



OPEN

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

Mireia Valles-Colomer ^{1,2,5,6}, Rodrigo Bacigalupe ^{1,2,6}, Sara Vieira-Silva ^{1,2,6}, Shinya Suzuki ^{1,3}, Youssef Darzi ^{1,2}, Raul Y. Tito ^{1,2}, Takuji Yamada ³, Nicola Segata ^{4,5}, Jeroen Raes ^{1,2,6} and Gwen Falony ^{1,2,6}

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, respectively¹. 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/hamper transmission can be formulated¹. Given reports of maternal inheritance of microbial strains^{2–5}, strain sharing among individuals sharing households⁶ and transmission events spanning multiple generations in animal models^{7–9}, similar considerations might apply when assessing the familial burden of conditions with a potential microbiota contribution, ranging from obesity¹⁰ to inflammatory bowel diseases¹¹.

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 indices (SBMI) of participants (an age- and sex-corrected version of the body mass index valid also in children^{12,13}) 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 cohousing 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 load¹⁴ 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, $R^2 = 4.3\%$, $P_{\text{adj}} = 2 \times 10^{-4}$) and age ($R^2 = 2.9\%$, $P_{\text{adj}} = 2 \times 10^{-4}$) 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 cohorts¹⁵, with stool moisture, a proxy of colonic transit time¹⁶, reflecting ecosystem development induced by nutrient depletion on passage through the gastrointestinal tract¹⁷. Additionally, we confirmed the negative associations between faecal water content and microbial load ($n = 101$, Spearman's test, $\rho = -0.25$, $P = 1.2 \times 10^{-2}$) as well as genus-level

¹Laboratory of Molecular Bacteriology, Department of Microbiology and Immunology, Rega Institute, Katholieke Universiteit Leuven, Leuven, Belgium.

²Center for Microbiology, Vlaams Instituut voor Biotechnologie, Leuven, Belgium. ³School of Life Science and Technology, Tokyo Institute of Technology, Tokyo, Japan. ⁴European Institute of Oncology Istituto di Ricovero e Cura a Carattere Scientifico, Milan, Italy. ⁵Present address: Department for Integrative Biology, University of Trento, Trento, Italy. ⁶These authors contributed equally: Mireia Valles-Colomer, Rodrigo Bacigalupe, Sara Vieira-Silva, Jeroen Raes, Gwen Falony. e-mail: jeroen.raes@kuleuven.be

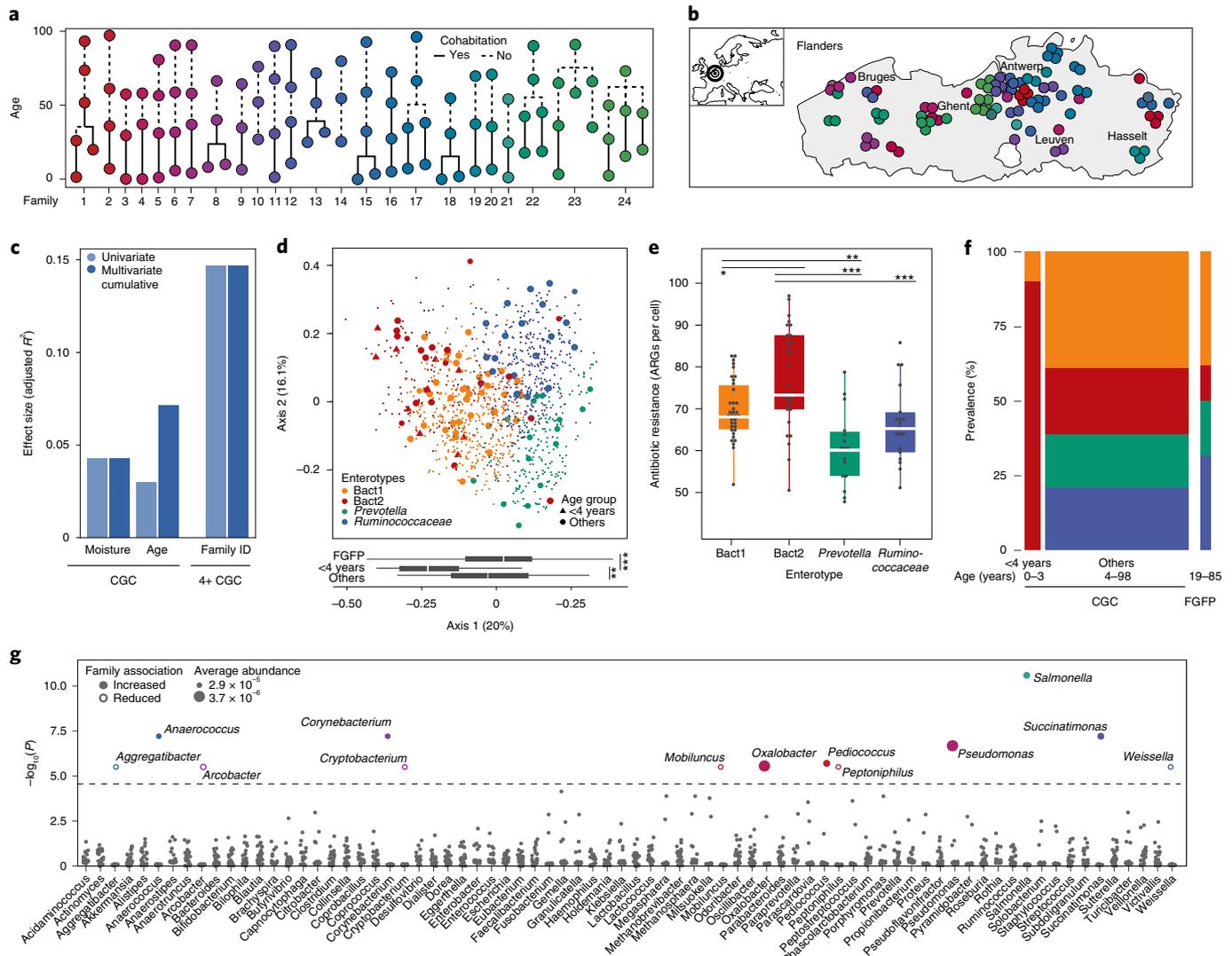


Fig. 1 | Familial structures and microbiome profiles. **a**, Family structures in the CGC cohort ($n=102$). The colours represent family IDs and the line type indicates current cohabitation status. **b**, Geographical distribution of the participants across Flanders (Belgium). Family IDs are coloured as in Fig. 1a. **c**, Cumulative effect sizes of significant covariates on microbiome community variation (dark blue, right bars; stepwise dbRDA on Bray-Curtis dissimilarity) compared to individual effect sizes assuming covariate independence (light blue, left bars; dbRDA on Bray-Curtis dissimilarity) in the CGC ($n=101$) and the ≥ 4 -year-old CGC (4+ CGC, $n=91$) cohorts (Supplementary Table 2). Age, SBMI, delivery mode, family ID, cohabitation status, medication use, antibiotic use, moisture content (%) and faecal calprotectin ($\mu\text{g g}^{-1}$) were tested as potential microbiome covariates. **d**, PCoA of interindividual differences (Bray-Curtis) in relative microbiome profiles of the CGC cohort ($n=101$ samples, larger dots) on a background dataset from a cross section of the Flemish population ($n=1,106$ samples, small dots). The distribution of samples along the first axis of the PCoA (bottom box plots) separated young children (<4 years old, $n=10$) from other individuals ($n=91$) and FGFP samples ($n=1,106$); Kruskal-Wallis with post-hoc Dunn test, $^{**}P_{\text{adj}} < 0.01$, $^{***}P_{\text{adj}} < 0.001$. 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 $1.5\times$ the interquartile range (IQR), with outliers beyond. Bact1: *Bacteroides* 1; Bact2: *Bacteroides* 2. **e**, Increased ARG load (n ARG/microbial load in Bact2 enterotyped samples; $n=101$, Kruskal-Wallis test, chi-squared = 26.7, $P=6.8\times 10^{-6}$; post-hoc Dunn test, $^{*}P_{\text{adj}} < 0.05$, $^{**}P_{\text{adj}} < 0.01$, $^{***}P_{\text{adj}} < 0.001$; Supplementary Table 3). 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 $1.5\times$ the IQR, with outliers beyond. Colour coding as in Fig. 1d. **f**, Increased prevalence of the Bact2 enterotype in young children ($n=10$) compared to other individuals ($n=91$) and FGFP samples ($n=1,106$); pairwise chi-squared test, $P_{\text{adj}} < 0.1$; Supplementary Table 4. Colour coding as in Fig. 1d. **g**, Seven bacterial genera displayed significantly higher abundances among members of specific families compared to the rest of the 4+ CGC dataset ($n=91$, two-sided Wilcoxon rank-sum test, $-\log_{10}(P) > 4.56$; Supplementary Table 5). The colours correspond to family IDs (as in Fig. 1a), while the circle sizes are proportional to the average abundance of the genus. The closed circles indicate genera with significantly increased abundances in a specific family; the open circles indicate those with decreased abundances in specific families.

microbiome richness ($n=101$, $\rho = -0.29$, $P = 3.7\times 10^{-3}$)¹⁸. 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^{19,20}, we assessed a potential association between age bins (young children <4 years old, $n=10$ versus others

(≥ 4 years old CGC, 4+ CGC), $n=91$) and quantitative genus-level microbiota composition. In a multivariate model, age bins were shown to have the largest effect size ($n=101$, genus-level stepwise dbRDA, $R^2=8.0\%$, $P_{\text{adj}}=1 \times 10^{-4}$; Supplementary Table 2), with moisture contributing an additional 1.7% to microbiome variance ($P_{\text{adj}}=1.1 \times 10^{-2}$). Hence, we confirmed that young children harbour a markedly different microbiota when compared to individuals with a fully matured colon ecosystem^{19,20}.

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 obesity²¹, inflammatory bowel disease^{14,22} and primary sclerosing cholangitis²², as well as among individuals with certain subtypes of multiple sclerosis²³ and depression²⁴. 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 markers²². To distinguish community states within the present CGC, we performed Dirichlet multinomial mixture (DMM) modelling²⁵ against the background of microbiome variation as observed in the Flemish Gut Flora Project (FGFP) dataset ($n=1,106$ population cohort)²⁶. 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 procedures²⁶. 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, $P=3.0 \times 10^{-3}$; post-hoc Dunn test, $P_{\text{adj}} < 0.05$ for Bact2 versus Bact1/*Prevotella*), lower genus-level richness ($n=101$, Kruskal–Wallis test, chi-squared=20.0, $P=1.6 \times 10^{-4}$; post-hoc Dunn test, $P_{\text{adj}} < 0.05$ 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, $P_{\text{adj}} < 0.05$ for Bact2 versus *Ruminococcaceae*; Extended Data Fig. 2 and Supplementary Table 1). With only a single participant scoring above the serum C-reactive protein (CRP) clinical threshold ($>15 \text{ mg l}^{-1}$) and 11 above the faecal calprotectin one ($>200 \mu\text{g g}^{-1}$; Supplementary Table 1)²⁷, we did not observe higher prevalence of systemic nor intestinal inflammation among participants hosting Bact2 ($n=77$, chi-squared=1.8, $P > 0.05$; $n=99$, chi-squared test=7.1, $P > 0.05$) in contrast with previous reports regarding associations in specific patient groups²². Although enterotype stratification was not significantly associated with participants' history of antimicrobial drug intake ($n=95$, chi-squared=5.0, $P > 0.05$), low microbial load Bact2 samples were proportionally enriched in antimicrobial resistance genes (ARGs; $n=101$, Kruskal–Wallis test, chi-squared=26.7, $P=6.8 \times 10^{-6}$; post-hoc Dunn test, $P_{\text{adj}} < 0.05$ for Bact2 versus Bact1/*Prevotella*/*Ruminococcaceae*; Fig. 1e and Supplementary Table 3). Enterotype distribution in the CGC cohort differed significantly from the proportions observed in the FGFP population cohort ($n=101$ versus $n=1,106$, chi-squared test=25.9, $P=9.98 \times 10^{-6}$), with the present dataset being characterized by a higher prevalence of Bact2 samples (29% versus 12%; pairwise chi-squared test=22.3, $P_{\text{adj}}=9.1 \times 10^{-6}$; Supplementary Table 4). More specifically, young children (<4 years old) displayed a

markedly higher prevalence of Bact2 configurations than observed both in the FGFP (90% versus 12%; $n=10$ versus 1,106, pairwise chi-squared test=48.7, $P_{\text{adj}}=1.2 \times 10^{-11}$) and the 4+ CGC (90% versus 22%; $n=10$ versus $n=91$, pairwise chi-squared test=16.9, $P_{\text{adj}}=1.6 \times 10^{-4}$; Fig. 1f and Supplementary Table 4).

Exclusion of young children reveals familial patterns in microbiota variation. To characterize cross-generational familial microbiome similarity, we first assessed variation in abundance patterns of microbial taxa and functions within and between families. Given their diverging microbiomes (Fig. 1d,f), we opted to exclude participants younger than 4 from these analyses, leaving us with an $n=91$ cohort (4+ CGC). Exclusion of young children resulted in family identifier being the sole significant microbiome covariate, accounting for 14.7% of genus-level compositional variation ($n=91$, genus-level stepwise dbRDA, $P_{\text{adj}}=5.5 \times 10^{-3}$; Supplementary Table 2 and Fig. 1c), exceeding the effects sizes of previously identified microbiome covariates²⁶. Hence, we conclude that among women with a mature colon microbial ecosystem, family-bound phylogenetic microbiome community patterns can be identified over multiple generations. Remarkably, no such significant association with family was observed when assessing interindividual variation in abundance patterns of core microbial metabolic pathway modules²⁸ ($n=91$, single dbRDA, $P=0.23$). These results are in accordance with the concept of a functionally redundant gut microbial ecosystem²⁸: while taxonomic profiles can vary substantially between individuals and even over time, taxa encode an overlapping core functional potential, ensuring stable interactions with the human host²⁹. Of note, ARG abundance profiles also did not differ significantly between families in the 4+ CGC dataset ($n=91$, single dbRDA, $P=0.27$). Next, we zoomed in on specific microbiome features rather than community-level variation, capitalizing on the availability of quantitative microbiome profiling (QMP)-based, metagenome-derived genus abundances. We found that seven genera occurred in higher abundances among members of specific families compared to the rest of the 4+ CGC cohort ($n=91$, two-sided Wilcoxon rank-sum test, $-\log_{10}(P) > 4.56$; Fig. 1g and Supplementary Table 5). While most of those family-associated genera could be qualified as low abundant (mean abundance $< 3 \times 10^6$ cells per gram of faeces, within 20% of the taxa with the lowest mean abundances in the dataset), 2 families were enriched in *Pseudomonas* (mean abundance = 3.74×10^6), an opportunistic pathogen³⁰, and *Oxalobacter* (mean abundance = 3.74×10^6), linked to kidney stone risk reduction³¹, respectively. As a complementary approach, we assessed whether prevalence (presence/absence) of species or functions appeared family-bound across 4+ CGC generations (non-random distribution in families across the cohort genealogy)³². 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 familiarity (GIF), $P_{\text{adj}} > 0.05$; 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 composition^{15,33}. 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 StrainPhlAn. This approach allowed us to characterize over 360 species across the CGC dataset (including samples from young children, $n=102$; Fig. 2 and Supplementary

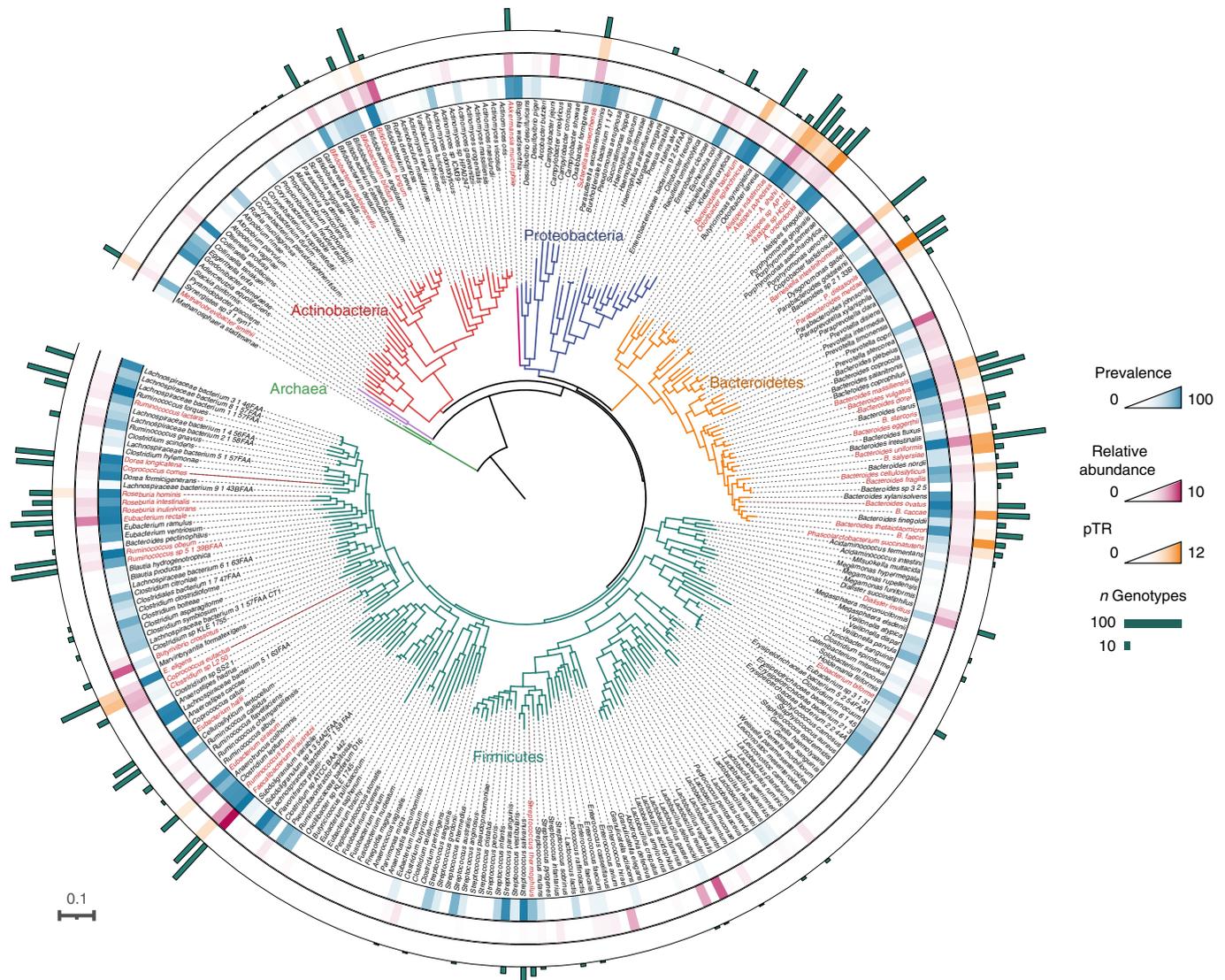


Fig. 2 | Maximum likelihood phylogenetic tree of the CGC species profiled in this study. All species detected within the CGC cohort and present in the PhyloPhlAn default database are shown; species profiled using StrainPhlAn are shown in red. Branches are coloured by phylum (Actinobacteria, red; Archaea, green; Bacteroidetes, yellow; Firmicutes, dark green; Proteobacteria, blue; Verrucomicrobia, pink), with the outer circles representing scaled prevalence (blue), relative abundance (percentage, pink) and pTRs (yellow). The number of genotypes recovered for each species is represented by bar plots (green).

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.²) 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, $R^2 = 0.304$, $P = 1 \times 10^{-3}$; 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, $P_{adj} < 0.05$; Supplementary Table 10). Of note, the overall distribution of IF distances showed a peak at $nGD = 0$ (that is, identical strains) whereas the BF one did not, suggesting a higher frequency of person-to-person transmissions and/or recent acquisition of microorganisms from a common source³⁴. 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 $n_{BF} = 63,287$,

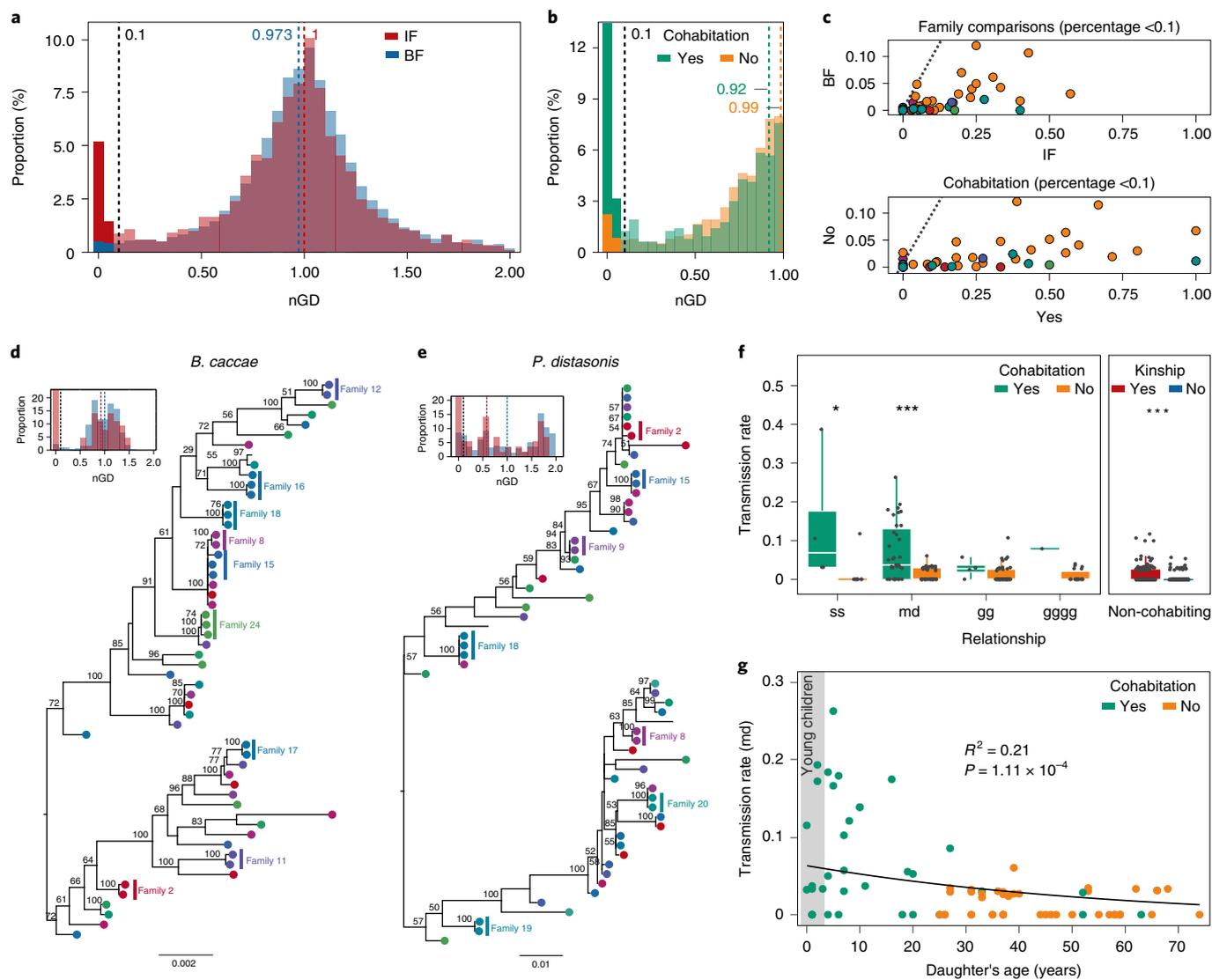


Fig. 3 | Transmissions of genotypes across family members. **a**, nGDs between all pairs of genotypes recovered. IF distances (red) present a lower overall median compared with BF distances (blue). **b**, nGDs for cohabiting individuals peak towards nGD = 0. **c**, Proportion of total pairwise comparisons 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. caccae* (**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, $n = 102$, $*P_{\text{adj}} < 0.05$, $***P_{\text{adj}} < 0.001$; Supplementary Table 15). Right: pTRs by kinship for non-cohabiting individuals (Wilcoxon rank-sum test, $n = 94$, $r = 0.24$, $P = 5.85 \times 10^{-5}$). 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 $1.5 \times$ the IQR, with outliers beyond. **g**, The pTR between mothers and daughters decreases with the daughter's age ($n = 78$ pairs, beta regression, $R^2 = 0.21$, $z = -3.87$, $P = 1.11 \times 10^{-4}$).

two-proportion test, chi-squared = 86.848, $P < 2.2 \times 10^{-16}$; 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%; $n_{\text{cohabitation}} = 633$ versus $n_{\text{LA}} = 1,817$, two-proportion test, chi-squared = 28.857, $P = 7.79 \times 10^{-8}$; 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^{33,35}. 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).

Bacteroidales species display the highest potential transmission rates. While sharing of strains can result from co-acquisition of bacterial species, increased frequencies of shared strains have been suggested to be indicative for transmission between individuals³⁶. In this study, based on the observed distribution of nGDs and opting for the most stringent among previously suggested similarity thresholds^{2,5,37}, we considered two genotypes to belong to the same strain when their nGDs < 0.10 (Fig. 3a and Extended Data Fig. 4). This cut-off allowed us to identify 1,958 strains in the CGC cohort, with 213 of them (belonging to 40 species) being involved in a total of 870 potential transmission events (Supplementary Table 9 and Supplementary Table 10). Such events were observed to occur significantly less frequently among unrelated individuals compared to family members (IF = 42.02% versus BF = 12.87%; $n_{IF} = 188$ versus $n_{BF} = 4,912$ pairs of individuals, chi-squared test = 125.87, $P < 2.2 \times 10^{-16}$). Potential intrafamilial transmission events were detected for 35 out of the 51 species genotyped, together representing more than 80% of the dominant genera in the gut microbiota (defined as the top 20% most abundant genera). To quantitatively explore these observations, we calculated potential transmission rates (pTRs) within as well as across species as the number of transmission/co-acquisition events detected divided by the maximum possible transmissions in a family (defined as the combinations of family members: maximum = $nCr(n = n$ members, $r = 2$)). Across species, average pTRs varied substantially between families (mean = 2.75%, range 0–9.85%; Supplementary Table 13), putatively reflecting differences in familial interaction patterns or habits such as hygiene practices³⁸. Within species, we observed the highest average pTRs for members of the order Bacteroidales (including *Parabacteroides distasonis* = 11.11% (0–50%), *Alistipes onderdonkii* = 9.58% (0–100%), *Bacteroides faecis* = 8.75% (0–50%), *Bacteroides caccae* = 8.61% (0–50%) and *Bacteroides salyersiae* = 8.33% (0–50%); Fig. 2 and Supplementary Table 13), in line with reports on their frequent transmission from mother to offspring⁵. To visualize IF strain sharing across the CGC dataset, we constructed a maximum likelihood phylogenetic tree based on the genotypes recovered within the cohort for each of the 35 species potentially transmitted between family members (Fig. 3d,e, Extended Data Fig. 5 and Supplementary Table 13). The highest numbers of potential IF transmission/co-acquisition events were detected for *B. caccae* and *P. distasonis*, shared between 15 and 14 pairs of individuals within 9 and 10 families, respectively (Fig. 3d,e).

Both kinship and cohabitation are associated with higher potential strain transmission. In the present dataset, kinship, cohabitation status and even age—all potential covariates of bacterial transmission frequency—emerged as closely related variables. For instance, 94% of participants under 30 years old reported living together with their mothers, with 96.3% of $n > 2$ households comprising a least 1 mother–<30-year-old daughter pair (Supplementary Table 1). While strain distribution was significantly associated with kinship (strain presence/absence microbiome profile variation, $n = 102$, Mantel test, $R^2 = 3.7\%$, $P = 6.5 \times 10^{-3}$), only cohabitation had a significant non-overlapping effect size (stepwise RDA, $R^2 = 8.1\%$, $P_{adj} = 1.3 \times 10^{-3}$; Supplementary Table 14). Within families, the highest pTRs were observed within sister ($n = 13$ pairs, mean = 5.17% (0–38.71%)) and mother–daughter pairs ($n = 78$, mean = 3.99% (0–26.32%)). While pTRs spanning multiple generations were markedly lower ($n = 49$, mean = 1.33% (0–10.71%) and $n = 19$, mean = 1.27% (0–7.9%) for pairs separated by 1 and 2 generations, respectively), only the differences between two (mother–daughter) and three generations (grandmother–granddaughter) were identified as significant within the limitations of our cohort ($n = 102$, Kruskal–Wallis test, chi-squared = 10.99, $P = 1.18 \times 10^{-2}$; post-hoc Dunn test, $P_{adj} = 1.36 \times 10^{-2}$; Supplementary Table 15). Both for sisters and mother–daughter pairs, the pTRs calculated

between pairs of individuals cohabiting were significantly higher than among their counterparts living apart (Wilcoxon rank-sum test, sisters, $r = 0.73$, $P_{adj} = 1.64 \times 10^{-2}$; mother–daughter pairs, $r = 0.47$, $P_{adj} = 1.56 \times 10^{-4}$; 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, $r = 0.24$, $P = 5.85 \times 10^{-5}$; 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.³⁹ 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, $P_{adj} < 0.01$; 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.³⁹ (re)analyses 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^{40,41}.

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, $R^2 = 0.21$, $z = -3.87$, $P = 1.11 \times 10^{-4}$; 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 Bacteroidetes, notably *B. caccae* (mean = 57.14%), *Bacteroides stercoris* (mean = 33.33%) 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 mothers⁴², 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^{2,19}. In this respect, given its low colonization resistance⁴³, 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 Universitair Ziekenhuis Brussel-Vrije Universiteit Brussel (BUN 143201215505) and the Commissie voor Medische Ethiek, Universitair Ziekenhuis/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 between 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.²⁶. Briefly, participants were asked to collect their faecal material (single defecation) 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 explorative study, CGC cohort size was similar to the number of participants included in previous publications^{2,39}. 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.³⁹ 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 ($5\text{ }\mu\text{m}$ 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.⁴⁴. 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 strategy⁴⁴; Extended Data Fig. 8). Cell counts were converted to microbial loads per gram of faecal material based on the exact weight of the aliquots analysed. 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 after lyophilization from approximately 0.2 g frozen aliquots of faecal material (-80°C). Faecal calprotectin concentrations were determined using the fCAL ELISA Kit (Bühlmann) on frozen faecal material (-80°C).

DNA extraction, sequencing and data preprocessing. Faecal DNA extraction and microbiota profiling was performed as described previously⁴⁵. 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/806R⁴⁶. 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 LotuS⁴⁷ v.1.565 to demultiplex the sequencing reads. Amplicon sequencing was used only for community typing to align with the FGFP 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 fastqc v.0.11.2 and Illumina adaptors and low-quality reads were removed using Trimmomatic⁴⁸ v.0.32 with the options ILLUMINACLIP:trimmomatic-0.32/adapters/NexteraPE-PE.fa:2:30:10:2, MAXINFO:40:0.70, HEADCROP:15 and MINLEN:40. High-quality reads were then decontaminated from phiX and human sequences using DeconSeq⁴⁹ v.0.4.3 and 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 DADA2⁵⁰ pipeline v.1.6.0 and the RDP classifier⁵¹ v.2.12 with default parameters. To obtain the 16S relative microbiome profiling (RMP) matrix, each sample was downsampled to 10,000 reads by random selection of reads. 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 MOCAT2⁵² pipeline (COG0012, COG0016, COG0018, COG0172, COG0215, COG0495, COG0525, COG0533, COG0541, COG0552). Paired-end reads were downsampled to even 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 = 4.98×10^{-9}). The resulting rarefied read counts were above 1.3×10^6 reads for all samples. Next, taxonomic classification of the rarefied reads into molecular operational taxonomic units (mOTUs) was performed with MOCAT2⁵² v.2.0.1 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 by 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 MetaPhlan2⁵³ and StrainPhlan2⁵⁶. Briefly, the preprocessed metagenomic reads were mapped against the MetaPhlan2 marker database using the metaphlan2 script with default parameters. Then, samples2markers.py was run to produce the gene marker file for each sample; gene marker files were parsed to StrainPhlan2 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)⁵⁴ using the Burrows–Wheeler Aligner⁵⁵ v.0.7.8 and the mapping was summarized into functional profiles by featureCounts⁵⁶ v.1.5.3, with the parameters --minOverlap 40 --pO. Gut metabolic module (GMM)²⁸ abundances were computed using Omixer-RPM v.1.0 (<https://github.com/raeslab/omixer-rpm>), with option -c 0.66 (66% coverage detection threshold). Coverage of the manually curated modules is calculated as 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 Database⁵⁷. GMM and ARG abundances were converted to quantitative abundance profiles (abundance per gram 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 StrainPhlan⁵⁶ 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 phages, viruses and viroids were discarded from further analysis. Gaps were removed from the alignments using T-Coffee⁵⁸ v.1.10.0 with option -action +rm_gap 1 so that only the covered core genome for that particular comparison was analysed; SNP-sites⁵⁹ 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-dists v.0.6 (<https://github.com/tseemann/snp-dists>). 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 taxa (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 others²⁵. To reconstruct the phylogenetic trees from the previously obtained core genome alignments, we used RAxML v.8.2.12⁶⁰ with the parameters *-f a* and *-m GTRGAMMA*. For the phylogenetic trees obtained with the *strainphlan.py* script, we set *bootstrap_raxml* to 100 and *marker_in_clade* to 0.2. Phylogenetic trees were rooted midpoint with the package ETE 3⁶¹. Finally, PhyloPhlAn v.3.0.60⁶² was used to produce a phylogenetic tree of all the species profiled using MetaPhlAn2 and the associated metadata were plotted using iTOL v.6⁶³. For the 51 species analysed at the strain level, we selected proximal representatives of taxa absent in the PhyloPhlAn 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 alignment⁶⁴, 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*⁶⁵ v.2.5.6, *phyloseq*⁶⁶ v.1.32.0, *FSA*⁶⁷ v.0.8.30, *coin*⁶⁸ v.1.3.1, *DirichletMultinomial*⁶⁹ v.1.30.0, *kinship2* (ref. 70) v.1.8.5, *FamAgg*⁷¹ v.1.16.0, *QuantPsc*⁷² v.1.5, *gmm*⁷³ v.1.6.5 and *ggplot2* (ref. 74) 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 ($\mu\text{g g}^{-1}$) to interindividual microbiota community variation was determined by single dbRDA on genus-level Bray–Curtis dissimilarity with the *capscale* function in the *vegan* R package⁷⁵. 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*⁶⁶. Enterotyping (or community typing) based on the DMM approach was performed in R using the *DirichletMultinomial*⁶⁹ package as described by Holmes et al.²⁵ on the RMP matrix. To increase accuracy, enterotyping was performed on a combined genus abundance matrix including the present dataset (*n* = 101) complemented with 1,106 samples from the FGFP²⁶ cohort rarefied to 10,000 reads. Microbiome interindividual variation was visualized by principal coordinate analysis (PCoA) using Bray–Curtis dissimilarity on the genus-level abundance matrix. The optimal number of Dirichlet components based on the Bayesian information criterion was four. The four FGFP clusters were named *Prevotella*, *Bacteroides* 1, *Bacteroides* 2 and *Ruminococcaceae* as described by Vandeputte et al.¹⁴. 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 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.

Microbiome transmission. Pedigrees were built using the *kinship2*⁷⁰ R package. The GIF was calculated using the *genealogicalIndexTest* function (*FamAgg*⁷¹ package) to assess family aggregation of specific microbiota traits across the kinship matrix; binomial tests (*binomialTest* function) were used to test for enrichment of specific traits in certain families.

Analyses of genetic distances. nGDs between pairs of genotypes recovered from individuals were annotated by familial relationship and current cohabitation status as follows: IF, BF, cohabitation and LA. Comparisons between the nGD distributions between groups (family and cohabitation status) were performed using the generalized method and two-proportions test in the *gmm* R package. Wilcoxon rank-sum tests were used to test the median differences of the nGDs between two groups (BF versus IF). We also used PERMANOVA (*adonis.test* function in *vegan*) to test for differences between BF and IF. A similar approach

was applied to evaluate the effects of kinship and cohabitation on the pairwise genetic distances between ARG sequences.

Identification of strains and calculation of pTRs. The nGDs between pairs of genotypes were used to define strains by grouping the pairs of genotypes from the same species with nGDs below 0.10. Transmission events between pairs of individuals were computed by adding up all species comparisons with nGD < 0.10 between any pair of individuals. pTRs between pairs of individuals were calculated by dividing the number of transmissions identified by the maximum possible transmissions in the pair (*n* shared strains per *n* species detected in any of the two individuals). The IF pTRs for each species were obtained by dividing the number of transmission events identified in a family for a certain species by the maximum possible transmissions in a family (maximum = *nCr* (*n* = *n* family members, *r* = 2)).

Analyses of pTRs. Statistical differences in the counts of transmissions among groups were evaluated using pairwise chi-squared tests. Spearman correlation analyses were performed 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 (betareg function in the *betareg* R package⁷⁶, v.3.1-4). The pTRs, with range = 0–1, were transformed to obtain rates in the range = 0–1 as $pTR = (pTR(n-1) + s)/n$, with *s* = 0.5 as recommended for beta regression⁷⁷. Nested model comparison was performed using a likelihood ratio test (*lrtest* in the *lmtest* R package⁷⁸, 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

- 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).
- 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).
- 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).
- Bäckhed, F. et al. Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe* **17**, 690–703 (2015).
- Shao, Y. et al. Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature* **574**, 117–121 (2019).
- Brito, I. L. et al. Transmission of human-associated microbiota along family and social networks. *Nat. Microbiol.* **4**, 964–971 (2019).
- Sonnenburg, E. D. et al. Diet-induced extinctions in the gut microbiota compound over generations. *Nature* **529**, 212–215 (2016).
- 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).
- Moeller, A. H., Suzuki, T. A., Phifer-Rixey, M. & Nachman, M. W. Transmission modes of the mammalian gut microbiota. *Science* **362**, 453–457 (2018).
- Zeller, M. & Daniels, S. The obesity epidemic: family matters. *J. Pediatr.* **145**, 3–4 (2004).
- Santos, M. P. C., Gomes, C. & Torres, J. Familial and ethnic risk in inflammatory bowel disease. *Ann. Gastroenterol.* **31**, 14–23 (2018).
- 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).

13. Di Angelantonio, E. et al. Body-mass index and all-cause mortality: individual-participant-data meta-analysis of 239 prospective studies in four continents. *Lancet* **388**, 776–786 (2016).
14. Vandeputte, D. et al. Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551**, 507–511 (2017).
15. Falony, G. et al. The human microbiome in health and disease: hype or hope. *Acta Clin. Belg.* **74**, 53–64 (2019).
16. Lewis, S. J. & Heaton, K. W. Stool form scale as a useful guide to intestinal transit time. *Scand. J. Gastroenterol.* **32**, 920–924 (1997).
17. Falony, G., Vieira-Silva, S. & Raes, J. Richness and ecosystem development across faecal snapshots of the gut microbiota. *Nat. Microbiol.* **3**, 526–528 (2018).
18. Vandeputte, D. et al. Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. *Gut* **65**, 57–62 (2016).
19. Beller, L. et al. Successional stages in infant gut microbiota maturation. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.06.25.450009> (2021).
20. Stewart, C. J. et al. Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* **562**, 583–588 (2018).
21. Vieira-Silva, S. et al. Statin therapy is associated with lower prevalence of gut microbiota dysbiosis. *Nature* **581**, 310–315 (2020).
22. Vieira-Silva, S. et al. Quantitative microbiome profiling disentangles inflammation- and bile duct obstruction-associated microbiota alterations across PSC/IBD diagnoses. *Nat. Microbiol.* **4**, 1826–1831 (2019).
23. Reynders, T. et al. Gut microbiome variation is associated to multiple sclerosis phenotypic subtypes. *Ann. Clin. Transl. Neurol.* **7**, 406–419 (2020).
24. Valles-Colomer, M. et al. The neuroactive potential of the human gut microbiota in quality of life and depression. *Nat. Microbiol.* **4**, 623–632 (2019).
25. Holmes, I., Harris, K. & Quince, C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS ONE* **7**, e30126 (2012).
26. Falony, G. et al. Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).
27. Tsirpanlis, G., Alevyzaki, F., Triantafyllis, G., Chatzipanagiotou, S. & Nicolaou, C. C-reactive protein: “cutoff” point and clinical applicability. *Am. J. Kidney Dis.* **46**, 368 (2005).
28. Vieira-Silva, S. et al. Species–function relationships shape ecological properties of the human gut microbiome. *Nat. Microbiol.* **1**, 16088 (2016).
29. Huttenhower, C. et al. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
30. Gellatly, S. L. & Hancock, R. E. W. *Pseudomonas aeruginosa*: new insights into pathogenesis and host defenses. *Pathog. Dis.* **67**, 159–173 (2013).
31. Kaufman, D. W. et al. *Oxalobacter formigenes* may reduce the risk of calcium oxalate kidney stones. *J. Am. Soc. Nephrol.* **19**, 1197–1203 (2008).
32. Khoury, M. J., Beaty, T. H. & Cohen, B. H. *Fundamentals of Genetic Epidemiology* (Oxford Univ. Press, 1993).
33. Song, S. J. et al. Cohabiting family members share microbiota with one another and with their dogs. *eLife* **2**, e00458 (2013).
34. Worby, C. J., Chang, H.-H., Hanage, W. P. & Lipsitch, M. The distribution of pairwise genetic distances: a tool for investigating disease transmission. *Genetics* **198**, 1395–1404 (2014).
35. Browne, H. P., Neville, B. A., Forster, S. C. & Lawley, T. D. Transmission of the gut microbiota: spreading of health. *Nat. Rev. Microbiol.* **15**, 531–543 (2017).
36. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27**, 626–638 (2017).
37. Korpela, K. et al. Selective maternal seeding and environment shape the human gut microbiome. *Genome Res.* **28**, 561–568 (2018).
38. Vandegriff, R. et al. Cleanliness in context: reconciling hygiene with a modern microbial perspective. *Microbiome* **5**, 76 (2017).
39. Costea, P. I. et al. Subspecies in the global human gut microbiome. *Mol. Syst. Biol.* **13**, 960 (2017).
40. Xie, H. et al. Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. *Cell Syst.* **3**, 572–584.e3 (2016).
41. Rothschild, D. et al. Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).
42. Sacri, A. S. et al. Transmission of acute gastroenteritis and respiratory illness from children to parents. *Pediatr. Infect. Dis. J.* **33**, 583–588 (2014).
43. Mughini-Gras, L. et al. Societal burden and correlates of acute gastroenteritis in families with preschool children. *Sci. Rep.* **6**, 22144 (2016).
44. Prest, E. L., Hammes, F., Köttsch, S., van Loosdrecht, M. C. M. & Vrouwenvelder, J. S. Monitoring microbiological changes in drinking water systems using a fast and reproducible flow cytometric method. *Water Res.* **47**, 7131–7142 (2013).
45. Tito, R. Y. et al. Dialister as a microbial marker of disease activity in spondyloarthritis. *Arthritis Rheumatol.* **69**, 114–121 (2017).
46. Walters, W. A. et al. PrimerProspector: de novo design and taxonomic analysis of barcoded polymerase chain reaction primers. *Bioinformatics* **27**, 1159–1161 (2011).
47. Hildebrand, F., Tadeo, R., Voigt, A. Y., Bork, P. & Raes, J. LotuS: an efficient and user-friendly OTU processing pipeline. *Microbiome* **2**, 30 (2014).
48. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
49. Schmieder, R. & Edwards, R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE* **6**, e17288 (2011).
50. Callahan, B. J. et al. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
51. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007).
52. Kultima, J. R. et al. MOCAT2: a metagenomic assembly, annotation and profiling framework. *Bioinformatics* **32**, 2520–2523 (2016).
53. Truong, D. T. et al. MetaPhlan2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
54. Li, J. et al. An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).
55. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
56. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
57. Jia, B. et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **45**, D566–D573 (2017).
58. Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217 (2000).
59. Page, A. J. et al. *SNP-sites*: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genom.* **2**, e000056 (2016).
60. Stamatakis, A. RAXML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
61. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
62. Asnicar, F. et al. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nat. Commun.* **11**, 2500 (2020).
63. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
64. Qin, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
65. Oksanen, J. et al. *vegan*: Community Ecology Package. R package version 2.2-1 <https://cran.r-project.org/web/packages/vegan/index.html> (2015).
66. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8**, e61217 (2013).
67. Ogle, D. H. *FSA*: Fisheries stock analysis. R package version 0.8.13 <https://cran.r-project.org/web/packages/FSA/index.html> (2017).
68. Hothorn, T., Hornik, K., van de Wiel, M. A. & Zeileis, A. A Lego system for conditional inference. *Am. Stat.* **60**, 257–263 (2006).
69. Morgan, M. DirichletMultinomial: Dirichlet-multinomial mixture model machine learning for microbiome data. R package version 1.18.0 <https://bioconductor.org/packages/release/bioc/html/DirichletMultinomial.html> (2017).
70. Sinnwell, J. P., Therneau, T. M. & Schaid, D. J. The kinship2 R package for pedigree data. *Hum. Hered.* **78**, 91–93 (2014).
71. Rainer, J. et al. FamAgg: an R package to evaluate familial aggregation of traits in large pedigrees. *Bioinformatics* **32**, 1583–1585 (2016).
72. Fletcher, T. D. QuantPsyc: Quantitative Psychology Tools. R package version 1.5 <https://cran.r-project.org/web/packages/QuantPsyc/index.html> (2012).
73. Chaussé, P. Computing generalized method of moments and generalized empirical likelihood with R. *J. Stat. Softw.* **34**, <https://doi.org/10.18637/jss.v034.i11> (2010).
74. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, 2009).
75. Oksanen, J. et al. *vegan*: Community Ecology Package. R package version 2.4-2 <https://cran.r-project.org/web/packages/vegan/index.html> (2017).
76. Cribari-Neto, F. & Zeileis, A. Beta regression in R. *J. Stat. Softw.* **34**, <https://doi.org/10.18637/jss.v034.i02> (2010).
77. Smithson, M. & Verkuilen, J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychol. Methods* **11**, 54–71 (2006).

78. Zeileis, A. & Hothorn, T. Diagnostic checking in regression relationships. *R News* 2, 7–10 (2002).

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. 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.²³. 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.

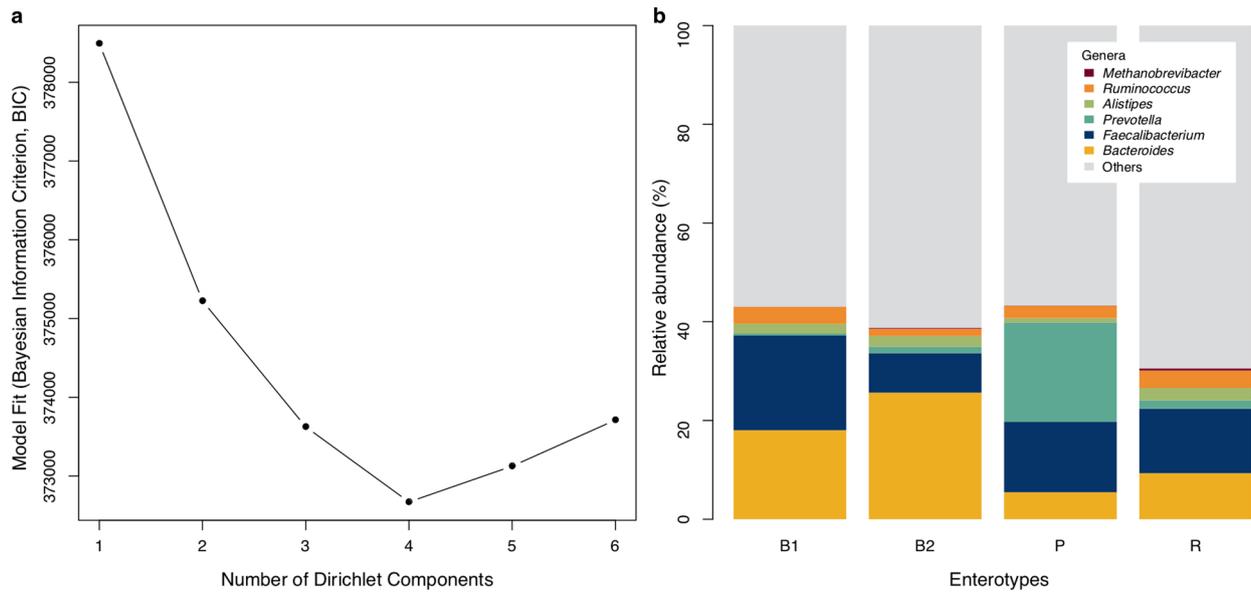
Reprints and permissions information is available at www.nature.com/reprints.

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

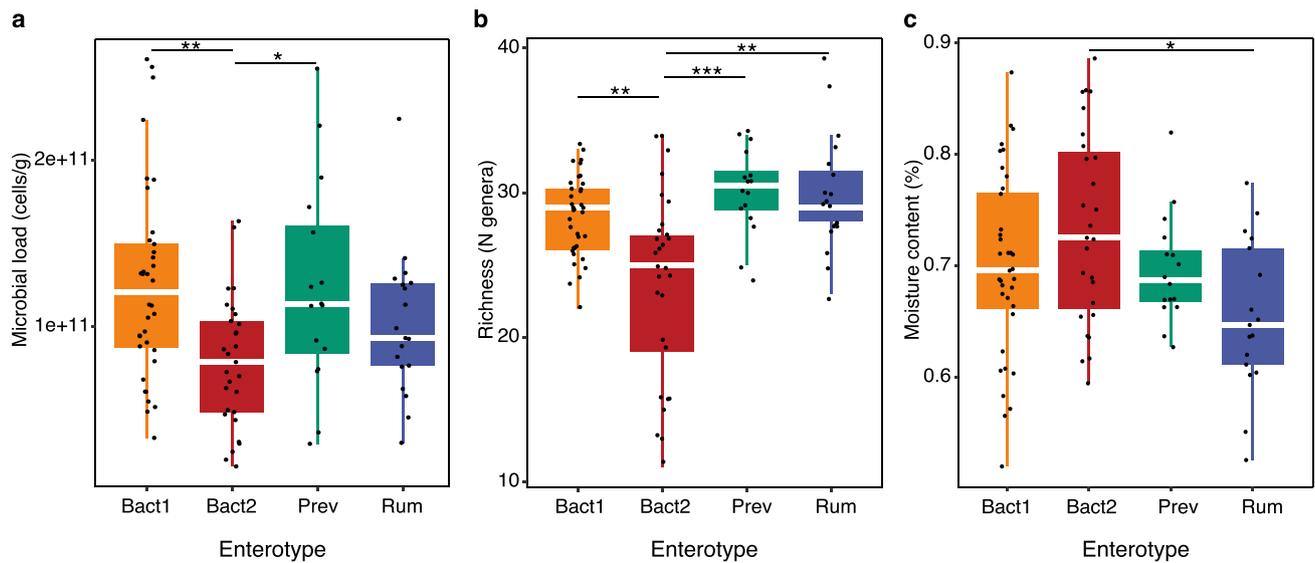


Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative 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/>.

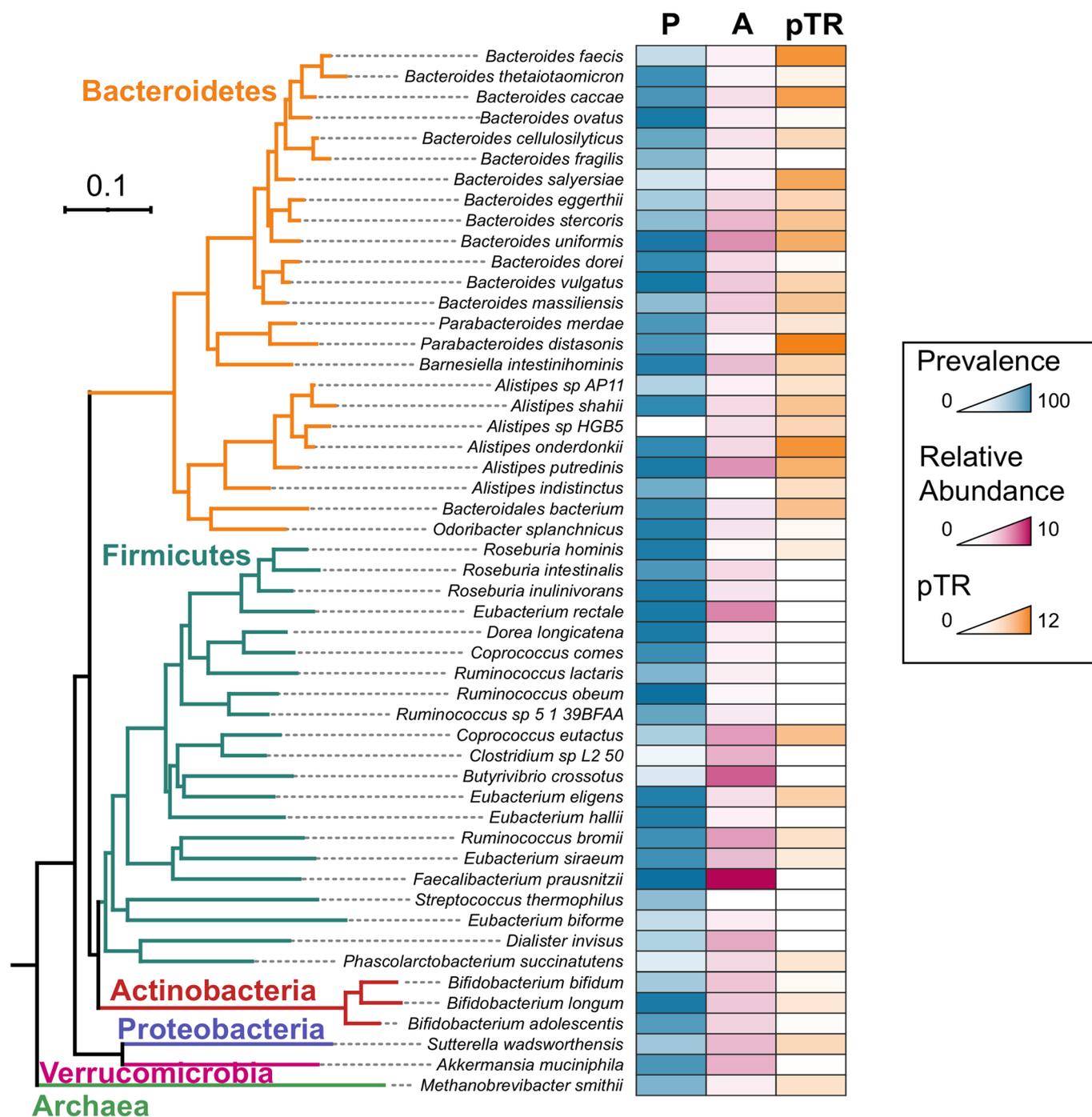
© The Author(s) 2021



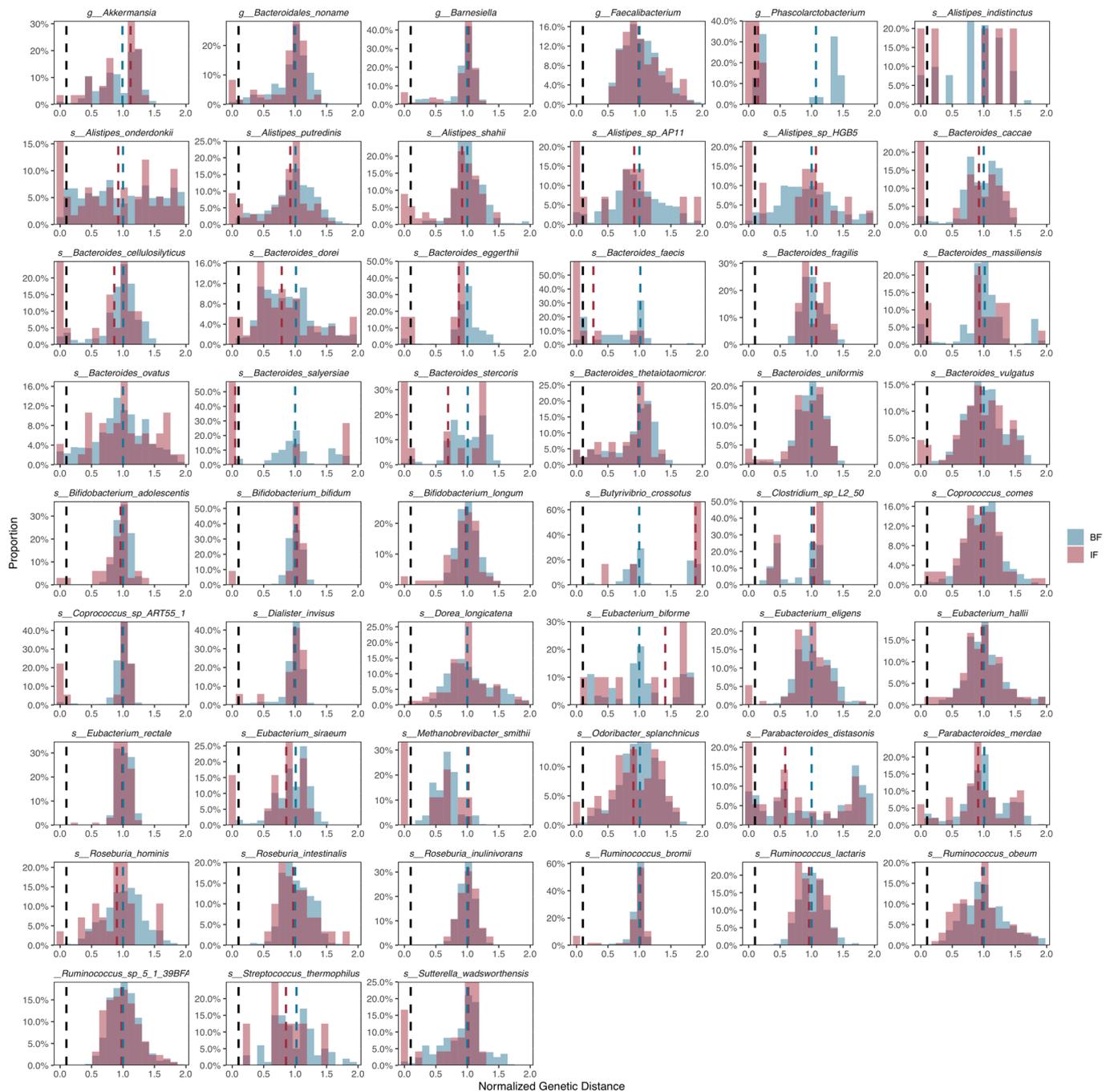
Extended Data Fig. 1 | Enterotype stratification by DMM community typing. (a) Identification of the optimal number of clusters (Dirichlet components) in the CGC dataset ($N = 101$) complemented with 1106 samples from the FGFP²⁶ cohort based on the Bayesian Information Criterion (BIC). (b) Barplot representation 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 ($N = 1207$).



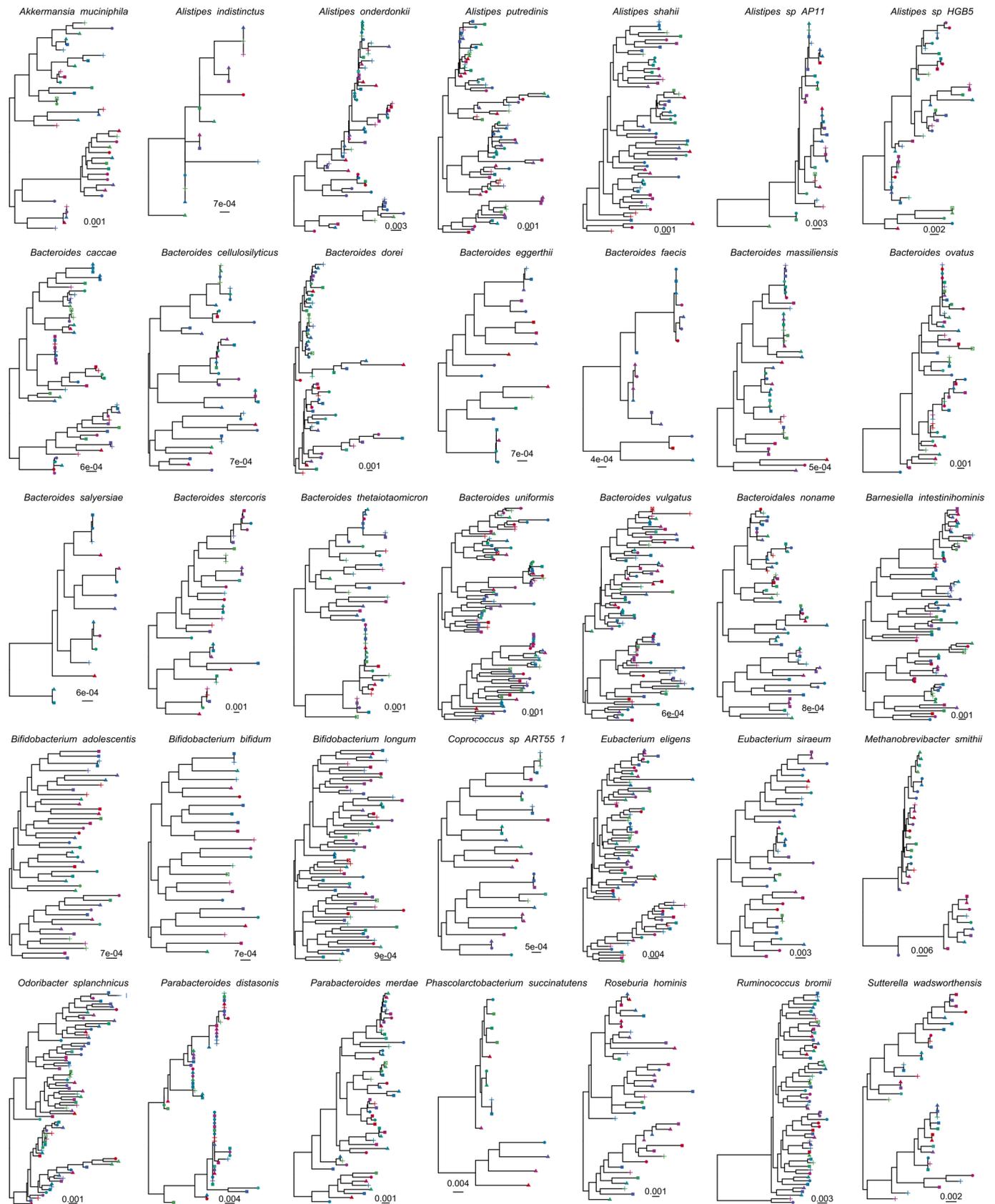
Extended Data Fig. 2 | Enterotype associations with microbial load, richness, and moisture content. (a) Reduced microbial load (microbial cells per gram of stool) in Bact2 enterotyped samples; N=101, KW test $\text{Chi}^2=13.9$, $P=3.0\text{e-}03$; pHD tests, $\text{adjP}<0.001$ (***), <0.01 (**), <0.05 (*); Supplementary Table 3. (b) Reduced richness (number of genera) in Bact2 enterotyped samples; N=101, KW test $\text{Chi}^2=20.0$, $P=1.6\text{e-}04$; pHD tests, $\text{adjP}<0.001$ (***), <0.01 (**), <0.05 (*); Supplementary Table 3. (c) Increased stool moisture content (%) in Bact2 enterotyped samples; N=101, KW test $\text{Chi}^2=8.8$, $P=0.03$; pHD tests, $\text{adjP}<0.001$ (***), <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 1.5x the interquartile range, with outliers beyond.



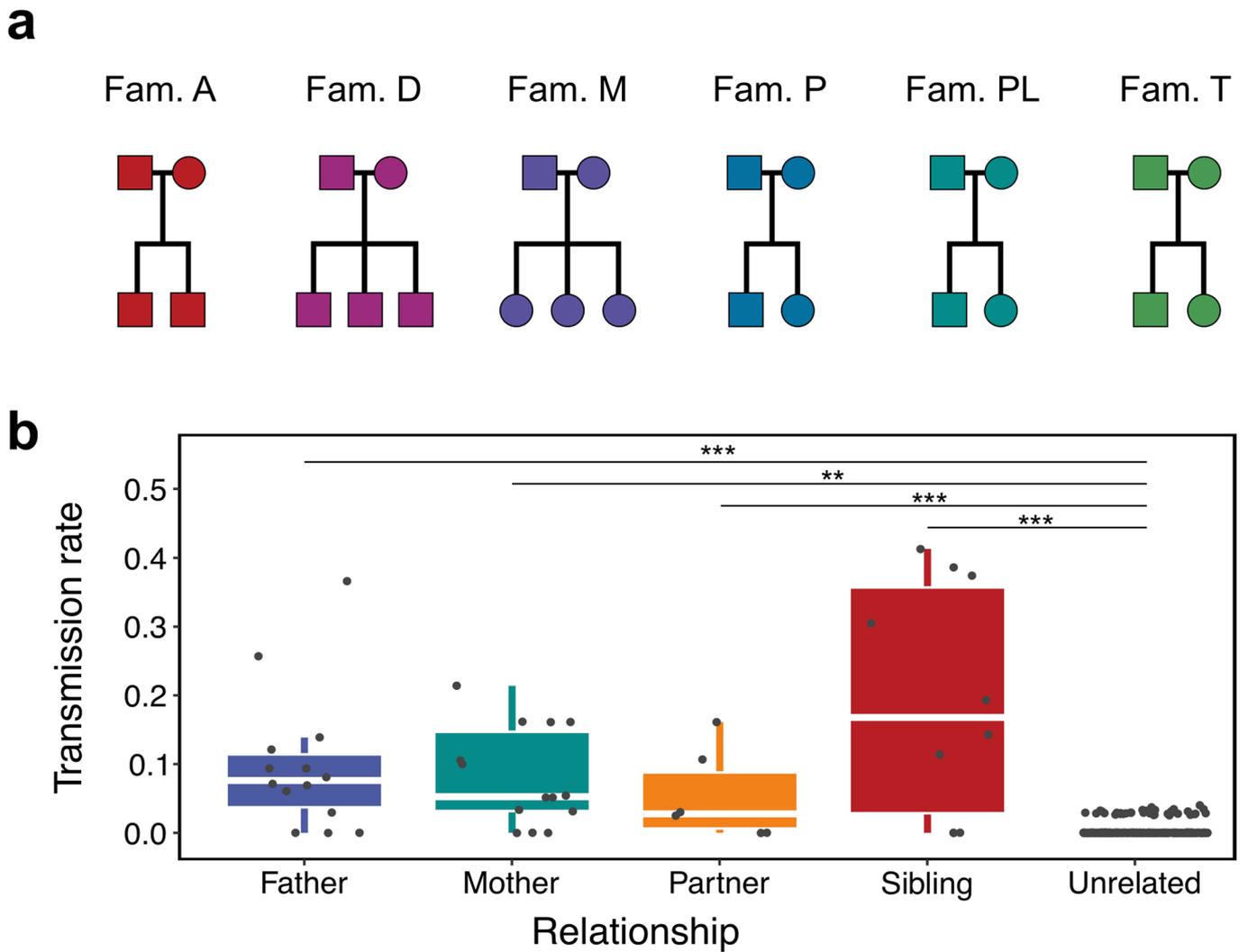
Extended Data Fig. 3 | Maximum-likelihood phylogenetic tree of the species present within the CGC cohort analysed using StrainPhlAn. Branches are coloured by phyla (Actinobacteria, red; Archaea, green; Bacteroidetes, yellow; Firmicutes, dark green; Proteobacteria, blue; Verrucomicrobia, pink). Boxes represent % prevalence (P, blue), % relative abundance (A, pink), and potential transmission rates (pTR, yellow).



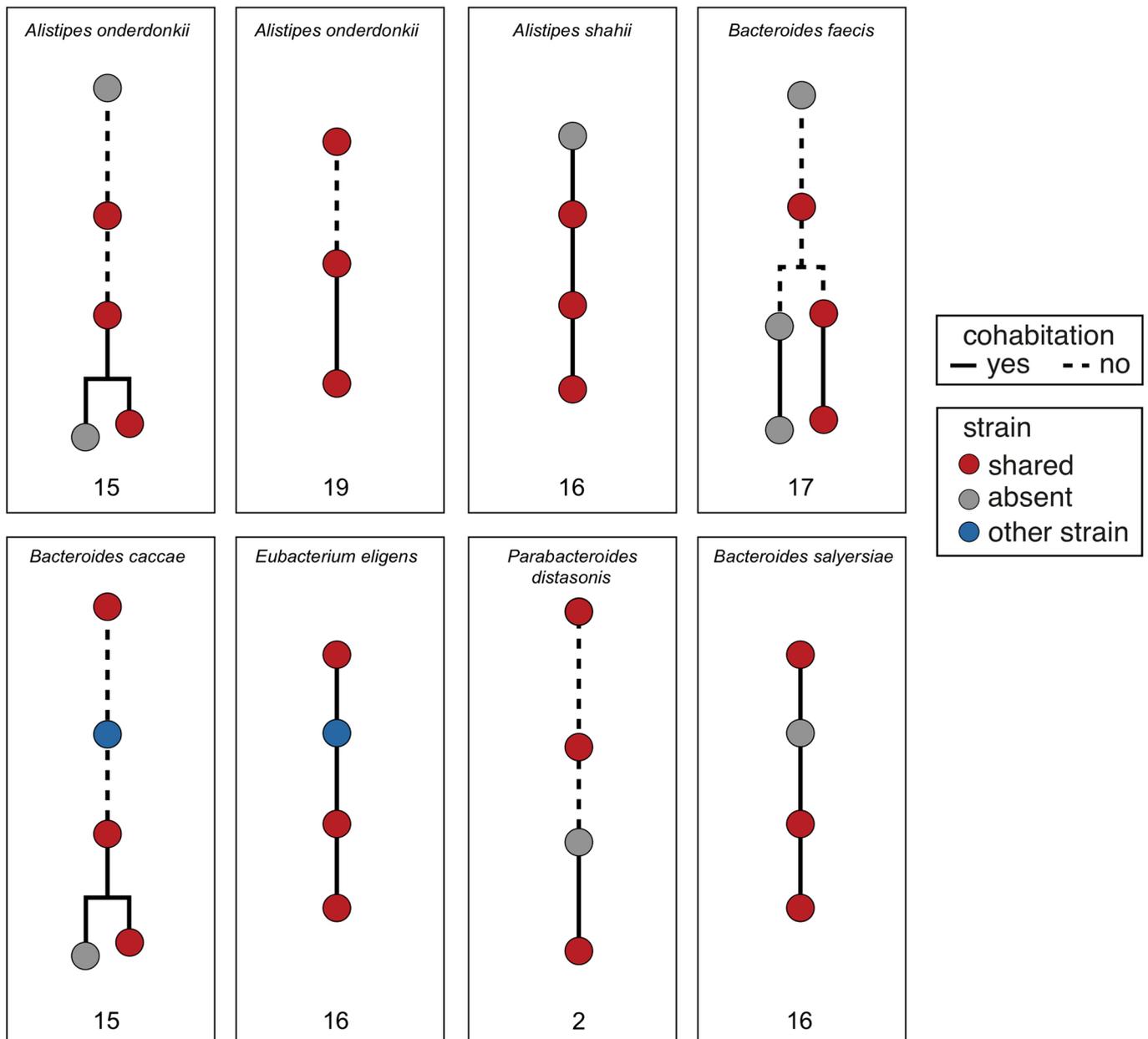
Extended Data Fig. 4 | Distribution of normalized genetic distances for each of the 51 species analysed. Distances between pairs of genotypes recovered from individuals of the same families are coloured in red, and those from individuals of different families are coloured in blue. Dashed vertical lines 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).



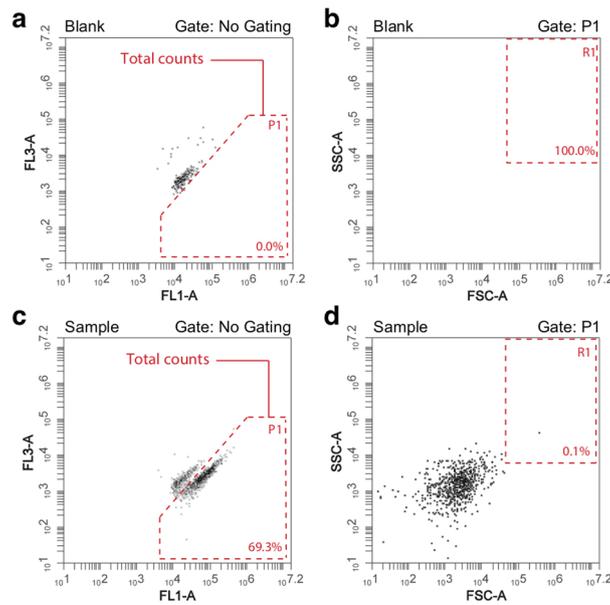
Extended Data Fig. 5 | Phylogenetic trees of species for which strain sharing was detected between at least two members of the same family. Tips are colored by family IDs as in Fig. 1 and shapes indicate generation number (0=Circle, 1=Triangle, 2=Square, 3=Cross).



Extended Data Fig. 6 | Summary results for cohabitating, non-related individuals. (a) Family structures in the Costea *et al.* study (N=26)³⁹. **(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.

n/a Confirmed

- 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.

Data analysis

16S data pre-processing was performed using LotuS (version 1.565, 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), QuantPsys (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.

Materials & experimental systems

Methods

- n/a Involved in the study
- Antibodies
- Eukaryotic cell lines
- Palaeontology and archaeology
- Animals and other organisms
- Human research participants
- Clinical data
- Dual use research of concern

- n/a Involved in the study
- ChIP-seq
- Flow cytometry
- MRI-based neuroimaging

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 participate 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 was 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.