Supplementary Information

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

Table of contents

Supplementary Notes

Supplementary Note 1. Detailed cohort and dataset description.

This study involved five European cohorts from the Netherlands, the Dutch Microbiome Project (DMP, *N* = 7,955), LifeLines-DEEP (LLD, *N* = 1,135, of which 338 were followed-up after an average of 4 years), the 500 Functional Genomics Project (500FG, *N* = 529), the 1000 Inflammatory Bowel Disease Project (IBD, *N* = 544), and 300-Obesity (300OB, *N* = 298), and an African cohort from Tanzania, the 300 Tanzania Functional Genomics Project (300TZFG, *N* = 320) (**Table S1**).

For LLD, there are sample sets taken at two time points: baseline time point samples (LLD1, *N* = 1,135 individuals) and samples from a subset of these participants taken at a follow-up time point around four years later (LLD2, *N* = 338). For the LLD2 faecal samples, two equal aliquots of the samples were processed by two different DNA isolation kits: the QIAamp Fast DNA Stool Mini Kit (FSK) or the QIAGEN AllPrep DNA/RNA kits (APK). We therefore have two datasets for LLD2: LLD2-APK and LLD2-FSK. The faecal samples of the 500FG cohort were also processed by FSK and APK, thus we also have two datasets for 500FG: 500FG-APK and 500FG-FSK. To make sure the SNV-MWAS was conducted in independent samples, we excluded the two LLD2 datasets and the 500FG-APK dataset from the SNV-MWAS (**Table S1**). If not specifically indicated, 500FG represents the 500FG-FSK dataset in the main text.

Supplementary Note 2. Population genetic structure in the gut microbiome.

The genomic variants can reflect the genetic diversity of species. To capture the intra-species genetic diversity, we conducted principal coordinate analysis (PCoA) on the genotype profiles of each microbial species separately. For the 251 species with >1000 samples included in the PCoA, the proportion of population genetic variance explained by the top 10 PCs was 4.63% on average, but with a wide range between 0.3% and 21.0% (**Fig. S2a** and **Table S5**). For instance, the top 10 PCs could explain >20% of genetic variations for a species from the order *Flavobacteriales* (species ID: 100451) and a species from family CAG-382 (species ID: 104081), suggesting that these species have a more pronounced population structure (**Fig. S2b,c**). The PCoA results revealed clear and distinct clusters for some species. For example, *Akkermansia muciniphila* (species ID: 102454), *Bacteroides fragilis* (species ID: 101337) and *Bacteroides caccae* (species ID: 102549) demonstrated well-defined clustering (**Fig. S2d–f**), in line with previous observations of distinct clades for these species [1,2]. Interestingly, different clusters of some species could be linked to the geographic origins of the samples. Samples from the Netherlands and Tanzania were distinctly separate for *Blautia obeum* (species ID: 100212), *Prevotella copri* (species ID: 102293), and *Agathobacter faecis* (species ID: 103694) (**Fig. S2j–l**). In contrast, some species, such as *Anaerostipes hadrus* (species ID: 100028), *Streptococcus salivarius* (species ID: 100113), and *Bifidobacterium adolescentis* (species ID: 102395), showed a more homogenous genetic makeup across samples (**Fig. S2g–i**).

The linkage disequilibrium (LD) structure of SNVs is related to the observed population structures of species. We further calculated the genome-wide LD (R^2) level for each of the 433 species. Genome-wide LD level varied across species, ranging from 0.002 to 0.922, with an average R^2 of 0.081 (**Table S6**). We then counted the proportion of SNV pairs showing moderate to elevated LD ($R^2 > 0.49$) in all randomly sampled SNV pairs for each species in nine datasets separately, and the proportions of linked SNV pairs of most species were similar in different datasets (**Fig. S3a**,**b** and **Table S7**). For some species, we observed high LD, i.e., that the majority of the SNVs in the genome were strongly linked, for example g__Butyricimonas (species ID:101655) and g__UBA7173 (species ID:100667) (**Fig. S3d**,**e**). Strong genome-wide long-range LD increases the false discovery rate and makes localization of genuine associations difficult, making it necessary to have systemic quality control and to inspect the LD structures of interesting signal regions.

Supplementary Note 3. Assessment of the impact of DNA isolation kits on metagenotyping and population structure.

For the LLD and 500FG cohorts, there are additional re-sequenced datasets that were included in metagenotyping but not in downstream association analysis. These two cohorts were profiled twice, once using microbial DNAs isolated using the QIAamp Fast DNA Stool Mini Kit (FSK) and a second time using DNA isolated with the QIAGEN AllPrep DNA/RNA kit (APK). This allowed us to assess the impact of the different DNA isolation methods on metagenotyping.

The FSK datasets of both LLD2 and 500FG were detected to have higher species numbers, higher total common SNV numbers, and higher average common SNV numbers than the APK datasets (**Fig.1c–e**). Further, previous studies have suggested that technical factors like the DNA isolation Kit used impact the enrichment of gut microbial species and lead to dramatic differences in gut microbiome composition. However, the impact of technical factors on within-species diversity and the detection of genetic variation remains unclear. The samples processed by two different DNA isolation kits in our cohorts allowed us to assess the impact of DNA isolation on within-species diversity and SNV detection. At population level, 203 species with at least 1,000 common SNVs between FSK and APK were detected (**Fig. S5a**), accounting for 74.63% and 82.19% of the species number in FSK and APK datasets, respectively, and reflecting the enrichment of different species by different DNA isolation kits. For the species available in both the FSK and APK datasets, the proportion of common SNVs detected in both varies across different species, ranging from 4.67% to 80.33%, with an average value of 38.32% (**Fig. S5b** and **c**).

We then assessed the consistency of metagenotypes between the FSK and APK datasets for each species in two ways. First, for the SNV sites captured by both FSK and APK kits, we calculated the correlation of the alternative allele proportion between all paired samples for each species, finding an average correlation of 89.57% (**Table S8**). Second, we converted the alternative allele proportion to 0, 1, or 2 to indicate no alternative allele, both alternative and reference allele, or only alternative allele detected, respectively. We then checked the rate of discordant genotypes and found an average rate for all species of 7.54% (**Fig. S5d**,**e** and **Table S8**), indicating that the

DNA isolation approach had a mild impact on metagenotyping. To further quantify the impact of the DNA isolation approach on the construction of gut microbial populational genetic structure, we conducted PCoA on LLD2 and 500FG samples and defined accuracy as the rate of paired samples from FSK and APK showing the least genetic dissimilarity among all sample pairs. The average accuracy was 82.19%, and more than 66.50% of species showed high accuracy (>80%) (**Fig. S5f,g** and **Table S8**), suggesting that the population structures of the majority of species are not strongly impacted by the DNA isolation approach used.

Supplementary Note 4. Correlation between gut microbial SNV-predicted age and other biological age markers.

To examine if gut microbial SNV-predicted age reflects unique aspects of biological ageing, independent of other biological age hallmarks, we obtained 10 biological age markers for the LLD1 cohort. These included three methylation-based age indices (epigenetic age models developed by Hannum et al. [3], Weidner et al. [4] and Horvath et al. [5]); signal joint T-cell receptor excision circles (sjTRECs) expression (CT values of a qPCR), which indicate the maturation of T cells; and the telomere length of six blood cell types: granulocytes, lymphocytes, B-cells ($CD45RA^+CD20^+$), naïve T cells ($CD45RA^+CD20^-$), memory T cells ($CD45RA^-$), and NK-cells/fully differentiated T cells ($CD45RA^+CD57^+$). The correlation (R^2) between actual age and these biological age indices ranged from 0.11 to 0.89. We then conducted pairwise partial correlation between our gut microbial SNV-predicted age and the 10 biological age markers, correcting for chronological age. Here we found that the gut microbial SNV-predicted age is independent of all other biological age indices (Partial Spearman's correlation test, $P > 0.05$; **Fig. 5f**), indicating that gut microbial SNV-predicted age may reflect a unique aspect of functional senescence not captured by other biological age markers.

Supplementary Figures

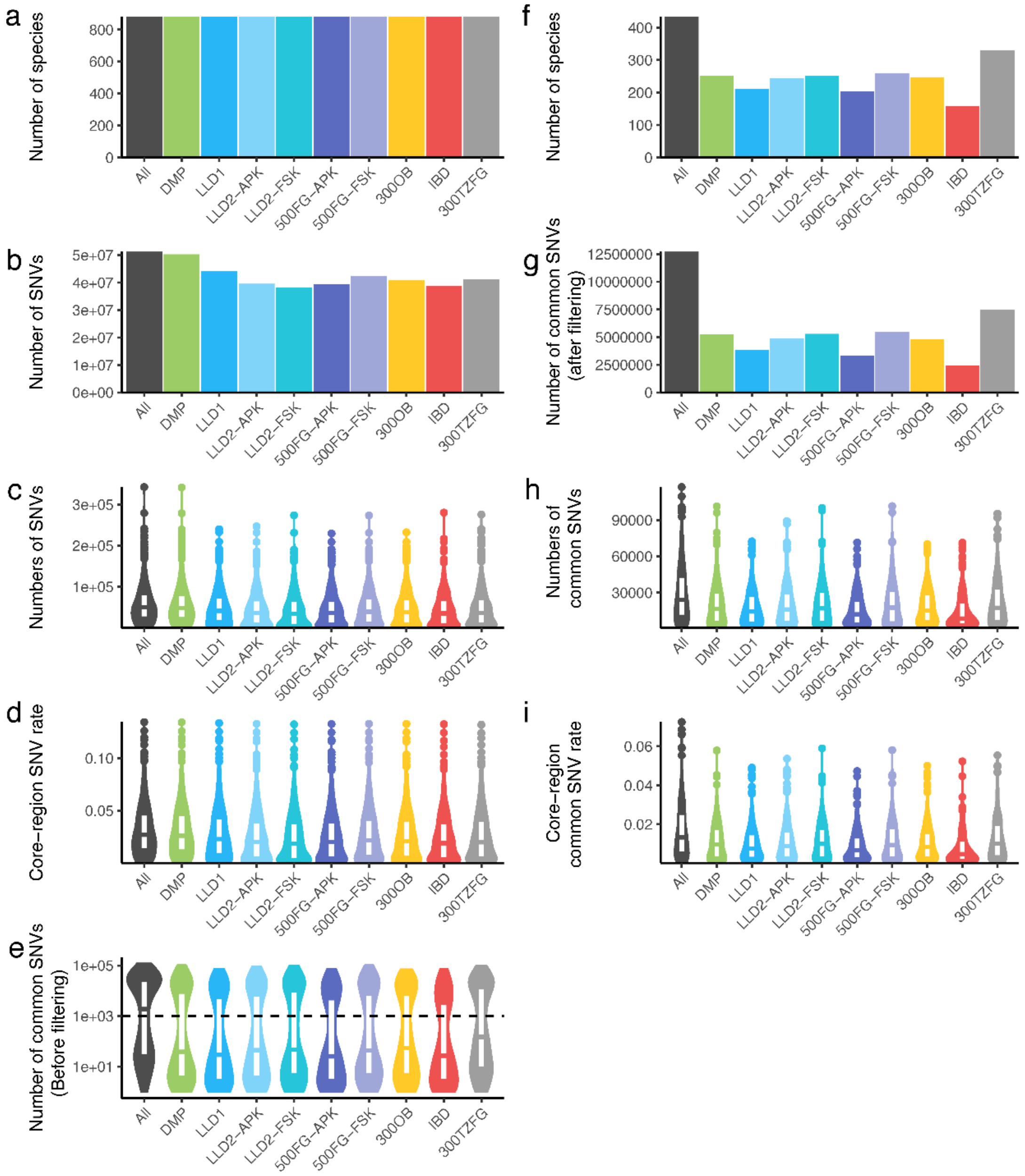

Fig. S1. Overview of microbial SNV data before and after filtering.

a, Total number of species detected with SNVs in each cohort before filtering.

b, Total number of gut microbial SNVs detected in each cohort before filtering.

c, Distributions of detected SNV numbers of each species in each cohort before filtering.

d, Proportion of detected SNVs in the core region of each species in each cohort before filtering.

e, Number of common SNVs in the core region of each species in each cohort before filtering. The distributions of common SNV numbers show two peaks that are separated at around 1,000 SNVs (dashed horizontal line).

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

g, Total numbers of common gut microbial SNVs kept for each cohort after filtering.

h, Distributions of common SNV numbers of each species in each cohort after filtering.

i, Proportion of common SNVs in the core region of each species in each cohort after filtering.

Inner boxplots of **c**, **d**, **e**, **h** and **i** represent summary statistics: the centre 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.

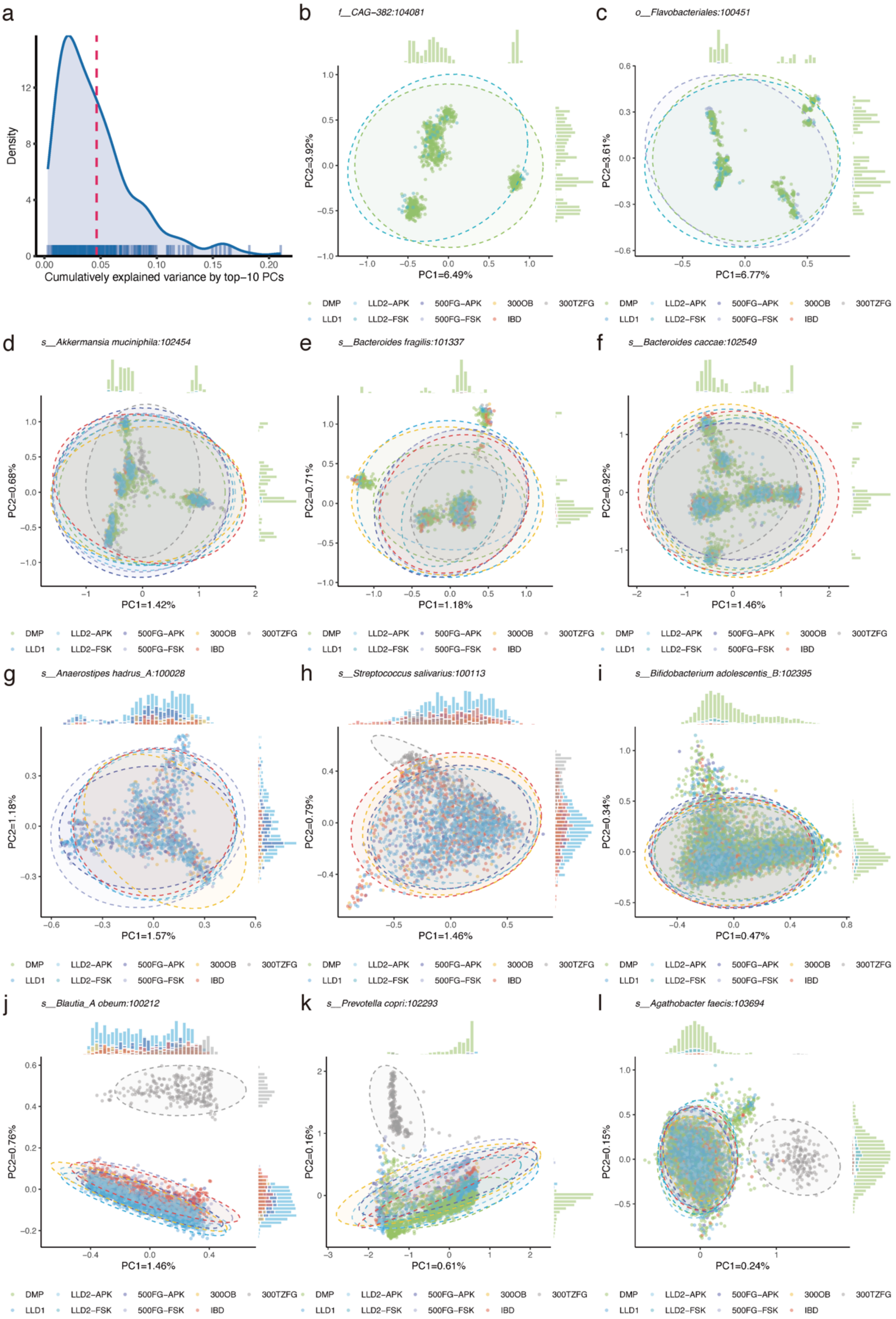

Fig. S2. Gut microbial population genetic structures.

a, Distribution of cumulatively explained variance by top 10 PCs of all species.

b and **c**, Populational genetic structures of f__CAG-382 (species ID: 104081) (**b**) and o__*Flavobacteriales* (species ID: 100451) (**c**), which showed highest cumulatively explained variance by the top 10 PCs.

d–**f**, Population genetic structures of *Akkermansia muciniphila* (species ID: 102454) (**d**), *Bacteroides fragilis* (species ID: 101337) (**e**), and *Bacteroides caccae* (species ID: 102549) (**c**), which showed clear clusters in PCoA plots.

g–**i**, Population genetic structures of *Anaerostipes hadrus* (species ID: 100028) (**g**), *Streptococcus salivarius* (species ID: 100113) (**h**), and *Bifidobacterium adolescentis* (species ID: 102395) (**i**), which showed continuous spread structure in the PCoA plot.

j–**l**, Population genetic structures of *Blautia obeum* (species ID: 100212) (**j**), *Prevotella copri* (species ID: 102293) (**k**), and *Agathobacter faecis* (species ID: 103694) (**l**), which showed geography-specific clusters in PCoA plot.

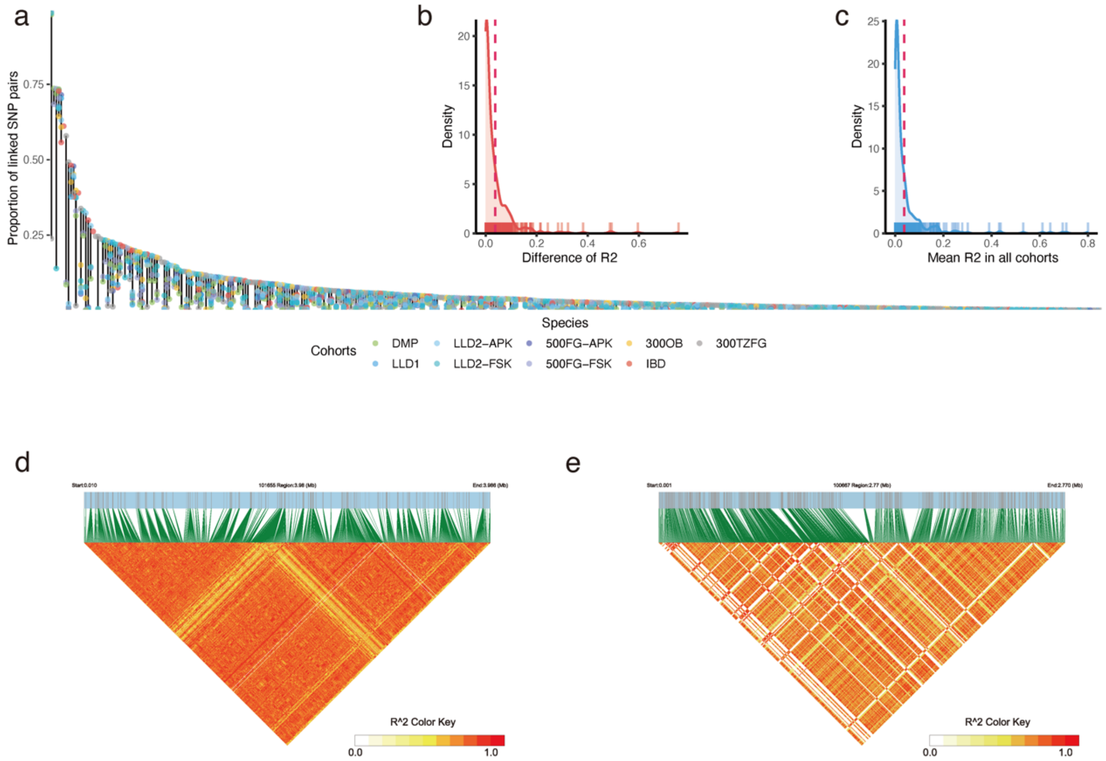

Fig. S3. LD structures of gut microbial genomes.

a, Proportion of linked SNV pairs in all randomly sampled SNV pairs of each species in different cohorts. Dot colours represent different cohorts. Vertical lines link the cohorts with the highest and lowest proportion of linked SNV pairs.

b, Distribution of the difference in the proportion of linked SNVs between the cohorts showing the highest and lowest proportion. Red vertical line indicates the mean value.

c, Distribution of average genome-wide LD (R^2). Red vertical line indicates the mean value.

d and **e**, Genome-wide LD heatmap of g__Butyricimonas (species ID:101655) (**d**) and g__UBA7173 (species ID:100667) (**e**).

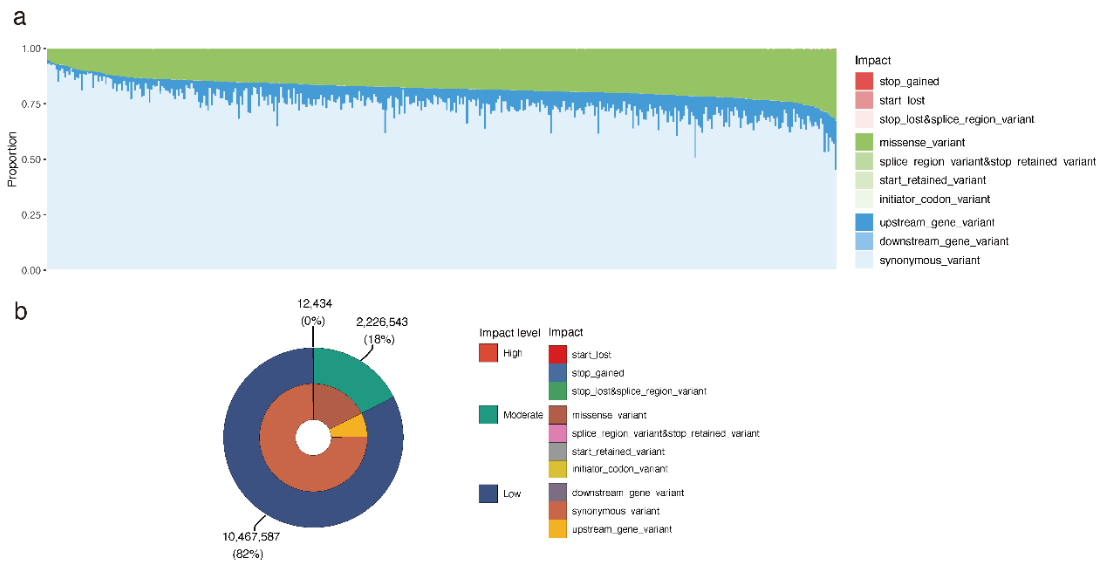

Fig. S4. Gut microbial SNV impact on gene product.

a, Proportion of SNV impact classes in each species (see key for impacts). Species are sorted by the proportion of low impact SNVs (upstream gene variants, downstream gene variants, and synonymous variants).

b, Overall proportion of gut microbial SNV impact classes and impact levels.

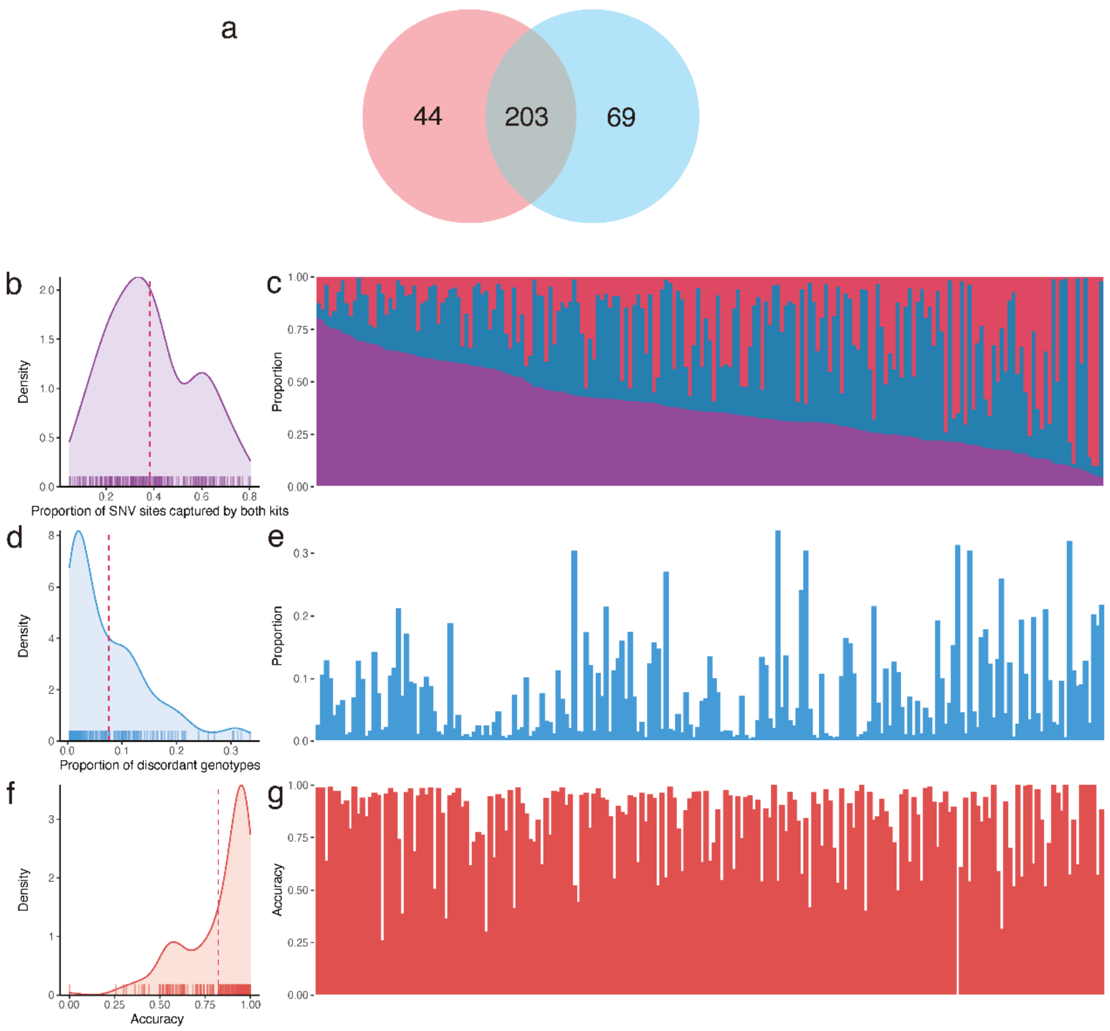

Fig. S5. Impact of DNA isolation kits on metagenotyping and gut microbial population genetic structure.

a, Number of species with >1000 common SNVs detected by APK (red) and FSK (blue). 203 species were captured by both kits.

b and **c**, We calculated the number of SNVs in union and intersection sets of the APK and FSK datasets for each species. The density plot (**b**) shows the distribution of the proportion of intersection SNV in union set SNVs. The red vertical line represents the mean proportion. The bar plot (**c**) shows the proportion of intersection SNVs (purple), APK-specific SNVs (red), and FSK-specific SNVs (blue) in the union set SNV. Each column is a species. Species were sorted by the proportion of intersection SNVs.

d and **e**, For each species, we calculated the average proportion of SNVs with a discordant genotype between the APK and FSK datasets of all samples. The density plot shows the distribution of the average proportion of the discordant genotypes (**d**). The red vertical line represents the mean proportion. The bar plot (**e**) shows the average proportion of SNVs with a discordant genotype between APK and FSK. Each column is a species. The order of species is identical to (**c**).

f and **g**, The accuracy was defined as the rate of samples that showed the least genetic dissimilarity with its corresponding paired sample processed by another DNA isolation kit. The density plot (**f**) shows the distribution of accuracy. The red vertical line represents the mean accuracy. The bar plot (**g**) shows the accuracy. Each column is a species. The order of species is identical to **c**.

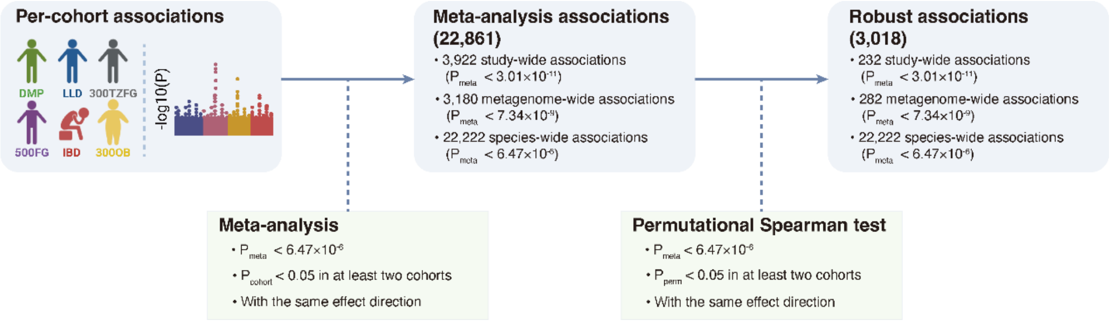
Per-cohort associations
DMP
LLD
300TZFG
500FG
IBD
300OB
-log10(P)
Meta-analysis associations (22,861)
• 3,922 study-wide associations (P_{meta} < 3.01×10^{-11})
• 3,180 metagenome-wide associations (P_{meta} < 7.34×10^{-9})
• 22,222 species-wide associations (P_{meta} < 6.47×10^{-6})
Robust associations (3,018)
• 232 study-wide associations (P_{meta} < 3.01×10^{-11})
• 282 metagenome-wide associations (P_{meta} < 7.34×10^{-9})
• 22,222 species-wide associations (P_{meta} < 6.47×10^{-6})
Meta-analysis
• P_{meta} < 6.47×10^{-6}
• P_{cohort} < 0.05 in at least two cohorts
• With the same effect direction
Permutational Spearman test
• P_{meta} < 6.47×10^{-6}
• P_{perm} < 0.05 in at least two cohorts
• With the same effect direction

Fig. S6. Workflow of meta-analysis and selection of robust associations.

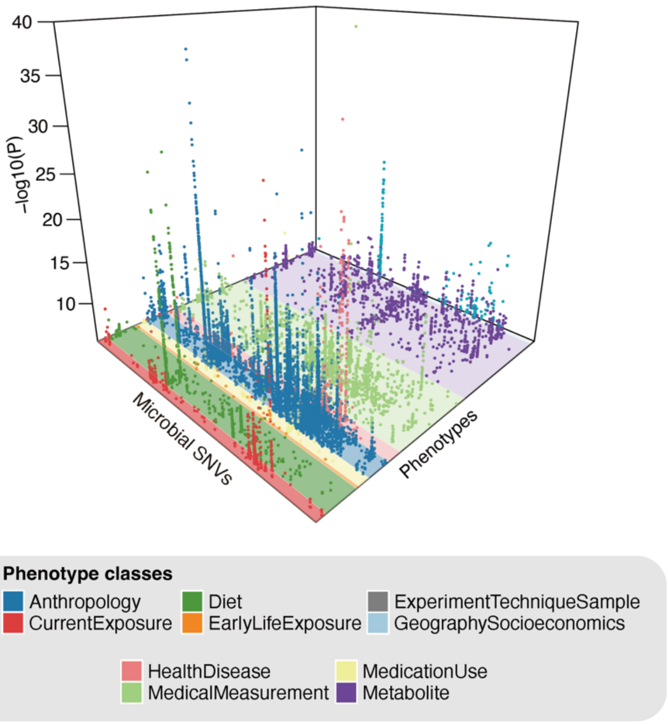

Fig. S7. Three-dimensional Manhattan plot showing 22,861 replicable associations across all phenotypes and all species before quality control. Colours indicate the categories of phenotypes.

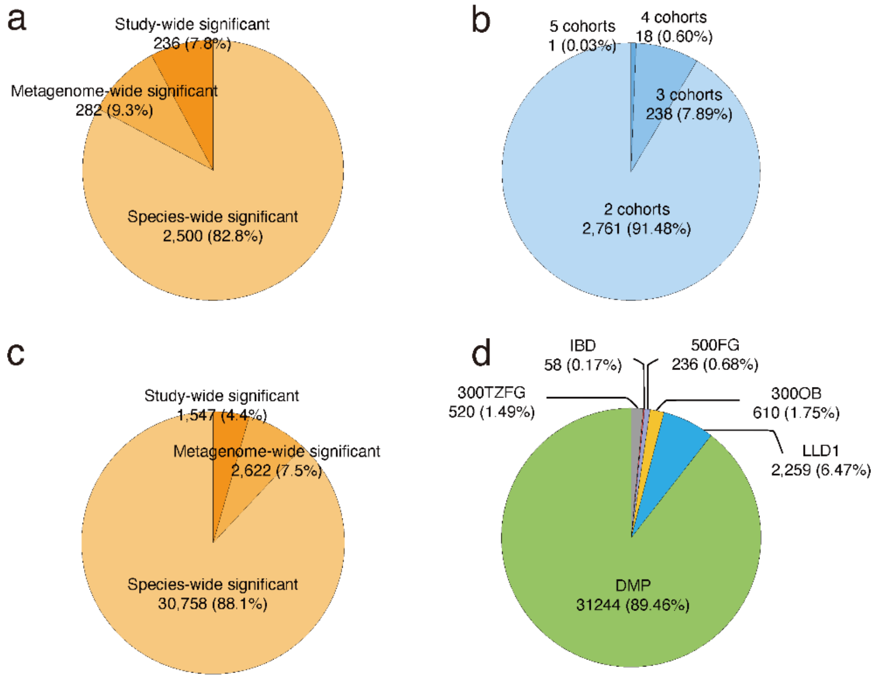

Fig. S8. Summary of SNV-MWAS signals.

a and **c**, Proportions of associations with different significance level for replicable associations (**a**) and single-cohort associations (**c**).

b, Proportion of replicable associations replicated in several cohorts.

d, Proportion of single-cohort associations in different cohorts.

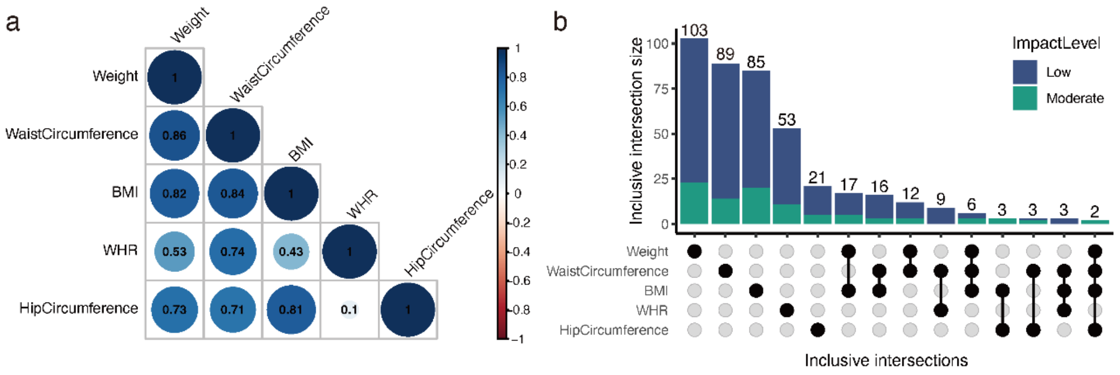

Fig. S9. Shared signals between body-shape-related phenotypes.

a, Correlations between body-shape-related phenotypes.

b, Shared associated SNV numbers between body-shape-related phenotypes.

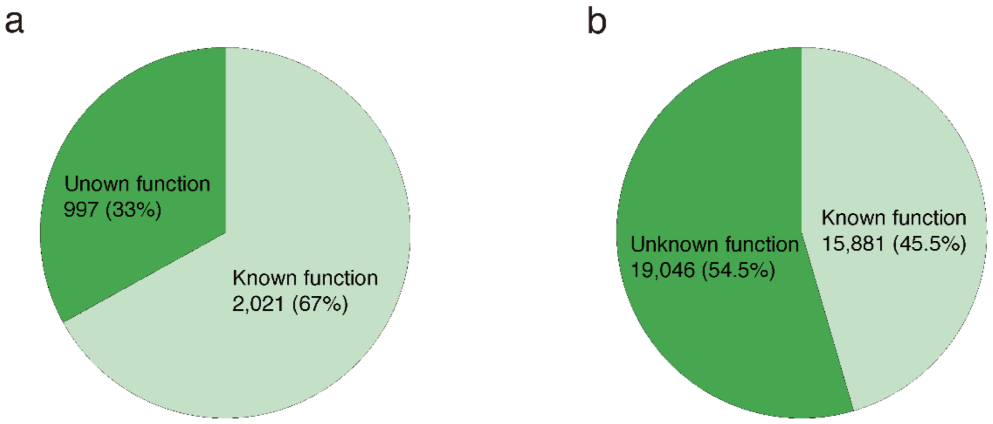

Fig. S10. Proportion of replicable associations involving gut microbial genes with known or unknown function.

a and **b**, Proportion of replicable associations involving gut microbial genes with known or unknown function in replicable associations (**a**) and single-cohort associations (**b**).

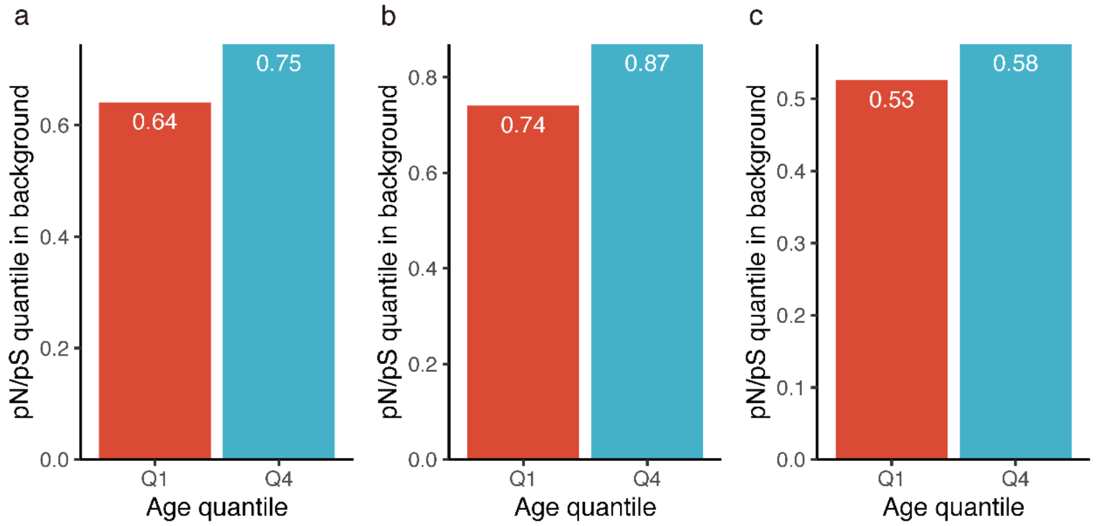

Fig. S11. Positive selection of host age on the sarA_2/oppA gene of *F. prausnitzii*_I.

a–c, In the DMP (**a**), LLD1(**b**) and 300TZFG (**c**) cohorts, we assigned the samples with ages in the first quantile (Q1) and fourth quantile (Q4) to younger and older groups, respectively, and used the median pN/pS of all genes of *F. prausnitzii*_I within each group as a background. We then compared the quantile of median pN/pS of sarA_2/oppA in the background pN/pS distribution between the older group (Q4) and young group (Q1). We also observed similar trends in the LLD1 and 300TZFG cohorts

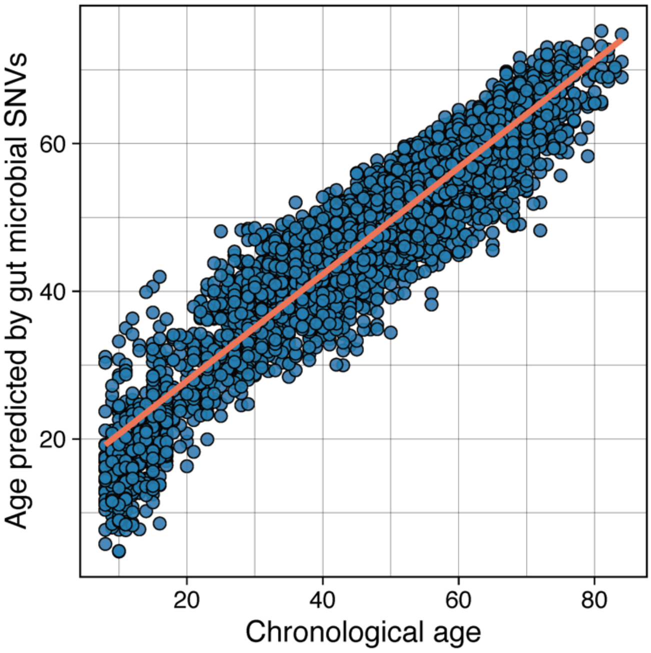

Fig. S12. Correlation between host age predicted by gut microbial SNVs and chronological age in the final fitted model in training dataset (*R^2* = 0.88).

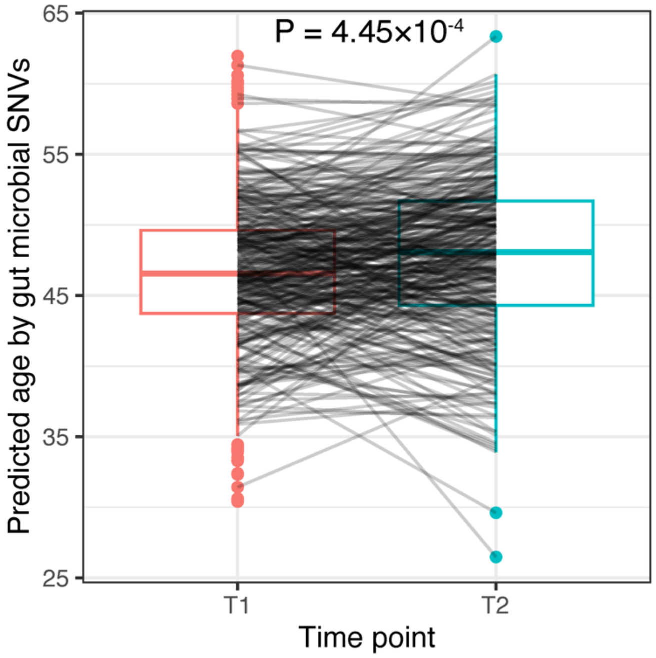

Fig. S13. Comparison of the age predicted by gut microbial SNVs for paired samples collected 4 years apart in LLD (One-tailed paired Wilcoxon rank sum test).

Boxplot represents summary statistics: the centre 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. Sample pairs from the same individuals at different time points are connected by grey lines.

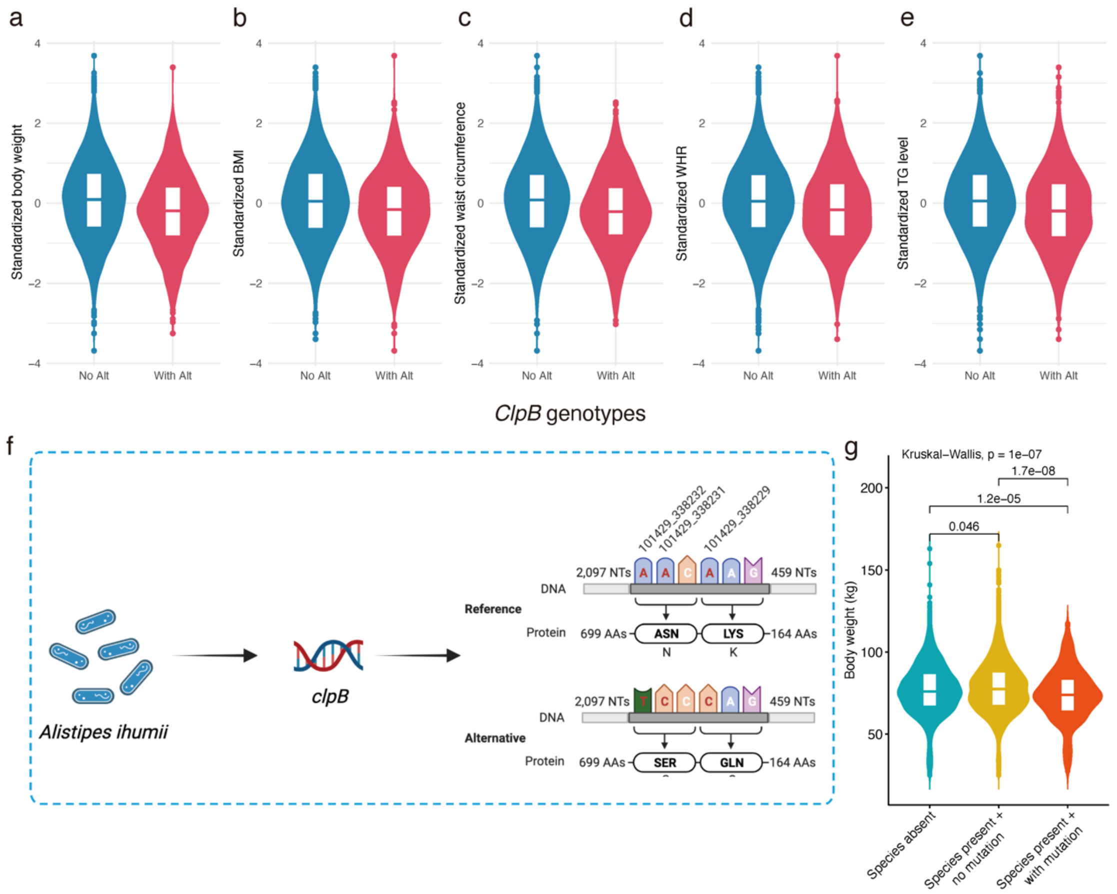

Fig. S14. Pleiotropic associations with SNVs in the *clpB* gene of *Alistipes ihumii*.

a-**e**, Associations between the *clpB* genotype of *A. ihumii* (Species ID: 101429) and body weight (**a**), BMI (**b**), waist circumference (**c**), waist-to-hip ratio (WHR) (**d**), and triglyceride (TG) level (**e**).

f, Amino acid changes in the ClpB protein of *A. ihumii* (Species ID: 101429) caused by three fully linked SNVs.

g, Comparison of body weight (kg) between host individuals without *A. ihumii*, with non-mutated *A. ihumii*, and with mutated *A. ihumii* (Kruskal-Wallis rank sum test). Inner boxplots of **a–e** and **g** represent summary statistics: the centre 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.

References

1. Karcher, N. *et al.* Genomic diversity and ecology of human-associated Akkermansia species in the gut microbiome revealed by extensive metagenomic assembly. *Genome Biol.* **22**, 209 (2021).

2. Wallace, M. J., Jean, S., Wallace, M. A., Burnham, C.-A. D. & Dantas, G. Comparative Genomics of Bacteroides fragilis Group Isolates Reveals Species-Dependent Resistance Mechanisms and Validates Clinical Tools for Resistance Prediction. *mBio* **13**, e03603-21 (2022).

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

4. 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).

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