Population Genetics in the Human Microbiome

Nandita R. Garud[1] and Katherine S. Pollard[2,3,4]
ngarud@ucla.edu, katherine.pollard@gladstone.ucsf.edu

[1]Department of Ecology and Evolutionary Biology, University of California, Los Angeles. [2]Gladstone Institutes. [3]Department of Epidemiology and Biostatistics, University of California, San Francisco. [4]Chan Zuckerberg Biohub

Key words: Population genetics, microbiome, population structure, recombination, adaptation, association studies

Abstract

While the ecological structure and function of the human microbiome have been extensively studied and associated with a number of diseases, the genetic diversity within species of microbes is less well understood. Yet, genetic mutations in the microbiome can confer biomedical traits, such as the ability to extract nutrients from food, metabolize drugs, evade antibiotics, and communicate with the host immune system. The population genetic processes by which these traits evolve are complex, in part due to interacting ecological and evolutionary forces in the microbiome. Advances in metagenomic sequencing, coupled with bioinformatics tools and population genetic models, provide the ability to widely quantify genetic variation in the microbiome and make inferences about how this diversity arises, evolves, and correlates with traits of both microbes and hosts. In this review, we explore the population genetic forces (mutation, recombination, drift, and selection) that shape microbiome genetic diversity within and between hosts, as well as efforts towards predictive models that leverage microbiome genetics.

Highlights

1. Genetic variation in host-associated microbiomes can be assayed in a high throughput manner with a variety of technologies.

2. Many bacterial species recombine extensively, even though they asexually reproduce.

3. Genetic diversity of many species within and across hosts is spatially structured.

4. Evidence for rapid adaptation within hosts is starting to emerge.

5. Modeling efforts are connecting microbiome genetic variants with host phenotypes, highlighting the biomedical importance of genetic variation in the microbiome.

Introduction

The human microbiome is composed of bacteria, archaea, viruses, and microbial eukaryotes living in our bodies. The taxonomic composition of these communities has been extensively studied and is significantly associated with a variety of diseases and traits [1]. However, each species in the microbiome is genetically heterogeneous, comprising individual cells whose genomes contain different mutations [2]. Widespread deployment of sequencing technologies (Box 1) has revealed that most microbiota harbor extensive genetic variation between hosts, within a host over time, and even within a host at a given time [2,3]. As in other species, this variation consists of single nucleotide variants (SNVs) [2], short insertions and deletions (indels) [4], and larger structural variants (SVs) [5], which include duplications, deletions, insertions, inversions, and can generate gene copy number variants (CNVs) [6]. There has been substantial progress towards quantifying genetic diversity in the human microbiome [2,6–8].

By contrast, our knowledge of population genetic processes that shape the human microbiome is nascent. Population genetics is a discipline that makes statistical inferences about the evolutionary events that gave rise to patterns of genetic variation across individuals of the same species. The main processes that determine the fate of a new mutation are *drift*, *selection*, *migration* (or *transmission*), and *recombination* [9]. These processes govern how new traits arise and spread among microbes, such as antibiotic resistance [10], drug side effects [11], pathogenic biofilm formation [12], and responses to diet, sanitation, health status [13]. However, the relative contributions of these forces to microbiome diversity–as well as how they fluctuate over time and across microbiota, genes, and hosts with different phenotypes–remain to be fully understood.

A population genetic view of the microbiome provides the opportunity to discover the genetic contribution to traits within microbial populations and to infer the processes creating and maintaining trait-associated genotypes. In this review, we focus on the population genetics of human-associated species, but note lessons learned from other environments. Much of the data discussed is from observational studies of natural populations, but we also discuss experimental methods (Box 2) and the use of model systems to establish causality and to interpret inferences from observational studies of human-associated communities.

Evolution versus ecology

To what extent do microbiota respond to their environment by *evolutionary* versus *ecological* processes? Ecological measurements here are defined as the presence or abundance of species and strains (e.g., strain replacements or shifts in species composition over time). Evolution, by contrast, refers to genetic modifications that accumulate on the same genetic background via mutation, recombination, selection, and drift. Both processes can produce within-species genetic changes in a microbiome (Figure 1A, Key Figure), but only evolutionary changes produce a succession of changes on the same genetic background over time.

It is commonly believed that evolutionary time scales are longer than ecological time scales [14–17]. If this is true, then ecological processes, such as changes in species and strain composition, could be the dominant force shaping diversity in the microbiome within a host's lifetime. However, short microbial generation times [18], large population sizes [19], and high mutation

rates [20] mean that microbes have the potential to accumulate genetic modifications so rapidly that they could impact simultaneous ecological processes. Indeed, mounting empirical evidence shows that microbes can evolve rapidly in experiments [20–28], marine and soil populations [29–32], and within hosts [33–35]. Thus, both evolutionary and ecological processes likely play an important role in shaping diversity in the microbiome over a hosts' lifetime.

However, it is unclear to what degree microbiomes respond to selective pressures by ecological versus evolutionary processes. Interestingly, the answer may not be either/or: ecological and evolutionary processes can interact and shape each other, especially when evolution is rapid and occurring on similar timescales as ecological processes [14–17]. For example, over time new ecological niches can form in an evolving clonal population [22], and, species sharing an environmental niche have higher rates of horizontal gene transfer (HGT) [30,36]. This review emphasizes evolutionary processes, but we highlight interactions between ecological and evolutionary processes and note that the two may be interdependent in the context of the human microbiome.

Population structure

The concept of a population is central to evolutionary biology [37]. Evolutionary changes are defined as modifications within a reproductively isolated, *genetically cohesive population* [37]. Within this defined group of individuals (i.e., cells or *lineages*), population genetic processes of drift, recombination and selection have the potential to impact all members of the population. For example, an adaptive mutation conferring a beneficial trait that *fixes* within a population may homogenize genetic diversity, thereby contributing to the genetic cohesion of the population [38,39]. Thus, delineation of a population could reveal the evolutionary history and ecological preferences of a species. Simultaneously, quantifying population structure, in which genetic variation is not randomly distributed, reveals physical or molecular barriers to gene flow. Population structure also provides an important control for genotype-phenotype analyses, such as genome-wide association studies (see below), since they can be confounded by cryptic relatedness amongst individuals [40].

Unfortunately, there is no consensus on how to define a population [37], and this is particularly challenging in the human microbiome given the pool of interacting microbes that may be exchanging DNA. Even standard microbial species definitions based on sequence identity of the 16S rRNA gene, panels of protein coding genes [41], or 95% genome-wide average nucleotide identity [42,43] suffer from incorrectly classifying species [44] and lack resolution to detect structure within species. However, fine-scale patterns of genetic diversity within and between hosts can help to identify population structure within species.

Structure within hosts

Assuming the null hypothesis that "everything is everywhere" [45], if the human body is colonized by a random sample of microbes from the global pool, then within-host genetic variation should resemble genetic variation among the broader population of microbes found across hosts. On average, bacterial genomes of the same species from different host stool samples have a difference every 100 base pairs (or, are 10^{-2}/bp diverged) [2,33,34,46]. However,

within-host diversity patterns range from ~10^{-6}/bp to ~10^{-2}/bp [33], suggesting that hosts are not colonized by a random sample. Instead, this wide range of diversity, coupled with peaks in intermediate allele frequencies within hosts [33,46] (Figure 1), reflects that people are rarely colonized by a single strain (with the possible exception of *Bacteroides fragilis* [47]), and instead are *oligo-colonized* [33,34,46–49] by one, two, three or a few detectable strains that are diverged from each other by ~10^{-2}/bp for most species (Figure 1B). Operationally, for the purposes of this review, strains are defined as genetically distinct collections of *lineages* that have not had sufficient time to recombine extensively to form a *genetically cohesive population*.

The oligo-colonization of hosts indicates that there may be some underlying ecological rules governing the host colonization process. For example, Verster et al. [47] showed that the type VI secretion system, which plays a role in interbacterial competition, may explain why hosts are colonized by one dominant strain of *B. fragilis*. As we discuss below, intra-host population structure is important to control for when performing population genetic analyses within hosts because both shifts in strain frequencies and evolutionary modifications can generate allele frequency changes within a species over time inside of a host.

Structure across hosts

Several well-studied host-associated bacteria show evidence for population structure across hosts from different parts of the world (or, *biogeography)*. For example, genetic diversity in *Helicobacter pylori* [50,51] and *Mycobacterium tuberculosis* [52,53] faithfully mirror human migration routes. However, until recently, it was not known how well gut commensal demographic patterns match our own. Recent work leveraging public metagenomic data from around the world showed that some gut commensals such as *Eubacterium rectale* and *Prevotella copri* have geographic structure mirroring that of humans [7,46,54]. However, many other species (e.g., *Bacteroides vulgatus, Bacteroides uniformis)* do not have diversity patterns that correlate with any one geographic region [7,46], even though many of these species co-speciated with primate hosts on longer time scales [55]. In fact, similar strains of many species are found on different continents [33,46], and simultaneously, very divergent strains of the same species are found in the same host [2,33].

Why do some species have geographic structure that mirror our own while others do not? One possibility is that some property of human hosts (e.g. antibiotic usage, travel, diet, and host genetics [56,57] as discussed below) could impose selective pressures on microbiota, resulting in signals of biogeography. However, while a pair of hosts can harbor closely related strains of any given species, this is not consistently true across all species shared by two hosts [33], suggesting that host factors are not the major drivers of microbiome biogeography, or that they exert different selective pressures across microbial species. Instead, a more likely explanation for biogeography patterns are properties of the microbes themselves that contribute to their distributions across hosts via differential transmission, colonization, and competition for niches. Traits that could influence transmission rates, for example, include limited dispersal, ability to survive oxygen [58], and modes of transmission (e.g. vertical transmission versus horizontal transmission). Interestingly, vertical transmissions of strains from mothers to infants [7,48,49,59,60] and within families [61] may not be the reason why some species show biogeography, since over our lifetimes resident strains of the most prevalent bacteria are replaced

[33]. Instead, horizontal transmission of microbes along broader social networks [62] coupled with limited dispersal is a likely mechanism by which signals of biogeography are generated. To test this, detailed analyses of the dispersal abilities of human-associated microbes and local transmission rates within and between communities are needed. Rare genetic variants and unique *haplotypes* are a promising means to track strains for such studies.

Recombination

Recombination in asexual prokaryotes is a process whereby close homologs or distantly related DNA is transferred from a donor to a recipient (also known as *Horizontal Gene Transfer*, or HGT) (Box 3). This unlinks loci from one another, allowing them to evolve independently and also results in the gain and/or loss of genetic material on different genomic backbones.

Quantifying recombination in the human microbiome is challenging. Initial work focused on detecting HGT across isolate genomes [36,63–66], leveraging the fact that transferred genes typically have distinct genomic signatures (Box 3). Within-species recombination events are more difficult to detect since the donor and recipient genomes are similar. However, recombination can be quantified by identifying phylogenetic inconsistencies across loci. For example, assuming each nucleotide mutates at most once in the whole population, if different phylogenetic trees parsimoniously describe the evolutionary history at different nucleotides, then, a recombination event likely occurred. These phylogenetic inconsistences due to recombination can be captured by the *four gamete test* which looks for at least four *haplotypes* harboring all four combinations of alleles at pairs of polymorphic nucleotides [67] (Box 3). The probability of such recombination events occurring increases with genomic distance between polymorphic sites. Thus, the decay in *linkage disequilibrium* (LD) over genomic distance, which quantifies correlations of alleles (nucleotides or gene presence), is also a signal of recombination (Box 3). Application of LD to microbiomes works best with long-range sequences from isolate genomes, single cell genomes, or long-read data [34,63,68–70], but it has also been applied to genotypes inferred from shotgun metagenomes [33,71,72].

It is generally thought that certain genomic regions (*mobile elements*) recombine more than others [73]. But recent analyses of patterns of LD in several human commensal microbiota and environmental microbes suggest that homologous recombination may affect the majority of loci in several species [33,70–72,74]. This highlights that rates and patterns of recombination vary across microbial species and across the genome [70,71,74–77]. Several examples of the range of recombination rates that bacteria can experience come from environmental samples. For instance, *Myxoccocus xanthus* from soil is a highly clonal species, indicating low recombination rates [76,78], whereas Vibrio from the ocean experience high recombination though some loci have high LD indicative of recent selection. Other bacteria have extreme levels of recombination, such as Cyanobacteria from a hot spring where virtually all loci are unlinked [31,79]. While we are starting to get a sense of the range of recombination experienced by human commensals [33,70,71], much work still remains to fully characterize this.

The relationship between recombination rate and genetic diversity is complex [77]. HGT frequently diversifies the gene content, or *pangenome*, within a species [80]. It also lowers divergence between species and can reduce diversity within species if horizontally transferred

DNA overwrites existing variation or is recently derived from a common low diversity source. In this manner, homologous recombination helps to create a genetically cohesive population [39,81]. By contrast, when recombination rates are sufficiently low, genomes can diversify into novel species [81]. For example, *B. vulgatus* and *Bacteroides dorei*, which were once defined as one species, are now considered two species that have recently diverged [82].

The ability to incorporate novel genetic material from the broader community and create new combinations of alleles may be particularly important for rapid adaptation to fluctuating environments in humans. The mode by which this happens is a topic of great interest [32,39]. Assuming an adaptive allele with selection coefficient *s* and a recombination rate *r* per generation per base pair, if $r \gg s$, an adaptive genetic variant could propagate through the population via a *gene-specific sweep*, whereby the trait recombines onto multiple genomic backgrounds or sweeps through the population on a plasmid [30,31,83–86]. Alternatively, if $r \ll s$, then the entire linked genome will rise to high frequency in a *genome-wide sweep*, spreading potentially deleterious alleles along with the beneficial allele [32,87] (Figure 1C). The relative frequencies of these two modes are yet to be fully quantified across human commensals.

Since recombination can facilitate adaptation by incorporating new beneficial genomic material, it is easy to make the mistake of concluding that recombination itself was positively selected. There are costs to recombination too: combinations of beneficial alleles can become unlinked [88], and deleterious material can become incorporated too, as observed with plasmids, which may impose costs despite also playing an important role in bringing in public goods from the broader community [89]. However, via the process of *purifying selection*, deleterious variants are purged from the population, thus making it seem retrospectively that everything incorporated by recombination is beneficial [90]. Thus, it is interesting to note that the original purpose of the enzymes involved in recombination is to repair and replicate DNA and not to facilitate gene transfer [90–92].

With growing interest in recombination in the human microbiome, many open questions remain. Although it is common practice to build phylogenetic trees in order to describe the evolutionary relationship between lineages [46,54,93], branch lengths in trees of highly recombining lineages may reflect the frequency with which lineages recombined with each other in the past rather than divergence time [70,76]. Detecting when this is the case is important for elucidating evolutionary mechanisms and determining if a tree is a useful representation of population history or not. It is also critical to determine how host-associated bacteria have the opportunity to recombine even when hosts are oligo-colonized by just a few strains [33,34,46–49]. Determining why some species like *H. pylori* have high rates of recombination and simultaneously a lot of geographic structure [94] is another focus of ongoing research.

Adaptation

Adaptation is a process whereby a population becomes better able to survive in its environment through changes in allele frequencies. Extensive laboratory experiments, such as the Long Term Experimental Evolution experiment [22,23,25,27], studies of host-associated pathogens [95–99], and natural populations of environmentally-associated microbes [30,31,83,85] have shown that

adaptation in microbial populations is common and rapid. However, until recently it was unknown how broadly this picture applies to the human microbiome.

In theory, commensal microbes in the human microbiome could be adapting rapidly because their large census sizes [100] and short generation times [18] result in a large daily mutational input [34,101]. However, ecological forces, including shifts in the abundance of species in the ecosystem or invasions of better-fit strains, could be more important for response to selective pressures in the human microbiome than in experimental evolution. It is also possible that there is less adaptation because microbiota have co-existed with their hosts for millions of years [55,102] and thus could have already evolved to be optimally adapted to the human body.

A picture of evolution on human-relevant timescales is starting to emerge in the microbiome. To identify evolutionary events, it is important to distinguish ecological scenarios, such as strain fluctuations within a host that could drive allele frequency changes, from true evolutionary changes on the background of a lineage. A few recent studies were able to identify rapid allele frequency changes associated with evolution on 6-month to 2-year time scales in the human gut microbiome either by computationally resolving the lineage structure within a host or sequencing isolates [33–35,103].

However, non-adaptive evolutionary forces can also change allele frequencies. For example, neutral and mildly deleterious alleles may increase under drift [35], potentially during population bottlenecks. But alleles that rise to high frequency rapidly in a sufficiently large population are more consistent with adaptation than neutral scenarios [9]. Still, linked non-adaptive alleles may also *hitchhike* to high frequencies with adaptive alleles [9].

One powerful strategy to identify allele frequency changes associated with adaptation versus neutral processes has been to look for an excess of non-synonymous versus synonymous fixations (*dN/dS*). This is an indication that a functional change may have been positively selected, the reason being that non-synonymous mutations change the amino acids of a protein and synonymous mutations are presumed to be neutral. This approach has been applied to many host-associated microbes, including fixed differences between lineages or strains (e.g., [104,105]) and polymorphism within a host (*pN/pS*) [2].

Genome-wide estimates of *pN/pS* and *dN/dS* are significantly less than 1 in the human microbiome [2,33,34], suggesting that *purifying selection* impacts the majority of the genome over long time scales. However, $dN/dS > 1$ within specific genes, genomic regions, or pathways has identified specific loci in the human microbiome experiencing adaptation. Demonstrating this approach, Zhao and Lieberman *et al.* [34] sequenced hundreds of isolates of *B. fragilis* from hosts over two years and identified 17 genes enriched for nonsynonymous mutations. Further supporting adaptation, these genes involved in cell envelope biosynthesis and polysaccharide utilization were undergoing parallel evolution in multiple hosts. Interestingly, these same loci show signals of purifying selection across hosts, suggesting that different selective forces act on the same loci on different time scales, potentially due to changing selective pressures.

Several other selection detection methods have been applied to microbes, with a focus on isolate genomes. For example, fixation index (F_{ST}) quantifies differences in allele frequencies due to

population structure by comparing allele frequencies in the joint population with allele frequencies in individual populations. F_{ST} can be driven both by non-adaptive population genetic processes like migration and drift, as well differential selection pressures. F_{ST} applied to human metagenomic samples has shown that that over time, samples from the same individual are genetically stable and have much lower F_{ST} than samples from different individuals [2]. Local regions of the genome with high F_{ST} are indicative of positive selection, as observed in soil bacteria [72] and *Plasmodium falciparum* [106,107].

The *dN/dS, pN/pS,* and F_{ST} statistics are powerful for detecting selection that has driven alleles to fixation. However, LD or *haplotype* statistics that account for correlations between pairs or many loci, are useful for detecting more recent selective sweeps that have not *fixed* [108]. LD and haplotype-based statistics have been successfully applied to recombining eukaryotes to detect adaptation [108–110] and could be successful in bacteria with high rates of recombination. For example, Rosen *et al.* showed that although a population of cyanobacteria has low genome-wide LD, individual loci showed elevated LD, consistent with selection and hitchhiking [31].

Although we are gaining evidence that human commensal microbiota can evolve, the tempo, and mode of adaptation in the microbiome remains to be fully characterized [111]. Theoretical work in population genetics has started to uncover multiple ways by which adaptation proceeds in multiple populations [101,112], such as *hard sweeps*, *soft sweeps*, polygenic adaptation, partial sweeps, and more. These different mechanisms of adaptation encode important features of population biology, such as the rate of adaptation, the mode of adaptation (e.g. from de novo mutations or pre-existing genetic variants segregating in the population), temporal fluctuations or spatial distribution in selective pressures, and the strength and timing of selection. These properties of selection are relevant to the study of adaptation in virtually any organism.

Yet our current understanding of the dynamics of positive selection in the microbiome is nascent. For example, it is unknown how commonly adaptation occurs through recombination-seeded adaptive events from distantly related members of the microbiome versus adaptation from *de novo* mutations. It is also unknown how common *partial sweeps* are. Such sweeps may be common (i) if environment and selective pressures fluctuate (e.g. due to medications, diet, urbanization, climate change), (ii) if there is a lot of spatial structure in the microbiome [113,114], or (iii) if lineages evolve different niche specializations [22,34]. Additionally, it is unknown whether *clonal interference* is common in the microbiome, whereby multiple lineages with the same fitness cannot outcompete each other until there is a clear fitness difference or drift takes over, as has been commonly observed in experimental evolution of bacteria [115,116]. Alternatively, it is unknown how common *soft sweeps* are in the microbiome, in which multiple lineages bearing independent instances of the same adaptive mutations rise in frequency simultaneously, as observed in gut microbiota of mice colonized with *E. coli* [117].

Studying adaptation in the human microbiome is an exciting frontier for both the microbiome and population genetics fields. Since population genetics has been typically studied one species at a time, there remains much to learn about adaptation in complex communities and ecology-evolution feedbacks. For example, how commonly does the evolution of a focal species result in permanent shifts in the ecological structure of the community [22]? And, does the ecological structure of the community accelerate or constrain adaptation [14,22,118]? Given the tight

connection to many diseases [119–123], the human microbiome affords us a unique system to analyze complex evolutionary and ecological dynamics and their impact on human health.

Connecting genotypes to phenotypes

Discovering the functional consequences of within species genetic variation in the microbiome is an important goal, both for interpreting the forces driving selective sweeps and for engineering the microbiome to improve human health.

Genetic variants in the microbiome have been causally linked to a range of microbial and host phenotypes including drug resistance [124–127], regulation of biofilm formation [12], conversion of commensals into pathogens [128], modulation of host immune responses [129,130], generation of disease-associated compounds from food [131], and metabolism of drugs [13]. The primary approach used to identify these SNVs and SVs involves performing isolate genome sequencing [132] or metagenomic analysis (e.g., [133]) of strains that harbor phenotypic variation and then using comparative genomics or phylogenetic regression methods to test for genetic variants that correlate with the phenotype. With abundant metagenomic data, we now can perform higher throughput genotype-phenotype associations at an unprecedented scale beyond isolate genomes to link genetic variation within hosts to their health. This opportunity is akin to how human genetics was transformed by genome-wide association studies (GWAS) spurred by the development of genotyping arrays.

As with other forms of GWAS, tests for association between microbiome variants and traits of communities or hosts have limitations. First, various confounding variables need to be accounted for to make accurate inferences. For example, population structure can create false genotype-phenotype associations if genetically distinct lineages are not evenly distributed across phenotype values [40]. To avoid confounding effects of population structure, it is critical to adjust for or explicitly model relatedness of microbes across samples [134–137]. Another confounder of microbiome GWAS is the potential pervasiveness of selection, since linked variants that have hitchhiked to high frequency make it difficult to identify the actual causal variant [138]. Another challenge is the huge sample sizes that may be required to detect statistical associations in the face of technical and biological variation. High-throughput assays for genotyping the microbiome may help make large studies feasible, as in human genetics. But in the near future GWAS will be most useful for variants with big effect sizes (e.g., large SV deleting a pathway). Finally, the utility of GWAS for improving our mechanistic understanding of host-microbe interactions is currently severely limited by how little we know about the functions of most microbiome genes and variants.

GWAS techniques have been applied to microbiome data in a variety of other ways. The plethora of case-control microbiome studies, for example, typically test for associations between microbial abundance and a host trait (e.g. [119,120,139]). Given that presence of some microbial species is a *heritable* trait [140–143], there is also a case for using GWAS to identify links between host genetic variants and microbial abundances [57,144], which most notably has identified an association between variants near the *LCT* locus and the abundance of *Bifidobacteria* [56,141,145,146]. However, it remains to be seen whether this association is due

to milk being available in the diets of individuals that are lactase persistent or driven by genotypes at the *LCT* locus itself. While these approaches associate the taxonomic composition of microbiomes to differences–genetic or phenotypic–between hosts, they do not leverage microbial population genetic variation and its ability to map such associations to genes and functions of microbes. An exciting direction is to explore host-microbe interactions through GWAS using both microbial and host genotypes [147], for example, to test the hypothesis that microbes carrying particular alleles preferentially associate with particular host genetic backgrounds, or to explore how microbial genotypes buffer or amplify human genotype-phenotype connections.

Concluding remarks

The emerging picture of microbiome evolution suggests that each of us harbors populations of microbes whose within-species genetic diversity is limited at any one time point but highly dynamic on time scales of days [34] to a few years [2,3]. These population genetic changes are both an opportunity (e.g., enabling digestion of new foods [148]) and a challenge (e.g., development of drug resistance [127,149] and maintaining stabile colonization of therapeutic strains [150,151]). Understanding how, when, and why microbes evolve is essential to leveraging them in medicine, industry, and agriculture.

To paint a clearer picture of the evolutionary dynamics in microbiome, we need new models that incorporate pervasive recombination, ecology-evolution interactions, and complex relationships between host genetics, highly dynamic microbiome genetics, and the human body's environment. Longitudinal modeling across host lifespan will help resolve open questions about recombination rates and patterns, as well as the mode, tempo, strength, and functional targets of selection.

We also must continue to develop bioinformatics methods for assaying genetic variation from different types of sequencing data, especially as the data becomes more abundant, higher resolution, and more complex. For example, longer reads with low error rates–or in combination with low-error short reads–will help resolve long-range linkage and guide assembly. Methods to estimate allele frequencies from metatranscriptomes will shed light on the functionality of different genotypes.

As microbiome GWAS becomes common, tools for editing microbiome genomes and testing hypotheses through experimental evolution (Box 2) will be essential for establishing causal relationships. Interpreting genetic variation and the results of GWAS also requires a massive effort to improve functional annotation of genes in the human microbiome, many of which are completely uncharacterized. Newly sequenced genomes are particularly unannotated.

Human genetics has progressed by producing a high-quality genome, assaying common variation, developing and deploying genotyping arrays at scale, and sequencing rare and complex variants. Microbiome genetics is just beginning this process ('see Outstanding Questions'), but has the advantage of leveraging existing methods, study designs, and theories. As these are adapted and expanded to study the complex populations of microbes living in our bodies, it will be exciting to

see microbiome population genetics move towards a predictive science and an important component of precision medicine.

Acknowledgements

We gratefully thank Benjamin Good, Kirk Lohmueller, Ami Bhatt, Pleuni Pennings, Sandeep Venkataram, Alison Feder, an anonymous reviewer, and members of the Lohmueller lab for their helpful feedback. Michael Fischbach helped estimate the proportion of human microbiome species that can be genetically engineered. Funding came from NSF (grant DMS-1563159), Gladstone Institutes, and the Chan Zuckerberg Biohub.

Figure 1, Key Figure. Ecological versus evolutionary processes in the microbiome. (A) Microbiota may respond to their environment by ecological shifts in species abundances (top, each species represented by a different colored circle) or invasions of genetically distinct lineages (middle). Alternatively, existing lineages may evolve genetic changes (yellow star) via mutation, recombination, selection, and drift (bottom). (B) An illustration of the genetic diversity of two genetically distinct lineages of one microbiome species colonizing a host at high frequency. In reality, there is a range of colonization scenarios across different hosts and species. On average, lineages from different individuals differ by 1 mutation per 100 base pairs genome-wide (a typical number; yellow stars). These lineages likely accumulated fixed genetic differences for millions of generations prior to colonization. Recombination and selection will have local effects on divergence. In a host that is *oligo-colonized* by these two lineages at equal abundance, this genome-wide divergence manifests as mutations at 50% in a mixed sample. These mutations populate the 0.5 frequency bin in a *site frequency spectrum (SFS)*. On the other hand, mutations that arose since colonizing a host are typically less frequent (e.g., ~1 per 100,000 base pairs) since they have had less time to accumulate. These compose the low-frequency bins of the SFS. Recently diverged lineages will have greater *haplotype homozygosity* than those of more diverged lineages. (C) Signatures of selective sweeps. In both scenarios, the sweeps are seeded by an adaptive recombinant fragment. Top: in a genome-wide sweep, the haplotype bearing the adaptive allele quickly rises to high frequency, resulting in low genome-wide diversity due to *hitchhiking* of linked variants. Bottom: in a gene-specific sweep, recombination events introduce the adaptive fragment onto multiple haplotypes, resulting in low diversity only at the adaptive locus.

Box 1: How do we obtain the necessary genomic information to study population genetics in the microbiome?

Genetic variants in the microbiome are assayed using a variety of approaches that vary on several dimensions including: whether they require culturing, number of species sampled concurrently, length of resulting haplotypes and reads, and cost.

Cultured isolate sequencing. Much of microbiome population genetics relies on identifying single nucleotide variants (SNVs) or structural variants (SVs) amongst sequenced and assembled genomes from cultured isolates. Since variants are linked in assembled genomes, studies of recombination, horizontal gene transfer, and nucleotide changes associated with evolution are feasible [34,35,76,117,127,149,152]. Throughput and difficult-to-culture species are limitations, but targeted culturing efforts [58,153–156], multiplexed sequencing [34], and microfluidics [157] are helping [158].

Metagenotyping. Shotgun metagenomics is the sequencing of pooled DNA from microbial communities. By aligning reads to genomes (or "pangenomes" of all genes observed in a species) and applying statistical models, both SNVs and SVs can be called [2,6,7,46,151,159]. Metagenotyping is relatively unbiased, does not require culturing, and can be used to study hundreds of species simultaneously for relatively low cost. However, alignment to a database misses novel genes and species, is sensitive to reads that align to multiple species or genes [5], and does not provide long-range linkage information, though probabilistic models can be used to computationally phase variants and resolve short-range haplotypes for abundant species [33]. Metagenotyping is distinct from methods that quantify gene or pathway abundance across all species from metagenomes [160] or use genes with co-varying abundance to define metagenomic linkage groups as estimates of species [120,161,162].

Metagenome-assembled-genomes (MAGs). Genomes can be recovered from metagenomes *de novo* using binning and assembly [120,161,162], which allows for characterization of new species and variant discovery in both known and novel species [163–165]. MAGs can rapidly expand the number of reference genomes and observed genetic diversity of a species, as for *Prevotella copri* which recently went from having one to >1,000 genomes from four distinct clades [54,133]. Chimeric and incomplete assemblies are limitations, especially for lower abundance species. Assembly is improving with chromatin capture scaffolding [166] and longer reads. Improved methods for assessing MAG quality and completeness will be helpful.

Single-cell and read-cloud sequencing. Flow cytometry [167] and microfluidics [157,168] can be used to capture and barcode single cells, small numbers of cells, or small numbers of DNA fragments [169] from a sample without culturing. Coupled with deep sequencing and specialized bioinformatics methods, these techniques assemble genome segments from which genetic variants, including mobile genes [63], can be identified. This strategy may soon produce genomes comparable to isolates with high throughput and no culture bias.

Box 2: In vitro experimental tools for microbiome population genetics

Experimental evolution. Observing microbes evolve in the laboratory is a useful method to test hypotheses about microbial population genetics in controlled settings. Longitudinal studies of cultured microbes under different conditions, typically starting from one or a few strains, have shown rapid adaptation from both *de novo* and existing variation. They enabled mutation and homologous recombination rates and patterns to be estimated. Experimental evolution also revealed that population genetic dynamics and targets of selection are often consistent across replicate experiments [20–28,117,170,171].

Reverse genetics. Genetic manipulation of microbes (e.g., gene knock out, mutation, or over-expression) helps to establish causal links between variants and phenotypes. This powerful approach is widely used in several genetically tractable model human microbiome species [172]. Establishing tools to edit the genomes of human microbiome species beyond the ~10% that can be engineered currently is a high priority. For difficult to culture species, future approaches to edit microbial genomes *in situ* within a human host could be transformative.

Strengths and limitations. Knowledge from these experimental systems needs to be tested in the human ecosystem to determine if conclusions about evolutionary mechanisms hold there. Also, the complexity of the human microbiome is difficult to model *in vitro*, though recent experimental evolution studies co-cultured different species to understand evolutionary and ecological interactions in controlled complex communities [173,174]. Experiments clarify what is possible in microbial evolution as well as the mechanisms of evolution, whereas observational data in human hosts sheds light on what actually evolved, but in a context where establishing mechanisms and causality is harder.

Box 3: Mechanisms and Genomic Signatures of Recombination

[Include Figure 2 in this box.]

(A) Bacteria take up DNA by three processes: conjugation (a direct transfer of DNA from one cell to another via cell-to-cell contact), competence (uptake of genetic material from the surrounding environment), and transduction (phage mediated transfer of DNA). Figure adapted from Redfield 2001 [90].

(B) Once inside the cell, the donor's DNA can be incorporated into the recipient's DNA by a variety of modes. Homologous recombination is a process where nearly identical strands of DNA are exchanged by crossovers (indicated with 'X') mediated by the *rec* protein [90]. However, in non-homologous recombination, where only the flanking regions match, or homology facilitated illegitimate recombination, where only one anchor locus matches, new genes may be introduced in the region in between the cross over points [175], resulting in horizontal gene transfer (HGT). HGT can also occur by transfer of plasmids and transposable elements, and it can occur across species boundaries but also within species [90]. Stars indicate mutations; rectangles indicate genes.

(C) HGT across species boundaries can be detected by identifying orthologous genes in multiple species with lower sequence divergence, distinct GC content, tetranucleotide composition, or codon usages [176] compared to the rest of the genome. However, within-species recombination may not leave these same signatures since the recombining genomes may be less diverged. Instead, gene gains and a decay in correlations of allelic states between loci are telltales of recombination. For example, the *four-gamete test* detects recombination by identifying all combinations of a pair of bi-allelic sites, which, assuming an infinite sites model, could not have arisen by multiple *de novo* mutations. *Linkage disequilibrium* captures correlations between sites, which decays over time and genomic distance due to recombination. There are many measures of LD, such as $D_{AB} = p_{AB} - p_A * p_B$, where p_A measures the frequency of allele A occurring at one locus, p_B measures the frequency of allele B occurring at another locus, and p_{AB} measures the frequency of the A and B alleles co-occurring on the same *haplotype*. The rate of decay in LD [33,71], as well as genealogical simulations [75], can be used to infer the ratio r/μ, which describes the relative contribution of recombination (r) versus mutation (μ) towards genetic diversity.

Outstanding Questions Box

1. What factors (e.g., *dispersal limitation*, transmission, population bottlenecks, and recombination rates) contribute to genetic structure within microbiome species across hosts and geographic regions? Why does structure differ so greatly across different microbiome species?
2. What is the tempo and mode of adaptation in the human microbiome within hosts and across hosts on different time scales? What are the forces driving adaptation and the targets of selection? Longitudinal data and more-complete reference genomes with annotations may be important for answering these questions.
3. How common are various mechanisms of recombination in the microbiome (e.g. plasmid transfer versus homologous recombination)? Does variability in recombination rates correlate with rates of adaptation or biogeography across microbiota?
4. Given that evolution in the human microbiome can occur rapidly on short time scales, does this evolution influence ecological processes and vice versa?
5. Which host and microbial traits are associated with genetic variation in the microbiome?
6. How large must sample sizes be for a well-powered microbiome GWAS? Answering this question requires estimating the strength of associations between microbiome genetic variants and phenotypes (effect size), as well as how variable these association are across hosts.
7. Microbiome genomics has largely focused on protein coding genes. Is there evidence of adaptation on mutations in gene regulatory elements and RNA genes?

Glossary

Biogeography – Spatial structure in allele frequencies either along the gut, or, across the earth's surface.
Clonal interference – When two genomes with identical fitness compete with each other, until a clear fitness difference arises, either via a subsequent mutation or recombination event.
Drift – Change in population allele frequencies due to random sampling.
Ecology – Interactions between species and strains and their environment.
Evolution – A change in allele frequency in a population driven by a population genetic force such as mutation, drift, migration, recombination, or selection.
Gene-specific sweep – A selective sweep in which only the gene or immediate locus bearing an adaptive allele rises to high frequency. This is common when $r \gg s$.
Genome-wide sweep – A selective sweep in which the entire genome sweeps to high frequency because it is linked to an adaptive allele. This is common when $r \ll s$.
Dispersal limitation – When the range of migration or dispersal of a species is smaller than the overall global area in which the species resides.
dN/dS – Ratio of nonsynonymous to synonymous fixations between two lineages. dN/dS > 1 indicates positive selection, < 1 indicates purifying selection, and ~1 indicates neutrality.
Fixation – the process by which an allele reaches 100% frequency in a population.
Four-gamete test – Test for recombination in which all four combinations of pairs of alleles at two loci are strong evidence for recombination.
FST – Fixation index measuring the difference in allele frequencies in two populations. FST values close to 1 indicate that the populations are very differentiated, and values close to 0 indicate similar populations.
Genetically cohesive population – a population that lacks barriers to gene flow.
Haplotype – Alleles and genes that are inherited together because they are linked on the same chromosome.
Hard Sweep – A selective sweep in which a single adaptive alleles rises to high frequency, likely because the population is mutation limited.
Heritability – Degree to which the variance in a phenotype can be attributed to genetics versus environmental factors.
Hitchhiking – when an allele changes frequency not because it is under selection, but rather, because it is linked with another allele that changes frequency due to selection.
Homozygosity – Probability of randomly observing identical alleles.
Horizontal Gene Transfer – Transfer of genes from one lineage or species to another, notably not by vertical transfer from parent to offspring.
Lineage – a descendant from a common ancestor. A lineage can harbor as few as just one point mutation differences from the ancestor. In this review, operationally strains are composed of lineages, and species are composed of strains.
Linkage disequilibrium – Non-random association of alleles due to non-neutral forces such as selection. Recombination can, over time, break up correlations of alleles.
Oligo-colonization – Colonization of a host by a small number of genetically distinct lineages.
Pangenome – The union of all genes found in any lineage of a given species.
Partial sweep – A selective sweep that has not reached 100% frequency in the population.
Purifying selection – Removal of deleterious alleles from a population.
Recombination – The exchange of DNA between two genomes.

Selection – A change in frequency of an allele due to fitness effects conferred by the allele.
Selective sweep – Rise in frequency of an adaptive allele and resulting loss in diversity in the vicinity of the selected allele due to hitchhiking.
Site frequency spectrum – Histogram of allele frequencies in a sample. This summary statistic is used in population genetics to infer demographic and selection parameters.
Soft sweep – A selective sweep in which multiple adaptive alleles at the same locus, but on distinct haplotypes, rise to high frequency simultaneously.
Transmission – the transfer of a microorganism from one host to another.
r – Recombination rate, measured in units of number of crossovers per generation per locus.
μ – Mutation rate, measured in units of number of mutations per generation per locus. In Bacteria, mu is estimated to be ~10^{-9} [20].

References

1 Cho, I. and Blaser, M.J. The human microbiome: At the interface of health and disease. , *Nature Reviews Genetics*. (2012)
2 Schloissnig, S. *et al.* (2013) Genomic variation landscape of the human gut microbiome. *Nature* 493, 45–50
3 Faith, J.J. *et al.* (2013) The long-term stability of the human gut microbiota. *Science (80- .).* DOI: 10.1126/science.1237439
4 Chen, J.Q. *et al.* (2009) Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Mol. Biol. Evol.* DOI: 10.1093/molbev/msp063
5 Zeevi, D. *et al.* (2019) Structural variation in the gut microbiome associates with host health. *Nature* DOI: 10.1038/s41586-019-1065-y
6 Greenblum, S. *et al.* (2015) Extensive strain-level copy-number variation across human gut microbiome species. *Cell* 160, 583–594
7 Nayfach, S. *et al.* (2016) An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.* DOI: 10.1101/gr.201863.115
8 Scholz, M. *et al.* (2016) Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat Methods* 13, 435–438
9 Gillespie, J.H. (2004) Population Genetics: A concise guide. *Johns Hopkins Univ. Press*
10 Jernberg, C. *et al.* Long-term impacts of antibiotic exposure on the human intestinal microbiota. , *Microbiology*. (2010)
11 Maurice, C.F. *et al.* (2013) Xenobiotics shape the physiology and gene expression of the active human gut microbiome. *Cell* DOI: 10.1016/j.cell.2012.10.052
12 Lohse, M.B. *et al.* Development and regulation of single-and multi-species Candida albicans biofilms. , *Nature Reviews Microbiology*. (2018)
13 Spanogiannopoulos, P. *et al.* The microbial pharmacists within us: A metagenomic view of xenobiotic metabolism. , *Nature Reviews Microbiology*. (2016)
14 Rudman, S.M. *et al.* What genomic data can reveal about eco-evolutionary dynamics. , *Nature Ecology and Evolution*. (2018)
15 MacColl, A. (2017) Eco-evolutionary Dynamics . By Andrew P. Hendry. Princeton (New Jersey): Princeton University Press. $65.00. xii + 397 p.; ill.; index. ISBN: 978-0-691-14543-3. 2017. . *Q. Rev. Biol.* DOI: 10.1086/693603
16 Hairston, N.G. *et al.* Rapid evolution and the convergence of ecological and evolutionary

time. , *Ecology Letters*. (2005)
17 Thompson, J.N. Rapid evolution as an ecological process. , *Trends in Ecology and Evolution*. (1998)
18 Korem, T. *et al.* (2015) Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science (80-.).* 349, 1101–1106
19 Sender, R. *et al.* (2016) Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol.* DOI: 10.1371/journal.pbio.1002533
20 Barrick, J.E. and Lenski, R.E. Genome dynamics during experimental evolution. , *Nature Reviews Genetics*. (2013)
21 Jerison, E.R. and Desai, M.M. Genomic investigations of evolutionary dynamics and epistasis in microbial evolution experiments. , *Current Opinion in Genetics and Development*. (2015)
22 Good, B.H. *et al.* (2017) The dynamics of molecular evolution over 60,000 generations. *Nature* 551, 45–50
23 Herron, M.D. and Doebeli, M. (2013) Parallel Evolutionary Dynamics of Adaptive Diversification in Escherichia coli. *PLoS Biol.* DOI: 10.1371/journal.pbio.1001490
24 Wiser, M.J. *et al.* (2013) Long-term dynamics of adaptation in asexual populations. *Science (80-.).* DOI: 10.1126/science.1243357
25 Lang, G.I. *et al.* (2013) Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature* DOI: 10.1038/nature12344
26 Sniegowski, P.D. *et al.* (1997) Evolution of high mutation rates in experimental populations of E. coli. *Nature* DOI: 10.1038/42701
27 Tenaillon, O. *et al.* (2016) Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature* DOI: 10.1038/nature18959
28 Lenski, R.E. *et al.* (2003) Rates of DNA sequence evolution in experimental populations of Escherichia coli during 20,000 generations. *J. Mol. Evol.* DOI: 10.1007/s00239-002-2423-0
29 Denef, V.J. and Banfield, J.F. (2012) In situ evolutionary rate measurements show ecological success of recently emerged bacterial hybrids. *Science (80-.).* DOI: 10.1126/science.1218389
30 Shapiro, B.J. *et al.* (2012) Population genomics of early events in the ecological differentiation of bacteria. *Science (80-.).* 336, 48–51
31 Rosen, M.J. *et al.* (2015) Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. *Science (80-.).* DOI: 10.1126/science.aaa4456
32 Bendall, M.L. *et al.* (2016) Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J.* DOI: 10.1038/ismej.2015.241
33 Garud, N.R. *et al.* (2019) Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLoS Biol* 17, e3000102
34 Zhao, S. *et al.* (2019) Adaptive Evolution within Gut Microbiomes of Healthy People. *Cell Host Microbe* DOI: 10.1016/j.chom.2019.03.007
35 Ghalayini, M. *et al.* (2018) Evolution of a dominant natural isolate of Escherichia coli in the human gut over the course of a year suggests a neutral evolution with reduced effective population size. *Appl. Environ. Microbiol.* DOI: 10.1128/AEM.02377-17
36 Smillie, C.S. *et al.* (2011) Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480, 241–244

37 Waples, R.S. and Gaggiotti, O. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. , *Molecular Ecology*. (2006)
38 Arevalo, P. *et al.* (2019) A Reverse Ecology Approach Based on a Biological Definition of Microbial Populations. *Cell* DOI: 10.1016/j.cell.2019.06.033
39 Cohan, F.M. Bacterial speciation: Genetic sweeps in bacterial species. , *Current Biology*. (2016)
40 Pritchard, J.K. and Donnelly, P. (2001) Case-control studies of association in structured or admixed populations. *Theor. Popul. Biol.* DOI: 10.1006/tpbi.2001.1543
41 Wu, D. *et al.* (2013) Systematic Identification of Gene Families for Use as "Markers" for Phylogenetic and Phylogeny-Driven Ecological Studies of Bacteria and Archaea and Their Major Subgroups. *PLoS One* DOI: 10.1371/journal.pone.0077033
42 Konstantinidis, K.T. and Tiedje, J.M. (2005) Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci.* DOI: 10.1073/pnas.0409727102
43 Konstantinidis, K.T. *et al.* (2006) , The bacterial species definition in the genomic era. , in *Philosophical Transactions of the Royal Society B: Biological Sciences*
44 Bobay, L.-M. and Ochman, H. (2017) Biological Species Are Universal across Life's Domains. *Genome Biol. Evol.* DOI: 10.1093/gbe/evx026
45 Baas-Becking, L.G.M. (1934) *Geobiologie of inleiding tot de milieukunde*,
46 Truong, D.T. *et al.* (2017) Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res* 27, 626–638
47 Verster, A.J. *et al.* (2017) The Landscape of Type VI Secretion across Human Gut Microbiomes Reveals Its Role in Community Composition. *Cell Host Microbe* DOI: 10.1016/j.chom.2017.08.010
48 Ferretti, P. *et al.* (2018) Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. *Cell Host Microbe* DOI: 10.1016/j.chom.2018.06.005
49 Yassour, M. *et al.* (2018) Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. *Cell Host Microbe* DOI: 10.1016/j.chom.2018.06.007
50 Linz, B. *et al.* (2007) An African origin for the intimate association between humans and Helicobacter pylori. *Nature* 445, 915–918
51 Falush, D. *et al.* (2003) Traces of human migrations in Helicobacter pylori populations. *Science (80-.).* 299, 1582–1585
52 Pepperell, C.S. *et al.* (2011) Dispersal of Mycobacterium tuberculosis via the Canadian fur trade. *Proc. Natl. Acad. Sci.* DOI: 10.1073/pnas.1016708108
53 Comas, I. *et al.* (2013) Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nat. Genet.* DOI: 10.1038/ng.2744
54 Tett, A. *et al.* (2019) The Prevotella copri complex comprises four distinct clades that are underrepresented in Westernised populations. *bioRxiv* DOI: 10.1101/600593
55 Moeller, A.H. *et al.* (2016) Cospeciation of gut microbiota with hominids. *Science (80-.).* 353, 380–382
56 Blekhman, R. *et al.* (2015) Host genetic variation impacts microbiome composition across human body sites. *Genome Biol* 16, 191
57 Goodrich, J.K. *et al.* (2017) The Relationship Between the Human Genome and Microbiome Comes into View. *Annu. Rev. Genet.* DOI: 10.1146/annurev-genet-110711-

155532
58 Browne, H.P. *et al.* (2016) Culturing of "unculturable" human microbiota reveals novel taxa and extensive sporulation. *Nature* 533, 543–546
59 Milani, C. *et al.* (2015) Exploring vertical transmission of bifidobacteria from mother to child. *Appl. Environ. Microbiol.* DOI: 10.1128/AEM.02037-15
60 Asnicar, F. *et al.* (2017) Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling. *mSystems* DOI: 10.1128/msystems.00164-16
61 Korpela, K. *et al.* (2018) Selective maternal seeding and environment shape the human gut microbiome. *Genome Res* 28, 561–568
62 Brito, I.L. *et al.* Transmission of human-associated microbiota along family and social networks. , *Nature Microbiology*. (2019)
63 Brito, I.L. *et al.* (2016) Mobile genes in the human microbiome are structured from global to individual scales. *Nature* 535, 435–439
64 Coyne, M.J. *et al.* (2014) Evidence of extensive DNA transfer between bacteroidales species within the human gut. *MBio* 5, e01305-14
65 Liu, L. *et al.* (2012) The human microbiome: A hot spot of microbial horizontal gene transfer. *Genomics* DOI: 10.1016/j.ygeno.2012.07.012
66 Lozupone, C.A. *et al.* (2008) The convergence of carbohydrate active gene repertoires in human gut microbes. *Proc. Natl. Acad. Sci.* DOI: 10.1073/pnas.0807339105
67 Hudson, R.R. and Kaplan, N.L. (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111, 147–164
68 Takuno, S. *et al.* (2012) Population genomics in bacteria: A case study of Staphylococcus aureus. *Mol. Biol. Evol.* DOI: 10.1093/molbev/msr249
69 Azim Ansari, M. and Didelot, X. (2014) Inference of the properties of the recombination process from whole bacterial genomes. *Genetics* DOI: 10.1534/genetics.113.157172
70 Sakoparnig, T. *et al.* (2019) Whole genome phylogenies reflect long-tailed distributions of recombination rates in many bacterial species. *bioRxiv* DOI: 10.1101/601914
71 Lin, M. and Kussell, E. (2019) Inferring bacterial recombination rates from large-scale sequencing datasets. *Nat. Methods* DOI: 10.1038/s41592-018-0293-7
72 Crits-Christoph, A. *et al.* (2019) Soil bacterial populations are shaped by recombination and gene-specific selection across a meadow. *bioRxiv* DOI: 10.1101/695478
73 Ghaly, T.M. and Gillings, M.R. Mobile DNAs as Ecologically and Evolutionarily Independent Units of Life. , *Trends in Microbiology*. (2018)
74 González-Torres, P. *et al.* (2019) Impact of Homologous Recombination on the Evolution of Prokaryotic Core Genomes. *MBio* DOI: 10.1128/mBio.02494-18
75 Vos, M. and Didelot, X. (2009) A comparison of homologous recombination rates in bacteria and archaea. *ISME J* 3, 199–208
76 Smith, J.M. *et al.* (1993) How clonal are bacteria? *Proc Natl Acad Sci U S A* 90, 4384–4388
77 Hanage, W.P. (2016) Not so simple after all: Bacteria, their population genetics, and recombination. *Cold Spring Harb. Perspect. Biol.* DOI: 10.1101/cshperspect.a018069
78 Wielgoss, S. *et al.* (2016) A barrier to homologous recombination between sympatric strains of the cooperative soil bacterium Myxococcus xanthus. *ISME J.* DOI: 10.1038/ismej.2016.34

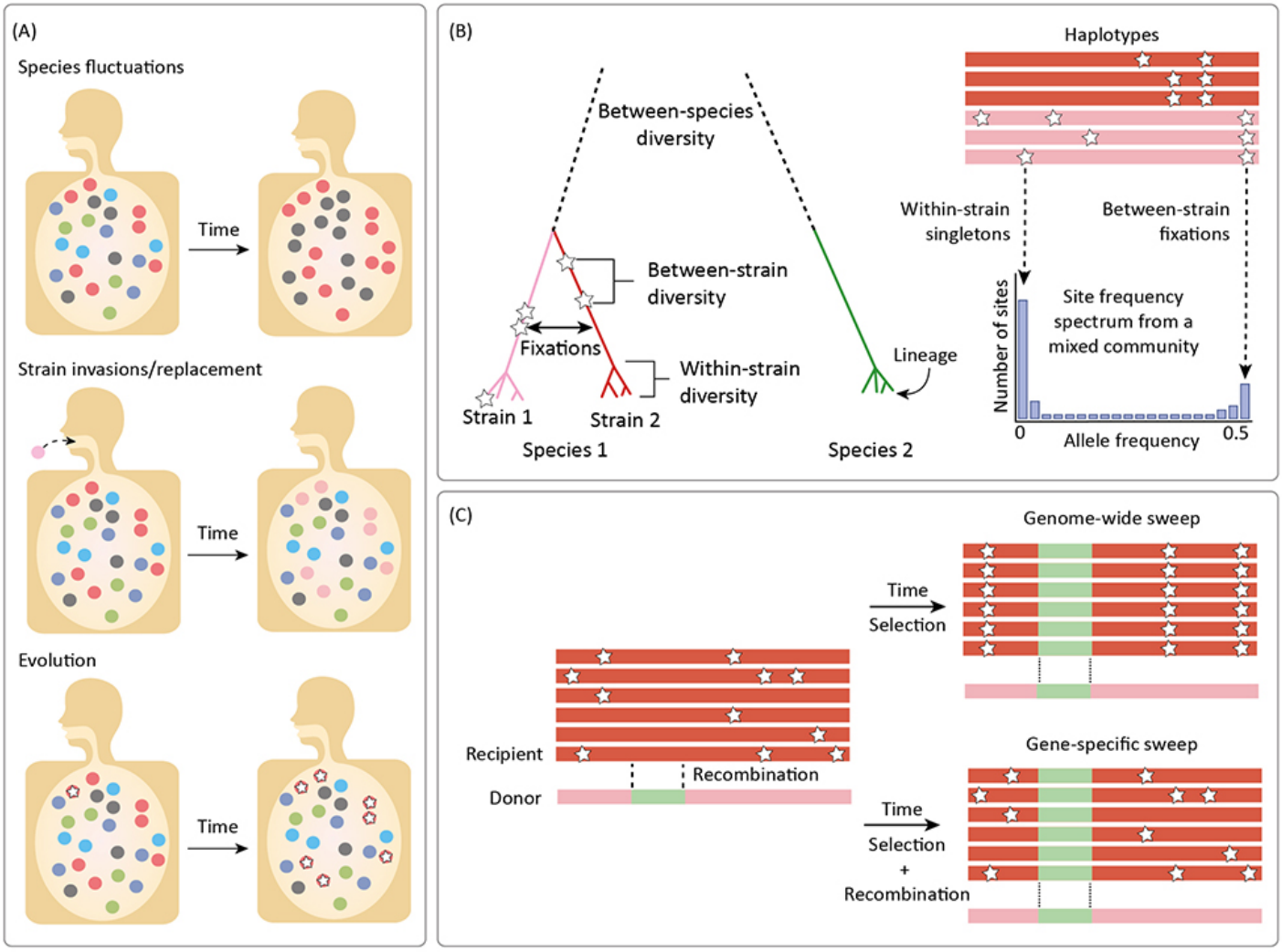
79 Rosen, M.J. *et al.* (2018) Probing the ecological and evolutionary history of a thermophilic cyanobacterial population via statistical properties of its microdiversity. *PLoS One* DOI: 10.1371/journal.pone.0205396
80 Lerat, E. *et al.* (2005) Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol.* DOI: 10.1371/journal.pbio.0030130
81 Wiedenbeck, J. and Cohan, F.M. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. , *FEMS Microbiology Reviews*. (2011)
82 Costea, P.I. *et al.* (2017) Subspecies in the global human gut microbiome. *Mol Syst Biol* 13, 960
83 Cadillo-Quiroz, H. *et al.* (2012) Patterns of gene flow define species of thermophilic Archaea. *PLoS Biol.* DOI: 10.1371/journal.pbio.1001265
84 Kumarasamy, K.K. *et al.* (2010) Emergence of a new antibiotic resistance mechanism in India, Pakistan, and the UK: A molecular, biological, and epidemiological study. *Lancet Infect. Dis.* DOI: 10.1016/S1473-3099(10)70143-2
85 Coleman, M.L. and Chisholm, S.W. (2010) Ecosystem-specific selection pressures revealed through comparative population genomics. *Proc. Natl. Acad. Sci.* DOI: 10.1073/pnas.1009480107
86 Lester, C.H. *et al.* (2006) In vivo transfer of the vanA resistance gene from an Enterococcus faecium isolate of animal origin to an E. faecium isolate of human origin in the intestines of human volunteers. *Antimicrob. Agents Chemother.* DOI: 10.1128/AAC.50.2.596-599.2006
87 Guttman, D.S. and Dykhuizen, D.E. (1994) Detecting selective sweeps in naturally occurring Escherichia coli. *Genetics*
88 Otto, S.P. and Lenormand, T. Resolving the paradox of sex and recombination. , *Nature Reviews Genetics*. (2002)
89 MacLean, R.C. and San Millan, A. Microbial Evolution: Towards Resolving the Plasmid Paradox. , *Current Biology*. (2015)
90 Redfield, R.J. (2001) Do bacteria have sex? *Nat. Rev. Genet.* DOI: 10.1038/35084593
91 Cox, M.M. *et al.* (2000) The importance of repairing stalled replication forks. *Nature* DOI: 10.1038/35003501
92 Cox, M.M. Recombinational DNA repair in bacteria and the RecA protein. , *Progress in nucleic acid research and molecular biology*. (1999)
93 Nayfach, S. *et al.* (2015) An integrated metagenomics pipeline for strain profiling reveals novel patterns of transmission and global biogeography of bacteria. *bioRxiv*
94 Yang, C. *et al.* (2018) Why panmictic bacteria are rare. *bioRxiv* DOI: 10.1101/385336
95 Palumbi, S.R. (2001) Humans as the world's greatest evolutionary force. *Science (80-.).* 293, 1786–1790
96 Feder, A.F. *et al.* (2016) More effective drugs lead to harder selective sweeps in the evolution of drug resistance in HIV-1. *Elife* DOI: 10.7554/eLife.10670
97 Lieberman, T.D. *et al.* (2014) Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. *Nat Genet* 46, 82–87
98 Zanini, F. *et al.* (2015) Population genomics of intrapatient HIV-1 evolution. *Elife* DOI: 10.7554/eLife.11282
99 Xue, K.S. *et al.* (2017) Parallel evolution of influenza across multiple spatiotemporal scales. *Elife* DOI: 10.7554/eLife.26875
100 Sender, R. *et al.* (2016) Revised estimates for the number of human and bacteria cells in

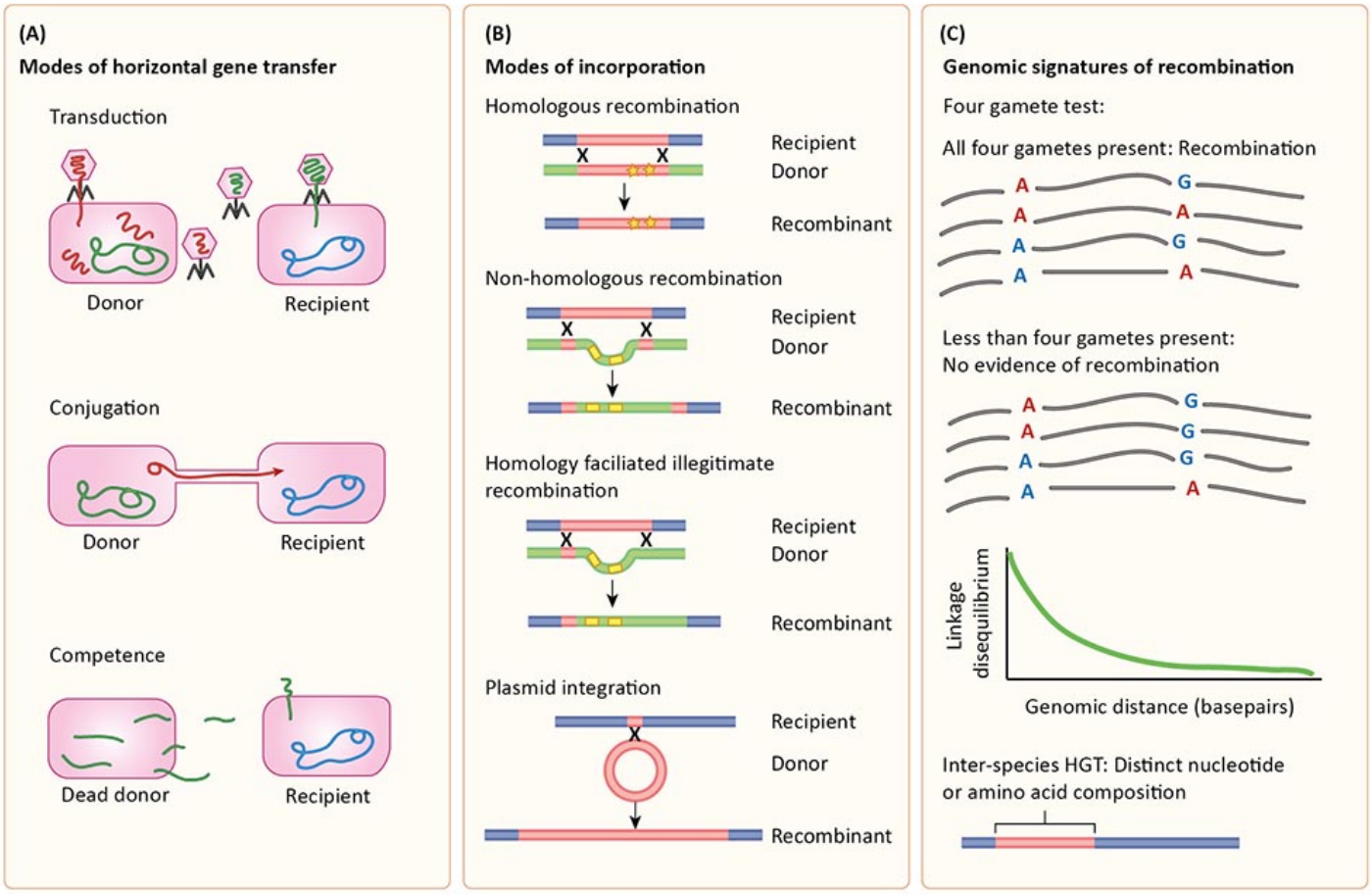
the body. *BioRxiv*
101 Pennings, P.S. and Hermisson, J. (2006) Soft sweeps II--molecular population genetics of adaptation from recurrent mutation or migration. *Mol. Biol. Evol.* 23, 1076–1084
102 Groussin, M. *et al.* (2017) Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. *Nat. Commun.* DOI: 10.1038/ncomms14319
103 Yaffe, E. and Relman, D.A. (2019) Tracking microbial evolution in the human gut using Hi-C. *bioRxiv* DOI: 10.1101/594903
104 Smith, N.H. *et al.* Bottlenecks and broomsticks: The molecular evolution of Mycobacterium bovis. , *Nature Reviews Microbiology*. (2006)
105 Lieberman, T.D. *et al.* (2011) Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nat. Genet.* DOI: 10.1038/ng.997
106 Simam, J. *et al.* (2018) Gene copy number variation in natural populations of Plasmodium falciparum in Eastern Africa. *BMC Genomics* DOI: 10.1186/s12864-018-4689-7
107 Cheeseman, I.H. *et al.* (2016) Population structure shapes copy number variation in malaria parasites. *Mol. Biol. Evol.* DOI: 10.1093/molbev/msv282
108 Garud, N.R. *et al.* (2015) Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps. *PLoS Genet.* DOI: 10.1371/journal.pgen.1005004
109 Sabeti, P.C. *et al.* Positive natural selection in the human lineage. , *Science*. (2006)
110 Voight, B.F. *et al.* (2006) A map of recent positive selection in the human genome. *PLoS Biol.* 4, e72
111 Simpson, G.G. (2019) *Tempo and Mode in Evolution*,
112 Pritchard, J.K. *et al.* (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr. Biol.* 20, R208-15
113 Donaldson, G.P. *et al.* Gut biogeography of the bacterial microbiota. , *Nature Reviews Microbiology*. (2015)
114 Tropini, C. *et al.* The Gut Microbiome: Connecting Spatial Organization to Function. , *Cell Host and Microbe*. (2017)
115 Lässig, M. *et al.* (2017) Predicting evolution. *Nat. Ecol. Evol.* DOI: 10.1038/s41559-017-0077
116 Gerrish, P.J. and Lenski, R.E. (2011) The fate of competing beneficial mutations in an asexual population.
117 Barroso-Batista, J. *et al.* (2014) The first steps of adaptation of Escherichia coli to the gut are dominated by soft sweeps. *PLoS Genet* 10, e1004182
118 Messer, P.W. *et al.* Can Population Genetics Adapt to Rapid Evolution? , *Trends in Genetics*. (2016)
119 Qin, J. *et al.* (2010) A human gut microbial gene catalog established by metagenomic sequencing Europe PMC Funders Group. *Nature* DOI: 10.1038/nature08821
120 Qin, J. *et al.* (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490, 55–60
121 Zhang, X. *et al.* (2015) The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat. Med.* DOI: 10.1038/nm.3914
122 Zeller, G. *et al.* (2014) Potential of fecal microbiota for early□stage detection of colorectal cancer. *Mol. Syst. Biol.* DOI: 10.15252/msb.20145645
123 Feng, Q. *et al.* (2015) Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat. Commun.* DOI: 10.1038/ncomms7528

124 Ferreiro, A. *et al.* (2018) Multiscale Evolutionary Dynamics of Host-Associated Microbiomes. *Cell* 172,
125 Forsberg, K.J. *et al.* (2012) The shared antibiotic resistome of soil bacteria and human pathogens. *Science (80-.).* DOI: 10.1126/science.1220761
126 Forslund, K. *et al.* (2013) Country-specific antibiotic use practices impact the human gut resistome. *Genome Res* 23, 1163–1169
127 Karami, N. *et al.* (2007) Transfer of an ampicillin resistance gene between two Escherichia coli strains in the bowel microbiota of an infant treated with antibiotics. *J. Antimicrob. Chemother.* DOI: 10.1093/jac/dkm327
128 Hacker, J. and Kaper, J.B. (2002) Pathogenicity Islands and the Evolution of Microbes. *Annu. Rev. Microbiol.* DOI: 10.1146/annurev.micro.54.1.641
129 Bron, P.A. *et al.* Emerging molecular insights into the interaction between probiotics and the host intestinal mucosa. , *Nature Reviews Microbiology*. (2012)
130 Needham, B.D. *et al.* (2013) Modulating the innate immune response by combinatorial engineering of endotoxin. *Proc. Natl. Acad. Sci.* DOI: 10.1073/pnas.1218080110
131 Koeth, R.A. *et al.* (2013) Intestinal microbiota metabolism of l-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nat. Med.* DOI: 10.1038/nm.3145
132 Haiser, H.J. *et al.* (2013) Predicting and manipulating cardiac drug inactivation by the human gut bacterium Eggerthella lenta. *Science (80-.).* DOI: 10.1126/science.1235872
133 De Filippis, F. *et al.* (2019) Distinct Genetic and Functional Traits of Human Intestinal Prevotella copri Strains Are Associated with Different Habitual Diets. *Cell Host Microbe* DOI: 10.1016/j.chom.2019.01.004
134 Yu, J. *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* DOI: 10.1038/ng1702
135 Chen, P.E. and Shapiro, B.J. The advent of genome-wide association studies for bacteria. , *Current Opinion in Microbiology*. (2015)
136 Falush, D. and Bowden, R. (2006) Genome-wide association mapping in bacteria? *Trends Microbiol.* DOI: 10.1016/j.tim.2006.06.003
137 Bradley, P.H. *et al.* (2018) Phylogeny-corrected identification of microbial gene families relevant to human gut colonization. *PLoS Comput. Biol.* DOI: 10.1371/journal.pcbi.1006242
138 Pasaniuc, B. and Price, A.L. Dissecting the genetics of complex traits using summary association statistics. , *Nature Reviews Genetics*. (2017)
139 Zhang, X. *et al.* (2015) The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat Med* 21, 895–905
140 Davenport, E.R. *et al.* (2015) Genome-wide association studies of the human gut microbiota. *PLoS One* DOI: 10.1371/journal.pone.0140301
141 Goodrich, J.K. *et al.* (2016) Genetic Determinants of the Gut Microbiome in UK Twins. *Cell Host Microbe* DOI: 10.1016/j.chom.2016.04.017
142 Lim, M.Y. *et al.* (2017) The effect of heritability and host genetics on the gut microbiota and metabolic syndrome. *Gut* DOI: 10.1136/gutjnl-2015-311326
143 Turpin, W. *et al.* (2016) Association of host genome with intestinal microbial composition in a large healthy cohort. *Nat. Genet.* DOI: 10.1038/ng.3693
144 Spor, A. *et al.* (2011) Unravelling the effects of the environment and host genotype on the gut microbiome. *Nat. Rev. Microbiol.* DOI: 10.1038/nrmicro2540
145 Goodrich, J.K. *et al.* Cross-species comparisons of host genetic associations with the

microbiome. , *Science*. (2016)
146 Wang, J. *et al.* (2016) Genome-wide association analysis identifies variation in Vitamin D receptor and other host factors influencing the gut microbiota. *Nat. Genet.* DOI: 10.1038/ng.3695
147 Wang, M. *et al.* (2018) Two-way mixed-effects methods for joint association analysis using both host and pathogen genomes. *Proc. Natl. Acad. Sci. U. S. A.* DOI: 10.1073/pnas.1710980115
148 Hehemann, J.H. *et al.* (2010) Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* DOI: 10.1038/nature08937
149 Gumpert, H. *et al.* (2017) Transfer and persistence of a multi-drug resistance plasmid in situ of the infant gut microbiota in the absence of antibiotic treatment. *Front. Microbiol.* DOI: 10.3389/fmicb.2017.01852
150 Li, S.S. *et al.* (2016) Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science (80-.).* 352, 586–589
151 Smillie, C.S. *et al.* (2018) Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation. *Cell Host Microbe* DOI: 10.1016/j.chom.2018.01.003
152 Lescat, M. *et al.* (2017) Using long-term experimental evolution to uncover the patterns and determinants of molecular evolution of an Escherichia coli natural isolate in the streptomycin-treated mouse gut. *Mol. Ecol.* DOI: 10.1111/mec.13851
153 Lagier, J.C. *et al.* (2012) Microbial culturomics: paradigm shift in the human gut microbiome study. *Clin Microbiol Infect* 18, 1185–1193
154 Forster, S.C. *et al.* (2019) A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nat. Biotechnol.* DOI: 10.1038/s41587-018-0009-7
155 Zou, Y. *et al.* (2019) 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat. Biotechnol.* DOI: 10.1038/s41587-018-0008-8
156 Poyet, M. *et al.* (2019) A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nat. Med.* DOI: 10.1038/s41591-019-0559-3
157 Ma, L. *et al.* (2014) Gene-targeted microfluidic cultivation validated by isolation of a gut bacterium listed in Human Microbiome Project's Most Wanted taxa. *Proc. Natl. Acad. Sci.* DOI: 10.1073/pnas.1404753111
158 Nelson, K.E. *et al.* (2010) A catalog of reference genomes from the human microbiome. *Science (80-.).* DOI: 10.1126/science.1183605
159 Segata, N. *et al.* (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* DOI: 10.1038/nmeth.2066
160 Franzosa, E.A. *et al.* (2018) Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* DOI: 10.1038/s41592-018-0176-y
161 Alneberg, J. *et al.* (2014) Binning metagenomic contigs by coverage and composition. *Nat. Methods* DOI: 10.1038/nmeth.3103
162 Parks, D.H. *et al.* (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* DOI: 10.1038/s41564-017-0012-7
163 Nayfach, S. *et al.* (2019) New insights from uncultivated genomes of the global human gut microbiome. *Nature* DOI: 10.1038/s41586-019-1058-x
164 Pasolli, E. *et al.* (2019) Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle.

Cell DOI: 10.1016/j.cell.2019.01.001
165 Almeida, A. *et al.* (2019) A new genomic blueprint of the human gut microbiota. *Nature* DOI: 10.1038/s41586-019-0965-1
166 Stewart, R.D. *et al.* (2018) Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat. Commun.* DOI: 10.1038/s41467-018-03317-6
167 Mou, X. *et al.* (2005) Flow-cytometric cell sorting and subsequent molecular analyses for culture-independent identification of bacterioplankton involved in dimethylsulfoniopropionate transformations. *Appl. Environ. Microbiol.* DOI: 10.1128/AEM.71.3.1405-1416.2005
168 Lan, F. *et al.* (2017) Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding. *Nat. Biotechnol.* DOI: 10.1038/nbt.3880
169 Bishara, A. *et al.* (2018) High-quality genome sequences of uncultured microbes by assembly of read clouds. *Nat. Biotechnol.* DOI: 10.1038/nbt.4266
170 Kawecki, T.J. *et al.* Experimental evolution. , *Trends in Ecology and Evolution*. (2012)
171 Frazao, N. *et al.* (2018) Sex overrides mutation in Escherichia coli colonizing the gut. *bioRxiv* DOI: 10.1101/384875
172 Anthony JF Griffiths, Jeffrey H Miller, David T Suzuki, Richard C Lewontin, and W.M.G. (2000) *An Introduction to Genetic Analysis, 7th edition*,
173 Hsu, R.H. *et al.* (2019) Rapid microbial interaction network inference in microfluidic droplets. *bioRxiv* DOI: 10.1101/521823
174 Rakoff-Nahoum, S. *et al.* (2016) The evolution of cooperation within the gut microbiota. *Nature* DOI: 10.1038/nature17626
175 Oliveira, P.H. *et al.* (2017) The chromosomal organization of horizontal gene transfer in bacteria. *Nat. Commun.* DOI: 10.1038/s41467-017-00808-w
176 Ravenhall, M. *et al.* (2015) Inferring Horizontal Gene Transfer. *PLoS Comput. Biol.* DOI: 10.1371/journal.pcbi.1004095


(A)
Species fluctuations
Time
Strain invasions/replacement
Time
Evolution
Time
(B)
Between-species diversity
Between-strain diversity
Fixations
Within-strain diversity
Strain 1
Strain 2
Species 1
Lineage
Species 2
Haplotypes
Within-strain singletons
Between-strain fixations
Site frequency spectrum from a mixed community
Number of sites
0
Allele frequency
0.5
(C)
Genome-wide sweep
Time
Selection
Recipient
Donor
Recombination
Gene-specific sweep
Time
Selection
+
Recombination


(A)
Modes of horizontal gene transfer
Transduction
Donor
Recipient
Conjugation
Donor
Recipient
Competence
Dead donor
Recipient
(B)
Modes of incorporation
Homologous recombination
Recipient
Donor
Recombinant
Non-homologous recombination
Recipient
Donor
Recombinant
Homology faciliated illegitimate recombination
Recipient
Donor
Recombinant
Plasmid integration
Recipient
Donor
Recombinant
(C)
Genomic signatures of recombination
Four gamete test:
All four gametes present: Recombination
Less than four gametes present:
No evidence of recombination
Linkage disequilibrium
Genomic distance (basepairs)
Inter-species HGT: Distinct nucleotide or amino acid composition