Published in final edited form as:
Nat Genet. 2018 November ; 50(11): 1524–1532. doi:10.1038/s41588-018-0224-7.

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

Daria V. Zhernakova[#1,2], **Trang H. Le**[#1], **Alexander Kurilshikov**[#1], **Biljana Atanasovska**[1,3], **Marc Jan Bonder**[1], **Serena Sanna**[1], **Annique Claringbould**[1], **Urmo Võsa**[1], **Patrick Deelen**[1,4], **LifeLines cohort study, BIOS consortium**, **Lude Franke**[1], **Rudolf A. de Boer**[5], **Folkert Kuipers**[3,6], **Mihai G. Netea**[7,8], **Marten H. Hofker**[3], **Cisca Wijmenga**[1,9,11], **Alexandra Zhernakova**[1,11], and **Jingyuan Fu**[1,3,11]

[1]University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, the Netherlands [2]Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, St. Petersburg, Russian Federation [3]University of Groningen, University Medical Center Groningen, Department of Pediatrics, Groningen, the Netherlands [4]University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Groningen, the Netherlands [5]University of Groningen, University Medical Center Groningen, Department of Cardiology, Groningen, the Netherlands [6]University of Groningen, University Medical Center Groningen, Department of Laboratory Medicine, Groningen, the Netherlands [7]Department of Internal Medicine and Radboud Center for Infectious Diseases, Radboud University Medical Center, Nijmegen, the Netherlands [8]Department for Genomics and Immunoregulation, Life and Medical Sciences Institute (LIMES), University of Bonn, 53115, Bonn, Germany [9]K.G. Jebsen Coeliac Disease Research Centre, Department of Immunology, University of Oslo, Oslo, Norway

[#] These authors contributed equally to this work.

Abstract

Correspondence should be addressed to J.F. (j.fu@umcg.nl) or A.Z. (a.zhernakova@umcg.nl).
[11]These authors jointly directed this work.

Accession codes
The LifeLines-DEEP metagenomics sequencing data are available at the European Genome-phenome Archive (EGA) with access code EGAS00001001704.

Author contributions
Conceptualization: M.H.H., C.W., A.Z. and J.F. Methodology: D.V.Z., T.L., A.K., M.J.B., A.Z and J.F. Software: D.V.Z., T.L, M.J.B. and J.F. Formal analysis: D.V.Z., T.L., A.K., B.A., M.J.B., A.C., U.V., P.D., A.Z., and J.F. Writing–original draft: D.V.Z., T.L., A.K., A.Z. and J.F. Writing–review and editing: B.A., M.J.B., S.S., A.C., U.V., P.D., L.F., R.A.B., F.K., M.N., C.W., A.Z and J.F. Visualization: D.V.Z., T.L., A.K. and J.F. Supervision: A.Z. and J.F. Project administration: M.H.H., A.Z. and J.F. Funding acquisition: L.F., R.D.B., F.K., M.N., M.H.H., C.W., A.Z. and J.F.

Competing financial interests statement
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 financial interests.

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

Despite a growing body of evidence, the role of the gut microbiome in cardiovascular diseases (CVDs) is still unclear. Here we present a systems-genome-wide and metagenome-wide association study on plasma concentrations of 92 CVD-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 contributes 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 reveals 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 diseases 1, type 2 diabetes 2, inflammatory bowel disease 3,4, and different types of cancer 5. 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-DEEP1 (n=1,178) and Lifelines-DEEP2 (n=86) (Online Methods, Fig. 1), for which we have genotype, metagenome, transcriptome and detailed phenotype data 6. These proteins were selected *a priori* based on their direct or indirect role in the development of cardiovascular diseases (CVD) (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 250kb 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 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 7 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), e.g. 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 8–11 and identified a 95% replication rate for *cis*-effects and an 88% replication rate for *trans*-effects, all with the same allelic direction (Supplementary Table 5). 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 12 (Supplementary Note 3, Supplementary Tables 2 and 6, Supplementary Fig. 5). 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 does provide evidence that a large amount of *trans*-regulation can happen at translation- or protein-level, e.g. through regulation of translational rate and protein secretion to blood, through post-translational modification or through protein-protein interactions, 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, Online Methods, 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) 13, epidermal growth factor receptor (EGFR) 14, members of the paraoxonases family (PON) 15,16, tumor necrosis factor receptor (TNF-R2) 17 and insulin-like growth factor (IGF) 18. For example, the gastric pathogen *Helicobacter pylori* stimulates expression of PAI and of uPA and its receptor (uPAR) in gastric epithelial cells 13 and can also affect epithelial signaling pathways partially through activation of EGFR 14.

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) 19 (Supplementary Note 4). While we did not observe any overlap between pQTLs and mbQTLs at the genome-wide significance level, several pQTLs for 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, 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 (Online Methods). *Cis*-pQTL effects explained an average of 14.9% of the variation of 66 of the *cis*-regulated proteins (range 0.35% to 73.3%). T*rans*-pQTL effects explained an average of 9.5% of the variation for 36 of the *trans*-regulated proteins (range 0.7% to 51.8%). Compared to host genetic factors, the effect of microbial features was smaller, contributing on average 3.2% of the variation for the 41 proteins (range 0.4% to 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 IL6-RA 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 epithelial cell adhesion molecule (Ep-CAM) and trefoil factor 3 (TFF3). 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 19q13.33 and the *TMIGD1* locus at 17q11.2 (Supplementary Table 14). Both the *FUT2* and *TMIGD1* genes are abundantly expressed in the intestine (Supplementary Fig. 10). Ep-CAM is a trans-membrane glycoprotein highly expressed in the colon and small intestine (according to GTEx data 20) and associated with both benign and malignant cell proliferation. Ep-CAM is overexpressed in cancer cells 21,22 and considered a plausible anti-cancer drug target 23–25. The *FUT2* gene encodes alpha-(1,2)-fucosyltransferase, which is involved in the creation of the H antigen precursor. The top *trans*-pQTL, SNP rs601338 in *FUT2* (r_{LLD1}=-0.13; P=3.9x10^{-6}), is a functional SNP. The rs601338-A allele creates a stop codon that results in a non-functional enzyme. Individuals homozygous for rs601338-A are not able to express the H-type1 oligosaccharide ligand to their mucus and are called ‘non-secretors’ for ABH antigens. FUT2 non-secretors have reduced susceptibility to multiple pathogens, including Norwalk virus 26 and *Helicobacter pylori* 27, but are at increased risk for Crohn’s disease 28. *FUT2* locus genotypes are also strongly associated with cancer markers 29 and to plasma level of B_{12}, which is clinically associated with CVD 30.

In our study, FUT2 non-secretor status correlated with a lower expression of Ep-CAM (r=-0.227, P=2.3x10^{-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 31,32, 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=2.4x10^{-4}) (Fig. 3B) for which both genetic and microbiome data are available 19. 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=9x10^{-6}, 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 ($P_{mediated}$=0.0084, P_{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, anti-proliferative and antineoplastic properties and are implicated in protection against colorectal cancer 35.

Other producers of butyrate include commensal bacteria *Clostridia*, which helps to maintain gut homeostasis 36. We also found the plasma level of Ep-CAM was associated with many species of the *Clostridiales* order, e.g. *Clostridium bartlettii* (r=-0.30, P=3.8x10^{-25}, FDR=4.7x10^{-21}) (Supplementary Table 9). Interestingly, we also detected suggestive interactions between *Clostridia 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 preQ$_0$ (r=-0.29, P=4.3x10^{-22}, FDR=1.7x10^{-17}), a newly discovered metabolite with anti-cancer activity 37. However, FUT2 secretor/non-secretor status did not appear to have any effect on the association between Ep-CAM and bacterial biosynthesis of anti-cancer compounds. Based on the methodology behind the Mendelian randomization approach, the lack of genetic association of *FUT2* with pre-Q$_0$ biosynthesis does not support a causal impact of Ep-CAM on pre-Q$_0$ biosynthesis. This may indicate that pre-Q$_0$ biosynthesis can affect Ep-CAM, in line with the anti-cancer properties of pre-Q$_0$.

PON3 links lipid oxidation to the gut microbiome

Proteins that function in the same pathways can have direct protein-protein interactions and/or form a co-abundant relationship. To visualize these relations, we combined Bayesian network analysis on protein co-abundance with experimentally validated protein-protein interactions 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 sub-cluster in the Bayesian network analysis. These proteins—PAI, PON3, uPA, IGFBP-1 and IGFBP-2, TFF3, 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=1.8x10^{-34}), which regulates adipogenesis and adipocyte metabolism 38, and insulin like growth factor binding protein 1 (IGFBP-1), which plays an important role in glucose and insulin metabolism 39 (r=-0.44; P=6.9x10^{-49}). Triglyceride concentration was mostly associated to IGFBP-2 (r=-0.35; P=2.0x10^{-31}), while HDL cholesterol level was mostly positively associated to PON3 (r=0.40; P=4.2x10^{-39}). Insulin was mostly associated to insulin binding proteins IGFBP-1 (r=-0.40; P=4.35x10^{-41}) and IGFBP-2 (r=-0.39; P=1.69x10^{-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 oxidation 40,41. 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=2.2x10^{-5}, FDR=0.002) and had the second highest association to beta-diversity (r^2=0.004, P=9.9x10^{-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 inflammatory bowel disease (IBD), including the BMI- and lipid-lowering species *Akkermansia muciniphila* 42,43 (r=0.12, P=7.8x10^{-5}, FDR=0.03), the butyrate-producing bacteria *Roseburia hominis* 44 (r=0.13, P=2.6x10^{-5}, FDR=0.01), and that hydrogen-removing archaeon *Methanobrevibacter smithii* 45(r=0.18, P=2.1x10^{-8}, FDR=3.3x10^{-5}) that has also been associated with BMI in the Lifelines-Deep cohort 46. 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 concordant direction (Supplementary Fig. 11A). Among them, one significant association was that observed for L-methionine biosynthesis (HOMOSER-METSYN-PWY, r=-0.17, P=3.4x10^{-8}, FDR=7.2x10^{-5}). L-methionine is an essential amino acid that must be obtained through diet that 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 47, although this has been debated in recent years 48. Another significant association we found was for a bacterial pathway involved in thiamine (vitamin B1) biosynthesis (PWY-7357, r=-0.16, P=2.3x10^{-7}, FDR=2.6x10^{-4}). Notably, a suggestive interaction was also observed between the thiamine biosynthesis pathway and the *cis*-pQTL SNP (rs10953142) of PON3 (ANOVA P=0.003) (Supplementary Table 15).

Serum level of plasminogen activator inhibitor (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 level can be induced by oxidized LDL 49,50. Consistent with PON3 preventing LDL oxidation 40, our data showed negative association between PON3 and PAI ($r=-0.33$, $P < 2.2x10^{-16}$), with both being associated to metabolic traits like 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, almost all bacterial species (8 out of 9) and pathways (38 out of 46) specifically associated to PON3 remained significant after regressing out BMI, suggesting PON3 has a more direct impact on<?> the gut microbiome (Supplementary Fig. 11B). Previous studies have shown that paraoxonases (PON1, 2 and 3) can display antioxidative and anti-inflammatory effects in the intestine 51. In patients with inflammatory gastrointestinal disorders, *PON3* gene expression in the intestine and colon was significantly reduced 52. Moreover, expression of human *PON1* in *Drosophila* resulted in a decrease in the superoxide anion level, which plays an important role in gut immunity, and eventually in altered bacterial colonization 15. Thus, the association between PON3 and the gut microbiome has revealed a novel functional link between lipid oxidation, intestinal health and gut microbiome.

Genetic-microbiome interaction in the gut-brain axis

The central nervous system (CNS) is emerging as an important factor in control of BMI and susceptibility to obesity and related morbidity53. In addition to proteins functioning in the intestine and in metabolism, we found some proteins associated with the gut microbiome that 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 6.6% of the variation in CNTN1 could be explained by two independent *cis*-pQTL SNPs, rs12811939 ($r=0.21$; $P=1.14x10^{-13}$) and rs17541621 ($r=0.18$; $P=7.47x10^{-10}$) (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 ($r=0.26$; $P=4.25x10^{-22}$) (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 ($r=0.13$, $P=2.5x10^{-5}$, $FDR=8.6x10^{-3}$, Fig. 5B). Chorismate is a precursor for tryptophan, which can be further converted to serotonin. 90% of tryptophan is synthesized in the gastrointestinal tract and is considered as a key signaling metabolite between the gut and the CNS 54. 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=1.6x10^{-5}) (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=2.75x10^{-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 55. To search for other genes affected by this SNP, we looked for it in a larger RNA-seq-based eQTL mapping 56. In addition to the eQTL for *TMEM8A*, we found that this SNP affects alternative exon usage (exon-ratio levels) of *AXIN1*. The gene *AXIN1* encodes a scaffold protein in the β-catenin destruction complex which, if disrupted, contributes to pathogenesis of various human diseases including colorectal carcinogenesis and IBD 57. *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 58. Moreover, evidence also supports vitamin D and the vitamin D receptor (VDR) as important regulators of IBD and colon cancer and they also play a role in maintenance of physiological level of AXIN1 57. This highlights the potential link between *AXIN1* and the gut microbiome.

Discussion

Cardiovascular diseases (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 as 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 disease including CVD 59. In line with this, Ep-CAM-deficient mice exhibited increased intestinal permeability and decreased iron transport 60, which may contribute to CVD susceptibility risk 59. 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 anti-cancer compound: a higher plasma level of Ep-CAM was associated with lower level of the bacterially produced anti-cancer metabolite preQ$_0$.

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 61. 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* 62. 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 low-density lipoprotein (LDL) contributes to the risk of CVD. LDL oxidation can be inhibited by paraoxonases that are bound to HDL and exert antioxidant properties. The paraoxonase (PON) family contains three members: PON1, 2 and 3. The role of PON1 in atherosclerosis has been well-established 63, it was recently identified as part of the cardiac secretome in post-myocardial infarction heart failure 64. 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 likely occurring via a different mechanism 65. 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 which modulate various biological processes and 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 Methods

Study cohorts

The Lifelines cohort is a large prospective cohort study from the north of the Netherlands. This study includes two sub-cohorts 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 Tigchelaar et al. 6 and Zhernakova et al. 66

For both these sub-cohorts multiple levels of omics data are available, including genotypes, gene expression, microbiome, and proteomics data 6,67.

Data generation and preprocessing

Profiling of proteomics data—Blood samples of study participants were collected and frozen at -80°C prior to 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, Uppsala, Sweden), a multiplex immunoassay for high-throughput detection of protein biomarkers in liquid samples 68. 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 with 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 (PCA) on adjusted protein levels. We found that the first 10 PCs were associated with genetic variants and therefore did not adjust for any PCs 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 earlier 66. 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) 69. MetaPhlan 2.2 reported 1,772

URLs

The LifeLines cohort study: www.lifelines.nl; Olink CVD II panel description: http://www.olink.com/products/cvd-iii-panel/; HUMAnN2 pipeline: http://huttenhower.sph.harvard.edu/humann2; UniRef50: http://www.uniprot.org/; MetaCyc metabolic pathway database: www.metacyc.org; Vegan R package: *http://CRAN.R-project.org/package=vegan*; the Human reference consortium (HRC v1.0) preparation checking tool: http://www.well.ox.ac.uk/~wrayner/tools/.

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 were 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 (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 datasets were log-transformed for subsequent analysis.

Expression data—Whole blood RNA-seq data was generated as described previously 56. In brief, paired end reads were aligned to the human genome using STAR aligner 70, allowing for no more than 7 mismatches. Gene expression estimation was performed using HTSeq-count 71 on uniquely mapping reads. Gene level expression data was normalized by TMM method 72, followed by $\log_2$ 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 was available for 1,293 samples.

Genotype data—Microarray genotype data for the LLD cohort was generated using CytoSNP and ImmunoChip assays as previously described 6. Quality control checks on the LLD cohort were performed using the Human reference consortium (HRC v1.0) preparation checking tool (v4.2.3) (see URLs section). We then uploaded the resulting VCF files to the Michigan Imputation Server 73. Phasing and imputation were performed using the option SHAPEIT for phasing, population EUR and mode Quality Control and Imputation, for all steps we used version R1 as reference 74.

Whole-genome-sequencing-based genotypes were generated for LLD2 as described previously 75. As the LLD2 participants were trios, we excluded first-degree relatives (offspring) for this study.

We further excluded SNPs that had imputation quality $r^2 < 0.5$, failed the Hardy-Weinberg equilibrium test ($P < 1x10^{-6}$), 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 pipeline 76, we performed QTL mapping by calculating Spearman rank correlation in LLD (1178 samples with both genotype and protein data) and LLD2 (86 samples with both genotype and protein data) datasets 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 false discovery rate (FDR) at 0.05.

Cis-pQTL mapping was done by testing SNP-gene pairs located within a ±250kb window. Proteins (corresponding Uniprot IDs) were converted to genes and genomic coordinates using Ensembl v75. One of the proteins (PECAM-1) mapped to an unmapped contig, 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 $P \leq 5\times10^{-8}$. Additionally, we added 2,706 genome-wide significant GWAS SNPs from a recent blood trait GWAS 77. SNP coordinates were lifted to hg19 using the *liftOver* command from the rtracklayer v1.34.1 R package 78 and subsequently standardized to match the GIANT 1000G p1v3 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. 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 it by removing *trans*-pQTL SNPs in linkage disequilibrium with $r^2 > 0.8$. 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 the large scale whole blood eQTL published by Westra et al. 79 and eQTLs data of various tissue types from GTEx v7 release 12.

Replication of published pQTLs—To replicate previously reported pQTLs, we considered all pQTL studies with reasonable sample sizes (n > 1,000) published in the two years prior to November 2017 (Sun et al. 2017 8, Suhre et al. 2016 9, Sun et al. 2016 10 and Folkersen et al. 2017 11). We first extracted pQTLs for the 92 proteins under study and tested all reported pQTLs in our dataset. 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—We then assessed the genetic effects of SNPs associated with coronary artery disease (CAD) reported by Nelson et al.7. 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 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 that 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 microbiome 80. We calculated alpha diversity using the Shannon diversity index, using the function *diversity()* from the R package *vegan* (see $^{\text{URLs}}$ 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 correction 81.

Association with beta-diversity

Beta diversity is the degree of differentiation of taxonomic communities between individuals (i.e. inter-individual gut microbiomes) and is usually defined by a between-community dissimilarity matrix. We calculated beta diversity using Bray-Curtis (BC) dissimilarity, using the function *vegdist()* from the R package *vegan* (see $^{\text{URLs}}$ section).

To quantify the variation in BC dissimilarity explained by an individual protein, Permutational Multivariate Analysis of Variance (PerMANOVA) 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 correction 81. PerMANOVA 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 n x n dissimilarity matrix). This test partitions the variation across a multivariate data cloud in response to one or more factors and uses permutations to determine P-value.

Association analysis with taxa and bacterial pathways

Pathway and taxa datasets were corrected for factors that can significantly alter the microbiome composition, including age, sex, read depth, antibiotic usage, irritable bowel syndrome and diarrhea. 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-Hochberg 82 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 ($Vcis$ and $Vtrans$) were estimated. Together, they make up variance explained by genetics ($Vg = Vcis + Vtrans$). We also calculated the combined variance explained by genetics by putting all *cis*- and *trans*-pQTLs in one model. The difference between the R^2 of the latter model and Vg is negligible (not shown), which confirms the expected independence contribution of *cis*- and *trans*-pQTLs.

Unlike the genetic data, the microbiome data has 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 model 83 that included all microbial pathways and species significantly associated with the respective protein, together with the Shannon diversity index and the first 5 PCs of Bray-Curtis dissimilarity as predictors. The independent and most dominant microbial features were selected automatically by L1-norm regularized quantile regression (the coefficients for non-informative features were automatically shrunk to 0). A Lambda regularization parameter was selected using ten-fold cross-validation for each protein independently, choosing from 100 lambdas. The variation of each protein explained by selected microbial features (Vm) was then estimated using a linear model.

The total variance explained by genetics and microbiome (Vt) 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 co-founding 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 effects on proteins deviates from an additive model. To do this we compared the total variation explained by genetic and microbial factors using the combined model (Vt) to the sum of the variation explained by genetics and the variation explained by microbiome estimated using two separated models ($Vg+Vm$). If $Vt=Vg+Vm$, the genetic and microbial factors were independent and exerted additive effect on proteins. If $Vt < Vg+Vm$, 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 regressing out genetic effects. If genetic and microbial effects were not independent, we hypothesized that the microbial association would vanish or decrease after regressing 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 Bonder et al. 84), we repeated microbial association analysis after regressing 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, hence the FDR was calculated using the according P-value, 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:

$$\begin{aligned}&\text{Model 1(no interaction term)}: \text{Protein} = \text{SNP} + \text{MB\_feature}\\&\text{Model 2(no interaction term)}: \text{Protein} = \text{SNP} + \text{MB\_feature} + \text{SNP} * \text{MB\_feature}\end{aligned}$$

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

For the protein models with significant interactions identified by ANOVA ($P < 0.05$), we reassessed the all interaction 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 protein-protein interactions and/or form co-abundance networks. To visualize protein-protein relationships, we plotted the protein network combining the experimentally validated protein-protein interactions (PPIs) from STRING 85 and the protein co-abundance network that was constructed using the function *bnfit()* from the R package for Bayesian Network Structure Learning, Parameter Learning and Inference (bnlearn) 86. The network was visualized using the program Cytoscape 87. 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* 88). Specifically, we reestimated the associations between each pair of traits while controlling for the third. Ep-CAM level was considered as 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 package *mediation* for R 89. 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 UMCG, ref.M12.113965.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

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 Jackie Dekens, Mathieu Platteel, Astrid Maatman and Jody Arends for management and technical support and Jackie Senior and Kate Mc Intyre for editing the manuscript. We thank Arnau Vich Vila and Rinse K. Weersma for discussions.

This project was funded by the Netherlands Heart Foundation (IN-CONTROL CVON grant 2012-03 to M.H.H., M.G.N., F.K., A.Z. and J.F. and CVON-DOSIS grant 2014-40 to R.D.B.), by the Top Institute Food and Nutrition, Wageningen, the Netherlands (TiFN GH001to C.W.), by the Netherlands Organization for Scientific Research (NWO) (NWO-VIDI 864.13.013 to J.F., NWO VIDI 917.13.350 to R.D.B., NWO-VIDI 016.178.056 to A.Z., NWO-VIDI 917.14.374 to L.F., NWO Spinoza Prize SPI 94-212 to M.N.G and NWO Spinoza Prize SPI 92-266 to C.W., NWO Gravitation Netherlands Organ-on-Chip Initiative (024.003.001) 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 and ERC Starting Grant 637640 to L.F.), by the Stiftelsen Kristian Gerhard Jebsen foundation (Norway) to C.W. and by the RuG 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.

References

1. Tang WHW, Hazen SL. The contributory role of gut microbiota in cardiovascular disease. J Clin Invest. 2014; 124:4204–11. [PubMed: 25271725]
2. Qin J, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 2012; 490:55–60. [PubMed: 23023125]
3. Knights D, et al. Complex host genetics influence the microbiome in inflammatory bowel disease. Genome Med. 2014; 6:107. [PubMed: 25587358]
4. Imhann F, et al. Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease. Gut. 2016; 67:108–119. [PubMed: 27802154]

5. Nakatsu G, et al. Gut mucosal microbiome across stages of colorectal carcinogenesis. Nat Commun. 2015; 6 8727.

6. Tigchelaar EF, et al. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. BMJ Open. 2015; 5

7. Nelson CP, et al. Association analyses based on false discovery rate implicate new loci for coronary artery disease. Nat Genet. 2017; 49:1385–1391. [PubMed: 28714975]

8. Sun BB, et al. Consequences Of Natural Perturbations In The Human Plasma Proteome. bioRxiv. 2017; :1–46. DOI: 10.1101/134551

9. Suhre K, et al. Connecting genetic risk to disease endpoints through the human blood plasma proteome. bioRxiv. 2016; 086793. doi: 10.1101/086793

10. Sun W, et al. Common Genetic Polymorphisms Influence Blood Biomarker Measurements in COPD. PLoS Genet. 2016; 12 e1006011.

11. Folkersen L, et al. Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. PLoS Genet. 2017; 13:1–21.

12. Aguet F, et al. Genetic effects on gene expression across human tissues. Nature. 2017; 550:204–213. [PubMed: 29022597]

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. 2008; 295:G431–41. [PubMed: 18599586]

14. Keates S, et al. cag+ Helicobacter pylori induce transactivation of the epidermal growth factor receptor in AGS gastric epithelial cells. J Biol Chem. 2001; 276:48127–34. [PubMed: 11604402]

15. Pezzulo AA, et al. Expression of Human Paraoxonase 1 Decreases Superoxide Levels and Alters Bacterial Colonization in the Gut of Drosophila melanogaster. PLoS One. 2012; 7

16. Stoltz DA, et al. Drosophila are protected from Pseudomonas aeruginosa lethality by transgenic expression of paraoxonase-1. J Clin Invest. 2008; 118:3123–31. [PubMed: 18704198]

17. Miller PG, Bonn MB, Franklin CL, Ericsson AC, McKarns SC. TNFR2 Deficiency Acts in Concert with Gut Microbiota To Precipitate Spontaneous Sex-Biased Central Nervous System Demyelinating Autoimmune Disease. J Immunol. 2015; 195:4668–4684. [PubMed: 26475926]

18. Yan J, et al. Gut microbiota induce IGF-1 and promote bone formation and growth. Proc Natl Acad Sci. 2016; 113:E7554–E7563. [PubMed: 27821775]

19. Bonder MJ, et al. The effect of host genetics on the gut microbiome. Nat Genet. 2016; 48:1407–1412. [PubMed: 27694959]

20. The Gtex Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013; 45:580–5. [PubMed: 23715323]

21. Trzpis M, McLaughlin PMJ, de Leij LMFH, Harmsen MC. Epithelial cell adhesion molecule: more than a carcinoma marker and adhesion molecule. Am J Pathol. 2007; 171:386–395. [PubMed: 17600130]

22. Baeuerle PA, Gires O. EpCAM (CD326) finding its role in cancer. Br J Cancer. 2007; 96:417–423. [PubMed: 17211480]

23. Gires O, Bauerle PA. EpCAM As a Target in Cancer Therapy. J Clin Oncol. 2010; 28:e239–e240. [PubMed: 20385979]

24. Kurtz J-E, Dufour P. Adecatumumab: an anti-EpCAM monoclonal antibody, from the bench to the bedside. Expert Opin Biol Ther. 2010; 10:951–958. [PubMed: 20426706]

25. Andersson Y, et al. Phase I trial of EpCAM-targeting immunotoxin MOC31PE, alone and in combination with cyclosporin. Br J Cancer. 2015; 113:1548–1555. [PubMed: 26554649]

26. Lindesmith L, et al. Human susceptibility and resistance to Norwalk virus infection. Nat Med. 2003; 9:548–53. [PubMed: 12692541]

27. Magalhães A, et al. Fut2-null mice display an altered glycosylation profile and impaired BabA-mediated Helicobacter pylori adhesion to gastric mucosa. Glycobiology. 2009; 19:1525–36. [PubMed: 19706747]

28. McGovern DPB, et al. Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. Hum Mol Genet. 2010; 19:3468–76. [PubMed: 20570966]

29. He M, et al. A genome wide association study of genetic loci that influence tumour biomarkers cancer antigen 19-9, carcinoembryonic antigen and α fetoprotein and their associations with cancer risk. Gut. 2014; 63:143–51. [PubMed: 23300138]

30. Hazra A, et al. Common variants of FUT2 are associated with plasma vitamin B12 levels. Nat Genet. 2008; 40:1160–2. [PubMed: 18776911]

31. Rausch P, et al. Colonic mucosa-associated microbiota is influenced by an interaction of Crohn disease and FUT2 (Secretor) genotype. Proc Natl Acad Sci U S A. 2011; 108:19030–5. [PubMed: 22068912]

32. Tong M, et al. Reprograming of gut microbiome energy metabolism by the FUT2 Crohn's disease risk polymorphism. ISME J. 2014; 8:2193–206. [PubMed: 24781901]

33. Ríos-Covián D, et al. Intestinal short chain fatty acids and their link with diet and human health. Frontiers in Microbiology. 2016; 7

34. Tanaka S, Yamamoto K, Yamada K, Furuya K, Uyeno Y. Relationship of enhanced butyrate production by colonic butyrate-producing bacteria to immunomodulatory effects in normal mice fed an insoluble fraction of Brassica rapa L. Appl Environ Microbiol. 2016; 82:2693–2699. [PubMed: 26921420]

35. Canani RB, et al. Potential beneficial effects of butyrate in intestinal and extraintestinal diseases. World J Gastroenterol. 2011; 17:1519–1528. [PubMed: 21472114]

36. Lopetuso LR, Scaldaferri F, Petito V, Gasbarrini A. Commensal Clostridia: leading players in the maintenance of gut homeostasis. Gut Pathog. 2013; 5:23. [PubMed: 23941657]

37. Xu D, et al. PreQ0 Base, an Unusual Metabolite with Anti-cancer Activity from Streptomyces qinglanensis 172205. Anticancer Agents Med Chem. 2015; 15:285–290. [PubMed: 25353335]

38. Goralski KB, et al. Chemerin, a Novel Adipokine That Regulates Adipogenesis and Adipocyte Metabolism. J Biol Chem. 2007; 282:28175–28188. [PubMed: 17635925]

39. Bae J-H, Song D-K, Im S-S. Regulation of IGFBP-1 in Metabolic Diseases. J lifestyle Med. 2013; 3:73–9. [PubMed: 26064841]

40. Witte I, Foerstermann U, Devarajan A, Reddy ST, Horke S. Protectors or Traitors: The Roles of PON2 and PON3 in Atherosclerosis and Cancer. J Lipids. 2012; 2012:1–12.

41. Kowalska K, Socha E, Milnerowicz H. Review: The role of paraoxonase in cardiovascular diseases. Annals of Clinical and Laboratory Science. 2015; 45:226–233. [PubMed: 25887882]

42. Everard A, et al. Cross-talk between Akkermansia muciniphila and intestinal epithelium controls diet-induced obesity. Proc Natl Acad Sci. 2013; 110:9066–9071. [PubMed: 23671105]

43. Schneeberger M, et al. Akkermansia muciniphila inversely correlates with the onset of inflammation, altered adipose tissue metabolism and metabolic disorders during obesity in mice. Sci Rep. 2015; 5 16643.

44. Tamanai-Shacoori Z, et al. Roseburia spp.: a marker of health? Future Microbiol. 2017; 12:157–170. [PubMed: 28139139]

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. 2016; 43:197–212. [PubMed: 26559904]

46. Mbakwa CA, et al. Gut colonization with *methanobrevibacter smithii* is associated with childhood weight development. Obesity. 2015; 23:2508–2516. [PubMed: 26524691]

47. McCully KS. Homocysteine and vascular disease. Nat Med. 1996; 2:386–9. [PubMed: 8597939]

48. Wierzbicki AS. Homocysteine and cardiovascular disease: a review of the evidence. Diabetes Vasc Dis Res. 2007; 4:143–50.

49. Chi YS, Bong CK, Hye KH, Hyun SL. Oxidized LDL activates PAI-1 transcription through autocrine activation of TGF-?? signaling in mesangial cells. Kidney Int. 2005; 67:1743–1752. [PubMed: 15840021]

50. Hong HK, Song CY, Kim BC, Lee HS. ERK contributes to the effects of Smad signaling on oxidized LDL-induced PAI-1 expression in human mesangial cells. Transl Res. 2006; 148:171–179. [PubMed: 17002919]

51. Précourt LP, et al. The three-gene paraoxonase family: Physiologic roles, actions and regulation. Atherosclerosis. 2011; 214:20–36. [PubMed: 20934178]

52. Rothem L, et al. Paraoxonases are associated with intestinal inflammatory diseases and intracellularly localized to the endoplasmic reticulum. Free Radic Biol Med. 2007; 43:730–739. [PubMed: 17664137]

53. Locke AE, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015; 518:197–206. [PubMed: 25673413]

54. Ruddick JP, et al. Tryptophan metabolism in the central nervous system: medical implications. Expert Rev Mol Med. 2006; 8:1–27.

55. Philonenko ES, et al. TMEM8 - A non-globin gene entrapped in the globin web. Nucleic Acids Res. 2009; 37:7394–7406. [PubMed: 19820109]

56. Zhernakova DV, et al. Identification of context-dependent expression quantitative trait loci in whole blood. Nat Genet. 2016; 49:139–145. [PubMed: 27918533]

57. Jin D, et al. Vitamin D receptor is a novel transcriptional regulator for Axin1. J Steroid Biochem Mol Biol. 2017; 165:430–437. [PubMed: 27601169]

58. Zhang Y, et al. Axin1 Prevents Salmonella Invasiveness and Inflammatory Response in Intestinal Epithelial Cells. PLoS One. 2012; 7:e34942. [PubMed: 22509369]

59. Bischoff SC, et al. Intestinal permeability--a new target for disease prevention and therapy. BMC Gastroenterol. 2014; 14:189. [PubMed: 25407511]

60. Kozan PA, et al. Mutation of EpCAM leads to intestinal barrier and ion transport dysfunction. J Mol Med (Berl). 2015; 93:535–45. [PubMed: 25482158]

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. 2014; 34:939–45. [PubMed: 24526693]

62. Rodriguez-Nunez I, et al. Nod2 and Nod2-regulated microbiota protect BALB/c mice from diet-induced obesity and metabolic dysfunction. Sci Rep. 2017; 7:548. [PubMed: 28373658]

63. Soran H, Schofield JD, Durrington PN. Antioxidant properties of HDL. Front Pharmacol. 2015; 6:222. [PubMed: 26528181]

64. Meijers WC, et al. The Failing Heart Stimulates Tumor Growth by Circulating Factors. Circulation. 2018; CIRCULATIONAHA.117.030816. doi: 10.1161/CIRCULATIONAHA.117.030816

65. Zhang C, et al. Studies on protective effects of human paraoxonases 1 and 3 on atherosclerosis in apolipoprotein E knockout mice. Gene Ther. 2010; 17:626–33. [PubMed: 20182519]

66. Zhernakova A, et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science (80-). 2016; 352:565–569.

67. Scholtens S, et al. Cohort Profile: LifeLines, a three-generation cohort study and biobank. Int J Epidemiol. 2015; 44:1172–80. [PubMed: 25502107]

68. Assarsson E, et al. Homogenous 96-Plex PEA Immunoassay Exhibiting High Sensitivity, Specificity, and Excellent Scalability. PLoS One. 2014; 9:e95192. [PubMed: 24755770]

69. Truong DT, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods. 2015; 12:902–903. [PubMed: 26418763]

70. Dobin A, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29:15–21. [PubMed: 23104886]

71. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015; 31:166–9. [PubMed: 25260700]

72. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010; 11:R25. [PubMed: 20196867]

73. Das S, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016; 48:1284–1287. [PubMed: 27571263]

74. McCarthy S, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016; 48:1279–83. [PubMed: 27548312]

75. The Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet. 2014; 46:818–25. [PubMed: 24974849]

76. Fehrmann RSN, 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. 2011; 7:e1002197. [PubMed: 21829388]

77. Astle WJ, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell. 2016; 167:1415–1429.e19. [PubMed: 27863252]

78. Lawrence M, Gentleman R, Carey V. rtracklayer: An R package for interfacing with genome browsers. Bioinformatics. 2009; 25:1841–1842. [PubMed: 19468054]

79. Westra H-J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013; 45:1238–1243. [PubMed: 24013639]

80. Magurran AE. Measuring biological diversity. Blackwell Publishing; Oxford: 2004.

81. Dunn OJ. Multiple Comparisons among Means. J Am Stat Assoc. 1961; 56:52–64.

82. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B. 1995; 57:289–300.

83. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. J R Stat Soc Ser B (Statistical Methodol. 2011; 73:273–282.

84. Bonder MJ, et al. The effect of host genetics on the gut microbiome. Nat Genet. 2016; 48:1407–1412. [PubMed: 27694959]

85. Szklarczyk D, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015; 43:D447–D452. [PubMed: 25352553]

86. Scutari M. Learning Bayesian Networks with the bnlearn R Package. J Stat Softw. 2010; 35:1–22. [PubMed: 21603108]

87. Shannon P, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. 2003; 13:2498–2504. [PubMed: 14597658]

88. Kim S. ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. Commun Stat Appl Methods. 2015; 22:665–674. [PubMed: 26688802]

89. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K. mediation : *R* Package for Causal Mediation Analysis. J Stat Softw. 2014; 59:1–38. [PubMed: 26917999]

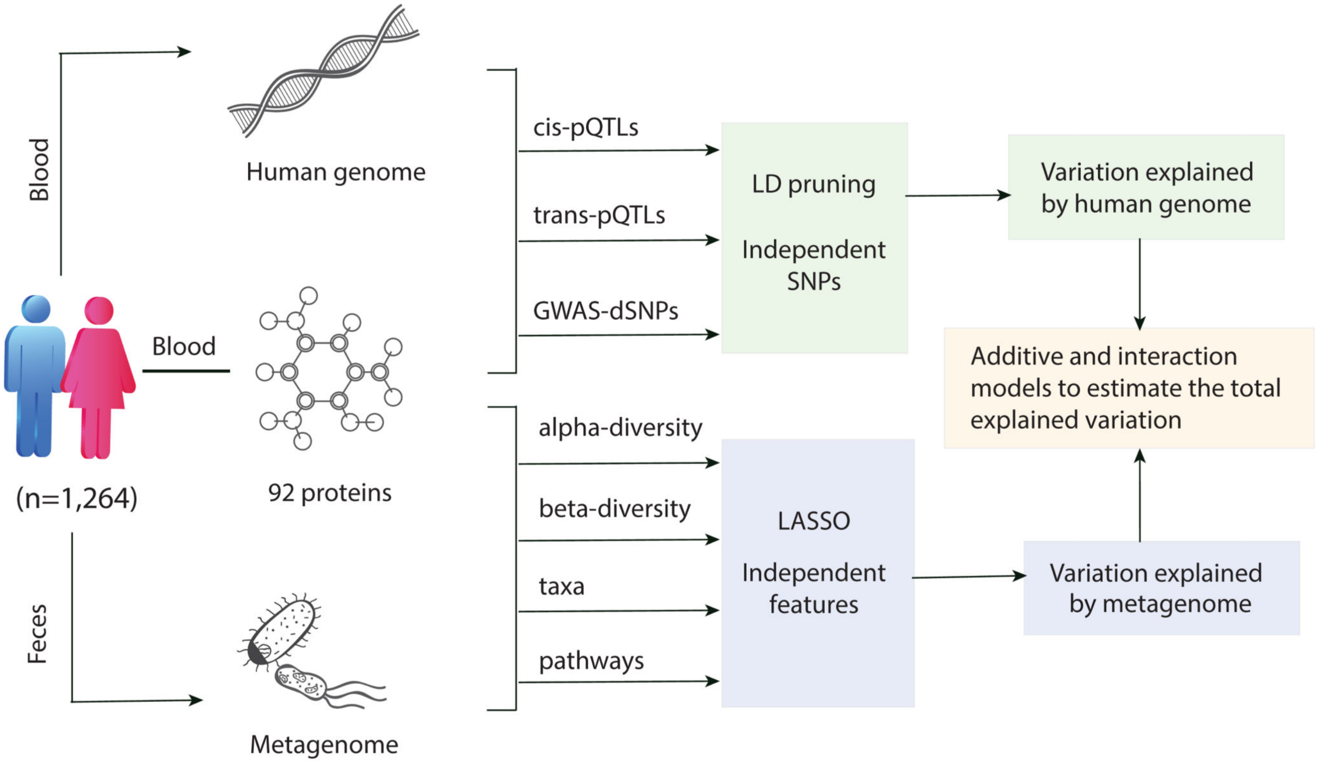

Fig. 1. Study analysis workflow.

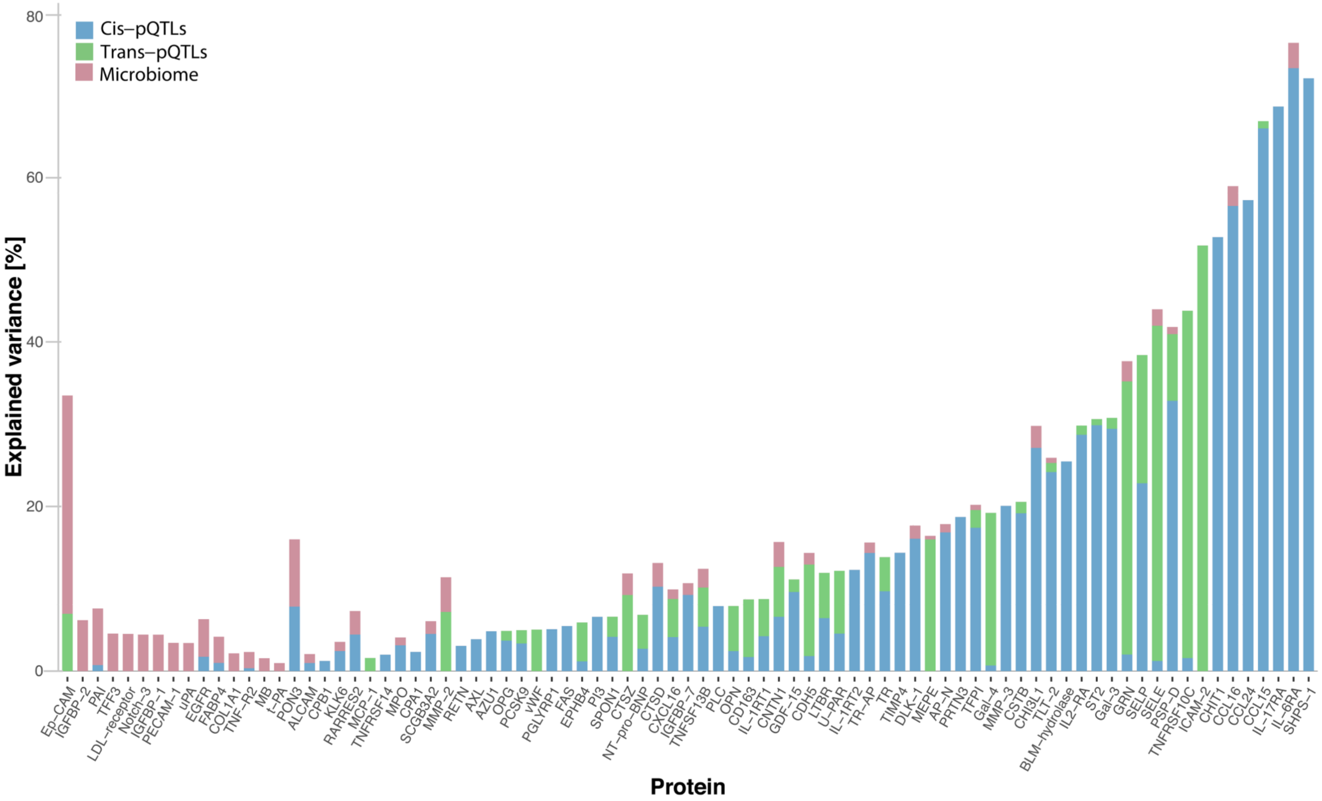

Fig. 2. Proportion of inter-individual variation explained by genetic and microbial factors.

Each bar represents a protein. Y-axis is the explained variation. Proportion of variation is separated into *cis*-pQTLs (blue), *trans*-pQTLs (green) and microbiome (red).

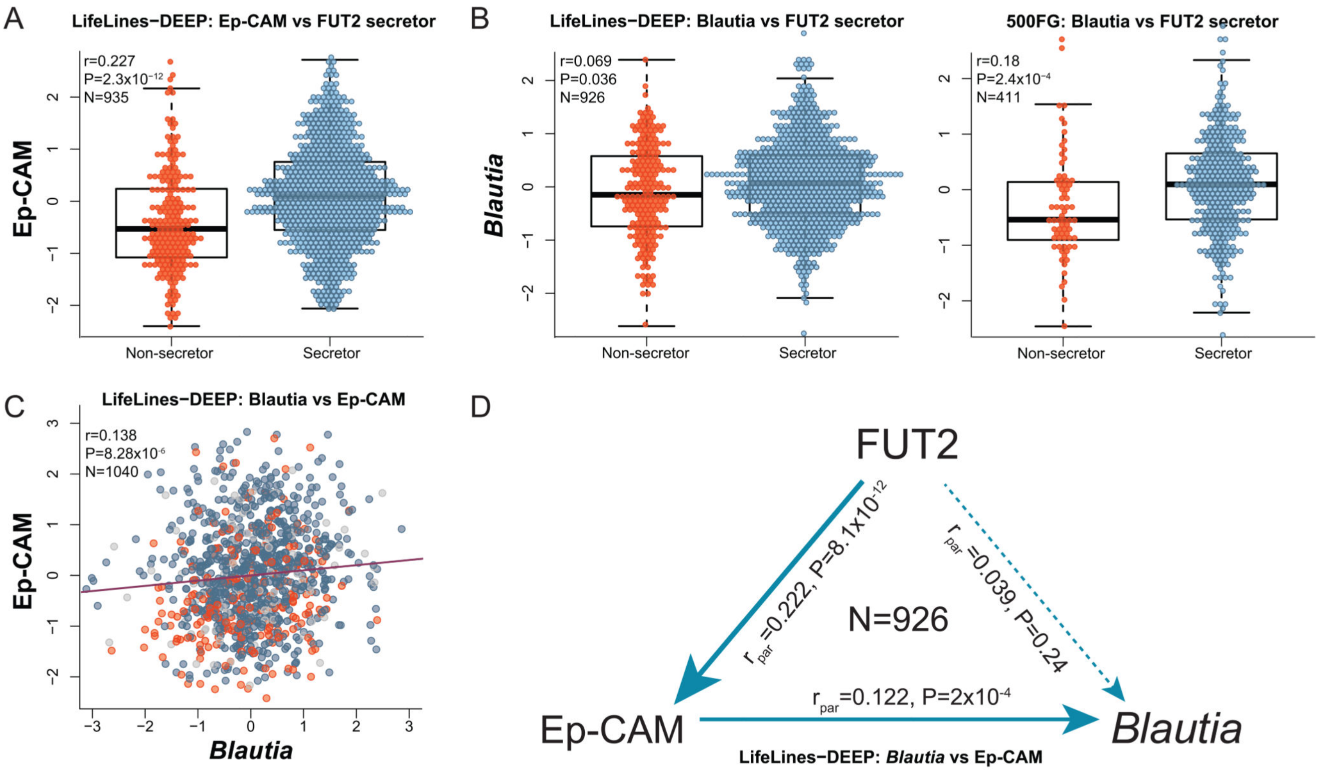

Fig. 3. Association of *FUT2*, Ep-CAM and *Blautia*.

On all plots, level of Ep-CAM and log-transformed abundance of *Blautia* genus are scaled to mean=0/SD=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 levels 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.

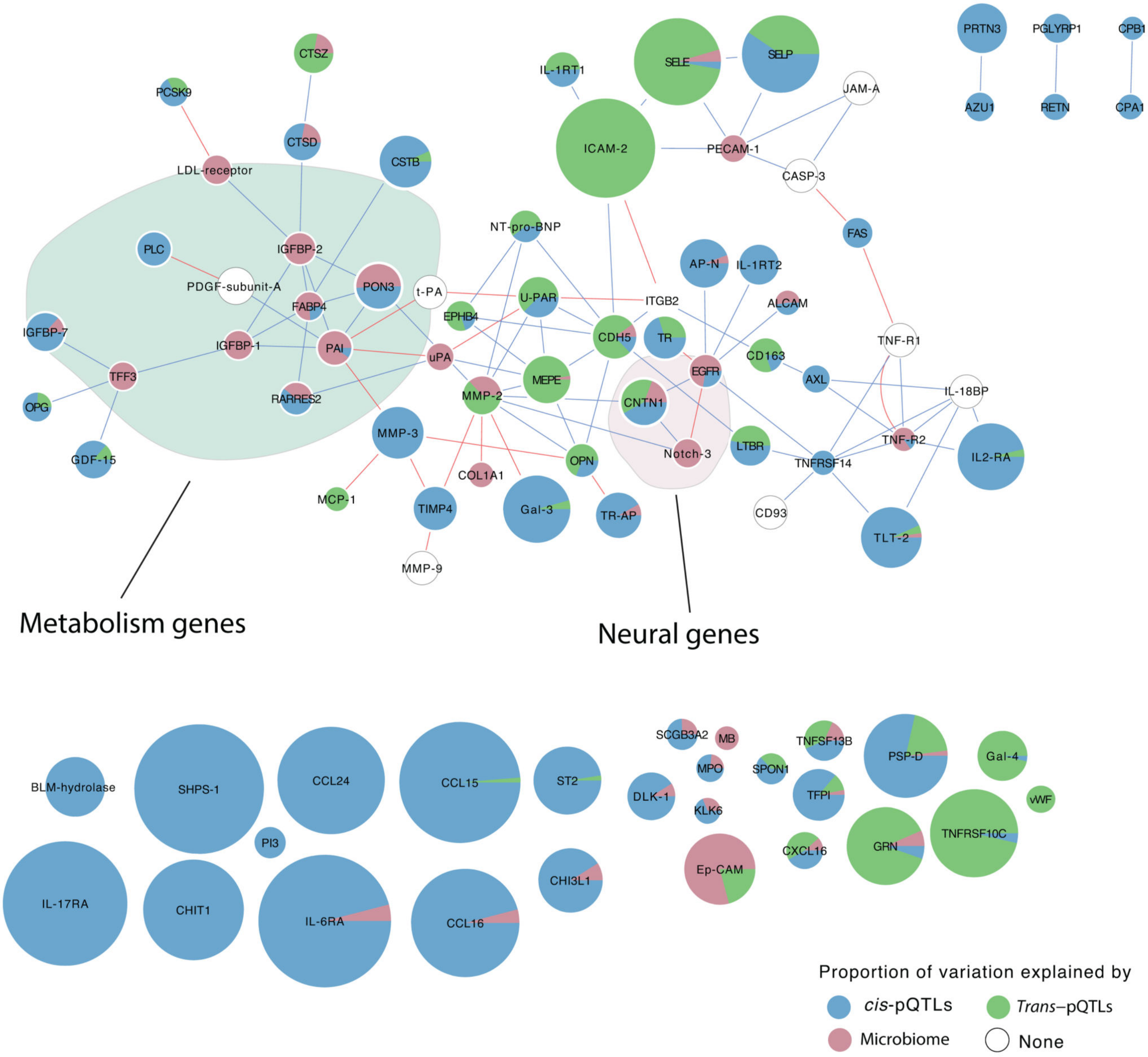

Fig. 4. Network of protein co-abundance and protein-protein interactions.

Each circle represents a protein. Circle size indicates the proportion of total explained variation. Large circles suggest a large amount of variation could be explained by genetic and/or microbial factors. Small circles suggest a small proportion of variation explained. The pie chart in each circle reflects the relative contributions of *cis*-pQTLs, *trans*-pQTLs and gut microbiome to the total explained variation. Blue edges refer to co-abundance of Bayesian network analysis. Red edges refer to experimentally verified protein-protein interaction. Two sub-networks are highlighted: a group of proteins involved in metabolism (pale green area) and a group of genes functioning in neural system (light gray area).

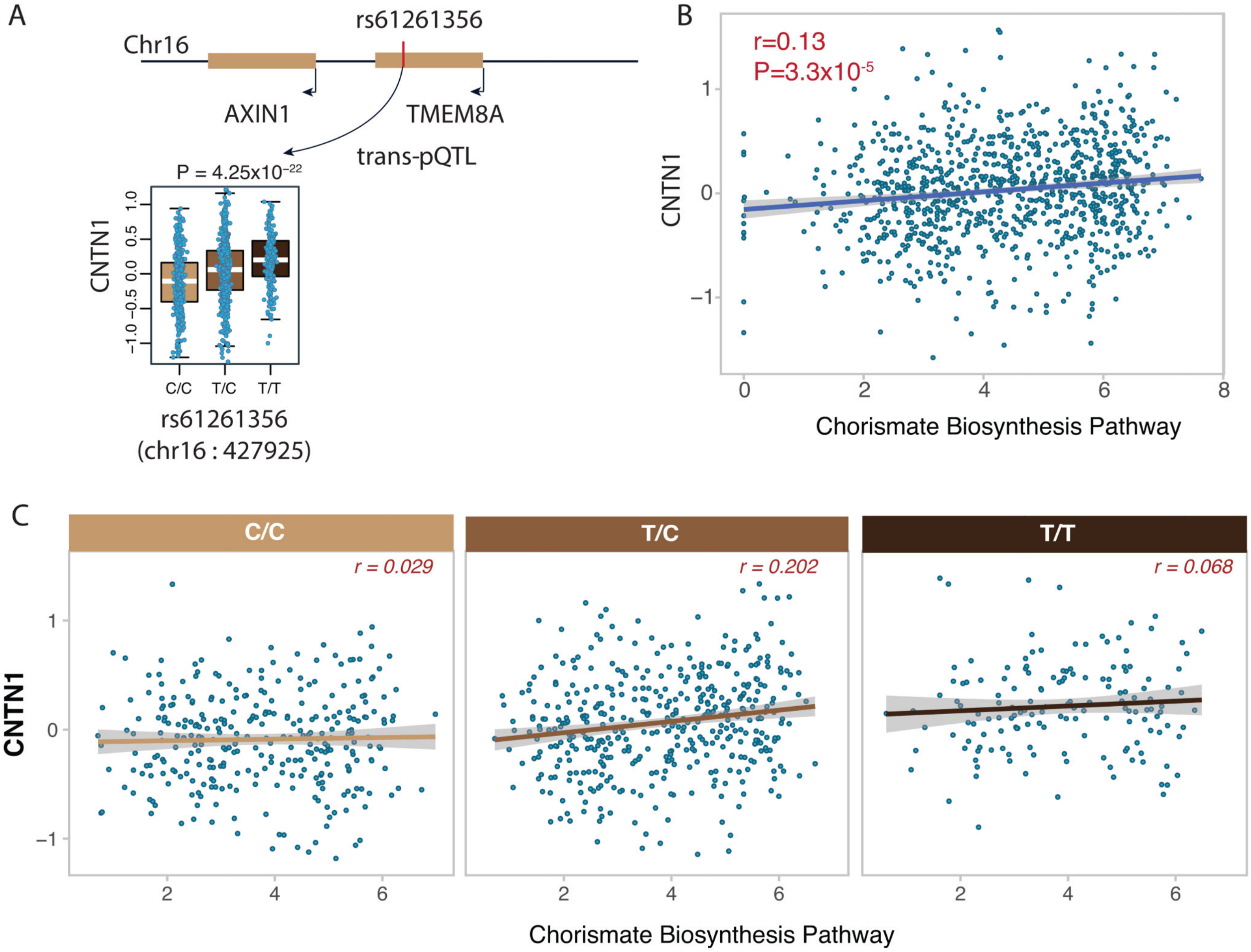

Fig. 5. Genetic-microbiome interaction for CNTN1.

A. *Trans*-pQTL effect of CNTN1 at SNP rs61261356 in the *TMEM8A/AXIN1* locus. **B.** Association between plasma level of CNTN1 and the bacterial chorismate biosynthesis pathway. **C.** Genetic-microbiome interaction in CNTN1. Association strength between CNTN1 and chorismate biosynthesis is different and significantly lower in homozygous individuals.