Individual variations in cardiovascular-disease-related protein levels are driven by genetics and gut microbiome

Daria V. Zhernakova1,2,10,12, Trang H. Le1,12, Alexander Kurilshikov1,12, Biljana Atanasovska1,3, Marc Jan Bonder1, Serena Sanna1, Anique Claringbould1, Urmo Võsa1, Patrick Deelen1,4, LifeLines cohort study5, BIOS consortium5, Lude Franke1,10, Rudolf A. de Boer6, Folkert Kuipers3,7, Mihai G. Netea6,9, Marten H. Hofker3,11, Cisca Wijmenga6,10,13, Alexandra Zhernakova1,13* and Jingyuan Fu1,3,13*

Despite a growing body of evidence, the role of the gut microbiome in cardiovascular diseases is still unclear. Here, we present a systems-genome-wide and metagenome-wide association study on plasma concentrations of 92 cardiovascular-disease-related proteins in the population cohort LifeLines-DEEP. We identified genetic components for 73 proteins and microbial associations for 41 proteins, of which 31 were associated to both. The genetic and microbial factors identified mostly exert additive effects and collectively explain up to 76.6% of inter-individual variation (17.5% on average). Genetics contribute most to concentrations of immune-related proteins, while the gut microbiome contributes most to proteins involved in metabolism and intestinal health. We found several host–microbe interactions that impact proteins involved in epithelial function, lipid metabolism, and central nervous system function. This study provides important evidence for a joint genetic and microbial effect in cardiovascular disease and provides directions for future applications in personalized medicine.

Advances in high-throughput deep-sequencing technology have revolutionized our understanding of the impact of the human genome and gut microbiome on human health. Genome-wide association studies (GWAS) and metagenome-wide association studies have provided evidence that the development of many complex diseases can be determined by the human genome, the microbiome, and their interactions. This has now been shown for cardiovascular diseases1, type 2 diabetes2, inflammatory bowel disease (IBD)3–4, and different types of cancer5. However, the impact of gut microbiome and genome–microbe interactions on molecular traits is largely unknown, which greatly limits our mechanistic understanding of microbial associations with complex diseases.

Circulating plasma proteins are often used as risk factors or biomarkers for various diseases. Understanding the impact of genetics, the gut microbiome, and their interactions on the inter-individual variation of circulating plasma proteins will provide deeper insights into host–microbe interactions in health and disease. Here, we present a systems genome and metagenome association analysis on 92 circulating plasma proteins in two subsets of the LifeLines Dutch population cohort, LifeLines-DEEP (n = 1,178) and LifeLines-DEEP2 (n = 86) (Methods and Fig. 1), for which we have genotype, metagenome, transcriptome, and detailed phenotype data6. These proteins were selected a priori based on their direct or indirect role in the development of cardiovascular diseases (CVDs) (Supplementary Table 1). However, most of the proteins have a broader impact on host health and are also found to be relevant to many other diseases (Supplementary Fig. 1). Thus, we believe that the knowledge gained here can be generalized to a wide spectrum of diseases and thereby impact other physiological processes and diseases.

Results

Associations to genetics. To estimate the effect of host genetics on protein levels, we first performed a local protein quantitative trait loci (cis-pQTL) analysis by testing SNPs located within 250 kb of the genes coding for the 92 proteins. This yielded 129 significant cis-pQTLs for 66 proteins at genome-wide false discovery rate (FDR) 0.05 level (Supplementary Table 2 and Supplementary Fig. 2). We then regressed out the cis-pQTL effects and conducted a trans-pQTL mapping in a genome-wide manner and then separately on disease- and trait-associated SNPs only, which together yielded 85...
Articles

NATURe GeNeTICS

VOL 50 | NOVEMBER 2018 | 1524–1532 | www.nature.com/naturegenetics

Fig. 1 Study analysis workflow. LD, linkage disequilibrium.

independent trans-pQTLs for 36 proteins (Supplementary Table 3 and Supplementary Fig. 3). Of these, 19 cis-pQTLs and 74 trans-pQTLs were associated with complex traits and diseases, including 10 cis- and 7 trans-regulated proteins known to be relevant for CVD (Supplementary Tables 2 and 3). In addition, we separately assessed associations to 422 putative CVD-associated SNPs and detected pQTL associations for 14 proteins (Supplementary Table 4 and Supplementary Note 1). These pQTLs could point to driver genes in CVD (Supplementary Note 2); for example, as can be seen in the pleotropic trans-pQTLs effect observed at the KLKB1 gene (Supplementary Fig. 4).

Next, we examined the power of our study by assessing the replication rate of previously reported pQTLs and trans-pQTLs. We found that only 64% of cis-pQTLs had at least one corresponding significant cis-eQTL, and 76% of these had the same allelic direction in blood from the same individuals or in other tissue types from the GTEx project. Our data also revealed novel pQTL associations including 36 cis-pQTLs for 25 proteins and 48 trans-pQTLs for 27 proteins (Supplementary Tables 2 and 3).

We found that only 64% of cis-pQTLs had at least one corresponding significant cis-eQTL, and 76% of these had the same allelic direction in blood from the same individuals or in other tissue types from the GTEx project. In contrast, none of the 85 trans-pQTLs were detectable at expression level, but this may be due to the power issue as the effect sizes of trans-eQTLs are known to be very modest. Despite this, our data do provide evidence that a large amount of trans-regulation can happen at translation- or protein-level; for example, through regulation of translational rate and protein secretion to blood, through post-translational modification, or through protein–protein interactions (PPIs), and these trans effects are not necessarily detectable at transcription level.

Associations to the gut microbiome. Next, we linked the variation in plasma proteins to the gut microbiome. We assessed the association of plasma protein concentration with various microbial features, including alpha diversity assessed by Shannon index, beta diversity assessed by Bray–Curtis distance, 355 bacterial taxa (including 164 bacterial species and 191 upper level taxa), and 438 bacterial MetaCyc pathways (Fig. 1, Methods, and Supplementary Tables 7–10). A total of 41 unique proteins were associated with at least one microbial feature at FDR 0.05 level: 2 proteins were associated to Shannon diversity index (Supplementary Table 7 and Supplementary Fig. 6), 6 proteins to beta-diversity (Supplementary Table 8 and Supplementary Fig. 6), 23 proteins to bacterial taxa (Supplementary Table 9), and 25 proteins to bacterial pathways (Supplementary Table 10). Although most of the microbial associations to proteins are novel, several proteins have previously been linked to the gut microbiome, including plasminogen activator inhibitor (PAI)-1 and urokinase plasminogen activator (uPA), epidermal growth factor receptor (EGFR), members of the paraoxonases family (PON)15,16, tumor necrosis factor receptor (TNF-R2), and insulin-like growth factor (IGF)17. For example, the gastric pathogen Helicobacter pylori stimulates expression of PAI and of uPA and its receptor (uPAR) in gastric epithelial cells18 and can also affect epithelial signaling pathways partially through activation of EGFR19.

In total, we identified host genetic and microbial features that underlie the inter-individual variation of 83 plasma proteins, of which 31 were found to be associated to both genetic and microbial factors (Supplementary Fig. 7). We further assessed to what extent the protein–microbiome associations are dependent on genetic effects by systematically comparing the pQTLs with previously reported microbiome QTLs (mbQTLs) (Supplementary Note 4). While we did not observe any overlap between pQTLs and mbQTLs at the genome-wide significance level, several pQTLs for epithelial cell adhesion molecule (Ep-CAM), PON3, PAI, and CHI3L1 showed suggestive associations with the corresponding microbial factors (Supplementary Table 11). After regressing out the effects of all significant pQTLs and mbQTLs for proteins and microbiome, respectively, the association strength between protein and microbiome remained highly comparable (Supplementary Fig. 8, Supplementary Note 4, and Supplementary Tables 12 and 13).

Protein variance explained by genetics and microbiome. We then quantified the proportion of explained variance for each protein by genetic and microbial factors separately and jointly (Methods). Cis-pQTL effects explained an average of 14.9% of the variation for 66 cis-regulated proteins (range 0.35–73.3%). Trans-pQTL effects explained an average of 9.5% of the variation for 36 trans-regulated proteins (range 0.7–51.8%). Compared to host genetic factors, the effect of microbial features was smaller, contributing on average 3.2% of the variation for 41 proteins (range 0.4–26.5%) (Fig. 2 and Supplementary Table 14). The effects of host genetic factors and
microbial features were mostly independent and showed additive effects (Supplementary Fig. 9 and Supplementary Note 5). They collectively explained an average of 17.5% of the variation, ranging from 1% of the variation in the concentration of tPA protein up to 76.6% of the variation in IL-6RA concentrations (Fig. 2). However, we also detected 21 significant genome–microbiome interactions for eight proteins (Supplementary Table 15 and Supplementary Note 5): 14 significant interactions for Ep-CAM and 1 interaction each for PON3, CNTN1, CHI3L1, TLT-2, PSP-D, CTSZ, and IL-6RA.

Ep-CAM as a mediator between host and microbiome. In contrast to most genes, which were more affected by genetics than by microbiome, plasma levels of proteins active in the intestinal epithelium appeared to be more linked to the gut microbiome, and these proteins include Ep-CAM and trefoil factor 3. The variation in Ep-CAM levels was mostly associated to microbial features (26.5%), with 7% of the variation explained by two trans-pQTL loci: the FUT2 locus at 17q11.2 (26.5%), with 7% of the variation explained by two trans-pQTLs (green), and microbiome (red).

In our study, FUT2 non-secretor status correlated with a lower expression of Ep-CAM ($r = -0.227$, $P = 2.3 \times 10^{-12}$) (Fig. 3a). Moreover, the FUT2 SNP is among the top trans-pQTL SNPs that showed suggestive associations to the gut microbiome (Supplementary Table 11). This is consistent with previously reported associations between FUT2 and microbial composition and function$^{14,15}$, suggesting an impact of host genetics on the gut microbiome. The strongest microbial association to the FUT2 SNP was observed for the abundance of the Blautia genus, which was also significantly correlated with Ep-CAM levels. We observed decreased levels of Blautia genus in non-secretors in our LifeLines-DEEP cohort ($r = -0.069$, $P = 0.036$) and replicated this in the independent 500FG cohort ($r = -0.18$, $P = 2.4 \times 10^{-4}$) (Fig. 3b), for which both genetic and microbiome data are available$^{26}$. In line with the genetic impact of FUT2 on both Ep-CAM concentration and Blautia abundance, higher Ep-CAM level was indeed associated with increased abundance of Blautia in the LifeLines-DEEP cohort ($r = 0.138$, $P = 9 \times 10^{-4}$, FDR $= 0.005$) (Fig. 3c). A partial correlation analysis suggested Ep-CAM as a likely mediator between FUT2 and Blautia (Fig. 3d), and mediation analysis estimated that 32.7% of the genetic effect of FUT2 on Blautia abundance is mediated by Ep-CAM level ($P_{\text{mediated}} = 0.0084$, $P_{\text{direct}} = 0.26$). Most species from Blautia genus are involved in the production of short-chain fatty acids$^{33,34}$. These metabolites, especially butyrate, have anti-inflammatory, antiproliferative, and antineoplastic properties and are implicated in protection against colorectal cancer$^{35}$.

Other producers of butyrate include the commensal bacteria Clostridia, which helps to maintain gut homeostasis$^{36}$. We also found that the plasma level of Ep-CAM was associated with many species of the Clostridiales order; for example, Clostridium bartletti;
Fig. 3 | Association of FUT2, Ep-CAM, and Blautia. On all plots, level of Ep-CAM and log-transformed abundance of Blautia genus are scaled to mean = 0/s.d. = 1. a, Association between plasma level of Ep-CAM and secretor/non-secretor status of ABH antigens encoded by a genetic variant rs601338 within FUT2 gene. b, Association between Blautia and secretor/non-secretor status in two independent cohorts. c, Association between plasma level of Ep-CAM and abundance of Blautia. Each circle represents an individual sample: secretor individuals (blue) and non-secretors (red). d, Partial correlation analysis among FUT2, Ep-CAM, and Blautia. Partial correlation coefficient and P value are labeled at each edge.

(r = −0.30, P = 3.8 × 10−25, FDR = 4.7 × 10−21) (Supplementary Table 9). Interestingly, we also detected suggestive interactions between *Clostridium* spp. and the FUT2 SNPs (Supplementary Table 15). Furthermore, of the 97 pathways significantly associated with Ep-CAM (Supplementary Table 10), higher plasma level of Ep-CAM was associated to a lower biosynthesis of preQ0 (r = −0.29, P = 4.3 × 10−25, FDR = 1.7 × 10−28), a newly discovered metabolite with anticancer activity52. However, FUT2 secretor/non-secretor status did not appear to have any effect on the association between Ep-CAM and bacterial biosynthesis of anticancer compounds. Based on the methodology behind the Mendelian randomization approach, the lack of genetic association of FUT2 with pre-Q0 biosynthesis does not support a causal impact of Ep-CAM on pre-Q0 biosynthesis. This may indicate that pre-Q0 biosynthesis can affect Ep-CAM, in line with the anticancer properties of pre-Q0.

PON3 links lipid oxidation to the gut microbiome. Proteins that function in the same pathways can have direct PPIs and/or form a co-abundant relationship. To visualize these relations, we combined Bayesian network analysis on protein co-abundance with experimentally validated PPIs to produce a joined network (Fig. 4). A group of proteins involved in lipid/glucose metabolism that were mainly associated to the gut microbiome also formed a subcluster in the Bayesian network analysis. These proteins—PAI, PON3, uPA, insulin-like growth factor binding protein 1 (IGFBP-1) and IGFBP-2, trefoil factor 3, low-density lipoprotein (LDL)-receptor, and RARRES2 (Fig. 4)—were also highly correlated to various risk factors for CVD measured in the LifeLines-DEEP cohort (Supplementary Table 16). Body mass index (BMI), for instance, was highly correlated to chemerin (RARRES2) (r = 0.37, P = 1.8 × 10−31), which regulates adipogenesis and adipocyte metabolism48, and IGFBP-1, which plays an important role in glucose and insulin metabolism49 (r = −0.44, P = 6.9 × 10−14). Triglyceride concentration was mostly associated to IGFBP-2 (r = −0.35, P = 2.0 × 10−35), while high-density lipoprotein (HDL) cholesterol level was mostly positively associated to PON3 (r = 0.40, P = 4.2 × 10−39). Insulin was mostly associated to insulin binding proteins IGFBP-1 (r = −0.40, P = 4.35 × 10−41) and IGFBP-2 (r = −0.39, P = 1.69 × 10−38).

Among these proteins, the gut microbiome showed the strongest association to blood levels of PON3 and PAI (Fig. 4). PON3 associates with HDL in the blood stream and inhibits LDL oxidation46. In our data, 16% of the variation of PON3 can be explained: 7.8% by cis-genetic effects and 8.2% by gut microbiome. At diversity level, PON3 was the top protein associated to Shannon diversity (r = 0.13, P = 2.2 × 10−5, FDR = 0.002) and had the second highest association to beta-diversity (r = 0.004, P = 9.9 × 10−5, FDR = 0.009) (Supplementary Fig. 6). At microbiome composition level, 14 unique taxa were significantly associated with PON3 at FDR < 0.05 (Supplementary Table 9). Many of these species are known to be associated with metabolic traits and IBD, including the BMI- and lipid-lowering species *Akkermansia muciniphila*52,53 (r = 0.12, P = 7.8 × 10−5, FDR = 0.03), the butyrate-producing bacteria *Roseburia hominis*44 (r = 0.13, P = 2.6 × 10−5, FDR = 0.01), and the hydrogen-removing archaeon *Methanobrevibacter smithii*45 (r = 0.18, P = 2.1 × 10−4, FDR = 3.3 × 10−3) that has also been associated with BMI in the LifeLines-Deep cohort46. At functional profile level, 63 pathways were associated to PON3 (Supplementary Table 10) and 97% of these can be significantly linked with more than one PON3-associated species with a consistent direction (Supplementary Fig. 11a). Among them, one significant association was observed for L-methionine biosynthesis (HOMOSER-METSYN-PWY, r = −0.17, P = 3.4 × 10−4, FDR = 7.2 × 10−3). L-methionine is an essential amino acid that must be obtained through diet and can be further metabolized to...
or from L-homocysteine through the S-adenosyl-L-methionine (SAM) cycle. L-homocysteine/SAM is considered to be a risk factor for CVD, although this has been debated in recent years. Another significant association we found was for a bacterial pathway involved in thiamine (vitamin B1) biosynthesis (PWY-7357, \( r = -0.16, P = 2.3 \times 10^{-7}, \text{FDR} = 2.6 \times 10^{-4} \)). Notably, a suggestive interaction was also observed between the thiamine biosynthesis pathway and the cis-pQTL SNP (rs10953142) of PON3 (analysis of variance (ANOVA) \( P = 0.003 \)) (Supplementary Table 15).

Serum level of PAI was another protein showing a strong association with the gut microbiome: 6.9% of the variation in PAI could be explained by the gut microbiome and only 0.74% by genetics. PAI is a key regulator in fibrinolysis, the process of breaking down fibrin and maintaining vessel patency. PAI also plays a role in inflammation, and its expression can be induced by oxidized LDL.

Consistent with PON3 preventing LDL oxidation, our data showed negative association between PON3 and PAI (\( r = -0.33, P < 2.2 \times 10^{-16} \)), with both being associated to metabolic traits such as BMI (Supplementary Table 16). We detected significant associations of PAI with 6 taxa and 38 pathways, 4 and 14 of which are shared with PON3, respectively (Supplementary Fig. 11b). Notably, the bacterial species and pathways shared between PON3 and PAI were likely to affect proteins through their impact on metabolism. After regressing out BMI, most shared associations vanished (Supplementary Fig. 11b). This provides a functional link between the gut microbiome, lipid metabolism, and obesity-related inflammation. However,
Genetic–microbiome interaction in the gut–brain axis. The central nervous system (CNS) emerging as an important factor in control of BMI and susceptibility to obesity and related morbidity\(^1\). In addition to proteins functioning in the intestine and metabolism, we found some proteins associated with the gut microbiome are active in neural systems. Among these, the most notable were contactin 1 (CNTN1) and NOTCH3, both proteins involved in the NOTCH signaling pathways that play important roles in the development of the neural system, and EGFR, an epidermal growth factor (Fig. 4). CNTN1, in particular, was associated to both genetics and the microbiome. Our data showed that 6.6% of the variation in CNTN1 could be explained by two independent cis-pQTL SNPs, rs12811939 \((r=0.21, P=1.14\times10^{-11})\) and rs17541621 \((r=0.18, P=7.47\times10^{-10})\) (Supplementary Table 2), and 6.1% of the variation was explained by a trans-pQTL SNP, rs61261356, located at the intronic region of the TMEM8A gene \((r=0.26, P=4.25\times10^{-22})\) (Fig. 5a). Moreover, 3% of the variation of CNTN1 was explained by the gut microbiome (Fig. 2), with a positive association with the bacterial pathway of chorismate biosynthesis \((r=0.13, P=2.5\times10^{-3}, \text{FDR}=8.6\times10^{-3}; \text{Fig. 5b})\). Chorismate is a precursor for tryptophan, which can be further converted to serotonin. Ninety percent of tryptophan is synthesized in the gastrointestinal tract and it is considered a key signaling metabolite between the gut and the CNS\(^4\). Interestingly, we also found a nominally significant interaction between chorismate biosynthesis and the trans-pQTL SNP at the TMEM8A locus \((P=0.04)\). The association between chorismate biosynthesis and CNTN1 is weaker and non-significant in individuals homozygous for the rs61261356 C/C \((r=0.029)\) and T/T \((r=0.068)\) genotypes as compared to T/C individuals \((r=0.20, P=1.6\times10^{-10}; \text{Fig. 5c})\).

To pinpoint the putative candidate gene at rs61261356, we checked blood eQTL data from the same samples and detected a cis-eQTL effect of this SNP on TMEM8A \((P=2.75\times10^{-28})\) (Supplementary Table 3). The role of TMEM8A is largely unknown. However, it is predicted to act as an adhesion protein that keeps T cells in resting state\(^5\). To search for other genes affected by this SNP, we looked for it in a larger RNA-sequencing-based eQTL map\(^\text{a}\). The gene AXIN1 encodes a scaffold protein in the β-catenin destruction complex that, if disrupted, contributes to pathogenesis of various human diseases including colorectal carcinogenesis and IBD\(^7\). AXIN1 shows a relatively abundant expression in brain and colon...
and plays a role in intestinal inflammation. A previous study demonstrated that Salmonella infection promotes the degradation and plasma sequestration of AXIN1, leading to bacterial invasiveness and inflammatory responses. Moreover, evidence also supports vitamin D and the vitamin D receptor as important regulators of IBD and colon cancer and they also play a role in maintenance of physiological level of AXIN1. This highlights the potential link between AXIN1 and the gut microbiome.

**Discussion**

CVDs are complex multifactorial diseases, and the processes underlying them include metabolism and inflammation, which relate to many other organs, including the liver, intestine, and CNS. CVD can therefore be considered a ‘systemic disease’ rather than a single-organ disease. In this study, we performed a systematic genome and metagenome analysis on plasma levels of 92 CVD-related proteins in 1,264 individuals. To our knowledge, this is the first effort to date to evaluate the roles of host genetics and microbiome composition, and their interaction, in regulating plasma protein concentrations in relation to CVD. We not only quantify what percentage of variation in plasma proteins can be collectively explained by genetics and the gut microbiome, but also provide insights into the mechanisms that underlie host–microbe interactions that relate to CVD. Our data show that genetic factors contribute most strongly to immune-related proteins, while the gut microbiome has a stronger relation to proteins involved in metabolism and obesity-related inflammation. We further demonstrate the interactions between the gut microbiome and host genetics, and these suggest novel microbial links to CVD via intestinal health, lipid oxidation, and the CNS.

Ep-CAM is an intestinal epithelial membrane protein that has a role in intercellular adhesion and immune defense against mucosal infections. Intestinal permeability has been suggested as a target for prevention and therapy in various diseases, including CVD. In line with this, Ep-CAM-deficient mice exhibited increased intestinal permeability and decreased ion transport, which may contribute to CVD susceptibility risk. In this respect, our study provides evidence for the interplay between genetics, gut microbiome, and the plasma level of Ep-CAM. We found that 26.5% of the variation in Ep-CAM levels can be explained by microbial factors, while 7% can be explained by two trans-pQTL loci, one of them in the FUT2 locus that is also associated to plasma vitamin B12 level, infectious diseases, and Crohn's disease. We showed that Ep-CAM can mediate the impact of FUT2 on the butyrate-producing bacteria Blautia. Our study also reported that many microbial associations with Ep-CAM were independent of FUT2, including the top association observed for bacterial biosynthesis of an anticancer compound: a higher plasma level of Ep-CAM was associated with lower level of the bacterially produced anticancer metabolite preQ0.

The CNS has an important regulatory role in the development of obesity, which, in turn, is a well-known risk factor for CVD. The CNTN1 gene encodes a key CNS regulator with a still-unclear role in the etiology of obesity and CVD. However, decreased plasma concentrations of CNTN1 have been proposed as a protein marker for new-onset CVD. A mouse experiment has also shown that deletion of the innate immunity antibacterial gene Nod2 induces dysbiosis and increases susceptibility to diet-induced obesity and metabolic dysfunction, as well as leading to an overall decrease in the expression of CNS genes including Cntn1. Our study demonstrates an association between a bacterial pathway related to metabolism of tryptophan and CNTN1 levels. Tryptophan is a well-established signaling metabolite in the gut–brain axis that is also related to appetite. Consistent with this, our study suggests a link between the gut microbiome, CNS, and obesity via tryptophan metabolism. Oxidation of LDL contributes to the risk of CVD. LDL oxidation can be inhibited by paraoxonases that are bound to HDL and exert antioxidant properties. The PON family contains three members: PON1, -2, and -3. The role of PON1 in atherosclerosis has been well-established and it was recently identified as part of the cardiac secretome in post-myocardial infarction heart failure. The role of PON3 is less clear, but PON3−/− knockout mice exhibited a reduced arterial lesion size comparable to that seen in PON1−/− mice, although this was probably occurring via a different mechanism. PON3 is one of the top proteins associated to microbial diversity, taxa, and pathways. Interestingly, most of its associations are independent of BMI, and one top associated bacterial pathway involves the biosynthesis of vitamin B1, which also exhibits antioxidant activity. Although the causal role of PON3 in gut microbiome and susceptibility risk for CVD still needs further evaluation, our data suggest a novel link between the antioxidative role of PON3 and the gut microbiome.

Altogether, our results demonstrate complex genetic–microbiome interplay in the regulation of circulating proteins that modulate various biological processes and demonstrate that these effects can be seen in many different organs and tissues. This study provides conceptual advances that lay important groundwork for future applications in personalized medicine, which will have to take into account both the genome and metagenome.

**Online content**

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability, and associated accession codes are available at https://doi.org/10.1038/s41588-018-0224-7.

Received: 9 November 2017; Accepted: 2 August 2018; Published online: 24 September 2018

**References**

1. Tang, W. H. W. & Hazen, S. L. The contributory role of gut microbiota in cardiovascular disease. *J. Clin. Invest.* 124, 4204–4211 (2014).
2. Qin, J. et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490, 55–60 (2012).
3. Knights, D. et al. Complex host genetics influence the microbiome in inflammatory bowel disease. *Genome Med.* 6, 107 (2014).
4. Imhann, F. et al. Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease. *Gut* 67, 108–119 (2016).
5. Nakatsu, G. et al. Gut mucosal microbiome across stages of colorectal carcinogenesis. *Nat. Commun.* 6, 8727 (2015).
6. Tiggeslaar, E. F. P. 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).
7. Nelson, C. P. et al. Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat. Genet.* 49, 1385–1391 (2017).
8. Sun, B. B. et al. Genomic atlas of the human plasma proteome. *Nature* 558, 73–79 (2018).
9. Suhre, K. et al. Genetic effects on gene expression across human tissues. *Nature* 558, 204–213 (2017).
10. Sun, W. et al. Common genetic polymorphisms influence blood biomarker measurements in COPD. *PLoS Genet.* 12, e1006011 (2016).
11. Folkerst, L. et al. Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. *PLoS Genet.* 13, 1–21 (2017).
12. Aguet, F. et al. Genetic effects on gene expression across human tissues. *Nature* 558, 204–213 (2017).
13. Kenny, S. et al. Increased expression of the urokinase plasminogen activator system by Helicobacter pylori in gastric epithelial cells. *Am. J. Physiol. Gastrointest. Liver Physiol.* 295, G431–G441 (2008).
44. Tamanoi-Shacoori, Z. et al. Roseburia spp.: a marker of health? *Future Microbiol.* 12, 157–170 (2017).
45. Gottlieb, K., Wacher, V., Sliman, J. & Pimentel, M. Review article: inhibition of methanogenic archaea by statins as a targeted management strategy for constipation and related disorders. *Aliment. Pharmacol. Ther.* 43, 197–212 (2016).
46. Mbaekwa, C. A. et al. Gut colonization with Methanobrevibacter smithii is associated with childhood weight development. *Obesity* 23, 2508–2516 (2015).
47. McCully, K. S. Homocysteine and vascular disease. *Nat. Med.* 2, 386–396 (1996).
48. Wierzbicki, A. S. Homocysteine and cardiovascular disease: a review of the evidence. *Diabetes Vasc. Dis. Res.* 4, 145–150 (2007).
49. Chi, Y. S., Bong, C. K., Hye, K. H. & Hyun, S. L. Oxidized LDL activates Pla2g1a transcription through autocrine activation of TGF-beta signaling in mesangial cells. *Kidney Int.* 67, 1743–1752 (2005).
50. Hong, H. K., Song, C. Y., Kim, B. C. & Lee, H. S. ERK contributes to the effects of Smad signaling on oxidized LDL-induced PAI-1 expression in human mesangial cells. *Transl. Res.* 148, 171–179 (2006).
51. Pricoupart, L. P. et al. The three-gene paraoxonase family: physiologic roles, actions and regulation. *Atherosclerosis* 214, 20–36 (2011).
52. Rothem, L. et al. Paraoxonases are associated with intestinal inflammatory diseases and intracelullarly localized to the endoplasmic reticulum. *Free Radic. Biol. Med.* 43, 730–739 (2007).
53. Locke, A. E. et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature* 518, 197–206 (2015).
54. Ruddick, J. P. et al. Tryptophan metabolism in the central nervous system: medical implications. *Expert Rev. Mol. Med.* 8, 1–27 (2006).
55. Philonenko, E. S. et al. TMEM8—A non-globin gene entrapped in the globin web. *Nucleic Acids Res.* 37, 7394–7406 (2009).
56. Zhernakova, D. V. et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* 49, 139–145 (2016).
57. Jin, D. et al. Vitamin D receptor is a novel transcriptional regulator for Axin1. *J. Steroid Biochem. Mol. Biol.* 165, 430–437 (2017).
58. Zhang, Y. et al. Axin1 prevents salmonella invasiveness and inflammatory response in intestinal epithelial cells. *PLoS One* 7, e34942 (2012).
59. Bischoff, S. C. et al. Intestinal permeability—A new target for disease prevention and therapy. *BMJ Gastroenterol.* 14, 189 (2014).
60. Kozan, P. A. et al. Mutation of EpCAM leads to intestinal barrier and ion transport dysfunction. *J. Mol. Med.* 93, 535–545 (2015).
61. Yin, X. et al. Protein biomarkers of new-onset cardiovascular disease: prospective study from the systems approach to biomarker research in cardiovascular disease initiative. *Arterioscler. Thromb. Vasc. Biol.* 34, 939–945 (2014).
62. Rodriguez-Nunez, I. et al. Nod2 and Nod2-regulated microbiota protect BALB/c mice from diet-induced obesity and metabolic dysfunction. *Sci. Rep.* 7, 548–553 (2017).
63. Soran, H., Schofield, J. D. & Durrington, P. N. Antioxidant properties of Pseudomonas aeruginosa 172205. *Appl. Environ. Microbiol.* 76, 2693–2699 (2010).

**Acknowledgements**

We dedicate this paper to the memory of Marten H. Hofker, who was a key person in developing the concept of this study. We thank participants and staff of the LifeLines-DEEP cohort for their collaboration. The study was approved by the UMCg review board, ref. M12.113965. We thank J. Dekens, M. Platteel, A. Maatman, and J. Arends for management and technical support and Jackie Senior and Kate Mc Irtyre for English editing. M.C. thank A. Vila and B. Wu for helpful discussions. This project was funded by the Netherlands Heart Foundation (IN-CONTROL CVON grant 2012-03 to M.H.H., M.G.N., E.K., A.Z.F., and J.F. and CVON-DOSIS grant 2014-40 to R.A.D.B.); by Top Institute Food and Nutrition, Wageningen, The Netherlands (TIEN GH01 to C.W.); by the Netherlands Organization for Scientific Research (NWO) (NWO-VIDI 864.13.013 to F.J.E.; NWO VIDI 917.13.350 to R.A.D.B., NWO VIDI 016.178.056 to A.Z.F., NWO-VIDI 917.14.374 to L.F.N. NWO Spinoza Prize SPI 9421 to M.G.N., NWO Spinoza Prize SPI 942–26 to C.W., and NWO Gravitation Netherlands Organ-on-Chip Initiative (014.003.001) to C.W.); and by the European Research Council (ERC) (FP7/2007-2013/ERC Advanced Grant Agreement 2012-322698 to C.W., ERC Consolidator Grant 310372 to M.G.N., ERC Starting Grant 715772 to A.Z.F., and ERC Starting Grant 637640 to L.F.N.), by the Stiftung Kristian Gerhard Jebsen (to C.W.); and by the RoG Investment Agenda Grant Personalized Health to C.W. A.Z. also holds a Rosalind Franklin Fellowship from the University of Groningen. D.V.Z. was supported by St. Petersburg State University (Genome Russia Grant no. 1.52.1647.2016). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
Author contributions
M.H.H., C.W., A.Z., and J.F. conceptualized the study. D.V.Z., T.L., A.K., M.J.B., A.Z., and J.F. designed the methodology. D.V.Z., T.L., M.J.B., and J.F. were responsible for the software. D.V.Z., T.L., A.K., B.A., M.J.B., A.C., U.V., P.D., A.Z., and J.F. carried out the formal analysis. D.V.Z., T.L., A.K., A.Z., and J.F. wrote the original draft of the manuscript. B.A., M.J.B., S.S., A.C., U.V., P.D., I.F., R.A.B., F.K., M.G.N., C.W., A.Z., and J.F. reviewed and edited the manuscript. D.V.Z., T.L., A.K., and J.F. were responsible for visualization. A.Z. and J.F. supervised the project. M.H.H., A.Z., and J.F. were responsible for project administration. I.F., R.A.D.B., F.K., M.G.N., M.H.H., C.W., A.Z., and J.F. were responsible for funding acquisition.

Competing interests
Dr. de Boer has received research grants and/or fees from AstraZeneca, Abbott, Bristol-Myers Squibb, Novartis, Roche, Trevena, and ThermoFisher GmbH.

Dr. de Boer is a minority shareholder of scPharmaceuticals, Inc. Dr. de Boer received personal fees from MandalMed Inc, Novartis, and Servier. All other authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0224-7.
Reprints and permissions information is available at www.nature.com/reprints.
Correspondence and requests for materials should be addressed to A.Z. or J.F.
Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Methods

Study cohorts. The LifeLines cohort is a large prospective cohort study from the north of the Netherlands. This study includes two subcohorts of LifeLines: LifeLines-DEEPI (LLL, n = 1,300) and LifeLines-DEEPI2 (LLD2, n = 119). The cohort comprises 58% he men and 42% males, the mean age is 45.04 years (± SD = 13.60), the mean BMI is 25.26 (SD = 4.18), and 12% of participants are obese (BMI >30). Common diseases within the cohort include high blood pressure (19%), anemia (15%), migraine (20%), irritable bowel syndrome (10%), and asthma (10%). The mean Framingham Risk Score for CVD is 8.6 for men and 5.7 for women, and 20% of individuals are current smokers. A detailed cohort description can be found in refs. 1 and 2.

For both of these subcohorts multiple levels of omics data are available, including genotypes, gene expression, microbiome, and proteomics data.

Data generation and preprocessing. Profiling of proteomics data. Blood samples of study participants were collected and frozen at −80 °C before proteomics measurement. Protein levels from EDTA plasma samples were measured in 1,447 participants in the LLD (n = 1,332) and LLD2 (n = 115) cohorts and determined using Olink Proseek Multiplex CVD III panel (Olink), a multiplex immunoassay for high-throughput detection of protein biomarkers in liquid samples31. This panel includes 92 proteins that are either established biomarkers or exploratory proteins with a high potential as new cardiovascular disease biomarkers. More detailed information can be found at the Olink website (see URLs section).

We normalized the protein data by applying a z-score transformation. To further identify confounding factors, we first calculated correlations of protein levels against phenotypes and detected significant correlations for age, sex, smoking status, oral contraceptive usage, and blood cell counts (basophils, eosinophils, erythrocytes, granulocytes, lymphocytes, monocytes, and thrombocytes). For subsequent analyses for each protein level, we considered adjusted protein levels residuals obtained from a linear regression model that included all the confounding factors as covariates. We further investigated for the presence of hidden confounders by principal component analysis on adjusted protein levels.

We found that the first ten principal components were associated with genetic variants and therefore did not adjust for any principal components in subsequent analyses. Finally, we estimated the effect of sample plate batches on protein levels, and detected no plate batch effect.

Metagenomic data. Fecal samples were collected within two weeks of blood sample collection, and microbial DNA isolation in LLD was performed as described earlier32. The metagenomics analysis in LLD2 samples followed the same procedure, except that fecal samples were not frozen but stored in RNAlater. We generated high-quality metagenomics sequencing data for 1,135 LLD samples and 119 LLD2 samples, with at least 15 million reads per sample.

The relative abundance of gut microbial taxonomic units was determined using metagenomic phylogenetic analysis (MetaPhlan 2.2)33. MetaPhlan 2.2 reported 1,772 microbial taxa in our data. We then filtered out low-abundance and rare taxa to only include taxa that accounted for at least 0.01% of total microbial composition and that present in more than 10% of samples in both LLD and LLD2, separately. This yielded a confined list of 355 taxa that accounted for 99.1% of microbial composition, including 164 microbial species and 191 taxa from upper levels. For all subsequent analyses, redundant taxa that provide the same information were removed (taxon–taxon correlation r > 0.99), leaving 267 unique taxa representatives.

Relative abundances of metabolic pathways were determined using the HUMAnN2 pipeline (see URLs section). Human genes were first removed and HUMAnN2 reported the abundances of 5,379,353 gene families from the UniProt Reference Clusters (UniRef50, see URLs section), which were further mapped to 773 microbial pathways from the MetaCyc metabolic pathway database (see URLs section). Rare pathways (present in fewer than 50% of the samples) microbial pathways were filtered out, leaving 536 out of the original 773 microbial pathways. These 536 pathways retained 98.9% of the original functional composition. Redundant pathways (pathway–pathway correlation r >0.99) were then removed, leaving 438 unique pathways for subsequent analysis.

Both taxa and pathway data sets were log-transformed for subsequent analysis.

Expression data. Whole-blood RNA-sequencing data were generated as described previously34. In brief, paired-end reads were aligned to the human genome using STAR aligner35, allowing for no more than seven mismatches. Gene expression estimation was performed using HTSeq-count36 on uniquely mapped reads. Gene level expression data were normalized by the trimmed mean of M-values (TMM) method37, followed by log, transformation, probe centering, and sample z-transformation. We regressed out the effects of age, sex, and blood cell counts (basophils, eosinophils, erythrocytes, granulocytes, lymphocytes, monocytes, and thrombocytes) from expression data using a linear regression model. Both proteomic and expression data were available for 1,293 samples.

Genotype data. Microarray genotype data for the LLD cohort were generated using CytoSNP and ImmunoChip assays as previously described38. Quality control checks on the LLD cohort were performed using the Haploview Reference Consortium (v1.0) preparation checking tool (v4.2.3) (see URLs section). We then uploaded the resulting VCF files to the Michigan Imputation Server39. Phasing and imputation were performed using the option SHAPEIT for phasing, population EUR, and the script Quality Control and Imputation; for all steps we used version R1 as reference40.

Whole-genome-sequencing-based genotypes were generated for LLD2 as described previously41. As the LLD2 participants were trios, we excluded first-degree relatives (offspring) for this study.

We further excluded SNPs that had imputation quality r² < 0.5, failed the Hardy–Weinberg equilibrium test (P < 1 × 10⁻⁸), or had a call rate < 95% or a minor allele frequency < 5%. In this way, we obtained genotype data for >7 million SNPs (Genome Build hg19) for 1,349 unrelated individuals: 1,262 from LLD and 87 from LLD2.

Statistical analysis. QTL mapping. Using a previously described pipeline42, we performed QTL mapping by calculating Spearman rank correlation in LLD (1,178 samples with both genotype and protein data) and LLD2 (86 samples with both genotype and proteome data) sets and combining the results using the weighted z-score method. To correct for multiple testing, we permuted genotype labels 250 times to create a null distribution that was then used to control FDR at 0.05.

cis-pQTL mapping was done by testing SNP–gene pairs located within a ± 250 kb window. Proteins (corresponding Uniprot identifiers) were converted to genes and genomic coordinates using Ensembl v75. One of the proteins (PECAM-1) mapped to a known genetic region and was thus removed from cis-pQTL mapping, but was used for trans-pQTL mapping. To extract all independent cis-pQTLs, we performed a step-wise conditional analysis for each protein. Specifically, we first performed a primary cis-pQTL mapping then regressed out the top SNP cis effect and ran cis-pQTL mapping again. This procedure was repeated stepwise until no more pQTLs could be detected at FDR < 0.05.

trans-pQTL mapping was run on protein data after regressing out all independent cis-pQTLs. To reduce the multiple testing burden, we first ran trans-pQTL mapping only on SNPs previously associated to complex traits and diseases. This SNP list contained 10,562 SNPs and was obtained as follows. Genetic risk factors were downloaded from three public repositories, the EBI GWAS Catalogue (downloaded 21.11.2016), the NIH GWAS Catalogue, and Immunobase (accessed 26.04.2016), with an applied significance threshold P < 5 × 10⁻⁶. Additionally, we added 2,706 genome-wide significant GWAS SNPs from a recent blood trait GWAS43. SNP coordinates were lifted to hg19 using the liftOver command from the trackracer v1.34.1 R package44 and subsequently standardized to match the GIANT 1000G p10 reference panel. After running a trans-pQTL mapping on disease-associated genetic risk factors, we also ran a genome-wide trans-pQTL mapping following the same procedure. A total of 195 trans-pQTLs were identified by the disease-associated analysis and 32 trans-pQTLs by the genome-wide analysis. We then combined these two lists and pruned by removing trans-pQTL SNPs in linkage disequilibrium with r² > 0.8. This resulted in 86 independent trans-pQTLs for 36 proteins.

We also compared genetic effects on protein and mRNA expression levels. To do this, we specifically tested all significant independent SNP–gene combinations, reported all eQTLs with FDR < 0.05, and compared the effect size and direction. We also used a large-scale whole-blood eQTL published previously45 and eQTL data of various tissue types from the GTEx v7 release46.

Replication of published pQTLs. To replicate previously reported pQTLs, we considered all pQTL studies with reasonable sample sizes (n >1,000) published in the 2 years before November of 2017 (see refs. 8–11). We first extracted pQTLs for the 92 proteins under study and tested all reported pQTLs in our data set. If a pQTL was significant at FDR 0.05 level and had the same allelic direction, we reported it as replicated. For each pQTL, we calculated the power for replication at significance level 0.009 (the least significantly replicated pQTL), estimating non-centrality parameters of chi-squared distributions reconstructed from P values and sample sizes reported in the aforementioned studies.

pQTL effects for variants associated with coronary artery disease (CAD). We then assessed the genetic effects of SNPs associated with CAD reported by Nelson et al.47. The paper reported 422 unique CAD-associated SNPs, including 73 SNPs for 72 established loci and 366 SNPs for 304 suggestive loci. In addition to single SNP association, we also tested the correlation between the combined CAD genetic risk score with protein levels using Spearman rank correlation analysis. To calculate the genetic risk score, we multiplied risk allele counts (for imputed genotypes we converted dosages to genotypes) by log odds ratios provided in the paper and summed for all reported SNPs. The FDR was controlled using the Benjamini–Hochberg approach.

Microbial association analysis. Association with alpha-diversity. Alpha diversity is the diversity of a community's taxonomic composition, in this case the intra-individual gut microbiome48. We calculated alpha diversity using the Shannon diversity index, using the function diversity() from the R package vegan (see URLs section), based on the abundances of 164 species that passed filtering in the LLD.
cohort. We used Spearman correlation to test associations. FDR was controlled at 0.05 using Bonferroni correction15.

Association with beta-diversity. Beta diversity is the degree of differentiation of taxonomic communities between individuals (that is, inter-individual gut microbiomes) and is usually defined by a between-community dissimilarity matrix. We calculated beta diversity using Bray–Curtis dissimilarity, using the function vegdist() from the R package vegan (see URL section).

To quantify the variation in Bray–Curtis dissimilarity explained by an individual protein, permutation multivariate ANOVA was carried out with the function adonis() from the R package vegan using 10,000 permutations, and FDR was controlled at 0.05 using Bonferroni correction15. Permutational multivariate ANOVA is a non-parametric test that allows us to perform association between vector data (the abundance of a specific protein of n subjects) and matrix data (pairwise x n distance matrix). This test partitions the variations across a multivariate data cloud in response to one or more factors and uses permutations to determine P value.

Association analysis with taxon and bacterial pathways. Pathway and taxon data sets were corrected for factors that can significantly alter the microbiome composition, including age, sex, race, dietary habits, antibiotic usage, irritable bowel syndrome, and diabetes. The cohort has <1% antibiotics users and 10% irritable bowel syndrome patients. Diarrhea was assessed by Bristol stool type. To correct for these factors, we first removed all antibiotic users and further adjusted other factors using a linear regression model.

To explore the association between 92 circulating proteins and microbiome features (267 unique taxa and 438 unique pathways), we performed pairwise Spearman rank correlation tests. Meta-analysis across the LLD and LLD2 cohorts was conducted using the weighted z-score method. Both Benjamini–Hochberg16 and Bonferroni-corrected P values are reported. This association analysis was performed on 999 samples from LLD and LLD2.

Variance explained by genetics and microbiome. To estimate the variance explained by genetic and microbial features, we confined our analysis to 926 participants in the LLD cohort who have microbiome, proteomics, and genetics data available. First, all significant SNPs from cis- and trans-pQTL analyses were chosen to represent genetic influence and their separate contributions to protein variance explained (\(V_{g}\) and \(V_{m}\)) were estimated. Together, they make up the variance explained by genetics (\(V_{g} = V_{m} + V_{g}\)). We also calculated the combined variance explained by genetics by putting all cis- and trans-pQTLs in one model. The difference between the R2 of the latter model and \(V_{g}\) is negligible (not shown), which confirms the expected independent contribution of cis- and trans-pQTLs. Unlike the genetic data, the microbiome data have a high degree of collinearity. We therefore first performed a feature selection procedure to extract only those microbial features that independently contributed to the variance of each protein. For this, we used a Lasso shrinkage model17 that included all microbial pathways and species significantly associated with the respective protein, together with the Shannon diversity index and the first five principal components of Bray–Curtis dissimilarity as predictors. The independent and most dominant microbial features were selected automatically by \(l_{1}\) norm regularized quantile regression (the coefficients for non-informative features were automatically shrunk to 0). A lambda regularization parameter was selected using tenfold cross-validation for each protein independently, choosing from 100 lambdas. The variation of each protein explained by selected microbial features \(V_{m}\) was then estimated using a linear model.

The total variance explained by genetics and microbiome \(V\) was calculated using a combined regression model that includes predictor-dependent shrinkage. No penalty was put on genetic predictors since we assumed genetic control to be primary and indispensable, while microbiome features went through automatic Lasso selection as described above.

Estimate additive or confounding effect between genetics and microbiome. As host genetics can affect the gut microbiome, we further estimated how likely it was that the combined genetic and microbial effect on proteins deviates from an additive model. To do this, we compared the total variation explained by genetic and microbial features using the combined model \(V\) to the sum of the variation explained by genetics and the variation explained by microbiome estimated using two separated models \((V_{g} + V_{m})\). If \(V_{g} = V_{m} + V_{g}\), the genetic and microbial features were independent and exerted additive effect on proteins. If \(V_{g} = V_{m} + V_{g}\), genetic and microbial effects were in part dependent.

If genetic and microbial effects were independent, we hypothesized that the power to detect microbial association would increase or remain similar after repressing out genetic effects. If genetic and microbial effects were not independent, we hypothesized that the microbial association would vanish or decrease after repressing out the genetic effects. Therefore, for each of the 73 proteins under genetic control (having significant pQTLs) and each of the 26 microbiome traits with mbQTLs (previously reported in ref. 84), we repeated microbial association analysis after repressing out genetic factors and compared the association strength before and after correcting for the genetic effect. As with the variance-explained model, this analysis was confined to 926 participants from the LLD cohort, and hence the FDR was calculated using the according P value and not the combined P value from LLD and LLD2.

Genetic and microbiome interaction analysis. To consider the relationship between genetic predisposition and the gut microbiome for 31 proteins significantly associated with both, interaction analysis was done for each protein as follows. For all possible combinations of each microbiome feature picked up by Lasso in the variance-explained model (except for alpha and beta diversity) and each SNP (cis or trans), we compared two models:

Model 1 (no interaction term): Protein = SNP + MB_feature

Model 2 (with interaction term): Protein = SNP + MB_feature + SNP × MB_feature

We utilized ANOVA to compare model 1 and model 2, testing whether the addition of the interaction term had a significant effect on the variability of the pre-existing model.

For the protein models with significant interactions identified by ANOVA (P <0.05), we reassessed all association terms, translating imputed SNP dosages to genotypes to identify non-additive effects of genetic–microbiome interactions.

Protein network. Proteins from the same pathways often exert PPIs and/or form co-abundance networks. To visualize protein–protein relationships, we plotted the protein network combining the experimentally validated PPIs from STRING41 and the protein co-abundance network that was constructed using the function bnlearn from the R package Bayesian network structure learning, parameter learning, and inference (bnlearn)42. The network was visualized using the program Cytoscape43. Node size reflected the amount of variance explained by genetics and microbiome together and the pie chart showed the percentage variance explained for cis-pQTLs, trans-pQTLs, and microbiome.

Partial correlation and mediation analysis between Ep-CAM, FUT2 genetic variants, and Blautia. To explore the relation between FUT2 secretor status, Ep-CAM level, and Blautia genus abundance, we used partial Spearman correlations (R package ppcor)44. Specifically, we re-estimated the associations between each pair of traits while controlling for the third. Ep-CAM level was considered a mediator of the effect of FUT2 secretor status on the level of Blautia. We estimated the proportion of FUT2 secretor effect on Blautia that mediated Ep-CAM level using R package mediation45. The significance of mediated and direct effects of FUT2 secretor status on Blautia was estimated using heteroskedasticity-consistent standard errors for quasi-Bayesian simulations and 10,000 permutations.

Microbial association analysis after correcting for BMI. A group of metabolism-related proteins, including PAI and PON3, were associated to the gut microbiome. We hypothesized that BMI might be a confounding variable. To further assess whether these microbial associations were related to or independent of BMI, we repeated microbial association analysis including BMI as a covariate.

Informed consent. The study was approved by the institutional review board of University Medical Center Groningen (UMCG), ref.M12.113965.

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

Data availability The LifeLines-DEEP metagenomics sequencing data are available at the European Genome-phenome Archive under accession EGA00001001704.

References
66. Zhermakova, A. et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science 352, 565–569 (2016).
67. Schoutens, S. et al. Cohort profile: LifeLines, a three-generation cohort study and biobank. Int. J. Epidemiol. 44, 1172–1180 (2015).
68. Assarsson, E. et al. Homogenous 96-Plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. PLoS One 9, e95192 (2014).
69. Truong, D. T. et al. MetaPhAAn for enhanced metagenomic taxonomic profiling. Nat. Methods 12, 902–903 (2015).
70. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).
71. Anders, S., Pyl, P. T. & Huber, W. HTSeq—A Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166–169 (2015).
72. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 11, R25 (2010).
73. Das, S. et al. Next-generation genotype imputation service and methods. Nat. Genet. 48, 1284–1287 (2016).
74. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. 48, 1279–1283 (2016).
75. The Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat. Genet. 46, 818–825 (2014).
76. Fehrmann, R. S. N. et al. Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS Genet. 7, e1002197 (2011).
77. Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. Cell 167, 1415–1429.e19 (2016).
78. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. Bioinformatics 25, 1841–1842 (2009).
79. Westra, H.-J. et al. Systematic identification of trans-eQTLs as putative drivers of known disease associations. Nat. Genet. 45, 1238–1243 (2013).
80. Magurran, A. E. Measuring Biological Diversity. (Blackwell Publishing, Oxford, 2004).
81. Dunn, O. J. Multiple comparisons among means. J. Am. Stat. Assoc. 56, 52–64 (1961).
82. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B 57, 289–300 (1995).
83. Tibshirani, R. Regression shrinkage and selection via the lasso: a retrospective. J. R. Stat. Soc. Ser. B 73, 273–282 (2011).
84. Bonder, M. J. et al. The effect of host genetics on the gut microbiome. Nat. Genet. 48, 1407–1412 (2016).
85. Szklarczyk, D. et al. STRINGv10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 43, D447–D452 (2015).
86. Scutari, M. Learning Bayesian networks with the bnlearn R package. J. Stat. Softw. 35, 1–22 (2010).
87. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498–2504 (2003).
88. Kim, S. ppcor: an R package for a fast calculation to semi-partial correlation coefficients. Commun. Stat. Appl. Methods 22, 665–674 (2015).
89. Tingley, D., Yamamoto, T., Hirose, K., Keele, L. & Imai, K. mediation: R package for causal mediation analysis. J. Stat. Softw. 59, 1–38 (2014).
Reporting Summary

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

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

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

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

| Data collection | No software was used for data collection |
| Data analysis | All data analyses were conducted using publicly available tools. The relative abundance of gut microbial taxonomic units was determined using Metagenomic Phylogenetic Analysis (MetaPhlan 2.2). The relative abundances of metabolic pathways were determined using the HUMAnN2 pipeline. QTL mapping pipeline was used for conducting both pQTL. The software and source-code is publicly available at https://github.com/molgenis/systemsgenetics/wiki/eQTL-mapping-analysis-cookbook and at https://github.com/molgenis/systemsgenetics/wiki/meQTL-mapping-analysis-cookbook, respectively. Most statistic analyses were conducted in R. The R packages and functions are described in details in the Online Method section. |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.
Data

Policy information about availability of data
All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The LifeLines-DEEP metagenomics sequencing data are available at the European Genome-phenome Archive (EGA) with accession code EGAS00001001704. The phenotype data can be requested through the Lifelines cohort study (www.lifelines.nl), following the standard procedure and the informed consent regulations of the UMCG's IRB that are described in the “Data and biomaterial access policy”.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences  ☐ Behavioural & social sciences  ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

The study has included a total of 1,264 individuals from the two Lifelines-DEEP cohorts. This is a study to ascertain to what extent individual’s variation in circulating proteins can be explained by genetics and gut microbiome. We chose a sample-size that as large as possible. We firstly included all 1,178 Lifelines-DEEP participants that have proteomics, genetics or metagenomic sequencing information available. To increase analysis power and validate findings, we further included extra 86 individuals (LifeLines-DEEP2).

Data exclusions

The following samples were excluded: family relatives and genetics outliers, and samples with low quality of metagenomics sequencing and/or genotyping data. For microbiome analysis, 11 antibiotics users were excluded too. The criteria for sample inclusion is described in the Online Methods section.

Replication

This study included two independent cohorts: 1,178 LLD1 participants and 86 LLD2 participants. We performed meta-analysis on both genetics analysis and metagenomics analysis. For genetics analysis, we also compared our pQTL results to previous four large-scale pQTLs studies. Our analysis showed a high replication rate: 95% for cis-effects and 88% for trans-effects, all with the same allelic direction.

Randomization

N.A. This is population-based association study. No randomization is needed.

Blinding

Not relevant for this study

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a Involved in the study
☐ Unique biological materials
☐ Antibodies
☐ Eukaryotic cell lines
☐ Palaeontology
☐ Animals and other organisms
☐ Human research participants

Methods

n/a Involved in the study
☐ ChIP-seq
☐ Flow cytometry
☐ MRI-based neuroimaging

Antibodies

Antibodies used

The proteomics data were measured using high-multiplex immunoassays by OLINK (Uppsala, Sweden, www.olink.com). This platform uses oligonucleotide-labeled antibody for protein measurement.

Validation

Commercialized and validated platform provided by the OLINK (Uppsala, Sweden, www.olink.com).
Human research participants

Policy information about studies involving human research participants

Population characteristics

The Lifelines cohort is a large prospective cohort study from the north of the Netherlands. This study includes two subcohorts of Lifelines: Lifelines-DEEP (LLD, n=1,500) and Lifelines-DEEP 2 (LLD2, n=119). The cohort contains 58% females and 42% males, the mean age is 45.04 (standard deviation (s.d.)=13.60), the mean body mass index (BMI) is 25.26 (s.d.=4.18) and 12% of participants are obese (BMI > 30). Common diseases within the cohort include high blood pressure (19%), anemia (15%), migraine (20%), irritable bowel syndrome (10%) and asthma (10%). The mean Framingham risk score for CVD is 8.6 for men and 5.7 for women, and 20% of individuals are current smokers. A detailed cohort description can be found in Tichelaar et al., BMJ Open, 2015.

Recruitment

This study was approved by the institutional ethical review board of the UMCG, M12.113965. All participants signed an informed consent prior to enrollment. All participants were recruited via the LifeLines organization (www.lifelines.nl) and followed the standard procedure.