Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Microbiome-wide PheWAS links gut microbial SNVs to human health and exposures

Jingyuan Fu

j.fu@umcg.nl

University Medical Center Groningen https://orcid.org/0000-0001-5578-1236

Daoming Wang

University Medical Center Groningen https://orcid.org/0000-0003-4623-8527

Sergio Andreu-Sánchez

University Medical Center Groningen

Haoran Peng

University Medical Center Groningen https://orcid.org/0009-0002-7779-7073

Angel Ruiz-Moreno

University Medical Center Groningen https://orcid.org/0000-0002-2798-5386

Daria Zhernakova

University Medical Center Groningen

Alexander Kurilshikov

University Medical Center Groningen

Ranko Gacesa

University of Groningen and University Medical Center Groningen https://orcid.org/0000-0003-2119-0539

Godfrey Temba

Radboud University Medical Center

Vesla Kullaya

Kilimanjaro Christian Medical University College

Ramnik Xavier

Massachusetts General Hospital https://orcid.org/0000-0002-5630-5167

Quirijn de Mast

Radboud University Medical Center

Leo Joosten

Radboud University Medical Center https://orcid.org/0000-0001-6166-9830

Niels Riksen

Radboud University Medical Center https://orcid.org/0000-0001-9197-8124

Joost Rutten

Radboud University Medical Center

Mihai Netea
Radboud University Nijmegen Medical Centre https://orcid.org/0000-0003-2421-6052
Serena Sanna
University Medical Center Groningen
Cisca Wijmenga
University Medical Centre Groningen https://orcid.org/0000-0002-5635-1614
Rinse Weersma
University of Groningen and University Medical Center Groningen https://orcid.org/0000-0001-7928-7371
Alexandra Zhernakova
University of Groningen

Biological Sciences - Article

Keywords: Metagenome-wide association study, Phenome-wide association study, Genome-wide association study, SNV, Gut microbiome, Metagenome, Genetic variation, Aging, Lineage effect, Host–microbe interaction

Posted Date: September 26th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-5063726/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: Yes there is potential Competing Interest. R.K.W. acted as a consultant for Takeda, received unrestricted research grants from Takeda, Johnson & Johnson, Tramedico and Ferring, and received speaker fees from MSD, Abbvie, and Janssen Pharmaceuticals. A.Z. received a speaker fee from Nestle. All other authors declare no competing interests.

Version of Record: A version of this preprint was published at Nature on November 4th, 2026. See the published version at https://doi.org/10.1038/s41586-026-11220-0.

Microbiome-wide PheWAS links gut microbial SNVs to human health and exposures

Authors

Daoming Wang[1], Sergio Andreu-Sanchez[1,2,*], Haoran Peng[1,2,*], Angel J. Ruiz-Moreno[1,2], Daria V. Zhernakova[1], Alexander Kurilshikov[1], Ranko Gacesa[1,3], Godfrey S. Temba[4,5,6], Vesla I. Kullaya[5,7], Lifelines Cohort Study, Human Functional Genomics Project, Ramnik J. Xavier[8], Quirijn de Mast[4,6], Leo A.B. Joosten[4,6,9], Niels P. Riksen[4], Joost H.W. Rutten[4], Mihai G. Netea[4,6,10,11], Serena Sanna[1,12], Cisca Wijmenga[1], Rinse K. Weersma[3], Alexandra Zhernakova[1], Jingyuan Fu[1,2,#]

Affiliations

[1] Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen 9713AV, the Netherlands

[2] Department of Pediatrics, University of Groningen, University Medical Center Groningen, Groningen 9713AV, the Netherlands

[3] Department of Gastroenterology and Hepatology, University of Groningen, University Medical Center Groningen, Groningen 9713 GZ, the Netherlands

[4] Department of Internal Medicine, Radboud University Medical Center, Nijmegen 6500 HB, the Netherlands

[5] Department of Medical Biochemistry and Molecular Biology, Kilimanjaro Christian Medical University College, P.O. Box 2240, Moshi, Tanzania

[6] Radboud Center for Infectious Diseases, Radboud University Medical Center, Nijmegen 6500 HB, the Netherlands

[7] Kilimanjaro Clinical Research Institute, Kilimanjaro Christian Medical Center, Moshi, Tanzania

[8] Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

[9] Department of Medical Genetics, Iuliu Haţieganu University of Medicine and Pharmacy, Cluj-Napoca 400000, Romania

[10] Department of Immunology and Metabolism, Life and Medical Sciences Institute, University of Bonn, Bonn 53113, Germany

[11] Human Genomics Laboratory, Craiova University of Medicine and Pharmacy, Craiova 200349, Romania

[12] Institute for Genetic and Biomedical Research, National Research Council, Cagliari, Italy

* These authors contributed equally

[#] Lead contact: J.fu@umcg.nl

Highlights

- We developed a bioinformatics framework to perform a lineage-aware single nucleotide variant–based metagenome-wide association study (SNV-MWAS).
- Through a phenome-wide association study (PheWAS), we provide a catalog of 37,945 associations with 237 host phenotypes and environmental exposures, offering insights into the relationship between host health and gut microbes at base pair scale.
- In meta-analysis, 718 SNVs of 91 species were associated with age, supporting variation in the gut microbiome as a hallmark of aging.
- We identified 25 associations between microbial SNVs and diet in both Dutch and Tanzanian cohorts.
- Associated microbial SNVs can reflect microbial evolution and adaptation to the host's physiological status, as we observed for an oligopeptide transportation gene of *Faecalibacteria prausnitzii* and host aging.
- SNVs in an α-Melanocyte-stimulating hormone-mimicking *clpB* gene of *Alistipes ihumii* are associated with host body weight and blood lipid levels.
- Associated microbial SNVs pinpoint microbial genes putatively involved in human health, as we saw for SNVs of the isoleucyl-tRNA synthetase gene of *Clostridium phoceensis,* which mediates the beneficial impact of diet quality on host health.

Abstract

The importance of the gut microbiome in human health has been widely recognized, with alterations in species abundances now reported for many different diseases. In recent years, observations of within-species genetic diversity and strain-level functional differences have urged the field to transition to analyzing microbes at the genetic level, taking advantage of genetic variants in the metagenome to pinpoint relevant genes and pathways. Here, we conducted a phenome-wide association study (PheWAS) of 12.7 million gut microbial single nucleotide variants (SNVs) with 244 host health phenotypes and environmental exposures in 10,781 individuals from six independent cohorts from Europe and Africa. We found 3,018 associations supported by at least two cohorts for 105 host factors, including age, sex and cardiovascular and metabolic parameters. Moreover, we found 34,927 significant associations in single-cohort analyses where SNV–phenotype pairs were only available for one cohort. All identified associations are detailed in an online catalog (https://fulab.web.rug.nl/SNV-MWAS/). The associated microbial SNVs inform mechanistic hypotheses about microbial functionality in human health and provide insights into microbial adaptation and evolution in response to changes in the environment and host physiology. For example, we identified 718 age-associated gut microbial SNVs across multiple cohorts and used these microbial SNVs to accurately predict host age. We also identified SNVs in microbial genes associated with lipid concentrations and BMI, including SNVs in the isoleucine-tRNA synthetase gene of *Clostridium phoceensis* and an α-Melanocyte-stimulating hormone-mimicking *clpB* gene of *Alistipes ihumii*. Mediation analysis provided evidence that microbial genetic variations may be under selection enforced by environmental factors, but also that they mediate the effects on the host, e.g. the isoleucyl-tRNA synthetase gene of *C. phoceensis* mediated dietary effects on four phenotypes. Overall, this study identifies new biological connections between human host and microbial genetic variants, exposing the functional layer of gut microbiome data hidden at base pair resolution and revealing microbial adaptation and evolution in relation to host lifestyle and physiology.

Keywords: Metagenome-wide association study, Phenome-wide association study, Genome-wide association study, SNV, Gut microbiome, Metagenome, Genetic variation, Aging, Lineage effect, Host–microbe interaction

INTRODUCTION

The diverse population of microbial species in the human gut forms a micro-ecological system. This ecosystem exhibits large variation, not only in its composition, but also in its genetic landscape. The within-species genetic variation determines strain differences in fitness, metabolic capacities, pathogenicity, and other biological properties. Thus, the microbiome research field is moving towards strain-level dissection of microbial contributions to human health. As these genetic differences may explain the heterogeneous findings of previous analyses based on microbial abundance or community composition, genetics-level analysis offers an opportunity to pinpoint microbial genes responsible for host–microbe interaction.

Different kinds of genetic variations exist within the gut microbiome, including thousands of structural variations (SVs) [1–4] and millions of single nucleotide variants (SNVs) [5,6]. SVs are genomic regions, ranging in size, that can be deleted, duplicated, transmitted, or reversed [3,4,7,8]. Microbial SVs have previously been associated with human BMI, bile acid (BA) concentrations, the human plasma metabolome, and host genetics, thereby pinpointing causal genes involved in BA biosynthesis and biotransformation, glucose metabolism, and N-acetylgalactosamine (GalNac) utilization [1,4,9,10]. SNVs are point mutations in gut microbial genomes that can also modify gene functionality, and thus also modulate host health. For example, an SNV in the *blc* gene of *Escherichia coli* can disrupt the intestinal barrier by decreasing the level of lysophosphatidylethanolamine, a metabolite helpful in remission of inflammatory bowel disease [11]. In the stomach, an SNV of the bacterial pathogen *Helicobacter pylori* can increase gastric cancer risk by enhancing proteolytic activity of serine protease HtrA[12]. Gut microbial SNVs have also been associated with cardiometabolic diseases, for example SNV distributions in glycosyl hydrolase genes of *Bacteroides coprocola* are are associated with type 2 diabetes [13].

Over the last two decades, genome-wide association studies (GWAS) have been widely conducted in human genetics to assist in the discovery of the genetic determinants underlying complex traits without prior-knowledge-based hypothesis [14]. The methodologies of human GWAS have now inspired microbial GWAS, which aim to discover microbial genetic determinants underpinning microbial phenotypes, particularly for infectious pathogens like *H. pylori* [15], *Mycobacterium tuberculosis* [16,17], and *Staphylococcus aureus* [18]. However, the high diversity and complexity of microbial genomes represents a major challenge for microbial GWAS [19,20], especially for those linking metagenome-based microbial data with host phenotypes without severe clinical manifestations [21]. Although microbiome studies based on whole-genome sequencing have now produced a lot of data on both host and gut microbes, gut microbial SNV-based metagenome-wide association studies (SNV-MWAS) on the human phenome and exposome remain rare [21]. Recently, a SNV-MWAS in an Israeli population

identified BMI-associated gut microbial SNVs, which could be replicated in a geographically independent Dutch cohort [21], providing support for SNV-MWAS as a method to identify health-relevant microbial genetic variants. Despite intriguing examples, there is still a lack of systematic exploration of the relationships of gut microbial SNVs with multiple host phenotypes and environmental exposures on phenome-wide association study (PheWAS) scale. Technical challenges have also hampered application of SNV-MWAS. For instance, microbial species possess strong and complex population genetic structures. When genetic differences between lineages or clades contribute to phenotypic differences, a series of lineage-specific genetic markers can be associated to the target phenotype, producing a strong lineage effect that would lead to false discoveries in association studies if the population structure is not properly controlled.

In this study, we genotyped 11,831 fecal metagenomic samples in five independent cohorts from the Netherlands and one from Tanzania. We then analyzed the 12.7 million common SNVs that were present in more than 10% of samples with a minor allele frequency (MAF) >5% in at least one cohort. We also developed an linear mixed model (LMM)-based analysis framework that controls for population genetic structure within species to ensure that the SNV–phenome associations we identify are not acting as a proxy for the cumulative effect of strain lineage. Our model takes genotype-dosage into account, considering different relative abundances of alleles in each metagenome sample. We then conducted a PheWAS to search for associations between the common metagenotypes and 244 host phenotype and exposure factors, using LMMs in each cohort, followed by a meta-analysis. We detected 3,018 associations supported by multiple cohorts and 34,927 single-cohort associations for phenotype–SNV pairs only available in one cohort, in total encompassing 237 of the 244 human phenotype and exposure factors analyzed. The full set of results is available in an online catalog.

RESULTS

Gut microbial SNV landscape

We obtained fecal metagenomic sequencing data for 11,831 samples from 10,781 individuals from six cohorts (**Fig. 1a**). This included five cohorts from the Netherlands, the Dutch Microbiome Project (DMP, *N* = 7,955) [22], Lifelines-DEEP (LLD, *N* = 1,135, plus 338 followed-up after an average of 4 years (designated LLD1 for baseline samples and LLD2 for follow-up samples)) [9,23], the 500 Functional Genomics Project (500FG, *N* = 529) [24,25], the 1000 Inflammatory Bowel Disease Project (IBD, *N* = 544) [26], and 300-Obesity (300OB, *N* = 298) [25,27], and one cohort from Tanzania, the 300 Tanzania Functional Genomics Project (300TZFG, *N* = 320) [28] (**Table S1**). For LLD2 and 500FG, we also have additional re-sequenced datasets from the same 251 and 459 individuals, respectively, included in metagenotyping. These were used to assess the impact of DNA extraction protocol on megagenotyping (but not in downstream association analysis) (**Fig. S1**, **Table S1** and **Supplementary Note 1**). Microbial DNA of two aliquots of each sample from LLD2 and 500FG were isolated using the QIAamp Fast DNA Stool Mini Kit (FSK) and the Qiagen AllPrep DNA/RNA kits (APK) [29]. The extensive phenotype information of corresponding samples was also curated and harmonized. Finally, we integrated the 244 factors available in the DMP or at least two cohorts for analysis, including 92 binary traits and 152 continuous or ordinal traits (**Fig. 1b** and **Table S2**). 131 phenotypes were available in at least two cohorts, and 113 were available for only one cohort (**Table S2**). The total sample size per phenotype ranged from 593 to 11,706 in the whole dataset (**Table S2**).

We detected 50,945,657 biallelic SNVs in the core genome region of 881 species using the GT-Pro tool [6], which conducts reference-panel-based metagenotyping using k-mer searching and exact matching. A larger number of SNVs were detected in the DMP, likely due to its large sample size (**Fig. S1a**,**b**). The total number of SNVs detected in at least one sample for each species ranged from 1 to 342,230 (average = 58,131) in the whole dataset (**Fig. S1c**). The average density of SNVs in the core genomic region of gut bacteria was 32 SNVs per kilobase pair (**Fig. S1d**), which is close to previous reports [30]. To keep only informative species for downstream analysis, we selected the “common” SNVs that were present in >10% of samples and had a MAF >5% in at least one cohort. Among the 50.95M SNVs detected, 12,761,497 SNVs were common (**Figures S1e** and **S2** and **Table S3**). As the distributions of common SNV numbers showed two peaks separated at around 1,000 SNVs (**Fig. S1e**), we selected species with at least 1000 common SNVs, leading to 433 species with 12,706,564 common SNVs (**Fig. 1c** and **d**, **Fig. S1f** and **g** and **Table S4**). After filtering, 300TZFG had the highest number of species for analysis (327) and the most common

SNVs (7,369,266) (**Figures 1c–e** and **S1f–h** and **Table S4**). The average density of common SNV sites in the core genomic region of gut bacteria was 16 SNVs per kilobase pair (**Figures 1f** and **S1i** and **Table S4**). The within-species population structure of the 433 species and the impact of technical factors are extensively described in **Supplementary Notes 2** and **3** (**Figures S2** and **S3** and **Tables S5–S8**).

To understand the functional consequences of the 12.7 million common gut microbial SNVs of the 433 species, we annotated genes of reference genomes and predicted the potential functional impact of SNVs on bacterial gene products, starting from SNV synonymity. SNVs were further classified into three impact levels according to their impact [31] (**Fig. S4a,b** and **Table S9**). The majority of SNVs were predicted to have a low impact, including the 75.06% of SNVs annotated to synonymous variants and the 7.20% annotated to upstream gene variants. In total, 17.52% of the common SNVs were categorized as having moderate impact (**Fig. S4b**), and these primarily consisted of missense or nonsynonymous variants, which alter amino acid sequences and potentially affect protein function. Just 0.1% of the common SNVs were identified as high-impact variants (**Fig. S4b**), including mutations that induce loss of function, premature stop codons, and shifts in reading frames. The distribution of high or moderate SNVs across different bacterial species varied, reflecting the genetic diversity and evolutionary pressures experienced by these species. For example, *Streptococcus thermophilus* exhibited the highest proportion of high or moderate impact SNVs (**Table S9**), suggesting a greater potential for adaptive changes in response to pressures.

Landscape of gut microbial SNV associations

We next performed a PheWAS on 244 factors through a gut microbial SNV-based metagenome-wide association study (SNV-MWAS) based on LMMs in each species, phenotype, and cohort separately. The LMM controls for the population structure and lineage effect by including the species-level between-sample genetic linkage matrix as a random effect covariate. Basic phenotypes like age, sex, and clean read count were included as fixed-effect covariates, except in the analyses for age and sex themselves. Given that both the reference and alternative alleles of an SNV may exist in the same sample, with differing abundances, we encoded the gut microbial genotypes to a dosage value ramging from 0 to 1 to represent the relative proportion of the alternative allele. To further control for false discoveries caused by inflated significance, we calculated the genomic inflation factor λ and excluded species–phenotype pairs with a high ($\lambda > 2$) or low ($\lambda < 0.5$) genomic inflation factor from the final association results of each cohort.

For the SNV–phenotype pairs available in at least two cohorts, we conducted a

sample-size-weighted meta-analysis to integrate the SNV-MWAS results derived from different cohorts. To ensure the independence of samples, the LLD2 and 500FG-APK datasets were not included in the meta-analysis. In total, we identified 22,861 significant associations ($P_{meta} < 6.47\times10^{-6}$) that could be supported by at least two cohorts at nominal significance level ($P_{cohort} < 0.05$) with the same association direction. This included 3,922 associations at metagenome-wide significance level ($P_{meta} < 7.34\times10^{-9}$) and 3,180 associations at study-wide significance level ($P_{meta} < 3.01\times10^{-11}$) (**Figures S6** and **S7**; see **Methods** for determination of significance cutoffs). Finally, we conducted a permutational Spearman correlation test (999 times) within each cohort to assess the robustness of these associations (see **Methods**) and further selected 3,018 robust associations ($P_{meta} < 6.47\times10^{-6}$ and $P_{cohort\text{-}permutation} < 0.05$; **Fig. S6**). Of these, 282 reached metagenome-wide significance level ($P_{meta} < 7.34\times10^{-9}$) and 232 reached study-wide significance level ($P_{meta} < 3.01\times10^{-11}$) (**Fig. S8a** and **Tables S10** and **S11**). 91.5% of the associations were supported by at least two cohorts, mostly DMP and LLD1 (**Fig. S8b** and **Table S11**), which is expected as these two cohorts have the highest sample sizes and share 40 available phenotypes (**Table S2**). Many commonly collected phenotypes, including age, sex and several anthropometric phenotypes, exhibited microbial SNV associations that can be supported by three or more cohorts (**Tables S2** and **S11**).

Age was associated with 718 SNVs of 91 species, showing the highest number of associations (**Fig. 2a,b** and **Tables S10** and **S11**). Many cardiometabolic-health-related phenotypes showed associations with the SNVs of >10 species, including five body shape phenotypes (body weight, waist circumference, BMI, waist-hip ratio (WHR), and hip circumference), four blood biochemical measurements (triglycerides (TG), creatinine, glucose, and high-density lipoprotein cholesterol (HDL-C)), and blood pressure (systolic blood pressure, diastolic blood pressure, and hypertension) (**Fig. 2b**). Although the body shape phenotypes are strongly inter-related (**Fig. S9a**), most associations are phenotype-specific. However, a few associations were shared between phenotypes, e.g., BMI and body weight shared 17 associated SNVs (**Fig. S9b**). Stool-related factors, including Bristol stool type and stool frequency, are often associated with overall gut microbiome composition [22,32], and we also observed Bristol stool type and stool frequency to be associated with the SNVs of 22 and 18 gut microbial species, respectively (**Fig. 2b**). Additionally, smoking, alcohol consumption, and daily calorie intake were associated with SNVs of >10 species, and the most significant meta-analysis association was between an SNV of a *Faecalibacterium* species (species ID: 102072) and recent 'on-a-diet' behavior ($P_{meta} = 1.34\times10^{-94}$; **Table S11**), which is a combination set of diet behaviors with a specific health-related consideration (e.g., energy-restricted diet, fat-restricted diet, sodium-restricted diet). Overall, these results reveal the shaping of gut microbial genomic variation by host

lifestyle factors (**Fig. 2b**).

Since a large number of SNV–phenotype pairs were available in only one cohort, we summarized these single-cohort associations and screened the robust associations using a permutational Spearman correlation test (999 times). This identified 34,927 associations between SNVs and host phenotypes in the cohorts that they specifically represent ($P_{cohort} < 6.47\times10^{-6}$; $P_{cohort\text{-}permutation} < 0.05$; **Table S12**), including 2,622 that reached metagenome-wide significance level ($P_{meta} < 7.34\times10^{-9}$) and 1,547 that reached study-wide significance level ($P_{meta} < 3.01\times10^{-11}$) (**Fig. S8c**). Since the DMP cohort had the largest number of available phenotypes and largest sample size, it contained the most (89.5%) single-cohort associations (**Fig. S8d**). Of the 31,244 associations in DMP, 2,476 reached study-wide significance and three dietary factors (daily calorie intake, daily fructose intake, and daily polysaccharides intake) showed the highest number of study-wide significant associations ($P_{meta} < 3.01\times10^{-11}$; **Table S10**).

Functional properties of associated gut microbial SNVs

We then examined the impact of the associated SNVs on gene-product function. Among the 3,018 meta-analysis associations and 34,927 single-cohort associations, 585 associations involved gut microbial SNVs with high level impact on gene function (stop-gained, start-lost, stop-lost, or splice region variants) and 6,364 associations involved SNVs with moderate level impact (missense, splice region or stop-retained, start-retained, or initiator codon variants) (**Fig. 3a,b**). For the species–phenotype pairs with >10 SNV associations, we compared the rate of high or moderate impact SNVs between associated SNVs and all SNVs in the corresponding species (**Fig. 3c,d**). Here we found that the rate of high or moderate impact SNVs of 31 and 216 species–phenotype pairs were higher in the associated SNVs than in all SNVs of the corresponding species in meta-analysis and single-cohort associations, respectively (**Tables S13** and **S14**). We then conducted an enrichment analysis and found that 10 meta-analysis and 14 single-cohort species–phenotype association pairs showed significant enrichment in high or moderate impact SNVs (Chi-squared test, False discovery rate (FDR) < 0.05; **Fig. 3e,f** and **Tables S13** and **S14**). In the meta-analysis associations, diastolic blood pressure associations showed the most significant enrichment of high or moderate impact SNVs in an *Agathobaculum* species (Species ID: 102287, the ID in the reference genome database of GT-Pro, see **Table S4**) (Chi-squared test, $P = 3.53\times10^{-6}$), followed by TG associations enriched in high or moderate impact SNVs in *Clostridium phoceensis* (Species ID: 101500) (Chi-squared test, $P = 6.12\times10^{-6}$). Age-related associations supported by multiple cohorts were significantly enriched in high or moderate impact SNVs in two species: a *Faecalibacterium* species

(Species ID: 102040) (Chi-squared test, P = 6.99×10^{-5}) and an *Anthropogastromicrobium* species (Species ID: 100193) (Chi-squared test, P = 1.17×10^{-3}). In the single-cohort associations, many associations related to exposure factors were significantly enriched in high or moderate impact SNVs (Chi-squared test, FDR < 0.05), including current exposures (current smoker) and dietary factors (daily carbohydrate intake, Lifelines diet quality score, daily alcohol intake, and daily calorie intake). Interestingly, two early life exposures (parent smoking during childhood and having pets before 15 years old) were also significantly enriched in high or moderate impact SNVs, indicating that the long-term impact of early life exposures also happens at the level of gut microbial genetics.

To provide insight into the functional mechanism behind these associations, we focused on the associations that involved SNVs located in gut microbial genes with known function, which accounted for 67.0% of the total meta-analysis associations and 45.5% of single-cohort associations (**Fig. S10a,b**). We then counted the number of associations involved in different enzyme classes at Enzyme Commission (EC) number level 2 (subclass), level 3 (sub-subclass), and level 4 (substrate classes) (**Fig. 3g,h**). For the associations supported by multiple cohorts, SNVs in genes that produce enzymes belonging o transferases transferring phosphorus-containing groups (EC: 2.7.-.-) showed the highest number of associations (180), followed by hydrolases acting on acid anhydrides (EC: 3.6.-.-) with 150 associations, and ligases forming carbon-oxygen bonds (EC: 6.1.-.-) with 131 associations (**Fig. 3g**). All the associations assigned to ligases forming carbon-oxygen bonds (EC: 6.1.-.-) are related to aminoacyl-tRNA synthetases (aaRSs) (EC: 6.1.1.-) (**Fig. 3g**), which showed the highest number of associations at sub-subclass level (level 3). At substrate level (level 4), SNVs in the isoleucine--tRNA synthetase (ileRS, EC: 6.1.1.5) gene showed the highest number of associations (**Fig. 3g**), including with sex, TG, HDL-C, and BMI (**Table S11**). Similarly, for the single-cohort associations, microbial genes encoding transferases transferring phosphorus-containing groups (EC: 2.7.-.-) and ligases forming carbon-oxygen bonds (EC: 3.6.-.-) showed the highest number of associated SNVs (**Fig. 3h**). Overall, the SNV-MWAS catalog underscored the impact of gut microbial SNVs on enzymes responsible for key biochemical processes and provided valuable clues for downstream functional studies at strain-level.

Microbial genetic adaptation with age

Meta-analysis identified 718 signals between host age and gut microbial SNVs at species-wide significance level ($P_{meta} < 6.47\times10^{-6}$; **Table S11**), including 131 missense or stop-gain SNVs and 587 synonymous or upstream gene SNVs. Of the 131 age-associated missense or stop-gain SNVs, six were associated with age in at least four

cohorts, with four of the six located in the *sarA_2* gene of *Faecalibacterium prausnitzii*_I (species ID: 102274). In the genome of *Faecalibacterium prausnitzii*_I, 22 SNVs were associated with host age ($P_{meta} < 6.47\times10^{-6}$; **Fig. 4a,b** and **Table S11**). The leading signal was from a missense SNV, 102274_1617604, at the *sarA_2 gene* (Gene ID: 102274_01456, $P_{meta} = 1.99\times10^{-22}$). In *sarA_2*, seven SNVs were associated with host age at species-wide significance level ($P_{meta} < 6.47\times10^{-6}$), and linkage disequilibrium (LD) between the leading SNV, 102274_1617604, and the other six SNVs was less than 0.2 (**Fig. 4c**). Strikingly, six out of the seven SNVs associated with host age at *sarA_2* were missense SNVs, with three associations significant in four cohorts and one significant in five cohorts (**Fig. 4d**). None of these associations were significant in 300OB, possibly because all 300OB participants are older than 60 years. In fact, the 300OB cohort showed the highest frequency of age-associated alleles of the five Dutch cohorts (**Fig. 4e,f**), thus it also demonstrated the enrichment of the age-associated allele in the *sarA_2* gene of *Faecalibacterium prausnitzii*_I in an older population. Individuals with and without the age-associated alleles in *sarA_2* gene showed differences in chronological age across the Dutch and Tanzanian cohorts (**Fig. 4d** and **g**), suggesting convergent genetic adaptation during host aging.

The *sarA_2* gene is a homologue of the *oppA* gene that encodes oligopeptide-binding lipoprotein, a membrane protein belonging to the ABC-type transporter system responsible for oligopeptide or oligosaccharide intake [33]. The *sarA_2/oppA* gene is in a gene cluster involved in oligopeptide transportation (**Fig. 4h**) and forms an out-membrane unit of the oligopeptide transporter complex, thus playing an important role in oligopeptide binding and transportation. To infer the impact of the age-associated missense SNVs on the function of the sarA_2/oppA protein, we predicted its 3D protein structure using AlphaFold3 [34]. This revealed that the missense SNVs cause amino acid changes in the cavity of sarA_2/oppA (**Fig. 4i,j**), which may influence oligopeptide recognition and binding affinity. Based on a structural comparative analysis with the oligopeptide transportation systems of *Mycobacterium tuberculosis* H37Rv (PDB: 8JA7) [35] and *Escherichia coli* (PDB: 6CVL) [36], the oligopeptide transportation system is a protein complex formed by five subunits encoded by the genes in the oligopeptide transportation gene cluster. To further understand the impact of the age-associated SNVs on the transportation capacity of sarA_2/oppA, we assembled all five subunits of the ABC-type transporter system based on the AlphaFold3-predicted protein structures of sarA_2/oppA and the other four subunits encoded by the *oppB*_4, *oppC*_4, *oppD*_4, and *oppF*_4 genes in the ABC-type transporter gene cluster (**Fig. 4k**). The predicted structure showed a tunnel-like cavity surrounded by the subunits, with the sarA_2/oppA protein acting as a gate for the transportation tunnel. Interestingly, the age-associated mutations were located in the cavity of sarA_2/oppA that is connected with the tunnel, which suggests that these mutations can modify the

transportation function of the protein complex. We also calculated a conservation score for the residues of sarA_2/oppA and found that all six age-associated SNVs lead to a shift of amino acid composition from major amino acids (1st rank in the amino acid usage in the corresponding position) to minor amino acids (2nd or 3rd rank in the amino acid usage) (**Fig. 4l**). This suggests that the shift of these amino acids may be related to a switch in protein function from main to alternative function.

To further understand whether the missense mutations in *oppA* of *F. prausnitzii*_I help the microbe to adapt to selective pressure within the intestine as the host ages, we calculated the rates of nonsynonymous (pN) and synonymous (pS) polymorphism and their ratio (pN/pS) in the gene for all samples. This ratio reflects overall direction and magnitude of selection acting on a single gene. In the DMP cohort, we assigned samples into quartiles according to age, assigning the first quantile (Q1) to the youngest group and the fourth quantile (Q4) to the oldest group. Within the young and old groups, respectively, we used the median pN/pS of all genes of *F. prausnitzii*_I as background and examined the quantile of median pN/pS of *sarA_2*/*oppA* in the background. We found that the quantile of median pN/pS of *sarA_2*/*oppA* was higher in the older group (Q4) compared to the young group (Q1), and we observed similar trends in LLD1 and 300TZFG (**Fig. S11**). This indicates that intestinal environmental changes related to host aging exert positive selection on the *sarA_2*/*oppA* gene of *F. prausnitzii*_I. Moreover, the presence of the age association signals across different populations indicates convergent adaptation of gut microbes within the host under common environmental pressure.

Gut microbial SNVs predict host age

As host age showed the highest number of associations with gut microbial SNVs, we constructed a machine-learning model to determine how well microbial SNVs could predict host age (see **Methods**). The DMP cohort was used as the training dataset. To ensure model robustness and generalizability, we conducted nested cross-validation and found that the gut microbial genotypes can predict host chronological age with a mean R^2 of 41.9% across 25 rounds of tests (**Fig. 5a**). We then selected 1,180 gut microbial SNVs as the final prediction features according to their mean importance across the nested cross-validation (see **Methods**; **Fig. 5b** and **Table S15**) and trained and fit the final model (**Fig. S12**). We first tested the model's performance in an independent dataset of 771 samples (500FG-FSK and LLD2-FSK) not included in the training process, and the model exhibited robust performance (R^2 = 33.9%) (**Fig. 5c**). To test if the model could also be applied to metagenotype datasets derived using a different DNA extraction kit, we tested the model's performance in 1,135 samples (LLD1) extracted with a different kit, and the model retained a degree of predictive

power (R^2 = 9.8%) (**Fig. 5d**). Although the model's performance dropped dramatically here, it is plausible that 282 of the 1,180 SNV features were not detected in LLD1 because the type of DNA isolation kit used impacts microbiome composition. To further validate the genotype-based age prediction model, we compared the genotype-predicted age in longitudinal samples (LLD1 and LLD2-APK) and found significantly higher predicted ages in the samples collected at the follow-up time point ~4 years later (LLD2) compared to the predicted age of the baseline time point (LLD1) (One-tailed paired Wilcoxon rank sum test, P = 4.45×10^{-4}, **Fig. S13**). We assessed the importance of the SNVs for age prediction using SHAP value, and found that the SNV 102274_1617604 in *sarA_2/oppA* gene was also among the top features with high predictive importance (**Fig. 5e**). We further found that the gut microbial SNV-predicted age (referred to as gmSNVage hereafter) is not associated with other biological age indices derived from host DNA methylome [37–39], telomere length, and T cell maturation [40] (see **Methods**; partial Spearman's correlation test, P > 0.05; **Fig. 5f** and **Supplementary Note 4**), indicating that gmSNVage may reflect unique aspects of functional senescence that are not captured by other biological age markers.

To better understand how gmSNVage reflects host physiological dysfunction during the aging process, we associated the gmSNVage with 131 intrinsic host phenotypes in LLD1, after correcting for chronological age using partial correlation. Here we found that 13 phenotypes showed nominally significant correlation with gmSNVage, independent of chronological age (partial Spearman's correlation test, P < 0.05; **Fig. 5g** and **Table S16**). The serum level of several gut-microbiota-derived metabolites showed significant correlation with gmSNVage, including three short-chain fatty acids (SCFAs) (propionate, butyrate, and acetate) and three BAs (lithocholic acid (LCA), glycoursodeoxycholic acid (GUDCA), and ursodeoxycholic acid (UDCA)). Interestingly, the direct correlations between gmSNVage and the SCFAs were higher than the chronological age correlations (Spearman correlation; **Fig. 5g–k** and **Table S16**), which suggests that gmSNVage better reflects the decrease of gut-microbiota-derived beneficial metabolites that are related to the regulation of metabolic homeostasis. To gain a holistic insight into the overall metabolic changes alongside the gmSNVage, we further linked gmSNVage to the levels of 1,183 blood metabolites in LLD1, after correcting for chronological age and other covariates (see **Methods**). This identified 68 and 72 blood metabolite levels positively and negatively associated with gmSNVage, respectively, independent of chronological age (Partial Pearson's correlation test, FDR < 0.05; **Fig. 5l** and **Table S17**). In line with the metabolite levels measured by the targeted approach, SCFAs including acetic acid and propionic acid showed a stronger negative correlation with gmSNVage compared to chronological age (**Fig. 5m** and **Table S18**). Several other beneficial components, including the essential amino acid L-phenylalanine and dethiobiotin (which exhibits biotin-like

activity [41]), also showed stronger negative correlations with gmSNVage (**Fig. 5m** and **Table S18**).

Gut microbial SNVs associate with host health

To zoom in on the gut microbial SNVs associated with host health-related phenotypes, we prioritized 294 meta-analysis associations involving high or moderate impact SNVs and 64 intrinsic phenotypes (excluding age and sex), with the highest number of signals coming from TG and body weight (**Table S11**). TG showed the most significant associations with missense SNVs in the *ileS* gene (Gene ID: 101500_01181) of *C. phoceensis* (Species ID: 101500) (P_{meta} = 2.32×10^{-16}; **Fig. 6a–c**), with the top signals supported in both DMP and LLD1 (**Fig. 6d–i**). Although the most significant association between TG and SNVs of *C. phoceensis* in LLD1 did not reach the species-wide significance cutoff, the leading signal peak also falls in the *ileS* gene region (**Fig. 6g–i**). Among 47 SNVs in the *ileS* gene of *C. phoceensis*, 25 showed significant meta-analysis associations, and significantly enriched missense SNVs (Fisher's exact test, P = 0.017) that effect sizes in DMP and LLD1 are shown in **Fig. 6j**. The individuals with the alternative alleles in the *ileS* gene of *C. phoceensis* showed a significantly lower blood TG level (**Fig. 6j,k**), suggesting that these alternative alleles have a potential beneficial impact. Notably, the SNVs of the *ileS* gene of *C. phoceensis* showed pleiotropic associations with many health-related phenotypes, including HDL-C level and BMI ($P_{meta} < 6.47\times10^{-6}$; **Fig. 6l**). In the DMP, the missense SNVs of the *ileS* gene also showed single-cohort associations with general health and several exposure factors, including diet quality scores and income level ($P_{DMP} < 6.47\times10^{-6}$; **Table S12**). The *ileS* gene encodes an isoleucine-tRNA synthetase (IleRS) that is a member of the aminoacyl-tRNA synthetase (aaRS) family. aaRSs are essential enzymes that attach amino acids to corresponding tRNAs, with the charged tRNAs transferring the attached amino acids to growing peptide chains during the translation process. Interestingly, TG level also showed associations with missense SNVs of two other aaRSs of *C. phoceensis*: cysteine-tRNA synthetase (CysRS) and aspartate-tRNA synthetase (AspRS) (**Table S11**). Gut microbial aaRSs can be secreted and execute non-translational functions [42]. To understand the potential impact of the SNVs on the aminoacyl-tRNA synthetases, we used AlphaFold3 to predict the protein structures of the IleRS, CysRS, and AspRS of C. *phoceensis*. In IleRS, we found that the amino acid changes caused by the associated missense SNVs are located in the surface region of the editing domain (**Fig. 6m,n**), suggesting that the TG-associated residuals may be related to the interaction of IleRS with other molecules.

Although the small sample size of 300TZFG limited our ability to replicate most associations discovered in the Dutch cohorts, we did observe 52 intrinsic-phenotype-

related SNV associations (excluding age and sex) that were supported by both the Dutch and Tanzanian cohorts ($P_{meta} < 6.47\times10^{-6}$, $P_{300TZFG} < 0.05$, and $P_{Dutch} < 0.05$; **Table S11**). We identified five body-weight-associated gut microbial missense SNVs in five different genes encoding 30S ribosomal protein S1, 30S ribosomal protein S2, GTP cyclohydrolase 1 type 2, multidrug resistance protein MdtA, and an uncharacterized protein (**Table S11**). GTP cyclohydrolase I is an enzyme involved in the biosynthesis pathway of folate (vitamin B9), the serum level of which is negatively associated with body weight and central adiposity [43].

In the single-cohort associations of the DMP, we identified three missense SNVs in the *clpB* gene of *Alistipes ihumii* (Species ID: 101429) that showed associations with BMI, TG, waist circumference, body weight, and WHR ($P_{DMP} < 6.47\times10^{-6}$; **Fig. S14a–e** and **Table S12**). These three SNVs were fully linked ($R^2 = 1$), i.e., they showed identical association effect sizes and P-values and caused changes in the amino acid residuals 700 and 701 from NK to SQ (**Fig. S14f**). The product of the *clpB* gene is a chaperone protein belonging to class 1 AAA^+ proteins that shows disaggregase activity [44]. Strikingly, bacterial clpB has been identified as an antigen-mimic of α-melanocyte-stimulating hormone (α-MSH), an endogenous peptide involved in regulation of host satiety. As one commercial strain of *Hafnia alvei* was previously demonstrated to reduce appetite and body weight through clpB [45], we compared body weight between individuals without *A. ihumii*, with non-mutated *A. ihumii*, and with mutated *A. ihumii.* Here we found that only individuals with mutated *A. ihumii* showed significantly lower body weight (Kruskal-Wallis rank sum test, $P = 1.70\times10^{-8}$; **Fig. S14g**) and lower daily calorie intake (one-sided t-test, P = 0.01) compared to individuals with non-muated *A. ihumii*. This suggests that *A. ihumii* may also play a role in host satiety and body weight regulation through the mutated clpB protein.

Gut microbial SNVs mediate exposure impact on host health

The highly variable gut microbial genomes are under selection via environmental exposures, and they adapt to survive through fitness-increasing point mutations. In total, we identified 258 meta-analysis associations and 16,975 single-cohort associations between gut microbial SNVs and environmental exposures (**Tables S11** and **S12**). Diet is one of the dominant factors shaping the gut microbial community, and we observed 129 meta-analysis associations between 11 dietary factors and SNVs of 50 species ($P_{meta} < 6.47\times10^{-6}$; **Table S11**), with 25 associations supported by both the Dutch and Tanzanian cohorts ($P_{meta} < 6.47\times10^{-6}$, $P_{300TZFG} < 0.05$, and $P_{LLD1} < 0.05$; **Table S11**). For example, fruit intake frequency was associated with an SNV in a serine hydroxymethyltransferase gene (Gene ID: 102509_00961) of a *Slackia* species (Species ID: 102509) that catalyzes the reversible conversion of glycine to

serine [46]. We also observed 6,405 single-cohort associations between gut microbial SNVs and dietary factors ($P_{meta} < 6.47\times10^{-6}$, $P_{300TZFG} < 0.05$, and $P_{LLD1} < 0.05$; **Table S11**). For example, a fat-restricted diet was associated with a missense SNV in a beta-galactosidase gene (Gene ID: 100039_02134; *lacZ_4*) of *F. prausnitzii* (Species ID: 100039) in DMP ($P_{DMP} = 9.74\times10^{-9}$). Smoking is another lifestyle factor influencing gut microbiome composition [22], with some gut bacteria known to degrade nicotine, reducing the nicotine concentration in the intestinal tract [47]. We identified 98 meta-analysis associations and 2,948 single-cohort associations between gut microbial SNVs and current smoking habit and passive exposure to smoking during early life (**Tables S11** and **S12**).

As an interface between the external environment and host, the gut microbiota can mediate the impact of environmental exposures on host health. Among the 14,205 gut microbial SNVs associated with exposure factors in the DMP, 992 are also associated with intrinsic host phenotypes. This allowed us to infer the mediatory role of gut microbial SNVs between exposures and host health factors. We conducted bidirectional mediation analysis to infer whether gut microbial SNVs mediated the impact of environmental exposures on health outcomes. As the DMP cohort has the largest sample size and largest number of available exposure factors, we focused on this cohort for the mediation analysis. We first identified 855 exposure–SNV–health linkage groups in which all three variables were correlated and then conducted bidirectional mediation analysis. In the first mediatory direction (Direction 1), we hypothesized that gut microbial genomes adapt to exposure factors and act as regulators that mediate the effects of exposure factors on host health, thus we treated gut microbial SNVs as mediators and the intrinsic host phenotypes as outcomes. In the second mediatory direction (Direction 2), we hypothesized that the internal environment of the host could mediate the effects of external environmental exposures on bacterial SNVs, thus we treated the host intrinsic phenotypes as mediators and the gut microbial SNVs as outcomes (**Fig. 7a**; Details see **Methods**). In total, we identified 390 groups of inferred *in silico* mediatory relationships, including 25 unidirectional mediation relationships in Direction 1, 81 unidirectional mediation relationships in Direction 2, and 284 bidirectional mediation relationships ($FDR_{mediation} < 0.05$ in Direction 1 or Direction 2; **Fig. 7b** and **Table S19**).

We further screened mediation relationships involving dietary factors and high or moderate SNVs in Direction 1. This led to a tripartite mediation network composed of 4 dietary factors, 18 gut microbial missense SNVs as mediators, and 7 host intrinsic phenotypes as outcomes ($FDR_{mediation} < 0.05$ in Direction 1; **Fig. 7c** and **Table S19**). Notably, 14 of the 18 mediator SNVs are located in the *ileS* gene of *C. phoceensis* that showed associations with TG (**Figures 6j** and **7c**). For example, SNV 101500_1146911 mediated the impacts of the Lifelines protein diet score and the

Lifelines diet score on general health, HDL-C level, and TG level (**Fig. 7c,d**), which suggest that the *ileS* gene of *C. phoceensis* may regulate host metabolic health and could thus be targeted through dietary intervention. Another missense SNV, 101356_1589924, located in the *dmdB* gene of *Lawsonibacter* sp000177015 (Species ID: 101356) mediated the effect of alcohol intake on waist circumference ($FDR_{mediation}$ < 0.05 in Direction 1; **Fig. 7c** and **e**). The *dmdB* gene encodes a subunit of 2,3-dimethylmalate dehydratase, which catalyze the conversion between (2R,3S)-2,3-dimethylmalate and dimethyl maleate. Dimethyl maleate is an oral administration drug for multiple sclerosis that upregulates the nuclear factor E2-related factor 2 (Nrf2) pathway important for adipocyte differentiation [48–51]. Another missense SNV, 100651_1999621, located in the *ptsA* gene of an *Oscillospiraceae* species (Species ID: 100651) mediated the effects of the Lifelines protein diet score and Lifelines diet score on urine creatinine level, which is related to kidney health ($FDR_{mediation}$ < 0.05 in Direction 1; **Fig. 7c** and **f**). *ptsA* encodes a transport system permease protein for phosphate, a uremic toxin that can accumulate in the body of patients with kidney disease [52]. Additionally, we observed that gut microbial SNVs mediated the impact of early life exposures on current health status. For instance, SNVs of the species GCA_000437915.1 (Species ID: 100933) mediated the impact of parental smoking during childhood on risk of heart rate complaints, migraine, and mental disease ($FDR_{mediation}$ < 0.05 in Direction 1; **Table S19**).

DISCUSSION

Over the last decade, both GWAS and metagenomics-based microbiome studies have enhanced our understanding of the roles of host genetics and the symbiotic microbiome in host health. However, the application of GWAS to metagenomic data remains limited. In this study, we took advantage of the growing number of large-scale, deeply phenotyped cohorts with gut microbial metagenomic sequencing data to conduct a PheWAS on 244 phenotype and exposures factors through gut microbial SNV-MWAS. This identified 3,018 meta-analysis associations between SNVs of 186 gut microbial species and 105 human phenotypes/exposures, and 34,927 significant associations from single-cohort analysis when phenotype–SNV pairs were only available for one cohort. All these associations are detailed in an online catalog (https://fulab.web.rug.nl/SNV-MWAS/).

We identified a large number of age-associated microbial SNVs, evidence that supports the gut microbiome as a hallmark of aging. We also identified lipid concentration- and BMI-associated SNVs in microbial genes, particularly an isoleucine-tRNA synthetase gene of *Clostridium phoceensis* and an α-Melanocyte-stimulating hormone-mimicking *clpB* gene of *Alistipes ihumii*. Mediation analysis also provided evidence that microbial genetic variations may be under the selection of environmental factors, as well as evidence that they may mediate the effects of diet on the host, e.g. through the isoleucyl-tRNA synthetase gene of *C. phoceensis*. These findings offer insights into the functional capacity of the gut microbiome that is still largely hidden in the fine resolution. They also reveal microbial adaptation and evolution in relation to host lifestyle and physiological status. This study thus not only provides a technical template for association studies based on gut microbial genetic variation, it yields fundamental resources and new hypotheses for functional studies on host–microbe interactions.

Many gut microbiome studies have primarily focused on compositional variations, providing much of our knowledge of the health value of the gut microbiome [22,32,53]. However, little is known about the relationships between human health and genetic variation of the gut microbiome, especially for SNVs. Previously, gut microbial SNV-MWAS was limited by several technical and resource barriers. As gut microbiome data is very sparse, with a high rate of missing data, gut microbial SNV-MWAS require large sample sizes to gain sufficient statistical power. Another challenge is multimapping of reads, which occurs when metagenomic reads map to multiple genomes of different microbial species. This issue can decrease the accuracy of metagenotyping and lead to spurious associations. Low coverage of low abundance species in metagenomic sequencing can hinder detection of SNVs, again reducing the statistical power of association studies. Additionally, the functions of most gut bacterial genes are still uncharacterized, and we lack powerful tools to help assess and interpret the

consequences of microbial SNVs. In this study based on large-scale, deeply phenotyped cohorts, we conducted metagenotyping using a k-mer searching and exact matching strategy that is more species-specific in the scenario of multi-alignment and more sensitive to low coverage species [6]. Deep-learning-based tools like AlphaFold then enabled us to explore the uncharacterized functions of bacterial genes and proteins [34,54] from the structural point of view, allowing us to demonstrate how these tools can assist in interpreting an SNV's impact on gene products.

Prior to this study, Zahavi et al had performed a pioneering study to explore the associations between BMI and gut microbial SNVs, based on metagenomic data. They discovered 17 independent associations that could be replicated in independent cohorts. However, their pilot study did not consider the within-species population genetic structure that may lead to the lineage effect and false discoveries in GWAS [14]. The genetic structures of gut microbial genomes are more complicated than human genomes due to the presence of long-range LD. Metagenomic sequencing is an untargeted method commonly used to collect the genetic information of a whole gut microbial community, but it can cause a high genotype missing rate, especially for low abundance species, when sequencing coverage is not high enough [19–21]. The construction of a population genetic structure based on metagenomic data, rather than on isolate assemblies, is also problematic due to the possibility of within-species strain mixture within the host individual [30,55]. Despite these difficulties, we adapted several solutions to make it feasible and plausible to construct and control population genetic structure in the metagenotype profiles. This allowed us to identify phylogenetically independent SNV association signals across genetically diverse strains. We calculated the kinship between samples based on the SNVs detected in both samples, with only the sample pairs with sufficient overlap in SNV detection retained for the construction of population genetic structure. Further, we calculated the quasi-phaseability of each species in each sample to assess if the sample is very likely dominated by one main strain within the corresponding species [56] and only kept samples dominated by one main strain for the association analysis.

The abundances of gut microbial taxa have been associated with the host's disease risk, and they have illuminated the importance of microbial diversity within a host's intestine [57,58]. However, the gut ecosystem is highly variable and sensitive to environmental exposures and changes in host status. By contrast, gut microbiome genetic variations, especially SNVs, have been found to be more stable than gut microbiome composition [9], with gut microbial strains of some taxa observed to be temporally stable within host individuals over 5 years [5,9,59]. We found that many early life exposure factors are with associated gut microbial SNVs, which may indicate long-term persistence of gut microbial strains that colonized the host gut during early life. The seeding of microbes in the infant gut mostly comes from maternal transmission,

677 living environment, food, and random transmission [60–64], and life exposures during
678 infancy exert the initial selection that allows seeding strains with favorable genotypes
679 to stably colonize [65]. The colonizing strains then mutate naturally, and early life
680 exposures further select strains with high fitness determined by specific genotypes [65,66].
681 The gut microbes adapted to or selected by early life exposures can further impact
682 long-term host health outcomes [67,68], and we observed gut microbial SNVs to mediate
683 the impact of early life exposures on host health, highlighting the crucial regulatory role
684 of gut microbial genetics in how early life exposures exert long-term health impact.
685 Unlike association studies based on host genetic variation, associations between
686 gut microbial genetic variation and host phenotypes could be bidirectional, making their
687 interpretation and mechanistic understanding of the identified associations more
688 challenging. To further understand the potential regulatory relationships behind the gut
689 microbial SNV associations, we leveraged external environmental exposure factors as
690 external variables to infer the mediatory role of gut microbial SNVs on intrinsic host
691 phenotypes. We found that the missense SNVs encoding IleRS mediate the impact of
692 diet on host health, which suggests an underexplored function of bacterial aaRS
693 beyond its role in translation. Growing evidence suggest that aaRSs do have biological
694 functions beyond translation [69–71]. In eukaryotes, aaRSs have evolved to have an
695 additional domain allowing functional versatility, and aaRSs have been associated to
696 signaling pathways related to cardiovascular development and immune response[70,71].
697 Thus far, studies of the non-translational roles of aaRSs have primarily focused on
698 eukaryotes, with little study on the role of microbial aaRSs in maintaining host
699 homeostasis. A recent study reported that the threonyl-tRNA synthetase (ThrRS)
700 secreted by *A. muciniphila* can activate host anti-inflammatory signaling to maintain
701 immune homeostasis [42], representing a first step toward understanding the non-
702 translational function of gut microbial aaRSs in regulating host health. The causal links
703 we inferred between SNVs in *ileS* and other aaRS genes of gut microbes also provide
704 more clues about the non-translational role of microbial aaRSs, and they could be a
705 promising target for downstream experimental follow-up.
706 We also identified gut microbial SNVs associated with host body weight in the *clpB*
707 gene, a gut microbial genetic determinant known to regulate host body weight [72–76].
708 Gut microbial ClpB significantly impacts host body weight by mimicking α-MSH (alpha-
709 melanocyte-stimulating hormone) [44,72,74,76,77], which is involved in regulating energy
710 balance and appetite. ClpB produced by certain gut bacteria can interact with
711 melanocortin receptors in the hypothalamus, mimicking or influencing the effects of α-
712 MSH, leading to reduced appetite and increased energy expenditure [44,74,77]. Previous
713 studies have shown that ClpB-producing bacteria reduce food intake and body weight
714 gain in rodents, and individuals with obesity often have lower levels of these bacteria
715 compared to lean individuals [75]. This has led to the exploration of probiotic therapies

containing ClpB-producing strains to regulate body weight in people with obesity or metabolic disorders [78]. Our finding identified a potential probiotic strain containing ClpB with the key amino acid residuals essential to α-MSH-like activity [79], although further *in vivo* research is needed to validate the impact of *A. ihumii* on host appetite and body weight. Many previous studies have highlighted the impact of gut microbes on host physiology through metabolic regulation, including biosynthesis and metabolism of beneficial or toxic metabolites[80,81]. However, the more diverse gut microbial molecular pathways regulating host physiology are not yet fully understood, and the genetic associations we identified provide clues about the potential impact of gut bacteria on the host through molecular mimicry [82].

Host aging is a complex, multifaceted process characterized by a variety of biological hallmarks, including genomic instability, telomere attrition, epigenetic alterations, loss of proteostasis, deregulated nutrient sensing, mitochondrial dysfunction, cellular senescence, stem cell exhaustion, and altered intercellular communication [83,84]. Recent studies have proposed gut microbiome dysbiosis as a new hallmark of aging [83,84] characterized by loss of commensal or keystone taxa, loss of stable community structure, and rise of pathobionts [85]. Delayed transition toward the age-related microbiome and increased gut microbiome composition uniqueness in host individuals have been identified as features of healthy aging [85,86]. Despite the recognition of gut microbiome dysbiosis in aging, the role of gut microbial genetic variations in the aging process has not been thoroughly described. Here, we discovered a large number of gut microbial SNVs associated host age that were consistently significant in both Dutch and Tanzanian cohorts. This finding underscores the potential universality of these microbial genetic changes across different populations and environments. We further constructed a prediction model that leverages the identified SNVs to predict host biological age, providing a novel tool to study the aging process. The age-associated gut microbial SNVs could arise from several factors. Selection pressures accumulated over lifespan from exposures like diet, medication, smoking, and other factors are likely contributors. Additionally, the immune system, which tends to become less efficient in its host defense and immunosurveillance function and more pro-inflammatory with age, can influence the genetic landscape of gut microbes, leading to the emergence of specific aging-associated SNVs. These age-associated gut microbial SNVs reflect the genetic adaptation of gut microbes during host aging, and they in turn further impact host metabolism and health. For instance, we observed that the age predicted by gut microbial SNVs was associated with a decline in beneficial metabolites including SCFAs, necessary amino acids, and vitamins, and it was independent of chronological age, highlighting gut microbial genetic variations as a new dimension of biological age.

Limitations of the study

We acknowledge several limitations of the current study. First, the construction of the metagenotyping panel of GT-Pro was based on the Unified Human Gastrointestinal Genome (UHGG) database [6,87]. However, the UHGG metagenome collections did not include assemblies from the DMP, the largest cohort in this study, which may mean the loss of Dutch-population-specific genomic regions and the ability to detect SNVs in these regions. In the core region shared between multiple populations, the Dutch population may also contain population-specific SNVs. Therefore, the panel-based metagenotyping we used may have impacted our ability to identify novel SNVs. Secondly, the reference panel mainly focuses on SNVs in core genome regions [6]. Considering that presence/absence of non-core-regions reflects another level of genetic information, SVs or copy-number variations, the SNVs in non-core-regions should also be considered valuable information. Thirdly, although we controlled for population structure by including a genetic kinship matrix in an LMM, the inclusion of candidate SNVs in the calculation of population structure may decrease statistical power due to “double fitting” of the candidate SNV in the model, an issue called proximal contamination [88]. The leave-one-chromosome-out scheme is the current popular solution to control the impact of proximal contamination in human GWAS [89]. However, as gut bacterial genomes only contain one main chromosome and long-range LD between distant SNVs is common in bacterial genomes [20], there is currently no efficient solution to control for proximal contamination for microbial GWAS. This calls for development of algorithm and tools specifically designed for microbial GWAS. Fourthly, this study mainly involved samples from the Netherlands. Although we included a Tanzanian cohort to test for replication of associations in a genetically and geographically distant population, its smaller sample size limited the power of the replication analysis. In future, we should integrate samples with diverse genetic and lifestyle backgrounds to enhance gut microbial SNV-MWAS in the context of multi-ancestry. On a positive note, we have provided catalogs of associations linking gut microbial SNVs to host phenotypes and exposures that may inspire new hypotheses for downstream functional study on human-microbe interaction. Culturomics and genetic engineering for the targeted generation of SNVs in gut microbial genomes will assist comparitive studies of the gut bacterial strains with different genotypes [90–92].

Supplemental Information

Supplementary Information is available for this paper.

Acknowledgements

We thank all the volunteers in the Lifelines cohort (https://www.lifelines.nl/) and the Human Functional Genomics Project (http://www.humanfunctionalgenomics.org/) for

their participation and the project staff for their help and management. We thank Kate Mc Intyre for critical reading and editing. We also thank the Genomics Coordination Center for providing data infrastructure and access to high-performance computing clusters. This study is supported by Netherlands Organization for Scientific Research (NWO)-VICI grant VI.C.202.022 (J.F.), NWO-VIDI grant 016.178.056 (A.Z.), NWO-VENI grant 222.016 (D.W.), NWO-VENI grant 194.006 (D.V.Z.), European Research Council (ERC)-Consolidator grant 101001678 (J.F.), ERC Starting Grant 715772 (A.Z.), Dutch Heart Foundation grant IN-CONTROL II (CVON2018-27 to J.F., A.Z., M.G.N., L.A.B.J., J.H.W.R, and N.P.R), and ZonMW ME/CFS research program (10091012110015 to A.Z. and 10091012110017 to A.K.). In addition, C.W. and J.F. are supported by the Netherlands Organ-on-Chip Initiative, an NWO Gravitation project (024.003.001) funded by the Ministry of Education, Culture and Science of the government of the Netherlands. J.F. is also supported by the AMMODO Science Award 2023 for Biomedical Sciences from Stichting Ammodo. J.F. and A.Z. are supported by the NWO Knowledge and Innovation Covenant grant (KICH1.LWV04.21.01). A.Z. and A.K. are further supported by the NWO Gravitation grant Exposome-NL (024.004.017). A.Z. is also supported by the EU Horizon Europe Program grant INITIALISE (101094099). S.S. is supported by the NextGeneration EU grant Project Age-It “Ageing Well in an Ageing Society” (DM MUR 1557 11.10.2022), ERC Starting Grant 2022 (101075624), and a NutrAGE grant (DM MUR 844 16.07.2021). R.K.W. is supported by the Seerave foundation and NWO. H.P is supported by a joint fellowship from the University Medical Center Groningen and the CSC with grant number CSC202208060107. The 300TZFG cohort received funding from the Joint Programming Initiative, A Healthy Diet for a Healthy Life (JPI-HDHL; project 529051018, TransMic) and ZonMw (the Netherlands Organization for Health Research and Development). Figures 1a was generated using BioRender.com, with publication licenses XX263AR9Z1.

Author Contributions

J.F. coordinated and supervised the study. J.F., D.W., S.A.-S., and A.K. conceptualized the study. D.W. conducted data generation, harmonization, and integration. D.W., S.A.-S., and D.V.Z. performed data analysis: D.W. for gut microbial SNV analysis, S.A.-S. for quasi-phaseability assessment, and D.V.Z. for human genetics analysis. R.G. processed metagenomic sequencing data in the Dutch Microbiome Project. H.P. and D.W. designed and constructed the database website. A.J.R.-M. and D.W. conducted protein-structure-based analysis. C.W., J.F., and A.Z. set up the Lifelines-DEEP cohort. C.W., J.F., A.Z., and R.W. set up the Dutch Microbiome Project. L.A.B.J., N.P.R., J.H.W.R., and M.G.N. set up the 500FG and 300OB cohorts. G.S.T., V.I.K, R.J.X., and Q.d.M. set up the 300TZFG cohort. D.W.

and J.F. drafted the manuscript. All authors reviewed and edited the manuscript.

Competing Interests

R.K.W. acted as a consultant for Takeda, received unrestricted research grants from Takeda, Johnson & Johnson, Tramedico and Ferring, and received speaker fees from MSD, Abbvie, and Janssen Pharmaceuticals. A.Z. received a speaker fee from Nestle. All other authors declare no competing interests.

Data and Code Availability

The raw metagenomic sequencing data of all cohorts are publicly available. Four are deposited at the European Genome–Phenome Archive: Dutch Microbiome Project (accession number EGAS00001005027), Lifelines-DEEP (accession number EGAD00001001991), 300-Obesity (accession number EGAD00001005083), and 1000IBD (accession number EGAD00001004194). The 500FG data is available at NCBI SRA under accession number PRJNA319574. The metagenomic data of 300TZFG is available in the NCBI BioProject under accession number PRJNA686265. The code used for the statistical analysis is available via https://github.com/GRONINGEN-MICROBIOME-CENTRE/SNV_MWAS/.

Ethical Approval

The Lifelines study was approved by the ethics committee of the University Medical Center Groningen (METc2007/152). All participants signed an informed consent form prior to enrollment. Additional written consents were signed by the DMP participants or legal representatives for children aged under 18 years. The LLD study was approved by the Institutional Ethics Review Board of the University Medical Center Groningen (UMCG) (ref. M12.113965), the Netherlands. The 300-OB study was approved by the Institutional Review Board (IRB) CMO Regio Arnhem-Nijmegen (nr. 46846.091.13). The 500FG study was approved by the Ethical Committee of Radboud University Nijmegen (NL42561.091.12, 2012/550). The inclusion of volunteers and experiments were conducted according to the principles expressed in the Declaration of Helsinki. All volunteers gave written informed consent before any material was taken. The 1000IBD study was approved by the IRB of the UMCG, Groningen (ref. M12.1139652008.338.), the Netherlands. The 300TZFG study was approved by the Ethical Committees of the Kilimanjaro Christian Medical University College (CRERC) (no. 936) and the National Institute for Medical Research (NIMR/HQ/R.8a/Vol. IX/2290) in Tanzania.

Main Figure Legends

Fig. 1. Overview of study design and microbial SNV data after filtering.

a, Workflow of this gut microbial SNV-based genome-wide association study (SNV-MWAS) on human phenotypes and environmental exposures. In this study we integrated whole-genome shotgun sequencing (WGS) data of 11,831 fecal samples from 10,781 individuals from six cohorts and harmonized 244 phenotype and exposure factors across these cohorts. We then conducted meta-genotyping for the WGS data from stool samples and generated the gut microbial genotype profile, keeping more than 12.7 million common biallelic microbial SNVs for the SNV-MWAS. The gut microbial SNVs associated with host phenotypes were used to construct a prediction model for host phenotype.

b, Overview of the 244 phenotype and exposure factors in this study. The proportions of samples with a non-missing value for the corresponding factor in each cohort are shown in the heatmap (color indicates available proportion). Each row represents a factor. Each column represents a cohort.

c, Numbers of gut microbial species with at least 1,000 common SNVs in each cohort.

d, Total number of common gut microbial SNVs kept for each cohort.

e, Distributions of common SNV numbers of each species in each cohort.

f, Proportion of common SNVs in the core region of each species in each cohort.

Inner boxplots of (**e**) and (**f**) represent summary statistics. The center line represents the median, the box hinges represent the lower and upper quartiles of the distribution, whiskers extend no further than 1.5× interquartile range from the hinges, and data beyond the end of the whiskers are outliers plotted as individual points.

Fig. 2. Summary of SNV-MWAS signals.

a, Number of meta-analysis associations with gut microbial SNVs for each phenotype at different significance levels, including species-wide ($P_{meta} < 6.47\times10^{-6}$), metagenome-wide ($P_{meta} < 7.34\times10^{-9}$) and study-wide ($P_{meta} < 3.01\times10^{-11}$) significance.

b, Heatmap showing the association signals between all gut microbial SNVs and host phenotypes. Each dot is one association signal. Dot and column background colors indicate different phenotype classes.

c, Number of single-cohort associations with gut microbial SNVs for each phenotype at species-wide significance ($P_{meta} < 6.47\times10^{-6}$).

Fig. 3. Functional properties of associated gut microbial SNVs.

a and **b**, Proportion of associated gut microbial SNVs with different impact levels on the gene product in meta-analysis associations (**a**) and single-cohort associations (**b**).

c and **d**, Proportion of high and moderate impact SNVs in phenotype–species pairs

with more than 10 meta-analysis associations (**c**) and single-cohort associations (**d**). Each column represents a species. Light blue bars in the background show the proportion of high or moderate impact common SNVs in the genomes of corresponding species to indicate the background rate of high or moderate impact SNVs. Dot color indicates the different phenotype classes. Each dot represents a phenotype–species association pair for the corresponding species. Y-axis indicates the proportion of high or moderate impact common SNVs in the associated SNVs of the corresponding phenotype–species pair. Dot size indicates the total number of associated SNVs of the corresponding phenotype–species pair.

e and **f**, Enrichment of high or moderate SNVs in meta-analysis associations (**e**) and single-cohort associations (**f**) for phenotype–species pairs with more than 10 associations. Blue indicates non-significant enrichment (Chi-squared test, FDR > 0.05). Red indicates significant enrichment (Chi-squared test, FDR < 0.05).

g, Number of meta-analysis associations involved in different enzyme classes at Enzyme Commission (EC) number level 2 (subclass), level 3 (sub-subclass) and level 4 (substrate classes). Enzyme classes with fewer than 20 associations are assigned to “Others”.

h, Number of single-cohort associations involved in different enzyme classes at EC number level 2 (subclass), level 3 (sub-subclass) and level 4 (substrate classes). Enzyme classes with fewer than 200 associations are assigned to “Others”.

Fig. 4. **Missense SNVs in the *sarA*/*oppA* gene of *Faecalibacterium prausnitzii* associate with host age.**

a, Manhattan plot showing the association between SNVs of *F. prausnitzii* (Species ID: 102274) and host age. Dot colors indicate the impact levels of gut microbial SNVs on gene product (blue = low, green = moderate, and red = high).

b, Q-Q plot of the P-values of associations between SNVs of *F. prausnitzii* (Species ID: 102274) and host age. Although the genomic inflation factor was 1.204, which is slightly higher than 1, the shape of the curve still suggest that the signals were not likely generated by randomness.

c, Expanded view of the association signal showing the low LD between the SNVs associated with host age in the *sarA_2* gene region (1,616,256–1,618,274 bp) of *F. prausnitzii* (Species ID: 102274). This suggests that the SNVs are independently associated with host age.

d, Heatmap of the significance of the associations between the seven SNVs and host age in six independent cohorts. $^{\#}$: $P < 6.47\times10^{-6}$; * : $P < 0.001$; . : $P < 0.05$.

e and **f**, Small bar plots show the frequency of age-associated alleles (**e**) and mean age (**f**) in six independent cohorts.

g, Violin plots show the difference in host age between samples with and without the

age-associated alleles of two SNVs (102274_1617484 and 102274_1617604) in six independent cohorts. Inner boxplots of (**g**) represent summary statistics. The center line represents the median, the box hinges represent the lower and upper quartiles of the distribution, whiskers extend no further than 1.5× interquartile range from the hinges, and data beyond the end of the whiskers are outliers plotted as individual points.

h, The oligo-peptide transportation gene cluster in the genome of *F. prausnitzii* (Species ID: 102274).

i and **j**, The predicted protein complex structure of the oligopeptide transporter formed by sarA/oppA, oppB, oppC, oppD, and oppF. (**i**) shows the oligopeptide transporter complex embedded into cell membrane. (**j**) shows the tunnel-like cavity space inside the oligopeptide transporter complex.

k, Zoomed-in view of the sarA/oppA subunit showing the amino acid residuals (red) affected by the missense SNVs associated with host age.

l, Probability of amino acids in the six age-associated oppA/sarA_2 residuals from the protein conservation analysis. Color indicates side-chain group chemistry type of amino acids, including acidic (red), basic (blue), hydrophobic (black), neutral (purple), and polar (green).

Fig. 5. **Host age predicted by gut microbial genotypes.**

a, The performance (R^2) distribution of gut microbial SNV-based age prediction models in 5x5 rounds of fivefold cross-validation. The vertical dashed line denotes the mean performance (R^2) of all rounds of cross-validation.

b, Selection of features by median importance across 25 rounds of nested cross-validation. The features are sorted by median importance, and we cumulatively sum the median importance values untill they reach 95% of the total sum of importance values (pink horizontal line). We then take the counted SNV features as the predictive features (from left to right up to the pink vertical line) for the final model.

c and **d**, Performance of final model for age prediction in the FSK test dataset (LLD2-FSK + 500FG-FSK, N = 711) (**c**) and the APK test dataset (LLD1, N = 1,135) (**d**).

e, SHapley Additive exPlanations (SHAP) for the final model showing the top 15 important features. The reference allele is added to the ID of gut microbial SNV features. Feature values are indicated as blue (low value) and red (high values).

f, Pairwise partial correlations between gut microbial SNV-predicted age and 10 biological age markers, correcting for chronological age. Blank or white cells indicate insignificant correlations (partial Spearman's correlation test, P > 0.05). Variables in each row (from top to bottom): three methylation-based age indicies (epigenetic age models developed by Hannum *et al.* [37], Weidner *et al.* [38] and Horvath *et al.* [39]); signal joint T cell receptor excision circles (sjTRECs) expression (CT values of a qPCR),

which indicates the maturation of T cells; and the telomere lengths of six blood cell types: lymphocytes, granulocytes, B-cells ($CD45RA^{+}CD20^{+}$), naïve T cells ($CD45RA^{+}CD20^{-}$), memory T cells ($CD45RA^{-}$) and NK-cells/fully differentiated T cells ($CD45RA^{+}CD57^{+}$); and gut microbial SNV-predicted age.

g, Associations between predicted age and host phenotypes adjusted for chronological age.

h and **j**, Associations between propionate level and predicted age and chronological age. Colors indicate if the correlations with phenotypes are higher for gut microbial SNV-predicted age (blue) or chronological age (purple).

i and **k**, Associations between acetate level and predicted and chronological age.

l, Associations between gut microbial SNV-predicted age and 1,183 blood metabolites, adjusted for chronological age. Dot color indicates super-classes of metabolites. Red horizontal dashed line denotes the significance cutoff (FDR < 0.05).

m, Metabolites show stronger correlation with gut microbial SNV-predicted age (blue) than with chronological age (purple) (Pearson's correlation test). * denotes significant correlations (Pearson's correlation test, FDR < 0.05).

Fig. 6. **Missense SNVs in the *ileS* gene of *Clostridium phoceensis* associate with blood TG level.**

a, **d**, and **g**, Manhattan plots showing the association between SNVs of *C. phoceensis* (Species ID: 101500) and blood TG level in the meta-analysis (**a**), DMP cohort (**d**), and LLD1 cohort (**g**). Dot color indicates the impact levels of gut microbial SNVs on gene product (blue = low, green = moderate, and red (very rare) = high).

b, **e**, and **h**, Q-Q plot of the P-values of associations between SNVs of *C. phoceensis* (Species ID: 101500) and blood TG level in the meta-analysis (**b**), DMP cohort (**e**), and LLD1 cohort (**h**).

c, **f**, and **i**, Expanded view of the association signal shows the low LD between the SNVs associated with blood TG level in the *ileS* gene region (1,145,770–1,148,577 bp) of *C. phoceensis* (Species ID: 101500) in the meta-analysis (**c**), DMP cohort (**f**), and LLD1 cohort (**i**).

j, Heatmap of the significance of the associations between the 15 missense SNVs in *ileS* and blood TG level in the DMP and LLD1 cohorts. The alternative alleles are added to the ID of gut microbial SNV. Cell colors show the effect size of the alternative allele. $^{\#}$: $P < 6.47\times10^{-6}$; * : $P < 0.001$; . : $P < 0.05$.

k, Violin plots show the difference in host age between samples with and without the TG-associated alleles of three SNVs (101500_1146911, 01500_1146916, and 01500_1146926) in DMP and LLD1. Inner boxplots of **k** represent summary statistics. The center line represents the median, the box hinges represent the lower and upper quartiles of the distribution, whiskers extend no further than 1.5× interquartile range

from the hinges, and data beyond the end of the whiskers are outliers plotted as individual points.

l, Heatmap showing the associations between the 15 TG-associated missense SNVs in *ileS* and HDL-C and BMI in meta-analysis. The alternative alleles are added to the ID of gut microbial SNV. Cell color indicates the effect size of the alternative allele. $^{\#}$: $P < 6.47\times10^{-6}$; * : $P < 0.001$; . : $P < 0.05$.

m and **n**, The predicted protein structure of the *ileS* in side view (**m**) and top view (**n**). The red amino acids show the positions affected by the missense SNVs associated with blood TG level.

Fig. 7. Mediation relationship inference between exposure factors, gut microbial SNVs, and intrinsic host phenotypes using bidirectional mediation analysis.

a, Framework of bidirectional mediation analysis between exposure factors, gut microbial SNVs, and intrinsic host phenotypes.

b, Number of inferred mediation relationships for Direction 1 (from gut microbial SNVs to intrinsic host phenotypes), Direction 2 (from intrinsic host phenotypes to gut microbial SNVs), and both.

c, Sankey diagram showing the inferred mediation relationship network of Direction 1.

d–**f**, Examples of mediation relationships between exposure factors, gut microbial SNVs, and intrinsic host phenotypes inferred by bidirectional mediation analysis. The beta coefficient and significance are labeled at each edge. The proportions of indirect effect (mediation effect) are labeled at the center of the ring charts.

METHODS

Cohort description

Dutch Microbiome Project

The Dutch Microbiome Project (DMP) consists of 8,719 individuals and is part of the Lifelines study, a multi-disciplinary prospective population-based cohort study that utilizes a unique three-generation design to examine health and health-related behaviors in 167,729 people living in the northern Netherlands. Lifelines employs a broad range of investigative procedures to assess the biomedical, socio-demographic, behavioral, physical, and psychological factors that contribute to health and disease, with a special focus on multi-morbidity and complex genetics [93]. Microbiome data generation for the DMP was described in Gacesa et al [22]. Briefly, fresh-frozen fecal samples were collected from participants of the Lifelines cohort. Microbial DNA was isolated using the QIAamp Fast DNA Stool Mini Kit (FSK, Qiagen) by the QIAcube automated sample preparation system (Qiagen). Metagenomic sequencing was performed at Novogene, China using the Illumina HiSeq 2000 sequencer.

Lifelines-DEEP cohort

Lifelines-DEEP (LLD) is a sub-cohort of Lifelines that was established to explore the risk factors of complex diseases. 1,539 individuals were included. In LLD, and multi-layers of omics data were collected at baseline in 2013 [23]. In the current study, high-quality metagenomic sequencing data were available for 1,135 individuals (474 males and 661 females). The average age of LLD participants was 45.04 years (18–81, SD = 13.60). The average BMI was 25.26 kg/m^2 (16.67–48.56, SD = 4.18). Microbiome data generation for LLD was described in Zhernakova et al [32]. In short, fresh-frozen fecal samples were collected and DNA was isolated with the AllPrep DNA/RNA Mini Kit (APK, Qiagen, cat. #80204). Sequencing was performed using the Illumina HiSeq platform at the Broad Institute, Boston. 1,135 metagenomic samples passed quality control.

Follow-up samples of LLD were collected for 338 individuals (150 males and 188 females) at a second time point. The duration between the two time points ranged from 3.33 to 3.92 years (mean = 3.53, SD = 0.12). At baseline, the mean age of participants was 48.2 years (18–80, SD = 11.7) and their mean BMI was 25.4 kg/m^2 (17.6–43.3, SD = 4.08). At follow-up, the mean age was 51.7 years (22–84, SD = 11.7), and the mean BMI was 25.6 (16.1–37.6, SD = 4.0). Two aliquots of each stool sample were processed and used for DNA isolation and library preparation using two kits, FSK (Qiagen) and APK (Qiagen, cat. #80204). WGS was performed on the Illumina HiSeq 2000 platform.

500 Functional Genomics cohort

The 500 Functional Genomics cohort (500FG) cohort is part of the Dutch Human Functional Genomics Project (DHFGP) and consists of 534 individuals. The average age of 500FG participants was 28.47 years old (18–75, SD = 13.69), and the average BMI was 22.75 kg/m^2 (15.09–35.43, SE = 2.85). The metagenomic data generation was described previously [24,25]. In short, DNA was isolated from fecal samples with the FSK (Qiagen) and APK (Qiagen, cat. #80204), and libraries were sequenced on the Illumina HiSeq 2000 platform.

300-Obesity cohort

The 300-Obesity (300-OB) cohort was established by Radboud University Medical Center, Nijmegen, the Netherlands [94]. In total, 298 older individuals (132 males and 166 females) with a high BMI >27 kg/m^2 were enrolled in 300-OB. The average age of 300-OB participants was 67.07 years old (54–81, SD = 5.39). The average BMI was 30.73 kg/m^2 (26.28–45.47, SD = 3.48). All participants were included between 2014 and 2016. The metagenomic sequencing of fecal samples was done using a similar protocol and analysis pipeline to LLD.

1000 Inflammatory Bowel Disease cohort

The 1000 Inflammatory Bowel Disease (1000IBD) cohort consists of patients with IBD recruited at the specialized IBD outpatient clinic of the University Medical Center Groningen in the Netherlands [26,95] IBD diagnosis was made based on accepted radiological, endoscopic, and histopathological evaluation. The 1000IBD cohort is 60.70% female and 39.30% male. The mean age of participants was 42.61 years old (18–82, SD = 14.69), and their mean BMI was 25.44 kg/m^2 (15–49.4, SD = 5.05).

300-Tanzania Functional Genomics cohort

For analysis among non-Europeans, we included 300-Tanzania Functional Genomics (300TZFG), a population cohort of 323 individuals from both rural and urban areas in the Kilimanjaro region of northern Tanzania. This study is part of the DHFGP. Metagenomic data generation was described previously [28]. In short, bacterial DNA was isolated using the APK (Qiagen), and libraries were sequenced on the Illumina NovaSeq 6000 platform. The average age of 300TZFG participants was 33.08 years old (17.76–66.43, SD = 11.25). The average BMI was 24.98 kg/m^2 (16.41–41.91, SD = 4.93).

Quantification of metabolites

Bile acid quantification

The levels of 15 BAs and C4 concentrations in fasting plasma were quantified by liquid chromatography-mass spectrometry procedures, as previously described [96,97]. In brief, for sample preparation, 25 µL plasma was mixed with 250 µL internal standard solution to precipitate proteins. Samples were centrifuged at 15800 × g, and the supernatant poured into a clean glass tube. The fluid was evaporated under nitrogen at 40 °C. Before measuring, samples were reconstituted in 200 µL 50% methanol in water. For the quantitative determination of BAs, we used a Nexera X2 Ultra High-Performance Liquid Chromatography system (SHIMADZU, Kyoto, Japan), coupled to a SCIEX QTRAP 4500 MD triple quadrupole mass spectrometer (SCIEX, Framingham, MA, USA) (UHPLC-MS/MS). The LC-MS/MS system is controlled by Analyst® MD 1.6.2 software. BAs were separated with an ACQUITY UPLC BEH C18 Column (1.7 µm, 2.1 mm × 100 mm) equipped with an ACQUITY UPLC BEH C18 VanGuard Pre-Column (1.7 µm, 2.1 mm × 5 mm) (Waters, Milford, MA, USA). Separation was achieved in 28 minutes using 10 mM ammonium acetate in 20% acetonitrile (mobile phase A) and 10 mM ammonium acetate in 80% acetonitrile (mobile phase B), flow 0.4 ml/min.

The proportions of 15 BAs (with suffix '_p') were calculated by dividing by total BA concentration. Additionally, eight different total BA concentrations and ratios were calculated (Chen *et al.*, 2020): (1) total BA (Total_BAs) = sum of all BA concentrations, (2) total primary BA (Total_primary_BAs) = sum of all primary BA concentrations, (3) total secondary BA (Total_secondary_BAs) = sum of all secondary BA concentrations, (4) ratio of Secondary BAs to primary BAs ratio (Secondary_primary_ratio) = Total_primary_BAs/Total_secondary_BAs, (5) ratio of CA to CDCA concentrations (CA_CDCA_ratio) = (CA + TCA + GCA)/(CDCA + TCDCA + GCDCA), (6) ratio of unconjugated BA to conjugated BA concentrations = (CA + CDCA + DCA + LCA)/(TCA + GCA + TCDCA + GCDCA + TDCA + GDCA + TLCA + GLCA), (7) ratio of dehydroxylated CA to deconjugated CA concentrations (CA_dehydro_deconju_ratio) = (DCA + TDCA + GDCA)/(CA + TCA + GCA) and (8) ratio of taurine conjugated BA to glycine conjugated BA concentrations (Taurine_glycine_ratio) = (TCA + TCDCA + TDCA + TLCA)/(GCA + GCDCA + GDCA + GLCA).

Short-chain fatty acid quantification

The stool level of SCFAs was measured using gas chromatography-mass spectrometry (GC-MS), as previously described [98,99].

Quantification of trimethylamine N-oxide and its precursors

Trimethylamine N-oxide and its precursors betaine, carnitine, deoxycarnitine, and choline were measured using LC-MS/MS, as previously described [100,101]. In brief, an online cleanup of the sample was carried out using an analytic column containing a C_{18} stationary phase, prior to its introduction into the mass spectrometer. This

stationary phase did not retain the analytes, but it did retain significant matrix interferences, like (phosphor-)lipids [102].

Untargeted blood metabolome quantification

Untargeted metabolic profiling of fasting plasma samples was conducted at General Metabolics using flow-injection time-of-flight mass spectrometry on an Agilent 6550Q-TOF system, as previously described [103–105].

Quantification of biological age parameters

Telomere length measurement

Telomere length was assessed using an automated multicolor flow-fluorescence in situ hybridization (Flow-FISH) technique as previously described [40,106]. In brief, white blood cells were obtained by lysing erythrocytes in whole blood with NH_4Cl. These white blood cells were then combined with bovine thymocytes, which have a known telomere length and serve as an internal reference. The cells underwent denaturation in formamide at 87° C, followed by hybridization with a fluorescein-labeled (CCCTAA)3 peptide nucleic acid probe specific to telomere sequences. They were then counterstained with the DNA dye LDS751. Flow cytometry was used to measure the fluorescence intensity in granulocytes, total lymphocytes, and lymphocyte subsets (identified by antibodies specific for CD20, CD45RA, and CD57). This fluorescence intensity, compared to internal control cells and unstained controls, was used to determine the median telomere length from duplicate measurements.

sjTREC measurement

Signal joint excision circles produced during T cell development (sjTREC) expression was quantified using qPCR, as previously described [40]. In brief, DNA was extracted from whole blood using Proteinase K digestion, phenol extraction, and isopropanol precipitation. We then used TaqMan qPCR on a ViiA™ 7 system (Life technologies) to quantify sjTREC expression. The qPCR was done on 75–100 ng of DNA in a 12 µl reaction mix with 18 µM of sjTREC primers (5′-TCGTGAGAACGGTGAATGAAG-3′ and 5′-CCATGCTGACACCTCTGGTT-3′) and albumin primers (5′-TGAACAGGCGACCATGCTT-3′ and 5′-CTCTCCTTCTCAGAAAGTGTGCATAT-3′) and 5 µM probes for sjTREC (5′-(FAM) CACGGTGATGCATAGGCACCTGC-3′ (TAMRA)) and albumin (5′-(FAM) TGCTGAAACATTCACCTTCCATGCAGA -3′ (TAMRA)). The PCR protocol was 50° C for 2 min, 95° C for 15 min, followed by 45 cycles of 95° C for 15 s and 60° C for 30 s. Reactions were done in duplicate. Fluorescence from probe cleavage was measured at the threshold cycle (CT). Valid results had a duplicate SD ≤ 1. Normalized sjTREC expression was calculated from the difference between average CT values of albumin and sjTREC.

Metagenomic sequencing and quality control

We removed host genome–contaminated reads and low-quality reads from the raw metagenomic sequencing data using KneadData (version 0.7.4), Bowtie2 (version 2.3.4.3) [107], and Trimmomatic (version 0.39) [108]. In brief, the data-cleaning procedure includes two main steps: (1) filtering out human genome–contaminated reads by aligning raw reads to the human reference genome (GRCh37/hg19) and (2) removing adaptor sequences and low-quality reads using Trimmomatic with default settings (SLIDINGWINDOW:4:20 MINLEN:50).

Quantification of reference genome abundance

We estimated the abundances of reference genomes in metagenomic samples using CoverM (version 0.7.0).

Metagenotyping

We characterized the SNVs for metagenomic samples using GT-Pro (version 1.0.1) based on the pre-constructed sck-mer database [6]. In brief, a catalog of 104 million biallelic SNVs in the core genomic region of 909 gut microbial species was initially constructed based on UHGG assemblies, the genotypes of the gut microbial SNVs were tagged by species-specific kmers, and a kmer-based gut microbiome metagenotype panel was constructed. GT-Pro directly conducted genotyping for gut microbiota from metagenomic sequencing data by kmer-searching and matching the genotype kmers of the metagenotype panel. The GT-Pro output was further converted to dosage genotype, indicating the proportion of alternative allele.

Quality control of the SNV profile

We performed quality control in each cohort separately, filtering out the SNVs with high missing rate (>90%) or low MAF (<0.05) to get the cohort-specific common SNVs. We then counted the number of SNVs retained for each species in each cohort, and removed species with a low common SNV number (<1000).

Annotation of microbial genes and SNVs

The protein-coding sequences of gut microbial reference genomes were predicted and annotated previously, and obtained from the UHGG assembly catalog. The functional effects of the common SNVs were annotated using SnpEff (version 5.1f) [31] and classified into impact levels according to their potential impact on the amino acid sequences of gene products as follows: high (stop-gained, start-lost, stop-lost, or splice region variant), moderate (missense, splice region or stop-retained, start-retained, or initiator codon variant) or low (upstream or downstream gene variant, synonymous variant).

Between-sample genetic relatedness and population genetic structure

For each species, we extracted the independent SNVs from the cleaned genotype dataset using bcftools (version 1.17) [109] with the parameter '*+prune -m r2=0.49 -w 10000bp*'. We then calculated the species-specific between-sample genetic dissimilarity $D_{i,j}$ measured by a Manhattan distance scaled by the number of shared SNVs within each sample pair using the *parallelDist*() function in R package *parallelDist* (version 0.2.6). Sample pairs sharing < 100 SNVs were skipped for the genetic dissimilarity calculation, and the corresponding $D_{i,j}$ were marked as missing values. Samples with missing $D_{i,j}$ values were then iteratively removed from the genetic distance matrices to maximize the size of the retained sample with no missing value. We further calculated the between-sample genetic relatedness, $R_{i,j}$ = 1 - $D_{i,j}$, and the matrix M_R containing $R_{i,j}$ was used as the representative of population genetic structure. We performed principal coordinate analysis on the genetic dissimilarity matrices M_D using the *pcoa()* function from R package *ape* (version 5.7-1) with the parameter *correction = "cailliez"*.

Linkage disequilibrium calculation

For each species, we randomly extracted 1000 SNVs from the cleaned genotype dataset using plink2 (version 2.00a3.3LM), with the parameters *--thin-count 1000*. We then calculated the LD as a squared correlation for all pairs of extracted independent SNVs using plink2 with the parameters *--r2 inter-chr yes-really --ld-window-r2 0*. We visualized the whole-genome LD structure using LDBlockShow (version 1.40) [110] with the parameters *-SeleVar 2 -NoShowLDist 100000000 -OutPdf -OutPng.*

Quasi-phaseability assessment

The co-existence of multiple strains of a species within the host may bias the position of the sample in the population genetic structure and decrease the power of association analysis. For a given species, samples in which >80% of the SNVs have the same allele in >90% of the reads were considered 'quasi-phaseable', meaning the sample is very likely dominated by one strain of the corresponding species. The criterium and concept of 'quasi-phaseability' were described [56,111].

SNV-based metagenome-wide association study and meta-analysis

Before the association analysis was conducted, the normality of continuous phenotypes was checked with a Shapiro-Wilk test, and most continuous phenotypes failed to meet the criteria of normal distribution (P > 0.05). Therefore, prior to the association analysis, we normalized all continuous phenotypes to follow a standard normal distribution (N~(0, 1)) using the empirical normal quantile transformation (EMQT) [112]. All phenotypes were then fitted with a LMM that included the kinship matrix

M_R of corresponding species as a random effect and age, sex, and clean read count as fixed-effect covariates, using EMMAX (version beta-07Mar2010) [113] with the parameters *emmax -d 10 -t genotype.file -p phenotype.file -c covariate.file -k kinship.file -o output -Z*. The association analyses were performed for each cohort and each species separately, with only the quasi-phaseable samples of corresponding species included for the association analysis. For the SNV–phenotype pairs available in two or more independent cohorts (DMP, LLD1, 500FG, 300OB, IBD and 300TZFG), a follow-up meta-analysis was performed in the independent datasets using the Metal software (version 2020-05-05) [114] with default options (weighting cohort-based P-values according to sample size). To ensure robustness and further decrease the chances of false discovery of identified associations, we conducted a permutational Spearman correlation test with 999 times permutation following the Monte Carlo procedure within each cohort to assess the robustness of significant meta-analysis associations and single-cohort associations using the *spearman_test()* function of R package *coin* (version 1.4-2). Regional association plots showing LD structures on the gene regions covering the association signals were generated using LDBlockShow (version 1.40) [110].

Determination of linkage significance

The threshold of study-wide significance was determined by the number of cohort–phenotype combinations and the total number of LD blocks in the reference metagenotype panel using the Bonferroni method: 0.05 / (244 phenotypes * 6,809,003 LD blocks) = 3.01×10^{-11}. Similarly, the metagenome-wide significance was determined without considering multiple cohorts and phenotypes: 0.05 / 6,809,003 LD blocks = 7.34×10^{-9}. We also adopted the species-wide significance as a suggestive significance threshold based on the average number of LD blocks per species in the reference metagenotype panel: 0.05 / 7,729 LD blocks = 6.47×10^{-6}.

Selective pressure

For a given gene in a given sample, we calculated the ratio of nonsynonymous to synonymous polymorphism rate (pN/pS) based on the read counts covering nonsynonymous and synonymous mutation sites. In each sample, genes with fewer than five mutations compared to reference genome were skipped for pN/pS analysis.

Protein 3D structure prediction and protein-structure-based analysis

The 3D structures of microbial proteins were modeled using the AI algorithm AlphaFold3 [54] and visualized in ChimeraX (version 1.8rc202405140121) and PyMOL (version 2.5.7). Detection of a channel in the oligopeptide transporter complex of *F. prausnitzii* was conducted using MOLE*online* (https://mole.upol.cz/, version 2.5) [115] in

the 'Pore mode' with the parameters *Probe Radium = 16* and *Interior Threshold = 0.3*. The conservation score of the sarA_2/oppA protein was determined using the online server of ConSurf (https://consurf.tau.ac.il/consurf_index.php, version 1.0) [116].

Gut microbial SNV-based age prediction model

We constructed a host age prediction model based on Light Gradient-Boosting Machine (lightGBM) implemented in R package *lightgbm* (version 3.3.5). To ensure generalizability of the model, we used *tidymodels* (version 1.1.0) to conduct nested cross-validation (CV) with a five times fivefold split for the outer loop and a fivefold split for the inner loop to assess the predictive power of gut microbial SNVs on host age. In all CV runs, we used a set of predetermined hyperparameters: *trees = 2000, tree_depth = 15, min_n = 30, sample_size = 0.6, mtry = round(0.2 * total_feature_number),* and *learning_rate = 0.01*. We estimated the importance of features using Shapley additive explanation (SHAP), which assigns each feature an importance value for a particular prediction. After 5 x 5 rounds of fivefold cross-validation, we calculated the median importance value of each SNV feature, sorted the SNV features by their median importance values from high to low, cumulatively summed the median importance values untill we reached 95% of the total sum of importance values, and took the counted SNV features as the predictive features for the final model.

Partial correlation analysis

The partial correlation test was conducted using the *pcor()* function of R package *ppcor* (version 1.1). For the partial correlations between gut microbial SNV age and other biological age indices and between gut microbial SNV age and host intrinsic phenotypes, we used partial Spearman's correlation, adjusting for chronological age. For the partial correlations between gut microbial SNV age and untargeted blood metabolome, we used a partial Pearsons correlation, adjusting for chronological age, sex, and blood creatinine level.

Bidirectional mediation analysis

Causal relationships between exposure factors, gut microbial SNVs, and host intrinsic phenotypes were inferred by bidirectional mediation analysis using R package *mediation* (version 4.5.0). To reduce the number of tests, prior to mediation analysis we identified exposure factors–gut microbial SNV–intrinsic host phenotype linkage groups in which all the variables correlated with each other as candidate groups with a potential causal relationship. A candidate group had to meet the following criteria: (1) the association between the intrinsic host phenotype and gut microbial SNV is significant in the DMP ($P_{DMP} < 6.47\times10^{-6}$) or meta-analysis ($P_{meta} < 6.47\times10^{-6}$), (2) the

association between intrinsic phenotype and exposure factors is nominally significant in the DMP ($P_{DMP} < 0.05$), and (3) the association exposure factor and gut microbial SNV is significant in the DMP ($P_{DMP} < 6.47\times10^{-6}$) or meta-analysis ($P_{meta} < 6.47\times10^{-6}$). We then performed bidirectional mediation analysis on the candidate variable groups following the framework shown in **Fig. 7a**. For the candidate groups involving continuous or ordinal phenotype variables, we used a linear model in each step of mediation analysis. For the candidate groups involving binary phenotype variables, we used a logistic regression model when the response variable was binary. Finally, the P-values of indirect effects were corrected by FDR-estimation.

References

1. Wang, D. *et al.* Characterization of gut microbial structural variations as determinants of human bile acid metabolism. *Cell Host Microbe* **29**, 1802-1814.e5 (2021).

2. Gulyaeva, A. *et al.* Discovery, diversity, and functional associations of crAss-like phages in human gut metagenomes from four Dutch cohorts. *Cell Rep.* **38**, 110204 (2022).

3. Zhernakova, D. V. *et al.* Host genetic regulation of human gut microbial structural variation. *Nature* **625**, 813–821 (2024).

4. Zeevi, D. *et al.* Structural variation in the gut microbiome associates with host health. *Nature* **568**, 43–48 (2019).

5. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45–50 (2013).

6. Shi, Z. J., Dimitrov, B., Zhao, C., Nayfach, S. & Pollard, K. S. Fast and accurate metagenotyping of the human gut microbiome with GT-Pro. *Nat. Biotechnol.* **40**, 507–516 (2022).

7. Chen, L. *et al.* Short- and long-read metagenomics expand individualized structural variations in gut microbiomes. *Nat. Commun.* **13**, 3175 (2022).

8. Jiang, X. *et al.* Invertible promoters mediate bacterial phase variation, antibiotic resistance, and host adaptation in the gut. *Science* **363**, 181–187 (2019).

9. Chen, L. *et al.* The long-term genetic stability and individual specificity of the human gut microbiome. *Cell* **184**, 2302-2315.e12 (2021).

10. Zhernakova, D. V. *et al.* Host genetic regulation of human gut microbial structural variation. *Nature* **625**, 813–821 (2024).

11. Zou, D. *et al.* A SNP of bacterial blc disturbs gut lysophospholipid homeostasis and induces inflammation through epithelial barrier disruption. *EBioMedicine* **52**, 102652 (2020).

12. Sharafutdinov, I. *et al.* A single-nucleotide polymorphism in Helicobacter pylori promotes gastric cancer development. *Cell Host Microbe* **31**, 1345-1358.e6 (2023).

13. Chen, Y. *et al.* Gut metagenomes of type 2 diabetic patients have characteristic single-nucleotide polymorphism distribution in Bacteroides coprocola. *Microbiome* **5**, 15 (2017).

14. Uffelmann, E. *et al.* Genome-wide association studies. *Nat. Rev. Methods Prim.* **1**, 59 (2021).

15. Berthenet, E. *et al.* A GWAS on Helicobacter pylori strains points to genetic variants associated with gastric cancer risk. *BMC Biol.* **16**, 84 (2018).

16. Farhat, M. R. *et al.* GWAS for quantitative resistance phenotypes in Mycobacterium tuberculosis reveals resistance genes and regulatory regions. *Nat. Commun.* **10**, 2128 (2019).

17. Coll, F. *et al.* Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. *Nat. Genet.* **50**, 307–316 (2018).

18. Young, B. C. *et al.* Panton–Valentine leucocidin is the key determinant of Staphylococcus aureus pyomyositis in a bacterial GWAS. *eLife* **8**, e42486 (2019).

19. Power, R. A., Parkhill, J. & Oliveira, T. de. Microbial genome-wide association studies: lessons from human GWAS. *Nat. Rev. Genet.* **18**, 41–50 (2017).

20. San, J. E. *et al.* Current Affairs of Microbial Genome-Wide Association Studies: Approaches, Bottlenecks and Analytical Pitfalls. *Front. Microbiol.* **10**, 3119 (2020).

21. Zahavi, L. *et al.* Bacterial SNPs in the human gut microbiome associate with host BMI. *Nat. Med.* **29**, 2785–2792 (2023).

22. Gacesa, R. *et al.* Environmental factors shaping the gut microbiome in a Dutch population. *Nature* **604**, 732–739 (2022).

23. Tigchelaar, E. F. *et al.* Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open* **5**, e006772 (2015).

24. Schirmer, M. *et al.* Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity. *Cell* **167**, 1125-1136.e8 (2016).

25. Horst, R. ter *et al.* Host and Environmental Factors Influencing Individual Human Cytokine Responses. *Cell* **167**, 1111-1124.e13 (2016).

26. Imhann, F. *et al.* The 1000IBD project: multi-omics data of 1000 inflammatory bowel disease patients; data release 1. *BMC Gastroenterol.* **19**, 5 (2019).

27. Kurilshikov, A. *et al.* Gut Microbial Associations to Plasma Metabolites Linked to Cardiovascular Phenotypes and Risk. *Circ. Res.* **124**, 1808–1820 (2019).

28. Stražar, M. *et al.* Gut microbiome-mediated metabolism effects on immunity in rural and urban African populations. *Nat. Commun.* **12**, 4845 (2021).

29. Fernández-Pato, A. *et al.* Choice of DNA extraction method affects stool microbiome recovery and subsequent phenotypic association analyses. *Sci. Rep.* **14**, 3911 (2024).

30. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27**, 626–638 (2017).

31. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6**, 80–92 (2012).

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

33. Hughes, A. M. *et al.* Peptide transport in Bacillus subtilis – structure and specificity in the extracellular solute binding proteins OppA and DppE. *Microbiology* **168**, (2022).

34. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 1–3 (2024) doi:10.1038/s41586-024-07487-w.

35. Liang, J. *et al.* Molecular recognition of trehalose and trehalose analogues by Mycobacterium tuberculosis LpqY-SugABC. *Proc. Natl. Acad. Sci.* **120**, e2307625120 (2023).

36. Nguyen, P. T., Lai, J. Y., Lee, A. T., Kaiser, J. T. & Rees, D. C. Noncanonical role for the binding protein in substrate uptake by the MetNI methionine ATP Binding Cassette (ABC) transporter. *Proc. Natl. Acad. Sci.* **115**, E10596–E10604 (2018).

37. Hannum, G. *et al.* Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Mol. Cell* **49**, 359–367 (2013).

38. Weidner, C. I. *et al.* Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biol.* **15**, R24 (2014).

39. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, 3156 (2013).

40. Andreu-Sánchez, S. *et al.* Genetic, parental and lifestyle factors influence telomere length. *Commun. Biol.* **5**, 565 (2022).

41. Rodriguez-Melendez, R., Lewis, B., McMahon, R. J. & Zempleni, J. Diaminobiotin and Desthiobiotin Have Biotin-Like Activities in Jurkat Cells. *J. Nutr.* **133**, 1259–1264 (2003).

42. Kim, S.-M. *et al.* Secreted Akkermansia muciniphila threonyl-tRNA synthetase functions to monitor and modulate immune homeostasis. *Cell Host Microbe* **31**, 1021-1037.e10 (2023).

43. Mlodzik-Czyzewska, M. A., Malinowska, A. M. & Chmurzynska, A. Low folate intake and serum levels are associated with higher body mass index and abdominal fat accumulation: a case control study. *Nutr. J.* **19**, 53 (2020).

44. Tennoune, N. *et al.* Bacterial ClpB heat-shock protein, an antigen-mimetic of the anorexigenic peptide α-MSH, at the origin of eating disorders. *Transl. Psychiatry* **4**, e458–e458 (2014).

45. Zolotarev, V. A. *et al.* Protein Extract of a Probiotic Strain of Hafnia alvei and Bacterial ClpB Protein Improve Glucose Tolerance in Mice. *Int. J. Mol. Sci.* **24**, 10590 (2023).

46. Holeček, M. Serine Metabolism in Health and Disease and as a Conditionally Essential Amino Acid. *Nutrients* **14**, 1987 (2022).

47. Chen, B. *et al.* Gut bacteria alleviate smoking-related NASH by degrading gut nicotine. *Nature* **610**, 562–568 (2022).

48. Baharnoori, M. *et al.* Altered adipokine levels are associated with dimethyl fumarate treatment in multiple sclerosis patients. *Mult. Scler. Relat. Disord.* **56**, 103311 (2021).

49. Seo, H.-A. & Lee, I.-K. The Role of Nrf2: Adipocyte Differentiation, Obesity, and Insulin Resistance. *Oxidative Med. Cell. Longev.* **2013**, 184598 (2013).

50. Wang, Z. *et al.* Nrf2 in adipocytes. *Arch. Pharmacal Res.* **43**, 350–360 (2020).

51. Chartoumpekis, D. V. *et al.* Nrf2 activation diminishes during adipocyte differentiation of ST2 cells. *Int. J. Mol. Med.* **28**, 823–8 (2011).

52. Favero, C. *et al.* Phosphate, Microbiota and CKD. *Nutrients* **13**, 1273 (2021).

53. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).

54. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

55. Zhao, C., Shi, Z. J. & Pollard, K. S. Pitfalls of genotyping microbial communities with rapidly growing genome collections. *Cell Syst.* **14**, 160-176.e3 (2023).

56. Rossum, T. V. *et al.* metaSNV v2: detection of SNVs and subspecies in prokaryotic metagenomes. *Bioinformatics* **38**, 1162–1164 (2021).

57. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).

58. VanEvery, H., Franzosa, E. A., Nguyen, L. H. & Huttenhower, C. Microbiome epidemiology and association studies in human health. *Nat. Rev. Genet.* **24**, 109–124 (2023).

59. Faith, J. J. *et al.* The Long-Term Stability of the Human Gut Microbiota. *Science* **341**, 1237439 (2013).

60. Tamburini, S., Shen, N., Wu, H. C. & Clemente, J. C. The microbiome in early life: implications for health outcomes. *Nat. Med.* **22**, 713–722 (2016).

61. Arrieta, M.-C., Stiemsma, L. T., Amenyogbe, N., Brown, E. M. & Finlay, B. The Intestinal Microbiome in Early Life: Health and Disease. *Front. Immunol.* **5**, 427 (2014).

62. Donald, K. & Finlay, B. B. Early-life interactions between the microbiota and immune system: impact on immune system development and atopic disease. *Nat. Rev. Immunol.* **23**, 735–748 (2023).

63. Valles-Colomer, M. *et al.* The person-to-person transmission landscape of the gut and oral microbiomes. *Nature* **614**, 125–135 (2023).

64. Andreu-Sánchez, S., Wu, J. & Fu, J. Beyond personal space: Unveiling the transmission pattern of the human gut and oral microbiome. *iMeta* **2**, e98 (2023).

65. Vatanen, T. *et al.* Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. *Nat. Microbiol.* **4**, 470–479 (2019).

66. Quercia, S. *et al.* From lifetime to evolution: timescales of human gut microbiota adaptation. *Front. Microbiol.* **5**, 587 (2014).

67. Sarkar, A., Yoo, J. Y., Dutra, S. V. O., Morgan, K. H. & Groer, M. The Association between Early-Life Gut Microbiota and Long-Term Health and Diseases. *J. Clin. Med.* **10**, 459 (2021).

68. Indrio, F. *et al.* Epigenetic Matters: The Link between Early Nutrition, Microbiome, and Long-term Health Development. *Front. Pediatr.* **5**, 178 (2017).

69. Kwon, N. H., Fox, P. L. & Kim, S. Aminoacyl-tRNA synthetases as therapeutic targets. *Nat. Rev. Drug Discov.* **18**, 629–650 (2019).

70. Guo, M. & Schimmel, P. Essential nontranslational functions of tRNA synthetases. *Nat. Chem. Biol.* **9**, 145–153 (2013).

71. Guo, M., Yang, X.-L. & Schimmel, P. New functions of aminoacyl-tRNA synthetases beyond translation. *Nat. Rev. Mol. Cell Biol.* **11**, 668–674 (2010).

72. Han, H. *et al.* From gut microbiota to host appetite: gut microbiota-derived metabolites as key regulators. *Microbiome* **9**, 162 (2021).

73. Vinot, N., Baghtchedjian, E., Picolo, C. & Lambert, G. Weight Management - Challenges and Opportunities. (2022) doi:10.5772/intechopen.103723.

74. Breton, J. *et al.* Gut Commensal E. coli Proteins Activate Host Satiety Pathways following Nutrient-Induced Bacterial Growth. *Cell Metab.* **23**, 324–334 (2016).

75. Arnoriaga-Rodríguez, M. *et al.* Gut bacterial ClpB-like gene function is associated with decreased body weight and a characteristic microbiota profile. *Microbiome* **8**, 59 (2020).

76. Dominique, M. *et al.* Effects of Bacterial CLPB Protein Fragments on Food Intake and PYY Secretion. *Nutrients* **13**, 2223 (2021).

77. Ericson, M. D., Schnell, S. M., Freeman, K. T. & Haskell-Luevano, C. A fragment of the Escherichia coli ClpB heat-shock protein is a micromolar melanocortin 1 receptor agonist. *Bioorg. Med. Chem. Lett.* **25**, 5306–5308 (2015).

78. Déchelotte, P. *et al.* The Probiotic Strain H. alvei HA4597® Improves Weight Loss in Overweight Subjects under Moderate Hypocaloric Diet: A Proof-of-Concept, Multicenter Randomized, Double-Blind Placebo-Controlled Study. *Nutrients* **13**, 1902 (2021).

79. Fetissov, S. O. & Hökfelt, T. On the origin of eating disorders: altered signaling between gut microbiota, adaptive immunity and the brain melanocortin system regulating feeding behavior. *Curr. Opin. Pharmacol.* **48**, 82–91 (2019).

80. Krautkramer, K. A., Fan, J. & Bäckhed, F. Gut microbial metabolites as multi-kingdom intermediates. *Nat. Rev. Microbiol.* **19**, 77–94 (2021).

81. Rooks, M. G. & Garrett, W. S. Gut microbiota, metabolites and host immunity. *Nat. Rev. Immunol.* **16**, 341–352 (2016).

82. Girdhar, K. *et al.* A gut microbial peptide and molecular mimicry in the pathogenesis of type 1 diabetes. *Proc. Natl. Acad. Sci.* **119**, e2120028119 (2022).

83. Cai, Y. *et al.* The landscape of aging. *Sci. China Life Sci.* **65**, 2354–2454 (2022).

84. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. Hallmarks of aging: An expanding universe. *Cell* **186**, 243–278 (2023).

85. Ghosh, T. S., Shanahan, F. & O'Toole, P. W. The gut microbiome as a modulator of healthy ageing. *Nat. Rev. Gastroenterol. Hepatol.* **19**, 565–584 (2022).

86. Wilmanski, T. *et al.* Gut microbiome pattern reflects healthy ageing and predicts survival in humans. *Nat. Metab.* **3**, 274–286 (2021).

87. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).

88. Listgarten, J. *et al.* Improved linear mixed models for genome-wide association studies. *Nat. Methods* **9**, 525–526 (2012).

89. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46**, 100–106 (2014).

90. Liang, J. & Tan, Y. Highly efficient CRISPR-mediated base editing for the gut Bacteroides spp. with pnCasBS-CBE. *Biotechnol. J.* **18**, e2200504 (2023).

91. Liu, L., Huang, Y. & Wang, H. H. Fast and efficient template-mediated synthesis of genetic variants. *Nat. Methods* **20**, 841–848 (2023).

92. Huang, Y. *et al.* High-throughput microbial culturomics using automation and machine learning. *Nat. Biotechnol.* **41**, 1424–1433 (2023).

93. Scholtens, S. *et al.* Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int. J. Epidemiology* **44**, 1172–1180 (2015).

94. Horst, R. ter *et al.* Sex-Specific Regulation of Inflammation and Metabolic Syndrome in Obesity. *Arter., Thromb., Vasc. Biol.* **40**, 1787–1800 (2019).

95. Vila, A. V. *et al.* Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. *Sci. Transl. Med.* **10**, eaap8914 (2018).

96. Eggink, H. M. *et al.* Complex interaction between circadian rhythm and diet on bile acid homeostasis in male rats. *Chronobiol. Int.* **34**, 1339–1353 (2017).

97. Hoogerland, J. A. *et al.* Glucose-6-Phosphate Regulates Hepatic Bile Acid Synthesis in Mice. *Hepatology* **70**, 2171–2184 (2019).

98. García-Villalba, R. *et al.* Alternative method for gas chromatography-mass spectrometry analysis of short-chain fatty acids in faecal samples. *J. Sep. Sci.* **35**, 1906–1913 (2012).

99. Sanna, S. *et al.* Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nat. Genet.* **51**, 600–605 (2019).

100. Li, X. S. *et al.* Untargeted metabolomics identifies trimethyllysine, a TMAO-producing nutrient precursor, as a predictor of incident cardiovascular disease risk. *JCI Insight* **3**, e99096 (2018).

101. Andreu-Sánchez, S. *et al.* Unraveling interindividual variation of trimethylamine N-oxide and its precursors at the population level. *iMeta* (2024) doi:10.1002/imt2.183.

102. Laan, T. van der *et al.* Fast LC-ESI-MS/MS analysis and influence of sampling conditions for gut metabolites in plasma and serum. *Sci. Rep.* **9**, 12370 (2019).

103. Fuhrer, T., Zampieri, M., Sévin, D. C., Sauer, U. & Zamboni, N. Genomewide landscape of gene–metabolome associations in Escherichia coli. *Mol. Syst. Biol.* **13**, 907 (2017).

104. Fuhrer, T., Heer, D., Begemann, B. & Zamboni, N. High-Throughput, Accurate Mass Metabolome Profiling of Cellular Extracts by Flow Injection–Time-of-Flight Mass Spectrometry. *Anal. Chem.* **83**, 7074–7080 (2011).

105. Chen, L. *et al.* Influence of the microbiome, diet and genetics on inter-individual variation in the human plasma metabolome. *Nat. Med.* **28**, 2333–2343 (2022).

106. Baerlocher, G. M., Vulto, I., Jong, G. de & Lansdorp, P. M. Flow cytometry and FISH to measure the average length of telomeres (flow FISH). *Nat. Protoc.* **1**, 2365–2376 (2006).

107. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

108. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

109. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).

110. Dong, S.-S. *et al.* LDBlockShow: a fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. *Brief. Bioinform.* **22**, bbaa227 (2020).

111. Costea, P. I. *et al.* Subspecies in the global human gut microbiome. *Mol. Syst. Biol.* **13**, 960 (2017).

112. Peng, B., Yu, R. K., DeHoff, K. L. & Amos, C. I. Normalizing a large number of quantitative traits using empirical normal quantile transformation. *BMC Proc.* **1**, S156 (2007).

113. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).

114. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

115. Pravda, L. *et al.* MOLEonline: a web-based tool for analyzing channels, tunnels and pores (2018 update). *Nucleic Acids Res.* **46**, W368–W373 (2018).

116. Yariv, B. *et al.* Using evolutionary data to make sense of macromolecules with a “face-lifted” ConSurf. *Protein Sci.* **32**, e4582 (2023).

Figures

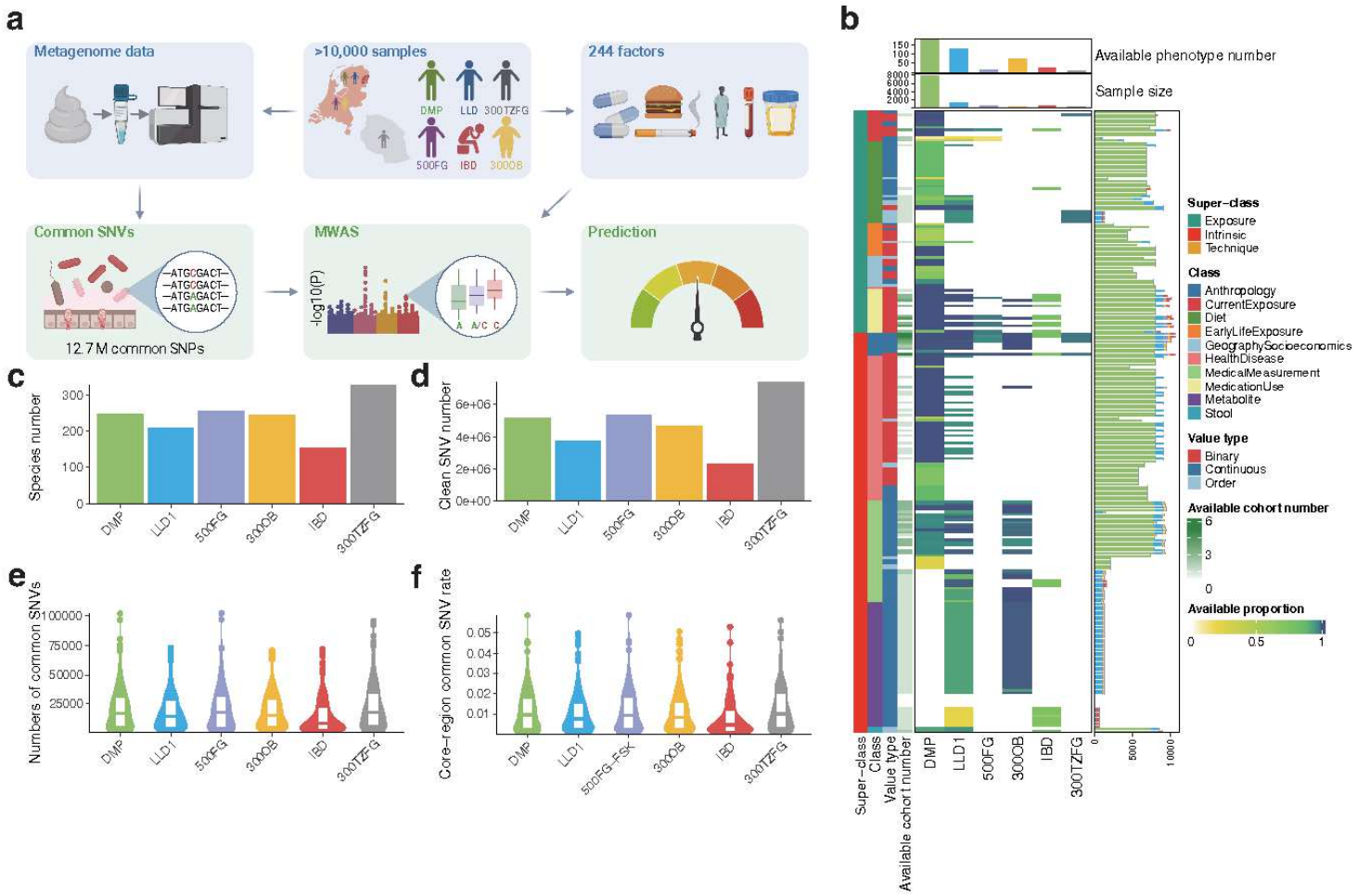


Figure 1

Overview of study design and microbial SNV data after filtering. a, Workflow of this gut microbial SNV-based genome-wide association study (SNV MWAS) on human phenotypes and environmental exposures. In this study we integrated whole-genome shotgun sequencing (WGS) data of 11,831 fecal samples from 10,781 individuals from six cohorts and harmonized 244 phenotype and exposure factors across these cohorts. We then conducted meta-genotyping for the WGS data from stool samples and generated the gut microbial genotype profile, keeping more than 12.7 million common biallelic microbial SNVs for the SNV-MWAS. The gut microbial SNVs associated with host phenotypes were used to construct a prediction model for host phenotype.

b, Overview of the 244 phenotype and exposure factors in this study. The proportions of samples with a non-missing value for the corresponding factor in each cohort are shown in the heatmap (color indicates available proportion). Each row represents a factor. Each column represents a cohort. c, Numbers of gut microbial species with at least 1,000 common SNVs in each cohort. d, Total number of common gut microbial SNVs kept for each cohort. e, Distributions of common SNV numbers of each species in each cohort. f, Proportion of common SNVs in the core region of each species in each

cohort. Inner boxplots of (e) and (f) represent summary statistics. The center line represents the median, the box hinges represent the lower and upper quartiles of the distribution, whiskers extend no further than 1.5× interquartile range from the hinges, and data beyond the end of the whiskers are outliers plotted as individual points.

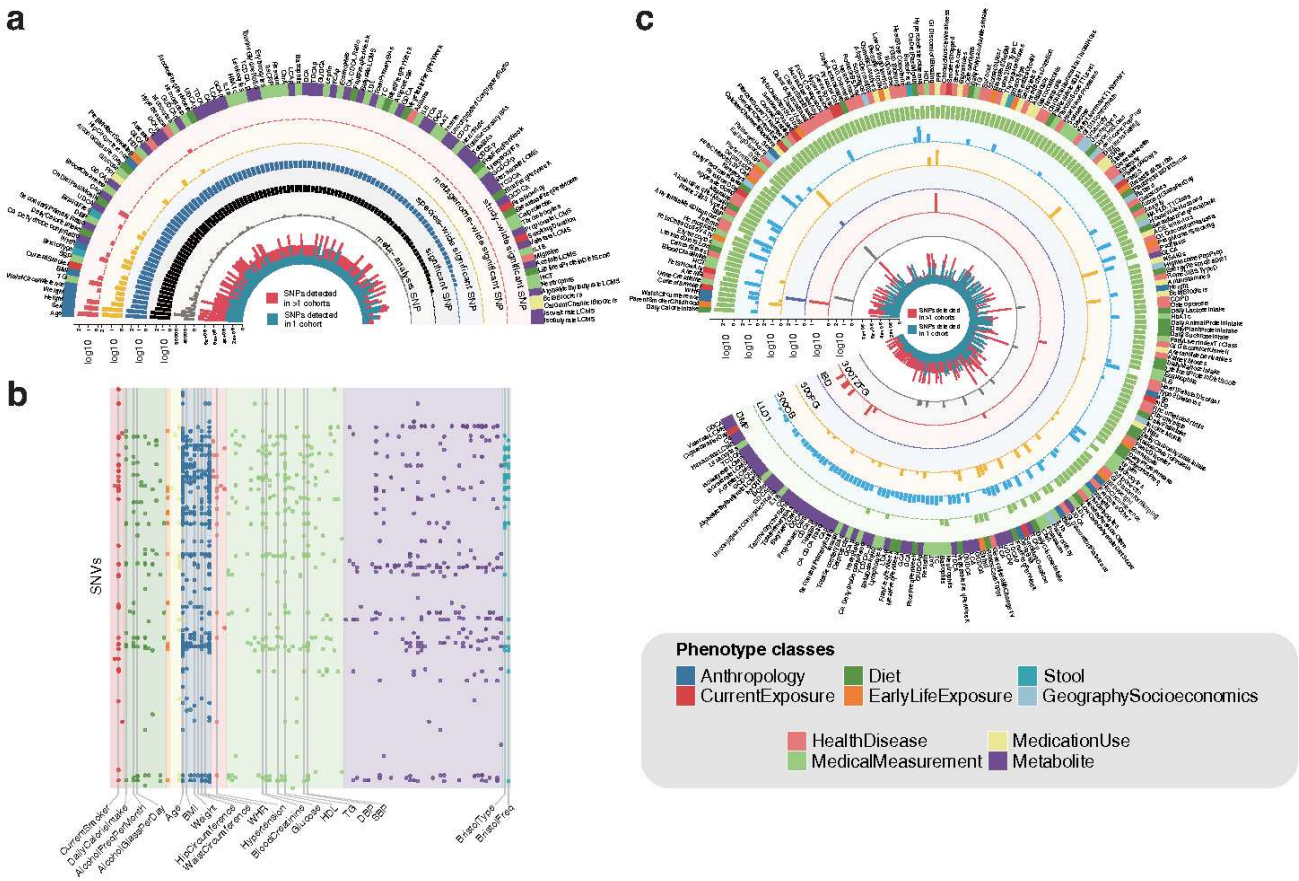


Figure 2

Summary of SNV-MWAS signals. a, Number of meta-analysis associations with gut microbial SNVs for each phenotype at different significance levels, including species-wide (Pmeta < 6.47×10-6), metagenome-wide (Pmeta < 7.34×10-9) and study-wide (Pmeta < 3.01×10-11) significance. b, Heatmap showing the association signals between all gut microbial SNVs and host phenotypes. Each dot is one association signal. Dot and column background colors indicate different phenotype classes. c, Number of single-cohort associations with gut microbial SNVs for each phenotype at species-wide significance (Pmeta < 6.47×10-6).

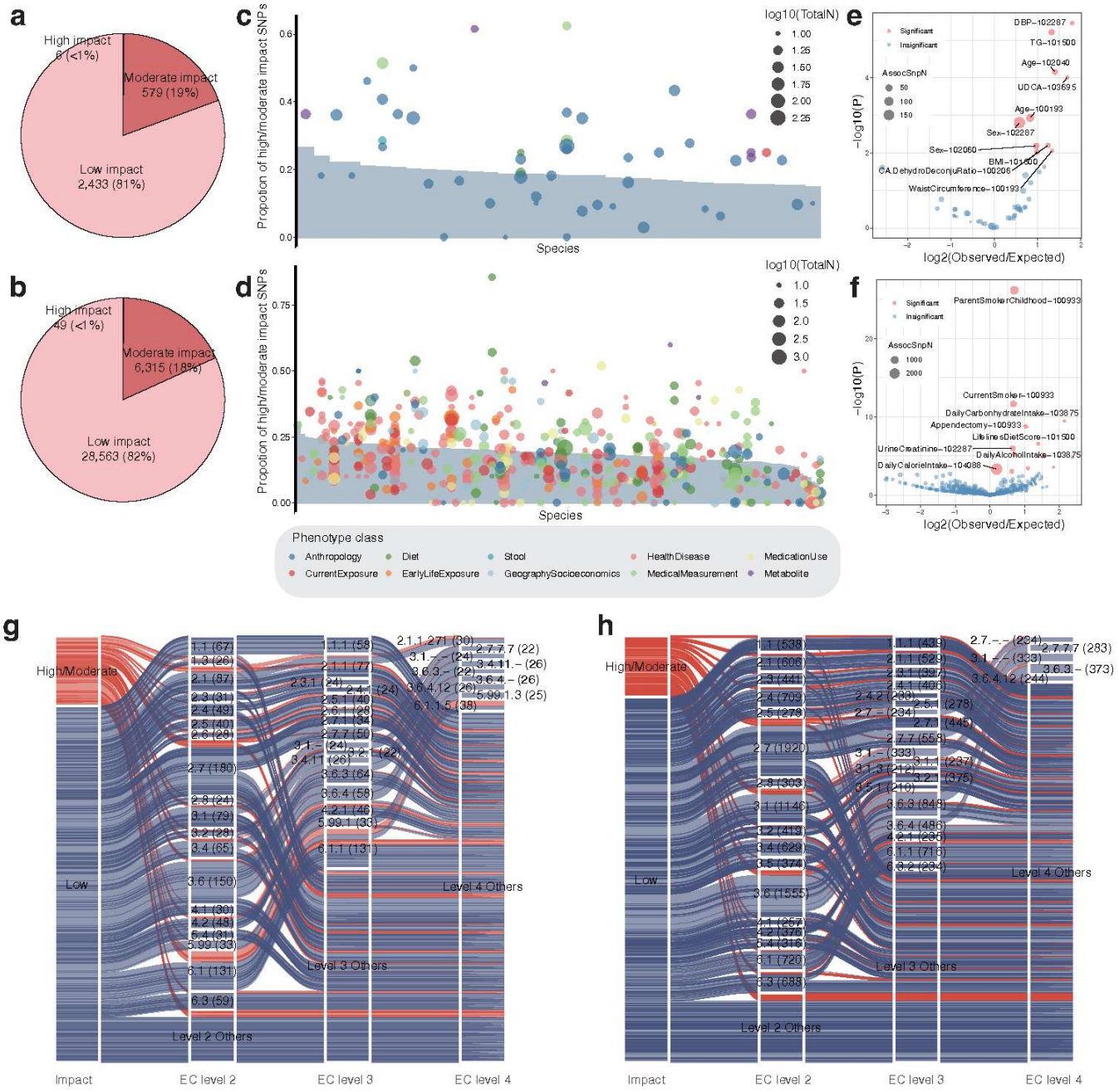


Figure 3

Functional properties of associated gut microbial SNVs. a and b, Proportion of associated gut microbial SNVs with different impact levels on the gene product in meta-analysis associations (a) and single-cohort associations (b). c and d, Proportion of high and moderate impact SNVs in phenotype–species pairs with more than 10 meta-analysis associations (c) and single-cohort associations (d). Each column represents a species. Light blue bars in the background show the proportion of high or moderate impact common SNVs in the genomes of corresponding species to indicate the background rate of high or moderate impact SNVs. Dot color indicates the different phenotype classes. Each dot represents a phenotype–species association pair for the corresponding species. Y-axis indicates the

proportion of high or moderate impact common SNVs in the associated SNVs of the corresponding phenotype–species pair. Dot size indicates the total number of associated SNVs of the corresponding phenotype–species pair. e and f, Enrichment of high or moderate SNVs in meta-analysis associations (e) and single-cohort associations (f) for phenotype–species pairs with more than 10 associations. Blue indicates non-significant enrichment (Chi-squared test, FDR > 0.05). Red indicates significant enrichment (Chi-squared test, FDR < 0.05). g, Number of meta-analysis associations involved in different enzyme classes at Enzyme Commission (EC) number level 2 (subclass), level 3 (sub-subclass) and level 4 (substrate classes). Enzyme classes with fewer than 20 associations are assigned to “Others”. h, Number of single-cohort associations involved in different enzyme classes at EC number level 2 (subclass), level 3 (sub-subclass) and level 4 (substrate classes). Enzyme classes with fewer than 200 associations are assigned to “Others”.

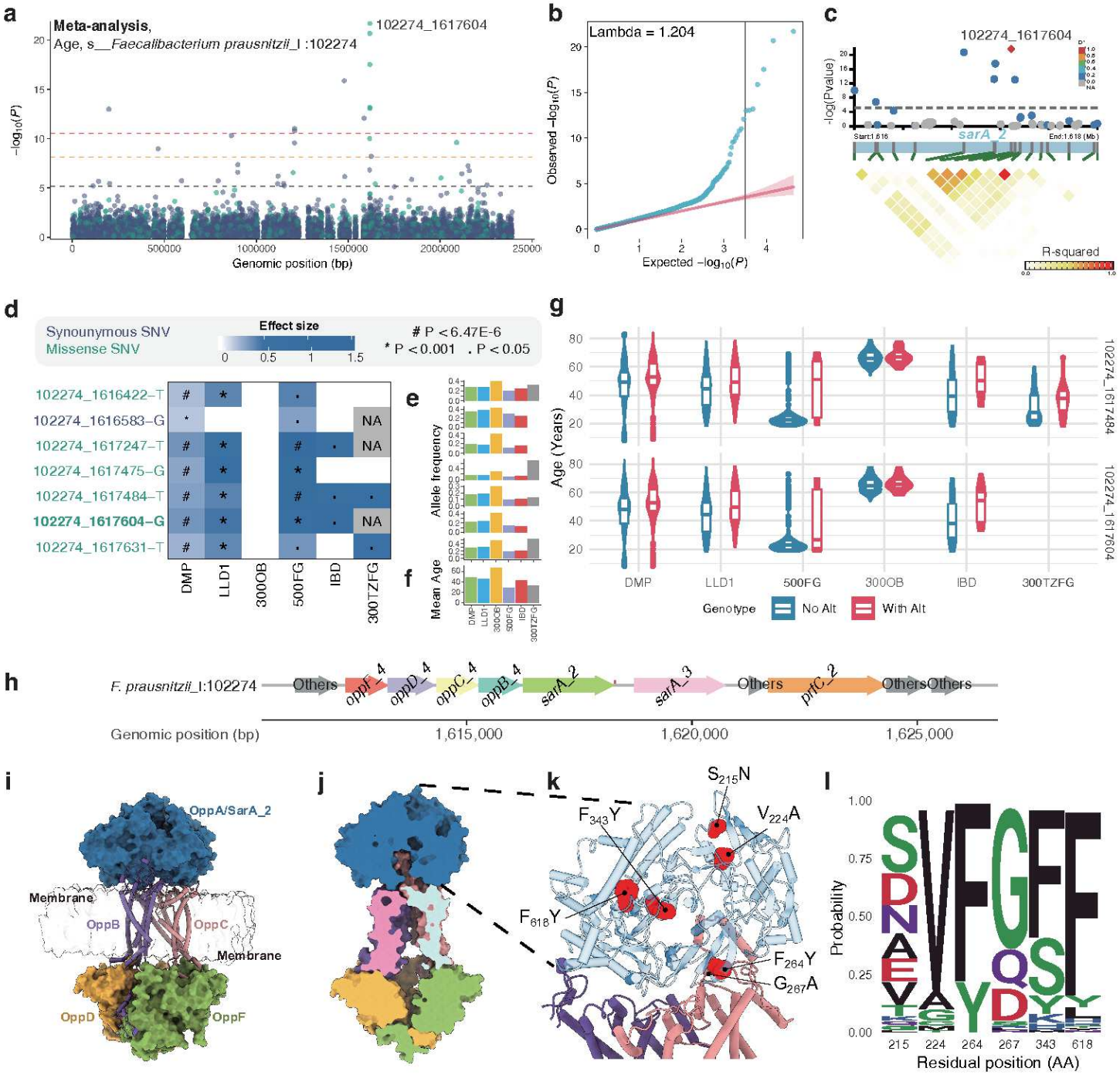


Figure 4

Missense SNVs in the sarA/oppA gene of Faecalibacterium prausnitzii associate with host age. a, Manhattan plot showing the association between SNVs of F. prausnitzii (Species ID: 102274) and host age. Dot colors indicate the impact levels of gut microbial SNVs on gene product (blue = low, green = moderate, and red = high). b, Q-Q plot of the P-values of associations between SNVs of F. prausnitzii (Species ID: 102274) and host age. Although the genomic inflation factor was 1.204, which is slightly higher than 1, the shape of the curve still suggest that the signals were not likely generated by randomness. c, Expanded view of the association signal showing the low LD between the SNVs associated with host age in the sarA_2 gene region (1,616,256–1,618,274 bp) of F. prausnitzii (Species

ID: 102274). This suggests that the SNVs are independently associated with host age. d, Heatmap of the significance of the associations between the seven SNVs and host age in six independent cohorts. # : P < 6.47×10-6; * : P < 0.001; . : P < 0.05. e and f, Small bar plots show the frequency of age-associated alleles (e) and mean age (f) in six independent cohorts. g, Violin plots show the difference in host age between samples with and without the age-associated alleles 947 of two SNVs (102274_1617484 and 102274_1617604) in six independent cohorts. Inner boxplots of (g) represent summary statistics. The center line represents the median, the box hinges represent the lower and upper quartiles of the distribution, whiskers extend no further than 1.5× interquartile range from the hinges, and data beyond the end of the whiskers are outliers plotted as individual points. h, The oligo-peptide transportation gene cluster in the genome of F. prausnitzii (Species ID: 102274). i and j, The predicted protein complex structure of the oligopeptide transporter formed by sarA/oppA, oppB, oppC, oppD, and oppF. (i) shows the oligopeptide transporter complex embedded into cell membrane. (j) shows the tunnel-like cavity space inside the oligopeptide transporter complex. k, Zoomed-in view of the sarA/oppA subunit showing the amino acid residuals (red) affected by the missense SNVs associated with host age. l, Probability of amino acids in the six age-associated oppA/sarA_2 residuals from the protein conservation analysis. Color indicates side-chain group chemistry type of amino acids, including acidic (red), basic (blue), hydrophobic (black), neutral (purple), and polar (green).

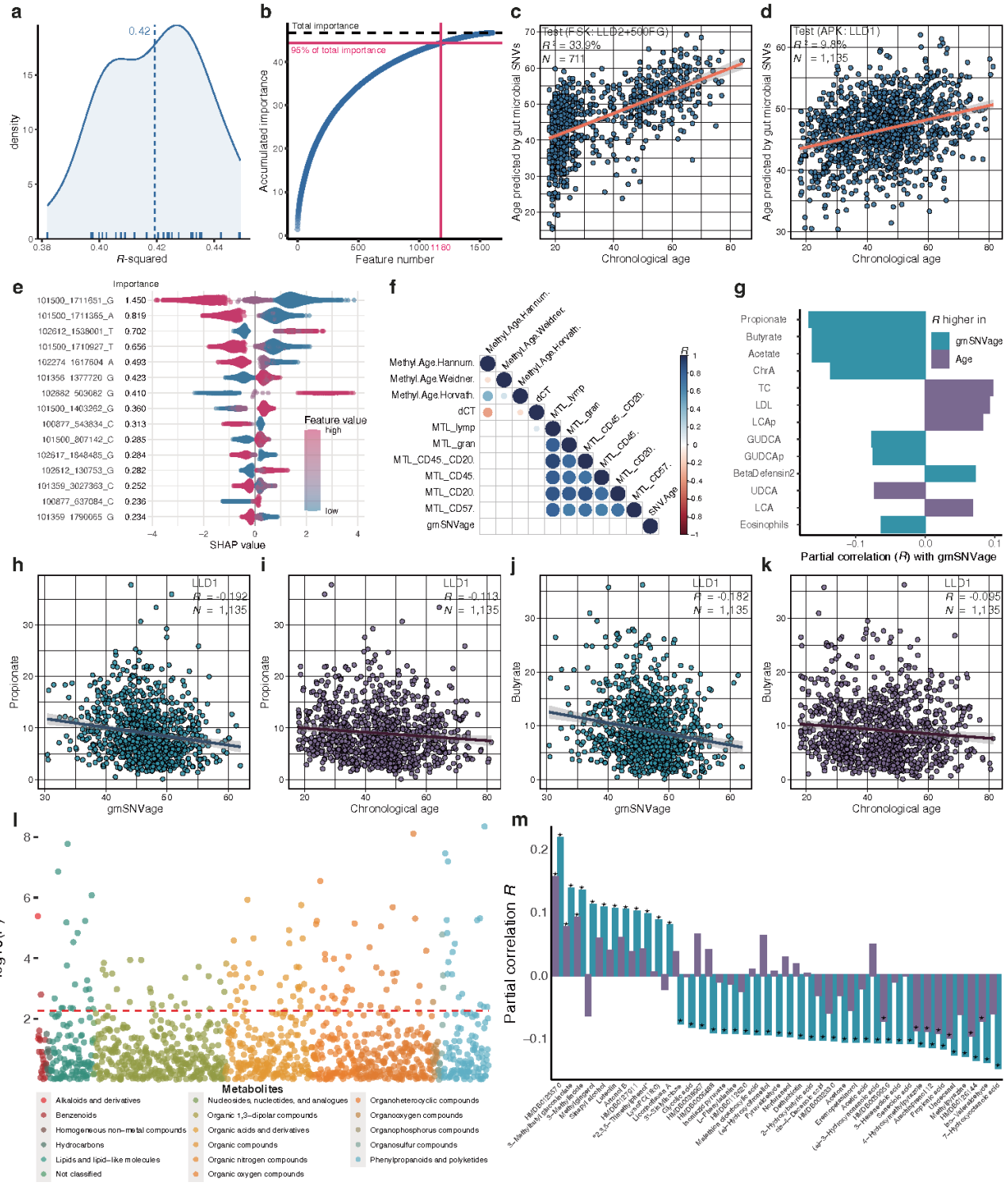


Figure 5

Host age predicted by gut microbial genotypes. a, The performance (R2) distribution of gut microbial SNV-based age prediction models in 5x5 rounds of fivefold cross-validation. The vertical dashed line denotes the mean performance (R2) of all rounds of cross-validation. b, Selection of features by median importance across 25 rounds of nested cross validation. The features are sorted by median importance, and we cumulatively sum the median importance values untill they reach 95% of the total sum of

importance values (pink horizontal line). We then take the counted SNV features as the predictive features (from left to right up to the pink vertical line) for the final model. c and d, Performance of final model for age prediction in the FSK test dataset (LLD2- FSK + 500FG-FSK, N = 711) (c) and the APK test dataset (LLD1, N = 1,135) (d). e, SHapley Additive exPlanations (SHAP) for the final model showing the top 15 important features. The reference allele is added to the ID of gut microbial SNV features. Feature values are indicated as blue (low value) and red (high values). f, Pairwise partial correlations between gut microbial SNV-predicted age and 10 biological age markers, correcting for chronological age. Blank or white cells indicate insignificant correlations (partial Spearman's correlation test, $P > 0.05$). Variables in each row (from top to bottom): three methylation-based age indicies (epigenetic age models developed by Hannum et al. 37, Weidner et al. 38 and Horvath et al. 39); signal joint T cell receptor excision circles (sjTRECs) expression (CT values of a qPCR), which indicates the maturation of T cells; and the 986 telomere lengths of six blood cell types: lymphocytes, granulocytes, B-cells (CD45RA+CD20+), naïve T cells (CD45RA+CD20−), memory T cells (CD45RA−) and NK-cells/fully differentiated T cells (CD45RA+CD57+); and gut microbial SNV-predicted age. g, Associations between predicted age and host phenotypes adjusted for chronological age. h and j, Associations between propionate level and predicted age and chronological age. Colors indicate if the correlations with phenotypes are higher for gut microbial SNV-predicted age (blue) or chronological age (purple). i and k, Associations between acetate level and predicted and chronological age. l, Associations between gut microbial SNV-predicted age and 1,183 blood metabolites, adjusted for chronological age. Dot color indicates super-classes of metabolites. Red horizontal dashed line denotes the significance cutoff (FDR < 0.05). m, Metabolites show stronger correlation with gut microbial SNV-predicted age (blue) than with chronological age (purple) (Pearson's correlation test). * denotes significant correlations (Pearson's correlation test, FDR < 0.05).

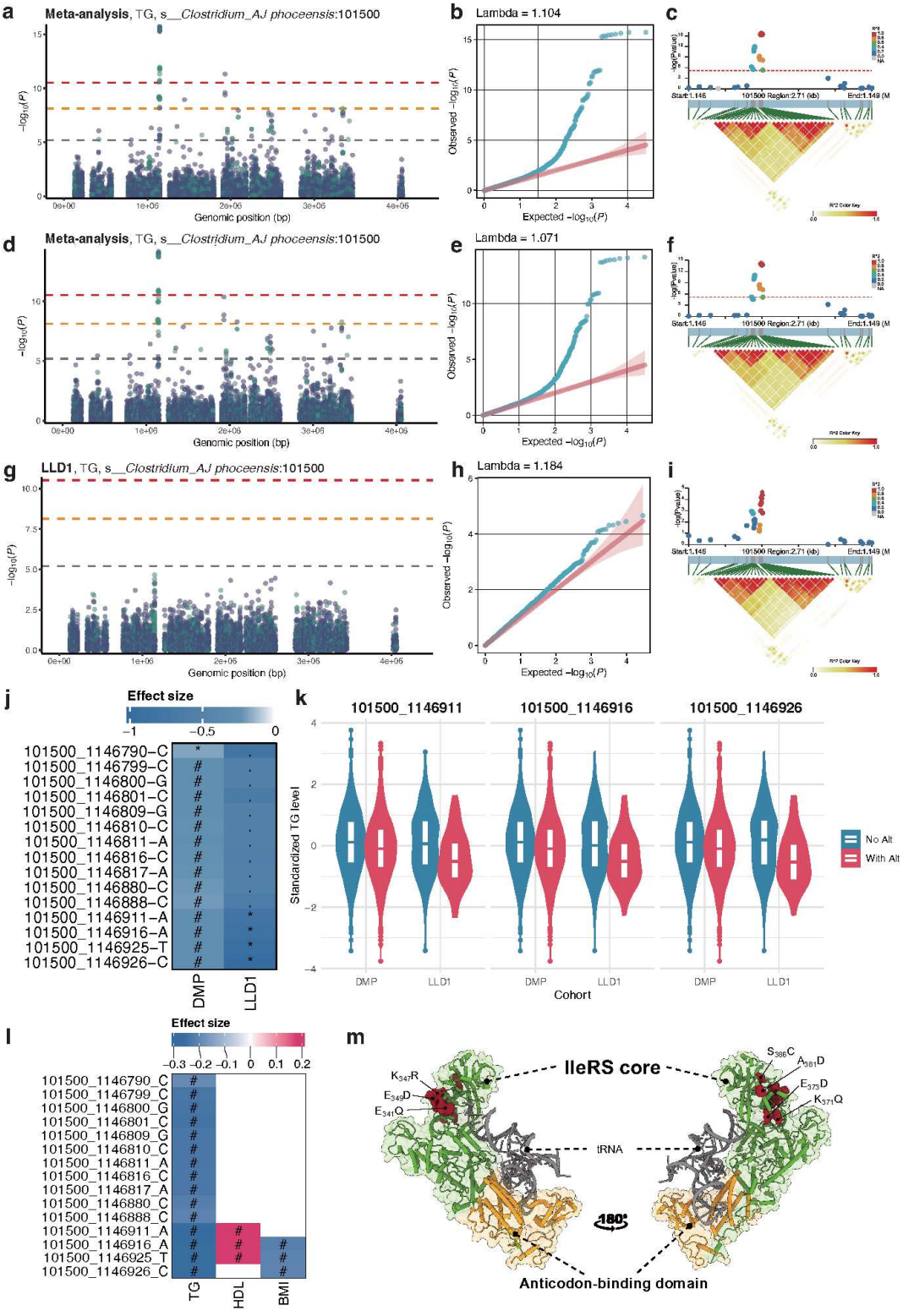


Figure 6

Missense SNVs in the ileS gene of Clostridium phoceensis associate with blood TG level. a, d, and g, Manhattan plots showing the association between SNVs of C. phoceensis (Species ID: 101500) and blood TG level in the meta-analysis (a), DMP cohort (d), and LLD1 cohort (g). Dot color indicates the impact levels of gut microbial SNVs on gene product (blue = low, green = moderate, and red (very rare) = high). b, e, and h, Q-Q plot of the P-values of associations between SNVs of C. phoceensis (Species ID:

101500) and blood TG level in the meta-analysis (b), DMP cohort (e), and LLD1 cohort (h). c, f, and i, Expanded view of the association signal shows the low LD between the SNVs associated with blood TG level in the ileS gene region (1,145,770–1,148,577 bp) of C. phoceensis (Species ID: 101500) in the meta-analysis (c), DMP cohort (f), and LLD1 cohort (i). j, Heatmap of the significance of the associations between the 15 missense SNVs in ileS and blood TG level in the DMP and LLD1 cohorts. The alternative alleles are added to the ID of gut microbial SNV. Cell colors show the effect size of the alternative allele. # : $P < 6.47\times10\text{-}6$; * : $P < 0.001$; . : $P < 0.05$. k, Violin plots show the difference in host age between samples with and without the TG-associated alleles of three SNVs (101500_1146911, 01500_1146916, and 01500_1146926) in DMP and LLD1. Inner boxplots of k represent summary statistics. The center line represents the median, the box hinges represent the lower and upper quartiles of the distribution, whiskers extend no further than 1.5× interquartile range

from the hinges, and data beyond the end of the whiskers are outliers plotted as individual points. l, Heatmap showing the associations between the 15 TG-associated missense SNVs in ileS and HDL-C and BMI in meta-analysis. The alternative alleles are added to the ID of gut microbial SNV. Cell color indicates the effect size of the alternative allele. # : $P < 6.47\times10\text{-}6$; * : $P < 0.001$; . : $P < 0.05$. m and n, The predicted protein structure of the ileS in side view (m) and top view (n). The red amino acids show the positions affected by the missense SNVs associated with blood TG level.

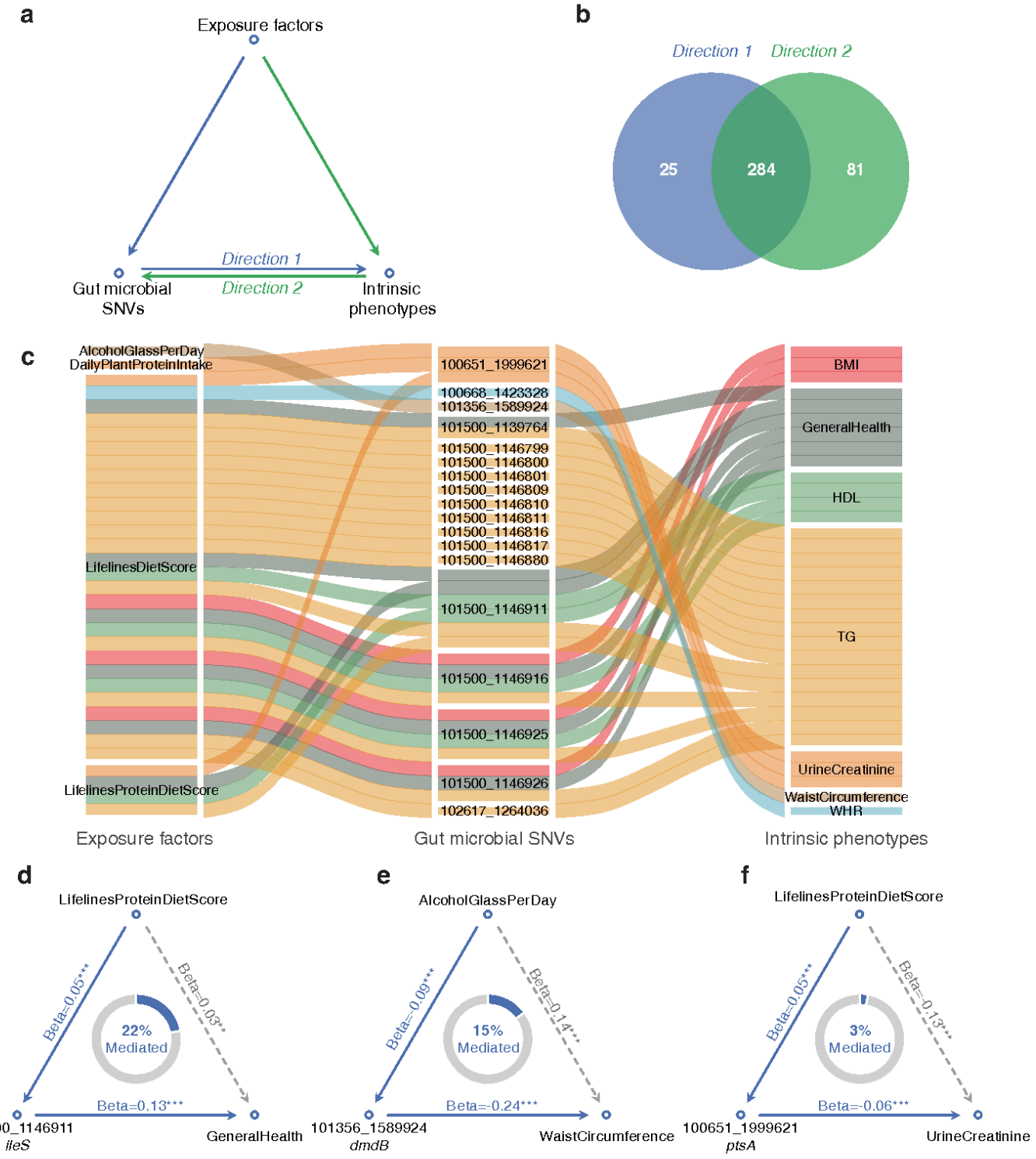


Figure 7

Mediation relationship inference between exposure factors, gut microbial SNVs, and intrinsic host phenotypes using bidirectional mediation analysis. a, Framework of bidirectional mediation analysis between exposure factors, gut microbial SNVs, and intrinsic host phenotypes. b, Number of inferred mediation relationships for Direction 1 (from gut microbial SNVs to intrinsic host phenotypes), Direction 2 (from intrinsic host phenotypes to gut microbial SNVs), and both. c, Sankey diagram showing the

inferred mediation relationship network of Direction 1. d–f, Examples of mediation relationships between exposure factors, gut microbial SNVs, and intrinsic host phenotypes inferred by bidirectional mediation analysis. The beta coefficient and significance are labeled at each edge. The proportions of indirect effect (mediation effect) are labeled at the center of the ring charts.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf
- SupplementaryTables.xlsx