Supplementary Text

Metagenomic sequencing of fecal DNA

Sequencing of fecal DNA samples was performed using 454 GS (Roche), Ion PGM/Proton (Life Technologies), and MiSeq (Illumina) according to the suppliers' protocols. For 454 GS, 5 μg of fecal DNA was sheared to obtain fragments ranging from 300 to 700 bp for the FLX Titanium platform and 500 to 1,000 bp for the FLX+ platform. The libraries were prepared using the GS FLX Titanium Rapid Library MID Adaptors Kit. For Ion PGM/Proton, 100 ng of fecal DNA was sheared to obtain fragments ranging from 350 to 470 bp and the library was prepared using the Ion Xpress Plus Fragment Library Kit. For the 454 and Ion PGM/Proton reads, artificially redundant reads were removed using a replicate filter if any sequences had ≥ 95% identity to other sequences with exactly the same starting point. Reads mapped to the human genome (HG19) with Newbler (version 2.7) were also removed. Finally, reads with an average Quality Value (QV) less than 20 or less than 75 bp in length were removed. For 150 bp paired-end sequencing of MiSeq, 20 ng of fecal DNA was sheared to obtain fragments ranging from 300 to 400 bp and the library was prepared using the TruSeq DNA Sample Prep Kit. For 300 bp paired-end sequencing by MiSeq, fecal DNA library was prepared using the Nextera DNA Sample Prep Kit. Any 5′ low quality (< 20 QV) bases in MiSeq reads were trimmed off. Reads having bases less than 20 QV for more than half of the read length and reads whose length was less than 50 bp were also filtered out. These procedures were done using the FASTX-Tool kit (http://hannonlab.cshl.edu/fastx_toolkit/). The filter-passed reads were then mapped to the human and phiX genomes using bowtie2 (version 2.2.1) and any mapped reads were removed. Sequencing statistics are summarized in the Supplementary Table S1. All bacterial metagenomic sequences were deposited in DDBJ under the accession number PRJDB3601.

Construction of microbial reference genomes

The microbial reference genome database was constructed as follows. First, genomes matching with either of the following criterion were selected as references; (i) genomes mapped with ≥ 10 metagenomic reads when 60 million metagenomic reads from 6 countries (Japan, China, Denmark, Spain, Sweden, and the United States) were mapped to the 25,085 genomes in GenBank/DDBJ/EBI using BLASTN with a 95% identity and 90% length coverage cut-off. (ii) genomes for which 16S rRNA gene sequences, identified with RNAmmer[48], were mapped with ≥ 10 in-house 16S rRNA V1-V2 region sequences of human microbes, which comprised of about 600 thousand reads, using BLASTN with a 85% identity and 90% length coverage cut-off. Second, typical known pathogenic species (*Bacillus anthracis*, *Bordetella pertussis, Burkholderia pseudomallei, Campylobacter coli, Campylobacter jejuni, Clostridium botulinum*, *C. chauvoei*, *C. tetani, Corynebacterium diphtheria,* *Francisella tularensis, Leptospira interrogans, Listeria monocytogenes, Mycobacterium abscessus, M. tuberculosis*, *Salmonella enterica, Shigella boydii, S. dysenteriae, S. flexneri, S. sonnei, Vibrio cholera, V. vulnificus,* and *Yersinia pestis*) and four genera of *Borrelia*, *Chlamydia*, *Mycoplasma*, and *Rickettsia* were excluded from the reference dataset. Since these pathogens are temporarily detected in the patients with infectious diseases, exclusion of them could improve the accuracy of taxonomic assignment of metagenomic reads from healthy individuals. Third, to reduce complexity and excess load in computing, for the species with ≥ 50 sequenced genomes at the strain level, some of them were excluded to the extent that the genomes still covered ≥ 99% of the total reads mapped to the species. These species included *Acinetobacter baumannii*, *Bacillus cereus*, *Bacteroides fragilis*, *Enterococcus faecalis*, *E. faecium*, *Escherichia coli*, *Helicobacter pylori*, *Klebsiella pneumonia*, *Peptoclostridium difficile*, *Propionibacterium acnes*, *Pseudomonas aeruginosa*, *Staphylococcus aureus*, *S. epidermidis*, *Streptococcus agalactiae*, *S. mutans*, *S. pneumonia*, *S. pyogenes*, and *S. suis*.

To reduce the species-level complexity in the database, the 16S rRNA gene sequences of the reference genomes were further clustered with a 98.8% identity cut-off[49], and the generated clusters were defined as single species. The numbers of mapped reads to the genomes in the same cluster were merged and their abundances were represented by the representative species of the cluster. For a few clusters such as *Streptococcus salivarius* and *Streptococcus thermophiles*, both of which have 16S rRNA gene sequences of ≥ 98.8% identity*,* we manually separated these clusters into different species. Of a few species, such as *Fusobacteroium nucleatum* and certain *E. coli* strains, that formed distinct clusters, even when the species' names were identical, the species/clusters were merged when a sufficient number of multi-hit reads were commonly shared among them. Several draft genomes lacking 16S rRNA gene sequences were assigned to the most similar species or clusters when the species' names were related and multi-hit reads were commonly shared among the genomes. Finally, we obtained the reference genome database comprised of a total of 6,149 genomes representing 2,373 clusters at the species level of Bacteria and Archaea (Supplementary Table S3).

PCR detection of *Methanobrevibacter smithii* in the Japanese individuals

M. smithii was detected by PCR using *M. smithii* 16S rRNA gene-specific primers 5′-ATGCACCTCCTCTCAGCTAGTC-3′ and 5′-AGAGGTACTCCCAGGGTAGAGG-3′. The primer sets were designed using Primer3[50]. PCR was conducted in 10.0 µL PCR solution containing 0.2 µL of template DNA, 0.02 µL of each primer, 1.0 µL of 10 × PCR buffer, 1.0 µL of dNTP mixture, 0.04 µL of Ex Taq polymerase (Takara Bio Inc., Otsu, Japan) and 7.52 µL of ddH_2O using GeneAmp PCR System 9700 (Applied Biosystems, Tokyo, Japan) with 40 cycles of denaturation (30 sec at 96°C), annealing (20 sec at 60°C), and elongation (3 min at 70°C). The PCR products were separated on 1.5% of agarose gels with a positive control from genomic DNA of *M. smithii* JCM 30028^T. PCR without DNA was also performed as negative control. Genomic DNA of *M. smithii* JCM 30028^T was obtained from Japan Collection of Microorganisms, RIKEN BRC.

Mapping of metagenomic reads to genomes and genes

One million (M) metagenomic reads per individual were mapped to the reference genomes using Bowtie2 with a 95% identity cut-off. For the SOLiD reads from RU, Bowtie[51] (version 0.12.7) was employed with the same threshold as previously. For several samples of which the number of reads was < 1 M, all of the reads for the individual were mapped to the reference genomes. In the 861 individuals selected, the minimum number of reads per individual was about 60,000, but Pearson's correlation coefficients (PCCs) between microbial compositions obtained from the mapping of 1 M and 60,000 reads from several same JP individuals was > 0.999, indicating that the number of reads per individual between 60,000 and 1 M did not significantly affect the results obtained from the mapping analysis in this study. The number of multi-hit reads that mapped to several different genomes with equal scores were divided among those genomes in proportion to the number of reads uniquely mapped to each genome. For genome *g*, we defined the abundance π_g as follows,

$$\pi_g = \frac{U_g + \sum_{r \in GtoR(g)} P_{r,g}}{l_g}$$

where U_g is the number of reads that are uniquely mapped to genome *g*, *GtoR(g)* is the set of reads that are equally mapped to several genomes including genome g, and l_g is the length of *g*. $P_{r,g}$ is the probability that a read *r* is assigned to genome *g*, and is calculated as follows,

$$P_{r,g} = \frac{U_g}{\sum_{g' \epsilon RtoG(r)} U_{g'}}$$

where *RtoG(r)* is a set of genomes to which a read *r* mapped. Approximately 66 % of the total reads from the 861 healthy individuals were mapped to the present reference genomes. NCBI taxonomy information was used for taxonomic assignment of phylum, genus, and species for each genome. Genomes that were not assigned to a particular rank were assigned to the higher taxonomic rank and designated 'unclassified higher class'. Mapping of metagenomic reads to gene sets was also performed under the same conditions as used for the reference genomes described above.

Assessment and comparison of different methodologies.

To evaluate the effect of different methodologies on the metagenomic analysis, the same fecal samples were subjected to sequencing with different sequencers, different DNA extraction methods, and different fecal sample storage conditions (Supplementary Table S5). The microbial compositions were calculated with the analytical method described above. Similarity of the microbial compositions at the genus level was evaluated using Pearson's correlation coefficient. Permutation test with 10,000 times randomization was conducted to test the statistical significance of the similarity between the data obtained by different methodologies and between individuals within and between countries.

Analysis of dietary intake data

Dietary intake information for 119 food items of the 12 countries was downloaded from the Food and Agriculture Organization Corporate Statistical database (FAOSTAT) (http://faostat3.fao.org/home/, as of June 2015). The averaged dietary intakes (g/capita/day) from 2002 to 2011 in the 12 countries were used for analysis. According to the Standard Tables of Food Composition in Japan, 2010[52], three major nutrient compositions (carbohydrates, lipids, and proteins) were calculated from the averaged dietary intakes of the 119 food items. Based on their nutrient similarities, hierarchical clustering was performed to group the 119 food items into nine food categories, where nutrient quantities were transformed to z-scores before clustering and dendrogram was generated using the Ward method and Spearman's correlation as dissimilarity.

References (continued)

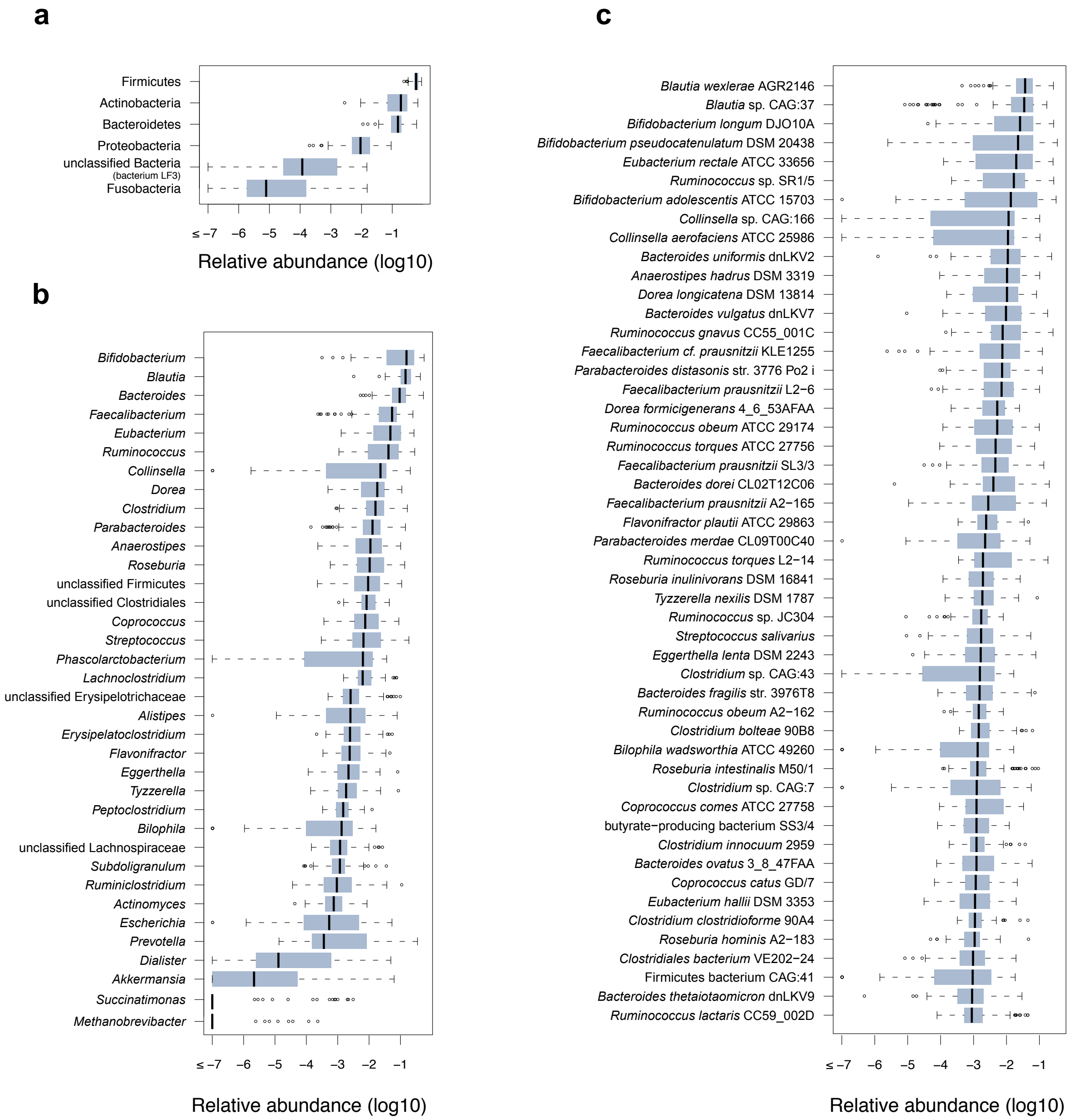
48. Lagesen, K., Hallin, P., Rodland, E. A., Staerfeldt, H. H., Rognes, T. and Ussery, D. W. 2007, RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.*, **35**, 3100-3108.

49. Sunagawa, S., Mende, D. R., Zeller, G., et al. 2013, Metagenomic species profiling using universal phylogenetic marker genes. *Nat. methods*, **10**, 1196-1199.

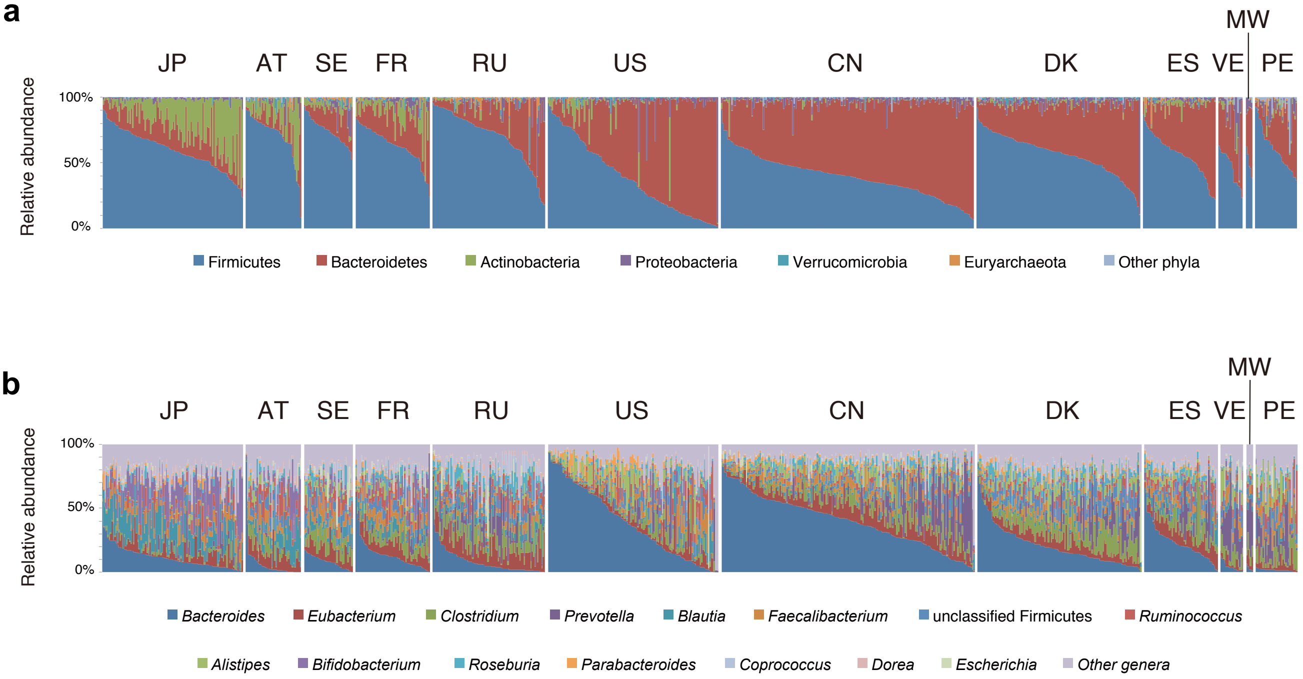
50. Rosen, S. and Skaletsky, H. 2000, Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365-386.

51. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. 2009, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

52. The Subdivision on Resources The Council for Science and Technology, MEXT, Japan. 2010, Standard Tables of Food Composition in Japan 2010 (in Japanese).

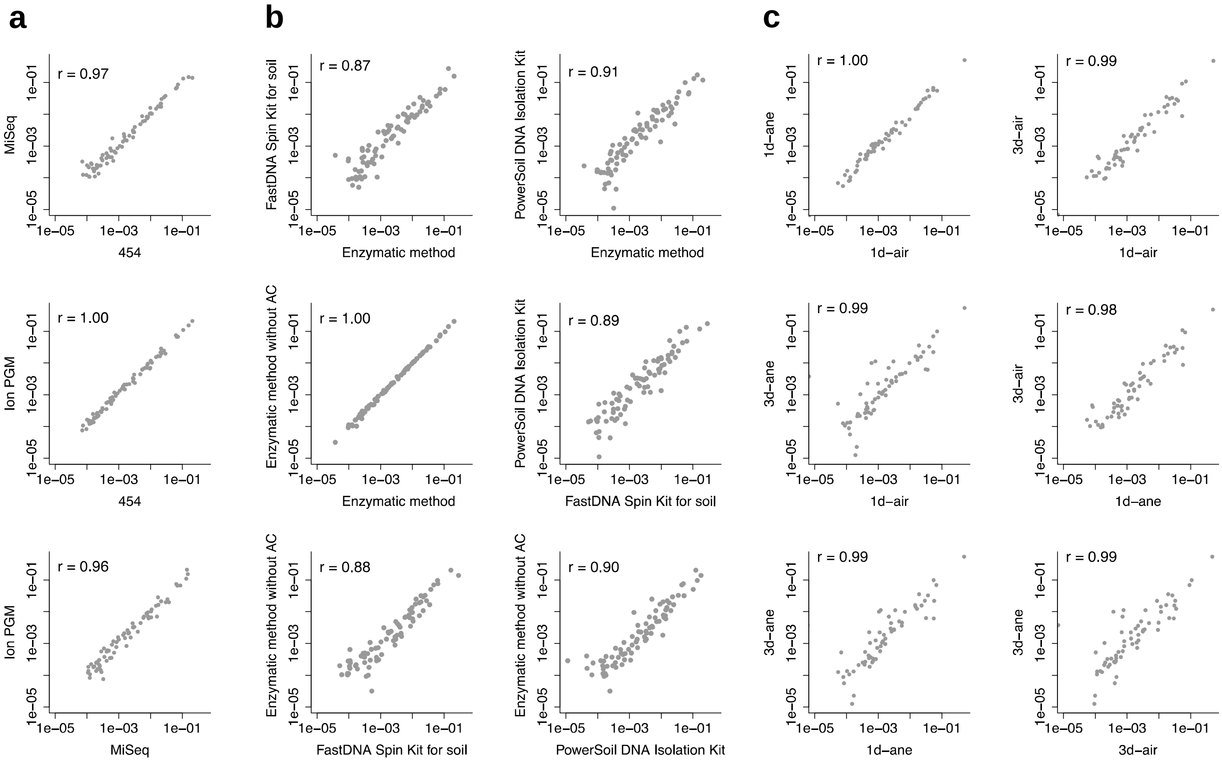


Supplementary Fig. S1. Microbial composition in the 106 JPGM.
The relative abundance of dominant microbes detected in the 106 JPGM is box-plotted at the phylum (a), genus (b) and species levels (c). The horizontal axes represent the log-transformed values of the relative abundance. Boxes represent the interquartile range (IQR) and the lines inside show the median. Whiskers denote the lowest and highest values within 1.5 times the IQR.



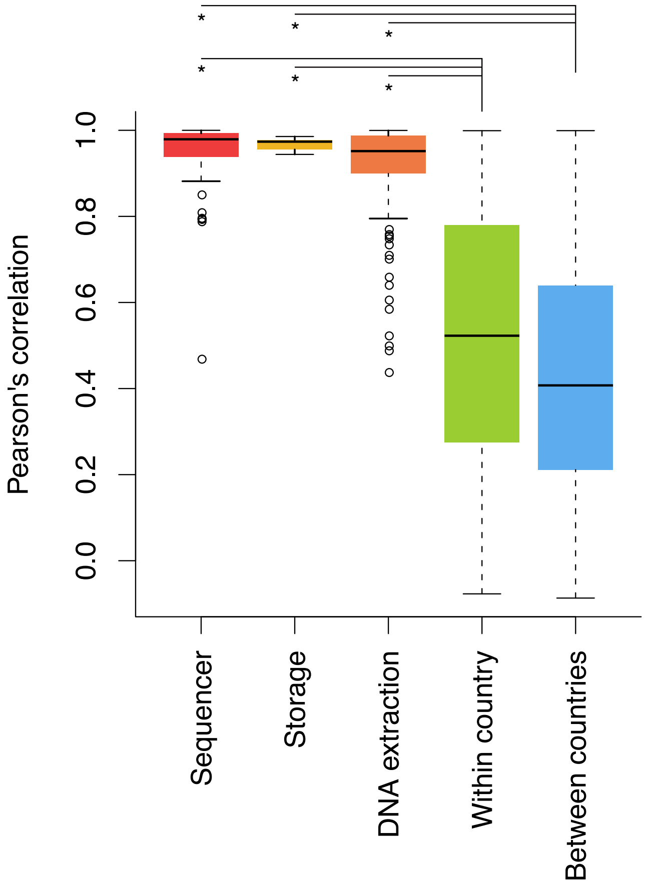
Supplementary Fig. S2. Microbial composition in the gut microbiome of the 861 healthy individuals from the 12 countries.

Microbial compositions of the gut microbiomes of the 861 individuals from the 12 countries at the phylum (a) and genus level (b) are shown. In each country, the individuals were rearranged in the ascending order of the abundance of Firmicutes (a) and *Bacteroides* (b).

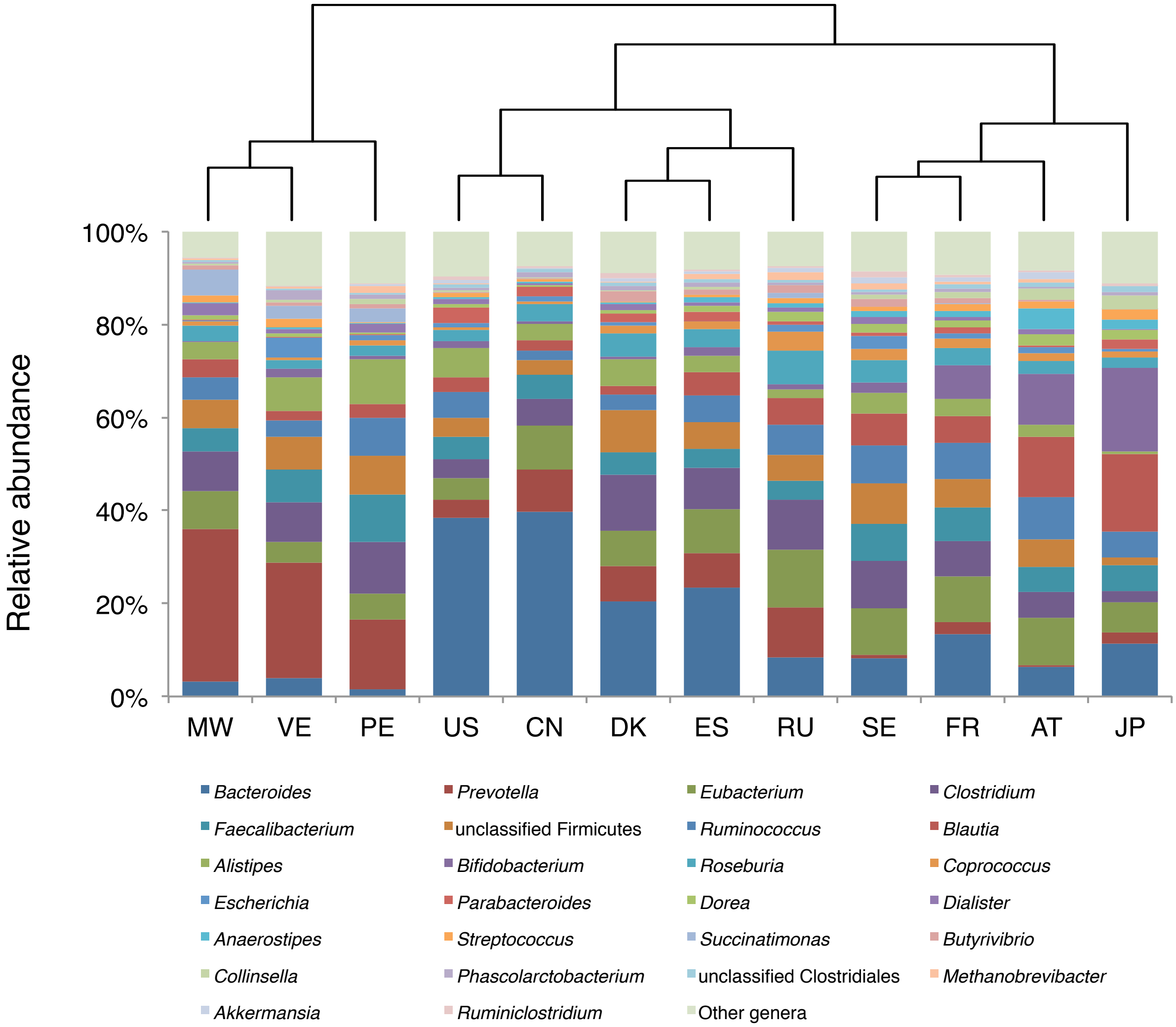


Supplementary Fig. S3. Assessment of variations in microbial compositions obtained from different methodologies.

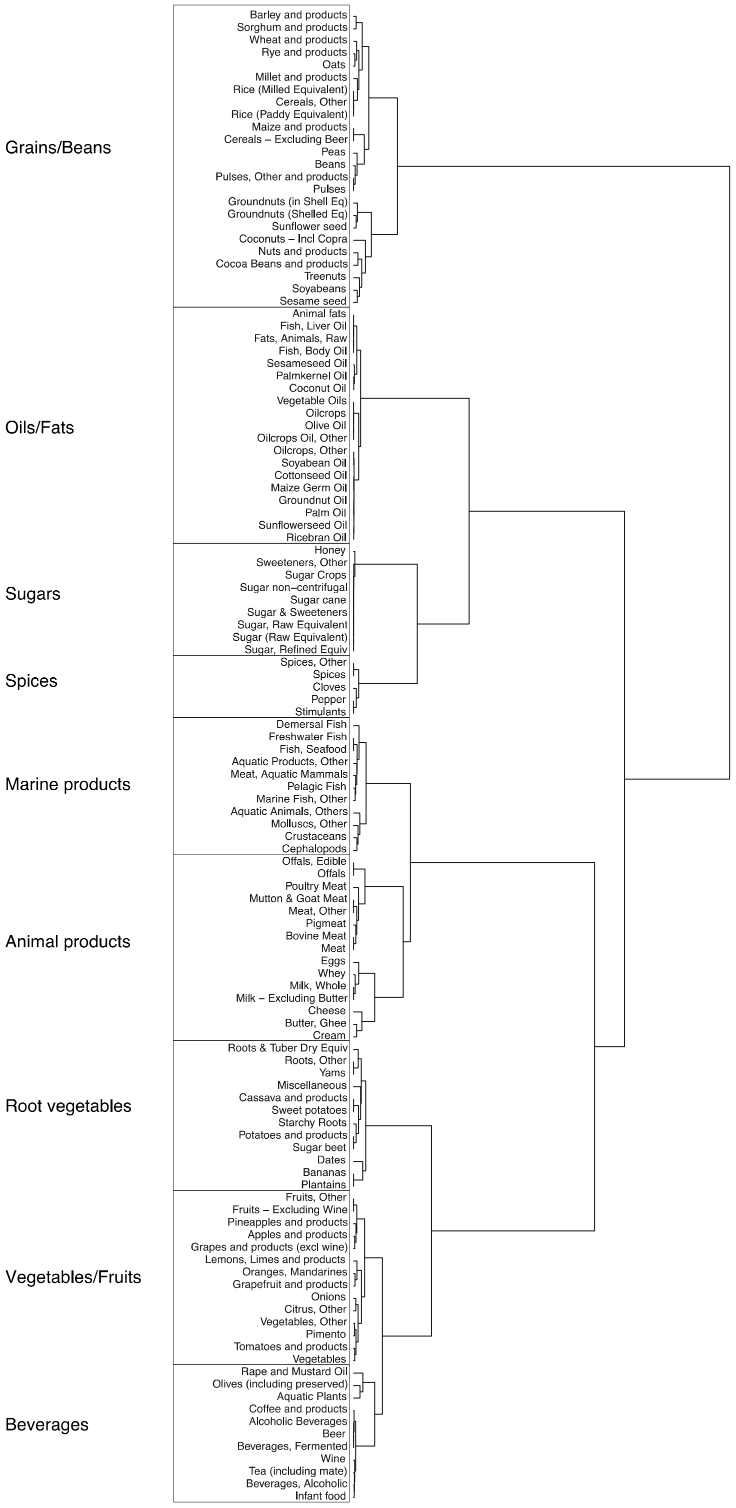
The genus-level average relative abundances obtained using different methodologies are plotted. Circles represent the genera with the average relative abundance ≥ 0.01%. Vertical and horizontal axes indicate the average relative abundance of each genus. (a) Comparison of three sequencers, Roche 454 GS, Illumina MiSeq and Ion PGM using 20 fecal DNA samples. (b) Comparison of four different DNA extraction methods using eight fecal samples. (c) Comparison of four different fecal storage conditions using three fecal samples. Abbreviations for fecal sample storage conditions are summarized in Supplementary Table S4.



Supplementary Fig. S4. Comparison of PCCs between microbial compositions obtained from three different methodologies, and those between individuals of within and between countries. PCCs between the microbial compositions in an individual obtained by different methodologies are shown in the left three boxes (red, yellow, and orange), and those between individuals within and between countries are shown in the right two boxes (green and blue). Several lowered PCCs observed in DNA extraction were not caused by a particular protocol, rather due to differences in the individual sample used. Asterisks represent $P < 0.05$.

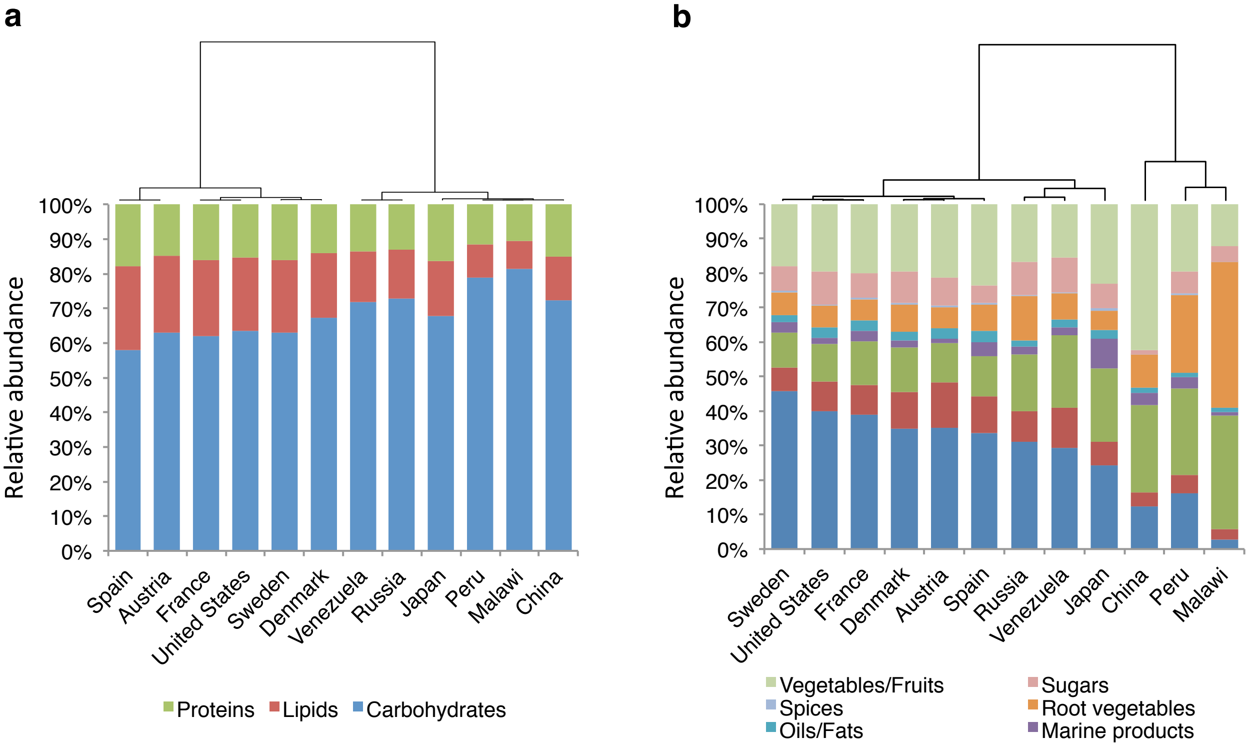


Supplementary Fig. S5. **Hierarchical clustering of the 12 countries based on the average microbial composition in the human gut microbiome.**
The top 26 genera with an average relative abundance of ≥ 0.5% are shown.

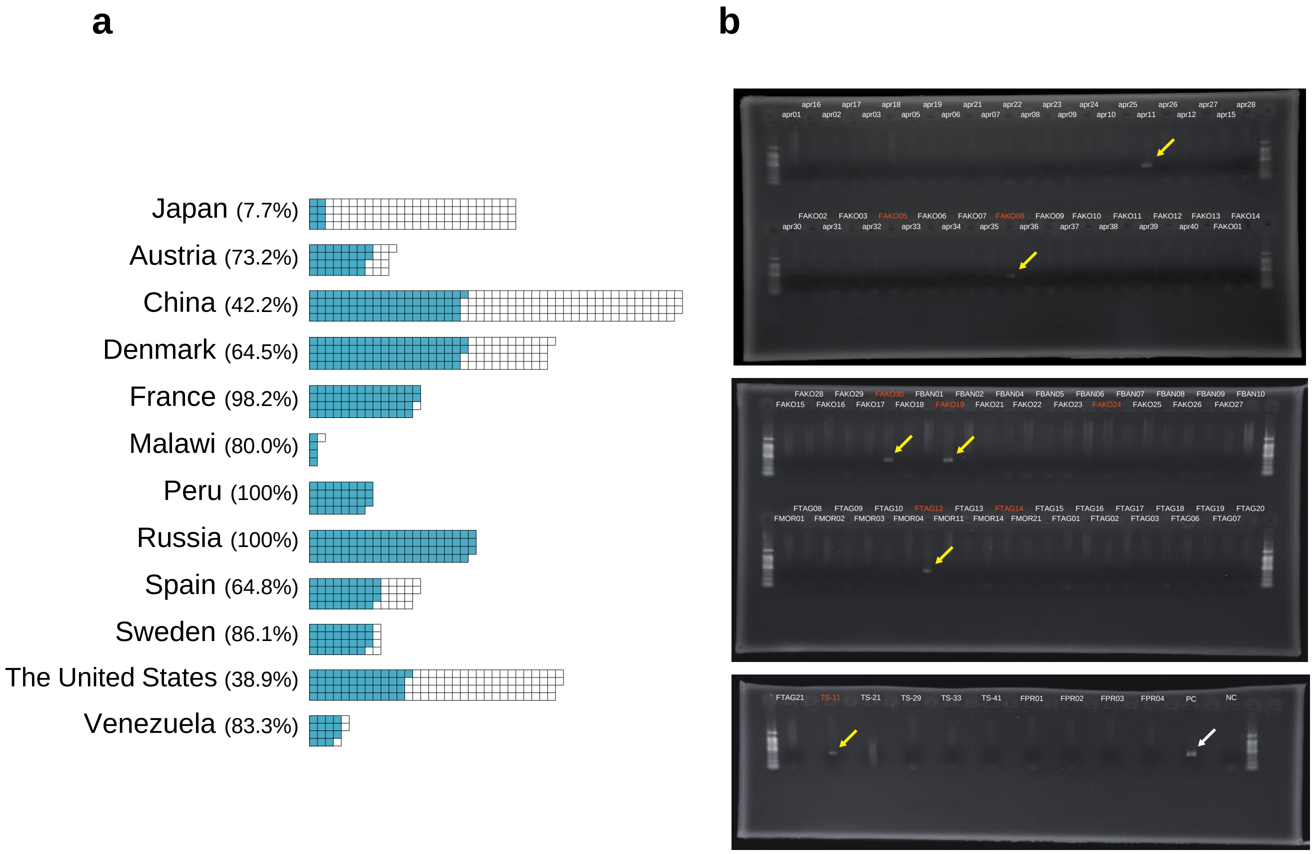


Supplementary Fig. S6. Grouping of 119 food items into nine food categories according to the compositional similarity of their nutrients.

The 119 food items in the FAOSTAT database were clustered based on the compositional similarity of nutrients of which levels were calculated according to a book of the Standard Tables of Food Composition in Japan 2010.

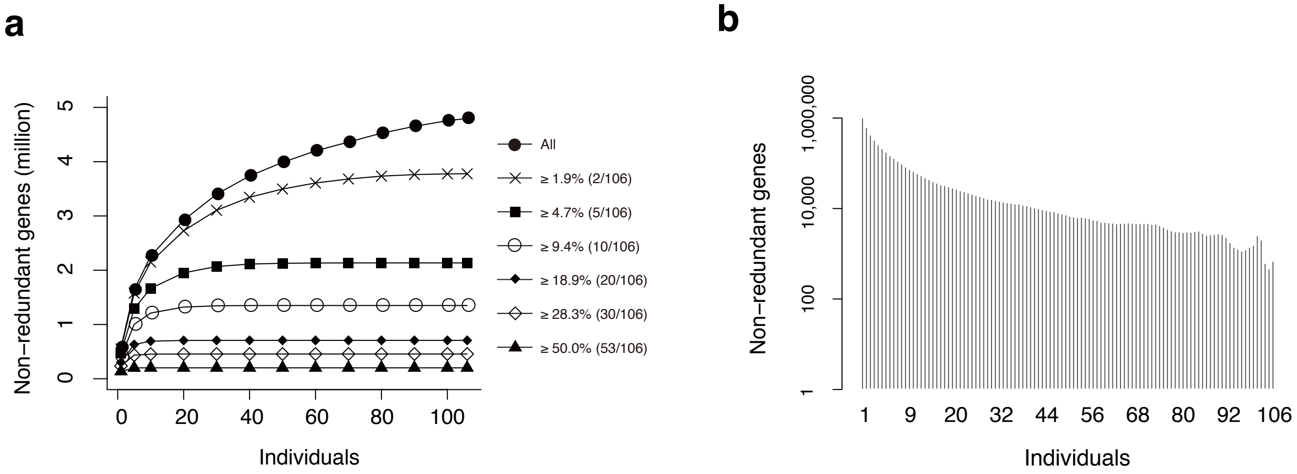


Supplementary Fig. S7. Hierarchical clustering of the 12 countries based on average dietary intake data in the 10 years from 2002 to 2011.
(a) The dendrogarm of the 12 countries based on the ratio of three main nutrients (carbohydrates, proteins, and lipids) is shown. (b) The dendrogarm of the 12 countries based on the ratio of nine food categories is shown.



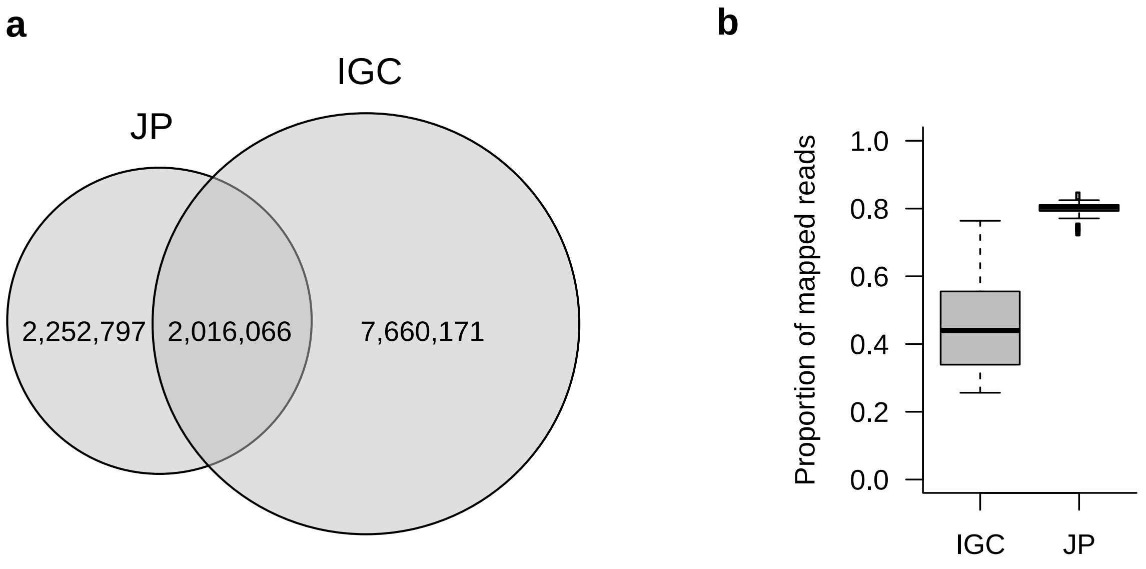
Supplementary Fig. S8. Detection of *M. smithii* in the human gut microbiome.

(a) Open and blue boxes indicate the individuals for which *M. smithii* was detected and undetected, respectively, by mapping of metagenomic reads to the reference genomes. Numbers in parentheses represent the proportion of the individuals *M. smithii* detected. (b) PCR detection of *M. smithii* in the 106 JP individuals. Individual's IDs are represented at each lane, and the ones indicated in orange were *M. smithii*-positive in the mapping analysis. Yellow arrows indicate the bands for the PCR product of *M. smithii*, of which the positive control (PC) is shown by a white arrow. NC, negative control.

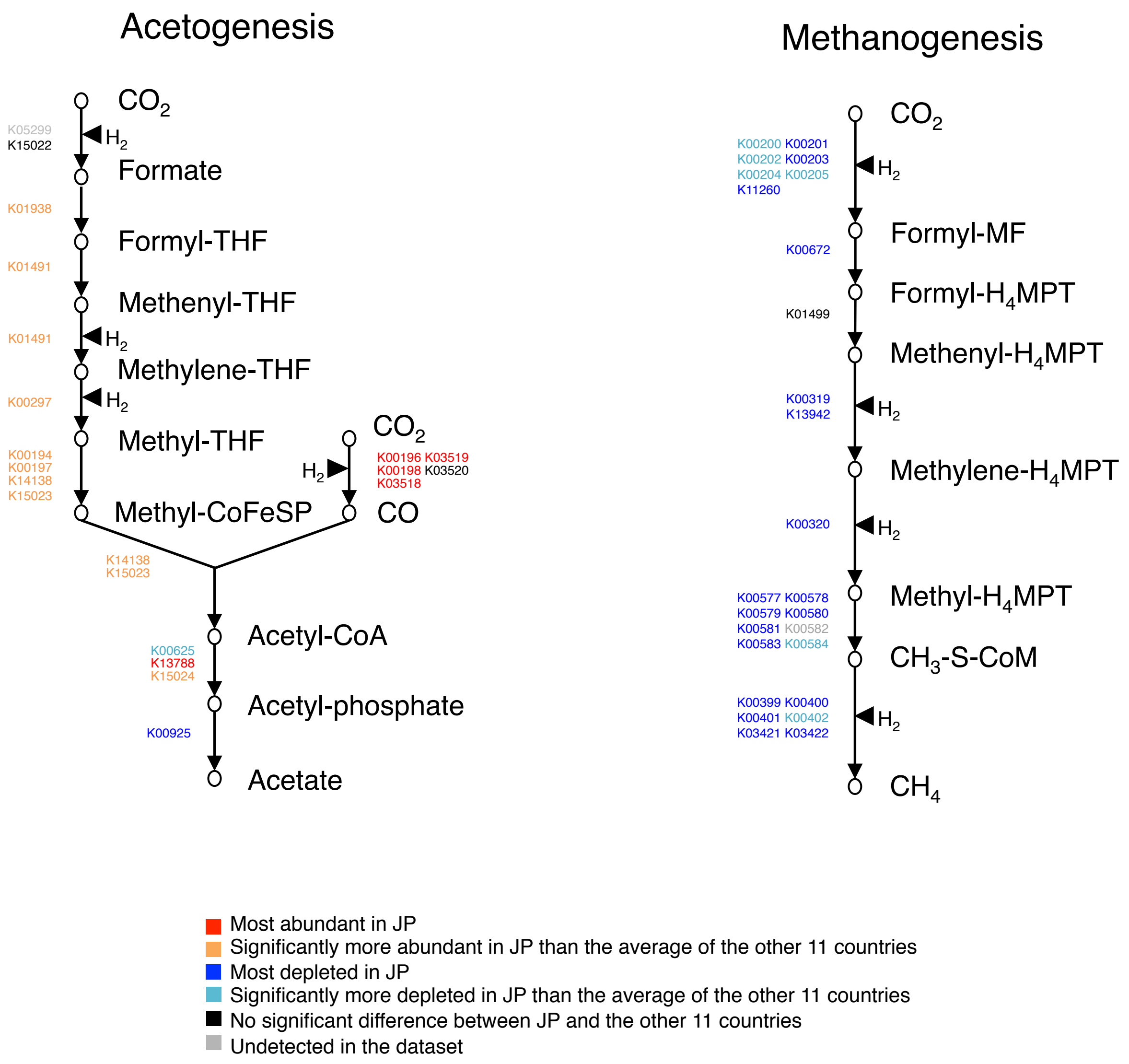


Supplementary Fig. S9. Non-redundant genes in the JPGM.

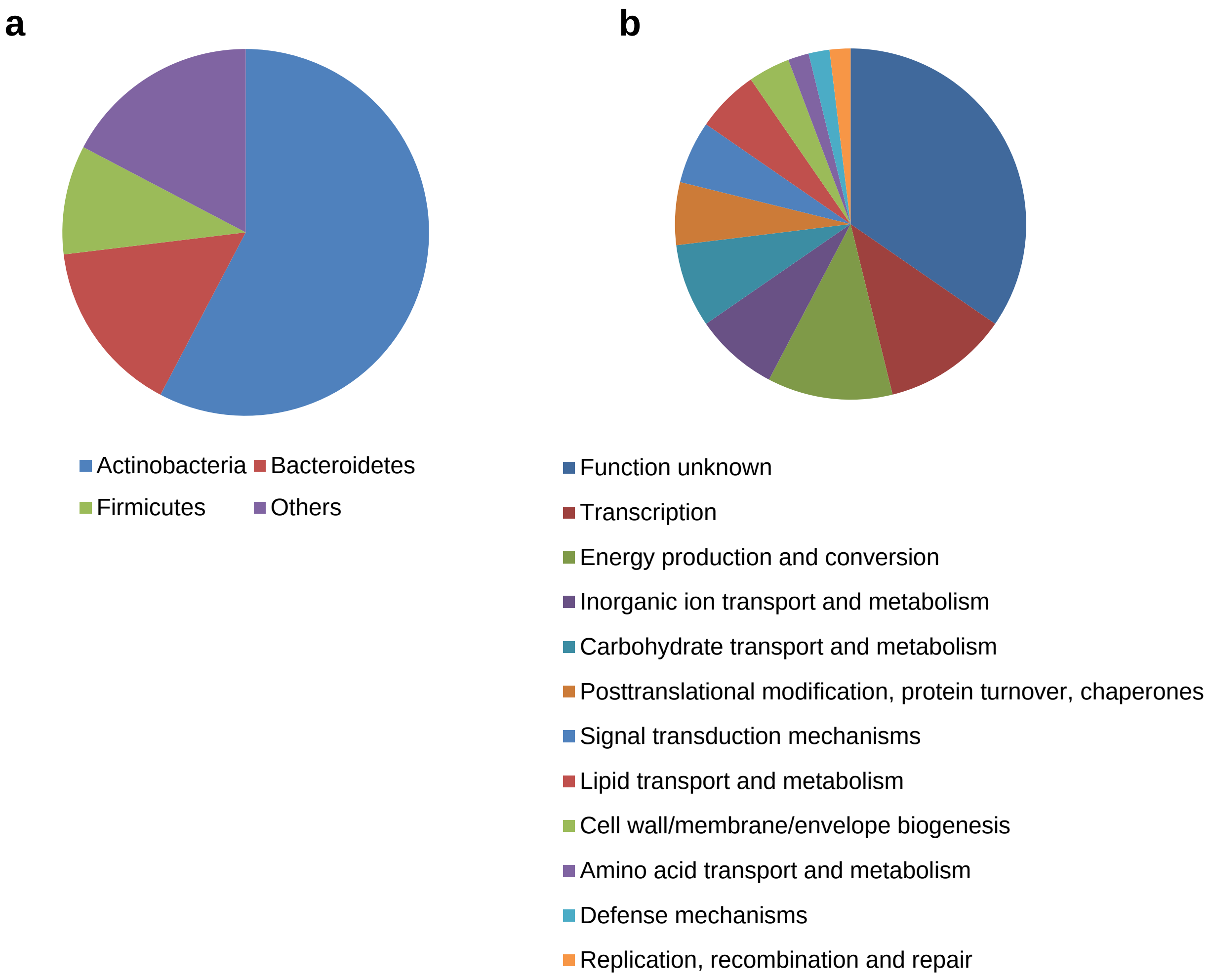
(a) The numbers of detected non-redundant genes plotted against the numbers of the JP individuals[19]. (b) The number of the non-redundant genes shared by each number of the individuals plotted against the number of individuals.



Supplementary Fig. S10. Comparison of the JP and IGC non-redundant gene sets.
(a) Venn diagram of the number of genes in both datasets. (b) Ratio of the mapped Japanese metagenomic reads to the JP and IGC gene sets.



Supplementary Fig. S11. Metabolic pathways of acetogenesis and methanogenesis.
Metabolic pathways for acetogenesis and methanogenesis and KOs involved are shown. Colors indicate differences in the abundance of the KOs shown in the figure.



Supplementary Fig. S12. Taxonomic and functional assignment of the 52 NOGs having a higher abundance in the JP cohort than in the other populations.
(a) Distribution of the 52 NOGs per phylum is shown. “Others” indicates more than two phyla. (b) Distribution of the 52 NOGs per functions is shown.