# Exclusive cataloging of near-complete genomes unveils novel functional insights into the human gut microbiome

Junyeong Ma[1#], Nayeon Kim[1#], Jun Hyung Cha[1], Wonjong Kim[1], Chan Yeong Kim[2], Sunmo Yang[1], Samuel Beck[3], and Insuk Lee[1,4*]

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

[2]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany.

[3]Department of Dermatology, Center for Aging Research, Chobanian & Avedisian School of Medicine, Boston University, Boston, USA.

[4]POSTECH Biotech Center, Pohang University of Science and Technology (POSTECH), Pohang 37673, Republic of Korea

[#]These authors contributed equally to this work

*Corresponding author:

Insuk Lee

Department of Biotechnology, College of Life Science and Biotechnology, Yonsei University, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea

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

## Abstract

Understanding the human gut microbiome requires comprehensive genomic catalogs, yet many existing ones lack geographic diversity and include numerous medium-quality metagenome-assembled genomes (MAGs) missing up to 50% of genomic regions, potentially leading to inaccurate functional insights into the microbiome. To address these limitations, we developed an enhanced Human Reference Gut Microbiome (HRGM2), a catalog of near-complete (NC) MAGs (completeness ≥90% and contamination ≤5%) and isolate genomes (collectively 'NC genomes') with extensive geographic sampling. HRGM2 comprises 155,211 non-redundant NC genomes from 4,824 prokaryotic species across 41 countries, representing a 66% increase in genome count and a 50% boost in species diversity over the NC genome-based Unified Human Gastrointestinal Genome catalog. Extensive sampling from Africa revealed previously undetectable Africa-associated phyla, highlighting microbial links between humans and the environment. Given the high species complexity of the human gut microbiota, DNA-based species profiling using HRGM2 outperforms marker-based profiling in accuracy. Profiling genomes with enhanced protein functional annotations provides reliable insights into the microbiome. Notably, with exclusive use of NC genomes, HRGM2 enhances assessments of microbiota metabolic capacities and supports genome-scale metabolic models (GEMs) of entire gut microbiota. These findings offer novel functional perspectives on the human gut microbiome, achievable only through high-quality genome cataloging.

## Introduction

Genome-resolved metagenomics has revolutionized microbiome research by enabling the assembly of uncultured microbial genomes from whole metagenome sequencing (WMS) data, leading to the generation of metagenome-assembled genomes (MAGs)[1]. This technical breakthrough has allowed for the cataloging of genomes from several thousand microbial species in the human gut microbiome, with over 80% of them remaining uncultured[2,3]. Despite significant advances in computational metagenomics, many MAGs still suffer from low completeness and high contamination due to the high complexity of the human gut microbiota. As a result, quality control is essential for initially generated MAGs, typically filtering for medium-quality (MQ) MAGs, defined as those with ≥50% completeness and ≤5% contamination, or for near-complete (NC) MAGs, defined as having ≥90% completeness and ≤5% contamination[4].

Most available genomic catalogs for the human gut microbiome are largely composed of MQ MAGs, with many species clusters consisting exclusively of MQ MAGs. For instance, the Unified Human Gastrointestinal Genome (UHGG)[3], a widely used genomic catalog for the human gut microbiome, contains 204,938 nonredundant genomes, of which 97,157 (47.4%) are MQ MAGs, and only 3,207 out of 4,644 (69%) species clusters include NC MAGs or isolate genomes (collectively referred to as 'NC genomes' hereafter). The reliance on incomplete genomes can lead to false hypotheses regarding the ecological and functional organization of the human gut microbiome. Missing genes in MQ MAGs may result in miscalculations of phylogenetic distances and erroneous conclusions in pangenome analyses. Additionally, species clusters composed exclusively of MQ MAGs are unlikely to represent real biological entities, as such clusters typically have very few MAGs. These limitations highlight the need for a genomic catalog of the human gut microbiome composed entirely of high-completeness genomes to ensure more accurate and reliable microbiome analyses.

Another limitation of current genomic catalogs of the human gut microbiome is their geographic and population bias. For instance, the UHGG catalog is heavily skewed towards data from Western countries and China. The issue of under-represented populations was partially addressed in the previous version of Human Reference Gut Microbiome (HRGM) catalog, which incorporated thousands of MAGs from Korea, Japan, and India[2]. However,

many geographic regions remain under-represented in genomic catalogs of the human gut microbiome. Including MAGs from these under-represented regions could provide novel ecological and functional insights into the diversity and organization of the human gut microbiome.

Here, we present an enhanced version of the HRGM catalog, HRGM2, based on a significantly expanded sample set covering 41 countries, including 10 previously unsampled countries in Southeast Asia, Africa, and South America. This catalog ensures high genome completeness by filtering exclusively for NC genomes. The resulting catalog contains 155,211 non-redundant NC genomes from 4,824 prokaryotic species. Geographic supplementation revealed novel population-associated phyla of the human gut microbiota. Contrary to the previous reports, DNA-based species-level taxonomic profiling of the human gut microbiota has been found to be more accurate with HRGM2 than marker-based profiling, particularly for samples with high species complexity. As expected, a comparison between conspecific MQ MAGs and NC genomes highlighted the limitations of MQ MAGs in taxonomic and functional microbiome analysis. The high completeness of genomes also enables more reliable assessments of metabolic independence[5] and facilitates the inference of metabolic interactions through supporting genome-scale metabolic models (GEMs) for the entire human gut microbiota[6].

## Results

### *Exclusive cataloging of NC genomes from the human gut microbiome in HRGM2*

We collected 5,625 new WMS samples from 31 human gut metagenomic datasets (**Supplementary Tables 1-2**) that were not covered in UHGG and the previous version of HRGM (v1). Contigs were derived from WMS sample reads using a *de novo* genome assembly pipeline, and these contigs were organized into MAGs through an ensemble binning method (**Extended Data Fig. 1a**). By integrating all available sources, we secured a total of 509,047 genomes, which included genomes from the previous version of HRGM (v1, n=316,079)[2], isolated genomes from the Broad Institute-OpenBiome Microbiome Library (BIO-ML, n=3,632)[7], the recently published human Gut Microbial Biobank (hGMB, n=115)[8],

and MAGs assembled from the 5,625 new WMS samples (n=189,221). These MAGs were then subjected to a quality control pipeline (**Extended Data Fig. 1b**), where we detected genomic chimerism using the Genome UNClutterer (GUNC)[9] and filtered out MAGs with clade separation scores (CSS) exceeding 0.45, resulting in 413,014 genomes. Subsequently, we assessed the completeness and contamination of the genomes using CheckM[10], selecting 224,351 near-complete (NC) genomes (completeness ≥ 90% and contamination ≤ 5%).

Lineage-specific workflow of CheckM occasionally underestimated completeness, especially for novel clades, due to the incorrect use of lineage-specific marker genes[11,12]. To address this, we re-evaluated genome quality using a universal bacterial marker gene set[10,13]. For the phylum Patescibacteria, which lacks many universal bacterial markers, we applied the Candidate Phyla Radiation (CPR) marker gene set[14] to reassess genome quality. After this reassessment, 6,281 additional NC genomes were retained, bringing the total to 230,632 NC genomes (**Supplementary Table 3**). After freezing this version of the genomic catalog, CheckM2[12] was published, addressing the issue of missing marker genes for certain lineages through machine learning. Reassessing genomes with CheckM2 confirmed that over 90% of the genomes retained using universal bacterial and CPR markers were still classified as NC genomes (**Extended Data Fig. 1c-d**).

We then dereplicated the 230,632 NC genomes into 4,824 species-level clusters, designating the highest-quality genome as the representative for each species cluster (**Supplementary Table 4**). After removing redundant genomes, we finalized an enhanced version of HRGM, HRGM2, cataloging 155,211 non-redundant NC genomes, representing 4,824 prokaryotic species (4,798 bacterial and 26 archaeal) in the human gut microbiome (**Fig. 1a**). Due to the challenges in recovering rRNA regions in MAGs from short-read sequencing, only 7,348 genomes (4.73% of non-redundant genomes) met the high-quality (HQ) criteria, which require an NC genome to contain 5S, 16S, and 23S rRNAs, as well as tRNAs for at least 18 of the 20 amino acids[4]. Specifically, only 2.08% of short-read sequencing-based MAGs were HQ, whereas 89.6% of HiFi long-read sequencing-based MAGs met the HQ criteria.

### *Extensive sampling reduced geographic bias and revealed Africa-associated phyla*

HRGM2 includes genomes assembled from 18,111 WMS samples collected from 41 countries **(Extended Data Fig. 2)**, which are 1.6 times more samples and 1.5 times more countries compared to the UHGG catalog **(Fig. 1b-c)**. This extensive sampling reduced the geographical bias in preexisting human gut microbiome catalogs. A total of 10,855 MAGs, generated from over 1,000 previously under-sampled WMS samples from Africa, were added to HRGM2. This represents a 9.3-fold increase compared to the 1,168 non-redundant NC genomes for the African gut microbiome in UHGG. This supplementation revealed two phyla enriched with microbial genomes from Africa: Elusimicrobiota and Spirochaetota **(Fig. 1d)**. The phylum Elusimicrobiota, also known as "Termite Group 1", was initially discovered in the gut of insects[15-17]. Subsequent studies have reported the presence of Elusimicrobiota in various habitats, such as groundwater[18], wastewater[19], sewage sludge, contaminated aquifer[20], and bog soil[21]. The phylum Spirochaetota is widely distributed in the environment and is associated with diseases including syphilis and Lyme disease[22,23]. Previous reports have indicated that Spirochaetota is highly abundant in the gut of Hadza hunter-gatherers in Africa[24] and that the relative abundance of Elusimicrobiota and Spirochaetota decreases with industrialization[25]. These geography-associated phyla underscore how environmental and microbial ecosystems influence human microbiome. The association of these phyla with Africa became detectable with the inclusion of novel genomes in HRGM2 (**Fig. 1e**), demonstrating that extensive sampling enabled the discovery of new Africa-associated phyla in the human gut microbiome.

### *Improved genome completeness of the HRGM2 enhances WMS read classification*

Considering only NC genomes, HRGM2 comprises 155,211 genomes from 4,824 species, marking a 1.7-fold and 1.5-fold increase over UHGG, respectively **(Fig. 1f-g)**. As a result, 99% of UHGG species clusters that include at least one NC genome (referred to as 'NC genome species') are also found within HRGM2 **(Fig. 1h)**. Furthermore, 93% and 79% of NC genome species from other prominent catalogs, such as those from the Weizmann Institute of Science (WIS)[26] and the Early-Life Gut Genomes (ELGG)[27], are also captured in HRGM2, despite these catalogs not being incorporated into HRGM2. In addition, HRGM2 adds 1,056 novel

NC genome species that are not present in UHGG. By comparison, only 3,207 of 4,644 species (69%) in UHGG have NC genomes. Notably, 55% of the 1,437 MQ MAG species (species containing only MQ MAGs) in UHGG contain only a single MAG, which may not sufficiently support the existence of these species in the human gut microbiota.

For 3,186 NC genome species shared between UHGG and HRGM2, HRGM2 includes, on average, 2.1 times more NC genomes per species compared to UHGG **(Fig. 1i)**. As a result, the pan-genome size of these conspecific species increased in HRGM2. Among 693 conspecific pairs that contain 10 or more NC genomes in both catalogs, we compared the size of the pan-genome. While the core-genome size remained similar between the two catalogs, the accessory-genome size increased by an average of ~1.3 times in HRGM2 (**Fig. 1j**).

Lastly, we assessed the taxonomic classification rates of WMS reads using the genomic catalogs of human gut microbiome. For this analysis, we aligned 2,624 WMS samples, which were not used in the construction of either catalog (**Supplementary Table 5**), using Kraken2[28] with a confidence score threshold of 0.2. We observed that although the majority of reads were mapped to both UHGG and HRGM2 genomes, at the same taxonomic level, the proportion of reads assigned to a taxonomic class was lower in UHGG compared to HRGM2 (**Fig. 1k**). This lower resolution in taxonomic classification for UHGG could be partially attributed to missing genomic regions in MQ genomes within UHGG. Additionally, when we divided the WMS samples by continent, HRGM2 consistently provided a higher proportion of sequencing reads assigned at the genus level compared to UHGG across all continents (**Fig. 1l**). These findings suggest that HRGM2 will improve the rate of WMS read classification and the accuracy of taxonomic classification for the human gut microbiome across diverse geographic regions.

### *Marker-based species-level taxonomic profiling is limited for fast-evolving species*

The most common application of a genomic catalog of the microbiome is assembly-free taxonomic profiling of given WMS samples. Two main approaches are used for WMS-based taxonomic profiling: DNA-based and marker-based methods. DNA-based methods classify sequencing reads into taxonomic groups using entire genome sequences. To process the large

volumes of genomic data in microbiome studies, ultrafast k-mer-based aligners such as Kraken2[28], often paired with Bracken[29] for species-level reassignments, are commonly used. On the other hand, marker-based methods rely on specific marker genes for each species. MetaPhlAn[30] is a widely used marker-based profiler. Custom databases for Kraken2-based pipelines can be easily created for any genome catalog, which is why resources like UHGG provide Kraken2 databases for representative genomes of cataloged prokaryotic species[3]. However, creating custom databases for MetaPhlAn is more labor-intensive because it requires deriving species-specific marker genes, which likely explains why Kraken2 databases are more widely offered in genomic catalogs.

To compare the performance of DNA-based or marker-based species-level taxonomic profiling with HRGM2, we provided both Kraken2 databases and a species-specific marker gene database. For HRGM2, we created two versions of Kraken2 databases: one based on representative genomes and another based on concatenated genomes for each species cluster. Additionally, we constructed a species-specific marker database for HRGM2 using the methodology of MetaPhlAn4[30] (**Extended Data Fig. 3a**). We identified marker genes for 4,209 of the 4,824 species, covering 87.3% of the cataloged species, with over 200 marker genes identified for 3,007 species (**Extended Data Fig. 3b**). However, species from the genus *Collinsella*, which show overdispersion in both the HRGM2 and UHGG phylogenetic trees, posed a challenge, as we failed to identify marker genes for 584 out of 771 species, representing 95% of the 615 species lacking species-specific marker genes (**Extended Data Fig. 3c**). This is likely due to rapid speciation, which generates distinct species with a high density of single nucleotide variations (SNVs)[2] but without gene-level diversification. The lack of marker genes in the genus *Collinsella* highlights a limitation of marker-based taxonomic profiling for rapidly evolving clades, potentially reducing the accuracy in species-level taxonomic profiling for these groups.

### ***DNA-based species profiling with HRGM2 outperforms marker-based profiling of the human gut microbiota***

To assess the accuracy of the species profiling pipelines, we used simulated WMS samples using HRGM2 genomes. To evaluate the impact of species complexity of the gut microbiota,

we generated simulated samples containing 40, 150, 600, 1,000, and 1,500 species, with each species consisting of either a single strain or multiple strains. We tested two versions of the Kraken2 custom database—one with representative species genomes and another with concatenated species genomes—along with the MetaPhlAn4 custom database based on species-specific marker genes. For DNA-based taxonomic abundance measurements, taxonomic groups with longer genomes may have a higher likelihood of read alignment[31]. Therefore, we also examined the effect of genome size normalization on the accuracy of the profiling. Additionally, since ambiguously assigned reads can significantly affect species abundance estimates[32], we evaluated the impact of adjusting the Kraken confidence threshold by excluding reads with a confidence score below 0.2. These steps allowed us to comprehensively assess the performance of different species-level taxonomic profiling methods across varying levels of species complexity.

In the comparison of computational efficiency, Kraken2-based pipelines were found to be approximately twice as fast as the MetaPhlAn4 pipeline, although MetaPhlAn4 requires significantly less memory (**Fig. 2a**). When we assessed the accuracy of species profiling, MetaPhlAn4 achieved the highest precision, while Kraken2-based pipelines using representative species genomes or operating without a confidence threshold demonstrated the lowest precision. However, the Kraken2-based pipeline utilizing concatenated species genomes with a confidence threshold achieved precision comparable to MetaPhlAn4, particularly in samples containing more than 600 species (**Fig. 2b, top**). This indicates that incorporating the species pangenome and managing false positives are crucial for improving the precision of species-level WMS read classification, especially as species complexity increases.

In contrast to precision, all Kraken2-based pipelines showed an almost perfect recall rate, while MetaPhlAn4 exhibited much lower recall, which declined further with increasing species complexity (**Fig. 2b, middle**). This suggests that using entire genome sequences offers significant advantages over marker genes alone for retrieving species from WMS sample. The lower recall observed in MetaPhlAn4 in highly complex communities may be attributed to the increasing number of species lacking species-specific marker genes such as those in the genus *Collinsella*.

Finally, the F1 score—which balances precision and recall—showed that the Kraken2-based pipeline with concatenated species genomes and a confidence threshold performed best in samples with more than 600 species, while MetaPhlAn4 excelled in less complex communities with 40 and 150 species (**Fig. 2b, bottom**). The effect of strain complexity (whether each species originates from single or multiple strains) was negligible in the overall analysis. Bray-Curtis similarity of species abundance profiles further supported the superiority of Kraken2-based pipeline with concatenated species genomes, genome-size normalization, and a confidence threshold in taxonomic profiling with HRGM2 (**Fig. 2c**). Using a total of 1,160 WMS samples from healthy human guts (**Supplementary Table 2**), we surveyed the number of HRGM2 species in each human gut and found that approximately 89% harbor more than 600 species (**Fig. 2d**). Given the high species complexity of the majority of human gut microbiota, we conclude that the Kraken2-based pipeline using concatenated species genomes, genome-size normalization, and a confidence threshold provides the most accurate species profiles for the human gut microbiota.

***Enhanced annotations of microbial proteins for Gene Ontology and Carbohydrate active enzymes***

Cataloging proteins encoded in microbial genomes and annotating their functions enhances our understanding of the functional capacity of the human gut microbiome. To construct a protein catalog for the human gut microbiome, we first identified 549,278,140 coding sequences from 230,632 redundant NC genomes. Redundancy in these coding sequences was reduced through clustering analysis. Identical sequences were removed before clustering, and the following identity thresholds were applied to define protein clusters: 100%, 95%, 90%, 70%, and 50%. Each threshold yielded approximately 87.0, 18.2, 12.4, 6.5, and 3.9 million protein clusters, respectively.

We utilized the genome and protein catalog of HRGM2 to predict functional categories **(Extended Data Fig. 4a)**. A common tool for protein functional annotation is eggNOG-mapper[33], which employs an orthology-based approach. This tool provides annotation of several functional terms, including KEGG Orthology (KO)[34], Gene Ontology (GO)[35], Protein families (PFAM)[36], Carbohydrate-active enzyme (CAZyme)[37] and Cluster of Ortholog

Groups (COG)[38]. However, due to the low annotation rate for some functional terms, especially GO, we explored deep learning-based GO predictions using DeepGOPlus[39]. This resulted in a significant increase—5.2, 18.7, and 5.9 folds—for biological process, cellular component, and molecular function annotation, respectively, compared to eggNOG-mapper **(Fig. 3a)**. Assuming microbial species within the same taxonomic clade share similar functional profiles, we assessed annotation accuracy by examining the taxonomic homogeneity of genomes clustered by their GO term profiles. We found that the GO profiles generated by DeepGOPlus outperformed those produced by eggNOG-mapper **(Fig. 3b)**.

CAZyme play a crucial role in the gut by breaking down various carbohydrates, including dietary fiber, starches, and polysaccharides, into various short-chain fatty acids (SCFAs) that impact host health[40]. Additionally, some CAZymes can degrade mucin, which protects the intestinal epithelium from pathogens. The annotation of CAZymes in HRGM2 proteins was enhanced using dbCAN3[41], which combines Hidden Markov Models (HMMs) and sequence alignment approaches for robust annotation. On average, the phyla Verrucomicrobiota and Bacteroidota exhibited the highest number of CAZyme gene families **(Extended Data Fig. 4b)**. However, CAZyme distribution varies more distinctly at lower taxonomic levels among different gut microbial species **(Fig. 3c)**. Within the CAZyme families, we identified the genus with the highest number of glycoside hydrolases (GHs) **(Extended Data Fig. 4c)**, enzymes critical for carbohydrate degradation[42]. In the Firmicutes phylum, genera from the Paenibacillaes order exhibited a high abundance of GH CAZymes, while in the Firmicutes_A phylum, genera such as *Robinsoniella*, *Murimonas*, and *Hungatella* showed a notable prevalence of these enzymes.

Given that CAZymes are primarily involved in the synthesis and degradation of carbohydrates, their activity is influenced by dietary habits[43]. Since HRGM2 was constructed using samples from various continents, we hypothesized that subspecies-level groups with distinct CAZyme profiles could emerge within a species based on diet style differences. Indeed, we identified several species with distinct CAZyme profiles between Western and non-Western countries **(Figure 3d, Extended Data Fig. 4d)**. Notably, *Faecalibacillus intestinalis* exhibited the most pronounced differences, with strains from non-Western countries containing more GH1, GH140, GH128, and GH55 enzymes, which are associated

with dietary fiber degradation[44] **(Extended Data Fig. 4e)**.

In summary, we enhanced the functional annotation of HRGM2 proteins using deep learning and probabilistic modeling approaches. In addition, CAZyme profiles of NC genomes revealed population structures between Western and non-Western countries within a species, suggesting that dietary adaptation is a key driver of genomic diversification in the human gut microbiota.

### ***NC genomes in HRGM2 enable reliable surveys of the human gut resistome and defensome***

Gut bacteria have evolved systems to cope with external threats, such as viruses and antibiotics. Due to their survival advantages, these systems are known to actively engage in horizontal gene transfer (HGT)[45,46]. Given that HRGM2 contains only NC genomes, it offers more reliable surveys of antibiotic and antiviral resistance gene distribution between core and accessory genomes. If a resistance system is found in the core genome, it suggests inheritance and widespread presence among species genomes. In contrast, its presence in accessory genome indicates acquisition through HGT. Using RGI[47], we predicted resistance genes from the pangenomes of 1,042 species with more than 10 NC genomes in HRGM2, focusing on phyla that have more than five species. We assessed the percentage of genomes within each species that carry specific antibiotic resistance genes (**Fig. 3e, Supplementary Table 6**). Our results confirm previous findings[48,49], demonstrating that many species within the Proteobacteria phylum exhibit resistance to multiple antibiotics, often attributed to genes that confer resistance through antibiotic efflux mechanism.

Species within the Actinobacteriota phylum exhibited a distinct pattern of antibiotic resistance, with certain lineages showing resistance to tetracycline and rifamycin. The rifamycin-resistant lineages were identified as *Bifidobacterium*, consistent with previous studies indicating that *Bifidobacterium adolescentis* is resistant to rifamycin through rpoB mutations [50]. Additionally, the tetracycline-resistant lineages were identified as belonging to the *Collinsella* genus, which harbors a ribosomal protection protein known as Tet(Q). In the Bacteroidota phylum, while few species exhibited antibiotic resistance in their core genome,

we found that many species possess resistance to multiple antibiotics in their accessory genome.

Given the potential for antibiotic-resistant species to transfer resistance to other bacteria through HGT, we focused on identifying species in which most genomes harbor multiple types of antibiotic resistance genes within their core genome. The majority of these species were from the genera *Escherichia*, *Salmonella*, *Klebsiella*, and *Enterobacter*, all of which belong to the Proteobacteria phylum. However, beyond these Proteobacteria, only a few species were found to carry multiple types of antibiotic resistance genes. For example, *Staphylococcus epidermidis*, known as methicillin-resistant *Staphylococcus epidermidis* (MRSE)[51] and *Staphylococcus aures*, known as methicillin-resistant *Staphylococcus aureus* (MRSA)[52], from Firmicutes phylum possess 10 and 6 types of resistance genes, respectively, in their core genome, rendering them resistant to multiple antibiotics. Additionally, the lesser-known bacterium BX12 sp014333425, belonging to the Firmicutes_A, was found to possess five types of antibiotic resistance genes in its core genome.

NC genomes in HRGM2 enabled reliable identification of antibiotics with the highest prevalence of resistance genes in the human gut microbiome. Notably, tetracycline resistance genes were the most prevalent, with 561 (53.8%) out of 1,042 species carrying resistance genes, at least as part of their accessory genome. This was followed by lincosamide and macrolide antibiotics, with resistance genes present in 278 (26.7%) species.

Finally, we surveyed bacterial defense systems against bacteriophages and found that, excluding the RM and Cas systems, most were typically spread as part of the accessory genome **(Fig. 3f)**. However, there were some notable outliers. For example, the Shango system was integrated into the core genome of species within the Actinobacteriota, while the Rst_PARIS system (Restriction of Short Tandem Repeats by Phage-Activated Repeat-Induced Synthesis) was present in the core genome of the Bacteroidota phylum. Most other defense systems were widely distributed across various bacterial phyla.

### ***MQ genomes may result in unreliable metabolic profiles of the gut microbiota***

The most unique feature of HRGM2 compared to UHGG is its exclusive cataloging of NC

genomes. Among the 4,644 species in UHGG, only 66% contain NC genomes, with 55% of the 204,938 non-redundant genomes being MQ **(Extended Data Fig. 5a)**. By median, 55% of genomes within each species cluster were MQ **(Extended Data Fig. 5b)**. This highlights the lower quality of representative and pangenomes in UHGG compared to HRGM2. The incompleteness of MQ genomes may lead to false negatives in functional profiles. To assess the impact of including MQ genomes in reference catalogs, we analyzed 327 species that contain only MQ genomes in UHGG but have at least one NC genome in HRGM2. We calculated the proportion of non-overlapping genomes by aligning representative genomes of conspecific pairs. The non-overlapping proportion was significantly higher in NC genomes, with 89% of MQ genomes overlapping with NC genomes, while 26% of NC genomes were not observed in conspecific MQ genomes **(Fig. 4a)**. A similar analysis at the protein level, clustering predicted proteins from each representative genome, showed consistent results **(Fig. 4b)**. Both genome coverage and protein-level analyses demonstrated missing portions in MQ genomes compared to conspecific NC genomes.

To investigate the impact of missing genomic regions in MQ genomes on functional analysis of the human gut microbiome, we examined the functions predominantly located in these regions. By predicting KEGG modules[53] for each representative genome, we identified functional units that were mainly detected only in conspecific pairs of NC genomes. Essential biosynthesis modules for amino acids, cofactors, vitamins, and nucleotide metabolism were predominantly missing in conspecific MQ genomes **(Fig. 4c)**, suggesting potential false negatives for core metabolic functions when MQ genomes are used for functional profiling.

We then compared the functional profiles of conspecific MQ and NC genomes within species, focusing on 543 UHGG species with at least 10 non-redundant MQ and NC genomes each. Functional distance between genomes was measured based on gene clusters within species. NC genomes exhibited similar functional profiles, while MQ genomes showed highly varied profiles with significantly greater functional distance among themselves compared to NC genomes **(Fig. 4d)**. This trend was consistent across all species tested **(Fig. 4e)**. A similar systematic bias in functional profiles due to reduced genome completeness was previously reported in the GTDB database as well[54]. This result indicates that functional differences between NC genomes reflect biological variation between strains, whereas the differences

between MQ genomes are largely due to technical variation. The phylogenetic distribution of species used in this analysis suggests these findings are not confined to specific clades **(Fig. 4f)**. These analyses underscore the importance of exclusive cataloging of NC genomes in the human gut microbiome reference.

***HRGM2 provides a reliable landscape of gut microbiota metabolic capacities***

The majority of prokaryotic species in the human gut environment have yet to be cultured, partly due to their metabolic dependence for growth. Understanding the landscape of metabolic capacities of gut microbial species therefore enhances our knowledge of their interactions and colonization. A recent study proposed estimating metabolic independence using 33 KEGG modules, 79% of which are involved in essential metabolite biosynthesis and are enriched in species that colonize effectively after fecal microbiota transplantation[5]. Species containing most of these modules were classified as high metabolic independence (HMI) species, while those lacking them were termed low metabolic independence (LMI) species. Using this framework, we evaluated the metabolic capacities of individual species in the human gut microbiota. As shown in earlier analysis, metabolic capacities based on MQ genomes are often underestimated, frequently classifying species as LMI. However, HRGM2, comprising only NC genomes, significantly reduces false negatives when profiling with the 33 KEGG modules, resulting in a more reliable view of gut microbiota metabolic capacities.

Among the 4,824 species in HRGM2, 751 (15.6%) were identified as HMI species and 388 (8.0%) as LMI species. Notably, UHGG species, many of which are based on MQ genomes, were found to have significantly fewer KEGG modules than HRGM2 species **(Extended Data Fig. 6a)**, leading to a lower proportion of HMI species and higher proportion of LMI species in UHGG **(Extended Data Fig. 6b)**. This finding reaffirms that genomic catalogs including MQ genomes tend to underestimate microbiota metabolic capacities.

Consistent with previous studies[5], HMI species had relatively larger genome sizes, while LMI species had smaller genome sizes **(Extended Data Fig. 6c)**. HMI species were assembled from a more diverse range of countries than non-HMI species **(Extended Data Fig. 6d)**, suggesting that species with high metabolic capacities are better adapted to various gut

environments. Moreover, metabolic independence was significantly associated with availability of isolated genomes **(Extended Data Fig. 6e)**, with 98.7% of LMI species remaining uncultured, likely due to their need for external assistance for growth, given their limited metabolic capacities.

Fifteen HMI clades, primarily composed of HMI species such as Enterobacterales and Lachnospirales, were identified, while seven LMI clades, mainly consisting of LMI species like TANB77 and RF39, were also identified **(Fig. 5a)**. Five of the seven LMI clades had temporary placeholder names and were poorly studied (o__RFN20, o__RF39, f__UBA1242, o__TANB77, and g__UMGS1491). Another LMI clade, the phylum Patescibacteria, is known for its very small genome (about 1 Mbp), largely uncultured status, and limited metabolic potential, indicating a parasitic lifestyle reliant on other bacteria[55]. These poorly understood LMI clades share similar features, suggesting that they may also survive through parasitism, and co-culture with other bacteria could facilitate their isolation.

***HRGM2 supports genome-scale metabolic models (GEMs) for entire gut microbiota***

Understanding metabolic interactions among microbes may enable modulation of microbiome composition and function, thereby influencing host physiology[6]. Modeling these interactions within the gut microbiota is emerging as a key technology for personalized nutrition and healthcare. Metabolic modeling primarily relies on genome-scale metabolic models (GEMs), comprehensive computational models that represent the entire metabolic network of an organism[56]. Traditionally, GEMs were reconstructed for culturable species with complete genomes to predict genetic content. More recently, GEMs have been developed for hundreds to thousands of human gut bacterial strains to study their metabolic interactions[57,58]. The surge in high-quality MAGs has also driven the development of various automated GEM reconstruction methods[59]. Since HRGM2 consists solely of NC genomes, it offers an ideal opportunity for community-scale metabolic modeling through automatic GEM reconstruction. Using the automatic GEM tool CarveMe[60], we constructed GEMs for the 4,824 species representative genomes.

Using these GEMs, we investigated metabolic interactions between gut microbial species.

For the analysis, we employed SMETANA[61], a tool that predicts the metabolic interaction potential of a microbial community using two scores: metabolic resource overlap (MRO) and metabolic interaction potential (MIP). MRO reflects the competition potential based on the similarity of nutrient requirements between species, while MIP indicates the cooperation potential based on the number of metabolites that can be exchanged between species. We validated the accuracy of these predicted interactions using strains that decolonize *Klebsiella pneumoniae* strain Kp-2H7 by competing for carbon sources[62]. Initially, 31 isolate gut bacterial strains were tested for decolonization of Kp-2H7, with 18 strains (F18-mix) showing greater decolonization effects and 13 strains (F13-mix) showing less. We anticipated that F18-mix strains would display lower cooperation potential (MIP) and higher competition potential (MRO) with Kp-2H7 compared to F13-mix strains. Although not statistically significant due to limited sample size, we observed the expected trend: F18-mix strains exhibited lower MIP and higher MRO with Kp-2H7 compared to F13-mix strains **(Extended Data Fig. 6f)**. This finding suggests that GEMs automatically reconstructed using HRGM2 genomes can support reliable predictions of metabolic interactions between microbes.

We focused on the metabolic interactions among 1,000 species most prevalent in the guts of healthy individuals from various countries **(Supplementary Table 7)**. A significant negative correlation was observed between MRO and MIP (**Fig. 5b**). Based on the distribution of MRO and MIP, we defined □cooperative pairs′ and □competitive pairs′. Pairs with MIP > 3 and MRO < 0.667 were classified as cooperative (N=121,679, 24.4%), while pairs with MIP < 3 and MRO > 0.667 were classified as competitive (N=98,686, 19.8%). We then examined the landscape of cooperative and competitive interactions among the 1,000 most prevalent species **(Fig. 5c)**. HMI species were more likely to compete with other HMI species, and LMI species tended to compete with other LMI species. Additionally, species of the same order were predominantly competitive with each other. We calculated the proportion of cooperative and competitive pairs within the same and different clades **(Fig. 5d)**, revealing that species pairs within the same clade were mostly competitive, consistent with previous findings[61]. The competitive relationship became more pronounced as the taxonomic resolution increased, reflecting greater metabolic similarity within closely related clades.

Interestingly, the two LMI clades, RF39 and TANB77 orders, were generally cooperative

with other clades—likely parasitic based on our hypothesis—except when interacting with other LMI clades **(Fig. 5c)**. When examining cooperative pair proportions between these two orders and other orders, we found that Peptostreptococcales, Lachnospirales, and Coriobacteriales had the highest cooperative pair proportions, while Veillonellales had the lowest **(Fig. 5e)**. The cooperative pair proportion remained largely unchanged when restricted to cultured species with available isolated genomes, except for Veillonellales, suggesting that clades with high cooperative pair proportions might help in culturing the RF39 and TANB77 orders. Notably, Peptostreptococcales and Coriobacteriales, which showed high cooperative pair proportions, had very few HMI species (2/29, 6.90%, and 0/73, 0%, respectively) **(Fig. 5c)** and, on average, contained only 58% and 61% of the 33 KEGG modules, respectively. Additionally, Bacteroidales, despite including the HMI clade *Bacteroides* genus, exhibited a low cooperative pair proportion.

We hypothesized that specific metabolites cooperate with TANB77 and RF39 orders. Using KEGG module analysis, we found that species capable of ornithine and arginine biosynthesis had significantly higher cooperative pair proportions with the two clades **(Fig. 5f, Extended Data Fig. 6g)**. Although PRPP biosynthesis was present in most species, including those with cooperative relationships, we did not focus on it due to its ubiquity, except in the two clades **(Extended Data Fig. 6g)**. The product of ornithine and arginine biosynthesis is arginine, which bacteria can use as a source of carbon and nitrogen, for ATP production in anaerobic conditions through the arginine deiminase pathway, or for biofilm regulation[63]. This suggests that TANB77 and RF39 rely on arginine for growth but are unable to synthesize it themselves **(Extended Data Fig. 6g)**, obtaining it from other bacteria. Notably, these analyses of metabolic interactions among microbiota were enabled by the exclusive use of NC genomes in the genomic catalog.

## Discussion

Building a complete reference set of genomes for commensal microbes is essential in advancing our understanding of the human microbiome. Despite significant efforts over the past several years to establish a comprehensive genomic catalog of the human gut

microbiome, challenges such as geographic bias and limited genome quality remain unresolved. In this study, we addressed these limitations by implementing extensive geographic sampling and exclusively using NC genomes in developing an enhanced catalog, HRGM2.

We significantly expanded the geographic range of samples to 41 countries, nearly 1.5-fold increase compared to the largest existing catalog, UHGG, which includes 28 countries. UHGG is heavily biased toward Western countries and China, leading to an underrepresentation of genomes from other geographic regions and limiting its ability to capture the full taxonomic diversity of human gut microbes. Through our extensive sampling, we identified two novel Africa-associated phyla, Elusimicrobiota and Spirochaetota, which highlighted microbial links between humans and their environments. These associations were absent from previous genomic catalogs due to limited sampling from African populations. This finding suggests that biased microbiome sampling may result in an incomplete taxonomic understanding of the human microbiota and that focused sampling of the underrepresented geographic regions and populations could be a promising strategy to further enhance HRGM2.

Another key improvement in HRGM2 is the exclusive use of NC genomes, which enhances downstream functional investigations of the microbiome. Most existing human gut microbiome catalogs include MQ genomes, with missing genomic regions that can lead to false negatives in functional profiles and potentially misrepresent microbiome characteristics. We provided several case studies illustrating the advantages of using only NC genomes for reliable functional profiling. These advantages include: (1) more comprehensive surveys of the resistome and defensome across taxonomic groups, (2) more reliable functional profiles across individual genomes (3) reduced false discovery of LMI microbes, and (4) enabling GEMs for the entire gut microbiota to analyze metabolic interactions. These findings suggest that genomic catalogs exclusively comprising NC genomes yield more accurate functional insights into the human gut microbiome.

To enable HRGM2-based taxonomic profiling with both DNA-based and marker-based methods, we developed custom databases for each approach. Contrary to a previous report[30], DNA-based species profiling using HRGM2 outperformed marker-based profiling for the

human gut microbiota with high species complexity (>600 species). Our survey of HRGM2 species in over a thousand healthy individuals showed that approximately 89% harbor more than 600 species, suggesting the Kraken2/Bracken pipeline as the preferred method for species profiling in the human gut microbiome using HRGM2. Several factors may explain this discrepancy: species-specific marker genes are challenging to generate for fast-evolving species, such as those in the *Collinsella* genus, leading to unresolved classification for these taxa; while the previous study used all-available genome databases for Kraken2, we used a gut-specific genome database, which may reduce false positives; we applied a strict confidence score cutoff of 0.2 for Kraken2 alignments, significantly improving accuracy; and finally, the previous report did not assess species profiling for microbiomes with species complexity over 600, while our population-scale survey indicated such complexity is typical of the human gut microbiota. Therefore, although both taxonomic profiling approaches are available for HRGM2-based species profiling, we recommend the Kraken2/Bracken pipeline with a confidence score threshold of 0.2 and genome size normalization.

HRGM2 has several limitations. First, while all genomes in HRGM2 meet NC criteria, a substantial number of genomes do not meet HQ standards. Although HRGM2 is higher quality than existing catalogs, constructing a truly complete catalog will require MAGs generated using platforms capable of long-read, high-accuracy sequencing, such as HiFi sequencing. Second, there are still underrepresented geographic populations. Our study addressed some gaps by adding MAGs from Africa, Southeast Asia, and South America; however, regions like West Asia, North Africa, and parts of Central and South America remain under-sampled, highlighting the need for further inclusion of MAGs from these areas. Finally, HRGM2 includes only prokaryotic genomes. With increasing interest in viruses and fungi within the human gut, a more comprehensive catalog would require the inclusion of genome sequences from these organisms as well.

## Methods

### *Sample collection and metagenome assembly*

We collected 5,625 whole-metagenomic shotgun sequencing (WMS) samples from 31 human

gut metagenomic datasets, which were not included in the Unified Human Gastrointestinal Genome (UHGG)[3] and the previous version of Human Reference Gut Microbiome (HRGM v1)[2], from various countries. The adapter sequences and low-quality bases of Illumina sequencing samples were removed using Trimmomatic v0.39[64]. Next, we aligned sample reads against the human reference genome (GRCh38.p13) using Bowtie2 v2.3.5[65] and removed the aligned reads as human contaminants. Most of the quality-controlled reads were assembled as contigs using metaSPAdes v3.14.0[66]. However, for a few samples, metaSPAdes did not terminate for unknown reasons, and metaSPAdes was only applicable to paired-end sequencing reads. For samples where metaSPAdes was unavailable, we used MEGAHIT v1.2.9[67] to assemble reads instead **(Supplementary Table 2)**. In the case of PacBio HiFi sequencing samples, we used Minimap2 v2.18-r1015[68] with the *-ax asm20* parameter for human contaminant removal and hifiasm-meta v0.2-r040[69] for assembly.

We employed the ensemble approach to generate genome bins. For each sample, reads were aligned to the assembled contigs with the aligner used to remove human contaminants, and alignment results were converted to sorted BAM using SAMtools v1.10[70]. Binning was then performed using three tools: MetaBAT2 v2.15[71], MaxBin2.0 v2.2.7[72], and CONCOCT v1.1.0[73]. The minimum contig size for binning was set to 1,000 bp, except for MetaBAT2, which requires at least 1,500 bp, so it was set to 1,500 bp. The binning results from the three tools were combined into a more robust bin set using the bin refinement module of MetaWRAP v1.3.1[74].

***Genome collection and bin quality control***

For HRGM2, we integrated genome bins assembled from 5,625 WMS samples with the genomes available in the previous version of HRGM. Additionally, we integrated isolated genomes from the Broad Institute-OpenBiome Microbiome Library (BIO-ML)[7] and a human Gut Microbial Biobank (hGMB)[8], which were recently published and not included in the previous version of HRGM.

To construct HRGM2 as a high-quality human gut microbiome catalog, we exclusively chose near-complete (NC) genomes. Initially, we identified genome chimerism using GUNC

v1.0.1[9]. Genomes with clade separation scores (CSS) greater than 0.45 were excluded as chimeric genomes, where 0.45 serves as the default threshold for GUNC. Subsequently, we calculated the completeness and contamination of the genomes using a lineage-specific workflow of CheckM v1.0.18[10]. We selected genomes with completeness ≥ 90% and contamination ≤ 5% as NC genomes. To identify completeness-underestimated clades, we performed taxonomic annotation for all genomes using GTDB-Tk v1.7.0[75] (database release GTDB R202[76]). Next, we define a clade as a completeness-underestimated clade if more than 95% of genomes within the clade, from phylum to genus, exhibit a completeness below 90%. Genomes within such clades were re-estimated for completeness and contamination by applying either bacterial or archaeal marker set, depending on their respective domain, to the taxonomic-specific workflow of CheckM. Genomes belonging to the phylum Patescibacteria were subject to quality reassessment using the CPR marker set[14]. Genomes meeting the criteria of completeness ≥ 90% and contamination ≤ 5% after reassessment were also considered NC genomes. The quality of these added genomes was validated with CheckM2 v0.1.2[12].

***Species clustering and genome dereplication***

We clustered NC genomes at the species level. Calculating pairwise distances for all genomes is excessively time-consuming, so we initially conducted preliminary clustering. For genomes with the same order according to GTDB taxonomic annotation, pairwise mash distances were calculated using Mash v2.3[77] with the '*-s 10000*' parameter. We performed average-linkage-based hierarchical clustering with a 0.1[3] cutoff to establish preliminary clusters. While the Mash algorithm is known for its speed, it may suffer from accuracy limitations when dealing with genome pairs exhibiting low coverage[78]. Therefore, we calculated average nucleotide identity (ANI) using ANImf[78] for each pair of genomes within each preliminary cluster. To address potential ANI overestimation stemming from local alignment, we imposed a minimum coverage threshold of 0.6[79] for each pair. If the coverage of a genome pair was below 0.6, we considered them to be different genomes. We then conducted average-linkage-based hierarchical clustering with an ANI threshold of 95% to cluster genomes at the species level[2,3,80]. To designate a representative genome within each species cluster, we computed a

genome intactness score ($S$)[2,3], $S = Completeness - 5 \times Contamination + 0.5 \times \log_{10}(N50)$, for each genome and selected the one with the highest $S$ value as the representative. To eliminate duplicate genomes, we performed average-linkage-based hierarchical clustering within each species cluster using ANI 99.9% and coverage 0.81 as thresholds. The coverage threshold of 0.81 was calculated by considering the minimum completeness of the NC genome, which is 90%. This threshold is the product of the minimum completeness of the two genomes.

***Taxonomic annotation of species and construction of the phylogenetic tree***

We performed taxonomic annotation of each species using GTDB-Tk v1.7.0[75] (database release GTDB R202[76]) for a representative genome of each species. GTDB-Tk predicted 120 and 122 universal marker genes for bacterial and archaeal genomes, respectively, and conducted multiple sequence alignments (MSA) for concatenated sequences of the predicted marker genes of the representative genome for each species. Based on the MSA, we constructed a phylogenetic tree using IQ-TREE v2.2.0.3[81]. The phylogenetic tree was visualized using iTOL v6[82].

***High-quality (HQ) genome annotation***

We checked if any of the redundant NC genomes met the HQ criteria. An HQ genome is defined as an NC genome that contains 5S, 16S, and 23S rRNAs, as well as tRNAs for at least 18 out of the 20 possible amino acids. The 5S, 16S, and 23S rRNAs for each genome were predicted using Barrnap v0.9[83]. The parameters “--kingdom bac” and “--kingdom arc” were applied for bacterial and archaeal genomes, respectively. An e-value threshold of 1e-05 was used for the prediction of 16S and 23S rRNAs, while an e-value threshold of 1e-04 was used for the prediction of 5S rRNA, considering its shorter length. For the tRNA search, tRNAscan-SE v2.0.9[84] was used, with the “-B” option applied for bacterial genomes and the “-A” option for archaeal genomes.

### *Identifying NC genome species overlap between HRGM2 and other catalogs*

We investigated the NC genome species (Species containing NC genomes) overlap between HRGM2 and other catalogs using NC representative genomes. A species pair was deemed identical if the ANI > 95% and coverage > 0.6 between the representative genome of HRGM2 and that of another catalog. Due to the computational burden associated with pairwise comparisons for all representative genomes, we only compared representative genomes that have the same order according to the GTDB taxonomic annotation.

### *Construction of pan-genomes for HRGM2 and comparison of pan-genome size with UHGG*

For 2,639 of the 4,824 species that contain more than one non-redundant genome, we constructed pan-genomes using Panaroo v1.3.0[85] with the parameters "--clean-mode strict -c 0.90 -f 0.5 --merge_paralogs --core_threshold 0.90". The core genes were defined as present in at least 90% of genomes, otherwise, accessory genes. We used GFF files generated by Prokka v1.14.6[86] as input to Panaroo. For each genome, Prokka was executed with the parameter "--kingdom Bacteria" or "--kingdom Archaea" depending on the domain.

To compare pan-genome sizes between UHGG and HRGM2, we identified 693 conspecific pairs among species with 10 or more non-redundant NC genomes in each catalog. Subsequently, GFF files were downloaded for all non-redundant NC genomes within these 693 UHGG species, and Panaroo was employed with same parameters to construct pan-genomes for each UHGG species. For each species, we computed the pan-genome size (number of gene clusters), core-genome size (number of core genes), and accessory-genome size (number of accessory genes). These metrics allowed for size comparisons to be conducted for each conspecific pair.

### *Construction of a species-specific marker gene database for HRGM2*

The species-specific marker gene database for HRGM2 was constructed by following the methodology of MetaPhlAn4[30]. Candidates for species-specific marker genes were identified through the clustering method. Initially, gene sequences were collected from the pan-

genomes of 2,639 species with multiple conspecific genomes. For the remaining species with one genome, all protein sequences from that genome predicted by Prokka were collected. The collected protein sequences were clustered into families with 90% identity using the "linclust[87]" function of MMseqs2 v13.45111[88]. Only protein families with lengths between 150 and 1,500 amino acids and averaging fewer than 1.5 copies per genome were included in subsequent steps. We identified core protein families for each species. "Coreness" was defined as the proportion of conspecific genomes within a species that possess a given protein family. A coreness threshold of 60% was set for species with fewer than 100 conspecific genomes, and 50% otherwise. Unique core protein families for each species were then identified. Core protein families that exhibited more than 1% coreness in other species were excluded from consideration. Finally, species-specific marker genes were identified from these candidates by alignment. The DNA sequences of these candidate genes were fragmented into 150 bp segments, which were aligned against all 155,211 non-redundant genomes using Bowtie2 with the "--sensitive" parameter. A candidate gene was considered to be present in a genome if at least one segment aligned to it. Based on this alignment, the coreness and uniqueness of the candidates were re-evaluated using the same criteria, and those still meeting these criteria were selected as species-specific markers. Up to 200 markers per species, which were the least shared with other species and the longest, were selected. Using these optimal markers, a custom database for MetaPhlAn4 was constructed using Bowtie2-build with the "--large-index" parameter.

***Simulation dataset generation for comparing taxonomic profiling methods***

To compare taxonomic profiling methods based on HRGM2, we generated 5 Gbp simulation samples using CAMISIM v1.3[89] with default parameters. Non-redundant genomes from HRGM2 were randomly selected to construct the simulation dataset. The complexity of the simulation community was systematically increased by selecting genomes in quantities of 40, 150, 600, 1,000, and 1,500. Additionally, the simulation data represented each species either as a single strain or as multiple strains. In the single strain scenario, one genome per species was chosen, whereas in the multiple strain scenario, up to five genomes per species were selected. Consequently, 10 different setups were organized based on the combination of five

levels of community complexity and two strain strategies. Four samples were simulated for each setup, resulting in a total of 40 samples.

### *Performance comparison of taxonomic profiling methods*

We compared the performance of taxonomic profiling methods based on HRGM2 using simulation data. For marker-based taxonomic profiling, MetaPhlAn4 v4.0.3 was employed with the HRGM2 custom marker database. For DNA-based taxonomic profiling, Kraken2 v2.1.2[28] and Bracken v2.8[29] were utilized with the HRGM2 custom database. The HRGM2 Kraken2 custom database was constructed in two versions: "Representative" and "Concatenated". The former was built with only the representative genome of each species, while the latter was constructed based on genomes concatenated with non-redundant genomes of each species. Kraken2 was run both with and without a confidence score of 0.2. For DNA-based taxonomic profiling, normalization of species abundance according to genome size was also considered. In the "Representative" version, species read counts were normalized by the representative genome size, whereas in the "Concatenated" version, by the average genome size of the species.

### *Cataloging human gut microbial proteins and functional annotations*

To construct a protein database for the human gut microbiome, we predicted coding sequences from NC genomes using Prokka. Proteins with identical sequences were pre-deduplicated, and clustering was performed at 100% identity using the “linclust” function of MMseqs2 with the "--min-seq-id 1" parameter, with other parameters set as previously described. Subsequently, clustering was sequentially repeated at 95%, 90%, 70%, and 50% identity by adjusting the "--min-seq-id" option to 0.95, 0.90, 0.70, and 0.50, respectively.

The representative protein sequences from the 90% identity clusters were functionally annotated using eggNOG-mapper v2.1.9[33], based on the eggNOG protein database v5.0.2[90]. While eggNOG-mapper provides functional annotations for several terms, we employed additional annotation methods to enhance the overall annotation quality. First, we utilized DeepGOPlus v1.0.1[39], which uses deep learning for gene ontology (GO) prediction, to

annotate representative protein sequences from 90% identity clusters as in eggNOG-mapper. Among the results of GO predictions, only the predictions with a confidence score exceeding 0.3 were retained. Furthermore, we excluded the term ‘cellular anatomical entity (GO:0110165)’, which refers to any structural component within a cell, as it annotated 89% of the protein clusters, leading to inflation of the prediction rate. Since DeepGOPlus was trained on data that includes protein sequences from eukaryotes, some predictions may correspond to functions specific to eukaryotic organisms, potentially leading to mispredictions for bacterial proteins. To address this, we retained only bacteria-associated GO terms in the predicted results. Using the UniRef90[91] database downloaded in September 2023, we identified proteins derived from bacteria, and designated the GO terms associated with these proteins as bacteria-associated GO terms.

Second, we annotated carbohydrate-active enzyme (CAZyme) families using run_dbcan v4.1.4, the standalone tool of dbCAN3[41], which employs a combination of HMMER, DIAMOND, and dbCAN_sub for prediction. We annotated 155,211 non-redundant genomes for the CAZyme families using the output of Prokka. To ensure higher confidence in the annotated results, only protein sequences predicted by at least two of the three approaches were classified into CAZyme families. Additionally, when the annotation results from each method differed, the results by HMMER were prioritized. If HMMER did not provide annotations, the results from dbCAN_sub were used instead.

Finally, we annotated defense functions against external threats, including antibiotics and viruses. Using the pan-genome of Panaroo outputs, we employed RGI v6.0.0[47] with CARD v3.2.5 and WildCARD v4.0.0 to predict antibiotic resistance genes for 4,824 species. Through RGI, each gene family is annotated using the Antibiotic Resistance Ontology (ARO). Certain ARO categories encompass scenarios where multiple genes act in concert to confer antibiotic resistance. Typical examples include the glycopeptide resistance gene cluster (Van) and MexAB-OprM. In these cases, multiple genes work together to confer resistance. This cooperative relation between proteins was inferred by CARD ontology. To avoid overestimating the antibiotic resistance potential of a gene when only one component gene is present, we defined that all genes working together must be present for resistance to function. However, for glycopeptide resistance gene clusters, which are composed of regulatory, core,

and accessory genes, the presence of antibiotic resistance is considered valid when all core genes are detected, even if accessory or regulatory genes are absent.

We analyze protein sequences from all 155,211 non-redundant genomes to predict the bacterial defense system against viruses using DefenseFinder v1.0.9[92]. We used a criterion of 80% to distinguish whether a defense system is present in the core or accessory genome for a species with more than 10 non-redundant genomes. Specifically, if more than 80% of the non-redundant genomes in a species have the system, it is classified as part of the core genome; otherwise, it is classified as part of the accessory genome.

***Comparison of GO prediction for human gut microbial proteins by DeepGOPlus and eggNOG-mapper***

Representative protein sequences from the 90% identity clusters in the protein catalog were functionally annotated by GO term using eggNOG-mapper and DeepGOPlus. For comparison, the results of GO prediction by eggNOG-mapper were filtered by the same method as the DeepGOPlus, leaving only bacteria-associated GO terms. To determine which method resulted in more accurate predictions, we assessed how well the functional GO profiles of each species reflected its taxonomy. Functional GO profiles for each species were constructed from the predicted GO terms for representative genomes of each species by eggNOG-mapper or DeepGOPlus. Then, we used agglomerative clustering with Jaccard distance to cluster the GO profiles generated by each method. Agglomerative clustering requires a hyperparameter to determine the total number of clusters, so we increased this value from 1 to 1,000 to evaluate the homogeneity of the clustering results. We evaluated the performance by checking whether species belonging to the same taxonomic lineage—phylum, order, family, and genus—are grouped into the same cluster.

***Strain level comparison of CAZyme families between Western and non-Western countries***

We first categorized countries in Europe and North America as Western and otherwise as non-Western. To identify species with different CAZyme families between strains from Western and non-Western countries, we analyzed profiles of the CAZyme family for all 155,211 non-

redundant genomes. These profiles indicate the number of CAZyme families present in each genome. Using the PERMANOVA test with Euclidean distance, we found that genomes from Western countries had significantly different CAZyme profiles compared to genomes from non-Western countries within the same species. The ‘adonis2’ function from the ‘vegan’ package in R was used to identify species with significant differences in CAZyme profiles between Western and non-Western genomes.

***Assessing the importance of the NC genomes***

To investigate the importance of NC genomes, we first identified 327 species that lacked NC genomes in UHGG but possessed NC genomes in HRGM2. For consistency, we considered representative genomes from each catalog as belonging to the same species if the ANI > 95% and the maximum coverage > 0.6. We then computed the non-overlapping genome proportion as 1 minus the coverage of each genome. To identify overlapping proteins, we used Prokka to predict proteins and clustered them using the “cluster” function of MMseqs2 with the option "--cov-mode 1 -c 0.8 --kmer-per-seq 80 --min-seq-id 0.50 --cluster-mode 2". The non-overlapping protein proportion was calculated by subtracting the overlapping protein proportion from 1.

To explore functions that are not primarily covered by the MQ genomes, we predicted KEGG modules[53] for each representative genome of the 327 conspecific pairs using METABOLIC v4.0[93]. For KEGG modules detected in at least 100 species, we calculated the proportions observed exclusively in NC genome, solely in MQ genome, and in both.

To compare the functional composition of NC genomes and MQ genomes within species, we selected UHGG species with at least 10 non-redundant MQ genomes and NC genomes each. To reduce computational burden, we randomly selected up to 50 NC and 50 MQ genomes per species, focusing on species where the disparity in the number of NC and MQ genomes was less than twofold to minimize bias. As a result, we analyzed 543 species comprising 41,072 genomes. We downloaded the GFF files for these genomes, used Panaroo to identify gene clusters for each genome, and calculated the functional distance between genomes using the Jaccard distance metric.

### *Assessment of metabolic independence*

The metabolic independence of species was evaluated using 33 KEGG modules referenced in Watson et al[5]. For 4,824 representative genomes, KEGG modules were predicted using METABOLIC v4.0[93]. A representative genome was classified as a high metabolic independence (HMI) species if it possessed more than 80% of these 33 KEGG modules, and as a low metabolic independence (LMI) species if it possessed 20% or fewer. At the clade level, from phylum to genus, clades with 10 or more species where over half of the species were either HMI or LMI were designated as HMI clades or LMI clades, respectively.

### *Reconstructing Genome-scale metabolic models and identifying metabolic interactions*

The genome-scale metabolic models (GEMs) for each non-redundant genome were reconstructed using CarveMe v1.5.2[60] with default parameters. Due to the high computational burden of testing metabolic interactions for all possible species pairs in HRGM2 (11,633,076 pairs), we focused on predicting pairwise metabolic interactions for the top 1,000 species most prevalent in the human gut. To select these prevalent species, we utilized 1,160 WMS samples from healthy individuals, evenly distributed across various countries and collected for HRGM2 **(Supplementary Table 2)**. Based on our recommended taxonomic profiling method (Kraken2/Bracken with the "Concatenated" version database, confidence score of 0.2, and genome size normalization), we calculated relative abundances of species. A species was considered present in a sample if its relative abundance exceeded 1e-06, and its prevalence across all samples was determined. For the top 1,000 species, the prevalence metrics were as follows: max = 100%, min = 28.53%, average = 60.34%, median = 56.25%. Pairwise metabolic interactions for these top 1,000 prevalent species were predicted using SMETANA v1.2.0[61], with GEMs reconstructed from representative genomes and the parameters "--flavor bigg -g --molweight." Based on the distribution of cooperation scores (metabolic interaction potential, MIP) and competition scores (metabolic resource overlap, MRO) for 499,500 pairs, we identified cooperative and competitive pairs. Pairs were classified as cooperative if MIP > 3 and MRO < 0.667, and as competitive if MIP < 3 and MRO > 0.667. The mathematical formulas for MIP and MRO are described in the original SMETANA publication.

**Data availability**

By accessing the web server, www.decodebiome.org/HRGM2/, users can browse and download all genomes for 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.

## Author Contributions

JM, NK, and IL conceived this study. JM and NK constructed the catalog and performed bioinformatics analysis. JHC, WK, and SB contributed to bioinformatic analysis. CYK provides technical and scientific advice. SY constructed the web server. IL supervised the project. JM, NK, and IL wrote the manuscript. All authors read and approved the final manuscript.

## Acknowledgements

This research was supported by the National Research Foundation funded by the Ministry of Science and ICT (2022M3A9F3016364, 2022R1A2C1092062 to I.L.). This work was also supported by the Technology Innovation Program (20022947) funded by the Ministry of Trade Industry & Energy (MOTIE, Korea). The work was supported in part by Brain Korea 21(BK21) FOUR program.

## References

1 Kim, N. *et al*. Genome-resolved metagenomics: a game changer for microbiome medicine. *Exp Mol Med* **56**, 1501-1512 (2024). https://doi.org:10.1038/s12276-024-01262-7

2 Kim, C. Y. *et al.* Human reference gut microbiome catalog including newly assembled

genomes from under-represented Asian metagenomes. *Genome Med* **13**, 134 (2021). https://doi.org:10.1186/s13073-021-00950-7

3 Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol* **39**, 105-114 (2021). https://doi.org:10.1038/s41587-020-0603-3

4 Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* **35**, 725-731 (2017). https://doi.org:10.1038/nbt.3893

5 Watson, A. R. *et al.* Metabolic independence drives gut microbial colonization and resilience in health and disease. *Genome Biol* **24**, 78 (2023). https://doi.org:10.1186/s13059-023-02924-x

6 Heinken, A., Basile, A., Hertel, J., Thinnes, C. & Thiele, I. Genome-Scale Metabolic Modeling of the Human Microbiome in the Era of Personalized Medicine. *Annu Rev Microbiol* **75**, 199-222 (2021). https://doi.org:10.1146/annurev-micro-060221-012134

7 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 (2019). https://doi.org:10.1038/s41591-019-0559-3

8 Liu, C. *et al.* Enlightening the taxonomy darkness of human gut microbiomes with a cultured biobank. *Microbiome* **9**, 119 (2021). https://doi.org:10.1186/s40168-021-01064-3

9 Orakov, A. *et al.* GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biol* **22**, 178 (2021). https://doi.org:10.1186/s13059-021-02393-0

10 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 (2015). https://doi.org:10.1101/gr.186072.114

11 Kim, C. Y., Ma, J. & Lee, I. HiFi metagenomic sequencing enables assembly of accurate and complete genomes from human gut microbiota. *Nat Commun* **13**, 6367 (2022). https://doi.org:10.1038/s41467-022-34149-0

12 Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat Methods* **20**, 1203-1212 (2023). https://doi.org:10.1038/s41592-023-01940-w

13 Wang, Y. *et al.* Genome-centric metagenomics reveals the host-driven dynamics and ecological role of CPR bacteria in an activated sludge system. *Microbiome* **11**, 56 (2023). https://doi.org:10.1186/s40168-023-01494-1

14 Brown, C. T. *et al*. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208-211 (2015). https://doi.org:10.1038/nature14486

15 Geissinger, O., Herlemann, D. P., Morschel, E., Maier, U. G. & Brune, A. The ultramicrobacterium "Elusimicrobium minutum" gen. nov., sp. nov., the first cultivated representative of the termite group 1 phylum. *Appl Environ Microbiol* **75**, 2831-2840 (2009). https://doi.org:10.1128/AEM.02697-08

16 Ikeda-Ohtsubo, W. & Brune, A. Cospeciation of termite gut flagellates and their bacterial endosymbionts: Trichonympha species and 'Candidatus Endomicrobium trichonymphae'. *Mol Ecol* **18**, 332-342 (2009). https://doi.org:10.1111/j.1365-294X.2008.04029.x

17 Zheng, H., Dietrich, C., Radek, R. & Brune, A. Endomicrobium proavitum, the first isolate of Endomicrobia class. nov. (phylum Elusimicrobia)--an ultramicrobacterium with an unusual cell cycle that fixes nitrogen with a Group IV nitrogenase. *Environ Microbiol* **18**, 191-204 (2016). https://doi.org:10.1111/1462-2920.12960

18 Meheust, R. *et al*. Groundwater Elusimicrobia are metabolically diverse compared to gut microbiome Elusimicrobia and some have a novel nitrogenase paralog. *ISME J* **14**, 2907-2922 (2020). https://doi.org:10.1038/s41396-020-0716-1

19 Zhang, L. *et al.* Metagenomic insights into the effect of thermal hydrolysis pre-treatment on microbial community of an anaerobic digestion system. *Sci Total Environ* **791**, 148096 (2021). https://doi.org:10.1016/j.scitotenv.2021.148096

20 Herlemann, D. P., Geissinger, O. & Brune, A. The termite group I phylum is highly diverse and widespread in the environment. *Appl Environ Microbiol* **73**, 6682-6685 (2007). https://doi.org:10.1128/AEM.00712-07

21 Uzun, M. *et al.* Recovery and genome reconstruction of novel magnetotactic Elusimicrobiota from bog soil. *ISME J* **17**, 204-214 (2023). https://doi.org:10.1038/s41396-022-01339-z

22 Seshadri, R. *et al*. Comparison of the genome of the oral pathogen Treponema denticola with other spirochete genomes. *Proc Natl Acad Sci U S A* **101**, 5646-5651 (2004). https://doi.org:10.1073/pnas.0307639101

23 Gupta, R. S., Mahmood, S. & Adeolu, M. A phylogenomic and molecular signature based approach for characterization of the phylum Spirochaetes and its major clades: proposal for a taxonomic revision of the phylum. *Front Microbiol* **4**, 217 (2013). https://doi.org:10.3389/fmicb.2013.00217

24 Carter, M. M. *et al*. Ultra-deep sequencing of Hadza hunter-gatherers recovers vanishing gut microbes. *Cell* **186**, 3111-3124 e3113 (2023). https://doi.org:10.1016/j.cell.2023.05.046

25 Angelakis, E. *et al*. Treponema species enrich the gut microbiota of traditional rural

populations but are absent from urban individuals. *New Microbes New Infect* **27**, 14-21 (2019). https://doi.org:10.1016/j.nmni.2018.10.009

26 Leviatan, S., Shoer, S., Rothschild, D., Gorodetski, M. & Segal, E. An expanded reference map of the human gut microbiome reveals hundreds of previously unknown species. *Nat Commun* **13**, 3863 (2022). https://doi.org:10.1038/s41467-022-31502-1

27 Zeng, S. *et al.* A compendium of 32,277 metagenome-assembled genomes and over 80 million genes from the early-life human gut microbiome. *Nat Commun* **13**, 5139 (2022). https://doi.org:10.1038/s41467-022-32805-z

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

29 Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *Peerj Comput Sci* **3**, e104 (2017). https://doi.org:10.7717/peerj-cs.104

30 Blanco-Miguez, A. *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat Biotechnol* **41**, 1633-1644 (2023). https://doi.org:10.1038/s41587-023-01688-w

31 Sun, Z. *et al.* Challenges in benchmarking metagenomic profilers. *Nat Methods* **18**, 618-626 (2021). https://doi.org:10.1038/s41592-021-01141-3

32 Wright, R. J., Comeau, A. M. & Langille, M. G. I. From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools. *Microb Genom* **9** (2023). https://doi.org:10.1099/mgen.0.000949

33 Cantalapiedra, C. P., Hernandez-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol* **38**, 5825-5829 (2021). https://doi.org:10.1093/molbev/msab293

34 Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, D353-D361 (2017). https://doi.org:10.1093/nar/gkw1092

35 Harris, M. A. *et al.* The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32**, D258-261 (2004). https://doi.org:10.1093/nar/gkh036

36 Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**, D412-D419 (2021). https://doi.org:10.1093/nar/gkaa913

37 Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* **42**, D490-495 (2014). https://doi.org:10.1093/nar/gkt1178

38 Galperin, M. Y. *et al*. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res* **49**, D274-D281 (2021). https://doi.org:10.1093/nar/gkaa1018

39 Kulmanov, M. & Hoehndorf, R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* **36**, 422-429 (2020). https://doi.org:10.1093/bioinformatics/btz595

40 Ney, L. M. *et al.* Short chain fatty acids: key regulators of the local and systemic immune response in inflammatory diseases and infections. *Open Biol* **13**, 230014 (2023). https://doi.org:10.1098/rsob.230014

41 Zheng, J. *et al*. dbCAN3: automated carbohydrate-active enzyme and substrate annotation. *Nucleic Acids Res* **51**, W115-W121 (2023). https://doi.org:10.1093/nar/gkad328

42 Wardman, J. F., Bains, R. K., Rahfeld, P. & Withers, S. G. Carbohydrate-active enzymes (CAZymes) in the gut microbiome. *Nat Rev Microbiol* **20**, 542-556 (2022). https://doi.org:10.1038/s41579-022-00712-1

43 Bhattacharya, T., Ghosh, T. S. & Mande, S. S. Global Profiling of Carbohydrate Active Enzymes in Human Gut Microbiome. *PLoS One* **10**, e0142038 (2015). https://doi.org:10.1371/journal.pone.0142038

44 Ducarmon, Q. R. *et al.* Large-scale computational analyses of gut microbial CAZyme repertoires enabled by Cayman. *bioRxiv*, 2024.2001.2008.574624 (2024). https://doi.org:10.1101/2024.01.08.574624

45 Kogay, R., Wolf, Y. I. & Koonin, E. V. Defence systems and horizontal gene transfer in bacteria. *Environmental Microbiology* **26**, e16630 (2024). https://doi.org:10.1111/1462-2920.16630

46 Tokuda, M. & Shintani, M. Microbial evolution through horizontal gene transfer by mobile genetic elements. *Microbial Biotechnology* **17**, e14408 (2024). https://doi.org:10.1111/1751-7915.14408

47 Alcock, B. P. *et al.* CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res* **51**, D690-D699 (2023). https://doi.org:10.1093/nar/gkac920

48 Gaurav, A., Bakht, P., Saini, M., Pandey, S. & Pathania, R. Role of bacterial efflux pumps in antibiotic resistance, virulence, and strategies to discover novel efflux pump inhibitors. *Microbiology (Reading)* **169**, 001333 (2023). https://doi.org:10.1099/mic.0.001333

49 Hassan, K. A. *et al*. Pacing across the membrane: the novel PACE family of efflux pumps is widespread in Gram-negative pathogens. *Res Microbiol* **169**, 450-454 (2018). https://doi.org:10.1016/j.resmic.2018.01.001

50 Lokesh, D., Parkesh, R. & Kammara, R. Bifidobacterium adolescentis is intrinsically resistant to antitubercular drugs. *Sci Rep* **8**, 11897 (2018). https://doi.org:10.1038/s41598-018-30429-2

51 Tang, B. *et al*. Characteristics of oral methicillin-resistant Staphylococcus epidermidis isolated from dental plaque. *Int J Oral Sci* **12**, 15 (2020). https://doi.org:10.1038/s41368-020-0079-5

52 Turner, N. A. *et al*. Methicillin-resistant Staphylococcus aureus: an overview of basic and clinical research. *Nat Rev Microbiol* **17**, 203-218 (2019). https://doi.org:10.1038/s41579-018-0147-4

53 Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res* **51**, D587-D592 (2023). https://doi.org:10.1093/nar/gkac963

54 Eisenhofer, R., Odriozola, I. & Alberdi, A. Impact of microbial genome completeness on metagenomic functional inference. *ISME Commun* **3**, 12 (2023). https://doi.org:10.1038/s43705-023-00221-z

55 Fujii, N. *et al*. Metabolic Potential of the Superphylum Patescibacteria Reconstructed from Activated Sludge Samples from a Municipal Wastewater Treatment Plant. *Microbes Environ* **37**, ME22012 (2022). https://doi.org:10.1264/jsme2.ME22012

56 Gu, C., Kim, G. B., Kim, W. J., Kim, H. U. & Lee, S. Y. Current status and applications of genome-scale metabolic models. *Genome Biol* **20**, 121 (2019). https://doi.org:10.1186/s13059-019-1730-3

57 Heinken, A. *et al*. Genome-scale metabolic reconstruction of 7,302 human microorganisms for personalized medicine. *Nat Biotechnol* **41**, 1320-1331 (2023). https://doi.org:10.1038/s41587-022-01628-0

58 Magnusdottir, S. *et al*. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat Biotechnol* **35**, 81-89 (2017). https://doi.org:10.1038/nbt.3703

59 Mendoza, S. N., Olivier, B. G., Molenaar, D. & Teusink, B. A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome Biol* **20**, 158 (2019). https://doi.org:10.1186/s13059-019-1769-1

60 Machado, D., Andrejev, S., Tramontano, M. & Patil, K. R. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Res* **46**, 7542-7553 (2018). https://doi.org:10.1093/nar/gky537

61 Zelezniak, A. *et al*. Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proc Natl Acad Sci U S A* **112**, 6449-6454 (2015). https://doi.org:10.1073/pnas.1421834112

62 Furuichi, M. *et al*. Commensal consortia decolonize Enterobacteriaceae via ecological control. *Nature* **633**, 878-886 (2024). https://doi.org:10.1038/s41586-024-07960-6

63 Scribani Rossi, C. *et al*. Nutrient Sensing and Biofilm Modulation: The Example of L-arginine in Pseudomonas. *Int J Mol Sci* **23**, 4386 (2022). https://doi.org:10.3390/ijms23084386

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

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

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

67 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 (2015). https://doi.org:10.1093/bioinformatics/btv033

68 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018). https://doi.org:10.1093/bioinformatics/bty191

69 Feng, X., Cheng, H., Portik, D. & Li, H. Metagenome assembly of high-fidelity long reads with hifiasm-meta. *Nat Methods* **19**, 671-674 (2022). https://doi.org:10.1038/s41592-022-01478-3

70 Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10** (2021). https://doi.org:10.1093/gigascience/giab008

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

72 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 (2016). https://doi.org:10.1093/bioinformatics/btv638

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

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

75 Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925-1927 (2019). https://doi.org:10.1093/bioinformatics/btz848

76 Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* **50**, D785-D794 (2022). https://doi.org:10.1093/nar/gkab776

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

78 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 (2017). https://doi.org:10.1038/ismej.2017.126

79 Saheb Kashaf, S., Almeida, A., Segre, J. A. & Finn, R. D. Recovering prokaryotic genomes from host-associated, short-read shotgun metagenomic sequencing data. *Nat Protoc* **16**, 2520-2541 (2021). https://doi.org:10.1038/s41596-021-00508-2

80 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 (2018). https://doi.org:10.1038/s41467-018-07641-9

81 Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530-1534 (2020). https://doi.org:10.1093/molbev/msaa015

82 Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* **49**, W293-W296 (2021). https://doi.org:10.1093/nar/gkab301

83 Seemann, T. *barrnap 0.9: rapid ribosomal RNA prediction*, <https://github.com/tseemann/barrnap> (2018).

84 Chan, P. P. & Lowe, T. M. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. *Methods Mol Biol* **1962**, 1-14 (2019). https://doi.org:10.1007/978-1-4939-9173-0_1

85 Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol* **21**, 180 (2020). https://doi.org:10.1186/s13059-020-02090-4

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

87 Steinegger, M. & Soding, J. Clustering huge protein sequence sets in linear time. *Nat Commun* **9**, 2542 (2018). https://doi.org:10.1038/s41467-018-04964-5

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

89 Fritz, A. *et al.* CAMISIM: simulating metagenomes and microbial communities. *Microbiome* **7**, 17 (2019). https://doi.org:10.1186/s40168-019-0633-6

90 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 Res* **47**, D309-D314 (2019). https://doi.org:10.1093/nar/gky1085

91 Suzek, B. E. *et al.* UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926-932 (2015). https://doi.org:10.1093/bioinformatics/btu739

92 Tesson, F. *et al.* Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat Commun* **13**, 2561 (2022). https://doi.org:10.1038/s41467-022-30269-9

93 Zhou, Z. *et al.* METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. *Microbiome* **10**, 33 (2022). https://doi.org:10.1186/s40168-021-01213-8

## Figures

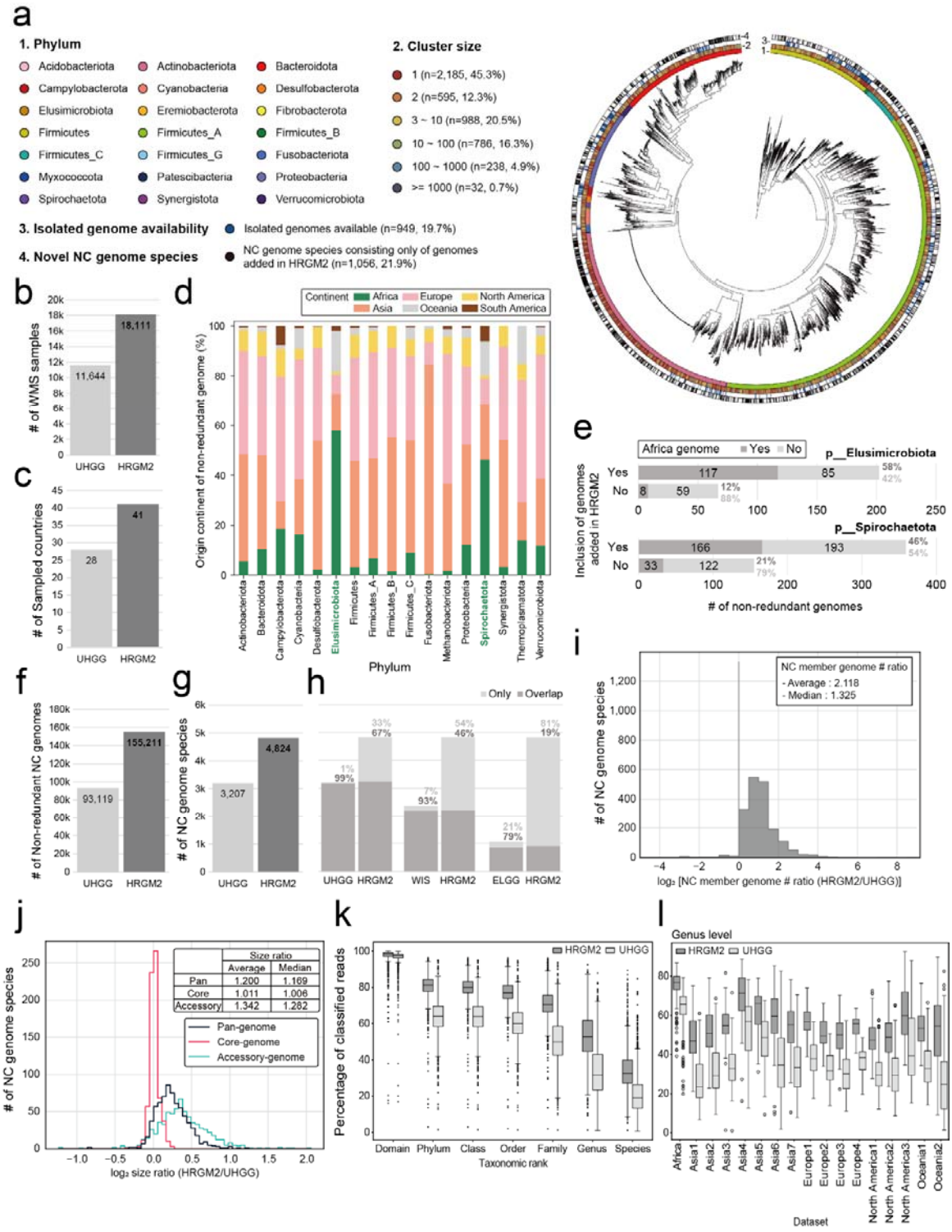

a
1. Phylum
Acidobacteriota
Actinobacteriota
Bacteroidota
Campylobacterota
Cyanobacteria
Desulfobacterota
Elusimicrobiota
Eremiobacterota
Fibrobacterota
Firmicutes
Firmicutes_A
Firmicutes_B
Firmicutes_C
Firmicutes_G
Fusobacteriota
Myxococcota
Patescibacteria
Proteobacteria
Spirochaetota
Synergistota
Verrucomicrobiota
2. Cluster size
1 (n=2,185, 45.3%)
2 (n=595, 12.3%)
3 ~ 10 (n=988, 20.5%)
10 ~ 100 (n=786, 16.3%)
100 ~ 1000 (n=238, 4.9%)
>= 1000 (n=32, 0.7%)
3. Isolated genome availability
Isolated genomes available (n=949, 19.7%)
4. Novel NC genome species
NC genome species consisting only of genomes added in HRGM2 (n=1,056, 21.9%)
b
# of WMS samples
11,644
18,111
UHGG
HRGM2
c
# of Sampled countries
28
41
d
Continent
Africa
Europe
North America
Asia
Oceania
South America
Origin continent of non-redundant genome (%)
Phylum
e
Africa genome
Yes
No
p__Elusimicrobiota
Inclusion of genomes added in HRGM2
p__Spirochaetota
# of non-redundant genomes
f
# of Non-redundant NC genomes
93,119
155,211
g
# of NC genome species
3,207
4,824
h
Only
Overlap
UHGG
HRGM2
WIS
ELGG
i
NC member genome # ratio
- Average : 2.118
- Median : 1.325
# of NC genome species
log₂ [NC member genome # ratio (HRGM2/UHGG)]
j
Size ratio
Average
Median
Pan 1.200 1.169
Core 1.011 1.006
Accessory 1.342 1.282
Pan-genome
Core-genome
Accessory-genome
# of NC genome species
log₂ size ratio (HRGM2/UHGG)
k
HRGM2
UHGG
Percentage of classified reads
Domain
Phylum
Class
Order
Family
Genus
Species
Taxonomic rank
l
Genus level
HRGM2
UHGG
Africa
Asia1
Asia2
Asia3
Asia4
Asia5
Asia6
Asia7
Europe1
Europe2
Europe3
Europe4
North America1
North America2
North America3
Oceania1
Oceania2
Dataset

**Fig. 1 | Overview of HRGM2. a**, Maximum-likelihood phylogenetic tree representing 4,798 bacterial species in HRGM2. Color strips from inner to outer circles represent phylum, number of non-redundant genomes in the species cluster, availability of isolated genomes, and whether the species cluster consists solely of genomes newly added by HRGM2. **b-c**, Comparison between UHGG and HRGM2 for (b) the number of WMS samples from which MAGs originate and (c) the number of countries providing WMS samples, **d**, The proportion of continental origin of non-redundant genomes within each phylum. The plot shows only the phyla with at least 50 non-redundant genomes with known continental origin. **e**, The number and proportion of genomes originating from Africa and non-Africa according to the inclusion of added genomes in HRGM2 from the phyla Elusimicrobiota (top) and Spirochaetota (bottom). **f-g,** Comparison between UHGG and HRGM2 for (f) the number of non-redundant near-complete (NC) genomes, and (g) the number of NC genome species. **h**, The number and percentage of NC genome species shared and not shared between HRGM2 and other human gut microbiome catalogs. **i-j**, Comparison between HRGM2 and UHGG for (i) the number of NC member genomes and (j) the pan-genome, core-genome, and accessory-genome sizes for conspecific pairs. **k-l**, Comparison of the percentage of classified reads between HRGM2 and UHGG at (k) each taxonomic rank and (l) the genus rank, with datasets stratified by continent. Box lengths represent the interquartile range of the data, and whiskers extend to the lowest and highest values within 1.5 times the interquartile range from the first and third quartiles, respectively. The center bar represents the median. All the outliers are shown in the plots.

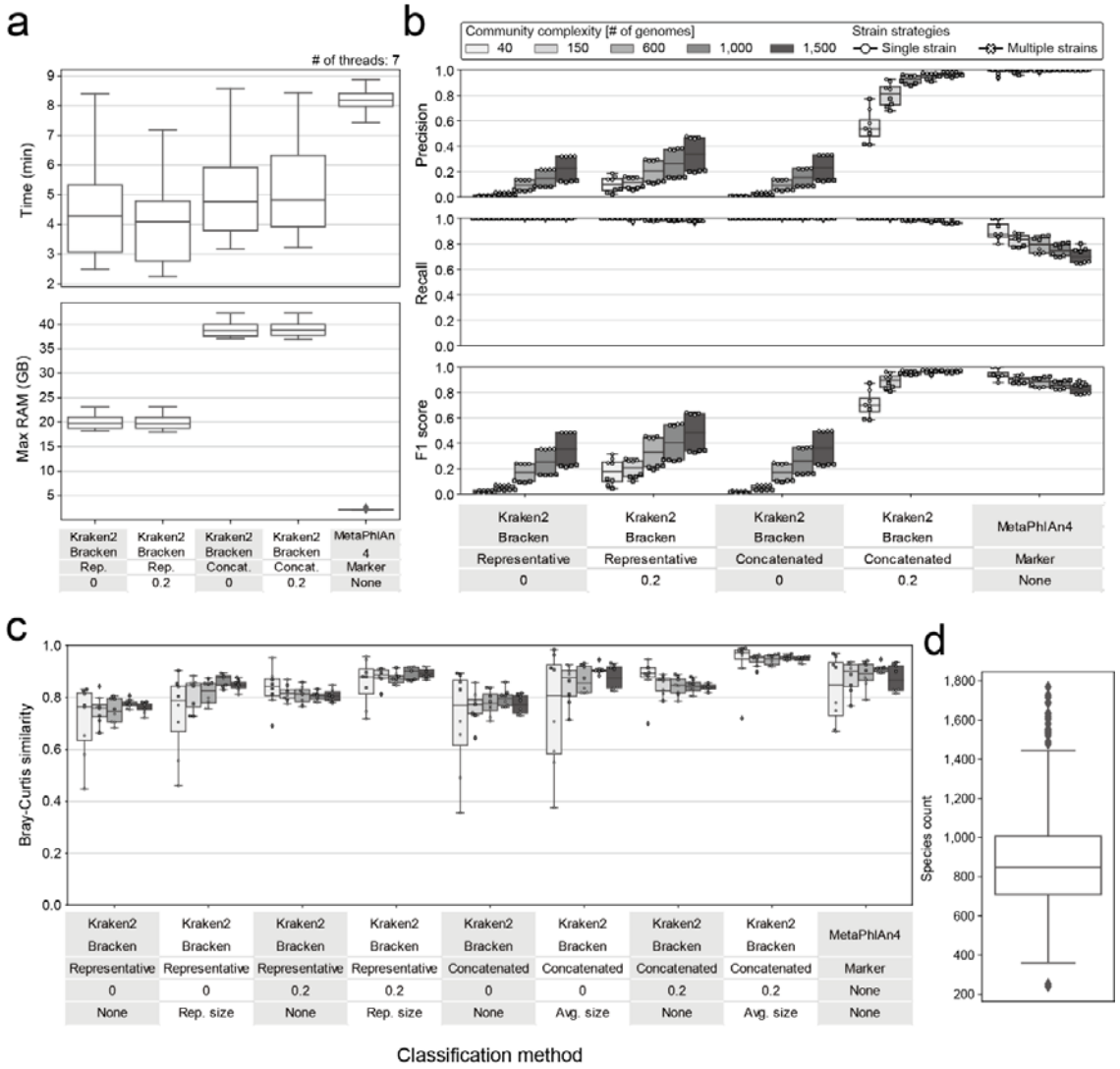
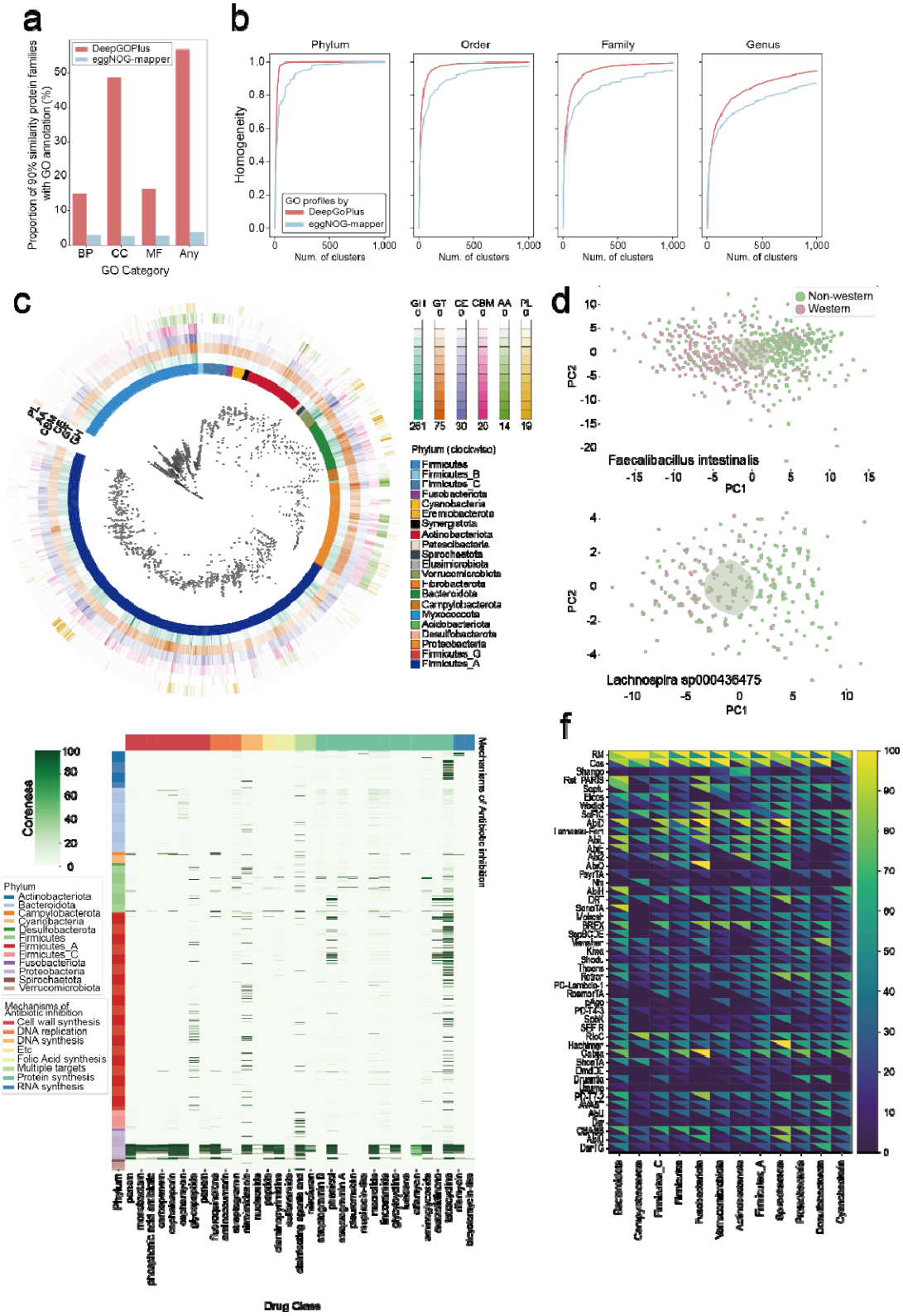



**Fig. 2 | Benchmarking of taxonomic classification methods based on HRGM2 for human gut microbiota. a-c**, Comparative analysis of taxonomic classification methods from three perspectives: (a) computational resources, (b) species detection, and (c) species abundance. Classification methods are represented by three layers in (a) and (b), and by four layers in (c). From the topmost layer, each layer represents the tool used for taxonomic classification, the database version, the confidence score, and the method of species size normalization, respectively. In (b) and (c), the color of the box indicates the number of genomes used to create the simulation data (community complexity), and the shape of the dot represents how many genomes a species is represented by (strain strategies). **d**, Number of HRGM2 species in 1,160 WMS samples from healthy human guts. The best pipeline according to our analysis was used, and to consider false positives, we assumed that a species was present in the sample if its normalized relative abundance was greater than 1e-06. In Fig. 2, box lengths represent the interquartile range of the data, and whiskers extend to the lowest and highest values within 1.5 times the interquartile range from the first and third quartiles, respectively. The center bar represents the median. All the outliers are shown in the plots.


a
DeepGOPlus
eggNOG-mapper
Proportion of 90% similarity protein families with GO annotation (%)
BP
CC
MF
Any
GO Category
b
Phylum
Order
Family
Genus
Homogeneity
GO profiles by
DeepGoPlus
eggNOG-mapper
Num. of clusters
c
GH
GT
CE
CBM
AA
PL
Phylum (clockwise)
Firmicutes
Firmicutes_B
Firmicutes_C
Fusobacteriota
Cyanobacteria
Desulfobacterota
Synergistota
Actinobacteriota
Patescibacteria
Spirochaetota
Elusimicrobiota
Verrucomicrobiota
Fibrobacterota
Bacteroidota
Campylobacterota
Myxococcota
Acidobacteriota
Desulfobacterota
Proteobacteria
Firmicutes_G
Firmicutes_A
d
Non-western
Western
PC1
PC2
Faecalibacillus intestinalis
Lachnospira sp000436475
e
Coreness
Phylum
Actinobacteriota
Bacteroidota
Campylobacterota
Cyanobacteria
Desulfobacterota
Firmicutes
Firmicutes_A
Firmicutes_C
Fusobacteriota
Proteobacteria
Spirochaetota
Verrucomicrobiota
Mechanisms of Antibiotic inhibition
Cell wall synthesis
DNA replication
DNA synthesis
Etc
Folic Acid synthesis
Multiple targets
Protein synthesis
RNA synthesis
Drug Class
f

1199

**Fig. 3 | Overview of the functional landscape of the human gut microbiome. a-b**, Comparison of Gene Ontology (GO) annotations by eggNOG-mapper and DeepGOPlus. (a) Annotation rates for each GO category (BP: Biological Process, CC: Cellular Component, MF: Molecular Function, Any: Any GO terms). (b) Taxonomic homogeneity of clusters resulting from agglomerative clustering based on GO annotations from either DeepGOPlus or eggNOG-mapper across different taxonomic levels. **c**, Phylogenetic tree of genera in HRGM2, indicating the average number of CAZyme families in genomes belonging to each genus. CAZyme classes; GH: Glycoside Hydrolase, GT: Glycosyl Transferase, CE: Carbohydrate Esterase, CBM: Carbohydrate Binding Module, AA: Auxiliary Activity, PL: Polysaccharide Lyase. **d**, PCoA of two species with the most distinct CAZyme profiles between Western and non-Western countries. **e**, Overview of antibiotic resistance gene in human gut microbiome. Coreness, which refers to the percentage of conspecific genomes within a species that have a resistance to each antibiotics. **f**, Overview of a defense system for viruses. Percentage of presence of each defense system as core or accessory genomes per phylum. At the bottom of each cell in the heatmap is the percentage of species where the defense system is present in their core genome, while at the top is the percentage of species where the defense system is present in their accessory genome.

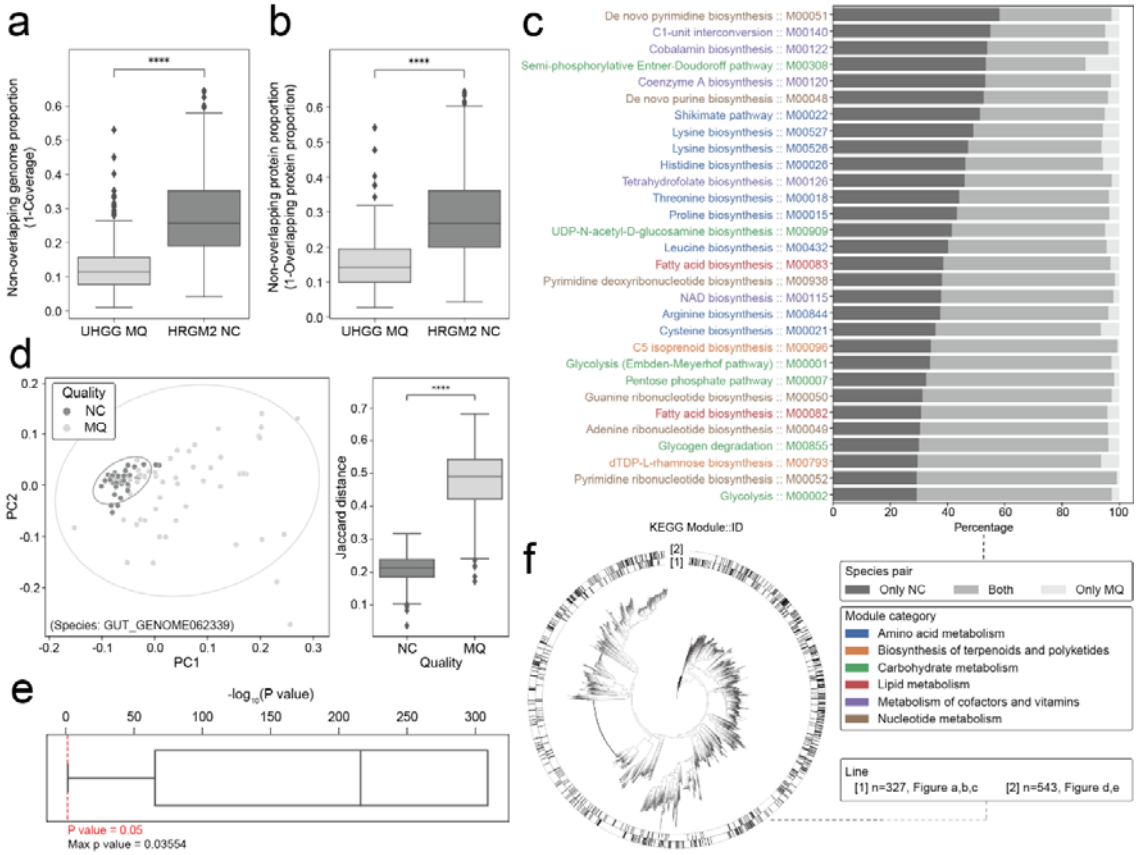



**Fig. 4 | The limitations of MQ genomes in functional profiling. a-b**, Comparison of non-overlapping (a) genome and (b) protein proportion for conspecific MQ and NC genomes. Differences were evaluated with two-sided Wilcoxon signed-rank test (****, $P$ < 1e-04). **c**, Top 30 KEGG modules with a high percentage detected only in NC genomes. **d**, PCoA plot of the functional composition profiles of MQ and NC genomes (left) and the functional distance between genomes within the same quality group (right) for the species 'GUT_GENOME062339' in UHGG. Ellipses in PCoA represent a 95% confidence interval. A one-tailed Mann-Whitney U test was performed to confirm if the functional distances between NC genomes were significantly smaller than between MQ genomes (****, $P$ < 1e-04). **e**, $P$-values of the 543 species evaluating the functional distance difference as in (**d**, right). To take the logarithm of $P$-values, $P$-values of 0 were replaced with the smallest non-zero $P$-value (9.11e-310). The red dotted line represents a $P$-value of 0.05. **f**, Position of the UHGG species used in the analyses on the UHGG phylogenetic tree. The inner and the outer black color strips represent the positions of the 327 species used in (**a-c**) and the 543 species used in (**d**, **e**), respectively. For boxplots in Fig. 4, box lengths represent the interquartile range of the data, and whiskers extend to the lowest and highest values within 1.5 times the interquartile range from the first and third quartiles, respectively. The center bar represents the median. All the outliers are shown in the plots.

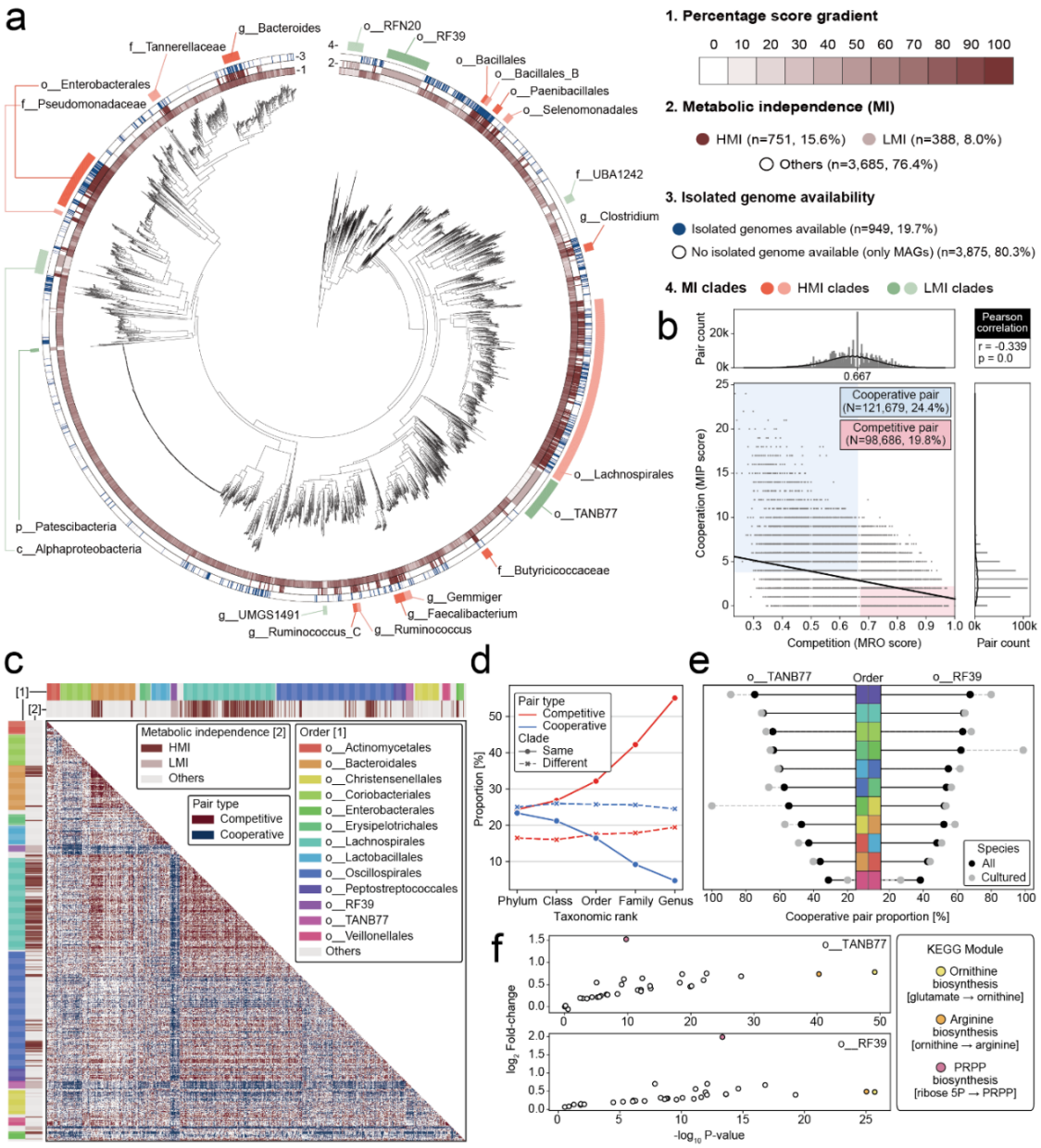
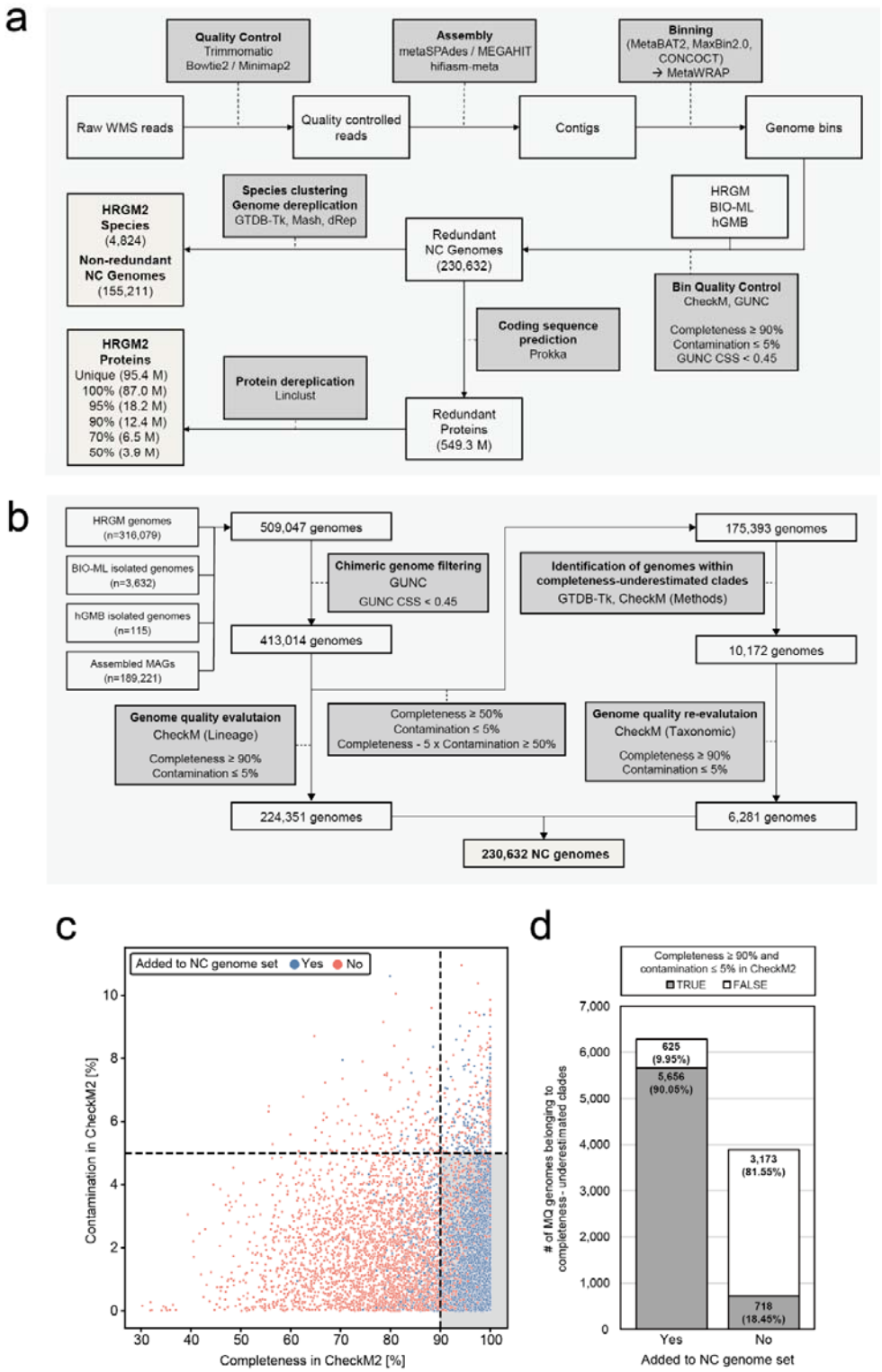


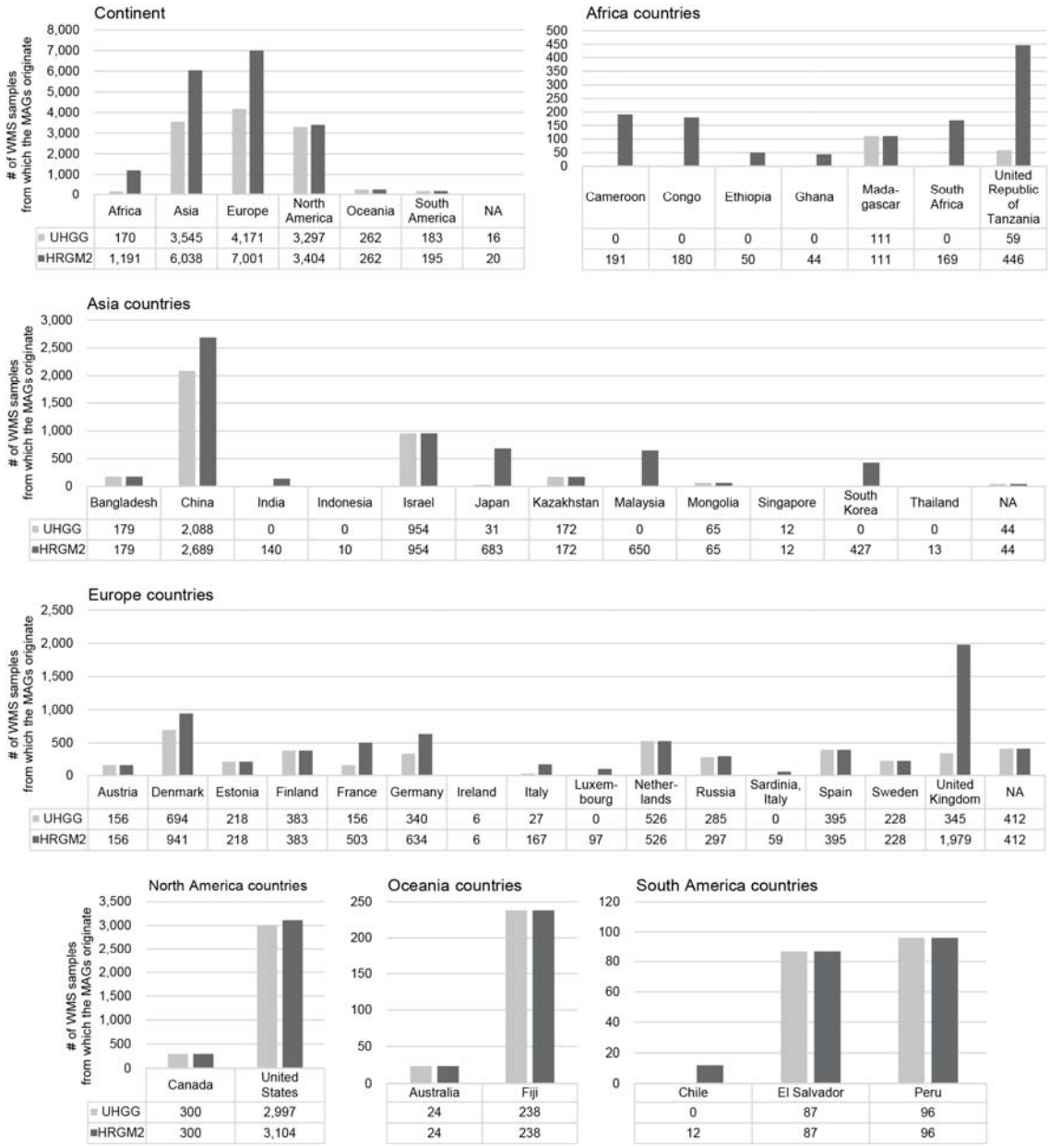

**Fig. 5 | Analysis of metabolic independence and interactions in the human gut microbiota. a**, A maximum-likelihood phylogenetic tree annotated with metabolic independence (MI) status. The inner to outer circles represent possession percentage of 33 KEGG modules used to assess metabolic independence, MI status, isolated genome availability, and designation of MI clades. **b**, A scatter plot showing distribution of metabolic resource overlap (MRO) and metabolic interaction potential (MIP) scores across all species pairs (n=499,500) among the top 1,000 species prevalent in healthy human guts globally. The blue and pink regions indicate cooperative (n=121,679, 24.4%) and competitive (n=98,686, 19.8%) pairs, respectively. A linear regression line highlights the relationship between MRO and MIP scores, with correlation coefficient and *P*-value determined by a Pearson correlation test in SciPy. **c**, A chart showing metabolic interactions, where outer color strips represent taxonomic order, and inner strips denote MI. Species are ordered by lineage, with orders containing fewer than 10 species grouped as "Others". **d**, Proportion of cooperative and competitive pairs within the same and across different clades, with "Clade" referring to ranks from phylum to genus. **e**, Cooperative pair

proportion between order TANB77 (left) / RF39 (right) and other orders is displayed. Each color represents an order, consistent with panel (**c**). "All" refers to the cooperative pair proportion across all species in an order, while "Cultured" is specific to cultured species with isolated genomes. **f**, Association between the presence of specific KEGG module and cooperative pair proportion within TANB77 (top) and RF39 (bottom) orders is shown. Fold-change is calculated by dividing the average cooperative pair proportion of species with a given KEGG module by those without it. Differences in cooperative pair distributions for species with and without KEGG modules were evaluated with a two-sided Mann-Whitney U test.

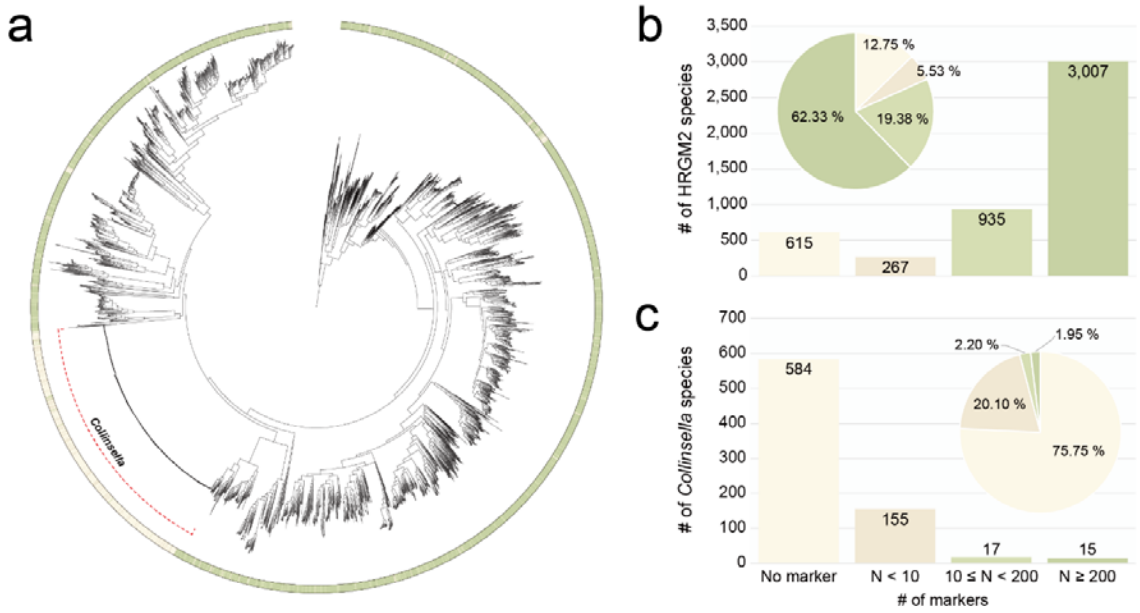
## Extended Data Figures



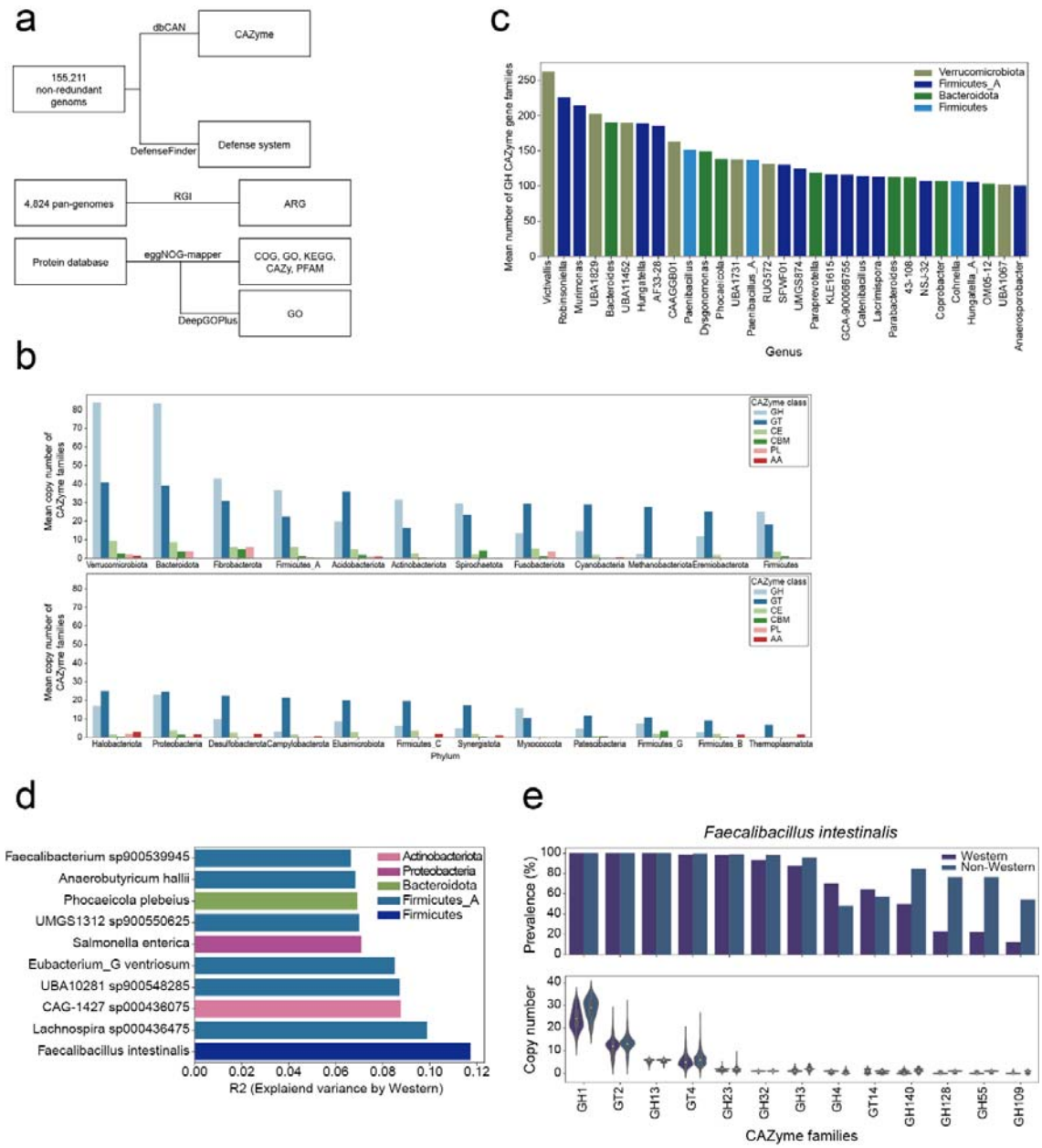
**Extended Data Fig. 1 | Bioinformatics pipelines for HRGM2 construction. a**, The overall pipeline for HRGM2 construction. **b**, Pipeline to control genome quality. **c**, CheckM2 assessment of 10,172 MQ genomes previously assessed by CheckM (completeness-underestimated). The 6,281 genomes that were added to the near-complete (NC) genome set based on assessment using universal bacterial markers and CPR markers are marked with blue dots. The vertical and horizontal dashed lines indicate the 90% completeness and 5% contamination threshold, respectively. The gray area includes genomes that meet the NC genome criteria according to CheckM2. **d**, Count and proportion of filtered in and out by universal bacterial markers and CPR markers that meet the NC criteria under CheckM2.



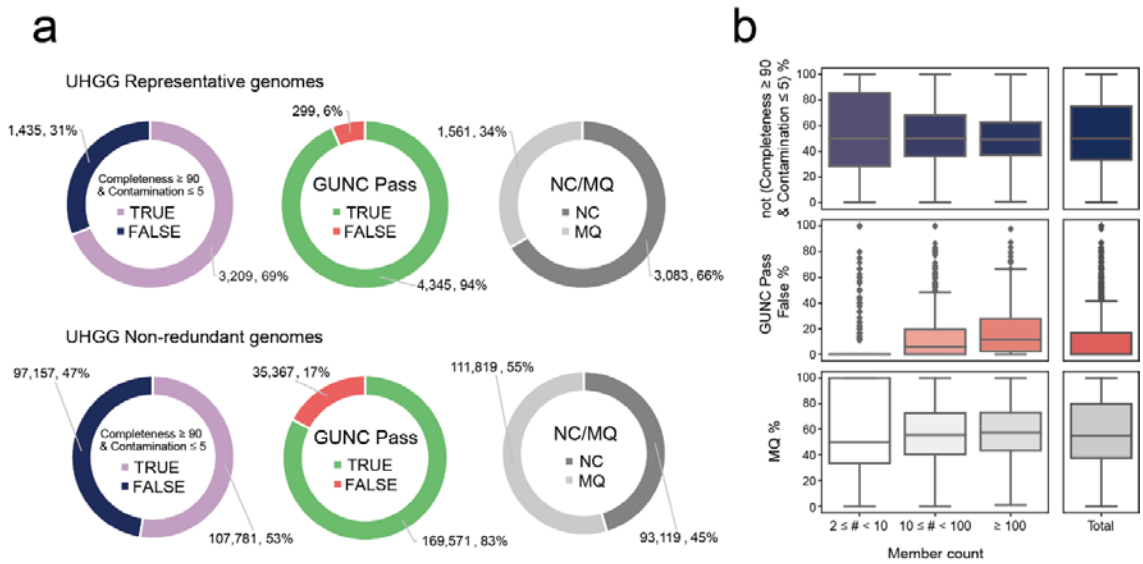

**Extended Data Fig. 2 | The number of WMS samples from which the MAGs originate in UHGG and HRGM2 by continent and country.**



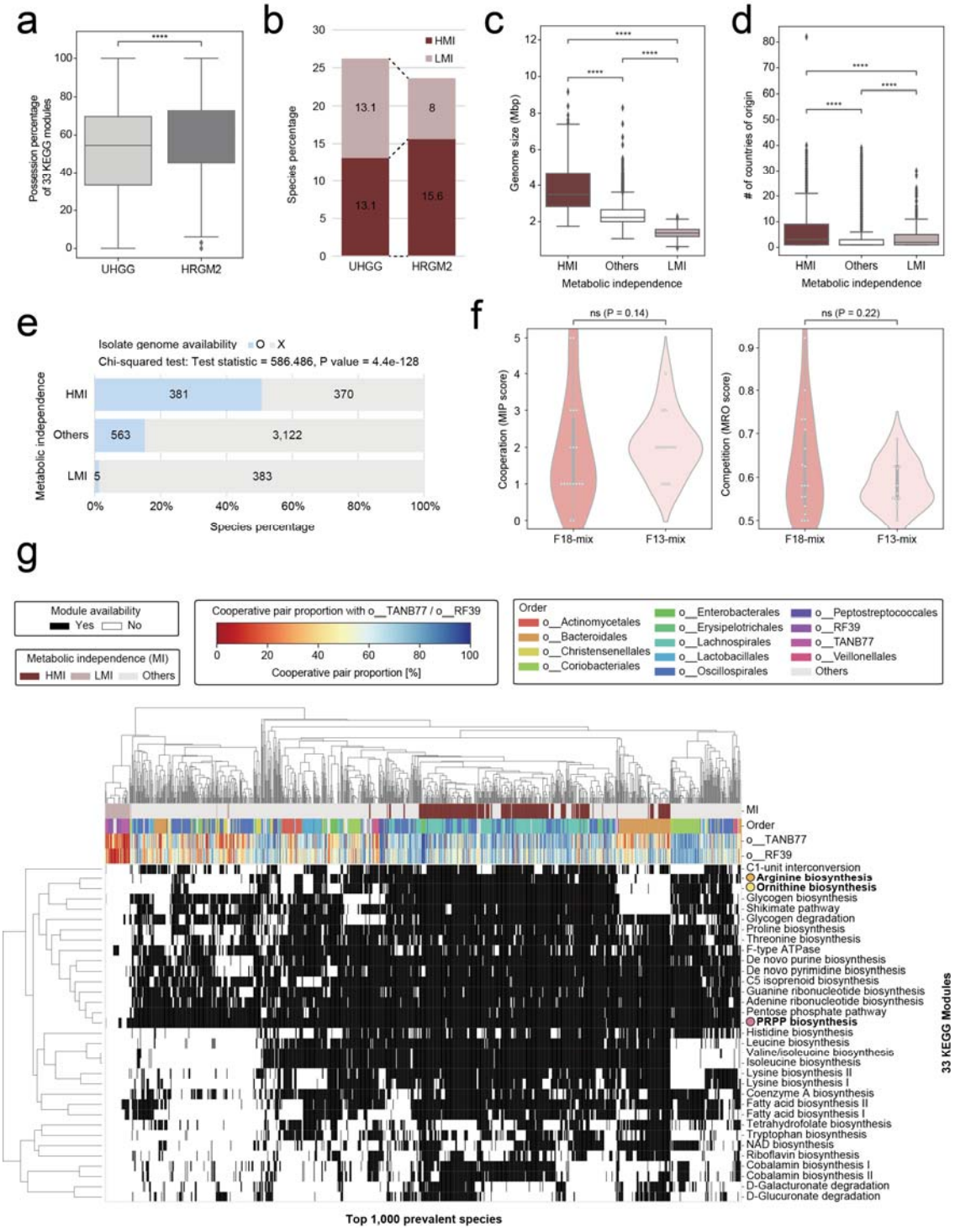

**Extended Data Fig. 3 | Summary of HRGM2 species-specific marker gene database. a**, Maximum-likelihood phylogenetic tree with annotations of the number of species-specific marker genes. The color of the strip according to the number of species-specific marker genes is the same as in (b, c). **b-c**, The number and proportion of species in (b) HRGM2 and (c) *Collinsella* genus according to the number of species-specific marker genes.




**Extended Data Fig. 4 | Functional landscape of human gut microbiome. a**, Summary of functional prediction pipeline in HRGM2. **b**, Average copy number of CAZyme families per phylum. GH: Glycoside Hydrolase, GT: Glycosyl Transferase, CE: Carbohydrate Esterase, CBM: Carbohydrate Binding Module, PL: Polysaccharide Lyase, AA: Auxiliary Activity. **c**, Average copy number of GH CAZyme families for genera with more than 100 GH CAZyme families. **d**, Explained variance by Western/non-Western categorization for the 10 species with the most distinct CAZyme profiles between Western and non-Western continents. In (**c**) and (**d**), the color of the bars represents phylum. **e**, Comparison of the prevalence and the copy number of each CAZyme family between *Faecalibacillus intestinalis* genomes from Western and non-Western countries.




**Extended Data Fig. 5 | Summary of genome qualities for UHGG. a**, Number and proportion of genomes with completeness ≥ 90% and contamination ≤ 5% (left pie chart), that passed GUNC (center pie chart), and that met the NC criteria (right pie chart), in UHGG representative genomes (up) and non-redundant genomes (down). **b**, Distribution of the percentage of genomes that are not completeness ≥ 90% and contamination ≤ 5% (top), that did not pass GUNC (middle), and that did not meet the NC criteria (bottom) for each UHGG species with at least two non-redundant genomes. The distributions are either categorized by the number of non-redundant genomes included in each species (left) or not (right). Box lengths represent the interquartile range of the data, and whiskers extend to the lowest and highest values within 1.5 times the interquartile range from the first and third quartiles, respectively. The center bar represents the median. All the outliers are shown in the plots.




**Extended Data Fig. 6 | Additional analyses for metabolic independence and interaction. a**, Possession percentage of 33 KEGG modules for UHGG and HRGM2 species. Statistical significance was assessed using a two-sided Mann-Whitney U test (****, $P$ < 1e-04), indicating notable differences in module possession between the two catalogs. **b**, HMI and LMI species percentage in UHGG and HRGM2. **c**, **d**, Representative genome size (**c**) and the number of countries where each species originated (**d**) of HMI, other, LMI species. $P$-value was

calculated using a two-sided Mann-Whitney U test (****, $P < 1e-04$). **e**, Number and percentage of species with available isolate genomes, categorized by metabolic independence. The association between metabolic independence and the availability of isolate genomes was assessed by a Chi-squared test. **f**, Violin plot showing distribution of MIP and MRO scores for F18-mix and F13-mix strains of Kp-2H7. *P*-value was calculated using a one-sided Mann-Whitney U test (ns: not significant, $P > 0.05$). **g**, Availability for 33 KEGG modules of the top 1,000 prevalent species. Species were annotated with four layers: the first and second layers (from the top) represent metabolic independence and taxonomic order (as in Fig. 5c); the third and fourth layers indicate the cooperative pair proportion for orders TANB77 and RF39. The three KEGG modules of primary interest in Fig. 5f are highlighted in bold. Clustering of species and modules was performed using Hamming distance, allowing visual comparison of module possession patterns across species. For boxplots in Extended Data Fig. 6, box lengths represent the interquartile range of the data, and whiskers extend to the lowest and highest values within 1.5 times the interquartile range from the first and third quartiles, respectively. The center bar represents the median. All the outliers are shown in the plots.