# Adaptive evolution within the gut microbiome of individual people

Shijie Zhao[1,2,3*], Tami D. Lieberman[1,3,4*,+,], Mathilde Poyet[1,3,4], Sean M. Gibbons[1,3,4], Mathieu Groussin[1,3], Ramnik J. Xavier[3,4,5], Eric J. Alm[1,3,4,+]

1. Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
2. Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
3. Center for Microbiome Informatics and Therapeutics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
4. Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
5. Gastrointestinal Unit and Center for Computational and Integrative Biology, Massachusetts General Hospital, Boston, Massachusetts, USA

* These authors contributed equally to this work.
+ Correspondence should be addressed to: E.J.A. (ejalm@mit.edu) or T.D.L. (tami@mit.edu)

## Abstract

Individual bacterial lineages stably persist for years in the human gut microbiome[1–3]. However, it is unknown if these lineages adapt during colonization of healthy people[2]. Here, we assess evolution within individual microbiomes by sequencing the genomes of 602 *Bacteroides fragilis* isolates cultured from 12 healthy subjects. We find that *B. fragilis* within-subject populations contain significant *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 seventeen genes. These seventeen 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[4], indicating that key

genes have not already been optimized for survival *in vivo*. This surprising 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. We conclude that *B. fragilis* adapts rapidly within the microbiomes of individual healthy people, with implications for microbiome stability and manipulation.

## Main Text

Billions of *de novo* mutations are generated daily within each person's gut microbiome[5–8] (**Table 1**). It is unknown if any of these mutations confer a significant adaptive benefit to the bacteria in which they emerge or, in contrast, all available mutations are deleterious or neutral. The latter possibility is supported by signals of long-term purifying selection in the microbiome[2,9]. These signals raise the possibility that millions of years of evolution within mammalian digestive systems[10,11] has exhausted all beneficial mutations. Yet, previous studies examined evolution at time scales much longer than a human lifespan. Therefore, it is possible that new mutations may still drive rapid adaptation within individual people.

Should adaptive mutations arise and be detectable within individual people, they are likely to indicate genes and pathways critical for long-term bacterial persistence in the human body[12,13,14]. The selective forces on these pathways might be common or person-specific, and their identification could guide microbiome-targeted therapies, including the selection and engineering of therapeutic bacteria for long-term colonization. To date, within-person evolution of the gut

microbiome has not been characterized, as it is difficult to distinguish *de novo* mutations from variants in homologous regions shared by co-colonizing bacteria using metagenomics alone[2]. Culture-based approaches, which enable single-cell level whole-genome comparisons, have been limited to a small number of isolates[1]. Further, it is often implicitly assumed that tracking within-person evolution requires sampling the same individual over many years. However, if bacteria diversify as they evolve, co-existing genotypes enable the inference of within-person evolution without long time-series[15].

To assess the degree to which gut commensals evolve and diversify during colonization, we used a culture-dependent approach and focused on *Bacteroides fragilis*, a prevalent and abundant commensal in the large intestine of healthy people[16]. We surveyed intra-species diversity within 12 healthy subjects (ages 22-37; **Supplementary Table 1**), sequencing the genomes of 602 *B. fragilis* isolates from 30 fecal samples. These fecal samples included longitudinal samples from 7 subjects spanning up to 2 years and single samples from 5 subjects (**Supplementary Table 2**). None of these isolates were enterotoxigenic[17] (Methods).

Isolate genomes from different subjects differed by more than 10,000 single nucleotide polymorphisms (SNPs), while genomes from the same subject differed by fewer than 100 SNPs (with one isolate exception; **Supplementary Fig. 1**). We conclude that each subject was dominated by a unique lineage, consistent with previous investigations of within-host *B. fragilis* diversity[4,16,18]. We refer to each major lineage by its host ID (e.g. L01 for Subject 01's lineage).

The SNP diversity was substantial within many lineages, allowing us to infer several years of within-person evolution. For each lineage, we assembled a draft genome using reads from all isolates, identified polymorphisms via alignment of short reads, and constructed a phylogeny (Methods, **Fig. 1a, Supplementary Fig. 2, 3**). Between 7 and 182 *de novo* SNPs were identified

per lineage (**Fig. 1b**). To estimate the age of the *B. fragilis* diversity within each subject, we calculated the average mutational distance of each population at initial sampling to its most recent common ancestor ($dMRCA_{T0}$). To convert $dMRCA_{T0}$ to units of time ($tMRCA_{T0}$), we estimated the rate at which *B. fragilis* accumulates SNPs in the human gut by comparing SNP contents across longitudinal samples from the same subject (molecular clock; **Fig. 1c**; Methods). Given our molecular clock estimate of 0.9 SNPs/genome/year, 11 of 12 subjects had values of $tMRCA_{T0}$ between 1.1-10 years (**Fig. 1d**). These values are consistent with an expansion from a single cell that existed years prior to the initial sampling, likely in the same subject.

One outlier, L08, had a significantly higher $dMRCA_{T0}$ (**Fig. 1d**, $P<0.001$, Grubb's test). This excess of mutations was due exclusively to an increase in a single type of mutation within one major sublineage (GC to TA transversions, $P<0.001$, Chi-square test), strongly suggesting that a hypermutation phenotype emerged within L08 (**Fig. 1e-f**). Hypermutation, an accelerated mutation rate usually due to a defect in DNA repair, is associated with adaptation and is commonly observed in laboratory experiments and during pathogenic infections[15,19–21]. To our knowledge, this is the first evidence of *in vivo* hypermutation in commensal bacteria. With these excess mutations (GC to TA) removed, the $dMRCA_{T0}$ for L08 was 6.9, compatible with within-person diversification.

Interestingly, each lineage's $tMRCA_{T0}$ was less than its subject's age, suggesting that these lineages colonized their subjects later in life, that adaptive or neutral sweeps purged diversity, or both. To determine if sweeps occur during colonization, we looked for mutations that fixed over time. We also examined how $tMRCA_T$ changes, where $tMRCA_T$ is defined as tMRCA of a population at a particular time point. We observed sweeps within 3 of the 7 lineages with longitudinal samples, and 2 of these 3 sweeps were associated with substantial decreases in

$tMRCA_T$ (**Supplementary Fig. 4**). Thus, sweeps appear to be common during colonization, and *B. fragilis* lineages likely resided longer in their hosts than suggested by $tMRCA_{T0}$.

We next assessed the contribution of horizontal evolution within the microbiome by identifying within-lineage mobile element differences (MEDs). We defined MEDs as DNA sequences with multi-modal coverage across isolates within a lineage (Methods). We found MEDs in 11 of the 12 lineages (**Fig. 1b**). These mobile elements include putative plasmids, integrative conjugative elements (ICEs), and prophages (**Supplementary Table 3**). We examined each MED's distribution across the phylogeny constructed using SNPs in the rest of the genome and used parsimony to categorize it as a gain or loss event. We inferred 10 elements gained, 12 lost, and 17 ambiguous loci in ~50 cumulative years of $tMRCA_{T0}$. This provided lower-bound estimates of ~0.05 gain/genome/year and ~0.04 loss/genome/year. We further estimated that MEDs change the *B. fragilis* genome by at least ~1.3 kbp gain/genome/year and ~1.9 kbp loss/genome/year. Thus, while gain and loss events are more rare than SNPs, they contribute more to nucleotide variation during *B. fragilis* evolution.

We reasoned that if these mobile elements were transferred from other species in the same microbiomes, we would observe evidence in metagenomes from the same stool communities. In particular, a transferred region should have increased coverage relative to the rest of the *B. fragilis* genome owing to its presence in other species. We leveraged stool metagenomes available from 10 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 with 38X relative coverage in the metagenomic samples (Methods; **Fig. 2a-b**). 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 strongly suggests that this prophage was acquired by *B. fragilis* during the sampling period.

The same approach helped us identify inter-species mobile element transfers of sequence regions present in all *B. fragilis* isolates of a given lineage. We identified candidate transfers in 3 subjects (**Supplementary Table 4**; **Fig. 2c**). One candidate, a putative integrative conjugative element (ICE), was confirmed in Subject 01 by culturing and sequencing 84 isolates of other *Bacteroides* species. This ICE was present in all *Bacteroides vulgatus, Bacteroides ovatus, and Bacteroides xylanisolvens* isolates (n=43, 25, and 4, respectively), but absent in all isolates of 2 other *Bacteroides* species (n=12). We found only 4 SNPs in this ICE among the four species, suggesting recent transfer among multiple species (**Fig. 2d, Supplementary Fig. 5,** Methods). This ICE contained a type VI secretion system (T6SS) of genetic architecture 2 (GA2)[22]. T6SSs of GA2 mediate inter-bacterial competition and have been shown to be shared by members of the same microbiome[16,23]. The sweep of this T6SS-containing ICE among 4 different species 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 changes.

To assess if adaptive selection was a significant driver of within-person *B. fragilis* evolution, we examined the identity of observed SNPs. We searched for within-person parallel evolution, a hallmark of positive selection[15]. We identified 17 genes mutated multiple times within a single subject, a significant deviation from a neutral model ($P<0.001$, **Fig. 3a**; **Supplementary Fig. 6**; Methods). These genes were significantly enriched for nonsynonymous mutations, as reflected by dN/dS, the normalized ratio of nonsynonymous to synonymous mutations, indicating that mutations in these genes were indeed adaptive (**Fig. 3b**).

Genes under parallel evolution reveal challenges to *B. fragilis* survival *in vivo*. The 17 genes include 5 involved in cell envelope biosynthesis, a dehydratase implicated in amino-acid metabolism, and 4 with unclear biological roles (**Fig. 3c)**. The remaining 7 genes all encode for homologs of SusC or SusD, a large group of outer-membrane polysaccharide importers (**Supplementary Table 5**). A typical *B. fragilis* lineage has 75 SusC/SusD pairs and their substrates are thought to be mainly complex yet unknown polysaccharides[24]. SusC proteins form homodimeric β-barrels capped with SusD lids[25], and the observed mutations were enriched at the interface between the barrel and lid (**Fig. 3d-e**). Notably, one of these SusC homologs (BF3581) has been shown to be critical for *B. fragilis* colonization in mice and its locus has been designated as commensal colonization factor (*ccf*)[4]. Its essentiality is thought to be related to binding to host-derived polysaccharides[4], and, therefore, mutations in Sus genes might reflect pressures to utilize host or diet-derived polysaccharides. Alternatively, the presence of Sus proteins in the outer membrane and their co-occurrence on this list with genes involved in cell envelope synthesis (**Fig. 3c,3f**) hints that selection on these genes might be driven by the pressure to evade the immune system or phage predation.

It is surprising that single amino acid changes in key genes of *B. fragilis* confer rapid adaptive advantages within individual people. These same genes show signatures of purifying selection across lineages separated by thousands of years (**Fig. 3g**). The discrepancy in signals between timescales implies that the selective forces acting on these genes are not constant and raises the possibility that adaptive mutations occurring *in vivo* may incur collateral fitness costs in the context of other selective forces[26,27]. This notion of competing selective forces is echoed by the well-described invertible promoters of *B. fragilis,* which enable rapid alternation between different outer-membrane presentations[28,29]. Interestingly, the invertible promoters control the

same major pathways that we identified as undergoing positive selection (capsule synthesis and polysaccharide importers)[28,30]. The non-constant selective forces driving these inversions and mutations might be specific to some people or lineages, recently introduced into the human population, present only at particular times (e.g. during early stages of colonization), or coexisting within individual people (**Fig. 3h**). We found evidence of both subject-specific and other selective forces. Three Sus genes (BF1802, BF1803, and BF3581) were each mutated multiple times within a subject, ($P < 0.003$ for each, Fisher's exact test), yet no times in other subjects. In contrast, six genes under selection were mutated in multiple lineages, with three genes even acquiring mutations at the same amino-acid residue in different lineages (BF4056, BF1708 and BF2755; **Fig. 3c**). Remarkably, a BF2755 mutation (Q100P) found polymorphic in 3 subjects was also in the ancestor of L12 and two publicly available genomes (**Supplementary Fig. 7**), suggesting a common and strong selective pressure on this amino acid.

Could competing selective forces create multiple coexisting niches for *B. fragilis* even within a same individual? We noticed that the two lineages with the largest $dMRCA_{T0}$ (L01 and L08) had long-branched, co-existing sublineages that might reflect niche-differentiation (**Supplementary Fig. 3a, Fig. 1d**). We closely examined L01's evolutionary history over a 537-day period, during which the relative abundance of *B. fragilis* did not substantially change, using 206 stool metagenomes (**Supplementary Fig. 8a**). We tracked 21 abundant SNPs whose evolutionary relationships were previously identified from isolate genomes and inferred the population dynamics of their corresponding sublineages (**Fig. 4a-c**; Methods). The relative ratio of the two major sublineages (SLs), SL1 and SL2, which diverged ~8 years prior to sampling, remained stable across the 1.5-year period (**Fig. 4c; Supplementary Fig. 8b**). SL1 showed signatures of rapid adaptation during this period, including mutations in genes under selection,

competition of mutations through clonal interference (e.g. between SL1-a and SL1-b, and within SL1-a), and a rapid sweep involving two SNPs related to Sus genes (SL1-a-1; **Fig. 4c-d**). The continued coexistence of SL1 and SL2 despite a sweep within SL1 is particularly striking and suggests frequency-dependent selection or occupation of distinct, perhaps spatially segregated, niches[31–34]. The fact that 11 of 12 intragenic mutations separating these sublineages are amino-acid changing furthers the notion that they are functionally distinct. Therefore, it is likely that *B. fragilis* niche-differentiation can occur within a single person.

Within the gut microbiome of individual people, *B. fragilis* acquires adaptive mutations in key genes, including polysaccharide importers and capsule synthesis genes, under the pressure of natural selection. While some of this adaptation, like that of opportunistic infections in the lungs of people with cystic fibrosis[15,35], may reflect common emerging selective forces, our results suggest that a subset of this adaptation is person-specific. Person-specific selection may contribute to observed microbiome stability[1], in that indigenous bacteria may be more adapted to an individual's ecosystem than foreign bacteria attempting to invade the microbiome later in life. Further work is required to identify whether rapid adaptation is specific to *B. fragilis* or a common feature of gut commensals, as well as how one species' evolution interacts with community composition and human health. The presence of strong selection within individuals' microbiomes suggests that the design of stably-colonizing probiotics and other microbiome manipulations may require personalized approaches based on genomewide profiling.

## 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, Kathryn Kauffman, 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, and Zhenrun Zhang for comments on the manuscript. This work was funded by a grant from the Broad Institute.

## 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*; S.M.G, R.J.X., and E.J.A. coordinated acquisition of metagenomic data. S.Z. and T.D.L. analyzed the data; S.Z., T.D.L., and E.J.A wrote the manuscript with input from all authors.

## References

1. Faith, J. J. *et al.* The Long-Term Stability of the Human Gut Microbiota. *Science.* **341,** 1237439–1237439 (2013).
2. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493,** 45–50 (2012).
3. Zoetendal, E. G., Akkermans, A. D. & De Vos, W. M. Temperature gradient gel electrophoresis analysis of 16S rRNA from human fecal samples reveals stable and host-specific communities of active bacteria. *Appl. Environ. Microbiol.* **64,** 3854–9 (1998).
4. Lee, S. M. *et al.* Bacterial colonization factors control specificity and stability of the gut microbiota. *Nature* **501,** 426–429 (2013).
5. Sender, R., Fuchs, S. & Milo, R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol.* **14,** 1–14 (2016).
6. Barrick, J. E. & Lenski, R. E. Genome dynamics during experimental evolution. *Nat. Rev. Genet.* **14,** 827–39 (2013).
7. Nayfach, S. & Pollard, K. S. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biol.* **16,** 51 (2015).
8. Korem, T. *et al.* Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science.* **349,** 1101–1106 (2015).
9. He, M. *et al.* Evolutionary dynamics of *Clostridium difficile* over short and long time scales. *Proc. Natl. Acad. Sci.* **107,** 7527–7532 (2010).
10. Groussin, M. *et al.* Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. *Nat. Commun.* **8,** 14319 (2017).
11. Goodrich, J. K. *et al.* Genetic Determinants of the Gut Microbiome in UK Twins. *Cell Host Microbe* **19,** 731–43 (2016).
12. Lieberman, T. D. *et al.* Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nat. Genet.* **43,** 1275–1280 (2011).
13. Barroso-Batista, J., Demengeot, J. & Gordo, I. Adaptive immunity increases the pace and predictability of evolutionary change in commensal gut bacteria. *Nat. Commun.* **6,** 8945 (2015).
14. Chattopadhyay, S. *et al.* High frequency of hotspot mutations in core genes of *Escherichia coli* due to short-term positive selection. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 12412–12417 (2009).
15. Lieberman, T. D. *et al.* Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. *Nat Genet* **46,** 82–87 (2014).
16. Verster, A. J. *et al.* The Landscape of Type VI Secretion across Human Gut Microbiomes Reveals Its Role in Community Composition. *Cell Host Microbe* **22,** 411–419.e4 (2017).
17. Chen, L. *et al.* VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.* **33,** D325–D328 (2004).
18. Yassour, M. *et al.* Natural history of the infant gut microbiome and impact of antibiotic treatment

on bacterial strain diversity and stability. *Sci. Transl. Med.* **8,** 343ra81-343ra81 (2016).
19. Giraud, A. Costs and Benefits of High Mutation Rates: Adaptive Evolution of Bacteria in the Mouse Gut. *Science.* **291,** 2606–2608 (2001).
20. Chu, N. D. *et al.* A Mobile Element in mutS Drives Hypermutation in a Marine *Vibrio*. *MBio* **8,** e02045-16 (2017).
21. Jolivet-Gougeon, A. *et al.* Bacterial hypermutation: clinical implications. *J. Med. Microbiol.* **60,** 563–573 (2011).
22. Coyne, M. J., Roelofs, K. G. & Comstock, L. E. 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,** 58 (2016).
23. Coyne, M. J. *et al.* Evidence of Extensive DNA Transfer between *Bacteroidales* Species within the Human Gut. *MBio* **5,** e01305-14 (2014).
24. Cerdeno-Tarraga, A. M. Extensive DNA Inversions in the *B. fragilis* Genome Control Variable Gene Expression. *Science (80-. ).* **307,** 1463–1465 (2005).
25. Glenwright, A. J. *et al.* Structural basis for nutrient acquisition by dominant members of the human gut microbiota. *Nature* **541,** 407–411 (2017).
26. Messer, P. W., Ellner, S. P. & Hairston, N. G. Can Population Genetics Adapt to Rapid Evolution? *Trends Genet.* **32,** 408–418 (2016).
27. Bell, G. Fluctuating selection: the perpetual renewal of adaptation in variable environments. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **365,** 87–97 (2010).
28. Krinos, C. M. *et al.* Extensive surface diversity of a commensal microorganism by multiple DNA inversions. *Nature* **414,** 555–558 (2001).
29. Porter, N. T., Canales, P., Peterson, D. A. & Martens, E. C. A Subset of Polysaccharide Capsules in the Human Symbiont *Bacteroides thetaiotaomicron* Promote Increased Competitive Fitness in the Mouse Gut. *Cell Host Microbe* **22,** 494–506 (2017).
30. Kuwahara, T. *et al.* Genomic analysis of *Bacteroides fragilis* reveals extensive DNA inversions regulating cell surface adaptation. *Proc. Natl. Acad. Sci.* **101,** 14919–14924 (2004).
31. Rocabert, C., Knibbe, C., Consuegra, J., Schneider, D. & Beslon, G. Beware batch culture: Seasonality and niche construction predicted to favor bacterial adaptive diversification. *PLOS Comput. Biol.* **13,** e1005459 (2017).
32. Chung, H. *et al.* Global and local selection acting on the pathogen *Stenotrophomonas maltophilia* in the human lung. *Nat. Commun.* **8,** 14078 (2017).
33. Good, B. H., McDonald, M. J., Barrick, J. E., Lenski, R. E. & Desai, M. M. The dynamics of molecular evolution over 60,000 generations. *Nature* (2017). doi:10.1038/nature24287
34. Plucain, J. *et al.* Epistasis and allele specificity in the emergence of a stable polymorphism in *Escherichia coli*. *Science (80-. ).* 1242862 (2014).
35. Smith, E. E. *et al.* Genetic adaptation by *Pseudomonas aeruginosa* to the airways of cystic fibrosis patients. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 8487–92 (2006).

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

| Number of bacteria per microbiome (cells/microbiome)[5] | Mutation rate of bacteria (SNP/nucleotide/replication)[6] | Average bacterial genome size (nucleotide/cell)[7] | Range of mean bacterial replication rate (replication/day)[8] | → | Estimated number of *de novo* mutation (SNP/microbiome/day) |
|---|---|---|---|---|---|
| $10^{13}$ -$10^{14}$ | $10^{-9}$ -$10^{-10}$ | 2-6×$10^{6}$ | 1-10 | | 2×$10^{9}$ - 6×$10^{12}$ |

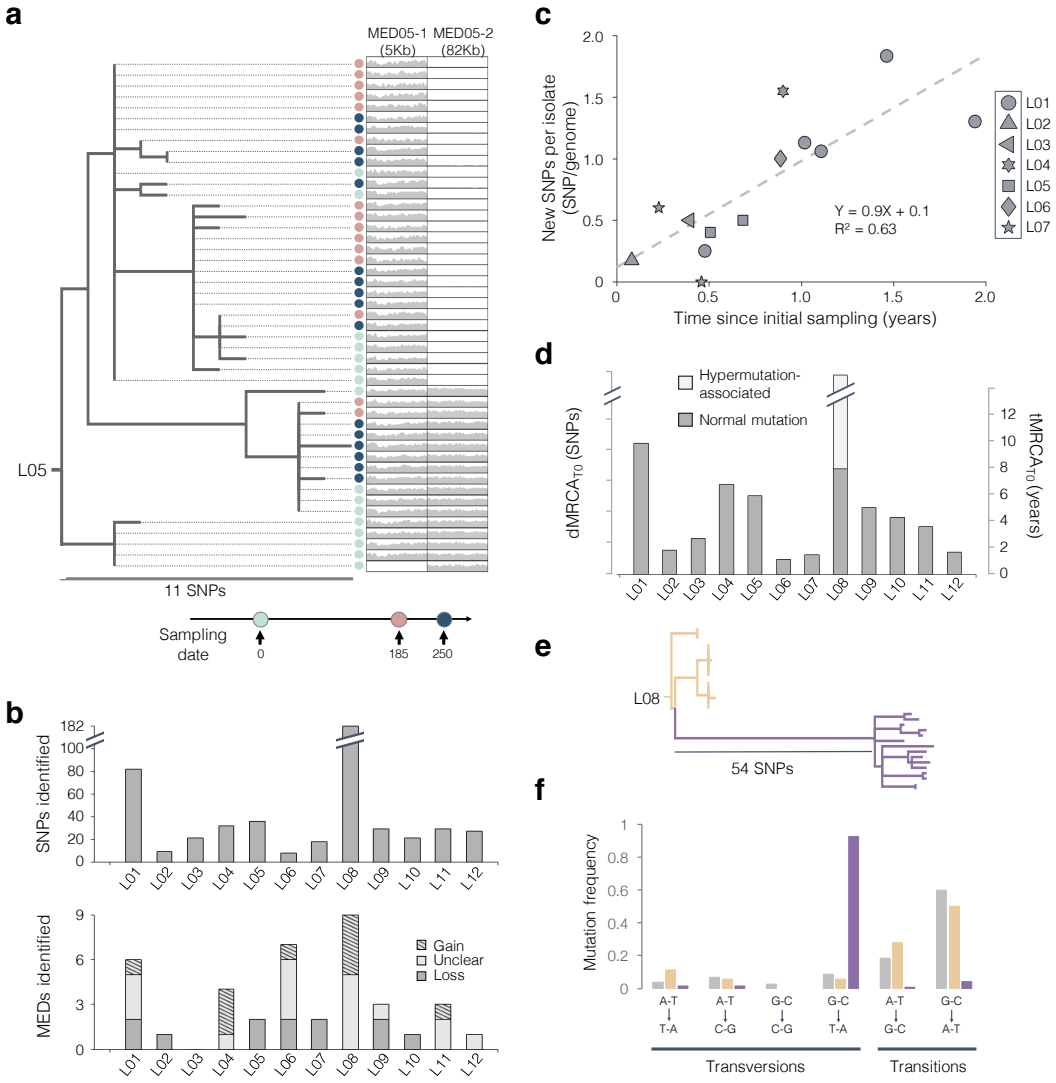



**Figure 1 | *B. fragilis* lineages diversify for years during colonization of healthy individuals via *de novo* SNPs and MEDs**. (**a**) The phylogeny of isolates from L05 is shown as an example. Light blue, pink, and dark blue circles indicate isolates taken at Day 0, 185, and 250, respectively. For each isolate, the relative coverage (compared to the mean genomewide) across the length of two identified mobile element differences (MEDs) is shown. For each isolate with an MED, the average relative coverage is ~1X. (**b**) The number of SNPs and MEDs identified for each lineage. MEDs were classified as gained (hatched), lost (dark gray), or unclear (light gray). (**c**) Estimate of the molecular clock for *B. fragilis.* Each shape represents the average number of new SNPs per isolate not present in the set of SNPs at initial sampling. (**d**) dMRCA$_{T0}$ and tMRCA$_{T0}$ for the 12 lineages. For L08, the contribution to these estimates is separated into that from hypermutation-induced (clear) and normal mutations (grey). (**e**) A rooted phylogeny of L08, with hypermutation sublineage (purple) and normal sublineages (yellow). (**f**) The spectrum of mutations in the hypermutation sublineage (purple), normal sublineages of L08 (yellow), and 11 other lineages (gray).

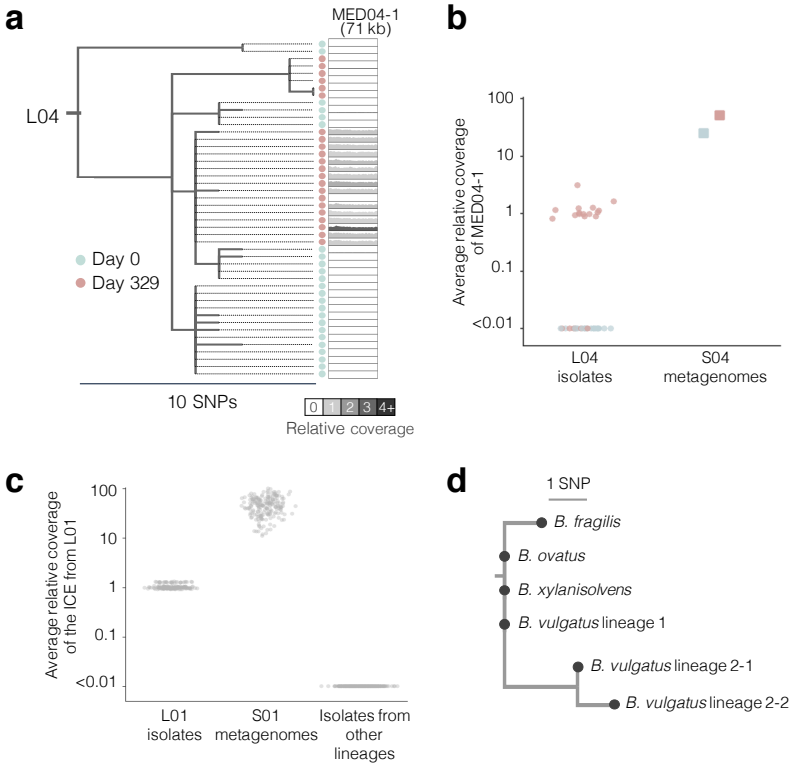



**Figure 2 | Mobile elements are transferred within the microbiome of individual subjects.** (**a**) The phylogeny of isolates from L04, illustrating gain of MED04-1. Blue and pink circles indicate isolates taken at Day 0 and 329, respectively. Shading of the MED region reflects the average relative coverage of the MED in that isolate. (**b**) Average relative coverage across the length of MED04-1, a prophage, in L04 isolates (circle) and Subject 04 metagenomic samples (square). Colors represent sampling dates as shown in (**a**). Isolates with this prophage had from 1X to 3X average coverage relative to the rest of the genome. (**c**) Average relative coverage of a putative integrative conjugative element (ICE) in isolates from L01, metagenomic samples from Subject 01, and isolates from other lineages. Isolates from the sample S01-0259 show slightly higher average relative coverage because genomic libraries of these isolates were prepared differently (Methods). (**d**) A rooted parsimonious phylogeny of the putative ICE across 4 species. Isolates that had identical ICE sequences and were from the same phylogenetic group are merged into a single node. In Subject 01, the *B. vulgatus* isolates were from 2 distinct lineages, one of which had 2 sublineages (**Supplementary Fig. 5**).

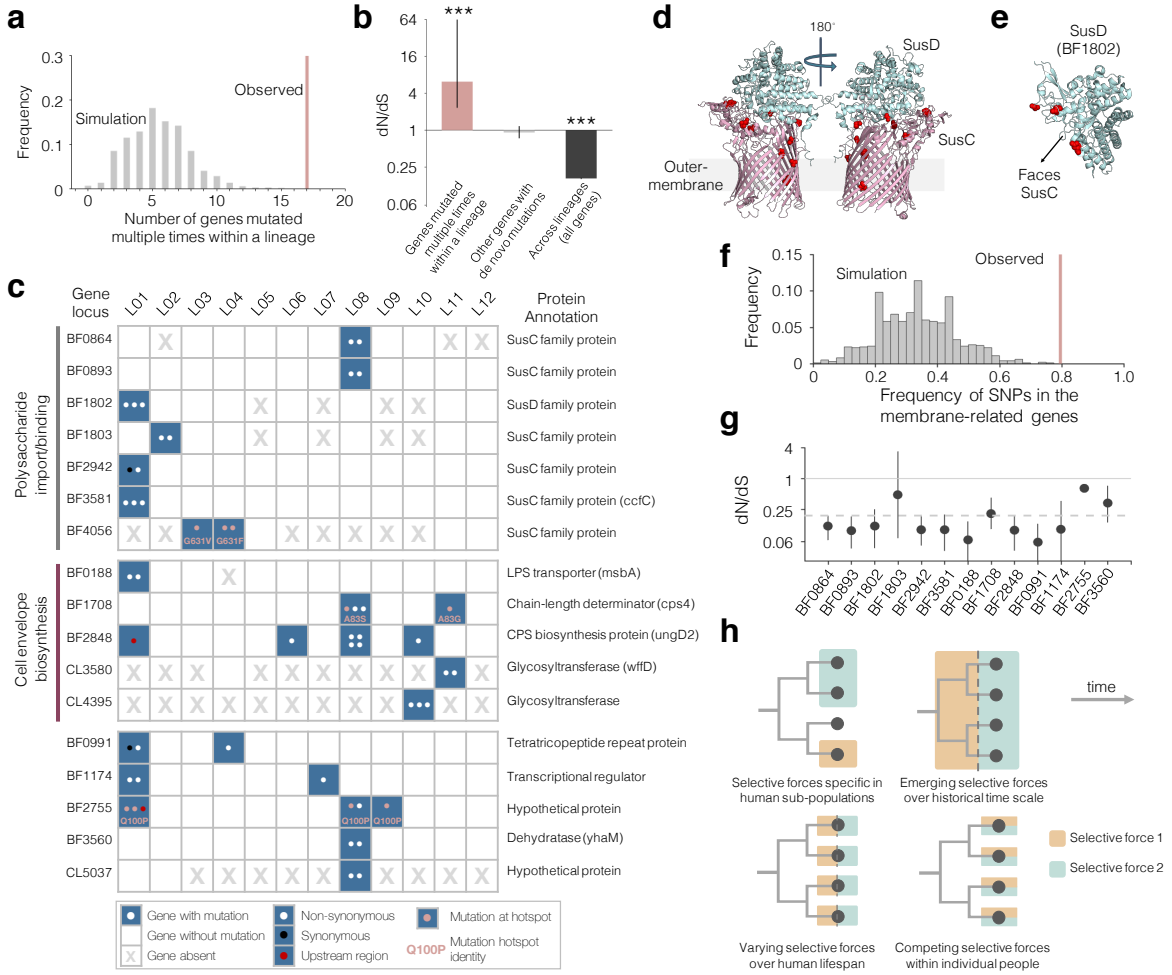
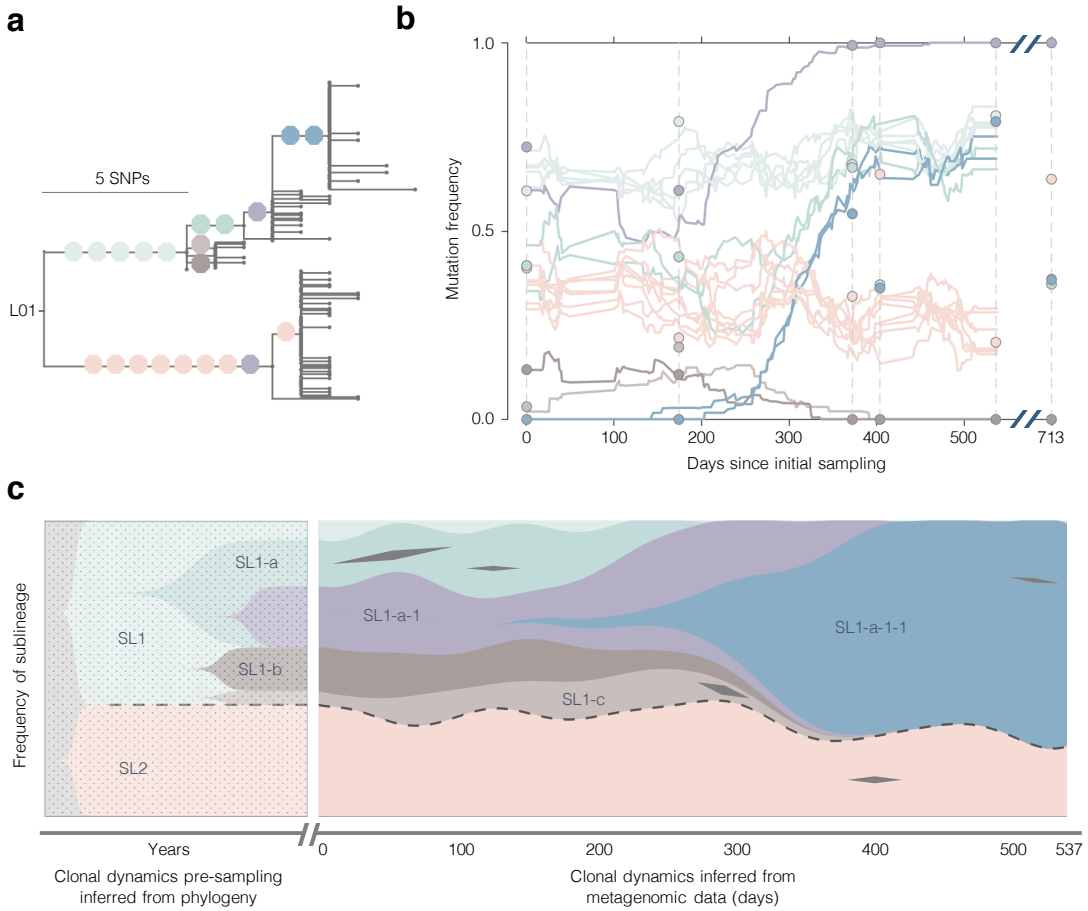



**Figure 3 | Genes involved in polysaccharide utilization and cell envelope biosynthesis are under selection to change during colonization.** (**a**) Simulated under a null model (grey) and observed (pink) number of genes mutated multiple times within at least one lineage (P<0.001, Methods). (**b**) Canonical signal for selection, dN/dS, shows a signature of adaptive evolution in genes mutated multiple times (P<0.001) but not for other *de novo* mutations. Mutations across lineages show a significant signature of purifying selection (P<0.001). Bars represent 95% confidence intervals for dN/dS. (**c**) The 17 genes mutated multiple times are grouped by their biological functions and labeled with their locus tags and the inferred functions of their encoded proteins (**Supplementary Table 5**). Genes absent in a lineage are indicated with a gray X. The number and types of SNPs are indicated. (**d-e**) Mutations in SusC and SusD homologs under selection were enriched at the interface between the proteins. (**f**) Simulated under a null model (grey) and observed (pink) frequency of mutations in membrane-related genes (P<0.001, Methods). (**g**) Genes mutated multiple times show significant signatures of purifying selection across lineages (for 13 genes with inter-lineage mutations). The dashed line represents the average dN/dS for all inter-lineage SNPs. (**h**) Four models that could account for the discrepancy of natural selection at different timescales.




d

| SL1 | SL2 | | SL1-a | SL1-a-1 | SL1-b |
|---|---|---|---|---|---|
| **BF1802-D526N -10 BF2848** | **BF3581-Q974R** | BF1873-A112V | CL10492-SYN | **BF2755-Q100P** | **BF0991-SYN** |
| BF0257-A61T | **BF3581-P240L** | CL02988-SYN | BF1401-L119W | **SL1-a-1-1** | **SL1-c** |
| BF0700-R41H | **BF2942-E769K** | CL02250-C105R | **(BF0991-A16V)** | **BF1802-T340M** | **BF1802-G312A** |
| CL01902-UP67 | **BF2755-Q100P** | Intergenic region | **(-165 BF2755)** | -260 BF0868 | **(BF3581-K751R)** |
| | BF3635-I260V | **(BF1174-K62*)** | | **(BF0188-D61N)** | |

**Figure 4 | Two sublineages coexisted at a stable relative frequency despite rapid adaptive dynamics within one sublineage, suggesting niche differentiation within L01.** (**a**) The phylogeny of isolates from L01. Branches with ≧4 isolates are labeled with colored octagons that represent individual SNPs. One SNP was inferred to have happened twice and is indicated in two locations (purple). (**b**) Frequencies of labeled SNPs over time in the *B. fragilis* population were inferred from 206 stool metagenomes (Methods). Colored circles represent SNP frequencies inferred from isolate genomes at particular time points. (**c**) The history of the sublineages carrying these SNPs prior to (left) and during (right) sampling was inferred (see Methods). The prior-to-sampling history is shaded to indicate temporal uncertainty. Sublineages are labeled with names and colored as in (**a**). Black diamonds represent transient SNPs from genes with multiple mutations. The two major sublineages, SL1 and SL2, are separated by dashed lines. (**d**) The identity of SNPs shown in (**c**) are listed in the table. SNPs in the 17 genes with multiple mutations in any subject are bolded and transient mutations in genes mutated multiple times are indicated with parentheses. Negative numbers indicate mutations upstream of the start of the gene. SYN indicates synonymous mutations.

# Methods

**Study cohort and sample collection**

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. All 12 subjects were healthy people screened by OpenBiome to minimize the potential for carrying pathogens and had ages between 22-37 years and body-mass indexes between 19.5-26.2 at initial sampling. Subjects were de-identified before receipt of samples. **Supplementary Table 1** contains detailed information about each subject.

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 -80C. 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 10 of the 12 subjects (352 stool samples in total). Detailed information about samples used for isolation, including handling conditions prior to sample receipt, is in **Supplementary Table 2** and information about samples used for metagenomic sequencing is in **Supplementary Table 6**.

**Library construction and Illumina sequencing**

Samples were serially diluted in phosphate-buffered saline (PBS) and cultured for *B. fragilis* on *Bacterodies* 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[36]. Libraries from one sample (S01-0259, Day 709) were prepared by the BioMicroCenter (BMC) at MIT using a similar protocol, with lower input DNA and a final Pippen size-selection step. 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 (**Supplementary Table 2**). Only isolates with average coverage of greater than 10 reads across the *B. fragilis* genome were included for analysis.

***de novo* assemblies of lineage genomes**

Reads were first trimmed and filtered using Cutadapt[37] and Sickle[38] (pe -f 20 -r 50). For each major lineage, we concatenated the first 0.25 million pairs of reads from each isolate, and we used this concatenated file as the input for *de novo* genome assembly via SPades v3.10.0 (parameter: --careful)[39]. Isolates prepared by the BMC, as well as a few isolates with apparent cross contamination (genome assembly built only using reads from an isolate was larger than 6MB; a typical *B. fragilis* genome assembly is ~5MB) were excluded in building assemblies. Isolates not used to build the genome assembly are indicated as such in the metadata associated with the uploaded raw data (see **Data availability**). Statistics of these genome assemblies are in

**Supplementary Table 1**. Assembly genomes were annotated using Prokka v1.11[40]. A genome assembly of the minor lineage from S10 was built using all reads from this isolate.

**Toxin detection**

We compared the genome assemblies of the 12 major lineages and 1 minor lineage to the Virulence Factors Database, which contains >2400 virulence factors[17], 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).

**Intra-subject and inter-subject SNPs**

To identify intra-subject mutations, trimmed and filtered short reads from isolates of the same subject were aligned to the lineage genome assembly using Bowtie2 (Alignment parameters: -X 2000 --no-mixed --very-sensitive --n-ceil 0,0.01 --un-conc). Candidate SNPs were identified using SAMtools[41] and filtered using custom filters modified from previous work[15]. 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; lower values indicate more agreement between reads) 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. Candidate positions in MEDs were also discarded (including homologous regions shared between MED01-1 and MED01-2). For lineage 10, the major allele frequency filter was set to 95%. Detailed information of intra-subject SNPs from the 12 subjects are listed in **Supplementary Tables 7-18.**

For Subject 10, reads from the minor lineage isolate were aligned to the genome of the major lineage to identify the number of intra-subject mutations between the minor and the major lineages. To estimate the distance between lineages from different subjects, we aligned all short reads to a publicly-available reference genome NCTC9343 (NCBI accession: CR626927.1) using the same methods for intra-subject mutation identification.

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

For each major lineage, a phylogeny of all isolates was built using a list of concatenated intra-subject SNPs and the closest lineage as an outgroup. While many filters were used for SNP calling, only the major nucleotide for each isolate at each called genomic position was used for phylogenetic inference. We used the dnapars program, a parsimony tree builder from PHYLIP v3.69 to infer the phylogeny[42]. 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**

To calculate $dMRCA_T$ (dMRCA of isolates from a particular time point T) for each subject at each time point, we counted the number of alleles that were different from ancestral

alleles for each isolate, assessing only SNP positions that were polymorphic among isolates from the particular time point, and averaged the results.

Collector curves for $dMRCA_T$ indicate that undersampling was a minor contributor to error in estimation of $dMRCA_T$ (**Supplementary Fig. 9)**. Interestingly, collector curves for the number of *de novo* SNPs reflect that the number of SNPs identified did not saturate (**Supplementary Fig. 10**).

**Mutation rate and tMRCA**

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 (**Fig. 1c**). The positive y-intercept reflects that new colonies from the same time point also bring in new SNPs, due to non-exhaustive sampling (**Supplementary Fig. 10**). $tMRCA_T$ was calculated by dividing $dMRCA_T$ by the estimated mutation rate (**Fig. 1d**).

**Identification of Mobile element difference (MED)**

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, that had low relative coverage (< 0.2X) at every nucleotide in at least one isolate, and that had >0.9X coverage at every nucleotide in at least one isolate. For L01, we excluded isolates from the last 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 manually separated these clusters of sequence regions into clusters with consistent coverage distribution patterns. Detailed information of all MEDs is in **Supplementary Table 3**.

**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 (**Supplementary Table 3; Fig. 1b**). 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 (**Supplementary Fig. 4f**). 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 $tMRCA_{T0}$. The following equation was used for gain and loss events, separately:

$$\frac{\sum_i \sum_j f_{ij}}{\sum_i tMRCA_{T0,i}}$$

To estimate the absolute contribution of gain and loss events to the size of *B. fragilis* genomes, we accounted for length of each MED ($L_{ij}$).

$$\frac{\sum_i \sum_j L_{ij} f_j}{\sum_i tMRCA_{T0,i}}$$

**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.

**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 1000X 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 30$^{th}$ percentile and 70$^{th}$ percentiles (median was not precise given the low coverage). To identify recent transfer events, we searched the genome for candidate regions >5000 nucleotides in length and in which the consensus genome from metagenomics was <0.02% different from the consensus genome from isolates. 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 **Supplementary Table 4**.

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 (**Supplementary Fig. 5c**), 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 (**Supplementary Table 4**). To test if the putative ICE was indeed transferred between species, we cultured and sequenced the genomes of 84 *Bacteroides* isolates from this subject. We examined 43 *Bacteroides vulgatus* isolates, 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 within a subject if, in at least one subject, the gene had multiple 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). To account for parallel evolution occurring at the same nucleotide position, we leveraged the phylogenies and counted each independent occurrence of a mutation separately. 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 and calculated how many genes showed a signature of selection (**Fig. 3a**). 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[43]. Detailed information for each gene under parallel evolution is in **Supplementary Table 5** and gene clusters are listed in **Supplementary Table 19.** Simulations performed for metrics of cross-subject parallel evolution did not yield additional signatures of adaptive evolution (**Supplementary Fig. 6**).

**dN/dS**

Mutations were categorized as synonymous (S) or non-synonymous (N) based on open-reading frame annotations created by Prokka[40]. dN/dS calculations were performed as previously described, normalizing for the spectrum of mutations observed within each set of genes[15]. 95% confidence intervals were calculated using binomial sampling.

**Annotation of genes under selection**

To discover homologs of the seventeen 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, using 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[44]. Cellular localizations were predicted using CELLO. Detailed information is in **Supplementary Table 5.**

**Mapping SusC and SusD mutations on protein structures**

Available crystal structures of a SusC homolog (BT1763) from *Bacteroides thetaiotaomicron*[25] and BF1802 from *B. fragilis* NCTC_9343[45] were used to visualize the mutations observed in Sus genes under parallel evolution. We aligned the 6 *B. fragilis* SusC proteins under parallel evolution and BT1763 using Clustal Omega from the EMBL-EBI web service[46] (default parameters). For all non-synomymous 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[25]. Non-synonymous mutations from Sus genes under parallel evolution are marked in red in **Fig. 3d** and **Fig. 3e**.

**Enrichment of membrane protein**

For all genes from the 12 major lineage genome assemblies, we used CELLO[47] 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 seventeen 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 seventeen 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 seventeen genes under selection, 79% of the SNPs are from membrane-related genes, a significant deviation from the null distribution ($P<0.001$, **Fig. 3f**).

**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 9 genes that were mutated only in one subject. The p-values for BF1802, BF3581, BF1803, are all less than 0.005, suggesting person-specific adaptation.

**Mutation dynamics**

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 sequencing depth was limited and *B. fragilis* represented only ~5% of reads on average, 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 (**Fig. 4b**).

To plot a schematic of the population dynamics of different sublineages (**Fig. 4c**), 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 370, 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.

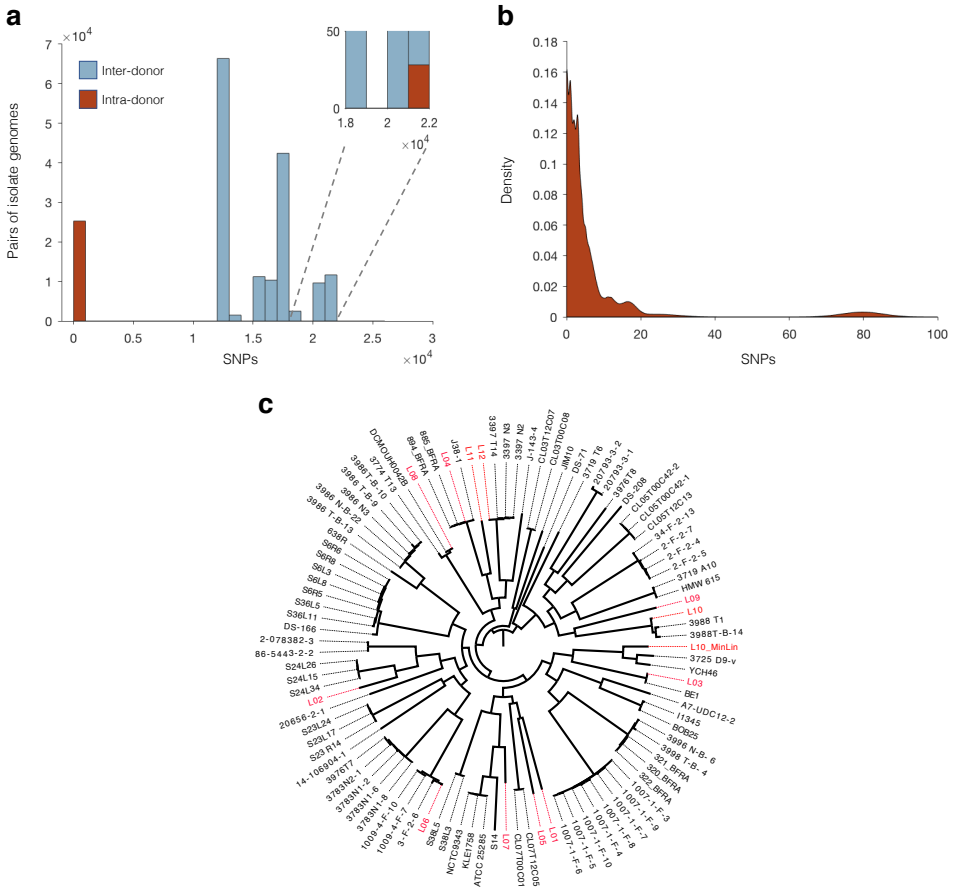
We also examined L03's dynamics during colonization using 75 metagenomics samples collected over 144 days (**Supplementary Fig. 11**). The same methods were used as described above, with the exception that mutations in $\geqq 3$ isolates were able to be tracked, owing to the higher relative abundance of *B. fragilis* in Subject 03. This schematic shows an expansion of a SNP and SNPs that decreases over time.

**Data availability**

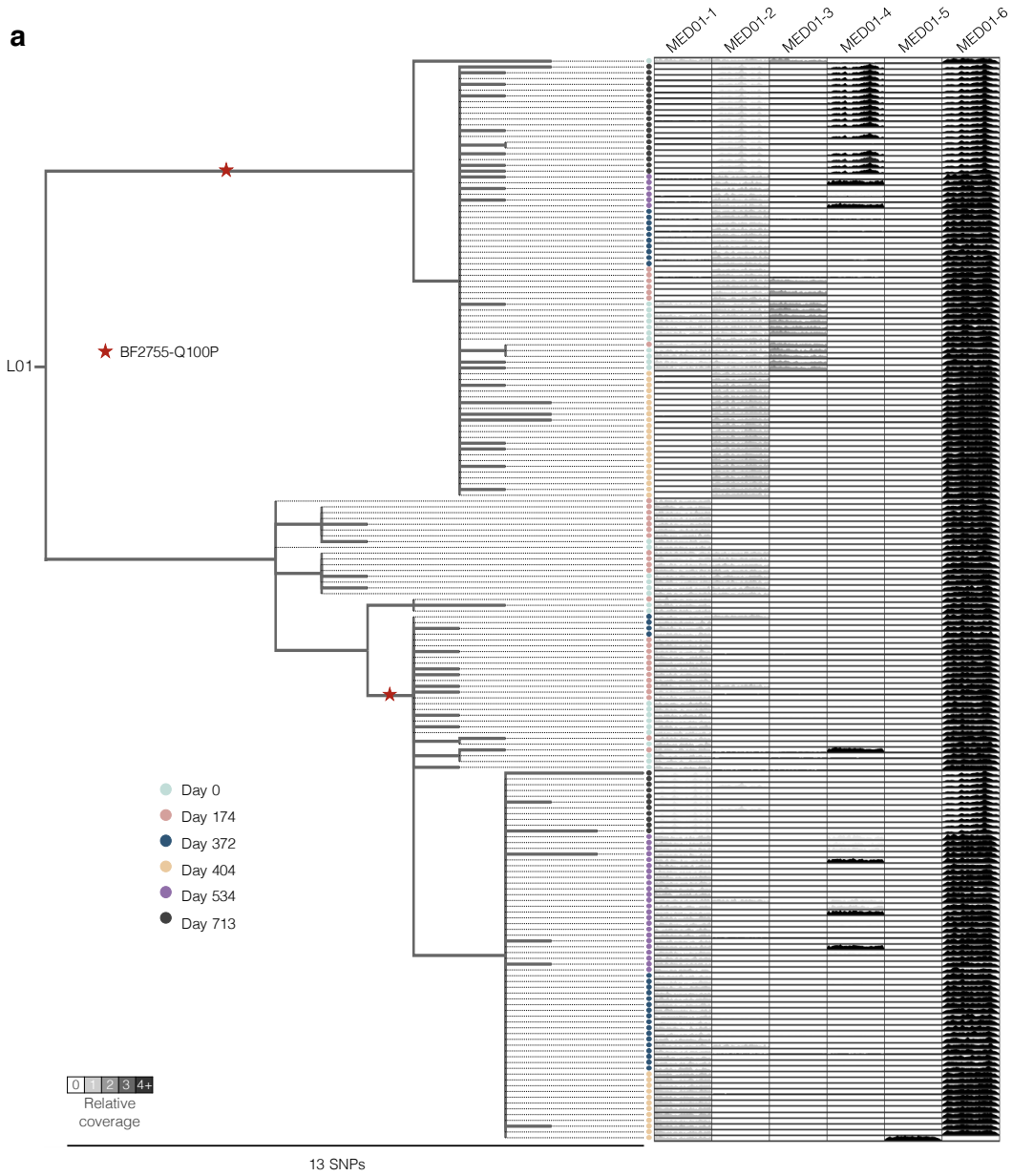
Data is in the process of being uploaded to public servers. FASTQ files for the 602 *B. fragilis* isolates, with adaptors removed and filtered for quality, will be uploaded to the SRA. BAM files of the 352 metagenomes aligned to *B. fragilis* lineage assemblies will also be available on the SRA. Lineage assemblies with annotations will be uploaded to NCBI.

# Methods references:

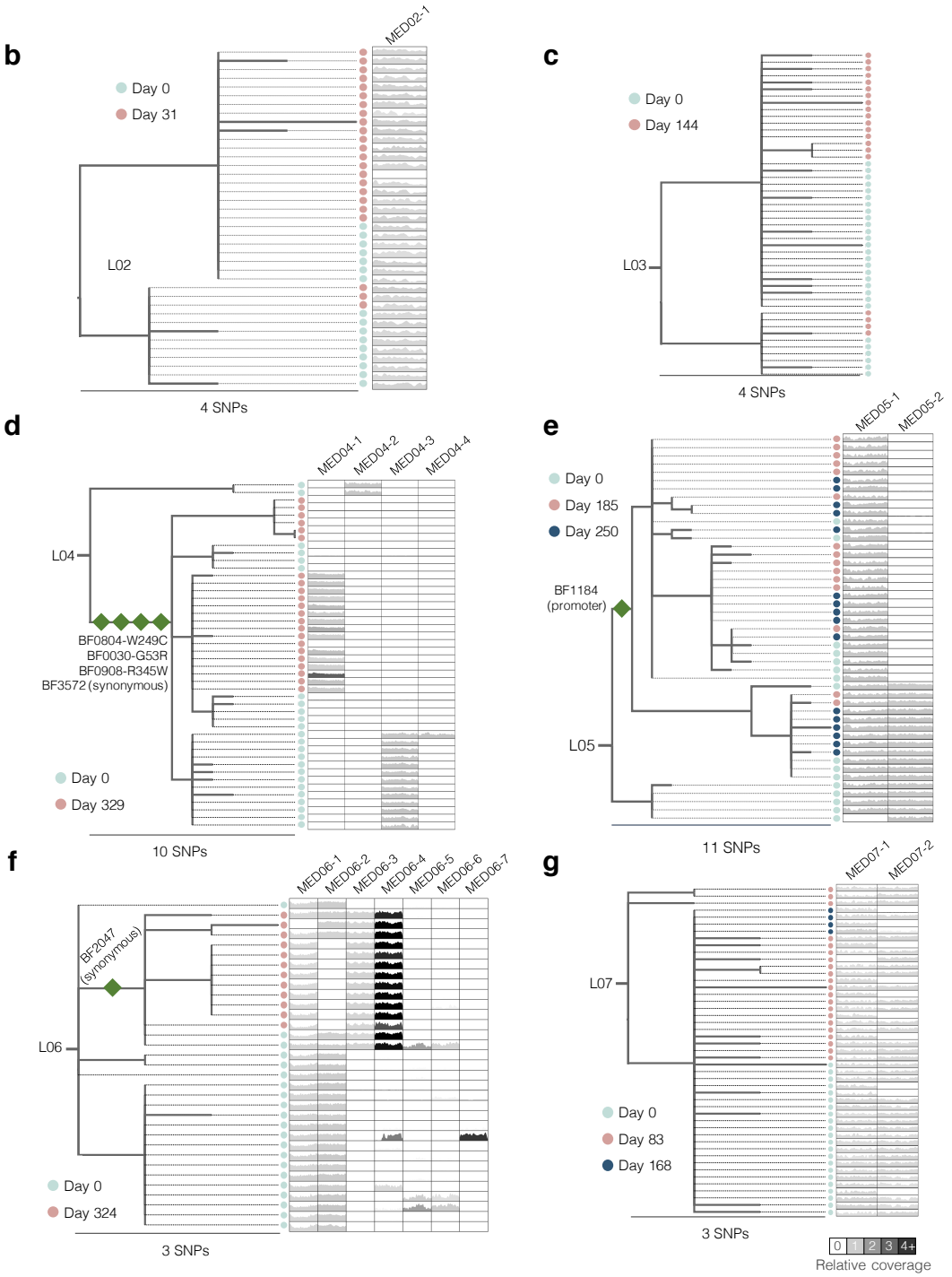

36. Baym, M. *et al.* Inexpensive Multiplexed Library Preparation for Megabase-Sized Genomes. *PLoS One* **10,** e0128036 (2015).
37. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* **17,** 10–12 (2011).
38. Joshi, N. A. & Fass, J. N. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33)[Software]. (2011).
39. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19,** 455–477 (2012).
40. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30,** 2068–2069 (2014).
41. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
42. Plotree, D. & Plotgram, D. PHYLIP-phylogeny inference package (version 3.2). *cladistics* **5,** 6 (1989).
43. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28,** 3150–3152 (2012).
44. Price, M. N. & Arkin, A. P. PaperBLAST: Text Mining Papers for Information about Homologs. *mSystems* **2,** e00039-17 (2017).
45. Joint Center for Structural Genomics. Crystal structure of a SusD superfamily protein (BF1802) from *Bacteroides fragilis* NCTC 9343 at 1.90 A resolution. *The Protein Data Bank* (2010). doi:10.2210/pdb3nqp/pdb
46. McWilliam, H. *et al.* Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res.* **41,** W597–W600 (2013).
47. Yu, C.-S., Chen, Y.-C., Lu, C.-H. & Hwang, J.-K. Prediction of protein subcellular localization. *Proteins Struct. Funct. Bioinforma.* **64,** 643–651 (2006).



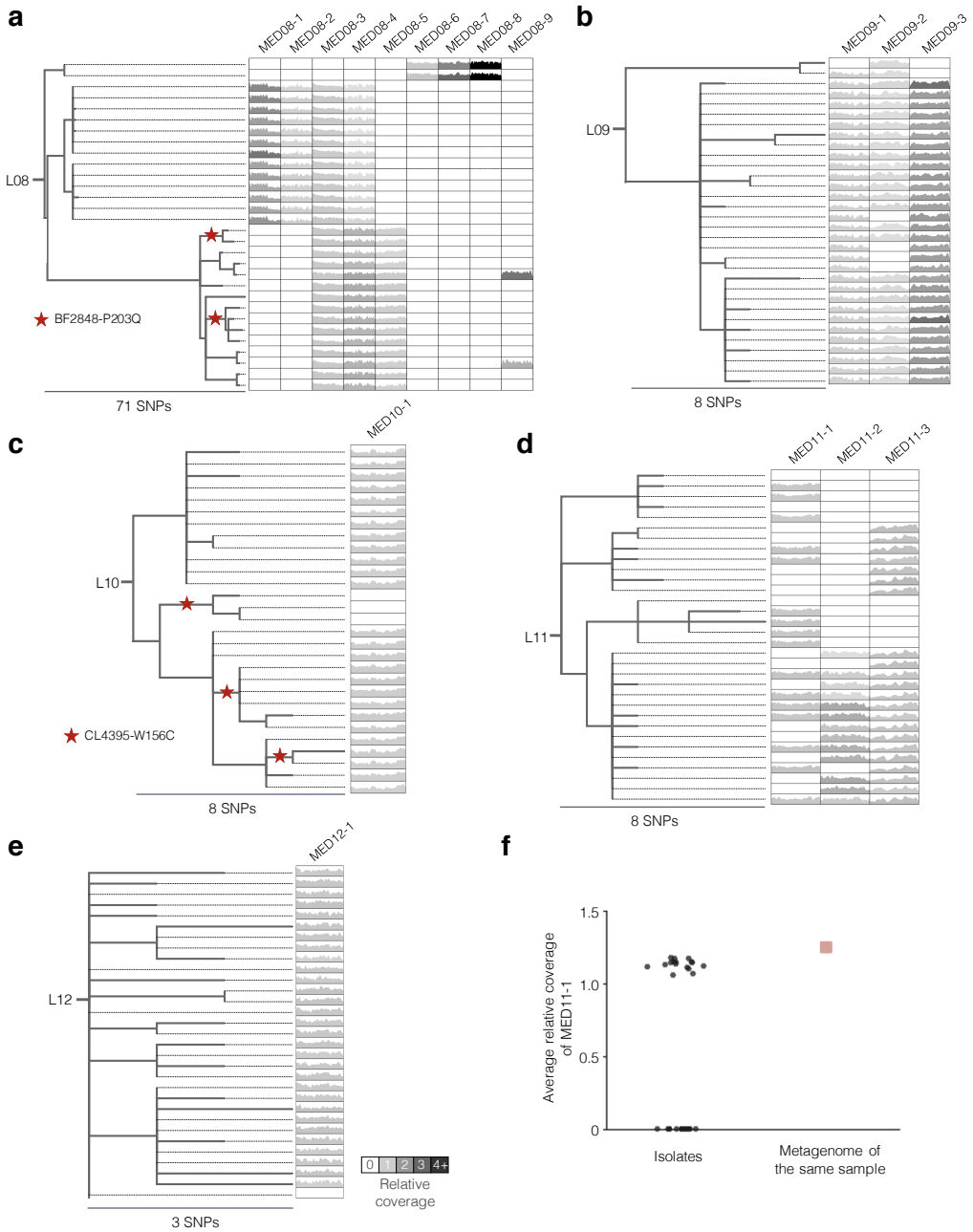

**Supplementary Figure 1 | Inter-subject and intra-subject mutational distances between pairs of isolates suggest that each individual subject has a dominant *B. fragilis* lineage.** (**a**) Histogram of the mutational distances between all pairs of isolates. Inter-subject pairs are shown in blue, while intra-subject pairs are in red. The bin size is 1000 SNPs. Twenty-eight intra-subject pairs are >22000 SNPs apart and emerged from one isolate from Subject 10 that was from a minor lineage. (**b**) Excluding this minor lineage, all intra-subject mutational distances were <100 SNPs. The probability distribution of intra-subject mutational distances, averaging across 12 subjects, is shown. (**c**) Phylogeny of genomes from 12 major lineages, 1 minor lineage from L10 and 88 references from NCBI. We clustered coding sequences from these 101 genomes with 95% similarity using CD-HIT and identified 277 genes present in all genomes. The number of shared genes is an underestimate, as the available genome assemblies had varying quality. We performed multiple sequence alignment for each shared gene using MAFFT v7.310[48] and concatenated the alignment files. A phylogenetic tree was constructed using the GTRGAMMAI model from RAxML v8.2.11 (parameters: -m GTRGAMMAI -p 12345 -# 20)[49].



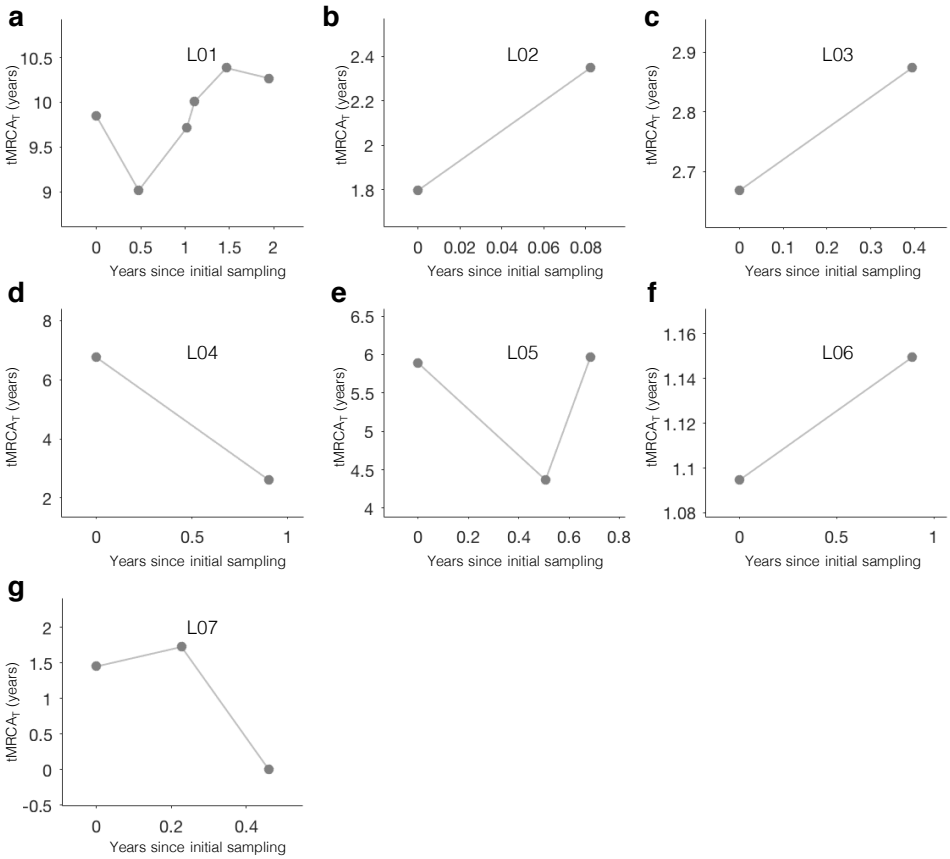

**Supplementary Figure 2 | Within-person *B. fragilis* evolution in subjects with longitudinal samples (continued on next page).** (**a-g**) The phylogeny for isolates from L01, L02, L03, L04, L05, L06, and L07, respectively. Colored circles represent isolates from samples collected at the indicated dates. For each isolate, the relative coverage across the 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 subject. In (**a**), isolates from Day 710 have different patterns of relative coverage across the MEDs because genomic libraries for these isolates were prepared differently (Method). Dark green diamonds indicate SNPs associated with sweeps and are labeled with the gene mutated and type of mutation. In (**g**), The SNP that was shared by all isolates from the latest time point (dark blue), yet polymorphic in isolates from the middle time point (pink), was not included as sweep, as it might be an artifact of undersampling (**Supplementary Fig. 9**). More details on the exact mutations and MEDs found are in **Supplementary Tables 7-18** and **Supplementary Table 3.**



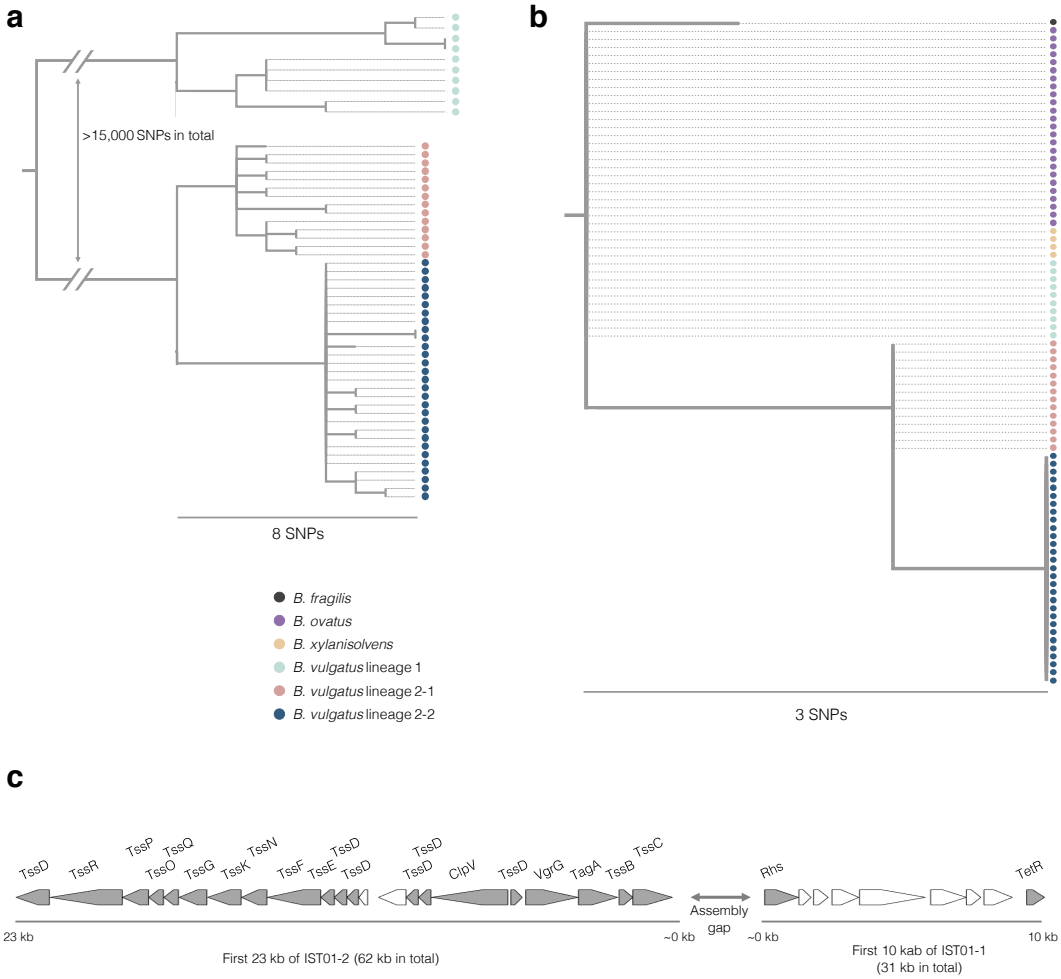

**Supplementary Figure 2 | Within-person *B. fragilis* evolution in subjects with longitudinal samples (continued from previous page).**



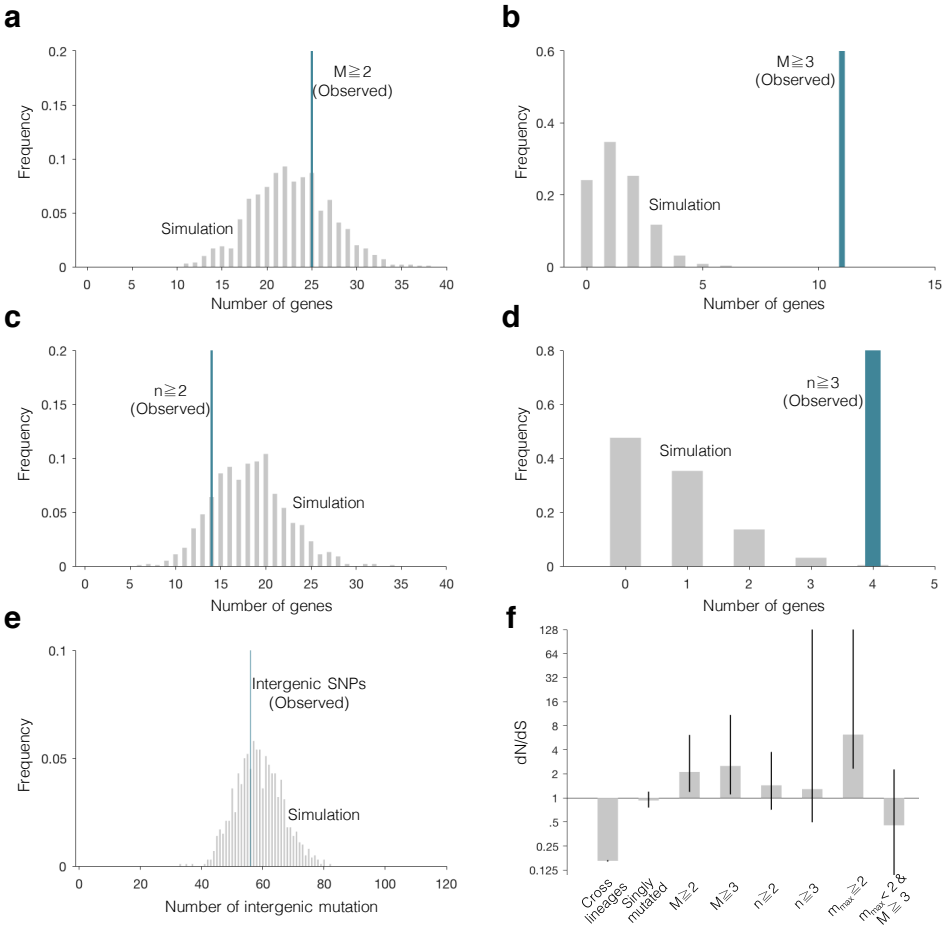

**Supplementary Figure 3 | Within-person *B. fragilis* evolution in subjects without longitudinal samples.** (**a-e**) The phylogeny for isolates from L08, L09, L10, L11 and L12, respectively. For each isolate, the relative coverage across the identified MEDs is shown. Shading of MED regions reflects the average relative coverage of the MED in that isolate. Red stars indicate when the same nucleotide mutation emerged multiple times within the same subject. (**d**) The presence/absence pattern of MED11-1 suggests many loss events on the phylogeny. (**f**) Notably, the relative coverage in the metagenome from the same sample is comparable to the relative coverage in individual isolates with the MED. This suggests that the MED may have been present in all cells and subsequently lost many times during or after stool collection from Subject 11. More details on the exact mutations and MEDs found are in **Supplementary Tables 7-18** and **Supplementary Table 3.**



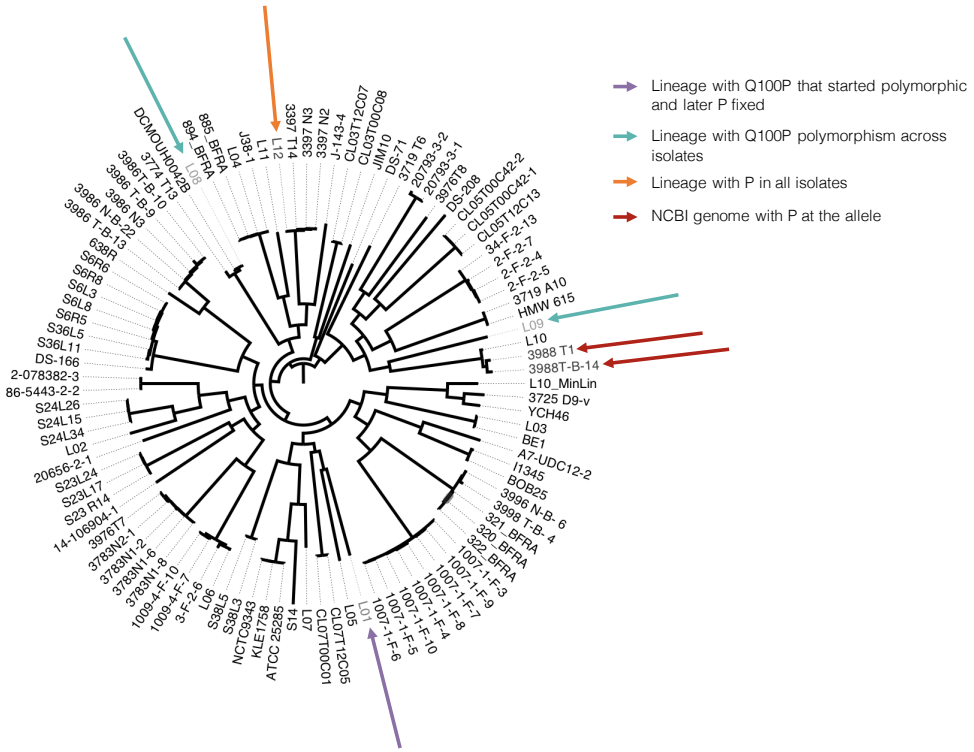

**Supplementary Figure 4 | Changes of tMRCA$_T$ over time.** (**a-g**) For each time point of each subject, we inferred the most recent common ancestor (MRCA) of just those isolates, and calculated tMRCA of the isolate population relative to that ancestor. In 2 subjects (panels **d, e**), tMRCA$_T$ decreased over time, associated with SNPs fixed in the same periods of time (**Supplementary Fig. 2d, e**). The decrease of tMRCA$_T$ in L07 (**g**) was possibly an artifact due to an undersampling of the last time point (**Supplementary Fig. 9g**) (**a**) Between time points 1 and 2 in L01, dMRCA$_T$ also decreased, but this decrease was due to changes in relative abundances of sublineages with different distances to the (same) MRCA (**Supplementary Fig. 2a**). (**f**) A sweep in L05 (**Supplementary Fig. 2f**) was not associated with a decrease in tMRCA$_T$, on account of the low initial value of tMRCA$_T$.



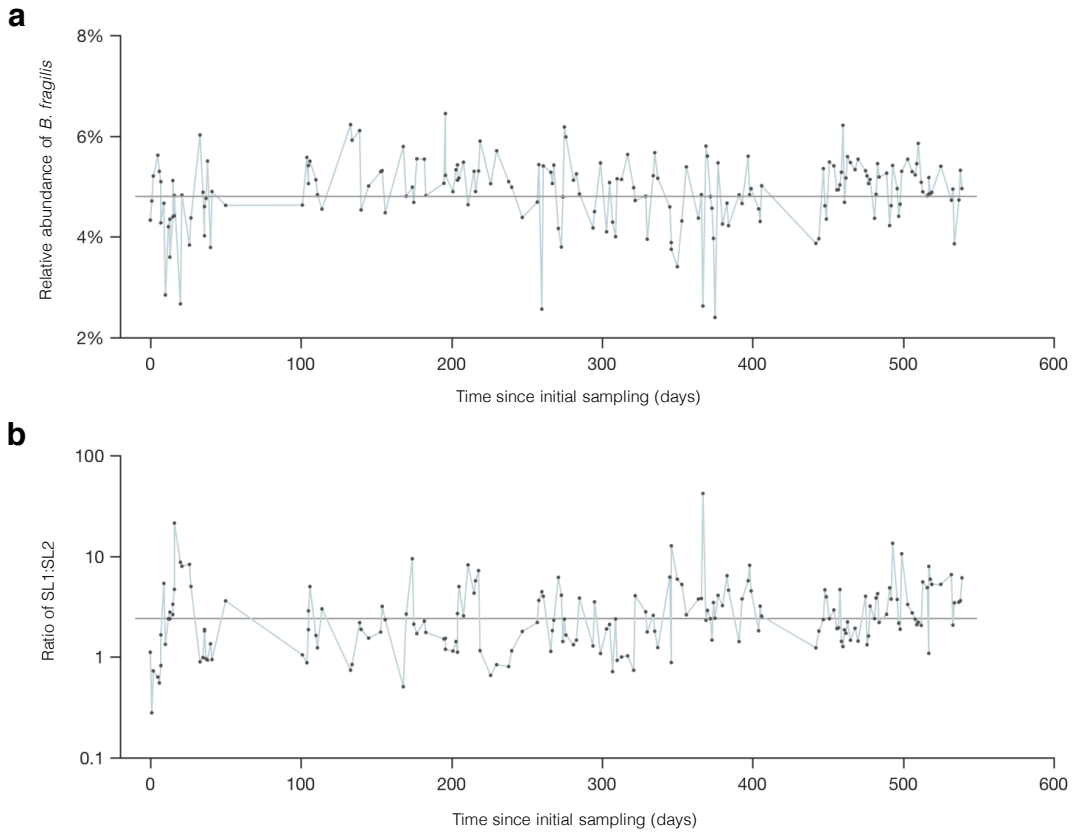

**Supplementary Figure 5 | Transfer of a putative integrative conjugative element with type 6 secretion system across *Bacteroides* species within Subject 01.** Analysis of the integrative conjugative element (ICE) found to be transferred in L01, identified from two candidate interspecies transfer regions (IST01-1 and IST-01-2, Methods). (**a**) 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 methods for *B. fragilis* SNP identification and evolutionary inference. We identified two *B. vulgatus* lineages that were separated by >15,000 SNPs. Within *B. vulgatus* lineage 2, we observed two sublineages. (**b**) A phylogeny was built using reads aligned to the ICE from all isolates of 4 *Bacteroides species* from Subject 01 (**Fig. 2d**). The sequences of IST-01 and IST-02 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 sequences 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 *B. fragilis* L01, each in a single isolate. **c**) ORF map of the type 6 secretion system of architecture 2 (T6SS-GA2) carried on this ICE. We aligned the ORFs from IST-01 and IST-02 to an annotated T6SS-GA2 from *Parabacteroides distasonis* CL03T12C09 (accession: JH976496.1). The first 10 kb of IST01-1 and the first 23 kb of IST01-2 had ORFs that are homologous to this T6SS-GA2. Grey pentagons represent conserved genes for T6SS-GA2[22].



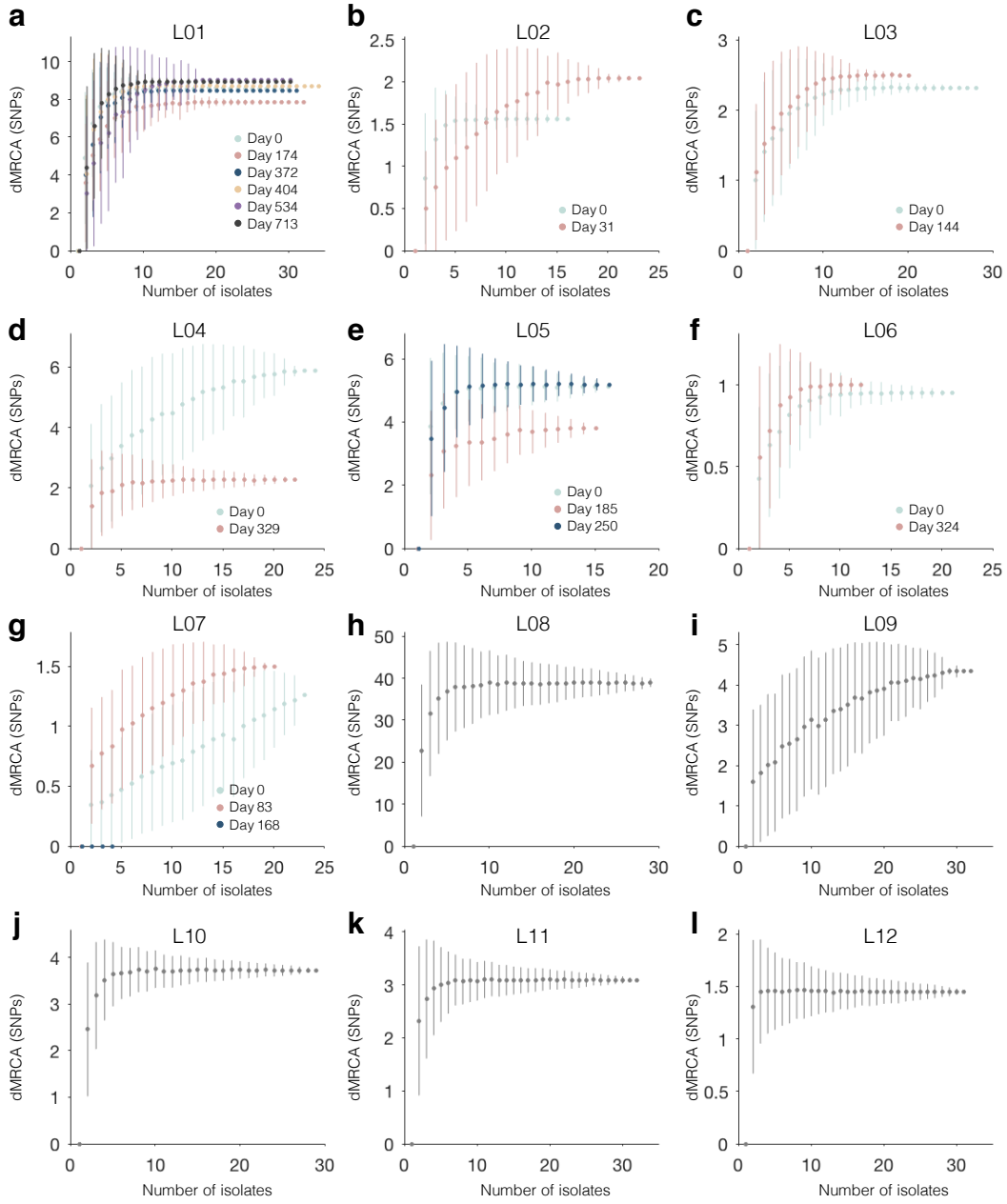

**Supplementary Figure 6 | Search for parallel evolution across lineages did not yield additional genes under selection.** We searched for genes mutated multiple times across lineages, counting the number of total SNPs obtained in each gene (M), the number of lineages a gene was mutated in (n), and the maximum number of mutation a given gene was mutated in any lineage ($m_{max}$). Simulations for all metrics were performed as described in the Methods. (**a**) A search with the criteria of M≧2 yielded results consistent with a null model. (**b**) When this threshold was increased to M≧3, 11 genes were observed. Interestingly, 9 of these genes were already discovered with the criteria used in the main text, $m_{max}$≧2. The 2 genes that are newly discovered with this metric ($m_{max}$<2 & M≧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 the entire group 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.



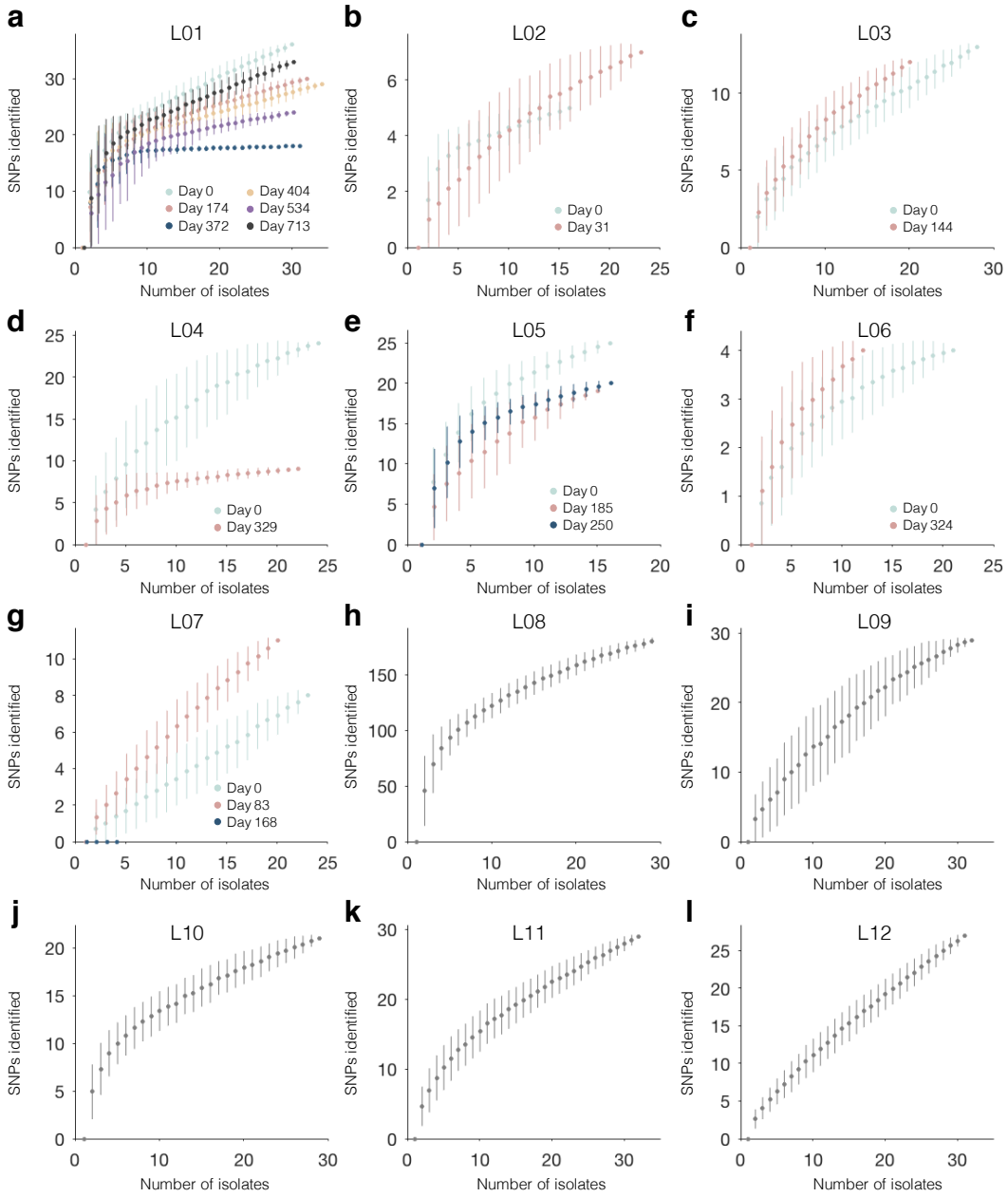

**Supplementary Figure 7 | Parallel evolution of Q100P in BF2755 across and within lineages.** For the 12 major lineages we investigated, two lineages had both isolates with a glutamine (Q) and isolates with a proline (P) at position 100 in the BF2755 protein (L08, L09). L01 started with two distinct Q100P mutations (**Supplementary Fig. 2a**) and later on the mutant genotype (P) fixed in the population. All isolates from the L12 lineage had a P at this position, suggesting it had fixed prior to or during colonization. The remaining 8 lineages did not have a mutation at this position. For 88 publicly available genomes, we blasted their genomes to the DNA sequence of BF2755, and examined the position of this mutation. Two lineages, 3988 T1 and 3988-B-14, encode for a P at this position.



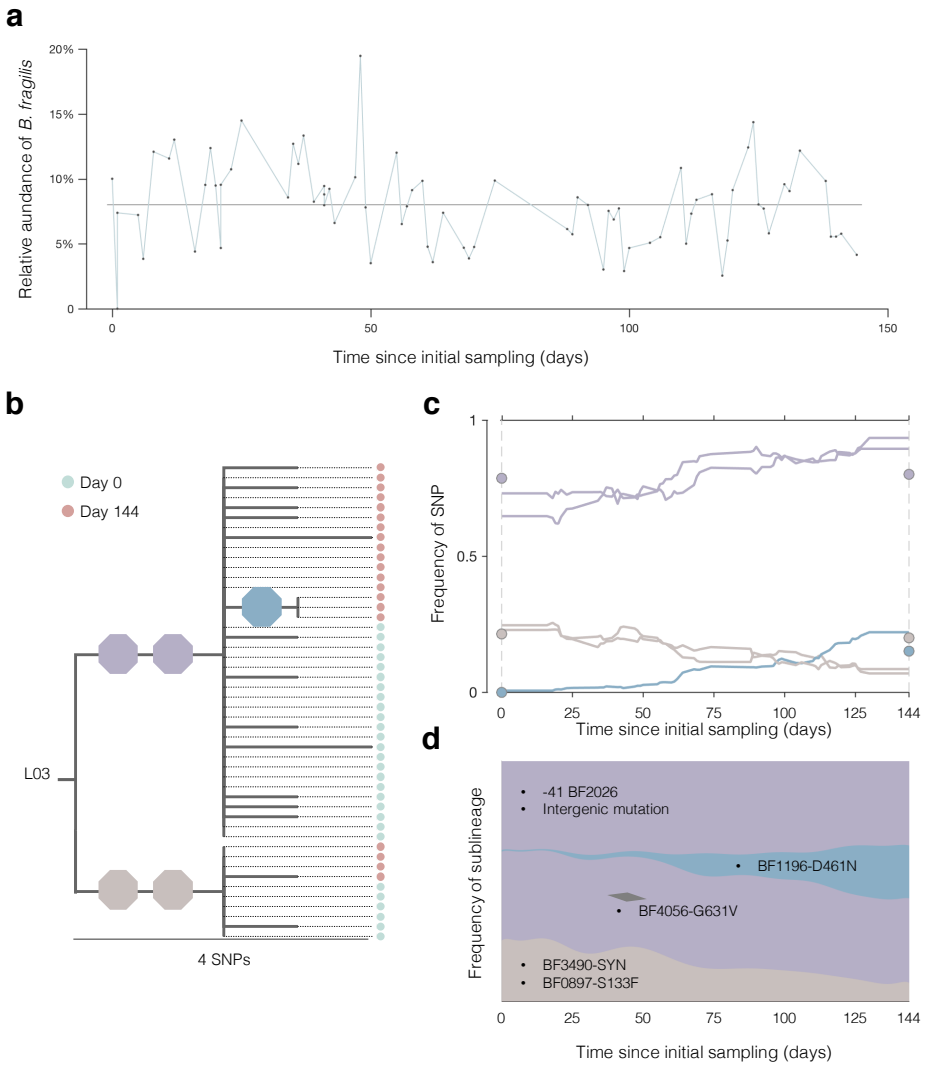

**Supplementary Figure 8 | Within Subject 01, relative abundance of *B. fragilis* and the ratio of SL1:SL2 were stable over time.** (**a**) For each metagenome from stool samples from Subject 01 (**Supplementary Table 6**), we calculated the percentage of metagenomic reads that aligned to the L01 genome 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 gray 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. We only plotted samples with more than 40 reads aligned to these SNP locations. The gray line indicates the mean across samples.




**Supplementary Figure 9 | Collector curves suggest sufficient sampling for $dMRCA_T$.** (**a-i**) For each lineage and time point, we created a collector curve for $dMRCA_T$ (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 $dMRCA_T$. For each x, we simulated 100 subsamples and computed the mean (dots) and standard deviation (bars) for the simulation results. $dMRCA_T$ were undersaturated only in 2 time points of L07 (0 and 168 Days).




**Supplementary Figure 10 | Number of SNPs identified depends on number of isolates collected.** (**a-i**) 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.




**Supplementary Figure 11 | Evolutionary dynamics of L03.** (**a**) The relative abundance of L03 *B. fragilis* inside Subject 03 was estimated in 75 metagenomes spanning 144 days, using the same method described in **Supplementary Fig. 8a**. (**b**) The phylogeny of isolates from L03. Branches with ≧3 isolates are labeled with colored octagons that represent individual SNPs. Circles represent individual isolates, and are colored according to sampling date. (**c**) Frequencies of labeled SNPs over time in the *B. fragilis* population were inferred from 75 stool metagenomes (Methods). Colored circles represent SNP frequencies inferred from isolate genomes at particular time points. (**d**) 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 (**b**). The black diamond represents a transient SNP from a gene that was identified as under positive selection (**Fig. 3c**).

# Supplementary Table 5: Genes under selection *in vivo*

| Gene locus* | Prokka annotation | Predicted biological role | Annotation in Figure 3 | Annotated homologs [organism] | Mutated lineage [locations] | Cellular Localization** | Notes |
|---|---|---|---|---|---|---|---|
| **BF0864** | TonB-dependent receptor SusC | Polysaccharide import/binding | SusC family protein | | L08: [V283M, E404D] | Outer Membrane | |
| **BF0893** | TonB-dependent receptor SusC | Polysaccharide import/binding | SusC family protein | | L08: [A702S, S189*] | Outer Membrane | |
| **BF1802** | SusD-like protein | Polysaccharide import/binding | SusD family protein | | L01: [G312A, T340M, D526N] | Outer Membrane | Upregulated in mice treated with human milk oligosaccharides[50] |
| **BF1803** | TonB-dependent receptor SusC | Polysaccharide import/binding | SusC family protein | | L02: [N293K, D572N] | Outer Membrane | Upregulated in mice treated with human milk oligosaccharides[50] |
| **BF2942** | TonB-dependent receptor SusC | Polysaccharide import/binding | SusC family protein | | L01: [E769K, S] | Outer Membrane | |
| **BF3581** | TonB-dependent receptor SusC | Polysaccharide import/binding | SusC family protein (ccfC) | **ccfC** [*Bacteroides fragilis*] | L01: [P240L, K751R, Q974R] | Outer Membrane | Shown to be important for colonization in mouse models[4] |
| **BF4056** | TonB-dependent receptor SusC | Polysaccharide import/binding | SusC family protein | | L03: [G631V]; L04: [G631F, G631F] | Outer Membrane | |
| **BF0188** | Lipid A export ATP-binding/permease protein MsbA | Cell envelope biosynthesis | ABC transporter msbA | **msbA** [*Bacteroides salyersiae*] | L01: [D61N, K485Q] | 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] | Periplasmic | The ortholog in *B. thetaiotaomicron* (BT1355) is in the capsule polysaccharide 4 locus, shown to be important for binding IgA[51]. |
| **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] | Inner Membrane | Deletion of this gene abrogates synthesis of 7 of the 8 capsular polysaccharides[52]. |
| **CL3580** | Hypothetical protein | Cell envelope biosynthesis | Glycosyltransferase (wffD) | **wffD** [*Escherichia coli*] | L11: [F52L]; [P165S] | Outer Membrane | |
| **CL4395** | Putative glycosyltransferase EpsH | Cell envelope biosynthesis | Glycosyltransferase | **cpsI** [*Prevotella sp.* oral taxon 299] | L10: [W156C, W156C, W156C] | 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] | Periplasmic | |
| **BF1174** | HTH-type transcriptional regulator CysL | Unknown gene regulation | Transcriptional regulator | **cysL** [*Bacillus subtilis subsp. subtilis str. 168*] | L01: [K62*, L165S]; L07: [Q162*] | 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] | 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] | Cytoplasmic | Also called csbB. Homologs have been implicated in cysteine metabolism[53] and serine metabolism[54], with connections to virulence. |
| **CL5037** | Hypothetical protein | Unknown | Hypothetical protein | | L08: [E53*, R228I] | 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**)
**Predicted by CELLO (Methods)

## Supplementary Tables

There are 18 Supplementary Tables uploaded in .xlsx format in a single zip file. Supplementary Table 5 is above in PDF format.

Supplementary Table 1: Subject information and per-lineage statistics
Supplementary Table 2: Stool samples used for culturing single-colony isolates
Supplementary Table 3: Mobile element difference (MED) information
Supplementary Table 4: Candidate inter-species transfers
Supplementary Table 6: Stool samples used for metagenomic sequencing and alignment results
Supplementary Table 7: *de novo* SNPs within L01
Supplementary Table 8: *de novo* SNPs within L02
Supplementary Table 9: *de novo* SNPs within L03
Supplementary Table 10: *de novo* SNPs within L04
Supplementary Table 11: *de novo* SNPs within L05
Supplementary Table 12: *de novo* SNPs within L06
Supplementary Table 13: *de novo* SNPs within L07
Supplementary Table 14: *de novo* SNPs within L08
Supplementary Table 15: *de novo* SNPs within L09
Supplementary Table 16: *de novo* SNPs within L10
Supplementary Table 17: *de novo* SNPs within L11
Supplementary Table 18: *de novo* SNPs within L12
Supplementary Table 19: Clustering of gene homologs from different lineages

## Supplementary references:


48. Yamada, K. D., Tomii, K. & Katoh, K. Application of the MAFFT sequence alignment program to large data - Reexamination of the usefulness of chained guide trees. *Bioinformatics* **32,** 3246–3251 (2016).
49. Stamatakis, A. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).
50. Marcobal, A. *et al. Bacteroides* in the infant gut consume milk oligosaccharides via mucus-utilization pathways. *Cell Host Microbe* **10,** 507–514 (2011).
51. 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).
52. 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).
53. 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).
54. Connolly, J. P. R. *et al.* A Highly Conserved Bacterial D-Serine Uptake System Links Host Metabolism and Virulence. *PLoS Pathog.* **12,** e1005359 (2016).