# Genome-wide associations of human gut microbiome variation and implications for causal inference analyses

David A Hughes[1,2*], Rodrigo Bacigalupe[3,4*], Jun Wang[3,4,5], Malte C Rühlemann[6], Raul Y. Tito[3,4], Gwen Falony[3,4], Marie Joossens[3,4], Sara Vieira-Silva[3,4], Liesbet Henckaerts[7,8], Leen Rymenans[3,4], Chloë Verspecht[3,4], Susan Ring[2,9], Andre Franke[6], Kaitlin H. Wade[1,2], Nicholas J. Timpson[1,2**], Jeroen Raes[3,4**].

*co-first authors

**co-corresponding authors

[1]MRC Integrative Epidemiology Unit at University of Bristol, Bristol, BS8 2BN, UK.

[2]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, BS8 2BN, UK.

[3]Department of Microbiology and Immunology, Rega Instituut, KU Leuven–University of Leuven, Leuven, Belgium.

[4]Center for Microbiology, VIB, Leuven, Belgium.

[5]Institute of Microbiology, Chinese Academy of Sciences, Chaoyang District, 100101 Beijing, China.

[6]Institute of Clinical Molecular Biology, Christian Albrechts University of Kiel, Kiel, Germany.

[7]KU Leuven-University of Leuven, Department of Microbiology, Immunology and Transplantation, Leuven, Belgium.

[8]KU Leuven-University Hospitals Leuven, Department of General Internal Medicine, Leuven, Belgium.

[9]Bristol Bioresource Laboratories (BBL), University of Bristol, UK.

## Abstract

Recent population-based[1–4] and clinical studies[5] have identified a range of factors associated with human gut microbiome variation. Murine quantitative trait loci[6], human twin studies[7] and microbiome genome-wide association studies (mGWAS)[1,3,8–12] have provided evidence for genetic contributions to microbiome composition. Despite this, there is still poor overlap in genetic association across human studies. Using appropriate taxon-specific models along with support from independent cohorts, we show association between human host genotype and gut microbiome variation. We also suggest that interpretation of applied analyses using genetic associations is complicated by the likely overlap between genetic contributions and heritable components of host environment. Using fecal derived 16S rRNA gene sequences and host genotype data from the Flemish Gut Flora Project (FGFP, n=2223) and two German cohorts (FoCus, n=950, PopGen n=717), we identify genetic associations involving multiple microbial traits (MTs). Two of these associations achieved a study-level p-value threshold of $1.57 \times 10^{-10}$; an association between *Ruminococcus* and rs150018970 near *RAPGEF1* on chromosome 9, and between *Coprococcus* and rs561177583 within *LINC01787* on chromosome 1. Exploratory analysis was undertaken using 11 other genome-wide associations with strong evidence for association (p-value < $2.5 \times 10^{-08}$) and a previously reported signal of association between rs4988235 (*MCM6*/*LCT*) and *Bifidobacterium*. Across these 14 SNPs there was evidence of signal overlap with other GWAS including those for age at menarche and cardiometabolic traits. Mendelian randomization (MR) analysis was able to estimate associations between MTs and disease (including *Bifidobacterium* and body composition), however in the absence of clear microbiome driven effects, caution is needed in interpretation. Overall, this work marks a growing catalog of genetic associations which will provide insight into the contribution of host genotype to gut microbiome. Despite this, the uncertain origin of association signals will likely complicate future work looking to dissect function or use associations for causal inference analysis.

## Main

Human host-microbiome mGWAS are still in their infancy and feature a paucity of overlap for even the most compelling signals across studies[13]. This is an observation influenced by environmental variables dominating microbial trait variation[1] and the complications of variation in sample collection, storage conditions, DNA extraction method, PCR primers, and amplicon versus shotgun sequencing[14]. While recent advances are improving resolution and reliability of microbiome profiles[15], inter-study analytical methodologies analyzing those profiles vary extensively (Table S1). Microbiota profiles, being the product of ecological sampling, are often zero-inflated with varying distributions across taxa. Consequently, variation in modeling, normalization procedures and choices of diversity indicators can greatly influence results across studies.

In an attempt to identify persistent signals of association between host genetic variation and human gut microbiome, we harmonized the analytical pipeline across three independent studies: an expanded release of the FGFP cohort (Flanders, Belgium; n=2,223) and two German cohorts (Food-Chain Plus[11] (FoCus; n=950) and the PopGen[16] cohort (n=717)). Of the initial 499 derived taxon abundances in FGFP (Table S2), 139, across all phylogenetic levels, met our analysis criteria and 92 were retained after identifying independent phenotypes (Methods; Fig. 1). Microbial taxa were described as relative abundance (AB) profiles and those with zero-inflated abundance distributions (67% or 62 of the 92 retained taxa) were described using a hurdle model[11]. That is, for taxa where more than 5% of individuals in FGFP had an abundance measurement of zero, we generated a presence/absence (P/A) phenotype and a zero-truncated (all zero values set as missing) abundance (AB) phenotype. We note that absence does not indicate non-existence, but that a taxon is not observed under the current sequencing depth. A comparison of data preparation methods led to the choices above, as alternatives used previously[1,7,11,12] failed to consistently account for skewness and categorical distributions across all outcomes in a GWAS context (Supplementary Fig. 1 and 2 and Supplementary Information: Distribution /model choice). In addition, we computed the total taxa richness present within each sample (α-diversity), compositional variation between samples (β-diversity) and the different community composition types (enterotypes). A total of 95 continuous (92 AB + three α-diversity), 62 binary (P/A), one multinomial (enterotype), and one multivariate (β-diversity) traits were carried forward to further analysis as microbial traits (MTs).

We first estimated the proportion of gut microbiota variation explained by genetic variation among individuals by estimating narrow sense genetic heritability ($h^2$) for each AB, P/A, and α-diversity trait in FGFP (Methods). Heritability ranged from 0 to 0.47, with 13 of the 157 continuous and binary MTs tested exhibiting non-zero estimates (likelihood ratio test p-value < 0.05; Fig. 1; Table S3). Eight of the 13 MTs noted above are from the phylum Firmicutes, five of which are from the family *Lachnospiraceae* and two from *Ruminococcaceae*. The most heritable MT observed was genus *Hespellia* ($h^2$ = 0.47, se = 0.18) of family *Lachnospiraceae*, class Clostridia. Among the highly prevalent genera (present in >98.5% of samples), the most heritable were *Dorea* ($h^2$ = 0.25, se = 0.14) and *Anaerostipes* (mean $h^2$ = 0.23, se = 0.13), key short-chain fatty acids-producing and health-associated genera[17]. Heritability estimates were also generated with log and box-cox transformed data to allow for comparison with previous work (Table S3). Heritability estimates derived from FGFP using other data transformations had Spearman's rho correlation coefficients of 0.95 (rank normal transformation (RNT) vs log2) and 0.53 (RNT vs box-cox; Extended Data Fig. 1a and 1b). Inter-study correlation coefficients of 0.28 and 0.23 were observed when comparing heritability estimates derived with similar data transformations[7,9] (Extended Data Fig. 1c and 1d). The low values are likely driven by poor power when estimating heritability across studies, though with temporal, local and individual environments influencing MT variation, heritability estimates may be inflated or deflated. Larger and/or environmentally controlled designs will be required to increase the power and accuracy of MT heritability estimates.

Associations between genetic variants and specific MTs were identified by fitting linear (AB + α-diversity; Fig. 2a), logistic (P/A; Fig. 2b), multinomial (enterotype), and multivariate (β-diversity) regressions assuming an additive genetic model and accounting for genotype uncertainty (Methods). In addition, human autosomal copy number variations (CNVs) were tested for associations to MTs in the FGFP data set, but none yielded strong evidence of association (Table S4; Methods). All single nucleotide variants at an inclusive association Score test p-value threshold of $<1x10^{-05}$ in the FGFP dataset (n=23,735) were taken forward into a targeted meta-analysis including two independent German cohorts. Three genera were not present in the German cohorts, a likely product of using a different hypervariable region of the 16S rRNA locus, limiting our meta-analysis to 153 MTs and 23,067 variants to analyze.

Two variants showed evidence of association that exceeded a study-wide meta-analysis p-value threshold of $1.57x10^{-10}$ (Table 1, Fig. 2c, Table S5, Extended Data Fig. 2). The strongest of these was between *Ruminococcus* (P/A) and rs150018970, an intergenic variant 33 kilo-bases upstream the *RAPGEF1* gene on chromosome 9. *RAPGEF1* encodes a protein factor that transduces signals from G-protein-coupled receptors (GPCRs), which are likely involved in the regulation of the physiology of the gastrointestinal tract[18]. GPCRs detect metabolites derived from commensal bacteria and have been proposed to be key mediators of host–microbial interactions[18] Relative to homozygous reference allele individuals, heterozygous individuals were less likely to have *Ruminococcus* (OR=0.111, 95% CI=0.062-0.197, meta-analysis p-value=$6.68x10^{-14}$), core members of the gut microbiota. The second association was *Coprococcus* (P/A) and rs561177583 on chromosome 1, sitting in the intron of the non-protein coding RNA *LINC01787*; with individuals heterozygous for the effect allele also being less likely to have *Coprococcus* in their sample (OR=0.161, 95% CI =0.09-0.28, meta-analysis p-value=1.10x10$^{-10}$). Despite the strength of their association, heterogeneity between the studies for these effects was high and further investigation is required to confirm and characterize these signals (Extended Data Fig. 2).

Following stringent filters for lead association signals, we also examined the properties of results with strong evidence for reliability. Satisfying a GWAS evidence threshold meta-analysis p-value < 2.5x10$^{-08}$, 11 associations showed low heterogeneity in meta-analysis (Table 1, Fig. 2c, Table S5, Extended Data Fig. 2). These contained the butyrate-producing genus *Butyricicoccus* associated with an eQTL (in multiple tissues including brain, data from GTEx portal) for *SLC5A11* on chromosome 16 (rs72770483, meta-analysis p-value=5.54x10$^{-10}$; Fig. 2f)[19]. *SLC5A11* encodes a sodium-dependent myo-inositol/glucose cotransporter[20], highly expressed in the brain and intestine, where it participates in appetite control and glycemic regulation[21,22]. Observations here suggest a role for *Butyricicoccus* in the formation of glycemic traits and is consistent with studies suggesting that butyrate-producing bacteria are associated with blood glucose regulation[23] and insulin sensitivity in mice[24]. Separately, P/A of *Veillonella* was associated with rs117338748 (meta-analysis p-value=2.42x10$^{-08}$; Fig. 2e) an eQTL for *LIPC* (in multiple tissues including thyroid, data from GTEx portal), which encodes the hepatic lipase enzyme involved in regulation of low-density lipoproteins (LDLs) and the transport of high-density lipoproteins (HDLs)[19]. Bacteria of the genus *Veillonella* are typical lactate fermenters that produce acetate, a substrate for lipogensis[25], and propionate, which inhibits lipid synthesis[26]. Using data from FGFP (Table S6) to test for association with LDL levels (Methods and Supplementary Information), we observed that presence of *Veillonella* was observationally associated with a decrease in LDL-cholesterol by 5.39mg/dL after accounting for sex, age and the top 10 genetic principle components (F-test, p-value=5.99x10$^{-04}$).

We assessed overlap with previously reported mGWAS for (1) equivalent SNP-to-MT associations and (2) the strongest association regardless of the MT in the FGFP mGWAS (Tables S7-S9). We found a strong association overlapping with two genetic variants reported previously[7]. Located on chromosome two, they are rs4988235 (*MCM6*/*LCT*) and rs6730157 (*RAB3GAP1*), 701 kilo-bases apart. Both variants are associated with *Bifidobacterium* abundance and located within a block of linkage disequilibrium (CEPH European : CEU r2 =

0.81)[27]. The association between rs4988235 and *Bifidobacterium* abundance is the only previously replicated mGWAS signal[1,7,8,10–12]. This association does not meet our study-wide or genome-wide threshold (meta-analysis β=-0.128, se=0.026, p-value=$1.34x10^{-06}$), but as the only association seen in multiple studies, it remains the most reliable host genetic contribution to MT[7]. To highlight regions that potentially contribute to multiple MTs, we additionally queried if previously reported variants (n = 522[1,7,8,10–12]) gave evidence of association with any MT. Of all results residing outside the *LCT* region, five SNP-MT associations have p-values that survive Bonferroni correction in this targeted analysis (Score test p-value < 0.05/522; Supplementary Information)[1].

We searched PhenoScanner V2[28] for previously identified associations across variants with strong evidence for reliability (Table 1) and for rs4988235 at the *MCM6* locus (Tables S10 and S11). Ten of those 14 variants have associations with other phenotypes at a p-value < $1x10^{-5}$ (Bonferroni p-value = 0.05/5000 phenotypes in database). Two of those variants have associations with other phenotypic traits that surpassed a p-value threshold of $2.5x10^{-8}$. They are unclassified Firmicutes:rs11788336 associated with age at menarche (p-value = $1.7x10^{-}10$) and *Bifidobacterium*:rs4988235 associated with total cholesterol (p-value =$3.98x10^{-14}$), low density lipoprotein (p-value = $3.22x10^{-11}$), forced vital capacity ($2.27x10^{-10}$), body fat percentage ($1.10x10^{-9}$), and numerous other obesogenic traits. To expand this bioinformatics screen, we also identified gene expression and biological pathway enrichments for the 1,361 genes closest to sites with improved evidence of association in meta-analysis (+/- 250kb) using GENE2FUNC and integrated hypergeometric tests of the FUMA platform[29] (Table S12). Gene expression enrichment (GTEx v7; Methods) was suggested for 30 tissues highlighted by brain (false discovery rate (FDR)=$2.25x10^{-34}$), kidney (FDR=$1.10x10^{-24}$), colon (FDR=$4.42x10^{-15}$), stomach (FDR=$7.62x10^{-14}$), and small intestine (FDR=$2.44x10^{-05}$). Abundant enrichment was observed in the GWAS Catalog for 438 gene sets, including the top category obesity-related traits (FDR=$1.12x10^{-31}$) and in addition, 137 Canonical Pathways showed evidence for enrichment (FDR < 0.05), Table S12).

Lastly, to explore the potential for associated microbiome variants to be used in causal inference methods[30], we performed a series of analyses to examine the utility of signals in a causal analysis framework known as Mendelian randomization (MR). Eleven metabolic health, inflammatory and neurological traits were selected *a priori* for this analysis and variants with strong evidence for reliability (Table 1 and the replicated lactase persistence variant, rs4988235) where used as proxies for microbiome variation in two-sample, bi-directional, MR analyses[30] (Extended Data Fig. 3). Analyses were able to estimate relationships between 5 MTs and 7 outcomes (Table 2, Tables S13 and S14). Amongst other hypothesis generating results, the strongest evidence from this analysis suggests that a lower *Bifidobacterium* abundance increases waist circumference (β=0.149SD, se=0.0290, Wald test p-value=2.82x10$^{-07}$) and body mass index (BMI) (β=0.124SD, se=0.027, Wald test p-value=3.37x10$^{-06}$) (Table 2, Table S13). However, other than analytical power and instrument strength, a complication of these analyses and potentially for all future MR analyses of this nature, is the likely impact of indirect or disease driven effects being upstream of microbiome variation. For example, each additional copy of the lactase persistent allele at rs4988235 is estimated to decrease *Bifidobacterium* abundance and as such, individuals liable to be lactase persistent have reduced *Bifidobacterium* abundance. This association could be the product of a direct effect of rs4988235 on *Bifidobacterium* abundance, however it could equally reflect a reverse effect where a host environment matching an ability to metabolize lactose alters microbiome profile. Consequently, using rs4988235 as a proxy marker for *Bifidobacterium* to generate estimates of causal effect(s) may be providing information about *Bifidobacterium* effects, but could equally be reporting on the impact of variation in dietary habits. Indeed, the prominence of host environmental effects in host mGWAS (exemplified by the abundant overlap with GWAS catalog traits - Table S11 and S12), may be a common theme observed in genetic signals and ultimately the most parsimonious explanation for apparently causal effects in naïve MR analysis.

Using a targeted meta-analysis framework, including the largest cross-sectional study with host genetics and microbiome data available, combined with distinct modeling of different MTs, we have detected evidence for host genetic associations to the gut microbiome. While environmental effects are likely to preside over host genetics as source of variation, this work illustrates that, even in the presence of unavoidable study-based heterogeneity, standardized and appropriate analytical protocols allow signal detection. We note that this study is limited to genus-level microbiome traits and that host-microbial interaction signals might be more pertinent at species or strain levels, but strain-level GWAS require much larger population sizes and metagenomic sequencing. Additionally, we have shown that associated loci can be deployed in frameworks designed to explore function and causality in otherwise observational associations between MTs and human phenotypes. To that end, future large-scale meta-analyses will likely advance this type of endeavor by providing larger catalogues of genetic variants associated with microbiome, however this approach is unlikely to be straightforward. It appears likely that signals captured in this type of mGWAS reflect a microbial footprint of disease or behavior and this complexity will need to be accounted for in future analyses aiming to use human genetics to target causal effects of the gut microbiome[31]. Despite this, with a further expanded catalog of reliable loci contributing to microbiome variation, there will be greater insight into the contribution of host genotype to gut microbiome variation and better understanding of the relationship between gut microbiota, host molecular biology and disease.

## Acknowledgements

FGFP procedures were approved by the medical ethics committee of the University of Brussels–Brussels University Hospital (approval 143201215505, 5/12/2012). A declaration concerning the FGFP privacy policy was submitted to the Belgian Commission for the Protection of Privacy. Written informed consent was obtained from all participants. The FGFP was funded with support of the Flemish government (IWT130359), the Research Fund–Flanders (FWO) Odysseus program (G.0924.09), the King Baudouin Foundation (2012-J80000-004), FP7 METACARDIS HEALTH-F4-2012-305312, VIB, the Rega Institute for Medical Research, and KU Leuven. We acknowledge the contribution of Flemish GPs and pharmacists to data and sample collection. Finally, we thank all FGFP volunteers for participating in the project. RB is funded by the Research Fund–Flanders (FWO) through a Postdoctoral Fellowship (1221620N). SVS is supported by Marie Curie Actions FP7 People COFUND–Proposal 267139 (acronym OMICS@VIB). NJT is a Wellcome Trust Investigator (202802/Z/16/Z), is supported by the University of Bristol NIHR Biomedical Research Centre (BRC-1215-20011) and works within the CRUK Integrative Cancer Epidemiology Programme (C18281/A19169). SR and NJT work within a unit that receives funding from the MRC and University of Bristol (MC_UU_12013/3, MC_UU_00011/6). DH is supported by the Wellcome Investigator Award (202802/Z/16/Z). KHW is supported by the Elizabeth Blackwell Institute for Health Research, University of Bristol and the Wellcome Trust Institutional Strategic Support Fund (204813/Z/16/Z). MR was funded by the German Research Foundation (DFG) Collaborative Research Center 1182 (CRC1182; Origin and Function of Metaorganisms).

## Author contributions

JR and NT conceived and designed the study. JW, RYT, GF, MJ, SVS and LH performed the sampling and meta-data collection. JW, RYT, LR and CV extracted and sequenced DNA. DAH, RB, KHW, JW, and MR performed the microbiome GWAS and MR analysis. DAH, RB, JW, KHW, NT, and JR wrote the manuscript. All authors revised and commented on the manuscript.

## Competing interests

Authors declare no conflict of interest.

## Tables and Figures

### Table 1: Meta-supported genetic variants.

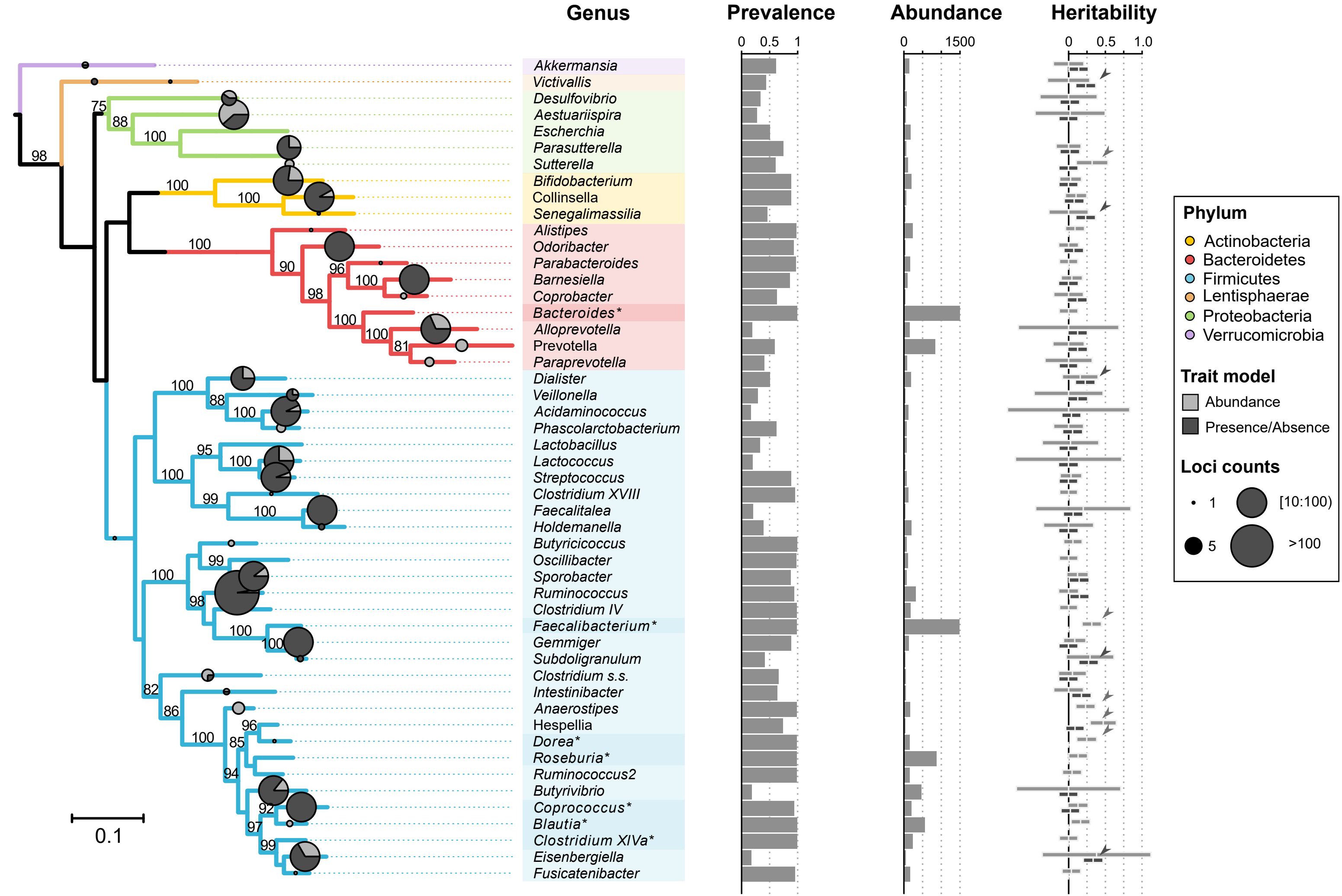
| Taxon | Model | rsID | SNP ID | Chromosome | Position (bp) | EAF | β | se | p-value | N | Q | hetP | I$^2$ | Closest gene |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **G_Ruminococcus** | P/A | rs150018970 | 9:134648925_G_A | 9 | 134,648,925 | 0.010 | 0.111 | 0.294 | 6.68x10$^{-14}$ | 3890 | 10.763 | 0.005 | 81.417 | *RAPGEF1* |
| **G_Coprococcus** | P/A | rs561177583 | 1:96741622_G_A | 1 | 96,741,622 | 0.012 | 0.161 | 0.283 | 1.10x10$^{-10}$ | 3890 | 22.036 | 0.000 | 90.924 | NA |
| **G_Butyricicoccus** | AB | rs55808472 | 16:24931691_G_A | 16 | 24,931,691 | 0.073 | 0.257 | 0.041 | 5.54 x10$^{-10}$ | 3890 | 1.248 | 0.536 | 0.000 | *ARHGAP17* |
| **F_Sutterellaceae** | P/A | rs4494297 | 11:44145588_G_T | 11 | 44,145,588 | 0.011 | 0.144 | 0.314 | 6.80 x10$^{-10}$ | 3890 | 3.478 | 0.176 | 42.497 | *EXT2* |
| **G_Dialister** | P/A | rs7118902 | 11:121440231_G_A | 11 | 121,440,231 | 0.306 | 0.734 | 0.053 | 4.14 x10$^{-09}$ | 3890 | 0.684 | 0.710 | 0.000 | *SORL1* |
| **G_u_F_Porphyromonadaceae** | AB | rs35980751 | 13:96011248_G_T | 13 | 96,011,248 | 0.259 | 0.196 | 0.034 | 5.07 x10$^{-09}$ | 2094 | 1.090 | 0.580 | 0.000 | *ABCC4* |
| **G_Parabacteroides** | AB | rs13207588 | 6:41519430_G_A | 6 | 41,519,430 | 0.229 | -0.180 | 0.031 | 6.89 x10$^{-09}$ | 3890 | 5.120 | 0.077 | 60.938 | *FOXP4* |
| **G_u_F_Erysipelotrichaceae** | P/A | rs6733298 | 2:56450856_A_G | 2 | 56,450,856 | 0.891 | 1.640 | 0.085 | 6.91 x10$^{-09}$ | 3890 | 1.739 | 0.419 | 0.000 | *CCDC85A* |
| **C_Gammaproteobacteria** | AB | rs116865000 | 15:95639861_G_A | 15 | 95,639,861 | 0.025 | 0.555 | 0.096 | 8.41 x10$^{-09}$ | 3213 | 1.847 | 0.397 | 0.000 | NA |
| **G_u_P_Firmicutes** | AB | rs11788336 | 9:111688387_T_C | 9 | 111,688,387 | 0.273 | -0.143 | 0.025 | 1.66 x10$^{-08}$ | 3485 | 5.388 | 0.068 | 62.882 | *IKBKAP* |
| **G_u_P_Firmicutes** | P/A | rs34656657 | 6:16613223_G_A | 6 | 16,613,223 | 0.022 | 0.294 | 0.218 | 1.82 x10$^{-08}$ | 3890 | 2.845 | 0.241 | 29.710 | *ATXN1* |
| **G_u_O_Bacteroidales** | P/A | rs116135844 | 4:168179343_G_A | 4 | 168,179,343 | 0.043 | 2.109 | 0.134 | 2.32 x10$^{-08}$ | 3890 | 1.969 | 0.374 | 0.000 | *SPOCK3* |
| **G_Veillonella** | P/A | rs117338748 | 15:58714239_G_A | 15 | 58,714,239 | 0.019 | 2.887 | 0.190 | 2.42 x10$^{-08}$ | 3890 | 2.109 | 0.348 | 5.178 | *LIPC* |

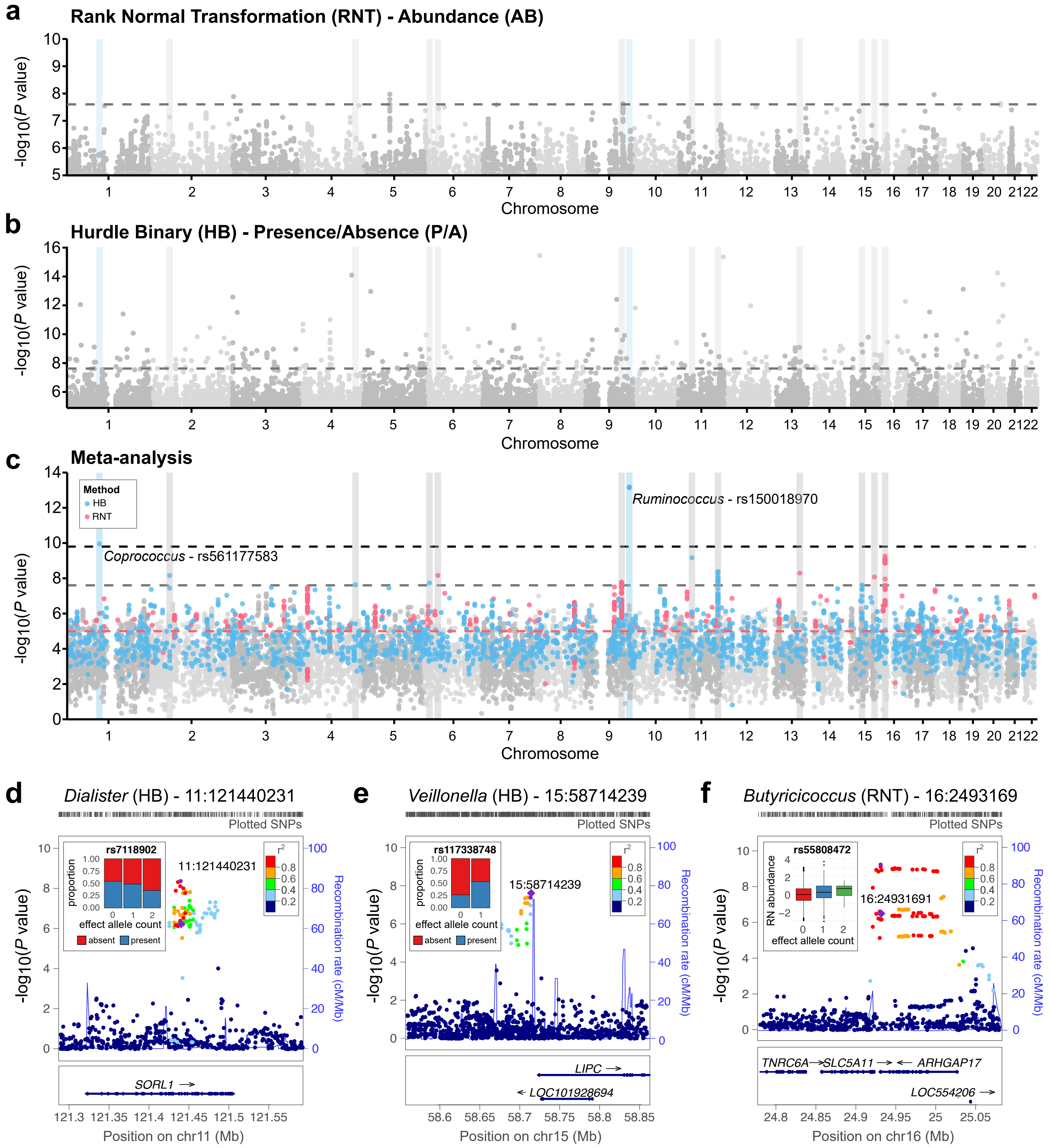
Genetic variants, representing LD-tagged loci, associated with 16S gut microbiome phenotypes at a meta-p-value smaller than 2.5x10$^{-08}$. Presented are the microbial taxa, the trait model type - abundance (AB) or presence/absence (P/A) - the reference SNP identifier (rsID), the SNP ID composed of chromosome, base pair (build hg19), alternative allele and effect allele, chromosome, position (bp), the effect allele frequency (EAF), the meta estimated effect size (β; presented as odds ratios for P/A, and in SD units of change for AB outcomes), standard error (se, in log(OR) scale for P/A outcomes), two-sided inverse variance fixed effect meta p-value, meta-sample size (N), Cochran's Q heterogeneity statistic (Q), the heterogeneity p-value (hetP), the proportion of variation among studies due to heterogeneity (I2) and the physically closest gene (+/- 250kb). P, C, O, F, G preceding taxa names indicate the classification levels phylum, class, order, family and genus, while u indicates unclassified. Genetic variants are sorted by significance (p-value), with the top two surpassing the study-wide meta-analysis p-value shaded in blue. Taxon names include abbreviated taxonomic levels, where “G” represents genus, “F” family, “O” order, “P” phylum, and “u” unclassified.

**Table 2: Bi-directional Mendelian Randomization.**

| | Exposure | model | Outcome | nSNP | MR method | beta | se | p-value |
|---|---|---|---|---|---|---|---|---|
| **MT -> disease** | G_Dialister | P/A | Alzheimer's disease* | 1 | Wald ratio | 0.810 | 0.055 | 1.35x10$^{-04}$ |
| | G_Butyricicoccus | AB | Inflammatory bowel disease* | 1 | Wald ratio | 0.748 | 0.126 | 2.17x10$^{-02}$ |
| | G_u_P_Firmicutes | AB | Waist circumference | 1 | Wald ratio | 0.070 | 0.033 | 3.34x10$^{-02}$ |
| | G_Butyricicoccus | AB | Alzheimer's disease* | 1 | Wald ratio | 0.790 | 0.116 | 4.17x10$^{-02}$ |
| | G_u_F_Erysipelotrichaceae | P/A | Type 2 diabetes* | 1 | Wald ratio | 0.906 | 0.049 | 4.48x10$^{-02}$ |
| | G_Dialister | P/A | Major depressive disorder* | 1 | Wald ratio | 1.160 | 0.074 | 4.63x10$^{-02}$ |
| | G_Bifidobacterium | AB | Waist circumference | 1 | Wald ratio | -0.149 | 0.029 | 2.82x10$^{-07}$ |
| | G_Bifidobacterium | AB | Body mass index | 1 | Wald ratio | -0.123 | 0.027 | 3.37x10$^{-06}$ |
| | G_Bifidobacterium | AB | Waist-to-hip ratio | 1 | Wald ratio | -0.060 | 0.029 | 3.74x10$^{-02}$ |
| **Disease -> MT** | Alzheimer's disease | P/A | G_Dialister* | 5 | IVW | 1.809 | 0.239 | 1.33x10$^{-02}$ |
| | Parkinson's disease | P/A | G_u_P_Firmicutes* | 1 | Wald ratio | 0.386 | 0.391 | 1.47x10$^{-02}$ |
| | Type 2 diabetes | P/A | G_u_P_Firmicutes* | 10 | IVW | 0.636 | 0.195 | 2.02x10$^{-02}$ |
| | Crohn's disease | P/A | G_u_P_Firmicutes* | 21 | IVW | 0.813 | 0.092 | 2.46x10$^{-02}$ |
| | Parkinson's disease | AB | G_Bifidobacterium | 1 | Wald ratio | 0.225 | 0.107 | 3.46x10$^{-02}$ |

Results from bi-directional Mendelian randomization analysis querying causal relationships between microbial traits (MTs) on each trait and each trait on MTs. The exposure identifies the independent variable in the analysis, while the outcome is the dependent variable. Presented are the trait model type - abundance (AB) or presence/absence (P/A) and the number of SNPs used as "instruments" for the exposure (nSNP). Primary MR results were limited to two MR models, namely the inverse variance weighted (IVW) and Wald ratio methods. All other models were considered sensitivity analysis and can be found in Table S13. The beta, se, and p-value provide the effect estimate (risk ratios for binary outcomes (*) and SD units of change for continuous outcomes), standard errors (in log(OR) scale for binary outcomes), and uncorrected two-sided model p-values for that analysis, respectively. Analyses were restricted to those MTs found in Table 1, with rank normalized *Bifidobacterium* (shaded in grey) as an addition. Sample sizes for each previously published GWAS disease trait can be found in Table S14.


Genus
Prevalence
Abundance
Heritability
0
0.5
1
0
1500
0
0.5
1.0
Akkermansia
Victivallis
Desulfovibrio
Aestuariispira
Escherchia
Parasutterella
Sutterella
Bifidobacterium
Collinsella
Senegalimassilia
Alistipes
Odoribacter
Parabacteroides
Barnesiella
Coprobacter
Bacteroides*
Alloprevotella
Prevotella
Paraprevotella
Dialister
Veillonella
Acidaminococcus
Phascolarctobacterium
Lactobacillus
Lactococcus
Streptococcus
Clostridium XVIII
Faecalitalea
Holdemanella
Butyricicoccus
Oscillibacter
Sporobacter
Ruminococcus
Clostridium IV
Faecalibacterium*
Gemmiger
Subdoligranulum
Clostridium s.s.
Intestinibacter
Anaerostipes
Hespellia
Dorea*
Roseburia*
Ruminococcus2
Butyrivibrio
Coprococcus*
Blautia*
Clostridium XIVa*
Eisenbergiella
Fusicatenibacter
98
75
88
100
100
100
100
90
96
98
100
100
100
81
100
88
100
95
100
100
99
100
100
99
98
100
100
82
86
100
96
85
94
92
97
99
0.1
Phylum
● Actinobacteria
● Bacteroidetes
● Firmicutes
● Lentisphaerae
● Proteobacteria
● Verrucomicrobia
Trait model
□ Abundance
■ Presence/Absence
Loci counts
1
5
[10:100)
>100


a
Rank Normal Transformation (RNT) - Abundance (AB)
-log10(P value)
5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome
b
Hurdle Binary (HB) - Presence/Absence (P/A)
-log10(P value)
6 8 10 12 14 16
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome
c
Meta-analysis
Method
HB
RNT
Ruminococcus - rs150018970
Coprococcus - rs561177583
-log10(P value)
0 2 4 6 8 10 12 14
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome
d
Dialister (HB) - 11:121440231
Plotted SNPs
rs7118902
proportion
0.00 0.25 0.50 0.75 1.00
0 1 2
effect allele count
absent present
11:121440231
r2
0.8 0.6 0.4 0.2
-log10(P value)
0 2 4 6 8 10
Recombination rate (cM/Mb)
0 20 40 60 80 100
SORL1 →
121.3 121.35 121.4 121.45 121.5 121.55
Position on chr11 (Mb)
e
Veillonella (HB) - 15:58714239
Plotted SNPs
rs117338748
proportion
0.00 0.25 0.50 0.75 1.00
0 1
effect allele count
absent present
15:58714239
r2
0.8 0.6 0.4 0.2
-log10(P value)
0 2 4 6 8 10
Recombination rate (cM/Mb)
0 20 40 60 80 100
LIPC →
← LOC101928694
58.6 58.65 58.7 58.75 58.8 58.85
Position on chr15 (Mb)
f
Butyricicoccus (RNT) - 16:2493169
Plotted SNPs
rs55808472
RN abundance
−2 0 2 4
0 1 2
effect allele count
16:24931691
r2
0.8 0.6 0.4 0.2
-log10(P value)
0 2 4 6 8 10
Recombination rate (cM/Mb)
0 20 40 60 80 100
TNRC6A→ SLC5A11 → ← ARHGAP17
LOC554206 →
24.8 24.85 24.9 24.95 25 25.05
Position on chr16 (Mb)

## Figure Legends

**Figure 1: MT-associated loci and heritability.** Maximum likelihood phylogenetic tree of gut microbial genera used in association analysis. Circle sizes along the branches and nodes of the tree indicate the number of loci observed to be associated to microbial traits at that genus, family, order, class, or phylum levels. Core genera, as defined by Falony *et al*[4], are starred. Bootstrap values over 60 are shown on branches. The length of the scale bar represents nucleotide substitutions per site. For illustrative purposes the number of associated loci, greater than 10 were set to the size 10, those larger than 100 were set to size 15. Bar plots on the right describe the prevalence (or proportional sample size n, where a prevalence of 1 = 2257) and the standardized mean abundance of the bacterial genera in the FGFP population, and the estimated heritability (bars represent standard errors, point estimates are in white; arrows identify those observations determined to be different from zero given a two-sided GCTA-GREML likelihood ratio test p-value less than 0.05) of AB and P/A traits. All summary statistics, sample sizes, and p-values illustrated here can be found in Table S3.

**Figure 2: Genomic variants associated with microbial traits**. Manhattan plots illustrating negative log10 two-sided (a) F-test or (b) chi-squared p-values derived from the FGFP cohort Score test association analysis for (**a**) rank normal transformed (RNT) abundances and (**b**) presence/absence (Hurdle Binary) states. The genome-wide threshold is indicated by the horizontal dashed line. (**c**) Manhattan plot for our targeted meta-analysis derived from Expectation-Maximization (em) parameter estimates. Sites that did not exhibit consistent effect estimates in meta-analysis are shaded in grey, while those sites that had a smaller two-sided inverse-variance fixed-effect meta-p-value than the FGFP (em) p-value are colored in blue and red for P/A (HB) and RNT (AB) traits, respectively. Loci that achieved study-wide significance are highlighted by a blue tower, while those exceeding the genome-wide threshold are in shaded in grey. Dashed lines indicate the study-level (black), conventional genome-wide (grey), and target-meta analysis (red) thresholds of $1.57\text{x}10^{-10}$, $2.5\text{x}10^{-8}$, and $1\text{x}10^{-5}$, respectively. LocusZoom plots of association results using the FGFP derived two-sided chi-squared (HB) or F-test (RNT) p-values for three top SNPs that achieved the conventional genome-wide level p-value in the meta-analysis (**d-f**). The LD estimates are color coded (D′ ≥0.3 to >0.4 in purple) and recombination rate is indicated by the blue lines

and the z-axis. The proximal genes to the top SNPs are indicated in the bottom panel. Genotype to microbial trait structure for the tagged variant at each locus is illustrated in the insert, with (**f**) illustrating a boxplot (identifying the mean, first and third quartiles and the 95% confidence intervals) of taxon abundance by genotypic state (homozygous non-effect allele (0), heterozygous (1), or homozygous effect allele (2)), while (**d**) and (**e**) inserts illustrate bar plots of the proportion of individuals in each genotypic state where the taxon is absent (red) or present (blue) in each of the observed genotypic states. Genotypic state sample sizes are (d) n0=1113, n1=927, n2=219, (e) n0=2177, n1=82, (f) n0=1948, n1=292, n2=19. Data used to generate these plots can be found on data.bris https://doi.org/10.5523/bris.22bqn399f9i432q56gt3wfhzlc) and Table S2.

**High resolution figures are included as separate files.

## Methods

### Study recruitment and sample collection

Individuals from the Flanders region of Belgium were recruited into the Flemish Gut Flora Project (FGFP) through public announcements in print and social media through the FGFP website (www.vib.be/darmflora), from January 2013 onwards. Volunteers provided informed consent by mail and FGFP procedures were approved by the medical ethics committee of the University of Brussels/Brussels University Hospital (approval 143201215505, 5/12/2012). A declaration concerning the FGFP privacy policy was submitted to the Belgian Commission for the Protection of Privacy. Additional information on the age, sex, height, weight BMI, waist hip ratio, and low-density lipoprotein distributions for FGFP cohort samples are provided in Supplementary Information (Table S6, Supplementary Fig. 3).

FGFP samples and data was collected as in Falony *et al*[4]. In short, stool samples were collected between June 2013 and April 2016 by mail. Sampling kits were sent to volunteers' home addresses and upon collection samples were stored at -18°C locally, cooled during delivery and again stored at -18°C upon arrival at a collection point until long-term storage was possible at -80°C at the research facility. A medical questionnaire was completed by each volunteers' general practitioner (GP). The GPs also took new measurements of volunteers' height, weight, hip and waist circumference, in addition to blood pressure and an eight-hour fasted blood sample.

Fecal DNA was extracted from the frozen fecal samples using the PowerMicrobiome RNA Isolation Kit (MOBIO Laboratories Inc.) following manufacturer's instructions, with the addition of a heating step (10min at 90°C) after vortexing/bead beating to increase DNA yield, and with the exclusion of DNA removal steps (steps 12 to 16 in the protocol). Further information on recruitment, sampling and DNA extraction can be found in Falony *et al*[4].

### Sequencing and microbiome data processing

For 2482 FGFP individuals, the V4 region of the 16S rRNA gene was amplified using the 515F/806R primer pair (GTGYCAGCMGCCGCGGTAA and GGACTACNVGGGTWTCTAAT, respectively), modified to contain a barcode sequence between each primer and the Illumina adaptor sequences to produce dual-barcoded libraries[32]. Size selection was performed using Agencourt AMPure to remove fragments below 200 bases. Sequencing was carried out on the Illumina HiSeq platform at the VIB Nucleomics core laboratory (Leuven, Belgium) with 500 cycles (sequencing kit HiSeq-Rapid SBS kit, version 2), producing 2x 250bp paired-end sequencing reads. After de-multiplexing with sdm as part of the LotuS pipeline[33] without allowing for mismatches, fastq sequences were further analyzed per sample using DADA2 pipeline (v. 1.6)[15]. In brief, after inspecting quality, sequences were trimmed to remove the primers and the first 10 bases after the primer, keeping only 200 bases and 130 for the R1 and R2 files, respectively. After merging paired sequences and removing chimeras, compositional matrices for each taxonomical level were carried out using the Ribosomal Database Project (RDP) training set ‘rdp_train_set_16’. Each sample was randomly down-sampled, also known as a rarefaction step, reducing the microbiome to a size of 10,000 reads. Classifications with low confidence at the genus level (<0.8) were organized in an arbitrary taxon of “unclassified_group”.

## Microbiome trait preparation

The DADA2 pipeline yielded count data for 499 taxa across five levels of the microbiota phylogeny from phylum to genus (Extended Data Fig. 4a and 4b). Quality control on an individual level was performed by constructing an initial multi-dimensional scaling plot using Bray Curtis distances derived from the vegdist() function of the vegan package, and the argument method=”bray”, followed by a Kruskal’s nMDS using the function isoMDS() from the MASS package using rarefaction data from genera level counts. Two individual samples were identified as outliers in their genus-level microbiome profiles in this analysis (Supplementary Fig. 4), with the outlier cut-off set at greater than or less than five standard deviations (SDs) from the population mean of both nMDS axes. These two individuals were removed from all subsequent analysis including all association analyses.

α- and β-diversity statistics were estimated using the rarefaction data for all 288 genera level taxa counts. The α-diversity statistics used are (1) the number of genera observed, defined as the number of non-zero counts observed across all genus-level taxa, (2) Shannon diversity as calculated with the function diversity() in the vegan package, using the arguments index = "shannon", MARGIN = 1, base = exp(1), and finally (3) Chao diversity estimated using the estimateR() function in the vegan package. β-diversity was estimated using a two-axis non-metric multidimensional scaling (nMDS) analysis using the function metaMDS() from the vegan package and the arguments, distance = "bray", k=2, try=20, trymax=50, and trace=FALSE. The stress of the nMDS was 0.209 (Extended Data Fig. 4c). Enterotyping (or community typing) based on Dirichlet multinomial mixtures (DMM) as previously described[34] was performed using the DirichletMultinomial package in R. This analysis was performed on the FGFP-genus-level relative abundance matrix rarefied to 10,000 reads (Extended Data Fig. 4c).

For association analysis, we retained any taxa that met two criteria - (1) make up ≥5% of the reads for at least 1 individual and (2) have ≥15% of individuals with non-zero data (Extended Data Fig. 5a; Tables S15 and S16). In total, 139 taxa across all phylogenetic levels met these criteria. However, given that lower phylogenetic level count data can be exactly, or very close to, the same as count data at higher phylogenetic levels, we wanted to eliminate any statistical redundancies in the mGWAS. As such, we estimated Pearson correlation coefficients among all taxa, and any taxa pair with a correlation coefficient greater than 0.985 had the higher taxon level removed from association analyses. Seventy-three of the taxa exhibited such a correlation with at least one other taxa (Extended Data Figs. 5b and 5c). After removing higher-level taxa, 92 taxa remained for association analyses. The FGFP data set was used to identify these 92 taxa.

Given the ecological, observational count nature of 16S data, many individuals contain zero counts for some taxa. As such, a common feature of this data is zero-inflation, which can prove problematic for data transformation and linear modeling (Supplementary Fig. 1 and 2). To account for this possibility, we identified those taxa that should go through a two-step hurdle analysis that includes a presence/absence (P/A) association analysis and a zero-truncated abundance (AB) mGWAS. To do so, the proportion of individuals that were zero (absent) for each taxon was estimated, and those with greater than 5% zeros were pushed through the hurdle analysis. First, all non-zero counts were turned into 1's for the binary P/A mGWAS and second, all zero counts were turned into NAs for the zero-truncated AB mGWAS. Sixty-two of the 92 retained taxa fit these criteria and were processed in this manner. The other 30 MTs were treated as simple abundance phenotypes and also denoted as AB. We note that the outcome of this procedure of course is a variety of different sample sizes in both the zero-truncated abundance phenotypes and between the number of absent individuals in P/A traits among taxa. This is an outcome that will introduce variability in power among the mGWAS performed here. Again, we note that only the FGFP data set was used to identify model type for each taxon.

In preparation for the association analysis, each abundance phenotype was rank normal transformed using the rntransform() function from the GenABEL[35] package and fit to a multivariate linear model, using the function lm() from the stats package, with the following covariates: the extraction type (drill or cut), the extraction year, the aliquot year (for 16S rRNA sequencing), the person performing the aliquot, the library preparation plate, genotype derived principle components 1-10, genotype predicted sex, and age. Residuals from this model were extracted using the function residuals() from the stats package and used in univariate linear modeling in the association analysis with genotypes, details below. Shapiro-Wilk W statistics for the raw and residualized data distributions can be found in Tables S17-S18. Analysis and preparation of the microbial trait data was carried out in R version 3.4.1 "Single Candle"[36].

### Observational analysis

To identify biological phenotypes that may be influenced by gut microbiome variation in the FGFP data set (generalized) linear models, as described above, where fit with age, sex, and the top ten principle components as covariates along with each of the microbial traits (MT) analysed in the GWAS (results not included, but available on request). Human phenotypes include blood lipids, glycemic traits, anthropomorphic traits, diet and Bristol stool score. To identify laboratory batch variables that may have influenced 16S microbiome variation, we set all available variables as dependent variables in univariate analysis, with each MT set as the response variable to identify those that should be included as covariates in the GWAS. Batch variables that exhibited independent effects on at least one MT are the extraction type (drill or cut), the extraction year, the aliquot year (for 16S rRNA sequencing), the person performing the aliquot, and the library preparation plate. Further information and results from these analyses can be found in Supplementary Information (Supplementary Fig. 5).

### Genotyping

A total of 2646 FGFP individuals were processed on two different arrays - the Human Core Exome v1.0 (N = 576 samples) and the Human Core Exome v1.1 (N = 2112 samples), which included repeat measurements. Allele calling was performed using GenomeStudio v2.0.4 following manufacturers default recommendations. While running GenomeStudio Log R Ratio (LRR) and B Allele Frequency (BAF) statistics were also extracted for copy number variant (CNV) calling with PennCNV[37]. Unmapped and duplicate positions were removed, and the two batches were merged into a single data set resulting in 545,535 overlapping markers.

Variant quality control (QC) steps included the removal of unmapped variants (n = 777), duplicated sites (n = 6899), variants with >5% missingness (n = 3445), those with Hardy-Weinberg equilibrium deviations p-values < $1x10^{-05}$ - after accounting for relatedness (n = 404), those with ambiguous alleles (n = 12,095), and those that are tri-allelic or allele flip errors (n = 1026). A total of 509,886 variants remained after QC. Sample QC included a cross check between genetically predicted sex and available sex information (117 mismatches), removal of array failed samples (n = 5), samples with >5% variant missingness (n = 53), samples with heterozygosity ± three SDs from the population mean estimate (n = 33), the removal of cryptically related (relatedness > 0.025) samples (n = 262) using the function in rel-cutoff in plink 1.9[38], and those with genotypic discordance among replicates (n = 8). Data was then merged with that from phase three of the 1000 Genomes Project to identify those individuals exhibiting ancestry components from populations outside of Western Europe (n = 34), using principal component analysis (PCA, Supplementary Fig. 6). After QC 2293 individuals remained (Supplementary Fig. 7), 2257 of which were retained given the availability of microbiome data.

FGFP genotype data was phased using SHAPEIT3[39] and imputed with IMPUTE2[40] using UK10K and all 1000 Genome Project phase 3 samples as the reference panel[41]. Following imputation the 39,168,681 SNPs were filtered to retain only those sites with a minor allele frequency greater than or equal to 1% and with an imputation quality score (INFO) greater than or equal to 0.3, as estimated with qctool v2.0 -snp-stats (www.well.ox.ac.uk/~gav/qctool). In total 7,711,310 SNPs were retained for the FGFP mGWAS. A flowchart of this genotyping quality control steps is available in Extended Data Fig. 6a.

To acquire insertion deletion variants, the data was also phased and imputed to Genome of Netherlands reference panel using Impute2 v2.3.0[42]. All indels were isolated from this imputed data set and run in addition to the imputation data set from above.

Copy number variants (CNVs) were called with PennCNV v1.0.4[37] using the perl script detect_cnv.pl. Cleaning was performed with the perl script clean_cnv.pl, and filtered with the script filter_cnv.pl using the flags --numsnp 5 --length 250 -qclrrsd 0.35 -qcnumcnv 716. Unique CNVs were defined by unique base pair start and stop locations. In total 35,020 unique CNVs were identified across the FGFP sample; 949 CNVs were shared across 1% or more individuals. Global CNV burden was estimated for each individual as the number of CNVs that do not equal the copy number count of 2. Insertions (>2) and deletions (<2) were treated the same. Regional CNV burden was calculated in sliding windows of 200 kilo-bases and estimated following the same rules as for global burden estimation.

### Heritability

Chip-based heritability was estimated for each microbiome presence/absence (62) abundance (92) and α-diversity metric (3) phenotype used in the association analysis, using the GCTA-GREML restricted maximum likelihood (REML) method, and a single genetic relationship matrix (GRM) as implemented in GCTA version 1.91.1beta[43]. The GREML power calculator was used to estimate the power to detect genetic covariation in the FGFP data set (Extended Data Fig. 7)[44]. For the abundance phenotypes, the residualized data used in the association analyses were used in the estimation of heritability. For binary phenotypes, the same covariates, mentioned above, were fit to the trait by GCTA. To produce the genetic relationship matrix (GRM) for running GCTA, we first identified all genotypes with an info (imputation quality) score greater than or equal to 0.9, a minor allele frequency greater than 0.05, and not deviating from Hardy-Weinberg equilibrium. Genotype probability score data was converted to hard call plink format data using qctools. SNP variation was linkage disequilibrium pruned using plink2 and the flag --indep-pairwise 50 5 0.45. Finally, the GRM was constructed using GCTA and the flags --grm-cutoff 0.025 --make-grm. In addition, for a more direct comparison with previously published studies[7], we also performed box-cox transformations of the abundance phenotypes and regressed out our covariates.

### Primary FGFP association analysis

Following genotype and microbiome QC, 2257 individuals remained, and 2223 remained after accounting for data missingness among covariates. All microbiome α-diversity, abundance and presence/absence associations analyses were performed using snptest.2.5.0[45]. All abundance traits were regressed on covariates (the aliquoting procedure, the extraction year, the aliquot year (for 16S rRNA sequencing), the person performing the aliquot, the library preparation plate, genotype derived principle components 1-10, genotype predicted sex, and age) and residuals were regressed on genotype probability data in a univariate fashion, assuming an additive genetic model and using the missing data likelihood score test in snptest (snptest flags: -frequentist 1 -method score and -use_raw_phenotypes). Presence/absence mGWAS were performed using the same covariates as those described above for abundance traits in a multivariate analysis again using the same snptest settings. These primary analyses were performed as a first pass signal detection step in order to determine signals to take forward for meta-analysis and to confirm the ability of score analyses to effectively rank the expectation–maximization (em) method (Supplementary Fig. 8). Association analyses for enterotype were run using a multinomial logistic regression for categorical traits as employed by snptest.2.5.4-beta3 and the flags -frequentist add -method newml and setting the -baseline_phenotype to "Bacteroides1". Finally, associations for β-diversity (a two axis MDS) were run using a bespoke R script and the function manova() from the stats package, in a multivariate analysis using the same covariates stated above and genotype dosages as derived by qctool v2.0. A flow chart of the mGWAS is provided in Extended Data Fig. 6b.

**Data preparation in German cohorts**

The FoCus[11] and PopGen[16] cohorts were genotyped using the Illumina Omni Express + Exome array and the Affymetrix Genome-Wide Human SNP Array 6.0, respectively. Genotyping QC and imputation in these two cohorts were performed following protocols defined here: https://github.com/alexa-kur/miQTL_cookbook#chapter-2-genotype-imputation. SNPs where filtered as a MAF of 0.01 and an INFO score of 0.3, as was done in the FGFP cohort. Microbiome census data for the Kiel based cohorts, targeting the 16S V1-V2 regions, was generated as described previously[11]. Data processing was performed using DADA2[15] modified for V1-V2 (https://github.com/mruehlemann/rep-cookbook/blob/master/scripts/Seq_dada2_V12_Kiel.R) following the same standardized workflow described in the above Microbiome trait preparation section. α- and β-diversity metrics, enterotypes, and abundance measures for association analysis were calculated as previously indicated. Three genera, *Escherichia Shigella*, *Hespellia*, and *Methanobrevibacter* were not present in the FoCus or PopGen cohorts (Supplementary Fig. 9). As such, in all three instances, their P/A and zero-truncated AB MTs were not available for association and inclusion in the meta-analysis. Association analyses were carried out as described below.

**Meta-analysis**

For the purposes of the meta-analysis, all outcomes were defined as described above in the "Primary FGFP association analysis" section, however, the expectation–maximization (em) method was used, rather than the score method, to account for genotype uncertainty and given the performance of "score" at low allele frequency and phenotype/trait group size (Supplementary Information).

The meta analyses were performed using the inverse-variance fixed effects method (method 1) as implemented in the software package META[46]. The imputation quality threshold for each SNP was set at 0.3. To identify loci or unique genomic regions defined by shared linkage disequilibrium, we clumped all meta-supported markers using plink, the flag –

clump and the p-values derived from the em meta-analysis. A locus or index tags was only identified if their meta-analysis p-value was < 0.0001.

Beta estimations (genotypic effects) for P/A traits are defined as an increase in the log odds ratio for each additional effect allele. For AB traits, beta is defined as a change in SD units for each effect allele carried. The study-wide p-value threshold was defined as a Bonferroni correction assuming 2 million independent genetic association tests across 159 mGWAS ($0.05 / (2x10^{06}$ x $159) = 1.57x10^{-10}$).

**Phylogenetic analysis**

Representative 16S rRNA gene sequences of all the genera identified were retrieved from the RDP database. Multiple sequence alignments were performed for all taxa and for genera included in the GWAS analyses using MUSCLE v.3.8[47]. The alignments were used to build maximum likelihood trees using FastTree v2.1.0[48] with default parameters. iTOL[49] was used for visualizing the trees with corresponding metadata, including the number of loci, prevalence of the MT, abundance of the MT and heritability of the AB and P/A trait(s) (Fig. 1 and Table S3).

**Functional annotation and enrichment**

Annotation of variants of interest was carried out with the biomaRt R package[50] with additional linkage disequilibrium based annotation and enrichment analysis with DEPICT[51]. When using biomaRt, we referenced the (feb2014.archive.ensembl.org) Ensembl 75 archive for GRCh37/hg19 coordinates and identified all genes and the closest gene within 250 kilo-bases up- and down-stream of each polymorphism. Given that DEPICT utilizes pre-computed LD structure from genotypes derived from 1000 Genomes Project Phase 1 CEU, GBR and TSI and HapMap Project release 2 and 3 CEU data, a substantial proportion of our SNPs of interest were not represented. As such, we identified tag SNPs for our SNPs that are also present in the DEPICT data set, when possible. To do so, we extracted dosage data for all variants +/- 200kb of our variant, using qctools, and then computing $r^2$ using a bespoke R script. We kept all variants with an $r^2 > 0.2$ with our SNP of interest and then queried if any of those tag SNPs existed in the DEPICT data. If more than one was present, we kept the one with the highest $r^2$ value. Subsequently, the list of reference SNP identifiers (rs-ids) composed of our SNPs of interest, when they existed in the DEPICT data, or alternatively tag SNPs, when they were present, were run through the DEPICT framework. Each MTs was run individually using FGFP variants associated at a threshold of $1x10^{-5}$. Results from these analyses, gene enrichment, tissue enrichment, and gene priority are available in Table S19, S20, S21, respectively.

The Gene-Tissue Expression portal (gtexportal.org) was used to determine if associated variants, specifically discussed in the text, were also expression quantitative trait loci (gtexportal.org).

When evaluating enrichment for all meta-supported variants as a unit, we extracted the closest gene (within +/- 250kb) for each variant and used GENE2FUNC (http://fuma.ctglab.nl/), and its integrated hypergeometric test, to identify tissue and pathways functional enriched, as reported in Table S12. The background set of genes, included in hypergeometric test, was 19,283 protein coding genes. The default parameter of GEN2FUNC in the FUMA platform[29].

All microbial traits analyzed were also analyzed under the GARFIELD[52] framework for identification of regulatory and tissue enrichments. This analysis uses the complete GWAS data set of a trait, so pooling data from multiple traits, as done above, was not carried out. Further, the FGFP results only were used in these analyses. The results of these analyses are available in Table S22.

Finally, variants with associations that met our study-wide and genome-wide confidence thresholds as well as those that may be deemed as replicated from other studies were passed through PhenoScanner V2, an online platform to screen for genotype-to-phenotype associations, expression quantitative loci, and methylation quantitative loci from previously published genome-wide -omics association analysis[28]. All of these results are provided in Table S11.

## Applied analyses

We undertook two-sample, bi-directional Mendelian randomization[30] analyses to estimate potentially causal relationships between gut MTs and 11 metabolic health, inflammatory and neurological traits. These were selected *a priori* as they have all been repeatedly been associated with variation in the gut microbiome and have been the focus of credible and accessible GWAS studies. They include waist circumference, waist-hip ratio, BMI, and type 2 diabetes; Crohn's disease, inflammatory bowel disease, ulcerative colitis and rheumatoid arthritis; and Alzheimer's disease, Parkinson's disease and major depressive disorder.

MR analyses interrogating the role of the gut microbiome on each of these outcomes were restricted to the gut MTs that had the greatest evidence of a host-genetic contribution -

where independent meta-derived genetic variants reaching a genome-wide threshold of $p<2.5\times10^{-08}$ were used as "instruments". In order to assess causality in relationships from microbiome to outcome, summary statistics for these genetic variants were obtained from publicly available genome-wide summary-level data for the 11 metabolic health, inflammatory and neurological traits. For analyses assessing causality in relationships from each trait to microbiome, independent genetic variants reaching a genome-wide threshold of $p<5\times10^{-08}$ in each respective GWAS were used as "instruments" for the relevant trait. Summary statistics were obtained from the current mGWAS meta-analysis.

Once all summary-level data was obtained, causal effect estimates were derived using the inverse variance weighted (IVW)[53] method (or the Wald ratio, if only 1 genetic variant was available) alongside sensitivity analyses including the weighted median[53], weighted mode[54] and MR-Egger[55] tests (if ≥3 genetic variants were available). All exploratory MR analyses were conducted using the TwoSampleMR package (https://github.com/MRCIEU/TwoSampleMR) in R version 3.4.1 with RStudio, created and provided by MR-Base (www.mrbase.org/)[56], a large-scale database of GWAS summary-level data and automated pipeline for two-sample MR analyses. Summary-level GWAS results for the 11 selected metabolic health, inflammatory and neurological traits were obtained from the following publications (indicated with pubmed ID): waist circumference and waist-to-hip ratio (25673412); BMI (25673413); type 2 diabetes (22885922); Crohn's disease, inflammatory bowel disease and ulcerative colitis (26192919); rheumatoid arthritis (24390342); Alzheimer's disease (24162737), Parkinson's disease (19915575) and major depressive disorder (22472876).

In analyses interrogating the impact of each continuous (AB) MT on each outcome, effect estimates represent the SD change for continuous outcomes (waist circumference, waist-to-hip ratio and BMI) or risk ratio for binary outcomes (type 2 diabetes, Crohn's disease, inflammatory bowel disease, ulcerative colitis, Alzheimer's disease, major depressive disorder, Parkinson's disease and rheumatoid arthritis) for a SD unit of AB MT phenotype. For analyses interrogating the impact of each binary (P/A) MT on each outcome, effect estimates represent the SD change for continuous outcomes or risk ratio for binary outcomes for a doubling of the genetic liability to presence (vs. absence) of each P/A MT phenotype. Results reaching a p-value threshold of ($p<0.05$) are presented (Table 2).

## Inter-study catalog

A catalog of previously published associations was compiled starting from the work of Rothschild et al[1], and includes Blekham et al[8], Davenport et al[9], Bonder et al[10], Goodrich et al[7], Turpin et al[12], and Wang et al[11]. Data are available in Tables S1, S7, and S8 (Supplementary Figs. 10 and 11).

## Code Availability

The full analysis pipeline is available at https://github.com/kul-fgfpgwas/rep-cookbook and includes four parts: (i) microbiome processing; (ii) genotype quality control and imputation; (iii) genome-wide association analysis and (iv) phylogenetic analysis.

## Data availability

All microbiome GWAS summary statistics are available online at the University of Bristol data repository, data.bris, at https://doi.org/10.5523/bris.22bqn399f9i432q56gt3wfhzlc. FGFP rarefaction count and transformed microbial trait data can be found in Supplementary Table 2. FGFP genotype data and host metadata from this study are not open but are available in accordance and in consent with ethical permission through managed access subject to a data use agreement with the Flemish Gut Flora Project and organised via Principal Investigator Jeroen Raes. The process of enquiry for data access is outlined as follows: Upon data request by email to jeroen.raes@kuleuven.be, the FGFP data access committee will evaluate access permission, which will be granted upon signature of a data use agreement between the governing legal entities. This is outlined on the study website http://www.raeslab.org/companion/fgfp-gwas/. Raw 16S data is available at the European Genome/Phenome Archive (https://ega-archive.org) under accession number EGAS00001004420. The datasets from Universitätsklinikums Schleswig-Holstein are available by application through their biobank (https://www.uksh.de/p2n/).

## References (both main text and methods)

1. Rothschild, D. *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).
2. McDonald, D. *et al.* American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems* **3**, (2018).
3. Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science (80-. ).* **352**, 565–569 (2016).
4. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science (80-. ).* **352**, 560–564 (2016).
5. Gilbert, J. A. *et al.* Current understanding of the human microbiome. *Nat. Med.* **24**, 392–400 (2018).
6. McKnite, A. M. *et al.* Murine Gut Microbiota Is Defined by Host Genetics and Modulates Variation of Metabolic Traits. *PLoS One* **7**, e39191 (2012).
7. Goodrich, J. K. *et al.* Genetic Determinants of the Gut Microbiome in UK Twins. *Cell Host Microbe* **19**, 731–743 (2016).
8. Blekhman, R. *et al.* Host genetic variation impacts microbiome composition across human body sites. *Genome Biol.* **16**, 1–12 (2015).
9. Davenport, E. R. *et al.* Genome-wide association studies of the human gut microbiota. *PLoS One* **10**, 1–22 (2015).
10. Bonder, M. J. *et al.* The effect of host genetics on the gut microbiome. *Nat. Genet.* **48**, 1407–1412 (2016).
11. 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).
12. Turpin, W. *et al.* Association of host genome with intestinal microbial composition in a large healthy cohort. *Nat. Genet.* **48**, 1413–1417 (2016).
13. Wang, J. *et al.* Meta-analysis of human genome-microbiome association studies: The MiBioGen consortium initiative. *Microbiome* **6**, 1–7 (2018).
14. 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).
15. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
16. Krawczak, M. *et al.* PopGen: Population-Based Recruitment of Patients and Controls for the Analysis of Complex Genotype-Phenotype Relationships. *Public Health Genomics* **9**, 55–61 (2006).
17. Ferreira-Halder, C. V., Faria, A. V. de S. & Andrade, S. S. Action and function of Faecalibacterium prausnitzii in health and disease. *Best Pract. Res. Clin. Gastroenterol.* **31**, 643–648 (2017).
18. Cohen, L. J. *et al.* Commensal bacteria make GPCR ligands that mimic human signalling molecules. *Nature* **549**, 48–53 (2017).
19. Consortium, Gte. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
20. Coady, M. J., Wallendorff, B., Gagnon, D. G. & Lapointe, J.-Y. Identification of a Novel Na +/myo -Inositol Cotransporter. *J. Biol. Chem.* **277**, 35219–35224 (2002).
21. Raffler, J. *et al.* Genome-Wide Association Study with Targeted and Non-targeted NMR Metabolomics Identifies 15 Novel Loci of Urinary Human Metabolic Individuality. *PLOS Genet.* **11**, e1005487 (2015).
22. Ugrankar, R., Theodoropoulos, P., Akdemir, F., Henne, W. M. & Graff, J. M.

Circulating glucose levels inversely correlate with Drosophila larval feeding through insulin signaling and SLC5A11. *Commun. Biol.* **1**, 110 (2018).

23. Puddu, A., Sanguineti, R., Montecucco, F. & Viviani, G. L. Evidence for the gut microbiota short-chain fatty acids as key pathophysiological molecules improving diabetes. *Mediators Inflamm.* **2014**, 162021 (2014).
24. Gao, Z. *et al.* Butyrate improves insulin sensitivity and increases energy expenditure in mice. *Diabetes* **58**, 1509–17 (2009).
25. Zambell, K. L., Fitch, M. D. & Fleming, S. E. Acetate and Butyrate Are the Major Substrates for De Novo Lipogenesis in Rat Colonic Epithelial Cells. *J. Nutr.* **133**, 3509–3515 (2003).
26. Nishina, P. M. & Freedland, R. A. Effects of Propionate on Lipid Biosynthesis in Isolated Rat Hepatocytes. *J. Nutr.* **120**, 668–673 (1990).
27. 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**, 3555–3557 (2015).
28. Kamat, M. A. *et al.* PhenoScanner V2: an expanded tool for searching human genotype–phenotype associations. *Bioinformatics* (2019). doi:10.1093/bioinformatics/btz469
29. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
30. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23**, R89-98 (2014).
31. Wade, K. H. & Hall, L. J. Improving causality in microbiome research: can human genetic epidemiology help? *Wellcome Open Res.* **4**, 199 (2020).
32. Tito, R. Y. *et al.* Population-level analysis of Blastocystis subtype prevalence and variation in the human gut microbiota. *Gut* gutjnl-2018-316106 (2018). doi:10.1136/gutjnl-2018-316106
33. Hildebrand, F., Tadeo, R., Voigt, A., Bork, P. & Raes, J. LotuS: an efficient and user-friendly OTU processing pipeline. *Microbiome* **2**, 30 (2014).
34. Holmes, I., Harris, K. & Quince, C. Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics. *PLoS One* **7**, e30126 (2012).
35. Karssen, L. C., van Duijn, C. M. & Aulchenko, Y. S. The GenABEL Project for statistical genomics. *F1000Research* **5**, 914 (2016).
36. R Core Team. R: A Language and Environment for Statistical Computing. (2016).
37. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–74 (2007).
38. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
39. O'Connell, J. *et al.* Haplotype estimation for biobank-scale data sets. *Nat. Genet.* **48**, 817–820 (2016).
40. Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet.* **5**, e1000529 (2009).
41. Consortium, the H. R. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
42. Deelen, P. *et al.* Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. *Eur. J. Hum. Genet.* **22**, 1321–1326 (2014).
43. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-

wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

44. Graw, S., Henn, R., Thompson, J. A. & Koestler, D. C. PwrEWAS: A user-friendly tool for comprehensive power estimation for epigenome wide association studies (EWAS). *BMC Bioinformatics* **20**, (2019).

45. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).

46. Liu, J. Z. *et al.* Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat. Genet.* **42**, 436–440 (2010).

47. Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

48. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One* **5**, e9490 (2010).

49. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* (2016). doi:10.1093/nar/gkw290

50. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).

51. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, (2015).

52. Iotchkova, V. *et al.* GARFIELD - GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction. *bioRxiv* 085738 (2016). doi:10.1101/085738

53. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* **40**, 304–314 (2016).

54. 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).

55. 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).

56. Hemani, G. *et al.* The MR-base platform supports systematic causal inference across the human phenome. *Elife* **7**, 1–29 (2018).