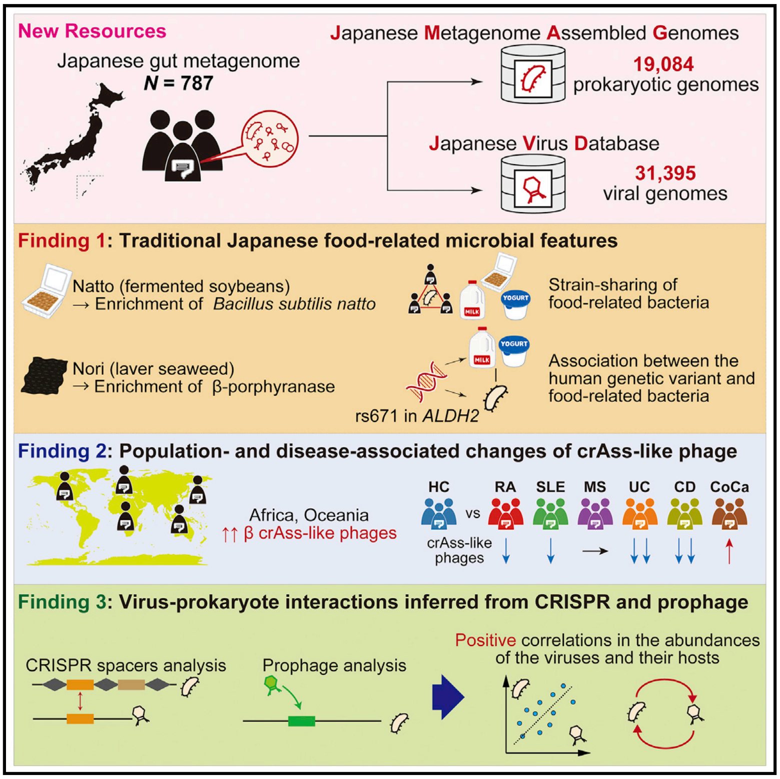
Cell Genomics

Resource

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

Graphical abstract


New Resources
Japanese gut metagenome
N = 787
Japanese Metagenome Assembled Genomes
19,084
prokaryotic genomes
Japanese Virus Database
31,395
viral genomes
Finding 1: Traditional Japanese food-related microbial features
Natto (fermented soybeans)
→ Enrichment of Bacillus subtilis natto
Strain-sharing of
food-related bacteria
MILK
YOGURT
Nori (laver seaweed)
→ Enrichment of β-porphyranase
Association between the
human genetic variant and
food-related bacteria
rs671 in ALDH2
Finding 2: Population- and disease-associated changes of crAss-like phage
Africa, Oceania
↑↑ β crAss-like phages
HC
vs
RA
SLE
MS
UC
CD
CoCa
crAss-like
phages
Finding 3: Virus-prokaryote interactions inferred from CRISPR and prophage
CRISPR spacers analysis
Prophage analysis
Positive correlations in the abundances
of the viruses and their hosts

Authors

Yoshihiko Tomofuji, Toshihiro Kishikawa, Yuichi Maeda, ..., Kiyoshi Takeda, Atsushi Kumanogoh, Yukinori Okada

Correspondence

ytomofuji@sg.med.osaka-u.ac.jp (Y.T.),
yokada@sg.med.osaka-u.ac.jp (Y.O.)

In brief

Tomofuji et al. reconstructed 19,084 prokaryotic and 31,395 viral genomes from Japanese gut metagenome shotgun sequencing data. They revealed the association between the gut microbiome and diet, populations, and diseases. Their genome catalog, JMAG and JVD, contributes to expanding the diversity of the microbial genome of a previously underrepresented population.

Highlights

- Assembly of 19,084 prokaryotic and 31,395 viral genomes from Japanese gut metagenome
- Traditional Japanese food-related features were observed in Japanese microbial genome
- crAss-like phages were associated with populations and diseases
- Abundances of bacteriophages and their hosts tended to be positively correlated

Tomofuji et al., 2022, Cell Genomics 2, 100219
December 14, 2022 © 2022 The Author(s).
https://doi.org/10.1016/j.xgen.2022.100219

Cell Genomics

CellPress
OPEN ACCESS

Resource

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

Yoshihiko Tomofuji,[1,2,*] Toshihiro Kishikawa,[1,3,4] Yuichi Maeda,[2,5,6] Kotaro Ogawa,[7] Yuriko Otake-Kasamoto,[8] Shuhei Kawabata,[9] Takuro Nii,[5,6] Tatsusada Okuno,[7] Eri Oguro-Igashira,[5,6] Makoto Kinoshita,[7] Masatoshi Takagaki,[9] Naoki Oyama,[10] Kenichi Todo,[7] Kenichi Yamamoto,[1,11,12] Kyuto Sonehara,[1,2] Mayu Yagita,[5,6] Akiko Hosokawa,[13] Daisuke Motooka,[2,14] Yuki Matsumoto,[14] Hidetoshi Matsuoka,[15] Maiko Yoshimura,[15] Shiro Ohshima,[15] Shinichiro Shinzaki,[8] Shota Nakamura,[2,14] Hideki Iijima,[8] Hidenori Inohara,[3] Haruhiko Kishima,[9] Tetsuo Takehara,[8] Hideki Mochizuki,[7] Kiyoshi Takeda,[6,16] Atsushi Kumanogoh,[2,5,17] and Yukinori Okada[1,2,12,18,19,20,21,*]

[1]Department of Statistical Genetics, Osaka University Graduate School of Medicine, 2-2 Yamadaoka, Suita 565-0871, Japan
[2]Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Suita 565-0871, Japan
[3]Department of Otorhinolaryngology-Head and Neck Surgery, Osaka University Graduate School of Medicine, Suita 565-0871, Japan
[4]Department of Head and Neck Surgery, Aichi Cancer Center Hospital, Nagoya 464-8681, Japan
[5]Department of Respiratory Medicine and Clinical Immunology, Osaka University Graduate School of Medicine, Suita 565-0871, Japan
[6]Laboratory of Immune Regulation, Department of Microbiology and Immunology, Osaka University Graduate School of Medicine, Suita 565-0871, Japan
[7]Department of Neurology, Osaka University Graduate School of Medicine, Suita 565-0871, Japan
[8]Department of Gastroenterology and Hepatology, Osaka University Graduate School of Medicine, Suita 565-0871, Japan
[9]Department of Neurosurgery, Osaka University Graduate School of Medicine, Suita 565-0871, Japan
[10]Department of Stroke Medicine, Kawasaki Medical School, Kurashiki 701-0192, Japan
[11]Department of Pediatrics, Osaka University Graduate School of Medicine, Suita 565-0871, Japan
[12]Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita 565-0871, Japan
[13]Department of Neurology, Suita Municipal Hospital, Suita 564-8567, Japan
[14]Department of Infection Metagenomics, Research Institute for Microbial Diseases, Osaka University, Suita 565-0871, Japan
[15]Department of Rheumatology and Allergology, NHO Osaka Minami Medical Center, Kawachinagano 586-8521, Japan
[16]WPI Immunology Frontier Research Center, Osaka University, Suita 565-0871, Japan
[17]Department of Immunopathology, Immunology Frontier Research Center, Osaka University, Suita 565-0871, Japan
[18]Laboratory for Systems Genetics, RIKEN Center for Integrative Medical Sciences, Tsurumi 230-0045, Japan
[19]Center for Infectious Disease Education and Research, Osaka University, Suita 565-0871, Japan
[20]Department of Genome Informatics, Graduate School of Medicine, the University of Tokyo, Tokyo, Japan
[21]Lead contact
*Correspondence: ytomofuji@sg.med.osaka-u.ac.jp (Y.T.), yokada@sg.med.osaka-u.ac.jp (Y.O.)
https://doi.org/10.1016/j.xgen.2022.100219

SUMMARY

We reconstructed 19,084 prokaryotic and 31,395 viral genomes from 787 Japanese gut metagenomes as Japanese metagenome-assembled genomes (JMAG) and Japanese Virus Database (JVD), which are large microbial genome datasets for a single population. Population-specific enrichment of the *Bacillus subtilis* and β-porphyranase among the JMAG could derive from the Japanese traditional food natto (fermented soybeans) and nori (laver), respectively. Dairy-related *Enterococcus_B lactis* and *Streptococcus thermophilus* were nominally associated with the East Asian-specific missense variant rs671:G>A in *ALDH2*, which was associated with dairy consumption. Of the species-level viral genome clusters in the JVD, 62.9% were novel. The β crAss-like phage composition was low among the Japanese but relatively high among African and Oceanian peoples. Evaluations of the association between crAss-like phages and diseases showed significant disease-specific associations. Our large catalog of virus-host pairs identified the positive correlation between the abundance of the viruses and their hosts.

INTRODUCTION

The human microbiome is a complex microbial community inhabiting the human body. The largest community of the human microbiota resides within the gut and they interact with the host's body via the immune system and metabolic reactions.[1] Thus, understanding the human gut microbiome is important not only in terms of microbiology but also for medicine.

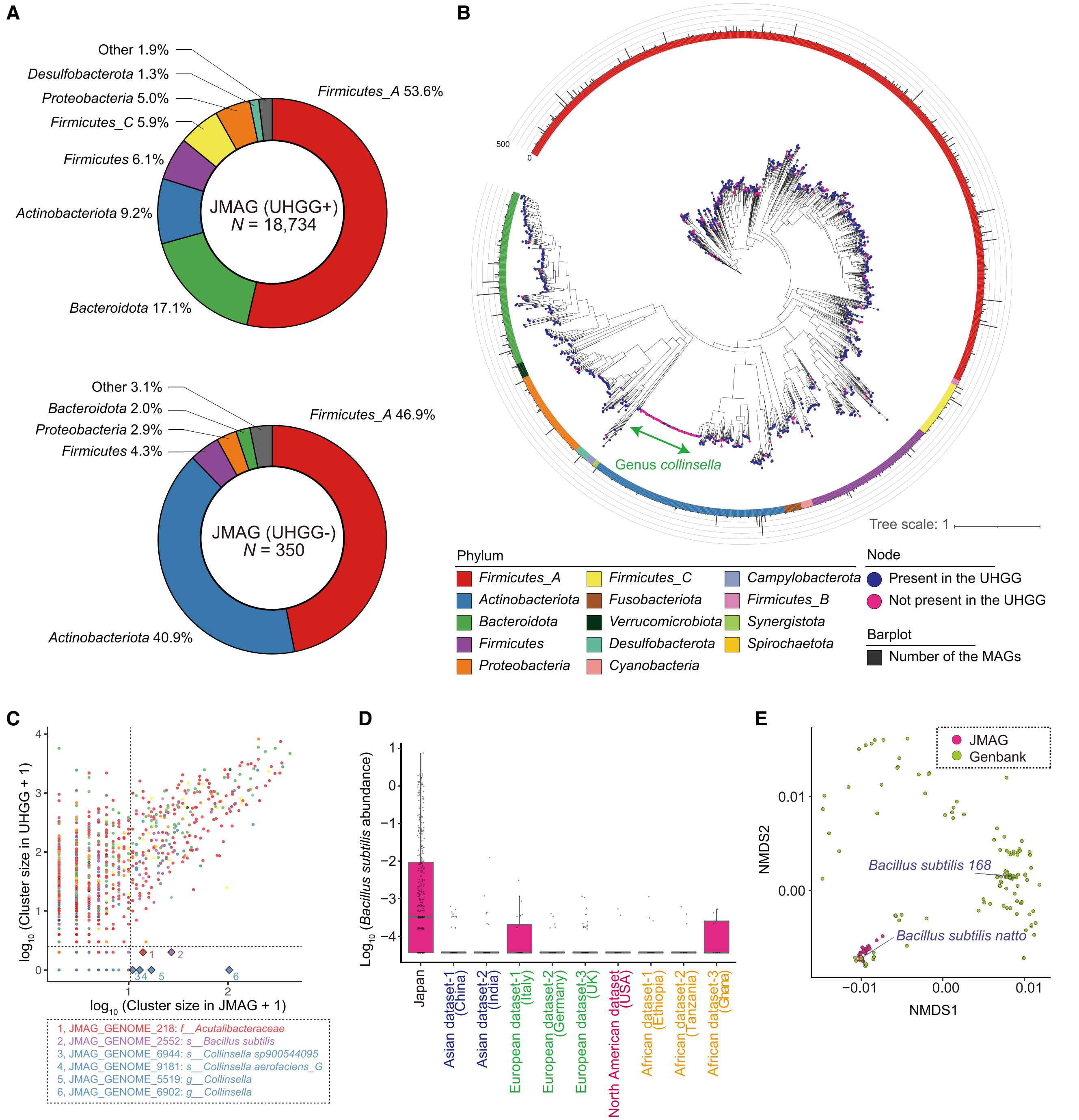


Figure 1. Phylogenetic analysis of the JMAG genomes

(A) A pie chart illustrating the phylum-level phylogenetic composition of the JMAG, which had corresponding species-level clusters in the UHGG (top) or not (bottom). The phyla comprising <1% of each genome set are collapsed into “Other.”

(B) A maximum-likelihood phylogenetic tree reconstructed from 1,267 species-level representative bacterial MAGs. The color of the nodes represents whether the species-level clusters have corresponding clusters in the UHGG (navy) or not (magenta). An outer ring represents phylum-level taxonomic annotation. A bar plot depicted in the periphery of the tree represents the number of the MAGs belonging to the same cluster of the representative genomes.

(C) A scatterplot represents the number of the JMAG genomes in the species-level clusters (x axis) and non-Japanese-derived UHGG genomes belonging to the corresponding species-level clusters (y axis). The colors of the dots represent phylum-level taxonomy. The six species-level clusters that contained ≥10 JMAG genomes and ≤1 UHGG genome are represented by rhombus.

(legend continued on next page)

In gut microbiome studies, the genomic sequences of the individual microbes are important resources that by themselves reflect the diversity and function of the gut microbiome and also can be utilized as the reference genomes for quantification with metagenome shotgun sequencing (MSS) data. Therefore, great efforts have been spent on expanding the catalog of the gut microbe genomes. In addition to culturing efforts,[2–4] genome assembly and binning from gut MSS data have greatly expanded the known diversity of the human gut microbiome.[5–7] These efforts to recover metagenome-assembled genomes (MAGs) from large-scale human MSS data enabled us to survey the previously unknown part of the gut microbiome, especially for unculturable prokaryotes. Recently, several microbial genome databases, including MAG datasets, were integrated, and a Unified Human Gastrointestinal Genome (UHGG) collection comprised of 4,644 species-level genomes, which represented >200,000 non-redundant reference genomes, was released as the currently most comprehensive atlas of the human gut prokaryotes.[8] However, current populational diversity of the prokaryotic genomes is still limited because the number of MAGs recovered from populations other than European, North American, and Chinese is relatively low. Therefore, reconstruction of the MAGs from currently underrepresented populations is warranted.

Although many of the gut microbiome studies have focused on the prokaryotes, viruses, mainly bacteriophages, are also highly abundant in the gut microbiome.[9] Bacteriophages infect bacteria and regulate the bacteriome by either lysing their hosts or altering their physiological functions. In addition to the mediating effects, gut viruses are thought to directly interact with our body via the immune system.[10,11] Various diseases, such as intestinal diseases[12,13] and metabolic diseases,[14,15] are associated with the gut virome. However, most of the human gut virome is still poorly characterized, partially because the traditional laboratory techniques, such as culturing, are typically low throughput and not applicable for some viruses. To overcome this problem, viral genomes have been recovered from the MSS data, and *de novo* assembly of the viral genomes greatly expanded the repertoire of the viral genomes and enable us to reveal a part of the gut virome.[16,17] For example, crAss-like phages, one of the major components of the human gut viromes, were first discovered in 2014 by cross-assembly of the MSS data.[18] Recently, a few studies recovered viral genomes from large-scale MSS data.[19,20] However, the diversity of the gut viral genomes is still not saturated and the current populational diversity of the viral genomes remains limited, as is the case of the prokaryotes.

The Japanese have unique dietary culture and habits, which have resulted in the unique features of the gut microbiome, such as the enrichment of the enzymes degrading seaweed-derived polysaccharides,[21] carbohydrate metabolism-related genes, and *Actinobacteria*, compared with other populations.[22] However, most of the previous studies utilized reference bacterial genomes for phylogenetic analyses. Thus, the existence of the gut microbes that were not covered by the reference dataset have not been fully evaluated. In addition, previous analysis of the gut microbial genes lacked the link between the genes and their genome of origins, which hindered us from understanding the taxonomic features of the microbial genes. Also, few studies have focused on the Japanese gut virome,[16] and there are only a small number of publicly available viral genomes recovered from the Japanese gut metagenome. Therefore, recovering MAGs and viral genomes from the Japanese gut metagenome is necessary for obtaining deep insights into the Japanese gut microbiome and complementing the microbial genome databases by increasing the populational diversity.

We recovered MAGs and viral genomes from the gut MSS data of 787 Japanese individuals.[23–27] Utilizing these reconstructed microbial genomes, we evaluated the existence of the microbial taxa and genes that were specific to the Japanese, revealed the association of the crAss-like phages with the populations and diseases, and expanded the current knowledge of the virus-prokaryote interaction. The reconstructed microbial genomes and related information are available to the scientific community.

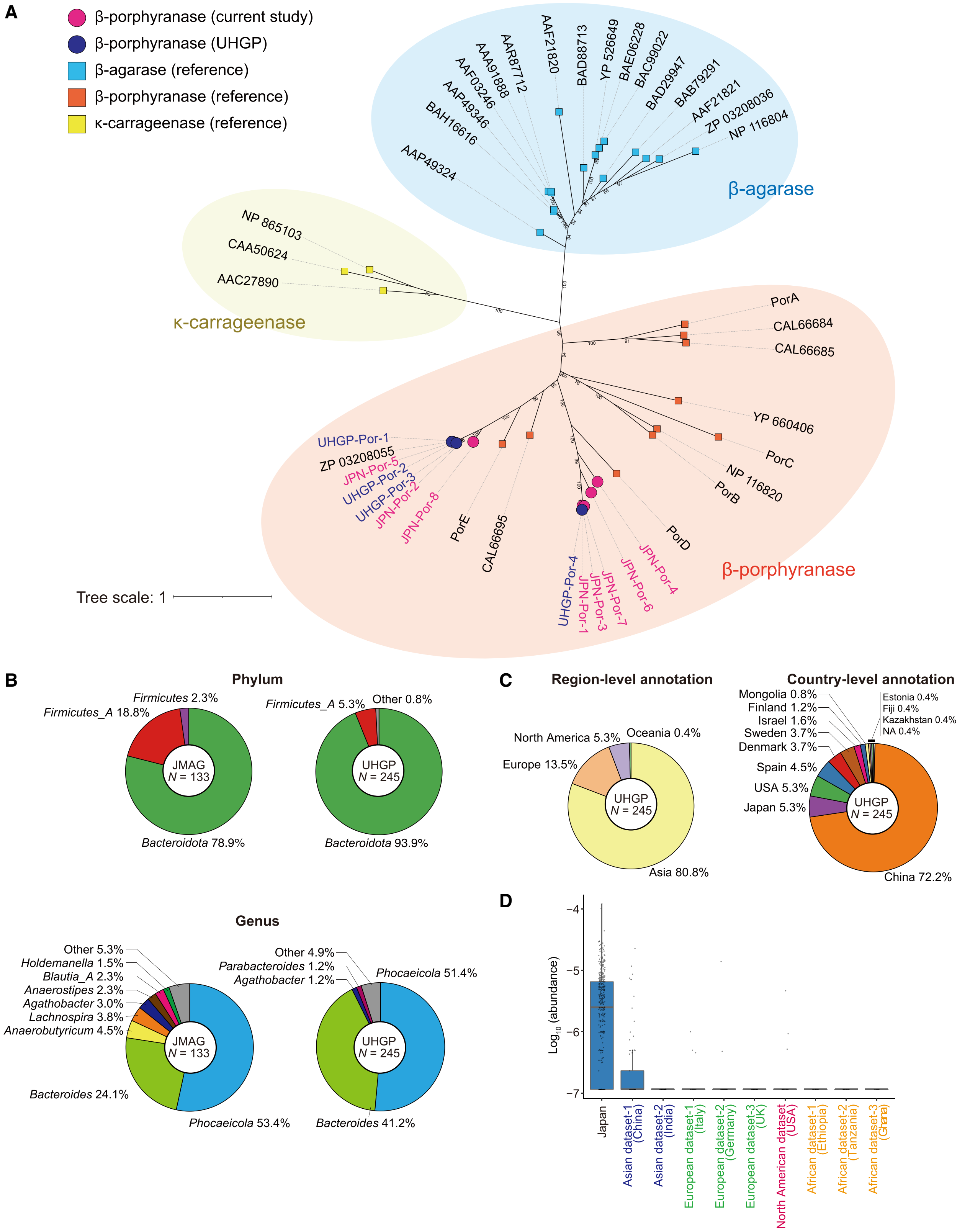
RESULTS

Reconstruction of MAGs from the Japanese MSS data

To recover MAGs from the Japanese gut, we performed a single-sample metagenomic assembly and binning on 787 Japanese gut MSS data[23–28] (Figure S1; Table S1). After the filtering based on the CheckM[29] (>50% genome completeness, <5% contamination, and an estimated quality score >50; STAR Methods), we obtained 19,084 MAGs that met or exceeded the medium quality defined by "minimum information about a metagenome-assembled genome" standard[30] ($\geq$50% genome completeness and <10% contamination; Figures S2A–S2J; Table S2; Data S1), and we call this set of the MAGs the JMAG (Japanese metagenome-assembled genomes). We refer to the 11,917 MAGs with >90% genome completeness and <5% contamination as near-complete following the UHGG.[8]

The JMAG genomes were then clustered into 1,273 species-level clusters based on the average nucleotide identity (ANI). Although some of the species-level clusters had corresponding clusters in the UHGG (1,040 clusters composed of 18,734 MAGs), others did not (233 clusters composed of 350 MAGs). We assigned taxonomic information to the JMAG genomes with GTDB-tk[31] and constructed a maximum-likelihood phylogenetic tree. Among the JMAG genomes presented in the UHGG, *Firmicutes_A*, *Bacteroidota*, and *Actinobacteriota* were frequent (Figure 1A). Among the JMAG genomes that did not present in the UHGG, the frequency of *Actinobacteriota* was higher than other MAGs, which reflected the high species-level diversity of the genus *Collinsella* (Figures 1A and 1B). To evaluate

(D) A boxplot of the *Bacillus subtilis* abundances (RPKM) in the different populations. 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]).
(E) A non-metric multidimensional scaling plot of the *Bacillus subtilis* genomes. The colors of the dots represent the derivation of the genomes. The *Bacillus subtilis natto* and *Bacillus subtilis 168* genomes in the GenBank are annotated and depicted in rhombus. IQR, interquartile range. See also Figures S1–S4, Table S2, and Data S1 and S2.



(legend on next page)

how representative the JMAG representative genomes were of Japanese gut microbial diversity, we mapped the gut MSS data against the 1,273 JMAG representative genomes. As for the Japanese gut MSS data, the mapping ratio to the 1,273 JMAG representative genomes was almost comparable with that of the 4,644 UHGG representative genomes despite the smaller number of the genomes in the JMAG than the UHGG (concordantly mapped read, 75.2% for the JMAG and 78.6% for the UHGG; overall mapped read, 82.6% for the JMAG and 86.4% for the UHGG; Figures S2K and S2L). Merging the 4,644 UHGG representative genomes and the 233 JMAG representative genomes that did not present in the UHGG only slightly improved the mapping ratio (concordantly mapped read, 79.1%; overall mapped read, 86.9%). Note that the differences in the mapping ratio between the JMAG and UHGG were larger in other populations than in Japan.

To evaluate whether the JMAG included the prokaryotic species that were underrepresented in the non-Japanese populations, we compared the number of the JMAG genomes in the species-level clusters and non-Japanese UHGG genomes belonging to the corresponding species-level clusters (Figure 1C). We found that six species-level clusters were enriched in the JMAG compared with the UHGG ($\geq$10 JMAG genomes and $\leq$1 UHGG genome). MAGs in these species-level clusters, especially the unclassified *Acutalibacteraceaem* and *Bacillus subtilis*, had several carbohydrate active enzymes (CAZymes) that were specific to these species-level clusters among the JMAG (Figures S3A and S3B), suggesting that they might have unique metabolic functions in the Japanese gut microbiome. These CAZymes were underrepresented in the Unified Human Gastrointestinal Protein (UHGP) (Figure S3B), currently the largest gut microbiome protein database, and tend to be more abundant in Japanese than other populations (Figure S3C). Therefore, the JMAG captured a part of the gut microbial features that were underrepresented in the previous studies.

Bacillus subtilis was frequently seen in the JMAG (26 MAGs), while only an isolated genome was included in the UHGG. *Bacillus subtilis* was more frequent in the Japanese than other datasets of different populations, also in the read-based quantification approach (Figure 1D). To reveal the phylogenetic characteristics of the *Bacillus subtilis* genomes in the JMAG and UHGG, we retrieved 162 *Bacillus subtilis* genomes that were available in the GenBank for comparative analyses. *Bacillus subtilis* genomes in the JMAG were closely placed to *Bacillus subtilis natto* by ANI-based non-metric multidimensional scaling analysis (Figures 1E, S4A, and S4B; Data S2). *Bacillus subtilis natto* is a key component of a Japanese traditional fermented food natto. Thus, it was suggested that *Bacillus subtilis* in the JMAG was *Bacillus subtilis natto* and its frequent presence in the JMAG compared with the UHGG was the result of the Japanese unique diet.

Taxonomic and population annotation of the β-porphyranase

To gain functional insights into the reconstructed MAGs, we predicted 43,043,613 hypothetical proteins in the JMAG genomes and functionally annotated them. Most of the predicted proteins were covered by the eggNOG-mapper for the frequently reconstructed taxa, such as *Firmicutes_A*, *Bacteroidota*, and *Actinobacteria*, while they included a significant number of the functionary uncharacterized proteins (Figure S5A). Both the database coverage ratio and functional annotation ratio (ratio of the proteins that had any eggNOG-mapper hit and functionally characterized COG annotation, respectively; STAR Methods) for some taxa, such as *Cyanobacteria* and *Verrucomicrobiota*, were relatively low (Figure S5A). We found the phylum specificity of a part of the proteins. For example, GH92 (dbCAN2) and *susD* (Kyoto Encyclopedia of Genes and Genomes [KEGG] gene) were predominantly derived from the *Bacteroidota* (Figures S5B–S5E). We merged the predicted protein sequences of the JMAG to the UHGP and evaluated the overlap between the two datasets by clustering at 100%, 95%, 90%, and 50% sequence identities. Among the clusters that included the predicted proteins in the JMAG, 46.1%, 19.6%, 16.2%, and 9.5% were solely detected in the JMAG (Figure S5F).

Among the proteins in the JMAG, we focused on β-porphyranase, which catalyzes the hydrolysis of the seaweed-derived polysaccharides, namely porphyran. A previous study identified β-porphyranase in the *Phocaeicola plebeius* (renamed from *Bacteroides plebeius*) genome and revealed that β-porphyranase was detectable in the Japanese gut but not in the European gut[21] because the Japanese eat nori made from *porphyra*. However, its taxonomic origin and populational pattern were not fully evaluated because of the limited availability of the gut MSS data at that time. We identified the putative β-porphyranase sequences in the JMAG and UHGP, and all of them were placed close to the known β-porphyranase sequences in a maximum-likelihood phylogenetic tree, suggesting that our analysis successfully discriminated the β-porphyranase from other related proteins (Figure 2A). Among the β-porphyranase sequences in the JMAG, three sequences (JPN-Por1, JPN-Por-2, and JPN-Por-5) were also included in the UHGG (amino acid identity [AAI] > 99%), while the other five sequences were solely included in the JMAG (Figure S6A). We detected the 133 and 245 β-porphyranase sequences in the JMAG (ratio = $133/43,043,613 = 3.09 \times 10^{-6}$) and UHGP (ratio = $245/625,255,473 = 3.92 \times 10^{-7}$), respectively, suggesting that

Figure 2. Phylogenetic and interpopulational analysis of the β-porphyranase in the JMAG and UHGP

(A) A maximum-likelihood phylogenetic tree of the β-porphyranase sequences detected in the JMAG and UHGG. The β-porphyranase proteins and their related proteins (β-agarase and κ-carrageenase) were also utilized to reconstruct the phylogenetic tree. The colors and shapes of the nodes represent the derivation and name of the genes.

(B) Pie charts representing the phylogenetic composition of the MAGs, which are linked to the β-porphyranase proteins at the phylum (top) and genus level (bottom). The results for the JMAG (left) and UHGG (right) are described separately. The phyla comprising <1% of each genome set are collapsed into "Other."

(C) Pie charts represent the origin of the β-porphyranase-linked MAGs at the region (top) and country (bottom) levels.

(D) A boxplot of the β-porphyranase abundances in the different populations. 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]). IQR, interquartile range. See also Figures S5 and S6.

β-porphyranase was more frequent among the Japanese-derived gut prokaryotic genomes than those mainly derived from other populations. We evaluated the taxonomic origin of β-porphyranase and found that the majority of the taxonomy was *Bacteroidota* (78.9% in the JMAG and 93.9% in the UHGP), although *Firmicutes_A*-derived β-porphyranase proteins were also detected in both the JMAG and UHGP (18.8% and 5.3%, respectively; Figure 2B). At the genus level, we detected *Phocaeicola* and *Bacteroides* as the major origins of β-porphyranase both in the JMAG and UHGP. We also evaluated the populational pattern of β-porphyranase in the UHGP and found that most of the β-porphyranase sequences were derived from the Asian population (Figure 2C). As for the country-level annotation, although the ratio of the β-porphyranase sequences in the Chinese population was lower than the Japanese population (ratio = 177/125,294,874 = 1.41 × 10^{-6} for China and 13/2,048,327 = 6.35 × 10^{-6} for Japan), it was still higher than other country-level annotations, such as the US, Spain, and Denmark (ratio = 13/113,161,322 = 1.15 × 10^{-7}, 11/43,819,760 = 2.51 × 10^{-7}, 9/59,342,818 = 1.52 × 10^{-7}, respectively; Figure 2C). β-porphyranase was more abundant in Japanese than other populations in the read-based quantification (Figures 2D and S6B). Thus, we replicated the high frequency of β-porphyranase in the Japanese gut metagenome, and newly revealed that β-porphyranase presented also in the gut metagenome of the Chinese population.

Strains of food-associated bacteria were shared among the Japanese population

Utilizing the species-level representative genomes of the JMAG and the original MSS data, we evaluated the sharing of the prokaryotic strains among the Japanese by inStrain.[32] We first performed per dataset analysis and found that strain sharing was reproducibly detected for 10 species in at least 3 datasets (Figure S7A) among the 1,273 species in the JMAG. As for these 10 species, we performed a strain-level comparison with all samples. We found that the majority of the individuals included in the analysis of the targeted species were involved in strain sharing for five species (Figures 3A and S7B), suggesting that strain sharing was relatively frequent for these species compared with the other species in the JMAG. Among the five species, *Bacillus subtilis* was considered to be derived from the Japanese traditional food natto as mentioned above. In addition, the other four species (*Bifidobacterium animalis*, *Enterocossus_B lactis*, *Lactobacillus paracasei*, and *Streptococcus thermophilus*) were reported to be associated with dairy products.[33,34] Thus, it was suggested that food-related bacteria tended to be shared among the population at the strain level.

A missense variant rs671:G>A in *ALDH2* is an East Asian-specific single-nucleotide polymorphism that is under the recent positive selection.[35] The A allele of rs671 causes alcohol intolerance and has various pleiotropic associations with diseases, clinical biomarkers, and dietary habits.[36,37] Since the consumption of natto and dairy was negatively and positively associated with the A allele of the rs671, respectively, we evaluated the association between the abundance of the five food-related bacterial species and the A allele of the rs671 (Figure 3B; Table S3; N = 546 in total). We found nominal associations for *Enterococcus_B lactis* (effect size = 0.270 and p = 0.034) and *S. thermophilus* (effect size = 0.122 and p = 0.048). Even when removing disease samples, the effect sizes for the *Enterococcus_B lactis* (effect size = 0.355 and p = 0.036) and *S. thermophilus* (effect size = 0.134 and p = 0.079) were consistent (Figure S7C). We performed a Mendelian randomization analysis[38] to evaluate the effect of the dietary habits on the bacterial abundances and found that increased intake of dairy products could increase the abundances of the *Enterococcus_B lactis* (effect size = 2.385 and p = 0.034) and *S. thermophilus* (effect size = 1.077 and p = 0.047; Table S4).

Reconstruction of viral genomes from Japanese MSS data

We recovered viral genomes from the 787 Japanese MSS data (Figure S8). The viral genomes were extracted from the assembled contigs with VirSorter[39] and VirFinder[40] and subjected to CheckV.[41] After CheckV, we retained the viral genomes that had ≥50% completeness and more viral genes than host genes. We obtained 31,395 viral genomes including 4,098 complete, 7,492 high-quality, and 19,805 medium-quality genomes (Table S5). We call this set of viral genomes as Japanese Virus Database (JVD). The 31,395 genomes were clustered into 12,213 clusters at ≥95% ANI, merged with the Gut Phage Database (GPD),[19] Metagenomic Gut Virus (MGV),[20] and taxonomic reference genomes (RefSeq and Yutin et al.[42]), and further clustered into 94,714 species-level viral operational taxonomic unit (vOTU) at ≥95% ANI. These species-level vOTUs were further clustered into 10,022 genus- and 2,577 family-level vOTUs based on the gene sharing ratio and AAI (STAR Methods; Tables S6 and S7). We assigned putative viral taxonomy to all the viral genomes based on the result of the clustering (Figure 4A). Siphoviridae (14.0%) and Myoviridae (9.3%) dominated the taxonomically annotated viruses, while crAss-like phages (2.6%) and Podoviridae (0.8%) also occupied a portion of the taxonomically annotated viral genomes. Salsmaviridae, a recently created viral family,[43] also occupied a part of the taxonomically annotated viral genomes (0.7%).

We evaluated the overlap between the JVD, previous studies (GPD and MGV), and reference genomes at the family, genus, and species levels. At the species level, the majority of the vOTUs that included the JVD genomes (62.9%) were not overlapped with the other databases (Figure 4B). Note that there was a relatively large overlap between the GPD and MGV because of the overlap of the original MSS dataset. In contrast, the majority of the family- and genus-level vOTUs were covered by the other databases (7.5% and 0.67% were novel, respectively).

We predicted and functionally annotated the protein sequences on the JVD viral genomes. The ratio of the proteins covered and functionally annotated by the current databases was lower for the crAss-like phages than the other viruses, possibly due to the relatively recent discovery and expansion of the crAss-like phage genomes (Figure S9A). Among the Virus Orthologous Groups[44] and KEGG[45] annotations of the JVD, typical viral proteins, such as the capsid proteins, terminase, and portal proteins, were observed as highly frequent proteins (Figures S9B and S9C). Among the KEGG pathways, virus-related pathways, such as DNA replication and homologous recombination, were frequently

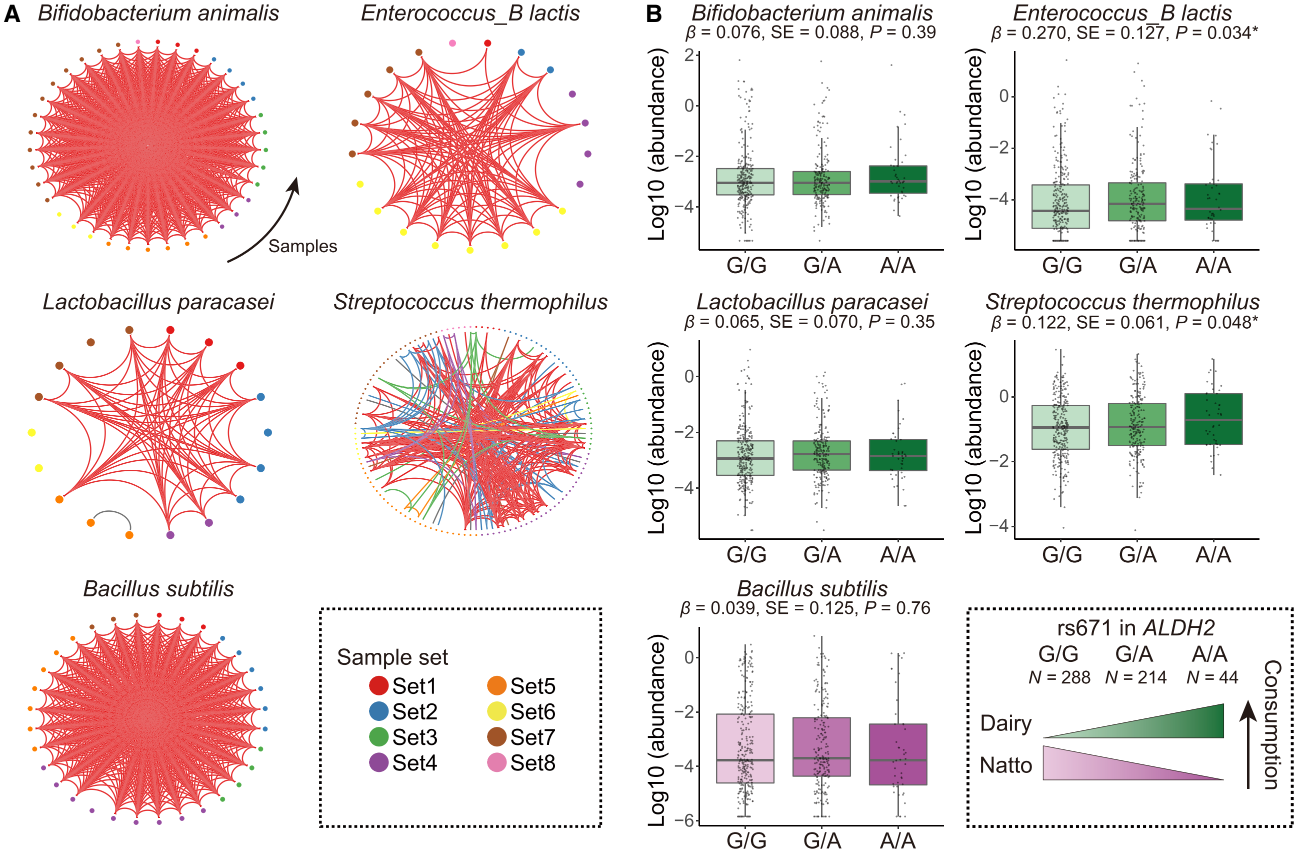


Figure 3. Strain-level analysis and association tests with rs671 for food-related bacterial species

(A) Circulized arc diagrams representing the strain sharing among the subjects for the five food-related bacterial species. The nodes represent the individuals with the detection of each species of bacterium and the edges represent the sharing of the bacterial strains. The colors of the nodes represent the dataset of the individuals. Independent strain-sharing networks are depicted in different colors. The gray edges mean that the strain is shared only between a pair of individuals.

(B) Boxplots represent the abundances (mean coverage) of the five food-related bacterial species stratified by the rs671 genotypes. The boxplots indicate 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]). The "A" allele of the rs671 in *ALDH2* had a negative association with natto consumption and a positive association with dairy consumption in the previous study.[37] *p < 0.05; IQR, interquartile range. See also Figure S7 and Tables S3 and S4.

seen (Figure S9D). In addition, we could see the taxonomic tendency of the KEGG gene and pathways, such as the relatively high occurrence of dUTP pyrophosphatase and pyrimidine metabolism-related proteins in the crAss-like phage genomes. We also detected some auxiliary metabolic genes[46] that potentially affect the metabolic function of their hosts (Figure S9E). Protein sequences were predicted also from the viral genomes in the GPD and MGV, merged with the JVD protein sequences, and clustered at 100%, 95%, 90%, and 50% amino acid sequence identity. Among the clusters that included the JVD proteins, 65.3%, 38.6%, 32.3%, and 19.4% were solely detected in the JVD, respectively (Figure S9F).

Interpopulational and case-control comparisons of the crAss-like phages

crAss-like phages were the bacteriophages that were reported to be abundant in the gut.[18] Since it was discovered in 2014 by a cross-assembly of the human gut metagenome data,[18] known diversity of the crAss-like phages has been expanded and now five subfamilies, namely $\alpha\gamma$, β, δ, ε, and ζ, are recognized.[42] We annotated the subfamily-level taxonomy to the crAss-like phage genomes based on the result of the genus-level vOTU clustering (Table S8). To validate the subfamily-level annotation, we made maximum-likelihood phylogenetic trees for the terminase (TerL), a marker protein of the crAss-like phages. The crAss-like phages belonging to the same subfamilies fell into the same clades, and those belonging to the same genus-level vOTU were placed closely (Figure 5A).

Then, we compared the subfamily-level composition of the crAss-like phage genomes among the various populational contexts. In the JVD, $\alpha\gamma$ followed by δ, ε, and ζ were frequent and β was minor among the crAss-like phage genomes. In the MGV, β crAss-like phages were also minor in Asia, Europe, and North America, as in the case of the JVD. In contrast, the composition of the β crAss-like phage genomes was significantly higher in Oceania and Africa than in Japan, Asia, Europe, and North America (Figure 5B; $P_{\mathrm{Fisher}} < 0.05/21 = 2.4 \times 10^{-3}$). A relatively higher prevalence of the β crAss-like phages in Africa was also

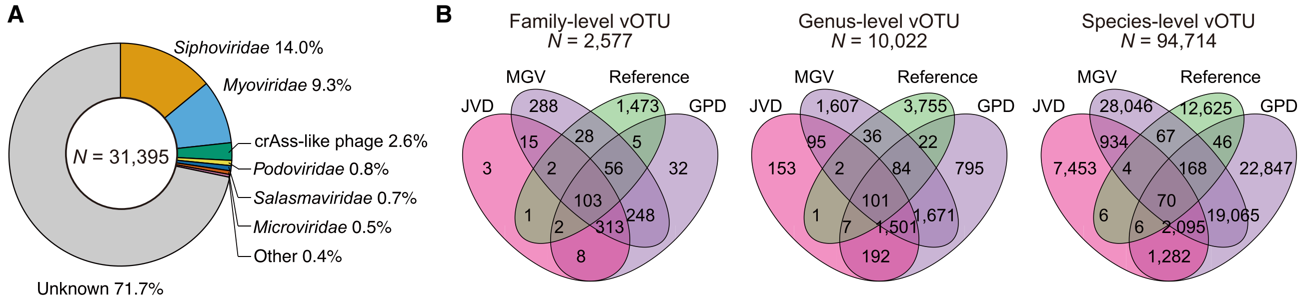


Figure 4. Reconstruction of the viral genomes and the comparison with the other databases

(A) A pie chart illustrating the family-level phylogenetic composition of the JVD. The families comprising <0.3% of the genomes are collapsed into "Other."

(B) Venn diagrams represent the sharing of the vOTUs (left, family level; middle, genus level; right, species level) among the different databases. Reference genomes are composed of RefSeq and Yutin et al.[42] (STAR Methods). See also Figures S8 and S9 and Tables S5, S6, and S7.

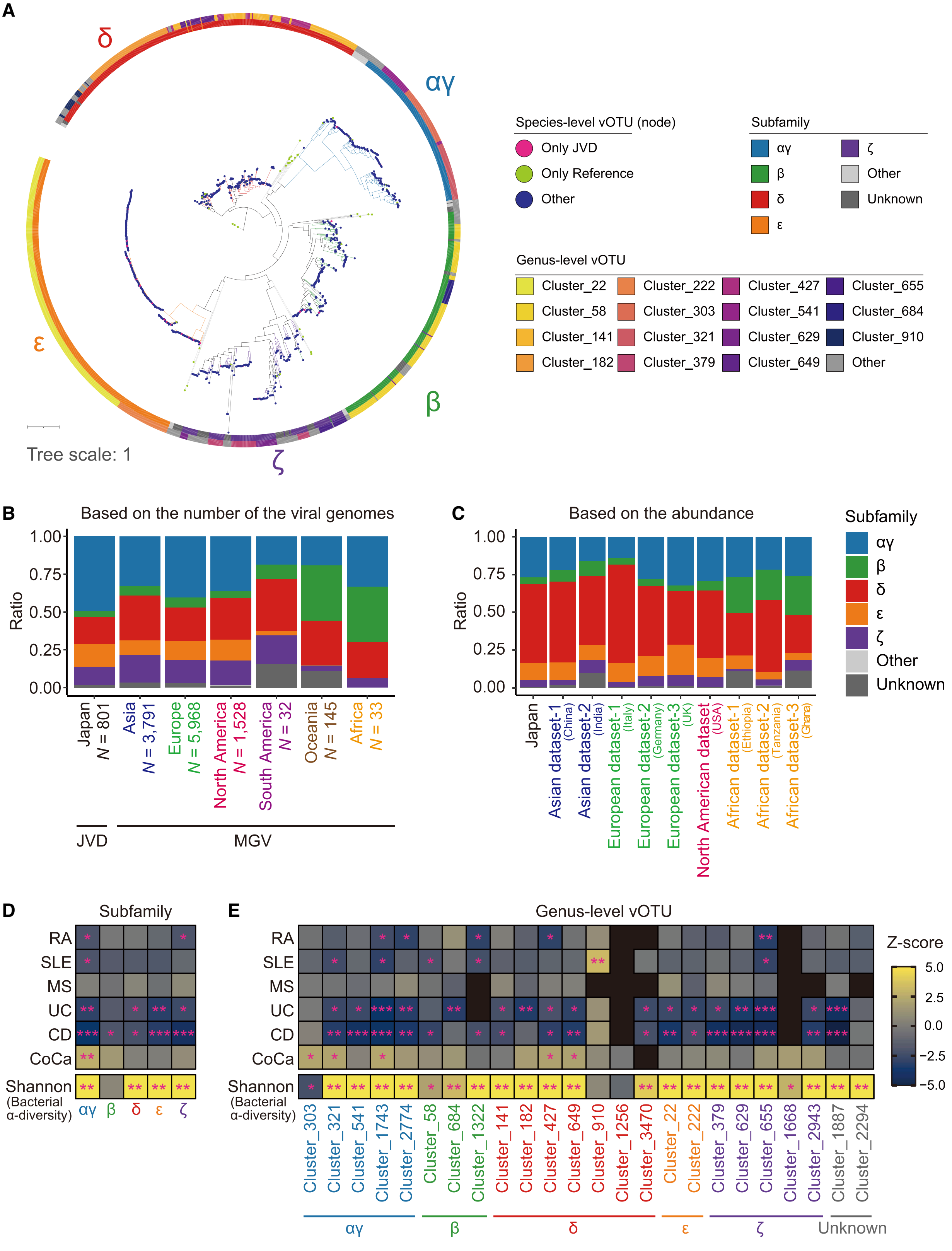
supported by the read-based quantification of the crAss-like phages (Figure 5C). Thus, it was suggested that the Japanese people's subfamily-level composition of the crAss-like phages was mostly similar to populations such as Asian, European, and North American, and β crAss-like phages were associated with the African and Oceanian populations. These results might reflect the differences in dietary habits.

Although crAss-like phages were assumed to be a core component of the healthy gut virome, their association to diseases had not been fully evaluated. Therefore, we evaluated the association between the subfamily- and genus-level vOTU of the crAss-like phages and affection status of the diseases, namely rheumatoid arthritis (RA) ($N_{\mathrm{Case}} = 113$, $N_{\mathrm{Control}} = 114$), systemic lupus erythematosus (SLE) ($N_{\mathrm{Case}} = 36$, $N_{\mathrm{Control}} = 205$), multiple sclerosis (MS) ($N_{\mathrm{Case}} = 30$, $N_{\mathrm{Control}} = 77$), ulcerative colitis (UC) ($N_{\mathrm{Case}} = 35$, $N_{\mathrm{Control}} = 40$), Crohn disease (CD) ($N_{\mathrm{Case}} = 39$, $N_{\mathrm{Control}} = 40$), and colorectal cancer (CoCa) ($N_{\mathrm{Case}} = 40$, $N_{\mathrm{Control}} = 39$; Figures 5D and 5E; Tables S9, S10, and S11). The $\alpha\gamma$, cluster_1743, cluster_1322, and cluster_655 crAss-like phages decreased at least nominally ($p < 0.05$) in both the RA and SLE patients. In MS patients, we could not detect any significant changes in the abundance of the crAss-like phages ($p > 0.05$). In patients with inflammatory bowel disease (IBD), namely UC and CD, most of the clades, including the $\alpha\gamma$, cluster_1743, and cluster_655 decreased ($p = 3.2 \times 10^{-3}$ and 3.0×10^{-4} for $\alpha\gamma$, $p = 1.9 \times 10^{-4}$ and 9.0×10^{-5} for cluster_1743, and $p = 3.5 \times 10^{-5}$ and 7.4×10^{-6} for cluster_655, respectively). In contrast, increases of some clades, such as the $\alpha\gamma$ crAss-like phages were observed in CoCa. Given that decreases of the diversity of the bacteria were reported for SLE,[25] UC, and CD,[47,48] but an increase was reported for CoCa,[49] we hypothesized that crAss-like phages were associated with the diversity of the bacteria. We evaluated the association between the crAss-like phage clades and Shannon index, which is a measurement of the diversity of the bacteria, and found that most of the clades were positively associated with the Shannon index (Figures 5D and 5E; Tables S10 and S11).

Virus-host interaction analysis with CRISPR, prophage, and co-abundance

CRISPR (clustered regularly interspaced short palindromic repeats) and CRISPR-associated (Cas) proteins comprise the CRISPR-Cas system, a prokaryotic adaptive immune system against predators such as bacteriophages.[50] The CRISPR-Cas system intakes short fragments of the viral sequences as CRISPR spacers to efficiently eject the viruses during subsequent infections. Thus, CRISPR sequences in the prokaryotic genomes are evidence of previous infections by viruses. Utilizing the CRISPR sequences in the JMAG genomes, we predicted the virus-prokaryote interaction. We detected 296,915 spacers in total, and 147,354 (49.6%) matched and 149,561 (50.4%) did not match the viral sequences recovered from the gut metagenome (Figure S10A). We then evaluated the taxonomic composition of the linked MAGs and viral targets of the CRISPR spacers, which reflected the host ranges of the viruses (Figure S10A; Table S12). For example, the major host of the crAss-like phages was *Bacteroidota*, while several crAss-like phages infected *Firmicutes_A*, as expected from previous studies.[20,42] We also searched the viral target sequences of the CRISPR spacers in the 286,997 UHGG genomes,[20] and 59% of the pairs of species-level vOTU and prokaryotic genus conferred from the analysis on the JMAG were replicated by the UHGG (Figure 6A). We also evaluated the virus-prokaryote interaction inferred from the proviral sequences in the JVD genomes (Figure S10B; Table S13). We got additional implications, such as the lack of the proviral sequences of the crAss-like phages and *Salasmaviridae*. The lack of proviral sequences of crAss-like phages in the JVD could reflect the lack of lysogeny of the crAss-like phages, as previously suggested.[20] As for *Salasmaviridae*, it was reported that *Salasmaviridae* follow a strict lytic life cycle with no evidence of lysogenic activity.[51] Thus, our large-scale analysis supported the previous implication for the newly classified virus.

Co-abundance analysis of the virus and prokaryote had been used for implicating virus-prokaryote interaction, but how much did it concordant to the result of the CRISPR-based and prophage-based analyses, which had not been well evaluated. Utilizing this large dataset, we evaluated the association between the abundances of viruses and prokaryotes stratified by the existence of supports from the CRISPR spacers in the JMAG (Figure S10C; Table S12). Inflation of the p values of the virus-prokaryote association tests was much more severe for the pairs supported by the CRISPR spacers than those without supports (Figure 6B). *Z* scores of the virus-prokaryote pairs supported by the CRISPR spacers were severely biased positively,



(legend on next page)

suggesting that the abundances of the viruses and their putative hosts tended to be positively correlated (Figure 6C). We performed the same analysis for the CRISPR sequences in the UHGG and the prophages in the JMAG and replicated the results obtained from the CRISPR sequences in the JMAG (Figures S10D–S10G).

Then, we performed the inter-database comparison of species-level vOTUs (JVD and MGV) and prokaryotic genome clusters (JMAG and UHGG) and integrated the results of these analyses based on the results of the CRISPR spacers. We calculated the odds ratio of the Japanese-derived genomes for each species-level vOTU and prokaryotic genome cluster. We found the enrichment of the CRISPR-supported virus-bacteria pairs that had the same sign of the log odds ratios for being Japanese derived (Figure 6D; STAR Methods). The log fold changes between the abundances in Japanese and other populations also tended to have same the signs for viruses and prokaryotes linked by the CRISPR spacers (Figure S10H). Thus, it was suggested that interpopulational differences of the viruses and their host were positively associated. For example, species-level vOTU 23245, which was frequently recovered and relatively abundant among the Japanese gut metagenome, infected *Blautia sp001304935*, which was also frequently recovered and relatively abundant among the Japanese gut metagenome (Figure 6E).

Virus-bacterium interaction network for crAss-like phages

Based on the result of the CRISPR analysis, we constructed a virus-bacterium interaction network of crAss-like phages (Figure 7A). The bacterial genera belonging to phylum *Bacteroidota*, such as *Parabacteroides*, *Prevotella*, *Bacteroides*, and *Phocaeicola* were highly connected to the crAss-like phages (Figure 7B), suggesting that the major host of the crAss-like phage was *Bacteroidota* as reported previously.[42] In addition, several *Firmicutes* were also present in the network. Although most of the crAss-like phage subfamilies infected various bacterial genera, ε crAss-like phages had strong preferences for the genus *Parabacteroides*.

DISCUSSION

In this study, we reconstructed the 19,084 MAGs and 31,395 viral genomes from the 787 Japanese gut MSS data. Utilizing these data, we performed a comparative analysis among databases, interpopulational and case-control comparisons of the crAss-like phages, and virus-prokaryote interaction analysis.

While a large part of the species-level diversity of the Japanese gut prokaryotes was covered by the UHGG catalog possibly due to the partially westernized dietary habits of the Japanese, some Japanese population-specific traditional diet-associated features of the gut microbiome, such as the presence of the *Bacillus subtilis natto* and enrichment of β-porphyranase, were identified. Natto is a Japanese traditional fermented food that is still widely consumed and expected as a potential probiotic food.[52] Although a previous 16S rRNA sequencing study suggested the presence of the family *Bacillacea* in the Japanese gut,[53] whether it was *Bacillus subtilis natto* was not confirmed due to the insufficient taxonomic resolution. Thus, our analysis suggested that the reconstruction of the MAG enabled us to evaluate *Bacillus subtilis natto* in the gut more accurately than 16S rRNA analysis and could be useful for future implementation of the probiotics.

β-Porphyranase is an enzyme that degrades seaweed-derived polysaccharides that are contained in the nori, a traditional Japanese food made from *porphyra*.[21] In our analysis, we confirmed the enrichment of the β-porphyranase in the Japanese gut with a large Japanese dataset, which had not been available in the previous study.[21] Although not as apparent as in the Japanese population, the frequency of β-porphyranase was relatively high in the Chinese population. Relative enrichment of β-porphyranase in the Chinese population could be because the Chinese population also eats *porphyra* as zicai or the long-standing traffic among East Asia.

Through strain-level analysis, we revealed that five strains of food-related bacterial species were reproducibly shared among the Japanese. A previous comparative analysis of gut-derived and food-derived MAGs revealed that the major source of several gut bacteria, including *L. paracasei* and *S. thermophilus*, was food.[33] Since the bacterial strains used for making fermented food are often determined by the manufacture, sharing of the strain for food-associated bacteria was expected when the major sources of the bacteria were food. rs671:G>A in *ALDH2* is the East Asian-specific missense variant that is associated with alcohol intolerance. We identified the positive association between the abundance of dairy-associated bacteria and A alleles of the rs671, which was also associated with high dairy consumption.[38] This finding suggested that human genetic variants could affect the gut microbiome via dietary habits, while we could not completely reject the possibility of the opposite (i.e., the high abundance of dairy consumption led to higher dairy consumption). Although not available for our datasets, future analysis with dietary information will be beneficial for deepening the insights into this association.

We mined the viral genomes from the MSS data. Among the taxonomically annotated viruses, *Siphoviridae*, *Myoviridae*, crAss-like phage, and *Podoviridae* were relatively frequent, as previously reported.[16,17,20,54] In addition, newly classified

Figure 5. Interpopulational and case-control comparisons of the crAss-like phages

(A) A maximum-likelihood phylogenetic tree reconstructed from the TerL proteins of the crAss-like phages. The color of the nodes represents the species-level clusters present only in the JVD (magenta), present only in the reference (green), or neither of the cases (navy). The outer rings represent the subfamily-level (inner) and genus-level (outer) taxonomic annotation of the crAss-like phages.

(B) A bar plot depicting the compositions of the subfamilies of the crAss-like phage genomes for the JVD and MGV. The genomes from the MGV are grouped according to their continental origin.

(C) A bar plot depicting the compositions of the subfamilies of the crAss-like phages calculated from the abundances (RPKM) in each group.

(D and E) Heatmaps represent the association of the crAss-like phages to the diseases (upper) and Shannon index (lower) at the subfamily (D) and genus (E) level, respectively. The colors indicate the *Z* score in each test. *p < 0.05. **p < 0.05/number of clades (per objective variables). ***p < 0.05/number of tests across all diseases (only for association tests for diseases). See also Tables S8, S9, S10, and S11.

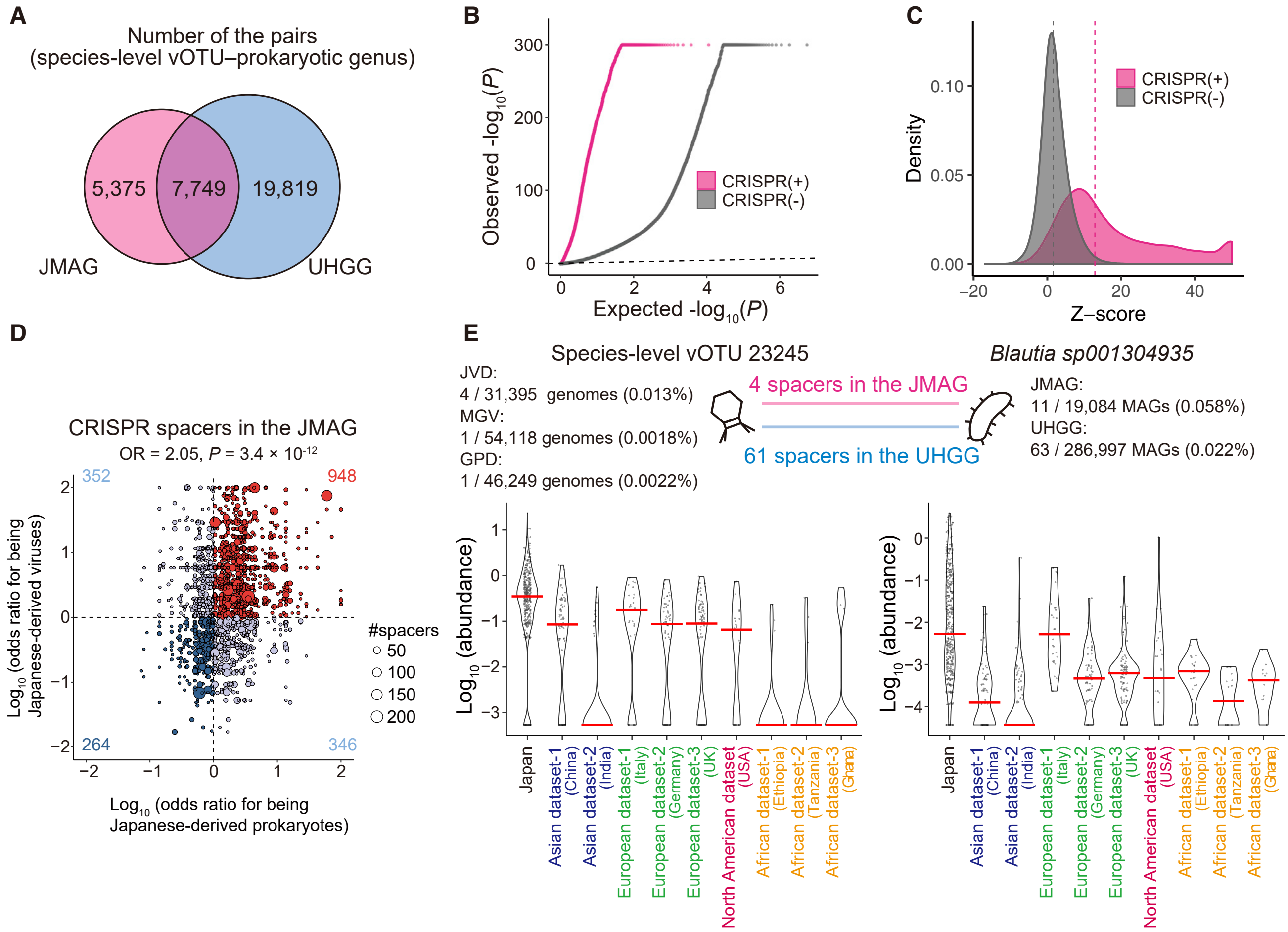


Figure 6. Virus-prokaryote interaction analysis based on the CRISPR and abundances

(A) Number of the pairs of the species-level vOTU and prokaryotic genus detected in the JMAG and UHGG.

(B) 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 JMAG (magenta) or not (gray). The x axis indicates $-\log_{10}(P)$ expected from the uniform distribution. The y axis indicates the observed $-\log_{10}(P)$. The diagonal dashed line represents $y = x$, which corresponds to the null hypothesis.

(C) A density plot of the Z score from the virus-prokaryote association analysis stratified by whether the virus-prokaryote pairs are supported by the CRISPR spacers 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.

(D) A scatterplot of the odds ratios for being the Japanese-derived viruses (y axis) and prokaryotes (x axis) for the virus-prokaryote pairs supported by the CRISPR in the JMAG. The size of the dots represents the number of spacers supporting the virus-prokaryote pairs. The horizontal and vertical dashed lines represent odds ratio = 0 for virus and prokaryote, respectively.

(E) Violin plots of the species-level vOTU 23245 (left) and *Blautia sp001304935* (right) abundances (RPKM) in each group. The red center lines indicate the median values. See also Figure S10 and Tables S12 and S13.

Salasmaviridae was also relatively frequent. As observed in the previous studies,[19,20] JVD included a significant amount of taxonomically unknown viruses possibly due to the underrepresentation of human gut phages in the taxonomic reference database. In contrast to prokaryotic genomes, a large part of the species-level diversity of the JVD was not covered by previous studies, such as GPD and MGV. This could be because of the enormous species-level diversity of the viruses, while differences in the populations and viral genome detection methods could also contribute.

We identified virus-prokaryote interaction by CRISPR and prophage analysis. Of the CRISPR spacers, 49.6% matched the viral sequence data composed of the JVD and the current largest gut virus databases (i.e., MGV and GPD), and future expansion of the viral sequence database may contribute to the further identification of the virus-prokaryote interaction. The abundance of the viruses and prokaryotes linked by the CRISPR spacers or proviral sequences was correlated positively in the gut. The Piggyback-the-Winner model,[55] in which phages take a lysogenic or pseudo-lysogenic cycle to "piggyback on" the success of their host rather than killing their host is supposed to be a major strategy for the gut virome.[56,57] Given that the lytic activities of the phages could result in a loss of positive correlation between the phages and their hosts,[58] our results could reflect the

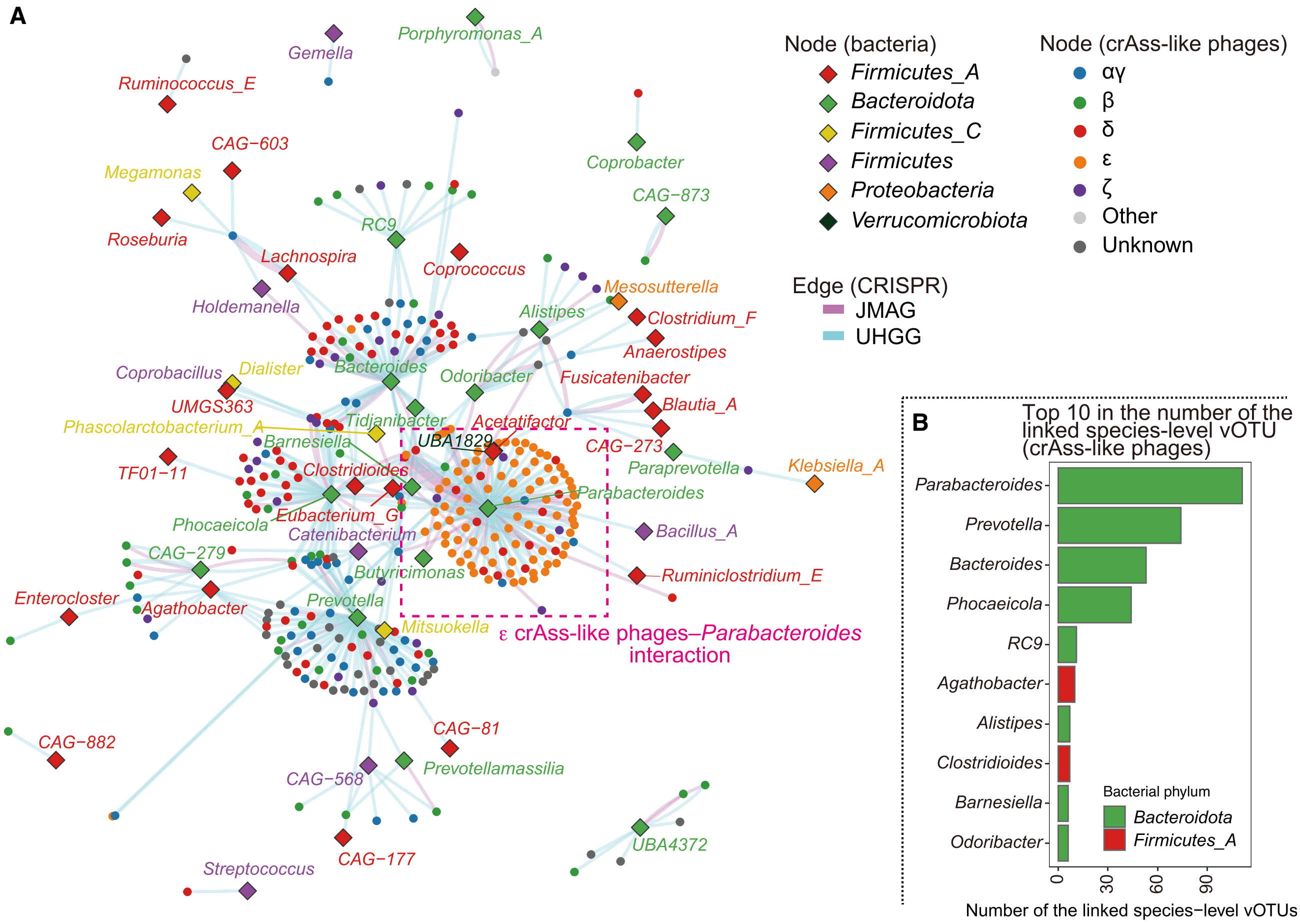


Figure 7. Network plot for the crAss-like phages and their predictive hosts

(A) A network plot of the CRISPR-based links (edges) between the species-level vOTU of the crAss-like phages (circle nodes) and bacterial genus (rhombus nodes). The color of the edges represents the derivation of the CRISPR spacers. The color of the circle nodes represents the subfamily level taxonomic annotations of the crAss-like phage genomes. The color of the rhombus nodes represents the genus level taxonomic annotations of the candidate hosts of the crAss-like phages. The magenta dashed box indicates the interaction between the ε crAss-like phages and *Parabacteroides*.

(B) A bar plot indicates the top 10 bacterial genera that have the highest number of the species-level crAss-like phage vOTU linked by the CRISPR spacers.

peaceful symbiosis as indicated in the Piggyback-the-Winner model. The interpopulational differences of the number of the recovered genome or read-based abundance had the same trend for the virus-prokaryote pairs supported by the CRISPR spacers. These results suggested that interpopulational differences of the viruses and their hosts were positively associated possibly because the abundances of the viruses and their hosts tended to be positively correlated.

At the subfamily level, the frequency of the recovery and read-based abundance of the β crAss-like phages were relatively high in the populations with the non-westernized dietary habits, such as African compared with populations with westernized dietary habits, including the Japanese. This result could reflect the impact of dietary habits on the crAss-like phages. In case-control comparisons of crAss-like phages, we revealed that several clades of the crAss-like phages decreased in RA, SLE, UC, and CD patients, but increased in CoCa patients. During the preparation of this manuscript, a study on Dutch cohorts reported decreases of the crAss-like phages in IBD.[59] Thus, decreases of the crAss-like phages in IBD could be a general event observed in multiple populations rather than a population-specific event. The diversity of the gut bacteriome has been reported to be associated with various diseases and is often suggested as a marker for microbiome health.[60] The positive association between the crAss-like phage abundances and bacterial diversity suggested that the abundance of the crAss-like phage could reflect the overall healthiness of the gut microbiome.

In virus-prokaryote interaction analysis, we could not find the proviral sequences of the crAss-like phages. Since the currently isolated two crAss-like phages (ΦcrAss001 and 002) neither possess lysogeny-associated genes nor can form stable lysogens,[61,62] this result could reflect the unique life cycle of crAss-like phages. Virus-prokaryote interaction analysis based on the CRISPR sequences predicted that the major host of the crAss-like phages was *Bacteroidota*, consistent with the previous finding.[42] Although most of the crAss-like phage subfamilies infected the various bacterial genus, ε crAss-like phages mostly exclusively infected the genus *Parabacteroides*. The limited host range might reflect relatively short evolutionary distances (length of the branches in phylogenetic trees) among the currently identified ε crAss-like phages.

In summary, we recovered the MAGs and viral genomes from the Japanese gut MSS data. Based on the recovered microbial genomes, we revealed the features of the Japanese gut metagenome, associations of the crAss-like phages to populations and diseases, and virus-prokaryote interactions. The reconstructed microbial genomes and related information are available at the National Bioscience Database Center (https://humandbs.biosciencedbc.jp). We believe that our dataset, which includes MAGs, viral genomes, and CRISPR spacers, will be a useful resource for future studies.

Limitation of the study

The JVD did not include viruses that were classified as RNA viruses or eukaryotic viruses because they were not efficiently detected by our pipeline due to the nature of the sequencing data and property of the virus detection pipeline. Future investigation on the other type of datasets such as *meta*-transcriptome data and further expansion of the reference databases will be beneficial to increase the known diversity of the gut virome.

Although the positively associated interpopulational differences were confirmed by the two different analyses (i.e., based on the number of the genomes and abundances) with the different outer datasets, batch/study effects were potential limitations of the current microbiome study focusing on the interpopulational differences. Ongoing efforts to collect and sequence stool samples from various populations in a unified framework will be promising.[63]

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Subject participation
- METHOD DETAILS
  - Sample collection and DNA extraction
  - Whole-genome shotgun sequencing
  - Quality control of sequencing reads
  - Reconstruction of MAGs
  - Analysis for the species-level representative MAGs
  - Comparative analysis of *Bacillus subtilis* genomes
  - Functional analysis of the MAGs
  - Analysis of the Japanese-specific species-level clusters
  - Strain-level analysis of the JMAG
  - Association tests between food-related bacteria and rs671
  - Reconstruction of viral genomes
  - Clustering and taxonomic annotation of the viral genomes
  - Functional analysis of the viral genomes
  - Subfamily-level annotation of the crAss-like phages
  - Interpopulational comparisons of the crAss-like phages
  - Case–control comparisons of the crAss-like phages
  - Virus–prokaryote association analysis based on the CRISPR and prophages
  - Virus–prokaryote association analysis based on the abundance
  - Comparison of the viral and prokaryotic numbers and abundances between Japanese and other populations
  - Network analysis of the crAss-like phages
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2022.100219.

ACKNOWLEDGMENTS

We would like to thank all the participants involved in this study. This research was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (19H01021 and 20K21834), the Japan Agency for Medical Research and Development (AMED; JP21km0405211, JP21ek0109413, JP21gm4010006, JP21km0405217, and JP21ek0410075), JST Moonshot R&D (JPMJMS2021 and JPMJMS2024), Takeda Science Foundation, Bioinformatics Initiative of Osaka University Graduate School of Medicine, and Grant Program for Next Generation Principal Investigators at Immunology Frontier Research Center (WPI-IFReC), Osaka University.

AUTHOR CONTRIBUTIONS

Y.T., T.K., and Y. Okada designed the study and conducted the data analysis. Y.T. and Y. Okada wrote the manuscript. Y.T., T.K., Y. Maeda, T.N., E.O., D.M., Y. Matsumoto, and S.N. conducted the experiments. Y.T., T.K., Y. Maeda, K.O., Y. Otake, S.K., T.N., T.O., E.O., M.K., M.T., N.O., K. Todo, K.Y., K.S., M. Yagita, A.H., H. Matsuoka, M. Yoshimura, S.O., S.S., and H. Iijima collected and managed the samples. H. Iijima, H. Inohara, H.K., T.T., H. Mochizuki, K. Takeda, A.K., and Y. Okada supervised the study.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 5, 2022
Revised: June 27, 2022
Accepted: October 31, 2022
Published: November 30, 2022

REFERENCES

1. Holmes, E., Li, J.V., Marchesi, J.R., and Nicholson, J.K. (2012). Gut microbiota composition and activity in relation to host metabolic phenotype and disease risk. Cell Metab. *16*, 559–564. https://doi.org/10.1016/j.cmet.2012.10.007.
2. Zou, Y., Xue, W., Luo, G., Deng, Z., Qin, P., Guo, R., Sun, H., Xia, Y., Liang, S., Dai, Y., et al. (2019). 1, 520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nat. Biotechnol. *37*, 179–185. https://doi.org/10.1038/s41587-018-0008-8.
3. Poyet, M., Groussin, M., Gibbons, S.M., Avila-Pacheco, J., Jiang, X., Kearney, S.M., Perrotta, A.R., Berdy, B., Zhao, S., Lieberman, T.D.,

et al. (2019). A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. Nat. Med. *25*, 1442–1452. https://doi.org/10.1038/s41591-019-0559-3.

4. Forster, S.C., Kumar, N., Anonye, B.O., Almeida, A., Viciani, E., Stares, M.D., Dunn, M., Mkandawire, T.T., Zhu, A., Shao, Y., et al. (2019). A human gut bacterial genome and culture collection for improved metagenomic analyses. Nat. Biotechnol. *37*, 186–192. https://doi.org/10.1038/s41587-018-0009-7.

5. Nayfach, S., Shi, Z.J., Seshadri, R., Pollard, K.S., and Kyrpides, N.C. (2019). New insights from uncultivated genomes of the global human gut microbiome. Nature *568*, 505–510. https://doi.org/10.1038/s41586-019-1058-x.

6. Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., et al. (2019). Extensive unexplored human microbiome diversity revealed by over 150, 000 genomes from metagenomes spanning age, geography, and lifestyle. Cell *176*, 649–662.e20. https://doi.org/10.1016/j.cell.2019.01.001.

7. Almeida, A., Mitchell, A.L., Boland, M., Forster, S.C., Gloor, G.B., Tarkowska, A., Lawley, T.D., and Finn, R.D. (2019). A new genomic blueprint of the human gut microbiota. Nature *568*, 499–504. https://doi.org/10.1038/s41586-019-0965-1.

8. Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z.J., Pollard, K.S., Sakharova, E., Parks, D.H., Hugenholtz, P., et al. (2021). A unified catalog of 204, 938 reference genomes from the human gut microbiome. Nat. Biotechnol. *39*, 105–114. https://doi.org/10.1038/s41587-020-0603-3.

9. Shkoporov, A.N., and Hill, C. (2019). Bacteriophages of the human gut: the "known unknown" of the microbiome. Cell Host Microbe *25*, 195–209. https://doi.org/10.1016/j.chom.2019.01.017.

10. Keen, E.C., and Dantas, G. (2018). Close encounters of three kinds: bacteriophages, commensal bacteria, and host immunity. Trends Microbiol. *26*, 943–954. https://doi.org/10.1016/j.tim.2018.05.009.

11. Guerin, E., and Hill, C. (2020). Shining light on human gut bacteriophages. Front. Cell. Infect. Microbiol. *10*, 481. https://doi.org/10.3389/fcimb.2020.00481.

12. Norman, J.M., Handley, S.A., Baldridge, M.T., Droit, L., Liu, C.Y., Keller, B.C., Kambal, A., Monaco, C.L., Zhao, G., Fleshner, P., et al. (2015). Disease-specific alterations in the enteric virome in inflammatory bowel disease. Cell *160*, 447–460. https://doi.org/10.1016/j.cell.2015.01.002.

13. Clooney, A.G., Sutton, T.D.S., Shkoporov, A.N., Holohan, R.K., Daly, K.M., O'Regan, O., Ryan, F.J., Draper, L.A., Plevy, S.E., Ross, R.P., and Hill, C. (2019). Whole-virome analysis sheds light on viral dark matter in inflammatory bowel disease. Cell Host Microbe *26*, 764–778.e5. https://doi.org/10.1016/j.chom.2019.10.009.

14. Ma, Y., You, X., Mai, G., Tokuyasu, T., and Liu, C. (2018). A human gut phage catalog correlates the gut phageome with type 2 diabetes. Microbiome *6*, 24. https://doi.org/10.1186/s40168-018-0410-y.

15. Zhao, G., Vatanen, T., Droit, L., Park, A., Kostic, A.D., Poon, T.W., Vlamakis, H., Siljander, H., Härkönen, T., Hämäläinen, A.M., et al. (2017). Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children. Proc. Natl. Acad. Sci. USA *114*, E6166–E6175. https://doi.org/10.1073/pnas.1706359114.

16. Fujimoto, K., Kimura, Y., Shimohigoshi, M., Satoh, T., Sato, S., Tremmel, G., Uematsu, M., Kawaguchi, Y., Usui, Y., Nakano, Y., et al. (2020). Metagenome data on intestinal phage-bacteria associations aids the development of phage therapy against pathobionts. Cell Host Microbe *28*, 380–389.e9. https://doi.org/10.1016/j.chom.2020.06.005.

17. Gregory, A.C., Zablocki, O., Zayed, A.A., Howell, A., Bolduc, B., and Sullivan, M.B. (2020). The gut virome database reveals age-dependent patterns of virome diversity in the human gut. Cell Host Microbe *28*, 724–740.e8. https://doi.org/10.1016/j.chom.2020.08.003.

18. Dutilh, B.E., Cassman, N., McNair, K., Sanchez, S.E., Silva, G.G.Z., Boling, L., Barr, J.J., Speth, D.R., Seguritan, V., Aziz, R.K., et al. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat. Commun. *5*, 4498. https://doi.org/10.1038/ncomms5498.

19. Camarillo-Guerrero, L.F., Almeida, A., Rangel-Pineros, G., Finn, R.D., and Lawley, T.D. (2021). Massive expansion of human gut bacteriophage diversity. Cell *184*, 1098–1109.e9. https://doi.org/10.1016/j.cell.2021.01.029.

20. Nayfach, S., Páez-Espino, D., Call, L., Low, S.J., Sberro, H., Ivanova, N.N., Proal, A.D., Fischbach, M.A., Bhatt, A.S., Hugenholtz, P., and Kyrpides, N.C. (2021). Metagenomic compendium of 189, 680 DNA viruses from the human gut microbiome. Nat. Microbiol. *6*, 960–970. https://doi.org/10.1038/s41564-021-00928-6.

21. Hehemann, J.-H., Correc, G., Barbeyron, T., Helbert, W., Czjzek, M., and Michel, G. (2010). Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. Nature *464*, 908–912. https://doi.org/10.1038/nature08937.

22. Nishijima, S., Suda, W., Oshima, K., Kim, S.-W., Hirose, Y., Morita, H., and Hattori, M. (2016). The gut microbiome of healthy Japanese and its microbial and functional uniqueness. DNA Res. *23*, 125–133. https://doi.org/10.1093/dnares/dsw002.

23. Kishikawa, T., Maeda, Y., Nii, T., Motooka, D., Matsumoto, Y., Matsushita, M., Matsuoka, H., Yoshimura, M., Kawada, S., Teshigawara, S., et al. (2020). Metagenome-wide association study of gut microbiome revealed novel aetiology of rheumatoid arthritis in the Japanese population. Ann. Rheum. Dis. *79*, 103–111. https://doi.org/10.1136/annrheumdis-2019-215743.

24. Kishikawa, T., Ogawa, K., Motooka, D., Hosokawa, A., Kinoshita, M., Suzuki, K., Yamamoto, K., Masuda, T., Matsumoto, Y., Nii, T., et al. (2020). A metagenome-wide association study of gut microbiome in patients with multiple sclerosis revealed novel disease pathology. Front. Cell. Infect. Microbiol. *10*, 585973. https://doi.org/10.3389/fcimb.2020.585973.

25. Tomofuji, Y., Maeda, Y., Oguro-Igashira, E., Kishikawa, T., Yamamoto, K., Sonehara, K., Motooka, D., Matsumoto, Y., Matsuoka, H., Yoshimura, M., et al. (2021). Metagenome-wide association study revealed disease-specific landscape of the gut microbiome of systemic lupus erythematosus in Japanese. Ann. Rheum. Dis. *80*, 1575–1583. https://doi.org/10.1136/annrheumdis-2021-220687.

26. Tomofuji, Y., Kishikawa, T., Maeda, Y., Ogawa, K., Nii, T., Okuno, T., Oguro-Igashira, E., Kinoshita, M., Yamamoto, K., Sonehara, K., et al. (2022). Whole gut virome analysis of 476 Japanese revealed a link between phage and autoimmune disease. Ann. Rheum. Dis. *81*, 278–288. https://doi.org/10.1136/annrheumdis-2021-221267.

27. Yachida, S., Mizutani, S., Shiroma, H., Shiba, S., Nakajima, T., Sakamoto, T., Watanabe, H., Masuda, K., Nishimoto, Y., Kubo, M., et al. (2019). Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat. Med. *25*, 968–976. https://doi.org/10.1038/s41591-019-0458-7.

28. Otake-Kasamoto, Y., Kayama, H., Kishikawa, T., Shinzaki, S., Tashiro, T., Amano, T., Tani, M., Yoshihara, T., Li, B., Tani, H., et al. (2022). Lysophosphatidylserines derived from microbiota in Crohn's disease elicit pathological Th1 response. J. Exp. Med. *219*, e20211291. https://doi.org/10.1084/jem.20211291.

29. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. *25*, 1043–1055. https://doi.org/10.1101/gr.186072.114.

30. Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., Schulz, F., Jarett, J., Rivers, A.R., Eloe-Fadrosh, E.A., et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat. Biotechnol. *35*, 725–731. https://doi.org/10.1038/nbt.3893.

31. Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. (2019). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy

Database. Bioinformatics *36*, 1925–1927. https://doi.org/10.1093/bioinformatics/btz848.

32. Olm, M.R., Crits-Christoph, A., Bouma-Gregson, K., Firek, B.A., Morowitz, M.J., and Banfield, J.F. (2021). inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nat. Biotechnol. *39*, 727–736. https://doi.org/10.1038/s41587-020-00797-0.

33. Pasolli, E., De Filippis, F., Mauriello, I.E., Cumbo, F., Walsh, A.M., Leech, J., Cotter, P.D., Segata, N., and Ercolini, D. (2020). Large-scale genome-wide analysis links lactic acid bacteria from food with the gut microbiome. Nat. Commun. *11*, 2610. https://doi.org/10.1038/s41467-020-16438-8.

34. Morandi, S., Cremonesi, P., Povolo, M., and Brasca, M. (2012). Enterococcus lactis sp. nov., from Italian raw milk cheeses. Int. J. Syst. Evol. Microbiol. *62*, 1992–1996. https://doi.org/10.1099/ijs.0.030825-0.

35. Okada, Y., Momozawa, Y., Sakaue, S., Kanai, M., Ishigaki, K., Akiyama, M., Kishikawa, T., Arai, Y., Sasaki, T., Kosaki, K., et al. (2018). Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. Nat. Commun. *9*, 1631. https://doi.org/10.1038/s41467-018-03274-0.

36. Sakaue, S., Kanai, M., Tanigawa, Y., Karjalainen, J., Kurki, M., Koshiba, S., Narita, A., Konuma, T., Yamamoto, K., Akiyama, M., et al. (2021). A cross-population atlas of genetic associations for 220 human phenotypes. Nat. Genet. *53*, 1415–1424. https://doi.org/10.1038/s41588-021-00931-x.

37. Matoba, N., Akiyama, M., Ishigaki, K., Kanai, M., Takahashi, A., Momozawa, Y., Ikegawa, S., Ikeda, M., Iwata, N., Hirata, M., et al. (2020). GWAS of 165, 084 Japanese individuals identified nine loci associated with dietary habits. Nat. Hum. Behav. *4*, 308–316. https://doi.org/10.1038/s41562-019-0805-1.

38. Hemani, G., Zheng, J., Elsworth, B., Wade, K.H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R., et al. (2018). The MR-Base platform supports systematic causal inference across the human phenome. Elife *7*, e34408. https://doi.org/10.7554/eLife.34408.

39. Roux, S., Enault, F., Hurwitz, B.L., and Sullivan, M.B. (2015). VirSorter: mining viral signal from microbial genomic data. PeerJ *3*, e985. https://doi.org/10.7717/peerj.985.

40. Ren, J., Ahlgren, N.A., Lu, Y.Y., Fuhrman, J.A., and Sun, F. (2017). VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome *5*, 69. https://doi.org/10.1186/s40168-017-0283-5.

41. Nayfach, S., Camargo, A.P., Schulz, F., Eloe-Fadrosh, E., Roux, S., and Kyrpides, N.C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat. Biotechnol. *39*, 578–585. https://doi.org/10.1038/s41587-020-00774-7.

42. Yutin, N., Benler, S., Shmakov, S.A., Wolf, Y.I., Tolstoy, I., Rayko, M., Antipov, D., Pevzner, P.A., and Koonin, E.V. (2021). Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. Nat. Commun. *12*, 1044. https://doi.org/10.1038/s41467-021-21350-w.

43. Walker, P.J., Siddell, S.G., Lefkowitz, E.J., Mushegian, A.R., Adriaenssens, E.M., Alfenas-Zerbini, P., Davison, A.J., Dempsey, D.M., Dutilh, B.E., García, M.L., et al. (2021). Changes to virus taxonomy and to the international code of virus classification and nomenclature ratified by the international committee on taxonomy of viruses (2021). Arch. Virol. *166*, 2633–2648. https://doi.org/10.1007/s00705-021-05156-1.

44. Grazziotin, A.L., Koonin, E.V., and Kristensen, D.M. (2017). Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic Acids Res. *45*, D491–D498. https://doi.org/10.1093/nar/gkw975.

45. Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30. https://doi.org/10.1093/nar/28.1.27.

46. Kieft, K., Zhou, Z., and Anantharaman, K. (2020). VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome *8*, 90. https://doi.org/10.1186/s40168-020-00867-0.

47. Franzosa, E.A., Sirota-Madi, A., Avila-Pacheco, J., Fornelos, N., Haiser, H.J., Reinker, S., Vatanen, T., Hall, A.B., Mallick, H., McIver, L.J., et al. (2019). Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nat. Microbiol. *4*, 293–305. https://doi.org/10.1038/s41564-018-0306-4.

48. Lloyd-Price, J., Arze, C., Ananthakrishnan, A.N., Schirmer, M., Avila-Pacheco, J., Poon, T.W., Andrews, E., Ajami, N.J., Bonham, K.S., Brislawn, C.J., et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature *569*, 655–662. https://doi.org/10.1038/s41586-019-1237-9.

49. Thomas, A.M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., Beghini, F., Manara, S., Karcher, N., Pozzi, C., et al. (2019). Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat. Med. *25*, 667–678. https://doi.org/10.1038/s41591-019-0405-7.

50. Deveau, H., Garneau, J.E., and Moineau, S. (2010). CRISPR/Cas system and its role in phage-bacteria interactions. Annu. Rev. Microbiol. *64*, 475–493. https://doi.org/10.1146/annurev.micro.112408.134123.

51. Stanton, C.R., Rice, D.T.F., Beer, M., Batinovic, S., and Petrovski, S. (2021). Isolation and characterisation of the bundooravirus genus and phylogenetic investigation of the Salasmaviridae bacteriophages. Viruses *13*, 1557. https://doi.org/10.3390/v13081557.

52. Wang, P., Gao, X., Li, Y., Wang, S., Yu, J., and Wei, Y. (2020). Bacillus natto regulates gut microbiota and adipose tissue accumulation in a high-fat diet mouse model of obesity. J. Funct.Foods *68*, 103923. https://doi.org/10.1016/j.jff.2020.103923.

53. Oki, K., Toyama, M., Banno, T., Chonan, O., Benno, Y., and Watanabe, K. (2016). Comprehensive analysis of the fecal microbiota of healthy Japanese adults reveals a new bacterial lineage associated with a phenotype characterized by a high frequency of bowel movements and a lean body type. BMC Microbiol. *16*, 284. https://doi.org/10.1186/s12866-016-0898-x.

54. Zuo, T., Sun, Y., Wan, Y., Yeoh, Y.K., Zhang, F., Cheung, C.P., Chen, N., Luo, J., Wang, W., Sung, J.J.Y., et al. (2020). Human-gut-DNA virome variations across geography, ethnicity, and urbanization. Cell Host Microbe *28*, 741–751.e4. https://doi.org/10.1016/j.chom.2020.08.005.

55. Silveira, C.B., and Rohwer, F.L. (2016). Piggyback-the-Winner in host-associated microbial communities. NPJ Biofilms Microbiomes *2*, 16010. https://doi.org/10.1038/npjbiofilms.2016.10.

56. Reyes, A., Haynes, M., Hanson, N., Angly, F.E., Heath, A.C., Rohwer, F., and Gordon, J.I. (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature *466*, 334–338. https://doi.org/10.1038/nature09199.

57. Moreno-Gallego, J.L., Chou, S.-P., Di Rienzi, S.C., Goodrich, J.K., Spector, T.D., Bell, J.T., Youngblut, N.D., Hewson, I., Reyes, A., and Ley, R.E. (2019). Virome diversity correlates with intestinal microbiome diversity in adult monozygotic twins. Cell Host Microbe *25*, 261–272.e5. https://doi.org/10.1016/j.chom.2019.01.019.

58. Faruque, S.M., Naser, I.B., Islam, M.J., Faruque, A.S.G., Ghosh, A.N., Nair, G.B., Sack, D.A., and Mekalanos, J.J. (2005). Seasonal epidemics of cholera inversely correlate with the prevalence of environmental cholera phages. Proc. Natl. Acad. Sci. USA *102*, 1702–1707. https://doi.org/10.1073/pnas.0408992102.

59. Gulyaeva, A., Garmaeva, S., Ruigrok, R.A.A.A., Wang, D., Riksen, N.P., Netea, M.G., Wijmenga, C., Weersma, R.K., Fu, J., Vila, A.V., et al. (2022). Discovery, diversity, and functional associations of crAss-like phages in human gut metagenomes from four Dutch cohorts. Cell Rep. *38*, 110204. https://doi.org/10.1016/j.celrep.2021.110204.

60. Mosca, A., Leclerc, M., and Hugot, J.P. (2016). Gut microbiota diversity and human diseases: should we reintroduce key predators in our

ecosystem? Front. Microbiol. *7*, 455. https://doi.org/10.3389/fmicb.2016.00455.

61. Shkoporov, A.N., Khokhlova, E.V., Fitzgerald, C.B., Stockdale, S.R., Draper, L.A., Ross, R.P., and Hill, C. (2018). ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. Nat. Commun. *9*, 4781. https://doi.org/10.1038/s41467-018-07225-7.

62. Guerin, E., Shkoporov, A.N., Stockdale, S.R., Comas, J.C., Khokhlova, E.V., Clooney, A.G., Daly, K.M., Draper, L.A., Stephens, N., Scholz, D., et al. (2021). Isolation and characterisation of ΦcrAss002, a crAss-like phage from the human gut that infects Bacteroides xylanisolvens. Microbiome *9*, 89. https://doi.org/10.1186/s40168-021-01036-7.

63. Rabesandratana, T. (2018). Microbiome conservancy stores global fecal samples. Science *362*, 510–511. https://doi.org/10.1126/science.362.6414.510.

64. Zhu, F., Ju, Y., Wang, W., Wang, Q., Guo, R., Ma, Q., Sun, Q., Fan, Y., Xie, Y., Yang, Z., et al. (2020). Metagenome-wide association of gut microbiome features for schizophrenia. Nat. Commun. *11*, 1612. https://doi.org/10.1038/s41467-020-15457-9.

65. Dhakan, D.B., Maji, A., Sharma, A.K., Saxena, R., Pulikkan, J., Grace, T., Gomez, A., Scaria, J., Amato, K.R., and Sharma, V.K. (2019). The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches. GigaScience *8*, giz004. https://doi.org/10.1093/gigascience/giz004.

66. Wirbel, J., Pyl, P.T., Kartal, E., Zych, K., Kashani, A., Milanese, A., Fleck, J.S., Voigt, A.Y., Palleja, A., Ponnudurai, R., et al. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat. Med. *25*, 679–689. https://doi.org/10.1038/s41591-019-0406-6.

67. Xie, H., Guo, R., Zhong, H., Feng, Q., Lan, Z., Qin, B., Ward, K.J., Jackson, M.A., Xia, Y., Chen, X., et al. (2016). Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. Cell Syst. *3*, 572–584.e3. https://doi.org/10.1016/j.cels.2016.10.004.

68. Tett, A., Huang, K.D., Asnicar, F., Fehlner-Peach, H., Pasolli, E., Karcher, N., Armanini, F., Manghi, P., Bonham, K., Zolfo, M., et al. (2019). The Prevotella copri complex comprises four distinct clades underrepresented in westernized populations. Cell Host Microbe *26*, 666–679.e7. https://doi.org/10.1016/j.chom.2019.08.018.

69. Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., Busk, P.K., Xu, Y., and Yin, Y. (2018). dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. *46*, W95–W101. https://doi.org/10.1093/nar/gky418.

70. BMTagger. ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/.

71. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

72. Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Lahti, L., Loman, N.J., Andersson, A.F., and Quince, C. (2014). Binning metagenomic contigs by coverage and composition. Nat. Methods *11*, 1144–1146. https://doi.org/10.1038/nmeth.3103.

73. Sieber, C.M.K., Probst, A.J., Sharrar, A., Thomas, B.C., Hess, M., Tringe, S.G., and Banfield, J.F. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat. Microbiol. *3*, 836–843. https://doi.org/10.1038/s41564-018-0171-1.

74. Buchfink, B., Reuter, K., and Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat. Methods *18*, 366–368. https://doi.org/10.1038/s41592-021-01101-x.

75. Olm, M.R., Brown, C.T., Brooks, B., and Banfield, J.F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. *11*, 2864–2868. https://doi.org/10.1038/ismej.2017.126.

76. Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol. Biol. Evol. *38*, 5825–5829. https://doi.org/10.1093/molbev/msab293.

77. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909. https://doi.org/10.1038/ng1847.

78. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274. https://doi.org/10.1093/molbev/msu300.

79. Letunic, I., and Bork, P. (2019). Interactive Tree of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. *47*, W256–W259. https://doi.org/10.1093/nar/gkz239.

80. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780. https://doi.org/10.1093/molbev/mst010.

81. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. *17*, 132. https://doi.org/10.1186/s13059-016-0997-x.

82. Wu, Y.-W., Simmons, B.A., and Singer, S.W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics *32*, 605–607. https://doi.org/10.1093/bioinformatics/btv638.

83. Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. *30*, 1575–1584. https://doi.org/10.1093/nar/30.7.1575.

84. Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ *7*, e7359. https://doi.org/10.7717/peerj.7359.

85. Bland, C., Ramsey, T.L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N.C., and Hugenholtz, P. (2007). CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics *8*, 209. https://doi.org/10.1186/1471-2105-8-209.

86. Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. *35*, 1026–1028. https://doi.org/10.1038/nbt.3988.

87. Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., and Zimin, A. (2018). MUMmer4: a fast and versatile genome alignment system. PLoS Comput. Biol. *14*, e1005944. https://doi.org/10.1371/journal.pcbi.1005944.

88. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797. https://doi.org/10.1093/nar/gkh340.

89. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421. https://doi.org/10.1186/1471-2105-10-421.

90. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575. https://doi.org/10.1086/519795.

91. Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics *27*, 863–864. https://doi.org/10.1093/bioinformatics/btr026.

92. Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and

translation initiation site identification. BMC Bioinformatics *11*, 119. https://doi.org/10.1186/1471-2105-11-119.

93. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069. https://doi.org/10.1093/bioinformatics/btu153.

94. Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., and Tyson, G.W. (2017). Recovery of nearly 8, 000 metagenome-assembled genomes substantially expands the tree of life. Nat. Microbiol. *2*, 1533–1542. https://doi.org/10.1038/s41564-017-0012-7.

95. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

96. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., and Korobeynikov, A. (2020). Using SPAdes de novo assembler. Curr. Protoc. Bioinformatics *70*, e102. https://doi.org/10.1002/cpbi.102.

97. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

98. Chan, P.P., Lin, B.Y., Mak, A.J., and Lowe, T.M. (2021). tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. *49*, 9077–9096. https://doi.org/10.1093/nar/gkab688.

99. Kawabata, S., Takagaki, M., Nakamura, H., Oki, H., Motooka, D., Nakamura, S., Nishida, T., Terada, E., Izutsu, N., Takenaka, T., et al. (2022). Dysbiosis of gut microbiome is associated with rupture of cerebral aneurysms. Stroke *53*, 895–903. https://doi.org/10.1161/STROKEAHA.121.034792.

100. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods *14*, 587–589. https://doi.org/10.1038/nmeth.4285.

101. Nishito, Y., Osana, Y., Hachiya, T., Popendorf, K., Toyoda, A., Fujiyama, A., Itaya, M., and Sakakibara, Y. (2010). Whole genome assembly of a natto production strain Bacillus subtilis natto from very short read data. BMC Genomics *11*, 243. https://doi.org/10.1186/1471-2164-11-243.

102. Sakaue, S., Yamaguchi, E., Inoue, Y., Takahashi, M., Hirata, J., Suzuki, K., Ito, S., Arai, T., Hirose, M., Tanino, Y., et al. (2021). Genetic determinants of risk in autoimmune pulmonary alveolar proteinosis. Nat. Commun. *12*, 1032. https://doi.org/10.1038/s41467-021-21011-y.

103. Bin Jang, H., Bolduc, B., Zablocki, O., Kuhn, J.H., Roux, S., Adriaenssens, E.M., Brister, J.R., Kropinski, A.M., Krupovic, M., Lavigne, R., et al. (2019). Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat. Biotechnol. *37*, 632–639. https://doi.org/10.1038/s41587-019-0100-8.

104. Yutin, N., Makarova, K.S., Gussow, A.B., Krupovic, M., Segall, A., Edwards, R.A., and Koonin, E.V. (2018). Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. Nat. Microbiol. *3*, 38–46. https://doi.org/10.1038/s41564-017-0053-y.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Fecal samples	This study	N/A
Human DNA extracted from blood	This study	N/A
Chemicals, peptides, and recombinant proteins		
Tris-HCl	NIPPON GENE	Cat#316-90385
SDS	Sigma Aldrich	Cat#28-3270
EDTA	Nacalai Tesque	Cat#06894-14
Phenol/chloroform/isoamyl alcohol	Nacalai Tesque	Cat#25970-56
TE saturated phenol	Nacalai Tesque	Cat#26829-96
Sodium acetate	Sigma Aldrich	Cat#28-1560
Isopropanol	JUNSEI	Cat#67-63-0
Ethanol	JUNSEI	Cat#64-19-5
RNA later	Thermo Fisher Scientific	Cat#AM7021
Critical commercial assays		
KAPA Hyper Prep Kit	illumina	Cat#KK8504
Glass beads (diameter 0.1 mm)	biospec	Cat#11079101
Deposited data		
Metagenome shotgun sequencing data	This study	National Bioscience Database Center (NBDC) Human Database: hum0197
Metagenome shotgun sequencing data	Kishikawa et al. 2020a[23]	National Bioscience Database Center (NBDC) Human Database: hum0197
Metagenome shotgun sequencing data	Kishikawa et al. 2020b[24]	National Bioscience Database Center (NBDC) Human Database: hum0197
Metagenome shotgun sequencing data	Tomofuji et al., 2021a[25]	National Bioscience Database Center (NBDC) Human Database: hum0197
Metagenome shotgun sequencing data	Tomofuji et al., 2021b[26]	National Bioscience Database Center (NBDC) Human Database: hum0197
Metagenome shotgun sequencing data	Otake et al., 2022[28]	National Bioscience Database Center (NBDC) Human Database: hum0197
Metagenome shotgun sequencing data	Yachida et al., 2019[27]	DDBJ Sequence Read Archive: DRA006684
Metagenome shotgun sequencing data	Zhu et al. 2020[64]	Europea Nucleotide Archive: ERP111403
Metagenome shotgun sequencing data	Dhakan et al. 2019[65]	Sequence Read Archive: SRP114847
Metagenome shotgun sequencing data	Thomas et al. 2019[49]	Sequence Read Archive: SRP136711
Metagenome shotgun sequencing data	Wirbel et al. 2019[66]	Europea Nucleotide Archive: ERP110064
Metagenome shotgun sequencing data	Xie et al. 2016[67]	Europea Nucleotide Archive: ERP010700
Metagenome shotgun sequencing data	Price et al. 2019[48]	Sequence Read Archive: SRP115494
Metagenome shotgun sequencing data	Tett et al. 2019[68]	Sequence Read Archive: SRP168387
Metagenome shotgun sequencing data	Tett et al. 2019[68]	Sequence Read Archive: SRP189832
Metagenome shotgun sequencing data	Tett et al. 2019[68]	Sequence Read Archive: SRP189572
RefSeq Virus	NCBI	https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/
Bacillus subtilis genomes	NCBI GenBank	https://www.ncbi.nlm.nih.gov/genbank/
crAss-like phage genomes	Yutin et al. 2021[42]	https://zenodo.org/record/4437596
CRISPR spacers	Nayfach et al. 2021[20]	https://portal.nersc.gov/MGV
CRISPR spacers in JMAG genomes	This study	National Bioscience Database Center (NBDC) Human Database: hum0197
dbCAN HMMdb v10	Zhang et al., 2018[69]	https://bcb.unl.edu/dbCAN2/index.php

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
GPD	Camarillo-Guerrero et al. 2021[19]	http://ftp.ebi.ac.uk/pub/databases/metagenomics/genome_sets/gut_phage_database/
JMAG	This study	National Bioscience Database Center (NBDC) Human Database: hum0197
JVD	This study	National Bioscience Database Center (NBDC) Human Database: hum0197
List of the AMGs	Kieft et al., 2020[46]	https://doi.org/10.1186/s40168-020-00867-0
MGV	Nayfach et al. 2021[20]	https://portal.nersc.gov/MGV
Scripts for recovering and analyzing microbial genomes	This study	https://doi.org/10.5281/zenodo.7053099 and https://github.com/ytomofuji
UHGG and UHGP	Almeida et al. 2021[8]	http://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/
VOG	Grazziotin et al., 2017[44]	https://vogdb.org
β-porphyranase sequences	Hehemann et al., 2010[21]	https://doi.org/10.1038/nature08937
Bacillus subtilis genomes	NCBI GenBank	https://www.ncbi.nlm.nih.gov/genbank/
Multiple sequence alignment files generated in this study (JMAG representative genomes, β-porphyranase, and TerL of crAss-like phages)	This study	https://doi.org/10.5281/zenodo.7053099
Software and algorithms		
Barrnap	https://github.com/tseemann/barrnap	https://github.com/tseemann/barrnap
bcl2fastq	Illumina	https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software/downloads.html
BMTagger	ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/[70]	ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/
bowtie2	Langmead and Salzberg, 2012[71]	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
CheckM	Parks et al., 2015[29]	https://github.com/Ecogenomics/CheckM
CheckV	Nayfach et al., 2021[41]	https://bitbucket.org/berkeleylab/checkv/
CONCOCT	Alneberg et al., 2014[72]	https://github.com/BinPro/CONCOCT
coverM	Queensland University of Technology Microbiome Research Group	https://github.com/wwood/CoverM
DAS Tool	Sieber et al., 2018[73]	https://github.com/cmks/DAS_Tool
DIAMOND	Buchfink et al., 2021[74]	https://github.com/bbuchfink/diamond
dRep	Olm et al., 2017[75]	https://github.com/MrOlm/drep
eggNOG-mapper	Cantalapiedra et al. 2021[76]	https://github.com/eggnogdb/eggnog-mapper
EIGENSTRAT	Price et al., 2006[77]	https://www.hsph.harvard.edu/alkes-price/software/
Ggraph	https://github.com/thomasp85/ggraph	https://github.com/thomasp85/ggraph
GTDB-tk	Chaumeil et al., 2019[31]	https://github.com/Ecogenomics/GTDBTk
Hmmer	http://hmmer.org/download.html	http://hmmer.org/download.html
inStrain	Olm et al., 2021[32]	https://github.com/MrOlm/instrain
Iqtree	Nguyen, L.-T et al., 2015[78]	http://www.iqtree.org
iTOL	Letunic & Bork, 2019[79]	https://itol.embl.de
MAFFT	Katoh & Standley, 2013[80]	https://mafft.cbrc.jp/alignment/software/
Mash	Ondov et al., 2016[81]	https://github.com/marbl/Mash
MaxBin	Wu et al., 2016[82]	https://sourceforge.net/projects/maxbin/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
MCL	Enright et al., 2002[83]	http://micans.org/mcl/
MetaBAT	Kang et al., 2019[84]	https://bitbucket.org/berkeleylab/metabat/src/master/
MinCED	Bland et al., 2007[85]	https://github.com/ctSkennerton/minced
MMseqs2	Steinegger & Söding, 2017[86]	https://github.com/soedinglab/MMseqs2
MUMmer	Marçais et al., 2018[87]	https://github.com/mummer4/mummer
muscle	Edgar, 2004[88]	https://drive5.com/muscle/downloads_v3.htm
ncbi-blast-plus	Camacho et al., 2009[89]	https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download
PLINK	Purcell et al., 2007[90]	https://www.cog-genomics.org/plink/
PRINSEQ	Schmieder and Edwards, 2011[91]	http://prinseq.sourceforge.net/
Prodigal	Hyatt et al., 2010[92]	https://github.com/hyattpd/Prodigal
Prokka	Seemann, 2014[93]	https://github.com/tseemann/prokka
Python	Python Software Foundation	https://www.python.org/downloads/release/python-376/
R	The R Foundation for Statistical Computing	https://www.r-project.org
RefineM	Parks et al., 2017[94]	https://github.com/dparks1134/RefineM
Samtools	Li et al., 2009[95]	http://www.htslib.org/download/
Script for clustering of the viral genomes	Nayfach et al., 2021[20]	https://github.com/snayfach/MGV
SPAdes	Prjibelski et al., 2020[96]	https://github.com/ablab/spades#sec5
Trimmomatic	Bolger et al., 2014[97]	http://www.usadellab.org/cms/?page=trimmomatic
tRNAScan-SE	Chan et al., 2021[98]	http://lowelab.ucsc.edu/tRNAscan-SE/
TwoSampleMR	Hemani et al., 2018[38]	https://mrcieu.github.io/TwoSampleMR/
Vegan	https://github.com/vegandevs/vegan	https://github.com/vegandevs/vegan
VirFinder	Ren et al., 2017[40]	https://github.com/jessieren/VirFinder
VirSorter	Roux et al., 2015[39]	https://github.com/simroux/VirSorter
Custom codes used in this study	This study	https://doi.org/10.5281/zenodo.7053099 and https://github.com/ytomofuji/JMAG_JVD

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Yukinori Okada (yokada@sg.med.osaka-u.ac.jp).

Materials availability

The materials that support the findings of this study are available from the corresponding authors upon reasonable request. Please contact the lead contact for additional information.

Data and code availability

The JMAG genomes, JVD genomes, and CRISPR sequences are available in NBDC Human Database (http://humandbs.biosciencedbc.jp/) with the accession number of hum0197. The JMAG genomes, JVD genomes, and CRISPR sequences can also be downloaded from the DNA DataBank of Japan (DDBJ) with the accession numbers provided in Table S14. Detailed metadata for the JMAG and JVD genomes are provided as Tables S2 and S5, respectively. The MSS data are under the controlled access in NBDC Human Database (http://humandbs.biosciencedbc.jp/) with the accession number of hum0197 to protect the participants' privacy. Applications from all the researchers who comply with the NBDC's data terms of use are quickly assessed and accepted. Multiple sequence alignment files for the maximum-likelihood phylogenetic trees (representative JMAG genomes, β-porphyranase, and TerL of crAss-like phages) are available in Zenodo (https://doi.org/10.5281/zenodo.7053099). Codes used for the analysis and instructions for downloading JMAG genomes, JVD genomes, and CRISPR sequences from DDBJ are available in GitHub (https://github.com/ytomofuji/JMAG_JVD) and Zenodo (https://doi.org/10.5281/zenodo.7053099).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Subject participation

818 Japanese gut metagenome sequencing data from 787 subjects were used in this study (Table S1). In addition, 432 gut metagenome sequencing data from various populations[48,49,64–68] were used for the comparative analyses. Although most of the data was derived from previous studies,[23–25,27,28] 136 Japanese sequencing data (included healthy control [HC], Unruptured cerebral aneurysm [UA], Sub-arachnoid hemorrhage [SAH], and stroke [ST] subjects) was newly obtained in this study. The newly recruited HC subjects were enrolled at the Osaka University Graduate School of Medicine. Participants with UA and SAH were recruited from the Osaka University, Osaka Neurological Institution, Hanwa Memorial Hospital, and Iseikai Hospital as previously described.[99] Participants with ST were recruited from the Osaka University.

Participants with extreme diets (e.g., strict vegetarians) were not included in the dataset. All subjects provided written informed consent before participation. Those who took antibiotics within a month were reported as the patients treated with antibiotics. The study protocol was approved by the ethics committees of Osaka University and related medical institutions.

METHOD DETAILS

Sample collection and DNA extraction

For the ST patients, fecal samples had been immediately frozen after production in an insulated container for storage at −20°C and subsequently stored at −80°C within 24 h after production. For the HCs, samples were stored at −80°C within 6 h after production. For the participants with UA, fecal samples were collected at home, immediately packed with frozen gel packs within insulated containers, and stored at −20°C. By the next day, the sample collection kits were returned by refrigerated shipping keeping at −20°C, and stored at −80°C until processing, as previously described.[99] For the participants with SAH, the fecal samples were collected within 48 h following admission and before the induction of antibiotics to minimize changes in the gut microbial community, as previously described.[99] Microbial DNA was extracted according to the previously described method.[23] Briefly, 0.3 g glass beads (diameter: 0.1 mm) (BioSpec) and 500 μL EDTA-Tris-saturated phenol were added to the suspension, and the mixture was vortexed vigorously using a FastPrep-24 (MP Biomedicals) at 5.0 power level for 30 s. After centrifugation at 20,000 g for 5 min at 4°C, 400 μL of supernatant was collected. Subsequently, phenol-chloroform extraction was performed, and 250 μL of supernatant was subjected to isopropanol precipitation. Finally, DNAs were suspended in 100 μL EDTA-Tris buffer and stored at −20°C.

Whole-genome shotgun sequencing

A shotgun sequencing library was constructed using the KAPA Hyper Prep Kit (KAPA Biosystems), and 150-bp paired-end reads were generated on NovaSeq 6000. The sequence reads were converted to the FASTQ format using bcl2fastq (version 2.19).

Quality control of sequencing reads

We followed a series of steps to maximize the quality of the datasets. The main steps in the quality control process were as follows: (i) trimming of low-quality bases, (ii) identification and masking of human reads, and (iii) removal of duplicated reads. We marked duplicate reads using PRINSEQ-lite[91] (version 0.20.4; -derep 1). We trimmed the raw reads to clip Illumina adapters and cut off low-quality bases at both ends using the Trimmomatic[97] (version 0.39; parameters: ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10:8:true LEADING:20 TRAILING:20 SLIDINGWINDOW:3:15 MINLEN:60). We discarded reads less than 60 bp in length after trimming. Next, we performed duplicate removal by retaining only the longest read among the duplicates. When there were multiple reads with the same sequences and length, we randomly selected one of the reads. As a final quality control step, we aligned the quality-filtered reads to the human reference genome (hg38) using bowtie2[71] (version 2.3.5.1) with default parameters and BMTagger[70] (version 3.101). We kept only the reads of which both paired ends failed to align in either tool.

Reconstruction of MAGs

The *de novo* assembly of the filtered paired-end reads into the contigs was conducted using SPAdes[96] (version 3.13.0) with the '—meta' option and the contigs longer than 2kbp were retained for subsequent binning. Then, filtered paired-end reads were mapped to the assembled contigs for quantifying the abundance of each contig with bowtie2 (version 2.3.5.1). Binning was performed per sample using three different tools with the default options; MaxBin[82] (version 2.2.6), MetaBAT[84] (version 2.12.1), and CONCOCT[72] (version 1.0.0). DAS Tool[73] (version 1.1.2) was used to integrate the results of the binning produced by the three tools. To refine the quality of the bins, we utilized RefineM[94] (version 0.1.2) and filtered out scaffolds with the divergent genomic properties or incongruent taxonomic classification (based on Genome Taxonomy Database release 95). Then, we evaluated the quality of the MAGs with CheckM[29] (version 1.0.12) using the 'lineage_wf' workflow to select only genomes that passed the following criteria; >50% genome completeness, <5% contamination, and an estimated quality score (completeness – 5 × contamination) > 50. After the filtering, we obtained 19,084 MAGs which were used for the subsequent analyses.

To evaluate the strain-level diversity of the MAGs, we mapped the filtered paired-end reads to the reconstructed MAGs with bowtie2 and calculated the average nucleotide diversity by inStrain[32] (version 1.5.4) per sample. Evaluation of the average nucleotide diversity was performed per dataset because it was originally reported to be affected by the sequencing batches. Then, the read coverages of the reconstructed MAGs in originated samples were calculated by coverM (version 0.6.1). We searched for the presence of rRNAs in each MAG by barrnap (version 0.9) with the following parameters; –kingdom bac (for MAGs determined as bacteria by CheckM), –kingdom arc (for MAGs determined as archaea by CheckM), –reject 0.8, –evalue 1e-3. tRNAs of the standard 20 amino acids were identified by tRNAScan-SE[98] (version 2.0.7) with the following parameters; -A (for MAGs determined as bacteria by CheckM), -B (for MAGs determined as archaea by CheckM).

Analysis for the species-level representative MAGs

The 19,084 reconstructed MAGs were clustered into estimated species-level clusters (ANI $\geq$95%) by dRep[75] (version 3.2.0) with the following parameters; -pa 0.9 -sa 0.95 -nc 0.30 -cm larger. Following score was calculated for each MAGs based on the output of CheckM and the genome with the highest score was selected as the representative genome for each species-level cluster; score = Completeness $-5 \times$ Contamination $+0.5 \times \log_{10}$(N50). After the dereplication at the species level, we obtained 1,273 species-level clusters and representative genomes. We then annotated taxonomy to the species-level representative genomes with GTDB-tk[31] (version 1.5.3) based on the Genome Taxonomy Database release 202. Taxonomy of the non-representative genomes was assigned according to the taxonomy of the representative genomes of their clusters. For the subsequent comparisons, the UHGG genomes were also subjected to taxonomic annotation because the reference database for GTDB-tk was updated from the version used in the original study.[8]

For each of the species-level representative genomes, we checked the existence of the same species-level clusters in the UHGG. First, we estimated the ANI between the 1,273 reconstructed MAGs and the 4,644 UHGG genomes by mash[81] (version 2.3) with the sketch size 1000. Based on the result of mash, we extracted pairs of the genomes with mash-based ANI $\geq$90%. Then, we calculated ANI for the extracted pairs of the genomes with the dnadiff function of MUMmer[87] (version 4.0.0.rc1). For each of the 1,273 MAGs, we assigned corresponding species-level genomes in the UHGG which had ANI $\geq$95%, aligned fraction $\geq$30%, and the highest ANI to the query MAGs.

Among the 1,273 species-level representative MAGs, we extracted 1,267 bacterial MAGs for the construction of a maximum-likelihood phylogenetic tree. A multiple sequencing alignment (MSA) of the core genes generated by GTDB-tk were subjected to the iqtree[78] (version 2.1.2). ‘LG + F + R10’ was chosen as the best-fit model by the ModelFinder[100] and constructed phylogenetic tree was visualized with iTOL[79] (version 6).

Comparative analysis of *Bacillus subtilis* genomes

To characterize the reconstructed *Bacillus subtilis* MAGs (26 genomes), we performed a comparative analysis with the *Bacillus subtilis* genomes retrieved from the GenBank (162 genomes) and UHGG (1 genome). We calculated pair-wide ANI for all the pairs of the *Bacillus subtilis* genomes with the dnadiff function of MUMmer. Then, we performed hierarchical clustering by the hclust function in the R (version 4.0.1) with the ‘method = ”average”’ option. After clustering, we extracted a cluster which included all the *Bacillus subtilis* MAGs in the JMAG by cutree function in the R with the ‘k = 10’ option. Then, we performed NMDS of the extracted cluster.

To confirm that *Bacillus subtilis* MAGs in the JMAG were closely related to *Bacillus subtilis natto*, we checked the genetic variations of the *degQ* promoter and *swrAA* (*yvzD*) coding regions which were previously reported to be different between *Bacillus subtilis natto* and *Bacillus subtilis* 168[101]. We made MSAs of these genomic regions from the *Bacillus subtilis natto* genomes, *Bacillus subtilis 168* genomes, and *Bacillus subtilis* genomes in JMAG by muscle[88] (version 3.8.31).

For the comparison of the microbial abundances in boxplots, only the HC samples were used. Quality-controlled reads were down-sampled to 1,000,000 paired-ends reads to adjust the differences of the library sizes between the datasets. Then, the down-sampled reads were mapped to the reference genome of the reconstructed 1,273 MAGs with bowtie2 and the abundances were calculated as Reads Per Kilobase of exon per Million mapped reads (RPKM) by coverM.

Functional analysis of the MAGs

Protein-coding genes for each of the 19,084 MAGs were predicted with Prokka[93] (version 1.14.6) with the specification of the kingdom annotated by CheckM. Predicted proteins were subjected to the eggNOG-mapper[76] (version 2.1.2) for the annotation of Cluster of Orthologous Groups (COG) and KEGG and the calculation of the database coverage ratio and functional annotation ratio. The database coverage ratio was defined as the ratio of the protein sequences which were assigned with any eggNOG-mapper hits including unknown functions. The functional annotation ratio was defined as the ratio of the protein sequences which were assigned with COG annotations other than S (Function unknown) and R (General function prediction only). Annotation of the carbohydrate-active enzymes (CAZyme) was performed separately with the hmmscan function in hmmer (version 3.1b2) and dbCAN HMMdb v10[69] was used as a reference hmm profile. E-values less than 1×10^{-18} were regarded as significant in the annotation of the CAZymes.

The predicted protein sequences on the MAGs were dereplicated with MMseqs2[86] (version 13.45111) with the following parameters; –cov-mode 1 -c 0.8 –kmer-per-seq 80 –min-seq-id 1. The Dereplicated set of the protein sequences were then merged with the UHGP-100 and subjected to further clustering with the following parameters; –cov-mode 1 -c 0.8 –kmer-per-seq 80. The ‘–min-seq-id’ option in the second clustering was set at 1, 0.95, 0.9, and 0.5 to dereplicate the protein sequences at 100%, 95%, 90%, and 50% amino acid sequence identity, respectively.

We identified the β-porphyranase sequences in the JMAG and UHGP. We first performed a blastp search with diamond[74] (version 2.0.4) ‘–ultra-sensitive’ mode. The dereplicated protein sequences for the JMAG and UHGP were queried against the β-porphyranase sequences identified in the previous study[21] and available in NCBI (PorA, PorB, PorC, PorD, and PorE). Since the β-porphyranase has high sequence similarity to other proteins such as β-agarase and κ-carrageenase, we set a relatively strict threshold for E-values ($<1 \times 10^{-40}$). In addition, we constructed a maximum-likelihood phylogenetic tree from the identified β-porphyranase sequences and other related proteins (i.e. β-porphyranase, β-agarase, and κ-carrageenase) published in the previous study[21] for confirming that our pipeline discriminated β-porphyranase from other related proteins. First, we made an MSA with MAFFT[80] (version 7.486) with the '–auto' parameter. Then, we generated a phylogenetic tree by iqtree with the 'VT + F + R4' model which was chosen as the best-fit model by the ModelFinder and visualized it with iTOL. To profile the taxonomic and geographic features of the β-porphyranase among the JMAG and UHGP, we extracted all the protein sequences which belong to the protein clusters of the β-porphyranase. For the calculation of the AAI between the β-porphyranase sequences, we performed an all vs all blastp search with the default setting and pident was used as the AAI. For the read-based quantification of the β-porphyranase, we translated and mapped the 1,000,000 paired-ends reads against the non-redundant β-porphyranase sequences in the JMAG and UHGP (Figure S6A), using the 'blastx' function in the diamond. We extracted the blastx hits with ≥95% identity and E-value $< 10^{-10}$. If the blastx had multiple hits, hits with the highest bitscore were selected. Abundance was calculated as a (total length of the alignment length of the query sequences)/(total sequencing length).

Analysis of the Japanese-specific species-level clusters

To identify Japanese-specific species-level clusters, we checked the (i) number of the JMAG genomes and (ii) number of the non-Japanese-derived MAGs contained in the corresponding UHGG clusters for all of the species-level clusters in the JMAG. The species-level clusters which contained ≥10 JMAG genomes and ≤1 UHGG genome were defined as the Japanese-specific species-level clusters. Based on the eggNOG-mapper annotation, we profiled the CAZyme profiles of the MAGs in these clusters. We extracted the CAZymes which satisfied (i) [within-cluster ratio of the MAGs which had the CAZymes] > 0.75, (ii) [within-cluster ratio of the MAGs which had the CAZymes] > 5 × [within-phylum ratio of the MAGs which had the CAZymes], and (iii) [within-cluster ratio of the MAGs which had the CAZymes] > 5 × [within-JMAG ratio of the MAGs which had the CAZymes]. We extracted the protein sequence clusters made by MMSeqs2 (dereplicated at 90% AAI) which included the extracted CAZymes. For the extracted protein sequence clusters, we checked the (number of the protein sequences from the Japanese-specific species-level cluster)/(number of the protein sequences in the JMAG) to evaluate the uniqueness of the CAZyme profiles of the Japanese-specific species-level clusters among the JMAG. We also checked the (number of the protein sequences in the JMAG)/(number of the protein sequences in the JMAG and UHGP) to evaluate the Japanese-specificity of the extracted CAZymes. For the read-based quantification of the CAZymes, we translated and mapped the 1,000,000 paired-ends reads against the extracted CAZyme sequences described in Figure S3A, using the 'blastx' function in the diamond. We extracted the blastx hits with ≥95% identity and E-value $< 10^{-10}$. If the blastx had multiple hits, hits with the highest bitscore were selected. Abundance was calculated as a (total length of the alignment length of the query sequences)/(total sequencing length). Only the HC samples were used for the calculation of the mean abundances.

Strain-level analysis of the JMAG

Reference prokaryotic genomes composed of the 1,273 species-level representative JMAG genomes were indexed with bowtie2. Then, we mapped the quality-controlled sequencing reads to the reference genomes with bowtie2. The mapped-read data were converted to bam format by samtools[95] (version 1.10) and individually subjected to the 'profile' function in inStrain with the '–database_mode' option. Then, the results of the 'profile' function were merged with the 'compare' function in inStrain per dataset because merging the results of all the samples was not computationally scaled. We set a threshold for the population ANI (popANI; a metric introduced by Olm et al.[32] to detect the strain-sharing) at ≥99.999% to define the sharing of the strain between two individuals according to the validation in the original study. As for the taxa for which strain sharing was detected in at least three datasets, the 'compare' function in inStrain was run with all the samples with the specification of the single taxa.

Association tests between food-related bacteria and rs671

We genotyped the 550 subjects using Infinium Asian Screening Array (Illumina, San Diego, CA, USA). This genotyping array was built using an East Asian reference panel including whole-genome sequences, which enabled effective genotyping in East Asian populations.

We applied stringent quality control filters to the genotyping dataset using PLINK[90] (version 1.90b4.4) as described elsewhere.[102] We confirmed that genotyping call rate was <0.98 for all the individuals. For pairs of closely related individuals (PI_HAT calculated by PLINK >0.185), we removed either of the related individuals. We confirmed that only the individuals of the estimated East Asian ancestry were included in this study, based on the principal component analysis with the samples of the HapMap project using EIGENSTRAT.[77] After the quality control procedures, we obtained the genotype data of rs671 for 546 subjects (Table S3).

As for the five bacterial species which satisfied (number of the samples involved in the strain-sharing)/(number of the samples used for the analysis of the target species) ≥ 0.5 in the strain-sharing analysis, we obtained the abundances. Note that samples with the usage of antibiotics were not included in this analysis. Quality-controlled reads were mapped to the reference genome of the reconstructed 1,273 MAGs with bowtie2, and the mean coverages of each genome calculated by coverM genome function were divided by 'total sequencing length/1,000,000,000' and subjected to the log transformation.

We evaluated the association between the bacterial abundances and the genotypes of rs671 by linear regression analysis with the following formula; normalized abundance of the bacterial abundance ~ rs671 genotype (dosage of the A allele) + age + sex + phenotype + dataset + total sequencing length. The significance of the associations was evaluated by Wald's test for the effect size of the rs671 genotype. In the sub-analysis without disease samples, we performed linear regression analysis with the following formula; normalized abundance of the bacterial abundance ~ rs671 genotype (dosage of the A allele) + age + sex + dataset + total sequencing length.

In the MR analysis for the five food-related bacteria, we used the result of the previous dietary habits GWAS in the Japanese population.[37] Since dairy (milk and yoghurt) and natto had genome-wide association ($p < 5 \times 10^{-8}$) only with the rs671 (effect size = 0.113 and $p = 6.4 \times 10^{-18}$ for milk; effect size = 0.113 and $p = 6.0 \times 10^{-21}$ for yoghurt; effect size = -0.114 and $p = 2.7 \times 10^{-24}$ for natto), we performed Wald's test as implemented in the TwoSampleMR package.[38]

Reconstruction of viral genomes

The assembled contigs longer than 5kbp were used for the detection of viral genomes by VirSorter[39] (version 1.0.6) and VirFinder[40] (version 1.1). VirSorter was performed using Viromes (–db 2) databases, and sequences sorted as viruses with the "most confident" prediction (category 1, 4) or "likely" prediction (category 2, 5) were extracted for further analysis. Contigs with the VirFinder score of ≥ 0.9 and $p < 0.01$ were also extracted for further analysis. We applied CheckV[41] (software version 0.7.0, database version 1.0) to all the viral sequences to estimate the completeness of the viral genomes and remove the flanking host regions on the assembled prophages. Subsequently, we checked the number of the viral genes and host genes based on the CheckV annotations. We extracted 31,395 viral genomes of which genome completeness >50% and the number of viral genes > the number of host genes for further analyses.

Clustering and taxonomic annotation of the viral genomes

The 31,395 viral genomes were clustered into species-level vOTUs at the 95% ANI and 85% alignment fraction of the shorter sequence as previously described.[20] We performed all vs all blast using the blastn function in the blast+[89] (version 2.5.0) with the '–max_target_seqs 10000′ option and the result were subjected to the greedy clustering with the previously published custom scripts.[20] After the clustering, we obtained 12,213 species-level vOTUs. Same clustering procedures were performed for the viral genomes with completeness >50% in the GPD and MGV.

We extracted all the representative viral genomes from the JVD, GPD, and MGV and they were merged with the RefSeq viral genomes and previously published crAss-like phage genomes (taxonomic reference genomes) for subsequent clustering and taxonomic annotation. The merged viral genomes were clustered into species-level vOTU as described above and resulted in 94,714 species-level vOTUs. We extracted representative genomes from each of the species-level vOTUs and clustered them into family- and genus-level vOTUs based on the gene sharing ratio and AAI as previously described.[20] The 94,714 viral genomes were subjected to prodigal[92] (version 2.6.3) with the '-p meta' option. Then all vs all blastp search by diamond was performed with the '–max_target_ seqs 10000 –evalue 1e-5' options. Then, pairwise gene-sharing and AAI were calculated for all the pairs of the viral genomes. For clustering, edges between viral genomes were filtered based on their minimum AAI and gene sharing ratio. We performed Markov clustering by MCL (version 14.137)[83] using the following parameters and thresholds for gene sharing ratio and AAI; inflation factors: 1.1, 1.4, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0; gene sharing ratio: 10, 15, 20, 30; AAI: 10, 15, 20, 25, 30, 40, 50, 60. We then selected the following filtering thresholds and MCL inflation factor that resulted in the highest accuracy[103] for the family- and genus-level annotations of the RefSeq viral genomes and previously published crAss-like phage genomes; genus-level vOTU: $\geq$40% AAI, $\geq$30% gene sharing ratio, inflation factor = 2.0; family-level vOTU: $\geq$25% AAI, $\geq$10% gene sharing ratio, inflation factor = 1.4. In this setting, accuracies were 0.77 for the genus-level vOTU and 0.68 for the family-level vOTU (Table S6).

Using the clustering results, we performed taxonomic annotation of the viral genomes based on the taxonomic information of the reference genomes. First, if the viral genome was clustered into a species-level vOTU which contained taxonomic reference genomes, we annotated the taxonomic information which was concordant among the $\geq$75% of the taxonomic reference genomes and higher than genus-level. Second, the same procedures were repeated based on the genus-level vOTU information for unannotated genomes. Third, unannotated viral genomes clustered into a family-level vOTU which contains $\geq$2 taxonomic reference genomes, we annotated the taxonomic information which was concordant among the $\geq$75% of the taxonomic reference genomes and higher than family-level. After the taxonomic annotation procedures, we annotated family-level taxonomy for the 8,873 of 31,395 newly reconstructed viral genomes. The 9,167 GPD genomes and 43,817 MGV genomes were taxonomically annotated at the family level both in the original studies and this study. Although the recent taxonomic modification of the Podoviridae lowered the overall family-level taxonomic concordance between the original study and this study (73.1% for the GPD and 83.9% for the MGV), high family-level taxonomic concordance was observed for viruses that were not annotated as Podoviridae in the previous studies (95.7% for the GPD and 97.9% for the MGV).

Functional analysis of the viral genomes

Protein-coding genes for each of the viral genomes were predicted by prodigal with the '-p meta' option. Predicted proteins were subjected to the eggNOG-mapper and hmmscan function in the hmmer against the VOG database[44] (E-value $< 1 \times 10^{-5}$) for the annotation of the KEGG and VOG and calculation of the database coverage ratio and functional annotation ratio. The database coverage ratio was defined as the ratio of the protein sequences which were assigned with any eggNOG or VOG hits including unknown functions. The functional annotation ratio was defined as the ratio of the protein sequences which were assigned with either COG annotations other than S (Function unknown) and R (General function prediction only) or VOG annotations other than Xu (Function unknown). AMGs among the KEGG genes were defined based on the previously published manually curated list of the AMGs.[46]

Predicted protein sequences on the viral genomes from the JVD, MGV, and GPD were dereplicated with MMseqs2 with the following parameters; –cov-mode 1 -c 0.8 –kmer-per-seq 80 –min-seq-id 1. The Dereplicated set of the protein sequences from the JVD, MGV, and GPD were then merged and subjected to further clustering with the following parameters; –cov-mode 1 -c 0.8 –kmer-per-seq 80. The '–min-seq-id' option in the second clustering was set at 1, 0.95, 0.9, and 0.5 to dereplicate the protein sequences at 100%, 95%, 90%, and 50% amino acid sequence identity, respectively.

Subfamily-level annotation of the crAss-like phages

We extracted the 1,378 putative crAss-like phage genomes which represented the species-level vOTUs. We annotate the subfamily of the crAss-like phage genomes based on the major annotation of the taxonomic reference genomes which were co-clustered into the same genus-level vOTUs. We checked the validity of the subfamily of crAss-like phages by constructing the maximum-likelihood phylogenetic trees for TerL, a previously reported marker gene.[42,104] First, we identified the TerL from the predicted protein sequences on the crAss-like phage genomes by the hmmsearch function in hmmer against previously constructed hmm profiles.[42] We then extracted the significant hits (E-value < 0.05) and constructed MSAs by MAFFT with the default parameter. The detection ratio of the TerL was 87.9%. We constructed phylogenetic trees from the MSAs by iqtree, and results were visualized by iTOL. 'LG + R9' model was selected as the best model by the ModelFinder.

Interpopulational comparisons of the crAss-like phages

For the interpopulational comparisons based on the number of the viral genomes, we utilized the viral genomes from the JVD and MGV. The geographical origin of the viral genomes from the MGV was defined in the original study. Fisher's exact tests were performed for the contingency tables made from the following four numbers; (i) number of the β crAss-like phage genomes from the target population, (ii) number of the non-β crAss-like phage genomes from the target population, (iii) number of the β crAss-like phage genomes from the reference population, and (iv) number of the non-β crAss-like phage genomes from the reference population. We performed 21 tests in total.

For the comparison of the crAss-like phage abundances, only the HC samples were used. The 94,714 representative genomes of the species-level vOTUs were indexed with bowtie2. Quality-controlled reads were down-sampled to 1,000,000 paired-ends reads to adjust the differences of the library sizes between the datasets and mapped to the reference genome with bowtie2. Abundances were calculated as Reads Per Kilobase of exon per Million mapped reads (RPKM) by coverM and summed up for each subfamily and genus-level vOTU of the crAss-like phages. Then the compositions among the crAss-like phages were calculated for each sample and averaged over the samples from the same groups.

Case–control comparisons of the crAss-like phages

Samples with the usage of the antibiotics were removed from case–control comparisons (Table S9). Quality-controlled reads were mapped to the reference genome with bowtie2. Abundances were calculated as mean coverage of each viral genome by coverM genome function, divided by 'total sequencing length/1,000,000,000', and summed up for each subfamily and genus-level vOTU of the crAss-like phages. Only the clades which satisfied the following three criteria were retained and subjected to the log transformation for subsequent analyses; (i) detected in >20% of samples used for case–control comparison (ii) detected in both case and control samples (iii) adjusted mean coverage ≥ 0.001. We evaluated the association between the crAss-like phages and the diseases (RA, SLE, MS, UC, CD, and CoCa) by logistic regression analysis with the following formula; disease state $\sim$ crAss-like phage abundance + age + sex + dataset + total sequencing length. The significance of the associations was evaluated by Wald's test for the effect size of the crAss-like phage abundance.

As for the tested clades, we also performed the association analyses with the Shannon index. Quality-controlled reads were down-sampled to 1,000,000 paired-ends reads to adjust the differences of the library sizes between the datasets. Then, the down-sampled reads were mapped to the reference genome of the reconstructed 1,273 MAGs with bowtie2 and the mean coverage matrix was calculated with coverM. The resulting matrix of the mean coverage was subjected to the diversity function in the R package vegan (version 2.5_6) to calculate the Shannon index. We evaluated the association between the crAss-like phages and Shannon index by linear regression analysis with the following formula; Shannon index $\sim$ crAss-like phage abundance + age + sex + phenotype + dataset + total sequencing length. The significance of the associations was evaluated by Wald's test for the effect size of the crAss-like phage abundances.

Virus–prokaryote association analysis based on the CRISPR and prophages

We predicted the CRISPR sequences on the reconstructed MAGs with MinCED[85] (version 0.4.2). Spacers within the predicted CRISPR sequences were queried against the viral contigs recovered from the gut metagenome data. Since the MGV and GPD have significant overlap, we performed MGV vs GPD blastn search and dereplicated at 100% ANI over 100% aligned fraction of the shorter sequences. Blast hits of the spacers with >95% ANI, end-to-end alignment, and spacer coverage >95% were retained for further analysis. For each spacer, we extracted all the blast hits with the highest bit-score, and if the phylum-, class-, order-, family-, genus-, or species-level taxonomy were consistent among more than a half of the blast subjects, taxonomic information of the target viruses was annotated to the spacer. If the taxonomy of the target of the spacer could not be determined, an ''ambiguous'' label was assigned. The same procedures were repeated for the CRISPR sequences identified within the UHGG in the previous study.[20]

In addition to the CRISPR sequences, we utilized prophage information to identify virus–prokaryote interaction. Among the proviral contigs determined by CheckV, we extracted contigs of which >50% were covered with host's sequences.

Virus–prokaryote association analysis based on the abundance

Samples with the usage of antibiotics and insufficient clinical information were removed from this analysis as done in the association analysis between the crAss-like phages and Shannon index. Quality-controlled reads were mapped to the reference genomes of the 1,273 JMAG genomes, 4,644 UHGG genomes, and 94,714 representative genomes of the species-level vOTUs with bowtie2, respectively. The mean coverage of each genome was calculated by the coverM genome function and divided by 'total sequencing length/1,000,000,000'.

We extracted the viruses and prokaryotes which were conferred to participate in the virus–prokaryote interaction from the JMAG CRISPR analysis, UHGG CRISPR analysis, and JMAG prophage analysis, respectively. Then, we retrieved the abundance of the viruses and prokaryotes which passed the following criteria; (i) detected in >20% of samples (ii) detected in all the datasets (iii) adjusted mean coverage ≥ 0.001. Abundances of the viruses and prokaryotes were subjected to the log transformation. We evaluated the association between the viruses and the prokaryotes by linear regression analysis with the following formula; prokaryotic abundance $\sim$ viral abundance + age + sex + phenotype + dataset + total sequencing length. The significance of the associations was evaluated by Wald's test for the effect size of the viral abundance.

Comparison of the viral and prokaryotic numbers and abundances between Japanese and other populations

For each species-level vOTU, we counted the number of viruses derived from Japanese and other populations from the JVD and MGV. Then we defined the odds ratio for being Japanese-derived as follows; (number of the Japanese-derived viruses belonging to the vOTU/number of the other viruses belonging to the vOTU)/(number of the Japanese-derived viruses not belonging to the vOTU/number of the other viruses not belonging to the vOTU). For each species-level cluster of the JMAG genome which have the corresponding cluster in the UHGG, we merged the JMAG and UHGG clusters and defined the odds ratio for being Japanese-derived as follows; (number of the Japanese-derived bacterial genomes belonging to the cluster/number of the other bacterial genomes belonging to the cluster)/(number of the Japanese-derived bacterial genomes not belonging to the cluster/number of the other bacterial genomes not belonging to the cluster).

We extracted species-level vOTU–prokaryotic cluster pairs which were detected in the JMAG CRISPR analysis. Among the pairs, we retained only pairs whose species-level vOTU were included in both the JVD and MGV and the prokaryotic cluster was present in both the JMAG and UHGG. Then, we evaluated the enrichment of the virus–bacteria pairs which had same sign of the log odds ratios for being Japanese-derived based on the Fisher's exact test.

We also checked the differences of the viruses-prokaryotes interaction by read count-based approach. We extracted species-level vOTU–prokaryotic cluster pairs which were detected in the JMAG CRISPR analysis. Quality-controlled reads were down-sampled to 1,000,000 paired-ends reads to adjust the differences of the library sizes between the datasets. Then, the down-sampled reads were mapped to the reference genomes and abundances were calculated as RPKM by coverM for the JMAG and viruses, respectively. For each comparison between Japan and other populations, only the pairs of which viruses and bacteria satisfied the following criteria were considered; (i) detected in >20% of samples (ii) detected in both of the populations (iii) RPKM ≥ 0.001. Fold changes between Japan and other populations were calculated for each species-level vOTU and prokaryotic cluster. Then, we evaluated the enrichment of the virus–bacteria pairs which had the same sign of the log fold-changes between the abundances in Japanese and other populations based on the Fisher's exact test.

Network analysis of the crAss-like phages

We extracted all the CRISPR spacers in the JMAG and UHGG genomes which supported the link between the crAss-like phage species-level vOTUs and MAGs. Then, we counted the number of the spacers which supported the link between the crAss-like phage species-level vOTUs and bacterial genus for the JMAG and UHGG, respectively. We constructed a network plot with the ggraph package (version 2.0.4). We specified 'kk' as a layout option to place nodes based on the spring-based algorithm by Kamada and Kawai. The bar plot was based on the combined number of the spacers in the JMAG and UHGG.

QUANTIFICATION AND STATISTICAL ANALYSIS

Please refer to figure legends and method details for details of statistical analysis. Number of the samples used in the analyses are described in Tables S1, S3, and S9.

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

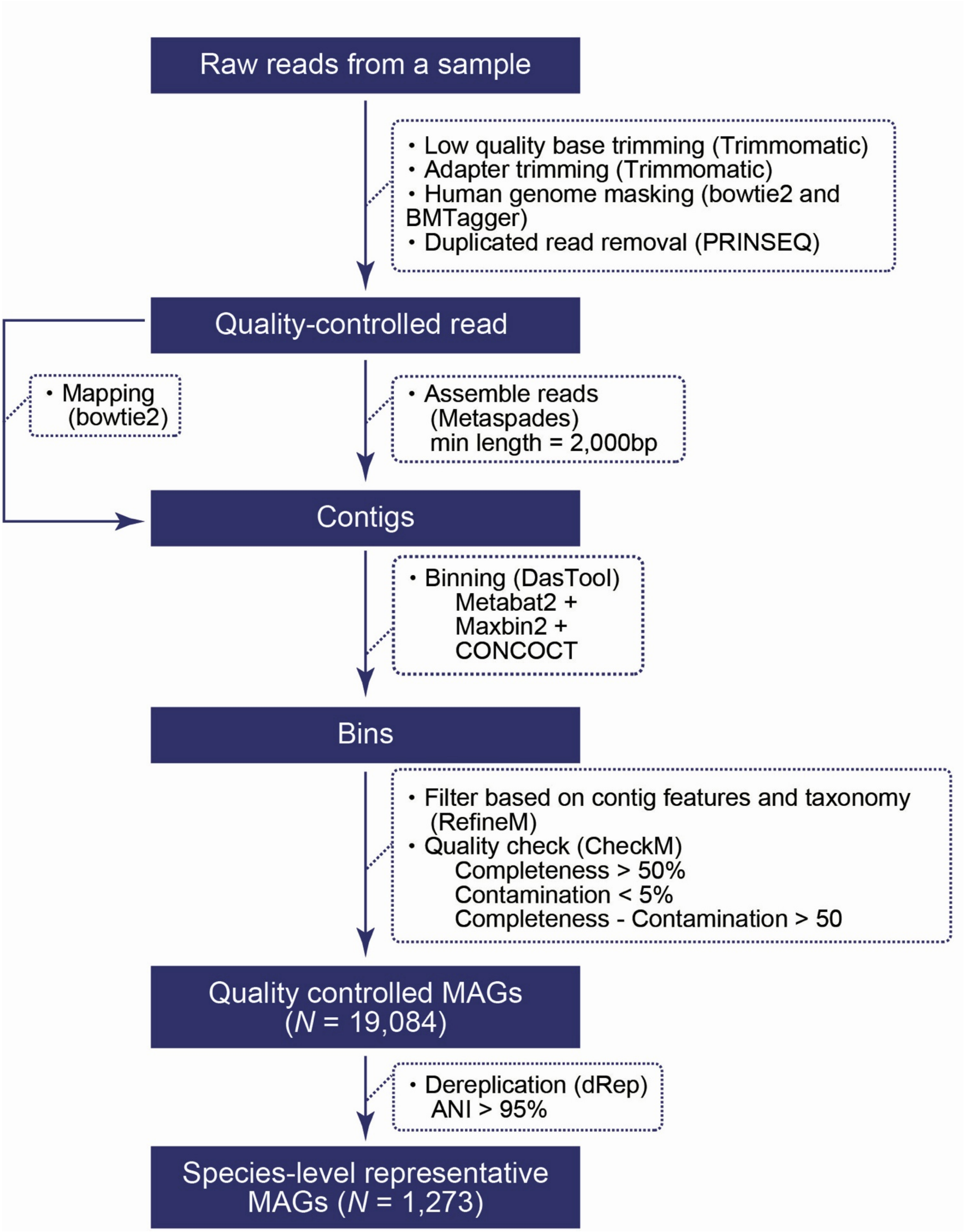

Raw reads from a sample
• Low quality base trimming (Trimmomatic)
• Adapter trimming (Trimmomatic)
• Human genome masking (bowtie2 and BMTagger)
• Duplicated read removal (PRINSEQ)
Quality-controlled read
• Mapping (bowtie2)
• Assemble reads (Metaspades) min length = 2,000bp
Contigs
• Binning (DasTool) 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.

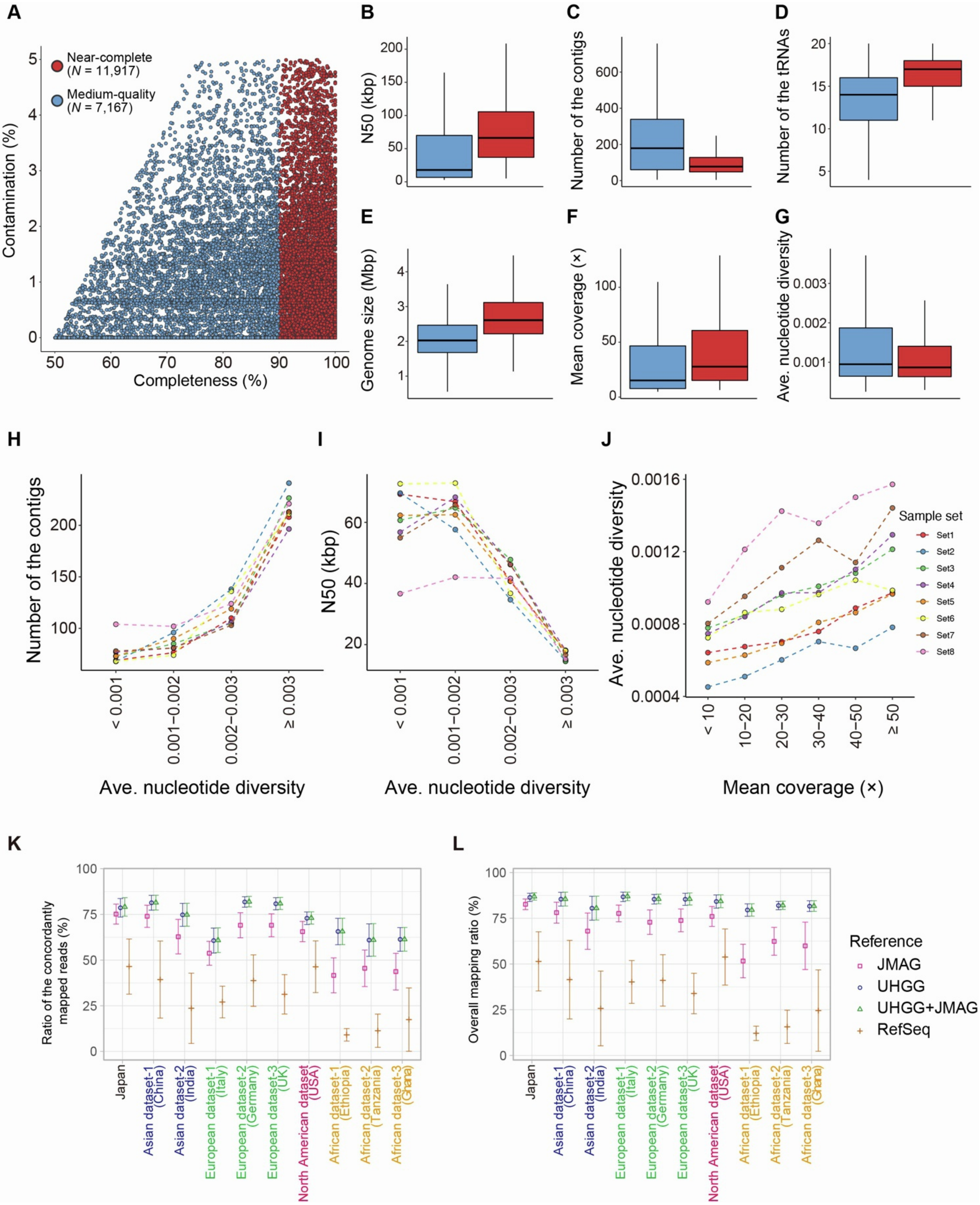


Figure S2. Metrics related to the quality of the MAGs, related to Figure 1

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.

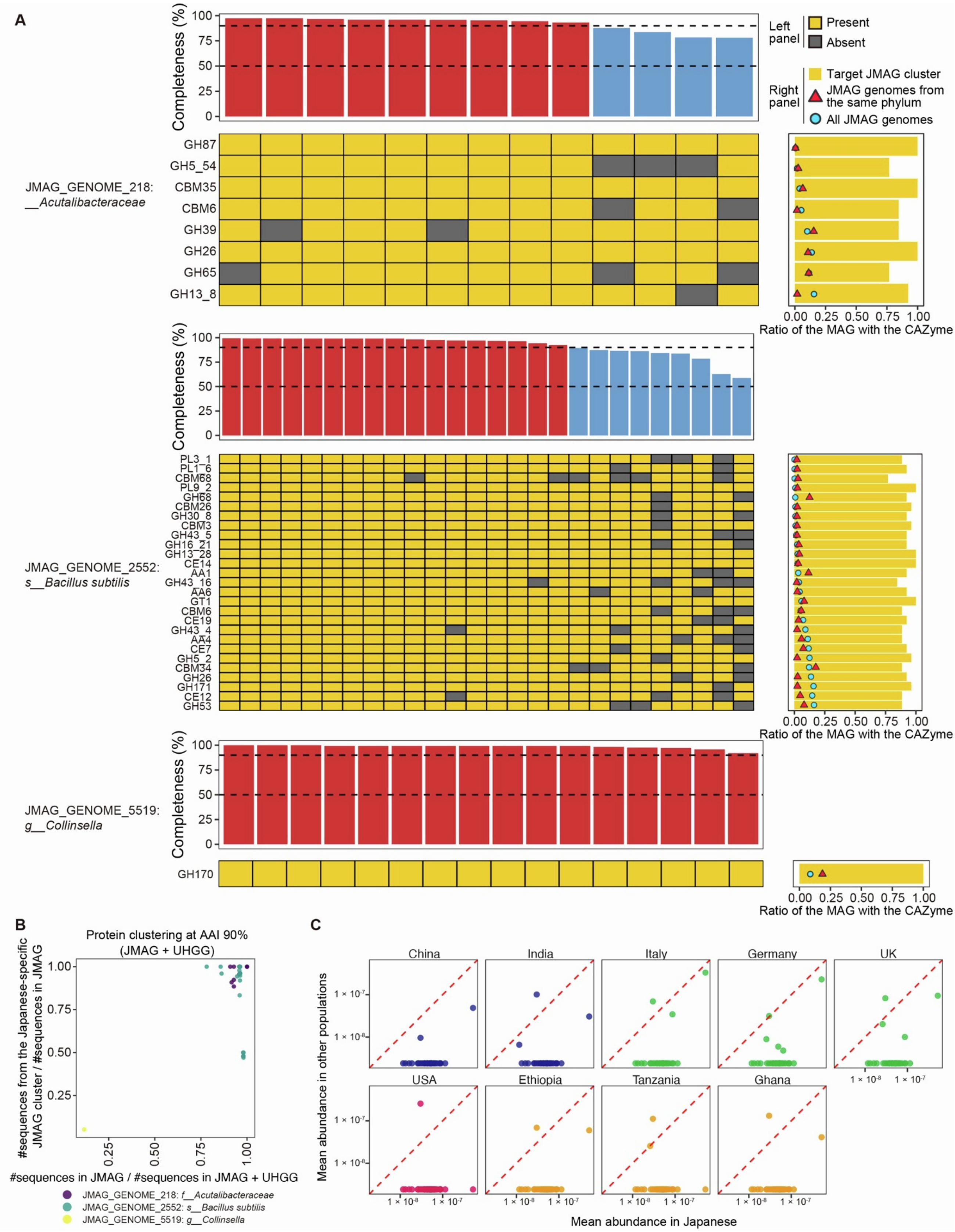


Figure S3. CAZyme profiles of the Japanese-specific species-level clusters, related to Figure 1

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.

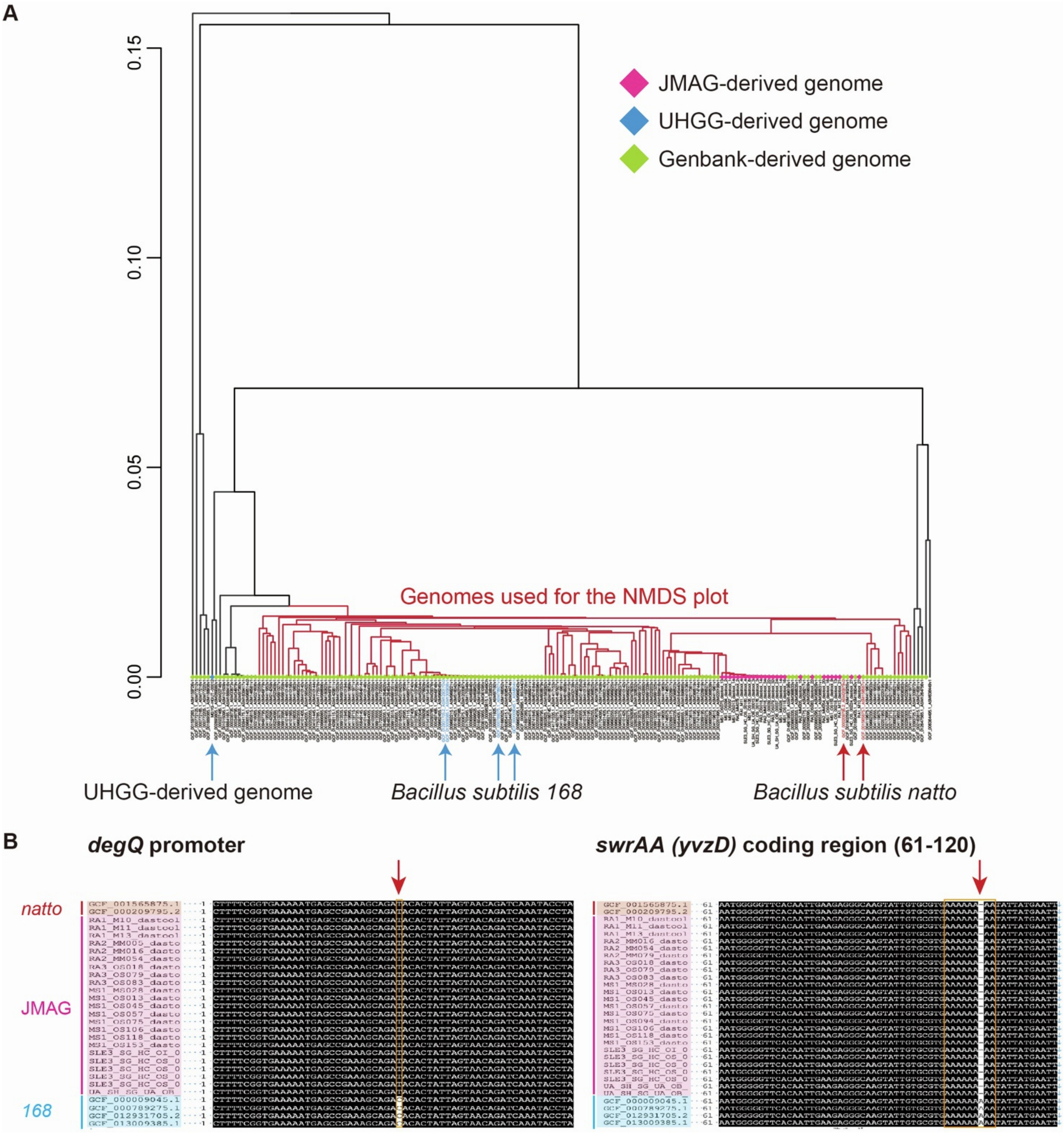


Figure S4. Comparison of the *Bacillus subtilis* genomes, related to Figure 1

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.

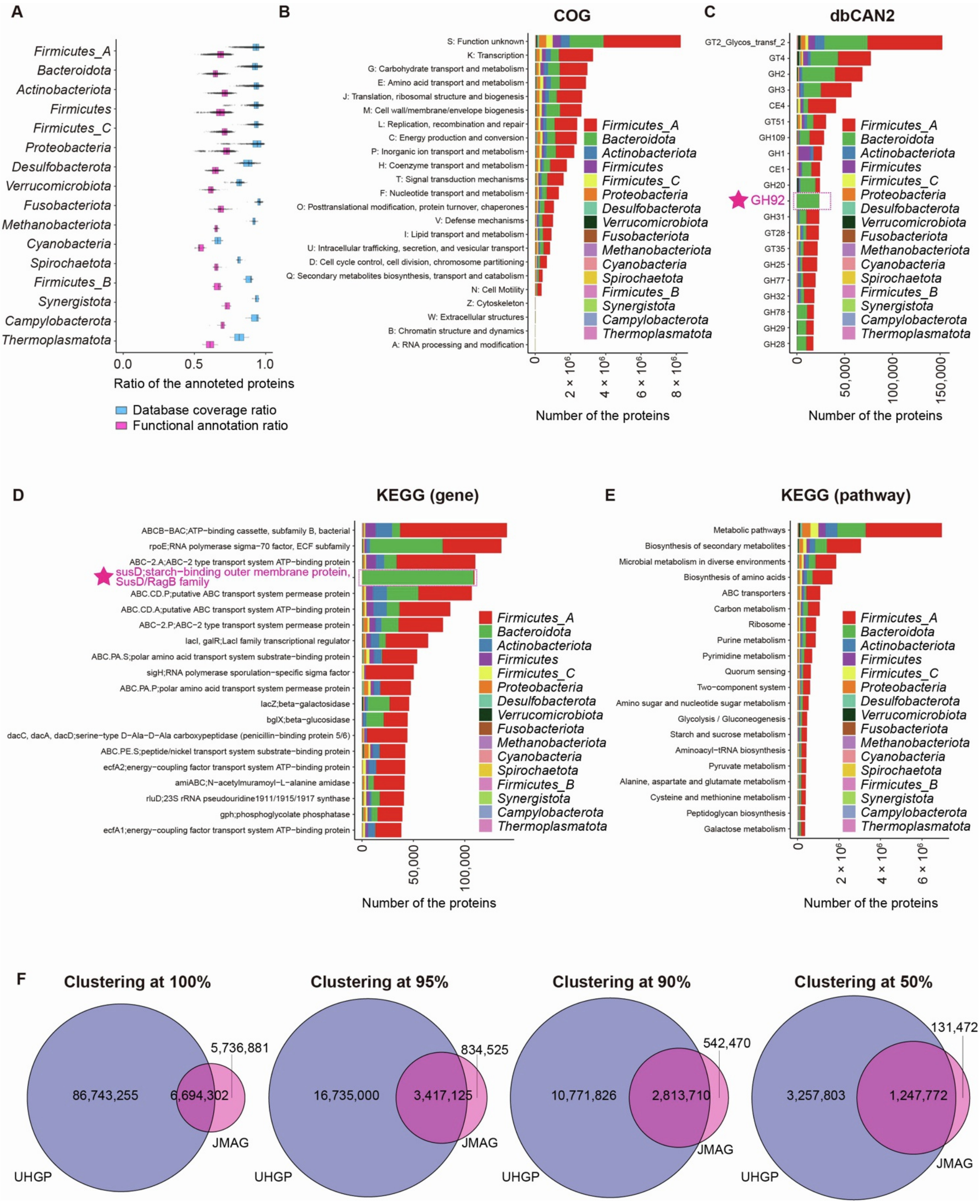


Figure S5. Annotation of the protein sequences in the JMAG, related to Figure 2

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.

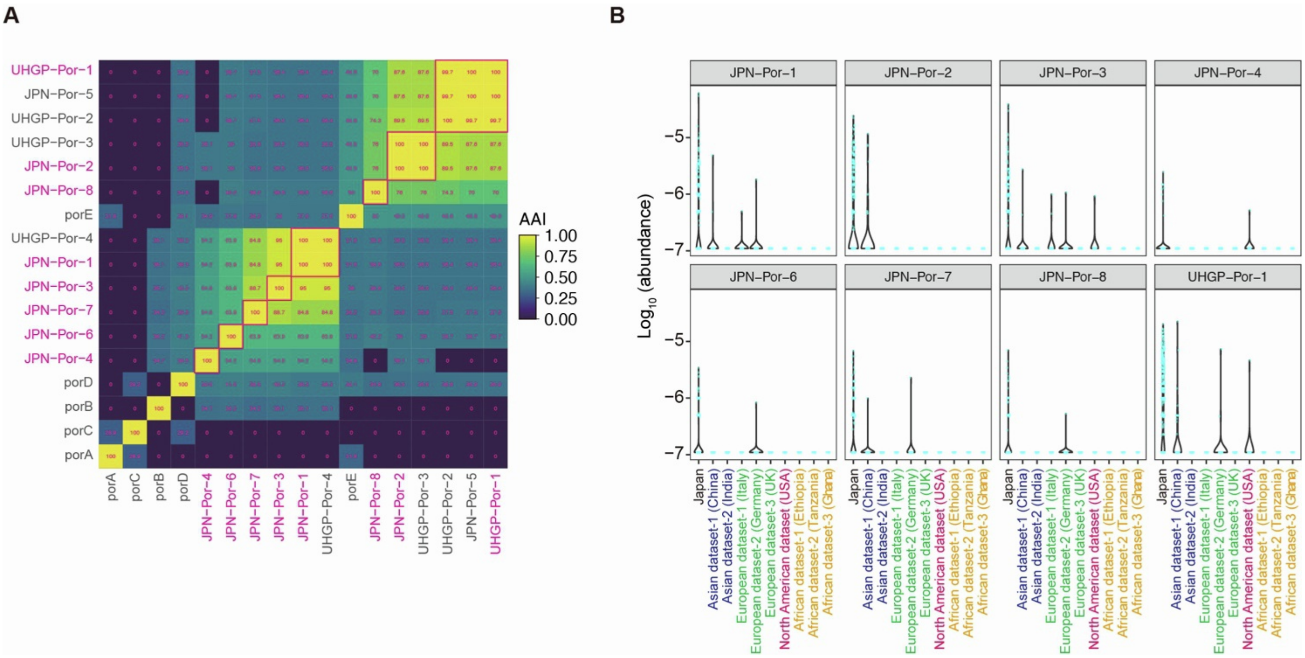


Figure S6. Evaluation of the β-porphyranase sequences in the JMAG and UHGP, related to Figure 2

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.

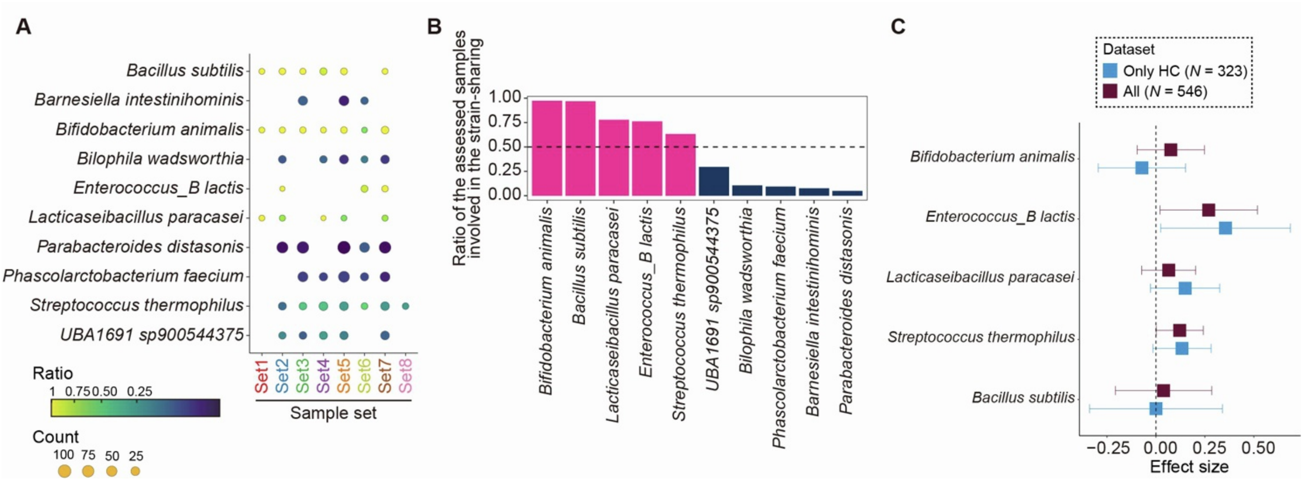
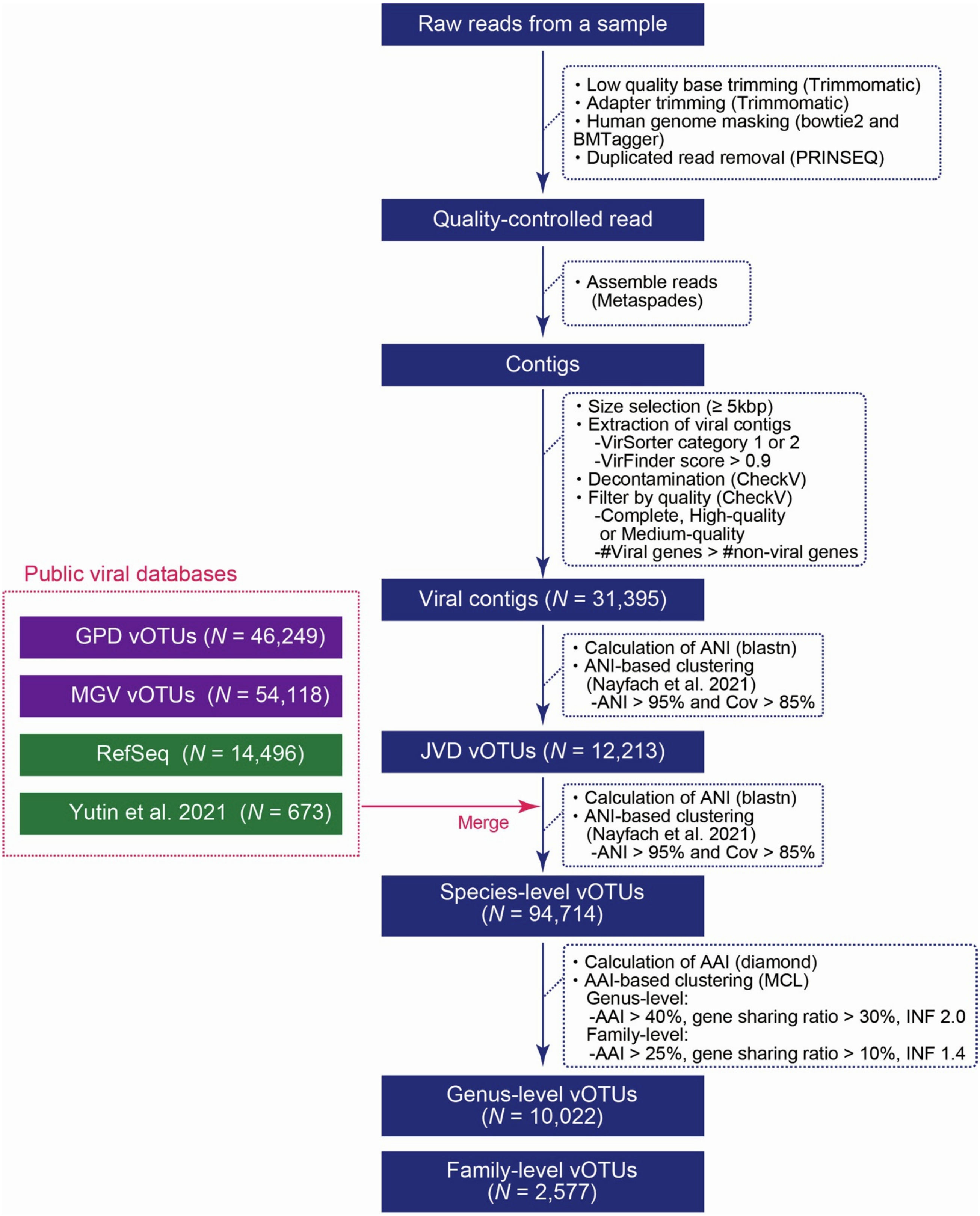


Figure S7. Sharing of the bacterial strains among the 787 Japanese people, related to Figure 3

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
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)
Public viral databases
GPD vOTUs (N = 46,249)
MGV vOTUs (N = 54,118)
RefSeq (N = 14,496)
Yutin et al. 2021 (N = 673)
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.

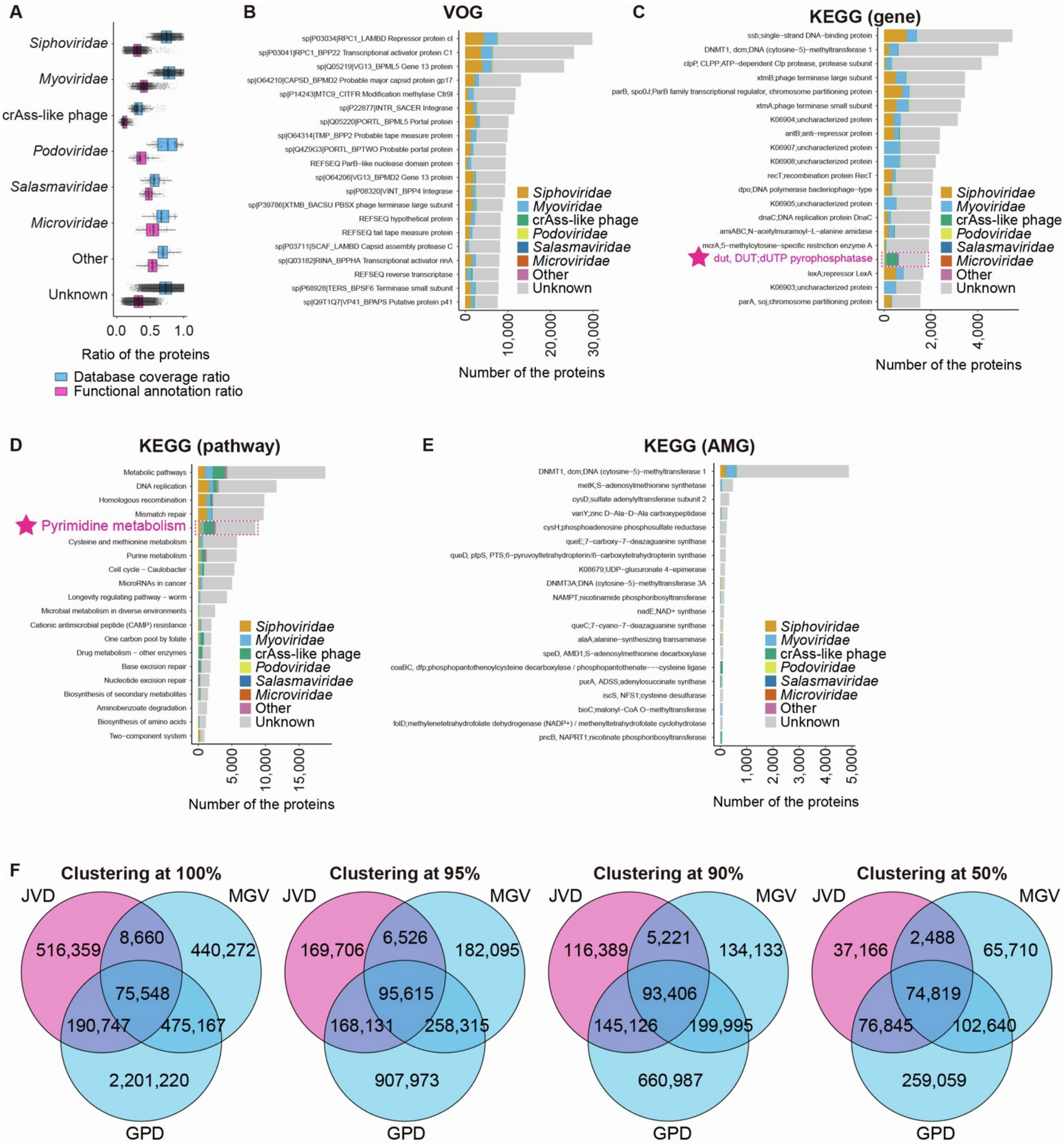


Figure S9. Annotation of the protein sequences in the JVD, related to Figure 4

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.

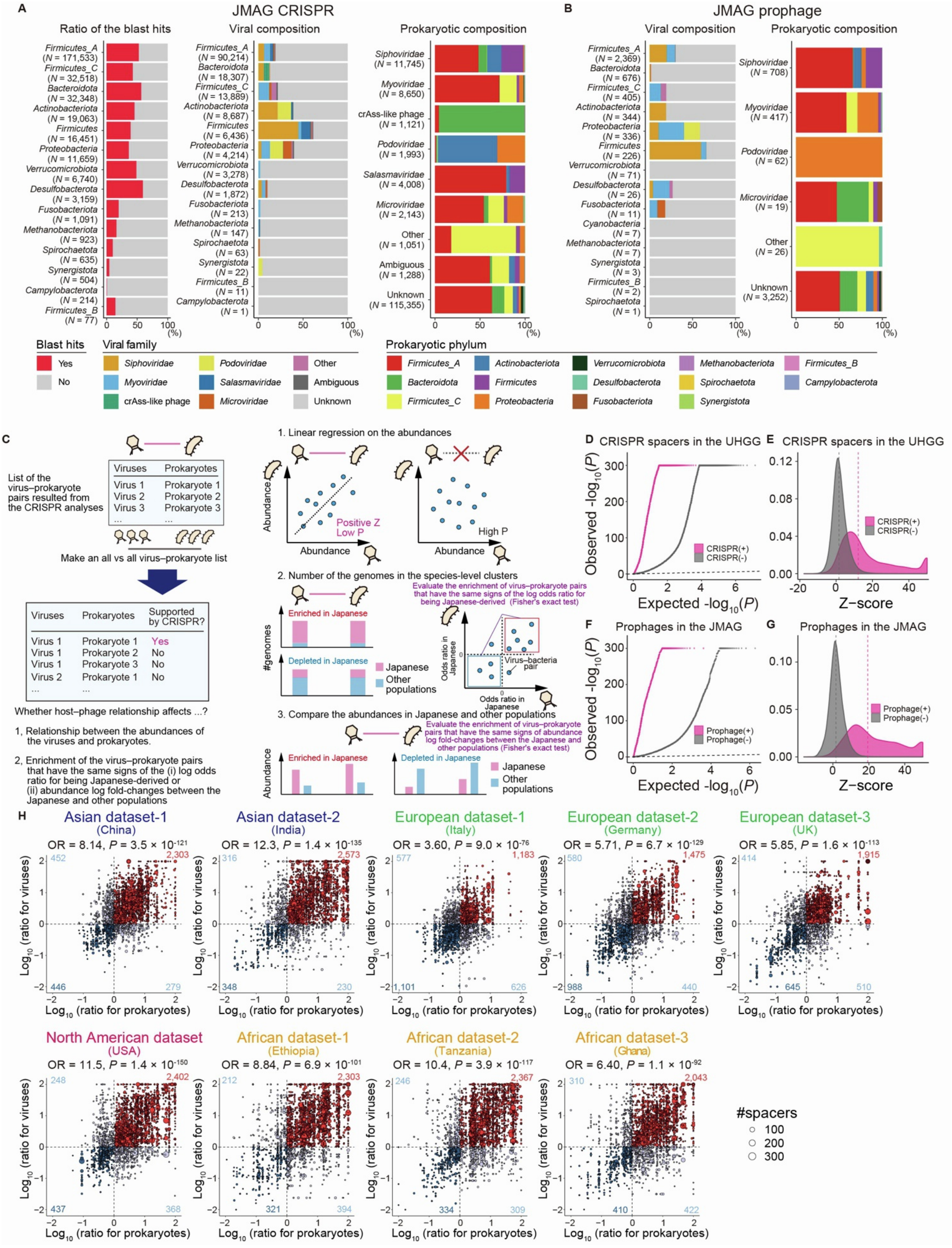


Figure S10. The taxonomic and populational analysis of the virus–prokaryote interaction, related to Figure 6

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**).