Genetics of human gut microbiome composition

Authors: Alexander Kurilshikov[1,†], Carolina Medina-Gomez[2,3,†], Rodrigo Bacigalupe[4,5,†], Djawad Radjabzadeh[2,†], Jun Wang[4,5,6,†], Ayse Demirkan[1,7§], Caroline I. Le Roy[8,§], Juan Antonio Raygoza Garay[9,10,§], Casey T. Finnicum[11,§], Xingrong Liu[12,§], Daria V. Zhernakova[1,13§], Marc Jan Bonder[1,§], Tue H. Hansen[14], Fabian Frost[15], Malte C. Rühlemann[16], Williams Turpin[9,10], Jee-Young Moon[17], Han-Na Kim[18,19], Kreete Lüll[20], Elad Barkan[21], Shiraz A. Shah[22], Myriam Fornage[23,24], Joanna Szopinska-Tokov[25], Zachary D. Wallen[26], Dmitrii Borisevich[14], Lars Agreus[27], Anna Andreasson[28], Corinna Bang[16], Larbi Bedrani[9], Jordana T. Bell[8], Hans Bisgaard[22], Michael Boehnke[29], Dorret I. Boomsma[30], Robert D. Burk[16,31,32], Annique Claringbould[1], Kenneth Croitoru[9,10], Gareth E. Davies[11,30], Cornelia M. van Duijn[33,34], Liesbeth Duijts[3,35], Gwen Falony[4,5], Jingyuan Fu[1,36], Adriaan van der Graaf[1], Torben Hansen[14], Georg Homuth[37], David A. Hughes[38,39], Richard G. Ijzerman[40], Matthew A. Jackson[7,41], Vincent W.V. Jaddoe[3,33], Marie Joossens[4,5], Torben Jørgensen[42], Daniel Keszthelyi[43,44], Rob Knight[45,46,47], Markku Laakso[48], Matthias Laudes[49], Lenore J. Launer[50], Wolfgang Lieb[51], Aldons J. Lusis[52,53], Ad A.M. Masclee[43,44], Henriette A. Moll[35], Zlatan Mujagic[43,44], Qi Qibin[17], Daphna Rothschild[21], Hocheol Shin[54,55], Søren J. Sørensen[56], Claire J. Steves[8], Jonathan Thorsen[22], Nicholas J. Timpson[38,39], Raul Y. Tito[4,5], Sara Vieira-Silva[4,5], Uwe Völker[37], Henry Völzke[57], Urmo Võsa[1], Kaitlin H. Wade[38,39], Susanna Walter[58,59], Kyoko Watanabe[60], Stefan Weiss[37], Frank U. Weiss[15], Omer Weissbrod[61], Harm-Jan Westra[1], Gonneke Willemsen[30], Haydeh Payami[26], Daisy M.A.E. Jonkers[43,44], Alejandro Arias Vasquez[25,62], Eco J.C. de Geus[30,63], Katie A. Meyer[64,65], Jakob Stokholm[22], Eran Segal[21], Elin Org[20], Cisca Wijmenga[1], Hyung-Lae Kim[66], Robert C. Kaplan[16,67], Tim D. Spector[8], Andre G. Uitterlinden[2,3,33], Fernando Rivadeneira[2,3], Andre Franke[16], Markus M. Lerch[15], Lude Franke[1], Serena Sanna[1,68], Mauro D'Amato[12,69,70,71], Oluf Pedersen[14], Andrew D. Paterson[72], Robert Kraaij[2,‡], Jeroen Raes[4,5,‡], Alexandra Zhernakova[1,‡,*]

Affiliations:

[1]Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
[2]Department of Internal Medicine, Erasmus MC University Medical Center, Rotterdam, the Netherlands
[3]The Generation R Study, Erasmus MC University Medical Center, Rotterdam, the Netherlands
[4]Department of Microbiology and Immunology, Rega Instituut, KU Leuven, Leuven, Belgium
[5]Center for Microbiology, VIB, Leuven, Belgium
[6]Institute of Microbiology, Chinese Academy of Sciences, Beijing, China
[7]Section of Statistical Multi-Omics, Department of Clinical & Experimental Medicine, School of Biosciences & Medicine, University of Surrey, Guildford, UK
[8]Department of Twin Research & Genetic Epidemiology, King's College London, London, UK
[9]Department of Medicine, University of Toronto, Toronto, Canada
[10]Division of Gastroenterology, Mount Sinai Hospital, Toronto, Canada
[11]Avera Institute of Human Genetics, Avera McKennan Hospital & University Health Center, Sioux Falls, USA
[12]Center for Molecular Medicine and Unit of Clinical Epidemiology, Department of Medicine Solna, Karolinska Institutet, Stockholm, Sweden

[13]Laboratory of Genomic Diversity, Center for Computer Technologies, ITMO University, St. Petersburg, Russia
[14]Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[15]Department of Medicine A, University Medicine Greifswald, Greifswald, Germany
[16]Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany
[17]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, USA
[18]Medical Research Institute, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea
[19]Department of Clinical Research Design and Evaluation, SAIHST, Sungkyunkwan University, Seoul, Republic of Korea
[20]Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia
[21]Department of Computer Science and Cell Biology, Weizmann Institute of Science, Rehovot, Israel
[22]COPSAC, Copenhagen University Hospital, Herlev-Gentofte, Copenhagen, Denmark
[23]Institute of Molecular Medicine McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, USA
[24]Human Genetics Center School of Public Health, The University of Texas Health Science Center at Houston, Houston, USA
[25]Department of Psychiatry, Radboudumc, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, the Netherlands
[26]Department of Neurology, University of Alabama at Birmingham, Birmingham, USA
[27]Division of Family Medicine and Primary Care, Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Stockholm, Sweden
[28]Stress Research Institute, Stockholm University, Stockholm, Sweden
[29]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, USA
[30]Biological Psychology, Vrije Universiteit, Amsterdam, the Netherlands
[31]Department of Pediatrics, Albert Einstein College of Medicine, Bronx, USA
[32]Department of Microbiology & Immunology, Albert Einstein College of Medicine, Bronx, USA
[33]Department of Epidemiology, Erasmus MC University Medical Center, Rotterdam, the Netherlands
[34]Nuffield Department of Population Health, University of Oxford, Oxford, UK
[35]Department of Pediatrics, Erasmus MC University Medical Center, Rotterdam, the Netherlands
[36]Department of Pediatrics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
[37]Department of Functional Genomics, Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany
[38]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK
[39]Population Health Sciences, Bristol Medical School, Bristol, UK
[40]Department of Endocrinology, Amsterdam University Medical Center, location VUMC, Amsterdam, the Netherlands
[41]Kennedy Institute of Rheumatology, University of Oxford, Oxford, UK

[42]Centre for Clinical Research and Disease Prevention, Bispebjerg/Frederiksberg Hospital, Capital Region of Copenhagen and Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[43]Division of Gastroenterology-Hepatology, Maastricht University Medical Center+, Maastricht, the Netherlands
[44]NUTRIM School of Nutrition and Translational Research in Metabolism, Maastricht University, Maastricht, the Netherlands
[45]Department of Pediatrics, University of California San Diego, La Jolla, USA
[45]Center for Microbiome Innovation, University of California San Diego, La Jolla, USA
[47]Center for Microbiome Innovation and department of Bioengeering, University of California San Diego, La Jolla, USA
[48]Institute of Clinical Medicine, Internal Medicine, University of Eastern Finland, Kuopio, Finland
[49]Department of Medicine I, University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany
[50]Laboratory of Epidemiology and Population Science, National Institute on Aging, Bethesda, USA
[51]Institute of Epidemiology, Kiel University, Kiel, Germany
[52]Department of Human Genetics, University of California, Los Angeles, Los Angeles, USA
[53]Department of Medicine, University of California, Los Angeles, Los Angles, USA
[54]Department of Family Medicine, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea
[55]Center for Cohort Studies, Total Healthcare Center, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea
[56]Department of Biology, University of Copenhagen, Copenhagen, Denmark
[57]Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany
[58]Department of Biomedical and Clinical Sciences, University of Linköping, Linköping, Sweden
[59]Department of gastroenterology, County Council of Östergötland, Linköping, Sweden
[60]Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Neuroscience Campus Amsterdam, VU University Amsterdam, Amsterdam, the Netherlands
[61]School of Public Health, Harvard University, Boston, USA
[62]Department of Human Genetics, Radboudumc, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, the Netherlands
[63]Amsterdam Public Health, Amsterdam UMC, Amsterdam, the Netherlands
[64]Department of Nutrition, University of North Carolina at Chapel Hill, Chapel Hill, USA
[65]Nutrition Research Institute, University of North Carolina at Chapel Hill, Kannapolis, USA
[66]Department of Biochemistry, Ewha Womans University School of Medicine, Seoul, Republic of Korea
[67]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, USA
[68]Istituto di Ricerca Genetica e Biomedica, National Research Council, Monserrato, Italy
[69]School of Biological Sciences, Monash University, Clayton, Australia
[70] Department of Gastrointestinal and Liver Diseases, Biodonostia Health Research Institute, San Sebastián, Spain
[71] Ikerbasque, Basque Science Foundation, Bilbao, Spain
[72]Genetics and Genome Biology, The Hospital for Sick Children Research Institute, Toronto, Canada

[†]denotes shared first authorship
[‡]denotes shared last authorship
[§]these authors contributed equally to this work
[*]corresponding author

Abstract

To study the effect of host genetics on gut microbiome composition, the MiBioGen consortium curated and analyzed whole-genome genotypes and 16S fecal microbiome data from 18,473 individuals (25 cohorts). Microbial composition showed high variability across cohorts: we detected only 9 out of 410 genera in more than 95% of the samples. A genome-wide association study (GWAS) of host genetic variation in relation to microbial taxa identified 30 loci affecting microbome taxa at a genome-wide significant (P<5x10-8) threshold. Just one locus, the lactase (*LCT*) gene region, reached study-wide significance (GWAS signal $P=8.6x10^{-21}$); it showed an age-dependent association with Bifidobacterium abundance. Other associations were suggestive ($1.94x10^{-10}<P<5x10^{-8}$) but enriched for taxa showing high heritability and for genes expressed in the intestine and brain. A phenome-wide association study and Mendelian randomization analyses identified enrichment of microbiome trait loci SNPs in the metabolic, nutrition and environment domains and indicated food preferences and diseases as mediators of genetic effects.

Main Text

Introduction

The gut microbiome is an integral part of the human holobiont and is often considered an organ in itself. More than 10^{13} microorganisms, in large part bacteria, make up the human gut microbiota[1], and in recent years, many studies have highlighted the link between its perturbations

and immune, metabolic, neurologic and psychiatric traits, as well as with drug metabolism and cancer[2]. Environmental factors, like diet and medication, are known to play a significant role in shaping the gut microbiome composition[3–5] although twin, family and population-based studies have provided compelling evidence that the genetic component also plays a role in determining the gut microbiota composition, and a proportion of bacterial taxa show heritability[6,7]. Several studies[8–10] have investigated the effect of genetics on microbiome composition through genome-wide association studies (GWAS) and have suggested dozens of genetic loci that affect the overall microbiome composition or the abundance of specific bacterial taxa. However, little cross-replication across these studies has been observed so far[11,12]. This may be due to a number of factors. First, methodological differences in the collection, processing, sequencing and annotation of stool microbiota are known to have significant effects on the results[13–15], and can therefore generate heterogeneity and a lack of reproducibility across studies. Second, most association signals reported so far are rather weak and prone to non-replication, suggesting that existing studies of 1,000-2,000 samples[8–10] are underpowered. Finally, some of the GWAS signals related to microbiome compositions may be population-specific, i.e. they may represent *bona fide* population differences in genetic structure and/or environment.

To address these challenges in microbiome GWAS studies and obtain valuable insights into the relationship between host genetics and microbiota composition, we set up the international consortium MiBioGen[12]. In this study, we have coordinated 16S rRNA gene sequencing profiles and genotyping data from 18,473 participants from 25 cohorts from the USA, Canada, Israel, South Korea, Germany, Denmark, the Netherlands, Belgium, Sweden, Finland and the UK. We performed a large-scale, multi-ethnic, genome-wide meta-analysis of the associations between autosomal human genetic variants and the gut microbiome. We explored the variation of microbiome composition across different populations and investigated the effects of differences in methodology on the microbiome data. Through the implementation of a standardized pipeline, we then performed microbiome trait loci (mbTL) mapping to identify genetic loci that affect the relative abundance (mbQTLs) or presence (microbiome Binary Trait loci, or mbBTLs) of microbial taxa. Finally, we focused on the biological interpretation of GWAS findings through Gene Set Enrichment Analysis (GSEA), Phenome-wide association studies (PheWAS) and Mendelian randomization (MR) approaches.

Results

Landscape of microbiome composition across cohorts

Our study included cohorts that were heterogeneous in terms of ethnic background, age, male/female ratio and microbiome analysis methodology. Twenty-one cohorts included samples of single ancestry, namely European origin (17 cohorts, N=13,399), Middle-Eastern (1 cohort, N=481), East Asian (1 cohort, N=811), Hispanic (1 cohort N=1,097) and African American (1 cohort, N=114), while four cohorts were multi-ethnic (N=2,571).

The age range was also diverse: 23 cohorts comprised adult or adolescent individuals (N=16,765) and two cohorts (N=1,708) consisted of children (Supplementary Materials, Fig. 1A, Tables S1, S2). The microbial composition of different cohorts was profiled by targeting three distinct variable regions of the 16S rRNA gene: V4 (10,413 samples, 13 cohorts), V3-V4 (4,211 samples, 6 cohorts) and V1-V2 (3,849 samples, 6 cohorts) (Fig. 1A). To account for coverage heterogeneity resulting from differences in sequencing depth and bacterial load across samples, all microbiome datasets were rarefied to 10,000 reads per sample. Next, we performed taxonomic classification using direct taxonomic binning instead of OTU clustering methods[12].

In general, cohorts were variable in their microbiome structure at various taxonomic levels. This variation may be largely driven by the heterogeneity between populations and technical differences derived from using diverse collection methods, DNA extraction protocols and sequencing strategies (Tables S1, S3). Out of these factors, we identified that the DNA extraction method was a principal contributor to heterogeneity, with a non-redundant effect size of 37% on the microbiome variation (measured as average genus-abundance per cohort; stepwise distance-based redundancy analysis $R2adj_{DNAext}=0.22$, $P_{adj}=1.5x10^{-3}$) (Table S4). The median richness at genus-level was significantly different across cohorts (Fig. 1F; pairwise Wilcoxon rank sum test; FDR<0.05), with the COMPULS and NTR cohorts exceeding the overall median of 90 genera by more than 15% and the HCHS/SOL cohort yielding the lowest median richness of 54 (SD:14) genera. The microbial Shannon diversity index also varied notably between study cohorts (Fig. 1G), with DanFund presenting the highest diversity index of 3.52 (SD:0.28), well above the PNP, HCHS/SOL, NGRC and KSCS cohorts, with median diversities below 2.5. The cohorts with the lowest and highest diversity, HCHS/SOL and DanFund, used specific DNA

extraction kits (NucleoSpin Soil and PowerLyzerPowerSoil, respectively) that were not used in any other cohorts, possibly contributing to their outlying alpha diversities (Table S3). In total, the different cohort ethnicities, 16S rRNA target-amplicons and DNA extraction methods accounted for 32% of the variance in the observed richness.

Combining all samples (N=18,473) resulted in a total richness of 410 genus-level taxonomic groups that had a relative abundance higher than 0.1% in at least one cohort. This observed total richness appears to be below the estimated saturation level (Fig. 1B), suggesting that a further increase in sample size and a higher sequencing depth are needed to capture the total gut microbial diversity (Fig. 1D). As expected, the core microbiota (the number of bacterial taxa present in over 95% of individuals) decreased with the inclusion of additional cohorts (see Methods, Fig. 1C). The core microbiota comprise nine genera, of which seven were previously identified as such[4], plus the genera *Ruminococcus* and *Lachnoclostridium* (Fig. 1E). Of these nine genera, the most abundant genus was *Bacteroides* (18.19% (SD:8.37)), followed by *Faecalibacterium* (6.28% (SD:2.24)), *Blautia* (3.48% (SD:2.76)) and *Alistipes* (2.93% (SD:1.47)). Among the European cohorts that compose the largest genetically and environmentally homogeneous cluster, the core microbiota also included *Ruminiclostridium*, *Fusicatenibacter*, *Butyricicoccus* and *Eubacterium*, genera which typically produce short-chain fatty acids[16].

Given the high heterogeneity of microbial composition across cohorts, we applied both per-cohort and whole study-filters for taxa inclusion in GWAS. Cohort-wise, the inclusion criteria for GWAS on bacterial abundance was that the taxon is present in more than 10% of samples from each cohort, while for the binary trait (bacterial presence vs absence) GWAS, a taxon had to be seen in at least 10% and maximally 90% of the cohort samples. Study-wide cutoffs for mbQTL mapping included an effective sample size of at least 3,000 samples and presence in at least three cohorts (see Methods). For mbBTLs, a mean abundance for a taxon of higher than 1% in the taxon-positive samples was required. This filtering resulted in 257 taxa being included in our analysis (Table S3).

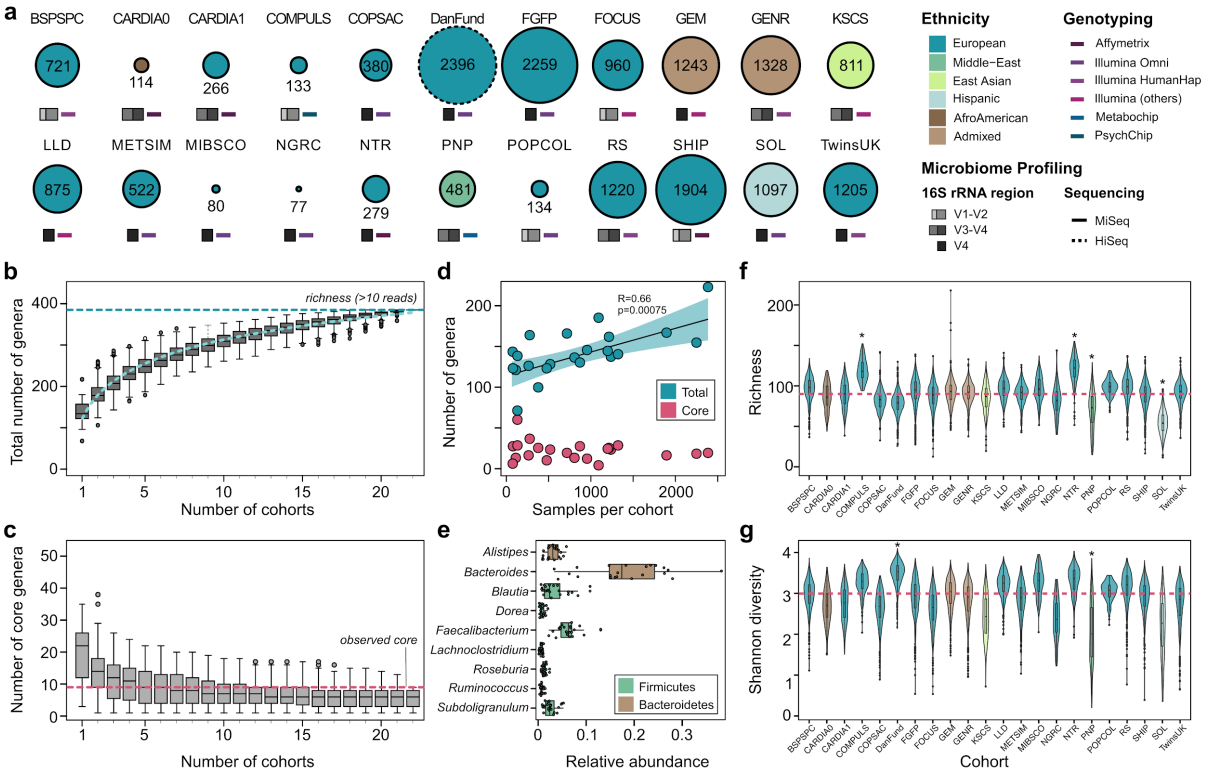


Figure 1. Diversity of microbiome composition across the MiBioGen cohorts. (A) Sample size, ethnicity, genotyping array and 16S profiling method. The SHIP/SHIP-TREND and GEM_v12/GEM_v24/GEM_ICHIP subcohorts are combined in SHIP and GEM, respectively (see Methods), resulting in the 22 cohorts depicted in the figure. **(B)** Total richness (number of genera with mean abundance over 0.1%, i.e. 10 reads out of 10,000 rarefied reads) by number of cohorts investigated. **(C)** Number of core genera (genera present in >95% of samples from each cohort) by number of cohorts investigated. **(D)** Correlation of cohort sample size with total number of genera. **(E)** Unweighted mean relative abundance of core genera across the entire MiBioGen dataset. **(F)** Per-sample richness across the 22 cohorts. Asterisks indicate cohorts that differ significantly from the others (pairwise Wilcoxon rank sum test; FDR<0.05). **(G)** Diversity (Shannon index) across the 22 cohorts, with the DanFund and PNP cohorts presenting higher and lower diversity in relation to the other cohorts (pairwise Wilcoxon rank sum test; FDR<0.05).

Heritability of microbial taxa and alpha diversity

We performed the estimation of heritability (H^2) of gut microbiome composition based on two twin cohorts included in our study (Table S5). The TwinsUK cohort, composed of 1,176 samples, including 169 monozygotic (MZ) and 419 dizygotic (DZ) twin pairs, was used to estimate H^2 using the ACE (additive genetic variance (A)/shared environmental factors (C)/ non-shared factors plus error (E)) model. The NTR cohort (only MZ twins, N=312, 156 pairs) was used to replicate the MZ intraclass correlation coefficient (ICC). None of the alpha diversity metrics (Shannon, Simpson and inverse Simpson) showed evidence for heritability (A<0.01, P=1). Among 159 bacterial taxa that were present in more than 10% of twin pairs (>17 MZ pairs, >41 DZ pairs in TwinsUK), 19 taxa showed evidence for heritability ($P_{nominal}$<0.05) (Fig. 2A). The ICC shows concordance between TwinsUK and NTR for these 19 bacterial taxa (R=0.25, P=0.0018, Fig. 2B).

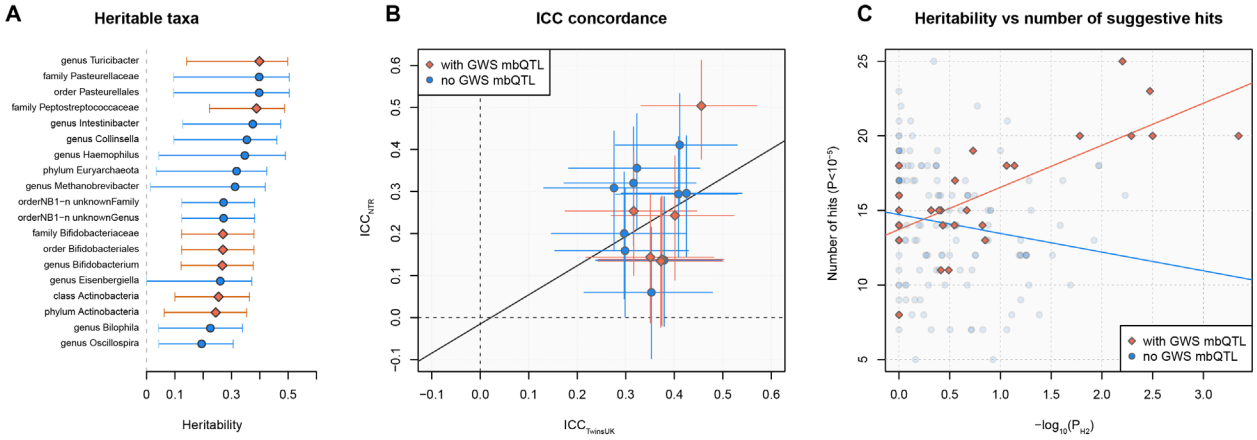


Figure 2. Heritability of microbiome taxa and its concordance with mbQTL mapping. (A) Microbial taxa that showed significant heritability in the TwinsUK cohort (P<0.05). Taxa with at least one genome-wide significant (GWS) mbQTL hit are marked red. Only taxa present in more than 10% of pairs (>17 MZ pairs, >41 DZ pairs) are shown. **(B)** Correlation of monozygotic ICC between TwinsUK and NTR cohort. Only taxa with significant heritability (P<0.05) that are present in both TwinsUK and NTR are shown. Red and blue dots indicate bacterial taxa with/without GWS mbQTLs ($P<5x10^{-8}$), respectively. **(C)** Correlation between heritability significance ($-\log_{10}P_{H2}$ TwinsUK) and the number of loci associated with microbial taxon at relaxed threshold ($P_{mbQTL}<1x10^{-5}$). Taxa with at least one GWS-associated locus are marked red. Error bars represent 95% confidence intervals.

The SNP-based heritability calculated from mbQTL summary statistics using linkage disequilibrium (LD) score regression showed no bacterial taxa passing the significance threshold, given the number of 204 tested taxa ($Z<3.67$, Table S5). The results of the SNP-based heritability and twin-based heritability showed significant correlation across the tested taxa ($R=0.252$, $P=4.8\times10^{-4}$).

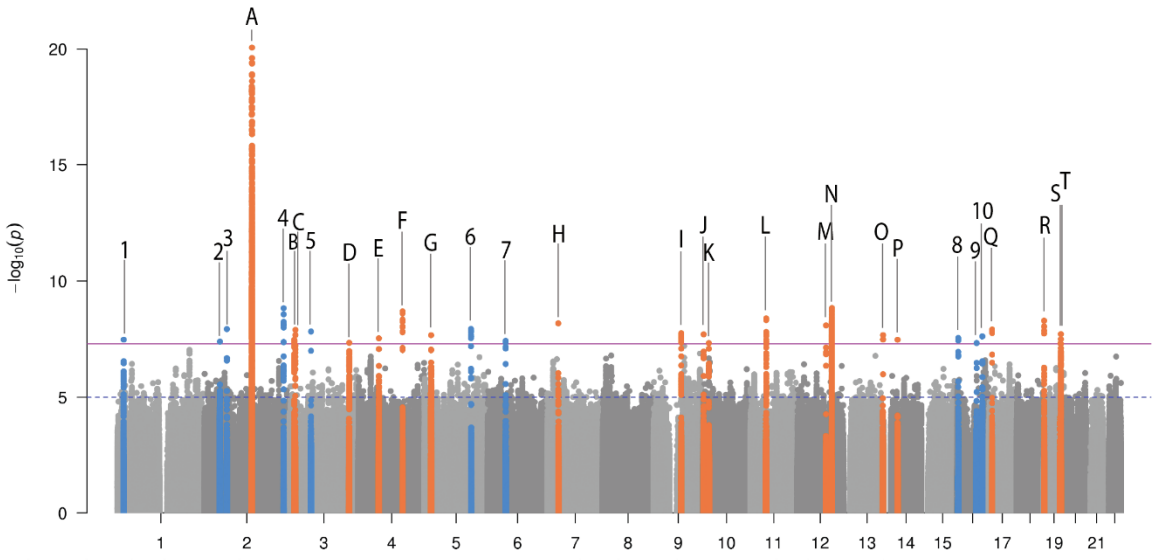
GWAS meta-analysis identified 30 loci associated with gut microbes at genome-wide significance level

First, we studied the genetic background of the alpha diversity (i.e. Simpson, inverse Simpson and Shannon diversity indices). We identified no significant hits in the meta-GWAS ($P>5\times10^{-8}$; Fig. S1, Table S6), which is in line with the observed lack of heritability for this trait in the twin cohorts.

Next, we used two separate GWAS meta-analysis approaches to explore the effect of host genetics on the abundance levels or presence/absence of bacterial taxa in the gut microbiota. We performed a mbQTL analysis (**mbQTL**) of the genetic effect on relative bacterial abundance, only including samples with non-zero abundance (see Methods). Additionally, we performed a mbBTL analysis gauging the presence/absence of each bacterial taxon. For each taxon, the analysis was performed in each cohort separately. Results from the different cohorts were meta-analyzed using a weighted z-score method.

In total, 18,473 samples with both microbiome 16S rRNA profiling data and imputed genetic data were included in the mbQTL mapping analysis. mbQTL mapping was performed for 211 taxa (131 genera, 35 families, 20 orders, 16 classes and 9 phyla) that passed taxon inclusion cutoffs (Online Methods, Table S3). At genome-wide significance level ($P<5\times10^{-8}$) we identified mbQTLs for 457 SNPs mapping to 20 distinct genetic loci, associated with 27 taxa (Figs. 3, S2, S3, Tables S7, S8). mbBTL mapping covered 173 taxa that passed taxon inclusion cutoffs (see Methods), including 105 genera, 31 families, 17 orders, 13 classes and 7 phyla. At $P<5\times10^{-8}$, 10 loci were found to be associated with presence/absence of bacterial taxa using a three-stage procedure (Pearson correlation of SNP dosage with microbial presence/absence followed by meta-analysis and validation using logistic regression adjusted for covariates, see Methods) (Fig.

3, Tables S7, S9). For one taxon, family *Peptococcaceae*, two independent mbBTLs were detected (Fig. 3, Table S7). The effect sizes of the leading SNPs at the 30 genome-wide significant loci were consistent across cohorts, with the exception of two mbQTLs presenting heterogeneity (Cochran's Q P<0.05), including the *LCT* association with phylum Actinobacteria and a cluster of related taxa (class Actinobacteria, order Bifidobacteriales, family *Bifidobacteriaceae* and genus *Bifidobacterium)* and the association in 3q26.31 (*FNDC3B,* leading SNP rs4428215) to the genus *Oxalobacter* (Fig. S4). In both the mbQTL and mbBTL mapping, the only association that passed strict study-wide significance cutoff ($P<1.94x10^{-10}$ for the total of 257 taxa included in the analysis) was observed between variants mapping to the *LCT* locus and the genus *Bifidobacterium* and related taxa at higher ranks (from family *Bifidobacteriaceae* to phylum Actinobacteria).



Locus	Pvalue	Type	Chr	rsID	Bacteria
A	8.63E-21	Quant	2	rs182549	genus Bifidobacterium
B	3.44E-08	Quant	3	rs9864379	order Gastranaerophilales
C	1.27E-08	Quant	3	rs75754569	genus Peptococcus
D	4.57E-08	Quant	3	rs4428215	genus Oxalobacter
E	2.92E-08	Quant	4	rs2199860	genus Porphyromonadaceae_unknown_genus_id959
F	2.00E-09	Quant	4	rs11098863	genus Enterorhabdus
G	2.16E-08	Quant	5	rs13163520	genus group Ruminococcus gnavus
H	6.66E-09	Quant	7	rs17159861	genus group Eubacterium coprostanoligenes
I	1.77E-08	Quant	9	rs602075	genus Allisonella
J	1.97E-08	Quant	10	rs12781711	genus RuminococcaceaeUCG013
K	4.70E-08	Quant	10	rs61841503	family Peptostreptococcaceae
L	4.08E-09	Quant	11	rs10769159	genus Ruminococcus1
M	8.13E-09	Quant	12	rs12320842	genus Faecalibacterium
N	1.46E-09	Quant	12	rs11110281	genus Streptococcus
O	2.12E-08	Quant	13	rs7322849	genus Bifidobacterium
P	3.37E-08	Quant	14	rs8009993	genus RuminococcaceaeUCG009
Q	1.22E-08	Quant	17	rs7211274	genus Erysipelatoclostridium
R	5.10E-09	Quant	19	rs67476743	genus Tyzzerella3
S	4.84E-08	Quant	19	rs830151	genus Candidatus Soleaferrea
T	1.93E-08	Quant	19	rs35866622	genus group Ruminococcus torques
1	3.34E-08	Binary	1	rs555115	genus Turicibacter
2	4.07E-08	Binary	2	rs57440956	genus Ruminiclostridium6
3	1.17E-08	Binary	2	rs56081657	genus Odoribacter
4	1.48E-09	Binary	2	rs7574352	family Peptococcaceae
5	1.50E-08	Binary	3	rs17066404	family Peptococcaceae
6	1.16E-08	Binary	5	rs62373497	genus Coprococcus1
7	3.79E-08	Binary	6	rs9474033	family Pseudomonadaceae
8	2.83E-08	Binary	16	rs437867	genus Lachnospiraceae UCG010
9	4.67E-08	Binary	16	rs17319026	genus Anaerostipes
10	2.42E-08	Binary	16	rs11647069	family Bacteroidales S24.7

Figure 3. Manhattan plot of the mbTL mapping meta-analysis results. MbQTLs are indicated by letters. MbBTLs are indicated by numbers.

Heterogeneity of microbiome composition reduces the power of genetic association analysis

The substantial variation in taxonomic composition driven by technical factors, including 16S domain and DNA extraction kits, has a significant effect on microbiome GWAS. For example, the genus *Bifidobacterium*, which showed the strongest genetic effect, was present in 93% of the samples in those cohorts that used the V4 domain of the 16S rRNA gene, but only in 78% and 62% of the samples sequenced by V3-V4 and V1-V2 domains, respectively. Similar to the 16S domain, the DNA isolation method showed a strong influence on *Bifidobacterium* abundance, which ranged from 35.7% to 100% depending on the DNA isolation kit (Table S3). Another example is the large effect of the sequencing domain on the presence of the Archaea, in particular genus *Methanobrevibacter*. The proportion of Archaea-positive individuals in cohorts sequenced by V3-V4 or V4 domains was around 25-35%, similar to the prevalence estimated using shotgun metagenomics sequencing[3], whereas Archaea were not detected at all in cohorts that used the V1-V2 domain. This lack of Archaea detection dramatically reduces the sample size for mbTL mapping and may well explain the lack of genome-wide significant mbTLs for this domain, despite its moderate heritability (H^2=0.319). In general, half of the 211 bacterial taxa that passed either the quantitative or binary mbTL filtering cutoff showed substantial differences in abundance or presence between 16S domains or DNA extraction methods (Table S3). However, our design did not always allow us to distinguish the causes of heterogeneity since the methodological discrepancy overlapped biological variance between cohorts, including ethnicity, age, body mass index (BMI) and study design. For example, most of the cohorts that used the V1-V2 16S domain had German ancestry, whereas the group of cohorts that used the V3-V4 domain was very diverse and included several non-European or multi-ethnic cohorts (Table S1). Despite the expected effects of microbiome heterogeneity on the heterogeneity of mbTLs effects, we did not observe this correlation for either genome-wide significant or suggestive mbTLs (Fig. S5A). However, the taxa with higher inter-cohort variation and smaller effective sample size showed smaller numbers of genome-wide significant ($P<5x10^{-8}$) and suggestive ($P<1x10^{-5}$) associated loci (Figs. S5B, S5C). Thus, the microbiome heterogeneity reduced the power of analysis but didn't induce heterogeneity of mbTL effects.

LCT locus association to *Bifidobacterium* is age- and ethnicity-dependent

Among the mbQTLs, the strongest association signal was seen for variants located in a large block of about 1.5Mb at 2q21.3, which includes the *LCT gene* and 12 other protein-coding genes. We found 317 SNPs ($P<5x10^{-8}$) from this locus that were associated with the genus *Bifidobacterium* and higher taxonomic ranks (family *Bifidobacteriaceae*, order Bifidobacteriales, class Actinobacteria, phylum Actinobacteria). This locus has been previously associated with the abundance of *Bifidobacterium* in Dutch[8], UK[7] and US[17] cohorts. Previous studies have also shown a positive correlation of *Bifidobacterium* abundance with the intake of milk products, but only in individuals homozygous for the low-function LCT haplotype, thereby indicating the role of gene–diet interaction in regulating *Bifidobacterium* abundance[8]. In our study, the strongest association was seen for rs182549 ($P=8.63x10^{-21}$), which is a perfect proxy for the functional *LCT* variant rs4988235 ($r^2=0.996$, D'=1 in European populations). This association showed evidence for heterogeneity across cohorts ($I^2=58.4\%$, Cochran's Q $P=1.32x10^{-4}$). The leave-one-out strategy showed that the cohort contributing the most to the detected heterogeneity was the $COPSAC_{2010}$ cohort, which includes children with an age range of 4-6 years (Table S2), which pulled effect estimates towards zero at a younger age (Figs. 4A, 4B). Once this study was excluded from the meta-analysis, heterogeneity was reduced ($I^2=48.2\%$, Cochran's Q P=0.004). A meta-regression analysis showed that age and ethnicity accounted for 11% of this heterogeneity ($P_{fixed}=3.76x10^{-21}$, $P_{random}=6.5x10^{-8}$), and graphical representation of this meta-regression (Fig. 4C) suggested a non-linear relationship between age and the mbQTL effect. The inclusion of quadratic and cubic terms of age in the model resulted in 42% of the heterogeneity being accounted for and showed evidence that the remaining residual heterogeneity was slight (Cochran's Q P=0.03) (Fig. 4C).

Following these observations, we decided to investigate the effect of age and ethnicity in the multi-ethnic GEM cohort, which has a comparable sample size for both infants and adults. In total, this cohort comprised 1,243 individuals, with an age range between 6 and 35 years, and nearly half of the participants 16 years or younger. Our analysis showed a significant SNP–age interaction on the level of *Bifidobacterium* abundance (P<0.05, see Methods). Those individuals homozygous for the rs182549*C/C genotype showed a higher abundance of the genus *Bifidobacterium* in the adult group, but not in the younger group (Fig. 4D). The age–genotype

interaction was significant in the GEM_v12 and GEM_ICHIP subcohorts, both comprising European individuals, while the GEM_v24 cohort, mainly composed of individuals of different Israeli subethnicities (see Methods) who live in Israel, showed neither an mbQTL effect ($Beta_{SNP}$=-0.002 [95%CI: -0.21, 0.21]) nor an interaction with age (P>0.1). The lack of an *LCT* mbQTL effect in adults was also replicated in another Israel cohort in the study (PNP, 481 adults, $Beta_{SNP}$=-0.20 [95%CI: -0.61, 0.20]). Thus the three cohorts that reported the lowest LCT effect sizes were two cohorts of Israeli ethnicity volunteered in Israel (GEM_v24, PNP) and a child cohort (COPSAC, $Beta_{SNP}$=-0.18 [95%CI: -0.36, -0.01]).

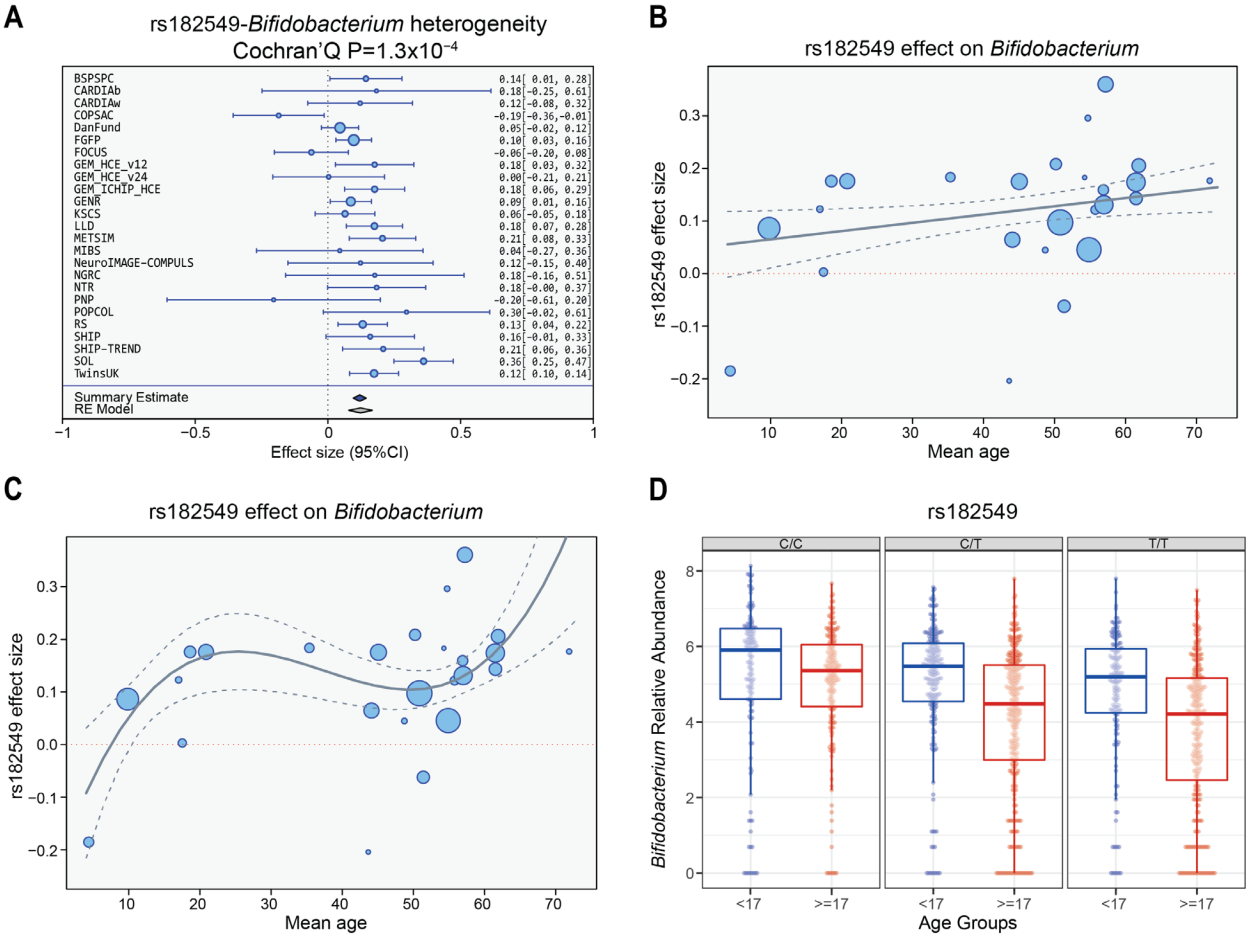


Figure 4. Association of the *LCT* locus (rs182549) with the genus *Bifidobacterium*. **(A)** Forest plot of effect sizes of rs182549 and abundance of *Bifidobacterium*. **(B)** Meta-regression of the association of mean cohort age and mbQTL effect size. **(C)** Meta-regression analysis of the effect of linear, squared and cubic terms of age on mbQTL effect size. **(D)** Age-dependence of

mbQTL effect size in the GEM cohort. Blue boxes include samples in the age range 6–16 years old. Red boxes include samples with age ≥17 years. The rs182549*T allele is a proxy of the rs4988235*C allele, which is associated to functional recessive hypolactasia.

Bacteria-associated loci are genetically enriched for genes related to metabolism

The remaining 29 loci that associated at the genome-wide significance level ($P<5x10^{-8}$) did not pass our strict cutoff of correction for the number of tested taxa ($P<1.46x10^{-10}$). However, these loci include functionally relevant variants (i.e. the *FUT2* gene suggested in earlier studies [18]), and overall showed concordance with the heritability of microbial taxa. Six out of nine taxa that showed the strongest evidence for heritability in the TwinsUK cohort ($P<0.01$) have genome-wide significant mbTLs (Fig. 2B): genus *Turicibacter*, family *Peptostreptococcaceae* and class Actinobacteria with its nested taxa: order Bifidobacteriales, family *Bifidobacteriaceae* and genus *Bifidobacterium*. For the taxa with genome-wide significant mbTLs, the number of independent loci associated with a relaxed threshold of $1x10^{-5}$ strongly correlated with heritability significance ($R=0.73$, $P=3.3x10^{-6}$, Fig. 2C), suggesting that more mbTLs would be identified for this group of bacteria with a larger sample size.

Of loci with an association that did not achieve our stringent study-wide threshold, but did pass the nominal genome-wide significance threshold, the strongest mbQTL included 66 SNPs located in the *UHRF1BP1L* locus (12q23.1) that associated with the *Streptococcus* genus and *Streptococcaceae* family (rs11110281, $P=1.46x10^{-9}$). Eight genes located in this locus were identified by FUMA as positional candidates, including the closest gene, *UHRF1BP1L,* which is expressed in adipose tissue, liver and skeletal muscle. None of these genes could be prioritized as a prominent functional candidate based on published data and co-expression networks [19]. In the LLD cohort, the *Streptococcus* genus and *Streptococcaceae* family were positively correlated with stool levels of inflammatory markers chromogranin A ($R_{Sp}=0.22$, $P_{adj}=1.89x10^{-7}$) and calprotectin ($R_{Sp}=0.16$, $P_{adj}=1.4x10^{-3}$) and with the intake of proton pump inhibitors (PPI) ($R_{Sp}=0.21$, $P_{adj}=9.42x10^{-7}$) (Table S10).

In mbQTL analysis, the ***FUT2-FUT1* locus** was associated to the abundance of the *Ruminococcus torques* genus group, a genus from the *Lachnospiraceae* family. The associated leading SNP (rs35866622 for *R. torques* group, $P=1.9x10^{-8}$) is a proxy for the functional variant

rs601338, which introduces a stop-codon in *FUT2* (r^2=0.8; D'=0.9 in European populations, according to LDlink)[20]. The other proxy of the functional *FUT2* SNP, rs281377 (r^2=0.71, D'=1, European populations), showed association to the *Ruminococcus gnavus* genus group in the binary analysis, although this signal was just above the genome-wide significance threshold ($P=5.79x10^{-8}$) (Table S9). *FUT2* encodes the enzyme alpha-1,2-fucosyltransferase, which is responsible for the secretion of fucosylated mucus glycans in the gastrointestinal (GI) mucosa[21]. Individuals homozygous for the stop-codon (rs601338*A, non-secretors) do not express ABO antigens on the intestinal mucosa. We observed that the tagging rs35866622*T (non-secretor) allele was associated with a reduced abundance of the *R. torques* group and a decreased presence of the *R. gnavus* group. *Ruminococcus sp.* are specialized in the degradation of complex carbohydrates[22], thereby supporting a link between genetic variation in the *FUT2* gene, levels of mucus glycans and the abundance of this taxa. When assessing the link between this variant and phenotypes in the LLD (N=875) and FGFP cohorts (N=2,259), the strongest correlation for the *R. torques* group was seen with fruit intake (LLD: R_{Sp}=-0.19, $P_{adj}=3.1x10^{-5}$; FGFP: R_{Sp}=-0.10, $P_{adj}=1.4x10^{-4}$, Tables S10, S11), which corresponds with the association of *FUT2* with food preferences, as discussed in results of PheWAS (see below).

Several other suggestive mbQTLs can be linked to genes potentially involved in host-microbiome crosstalk. One of them includes eight SNPs in 9q21 (top-SNP rs602075, $P=1.77x10^{-8}$) associated with abundance of *Allisonella*. The 9q21 locus includes genes *PCSK5*, *RFK* and *GCNT1*, of which *RFK* encodes the enzyme that catalyzes the phosphorylation of riboflavin (vitamin B2) and *GCNT1* encodes a glycosyltransferase involved in biosynthesis of mucin. These products play major roles in the host-microbiota interactions within the intestine, used by bacteria for their metabolism but also involved in the regulation of the immune defense[24]. Another associated locus on 10p13 (rs61841503, $P=4.7x10^{-8}$), which affects the abundance of the heritable family *Peptostreptococcaceae,* is located in the *CUBN* gene, the receptor for the complexes of cobalamin (vitamin B12) with gastric intrinsic factor (the complex required for absorption of cobalamin). *CUBN* is expressed in the kidneys and the intestinal epithelium and is associated with B12-deficient anemia and albuminuria[25]. Cobalamin is required for host–microbial interactions[26], and supplementation with cobalamin for seven days induced a substantial shift in the microbiota composition of an *in vitro* colon model[27]. These associations

suggest that some members of the gut microbiome community might be affected by genetic variants that regulate the absorption and metabolism of vitamins B2 and B12.

Among mbBTLs, the strongest evidence for association was seen for a block of 10 SNPs (rs7574352, $p=1.48x10^{-9}$) associated with the family *Peptococcaceae*, a taxon negatively associated with stool levels of gut inflammation markers chromogranin A (LLD: R_{Sp}=-0.31, $P_{adj}=4.4x10^{-18}$, Table S10) and calprotectin (LLD: R_{Sp}=-0.11, P_{adj}=0.058) and with ulcerative colitis (FGFP: R_{Sp}=-0.06, P_{adj}=0.09, Table S11). The association block is located in the intergenic region in the proximity (220kb apart) of *IRF1*, which is involved in insulin resistance and susceptibility to type 2 diabetes (T2DM)[28]. Another genus, *Turicibacter*, which was the most heritable taxon determined by the twin analysis, was associated with rs555115 ($P=3.34x10^{-8}$), which is located in *IGSF21*, an immunoglobulin superfamily gene. *Turicibacter* is associated with decreased stool frequency and higher tea intake in the LLD cohort (Table S10) and is negatively associated with smoking in the FGFP (Table S11). The genus *Anaerostipes* was observed to be linked with rs17319026 ($P=4.67x10^{-8}$), located in carboxylesterase 5A (*CES5A*), which is involved in xenobiotic metabolism. Finally, the prevalence of the *Lachnospiraceae* family was associated with SNPs located in the olfactory receptor family 1 subfamily F member 1 (*OR1F1*). Although no associations have been reported between this SNP or the bacteria and food-related phenotypes, this gene is one of the olfactory receptors that regulates the perception of smell, which, in turn, might influence food preferences.

mbTLs are enriched for genes expressed in intestine and brain and associated with metabolic phenotypes

To systematically explore the potential functions of the identified mbTLs, we performed a functional mapping analysis, GSEA and PheWAS, followed by Bayesian colocalization analysis and genetic correlation of *Bifidobacterium* abundance to various traits. FUMA annotation (Functional Mapping and Annotation of GWAS, see Methods) of associated loci returned 130 positional and eQTL genes from 20 mbQTLs. GSEA on the 130 positional and eQTL genes suggested the 20 mbQTL loci to be enriched for genes expressed in the small intestine (terminal ileum) and brain (substania nigra) (Figs. S6A, B). The positional candidates for mbBTLs did not show any enrichment in GSEA analysis.

To systematically assess the biological outcomes of the mbTLs, we looked up the 30 mbTLs in the summary statistics for 4,155 complex traits and diseases, using the GWAS ATLAS[29]. First, we performed the analysis on single SNP overlap; next, we performed a gene-based analysis, and finally a phenotype domain enrichment analysis among mbTL hits. In the single SNP analysis of 30 mbTLs, five were associated with one or more phenotypes with $P<5x10^{-8}$ (Table S12). The loci associated with the most phenotypes were: rs182549 (*LCT*) and rs35866622 (*FUT1/ FUT2*), followed by rs4428215 (*FNDC3B*) from the mbQTLs and rs11647069 (*PMFBP1)* and rs9474033 (*PKHD1*) from the mbBTLs.

Rs182549 (LCT, *Bifidobacterium*) was associated with multiple dietary and metabolic phenotypes, and the causal involvement of the SNP across pairs of traits was confirmed by colocalization test (PP.H4.abf > 0.9) for 58 out of 60 phenotypes. The rs182549*C allele, which predisposes to lactose intolerance, was negatively associated with obesity[30] and positively associated with several nutritional phenotypes, T2DM diagnosis (OR=1.057 [95%CI:1.031, 1.085], $P=1.74x10^{-5}$) and family history of T2DM (paternal: OR=1.054 [95%CI:1.035, 1.073], $P=1.41x10^{-8}$; maternal: OR=1.035 [95%CI:1.016, 1.053], P=0.0002, siblings: OR=1.03 [95%CI:1.009, 1.052]) in the UK Biobank cohort[29]. Moreover, the functional *LCT* SNP rs4988235 variant was in strong linkage disequilibrium with top mbQTL hit and associated with 1,5-anhydroglucitol ($P=4.23x10^{-28}$)[31], an indicator of glycemic variability[32]. The strongest genetic correlation of *Bifidobacterium* was with raw vegetable intake (rg=0.36, P=0.0018), although this correlation was not statistically significant after correction for multiple testing.

Rs35866622*C (proxy of the secretor allele in *FUT1*/*FUT2* locus) was positively associated with fish intake and height. The secretor allele was negatively associated with the risk of cholelithiasis and Crohn's disease, alcohol intake frequency, high cholesterol and waist-to-hip ratio (adjusted for BMI) with PP.H4.abf > 0.9. The gene-based analysis indicated a strong link of the *LCT* locus with metabolic traits ($P=5.7x10^{-9}$ for BMI), whereas the *FUT1*/*FUT2* locus showed to contain several nutritional ($P=1.26x10^{-20}$ for oily fish intake), immune-related ($P=1.73x10^{-12}$ for mean platelet volume), gastrointestinal (cholelithiasis, $P=8.77x10^{-14}$) and metabolic signals (high cholesterol, $P=1.13x10^{-13}$) (Fig. 5, Table S13).

Next, we performed a phenotype domain enrichment analysis (described in Methods) to see if any of the phenotype domains were enriched among the PheWAS signals (Fig. 5). Overall, the top loci were enriched with signals associated with the metabolic domain supported by 10

mbTLs, followed by nutritional, cellular, immunological, psychiatric, ophthalmological and respiratory traits and the activities domain (Fig. 5, Table S14).

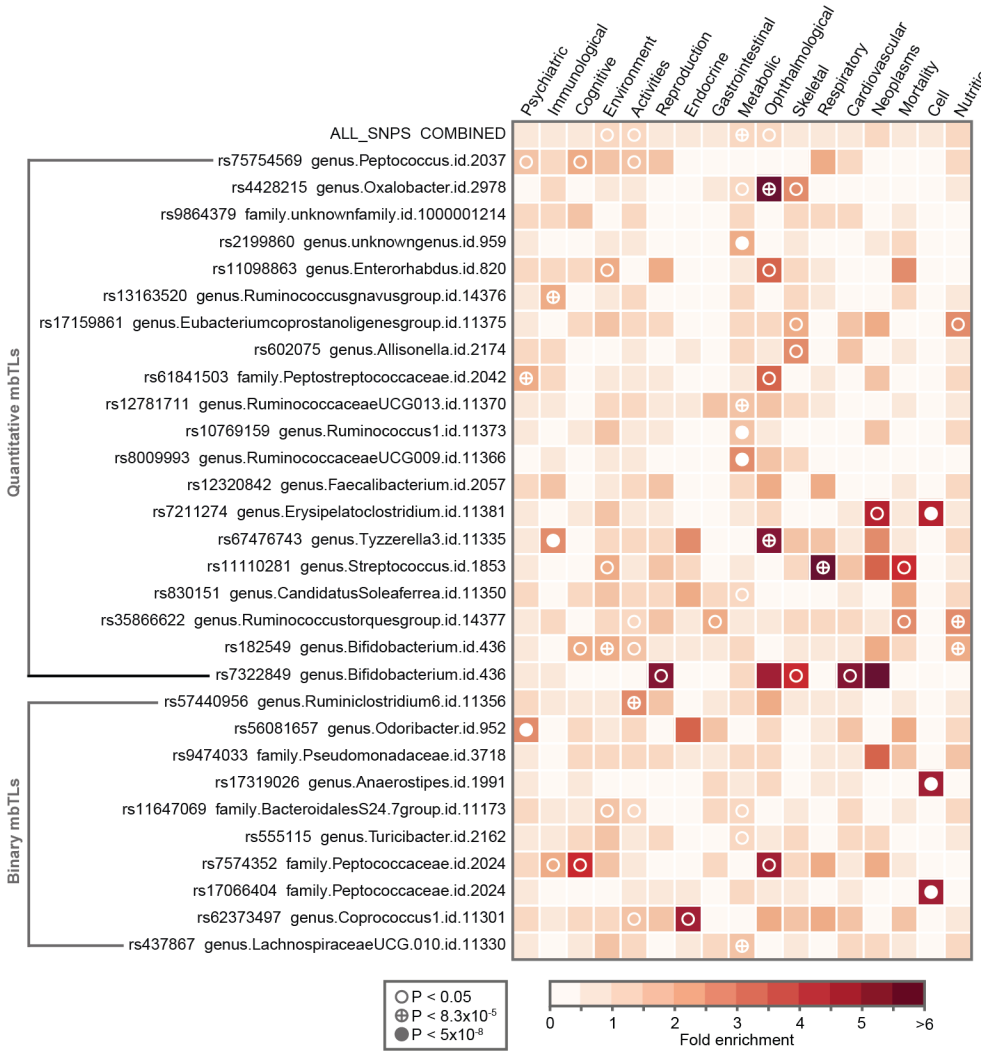


Figure 5. Phenome-wide association study (PheWAS) domain enrichment analysis. The analysis covered top-SNPs from 30 mbTLs and 20 phenotype domains. Three thresholds for multiple testing were used: 0.05, 8.3×10^{-5} (Bonferroni adjustment for number of phenotypes and genotypes studied) and 5×10^{-8} (an arbitrary genome-wide significance threshold). Only categories with at least one significant enrichment signal are shown.

Mendelian randomization (MR) analysis

To identify the potential causal links between gut microbial taxa and phenotypes, we performed bi-directional two-sample MR analyses using the TwoSampleMR package[33]. We focused on two groups of phenotypes: (1) diseases, including autoimmune, cardiovascular, metabolic and psychiatric diseases and (2) nutritional phenotypes.

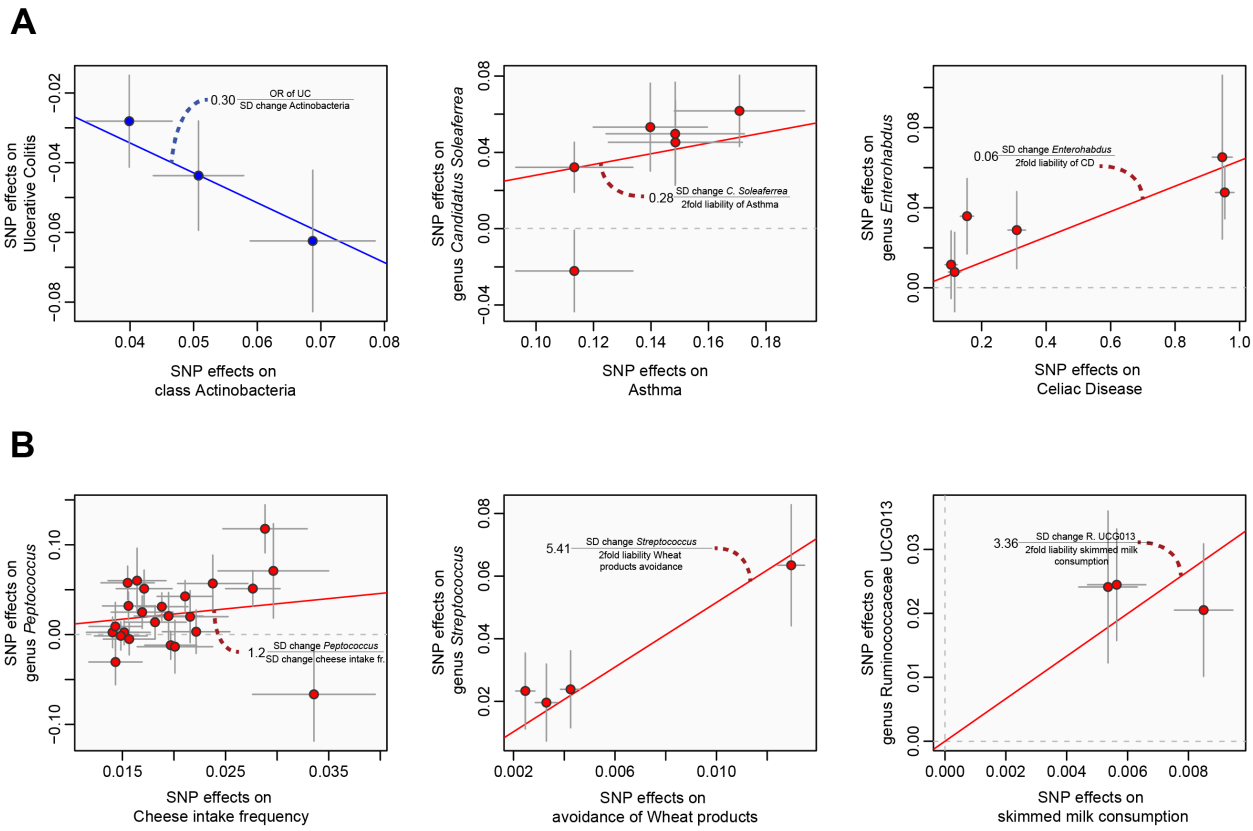


Figure 6. Mendelian randomization (MR) analysis. The X-axes show the SNP-exposure effect and the Y-axes show the SNP-outcome effect. The figures show only the significant results (P_{adj}<0.05). Blue dots represent microbial trait as exposure, red dots used in the graphs with microbial trait as outcome. **(A)** MR analysis of microbiome taxa abundance with diseases. **(B)** MR analysis of microbiome taxa abundance with dietary preference factors.

We used GWAS summary statistics of complex traits in conjunction with our meta-analysis results (30 mbTL loci) to find complex traits (exposure) that suggest a causal relationship to microbiome composition (outcome) and, in the reverse direction, to find bacterial taxa (exposure) that affect complex traits (outcome). The complexity of mechanisms by which

host genetics affect microbiome composition and the limited impact of genetic variants on microbial taxa variability require caution when performing and interpreting causality estimation using MR analysis[34]. We therefore carried out several sensitivity analyses and excluded any results that showed evidence of being confounded by pleiotropy (see Methods).

When using complex traits as the exposure, we found two diseases that showed evidence for a causal relationship with microbiome composition (based on a multiple testing adjusted P-value, $P_{BHadj}<0.05$, Table S15). Specifically, our MR results suggest that doubling the genetic liability of asthma increases the abundance of the genus *Candidatus Soleaferrea* by 0.28 standard deviations (SD) (standard error (SE) of the estimate=0.01, $P_{BHadj}=0.01$, Table S15) and that doubling genetic liability of celiac disease increases the abundance of the genus *Enterorhabdus* by 0.06 SD (SE=0.01, $P_{BHadj}=0.006$, see Table S15). In line with our observations, members of the *Enterorhabdus* genus have been shown to play a role in gluten metabolism[35]. In the reverse direction, with bacterial exposure affecting disease outcome, we found evidence that a higher abundance of the class Actinobacteria may have a protective effect on ulcerative colitis (OR=0.30 [95%CI: 0.23-0.38] for each SD increase in bacteria abundance, $P_{BHadj}=0.003$) (Fig 6A).

Next, we assessed the causal direction between dietary preference factors from the UK Biobank and the abundance of gut microbial taxa. While instrument strength for MR analysis was comparable for both microbiome and nutritional phenotypes (median F statistic per trait was 36 and 41, respectively, Table S15), there was no evidence of causal relationships leading from bacterial taxa to dietary preference factors. In the opposite direction, four nutritional phenotypes showed evidence of a causal effect on the abundance of several taxa. Doubling the genetic liability of avoiding wheat products and of consuming hot drinks was associated with an 5.41 SD increase and an 0.95 SD decrease in the abundance of *Streptococcus*, respectively (SE=1.21, PBHadj=2.7x10-3 and SE=0.27, PBHadj=0.39). Doubling the genetic liability of consuming skimmed milk was associated with a 3.36 SD increase in the abundance of the family Ruminococcaceae UCG013 (SE=0.88, PBHadj=0.01), while each SD increase in cheese intake frequency was associated with 1.2 increase in abundance of *Peptococcus* (SE=0.31, PBHadj=1.1x10-2) (Fig. 6B). Several interesting associations were detected at a more relaxed threshold ($P_{BHadj}<0.1$) (Table S16), including the potential effect of alcohol intake frequency on

increased abundance of *Ruminococcus*1 (0.37 SD for each SD increase in intake frequency, SE=0.12, P_{BHadj}=0.09), a causal link that has also been observed in animals[36].

Discussion

We report here on the relationship between host genetics and gut microbiome composition in 18,473 individuals from 25 population-based cohorts of European, Hispanic, Middle Eastern, Asian and African ancestries. We have estimated the heritability of the human gut microbiome and the effect of host genetics on the presence and abundance of individual microbial taxa (mbTL analysis) profiled using 16S rRNA gene sequencing. We studied the heterogeneity of the mbTLs signals and characterized the impact of technical and biological factors on their effect size. Lastly, we explored the relevance of the identified mbTLs to human disease and health-related traits using gene-set enrichment analysis, phenome-wide association studies and Mendelian randomization approaches.

Our large, multi-ethnic study allowed for an informative investigation of the human gut microbiome, providing a snapshot of its composition across different ethnicities and geographic locations. However, our analyses were complicated by two factors: the large heterogeneity in the data that reflects biological differences in the cohorts' ethnicities and ages and the methodological effects introduced by the different approaches used for collecting and processing samples. Overall, eight different methods were used to extract DNA from fecal samples, which strongly influenced the proportions of identified bacteria[13]. In addition, three different variable regions of 16S rRNA gene were chosen for amplification, which also vastly misrepresent microbiome composition[37]. Together with the variation in participants' ethnicity, age and BMI across cohorts, this led to a remarkable variation in microbiome richness, diversity and composition. These variations are likely also influenced by differences in diet, medication, lifestyle and other factors known to affect the microbiome composition[3,4], but these data were not available for all the cohorts and therefore not included in our analysis. Of the 410 genera identified in all cohorts with a presence rate higher than 1%, only nine genera were found in more than 95% of samples; these form the core microbiome. This combination of technical and biological heterogeneity led to the substantial variation seen in the microbiome composition, which reduced the power of mbTL analysis. Despite the large total sample size of 18,473

participants, the actual power to detect taxon-specific mbQTLs was limited by the taxon presence rate; excluding taxa with a sample size less than 3,000 individuals resulted in 211 taxa for mbQTL and 173 taxa for mbBTL mapping.

We performed three types of genetic analysis of microbiome composition: we analyzed the effect of genetics on alpha diversity, bacterial abundance and bacterial presence. We did not identify a genetic effect on bacterial diversity, which was expected given the lack of detectable heritability of this trait. Thirty taxon-specific mbTLs were identified at the genome-wide significance level of $5x10^{-8}$. Most of them (20 out of 30) affected taxa abundance, while the remaining 10 affected taxa presence. Even with our large sample size, the number of mbTLs revealed and their significance is rather modest. Despite the number of microbial traits (257 taxa) included in the study, only one locus passed a conservative study-wide significance threshold of $P<1.94x10^{-10}$: the association of the *LCT* locus with *Bifidobacterium* ($P=8.63x10^{-21}$). The statistical significance of the top-SNPs for the rest of the mbTLs lay in the suggestive zone, between a study-wide threshold of $P<1.94x10^{-10}$ and a nominal genome-wide significance of $P<5x10^{-8}$. However, our mbTL mapping results are concordant with heritability analyses: the heritable taxa tend to have more genome-wide significant loci and suggestively associated loci, and twin-based heritability is significantly correlated with SNP-based heritability. Our results confirm that only a subset of gut bacteria are heritable and that the genetic architecture affecting the abundance of heritable taxa is complex and polygenic. They also confirm the claim that the overall impact of host genetics on gut microbiome composition is rather modest compared to the effect of the environment[5].

The large heterogeneity in microbiome composition we have revealed points to the need to formulate guidelines for future microbiome GWAS. Standardized methodology should be aspired to, not only in computational analysis pipelines, but also in sample collection and storage, and especially DNA extraction and 16S PCR primers. Furthermore, we would advise researchers to aim for homogeneity in participants' ages and ethnicities, given the age- and ethnicity/geography-dependent mbQTL heterogeneity of *Bifidobacterium* and *Oxalobacter* mbQTLs. Finally, the sample size and sequencing depth needs to be further increased in order to capture a wider range of taxa with lower abundance and prevalence.

The strongest association determined in this study was that between the *LCT* locus and the *Bifidobacterium* genus. It has been shown that the functional SNP in the *LCT* locus,

rs4988235, determines not only the abundance of the *Bifidobacterium* genus, but also the strength of association between the genus and milk/dairy product intake[8]. Here, we showed the age-dependent nature of the *LCT-Bifidobacterium* association – the effect is weaker in children and adolescents and in populations of non-European ancestry, which is in line with current knowledge on lactose intolerance[38,39], while the strongest effect was observed in the HCHS/SOL cohort which comprises individuals of Hispanic/Latin American ethnicity and shows the highest prevalence of the rs182549*C/C genotype (683 out of 1,097 individuals). The lactose intolerant allele of the top *LCT* SNP (rs182549*C, associated with increased abundance of Bifidobacteria) was also associated with increased T2DM diagnosis and risk in family members in the UK Biobank study[29]. On the other hand, the T (lactose tolerant) allele of the same SNP is associated with increased BMI and waist circumference, but not with circulating concentration of glucose or insulin. Thus, the possible association between lactose intolerance and T2DM might be mediated through lower calcium intake, which is known to increase the risk of T2DM[40], rather than through BMI.

To explore the potential functional effects of mbTLs on health-related traits, we used GSEA, PheWAS and MR approaches. The GSEA indicated enrichment of mbQTLs for genes expressed in the small intestine and brain. Both tissues are known to be strongly connected to microbiome composition, so our results support the role of the gut microbiome on the gut-brain axis, and likely in gastrointestinal, brain and mood disorders, which have been the focus of several studies, e.g.[41–43]. The PheWAS analysis revealed a significant overlap between the genetic effect on gut microbes and a broad range of host phenotypes, including metabolic traits (6 mbTLs), cell signaling traits (3 mbTLs), immunological traits (2 mbTLs), nutritional phenotypes (2 mbTLs), psychiatric traits (2 mbTLs) and other phenotype groups (Table S14). The PheWAS enrichment analysis indicated that genetic determinants of bacterial abundance are also involved in regulating host metabolism, particularly for obesity-related traits. Among the interesting bacteria, earlier studies have linked the relative abundances of *Ruminococcus*[44], *Lachnospiraceae*[45] and *Ruminococcaceae*[46] to obesity.

Genetic anchors to microbiome variation also allow for estimation of causal links with complex traits through MR approaches[47–49]. Our results indicate that some autoimmune diseases, in particular asthma and celiac disease, may affect the abundance of specific bacterial taxa. Moreover, our MR results provide evidence of a protective effect from Actinobacteria in

ulcerative colitis, a link that was proposed in several cross-sectional studies that reported an increased abundance of Actinobacteria in healthy individuals compared to IBD patients[50,51], although these results were not always consistent across studies[52,53]. Consistent with our observation, the most abundant species from the class Actinobacteria, *Bifidobacterium*, was previously shown to have a beneficial effect on ulcerative colitis in a clinical trial[52,54].

Finally, we observed that nutritional phenotypes are also causally linked to changes in bacteria abundance, but not *vice versa*. However, it is important to realize that there are likely to be other causal relationships between bacteria and phenotypes that we were unable to detect due to the limited power of the microbiome GWAS, the modest size of a genetic component for either the bacteria or the nutritional phenotype, or the presence of more complex bi-directional relationships that are difficult to detect with the current data. Nonetheless, our results generally support the hypothesis that the genetic effect on the microbiome composition can be mediated by diet and is affected by diseases. The PheWAS analysis indicated that SNPs associated with the *LCT* and *FUT2* loci are also associated to dietary preference factors, including fish, cereal, bread, alcohol, vegetable and ground coffee intake, along with other diet-related phenotypes. MR analysis indicated a causal effect of dietary preference factors on several bacteria, but not *vice versa*; we did not identify any bacteria that affect nutritional phenotypes. The genes found to be associated with mbTLs also included olfactory receptors and genes involved in the absorption and metabolism of B2 and B12 vitamins, such as *RFK*, *CUBN* and *OR1F1*. It has been proposed that the genetic determinants of dietary factors affect the gut microbiome composition[55]; our results confirm this hypothesis and we have indicated links between certain nutritional categories and specific microbial taxa.

In summary, here we report the largest study to date to investigate the genetics of human microbiome across multiple ethnicities. The observed heterogeneity and high inter-individual variability of the microbiome substantially reduces the statistical power of the analysis. Consequently, similar to early GWAS studies, we reported limited number of associated loci. Nevertheless, our results point to causal relationships between specific loci and bacterial taxa and health-related traits. Heritability estimates suggest that these associations are likely part of a larger spectrum, currently undetectable in the existing sample size. This warrants future studies that should take advantage of larger sample sizes, harmonized protocols and more advanced microbiome analysis methods, including metagenomics sequencing instead of 16S profiling and

quantifying bacterial cell counts. Given the essential role of the gut microbiome in the metabolism of food and drugs, our results contribute to the development of personalized nutrition and medication strategies based on both host genomics and microbiome data.

Acknowledgements

Authors declare no conflict of interest.

We thank Jackie Senior and Kate McIntyre for editing the manuscript.

The funding and acknowledgements per cohorts are given in Cohort Acknowledgements section of Supplementary materials.

All GWAS summary statistics is available on www.mibiogen.org built using MOLGENIS framework[56].

References

1. Sender, R., Fuchs, S. & Milo, R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLOS Biol.* **14**, e1002533 (2016).
2. Gilbert, J. A. *et al.* Current understanding of the human microbiome. *Nat. Med.* **24**, 392–400 (2018).
3. Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).
4. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).
5. Rothschild, D. *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).
6. Goodrich, J. K. *et al.* Human Genetics Shape the Gut Microbiome. *Cell* **159**, 789–799 (2014).
7. Goodrich, J. K. *et al.* Genetic Determinants of the Gut Microbiome in UK Twins. *Cell Host Microbe* **19**, 731–743 (2016).
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. Kurilshikov, A., Wijmenga, C., Fu, J. & Zhernakova, A. Host Genetics and Gut Microbiome: Challenges and Perspectives. *Trends Immunol.* **38**, (2017).
12. Wang, J. *et al.* Meta-analysis of human genome-microbiome association studies: the

MiBioGen consortium initiative. *Microbiome* **6**, 101 (2018).

13. Sinha, R. *et al.* Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. *Nat. Biotechnol.* **35**, 1077–1086 (2017).
14. Sinha, R., Abnet, C. C., White, O., Knight, R. & Huttenhower, C. The microbiome quality control project: baseline study design and future directions. *Genome Biol.* **16**, 276 (2015).
15. Vandeputte, D., Tito, R. Y., Vanleeuwen, R., Falony, G. & Raes, J. Practical considerations for large-scale gut microbiome studies. *FEMS Microbiol. Rev.* **41**, S154–S167 (2017).
16. Louis, P., Young, P., Holtrop, G. & Flint, H. J. Diversity of human colonic butyrate-producing bacteria revealed by analysis of the butyryl-CoA:acetate CoA-transferase gene. *Environ. Microbiol.* **12**, 304–314 (2010).
17. Blekhman, R. *et al.* Host genetic variation impacts microbiome composition across human body sites. *Genome Biol.* **16**, 191 (2015).
18. Zhernakova, D. V *et al.* Individual variations in cardiovascular-disease-related protein levels are driven by genetics and gut microbiome. *Nat. Genet.* **50**, 1524–1532 (2018).
19. Deelen, P. *et al.* Improving the diagnostic yield of exome- sequencing by predicting gene–phenotype associations using large-scale gene expression analysis. *Nat. Commun.* **10**, 2837 (2019).
20. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555–7 (2015).
21. Kashyap, P. C. *et al.* Genetically dictated change in host mucus carbohydrate landscape exerts a diet-dependent effect on the gut microbiota. *Proc. Natl. Acad. Sci.* **110**, 17059–17064 (2013).
22. Crost, E. H. *et al.* Mechanistic Insights Into the Cross-Feeding of Ruminococcus gnavus and Ruminococcus bromii on Host and Dietary Carbohydrates. *Front. Microbiol.* **9**, 2558 (2018).
23. Magnúsdóttir, S., Ravcheev, D., de Crécy-Lagard, V. & Thiele, I. Systematic genome assessment of B-vitamin biosynthesis suggests co-operation among gut microbes. *Front. Genet.* **6**, (2015).
24. Yoshii, K., Hosomi, K., Sawane, K. & Kunisawa, J. Metabolism of Dietary and Microbial Vitamin B Family in the Regulation of Host Immunity. *Front. Nutr.* **6**, (2019).
25. Haas, M. E. *et al.* Genetic Association of Albuminuria with Cardiometabolic Disease and Blood Pressure. *Am. J. Hum. Genet.* **103**, 461–473 (2018).
26. Rowley, C. A. & Kendall, M. M. To B12 or not to B12: Five questions on the role of cobalamin in host-microbial interactions. *PLOS Pathog.* **15**, e1007479 (2019).
27. Xu, Y. *et al.* Cobalamin (Vitamin B12) Induced a Shift in Microbial Composition and Metabolic Activity in an in vitro Colon Simulation. *Front. Microbiol.* **9**, (2018).
28. Gysemans, C. *et al.* Interferon regulatory factor-1 is a key transcription factor in murine beta cells under immune attack. *Diabetologia* **52**, 2374–2384 (2009).
29. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).
30. Nicklas, T. A. *et al.* Self-perceived lactose intolerance results in lower intakes of calcium and dairy foods and is associated with hypertension and diabetes in adults. *Am. J. Clin. Nutr.* **94**, 191–8 (2011).

31. Shin, S.-Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**, 543–550 (2014).
32. Suhre, K. *et al.* Metabolic Footprint of Diabetes: A Multiplatform Metabolomics Study in an Epidemiological Setting. *PLoS One* **5**, e13953 (2010).
33. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7**, (2018).
34. Wade, K. H. & Hall, L. J. Improving causality in microbiome research: can human genetic epidemiology help? *Wellcome Open Res.* **4**, 199 (2019).
35. Zhang, L. *et al.* Effects of Gliadin consumption on the Intestinal Microbiota and Metabolic Homeostasis in Mice Fed a High-fat Diet. *Sci. Rep.* **7**, 44613 (2017).
36. Posteraro, B. *et al.* Liver Injury, Endotoxemia, and Their Relationship to Intestinal Microbiota Composition in Alcohol-Preferring Rats. *Alcohol. Clin. Exp. Res.* **42**, 2313–2325 (2018).
37. Brooks, J. P. *et al.* The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiol.* **15**, 66 (2015).
38. Coluccia, E. *et al.* Congruency of Genetic Predisposition to Lactase Persistence and Lactose Breath Test. *Nutrients* **11**, 1383 (2019).
39. Lapides, R. A. & Savaiano, D. A. Gender, Age, Race and Lactose Intolerance: Is There Evidence to Support a Differential Symptom Response? A Scoping Review. *Nutrients* **10**, (2018).
40. Pittas, A. G., Lau, J., Hu, F. B. & Dawson-Hughes, B. The Role of Vitamin D and Calcium in Type 2 Diabetes. A Systematic Review and Meta-Analysis. *J. Clin. Endocrinol. Metab.* **92**, 2017–2029 (2007).
41. Valles-Colomer, M. *et al.* The neuroactive potential of the human gut microbiota in quality of life and depression. *Nat. Microbiol.* **4**, (2019).
42. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
43. Vich Vila, A. *et al.* Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. *Sci. Transl. Med.* **10**, eaap8914 (2018).
44. Ottosson, F. *et al.* Connection Between BMI-Related Plasma Metabolite Profile and Gut Microbiota. *J. Clin. Endocrinol. Metab.* **103**, 1491–1501 (2018).
45. Tun, H. M. *et al.* Roles of Birth Mode and Infant Gut Microbiota in Intergenerational Transmission of Overweight and Obesity From Mother to Offspring. *JAMA Pediatr.* **172**, 368–377 (2018).
46. Finnicum, C. T. *et al.* Metataxonomic Analysis of Individuals at BMI Extremes and Monozygotic Twins Discordant for BMI. *Twin Res. Hum. Genet.* **21**, 203–213 (2018).
47. Sanna, S. *et al.* Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nat. Genet.* **51**, 600–605 (2019).
48. Jia, J. *et al.* Assessment of Causal Direction Between Gut Microbiota–Dependent Metabolites and Cardiometabolic Health: A Bidirectional Mendelian Randomization Analysis. *Diabetes* **68**, 1747–1755 (2019).
49. Yang, Q., Lin, S. L., Kwok, M. K., Leung, G. M. & Schooling, C. M. The Roles of 27 Genera of Human Gut Microbiota in Ischemic Heart Disease, Type 2 Diabetes Mellitus, and Their Risk Factors: A Mendelian Randomization Study. *Am. J. Epidemiol.* **187**, 1916–1922 (2018).
50. Rinninella, E. *et al.* What is the Healthy Gut Microbiota Composition? A Changing

Ecosystem across Age, Environment, Diet, and Diseases. *Microorganisms* **7**, 14 (2019).

51. Plichta, D. R., Graham, D. B., Subramanian, S. & Xavier, R. J. Therapeutic Opportunities in Inflammatory Bowel Disease: Mechanistic Dissection of Host-Microbiome Relationships. *Cell* **178**, 1041–1056 (2019).
52. Frank, D. N. *et al.* Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl. Acad. Sci.* **104**, 13780–13785 (2007).
53. Morgan, X. C. *et al.* Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* **13**, R79 (2012).
54. Tursi, A. *et al.* Treatment of Relapsing Mild-to-Moderate Ulcerative Colitis With the Probiotic VSL#3 as Adjunctive to a Standard Pharmaceutical Treatment: A Double-Blind, Randomized, Placebo-Controlled Study. *Am. J. Gastroenterol.* **105**, 2218–2227 (2010).
55. Goodrich, J. K., Davenport, E. R., Waters, J. L., Clark, A. G. & Ley, R. E. Cross-species comparisons of host genetic associations with the microbiome. *Science* **352**, 29–32 (2016).
56. Swertz, M. A. *et al.* The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button. *BMC Bioinformatics* **11**, S12 (2010).
57. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2012).
58. Cole, J. R. *et al.* The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* **37**, D141-5 (2009).
59. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
60. Howie, B., Marchini, J. & Stephens, M. Genotype Imputation with Thousands of Genomes. *G3* **1**, 457–470 (2011).
61. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
62. Carmi, S. *et al.* Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. *Nat. Commun.* **5**, 4835 (2014).
63. Deelen, P. *et al.* Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Res. Notes* **7**, 901 (2014).
64. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
65. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* (2016). doi:10.1038/ng.3538
66. Cochran, W. G. The Combination of Estimates from Different Experiments. *Biometrics* **10**, 101 (1954).
67. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
68. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
69. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
70. Võsa, U. *et al.* Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv* 447367 (2018). doi:10.1101/447367

71. Pers, T. H., Timshel, P. & Hirschhorn, J. N. SNPsnap: a Web-based tool for identification and annotation of matched SNPs. *Bioinformatics* **31**, 418–420 (2015).
72. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–41 (2015).
73. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
74. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012).
75. Karlsson, F. H. *et al.* Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498**, 99–103 (2013).
76. Koeth, R. a *et al.* Intestinal microbiota metabolism of l-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nat. Med.* **19**, 576–85 (2013).
77. Coit, P. & Sawalha, A. H. The human microbiome in rheumatic autoimmune diseases: A comprehensive review. *Clin. Immunol.* **170**, 70–79 (2016).
78. Vatanen, T. *et al.* Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. *Cell* **165**, 842–853 (2016).
79. O'Mahony, S. M., Clarke, G., Borre, Y. E., Dinan, T. G. & Cryan, J. F. Serotonin, tryptophan metabolism and the brain-gut-microbiome axis. *Behav. Brain Res.* **277**, 32–48 (2015).
80. RAPID GWAS OF THOUSANDS OF PHENOTYPES FOR 337,000 SAMPLES IN THE UK BIOBANK. Available at: http://www.nealelab.is/uk-biobank.
81. Bowden, J. *et al.* A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Stat. Med.* **36**, 1783–1802 (2017).
82. Hartwig, F. P., Davey Smith, G. & Bowden, J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int. J. Epidemiol.* **46**, 1985–1998 (2017).
83. 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).
84. Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* **50**, 693–698 (2018).
85. Shim, H. *et al.* A Multivariate Genome-Wide Association Analysis of 10 LDL Subfractions, and Their Response to Statin Treatment, in 1868 Caucasians. *PLoS One* **10**, e0120758 (2015).
86. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data. *Genet. Epidemiol.* **37**, 658–665 (2013).
87. von Rhein, D. *et al.* The NeuroIMAGE study: a prospective phenotypic, cognitive, genetic and MRI study in children with attention-deficit/hyperactivity disorder. Design and descriptives. *Eur. Child Adolesc. Psychiatry* **24**, 265–281 (2015).
88. Naaijen, J. *et al.* COMPULS: design of a multicenter phenotypic, cognitive, genetic, and magnetic resonance imaging study in children with compulsive syndromes. *BMC Psychiatry* **16**, 361 (2016).
89. Fernández-Calleja, J. M. S. *et al.* Non-invasive continuous real-time in vivo analysis of microbial hydrogen production shows adaptation to fermentable carbohydrates in mice.

Sci. Rep. **8**, 15351 (2018).

90. Bisgaard, H. *et al.* Deep phenotyping of the unselected COPSAC 2010 birth cohort study. *Clin. Exp. Allergy* **43**, 1384–1394 (2013).
91. Bisgaard, H. The Copenhagen Prospective Study on Asthma in Childhood (COPSAC): design, rationale, and baseline data from a longitudinal birth cohort study. *Ann. Allergy, Asthma Immunol.* **93**, 381–389 (2004).
92. Stokholm, J. *et al.* Maturation of the gut microbiome and risk of asthma in childhood. *Nat. Commun.* **9**, 141 (2018).
93. Dantoft, T. M. *et al.* Cohort description: The Danish study of Functional Disorders. *Clin. Epidemiol.* **Volume 9**, 127–139 (2017).
94. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6**, 1621–1624 (2012).
95. Kooijman, M. N. *et al.* The Generation R Study: design and cohort update 2017. *Eur. J. Epidemiol.* **31**, 1243–1264 (2016).
96. Medina-Gomez, C. *et al.* Challenges in conducting genome-wide association studies in highly admixed multi-ethnic populations: the Generation R Study. *Eur. J. Epidemiol.* **30**, 317–330 (2015).
97. Radjabzadeh, D. *et al.* Diversity, compositional and functional differences between gut microbiota of children and adults. *Sci. Rep.* **10**, 1040 (2020).
98. Hamza, T. H. *et al.* Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. *Nat. Genet.* **42**, 781–5 (2010).
99. Hill-Burns, E. M. *et al.* Parkinson's disease and Parkinson's disease medications have distinct signatures of the gut microbiome. *Mov. Disord.* **32**, 739–749 (2017).
100. Willemsen, G. *et al.* The Netherlands Twin Register Biobank: A Resource for Genetic Epidemiological Studies. *Twin Res. Hum. Genet.* **13**, 231–245 (2010).
101. Walter, S. A., Kjellström, L., Nyhlin, H., Talley, N. J. & Agréus, L. Assessment of normal bowel habits in the general adult population: the Popcol study. *Scand. J. Gastroenterol.* **45**, 556–566 (2010).
102. Kjellström, L. *et al.* A randomly selected population sample undergoing colonoscopy. *Eur. J. Gastroenterol. Hepatol.* **26**, 268–275 (2014).
103. Ikram, M. A. *et al.* The Rotterdam Study: 2018 update on objectives, design and main results. *Eur. J. Epidemiol.* **32**, 807–850 (2017).
104. Volzke, H. *et al.* Cohort Profile: The Study of Health in Pomerania. *Int. J. Epidemiol.* **40**, 294–307 (2011).
105. Frost, F. *et al.* Impaired Exocrine Pancreatic Function Associates With Changes in Intestinal Microbiota Composition and Diversity. *Gastroenterology* **156**, 1010–1015 (2019).
106. LaVange, L. M. *et al.* Sample Design and Cohort Selection in the Hispanic Community Health Study/Study of Latinos. *Ann. Epidemiol.* **20**, 642–649 (2010).
107. Sorlie, P. D. *et al.* Design and Implementation of the Hispanic Community Health Study/Study of Latinos. *Ann. Epidemiol.* **20**, 629–641 (2010).
108. Conomos, M. P. *et al.* Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos. *Am. J. Hum. Genet.* **98**, 165–184 (2016).
109. Marotz, C. *et al.* DNA extraction for streamlined metagenomics of diverse environmental samples. *Biotechniques* **62**, (2017).

110. Verdi, S. *et al.* TwinsUK: The UK Adult Twin Registry Update. *Twin Res. Hum. Genet.* 1–7 (2019). doi:10.1017/thg.2019.65