Cell Host & Microbe, Volume *25*

Supplemental Information

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

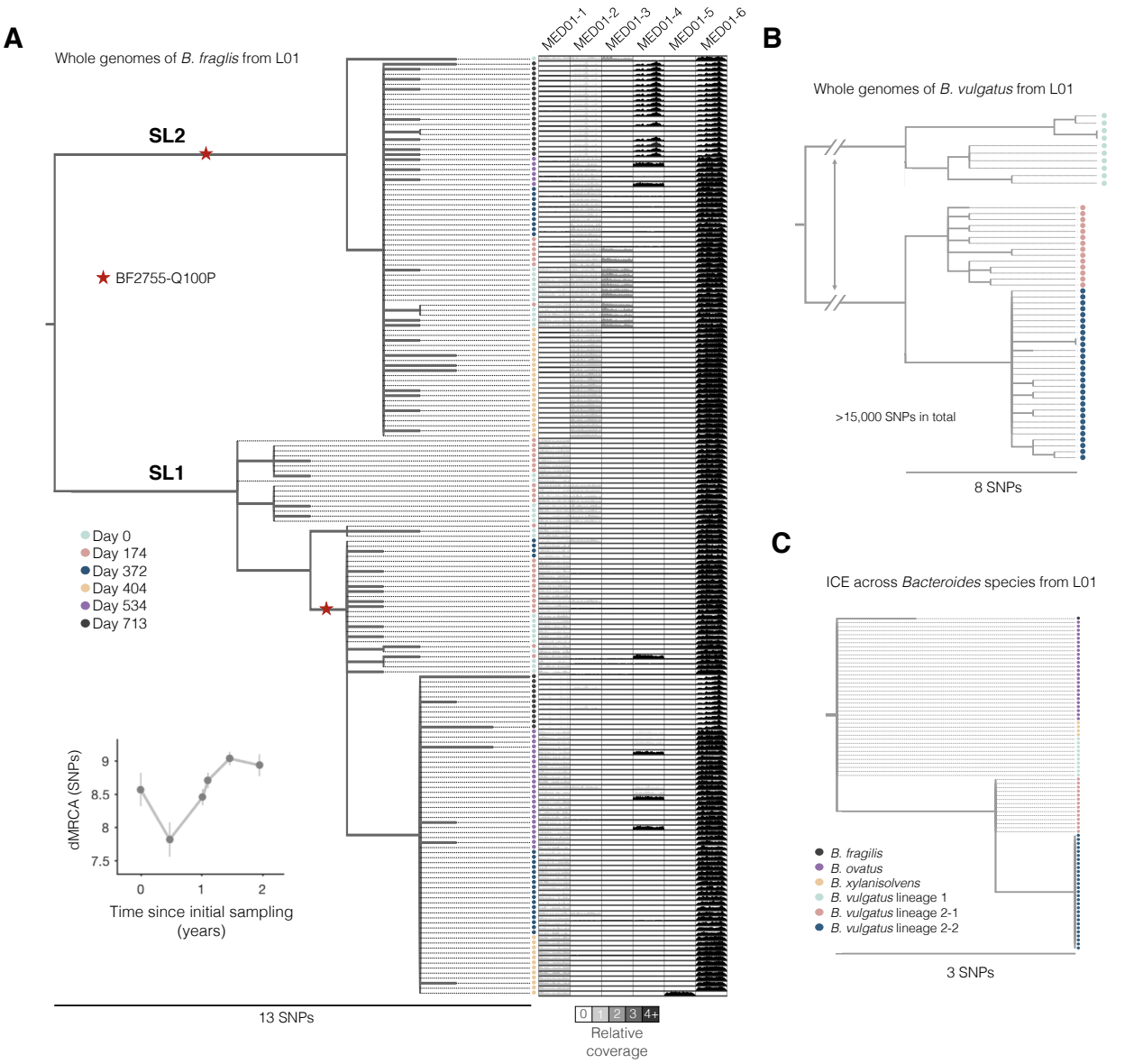

Figure S1 | Within-person evolution of *Bacteroides* from Subject 01, Related to Figure 2

(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 S7** and **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.

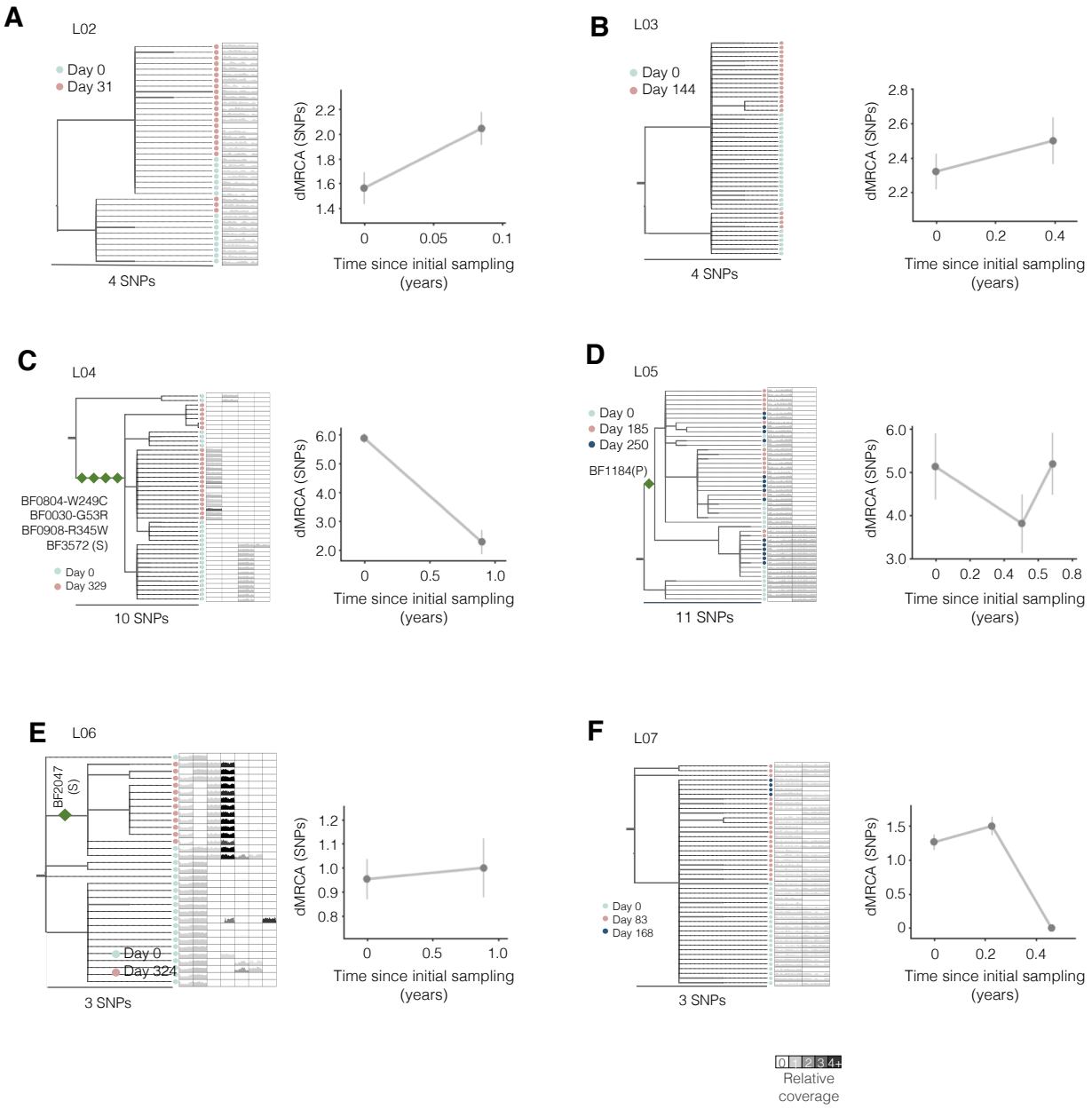

Figure S2 | Within-person *B. fragilis* evolution in L02-L07, Related to Figure 2

(**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 S7** and **Table S3.**

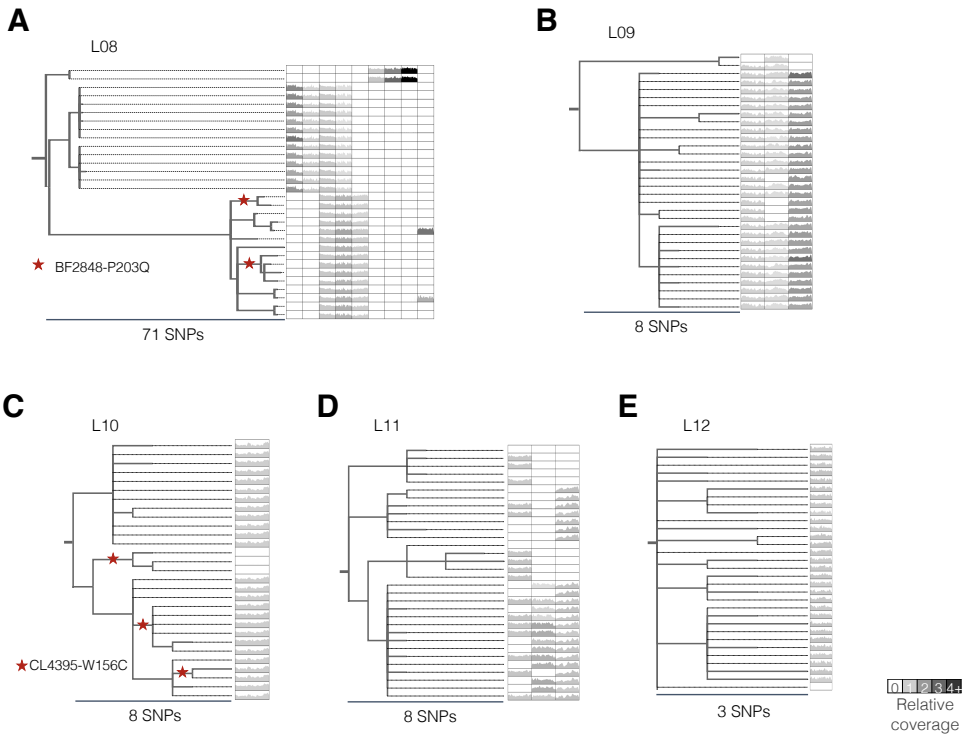

Figure S3 | Within-person *B. fragilis* evolution in L08-L12, Related to Figure 2

(**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**.

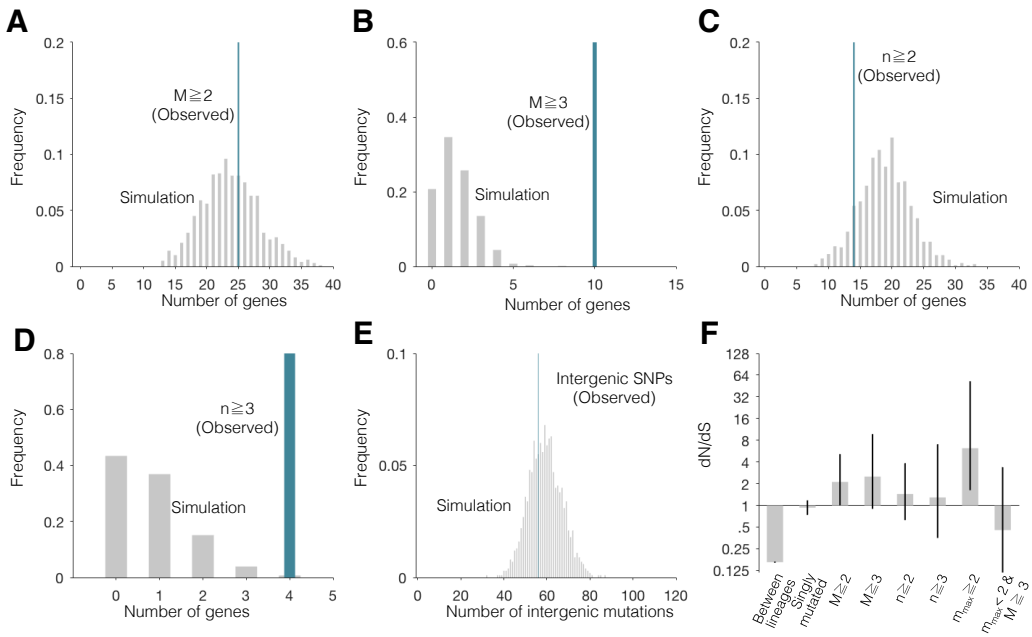

Figure S4 | Search for parallel evolution across lineages did not yield additional genes under selection, Related to Figure 4

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≧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 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.

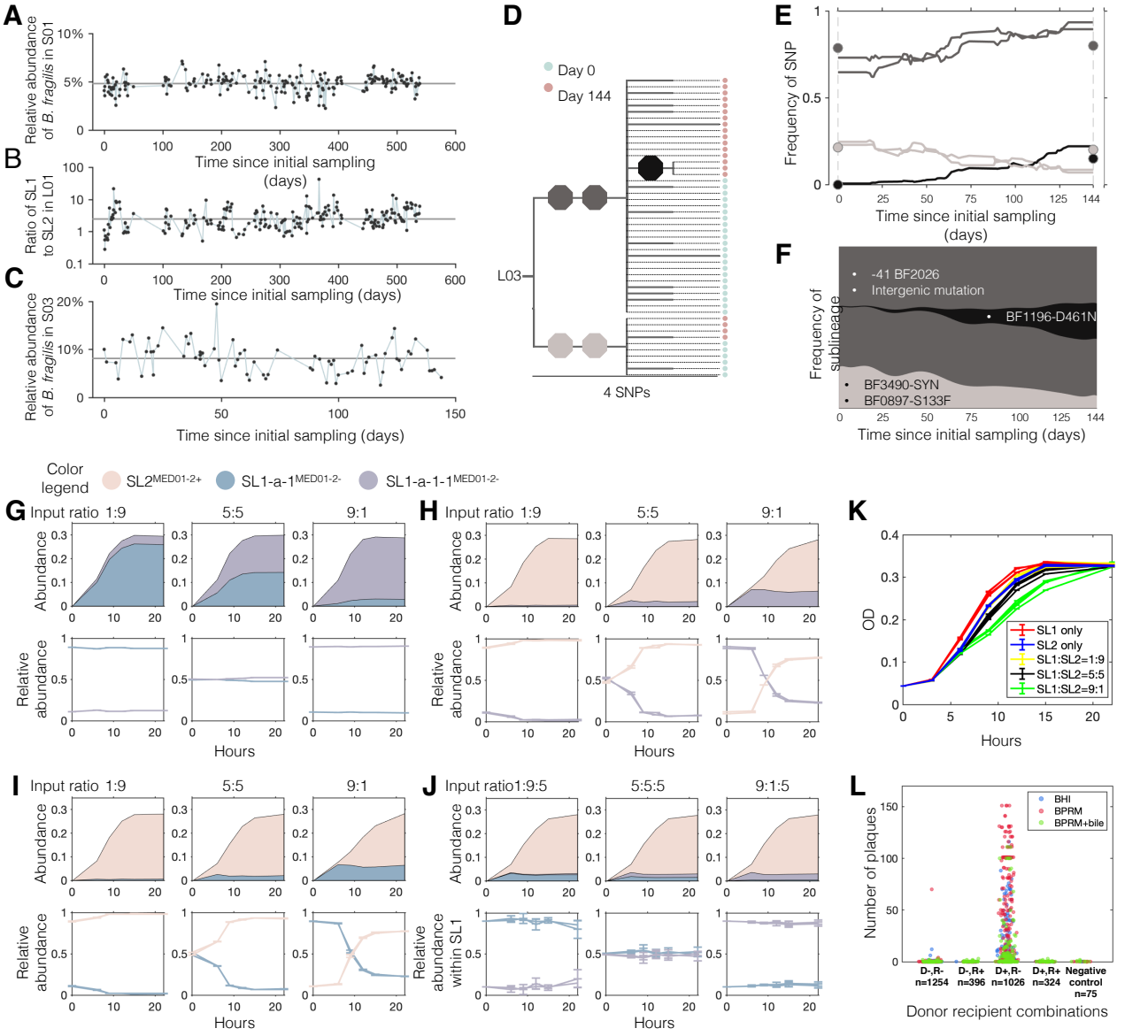

Figure S5 | Evolutionary dynamics of L01 and L03 and the phage-mediated competition between L01 sublineages; Related to Figure 5

(A) For each metagenome from stool samples from Subject 01 (**Table S8**), 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).

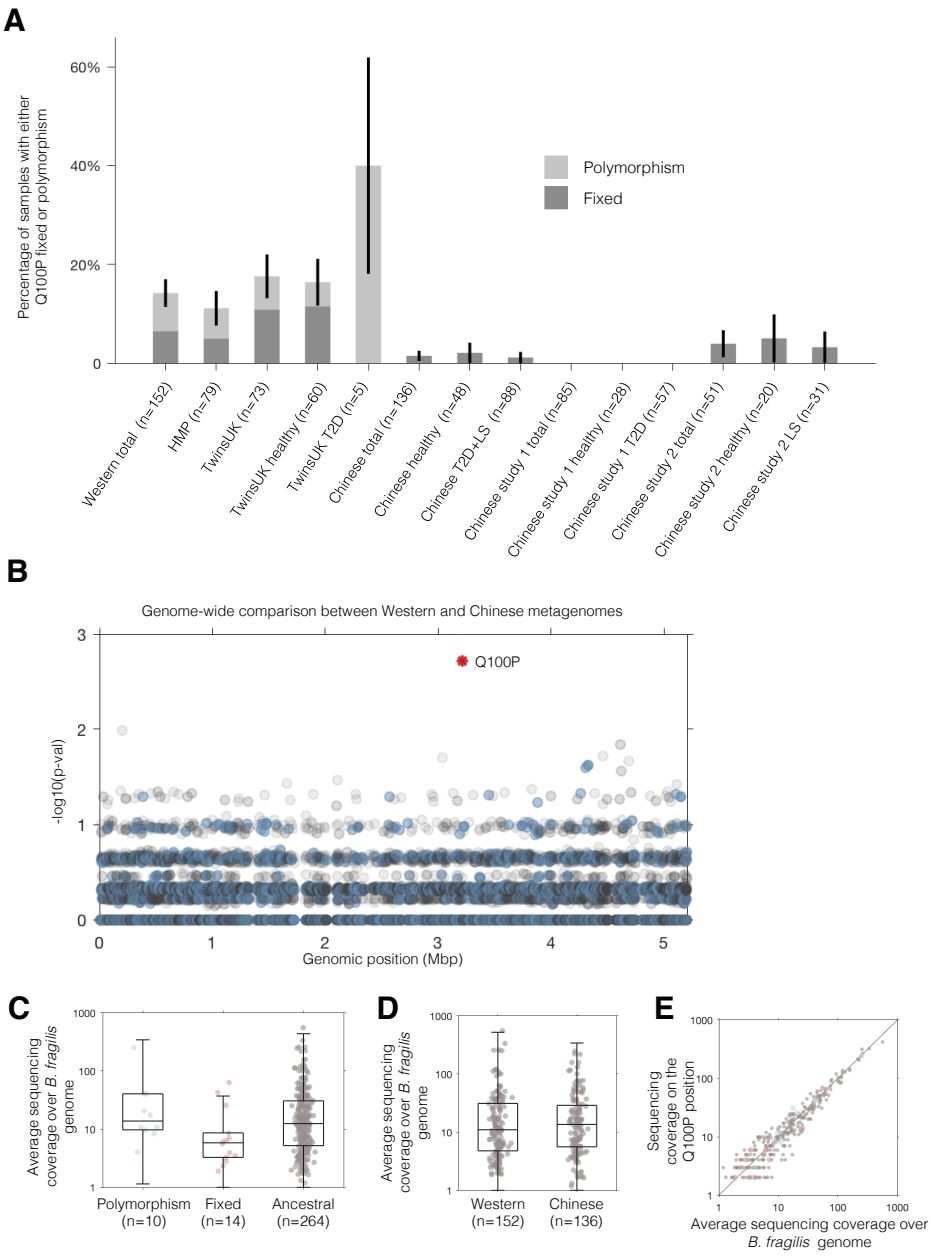

Figure S6 | BF755 Q100P difference between Chinese and Western populations are robust to subject health conditions and are the most significant difference; Related to Figure 6

(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**).

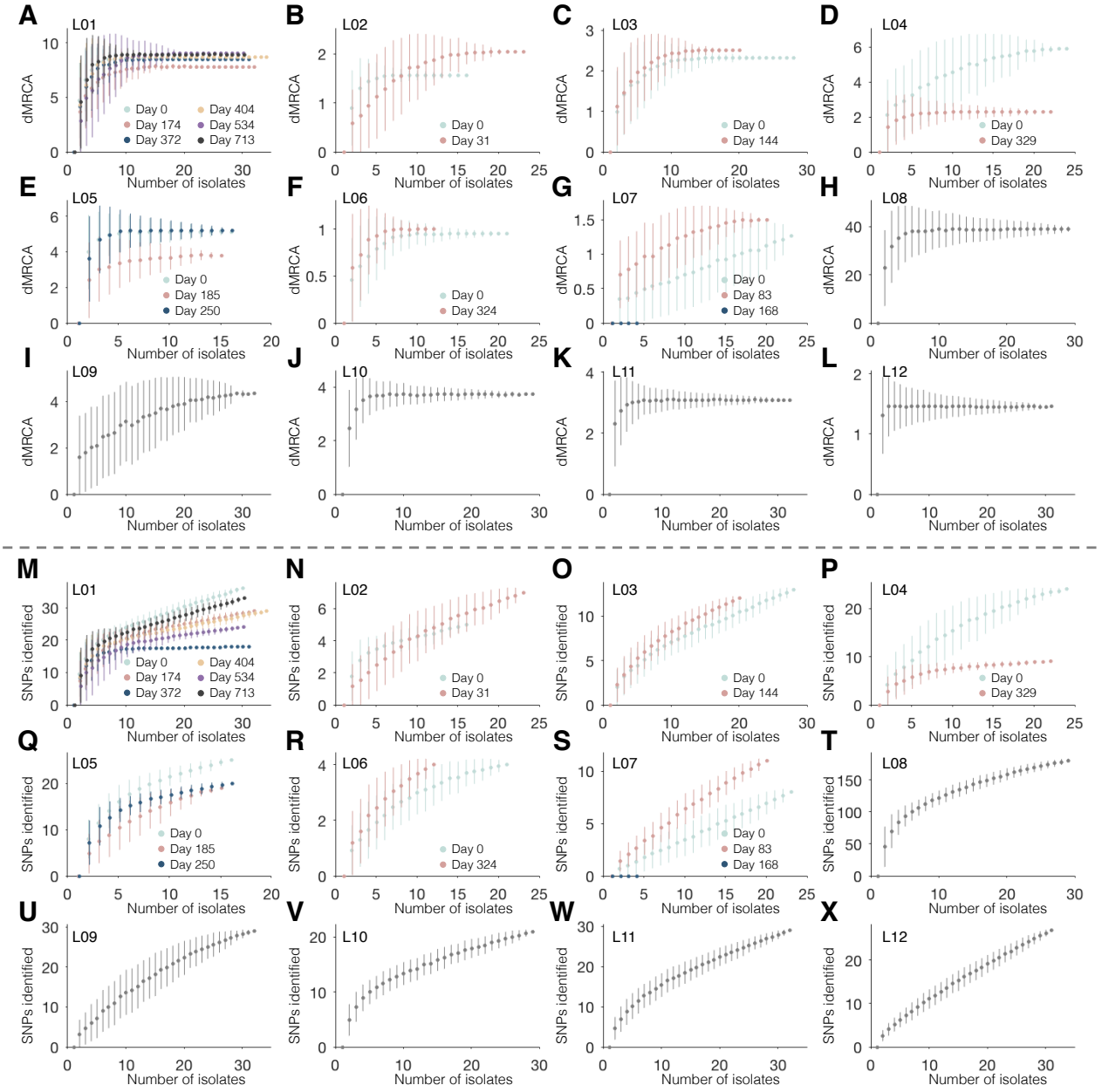

Figure S7 | Collector curves suggest sufficient sampling for dMRCA, yet numbers of SNPs identified depends on number of isolates collected, Related to STAR Methods

(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.