Selective maternal seeding and environment shape the human gut microbiome

Katri Korpela[1,2], Paul Costea[1], Luis Pedro Coelho[1], Stefanie Kandels-Lewis[1], Gonneke Willemsen[3], Dorret I Boomsma[3], Nicola Segata[4], Peer Bork[1]

[1] European Molecular Biology Laboratory, Heidelberg, Germany
[2] Department of Bacteriology and Immunology, Immunobiology Research Program, University of Helsinki, Helsinki, Finland
[3] Department of Biological Psychology, Vrije Universiteit, Amsterdam, The Netherlands
[4] Centre for Integrative Biology, University of Trento, Trento, Italy

Correspondence:
Peer Bork, bork@embl.de, Meyerhofstraße 1, 69117 Heidelberg, Germany

Running title: Maternal seeding of the human gut microbiome

Keywords: metagenomics, microbiota, strain, transmission, single nucleotide variant

Vertical transmission of bacteria from mother to infant at birth is postulated to initiate a life-long host-microbe symbiosis, playing an important role in early infant development. However, only the tracking of strictly defined unique microbial strains can clarify where the intestinal bacteria come from, how long the initial colonisers persist and whether colonisation by other strains from the environment can replace existing ones. Using rare single nucleotide variants in faecal metagenomes of infants and their family members, we show strong evidence of selective and persistent transmission of maternal strain populations to the vaginally born infant, and their occasional replacement by strains from the environment, including those from family members, in later childhood. Only strains from the classes Actinobacteria and Bacteroidia, which are essential components of the infant microbiome, are transmitted from the mother and persist for at least one year. In contrast, maternal strains of Clostridia, a dominant class in the mother's gut microbiome, are not observed in the infant. Caesarean-born infants show a striking lack of maternal transmission at birth. After the first year, strain influx from the family environment occurs, and continues even in adulthood. Fathers appear to be more frequently donors of novel strains to other family members than receivers. Thus, the infant gut is seeded by selected maternal bacteria, which expand to form a stable community, with a rare but stable continuing strain influx over time.

Introduction

The infant's intestinal microbiota are believed to guide the development of the host (Houghteling and Walker 2015 and references therein), and to be transmitted from the mother at birth (Cho and Blaser 2012; Funkhouser and Bordenstein 2013; Bäckhed et al. 2015; Dominguez-Bello et al. 2015). Transmission of bacteria is even suggested to represent a form of epigenetic inheritance (Gilbert 2014). The infant's immune system develops in tight interaction with the intestinal microbiome, which guides its maturation and may significantly influence the risk of immune-related diseases (Houghteling and Walker 2015). The early gut colonisers are therefore thought to play a fundamental role in the long-term health of the child (Houghteling and Walker 2015). However, despite the consensus view of maternal inheritance of the microbiome at birth, the source of the early colonisers has never been exhaustively studied. Targeted cultivation-based investigations have shown that mother-infant pairs often share strains of bifidobacteria and lactobacilli (Tannock et al. 1990; Matsumiya et al. 2002; Jiménez et al. 2008; Albesharat et al. 2011; Makino et al. 2013; Jost et al. 2014). Large scale 16S rRNA and metagenomic surveys have found that vaginally born infants often harbour species that can also be detected in the mother (Bäckhed et al. 2015; Dominguez-Bello et al. 2015), thus hinting at broad vertical transmission of the microbiota, but the same species can also be shared by unrelated individuals. Based on the taxonomic profile of the infant gut, it is clear that vaginal microbiota, consisting mostly of lactobacilli (Ravel et al. 2011), cannot account for the majority of infant gut species. The same is true for breast milk, as the overlap with infant gut taxa is minor (Asnicar et al. 2017). The most likely source of bacteria to the infant is the mother's gut, which contains most of the species present in the infant gut, albeit at different relative abundances (Bäckhed et al. 2015; Asnicar et al. 2017).

In order to track bacterial transmission events, high taxonomic resolution at the genomic level is required (Schloissnig et al. 2013; Zhu et al. 2015; Li et al. 2016; Nayfach et al. 2016; Truong et al. 2017). The nucleotide variation of bacterial genomes allows high-resolution fingerprinting of bacteria shared between individuals or within individuals over time. However, detecting the same single nucleotide variants (SNV) in two samples in not evidence of transmission as common SNVs (and common strains) can prevail in the population (Yassour et al. 2016). Only the detection of rare variants shared exclusively between two individuals assures true transmission. Recently, the ability of tracking maternal transmission and intra-individual stability of strain populations, measured as shared SNVs in metagenomics data, has been demonstrated (Nayfach et al. 2016; Yassour et al. 2016; Asnicar et al. 2017). As the infant's intestinal microbial composition changes considerably during the first year of life (Bäckhed et al. 2015) and beyond, the fate of maternal strains and the impact of family members or environment in general still remain to be uncovered. Recent data suggest that some postnatal colonisation from the environment may occur (Yassour et al. 2016), but it is currently unknown if there is a restricted age window, during which colonisation is possible. We here compiled a cohort of family members that allowed us to assess strain persistence and intra-family strain transmissions at birth and later in life, identifying strain transmission at different ages up to adulthood.

Results

Tracking strain sharing between metagenomes using rare marker SNVs

To identify the source and persistence of bacterial strains in children and their family members, we adapted genomic variation approaches (Schloissnig et al. 2013; Li et al. 2016) to monitor SNVs in faecal metagenomes from several European and North American cohorts, totalling 695 samples from 307 individuals in 159 families with children and adults of different ages, sampled at multiple time points (Supplemental Table S1). The cohorts included 139 infants of whom 25 were born by Caesarean section (c-section). We collected publicly available metagenomic data on Swedish, Italian, and US mother-infant pairs (Bäckhed et al. 2015; Asnicar et al. 2017; Chu et al. 2017) and unrelated adults (Human Microbiome Project Consortium 2012; Le Chatelier et al. 2013; Zeller et al. 2014) from the US, France, Denmark and Spain, and amended these data sets with newly generated data from 10 German and Dutch families (of European ethnicity) with children of various ages (Supplemental Table S2) to assess potential transmission between family members at different stages in life. In addition, we profiled publically available data of 139 unrelated US adults, followed in time over one year (Human Microbiome Project Consortium 2012). The number of compared pairs of individuals by relation is presented in Supplemental Table S3. Although we pooled data from different cohorts, we did not find clear batch effects in the observed species composition (Supplemental Fig. S1). Rather, as expected, the species composition was strongly associated with age (Supplemental Fig. S1).

As it is currently not possible to fully resolve multiple individual bacterial strains using shotgun metagenomic data, we considered SNV species profiles to represent strain populations, whereby multiple strains of the same species may occur within individuals, some of which may be shared between individuals or at different time points. Given the complexity of SNV analysis in metagenomes and various sources of errors (Schloissnig et al. 2013; Li et al. 2016), we used a very conservative approach to define strain sharing. Namely, we only tracked rare marker SNVs (hereafter rmSNVs) that were not shared with any non-family-member to avoid analysing uninformative common SNVs prevalent in the population. In addition to the analysed families, we included 884 deeply sequenced metagenomes of Europeans and North Americans (Human Microbiome Project Consortium 2012; Le Chatelier et al. 2013; Zeller et al. 2014) as a background population to exclude common variants that do not unambiguously ensure shared origin. For an illustration of the definition of rmSNVs, see Supplemental Fig. S2. For each pairwise sample comparison (between any two samples A and B), we defined as the reference population, a sample set that excluded all samples from the donors of samples A and B, and from their family members. For sample A, we identified exclusive alleles (rmSNVs) that were not found in the reference population containing all samples from non-family-members of the compared pair of individuals. Each comparison had a different reference population, and a different set of rmSNVs. As even in this restricted set, a small fraction of rmSNVs was shared by chance between pairs of unrelated individuals, i.e. from other families (mean similarity, 0.2%), we required >20% rmSNV sharing between compared samples as evidence of strain sharing (see Methods).

Maternal transmission of strains and their persistence during infant development

Thirty-four species had high genomic coverage (>10x) in at least two family members to allow rigorous statistical analysis (Supplemental Fig. S3). To track vertical strain transmission at birth in vaginally delivered infants in regard to those species, we compared the rmSNV profiles of the neonates to those of their mothers. In 87% of the 55 vaginally born neonates whose strain populations could be compared to those of their mothers, we found strong evidence of maternal transmission of bacteria (Fig. 1A), as indicated by high rmSNV similarity between neonates and their mothers (median 92%). While the total microbiota composition of these neonates did not resemble that of their mothers (mean correlation = 0.29 between mother-neonate pairs; Fig. 1B), the similarity to the mother in composition increased with age as the child's microbiota matured ($p<0.0001$, Fig. 1B). As it was never significantly higher than the similarity to unrelated mothers, the increase in similarity to the mothers likely reflects the changes in the gut environment due to child development and diet on the microbiota composition.

After establishing colonisation with maternal strains, we studied strain persistence, tracking rmSNVs within individuals over time (Fig. 2). The rmSNV similarity between time points of the same individual was extremely high (median 90% similarity within a year) at all ages, even in infants (Fig. 2A), despite the considerable species-level compositional fluctuation observed there (Fig. 2B). This is consistent with generally high strain stability within individuals observed in previous studies (Schloissnig et al. 2013; Yassour et al. 2016). In infants, maternal strains were very stable over the first year of life, but the originally non-maternal strains were nearly always replaced by the age of 12 months, as indicated by close to zero intra-individual rmSNV similarity (Fig. 2A). In children and adults (n=304), strain population similarity between time points declined gradually over time, with 13% rmSNV change per year on average (Fig. 2C), while the total species-level composition remained at close to 90% similarity over time spans of several years (Fig. 2D).

Selective seeding by maternal Actinobacteria and Bacteroidia

As the maternally derived strains appeared to be more stable than the non-maternal strains in the infants, we considered that different taxa might be differentially transmissible. Of the 34 species analysed, 31 belong to only three bacterial classes, Actinobacteria (phylum Actinobacteria), Bacteroidia (phylum Bacteroidetes), and Clostridia (phylum Firmicutes), that together make up 90% of the assignable relative species abundance. We therefore focus on these three classes. Adults had a considerable abundance of Clostridia (mean relative abundance 34% of the total), and Bacteroidia (50%) and a low abundance of Actinobacteria (4%) (Supplemental Fig. S1, S4). The microbiota of infants was dominated by Actinobacteria and Bacteroidia (together representing 59% of the total assignable abundance), with low abundance of Clostridia (8%).

Distinct phylogenetic patterns were indeed evident in the transmissibility of strains from mother to neonate (Fig. 3A, Supplemental Table S4). Strain populations from Actinobacteria and Bacteroidia were nearly always shared between mother and vaginally born neonate: the strain populations in 97% of the neonates with tracked Actinobacteria (n=36) and in 93% of those with tracked Bacteroidia species (n=29)

were identified as maternal (Fig. 3A), indicating transmission during birth (Supplemental Table S4). Maternal strains of Clostridia were never observed in any neonate with tracked Clostridia species (n=12) in the first days of life (Fig. 3A, Supplemental Table S4) albeit the abundance of Clostridia in neonates was generally low (Supplemental Fig. S1, S4). Of the few other species that did not belong to these three classes (Supplemental Fig. S5A, Supplemental Table S4), *Akkermansia muciniphila* (phylum Verrucomicrobia), and *Dialister invisus* (phylum Firmicutes) were not transmitted at birth, while *Collinsella aerofaciens* was (phylum Actinobacteria, close relative of Actinobacteria). Thus, although only about a half of the relative abundance of the mother's intestinal bacteria seem transmissible at birth (Actinobacteria and Bacteroidia together), the maternal strains of these two classes represent the vast majority of abundance in the neonate, with an expansion of environmental strains of Clostridia later in life.

Bacteroidia strain populations showed higher stability than Actinobacteria and Clostridia ($p<0.0001$, Fig. 3B), suggesting that they are less frequently replaced by new strains. Maternal Bacteroidia were still observable after one year in 94% of the infants with maternal Bacteroidia strains (n=46), compared to 81% of the infants retaining traceable maternal Actinobacteria strains (n=37; Fig. 3B). Diet changes did not influence the stability of maternal SNVs (Supplemental Fig. S6, $p=0.45$). All infants initially received breast milk (exclusively or supplemented with formula), and 14 infants later transitioned to formula feeding before solid foods by age 12 months. After the age of 12 months, rmSNV similarity to the mother gradually declined, roughly following the same time course as the decline in intra-individual rmSNV similarity in adults (Fig. 3A, Supplemental Fig. S5A).

Clostridia strains, as well as strains from *Dialister* and *Akkermansia*, appeared to colonise the infants persistently only after the first year, and even then, *Akkermansia* was unstable in a subset of the individuals (Fig. 3B, Supplemental Fig. S5B). The persistent colonisation by these taxa coincided with a shift to a diet containing plant polysaccharides, the preferred substrate of Clostridia and *Dialister*, which likely enabled their stable colonisation.

The transient non-maternal strains appeared to be partly replaced by maternal ones during the first year (Fig. 1A), demonstrated by the high rmSNV similarity of the originally non-maternally derived species to the maternal strains at 12 months, implying post-natal colonisation by strains from the environment. We therefore further investigated the potential influx of strains from the environment in later stages of life by also tracking rmSNVs between father-child, sibling and spouse pairs and considering them as indicators of colonization by environmental strains.

Strain transmission between family members in adulthood

The influence of the family environment peaked at age 2-10 years (Fig. 3C, Supplemental Fig. S5C, n=12). After the age of 10 years, strain sharing was generally low (Fig. 3C, Supplemental Fig. S5C, n=27), although still significantly more frequent than with unrelated individuals (Fig. 4A, $p<0.0001$), concordant with results from adult twins (Rothschild et al. 2017). Strain sharing between family members was stable over at least one month, indicating that the shared strains were resident microbes. All family

member pairs had similar levels of rmSNV sharing, apart from mother-child pairs, which shared more rmSNVs, perhaps a reflection of the maternal seeding. Twins generally did not have more similar rmSNV profiles than non-twin siblings (Fig. 4A). An exception was Clostridia, showing higher rmSNV similarity between twins than mother-child pairs, likely due to the lack of maternal seeding and the strong impact of shared environment for this class (Fig. 4A). In some cases siblings had (non-significantly) higher rmSNV similarity than twins, which is likely due to chance. Overall the results are in line with a recent study on adult twins in which monozygotic twins did not share more strains than dizygotic twins, indicating a lack of genetic influence on strain sharing (Xie et al. 2016; Rothschild et al. 2017).

Another indication for the retaining of maternal strains later in life is the exclusive sharing of many rmSNVs, i.e. the rmSNVs shared with mother were not observed in other family members, as indicated by the exclusively shared rmSNVs (ErmSNVs) in Fig. 4A. In mother-child pairs, the ErmSMVs encompassed nearly all of the shared rmSNVs. This is in contrast to the shared rmSNVs between twins, siblings, father-child pairs, and spouses that were often also shared with a third family member (ErmSNV sharing is often much lower than the total rmSNV sharing in Fig. 4A), implying a source in the shared household environment, with no evidence for direct transmission from a particular person.

To assess the directionality of strain transmissions within families, we tracked the emergence of novel SNVs that were not detected in previous samples of the focal individual, now extending the analysis from rmSNVs to all SNVs to increase statistical power. As the newly observed SNVs in one family member were indeed often present in previous samples of one or several family members, we assumed strain transmission between family members (Fig. 4B). Surprisingly, the father was significantly more often than the mother or the children the likely donor of intra-family transmissions, including transmissions between the parents and between parent and child ($p<0.0001$), and less often the recipient ($p<0.0001$). Thus the fathers in the 10 families with at least two siblings analysed appeared to bring most of the novel strains into the family environment.

C-section prevents maternal seeding

As maternal selective seeding is likely to be affected by birth mode, we also analysed 25 infants and six 2-10 year olds in our cohorts that were born by caesarean section, which were excluded in the prior analyses. These children showed strikingly different strain transmission and persistence patterns compared to the vaginally born (Supplemental Fig. S7, S8). The microbiome in the caesarean-delivered infants was mostly devoid of the maternally transmitted seeding classes Actinobacteria and Bacteroidia during the first months; species of Bacteroidia in particular were consistently missing (Supplemental Fig. S7A), as observed in other cohorts (Bokulich et al. 2016; Yassour et al. 2016). Only 6 of the 15 Caesarean-delivered neonates had any species overlap with the mothers to allow SNV analysis, and we did not observe sharing of a single strain (Supplemental Fig. S7B), clearly implicating vaginal birth as the main transmission route. The likelihood of having maternal strains was 0.87 in the vaginally born neonates and 0 in the caesarean-born (χ^2 test, $p<0.0001$). During their first year of life, the caesarean-delivered infants gradually acquired maternal Actinobacteria and

Bacteroidia strains, which replaced the original non-maternal ones (Supp. Fig. S7B). In addition to the absence of maternal strains, the caesarean-delivered infants showed higher strain flux than the vaginally born ones, particularly regarding Bacteroidia strains ($p<0.0001$, Supplemental Fig. S8).

Discussion

Taken together, our results indicate that maternal transmission of bacteria is a controlled process, in which only a minor, selected part of the maternal microbiome colonises the neonate and expands to form the stable seeding community of the infant microbiome. The results demonstrate that colonisation of the infant gut is a selective process, rather than randomly determined by the strains that the infant is exposed to. The selection most likely occurs in the infant gut after initial inoculation by maternal faecal microbes, and may be due to the milk-based diet, which Actinobacteria and Bacteroidia species are able to utilise, thus gaining a selective advantage (Sela and Mills 2010; Marcobal et al. 2011). These taxa are considered important for healthy metabolic and immunological development in infants (Mazmanian et al. 2005; Korpela et al. 2017).

The high stability of the maternal strains in the infant gut demonstrates the importance of maternal seeding. It was recently suggested that certain species colonise the infant once and remain stable thereafter while others spread in the population, colonising each individual several times (Yassour et al. 2016). Our results further this idea, identifying the maternally derived strains to be stable and the non-maternal ones to be replaceable. Generally, the results may indicate that different symbiotic bacterial species have evolved different transmission strategies, some relying on vertical transmission at birth and others on horizontal transmission in later life. Horizontal transfer via the environment could be facilitated by endospore formation, which is common among Clostridia.

Surprisingly, infants and children are not more commonly colonised by novel strains than adults. After initial colonisation, the maternal strains persist in infants. This implies that the transmitted part of the maternal gut microbiome may have a protective function, preventing the influx of environmental conspecific strains, which mayx have a higher risk of carrying unwanted properties. It is currently not clear why the maternal strains are more stable than non-maternal conspecific strains. A candidate explanation is breast milk, which contains maternal immunoglobulins as well as oligosaccharide structures similar to those present in the maternal gut mucosa. Maternal bacterial strains may thus have an advantage as they are pre-selected to be compatible with these breast milk components. However, transitions to formula feeding and to solid foods during the first year did not reduce the stability of the maternal strains. Evidently the bacteria, especially Bacteroidia, were able to adapt to the new dietary pattern. Bacteroidia species are known to have a diverse repertoire of carbohydrate-active enzymes, which gives them considerable substrate flexibility (El Kaoutari et al. 2013).

Over the years, strain similarity to the mother declines, most likely due to gradual strain replacement, although mutations may contribute to this, as well. Additionally, several species, including members of Clostridia and *Akkermansia muciniphila* appear to colonise persistently only after the first year of life and are not derived from the mother. Such behaviour of late-colonizers may be explained by their high degree of

specialisation to the conditions in the adult gut, rendering them unable to persistently colonise the neonate. Due to the frequent absence or very low abundance of Clostridia in the neonates, we cannot refute occasional transmission from the mother, but the consistent lack of maternal Clostridia rmSNVs in children of all ages, despite commonly observed paternal and fraternal strains, makes transmission at birth highly unlikely or at least extremely rare.

Our data show the dramatic effect of cesarean section on infant gut colonization: these infants fail to receive maternal strains at birth and instead show a high degree of strain flux in early life, comparable to the flux observed in non-transmitted species in vaginally born infants. This suggests that there is initially a mismatch between the bacteria and the host, which is gradually resolved as the section-born infants acquire maternal strains postnatally from the environment. How the disruption of bacterial transmission affects the developing immune system is currently not known. Despite the low number of infants studied, these results do not support the generality of the recently suggested intrauterine bacterial transmission via the placenta (Aagaard et al. 2014) or the entero-mammary route of transmission (Funkhouser and Bordenstein 2013; Jost et al. 2014; Rodriguez et al. 2014), as these should affect infants regardless of birth mode. It remains unclear whether this early fluctuation implies an increased risk of colonisation by undesired strains with respective health consequences.

While the species composition of gut microbiota has been reported to be influenced by host genetics (Xie et al. 2016), our results demonstrate that the identity of the strains within a species is dependent on the environment (Rothschild et al. 2017). Family members are likely the most important environmental source of human gut microbes, and bacterial transmission between family members occurs also in later stages of life. The frequency, the constraints, and the functional consequences of strain transmission between family members and from other environmental sources still need to be investigated.

Methods

Data

We monitored bacterial strains in gut metagenomes of 100 Swedish mother-infant pairs[5], including 15 caesarean-delivered infants, 42 US infants 10 of which were caesarean-delivered, and 5 vaginally born Italian infants. The Swedish infants had been sampled during their first postnatal week, and at ages 4 and 12 months, the mothers only during the infant's first week. The US infants and mothers were sampled at birth and at 6 weeks. The Italian infants and mothers were sampled at birth, at 7 months, and at 1 year. In addition, we sequenced metagenomes of 8 German and 2 Dutch families with children of different ages, including 2 families with infants, 2 families with adult children and 5 families with twins of different ages, some of which were caesarean-delivered (Supplemental Table S2). As an additional adult cohort, we analysed 139 female HMP participants sampled at one to three time points with 6-12 month intervals. As the background population, we included 884 samples from different European cohorts.

We categorised the samples into 6 age groups: <1 week ("Neonate", N=113), >1 week to 6 months ("6 months", N=139), 6 months to 2 years ("12 months", N=105), 2-10 years (N=12), 10-25 years (N=60), >25 years (N=232).

Sample collection, DNA extraction, sequencing

Samples were collected fresh and immediately frozen at -20 degrees until arriving to the laboratory where they were kept in long-term storage at -80 degrees. Genomic DNA was extracted from frozen faecal samples as previously described (Zeller et al. 2014) using a GNOME® DNA Isolation Kit (MP Biomedicals). Libraries were generated and shotgun sequenced on the Illumina HiSeq 2000/2500 (Illumina, San Diego, CA, USA) platform, in a paired-end sequenced setup with 100 bp read lengths at the Genomics Core Facility, European Molecular Biology Laboratory.

Species abundance estimation

All samples were processed with the same computational protocol. Reads were quality filtered and screened against the human genome sequence for removing contamination as previously described (Zeller et al. 2014). Sequencing reads were then mapped to a reference set consisting of 1753 genomes, each representative of one specI cluster (Mende et al. 2013) using MOCAT (Kultima et al. 2012) with default parameters. Specifically, reads were mapped at 97% identity and multiple mappers were discarded. Computation of genome coverage for each specI cluster was performed using qaCompute (https://github.com/CosteaPaul/qaTools), resulting in estimations of both horizontal and vertical coverage per sample, per genome. The abundance of species was estimated based on the genome coverages and transformed into relative abundances. Species composition similarity between samples was calculated as Pearson correlation of the log-transformed relative abundances.

SNV sharing

Determination of SNVs was performed using the metaSNP tool (available at https://git.embl.de/costea/metaSNV) with default parameters. The number of SNVs varied by species, ranging from 5 to 116915 (mean 22308) for Clostridia, 3133- 30823 (mean 20695) for Bacteroidetes, and 2-80287 (mean 46697) for Actinobacteria. Sharing of SNVs between samples was calculated as the number of shared SNVs divided by the number of positions that were detected in both samples. In cases where a species was not detected in a sample (genome coverage was 0), SNV sharing of that species between that sample and all other samples was set to 0.

To ensure that the shared SNVs were indeed of the same origin, the main part of the study was based on marker SNVs (rmSNVs) excluding all SNVs that were shared between unrelated individuals (apart from the pair of individuals being compared, even if they were unrelated), as these are uninformative for strain source tracking. The reference population thus contained all samples from individuals not related to the donors of the pair of samples (sample A and sample B) currently being compared. For sample A, we defined rmSNVs as the SNVs, which were not present in the reference population (keeping in mind that the reference population excludes sample B, even if the donor of sample B is not related to the donor of sample A). These rmSNVs were

then compared to the SNVs in sample B. SNV sharing was only assessed if >100 marker-positions were available for the compared pair of samples.

We initially monitored the SNVs in 108 species, but selected for detailed analysis a subset of 34 based on sufficient genome coverage (>40%). To make sure the observed inter-sample rmSNV similarity was not influenced by the observed number of rmSNVs in the samples or the abundance of the species, we checked for correlations between rmSNV sharing and rmSNV richness and species abundance. For the classes Actinobacteria, Bacteroidia, and Clostridia, the observed number of rmSNVs or the abundance of the species were not associated with rmSNV sharing. However, for other classes, the observed rmSNV similarity was not independent of rmSNV richness or species abundance, suggesting that for many cases the rmSNV similarity was underestimated due to insufficient sampling, despite the high sequencing depth. We therefore excluded these taxa from the analysis.

SNV frequencies within species showed often bimodal distributions, peaking at <20% or >80%. This indicates that individuals mostly harboured one dominant strain, either the one with the reference SNV or the one with the non-reference SNV. However, particularly among the infants, there were also many cases with unimodal or even distributions of SNV frequencies, indicating that having two or more strains simultaneously was common. Therefore, we considered the possibility that <100% SNV similarity could indicate the sharing of one strain, which coexisted with a non-shared strain. To establish a cut-off for reliably demonstrating strain sharing (with the possibility of coexisting non-shared strains), we assessed rmSNV similarities between unrelated individuals. Unrelated individuals rarely shared rmSNVs. The frequency of >20% similarity between unrelated pairs was 0.5% (Supplemental Fig. S9). We used this conservative cut-off for defining strain sharing: rmSNV similarity <20% was deemed insufficient evidence of strain sharing, resulting in 0.005 false discovery rate.

For more detailed analysis on intra-family strain sharing, we further tracked a subset of rmSNVs that were exclusively shared between the compared pair of individuals, and with no other family members. We compared the family-specific and the exclusive rmSNV similarities within a pair of individuals to establish if the shared rmSNVs were likely a result of direct transmission (all shared rmSNVs were shared exclusively), or more likely came from a common source in the family environment (shared rmSNVs were also shared with other family members).

To establish directionality of strain transmission, we compared novel SNVs not seen in previous samples of the same individual to SNVs in previous samples of family members, including all SNVs to maximise the number of novel SNVs observed. Strain transmission from a family member was assumed to have occurred if novel SNVs appearing in an individual were observed in previous samples of a family member. Novel SNVs were rare so nove rmSNVs were too rare for robust analysis. However, in a few individual we were able to trace novel rmSNVs to a family member, validating the concept.

The number of rmSNVs compared by species and type of relation between individuals is shown in Supplemental Fig. S10.

Statistical tests

Significance of group differences in SNV similarity (proportion of shared rmSNVs) was assessed using beta-regression (package betareg (Cribari-Neto and Zeileis 2010) in R (R

480 Core Team 2015)). The significance of trends and group differences in compositional
481 similarity, which was normally distributed, was assessed using linear models.
482
483

Data Access

The newly generated metagenomes from this study have been submitted to the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under accession number PRJEB24041.

Acknowledgements

We thank members of the Bork group for constructive discussion on the manuscript, Yan Yuan for IT support, Rajna Hercog and Vladimir Benes at EMBL Genomic Core Facility for the sequencing work. The work was supported by EMBL and grants to PB from ERC (EC/H2020/ES/ERC-AdG-669830) and to KK from the Academy of Finland (297765).

Disclosure declaration

The authors declare no competing financial interests.

503 References
504
505 Aagaard K, Ma J, Antony K, Ganu R, Petrosino J, Versalovic J. 2014. The placenta
506 harbors a unique microbiome. *Sci Transl Med* **6**: 237ra65.
507 Albesharat R, Ehrmann MA, Korakli M, Yazaji S, Vogel RF. 2011. Phenotypic and
508 genotypic analyses of lactic acid bacteria in local fermented food, breast milk
509 and faeces of mothers and their babies. *Syst Appl Microbiol* **34**: 148-155.
510 Asnicar F, Manara S, Zolfo M, Truong DT, Scholz M, Armanini F, Ferretti P, Gorfer V,
511 Pedrotti A, Tett A et al. 2017. Studying Vertical Microbiome Transmission from
512 Mothers to Infants by Strain-Level Metagenomic Profiling. *mSystems* **2**: e00164-
513 16.
514 Bäckhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, Li Y, Xia Y,
515 Xie H, Zhong H *et al*. 2015. Dynamics and stabilization of the human gut
516 microbiome during the first year of life. *Cell host & microbe* **17**: 690-703.
517 Bokulich N, Chung J, Battaglia T, Henderson N, Jay M, Li H, Lieber A, Wu F, Perez-
518 Perez G, Chen Y et al. 2016. Antibiotics, birth mode, and diet shape microbiome
519 maturation during early life. *Sci Trans Med* **8**: 343.
520 Cho I, Blaser MJ. 2012. The human microbiome: at the interface of health and disease.
521 *Nat Rev Genetics* **13**: 260-270.
522 Chu D, Ma J, Prince A, Antony K, Seferovic MD, Aagard K 2017. Maturation of the
523 infant microbiome community structure and function across multiple body sites
524 and in relation to mode of delivery. *Nature Medicine* **23**: 314-326.
525 Cribari-Neto F, Zeileis A. 2010. Beta Regression in R. *Journal of Statistical Software*
526 **32**: 1-24.
527 Dominguez-Bell MG, Costello EK, Contreras M, Magris M, Hidalgo G, Fierer N,
528 Knight R. 2010. Delivery mode shapes the acquisition and structure of the initial
529 microbiota across multiple body habitats in newborns. *PNAS* **107**: 11971-11975.
530 El Kaoutari A, Armougom F, Gordon J, Raoult D, Henrissat B. 2013. The abundance
531 and variety of carbohydrate-active enzymes in the human gut microbiota. *Nat*
532 *Rev Microbiol* **11**: 497–504.
533 Funkhouser LJ, Bordenstein SR. 2013. Mom knows best: the universality of maternal
534 microbial transmission. *PLoS Biol* **11**: e1001631.
535 Gilbert, SF. 2014. A holobiont birth narrative: the epigenetic transmission of the human
536 microbiome. *Frontiers in genetics* **5**: 282.
537 Houghteling PD, Walker WA. 2015. Why is initial bacterial colonization of the intestine
538 important to infants' and children's health? *J Pediatr Gastroenterol Nutr* **60**:
539 294-307.
540 Human Microbiome Project Consortium. 2012. Structure, function and diversity of the
541 healthy human microbiome. *Nature* **486**: 207-214.
542 Jiménez E, Delgado S, Maldonado A, Arroyo R, Albújar M, García N, Jariod M,
543 Fernández L, Gómez A, Rodríguez JM. 2008. Staphylococcus epidermidis: a
544 differential trait of the fecal microbiota of breast-fed infants. *BMC Microbiol* **8**:
545 1.
546 Jost T, Lacroix C, Braegger CP, Rochat F, Chassard C. 2014. Vertical mother–neonate
547 transfer of maternal gut bacteria via breastfeeding. *Environ Microbiol* **16**: 2891-
548 2904.
549 Korpela K, Zijlmans M, Kuitunen M, Kukkonen K, Savilahti E, Salonen A, de Weerth
550 C, de Vos W. 2017. Childhood BMI in relation to microbiota in infancy and
551 lifetime antibiotic use. *Microbiome* **5**: 26.

552 Kultima J, Sunagawa S, Li J, Chen W, Chen H, Mende D, Arumugam M, Pan Q, Liu B,
553 Qin J, et al. 2012. MOCAT: A Metagenomics Assembly and Gene Prediction
554 Toolkit. *Plos One* **7:** e47656.
555 Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, Almeida M,
556 Arumugam M, Batto J-M, Kennedy S et al. 2013. Richness of human gut
557 microbiome correlates with metabolic markers. *Nature* **500**: 541–546.
558 Li S, Zhu A, Benes V, Costea P, Hercog R, Hildebrand F, Huerta-Cepas J, Nieuwdorp
559 M, Salojärvi J, Voigt A, et al. 2016. Durable coexistence of donor and recipient
560 strains after fecal microbiota transplantation. *Science* **352**: 586-589.
561 Makino H, Kushiro A, Ishikawa E, Kubota H, Gawad A, Sakai T, Oishi K, Martin R,
562 Ben-Amor K, Knol J, Tanaka R. 2013. Mother-to-infant transmission of
563 intestinal bifidobacterial strains has an impact on the early development of
564 vaginally delivered infant's microbiota. *PloS One* **8**: e78331.
565 Marcobal A, Barboza M, Sonnenburg E, Pudlo N, Martens E, Desai P, Lebrilla C,
566 Weimer B, Mills D, German J et al. 2011. Bacteroides in the Infant Gut
567 Consume Milk Oligosaccharides via Mucus-Utilization Pathways. *Cell Host &*
568 *Microbe* **10**: 507-514.
569 Matsumiya Y, Kato N, Watanabe K, Kato H. 2002. Molecular epidemiological study of
570 vertical transmission of vaginal Lactobacillus species from mothers to newborn
571 infants in Japanese, by arbitrarily primed polymerase chain reaction. *J Infect*
572 *Chemotherapy* **8**: 43-49.
573 Mazmanian SK, Liu CH, Tzianabos AO, Kasper DL. 2005. An immunomodulatory
574 molecule of symbiotic bacteria directs maturation of the host immune system.
575 *Cell* **122**: 107-118.
576 Mende DR, Sunagawa S, Zeller G, Bork P. 2013. Accurate and universal delineation of
577 prokaryotic species. *Nature Methods* **10**: 881-884.
578 Nayfach S, Rodriguez-Mueller B., Garud, N. & Pollard, K. S. 2016. An integrated
579 metagenomics pipeline for strain profiling reveals novel patterns of bacterial
580 transmission and biogeography. *Genome Research* **26**: 1612-1625.
581 R Core Team. 2015. R: A language and environment for statistical computing. R
582 Foundation for Statistical Computing, Vienna, Austria. http://www.R-
583 project.org/.
584 Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SS, McCulle SL, Karlebach S, Gorle
585 R, Russell J, Tacket CO et al. 2011. Vaginal microbiome of reproductive-age
586 women. *PNAS* **108**: 4680-4687.
587 Rodriguez JM. 2014. The origin of human milk bacteria: is there a bacterial entero-
588 mammary pathway during late pregnancy and lactation? *Adv Nutr* **5**: 779-784.
589 Rothschild D, Weissbrod O, Barkan E, Korem T, Zeevi D, Costea P, Godneva A, Kalka
590 I, Bar N, Zmora N, et al. 2017. Environmental factors dominate over host
591 genetics in shaping human gut microbiota composition. *bioRχiv* pre-print
592 https://doi.org/10.1101/150540.
593 Schloissnig S, Arumugam M, Sunagava S, Mitreva M, Tap J, Zhu A, Waller A, Mende
594 D, Kultima J, Martin J, et al. 2013. Genomic variation landscape of the human
595 gut microbiome. *Nature* **493**: 45-50.
596 Sela D, Mills D. 2010. Nursing our microbiota: molecular linkages between
597 bifidobacteria and milk oligosaccharides. *Trends Microbiol* **18**: 298-307.
598 Tannock GW, Fuller R, Smith SL, Hall MA. 1990. Plasmid profiling of members of the
599 family Enterobacteriaceae, lactobacilli, and bifidobacteria to study the
600 transmission of bacteria from mother to infant. *J Clin Microbiol* **28**: 1225-1228.

Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N 2017. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27**: 626-638.

Xie H, Guo R, Zhong H, Feng Q, Lan Z, Qin B, Ward K, Jackson M, Xia Y, Chen X et al. 2016. Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut Microbiome. *Cell Systems* **3**: 572-584.

Yassour M, Vatanen T, Siljander H, Hämäläinen A-M, Härkönen T, Ryhänen S, Franzosa E, Vlamakis H, Huttenhower C, Gevers D et al. 2016. Natural history of the infant gut microbiome and impact of antibiotic treatments on strain-level diversity and stability. *Sci Transl Med* **8**: 343.

Zeller G, Tap J, Voigt A, Sunagawa S, Kultima J, Costea P, Amiot A, Böhm J, Brunetti F, Habermann N, et al. 2014. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol Syst Biol* **10**: 766.

Zhu A, Sunagawa S, Mende DR, Bork P. 2015. Inter-individual differences in the gene content of human gut bacterial species. *Genome Biol* **16:** 1.

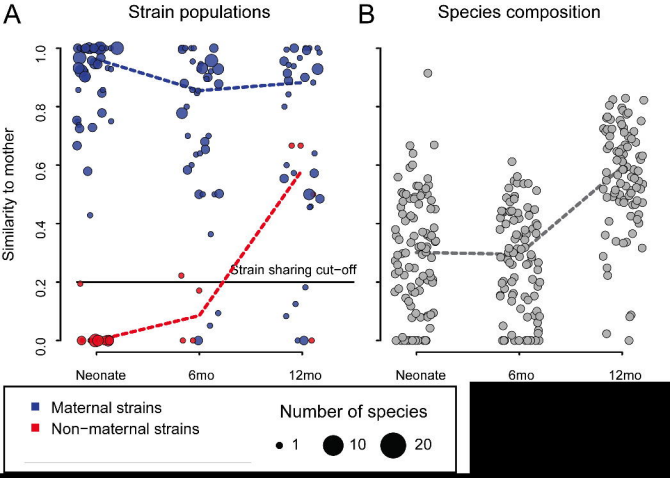
Figure Legends

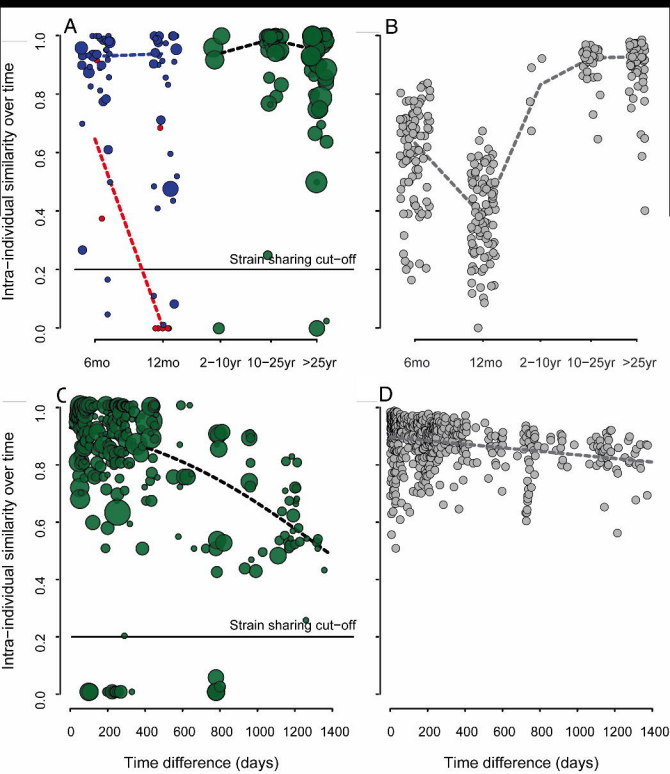
Figure 1. Similarity to mother over time, of vaginally born infants (n=114), measured as rmSNV similarity (A) and species composition similarity (B). Strains were classified as originally maternal if rmSNV similarity to mother was >0.2 (strain sharing cut-off) in the first neonatal sample available for each pair. Maternal/non-maternal strains were summarised in each circle as median per individual. Even if the similarity to the mother increased in later time points, the original distinction of maternal/non-maternal was kept for each child-mother pair through all further time points.

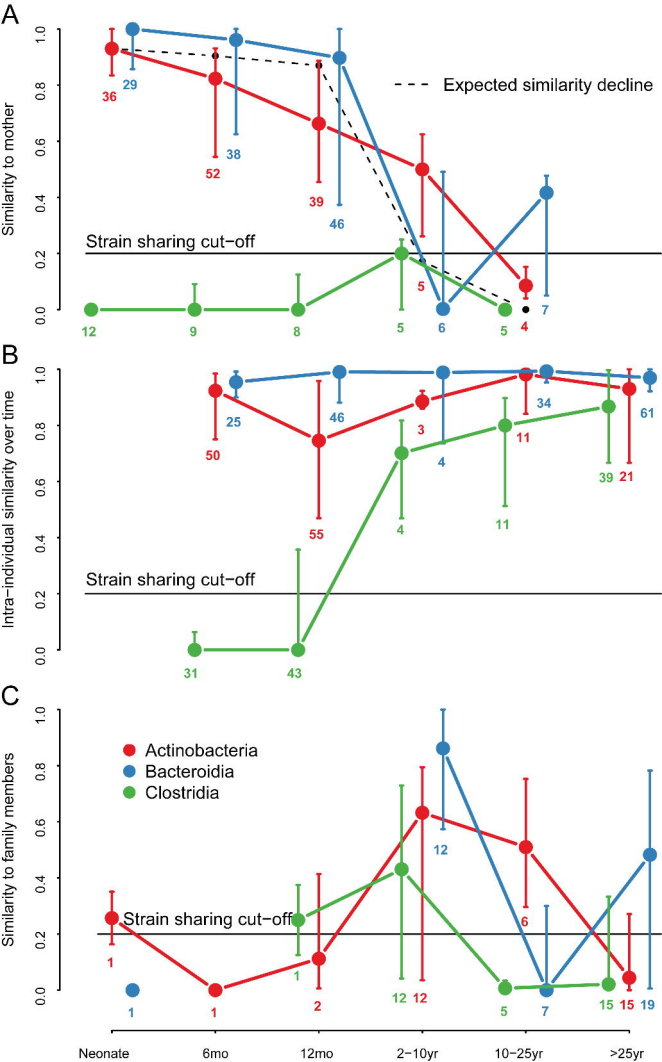
Figure 2. Intra-individual similarity over time. Panels A and B show similarity to previous time points at different ages (median over all time points per individual, including vaginally born infants and children, and adults irrespective of birth mode, N=282). Panels C and D show intra-individual similarity by time difference between samples in >1 year olds (N=304).

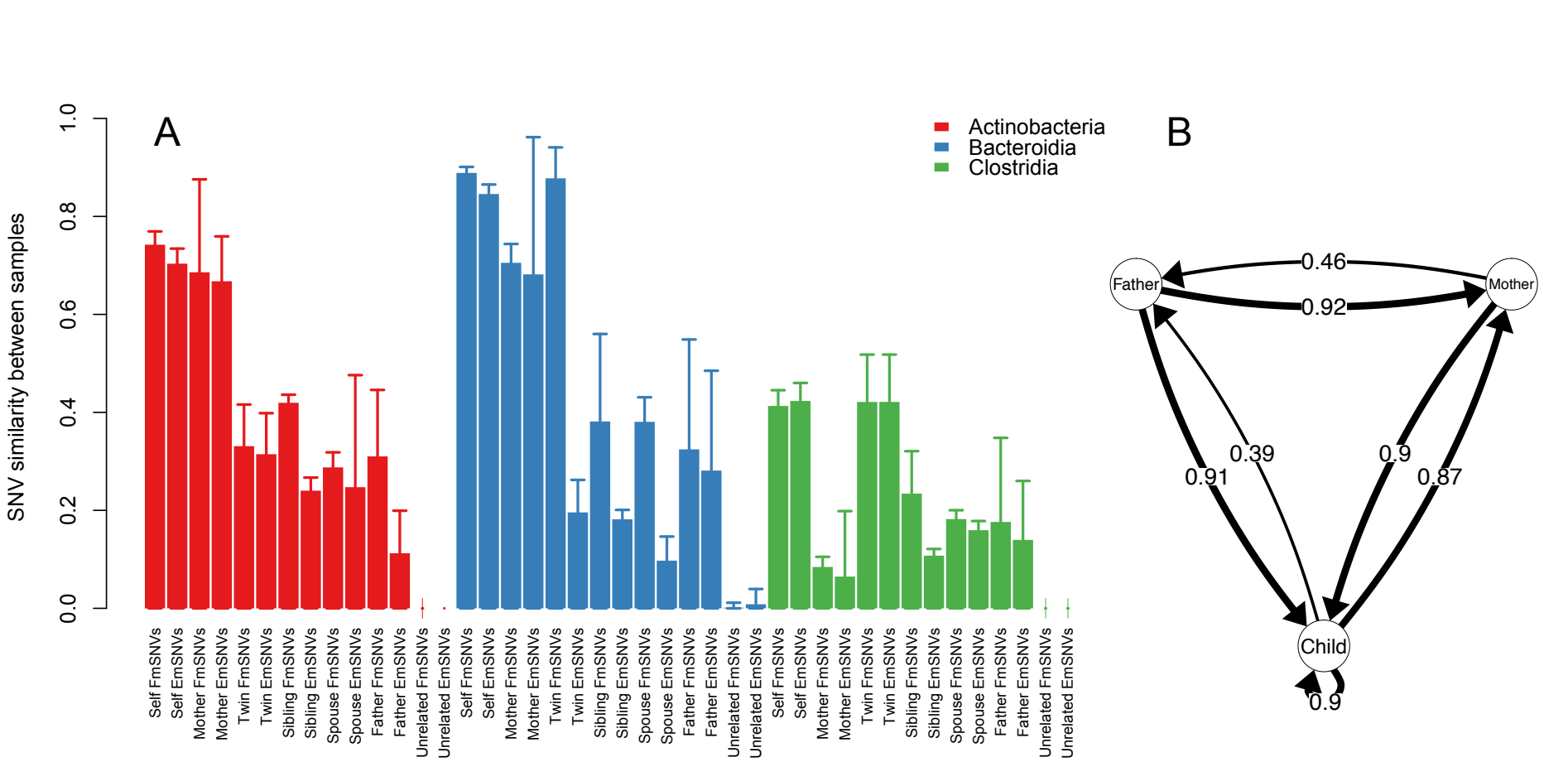
Figure 3. Phylogenetic signal in strain transmission and stability. (A) Strain population (rmSNV profile) similarity of vaginally born infants to mother. (B) Intra-individual rmSNV similarity over time in vaginally born infants and children, and adults regardless of birth mode. (C) Strain population (rmSNV) similarity to other family members. The symbols represent age group medians ± inter-quartile range per bacterial class. The number of individuals included in each group is indicated. Expected similarity decline is based on the intra-individual similarity decline in Fig. 2C.

Figure 4. Strain similarity and transmissions within families. The bars indicate mean±sd. Two sets of rare marker SNVs were defined per pair of individuals: rare marker SNVs, which were allowed to be shared with other family members (marked here as Family marker SNVs, FmSVs), and exclusive marker SNVs (EmSNVs), which were only shared between the compared pair. The arrows indicate the cases where novel SNVs emerging in the focal individual were observed in previous time points of a given family member. The arrow from child to child indicates transmission between siblings. The numbers on the arrows represent the proportion of identified transmission events between the indicated pair of family members (number of observed transmission events / total number of family member comparisons).


A
Strain populations
B
Species composition
Similarity to mother
1.0
0.8
0.6
0.4
0.2
0.0
Strain sharing cut−off
Neonate
6mo
12mo
Neonate
6mo
12mo
Maternal strains
Non−maternal strains
Number of species
1
10
20


A
B
C
D
Intra−individual similarity over time
0.0
0.2
0.4
0.6
0.8
1.0
Strain sharing cut−off
6mo
12mo
2−10yr
10−25yr
>25yr
0
200
400
600
800
1000
1200
1400
Time difference (days)
Maternal strains
Non−maternal strains
Strains of unknown origin
Number of species
1
10
20


A
Similarity to mother
1.0
0.8
0.6
0.4
0.2
0.0
Expected similarity decline
Strain sharing cut−off
36
29
52
38
39
46
5
7
12
9
8
5
6
5
4
B
Intra−individual similarity over time
Strain sharing cut−off
50
25
46
3
4
34
61
55
4
11
39
21
31
43
C
Similarity to family members
Actinobacteria
Bacteroidia
Clostridia
Strain sharing cut−off
1
1
1
1
2
12
12
12
6
5
7
15
15
19
Neonate
6mo
12mo
2−10yr
10−25yr
>25yr


A
B
Actinobacteria
Bacteroidia
Clostridia
SNV similarity between samples
0.0
0.2
0.4
0.6
0.8
1.0
Self FmSNVs
Self EmSNVs
Mother FmSNVs
Mother EmSNVs
Twin FmSNVs
Twin EmSNVs
Sibling FmSNVs
Sibling EmSNVs
Spouse FmSNVs
Spouse EmSNVs
Father FmSNVs
Father EmSNVs
Unrelated FmSNVs
Unrelated EmSNVs
Father
Mother
Child
0.46
0.92
0.91
0.39
0.9
0.87
0.9

Selective maternal seeding and environment shape the human gut microbiome

Katri Korpela, Paul Igor Costea, Luis Pedro Coelho, et al.

Genome Res. published online March 1, 2018
Access the most recent version at doi:10.1101/gr.233940.117