/nrg.2016.104

165. Good BH, Hallatschek O (2018) Effective models and the search for quantitative principles in microbial evolution. *Current opinion in microbiology*, 45:203–212.

166. Wiuf C, Hein J (1999) Recombination as a Point Process along Sequences. *Theoretical Population Biology*, 55(3):248–259. https://doi.org/10.1006/tpbi.1998.1403

167. McVean GAT, Cardin NJ (2005) Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1387–1393. https://doi.org/10.1098/rstb.2005.1673

168. De Maio N, Wilson DJ (2017) The Bacterial Sequential Markov Coalescent. *Genetics*, 206(1):333–343. https://doi.org/10.1534/genetics.116.198796

169. Zahrt TC, Maloy S (1997) Barriers to recombination between closely related bacteria: MutS and RecBCD inhibit recombination between Salmonella typhimurium and Salmonella typhi. *Proceedings of the National Academy of Sciences*, 94(18):9786–9791. https://doi.org/10.1073/pnas.94.18.9786

170. Durbin R (1998) Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.

171. Brito IL (2021) Examining horizontal gene transfer in microbial communities. *Nature Reviews Microbiology*, 19(7):442–453. https://doi.org/10.1038/s41579-021-00534-7

172. Smith JM (1991) The population genetics of bacteria. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 245(1312):37–41. https://doi.org/10.1098/rspb.1991.0085

173. Majewski J, Cohan FM (1999) Adapt Globally, Act Locally: The Effect of Selective Sweeps on Bacterial Sequence Diversity. *Genetics*, 152(4):1459–1474.

174. Hallatschek O (2018) Selection-Like Biases Emerge in Population Models with Recurrent Jackpot Events. *Genetics*, 210(3):1053–1073. https://doi.org/10.1534/genetics.118.301516

175. National Research Council, Division on Earth and Life Studies, Board on Life Sciences, Committee on a New Biology for the 21st Century: Ensuring the United States Leads the Coming Biology Revolution (2009) A New Biology for the 21st Century.

176. Bialek W, Botstein D (2004) Introductory science and mathematics education for 21st-Century biologists. *Science (New York, N.Y.)*, 303(5659):788–790. https://doi.org/10.1126/science.1095480

177. Phillips R, Kondev J, Theriot J, Garcia H, Kondev J (2012) Physical Biology of the Cell.

178. Crouch CH, Heller K (2014) Introductory physics in biological context: An approach to improve introductory physics for life science students. *American Journal of Physics*, 82(5):378–386. https://doi.org/10.1119/1.4870079

179. O'Leary ES, Sayson HW, Shapiro C, Garfinkel A, Conley WJ, Levis-Fitzgerald M, Eagan MK, Van Valkenburgh B (2021) Reimagining the Introductory Math Curriculum for Life Sciences Students. *CBE—Life Sciences Education*, 20(4):ar62. https://doi.org/10.1187/cbe.20-11-0252

180. Hartl DL, Clark AG (2006) Principles of Population Genetics.

181. Durrett R (2002) Probability Models for DNA Sequence Evolution.

182. Zeevi D, Korem T, Godneva A, Bar N, Kurilshikov A, Lotan-Pompan M, Weinberger A, Fu J, Wijmenga C, Zhernakova A, Segal E (2019) Structural variation in the gut microbiome associates with host health. *Nature*, 568(7750):43–48. https://doi.org/10.1038/s41586-019-1065-y

183. Zhao S, Lieberman TD, Poyet M, Kauffman KM, Gibbons SM, Groussin M, Xavier RJ, Alm EJ (2019) Adaptive evolution within gut microbiomes of healthy people. *Cell host & microbe*, 25(5):656–667.

184. Vatanen T, Plichta DR, Somani J, Münch PC, Arthur TD, Hall AB, Rudolf S, Oakeley EJ, Ke X, Young RA, Haiser HJ, Kolde R, Yassour M, Luopajärvi K, Siljander H, Virtanen SM, Ilonen J, Uibo R, Tillmann V, Mokurov S, Dorshakova N, Porter JA, McHardy AC, Lähdesmäki H, Vlamakis H, Huttenhower C, Knip M, Xavier RJ (2019) Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. *Nature Microbiology*, 4(3):470–479. https://doi.org/10.1038/s41564-018-0321-5

185. Kurt F, Leventhal GE, Spalinger MR, Anthamatten L, Bieberstein PR von, Rogler G, Lacroix C, Wouters T de (2021) Co-cultivation is a powerful approach to produce a robust functionally designed synthetic consortium as a live biotherapeutic product (LBP). :2021.10.13.464188. https://doi.org/10.1101/2021.10.13.464188

186. Backman JD, Li AH, Marcketta A, Sun D, Mbatchou J, Kessler MD, Benner C, Liu D, Locke AE, Balasubramanian S, Yadav A, Banerjee N, Gillies CE, Damask A, Liu S, Bai X, Hawes A, Maxwell E, Gurski L, Watanabe K, Kosmicki JA, Rajagopal V, Mighty J, Jones M, Mitnaul L, Stahl E, Coppola G, Jorgenson E, Habegger L, Salerno WJ, Shuldiner AR, Lotta LA, Overton JD, Cantor MN, Reid JG, Yancopoulos G, Kang HM, Marchini J, Baras A, Abecasis GR, Ferreira MAR (2021) Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature*, 599(7886):628–634. https://doi.org/10.1038/s41586-021-04103-z

187. Bolton KL, Ptashkin RN, Gao T, Braunstein L, Devlin SM, Kelly D, Patel M, Berthon A, Syed A, Yabe M, Coombs CC, Caltabellotta NM, Walsh M, Offit K, Stadler Z, Mandelker D, Schulman J, Patel A, Philip J, Bernard E, Gundem G, Ossa JEA, Levine M, Martinez JSM, Farnoud N, Glodzik D, Li S, Robson ME, Lee C, Pharoah PDP, Stopsack KH, Spitzer B, Mantha S, Fagin J, Boucai L, Gibson CJ, Ebert BL, Young AL, Druley T, Takahashi K, Gillis N, Ball M, Padron E, Hyman DM, Baselga J, Norton L, Gardos S, Klimek VM, Scher H, Bajorin D, Paraiso E, Benayed R, Arcila ME, Ladanyi M, Solit DB, Berger MF, Tallman M, Garcia-Closas M, Chatterjee N, Diaz LA, Levine RL, Morton LM, Zehir A, Papaemmanuil E (2020) Cancer therapy shapes the fitness landscape of clonal hematopoiesis. *Nature Genetics*, 52(11):1219–1226. https://doi.org/10.1038/s41588-020-00710-0