Human reference gut microbiome comprising 5,414 prokaryotic species, including newly assembled genomes from under-represented Asian metagenomes

Chan Yeong Kim[1†], Muyoung Lee[1†], Sunmo Yang[1], Kyungnam Kim[2], Dongeun Yong[2], Hye Ryun Kim[3], and Insuk Lee[1*]

[1] Department of Biotechnology, College of Life Science & Biotechnology, Yonsei University, Seoul 03722, Korea

[2] Department of Laboratory Medicine, Research Institute of Bacterial Resistance, College of Medicine, Yonsei University, Seoul 03722, Korea

[3] Division of Medical Oncology, Department of Internal Medicine, Yonsei Cancer Center, College of Medicine, Yonsei University, Seoul 03722, Korea

[†] These authors contributed equally to this work

* Corresponding author:

Insuk Lee

Tel: +82-10-4186-8706, E-mail: insuklee@yonsei.ac.kr

Short title: Human reference gut microbiome

Key words: metagenomic shotgun sequencing, human gut microbiome, metagenome-assembled genome

Abstract

Metagenome sampling bias for geographical location and lifestyle is partially responsible for the incomplete catalog of reference genomes of gut microbial species. Here, we present a substantially expanded microbiome catalog, the Human Reference Gut Microbiome (HRGM). Incorporating newly assembled 29,082 genomes from 845 fecal samples collected from three under-represented Asian countries—Korea, India, and Japan—the HRGM contains 232,098 non-redundant genomes of 5,414 representative prokaryotic species, >103 million unique proteins, and >274 million single-nucleotide variants. This is an over 10% increase from the largest reference database. The newly assembled genomes were enriched for members of the *Bacteroidaceae* family, including species associated with high-fiber and seaweed-rich diet. Single-nucleotide variant density was positively associated with the speciation rate of gut commensals. Ultra-deep sequencing facilitated the assembly of genomes of low-abundance taxa, and deep sequencing (>20 million read pairs) was needed for the profiling of low-abundance taxa. Importantly, the HRGM greatly improved the taxonomic and functional classification of sequencing reads from fecal samples. Finally, mapping homologous sequences for human auto-antigens onto the HRGM genomes revealed the association of commensal bacteria with high cross-reactivity potential with autoimmunity. The HRGM (www.mbiomenet.org/HRGM/) will facilitate the identification and functional analysis of disease-associated gut microbiota.

Introduction

Human gut microbiome is considered the “second human genome” and plays a crucial role in various diseases[1,2]. Therefore, targeting gut microbes and their functional elements may provide novel therapeutic opportunities. The assembly of human reference genome, together with a catalog of protein-coding genes and genomic variants, led us to the era of genomic medicine. Likewise, transformation of human medicine by harnessing the gut microbes requires the cataloging of reference microbial genomes and their encoded functional elements. Conventional approaches for microbial genome assembly require microbial isolation and culture. Indeed, with the development of culturomics technology, the number of culturable gut microbes has increased greatly[3-6]. However, the culturable taxa are biased toward specific clades, and a large portion of the human gut microbiome remains unculturable[7-9]. To address this, culture-independent methods of metagenome assembly from whole-metagenomic shotgun sequencing (WMS) data have been developed.

Recently, three independent studies have consecutively released large collections of prokaryotic genomes, including many based on metagenome assembly[8-10]. The metagenome-assembled genomes (MAGs) from these studies were then combined with the genomic information deposited in public databases to generate integrated catalogs of prokaryotic genomes and proteins in the human gut[11], the Unified Human Gastrointestinal Genome (UHGG) and Unified Human Gastrointestinal Protein (UHGP) catalogs, respectively. The UHGG contains 204,938 non-redundant genomes that represent 4,644 prokaryotic species and the UHGP catalogs approximately 95 million unique proteins.

Despite the latest advances, the current human gut microbiome catalog is incomplete, partially because the metagenome sampling is biased for geographical location and lifestyle. Specifically, the UHGG is strongly biased towards fecal samples collected in China, Denmark, Spain, and the US. In the present study, to account for the under-sampling of certain metagenomes, we assembled genomes from fecal samples collected from Korea, India, and Japan. Since the genome assembly of low-abundance species in most human fecal samples may require a much deeper sequencing than usually employed, we performed ultra-deep WMS (>30 Gbp or >100 million read pairs) of 90 fecal samples collected from Korea. We then collected public WMS data for 110 fecal samples from India and 805 fecal samples from Japan. We consequently assembled 29,082 prokaryotic genomes, and combined them with the UHGG genomes to generate the Human Reference Gut Microbiome (HRGM), which substantially expands the list of representative species, genomes, proteins, and single-nucleotide variants (SNVs) in the human gut microbiome. The HRGM is a freely available resource and will be invaluable to therapeutic targeting of the gut microbiota.

Results

Assembly of gut microbial genomes from Korea, India, and Japan

We assembled prokaryotic genomes using an in-house bioinformatics pipeline (**Supplementary Fig. 1a, Methods**), which is more exhaustive than similar approaches[8-11] (**Supplementary Table 1**). For instance, we adopted an ensemble method for binning assembled contigs, as it showed better performance than individual binning tools[12,13]. We hypothesized that metagenomes harbored by individuals from under-represented geographical locations and lifestyles would expand the current catalog of human gut microbiome. Therefore, we performed *de novo* genome assembly of fecal samples from three Asian countries: Korea, India, and Japan (referred to here as KIJ samples, **Supplementary Table 2**). At the start of the current study, WMS data for 805 and 110 fecal samples from Japan and India, respectively, were publicly available but not included in the UHGG[14,15]. To complement these data, we generated WMS data for fecal samples collected from 90 donors recruited in Korea. We set the minimum completeness at 50% and the maximum contamination at 5% for genomes of minimum quality. We divided the genome bins into two groups: high quality (HQ) genomes with ≥90% completeness and ≤5% contamination, and medium quality (MQ) genomes (the remaining genomes). This yielded 29,082 KIJ sample MAGs: 7,767 from Korea, 563 from India, and 20,752 from Japan.

Ultra-deep sequencing facilitates the genomic assembly of low-abundance taxa

To investigate the impact of metagenome sequencing depth on *de novo* genome assembly, we performed ultra-deep sequencing of the 90 Korean fecal samples (>30 Gbp or >100 million read pairs); the depth was approximately 5-fold deeper than the normal sequencing depth (**Fig. 1a**). Despite sequencing at the normal depth, fecal samples from Japan had a larger total read length than Korean samples because of a much larger sample size (**Fig. 1b**). For nine of the 90 Korean samples, approximately 60 Gbp was sequenced for the study of sequencing depth effect on genome assembly. We then generated 81 simulated WMS datasets (9 different depths for each of the 9 original samples with ~60 Gbp depth) and used the same pipeline of *de novo* genome assembly for all samples. As expected, the number of HQ and MQ genomes increased with the increasing sequencing depth. However, the growth rate simultaneously decreased and the proportion of HQ genomes became stable after the initial phase of rapid growth (**Fig. 1c**). Next, we investigated whether the increased sequencing depth improved the quality of assembled genomes. We compared the assembly quality of MAGs for the same species in two different simulated samples at adjacent sequencing depths (**Supplementary Fig. 2**; **Methods**). The quality of MAGs from the greater sequencing depth was significantly higher than that of genomes from the lower sequencing depth in terms of completeness, contamination, N50, and genome size (**Fig. 1d,e; Supplementary Fig. 3a,b**). However, the degree of improvement of the assembly quality diminished as the sequencing depth increased.

We then examined the effect of sequencing depth using the actual WMS data for KIJ samples. The number of HQ and MQ genomes assembled from each sample was highest in the ultra-deep sequenced samples from Korea (**Fig. 1f**). However, the proportion of HQ genomes in samples from Korea and Japan was not significantly different (**Fig. 1g; Supplementary Fig. 3c**). Notably, the genome assembly yield, i.e., the number of assembled genomes divided by the total sequencing length, was highest for samples from Japan (**Fig. 1g**). This suggests that sequencing hundreds of samples at a depth of 5–10 Gbp may constitute the most effective strategy for cataloging MAGs for a given population.

The ultra-deep sequencing may be advantageous for the genome assembly for low-abundance taxa. To test this, we compared MAGs exclusively assembled from each country but not included in the UHGG, i.e., 224, 388, and 18 genomes from Korea, Japan, and India, respectively. We then estimated their relative abundance in fecal samples in an independent population of 926 fecal samples from the US[16], using Kraken2[17]. The genomes assembled exclusively from Korean samples shifted towards low-abundance taxa compared with genomes assembled from samples from other countries (**Fig. 1h**), which confirmed the original hypothesis.

Cataloging reference genomes of 5,414 prokaryotic species from the human gut

To construct the most comprehensive reference database for the human gut microbiome, we integrated the newly generated 29,082 MAGs from KIJ samples with the UHGG genomes using dereplication approach (**Supplementary Fig. 1b**, **Methods**). Dereplication of the 29,082 MAGs resulted in 2,199 clusters of genomes. We selected a representative genome from each cluster to catalog the genomes for 2,199 representative species, which we then integrated with 4,644 representative genomes from the UHGG, via dereplication, resulting in 5,414 clusters of genomes. Finally, we selected 5,414 representative genomes and assigned their phylogenetic classifications using GTDB-Tk[18] (**Fig. 2**). Among these representative genomes, 4,531 (83.7%) genomes were exclusively assembled from metagenomic data, which confirmed the notion that the major portion of the human gut microbiome has not yet been isolated. We identified 16S rRNA sequences in 2,542 representative genomes (47%) (**Supplementary Fig. 4**), covering the majority of phylogenetic clades. Unlike conventional databases of 16S rRNA sequences, the new database provides opportunities for functional interpretation of the detected taxa because it contains genomes corresponding to the 16S rRNA sequences.

The inclusion of MAGs from KIJ samples in the new database allowed several improvements on the UHGG. First, we reduced the data bias toward China among Asian countries (**Supplementary Fig. 5a**). Second, we expanded the total number of non-redundant reference genomes by 13.25% and the number of representative species by 16.6% increase (**Supplementary Table 3**). Among the 5,414 representative genomes, 780 genomes were assembled from KIJ samples only, and 536 representative genomes from the UHGG were replaced with new MAGs from KIJ samples. Hence, 1,316 representative genomes (28.3%) were updated in the HRGM (**Supplementary Fig. 5b**).

New MAGs from Korea, India, and Japan are associated with diet-related lifestyles

Notably, *Bacteroidaceae* family (**Fig. 3,** redtree branches) was enriched in the updated MAGs (P < 0.001, Fisher's exact test). Almost half the genomes from this family are from the *Bacteroides* genus and approximately two-thirds of the other half are from the *Prevotella* genus (**Supplementary Fig. 6**). Interestingly, three widely dispersed regions in the phylogenetic tree were highly enriched in the updated genome set. The first region ("a") encompasses a portion of the *Prevotella* genus and includes 30 genomes annotated as *Prevotella copri*. Accordingly, westernized populations with a typically high-fat and low-complex carbohydrate diet exhibit low prevalence and diversity of *P. copri* compared with non-westernized populations[19]. The second region ("b") encompasses a portion of the *Bacteroides* genus and includes 22 genomes annotated as *Bacteroides plebeius*. This species is typically found in Japanese subjects whose diet includes seaweed-rich food, such as sushi[20]. It has been suggested that *B. plebeius* harbors genes encoding an enzyme specific for algal carbohydrates, acquired from marine microbes. The third region ("c") also encompasses a portion of the *Bacteroides* genus and includes 12 genomes annotated as *Bacteroides vulgatus*, which is typically present in the human distal gut, where undigested plant polysaccharides and proteins exist in large quantities[21]. Together, these observations indicate that the new MAGs from KIJ samples are associated with the diet-related lifestyles in Japan and Korea.

SNV density is positively associated with the speciation rate of gut commensals

We then aligned genomes of species clusters containing ≥3 genomes with the representative genome and mapped SNVs (**Methods**). This yielded 274,543,071 SNVs from 2,821 species clusters, representing 10.07% and 13.34% increases, respectively, from the UHGG. The Actinobacteriota phylum had the highest SNV density (**Fig. 3a**). Phylogenetically overdispersed branches of Actinobacteriota species were apparent in both, the HRGM and UHGG. The majority of genomes from the overdispersed tree region belonged to the *Collinsella* genus. We divided these genomes into ones from a tree region with a modest phylogenetic dispersion (MD, 20 genomes) and those with a high phylogenetic dispersion (HD, 619 genomes) (**Fig. 3b**). Although the majority of genomes were not annotated at the species level, *Collinsella aerofaciens* was enriched in the HD group and other known *Collinsella* species were enriched in the MD group (**Fig. 3c**). SNV density in HD group was significantly higher than that of MD group (**Fig. 3d**).

SNV, a within-species genetic variation, is a major mechanism for the adaptation of commensal species to a distinct host environment. Wide dispersion of species branches indicates rapid speciation. Accordingly, high SNV density for a species with an overdispersed tree may indicate that the degree of within-species genetic variation may be positively associated with the speciation rate of gut commensals. To test this, we examined the correlation between SNV density of representative species and their phylogenetic distance to the five nearest species. The branch length to the neighboring species in the phylogenetic tree of a species that arose during rapid speciation tends to be short. We observed an inverse correlation between the average phylogenetic distance to the five nearest species and their SNV density (**Fig. 3e**), and a significantly higher SNV density for the top 10% species with shorter phylogenetic distance to the nearest five species than those for the bottom 90% species (**Fig. 3f**). This supports the model of a positive correlation of SNV density and the speciation rate of gut commensals.

Functional landscape of 103 million proteins from human gut prokaryotes

Information on proteins encoded in the human gut microbes will facilitate the functional characterization of disease-associated microbiota. Using an in-house computational pipeline for cataloging human gut prokaryotic proteins (**Supplementary Fig. 1c** and **Supplementary Fig. 7**), we first identified 64,661,728 CDS (coding sequences) from 29,082 genomes from KIJ samples using Prodigal[22]. To reduce redundancy in the protein catalog, we first executed CD-HIT[23] at 100% similarity level and then combined with proteins cataloged by the UHGP-100[11]. The consolidated protein catalog was next consecutively clustered by CD-HIT at lower sequence similarity levels: 95%, 90%, 70%, and 50%. This led to approximately 103.7, 20.0, 14.8, 8.5, and 4.7 million proteins at the sequence similarity levels of 100%, 95%, 90%, 70%, and 50%, respectively.

Unexpectedly, we observed that the UHGP contains proteins that are 100% identical, even in a catalog at 50% sequence similarity level. For instance, among the UHGP-50 proteins, GUT_GENOME232012_01109 and GUT_GENOME231777_00918 have an identical amino acid sequence. We identified 8,663, 82,507, 243,362, and 75,620,150 proteins that are redundant at 100% similarity in the UHGP-50, UHGP-90, UHGP-95, and UHGP-100, respectively. Exclusion of the UHGP proteins that were 100% identical revealed that the HRGM contains more proteins than UHGP at all levels of sequence similarity except for 50% (**Supplementary Table 3**).

To facilitate the functional interpretation of gut microbiome profiles, we next annotated functional genomic elements and proteins in the HRMG. We predicted and annotated non-coding RNAs and functional peptides, using Prokka[24]; antibiotic resistance genes, using RGI[25]; biosynthetic gene clusters, using antiSMASH[26]; and 16S rRNA regions, using barrnap[27]. For functional annotation of proteins, we used eggNOG-mapper[28]. Notably, the landscape of antibiotic resistance ontology revealed that phylogenetically close species in the human gut tend to share antibiotic resistance mechanisms (**Supplementary Fig. 8**). A significantly large portion of the human gut prokaryotic proteins has not yet been functionally annotated. For the HRGM protein catalogs at 100%, 95%, 90%, 70%, and 50% similarity levels, 13.13%, 28.05%, 29.17%, 36.35%, and 47.62% of proteins, respectively, had no functional annotation, according to eggNOG-mapper. This effect appears to be amplified by redundant proteins, resulting in a reduced annotation rate at low similarity level. Further, the annotation rate of proteins that are shared by many species is higher than that of species-specific proteins (**Supplementary Fig. 9**).

HRGM improves taxonomic and functional classification of sequencing reads

According to a recent benchmark study, whole-DNA–based methods outperform marker-based methods for taxonomic classification of metagenomic sequencing reads[29]. The performance of whole-DNA–based methods relies on the quality of the reference genome database. The standard databases lack numerous genomes of species that exist in the human gut, which leads to false-negatives, while including many genomes from other microbial communities, which leads to false-positives[29]. We hypothesized that the HRGM, which is specific to the human gut microbiome and more comprehensive than other databases, can improve the taxonomic classification of sequencing reads. We used Kraken2[17] to compare the taxonomic classification of three genome databases: a standard database that contains RefSeq[30] complete genomes (RefSeq CG) of bacterial, archaeal, and viral domains; the UHGG-based database; and the HRGM-based database. To generate independent test datasets, we compiled WMS data for 1,022 fecal samples from the US, Cameroon, Luxembourg, and Korea, which were not included in the UHGG nor HRGM. We then evaluated the efficacy of Kraken2 classification based on the proportion of classified reads (**Methods**). The classification efficacy using the UHGG and HRGM-based databases was substantially higher than that of the standard database (**Fig. 4a,b**, $P < 0.001$, two-sided Wilcoxon signed-rank test). In addition, the variance of the read classification rate of custom databases was significantly smaller than that of the standard database, except for the Cameroon population (**Fig. 4a**, $P < 0.001$, Brown-Forsythe test). Importantly, the classification efficacy of the HRGM-based database was significantly improved compared with that of the UHGG-based database for the four test samples (**Fig. 4a,c**, $P < 0.001$, two-sided Wilcoxon signed-rank test), which suggests that the updated reference genome database improves taxonomic classification of the gut metagenomic sequencing data.

Next, we investigated the efficacy of functional classification based on the number of aligned sequencing reads from reference protein databases. Because of the extremely large number of reference proteins, we used only 40 samples randomly selected from the 1,022 fecal samples (10 samples from each population), and aligned the sequencing reads with the UHGP-95 and HRGM-95 protein catalogs (**Methods**). The number of aligned reads was 1.31% higher, on average, with HRGM-95 in all tested samples than with UHGP-95 (**Fig. 4d**), although HRGM-95 contains 0.4% more proteins than UHGP-95.

Taken together, the newly assembled genomes from under-represented Asian countries significantly improve the genome and protein databases for metagenomic analysis of both, taxonomic and functional profiling.

Reliable taxonomic profiling of low-abundance taxa requires deep sequencing

Taxonomic profiles obtained by shallow sequencing (0.5–2 million reads) highly correlate with those obtained by ultra-deep sequencing (2.5 billion reads)[31]. However, this evaluation is based on entire taxa, in which highly abundant or core taxa govern the correlation measure. Further, low-abundance taxa likely play important, as yet unknown, biological roles in the gut microbial communities[32,33]. We therefore evaluated the impact of sequencing depth on the reliability of taxonomic profiling for different ranges of taxon abundance. We generated a simulated dataset at various sequencing depths 16 new Korean fecal samples, and not included in the HRGM. We then stratified the taxonomic features into eight different groups, according to the mean relative abundance (**Fig. 5a,b**). We calculated the mean Pearson correlation coefficient (*PCC*) and the mean Spearman correlation coefficient (*SCC*) between the taxonomic profiles at different sequencing depths for different mean relative abundances (**Methods**). The taxonomic profile similarity between two groups showed increasing *PCC* and *SCC* with an increasing sequencing depth. For example, >10 million read pairs (3 Gbp) may need to have taxonomic profiles that highly correlate (*PCC* > 0.9) with those based on 80 million read pairs (25 Gbp) to account for the features with lowest 13.92% of relative abundance (relative abundance < 1e–06) (**Fig. 5c and Supplementary Fig. 10a**). For *SCC* > 0.9, the required sequencing depth increased to 20 million read pairs (6 Gbp) for taxonomic features with a similar level of relative abundance (**Fig. 5b and Supplementary Fig. 10b**). Overall, these observations suggest that deep sequencing (>20 million read pairs) may be required to obtain reliable taxonomic profiles of low-abundance taxa.

Sequencing 30 Gbp is optimal for functional profiling of the human gut microbiome

Next, using the protein catalog, we investigated the optimal sequencing depth for functional profiling of the human gut microbiome. Since the detection of gene content generally requires a much deeper sequencing depth than that for the detection of genomes, we analyzed the WMS data for five Korean fecal samples at a depth of approximately 200 million read pairs (60 Gbp) (**Methods**). The number of the detected coding genes initially grew rapidly as the sequencing depth increased, but later approached the estimated maximum count (**Supplementary Fig. 11a**). The curves fitted well (R^2 > 0.99) two-site saturation models[34], and we hence estimated the maximum number of coding genes for each sample using the regression model. Interestingly, the estimated maximum gene counts in the samples differed, reflecting the different alpha diversity of the microbial community. However, all samples showed very similar normalized maximum gene count curves, with over 80% of the gut microbial coding genes detected by sequencing 30 Gbp or 100 million read pairs in all samples (**Supplementary Fig. 11b**). Sequencing another 30 Gbp would fail to detect 90% of the maximum gene count. Therefore, 100 million read pairs is the optimal sequencing depth for the best trade-off between the sequencing cost and the gain-of-functional information for WMS-based studies of the human gut microbiome.

Profiling cross-reactivity potential identifies autoimmune-associated commensals

Microbial peptides homologous to the host auto-antigens may stimulate host immune cells and, hence, the hypothesis of molecular mimicry has emerged as a mechanism underlying autoimmune diseases[35]. To systematically evaluate this hypothesis, we mapped microbial peptide sequences homologous to the human self-antigens involved in autoimmune diseases onto the genomes of HRGM representative species. We first compiled autoimmune disease-related antigen set from the Immune Epitope Database (IEDB)[36], and then used it for homology-searches of microbial peptide sequences from 5,414 representative species. We thus identified species with a high cross-reactivity potential based on the density of the encoded cross-reactive epitopes. Because the number of epitope-containing genes (ECG) increased as the number of coding genes increased (**Fig. 6a**), we divided the ECG count by the total number of genes for each species. Some human gut commensals had a relatively high cross-reactivity potential (**Fig. 6b,c, Methods**). On the genus level, *Akkermansia*, *Alistipes*, *Bifidobacterium*, *Lawsonibacter*, *Oscillibacter*, *Prevotella*, and *Sutterella* have a high cross-reactivity potential (**Fig. 6d**). Indeed, many of them are associated with autoimmune diseases. For example, *Akkermansia muciniphila* is abundant in the enthesitis-related arthritis patients[37], while *Bifidobacterium* is enriched in these[37] and inflammatory bowel disease (IBD) patients[38]. Increased abundance of *Oscillibacter* is accompanied by increased levels of interleukin 6[39], a pro-inflammatory cytokine that can disrupt the immune homeostasis and increase the risk of autoimmune diseases. The abundance of intestinal *Prevotella copri* is strongly correlated with the risk of arthritis[40] and *Sutterella wadsworthensis* is enriched in ulcerative colitis patients who do not respond to fecal microbiota transplantation[41]. These suggests that cross-reactivity potential of commensal genomes is predictive for human gut microbiota associated with autoimmune diseases.

Discussion

In the present study, we constructed an improved catalog of the human reference gut prokaryotic genomes and their proteins, by including MAGs from fecal metagenomes from under-represented Asian countries. Inclusion of the newly assembled genomes expanded the catalog size by over 10%. In addition, we demonstrated that database expansion also significantly improved the taxonomic and functional classification of sequencing reads. Many new MAGs were associated with diet-related lifestyles at the sampled geographic locations.

Therefore, complementation of metagenome datasets to account for under-sampled geographical locations and lifestyles might be an effective strategy for improving the human reference gut microbiome.

We also demonstrated that the analysis of microbial DNA and peptide sequences facilitates the understanding of gut commensal speciation and interactions with the host immunity. The colonizing commensal microbes adjust to their host environment via genetic changes and selection, which lead to genetic variation within species. We cataloged the SNVs of conspecific genomes and found that the SNV density of gut prokaryotic species is inversely correlated with the phylogenetic distance to their neighboring species. This may suggest that the degree of within-species genetic variation is positively associated with the speciation rate of gut commensal microbes. Whether SNV actually enhances the speciation rate should be addressed in future investigations. Finally, we showed that systematic analysis of microbial peptide sequences homologous to the host auto-antigens allows the prediction of gut microbial taxa potentially associated with autoimmune disease via the mechanism of molecular mimicry. Such analysis is only possible if microbial protein sequences are available with the corresponding taxonomic information.

As the WMS analysis for population-wide human gut microbiome profiling increases in popularity, the choice of sequencing depth is an important factor to consider in study design. Here, we demonstrated that deep sequencing (>20 million read pairs) is necessary for reliable taxonomic profiling of low-abundance commensals. The current knowledge of human gut microbiome is biased towards core taxa that are usually highly abundant. Low sequencing depth (e.g., 0.5–2 million read pairs) may be sufficient for the profiling of core taxa, but not those with low abundance. Deep sequencing may therefore be required for the WMS-based analysis of human gut microbiome to investigate the function of relatively unexplored low-abundance species. Accordingly, the current study provides the guidelines for the choice of sequencing-depth for the analysis of human gut microbiome for different purposes.

In conclusion, the HRGM database, which contains information on various biological entities, from DNA and protein sequences to pan-genomes of species, is a versatile resource for functional dissection of disease-associated gut microbiota. The data will be available via a web server (www.mbiomenet.org/HRGM/) and will be periodically updated as new WMS data for fecal samples become publically available.

Methods

Sequencing fecal metagenome samples from Korea, India, and Japan

WMS data for fecal samples from India and Japan were obtained from published studies[14,15]. Fecal WMS data for India were generated for 110 healthy donors in North-Central and Southern India[14]. Although the sequencing depth was relatively low (1.2 Gbp on average), it was expected that many novel genomes would be assembled because MAGs from India are not included in the existing catalogs. By contrast, 805 MAGs from Japan are included in the UHGG. However, it was expected that the inclusion of the recently published deep-sequencing WMS data for 645 Japanese fecal samples (6.5 Gbp on average)[15] would greatly expand the number of MAGs for Japan. In addition, ultra-deep WMS data (31 Gbp on average) were generated for fecal samples from 90 Koreans recruited by the Severance Hospital (Seoul, Korea; IRB No 4-2020-0309 and IRB No 4-2017-0788). Written informed consent was obtained before the study. The UHGG does not contain any MAGs from Korea.

The libraries were prepared as described in the TruSeq Nano DNA Library Prep Reference Guide (Illumina #15041110). Briefly, 100 ng DNA was fragmented using LE220 Focused ultrasonicator (Covaris, Inc.). Fragmented DNA was end-repaired and approximately 350-bp fragments were obtained after size selection. After adapter ligation, eight PCR cycles were performed. Library quantification was performed as described in the Kapa Illumina Library Quantification Kit (Kapa Biosystems, #KK4854). Next, 150 bp ×2 paired-end sequencing was performed using Illumina HiSeq4000. In summary, new WMS data for 845 fecal samples collected from Korea, India, and Japan were obtained. The total read length was 7.2 Tbp. All samples used in the current study are described in **Supplementary Table 2**.

Metagenome assembly and binning

The adapter sequences were trimmed, and low-quality bases and short reads were removed from WMS data using Trimmomatic v0.39[42]. Next, the reads were aligned with the human genome GRCh38.p7 using Bowtie2 v2.3.5[43], and the aligned reads were then removed. The majority of quality-controlled reads were assembled as contigs using metaSPAdes[44], which is a metagenome-specific pipeline of SPAdes v3.13.0. For unknown reasons, and regardless of sample size, metaSPAdes runtime was excessively long for 107 samples. In those cases, MEGAHIT v1.2.8[45] was used (**Supplementary Table 2**).

Genome bins were generated using the ensemble approach and three binning tools: MetaBAT2 v2.13[46], MaxBin2.0 v2.2.6[47], and CONCOCT v1.1.0[48]. First, the reads from each sample were first aligned with the assembled contigs from the previous step using Bowtie2, and the three binning programs were initiated. The minimum size of a contig for binning was set at 1,000 bp, except for MetaBAT2, which requires at least 1,500 bp. The three binning predictions were combined for improved binning results using the bin refinement module of MetaWRAP v1.2.2[12], which uses CheckM v1.0.18[49] to evaluate the quality of genome bins in terms of completeness and contamination rate. The minimum completeness was set at 50%, the maximum contamination at 5%, and the minimum quality score ($Completeness - 5 \times Contamination$) at 50. The same threshold values for CheckM results were applied during the construction of the UHGG. This resulted in 7,767 genomes from Korean samples, 563 genomes from Indian samples, and 20,752 genomes from Japanese samples (29,082 genomes in total). The genome bins were divided into two groups: HQ, bins with over 90% completeness and less than 5% contamination; and MQ, bins with 50–90% completeness and less than 5% contamination.

Generation of genomic species clusters

Groups of genomes that corresponded to species were generated using a two-step iterative procedure. Preliminary clustering was performed using Mash v2.2[50] algorithm. Mash distances were calculated for all possible pairs of genomes using the “-s 10,000” parameter. Next, the average-linkage–based hierarchical clustering was performed, at a cutoff of 0.2. Mash algorithm is sufficiently fast to calculate all-by-all distances for hundreds of thousands of genomes in a timely manner. However, this compromises the accuracy, especially for low-coverage genome pairs[51], which are common in MAGs. Therefore, to improve cluster quality, ANImf[51] was calculated for every pair of genomes within each initial cluster. To avoid the over-estimation of ANI by local alignment, a minimum coverage threshold was applied for each pair. The coverage cutoff of genome A and genome B was determined at *min(0.8,* *Completeness of genome A* × *Completeness of genome B)*. If the alignment coverage between two genomes was lower than the cutoff, they were regarded as different genomes. The genomes were then clustered using the average linkage-based hierarchical clustering at a cutoff of 0.05 (or 95% identity), which is a widely accepted ANI threshold for species-level boundary[4,9-11,52]. The genome intactness score (*S*)[9,11], $S = Completeness - 5 \times Contamination + 0.5 \times \log_{10}(N50)$, was then calculated. For clusters containing more than two genomes, a genome with the highest *S* was selected as the representative genome for the cluster. The above two-step procedure was iterated until the clusters ceased to change. Hence, 2,199 species clusters were generated for 29,082 genomes from KIJ samples, with eight iterations of the aforementioned procedure (**Supplementary Fig. 1a**). Finally, the 2,199 genomes were combined with 4,644 genomes from the UHGG, generating 5,414 species clusters for the HRGM at the fourth iteration (**Supplementary Fig. 1b**).

Non-redundant genome counting

To count the number of non-redundant genomes, the redundant genomes were removed, similar to what was done for the UHGG pipeline[11]. First, the pairwise genome distance was calculated using Mash[50] and the entire genomes were clustered using average-linkage–based hierarchical clustering, with a 0.001 cutoff (Mash ANI 99.9%). To reduce the computation time, the hierarchical clustering was performed only for the connected components with the distance of 0.1, because it is highly unlikely that genomes that are not within the distance of 0.1 are clustered together by a distance of 0.001. In the process, 22,761 genomes were clustered into 8,508 conspecific genome bins. Multiple genomes from the same sample in the same species bin were counted only once.

Taxonomic and functional annotation of representative species genomes

The taxonomic annotation of 5,414 representative species genomes was performed using the "classify_wf" function of GTDB-Tk v1.0.2[18]. The reference version was GTDB R04-RS89, released in June 2019. Genomic features, such as CDS, rRNA, and tRNA, were identified and annotated in each genome using Prokka v1.14.5[24] with "--kingdom Bacteria" and "--kingdom Archaea" parameters for the bacterial and archaeal genomes, respectively. With the protein sequences predicted by Prokka, the antibiotic resistance genes were annotated using RGI v5.1.0[25] with default parameters. The landscape of antibiotic resistance potential of 5,414 species-representative genomes is depicted in **Supplementary Fig. 8**. Finally, the secondary metabolite gene cluster was annotated using antiSMASH v5.1.2[26]. For the full-featured annotation, the "--cb-general, --cb-knownclusters, --cb-subclusters, --asf, --pfam2go, --smcog-trees, --cf-create-clusters" parameters were set.

To render the HRGM useful for the 16S rRNA sequencing-based metagenomic analysis, the 16S rRNA regions for 5,414 representative species genomes were predicted using barrnap v0.9[27] tool and the "--evalue 1e-05" parameter, and "--kingdom bac" and "--kingdom arc" parameters for bacterial and archaeal genomes, respectively. The 16S rRNA sequences were thus directly predicted from 1,364 representative species genomes. For the remaining 4,050 representative species, the search for 16S rRNA sequences was expanded to their conspecific genomes. The barrnap analysis was used for the genomes from KIJ samples and pre-established 16S rRNA region annotations were used for the genomes from the UHGG. Within the expanded search space, 16S rRNA sequences were identified for 1,178 additional genomes. Consequently, 16S rRNA sequences were generated for 2,542 species in the HRGM (**Supplementary Fig. 4**).

Cataloging SNVs

For the species bins with more than three genomes, SNVs were identified using the codes provided by the UHGG[11]. Briefly, non-representative genomes were aligned with the representative genome in the species bin using nucmer 4.0.0beta2[53]. Best bi-directional alignments were identified using the delta-filter program and "-q –r" options, and SNVs were annotated using the show-snp program; nucmer, delta-filter, and show-snp are software packages of MUMmer v3[54]. For each species bin (G) whose representative genome is r, the number of SNV per kb was calculated as follows:

$$SNV\ per\ kb = \frac{\sum_{g\in(G-\{r\})} \frac{\#\,SNV_{r,g}}{Aligned\ length_{r,g}/1000}}{n(G)-1}$$

SNV per kb was only calculated for 1,521 species bins with ≥10 genomes to avoid bias. For the 1,521 genomes, the average phylogenetic distance to the five nearest species was calculated using the IQ-Tree[55].

Cataloging gut prokaryotic proteins and their functional annotation

Overall, 64,661,728 CDS were identified in 29,082 genomes from the KIJ set using Prodigal v2.6.3[22] and “-c -m -p single” parameters. Since many proteins were derived from conspecific genomes, the catalog may have included many homologous proteins. To reduce the redundancy in the protein catalog, CD-HIT v4.8.1[23] was adopted. To reduce CD-HIT running time, identical proteins were first clustered and then CD-HIT was executed at 100% similarity level. The cataloged proteins were then combined with those in UHGP-100[11]. The consolidated protein catalog was subsequently submitted to CD-HIT clustering analysis at five different sequence similarity levels, 100%, 95%, 90%, 70%, and 50%. For accurate and efficient clustering, a multi-step iterative clustering method recommended by the CD-HIT tutorial was adopted. For instance, the CD-HIT-95 protein catalog (a 95% similarity level protein catalog) was constructed based on CD-HIT-100 proteins, and the CD-HIT-90 protein catalog was constructed based on CD-HIT-95 proteins. This resulted in approximately 103.7 million, 20.0 million, 14.8 million, 8.5 million, and 4.7 million proteins at the sequence similarity levels of 100%, 95%, 90%, 70%, and 50%, respectively. The overall pipeline for protein catalog construction is depicted in **Supplementary Fig. 7**.

Representative protein sequences in the five protein catalogs were functionally annotated using eggNOG-mapper v2.0.1[28], which is based on the eggNOG protein database v5.0[56]. The resultant annotations include eggNOG orthologs and functional terms from several databases, including Gene Ontology (GO)[57] and Kyoto Encyclopedia of Genes and Genomes (KEGG)[58]. Further, for each protein cluster, taxonomic origins of all member proteins and the lowest common ancestor of the cluster were tracked and annotated.

The numbers of shared species and shared phyla of proteins in the HRGM-50 protein catalogs were annotated based on the taxonomic annotation of member proteins. The number of shared species was binned at the bin size of 10, then the annotation rate for each protein bin was calculated as the number of annotated proteins divided by the number of proteins in the bin.

Reconstruction of the phylogenetic tree

For the bacterial and archaeal genomes, 120 and 122 universal marker genes, respectively, were predicted by the GTDB-Tk[18]. Using the concatenated sequences of marker genes, the maximum-likelihood tree was generated using IQ-TREE[55]. The phylogenetic tree of bacterial genomes was visualized using iTOL[59].

Kraken2 databases

The Kraken2 v2.0.8-beta[17] custom database for the HRGM representative genomes was prepared based on the taxonomic annotations in GTDB-TK[18]. When two or more genomes were annotated to the same taxon, they were discriminated at the succeeding lower rank. For example, if *genome a* and *genome b* were both annotated to *species_A*, *genome a* and *genome b* were annotated as *Species_A;strain_1* and *Species_A;strain_2*, respectively. By doing so, the user can select a taxonomic rank, thereby measuring species abundances together or individually.

The Kraken2 database for the UHGG[11] was downloaded from UHGG FTP on March 6, 2020. The Kraken2 standard database was downloaded and constructed using "kraken2-build --standard" command on July 14, 2020.

Measuring taxonomic classification rate of sequencing reads

WMS data were compiled for publicly available data for 926, 54, and 26 fecal samples from the US[16], Cameroon[60], and Luxembourg[61,62], respectively. WMS data for 16 fecal samples collected from Korea, which were not included in the HRGM, were also used. These 1,022 fecal samples were neither used for the UHGG nor for the HRGM. The data were pre-processed and taxonomically classified using Kraken2 with standard database, UHGG-based database, and HRGM-based database. The taxonomic classification rate was then calculated based on the proportion of aligned sequence reads in a sample with respect to the database.

Measurement of functional classification rate of sequencing reads

The functional classification rate of sequencing reads was determined based on the number of aligned reads against the protein catalog. For the analysis, WMS data were randomly selected for ten fecal samples from the Cameroon, Korea, US, and Luxembourg cohorts (the same samples were used for the taxonomic classification assessment). After pre-processing, 40 samples were aligned with the UHGP-95 and HRGM-95 protein databases using blastx of DIAMOND v0.9.35.136 [63]. The results were filtered at >80% query coverage (read coverage) and >95% alignment identity thresholds. A pair of reads was treated as two independent reads. For multiple alignments of a read, only the best alignments by bit score and e-value were considered.

Finding the optimal sequencing depth for gene-level analysis of the gut microbiome

For five Korean fecal samples, WMS data generated at a sequencing depth of >60 Gbp, the reads were aligned against the HRGM-95 protein database using blastx of DIAMOND[63]. Alignment results with >80% read coverage and 80% identity were included in further analysis. For each sample, the number of detected genes with at least one aligned read was counted by iteratively removing 1000 randomly selected reads. The number of the detected genes for a given sequencing depth exhibited a saturation curve. The curve fitted well ($R^2 >$ 0.99 for all samples) the two-site binding model[34]. The required sequencing depth for a given gene coverage was determined based on the estimated maximum number of genes according to the equation.

Evaluation of the effect of sequencing depth on *de novo* genome assembly

Nine Korean samples with sequencing depth of >52.5 Gbp (**Supplementary Table 2**) were selected for analysis. Then, 0.5, 2.5, 5, 10, 20, 40, 80, 125, and 175 million read pairs were randomly sampled from each of these samples. As the average read-pair length was 300 bp, the sequencing depths of these random samples corresponded to 150 Mbp, 750 Mbp, 1.5 Gbp, 3 Gbp, 6 Gbp, 12 Gbp, 24 Gbp, 37.5 Gbp, and 52.5 Gbp, respectively (**Supplementary Fig. 2**). For the 81 simulated samples (9 samples × 9 depths), *de novo* genome assembly was performed using the same pipeline as that used for the database construction.

Two adjacent sequencing depths (e.g., 125 *vs.* 175 million read pairs) were compared to evaluate the effect of sequencing depth on the *de novo* genome assembly. Samples with a greater sequencing depth may yield more MAGs with over 50% completeness, yet with a lower average quality, than those with a lower sequencing depth because of MAGs that barely pass the completeness threshold. Therefore, instead of the average quality scores of all assembled genomes, two genomes assembled at different sequencing depths for the same species clusters were compared. Mash[50] clustering of genomes from two random samples was performed for a comparison based on the average-linkage–based hierarchical clustering, at a threshold of 0.1 (90% identity). Mash clustering was sufficient for clustering conspecific genomes in the simulated samples. Indeed, no cluster had more than two genomes from the same sequencing depth. The assembly quality (completeness, contamination, N50, and genome size) of conspecific genomes at adjacent sequencing depths was then compared.

Evaluation of the effect of sequencing depth on taxonomic profiling

To avoid overestimation of performance, WMS data for 16 Korean fecal samples that have not been used for the HRGM construction and generated at a sequencing depth of >24.5 Gbp were used. From each of the 16 samples, 1, 5, 10, 20, 40, 60, and 80 million read pairs that corresponded to 300 Mbp, 1.5 Gbp, 3 Gbp, 6 Gbp, 12 Gbp, 18 Gbp and 24 Gbp, respectively, were randomly sampled. Taxonomic profiling was then conducted using Kraken2 and the HRGM-based database. Based on the hypothesis that profiling of low-abundance taxa is more affected by sequencing depth than abundant ones, the taxonomic features were stratified at eight different levels of relative abundance, ranging from 1e–07 to 1 with every ten-fold increase (**Fig. 5a,b**). *PCC* and *SCC* between the taxonomic profiles at different sequencing depths were then calculated for each group of features for different levels of relative abundance.

Profiling cross-reactivity potential of the gut prokaryotic genomes

Epitope sequences from autoimmune disease-related self-antigen were compiled from IEDB[36]. “Epitope: Linear epitope”, “Antigen: Organism: Homo sapiens”, “Host: Homo sapiens”, and “Disease: Autoimmune Disease” filters from the IEDB web portal were applied. Epitope sequences that required post-translational modification (e.g., citrullination and deamination) and epitopes shorter than five amino acids were removed. Next, 24,461 unique epitope sequences were aligned with the protein sequences encoded by 5,414 species representatives using BLASTP [64]. For meticulous alignment of short peptide sequences, "-word_size 4", "-evalue 10000", and "-max_target_seqs 100000" options were applied. For every epitope-to-gene pairwise alignment, the Alignment Score (*AS*) was calculated, as follows:

$$AS = (\text{match length - gap length}) / \text{epitope length}$$

AS = 1 alignments were used and the number of protein-coding genes of autoimmune disease epitopes was calculated for every representative species. The number of ECGs was positively correlated with the number of genes. Therefore, the number of ECGs was normalized to the number of genes. To identify epitope-enriched taxonomic clades, EGC per gene of each taxonomic group were compared with the entire 5,414 genomes, and Mann–Whitney P-values and fold-change were calculated.

Data availability

Raw metagenomic sequencing data are available from the Sequence Read Archive (accession number will be released upon publication). By accessing the web server, www.mbiomenet.org/HRGM/, users can browse and download all genomes for 5,414 representative species, their annotations, and metadata, including geographical origin, taxonomy, genomic content, and genome statistics. The five classes of protein catalogs, 16S rRNA sequences, and SNVs are also provided with their functional annotation and taxonomic origin.

References

1 Shreiner, A. B., Kao, J. Y. & Young, V. B. The gut microbiome in health and in disease. *Curr Opin Gastroenterol* **31**, 69-75, doi:10.1097/MOG.0000000000000139 (2015).

2 Thursby, E. & Juge, N. Introduction to the human gut microbiota. *Biochem J* **474**, 1823-1836, doi:10.1042/BCJ20160510 (2017).

3 Zou, Y. *et al.* 1,520 reference genomes from cultivated human gut bacteria enable

functional microbiome analyses. *Nat Biotechnol* **37**, 179-185, doi:10.1038/s41587-018-0008-8 (2019).

4 Forster, S. C. *et al.* A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nat Biotechnol* **37**, 186-192, doi:10.1038/s41587-018-0009-7 (2019).

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

6 Browne, H. P. *et al.* Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature* **533**, 543-546, doi:10.1038/nature17645 (2016).

7 Eckburg, P. B. *et al.* Diversity of the human intestinal microbial flora. *Science* **308**, 1635-1638, doi:10.1126/science.1110591 (2005).

8 Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499-504, doi:10.1038/s41586-019-0965-1 (2019).

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

10 Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649-662 e620, doi:10.1016/j.cell.2019.01.001 (2019).

11 Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol*, doi:10.1038/s41587-020-0603-3 (2020).

12 Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158, doi:10.1186/s40168-018-0541-1 (2018).

13 Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol* **3**, 836-843, doi:10.1038/s41564-018-0171-1 (2018).

14 Dhakan, D. B. *et al.* The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches. *Gigascience* **8**, doi:10.1093/gigascience/giz004 (2019).

15 Yachida, S. *et al.* Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat Med* **25**, 968-976, doi:10.1038/s41591-019-0458-7 (2019).

16 Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655-662, doi:10.1038/s41586-019-1237-9 (2019).

17 Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol* **20**, 257, doi:10.1186/s13059-019-1891-0 (2019).

18 Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, doi:10.1093/bioinformatics/btz848 (2019).

19 Thomas, A. M. *et al.* Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat Med* **25**, 667-678, doi:10.1038/s41591-019-0405-7 (2019).

20 Ledford, H. A genetic gift for sushi eaters. *Nature*, doi:10.1038/news.2010.169 (2010).

21 Xu, J. *et al.* Evolution of symbiotic bacteria in the distal human intestine. *PLoS Biol* **5**, e156, doi:10.1371/journal.pbio.0050156 (2007).

22 Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119, doi:10.1186/1471-2105-11-119 (2010).

23 Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659, doi:10.1093/bioinformatics/btl158 (2006).

24 Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069, doi:10.1093/bioinformatics/btu153 (2014).

25 Alcock, B. P. *et al.* CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res* **48**, D517-D525, doi:10.1093/nar/gkz935 (2020).

26 Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res* **47**, W81-W87, doi:10.1093/nar/gkz310 (2019).

27 Seemann, T. barrnap 0.9: rapid ribosomal RNA prediction.

28 Huerta-Cepas, J. *et al.* Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol Biol Evol* **34**, 2115-2122, doi:10.1093/molbev/msx148 (2017).

29 Ye, S. H., Siddle, K. J., Park, D. J. & Sabeti, P. C. Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell* **178**, 779-794, doi:10.1016/j.cell.2019.07.010 (2019).

30 O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-745, doi:10.1093/nar/gkv1189 (2016).

31 Hillmann, B. *et al.* Evaluating the Information Content of Shallow Shotgun Metagenomics. *mSystems* **3**, doi:10.1128/mSystems.00069-18 (2018).

32 Claussen, J. C. *et al.* Boolean analysis reveals systematic interactions among low-abundance species in the human gut microbiome. *PLoS Comput Biol* **13**, e1005361, doi:10.1371/journal.pcbi.1005361 (2017).

33 Benjamino, J., Lincoln, S., Srivastava, R. & Graf, J. Low-abundant bacteria drive compositional changes in the gut microbiota after dietary alteration. *Microbiome* **6**, 86, doi:10.1186/s40168-018-0469-5 (2018).

34 Wang, Z. X. & Jiang, R. F. A novel two-site binding equation presented in terms of the total ligand concentration. *FEBS Lett* **392**, 245-249, doi:10.1016/0014-5793(96)00818-6 (1996).

35 Zhang, X., Chen, B. D., Zhao, L. D. & Li, H. The Gut Microbiota: Emerging Evidence in Autoimmune Diseases. *Trends Mol Med* **26**, 862-873, doi:10.1016/j.molmed.2020.04.001 (2020).

36 Vita, R. *et al.* The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res* **47**, D339-D343, doi:10.1093/nar/gky1006 (2019).

37 Stoll, M. L. *et al.* Altered microbiota associated with abnormal humoral immune responses to commensal organisms in enthesitis-related arthritis. *Arthritis Res Ther* **16**, 486, doi:10.1186/s13075-014-0486-0 (2014).

38 Wang, W. *et al.* Increased proportions of Bifidobacterium and the Lactobacillus group and loss of butyrate-producing bacteria in inflammatory bowel disease. *J Clin Microbiol* **52**, 398-406, doi:10.1128/JCM.01500-13 (2014).

39 Claesson, M. J. *et al.* Gut microbiota composition correlates with diet and health in the elderly. *Nature* **488**, 178-184, doi:10.1038/nature11319 (2012).

40 Scher, J. U. *et al.* Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. *Elife* **2**, e01202, doi:10.7554/eLife.01202 (2013).

41 Paramsothy, S. *et al.* Specific Bacteria and Metabolites Associated With Response to Fecal Microbiota Transplantation in Patients With Ulcerative Colitis. *Gastroenterology* **156**, 1440-1454 e1442, doi:10.1053/j.gastro.2018.12.001 (2019).

42 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).

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

44 Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* **27**, 824-834, doi:10.1101/gr.213959.116 (2017).

45 Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: an ultra-fast

single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674-1676, doi:10.1093/bioinformatics/btv033 (2015).

46 Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359, doi:10.7717/peerj.7359 (2019).

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

48 Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat Methods* **11**, 1144-1146, doi:10.1038/nmeth.3103 (2014).

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

50 Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**, 132, doi:10.1186/s13059-016-0997-x (2016).

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

52 Jain, C., Rodriguez, R. L., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**, 5114, doi:10.1038/s41467-018-07641-9 (2018).

53 Marcais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol* **14**, e1005944, doi:10.1371/journal.pcbi.1005944 (2018).

54 Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12, doi:10.1186/gb-2004-5-2-r12 (2004).

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

56 Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research* **47**, D309-D314, doi:10.1093/nar/gky1085 (2018).

57 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29, doi:10.1038/75556 (2000).

58 Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44**, D457-462, doi:10.1093/nar/gkv1070 (2016).

59 Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127-128, doi:10.1093/bioinformatics/btl529 (2007).

60 Lokmer, A. *et al.* Use of shotgun metagenomics for the identification of protozoa in the gut microbiota of healthy individuals from worldwide populations with various industrialization levels. *PLoS One* **14**, e0211139, doi:10.1371/journal.pone.0211139 (2019).

61 Schmidt, T. S. *et al.* Extensive transmission of microbes along the gastrointestinal tract. *Elife* **8**, doi:10.7554/eLife.42693 (2019).

62 Heintz-Buschart, A. *et al.* Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat Microbiol* **2**, 16180, doi:10.1038/nmicrobiol.2016.180 (2016).

63 Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59-60, doi:10.1038/nmeth.3176 (2015).

64 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).

Competing interests

The authors declare no competing interests.

Author contributions

CYK and IL conceived this study. CYK and ML constructed the catalog and performed bioinformatics analysis. SY constructed the web server. KK, DY, and HRK organized the study cohorts and provided the fecal samples. IL supervised the project. CYK, ML, and IL wrote the manuscript.

Acknowledgments

This research was supported by the National Research Foundation funded by the Ministry of Science and ICT (2018R1A5A2025079, 2018M3C9A5064709, 2019M3A9B6065192) to IL. We appreciate the assistance from the KOBIC Research Support Program.

Figure legends

Fig. 1 | Effect of sequencing depth on *de novo* genome assembly. a, Sequencing depth of samples from Korea, Japan, and India. Red data points, nine samples used for the generation of simulated samples for different sequencing depths. **b,** Total read length of samples from Korea, Japan, and India. **c,** The average number of HQ and MQ genomes (left axis) and the proportion of HQ genomes (right axis) from nine samples. **d,e,** Completeness (d) and N50 (e) of assembled genomes from lower sequencing depth (left box of each column) and greater sequencing depth (right box of each column). **f,** The number of the assembled genomes from Korea, Japan, and India. **g,** Total number of the assembled genomes from Korea, Japan, and India, and genome assembly yields. **h,** The relative abundance of 224 Korea-specific, 338 Japan-specific, and 18 India-specific assembled genomes in independent fecal samples from the US (n = 926). *P*-values were calculated by two-sided Mann–Whitney U test (**: $P < 0.01$; ***: $P < 0.001$).

Fig. 2 | Phylogenetic tree of 5,386 representative genomes of prokaryotic species from the human gut contained in the HRGM. Maximum-likelihood phylogenetic tree reconstructed from 120 bacterial marker genes (**Methods**). Representative genomes were annotated by their isolated genome availability (1st layer from the inside), phylum classification (2nd layer), whether they were from UHGG or assembled from KIJ samples (3rd layer), 16S rRNA sequence availability (4th layer), and genome completeness (the outermost layer). Red branches represent 410 genomes from the *Bacteroidaceae* family that are enriched in the representative genome set updated by including KIJ samples.

Fig. 3 | SNV density analysis of the relationship between within-species variation and speciation of gut microbes. a, The number of SNVs per kb pair of the aligned region. SNV density is summarized for each phylum. Boxes are sorted by the median. Arc, archaeal phylum. **b,** The phylogenetic tree for Actimobacteriota phylum. Inside annotation indicates the *Collinsella* genus, divided into *Collinsella* with modest phylogenetic dispersion (MD

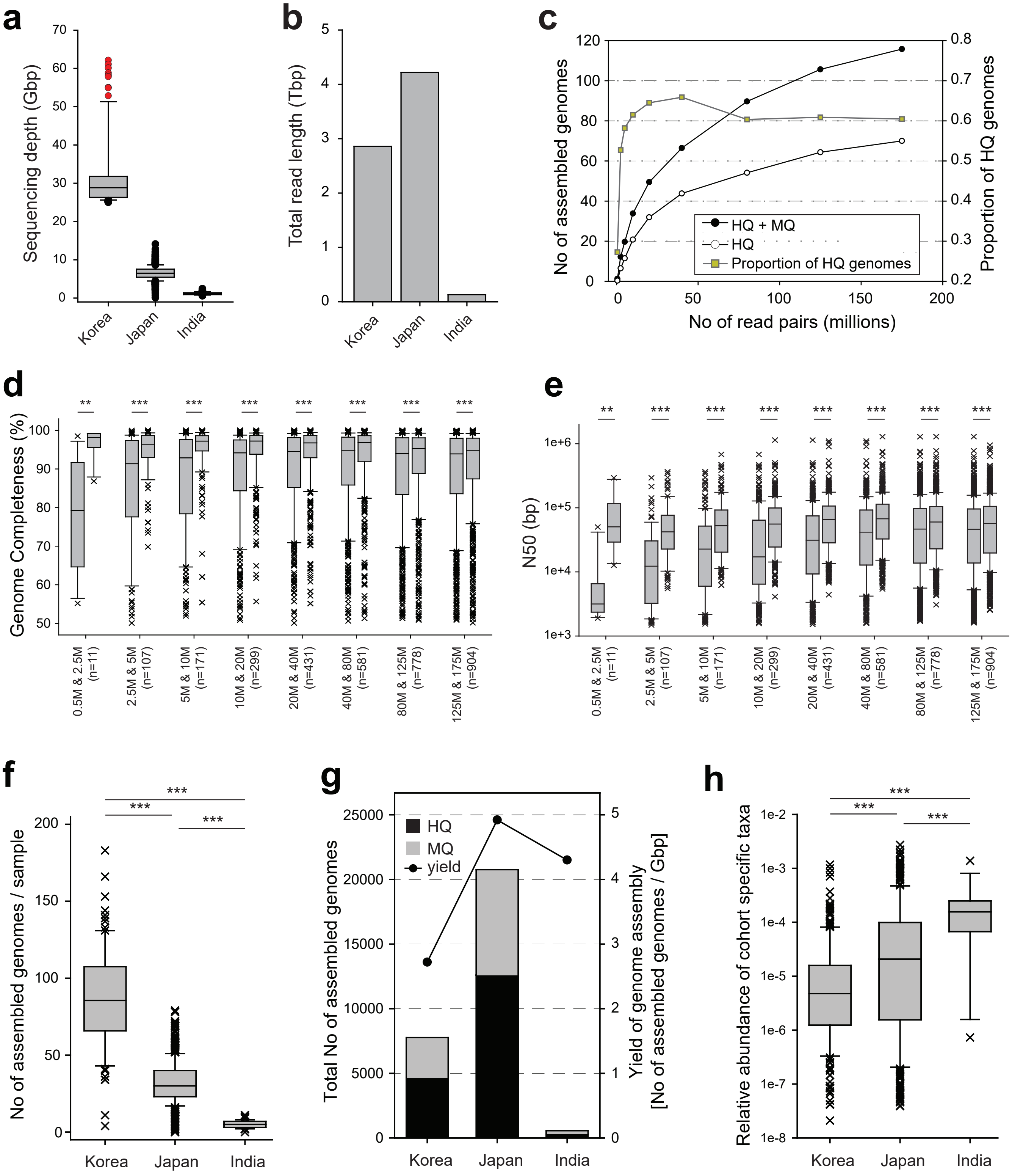
Collinsella, Red) and *Collinsella* with high phylogenetic dispersion (HD *Collinsella*, Orange). Black annotations in the outer circle represent *Collinsella aerofaciens*, *Collinsella aerofaciens_A*, *Collinsella aerofaciens_E*, and *Collinsella aerofaciens_F*, according to the GTDB-TK annotation. **c,** GTDB-TK based taxonomic annotation of MD *Collinsella* and HD *Collinsella*. **d,** SNV density of HD *Collinsella*, MD *Collinsella*, Non-*collinsella* actinobacteriota, and other species. **e,** Scatter plot analysis of SNV density and average phylogenetic distance to the five nearest species of each representative species. Orange points denote species of HD *Collinsella* and black points represent other species. **f,** Comparison of SNV density between the top 10% and bottom 90% species sorted from the lowest average phylogenetic distance to the five nearest species. Statistical significance was calculated by two-sided Mann–Whitney U test (n.s.: not significant; *: $P < 0.05$; ***: $P < 0.001$).

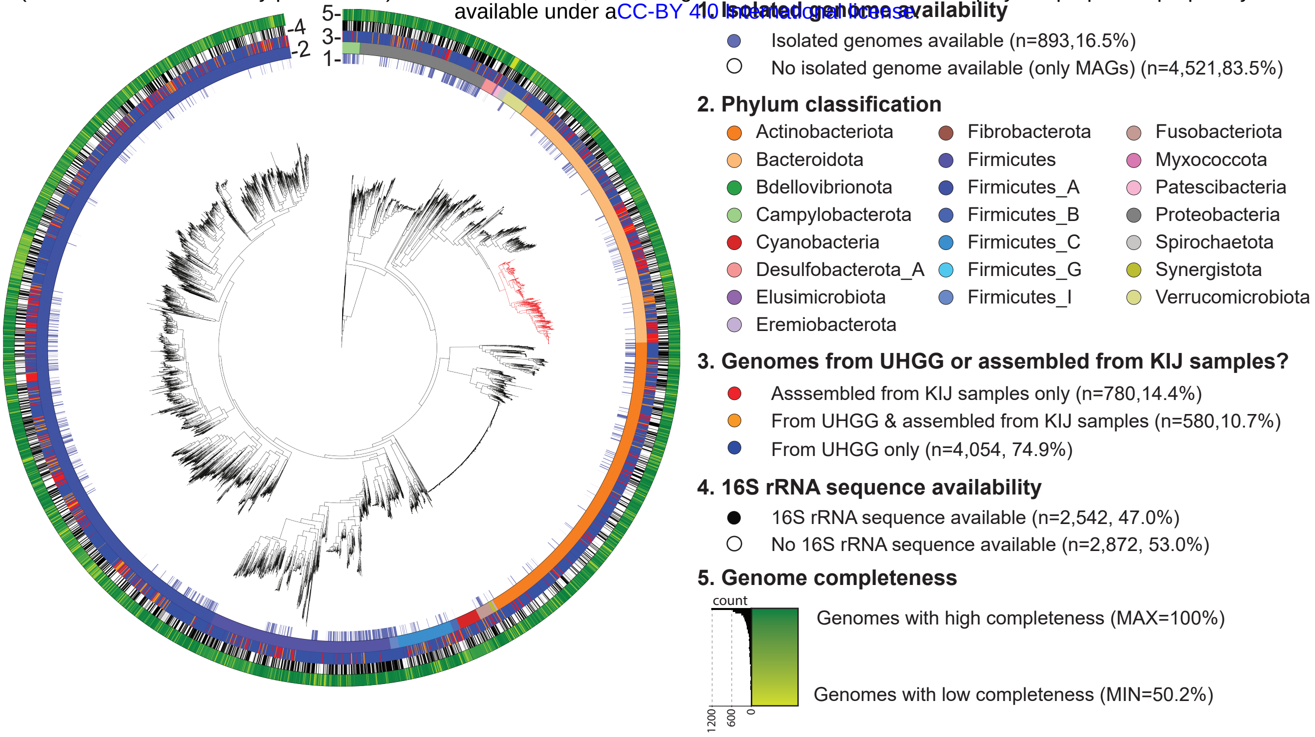
Fig. 4 | Effect of HRGM on taxonomic and functional classification of sequencing reads. a, Proportion of taxonomically classified sequencing reads of WMS data from four different populations. The significance of the improvement was calculated by Wilcoxon signed-rank test. Brown–Forsythe test was used to evaluate the decrease of variance. **b,c,** Percent improvement of the read classification proportion in HRGM-based database compared with the standard database (b) and the UHGG-based database (c). **d,** The number of reads aligned to the UHGP-95 and HRGM-95 protein catalogs. Statistical significance was calculated by using Wilcoxon signed-rank test.

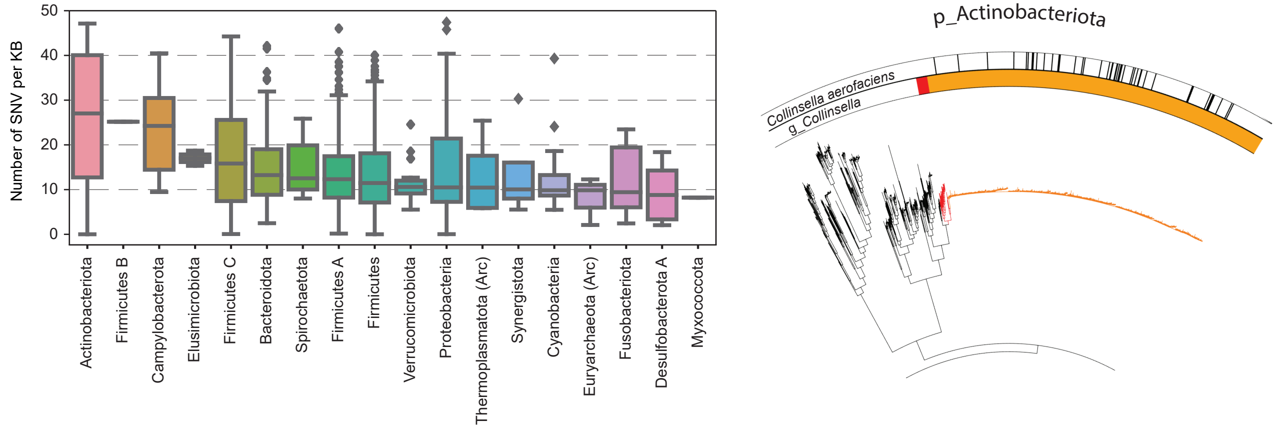
Fig. 5 | Effect of sequencing depth on the reliability of taxonomic profiles. a, The distribution of taxonomic features over different mean relative abundances. **b**, The cumulative proportion of taxonomic features at different thresholds of mean relative abundance. **c,d,** Pearson correlation coefficient (*PCC*) (c) and Spearman correlation coefficient (*SCC*) (d) of the taxonomic profiles at the given sequencing depth and 80M fragments. The x-axis (the mean relative abundance threshold) indicates the upper boundary of the mean relative abundance.

Fig. 6 | Landscape of cross-reactivity potential of gut prokaryotic genomes. a, The number of genes and autoimmune epitope sequence-containing genes (ECG) in 5,414 genomes of species representatives. Red and orange points, species with the top 1% and 5% ECG per gene, respectively. **b,** Volcano plot of the enrichment of ECG density. Taxonomic clades with positive log2 fold-change and P < 1e–5 are highlighted with different colors.

Taxonomic clades denoted by the same color have an inclusive relationship (e.g., *g_Prevotella* belongs to *f_Bacteroidaceae*), with the exception of p_Bacteroidota, c_Bacteroidia, and o_Bacteroidales. The first character of each clade name indicates the taxonomic levels (p: phylum; c: class; o: order; f: family; and g: genus). **c,** The red-highlighted area from (b). **d,** Maximum-likelihood phylogenetic tree with taxonomic annotations of clades with high ECG density. The first layer represents clades with the top 1% (red) and 5% (orange) ECG density [annotations and color designations are the same as in (**a**)]. The second and third layers represent enriched taxonomic clades in the volcano plot [taxonomic annotations and color designations are the same as in (b) and (c)]. The second layer represents above-genus level annotations. The third layer represents genus-level taxonomic clades.


a
Sequencing depth (Gbp)
0
10
20
30
40
50
60
70
Korea
Japan
India
b
Total read length (Tbp)
0
1
2
3
4
5
Korea
Japan
India
c
No of assembled genomes
0
20
40
60
80
100
120
Proportion of HQ genomes
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
50
100
150
200
No of read pairs (millions)
HQ + MQ
HQ
Proportion of HQ genomes
d
Genome Completeness (%)
50
60
70
80
90
100
**
***
***
***
***
***
***
***
0.5M & 2.5M (n=11)
2.5M & 5M (n=107)
5M & 10M (n=171)
10M & 20M (n=299)
20M & 40M (n=431)
40M & 80M (n=581)
80M & 125M (n=778)
125M & 175M (n=904)
e
N50 (bp)
1e+3
1e+4
1e+5
1e+6
**
***
***
***
***
***
***
***
0.5M & 2.5M (n=11)
2.5M & 5M (n=107)
5M & 10M (n=171)
10M & 20M (n=299)
20M & 40M (n=431)
40M & 80M (n=581)
80M & 125M (n=778)
125M & 175M (n=904)
f
No of assembled genomes / sample
0
50
100
150
200
***
***
***
Korea
Japan
India
g
Total No of assembled genomes
0
5000
10000
15000
20000
25000
Yield of genome assembly [No of assembled genomes / Gbp]
0
1
2
3
4
5
HQ
MQ
yield
Korea
Japan
India
h
Relative abundance of cohort specific taxa
1e-8
1e-7
1e-6
1e-5
1e-4
1e-3
1e-2
***
***
***
Korea
Japan
India

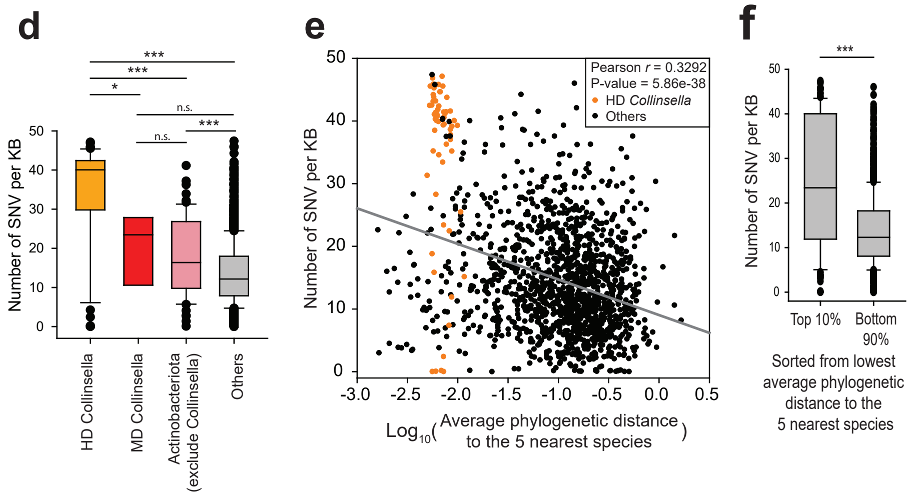

5-
3-
1-
-4
-2
1. Isolated genome availability
● Isolated genomes available (n=893,16.5%)
○ No isolated genome available (only MAGs) (n=4,521,83.5%)
2. Phylum classification
● Actinobacteriota
● Bacteroidota
● Bdellovibrionota
● Campylobacterota
● Cyanobacteria
● Desulfobacterota_A
● Elusimicrobiota
● Eremiobacterota
● Fibrobacterota
● Firmicutes
● Firmicutes_A
● Firmicutes_B
● Firmicutes_C
● Firmicutes_G
● Firmicutes_I
● Fusobacteriota
● Myxococcota
● Patescibacteria
● Proteobacteria
● Spirochaetota
● Synergistota
● Verrucomicrobiota
3. Genomes from UHGG or assembled from KIJ samples?
● Asssembled from KIJ samples only (n=780,14.4%)
● From UHGG & assembled from KIJ samples (n=580,10.7%)
● From UHGG only (n=4,054, 74.9%)
4. 16S rRNA sequence availability
● 16S rRNA sequence available (n=2,542, 47.0%)
○ No 16S rRNA sequence available (n=2,872, 53.0%)
5. Genome completeness
count
1200
600
0
Genomes with high completeness (MAX=100%)
Genomes with low completeness (MIN=50.2%)

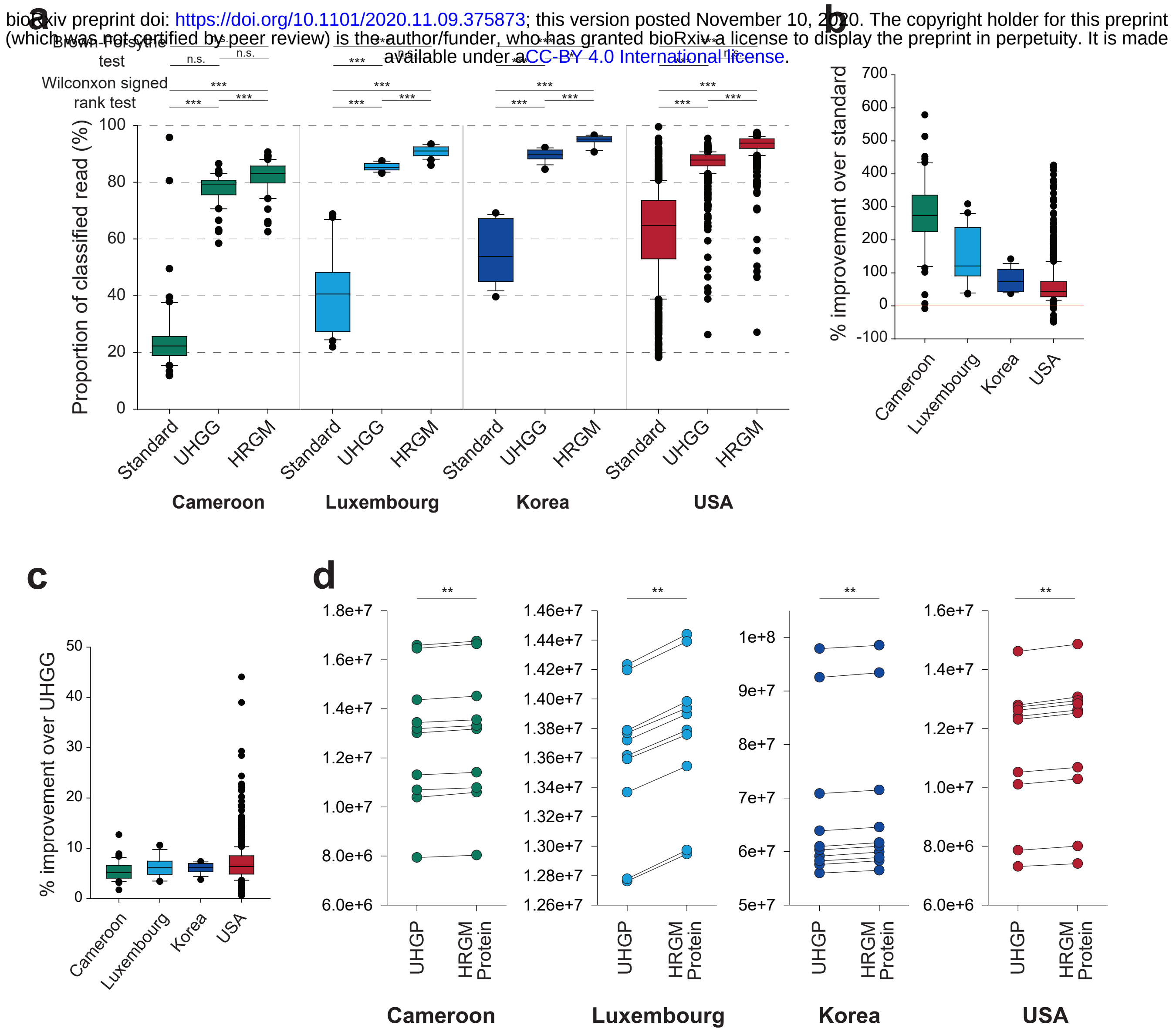


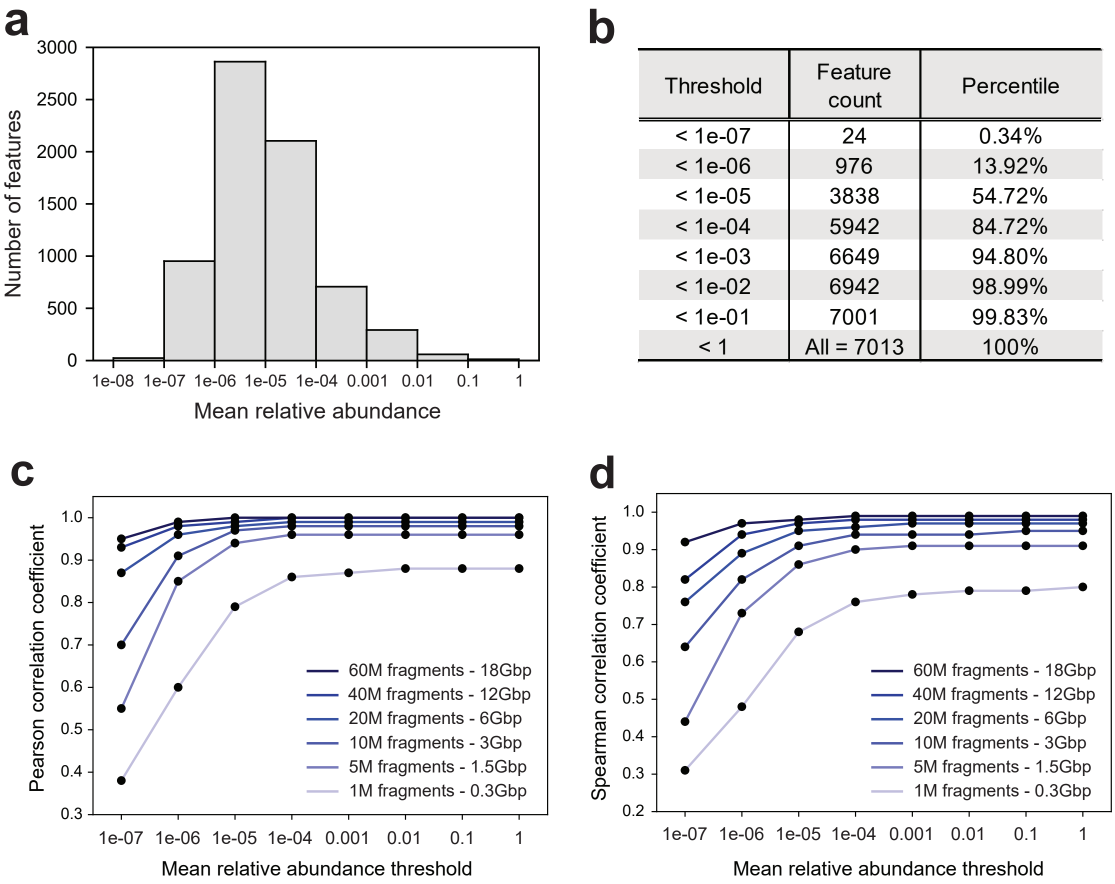
c

Collinsella **with modest diversity (🟥)**	
Species	**Count**
Collinsella vaginalis	1
Collinsella bouchesdurhonensis	1
Collinsella provencensis	1
Collinsella tanakaei	1
Collinsella ihuae	1
Collinsella stercoris	1
Collinsella phocaeensis	1
Collinsella intestinalis	1
Unclassified *Collinsella*	12

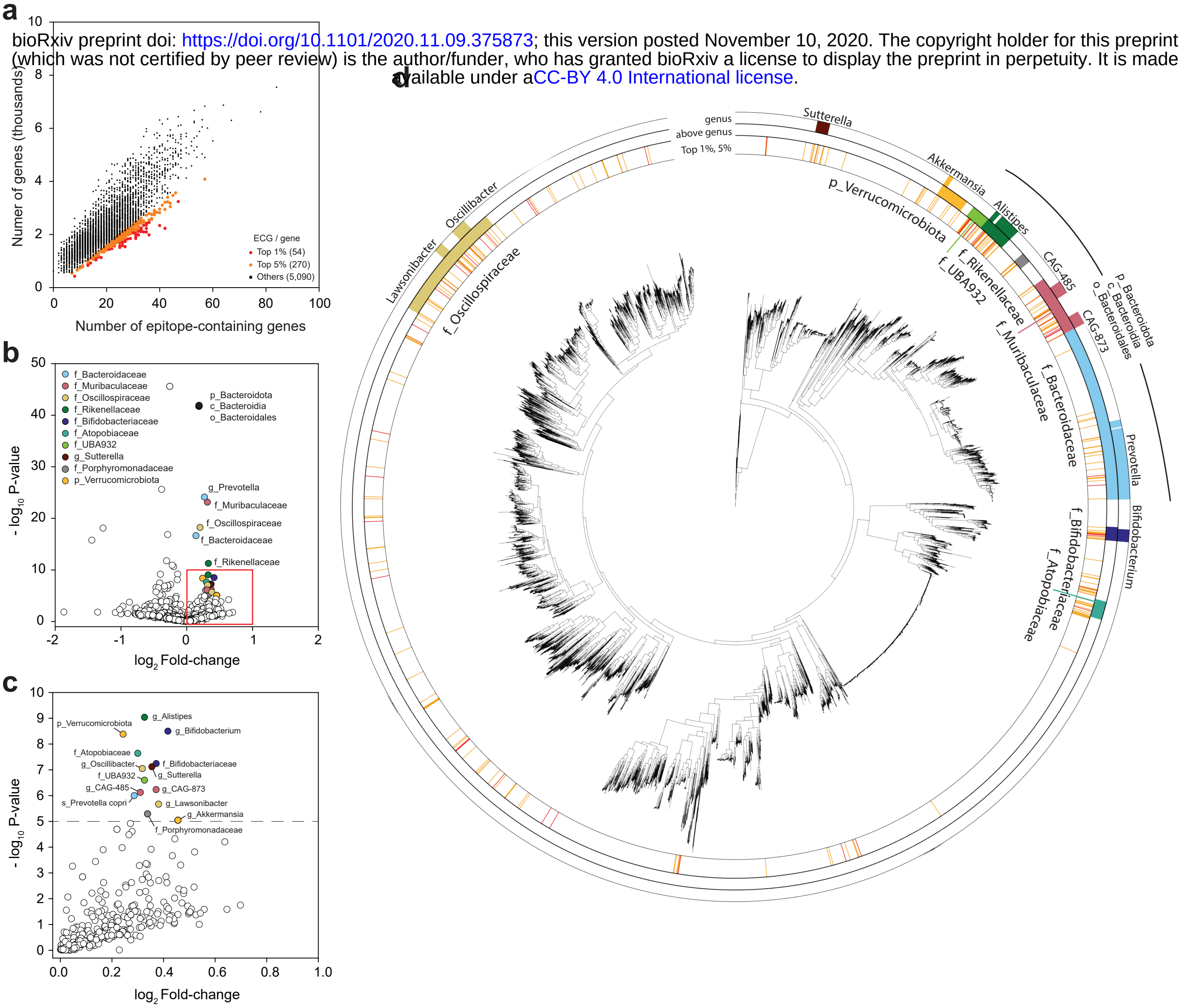
Collinsella **with high diversity (🟧)**	
Species	**Count**
Collinsella aerofaciens	4
Collinsella aerofaciens_A	2
Collinsella aerofaciens_E	1
Collinsella aerofaciens_F	35
Collinsella sp000434535	1
Collinsella sp000763055	1
Collinsella sp002232035	41
Collinsella sp002391315	1
Collinsella sp003487125	2
Unclassified *Collinsella*	531




a
Brown-Forsythe test
n.s.
n.s.
n.s.
***
n.s.
***
***
***
***
n.s.
***
Wilconxon signed rank test
***
***
***
Proportion of classified read (%)
100
80
60
40
20
0
Standard
UHGG
HRGM
Cameroon
Luxembourg
Korea
USA
b
% improvement over standard
700
600
500
400
300
200
100
0
-100
Cameroon
Luxembourg
Korea
USA
c
% improvement over UHGG
50
40
30
20
10
0
Cameroon
Luxembourg
Korea
USA
d
**
1.8e+7
1.6e+7
1.4e+7
1.2e+7
1.0e+7
8.0e+6
6.0e+6
1.46e+7
1.44e+7
1.42e+7
1.40e+7
1.38e+7
1.36e+7
1.34e+7
1.32e+7
1.30e+7
1.28e+7
1.26e+7
1e+8
9e+7
8e+7
7e+7
6e+7
5e+7
1.6e+7
1.4e+7
1.2e+7
1.0e+7
8.0e+6
6.0e+6
UHGP
HRGM Protein
Cameroon
Luxembourg
Korea
USA



Threshold	Feature count	Percentile
< 1e-07	24	0.34%
< 1e-06	976	13.92%
< 1e-05	3838	54.72%
< 1e-04	5942	84.72%
< 1e-03	6649	94.80%
< 1e-02	6942	98.99%
< 1e-01	7001	99.83%
< 1	All = 7013	100%


a
Numer of genes (thousands)
10
8
6
4
2
0
0
20
40
60
80
100
Number of epitope-containing genes
ECG / gene
Top 1% (54)
Top 5% (270)
Others (5,090)
b
50
40
30
20
10
0
-2
-1
0
1
2
- log10 P-value
log2 Fold-change
f_Bacteroidaceae
f_Muribaculaceae
f_Oscillospiraceae
f_Rikenellaceae
f_Bifidobacteriaceae
f_Atopobiaceae
f_UBA932
g_Sutterella
f_Porphyromonadaceae
p_Verrucomicrobiota
p_Bacteroidota
c_Bacteroidia
o_Bacteroidales
g_Prevotella
f_Muribaculaceae
f_Oscillospiraceae
f_Bacteroidaceae
f_Rikenellaceae
c
10
9
8
7
6
5
4
3
2
1
0
0.0
0.2
0.4
0.6
0.8
1.0
- log10 P-value
log2 Fold-change
p_Verrucomicrobiota
g_Alistipes
g_Bifidobacterium
f_Atopobiaceae
g_Oscillibacter
f_Bifidobacteriaceae
f_UBA932
g_Sutterella
g_CAG-485
g_CAG-873
s_Prevotella copri
g_Lawsonibacter
g_Akkermansia
f_Porphyromonadaceae
d
genus
above genus
Top 1%, 5%
Sutterella
Akkermansia
Alistipes
CAG-485
CAG-873
Prevotella
Bifidobacterium
Oscillibacter
Lawsonibacter
p_Verrucomicrobiota
f_UBA932
f_Rikenellaceae
f_Muribaculaceae
f_Bacteroidaceae
p_Bacteroidota
c_Bacteroidia
o_Bacteroidales
f_Bifidobacteriaceae
f_Atopobiaceae
f_Oscillospiraceae