Rapid evolution and strain turnover in the infant gut microbiome

Daisy W. Chen and Nandita R. Garud

Genome Res. published online May 11, 2022
Access the most recent version at doi:10.1101/gr.276306.121

Rapid evolution and strain turnover in the infant gut microbiome

Daisy W. Chen[1,2] and Nandita R. Garud[3,4,*]

1. Computational and Systems Biology, University of California, Los Angeles
2. Bioinformatics and Systems Biology Program, University of California, San Diego
3. Department of Ecology and Evolutionary Biology, University of California, Los Angeles
4. Department of Human Genetics, University of California, Los Angeles

* correspondence to ngarud@ucla.edu

ABSTRACT

While the ecological dynamics of the infant gut microbiome have been intensely studied, relatively little is known about evolutionary dynamics in the infant gut microbiome. Here we analyze longitudinal fecal metagenomic data from >700 infants and their mothers over the first year of life and find that the evolutionary dynamics in infant gut microbiomes are distinct from that of adults. We find evidence for more than 10-fold increase in the rate of evolution and strain turnover in the infant gut compared to healthy adults, with the mother-infant transition at delivery being a particularly dynamic period in which gene loss dominates. Within a few months after birth, these dynamics stabilize, and gene gains become increasingly frequent as the microbiome matures. We furthermore find that evolutionary changes in infants show signatures of being seeded by a mixture of *de novo* mutations and transmissions of pre-evolved lineages from the broader family. Several of these evolutionary changes occur in parallel across infants, highlighting candidate genes that may play important roles in the development of the infant gut microbiome. Our results point to a picture of a volatile infant gut microbiome characterized by rapid evolutionary and ecological change in the early days of life.

INTRODUCTION

The infant gut microbiome is a rapidly changing ecosystem that plays a crucial role in infant health, including milk digestion (Sela et al. 2008), immune system development (Smith et al. 2013), and prevention of colonization of pathogens (Nicholson et al. 2012; Sela et al. 2008). Given its importance, there has been substantial research focus on the ecological and functional maturation of the infant gut microbiome (Stewart et al. 2018; Niu et al. 2020; Sprockett et al. 2020; Chu et al. 2017; Palmer et al. 2007; Vallès et al. 2014; Yassour et al. 2016; Koenig et al.

2011; Korpela et al. 2018; Ferretti et al. 2018; Bäckhed et al. 2015) as well as the extent to which mode of delivery, feeding, and exposure to antibiotics alters its development (Shao et al. 2019; Karlsson et al. 2013; Bokulich et al. 2016; Dominguez-Bello et al. 2010; Fehr et al. 2020; Koenig et al. 2011; Mitchell et al. 2020; Yassour et al. 2016). Strain tracking methods using single nucleotide variants (SNVs) have revealed that mothers (Ferretti et al. 2018; Mitchell et al. 2020; Nayfach et al. 2016; Yassour et al. 2018), the family at large (Korpela et al. 2018; Hildebrand et al. 2021), and hospitals (Raveh-Sadka et al. 2016; Brooks et al. 2017) play a critical role in seeding the infant microbiome. By contrast, how these lineages evolve once they colonize the infant remains unknown.

Evolutionary changes are important to characterize because the rise of new genetic variants ultimately drives the emergence of new traits. For example, genetic variants in the human gut microbiome are known to confer traits such as the ability to digest food (Kenny et al. 2020; Hehemann et al. 2010), metabolize drugs (Spanogiannopoulos et al. 2016), and evade antibiotics (Gumpert et al. 2017). Recent studies illustrate the pervasiveness of evolutionary changes in adult gut microbiomes, which include changing SNV frequencies and horizontal gene transfers on short time scales of just a few days, weeks, and months (Garud et al. 2019; Zhao et al. 2019; Yaffe and Relman 2020; Groussin et al. 2021; Poyet et al. 2019; Roodgar et al. 2021; Jiang et al. 2019; Ghalayini et al. 2018). However, rapid evolutionary changes within species frequently are not associated with changes in species relative abundances (Roodgar et al. 2021), illustrating that intra-species genetic variation in the microbiome can reveal changes in the microbiome that species abundances cannot. Given the functional importance of genetic variation in the microbiome, it is necessary that we characterize the typical evolutionary dynamics that occur over the course of infant maturation.

There is reason to believe that the targets of selection as well as the tempo and mode of evolution in infant gut microbiomes differ substantially from those of adults. Compared to adults, microbes in infants confront unique selective pressures. For example, infants have an immature immune system compared to adults (Sjögren et al. 2009) and a vastly simpler diet composed primarily of milk in the early months of life before transitioning to solid foods. Moreover, infants harbor simpler ecological communities with low levels of richness (Bäckhed et al. 2015) which could alter the overall rate of evolution in the community (Venkataram et al. 2021; Post and Palkovacs 2009). In particular, rates of horizontal gene transfer, a common

feature of the adult gut microbiome (Groussin et al. 2021; Garud et al. 2019; Lin and Kussell 2019; Sakoparnig et al. 2021), could be reduced in infants given the simpler community since the broader ecosystem may serve as a reservoir for pre-adapted material. Given these differences, it is possible that infant-specific evolutionary changes are not only present but also necessary for gut microbiota to successfully colonize their new environment.

Here, we track the evolutionary and ecological dynamics of the gut microbiome in >700 infants and their mothers over a span of 1 year after birth. In this study, we assess changes in rates of evolution and strain replacement with life stage. Additionally, we assess evidence for parallelism of evolution across infants to identify candidate genes that may play important roles in the developing infant gut microbiome. By gaining an understanding of both the ecological and evolutionary dynamics of the developing infant gut microbiome, we can better understand the fundamental processes that contribute to the maturation of this complex ecosystem at a particularly volatile and critical juncture.

RESULTS

Microbiome diversity rapidly grows in the first year of life

To quantify the evolutionary dynamics of gut microbiota in infants, we analyzed fecal metagenomes from four cohorts (Shao et al. 2019; Ferretti et al. 2018; Yassour et al. 2018; Bäckhed et al. 2015), totaling 2399 samples from 762 healthy infants and 337 mothers. Additionally, we analyzed 249 healthy adults from the Human Microbiome Project (HMP) (Lloyd-Price et al. 2017; The Human Microbiome Project Consortium 2012), and 185 healthy adults from Qin et al. (2012) (**Table 1**) to compare the evolutionary dynamics in adults versus infants. These datasets were chosen because of the availability of deeply sequenced longitudinal data from a large panel of healthy individuals. Infants were longitudinally sampled at 2-7 timepoints ranging from birth (meconium) to 1 year post-delivery across cohorts (**Figure S1**), with dense samples within the first week and month of life, as well as every month thereafter. Mothers were sampled within 1 week post-delivery, while HMP adults were sampled 1-3 times over a time span of approximately 1 year.

Dataset	# Hosts	Host Type	# Samples	# Timepoint(s) per host

HMP	249	U.S. adults	469	1-3
Qin et al. 2012	185	Chinese adults	185	1
Backhed et al. 2015	98	Swedish infants and mothers	391	4
Yassour et al. 2018	44	Finnish infants and mothers	213	6
Ferretti et al. 2018	25	Italian infants and mothers	119	6
Shao et al. 2019	600	UK infants and mothers	1676	2-7

Table 1. Datasets analyzed in this study. For the infant gut microbiome datasets, "host" refers to all samples from mothers and infants in the same dyad. For the Yassour et al. (2018) dataset, only mothers at time of delivery were included in our analyses.

We used a reference-based mapping approach (Nayfach et al. 2016) to call single nucleotide variants (SNVs) and gene copy number variants (CNVs) for sufficiently abundant and prevalent species in our datasets (**Methods**). Summaries of diversity in infant microbiomes at the species and sub-species levels have been reported previously (Korpela et al. 2018; Shao et al. 2019; Bäckhed et al. 2015; Yassour et al. 2018; Ferretti et al. 2018; Nayfach et al. 2016). Here, we revisit diversity patterns at the species and nucleotide level to investigate how gut microbiome community complexity changes in the infant over the first year of life. We then leverage these observations to infer the lineage structure within metagenomic samples to be able to make evolutionary inferences.

To understand how diversity changes over life stage, we first revisited Shannon alpha diversity of gut microbiota from birth to adulthood (**Figure S2**). Alpha diversity increases with the age of the infant, approaching levels observed in adults (HMP and Qin et al) by month 12, but generally does not surpass that of mothers at time of delivery. Among these datasets, mothers at time of delivery have higher alpha diversity than non-pregnant female adults (*p* value 0.035, GLMM) (**Figure S3**), suggesting that pregnancy significantly increases microbiome diversity of the mother.

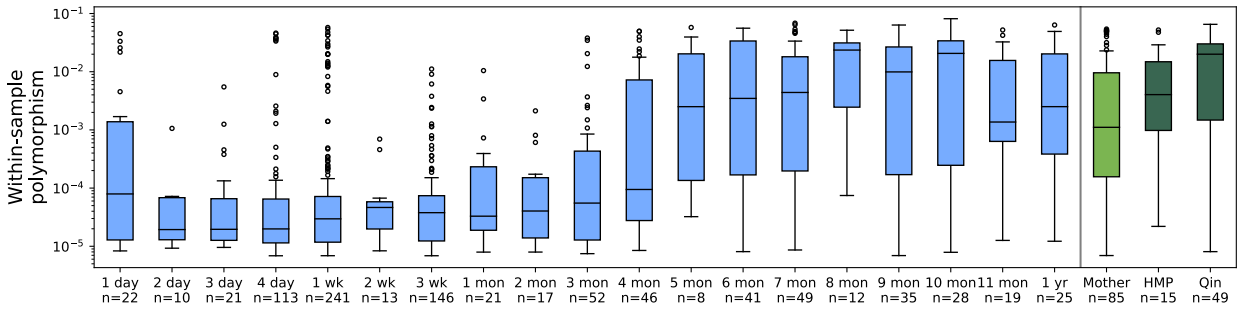

Figure 1. Polymorphism rates for *E. coli* over the course of the first year of life. Shown are within-sample nucleotide polymorphism rates per base pair for the most prevalent species among infants in our dataset, *E. coli.* **Figures S4 and S5** show distributions for the next 12 most prevalent species. Here, within-sample nucleotide polymorphism was quantified as the fraction of synonymous sites in core genes with allele frequencies between 0.2 and 0.8 (see **Methods**).

We next asked whether within-species genetic diversity also increases with life stage. Here we measure polymorphism rate, measured by the fraction of synonymous sites in core genes with an intermediate frequency polymorphism between 0.2 and 0.8 (**Supplemental Text**). We plot polymorphism levels for the most prevalent species in our dataset in **Figure 1** and **S4**. Although polymorphism levels are generally not significantly different between datasets for the same life stage (**Figure S6, Supplemental Text, Table S1**), we also plot the polymorphism within the Shao et al. dataset only (**Figure S5**), for which there are maximal timepoints available. Some species, including *Escherichia coli, Bifidobacterium longum* and *Bifidobacterium breve*, experience increases in nucleotide diversity over time (**Figures 1** and **S4, S5**). Specifically, coincident with the average time of the transition to solid foods at ~5-6 months of age (Van Dijk et al. 2012), median polymorphism rates increase sharply by an order of magnitude from $<10^{-3}$ /bp to $\sim10^{-3}$-10^{-2}/bp. This sudden increase in polymorphism rates is likely driven by increasing numbers of strains colonizing the host and not evolution of the resident strains. A mutation rate of 10^{-9}/bp (Barrick and Lenski 2013) and generation time of ~1-2 generations/day (Sender et al. 2016) are too low to produce such rapid increases in levels of polymorphism, whereas 'oligo-colonization' of multiple strains of the same species is a far more likely candidate (Garud et al. 2019).

Rates of polymorphism do not consistently increase with age for all species. For example, *Bacteroides fragilis* and *Bifidobacterium bifidum* have consistently low median within-host polymorphism of 10^{-4}/bp over most time points in the first year of life. Other species do not display consistent trends and instead show wide variation in levels of polymorphism over time, likely a reflection of the stochasticity of the colonization process (**Figures S4 and S5**). Thus, the ecological forces determining colonization success and lineage structure likely vary from species to species.

Since hosts are often oligo-colonized by multiple genetically distinct strains of the same species (Garud et al. 2019; Truong et al. 2015), fluctuations in strain frequencies can confound the detection of evolutionary changes from shotgun metagenomic data as both can generate SNV

and gene differences over time (Garud et al. 2019). To confidently distinguish SNV and gene changes (e.g. horizontal gene transfers) due to evolution from fluctuations in strain frequencies, we leveraged a quasi-phasing approach that we previously developed (Garud et al. 2019) to assign genotypes to individual lineages for each species (**Methods, Supplemental Text).** This approach can be applied to species in samples with sufficiently simple lineage structures, i.e. having a single dominant strain such that alleles can be confidently assigned to that dominant lineage; specifically, the probability of incorrectly inferring the allelic state of a lineage at any given site is bounded (**Supplemental Text**). In practice, quasi-phasing excludes samples with high proportions of intermediate frequency polymorphisms, which have a high probability of being misassigned to the correct lineage. With quasi-phaseable (QP) samples, evolutionary changes that occur on the background of a given lineage can be tracked and distinguished from SNV and gene changes due to fluctuations of genetically distinct strains.

Figures S7 and S8 shows the distribution of QP samples (host × species quasi-phaseable lineage) across sufficiently prevalent gut bacterial species in the combined dataset and in individual datasets, respectively. Among infants, there are 7,063 QP samples; among mothers, there are 1,159 QP samples; and among HMP adults there are 3,544 QP samples from 217 candidate species. To quantify evolutionary changes, we first identified QP sample pairs for consecutive time points from the same infant or adult host as well as mother-infant comparisons from the same dyad. This yielded a total of 2,184 infant-infant, 241 mother-infant (where the infant sample was within the first week of life), and 1,296 adult-adult HMP QP pairs (combinations of host, species, and timepoint pair) across 176 of the 217 most prevalent species. This large number of QP pairs enables tracking of gut microbiota evolutionary dynamics across life stages. To infer evolutionary changes, we identified SNVs that change in allele frequency from ≤0.2 to ≥0.8 in the core and accessory parts of the genome between pairs of QP time points (**Methods and Supplemental Text**). By biasing our results towards extreme allele frequency changes, the expected number of false positive SNV changes due to sampling error is < 1 (**Supplemental Text**).

Elevated rates of evolutionary change and strain turnover in the infant compared to adult gut microbiome

We quantified SNV differences between sampled time points in four categories of life

stages: infants sampled over < 3 month intervals (which generally fall within the first 3 months after birth), infants sampled over ~3-6 month intervals (which generally fall after the first 3 months of life), pairs of mother and infant samples within the first week of life, and HMP adults sampled ~6 months apart. In the mother-infant category, we treated mothers as one time point and infants as another for a given dyad, where mother time points fall within the first week.

SNV differences between timepoints in a QP pair can arise from a combination of two processes: strain replacement and evolutionary modification (Garud et al. 2019). Reflecting this, SNV differences are distributed bimodally across all four life stage categories (**Figure 2A**). Most within-host QP sample pairs (83%) experience zero SNV changes over timescales of a week or less, but a small percentage (9%) undergo a small number of SNV changes (≤20). An even smaller percentage of hosts harbor ~10^4 SNV differences (7%), which is on the same order of magnitude of the number of SNV differences between unrelated hosts. This between-host comparison serves as a helpful reference for typical nucleotide divergence between resident and invading strains. Thousands of SNV changes accumulating within hosts over 6 month time scales are unlikely to have arisen from evolutionary diversification of a lineage within a host and are instead consistent with strain replacement (Garud et al. 2019). Thus, we classified the samples experiencing ≤20 SNV changes as undergoing evolutionary modifications and samples experiencing >500 SNV changes as undergoing strain replacements.

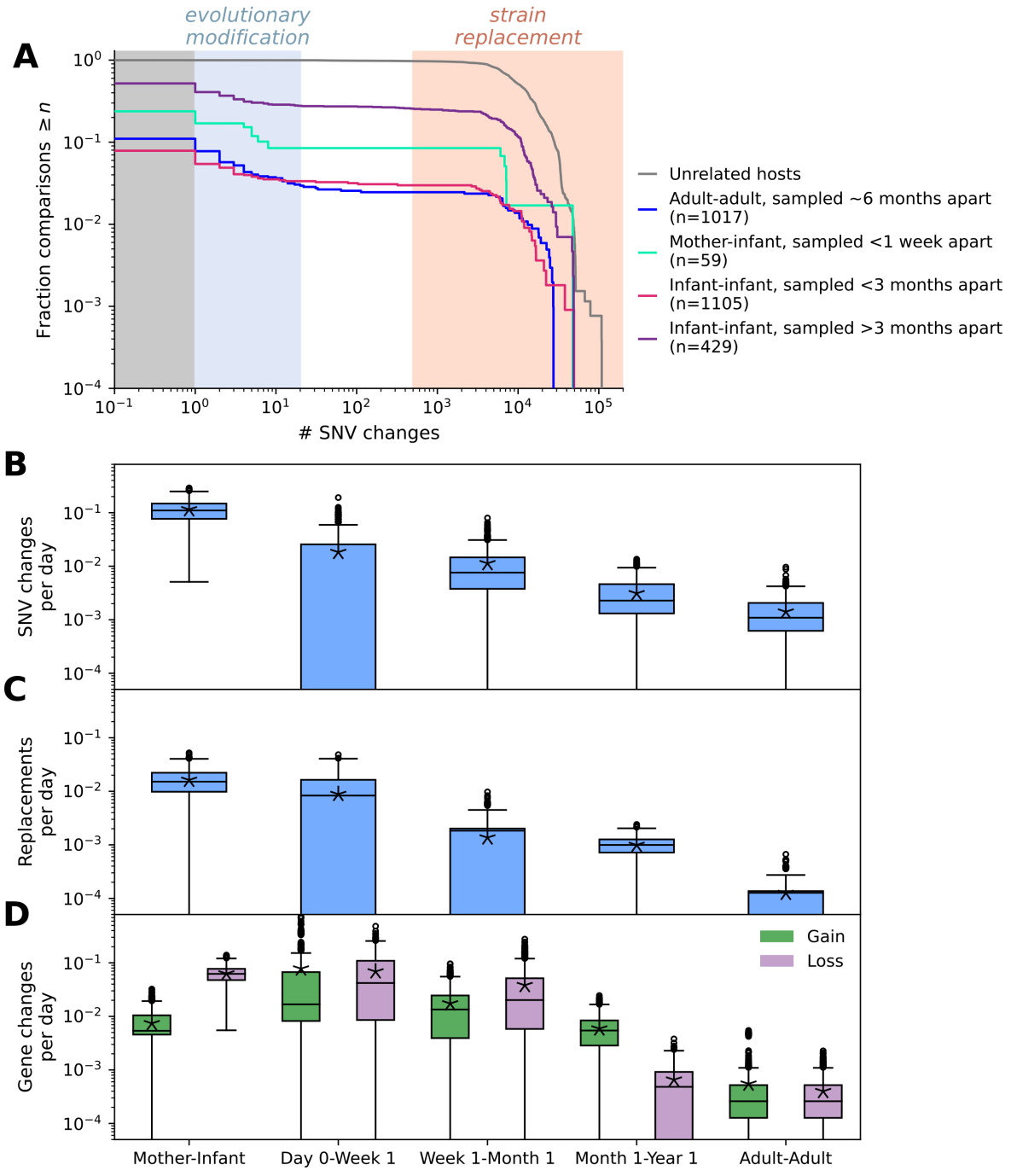

Figure 2. Rates of evolutionary change and replacement decay with life stage. (**A**) Survival distributions of number of SNV differences (defined as changes in allele frequency from <0.2 to >0.8 between two time points) for each life stage. (**B**) Rates of number of SNV changes per day for QP pairs experiencing evolutionary modification, (**C**) rates of number of strain replacements per day for all QP pairs, and (**D**) rates of gene gains and loss per day for QP pairs experiencing evolutionary modification. Asterisks indicate the mean rate computed over all QP pairs in a life stage. To assess variance, rate estimates for each life stage were bootstrapped 1000 times by subsampling with replacement 40 out of all QP pairs available. The mother-infant category only includes comparisons between mother and her baby sampled within one week of life. To address

potential confounding of amount of time between sampling, we compared rates for infants and HMP adults sampled 4-8 months apart (**Figure S10**). In **B-D**, all sample sizes are ≥ 53 for each life stage. Exact sample sizes and statistical significance of permutation tests comparing all pairs of life stages are reported in **Table S2**. In **Figure S9**, a scaling relationship between changes and life stage is assessed.

Figure 2A shows that infants sampled ~3-6 months apart have elevated proportions of modification (24.2%) and replacements (25.6%) compared to adults. By contrast, among HMP adults sampled on ~6 month time scales, only 8.1% of resident populations undergo modification and an even smaller percentage (2.5%) undergo replacement (Garud et al. 2019). Mother-infant dyads sampled within the first week also have elevated proportions of modification (14.5%) and replacement (11.3%).

The increased proportions of modifications and replacement events in infants relative to adults suggest that ecological and evolutionary processes occur more frequently and rapidly early in life compared to in adulthood. To confirm this, we next quantified per-day rates of evolutionary modification and replacement (**Figure 2B-D)** over various life stages, aggregated over species (**Methods**).

Both modification and replacement rates decay rapidly with life stage and in fact follow a scaling relationship with the age of the host (**Figures 2** and **S9**). During the mother to infant transition, average modification rates are ~0.1 SNV changes/day. Within infants during the first week of life, the average rate drops 10-fold to ~0.02 SNV changes/day (p value = 0.012, permutation test), and by adulthood, the average rate drops another ~10-fold to ~0.001 SNV changes/day (p value < 10^{-4}, permutation test), consistent with rates previously estimated in adults (Zhao et al. 2019; Didelot et al. 2012). Similarly, replacement rates also rapidly decay with the age of the host. During the mother-infant transition, the average strain replacement rate is $1.6 \cdot 10^{-2}$ replacements/QP pair/day. Then, within the first week of life, the average replacement rate drops to $8.9 \cdot 10^{-3}$ replacements/day (p value =0.16, permutation test), and finally in adulthood it drops to $1.3 \cdot 10^{-4}$ replacements/day (p value < 10^{-4}, permutation test).

We next quantified rates of gene gains and loss per day, representing another mode of evolutionary change common in adult microbiomes (Garud et al. 2019; Groussin et al. 2021; Zhao et al. 2019; Smillie et al. 2011; Lin and Kussell 2019; Yaffe and Relman 2020; Zlitni et al. 2020; Coyne et al. 2014) (**Methods**). Gene gains and losses represent an acquisition or loss of a gene in the most common lineage relative to the most common lineage at a prior time point. Like

the modification and replacement rates, gene change rates also rapidly decay with life stage and follow a scaling relationship (**Figures 2D and S9**). However, the rate of gene loss decays more quickly than the rate of gene gain with life stage after the mother-infant transition; furthermore, during the mother-infant transition, gene losses exceed gains by almost 10-fold (an average of 0.06 gene losses/day versus 0.007 gene gains/day (p value=0.002, permutation test). This trend reverses later in life; between 1 month and 1 year after birth, gains exceed losses by almost 10-fold (an average of $5.9 \cdot 10^{-3}$ gains/day versus $6.6 \cdot 10^{-4}$ losses/day, p value=0.008, permutation test). Eventually, by adulthood, losses and gains occur at similar rates of $\sim 5 \cdot 10^{-4}$ changes/day (p value = 0.33, permutation test). This suggests that during the initial host colonization process at birth, diversity at the gene level is reduced through gene loss, but recovers gradually as the infant gut matures before eventually reaching an equilibrium in adulthood. As discussed further below, we note that it is not possible to exclude the possibility that the gene gains and losses observed in infants represent pre-evolved lineages transmitted from the mother or another family member at a later time point.

Since earlier life stages are generally sampled at shorter consecutive time intervals in our dataset, we assessed whether duration of time between sampling points, rather than life stage, could explain the faster rates of modification, strain replacement, and gene changes in infants. We found that infants sampled over 4-8 month timescales had significantly elevated SNV change (for modification events), gene change, and strain replacement rates compared to adults sampled on similar time scales (p values <0.02, permutation test, **Figure S10**), confirming that age, rather than duration of the sampling interval, is a major driver of faster evolution in the infant gut microbiome.

Finally, we assessed whether birth or feeding mode is associated with changes in rates in evolution or strain replacement (**Figure S11**). To perform this analysis, we focused on QP pairs for the age ranges of day 0 to week 1 and week 1 to month 1 because these had the most samples available in the C-section versus vaginal classes and breast versus formula fed classes. Rates of gene gain and loss are significantly elevated among C-section infants compared to vaginally born in both life stages (p values <0.05, permutation tests), but SNV change and strain replacement rates do not significantly differ with delivery mode. Gene losses were also elevated among formula fed infants compared to breast fed infants during the day 0 to week 1 life stage (p value <0.05, permutation test), but not in the older life stage. The consistent elevation in gene gain and

loss rates for C-section infants across both life stages suggests that ongoing gene changes may play an important role in the establishment of the microbiomes of C-section infants.

Evolutionary changes in the infant are likely seeded by a mixture of *de novo* mutations and standing genetic variation

To further probe the evolutionary origins of the SNV changes in infants, we asked whether they arise in infants from *de novo* mutation or pre-existing standing genetic variation that may have been seeded by recombination and also could have potentially arisen in the broader family unit before being transmitted to the infant. To assess evidence for the two scenarios, we combined the 428 SNV changes occurring in the 154 modification events of infant-infant QP pairs and then assessed the prevalence of sweeping (or "derived") alleles (**Figure 3B**). Here, "prevalence" is defined as proportion of HMP adults that harbor the sweeping allele. We compare the prevalence of sweeping alleles with that of a null distribution of randomly selected sites assuming *de novo* mutation (**Methods**). Additionally, we compare the prevalences of sweeping alleles during the mother to infant transition (**Figure 3D**) and in adults (**Figure 3F**).

The observed distribution of prevalences shows interesting departures from the null expectation. Specifically, under the null, the majority of sweeping alleles are expected to be rare in the broader population. By contrast, the observed distribution of prevalences is bimodal, with infant sweeping alleles being completely absent from adults or present in virtually all HMP adults (**Figure 3B**). This is distinct from the distribution observed in adults in which sweeping SNVs are also enriched for intermediate prevalences (0.1-0.5) (**Figure 3F**). The bimodal distribution in infants suggests that evolutionary changes occurring within infants are either *de novo* mutations that are absent among adults or reversions to the consensus state in adults, possibly reflecting infant-specific adaptations in the former case or reversions of maladaptations in the latter case.

The majority of infant sweeping alleles that are absent among unrelated adults are private to the infant in which the allele sweeps and are rarely found in the infant's respective mother (**Figure S14**). Specifically, only 2% of the 186 infant sweeping alleles that are absent in HMP hosts are found in mothers. These rare infants sweeping SNVs have a d_N/d_S value of 1.3 (CI 0.95 to 2.37, **Figure S13**, **Methods**), indicating that they are indeed potentially adaptive. However,

they could also be recent deleterious mutations that reached high frequency by chance during a bottleneck associated with transmission to the infant. In several instances, we observe that private nonsynonymous mutations that arise early in an infant's life revert to an allelic state that is present in all other HMP adults later in life (**Figure 4A**), suggesting that some of these initial mutations may indeed be deleterious. Together, the preponderance of rare alleles with an overall d_N/d_S >1 as well as frequent reversions of these rare alleles to prevalent allelic states indicate that *de novo* mutations are likely common in infants.

However, the rest of the infant sweeping alleles, most of which are present in the majority of HMP adults (**Figure 3B**), have a d_N/d_S value of ~0.6. This lower d_N/d_S likely reflects that purifying selection has had sufficient time to purge deleterious variants from the population, indicating that *de novo* mutation is not the only mode of evolution in the infant. Instead, the high prevalences and lower d_N/d_S value suggest that in many instances, sweeps from standing variation seeded by recombination events may be common in the infant as well, as has been observed in adults (Garud et al. 2019). Consistent with recombination playing an important role, the gene gains observed in infants cannot have arisen via *de novo* mutation and instead must have arisen via recombination events. In fact, the increase in rate of gene gains over the first year of life indicates that recombination likely plays a significant role in recovering diversity that is lost during the mother to infant transition. Indeed, **Figure 3C** shows few genes are gained during the mother to infant transition in contrast to the first year of life (**Figure 3A**). Over the first year of life, a larger proportion of genes gained in infants compared to those gained in adults have a high prevalence (e.g. >0.9) (**Figures 3A** and **3E**), suggesting that common genes play an important role in recovery of diversity.

It is possible that several of these evolutionary events observed in infants occurred in the mother or broader family unit before the strain was transmitted to the infant. In two mother-infant dyads, multiple alleles sweep to high frequency in the infant after birth before later reverting to allelic states found in the mother (**Figure 4B**). It is unlikely that these are *de novo* mutations that occurred twice in succession since >4 SNVs change twice and include synonymous sites. Notably, multiple haplotype configurations are observed over time in these two dyads, suggesting that multiple variants of the same lineage circulate among the members of the family unit. This suggests that infants may be seeded multiple times by their family members and likely experience significant flux of variants of the same strain over the course of their first

year.

We conclude that multiple processes contribute to SNV changes within infants. During infancy, new *de novo* mutations arise that are generally absent from the broader population. Additionally, more prevalent allelic changes in the infant are likely seeded by recombination events. Finally, there is evidence for ongoing transmission between mother and infants well after birth, which may also seed evolutionary changes in the infant.

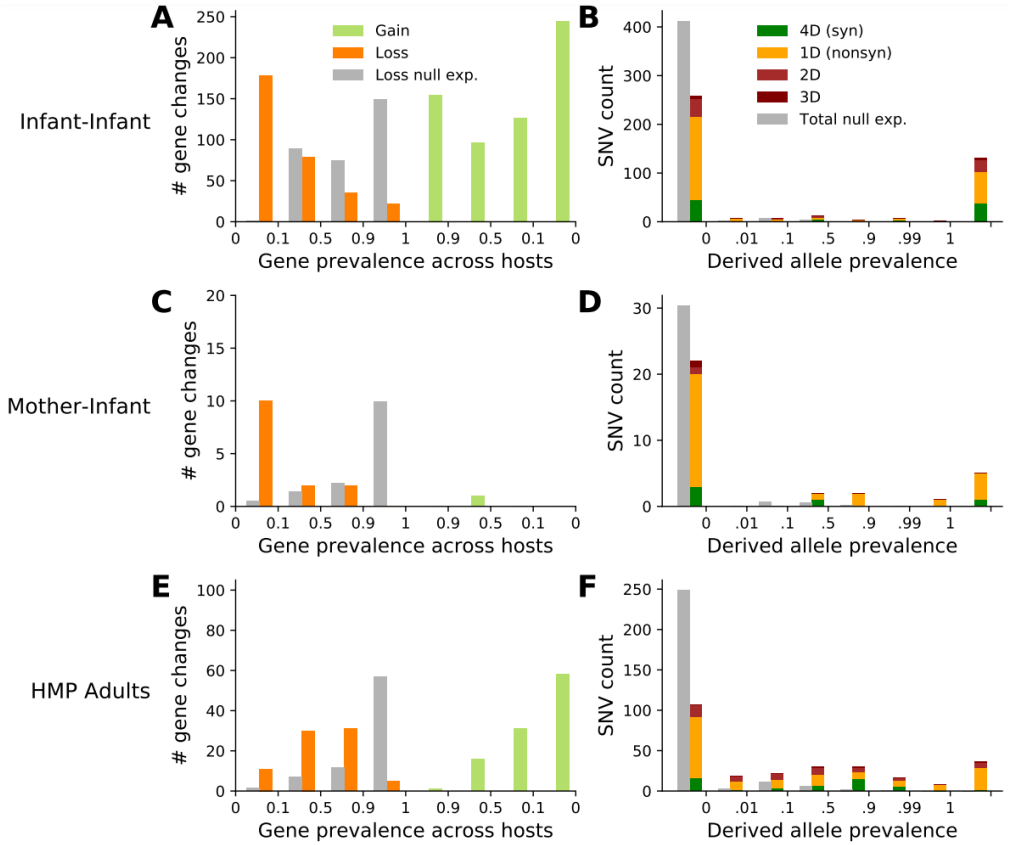

Figure 3. Prevalences of gene and SNV changes. Distribution of prevalences of gene (**A**, **C**, and **E**) and SNV changes (**B**, **D**, and **F**) occurring in infant-infant, mother-infant and HMP adult-adult QP pairs. Here, prevalence is defined as the fraction of HMP adults that harbor a SNV or gene. Additional cohorts are used to compute prevalence in **Figures S12, S13** and **S14**. The null distributions in grey show the expectation for random *de novo* events (**Methods**). By definition, gene gains cannot occur *de novo*. In B, D and F, the prevalence bin <0 indicates that no HMP adult harbors the sweeping allele other than the person in which the allele arises, and the prevalence bin >1 indicates that the allele is present in all HMP adults.

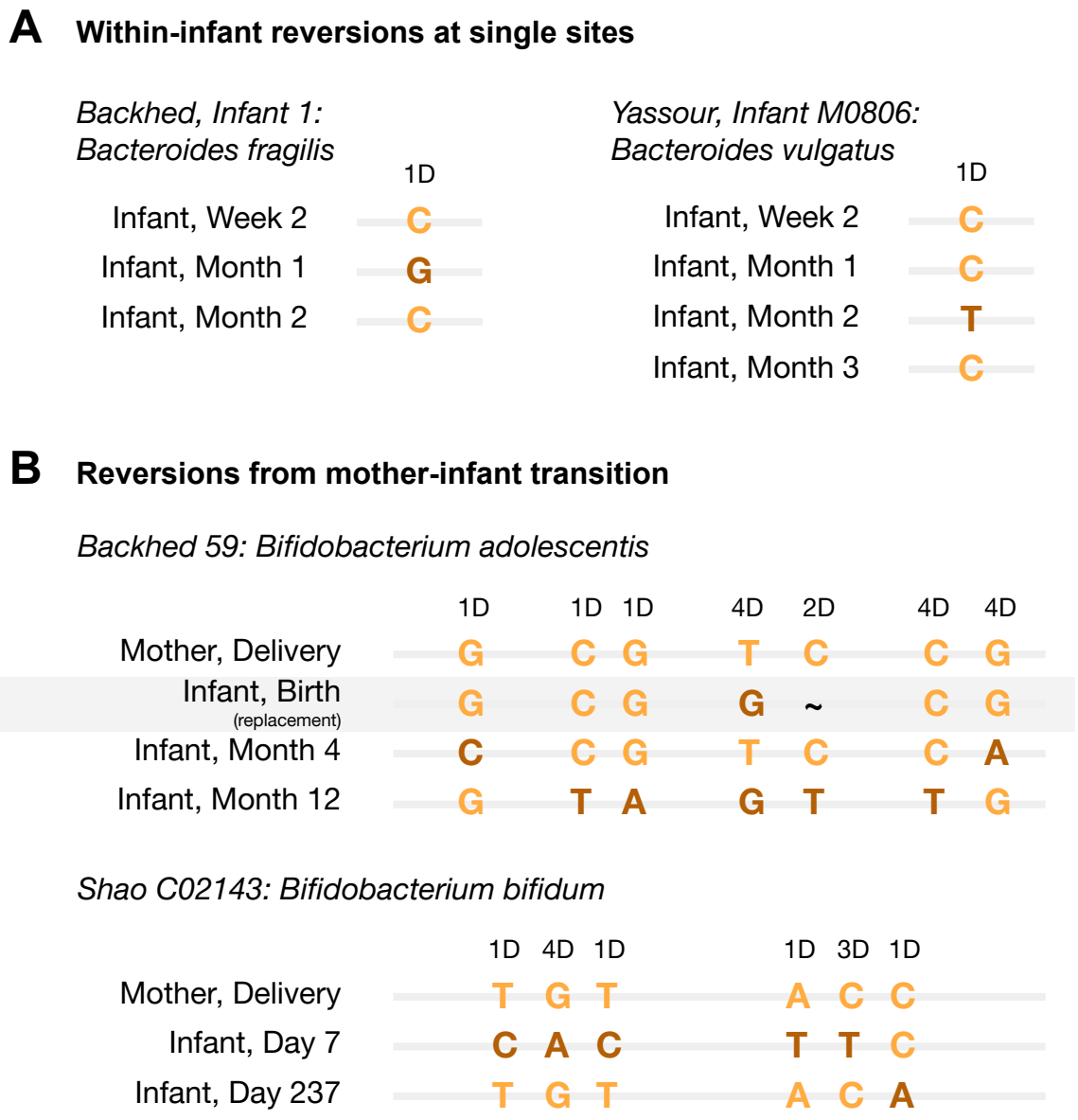

Figure 4: Reversion events in infant gut microbiomes. There were 371 host × species instances with at least 3 QP samples across available infant timepoints and mother at delivery. Of these, 13 experienced modifications at 2 nonoverlapping timepoint pairs. Seven of these 13 experienced reversions; here we illustrate the haplotypes corresponding to these reversions. Sites are ordered by gene's position in the reference genome. Sites are annotated as being in a 1D, 2D, 3D, or 4D position, which indicates codon degeneracy. For example, 1D indicates that the site is 1-fold degenerate (any nucleotide difference results in an amino acid change), whereas 4D indicates that the site is 4-fold degenerate (any nucleotide difference will not result in an amino acid change). A 2D and 3D site indicate that either 2 or 3 possible nucleotide changes, respectively, can be tolerated before the amino acid is changed. (**A**) In 5 of these cases, there is a reversion at a single nonsynonymous site to an allelic state that is prevalent in adults; two

examples are shown. (**B**) In two instances, multiple nucleotides change at both synonymous and nonsynonymous sites. In the host 'Backhed 59', a replacement occurs at birth with respect to mother, and then by month 4 it reverts back to the strain harbored by mother.

Parallelism of SNV changes across hosts

We next scanned for putative loci experiencing positive selection in the infant gut microbiome. To do so, we leveraged a signature of positive selection known as parallel evolution in which independent mutations in the same gene sweep to high frequency in multiple distinct hosts. Signatures of parallel evolution have been successfully used to detect positive selection in human-associated microbes (Lieberman et al. 2011; Zhao et al. 2019; Feder et al. 2021) as well as in laboratory experiments (Wichman et al. 1999; Woods et al. 2006; Barroso-Batista et al. 2014).

To identify candidates for parallel evolution, we counted the number of mutations that each gene class (defined as a unique PATRIC gene description, **Methods**) has accumulated across hosts and compared this number with that of a null in which the mutations are randomized across metagenomes across hosts (**Figure 5A**) (**Methods**). Under the null, fewer than 1 mutation is expected in a given gene class, reflecting the massive size of the metagenome relative to the number of evolutionary changes observed (**Table S3**). By contrast, several gene classes have >1 mutations in the observed data, sometimes occurring in a single host and sometimes in multiple hosts (**Table S3**). To bias our candidate list towards gene classes that are mutated recurrently in multiple hosts, we imposed a threshold requirement that gene classes must undergo an evolutionary modification in at least 4 distinct hosts. With these criteria, we found a total of 14 distinct gene classes (excluding hypothetical proteins) experiencing parallel evolution (**Figure 5B**). Some of these gene classes additionally acquired multiple mutations within hosts; for example, a TonB-dependent transporter part of the SusC family acquired a total of 32 SNV changes across 26 unique hosts (**Figure 5B)**.

The mutations in these 14 gene classes have a combined d_N/d_S of ~1.0. Although this value of d_N/d_S is consistent with a model without positive or negative selection, it is significantly higher than the d_N/d_S for other gene classes that are not as recurrently mutated (~0.6) (p value=0.001, permutation test), and is even greater than the d_N/d_S for SNVs that differ in replacement events (~0.1) (**Figure 5C**). This suggests that a higher proportion of mutations in

these recurrently evolving gene classes may be potentially adaptive.

Two of the top ranking gene classes have been previously discovered to undergo parallel evolution, most prominently genes in the starch utilization system (SusC and SusD) family (Zhao et al. 2019). Of notable interest is the previously mentioned TonB-dependent transporter part of the SusC family, which is implicated in processing of complex glycans (Martens et al. 2009) and degradation of breast milk-associated human milk oligosaccharides (Sela et al. 2008), is associated with the colonization of infant gut microbiomes (Yassour et al. 2018), and is posited to undergo adaptation in infant microbiomes (Kujawska et al. 2020; Sela et al. 2008). Another notable candidate gene class undergoing parallel evolution in our dataset includes beta-galactosidase, which was previously shown to be enriched in infants (Kujawska et al. 2020; Ambrogi et al. 2019; Sela et al. 2008; Duranti et al. 2019; Lawson et al. 2020) and is involved in the digestion of breast milk (Kitaoka 2012).

Other gene classes undergoing parallel evolution include transcriptional regulators, membrane transporters, and histidine kinases. In particular, the periplasmic ligand-binding sensor domain/histidine kinase is mutated 12 times in infants but never in adults. While, to our knowledge, these gene classes have not previously been implicated in microbiome adaptation or infant microbiomes, the recurrent evolutionary changes across multiple hosts provide interesting opportunities for follow up analyses.

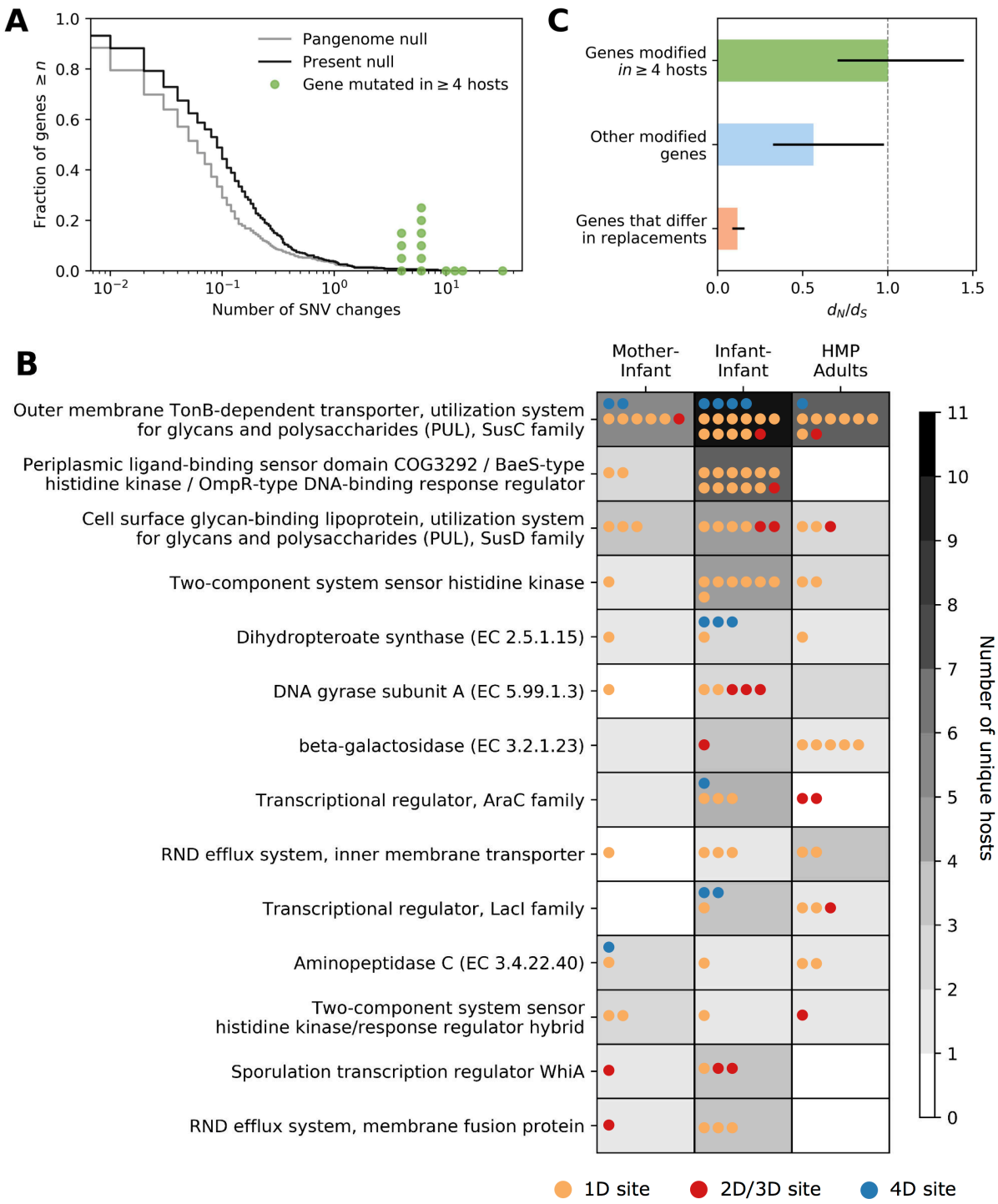

Figure 5: Parallelism of SNV changes across hosts. (A) Mutations are randomized across the metagenomes across hosts to construct two null distributions to assess parallelism (**Methods**). Observed SNV change counts are shown with orange dots for the 14 gene classes found to mutate in parallel in ≥4 hosts. In the cases where there are multiple orange dots, there are multiple gene classes experiencing the same number of SNV changes. **(B)** SNV changes in these 14 gene classes found to mutate in parallel in ≥4 hosts are partitioned by life stage and colored

by degeneracy (e.g. 1D indicates a nonsynonymous site and 4D indicates a synonymous site). Some functional classes are also mutated multiple times within hosts. For inclusivity, the mother-infant age class here includes all QP sample pairs in which the earliest infant timepoint is taken, irrespective of whether the infant was sampled in the first week of life. This avoids overlapping time points with the infant-infant age group. **(C)** d_N/d_S of SNV changes in the 14 gene classes found to mutate in parallel in ≥4 hosts compared with d_N/d_S of SNV changes in all other gene classes and d_N/d_S of sites that differ in strain replacements. 95% confidence intervals for bootstrapped d_N/d_S values are reported as black bars.

DISCUSSION

To date, there has been substantial focus on the ecological and functional development of the infant gut microbiome (Stewart et al. 2018; Bäckhed et al. 2015; Niu et al. 2020; Sprockett et al. 2020; Chu et al. 2017; Palmer et al. 2007; Vallès et al. 2014; Yassour et al. 2018; Ferretti et al. 2018; Koenig et al. 2011), but relatively little is known about evolutionary dynamics of microbiota in the infant gut microbiome. Here we examined the evolutionary dynamics of gut microbiota in a cohort of >700 infants (Yassour et al. 2018; Ferretti et al. 2018; Bäckhed et al. 2015; Shao et al. 2019) and compared them with that of adults. We found that the initial days after birth are marked by more than 10-fold increased rates of evolutionary modification and replacement of dominant resident strains relative to typical rates observed in adults. Over time, these rates settle, but not without rapid recovery of diversity initially lost during the mother to infant transmission period via elevated rates of gene gains in the first year of life. Many of these evolutionary events show signatures of *de novo* mutation that may potentially be adaptive, but ongoing transmission from the broader family also plays a significant role in seeding evolutionary changes in the infant gut microbiome over time.

Perhaps it is not surprising that infancy is marked by particularly volatile rates of evolutionary change and strain turnover that eventually stabilize with time, given that ecological succession at the species level is also rapid and dynamic in the early days of life. However, with a mechanistic understanding of how individual lineages change at a nucleotide level over time, we may be able to better understand the ecological succession in the infant gut microbiome. For example, some species such as *E. coli* undergo two orders of magnitude increase in nucleotide polymorphism levels over just a span of 1-2 months (**Figure 1**), while other species such as *B. vulgatus* and *E. faecalis* exhibit idiosyncratic patterns at different life stages (**Figures S4 and S5**). Community-wide level statistics like Shannon alpha diversity may miss these sharp

transitions in sub-species diversity because Shannon alpha diversity is computed as an aggregate of abundances across species. Species-specific signatures of ecological and evolutionary change at the nucleotide level may be important for uncovering functional shifts in the microbiome and understanding the roles each species plays in maturation of the human gut microbiome.

We note that several studies (e.g. Niu et al. 2020; Donovan 2020) use the term 'evolution' to describe ecological changes in species composition over time. In our paper, we define evolution as genetic change on the background of a resident lineage on top of which additional changes can accumulate, and we distinguish such evolutionary changes from strain replacement (Garud and Pollard 2020). By distinguishing evolution from strain replacement, we can start to understand at a molecular level the genetic variants needed for microbes to survive in the gut as well as the mechanisms through which they arise. Additionally, adaptation and co-evolution of infant gut bacteria have been described in the literature in reference to enrichment of genes that play a role in milk digestion in genera like *Bifidobacterium* (Milani et al. 2015; Duranti et al. 2019, 2017; Sela et al. 2008). Here, we do not examine longer-term evolutionary forces that result in functional enrichment in certain strains of gut bacteria, instead focusing on short term within-host changes. Despite this, among the genes experiencing parallel evolution across multiple infants are those implicated in milk digestion, suggesting that short-term forces within hosts could be contributing to signals of adaptation that accrue over longer periods.

Although our analysis distinguishes evolution from strain replacement, some evolutionary modifications detected in infants reflect migrations of lineages that may have evolved recently in the mother or another family member. Previous studies have found extensive evidence for transmission of strains from mother and infant as well as among other family members over time (Ferretti et al. 2018; Korpela et al. 2018; Koo et al. 2020; Yassour et al. 2018; Mitchell et al. 2020; Shao et al. 2019; Asnicar et al. 2017; Milani et al. 2015; Siranosian et al. 2020; Hildebrand et al. 2021). In our study, by analyzing QP samples, we were able to recover evidence of transmission of lineages diverged by only a few SNVs, reflecting evolutionary modifications that occurred in the recent past in the broader family unit. If families indeed circulate strains among each other over several decades, it may then be worth considering the broader family's microbiome as a larger, inter-connected microbiome. Additional data from multiple co-habiting individuals is needed to fully understand the extent of evolution and transmission of closely related lineages among family members (Korpela et al. 2018; Hildebrand

et al. 2021). Moreover, data from multiple body sites over multiple time points (Ferretti et al. 2018; Mitchell et al. 2020) can reveal additional sources of strains colonizing the infant.

It is possible that the genetic changes observed during the mother to infant transmission are confounded by yet another process: bottlenecks. During a bottleneck event, deleterious alleles can rise in frequency due to drift. However, recent work suggests that fluid flow in the adult colon cannot create bottlenecks that can result in large fluctuations in allele frequencies (Ghosh and Good 2021) and that, instead, natural selection is a more plausible mechanism of genetic change occurring over short time scales. Still, the mother-infant transition represents a unique colonization process in which a sterile or nearly sterile environment is colonized by microbes for the first time. Thus, the strength of bottleneck may be stronger in the infant than in established adult guts. In support of bottlenecks potentially playing a role in the evolutionary dynamics of the infant, we observe, in five host × species instances, private single nucleotide mutations at non synonymous sites reaching high frequency within infants that later revert back to the prevalent allelic state observed in adults. This suggests that the initial mutation was either temporarily adaptive or rose to high frequency due to drift and then reverted due to deleterious selective effects. There are, however, hallmarks of positive selection driving evolutionary change within infants. d_N/d_S of rare SNV changes in infants is >1, and many genes are mutated in parallel across multiple hosts, which also harbor elevated levels of d_N/d_S ~1 compared to other sweeping SNVs. Future work resolving the role of bottlenecks in the infant colonization process will be needed to fully understand the evolutionary processes taking place in infant microbiomes.

A future goal in the field is to understand the functional consequences of evolutionary changes in the gut microbiome. Although our focus was on understanding signatures of evolution over life stage, we did find that C-section infants have elevated rates of gene gains and losses compared to vaginally born infants (**Figure S11**). Given that C-section babies experience a disruption in transmission of microbes from their mothers (Mueller et al. 2015; Shao et al. 2019), the elevated rates of gene gains and loss in C-section infants pose the possibility that these microbes may need to evolve genetic adaptations to survive in the infant gut. Future work investigating the relative importance of evolutionary change in infants with different birth modes, as well as other attributes such as feeding mode and medication, will be needed to fully understand how the microbiome forms in these early days as well as their impact on infant health.

Another potential area for future work is to understand changes in the gut microbiome during pregnancy. In this paper, we computed species diversity in delivering mothers in four cohorts (Yassour et al. 2018; Ferretti et al. 2018; Bäckhed et al. 2015; Shao et al. 2019) and compared levels with that of healthy non-pregnant female adults in the HMP (Lloyd-Price et al. 2017; Methé et al. 2012) and (Qin et al. 2012) (**Figure S3**). We found that alpha diversity is significantly higher among mothers compared to non-pregnant females. The increase in alpha diversity is finding is consistent with previous findings by (Jašarević et al. 2017). However, (Goltsman et al. 2018) find that Shannon diversity decreases and (Koren et al. 2012) find that Faith's phylogenetic diversity decreases in pregnancy. Given the importance of being able to predict microbiome changes associated with pre-term birth (Vinturache et al. 2016) and the resulting impacts on the developing infant gut microbiome, future studies examining microbiome ecological and evolutionary biomarkers of pregnancy and delivery are needed.

The finding that many species in the infant gut microbiome rapidly evolve is important for understanding how the microbiome is assembled early in life. For example, it will be important to understand how these frequent evolutionary changes in the infant gut impact the persistence of lineages, ecological interactions, and the overall development of the gut microbiome. By incorporating evolution into our understanding of the development of the infant gut microbiome, we may be able to better understand the functional impact of the microbiome on human health.

METHODS

Data

The raw sequencing reads for the metagenomic samples used in this study were downloaded from Bäckhed et al. (2015) (accession number PRJEB6456); Ferretti et al. (2018) (accession number PRJNA352475); Yassour et al. (2018) (accession number PRJNA475246); Shao et al. (2019) (accession number PRJEB32631); The Human Microbiome Project Consortium 2012 and Lloyd-Price et al. (2017) (URL: https://aws.amazon.com/datasets/human-microbiome-project/); and Qin et al. (2012) (accession number PRJNA422434).

Estimation of species, gene, and SNV content of metagenomic samples

We used MIDAS (Metagenomic Intra-Species Diversity Analysis System, version 1.2, downloaded on November 21, 2016) (Nayfach et al. 2016) to estimate within-species nucleotide and gene content of each metagenomic shotgun sequencing sample. MIDAS utilizes a reference database of 31,007 bacterial genomes, clustered into 5,952 species, which covers roughly 50% of species found in "urban" human stool metagenomes. We followed the parameters described in Garud et al. (2019) and below to estimate species abundances, SNVs, and gene copy numbers variants (CNVs) with MIDAS:

Estimation of species content:

A major goal in this work is to infer evolutionary changes from metagenomic data. To do so, we mapped reads to call SNVs and CNVs to infer evolutionary changes. However, to avoid spurious inferences of allele frequency changes due to mismapping of reads to regions of the genome shared by multiple species, we constructed a personal reference database for each host comprised of the union of all species present at one or more timepoints. This per-host reference database was constructed to be as inclusive as possible to prevent reads from being "donated" to reference genome, while also being selective to prevent a reference genome from "stealing" reads from a species truly present.

To estimate the species abundances for each host × timepoint sample, we mapped the reads to a set of single copy marker genes (Wu et al. 2013; Nayfach et al. 2016)belonging to the 5,952 species. A species was considered present in a given sample if it had an average marker gene coverage ≥ 3. Next, we determined a single reference database for each host by including all species present at one or more timepoints with coverage ≥3.

Estimation CNV content:

We estimated CNV content by mapping reads to the pangenome for each species in each per-host reference database using Bowtie 2 (Langmead and Salzberg 2012) with default MIDAS settings (local alignment, MAPID≥94.0%, READQ≥20, and ALN_COV≥0.75). The average coverage for each gene was estimated by dividing the total number of reads mapped to a given gene by the gene length. Among these genes were a panel of universal single-copy genes. The copy number of a given gene (c) was then estimated by taking the ratio between its coverage and the median single marker gene coverage.

These copy number values were used estimate the prevalence of genes in the broader population, defined as the fraction of samples with copy number $c \leq 3$ and $c \geq 0.3$ (conditional on the mean single gene marker coverage being $\geq 5\times$). For each species, we computed "core genes", defined as genes in the MIDAS reference database that are present in at least 90% of samples within a given cohort (infants or adults). We separated the computation of core genes by cohort because several species that are prevalent in infants are absent in adults, and vice versa. In our computation of within-host polymorphism rates, we analyzed the union of core genes found in adults (mother + HMP) and infants.

Genes shared across species boundaries can result in read stealing and read donating, potentially confounding evolutionary inferences. Thus, we used a 'blacklist' of genes that are shared across species boundaries that was constructed in Garud et al. 2019. Briefly, this blacklist was constructed by using USEARCH (Edgar 2010) to cluster all genes in human-associated reference genomes with a 95% identity threshold. Additionally, since some genes may be absent from the MIDAS database that may also be shared across species boundaries, we implemented another filter in Garud et al. 2019 in which genes with $c \geq 3$ in at least one sample in our cohort was excluded from further analysis to avoid examining common genes.

Estimation of SNV content:

We next estimated SNV content. Below we describe the thresholds and parameters used in Garud et al. 2019 and in this paper. To call SNVs, we mapped reads to a single representative reference genome as per the default MIDAS software. Reads were mapped with Bowtie 2 (Langmead and Salzberg 2012), with default MIDAS mapping thresholds: global alignment, MAPID≥94.0%, READQ≥20, ALN_COV≥0.75, and MAPQ≥20. We excluded species from further analysis if reads mapped to ≤ 40% of their genome. We further excluded samples from further analysis if they had low median read coverage ($\overline{D}$) at protein coding sites. Specifically, samples with $\overline{D} < 5$ of across all protein coding sites with nonzero coverage were excluded.

Since a large component of the analyses performed here were to infer evolutionary changes between time points, additional bioinformatic filters were imposed. First, in addition, to excluding the blacklisted genes, to further avoid read stealing and donating from generating fluctuations in allele frequencies, we masked sites in a given sample if $D < 0.3\overline{D}$ or $D > 3\overline{D}$ as these sites harbor coverage anomalously low or high compared to the genome-wide average $\overline{D}$.

We also imposed another filter in which a SNV difference between two samples from the same host was called only if the successive values of $D/\overline{D}$ were within a factor of 3 as we expect that samples from the same host will have a smaller range of coverage fluctuations over time than samples from different hosts. An additional coverage threshold requirement of 20 reads /site was imposed for calling SNVs for analyses below.

Quasi-phasing and inference of rates of SNV changes, gene changes, and replacements over time

We followed the approach in Garud et al. (2019) to identify "quasi-phaseable" (QP) samples (see **SI Text 1** for an explanation of QP samples). We then identified SNV and gene changes between pairs of QP samples collected from consecutive time points with available data from the same host (see **SI Text 1** for the expected rate of false positive SNV changes due to sampling error.). A consecutive time point consists of two time points in the mother-infant dyad that are both QP for a given species such that there is no intervening timepoint that is also QP. Note that "mother-infant" timepoint pairs were restricted to comparisons between mother timepoints at delivery and infant timepoints within the first week in order to best approximate gut microbiome dynamics of the mother-infant transition at birth and exclude changes that occur later on within the infant. The only exception to this "mother-infant" definition is in **Figure 5B**, in which the mother-infant age class includes all QP sample pairs comprising mothers and the earliest infant timepoint, irrespective of whether the infant was sampled in the first week of life. This enables all data to be considered for the parallelism analysis.

SNV changes were computed by identifying SNV allele frequency changes from ≤ 0.2 to ≥ 0.8. With these strict thresholds as described in **SI Text 1**, << 1 SNV change per genome is expected by chance due to sampling error. Since SNV changes can be generated by either modification or replacement events, we classified QP pairs as undergoing a modification or no change if they had <= 20 SNV differences, or a replacement if they had >=500 SNV differences.

Additionally, we assessed evolutionary gene gains and losses between timepoints not experiencing replacements. These were computed by identifying genes with copy number $c \leq 0.05$ (indicating gene absence) in one sample and $0.6 \leq c \leq 1.2$ in another (indicating single copy gene presence), reflecting parameters used in Garud et al. 2019.

Per-day rates of SNV changes, replacements, and gene changes

To quantify per-day modification rates, we divided the number of SNV changes and gene changes, respectively, observed among non-replacement QP pairs by the number of days elapsed between sampling time points. To quantify per-day replacement rates, we divided the number of replacement events by the number of available QP pairs and the number of days elapsed between sampling timepoints. In the case of mother-infant time point pairs, while mothers were generally sampled around the time of birth, some were sampled within a few days (up to a week) after birth in the Backhed dataset. In cases where the mother's exact time of sampling was known, the sum of time since birth for the mother and infant samples was used as time elapsed between sampling timepoints. For the Shao, Ferretti, and Yassour datasets, the mother was assumed to have been sampled on the day of birth.

To assess variance, we performed bootstrapping of rates by repeatedly taking random subsamples of size 50 (with replacement) of all QP pairs in a category. In each bootstrap, the total number of SNV changes, gene changes, and replacements, respectively, were summed across the samples, and divided by the total duration elapsed between time point pairs for the sample.

To quantify the scaling relationship between rate of change versus life stage, we performed a linear regression on logged mean SNV change, replacement and gene gain/loss rates computed per timepoint pair category as a function of days since birth. To compute days since birth, the median timepoint was computed for each life stage (mother-infant within the first week, day 0-week 1, week 1-month 1, month 1-year 1). In the case of HMP adults, a value of 40 years was assigned, reflecting the age range of HMP participants (Lloyd-Price et al. 2017; Methé et al. 2012).

Permutation tests

To assess whether rates of SNV changes, gene changes or replacements are significantly different between age categories (**Figures 2** and **S10**, **Table S2**), or between C-section versus vaginally born infants (**Figure S11**), or between formula versus breast fed infants (**Figure S11**), we performed permutation tests. For the permutation tests, labels (e.g. C-section versus vaginal, or infant versus adult) were shuffled for 10,000 trials and in each trial the difference in rates between the two categories being considered was computed. The resulting *p* value was computed

by assessing the quantile of the observed difference in rates from the distribution generated from permuted data.

Prevalences of sweeping alleles

Derived allele prevalences were computed with respect to three prevalence cohorts: HMP adults, infants, and mothers (**Figures 3, S13, S14**). Replicating the analysis in Garud et al. (2019), we define population prevalence of an allele as the fraction samples where the majority of the reads at a given site (minimum coverage of at least 20×) harbor the allele. If a host had multiple time points, that host's contribution to total prevalence was the fraction of timepoints possessing the allele. Private SNVs were assigned a prevalence of 0. We computed a null distribution for SNV change prevalences by randomly drawing the number of SNV modifications observed in data from all SNV opportunities in the genome, bootstrapped 10 times.

Prevalences of gene gains and losses

Gene prevalences were computed with respect to three prevalence cohorts of HMP adults, infants and mothers. As described above we defined population prevalence of a gene as the proportion of all samples harboring the gene with copy number ≤ 3 and ≥ 0.3; We computed a null distribution for prevalences of gene losses by randomly drawing the number of genes lost from all genes present in the pangenome, bootstrapped 10 times. The null expectation for gene gains was zero across prevalence bins as there are no *de novo* gene gains by definition.

d_N/d_S computation for changing SNVs

To investigate whether or not SNV changes in modification events are adaptive, we estimated d_N/d_S. d_N/d_S was computed as the ratio between number of observed nonsynonymous (1D) SNV changes divided by nonsynonymous opportunities and number of observed synonymous (4D) SNV changes divided by nonsynonymous opportunities per QP pair under consideration. We also bootstrapped d_N/d_S estimates by sampling a binomial distribution 10,000 times with number of trials n equal to total number of nonsynonymous or synonymous SNV changes and success probability p equal to the proportion that are either nonsynonymous or synonymous; this resulted in bootstrapped nonsynonymous and synonymous SNV change counts

which were divided by the same nonsynonymous and synonymous SNV opportunities. We reported 95% confidence interval for d_N/d_S estimates by the 2.5% and 97.5% quantiles of the bootstrapped d_N/d_S values.

Parallelism of evolutionary SNV changes across hosts

To assess parallelism of evolutionary changes across hosts, we enumerated the number of observed SNV changes per PATRIC gene description. A PATRIC gene description is a unique string that describes a gene. This same string can potentially be present across multiple species (**Table S3**), though as described above, genes shared across species boundaries (i.e. possess >=95% nucleotide identity with a gene in a different species in the MIDAS database) were filtered from our dataset.

We evaluated whether the observed number of SNV changes per gene class is greater than expected under two null distributions. The two null distributions were constructed as follows: the number of observed changes were randomly distributed across (1) all genes present at either timepoint for a given QP pair and (2) and all genes that present in the MIDAS pangenome for the species harboring the SNV changes. We bootstrapped the null distribution estimates 100 times.

Software availability: All computer code for this paper is available at https://github.com/garudlab/mother_infant and the associated Supplemental Code file.

Competing interest statement: The authors declare no competing financial interests.

ACKNOWLEDGEMENTS

We sincerely thank William R. Shoemaker, Richard Wolff, Alison Feder, Sandeep Venkataram, and Leah Briscoe for their critical comments on the manuscript and Benjamin Good and Katherine Pollard for early discussions on this project. We thank Michael Tsiang for his advice on statistical tests. We thank members of the Garud lab for their feedback during the development of this paper. NRG is received support from the Paul Allen Frontiers Group, a University of California Hellman fellowship, a UCLA Faculty Career Development award, and the Research Corporation for Science Advancement. DWC received funding support from NIH R25 MH 109172.

REFERENCES

Ambrogi V, Bottacini F, O'Sullivan J, O'Connell Motherway M, Linqiu C, Schoemaker B, Schoterman M, van Sinderen D. 2019. Characterization of GH2 and GH42 β-galactosidases derived from bifidobacterial infant isolates. *AMB Express* **9**.

Asnicar F, Manara S, Zolfo M, Truong DT, Scholz M, Armanini F, Ferretti P, Gorfer V, Pedrotti A, Tett A, et al. 2017. Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling. *mSystems* **2**.

Bäckhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, Li Y, Xia Y, Xie H, Zhong H, et al. 2015. Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe*.

Barrick JE, Lenski RE. 2013. Genome dynamics during experimental evolution. *Nat Rev Genet* **14**: 827–839.

Barroso-Batista J, Sousa A, Lourenço M, Bergman ML, Sobral D, Demengeot J, Xavier KB, Gordo I. 2014. The First Steps of Adaptation of Escherichia coli to the Gut Are Dominated by Soft Sweeps. *PLoS Genet* **10**: e1004182. http://www.ncbi.nlm.nih.gov/pubmed/24603313.

Bokulich NA, Chung J, Battaglia T, Henderson N, Jay M, Li H, Lieber AD, Wu F, Perez-Perez GI, Chen Y, et al. 2016. Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Sci Transl Med* **8**.

Brooks B, Olm MR, Firek BA, Baker R, Thomas BC, Morowitz MJ, Banfield JF. 2017. Strain-resolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome. *Nat Commun* **8**.

Chu DM, Ma J, Prince AL, Antony KM, Seferovic MD, Aagaard KM. 2017. Maturation of the infant microbiome community structure and function across multiple body sites and in relation to mode of delivery. *Nat Med* **23**.

Coyne MJ, Zitomersky NL, McGuire AM, Earl AM, Comstock LE. 2014. Evidence of extensive DNA transfer between bacteroidales species within the human gut. *MBio* **5**: e01305-14. http://www.ncbi.nlm.nih.gov/pubmed/24939888.

Didelot X, Eyre DW, Cule M, Ip CLC, Ansari MA, Griffiths D, Vaughan A, O'Connor L, Golubchik T, Batty EM, et al. 2012. Microevolutionary analysis of Clostridium difficile genomes to investigate transmission. *Genome Biol* **13**.

Dominguez-Bello MG, Costello EK, Contreras M, Magris M, Hidalgo G, Fierer N, Knight R. 2010. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc Natl Acad Sci U S A* **107**.

Donovan SM. 2020. Evolution of the gut microbiome in infancy within an ecological context. *Curr Opin Clin Nutr Metab Care* **23**.

Duranti S, Lugli GA, Mancabelli L, Armanini F, Turroni F, James K, Ferretti P, Gorfer V, Ferrario C, Milani C, et al. 2017. Maternal inheritance of bifidobacterial communities and bifidophages in infants through vertical transmission. *Microbiome* **5**.

Duranti S, Lugli GA, Milani C, James K, Mancabelli L, Turroni F, Alessandri G, Mangifesta M, Mancino W, Ossiprandi MC, et al. 2019. Bifidobacterium bifidum and the infant gut microbiota: an intriguing case of microbe-host co-evolution. *Environ Microbiol* **21**.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**.

Feder AF, Pennings PS, Petrov DA. 2021. The clarifying role of time series data in the

population genetics of HIV. *PLoS Genet* **17**.
Fehr K, Moossavi S, Sbihi H, Boutin RCT, Bode L, Robertson B, Yonemitsu C, Field CJ, Becker AB, Mandhane PJ, et al. 2020. Breastmilk Feeding Practices Are Associated with the Co-Occurrence of Bacteria in Mothers' Milk and the Infant Gut: the CHILD Cohort Study. *Cell Host Microbe* **28**.
Ferretti P, Pasolli E, Tett A, Asnicar F, Gorfer V, Fedi S, Armanini F, Truong DT, Manara S, Zolfo M, et al. 2018. Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. *Cell Host Microbe* **24**: 133-145.e5.
Garud NR, Good BH, Hallatschek O, Pollard KS. 2019. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLoS Biol* **17**: e3000102. http://www.ncbi.nlm.nih.gov/pubmed/30673701.
Garud NR, Pollard KS. 2020. Population Genetics in the Human Microbiome. *Trends Genet* **36**: 53–67.
Ghalayini M, Launay A, Bridier-Nahmias A, Clermont O, Denamur E, Lescat M, Tenaillona O. 2018. Evolution of a dominant natural isolate of Escherichia coli in the human gut over the course of a year suggests a neutral evolution with reduced effective population size. *Appl Environ Microbiol* **84**.
Ghosh OM, Good BH. 2021. Emergent evolutionary forces in spatial models of luminal growth in the human gut microbiota. *bioRxiv*.
Goltsman DSA, Sun CL, Proctor DM, DiGiulio DB, Robaczewska A, Thomas BC, Shaw GM, Stevenson DK, Holmes SP, Banfield JF, et al. 2018. Metagenomic analysis with strain-level resolution reveals fine-scale variation in the human pregnancy microbiome. *Genome Res* **28**.
Groussin M, Poyet M, Sistiaga A, Kearney SM, Moniz K, Noel M, Hooker J, Gibbons SM, Segurel L, Froment A, et al. 2021. Elevated rates of horizontal gene transfer in the industrialized human microbiome. *Cell* **184**.
Gumpert H, Kubicek-Sutherland JZ, Porse A, Karami N, Munck C, Linkevicius M, Adlerberth I, Wold AE, Andersson DI, Sommer MOA. 2017. Transfer and persistence of a multi-drug resistance plasmid in situ of the infant gut microbiota in the absence of antibiotic treatment. *Front Microbiol* **8**.
Hehemann JH, Correc G, Barbeyron T, Helbert W, Czjzek M, Michel G. 2010. Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464**: 908–912.
Hildebrand F, Gossmann TI, Frioux C, Özkurt E, Myers PN, Ferretti P, Kuhn M, Bahram M, Nielsen HB, Bork P. 2021. Dispersal strategies shape persistence and evolution of human gut bacteria. *Cell Host Microbe* **29**.
Jašarević E, Howard CD, Misic AM, Beiting DP, Bale TL. 2017. Stress during pregnancy alters temporal and spatial dynamics of the maternal and offspring microbiome in a sex-specific manner. *Sci Rep* **7**.
Jiang X, Brantley Hall A, Arthur TD, Plichta DR, Covington CT, Poyet M, Crothers J, Moses PL, Tolonen AC, Vlamakis H, et al. 2019. Invertible promoters mediate bacterial phase variation, antibiotic resistance, and host adaptation in the gut. *Science (80-)* **363**.
Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B, Nielsen J, Bäckhed F. 2013. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498**: 99–103. http://www.ncbi.nlm.nih.gov/pubmed/23719380.
Kenny DJ, Plichta DR, Shungin D, Koppel N, Hall AB, Fu B, Vasan RS, Shaw SY, Vlamakis H,

Balskus EP, et al. 2020. Cholesterol Metabolism by Uncultured Human Gut Bacteria Influences Host Cholesterol Level. *Cell Host Microbe* **28**: 245-257.e6.
Kitaoka M. 2012. Bifidobacterial enzymes involved in the metabolism of human milk oligosaccharides. *Adv Nutr* **3**.
Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R, Angenent LT, Ley RE. 2011. Succession of microbial consortia in the developing infant gut microbiome. *Proc Natl Acad Sci U S A* **108**.
Koo H, McFarland BC, Hakim JA, Crossman DK, Crowley MR, Rodriguez JM, Benveniste EN, Morrow CD. 2020. An individualized mosaic of maternal microbial strains is transmitted to the infant gut microbial community. *R Soc Open Sci* **7**.
Koren O, Goodrich JK, Cullender TC, Spor A, Laitinen K, Kling Bäckhed H, Gonzalez A, Werner JJ, Angenent LT, Knight R, et al. 2012. Host remodeling of the gut microbiome and metabolic changes during pregnancy. *Cell* **150**.
Korpela K, Costea P, Coelho LP, Kandels-Lewis S, Willemsen G, Boomsma DI, Segata N, Bork P. 2018. Selective maternal seeding and environment shape the human gut microbiome. *Genome Res* **28**: 561–568. http://www.ncbi.nlm.nih.gov/pubmed/29496731.
Kujawska M, La Rosa SL, Roger LC, Pope PB, Hoyles L, McCartney AL, Hall LJ. 2020. Succession of Bifidobacterium longum Strains in Response to a Changing Early Life Nutritional Environment Reveals Dietary Substrate Adaptations. *iScience* **23**.
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**.
Lawson MAE, O'Neill IJ, Kujawska M, Gowrinadh Javvadi S, Wijeyesekera A, Flegg Z, Chalklen L, Hall LJ. 2020. Breast milk-derived human milk oligosaccharides promote Bifidobacterium interactions within a single ecosystem. *ISME J* **14**: 635–648.
Lieberman TD, Michel JB, Aingaran M, Potter-Bynoe G, Roux D, Davis MR, Skurnik D, Leiby N, Lipuma JJ, Goldberg JB, et al. 2011. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nat Genet* **43**: 1275–1280.
Lin M, Kussell E. 2019. Inferring bacterial recombination rates from large-scale sequencing datasets. *Nat Methods* **16**: 199–204.
Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, Brady A, Creasy HH, McCracken C, Giglio MG, et al. 2017. Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**: 61–66.
Martens EC, Koropatkin NM, Smith TJ, Gordon JI. 2009. Complex glycan catabolism by the human gut microbiota: The bacteroidetes sus-like paradigm. *J Biol Chem* **284**.
Methé BA, Nelson KE, Pop M, Creasy HH, Giglio MG, Huttenhower C, Gevers D, Petrosino JF, Abubucker S, Badger JH, et al. 2012. A framework for human microbiome research. *Nature* **486**: 215–221. http://www.ncbi.nlm.nih.gov/pubmed/22699610.
Milani C, Mancabelli L, Lugli GA, Duranti S, Turroni F, Ferrario C, Mangifesta M, Viappiani A, Ferretti P, Gorfer V, et al. 2015. Exploring vertical transmission of bifidobacteria from mother to child. *Appl Environ Microbiol*.
Mitchell C, Hogstrom L, Bryant A, Bergerat A, Cher A, Pochan S, Herman P, Carrigan M, Sharp K, Huttenhower C, et al. 2020. Delivery mode impacts newborn gut colonization efficiency. *Cell Reports Med*.
Mueller NT, Bakacs E, Combellick J, Grigoryan Z, Dominguez-Bello MG. 2015. The infant microbiome development: Mom matters. *Trends Mol Med* **21**.
Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. 2016. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and

biogeography. *Genome Res* **26**: 1612–1625.
Nicholson JK, Holmes E, Kinross J, Burcelin R, Gibson G, Jia W, Pettersson S. 2012. Host-gut microbiota metabolic interactions. *Science (80-)* **336**: 1262–1267. http://www.ncbi.nlm.nih.gov/pubmed/22674330.
Niu J, Xu L, Qian Y, Sun Z, Yu D, Huang J, Zhou X, Wang Y, Zhang T, Ren R, et al. 2020. Evolution of the Gut Microbiome in Early Childhood: A Cross-Sectional Study of Chinese Children. *Front Microbiol* **11**.
Palmer C, Bik EM, DiGiulio DB, Relman DA, Brown PO. 2007. Development of the human infant intestinal microbiota. *PLoS Biol* **5**.
Post DM, Palkovacs EP. 2009. Eco-evolutionary feedbacks in community and ecosystem ecology: Interactions between the ecological theatre and the evolutionary play. *Philos Trans R Soc B Biol Sci* **364**.
Poyet M, Groussin M, Gibbons SM, Avila-Pacheco J, Jiang X, Kearney SM, Perrotta AR, Berdy B, Zhao S, Lieberman TD, et al. 2019. A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nat Med* **25**: 1442–1452.
Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D, et al. 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**: 55–60. http://www.ncbi.nlm.nih.gov/pubmed/23023125.
Raveh-Sadka T, Firek B, Sharon I, Baker R, Brown CT, Thomas BC, Morowitz MJ, Banfield JF. 2016. Evidence for persistent and shared bacterial strains against a background of largely unique gut colonization in hospitalized premature infants. *ISME J* **10**.
Roodgar M, Good BH, Garud NR, Martis S, Avula M, Zhou W, Lancaster SM, Lee H, Babveyh A, Nesamoney S, et al. 2021. Longitudinal linked-read sequencing reveals ecological and evolutionary responses of a human gut microbiome during antibiotic treatment. *Genome Res* **31**.
Sakoparnig T, Field C, van Nimwegen E. 2021. Whole genome phylogenies reflect long-tailed distributions of recombination rates in many bacterial species. *Elife*.
Sela DA, Chapman J, Adeuya A, Kim JH, Chen F, Whitehead TR, Lapidus A, Rokhsar DS, Lebrilla CB, German JB, et al. 2008. The genome sequence of Bifidobacterium longum subsp. infantis reveals adaptations for milk utilization within the infant microbiome. *Proc Natl Acad Sci U S A* **105**: 18964–18969.
Sender R, Fuchs S, Milo R. 2016. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol* **14**.
Shao Y, Forster SC, Tsaliki E, Vervier K, Strang A, Simpson N, Kumar N, Stares MD, Rodger A, Brocklehurst P, et al. 2019. Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature* **574**: 117–121.
Siranosian BA, Tamburini FB, Sherlock G, Bhatt AS. 2020. Acquisition, transmission and strain diversity of human gut-colonizing crAss-like phages. *Nat Commun* **11**.
Sjögren YM, Tomicic S, Lundberg A, Böttcher MF, Björkstén B, Sverremark-Ekström E, Jenmalm MC. 2009. Influence of early gut microbiota on the maturation of childhood mucosal and systemic immune responses: Gut microbiota and immune responses. *Clin Exp Allergy* **39**: 1842–1851.
Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. 2011. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**: 241–244. http://www.ncbi.nlm.nih.gov/pubmed/22037308.

Smith PM, Howitt MR, Panikov N, Michaud M, Gallini CA, Bohlooly-y M, Glickman JN, Garrett WS. 2013. The Microbial Metabolites, Short-Chain Fatty Acids, Regulate Colonic Treg Cell Homeostasis. *Science (80-)* **341**.

Spanogiannopoulos P, Bess EN, Carmody RN, Turnbaugh PJ. 2016. The microbial pharmacists within us: A metagenomic view of xenobiotic metabolism. *Nat Rev Microbiol* **14**: 273–287.

Sprockett DD, Martin M, Costello EK, Burns AR, Holmes SP, Gurven MD, Relman DA. 2020. Microbiota assembly, structure, and dynamics among Tsimane horticulturalists of the Bolivian Amazon. *Nat Commun* **11**.

Stewart CJ, Ajami NJ, O'Brien JL, Hutchinson DS, Smith DP, Wong MC, Ross MC, Lloyd RE, Doddapaneni HV, Metcalf GA, et al. 2018. Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* **562**.

The Human Microbiome Project Consortium. 2012. A framework for human microbiome research. *Nature* **486**.

Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* **12**: 902–903. http://www.ncbi.nlm.nih.gov/pubmed/26418763.

Vallès Y, Artacho A, Pascual-García A, Ferrús ML, Gosalbes MJ, Abellán JJ, Francino MP. 2014. Microbial Succession in the Gut: Directional Trends of Taxonomic and Functional Change in a Birth Cohort of Spanish Infants. *PLoS Genet* **10**.

Van Dijk M, Hunnius S, Van Geert P. 2012. The dynamics of feeding during the introduction to solid food. *Infant Behav Dev* **35**.

Venkataram S, Kuo H-Y, Hom EFY, Kryazhimskiy S. 2021. Early adaptation in a microbial community is dominated by mutualism-enhancing mutations. *bioRxiv*.

Vinturache AE, Gyamfi-Bannerman C, Hwang J, Mysorekar IU, Jacobsson B. 2016. Maternal microbiome - A pathway to preterm birth. *Semin Fetal Neonatal Med* **21**.

Wichman HA, Badgett MR, Scott LA, Boulianne CM, Bull JJ. 1999. Different trajectories of parallel evolution during viral adaptation. *Science (80-)* **285**.

Woods R, Schneider D, Winkworth CL, Riley MA, Lenski RE. 2006. Tests of parallel molecular evolution in a long-term experiment with Escherichia coli. *Proc Natl Acad Sci U S A* **103**.

Wu D, Jospin G, Eisen JA. 2013. Systematic Identification of Gene Families for Use as "Markers" for Phylogenetic and Phylogeny-Driven Ecological Studies of Bacteria and Archaea and Their Major Subgroups. *PLoS One* **8**.

Yaffe E, Relman DA. 2020. Tracking microbial evolution in the human gut using Hi-C reveals extensive horizontal gene transfer, persistence and adaptation. *Nat Microbiol* **5**.

Yassour M, Jason E, Hogstrom LJ, Arthur TD, Tripathi S, Siljander H, Selvenius J, Oikarinen S, Hyöty H, Virtanen SM, et al. 2018. Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. *Cell Host Microbe* **24**: 146-154.e4.

Yassour M, Vatanen T, Siljander H, Hämäläinen AM, Härkönen T, Ryhänen SJ, Franzosa EA, Vlamakis H, Huttenhower C, Gevers D, et al. 2016. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Sci Transl Med* **8**.

Zhao S, Lieberman TD, Poyet M, Kauffman KM, Gibbons SM, Groussin M, Xavier RJ, Alm EJ. 2019. Adaptive Evolution within Gut Microbiomes of Healthy People. *Cell Host Microbe* **25**: 656-667.e8.

Zlitni S, Bishara A, Moss EL, Tkachenko E, Kang JB, Culver RN, Andermann TM, Weng Z, Wood C, Handy C, et al. 2020. Strain-resolved microbiome sequencing reveals mobile

elements that drive bacterial competition on a clinical timescale. *Genome Med* **12**.