OPEN

Variation and transmission of the human gut microbiota across multiple familial generations

Mireia Valles-Colomer, Rodrigo Bacigalupe, Sara Vieira-Silva, Shinya Suzuki, Youssef Darzi, Raul Y. Tito, Takuji Yamada, Nicola Segata, Jeroen Raes and Gwen Falony

Although the composition and functional potential of the human gut microbiota evolve over the lifespan, kinship has been identified as a key covariate of microbial community diversification. However, to date, sharing of microbiota features within families has mostly been assessed between parents and their direct offspring. Here we investigate the potential transmission and persistence of familial microbiome patterns and microbial genotypes in a family cohort (n = 102) spanning 3 to 5 generations over the same female bloodline. We observe microbiome community composition associated with kinship, with seven low abundant genera displaying familial distribution patterns. While kinship and current cohabitation emerge as closely entangled variables, our explorative analyses of microbial genotype distribution and transmission estimates point at the latter as a key covariate of strain dissemination. Highest potential transmission rates are estimated between sisters and mother–daughter pairs, decreasing with increasing daughter's age and being higher among cohabiting pairs than those living apart. Although rare, we detect potential transmission events spanning three and four generations, primarily involving species of the genera Alistipes and Bacteroides. Overall, while our analyses confirm the existence of family-bound microbiome community profiles, transmission or co-acquisition of bacterial strains appears to be strongly linked to cohabitation.

The characterization of the acquisition and maturation of the human gut microbiota over the lifespan is of key importance for future clinical translation of microbiome research. Assessing transmissibility of bacterial strains and determining whether they are passed on at birth or acquired only later in life will support the development of guidelines to facilitate or hamper transmission depending on their beneficial or risk profile, respectively. Based on such a timeline and depending on whether the acquisition of a specific strain should be considered a health benefit or rather a risk factor with respect to disease development, guidelines to facilitate/hampen transmission can be formulated. Given reports of maternal inheritance of microbial strains, strain sharing among individuals sharing households and transmission events spanning multiple generations in animal models, similar considerations might apply when assessing the familial burden of conditions with a potential microbiota contribution, ranging from obesity to inflammatory bowel diseases.

Results and Discussion

Microbiome variation in a multigenerational family cohort is associated with age. To explore the persistence of transmittable microbial features across generations in the human host, we assembled a unique dataset of stool samples from women belonging to 24 multigenerational families living in the region of Flanders (Belgium) with accompanying metadata covering anthropometrics, delivery mode, cohabitation status, levels of systemic and local inflammation markers and use of medication (Fig. 1a,b and Supplementary Table 1). One hundred and two healthy individuals (aged 0–98, median = 37.5, born between 1917 and 2016) were sampled between November 2015 and November 2016. The standardized body mass index (SBMI) of participants (an age- and sex-corrected version of the body mass index valid also in children) varied between 7 and 56 (median = 37) with most individuals falling within the normal range (n = 60 out of 87; normal range 30–39). Ninety-nine (n = 99 out of 102) were born by vaginal delivery. Family structures ranged from 3 up to 5 generations (median = 4), presenting different degrees of multigenerational cohabitation and geographical dispersion.

Exploring host or environmental factors significantly contributing to interindividual microbiome variation in our cross-generational cohort (CGC), we combined shotgun metagenomic sequencing data with flow cytometry measurements of faecal microbial load to construct quantitative microbial abundance profiles. Within the limitations of the CGC cohort, stool moisture (n = 101, stepwise distance-based redundancy analysis (dbRDA) at the genus-level Bray–Curtis dissimilarity, R2 = 4.3%, P = 2 × 10^-4) and age (R2 = 2.9%, P = 2 × 10^-4) were identified as the only metadata variables with non-redundant explanatory power over quantitative microbiome variation (Fig. 1c and Supplementary Table 2). These findings align with previous reports on proportional microbiome variation in population cohorts, with stool moisture, a proxy of colonic transit time, reflecting ecosystem development induced by nutrient depletion on passage through the gastrointestinal tract. Additionally, we confirmed the negative associations between faecal water content and microbial load (n = 101, Spearman’s test, ρ = −0.25, P = 1.2 × 10^-2) as well as genus-level...
microbiome richness ($n=101$, $\rho=-0.29$, $P=3.7 \times 10^{-4}$). Although 21 participants reported to have taken antibiotics during the 12 months before sampling, we did not observe a significant impact of (history of) antibiotic therapy on microbiome composition in the present CGC cohort (Supplementary Table 2). Following up on reports of altered microbial ecosystem configurations in early childhood\cite{1,2}, we assessed a potential association between age bins (young children <4 years old, $n=10$ versus others...
The Bacteroides2 enterotype is highly prevalent among young children. Recently, we identified a faecal microbiota community type with high prevalence in cohorts of individuals with obesity14, inflammatory bowel disease12,23 and primary sclerosing cholangitis13, as well as among individuals with certain subtypes of multiple sclerosis24 and depression25. Common features of this potentially dysbiotic Bacteroides2 (Bact2) enterotype include low compositional richness, low faecal cell counts and high and low proportional abundances of the Bacteroides and Faecalibacterium genera, respectively. In general, Bact2-enterotyped individuals present looser stools and higher (both intestinal and systemic) inflammation markers26. To distinguish community states within the present CGC, we performed Dirichlet multinomial mixture (DMM) modelling27 against the background of microbiome variation as observed in the Flemish Gut Flora Project (FGFP) dataset (n = 1,106 population cohort)28. To this end and to preclude community clustering driven by methodological differences, the CGC dataset was additionally profiled using 16S ribosomal RNA gene amplicon sequencing following FGFP procedures29. The resulting amplicon profiles were only used for the purpose of enterotyping. Applying probabilistic models to group samples potentially originating from the same community, DMM-based stratification reproducibly identifies microbiome configurations across datasets without making any claims regarding the putative discrete nature of the strata detected. Microbiomes were observed to stratify over four previously described enterotypes14, labelled as Bacteroides1 (Bact1), Bact2, Prevotella and Ruminococcaceae (Fig. 1d and Extended Data Fig. 1). Bact2 samples diverged from their non-Bact2 counterparts, displaying lower microbial load (n = 101, Kruskal–Wallis test, chi-squared = 13.9, P = 3.0 × 10^{-4}; post-hoc Dunn test, P_{adj} < 0.05 for Bact2 versus Bact1 (Prevotella), lower genus-level richness (n = 101, Kruskal–Wallis test, chi-squared = 20.0, P = 1.6 × 10^{-4}; post-hoc Dunn test, P_{adj} < 0.05 for Bact2 versus Bact1 (Prevotella/Ruminococcaceae) and higher stool moisture content (n = 101, Kruskal–Wallis test, chi-squared = 8.8, P = 0.03; post-hoc Dunn test, P_{adj} < 0.05 for Bact2 versus Ruminococcaceae; Extended Data Fig. 2 and Supplementary Table 1). With only a single participant (mean abundance of one member species or functions appearing family-bound across 4+ CGC generations (non-random distribution in families across the cohort genealogy)30). None of the features evaluated (species, core functions and ARGs) were shared more frequently between related individuals than expected by chance in the cohort (n = 91, genealogical index of familiality (GIF), P_{adj} > 0.05; Supplementary Table 6).

Family members share closely related bacterial genotypes. The detection of familial microbiome community patterns does not necessarily reflect actual transmission of microorganisms across generations but could also result from shared genetic backgrounds and cultural transmission of lifestyle and dietary habits selecting for a similar microbial composition11,31. To infer potential exchange or co-acquisition of microbial strains between members of the same family, we recovered representative genotypes (consensus genetic sequences resulting from concatenation of marker genes with complete coverage) of species present with sufficient coverage in the unrarefied CGC faecal shotgun metagenomes using StrainPhAn. This approach allowed us to characterize over 360 species across the CGC dataset (including samples from young children, n = 102; Fig. 2 and Supplementary Table 1).
Table 7). Focusing on species detected at least 3 times within a single family and having a core genome alignment higher than 1,000 base pairs (bp), we restricted our analyses to 2,374 genotypes representing 51 species (median genotypes per species = 44, range = 13–92; Extended Data Fig. 3 and Supplementary Table 8), together constituting a substantial fraction of the CGC metagenomes (median = 77.04%, range = 7.35–91.85%; Supplementary Table 1). For each species, we calculated the genetic distances between all pairs of genotypes recovered as the number of single-nucleotide polymorphisms (SNPs) (Supplementary Table 9). Overall, for these 51 species, the normalized genetic distances (nGDs) (normalized by the median intraspecies genetic distance as proposed by Ferretti et al.7) between genotypes recovered from family members (intrafamily (IF) were lower than those observed between non-related individuals (between-family (BF)); median nGD_{IF} = 0.973 versus nGD_{BF} = 1; n = 102, permutational multivariate analysis of variance (PERMANOVA) on median nGDs, $R^2 = 0.304$, $P = 1 \times 10^{-5}$; Fig. 3a), indicating that more similar strains could be found within than across families. Analysed per species, a similar pattern was observed for 13 out of the 51 taxa genotyped (PERMANOVA, $P_{ab} < 0.05$; Supplementary Table 10). Of note, the overall distribution of IF distances showed a peak at nGD = 0 (that is, identical strains) whereas the BF did not, suggesting a higher frequency of person-to-person transmissions and/or recent acquisition of microorganisms from a common source7. Estimating the proportion of genotype pairs falling within this nGD = 0 peak by fitting a Gaussian mixture model, we confirmed the fraction of high-similarity pairs to be significantly higher between related participants than non-family members (IF = 5.71% versus BF = 2.06%; $n_{IF} = 2,450$ versus $n_{BF} = 63,287$, respectively).
Fig. 3 | Transmissions of genotypes across family members. a, nGDs between all pairs of genotypes recovered. IF distances (red) present a lower overall median compared with BF distances (blue). b, nGDs for cohabiting individuals peak towards nGD = 0. c, Proportion of total pairwise comparisons with nGDs < 0.1. Top: Comparisons based on relatedness, with higher proportions of IF (x axis) compared with BF (y axis) for all taxa. Bottom: Cohabitation-based comparisons, with cohabiting (x axis) participants showing higher proportions compared to non-cohabiting (y axis) individuals. Taxa are coloured by phylum (Actinobacteria, red; Bacteroidetes, yellow; Euryarchaeota, green; Firmicutes, blue; Proteobacteria, violet; Verrucomicrobia, pink). The dashed line indicates 1:1 proportions. d, e, Maximum likelihood phylogenetic trees of species exhibiting the highest number of IF transmissions (Supplementary Table 13): B. caceae (d) and P. distasonis (e). Nodes are coloured by family ID (colour-coded as in Fig. 1a); family members that exchange strains are shaded. The vertical lines in the histograms indicate nGDs for IF (red) and BF (blue). f, Left: pTRs by relationship (ss, sister; md, mother–daughter; gg, grandmother–granddaughter; gggg, great-grandmother–great-granddaughter) and cohabitation status (green = yes, yellow = no) (Wilcoxon rank-sum test, n = 102, *P adj < 0.05, **P adj < 0.001; Supplementary Table 15). Right: pTRs by kinship for non-cohabiting individuals (Wilcoxon rank-sum test, n = 94, r = 0.24, P = 5.85 × 10−2). g, The body of the box plot represents the first and third quartiles of the distribution and the median line. The whiskers extend from the quartiles to the last data point within 1.5× the IQR, with outliers beyond. g, The pTR between mothers and daughters decreases with the daughter’s age (n = 78 pairs, beta regression, R² = 0.21, z = −3.87, P = 1.11 × 10−4).

Two-proportion test, chi-squared = 86.848, P < 2.2 × 10−15; Fig. 3a). Similarly, family members sharing a household cohabitation presented a significantly higher proportion of closely related genotypes compared to those living apart (LA) (cohabitation = 14.27% versus LA = 1.81%; nIF cohabitation = 633 versus nIF LA = 1,817, two-proportion test, chi-squared = 28.857, P = 7.79 × 10−4; Fig. 3b). This finding aligns with the hypothesis of a higher probability of transmission or co-acquisition of gut microbes among household members due to the closeness and frequency of their contacts. Both within family and household, highly similar genotypes primarily belonged to the phylum Bacteroidetes (Fig. 3c and Supplementary Table 10). Applying a similar approach on ARGs, we additionally computed all pairwise genetic distances between ARG sequences retrieved from CGC individuals (n = 533 ARG clusters). Evaluating the distribution of nGDs between ARG variants within and between families and among family members living together or apart, the differences observed (uncorrected for multiple testing; PERMANOVA, P < 0.05; Supplementary Table 11) corresponded to more closely related sequences shared by family members (12.31%, n = 64 out of 520) and participants living together (15.13%, n = 59 out of 390).
**Articles Nature Microbiology**

**Articles**

The highest average pTRs for members of the order Bacteroidales were identified as significant within the limitations of our cohort (Wilcoxon rank-sum test, $r = 0.73$, $P_{\text{adj}} = 1.64 \times 10^{-2}$; mother–daughter pairs, $r = 0.47$, $P_{\text{adj}} = 1.56 \times 10^{-4}$; Fig. 3f and Supplementary Table 15), again indicative of cohabitation potentially promoting exchange of gut bacteria. However, overall, the pTRs for pairs of non-cohabiting family members was higher compared to non-related individuals (Wilcoxon rank-sum test, $r = 0.24$, $P = 5.85 \times 10^{-2}$; Fig. 3f and Supplementary Table 15). To gain a better understanding of the impact of cohabitation on strain sharing or potential transmission events, we reanalysed a family cohort assembled by Costea et al. consisting of 26 individuals belonging to 6 households (parents and offspring; Extended Data Fig. 6a). Applying the methodology described above, 43 species covering 498 strains were considered eligible for pTR analysis (Supplementary Table 17). Distinguishing between strains being shared among cohabiting related individuals (mother/father–offspring, $n$ pairs $= 28$) and between partners (father–mother, $n$ pairs $= 6$), we found that both categories exhibited higher pTRs than non-related, non-cohabiting individuals ($n = 26$, Kruskal–Wallis test, chi-squared $= 105.65$, $P < 2.2 \times 10^{-16}$; post-hoc Dunn test, $P_{\text{adj}} < 0.01$; Extended Data Fig. 6b and Supplementary Table 18), albeit with smaller effect and sample sizes for partners. Hence, while our analyses identified kinship as a key covariate of genus-level microbiota community differentiation, both CGC and the Costea et al. reanalyses do not exclude cohabitation to be the driving factor in transmission or co-acquisition of individual microbiome features, which is in line with the findings of recent studies on gut ecosystem heritability.

Potentially reflecting the physical intimacy of their relation, we detected the highest average pTRs between mothers and daughters in pairs comprising younger children, with frequencies steadily decreasing with age ($n = 78$ pairs, beta regression, $R^2 = 0.21$, $z = -3.87$, $P = 1.11 \times 10^{-2}$; Fig. 3g)—an association again clearly linked to cohabitation, although the addition of this parameter did not significantly improve the correlation ($R^2 = 0.24$, model comparison likelihood ratio test $P = 0.16$). Also, among the species shared between mothers and young children, the largest pTRs were observed for Bacteroides, notably *B. caccae* (mean = 57.1%), *Bacteroides stercoris* (mean = 33.3%) and *P. distasonis* (mean = 28.57%; Supplementary Table 16). Although our analyses did not allow to resolve directionality, with pTRs also reflecting potential transmission from daughters to mothers, our findings do not contradict the hypothesis of the maternal gut ecosystem being a contributor to primary succession events that constitute microbiota maturation processes in young children. In this respect, given its low colonization resistance, the immature nature of the infant and toddler microbiota can be expected to facilitate inclusion of exogenous microbiome features, acquired through both vertical and horizontal transmission or originating from environmental sources. Finally, we observed four strains belonging to the species *A. onderdonkii* (two strains), *Alistipes shahii* and *B. faecis* to be present in three consecutive generations in four families, potentially reflecting persistent niche colonization across generations (Extended Data Fig. 7). In addition, the strains of four other species were detected across three non-consecutive generations in three families. *B. salyersiae* and *P. distasonis* remained undetected in one of the intermediate levels, while a different strain of *B. caccae* and *Eubacterium eligens* were found at the grandmother level (Extended Data Fig. 7).

**Conclusion**

Our explorative analyses of gut microbiota variation across generations confirmed the microbiome of young children to be fundamentally divergent from more developed configurations, with familial community structures only emerging on ecosystem maturation. Although the impact of kinship was additionally reflected in a higher frequency of strain sharing between family members...
compared to unrelated individuals, estimations of pTRs identified cohabitation as a key covariate of strain distribution. In line with these findings, we observed IF pTRs to decrease both with degree of kinship and age difference, with potential transmission events across generations being rare but detectable. Shared strains predominantly belonged to the Bacteroidales order. Overall, while our analysis does not exclude cross-generational transmission of strains resulting from maternal inheritance, strain sharing was most frequently detected among first-degree relatives sharing a household.

Methods

Ethical compliance. All experimental protocols were approved by the Medical Ethics Committee Universiteit Ziekenhuis Brussel- Vrije Universiteit Brussel (BUN 14312015/501/05) and the Ethical Committee for Medical Ethics of Katholieke Universiteit Leuven (S58125). Study design complied with all relevant ethical regulations, aligning with the Declaration of Helsinki (2013 version) and in accordance with Belgian privacy legislation. Written informed consent was obtained from all adult participants and from the parents of underage participants. Participants did not receive compensation for their participation in the study.

Sample collection. The cohort included 102 female participants belonging to families with at least 3 generations of women (n = 24 families, median = 4 generations per family). Sampling took place before November 2015 and November 2016 and all participants signed a statement of informed consent. A limited set of data, including participant’s birth date, height, weight, delivery mode, antibiotic use over the last months and family structure was collected at enrolment (Supplementary Table 1). Faecal sample collection and blood analyses were performed as in Falony et al.10. Briefly, participants were asked to collect their faecal material (single defection) in a plastic vial, place the vial in a labelled non-transparent ziplock bag and freeze it at −20 °C immediately after collection. Frozen samples were transported within 72 h to the research facility and stored at −80 °C. Blood samples were drawn by a study nurse and analysed by an independent certified clinical laboratory (Centrum voor Medische Analyse, Belgium). Participants were asked to refrain from calorie intake for 8 h before blood sampling.

Statistics and reproducibility. While no statistical method was used to predetermine sample sizes for the present exploratory study, CCG cohort size was similar to the number of participants included in previous publications12–14. Data exclusions are specified and justified for each of the analyses presented. Experiments were not randomized in this explorative cross-sectional study but the Costea et al.15 dataset was used to replicate findings. No intervention was performed on participants; thus, they were not randomly allocated into study groups. Data collection and analysis were not performed blinded to the conditions of the study set-up.

Faecal sample characterization. To assess microbial loads in faecal samples, 0.2 g frozen (−80 °C) aliquots were diluted 100,000 times in physiological solution (8.5 g l−1 NaCl; VWR International). Samples were filtered using a sterile syringe filter (5 μm pore size; Sartorius Stedim Biotech) and 1 ml of the resulting microbial cell suspension was stained with 1 μl of SYBR Green I (1:100 dilution in dimethyl sulfoxide; shaded 15 min incubation at 37 °C; 10,000 concentrate; Thermo Fisher Scientific). Microbial cell count (n = 101; Supplementary Table 1) was performed using an Accuri C6 flow cytometer (BD Biosciences) based on Prest et al.40. Fluorescence events were recorded using the FL1 533/30 nm and FL3 > 670 nm optical detectors; forward and sideward scattered light signals were collected. The BD Accuri Cflow software v.1.0.264.21 was used to gate and separate the microbial fluorescence events on the FL1/FL3 density plot from the faecal sample background. A threshold value of 2,000 was applied to the FL1 channel. To exclude any remaining background events, gated fluorescence events were evaluated on the forward/sideward density plot. Instrument and gating settings were kept identical for all samples (fixed staining/gating strategy40; Extended Data Fig. 8). Cell counts were converted to microbial loads per gram faecal material based on the exact weight of the aliquots. Measurements were performed in duplicate; if the number of events recorded differed by more than 10%, a third replicate was measured. One sample was excluded from cell counting due to insufficient faecal material to perform the measurements. Moisture content was determined as the percentage of mass loss excluded from cell counting due to insufficient faecal material to perform the measurements. Moisture content was determined as the percentage of mass loss.

DNA extraction, sequencing and data preprocessing. Faecal DNA extraction and microbiota profiling was performed as described previously41. Briefly, DNA was extracted from faecal material using the MoBio PowerMicrobiome RNA Isolation Kit, with the addition of 10 min incubation at 90 °C after the initial vortexing step. For amplicon sequencing, the V4 region of the 16S rRNA gene was amplified with the primer pair 515F/806R42. Sequencing was performed on the Illumina MiSeq platform to generate paired-end reads of 250 bases in length in each direction. 16S data preprocessing was performed using QIIME2 v.1.56.5 to demultiplex the sequencing reads. Amplicon sequencing was used only for community typing to align with the IFGFP dataset.

Whole-metagenome shotgun sequencing was performed using the Illumina HiSeq 2500 system (151 bp paired-end reads; Novogene). Paired-end reads were first quality-checked using fastq: v0.11.2 and Illumina adaptors and low-quality reads were removed using Trimmomaticv.0.32 with the options ILLUMINACLIP:trimmomatic-0.32/adapters/NexteraPE-PE.fa:2:30:10, IIADAPTER:101:80:70, HEADCROP:15 and MINLEN:40. High-quality reads were then decontaminated from phiX and human sequences using DeconSeqv0.4.3 and removed broken pairs of reads (pairs for which one member was removed during filtering) were identified and removed using a custom script, available at https://github.com/raeslab/raeslab-utils/.

Relative and quantitative microbiome taxonomic profiling. Taxonomical assignment of preprocessed 16S data was performed using the DADA2v.1.6.0 and the RDP classifierv.2.12 with default parameters. To obtain the 16S relative microbiome profiling (rMP) matrix, each sample was downsized to 10,000 reads by random selection of samples. Samples with less than 10,000 reads were excluded (1 sample) from the analyses.

Using sequencing data decontaminated from phiX and human sequences to generate the shotgun QMP matrix, shotgun sampling size was defined as the average abundance of ten universal single-copy marker genes of the MOCA2v pipeline (COG0012, COG0016, COG0018, COG1017, COG2015, COG0495, COG0525, COG0533, COG0541, COG0552). Paired-end reads were downsized to equivalent sampling depth (ratio between sampling size and microbial load, that is, the average total cell count per gram of frozen faecal material) by random selection of the reads to equate the minimum observed sampling depth in the dataset (minimum sampling depth = 4.98 × 104). The resulting rarefied read counts were above 1.3 × 104 reads for all samples. Next, taxonomical classification of the rarefied read counts was performed using the RDP classifierv.2.12 based on the abundances of the single-copy marker genes, with default parameters and skipping any filtering or trimming steps. mOTUs were then aggregated into species and genera using mOTU taxonomic annotation (mOTU.v1 database). Microbiome profiles were converted to the numbers of cells per gram by dividing the total mOTU linkage group abundance in the sample (including mOTUs with no phylogenetic assignment) and multiplying by the number of cells per gram of faeces. In addition, taxonomic profiling at the species and strain levels were performed using MetaPhAn2v and StrainPhAn2.43. Briefly, the preprocessed metagenomic read matrices were mapped against the MetaPhAn2 marker database using the metaplant2 script with default parameters. Then, samples/markers.py was run to produce the gene marker file for each sample; gene marker files were parsed to StrainPhAn2 to identify the taxa detected in each metagenomic sample. Rarefied abundances at the genus, species and strain levels were also converted into number of cells per gram as described for the mOTUs.

Quantitative microbiome functional profiling. QMP rarefied reads were mapped on the integrated gene catalogue (IGC)44 using the Burrows–Wheeler Alignerv.0.7.8 and the mapping was summarized into functional profiles by featureCounts v.1.5.3, with the parameters --minOverlap 40. Gut metabolic module (GMM)45 abundances were computed using Omixzer-RPM v.1.0 (https://github.com/raeslab/omixzer-rpm), with option <0.66 (66% coverage detection threshold). Coverage of the most abundantly curated metabolic pathways was used to determine the number of pathway steps for which at least one of the orthologous groups is found in a metagenome, divided by the total number of steps constituting the module. The rarefied reads mapped on the IGC were also annotated with ARGs using the Comprehensive Antibiotic Resistance Database46. GMM and ARG abundances were converted to quantitative abundance profiles (abundance per genome of faeces) by dividing by total mOTU linkage group abundance in the sample (including mOTUs with no phylogenetic assignment) and multiplying by the number of cells per gram of faeces.

Identification of species-representative genotypes. To identify the species genotypes in the dataset, we used StrainPhAn on the original, non-rarefied reads to produce covered core alignments of marker genes as indicated above. As such, the consensus genetic sequence resulting from the concatenation of marker genes for each species and individual is referred to as genotype. Taxonomic groups corresponding with phases, viruses and viroids were discarded from further analysis. Gaps were removed from the alignments using T-Coffee58 v.11.00 with option -action =rm_gap 1 so that only the covered core genome for that particular comparison was analysed; SNP-sites v.2.5.1 was used to obtain the alignments of SNPs. Only alignments that contained 3 or more samples from at least 1 family and core genome sizes of 1,000 bp were kept.

Genetic distances and phylogenetic analysis. Core genome alignments were used to compute the pairwise genetic distances between all genotypes of each species by using snp-distsv.0.6 (https://github.com/seeemann/snp-distsv). The genetic distances, calculated as the number of SNPs between pairs of genotypes, were divided by the length of the core genomes to obtain the number of SNPs
per megabase. In addition, distances were normalized by the median genetic distance of each taxon (nGDs). We considered that two genotypes belonged to the same strain if their nGD was below the stringent threshold of 0.10, as used by others. To reconstruct the phylogenetic tree from the previously obtained core genome alignments, we used RAxML v.8.2.12 with the parameters -f a and -m GTRGAMMA. For the phylogenetic trees obtained with thestrainphln.py script, we set bootstrap_raxml to 100 and marker_in_clade to 0.2. Phylogenetic trees were rooted midpoint with the package ETE 3. Finally, PhyloPhAn v.3.60.80 was used to produce a phylogenetic tree of all the species profiled using MetaPhAn2 and the associated metadata were plotted using iTOL v6. For the 51 species analysed at the strain level, we selected proximal representatives of taxa absent in the PhyloPhAn database.

Antimicrobial resistance genes. The presence of sequence-identical antimicrobial resistance genes (ARGs) across individuals was assessed by extracting consensus sequences corresponding with ARGs from the IGC alignment, filtering by gene length coverage above 99% and 5 reads of minimum depth. Next, for each gene, we computed the pairwise genetic distances between pairs of individuals, as described for genotypes.

Statistical analyses. Statistical analyses were performed in R using the packages vegan v.2.5-6, phyloseq v.1.32.0, FSA v.0.8.30, coin v.1.3-1, DirichletMultinomial v.1.3.0, kinship2 (ref. 7) v.1.8.5, FamAgg v.1.16.0, QuantPsyc v.1.5, gmm v.1.6.5 and ggplot2 (ref. 7) v.3.3.2. Non-parametric statistical tests were used because data did not follow normality or equal variance assumptions. All P values were corrected for multiple testing using the Benjamini–Hochberg method (reported as P_{adj}) unless specified otherwise and significance was defined as P < 0.05 and P_{adj} < 0.05.

Microbiota community variation explained by metadata variables. Contribution of metadata variables (age, SBMI, delivery mode, family ID, cohabitation status, medication use, antibiotic use, moisture content (%) and faecal calprotectin (μg g^-1)) to interindividual microbiota community variation was determined by single dbRDA on genus-level Bray–Curtsis dissimilarity with the capsule function in the vegan R package. The cumulative contribution of metadata variables was determined by forward model selection on dbRDA with the ordiR2step function in vegan, with variables that showed a significant contribution to microbiota community variation (P_{adj} < 0.05) in the previous step.

Faecal microbiome-derived features and visualization. Observed genus richness was calculated on the QMP matrix using phyloseq. Enterotyping (or community typing) based on the DMN approach was performed in R using the DirichletMultinomial v.1 package as described by Holmes et al. on the RMP matrix. To increase accuracy, enterotyping was performed on a combined genus abundance matrix including the present dataset (n = 101) complemented with 1,100 samples from the FGPG cohort rarefied to 10,000 reads. Microbiome interindividual variation was visualized by principal coordinate analysis (PCoA) using Bray–Curtsis dissimilarity on the genus-level abundance matrix. The optimal number of Dirichlet components based on the Bayesian information criterion was four. The four FGPG clusters were named Prevotella, Bacteroides 1, Bacteroides 2 and Ruminococcaceae as described by Vanderputte et al. The first has high relative abundance of Prevotella and the fourth has the highest genus-level richness, while the other two are dominated by the Bacteroides genus, with Bacteroides 2 also harbouring reduced Faecalibacterium abundance.

Microbiome and metadata associations. Taxa unclassified at the genus level or present in less than 10% of samples were excluded from the statistical analyses. Spearman correlations were used for rank–order correlations between continuous variables, including genera abundances, microbial loads, CRP and age. Wilcoxon rank-sum tests were used to test the differences of continuous variables between two different groups. For more than two groups, Kruskal–Wallis tests were used to test the differences of continuous variables between two different groups. For more than two groups, Kruskal–Wallis tests with post-hoc Dunn tests were applied. Statistical differences in the proportions of categorical variables (enterotypes) among groups were evaluated using pairwise chi-squared tests. Spearman correlation analyses were used to identify correlations between pTRs and continuous variables, including family size, carriers and prevalence of species in the population. To model the pTRs between mother and daughter pairs in relation to age, a generalized regression with beta response distribution (for response variables bound between 0 and 1) was fitted by maximum likelihood (betagen function in the betareg R package). The pTRs, with range 0–1, were transformed to obtain rates in the range 0–1 as pTR = (pTR(n−1)+1)/n, with s = 0.5 as recommended for betagen model. Nested model comparison was performed using a likelihood ratio test (lrtest in the lmtest R package) v.0.9.38. For comparisons of pTRs between different types of kinship (sister, mother–daughter, grandmother–granddaughter), a Wilcoxon rank-sum test was used if only two groups were compared; a Kruskal–Wallis test with post-hoc Dunn test was used if more than two groups were compared.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
The raw amplicon sequencing data and shotgun metagenomics sequencing data reported in this study have been deposited in the European Genome-phenome Archive under accession nos. EGAS00001005651 and EGAS00001005649.

Code availability
The custom script used to identify and remove broken pairs of reads (pairs for which one member was removed during filtering) is available at https://github.com/raeslab/raeslab-utils/.

Received: 6 July 2021; Accepted: 10 November 2021; Published online: 30 December 2021

References
1. Mueller, N. T. et al. Bacterial baptism: scientific, medical, and regulatory issues raised by vaginal seeding of C-section-born babies. J. Law Med. Ethics 47, 568–578 (2019).
2. Ferretti, P. et al. Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome. Cell Host Microbe 24, 133–145.e5 (2018).
3. Yassour, M. et al. Strain-level analysis of mother-to-child bacterial transmission during the first few months of life. Cell Host Microbe 24, 146–154.e4 (2018).
4. Räckhed, F. et al. Dynamics and stabilization of the human gut microbiome during the first year of life. Cell Host Microbe 17, 690–703 (2015).
5. Shao, Y. et al. Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. Nature 574, 117–121 (2019).
6. Brito, I. L. et al. Transmission of human-associated microbiota along family and social networks. Nat. Microbiol. 4, 964–971 (2019).
7. Sonnenburg, E. D. et al. Diet-induced extinctions in the gut microbiota compound over generations. Nature 529, 212–215 (2016).
8. Morimoto, J., Simpson, S. J. & Ponton, F. Direct and trans-generational effects of male and female gut microbiota in Drosophila melanogaster. Biol. Lett. 13, 20160966 (2017).
9. Moeller, A. H., Suzuki, T. A., Phifer-Rixey, M. & Nachman, M. W. Transmission modes of the mammalian gut microbiota. Science 362, 453–457 (2018).
10. Zeller, M. & Daniels, S. The obesity epidemic: family matters. J. Pediatr. 145, 10–14 (2004).
11. Santos, M. P. C., Gomes, C. & Torres, J. Familial and ethnic risk in inflammatory bowel disease. Ann. Gastroenterol. 31, 14–23 (2018).
12. Kromeyer-Hauschild, K. et al. Perzentile für den Body-mass-Index für das Kindes- und Jugendalter unter Heranziehung verschiedener deutscher Stichproben. Monatsschr. Kinderheilkd. 149, 807–818 (2001).

| VOL 7 | JanUaRy 2022 | 87–96 | www.nature.com/naturemicrobiology |
Acknowledgements
We thank all study participants for their commitment, M. Joossens, L. De Sutter and S. Janssens for assisting in sample collection and L. Rymenans, C. Verspecht and T. T. D. Nguyen for their efforts in sample analysis.

Author contributions
J.R. and G.F. conceived the study. M.V.-C., R.B., S.V.-S., Y.D., J.R. and G.F. designed the experiments. M.V.-C. performed the flow cytometry analysis and determined the moisture content. M.V.-C. and R.B. performed the metagenomic and genome-based analyses. M.V.-C., R.B., S.V.-S., S.S., R.Y.T., T.Y., N.S., J.R. and G.F. planned and executed the statistical analyses. M.V.-C., R.B., S.V.-S., J.R. and G.F. drafted the manuscript. All authors revised the article and approved the final version for publication. M.V.-C., R.B., S.V.-S. and R.Y.T. are funded by (post)doctoral fellowships from the Research Fund-Flanders (Fonds Wetenschappelijk Onderzoek-Vlaanderen (FWO) 1110918N, 1221620N, 12K5116N and 1234321N, respectively). S.S. is supported by a Japan Society for the Promotion of Science KAKENHI grant no. 17J10014. The Raes lab is supported by the Vlaams Instituut voor Biotechnologie (VIB), Katholieke Universiteit (KU) Leuven, Rega Institute for Medical Research and by the FWO and Fonds de la Recherche Scientifique under EOS Project no. 30770923.

Competing interests
M.V.-C., S.V.-S., J.R. and G.F. and are inventors on the patent application PCT/EP2018/084920 in the name of VIB VZW, KU Leuven, KU Leuven R&D and Vrije Universiteit Brussel covering microbiome features associated with inflammation described in Vieira-Silva et al. The other authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41564-021-01021-8.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-021-01021-8.

Correspondence and requests for materials should be addressed to Jeroen Raes.

Peer review information Nature Microbiology thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021
Extended Data Fig. 1 | Enterotype stratification by DMM community typing. (a) Identification of the optimal number of clusters (Dirichlet components) in the CGC dataset (N = 101) complemented with 1106 samples from the FGFP cohort based on the Bayesian Information Criterion (BIC). (b) Barplot representation of the average relative abundance of a few representative genera split into the four enterotypes identified by DMM community typing on the combined CGC and FGFP sets (N = 1207).
Extended Data Fig. 2 | Enterotype associations with microbial load, richness, and moisture content. (a) Reduced microbial load (microbial cells per gram of stool) in Bact2 enterotyped samples; N = 101, KW test Chi² = 13.9, P = 3.0e-03; phD tests, adjP < 0.001(***), < 0.01(**), < 0.05(*); Supplementary Table 3. (b) Reduced richness (number of genera) in Bact2 enterotyped samples; N = 101, KW test Chi² = 20.0, P = 1.6e-04; phD tests, adjP < 0.001(***), < 0.01(**), < 0.05(*); Supplementary Table 3. (c) Increased stool moisture content (%) in Bact2 enterotyped samples; N = 101, KW test Chi² = 8.8, P = 0.03; phD tests, adjP < 0.001(***), < 0.01(**), < 0.05(*); Supplementary Table 3. The body of all box plots represent the first and third quartiles of the distribution and the median line. The whiskers extend from the quartiles to the last data point within 1.5× the interquartile range, with outliers beyond.
Extended Data Fig. 3 | Maximum-likelihood phylogenetic tree of the species present within the CGC cohort analysed using StrainPhlAn. Branches are coloured by phyla (Actinobacteria, red; Archaea, green; Bacteroidetes, yellow; Firmicutes, dark green; Proteobacteria, blue; Verrucomicrobia, pink). Boxes represent % prevalence (P, blue), % relative abundance (A, pink), and potential transmission rates (pTR, yellow).
Extended Data Fig. 4 | Distribution of normalized genetic distances for each of the 51 species analysed. Distances between pairs of genotypes recovered from individuals of the same families are coloured in red, and those from individuals of different families are coloured in blue. Dashed vertical lines represent the threshold used to define two genotypes belong to the same strain (nGD < 0.10; in black), intra-family median distances (red) and between-family median distances (blue).
Extended Data Fig. 5 | Phylogenetic trees of species for which strain sharing was detected between at least two members of the same family. Tips are colored by family IDs as in Fig. 1 and shapes indicate generation number (0 = Circle, 1 = Triangle, 2 = Square, 3 = Cross).
Extended Data Fig. 6 | Summary results for cohabitating, non-related individuals. (a) Family structures in the Costea et al. study (N = 26). (b) pTRs by relationship (KW, N = 26, Chi2 = 105.65, P < 2.2e-16; PhD tests for IF adjP > 0.05; IF groups vs unrelated adjP < 0.001(**), < 0.01(**); Supplementary Table 18). The body of the box plot represents the first and third Quartiles of the distribution and the median line. The whiskers extend from the quartiles to the last data point within 1.5x the interquartile range, with outliers beyond.
Extended Data Fig. 7 | Strains shared across three consecutive generations (top) and across four generations but missing one intermediate level (bottom). Numbers indicate family IDs for which strain sharing over more than two generations was observed.
Extended Data Fig. 8 | Illustration of flow cytometry gating strategy. A fixed gating/staining approach was applied. Both blank and sample solutions were stained with SYBR Green I. (a) FL1-A/FL3-A acquisition plot of a blank sample (0.85% w/v physiological solution) with gate boundaries indicated. A threshold value of 2000 was applied on the FL1 channel. (b) Secondary gating was performed on the FSC-A/SSC-A channels to further discriminate between debris/background and microbial events. (c, d) FL1-A/FL3-A count acquisition of a faecal sample with secondary gating on FSC-A/SSC-A channels based on blank analyses. Total counts were defined as events registered in the FL1-A/FL3-A gating area excluding debris/background events observed in the FSC-A/SSC-A R1 gate. The flow rate was set at 14 microliters per minute and the acquisition rate did not exceed 10,000 events per second. Each panel reflects the events registered during a 30 s acquisition period. Cell counts were determined in duplicate starting from a single biological sample.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
- Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

No software was used for microbiome data collection. To assess faecal microbial loads, flow cytometry analysis was performed using the BD Accuri CFlow software (v1.0.264.21) for gating and event counting.
16S data pre-processing was performed using LotuS (version 1.5.65, used for demultiplexing sequencing reads) and the DADA2 pipeline (version 1.6.0), with the RDP classifier (version 2.12) for taxonomy assignment.

For analysis of shotgun sequencing data, paired-end reads were quality checked using fastqc (version 0.11.2), and Illumina adapters and low-quality reads trimmed with Trimmomatic (version 0.32), decontaminated from phiX and human sequences using DeconSeq (version 0.4.3), and broken pairs were fixed using a custom Biopython script (available at https://github.com/raeslab/raeslab-utils/).

Taxonomic classification of the rarefied reads into mOTUs was performed with MOCAT2 (version 2.0.1). Taxonomic profiling at the species and strain levels were performed using MetaPhlAn2 and StrainPhlAn2. Core alignments were computed by removing gaps using T-Coffee v11.00, and SNP-sites v2.5.1 was used to obtain SNP alignments. Pairwise genetic distances between all genotypes of each species were computed with snp-dists v0.6. Phylogenetic trees were computed with RAxML v8.2.12.

For functional profiling, QMP-rarefied reads were mapped on the integrated gene catalogue (IGC) using BWA (version 0.7.8), and the mapping was summarized into functional profiles by featureCounts (version 1.5.3, with parameters --minOverlap 40 –pO). GMM (gut metabolic module) abundances were computed using Omixer-RPM v1.0 (https://github.com/raeslab/omixer-rpm).

Data analysis and graphical representations were performed using R, a free software environment for statistical computing, with packages vegan (v2.5.6), phyloseq (v1.32.0), FSA (v0.8.30), coin (v1.3.1), DirichletMultinomial (v1.30.0), kinship2 (v1.8.5), FamAgg (v1.16.0), QuantPsyc (v1.5), gmm (v1.6.5), ggplot2 (v3.3.2), and lmtest (v0.9.38).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw amplicon sequencing data and shotgun metagenomics sequencing data reported in this study have been deposited in European Genome-phenome Archive with accession codes EGAS00001005651 and EGAS00001005649 respectively.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | No sample size calculation was performed prior to cohort recruitment. This is the first study on multigeneration microbiota transmission and therefore no previous information was available. However, previous studies on mother-infant transmission showed similar sample sizes were sufficient (PMID: 28144631, 30001516, 30001517). |
| Data exclusions | No data were excluded from the analyses. |
| Replication | We used a published dataset (Costea et al, 2017) to replicate the main findings of the study. |
| Randomization | Not applicable: this was a cross-sectional study limited in size, not a randomized study. No intervention was performed on subjects, and therefore no random allocation into groups. |
| Blinding | This was a data-driven cross-sectional study. As in similar studies, data collection and analysis were not performed blinded to the conditions of the study set-up. |

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.
Human research participants

Policy information about studies involving human research participants

Population characteristics

A complete description of the study participants can be found in Table S1.
- 102 generally-healthy female individuals belonging to 24 families ([3:5] generations/family) were enrolled in the study.
- Age range: [0:98]
- Vaginal delivery: 99 yes:3 no

Recruitment

The FGFP recruitment channels (social media, newsletters, appearances in popular media) were used to enroll any interested women from families with at least three generations of women in Flanders. All families who volunteered to participated and followed the inclusion criteria were included in the studies, no recruitment bias is expected. Recruitment took place between November 2015 and November 2016.

Ethics oversight

All experimental protocols were approved by the Medical Ethics Committee UZ Brussels-VUB (BUN 143201215505) and the Commissie Medische Ethiek, UZ/KU Leuven (S58125). Study design complied with all relevant ethical regulations, aligning with the Declaration of Helsinki and in accordance with Belgian privacy legislation. Written informed consent was obtained from all adult participants, and from the parents of under-aged participants.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:
- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a ‘group’ is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

0.2 g frozen (-80°C) faecal aliquots were dissolved in physiological solution (8.5 g/L NaCl; VWR International, Germany) to a total volume of 100 mL. Subsequently, the slurry was diluted 1000 times. Samples were filtered using a sterile syringe filter (pore size of 5 μm), and 1 mL of the resulting microbial cell suspension was stained with 1 μL SYBR Green I (1:100 dilution in DMSO; shaded 15 min incubation at 37°C; 10,000 concentrate).

Instrument

C6 Accuri flow cytometer (BD Biosciences, New Jersey, USA).

Software

BD Accuri CFlow software v1.0.264.21 (BD Biosciences, New Jersey, USA).

Cell population abundance

Not applicable: no sorting of the fractions was performed.

Gating strategy

Fluorescence events were monitored using the FL1 533/30 nm and FL3 >670nm optical detectors. In addition, forward and sideward-scattered light were collected. The BD Accuri CFlow software was used to gate and separate the microbial fluorescence events on the FL1/FL3 density plot from background. A threshold value of 2000 was applied on the FL1 channel. The gated fluorescence events were evaluated on the forward/sideward density plot, to exclude remaining background events. Instrument and gating settings were kept identical for all samples.

Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.