Article

Environmental factors shaping the gut microbiome in a Dutch population

https://doi.org/10.1038/s41586-022-04567-7

Received: 23 October 2020

Accepted: 18 February 2022

Published online: 13 April 2022

Check for updates

R. Gacesa[1,2,10], A. Kurilshikov[2,10], A. Vich Vila[1,2], T. Sinha[2], M. A. Y. Klaassen[1,2], L. A. Bolte[1,2], S. Andreu-Sánchez[2,3], L. Chen[2,3], V. Collij[1,2], S. Hu[1,2], J. A. M. Dekens[2,4], V. C. Lenters[5], J. R. Björk[1,2], J. C. Swarte[1,2], M. A. Swertz[2,6], B. H. Jansen[1,2], J. Gelderloos-Arends[2], S. Jankipersadsing[2], M. Hofker[3,12], R. C. H. Vermeulen[5,7], S. Sanna[2,8], H. J. M. Harmsen[9,11], C. Wijmenga[2,11], J. Fu[2,3,11]✉, A. Zhernakova[2,11]✉ & R. K. Weersma[1,11]✉

The gut microbiome is associated with diverse diseases[1–3], but a universal signature of a healthy or unhealthy microbiome has not been identified, and there is a need to understand how genetics, exposome, lifestyle and diet shape the microbiome in health and disease. Here we profiled bacterial composition, function, antibiotic resistance and virulence factors in the gut microbiomes of 8,208 Dutch individuals from a three-generational cohort comprising 2,756 families. We correlated these to 241 host and environmental factors, including physical and mental health, use of medication, diet, socioeconomic factors and childhood and current exposome. We identify that the microbiome is shaped primarily by the environment and cohabitation. Only around 6.6% of taxa are heritable, whereas the variance of around 48.6% of taxa is significantly explained by cohabitation. By identifying 2,856 associations between the microbiome and health, we find that seemingly unrelated diseases share a common microbiome signature that is independent of comorbidities. Furthermore, we identify 7,519 associations between microbiome features and diet, socioeconomics and early life and current exposome, with numerous early-life and current factors being significantly associated with microbiome function and composition. Overall, this study provides a comprehensive overview of gut microbiome and the underlying impact of heritability and exposures that will facilitate future development of microbiome-targeted therapies.

Alterations in gut microbiota composition and function are associated with a broad range of human health disorders, including gastrointestinal and metabolic diseases and mental disorders[1,2]. The influence of gut bacteria and microbial pathways on host metabolism and immunity, together with the fact that the microbiota can be modified, have heightened interest in developing microbiome-targeted therapies[3] and multiple microbiome-targeting therapies are currently in clinical trials[4]. However, the characteristics of a healthy microbiome remain largely unclear, as does the extent to which the gut microbiome is driven by intrinsic factors (such as genetics) versus modifiable environmental factors (such as pollutants, diet or lifestyle) or health-related factors (such as gut disorders or body mass index (BMI)) that might be amenable to microbiome-targeted therapies.

The capacity to define a healthy microbiome has been hampered by differences in the processing of faecal samples between studies and the large interindividual variation in microbiome composition. Population-based studies have shown that the interindividual variation is partially accounted for by diet, lifestyle, host genetics and environmental factors, including early-life exposures[5–7]. However, in-depth integrative analyses of large, standardized, cohorts with well-defined phenotypes remain scarce, even though this integrative perspective is essential for disentangling meaningful host–microbiota associations and identifying potential targets for microbiota-directed interventions.

The Dutch Microbiome Project

To address these issues, we initiated the Dutch Microbiome Project (DMP) within Lifelines, a three-generational population cohort and biobank from the northern Netherlands with well-defined phenotypes. In the DMP, we characterized the composition and function of the gut microbiota of 8,208 individuals (age range 8–84 years, 57.4% female, 99.5% Dutch European ancestry; 4,745 of the individuals were

[1]University of Groningen and University Medical Center Groningen, Department of Gastroenterology and Hepatology, Groningen, The Netherlands. [2]University of Groningen and University Medical Center Groningen, Department of Genetics, Groningen, The Netherlands. [3]Department of Pediatrics, University of Groningen and University Medical Center Groningen, Groningen, The Netherlands. [4]Center of Development and Innovation, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands. [5]University Medical Centre Utrecht, Julius Center for Health Sciences and Primary Care, Utrecht, The Netherlands. [6]University of Groningen and University Medical Center Groningen, Genomics Coordination Center, Groningen, The Netherlands. [7]Utrecht University, Institute for Risk Assessment Sciences (IRAS), Department of Population Health Sciences, Utrecht, The Netherlands. [8]Institute for Genetic and Biomedical Research (IRGB), National Research Council (CNR), Cagliari, Italy. [9]Department of Medical Microbiology and Infection prevention, University of Groningen and University Medical Center Groningen, Groningen, The Netherlands. [10]These authors contributed equally: R. Gacesa, A. Kurilshikov. [11]These authors jointly supervised this work: H. J. M. Harmsen, C. Wijmenga, J. Fu, A. Zhernakova & R. K. Weersma. [12]Deceased: M. Hofker. ✉e-mail: j.fu@umcg.nl; sasha.zhernakova@gmail.com; r.k.weersma@umcg.nl

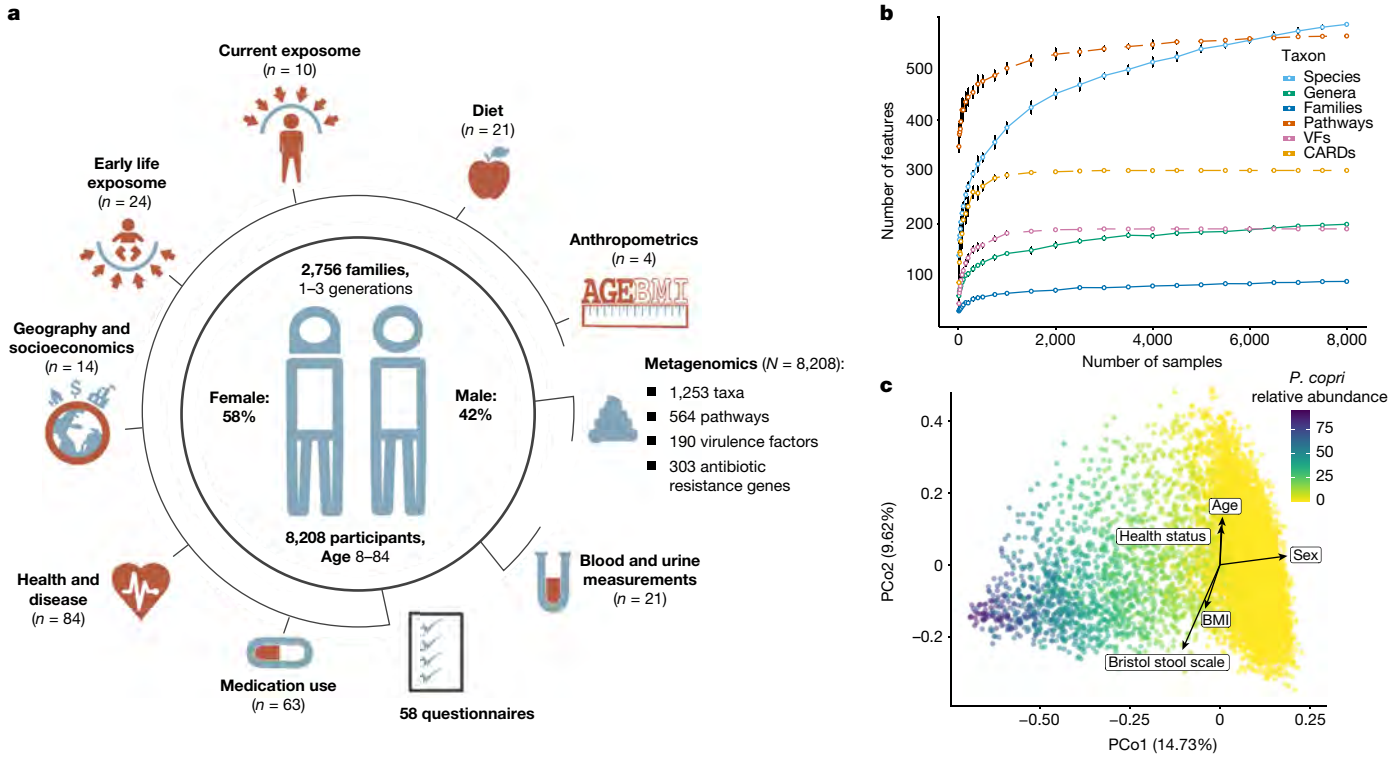


Fig. 1 | Summary of the Dutch Microbiome Project. a, Graphical summary of the cohort and overview of available metadata (*n* = number of variables collected, *N* = sample size). **b**, Number of microbial features discovered in relation to sample size. Dots denote mean values. Error bars display SD of 100 resamplings. **c**, PCoA visualizing the beta-diversity of the cohort. Colour indicates relative abundance of *P. copri*. Arrows indicate influence of self-reported health, anthropometrics and faecal sample metadata. CARDs, antibiotic resistance gene families from the Comprehensive Antibiotic Resistance Database; PCo, principal coordinate; VFs, bacterial virulence factors.

clustered into 2,756 families; Fig. 1a, Supplementary Table 2d). The dataset (processed by Metaphlan2 and HUMAnN2) contained 1,253 taxa (4 kingdoms, 21 phyla, 35 classes, 62 orders, 128 families, 270 genera and 733 species) and 564 metabolic pathways, including 257 archaeal and bacterial taxa and 277 pathways with a relative abundance higher than 0.01 and present in more than 5% of individuals.

Our sample size enabled us to cover more than 90% of the total expected number of microbial functional features estimated by bootstrap analysis: 564 out of 574.4 (standard error = 4.2) expected MetaCyc pathways, 190 out of 190 (standard error = 0.05) virulence factors, 303 out of 303 (standard error = 0.03) antibiotic resistance genes and 128 out of 136.5 (standard error = 2.7) microbial taxa at family level or higher (Fig. 1b). By subsampling the cohort, we estimated that the presence rates of these microbial features become relatively stable (within 90% of the numbers observed for the whole cohort) when at least 40% of the cohort is sampled (approximately 3,300 samples). However, the number of microbial species discovered continued to increase with increased sample size, with the total number of species in the population estimated to be 600 (standard error = 27) at 25,000 samples, suggesting that other rare microbial species remain undiscovered (Fig. 1b, Extended Data Fig. 1). Gut microbiota composition was highly variable across the population, for example the relative abundance of Bacteroidetes ranged from 5% to more than 95% (Extended Data Fig. 2c). The most abundant microbial pathways were significantly less variable than the majority of phyla (one-sided *F*-test of variances false discovery rate (FDR) < 0.05, Extended Data Fig. 2d, Supplementary Table 2g).

Core and keynote species

To pinpoint microbial species and pathways potentially critical for the organization and maintenance of the gut ecosystem, we investigated our cohort for microbial taxa present in more than 95% of individuals (we designated these 'core microbes') and taxa that form central nodes in microbial co-abundance networks[8] (designated 'keystone features'). We identified nine core species (*Subdoligranulum sp.*, *Alistipes onderdonkii*, *Alistipes putredinis*, *Alistipes shahii*, *Bacteroides uniformis*, *Bacteroides vulgatus*, *Eubacterium rectale*, *Faecalibacterium prausnitzii* and *Oscillibacter sp.*) that are highly consistent with those found in UK, US, European and non-western populations (Supplementary Table 1a). We also identified 28 species and 53 pathways as potential keystone features defined by more than 109 and 337 significant co-abundances, respectively (empirical FDR < 0.05). The networks defined by these features showed 20.2% overlap for species and 25.3% overlap for pathways.

Five of the nine identified core microbes (*A. putredinis*, *A. shahii*, *F. prausnitzii*, *Oscillibacter sp.* and *Subdoligranulum sp.*) are also keystone species, implying that they also have central roles in the gut microbiome ecosystem in the Dutch population. For example, *F. prausnitzii*, a major butyrate producer that is depleted in many chronic diseases[9], shows significant co-abundance with the majority of *Bacteroidetes* and *Bifidobacterium* species (Supplementary Table 1e). However, we also identified potential keystone species with low prevalence (≤10%), including *Ruminococcus gnavus* and multiple species from the genus *Clostridium*, that are positively associated with multiple diseases (Supplementary Table 3b), consistent with previous studies[10,11].

Prevotella copri defines clustering

We examined the microbiome data clustering in our cohort using principal coordinate analysis (PCoA) and identified that the first principal coordinate is driven by *Prevotella copri* (Spearman $r = 0.68$, P value = 3.6×10^{-180}; Fig. 1c, Supplementary Table 9). This bacterium is bimodally distributed in our cohort and defines two clusters on the

basis of its presence or absence (Extended Data Fig. 3a, b). As suggested by previous studies[12], we observed that the cluster with high abundance of *P. copri* associates with a lower risk of irritable bowel syndrome (IBS) (odds ratio = 0.72, 95% confidence interval 0.61–0.86). We also found that *P. copri* positively associates with general health (odds ratio = 1.24, 95% confidence interval 1.11–1.40, FDR < 0.05; Extended Data Fig. 3c). Although previous studies reported distinct enterotypes dominated by *Bacteroides*, *Prevotella* and *Ruminococcaceae*[13], we only observed two such clusters, possibly because our cohort is ethnically uniform and from a constrained geographic area.

The PCoA of functional potential was highly correlated with many pathways rather than any individual one, and the top features explaining variance were queuosine biosynthesis, peptidoglycan biosynthesis and L-isoleucine biosynthesis pathways (Supplementary Table 9). Similar to pathways, the PCoA of virulence factors was correlated with multiple gene families that encode various bacterial functions, including a flagellin involved in bacterial invasion of intestinal cells (gene family VF0114, Spearman $r = -0.60$, P value ≈ 0.0), bacterial siderophores (VF0136, VF0228 and VF0256; Spearman $r = 0.55$, 0.58 and 0.51, respectively; P values < 1.0×10^{-100}), a secretion system (VF0333; Spearman $r = 0.76$, P value ≈ 0.0) and bacterial adherence factors (VF0221 and VF0404; Spearman $r = 0.581$ and 0.561, respectively; P values < 1.0×10^{-100}). By contrast, the PCoA of antibiotic resistance genes was dominated by just three gene families that confer resistance to tetracycline antibiotics: gene families encoding ribosomal protection proteins ARO_30001914 and ARO_3000191 (Spearman $r = 0.81$ and −0.68, respectively; P values ≈ 0.0) and the gene family encoding efflux pump, ARO_3000567 (Spearman $r = 0.50$, P value ≈ 0.0).

Cohousing dominates over heritability

We next explored the relative contributions of family structure, cohabitation and other exposome factors in shaping the gut microbiome. We used the multi-generational family structure of our cohort to estimate the heritability of microbial taxa and identified 17 heritable taxa (6.6% of the tested taxa) at FDR-corrected empirical P value < 0.1 (determined using 30,000 permutations) (Fig. 2a, Supplementary Table 5). The highest heritability was observed for Proteobacteria ($h^2 = 0.308$, where h = narrow sense heritability (the proportion of variance in the abundance of microbial taxon caused by additive genetic effect)), followed by *Akkermansia muciniphila* ($h^2 = 0.302$) with its higher-level taxa Bacteroidaceae ($h^2 = 0.299$), Bacteroidaceae species *Parabacteroides goldsteinii* ($h^2 = 0.266$) and *Bacteroides coprocola* ($h^2 = 0.228$), *Bifidobacterium longum* ($h^2 = 0.247$), the genus *Phascolarctobacterium* ($h^2 = 0.245$) and a genus-level cluster from the Clostridiales order ($h^2 = 0.237$). Among microbial pathways, only 7 were heritable at FDR-corrected empirical P value < 0.1, including the lipid IV_A biosynthesis pathway (NAGLIPASYN-PWY), two pathways of pyridoxal 5-phosphate biosynthesis (PWY0-845 and PYRIDOXSYN-PWY), the isoleucine biosynthesis II pathway (PWY-5101) and the pre-quinone biosynthesis pathway (PWY-6703). The heritability of pathways and taxa showed a degree of concordance: the NAGLIPASYN-PWY is highly correlated with Proteobacteria ($R = 0.64$), and the other heritable pathways are mostly linked with the family Bacteroidaceae and some Bacteroidaceae species.

Heritability of some of these taxa have been observed previously. *Akkermansia* and *Bifidobacterium* and some genera from the order Clostridiales have been reported to be heritable in a study of UK twins[14], and the heritability of members of the Bacterioidaceae family, including *Bacteroides* and *Parabacteroides*, was reported in a family-based analysis in a Canadian population[15]. Our results did not replicate some of the heritable taxa from these studies, possibly owing to differences in techniques used (use of 16S versus metagenomic sequencing and differing DNA isolation methods) or reference databases—that is, some taxa were not present in the reference databases (such as Christensenellaceae in the Metaphlan2 database), did not pass the 5% presence threshold (the genus *Turicibacter*), or showed very low power for detecting heritability owing to a low presence rate (Euryarchaeota, Christensenellaceae profiled using the Metaphlan3 database).

Cohabitation has a much larger effect than heritability, with 125 out of 257 taxa (48.6%) being significantly affected by cohabitation, and history of cohabitation explained significant variance in 22 taxa (8.6%) (Supplementary Table 5a). The microbiomes of cohabiting participants were also more similar than those of participants living separately, regardless of the relatedness of the participants, with the microbiomes of parent–child pairs, sibling pairs and unrelated partners all resembling each other more than those of non-cohabiting participants (Fig. 2c–e, permutation-based Wilcoxon test FDR < 1.0×10^{-5}). We further observed similar patterns in the compositions of microbial pathways, virulence factors and antibiotic resistance genes (Extended Data Fig. 4a–c).

These results indicate that whole-microbiome composition is significantly influenced by cohabitation, with genetics having a smaller role, although a small subset of microbiota species—such as *A. muciniphila*, *B. longum* and Bacteriodaceae—are significantly heritable and show a degree of replication across cohorts. However, given the setting of our cohort (Methods) and the large cohabitation effect observed for several taxa and pathways, our estimates of heritability may be slightly inflated.

Overview of associations

We then explored the associations of microbial features to 241 measurements including technical factors, anthropometrics, early-life and current exposome, diet, self-reported diseases and use of medication, medical measurements and socioeconomics (Extended Data Fig. 5). These phenotypes explained 12.9% of microbiome taxonomic composition and 16.3% of microbiome function, with the largest contribution coming from technical factors, stool characteristics, diseases, use of medication and anthropometrics (Fig. 3b, Supplementary Tables 4a–c).

After correcting for technical factors, we observed 4,530 associations of phenotypes with taxa, 5,224 with pathways, 1,848 with antibiotic resistance genes and 385 with virulence factors (Supplementary Table 3g, Extended Data Fig. 5). Individually, the largest number of associations were with keystone and core taxa, including *Flavonifractor salivarius*, *F. prausnitzii*, *Alistipes senegalensis* and *Clostridium* and *Subdogranulum* species (Supplementary Tables 3b, h). *A. senegalensi* has previously been associated with Crohn's disease and hepatitis B virus–related acute chronic liver failure[16]. In our cohort, *A. senegalensi* was associated with 43 phenotypes (FDR < 0.05), highlighting a potential role across multiple diseases. Similarly, *Clostridium asparagiforme*, which has been associated with type 2 diabetes (T2D), hypertension and ankylosing spondylitis[10,17], was associated with an additional 23 unrelated diseases (FDR < 0.05). Further associations are discussed below. Supplementary Tables 3a–h provide a complete overview.

Age, sex and BMI ranked among the top phenotypes in our analysis of interindividual variation of beta-diversity, explaining 0.6%, 0.53% and 0.32% of individual variation, respectively. Bristol stool scale explained the largest proportion of beta-diversity ($R^2 = 0.77\%$, FDR = 0.012), with sampling season also explaining a significant proportion of variance ($R^2 = 0.36\%$, FDR = 0.012, Supplementary Table 4a), together highlighting the importance of assessing faecal consistency and collection timeframe effects in microbiome studies. On the basis of these results, we included corrections for age, sex, BMI, Bristol stool scale and sampling season into our association models alongside the corrections for technical factors (Supplementary Tables 3a–e).

Definition of healthy and unhealthy microbiomes

To define the microbiome signatures of health and disease, we associated microbiome features with self-reported health and 81 diseases with

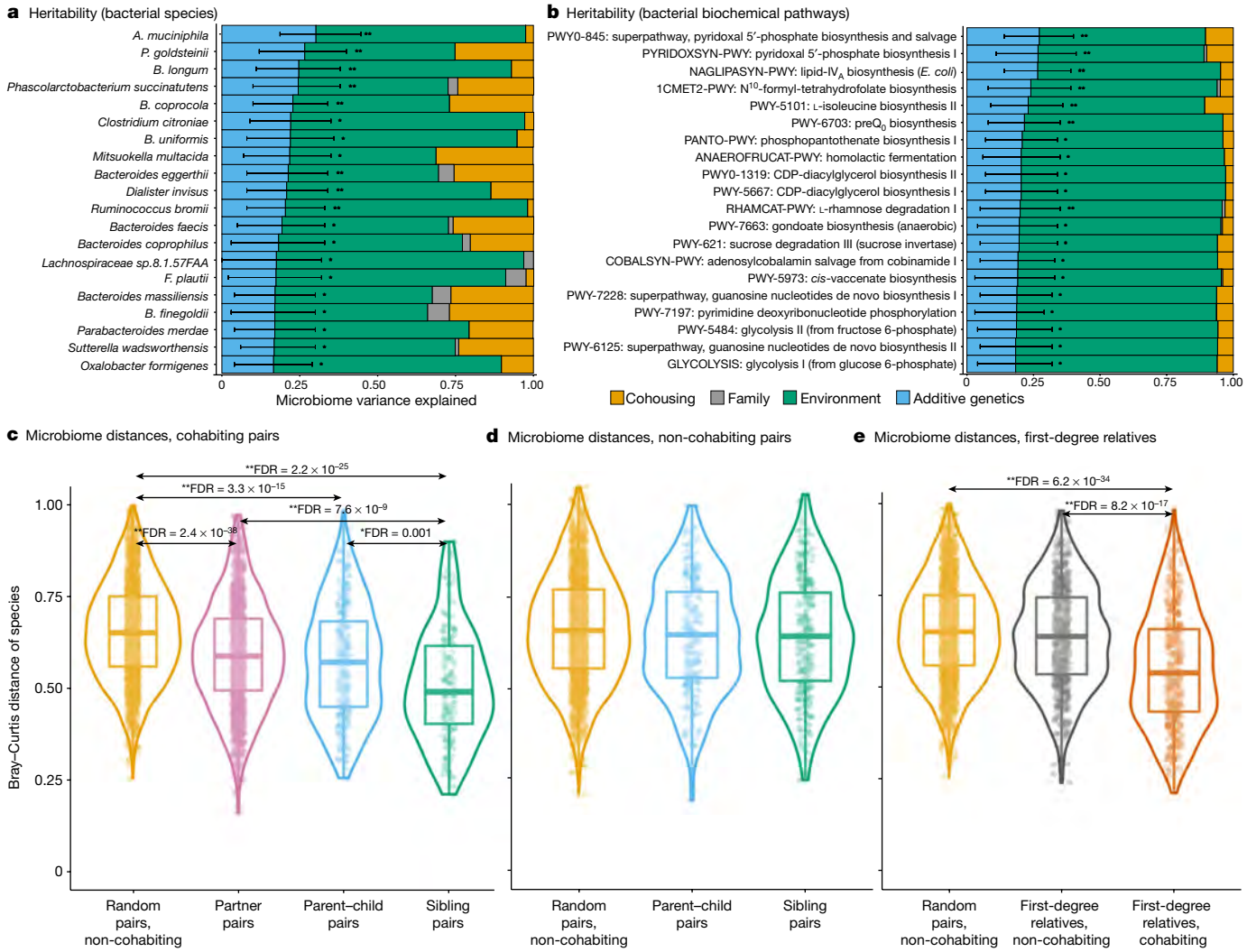


Fig. 2 | Heritability and effect of cohabitation on the gut microbiome. a, Top 20 heritable species. **b**, Top 20 heritable pathways. **Taxa significantly heritable at FDR < 0.1; *taxa with nominally significant heritability ($P < 0.05$). Error bars show 95% confidence interval of heritability. Results in **a**, **b**, were calculated from 3,571 distinct individuals from 1,432 families. **c**–**e**, Pairwise microbiome distance comparisons. Centre line is the median, box limits indicate upper and lower quartiles, whiskers show 1.5× interquartile range, points indicate outliers and the outline displays the distribution of the data. Bray–Curtis dissimilarities were calculated using microbial species of groups of random non-cohabiting pairs ($n = 2{,}000$) compared with cohabiting partners ($n = 1{,}710$), parent–child pairs ($n = 285$) and sibling pairs ($n = 144$) (**c**), random pairs ($n = 2{,}000$) compared with non-cohabiting parent–child pairs ($n = 301$) and sibling pairs ($n = 299$) (**d**) and random pairs ($n = 2{,}000$) compared with non-cohabiting first-degree relative pairs ($n = 600$) and cohabiting first-degree relative pairs ($n = 429$) (**e**). In **c**–**e**, **FDR < 1.0×10^{-5}, *FDR < 0.05 (permutation-based Wilcoxon test, two-sided).

at least 20 cases (Supplementary Table 2d). This identified 1,206 significant associations with bacterial taxa, 1,182 with microbial pathways, 390 with antibiotic resistance genes and 76 with bacterial virulence factors (FDR < 0.05, Supplementary Tables 3b–f). Different diseases had different numbers of associations, and the strongest signatures were observed for cardiovascular and metabolic disorders, such as non-alcoholic fatty liver disease and T2D, and for gastrointestinal disorders including inflammatory bowel disease and IBS (Supplementary Table 3f). We observed consistent microbiome–disease patterns across the majority of diseases (Fig. 4a), enabling us to pinpoint microbiome signatures that were shared between unrelated diseases as well as features that define a healthy (that is, absence of disease) microbiome.

The shared microbiome signatures of diseases (see Supplementary Discussion for details) mainly consisted of increases in *Anaerotruncus*, *Ruminococcus*, *Bacteroides*, *Holdemania*, *Flavonifractor*, *Eggerthella* and *Clostridium* species and decreases in *Faecalibacterium*, *Bifidobacterium*, *Butyrivibrio*, *Subdoligranulum*, *Oxalobacter*, *Eubacterium* and *Roseburia*. Gut microbiome pathways shared across unrelated diseases consisted mainly of increases in biosynthesis of L-ornithine, ubiquinol and menaquinol, enterobacterial common antigen, Kdo-2-lipid-A and molybdenum cofactor and decreases in biosynthesis of amino acids, deoxyribonucleosides and nucleotides, anaerobic energy metabolism and fermentation to short chain fatty acids (mainly butanoate). Virulence factors were increased in some diseases, including T2D and gastrointestinal disorders, with the largest effects observed for bacterial adherence and iron-uptake factors (VF036, VF0228, VF0236, VF0404 and VF0394). We further validated these signals by constructing L1/L2 regularized regression models for prediction of the 36 most common diseases and identified high consistency with association analysis, with 22 out of 31 species selected by the models associated with more than 5 diseases (Supplementary Discussion). The predictive features showed high correlation despite the low disease comorbidities in our cohort (Fig. 4b, Supplementary Table 6e).

Finally, we calculated the recently developed Gut Microbiome Health Index[18] (GMHI) and identified a significant difference between healthy and unhealthy individuals (P value = 6.16×10^{-22}; Extended

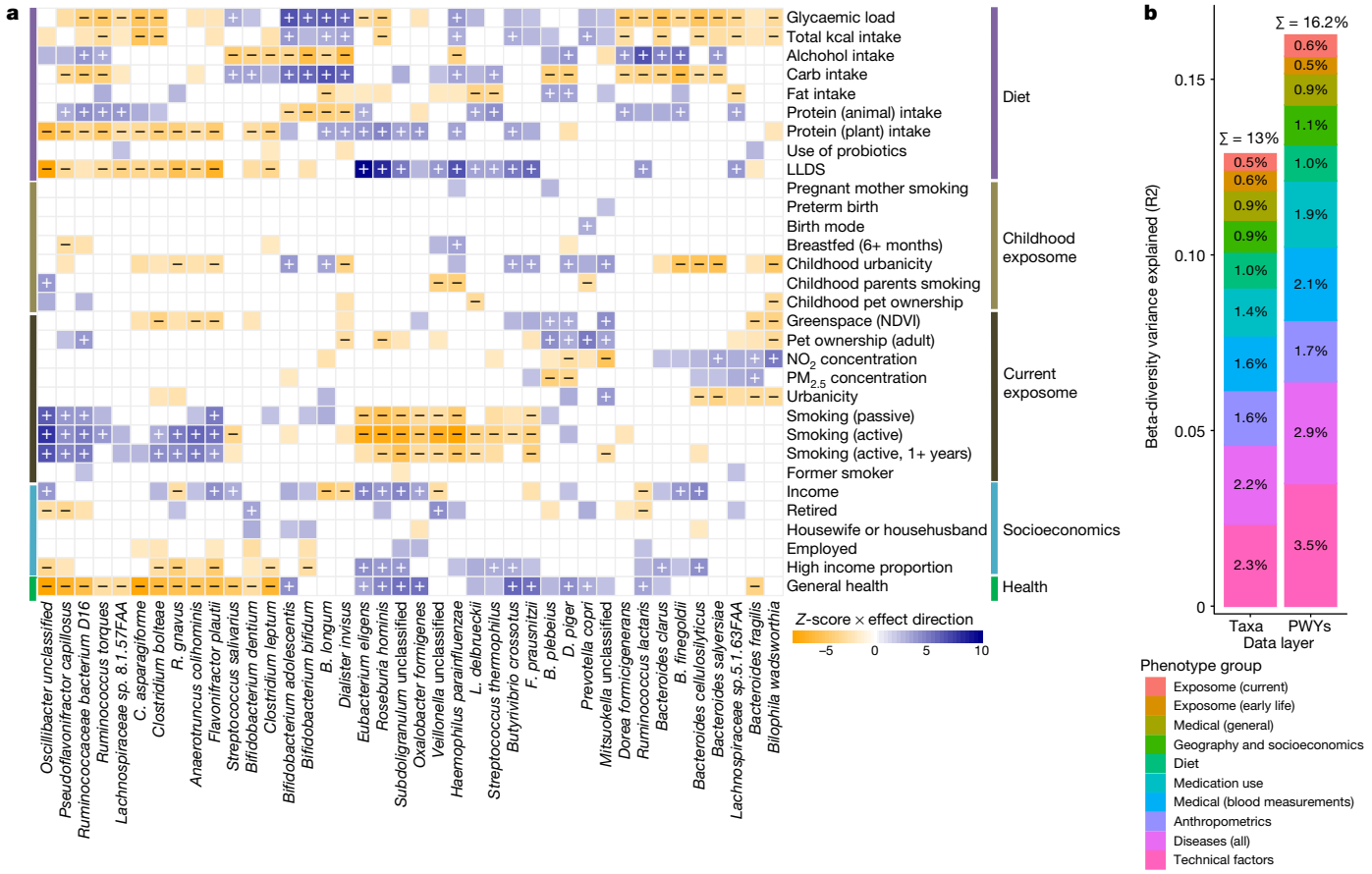


Fig. 3 | Microbiome–phenotype associations. a, Microbiome–phenotype associations for diet, childhood and current exposome and socioeconomics in comparison to healthy microbiome signature. Top 40 microbial species with the highest number of significant associations are clustered by association *Z*-score using hierarchical clustering and coloured by direction of effect (blue, positive; orange, negative), with associations significant at study-wide FDR < 0.05 marked with plus and minus for positive and negative correlations, respectively. Coloured associations without a mark indicate nominally significant associations ($P < 0.05$). **b**, Variance in microbiome composition and function explained by phenotype groups in multivariate PERMANOVA analysis. NDVI, Normalized Difference Vegetation Index; PWYs, pathways.

Data Fig. 6). Our microbiome signature of health replicated 43 out of 50 GMHI signals[18] at genus or species level (Supplementary Table 10). In addition, we identified 55 microbiome–health associations that were not seen in the GMHI study, including with species from *Butyrivibrio*, *Akkermansia* and *Prevotella* genera (Supplementary Tables 3b, 10) that were previously linked to gastrointestinal disorders but not to other diseases[1,2].

The microbiome reflects diseases and treatments

We observed high consistency in the microbiome effects of common non-communicable diseases and the medications used to treat them (Extended Data Fig. 7, Supplementary Tables 3a–g). Among these, we identified the strongest effects for proton pump inhibitors (PPIs), antibiotics, biguanide antidiabetics, osmotic laxatives and intestinal anti-inflammatory agents (84, 56, 47 and 32 associations with microbial taxa, respectively, at FDR < 0.05). To disentangle the effects of diseases and medications, we performed multivariate linear regression of three common diseases and their corresponding common medications: antidiabetics in T2D, selective serotonin reuptake inhibitors (SSRIs) in depression and PPIs in functional gastrointestinal disorders and IBS. We observed that effect of diseases and corresponding drugs are associated with microbiome with high consistency, even when conditioned on each other (Supplementary Table 7), indicating that the unhealthy gut microbiome signature reflects both the disease and the associated medication.

Childhood is linked to the adult microbiome

As the first two to three years of life are crucial for microbiome development, we examined the influence of early-life (less than four years of age) factors on the adult microbiome. We identified 106 associations with taxa, 30 with pathways, 22 with antibiotic resistance genes and 2 with virulence factors (FDR < 0.05), with only minimal effects observed for birth mode, breastfeeding and preterm birth (Fig. 3b, Extended Data Fig. 8). Childhood living environment (scale from 1, rural to 5, highly urban) was significantly associated with the adult microbiome (54, 8 and 7 associations with taxa, pathways and CARDs, respectively, at FDR < 0.05) despite a very low correlation between childhood and adult urbanicity (Spearman $r < 0.015$).

Rural childhood environment was associated with an increase in various bacteria, including *P. copri*, *F. prausnitzii*, *Rothia mucilaginosa* and species from the *Bifidobacterium* and *Mitsuokella* genera, whose abundances were also associated with increased general health. By contrast, the gut abundances of multiple species from genera *Bacteroides*, *Alistipes* and *Biophila* were reduced in participants with an urban childhood environment.

Parental smoking was associated with the microbiome of their children (15, 9 and 4 associations with taxa, pathways and CARDs, respectively, at FDR < 0.05). Here we observed associations between parental smoking and decreased abundances of *Veillonella* and *Oscillibacter* species and of *P. copri* ($R^2 = 0.0012$, FDR = 0.033), consistent with observations for current smokers. Finally, childhood pet ownership was

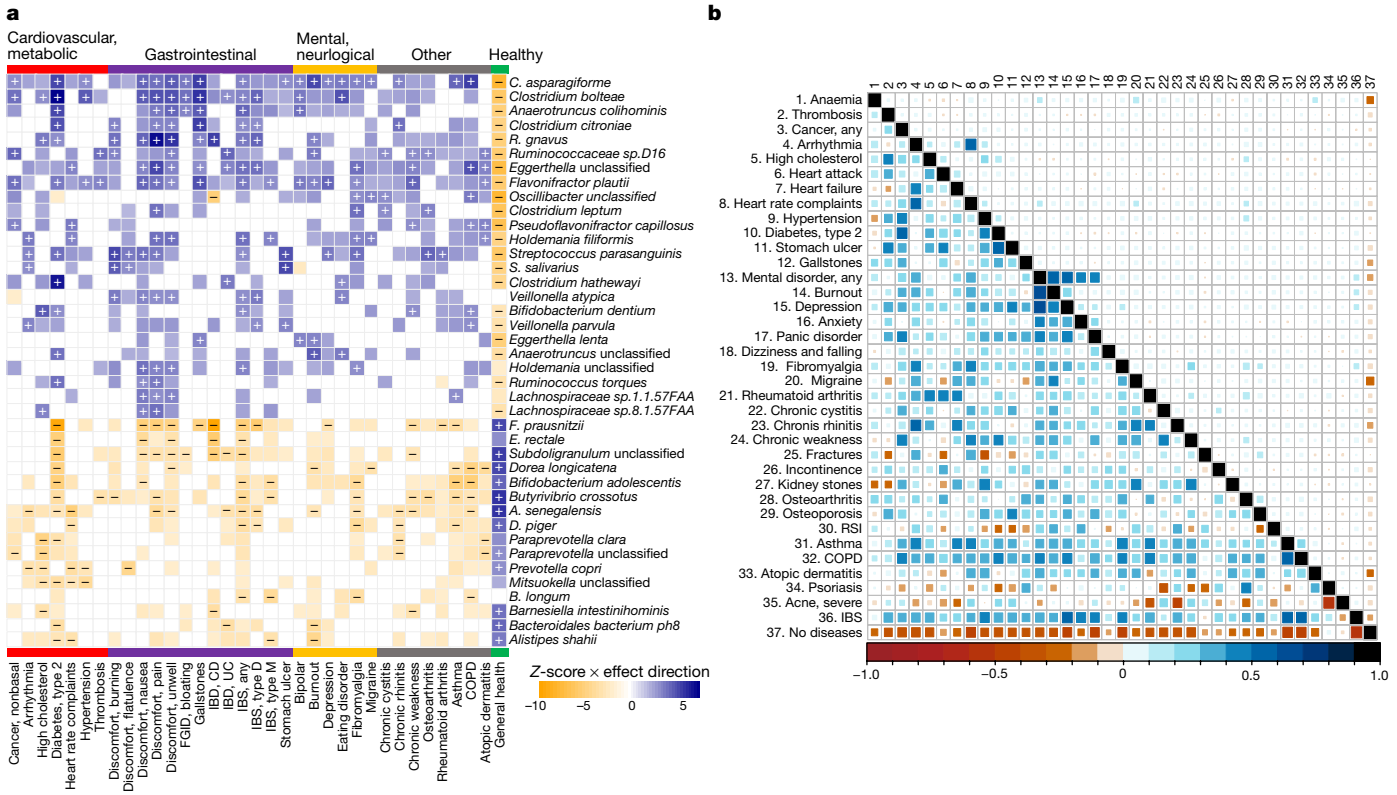


Fig. 4 | Microbiome signatures of health and diseases. a, Heat map of microbial species associated with categories of diseases and health status. Diseases are sorted and labelled by disease type. Top 40 microbial species with the highest number of associations are clustered by association *Z*-score (indicated by colour intensity) using hierarchical clustering. Associations are coloured by direction of effect (blue, positive; orange, negative), with associations significant at study-wide FDR < 0.05 marked with plus and minus for positive and negative correlations, respectively. Coloured associations without a label indicate nominally significant associations ($P < 0.05$, no multiple-testing correction). **b**, Comparison of Pearson correlations between signatures predictive for diseases (bottom left triangle) and comorbidities of these diseases in the cohort (top right triangle). CD, Crohn's disease; COPD, chronic obstructive pulmonary disease; IBD, inflammatory bowel disease; IBS, irritable bowel syndrome; RSI, repetitive strain injury; UC, ulcerative colitis.

associated with the adult microbiome (7 associations at FDR < 0.05), including decreases in *Alistipes finegoldii*, *Lactobacillus delbrueckii* and *Dialister* and *Bilophila* species observed in participants who had childhood pets.

Exposome is associated with gut microbiome

We studied environmental factors at time of sampling and identified shared association patterns of healthy microbiome signatures with pet ownership, rural living environment and greenspace surface area in living environment (Fig. 3b), including increases in *P. copri*, *Bacteroides plebeius*, *Desulfovibrio piger* and *Mitsuokella* species, and decreases in *Bacteroides fragilis* and *Bilophila wadsworthia*. These signals contrast with the microbiome associations of increased measurements of NO_2 and small particulate matter pollutants, which are negatively associated with health (Extended Data Fig. 9). Smoking phenotypes, including current active and passive smoking and a history of smoking, showed consistent directions of microbiome associations, which also matched the signatures of microbiome–disease associations (Extended Data Fig. 9). Active smoking was associated with 41 species and 84 pathways (FDR < 0.05, Fig. 3b), with 60% of these also being associated with smoking history, suggesting a long-lasting effect of smoking. Of note, 15 of these were also associated with passive smoking, highlighting the need to consider passive smoking in disease risk models.

We observed 220 associations (FDR < 0.05) between socioeconomic factors (for example, monthly income and neighbourhood income) and gut microbiome, of which 72 were between bacterial abundances and monthly income; higher income was associated with a healthy microbiome signature. Income showed low, but significant, correlations with neighbourhood greenspace area, rural living environment and Lifelines diet score (LLDS) (Spearman correlations 0.22, 0.17 and 0.07, respectively; correlation test FDRs < 1.0×10^{-6}), implying that the microbiome–income association probably reflects a combination of factors, including healthier diet and lifestyle and less-urban living environment. These results support the hypothesis that gut microbiome might be a mediator of the well-known differences in health across the socioeconomic spectrum[19].

We identified 378 associations between 20 dietary factors and 82 species (Supplementary Table 3b). Diet was also found to be relatively stable over the 5-year period between food frequency questionnaire (FFQ) collection and faecal sampling, with 30 FFQ items that represent major dietary items and dietary trends being more than 95% conserved across measurements and the remaining items being more than 50% conserved (Supplementary Table 8, Supplementary Discussion).

The LLDS, a diet score based on international nutrition literature, showed the highest number of associations (79 associations with taxa, 44 with pathways, 20 with antibiotic resistance genes and 8 with virulence factors at FDR < 0.05), followed by total alcohol intake, glycaemic load, protein diet score (reflecting quantity and source of protein) and total carbohydrate intake. The LLDS and protein intake scores showed association patterns that overlapped with microbiome signals of increased general health—for example, decreases in *Clostridia* species and increases in *Butyrivibrio* and *Roseburia* genera and pathways involved in ubiquinol and menaquinol synthesis. By contrast, total dietary carbohydrate intake and glycaemic load showed the opposite associations (Fig. 3b, Extended Data Fig. 10).

Discussion

In addition to confirming known microbiome associations with age, sex and BMI[20], our results highlight the importance of frequently omitted confounders related to stool samples[21] (stool consistency), sampling season[22] and sample processing[23] (DNA concentration or sequencing batch). This is especially important when studying diseases in which age, sex, BMI and faecal consistency are associated with the disease or exploring phenotypes with seasonal variation such as diet[24], physical activity[25] and diseases such as allergies, flu and common cold.

Our observation that the microbiome is primarily associated with cohabitation and environment rather than genetic relatedness corroborates previous studies that identified limited overall microbiome heritability[26], shared microbiome patterns between cohabiting family members and their pets[27] and microbial divergence in twins who stopped cohabiting[28]. These results suggest that bacteria with low heritability, such as *Ruminococcus*, *Streptococcus* and *Veillonella* species, might be more susceptible to microbiome-altering therapies than more heritable bacteria, such as *Akkermansia, Collinsella* and *Bacteroides* species.

By comparing associations between microbiome, health and diverse diseases, we identified a common signal for gut dysbiosis (Fig. 4a) that was largely consistent with a previous study[18]. The existence of shared dysbiosis has considerable implications for microbiome research and microbiota-targeting diagnostics and therapies. Shared dysbiosis implies that the gut microbiome is a biomarker of general health, as supported by our prediction models and previous studies[18,29], but also complicates microbiome-based diagnosis of individual diseases. As single-disease models might be confounded by signals shared across unrelated diseases, testing such models for specificity in mixed-disease cohorts will be an important step before clinical implementation. The shared microbiome signature also suggests that microbiome-targeting interventions could improve overall human health. This is supported by our observations that lifestyle factors generally considered healthy—for example, adherence to current dietary recommendations and not smoking—associate with microbiome patterns similar to those associated with general health. Although microbiome–drug interactions are well described in vitro[30] and have been characterized in vivo for antibiotics, PPIs[31] and antidiabetics[32], our results suggest that general microbiome dysbiosis arises as a result of a combination of drug and disease effects, implying that many currently understudied drugs, such as SSRIs, might have a negative effect on the gut microbiome.

The presence of disease-like microbiome signatures in population participants not reporting diseases or medication use indicates the presence of pre-clinical 'hidden' disorders in the population and suggests that gut dysbiosis might precede clinical onset of chronic diseases such as T2D. Although this hypothesis requires experimental validation or analysis of long-timeframe longitudinal cohorts, our observations suggest that the gut microbiome might be used to monitor long-term health and detect disorders in the pre-clinical stage.

Linking healthy and unhealthy microbiome patterns to childhood and current exposome, diet and socioeconomics, we observed that healthier diet, childhood and current exposures to rural environment and pets, exposure to green space and higher income share signals with healthy microbiome patterns. These observations support the microbiome diversity hypothesis (also called the hygiene hypothesis), which postulates that reduced exposure to microbiota contributes to an increase in the frequency of autoimmune and allergic diseases[33]. Notably, whereas the classic hygiene hypothesis focuses on pathogens and early-life exposures, our results suggest that adult exposures also contribute to healthy or unhealthy microbiome patterns and that the environment shapes the microbiome throughout life, meaning that microbiome-targeted therapies could be effective throughout an individual's life. Furthermore, we identified negative correlations of diet scores, pets and rural environment with opportunistic pathogens such as *Clostridia* species as well as positive correlations with commensals such as butyrate producers from *Bacteroides*, *Alistipes* and *Faecalibacterium*, implying that exposure to pathogens and commensals from the environment has an important role in establishing a healthy gut ecosystem.

We also observed that smoking, a high-carbohydrate diet and exposure to NO_2 and small particulate matter ($PM_{2.5}$) are positively correlated with disease-linked *Clostridia* and *Ruminococcus* species. Whereas air pollutants have been associated with gastrointestinal diseases such as IBS[34] and have been shown to affect the gut microbiota of mice, their effects on the human gut microbiota remain largely unexplored[35]. Our results agree with a previous analysis of association of air pollution in an American cohort[36] and suggest that air pollutants negatively affect the human gut microbiota and might increase the risk of gastrointestinal diseases by contributing to general dysbiosis. This is further supported by our observation that the presence of IBS correlates with $PM_{2.5}$ pollutants (Spearman $r = 0.15$), whereas the correlations between pollutants and other diseases, socioeconomic factors and diet are very low (Spearman $r < 0.05$) (Supplementary Table 2c).

We found that childhood exposures to smoking, pets and rural environment are associated with the adult microbiome. Although the effect sizes for these associations were lower than for current exposures, the effect directions and patterns were consistent, suggesting that environmental exposures can have a long-lasting effect and that the microbiome reflects an individual's history of exposures. This is further supported by our finding that former smokers still showed microbiome associations similar to those of current smokers, albeit with lower effect sizes.

We measured a broad range of 241 phenotypes, but we could only explain around 15% of the interindividual variation in microbiome composition and function, which is consistent with previous large-scale studies[5,26,37]. This implies that the gut microbiome is highly individual, contains rare taxa that might be difficult to disentangle from artefacts in the data[38], and that our current understanding of the factors that shape the gut microbiome is still limited. This low explanatory power might also reflect the use of database-centred microbiome classification which, while facilitating standardization of studies and a low false-positive rate, precludes identification of uncatalogued microbes[39]. Future quantification of 'missing variance', potentially by assembly-based and database-independent methods and longitudinal studies, will have a critical role in development of microbiome-targeting diagnostics and therapies.

Conclusion

We have generated and analysed a large, multi-generational gut microbiome cohort that has been collected and profiled in a highly standardized matter and linked it to extensive phenotype data. We define and describe a gut dysbiosis shared across diverse diseases and identify links between this dysbiosis and heritability, childhood and current exposome, lifestyle and socioeconomics. This study demonstrates the power of large-scale, well-phenotyped cohorts for dissecting the links between gut microbiome, health, genetics and environment and provides a rich resource for future studies for microbiome-directed interventions.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-04567-7.

1. Lynch, S. V. & Pedersen, O. The human intestinal microbiome in health and disease. *N. Engl. J. Med.* **375**, 2369–2379 (2016).
2. Liang, D., Leung, R. K., Guan, W. & Au, W. W. Involvement of gut microbiome in human health and disease: brief overview, knowledge gaps and research opportunities. *Gut Pathog.* **10**, 3 (2018).
3. Zmora, N., Soffer, E. & Elinav, E. Transforming medicine with the microbiome. *Sci. Transl. Med.* **11**, eaaw1815 (2019).
4. Garber, K. First microbiome-based drug clears phase III, in clinical trial turnaround. *Nat. Rev. Drug Discov.* **19**, 655–656 (2020).
5. Zhernakova, A. et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).
6. Gaulke, C. A. & Sharpton, T. J. The influence of ethnicity and geography on human gut microbiome composition. *Nat. Med.* **24**, 1495–1496 (2018).
7. Vatanen, T. et al. Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. *Nat. Microbiol.* **4**, 470–479 (2019).
8. Berry, D. & Widder, S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Front. Microbiol.* **5**, 219 (2014).
9. Miquel, S. et al. *Faecalibacterium prausnitzii* and human intestinal health. *Curr. Opin. Microbiol.* **16**, 255–261 (2013).
10. Huang, R. et al. Metagenome-wide association study of the alterations in the intestinal microbiome composition of ankylosing spondylitis patients and the effect of traditional and herbal treatment. *J. Med. Microbiol.* **69**, 797–805 (2020).
11. Kandeel, W. A. et al. Impact of *Clostridium* bacteria in children with autism spectrum disorder and their anthropometric measurements. *J. Mol. Neurosci.* **70**, 897–907 (2020).
12. Tap, J. et al. Identification of an intestinal microbiota signature associated with severity of irritable bowel syndrome. *Gastroenterology* **152**, 111–123.e8 (2017).
13. Costea, P. I. et al. Enterotypes in the landscape of gut microbial community composition. *Nat. Microbiol.* **3**, 8–16 (2018).
14. Goodrich, J. K. et al. Genetic determinants of the gut microbiome in UK twins. *Cell Host Microbe* **19**, 731–743 (2016).
15. GEM Project Research Consortium. Association of host genome with intestinal microbial composition in a large healthy cohort. *Nat. Genet.* **48**, 1413–1417 (2016).
16. Wang, K. et al. Gut microbiota as prognosis markers for patients with HBV-related acute-on-chronic liver failure. *Gut Microbes* **13**, 1–15 (2021).
17. Wang, P. et al. Cigarette smoking status alters dysbiotic gut microbes in hypertensive patients. *J. Clin. Hypertens.* **23**, 1431–1446 (2021).
18. Gupta, V. K. et al. A predictive index for health status using species-level gut microbiome profiling. *Nat. Commun.* **11**, 1–16 (2020).
19. Bowyer, R. et al. Socioeconomic status and the gut microbiome: a TwinsUK cohort study. *Microorganisms* **7**, 17 (2019).
20. Yatsunenko, T. et al. Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
21. 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).
22. Davenport, E. R. et al. Seasonal variation in human gut microbiome composition. *PLoS ONE* **9**, e90731 (2014).
23. The Microbiome Quality Control Project Consortium. Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. *Nat. Biotechnol.* **35**, 1077–1086 (2017).
24. Toorn, J. E. et al. Seasonal variation of diet quality in a large middle-aged and elderly Dutch population-based cohort. *Eur. J. Nutr.* **59**, 493 (2020).
25. Tucker, P. & Gilliland, J. The effect of season and weather on physical activity: a systematic review. *Public Health* **121**, 909–922 (2007).
26. Rothschild, D. et al. Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).
27. Song, S. J. et al. Cohabiting family members share microbiota with one another and with their dogs. *eLife* **2**, e00458 (2013).
28. 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).
29. Oh, M. & Zhang, L. DeepMicro: deep representation learning for disease prediction based on microbiome data. *Sci. Rep.* **10**, 6026 (2020).
30. Zimmermann, M., Zimmermann-Kogadeeva, M., Wegmann, R. & Goodman, A. L. Mapping human microbiome drug metabolism by gut bacteria and their genes. *Nature* **570**, 462–467 (2019).
31. Imhann, F. et al. Proton pump inhibitors affect the gut microbiome. *Gut* **65**, 740–748 (2016).
32. Wu, H. et al. Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug. *Nat. Med.* **23**, 850–858 (2017).
33. Bach, J. F. The hygiene hypothesis in autoimmunity: the role of pathogens and commensals. *Nat. Rev. Immunol.* **18**, 105–120 (2018).
34. Marynowski, M. Role of environmental pollution in irritable bowel syndrome. *World J. Gastroenterol.* **21**, 11371 (2015).
35. Dujardin, C. E. et al. Impact of air quality on the gastrointestinal microbiome: a review. *Environ. Res.* **186**, 109485 (2020).
36. Fouladi, F. et al. Air pollution exposure is associated with the gut microbiome as revealed by shotgun metagenomic sequencing. *Environ. Int.* **138**, 105604 (2020).
37. Manor, O. et al. Health and disease markers correlate with gut microbiome composition across thousands of people. *Nat. Commun.* **11**, 5206 (2020).
38. Sogin, M. L., Morrison, H., McLellan, S., Welch, D. & Huse, S. The rare biosphere: sorting out fact from fiction. *Genome Biol.* **11**, I19 (2010).
39. Sun, Z. et al. Challenges in benchmarking metagenomic profilers. *Nat. Methods* **18**, 618–626 (2021).
40. Scholtens, S. et al. Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int. J. Epidemiol.* **44**, 1172–1180 (2015).
41. Tigchelaar, E. F. et al. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open* **5**, e006772 (2015).
42. Siebelink, E., Geelen, A. & de Vries, J. H. M. Self-reported energy intake by FFQ compared with actual energy intake to maintain body weight in 516 adults. *Br. J. Nutr.* **106**, 274–281 (2011).
43. Brouwer-Brolsma, E. M. et al. A National Dietary Assessment Reference Database (NDARD) for the Dutch population: rationale behind the design. *Nutrients* **9**, 1136 (2017).
44. Willett, W. C. et al. Reproducibility and validity of a semiquantitative food frequency questionnaire. *Am. J. Epidemiol.* **122**, 51–65 (1985).
45. Vinke, P. C. et al. Development of the food-based Lifelines Diet Score (LLDS) and its application in 129,369 Lifelines participants. *Eur. J. Clin. Nutr.* **72**, 1111–1119 (2018).
46. Møller, G. et al. A protein diet score, including plant and animal protein, investigating the association with HbA1c and eGFR—the PREVIEW project. *Nutrients* **9**, 763 (2017).
47. Leeming, E. R., Johnson, A. J., Spector, T. D. & Le Roy, C. I. Effect of diet on the gut microbiota: rethinking intervention duration. *Nutrients* **11**, 2862 (2019).
48. Eeftens, M. et al. Development of land use regression models for $PM_{2.5}$, $PM_{2.5}$ absorbance, PM_{10} and PM_{coarse} in 20 European study areas; results of the ESCAPE project. *Environ. Sci. Technol.* **46**, 11195–11205 (2012).
49. Beelen, R. et al. Development of NO_2 and NO_x land use regression models for estimating air pollution exposure in 36 study areas in Europe—the ESCAPE project. *Atmos. Environ.* **72**, 10–23 (2013).
50. Eeftens, M. et al. Stability of measured and modelled spatial contrasts in NO_2 over time. *Occup. Environ. Med.* **68**, 765–770 (2011).
51. Ford, A. C. et al. Validation of the Rome III criteria for the diagnosis of irritable bowel syndrome in secondary care. *Gastroenterology* **145**, 1262–1270.e1 (2013).
52. Angulo, P. et al. The NAFLD fibrosis score: a noninvasive system that identifies liver fibrosis in patients with NAFLD. *Hepatology* **45**, 846–854 (2007).
53. Bedogni, G. et al. The fatty liver index: a simple and accurate predictor of hepatic steatosis in the general population. *BMC Gastroenterol.* **6**, 33 (2006).
54. Imhann, F. et al. The 1000IBD project: multi-omics data of 1000 inflammatory bowel disease patients; data release 1. *BMC Gastroenterol.* **19**, 5 (2019).
55. McIver, L. J. et al. bioBakery: a meta'omic analysis environment. *Bioinformatics* **34**, 1235–1237 (2018).
56. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
57. Truong, D. T. et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
58. Franzosa, E. A. et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).
59. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
60. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
61. Swertz, M. A. et al. The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button. *BMC Bioinf.* **11**, S12 (2010).
62. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **14**, 927–930 (2003).
63. Hsieh, T. C., Ma, K. H. & Chao, A. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). *Methods Ecol. Evol.* **7**, 1451–1456 (2016).
64. Kaminski, J. et al. High-specificity targeted functional profiling in microbial communities with ShortBRED. *PLoS Comput. Biol.* **11**, e1004557 (2015).
65. Chen, L. et al. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.* **33**, D325–D328 (2005).
66. Jia, B. et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **45**, D566–D573 (2017).
67. Ziyatdinov, A. et al. lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals. *BMC Bioinf.* **19**, 68 (2018).
68. Aitchison, J. The Statistical Analysis of Compositional Data. *J. R. Stat. Soc. Series B Stat. Methodol.* **44**, 139–177 (1982).
69. Turnbaugh, P. J. et al. The Human Microbiome Project. *Nature* **449**, 804–810 (2007).
70. Qin, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
71. Goodrich, J. K. et al. Human genetics shape the gut microbiome. *Cell* **159**, 789–799 (2014).
72. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8**, e1002687 (2012).
73. Chen, L. et al. Gut microbial co-abundance networks show specificity in inflammatory bowel disease and obesity. *Nat. Commun.* **11**, 4018 (2020).
74. Arumugam, M. et al. Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011).

Methods

Population cohort and metadata collection

The Lifelines Dutch Microbiome Project (DMP) cohort was developed as a part of Lifelines cohort study. Lifelines is a multi-disciplinary prospective population-based cohort study which utilizes a unique three-generation design to examine health and health-related behaviours in 167,729 people living in the northern Netherlands. Lifelines employs a broad range of investigative procedures to assess the biomedical, socio-demographic, behavioural, physical and psychological factors that contribute to health and disease, with a special focus on multi-morbidity and complex genetics[40,41]. To form the DMP cohort, 8,719 distinct fresh-frozen faecal and blood samples were collected from Lifelines participants in 2015 and 2016 (one sample per individual). Whole-genome shotgun sequencing was performed on one aliquot from each of 8,534 faecal samples, and 8,208 were retained for downstream analysis after stringent quality control. Metadata information collected from the participants was grouped into the following categories: family structure, diseases, gastrointestinal complaints, general health score, medication use, anthropometrics, birth-related factors, reported childhood (<16 years) exposures, current exposome (air pollutants, greenspace, urbanicity, pets and smoking), socioeconomic characteristics and diet (Supplementary Table 2d).

Informed consent

The Lifelines study was approved by the medical ethical committee of the University Medical Center Groningen (METc number: 2017/152). Additional written consent was signed by all DMP participants or their parents or legal representatives (for children aged under 18).

Metadata

Metadata was collected by questionnaires and curated as described previously[41] and below. We included 241 phenotypes from a broad range of categories, including socioeconomic factors, self-reported diseases and medications, quality of life, mental health, education and employment, nutrition, smoking, stress and childhood environmental factors. Questionnaires were developed and processed by the Lifelines cohort study[41] as described at www.lifelines.nl. Additional in-depth data curation and acquisition was performed to assess dietary intake, air pollution and environmental exposures, medication use and gut health, as described below.

Diet

Habitual diet was assessed through a semiquantitative FFQ collected 4 years prior to DMP faecal sampling[41]. The FFQ was designed and validated by the division of Human Nutrition of Wageningen University, using standardized methods[42]. It assesses how often a food was consumed over the previous month on a scale ranging from 'never' to '6–7 days per week' and the usual amount taken. Average daily nutrient intake was calculated using the Dutch Food Composition database (NEVO, RIVM) and a mono- and disaccharide-specific food composition table[43], resulting in the generation of data on 21 dietary factors. Energy adjustment was performed by means of the nutrient density method[44]. Published dietary scores and inter-nutrient ratios were calculated as indicators of dietary quality and composition[45,46].

To validate the assumed stability of FFQs across time[43,47], we used questionnaires from 128,501 Lifelines participants to study diet consistency between the baseline questionnaire collected 4 years prior to the DMP study and a second smaller-scope nutrient-specific questionnaire collected concurrently with DMP faecal sampling. 65 dietary questions, reflecting consumption of major food categories such as fruits, vegetables, fish, meat, bread, grains and sweets, as well as special dietary regimes (e.g. vegan or macrobiotic diet), were compared between the first and second time point. The majority of dietary items were available for > 44,000 individuals at both time points (Supplementary Table 8). For each food item, we computed the mean (absolute) change between results of FFQs at time point 1 and time point 2. The FFQ results were encoded as numbers corresponding to FFQ answers (e.g. 1 = participant reports "never" consuming the food item, 4 = the food item is consumed "always" or "every day"). The changes were calculated as:

$$\mathrm{MC}_i = \frac{1}{n} \times \sum_{j=1}^{n} x_{ij} - y_{ij}$$

where MC is mean (absolute) change, n is the number of individuals who answered FFQs at the two time points for food item i, and x and y represent results of baseline and follow-up FFQs. Each of the items in the sum corresponds to FFQ results of one participant (Supplementary Table 8).

Exposome

Elements of the exposome, neighbourhood urbanicity and income were assessed for the participant's home address at the time of faecal sampling. Exposure to two air pollutants, particulate matter with aerodynamic diameter ≤ 2.5 μm ($PM_{2.5}$) and nitrogen dioxide (NO_2), was assigned based on land-use regression models developed in the European Study of Cohorts for Air Pollution Effects (ESCAPE) project[48,49]. These estimates are based on measurement data from 2009 and reflect long-term ambient air pollution exposures[50].

Greenspace was assigned using the NDVI, which reflects the average density of green vegetation within a 100 m circular buffer around the participant's residential address. The NDVI was derived from a LANDSAT 5 (Thematic Mapper) satellite image taken in 2016 and captures the density of green vegetation at a spatial resolution of 30 × 30 m based on land surface reflectance of the visible (red) and near-infrared parts of the spectrum.

Neighbourhood urbanicity was assigned based on a five-category scale of surrounding address density developed by Statistics Netherlands (1, very urban, ≥2,500 addresses per km^2 to 5, very rural, <500 addresses per km^2, data for 2015). Neighbourhood income was considered a proxy for neighbourhood socioeconomic position and defined as the proportion of individuals with low (<40th percentile) income (Statistics Netherlands, data for 2015; https://opendata.cbs.nl/#/CBS/en/). Childhood neighbourhood urbanicity was defined based on the self-reported answer to the Lifelines biobank questionnaires question "What is the best description of the place where you lived most of the time when you were younger than 5 years old?", with possible answers being 'farm', 'rural/village', 'small town or large village', 'suburb of a large city' and 'city centre'.

Stool characteristics, diseases and medication

For 7 days in the week of stool sample collection, DMP participants recorded a bowel movement diary, Bristol stool scale, daily medication use and gastrointestinal symptoms daily, and these records were used to extract information on stool frequency, stool characteristics, drug use and gastrointestinal symptoms during the week of stool collection. The validated ROME III questionnaires[51] were used to characterize functional gastrointestinal disorders, and participants were classified as having either no functional gastrointestinal diseases or IBS, functional diarrhoea, functional constipation or functional bloating. Information about the presence of other diseases was self-reported and collected using Lifelines questionnaires. Diseases were grouped into 11 disease categories. The presence of cancer was grouped into a separate category defined as 'any cancer', independent of cancer type. Non-alcoholic fatty liver disease fibrosis score[52] and fatty liver index[53] were calculated from the anthropometrics and blood measurements, as described previously[52,53]. Diseases with < 20 cases were excluded from further analysis. Self-reported medications were grouped into categories based on Anatomical Therapeutic Chemical classification (ATC codes) at the most specific ATC level (5-digit ATC code if possible).

ATC categories with < 20 users were grouped into a higher (4-digit or 3-digit) ATC class. Categories with < 20 individuals that could not be grouped according to ATC classification were excluded from further analysis. In total, 62 drug groups were included (Supplementary Table 2a).

Faecal sample collection, DNA extraction and sequencing

Faecal sample collection was performed by participants at home. Participants were asked to freeze stool samples within 15 min of stool production. Frozen samples were collected by Lifelines personnel, transported to the Lifelines biorepository on dry ice and stored at −80 °C until DNA extraction. Microbial DNA was isolated with the QIAamp Fast DNA Stool Mini Kit (Qiagen), according to the manufacturer's instructions, using the QIAcube (Qiagen) automated sample preparation system. Library preparation for samples with total DNA yield lower than 200 ng (as determined by Qubit 4 Fluorometer) was performed using NEBNext Ultra DNA Library Prep Kit for Illumina, while libraries for other samples were prepared using NEBNext Ultra II DNA Library Prep Kit for Illumina. Metagenomic sequencing was performed at Novogene, China using the Illumina HiSeq 2000 platform to generate approximately 8 Gb of 150 bp paired-end reads per sample (mean = 7.9 Gb, s.d. = 1.2 Gb).

Profiling microbiome composition and function

Metagenomes were profiled consistent with previous data analysis of 1000IBD[54] and Lifelines-DEEP[41] cohorts, as follows. KneadData tools (v0.5.1)[55] were used to process metagenomic reads (in fastq format) by trimming the reads to PHRED quality 30 and removing Illumina adapters. Following trimming, the KneadData integrated Bowtie2 tool (v2.3.4.1)[56] was used to remove reads that aligned to the human genome (GRCh37/hg19).

Taxonomic composition of metagenomes was profiled by MetaPhlAn2 tool (v2.7.2)[57] using the MetaPhlAn database of marker genes mpa_v20_m200. Profiling of genes encoding microbial biochemical pathways was performed using the HUMAnN2 pipeline (v0.11.1)[58] integrated with the DIAMOND alignment tool (v0.8.22)[59], UniRef90 protein database (v0.1.1)[60] and ChocoPhlAn pan-genome database (v0.1.1)[58]. As a final quality control step, samples with unrealistic microbiome composition (eukaryotic or viral abundance >25% of total microbiome content or total read depth <10 million) were excluded, leaving 8,208 samples for further analyses. Analyses were performed using locally installed tools and databases on CentOS (release 6.9) on the high-performance computing infrastructure available at our institution and using the MOLGENIS data platform[61].

In total, we detected 1,253 taxa (4 kingdoms, 21 phyla, 35 classes, 62 orders, 128 families, 270 genera and 733 species) and 564 pathways in at least one of the samples in the quality-controlled dataset. To deal with sparse microbial data in the downstream analysis, we focused on bacterial and archaeal species/pathways with a mean relative abundance >0.01% that were present in at least 5% of participants. This yielded 257 taxa (6 phyla, 11 classes, 15 orders, 30 families, 59 genera and 136 species) and 277 pathways. Together, these microbial features accounted for 97.86% and 87.82% of the average taxonomic and functional compositions, respectively.

Based on the abundance profiles of the taxa that passed the filtering process, we calculated alpha diversity, as measured by richness and Shannon entropy, at family-, genus- and species-level using specnumber and the function diversity, respectively, in R package vegan (v.3.6.1)[62]. Rarefaction and extrapolation (R/E) sampling curves for estimation of total richness of species and genera in the population were constructed using a sample size–based interpolation/ extrapolation algorithm implemented in the iNEXT package for R[63].

Profiling of bacterial virulence factor and antibiotic resistance genes

Metagenomes were searched for bacterial virulence factors using the shortBRED toolkit (v0.9.5)[64] and the virulence factors of pathogenic bacteria (VFDB) core dataset of DNA sequences[65] (downloaded on 1 November 2018). The shortBRED tool shortbred_identify.py (v0.9.5) was used to identify unique markers for virulence factors, with the UniRef90 database (downloaded on 1 November 2018) used as negative control, and the shortbred_quantify.py tool (v0.9.5) was used to perform a quantification of these markers in metagenomes. Quantification of antibiotic resistance genes was performed using the shortBRED tool shortbred_quantify.py (v0.9.5), with markers generated using shortbred_identify.py (v0.9.5) on the CARD database of bacterial antibiotic resistance genes[66] (downloaded 1 November 2018), with the UniRef90 database used as negative control. This identified 190 virulence factors and 303 antibiotic resistance gene families, of which 47 virulence factors and 98 antibiotic resistance genes were present in at least 5% of participants with a relative abundance > 0.01%. These accounted for 95.22% of virulence factor composition and 98.08% of antibiotic resistance composition, respectively.

Estimation of heritability of microbiome

We estimated the heritability of bacterial taxa and pathways using linear mixed models. In particular, we fitted the following model using the function *relmatLmer* from the *lme4qtl* package (v.0.1.10)[67] for R:

$$Y \sim \text{age} + \text{age}^2 + \text{sex} + \text{read depth} + \text{stool frequency} + 1|\text{ID} + 1|\text{FAM} + 1|\text{cohousing}$$

where Y is the relative abundance of the bacterial taxon or pathway, transformed using the centred additive log-ratio (CLR) transformation, with the geometric mean calculated on the taxonomic level of species used as denominator[68]. Age, age^2, sex, read depth and stool frequency are participant-specific factors modelled as fixed-effect covariates, while the remaining three terms are random effects representing a polygenic additive effect (1|ID, equivalent to twice the kinship matrix), history of a familial shared environment (1|FAM, implemented as unique family identifier) and current cohabitation (1|cohousing, implemented as unique housing location identifier). We note that this model may still provide heritability estimates that are slightly inflated due to residual correlation between the polygenic additive effect and the cohabitation effect in our dataset, where 42% of participants share the same household. Narrow-sense heritability was estimated as the proportion of variance explained by the polygenic additive effect over total variance, using the profile function of the same R package. We restricted the analysis of heritability to the relative abundances of the 257 microbial taxa and 277 pathways present in at least 5% of individuals and focused on 3,571 individuals in 1,432 families in which at least two individuals had available microbiome data. In total, the analysis included 1,004 first-degree relative pairs, 210 second-degree relative pairs and 85 pairs with third-degree or more distant relationships.

The significance of random variables in heritability models (heritability, cohabitation and history of a familial shared environment) was assessed using a permutation test to determine empirical *P* value. For each microbiome feature, the link between the participant ID and relative abundance of microbial taxa was randomly permuted 30,000 times, while retaining the data structure of the rest of data to maintain families, cohabitation and other associations in the dataset. To maintain consistency, each microbiome feature was permuted using the set of identical randomization seeds. A heritability model was constructed for each permuted dataset, and the empirical *P* values for heritability, family ID and cohabitation were calculated as a proportion of the values of these variables higher than, or equal to, values of heritability, family ID or cohabitation in the model constructed from non-permuted data. Significance of fixed effects was calculated using type-II analysis of variance using Wald tests implemented in the anova function for R package car, while the significance of random effects was calculated using likelihood ratio tests implemented in the ranova function for R package lmerTest. Confidence intervals were estimated using the

function profile from the R stats package. As this approach could not estimate confidence intervals for taxa or pathway models for which one or more random effects estimate was approximately 0, we estimated the confidence intervals of these traits using reduced models without random variables with effect size $\leq 1.0 \times 10^{-3}$. All empirical *P* values were corrected for multiple testing using Benjamini–Hochberg correction, and results with an empirical FDR < 0.1 were considered significant.

Estimation of the effect of cohabitation on microbiome

We estimated the effect of cohabitation on overall microbiome composition, function, antibiotic resistance genes and virulence factors by comparing the beta-diversities of the microbiomes of cohabiting study participants (1,710 unrelated pairs, 285 parent–child pairs and 144 sibling pairs) to those of participants who did not share housing (2,000 unrelated pairs, 301 parent–child pairs and 299 sibling pairs). Microbiome distance was calculated for each pair using Bray–Curtis dissimilarity, and mean dissimilarities within groups were compared using Mann-Whitney U tests using the Benjamini–Hochberg correction to control multiple testing FDR. Results were considered significant at FDR < 0.05.

Calculation of microbiome–phenotype associations

The proportion of variance in microbiome composition that can be explained by individual phenotypes was calculated by permutational multivariate analysis of variance using distance matrices (adonis) implemented in the adonis function of R package vegan (v.2.4–6)[62]. Analysis was performed on the microbiome beta-diversity (Bray–Curtis distance matrix calculated using relative abundances of microbial species) and separately for each phenotype using univariate adonis with 20,000 permutations. To calculate the proportion of microbiome functional potential explained by individual phenotypes, an equivalent analysis was performed on the Bray–Curtis distance matrix calculated using relative abundances of MetaCyc microbial biochemical pathways.

The total proportion of microbiome composition variance and function explained by groups of phenotypes was calculated by multivariate adonis analyses. These analyses included all phenotypes that showed significant (FDR < 0.05) association with microbiome composition or function in the univariate analyses.

The associations between each individual microbiome feature (Shannon alpha diversity index, microbial taxa, MetaCyc pathways, virulence factor and antibiotic resistance gene families) and each phenotype were calculated using linear regression. To correct for potential batch effects and confounders, the regression model also included age, sex, BMI, Bristol stool scale and technical factors (DNA concentration, sequencing read depth, sequencing batch and sampling season). The microbiome data was transformed using the CLR transformation to break compositionality of the data and normalize skewed distributions of microbiome features.

All microbial taxa, regardless of taxonomic level, were CLR-transformed using the geometric mean of the relative abundance of microbial species as the CLR denominator. The other microbiome data layers (pathways, virulence factor and antibiotic resistance genes) used geometric mean of relative abundances of these features as the CLR denominator. As CLR transformation cannot be applied to zero values, zeros in the tables were adjusted by adding half of the lowest value in the table to each cell. Benjamini-Hochberg correction was used to control for multiple testing, with the number of tests equal to the number of feature–phenotype pairs tested (64,764 tests for taxa, 87,444 for pathways, 111,889 for VFDB and 24,444 for CARD). Results were considered significant at FDR < 0.05.

Quantification of microbiome–disease–drug interactions

To disentangle interactions between the gut microbiome, medication and diseases, we explored the effect of a selection of drug–disease pairs for three common diseases and the drugs used to treat them: functional gastrointestinal disorders PPIs (*N* = 545 (disease only), 375 (drug only) and 108 (overlap)), T2D with antidiabetic drugs (*N* = 83 (disease only), 10 (drug only) and 98 (overlap)) and depression with SSRIs (*N* = 268 (disease only), 158 (drug only) and 78 (overlap)). For each disease–drug pair, we used multivariate linear regression including the drug and disease as independent variables and microbiome traits as outcomes to calculate the conditional effects of drugs and diseases on the microbial traits. Details of the models and approaches used are described in 'Calculation of microbiome–phenotype associations'.

Definition of core microbiome and prediction of keystone microbiome features

We used a bootstrapping-based selection approach to identify the expected number of microbial traits in the DMP. We created randomized subsamples with a size of 1% to 100% of the cohort, one hundred times for each subsampling size, and calculated mean and standard deviation of the number of microbial features for each subsampling level. The expected cohort richness was estimated using the specpool function in vegan package v.2.5.7 using the bootstrap method. Microbial features with a prevalence >95% in the cohort were defined as the core microbiome, which included 9 core microbial species and 143 core microbial pathways (Supplementary Table 1a, b). To validate the consistency of our approach, we compared the results to previously published studies that defined the core microbiota in UK, US, European and non-Western populations[5,20,69–71].

To analyse the structure of the microbiome community, we constructed microbial species and pathway co-abundance networks using SparCC, as previously published[72,73]. Relative abundances of taxa were converted to estimated read counts by multiplying abundance percentages by total sequenced reads per sample after quality control. For pathway analysis, the read counts (RPKM) from HUMAnN2 were used directly for SparCC.

Co-abundances were deemed significant at an empirical FDR < 0.05, calculated using 100 permutations. In each permutation, the abundance of each microbial feature was randomly shuffled across samples. This identified 6,473 species and 55,407 pathway co-abundances at empirical FDR < 0.05. Features that ranked in the top 20% in the number of network connections (node degree) were considered keystone species or pathways, resulting in 28 keystone species and 53 keystone pathways (Supplementary Table 1e).

Identification of microbiome clusters

To identify microbial clusters and assess the presence of gut enterotypes in our cohort, we performed the partitioning around the medoid method on the relative abundances of microbial species and used the Calinski–Harabasz index to select the optimal number of clusters, as previously published in a study of gut enterotypes[74]. Enrichment of phenotypes in each cluster was assessed by logistic regression in R.

Calculation of microbiome signatures predictive of diseases and health

We calculated the microbial signatures predictive of the 36 most common (number of cases > 100) diseases in our dataset. In addition, we defined a healthy phenotype as absence of any self-reported disease, with 2,937 (36%) out of 8,208 individuals defined as healthy.

We used fivefold cross-validation to train prediction models for common diseases, with 4 out of 5 of the data used as a training set and 1 out of 5 as a test set. Next, we performed elastic net L1/L2 regularized regression (R package glmnet v.4.0) on the training set, using Shannon diversity, CLR-transformed microbial taxa, CLR-transformed MetaCyc bacterial pathways and age, sex and BMI as fixed covariates. Within each fold, the model for each disease was calculated independently using nested tenfold cross-validation to select the optimal lambda penalization factor (at L1/L2 mixing parameter alpha fixed at 0.5). The lambda with minimal cross-validation error was used in the downstream analysis.

In total, we defined three probabilistic models: a 'null' signature that only includes effects of general covariates (age, sex and BMI), a 'microbiome' signature that includes all selected microbiome features and a 'combined' signature that includes both the effects of microbiome features and general covariates. Correlations of predictive signatures of diseases were calculated as Pearson correlation of predicted values for diseases from the test set of each fold, while disease comorbidities were calculated as Pearson correlations of the presence of diseases in the population (encoded 0 for controls and 1 for presence of disease).

Reporting summary

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

Data availability

The raw microbiome sequencing data, processed microbiome data (including taxonomy, pathway, virulence factor and antibiotic resistance gene profiles) and basic phenotypes (including age, sex and BMI) used in this study are available at the European Genome-Phenome Archive under accession EGAS00001005027. These datasets can be accessed from https://forms.gle/eHeBdXJMXbVvCJRc8 or by email from the corresponding author (R.K.W.) at the address listed at the EGA data access committee EGAC00001001996. The phenotype data can be requested, for a fee, by filling the application form at https://www.lifelines.nl/researcher/how-to-apply/apply-here. Lifelines will not charge an access fee for controlled access to the full dataset used in the manuscript (including phenotype and sequencing data), for the specific purpose of replication of the results presented in this Article or for further assessment by the reviewers, for a period of three months. Researchers interested in such a replication study or review assessment can contact Lifelines at research@lifelines.nl. Source data are provided with this paper.

Code availability

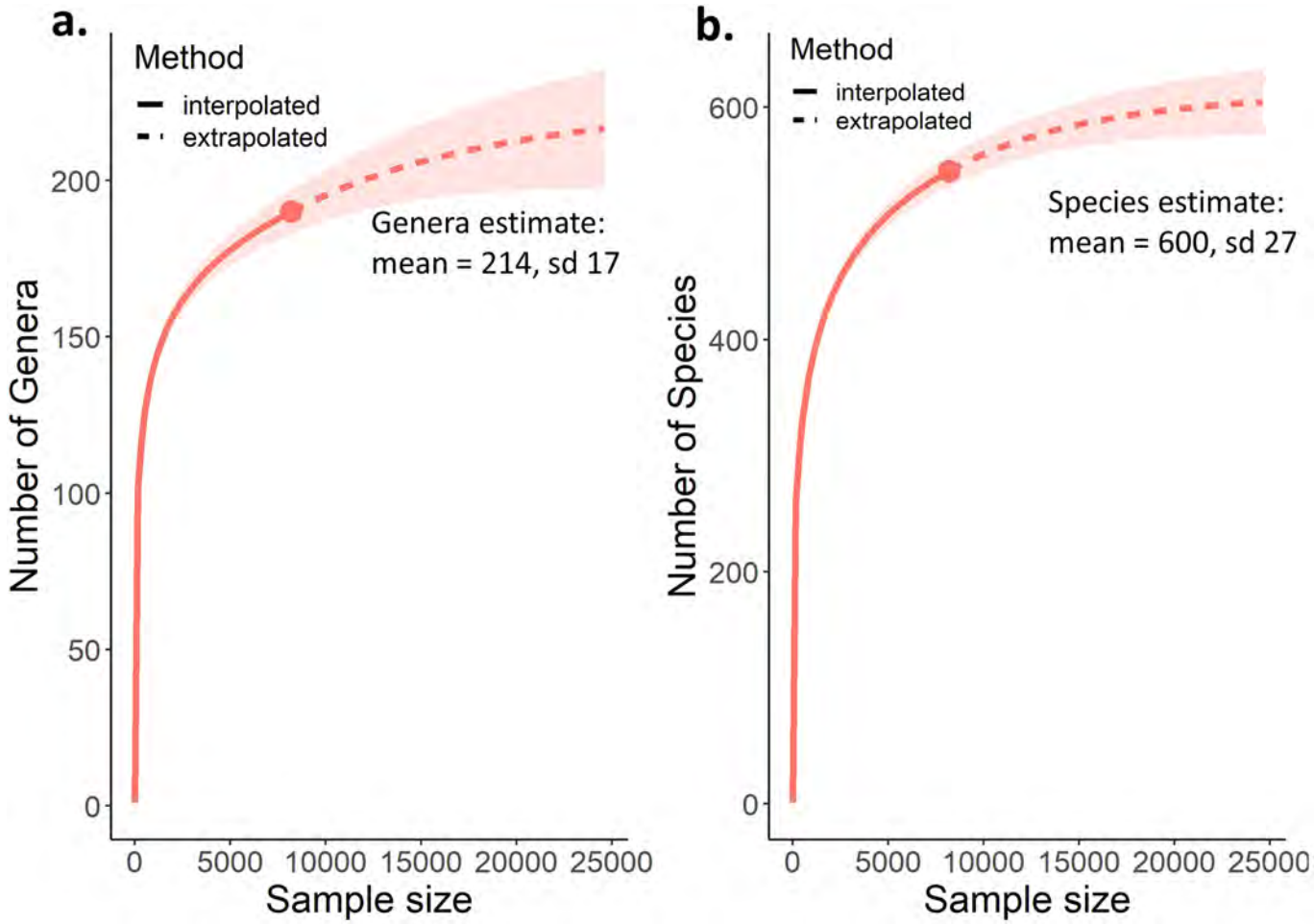
Open source codes and scripts used for the analyses or figures are available at the GitHub repository (https://github.com/GRONINGEN-MICROBIOME-CENTRE/DMP) and Zenodo (https://doi.org/10.5281/zenodo.5910709). To facilitate the re-use of the codes, the repository also includes example datasets that enable users to test the codes without the need to apply for access to phenotypes.

Acknowledgements We acknowledge and thank the late M. Hofker who initiated the Lifelines DAG3/Dutch Microbiome Project. We acknowledge the services of Lifelines Cohort Study, the contributing research centres delivering data to Lifelines and all the study participants. The Lifelines Biobank initiative has been made possible by subsidies from the Dutch Ministry of Health, Welfare and Sport, the Dutch Ministry of Economic Affairs, the University Medical Center Groningen (UMCG the Netherlands), the University of Groningen and the Northern Provinces of the Netherlands. We thank the Center for Information Technology of the University of Groningen (RUG) for their support and for providing access to the Peregrine high performance computing cluster, the Genomics Coordination Center (UMCG and RUG) for their support and for providing access to Calculon and Boxy high-performance computing clusters and the MOLGENIS team for data management and analysis support. Metagenomics library preparation and sequencing was done at Novogene. We also thank K. Mc Intyre for English and content editing and Tania Ballve Fernandez for illustration of Fig. 1a. Sequencing of the cohort was funded by a grant from CardioVasculair Onderzoek Nederland (CVON 2012-03) to M.H., J.F. and A.Z. R.G., H.J.M.H. and R.K.W. are supported by the collaborative TIMID project (LSHM18057-SGF) financed by the PPP allowance made available by Top Sector Life Sciences & Health to Samenwerkende Gezondheidsfondsen (SGF) to stimulate public–private partnerships and co-financed by health foundations that are part of the SGF. R.K.W. is supported by the Seerave Foundation and the Dutch Digestive Foundation (16-14). A.Z. is supported by European Research Council (ERC) Starting Grant 715772, Netherlands Organization for Scientific Research (NWO) VIDI grant 016.178.056, CVON grant 2018-27 and NWO Gravitation grant ExposomeNL 024.004.017. JF is supported by the Dutch Heart Foundation IN-CONTROL (CVON2018-27), the ERC Consolidator grant (grant agreement No. 101001678), NWO-VICI grant VI.C.202.022, and the Netherlands Organ-on-Chip Initiative, an NWO Gravitation project (024.003.001) funded by the Ministry of Education, Culture and Science of the government of The Netherlands. C.W. is further supported by an ERC advanced grant (ERC-671274) and an NWO Spinoza award (NWO SPI 92-266). L.C. is supported by a joint fellowship from the University Medical Center Groningen and China Scholarship Council (CSC201708320268) and a Foundation De Cock-Hadders grant (20:20-13). M.A.S. is supported by NWO VIDI grant 016 and EUCAN-connect, a project funded by European Commission H2020 grant 824989.

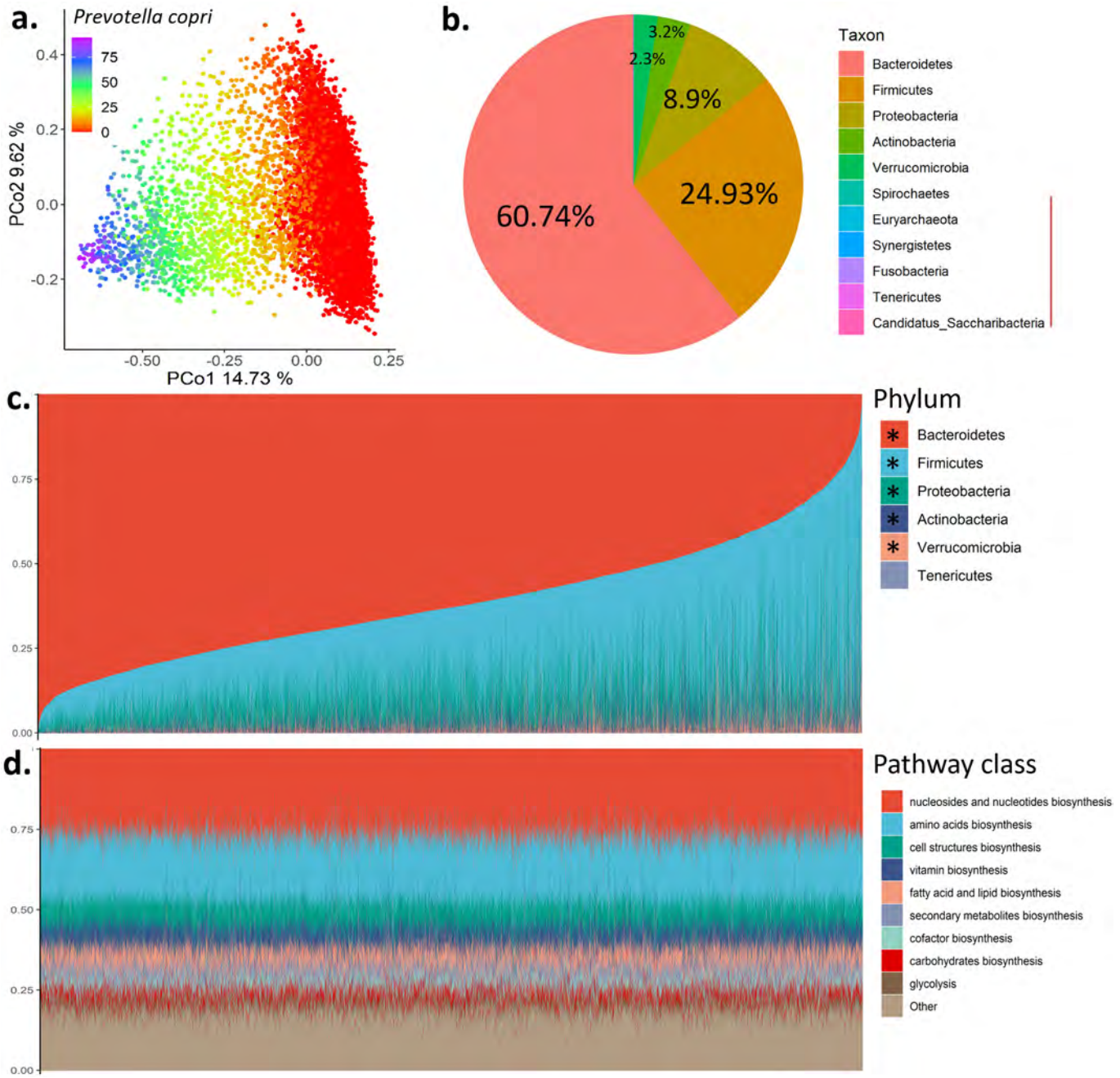
Author contributions R.G. designed and implemented the metagenomic data analysis pipelines, analysed metagenomic data, performed heritability analysis and drafted the manuscript. A.K. designed the prediction models and implemented statistical methods for association analyses and assisted in drafting of the manuscript. A.V.V., L.C., V.C., S.H., M.A.Y.K., S.A.-S., J.R.B., L.A.B., V.C.L., T.S., M.H., J.C.S. and S.S. assisted in other statistical analyses, interpretation of data and drafting of the manuscript. M.A.S. provided data stewardship and analysis infrastructure. B.H.J., J.A.M.D., S.J. and J.G.-A. collected data, assisted in study planning and critically reviewed the manuscript. S.S. supervised and coordinated heritability analysis. R.C.H.V. provided the air pollution data and supervised the air pollution analysis. H.J.M.H., A.Z., R.K.W., J.F. and C.W. conceived, coordinated and supported the study. All authors critically revised and approved the manuscript.

Competing interests The authors declare no competing interests.

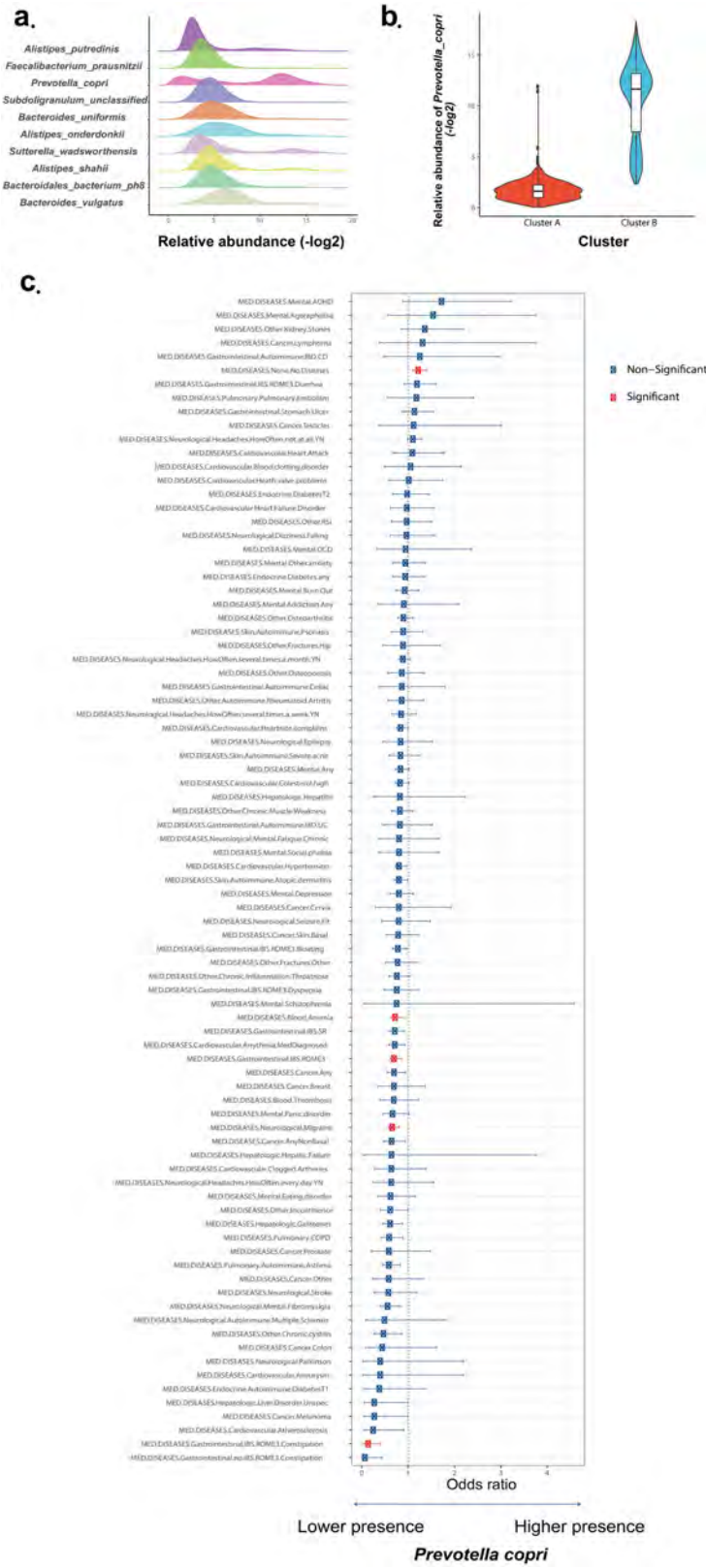
Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-022-04567-7.
Correspondence and requests for materials should be addressed to J. Fu, A. Zhernakova or R. K. Weersma.
Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer review reports are available.



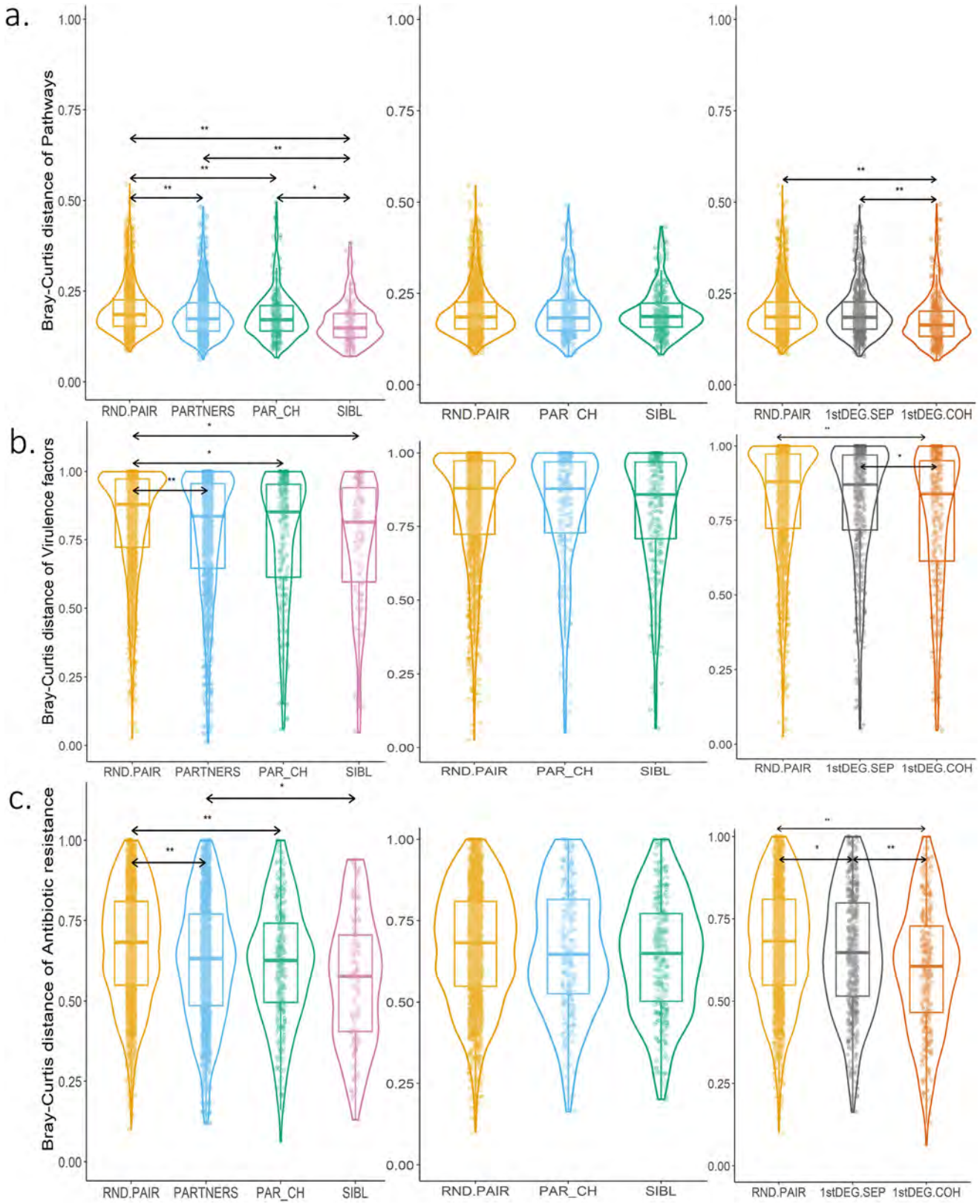
Extended Data Fig. 1 | Estimation of total number of species and genera in the DMP population. Figure shows rarefaction and extrapolation sampling curve for **a**, genera and **b**, species richness calculated using Hill numbers implemented in the iNEXT package for R. The extrapolated part of rarefaction curve is shown dotted. The SD of the estimate is shaded and the asymptotic richness estimate is shown.



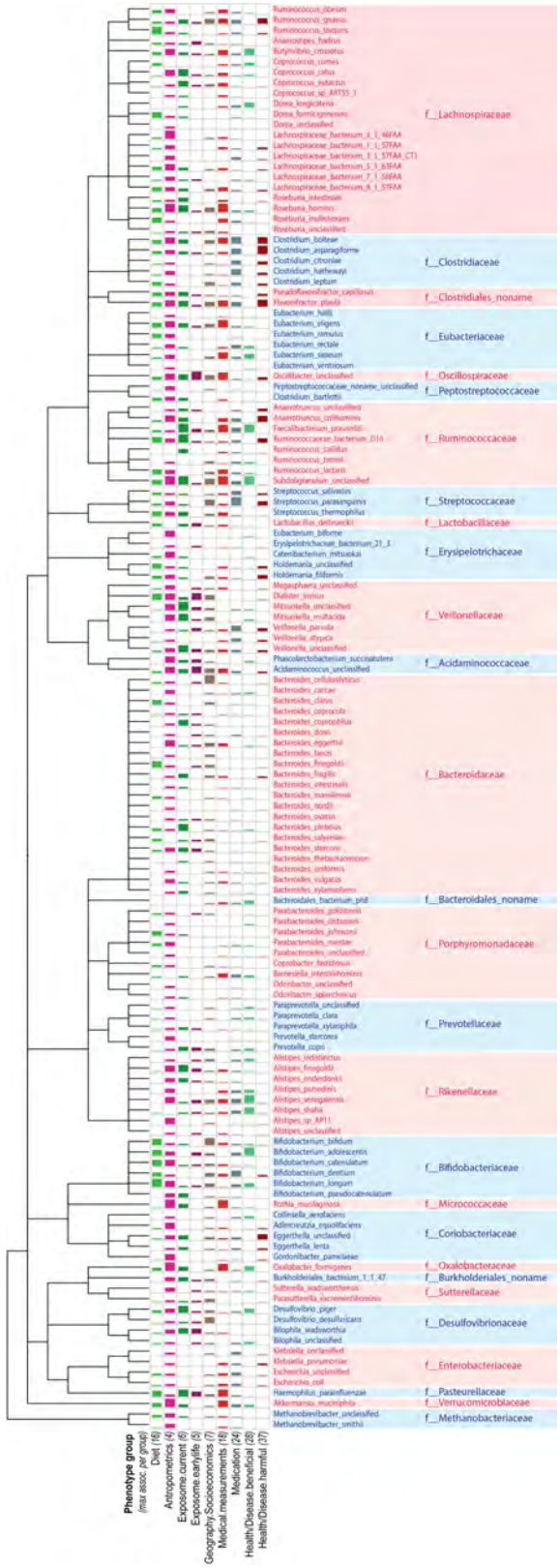
Extended Data Fig. 2 | Overview of DMP microbiome composition and function. a, First two principal coordinates of the Bray-Curtis distance matrix calculated on microbial species of the DMP cohort, coloured by the relative abundance of *Prevotella copri* bacterium. **b**, Average relative abundances of bacterial phyla present in > 0.1% of the DMP cohort. Red vertical line indicates rare phyla (abundance < 0.1%). **c**, Phylum-level composition of all samples in the cohort, sorted by abundance of phylum Bacteroidetes. Each vertical line indicates one sample. * phylum has significantly higher variance when compared to each of pathway classes (one-sided F test of variances FDRs < 0.05, Supplementary Table 1G) **d**, Relative abundances of the top 10 MetaCyc pathways of all samples (sorted to match panel c). Each vertical line indicates one sample. The means of standard deviations of taxa and pathways were found to be significantly different (mean(sd(tax_1),...,sd(tax_n)) - mean(sd(pwy_1),..., sd(pwn_m) > 0, two-sided permutation test (1,000 permutations) $P < 1.0 \times 10^{-3}$). All panels show results generated from n = 8,208 independent samples.



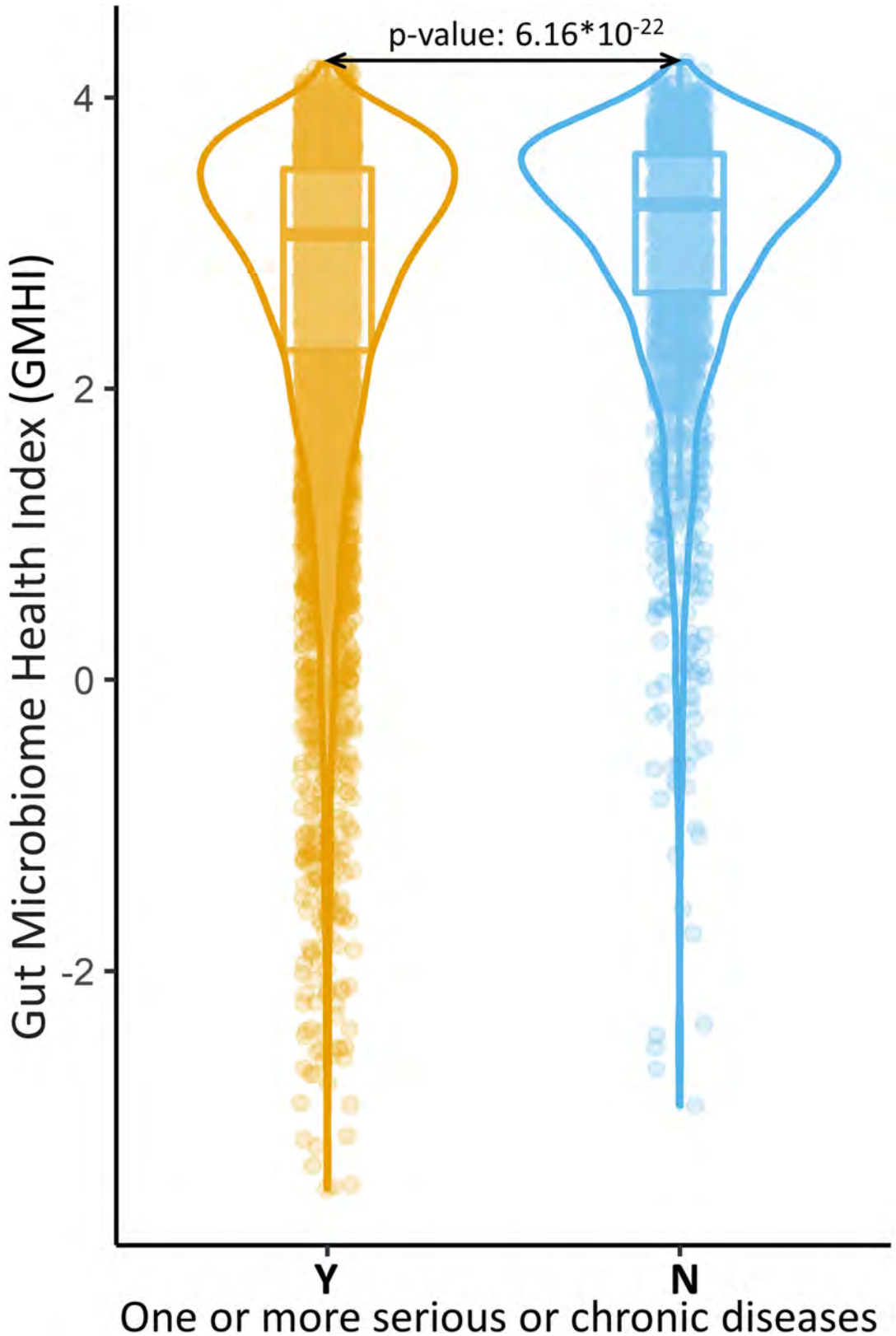
Extended Data Fig. 3 | Clusters determined by bi-modally distributed *Prevotella copri*. **a**, Density plots of log_2-transformed relative abundances of the 10 most abundant bacterial species. **b**, Log_2-transformed relative abundance of *Prevotella copri* per microbiome cluster (n (cluster1, red) = 6,346, n (cluster2, blue) = 1,862; boxplot: centre line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers; outer line, distribution of data). The clusters were determined using the partitioning around the medoid method on the relative abundances of microbial species. **c**. Association of *P. copri* with metadata (n = 8,208 independent samples; dot, mean; lines, 95% confidence intervals).



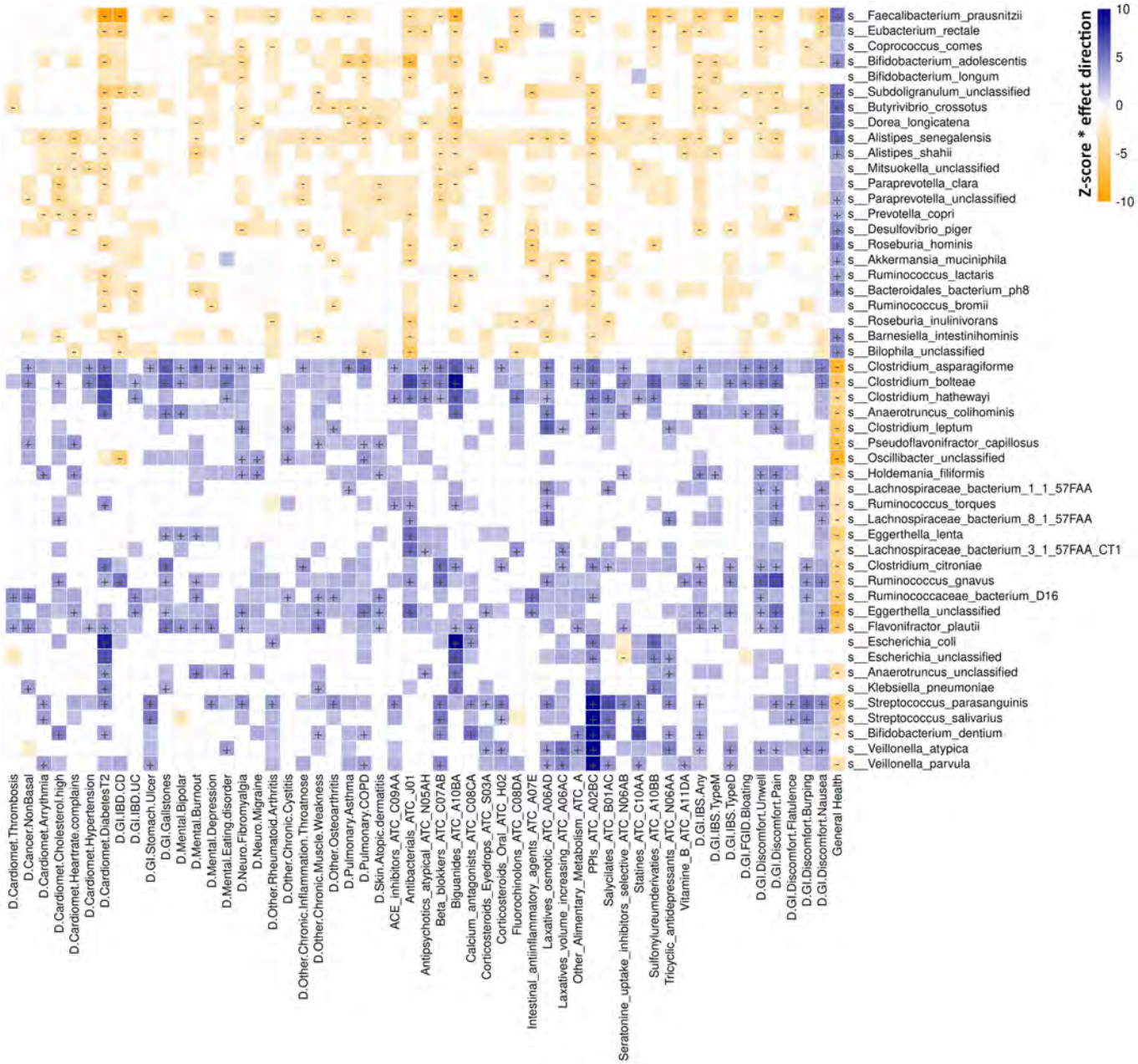
Extended Data Fig. 4 | Bray-Curtis distances of microbiome features of cohabiting and non-cohabiting participants. Pairwise microbiome Bray-Curtis dissimilarity comparisons of groups of random, non-cohabiting pairs (RND.PAIR, n = 2,000) compared to cohabitating partners (PARTNERS, n = 1,710); cohabiting parent–child pairs (PAR_CH, n = 285) and cohabiting siblings (SIBL, n = 144); and random pairs (n = 2,000) compared to non-cohabiting 1st-degree relatives (1stDEG.SEP, n = 600) and cohabiting 1st-degree relatives (1stDEG.COH, n = 429). **a**, MetaCyc pathways. **b**, Virulence factor gene families. **c**, antibiotic resistance gene families (centre line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers; outer line: distribution of data). Significantly different groups are marked with ** for FDR < 1.0e−5 or * for FDR < 0.05 (two-sided Wilcoxon test).



Extended Data Fig. 5 | Overview of microbiome–phenotype associations. Figure shows the number of study-wide significant associations (FDR < 0.05) per phenotype group, clustered by taxonomy. Bar heights represent the number of associations relative to the maximal number of associations for the phenotype group.

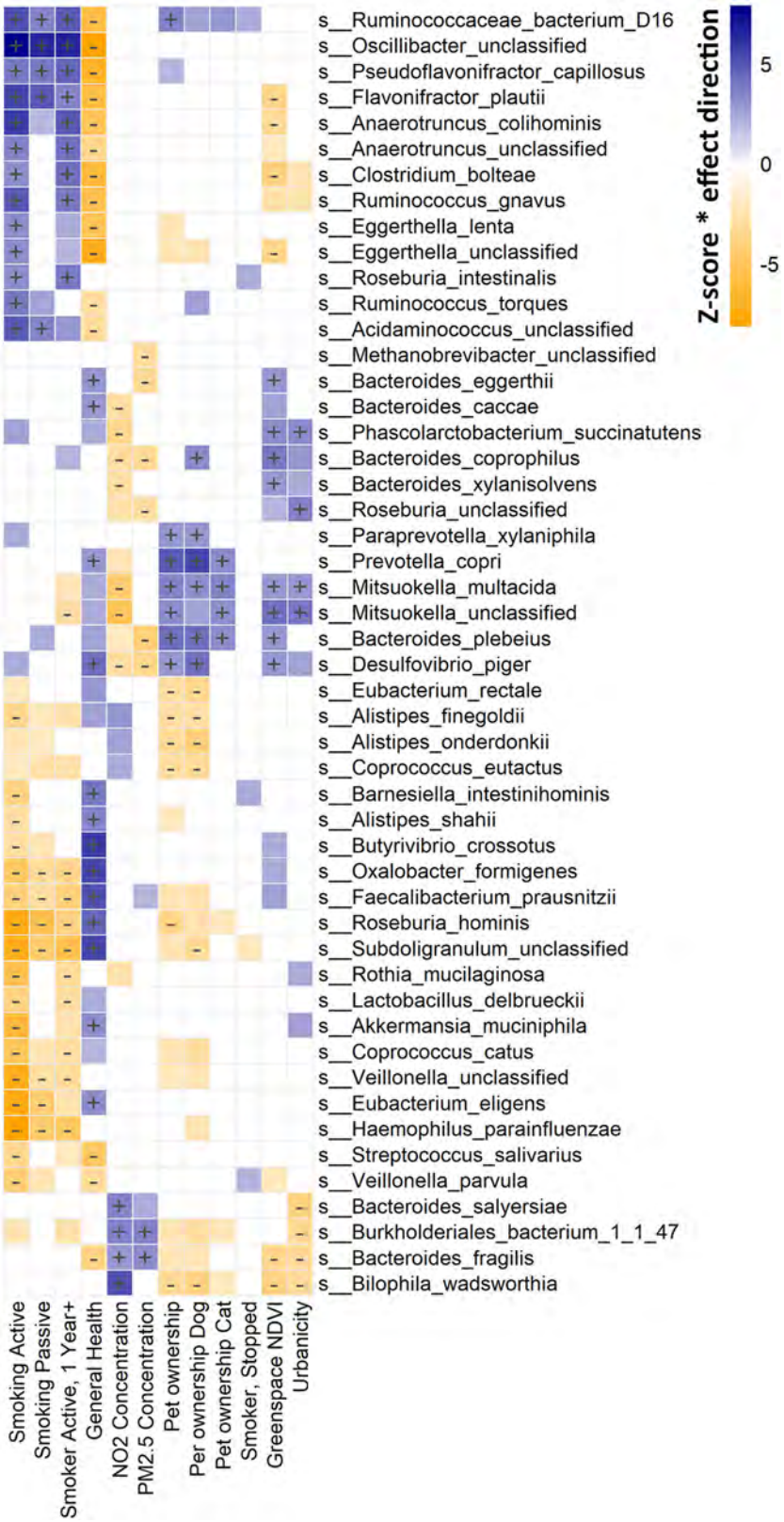


Extended Data Fig. 6 | Gut Microbiome Health Index calculated for DMP cohort. Box-plots of the Gut Microbiome Health Index (GMHI) for healthy participants of the DMP cohort samples (Y, n = 1,876 independent participants) vs participants who reported one or more diseases (N, n = 6,332 independent participants) (centre line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers; outer line, distribution of data). P-value is shown for two-sided Wilcoxon rank-sum test.

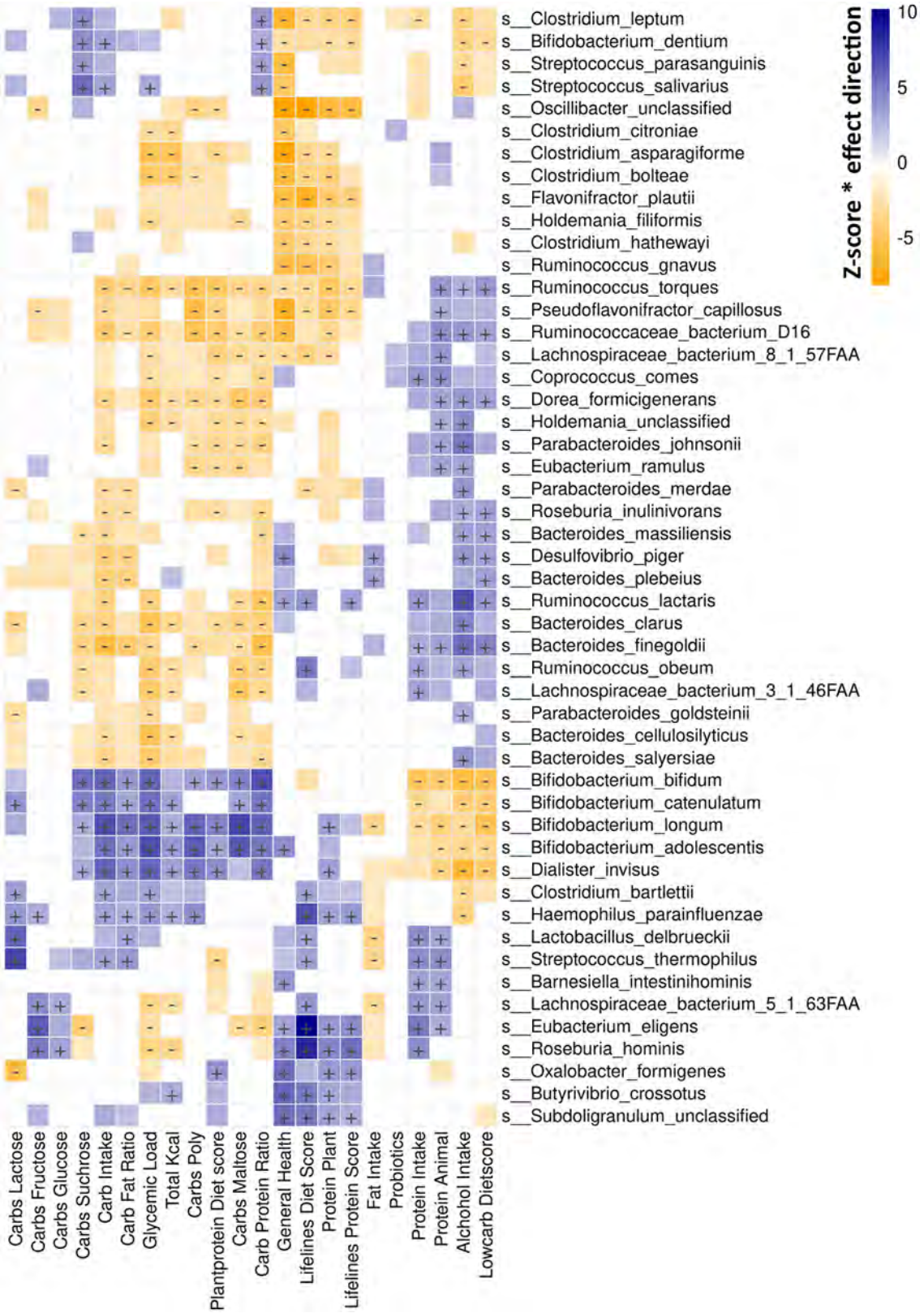


Extended Data Fig. 7 | Microbiome associations with diseases and medication use. Heatmap of microbiome–phenotype associations, with microbial species clustered by Z scores (multivariate linear regression of CLR-transformed relative abundance of taxa, correcting for age, Sex, BMI, Bristol stool scale of the faecal sample and technical factors (DNA concentration, sequencing read depth, sequencing batch and sampling season)) using hierarchical clustering and coloured by the direction of association. Study-wide significant associations (Benjamini-Hochberg corrected p-value < 0.05) are marked with +/−. Coloured associations without a label indicate nominally significant associations (Benjamini-Hochberg corrected p-value < 0.05).

Extended Data Fig. 8 | Microbiome association with early-life exposures. Heatmap of microbiome–phenotype associations, with microbial species clustered by Z scores (multivariate linear regression of CLR-transformed relative abundance of taxa, correcting for age, Sex, BMI, Bristol stool scale of the faecal sample and technical factors (DNA concentration, sequencing read depth, sequencing batch and sampling season)) using hierarchical clustering and coloured by the direction of association. Study-wide significant associations (Benjamini-Hochberg corrected p-value < 0.05) are marked with +/−. Coloured associations without a label indicate nominally significant associations (Benjamini-Hochberg corrected p-value < 0.05).



Extended Data Fig. 9 | Microbiome association with smoking, pollutants and greenspace. Heatmap of microbiome–phenotype associations, with microbial species clustered by Z scores (multivariate linear regression of CLR-transformed relative abundance of taxa, correcting for age, Sex, BMI, Bristol stool scale of the faecal sample and technical factors (DNA concentration, sequencing read depth, sequencing batch and sampling season)) using hierarchical clustering and coloured by the direction of association. Study-wide significant associations (Benjamini-Hochberg corrected p-value < 0.05) are marked with +/−. Coloured associations without a label indicate nominally significant associations (Benjamini-Hochberg corrected p-value < 0.05).



Extended Data Fig. 10 | Microbiome association with diet. Heatmap of microbiome–phenotype associations, with microbial species clustered by Z scores (multivariate linear regression of CLR-transformed relative abundance of taxa, correcting for age, Sex, BMI, Bristol stool scale of the faecal sample and technical factors (DNA concentration, sequencing read depth, sequencing batch and sampling season)) using hierarchical clustering and coloured by the direction of association. Study-wide significant associations (Benjamini-Hochberg corrected p-value < 0.05) are marked with +/−. Coloured associations without a label indicate nominally significant associations (Benjamini-Hochberg corrected p-value < 0.05).

nature portfolio

Corresponding author(s): R. K. Weersma

Last updated by author(s): Feb 4, 2022

Reporting Summary

Nature Portfolio 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 Portfolio 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	For full details see Methods; KneadData tools (v0.5.1), Bowtie2 tool (v2.3.4.1), MetaPhlAn2 tool (v2.7.2), HUMAnN2 (v0.11.1), DIAMOND alignment tool (v0.8.22), shortBRED toolkit (v0.9.5) were used to process the microbiome sequencing data.
Data analysis	Custom codes and scripts are avaliable at GitHub: https://github.com/GRONINGEN-MICROBIOME-CENTRE/DMP and Zonedo: https://doi.org/10.5281/zenodo.5910709. Other software used: Microsoft Excel v 14.0.4734, R version 3.6.0 (2019-04-26), R packages: vegan (v.2.5-6), iNEXT (v.2.0.20), lme4qtl (v.0.1.10), car (v.3.0-12), lmerTest (v.3.1-3), stats (v.3.6.0), glmnet (v.4.0)

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 Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Access to raw microbiome sequencing data and basic phenotypes:
The raw microbiome sequencing data, processed microbiome data (including taxonomy, pathway, VF and antibiotic resistance gene profiles) and basic phenotypes (including age, sex and BMI) used in this study are available at the European Genome-Phenome Archive (EGA, https://ega-archive.org/) as EGA study

EGAS00001005027 (https://ega-archive.org/studies/EGAS00001005027). These datasets require a minimal access procedure (data access form at https://forms.gle/eHeBdXJMXbVvCJRc8 or email request to corresponding author at address listed on EGA data access committee EGAC00001001996: https://ega-archive.org/dacs/EGAC00001001996) to ensure that the data is being requested for research/scientific purposes only and thus complies with the informed consent signed by Lifelines participants, which specifies that the collected data will not be used for commercial purposes. Submitted data access forms will be evaluated by the data access committee and Lifelines, and a response to requests will be given within two weeks. For requests from verified academic parties, access will be given without further delay. For requests from commercial parties, Lifelines will perform a pre-DPIA (Data Privacy Impact Assessment) to assess the risks of the proposed processing of personal data (e.g. purpose, storage, access, archiving, etc.) with respect to the GDPR (EU privacy laws) subject rights. Based on the outcome of the pre-DPIA, Lifelines will decide whether sharing data with the commercial entity is allowed and/or whether additional measures have to be taken.

Access to other phenotype data:
To ensure adherence to participant's privacy and informed consent, the rights of participants as described in the GDPR and Lifelines biobank regulations, the complete phenotype data cannot be provided open-access and is available from the Lifelines under controlled-access in a secure Lifelines Workspace or High Performance Cluster (HPC) environment. As Lifelines is a non-profit organization dependent on (governmental) subsidies, a fee is required to cover the costs of controlled data access and supporting infrastructure.
In brief, the step-by-step data access procedure is as follows: 1) Data is requested by filling the application form to request "Available Lifelines-data" at https://www.lifelines.nl/researcher/how-to-apply/apply-here, 2) Lifelines will evaluate project proposals to ensure compliance with the Lifelines data access policy, informed consent of Lifelines participants and the GDPR and that the data is being requested for non-commercial research, 3) upon approval, Lifelines will send Data and Material Transfer Agreement (DMTA) contracts to the applicants and 4) after the required contracts are signed, Lifelines will provide access to data via the Workspace or HPC and link the raw and processed DMP sequencing data to the Lifelines phenotypes. Lifelines strives to accomplish steps 2–4 at 2-weeks per step, assuming that no extra actions by the applicant or Lifelines are required.
The fee for data access on the HPC is €3,500 for one year and the fee for the Lifelines Workspace environment is €4,500 for one year, or less for shorter periods of time. There are no restrictions on downstream re-use of aggregated, non-identifiable results (as approved by Lifelines), nor are there authorship requirements, but Lifelines does request that it is acknowledged in publications using these data. The data access policy and data access fees are described in detail at https://www.lifelines.nl/researcher/how-to-apply. Further information and an example DMTA (which includes details on how to acknowledge the use of Lifelines data in publications) can be obtained from Lifelines at https://www.lifelines.nl/researcher/how-to-apply/information-request or by contacting Lifelines at research@lifelines.nl. Finally, Lifelines will not charge an access fee for controlled-access to the full dataset used in the manuscript (including phenotype and sequencing data) for a period of 3 months, for the specific purpose of replication of the results presented in the current manuscript or for further assessment by the reviewers. Researchers interested in such a replication study or review assessment can contact Lifelines at research@lifelines.nl.

Other supporting data: Source data for Figures 1-4 are provided with the paper. The authors declare that all other data supporting the findings of this study are available within the paper and its supplementary information files.

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	Sample size calculation was not performed before the study because of unknown effect sizes of many of studied factors. Instead the study focused on obtaining the largest possible sample size to capture microbiome variation in the population. As previous studies have identified significant microbiome-disease associations with 50-100 samples, our sample size was deemed sufficient to capture significant associations with moderately rare traits (prevalence >=1%).
Data exclusions	Out of 8,719 samples, 8,534 samples were successfully sequenced. Sequenced samples with with eukaryotic or viral abundance > 25% of total microbiome content or total read depth < 10 million were excluded, and 8,208 samples were retained for statistical analyses.
Replication	As this was a hypothesis generating study, no explicit replication attempts were made, but we compared our results to previously published studies where possible. These comparisons are described in details in the manuscript and include: 1) core gut microbiome species identified in our study were found to be highly consistent with those found in previous studies of UK, US and European and non-western populations (https://pubmed.ncbi.nlm.nih.gov/27126040/, https://www.nature.com/articles/nature06244/, https://pubmed.ncbi.nlm.nih.gov/22699611/, https://pubmed.ncbi.nlm.nih.gov/20203603/, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4255478/) (Supplementary Table 1a). 2) We compared results of our microbiome heritability analysis to previously published data (https://pubmed.ncbi.nlm.nih.gov/27173935/, https://pubmed.ncbi.nlm.nih.gov/27694960/) and found moderate consistency (heritability of Genera Akkermansia, Bifidobacterium, Bacteroides and Parabacteroides, as well as some of the genera from order Clostridiales), but did not replicate certain previous results (family Christensenellaceae, genus Turicibacter), possibly due to differences in cohort ethnicities, experimetnal methodology (we used metagenomic sequencing as opposed to previous studies using 16S sequencing) and used software and databases (e.g. Christensenellaceae was not present in Metaphlan2 database we used for classification and showed very low prevalence (< 1%) in our cohort when using Metaphlan3 database. 3) As described in the manuscript, we performed 5-fold cross-validation of disease-prediction models on our data as internal replication / test of data consistency. Cross-validation found that results are highly consistent across the folds (for example, prediction of no-disease status had AUC of 0.572 with standard error = 0.005); 4) We calculated Gut Microbiome Health Index (GMHI, https://pubmed.ncbi.nlm.nih.gov/32934239/) using our data and found high consistency with previously published results, with 43/50 GMHI signals replicating across the two studies at genus- or species-level (Supplementary Table 10).

Randomization	Study did not include any interventions and thus the randomization (as used in clinical trials or intervention studies) was not appropriate for this study. However, samples were randomly assigned to batches for all procedures which requires batch processing (DNA isolation, metagenomic sequencing and data preprocessing). As such, we do not anticipate any bias which might be caused by the potential lack of randomization.
Blinding	Study did not include any interventions and thus the conventional blinding (as used in clinical trials or intervention studies) was not appropriate for this study. However, the experimental procedures (DNA isolation, metagenomic sequencing) were in-effect blinded by samples being given group-unrelated identifiers and barcodes and by staff performing these procedures being unaware of sample metadata. As the data processing was performed computationally using standardized pipeline, it was effectively blinded as identical settings and software were used to process all data. The only unblinded part of the study was initial metadata collection by study data collection personnel, who due to the design of the study (data collection from medical records and self-reporting of data by study participants) cannot be blinded. We do anticipate any resulting bias which might be caused by the lack of blinding.

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

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

Methods

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	Lifelines is a multi-disciplinary prospective population-based cohort study using a unique three-generation design to examine the health and health-related behaviours of people living in the North of the Netherlands. The participants age-range was 8 - 84 years, 57.4% were female and 4,745 individuals clustered into 2,756 families.The participants were largely (99.5%) of Dutch European ancestry. A total of 241 host and environmental factors, including physical and mental health, medication use, diet, socioeconomic factors and childhood and current exposome were collected. These variables are described in detail, including summary statistics, in Methods section and Supplementary tables.
Recruitment	Between 2006 and 2013, inhabitants of the northern part of the Netherlands and their families were invited to participate in the Lifelines cohort study. Initially, eligible participants between 25 and 50 years of age were recruited through their general practitioner, and these individuals were then asked to indicate whether their family members (parents, partner, children, parents-in-law) would also be willing to participate in the study, if so they were sent an invitation to participate. In addition, other interested individuals could self-register as participants via the website during a limited period of time. This approach has resulted in a three-generation cohort of over 167,000 participants. DMP cohort was formed from volunteers from Lifelines cohort study who donated fecal samples.
Ethics oversight	The Lifelines study was approved by the medical ethical committee from the University Medical Center Groningen (METc number: 2017/152). Additional written consent was signed by all DMP participants or their parents or legal representatives (for children aged under 18).

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