Supplemental Materials

Contents:

Computational Methods and Analysis

List of Figures:

List of Tables:

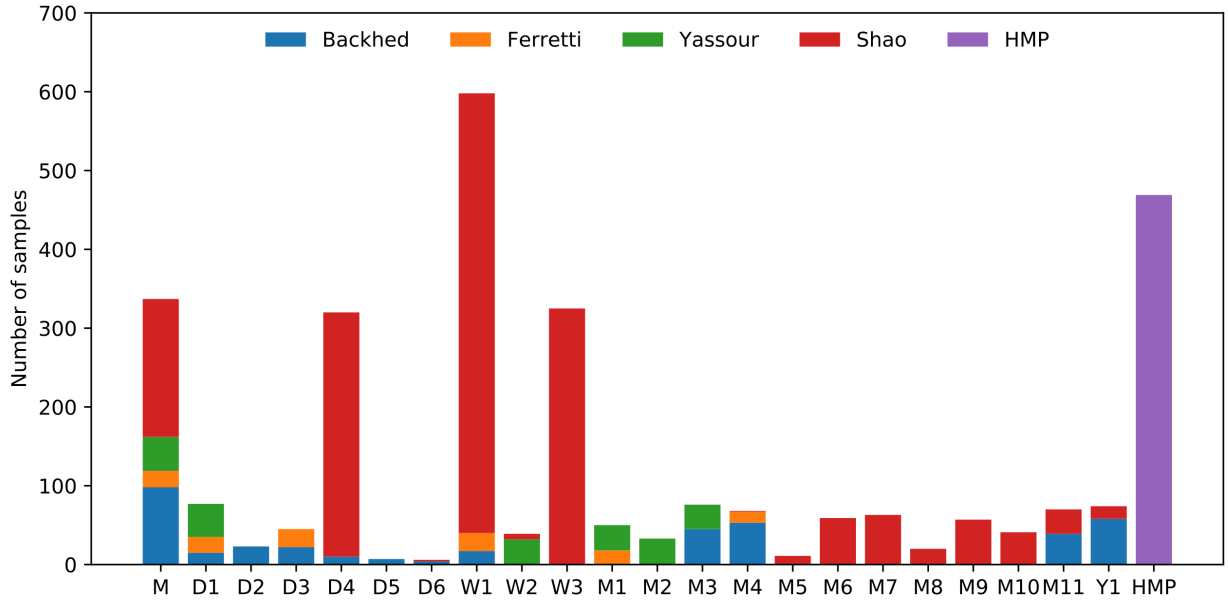

Figure S1: Sampling timepoints and sample sizes for infants and mothers analyzed in this paper. M indicates mother. Only those mothers sampled around the time of delivery were included in this analysis. "D1" indicates day 1, "W1" indicates week 1, "M1" indicates month 1, "Y1" indicates year 1, and so on. "M" indicates mother samples at delivery and "HMP" includes all HMP1-2 adult samples. Some infants from Shao et al. 2019 were sampled multiple times between day 7 and 14, and thus were grouped in the Week 1 category.

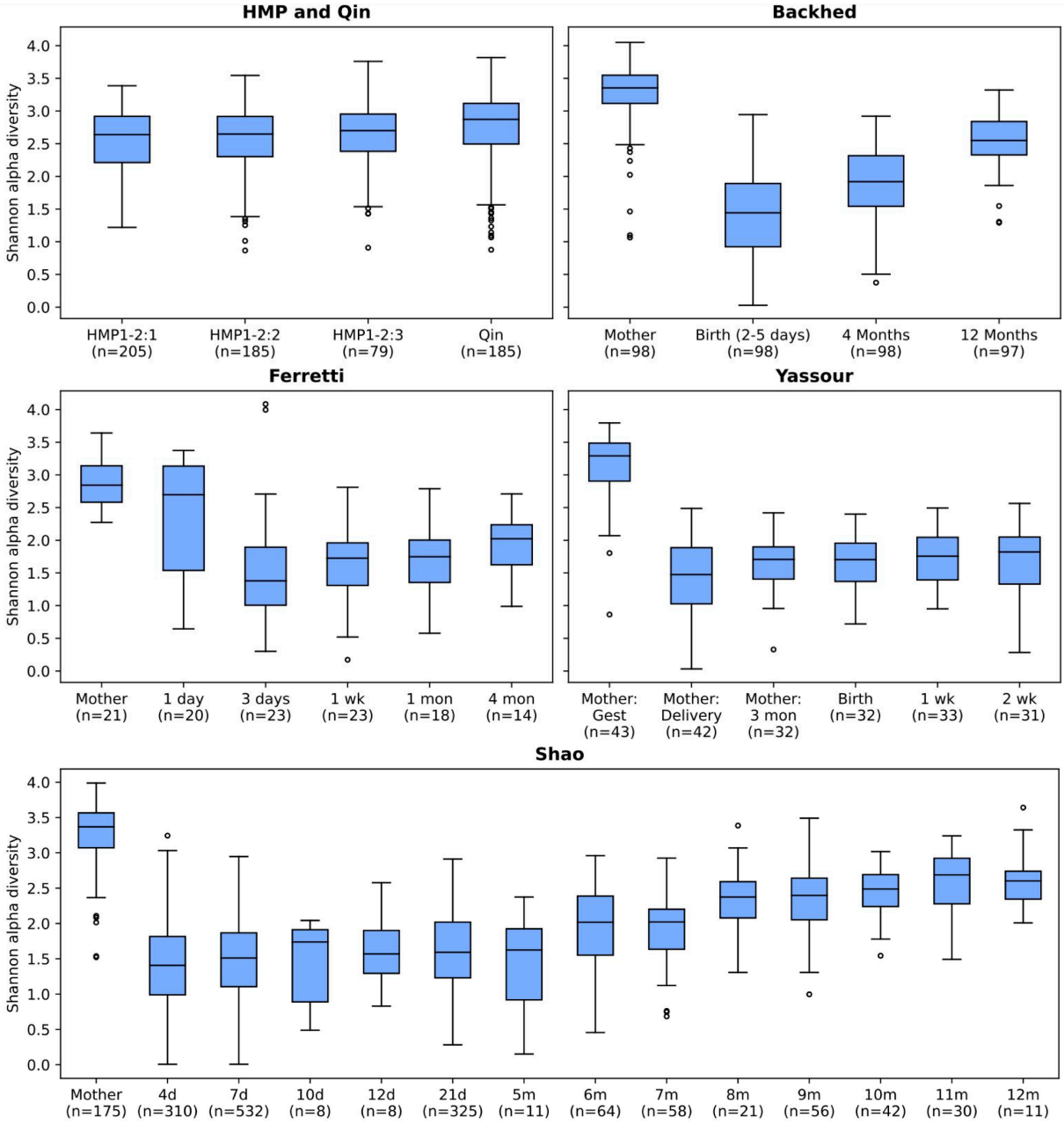

Figure S2: Shannon alpha diversity displayed by dataset. Samples sizes (n) are shown in parentheses. Overall, alpha diversity increases as a function of age of the host (*p* value = 2*10^-16, glmmTMB(alpha_div ~ day + (1|sample) + (1|subject) + (1|dataset)), using all samples from infants).

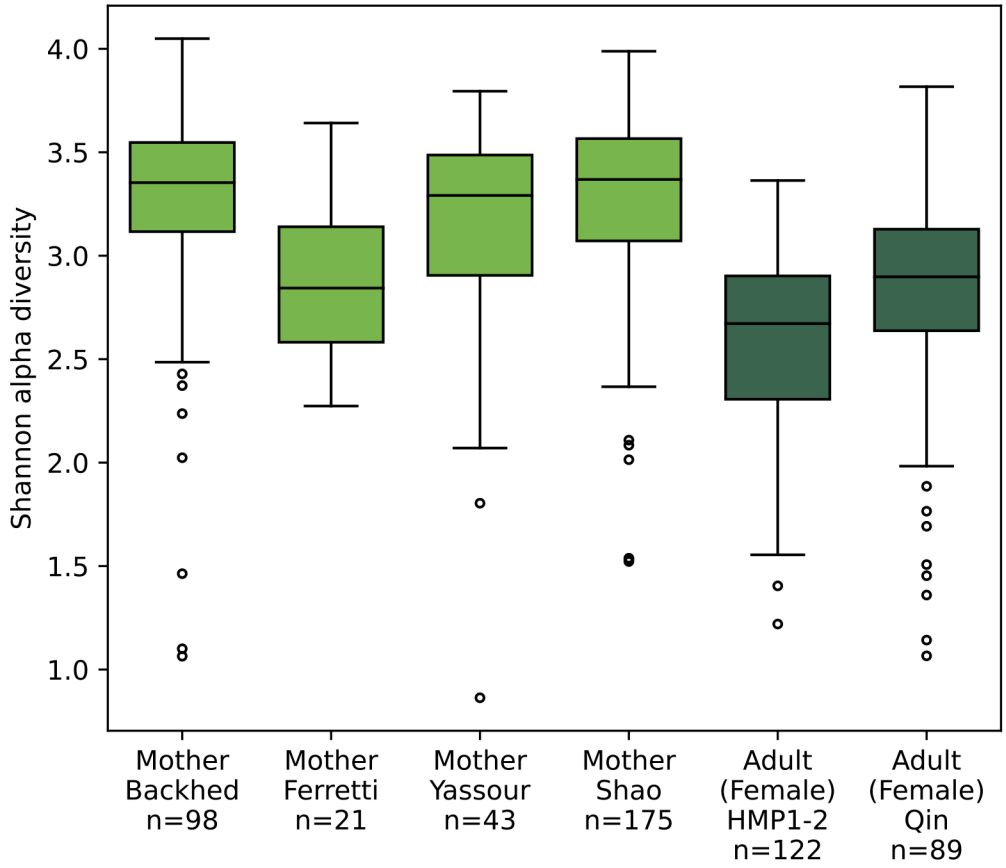

Figure S3: Shannon alpha diversity in mothers' gut microbiomes at time of delivery versus healthy non-pregnant women. Plotted are alpha diversity values for mothers at time of delivery in the Backhed, Ferretti, Yassour, and Shao datasets, and non-pregnant women in the HMP and Qin et al. datasets. To assess whether pregancy results in a significant difference in alpha diversity, we fit a GLMM with pregnancy status as the predictor and study and host as random effects (alpha diversity ~ pregnancy status + (1|study) + (1|host)) and obtained a coefficient of 0.44 for pregnancy status with a *p* value of 0.035.

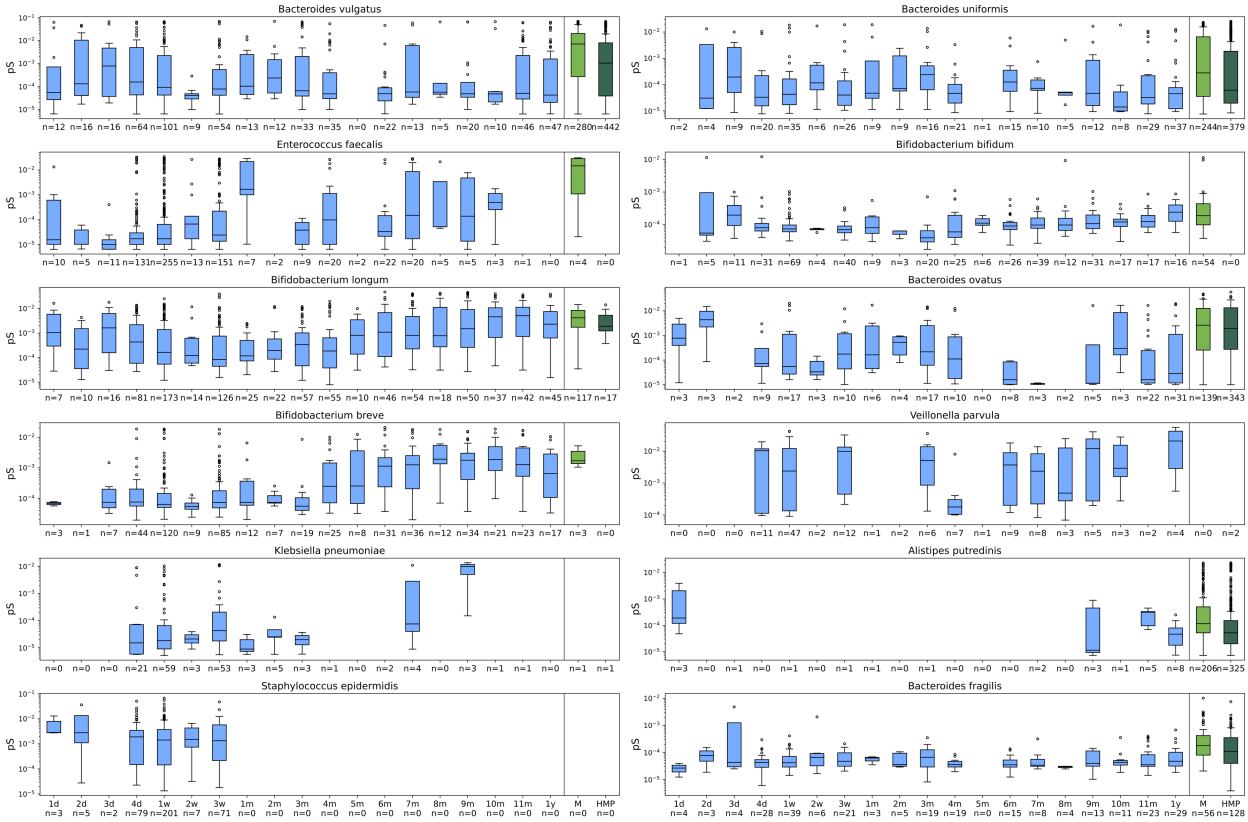

Figure S4: Within-sample polymorphism over different life stages for the most prevalent species in our dataset. The next 12 most prevalent species, after *E. coli*, which is shown in **Figure 1B**, are displayed here. The same life stages as in **Figure 1** are shown, where "B" indicates birth, "D1" indicates day 1, "W1" indicates week 1, "M1" indicates month 1, "Y1" indicates year 1, and so on. "M" indicates mother samples at delivery and "HMP" includes all HMP1-2 adult samples. pS indicates synonymous polymorphism rate, where a polymorphism was defined as a site with allele frequency between 0.2 and 0.8.

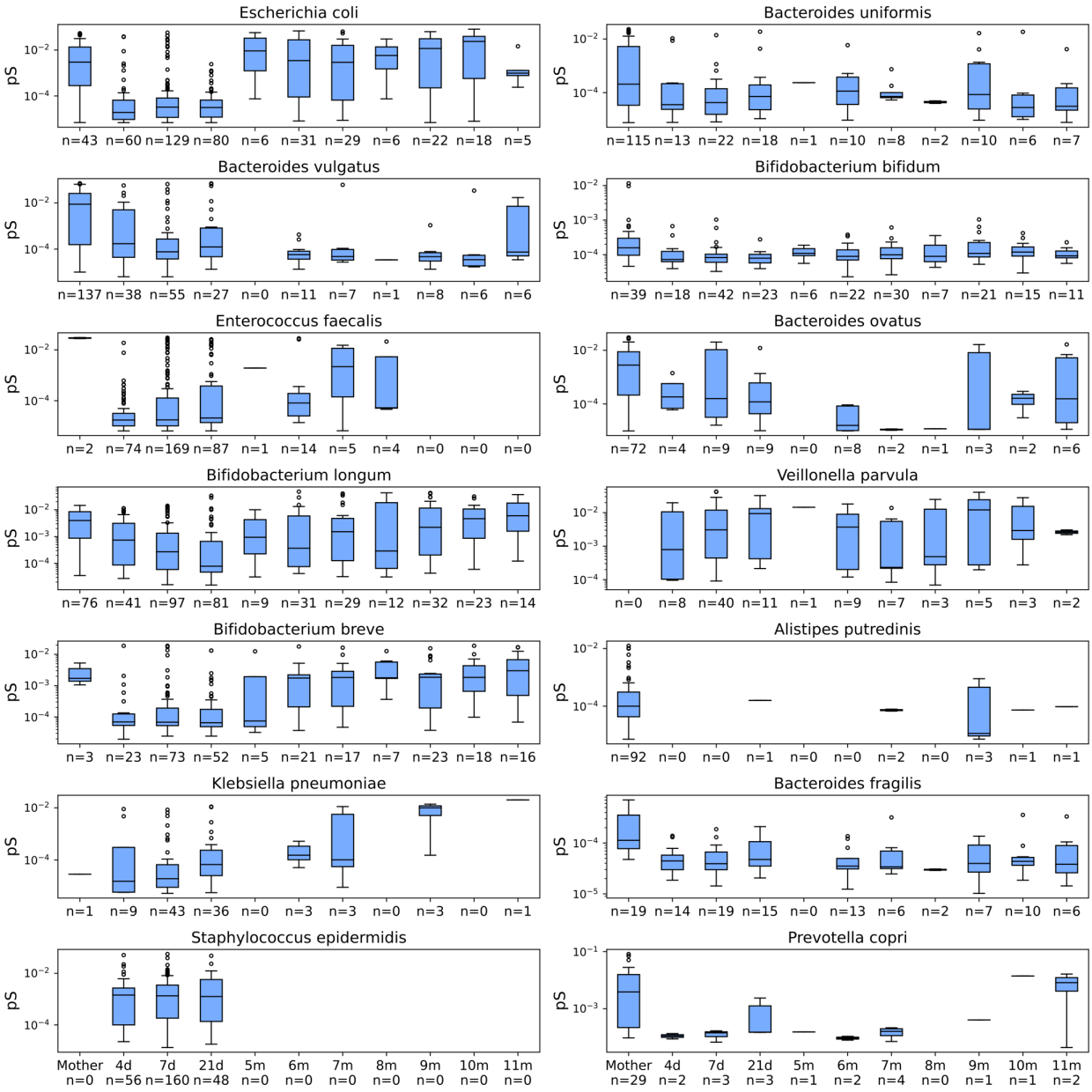

Figure S5: Within-sample polymorphism in the Shao et al. data set. To assess if polymorphism rates generally follow the same trends within a dataset versus across datasets, we we examine polymorphism rates for the 14 most prevalent species in our dataset using data from Shao et al only.

Polymorphism as a function of age across datasets

To assess the effect of dataset on polymorphism, we compared distributions of polymorphism per sample-species pair for matched age categories across different datasets for the 13 most prevalent species (in infants). Up to 12 pairwise comparisons for each species were considered for the following age categories: 1 day (Backhed and Ferretti), 3 days (Backhed and Ferretti), 4 days (Backhed and Yassour), 1 week (Ferretti, Yassour and Shao), 2 weeks (Yassour and Shao), 1 month (Ferretti, Yassour and Shao), 4 months (Backhed and Ferretti), and 12 months (Backhed and Shao). Due to non-normal polymorphism distributions, we used the Mann Whitney U statistic to test for significant differences in polymorphism between datasets. Of 18 tests across the 13 species satisfying a minimum sample size requirement of 10, one test was significant (Backhed vs. Ferretti at 12 months, P=0.00018) using a Bonferroni-corrected significance level $\alpha=0.05/18$. Of 45 tests satisfying a laxer sample size threshold of 5, only the aforementioned comparison remained significant. All test results are reported in **Table S1**.

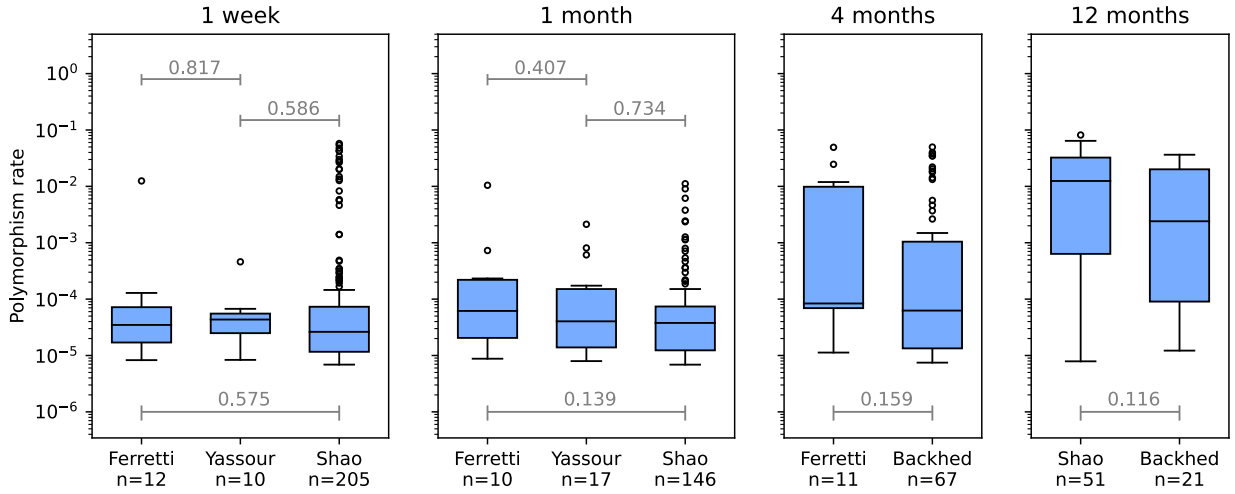

Figure S6: Between-dataset comparisons of *E. coli* polymorphism levels for matched life stages. *P* values from Mann-Whitney *U* tests are reported in grey.

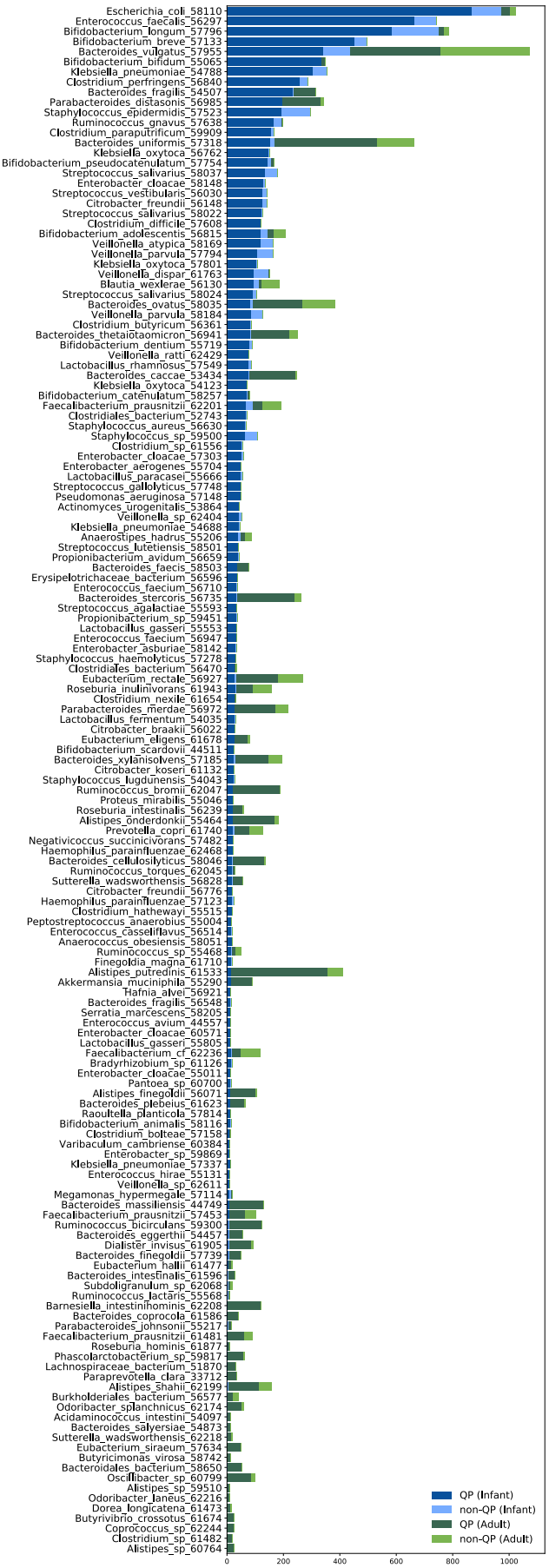
Escherichia_coli_58110
Enterococcus_faecalis_56297
Bifidobacterium_longum_57796
Bifidobacterium_breve_57133
Bacteroides_vulgatus_57955
Bifidobacterium_bifidum_55065
Klebsiella_pneumoniae_54788
Clostridium_perfringens_56840
Bacteroides_fragilis_54507
Parabacteroides_distasonis_56985
Staphylococcus_epidermidis_57523
Ruminococcus_gnavus_57638
Clostridium_paraputrificum_59909
Bacteroides_uniformis_57318
Klebsiella_oxytoca_56762
Bifidobacterium_pseudocatenulatum_57754
Streptococcus_salivarius_58037
Enterobacter_cloacae_58148
Streptococcus_vestibularis_56030
Citrobacter_freundii_56148
Streptococcus_salivarius_58022
Clostridium_difficile_57608
Bifidobacterium_adolescentis_56815
Veillonella_atypica_58169
Veillonella_parvula_57794
Klebsiella_oxytoca_57801
Veillonella_dispar_61763
Blautia_wexlerae_56130
Streptococcus_salivarius_58024
Bacteroides_ovatus_58035
Veillonella_parvula_58184
Clostridium_butyricum_56361
Bacteroides_thetaiotaomicron_56941
Bifidobacterium_dentium_55719
Veillonella_ratti_62429
Lactobacillus_rhamnosus_57549
Bacteroides_caccae_53434
Klebsiella_oxytoca_54123
Bifidobacterium_catenulatum_58257
Faecalibacterium_prausnitzii_62201
Clostridiales_bacterium_52743
Staphylococcus_aureus_56630
Staphylococcus_sp_59500
Clostridium_sp_61556
Enterobacter_cloacae_57303
Enterobacter_aerogenes_55704
Lactobacillus_paracasei_55666
Streptococcus_gallolyticus_57748
Pseudomonas_aeruginosa_57148
Actinomyces_urogenitalis_53864
Veillonella_sp_62404
Klebsiella_pneumoniae_54688
Anaerostipes_hadrus_55206
Streptococcus_lutetiensis_58501
Propionibacterium_avidum_56659
Bacteroides_faecis_58503
Erysipelotrichaceae_bacterium_56596
Enterococcus_faecium_56710
Bacteroides_stercoris_56735
Streptococcus_agalactiae_55593
Propionibacterium_sp_59451
Lactobacillus_gasseri_55553
Enterococcus_faecium_56947
Enterobacter_asburiae_58142
Staphylococcus_haemolyticus_57278
Clostridiales_bacterium_56470
Eubacterium_rectale_56927
Roseburia_inulinivorans_61943
Clostridium_nexile_61654
Parabacteroides_merdae_56972
Lactobacillus_fermentum_54035
Citrobacter_braakii_56022
Eubacterium_eligens_61678
Bifidobacterium_scardovii_44511
Bacteroides_xylanisolvens_57185
Citrobacter_koseri_61132
Staphylococcus_lugdunensis_54043
Ruminococcus_bromii_62047
Proteus_mirabilis_55046
Roseburia_intestinalis_56239
Alistipes_onderdonkii_55464
Prevotella_copri_61740
Negativicoccus_succinicivorans_57482
Haemophilus_parainfluenzae_62468
Bacteroides_cellulosilyticus_58046
Ruminococcus_torques_62045
Sutterella_wadsworthensis_56828
Citrobacter_freundii_56776
Haemophilus_parainfluenzae_57123
Clostridium_hathewayi_55515
Peptostreptococcus_anaerobius_55004
Enterococcus_casseliflavus_56514
Anaerococcus_obesiensis_58051
Ruminococcus_sp_55468
Finegoldia_magna_61710
Alistipes_putredinis_61533
Akkermansia_muciniphila_55290
Hafnia_alvei_56921
Bacteroides_fragilis_56548
Serratia_marcescens_58205
Enterococcus_avium_44557
Enterobacter_cloacae_60571
Lactobacillus_gasseri_55805
Faecalibacterium_cf_62236
Bradyrhizobium_sp_61126
Enterobacter_cloacae_55011
Pantoea_sp_60700
Alistipes_finegoldii_56071
Bacteroides_plebeius_61623
Raoultella_planticola_57814
Bifidobacterium_animalis_58116
Clostridium_bolteae_57158
Varibaculum_cambriense_60384
Enterobacter_sp_59869
Klebsiella_pneumoniae_57337
Enterococcus_hirae_55131
Veillonella_sp_62611
Megamonas_hypermegale_57114
Bacteroides_massiliensis_44749
Faecalibacterium_prausnitzii_57453
Ruminococcus_bicirculans_59300
Bacteroides_eggerthii_54457
Dialister_invisus_61905
Bacteroides_finegoldii_57739
Eubacterium_hallii_61477
Bacteroides_intestinalis_61596
Subdoligranulum_sp_62068
Ruminococcus_lactaris_55568
Barnesiella_intestinihominis_62208
Bacteroides_coprocola_61586
Parabacteroides_johnsonii_55217
Faecalibacterium_prausnitzii_61481
Roseburia_hominis_61877
Phascolarctobacterium_sp_59817
Lachnospiraceae_bacterium_51870
Paraprevotella_clara_33712
Alistipes_shahii_62199
Burkholderiales_bacterium_56577
Odoribacter_splanchnicus_62174
Acidaminococcus_intestini_54097
Bacteroides_salyersiae_54873
Sutterella_wadsworthensis_62218
Eubacterium_siraeum_57634
Butyricimonas_virosa_58742
Bacteroidales_bacterium_58650
Oscillibacter_sp_60799
Alistipes_sp_59510
Odoribacter_laneus_62216
Dorea_longicatena_61473
Butyrivibrio_crossotus_61674
Coprococcus_sp_62244
Clostridium_sp_61482
Alistipes_sp_60764
QP (Infant)
non-QP (Infant)
QP (Adult)
non-QP (Adult)
0
200
400
600
800
1000
Number of QP samples per species

Figure S7: Number of QP versus non-QP samples. The 153 species that have greater than 10 QP samples across all hosts are shown; species are ordered by the number of QP samples in infants

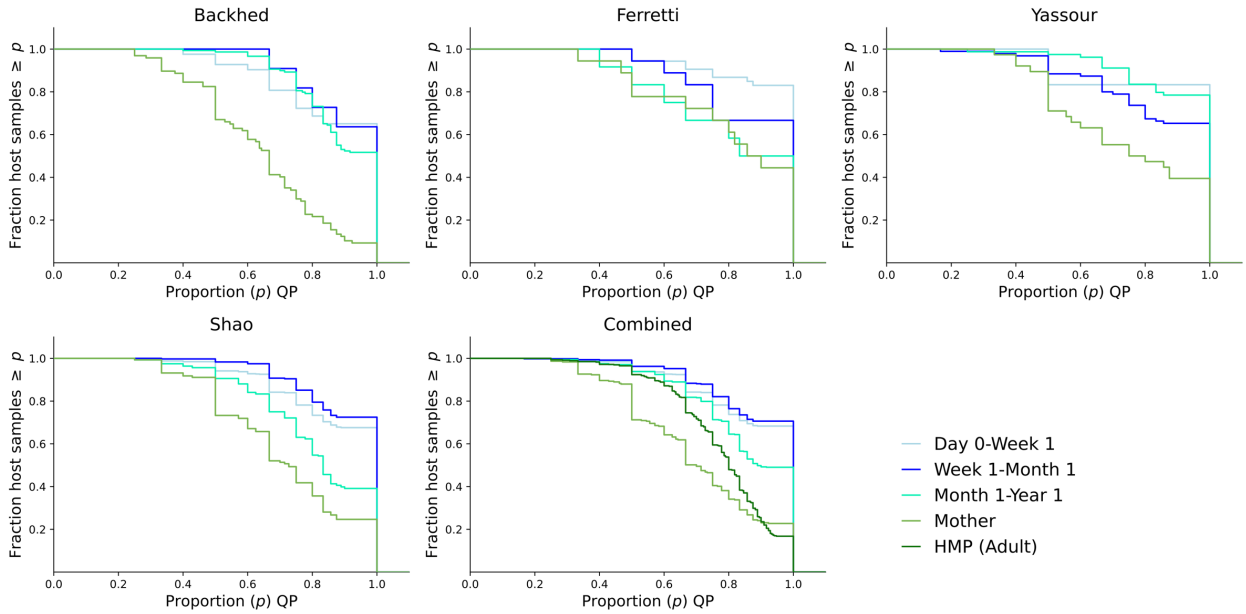

Figure S8: **Proportions of high coverage species that are quasi-phaseable (QP) per host sample, categorized by life stage.** Other than 'Mother' and 'HMP', category label refers to life stage within infants. Here, 'host sample' is defined as a single metagenomic sample. Distributions are plotted per infant dataset, as well as in the combined dataset.

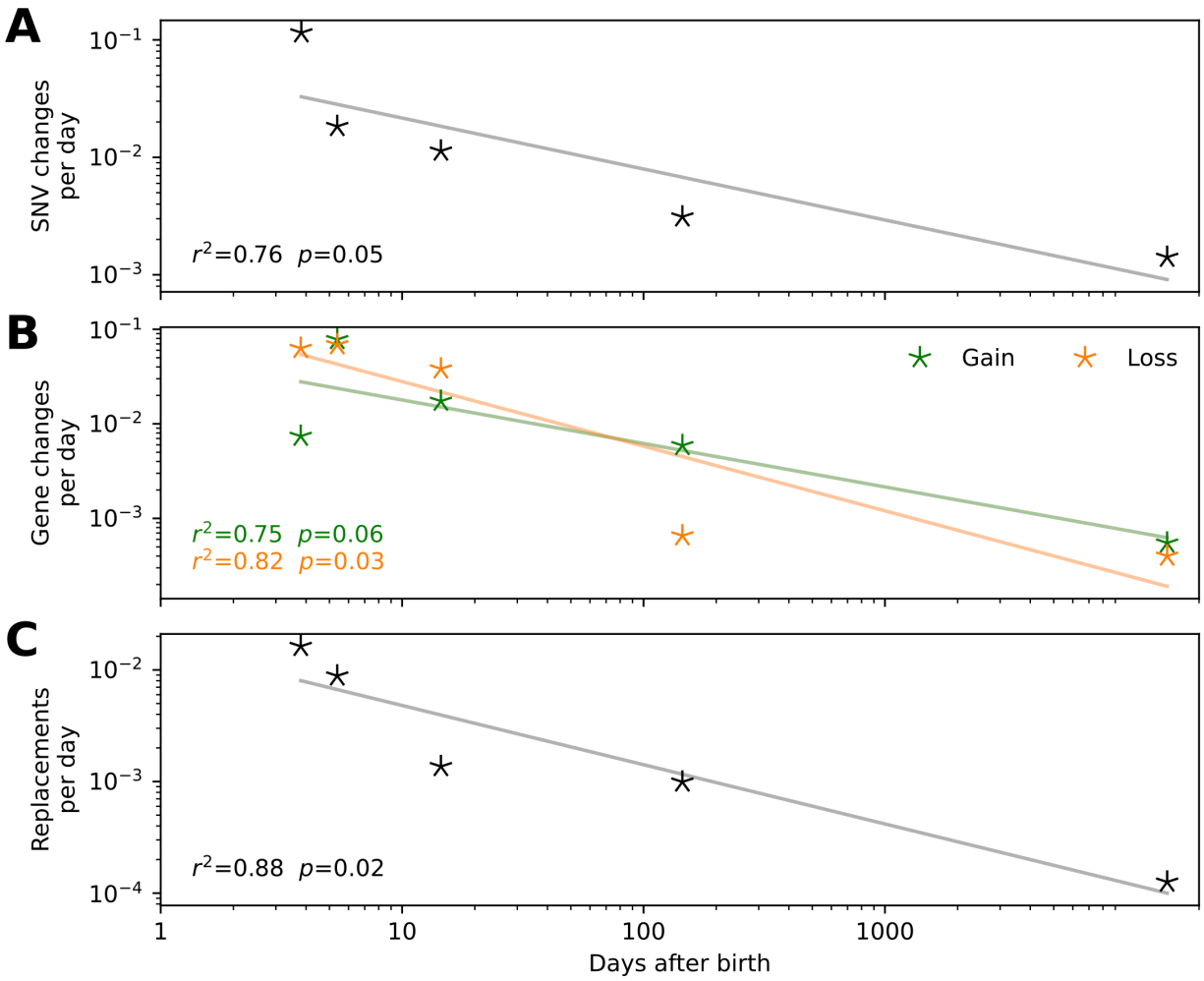

Figure S9: **Decay in rates of SNV change, gene gain, gene loss and replacement rates over life stage.** The same timepoint pair categories as in **Figure 2** were considered: mother-infant within the first week, infant-infant within the first week, infant-infant week 1-month 1, infant-infant month 1-year 1, and adult-adult. Days after birth were assigned to be the median day in a given life stage interval. Mothers were assigned day 0, infant meconium samples were assigned day 1, and adult timepoints were arbitrarily assigned to be approximately 40 years after birth. A linear model was fit with either rate of evolutionary change or replacement as the response variable and days after birth as the predictor. *P* values are reported, as are the correlation coefficients, r^2.

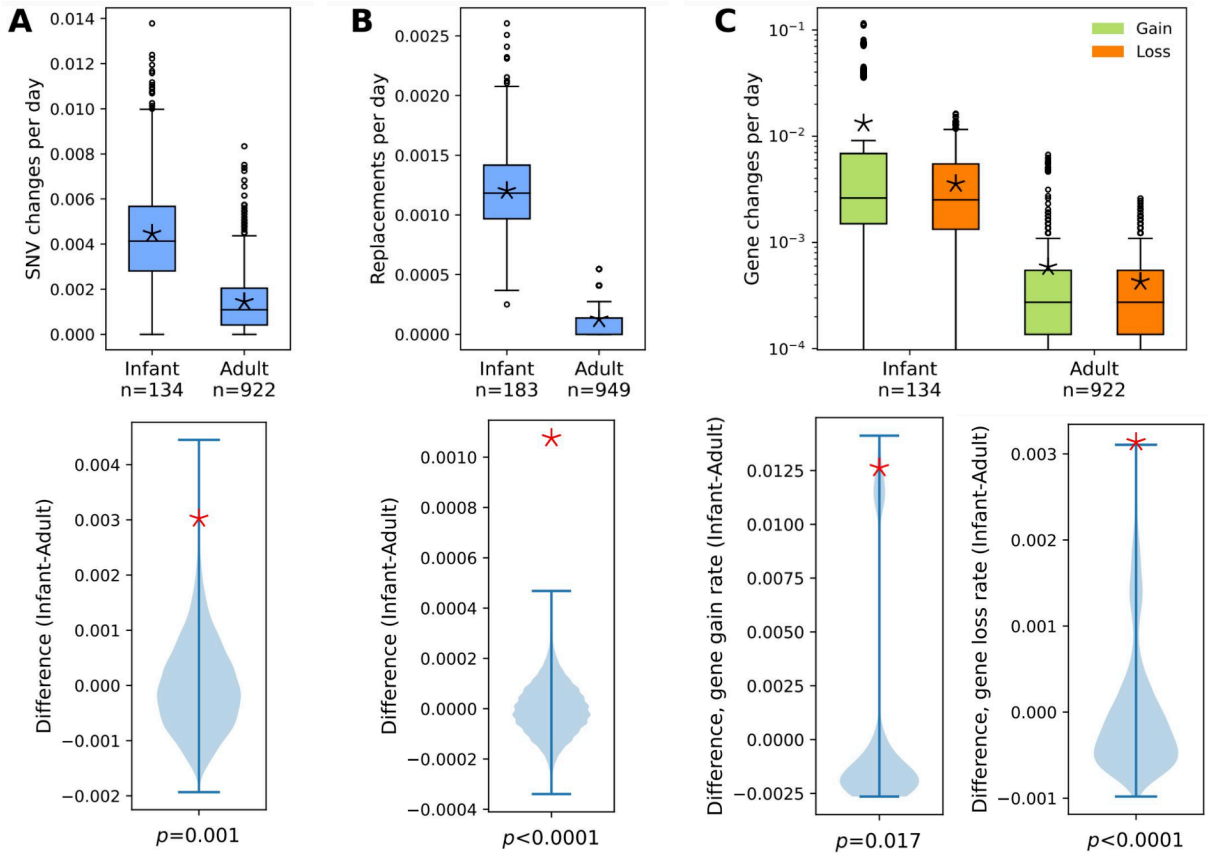

Figure S10: **Evolution and replacement rates in infants versus adults matched for duration of sampling.** Shown are 1000 bootstrapped (**A**) SNV change, (**B**) replacement, and (**C**) gene change rates for all infant-infant vs. HMP adult-adult QP pairs matched for duration between timepoints of 4 to 8 months. To assess if there are significant differences between infant and adult distributions for SNV change, replacement, gene gain, and gene loss rates, we performed a permutation test consisting of 10,000 permutations. The distributions of the difference in rates between infants and adults are plotted, with the observed value indicated with an asterix. *P* values are reported below each permutation distribution plot. Adjusted *p* values with the Benjamini-Hochberg method are 0.0013, 0.0002, 0.017, and 0.0002, respectively.

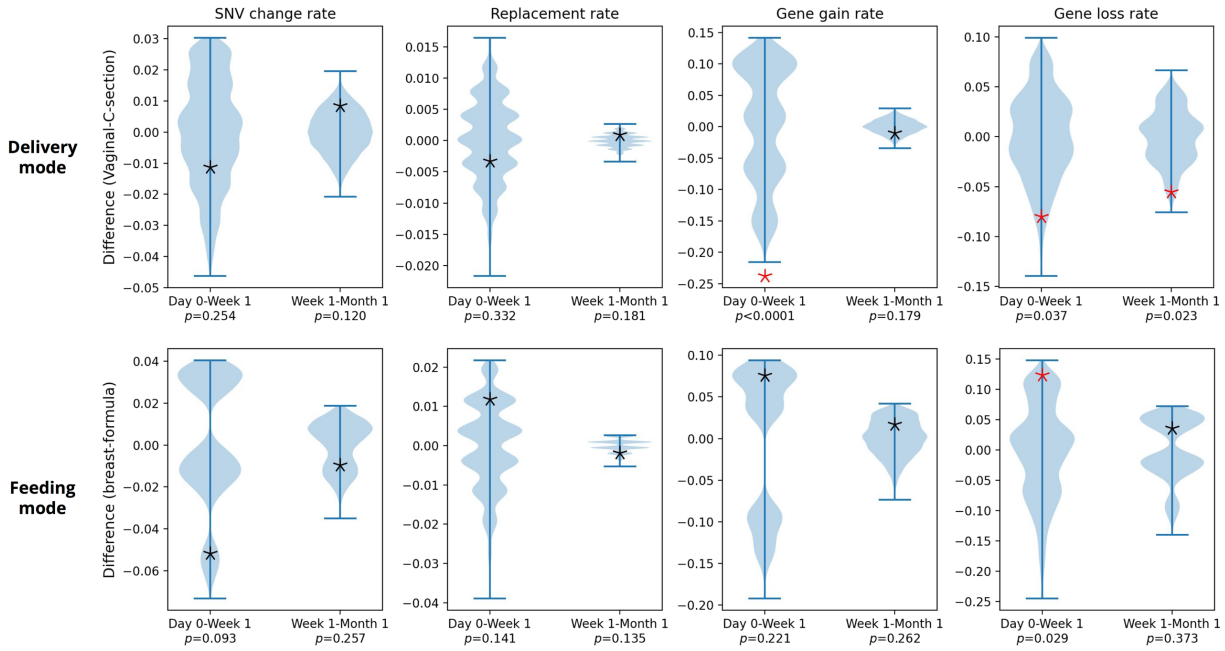

Figure S11: **Difference in rates of evolution and strain replacement for C-section versus vaginally born and breast versus formula fed babies.** We assessed whether rates of SNV changes associated with evolutionary modifications, strain replacement, gene gains, and gene loss are significantly different between C-section and vaginally born babies as well as between breast versus formula-fed babies. In the breast versus formula comparison, we did not include babies on mixed diets. We performed 10,000 permutations, and the resulting distributions are shown above. The astrix shows the observed difference, and reported are associated *p* values. In red are those *p* values that pass the 0.05 significance threshold. Note that only the difference in gene gain rate for delivery mode for day 0 – week 1 survives multiple hypothesis correction with a Bonferroni adjusted *p* value. We analyzed infants in the day 0 to week 1 and week 1 to month 1 categories because these had the maximal number of samples available. The sample sizes were as follows: Day 0-Week 1: Vaginal: 127; C-section: 52; Breast: 112; Formula: 67. Week 1-Month 1: Vaginal: 149; C-section: 59; Breast: 136; Formula: 72.

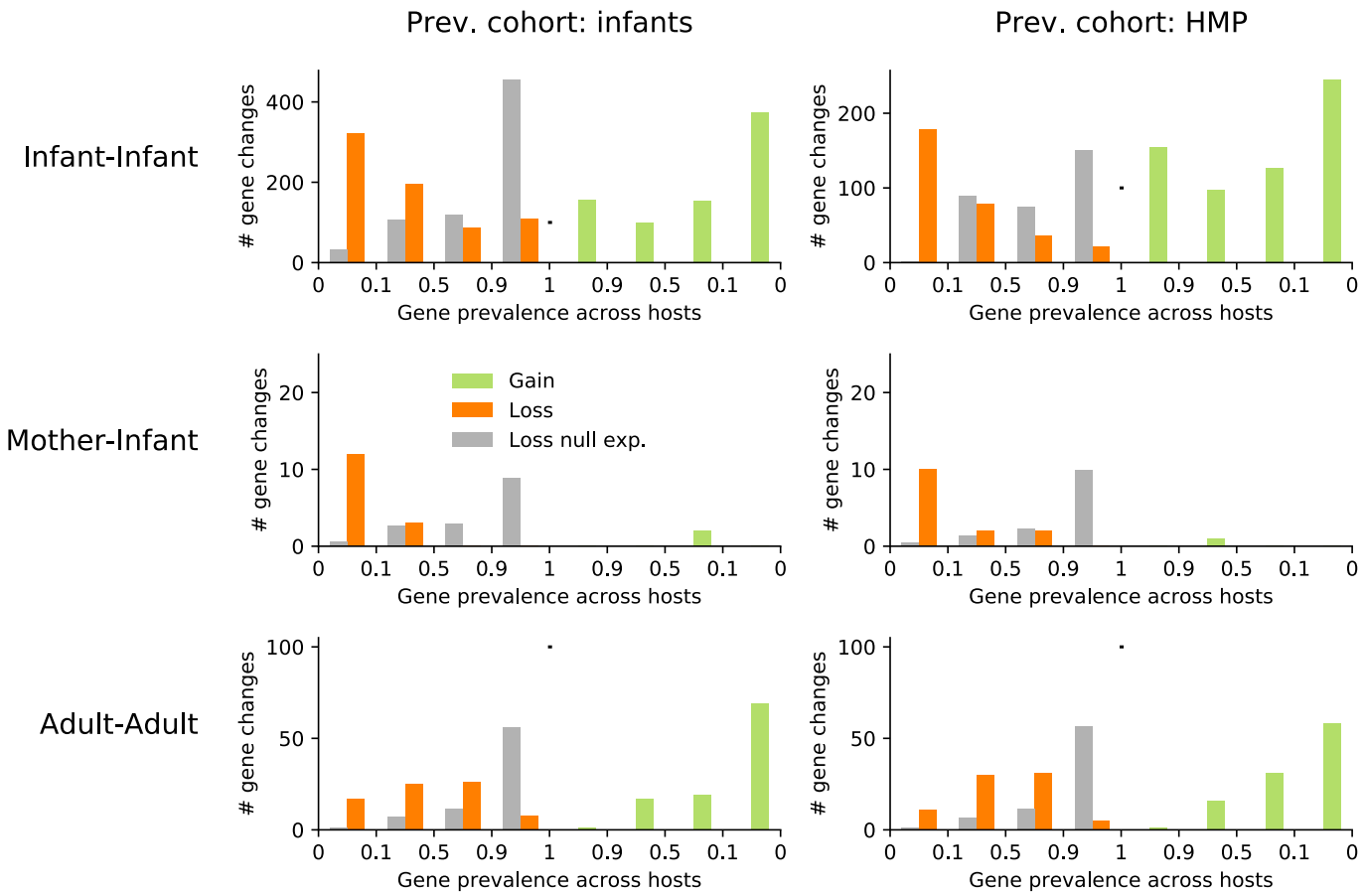

Figure S12: Prevalence of genes that are gained or lost in modification events with respect to infant and HMP adult cohorts.

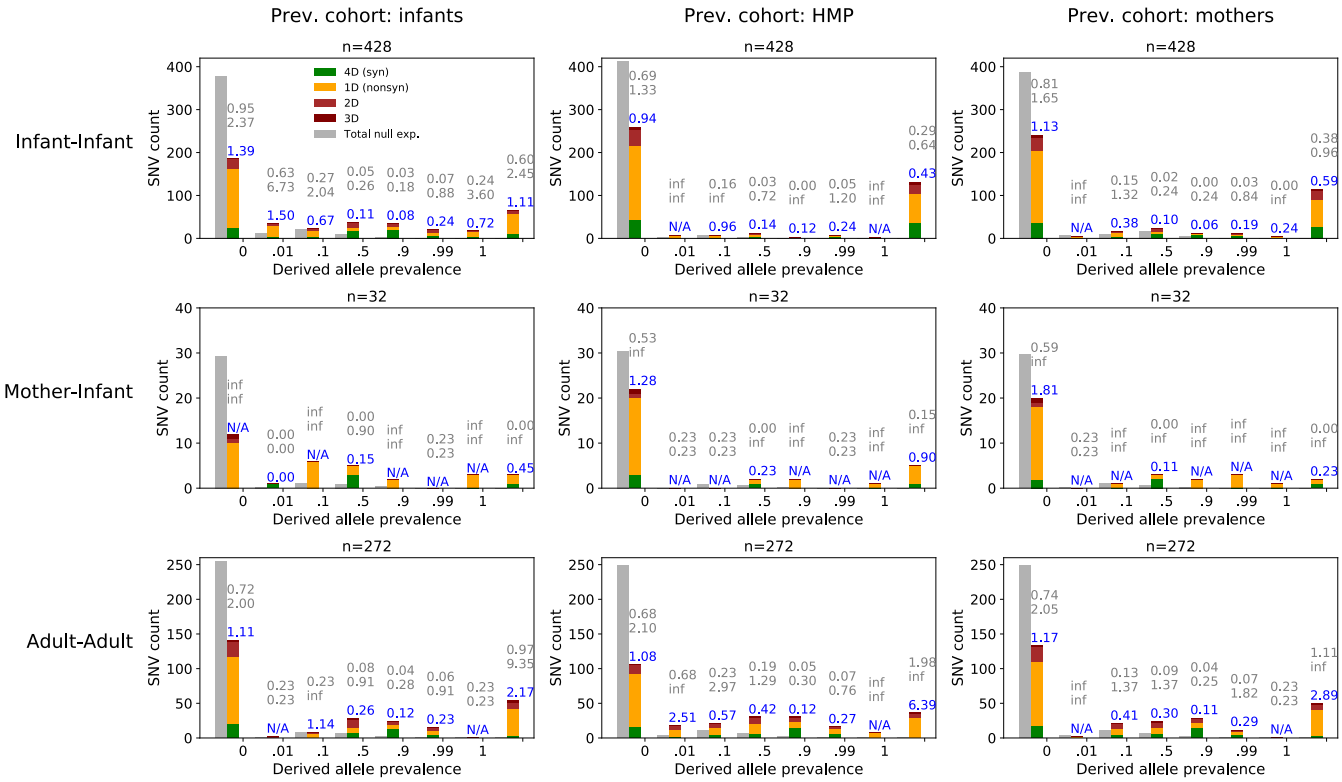

Figure S13: Prevalence of sweeping SNVs in the infant, HMP adult, and mother cohorts. The top left plot shows the prevalence of SNVs sweeping in infants with respect to a prevalence cohort defined by infants. By contrast, the top right plot shows the prevalence of same SNVs sweeping in infants, but with respect to a prevalence cohort defined by mothers. d_N/d_S of each prevalence bin are reported in blue with 95% confidence intervals reported in gray.

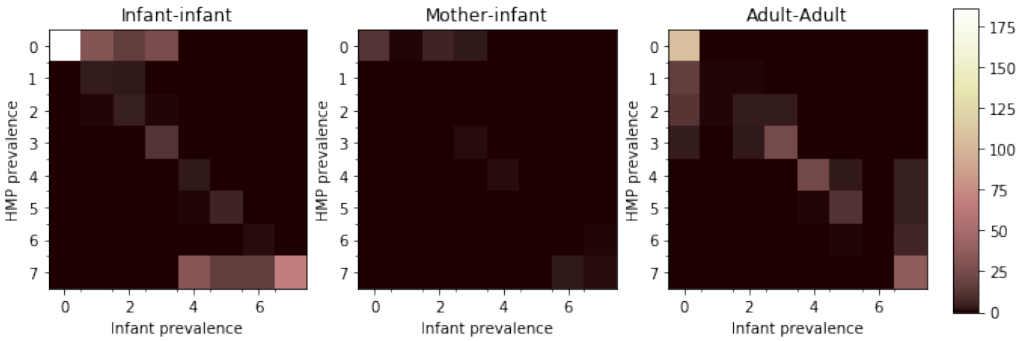

Figure S14: Comparison of HMP (adult) and infant prevalences of sweeping alleles involved in putative modification events. The cohort in which the sweeping allele was identified is indicated at top (e.g 'Infant-Infant', 'Mother-Infant', 'Adult-Adult'). The colorbar indicates number of SNV changes.

Quasi-phasing

A major goal in this paper is to infer evolutionary changes from metagenomic samples. Here, an evolutionary change represents a change in allele frequency over time among lineages belonging to a strain. Unfortunately, two major potential confounders of evolutionary changes in metagenomic data are strain fluctuations and sampling error, since both can generate allele frequency changes.

To control for these potential confounders, we "quasi-phase" samples with sufficiently simple lineage structures. Quasi-phasing means that pairs of alleles can be confidently assigned to a single lineage's genome. In doing so, we can identify evolutionary changes that accrue on the background of a single lineage. The approach we take is similar to that of Truong et al. 2017 in which a dominant allele is assigned to a dominant strain, but in Garud, Good et al. 2019 we put bounds on the error for phasing (for further statistical details, please see the supplement of Garud, Good et al. 2019).

To quasi-phase, we leverage knowledge about the lineage structure of a given species within a host. As described in Garud, Good et al. 2019, hosts are typically colonized by a small handful of genetically distinct lineages belonging to the same species. In **Figure S15**, we plot a distribution of allele frequencies for the common species *Bacteroides vulgatus* in three infant samples from Backhed et al. 2015, which as we describe below, illustrate a range of typical within-host lineage structures.

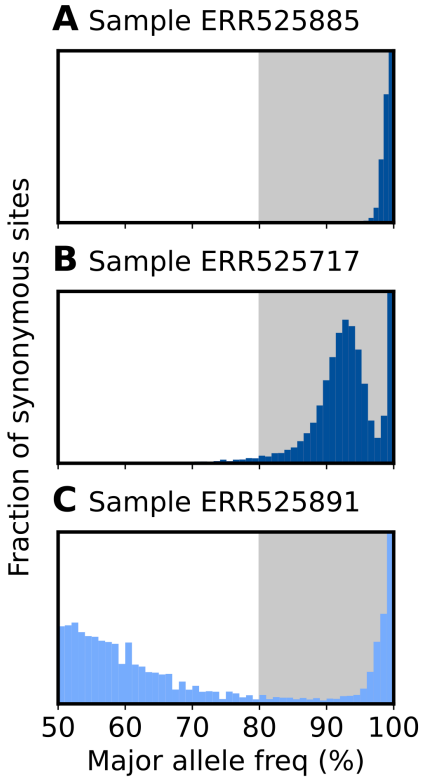

Figure S15: Allele frequency distributions for three infant hosts. Plotted are allele frequency distributions for the species *B. vulgatus* from three infant hosts from the Backhed et al. dataset. Only major allele frequencies are plotted. Figures (A) and (B) depict distributions from quasi-phaseable samples, in which the majority of polymorphic sites have a within-host frequency ≥0.8. The sample in figure (C) is not quasi-phaseable since a large fraction of sites have a frequency < 0.8.

In **Figure S15A**, there is a mass of sites with allele frequencies (*f*) close to 0. As described in Garud, Good et al. 2019, these sites are likely is comprised of a mixture of sequencing errors and low frequency mutations that have arisen due to the expansion of a single lineage within a host. By contrast, in Figures S1B and C there are a mass of sites with allele frequencies close to 0 and another mass of sites with allele frequencies peaked at intermediate frequencies. As described in Garud, Good et al. 2019, these peaked distributions are inconsistent with a single lineage expanding within a host and instead represent multiple, divergent lineages present within a host. The allele frequencies at which these distributions are peaked are representative of the relative frequencies at which lineages are colonizing the host.

In scenario A, a new mutation that arises represents a true evolutionary modification rather than a strain fluctuation, because there is only one strain colonizing the host. However, in scenarios B and C, a shift in frequency of the multiple strains could also generate an allele frequency change. In scenarios B and C we attempt to solve this problem by identifying samples in which a nucleotide confidently to a single lineage's haplotype, or, in other words, can be 'quasi-phased'.

Scenarios A and B are fairly straightforward to quasi-phase. In this scenario, there is typically a single dominant allele (**Figure S16**), which can be assigned to the dominant lineage. Even in the case where an allele is mis-assigned, there will still be a lineage harboring both alleles, even if it is not the dominant lineage (**Figure 16**, Garud et al. 2019, SI text). However, in scenario C, two strains are present at roughly 50% frequency. In this scenario, there is no allele that is dominant and there is a ~50% chance of incorrectly assigning an allele to the dominant lineage (**Figure S16**).

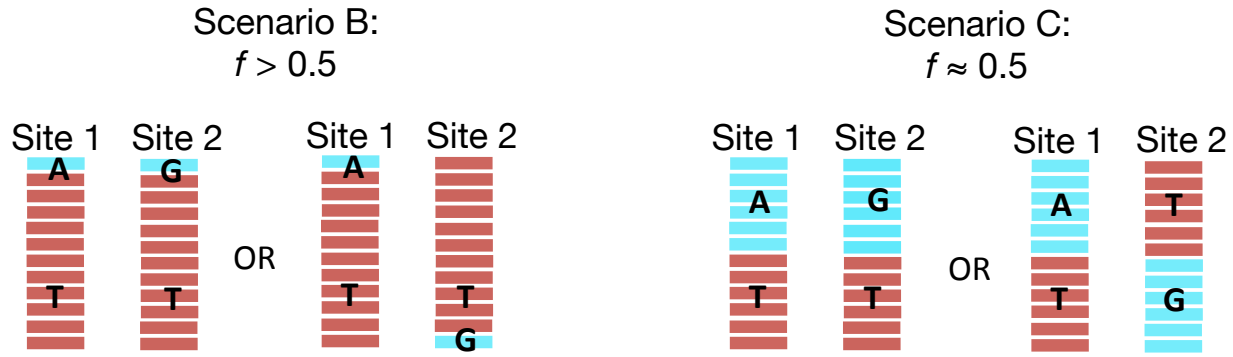

Figure S16: Quasi-phasing of two hypothetical samples with difference allele frequencies. Hypothetical read pileups for two sites in Scenarios B and C. In Scenario B, there is a dominant lineage present at >80% frequency, whereas in Scenario C two lineages colonize the host at roughly 50% frequency (Fig S1). In Scenario B, the dominant alleles at both sites 1 and 2 can be assigned to the dominant lineage. If there is a phasing error, there still will be a fraction of cells that possess the pair of Ts on the same haplotype (see Garud, Good et al. 2019). By contrast, in Scenario C, there is a 50% chance that the alleles will be assigned to the incorrect lineage. In this scenario, it is unlikely that there will be any lineage that harbors both Ts.

Thus, samples with a large number of intermediate frequency alleles are more suspect to phasing errors. Quasi-phaseable samples are those that have few alleles at intermediate frequency. To identify quasi-phaseable samples, we wish to identify an allele frequency cutoff, f^*, that signifies the upper bound of what constitutes 'intermediate frequency'. With such a cutoff, we can then assess the probability that an observed frequency, $\hat{f}$, is greater than f^* given k alternate alleles and D number of reads and a true allele frequency of f:

Eq 1:

$$\Pr[\hat{f} \geq f^* \mid D, f] = \sum_{k > f^* D} \binom{D}{k} f^k (1-f)^{D-k}$$

As described in Garud, Good et al. 2019, this probability can be computed across the genome to obtain a genome-wide error rate of incorrectly phasing an allele. With sufficient depth D, which we assign to be a minimum of 20 in our analysis, and a sufficiently high f^*, which we set to be 0.8, sampling error is minimized. Quasi-phaseable samples are identified if they contain sufficiently low numbers of sites with $\hat{f}$<f* , as described in greater detail in Garud, Good et al. 2019.

3. False positive rate for SNV changes

We next quantified SNV changes between quasi-phaseable time point pairs from the same host. To do so, we identified extreme allele frequency changes from ≤0.2 to ≥0.8 (or vice versa) between two samples, S1 and S2 (**Figure S17**).

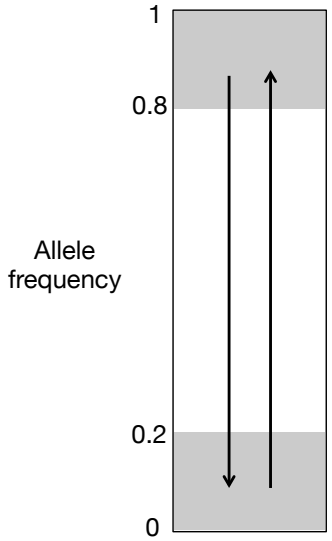

Figure S17: Schematic of allele frequency changes detected. To detect evolutionary changes within a host over time, we identified sites that changed allele frequency from <=0.2 to >=0.8, or vice versa.

We compute the probability of observing an allele frequency due to sampling error (P_{err}) as follows:

Eq2:

$$P_{err} = \int \left(P\left(\widehat{f_1} \leq 0.2 \middle| D_1, f\right) * \left(\widehat{f_2} \geq 0.8 \middle| D_2, f\right) + P\left(\widehat{f_1} \geq 0.8 \mid D_1, f\right) * \left(\widehat{f_2} \leq 0.2 \middle| D_2, f\right)\right) * P(D_1, D_2, f)\ dD_1 dD_2 df$$

Where $\widehat{f_1}$ and $\widehat{f_2}$ are allele frequencies in samples 1 and 2, respectively, and D_1 and D_2 are read depths in samples 1 and 2 respectively. f is assumed to be the same in both samples 1 and 2 under the null hypothesis where there is no evolutionary change. $P(\hat{f}_1 \geq f^*)$ is computed as eq1. $P(D_1, D_2, f) \sim P(D_1)P(D_2)P(f)$ and is estimated empirically from the data as described in Garud et al. 2019 SI text 1.

To compute a genome-wide false positive rate, we can multiply the per-site error rate by the length of the genome, L to estimate the total expected number of false positives:

$N_{err} = P_{err} * L$

For a depth of 20 in samples 1 and 2 (the minimum depth we require) and a true allele frequency f=0.2 in both samples, the probability of observing an allele frequency change by chance from 0.2 to 0.8 is 1.7*10^-8. Multiplying this by a mean genome size of 10^6, the expected number of false positives for a given genome is 0.017, which is <<1.

A similar logic as in Eq2 is applied for inferring gene changes, as described in greater depth in Garud, Good et al. 2019.