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

Lopera-Maya, Esteban A; Kurilshikov, Alexander; van der Graaf, Adriaan; Hu, Shixian; Lifelines Cohort Study; Andreu-Sánchez, Sergio; Chen, Lianmin; Vila, Arnau Vich; Gacesa, Ranko; Sinha, Trishla

Published in:
Nature genetics

DOI:
10.1038/s41588-021-00992-y

Publication date:
2022

nature genetics

ARTICLES

https://doi.org/10.1038/s41588-021-00992-y

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

Esteban A. Lopera-Maya[1,8], Alexander Kurilshikov[1,8], Adriaan van der Graaf[1,8], Shixian Hu[1,2,8], Sergio Andreu-Sánchez[1,3], Lianmin Chen[1,3], Arnau Vich Vila[1,2], Ranko Gacesa[1,2], Trishla Sinha[2], Valerie Collij[1,2], Marjiolein A. Y. Klaassen[1,2], Laura A. Bolte[1,2], Milla F. Brandao Gois[1], Pieter B. T. Neerincx[1,4], Morris A. Swertz[1,4], LifeLines Cohort Study*, Hermie J. M. Harmsen[5], Cisca Wijmenga[1], Jingyuan Fu[1,3], Rinse K. Weersma[2], Alexandra Zhernakova[1,9] ✉ and Serena Sanna[1,6,9] ✉

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 in 7,738 participants of the Dutch Microbiome Project. 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, whereas 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.

The human intestinal microbial community contains trillions of microorganisms that play an important role in maintaining normal gut function and immune homeostasis[1]. Emerging evidence shows that gut microbial composition alterations are associated with the pathogenesis of human diseases, including gastrointestinal disorders, metabolic syndrome, cardiovascular diseases and other conditions[2,3].

Many environmental factors influence the gut microbiome, including diet, medication usage[4,5] and host genetics. Heritability studies have estimated that human genetics could explain from 1.9% to 8.1% of gut microbiome variation[6,7]. This observation drove the first efforts to identify genomic loci that influence gut microbiota through genome-wide association studies (GWASs). These early gut microbiome GWASs identified several microbial quantitative trait loci (mbQTLs) located in genes related to the intestinal mucosal barrier, immune response and drug and food metabolism[8–11]. 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[12]. A recent large-scale genome-wide meta-analysis of 24 cohorts replicated the association between *Bifidobacterium* abundance and the lactase (*LCT*) gene locus[13], which had previously been reported in single-cohort studies[6,14]. 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[13]. Nonetheless, meta-analyses of mbQTL studies are still underpowered due to the high levels of heterogeneity between cohorts. 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 abundances is essential for a further understanding of an individual's microbiome; pathways may be shared across distant microbial species and have the same biological effect[15]. mbQTL studies using shotgun metagenomic sequencing in large cohorts are therefore needed to overcome the variability in microbiome definition to reveal robust associations.

For a broader and deeper understanding of host–microbiota interactions, here, we use shotgun metagenomic sequencing on feces from 7,738 individuals of the Dutch Microbiome Project (DMP)[16] and match their imputed genotypes to differences in taxa and pathway abundances. By comparing our results with summary

[1] Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. [2] Department of Gastroenterology and Hepatology, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. [3] Department of Pediatrics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. [4]University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Groningen, the Netherlands. [5]Department of Medical Microbiology and Infection Prevention, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands. [6]Institute for Genetic and Biomedical Research (IRGB), National Research Council (CNR), Cagliari, Italy. [8]These authors contributed equally: Esteban A. Lopera-Maya, Alexander Kurilshikov, Adriaan van der Graaf, Shixian Hu. [9]These authors jointly supervised this work: Alexandra Zhernakova, Serena Sanna. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: sasha.zhernakova@gmail.com; serena.sanna@irgb.cnr.it

statistics from other independent studies[13,17,18], 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 (MR). Finally, we carry out a power analysis showing how microbiome studies, even at the current sample size, are underpowered to reveal the complex genetic architecture by which host genetics regulates the gut microbiome.

Results

Genome-wide associations with bacterial taxa and 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[19] to test their association with 207 taxa and 205 bacterial pathways in 7,738 individuals from the DMP cohort (Methods and Supplementary Table 1)[19]. There was no 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 single nucleotide polymorphism (SNP)–trait associations at 24 independent loci at a genome-wide P value threshold of 5×10^{-8} (Fig. 1 and Supplementary Table 2). Genetic variants at two loci passed the more stringent study-wide threshold of 1.89×10^{-10} that accounts for the number of independent tests performed (Methods).

The strongest signal was seen for rs182549 located in an intron of *MCM6*, a perfect proxy of rs4988235 ($r^2=1$, 1000 Genomes Project European populations), one of the variants known to regulate the *LCT* gene and responsible for lactase persistence in adults (ClinVar accession RCV000008124). 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* ($P=7.6\times10^{-14}$) and *Bifidobacterium longum* ($P=3.2\times10^{-08}$), as well as decreased abundances of higher-level taxa (Supplementary Table 2 (ref. [5])). Associations at this locus were also seen for other taxa of the same genus but at lower levels of significance (*Bifidobacterium catenulatum*, $P=3.9\times10^{-5}$) and for species of the *Collinsella* genus (Extended Data Fig. 1). The genetic association at the *LCT* locus has been previously described, albeit only at the genus level, in Dutch, UK and US cohorts[6,8,14], as well as in a recent large-scale meta-analysis[13].

The second locus that passed study-wide significance consisted 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 *Bifidobacterium bifidum* (rs8176645, $p=5.5\times10^{-15}$) and *Collinsella aerofaciens* (rs550057, $P=2.0\times10^{-8}$, $r^2=0.59$ with rs8176645 in 1000 Genomes Project Europeans) and higher-order taxa (rs550057, genus *Collinsella*, $P=9.3\times10^{-11}$; family Coriobacteriaceae, $P=3.01\times10^{-9}$; order Coriobacteriales, $P=3.03\times10^{-9}$) (Extended Data Fig. 1). Interestingly, the metabolic pathway representing the bacterial degradation of lactose and galactose was also associated with the *ABO* locus (Metacyc ID LACTOSECAT-PWY: lactose and galactose degradation I, rs507666, $P=5.38\times10^{-15}$). Associations of this locus with the genus *Collinsella* and the metabolic pathway LACTOSECAT-PWY have been recently described[18,20].

Association at *LCT* affects multiple taxa and 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) and lactose-intolerant (LI) individuals. Indeed, all seven taxa associated with the *LCT* locus at genome-wide significance showed a stronger association signal when we used a dominant model (all associations $P<2\times10^{-27}$), with increased taxa abundance in LI 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 value > 0.05), whereas smaller 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* were 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.

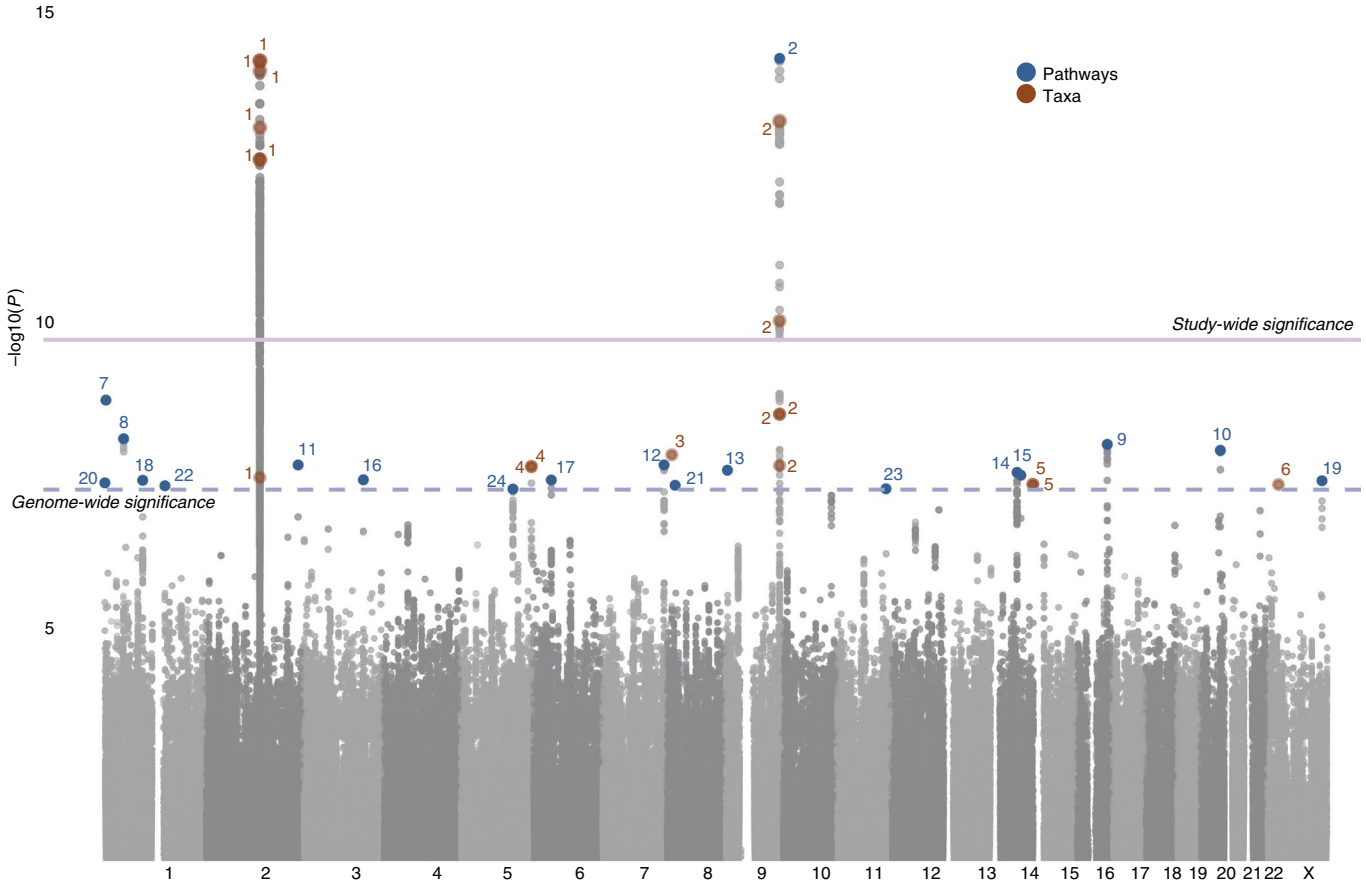
We further tested the other 200 taxa for this SNP and the dominant model. Intriguingly, we observed suggestive association ($P<1\times10^{-4}$) at rs182549 with taxa that were associated with the *ABO* locus in our GWAS (*Collinsella* genus and species *B. bifidum* and *C. aerofaciens*) and for the species *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 with LP individuals (Fig. 2 and Extended Data Fig. 2). The associations seen with several taxa suggest that this locus has a wide-ranging effect on microbiome composition.

Finally, when comparing the abundance of bacterial pathways between the LI and LP groups, we observed a higher abundance of the LACTOSECAT-PWY in LI individuals (effect = +0.300 in s.d. units, s.e. = 0.049, $P=1.02\times10^{-9}$). This is not surprising given that in our dataset, this pathway correlates mostly with class Actinobacteria and species *B. adolescentis* (Spearman correlation [r_S], 0.73 and 0.69, respectively), which are both associated with SNPs at the *LCT* locus.

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 (Methods). The majority of the individuals were either type A (40%) or type O (48%), as expected[21]. Genetic associations at this locus could be explained by differences between individuals with non-O blood type and O blood type. Individuals with blood type O had the highest observed abundance of *B. bifidum* compared with other blood-type groups (Wilcoxon test blood type, A vs. O: $P=2.3\times10^{-14}$, blood type B vs. O: $P=0.007$, blood type AB vs. O: $P=0.006$), whereas higher abundances of *Collinsella* and the pathway LACTOSECAT-PWY were observed in individuals with blood type A compared with individuals with blood type O (Wilcoxon test for *Collinsella*, blood type O vs. A: $P=2.8\times10^{-9}$, for metabolic pathway, blood type O vs. A: $P=5\times10^{-14}$) (Fig. 3).

Notably, all these associations were present only in individuals able to expose A/B antigens to gut bacteria (secretors) and were absent in nonsecretors, with secretor status being determined by a functional variant in the *FUT2* gene (Fig. 3 and Supplementary Note). This observation is in line with recent studies showing that association at the *ABO* locus with microbiome depends on *FUT2* genotypes[18,22]. Associations of functional variants in the *FUT2* gene with other bacterial taxa were observed in a recent meta-analysis[13], but none of the bacterial taxa or pathways analyzed showed significant association at this locus in our cohort.

The novel associations with species levels at *ABO*, including that with *B. bifidum*, are intriguing. Early genomic analyses suggest that among *Bifidobacterium*, *B. bifidum* is particularly adapted to the human gastrointestinal mucosa because of a unique set of genes encoding for enzymes involved in the degradation and utilization of mucin, the main component of gastrointestinal mucosa[23,24]. The degradation activity of the mucosa, where antigens are secreted, could explain the association with differences in *B. bifidum* abundance at *ABO* modulated by *FUT2*. This mechanism has been proposed for the association with *Collinsella* genus[18], and it is likely to also apply to the novel associations we identified with *C. aerofaciens* and LACTOSECAT-PWY pathway abundances.



Locus ID	Genes (variant function)	rs ID	*P* value	Trait
1	*MCM6* (intronic)	rs182549	5.86×10^{-15}	f. *Bifidobacteriaceae*
1	*MCM6* (intronic)	rs182549	5.91×10^{-15}	o. Bifidobacteriales
1	*MCM6* (intronic)	rs182549	8.59×10^{-15}	g. *Bifidobacterium*
2	*ABO* (intronic)	rs8176645	5.54×10^{-14}	s. *Bifidobacterium bifidum*
1	*DARS1* (intronic)	rs6754311	7.06×10^{-14}	s. *Bifidobacterium adolescentis*
1	*MCM6* (intronic)	rs182549	2.32×10^{-13}	c. Actinobacteria
1	*MCM6* (intronic)	rs182549	2.32×10^{-13}	p. Actinobacteria
2	*ABO* (intronic)	rs550057	9.34×10^{-11}	g. *Collinsella*
2	*ABO* (intronic)	rs550057	3.01×10^{-9}	f. *Coriobacteriaceae*
2	*ABO* (intronic)	rs550057	3.03×10^{-9}	o. Coriobacteriales
2	*ABO* (intronic)	rs550057	2.05×10^{-8}	s. *Collinsella aerofaciens*
3	*SGCZ* (intron)	rs28450200	1.37×10^{-8}	s. *Alistipes sp AP11*
4	*COL23A1* (intron)	rs73344869	2.10×10^{-8}	f. *Rikenellaceae*
4	*COL23A1* (intron)	rs73344869	2.14×10^{-8}	g. *Alistipes*
1	*RH3DM1* (missense)	rs1446585	3.18×10^{-8}	s. *Bifidobacterium longum*
5	*TNFAIP2* (intronic); *LINC0067* (intronic)	rs2282037	4.10×10^{-8}	c. Negativicutes
5	*TNFAIP2* (intronic); *LINC0067* (intronic)	rs2282037	4.10×10^{-8}	o. Selenomonadales
6	*DMD* (intron)	rs921896	4.15×10^{-8}	s. *Escherichia* unclassified

Locus ID	Genes (variant function)	ID	pval	Trait
2	*ABO* (intronic)	rs507666	5.38×10^{-15}	LACTOSECAT-PWY: lactose and galactose degradation I
7	Intergenic	rs7513344	1.79×10^{-9}	PWY-6151: S-adenosyl L methionine cycle I
8	*AC105942.1* (intronic)	rs10159356	7.52×10^{-9}	PWY-REDCITCYC: TCA cycle VIII (*Helicobacter*)
9	*WWOX* (intron)	rs9927590	9.28×10^{-9}	PWY-5088: L-glutamate degradation VIII to propanoate
10	*SALL4* (intron)	rs79676117	1.16×10^{-8}	PWY-7234: inosine-5'-phosphate biosynthesis III
11	*AC010149.1* (intron)	rs11898768	2.00×10^{-8}	PWY-6708: ubiquinol-8 biosynthesis
12	*INSIG*; *BLACE* (intergenic)	rs36182340	2.00×10^{-8}	P105-PWY: TCA cycle IV (2-oxoglutarate decarboxylase)
13	*LINC02851* (upstream)	rs820496	2.43×10^{-8}	ASPASN-PWY: superpathway of L-aspartate and L-asparagine biosynthesis
14	*SYNE2* (intronic)	rs17824268	2.63×10^{-8}	PWY-6612: superpathway of tetrahydrofolate biosynthesis
15	*NUMB* (intronic)	rs148280750	2.93×10^{-8}	PYRIDNUCSYN-PWY: NAD biosynthesis I from aspartate
16	*TSC22D2* (intergenic)	rs1584586	3.47×10^{-8}	PWY-841: superpathway of purine nucleotides de novo biosynthesis I
17	*RCAN2* (intron)	rs59657730	3.50×10^{-8}	GLYCOCAT.PWY: glycogen degradation I (bacterial)
18	*CNN3*; *ALG14* (intergenic)	rs12137024	3.53×10^{-8}	HISDEG-PWY: L-histidine degradation I
19	*AL451048.1* (intron)	rs11795855	3.59×10^{-8}	ENTBACSYN-PWY: enterobactin biosynthesis
20	*CDK11B* (upstream)	rs11486024	3.89×10^{-8}	TRPSYN PWY: L-tryptophan biosynthesis
21	*SLC25A37* (intron)	rs2137304	4.26×10^{-8}	PWY-6353: purine nucleotides degradation II (aerobic)
22	*RPRD2* (intergenic)	rs78778914	4.34×10^{-8}	HISTSYN PWY: L-histidine biosynthesis
23	*GRAMD1B* (intronic)	rs2714053	4.85×10^{-8}	PWY-6147: 6-hydroxymethyl dihydropterin diphosphate biosynthesis I
24	*IL4*; *KIF3A* (intergenic)	rs11242123	4.89×10^{-8}	PWY-7209: superpathway of pyrimidine ribonucleosides degradation

Fig. 1 | Genome-wide association scan results. Manhattan plot of host genomic associations with bacterial taxa and bacterial pathway abundances with at least one genome-wide significant association ($P < 5 \times 10^{-8}$). The *y* axis shows the −log10 transformation of the association *P* value observed at each tested variant. The *x* axis shows the genomic position of variants. The thresholds of study-wide ($P = 1.89 \times 10^{-10}$) and genome-wide ($P = 5 \times 10^{-8}$) significance are shown with horizontal lines. Independent SNP–trait associations reaching genome-wide significance are listed in the tables and labeled on the Manhattan plot. The colors of associated hits indicate whether they represent an association with taxonomy or a bacterial pathway, as indicated in the key within the image. c., class; f., family; g., genus; p., phylum; o., order; s., species.

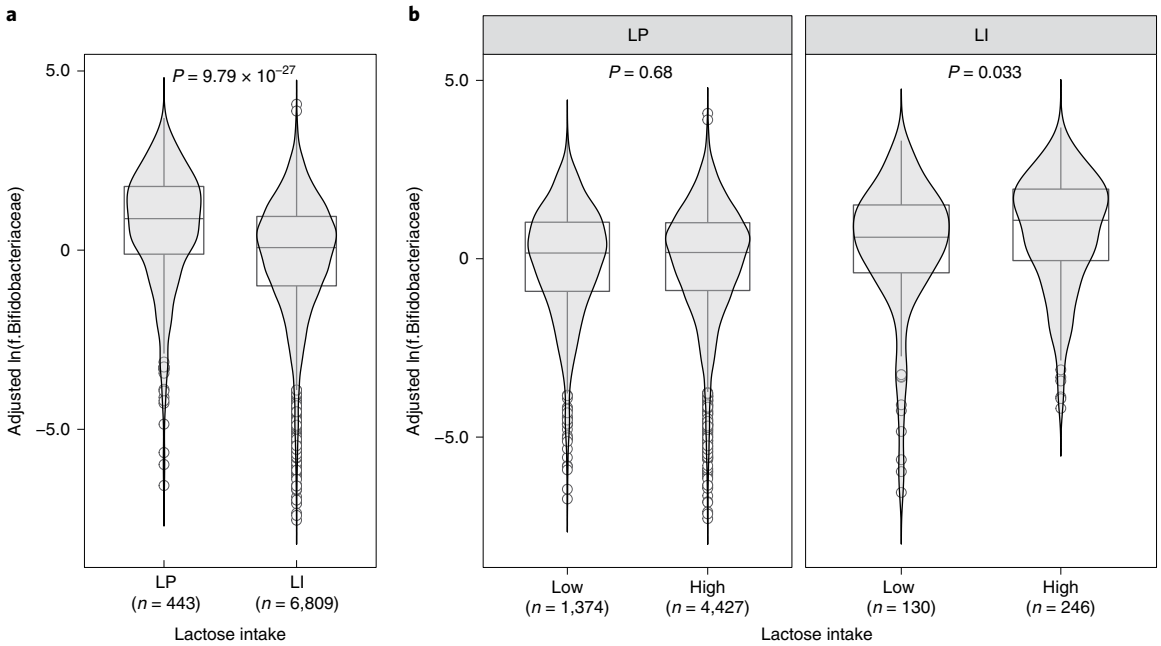


Fig. 2 | Association at the LCT locus and interaction with lactose intake. a,**b**, Comparison of f. Bifidobacteriaceae relative abundance between groups of LP (rs182549 C/T or T/T) and LI (rs182549 C/C) participants (**a**) and stratified among individuals with low or high daily lactose intake levels (**b**). Lactose intake was corrected for daily calorie consumption. The *y* axis represents the relative abundance of the microbial feature, natural log-transformed and adjusted by age and sex. Density distribution is displayed with violin plots, whereas boxplots represent summary statistics; the center line represents the median, the box hinges represent the lower and upper quartiles (percentiles 25 and 75) of the distribution, the upper whisker extends to the maximum value no further than 1.5× interquartile range (IQR) from the upper hinge, the lower whisker extends to the minimum value no further than 1.5× IQR from the lower hinge and data beyond the end of the whiskers are outliers plotted as individual points. Lactose intake levels were defined as low if less than the first quartile and high if greater than or equal to the first quartile. Lactose intake was only available for 5,801 of the 6,809 LP participants and 376 of the 443 LI participants. *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 was not significant (Extended Data Fig. 2). *n*, number of participants.

Genetic associations may be modulated by diet. Gut microbiome composition and function are known to be affected by several factors, including sex, body mass index (BMI), diet and medication usage[4]. None of our 37 SNP–trait genome-wide significant associations were attenuated when including BMI, medication usage, stool frequency or stool consistency as covariates (Methods; for all comparisons, Cochran's *Q* for difference in effect size $P > 0.05$), indicating that these associations are independent and not confounded by these factors (Supplementary Table 4a). Furthermore, none showed evidence for being a sex-specific effect, although five 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, considering the dominant inheritance model at *LCT* and the observed dependence on secretor status at *ABO*. We considered dietary factors previously associated (false discovery rate (FDR) < 0.05) with microbial taxa and pathway abundances that show genome-wide association signals in the *ABO* and *LCT* loci (Methods)[16]. In an analysis that included age, sex and 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 with *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 taxon 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 $P = 0.03$) (Fig. 3 and Supplementary Table 6), a finding that is consistent with previous reports[8,18]. In contrast, there was no evidence for interaction with diet at the *ABO* locus (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[18].

Taxa and pathways genetic signatures are likely polygenic. None of the 22 other loci that showed suggestive association at $P < 5 \times 10^{-8}$ (Supplementary Table 2) were reported previously. The majority (18 loci) 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. The associated regions harbor genes and variants associated with metabolic and immune phenotypes, thus providing intriguing links with microbiome and diseases, as described in Supplementary Note.

We sought to replicate these suggestive 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

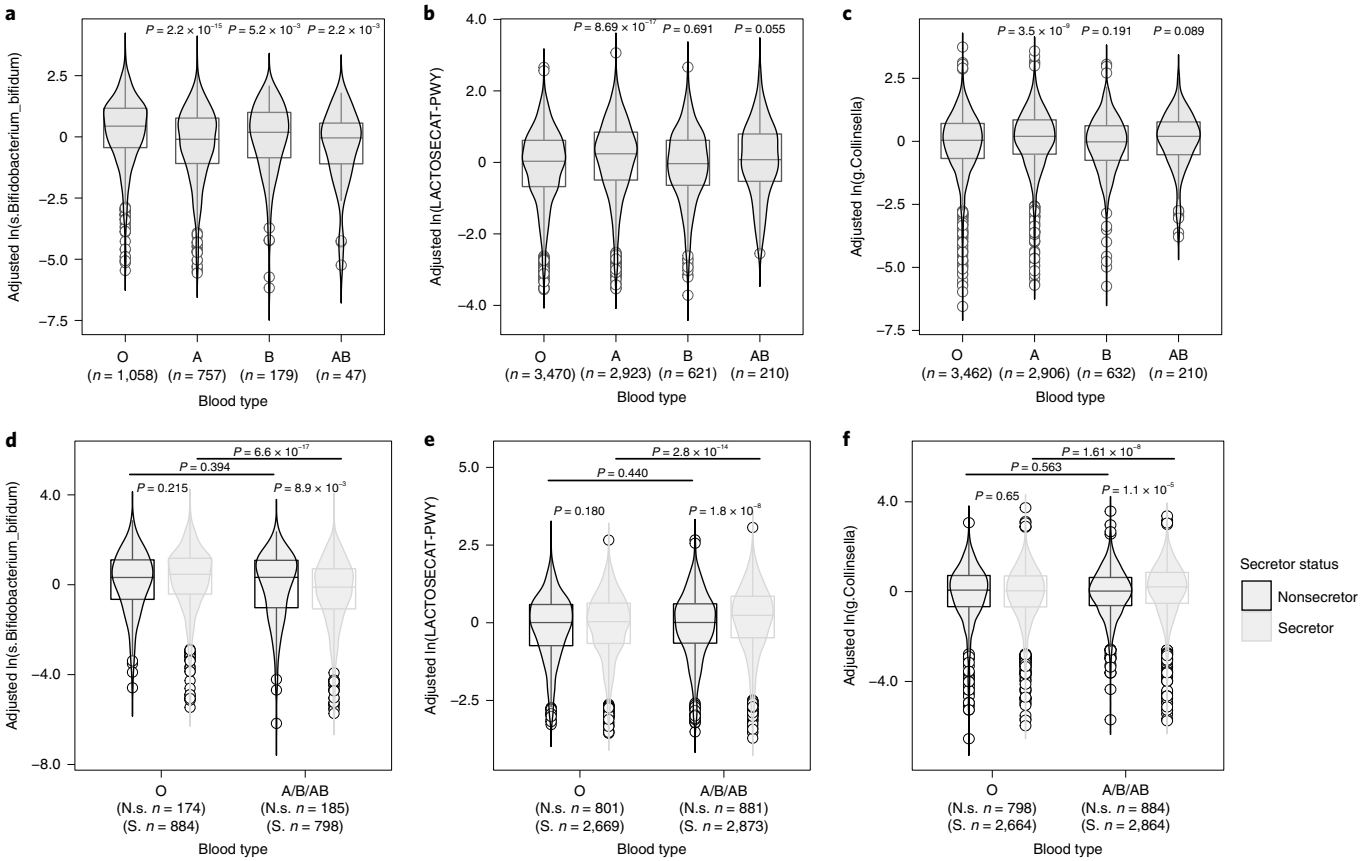


Fig. 3 | Association with blood types and interaction with *FUT2*. a–f, Comparison of the relative abundances of microbiome features associated with SNPs at the *ABO* locus between the inferred blood types (O, A, B and AB), with O as the reference group (**a**–**c**), and between secretors and nonsecretors, stratified again by the blood types grouped by the presence (A/B/AB) or absence (O) of terminal sugars (**d**–**f**). The *y* axis represents the relative abundance of the microbial feature, natural log-transformed and adjusted by age and sex. Density distribution is displayed with violin plots, whereas boxplots represent summary statistics; the center line represents the median, the box hinges represent the lower and upper quartiles (percentiles 25 and 75) of the distribution, the upper whisker extends to the maximum value no further than 1.5× IQR from the upper hinge, the lower whisker extends to the minimum value no further than 1.5× IQR from the lower hinge and data beyond the end of the whiskers are outliers plotted as individual points. In the top panels, the distribution across all blood types is also significantly different for all features (ANOVA: $P=6.7\times10^{-14}$ for species *B. bifidum*, $P=5.4\times10^{-9}$ for LACTOSECAT-PWY and $P=1.4\times10^{-13}$ for genus *Collinsella*). *P* values shown at the top were obtained from a two-sided Wilcoxon ranked test. S., secretors; N.s., nonsecretors; *n*, number of participants; s., species; g., genus.

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 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 $P<3.7\times10^{-6}$). For the *ABO* locus, we found significant replication for the *Collinsella* genus ($P<2\times10^{-5}$ in MiBioGen) and replication at only nominal significance for the bacterial pathway LACTOSECAT-PWY and *B. bifidum* species ($P<0.05$ in LL-DEEP) (Supplementary Tables 7 and 8). The association at *ABO* with *B. bifidum* does not reach the multiple-testing–adjusted threshold for replication. Therefore, although 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 ($P=2.4\times10^{-5}$) and genus *Alistipes* ($P=9.3\times10^{-6}$).

To explore whether the association signals at lower levels of significance are enriched in heritable bacteria, which would indicate if it is 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 ($r_S=0.248$, $P=3.1\times10^{-4}$) and pathway ($r_S=0.263$, $P=1.3\times10^{-4}$) heritability with the number of suggestive ($P<1\times10^{-5}$) loci identified in our GWAS (Fig. 4). The correlation for pathways remained significant when increasing the mbQTL threshold to $P<5\times10^{-4}$ and removing the *LCT* and *ABO* loci from the analyses ($r_S=0.541$, $P=5.3\times10^{-17}$).

We also evaluated whether we could replicate any of the association signals outside the *LCT* and *ABO* loci that were reported in a recent and similarly sized Finnish population study[18]. After extracting all associations with $P<1\times10^{-4}$ 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;

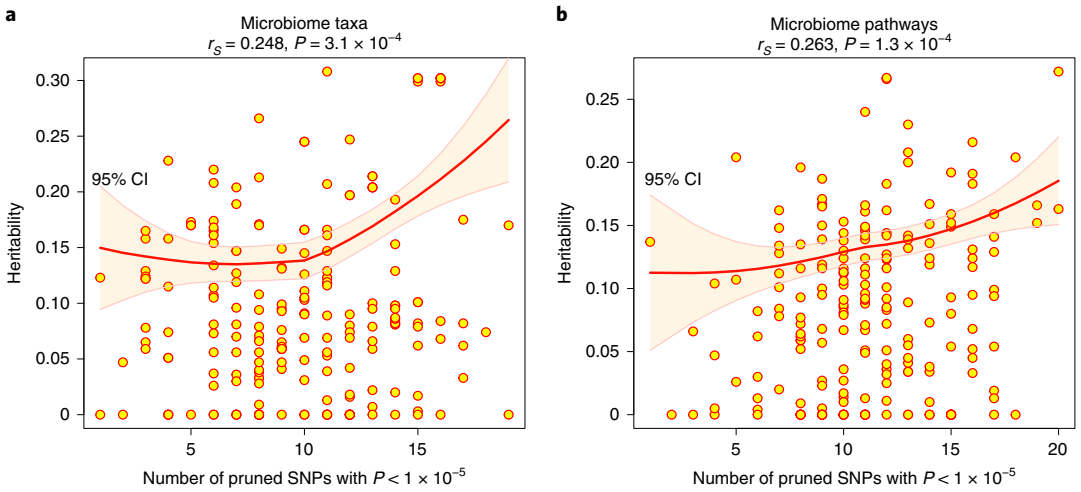


Fig. 4 | Weighted Spearman correlation between estimated heritability and number of suggestive loci. Each point represents one taxon or pathway. The bold line at the center of the bands represents the locally estimated scatterplot smoothing fit, whereas the bands indicate the 95% regression confidence interval (CI). The number of independent suggestive mbQTLs was obtained using linkage disequilibrium (LD) pruning (Methods). **a**, Weighted Spearman correlation of the number of suggestive mbQTLs and family-based heritability for microbiome taxa. **b**, Weighted Spearman correlation of number of suggestive mbQTLs and family-based heritability for microbiome pathways.

the minor allele at this variant (through proxy SNP rs632222) was associated with a decreased abundance of the *Desulfovibrio* genus ($P = 8.7 \times 10^{-5}$) in our study (Supplementary Table 9). This corroborates the allelic effects seen in the Finnish study for the abundances of phylum Desulfobacterota A, class Desulfovibrionia, order Desulfovibrionales and family Desulfovibrionaceae.

Taxa abundances may modulate salt intake and triglycerides. 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 bidirectional two-sample MR analyses (Methods). We analyzed 78 phenotypes representing autoimmune diseases, cardiometabolic diseases and related risk factors, as well as food preferences (Supplementary Table 10) and the 37 microbiome features associated with at least one variant at genome-wide significance in our GWAS. None of the causal relationships were significant at FDR < 0.05. At FDR < 0.1, we observed three causal relationships in the direction from microbiome to phenotypes, suggesting that variation in microbiome abundance can influence salt intake and triglyceride levels (Supplementary Note, Supplementary Table 11 and Extended Data Fig. 3). There was no evidence for these relationships being affected by pleiotropy, and results were very similar when we performed a polygenic risk score analysis in the UK Biobank cohort (Supplementary Table 12 and Supplementary Note).

Very large cohorts are necessary in future studies. Analysis of core microbiota in different populations indicates that only a few bacteria were present in >95% of studied individuals[5,13,16], drastically reducing the effective sample sizes for analyses. We performed power calculations taking into account the interindividual variations in microbiome composition and concluded that a sample size comparable to our study (~8,000 participants) is only sufficient to identify associations with taxa present in >80% individuals, which comprise only 7% of all identified taxa in the cohort (Methods and Supplementary Note). For bacteria present in >20% of the samples, >50,000 participants are necessary to identify an effect size similar to that of *LCT* and *ABO* (Fig. 5).

Discussion

We carried out the largest GWAS 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 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 ($P < 5 \times 10^{-8}$) at 22 other loci for four taxa and 18 bacterial pathways. None were affected by major confounders of the gut microbiome such as medication usage, diet and BMI. Finally, we used an MR approach to pinpoint causal links among gut microbiome composition, complex traits and food intake habits.

The association between the *LCT* locus and the gut microbiome remains the most robust genetic association identified to date. Associations at *LCT* 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[6,8,13,14,22]. We recapitulate that an increase of *Bifidobacterium* was more evident in LI individuals who consume milk or milk-derived products[13,25]. In addition, given that the resolution of metagenomic sequencing allows for species-level characterization of microbiome profiles, we showed that this effect was mainly attributable to the species *B. longum, B. adolescentis, B. catenulatum* and *B. bifidum*, which was also corroborated by two recent studies in Finnish and US Hispanic/Latino populations[18,26].

Another study-wide association was at the *ABO* locus. Associations with microbiome composition and blood types were observed in previous experimental studies[27,28]. Genetic associations at *ABO* have been reported previously in populations of different ethnicities[22,29] and in nonhuman species, including pigs[30]. A deletion at this locus that inactivates the *ABO* acetylgalactosaminyltransferase has been shown to change porcine microbiome composition by altering intestinal *N*-acetylgalactosamine concentrations and consequently reducing the abundance of *Erysipelotrichaceae* strains, which have the capacity to import and catabolize *N*-acetylgalactosamine[31]. We did not detect any evidence of interaction with diet at the *ABO* locus, although this could be due to limitations in available information, as the recording of dietary

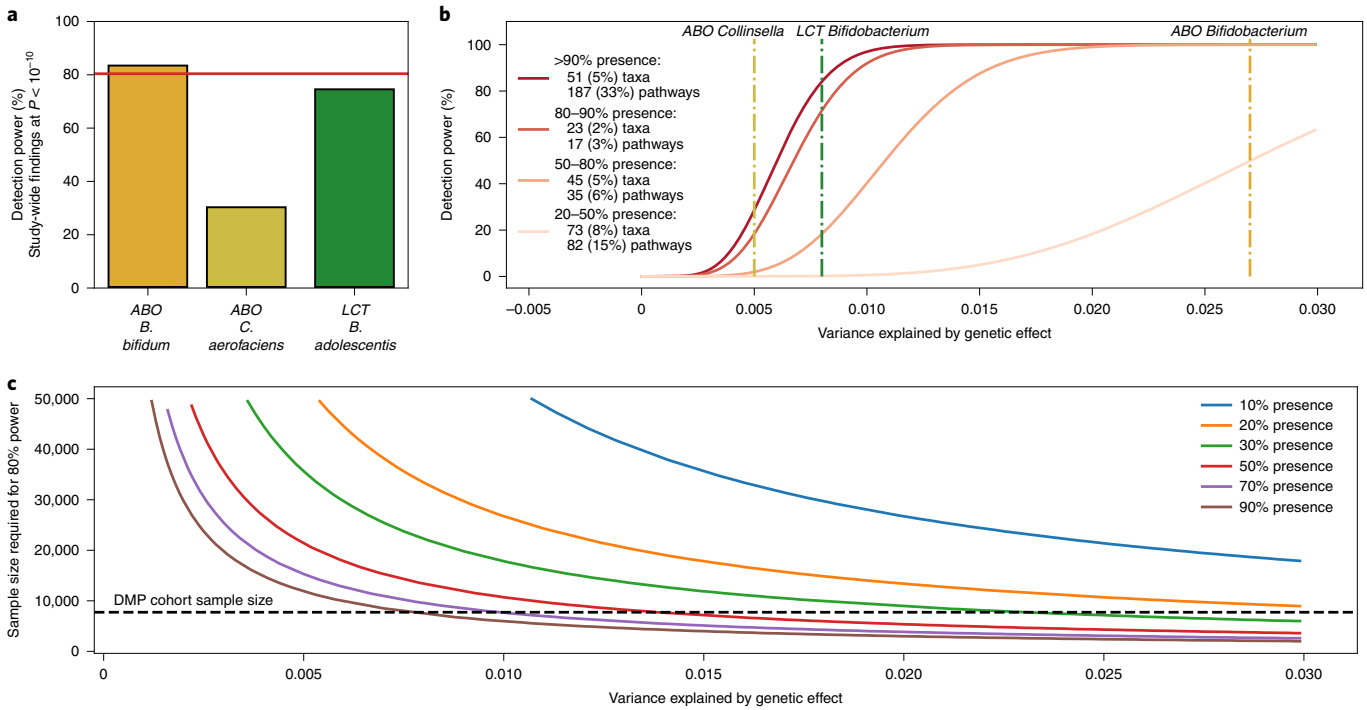


Fig. 5 | Power analysis taking into account bacterial presence levels. a, Statistical power (at alpha = 1×10^{-10}) in our dataset for SNPs at the *ABO* and *LCT* loci, considering the associated taxonomies. Red horizonal line shows 80% detection power. **b**, Power (*y* axis) to detect an effect at significance level alpha = 1×10^{-10} depending on the variance explained by a genetic variant (*x* axis). Colored lines distinguish power under different levels of bacterial or pathway prevalence, which reduces effective sample size (lines correspond to lower bounds of prevalence ranges). Vertical dashed lines indicate effect size of associations at the *ABO* and *LCT* loci and are colored as in **a**. Prevalence, absolute numbers and percentages of taxa and pathways are estimated in the DMP (ref. [16]). **c**, Sample size needed to identify an association at 80% power level across different bacterial presence levels and alpha = 1×10^{-10}. Horizontal dashed line indicates the sample size of the DMP cohort.

information and stool collection were done at different times. We did, however, find that associations at this locus depend on secretor status that is determined by a nonsense mutation at the *FUT2* gene[22] 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 *ABO* also showed evidence of association at *LCT* independently of blood type (interaction $P > 0.05$ for all taxa), indicating a common, independent action of these two loci in contributing to the growth of the associated bacteria. The most compelling hypothesis is that the availability of sugars in the gut, via undigested lactose in LI individuals or secretion of antigens with accessible glycans in non–O blood-type secretors, provides direct energy sources for these bacteria. This is further supported by the observation that LI 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 Fig. 4), would be subjected to competition in the environment, as is the case of non–O blood secretors. Of note, a similar pattern of association at this locus was found in a recent study[22], 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. Although 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. Although we have shown that in general the microbiome remains fairly stable in an individual after 4 years, with interindividual differences being larger than interindividual differences[32], short-term changes in diet, especially those introducing drastic shifts, can perturbate microbiome composition and function. These cannot be taken in account by our analyses. Capturing these shifts would not be easy even with frequency food questionnaires recorded at 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 highly differentiated among human populations due to positive selection of the lactase-persistence phenotype. It has been estimated that strong selection occurred within the past 5,000–10,000 years, consistent with an advantage of lactase persistence and the ability to digest milk in the setting of dairy farming[33]. Variants at this locus have been linked, through GWASs, to not only food habits and metabolic phenotypes but also immune cell populations[34]. The *ABO* locus is evolutionarily highly differentiated; it has been shown to have experienced balancing selection in the last three million years in many primate species[35]. Several evolutionary sources of selective pressure have been proposed, including via infections by pathogens such as malaria[36] and cholera[37]. *ABO* variants have also been linked to cardiometabolic traits, cytokine levels and white and red blood cell levels[38,39]. Therefore, host–microbiome interactions are likely shaped by human–microbe coevolution and survival, probably through a balance between food availability for gut bacteria and enhanced immune response of the host. Better understanding of these interactions will expand our current knowledge of human evolution[40]. From this perspective, it will be crucial to compare genetic studies

of the gut microbiome in diverse populations with different genetic backgrounds. 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[41].

To explore the causality of the relations of the microbiome with complex traits and food preferences, we performed bidirectional MR analysis using 56 dietary traits, 16 diseases and 5 biomarkers. At FDR < 0.1, we observed that a genetically determined increase in the abundance of genus *Alistipes* and its family Rikenellaceae led to decreased consumption of salt, although the limited impact of genetic variants on both microbiome composition and dietary preferences requires caution when interpreting causality estimation by MR[42,43]. Although it is known that dietary changes have a strong effect on microbiome composition[44], it is intriguing to suggest that genetically determined variation of the microbiome might affect food preferences. This is supported by bacterial genetic variations in salt tolerance[45] and the established knowledge that the composition of gut microbiome can predict the effect of food items on host metabolism[46]. There is additional evidence supporting a role for the microbiome in influencing food preferences. A perfect proxy of rs642387 (rs503397, $r^2 = 0.99$ in 1000 Genomes Project Europeans) was recently reported to be associated with the family Desulfovibrionaceae and related taxa in a Finnish population and replicated in our study and has been associated with bitter alcoholic beverage consumption in an independent cohort[47]. Another study found an increase in family Desulfovibrionaceae in individuals with high alcohol consumption[48], and family Desulfovibrionaceae, genus *Desulfovibrio* and other related taxa were also associated with increased consumption of alcohol in our DMP cohort[16], supporting a pleiotropic effect of this locus on both microbiome composition and alcohol intake. Combined with our findings, this suggests that gut microbiota could influence an individual's food preferences by mediating the downstream effect of the consumption of different products.

In addition to the two study-wide significant loci, we also observed 22 loci at genome-wide significance and many more at more lenient thresholds. The observed correlation 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. These loci remain under the detection limit in this study. According to our power estimates, sample sizes would need to be increased by orders of magnitude to elucidate the genetic architecture of microbiome traits, especially of rarer bacteria. 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[49–51].

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/s41588-021-00992-y.

Received: 1 December 2020; Accepted: 19 November 2021;
Published online: 3 February 2022

References

1. Valdes, A. M., Walter, J., Segal, E. & Spector, T. D. Role of the gut microbiota in nutrition and health. *BMJ* **361**, k2179 (2018).
2. Hall, A. B., Tolonen, A. C. & Xavier, R. J. Human genetic variation and the gut microbiome in disease. *Nat. Rev. Genet.* **18**, 690–699 (2017).
3. Fan, Y. & Pederson, O. Gut microbiota in human metabolic health and disease.*Nat. Rev. Microbiol.* **19**, 55–71 (2020).
4. Zhernakova, A. et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).
5. Falony, G. et al. Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).
6. Goodrich, J. K. et al. Genetic determinants of the gut microbiome in UK twins. *Cell Host Microbe* **19**, 731–743 (2016).
7. Rothschild, D. et al. Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).
8. Bonder, M. J. et al. The effect of host genetics on the gut microbiome. *Nat. Genet.* **48**, 1407–1412 (2016).
9. Wang, J. et al. Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. *Nat. Genet.* **48**, 1396–1406 (2016).
10. Turpin, W. et al. Association of host genome with intestinal microbial composition in a large healthy cohort. *Nat. Genet.* **48**, 1413–1417 (2016).
11. Hughes, D. A. et al. Genome-wide associations of human gut microbiome variation and implications for causal inference analyses. *Nat. Microbiol.* **5**, 1079–1087 (2020).
12. Kurilshikov, A., Wijmenga, C., Fu, J. & Zhernakova, A. Host genetics and gut microbiome: challenges and perspectives. *Trends Immunol.* **38**, 633–647 (2017).
13. Kurilshikov, A. et al. Large-scale association analyses identify host factors influencing human gut microbiome composition. *Nat. Genet.* **53**, 156–165 (2021).
14. Blekhman, R. et al. Host genetic variation impacts microbiome composition across human body sites. *Genome Biol.* **16**, 191 (2015).
15. Vieira-Silva, S. et al. Species–function relationships shape ecological properties of the human gut microbiome. *Nat. Microbiol.* **1**, 1–8 (2016).
16. Gacesa, R. et al. The Dutch Microbiome Project defines factors that shape the healthy gut microbiome. Preprint at *bioRxiv* https://doi.org/10.1101/2020.11.27.401125 (2020).
17. Sanna, S. et al. Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nat. Genet.* **51**, 600–605 (2019).
18. Qin, Y. et al. Combined effects of host genetics and diet on human gut microbiota and incident disease in a single population cohort. Preprint at *medRxiv* https://doi.org/10.1101/2020.09.12.20193045 (2020).
19. Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
20. Liu, X. et al. A genome-wide association study for gut metagenome in Chinese adults illuminates complex diseases. *Cell Discov.* **7**, 9 (2021).
21. Van Der Heide, H. M., Magnee, W. & Van Loghem, J. J. Blood group frequencies in the Netherlands. *Am. J. Hum. Genet.* **3**, 344–347 (1951).
22. Rühlemann, M. C. et al. Genome-wide association study in 8,956 German individuals identifies influence of ABO histo-blood groups on gut microbiome. *Nat. Genet.* **53**, 147–155 (2021).
23. Turroni, F. et al. Bifidobacterium bifidum as an example of a specialized human gut commensal. *Front. Microbiol.* **5**, 437 (2014).
24. Turroni, F., Milani, C., van Sinderen, D. & Ventura, M. Genetic strategies for mucin metabolism in *Bifidobacterium bifidum* PRL2010: an example of possible human-microbe co-evolution. *Gut Microbes* **2**, 183–189 (2011).
25. Bonder, M. J. et al. Genetic and epigenetic regulation of gene expression in fetal and adult human livers. *BMC Genomics* **15**, 860 (2014).
26. Moon, Jee-Young et al. Milk intake, host LCT genotype and gut Bifidobacteria in relation to obesity: results from the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). *Circulation* **141**, AP459 (2020).
27. Arnolds, K. L., Martin, C. G. & Lozupone, C. A. Blood type and the microbiome: untangling a complex relationship with lessons from pathogens. *Curr. Opin. Microbiol.* **56**, 59–66 (2020).
28. Mäkivuokko, H. et al. Association between the ABO blood group and the human intestinal microbiota composition. *BMC Microbiol.* **12**, 94 (2012).
29. Liu, X. et al. Inter-determination of blood metabolite levels and gut microbiome supported by Mendelian randomization. Preprint at *bioRxiv* https://doi.org/10.1101/2020.06.30.181438 (2020).
30. Motta, V., Luise, D., Bosi, P. & Trevisi, P. Faecal microbiota shift during weaning transition in piglets and evaluation of AO blood types as shaping factor for the bacterial community profile. *PLoS One* **14**, e0217001 (2019).
31. Yang, H., et al. An ancient deletion in the ABO gene affects the composition of the porcine microbiome by altering intestinal N-acetyl-galactosamine concentrations. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/2020.07.16.206219v1 (2020).
32. Chen, L. et al. The long-term genetic stability and individual specificity of the human gut microbiome. *Cell* **184**, 2302–2315 (2021).

33. Bersaglieri, T. et al. Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74**, 1111–1120 (2004).
34. Auer, P. L. et al. Imputation of exome sequence variants into population- based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project. *Am. J. Hum. Genet.* **91**, 794–808 (2012).
35. Ségurel, L. et al. The ABO blood group is a trans-species polymorphism in primates. *Proc. Natl Acad. Sci. USA* **109**, 18493–18498 (2012).
36. Band, G. et al. Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. *Nat. Commun.* **10**, 5732 (2019).
37. Barua, D. & Paguio, A. S. ABO blood groups and cholera. *Ann. Hum. Biol.* **4**, 489–492 (1977).
38. Watanabe, K. et al. A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).
39. Naitza, S. et al. A genome-wide association scan on the levels of markers of inflammation in sardinians reveals associations that underpin its complex regulation. *PLoS Genet.* **8**, e1002480 (2012).
40. Suzuki, T. A. et al. The role of the microbiota in human genetic adaptation. *Science* **370**, eaaz6827 (2020).
41. Parks, D. H. et al. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* **38**, 1079–1086 (2020).
42. Wade, K. H. & Hall, L. J. Improving causality in microbiome research: can human genetic epidemiology help? *Wellcome Open Res.* **4**, 199 (2020).
43. Pirastu, N. et al. Using genetics to disentangle the complex relationship between food choices and health status. Preprint at *bioRxiv* https://doi.org/10.1101/829952 (2019).
44. Wang, C. et al. High-salt diet has a certain impact on protein digestion and gut microbiota: a sequencing and proteome combined study. *Front. Microbiol.* **8**, 1838 (2017).
45. Culligan, E. P. et al. Combined metagenomic and phenomic approaches identify a novel salt tolerance gene from the human gut microbiome. *Front. Microbiol.* **5**, 189 (2014).
46. Zeevi, D. et al. Personalized nutrition by prediction of glycemic responses. *Cell* **163**, 1079–1094 (2015).
47. Zhong, V. W. et al. A genome-wide association study of bitter and sweet beverage consumption. *Hum. Mol. Genet.* **28**, 2449–2457 (2019).
48. Bjørkhaug, S. T. et al. Characterization of gut microbiota composition and functions in patients with chronic alcohol overconsumption. *Gut Microbes* **10**, 663–675 (2019).
49. Sanna, S. et al. Common variants in the GDF5-UQCC region are associated with variation in human height. *Nat. Genet.* **40**, 198–203 (2008).
50. Weedon, M. N. et al. A common variant of HMGA2 is associated with adult and childhood height in the general population. *Nat. Genet.* **39**, 1245–1250 (2007).
51. Loos, R. J. F. et al. Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat. Genet.* **40**, 768–775 (2008).

LifeLines Cohort Study

Raul Aguirre-Gamboa[1], Patrick Deelen[1], Lude Franke[1], Jan A. Kuivenhoven[3], Esteban A. Lopera-Maya[1], Ilja M. Nolte[7], Serena Sanna[1,6], Harold Snieder[7], Morris A. Swertz[1,4], Judith M. Vonk[7] and Cisca Wijmenga[1]

[7] Department of Epidemiology, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands.

Methods

Cohort description. LifeLines is a multidisciplinary 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, sociodemographic, behavioral, physical and psychological factors that contribute to the health and disease of the general population, with a special focus on multimorbidity and complex genetics[52,53]. During the first follow-up visit, all participants were invited to participate in a parallel project, the DMP, on a voluntary basis. The goal of this project is to evaluate the impact of different exposures and lifestyles on gut microbiota composition[16]. A subset of 10,000 LifeLines participants were included; for 8,719 of these participants, both feces and phenotype information were collected. Of these, samples from 8,208 participants were retained for downstream analysis after stringent quality control[16]. Distribution of age, gender and location within the three north provinces was similar to that observed in the total LifeLines cohort[53].

The LifeLines study was approved by the medical ethical committee from the University Medical Center Groningen (METc number 2017/152). All LifeLines participants signed a written consent prior participation. Additional written consents were signed by the DMP participants or legal representatives for children aged under 18 years. This study complies with all relevant ethical regulations.

Genome characterization. Genotyping of 38,030 LifeLines participants was carried out using the Infinium Global Screening Array 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 panel v.1.1[54], as described elsewhere[55]. To avoid population stratification, we analyzed only European samples. We selected 43,587 genetic markers by applying LD pruning on the genotyped data (sliding window of 1 Mb, LD $r^2 < 0.2$, step = 5) and used them for a principal-component analysis projecting the 1000 Genomes Project (all populations) and GoNL cohorts a population references[56,57]. Europeans were then identified as samples clustering together with European populations according to the first two principal components or <3 s.d. away from the most distant European samples in the reference. In total, 35 non-European participants were removed. Quality-controlled genotype information was obtained for 7,738 of the DMP participants for whom quality-controlled microbiome data and BMI were also available. Of these, 58.1% were females, and ages ranged from 8 to 84 years (mean, 48.5 years). The mean BMI value was 25.58 (range, 13.10 to 63.70).

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, frozen within 15 min after production and 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) 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 DNA Library Prep Kit (Illumina), and samples with higher DNA yield were prepared using NEBNext Ultra II DNA Library Prep Kit (Illumina). Shotgun metagenomic sequencing was carried out using the Illumina HiSeq 2000 platform at Novogene. Metagenomic sequencing data was profiled following methods used in other cohorts, as described previously[16,52]. Low-quality reads (PHRED quality ≤30), adapters and host sequences were removed using KneadData tools v.0.5.1. Taxonomic composition was determined with MetaPhlAn2 v.2.7.2 (ref. [58]). Characterization of biochemical pathways was performed with the HUMAnN2 pipeline v.0.11.1 (ref. [59]), integrated with the UniRef90 v.0.1.1 protein database[60], the ChocoPhlAn pangenome database and the DIAMOND alignment tool v.0.8.22 (ref. [61]). After quality control (where 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 1,000 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 semiquantitative Food Frequency Questionnaire designed and validated by the division of Human Nutrition of Wageningen University as described before in Siebelink et al. and Gacesa et al.[16,62]. The Food Frequency Questionnaire data were 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) with 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 *B. longum* and *B. adolescentis*, given their association with the *LCT* region in our study. Participants with an implausible caloric intake (<800 or >3,934 kcal/d for males and <500 or >2,906 kcal/d for females)[63] 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.0.38 (ref. [19]), with age, sex and the genetic relationship matrix 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 genetic relationship matrix 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, $r^2 < 0.2$, sliding window = 500 kb).

Definition of the study-wide significant threshold. To estimate the number of independent phenotypes assessed, we used principal-component analysis 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 ($5 \times 10^{-8}/264 = 1.89 \times 10^{-10}$).

Association using a 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 SAIGEgds[64] 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.[65]. 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/nonsecretor) locus, we grouped individuals into two groups (non-O blood type and blood type O) to distinguish production or nonproduction of antigens and compared pairs using a two-sided Wilcoxon test. All analyses were done using base R v.3.6.1 (https://www.R-project.org/).

Effects of potential confounders. 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 group of 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 (nonuser) 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 < 0.05[16]. 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/nonsecretor 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 v.3.6.1, other statistical tests were as implemented in packages rstatix v.0.5.0 and ggpubr v.0.3.0 and package RNOmni v.0.7.1 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 16 S 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 DMP (refs. [13,17]). In this study, a different DNA isolation procedure (AllPrep Kit) was used, and taxonomies were defined using the Bracken pipeline, which may explain why *C. aerofaciens* 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 16 S 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 data[18]. We searched for all SNPs outside the *LCT* and *ABO* loci or any proxy ($r^2 > 0.8$) in our dataset and selected all SNP–taxonomy pairs that showed a $P < 1 \times 10^{-4}$ with at least one 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.

Heritability estimates and 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 cohort[16]. 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 1 Mb using Plink v.1.9[66]) for all SNPs at the three different thresholds: $P < 5 \times 10^{-4}$, $P < 1 \times 10^{-4}$ and $P < 5 \times 10^{-5}$. 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 v.1.9.1 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 P value estimates. The regression lines in Fig. 4 were fit using the LOESS (locally estimated scatterplot smoothing) function (base R package v.4.0.3) with span and degree parameters set to 1.

MR analysis. To evaluate potential causal relationships between the gut microbiome and other common traits, we performed MR analyses that combined the summary statistics of the microbiome with publicly available summary statistics on food preferences, autoimmune and cardiovascular diseases and other cardiometabolic 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 database[67].

We performed a bidirectional 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 rs ID, position and alleles with the publicly available summary statistic variants. We then selected instruments using the clump_data() function of the TwoSampleMR package v.0.5.5 (ref.[68]). The publicly available summary statistics were clumped using a P-value threshold of $< 5 \times 10^{-8}$ and otherwise standard settings ($r^2 < 0.001$, 10-Mb window size). Due to the limited statistical significance of the microbiome traits, we performed P value clumping at a less stringent $P < 5 \times 10^{-6}$ threshold. If fewer than three variants were clumped, then we removed the trait combination from analysis.

The MR analysis was done using the TwoSampleMR v.0.5.5 package. We first selected trait combinations that passed the Benjamini–Hochberg FDR threshold of 0.1 (corresponding to $P = 2.805 \times 10^{-5}$) in the inverse-variance weighting test, resulting in one suggestively causal trait combination. We further checked that the trait combinations were unlikely to be driven by pleiotropy based on two criteria: (1) the Egger regression intercept was nominally significant (indicating the presence of horizontal pleiotropy), and (2) the weighted median results were not nominally significant (indicating that no single variant influences the result)[69,70].

Power analysis. We calculated the variance explained at our loci using the formula described in Teslovich et al.[71], which takes into account MAF, effect size, s.e. 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 (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 (1,000, 1,050, …, 50,000) and kept the lowest sample size that had power >80%.

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

Data availability

Raw sequencing microbiome data are available at European Genome-Phenome archive (accession number EGAS00001005027). Genotyping data and participant metadata are not publicly available to protect participants' privacy, and neither can be deposited in public repositories to respect the research agreements in the informed consent. The data can be accessed by all bona-fide researchers with a scientific proposal by contacting the LifeLines Biobank (instructions at https://www.lifelines.nl/researcher/how-to-apply). Researchers will need to fill in an application form that will be reviewed within 2 weeks. If the proposed research complies with LifeLines regulations, such as noncommercial use and warranty of participants' privacy, then researchers will receive a financial offer and a data and material transfer agreement to sign. In general, data will be released within 2 weeks after signing the offer and data and material transfer agreement. The data will be released in a remote system (the LifeLines workspace) running on a high-performance computer cluster to ensure data quality and security. The full GWAS summary statistical data for all 207 taxa and 205 pathways are instead available for direct download at NHGRI-EBI GWAS Catalog (https://www.ebi.ac.uk/gwas/) under the study accession numbers GCST90027446-GCST90027857 (accession numbers for each specific taxa and pathways can be found in Supplementary Table 13) or at https://dutchmicrobiomeproject.molgeniscloud.org. The processed microbiome data (taxonomy and pathway abundance per individual) can also be downloaded after filling in a request form available at the same website and after signing a data access agreement. This study also used the following databases: UniRef90 v.0.1.1 protein database and the ChocoPhlAn pangenome databases available within the Humann2 pipeline (https://huttenhower.sph.harvard.edu/humann2/), the Genome Taxonomy Database (https://gtdb.ecogenomic.org/) and the IEU GWAS database (https://gwas.mrcieu.ac.uk/). All other data supporting the findings of this study are available within the paper and Supplementary Note.

References

52. 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).
53. Scholtens, S. et al. Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int. J. Epidemiol.* **44**, 1172–1180 (2015).
54. the Haplotype Reference Consortium. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
55. Lopera Maya, E. A. et al. Lack of association between genetic variants at ACE2 and TMPRSS2 genes involved in SARS-CoV-2 infection and human quantitative phenotypes. *Front. Genet.* **11**, 613 (2020).
56. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
57. Francioli, L. C. et al. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46**, 818–825 (2014).
58. Truong, D. T. et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
59. Franzosa, E. A. et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).
60. Suzek, B. E. et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinforma. Oxf. Engl.* **31**, 926–932 (2015).
61. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
62. 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).
63. Willett, W. C. *Nutritional Epidemiology* (Oxford Univ. Press, 2012).

64. Zheng, X. et al. SAIGEgds: an efficient statistical tool for large-scale PheWAS with mixed models. *Bioinformatics* **37**, 728–730 (2021).
65. The Severe Covid-19 GWAS Group Genome-wide association study of severe COVID-19 with respiratory failure. *N. Engl. J. Med.* **383**, 1522–1534 (2020).
66. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
67. Elsworth, B. et al. The MRC IEU OpenGWAS data infrastructure. Preprint at *bioRxiv* https://doi.org/10.1101/2020.08.10.244293 (2020).
68. Hemani, G. et al. The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7**, e34408 (2018).
69. Burgess, S. & Thompson, S. G. *Mendelian Randomization Methods for Using Genetic Variants in Causal Estimation* (CRC Press, 2015).
70. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).
71. Teslovich, T. M. et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).

Acknowledgements

We acknowledge the services of the LifeLines Cohort Study, the contributing research centers delivering data to LifeLines and all the study participants. The LifeLines initiative was made possible by subsidy from the Dutch Ministry of Health, Welfare and Sport; the Dutch Ministry of Economic Affairs; the University Medical Center Groningen (UMCG); the University of Groningen (UG) and the Provinces of the North of the Netherlands (Drenthe, Friesland and Groningen). This project was carried out under LifeLines project number OV18_0464. We thank Mathieu Plateel and Jody Geelderloos-Arends for their contribution in genotyping the LifeLines samples, Kate McIntyre for help developing the manuscript, Marije van der Geest for setting up the website for sharing summary statistics and Patrick Deelen for discussion of results. We also thank the UMCG Genomics Coordination Center, the UG Center for Information Technology and their sponsors (BBMRI-NL and TarGet) for storage and computational infrastructure and Novogene for providing gut metagenome sequencing of all DMP samples. Finally, we thank the UK Biobank for making their resource available. Analyses of UK Biobank data described in this work were carried out under project number 48548 to C.W. The generation and management of genotype data for the LifeLines Cohort Study was supported by the UMCG Genetics LifeLines Initiative. Genotyping quality control was supported by UMCG (HAP grant CD017.0031/ronde 2017-2/nr 324). Metagenomics sequencing of the cohort was mainly funded by the CardioVasculair Onderzoek Nederland (CVON) (grant CVON 2012-03) to M. Hofken (who died in 2016), J.F. and A.Z., as well as other grants to R.K.W. and C.W. (listed below). This work was further supported by the collaborative TIMID project LSHM18057-SGF financed by the allowance made available by Top Sector Life Sciences & Health to Samenwerkende Gezondheidsfondsen to stimulate public/private partnerships and cofinancing by health foundations that are part of the Samenwerkende Gezondheidsfondsen (R.K.W.); the the Seerave Foundation (R.K.W.); European Research Council (ERC) starting grant 715772 (S.Z.), consolidator grant 101001678 (J.F.) and advanced grant ERC-671274 (C.W.); Netherlands Organization for Scientific Research VIDI grant 016.178.056 (A.Z.), gravitation grant ExposomeNL 024.004.017 (A.Z.), VICI grant VI.C.202.022 (J.F.), gravitation grant The Netherlands Organ-on-Chip Initiative 024.003.001 (C.W.) and Spinoza award NWOSPI 92-266 (C.W.); CVON grant 2018-27 (A.Z. and J.F.); the EurHealth-1Health INTERREG V A 202085 project (H.J.M.H.); Foundation De Cock-Hadders grant 20:20-13 (L.C.); a joint fellowship from the UMCG and China Scholarship Council (CSC201708320268 to L.C.); and Colciencias fellowship ed.783 (E.A.L.-M.).

Author contributions

E.A.L.-M., A.K. and S.H. performed genetic analyses. A.v.d.G. performed MR and power analyses. A.K., S.H., L.C., A.V.V. and R.G. processed microbiome data. S.A.-S., T.S., V.C., M.A.Y.K., L.A.B. and M.F.B.G. processed samples meta-data. P.B.T.N. and M.A.S. provided resources on the HP computing cluster and assisted with data management. H.J.M.H., C.W., J.F., R.K.W. and A.Z. provided funding resources and designed the DMP. A.Z. and S.S. supervised the study. E.A.L.-M., A.K., A.v.d.G., S.H., S.A.S., A.Z. and S.S. drafted the manuscript. All authors were involved in data interpretation and provided critical input to the manuscript draft.

Competing interests

The authors declare no competing interests.

Additional information

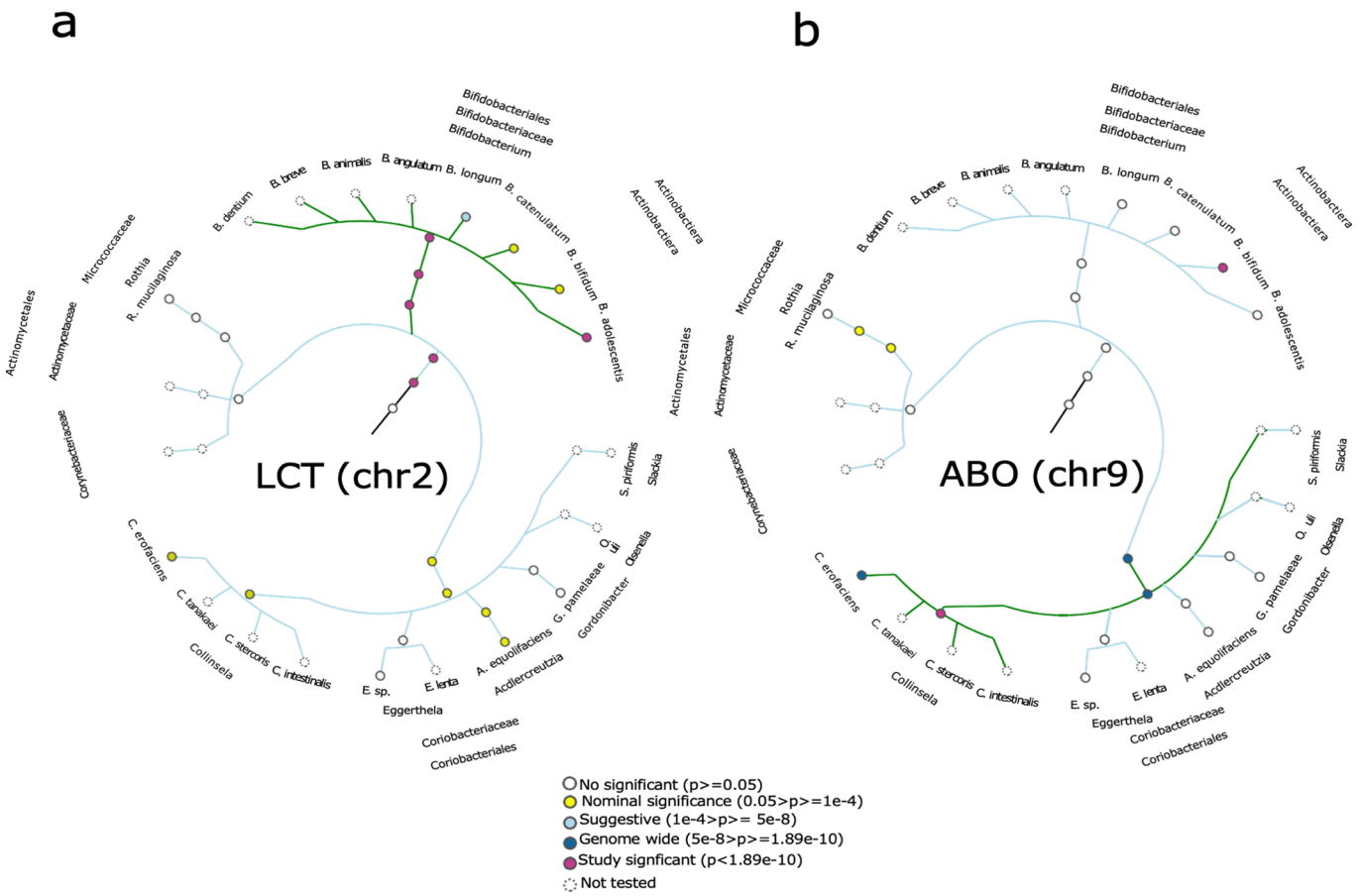
Extended data is available for this paper at https://doi.org/10.1038/s41588-021-00992-y.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00992-y.

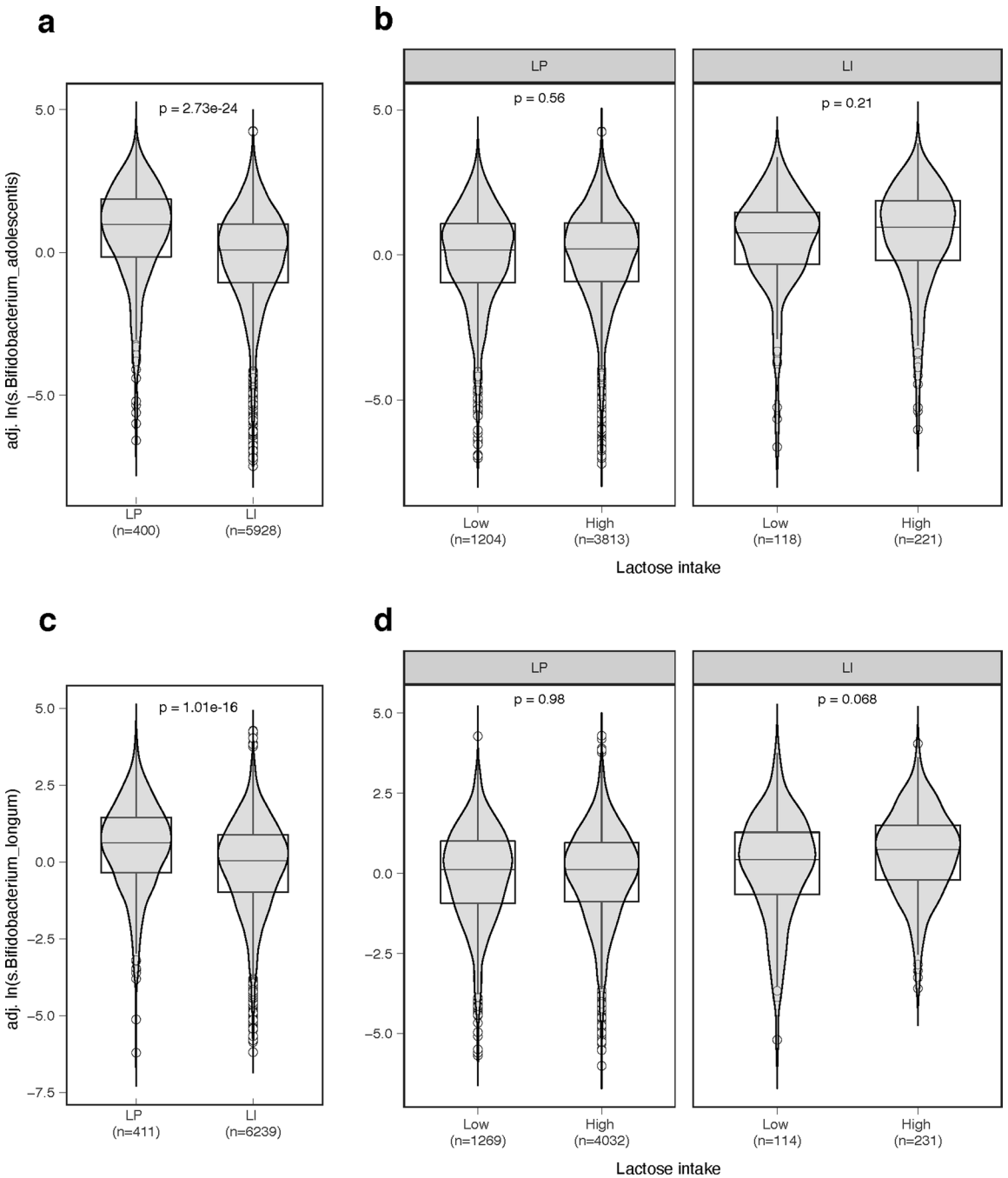
Correspondence and requests for materials should be addressed to Alexandra Zhernakova or Serena Sanna.

Peer review information *Nature Genetics* thanks A. Franke and T. Zhang for their contribution to the peer review of this work.

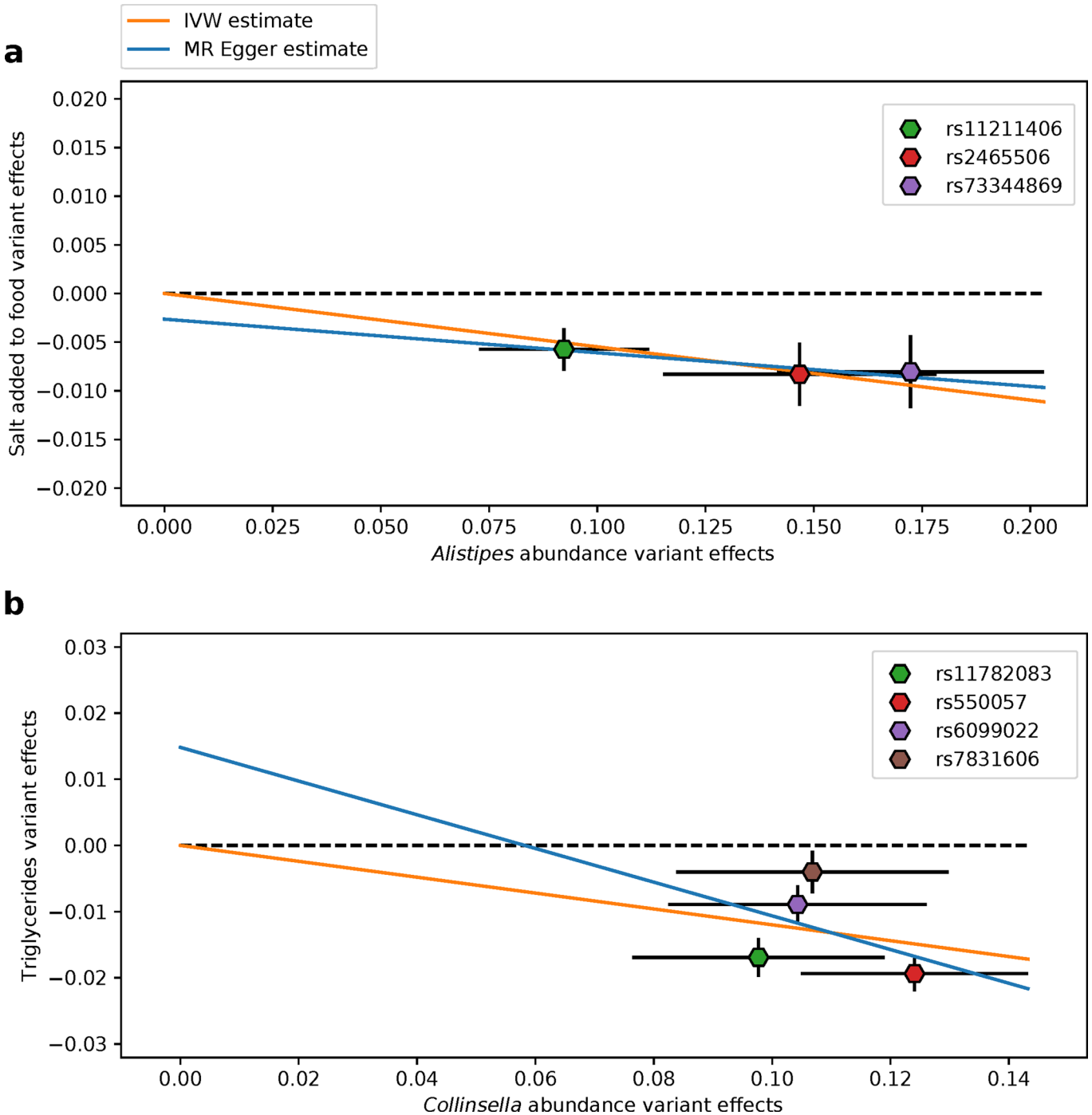
Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | Cladogram plot tree of taxonomic relations between bacteria of the class Actinobacteria and their associations with host genetics. Each node shows a taxonomic level (from outside to inside: phylogenetic group, phylum, class, order, family, genus and species). Note that branch lengths do not represent phylogenetic distance. Inner labels represent genetic locus. External labels represent the clade. Nodes with dotted lines indicate that the GWAS was not performed for that taxa. Node color corresponds to different levels of significance as described in the legend. **a**, Depicts associations detected at the *MCM6/LCT* locus with each taxa, using the most significant p-value observed between rs4988235 and rs182549. **b**, Depicts associations at the *ABO* locus with each taxa, using the most significant p-value observed between rs8176645 and rs550057.



Extended Data Fig. 2 | Association at the *LCT* locus and interaction with lactose intake in other members of family *Bifidobacteriaceae*. Relative abundances of taxa, natural log-transformed and adjusted by age and sex, compared between LP (rs182549 C/T or T/T) and LI (rs182549 C/C) participants and among individuals with low or high daily lactose intake levels. The *y* axis represents the relative abundance of the microbial feature, natural log-transformed and adjusted by age and sex. Density distribution is displayed with violin plots, while boxplots represent summary statistics: the center line represents the median, the box hinges represent the lower and upper quartiles (percentiles 25 and 75) of the distribution, the upper whisker extends to the maximum value no further than 1.5*IQR (where IQR is the interquartile range) from the upper hinge, the lower whisker extends to the minimum value no further than 1.5*IQR from the lower hinge, and data beyond the end of the whiskers are outliers plotted as individual points. **a** and **c**, Relative abundances for the taxa between LP and LI participants. **b** and **d**, Comparisons of abundance between lactose intake levels, low (<first quartile) and high (≥ first quartile), stratified by lactose persistence status. The distributions are shown for *s. Bifidobacterium adolescentis* (top) *and s. Bifidobacterium longum* (bottom). *P*-values were obtained with a two-sided Wilcoxon rank test. n: number of participants.



Extended Data Fig. 3 | Graphical representation of MR results with a Benjamini–Hochberg FDR *q* value < 0.1. a, Effect size in standard deviation units of 3 variants associated with *Alistipes* abundance changes that were used as instrumental variables (effects estimated on 7,728 independent samples) (x-axis) versus effect size in standard deviation units of the same variants for salt intake (estimated effects estimated on 462,630 independent samples) (y-axis). Error bars represent standard errors (SE) of each effect size (beta + SE and beta-SE). The orange and blue lines represent lines whose slope is the causal estimate from MR methods IVW and Egger, respectively. **b**, A plot similar to **a**, but the *x* axis is the effect size in standard deviation units for instrumental variants selected for *Collinsella* (effects estimated on 7,210 independent samples) abundance and on the y-axis for Triglyceride levels (effects estimated on 343,992 independent individuals).

nature research

Corresponding author(s): Serena Sanna

Last updated by author(s): Sep 25, 2021

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	We used genetic and microbiome data for which data processing and QC was already performed in other studies. For analysing microbiome sequencing data we used KneadData tools v0.5.1. For defining taxonomic composition we used MetaPhlAn2 v2.7.2. For the characterization of biochemical pathways was performed with the HUMAnN2 pipeline v0.11.1, integrated with the UniRef90 v0.1.1 protein databas, the ChocoPhlAn pangenome database and the DIAMOND alignment tool v0.8.22 For genome-wide association analyses, we used the software SAIGE v.0.38 . For association models with confounders, and for analyses of blood groups we used the linear function lm() in R (v3.6.1) and packages rstatix v0.5.0 and ggpubr v0.3.0 and package RNOmni v0.7.1 For identifying independent variants and for polygenic risk score calculations we used the software PLINK 1.9. For causal inferences analyses, we used the "two-sample MR" R package (v0.5.5).

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/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

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

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

Raw sequencing microbiome data is available at EGA (accession number EGAS00001005027). Genotyping data and participant metadata are not publicly available to protect participants' privacy, neither can be deposited in public repositories to respect the research agreements in the informed consent. The data can be accessed by all bona-fide researchers with a scientific proposal by contacting the Lifelines Biobank (instructions at: https://www.lifelines.nl/researcher/ how-to-apply). Researchers will need to fill in an application form that will be reviewed within two weeks. If the proposed research complies with Lifelines regulations, such as non-commercial use and warranty of participants' privacy, researchers will receive a financial offer and a Data and Material transfer agreement (DMTA) to sign. In general, data will be released within two weeks after signing the offer and DMTA. The data will be released in a remote system (the Lifelines workspace) running on a high-performance computer cluster to ensure data quality and security. The full GWAS summary statistics for all 207 taxa and 205 pathways are instead available for direct download at GWAS Catalogue Globus ((http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90027001-GCST90028000/; accession numbers for each specific taxa and pathways can be found in Supplementary Table 13) or at https://dutchmicrobiomeproject.molgeniscloud.org. The processed microbiome data (taxonomy and pathway abundance per individual) can also be downloaded after filling in a request form available at the same website and after signing a data access agreement. This study also used the following databases: UniRef90 v0.1.1 protein database and the ChocoPhlAn pangenome databases available within the Humann2 pipeline (https://huttenhower.sph.harvard.edu/humann2/), the Genome Taxonomy Database (https://gtdb.ecogenomic.org/) and the IEU GWAS DATABASE (https://gwas.mrcieu.ac.uk/). 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 for this study was no predetermined; we used all available samples with genetic and gut microbiome information
Data exclusions	We used preQCed genotyping and gut microbiome data. For this study, we further excluded non European samples and samples for which BMI was not available. On the microbiome data we excluded very rare (present in <1000 samples) microbiomal taxa or pathways on this set of samples, as well as highly redundant pathways (Spearman correlation >0.95). We also excluded genetic variants with minor allele frequency <0.05.
Replication	we looked for replication of our estimates in two independent cohorts previously published (the MiBIOGEN and LL-DEEP cohorts)
Randomization	this is a population-cohort study and not an intervention study. Thus randomization is not applicable
Blinding	this is a population-cohort study and not an intervention study. Thus blinding is not applicable

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
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

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	This study used a subset of 8,208 volunteers from the Lifelines population cohort, enrolled in a parallel project: The Dutch Microbiome Project (DMP). QCed genotype and microbiome information was obtained for 7,738 of DMP participants. Of these, 58.1% were females, and with ages ranging from 8 to 84 years (mean 48.5 years)
Recruitment	Volunteers of the Dutch Microbiome Project were recruited independently of this study on voluntary participation after an invitation letter and who were willing to collect stool samples. We acknowledge that this form of recruitment may discourage volunteers that suffer from chronic or debilitating diseases at advanced stage. This bias is unlikely to affect our results who are reflecting an association present at the general population and not specifically correlated to a disease.
Ethics oversight	We did no collected new data for this study. We used available data from the Lifelines cohort which was approved by the medical ethical committee from the University Medical Center Groningen (METc number: 2017/152).

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