Cell Genomics, Volume *2*

## Supplemental information

# Prokaryotic and viral genomes recovered from 787 Japanese gut metagenomes revealed microbial features linked to diets, populations, and diseases

Yoshihiko Tomofuji, Toshihiro Kishikawa, Yuichi Maeda, Kotaro Ogawa, Yuriko Otake-Kasamoto, Shuhei Kawabata, Takuro Nii, Tatsusada Okuno, Eri Oguro-Igashira, Makoto Kinoshita, Masatoshi Takagaki, Naoki Oyama, Kenichi Todo, Kenichi Yamamoto, Kyuto Sonehara, Mayu Yagita, Akiko Hosokawa, Daisuke Motooka, Yuki Matsumoto, Hidetoshi Matsuoka, Maiko Yoshimura, Shiro Ohshima, Shinichiro Shinzaki, Shota Nakamura, Hideki Iijima, Hidenori Inohara, Haruhiko Kishima, Tetsuo Takehara, Hideki Mochizuki, Kiyoshi Takeda, Atsushi Kumanogoh, and Yukinori Okada

**Supplemental figures**

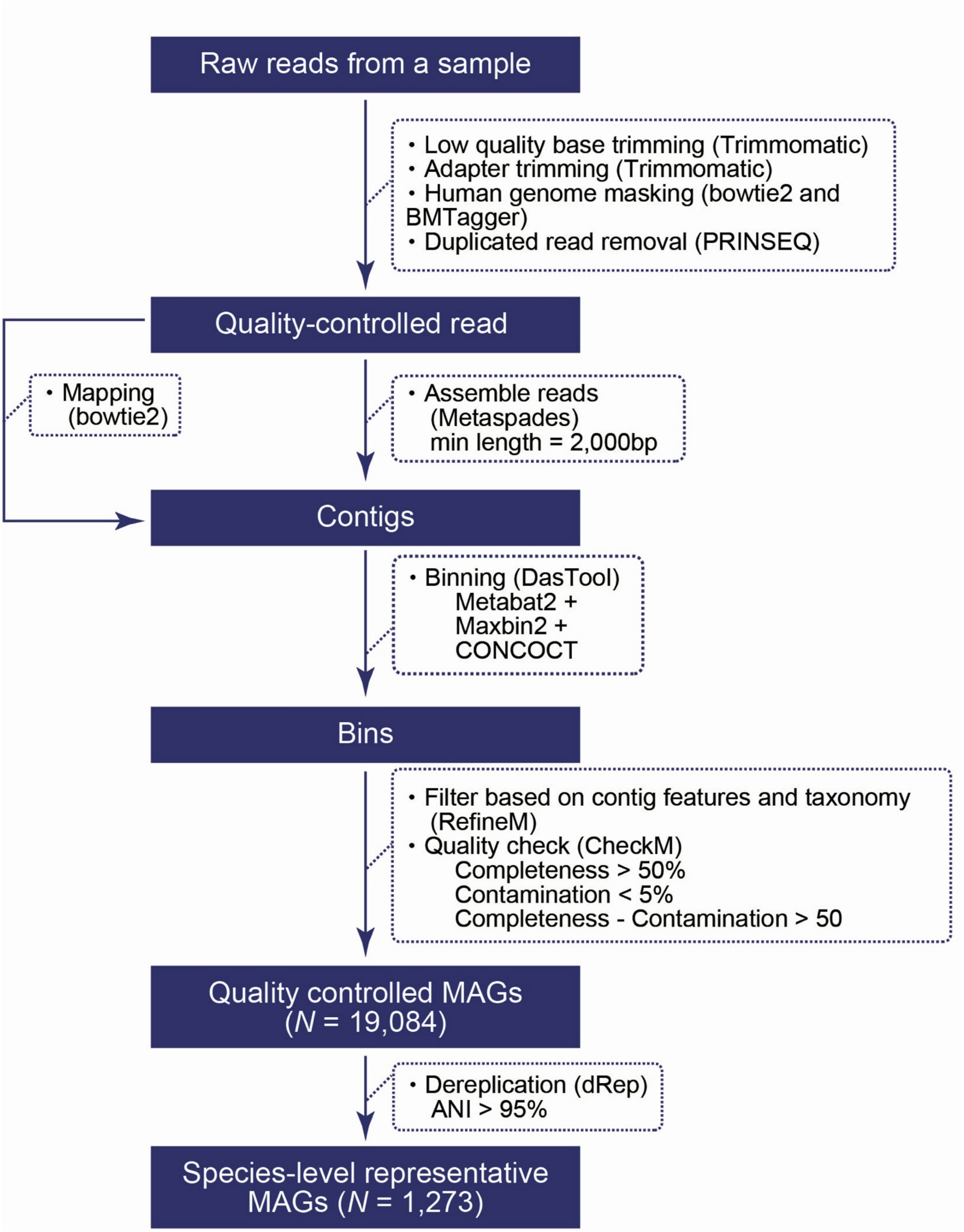
Raw reads from a sample
• Low quality base trimming (Trimmomatic)
• Adapter trimming (Trimmomatic)
• Human genome masking (bowtie2 and BMTagger)
• Duplicated read removal (PRINSEQ)
Quality-controlled read
• Mapping (bowtie2)
• Assemble reads (Metaspades) min length = 2,000bp
Contigs
• Binning (DasTool) Metabat2 + Maxbin2 + CONCOCT
Bins
• Filter based on contig features and taxonomy (RefineM)
• Quality check (CheckM) Completeness > 50% Contamination < 5% Completeness - Contamination > 50
Quality controlled MAGs (N = 19,084)
• Dereplication (dRep) ANI > 95%
Species-level representative MAGs (N = 1,273)

**Figure S1. A pipeline for reconstructing MAGs, related to Figure 1**

The metagenome shotgun sequencing data of the Japanese gut microbiome were processed following this pipeline, resulting in 19,084 MAGs comprising 1,273 species-level clusters. ANI, average nucleotide identity; MAG, metagenome-assembled genome.

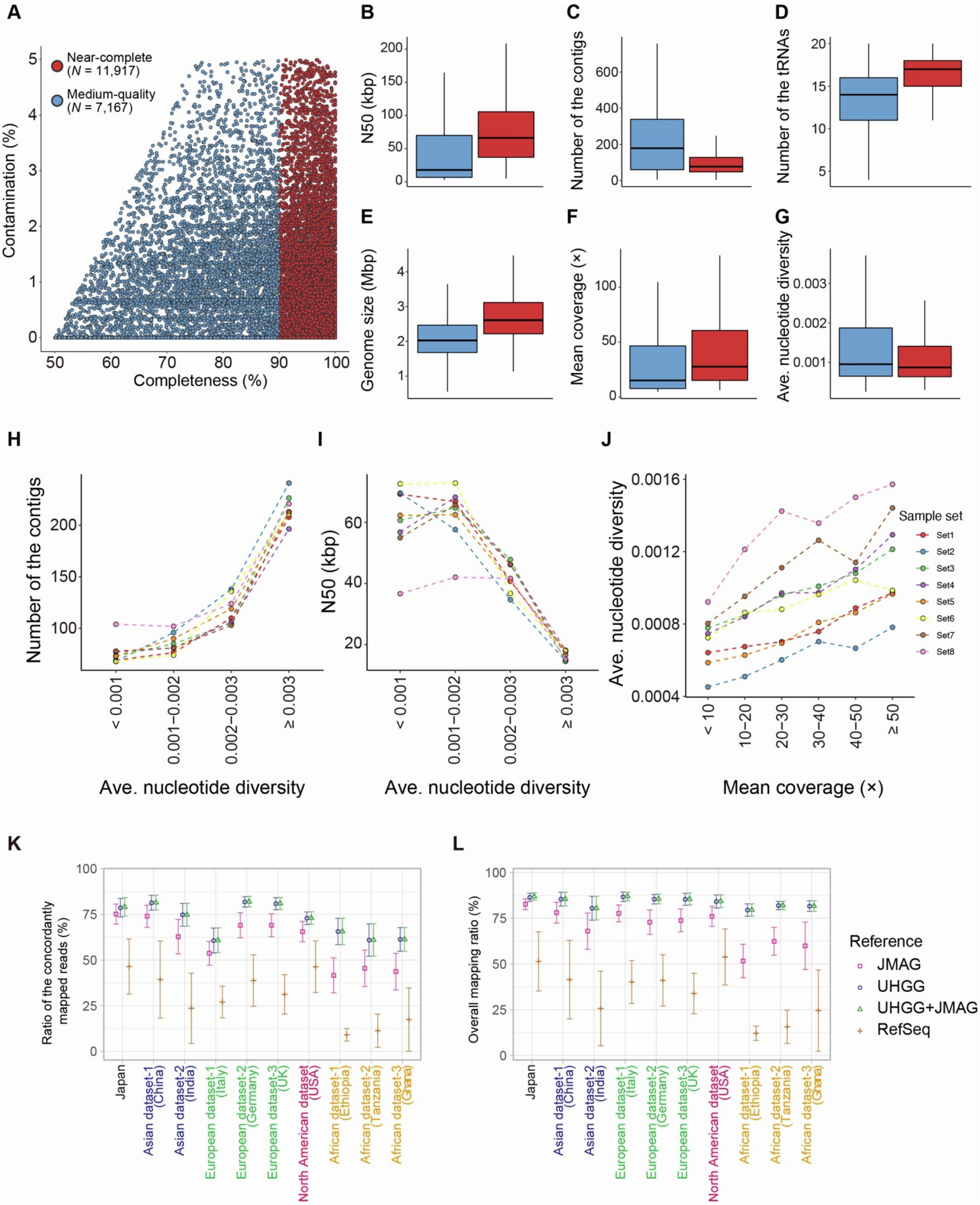

**Figure S2. Metrics related to the quality of the MAGs, related to Figure 1**

**A,** A scatter plot for the completeness and contamination of the 19,084 MAGs recovered from the Japanese gut metagenome. The colors of the dots represent the quality of the MAGs (red: near-complete, blue: medium-quality). **B-G,** boxplots for the N50 (B), number of the contigs (C), number of the non-redundant tRNAs (D), genome size (E), mean coverage (F), and average nucleotide diversity (G) of the MAGs stratified by the quality (red: near-complete, blue: medium-quality). The boxplot indicates the median values (center lines) and IQRs (box edges), with the whiskers extending to the most extreme points within the range between (lower quantile − [1.5 × IQR]) and (upper quantile + [1.5 × IQR]). **H-J,** The relationship between the average nucleotide identity and the number of the contigs (H), the average nucleotide identity and N50 (I), and the mean coverage and average nucleotide identity (J) per dataset. The colors of the dots and dashed lines indicate the sample sets. **K,L,** The dots represent the mean of the ratio of the concordantly mapped reads (K) or overall mapping ratio (L) for the metagenome shotgun sequencing dataset from several populations. The shapes and colors of the dots represent the reference genome databases. The error bars represent the standard errors. Ave., average; IQR, interquartile range; JMAG, Japanese Metagenome Assembled Genomes; MAG, metagenome-assembled genome; tRNA, transfer RNA; UHGG, Unified Human Gastrointestinal Genome.

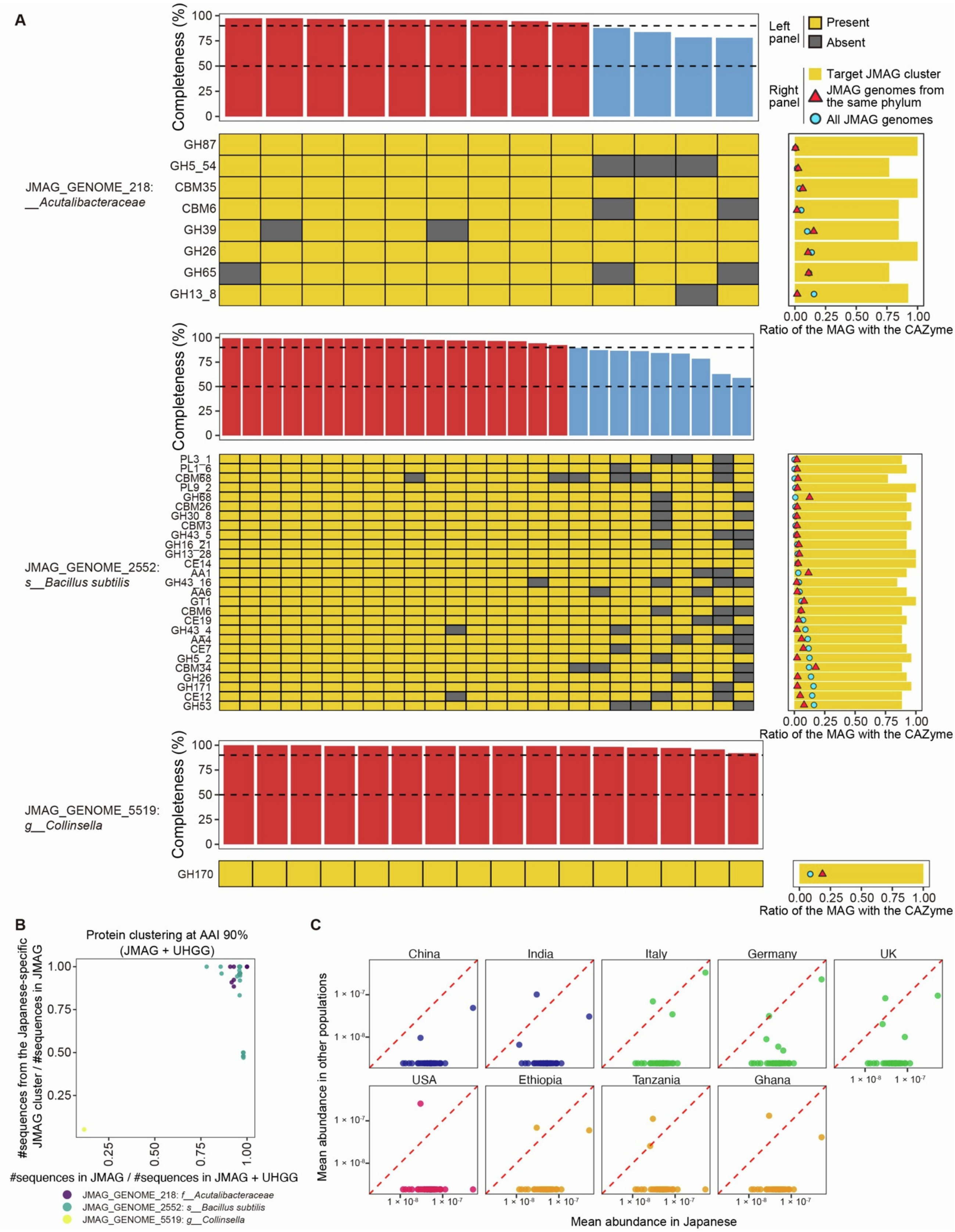

**Figure S3. CAZyme profiles of the Japanese-specific species-level clusters, related to Figure 1**

**A,** Tile plots represent the presence (yellow) or absence (grey) of the CAZymes for the MAGs belonging to the Japanese-specific species-level clusters. The barplots represent the completeness of the MAGs (top) and the within-cluster ratio of the MAGs which had the CAZymes (right). The dashed lines in the barplots represent 90% and 50% completeness (top). The red and blue dots in the barplots represent the within-phylum or within-JMAG ratio of the MAGs which had the CAZymes (right). Only the CAZymes which satisfy (i) [within-cluster ratio of the MAGs which have the CAZymes] > 0.75, (ii) [within-cluster ratio of the MAGs which have the CAZymes] > 5 × [within-phylum ratio of the MAGs which have the CAZymes], and (iii) [within-cluster ratio of the MAGs which have the CAZymes] > 5 × [within-JMAG ratio of the MAGs which have the CAZymes] are depicted. Note that three Japanese-specific species-level clusters which are not described do not have CAZymes that satisfy the above criteria. **B,** A scatter plot for the protein clusters made from the JMAG and UHGP (at 90% AAI) which include CAZymes described in (A). The x-axis indicates the (number of the protein sequences from each Japanese-specific MAG cluster) / (number of the protein sequences in the JMAG) indicating the uniqueness of the CAZymes of the Japanese-specific species-level clusters among the JMAG genomes. The y-axis indicates the (number of the protein sequences in the JMAG) / (number of the protein sequences in the JMAG and UHGP) indicating the Japanese-specificity of the extracted CAZymes. **C,** Scatter plots indicating the abundances of the CAZymes described in (A). The x-axis indicates the mean abundances in the Japanese, and the y-axis indicates the mean abundances in the non-Japanese. CAZyme, Carbohydrate Active Enzyme; JMAG, Japanese Metagenome Assembled Genomes; MAG, metagenome-assembled genome; UHGP, Unified Human Gastrointestinal Protein.

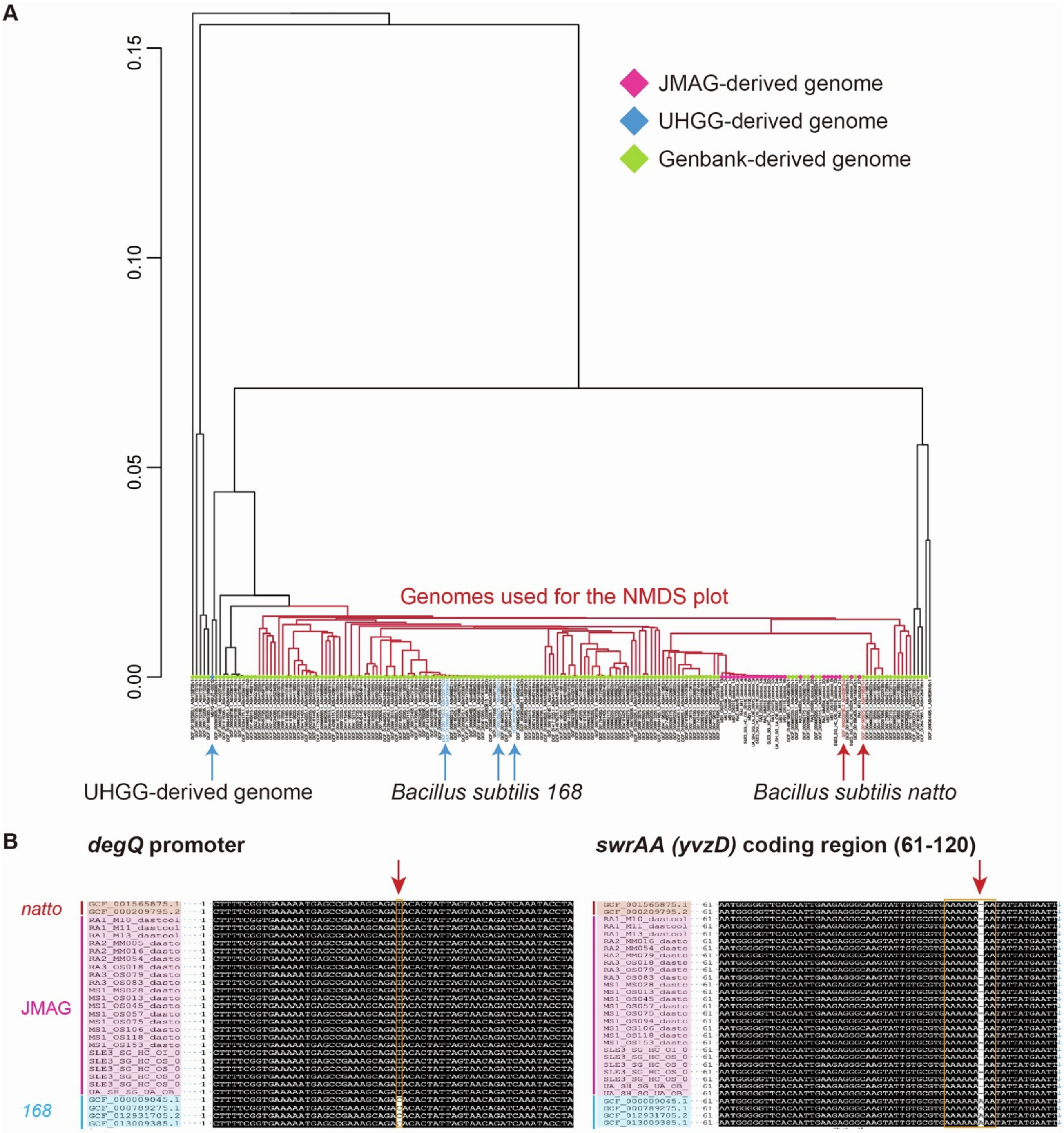

**Figure S4. Comparison of the *Bacillus subtilis* genomes, related to Figure 1**

**A,** A dendrogram represents the result of hierarchical clustering of the *Bacillus subtilis* genomes based on the average nucleotide identity. The colors of the nodes indicate the derivation of the genomes. A UHGG-derived genome, *Bacillus subtilis 168* genomes, *Bacillus subtilis natto* genomes are annotated with arrows. Ten clusters are defined and a cluster includes all the JMAG-derived genomes that are subsequently used for dimension-reduction analysis (**Figure 1E;** highlighted by red). **B,** Comparisons of the sequences of the JMAG-derived *Bacillus subtilis* genomes, *Bacillus subtilis natto* genomes, and *Bacillus subtilis 168* genomes at the *degQ* promoter (left) and *swrAA* (also called *yvzD*) coding region (right). Polymorphic sites are enclosed with yellow rectangles. JMAG, Japanese Metagenome Assembled Genomes; NMDS, Non-metric Multidimensional Scaling; UHGG, Unified Human Gastrointestinal Genome.

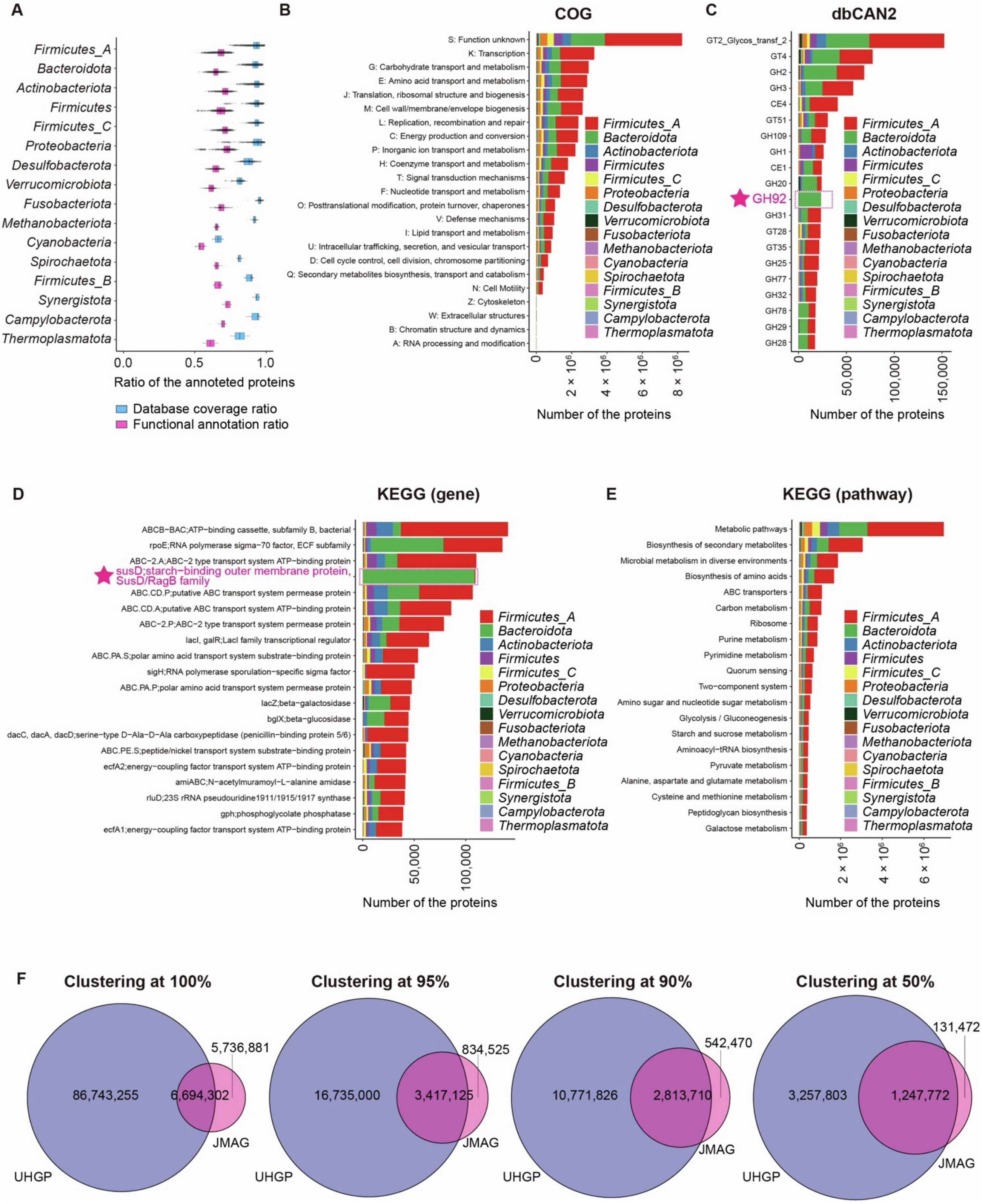

**Figure S5. Annotation of the protein sequences in the JMAG, related to Figure 2**

**A,** A boxplot represents the ratio of the proteins in the JMAG which have any eggNOG-mapper hits (database coverage ratio, cyan) and functional COG annotation (functional annotation ratio, magenta) per bacterial phylum. The boxplot indicates the median values (center lines) and IQRs (box edges), with the whiskers extending to the most extreme points within the range between (lower quantile − [1.5 × IQR]) and (upper quantile + [1.5 × IQR]). **B,** COG annotations of the JMAG proteins. The colors represent the taxonomic annotation of the MAGs linked to the proteins. **C-E,** Top 20 frequent dbCAN2 (C), KEGG gene (D), and KEGG pathway (E) annotations of the JMAG proteins. The colors represent the taxonomic annotation of the MAGs linked to the proteins. **F,** Venn diagrams represent the results of the clustering of the predicted proteins in the JMAG and UHGP at 100%, 95%, 90%, and 50% identity of amino-acid sequences. COG, Cluster of Orthologous Groups; JMAG, Japanese Metagenome Assembled Genomes; KEGG, Kyoto Encyclopedia of Genes and Genomes; MAG, metagenome-assembled genome; UHGP, Unified Human Gastrointestinal Protein.

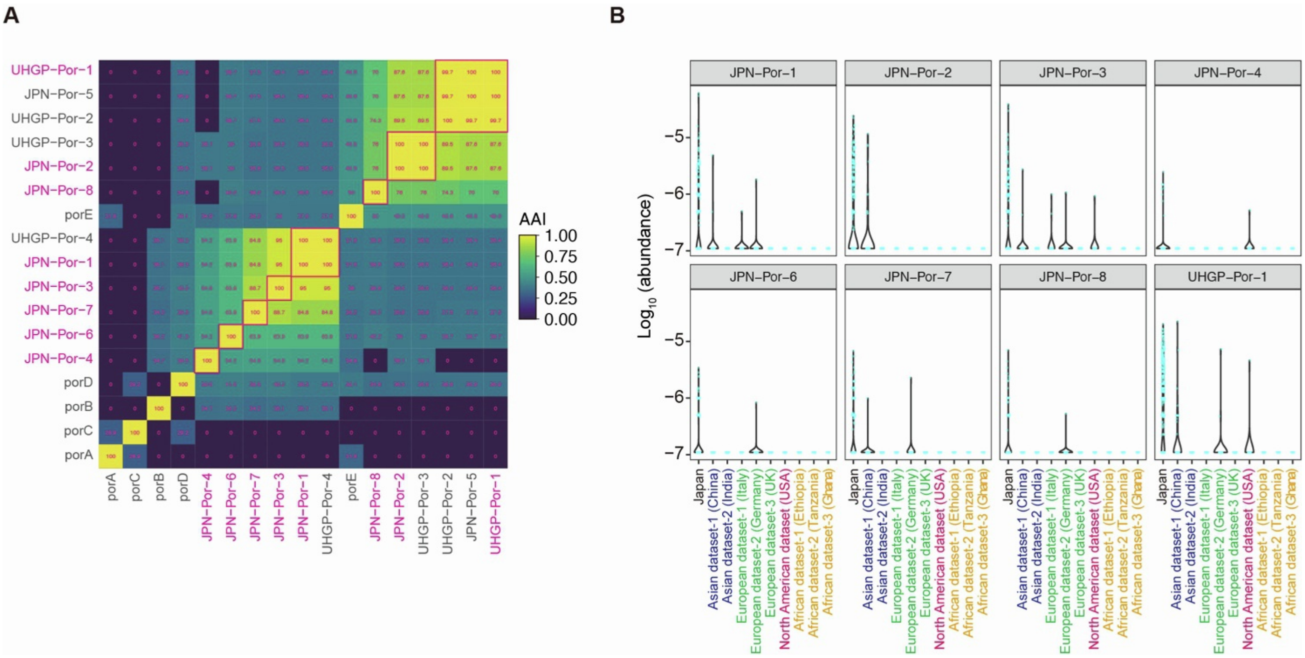

**Figure S6. Evaluation of the β-porphyranase sequences in the JMAG and UHGP, related to Figure 2**

**A,** A heatmap indicating the AAI between the β-porphyranase sequences in the JMAG and UHGP. β-porphyranase sequences available in NCBI (PorA, PorB, PorC, PorD, and PorE) are also indicated. Groups of the highly similar JMAG- and UHGG-derived sequences (AAI > 99%) are enclosed with the magenta square. The non-redundant set of the β-porphyranases is indicated with magenta. **B,** Violin plots indicate the abundances of the non-redundant β-porphyranase sequences indicated in (A). AAI, amino acid identity; JMAG, Japanese Metagenome Assembled Genomes; NCBI, National Center for Biotechnology Information; UHGP, Unified Human Gastrointestinal Protein.

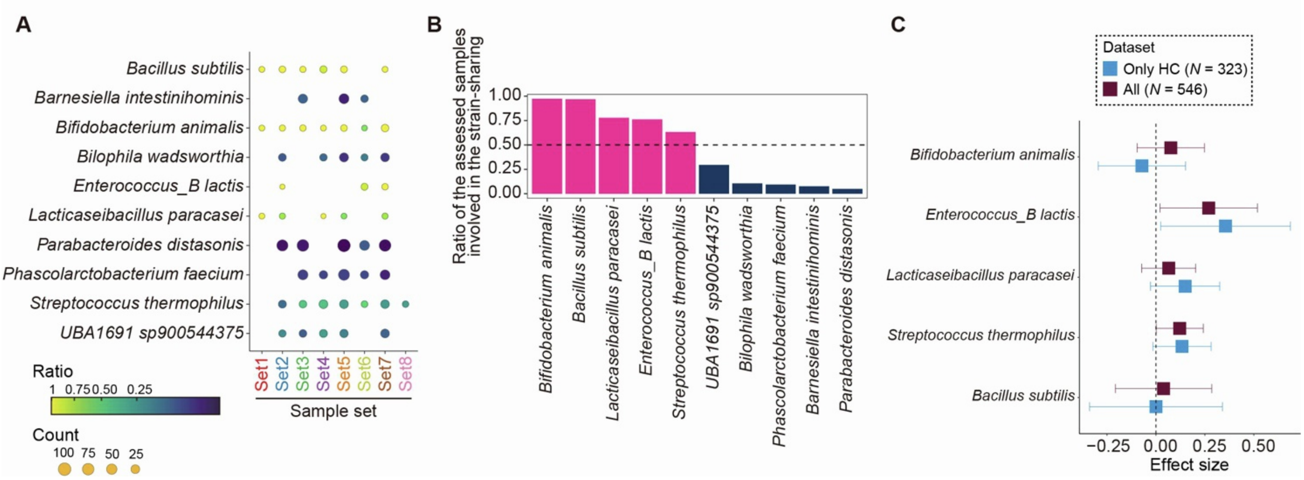

**Figure S7. Sharing of the bacterial strains among the 787 Japanese people, related to Figure 3**

**A,** A dot plot represents the sharing of bacterial strains per dataset. The colors of the dots represent the ratio of the individuals involved in the strain-sharing (popANI ≥ 99.999%). The sizes of the dots represent the number of individuals for which the bacterial species are detected by inStrain. Only the bacterial species for which the sharing of the strains are detected in at least three datasets are indicated. **B,** The ratio of the assessed samples involved in the strain-sharing is indicated as a barplot. The dashed horizontal line indicates (number of the samples involved in the strain-sharing) / (number of the samples used for the analysis of the target species) = 0.5. **C,** A forest plot for the sub-analyses of the association between the rs671 and food-associated bacteria. The effect sizes of the analyses with or without disease samples are indicated. The boxes indicate the point estimates, and the error bars indicate the 95% confidence interval. HC, healthy control.

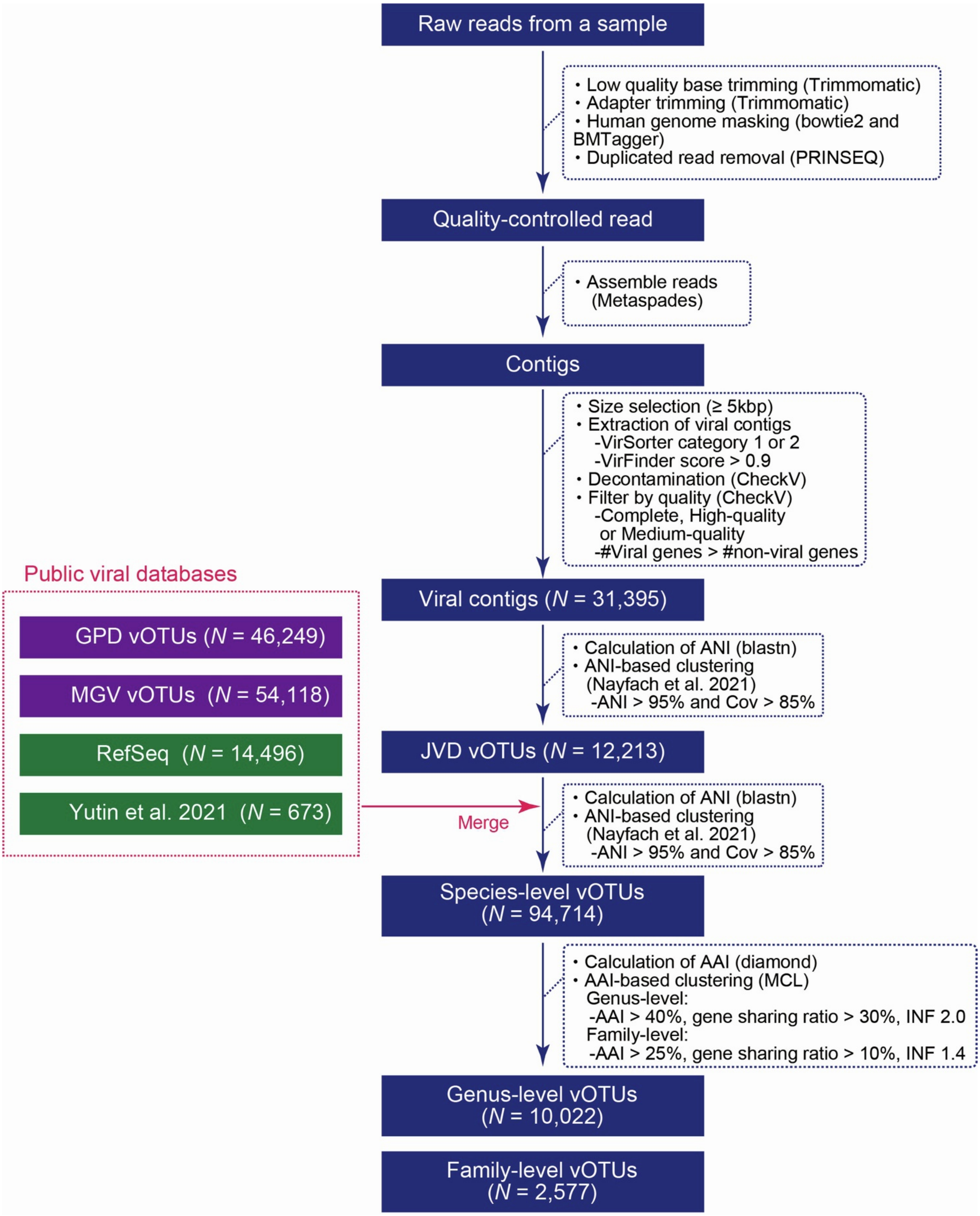
Raw reads from a sample
• Low quality base trimming (Trimmomatic)
• Adapter trimming (Trimmomatic)
• Human genome masking (bowtie2 and BMTagger)
• Duplicated read removal (PRINSEQ)
Quality-controlled read
• Assemble reads (Metaspades)
Contigs
• Size selection (≥ 5kbp)
• Extraction of viral contigs
-VirSorter category 1 or 2
-VirFinder score > 0.9
• Decontamination (CheckV)
• Filter by quality (CheckV)
-Complete, High-quality or Medium-quality
-#Viral genes > #non-viral genes
Public viral databases
GPD vOTUs (N = 46,249)
MGV vOTUs (N = 54,118)
RefSeq (N = 14,496)
Yutin et al. 2021 (N = 673)
Viral contigs (N = 31,395)
• Calculation of ANI (blastn)
• ANI-based clustering (Nayfach et al. 2021)
-ANI > 95% and Cov > 85%
JVD vOTUs (N = 12,213)
Merge
• Calculation of ANI (blastn)
• ANI-based clustering (Nayfach et al. 2021)
-ANI > 95% and Cov > 85%
Species-level vOTUs (N = 94,714)
• Calculation of AAI (diamond)
• AAI-based clustering (MCL)
Genus-level:
-AAI > 40%, gene sharing ratio > 30%, INF 2.0
Family-level:
-AAI > 25%, gene sharing ratio > 10%, INF 1.4
Genus-level vOTUs (N = 10,022)
Family-level vOTUs (N = 2,577)

**Figure S8. A pipeline for detecting viral genomes, related to Figure 4**

The metagenome shotgun sequencing reads of the Japanese gut microbiome were processed following this pipeline, resulting in 31,395 genomes comprising 12,213 JVD vOTUs. The JVD vOTUs were merged with other databases, namely GPD, MGV, RefSeq, and Yutin et al. 2021, and clustered into 94,714 species-level vOTUs. The databases indicated in purple are the viral genome databases recovered from the human gut metagenome, and those indicated in green are the viral genome databases with curated taxonomy. The 94,714 species-level vOTUs were further clustered into 10,022 genus- and 2,577 family-level vOTUs. ANI, average nucleotide identity; AAI, average amino acid identity; Cov, coverage; GPD, Gut Phage Database; INF, inflation factor; JVD, Japanese Virus Database; MCL, markov clustering; MGV, Metagenomic Gut Virus; vOTU, viral operative taxonomic unit.

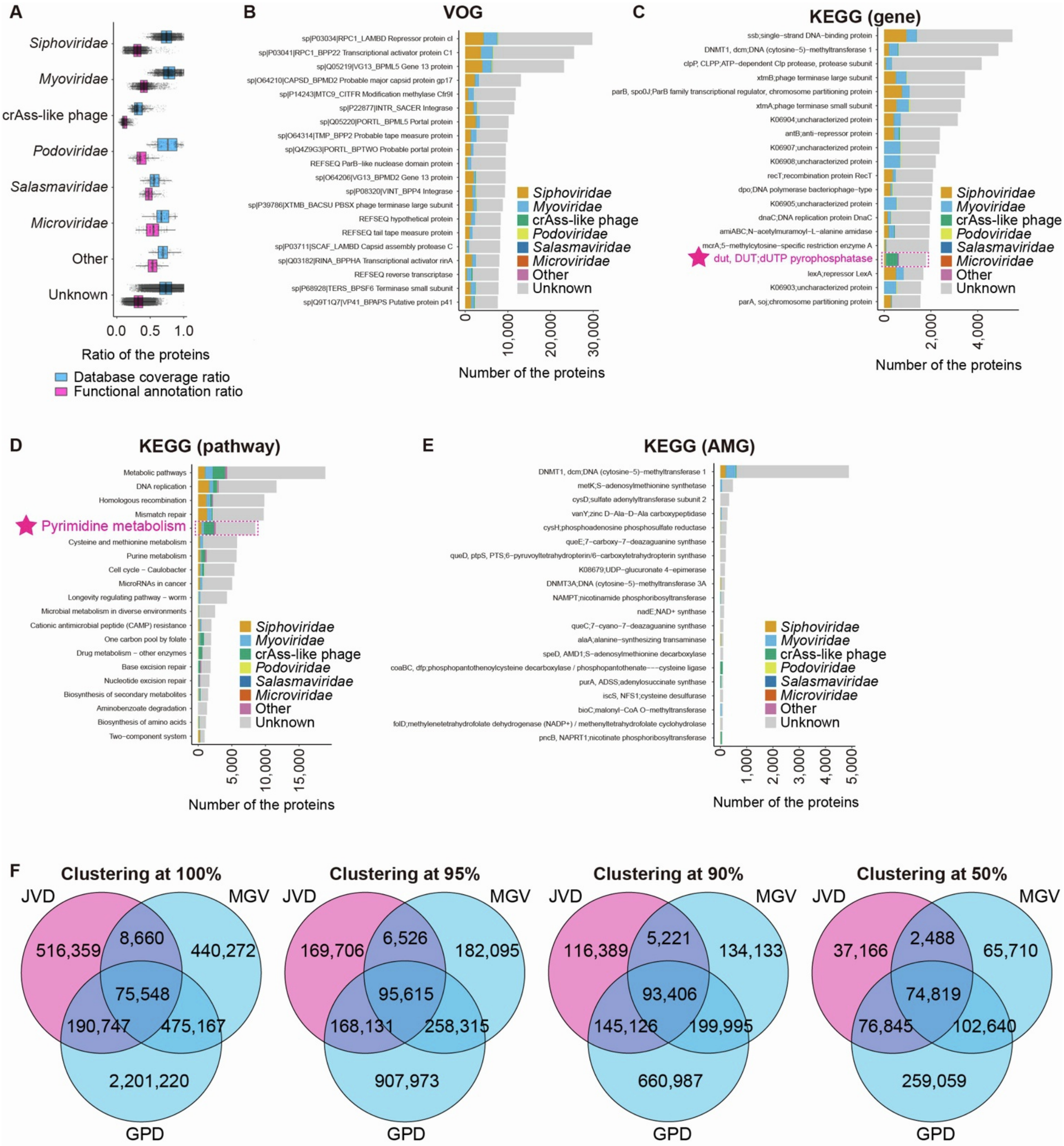

**Figure S9. Annotation of the protein sequences in the JVD, related to Figure 4**

**A,** A boxplot represents the ratio of the annotated proteins in the JVD which have any eggNOG-mapper or VOG hits (database coverage ratio, cyan) and functional COG or VOG annotation (functional annotation ratio, magenta) per viral family. The boxplot indicates the median values (center lines) and IQRs (box edges), with the whiskers extending to the most extreme points within the range between (lower quantile − [1.5 × IQR]) and (upper quantile

+ [1.5 × IQR]). **B-E,** Top 20 frequent VOG (B), KEGG gene (C), KEGG pathway (D), and KEGG AMG (E) annotations of the JVD proteins. The colors represent the taxonomic annotation of the viral genomes linked to the proteins. **F,** Venn diagrams represent the results of the clustering of the predicted proteins in the JVD, GPD, and MGV at 100%, 95%, 90%, and 50% identity of amino-acid sequences. AMG, auxiliary metabolic genes; GPD, Gut Phage Database; IQR, interquartile range; JVD, Japanese Virus Database; KEGG, Kyoto Encyclopedia of Genes and Genomes; MGV, Metagenomic Gut Virus; VOG, Virus orthologous groups.

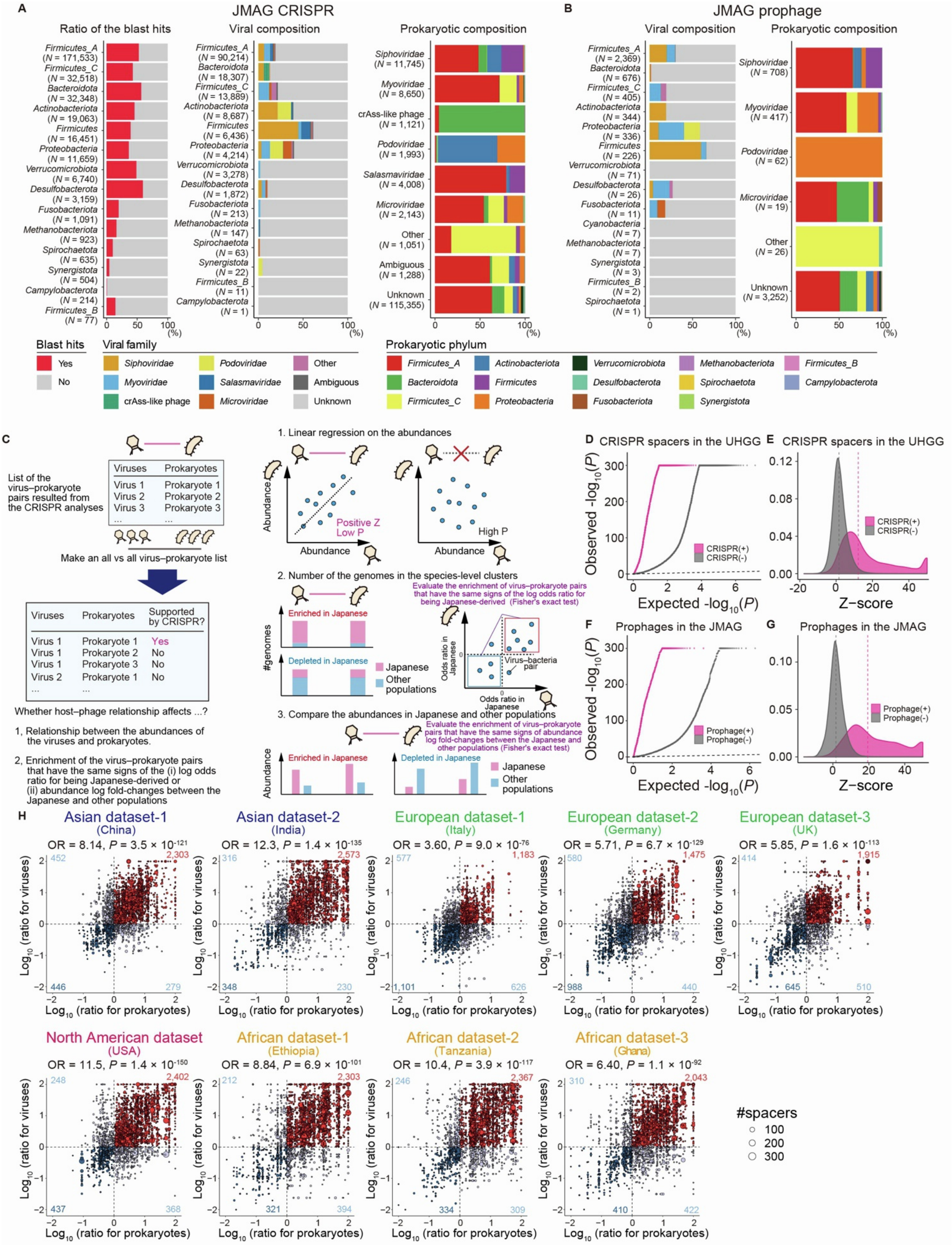

**Figure S10. The taxonomic and populational analysis of the virus–prokaryote interaction, related to Figure 6**

**A,** Barplots represent the ratio of the blast hits (left), composition of the family-level taxonomy of targeted viruses (middle), and composition of the phylum-level taxonomy of the linked MAGs (right) of the JMAG CRISPR spacers. The colors of the bar represent whether the spacer sequences have blast hits or not (left) and the taxonomic annotation of the microbes (middle, right), respectively. **B,** Barplots represent the composition of the family-level taxonomy of the prophages (left) and composition of the phylum-level taxonomy of the linked MAGs of the prophages (right) in the JMAG. The colors of the bar represent the taxonomic annotation of the microbes. **C,** A graphical abstract for the CRISPR analysis. Viruses and prokaryotes involved in the virus–prokaryote interaction supported by the CRISPR are listed up (left). All vs all virus–prokaryote pairs made from the list are classified by whether they are supported by the CRISPR or not. Then, linear regression analysis between the viral and prokaryotic abundances is performed (right-top), and the results are evaluated with the stratification based on the supports from the CRISPR. Utilizing all the virus–prokaryote pairs supported by the CRISPR, we evaluate the enrichment of the virus–prokaryote pairs which have the same trends of the interpopulational differences based on the number of the genomes (right-middle) and abundances (right-bottom). **D,** A quantile–quantile plot of the p-values from the virus–prokaryote association analysis stratified by whether the virus–prokaryote pairs are supported by the CRISPR spacers in the UHGG (magenta) or not (gray). The x-axis indicates $-\log_{10}(P)$ expected from a uniform distribution. The y-axis indicates the observed $-\log_{10}(P)$. The diagonal dashed line represents $y = x$, which corresponds to the null hypothesis. **E,** A density plot of the Z-score from the virus–prokaryote association analysis stratified by whether the virus–prokaryote pairs are supported by CRISPR spacers in the UHGG (magenta) or not (gray). The upper limit of the Z-score is set at 50. The diagonal dashed line represents $y = x$, which corresponds to the null hypothesis. The vertical dashed lines indicate the mean of the Z-score for each group of the virus–prokaryote pair. **F,** A quantile–quantile plot of the p-values from the virus–prokaryote association analysis stratified by whether the virus–prokaryote pairs are supported by the proviral sequences in the JMAG (magenta) or not (gray). The x-axis indicates $-\log_{10}(P)$ expected from a uniform distribution. The y-axis indicates the observed $-\log_{10}(P)$. The diagonal dashed line represents $y = x$, which corresponds to the null hypothesis. **G,** A density plot of the Z-score from the virus–prokaryote association analysis stratified by whether the virus–prokaryote pairs are supported by the proviral sequences in the JMAG (magenta) or not (gray). The upper limit of the Z-score is set at 50. The diagonal dashed line represents $y = x$, which corresponds to the null hypothesis. The vertical dashed lines indicate the mean of the Z-score for each group of the virus–prokaryote pair. **H,** Scatter plots of the fold-changes between the RPKM in Japanese and other populations for viruses (y-axis) and prokaryotes (x-axis). Fold-changes were calculated for the virus–prokaryote pairs supported by the CRISPR spacers in the JMAG (**STAR Methods**). The size of the dots represents the number of spacers supporting the virus–prokaryote pairs. The horizontal and vertical dashed lines represent log fold-change = 0 for virus and prokaryote, respectively. JMAG, Japanese Metagenome Assembled Genomes; JVD, Japanese Virus Database; MAG, metagenome-assembled genome; MGV, Metagenomic Gut Virus; OR, odds ratio; RPKM, Reads Per Kilobase of exon per Million mapped reads; UHGG, Unified Human Gastrointestinal Genome; vOTU, viral operative taxonomic unit.

## Supplemental data

### Data S1. Metrics represent the quality of the MAGs, related to Figure 1.

Most of the MAGs with >90% genome completeness and <5% contamination did not satisfy the MIMAG's criteria for high-quality genomes due to the difficulty in assembling the ribosomal RNA (rRNA) regions of the prokaryotic genomes (recovered in 3,544 MAGs for 5S rRNA, 200 MAGs for 16S rRNA, and 347 MAGs for 23S rRNA). Therefore, we refer to the 11,917 MAGs with >90% genome completeness and <5% contamination as near-complete following the UHGG[8]. The near-complete MAGs tended to have higher contiguity, more transfer RNA (tRNA), and longer genome size than the medium-quality MAGs (**Figure S2B-E**). The lower coverage and higher average nucleotide diversity of the medium-quality MAGs than the near-complete MAGs suggested that prokaryotes with the low coverage and high strain-level diversity were difficult to assemble as previously suggested[5] (**Figure S2F,G**). The difficulty for assembling prokaryotes with the high strain-level diversity was also reflected in the negative associations between the average nucleotide diversity and the contiguity of the MAGs (**Figure S2H,I**) and the requirement of the high read coverage for assembling the genomes with high strain-level diversity (**Figure S2J**).

### Data S2. Characterization of the *Bacillus subtilis* genomes in the JMAG, related to Figure 1

We calculated the pairwise ANI for the 189 *Bacillus subtilis* genomes derived from the JMAG, UHGG, and Genbank, and performed hierarchical clustering (**Figure S4A**). The *Bacillus subtilis* genomes derived from the JMAG were clustered close to the *Bacillus subtilis natto* and distantly from the UHGG-derived *Bacillus subtilis* and a laboratory strain *Bacillus subtilis 168*. In a non-metric multidimensional scaling (NMDS) analysis, the JMAG-derived *Bacillus subtilis* genomes were in proximity to the *Bacillus subtilis natto* genomes (**Figure 2D**). In addition, we evaluated the genetic polymorphism of the *degQ* promoter and *swrAA* (also known as *yvzD*) coding regions which were important for the production of γ-poly-DL-glutamic acid, a source of the unique and sticky texture of natto[87]. All of the JMAG-derived *Bacillus subtilis* had the same genotype as the *Bacillus subtilis natto* genomes (**Figure S4B**).