Evolutionary Dynamics of the Human Gut Microbiome

by

Shijie Zhao

B.S., Peking University (2013)

Submitted to the Biology Graduate Program in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy in Biology

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

~~May 2019~~ [June 2019]

Signature redacted

Signature of Author: ______________________________

Shijie Zhao
Department of Biology

Signature redacted

Certified by: ______________________________

Eric J. Alm
Professor of Biological Engineering
Thesis Advisor

Signature redacted

Accepted by: ______________________________

Amy Keating
Professor of Biology
Co-Director, Biology Graduate Committee

Evolutionary Dynamics of the Human Gut Microbiome

by
Shijie Zhao

Submitted to the Department of Biology on May 24, 2019
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Biology

Abstract

The composite members of the human gut microbiome encounter a myriad of selective pressures from the host environment and other microbial members in the ecosystem. Understanding the evolutionary dynamics of microbial species in the gut microbiome requires sequencing information that differentiates strains and even single cells. In this thesis, I present efforts that investigate the evolution of bacterial strains in their complex natural environments. In the first project, I discover that a commensal species, *Bacteroides fragilis*, undergoes within-person adaptive evolution in the absence of antibiotics. Combining culture-based whole genome sequencing with metagenomes, I uncover genes important to *B. fragilis* survival in the human gut microbiome and describe evolutionary dynamics within individuals and across populations. In the second project, I developed a strain-tracking method that predicts personal microbiomes. Using this method to track closely-related strains, I discover signals of adaptive evolution for *Bacteroidetes* strains, potentially over decades of colonization in adult twins. In the final project, this strain-tracking method is applied to advance the analysis of microbial transmission within social networks of Fiji islanders. These projects demonstrate the power of genome-resolved and strain-resolved methods in revealing insights of evolutionary dynamics of the gut microbiome. Future studies are expected to further investigate other taxonomical groups in depth and technical breakthroughs are needed to improve the throughput of evolutionary studies of complex systems like the gut microbiome.

Thesis Supervisor: Eric Alm
Title: Professor of Biological Engineering

Acknowledgement

I greatly acknowledge my thesis advisor Eric Alm, whom I thank wholeheartedly for his mentorship and advocacy. In the Alm lab, I had the chance to explore risky projects independently and consistently gain supports from Eric. He taught me how to navigate complex issues and come up with out-of-box solutions. Science is not always easy and full of challenges and obstacles; Eric's optimism has always been a positive influence to support me in going through the difficult times.

Not only a collaborator but more of a mentor, I can't be more appreciative to Tami Lieberman. We began to work together when we first formalized the idea of investigating the microbiomc through the lens of evolution. She trained me how to design studies, perform experiments, run rigorous computational analysis and write solid scientific manuscripts. Her pursuit of excellence in science has also inspired me to achieve the highest standards. Without her generosity and mentorship, my journey through the graduate school would have been much more difficult.

I am also thankful to Eric for assembling a group full of generous, collaborative and bright minds. Graduate school has been much more fun and comfortable because of these people. I collaborated with Thomas Gurry and Sean Gibbons on various projects and received enormous amounts of support, training, and advice from them. I overlapped early in my graduate school with Chris Smillie, Mark Smith, and Ilana Brito, and I thank them for leading me into the exciting world of the microbiome. Thanks to Mathilde Poyet and Mathieu Groussin, who provided lots of inputs in my major projects and have been an amazing source for inspirations. I want to give special thanks to the peer graduate students in the Alm lab: Xiaoqian Yu, Sean Kearney, Claire Duvallet, Nathaniel Chu, Scott Olesen, Sarah Spencer, and Sarah Bi. It has always been inspiring to discuss science and life with them. I also want to thank Chengzhen Dai and Xiaoqian Yu for their contributions to several of my projects and to proofreading my thesis.

Beyond Alm lab, I'd like to thank my committee members, Amy Keating, Gene-Wei Li, and Michael Baym, for supporting my journey of graduate school with scientific and career advice. I can't send enough thanks to Amy Keating, who has generously given me career supports and advice when those were most needed. I also want to thank my rotation lab advisors or mentors—Jing-ke Weng, Chris Burge, Jeff Gore, and Jonathan Friedman—for their generosity and the precious learning experience.

Contents

Chapter 1

Introduction

1.1 Fundamental questions in the evolutionary dynamics of the human gut microibome

The human gut microbiome is a dynamic and complex ecosystem consisting of four major phyla of micro-organisms: *Bacteroidetes*, *Firmicutes, Proteobacteria*, and *Actinobacteria*, and has large diversity with hundreds of distinct species (Lloyd-Price et al. 2017; Arumugam et al. 2011). Gut microbiome helps human resist pathogen invasion (Britton and Young 2014), harvest otherwise inaccessible nutrients (Peter J Turnbaugh et al. 2009), and modulate host behaviors (Gilbert et al. 2016). In addition, the gut microbiome has been shown to play roles in complex diseases such as inflammatory bowel diseases (IBD), obesity, diabetes and autism (Wlodarska, Kostic, and Xavier 2015; Methé et al. 2012), and has recently been shown to significantly impact the treatment of cancer immune therapies (W. Li et al. 2019; Routy, B.. et al. 2018; Aquino-Michaels et al. 2015). Although the ecological dynamics of the intestinal microbiome has been extensively studied in recent years, relatively little has been done to understand how our microbial inhabitants evolve during their colonization of the gut. Characterizing within-host evolution of the microbiome will provide insights into the selective pressures and ecological forces encountered in the gut.

The gut microbiome has enormous potential for evolution during the colonization of human hosts. About 10^{11} *de novo* mutations are estimated to emerge in the microbiome of an individual host per day: a total of 10^{14} bacterial cells live in the human gut, a typical bacterial genome size is on the order of 10^{6} bp, the bacterial replication error rate is 10^{-9} per base pair, and gut bacteria replicate on average once per day (Sender, Fuchs, and Milo 2016; Biek et al. 2015; Good et al.

2017). Although most of the mutations may be lost from the populations due to random drift and/or deleterious effects, there is still enormous potential for evolution. In addition, comparative genomic studies showed evidence that microbiome has co-evolved with human hosts for millions of years, mainly through mutations in genes involved in host interaction and nutrient utilization (J. Xu et al. 2007; Walter and Ley 2011; I. L. Brito et al. 2016). Evolution of bacterial species within a timescale of host lifespan is the building block of those long-term evolutionary events.

However, fundamental questions remain about the evolutionary dynamics of commensal species living in the gut microbiome. In particular, it is unknown whether adaptive evolution or purifying evolution dominates the within-person evolution of commensal gut species. Studies on pathogenic species have shown that within-host evolution is driven by strong host selection (Ley et al. 2008; Didelot et al. 2016). However, it is unclear if the lessons learned from pathogens can be generalized to members of our gut microbiome. In contrast to pathogens, whose natural environment is clearly outside the human body, the majority of bacterial species in the gut microbiome are commensal (neutral or beneficial to the host) and they have been co-evolving with human hosts for millions of years (Walter and Ley 2011). It is possible that the long history of colonization has exhausted adaptive mutations, and neutral or purifying selection has been dominating commensal species. On the other hand, an individual's gut ecosystem is a personalized and dynamic system, factors like pressures from the host immune system or diet may impose strong selections for contemporary adaptive mutations. We need empirical evidence to reveal the dominating force for within-person evolution of the commensal gut microbiome.

Understanding the selective pressures and ecological forces shaping the within-host evolution of the intestinal microbiome may fundamentally change our view of how personalized each individual's microbiome is and may have significant translational implications. Recent years have witnessed the emergence of fecal microbiota transplant (FMT) as a highly efficient treatment for recurrent *C.difficile* infections (Crum-Cianflone, Sullivan, and Ballon-Landa 2015). Inspired by this novel treatment, several clinical trials have been initiated to test microbiome therapeutics on IBD patients (Grinspan and Kelly 2015), and attempts have been made to develop synthetic microbiome drugs to treat various diseases (Petrof and Khoruts 2014; Routy, B. et al. 2018). Knowledge of adaptive evolution in the gut may reveal mutations that increase the fitness of our personalized microbiomes (Klemm et al. 2016; Coombes 2016). These adaptations may provide orthogonal insights to optimize therapeutics design and even underscore the importance of an autologous therapeutics (microbiome-based drugs derived from patients themselves before disease onset) (Grinspan and Kelly 2015). Moreover, mechanistic understandings of the microbiome within-host evolution may inspire future therapeutic direction by targeting genes under strong selective pressures (Lieberman et al. 2011; Lee et al. 2013; Donaldson et al. 2018).

1.2 Methods to investigate the within-person evolutionary dynamics of gut microbiome

Technical limitations have hindered attempts to characterize the within-host evolution process of microbiome. The standard approach for the microbiome field—metagenomic sequencing—is inefficient, when used alone, in resolving mutations emerged within the human gut (P.J. Turnbaugh et al. 2007; Methé et al. 2012). Mutations between different bacterial cells can be identified by aligning metagenomic reads to reference genomes, but SNPs identified this way may arise from homologous regions shared by closely-related lineages (Schloissnig et al. 2013). In addition, it is difficult to differentiate low frequency *de novo* mutations to Illumina sequencing errors. Previous attempts using metagenomes have thus been focused on using comparative genomic analysis to analyze strains carried by different human hosts (Schloissnig et al. 2013). Those genomes are usually separated by thousands of years of evolution, and these studies reveal long-term evolutionary dynamics. Culture-based whole-genome sequencing circumvents the limitations of metagenomes by enabling precise identifications of mutations between recently diverged genotypes and phylogenetic inference. However, due to the high complexity of the microbiome, previous reports have not optimize the sampling schemes to gain enough number of isolates at the right genetic distance for investigating within-person evolution (Forster et al. 2019; Zou et al. 2019; Browne et al. 2016; J. J. Faith et al. 2013).

In this thesis, I systematically study the within-host evolution of commensal species from the human gut microbiome using a combinatorial approach. The culture-based whole-genome sequencing framework is adopted in large scale to investigate a single bacterial species—*Bacteroides fragilis* (Chapter 2). Although metagenomes are of low-efficiency at detecting *de novo* SNPs, they are valuable sources to track the mutational dynamics. I, therefore, combine the SNPs identified via whole-genome analysis with metagenomes from the same human subjects and describe the evolutionary dynamics within subjects. In addition, I develop a strain-tracking method that rapidly and accurately detects closely-related strains from different metagenomic samples. This method allows me to identify strains shared by family members potentially over decades. SNP analysis is performed for these strains and within-person evolutionary history is revealed (Chapter 3).

1.3 *Bacteroides fragilis* undergoes within-person adaptive evolution

In the second chapter, I present one of the pioneering works that investigated the within-person evolutionary processes of one prevalent and abundant species—*Bacteroides fragilis*. Six hundred and two *B. fragilis* isolates are cultured and sequenced from twelve healthy donors' microbiomes. We find that *B. fragilis* within-subject populations contain substantial *de novo* nucleotide and mobile element diversity, which preserve years of within-person evolutionary history. This evolutionary history contains signatures of within-person adaptation to both subject-specific and common selective forces, including parallel mutations in sixteen genes. These sixteen genes are involved in cell-envelope biosynthesis and polysaccharide utilization, as well as yet under-characterized pathways. Notably, one of these genes has been shown to be critical for *B. fragilis* colonization in mice (Lee et al. 2013), indicating that key genes have not already been optimized for survival *in vivo*. This lack of optimization, given historical signatures of purifying selection in these genes, suggests that varying selective forces with discordant solutions act upon *B. fragilis in vivo*. Remarkably, in one subject, two *B. fragilis* sublineages coexisted at a stable relative frequency over a 1.5-year period despite rapid adaptive dynamics within one of the sublineages. This stable coexistence suggests that competing selective forces can lead to *B. fragilis* niche-differentiation even within a single person. By mining publicly available deeply sequenced metagenomes from different countries, we identify an evolutionary signal that is enriched in Western metagenomes than Chinese. We conclude that *B. fragilis* adapts rapidly within the microbiomes of individual healthy people, providing a new route for the discovery of key genes in the microbiome and implications for microbiome stability and manipulation.

1.4 A strain tracking method detects personal microbiome and signature of adaptive evolution for *Bacteroides* species

In the third chapter, I introduce a metagenome-based approach that facilitates the investigation of the evolution of microbial strains in the metagenomes. This method—DonorFinder—can rapidly compare strains from different metagenome samples and identify closely-related strains from different metagenomes. DonorFinder assumes that many commensal strains carried by different individuals are distinct in accessory genomic contents and are stably colonizing for years. This interpersonal variability of strains thus helps us understand the personal signatures of the microbiome and the transmission of strains between individuals. DonorFinder achieves near-perfect specificity and sensitivity in predicting metagenome donors and discovers a pair of metagenomes with labels switched in the HMP metagenomes. In addition, we apply DonorFinder to metagenomes from TwinUK registry and discover 6 closely-related *Bacteroidetes* strains that are shared between the twins. Analyses of point mutations swept in either twin in these strains indicate genome-wide within-person adaptive evolution during the time that these strains diverged between the twins' microbiomes.

1.5 Additional application of the strain tracking method reveals social networking of Fiji islanders

In the fourth chapter, a modified version of DonorFinder is applied to a metagenome dataset from Fiji Islanders. In a closed society of 287 people from the Fiji Islands, we examined how bacterial strains transmit across people via oral and gut microbiomes. Strain-level variations are tracked using both SNP-based analysis and a modified version of DonorFinder. Using both methods, we find strong transmission patterns enriched within households and between spouses. We also find that the host gender is strongly associated with strain-sharing patterns and the transmission patterns of gut and oral microbiomes are not necessarily the same.

Chapter 2

Adaptive Evolution within Gut Microbiomes of Healthy People

Shijie Zhao, Tami D. Lieberman, Mathilde Poyet, Kathryn M. Kauffman, Sean M. Gibbons, Mathieu Groussin, Ramnik J. Xavier and Eric J. Alm

The contents of this chapter have been published as: "Adaptive evolution within gut microbiomes of healthy people" at *Cell Host & Microbe* on April 23, 2019.

Abstract

Natural selection shapes bacterial evolution in all environments. However, the extent to which commensal bacteria diversify and adapt within the human gut remains unclear. Here, we combine culture-based population genomics and metagenomics to investigate the within-microbiome evolution of *Bacteroides fragilis*. We find that intra-individual *B. fragilis* populations contain substantial *de novo* nucleotide and mobile element diversity, preserving years of within-person history. This history reveals multiple signatures of within-person adaptation, including parallel evolution in sixteen genes. Many of these genes are implicated in cell-envelope biosynthesis and polysaccharide utilization. Tracking evolutionary trajectories using near-daily metagenomic sampling, we find evidence for years-long coexistence in one subject despite adaptive dynamics. We used public metagenomes to investigate one adaptive mutation common in our cohort and found that it emerges frequently in Western, but

not Chinese microbiomes. Collectively, these results demonstrate that *B. fragilis* adapts within individual microbiomes, pointing to factors that promote long-term gut colonization.

2.1 Introduction

The human gut microbiome harbors a large potential for within-person bacterial evolution and adaptation. Commensals can stably colonize a person for decades (Faith *et al.*, 2013), and during this time billions of bacterial mutations are generated daily (**Table 1**). Should adaptive mutations arise and be detectable within individual microbiomes, they are likely to indicate genes and pathways whose fine tuning is critical for long-term bactcrial persistence in the human body (Feliziani *et al.*, 2014; Lieberman *et al.*, 2011). In bacteria, adaptive evolution can be detected by the independent recurrence of similar mutations in genes under selection (parallel evolution or convergent evolution) or by an increase in mutational frequency that is inconsistent with neutral drift (Lieberman *et al.*, 2011; Wichman *et al.* , 2012; Woods *et al.*, 2006). The selective forces driving within-person adaptation might be person-specific, exposure-specific (e.g. diet), or widespread, and their identification could guide microbiome-targeted therapies—including the selection or engineering of therapeutic bacteria for long-term colonization. Additionally, within-person adaptation, if it occurs, may contribute to the stability of microbiome communities and their resilience to invasion (Martínez *et al.*, 2018).

However, relatively little is known about how commensals evolve within humans. To date, identification of contemporary adaptive point mutations has only been described during infections and in laboratory experiments. In these cases, the bacteria under study were exposed to environmental conditions clearly novel to them: the presence of antibiotics (Mwangi *et al.*, 2007; Snitkin *et al.*, 2013), a new host species (Didelot *et al.*, 2016), or artificial laboratory environments (Barrick *et al.*, 2009). However, human commensal bacteria have been colonizing mammalian digestive tracts for potentially hundreds of thousands of years (Moeller *et al.*, 2016; Groussin *et al.*, 2017). After long periods of evolution in a relatively unchanging environment, only neutral or very weakly beneficial mutations are expected to be available (Wiser *et al.*, 2013; Didelot *et al.*, 2016). Consistent with this expectation, investigations into healthy carriage of commensals have not revealed signals of within-person adaptive evolution (Golubchik *et al.*, 2013; Ghalayini *et al.*, 2018) and several studies have found signals of long-term purifying selection in the gut microbiome (Schloissnig *et al.*, 2012; He *et al.*, 2010). Yet, gut microbiomes are heterogeneous and individualized ecosystems that may vary over time *(*Lloyd-Price *et al.*, 2017). Encounters with other microorganisms, host immune systems, and diets may impose novel selective pressures on bacteria, and it is possible that these variable forces provide the potential for within-person genomic adaptation of certain commensal species (Nemergut *et al.*, 2013). Empirical data is needed to understand whether the environments within and between human gut microbiomes are variable enough to enable adaptation within individual people. To date, technical challenges have limited characterization of within-person evolution in the gut microbiome. One major challenge of metagenomics is discriminating *de novo* mutations (those that arise within an individual) from variants in homologous regions shared by co-colonizing bacteria (e.g. multiple-strain colonization or the presence of closely related species with shared mobile element) (Schloissnig *et al.*, 2012). Moreover, it is difficult to resolve the phylogenetic relationships between *de novo* SNPs using metagenomic-based approaches (Garud *et al.*, 2017). Culture-based whole-genome sequencing circumvents these limitations by enabling precise measurements of mutational distances between coexisting genotypes and phylogenetic inference. However, culture-based approaches have so far been limited to a small number of closely-related isolates from the gut microbiome (Faith *et al.*, 2013).

Here, we systematically characterize the within-host evolution and adaptation of *Bacteroides fragilis,* a prevalent commensal in the large intestine of healthy people. We use culture-based population genomics to identify *de novo* mutations and complement these analyses with comparisons to metagenomic data. We find extensive within-person diversification and multiple signals of adaptation, including within-person parallel evolution in 16 genes. Our findings provide a genome-wide understanding of *B. fragilis* within-person evolution, highlight the potential of commensals to adapt to individual microbiomes, and provide a roadmap for discovering genes important to commensal gut colonization and persistence.

2.2 Results

Within-person *B. fragilis* diversity is consistent with a single colonization event

We set out to survey intra-species diversity and evolution of *B. fragilis* within 12 healthy subjects, all donors to the OpenBiome stool bank (ages 22-37; **Table S1**). A total of 30 fecal samples from these subjects were studied. These fecal samples included longitudinal samples from 7 subjects spanning up to 2 years and single samples from 5 subjects (**Table S2**). Subjects did not take antibiotics for at least 3 months prior to initial sampling or during longitudinal sampling. We sequenced the genomes of 602 *B. fragilis* isolates cultured from 30 fecal samples. Each isolate was derived from an independent single cell in the original microbiome community.

Previous investigations have suggested that each person's *B. fragilis* population is dominated by a single strain (Lee *et al.*, 2013; Verster *et al.*, 2017). To confirm this in our donor population, we compared all 602 isolates via alignment of short reads to a public *B. fragilis* reference (Methods). We identified single nucleotide polymorphisms (SNPs) between these 602 isolates and built a phylogeny for these isolates. Isolate genomes from different subjects differed by more than 10,000 SNPs, while genomes from the same subject differed by fewer than 100 SNPs (with one isolate exception; **Figures 1A-1B**). This pattern confirms that each subject was colonized by a unique lineage.

***B. fragilis* populations diversify for years within individuals, with occasional sweeps**

To ascertain if the sublineage diversity present in each person could have emerged within the subject's lifetime, we estimated the coalescence time of each person's *B. fragilis* population. To include mutations in accessory genomic regions, we built a draft genome for each lineage using reads from all isolates. We then identified polymorphisms and constructed person-specific phylogenies using these draft genomes (Methods**, Figures 2A** and **S1-S3**). This sensitive approach detected between 8 and 182 polymorphic positions per subject (**Figure 2B**), and it enabled us to estimate the rate at which *B. fragilis* accumulates SNPs in the human gut (**Figures 2C-2D**; Methods). Our molecular clock estimate of ~0.9 SNPs/genome/year is within the range of what has been reported for bacterial species during infections of humans (Didelot *et al.*, 2012). Combining this rate and each population's phylogeny, we inferred that 11 of 12 lineages had *B. fragilis* populations that emerged from an ancestral cell between ~1.1-10 years before the initial sampling (time to most recent common ancestor, tMRCA; **Figure 2E**). These values are consistent with an expansion from a single cell that existed years prior to the initial sampling. Given the low acquisition rate of *Bacteroides* (Faith *et al.*, 2013), it is likely that the sublineage diversity emerged within each subject. We conclude that a typical *B. fragilis* population diversifies for years within the human gut.

One lineage, L08, was an outlier with an estimated tMRCA of 43, and we suspected that this high estimate of tMRCA was due to hypermutation. Hypermutation is an excess of mutations due to a defect in DNA repair, is commonly observed in laboratory experiments and during pathogenic infections, and is associated with adaptation (Giraud, 2001; Jolivet-Gougeon *et al.*, 2011; Marvig *et al.*, 2013; Lieberman *et al.*, 2014; Chu *et al.*, 2017). To test this hypothesis, we examined the type of mutations accumulated and the intrapersonal phylogeny. We found that the excess of mutations in L08 relative to other subjects was due solely to an increase in GC to TA transversions within one sublineage, supporting hypermutation (P<0.001, Chi-squared test, **Figure 2F**) (Jolivet-Gougeon *et al.*, 2011). The topology of the rooted phylogeny and the tMRCA of non-hypermutator sublineages (9.9 SNPs/genome) suggest that the hypermutation phenotype emerged within this subject (**Figure 2F**).

We noticed that estimates of divergence time were substantially smaller than each subject's age. These low values are consistent with colonization later in life, as well as early life colonization followed by loss of diversity through a neutral or adaptive sweep of a single sublineage. Consistent with the later scenario, we lost the ability to detect some sublineages in 3 of the 7 lineages using longitudinal samples over time (**Figures S2C, S2D and S2F**). Thus, the low values of tMRCA may have emerged because sweeps occasionally purge within-person *B. fragilis* population genetic diversity. We examine the role of adaptation, which might have driven these sweeps, in a later section.

Detection of mobile element transfer within individual microbiomes

We next assessed the relative contribution of horizontal gene transfer to within-person evolution of *B. fragilis.* We identified within-lineage mobile element differences (MEDs), which we define as DNA sequences with multi-modal coverage across isolates (Methods). We found MEDs in 11 of the 12 lineages (**Figure 2B**), including putative plasmids, integrative conjugative elements (ICEs), and prophages (**Table S3**). Using parsimony, we inferred 10 MEDs gained, 12 lost, and 17 ambiguous loci in ~50 cumulative years of evolution (using tMRCAs at initial samplings). This provided lower-bound event estimates of ~0.05 gain/genome/year and ~0.04 loss/genome/year and genomic change estimates of ~1.3 kbp gain/genome/year and ~1.9 kbp loss/genome/year. We did not find evidence of homologous recombination in these 12 lineages.

To identify MEDs transferred from the microbiome, we compared isolate genomes and metagenomes from the same subjects. We reasoned that a transferred region should have increased coverage in the metagenome compared to the rest of the *B. fragilis* genome, owing to its presence in other species (**Table S4**, Methods). We leveraged stool metagenomes available from 8 subjects, scanning for genomic regions with high relative coverage and high identity (>3X and >99.98%, respectively; Methods). We found evidence of one inter-species MED transfer within Subject 04 (38X relative coverage in the metagenomes; Methods; **Figures 3A-3B**). This MED, a putative prophage, was absent from all isolates at Day 0 yet present in 68% of isolates at Day 329. This combination of longitudinal genomic and metagenomic evidence suggests that this prophage was acquired by *B. fragilis* during the sampling period.

This same approach enabled us to identify three additional putative inter-species transfer events (**Table S4**; **Figure 3C**). We detected no difference in coverage between isolates for these regions (no MED), but an excess of coverage in the metagenomes. We confirmed one candidate from Subject 01, an integrative conjugative element (ICE) containing a type VI secretion system (Coyne *et al.*, 2016) (T6SS), by culturing and sequencing 94 isolates of other *Bacteroides* species from this subject. This ICE was present in 3 species (82 isolates) and harbored only 4

SNPs across these species, suggesting recent transfer (**Figures 3D** and **S1B-S1C**; Methods). T6SSs mediate inter-bacterial competition and have been shown to be shared by members of the same microbiome (Coyne *et al.*, 2014; Verster *et al.*, 2017). The prevalence of this ICE in this subject suggests it confers a strong selective advantage to its recipient species. In general, however, there are limited statistical tools for distinguishing adaptation from neutral evolution for mobile element exchanges.

Parallel evolution reveals genes involved in within-person adaptation

To systematically assess if adaptive mutations were a significant driver of within-person *B. fragilis* evolution, we searched for genes that underwent parallel evolution. Parallel evolution is the independent emergence of similar mutations on closely related genetic backgrounds, is a hallmark of positive selection, and is often used to identify putative targets of natural selection (Lieberman *et al.*, 2011; Wichman *et al.* , 2012; Woods *et al.*, 2006). We searched for genes that accumulated recurrent mutations within at least one person, leveraging the phylogeny to only include those events in which distinct mutations occurred in different sublineages (**Figure 4B**). We identified 16 such genes from the 12 lineages (**Figures 4B** and **4C**). This represents a significant deviation from a neutral model in which mutations occur randomly on the genome (P<0.001, **Figures 4C**; Methods). To confirm that adaptation, rather than mutational bias, was driving this clustering of mutations, we examined how many of the mutations encoded for an amino-acid change and compared this distribution to a neutral model (dN/dS, a canonical measure of selection). We found a significant enrichment for nonsynonymous mutations for these 16 genes, indicating adaptation (**Figure 4D**, dN/dS = 6.03, CI = (1.57, 51.3); Methods). We did not discover additional genes under adaptive evolution when including a search for parallel evolution across lineages (**Figures S4A-S4F**). We therefore conclude that some or all of these 16 genes underwent adaptive evolution within these subjects.

We found evidence of both subject-specific selection and selective forces shared across multiple subjects. Supporting person-specific selection, three Sus genes (BF1802, BF1803, and BF3581) were each mutated multiple times within one subject ($P < 0.003$ for each, Fisher's exact test) and no times in other subjects. In contrast, five genes under selection were mutated in multiple subjects, with two genes even acquiring mutations at the same amino-acid residue in different subjects (BF1708 and BF2755; **Figure 4B**). We discuss one of these mutations in detail in a following section.

Genes under parallel evolution are involved in polysaccharide utilization and cell envelope biosynthesis

The genes under parallel adaptive evolution reveal insights into the challenges to *B. fragilis* survival *in vivo*. The 16 genes include 5 involved in cell envelope biosynthesis, a dehydratase implicated in amino-acid metabolism, and 4 with unclear biological roles (**Figure 4B**). The remaining 6 genes all encode for homologs of SusC or SusD, a large group of outer-membrane polysaccharide importers (**Table S5**). A typical *B. fragilis* lineage has 75 distinct SusC/SusD pairs (out of ~4300 genes) and their main substrates are thought to be complex polysaccharides (Cerdeno-Tarraga, 2005; Martens *et al.*, 2009). SusC proteins form homodimeric β-barrels capped with SusD lids (Glenwright *et al.*, 2017), and the observed mutations were enriched at the interface between the barrel and lid (**Figure 4E**; $P<0.001$, Methods).

Notably, one of these SusC homologs (BF3581) has been shown to be critical for IgA-mediated colonization in mice. This locus has been designated as commensal colonization factor *(ccf)* (Lee *et al.*, 2013) and was the most significant locus discovered in a genome-wide screen for colonization determinants. The essentiality of the *ccf* locus is thought to be related to its regulation of capsular polysaccharide synthesis genes (Donaldson *et al.*, 2018). Therefore, while mutations altering Sus proteins might reflect pressures to utilize host or diet-derived polysaccharides (Martens *et al.*, 2009), selection on these genes might also reflect pressure to modify the *B. fragilis* cell envelope directly or indirectly. Additionally, the presence of Sus proteins in the outer membrane and their co-occurrence on this list with genes involved in cell envelope synthesis (**Figure 4B**) hints that selection on these genes might be driven by the pressure to evade the immune system (Merino and Tomás 2015) or phage predation (Stummeyer *et al.*, 2006).

Dense time-series reveals evolutionary dynamics and stable co-existence of sublineages

To better understand within-person evolutionary dynamics, we made use of the available densely sampled metagenomic time-series from Subject 01 and Subject 03. We closely examined the evolutionary dynamics for each lineage by tracking abundant SNPs, whose evolutionary relationships were previously identified from comparing isolate genomes. We inferred the population dynamics of sublineages, defined as clades with linked SNPs (Methods). These densely sampled time-series allowed us to track dynamics of *de novo* SNPs and to assess the strength of selection upon these mutations.

In both L01 and L03, we found SNPs that steadily increased in frequency, suggesting a fitness advantage of the lineages carrying them relative to their ancestors (**Figures 5 and S5)**. Given the large population sizes of *B. fragilis* in these subjects ($>10^{11}$; **Figure S5**), these relatively rapid rises in frequency are incompatible with neutral drift (Moran 1957). In L01, two linked mutations emerged around day 150 and swept a major sublinage (SL1) around day 400, increasing in frequency at a rate of 1.9% daily (**Figures 5A-5B**). One of these mutations was an amino-acid change in BF1802, a gene previously identified as under parallel evolution within this subject (**Figures 5C-5D**). The other mutation was 260 nucleotides upstream of a SusC gene not identified as under parallel evolution. In L03, the frequency of an amino-acid changing mutation in a glycosyltransferase (BF1196) rose from 0.5% at day 0 to 21% at day 144, corresponding to an average daily increase of 2.6% (**Figures 5E-5F**). While BF1196 did not show a signal for within-person parallel evolution, it was also mutated once in L10, suggesting this is an additional gene that may be under selection. Assuming that *B. fragilis* divides between 1-10 times per day, we estimate that these mutations provide a fitness gain (selection coefficients) of 0.2-2% for the two sweeping mutations from L01 combined, and 0.3-3% for the L03 mutation (Methods). These estimates are further evidence of adaptive evolution occurring within individuals in the absence of antibiotics.

Notably, in L01, the ratio between two major sublineages remained stable throughout the sampling period, despite the mutational sweep within SL1 (**Figure 5C**). We estimate that these major sublineages diverged ~8 years prior to sampling. This persistent coexistence suggests that the sweeping genotype, while 0.2-2% more fit than other genotypes from SL1, are not more fit than bacteria from SL2. This might result from frequency-dependent selection, ecological cross-feeding, or the occupation of distinct, perhaps spatially segregated, niches (Plucain *et al.*, 2014; Chung *et al.*, 2017; Good *et al.*, 2017; Rocabert *et al.*, 2017). The fact that 11 of 12 intragenic mutations separating these sublineages are amino-acid changing furthers the notion that they are functionally distinct.

To test if the two sublineages stably coexist *in vitro*, we competed combinations of isolates from different sublineages *in vitro* (Methods). Tracking their ratios using targeted amplicon sequencing, we found that both selected isolates from SL2 quickly outcompeted both selected isolates from SL1 (**Figure 5A; Figures S5G-S5J**). The growth profiles suggested active killing of SL1 in the presence of SL2 (**Figure S5K**). We noticed that all isolates from SL2 carried a prophage-like genomic element (MED01-2+), while only 14 of the 111 SL1 isolates were MED01-2+ (**Figures 5E** and **S1A**), and the above tested SL1 isolates both lacked this element (MED01-2-). To test the importance of MED01-2, we performed additional competition experiments including SL1 isolates that were MED01-2+ (Methods). We observed that, regardless of the sublineage-background, MED01-2+ isolates quickly outcompeted MED01-2- isolates, (**Figures 5F** and **5G**). In contrast, we observed stable coexistence of SL1 and SL2 when both competing isolates were MED01-2+ (**Figure 5H**). These results supported a pivotal role of MED01-2. To confirm that MED01-2 is a prophage and is responsible for SL2's *in vitro* competitive advantage, we performed phage plaque assays using 1000 donor-recipient combinations from L01 (40 donor isolates and 25 recipient isolates, Methods). Filtrates of MED01-2+ isolates from either SL1 or SL2 formed plaques on lawns of MED01-2- bacteria, but almost no plaques were found for other combinations (**Figure S5L** and **Table S6**). These results are consistent with an advantage of MED01-2+ isolates mediated by prophage-dependent killing.

These *in vitro* results are at odds with the observed within-person population dynamics. The years-long coexistence of SL1 and SL2—including SL1 isolates lacking MED01-2— suggests a balancing advantage for SL1 isolates that is not captured by our experiments. Alternatively, MED01-2 may provide a much weaker fitness advantage within Subject 01. These experimental results reflect the challenge of reconstructing within-person dynamics *in vitro* and highlight the power of dense and deeply analyzed timeseries for observing within-person evolutionary and ecological dynamics.

Parallel evolution in BF2755 is enriched in Western populations relative to Chinese populations

Lastly, we further investigated an amino acid change that had a high incidence across subjects. The mutant allele emerged four independent times across three subjects and was found in all isolates from L12 (Q100P mutation in BF2755, glutamine to proline). The function of BF2755 is unknown, but it is predicted to be periplasmic (Yu *et al.*, 2014) and has structural similarity to a beta-lactamase inhibitor (Das *et al.*, 2010). The high incidence of this mutation provided the opportunity to investigate its prevalence across human populations. We leveraged four available deeply-sequenced metagenome datasets: two from China (Qin *et al.*, 2012; Qin *et al.*, 2014), one from the USA (Lloyd-Price *et al.*, 2017), and one from the UK (Xie *et al.*, 2016) (Methods).

Unexpectedly, the mutant allele was at high prevalence in Western samples but nearly absent in the Chinese samples. Among Western metagenomes with evidence of *B. fragilis,* 15% had reads supporting the Q100P mutation, compared with only 1.5% in Chinese metagenomes (n=162 and n=136, respectively). This between-population difference was significant (**Figure 6A,** p<0.0001, Fisher's exact test) and robust to subject health status metadata (**Figure S6A**). To rule out the possibility that this difference was due to limited dispersal of a strain carrying this allele within Western populations, we reconstructed the evolutionary relationships among the *B. fragilis* strains within each metagenome (**Figure 6B**). We found that the occurrences of the Q100P mutations were on distinct and independent *B. fragilis* backgrounds (**Figure 6B**). In addition, 10 out of the 26 Western individuals with the derived allele showed evidence of coexistence of this mutation with the ancestral allele. This polymorphism, given that only a single lineage of *B. fragilis* colonizes each person (Lee *et al.*, 2013; Verster *et al.*, 2017), supports independent emergence of this mutation within each of these individuals. Further, a genome-wide search showed that this mutation is the most different locus between Western and Chinese populations (**Figure S6B**). In total, this data suggests a selective pressure to change this residue that is enriched in Western populations relative to Chinese populations.

2.3 Discussion

B. fragilis populations are dominated by single lineages (**Figure 1A**) which diversify within each individual to form coexisting sublineages (**Figure 2A**). Here, we report multiple lines of evidence that these sublineages acquire *de novo* mutations with significant beneficial effects, in the absence of antibiotic treatment and despite perhaps hundreds of thousands of years in mammalian digestive tract. This evidence includes: (1) independent, parallel acquisition of point mutations in the same gene among co-existing sublineages within individuals, concentrated in a few key pathways (**Figure 4B**); (2) an enrichment of amino-acid changing mutations relative to amino-acid preserving mutations, compared to a neutral model, in the target genes of parallel evolution (**Figure 4D**); and (3) rapid and continuous increases in the frequency of a few mutations (~2% daily increase, **Figures 5B-5C** and **S5E-S5F**). Adaptation of *B. fragilis* appears to be common feature of within-person *B. fragilis* evolution; 9 of 12 subjects had at least one mutation in the 16 genes we identified as under parallel evolution. The tempo of evolution observed here enables the straightforward identification of genes contributing to within-host adaptation, and therefore to long-term colonization in the microbiome, from either longitudinal sampling or investigation of many coexisting isolates.

This study was limited to a single species, and we hope that it will inspire similar studies for a variety of commensal organisms. Additional studies are needed to identify whether rapid adaptation is specific to *B. fragilis* or a common feature of gut commensals. Evidence that our results may be generalizable is provided by a recent study using metagenomics to track microbiome evolution across species (Garud *et al.*, 2017). This study detected that, averaged across species, single nucleotide variants at low frequency in the human population had a value of dN/dS consistent with either neutrality or adaptation—hinting at a possible microbiome-wide signature of adaptive evolution. In contrast, an investigation into *E. coli* within-microbiome evolution in one person uncovered only signatures of neutral diversity (Ghalayini *et al.*, 2018). While there are many possible explanations for the discrepancy between this finding for *E. coli* and our results, we speculate that genetic drift plays a larger role for *E. coli* due to its low population size within microbiomes (Lloyd-Price *et al.*, 2017). Future studies may identify taxonomic groups, bacterial life history strategies, human disease states, or other features that determine within-person evolutionary dynamics of commensals.

Selective forces that drive within-person adaptation

We report 16 genes in which adaptive mutations are concentrated, which warrant further study and whose identities provide hints about the nature of within-person selection. Six of the genes identified as under selection are members of the SusC/SusD family of nutrient import proteins. One pair of SusC/SusD genes (BF1802 and BF1803) have orthologs in *Bacteroides thetaiotaomicron* shown to be upregulated by milk oligosaccharides (Marcobal *et al.*, 2011) (**Table S5**). It is possible that some of the selective pressures driving mutations in these genes are in response to host diet. On the other hand, many of these genes are implicated in outer-membrane biosynthesis or encode for nutrient importers which sit in the outer membrane. In particular, a cell-envelope biosynthesis gene (BF2848) essential for the biosynthesis of 7 out of the 8 capsule polysaccharides was mutated in 3 lineages (Coyne *et al.*, 2008) (**Figure 4B**). We speculate that these genes are under pressure to evade phage predation or alter interaction with the immune system (Merino and Tomás 2015; Stummeyer *et al.*, 2006).

These same major pathways (capsule synthesis and SusC/SusD loci) are also controlled by invertible promoters in *B. fragilis*. At these loci, inducible integrases vary which gene in of a set of homologs is driven by a particular promoter. Using this mechanism and additional regulation, each *B. fragilis* isolate expresses only 1 of 8 capsule polysaccharides at a time (Kuwahara *et al.*, 2004; Cerdeno-Tarraga, 2005). It is interesting that the variation provided by invertible promoters does not preclude *de novo* mutations in these genes from contributing to within-person adaptation. More importantly, this overlap suggests that further elucidation of the pressures driving variation at these loci *in vivo* will illuminate the pressures driving within-person evolution.

Evolutionary dynamics within and across human subjects

The same genes identified here as under positive selection within individual people show signatures of purifying selection across lineages separated by thousands of years (**Figure 6C**; Methods). The discrepancy in signals between timescales raises the possibility that adaptive mutations in *B. fragilis* may incur collateral fitness costs in the context of other selective forces (e.g., following transmission to a new human host or invasion by a new species). We propose four scenarios that might reconcile the discrepancy between timescales (**Figure 6D**). The non-constant selective forces could be (1) specific to some people or lineages, (2) recently introduced into the human population (emerging), (3) present only at particular times during colonization, or

(4) coexisting within individual people. These models are not mutually exclusive and are agnostic to whether these forces are ecological or abiotic in nature. Our study, which was limited to 12 subjects, points to the existence of multiple of these non-constant selective forces.

A point of particular interest is whether the selective forces driving adaptation are person specific. In support of person specific selection, 11 of the 16 identified genes had mutations in only one subject. In particular, all six Sus genes under selection were mutated only in a single subject each. Furthermore, we did not find additional genes under adaptive evolution by grouping mutations from all subjects together (**Figures S4A-S4F**). We therefore speculate that person-specific or lineage-specific selection play important roles in shaping within-person evolution of the microbiome.

We also find evidence supporting other modes of contemporary selection. Five genes present signs of common selective forces (**Figure 4B**). Our finding of an amino acid frequently mutated in Western, but not Chinese, microbiomes, hints to a selective pressure that is enriched in Western populations (**Figures 6A-6B**). Studies tracking larger numbers of human subjects, as well as those tracking the same lineage in independent hosts (e.g. following fecal transplant), are needed to unravel the nature and specificity of pressures driving adaptation in these genes.

2.4 Declaration

Acknowledgements

We thank OpenBiome for providing stool samples and Hera Vlamakis, Paige Swanson, Timothy Arthur, Julian Avila-Pacheco, and Xiaofang Jiang for their assistance in obtaining samples and data. We are grateful to the BioMicroCenter at MIT and Microbial Omics Core at the Broad Institute for their assistance with library preparation and sequencing, Sean Kearney, Alexandra Stanton, Alexa Garcia and Nadine Fornelos Martins for experimental assistance, and Vicki Mountain, Katya Frois-Moniz, and Shandrina Burns for administrative assistance. We thank members of the Alm lab for helpful discussions and Kevin Roelofs, Xiaoqian Yu, Zhenrun Zhang, Chengzhen Dai, and Chris Smilie for comments on the manuscript. This work was funded by a grant from the Broad Institute. T.D.L. acknowledges support from Boehringer Ingelheim.

Author contributions

S.Z., T.D.L., and E.J.A. designed the study; S.Z. performed *B. fragilis* experiments; M.P. and M.G. performed experiments for other *Bacteroides*; K.M.K performed the phage experiments; S.M.G, R.J.X., and E.J.A. coordinated acquisition of metagenomic data. S.Z. and T.D.L. analyzed the data and wrote the manuscript with input from all authors.

Declaration of interests

Eric Alm is a co-founder and shareholder of Finch Therapeutics, a company that specializes in microbiome-targeted therapeutics.

2.5 Methods

Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
Stool samples from OpenBiome	This paper	N/A
Critical Commercial Assays		
Nextera DNA Library Preparation Kit	Illumina	FC-121-1031
MoBio PowerSoil kits	Qiagen	12955-4
PureLink Pro 96 Genomic Purification Kit	Thermo Fisher Scientific	K182104A
Bacteroidies Bile Esculin plates	BD	221836
Nextera XT DNA Library Preparation kit	Illumina	FC-131-1096
Deposited Data		
Raw sequencing data for isolate whole genomes	This paper	NCBI-SRA BioProject: PRJNA524913
Raw sequencing data for competition experiments	This paper	NCBI-SRA BioProject: PRJNA524913
BAM files of the 352 metagenomes	This paper	NCBI-SRA BioProject: PRJNA524913
Lineage assemblies with gene annotations	This paper	NCBI-SRA BioProject: PRJNA524913
Oligonucleotides		
Targeted Amplicon 1 Forward Primer: ATCTTCTATCGCCTGCCGTG	This paper	N/A
Targeted Amplicon 1 Reverse Primer: CGTGTATTCCGCCCTCTACC	This paper	N/A
Targeted Amplicon 2 Forward Primer: GCCAAAAACAAGGCAAATGACG	This paper	N/A
Targeted Amplicon 2 Reverse Primer: GGTCGCTTCCTTACGGGTAT	This paper	N/A
Software and Algorithms		
Cutadapt (version 1.9.1)	(Martin 2011)	https://cutadapt.readthedocs.io/en/stable/
Sickle	(Joshi and Fass 2011)	https://github.com/najoshi/sickle
Bowtie2 (version 2.2.6)	(Langmead and Salzberg 2012)l	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
SAMtools (version 1.2)	(H. Li et al. 2009)	http://samtools.sourceforge.net/
Spades (version 3.10.0)	(Bankevich et al. 2012)	https://github.com/ablab/spades
Prokka (version 1.11)	(Seemann 2014)	https://github.com/tseemann/prokka
FigTree (version 1.4.3)	N/A	http://tree.bio.ed.ac.uk/software/figtree/
PHYLIP (version 3.69)	(Plotree and Plotgram 1989)	http://evolution.genetics.washington.edu/phylip.html
CD-HIT	(Fu *et al.*, 2012)	https://github.com/weizhongli/cdhit
PyMol (version 2.2.3)	(Schrödinger, LLC 2015)	https://pymol.org/2/

PaperBLAST	(Price and Arkin 2017)	http://papers.genomics.lbl.gov/cgi-bin/litSearch.cgi
CELLO	(C.-S. Yu et al. 2014)	http://cello.life.nctu.edu.tw/
Consurf	(Ashkenazy *et al.*, 2010)	http://consurf.tau.ac.il/2016/
Clustal omega	(McWilliam et al. 2013)	https://www.ebi.ac.uk/Tools/msa/clustalo/
PDBePISA	(Krissinel and Henrick 2007)	http://www.ebi.ac.uk/pdbe/pisa/
Other		
Virulence Factors Database	(Chen et al. 2004)	http://www.mgc.ac.cn/VFs/

Contact for reagent and resource sharing

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Eric J. Alm (ejalm@mit.edu).

Experimental model and subject details

Stool samples were obtained from OpenBiome, a non-profit stool bank, under a protocol approved by the institutional review boards at MIT and the Broad Institute (# 1510271631). All 12 subjects were healthy people screened by OpenBiome to minimize the potential for carrying pathogens and had ages between 22 and 37 years and body-mass indexes between 19.5 and 26.2 at initial sampling. Subjects were de-identified before receipt of samples. **Table S1** contains detailed information about each subject.

Study cohort and sample collection

OpenBiome received and processed fresh stool donations within 6 hours of generation. Most samples were homogenized in a buffer containing 12.5% glycerol and 0.9% sodium chloride by mass (relative ratio of buffer to stool was either 10:1 or 2.5:1 volume/mass). Some samples were homogenized in proprietary buffers (1:1 volume/mass). Homogenized samples were passed through a 330-micron filter and stored at -80°C. Subjects 01-07 had multiple samples from which *B. fragilis* was selectively cultured, with time-series spanning 31 to 709 days. For Subjects 08-12, only one sample was selectively cultured for *B. fragilis*. Metagenomic sequencing was performed on stool samples from 8 of the 12 subjects (319 stool samples in total). Detailed information about samples used for isolation, including handling conditions prior to sample receipt, is in **Table S2**.

Library construction and Illumina sequencing

Samples were serially diluted in phosphate-buffered saline (PBS) and cultured for *B. fragilis* on *Bacteroidies* Bile Esculin plates (BD 221836) in an anaerobic environment. Single colonies suspected of being *B. fragilis* based on colony morphology were re-suspended in 50μL of PBS with 0.1% L-cysteine. For future characterization, 15μL of the re-suspension was mixed with 15μL of 50% glycerol and stored at -80°C. DNA was extracted from the remaining 35μL using the PureLink *Pro* 96 genomic purification kit, following the manufacturer's instructions. Genomic DNA libraries were constructed and barcoded using a modified version of the Illumina Nextera protocol (Baym *et al.*, 2015) (Library Prep. 1). Libraries from one sample (S01-0259, Day 709) were prepared by the BioMicroCenter at MIT using a different protocol, with lower input DNA and a final Pippin size-selection step (Library Prep. 2). Genomic libraries were sequenced either on the Illumina Hiseq platform with paired-end 100-bp reads or on the Illumina Nextseq platform with paired-end 75-bp reads by the Broad Institute Genomics Platform (**Table S2**).

SNP-calling and identification of major lineages

To estimate the distance between isolates across subjects and identify major lineages, we aligned all short reads to publicly available reference genome NCTC9343 (NCBI accession: CR626927.1) and identified SNPs. Reads were first trimmed and filtered using Cutadapt (Martin 2011) and Sickle (Joshi and Fass, 2011) (pe -q 20 -l 50) and aligned using Bowtie2 (Alignment parameters: -X 2000 --no-mixed --very-sensitive --n-ceil 0,0.01 --un-conc) (Langmead and Salzberg 2012). Isolates for which more than 70% of reads aligned to the reference and which had average coverage of greater than 10 reads across the genome were included for analysis (These filters excluded 1 isolate from subject 10 and 13 isolates from subject 06). Candidate SNPs were identified using SAMtools (Li *et al.*, 2009) and filtered using methods from previous work (Lieberman *et al.*, 2014). In particular, genomic positions were considered to be candidate SNP positions if at least one pair of isolates was discordant on the called base and both members of the pair had: FQ scores (produce by SAMtools) less than −60, at least 7 reads that aligned to each of the forward strand and reverse strand, and a major allele frequency of at least 90%. If the median coverage across samples at a candidate position was less than 10 reads or if 33% or more of the isolates failed to meet filters described above, this position was discarded. For each SNP position identified, a nucleotide call was assigned to each isolate using the major allele call across reads for that isolate at that position. If fewer than 7 reads aligned to either forward or reverse strand of a position in an isolate, or the major allele frequency was smaller than 90%, an ambiguous call was assigned to the isolate at that SNP position. See "code availability" for more information.

We generated a neighbor-joining tree from the concatenated list of variable positions from conserved genomic regions present in all *B. fragilis* isolates from all subjects. When computing the distance between each pair of isolates, we only used variable positions that had unambiguous nucleotide calls from both isolates. This tree showed 12 major clades corresponding to the 12 subjects and one minor clade containing a single isolate from Subject 10 (**Figure 1A**). Within each major clade, all isolates differed from one another by fewer than 100 SNPs. We therefore operationally defined a lineage as a set of isolates that differ by fewer than 100 SNPs and refer to specific genotypes within a lineage as sublineages. All lineages differed by over 10,000 mutations (**Figure 1B**); given the molecular clock estimated by this work, this represents at least thousands of years of evolutionary distance.

De novo assemblies of lineage genomes and within-lineage SNP identification

To enable us both to detect variants within genes carried only in a subset of lineages and to detect gains and losses of genomic regions that are specific to single lineages, we created a pan-genome for all isolates from each major lineage. For each major lineage, we concatenated reads (trimmed and filtered) from all isolates and used this concatenated file as the input for *de novo* genome assembly via Spades v3.10.0 (parameter: --careful) (Bankevich *et al.*, 2012). To limit the memory required for assembly, we used 0.25 million pairs of reads from each isolate (~7x coverage). Isolates prepared by the Library Prep. 2, as well as a few isolates with apparent cross contamination (genome assemblies built only using reads from single isolates were larger than 6MB) were excluded in building assemblies. Isolates not used to build the genome assemblies are indicated as such in the metadata associated with the uploaded raw data (see **Data availability**). Statistics of these genome assemblies are in **Table S1**. Assembly genomes were annotated using Prokka v1.11 (Seemann 2014). Lineage pan-genomes successfully assembled regions present in only a single isolate (e.g. **Figures S1A**, **S2C**, **S2E** and **S3A**) and enabled detection of mutations that would have been missed by comparison to a single reference (e.g. mutations in CL4395, **Figure 4B**). A genome assembly of the minor lineage from Subject 10 was built using all reads from this isolate.

Within-lineage mutations were identified by alignment of short reads to the corresponding lineage genome assembly, using the same parameters as described in the previous section. For lineage 10, the major allele frequency filter was set to 95% to exclude an apparent false positive. Candidate positions in MEDs were also discarded (see below for information on MED identification). Detailed information of intra-subject SNPs from the 12 subjects is listed in **Tables S6**.

The gene content across the 12 major lineage genomes and the NCTC9343 reference varied between 10%-20%.

Toxin detection

None of the *B. fragilis* genome assemblies showed evidence of pathogenicity. We compared the genome assemblies of the 12 major lineages and 1 minor lineage to the Virulence Factors Database, which contains >2400 virulence factors (Chen *et al.*, 2004), via BLAST using a threshold bit score of 200. We found only two hits to the database: Cps4J in L11 and ospC4 in L01. Both hits were not toxins previously characterized for *B. fragilis*. In contrast, this method identified 171 hits to known *B. fragilis*-related toxins from 30 out of 88 *B. fragilis* genomes from National Center for Biotechnology Information (NCBI).

Phylogeny of isolates from each *B. fragilis* lineage and identification of ancestral alleles

We used parsimony to reconstruct the evolutionary relationship between isolates from the same lineage. For each major lineage, a phylogeny of all isolates was built using a list of concatenated intra-subject SNPs, the closest lineage as an outgroup, and the dnapars program from PHYLIP v3.69 (Plotree and Plotgram, 1989). When parsimony could not resolve which allele was more likely to be ancestral, we inferred the ancestral allele to be the majority nucleotide at this genomic position across all other lineages with this genomic region. If a region was unique to a lineage, we assigned the ancestral allele that minimized the average mutational distances to the most recent common ancestor (dMRCA) for all isolates (3 cases).

dMRCA of each *B. fragilis* major lineage, molecular clock, and tMRCA

To calculate dMRCA for each subject at each time point, we counted the number of positions at which the called allele was different than the ancestral allele for each isolate, assessing only SNP positions that were polymorphic among isolates from the particular time point, and averaged the results. For each lineage with multiple time points, we computed the average number of new SNPs brought in per isolate from a later time point compared to the collection of SNPs identified at the initial time point. We then used linear regression to estimate the rate of evolution. The slope of the regression is our estimation of the evolutionary rate (**Figure 2C**). This method allows us to combine longitudinal data from different lineages to compute a molecular clock. In addition, we computed a molecular clock for L01, used tip-to-root distances overtime and obtained similar estimate (**Figure 2D**).

Each tMRCA was calculated by dividing dMRCA by the estimated molecular clock (**Figure 2E**). We stress that tMRCA is not an estimate of time to colonization, but simply an estimate of the age of the coexisting diversity. While potential systematic false negative and false positive SNPs may have impacted tMRCA values, these sources of error would have had a similar impact on our molecular clock estimation, as SNP-calling was consistent throughout. Other possible sources of error in estimating tMRCA include incorrect designation of ancestral versus derived allele and undersampling of the population, though collector curves for dMRCA indicate that sampling was usually sufficient (**Figures S7A-S7L**). Interestingly, collector curves for the number of *de novo* SNPs reflect that the number of SNPs identified did not saturate (**Figures S7M-S7X**).

Mutation spectrum of hypermutator sublineage

SNPs were categorized into 6 types, based on the chemical nature of the single nucleotide changes (**Figure 2F**). For L08, we computed the frequency of each type separately for the hypermutator sublineage and non-hypermutator sublineages (**Figure 2F**, purple and yellow bars). For the remaining lineages (L01-L07 and L09-L12), we computed the mutation spectrum for each lineage and then computed the mean and standard deviation of each of the 6 types (**Figure 2F**, gray bars). The mutation spectrum was significantly different between the hypermutator sublineage and the non-hypermutator sublineages (Chi-squared test, $P<0.001$), as well as the mean across the other 11 lineages (Chi-squared test, $P<0.001$). No significant difference was found between the 11 other lineages and the non-hypermutator sublineages from L08 (Chi-squared test, $P=0.4$). When excluding the GC-TA type of mutation from the analysis, we found no significant difference between the hypermutator sublineage in L08 from the 11 other lineages ($P=0.11$, Chi-squared test), suggesting that the hypermutation phenotype was exclusively due to an increase in GC-TA mutations.

Metagenomic library construction and Illumina sequencing

Genomic DNA was extracted from stool samples for metagenomic sequencing by the Microbial Omics Core at the Broad Institute using MoBio PowerSoil kits (Qiagen 12955-4) according the manufacturer's instructions. Genomic DNA libraries were constructed and barcoded by the Broad Technology Labs from 100-250pg of DNA using the Nextera XT DNA Library Preparation kit (Illumina) according to the manufacturer's recommended protocol, with reaction volumes scaled accordingly. Pooled libraries were sequenced on the HiSeq platform with paired-end 100bp reads by the Broad Technology Labs.

Identification of MEDs

We aligned short reads to the assembled genome of each major lineage as above and identified candidate regions that were at least 500nt in length, had low relative coverage (< 0.2X) at every nucleotide in at least one isolate, and had >0.9X coverage at every nucleotide in at least one isolate. For L01, we excluded isolates from the final time point, as these isolates' genomic libraries were prepared differently than the other isolates and therefore had different coverage pattern genomewide.

To account for the fact that single mobile elements could have been separated into multiple pieces in the genome assembly, we grouped regions suspected to emerge from the same event. We clustered sequences that had identical presence/absence patterns across all isolates, where presence was defined by >0.4X average relative coverage over the region. On 3 occasions, we noticed regions that had the same presence/absence pattern but had different coverage distribution across isolates, suggesting they came from distinct mobile elements. In these cases, we separated these clusters of sequence regions into clusters with consistent coverage distribution patterns. Detailed information of all MEDs is in **Table S3**.

MED gain and loss rates

We used parsimony to infer whether a MED was a gain or loss event. For each MED, we inferred events on the phylogenetic tree generated from whole genome data. If a single change of one type (e.g. gain) could explain the distribution, but more events were required for the other type (e.g. loss), the MED was categorized as such (**Table S3; Figure 2B**). Seventeen MEDs were classified as unknown because either: multiple gain or multiple loss events were required to explain the distribution (e.g. MED01-2); or both a single gain event and a single loss event were consistent with the distribution. Interestingly, one putative MED from L11 appeared to have been lost many times among isolates during culture (**Figure S3D**). To estimate lower bounds for the rates at which gain and loss events change *B. fragilis* genomes, we weighted each observed MED *j* by its frequency within lineage *i* (*f*$_{ij}$). We then divided the weighted sum of events by the total time of diversification, estimated by the sum of tMRCA at initial sampling. The following equation was used for gain and loss events, separately: $_{i}\,_{j} f_{ij}\ /\ _{i} tMRCA_{T0,i}$. To estimate the absolute contribution of gain and loss events to the size of *B. fragilis* genomes, we accounted for length of each MED (*L*$_{ij}$): $_{i}\,_{j}\ (L_{ij} f_j)\ /\ _{i} tMRCA_{T0,i}$

Inter-species mobile element transfer

For each lineage, we scanned the assembled genome for regions with high average relative coverage when aligning metagenomic reads to the lineage genome assembly (>3X). The coverage of metagenomic reads over the *B. fragilis* assembly varied over as much as 1000 folds due to reads from homologous regions of different species. Therefore, to normalize against the true expected coverage of the *B. fragilis* genome, we divided observed coverage at each position by the mean coverage across positions between the 30th percentile and 70th percentiles (median was not precise given the low coverage in some samples). To identify recent transfer events, we searched the genome for candidate regions >5000 nucleotides in length and in which the consensus genome from metagenomes was <0.02% different from the consensus genome from isolates of the same subject. We found 14 candidate regions in 3 lineages. We found only two candidate regions that overlapped with MEDs, all of which were in Subject 04 (representing one MED). Information about these candidate regions is listed in **Table S4**.

We identified two genomic regions (31 Kb and 62 Kb, respectively) that were candidates for inter-species mobile element transfer in Subject 01. These two regions contained distinct ORFs homologous to conserved genes from type 6 secretion system of genomic architecture 2 (**Figures S1B-S1C**), consistent with a single transfer event. This transfer event was inferred to be an integrative conjugative element (ICE) because it contains the *tra* genes associated with integrative conjugative elements and a tRNA gene at one edge of a transfer region (**Table S4**). To test if the putative ICE was indeed transferred between species, we cultured and sequenced the genomes of 94 *Bacteroides* isolates from this subject. We examined 53 *Bacteroides vulgatus* isolates (43 isolates one *B. vulgatus* lineage, 10 isolates from a different *B. vulgatus* lineage, **Figures S1B-S1C**), 25 *Bacteroides ovatus* isolates, 4 *Bacteroides xylanisolyens* isolates, 10 *Bacteroides stercoris* isolates and 2 *Bacteroides salyersiae* isolates. We sequenced these isolates as described for *B. fragilis* and aligned reads to the mobile element candidates, using the same parameters for *B. fragilis.* Strikingly, both genomic regions were present (average coverage >10 reads) in all *B. ovatus, B. xylanisolyens,* and *B. vulgatus* isolates profiled, but absent in all isolates of the other two species. The perfect co-occurrence of these two genomic regions further supports that they were from a single transfer event.

Parallel evolution

We counted a gene as under parallel evolution if, in at least one subject, the gene had multiple independent SNPs and more than 1 SNP per 2,000 bp (to account for the fact that long genes are more likely to be mutated multiple times by chance). Cases in which two SNPs in the same gene always occurred together in the same isolates were not included as parallel evolution (one case from L04). To identify nucleotide positions that mutated multiple independent times within a person, we leveraged the parsimony phylogenies described above. We inferred the genotypes of all internal nodes using the parsimony assumption and counted the number of mutation events. This method identified 3 nucleotides that were mutated multiple times within an individual (**Figures S1A**, **S3A**, and **S3C**). All genes under parallel nucleotide evolution also underwent parallel evolution involving distinct amino acid residues within at least one lineage. To determine whether the number of genes under parallel evolution represented a significant departure from what would be expected in a neutral model, we performed for each subject 1,000 simulations in which we randomly shuffled the mutations found across the lineage genome assembly and calculated how many genes showed a signature of within-person parallel evolution (**Figure 4C**). To compare genes from different assemblies, coding sequences identified by Prokka from all lineages were clustered using CD-HIT with at least 98% identity and 90% coverage (Fu *et al.*, 2012). Detailed information for each gene under parallel evolution is in **Table S5**. Simulations performed for metrics of cross-subject parallel evolution did not yield additional signatures of adaptive evolution (**Figures S4A-S4F**).

dN/dS

Mutations were categorized as synonymous (S) or non-synonymous (N) based on open-reading frame annotations created by Prokka (Seemann 2014). To calculate dN/dS for sets of *de novo* mutations emerged within subjects (**Figure 4D**, first two categories), we normalized the observed N/S ratios by the expected N/S ratios. For any given set of SNPs, we calculated the expected N/S for these SNPs, accounting for both (1) the different probabilities of acquiring nonsynonymous mutations for different types of mutations and (2) the codon compositions of the genes in which these SNPs occurred. This method is similar to what we have done previously (Lieberman *et al.*, 2014), but accounts for different codon composition between genes. 95% confidence intervals were calculated using binomial sampling.

To compute dN/dS for mutations across lineages (**Figure 4D**, third category), we leveraged publicly available sequences. We downloaded fastq files of 55 publicly available *B. fragilis* isolate sequencing runs. We then identified mutations across these genomes and the 12 major lineages from this study (one isolate per lineage) using the same approach and parameters described above (Identification of major lineages and SNPs). The NCTC9343 genome was used as reference and ancestor. Expected N/S ratio was calculated with the same method described above, using all the SNPs identified across lineages.

We calculated dN/dS for cross-lineage mutations in individual genes (**Figure 6C**). Since lineages are separated by tens of thousands of SNPs (**Figure 1**) and the molecular clock for *B. fragilis* is ~1 SNP/genome/year (**Figure 2C-D**), this metric reflects selection over thousands of years. Expected N/S ratio was calculated with the same method described above, using only cross-lineage SNPs identified within the particular genes. For 3 genes not present in the NTCT9343 genome (**Figure 4B**), we used the *de novo* assemblies to recruit reads from the publicly available sequences. No cross-lineage SNPs were identified in these 3 genes and dN/dS was not reported for these genes.

Annotation of genes under selection

To discover homologs of the sixteen genes under within-person parallel evolution, we used blastp to search against the RefSeq database, excluding proteins from *B. fragilis* genomes. Top hits with 3-4 letter gene names were searched against the *B. fragilis* genome to confirm whether they are true orthologs. We used the organisms from which these gene names were initially described to avoid false propagation of misannotation. We also used PaperBLAST to aid in identifying candidate gene names (Price and Arkin, 2017). Cellular localizations were predicted using CELLO (Yu *et al.*, 2014).

Conservation scores for each mutated residue was predicted using the Consurf web service (Ashkenazy *et al.*, 2010). For each gene, we used blastp to find homologs from the RefSeq database (first 100 hits; sequence similarity from 35% to 95%; query coverage > 80%). A multiple sequence alignment (MSA) was created using Clustal omega from the EMBL-EBI web service (McWilliam *et al.*, 2013) (default parameters). We then used each MSA to generate conservation score at each amino-acid residue using Consurf (default parameters). Detailed information is in **Table S5**.

SusC and SusD protein structures and interface residues

Available crystal structures of a SusC homolog (BT1763) from *Bacteroides thetaiotaomicron* (Glenwright *et al.*, 2017) was used to visualize the mutations observed in Sus genes under parallel evolution. We aligned the five *B. fragilis* SusC proteins under parallel evolution and BT1763 using Clustal Omega from the EMBL-EBI web service (McWilliam *et al.*, 2013) (default parameters). For all non-synonymous mutations, we identified their aligned positions on the BT1763 crystal structure. Two amino acid residues aligned to the first 211 amino-acid region, which encodes for a plug domain and is not available in the crystal structure of BT1763 (Glenwright *et al.*, 2017). Eight non-synonymous mutations from Sus genes under parallel evolution are marked in red in **Figure 4E**, using PyMol software (Schrödinger, LLC, 2015).

To test if the mutated residues were enriched at the interface between SusC and SusD, we used the PDBePISA web service (Krissinel and Henrick, 2007) (default parameters) to classify residues on the BT1763 crystal structure as in contact or not in contact with the SusD homolog. Of 806 residues, 119 were inferred to be interface residues. Among the 8 residues that were mutated in parallel, 4 of them were predicted to be interface residues in both programs, a significant enrichment (P=0.02, Fisher's exact test). A similar result was obtained using the PyMol function InterfaceResidues (cutoff=1.0; P=0.02, Fisher's exact test).

Enrichment of membrane proteins

For all genes from the 12 major lineage genome assemblies, we used CELLO (Yu *et al.*, 2014) to predict the cellular localization. Genes were considered to be membrane-related if they were annotated as inner membrane, periplasmic, or outer membrane. To compare our observation to the null expectation, we performed simulations. For each of the sixteen genes, we randomly selected one gene from the genome assembly of the lineage in which parallel evolution was identified. If a gene had parallel mutation in multiple lineages, we randomly chose one of the lineages. The cellular localization of *n* SNPs was assigned based on the CELLO prediction of this randomly picked gene, where *n* is the number of SNPs the original gene had across lineages. The proportion of SNPs from membrane-related genes was inferred using all sixteen such randomly picked genes (repeat genes not allowed). This procedure was repeated 1000 times to draw a null distribution of proportion of membrane-related SNPs. We calculated that in the sixteen genes under selection, 79% of the SNPs are from membrane-related genes, a significant deviation from the null distribution (Binomial test, P<0.001).

Signatures of subject-specific adaptation

Fisher's exact statistic was used to test subject-specific adaptation, comparing the number of SNPs in a tested gene within a particular lineage, the number of SNPs in other genes within this lineage, the number of SNPs in this gene from all other lineages combined, and the number of SNPs in other genes from all other lineages combined. We tested 10 genes that were present in multiple subjects but mutated only in one subject. The p-values for BF1802, BF3581, BF1803, are all less than 0.005, suggesting person-specific adaptation.

Mutation dynamics from metagenomes

Metagenomic reads from Subject 01, acquired as described above, were aligned to the assembled genome of L01 using the same parameters described for aligning isolates reads. We tracked the frequency of each SNP found in 4 or more isolates from L01; SNPs found in fewer isolates were not abundant in the metagenomes. For each of the 21 SNPs that met this threshold, we calculated the frequency of reads at each position that agreed with the mutation (derived) allele. As the total metagenomics sequencing coverage was limited and *B. fragilis* represented only ~5% of reads on average (**Figure S5A**), not every SNP was covered at every time point. For each SNP, we visualized its dynamics by using time points with non-zero read counts and smoothing the trajectory using the Savitzky-Golay method with a span of 25 and degree of 0 (**Figure 5B**).

To plot a schematic of the population dynamics of different sublineages (**Figure 5C**), we averaged frequencies of SNPs that were shared by a particular sublineage to estimate the relative abundance of this sublineage. To fill the time points where no stool community was sampled, we generated a continuous relative abundance trajectory for each sublineage using Fourier curve fitting (Matlab model fourier8). To visualize parent and child sublineages separately, we subtracted the relative abundance of a parent sublineage by the sum of relative abundances of its child sublineages. When the combined relative abundance of child sublineages exceeded that of their parent sublineage, we set the frequency of the parent sublineage to 0. After Day 180, we manually set the frequency of the SL1 parent genotype to zero, and reduced discontinuities caused by this assignment by an additional Fourier curve fitting step (Matlab parameter: fourier8). The imputed relative frequencies were then renormalized so that they sum up to 1. We also examined L03's dynamics during colonization using 74 metagenomes collected over 144 days (**Figures S5C-S5F**). The same methods were used as described above, with the exception that mutations in at least 3 isolates were able to be tracked, owing to the higher relative abundance of *B. fragilis* in Subject 03 (**Figure S5C**).

Selection coefficient was inferred using $(1 + s)^g = f$, where f represents the change in genotype frequency, g represents the number of generations and s represents the selection coefficient.

Competition experiments

We performed competition experiments using pairs or trios of isolates from different L01 sublineages. Frozen stocks were restreaked on brain heart infusion plates (Sigma-Aldrich 53286-500G) supplemented with haemin and vitamin K (BHIS) and revived for two days. Isolates were cultured concurrently using the following procedure in order to ensure reproducibility. Single colonies were inoculated in 1 mL of BHIS liquid media (hour -64). After 24 hours of growth, each pure culture was diluted 1:100 into 1 mL of BHIS liquid media and grown for another 24 hours. At hour -16, each pure culture was diluted 1:5 and grown for another 16 hours. All operations were performed in an anaerobic chamber and bacteria were grown at 37 °C.

Synchronized and saturated pure cultures were mixed at hour 0. Co-cultures were diluted 1:100 in 1 mL of BHI liquid media and grown at 37 °C anaerobically. At indicated points, 80 μL aliquots of each co-culture was taken for OD measurement and targeted amplicon sequencing. For the experiments shown in **Figures S5G-S5K**, time points were taken at 0, 6, 9, 12, 15 and 22 hours. For the experiments shown in **Figures 5E-5G**, we passaged the co-culture for another round of dilution at hour 18, and timepoints were taken at 0, 9, 18 and 27 hours.

Targeted amplicon sequencing

To determine the relative abundances of different sublineages in co-cultures, we picked two mutations from BF1802 that distinguished sublineages: D526N (T to C) mutation distinguished SL1 from SL2, and T340M (A to G) separated SL1-a-1-1 from all other sublineages. We designed two sets of primers that covered these mutations: 5'-ATCTTCTATCGCCTGCCGTG-3' and 5'-CGTGTATTCCGCCCTCTACC-3' for D526N and 5'-GCCAAAAACAAGGCAAATGACG-3' and 5'-GGTCGCTTCCTTACGGGTAT-3' for T340M. Each primer was linked to an Illumina adapter overhang nucleotide sequence (See online manual: Illumina 16S Metagenomic Sequencing Library Preparation). The co-culture was first incubated in alkaline PEG solution at 95 °C for 10 minutes (Chomczynski and Rymaszewski 2006). The target sequences were amplified individually using the KAPA HiFi HotStart Ready Mix, 2 μL lysis product, and 0.5μM of forward and reverse primers. Libraries were diluted 30X and barcoded using 2.5 μL diluted PCR products as template for PCR, the

KAPA HiFi HotStart Ready Mix, and 0.5μM Nextera primers (Baym *et al.*, 2015). Amplicon sequencing libraries were sequenced on the Illumina Miseq platform with paired-end 250-bp reads by the Broad Institute Genomics Platform. Sequencing reads were aligned to the assembled genome of L01 using the same parameters described for aligning isolates reads. Relative abundances were inferred by counting the number of nucleotides assigned to different sublineages at the targeted mutation loci.

Phage plaque assay

All pairs of donor-recipient assays were performed on three different media: BHIS, BPRM and BPRM+Bile (Media recipes can be found in **Table S6**). At hour 0, selected isolates from the freezer were restreaked on three different media plates. At hour 48, 10 colonies from each restreak were picked and inoculated into 500 μL of the corresponding liquid media. We then transferred 10 μL of the well-mixed pre-inoculum into 3.5 ml of media in a deep well 48-well culture block. Media for overnight cultures was aliquoted into tubes and culture blocks aerobically and these were transferred into the anaerobic chamber immediately prior to inoculation. To prepare donor filtrates at hour 73, we transferred 200 μL of donor cultures to 0.22 μm filter-bottom plate wells (MED Millipore MSGVS2210) attached to a receiver plate (Greiner Bio-One #651261) and centrifuged (3,200 rcf for 45 minutes) them in an aerobic environment. Lawns of recipient strains were generated using tube-less agar overlay approach using 130 μL of overnight culture with 3.2 mL of molten top agar, and 32 mL bottom agar plates, for each media respectively (Kauffman and Polz 2018). Lawns of recipient strains were prepared at hour 74, 75 and 76 for BHIS, BPRM and BPRM+Bile respectively. Waiting for 20 minutes until top agar solidified, 4 μL of donor filtrates were pipetted onto the surface of each recipient lawn. Following drying of the drop spots, the plates were transferred to incubator at

37 °C in an anaerobic chamber to form phage plaques. Counting results are summarized in **Table S6.**

Identification of mutations in publicly available metagenomes

Four datasets were collected: the Human Microbiome Project (Lloyd-Price *et al.*, 2017) (536 samples from 250 subjects; http://hmpdacc.org), the TwinsUK study (Xie *et al.*, 2016) (250 subjects; ERP010708), a Chinese type 2 diabetes study (Qin *et al.*, 2012) (368 subjects; SRA045646 and SRA050230) and a Chinese liver cirrhosis study (Qin *et al.*, 2014) (237 subjects; ERP005860). These datasets were chosen because they are deeply sequenced, have large sample sizes and have comparable collective sample sizes from both Western countries and China (**Figures S6C-S6E**). For each sample, metagenomic reads were filtered and aligned to the *B. fragilis* reference genome (NCTC9343) as above. For HMP subjects with multiple samples, only the sample with highest average coverage over *B. fragilis* genome was included. Alignment information for positions previously identified as *de novo* SNPs or inter-lineage SNPs were examined across metagenomes (56,272 SNP positions). Samples with average sequencing coverage <1 or with potential multiple-lineage colonization (>3% of positions with major allele frequency <95%) were discarded. In total, 347 samples passed our filters (n=90, 81, 100, and 76 for the four datasets, respectively). To minimize false positive polymorphisms emerging from homologous regions in other organisms, for each sample, genomic positions with average mapping quality < 41.9 (>95% of reads having maximum mapping quality) or with coverage outside the 1%-99% quantile of genome-wide coverage were masked. For the Q100P mutation position from BF2755 (nucleotide position 3213109 in the NCTC9343 genome), 288 of the 347 samples met our filters. For a given sample, a variable position was defined as polymorphic if the major allele frequency was between 50% and 95%.

We also searched for other potential mutations under population-specific selective pressure. We examined SNP positions in which >80% samples had sufficient mapping quality and more than 1 read covering that position (23,395 SNP positions in total, also used to build phylogeny in **Figure 6B**). We did not find SNPs with a comparable signal to the Q100P mutation (**Figure S6B**)

Quantification and statistical analysis

Statistical significance was calculated using Fisher's exact text, Mann-Whitney U-test, Chi-squared test, Binomial test and simulations as reported in the text.

Data and software availability

FASTQ files for the 602 *B. fragilis* isolates and the 667 targeted amplicon sequencing reactions, with adaptors removed and filtered for quality, as well as the BAM files of the 352 metagenomes aligned to *B. fragilis* lineage assemblies, are available from NCBI Sequence Read Archive (BioProject PRJNA524913). Commented MATLAB and Python scripts are available at https://github.com/shijiezhao/Within-person-evolution-of-Bacteroides-fragilis.

2.6 Figure titles and legends

Figure 1 | Each subject's *B. fragilis* population is dominated by a single lineage.

(**A**) Phylogenetic reconstruction shows that isolates cluster by subject (n=602), with one outlier isolate from Subject 10. Isolates are colored according to subject. **(B)** Isolates from same subjects generally differ by < 100 single nucleotide differences (SNPs) while isolates from different subjects differ by >10,000 SNPs. Mutational distances between all pairs of isolates. *Inset*: Intra-subject pairs separated by >18,000 SNPs all involve the outlier isolate from Subject 10.

Figure 2 | *B. fragilis* lineages diversify for years in healthy individuals via *de novo* SNPs and MEDs.

(**A**) The phylogeny of isolates from L05 is shown as an example, demonstrating both SNP and mobile element differences (MEDs; see also **Figures S1-S3**). Thin lines connect each isolate to a colored circle, which indicates the timepoint of isolation. Relative coverage (compared to the mean genomewide) across two MEDs is also shown. (**B**) The number of SNPs and MEDs identified for each lineage. (**C-D**) Estimate of the *B. fragilis* molecular clock using two different methods. (**C**) Each shape represents the average number of new SNPs per isolate from the indicated timepoint not present in the set of SNPs at initial sampling. (**D**) Estimate of molecular clock using root-to-tip distances for L01 only. (**E**) Distance and inferred time to most recent common ancestor at initial sampling (dMRCA and tMRCA, respectively). Gray dots represent individual isolates and bars represent averages. For L08, purple dots represent hypermutator isolates, and the average presented excludes these. (**F**) The spectrum of mutations in the hypermutator sublineage (purple) differs substantially from that of the normal sublineages of L08 (yellow) and 11 other lineages (gray). Error bars represent standard deviation across the 11 other lineages. *Inset*: Phylogeny for L08.

Figure 3 | Mobile elements are transferred within the microbiome of individual people.

(**A-B**) The phylogeny of isolates from L04 illustrates the gain of MED04-1 over time. Shading reflects the average relative coverage of the MED (compared to the mean genomewide). (**B**) Average relative coverage across the length of MED04-1 for different samples (46 isolates and 2 metagenomes). Colors are as indicated in (**A**). (**C-D**) Combining isolate whole genomes and metagenomes reveals an inter-species mobile element transfer event. (**C**) Isolates from L01 (n=187) show ~1X relative coverage of a putative integrative conjugative element (ICE), while isolates from other 11 lineages show relative coverage close to zero (n=415). Metagenomic libraries from all time points of L01 (n=206) show high relative coverage of this (ICE). (**D**) A rooted parsimonious phylogeny of the putative ICE across 4 species. Isolates with identical ICE sequences from a same phylogenetic group were merged into a single node (see also **Figures S1B-S1C**).

Figure 4 | Genes involved in polysaccharide utilization and cell envelope biosynthesis undergo parallel adaptive evolution within individual subjects.

(**A**) An example gene under parallel evolution from L02 is shown, demonstrating that observed mutations are of independent origin and occur in distinct isolates. Nodes represent individual isolates and are colored by sampling dates. (**B**) A total of 16 genes were identified as undergoing parallel evolution in the 12 lineages. These 16 genes are grouped by inferred function (**Table 5**). Each dot in the table represents an independent mutation event, colored by type of mutation. (**C**) The number of genes mutated in parallel within at least one lineage deviates significantly from neutral simulations ($P<0.001$, Methods). (**D**) A classic signature of selection, dN/dS, indicates adaptive evolution in genes under parallel evolution ($P<0.001$, Binomial test), but not for other genes mutated within subjects. Mutations across lineages show a significant signature of purifying selection ($P<0.001$, Binomial test). Error bars represent 95% confidence intervals. (**E**) Mutations in SusC homologs under selection were enriched at the interface between the proteins ($P< 0.001$, Binomial test, Methods).

Figure 5 | Evolutionary dynamics over a 1.5 year sampling period reveals a steady increase in mutational frequencies and a stable coexistence of two sublineages.

(**A-C**) We combined 206 stool metagenomes and 187 isolate whole genomes to infer evolutionary dynamics within L01. (**A**) Branches with at least 4 isolates are labeled with colored squares that represent individual SNPs. One SNP was inferred to have happened twice and is indicated in both locations (purple). (**B**) Frequencies of labeled SNPs were inferred from metagenomes. Circles represent SNP frequencies inferred from isolate genomes. (**C**) We combined these data types to infer the trajectory of sublineages prior to and during sampling. Sublineages are labeled with names and colored as in (**A**). The two major sublineages, SL1 and SL2, are separated by dashed lines. Black diamonds represent transient SNPs from genes presented in **Figure 4**. (**D**) The identity of SNPs shown in (**B-C**). SNPs in the 16 genes under positive selection are bolded and transient mutations in these genes are indicated with parentheses. Negative numbers indicate mutations upstream of the start of the gene. (**E**) All isolates from SL2 (n=76), but only 13% from SL1 (n=111) carry putative prophage MED01-2. (**F-H**) Relative abundances of pairs of isolates during competition assays, over two rounds of passages. Dashed lines represent 1:100 dilution at hour 18. Each line represents the average of 3 technical replicates, and error bars represent standard error of the mean.

Figure 6 | Comparison to published metagenomes reveals a mutation that emerges independently and frequently in Western, but not Chinese populations

(**A**) We examined the prevalence of a common amino acid change in available metagenomes. The percentage of metagenome samples with a polymorphism or fixed proline at this position was greater in Western populations than in Chinese populations (n=152, 136 respectively). Error bars represent standard error. (**B**) A neighbor-joining phylogeny of inferred *B. fragilis* genotypes within public metagenomes demonstrates that this mutation emerged independently and repeatedly. Phylogeny is shown as a dendrogram to better visualize the independent emergence of Q100P mutations. **(C)** Between lineages, 12 genes under parallel evolution and with inter-lineage mutations show significant signatures of purifying selection (dN/dS>1, Binomial test, Methods). BF2755 does not show signs of purifying selection for inter-lineage mutations. This analysis represents tens of thousands of years of evolution (Methods), in contrast to **Figure 4D**. Error bars represent 95% confidence interval. The dashed line represents the average dN/dS for all inter-lineage SNPs. (**D**) Four models that could account for the discrepancy of natural selection at different timescales.

Table 1 | Estimation of the number of mutations occurring daily within the human microbiome

Number of bacteria (cells/microbiome) (Sender *et al.*, 2016)	**Mutation rate (SNP/nucleotide/replication) (Barrick and Lenski 2013)**	**Bacterial genome size (nucleotide/cell) (Nayfach and Pollard 2015)**	**Range of replication rate (replication/day) (Korem *et al.*, 2015)**	**→**	**Estimated number of *de novo* mutations (SNP/microbiome/day)**
10^{13}-10^{14}	10^{-10}-10^{-9}	2-6×10^{6}	1-10		2×10^{9} - 6×10^{12}

Figure S1 | Within-person evolution of *Bacteroides* from Subject 01, Related to Figure 2

(A) The phylogeny for isolates from *B. fragilis* is shown. Colored circles represent isolates from samples collected at the indicated dates. For each isolate, the relative coverage across identified MEDs is shown. Shading of MED regions reflects the average relative coverage of the MED in that isolate. Red stars indicate when the same nucleotide mutation emerged multiple times within the same lineage (inferred via parsimony, Methods). More details on the exact mutations and MEDs found are in Table S3. *Inset:* dMRCA values across sampling times. (B-C) Analysis of the integrative conjugative element (ICE) found to be transferred in Subject 01, identified from two candidate interspecies transfer regions (IST01-1 and IST01-2, Methods). (B) A phylogeny was constructed for all *B. vulgatus* isolates cultured from Subject 01, using a publicly available reference genome (GCF_000012825.1) and the same parameters and methods for *B. fragilis* SNP identification and evolutionary inference. (C) A phylogeny was built using reads aligned to the ICE from all isolates of 4 *Bacteroides* species from Subject 01 (Figure 3D). The sequences of IST01-1 and IST01-2 in the L01 assembly were used as the reference and the same methods were used as for *B. fragilis* SNP evolutionary inference. Among the 4 SNPs identified, we found 2 SNP locations whose 200-bp flanking sequence had matches in NCBI with >85% similarity, and we used these alleles as outgroups to root the tree. For the remaining 2 SNP locations, we assigned ancestral alleles that minimized the variance of dMRCA of all isolates. Colors represent isolates from the same phylogenetic group. The consensus ICE sequence in the L01 *B. fragilis* genome is represented by a single circle (black). We note that three SNPs were identified within this ICE in *B. fragilis* L01, each in a single isolate.

Figure S2 | Within-person *B. fragilis* evolution in L02-L07, Related to Figure 2

(**A-F**) The phylogeny for isolates from L02 to L7, respectively. Colored circles represent isolates from samples collected at the indicated dates. For each isolate, the relative coverage across identified MEDs is shown. Shading of MED regions reflects the average relative coverage of the MED in that isolate. Dark green diamonds indicate SNPs associated with putative sweeps and are labeled with gene ID and type of mutation. Within-sample dMRCA changes over time are shown adjacent to the phylogeny. In (**F**), the SNP that was shared by all isolates from the latest time point (dark blue) was not included as a sweep because it might be an artifact of undersampling at the later time point (**Figure S7G**). More details on the exact mutations and MEDs identified from these lineages are in **Table S3.**

Figure S3 | Within-person *B. fragilis* evolution in L08-L12, Related to Figure 2

(**A-E**) The phylogeny for isolates from L08 to L12, respectively. All lineages were sampled once. For each isolate, the relative coverage across identified MEDs is shown. Shading of MED regions reflects the average relative coverage pattern of the MED in that isolate. Red stars indicate when the same nucleotide mutation emerged multiple times within the same lineage (inferred via parsimony, Methods). (**D**) The presence/absence pattern of MED11-1 suggests many loss events on the phylogeny. More details on the exact mutations and MEDs identified from these lineages are in **Table S6** and **Table S3**.

Figure S4 | Search for parallel evolution across lineages did not yield additional genes under selection, Related to Figure 4

We searched for genes mutated multiple times across lineages, counting the number of total SNPs obtained in each gene (M), the number of lineages a gene was mutated in (n), and the maximum number of mutation a given gene was mutated in any lineage (m_{max}). Simulations were performed as described in the Methods. (**A**) A search with the criteria of $M \geqq 2$ yielded results consistent with a null model. (**B**) When this threshold was increased to $M \geqq 3$, 11 genes were observed. Interestingly, 9 of these genes were already discovered with the criteria used in the main text, $m_{max} \geqq 2$. The 2 genes that are newly discovered with this metric ($m_{max}<2$ & $M \geqq 3$) do not show a signal for positive selection (**F**). (**C-D**) Similar results were obtained for the metric n, with the only 2 new genes discovered being identical to the analysis in (**A-B**). Further, dN/dS of genes discovered with the n metric did not show a significant signal for adaptive evolution (**F**). (**E**) The number of intergenic mutations is consistent with a null model. (**F**) dN/dS calculated across groups of genes defined with various metrics for parallel evolution. Together, these results are consistent with the evidence of person-specific selection forces found in the main text and suggest that when a selection pressures is shared across subjects, it can usually be detected from just studying a single subject.

Figure S5 | Evolutionary dynamics of L01 and L03 and the phage-mediated competition between L01 sublineages; Related to Figure 5

(**A**) For each metagenome from stool samples from Subject 01, we calculated the percentage of metagenomic reads that aligned to the L01 genome assembly and plotted it against the time of sample collection. Reads potentially from other species (in regions with >5X median coverage) were excluded. This percentage estimates the relative abundance of *B. fragilis* in the stool community. The black line indicates the mean across samples. (**B**) For each sample, the ratio of SL1:SL2 was estimated using total number of reads aligned to alleles corresponding to either sublineage at the SNPs that separate them. Samples with fewer than 40 reads aligned to these SNP locations were excluded. The black line indicates the mean across samples. (**C**) The relative abundance of L03 *B. fragilis* inside Subject 03 was estimated in 74 metagenomes spanning 144 days, using the same method described in (**A**). (**D**) The phylogeny of isolates from L03. Branches with at least 3 isolates are labeled with colored octagons that represent individual SNPs. Circles represent individual isolates and are colored according to sampling date. (**E**) Frequencies of labeled SNPs over time in the *B. fragilis* population were inferred from 74 stool metagenomes (Methods). Colored circles represent SNP frequencies inferred from isolate genomes at particular time points. (**F**) The evolutionary history of sublineages during sampling was inferred (see Methods). Sublineages are defined by their signature SNPs and labeled with the identity of SNPs and colored as in (**D**). (**G-J**) We picked one SL1-a-1 isolate, one SL-1-a-1-1 isolate and two SL2 isolates to perform competition experiments. Neither of the two SL1 isolates carried MED01-2. We performed multiple competitions and treated those with different SL2 isolates as biological replicates. Saturated and synchronized pure cultures of the indicated isolates were mixed at the indicated ratios diluted 1:100 to begin the competition. Relative abundances were estimated using targeted amplicon sequencing (Methods), and OD measurements were used to convert these to absolute abundances. Absolute abundances are displayed as the average of replicates (top panels) and relative abundances are displayed separately for each replicate (bottom panels). (**G**) Competition between SL1-a-1 and SL1-a-1-1, both were MED01-2-, showed stable coexistence over 22 hours. (**H-I**) Both SL1-a-1-1 isolate and SL1-a-1 were outcompeted by SL2 within 22 hours. (**J**) In the trio competition, both SL1-a-1 and SL1-a-1-1 were outcompeted by SL2 and their ratio did not change over time, suggesting that SL1-a-1-1 did not have obvious advantage over SL1-a-1 in this experimental setting. (**K**) Growth curves for pure cultures and competition co-cultures show that mixtures of SL1 and SL2 had slower overall growth than pure cultures, suggesting actively killing of SL1 by SL2. (**L**) Phage plaque assay showed that the isolates with MED01-2+ formed phage plaques on isolates that are MED01-2-. Each dot represents the number of plaques formed for a distinct donor-recipient pair, color coded by the media the recipient was grown on (Methods). Results are grouped by the donor-recipient pair. The difference between the D+,R- group and each of the other four groups are all significant ($P < 5\times10^{-12}$, Mann-Whitney U test). Between the other four groups, there are no significant differences ($P > 0.15$, Mann-Whitney U test).

Figure S6 | BF755 Q100P difference between Chinese and Western populations are robust to subject health conditions and are the most significant difference; Related to Figure 6

(A) For all four datasets, we inferred the total fraction of samples with Q100P polymorphism or fixed for subjects with different disease conditions. The HMP consists of healthy subjects. TwinsUK subjects are elderly people and a small fraction of them are diagnosed with diabetes. For the two Chinese studies, patients and healthy controls were plotted separately. Light gray represents the fraction of samples with Q100P polymorphism, and dark gray represents fraction of samples with P mutation fixed (stacked bar chart). Error bars represent standard error of percentage of samples with either fixed or polymorphic mutation. We do not find any association between subject health and the prevalence of Q100P mutation. **(B)** Manhattan plot shows that the Q100P mutation in BF2755 is the only mutation that is under differential selective pressure between Western and Asian populations. Each dot represents the p-value of a Fisher's exact test of a variable position on the B. fragilis genome, comparing the number of samples with polymorphism between Western (TwinsUK and HMP) and Chinese metagenome datasets. Gray dots are synonymous mutations positions while blue dots are non-synonymous mutation positions, the red dot represents Q100P mutation from gene BF2755. **(C)** Among samples passing filters, those that were polymorphic did not have different sequencing coverage relative to those with the ancestral allele (Q, p=0.37, Mann-Whitney test). Samples with all reads pointing to P had slightly lower coverage comparing to samples with the ancestral allele (p=0.03, Mann-Whitney test). (**D**) Western samples and Chinese samples have similar overall coverage (p=0.49, Mann-Whitney test). (**E**) Coverage over the Q100P position is comparable with genomewide average coverage for the included metagenome samples. Colors scheme is the same with panel (**C**).

Figure S7 | Collector curves suggest sufficient sampling for dMRCA, yet numbers of SNPs identified depends on number of isolates collected, Related to STAR Methods

(**A-L**) For each lineage and time point, we created a collector curve for dMRCA (one curve if the lineage was sampled once). For an isolate population from a particular time point, we subsampled the population to x isolates (0<x<n, n = total number of isolates at the time point), reconstructed the MRCA, and recomputed dMRCA. For each x, we simulated 100 subsamples and computed the mean (dots) and standard deviation (bars) for the simulation results. dMRCA was undersaturated only in 2 time points from L07 (0 and 168 Days). (**M-X**) For each lineage and time point, we created a collector curve for the number of SNPs identified (one curve if the lineage was sampled once). For an isolate population from a particular time point, we subsampled the population to x isolates (0<x<n, n = total number of isolates at the time point), and recomputed the number of SNPs identified. For each x, we simulated 100 subsamples and computed the mean (dots) and standard deviation (bars) for the simulation results.

Figure 1

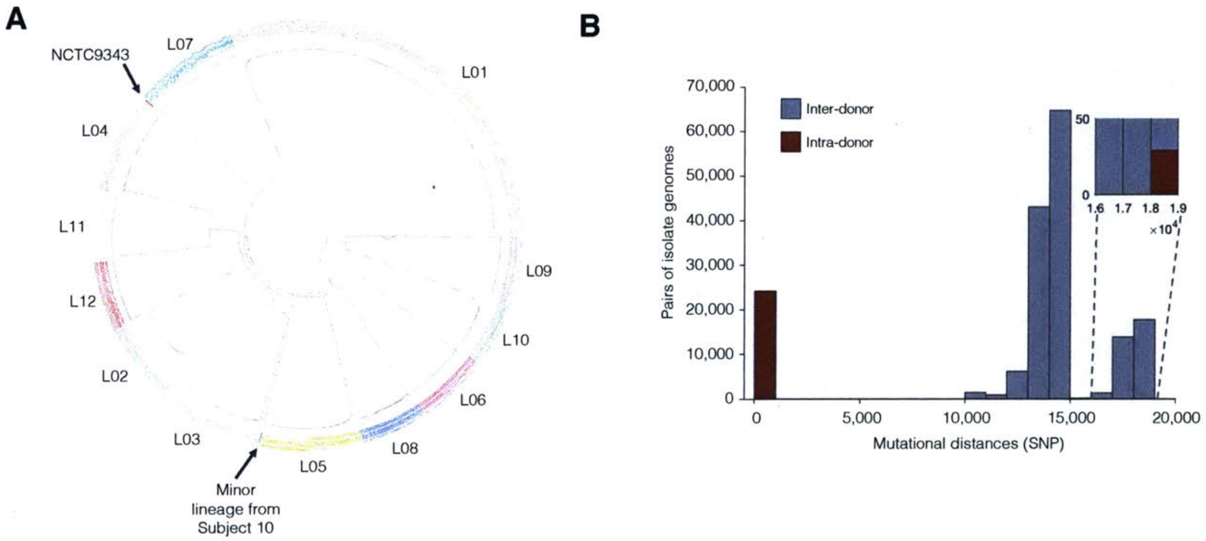

A
NCTC9343
L07
L01
L04
L11
L09
L12
L10
L02
L06
L03
L08
L05
Minor lineage from Subject 10
B
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
Pairs of isolate genomes
Inter-donor
Intra-donor
50
0
1.6
1.7
1.8
1.9
×10⁴
0
5,000
10,000
15,000
20,000
Mutational distances (SNP)

Figure 2

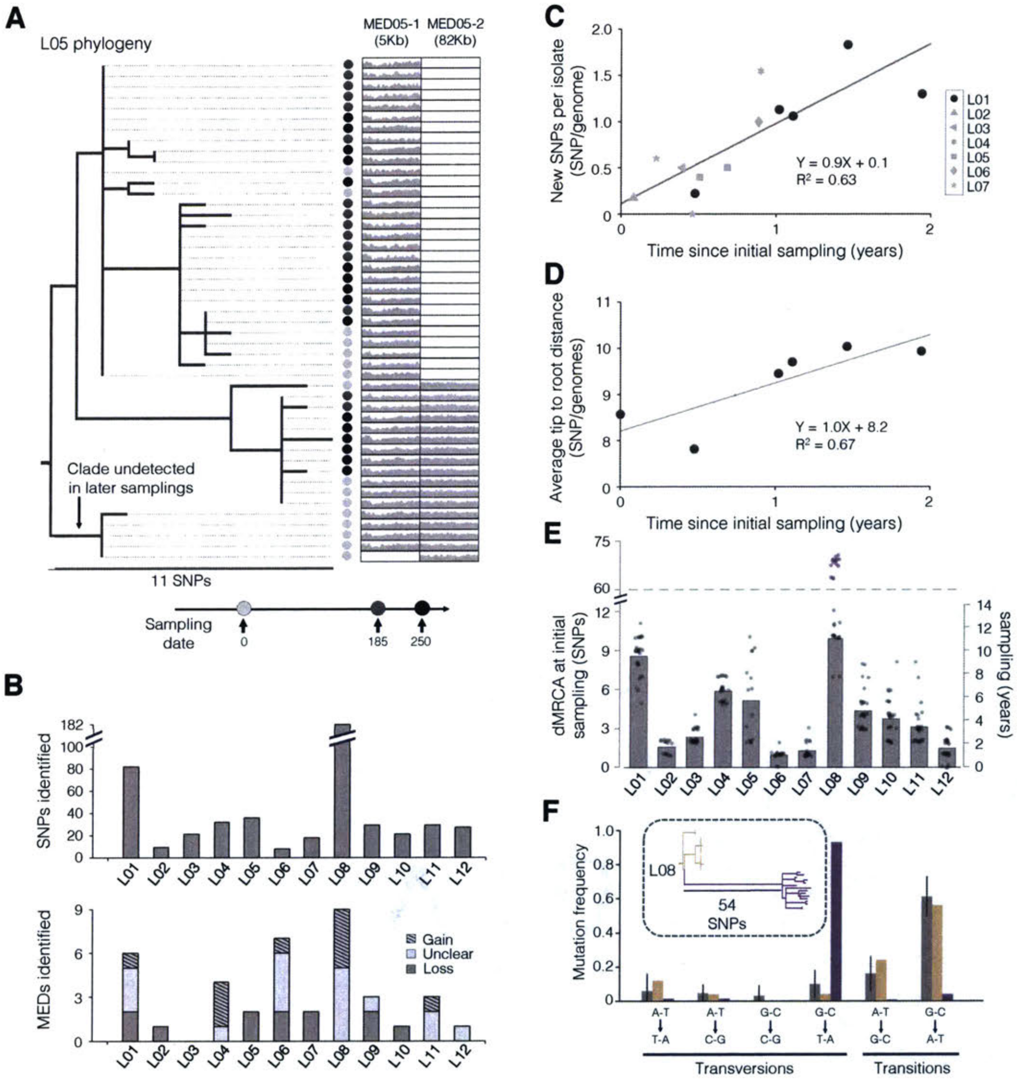

A
L05 phylogeny
MED05-1 (5Kb)
MED05-2 (82Kb)
Clade undetected in later samplings
11 SNPs
Sampling date
0
185
250
B
SNPs identified
182
100
80
60
40
20
0
L01 L02 L03 L04 L05 L06 L07 L08 L09 L10 L11 L12
MEDs identified
9
6
3
0
Gain
Unclear
Loss
L01 L02 L03 L04 L05 L06 L07 L08 L09 L10 L11 L12
C
New SNPs per isolate (SNP/genome)
2.0
1.5
1.0
0.5
0
Y = 0.9X + 0.1
R² = 0.63
L01
L02
L03
L04
L05
L06
L07
0
1
2
Time since initial sampling (years)
D
Average tip to root distance (SNP/genomes)
11
10
9
8
0
Y = 1.0X + 8.2
R² = 0.67
0
1
2
Time since initial sampling (years)
E
75
60
12
9
6
3
0
dMRCA at initial sampling (SNPs)
tMRCA at initial sampling (years)
14
12
10
8
6
4
2
0
L01 L02 L03 L04 L05 L06 L07 L08 L09 L10 L11 L12
F
Mutation frequency
1.0
0.8
0.6
0.4
0.2
0
L08
54 SNPs
A-T → T-A
A-T → C-G
G-C → C-G
G-C → T-A
A-T → G-C
G-C → A-T
Transversions
Transitions

Figure 3

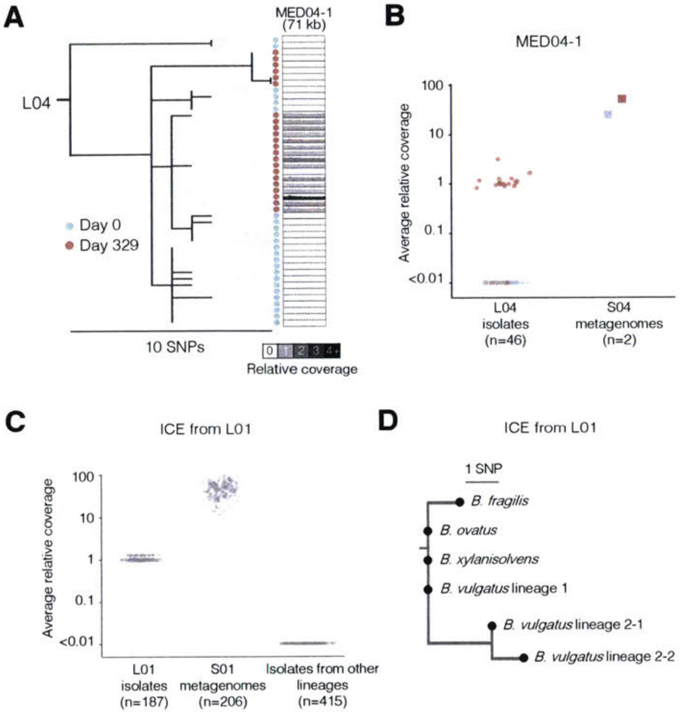

A
MED04-1
(71 kb)
L04
Day 0
Day 329
10 SNPs
0
1
2
3
4+
Relative coverage
B
MED04-1
Average relative coverage
100
10
1
0.1
<0.01
L04
isolates
(n=46)
S04
metagenomes
(n=2)
C
ICE from L01
Average relative coverage
100
10
1
0.1
<0.01
L01
isolates
(n=187)
S01
metagenomes
(n=206)
Isolates from other
lineages
(n=415)
D
ICE from L01
1 SNP
B. fragilis
B. ovatus
B. xylanisolvens
B. vulgatus lineage 1
B. vulgatus lineage 2-1
B. vulgatus lineage 2-2

Figure 4

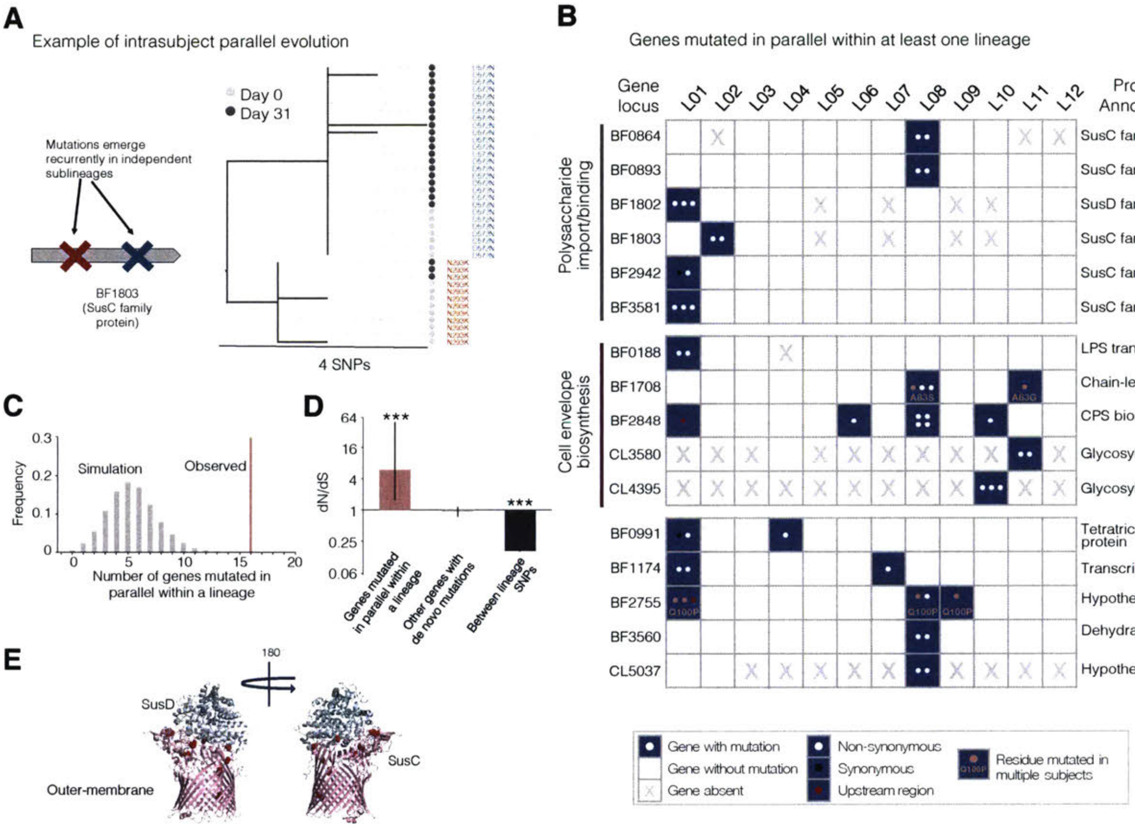

A
Example of intrasubject parallel evolution
Day 0
Day 31
Mutations emerge recurrently in independent sublineages
BF1803 (SusC family protein)
4 SNPs
B
Genes mutated in parallel within at least one lineage
Gene locus
L01 L02 L03 L04 L05 L06 L07 L08 L09 L10 L11 L12
Protein Annotation
Polysaccharide import/binding
BF0864 SusC family protein
BF0893 SusC family protein
BF1802 SusD family protein
BF1803 SusC family protein
BF2942 SusC family protein
BF3581 SusC family protein (ccfC)
Cell envelope biosynthesis
BF0188 LPS transporter (msbA)
BF1708 Chain-length determinator (cps4)
A83S
A83G
BF2848 CPS biosynthesis protein (ungD2)
CL3580 Glycosyltransferase (wffD)
CL4395 Glycosyltransferase
BF0991 Tetratricopeptide repeat protein
BF1174 Transcriptional regulator
BF2755 Hypothetical protein
Q100P
BF3560 Dehydratase (yhaM)
CL5037 Hypothetical protein
Gene with mutation
Gene without mutation
Gene absent
Non-synonymous
Synonymous
Upstream region
Residue mutated in multiple subjects
C
Frequency
0.3
0.2
0.1
0
Simulation
Observed
0 5 10 15 20
Number of genes mutated in parallel within a lineage
D
dN/dS
64
16
4
1
0.25
0.06
***
***
Genes mutated in parallel within a lineage
Other genes with de novo mutations
Between lineage SNPs
E
180
SusD
SusC
Outer-membrane

Figure 5

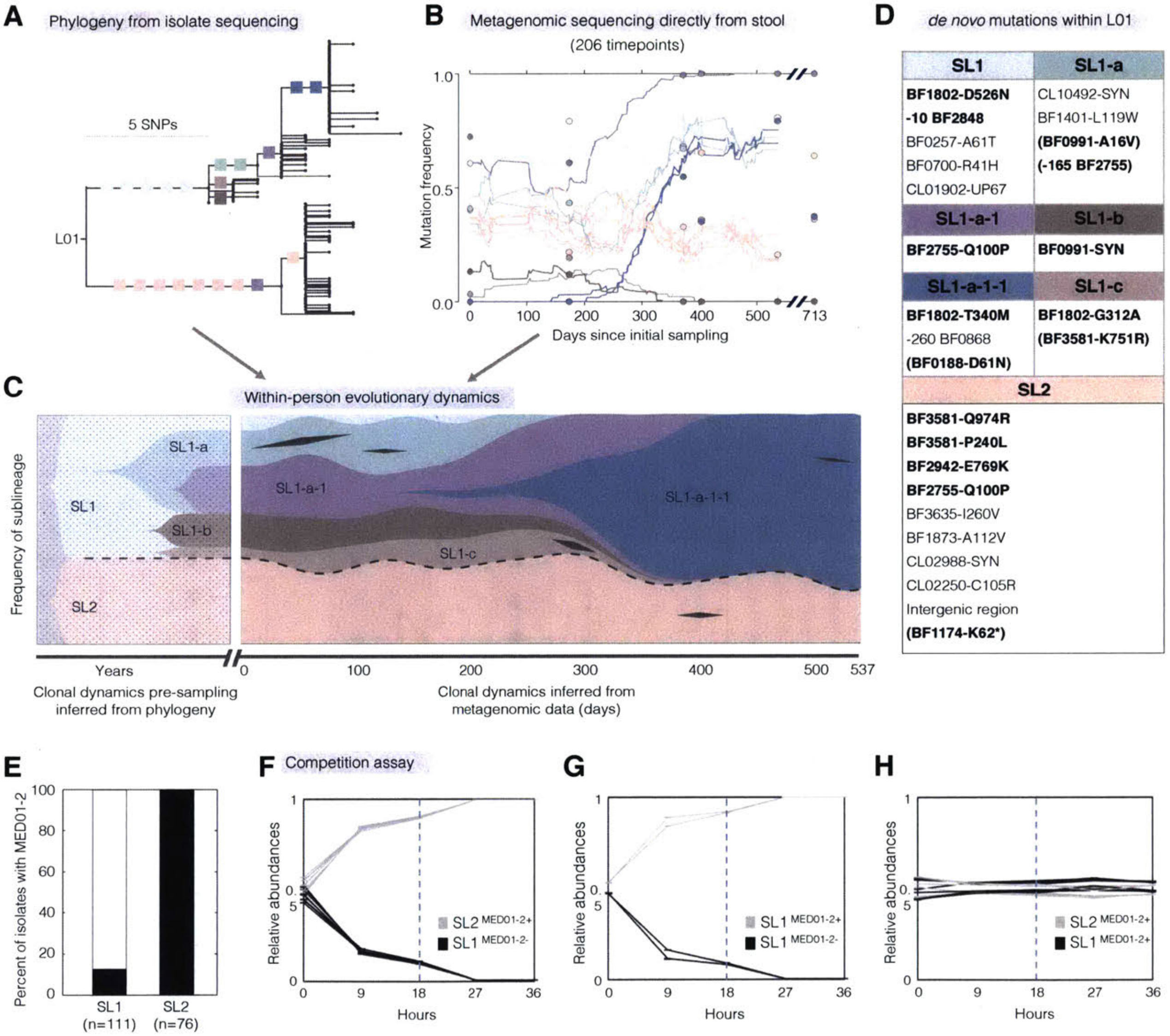

A
Phylogeny from isolate sequencing
5 SNPs
L01
B
Metagenomic sequencing directly from stool
(206 timepoints)
Mutation frequency
1.0
0.5
0.0
0
100
200
300
400
500
713
Days since initial sampling
D
de novo mutations within L01
SL1
BF1802-D526N
-10 BF2848
BF0257-A61T
BF0700-R41H
CL01902-UP67
SL1-a
CL10492-SYN
BF1401-L119W
(BF0991-A16V)
(-165 BF2755)
SL1-a-1
BF2755-Q100P
SL1-b
BF0991-SYN
SL1-a-1-1
BF1802-T340M
-260 BF0868
(BF0188-D61N)
SL1-c
BF1802-G312A
(BF3581-K751R)
SL2
BF3581-Q974R
BF3581-P240L
BF2942-E769K
BF2755-Q100P
BF3635-I260V
BF1873-A112V
CL02988-SYN
CL02250-C105R
Intergenic region
(BF1174-K62*)
C
Within-person evolutionary dynamics
Frequency of sublineage
SL1-a
SL1
SL1-b
SL2
SL1-a-1
SL1-a-1-1
SL1-c
Years
Clonal dynamics pre-sampling
inferred from phylogeny
0
100
200
300
400
500
537
Clonal dynamics inferred from
metagenomic data (days)
E
Percent of isolates with MED01-2
100
80
60
40
20
0
SL1
(n=111)
SL2
(n=76)
F
Competition assay
Relative abundances
1
0.5
0
0
9
18
27
36
Hours
SL2 MED01-2+
SL1 MED01-2-
G
Relative abundances
1
0.5
0
0
9
18
27
36
Hours
SL1 MED01-2+
SL1 MED01-2-
H
Relative abundances
1
0.5
0
0
9
18
27
36
Hours
SL2 MED01-2+
SL1 MED01-2+

Figure 6

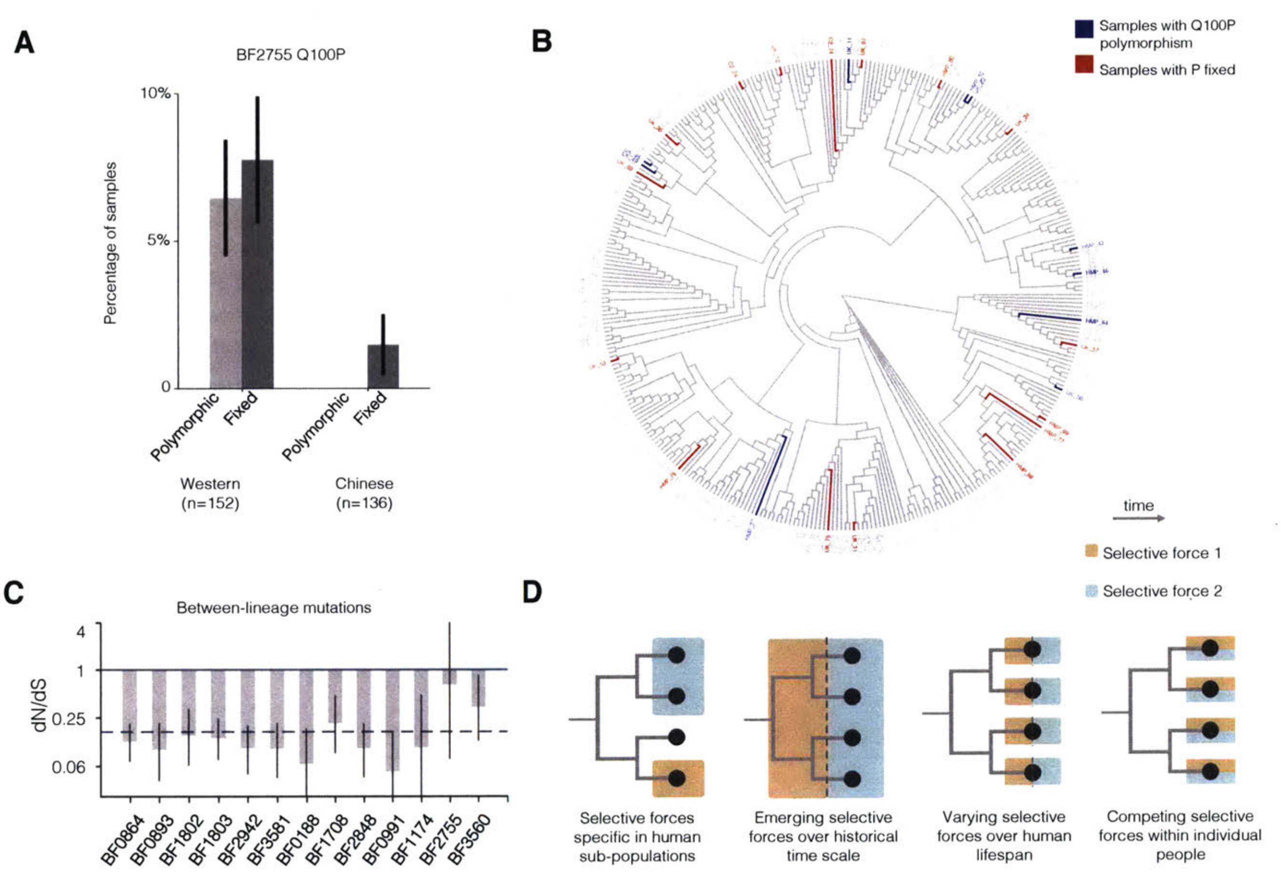

A
BF2755 Q100P
10%
5%
0
Percentage of samples
Polymorphic
Fixed
Polymorphic
Fixed
Western
(n=152)
Chinese
(n=136)
B
Samples with Q100P polymorphism
Samples with P fixed
C
Between-lineage mutations
4
1
0.25
0.06
dN/dS
BF0864
BF0893
BF1802
BF1803
BF2942
BF3581
BF0188
BF1708
BF2848
BF0991
BF1174
BF2755
BF3560
D
time
Selective force 1
Selective force 2
Selective forces specific in human sub-populations
Emerging selective forces over historical time scale
Varying selective forces over human lifespan
Competing selective forces within individual people

Figure S1

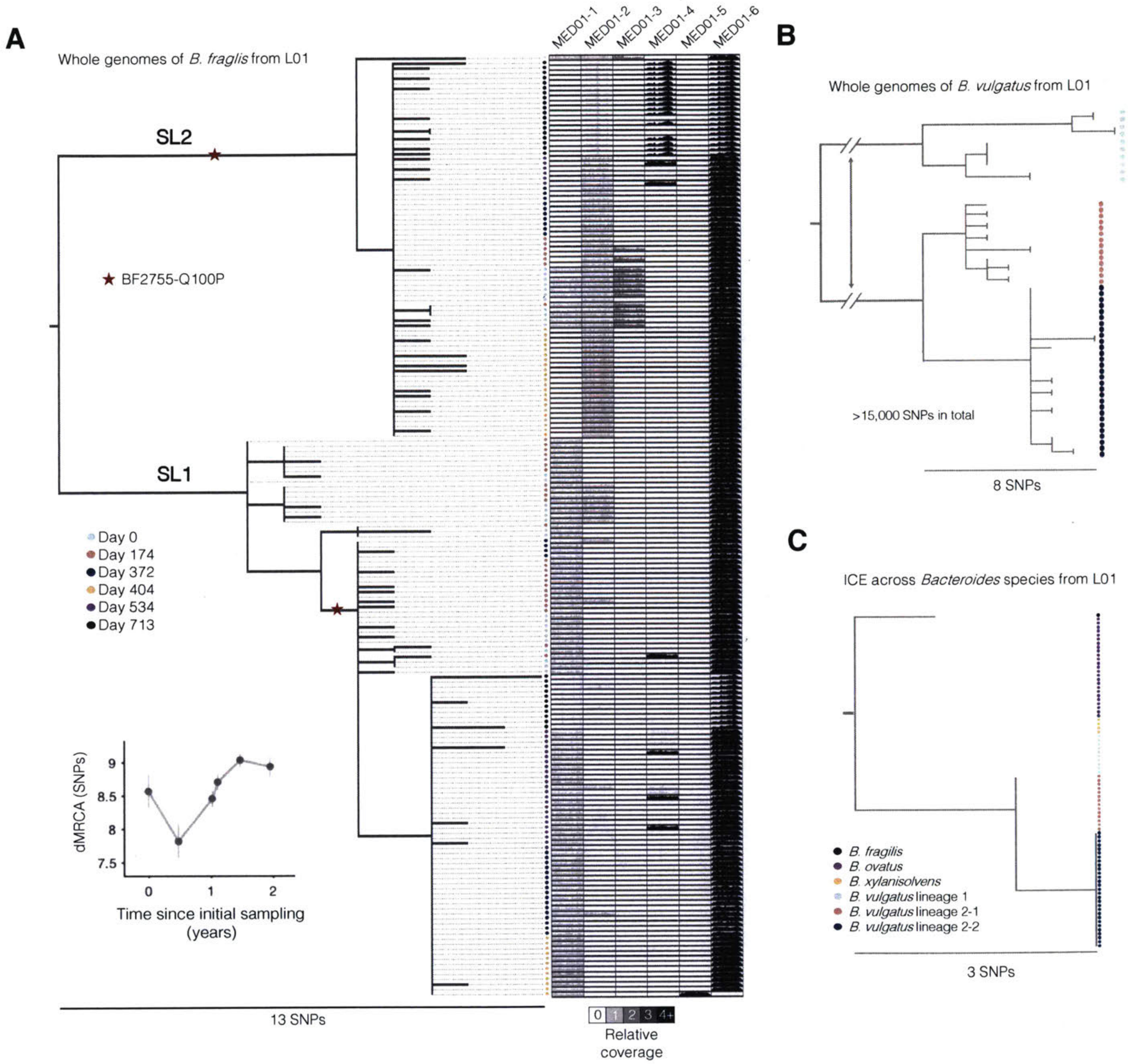

A
Whole genomes of B. fraglis from L01
MED01-1
MED01-2
MED01-3
MED01-4
MED01-5
MED01-6
SL2
★ BF2755-Q100P
SL1
Day 0
Day 174
Day 372
Day 404
Day 534
Day 713
dMRCA (SNPs)
9
8.5
8
7.5
0
1
2
Time since initial sampling (years)
13 SNPs
0 1 2 3 4+
Relative coverage
B
Whole genomes of B. vulgatus from L01
>15,000 SNPs in total
8 SNPs
C
ICE across Bacteroides species from L01
B. fragilis
B. ovatus
B. xylanisolvens
B. vulgatus lineage 1
B. vulgatus lineage 2-1
B. vulgatus lineage 2-2
3 SNPs

Figure S2

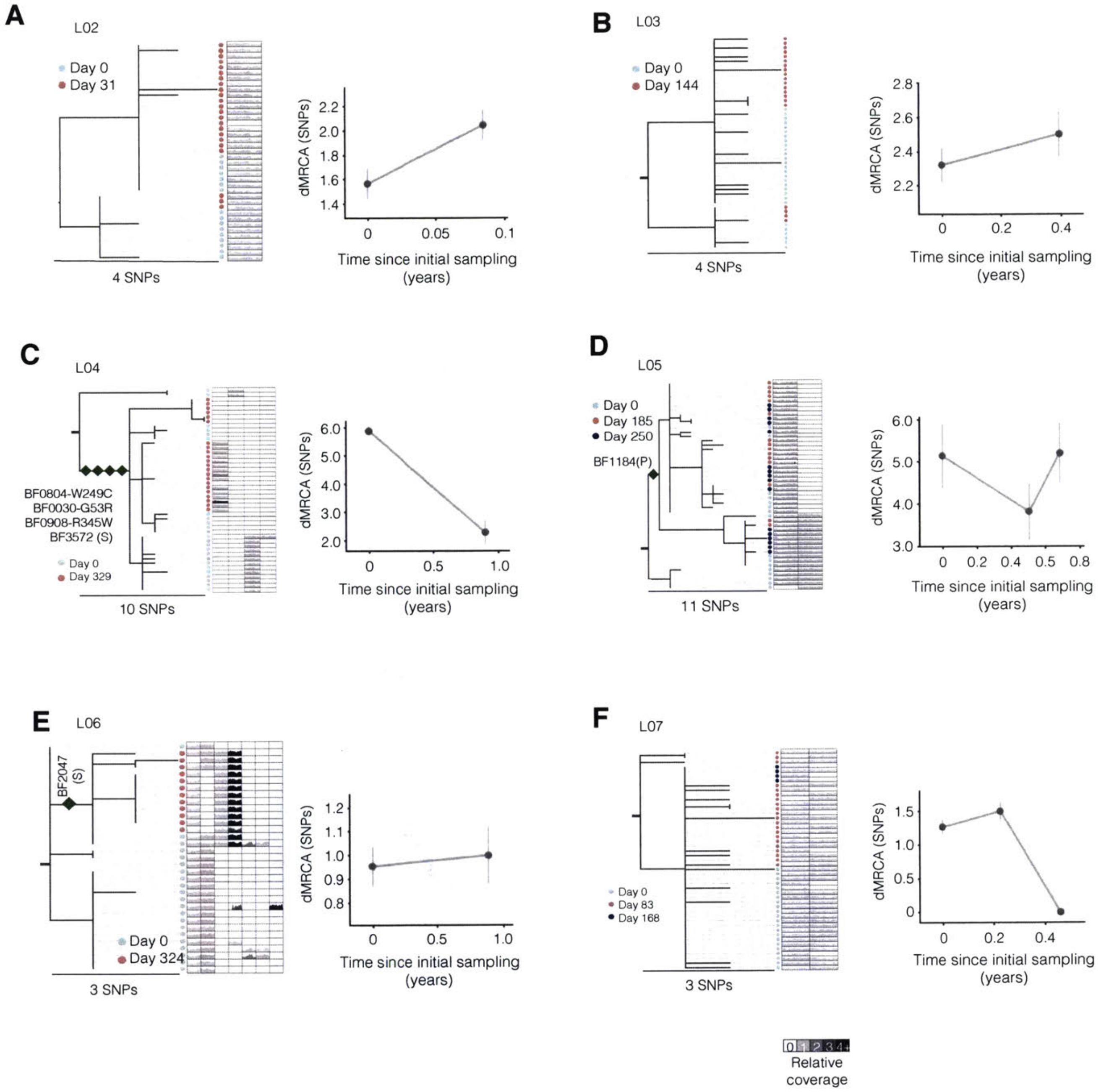

A
L02
Day 0
Day 31
4 SNPs
dMRCA (SNPs)
2.2
2.0
1.8
1.6
1.4
0
0.05
0.1
Time since initial sampling (years)
B
L03
Day 0
Day 144
4 SNPs
dMRCA (SNPs)
2.8
2.6
2.4
2.2
0
0.2
0.4
Time since initial sampling (years)
C
L04
BF0804-W249C
BF0030-G53R
BF0908-R345W
BF3572 (S)
Day 0
Day 329
10 SNPs
dMRCA (SNPs)
6.0
5.0
4.0
3.0
2.0
0
0.5
1.0
Time since initial sampling (years)
D
L05
Day 0
Day 185
Day 250
BF1184(P)
11 SNPs
dMRCA (SNPs)
6.0
5.0
4.0
3.0
0
0.2
0.4
0.5
0.8
Time since initial sampling (years)
E
L06
BF2047 (S)
Day 0
Day 324
3 SNPs
dMRCA (SNPs)
1.2
1.1
1.0
0.9
0.8
0
0.5
1.0
Time since initial sampling (years)
F
L07
Day 0
Day 83
Day 168
3 SNPs
dMRCA (SNPs)
1.5
1.0
0.5
0
0
0.2
0.4
Time since initial sampling (years)
0 1 2 3 4+
Relative coverage

Figure S3

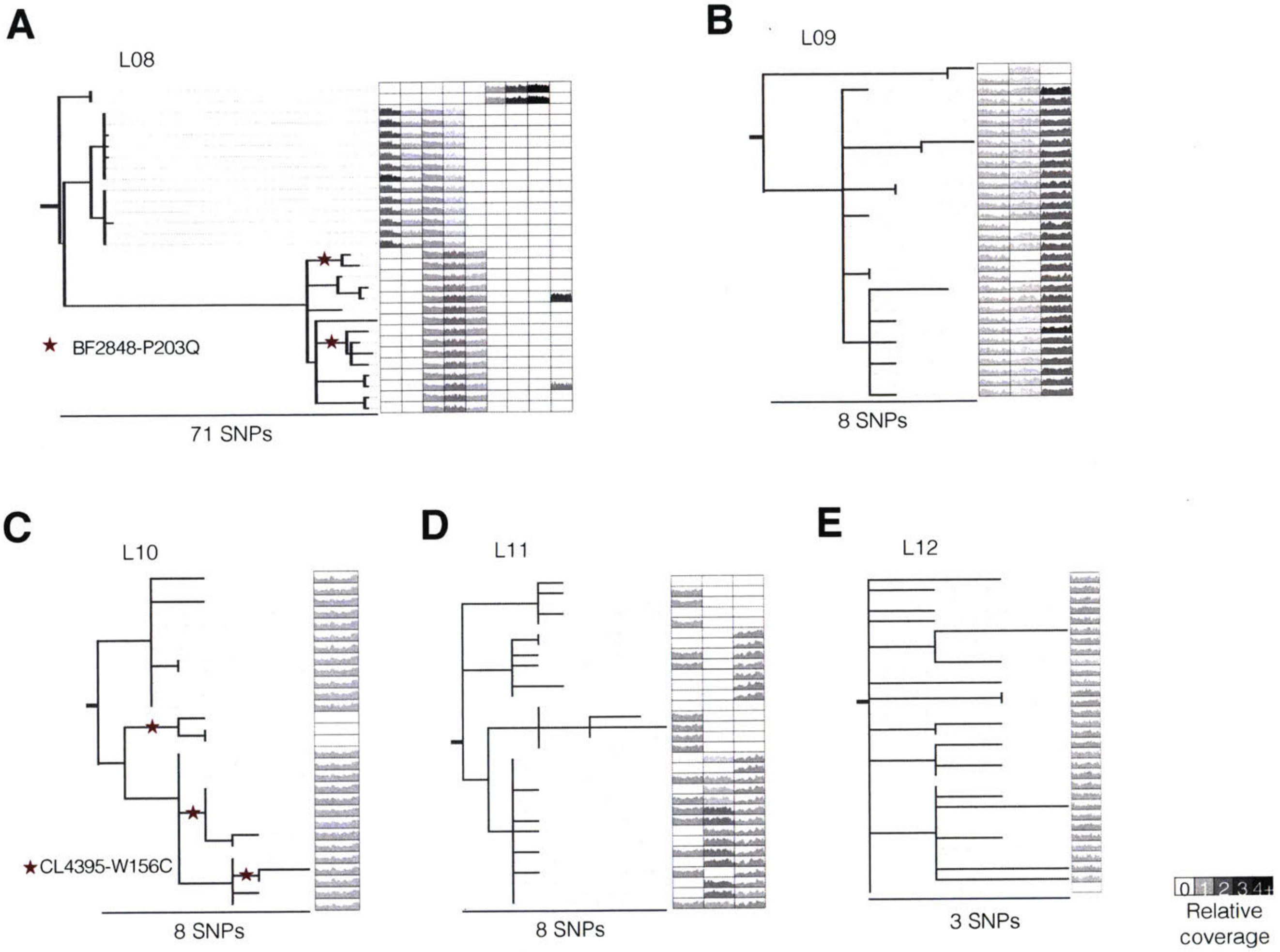

A
L08
★ BF2848-P203Q
71 SNPs
B
L09
8 SNPs
C
L10
★CL4395-W156C
8 SNPs
D
L11
8 SNPs
E
L12
3 SNPs
0 1 2 3 4+
Relative coverage

Figure S4

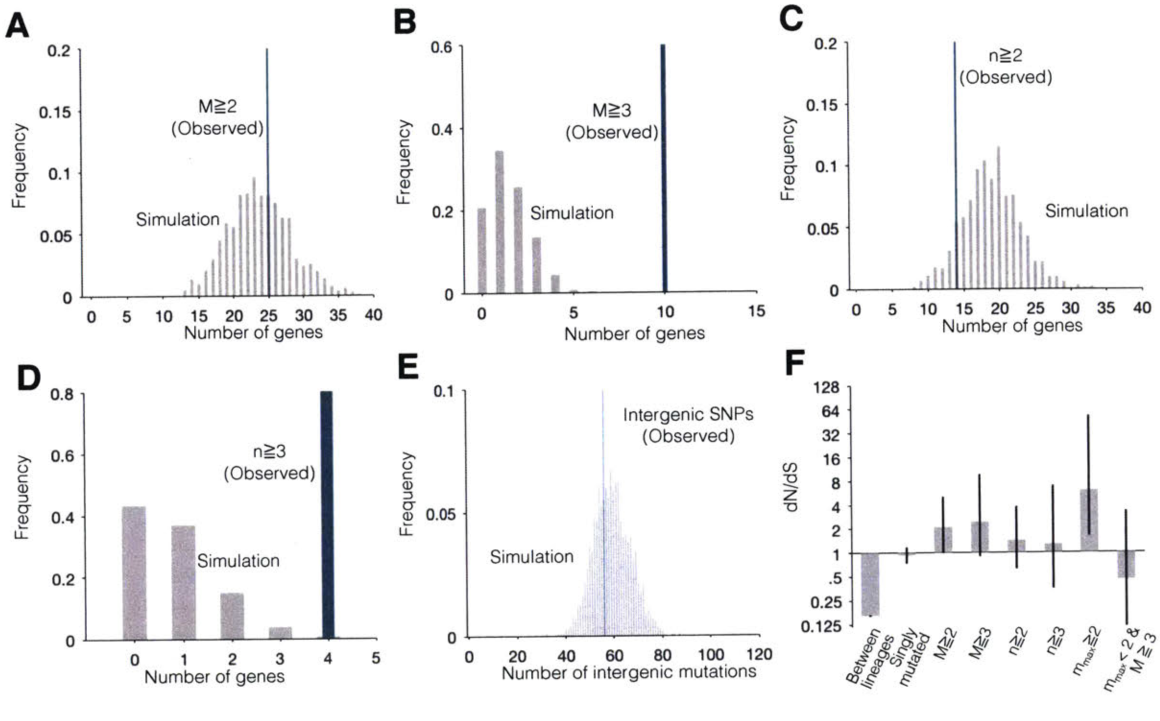

A
Frequency
0.2
0.15
0.1
0.05
0
M≧2
(Observed)
Simulation
0 5 10 15 20 25 30 35 40
Number of genes
B
Frequency
0.6
0.4
0.2
0
M≧3
(Observed)
Simulation
0 5 10 15
Number of genes
C
Frequency
0.2
0.15
0.1
0.05
0
n≧2
(Observed)
Simulation
0 5 10 15 20 25 30 35 40
Number of genes
D
Frequency
0.8
0.6
0.4
0.2
0
n≧3
(Observed)
Simulation
0 1 2 3 4 5
Number of genes
E
Frequency
0.1
0.05
0
Intergenic SNPs
(Observed)
Simulation
0 20 40 60 80 100 120
Number of intergenic mutations
F
dN/dS
128
64
32
16
8
4
2
1
.5
0.25
0.125
Between lineages
Singly mutated
M≧2
M≧3
n≧2
n≧3
m_{max}≧2
m_{max} < 2 & M ≧ 3

Figure S5

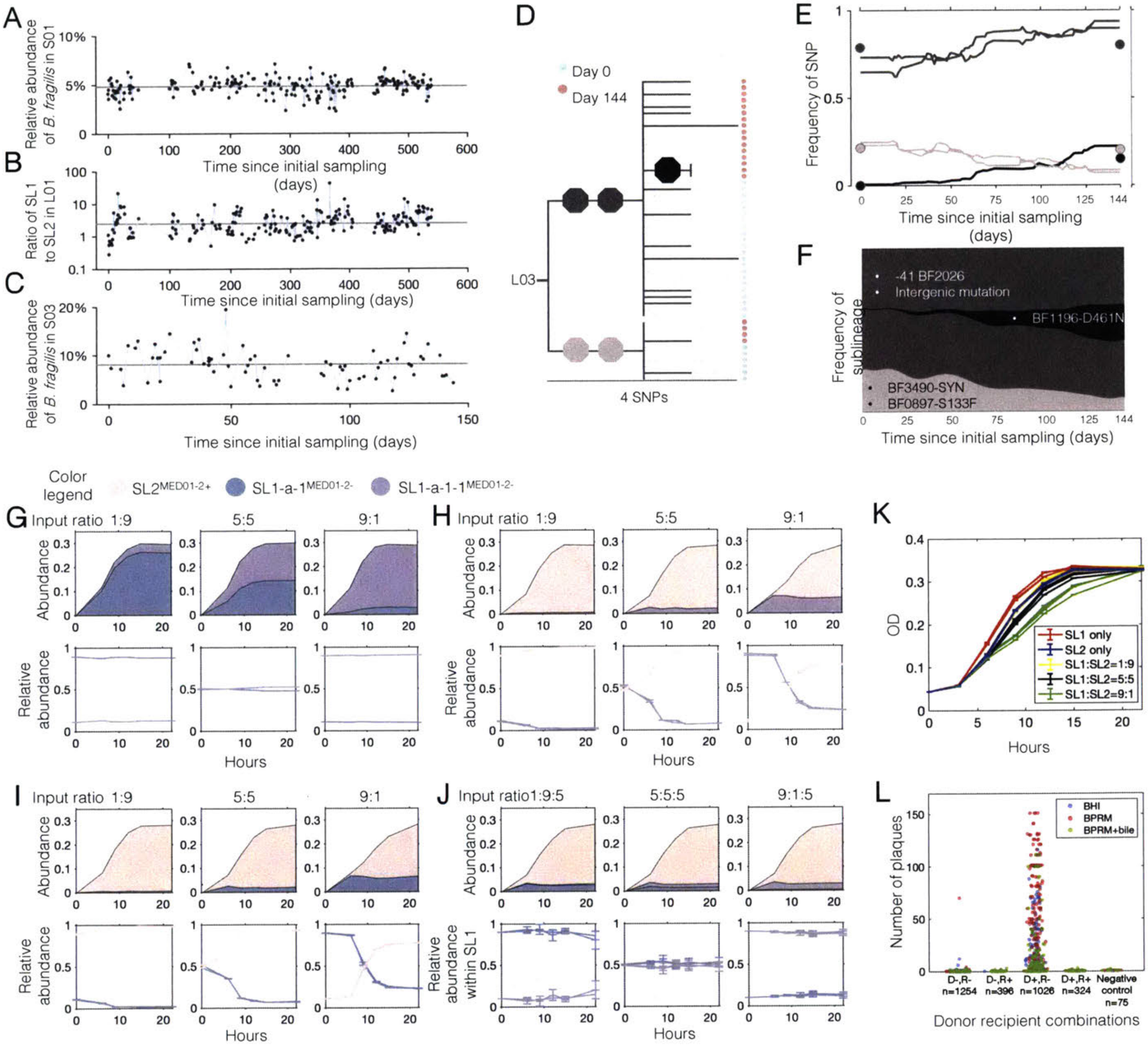

A
Relative abundance of B. fragilis in S01
10%
5%
0
0 100 200 300 400 500 600
Time since initial sampling (days)
B
Ratio of SL1 to SL2 in L01
100
10
1
0.1
0 100 200 300 400 500 600
Time since initial sampling (days)
C
Relative abundance of B. fragilis in S03
20%
10%
0
0 50 100 150
Time since initial sampling (days)
D
Day 0
Day 144
L03
4 SNPs
E
Frequency of SNP
1
0.5
0
0 25 50 75 100 125 144
Time since initial sampling (days)
F
Frequency of sublineage
-41 BF2026
Intergenic mutation
BF1196-D461N
BF3490-SYN
BF0897-S133F
0 25 50 75 100 125 144
Time since initial sampling (days)
Color legend
SL2MED01-2+
SL1-a-1MED01-2-
SL1-a-1-1MED01-2-
G
Input ratio 1:9
5:5
9:1
Abundance
0.3 0.2 0.1 0
Relative abundance
1 0.5 0
0 10 20
Hours
H
Input ratio 1:9
5:5
9:1
Abundance
Relative abundance
Hours
K
OD
0.4 0.3 0.2 0.1 0
0 5 10 15 20
Hours
SL1 only
SL2 only
SL1:SL2=1:9
SL1:SL2=5:5
SL1:SL2=9:1
I
Input ratio 1:9
5:5
9:1
Abundance
Relative abundance
Hours
J
Input ratio1:9:5
5:5:5
9:1:5
Abundance
Relative abundance within SL1
Hours
L
Number of plaques
150 100 50 0
BHI
BPRM
BPRM+bile
D-,R- n=1254
D-,R+ n=396
D+,R- n=1026
D+,R+ n=324
Negative control n=75
Donor recipient combinations

Figure S6

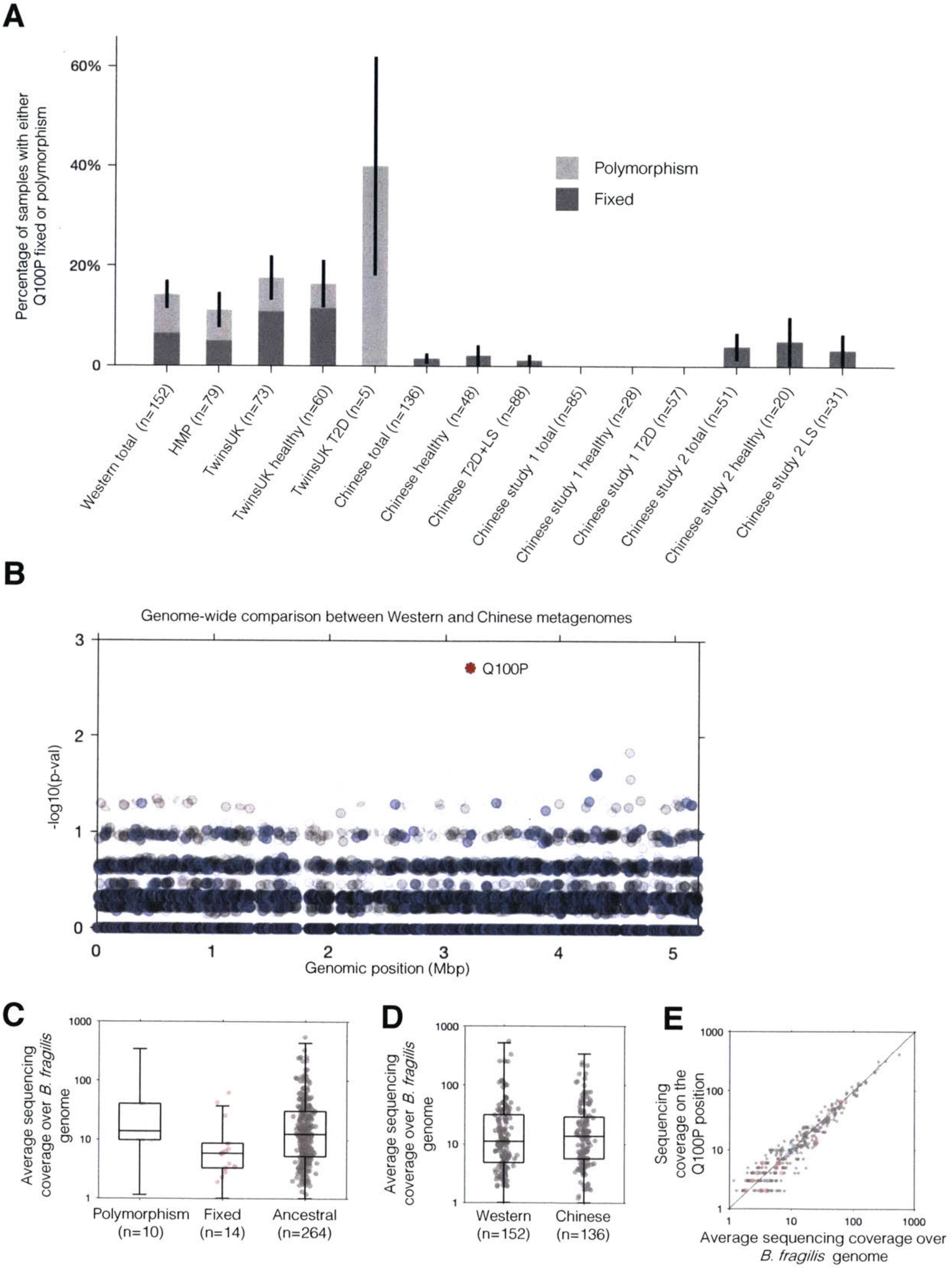
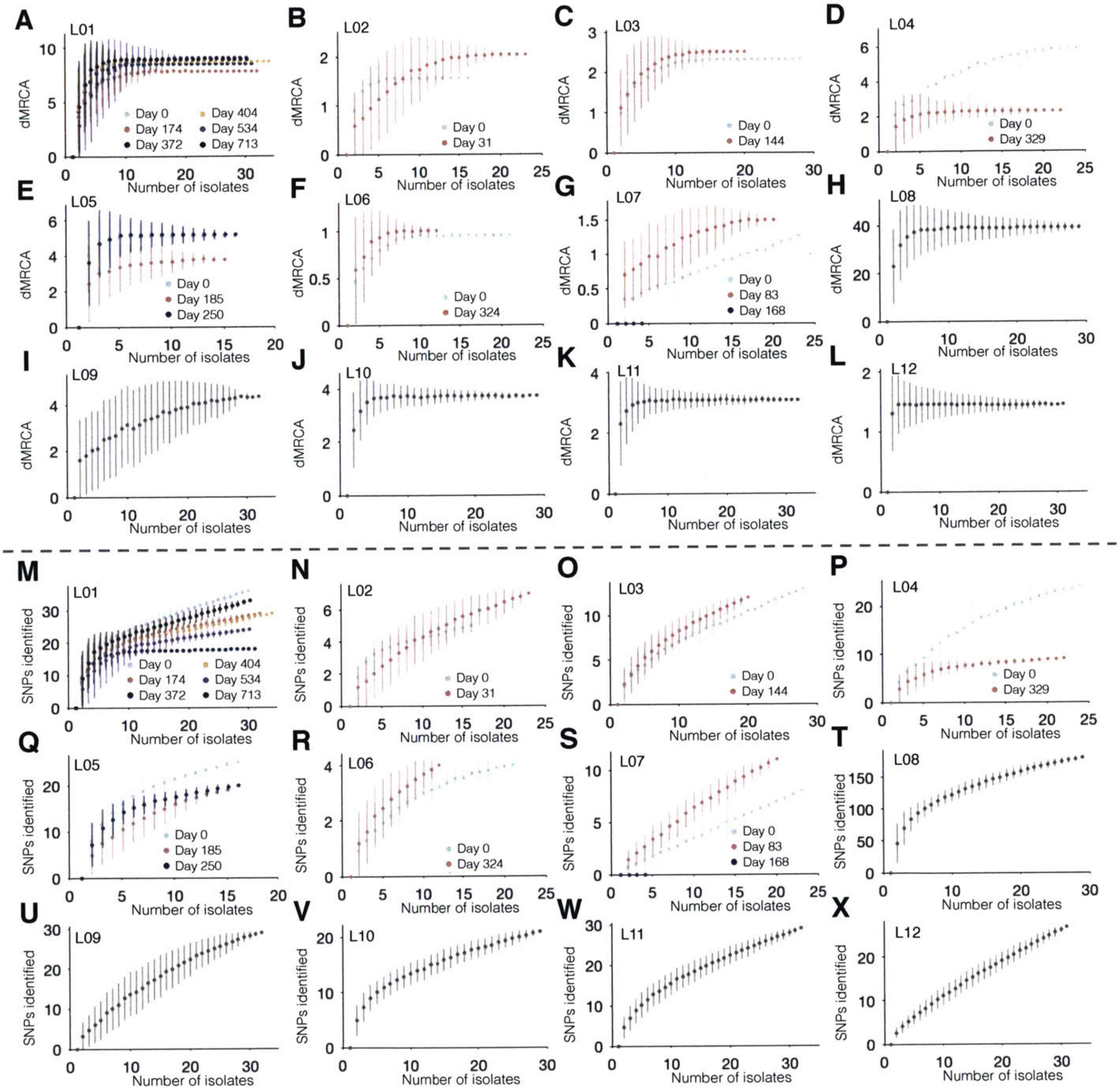

A
Percentage of samples with either Q100P fixed or polymorphism
60%
40%
20%
0
Polymorphism
Fixed
Western total (n=152)
HMP (n=79)
TwinsUK (n=73)
TwinsUK healthy (n=60)
TwinsUK T2D (n=5)
Chinese total (n=136)
Chinese healthy (n=48)
Chinese T2D+LS (n=88)
Chinese study 1 total (n=85)
Chinese study 1 healthy (n=28)
Chinese study 1 T2D (n=57)
Chinese study 2 total (n=51)
Chinese study 2 healthy (n=20)
Chinese study 2 LS (n=31)
B
Genome-wide comparison between Western and Chinese metagenomes
-log10(p-val)
3
2
1
0
Q100P
0
1
2
3
4
5
Genomic position (Mbp)
C
Average sequencing coverage over B. fragilis genome
1000
100
10
1
Polymorphism (n=10)
Fixed (n=14)
Ancestral (n=264)
D
Average sequencing coverage over B. fragilis genome
1000
100
10
1
Western (n=152)
Chinese (n=136)
E
Sequencing coverage on the Q100P position
1000
100
10
1
1
10
100
1000
Average sequencing coverage over B. fragilis genome

Figure S7


A
L01
dMRCA
Number of isolates
Day 0
Day 174
Day 372
Day 404
Day 534
Day 713
B
L02
Day 0
Day 31
C
L03
Day 0
Day 144
D
L04
Day 0
Day 329
E
L05
Day 0
Day 185
Day 250
F
L06
Day 0
Day 324
G
L07
Day 0
Day 83
Day 168
H
L08
I
L09
J
L10
K
L11
L
L12
M
L01
SNPs identified
Day 0
Day 174
Day 372
Day 404
Day 534
Day 713
N
L02
Day 0
Day 31
O
L03
Day 0
Day 144
P
L04
Day 0
Day 329
Q
L05
Day 0
Day 185
Day 250
R
L06
Day 0
Day 324
S
L07
Day 0
Day 83
Day 168
T
L08
U
L09
V
L10
W
L11
X
L12

2.7 Supplementary tables

Table S1

Table S1: Subject information and per-lineage statistics

Donor ID	**S01**	**S02**	**S03**	**S04**	**S05**	**S06**	**S07**	**S08**	**S09**	**S10**	**S11**	**S12**
Donor information												
Age at first sample (years)	28	22	37	27	27	36	29	26	35	25	31	32
Sex	M	M	M	F	F	M	M	M	M	F	F	F
BMI	23	26	26	21	21	22	22	23	24	22	21	20
Per-donor sample information												
Number of stool samples for culturing	10	3	2	2	3	2	3	1	1	1	1	1
Number of time points for culturing*	6	2	2	2	3	2	3	1	1	1	1	1
Time between first and last culture samples (days)	713	31	144	329	250	324	168	0	0	0	0	0
Number of metagenome samples	206	29	74	2	0	2	0	0	0	2	2	2
Time between first and last metagenome samples (days)	539	147	144	329	0	324	0	0	0	172	133	125
Within-person evolution summary statistics												
Lineage ID	L01	L02	L03	L04	L05	L06	L07	L08	L09	L10**	L11	L12
Number of SNPs	81	9	21	32	36	8	18	182***	29	21	29	27
Number of gain events	1	0	0	3	0	1	0	4	0	0	1	0
Number of loss events	2	1	0	0	2	2	2	0	2	1	0	0
Number of mobile element differences of unclear direction	3	0	0	1	0	4	0	5	1	0	2	1
dMRCAT0 (SNPs/genome)	8.57	1.56	2.32	5.88	5.13	0.95	1.26	38.9****	4.34	3.72	3.09	1.45
Tajima's D	-0.36	-0.30	-0.66	-0.53	-0.21	-0.10	-0.93	0.01	-0.70	-0.12	-0.45	-0.76
Genome assembly statistics												
Number of contigs (>500 bp)	428	242	84	83	98	157	84	112	164	117	105	98
Size of the largest contig (bases)	504979	294815	732334	504646	367565	716540	506896	479793	260986	578609	474008	442488
Genome Size (from contigs>500 bp, bases)	5480571	5366007	5319470	5230166	5348204	5202614	5191962	5331732	5334688	4897447	5235138	5287402
GC (%)	43.27	43.13	43.12	43.23	43.4	43.3	43.17	43.34	43.74	43.26	43.14	43.13
N50 (bases)	190337	142996	221751	169000	151088	189635	213368	197256	102101	241706	157794	170662

* Samples were defined as from the same time point if they were sampled within a 5-day window

** One isolate from S10 was not in linage L10 (see Extended Data Figure 1)

*** 35 SNPs if suspected hypermutator-induced mutations are removed

**** dMRCAT0=9.9 if suspected hypermutator-induced mutations are removed

N/A = Not available

Table S2

Table S2: Stool samples used for culturing single-colony isolates

Sample ID	Sampling date	Timepoint	Number of isolates	Average fold	Percentage reads	Metagenomic	Sequencing platform	Read length (bp,	Buffer per 1g stool
S01-0001	0	S01-1	15	30.75	98.12%	Yes	HiSeq	100	10
S01-0002	1	S01-1	15	42.15	98.33%	Yes	HiSeq	100	10
S01-0072	174	S01-2	16	40.99	98.38%	Yes	HiSeq	100	10
S01-0073	175	S01-2	16	39.71	98.44%	Yes	HiSeq	100	2.5
S01-0155	372	S01-3	15	33.59	97.69%	Yes	HiSeq	100	2.5
S01-0156	373	S01-3	16	41.84	97.76%	Yes	HiSeq	100	***
S01-0171	404	S01-4	34	59.99	98.51%	Yes	NextSeq	75	2.5
S01-0228	534	S01-5	15	40.72	97.40%	Yes	HiSeq	100	2.5
S01-0231	539	S01-5	15	37.11	97.91%	Yes	HiSeq	100	***
S01-0259*	713	S01-6	30	109.12	99.06%	No	HiSeq	100	2.5
S02-0001	0	S02-1	7	59.21	95.88%	Yes	NextSeq	75	10
S02-0003	2	S02-1	9	56.61	97.41%	Yes	NextSeq	75	10
S02-0024	31	S02-2	23	62.84	97.39%	Yes	NextSeq	75	10
S03-0001	0	S03-1	28	62.67	98.69%	Yes	NextSeq	75	10
S03-0090	144	S03-2	20	96.58	98.87%	Yes	NextSeq	75	10
S04-0006	0	S04-1	24	85.26	98.46%	Yes	NextSeq	75	10
S04-0107	329	S04-2	22	90.20	98.33%	Yes	NextSeq	75	10
S05-0002	0	S05-1	16	59.50	98.80%	No	NextSeq	75	10
S05-0068	185	S05-2	15	28.21	98.73%	No	NextSeq	75	10
S05-0105	250	S05-3	16	48.22	98.64%	No	NextSeq	75	10
S06-0001	0	S06-1	21	54.87	94.89%	Yes	NextSeq	75	10
S06-0122	324	S06-2	12	68.70	98.23%	Yes	NextSeq	75	10
S07-0001	0	S07-1	23	65.91	98.74%	No	NextSeq	75	10
S07-0068	83	S07-2	20	65.57	98.75%	No	NextSeq	75	10
S07-0134	168	S07-3	4	80.01	98.74%	No	NextSeq	75	10
S08-0009	0	S08-1	30	55.60	98.34%	Yes	NextSeq	75	10
S09-0001	0	S09-1	32	78.18	98.24%	Yes	NextSeq	75	10
S10-0039	0	S10-1	30**	81.44	98.48%	Yes	NextSeq	75	10
S11-0001	0	S11-1	32	77.80	98.71%	Yes	NextSeq	75	10
S12-0002	0	S12-1	31	60.86	98.61%	Yes	NextSeq	75	10

* Libraries prepared differently (by BioMicroCenter at MIT)

** One of the isolates from this sample was from a minor lineage

*** This sample was prepared by OpenBiome using proprietary stool formulation 1

Table S3

Table S3: Mobile element difference (MED) information

Summary information of each MED group

Name of the MED	Total Length estimated (bp)	Present in % of isolates	Inference of gain or loss
MED01-1*	8142	66.0%	loss
MED01-2*	8952	48.1%	unknown
MED01-3	13173	7.5%	unknown
MED01-4	5026	15.5%	unknown
MED01-5	1807	0.5%	gain
MED01-6	1645	99.5%	loss
MED02-1	1061	97.4%	loss
MED04-1	71205	32.6%	gain
MED04-2	20906	4.3%	unknown
MED04-3	19089	28.3%	gain
MED04-4	2389	2.2%	gain
MED05-1	5071	97.9%	loss
MED05-2	82461	36.2%	loss
MED06-1	41834	97.0%	loss
MED06-2	58606	72.7%	loss
MED06-3	92253	42.4%	unknown
MED06-4	4168	48.5%	unknown
MED06-5	8791	9.1%	unknown
MED06-6	1329	9.1%	unknown
MED06-7	1073	3.0%	gain
MED07-1	3335	97.9%	loss
MED07-2	21018	76.6%	loss
MED08-1	38315	43.3%	gain
MED08-2	9310	43.3%	gain
MED08-3	80644	93.3%	unknown
MED08-4	14373	93.3%	unknown
MED08-5	24705	50.0%	gain
MED08-6	32102	6.7%	unknown
MED08-7	43112	6.7%	unknown
MED08-8	11150	6.7%	unknown
MED08-9	1568	6.7%	gain
MED09-1	9455	96.9%	loss
MED09-2	571	87.5%	loss
MED09-3	1202	93.8%	unknown
MED10-1	44181	89.7%	loss
MED11-1	55204	50.0%	loss
MED11-2	2751	40.6%	gain
MED11-3	882	68.8%	unknown
MED12-1	3000	96.8%	unknown

* MED01-1 and MED01-2 are hypothesized to be prophages, as they both include phage signature genes XerD and Structural protein P5. MED01-1 and MED01-2 share ~42 kbp homologous regions, which are not included in this table.

Table S4

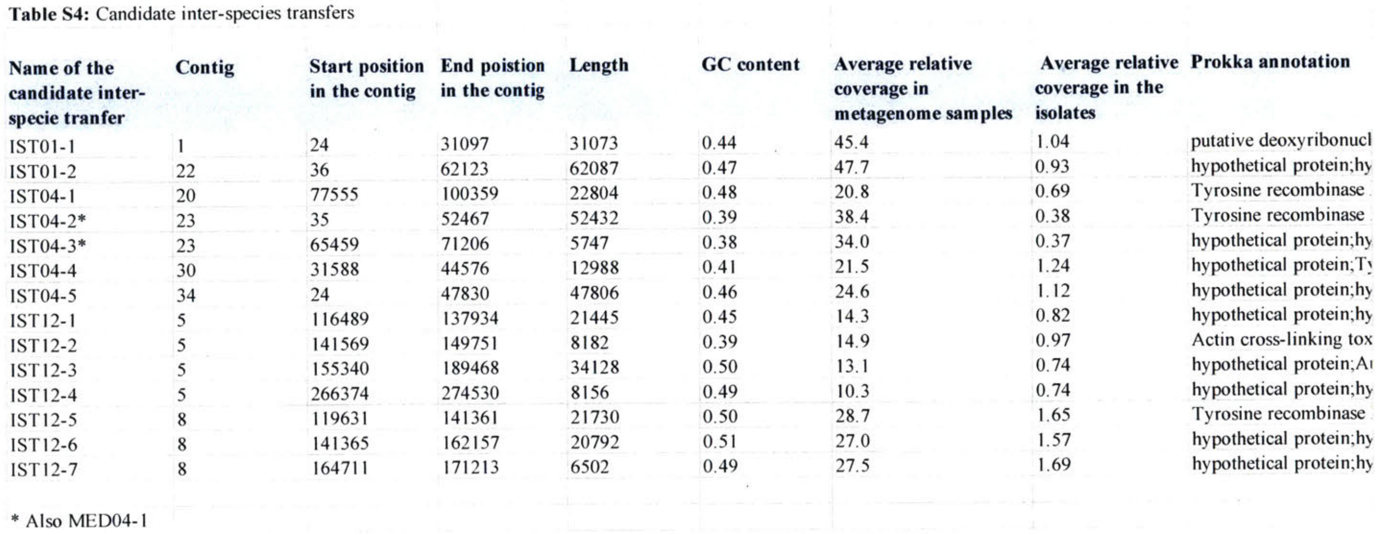


Table S4: Candidate inter-species transfers

Name of the candidate inter-specie tranfer	Contig	Start position in the contig	End poistion in the contig	Length	GC content	Average relative coverage in metagenome samples	Average relative coverage in the isolates	Prokka annotation
IST01-1	1	24	31097	31073	0.44	45.4	1.04	putative deoxyribonucl
IST01-2	22	36	62123	62087	0.47	47.7	0.93	hypothetical protein;hy
IST04-1	20	77555	100359	22804	0.48	20.8	0.69	Tyrosine recombinase
IST04-2*	23	35	52467	52432	0.39	38.4	0.38	Tyrosine recombinase
IST04-3*	23	65459	71206	5747	0.38	34.0	0.37	hypothetical protein;hy
IST04-4	30	31588	44576	12988	0.41	21.5	1.24	hypothetical protein;Ty
IST04-5	34	24	47830	47806	0.46	24.6	1.12	hypothetical protein;hy
IST12-1	5	116489	137934	21445	0.45	14.3	0.82	hypothetical protein;hy
IST12-2	5	141569	149751	8182	0.39	14.9	0.97	Actin cross-linking tox
IST12-3	5	155340	189468	34128	0.50	13.1	0.74	hypothetical protein;A
IST12-4	5	266374	274530	8156	0.49	10.3	0.74	hypothetical protein;hy
IST12-5	8	119631	141361	21730	0.50	28.7	1.65	Tyrosine recombinase
IST12-6	8	141365	162157	20792	0.51	27.0	1.57	hypothetical protein;hy
IST12-7	8	164711	171213	6502	0.49	27.5	1.69	hypothetical protein;hy

* Also MED04-1

Table S5

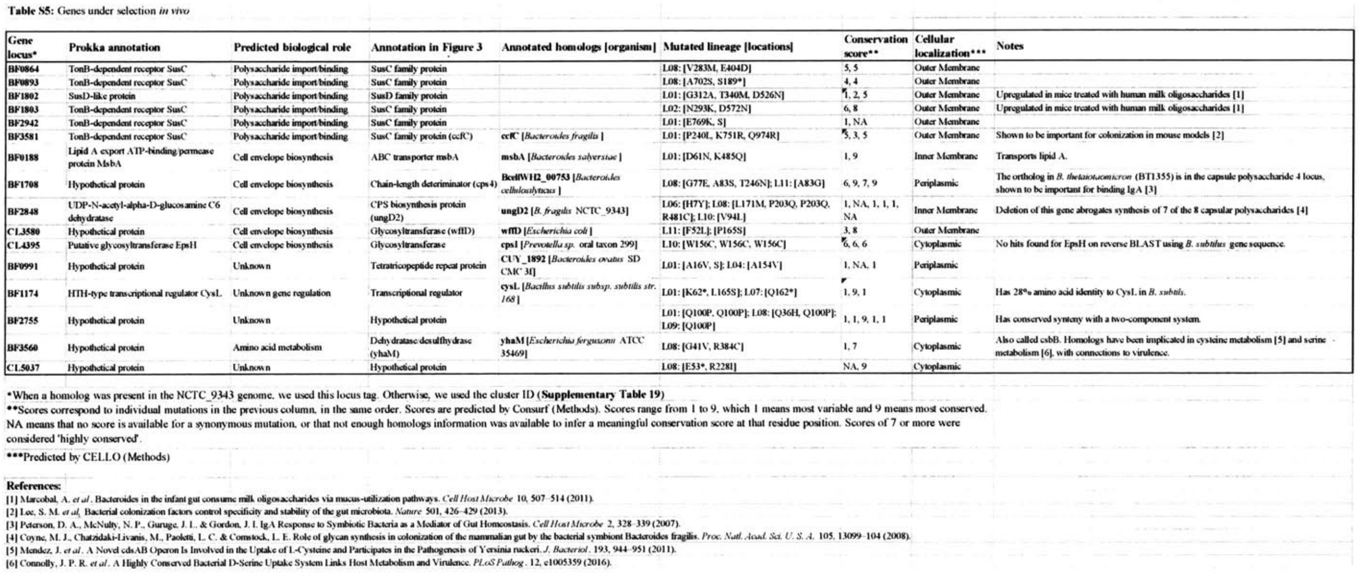


Table S5: Genes under selection *in vivo*

Gene locus*	Prokka annotation	Predicted biological role	Annotation in Figure 3	Annotated homologs [organism]	Mutated lineage [locations]	Conservation score**	Cellular localization***	Notes
BF0864	TonB-dependent receptor SusC	Polysaccharide import/binding	SusC family protein		L08: [V283M, E404D]	5, 5	Outer Membrane	
BF0893	TonB-dependent receptor SusC	Polysaccharide import/binding	SusC family protein		L08: [A702S, S189*]	4, 4	Outer Membrane	
BF1802	SusD-like protein	Polysaccharide import/binding	SusD family protein		L01: [G312A, T340M, D526N]	1, 2, 5	Outer Membrane	Upregulated in mice treated with human milk oligosaccharides [1]
BF1803	TonB-dependent receptor SusC	Polysaccharide import/binding	SusC family protein		L02: [N293K, D572N]	6, 8	Outer Membrane	Upregulated in mice treated with human milk oligosaccharides [1]
BF2942	TonB-dependent receptor SusC	Polysaccharide import/binding	SusC family protein		L01: [E769K, S]	1, NA	Outer Membrane	
BF3581	TonB-dependent receptor SusC	Polysaccharide import/binding	SusC family protein (ccfC)	**ccfC** [*Bacteroides fragilis*]	L01: [P240L, K751R, Q974R]	5, 3, 5	Outer Membrane	Shown to be important for colonization in mouse models [2]
BF0188	Lipid A export ATP-binding/permease protein MsbA	Cell envelope biosynthesis	ABC transporter msbA	**msbA** [*Bacteroides salyersiae*]	L01: [D61N, K485Q]	1, 9	Inner Membrane	Transports lipid A.
BF1708	Hypothetical protein	Cell envelope biosynthesis	Chain-length deteriminator (cps4)	**BcellWH2_00753** [*Bacteroides cellulosilyticus*]	L08: [G77E, A83S, T246N]; L11: [A83G]	6, 9, 7, 9	Periplasmic	The ortholog in *B. thetaiotaomicron* (BT1355) is in the capsule polysaccharide 4 locus, shown to be important for binding IgA [3]
BF2848	UDP-N-acetyl-alpha-D-glucosamine C6 dehydratase	Cell envelope biosynthesis	CPS biosynthesis protein (ungD2)	**ungD2** [*B. fragilis* NCTC_9343]	L06: [H7Y]; L08: [L171M, P203Q, P203Q, R481C]; L10: [V94L]	1, NA, 1, 1, 1, NA	Inner Membrane	Deletion of this gene abrogates synthesis of 7 of the 8 capsular polysaccharides [4]
CL3580	Hypothetical protein	Cell envelope biosynthesis	Glycosyltransferase (wfID)	**wfID** [*Escherichia coli*]	L11: [F52L]; [P165S]	3, 8	Outer Membrane	
CL4395	Putative glycosyltransferase EpsH	Cell envelope biosynthesis	Glycosyltransferase	**cpsI** [*Prevotella sp.* oral taxon 299]	L10: [W156C, W156C, W156C]	6, 6, 6	Cytoplasmic	No hits found for EpsH on reverse BLAST using *B. subtilus* gene sequence.
BF0991	Hypothetical protein	Unknown	Tetratricopeptide repeat protein	**CUY_1892** [*Bacteroides ovatus* SD CMC 3f]	L01: [A16V, S]; L04: [A154V]	1, NA, 1	Periplasmic	
BF1174	HTH-type transcriptional regulator CysL	Unknown gene regulation	Transcriptional regulator	**cysL** [*Bacillus subtilis subsp. subtilis str. 168*]	L01: [K62*, L165S]; L07: [Q162*]	1, 9, 1	Cytoplasmic	Has 28% amino acid identity to CysL in *B. subtils.*
BF2755	Hypothetical protein	Unknown	Hypothetical protein		L01: [Q100P, Q100P]; L08: [Q36H, Q100P]; L09: [Q100P]	1, 1, 9, 1, 1	Periplasmic	Has conserved synteny with a two-component system.
BF3560	Hypothetical protein	Amino acid metabolism	Dehydratase-desulfhydrase (yhaM)	**yhaM** [*Escherichia fergusonii* ATCC 35469]	L08: [G41V, R384C]	1, 7	Cytoplasmic	Also called csbB. Homologs have been implicated in cysteine metabolism [5] and serine metabolism [6], with connections to virulence.
CL5037	Hypothetical protein	Unknown	Hypothetical protein		L08: [E53*, R228I]	NA, 9	Cytoplasmic	

*When a homolog was present in the NCTC_9343 genome, we used this locus tag. Otherwise, we used the cluster ID (**Supplementary Table 19**)

**Scores correspond to individual mutations in the previous column, in the same order. Scores are predicted by Consurf (Methods). Scores range from 1 to 9, which 1 means most variable and 9 means most conserved. NA means that no score is available for a synonymous mutation, or that not enough homologs information was available to infer a meaningful conservation score at that residue position. Scores of 7 or more were considered 'highly conserved'.

***Predicted by CELLO (Methods)

References:

[1] Marcobal, A. *et al*. Bacteroides in the infant gut consume milk oligosaccharides via mucus-utilization pathways. *Cell Host Microbe* 10, 507–514 (2011).

[2] Lee, S. M. *et al*, Bacterial colonization factors control specificity and stability of the gut microbiota. *Nature* 501, 426–429 (2013).

[3] Peterson, D. A., McNulty, N. P., Guruge, J. L. & Gordon, J. I. IgA Response to Symbiotic Bacteria as a Mediator of Gut Homeostasis. *Cell Host Microbe* 2, 328–339 (2007).

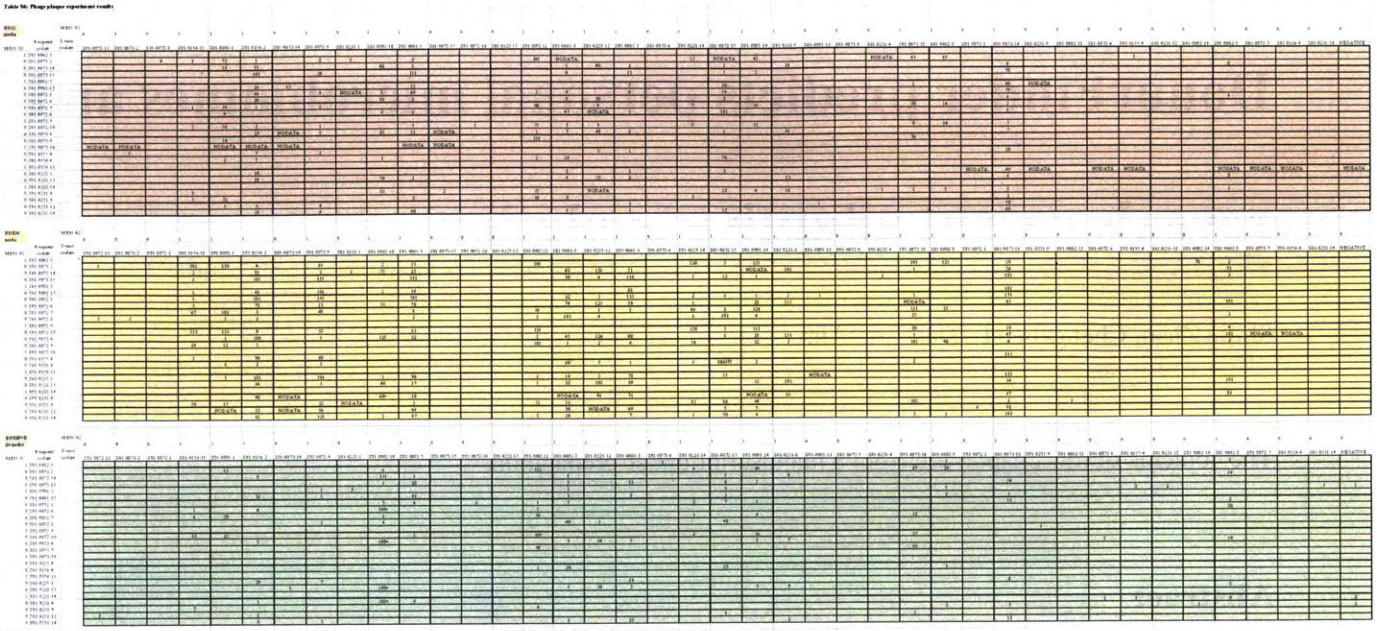
[4] Coyne, M. J., Chatzidaki-Livanis, M., Paoletti, L. C. & Comstock, L. E. Role of glycan synthesis in colonization of the mammalian gut by the bacterial symbiont Bacteroides fragilis. *Proc. Natl. Acad. Sci. U. S. A.* 105, 13099–104 (2008).

[5] Mendez, J. *et al*. A Novel cdsAB Operon Is Involved in the Uptake of L-Cysteine and Participates in the Pathogenesis of Yersinia ruckeri. *J. Bacteriol*. 193, 944–951 (2011).

[6] Connolly, J. P. R. *et al*. A Highly Conserved Bacterial D-Serine Uptake System Links Host Metabolism and Virulence. *PLoS Pathog*. 12, e1005359 (2016).

Table S6

Table S6: Phage plaque experiment results



Data recording rules

1. Each donor-recipient combination was counted on Day 1 (17-20 hours), Day 2 (43 hours) and Day [illegible], the maximum number of plaques was listed in the table
2. Blanks represent no plaques
3. Datapoints visually marked as potentially representing cross-contaminating plaques from neighboring drops were marked as "NODATA"

Media recipe
Brain heart infusion (BHI) growth media was prepared in water at 37 g/L Brain Heart Infusion [illegible]
[illegible]
EHEM+bile medium was prepared as was EHEM with the exception of the addition of [illegible]

Chapter 3

DonorFinder predicts personal microbiomes and reveals recent adaptive evolution

Shijie Zhao, Chengzhen L. Dai, Eric Alm

The contents of this chapter are from a manuscript under preparation.

Abstract

The human gut microbiomes are individualized ecosystems that consist of bacterial strains stably colonizing for up to years. The interpersonal variability of strains can help understand the personal signatures of the microbiome and the transmission of strains between individuals. Here, we introduce a reference-based microbiome strain tracking approach (DonorFinder) that determines whether distinct metagenomes harbor closely-related strains and classifies if a pair of metagenomes belong to the same donor. Focusing on a set of 30 species, DonorFinder achieves >96% specificity and ~100% sensitivity in predicting metagenome donors and discovers a pair of metagenomes with labels switched in the Human Microbiome Project. Applying DonorFinder to the metagenomes of adult twins from the TwinUK registry, we identify 6 cases of closely-related strains carried by both twins. Identification of point mutations in these shared strains reveals evidence of genome-wide within-person adaption, potentially over decades of colonization. Overall, we demonstrate that DonorFinder predicts closely-related strains from metagenomic samples, with applications to revealing microbiome donors and recent evolution events.

3.1 Introduction

The human gut microbiome harbors a complex community of microbial species stably colonizing for years or even decades (Lloyd-Price et al. 2017; J. Faith et al. 2013). Although different individuals from a human population tend to have a similar set of species, the strains carried by distinct people – defined by genomic variations such as single nucleotide polymorphisms (SNPs) and gene presence/absence – are usually person-specific (Truong et al. 2017; Scholz et al. 2016). Thus, understanding the composition and identity of strains in a metagenome has enabled the prediction of individual microbiome donors with up to 80% accuracy (Franzosa et al. 2015; Hampton-Marcell, Lopez, and Gilbert 2017). In addition, tracking closely-related strains paves the way to understanding how microbial strains are transmitted between family members, across social networks and after fecal microbiota transplantation (Smillie et al. 2018; Ferretti et al. 2018; Ilana L. Brito et al. 2019). Monitoring strains from the gut microbiome, therefore, has the potential to reveal fine-scale interactions between microbial species and their human hosts.

To date, various computational methods have been developed to resolve strains from metagenomic samples. Several methods resolve strains via identification of SNPs across different metagenomes by aligning short reads to targeted species reference genomes (Luo et al. 2015; Truong et al. 2017; Costea et al. 2017; Smillie et al. 2018). However, as the process of calling SNPs is prone to false positives, high-resolution strain profiling usually requires careful and iterative tuning of filtering parameters (Ilana Lauren Brito and Alm 2016). Another commonly-used approach involves the analysis of pangenomes, particularly the accessory genomic regions carried by distinct strains (Ilana Lauren Brito and Alm 2016). Unlike SNP-based methods, pangenome-based approaches are more robust to parameter changes, yet they still require a large database of pangenomes and substantial computational resources (Scholz et al. 2016; Zhu et al. 2015).

Here, we introduce a flexible and lean method—DonorFinder—that uses a single reference for individual species to compare strain identities between metagenomes and classifies if two metagenomes are from the same individual. Our approach is based on the assumption that the accessory genomes of microbial species are highly individual-specific and stable over time (**Figure 1A**). We compare the accessory genomes of 40 prevalent gut bacteria species and show empirical data supporting this assumption for 25 out of the 40 species. We also design a classification rule that leverages these comparisons to predict whether two metagenomes belong to the same individual. We demonstrate near-perfect specificity and sensitivity of DonorFinder using metagenomes from the Human Microbiome Project (HMP) and the Broad Next 10 (BN10) project. We find evidence supporting a mislabeling of donor IDs in a pair of the HMP metagenomes. When applied to members of the same family, DonorFinder fails to differentiate samples between certain family members, consistent with the occurrence of strain transmission between family members. Lastly, when applying DonorFinder to track strains from a dataset of adult twins, we find that these twins can share strains for potentially over decades of colonization and discover signatures for adaptive evolution in these shared strains. The datasets and code used in this work are available for download from https://github.com/shijiezhao/DonorFinder.

3.2 Results

Accessory genome difference (AGD) as a metric to define inter-sample strain variance

We developed a bioinformatic workflow (DonorFinder) to achieve strain-level comparison by first aligning metagenomic reads against a single well-assembled reference genome for individual species. Specifically, metagenomics reads are aligned and compared against a curated set of 40 abundant or widely-studied representative gut bacterial species (Lloyd-Price et al., 2017; Xie et al., 2016, **Table S1**). To quantify the differences of strains between metagenomes, we developed a metric that estimates the fraction of a reference genome that is variable between two metagenomes. For any given species, we calculated the relative sequencing depth for every 5 kb genomic window within a metagenomic sample and compared the relative sequencing depth of each genomic window between sample pairs (Methods). If a genomic window was present with a relative sequencing depth of >0.5X in one sample but was present with <0.05X in another sample, this genomic region is designated as a differential region. The fraction of these regions in the reference genome is defined as the accessory genome difference (AGD; Methods).

To demonstrate how AGDs can reveal personalized strain signatures, we examined *Bacteroides vulgatus*, a prevalent species inhabits the large intestine (Yatsunenko et al. 2012), across HMP metagenomic samples. A pair of distinct metagenomic samples from a same HMP donor had an AGD of 0 (**Figure 1B**), while a pair of metagenomes from two different HMP subjects showed an AGD of 0.040 (**Figure 1C**). We estimated AGDs for all pairwise HMP metagenomes for *B. vulgatus* and observed a clear difference between the inter-subject and intra-subject AGD profiles (Methods; **Figures 1D**). We generated a receiver operating characteristic (ROC) curve and calculated the area under curve (AUC) to be 0.989 (**Figure 1E**). We picked an AGD cutoff that maximized Youden's index (sensitivity + specificity -1; Methods) for *B. vulgatus.* DonorFinder designates that two metagenomes have personalized signature for *B. vulgatus* if the AGD of this species is smaller than the cutoff. Expanding this inter-subject and intra-subject AGD profiles comparison to all of 40 species, we calculated an AUC and a species-specific AGD cutoff for each species (**Figure S1**, **Table S1**).

DonorFinder predicts personal microbiomes for distinct people

We reasoned that a pair of metagenomes from the same individual can usually share strains for multiple species, while a pair of unrelated metagenomes are unlikely to have personalized signature for multiple species (Lloyd-Price et al. 2017; Franzosa et al. 2015). We therefore designed a classification rule that when DonorFinder predicts that more than two species share personalized signatures, these two samples are predicted as from the same donor (**Figure 2**, Methods). To minimize false predictions for individual species, we excluded species that with an AUC < 0.975 in the AGD analysis (**Figure S1**; **Table S1**).

To demonstrate the performance of DonorFinder, we used it to predict donors for all pairs of 535 HMP metagenomic samples. These samples are from 250 distinct human subjects and 161 of the subjects have more than one samples (Lloyd-Price et al. 2017). When comparing all pairwise samples from HMP, our classifier provided us with a sensitivity of 95.79% and a specificity of 99.99% (**Figure 2A**, Methods). To validate DonorFinder with an independent test dataset, we applied DonorFinder to the Broad Next 10 datasets, consisting of 410 metagenomic samples from 50 distinct individuals, and achieved 100% specificity and 100% sensitivity (**Figure 2B**).

From the HMP dataset, we noticed a potential mislabeling of donor IDs for a pair of metagenomic samples. We observed that the sample SRS045244 from subject 763880905 is predicted by DonoFinder to belong to share a microbiome donor with two samples from subject 763536994 (SRS014287 and SRS062427); meanwhile, the other sample from subject 763880905 (SRS014948) was matched to the remaining sample from subject 763536994 (SRS050422, **Figure 2C**). Given that the empirical estimation of false negative rate is < 5% and false positive rate is < 0.02% (**Figure 2A**), we estimated that the probability of observing these donor matching patterns is <10^{-17} (**Figure 2C**). However, if we assumed that the donor labels of sample SRS050422 and SRS045244 were shuffled, the estimated probability of observing the predicted pairings is ~1 (**Figure 2D**). This analysis suggested that there is a mislabeling in the HMP metagenomes and we offered a parsimonious solution to correct that. After correcting for this putative label-shuffling, DonorFinder had an updated sensitivity of 96.4% for the HMP metagenomes.

DonorFinder is limited when applying to samples from family members

Both HMP and BN10 donors consist of mostly unrelated individuals from the US, and it is therefore expected that distinct donors carry distinct strains. However, members from a same household may share strains to extensive levels, especially for children whose microbiome may derive directly from the parents (Ferretti et al. 2018). To test whether DonorFinder can be applied to people from a same family, we examined the metagenomes of an 8-member family: mother, father, and six children of ages 0, 2, 4, 6, 8, and 10 years old (Schloss et al. 2014). Each family member had from 1 to 3 metagenomic samples available (**Figure S2**). While most pairs of family members can be successfully separated by DonorFinder, we found that certain family members shared closely-related strains that complicated the classification accuracy. In particular, the microbomes of the 4-year old, 6-year old and 8-year old children are predicted to be from the same individual (**Figure S2**). These three children shared similar strain signatures for multiple species, including *Bifidobacterium* and *Bacteroides* strains. While such results suggest that our method is limited in differentiating the microbiomes of family members, they demonstrate the ability of DonorFinder in tracking the transmission of strain.

Adult twins share closely-related strains at low frequency, potentially for decades

We next explored the ability of DonorFinder to track strains shared by different human subjects over longer period of time (potential transmission events). We selected a dataset from adult twins in the UK Twin Registry, including 125 pairs of adult twins between their 50s to 70s (Xie et al. 2016). We first calculated the inter-twin AGDs for the 25 species that are used in DonorFinder predictor. While majority of the species do not have inter-twin AGDs smaller than the species-specific cutoffs, we nonetheless identify 27 cases in which a species shared a personalized signature between twins (**Table S2**).

To validate if these identified personalized signatures reveals real transmission events between the twins, we examined the evolutionary history of these strains to rule out apparent false positives. We identified genome-wide distribution of SNPs for these strains by searching for nucleotide positions in which the major alleles are discordant between twins (Methods). Species with a signature of multiple-strain colonization, defined by an excess of genomic positions with major allele frequency smaller than 0.95, are excluded (Methods). We also excluded twin-species combinations containing genomic regions with more than 20 SNPs/Kb (Methods, **Table S2**, **Figure S3**). Given that the documented molecular clock for bacterial species in natural environment ranges from 0.5-5 SNPs/year, these numbers of SNPs are inconsistent with recent transmission events and are likely false positives or complicated by homologous recombination events. After filtering, we had 6 cases of shared closely-related strains showing evidence of recently-emerged mutations. Our analysis suggested between 4 to 74 SNPs separating the strains harbored by distinct twins (**Figure 4A-B**, **Table S2**). It is worth noting that our SNP analysis can only detect mutations that reached high frequency within either twin's microbiome. Given the molecular clocks for bacterial species (Didelot et al. 2016), these numbers of SNPs suggest years to decades of evolutionary divergence between the twins. Since these twins usually have been living apart for 30-50 years, it is thus likely that these mutations emerged and accumulated independently within the gut of each twin and some shared strains may have been colonizing both twins for decades.

Strains shared by twin pairs show signatures of adaptive within-person evolution

The shared strains between adult twins and the recently emerged SNPs (years to decades) provide an opportunity to investigate the within-person evolutionary process of these strains. Since our method only included SNPs with major allele frequency larger than 80%, these point mutations have partially or completely swept one of the twin's microbiome. We identified 6 strains that passed our filtering criteria.

To examine if these point mutations reflect adaptive evolution within these twin subjects, we calculated the canonical measure of selection, dN/dS, for mutations arising from a same species. dN/dS is the normalized ratio of non-synonymous mutations to synonymous mutations and is a canonical measure of selection (Methods). For all species that we tested, we found the values dN/dS are larger than or very close to 1(**Figure 4B**). When combining SNPs identified from all these species, we obtained a dN/dS that is significantly bigger than 1, suggesting genome-wide adaptive evolution dominates the within-person evolution for these species. We therefore conclude that the mutations that swept in these twins were driven by adaptive evolution.

3.3 Discussion

Here, we introduce a new metagenomic analysis framework (DonorFinder) for the rapid identification of closely-related strains between metagenomes. Comparing to other strain-tracking methods, DonorFinder has a lean implementation and is flexible to be modified for different tasks. Leveraging the assumption that unrelated human subjects carry strains with unique accessory genome profiles, we build DonorFinder with 25 species that have distinct intra-personal and inter-personal AGD profiles. Such differences allow us to infer with confidence whether two metagenomic samples share strains and further predict personal microbiomes with near-perfect accuracy. DonorFinder performs particularly well for HMP and BN10 metagenomes, with only a few cases of misclassification in HMP samples. Our sensitivity and specificity are better compared to a previously reported classifier with 80% accuracy in recovering HMP microbiome donors (Franzosa et al. 2015). Our results suggest that this method may have applications in microbiome-based forensics and tracking transmissions. In addition, we find at least one convincing case that a pair of metagenomes were incorrectly labeled. This mislabeling has been hinted in the supplementary materials from a previous report from the Bork lab (Schloissnig et al. 2013).

Our method also enables us to track closely-related strains across metagenome samples and helps identify strains shared by twin pairs, potentially over decades of colonization. Further analysis of the point mutations between the twin pairs revealed evidence that these shared strains experience genome-wide adaptive evolution. Our analysis only accounts for mutations that nearly sweep either twin and is likely missing mutations that are present at medium or low frequencies. In addition, we identified only 6 shared strains and these strains are all from the *Bacteroidetes* phylum. Nonetheless, our results demonstrated that adaptive evolution may dominates at short timescale genome-wide. This is striking given compelling evidence that purifying selection dominates evolution at timescales for over thousands of years (Zhao et al. 2019; Schloissnig et al. 2013). To solve this discrepancy, we propose two theoretical scenarios to reconcile signals from the two timescales. One possibility is that many strains carried by an individual will be lost over transmission between human populations, thus within-person adaptive mutations rarely transmitted to new human hosts. Another possibility is that within-person adaptive mutations are person-specific and usually lead to selective disadvantages in new human hosts, and over time, these adaptive mutations will be selected against by natural forces. Future studies with larger sample sizes and more species diverse taxonomical groups are needed to test these hypotheses.

2.4 Declaration

Acknowledgements

We thank Fangqiong Ling for inspiring us to explore predicting human host using metagenome data. We are grateful to Tami Lieberman, Sean Gibbons, Sean Kearney, Thomas Gurry and Xiaoqian Yu for helpful discussions. We are thankful to Xiaofang Jiang, Mathieu Groussin and Mahilde Poyet for obtaining and preprocessing BN10 metagenomes.

Author contributions

S.Z. and E.J.A designed the study; S.Z. performed the metagenomic analysis; S.Z. and C.Z.D. wrote the manuscript with the input from E.J.A.

3.5 Methods

Metagenomic datasets used in this study

We considered three publicly available datasets for this study: the Human Microbiome Project (Lloyd-Price et al. 2017) (535 samples from 250 sbujects://hmpdacc.org), the TwinsUK study (Xie et al. 2016) (250 samples from 250 subjects; ERP010708), and a gut microbiome study of an eight member family (Schloss et al. 2014) (15 samples from 8 family members). We also included datasets from the Broad Next 10 project (410 samples from 50 subjects); the manuscript of the BN10 resource paper is under review and the BN10 metagenomes will be public upon acceptance.

Reference genomes

Our accessory genome comparison requires that this strain has adequate sequencing depths from both metagenomic samples. To meet this criterion, we manually curated species that are abundant and prevalent in the HMP and TwinsUK datasets and included some well-characterized species found in the gut microbiome (e.g., *E. coli*), totaling 40 species. A representative reference genome for each species was obtained from NCBI and a single fasta file was generated that contains these 40 genomes. To simplify downstream analysis, for references with multiple scaffolds, we connected the sequences from different scaffolds to form an artificial single contig. The list of references used in this study can be found in **Table S1**.

Metagenomic reads alignment

Metagenomic reads were trimmed and filtered using Cutadapt and Sickle (Martin 2011; Joshi and Fass 2011). The filtered reads were aligned to the combined reference genome using Bowtie2 (Langmead and Salzberg 2012) (Parameters: -X 2000, --no-mixed, --very-sensitive, --n-ceil 0,0.01, --un-conc). Alignment files (sam format) were converted to pileup files using Samtools (H. Li et al. 2009) (Step 1: samtools view -bS; Step 2: samtools sort; Step 3: samtools mpileup -q30 -x -s -O -d3000). From these pileup files, we extracted information about read depth at each genomic position from the combined reference.

AGD calculation

AGD is defined as the fraction of accessory genomic regions that are different between two metagenomic samples. AGD is used to quantify strain-strain distance between a pair of metagenomes. We first divided the single contig (Methods: Reference genomes) for the targeted species into 5Kb genomic windows. Average sequencing depth was calculated for each of the genomic windows from either metagenome. A genomic region is designated as different when its sequencing depth is lower than 5% of the average sequencing depth in one sample but is higher than 50% of the average in the other sample. To avoid inaccurate estimation of average sequencing depth due to abnormal alignment at mobile genomic regions, average sequencing depth is defined as the mean sequencing depth for regions that are in the 25% and 75% percentile for a given sample.

For each species, we generated a cutoff that maximizes the differentiation of inter-subject and intra-subject AGD profiles (**Figure 1D**, **Table S1**)

DonorFinder predictor

DonorFinder predictor includes 25 species that the AGD analysis shows AUC > 0.975. These species all have sharply different intra-subject and inter-subject AGD profiles. For a given pair of metagenomes, we performed metagenomic alignment to the 25 species references and calculated the AGD for each species between the two metagenomes. For each of the 25 species, the AGD is compared to the species-specific AGD cutoff (**Table S1**). If AGD is smaller than the cutoff, this species is classified to have personalized signature between the two samples. If more than two species share personalized signature between the two metagenomes, DonorFinder predicts that these two samples belong to a same stool donor (**Figure 2**).

Identification of mutations between twins

For each species that share personalized signature between a twin pair, candidate SNPs were identified using SAMtools and filtered using filters optimized from previous work (Lieberman et al. 2014, 2011). In particular, genomic positions were considered to be potential SNP positions if the twin pair were discordant on the called base and both samples had: FQ score less than 30, at least 1 read aligning either forward strand or reverse strand, and a major allele frequency of at least 80%. The median coverage across samples must be more than 1 read. Samples with potential multiple-strain colonization are discarded in the analysis (>3% of the variable positions have <95% major allele frequency). In addition, regions that are not within 50%-200% of average sequencing depth of the genome are discarded, as these polymorphisms are likely from species that share homologous sequence to the reference. Detailed information of between-twins SNPs for the shared strains are listed in **Table S2.**

dN/dS

Mutations were categorized as synonymous (S) or non-synonymous (N) based on open-reading frame annotations from the genbank files of the reference genomes. To calculate dN/dS for sets of *de novo* mutations (**Figure 4**, **Table S2**), we normalized the observed N/S ratios by the expected N/S ratios (Zhao et al. 2019). For any given set of SNPs, we calculated the expected N/S for these SNPs, accounting for both (1) the different probabilities of acquiring nonsynonymous mutations for different types of mutations and (2) the codon compositions of the genes in which these SNPs occurred.

3.6 Figure and Table legends

Figure 1 | Accessory genome difference (AGD) as a metric to define inter-sample strain variance for *B. vulgatus*

(A) An example showing that *B. vulgatus* strains from distinct human subjects differ in accessory genomes. Sequencing depths over the *B. vulgatus* reference are presented for 4 HMP metagenomes. Genomic regions that are differentially present between the samples are colored in red; genomic regions that are present in all four metagenomes are colored in gray.

(**B**) Graphical illustration of calculating AGD for *B. vulgatus* for a pair of metagenomic samples from a same subject. Each dot represents the sequencing depths of a 5Kb genomic window.

(**C**) Graphical illustration of calculating AGD for *B. vulgatus* for a pair of metagenomic samples from two different subjects. Each dot represents the sequencing depth of a 5Kb genomic window. Genomic windows that are differentially present between the two samples are colored in red (Methods)

(**D**) Density histograms for intra-subject AGD profile (red) and inter-subject AGD profile (green) of *B. vulgatus*.

(**E**) ROC analysis for the AGD profiles of *B. vulgatus*. To obtain (sensitivity, specificity) sets to draw the curve, we set cutoffs from 0 to 1 with 0.0001 intervals.

Figure 2 | DonorFinder predicts personal microbiomes for distinct people

For a given pair of metagenomes, inter-sample AGDs are calculated for each of the 25 species (**Table S1**). For each species, the inter-sample AGD is compared to the species-specific AGD cutoff (**Table S1**) and DonofFinder predicts the metagenomes share personalized signature for this species if the AGD is smaller than the species-specific cutoff. When more than two species share personalized signature between the two metagenomes, DonorFinder predicts that these two samples belong to a same stool donor.

Figure 3 | DonorFinder achieves >96% sensitivity and ~100% specificity in predicting metagenome owners

(**A**) Contingency table shows the results of using DonorFinder to predict if a pair of HMP metagenomes belong to a same donor. Numbers in the parentheses are after correcting for misclassifications due to the putative mislabeling illustrated in (**C**) and (**D**).

(**B**) Contingency table shows the results of using DonorFinder to predict if a pair of BN10 metagenomes are from a same donor.

(**C**) Predictions for a set of 5 HMP metagenomes with apparent misclassifications. Samples labeled by HMP as from different donor subjects are labeled with different colors. Given the empirical false positive rate and false negative rate, the probability of observing these predictions is $< 10^{-17}$.

(**D**) If shuffling the donor IDs of SRS050422 and SRS045244, the probability of observing the DonorFinder predictions is ~1. The colors of sample names are recolored according to donor IDs.

Figure 4 | Strains shared by twin pairs show signatures of adaptive within-person evolution

(**A**) - (**B**) Phylogenies of a *B. caccae* strain shared between the twin pair P121 and a *B. cellulosilyticus* strain shared between the twin pair P58. Diamonds represent mutations in genes with more than one SNPs identified within the twins. Individual genes are colored with distinct colors and annotated.

(**C**) dN/dS is calculated for SNPs identified in each individual species and combined. dN/dS is also calculated for combined SNPs excluding *B. intestinihominis*, as this species appears to be a hypermutator. Error bars represent 95% confidence intervals.

Figure S1 | Accessory genome difference (AGD) as a metric to define inter-sample strain variance for *B. adolesentis*

(A) An example showing that *B. adolesentis* strains from different human subjects are different in accessory genomes. Sequencing depths over the *B. adolesentis* reference are presented for 4 HMP metagenomes. Genomic regions that are differentially present between the samples are colored in red; genomic regions that are present in all four metagenomes are colored in gray.

(**B**) Graphical illustration of calculating AGD for *B. adolesentis* for a pair of metagenomic samples from a same subject. Each dot represents the sequencing depth of a 5Kb genomic window.

(**C**) Graphical illustration of calculating AGD for *B. adolesentis* for a pair of metagenomic samples from two different subjects. Each dot represents the sequencing depths of a 5Kb genomic window. Genomic windows that are differentially present between the two samples are colored in red (Methods)

(**D**) Density histograms for intra-subject AGD profile (red) and inter-subject AGD profile (green) of *B. adolesentis*.

(**E**) ROC analysis for the AGD profiles of *B. vulgatus*. To obtain (sensitivity, specificity) sets to draw the curve, we set cutoffs from 0 to 1 with 0.0001 intervals.

Figure S2 | DonorFinder is limited when applying to samples from family members

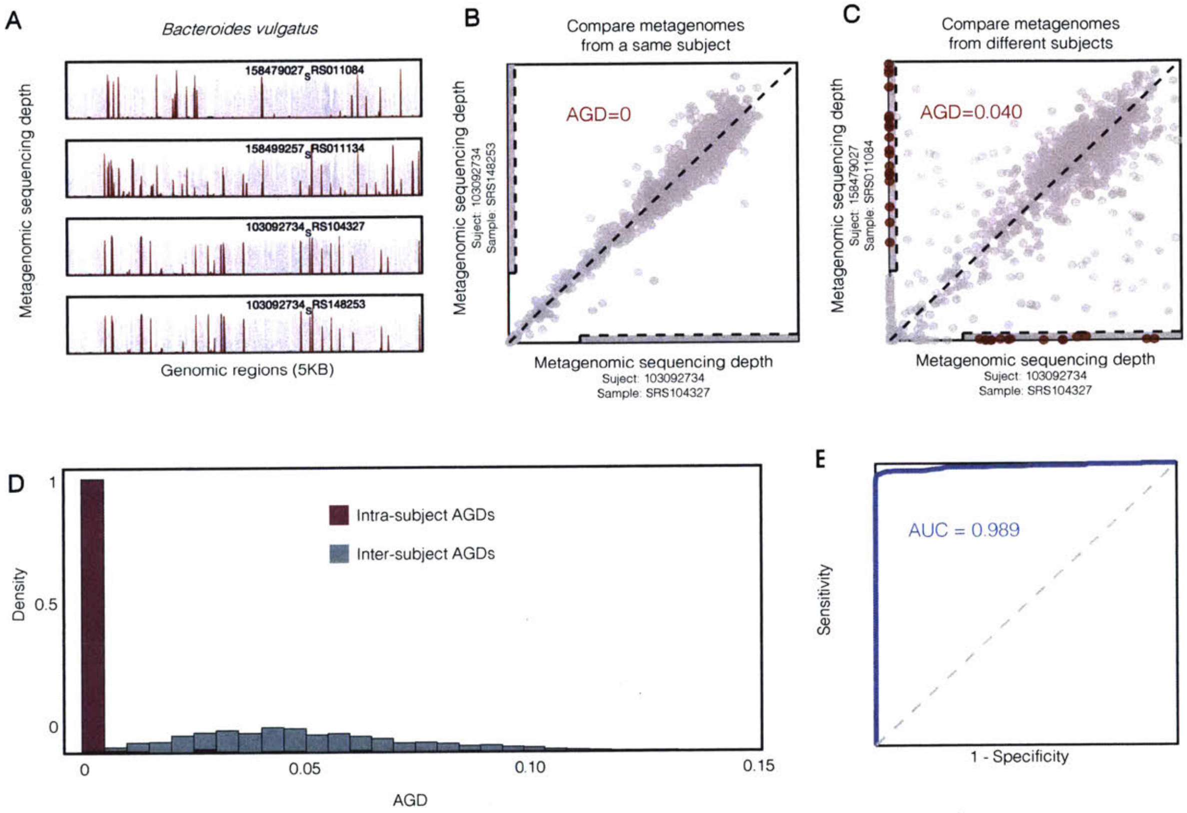
DonorFinder is applied to 15 metagenomes from 8 family members. Columns and rows represent distinct metagenomic samples. Row labels and column lables represent the identity of the family member. If two metagenomes are predicted by DonorFinder as from a same donor, they are colored with green color in the heatmap. We notice that DonoFinder cannot distinguish metagenomes from the 4-year old, 6-year old and 8-year old children.

Figure S3 | AO and AP between UK twin pair P126 shows signs for recombination or false positive due to

The *Alistipes onderdonkii* and *Alistipes putredinis* strains predicted by DonorFinder as having personalized signature between the twins. Both genomes contain regions enriched for SNPs (>20 SNPs/Kb), suggesting that these two species underwent homologous recombination or they are not closely-related strains.

3.7 Figures and Tables

Figure 1



Figur 2

DonorFinder predicts whether two metagenomes are from a same donor

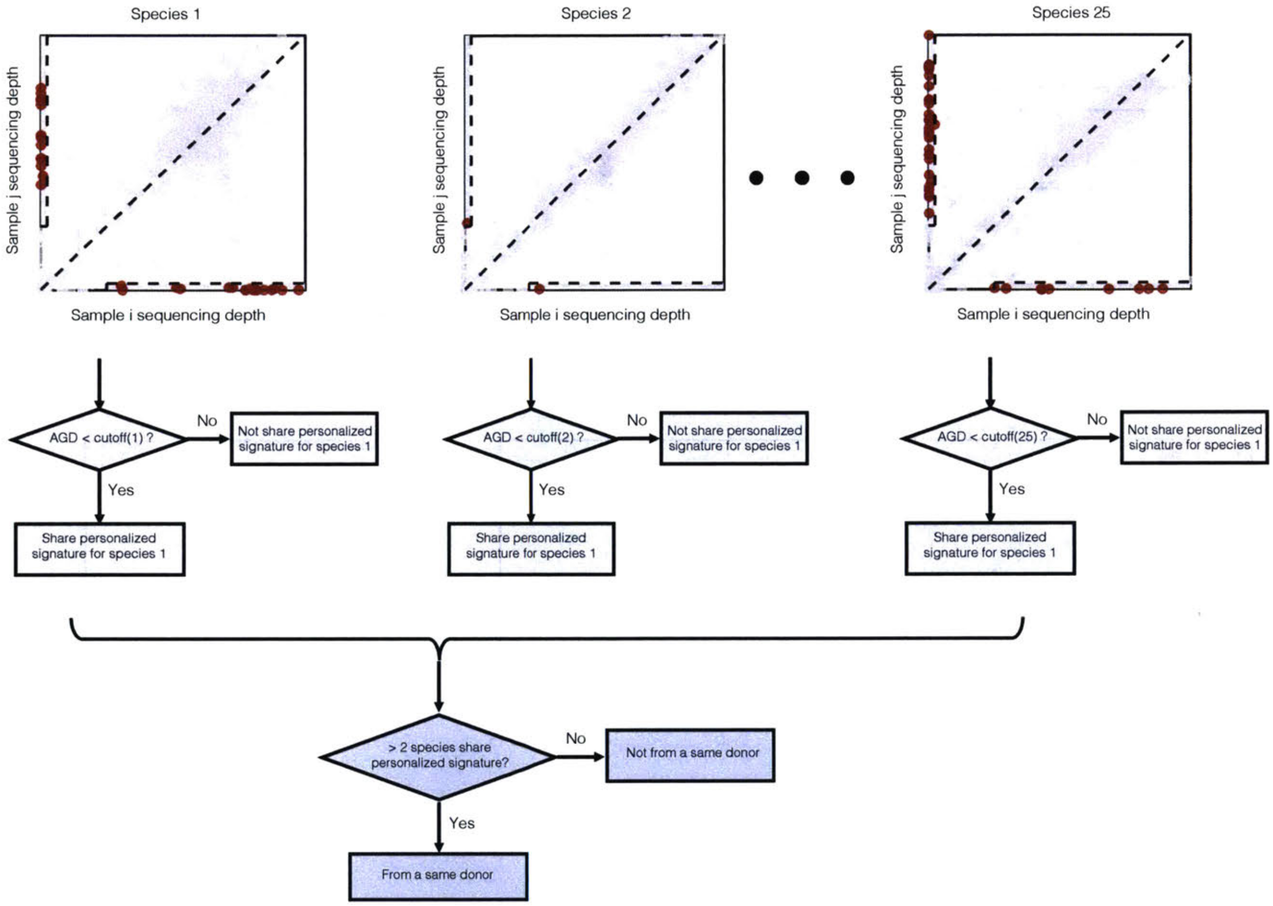

Species 1
Sample j sequencing depth
Sample i sequencing depth
Species 2
Sample j sequencing depth
Sample i sequencing depth
Species 25
Sample j sequencing depth
Sample i sequencing depth
AGD < cutoff(1) ?
No
Not share personalized signature for species 1
Yes
Share personalized signature for species 1
AGD < cutoff(2) ?
No
Not share personalized signature for species 1
Yes
Share personalized signature for species 1
AGD < cutoff(25) ?
No
Not share personalized signature for species 1
Yes
Share personalized signature for species 1
> 2 species share personalized signature?
No
Not from a same donor
Yes
From a same donor

Figure 3

A

DonorFinder performance on HMP metagenomes

	Pairs belongs to a same individual	Pairs from different individuals
Predicted as from a same individual	478 (481)	19 (16)
Predicted as from different individuals	21 (18)	142335 (142330)

- Sensitivity: 95.79% (96.39%)
- Specificity: 99.99% (99.99%)

B

DonorFinder performance on BN10 metagenomes

	Pairs belongs to a same individual	Pairs from different individuals
Predicted as from a same individual	24289	0
Predicted as from different individuals	0	59556

- Sensitivity: 100%
- Specificity: 100%

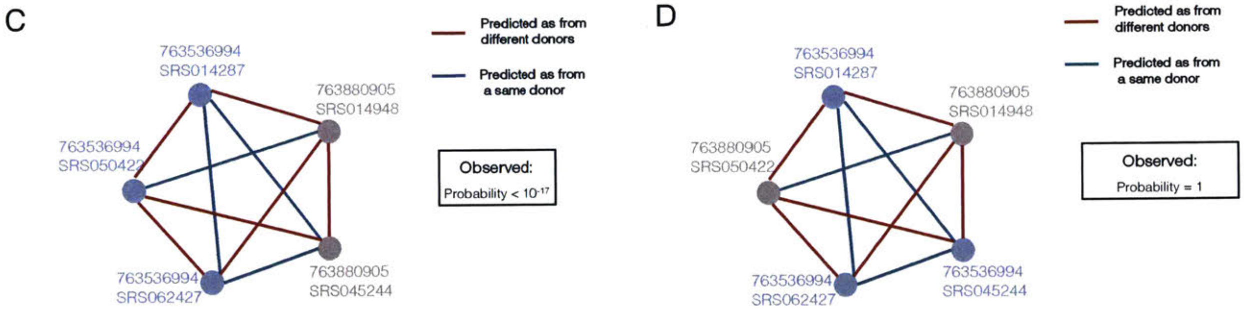


Figure 4

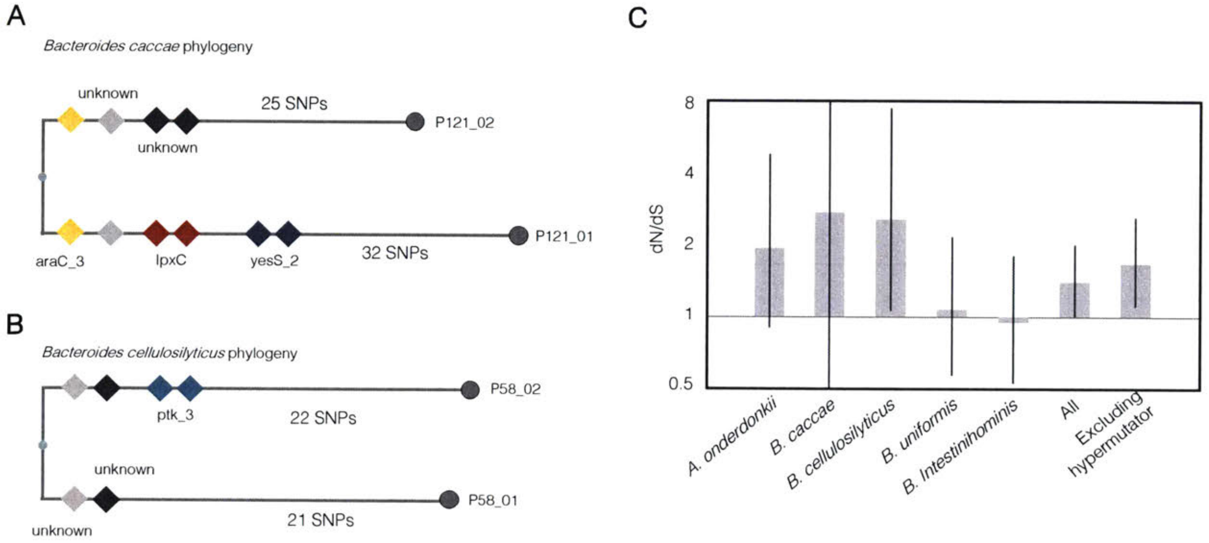

A
Bacteroides caccae phylogeny
unknown
25 SNPs
unknown
P121_02
araC_3
lpxC
yesS_2
32 SNPs
P121_01
B
Bacteroides cellulosilyticus phylogeny
P58_02
ptk_3
22 SNPs
unknown
P58_01
unknown
21 SNPs
C
8
4
2
1
0.5
dN/dS
A. onderdonkii
B. caccae
B. cellulosilyticus
B. uniformis
B. Intestinihominis
All
Excluding hypermutator

Figure S1

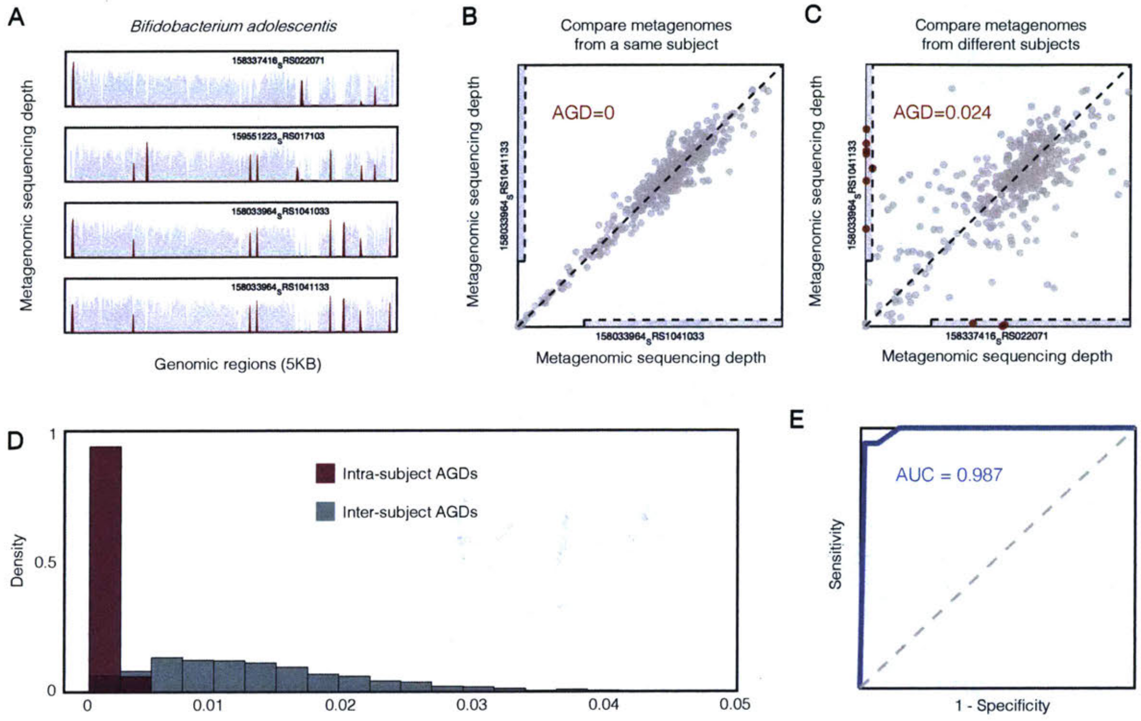

A
Bifidobacterium adolescentis
Metagenomic sequencing depth
158337416_sRS022071
159551223_sRS017103
158033964_sRS1041033
158033964_sRS1041133
Genomic regions (5KB)
B
Compare metagenomes from a same subject
AGD=0
Metagenomic sequencing depth
158033964_sRS1041133
158033964_sRS1041033
Metagenomic sequencing depth
C
Compare metagenomes from different subjects
AGD=0.024
Metagenomic sequencing depth
158033964_sRS1041133
158337416_sRS022071
Metagenomic sequencing depth
D
Density
1
0.5
0
Intra-subject AGDs
Inter-subject AGDs
0
0.01
0.02
0.03
0.04
0.05
E
AUC = 0.987
Sensitivity
1 - Specificity

Figure S2

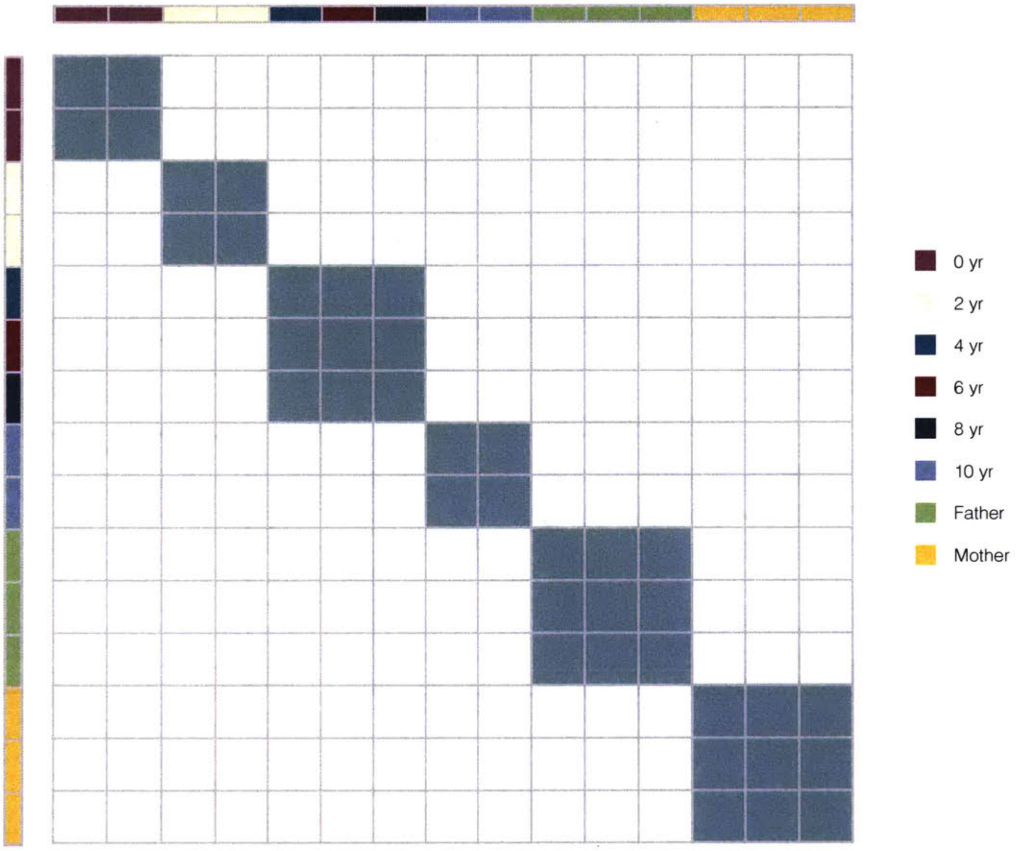


Figure S3

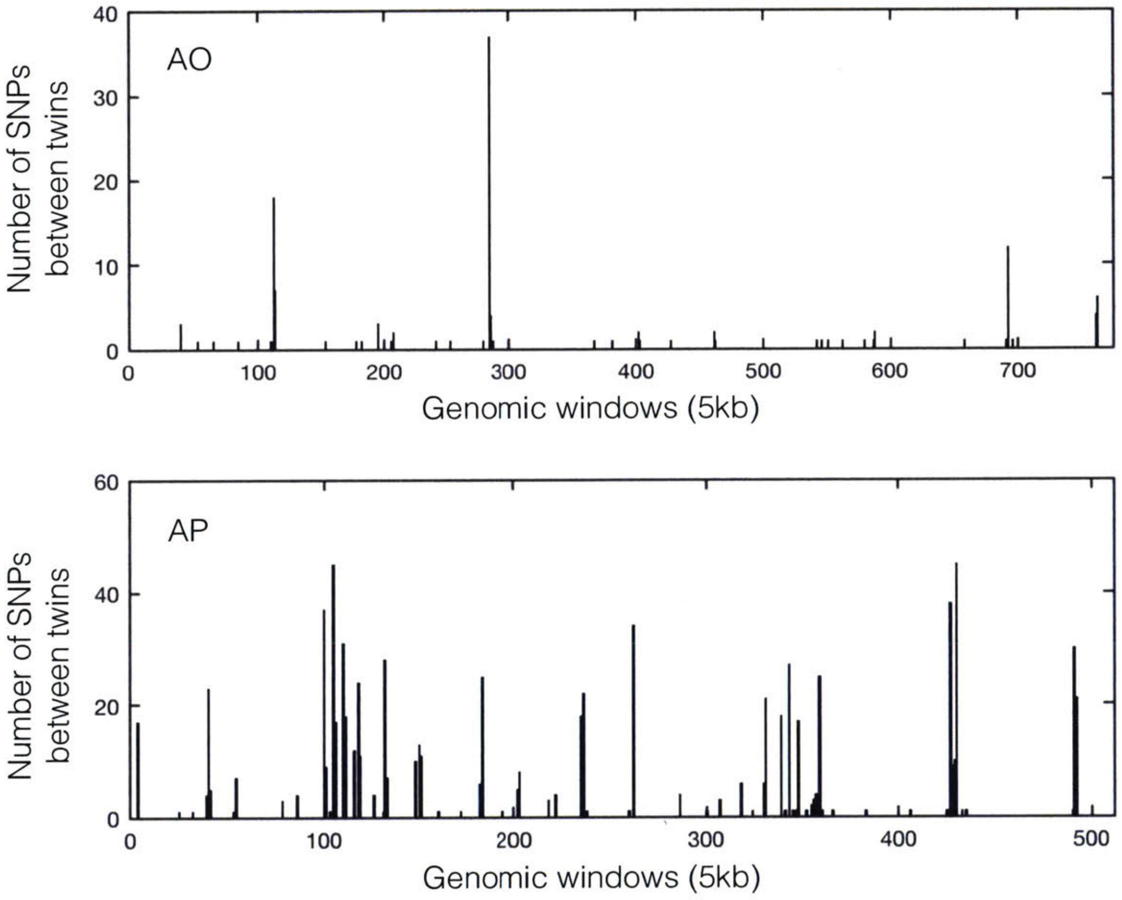

AO
Number of SNPs between twins
0
10
20
30
40
0
100
200
300
400
500
600
700
Genomic windows (5kb)
AP
Number of SNPs between twins
0
20
40
60
0
100
200
300
400
500
Genomic windows (5kb)

Table S1: Species used in DonorFinder; related to Figure 1 and Figure 2

Species	AUC	cutoff	TP rate	TF rate
[Eubacterium] eligens	0.948	0.0036	96.0%	84.8%
[Eubacterium] rectale	0.944	0.0044	96.8%	88.3%
Acidaminococcus intestini	0.978	0.0001	95.6%	100.0%
Akkermansia muciniphila	0.934	0.0001	95.0%	92.5%
Alistipes finegoldii	0.979	0.0027	99.7%	95.5%
Alistipes onderdonkii	0.988	0.0039	98.3%	97.2%
Alistipes putredinis	0.976	0.0001	99.5%	95.0%
Alistipes shahii	0.973	0.0054	99.1%	94.6%
Bacteroides caccae	0.979	0.0011	97.8%	95.9%
Bacteroides cellulosilyticus	0.989	0.0022	99.9%	98.2%
Bacteroides dorei	0.983	0.0085	99.3%	94.7%
Bacteroides eggerthii	1.000	0.0001	100.0%	100.0%
Bacteroides fragilis	0.974	0.0029	99.8%	93.5%
Bacteroides helcogenes	NA	NA	NA	NA
Bacteroides massiliensis	1.000	0.0087	98.9%	99.2%
Bacteroides ovatus	0.983	0.0055	98.7%	95.0%
Bacteroides stercoris	0.992	0.0075	99.5%	99.1%
Bacteroides thetaiotaomicron	0.999	0.0056	99.4%	98.1%
Bacteroides uniformis	0.989	0.0043	99.3%	96.4%
Bacteroides vulgatus	0.983	0.003	99.9%	96.9%
Barnesiella intestinihominis	0.980	0.0015	98.8%	94.6%
Bifidobacterium adolescentis	0.987	0.0001	98.5%	94.1%
Bifidobacterium longum	0.968	0.0001	93.7%	100.0%
Collinsella aerofaciens	0.979	0.0001	100.0%	92.3%
Coprococcus comes	0.994	0.0001	98.9%	100.0%
Dialister invisus	0.923	0.0001	97.4%	89.2%
Dorea formicigenerans	1.000	0.0001	100.0%	100.0%
Escherichia coli	0.608	0.0022	90.1%	50.0%
Faecalibacterium prausnitzii	0.941	0.0033	93.3%	91.2%
Odoribacter splanchnicus	0.990	0.0035	99.9%	98.1%
Parabacteroides distasonis	0.989	0.0053	98.4%	97.6%
Parabacteroides merdae	0.990	0.0034	98.6%	95.8%
Paraprevotella clara	0.998	0.0024	99.5%	98.3%
Parasutterella excrementihominis	0.987	0.0018	99.0%	99.0%
Roseburia hominis	0.918	0.007	94.3%	88.5%
Roseburia intestinalis	0.973	0.0091	98.9%	90.0%
Roseburia inulinivorans	0.971	0.0062	96.0%	85.7%
Ruminococcus bromii	NA	NA	NA	NA
Sutterella wadsworthensis	0.973	0.0001	94.6%	100.0%
Tyzzerella nexilis	0.984	0.0001	96.8%	100.0%

Table S2: closely-related strains shared between adult twins; related to Figure 4

Twin Pairs	Species	Conclusion from SNP analysis	Number of SNPs swept in twin 1
P64	Alistipes onderdonkii	Possibly mixed strains	NA
P97	Alistipes onderdonkii	Closely-related strains	8
P113	Alistipes onderdonkii	Possibly mixed strains	NA
P119	Alistipes onderdonkii	Possibly mixed strains	NA
P126	Alistipes onderdonkii	Possibly with recombination	NA
P126	Alistipes putredinis	Possibly with recombination	NA
P121	Bacteroides caccae	Closely-related strains	57
P58	Bacteroides cellulosilytic	Closely-related strains	43
P73	Bacteroides ovatus	Possibly mixed strains	NA
P15	Bacteroides uniformis	Closely-related strains	4
P19	Bacteroides uniformis	Possibly mixed strains	NA
P27	Bacteroides uniformis	Closely-related strains	56
P29	Bacteroides uniformis	Possibly mixed strains	NA
P35	Bacteroides uniformis	Possibly mixed strains	NA
P54	Bacteroides uniformis	Possibly mixed strains	NA
P89	Bacteroides uniformis	Possibly mixed strains	NA
P111	Bacteroides uniformis	Possibly mixed strains	NA
P7	Bacteroides vulgatus	Possibly mixed strains	NA
P15	Bacteroides vulgatus	Possibly mixed strains	NA
P29	Bacteroides vulgatus	Possibly mixed strains	NA
P45	Bacteroides vulgatus	Possibly mixed strains	NA
P62	Bacteroides vulgatus	Possibly mixed strains	NA
P70	Bacteroides vulgatus	Possibly mixed strains	NA
P100	Bacteroides vulgatus	Possibly mixed strains	NA
P103	Bacteroides vulgatus	Possibly mixed strains	NA
P15	Barnesiella intestinihomi	Closely-related strains	74
P47	Collinsella aerofaciens	Possibly mixed strains	NA

Chapter 4

Transmission of human-associated microbiota along family and social networks

Ilana L. Brito, Thomas Gurry, Shijie Zhao, Katherine Huang, Sarah K. Young, Terrence P. Shea, Waisea Naisilisili, Aaron P. Jenkins, Stacy D. Jupiter, Dirk Gevers and Eric J. Alm

The contents of this chapter have been published as: "Transmission of human-associated microbiota along family and social networks" at *Nature Microbiology* on March 25, 2019.

Statement of author contribution: This study was conceived by I.L.B. and E.J.A. The study was designed by I.L.B., A.P.J., S.P.J. and E.J.A. Raw data were collected by I.L.B. and W.N. Metagenomic assemblies and metrics were developed and assessed by I.L.B., T.G., S.Z., K.H., S.K.Y., T.P.S., D.G. and E.J.A. Final analyses were performed by I.L.B. and T.G. The paper was written by I.L.B., T.G. and E.J.A.

Abstract

The human microbiome, described as an accessory organ because of the crucial functions it provides, is composed of species that are uniquely found in humans. Yet, surprisingly little is known about the impact of routine interpersonal contacts in shaping microbiome composition. In a relatively 'closed' cohort of 287 people from the Fiji Islands, where common barriers to bacterial transmission are absent, we examine putative bacterial transmission in individuals' gut and oral microbiomes using strain-level data from both core single-nucleotide polymorphisms and flexible genomic regions. We find a weak signal of transmission, defined by the inferred sharing of genotypes, across many organisms that, in aggregate, reveals strong transmission patterns, most notably within households and between spouses. We were unable to determine the

directionality of transmission nor whether it was direct. We further find that women harbour strains more closely related to those harboured by their familial and social contacts than men, and that transmission patterns of oral-associated and gut-associated microbiota need not be the same. Using strain-level data alone, we are able to confidently predict a subset of spouses, highlighting the role of shared susceptibilities, behaviours or social interactions that distinguish specific links in the social network.

4.1 Main text

Host specificity rather than generalist life histories dominate in the colonization of the gut. Thus, colonization probably depends on direct interpersonal interactions where individuals are exposed to other humans' microbiota. Nevertheless, the extent to which regular, repeated bacterial exposures result in transmission is unknown. Mother-to-child transmission can be detected early in life (S. et al. 2016; Ferretti et al. 2018), but these patterns fade, whereas other factors—environment (Rothschild et al. 2018), behaviours and genetics (Goodrich et al. 2014)—impact the strain-level composition of each adult's microbiome (Yatsunenko et al. 2012; Xie et al. 2016). The human microbiome remains remarkably stable in composition over days (David et al. 2014) and even years, at the level of strains (Xie et al. 2016; J. Faith et al. 2013), raising the question: do we exchange oral and gut commensals with our closest family and friends?

Here, we take advantage of rich familial and social network data obtained as part of the Fiji Comunity Microbiome Project (FijiCOMP) (Figure 1a and Supplementary Tables 1 and 2) to explore the role of transmission in human populations with strain-level resolution. Our data consist of shotgun metagenomic sequences from 287 people living in 5 agrarian villages in the Fiji Islands (Supplementary Tables 3 and 4). Paired gut and oral microbiome samples were deeply sequenced to enable molecular epidemiological analyses. The presence of locally endemic bacterial disease suggests that commensal bacteria may also spread widely within the community. Owing to the relative isolation of these villages and the reliance on local food and water, we hypothesized that, with comprehensive sampling of eligible individuals in each village, we could capture all human sources and sinks of human-associated bacteria, enabling the tracking of strains within this comparatively 'closed' network.

The bacteria present in the FijiCOMP microbiomes are largely distinct from those in existing databases (I. L. Brito et al. 2016), resulting in poor read alignments to reference genomes (Supplementary Figure 1). Thus, we binned reads derived from oral or gut microbiomes using latent strain analysis (LSA) (Cleary et al. 2015), and de novo assembled a set of draft genomes (Supplementary Table 5). There were little-to-no detectable differences in species-level sharing than expected by chance across any relationship type in either the gut or oral microbiome samples (Figure 1b,c and Supplementary Figure 2), a finding at odds with that of households in Kenya (Mosites et al. 2017), Israelis (Rothschild et al. 2018) or metropolitan Americans (Yatsunenko et al. 2012), yet one that may reflect the high contact rates between individuals in this cohort.

To achieve strain-level resolution within individuals' microbiomes, we employed two orthogonal approaches, focusing on either polymorphisms in core proteins, or the presence or absence of flexible genomic regions. The former involved aligning sequencing reads to sets of core genes from each of the assemblies (Supplementary Table 5), similar to several established methods (Rothschild et al. 2018; Goodrich et al. 2014), adjusted for use within the context of a social network. Specifically, we calculated the Manhattan distances between pairs of individual's putative genotypes, inferred by the dominant single-nucleotide polymorphism (SNP) at each polymorphic position in the alignment. For individuals in the same village, household members or non-nuclear connections, we compared the distances for each genome of all connected pairs and a balanced random subset of unconnected pairs, whereas we simply shuffled the associations of spouses and mother–child pairs. We performed 100 bootstraps of the unconnected pairs or shuffles, each time tallying the number of genomes for which the median Manhattan distance was lower in connected individuals versus unconnected individuals (Figure 1b,c). We next implemented an alternate strategy, largely based on the previous observation that flexible genomic regions may be highly personalized (Franzosa et al. 2015). Coverage of 1-kb windows of contigs over 10 kb were compared across pairs of individuals. Shared genotypes were defined by the complete lack of outlying 1-kb regions present in one individual and absent in the other (Supplementary Figure 3). We tallied the number of assembled genomes more frequently shared in each relationship type in over 100 shuffles or bootstraps, again controlling for class imbalances, resulting in the distributions in Figure 1.

Transmission, loosely defined by shared inferred genotypes, has been observed for strains within the gut microbiomes of mother–child pairs (Segata et al. 2018), albeit most notably in the first year of life, in cases in which faecal material was used for transplantation (Smillie et al. 2018), and between twin pairs (Xie et al. 2016). Within the village setting, we are unable to determine whether strain transfer is direct or indirect, or from a common source, nor can we infer its directionality. However, we refer to the presence of shared genotypes as 'transmission' as the putative explanation for the observed patterns. Here, consistent patterns of transmission were revealed across individuals' social networks in both gut and oral microbiomes, independent of the metric used (Figure 1b,c and distributions of P values in Supplementary Figure 4). Household members showed high levels of strain similarities in their gut microbiomes, across mother–child pairs and, most notably, among spouses, who share no genetic relatedness. The length of cohabitation was positively correlated, albeit weakly, with the measure of strain dissimilarities (Supplementary Figure 5), which may reflect long-term changes in intimacy or lifestyle.

The signal varies across our two metrics, potentially highlighting interactions in which organisms versus mobile genetic elements are transmitted between individuals. Using a set of gut microbiome mobile genes previously identified in the FijiCOMP cohort (I. L. Brito et al. 2016), we find mobile genes weakly shared between spouses (Supplementary Figure 6). Using strain-level metrics, the transmission signals are robust. Transmission within villages in both gut and oral microbiomes was detectable in core gene SNPs, even when we rarefied the number of village pairs from over 1,000 down to 10 pairs each of connected and unconnected individuals (Supplementary Figure 7). Furthermore, our results were consistent even when we reduced the number of genomes considered using only those genomes from LSA-informed assemblies with low putative contamination (Supplementary Figure 8). In all cases, shuffling network relations, while retaining network architecture, ablated observable transmission patterns (Supplementary Figure 9).

We next examined the contributions of specific organisms, as familial transmission has been previously observed for certain gut and oral commensals (Goodrich et al. 2014). There was no consistent signal of transmission across any single phyla (Supplementary Figure 10). Instead, each pair of connected individuals had a unique signature of shared organisms (Figure 2a,b and Supplementary Figs. 11–18), suggesting that transmission may be largely driven by chance events and indirect transfer. Interestingly, the fidelity of our LSA-informed assemblies did not strongly affect our results, as transmission may still be observed even if core genes are shuffled between assemblies (Supplementary Figure 19), supporting the notion that signatures of transmission are distributed broadly over many strains. Microbiome functional profiles also failed to capture transmission signals (Supplementary Figure 20), although this does not negate the potential contributions of individual virulence-associated or transmission-associated genes contributing to transmissibility. We hypothesized that perhaps the abundance of each organism would be indicative of its overall transmissibility, favouring a mass-action model of transmission, yet this was not the case (Figure 2c,d).

These findings lead to an apparent paradox: if most bacteria are transmitted directly between members of the community, then why don't we observe clearer patterns of transmission? We believe there are several factors that contribute to the 'diffuse' signal for transmission observed across this population. First, despite this relatively 'closed' network of individuals, there are inherent difficulties in capturing the full range of individuals' contacts and exposures. Our best approximations of direct transfer may be far from actual events, where indirect transfer between individuals outside the network or transmission from unknown and unsampled environmental reservoirs may play a consequential role. Second, we focus on a snapshot in time, not knowing a priori what types of interpersonal interactions result in transfer nor whether transmission occurs during particularly volatile points in an individual's microbiome history. Third, despite our achieved sequencing depth, perhaps longer-read sequencing or a massive increase in sequencing depth is required to achieve greater strain resolution. We reached the limit of detecting transmission when we rarefied samples to 5 million reads (Supplementary Figure 21). Last, this community may actually be more prone to transmission between a wide range of community members, even when compared to other non-industrialized populations. This is best illustrated by regular gatherings to drink kava, in which a communal vessel and cup are shared.

Borrowing from the framework of disease ecology, we sought to test the effect of specific individuals within the social network on overall network-level transmission. 'Superspreading' is a phenomenon observed for the transmission of diseases, such as severe acute respiratory syndrome and human immunodeficiency virus, in which the majority of the transmission observed is attributable to a relatively small number of people (Lloyd-Price et al. 2017). Across our cohort, there were detectable differences in transmission per individual of both stool and saliva (Figure 3a–c,e and Supplementary Figure 22). As we cannot determine the direction of transmission, we refer to this phenomenon as 'supersharing' in this cohort. Supersharing was largely agnostic to the individual's read depth, once a threshold is achieved for obtaining accuracy in Manhattan distances (Supplementary Figure 23). Interestingly, individuals who were strong supersharers of gut microbiota were not the same as those of oral microbiota (Figure 3g), revealing differences between the transmission routes of commensals. There was also no specific association with individual's overall sharing and their network positions, either in terms of the number of connections ('degree') or the centrality (measured by 'betweenness') (Supplementary Figure 24).

Surprisingly, sharing of both gut and oral microbiota was more associated with females in the network ($P < 0.005$ for gut microbiomes; $P < 0.05$ for oral microbiomes, Pearson correlation; Figure 3d–f and Supplementary Figure 25), yet had no relationship with age (Supplementary Figure 26). Although gender-related differences in pathogenic bacterial transmission are well known, as are the myriad factors that affect exposure and susceptibility, these are less well understood for commensal microbiota with no clear mechanisms of transmission. Nevertheless, exposure risks may be associated with occupations and behaviours that are highly gendered within this cohort (housekeeping: $P < 10-15$; farming and fishing: $P < 10-15$; caring for ill family members: $P < 0.05$; and soap usage: $P < 0.05$, chi-squared tests). It remains to be determined how the transmission observed in this low-income, agrarian population would compare to a population living in an industrialized nation, where interventions such as the use of antiseptics, disinfectants and antibiotics, sanitation infrastructure and food safety restrictions may influence the transmission of commensal bacteria.

We next asked whether strain-level information alone could be used to predict specific social relationships. We implemented a machine learning approach that utilized organism abundances, core SNP profiles, flexible region similarity or combinations thereof, without considering demographics. Our household predictions were moderately accurate (area under the curve (AUC) = 0.64 ± 0.02 and 0.61 ± 0.01 for gut and oral microbiomes, respectively), whereas our model to predict spousal relationships performed better (AUC = 0.70 ± 0.03 and 0.72 ± 0.02 for gut and oral microbiomes, respectively) (Figure 4 and Supplementary Figure 27). Despite the poorer overall performance of our household models, the predictions seemed to be dependent on the network structure, as all of the relationships within some households were accurately predicted, in both gut and oral samples. Remarkably, our model reveals that close to 25% of spouses are exceedingly easy to predict with high confidence (Figure 4c,d). Within the household network, some of these spousal pairs were obscured, highlighting the subtle nature of these transmission signals. Why certain couples are easier to predict than others is unknown, but may reflect shared susceptibilities, specific behaviours or the relative importance of extra-marital relationships. Interestingly, spouses have been found to share immune repertoires (Carr et al. 2016) and households display family-specific signatures (Scott et al. 2014), providing evidence for shared exposures.

Although it is well established that shared environments significantly affect the gut microbiome composition and phenotype of isogenic mice (Aquino-Michaels et al. 2015; Rosshart et al. 2017) and that social interactions shape wild primate microbiomes (Moeller, Foerster, et al. 2016), this work opens the door to understanding the process of transmission and its implications in human society. Within this small community of individuals with relatively homogeneous living environments, diets and microbiomes, bacterial DNA alone can be used to accurately predict certain intimately linked pairs of individuals. Our research begins to tease apart relevant transmission patterns evident in a social network and a role for gender in commensal transmission, revealing that long-term intimate interactions that occur later in life, such as through marriage or cohabitation, can result in stochastic transmission events in both the gut and the oral microbiomes. Given the wide array of microbiome-associated health conditions, this study further hints at the possibility that diffuse transmission patterns of pathogenic or protective commensals may contribute to the overall health status of individuals.

4.2 Methods

Social network construction

The FijiCOMP consisted of interviewing and sampling the gut and oral microbiomes of almost 300 individuals living in 5 village communities in 2 districts approximately 50 miles away from one another on Vanua Levu in the Fiji Islands. The sampling all took place within a 4-week period, each village taking approximately 1–2 weeks. Institutional Review Board approval was received from Institutional Review Boards at Columbia University (New York City, NY, USA), the Massachusetts Institute of Technology (Cambridge, MA, USA) and the Broad Institute (Cambridge, MA, USA), and ethics approvals were received from the Research Ethics Review

Committees at the Fiji National University and the Ministry of Health in the Fiji Islands. Informed consent was obtained from all study participants.

As part of the survey, each head of the household was asked to draw their family trees, including all members of their household, even if they are not related. Individuals were specifically asked to name their spouse, if married, and the number and ages of their children. We inferred the number of years a married couple lived together by the age of their oldest child. We excluded 6 of the 63 couples from our analysis of the time they lived together because either they did not have any children or their children's ages were inconsistent (for example, if children came from a previous marriage). As houses commonly have names rather than specific addresses in these villages, individuals were asked the name of the house in which they live. Responses by individuals were cross-referenced for consistency and ambiguous links were removed from our analysis. Minor discrepancies, such as slight differences between spouses in the reporting of their children's ages that differed by 1 year were permitted. Individuals were further asked to provide the names of five individuals with whom they spent the most amount of time. Although the individuals mentioned the type of relationship (for example, mother–child, cousin, sister-in-law, friend, classmate and churchmate, among others), these relationship types were not solely relied on to define a particular relationship type. In a small number of examples, individuals cited social interaction with a third party whose identity could not be verified and were therefore not included in our analysis. In addition, some individuals mentioned siblings or parent–child relationships that could not be verified, so these were also counted as merely social interactions. This resulted in 489 unique social or familial interactions, in addition to household-level interactions. For the purposes of anonymity, the ages of individuals were rounded to the nearest 5-year increment and the number of children per person was not reported. Not all children of each family were surveyed, either because the children did not meet the inclusion criteria (they needed to be at least 8 years of age) or because they were inaccessible during the time when we were sampling. The social network was plotted using R package igraph (v.1.0.1). Network metrics (that is, betweenness and degree) were calculated using igraph standard functions.

Additional information was obtained from all participants, including having individuals name their occupation (of which domestic duties, farmer and fisherman were all possible answers), whether the individual had cared for a sick family member in the past year and whether they used soap (with possible answers: always, sometimes and never).

Alignments and identification of SNPs

We calculated the Manhattan distances between the dominant SNPs within pairs of individuals' core gene alignments. This involved aligning each individual's reads to core genes in the assembled LSA partitions, extracting polymorphic loci and determining the dominant allele at each locus. For each pair of individuals, we computed the Manhattan distance at each locus, averaged these distances across loci and computed this quantity for every assembled genome. These distances were then used for the network comparisons described in the 'Network comparisons' section.

More precisely, quality-filtered, dereplicated metagenomic data sets (on average, over 52 million and 10 million reads for our gut and oral microbiomes, respectively), devoid of human genetic material (filtered as described in Brito et al.13), were partitioned before assembly using LSA14 according to covarying k-mer content across samples. Read partitions were then assembled using Velvet (Bankevich et al. 2012). Sets of core genes were identified using AMPHORA2 (Wu and

Scott 2012). Core genes were assigned taxonomies using genera-level best hits using BLAST+ against the NCBI nr database. Partitions with complete (31 single-copy genes for bacteria) or near-complete gene sets of AMPHORA genes deriving from the same genera were retained for analysis (Supplementary Tables 3 and 4). If a core gene set contained more than two of the same assembled gene, we removed both copies of that gene.

Each individual's samples were then aligned to the sets of core genes using BWA-MEM (Heng Li 2013). Reads were subjected to more stringent trimming using TRIMMOMATIC (Bolger, Lohse, and Usadel 2014) (in addition to trailing low-quality base pairs (quality < 4), we also implemented a sliding window, trimming when the quality was <15). Reads were then aligned to regions that included one read length (100 bp) downstream and upstream of each core gene to avoid edge effects within the alignments. One-hundred base pairs from each end of the alignment, regardless of whether the gene was positioned at the end of the contig, were then trimmed from the final pile-up. Reads were filtered to retain those with >40% of the length aligning at 90, 95, 97 or 99% identity. A lower cut-off was chosen to capture a wide variety of strains for each alignment. Setting a lower threshold would be more inclusive of strains more distantly related to the reference, which would only obfuscate a signal for a given species should it include too distant strains. Previous work19 estimated the species boundary at approximately 85–90% identity in core genes (analogous to ~97% in the 16S rRNA gene). Ninety percent identity also resulted in the most consistent coverage across core genes, and it was therefore chosen for all subsequent SNP-level calculations. Reads with soft or hard clipping were removed. To further validate our gene sets, we filtered out genes with abnormal coverage relative to the rest of the gene set. We expected the depth of each gene to be uniform across a genome, and sequencing depth to be Poisson distributed at each locus. To avoid including genes within species' genomes recruiting abnormal numbers of reads compared to the remainder of the genome (and thus more likely to be recruiting reads from other species), we computed a chi-squared goodness-of-fit test for each gene between the empirical coverage distribution and the equivalent Poisson distribution of the same mean. Genes with a median $P < 0.05$ across subjects were discarded from any subsequent analysis. Results were mostly bimodal, where most genes fit the equivalent Poisson distribution very well, giving us confidence that reads were being recruited uniformly across the full length of the considered genes.

To calculate genome-wide statistics (Figure 1, left), we built a table of the median coverage across the SNP tables within the core genes, across different assemblies. Then, for each pair of people, we counted the number of these genomes that they shared and compared that between related and a balanced set of unlinked pairs.

Polymorphic loci were then identified from the alignment, resulting in a counts matrix for each genome containing read counts for each allele at each locus in each individual. We retained the dominant allele for each individual (the allele with the highest number of read counts) at each site, then computed the Manhattan distance between each individual's dominant allele at each site and averaged these distances across each genome to obtain an average Manhattan distance per SNP for each genome in a given pair of individuals. For each pair of individuals in a given social network (for example, the same household), this average Manhattan distance per SNP was computed for every genome, and the median distance for a given genome compared to the median distance observed in unrelated pairs of individuals. This calculation is described in more detail in the 'Network comparisons' section.

As a comparison, we also ran the quality-filtered forward metagenomic reads through the MetaPhlan2 (Truong et al. 2015) pipeline.

Abundance comparisons of 1-kb windows in assembled genomes

Contigs under 10 kb were removed from LSA-assembled draft genomes. Reads were aligned to contigs with 95% identity. Reads with hard and soft clipping were removed, as were supplementary alignments. We only considered pairs in which both individuals had a median coverage of 10 or more across the genome. One-kilobase regions were considered present in an individual and absent in another if its coverage was greater than the median in the first individual, and lower than one-thousandth of the median in the other. Pairs of individuals were considered to share the same strain if there were no such 1-kb regions across the entire genome (that is, all regions were either present or absent in both individuals) and that it was present with a median coverage of 10 or more in both individuals.

Mobile genetic element analysis

For Supplementary Figure 6, we used the abundances of mobile genes identified in Brito et al.13 to determine whether there was a transmission signal. We calculated the Jensen–Shannon divergence between all pairs and compared the number of pairs within each group with a balanced, subsampled group.

Functional contribution to transmission

Genes in the LSA-assembled genomes were first clustered at 90% identity using CD-HIT (Fu et al. 2012). Representative genes were then annotated using DIAMOND (Buchfink, Xie, and Huson 2014) against the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (release

73.0). Abundances for each gene were then calculated as the median read depth across genes with over 85% coverage. Abundances were summed for each functional gene family (represented by a single KO number). For each pair, the Jensen–Shannon divergence was calculated.

Network comparisons

Network comparisons on the mean pairwise SNP distance were performed by comparing the median value of the mean pairwise distance per SNP in related pairs with those in unrelated pairs for each genome. If a genome's median pairwise distance was lower in related pairs than in unrelated pairs, it was counted as a positive hit for related, and vice versa. The total number of genomes that fell in favour of related and unrelated was then compared. Similar analyses were performed comparing sharing of 1-kb windows in assembled genomes. A genome was assigned a positive hit for related if the number of related pairs sharing the same strain of that genome exceeded the number of unrelated pairs sharing the same strain, and vice versa. To avoid artefacts arising from the fact that the number of unrelated pairs often vastly exceeds the number of related pairs, we downsampled each of the sets of unrelated pairs 100 times, resulting in the P value distributions observed in Supplementary Figure 4.

Networks considered were spousal relationships (spouses versus non-spouse), household relationships (same versus different household), mother–child relationships (mother–child versus non-mother–child), any social network connection (any connection versus no connection) and village (same versus different village). To ensure fair comparisons in the case of spousal relationships, a set of non-spousal pairs was constructed by considering all pairs possible between males of one marriage with females of a different marriage. Similarly, in the case of household relationships, a set of different household pairs was constructed by considering all pairs possible between members of one household and members of another. In addition, comparisons were also made between randomized networks of related and unrelated pairs, in which the identity of the network's nodes was shuffled but the connections preserved, thus preserving the structure of the network.

Social network predictions

For each pair of individuals, we created feature vectors containing the mean pairwise SNP distance for each genome, the relative abundance of that genome in each individual, the number of shared genomes using 1-kb outlier regions and true or false values for whether a given genome was considered to be the same strain in both individuals using the 1-kb outlier regions. These features were then used to train random forest classifiers to predict spousal and household connections, in which class-balanced data sets were constructed by downsampling the number of unrelated pairs to equal the number of related pairs (spouse/non-spouse; same household/different household). To train the random forest classifiers on different data than those used in the predictions, we performed a threefold split of the related pairs and trained on two-thirds of the data while predicting on the remaining one-third. Predictions from the three separate test sets were combined. Receiver operating characteristic (ROC) curves were constructed from the average of ten sets of threefold cross-validation, and P values were computed for the resulting AUCs using a Mann–Whitney U-statistic on the confusion matrices.

Code availability

The code for the analyses in this paper start with an alignment table in the form of a Python dictionary containing individual core genes as its highest-level keys, where for each core gene there is a M × N × 4 numpy array, for M subjects, N loci and four different alleles (A, G, C and T). The code for filtering these alignment tables into SNP tables and Manhattan distance calculations, and scripts for identifying non-shared mobile genetic elements from 1-kb regions are posted on GitHub at https://github.com/thomasgurry/fijiComp_transmission.

Data availability

Additional information on the samples can be obtained from www.fijicomp.org. All samples may be downloaded from the NCBI Short Read Archive under Bioproject PRJNA217052. Note that the name for sample SRS475548 in the Short Read Archive was incorrectly entered; it should be the oral microbiome sample from M2.33, not W2.33. All accession numbers are listed in Supplementary Table 1. Sample collection was voluntary; thus, not all of the individuals have oral and gut microbiome samples associated with the surveys.

Additional information

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

4.3 Figures titles and legends

Figure 1 | Household membership results in shared bacterial lineages

a, The family and social network of the FijiCOMP cohort, coloured by village membership. Four villages are in the same district, whereas the fifth village is in a different district. Spousal relationships are designated by edges coloured red, whereas mother–child relationships are designated by green edges. Grey edges represent all other familial or social network relations. b,c, In the gut (b) and oral (c) microbiome samples, the number of shared genomes, the number of genomes with shared core gene SNP profiles, determined by Manhattan distances, and the number of genomes sharing flexible genomic regions, determined by 1-kb genomic windows, were significantly associated with pairs of linked, rather than unlinked, members of the social network. A 'genome' refers to each assembled set of core proteins for each species (left and middle), or to each assembled LSA partition (right). Any connection refers to friendship or distant familial connections in the network, excluding nuclear family and household connections. Full P value distributions for the distributions shown are in Supplementary Figure 4. The violin plot distributions represent results from comparing the linked pairs in a given social network (red) or the shuffled network (grey) with n = 100 independent sets of the unlinked pairs obtained by bootstrapping. Whiskers inside the violin extend to points within 1.5 interquartile ranges of the lower and upper quartiles for a distribution, and the centre points represent its median. The numbers of linked pairs for each network (stool or saliva) are as follows: household (101 out of 224); spouse (29 out of 36); mother–child pairs (24 out of 50); any connection (116 out of 169); and village (3,711 out of 8,486).

Figure 2 | Organisms vary in their transmissibility across the social network

a,b, The mean Manhattan distance, prevalence (the number of individuals who harbour that organism), log10[mean abundance] and phyla are plotted for organisms in the gut (n = 29) (a) and oral (n = 36) (b) microbiomes of spouses. c,d, The mean abundance of each organism across each pair of individuals is plotted against the Manhattan distance of that organism for that pair of individuals in the gut (n = 1,988) (c) and oral (n = 1,111) (d) microbiomes. Linear regressions are plotted in red.

Figure 3 | Some individuals are 'supersharers'

a,**b**, For each person in the network, the average distance, defined as the median of the mean Manhattan distances across all genomes to all directly connected individuals, is plotted for organisms within the gut (**a**) and oral (**b**) microbiomes. The arrows point to examples in which the sharing patterns of individuals are different for gut and oral microbiota. The red and blue in plots **a** and **b** match the values plotted in parts **c** and **e**, respectively. **c**,**e**, Histograms of the average distances for each individual to all of their directly connected individuals is plotted for individuals' gut (*n* = 173) (**c**) and oral (*n* = 243) (**e**) microbiomes. **d**,**f**, The distribution of the average distances for each individual to all of their directly connected individuals is plotted for female and male individuals' gut (*n* = 173) (**d**) and oral (*n* = 243) (**f**) microbiomes. Boxes indicate the upper and lower quartiles, the whiskers extend to the highest and lowest values excluding outliers, and the centre lines indicate the medians. *P* values were obtained from one-tailed Wilcoxon rank-sum tests. **g**, Each individual's median of the mean Manhattan distances to all individuals within the same village is plotted for their gut and oral microbiomes (*n* = 142).

Figure 4 | Machine learning predicts a subset of spouses with high confidence.

a,b, ROC curves for a random forest model predicting household membership based on shared gut (a) or oral (b) microbiome strain-level data are plotted for models using

SNP profiles, shared flexible regions, both, or both with organismal abundances. Random forest models were constructed from 1,000 decision trees and without constraint on maximum tree depth. The dotted line shows an ROC in which false positives equal false negatives. The legend reports the means and standard deviations for each classifier's AUC. c,d, The social network plotted with predicted true-positive household pairs and false-negative household pairs using gut (c) or oral (d) microbiome data.e,f, ROC curves for a random forest model predicting spousal relationships based on shared gut (e) or oral (f) microbiome strain-level data are plotted for models using SNP profiles, shared flexible regions, both, or both with organismal abundances. Random forest models were constructed from 1,000 decision trees and without constraint on maximum tree depth. The dotted line shows an ROC in which false positives equal false negatives. The legend reports the means and standard deviations for each classifier's AUC. g,h, The social network plotted with predicted true-positive spousal pairs and false-negative spousal pairs using gut (g) or oral (h) microbiome data.

Supplementary Figure 1 | Alignments of FijiCOMP reads to the HMP reference genomes and LSAbinned FijiCOMP assemblies.

The total number of primary read alignments of (a) gut (N=176) and (b) oral FijiCOMP metagenomes (N=244) (each represented by a red circle) to either the 2,191 reference genomes that make up the Human Microbiome Project (HMP) reference genome collection (downloaded from https:// www.hmpdacc.org on August 3, 2018) or the complete collection of LSA-binned FijiCOMP assemblies. Reads were filtered at 95% identity. Boxplots are drawn with lines at the median, whiskers that show quartile values, and black circles representing those samples that fall outside the quartile values.

Supplementary Figure 2 | Transmission of species using MetaPhlan2-annotated metagenomes.

The number of MetaPhlan2-annotated species shared between related pairs and 100 bootstraps of unrelated pairs are shown for the (a) gut and (b) oral microbiomes. The violin plot distributions represent results from comparing the linked pairs in a given social network (red) or the shuffled network (gray) with N=100 independent sets of the unlinked pairs obtained by bootstrapping. Whiskers inside the violin extend to points within 1.5 IQRs of the lower and upper quartile for a distribution, and center points represent its median. The numbers of linked pairs for each network (gut/oral) are as follows: household (101/224); spouse (29/36); mother-child pairs (24/50); any connection (116/169); village (3,711/8,486). Red violin

Supplementary Figure 3 | Example of 1kb windows in a pair of individuals that share and do not share a strain.

Median read depths, averaged across all sites within each 1kb window, are plotted for two individuals. The example shown here is gut microbiome partition 3558 for two individuals who share (left) the organisms versus two individuals who are not counted as sharing the organism (right). Outliers are shown in red. One partition is shown for two pairs of people, those that (a) share the same strain; and (b) those that do not. To consider that two individuals shared the same strain, they were not allowed to have any outlying 1kb windows.

Supplementary Figure 4. Distributions of p-values for the down-sampled networks.

p-values for the core gene SNPs and 1kb windows shown in Figure 1 are the median p-values are for one-sided binomial tests between the linkèd members and subsampled unlinked members. Here, we show the full range of log(p-values), without multiple test correction, for the SNPs and 1kb windows in the (a) gut and (b) oral microbiomes for 100 subsampled networks each. The violin plot distributions represent results from comparing the linked pairs in a given social network (red) or the shuffled network (gray) with N=100 independent sets of the unlinked pairs obtained by bootstrapping. The numbers of linked pairs for each network (gut/oral) are as follows: household (101/224); spouse (29/36); mother-child pairs (24/50); any connection (116/169); village (3,711/8,486).

Supplementary Figure 5 | Strain transmission in spouses according to the length of cohabitation.

The length of cohabitation, inferred by the age of the couple's oldest child, is plotted for each spousal pair against the mean Manhattan distances across genomes for that pair in the (a) gut (N=25) and (b) oral (N=18) microbiomes. Linear regressions are shown in red.

Supplementary Figure 6 | Transmission of mobile genes is mildly detectable between spouses using a previously identified set of mobile genes.

We calculated the JensenShannon divergence for the vectors of abundances of mobile genes identified in Brito et al., (2016) in individuals' gut microbiomes across social networks. The violin plot distributions represent results from comparing the linked pairs in a given social network (red) or the shuffled network (gray) with N=100 independent sets of the unlinked pairs obtained by bootstrapping. Whiskers inside the violin extend to points within 1.5 IQRs of the lower and upper quartile for a distribution, and center points represent its median. The numbers of linked pairs for each network are as follows: household 101; spouse 29; mother-child pairs 24; any connection 116; village 3,711.

Supplementary Figure 7 | Transmission is detected even at small numbers of pairs in our network.

We rarefied the number of people considered in each iteration of our network analysis to determine the detection limit for the observed transmission patterns, to N=10, 20 50, 100, 500, or 1,000 pairs of individuals for the gut microbiome (left) and oral microbiome (right) samples. Whiskers inside boxplots extend to points within 1.5 IQRs of the lower and upper quartile for a distribution, and center points represent its median. The box plot distributions represent results from comparing the linked pairs in a given social network (red) or the shuffled network (gray) with N=100 independent sets of the unlinked pairs obtained by bootstrapping.

Supplementary Figure 8 | Transmission signals are maintained with fewer, albeit higher quality, partitions.

We reanalyzed 1kb segments from partitions with low amounts of putative contamination (less than 10% as determined by CheckM) for signals of transmission within the FijiCOMP cohort. 46 out of 207 gut microbiome partitions were removed, in addition to 34, for which CheckM was unable to run. 258 partitions out of 1,091 oral microbiome partitions were removed. The number of genomes shared according to 1kb segments for the (a) gut and (b) oral microbiomes are shown. The violin plot distributions represent results from comparing the linked pairs in a given social network (red) or the shuffled network (gray) with N=100 independent sets of the unlinked pairs obtained by bootstrapping. The numbers of linked pairs for each network (gut/oral) are as follows: household (101/224); spouse (29/36); mother-child pairs (24/50); any connection

(116/169); village (3,711/8,486). Whiskers inside the violin extend to points within 1.5 IQRs of the lower and upper quartile for a distribution, and center points represent its median.

Supplementary Figure 9 | Transmission signals are ablated with the shuffled networks. First the network was shuffled, maintaining the overall network structure. In the (a) gut and (b) oral microbiomes, the number of shared genomes (left); the number of genomes with shared core gene SNP profiles, determined by Manhattan distances (middle); or the number of genomes sharing flexible genomic regions, determined by 1kb genomic windows (right), whose data are represented in Figure 1; are shown for linked and subsampled unlinked members of the social network as indicated for the shuffled networks. The violin plot distributions represent results from comparing the linked pairs in a given social network (red) or the shuffled network (gray) with N=100 independent sets of the unlinked pairs obtained by bootstrapping. The numbers of linked pairs for each network (gut/oral) are as follows: household (101/224); spouse (29/36); mother-child pairs (24/50); any connection (116/169); village (3,711/8,486). Whiskers inside the violin extend to points within 1.5 IQRs of the lower and upper quartile for a distribution, and center points represent its median.

Supplementary Figure 10 | Phylogeny does not correlate with transmission patterns. In the (a) gut and (b) oral microbiome samples, the number of genomes with shared core gene SNP profiles, determined by Manhattan distances, were determined for organisms according to their phylum. For each social group, we compared 'related' versus 'unrelated', shorthand for directly linked and unlinked in the network. There were no consistent phyla associated with greater sharing across social contexts in either gut or oral microbiomes. The box plot distributions represent results from comparing the linked pairs in a given social network (red) or the shuffled network (gray) with N=100 independent sets of the unlinked pairs obtained by bootstrapping. Whiskers inside boxplots extend to points within 1.5 IQRs of the lower and upper quartile for a distribution, and center points represent its median. The numbers of linked pairs for each network (gut/oral) are as follows: household (101/224); spouse (29/36); mother-child pairs (24/50); any connection (116/169); village (3,711/8,486).

Supplementary Figure 11 | Heatmaps show a dispersed transmission signal across organisms and pairs of individuals.

For the core gene SNPs, a z-score of the Manhattan distance was calculated for each organism across the gut microbiomes of pairs of individuals living in the same household.

Supplementary Figure 12 | Heatmaps show a dispersed transmission signal across organisms and pairs of individuals.

For the core gene SNPs, a z-score of the Manhattan distance was calculated for each organism across the gut microbiomes of spouses.

Supplementary Figure 13 | Heatmaps show a dispersed transmission signal across organisms and pairs of individuals.

For the core gene SNPs, a z-score of the Manhattan distance was calculated for each organism across the oral microbiomes of pairs of individuals living in the same household.

Supplementary Figure 14 | Heatmaps show a dispersed transmission signal across organisms and pairs of individuals.

For the core gene SNPs, a z-score of the Manhattan distance was calculated for each organism across the oral microbiomes of spouses.

Supplementary Figure 15 | Heatmaps show a dispersed transmission signal across organisms and pairs of individuals.

For the 1kb genomic regions, we simply plot the presence and absence of a shared partition for gut microbiomes in pairs of individuals living in the same households.

Supplementary Figure 16 | Heatmaps show a dispersed transmission signal across organisms and pairs of individuals.

For the 1kb genomic regions, we simply plot the presence and absence of a shared partition for gut microbiomes in spouses.

Supplementary Figure 17 | Heatmaps show a dispersed transmission signal across organisms and pairs of individuals. For the 1kb genomic regions, we simply plot the presence and absence of a shared partition for oral microbiomes in pairs of individuals living in the same households

Supplementary Figure 18 | Heatmaps show a dispersed transmission signal across organisms and pairs of individuals. For the 1kb genomic regions, we simply plot the presence and absence of a shared partition for oral microbiomes in spouses.

Supplementary Figure 19 | Shuffling core genes does not ablate signals of transmission.

We shuffled the core genes that contributed to each of the genomes in the (a) gut and (b) oral microbiomes. Briefly, an equal number of new, synthetic partitions were created from the original LSA partitions, and core genes from the originals were distributed randomly to the new partitions, with the constraint that a partition could only have one copy of a given core gene. The violin plot distributions represent results from comparing the linked pairs in a given social network (red) or the shuffled network (gray) with N=100 independent sets of the unlinked pairs obtained by bootstrapping. Whiskers inside the violin extend to points within 1.5 IQRs of the lower and upper quartile for a distribution, and center points represent its median. The numbers of linked pairs for each network (gut/oral) are as follows: household (101/224); spouse (29/36); mother-child pairs (24/50); any connection (116/169); village (3,711/8,486).

Supplementary Figure 20 | Functional profiles are not indicative of transmission routes. We calculated the Jensen-Shannon divergence for abundances of genes in the (a) gut and (b) oral microbiomes, aggregated by KEGG pathway, between pairs according to each relationship type. Classes were balanced for comparison. The violin plot distributions represent results from comparing the linked pairs in a given social network (red) or the shuffled network (gray) with N=100 independent sets of the unlinked pairs obtained by bootstrapping. Whiskers inside the violin extend to points within 1.5 IQRs of the lower and upper quartile for a distribution, and center points represent its median. The numbers of linked pairs for each network (gut/oral) are as follows: household (101/224); spouse (29/36); mother-child pairs (24/50); any connection (116/169); village (3,711/8,486).

Supplementary Figure 21 | Transmission signals are not reliable detected after rarefying read counts.

We rarefied individuals' metagenomic libraries to 5 million quality-filtered reads and re-computed the number of shared genomes in the (a) gut and (b) oral microbiome according to their shared core gene SNP profiles, determined by Manhattan distances, and compared these with the shuffled social network computed using this many reads. The violin plot distributions represent results from comparing the linked pairs in a given social network (red) or the shuffled network (gray) with N=100 independent sets of the unlinked pairs obtained by bootstrapping. Whiskers inside the violin extend to points within 1.5 IQRs of the lower and upper quartile for a distribution, and center points represent its median. The numbers of linked pairs for each network (gut/oral) are as follows: household (101/224); spouse (29/36); mother-child pairs (24/50); any connection (116/169); village (3,711/8,486).

Supplementary Figure 22 | Distributions of Manhattan distances between all direct linkages for each individual.

For simplicity's sake, distributions of Manhattan distances for (a) gut and (b) oral microbiomes for individuals living in one of the villages are shown (N=51). Whiskers inside boxplots extend to points within 1.5 IQRs of the lower and upper quartile for a distribution, and center points represent its median.

Supplementary Figure 23 | Supersharing is agnostic to read depth.

The number of quality-filtered reads per individual's (a) gut or (b) oral microbiomes are plotted against their median of mean Manhattan distances across all genomes to all of their directly connected individuals (N=141).

Supplementary Figure 24 | Network statistics do not capture 'supersharing' metrics.

Node degree and 'betweenness' are plotted for each person in the (a,b) gut and (c,d) oral microbiomes.

Supplementary Figure 25 | Females average strains are closer to their female links.

The distribution of mean Manhattan distances for each individual to all of their directly connected female or male linkages is plotted for female (N=125) and male (N=118) individuals' (a) gut and (b) oral microbiomes. Boxes indicate the upper and lower quartiles, whiskers extend to highest and lowest values excluding outliers, and center lines indicate medians. P-values reflect a two-tailed Wilcoxon Rank-sum test.

Supplementary Figure 26 | Age is unrelated to 'supersharing'.

Individuals are plotted according to their age and their mean Manhattan distance to all individuals they are connected to in the (a) gut and (b) oral microbiomes.

Supplementary Figure 27 | Precision-recall curves for predictive family and social interaction models.

Precision-recall curves for the models in Figure 4, namely (a) gut microbiomes of household members; (b) gut microbiomes of spouses; (c) oral microbiomes of household members; and (d) oral microbiomes of spouses.

Supplementary tables are not included due to space limit and can be found at

https://www.nature.com/articles/s41564-019-0409-6

Figure 1

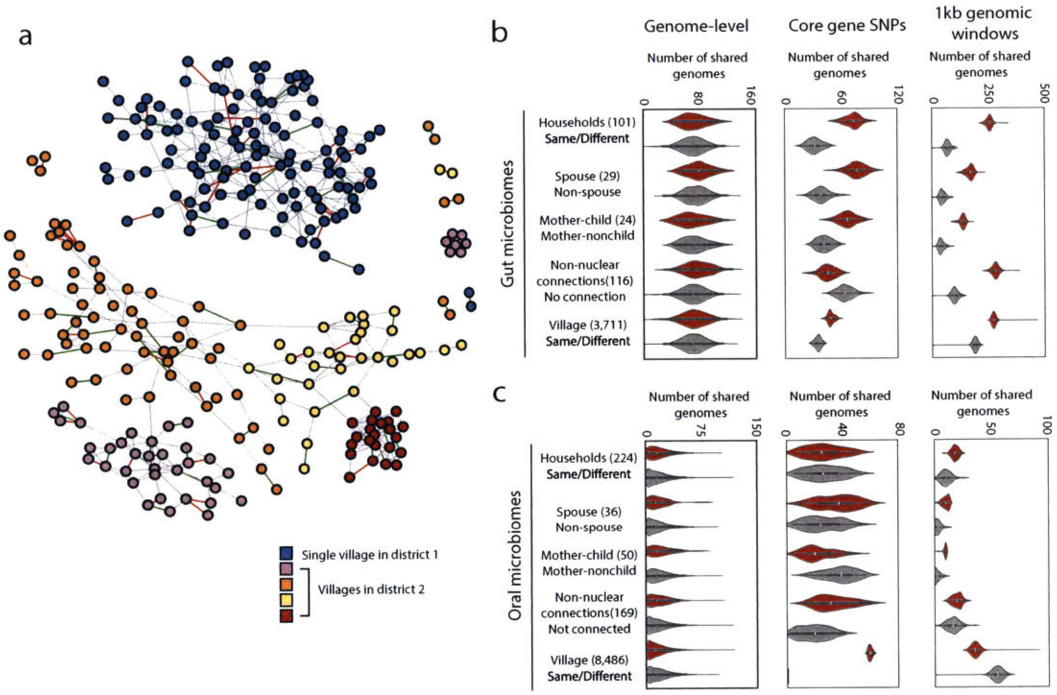


Figure 2

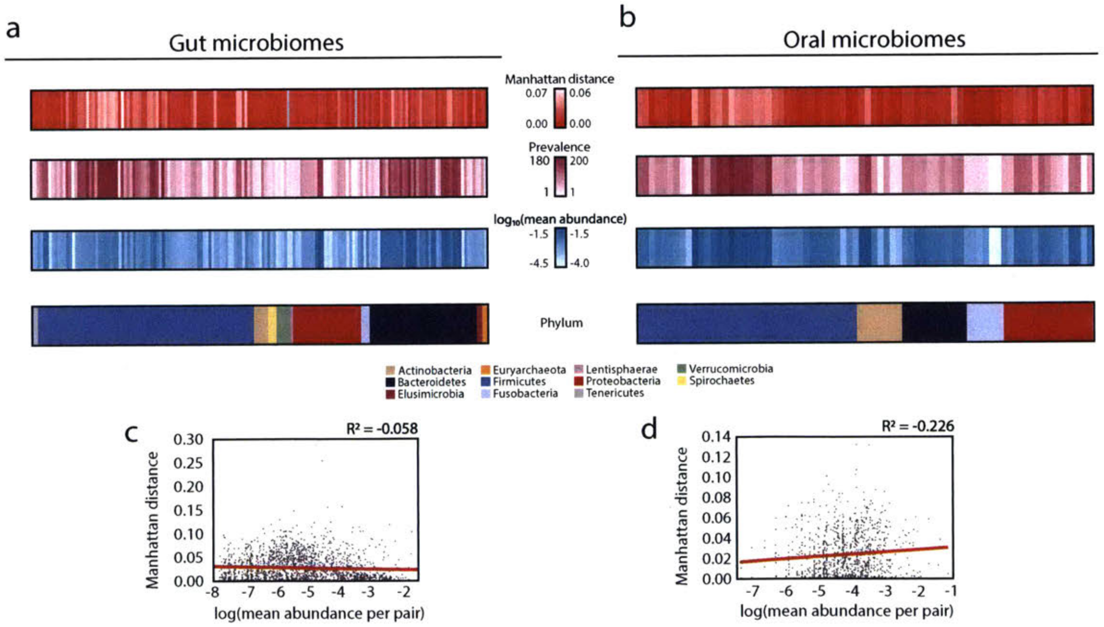


Figure 3

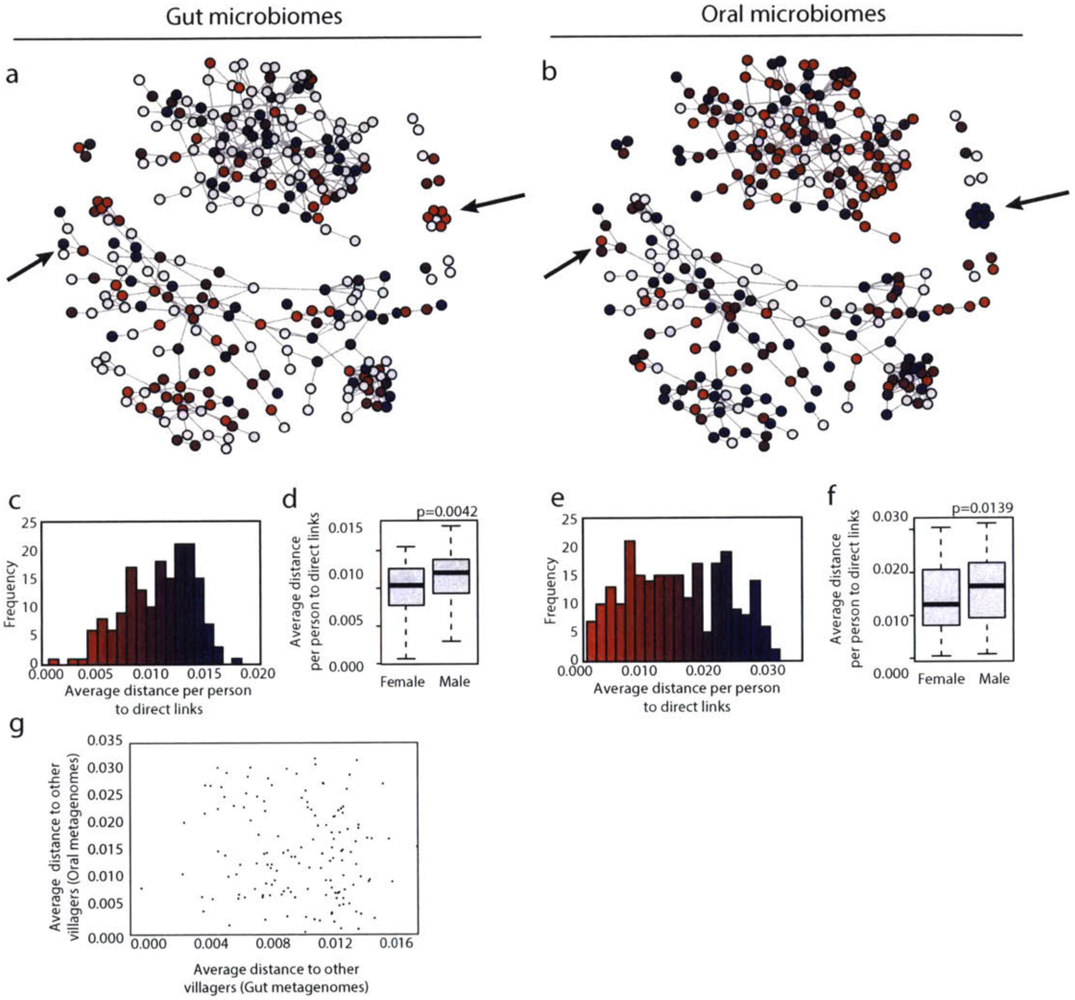
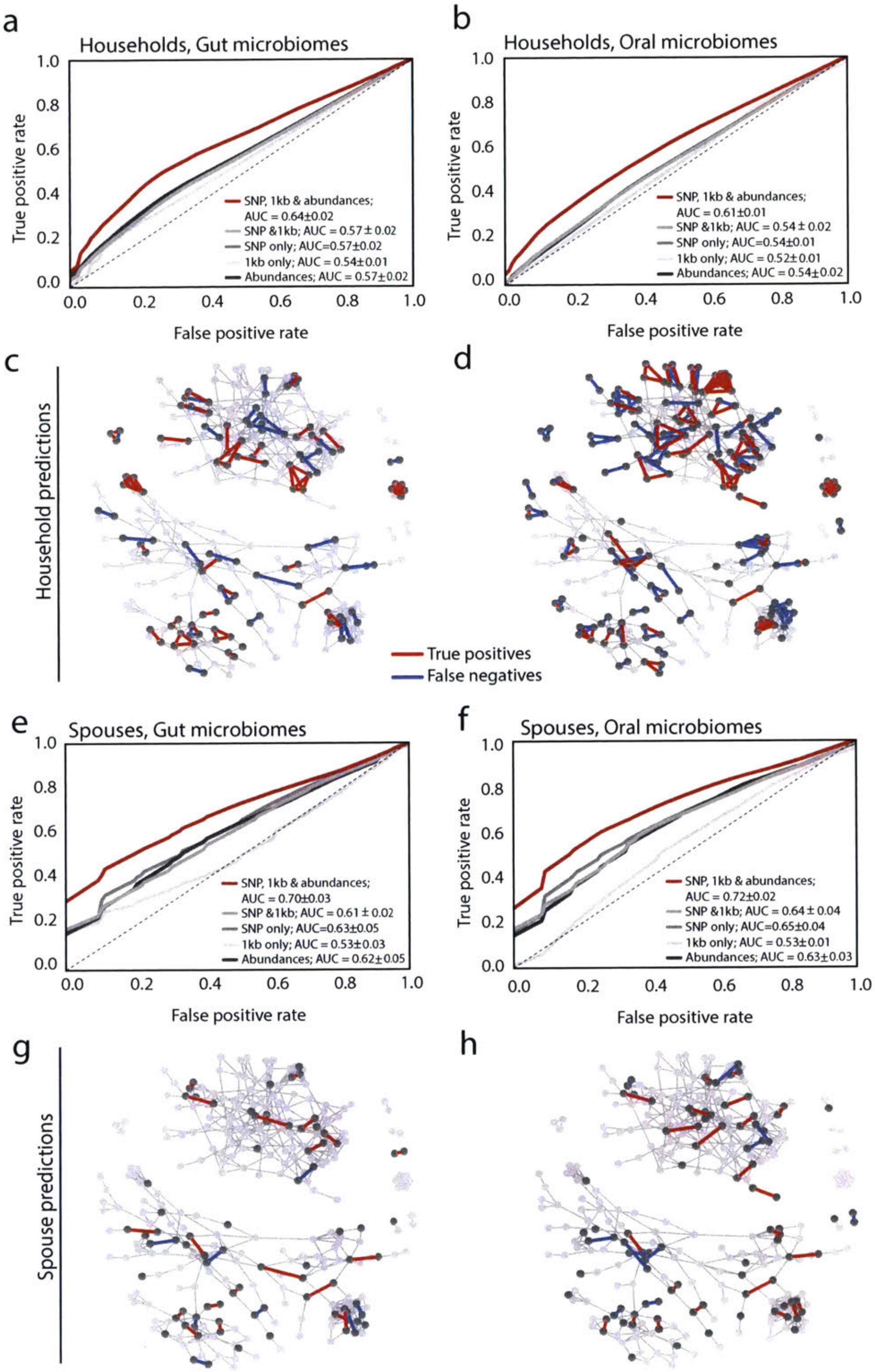
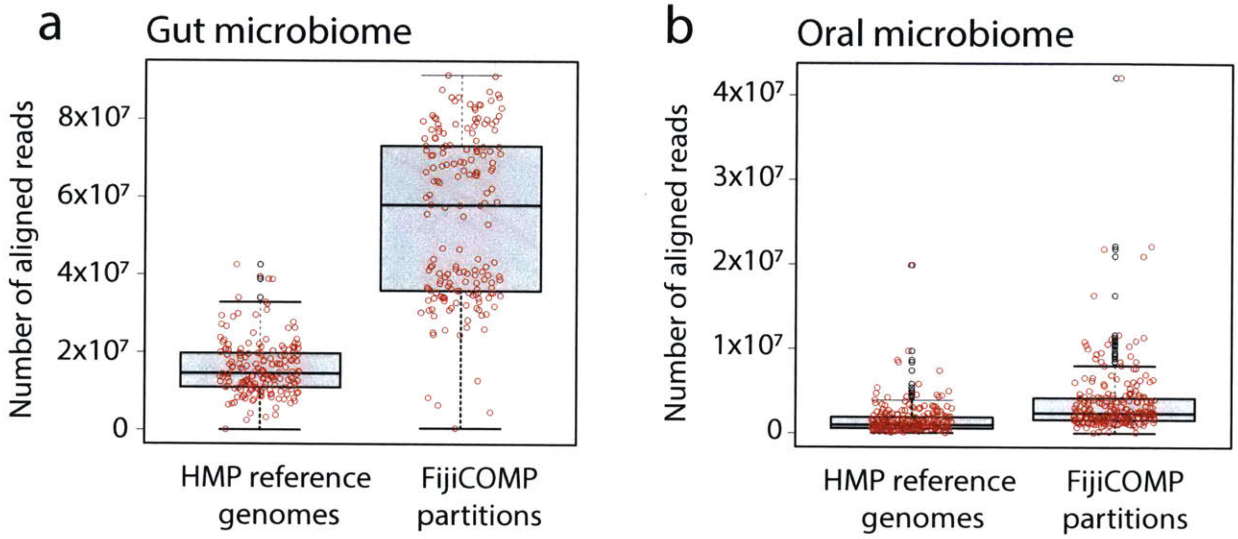

Gut microbiomes
Oral microbiomes
a
b
c
Frequency
25
20
15
10
5
0
0.000 0.005 0.010 0.015 0.020
Average distance per person to direct links
d
p=0.0042
Average distance per person to direct links
0.015
0.010
0.005
0.000
Female Male
e
Frequency
25
20
15
10
5
0
0.000 0.010 0.020 0.030
Average distance per person to direct links
f
p=0.0139
Average distance per person to direct links
0.030
0.020
0.010
0.000
Female Male
g
Average distance to other villagers (Oral metagenomes)
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0.000 0.004 0.008 0.012 0.016
Average distance to other villagers (Gut metagenomes)

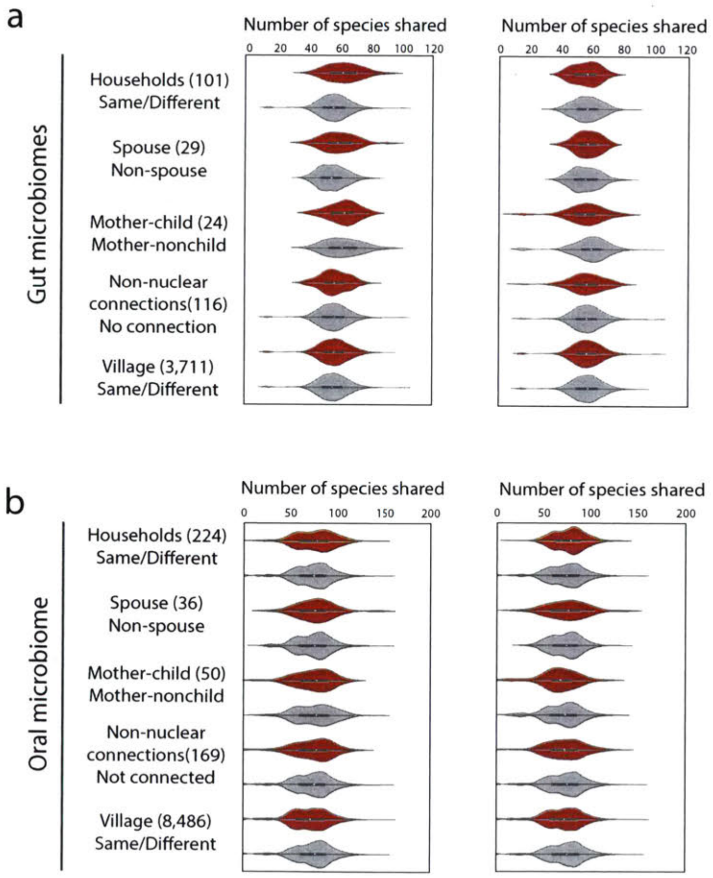
Figure 4


a
Households, Gut microbiomes
1.0
0.8
0.6
0.4
0.2
0.0
True positive rate
SNP, 1kb & abundances; AUC = 0.64±0.02
SNP &1kb; AUC = 0.57 ± 0.02
SNP only; AUC=0.57±0.02
1kb only; AUC = 0.54±0.01
Abundances; AUC = 0.57±0.02
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
b
Households, Oral microbiomes
1.0
0.8
0.6
0.4
0.2
0.0
True positive rate
SNP, 1kb & abundances; AUC = 0.61±0.01
SNP &1kb; AUC = 0.54 ± 0.02
SNP only; AUC=0.54±0.01
1kb only; AUC = 0.52±0.01
Abundances; AUC = 0.54±0.02
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
c
Household predictions
d
True positives
False negatives
e
Spouses, Gut microbiomes
1.0
0.8
0.6
0.4
0.2
0.0
True positive rate
SNP, 1kb & abundances; AUC = 0.70±0.03
SNP &1kb; AUC = 0.61 ± 0.02
SNP only; AUC=0.63±0.05
1kb only; AUC = 0.53±0.03
Abundances; AUC = 0.62±0.05
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
f
Spouses, Oral microbiomes
1.0
0.8
0.6
0.4
0.2
0.0
True positive rate
SNP, 1kb & abundances; AUC = 0.72±0.02
SNP &1kb; AUC = 0.64 ± 0.04
SNP only; AUC=0.65±0.04
1kb only; AUC = 0.53±0.01
Abundances; AUC = 0.63±0.03
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
g
Spouse predictions
h

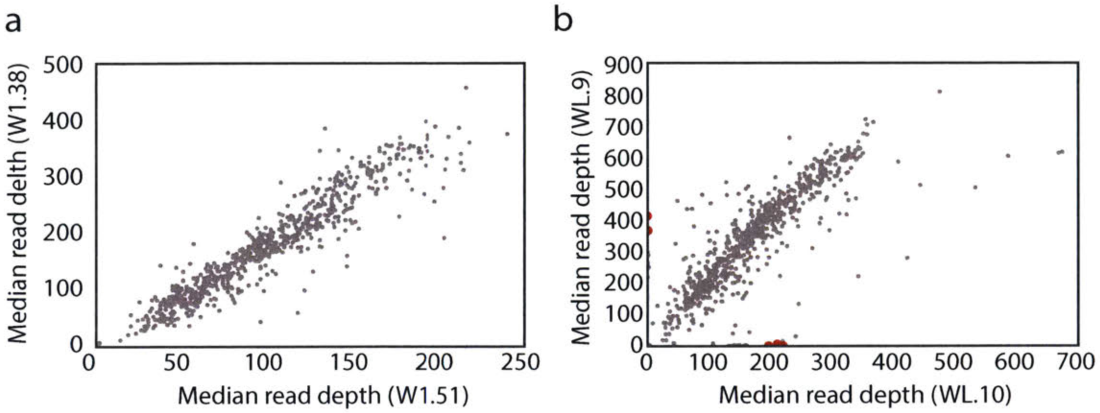
Supplementary Figure 1



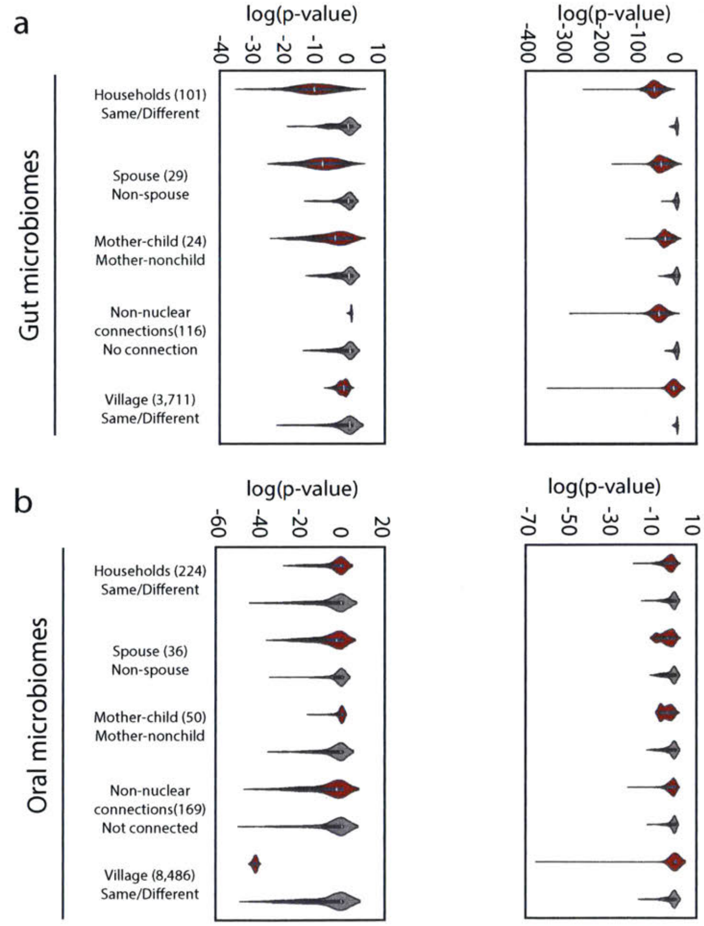
Supplementary Figure 2



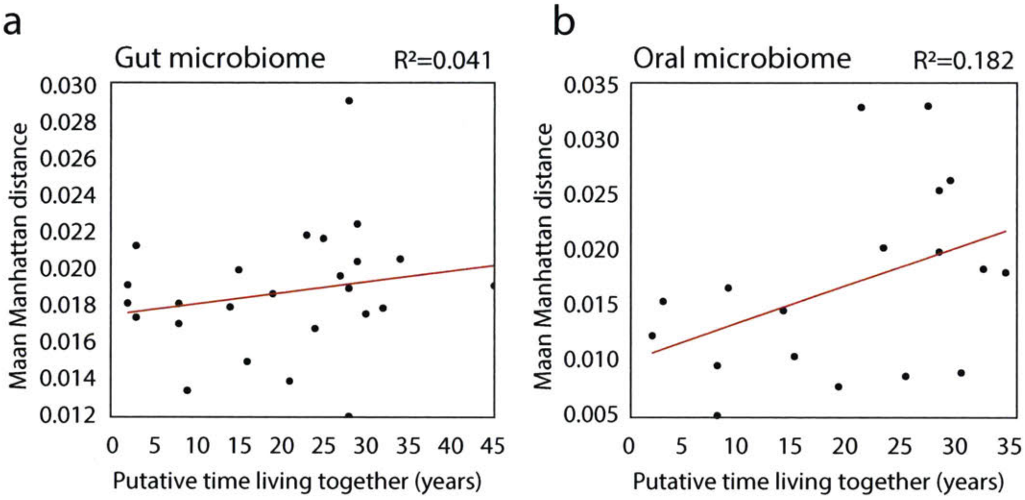
Supplementary Figure 3



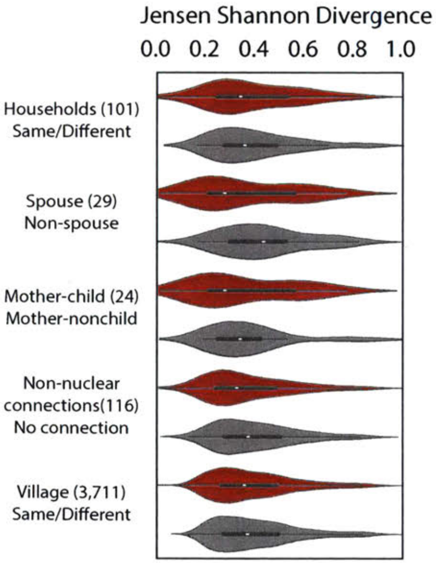
Supplementary Figure 4



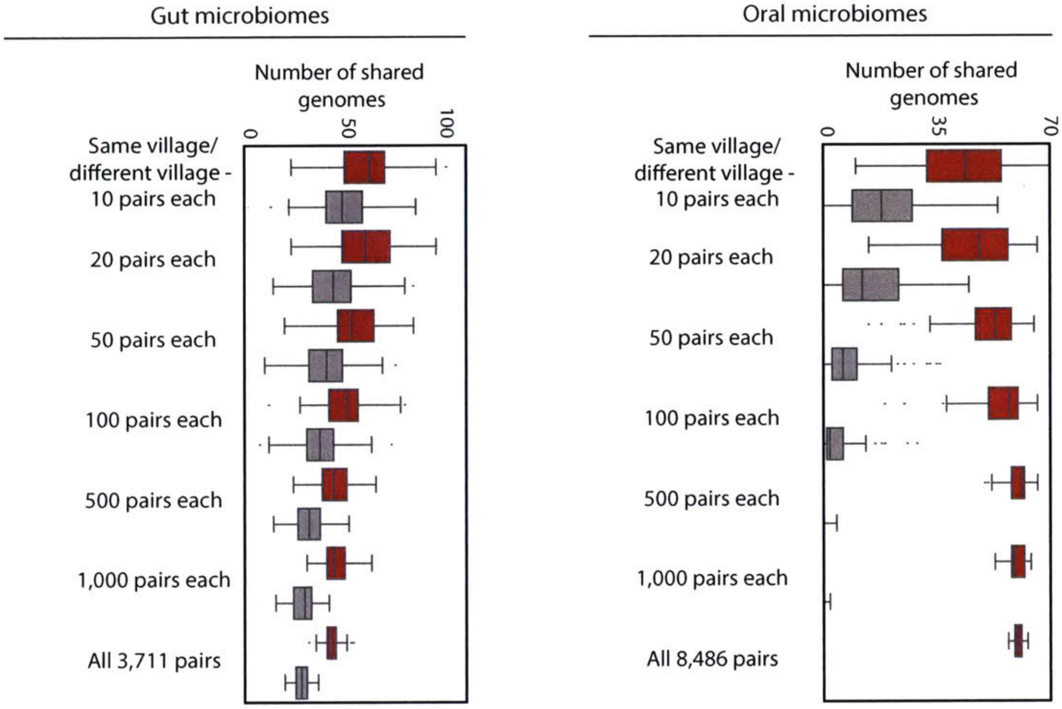
Supplementary Figure 5



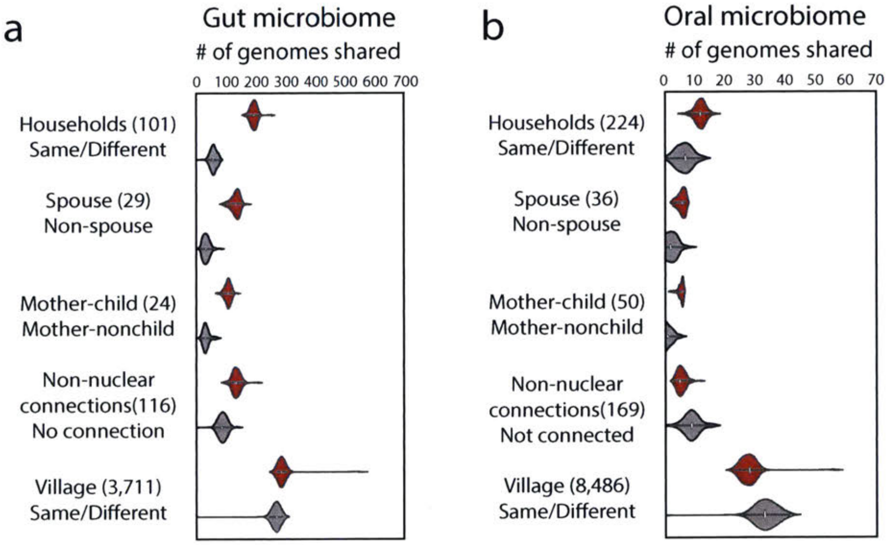
Supplementary Figure 6



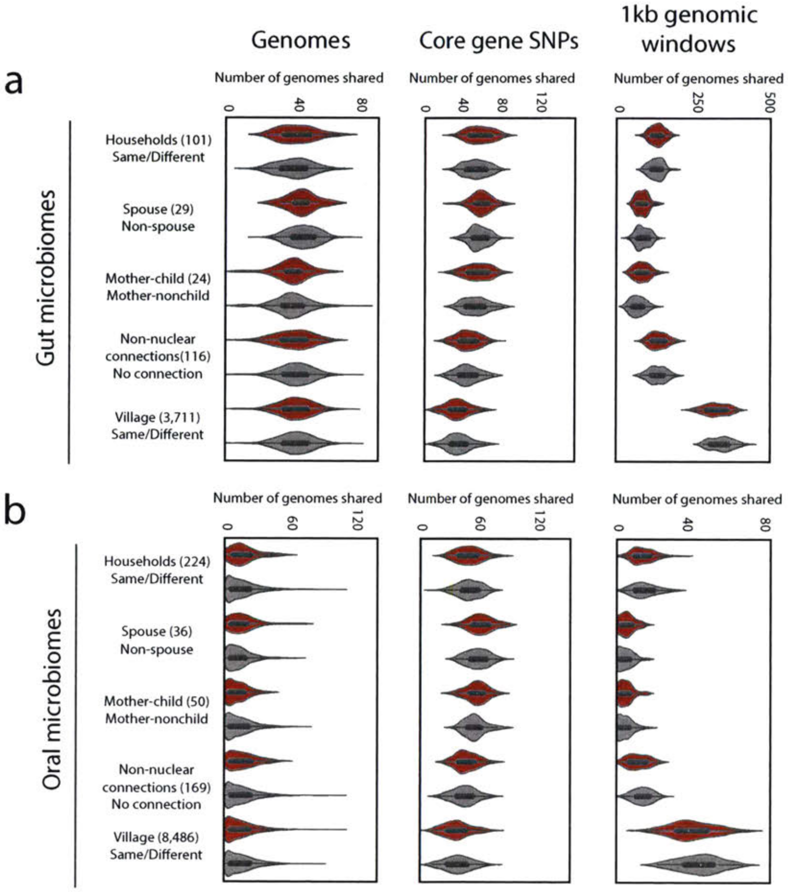
Supplementary Figure 7



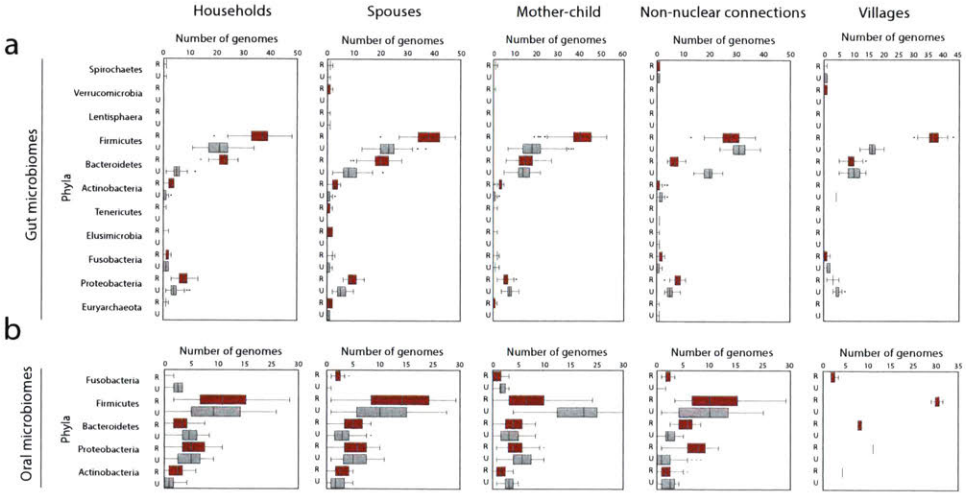
Supplementary Figure 8



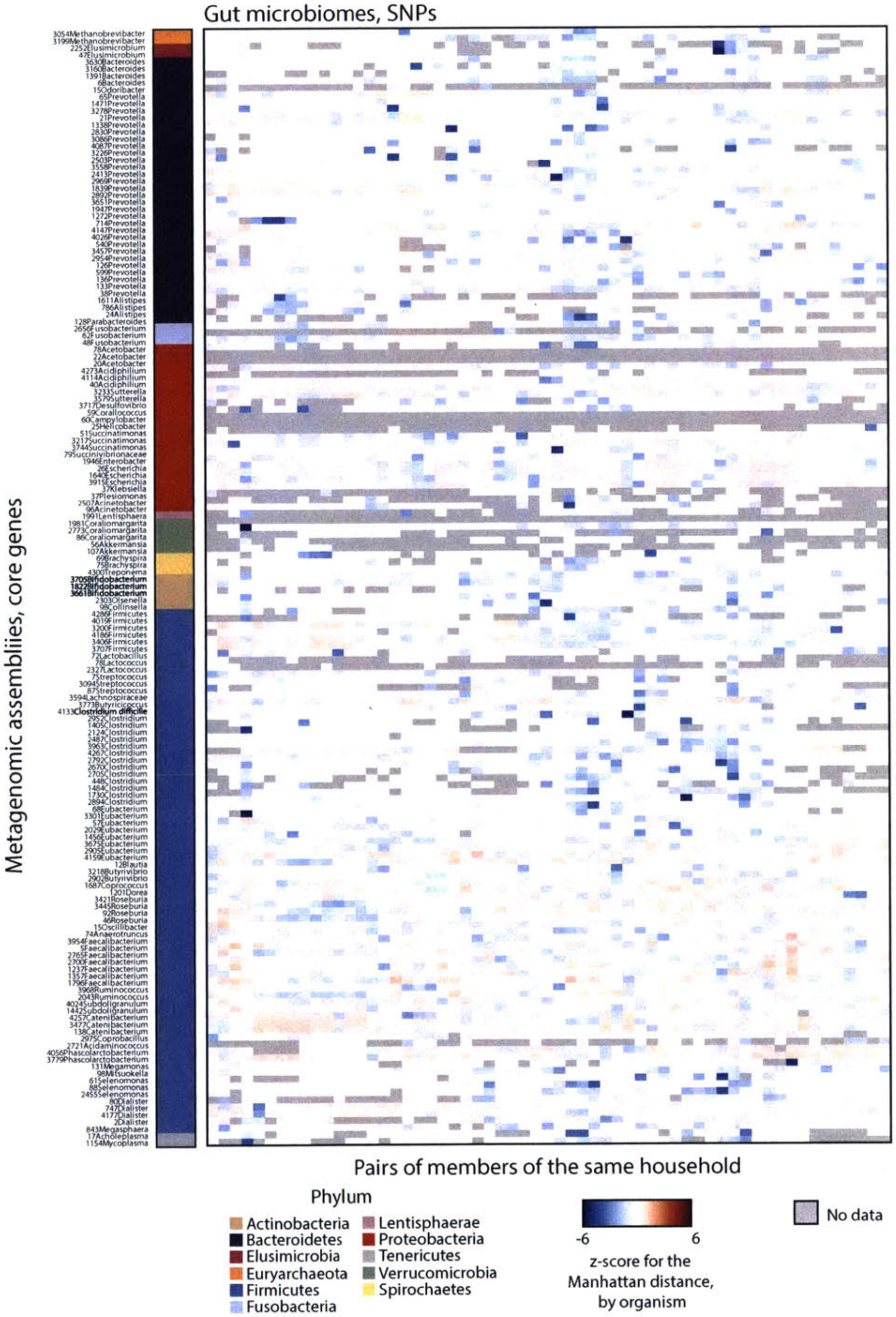
Supplementary Figure 9



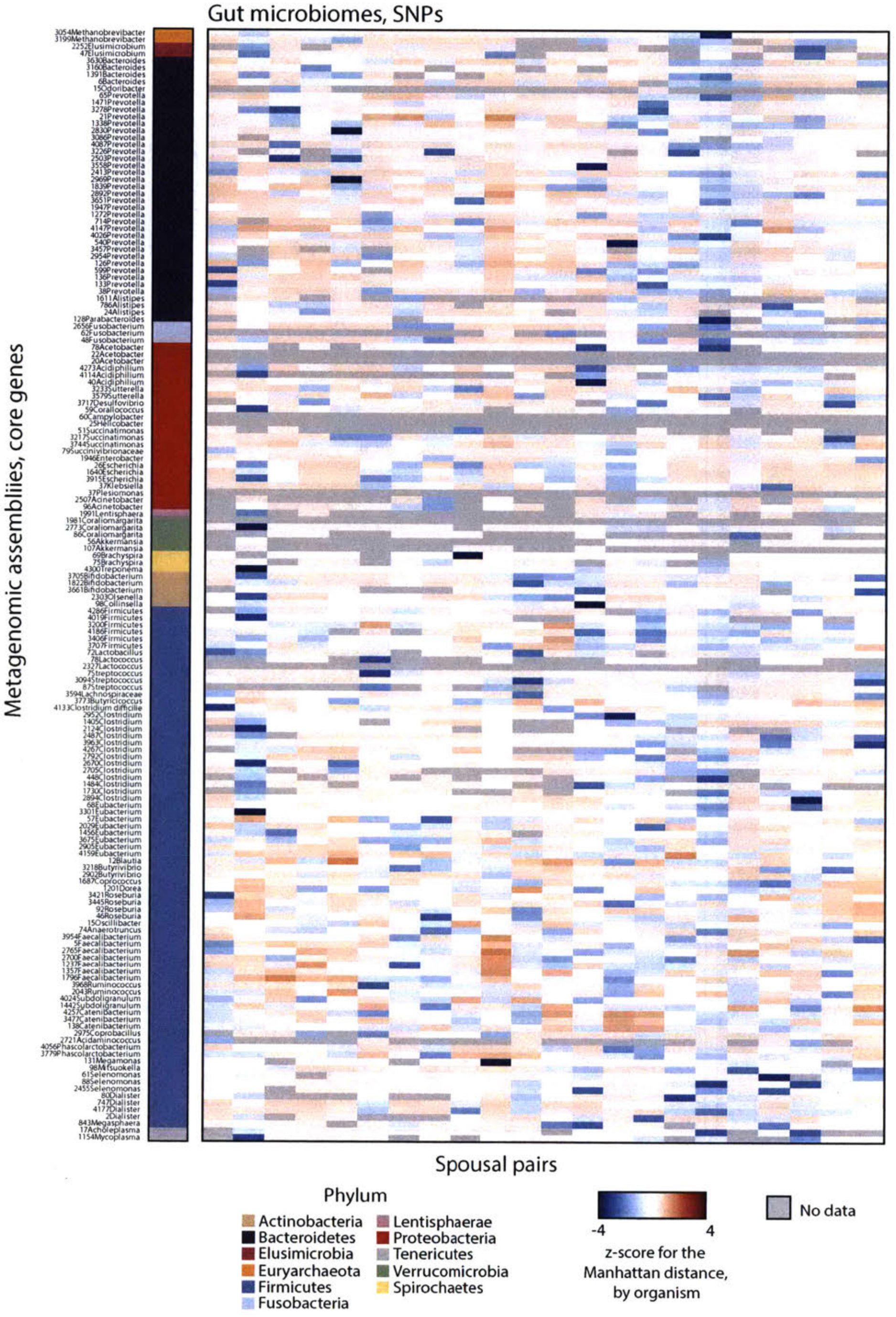
Supplementary Figure 10



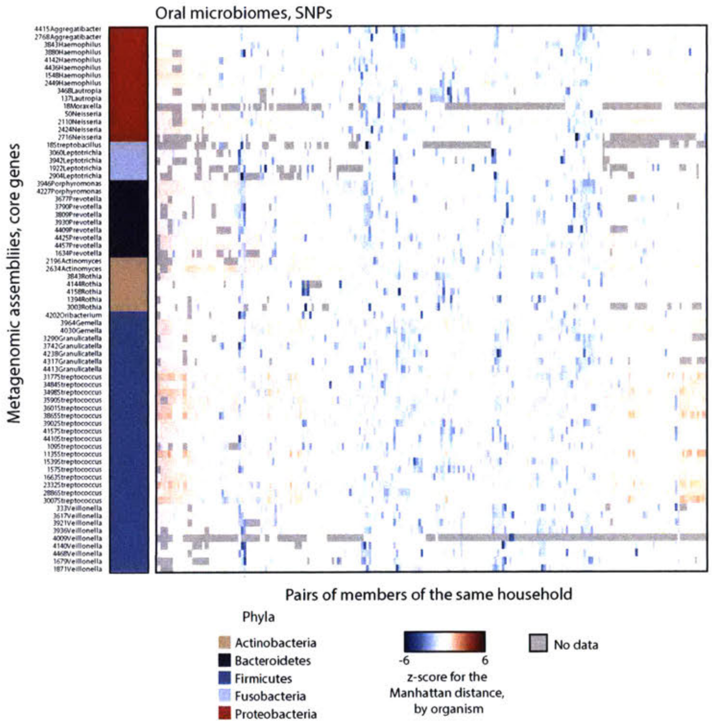
Supplementary Figure 11



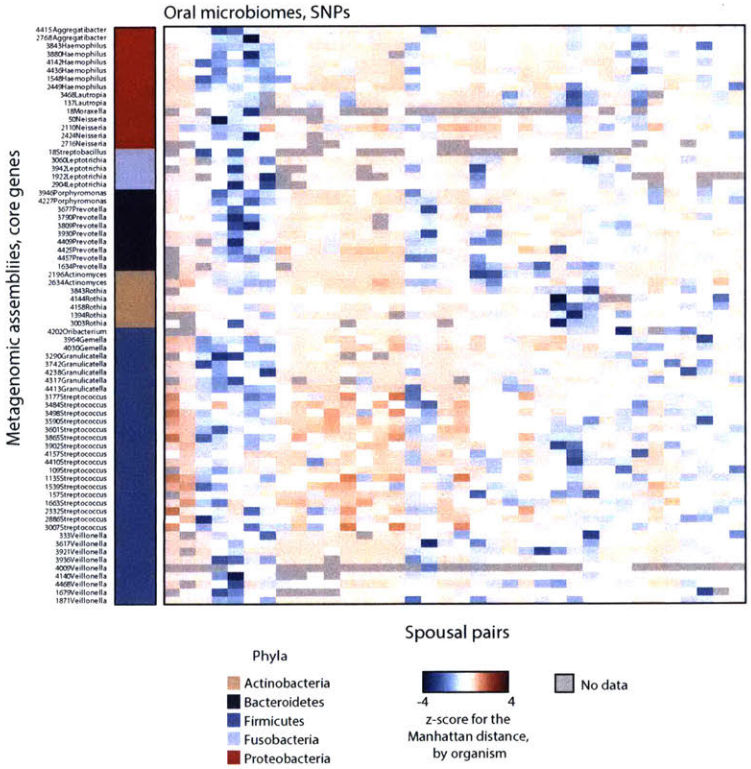
Supplementary Figure 12



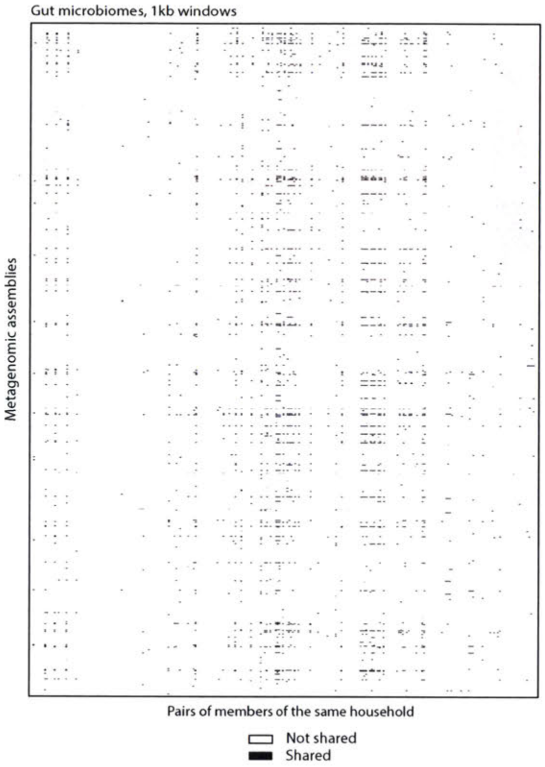
Supplementary Figure 13



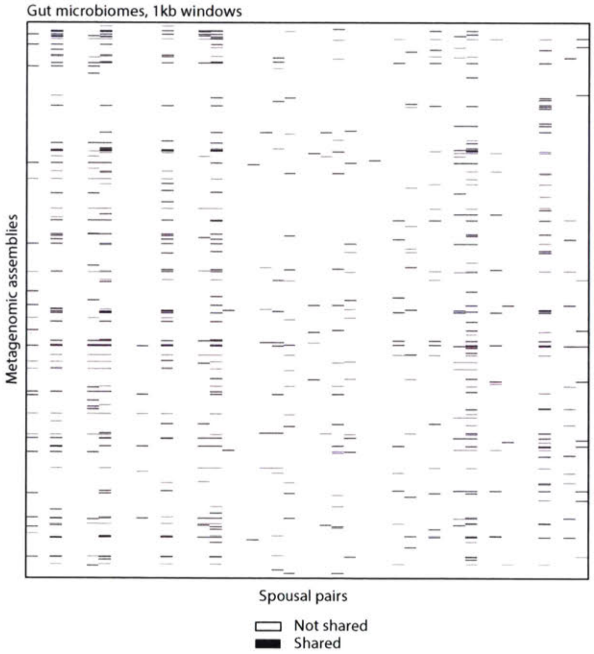
Supplementary Figure 14



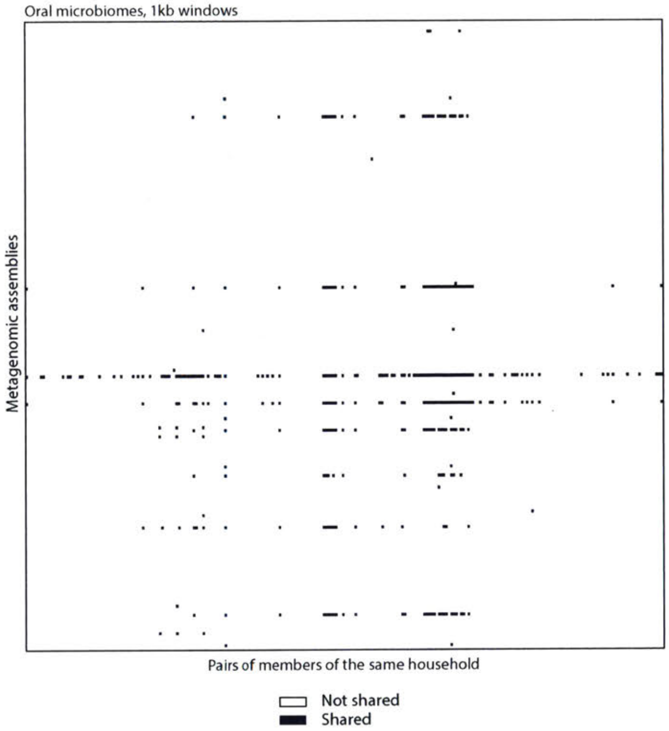
Supplementary Figure 15



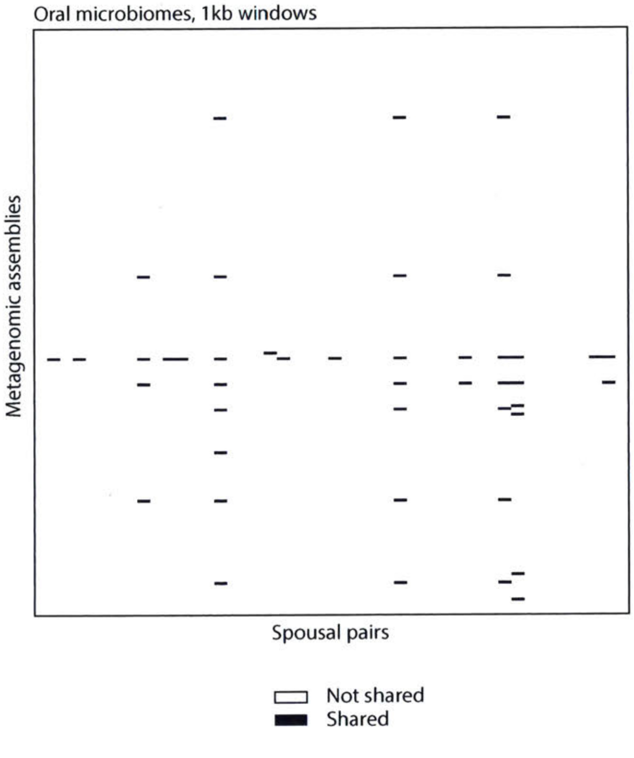
Supplementary Figure 16



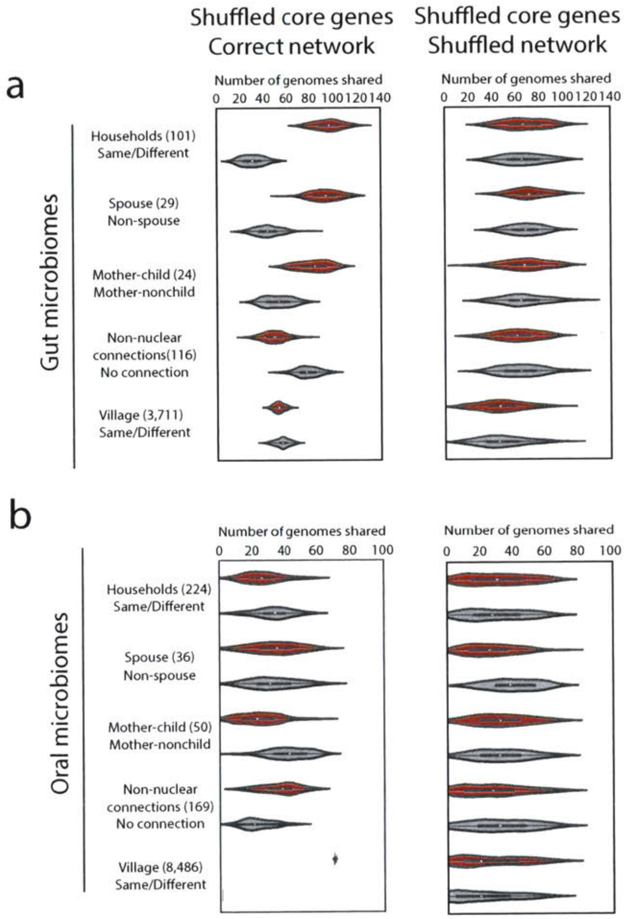
Supplementary Figure 17



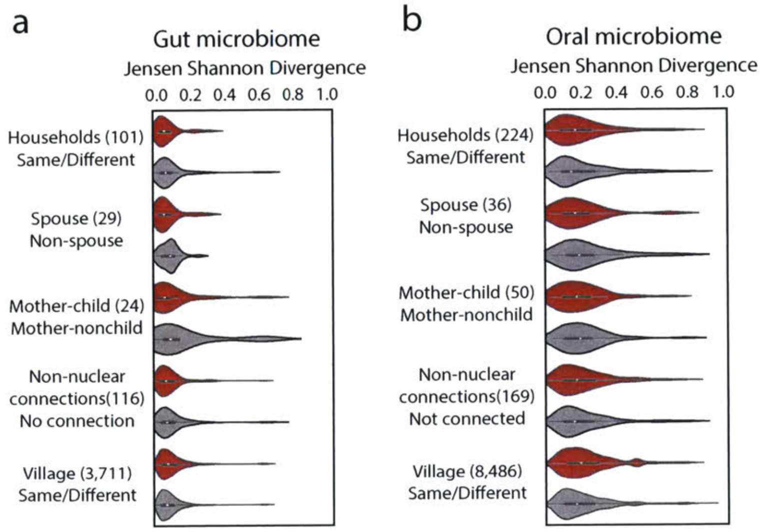
Supplementary Figure 18



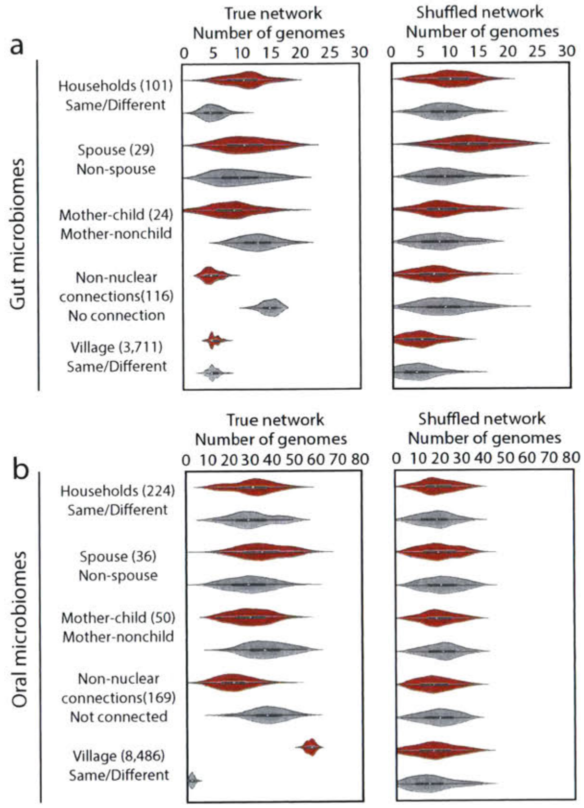
Supplementary Figure 19



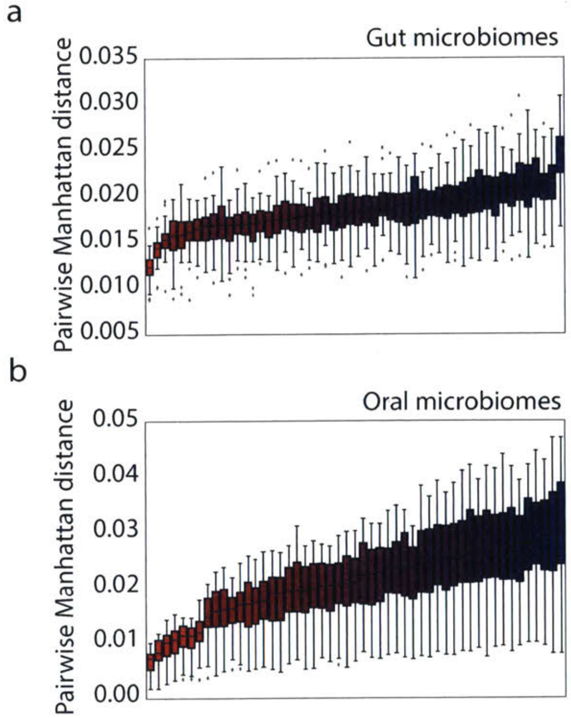
Supplementary Figure 20



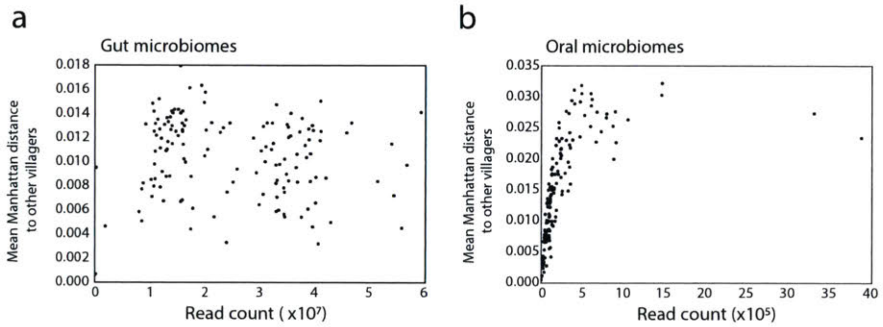
Supplementary Figure 21



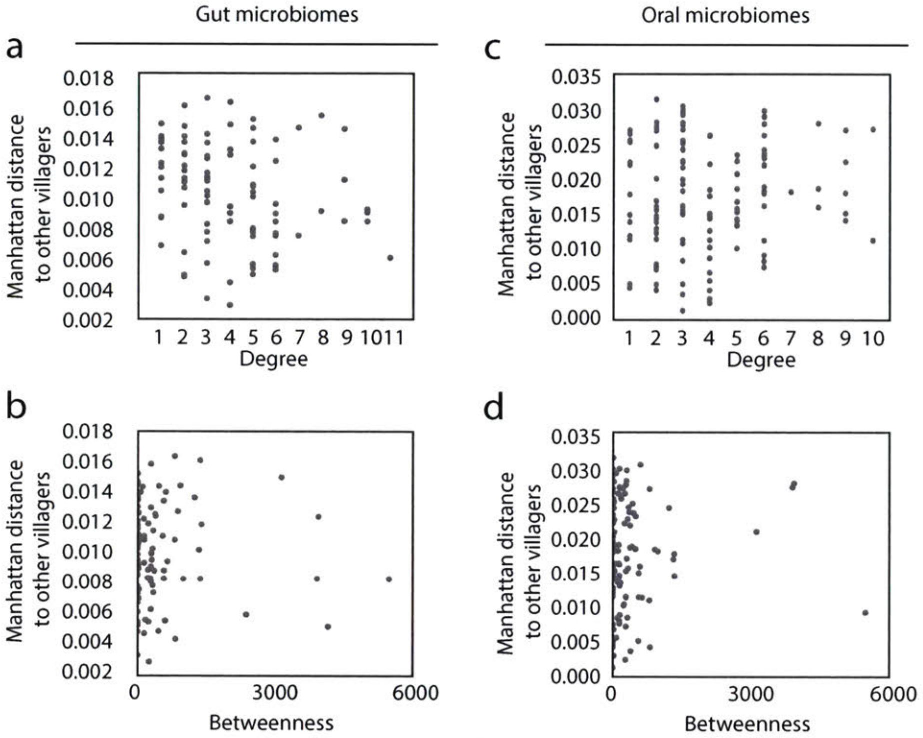
Supplementary Figure 22



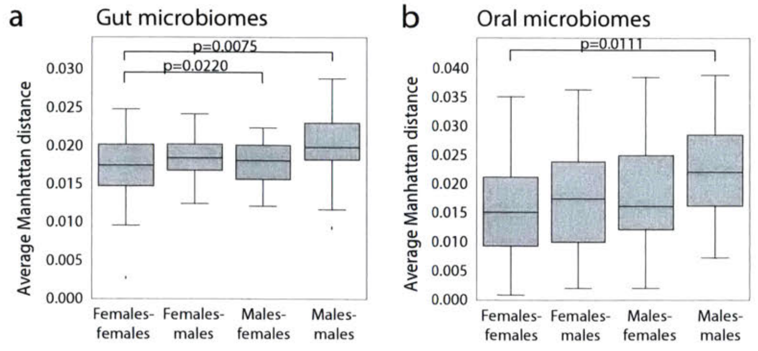
Supplementary Figure 23



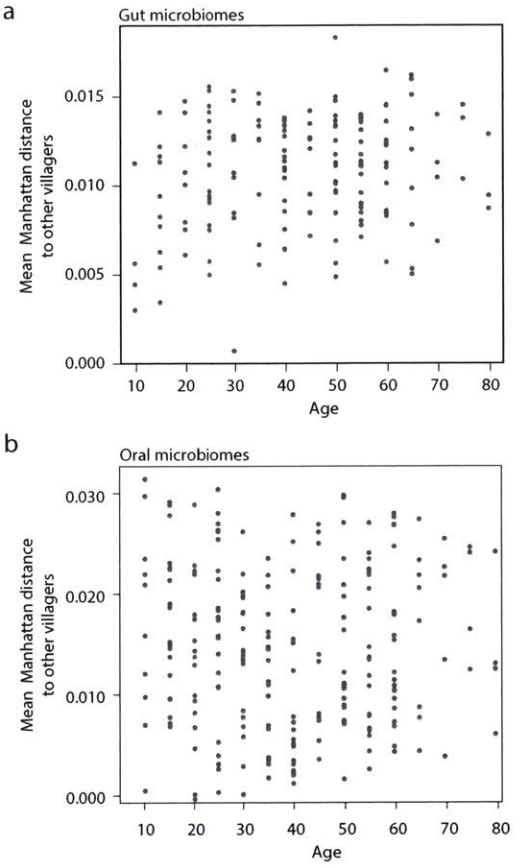
Supplementary Figure 24



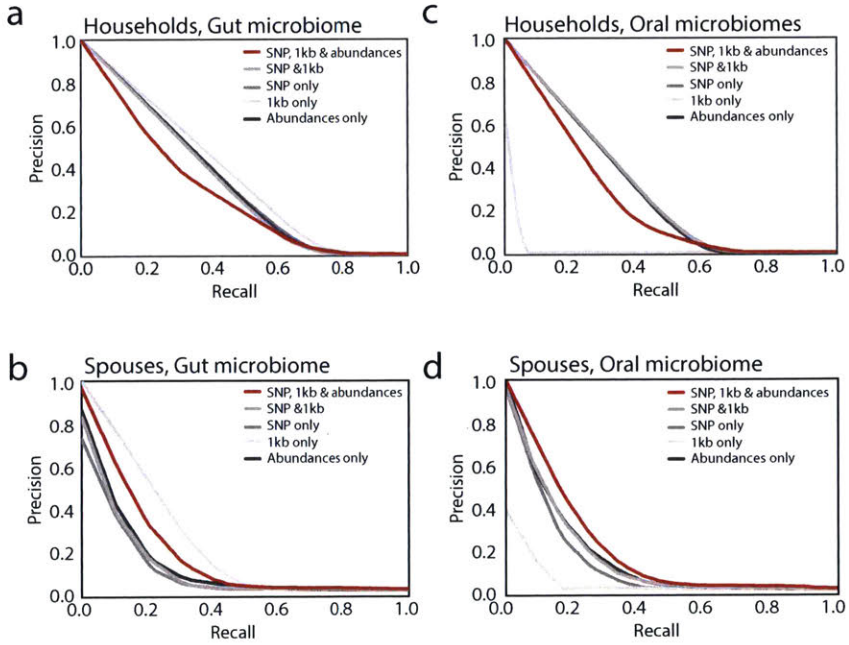
Supplementary Figure 25



Supplementary Figure 26



Supplementary Figure 27



Chapter 5

Discussion

In this thesis, I present the description of rapid within-person adaptation for a bacterial species, *Bacteroides fragilis*, whose native niche is the human intestine, as well as the most time-resolved description of bacterial within-person evolutionary dynamics to date. Within the gut microbiome of individual people, *B. fragilis* acquires adaptive point mutations in key genes, including polysaccharide importers and capsule synthesis genes, under the pressure of natural selection (Chapter 2). Continuing adaptation suggests there is no single optimal *B. fragilis* sequence for survival in the human microbiome and points to competing selective forces. From the TwinUK metagenomes, we also discovered 6 *Bacteroidetes* strains undergoing adaptive evolution during colonization periods up to decades (Chapter 3). It is therefore likely that species in the *Bacteroidetes* phylum share similar within-person adaptive evolutionary patterns.

Nonetheless, additional studies are necessary to show whether adaptive within-person evolution is specific to *B. fragilis* (and *Bacteroidetes* in general) or is a common feature of gut commensals. A recent evolutionary analysis of multiple commensal species using HMP metagenomes provides hope that our results are generalizable (Garud *et al.*, 2017). Their results hint that SNPs detected with low frequency provide a dN/dS value that may be compatible with positive selection. However, conclusions might still be complicated, as another recent investigation into *E. coli* evolution detected only signatures of neutral diversity (Ghalayini *et al.*, 2018). Many factors can contribute to the discrepancy between the conclusions for *E. coli* and our results. For example, *E. coli* has a much smaller population size in the micrbiome compared to those of *Bacteroidetes* strains (Lloyd-Price et al. 2017), and genetic drift may play a major role in *E. coli*'s evolution. To draw a more complete picture of within-person evolution of the commensal microbiome, I am currently applying culture-based evolutionary approaches to more species, including *Parabacteroides distasonis*, *Bifidobacterium longum*, *Bifidobacterium adolescentis* and *Escherichia coli*. I will further explore whether evolutionary dynamics is different between healthy subjects and IBD patients. This study may provide both fundamental insights into the dynamics of human microbiomes and provide a discovery route for understanding genes and pathways important to bacterial survival within the microbiome; many of these species lack efficient genetic tools to characterize gene functions (Lee et al. 2013)(Price et al. 2018). We may also determine how taxonomy, bacterial life strategies, and human health states affect evolutionary dynamics of commensals within micrbiomes.

Two complementary approaches—Culture-based and culture-independent—are introduced in this thesis to study the evolutionary dynamics of gut microbiome. Culture-based population genomics allows precise detection of mutations emerged within-person; culture-independent metagenomic sequencing is broadly applied and has a huge number of public datasets (Almeida et al. 2019; Pasolli et al. 2019). While metagenome alone falls short in revealing high-resolution evolutionary signals, we manage to develop two strategies to utilize metagenomic samples. In the first strategy, we identify SNPs with culture-based genomics and combine these SNPs with metagenomes to track the dynamics of different genotypes over time. Alternatively, we use DonorFinder to track closely related strains from metagenomes and identify strains transmitted in recent history. Strains thus identified have been diverged for the right amount of time for observing insights of within-person evolution. Culture-based approach can be labor-intense and is not yet optimized for many species. Culture-independent metagenomes lack the resolution to differentiate low-frequency mutations, which may constitute the majority of the information for within-person evolution. A future might simply be high-throughput single cell microbiome sequencing techniques. Current single cell techniques lack in throughput and usually have low genome completeness for individual cells (L. Xu et al. 2016; F. B. Yu et al. 2017). To advance single cell technique in the microbiome field, I am collaborating with David Weitz's lab to develop high-throughput experimental and computational pipelines. We endeavor to significantly increase the number of cells sequenced from a single microbiome sample with drop-based microfluidics, obtaining tens of thousands of single cell-level genomes for many bacterial lineages. Although individual single cell genomes likely still have relatively low quality and completeness, the large number of cells may compensate for this weakness and provide the opportunity to identify intra-strain variations via proper inferences.

Should within-person adaptive evolution be a common feature of gut commensals, as it is for many opportunistic pathogens of the cystic fibrosis lung (Smith *et al.*, 2006; Lieberman *et al.*, 2014; Chung *et al.*, 2017), it may have far-reaching implications for the microbiome field. Within-person evolution, in addition to ecological forces, may need to be considered as a possible driver of community dynamics, such as increases or decreases in species abundances over time. In particular, the eco-evolutionary force of monopolization—in which adaptation to a unique local environment enables early colonizers to prevent subsequent invasion by new potential colonizers (De Meester *et al.*, 2016)—may need more attention in the microbiome field. Monopolization may be responsible for the observed stability of individual lineages in the microbiome and the microbiome's ability to provide colonization resistance (Faith *et al.*, 2013; Martínez *et al.*, 2018). Further, pressures specific to individuals or populations may necessitate the need for careful selection or engineering of probiotic strains to maximize the potential for long-term colonization. Future work is needed to understand the importance of within-person evolution to the design of microbial-based therapeutics, as well as its interplay with ecological forces. Our work demonstrates the power of culture-based evolutionary approaches for providing insights into the dynamics of human microbiomes and for discovering genes and pathways critical to bacterial survival within the microbiome.

References

Almeida, Alexandre, Alex L. Mitchell, Miguel Boland, Samuel C. Forster, Gregory B. Gloor, Aleksandra Tarkowska, Trevor D. Lawley, and Robert D. Finn. 2019. "A New Genomic Blueprint of the Human Gut Microbiota." *Nature*. https://doi.org/10.1038/s41586-019-0965-1.

Aquino-Michaels, K., Y. Man Lei, M.-L. Alegre, Jason B Williams, Bana Jabri, Nathaniel Hubert, Thomas F Gajewski, et al. 2015. "Commensal Bifidobacterium Promotes Antitumor Immunity and Facilitates Anti-PD-L1 Efficacy." *Science* 350 (6264): 1084–89. https://doi.org/10.1126/science.aac4255.

Arumugam, Manimozhiyan, Jeroen Raes, Eric Pelletier, Denis Le Paslier, Takuji Yamada, Daniel R Mende, Gabriel R Fernandes, et al. 2011. "Enterotypes of the Human Gut Microbiome." *Nature* 473 (7346): 174–80. https://doi.org/10.1038/nature09944.

Ashkenazy, Haim, Elana Erez, Eric Martz, Tal Pupko, and Nir Ben-Tal. 2010. "ConSurf 2010: Calculating Evolutionary Conservation in Sequence and Structure of Proteins and Nucleic Acids." *Nucleic Acids Research* 38 (SUPPL. 2): 529–33. https://doi.org/10.1093/nar/gkq399.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, et al. 2012. "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing." *Journal of Computational Biology* 19 (5): 455–77. https://doi.org/10.1089/cmb.2012.0021.

Barrick, Jeffrey E., Dong Su Yu, Sung Ho Yoon, Haeyoung Jeong, Tae Kwang Oh, Dominique Schneider, Richard E. Lenski, and Jihyun F. Kim. 2009. "Genome Evolution and Adaptation in a Long-Term Experiment with Escherichia Coli." *Nature* 461 (7268): 1243–47. https://doi.org/10.1038/nature08480.

Barrick, Jeffrey E, and Richard E Lenski. 2013. "Genome Dynamics during Experimental Evolution." *Nature Reviews. Genetics* 14 (12): 827–39. https://doi.org/10.1038/nrg3564.

Baym, Michael, Sergey Kryazhimskiy, Tami D. Lieberman, Hattie Chung, Michael M. Desai, and Roy Kishony. 2015. "Inexpensive Multiplexed Library Preparation for Megabase-Sized Genomes." Edited by Stefan J. Green. *PLOS ONE* 10 (5): e0128036. https://doi.org/10.1371/journal.pone.0128036.

Biek, Roman, Oliver G. Pybus, James O. Lloyd-Smith, and Xavier Didelot. 2015. "Measurably Evolving Pathogens in the Genomic Era." *Trends in Ecology and Evolution* 30 (6): 306–13. https://doi.org/10.1016/j.tree.2015.03.009.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30 (15): 2114–20. https://doi.org/10.1093/bioinformatics/btu170.

Brito, I. L., S. Yilmaz, K. Huang, L. Xu, S. D. Jupiter, A. P. Jenkins, W. Naisilisili, et al. 2016. "Mobile Genes in the Human Microbiome Are Structured from Global to Individual Scales." *Nature* 535 (7612): 435–39. https://doi.org/10.1038/nature18927.

Brito, Ilana L., Thomas Gurry, Shijie Zhao, Katherine Huang, Sarah K. Young, Terrence P. Shea, Waisea Naisilisili, et al. 2019. "Transmission of Human-Associated Microbiota along Family and Social Networks." *Nature Microbiology*. https://doi.org/10.1101/540252.

Brito, Ilana Lauren, and Eric John Alm. 2016. "Tracking Strains in the Microbiome: Insights from Metagenomics and Models." *Frontiers in Microbiology* 7 (May): 1–8. https://doi.org/10.3389/fmicb.2016.00712.

Britton, Robert A., and Vincent B. Young. 2014. "Role of the Intestinal Microbiota in Resistance to Colonization by Clostridium Difficile." *Gastroenterology* 146 (6): 1547–53. https://doi.org/10.1053/j.gastro.2014.01.059.

Browne, Hilary P., Samuel C. Forster, Blessing O. Anonye, Nitin Kumar, B. Anne Neville, Mark D. Stares, David Goulding, and Trevor D. Lawley. 2016. "Culturing of 'Unculturable' Human Microbiota Reveals Novel Taxa and Extensive Sporulation." *Nature* 533 (7604): 543–46. https://doi.org/10.1038/nature17645.

Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2014. "Fast and Sensitive Protein Alignment Using DIAMOND." *Nature Methods* 12 (1): 59–60. https://doi.org/10.1038/nmeth.3176.

Carr, Edward J., James Dooley, Josselyn E. Garcia-Perez, Vasiliki Lagou, James C. Lee, Carine Wouters, Isabelle Meyts, et al. 2016. "The Cellular Composition of the Human Immune System Is Shaped by Age and Cohabitation." *Nature Immunology* 17 (4): 461–68. https://doi.org/10.1038/ni.3371.

Cerdeno-Tarraga, A. M. 2005. "Extensive DNA Inversions in the B. Fragilis Genome Control Variable Gene Expression." *Science* 307 (5714): 1463–65. https://doi.org/10.1126/science.1107008.

Chen, L., Jian Yang, Jun Yu, Zhijian Yao, Lilian Sun, Yan Shen, and Qi Jin. 2004. "VFDB: A Reference Database for Bacterial Virulence Factors." *Nucleic Acids Research* 33 (Database issue): D325–28. https://doi.org/10.1093/nar/gki008.

Chomczynski, Piotr, and Michal Rymaszewski. 2006. "Alkaline Polyethylene Glycol-Based Method for Direct PCR from Bacteria, Eukaryotic Tissue Samples, and Whole Blood." *BioTechniques* 40 (4): 454–58. https://doi.org/10.2144/000112149.

Chu, Nathaniel D, Sean A Clarke, Sonia Timberlake, Martin F Polz, Alan D Grossman, and Eric J Alm. 2017. "A Mobile Element in MutS Drives Hypermutation in a Marine Vibrio." *MBio* 8 (1): e02045-16. https://doi.org/10.1128/mBio.02045-16.

Chung, Hattie, Tami D. Lieberman, Sara O. Vargas, Kelly B. Flett, Alexander J. McAdam, Gregory P. Priebe, and Roy Kishony. 2017. "Global and Local Selection Acting on the Pathogen Stenotrophomonas Maltophilia in the Human Lung." *Nature Communications* 8 (January): 14078.

https://doi.org/10.1038/ncomms14078.

Cleary, Brian, Ilana Lauren Brito, Katherine Huang, Dirk Gevers, Terrance Shea, Sarah Young, and Eric J Alm. 2015. “Detection of Low-Abundance Bacterial Strains in Metagenomic Datasets by Eigengenome Partitioning.” *Nature Biotechnology* 33 (10): 1053–60. https://doi.org/10.1038/nbt.3329.

Coombes, Brian K. 2016. “Bacterial Evolution: Making a Host-Adapted Bacterium.” *Nature Microbiology* 1 (3): 16010. https://doi.org/10.1038/nmicrobiol.2016.10.

Costea, Paul Igor, Robin Munch, Luis Pedro Coelho, Lucas Paoli, Shinichi Sunagawa, and Peer Bork. 2017. “MetaSNV: A Tool for Metagenomic Strain Level Analysis.” *PLoS ONE* 12 (7): 1–9. https://doi.org/10.1371/journal.pone.0182392.

Coyne, Michael J., Kevin G. Roelofs, and Laurie E. Comstock. 2016. “Type VI Secretion Systems of Human Gut Bacteroidales Segregate into Three Genetic Architectures, Two of Which Are Contained on Mobile Genetic Elements.” *BMC Genomics* 17 (1): 58. https://doi.org/10.1186/s12864-016-2377-z.

Coyne, Michael J, Maria Chatzidaki-Livanis, Lawrence C Paoletti, and Laurie E Comstock. 2008. “Role of Glycan Synthesis in Colonization of the Mammalian Gut by the Bacterial Symbiont Bacteroides Fragilis.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (35): 13099–104. https://doi.org/10.1073/pnas.0804220105.

Coyne, Michael J, Naamah Levy Zitomersky, Manson Mcguire, Levy Zitomersky, Ashlee M Earl, and E Comstock. 2014. “Evidence of Extensive DNA Transfer between Bacteroidales Species within the Human Gut.” *MBio* 5 (3): e01305-14. https://doi.org/10.1128/mBio.01305-14.Editor.

Crum-Cianflone, Nancy F., Eva Sullivan, and Gonzalo Ballon-Landa. 2015. “Fecal Microbiota Transplantation and Successful Resolution of Multidrug-Resistant-Organism Colonization.” *Journal of Clinical Microbiology* 53 (6): 1986–89. https://doi.org/10.1128/JCM.00820-15.

Das, Debanu, Robert D. Finn, Dennis Carlton, Mitchell D. Miller, Polat Abdubek, Tamara Astakhova, Herbert L. Axelrod, et al. 2010. “The Structure of BVU2987 from Bacteroides Vulgatus Reveals a Superfamily of Bacterial Periplasmic Proteins with Possible Inhibitory Function.” *Acta Crystallographica Section F: Structural Biology and Crystallization Communications* 66 (10): 1265–73. https://doi.org/10.1107/S1744309109046788.

David, Lawrence A, Arne C Materna, Jonathan Friedman, Maria I Campos-Baptista, Matthew C Blackburn, Allison Perrotta, Susan E Erdman, and Eric J Alm. 2014. “Host Lifestyle Affects Human Microbiota on Daily Timescales.” *Genome Biology* 15 (7): R89. https://doi.org/10.1186/gb-2014-15-7-r89.

Didelot, Xavier, David W Eyre, Madeleine Cule, Camilla LC Ip, M Ansari, David Griffiths, Alison Vaughan, et al. 2012. “Microevolutionary Analysis of Clostridium Difficile Genomes to Investigate

Transmission." *Genome Biology* 13 (12): R118. https://doi.org/10.1186/gb-2012-13-12-r118.

Didelot, Xavier, A. Sarah Walker, Tim E. Peto, Derrick W. Crook, and Daniel J. Wilson. 2016. "Within-Host Evolution of Bacterial Pathogens." *Nature Reviews Microbiology* 14 (3): 150–62. https://doi.org/10.1038/nrmicro.2015.13.

Donaldson, G. P., M. S. Ladinsky, K. B. Yu, J. G. Sanders, B. B. Yoo, W. C. Chou, M. E. Conner, et al. 2018. "Gut Microbiota Utilize Immunoglobulin a for Mucosal Colonization." *Science* 360 (6390): 795–800. https://doi.org/10.1126/science.aaq0926.

Faith, J., J. Guruge, M. Charbonneau, S. Subramanian, H. Seedorf, A. Goodman, J. Clemente, et al. 2013. "The Long-Term Stability of the Human Gut Microbiota." *Science* 341 (6141): 1237439–1237439. https://doi.org/10.1126/science.1237439.

Faith, Jeremiah J, Janaki L Guruge, Mark Charbonneau, Sathish Subramanian, Henning Seedorf, Andrew L Goodman, Jose C Clemente, et al. 2013. "The Long-Term Stability of the Human Gut Microbiota" 340 (July). https://doi.org/10.1126/science.1237439.

Feliziani, Sof??a, Rasmus L. Marvig, Adela M. Luj??n, Alejandro J. Moyano, Julio A. Di Rienzo, Helle Krogh Johansen, S??ren Molin, and Andrea M. Smania. 2014. "Coexistence and Within-Host Evolution of Diversified Lineages of Hypermutable Pseudomonas Aeruginosa in Long-Term Cystic Fibrosis Infections." *PLoS Genetics* 10 (10). https://doi.org/10.1371/journal.pgen.1004651.

Ferretti, Pamela, Edoardo Pasolli, Adrian Tett, Francesco Asnicar, Valentina Gorfer, Sabina Fedi, Federica Armanini, et al. 2018. "Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome." *Cell Host and Microbe* 24 (1): 133–145.e5. https://doi.org/10.1016/j.chom.2018.06.005.

Forster, Samuel C., Nitin Kumar, Blessing O. Anonye, Alexandre Almeida, Elisa Viciani, Mark D. Stares, Matthew Dunn, et al. 2019. "A Human Gut Bacterial Genome and Culture Collection for Improved Metagenomic Analyses." *Nature Biotechnology* 37 (2): 186–92. https://doi.org/10.1038/s41587-018-0009-7.

Franzosa, Eric A., Katherine Huang, James F. Meadow, Dirk Gevers, Katherine P. Lemon, Brendan J. M. Bohannan, and Curtis Huttenhower. 2015. "Identifying Personal Microbiomes Using Metagenomic Codes." *Proceedings of the National Academy of Sciences* 112 (22): E2930–38. https://doi.org/10.1073/pnas.1423854112.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. "CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data." *Bioinformatics* 28 (23): 3150–52. https://doi.org/10.1093/bioinformatics/bts565.

Garud, Nandita R., Benjamin H. Good, Oskar Hallatschek, and Katherine S. Pollard. 2017. "Evolutionary Dynamics of Bacteria in the Gut Microbiome within and across Hosts." *Doi.Org*, 210955. https://doi.org/10.1101/210955.

Ghalayini, Mohamed, Adrien Launay, Antoine Bridier-Nahmias, Olivier Clermont, Erick Denamur, Mathilde Lescat, and Olivier Tenaillon. 2018. "'Evolution of a Dominant Natural Isolate of *Escherichia Coli* in the Human Gut over a Year Suggests a Neutral Evolution with Reduced Effective Population Size.'" *Applied and Environmental Microbiology*, no. January: AEM.02377-17. https://doi.org/10.1128/AEM.02377-17.

Gilbert, Jack A., Robert A. Quinn, Justine Debelius, Zhenjiang Z. Xu, James Morton, Neha Garg, Janet K. Jansson, Pieter C. Dorrestein, and Rob Knight. 2016. "Microbiome-Wide Association Studies Link Dynamic Microbial Consortia to Disease." *Nature* 535 (7610): 94–103. https://doi.org/10.1038/nature18850.

Giraud, Antoine, Antoine Giraud, Ivan Matic, Olivier Tenaillon, Antonio Clara, Miroslav Radman, and Michel Fons. 2001. "Costs and Benefits of High Mutation Rates: Adaptive Evolution of Bacteria in the Mouse Gut." *Science* 291 (5513): 2606–8. https://doi.org/10.1126/science.1056421.

Glenwright, Amy J., Karunakar R. Pothula, Satya P. Bhamidimarri, Dror S. Chorev, Arnaud Baslé, Susan J. Firbank, Hongjun Zheng, et al. 2017. "Structural Basis for Nutrient Acquisition by Dominant Members of the Human Gut Microbiota." *Nature* 541 (7637): 407–11. https://doi.org/10.1038/nature20828.

Golubchik, Tanya, Elizabeth M. Batty, Ruth R. Miller, Helen Farr, Bernadette C. Young, Hanna Larner-Svensson, Rowena Fung, et al. 2013. "Within-Host Evolution of Staphylococcus Aureus during Asymptomatic Carriage." *PLoS ONE* 8 (5): 1–14. https://doi.org/10.1371/journal.pone.0061319.

Good, Benjamin H., Michael J. McDonald, Jeffrey E. Barrick, Richard E. Lenski, and Michael M. Desai. 2017. "The Dynamics of Molecular Evolution over 60,000 Generations." *Nature*, October. https://doi.org/10.1038/nature24287.

Goodrich, Julia K., Jillian L. Waters, Angela C. Poole, Jessica L. Sutter, Omry Koren, Ran Blekhman, Michelle Beaumont, et al. 2014. "Human Genetics Shape the Gut Microbiome." *Cell* 159 (4): 789–99. https://doi.org/10.1016/j.cell.2014.09.053.

Grinspan, Ari M., and Colleen R. Kelly. 2015. "Fecal Microbiota Transplantation for Ulcerative Colitis: Not Just Yet." *Gastroenterology* 149 (1): 15–18. https://doi.org/10.1053/j.gastro.2015.05.030.

Groussin, Mathieu, Florent Mazel, Jon G. Sanders, Chris S. Smillie, Sébastien Lavergne, Wilfried Thuiller, and Eric J. Alm. 2017. "Unraveling the Processes Shaping Mammalian Gut Microbiomes over Evolutionary Time." *Nature Communications* 8 (February): 14319. https://doi.org/10.1038/ncomms14319.

Hampton-Marcell, Jarrad T., Jose V. Lopez, and Jack A. Gilbert. 2017. "The Human Microbiome: An Emerging Tool in Forensics." *Microbial Biotechnology* 10 (2): 228–30. https://doi.org/10.1111/1751-7915.12699.

He, M., M. Sebaihia, T. D. Lawley, R. A. Stabler, L. F. Dawson, M. J. Martin, K. E. Holt, et al. 2010.

"Evolutionary Dynamics of Clostridium Difficile over Short and Long Time Scales." *Proceedings of the National Academy of Sciences* 107 (16): 7527–32. https://doi.org/10.1073/pnas.0914322107.

Jolivet-Gougeon, A., B. Kovacs, S. Le Gall-David, H. Le Bars, L. Bousarghin, M. Bonnaure-Mallet, B. Lobel, F. Guille, C.-J. Soussy, and P. Tenke. 2011. "Bacterial Hypermutation: Clinical Implications." *Journal of Medical Microbiology* 60 (5): 563–73. https://doi.org/10.1099/jmm.0.024083-0.

Joshi, N A, and J N Fass. 2011. "Sickle: A Sliding-Window, Adaptive, Quality-Based Trimming Tool for FastQ Files (Version 1.33)[Software]."

Kauffman, Kathryn M., and Martin F. Polz. 2018. "Streamlining Standard Bacteriophage Methods for Higher Throughput." *MethodsX* 5: 159–72. https://doi.org/10.1016/j.mex.2018.01.007.

Klemm, Elizabeth J., Effrossyni Gkrania-Klotsas, James Hadfield, Jessica L. Forbester, Simon R. Harris, Christine Hale, Jennifer N. Heath, et al. 2016. "Emergence of Host-Adapted Salmonella Enteritidis through Rapid Evolution in an Immunocompromised Host." *Nature Microbiology* 1 (3): 15023. https://doi.org/10.1038/nmicrobiol.2015.23.

Korem, T., D. Zeevi, J. Suez, A. Weinberger, T. Avnit-Sagi, M. Pompan-Lotan, E. Matot, et al. 2015. "Growth Dynamics of Gut Microbiota in Health and Disease Inferred from Single Metagenomic Samples." *Science* 349 (6252): 1101–6. https://doi.org/10.1126/science.aac4812.

Krissinel, Evgeny, and Kim Henrick. 2007. "Inference of Macromolecular Assemblies from Crystalline State." *Journal of Molecular Biology* 372 (3): 774–97. https://doi.org/10.1016/j.jmb.2007.05.022.

Kuwahara, Tomomi, Atsushi Yamashita, Hideki Hirakawa, Haruyuki Nakayama, Hidehiro Toh, Natsumi Okada, Satoru Kuhara, Masahira Hattori, Tetsuya Hayashi, and Yoshinari Ohnishi. 2004. "Genomic Analysis of Bacteroides Fragilis Reveals Extensive DNA Inversions Regulating Cell Surface Adaptation." *Proceedings of the National Academy of Sciences* 101 (41): 14919–24. https://doi.org/10.1073/pnas.0404172101.

Langmead, Ben, and Steven L Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

Lee, S. Melanie, Gregory P. Donaldson, Zbigniew Mikulski, Silva Boyajian, Klaus Ley, and Sarkis K. Mazmanian. 2013. "Bacterial Colonization Factors Control Specificity and Stability of the Gut Microbiota." *Nature* 501 (7467): 426–29. https://doi.org/10.1038/nature12447.

Ley, Ruth E, Catherine A Lozupone, Micah Hamady, Rob Knight, and Jeffrey I Gordon. 2008. "Worlds within Worlds: Evolution of the Vertebrate Gut Microbiota." *Nature Reviews Microbiology* 6 (10): 776–88. https://doi.org/10.1038/nrmicro1978.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.

Li, Heng. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM" 00 (00): 1–3. http://arxiv.org/abs/1303.3997.

Li, Weina, Yu Deng, Qian Chu, and Peng Zhang. 2019. "Gut Microbiome and Cancer Immunotherapy." *Cancer Letters* 447 (December 2018): 41–47. https://doi.org/10.1016/j.canlet.2019.01.015.

Lieberman, Tami D, Kelly B Flett, Idan Yelin, Thomas R Martin, Alexander J Mcadam, Gregory P Priebe, and Roy Kishony. 2014. "Genetic Variation of a Bacterial Pathogen within Individuals with Cystic Fibrosis Provides a Record of Selective Pressures." *Nat Genet* 46 (1): 82–87. https://doi.org/10.1038/ng.2848.

Lieberman, Tami D, Jean-Baptiste Michel, Mythili Aingaran, Gail Potter-Bynoe, Damien Roux, Michael R Davis, David Skurnik, et al. 2011. "Parallel Bacterial Evolution within Multiple Patients Identifies Candidate Pathogenicity Genes." *Nature Genetics* 43 (12): 1275–80. https://doi.org/10.1038/ng.997.

Lloyd-Price, Jason, Anup Mahurkar, Gholamali Rahnavard, Jonathan Crabtree, Joshua Orvis, A. Brantley Hall, Arthur Brady, et al. 2017. "Strains, Functions and Dynamics in the Expanded Human Microbiome Project." *Nature* 550 (7674): 61–66. https://doi.org/10.1038/nature23889.

Luo, Chengwei, Rob Knight, Heli Siljander, Mikael Knip, Ramnik J Xavier, and Dirk Gevers. 2015. "ConStrains Identifies Microbial Strains in Metagenomic Datasets." *Nature Biotechnology* 33 (10): 0–4. https://doi.org/10.1038/nbt.3319.

Marcobal, Angela, Mariana Barboza, Erica D. Sonnenburg, Nicholas Pudlo, Eric C. Martens, Prerak Desai, Carlito B. Lebrilla, et al. 2011. "Bacteroides in the Infant Gut Consume Milk Oligosaccharides via Mucus-Utilization Pathways." *Cell Host and Microbe* 10 (5): 507–14. https://doi.org/10.1016/j.chom.2011.10.007.

Martens, Eric C., Nicole M. Koropatkin, Thomas J. Smith, and Jeffrey I. Gordon. 2009. "Complex Glycan Catabolism by the Human Gut Microbiota: The Bacteroidetes Sus-like Paradigm." *Journal of Biological Chemistry* 284 (37): 24673–77. https://doi.org/10.1074/jbc.R109.022848.

Martin, Marcel. 2011. "Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads." *EMBnet J* 17 (May): 10–12. http://journal.embnet.org/index.php/embnetjournal/article/view/200.

Martínez, Inés, Maria X. Maldonado-Gomez, João Carlos Gomes-Neto, Hatem Kittana, Hua Ding, Robert Schmaltz, Payal Joglekar, et al. 2018. "Experimental Evaluation of the Importance of Colonization History in Early-Life Gut Microbiota Assembly." *ELife* 7: 1–26. https://doi.org/10.7554/eLife.36521.

Marvig, Rasmus Lykke, Helle Krogh Johansen, Søren Molin, and Lars Jelsbak. 2013. "Genome Analysis of a Transmissible Lineage of Pseudomonas Aeruginosa Reveals Pathoadaptive Mutations and Distinct Evolutionary Paths of Hypermutators." *PLoS Genetics* 9 (9). https://doi.org/10.1371/journal.pgen.1003741.

McWilliam, Hamish, Weizhong Li, Mahmut Uludag, Silvano Squizzato, Young Mi Park, Nicola Buso, Andrew Peter Cowley, and Rodrigo Lopez. 2013. "Analysis Tool Web Services from the EMBL-EBI." *Nucleic Acids Research* 41 (W1): W597–600. https://doi.org/10.1093/nar/gkt376.

Meester, Luc De, Joost Vanoverbeke, Laurens J. Kilsdonk, and Mark C. Urban. 2016. "Evolving Perspectives on Monopolization and Priority Effects." *Trends in Ecology and Evolution* 31 (2): 136–46. https://doi.org/10.1016/j.tree.2015.12.009.

Merino, Susana, and Juán M Tomás. 2015. "Bacterial Capsules and Evasion of Immune Responses." *ELS*, no. September: 1–10. https://doi.org/10.1002/9780470015902.a0000957.pub4.

Methé, Barbara A., Karen E. Nelson, Mihai Pop, Heather H. Creasy, Michelle G. Giglio, Curtis Huttenhower, Dirk Gevers, et al. 2012. "A Framework for Human Microbiome Research." *Nature* 486 (7402): 215–21. https://doi.org/10.1038/nature11209.

Moeller, Andrew H., Steffen Foerster, Michael L. Wilson, Anne E. Pusey, Beatrice H. Hahn, and Howard Ochman. 2016. "Social Behavior Shapes the Chimpanzee Pan-Microbiome." *Science Advances* 2 (1): e1500997. https://doi.org/10.1126/sciadv.1500997.

Moeller, Andrew H, Alejandro Caro-Quintero, Deus Mjungu, Alexander V Georgiev, Elizabeth V Lonsdorf, Martin N Muller, Anne E Pusey, Martine Peeters, Beatrice H Hahn, and Howard Ochman. 2016. "Cospeciation of Gut Microbiota with Hominids." *Science (New York, N.Y.)* 353 (6297): 380–82. https://doi.org/10.1126/science.aaf3951.

Moran, P.A.P. 1957. "Random Processes in Genetics." *Mathematical Proceedings of the Cambridge Philosophical Society*, no. April 1957: 60–71.

Mosites, Emily, Matt Sammons, Elkanah Otiang, Alexander Eng, Cecilia Noecker, Ohad Manor, Sarah Hilton, et al. 2017. "Microbiome Sharing between Children, Livestock and Household Surfaces in Western Kenya." *PLoS ONE* 12 (2): 1–15. https://doi.org/10.1371/journal.pone.0171017.

Mwangi, M M, S W Wu, Y Zhou, K Sieradzki, H de Lencastre, P Richardson, D Bruce, et al. 2007. "Tracking the in Vivo Evolution of Multidrug Resistance in Staphylococcus Aureus by Whole-Genome Sequencing." *Proc.Natl.Acad.Sci.U.S.A* 104 (0027–8424 (Print)): 9451–56. https://doi.org/10.1073/pnas.0609839104.

Nayfach, Stephen, and Katherine S Pollard. 2015. "Average Genome Size Estimation Improves Comparative Metagenomics and Sheds Light on the Functional Ecology of the Human Microbiome." *Genome Biology* 16 (1): 51. https://doi.org/10.1186/s13059-015-0611-7.

Nemergut, D. R., S. K. Schmidt, T. Fukami, S. P. O'Neill, T. M. Bilinski, L. F. Stanish, J. E. Knelman, et al. 2013. "Patterns and Processes of Microbial Community Assembly." *Microbiology and Molecular Biology Reviews* 77 (3): 342–56. https://doi.org/10.1128/MMBR.00051-12.

Pasolli, Edoardo, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, et al. 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.

Petrof, Elaine O., and Alexander Khoruts. 2014. "From Stool Transplants to Next-Generation Microbiota Therapeutics." *Gastroenterology* 146 (6): 1573–82. https://doi.org/10.1053/j.gastro.2014.01.004.

Plotree, DOTREE, and DOTGRAM Plotgram. 1989. "PHYLIP-Phylogeny Inference Package (Version 3.2)." *Cladistics* 5 (163): 6.

Plucain, Jessica, Thomas Hindré, Mickaël Le Gac, Olivier Tenaillon, Stéphane Cruveiller, Claudine Médigue, Nicholas Leiby, William R Harcombe, Christopher J Marx, and Richard E Lenski. 2014. "Epistasis and Allele Specificity in the Emergence of a Stable Polymorphism in Escherichia Coli." *Science*, 1242862.

Price, Morgan N., and Adam P. Arkin. 2017. "PaperBLAST: Text Mining Papers for Information about Homologs." Edited by Morgan G. I. Langille. *MSystems* 2 (4): e00039-17. https://doi.org/10.1128/mSystems.00039-17.

Price, Morgan N., Kelly M. Wetmore, R. Jordan Waters, Mark Callaghan, Jayashree Ray, Hualan Liu, Jennifer V. Kuehl, et al. 2018. "Mutant Phenotypes for Thousands of Bacterial Genes of Unknown Function." *Nature* 557 (7706): 503–9. https://doi.org/10.1038/s41586-018-0124-0.

Qin, Junjie, Yingrui Li, Zhiming Cai, Shenghui Songgang Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, et al. 2012. "A Metagenome-Wide Association Study of Gut Microbiota in Type 2 Diabetes." *Nature* 490 (7418): 55–60. https://doi.org/10.1038/nature11450.

Qin, Nan, Fengling Yang, Ang Li, Edi Prifti, Yuanting Yanfei Chen, L. Shao, Jing Guo, et al. 2014. "Alterations of the Human Gut Microbiome in Liver Cirrhosis." *Nature* 513 (7516): 59–64. https://doi.org/10.1038/nature13568.

Rocabert, Charles, Carole Knibbe, Jessika Consuegra, Dominique Schneider, and Guillaume Beslon. 2017. "Beware Batch Culture: Seasonality and Niche Construction Predicted to Favor Bacterial Adaptive Diversification." Edited by Mark M. Tanaka. *PLOS Computational Biology* 13 (3): e1005459. https://doi.org/10.1371/journal.pcbi.1005459.

Rosshart, Stephan P., Brian G. Vassallo, Davide Angeletti, Diane S. Hutchinson, Andrew P. Morgan, Kazuyo Takeda, Heather D. Hickman, et al. 2017. "Wild Mouse Gut Microbiota Promotes Host Fitness and Improves Disease Resistance." *Cell* 171 (5): 1015–1028.e13. https://doi.org/10.1016/j.cell.2017.09.016.

Rothschild, Daphna, Omer Weissbrod, Elad Barkan, Alexander Kurilshikov, Tal Korem, David Zeevi, Paul I. Costea, et al. 2018. "Environment Dominates over Host Genetics in Shaping Human Gut Microbiota." *Nature* 555 (7695): 210–15. https://doi.org/10.1038/nature25973.

Routy, B., Le Chatelier, E., Derosa, L., Duong, C., et al., Le Chatelier E., Derosa L., Duong C.P.M., Alou M.T., Daillère R., Fluckiger A., et al. 2018. "Gut Microbiome Influences Efficacy of PD-1-Based

Immunotherapy against Epithelial Tumors." *Science* 359 (6371): 91–97. https://doi.org/10.1126/science.aan3706.

S., Nayfach, Rodriguez-Mueller B., Garud N., and Pollard K.S. 2016. "An Integrated Metagenomics Pipeline for Strain Profiling Reveals Novel Patterns of Bacterial Transmission and Biogeography." *Genome Research* 26 (11): 1612–25. https://doi.org/10.1101/gr.201863.115.

Schloissnig, Siegfried, Manimozhiyan Arumugam, Shinichi Sunagawa, Makedonka Mitreva, Julien Tap, Ana Zhu, Alison Waller, et al. 2013. "Genomic Variation Landscape of the Human Gut Microbiome." *Nature* 493 (7430): 45–50. https://doi.org/10.1038/nature11711.

Schloss, Patrick D., Kathryn D. Iverson, Joseph F. Petrosino, and Sarah J. Schloss. 2014. "The Dynamics of a Family's Gut Microbiota Reveal Variations on a Theme." *Microbiome* 2 (1): 1–13. https://doi.org/10.1186/2049-2618-2-25.

Scholz, Matthias, Doyle V. Ward, Edoardo Pasolli, Thomas Tolio, Moreno Zolfo, Francesco Asnicar, Duy Tin Truong, Adrian Tett, Ardythe L. Morrow, and Nicola Segata. 2016. "Strain-Level Microbial Epidemiology and Population Genomics from Shotgun Metagenomics." *Nature Methods* 13 (5): 435–38. https://doi.org/10.1038/nmeth.3802.

Schrödinger, LLC. 2015. "The {PyMOL} Molecular Graphics System, Version~1.8."

Scott, N. M., J. Hampton-Marcell, K. M. Handley, S. Lax, M. K. Gibson, N. A. Hasan, W. Van Treuren, et al. 2014. "Longitudinal Analysis of Microbial Interaction between Humans and the Indoor Environment." *Science* 345 (6200): 1048–52. https://doi.org/10.1126/science.1254529.

Seemann, Torsten. 2014. "Prokka: Rapid Prokaryotic Genome Annotation." *Bioinformatics* 30 (14): 2068–69. https://doi.org/10.1093/bioinformatics/btu153.

Segata, Nicola, Katri Korpela, Peer Bork, Luis Pedro Coelho, Stefanie Kandels-Lewis, Paul Costea, Gonneke Willemsen, and Dorret I. Boomsma. 2018. "Selective Maternal Seeding and Environment Shape the Human Gut Microbiome." *Genome Research*, 561–68. https://doi.org/10.1101/gr.233940.117.

Sender, Ron, Shai Fuchs, and Ron Milo. 2016. "Revised Estimates for the Number of Human and Bacteria Cells in the Body." *PLoS Biology* 14 (8): 1–14. https://doi.org/10.1371/journal.pbio.1002533.

Smillie, Christopher S., Jenny Sauk, Dirk Gevers, Jonathan Friedman, Jaeyun Sung, Ilan Youngster, Elizabeth L. Hohmann, et al. 2018. "Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation." *Cell Host and Microbe* 23 (2): 229–240.e5. https://doi.org/10.1016/j.chom.2018.01.003.

Smith, Eric E, Danielle G Buckley, Zaining Wu, Channakhone Saenphimmachak, Lucas R Hoffman, David A D'Argenio, Samuel I Miller, et al. 2006. "Genetic Adaptation by Pseudomonas Aeruginosa to the Airways of Cystic Fibrosis Patients." *Proceedings of the National Academy of Sciences of the*

United States of America 103 (22): 8487–92. https://doi.org/10.1073/pnas.0602138103.

Sniegowski, Paul D., Philip J. Gerrish, and Richard E. Lenski. 1997. "Evolution of High Mutation Rates in Experimental Populations of E. Coli." *Nature* 387 (6634): 703–5. https://doi.org/10.1038/42701.

Snitkin, Evan S, Adrian M Zelazny, Jyoti Gupta, Nisc Comparative, Sequencing Program, Tara N Palmore, Patrick R Murray, and Julia A Segre. 2013. "Genomic Insights into the Fate of Colistin Resistance and Acinetobacter Baumannii during Patient Treatment." *Genome Research* 23: 1155–62. https://doi.org/10.1101/gr.154328.112.Park.

Stummeyer, Katharina, David Schwarzer, Heike Claus, Ulrich Vogel, Rita Gerardy-Schahn, and Martina Mühlenhoff. 2006. "Evolution of Bacteriophages Infecting Encapsulated Bacteria: Lessons from Escherichia Coli K1-Specific Phages." *Molecular Microbiology* 60 (5): 1123–35. https://doi.org/10.1111/j.1365-2958.2006.05173.x.

Truong, Duy Tin, Eric A Franzosa, Timothy L Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. 2015. "MetaPhlAn2 for Enhanced Metagenomic Taxonomic Profiling." *Nature Methods* 13 (1): 101–101. https://doi.org/10.1038/nmeth0116-101b.

Truong, Duy Tin, Adrian Tett, Edoardo Pasolli, Curtis Huttenhower, and Nicola Segata. 2017. "Microbial Strain-Level Population Structure & Genetic Diversity from Metagenomes." *Genome Research* 27 (4): 626–38. https://doi.org/10.1101/gr.216242.116.

Turnbaugh, P.J., R.E. Ley, M. Hamady, C.M. Fraser-Liggett, R. Knight, and J.I. Gordon. 2007. "Feature The Human Microbiome Project." *Nature* 449 (October): 804–10. https://doi.org/10.1038/nature06244.

Turnbaugh, Peter J, Vanessa K Ridaura, Jeremiah J Faith, Federico E Rey, Rob Knight, and Jeffrey I Gordon. 2009. "The Effect of Diet on the Human Gut Microbiome: A Metagenomic Analysis in Humanized Gnotobiotic Mice." *Science Translational Medicine* 1 (6): 6ra14. https://doi.org/10.1126/scitranslmed.3000322.

Verster, Adrian J., Benjamin D. Ross, Matthew C. Radey, Yiqiao Bao, Andrew L. Goodman, Joseph D. Mougous, and Elhanan Borenstein. 2017. "The Landscape of Type VI Secretion across Human Gut Microbiomes Reveals Its Role in Community Composition." *Cell Host & Microbe* 22 (3): 411–419.e4. https://doi.org/10.1016/j.chom.2017.08.010.

Walter, Jens, and Ruth Ley. 2011. "The Human Gut Microbiome: Ecology and Recent Evolutionary Changes." *Annual Review of Microbiology* 65 (1): 411–29. https://doi.org/10.1146/annurev-micro-090110-102830.

Wichman, H A, M R Badgett, L A Scott, and C M Boulianne. 2012. "Different Trajectories of Parallel Evolution During Viral Adaptation" 285 (5426): 422–24.

Wiser, M J, N Ribeck, R E Lenski, J S Littell, C J Muller, K a Dunne, a V Vecchia, et al. 2013. "Asexual

Populations" 342 (December): 1364–67.

Wlodarska, Marta, Aleksandar D. Kostic, and Ramnik J. Xavier. 2015. "An Integrative View of Microbiome-Host Interactions in Inflammatory Bowel Diseases." *Cell Host and Microbe* 17 (5): 577–91. https://doi.org/10.1016/j.chom.2015.04.008.

Woods, Robert, Dominique Schneider, Cynthia L Winkworth, Margaret a Riley, and Richard E Lenski. 2006. "Tests of Parallel Molecular Evolution in a Long-Term Experiment with Escherichia Coli." *Proceedings of the National Academy of Sciences of the United States of America* 103 (24): 9107–12. https://doi.org/10.1073/pnas.0602917103.

Wu, Martin, and Alexandra J. Scott. 2012. "Phylogenomic Analysis of Bacterial and Archaeal Sequences with AMPHORA2." *Bioinformatics* 28 (7): 1033–34. https://doi.org/10.1093/bioinformatics/bts079.

Xie, Hailiang, Ruijin Guo, Huanzi Zhong, Qiang Feng, Zhou Lan, Bingcai Qin, Kirsten J. Ward, et al. 2016. "Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut Microbiome." *Cell Systems* 3 (6): 572–584.e3. https://doi.org/10.1016/j.cels.2016.10.004.

Xu, Jian, Michael A. Mahowald, Ruth E. Ley, Catherine A. Lozupone, Micah Hamady, Eric C. Martens, Bernard Henrissat, et al. 2007. "Evolution of Symbiotic Bacteria in the Distal Human Intestine." *PLoS Biology* 5 (7): 1574–86. https://doi.org/10.1371/journal.pbio.0050156.

Xu, Liyi, Ilana L. Brito, Eric J. Alm, and Paul C. Blainey. 2016. "Virtual Microfluidics for Digital Quantification and Single-Cell Sequencing." *Nature Methods* 13 (9): 759–62. https://doi.org/10.1038/nmeth.3955.

Yatsunenko, T, F E Rey, M J Manary, I Trehan, M G Dominguez-Bello, M Contreras, M Magris, et al. 2012. "Human Gut Microbiome Viewed across Age and Geography." *Nature* 486 (7402): 222–27. https://doi.org/10.1038/nature11053.

Yu, Chin-Sheng, Chih-Wen Cheng, Wen-Chi Su, Kuei-Chung Chang, Shao-Wei Huang, Jenn-Kang Hwang, and Chih-Hao Lu. 2014. "CELLO2GO: A Web Server for Protein SubCELlular LOcalization Prediction with Functional Gene Ontology Annotation." *PLoS ONE* 9 (6): e99368. https://doi.org/10.1371/journal.pone.0099368.

Yu, Feiqiao Brian, Paul C Blainey, Frederik Schulz, Tanja Woyke, Mark A Horowitz, and Stephen R Quake. 2017. "Microfluidic-Based Mini-Metagenomics Enables Discovery of Novel Microbial Lineages from Complex Environmental Samples." *ELife* 6: 1–20. https://doi.org/10.7554/elife.26580.

Zhao, Shijie, Tami D Lieberman, Mathilde Poyet, Mathieu Groussin, Ramnik J Xavier, Eric J Alm, Shijie Zhao, et al. 2019. "Adaptive Evolution within Gut Microbiomes of Article Adaptive Evolution within Gut Microbiomes of Healthy People." *Cell Host and Microbe*, 1–12. https://doi.org/10.1016/j.chom.2019.03.007.

Zhu, Ana, Shinichi Sunagawa, Daniel R. Mende, and Peer Bork. 2015. "Inter-Individual Differences in the Gene Content of Human Gut Bacterial Species." *Genome Biology* 16 (1): 1–13. https://doi.org/10.1186/s13059-015-0646-9.

Zou, Yuanqiang, Wenbin Xue, Guangwen Luo, Ziqing Deng, Panpan Qin, Ruijin Guo, Haipeng Sun, et al. 2019. "1,520 Reference Genomes From Cultivated Human Gut Bacteria Enable Functional Microbiome Analyses." *Nature Biotechnology* 37 (2): 179–85. https://doi.org/10.1038/s41587-018-0008-8.