Supplementary Figures

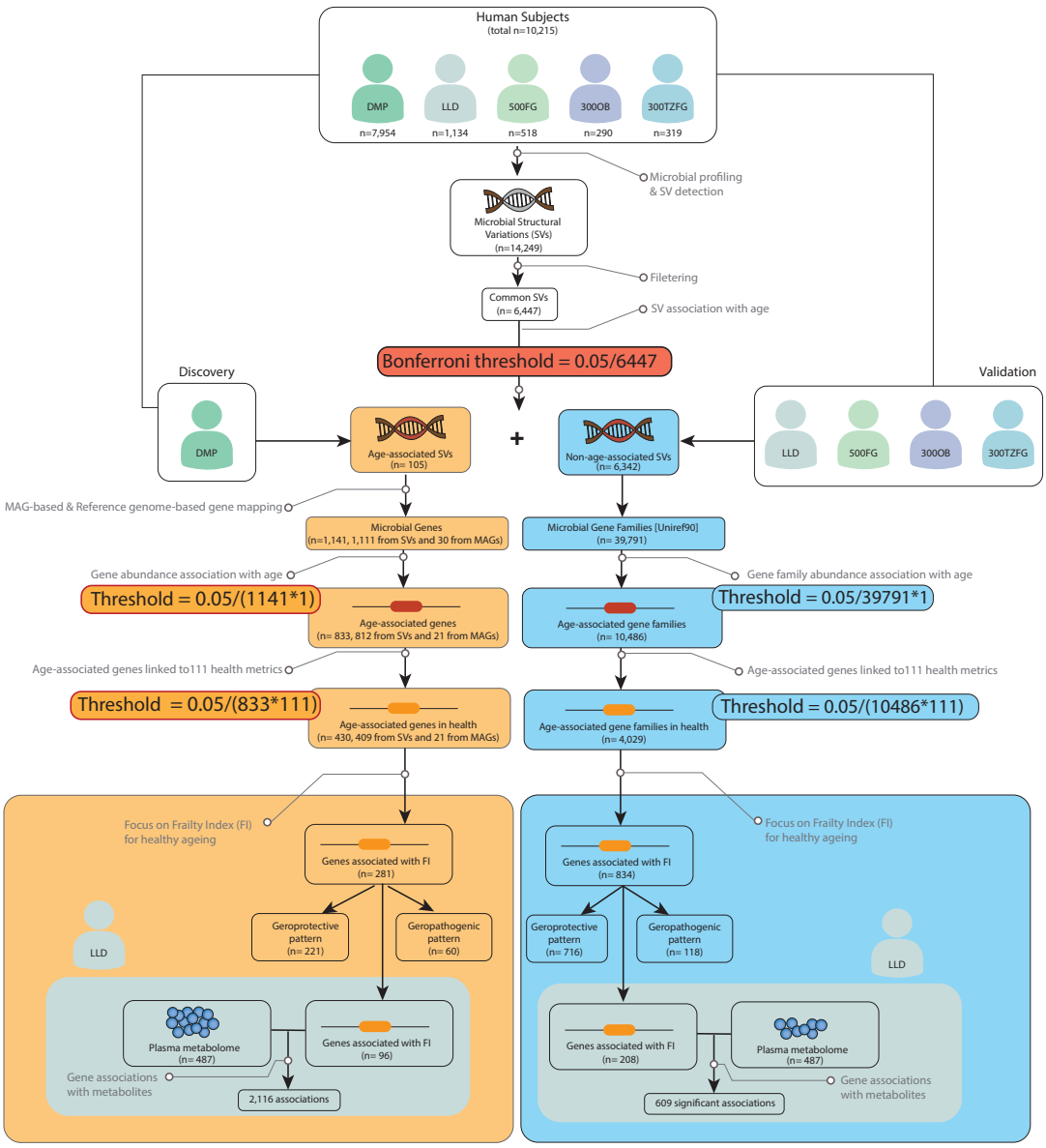

Figure S1. Study workflow. This study explored the relationship between host age and gut microbial structural variations (SVs) in 10,215 individuals across five cohorts (the Dutch Microbiome Project (DMP), N = 7,954; Lifelines-DEEP (LLD), N = 1,134; the 500 Functional Genomics Project (500FG), N = 518, 300-Obesity (300OB), N = 290; and 300 Tanzania functional genomics (300TZFG), N = 319). A total of 14,249 SVs were profiled, including 6,447 found to be common SVs after filtering. Using the DMP as the discovery cohort and the other cohorts for validation, we conducted an association analysis between SVs and age and identified 105 replicable age-associated SVs. To map genes from SVs, we used both reference genomes and metagenome-assembled genomes (MAGs) (especially for *Oscillibacter sp. ER4*). As a result, 1,111 genes were extracted from age-associated SVs, along with 30 additional age-associated genes specifically from *Oscillibacter sp. ER4* MAGs. We profiled the abundance of these 1,141 genes for the entire metagenomic sample, and 833

showed a significant association with host age. Among these 830 genes, 430 were further associated with at least one health metric, including 281 linked to frailty index (FI). For 6,342 SVs not associated with age, we further extracted and profiled abundance levels for 39,791 families. Out of them, 10,486 families were associated with age. 4,029 of them were further associated with at least one health matric, including 834 linked to FI. These age and FI-associated gene and gene families were classified as having either a geroprotective or geropathogenic pattern. In addition, we linked them to plasma metabolite levels in the LLD cohort, if data available and FI associations replicable.

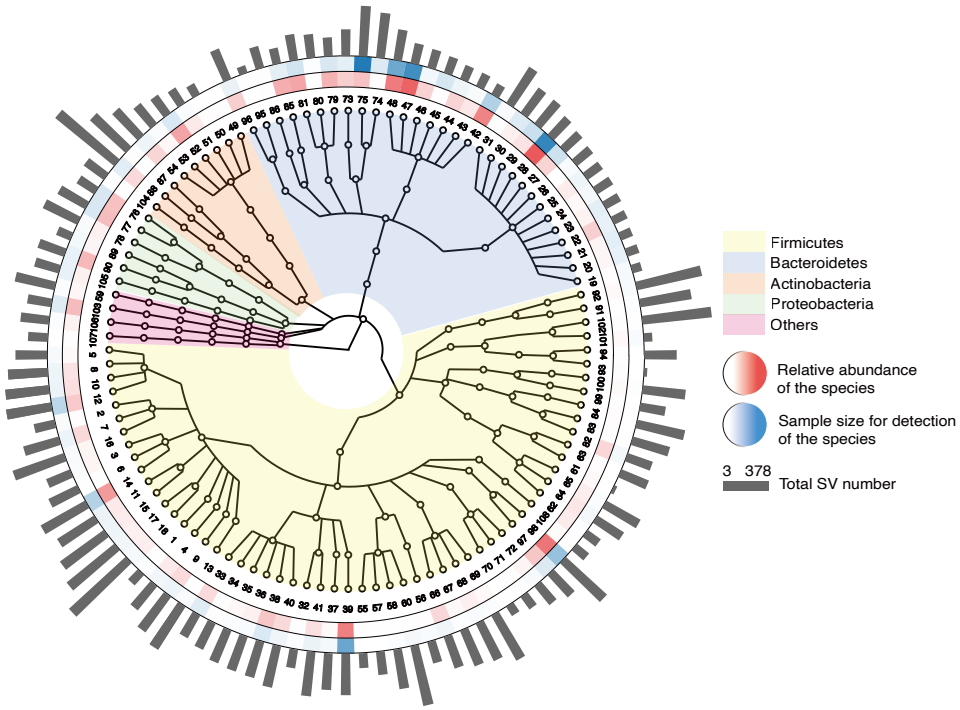

Figure S2. Cladogram of the SVs profiled in our study. In total, 14,249 SVs were identified in the genomes of 108 reference species, including 10,350 dSVs and 3,899 vSVs, with 3–378 SVs found per species. Inner cladogram indicates the taxonomic relationship of the 108 reference species. Different background colours indicate different phyla. The species number corresponds to the taxonomic ID as detailed in Table S1. Red boxes indicate the average relative abundance of the species in the five cohorts. Blue boxes indicate the number of samples in which each species was detected. The outside bar shows the total SV number detected per species in the five cohorts.

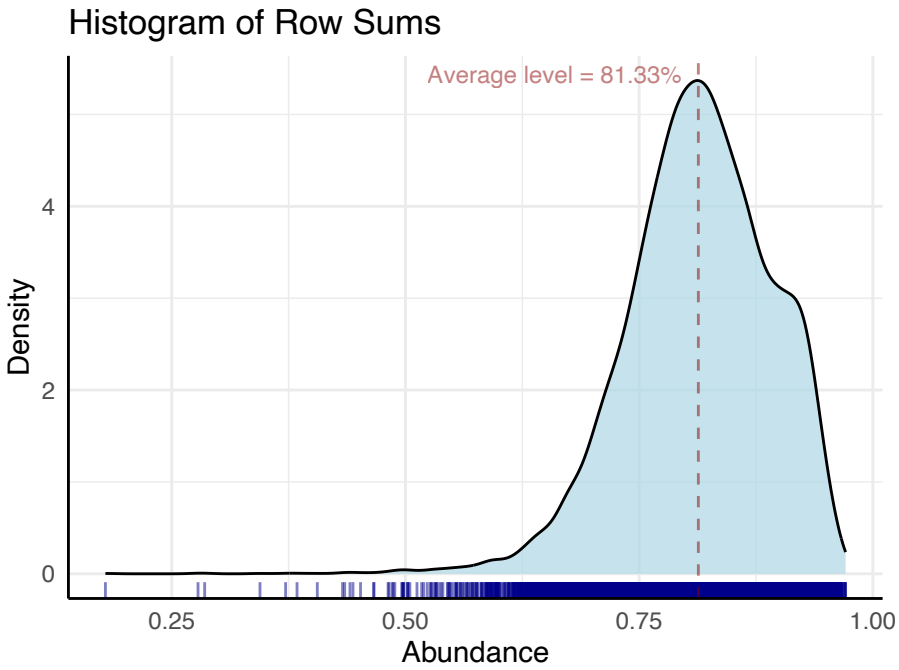

Figure S3. Distribution of the entire microbial composition made up by the 108 species in which SVs were profiled. Together, these 108 species make up 81.33% of the entire microbial composition on average (range 17.90–97.09% in different hosts). The x-axis indicates the collective abundance of 108 species, and the y-axis indicates the density of sample distribution.

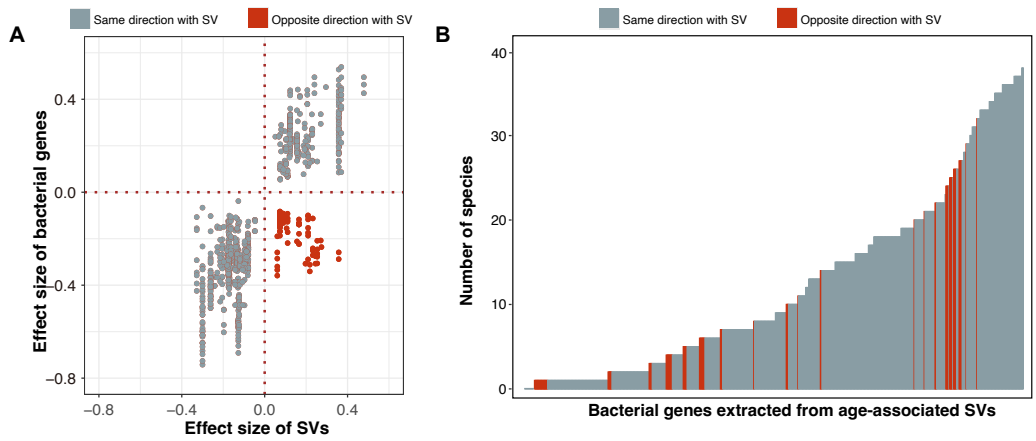

Figure S4. Consistency between gene–age and SV–age associations. (A) Scatter plot showing the relationship between the effect sizes of age associations for 812 age-associated bacterial genes and the SVs from which they were extracted. Each point represents one gene. Grey = same direction. Red = opposite direction. (B) Bar plot showing the number of species significantly correlated to the abundance of each bacterial gene (genes are extracted from age-associated SVs). Genes are ordered by the number of correlated species, with colours indicating whether the direction of the gene's age association matches (red) or differs (grey) from that of the parent SV.

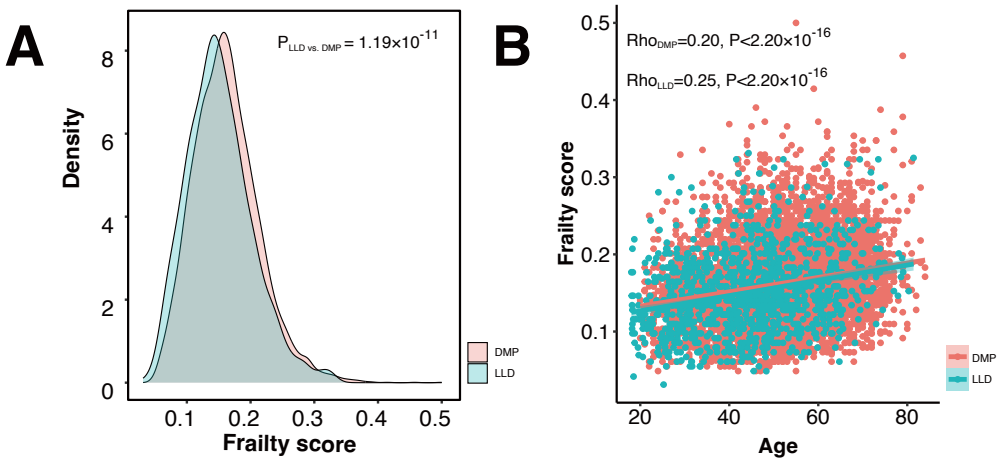

Figure S5. Frailty index (FI) and general health score information. (A) The distribution of the FI in the DMP and LLD cohorts. The FI level is significantly higher in the DMP cohort compared to the LLD cohort (Wilcoxon test, $P_{LLD\ vs.\ DMP} = 1.19\times10^{-11}$). (B) Dot-plots for correlations between age and FI in the DMP and LLD cohorts (Spearman correlation test, $P_{LLD} < 2.20\times10^{-11}$, $P_{DMP} < 2.20\times10^{-11}$). Each dot represents one sample. Red and green dots represent DMP and LLD, respectively.

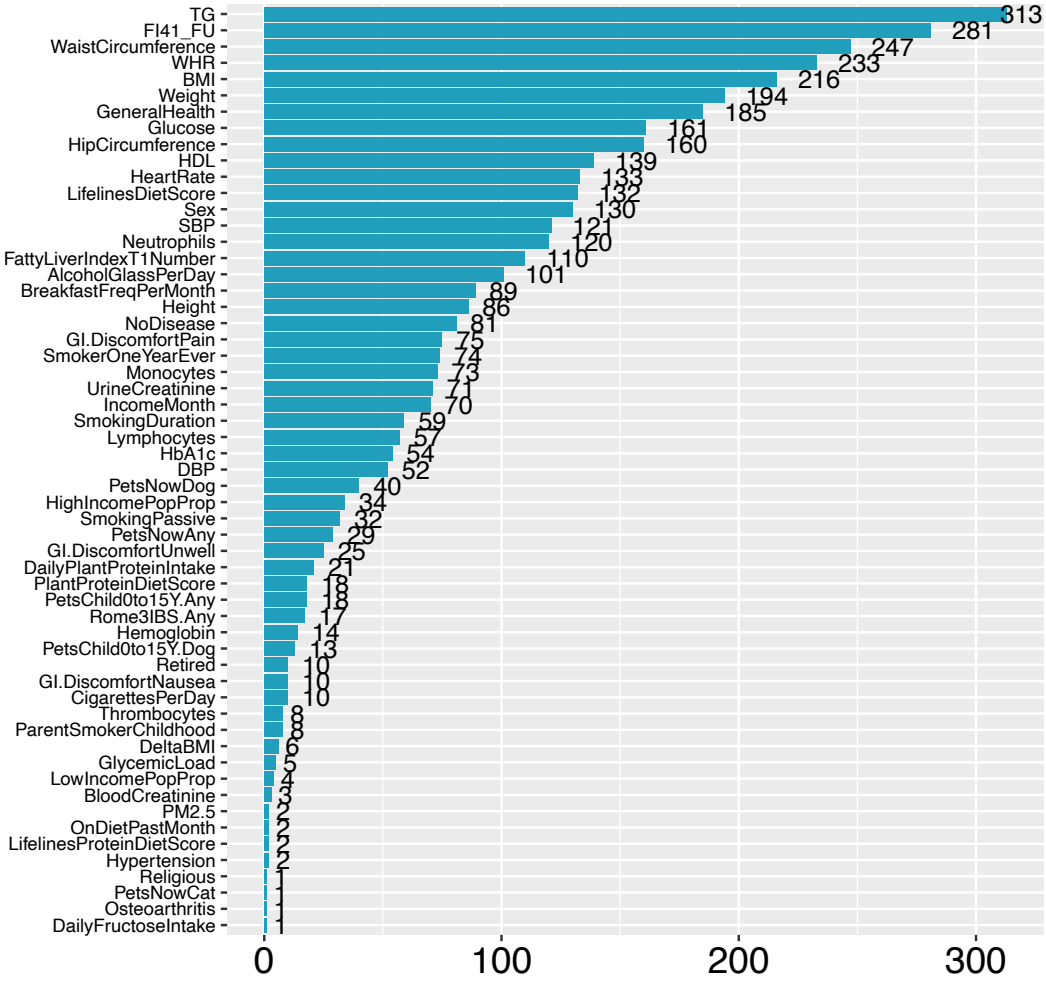

Figure S6. Number of significant associations between health metrics and age-associated genes. Bar plot shows the number of significant associations with bacterial genes (x-axis) across different phenotypes (y-axis). The definition of the abbreviations in this figure are indicated in the **Table S18**.

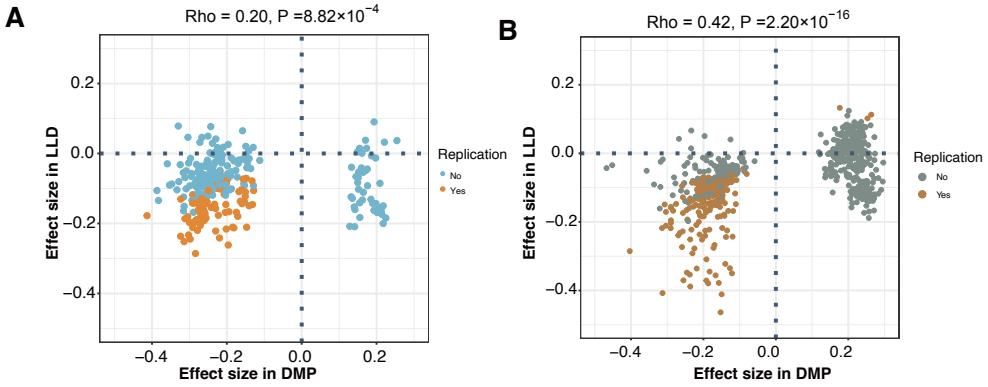

Figure S7. Validation of frailty index (FI)-associated genes in the LLD cohort. (A) Scatter plot showing the effect sizes of gene–FI associations in the DMP cohort (x-axis) versus the LLD cohort (y-axis). Each point represents a bacterial gene (genes are extracted from age-associated SVs), with colours indicating genes that were replicable (blue) or non-replicable (orange) between cohorts. (B) As in (A) but highlighting gene families extracted from non-age-associated SVs.

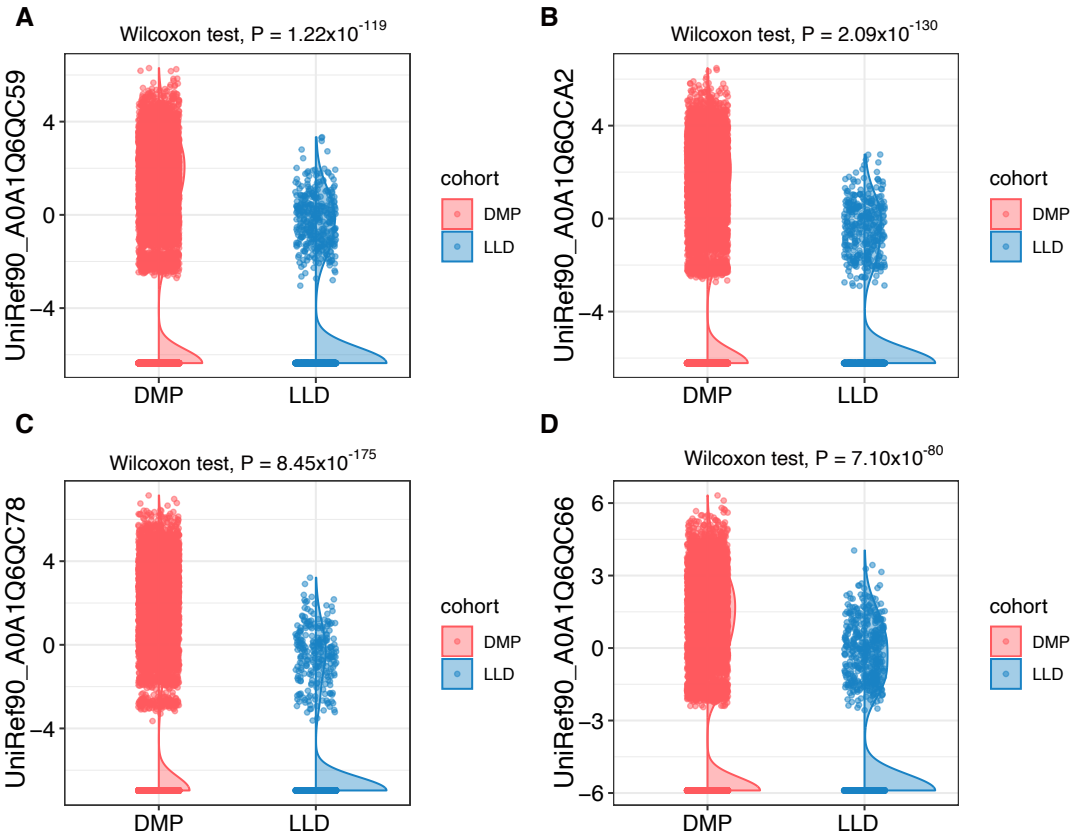

Figure S8. Distribution of four genes related to biotin biosynthesis (log-transformed) in the DMP and LLD cohorts. (A) UniRefef90_A0A1Q6QC59. (B) UniRefef90_A0A1Q6QCA2. (C) UniRefef90_A0A1Q6QC78. (D) UniRefef90_A0A1Q6QC66.

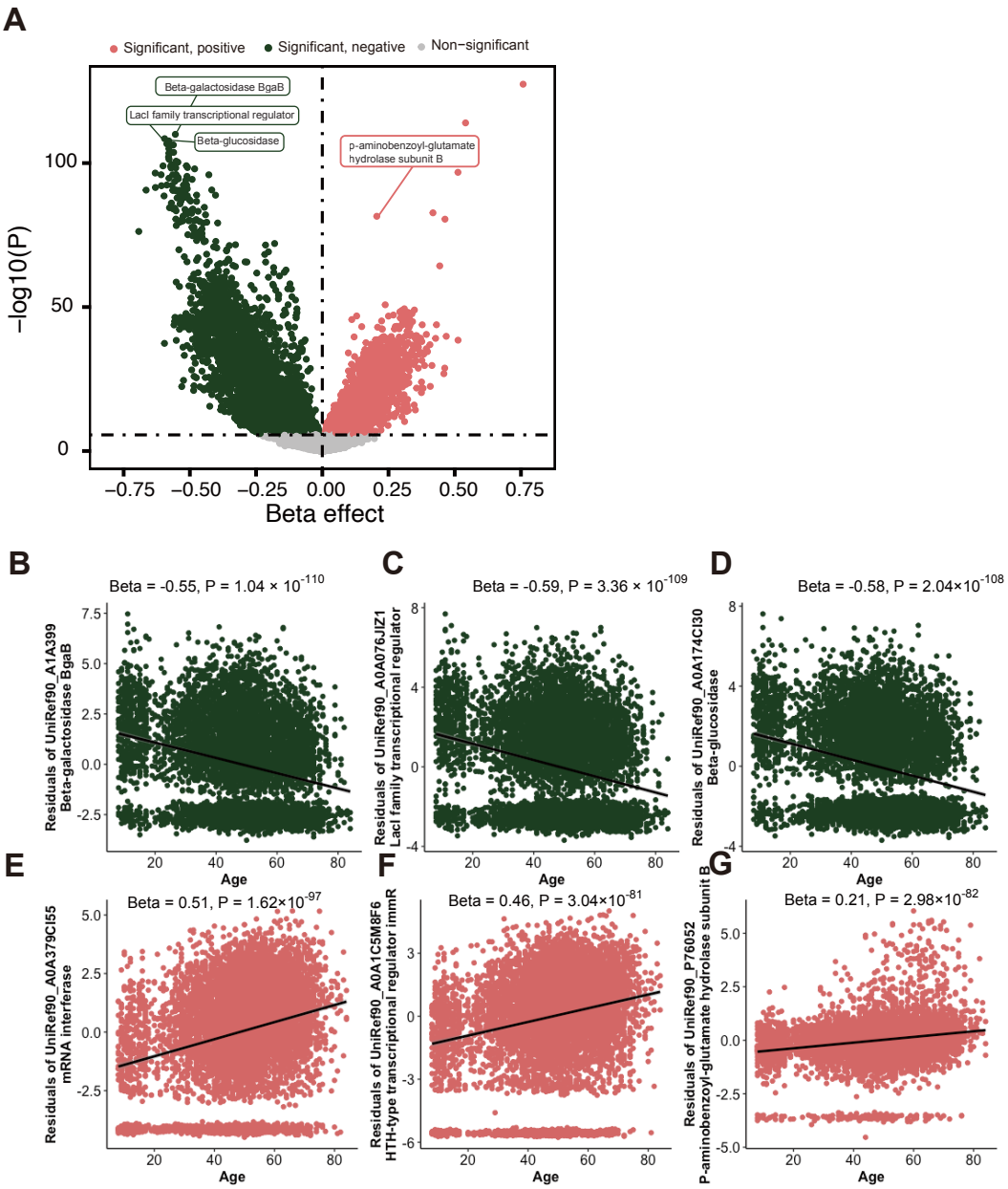

Figure S9. Gene families in non-age-associated SVs show associations with host age. (A) Volcano plot showing the association between bacterial gene families and host age. X-axis represents the beta effect for the association. Y-axis represents the P value for the association. Dot colour indicates association direction (red = positive, dark green = negative, grey = signal did not pass multiple-hypothesis correction). The functional annotations of the top signals are indicated in the boxes. (B−G) Scatter plots showing the strongest positive and negative associations between age and (B) UniRef90_A1A399, (C) UniRef90_A0A076JIZ1, (D) UniRef90_A0A174CI30, (E) UniRef90_A0A379CI55, (F) UniRef90_A0A1C5M8F6 and (G) UniRef90_P76052. In (B−G), the y-axis refers to the residual of log-transformed gene abundance after correcting for covariates (including sex, read counts and DNA concentration), each dot represents one sample, and x-axis refers to age. Chronological age is shown here for visualization, however age was empirical-normal-quantile-transformed for association analysis.

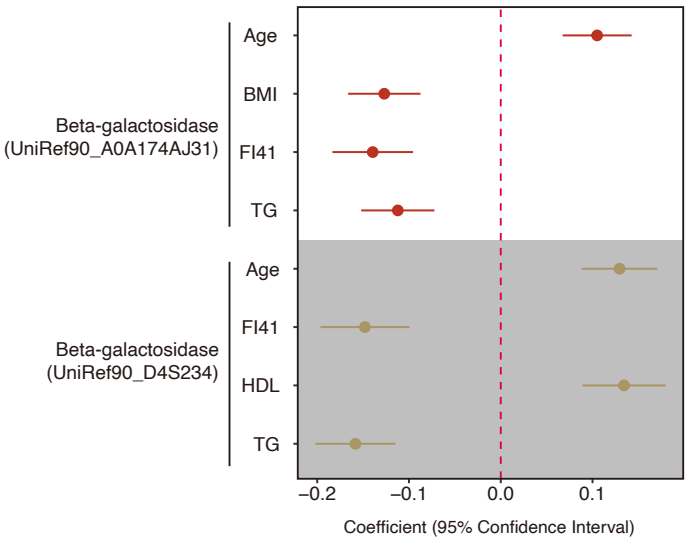

Figure S10. Age-associated gene families are potentially indicative of host health. Forest plot indicating the coefficients for the associations of bacterial gene families (extracted from non-age-associated SVs) with age- and health-related metrics.

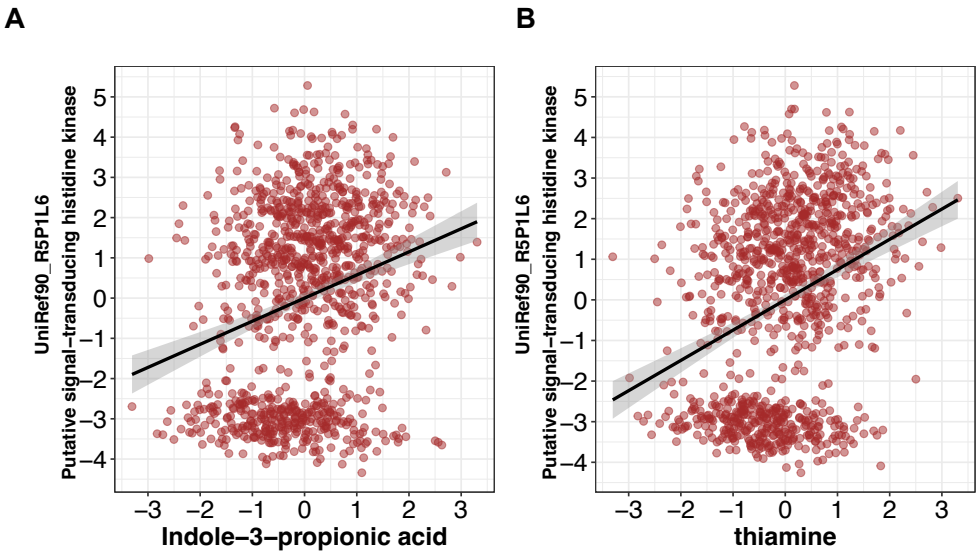

Figure S11. Age-related bacterial gene families associate with the blood level of tryptophan-related metabolites and thiamine. (A−B) Scatter plots showing the top associations of UniRef90_R5P1L6 with (A) indole-3-propionic acid and (B) thiamine. Metabolite levels were empirical-normal-quantile transformed. Gene family abundance was log-transformed and then corrected for covariates including age, sex and read counts. Each dot represents one sample.

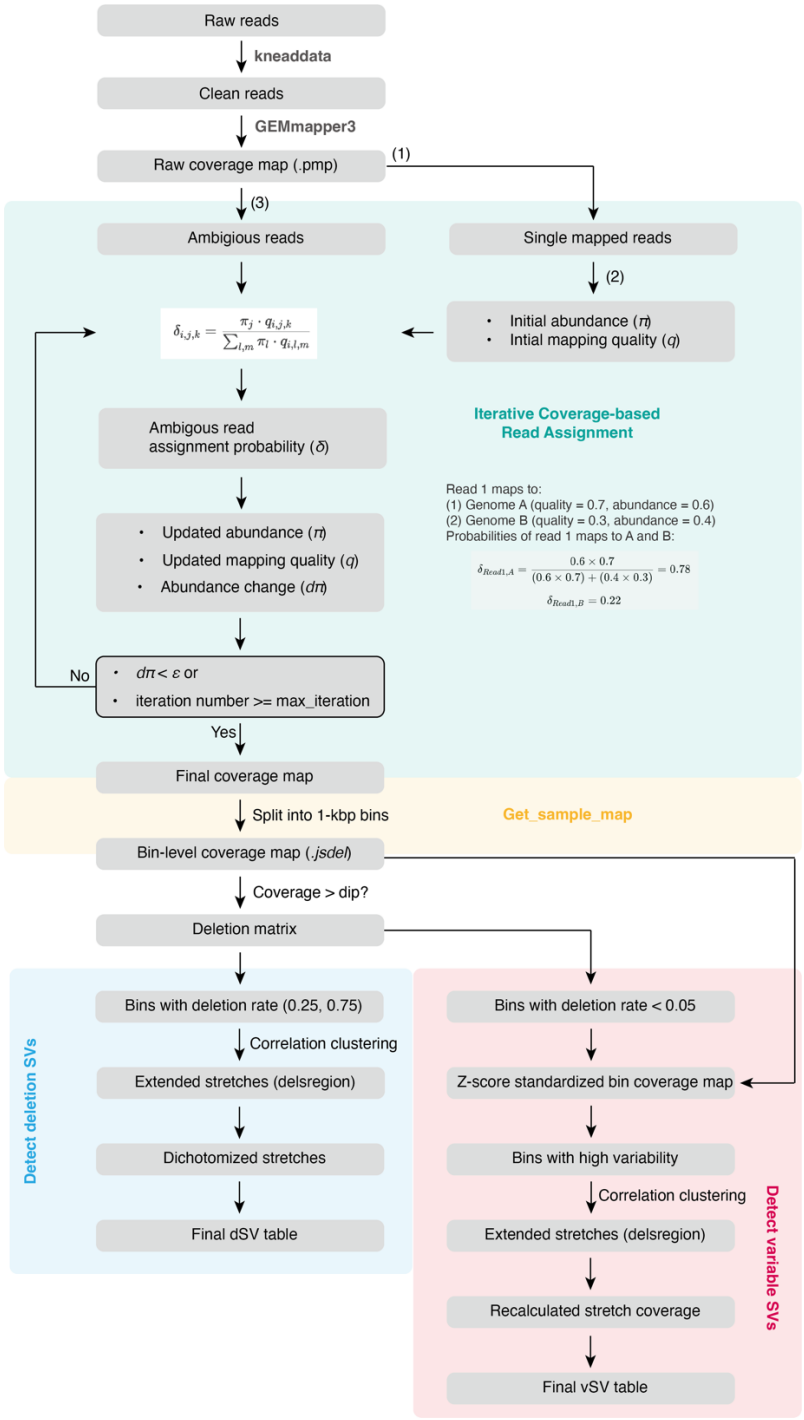

Figure S12. Overview of SV Detection in SGVFinder. Step 1. ***Read mapping & coverage calculation.*** Raw metagenomic sequencing reads are processed using Kneaddata to remove low-quality reads and host contamination. Reads are then mapped to the reference genome database using GEMmapper. A raw coverage map is generated representing how sequencing reads are distributed across microbial genomes. **Step 2.** ***Resolve ambiguous reads with multiple alignments according to mapping quality and genomic coverage using the ICRA and reassign ambiguous reads to the most likely reference with high accuracy.*** **2.1** Based on the raw coverage map generated by GEMmapper, mapped reads are divided into two categories: *single-mapped reads* that align uniquely to a single location in the genome and *ambiguously mapped reads* that align to multiple locations across different genomes or strains. **2.2** The ICRA

algorithm resolves ambiguous read mappings by assigning reads probabilistically based on species abundance and mapping quality. For each ambiguous read R_i, ICRA computes the probability (δ) that it originated from genome j at position k:

$$\delta_{i,j,k} = \frac{\pi_j \cdot q_{i,j,k}}{\sum_{l,m} \cdot \pi_l \cdot q_{i,l,m}}$$

Where π_j = initial abundance estimates of genome j (calculated from uniquely mapped reads), $q_{i,j,k}$ = mapping quality score for read i aligning to genome j at position k, and l, m = all possible genomes and alignment positions for read i. **2.3** After initial reads assignment, ICRA updates species abundance π_j and mapping quality $q_{i,j,k}$ by incorporating the probabilistically assigned ambiguous reads. ICRA then calculates the change in abundance $\Delta\pi$. **2.4** The new π values are used to recalculate probabilities δ for all ambiguous reads. This process is repeated until the change in abundance ($\Delta\pi$) falls below a predefined threshold (ε) OR a maximum number of iterations is reached. Once convergence is achieved, a final coverage map is generated for each genome, which is then used for SV detection. ***Step 3. Splitting the reference genomes of each microbial species into 1-kbp genomic bins and examining the coverage of genomic bins across all samples***. To determine deletion SVs (dSVs) within each species, the genomic bins are classified as deleted or retained in each sample, with those deleted in 25–75% of samples kept in the analysis as raw dSVs. Raw dSVs that are highly correlated in co-occurrence are further merged into larger SV regions to produce the final dSV profile. To determine variable SVs (vSVs) within each species, the coverage of genomic bins within each sample is standardized using the Z-score approach. Each bin is then assessed across all samples, and those that are highly variable are kept as raw vSVs. Raw vSVs that are highly correlated in standardized coverage are further merged into large SV regions to produce the final vSV profile.

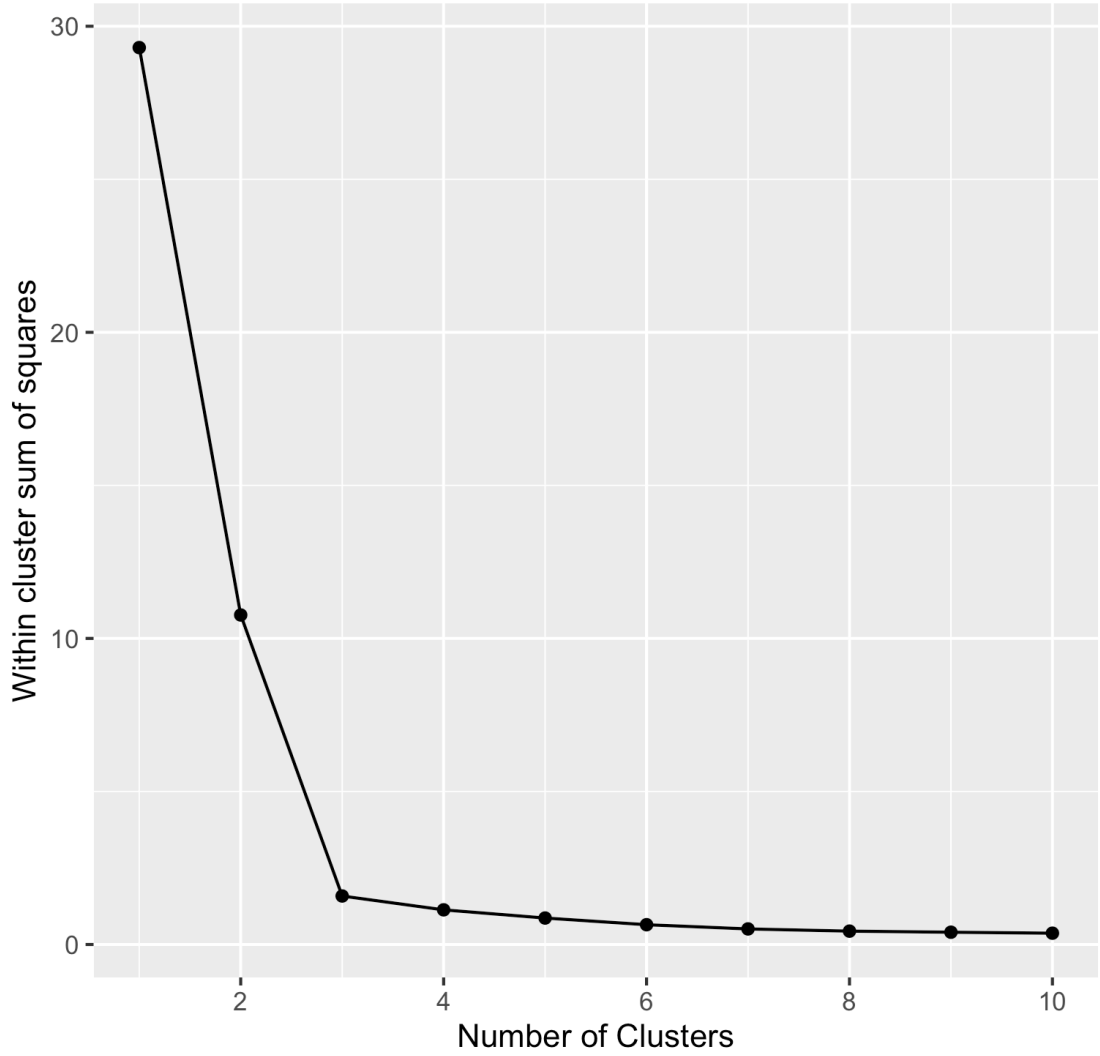

Figure S13. Elbow graph showing the within-cluster-sum-of-square values (y-axis) corresponding to different numbers of clusters (x-axis). Based on a Jaccard distance matrix of the prevalence of genes in metagenome-assembled genomes, the within-cluster sum of squares shows the largest change around three clusters when applying the K-means algorithm, indicating three as the optimal number of clusters (k=3).