nature genetics

The effect of host genetics on the gut microbiome

Marc Jan Bonder[1,19], Alexander Kurilshikov[1–3,19], Ettje F Tigchelaar[1,4], Zlatan Mujagic[4,5], Floris Imhann[6], Arnau Vich Vila[6], Patrick Deelen[1,7], Tommi Vatanen[8,9], Melanie Schirmer[8,10], Sanne P Smeekens[11,12], Daria V Zhernakova[1], Soesma A Jankipersadsing[1,13], Martin Jaeger[11,12], Marije Oosting[11,12], Maria Carmen Cenit[1,18], Ad A M Masclee[5], Morris A Swertz[1,7], Yang Li[1], Vinod Kumar[1], Leo Joosten[11,12], Hermie Harmsen[14], Rinse K Weersma[6], Lude Franke[1], Marten H Hofker[13], Ramnik J Xavier[8,15–17], Daisy Jonkers[5], Mihai G Netea[11,12], Cisca Wijmenga[1], Jingyuan Fu[1,13,20] & Alexandra Zhernakova[1,4,20]

The gut microbiome is affected by multiple factors, including genetics. In this study, we assessed the influence of host genetics on microbial species, pathways and gene ontology categories, on the basis of metagenomic sequencing in 1,514 subjects. In a genome-wide analysis, we identified associations of 9 loci with microbial taxonomies and 33 loci with microbial pathways and gene ontology terms at $P < 5 \times 10^{-8}$. Additionally, in a targeted analysis of regions involved in complex diseases, innate and adaptive immunity, or food preferences, 32 loci were identified at the suggestive level of $P < 5 \times 10^{-6}$. Most of our reported associations are new, including genome-wide significance for the C-type lectin molecules *CLEC4F–CD207* at 2p13.3 and *CLEC4A–FAM90A1* at 12p13. We also identified association of a functional *LCT* SNP with the *Bifidobacterium* genus ($P = 3.45 \times 10^{-8}$) and provide evidence of a gene–diet interaction in the regulation of *Bifidobacterium* abundance. Our results demonstrate the importance of understanding host–microbe interactions to gain better insight into human health.

The gut microbiome is considered to be our second genome and is linked to many common diseases. Numerous intrinsic and extrinsic factors, including diet and medication, affect its composition[1–6]. Studies in mice[7,8] and in human twins[6,9] have observed substantial heritability for some bacteria. Such a genetic effect has not, thus far, been investigated in humans on a large scale[10,11]. Several human and mouse studies have identified interactions between host genetics and the microbiome in relation to disease phenotypes[12–14]. For example, *NOD2* and *CARD9* risk alleles associated with inflammatory bowel disease (IBD) only become manifest when triggered by the gut microbiome[15–17], indicating that genetic predisposition to diseases can depend on the microbiome.

In this study, we aimed to identify genomic loci that influence the gut microbiome in humans. We performed a three-stage association analysis using three Dutch cohorts—LifeLines-DEEP[18] ($n = 984$) as the discovery stage and 500FG[19] ($n = 425$) and MIBS-CO[20] ($n = 105$) as replication stages—followed by meta-analysis (Online Methods). Genotyping and metagenome shotgun sequencing were performed for all samples using uniform methods and pipelines. Our full data set includes 1,514 individuals. Analysis was performed in two branches: a genome-wide branch and a focused branch for which we selected SNPs from four sources—genome-wide association study (GWAS) hits for immune and metabolic traits, innate immunity bacterial-sensing genes, genes in the major histocompatibility complex (MHC) locus, and genes related to food metabolism and preferences.

We first performed genome-wide analysis of the association between common SNPs (minor allele frequency (MAF) > 0.05) and microbial taxonomies and functional units present in at least 25% of individuals (219 taxonomies, 636 MetaCyc pathways and 661 Gene Ontology (GO) terms with at least 2,000 genes (GO2000)) (**Supplementary Table 1**). For each trait, only individuals with nonzero values were included in the analysis (Online Methods). SNPs associated at $P < 5 \times 10^{-5}$ in LifeLines-DEEP and replicated at $P < 0.01$ with the same allelic direction were included in the meta-analysis (**Fig. 1** and

[1]University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, the Netherlands. [2]Institute of Chemical Biology and Fundamental Medicine SB RAS, Novosibirsk, Russia. [3]Department of Natural Sciences, Novosibirsk State University, Novosibirsk, Russia. [4]Top Institute Food and Nutrition, Wageningen, the Netherlands. [5]Division of Gastroenterology-Hepatology, Department of Internal Medicine, NUTRIM School of Nutrition and Translational Research in Metabolism, Maastricht University Medical Center, Maastricht, the Netherlands. [6]University of Groningen, University Medical Center Groningen, Department of Gastroenterology and Hepatology, Groningen, the Netherlands. [7]University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Groningen, the Netherlands. [8]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [9]Department of Computer Science, Aalto University School of Science, Espoo, Finland. [10]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. [11]Department of Internal Medicine, Radboud University Medical Center, Nijmegen, the Netherlands. [12]Radboud Center of Infectious Diseases, Radboud University Medical Center, Nijmegen, the Netherlands. [13]University of Groningen, University Medical Center Groningen, Department of Pediatrics, Groningen, the Netherlands. [14]University of Groningen, University Medical Center Groningen, Department of Medical Microbiology, Groningen, the Netherlands. [15]Center for Computational and Integrative Biology, Massachusetts General Hospital, Boston, Massachusetts, USA. [16]Gastrointestinal Unit, Massachusetts General Hospital, Boston, Massachusetts, USA. [17]Center for the Study of Inflammatory Bowel Disease, Massachusetts General Hospital, Boston, Massachusetts, USA. [18]Present address: Department of Pediatrics, Dr. Peset University Hospital, Valencia, Spain. [19]These authors contributed equally to this work. [20]These authors jointly directed this work. Correspondence should be addressed to A.Z. (a.zhernakova@umcg.nl), J.F. (j.fu@umcg.nl) or C.W. (c.wijmenga@umcg.nl).

Received 11 April; accepted 10 August; published online 3 October 2016; doi:10.1038/ng.3663

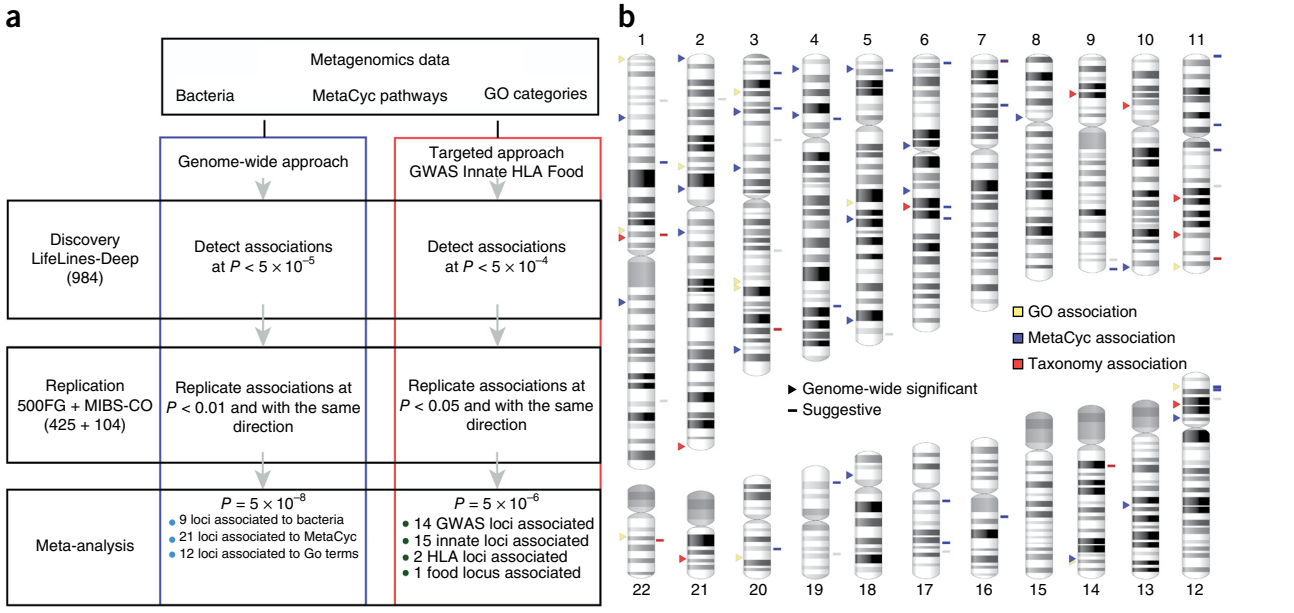


Figure 1 Data analysis workflow and association summary. (**a**) Analysis overview highlighting the steps taken and selections made. QTLs, quantitative trait loci. (**b**) Overview of the loci in the human genome that influence the gut microbiome.

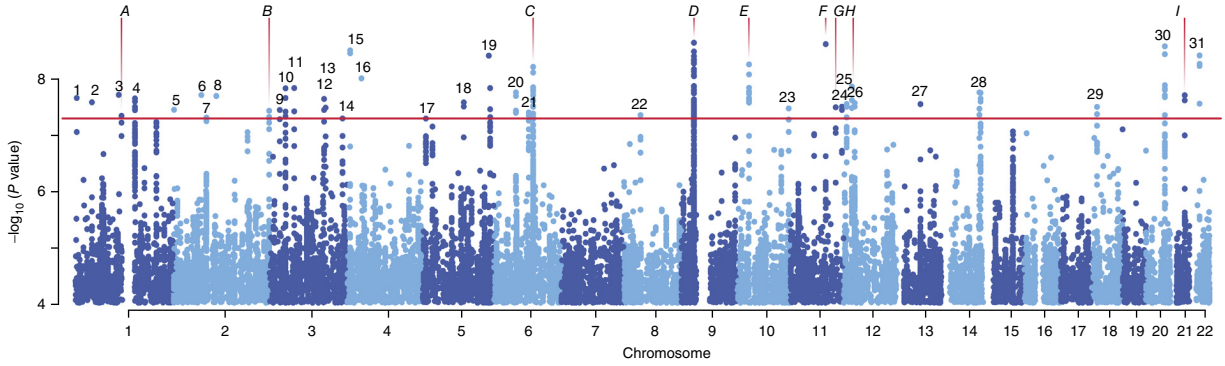
Supplementary Table 2). In this analysis, 58 SNPs at nine loci were associated with microbial taxa (**Fig. 2**, **Supplementary Fig. 1** and **Supplementary Table 3**) and a total of 33 loci were associated with functional units, with 21 loci associated with MetaCyc pathways and 12 loci associated with GO2000 groups (**Fig. 2**, **Supplementary Figs. 2** and **3**, and **Supplementary Table 3**), at $P < 5 \times 10^{-8}$ (false discovery rate (FDR) < 12%) (**Supplementary Table 4a**,**b**).

The strongest taxonomical association was observed for genus *Blautia* and family Methanobacteriaceae, for which substantial heritability was reported in twin studies (0.34 and 0.22, respectively)[6]. Bacteria in the *Blautia* genus are immunogenic epithelial-barrier-associated microbes[21] linked to abnormal Paneth cell counts[22], Crohn's disease and primary sclerosing cholangitis[23]. The associated block is a large intergenic region, located upstream of the *LINGO2* gene, which has been associated with body mass index (BMI), obesity and motion sickness[24–26]. Methanobacteriaceae have been linked to BMI and lipid levels[6,27]; the Methanobacteriaceae-associated SNPs at 6q16.1 map to an extended long noncoding RNA (lncRNA), RP11-436D23.1. Two genome-wide significant hits, at 2q37.3 and 10p12.1, were observed for *Dialister invisus*. Individuals with higher levels of *D. invisus* showed diet-dependent improvement in their inflammatory and cytokine profiles[28]. Sutterellacea abundance was associated with a SNP in the *VANGL1* gene ($P = 4.5 \times 10^{-8}$); high abundance of *Sutterella* bacteria has previously been associated with pouch health[29,30]. *VANGL1* is highly expressed in gut tissues, is involved in wound healing of the intestinal mucosa[31] and mediates the invasiveness and progression of colon cancer cells[32,33].

For MetaCyc pathways, the strongest association was observed for a pathway involved in plant-derived steroid degradation (PWY-6948). Plant sterols have been suggested to have a beneficial effect on metabolic syndromes[34]. Abundance of reads mapping to genes in this pathway was associated with SNPs in two independent loci: the *SORCS2* gene ($P = 3.1 \times 10^{-9}$) at 4p16.1, which is also associated with insulin-like growth factors[35], and the *SLIT3* gene ($P = 3.8 \times 10^{-9}$) at 5q35, which is also associated with obesity[36]. Another strong association was seen for the bile acid metabolism pathway (PWY-6518) ($P = 9.7 \times 10^{-9}$) with SNPs in the *ARAP1* gene, which is involved in focal adhesion and also associated with type 2 diabetes[37]. Of the 12 loci associated with GO terms, 2 were located near clusters of C-type lectin domain family 4 genes: the *CLEC4F–CD207* (*CLEC4K*) genes at 2p13.3 and the *CLEC4A–FAM90A1* locus at 12p13. These genes encode factors involved in cell adhesion, cell signaling, inflammation and immune response that are known to bind intestinal microbiota and to modulate the production of proinflammatory cytokines when activated. Several SNPs at the *CLEC4A–FAM90A1* locus correlated with the GO term 'riboflavin biosynthetic process'. Riboflavin is a redox mediator that can stimulate the growth of certain gut bacteria, including *Faecalibacterium prauznitsii*[38]; its metabolism is also increased in ileal Crohn's disease[39].

Overall, we observed that the genetic variants associated with functional profiles showed little overlap with the taxonomy-associated SNPs. Studies have shown that microbial composition can be highly variable, even when individuals have similar functional profiles[40]. We investigated whether associated pathways and GO terms could be driven by specific microbial taxonomies by assessing pathway–bacteria and GO term–bacteria correlations (**Supplementary Fig. 4** and **Supplementary Tables 5** and **6**). We observed that most associated pathways and GO terms were driven by several microbial taxonomies. This partly explains the low overlap between taxonomy- and function-associated loci and suggests that host genetics can directly shape the functional composition of the microbiome. Another explanation for this low overlap could be limitation of analysis power. The average zero rate (percentage of samples with no reads in a category (bacterial taxonomy or pathway)) per entry for functional data was 5%, in contrast to a rate of 35% for taxonomic abundance (**Supplementary Table 1**). The larger sampling for microbial function leads to higher discovery power.

We further performed a targeted analysis using less stringent thresholds: $P < 5 \times 10^{-4}$ at discovery, $P < 0.05$ at replication and $P < 5 \times 10^{-6}$ in the meta-analysis (**Fig. 1** and **Supplementary Table 2**), corresponding to FDR < 1% (**Supplementary Table 4c**,**d**), on four categories of relevant immune response– and metabolism-related genes: genes with GWAS SNPs for immune and metabolic traits;



	SNP	Microbial taxon, MetaCyc pathway or GO term		SNP	Microbial taxon, MetaCyc pathway or GO term
1	rs199545687	Microtubule (GO:0005874)	21	rs6933130	L-methionine biosynthesis (PWY-5345)
2	rs12563071	Ubiquinol-8 biosynthesis (UBISYN-PWY)	22	rs7016086	Ubiquinol-7 biosynthesis (PWY-5873)
3	rs958798	Transcription regulation from RNA Pol II (GO:0045944)	23	rs1041530	Seleno-compound metabolism (PWY-6395)
4	rs12048670	Polyamine biosynthesis II (POLYAMINSYN3-PWY)	24	rs11606643	Innate immune response (GO:0045087)
5	rs4553849	Creatinine degradation II (PWY-4722)	25	rs9669179	Riboflavin biosynthetic process (GO:0009231)
6	rs6546647	Embryonic morphogenesis (GO:0048598)	26	rs7133214	L-methionine salvage cycle I (PWY-7528)
7	rs2084597	Starch degradation III (PWY-6731)	27	rs7992913	L-isoleucine biosynthesis I (PWY-3001)
8	rs17770672	Demethylmenaquinol-9 biosynthesis (PWY-5862)	28	rs74773701 + rs4144435	P562-PWY + GO:0042823
9	rs2166811	Oxidoreductase activity (GO:0016706)	29	rs113062739	Demethylmenaquinol-9 biosynthesis (PWY-5862)
10	rs4973961	Thiamine biosynthesis (PWY-7282)	30	rs17789629	Sulfuric ester hydrolase activity (GO:0008484)
11	rs924067	4-chlorobenzoate degradation (PWY-6215)	31	rs2285198	Cell proliferation (GO:0008283)
12	rs1497266	Thiamine diphosphate biosynth. proc. (GO:0009229)	*A*	*rs12137699*	Family Sutterellaceae
13	rs10935496	Drug transmembrane transporter act. (GO:0015238)	*B*	*rs7605872*	Species *Dialister invisus*
14	rs35598536	Chorismate biosynthesis I (ARO-PWY)	*C*	*rs4548017*	Class Methanobacteria
15	rs10012347	Sitosterol degradation (PWY-6948)	*D*	*rs10813066*	Genus *Blautia*
16	rs12645801	Glycocholate metabolism (PWY-6518)	*E*	*rs1889714*	Species *Dialister invisus*
17	rs1666789	PROTOCATECHUATE-ORTHO-CLEAVAGE-PWY	*F*	*rs16913594*	Species *Bacteroides xylanisolvens*
18	rs78533343	Aerobic respiration I (PWY-3781)	*G*	*rs17115310*	Family Acidaminococcaceae
19	rs2163761 + rs56879175	PWY-6948 + GO:0003697	*H*	*rs10743315*	Species Lachnospiraceae bacterium 1 1 57FAA
20	rs9475677	Methylaspartate cycle (PWY-6728)	*I*	*rs2834288*	Family Oscillospiraceae

Figure 2 Manhattan plot of genome-wide associations with microbes and functional units (MetaCyc pathways and GO2000 terms). The Manhattan plot shows QTLs for microbial abundance, labeled with letters, and QTLs for microbial function, labeled with numbers. Details for the associations are given in the table below; QTLs for microbial abundance are italicized. The red line corresponds to a significance threshold of 5×10^{-8}.

genes encoding innate immunity molecules involved in microbial sensing; genes in the MHC locus, representing the adaptive immune system; and genes involved in food tolerance and preferences (**Supplementary Tables 7** and **8**).

In the analysis of GWAS hits, 14 loci were associated with the gut microbiome (**Supplementary Table 9**). The strongest signal ($P = 1.3 \times 10^{-7}$) was observed for the GO2000 term 'cell–cell signaling', associated with rs2155219 and located in the *C11orf30–LRRC32* locus, which has been associated with multiple immune-related phenotypes, including IBD and allergy[15,41]. This GO term is a general functional category, but its abundance was correlated with the abundance of two IBD-associated bacteria[42], *Coprococcus comes* ($r_S = 0.55$, $P = 4.64 \times 10^{-86}$) and Proteobacteria ($r_S = -0.42$, $P = 7.10 \times 10^{-49}$). Three other IBD-associated genes were also associated with the gut microbiome (**Supplementary Table 9**): *CCL2*, which contributes to microbiota-induced inflammation in mice on a high-fat diet[43]; *DAP2*, which is a negative regulator of autophagy; and *IL23R*, which mediates IL-23–induced type 17 helper T cell (T_H17) responses, which are crucial for mucosal immunity[44,45]. We also observed association of a SNP implicated in Behçet's disease (rs1800871), located in the *IL10* gene, with the cullin–RING ubiquitin ligase complex (GO:0031461). IL-10 is an important immunoregulator of the gut, and mutations in the corresponding gene have been reported in cases of severe IBD[46]. Our findings, together with epidemiological evidence for the

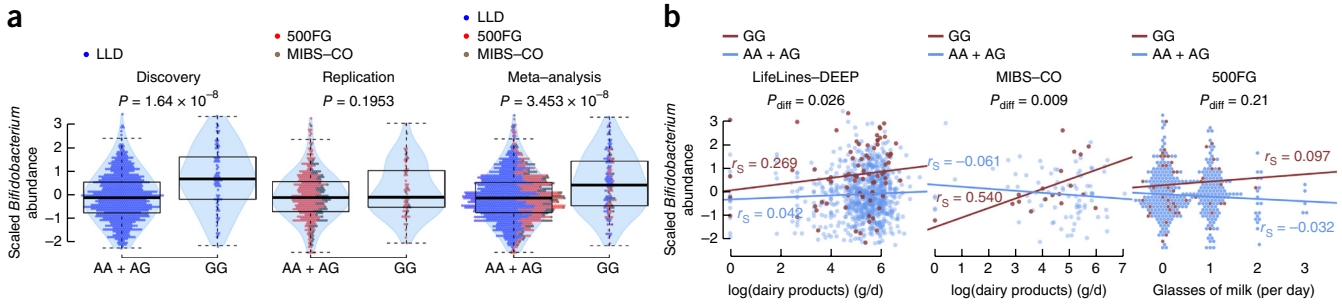


Figure 3 Complex interaction between a functional *LCT* variant, dairy intake and *Bifidobacterium* abundance. (**a**) Microbial QTL plots showing association of a functional *LCT* SNP (rs4988235) with the abundance of *Bifidobacterium*. The three plots show the effect of genotype at the functional SNP on *Bifidobacterium* abundance. The plots are a combination of violin plots and box plots; box plots show the median and the 25% and 75% quantiles. LLD, LifeLines-DEEP. (**b**) Interaction of *LCT* genotype, intake of dairy products and *Bifidobacterium* abundance in the three cohorts. Data on total dairy consumption were unavailable for the 500FG cohort, but the same trend was observed with respect to data on the number of glasses of milk drunk per day in this cohort. Pdiff shows the significance of the difference in the correlation of diary product consumption and Bifidobacterium abundance between the two genotype groups, GG versus AA and AG (Online Methods).

role of infections in Behçet's disease[47], encourage further investigation into the gut microbiome in these patients.

We also identified associations with several metabolic loci, including SNPs located in two genes encoding lipid transfer proteins, *PLTP* and *APOE*[48,49], and in the *PPARG* gene, which encodes a regulator of fatty acid storage and glucose metabolism involved in lipodystrophy and type 2 diabetes[50–52]. Finally, we observed an interesting association between the abundance of *Lactococcus* bacteria and a SNP associated with body fat distribution (rs2294239) that affects expression of the nearby *ZNRF3* gene. *ZNRF3* is a negative regulator of Wnt signaling in Paneth cells[53,54] and acts as a tumor suppressor in gastric cancer[55]. Thus, association of the microbiome with several GWAS SNPs suggests that the link between host genetics and immunological and metabolic phenotypes can be mediated by gut microbiota.

To investigate the effect of variation in host immune genes on the microbiome, we selected key molecules for microbial sensing in innate immunity on the basis of previous studies[56–58] (**Supplementary Table 8**). Furthermore, we imputed all SNPs, amino acids and alleles for *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DQ*, *HLA-DP* and *HLA-DR* genes located in the MHC region. Only two genetic variants in the MHC region were associated: an amino acid encoded in the *HLA-B* gene and SNP rs3873352 located near the *MUC22* gene (**Supplementary Table 10**). Many more associations were observed with innate immunity (**Supplementary Table 11**). Fifteen loci encoding sensors involved in innate immunity were associated with microbiome composition or function at $P < 5 \times 10^{-6}$. In particular, the *NOD2* locus (a strong risk locus for Crohn's disease, previously linked with microbiome composition[16]) was associated with the MetaCyc pathway abundance of enterobactin biosynthesis. This pathway showed strong correlation with the abundance of *Escherichia coli* (**Supplementary Table 5**). Enterobactin produced by *E. coli* inhibits myeloperoxidase (MPO), a bactericidal host enzyme, thereby providing a survival advantage that allows *E. coli* to bypass host innate immune responses in inflammatory gut disease[59]. Moreover, SNPs in the *NOD1* gene were associated with several pathways or GO categories, most of which were also driven by Enterobacteriales and, in particular, by *E. coli*. This bacterium is commonly associated with gut inflammation and Crohn's disease[60,61].

Consistent with the genome-wide associations detected for the CLEC gene clusters, targeted analysis found associations with two other CLEC loci: the CLEC gene clusters at 12p13.3 (*CLEC6A*, *CLEC4E* and *CLEC7A*) and *CD209–CLEC4G* (dectin-1) at 19p13.2 (**Supplementary Table 11**). C-type lectin receptors, together with Toll-like and NOD-like receptors, have a major role in microbial recognition and in activation of the inflammatory reactions that are essential for controlling infections[62]. Mice lacking *Clec7a* (dectin-1) have increased severity of chemically induced colitis, and polymorphisms in the *CLEC7A* gene in humans are linked to increased severity of ulcerative colitis[63]. In general, our analysis indicated a stronger link of microbiota composition and function with genes involved in innate receptors than with adaptive immunity genes.

As it is known that food intake can be affected by the host genome, we also selected SNPs from seven genes involved in food metabolism and/or consumption preferences for analysis (**Supplementary Table 8**). SNPs located in the *LCT* locus were associated with five GO terms at $P < 5 \times 10^{-6}$ (**Supplementary Table 12**). The *LCT* locus has been widely studied in relation to adult-type lactose intolerance (hypolactasia) and the inability of adults to digest milk[64,65]. The functional SNP rs4988235 tags the primary haplotype associated with hypolactasia in European populations and predisposes to hypolactasia when homozygous for the G allele[66]. Given previous observations linking this functional variant to *Bifidobacterium*[9,10], we used a recessive model to investigate the relationship between rs4988235 and abundance of *Bifidobacterium*. Indeed, we observed that the GG genotype was related to high abundance of *Bifidobacterium* ($P = 3.45 \times 10^{-8}$) (**Fig. 3a**). Next, we tested whether this haplotype and *Bifidobacterium* abundance were associated with the consumption of dairy products. Quantitative information on dairy intake was available for the LifeLines-DEEP and MIBS-CO cohorts. We did not observe a significant difference in dairy product consumption among individuals with different genotypes at rs4988235, nor was there any correlation between dairy product consumption and *Bifidobacterium* abundance (**Supplementary Fig. 5**). However, we found that the abundance of *Bifidobacterium* was dependent on an interaction between genotype and intake of dairy products. In individuals with the GG genotype, a relationship between *Bifidobacterium* abundance and milk product consumption was observed in both the LifeLines-DEEP and MIBS-CO cohorts ($P_{meta} = 0.0144$) (**Fig. 3b**). In the 500FG cohort, a similar trend was observed, although it was not significant (**Fig. 3b**). In summary, we detected the recessive effect of an *LCT* functional variant on the abundance of *Bifidobacterium* and found evidence of interaction between host genetics and diet in regulating microbiome composition.

In conclusion, we performed a GWAS of gut microbiome composition and function, accessed by metagenomic sequencing in a large population-based cohort of 1,514 individuals. We identified multiple associations of genetic variants with human gut microbiome composition and function, which highlight the role of sensor molecules in innate immunity, particularly C-type lectins, in gut homeostasis and provide evidence for an interaction between host genetics, diet and the microbiome. Identifying associations between human genetics and the gut microbiome, and exploring their interactions, can provide insights into the role of the microbiome in complex diseases and drive the development of therapies to modulate the microbiome toward better health.

URLs. Human Functional Genomics Project, http://www.humanfunctionalgenomics.org/; HUMAnN2, http://huttenhower.sph.harvard.edu/humann2; MetaCyc metabolic pathway database, http://www.metacyc.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The LifeLines-DEEP and MIBS metagenomics sequencing data are available at the European Genome-phenome Archive (EGA); LifeLines-DEEP, EGAS00001001704; MIBS, EGAS00001001924. The 500FG data are available at the Sequence Read Archive (SRA): PRJNA319574.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the participants and the staff of LifeLines-DEEP, 500FG and MIBS for their collaboration. We thank J. Dekens, M. Platteel, J. Pietersma and A. Maatman for management and technical support, and K. McIntyre and J. Senior for editing the manuscript.

This project was funded by grants from Top Institute Food and Nutrition, Wageningen, to C.W. (TiFN GH001), the Netherlands Organization for Scientific Research to J.F. (NWO-VIDI 864.13.013), L.F. (ZonMW-VIDI 917.14.374) and R.K.W. (ZonMW-VIDI 016.136.308), and CardioVasculair Onderzoek Nederland to M.H.H., M.G.N., A.Z. and J.F. (CVON 2012-03). A.Z. holds a Rosalind Franklin

Fellowship (University of Groningen). This research received funding from the European Research Council under the European Union's Seventh Framework Programme: C.W. is supported by FP7/2007-2013/ERC Advanced Grant (agreement 2012-322698) and a Spinoza Prize from the Netherlands Organization for Scientific Research. M.G.N. holds an ERC Consolidator Grant (310372). L.F. has an FP7/2007-2013 grant (agreement 259867) and an ERC Starting Grant (637640, ImmRisk). Y.L. holds a Netherlands Organization for Scientific Research VENI grant (863.13.011).

AUTHOR CONTRIBUTIONS

Conceptualization: A.Z., J.F., C.W. and M.J.B. Methodology: M.J.B., A.K., L.F., J.F., P.D., T.V. and M.S. Software: M.J.B., A.K., L.F., J.F., P.D., M.A.S. and D.V.Z. Formal analysis: M.J.B., A.K., J.F. and A.Z. Investigation: A.Z., J.F., M.J.B., A.K., F.I., D.V.Z., S.A.J., A.V.V., E.F.T., H.H. and M.C.C. Resources: C.W., A.Z., L.F., J.F., M.A.S., M.G.N., R.J.X., L.J. and A.A.M.M. Data curation: M.J.B., A.K., J.F., P.D., L.F., S.A.J. and Y.L. Writing–original draft: A.Z., J.F., M.J.B., A.K. and C.W. Writing–review and editing: M.J.B., A.K., E.F.T., Z.M., F.I., A.V.V., P.D., T.V., M.S., S.P.S., D.V.Z., S.A.J., M.J., M.O., M.A.S., M.C.C., Y.L., V.K., H.H., R.K.W., L.F., M.H.H., D.J., M.G.N., C.W., J.F. and A.Z. Visualization: A.K., M.J.B., A.Z. and J.F. Supervision: A.Z., J.F., C.W., L.F., R.K.W. and M.H.H. Project administration: A.Z., J.F., C.W., L.F., M.G.N., D.J., A.A.M.M. and S.P.S. Funding acquisition: A.Z., J.F., C.W., L.F., M.G.N., D.J. and A.A.M.M.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Tigchelaar, E.F. *et al.* Gut microbiota composition associated with stool consistency. *Gut* **65**, 540–542 (2015).
2. Imhann, F. *et al.* Proton pump inhibitors affect the gut microbiome. *Gut* **65**, 740–748 (2015).
3. Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).
4. David, L.A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563 (2014).
5. Scott, K.P., Gratz, S.W., Sheridan, P.O., Flint, H.J. & Duncan, S.H. The influence of diet on the gut microbiota. *Pharmacol. Res.* **69**, 52–60 (2013).
6. Goodrich, J.K. *et al.* Human genetics shape the gut microbiome. *Cell* **159**, 789–799 (2014).
7. Org, E. *et al.* Genetic and environmental control of host–gut microbiota interactions. *Genome Res.* **25**, 1558–1569 (2015).
8. Leamy, L.J. *et al.* Host genetics and diet, but not immunoglobulin A expression, converge to shape compositional features of the gut microbiome in an advanced intercross population of mice. *Genome Biol.* **15**, 552 (2014).
9. Goodrich, J.K. *et al.* Genetic determinants of the gut microbiome in UK Twins. *Cell Host Microbe* **19**, 731–743 (2016).
10. Blekhman, R. *et al.* Host genetic variation impacts microbiome composition across human body sites. *Genome Biol.* **16**, 191 (2015).
11. Davenport, E.R. *et al.* Genome-wide association studies of the human gut microbiota. *PLoS One* **10**, e0140301 (2015).
12. Srinivas, G. *et al.* Genome-wide mapping of gene–microbiota interactions in susceptibility to autoimmune skin blistering. *Nat. Commun.* **4**, 2462 (2013).
13. Parks, B.W. *et al.* Genetic control of obesity and gut microbiota composition in response to high-fat, high-sucrose diet in mice. *Cell Metab.* **17**, 141–152 (2013).
14. McKnite, A.M. *et al.* Murine gut microbiota is defined by host genetics and modulates variation of metabolic traits. *PLoS One* **7**, e39191 (2012).
15. Jostins, L. *et al.* Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
16. Knights, D. *et al.* Complex host genetics influence the microbiome in inflammatory bowel disease. *Genome Med.* **6**, 107 (2014).
17. Lamas, B. *et al.* CARD9 impacts colitis by altering gut microbiota metabolism of tryptophan into aryl hydrocarbon receptor ligands. *Nat. Med.* **22**, 598–605 (2016).
18. Tigchelaar, E.F. *et al.* Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open* **5**, e006772 (2015).
19. Netea, M.G. *et al.* Understanding human immune function using the resources from the Human Functional Genomics Project. *Nat. Med.* **22**, 831–833 (2016).
20. Mujagic, Z. *et al.* Small intestinal permeability is increased in diarrhoea predominant IBS, while alterations in gastroduodenal permeability in all IBS subtypes are largely attributable to confounders. *Aliment. Pharmacol. Ther.* **40**, 288–297 (2014).
21. Juste, C. *et al.* Bacterial protein signals are associated with Crohn's disease. *Gut* **63**, 1566–1577 (2014).
22. Liu, T.-C. *et al.* O-011 Paneth cell phenotypes define a subtype of pediatric Crohn's disease through alterations in host–microbial interactions. *Inflamm. Bowel Dis.* **22** (Suppl. 1), S4 (2016).
23. Torres, J. *et al.* The features of mucosa-associated microbiota in primary sclerosing cholangitis. *Aliment. Pharmacol. Ther.* **43**, 790–801 (2016).
24. Hromatka, B.S. *et al.* Genetic variants associated with motion sickness point to roles for inner ear development, neurological processes and glucose homeostasis. *Hum. Mol. Genet.* **24**, 2700–2708 (2015).
25. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
26. Williams, A.L. *et al.* Sequence variants in *SLC16A11* are a common risk factor for type 2 diabetes in Mexico. *Nature* **506**, 97–101 (2014).
27. Fu, J. *et al.* The gut microbiome contributes to a substantial proportion of the variation in blood lipids. *Circ. Res.* **117**, 817–824 (2015).
28. Martínez, I. *et al.* Gut microbiome composition is linked to whole grain–induced immunological improvements. *ISME J.* **7**, 269–280 (2013).
29. Tyler, A.D. *et al.* Characterization of the gut-associated microbiome in inflammatory pouch complications following ileal pouch-anal anastomosis. *PLoS One* **8**, e66934 (2013).
30. Landy, J. *et al.* Variable alterations of the microbiota, without metabolic or immunological change, following faecal microbiota transplantation in patients with chronic pouchitis. *Sci. Rep.* **5**, 12955 (2015).
31. Kalabis, J., Rosenberg, I. & Podolsky, D.K. Vangl1 protein acts as a downstream effector of intestinal trefoil factor (ITF)/TFF3 signaling and regulates wound healing of intestinal epithelium. *J. Biol. Chem.* **281**, 6434–6441 (2006).
32. Bae, J.A. *et al.* An unconventional KITENIN/ErbB4-mediated downstream signal of EGF upregulates c-Jun and the invasiveness of colorectal cancer cells. *Clin. Cancer Res.* **20**, 4115–4128 (2014).
33. Lee, S. *et al.* Expression of KITENIN in human colorectal cancer and its relation to tumor behavior and progression. *Pathol. Int.* **61**, 210–220 (2011).
34. Klingberg, S. *et al.* Inverse relation between dietary intake of naturally occurring plant sterols and serum cholesterol in northern Sweden. *Am. J. Clin. Nutr.* **87**, 993–1001 (2008).
35. Kaplan, R.C. *et al.* A genome-wide association study identifies novel loci associated with circulating IGF-I and IGFBP-3. *Hum. Mol. Genet.* **20**, 1241–1251 (2011).
36. Liu, Y.J. *et al.* Genome-wide association scans identified *CTNNBL1* as a novel gene for obesity. *Hum. Mol. Genet.* **17**, 1803–1813 (2008).
37. Strawbridge, R.J. *et al.* Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes* **60**, 2624–2634 (2011).
38. Khan, M.T., Browne, W.R., van Dijl, J.M. & Harmsen, H.J.M. How can *Faecalibacterium prausnitzii* employ riboflavin for extracellular electron transfer? *Antioxid. Redox Signal.* **17**, 1433–1440 (2012).
39. Morgan, X.C. *et al.* Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* **13**, R79 (2012).
40. Lozupone, C.A., Stombaugh, J.I., Gordon, J.I., Jansson, J.K. & Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* **489**, 220–230 (2012).
41. Bønnelykke, K. *et al.* Meta-analysis of genome-wide association studies identifies ten loci influencing allergic sensitization. *Nat. Genet.* **45**, 902–906 (2013).
42. Gevers, D. *et al.* The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* **15**, 382–392 (2014).
43. Caesar, R., Tremaroli, V., Kovatcheva-Datchary, P., Cani, P.D. & Bäckhed, F. Crosstalk between gut microbiota and dietary lipids aggravates WAT inflammation through TLR signaling. *Cell Metab.* **22**, 658–668 (2015).
44. Yahiro, K. *et al.* DAP1, a negative regulator of autophagy, controls SubAB-mediated apoptosis and autophagy. *Infect. Immun.* **82**, 4899–4908 (2014).
45. Koren, I., Reem, E. & Kimchi, A. DAP1, a novel substrate of mTOR, negatively regulates autophagy. *Curr. Biol.* **20**, 1093–1098 (2010).
46. Glocker, E.O. *et al.* Inflammatory bowel disease and mutations affecting the interleukin-10 receptor. *N. Engl. J. Med.* **361**, 2033–2045 (2009).
47. Galeone, M., Colucci, R., D'Erme, A.M., Moretti, S. & Lotti, T. Potential infectious etiology of Behçet's disease. *Patholog. Res. Int.* **2012**, 595380 (2012).
48. Huuskonen, J., Olkkonen, V.M., Jauhiainen, M. & Ehnholm, C. The impact of phospholipid transfer protein (PLTP) on HDL metabolism. *Atherosclerosis* **155**, 269–281 (2001).
49. Getz, G.S. & Reardon, C.A. Apoprotein E as a lipid transport and signaling protein in the blood, liver, and artery wall. *J. Lipid Res.* **50** (Suppl.), S156–S161 (2009).
50. Hevener, A.L. *et al.* Muscle-specific *Pparg* deletion causes insulin resistance. *Nat. Med.* **9**, 1491–1497 (2003).
51. Hegele, R.A., Cao, H., Frankowski, C., Mathews, S.T. & Leff, T. PPARG F388L, a transactivation-deficient mutant, in familial partial lipodystrophy. *Diabetes* **51**, 3586–3590 (2002).
52. Doney, A.S.F. *et al.* Association of the Pro12Ala and C1431T variants of *PPARG* and their haplotypes with susceptibility to type 2 diabetes. *Diabetologia* **47**, 555–558 (2004).
53. Hao, H.-X. *et al.* ZNRF3 promotes Wnt receptor turnover in an R-spondin–sensitive manner. *Nature* **485**, 195–200 (2012).
54. Farin, H.F. *et al.* Visualization of a short-range Wnt gradient in the intestinal stem-cell niche. *Nature* **530**, 340–343 (2016).
55. Zhou, Y. *et al.* ZNRF3 acts as a tumour suppressor by the Wnt signalling pathway in human gastric adenocarcinoma. *J. Mol. Histol.* **44**, 555–563 (2013).
56. Casals, F. *et al.* Genetic adaptation of the antibacterial human innate immunity network. *BMC Evol. Biol.* **11**, 202 (2011).
57. Bunge, J., Willis, A. & Walsh, F. Estimating the number of species in microbial diversity studies. *Annu. Rev. Stat. Appl.* **1**, 427–445 (2014).

58. Caballero, S. & Pamer, E.G. Microbiota-mediated inflammation and antimicrobial defense in the intestine. *Annu. Rev. Immunol.* **33**, 227–256 (2015).
59. Singh, V. *et al.* Interplay between enterobactin, myeloperoxidase and lipocalin 2 regulates *E. coli* survival in the inflamed gut. *Nat. Commun.* **6**, 7113 (2015).
60. Nguyen, H.T.T. *et al.* Crohn's disease–associated adherent invasive *Escherichia coli* modulate levels of microRNAs in intestinal epithelial cells to reduce autophagy. *Gastroenterology* **146**, 508–519 (2014).
61. Sadaghian Sadabad, M. *et al.* The *ATG16L1*-T300A allele impairs clearance of pathosymbionts in the inflamed ileal mucosa of Crohn's disease patients. *Gut* **64**, 1546–1552 (2014).
62. Dambuza, I.M. & Brown, G.D. C-type lectins in immunity: recent developments. *Curr. Opin. Immunol.* **32**, 21–27 (2015).
63. Iliev, I. D. *et al.* Interactions between commensal fungi and the C-type lectin receptor Dectin-1 influence colitis. *Science* **336**, 1314–1317 (2012).
64. Enattah, N.S. *et al.* Evidence of still-ongoing convergence evolution of the lactase persistence T-13910 alleles in humans. *Am. J. Hum. Genet.* **81**, 615–625 (2007).
65. Tishkoff, S.A. *et al.* Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* **39**, 31–40 (2007).
66. Troelsen, J.T. Adult-type hypolactasia and regulation of lactase expression. *Biochim. Biophys. Acta* **1723**, 19–32 (2005).

ONLINE METHODS

Cohort descriptions. This study uses three independent Dutch population cohorts: the LifeLines-DEEP cohort as a discovery cohort and the 500 Functional Genomics cohort (500FG) and the controls in the Maastricht IBS cohort (MIBS-CO) as replication cohorts. The LifeLines-DEEP cohort consists of 1,539 individuals from the three northern provinces of the Netherlands (636 males and 903 females, age range 18–84 years)[1]. The replication cohorts comprise 534 individuals from the 500FG cohort (from the Human Functional Genomics Study; 237 males and 296 females, age range 18–75 years) and 105 healthy controls from the MIBS cohort (42 males and 63 females, age range 19–71 years)[20]. We confined our analysis to the set of individuals with both high-quality genotype and high-quality microbiome data: 984 individuals from LifeLines-DEEP, 425 individuals from 500FG and 105 controls from MIBS (MIBS-CO).

For all three cohorts, extensive phenotype information was available, including food intake. For 500FG, this was based on questionnaire data for selected food items. LifeLines-DEEP and MIBS used a validated food frequency questionnaire[67,68].

Informed consent. The LifeLines-DEEP and MIBS-CO studies were approved by the institutional ethics review boards of the UMCG and MUMC (ClinicalTrials.gov NCT00775060). The 500FG study was approved by the Ethical Committee of Radboud University Nijmegen (NL42561.091.12, 2012/550). All participants signed an informed consent form. This study was approved by the institutional ethical review board of the UMCG, M12.113965.

Genotyping. For all three cohorts, genome-wide genotyping was performed and the remaining SNPs and HLA alleles were imputed. Genotyping of the LifeLines-DEEP samples is described in Tigchelaar *et al.*[18]. In short, individuals were genotyped using both the HumanCytoSNP-12 BeadChip and Immunochip, a customized Illumina Infinium array. The data were merged and subsequently imputed using the Genome of the Netherlands (GoNL) data set[69]. We removed ancestry outliers and genetically related participants to include a total of 1,268 individuals in the final genetic study.

For the 500FG and MIBS cohorts, DNA samples from 516 500FG and 288 MIBS (cases and controls) individuals were genotyped using the commercially available SNP chip Illumina HumanCoreExome-12 v1.1. Genotype calling was performed with Opticall 0.7.0 (ref. 70) using the default settings. Samples with a call rate ≤0.99 were excluded from the data set, as were variants with a Hardy–Weinberg equilibrium P value ≤0.0001, call rate ≤0.99 or MAF ≤0.001. Ancestry outliers were identified from multidimensional scaling plots of samples merged with 1000 Genomes Project data and were then removed from our data set. This filtering step resulted in two high-quality genetic data sets of 487 500FG individuals and 287 MIBS individuals. The final data contained genotype information for 518,980 variants in both data sets for further imputation. The strands of the variants were aligned, and identifiers were updated to a combined reference of 1000 Genomes Project phase 3 v5 (ref. 71) and the GoNL data set[69] using Genotype Harmonizer[72]. The data were phased using SHAPEIT2 v2.r644 (ref. 73) with the combined reference panel. Finally, these data were imputed using IMPUTE2 (ref. 74) with GoNL as the reference panel[75].

Further, we imputed the HLA region in the three cohorts using the T1DGC reference derived from the SNP2HLA v 1.0.2 (ref. 76) reference and developed by the Type 1 Diabetes Genetics Consortium (TIDGC)[76]. We converted the SNP2HLA reference to a build 37 IMPUTE2 reference and separately harmonized and imputed our data sets using IMPUTE2 (ref. 74). In total, this study tested associations for 8.1 million SNPs in the genome-wide analysis and 8,606 variants in the HLA region.

Metagenomic sequencing. *Sequencing and read filtering.* Metagenomic sequencing was performed for 1,179 LifeLines-DEEP samples, 520 500FG samples and 626 MIBS samples (cases and controls) using the Illumina HiSeq platform. Within 2 weeks of participants giving a blood sample, they collected fecal samples at home and stored them immediately at −20 °C. After transport to the research laboratory on dry ice, fecal samples were stored at −80 °C. Aliquots were made, and DNA was isolated with the AllPrep DNA/RNA Mini kit (Qiagen, 80204) with the addition of mechanical lysis. Reads were quality filtered, and adaptor removal was performed using Trimmomatic (v.0.32)[77]. An average of 3.0 Gb of data (around 32.3 million high-quality reads) was obtained per sample. Reads belonging to the human genome were removed by mapping the data to the human reference genome (version NCBI37) with Bowtie2 (v.2.1.0). Finally, 99 samples with fewer reads than our threshold of 15 million reads were removed (44 samples from LifeLines-DEEP, 10 samples from 500FG and 45 samples from MIBS).

Microbial data processing. The profile of microbial composition was determined using MetaPhlan 2.2 (ref. 78), which uses a set of ~1 million markers (average of 184 marker genes for each microbial clade) from >7,500 species. MetaPhlan 2.2 reported the abundance of 1,772 microbial taxonomies in our data, including 21 phyla, 32 classes, 50 orders, 108 families, 235 genera and 678 species from four different domains. We further normalized the taxonomy data using an arcsin square root transformation and corrected the normalized nonzero data for age, sex and read depth.

Functional profiling was performed using HUMAnN2, which maps reads to a customized database of functionally annotated pan-genomes. This analysis showed the abundance of 5,379,353 gene families from the UniProt Reference Clusters that were further mapped to 773 pathways from the MetaCyc metabolic pathway database. We grouped gene families on the basis of GO[79,80]. Using the hierarchy in GO, we isolated a subset of 'informative' GO terms, which we defined as those associated with >2,000 proteins for which no descendant term was associated. This yielded a set of 611 non-redundant GO terms for subsequent analysis. Details about the grouping can be found in Vatanen *et al.*[81]. For the gene counts for MetaCyc pathways and GO2000 terms, we first quantile normalized the data and then corrected the normalized nonzero data for age, sex and read depth.

Relating abundances of microbes, pathways and genes. As we extracted multiple levels of information on microbial abundance and functional entities from the metagenomics data, we also investigated the correlations between microbial taxonomies and pathways/GO terms using Spearman's rank correlation in the R 'base' package. Co rPvalues were Bonferroni corrected for the number of tests performed in each case (number of microbial taxonomies on the specified taxonomic level multiplied by the number of pathways or GO terms). All correlations mentioned were significant after correction for multiple testing.

Quantitative trait locus association analysis. We confined genetic analysis to the taxonomies, MetaCyc pathways and GO2000 terms that were present in at least 25% of the samples in our data sets. In total, 236 taxonomies, 636 MetaCyc pathways and 611 GO2000 terms were included for analysis in the LifeLines-DEEP discovery cohort. We were able to perform replication analysis for 219 taxonomies, 636 MetaCyc pathways and 611 GO2000 terms in the replication cohorts.

To link microbial composition and function to genetic variation, normalized abundance values of taxonomies, GO2000 terms and MetaCyc pathways were treated as quantitative traits and we used rank-based Spearman correlation analysis to identify association with the genotype of each SNP. For each trait, only individuals with nonzero values were included. These analysis steps have been incorporated into our QTL mapping pipeline[82], along with various quality control steps. For this specific application, microbiome QTL mapping, we have included a filter for zero counts, to only incorporate nonzero values during the analysis. We chose to do so because zero counts can cause the data distribution to depart from normality, thus inducing a bias when correcting for the effect of age, sex and sequence depth using a linear model. In such a way, the number of samples can differ per taxon or bacterial pathway, thereby affecting analysis power. The common taxa or pathways have more nonzero values, yielding more power to identify QTLs.

We preformed QTL mapping in three steps. Association was first tested in the LifeLines-DEEP cohort. It was then replicated by a meta-analysis in the 500FG and MIBS cohorts. Finally, we performed a full meta-analysis over all three data sets. In this way, we performed genetic analysis at both the genome-wide level and the targeted gene level (**Fig. 1** and **Supplementary Table 2**).

Genome-wide association. At the genome-wide level, we performed association analysis for all 8.1 million SNPs. Associations with bacteria, pathways

and GO2000 terms were first identified at a *P*-value threshold of 5×10^{-5} in the discovery cohort. We then replicated these associations in the replication cohorts and selected associations that were replicated at $P < 0.01$ in the same allelic direction as in the discovery cohort. For these associations, we then performed a meta-analysis using a weighted *z* approach to combine association signals across the three cohorts. Only associations at $P < 5 \times 10^{-8}$ have been reported.

Targeted gene approach. For this analysis, we selected SNPs from four different categories: (i) SNPs associated with GWAS hits for immune and metabolic phenotypes, (ii) genes involved in innate microbial sensing, (iii) genes involved in adaptive microbial sensing (MHC locus) and (iv) genes involved in food metabolism and food preferences.

For GWAS-associated SNPs, we extracted immune response– and metabolism-related SNPs from the GWAS catalog and from recent Immunochip and Metabochip studies listed in Bonder *et al.*[83]. The selection of GWAS SNPs and their associated diseases is presented in **Supplementary Table 7**. For genes involved in innate microbial sensing, we included all 26 pathogen receptor and adaptor molecules described in Casals *et al.*[56]. We also added *MUC2* and *NLRP6* (their roles in bacterial sensing were recently described in Caballeor *et al.*[58]). We also included C-type lectin molecules for their role in pathogen recognition[84] and three viral recognition receptors, MDA5 (encoded by *IFIH1*), RIG1 (encoded by *RARRES3*) and MAVS[85]. For adaptive immunity, we included all the genes from the MHC locus. We performed imputation of amino acids and allelic variants of the MHC molecules, as described above. For food tolerance and food preferences, we included loci on the basis of their association in GWAS and other literature studies (**Supplementary Table 8**).

For candidate genes in innate and adaptive immunity and food-related genes, we defined a region of 250 kb around the genes and tested all SNPs in this region. We selected a total of 76,444 SNPs for the targeted gene approach, in which we adopted the same methods as for the GWAS analysis, but with a less stringent cutoff. Associations with bacteria, pathways and GO2000 terms were first identified at a *P*-value threshold of 5×10^{-4} in the discovery cohort. We then replicated these associations in the replication cohorts and selected associations that were replicated at $P < 0.01$ in the same allelic direction as in the discovery cohort. For these associations, we further performed a meta-analysis using a weighted *z* approach to combine association signals across the three cohorts. Only associations at $P < 5 \times 10^{-6}$ have been reported.

Statistical significance. In our analysis, we use a conservative three-stage approach to identify associations between the gut microbiome and host genetics. For genome-wide analysis, we assumed that 10^9 unrelated pairwise tests were performed between 1 million unrelated SNPs and approximately 1,000 unrelated microbial traits. In the first stage, we selected all associations at $P < 5 \times 10^{-5}$ in the discovery data set. Therefore, we expect 50,000 false positives ($5 \times 10^{-5} \times 10^9$ tests performed). In the second stage we used a replication threshold of 0.01. After the second stage, we expect to have 500 false positive results. In the third stage, we performed a meta-analysis using the weighted *z*-score approach[86] over both the discovery and replication cohorts, here using a more stringent threshold of $P < 5 \times 10^{-8}$.

Multiple *P*-value combinations, in the first and second stages, can reach the final desired meta *P*-value threshold (**Supplementary Table 4a**,**b**). For example, if the *P* value in the first stage is 5×10^{-5}, the *P* value in the second stage needs to be lower than 3×10^{-5} to meet the meta *P*-value threshold of 5×10^{-8}. In the case of the aforementioned *P*-value thresholds, the expected number of false positive results would be 1.5. The maximum number of false positive results in our three-stage approach is 5, depending on the combination of first- and second-stage *P* values (**Supplementary Table 4b**). As we found 42 associations, we expect an FDR of 12%.

For the targeted analysis, we report all associations that passed the meta *P*-value threshold of $P < 5 \times 10^{-6}$. Given our setup, explained above, and application of the chosen thresholds ($P < 5 \times 10^{-4}$ in the initial discovery, $P < 0.05$ in the first replication and the final $P < 5 \times 10^{-6}$), our sample sizes, the 85,405 targeted SNPs and the roughly 1,000 unrelated microbial traits, the chance of the association being sporadic is very minor (**Supplementary Table 4c**,**d**). Given that we found 32 independent loci at this level of significance, we expect our FDR to be <1%.

Analysis of the *LCT* locus. The genetic variant rs4988235 at the *LCT* locus is known to have a recessive effect on lactose intolerance. The homozygous GG genotype will result in lactase deficiency, and individuals with a GG genotype are susceptible to lactose intolerance, whereas individuals with the AA or AG genotype will have efficient LCT enzyme. Thus, we specifically assessed the recessive effect of the functional SNP rs4988235 in the *LCT* locus by dividing the individuals into an LCT-deficient group (GG genotype) and an LCT-efficient group (AA or AG genotype). Microbial abundance was compared between the two groups using the Spearman correlation test in the R 'base' package. Association was assessed in the discovery and replication cohorts, followed by a meta-analysis using the weighted sum of *z* method in R package 'metap' v.0.6-2.

Further, we specifically tested the association between bacteria and dairy product intake for the functional *LCT* SNP. To do this, we calculated the Spearman correlation between dairy product consumption and *Bifidobacterium* abundance in groups of LCT-deficient and LCT-efficient individuals separately. For the LifeLines-DEEP and MIBS cohorts, dairy intake was extracted from the food frequency questionnaire (units: g/day). For the 500FG cohort, milk intake was reported as glasses of milk per day. Comparison of correlations between groups of LCT-efficient and LCT-deficient individuals was performed using Fisher's *r*-to-*z* transformation with a one-tailed test in R package 'psych' v.1.5.8.

For the *LCT* associations described above, we selected only participants for whom we had genetic, microbiome and food frequency questionnaire data available (923 individuals from LifeLines-DEEP, 397 individuals from 500FG and 101 individuals from MIBS-CO).

Functional annotation of SNPs. For the associated SNPs, we further characterized their functions by looking into their association with complex traits and effect on gene expression using multiple eQTL data sets. Multitissue eQTL results from the GTEx Consortium[87] were used in the annotation, including sigmoid colon, transverse colon, small intestine, esophagus, gastroesophageal junction, mucosa and muscularis from the esophagus, and from the eQTL data set in a large number of blood samples derived from Zhernakova *et al.*[88].

67. Streppel, M.T. *et al.* Relative validity of the food frequency questionnaire used to assess dietary intake in the Leiden Longevity Study. *Nutr. J.* **12**, 75 (2013).
68. Siebelink, E., Geelen, A. & de Vries, J.H.M. Self-reported energy intake by FFQ compared with actual energy intake to maintain body weight in 516 adults. *Br. J. Nutr.* **106**, 274–281 (2011).
69. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46**, 818–825 (2014).
70. Shah, T.S. *et al.* optiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants. *Bioinformatics* **28**, 1598–1603 (2012).
71. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
72. Deelen, P. *et al.* Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Res. Notes* **7**, 901 (2014).
73. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
74. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 (Bethesda)* **1**, 457–470 (2011).
75. 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).
76. Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One* **8**, e64683 (2013).
77. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
78. Truong, D.T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
79. Dimmer, E.C. *et al.* The UniProt-GO Annotation database in 2011. *Nucleic Acids Res.* **40**, D565–D570 (2012).
80. Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **43**, D1049–D1056 (2015).
81. Vatanen, T. *et al.* Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. *Cell* **165**, 842–853 (2016).
82. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).

doi:10.1038/ng.3663

83. Bonder, M.J., Luijk, R., Zhernakova, D.V. & Moed, M. Disease variants alter transcription factor levels and methylation of their binding sites. Preprint at *bioRxiv* http://dx.doi.org/10.1101/033084 (2015).
84. Robinson, M.J., Sancho, D., Slack, E.C., LeibundGut-Landmann, S. & Reis e Sousa, C. Myeloid C-type lectins in innate immunity. *Nat. Immunol.* **7**, 1258–1265 (2006).
85. Reikine, S., Nguyen, J.B. & Modis, Y. Pattern recognition and signaling mechanisms of RIG-I and MDA5. *Front. Immunol.* **5**, 342 (2014).
86. Whitlock, M.C. Combining probability from independent tests: the weighted *Z*-method is superior to Fisher's approach. *J. Evol. Biol.* **18**, 1368–1373 (2005).
87. Ardlie, K.G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
88. Zhernakova, D.V. *et al.* Hypothesis-free identification of modulators of genetic risk factors. Preprint at *bioRxiv* http://dx.doi.org/10.1101/033217 (2015).