Chapter 6

Effect of host genetics on the gut microbiome in 7,738 participants of the Dutch Microbiome Project


§ shared first authors
* shared last authors

Nature Genetics, in press
Abstract

Host genetics are known to influence the gut microbiome, yet their role remains poorly understood. To robustly characterize these effects, we performed a genome-wide association study of 207 taxa and 205 pathways representing microbial composition and function within the Dutch Microbiome Project, a population cohort of 7,738 individuals from the northern Netherlands. Two robust, study-wide significant ($p<1.89 \times 10^{-10}$) signals near the LCT and ABO genes were found to be associated with multiple microbial taxa and pathways and were replicated in two independent cohorts. The LCT locus associations seemed modulated by lactose intake, while those at ABO could be explained by participant secretor status determined by their FUT2 genotype. Twenty-two other loci showed suggestive evidence ($p<5 \times 10^{-8}$) of association with microbial taxa and pathways. At a more lenient threshold, the number of loci we identified strongly correlated with trait heritability, suggesting that much larger sample sizes are needed to elucidate the remaining effects of host genetics on the gut microbiome.

Introduction

The human intestinal microbial community contains trillions of microorganisms that play an important role in maintaining normal gut function and immune homeostasis. Emerging evidence has shown that alterations in gut microbial composition are associated with the pathogenesis of many human diseases, including gastrointestinal disorders, metabolic syndrome, cardiovascular diseases and other conditions.

The gut microbiome is influenced by many factors, such as environmental factors including diet and medication usage, but also by host genetics. Heritability studies in twins and families have highlighted that human genetics contributes to gut microbial variation, showing heritability estimates ranging from 1.9% to 8.1%. This observation drove the first efforts to identify genomic loci that influence gut microbiota through genome-wide association studies (GWAS). These early gut microbiome GWAS identified several microbial quantitative trait loci (mbQTL) located in genes related to the intestinal mucosal barrier, immune response and drug and food metabolism. However, the reproducibility of these findings has been limited by differences in data processing methodologies, modest sample sizes and strong environmental effects, which, taken together, limit the detection of robust host genetic associations.

A recent large-scale genome-wide meta-analysis of 24 cohorts replicated the association between Bifidobacterium abundance and the lactase (LCT) gene locus, which had previously only been reported in single-cohort studies. Other suggestive mbQTLs identified in this broad meta-analysis were proportional to heritability estimates from independent twin studies, indicating that additional loci found at lenient levels of significance are likely to be real but larger sample sizes are needed to reach sufficient statistical power.

Nonetheless, meta-analyses of mbQTL studies still lack resolution due to the high levels of heterogeneity between cohorts, which can reduce power in meta-analysis. On top of this, many existing cohorts rely on 16S rRNA measurements, which do not allow for bacterial identification at species-level resolution or for identification of bacterial pathway abundances. Indeed, measuring both species- and pathway-level abundances is essential for a further understanding of the function of an individual’s microbiome. Two closely related bacterial species can have very different functions in the gut microbiome, yet pathways may be shared across distant microbial species and have roughly the same biological effect. mbQTL studies using...
shotgun metagenomic sequencing in large population cohorts are therefore needed to overcome the high variability in microbiome definition and reveal more robust and refined associations.

To enable a broader and deeper understanding of host‒microbiota interactions, we used shotgun metagenomic sequencing on feces from 7,738 individuals of the Dutch Microbiome Project (DMP) and matched interactions, we used shotgun metagenomic sequencing on feces from 7,738 individuals of the Dutch Microbiome Project (DMP) and matched their imputed genotypes to differences in taxa and pathway abundances. By comparing our results with summary statistics from other independent studies, we identify novel host‒microbiota interactions. Furthermore, we explore the impact of potential confounding factors in modulating these genetic effects and identify potential diet-dependent host‒microbiota interactions. We further assess the potential causal relationships between the gut microbiome and dietary habits, biomarkers and disease using Mendelian randomization. Finally, we carry out a power analysis that shows how microbiome studies, even at the current sample size, are not well powered to reveal the complex genetic architecture by which host genetics can regulate the gut microbiome.

Results

Genome-wide association identifies known and novel associations with several bacterial taxonomies and bacterial pathways

We investigated 5.5 million common (minor allele frequency (MAF) > 0.05) genetic variants on all autosomes and the X chromosome, using linear mixed models to test their association and following an additive genetic model, with 207 taxa and 205 bacterial pathways in 7,738 individuals from the DMP cohort (Methods, Supplementary Table 1). Our quality control steps did not detect evidence for test statistic inflation (median genomic lambda 1.002 (range 0.75–1.03) for taxa and 1.004 (range 0.87–1.04) for pathways). We identified 37 SNP‒trait associations at 24 independent loci at a genome-wide \( p \)-value threshold of \( 5 \times 10^{-8} \) (Figure 1, Supplementary Table 2). Genetic variants at two loci passed the more stringent study-wide threshold of \( 1.89 \times 10^{-10} \) that accounts for the number of independent tests performed (Methods). The other 22 loci were associated at \( p \)-values between \( 5 \times 10^{-8} \) and \( 1.89 \times 10^{-10} \), with four associated to taxa and 18 to pathways. We refer to these 22 loci as “suggestive” and discuss them below.

The strongest signal was seen for rs182549 located in an intron of MCM6, a perfect proxy of rs4988235 \((r^2 = 1, 1000 Genome European populations)\), one of the variants known to regulate the LCT gene and responsible for lactase persistence in adults (ClinVar accession RCV000081214). The T allele of rs182549, which confers lactase persistence through a dominant model of inheritance, was found to be associated with decreased abundances of the species _Bifidobacterium adolescentis_ \((p = 7.6 \times 10^{-14})_ and _Bifidobacterium longum_ \((p = 3.2 \times 10^{-9})_ , as well as decreased abundances of higher-level taxa (phyla Actinobacteria \( p = 2.3 \times 10^{-13}\); class Actinobacteria, \( p = 2.3 \times 10^{-13}\); order Bifidobacteriales, \( p = 5.9 \times 10^{-15}\); family Bifidobacteriaceae, \( p = 5.8 \times 10^{-15}\); genus _Bifidobacterium_, \( p = 8.59 \times 10^{-15}_)_. Associations at this locus were also seen for other taxa of the same genus, but at lower levels of significance (_Bifidobacterium catenulatum_, \( p = 3.9 \times 10^{-5}\), and for species of the _Collinsella_ genus (Supplementary Figure 1). The genetic association at the LCT locus has been previously described, albeit only at the genus level, in Dutch, UK and US cohorts, as well as in a recent large-scale meta-analysis. The presence of the LCT association across several taxa suggests that this locus has a wide-ranging effect on microbiome composition.

The second locus that passed study-wide significance consists of genetic variants near the _ABO_ gene. _ABO_ encodes the BGAT protein, a histo-blood group ABO system transferase. Associations found at this locus include species _B. bifidum_ \((rs8176645, p = 5.5 \times 10^{-15})_ and _Collinsella aerofaciens_ \((rs550057, p = 2.0 \times 10^{-9}; r^2 = 0.59\) with rs8176645 in 1000 Genomes Europeans) and higher order taxa \((rs550057, genus Collinsella, p = 9.3 \times 10^{-11}; family Coriobacteriaceae, p = 3.01 \times 10^{-9}; order Coriobacteriales, p = 3.03 \times 10^{-9})_ (Supplementary Figure 1). Interestingly, the metabolic pathway representing the bacterial degradation of lactose and galactose was also associated to the _ABO_ locus (Metacyc ID: "LACTOSECAT-PWY: lactose and galactose degradation", rs507666, \( p = 5.38 \times 10^{-15}\)), was recently described. To our knowledge, the associations with _B. bifidum_ and _C. aerofaciens_ are novel.
Association at LCT reflects the lactase persistence inheritance model and affects multiple taxonomic levels and bacterial pathways

Given that lactose tolerance is inherited in a dominant fashion, we tested the associations found in this locus using a dominant model for the alternative allele at SNP rs182549 and thereby compared lactase persistent (LP) to lactose intolerant (LI) individuals. Indeed, all seven taxa associated to the LCT locus at genome-wide significance showed a stronger association signal when we used a dominant model (all associations $p < 2 \times 10^{-27}$), with decreased taxa abundance in LP individuals (Supplementary Table 3). The associations seen at the family level could mostly be accounted for by species B. adolescentis (no significant difference in effect size, Cochran’s $Q$ $p$ values>0.05), while lower effects were seen for species B. longum and B. bifidum (Cochran’s $Q$ $p$ values when comparing effect sizes with those observed for B. adolescentis: 0.018 and 0.003). Moreover, the association with these species remained unchanged when adding B. adolescentis to the association models, indicating that the associations are independent and not driven by species correlation.

We further tested the other 200 taxa for this SNP and the dominant model. Intriguingly, we also observed suggestive association ($p < 1 \times 10^{-4}$) with rs182549 for taxa that were associated to ABO locus in our GWAS analysis (Collinsella genus and species B. bifidum and C. aerofaciens) and for the species Roseburia inulinivorans of the family Lachnospiraceae (Supplementary Table 3). For all but Roseburia inulinivorans, there was a consistent direction of effect across the associated taxa, with increased abundance in LI compared to LP individuals (Figure 2 and Supplementary Figure 2).

Finally, when comparing the abundance of bacterial pathways between LI and LP groups, we also observed a higher abundance of the LACTOSECAT-PWY in LI individuals (effect $= +0.300$ in standard deviation (SD) units, standard error (SE) $= 0.49$, $p = 1.02 \times 10^{-9}$). This is not surprising given that in our dataset this pathway correlates mostly with class Actinobacteria and species B. adolescentis (Spearman correlations $r_S$: 0.73 and 0.69, respectively), which are both associated with SNPs at the LCT locus.
LACTOSECAT-PWY in blood type A or AB individuals compared to individuals with blood type O (Wilcoxon test for Collinsella, blood type O vs. A: \( p = 2.8 \times 10^{-9} \), 0 vs. AB: \( p = 0.11 \); Wilcoxon test for metabolic pathway, 0 vs. A: \( p = 5 \times 10^{-14} \), 0 vs. AB: \( p = 0.038 \)). In contrast, blood type O individuals had the highest observed abundance of *B. bifidum* compared to other blood type groups (Wilcoxon test blood type A vs. O: \( p = 2.3 \times 10^{-14} \), B vs. O: \( p = 0.007 \), AB vs. O: \( p = 0.006 \)) (Figure 3).

Figure 2. Association at the LCT locus and interaction with lactose intake.

Comparison of f. Bifidobacteriaceae relative abundance a) between groups of Lactase Persistent (LP, rs182549 C/T or T/T) and Lactose Intolerant (LI, rs182549 C/C) participants and b) stratified among individuals with low or high daily lactose intake levels. Lactose intake was corrected for daily calorie consumption. Relative abundance was natural log-transformed and adjusted by age and sex. Distribution is displayed with violin plots, where each inner box represents the first and third quartiles of the distribution and the middle line represents the median. Lactose intake levels were defined as low if < first quartile and high if ≥ first quartile. Lactose intake was only available for 5,801 of the 6,809 LP samples and for 376 of the 443 LI samples. P-values were obtained with a two-sided Wilcoxon rank test. Species in the same family shared similar distributions, although the difference in distribution within the LI group is not significant (Supplementary Figure 2).

Associations at ABO are dependent on secretor status

To further understand the mechanisms underlying the association signals at the ABO locus, we derived blood-group types based on the genotype status of three genetic variants (see Methods). The majority of the individuals were either type A (40%) or type O (48%), reflecting estimates from previous studies\(^1\). Genetic associations at this locus could be explained by differences between non-O blood type individuals and 0 blood type individuals. We observed higher abundances of *Collinsella* and the metabolic pathway

Figure 3. Association with blood types and interaction with FUT2.

Relative abundances of microbiome features found to be associated with SNPs at the ABO locus are compared between the inferred blood types (O, A, B and AB) with O as reference group (panels a, b and c) and between secretors (light blue) and non-secretors (dark blue) stratified again by the blood types grouped by the presence (A/B/AB) or absence (O) of terminal sugars (panels d, e and f). The natural log-transformed relative abundance, adjusted by age and sex, and is shown using a violin plot in which the inner box represents the first and third quartiles of the distribution and the middle line represents the median. In the top panels, the distribution across all blood types is also significantly different for all features (ANOVA: \( p = 6.7 \times 10^{-9} \).
Genetic associations may be modulated by diet

Gut microbiome composition and function is known to be affected by several factors including sex, body-mass index (BMI), diet and medication usage. None of our genome-wide significant associations were attenuated when including BMI, medication usage, stool frequency or stool consistency as covariates (Methods, all comparisons, Cochran’s Q for difference in effect size $p > 0.05$), indicating that the genetic associations are independent and not confounded by these factors (Supplementary Table 4a). Furthermore, we did not find any sex-specific associations for the 37 SNP–trait associations, although five associations did exhibit a smaller genetic effect in females compared to males (Supplementary Table 4b).

We also investigated the effect of diet at the LCT and ABO loci, taking into account the dominant inheritance model at LCT and the observed dependence on secretor status at ABO. Based on 21 dietary factors derived from a food questionnaire recorded 4 years prior to microbiome collection, we considered dietary factors previously associated (FDR < 0.05) with microbial taxa and pathway abundances that show genome-wide association signals in the ABO and LCT loci (see Methods)\(^{18}\). In an analysis that included age, sex, genetic and dietary factors, the dietary factors did not significantly attenuate the effect of the genetic components, suggesting that diet is not a source of bias in these genetic associations (Supplementary Table 5). Nonetheless, diet remained an important factor after correction for genetic factors. Four taxa and one pathway associated to LCT and ABO SNPs were statistically associated with at least one dietary factor ($p < 0.05$) (Supplementary Table 5), with a maximum of 16 factors found for B. longum. We further tested these associated dietary (44 diet–microbiome pairs) factors for interaction with genetics and detected evidence for a gene–diet interaction for only one bacteria at the LCT locus. Specifically, we observed an increased abundance of the Bifidobacteriaceae family in LI individuals who consumed larger amounts of lactose or dairy (interaction term $p = 0.03$ (Figure 3, Supplementary Table 6), a finding that is consistent with previous reports\(^5,18\). In contrast, there was no evidence for interaction with diet at the ABO locus for the B. bifidum species, the Collinsella genus and the bacterial pathway LACTOSECAT-PWY (interaction term $p > 0.05$ (Supplementary Table 6). This could be attributable to the limited accuracy of our diet scores, information that was recorded 4 years prior to microbiome collection. Interaction between fiber intake and genetic variants at this locus, when it is associated with Collinsella genus, have been reported in other populations\(^5,18\).
Association signals reveal a polygenic architecture of microbiome taxa and pathways and replicate in other independent cohorts

Apart from the study-wide significant associations at the LCT and ABO gene regions, 22 other loci showed suggestive association at \( p < 5\times 10^{-8} \) (Supplementary Table 2), none of which were reported previously. The majority (18) were associated with bacterial pathways that could not have been directly quantified in studies using 16S rRNA data, the methodology predominantly used in microbiome genetic studies to date. These associations included several genomic loci located in genes associated with human immune or metabolic disorders and mainly affected microbial amino acids (5), vitamin (5), nucleotide (4), TCA cycle (2), carbohydrate (1) and enterobactin toxin (1) pathways. These SNP-pathway associations provide insights into potential relationships between host phenotypes and gut microbial metabolic functions. For instance, SNP rs9927590 in 16q23.1 (in the WWOX gene), which has been reported to be involved in stenosis of Crohn’s disease \(^{37,38}\), was associated with a metabolic pathway involved in short chain fatty acid (SCFA) production (PWY-5088, \( p = 9.2\times 10^{-7} \)). SCFAs have been shown to boost intestinal barrier functions, reduce local inflammation and decrease the risk of immune and metabolic diseases, including type 1 diabetes \(^{29}\). Intriguingly, the gene ALG14 has been related to pneumonia \(^{39}\), while RPRD2 has been associated with immune diseases like asthma. Furthermore, another example is the association of SNP rs1584586 (3q25.1), near the type 2 diabetes-associated gene TSC22D2 \(^{40}\), with the purine processing pathway (PWY-841, \( p = 3.47\times 10^{-4} \)) and of SNP rs59657730 (6p12.3) within metabolic health related gene RCAN2, with microbial glycogen degradation (GLYCOCAT.PWY, \( p = 3.50\times 10^{-8} \)) support a role for host–microbiome interaction in regulating metabolic function. Four additional suggestive associations for taxa were identified, including signals in 8p22 (in the SGCZ gene related to diabetes) \(^{41}\) and in Xp21.1 (in the DMD gene), although little is known about the function of the associated species.

We sought to replicate our genome-wide significant signals using summary statistics from other independent cohorts in which microbiome data was characterized using either 16S rRNA (the MiBioGen study) \(^{13}\) or metagenomic sequencing (the LL-DEEP cohort) \(^{17}\). In the MiBioGen study, a meta-analysis of 24 cohorts comprising up to 18,340 individuals, the 16S rRNA measurements do not allow for the evaluation of the abundance of bacterial species and pathways and the X chromosome was not analyzed. Consequently, only 10 of our 18 SNP-taxa pairs could be tested in MiBioGen and no pathways were testable. In LL-DEEP, a genome-wide microbiome association study on 952 individuals, we extracted information for the majority of the SNP-taxa pairs (14/18) and SNP-pathway associations (18/19). Unfortunately, the power to replicate the associations in LL-DEEP was limited due to the small sample size. In both studies, we observed significant replication of the study-wide significant loci, LCT and ABO, using a conservative Bonferroni threshold of \( p < 0.0015 \) (equivalent to 32 SNP-trait pairs tested). All seven taxa associated with SNPs near LCT were replicated with consistent allelic effect directions (all \( p < 3.7\times 10^{-1} \)). For the ABO locus, we found significant replication for the Collinsella genus (\( p < 2\times 10^{-6} \) in MiBioGen) and replication at only nominal significance for the metabolic pathway LACTOSECAT-PWY and B. bifidum species (\( p < 0.05 \) in LL-DEEP) (Supplementary Table 7 and Supplementary Table 8). The association at ABO with B. bifidum has

In other loci, associations also point to immune and metabolic function. For example, two SNPs, rs12137024 (1p21.3, close to gene ALG14, \( p = 3.53\times 10^{-8} \)) and rs78778914 (1q21.2, close to gene RPRD2, \( p = 4.34\times 10^{-4} \)), were associated with the L-histidine processing pathway. L-histidine is usually catabolized by microbiota into glutamate, which is considered as an “immunonutrient” serving as major fuel for immune cells including lymphocytes and macrophages \(^{37,38}\). Notably, methyl diphosphate is a precursor of folate (Vitamin B9) that is predominantly used in microbiome genetic studies to date. These SNP-pathway associations included several genomic loci located in genes associated with human immune or metabolic disorders and mainly affected microbial amino acids (5), vitamin (5), nucleotide (4), TCA cycle (2), carbohydrate (1) and enterobactin toxin (1) pathways. These SNP-pathway associations provide insights into potential relationships between host phenotypes and gut microbial metabolic functions. For instance, SNP rs12137024 (1p21.3, close to gene ALG14, \( p = 3.53\times 10^{-8} \)). This SNP associated with a metabolic pathway involved in short chain fatty acid (SCFA) production (PWY-5088, \( p = 9.2\times 10^{-7} \)). SCFAs have been shown to boost intestinal barrier functions, reduce local inflammation and decrease of folate production from the gut microbiota.
not been reported previously and it does not reach the multiple-testing adjusted threshold for replication. Therefore, while the consistent direction of effects is encouraging, we cannot exclude the possibility for this signal to be a false positive. None of the other SNP-taxa or pathway pairs were replicated in MiBioGen or LL-DEEP. Interestingly, another independent SNP in the COL23A1 gene (rs11958296; \( r^2 = 0.1 \) with rs10447306 from our study) shows association in MiBioGen to the abundance of the same taxa: family Rikenellaceae (\( p = 2.4 \times 10^{-5} \)) and genus Alistipes (\( p = 9.3 \times 10^{-4} \)).

To explore if the association signals at lower levels of significance are enriched in heritable bacteria, which would indicate if it were possible to detect more genome-wide significant mbQTLs by further increasing the sample size, we investigated the correlation of taxa and pathway heritability estimations from family-based analysis with the number of suggestively associated loci for each taxon and pathway. Here we observed a positive and significant correlation for both taxonomic (\( r_s = 0.248, p = 3.2 \times 10^{-5} \)) and pathway (\( r_s = 0.263, p = 1.3 \times 10^{-4} \)) heritability with the number of suggestive (\( p < 1 \times 10^{-5} \)) loci identified in our GWAS (Figure 4). The correlation for pathways remained significant when increasing the mbQTL threshold to \( p < 5 \times 10^{-4} \) (\( r_s = 0.541, p = 6.3 \times 10^{-11} \)) and when further removing the LCT and ABO loci from the analyses (\( r_s = 0.540, p = 6.3 \times 10^{-11} \)).

We also evaluated if we could replicate any of the genome-wide significant signals outside the LCT and ABO loci that were reported in a recent and similarly-sized Finnish population study. After extracting all associations with \( p < 1 \times 10^{-4} \) in our dataset, we identified 3 out of 451 genome-wide significant SNPs from the Finnish study using direct or proxy (\( r^2 > 0.8 \)) information (279 of the SNPs reported in this study were not included in our dataset because their MAF was < 0.05). For one SNP, rs642387, we identified an association with consistent allelic effect for similar taxa: family Rikenellaceae (\( p = 3.1 \times 10^{-4} \)) and genus Alistipes (\( p = 3.2 \times 10^{-4} \)).

Mendelian randomization suggests salt intake and triglycerides are controlled by microbiome abundance

To investigate the causal relationships between microbiome composition/function and complex traits, and vice-versa, we used publicly available summary statistics in conjunction with our mbQTL results to perform two-sample Mendelian randomization (MR) analyses (see Methods). We focused on 78 phenotypes representing autoimmune diseases, cardiometabolic diseases, and related risk factors, as well as food preferences (Supplementary Table 10), and on 37 microbiome features that were associated with at least one variant at genome-wide significance (\( p < 5 \times 10^{-8} \)). None were significant at FDR < 0.05. At FDR < 0.1 (corresponding to a \( p = 2.7 \times 10^{-5} \) obtained from the inverse-variance weighted (IVW) MR test), we observed three causal relationships in the direction from microbiome to phenotypes. Two of the three are based on the same SNPs used as genetic instruments: genetic predispositions to a higher abundance of family Rikenellaceae and genus Alistipes are linked to lower consumption...
of salt (‘Salt added to food’ ordinal questionnaire phenotype, both traits: IVW \( p = 2.4 \times 10^{-5} \), causal effect = -0.06 SDs of salt intake for each SD increase in Rikenellaceae or Alistipes abundance, based on three genetic instruments located on different chromosomes) (Supplementary Table 11) (Supplementary Figure 3a,b). This correlation was not significant in the opposite direction (IVW \( p = 0.32 \) for both), suggesting that Rikenellaceae abundance may have an effect on the increased consumption of salt without being tagged by a shared causal factor. The other potential causal relationship that passed our FDR < 0.1 threshold was between the genus Collinsella and triglyceride levels. We found that a genetic predisposition to a one SD increase in the relative abundance of Collinsella was associated with a 0.12 SD decrease in triglyceride levels (\( p = 5.7 \times 10^{-5} \), FDR = 0.084) (Supplementary Table 11) (Supplementary Figure 3). All of these causal estimates were consistent when using other MR tests (weighted median and MR Egger) (Methods and Supplementary Table 11). Furthermore, there was no evidence for the causal effect being affected by pleiotropy (Egger intercept \( p = 0.74 \) for Rikenellaceae and Alistipes and 0.755 for Collinsella) (Supplementary Table 11). The results found in the two-sample MR were very similar when we performed a polygenic risk score analysis in the UK Biobank cohort (Supplementary Notes).

While all these potential causal effects are intriguing and exhibit robustness to sensitivity analyses, we recognize that the limited availability of genetic variants and the high FDR threshold dictate caution in their interpretation.

**Power analysis indicates that larger sample sizes are necessary to identify host genetic effects for microbial features with low prevalence**

One of the complicating factors in any genome-wide association study of the gut microbiome is the variable detection rate of the pathways and taxa measured, which reduces the effective sample size and consequently statistical power. In our study, the additive effect of associated variants at LCT explains 0.8% of the variance for species B. longum and B. adolescentis, which are present in >80% of the samples, and this allowed us to have ~70% power to detect the association at \( p < 1 \times 10^{-10} \) (Figure 5). In contrast, we had no power (~5%) to detect the association for species B. bifidum and B. catenulatum at this level of significance, as they were present in only 14% and 26% of samples, respectively. For B. bifidum, despite it being a rare species, we were able to detect a study-wide significant association at ABO because the effect was substantial (2.5% of the variance explained by SNP rs8176645, power = 83%), but we were underpowered to see the association at this locus with species C. aerofaciens, which was present in only 45% of the samples, and for which the additive effect of ABO variants was smaller (0.5% variance explained) (Figure 5). In a power analysis, we estimated that our sample size (7,738 individuals) is underpowered to detect genetic effects for taxa that are present in <80% of samples when considering an effect size comparable to or smaller than the effect of LCT for B. adolescentis. This is striking, considering that 93% of the taxa we identified in this cohort are present in <80% of the samples (Methods). We estimate that to find a similar effect size for at least 20% of gut microbiome composition (approximately 20% of the taxa are present in at least 20% of the samples in our cohort) would require a cohort of ~50,000 individuals.
Discussion

We carried out the largest genome-wide association study of gut microbiome composition and function in a single population by analyzing metagenomic sequencing data in 7,738 volunteers from the northern Netherlands. We recapitulated genetic associations at two known loci, LCT and ABO, and the resolution of our metagenomic sequencing allowed us to pinpoint novel associations with species (B. adolescentis at LCT, B. bifidum and C. aerofaciens at ABO) and bacterial pathways (LACTOSECAT-PWY at ABO) in these loci. Furthermore, we identified associations ($p < 5 \times 10^{-4}$) at 22 other loci for four taxa and 18 bacterial pathways. None of these associations were affected by major confounders of the gut microbiome such as medication usage, diet and BMI. Finally, we used an MR approach to pinpoint potential causal links between gut microbiome composition, complex traits and food intake habits.

The strongest association with the gut microbiome was found at the LCT locus, which remains the most robust genetic association with the gut microbiome identified to date. Associations at this locus with Bifidobacterium have been consistently reported in studies of different ethnicities, across a range of sample sizes, and in studies using different technologies and protocols for gut microbiome characterization. We also recapitulate that an increase of Bifidobacterium was more evident in L1 individuals who were consuming milk or milk-derived products. In addition, given that the resolution of metagenomic sequencing allows for species-level characterization of microbiome profiles, we could show that this effect was mainly attributable to the species B. longum, B. adolescentis, B. catenulatum and B. bifidum, an observation that was corroborated by two recent studies in Finnish and US Hispanic/Latino populations.

We found another study-wide association in the ABO gene locus with changes in abundance of several taxa and pathways. Associations with microbiome composition and blood types were observed in previous experimental studies. Genetic associations at the ABO locus have also been reported previously for populations of different ethnicities and in non-human species. For example, associations with Bacteroides and Faecalibacterium were reported in a study of five German cohorts that used 16S rRNA sequencing for gut microbiome characterization and with the microbial pathway of lactose and galactose degradation in a cohort of 3,432 Chinese individuals. The importance of ABO in determining host-microbiome interaction has also been recently reported in pigs. A deletion at this locus that inactivates the ABO acetyl-galactosaminyl-transferase has been shown to change the porcine microbiome composition by altering intestinal N-acetyl-galactosamine concentrations and consequently reducing the abundance of Erysipelotrichaceae strains, which have the capacity to import and catabolize N-acetyl-galactosamine. In our analysis, the strongest associations were with B. bifidum, which has not been reported before, and with the Collinsella genus, which was recently reported, and at genome-wide significance with C. aerofaciens. We did not detect any evidence of an interaction with diet at this locus, although this could be due to limitations in available information as the recording of dietary information was done at different times than the stool collection. We did, however, find that associations at this locus depend on secretor status that is determined by a nonsense mutation at the FUT2 gene and thus on the host’s ability to incorporate antigens into bodily fluids that are released in the gut. Intriguingly, we observed that taxa associated with the ABO locus also showed evidence of association at the LCT locus, this association being independent of blood type (interaction pvalue >0.05 for all taxa), indicating a common, independent action of these two loci in contributing to the growth of these bacteria. The most compelling hypothesis is that the availability of sugars in the gut, via undigested lactose in L1 individuals or secretion of antigens with accessible glycans in non-O blood type secretors, provides direct energy sources for most of these bacteria. This is further supported by the observation that L1 individuals and non-O blood type secretors were both associated with the increased abundance of a bacterial pathway for lactose and galactose degradation. However, this mechanism would not fully explain the opposite direction of the association at ABO seen for B. bifidum, which, under this hypothesis and considering its adaptation to normal gastrointestinal mucosa (apparently independent of H-antigen secretion, as shown in Figure 4), would be subjected to competition in the environment, such as in the case of non-O blood secretors subjects. Of note, a similar pattern of complex association at this locus was found in a recent study of German cohorts, where a branch of the Bacteroides genus represented by OTU97_12, OTU99_12, and TestASV_13, showed association with an inverse relationship between their prevalence and the non-O blood type group and instead a positive relationship with prevalence of OTU97_27. While these opposite associations could be explained by antigen degradation activity of certain species and consequent environment competition for others, more studies are needed to clarify the complex mechanisms involved.
We acknowledge that anachronistic diet information is a limitation of our study. While we have shown that in general the microbiome remains fairly stable in an individual after 4 years, with inter-individual differences being larger than inter-individual differences <https://doi.org/10.1016/j.cell.2021.03.024>, short-term changes in diet, especially those introducing drastic shifts, can perturbate both microbiome composition and function. These could not be taken in account by our analyses. Capturing these shifts would not be easy even with frequency food questionnaires recorded at the time of sample collection; ideally, real-time extensive recording in weeks preceding microbiome collection should be implemented in future biobanks.

The strongest mbQTLs we identified reside in genes under selective pressure. The LCT gene is a highly differentiated gene among human populations due to positive selection conferred by the lactase persistence phenotype. It has been estimated that strong selection occurred within the past 5,000–10,000 years, consistent with there being an advantage to lactase persistence and the ability to digest milk in the setting of dairy farming<sup>50</sup>. Variants at this locus have been linked through GWAS to not only food habits and metabolic phenotypes, but also to immune cell populations<sup>51</sup>. The ABO locus is evolutionarily highly differentiated; it has been shown to have experienced balancing selection for the last 3 million years in many primate species<sup>52</sup>. Several evolutionary sources of selective pressure have been proposed, including via infections by pathogens such as malaria<sup>53</sup> and cholera<sup>54</sup>. ABO variants have also been linked to cardiometabolic traits, white and red blood cell levels and cytokine levels<sup>55,56</sup>. Therefore, host-microbe interactions are likely to be shaped by human-microbe co-evolution and survival, probably through a balance between food availability for gut bacteria and enhanced immune response of the host. A better understanding of these interactions will expand our current knowledge of human evolution<sup>57</sup>.

From this perspective, it will be crucial to compare genetic studies of the gut microbiome in diverse populations with different genetic backgrounds in order to understand the complexity of host-microbe interactions. This requires community efforts to standardize the definition of taxonomies and to standardize measurement methodologies in order to facilitate comparison between cohorts. For example, in our attempt to replicate the findings from the Finnish population cohort, only a limited number of taxa could be directly matched or connected through the Genome Taxonomy Database<sup>58</sup>.

To explore the causality of the relations of the microbiome with complex traits and food preferences, we performed bi-directional MR analysis using 56 dietary traits, 16 diseases and 5 biomarkers. At FDR < 0.1, we observed three causal relationships. Although the limited impact of genetic variants on both microbiome composition and dietary preferences requires caution when interpreting causality estimation by MR<sup>59</sup>, it is intriguing to observe a causal role for microbial composition on food preferences. Here we observed that an increase in the abundance of genus Alistipes and its family Rikenellaceae led to decreased consumption of salt. Although multiple studies have shown that dietary changes, such as variations in salt and meat intake, have a strong effect on microbiome composition<sup>60,61</sup>, it is intriguing to suggest that genetically determined variations in the microbiome might affect individual food preferences. This would be supported by bacterial genetic variations in salt tolerance<sup>62</sup> and by the established knowledge that the composition of gut microbiome can accurately predict the effect of food items on host metabolism<sup>63</sup>. There was additional evidence from our results to support a role for the microbiome in influencing food preferences. A perfect proxy of rs642387 (rs503397, $r^2 = 0.99$ in 1000 Genomes Europeans), which was recently reported to be associated with the family Desulfovibrionaceae and related taxa in a Finnish population and replicated in our study, had previously been associated to bitter alcoholic beverage consumption in an independent cohort<sup>64</sup>. Another study found an increase in family Desulfovibrionaceae in individuals with high alcohol consumption<sup>65</sup>, and family Desulfovibrionaceae, genus Desulfovibrio and other related taxa were also associated with increased consumption of alcohol in our DMP cohort<sup>16</sup>, supporting a pleiotropic effect of this locus on both microbiome and alcohol intake. These studies, together with our findings, suggest that gut microbiota could indirectly influence an individual’s food preferences by mediating the downstream effect of the consumption of different products.

Understanding the differences in microbial pathway abundance is crucial for deeply understanding the underlying function of the gut microbiome. Microbiome measurements that target single organisms or taxonomic groups may be inefficient because two strains of one species may differ in their functional capacities. It is therefore essential to evaluate not only the taxonomic but also the functional composition of the microbiome. To our knowledge, we report the largest single study to investigate host genetic effects on the human gut microbiome function using metagenomics sequencing and thereby perform genome-wide association analysis of microbial pathways. Our analyses revealed several loci associated with
microbial pathways, although only one passed stringent study-wide significance. Functional pathway abundances are also an approximate indicator of metabolite expression. For example, we report an association with PWY-5088, which is involved in propionate production. Although we do not have SCFA levels in our samples for direct validation, we have measured propionate levels in feces in the LL-DEEP cohort and using summary statistics from our previous study, we see consistent directions of effect for the association with PWY-5088 (allele T was associated with decreased levels of propionate), but the association was not significant, perhaps due to the small sample size (N=897, p-value = 0.08). It also should be noted that pathway abundance, even if not necessarily representing the expression, might represent the gene load and abundance of the bacterial taxa that are able to execute this metabolic pathway, which by itself is an important indicator for the metabolic capacity of a given gut microbiomal community. We believe more metabolite screening should be undertaken once more robust and consistent replications are detected for microbial pathways by other genetic studies.

In addition to the two study-wide significant loci, we also observed 22 loci at genome-wide significance and many more at a more lenient threshold. The correlation we observe between the heritability of microbial taxa and pathways and the number of suggestively associated loci indicates that mbQTLs with smaller effects are likely to exist. Those loci remained under the detection limit in the current study, and the sample size would need to be increased by orders of magnitude to discover their role. In fact, gut microbiome composition is characterized by high inter-individual variation. Analysis of core microbiota in different populations indicates that only a few bacteria are present in >95% of studied individuals, combined with harmonized methodology to reduce technical biases. We conclude that joint efforts that combine tens of thousands of larger cohorts are therefore required to identify genetic contribution to rarer bacteria are present in >95% of studied individuals and concluded that only 7% of all identified taxa in the cohort. For bacteria present in >20% of the samples, >50,000 participants would be necessary to identify an effect size similar to that of LCT and ABO. Much larger cohorts are therefore required to identify genetic contribution to rarer bacteria. We conclude that joint efforts that combine tens of thousands of individuals, combined with harmonized methodology to reduce technical bias, will be needed to characterize more than a few major loci, as has also been the case for genetic studies of much more heritable quantitative traits such as BMI (heritability ~40%), height (heritability ~80%) and other human phenotypes.

Methods

Cohort description

Lifelines is a multi-disciplinary prospective population-based cohort study with a unique three-generation design that is examining the health and health-related behaviors of 167,729 people living in the north of the Netherlands. Lifelines employs a broad range of investigative procedures to assess the biomedical, socio-demographic, behavioral, physical and psychological factors that contribute to the health and disease of the general population, with a special focus on multi-morbidity and complex genetics. During the first follow-up visit, all participants were invited to participate in a parallel project, The Dutch Microbiome Project (DMP), on a voluntary basis. The goal of this project is to evaluate the impact of different exposures and life-styles on gut microbiota composition. A subset of 8,719 Lifelines participants agreed to participate and feces were collected. Of these, 8,208 were retained for downstream analysis after stringent quality control. Distribution of age, gender and location within the three north provinces was similar to that observed in the total Lifelines cohort.

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

Genome characterization

Genotyping of 38,030 Lifelines participants was carried out using the Infinium Global Screening array® (GSA) MultiEthnic Diseases version, following the manufacturer’s protocols, at the Rotterdam Genotyping Center and the Department of Genetics of the University Medical Center Groningen. Here, we used available quality-controlled genotyping data imputed with Haplotype Reference Consortium (HRC) panel v1.117, as described elsewhere. To avoid population stratification, we analyzed only European samples. We selected 43,587 genetic markers by applying linkage disequilibrium (LD) pruning on the genotyped data (sliding window of 1Mb, linkage disequilibrium r2<0.2, step = 5) and used them for a Principal Component Analysis (PCA) projecting the
Microbiome characterization

The gut microbiome was characterized from stool samples as described in Gacesa et al.16. In brief, stool samples were collected by participants and frozen within 15 minutes after production and then transported on dry ice to the Lifelines facility to be stored at -80°C. Microbial DNA was extracted using the QIAamp Fast DNA Stool Mini Kit (Qiagen, Germany), following the manufacturer’s instructions. Samples with a total DNA yield lower than 200 ng (as determined by Qubit 4 Fluorometer) were prepared using NEBNext® Ultra™ II DNA Library Prep Kit for Illumina®. Shotgun metagenomic sequencing was carried out using the Illumina HiSeq 2000 platform at Novogene, China. Metagenomic sequencing data was profiled following methods previously used in other cohorts, as described in16,75. Low quality reads (PHRED quality ≤ 30), adapters and host sequences were removed using KneadData tools v0.5.1. Taxonomic composition was determined with MetaPhlAn2 v2.7.2.76. Characterization of biochemical pathways was performed with the HUMAnN2 pipeline v0.11.177, integrated with the UniRef90 v0.1.1 protein database78, the ChocoPhlAn pan-genome database and the DIAMOND alignment tool v0.8.22.79. After quality control (samples with eukaryotic or viral abundance ≤ 25% and total read depth ≥ 10 million were retained), we had information on 950 microbial taxa and 559 functional pathways. For this study, we focused only on bacterial taxa and functional pathways with mean relative abundance >0.001% across all samples and present in at least 1000 of the 7,738 participants, which resulted in a list of 207 taxonomies (5 phyla, 10 classes, 13 orders, 26 families, 48 genera and 105 species) and 328 bacterial pathways. Furthermore, we removed redundant pathways by discarding one pathway among pairs that were highly correlated ($r_s > 0.95$), as well as pathways not previously described in bacteria that could thus be coming from sources other than bacteria, resulting in 205 pathways for genetic analyses.

Diet phenotypes definition

Dietary habits were assessed using a semi-quantitative Food Frequency Questionnaire (FFQ) designed and validated by the division of Human Nutrition of Wageningen University as described before in Siebelink et al and Gacesa et al.16,80. The FFQ data was collected 4 years prior to fecal sampling, and supplementary questionnaires were collected concurrent with fecal sampling, with the stability of long-term dietary habits between time points assessed as described in Gacesa et al.16. We analyzed the dietary factors that were previously found to be associated (FDR < 0.05) to the microbiome features in our study that had a genome-wide significant signal in the ABO and LCT loci. We also analyzed lactose intake for the species Bifidobacterium longum and Bifidobacterium adolescentis, given their association with the LCT region in our study. Participants with an implausible caloric intake (< 800 or > 3934 kcal/day for males, < 500 or > 2906 kcal/day for females)81 were not included in these analyses.

GWAS analysis method

Genome-wide association analysis was performed in 7,738 European samples for 412 features (205 functional pathways and 207 microbial taxa), investigating genetic additive effects using allele dosages for 5,584,686 genetic variants with MAF > 0.05 and information score > 0.4 on the autosomes (chromosomes 1-22) and the X chromosome. We focused on the quantitative dimensions of relative bacterial and pathway abundances, treating all zero values as missing data. We used natural-log-transformed abundances and regressed these in a linear mixed model using SAIGE v.0.3819, with age, sex and the genetic relationship matrix (GRM) among participants as covariates. We used the standard settings of SAIGE, which applies inverse-rank normalization to the traits prior to the association analyses. The GRM was built with SAIGE using a set of 54,565 SNPs selected from the total set of quality-controlled SNPs directly genotyped and filtered for allele frequency and redundancy (MAF ≥ 0.05, r2 < 0.2, sliding window = 500Kb).

Definition of the study-wide significant threshold

To estimate the number of independent phenotypes assessed, we used PCA on the matrix of 412 microbiome features (207 taxonomies and 205 pathways) available for GWAS analysis to decompose variability in independent components (axes). We estimated that 264 components are needed to explain 90% of the microbiome variance. We then defined our study-wide significant p-value threshold by correcting the genome-wide significance threshold for this factor (5x10^{-8}/264 = 1.89x10^{-10}).
Association using dominant model
To evaluate association using a dominant model on SNP rs182549 at the LCT locus, we used best-guess genotypes and converted T/C to T/T. Association analysis was then run for all taxa as well as the LACTOSECAT-PWY pathway using SAIGE v0.7.1 and the same covariates and transformation used for the GWAS analysis.

Inference of blood groups
We estimated blood groups from genotyped and imputed data following the scheme of Ellinghaus et al.96. Specifically, we used the absence of the rs8176719 insertion to define blood-type allele O1, the T allele of rs41302905 to define blood-type allele O2 and the T allele of rs8176746 to define the blood-type allele B (instead of rs8176747). Diploid individuals O1O1, O2O2 and O1O2 were considered blood type O. Diploid individuals O1B, O2B and BB were considered blood type B. Absence of the alleles mentioned above was used to define blood-type allele A. To evaluate differences across blood types, we compared the mean relative abundance of microbiome features in individuals with A, B and AB blood type to that in individuals with the O blood type using a two-sided Wilcoxon test. To evaluate the interaction with the rs601338 FUT2 (secretor/non-secretor) locus, we grouped individuals into two groups (non-O blood type and blood type O) to distinguish production or non-production of antigens, and compared pairs using a two-sided Wilcoxon test. All analyses were done using base R v3.6.1 (https://www.R-project.org/).

Effects of potential confounders on significantly associated loci
We evaluated the robustness of genome-wide associated signals by incorporating the following potential confounders into our statistical model: medication usage, anthropometric data and stool frequency and consistency data (collection and processing was described in Gacesa et al.16). We analyzed the effects of the following medication groups: proton pump inhibitors (ATC A02BC, N = 130), laxatives (osmotic ATC A06AD, N = 44; volume increasing ATC A06AC, N = 77), one group of antibacterials (ATC J01, N = 24) and other anti-infectives (ATC J, N = 39). The other group of medication considered was antibiotic use in the 3 months prior to stool collection (N = 450). For each of these medications, we created dichotomous variables for all participants coded as 0 (non-user) or 1 (user). The other factors included were BMI, stool frequency and stool consistency (mean Bristol stool scale). All models also incorporated age and sex as covariates and were run only for the genome-wide-significant SNP-trait pairs (Supplementary Table 1) using the same software used for GWAS (SAIGE, Zhou et al.19). To evaluate the impact of these covariates on the genetic signals, we used Cochran’s Q heterogeneity test to compare the effect size obtained by the covariate-inclusive model and the basic model (that only includes age, sex and the genetic variant). To evaluate the impact of sex, we ran the SNP-association analysis in SAIGE separately for males and females using only age as a covariate. For each genetic variant, differences in effect size in males and females were tested using Cochran’s Q heterogeneity test.

Interaction analyses
We used a three-step procedure to evaluate gene-diet interactions for all the taxa and pathways associated with SNPs at the LCT and ABO loci. First, we extracted the variables representing dietary habits that had previously shown significant association with these microbial traits at FDR<0.0516. For the genus Collinsella and pathway LACTOSECAT-PWY, no dietary factors were found at this FDR threshold. We therefore considered the same dietary factors associated with B. bifidum in the analyses, given that they showed similar patterns of genetic association. Next, we added these variables to the basic genetic model (feature ~ age + sex + genetic variant) to confirm their association at (at least) nominal significance level (p < 0.05) while accounting for the associated genetic variant(s). Finally, for the dietary variables showing nominal significance, we evaluated the interaction with the genetic variant(s) by including an interaction term into the association model. For the LCT locus, we considered a binary variable to distinguish two groups of genotypes at SNP rs182549 (C/C vs. C/T and T/T) according to the dominant inheritance model at this locus. For the ABO locus, we used a binary definition of blood type (blood type O vs A/B/AB) and also considered the effect of the rs601338 genotype in the FUT2 gene (defining secretor/non-secretor individuals). All microbiome features were inverse-rank normalized before analyses, and age and sex were added as covariates as in the main GWAS analysis. All models were fit using the lm() function from base R v3.6.1, other statistical tests were as implemented in packages rstatix v0.5.0 and ggpubr v0.3.0 and package RNOmni v0.7.1 was used for the inverse-rank normalization.

Replication in other cohorts and data sets
We looked for replication of our results using summary statistics from two independent studies: a genome-wide meta-analysis of 165 rRNA data from 24 cohorts (the MiBioGen consortium) and a genome-wide study on metagenomics data in the LL-DEEP cohort, another subset of the Lifelines
cohort with data generated 4 years before the DMP and in which 255 participants were also later enrolled in DAG311,17. In this study, a different DNA isolation procedure (All prep) was used and taxonomies were defined using the Bracken pipeline, which may explain why C. areofaciens was not identified. In the MiBioGen study, we could not look at SNPs associated with species or pathways, as these microbiome features cannot be defined using 16S data, or at SNPs on the X chromosome, as they were not analyzed in this study. In the second study, we searched for the exact same taxonomy or pathway, but similarly to MiBioGen, X chromosomal variants were not tested and some taxonomies were not defined due to the differences in metagenomic data processing pipelines.

Next to the replication of our findings, we also evaluated whether the genome-wide signals reported in a recent genome-wide study of microbiome taxa from a Finnish cohort were replicable in our data18. We searched for all genome-wide signals reported in a recent genome-wide study of microbiome Next to the replication of our findings, we also evaluated whether the metagenomic data processing pipelines.

test and some taxonomies were not defined due to the differences in or pathway, but similarly to MiBioGen, X chromosomal variants were not identified. In the MiBioGen study, we could not look at SNPs associated with species or pathways, as these microbiome features cannot be defined using the Bracken pipeline, which may explain why C. areofaciens was not DNA isolation procedure (All prep) was used and taxonomies were defined with species or pathways, as these microbiome features cannot be defined using 16S data, or at SNPs on the X chromosome, as they were not analyzed in this study. In the second study, we searched for the exact same taxonomy in our cohort (Supplementary Table 9). We then looked at these associated taxa in the respective cohorts and compared them visually and with the aid of the Genome Taxonomy Database (https://gtdb.ecogenomic.org/) to determine if they were the same bacterial taxa or taxa from the same taxonomic branch.

**Correlation between heritability estimates and the number of associated loci**

To analyze the correlation between family-based heritability and the number of suggestive mbQTLs, we used narrow-sense heritability estimates for taxa and pathways that accounted for household environment sharing and previously derived for this cohort16. We then calculated the number of independent mbQTLs per microbial trait by performing LD pruning (\(r^2 < 0.1\) in our data set, window size 1Mb using Plink v1.941) for all SNPs at the three different thresholds: \(p < 5 \times 10^{-8}\), \(< 1 \times 10^{-4}\) and \(< 5 \times 10^{-5}\). The association of heritability and the number of mbQTLs was calculated in R v.4.0.3 using a weighted Spearman correlation from the `wCorr` v1.9.1 package, with each taxon or pathway treated as a data point. The weights used in calculating the correlation were inversely proportional to the Z scores calculated from heritability pvalue estimates. The regression lines in Figure 4 were fit using the loess function (locally estimated scatterplot smoothing, base R package v4.0.3) with span and degree parameters set to 1.

**Mendelian randomization analysis**

To evaluate potential causal relationships between the gut microbiome and other common traits, we performed Mendelian randomization (MR) analyses that combined the summary statistics of the microbiome with publicly available summary statistics on food preferences, autoimmune and cardiovascular diseases and other cardio-metabolic traits. We analyzed the 37 microbiome features (pathways and taxa) with at least one variant passing the \(p < 5 \times 10^{-8}\) threshold (Supplementary Table 2) and combined these with 78 publicly available summary statistic datasets retrieved using the IEU GWAS DATABASE85.

We performed a bi-directional MR analysis, first testing if microbiome traits causally affect a phenotype and then testing if phenotypes can causally affect the microbiome traits. For each comparison, we intersected the microbiome variants (MAF > 0.05) by rsID, position and alleles with the publicly available summary statistic variants. We then selected instruments using the ‘clump_ data()’ function of the TwoSampleMR package (v0.5.5)86. The publicly available summary statistics were clumped using a \(p\)-value threshold of \(< 5 \times 10^{-4}\) and otherwise standard settings (\(r^2 < 0.001, 10mb \)window size). Due to the limited statistical significance of the microbiome traits, we performed \(p\)-value clamping at a less stringent \(p < 5 \times 10^{-4}\) threshold. If fewer than three variants were clumped, we removed the trait combination from analysis.

The MR analysis was done using the TwoSampleMR v0.5.5 package. We first selected trait combinations that passed the Benjamini-Hochberg FDR threshold of 0.1 (\(p = 2.805 \times 10^{-5}\)) in the inverse-variance weighting (IVW) test, resulting in one suggestively causal trait combination. We further checked that the significant trait combinations were unlikely to be driven by pleiotropy based on two criteria: i) the Egger regression intercept was nominally significant (indicating the presence of horizontal pleiotropy) and ii) the weighted median results were not nominally significant (indicating that no single variant influences the result)97,98.

**Power analysis**

We calculated the variance explained at our loci using the formula described in Teslovich et al.99, which takes into account MAF, effect size, standard error and sample size. We then performed a power analysis based on a linear model of association, considering different genetic effect sizes (variance explained) and sample sizes (N) <https://genome.sph.umich.edu/wiki/
Power_Calculations:_Quantitative_Traits>. We performed a sample size calculation by doing a grid search in the sample size sequence: \([1000, 1050, \ldots, 50000]\) and kept the lowest sample size that had power > 80%.

**LifeLines Cohort Study - group authors genetics**

Raul Aguirre-Gamboa, Patrick Deelen, Lude Franke, Jan A Kuivenhoven, Esteban A Lopera Maya, Ilja M Nolte, Serena Sanna, Harold Snieder, Morris A Swertz, Judith M

**References**


