Genomic diversity and underlying metabolic variations of the human gut microbiome between populations at single nucleotide resolution

Quanbin Dong[1,2†], Beining Ma[1,3†], Xiaofeng Zhou[4,5,6†], Pan Huang[7,8†], Mengke Gao[1], Shuai Yang[1], Yuwen Jiao[2], Yan Zhou[2], Zhun Shi[9], Qiufeng Deng[1], Dongxu Hua[1], Xiuchao Wang[1], Lu Liu[1], Chengcheng Zhang[7,8], Chuan Zhang[7,8], Mengmeng Kong[4,5,6], Chang He[4,5,6], Tingting Wu[1], Huayiyang Zou[1], Jing Shi[1], Yanhui Sheng[3], Yifeng Wang[1], GMR Consortium, Liming Tang[2], Shixian Hu[10], Huanzi Zhong[9], Wei Sun[1], Wei Chen[7,8], Qixiao Zhai[7,8]*, Xiangqing Kong[1,3]*, Yan Zheng[4,5,6]* & Lianmin Chen[1,2]*

[1] Department of Cardiology, The First Affiliated Hospital of Nanjing Medical University, Nanjing Medical University, Nanjing, China

[2] Department of Gastroenterology, The Third Affiliated Hospital of Nanjing Medical University, Nanjing Medical University, Changzhou, China

[3] Cardiovascular Research Center, The Affiliated Suzhou Hospital of Nanjing Medical University, Nanjing Medical University, Suzhou, China

[4] State Key Laboratory of Genetic Engineering, School of Life Sciences, Human Phenome Institute, Fudan University, Shanghai, China

[5] Department of Cardiology, Zhongshan Hospital, Fudan University, Shanghai Institute of Cardiovascular Disease, Shanghai, China

[6] Ministry of Education Key Laboratory of Public Health Safety, School of Public Health, Institute of Nutrition, Fudan University, Shanghai, China

[7] State Key Laboratory of Food Science and Resources, Jiangnan University, Wuxi, China

[8] School of Food Science and Technology, Jiangnan University, Wuxi, China

[9] BGI Genomics, Shenzhen, China

[10] Institute of Precision Medicine, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

†These authors contributed equally and share the first author

*Correspondence should be addressed to: zhaiqixiao@jiangnan.edu.cn, kongxq@njmu.edu.cn, yan_zheng@fudan.edu.cn & lianminchen@njmu.edu.cn

Abstract

Understanding microbial genomic diversity across populations is crucial for advancing personalized medicine and addressing regional health disparities. However, with 70% of current gut microbial genome data from Europeans and the Americans, regions like East Asia, including China, remain underrepresented, limiting our understanding of global microbial genomic diversity and its health impacts.

Here, we present the Global gut Microbiome Reference (GMR), a comprehensive dataset of 478,588 high-quality microbial genomes from Chinese (247,134) and non-Chinese (231,454) populations, covering 6,664 species, 26.4% of which are potential newly classified. Analysis of single nucleotide variations (SNVs) across 304 globally prevalent species revealed that 24.7%, especially from the *Lachnospiraceae* and *Ruminococcaceae* families, showed significant genetic differences between Chinese and non-Chinese populations, primarily linked to antibiotic resistance, virulence, and other health-related functions. By integrating these new species into the MetaPhlAn4 database, we enhanced species profiling by up to 22.8%, and established replicable abundance associations to host phenotypic traits between cohorts. Additionally, we found that microbial strain diversity at the SNV level, independent of abundance, was associated with host phenotypic traits, such as cardiometabolic-related risk factors like obesity. In vitro experiments with *Bacteroides fragilis_A* further highlighted the regulatory role of microbial SNVs in metabolites related to cardiometabolic health.

These findings emphasize the need to consider gut microbial genetic diversity in population studies for better insights into regional microbiome differences and their health implications.

Introduction

The human gut colonized by trillions of microbes with diverse metabolic activities that are crucial to human health[1]. Beyond mere diversity and abundance, microbial genomic variations, which contribute to metabolic capacities, provide a finer-resolution view and mechanistic understanding of host-microbiome interactions in the context of human health[2-5]. For instance, within-species microbial single nucleotide variations (SNVs) have been linked to obesity[6], and strain-resolved microbial abundances have significantly improved prediction models of immune checkpoint blockade response across cancer types[7].

Importantly, prominent microbial species, although globally distributed, exhibit significant codiversification, developing population-specific characteristics related to host dependence[8]. Therefore, understanding microbial genomic diversity and metabolic variations across major world populations can enhance personalized approaches in microbial discovery and treatment, particularly in addressing regional health disparities.

Nevertheless, several critical issues remain unaddressed, impeding a comprehensive understanding of global gut microbial genomic diversity and the underlying functional variations. First, currently available gut microbial genomes are predominantly derived from western populations. According to the Unified Human Gastrointestinal Genome (UHGG)[9] version 2 collection with 289,232 worldwide gut microbial genomes released in late 2023, only about 30% of them come from non-western populations. As a major representative of eastern populations, significant efforts have been devoted to generating comprehensive gut microbial genomes for Chinese[9-11]. However, the sample size remains limited compared to the large population. Second, genomic characterization of well-known microbes such as *Prevotella copri*[12] and *Akkermansia muciniphila*[13] has revealed significant global phylogenetic and functional diversity. However, a systematic comparison of within-species genomic diversity across prominent microbes in major populations, and the potential physiological relationships between host and regional-specific microbes are still lacking. Third,

while within-species genetic variations, such as nonsynonymous mutations, have been associated with differences in their encoded metabolic capacities[3], these associations lack experimental confirmation to determine whether specific genomic variations in the gut microbiome are driving forces behind host health-relevant metabolites.

In this study, we present a comprehensive Global gut Microbial Reference (GMR) collection, comprising nearly half a million high-quality genomes. This collection integrates existing datasets[9-11, 14, 15] and includes an additional 6,657 Chinese metagenomic samples to enhance the representation of gut microbial genomic diversity across major world populations. We systematically characterized the within-species phylogenetics of prominent microbes between Chinese and non-Chinese populations, identifying region-specific genomic components encoding antibiotic resistance and other human health relevant functionalities. For potential novel microbial species, we incorporated their genomes into the MetaPhlAn4[16] taxonomic classification database (released in March 2024) for population-level profiling, establishing their compositional and genomic relationships with diverse host physiological traits. Finally, we characterized within-species microbial strain diversity associations to host physiological traits at the SNVs level, and further assessed the metabolic regulatory roles of SNVs in *Bacteroides fragilis_A* strains relevant to cardiometabolic health, using *in vitro* experimental settings.

Results

Extend eastern gut microbial genomes with a quarter million from Chinese populations

Nearly 300,000 human gut microbial genomes were included in the late 2023 release of the Unified Human Gastrointestinal Genome (UHGG)[9] version 2, but only 30% originated from non-western populations. To enhance the representation of eastern populations and our understanding of global gut microbial genomics diversity and its health impacts, we assembled extra 6,657 gut metagenomic samples from the Chinese population as of February 2024[10, 17, 18, 19], which have not yet been included in the latest UHGG collection. Using a single-sample metagenomic assembly strategy (**Fig.**

1a), a total of 180,719 metagenome-assembled genomes (MAGs) were obtained. We then incorporated existing microbial genomes collections, including UHGG[9], the Human Gut Microbial Biobank (hGMB)[14], the Culturable Genome Reference (CGR2)[15] and IMGG[11], with these newly assembled genomes (**Fig. 1a**).

To standardize genome quality across all MAGs, we applied UHGG criteria[9, 20] (completeness > 50%, contamination < 5%, and quality score > 50), resulting in a total of 478,588 medium-to-high quality microbial genomes, with 247,134 (51.6%) originated from Chinese populations (**Fig. 1b; Supplementary Table 1**). It is worth noting that the assembled genomes were of high quality, with the median completeness and contamination rates of genomes of Chinese origin are 90.27% (interquartile range, IQR: 77.01–95.97%) and 0.88% (IQR: 0.08–1.90%), respectively. For non-Chinese genomes, these values are 90.60% (IQR: 77.47–96.51%) and 0.85% (IQR: 0.11–1.79%), respectively (**Fig. 1c; Supplementary Fig 1a&b; Supplementary Table 1**). Additionally, both median genome size (2.27Mb vs. 2.21Mb, $P_{Wilcoxon}=3.10\times10^{-35}$) and N50 length (30Kb vs. 27Kb, $P_{Wilcoxon}=1.19\times10^{-109}$) of Chinese-origin genomes were slightly higher than non-Chinese genomes (**Supplementary Fig. 1c&d; Supplementary Table 1**). We refer to this largest integrated collection as the Global gut Microbial Reference (GMR) set. The GMR set will facilitate the decoding of global gut microbial phylogenetic and functional diversity and its implications for regional health disparities related to the gut microbiome.

Taxonomic classification of GMR revealed 6,664 species clusters with 1,762 being identified as potential new species

We then assessed the taxonomic distribution of microbial genomes in the GMR using a two-step approach. In brief, we first annotated all 478,588 genomes at the species level using the Genome Taxonomy Database (GTDB, R214)[21, 22]. Among these, 475,110 genomes were successfully assigned to 4,902 species. For the remaining 3,478 genomes that could not be annotated at the species level, we clustered them using dRep[23] with an ANI threshold of > 95%, resulting in 1,762 clusters (**Fig. 2a**).

Notably, among the 1,762 clusters that could not be assigned to known species, 101 clusters (1.52%) and 4 clusters (0.06%) could not even be classified at the genus or family levels, respectively (**Fig. 2b**). In total, we obtained 6,664 species-level clusters, comprising 6,624 bacterial and 40 archaeal species (**Supplementary Table 2**). The most prevalent bacterial phyla were *Bacillota_A* (2,585 species), *Actinomycetota* (1,103 species), and *Bacteroidota* (929 species) (**Fig. 2c**). Among them, *Actinomycetota* has the highest proportion (41.32%) of unknown species discovered (**Fig. 2c**). This highlights the vast, unexplored microbial diversity within the human gut.

To explore the functional capacities encoded by microbial genomes in our GMR set, we predicted protein-coding sequences (CDS) for all 478,588 genomes, resulting in a dataset of 1,028,234,783 genes, with an average of 2,148 CDS per genome (**Fig. 2d; Supplementary Table 3**). Gene function annotation, performed using Prokka[24], which integrates databases including UniProt, RefSeq, Pfam and TIGRFAM, revealed that the median proportion of genes with unknown function was 47.3% (IQR: 43.3–51.5%) for Chinese genomes and 48.2% (IQR: 43.7–52.9%) for non-Chinese genomes (**Fig. 2e; Supplementary Table 3**). This indicates that nearly half of the microbial genes in the human gut remain unexplored (**Fig. 2f)**, consistent with previously reported[25]. Notably, we found that, on average, species in the *Thermoplasmatota* phylum had the highest percentage of unknown functional genes (**Supplementary Fig. 2)**, suggesting that species belonging to this clade have great potential for functional exploitation. Consequently, the GMR provides a comprehensive catalog of gut microbiome proteins, offering a valuable resource for future functional metagenomic analyses.

Taxonomic markers from new species boost populational profiling by up to 23%

For the 1,762 newly assembled species, to enhance the detection of these microbes in population studies and provide a more comprehensive gut microbial composition profile, we integrated MAGs from our GMR set into the MetaPhlAn4 taxonomic classification database[16], released in March 2024. Among the 6,664 species-level

clusters, 2,465 are represented by a single genome, and 929 by two genomes (**Fig. 3a**; **Supplementary Fig. 3; Supplementary Table 4**). For the remaining 3,270 species clusters with at least three genomes per cluster, we classified their species-level specific markers.

Following the MetaPhlAn4 taxonomic database construction pipeline, we first identified core coding sequences (CDS) for each species, averaging 1,626 core genes per species (**Fig. 3b; Supplementary Table 5**). We then screened these core genes to identify species-specific marker genes, finding an average of 676 per species (**Fig. 3b; Supplementary Table 5**). Based on the MetaPhlAn4 guidelines, we retained a maximum of 200 marker genes for each species. For 493 species with fewer than 10 species-specific marker genes and 2,198 species overlapping with the existing MetaPhlAn4 taxonomic database, we excluded them from the database construction (**Fig. 3c)**. Ultimately, 96,156 marker genes from 579 non-overlapped species were integrated into the MetaPhlAn4 database to build our customized database (**Supplementary Table 6**).

To evaluate the performance of our customized database in profiling gut microbial composition, we analyzed an independent cohort of 1,076 publicly available fecal metagenomic samples from a Chinese cohort[26]. Using MetaPhlAn4 with default parameters and our customized taxonomic classification database, we performed microbial abundance profiling. On average, integrating the GMR database increased the reads mapping rate by 1.5% per sample (**Supplementary Fig. 4**). In total, we detected 2,669 species across all samples, including 326 newly added species. Filtering for species present in at least 1% of samples, 150 out of 1,325 species were from our new additions, indicating a 22.8% increase compared to the existing MetaPhlAn4 database (**Fig. 3d**). With a filtering threshold set from 1% to 10%, our customized database consistently showed an average 9.5% improvement in the total number of species detected. The species that showed an increased detection rate at the family level primarily belong to *Acutalibacteraceae*, *CAG-508* and *Ruminococcaceae*, as well as 13 other families in total (Hypergeometric test, FDR < 0.05)

(**Supplementary Fig. 5; Supplementary Table 7**). Notably, 97.58% of individuals had at least one newly added species detected, indicating that these species are widely present in Chinese populations (**Supplementary Fig. 6**). These results highlight the enhanced capability of our customized taxonomic classification database for species abundance identification in Chinese populations.

Abundance of newly profiled species widely associated with host phenotypic traits

Next, we assessed the potential biological significance of these newly profiled species in the CGMR cohort[10]. Using our customized database, we detected 5,310 species in 3,234 participants, including 465 newly added species (**Supplementary Fig. 7; Supplementary Table 8**). On average, 201 species and 11 new species were detected per sample (**Fig. 4a**). Unsurprisingly, 97.8% of individuals had at least one newly added species detected (**Fig. 4a; Supplementary Table 9**).

By using spearman correlation to associate the abundance of 54 out of the 465 newly added species with 22 phenotypic traits (each with at least 5% prevalence), we identified 325 significant associations (FDR < 0.05, **Fig. 4b; Supplementary Table 10**), ranging from intrinsic factors to diverse dietary habits. Among them, age showed the most associations with the abundance of these novel species (**Supplementary Fig. 8**). To assess the robustness of these associations, we used an independent cohort consisting of 551 participants[19]. We identified 318 associations for 20 overlapping phenotypes (**Supplementary Table 11**). Heterogeneity analysis further revealed that 295 out of 318 (92.8%) associations were consistent between the two cohorts ($P_{Cochran\text{-}Q} > 0.05$) (**Fig. 4c; Supplementary Table 11**).

We then assessed the potential confounding effects of the identified associations, noting that multiple phenotypic traits were associated with the same species (**Fig. 4b**). Using MaAsLin2[27], a tool for addressing multivariable associations between host phenotypes and gut microbes, we identified 64 robust associations between 39 species and 16 phenotypes among those 295 associations (**Supplementary Table 10**). These associations are primarily related to urban or rural residency, defecation frequency,

age, lifestyle, and dietary habits. For example, the abundance of *Pseudoflavonifractor_A sp018374635* was negatively associated with age ($P = 9.13\times10^{-4}$, **Fig. 4d**), while there was a positive association between defecation frequency and the abundance of *Streptococcus infantis_I* ($P = 1.44\times10^{-3}$, **Fig. 4e**). Additionally, we observed a significantly higher abundance of a *Streptococcus* species *GT_GMRSGB00117* in urban areas compared to rural areas ($P = 1.44\times10^{-2}$, **Supplementary Fig. 9a; Supplementary Table 10**). Regarding lifestyle, the frequency of drinking was positively associated with the abundance of *Nanosynbacter sp900556355* ($P = 2.47\times10^{-2}$) (**Supplementary Fig. 9b; Supplementary Table 10**). These results demonstrate that our customized database can uncover previously undetected relationships between microbes and host phenotypic traits, laying the foundation for further exploration of the relationship between microbial abundance and host disease.

Genetic stratification of predominant species between populations is related to antibiotic resistance and virulence

For 2,145 out of 6,664 species that present in both Chinese and non-Chinese populations (**Fig. 5a; Supplementary Table 12**), we questioned their potential genomic differences and underlying functionalities. For this, we focused on 304 species with at least 100 MAGs and a proportion of at least 30% in either Chinese and non-Chinese populations (**Fig. 5b**). On average, the pangenome size of these species was 25.96M bp, with an average of 399,764 SNVs detected (**Supplementary Fig. 10**). Next, we compared the within- and between-population genetic dissimilarity per species based on their SNV sites using Kimura 2-parameter distance matrix. Given its importance in previous studies, *Prevotella copri* serves as a reference for understanding the population structuring of gut species. Previous studies have confirmed that *P. copri* exhibits a genetic structure linked to geographic origin and lifestyle[12], with Chinese-origin genomes displaying distinct lineages[28]. Consistent with this, we observed a clear stratification of genomes originating from Chinese individuals (**Supplementary Fig. 11**). Next, we systematically evaluated the genetic

structure of all these 304 predominant gut species between Chinese and non-Chinese populations. Overall, 75 out of the 304 species (24.7%) exhibited significant population stratification (Kruskal-Wallis test, FDR < 0.05, **Fig. 5c&d; Supplementary Figure 12**). Notably, 33 of these species belong to the *Lachnospiraceae* family, and 9 species belong to the *Ruminococcaceae* family, indicating that species from these families showed populational specificity (Hypergeometric test, $P = 1.15 \times 10^{-15}$ and $P = 1.06 \times 10^{-3}$, respectively) (**Fig. 5c**).

Top species exhibiting the most distinct clusters between populations included *Acetatifactor intestinalis*, *RUG115 sp900066395*, *Fusicatenibacter saccharivorans*, *Gemmiger qucibialis* and *Lachnospira eligens_A* (**Supplementary Fig. 12**). For instance, *Acetatifactor intestinalis*, showed two distinct clusters between Chinese and non-Chinese populations (Fisher exact test, $P < 2.2 \times 10^{-16}$) (**Fig. 5e, Supplementary Fig. 13a**). To further characterize the genomic components potentially driving the observed genetic stratification in *A. intestinalis* between populations, we correlated 41,928 SNV sites with the two t-SNE (t-distributed stochastic neighbor embedding, see Methods) genetic coordinates of the species. This analysis revealed 9,155 associations (FDR < 0.05) corresponding to 2,421 coding sequences (CDSs) (**Fig. 5f; Supplementary Fig. 13b; Supplementary Table 13**). By further annotating these CDSs, we observed a wide range of encoded functionalities, from metabolic processes to antibiotic resistance (**Supplementary Table 13**). For example, we identified 3 out of 15 nonsynonymous SNV sites on the *mdtC* gene that were significantly associated with the two t-SNE coordinates based on the species-level genomic dissimilarity matrix (**Fig. 5g**). The *mdtC* gene encodes a protein known as a multidrug transporter, which confers antibiotic resistance by either exporting antibiotics out of the cell or reducing their toxicity within the cell[29].

We also observed significant genetic stratification in *RUG115 sp900066395* between populations (**Fig. 5h**; **Supplementary Fig. 14a**), which is driven by 8,232 out of 63,094 SNVs detected in this species (FDR < 0.05, **Fig. 5i**; **Supplementary Fig. 14b**), including SNVs located in virulence genes. For instance, three SNV sites in the *clpC*

gene (**Fig. 5j**), a well-known virulence gene[30], were significantly associated with the two t-SNE coordinates of *RUG115 sp900066395* (**Supplementary Fig. 15a-d**). These findings demonstrate that commonly presented species between populations harbor population-specific genomic components in relation to functionalities such as antibiotic resistance and virulence, which might be relevant for regional health disparities.

Within-species genetic diversity is associated with host physiological traits and lifestyles

In addition to the genomic stratification of predominant species between populations implanted potential health-relevant encoded functionalities for regional health disparities, we further checked if within-species genomic variations also related to host physiology traits. For this, we associated the within-species genetic dissimilarities between pairs of individuals and their phenotypic differences in CGMR. We first filtered the cohort to identify species for which genomes could be constructed in at least 1% individuals, resulting in 451 species. We then performed correlation analysis between the genetic distances of these species and the phenotypic traits of host, identifying 1,605 significant associations at FDR < 0.05 between 368 species and 22 phenotypes (**Fig. 6a; Supplementary Table 14**).

In general, within-species genetic dissimilarities related host phenotypes mainly enriched for intrinsic factors such as age, BMI and dietary habits (**Fig. 6a**). Besides, we observed that participants with more differences in their phenotypes tended to show more variation in the genetic makeups of gut microbial species. For instance, we found that more differences in the frequency of probiotic consumption were associated with more genetic variations in *Ruminococcus_B gnavus* ($r_{spearman}$ = 0.14, P = 9.52×10^{-30}, **Fig. 6b**). Another example is we observed that variations in BMI, which are known to influence cardiovascular metabolism, were associated with the genetic dissimilarities of *Bacteroides fragilis_A* ($r_{spearman}$ = -0.11, P = 1.05×10^{-5}, **Fig. 6c**).

Given the large within species genetic diversity observed, we further investigated whether these bacteria could form sub-clades, hypothesizing that main types of

microbial strains with similar genetic makeup may lead to unique functional attributes that potentially influence host physiology. For 273 out of 451 species, we could see distinct subspecies clades based of their phylogenetic trees, including 59 *Lachnospiraceae* members. By further linking species clades with above mentioned host phenotypic traits, we characterized 66 associations between 56 species with distinct sub-clades and 20 host phenotypic traits (FDR < 0.05, **Supplementary Table 15**), mainly related to age and dietary habit.

The strongest association with age was observed for *Faecalibacterium prausnitzii_D*, a genetic diverse species that widely associated with host cardio metabolism[5], which stratifies into two clades (**Fig. 6d**). After comparing the ages of hosts in the two clades, adjusting for gender and BMI, we found that the age of hosts in *F. prausnitzii_D* clade 2 was significantly higher than in clade 1 ($P = 9.8\times10^{-7}$, **Fig. 6e**). To further zoom into the potential functionalities of the for *F. prausnitzii_D* species clades, we checked associations between SNVs driving the populational stratification and age of their host, identifying 77 SNVs significantly associated with age (FDR < 0.05, **Supplementary Table 16**), including genes responsible for enzyme synthesis and energy metabolism. For example, two synonymous SNVs (at positions 150 and 1,254) in the *cbiA* gene (**Fig. 6f&g**), strongly correlated with age. *CbiA* encodes cobyrinate a,c-diamide synthase, a key enzyme in vitamin B12 biosynthesis[31]. Vitamin B12 is a key factor linked to age-related health conditions[32, 33]. These results underscore the critical role of microbial strains, rather than species in host-microbe interactions and their potential influence on physiological traits of their host.

Microbial genetic diversity is unlikely biased by abundance variations

We have observed that inter-individual variations in both the gut microbial species abundance and genetic makeup are associated with host physiological traits and lifestyles. We then questioned if the observed microbial genetic associations to host phenotypic traits are biased by their abundance differences, or the two layers of microbial variations showed independency.

To test this, by screening for species with both abundance and MAGs in our CGMR

cohort (945 species), and filtering those profiled in at least 10 participants, we retained 232 species (**Fig. 7a &b**). For these species, 74.14% (172 out of 232) of them with their genomic dissimilarity not significantly associated with abundance difference (FDR > 0.05, **Fig. 7c; Supplementary Table 17**), suggesting that microbial genetic variability provides an extra layer of information which is independent of microbial abundance. In contrast, we observed that 60 species (25.86%) with their genetic dissimilarities associated with abundance difference (FDR < 0.05, **Fig. 7 c&d**). Among these, the within-species genomic dissimilarity and abundance of 43 species (71.7%) were simultaneously correlated with at least one phenotype (**Supplementary Table 18**).

For instance, we observed differential genomic dissimilarity of *Bacteroides caccae* and variations in urban or rural ($P_{Wilcoxon} = 2.31\times10^{-5}$, **Fig. 7e; Supplementary Table 19**). We also found differential overall abundance of *B. caccae* and residential location ($P_{Wilcoxon} = 1.08\times10^{-7}$, **Fig. 7f; Supplementary Table 19**); However, genomic and abundance dissimilarities were not found to be correlated ($r_{spearman} = -0.008$, $P = 0.25$). Another example is, we observed that *Roseburia hominis*, which genomic dissimilarity not significantly associated with abundance ($r_{spearman} = 0.002$, $P = 0.48$), exhibited a correlation between genomic difference and defecation frequency ($r_{spearman} = 0.02$, $P = 1.16\times10^{-10}$, **Fig. 7g; Supplementary Table 19**), while its abundance associated with defecation frequency ($r_{spearman} = -0.14$, $P = 1.95\times10^{-8}$, **Fig. 7h; Supplementary Table 19**). These results further suggest that for some microbes their relationships with host phenotypic traits are not solely dependent on specific strains but also their abundances.

Microbial SNVs may regulate differential metabolic capacity in *Bacteroides fragilis_A* isolates

Since we have characterized hundreds of associations between within species genetic diversity that driven by SNVs and a diverse range of host phenotypic traits. To further confirm whether within species SNVs are associated with the metabolic capacity of different strains, we focused on *Bacteroides fragilis_A*, a genetic diverse species that

has been shown to be related to host cardiometabolic phenotypes by us recently[10]. In this study, we also found that its genetic diversity is associated with variation in host BMI (**Fig. 6c**), a key risk factor influencing cardiometabolic health. By further plotting the phylogenetic trees of 95 MAGs of this species from our CGMR cohort, as well as three isolates from our lab (FSXLVLS8E1, FSHXHZ1E1, FSXLVWS8E1) with whole genome sequenced, we revealed three distinct clades with one of the three isolates available for each (**Fig. 8a**).

We then carried out *in vitro* fermentation experiments with the 3 genetically different isolated from the *B. fragilis_A* species to assess their differential metabolic capacity potentially driven by SNVs with the same GMM culture medium (see Methods). For each isolate, 10 technical repeats were generated and an amount of 1×10^{8} CFU was injected into a 250mL bottle containing 150 mL medium. The cultures were incubated at 37□ in anaerobic environment. After reaching the logarithmic growth phase, the bacterial cultures were harvested for untargeted metabolomic analysis and a total of 623 metabolites were profiled (**Supplementary Table 20**). Concurrently, we identified 129 SNVs presented in at least two of the three isolates (**Fig. 8b; Supplementary Table 21**). Notably, these SNVs were enriched in genes involved in carbohydrate metabolism, such as beta-galactosidase and glucose-6-phosphate 1-dehydrogenase, suggesting differential energy utilization (**Fig. 8c**).

By associating metabolite levels to SNVs of the 3 isolates, 11,549 significant associations were observed (FDR < 0.05, **Fig. 8d; Supplementary Table 22**), suggesting the potential of microbial SNVs in regulating metabolic capacities. For instance, we found two SNVs in the *BHNJOMMA_04139* gene that were significantly associated with prostaglandin F3α levels (**Supplementary Table 22**), which plays an important role in the regulation of cardiovascular homeostasis and the occurrence of major cardiovascular diseases[34, 35]. Specifically, strains with the T base at position 159 exhibited significantly lower prostaglandin F3α levels compared to those with the A base ($P_{Wilcoxon}=1.1\times10^{-5}$) (**Fig. 8e)**. Similarly, strains with either the C or G base at position 407 also showed significant differences in prostaglandin F3α levels

($P_{Wilcoxon}$=1.1×10^{-5}) (**Fig. 8f).**

Due to the lack of functional annotation for this gene based on Prokka, we then employed AlphaFold3[36] to predict the protein structure and assess the impact of these SNVs. Our analysis revealed that the two strains harbored different amino acids at positions 159 and 407: Asparagine versus Lysine and Alanine versus Glycine, respectively. These amino acid substitutions resulted in modest changes in protein molecular weight, instability index, and grand average of hydropathicity (GRAVY), suggesting alterations in protein stability and hydrophilicity (**Fig. 8g**). Protein structure prediction indicated potential oxidoreductase activity (**Fig. 8g**), targeting the aldehyde or oxygen group of a donor molecule. Considering the multiple oxidation-reduction steps involved in prostaglandin biosynthesis[37], these genetic variants may potentially influence prostaglandin synthesis by altering the function of the encoded protein.

Another example is, we identified a SNV in the *BHNJOMMA_01496* gene. This variant was associated with altered levels of taurochenodeoxycholic acid ($P_{Wilcoxon}$= 1.1×10^{-5}) and taurocholic acid ($P_{Wilcoxon}$= 1.1×10^{-5}) (**Fig. 8 h&i**). Structural modeling using AlphaFold3 predicted that the resulting amino acid substitution (Glutamine to Histidine at position 1,869) disrupted a hydrogen bond with Glutamine 443 (**Fig. 8j**), potentially impacting protein stability and hydrophilicity. Computational analysis suggests that this protein binds carbohydrate derivatives, indicating a role in carbohydrate metabolism. Thus, this SNP may affect carbohydrate processing, potentially influencing bile acid synthesis[38]. Given the critical role of bile acids in cardiovascular diseases, such as coronary heart disease and non-alcoholic fatty liver disease[39-41], our findings suggested that even a single nucleotide change may alter the protein structure and function, thereby impacting host cardiovascular metabolism. Taken together, our study underscores the importance of single nucleotide variation in shaping the metabolic potential of bacterial strains, with potential implications for host cardiovascular health.

Discussion

In this study, we constructed a comprehensive Chinese gut microbiome reference dataset comprising 247,134 middle-to-high quality genomes, significantly expanding our previous CGMR[10] database. By integrating existing datasets and incorporating a large number of Chinese metagenomic samples, we broaden the representation of non-Western populations, a critical gap identified in previous resources. This expansion reveals a unique microbial resource within the Chinese population, highlighting the importance of geographically diverse sampling. While the latest MetaPhlAn4[16] database, a widely used software for microbial comparative analysis, encompasses marker information for over 36,000 species, our integration of the GMR database into it has expanded the database's reach by an additional 579 species. The newly constructed database significantly enhanced our ability to characterize the microbial composition of the CGMR cohort, leading to an average increase of 8.5% in the number of detectable species. Notably, several previously undiscovered species were identified, and their abundance was found to be correlated with specific host phenotypes, including age and dietary habits. Employing an entirely independent Chinese cohort sample dataset[42] for species abundance identification, we also found that the new database can enhance species detection rates by up to 22.8%. This enhanced detection capability, coupled with the identification of significant correlations between newly identified species and host phenotypes, paves the way for more comprehensive investigations into the functional roles of the gut microbiome in health and disease.

We further delved into the genomic diversity of prevalent gut bacteria, revealing pronounced population-level stratification within *Lachnospiraceae* and *Ruminococcaceae*, driven in part by nucleotide variations in antibiotic resistance and virulence genes. Based on our CGMR cohort, we further found that the genomic stratification within these species is correlated with host phenotypes, with more genetic dissimilarity corresponding to larger host phenotype variations. These findings underscore the critical role of strain-level resolution in understanding microbiome-host interactions.

Recent research has found that even a single nucleotide substitution within a bacterial genome can have profound effects on its functionality[43-45]. Single-nucleotide polymorphisms (SNPs) can, for instance, equip bacteria with antibiotic resistance[46] or the ability to infect novel host species[43]. The ubiquity of SNPs in the microbiome[44, 45, 47], has prompted investigations into their potential influence on drug metabolism and individual variations in drug response[48]. In this study, we employed three isolates of *Bacteroides fragilis_A* to corroborate the role of SNPs in modulating bacterial metabolite production. These single-base modifications have the potential to induce alterations in the structural conformation of encoded proteins, thereby impacting their functionality in cardiometabolic health. Our study emphasizes the importance of base-level resolution for understanding host-microbiome interactions.

Despite the significant advancements achieved, our study also underscores ongoing limitations. A substantial portion of the identified species remains uncharacterized, and a large percentage of genes lack functional annotation, suggesting that our understanding of this complex ecosystem is still very limited[49]. In our collection, 36.99% of the species were represented as singletons containing only one genome, limiting the strain level diversity analysis of these species. While we have identified and experimentally validated that specific site mutations in isolates can influence metabolite production, their impact on host health requires further investigation. Moreover, genome-wide association studies (GWAS) involving a more comprehensive set of microbes are warranted and require validation in independent cohorts.

In conclusion, this study presents a substantial contribution to gut microbiome research by establishing a comprehensive GMR database and investigating microbe-host phenotype associations at strain and SNV resolution. By prioritizing underrepresented populations, focusing on functional validation, the GMR collection holds immense potential to advance personalized microbiome-based therapeutic strategies and address global health disparities.

Acknowledgements

We thank the participants and management staffs of the study for their supports. This project was funded by the National Natural Science Foundation of China (NSFC, 32394052, 32270077 and Excellent Young Scientists Fund Program Overseas-2022); Jiangsu Shuangchuang Project (Medical Innovation Team, Medical Expert & JSSCBS20221815); the Natural Science Foundation of Jiangsu (BK20220709); High-level Talent Cultivation Program (CZ0121002010039 & YNRCZN0301) and Major Basic Research Fund (2024) of Jiangsu Province Hospital; Nanjing Medical University (CMCM202204 and GSKY20210105); Suzhou Science and Technology Development Plan Foundation (SKY2021010); Gusu Health Talent Program (GSWS202206); Development of Jiangsu Higher Education Institutions Priority Academic Program (PAPD). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

L.C, Y.Z., X.K., and Q.Z. contributed to conceptualization and funding. The CGMR Consortium Initiative staffs contributed to data and sample collection. Q.D., B.M., X.Z. and L.C. contributed to data analysis. Q.D., L.C., Y.Z., X.K., Q.Z., B.M., X.Z., and P.H. drafted the manuscript. M.G., S.Y., Y.J., Y.Z., Z.S., Q.D., D.H., X.W., L.L., C.Z., C.Z., M.K., C.H., T.W., H.Z., J.S., Y.S., Y.W., L.T., S.H., H.Z., W.S., and W.C. contributed to discussion of the content. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Methods

Overview of the metagenomic datasets

We compiled publicly available prokaryotic genomes sampled from the Chinese gut as of February 2024. The isolates from Chinese individuals are collected from CGR2[15] (https://db.cngb.org/codeplot/datasets/CGR2) and hGMB[14] (https://hgmb.nmdc.cn/) collections, which are the largest microbial culture repositories in China. A total of 104 and 1,804 isolates were collected from these datasets, respectively. The metagenomic data were primarily obtained from studies by Huang et al.[10] (https://db.cngb.org/search/project/CNP0004122/), Qi et al.[19] (https://ehp.niehs.nih.gov/doi/10.1289/isee.2024.0796), Zhang et al.[17] (https://www.ebi.ac.uk/ena/browser/view/PRJEB65297), and Li et al.[18] (https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=772518). In total, we downloaded a total of 76.8 TB of data from 6,657 metagenomic samples and assembled MAGs using the methods described below. In addition, we included a recently published dataset of 6,729 MAGs from Inner Mongolians[11] (https://doi.org/10.6084/m9.figshare.19661523) and the complete set of 289,232 genomes from the UHGG[9] dataset (human gut, v2.0.1, https://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/human-gut/v2.0.1/).

Metagenomic *de novo* assembly and binning

To maximize the quantity of reconstructed genomes, we employed a single-sample assembly strategy following the recommendations of Stephen *et al.*[50] and Pasolli *et al.*[51]. The raw metagenomic reads were pre-processed through a pipeline with the following key steps: 1). trimming of low-quality bases and adapter sequences with Trimmomatic (v.0.33)[52]; 2). removing duplicate reads with Trf (v.4.09)[53]; 3). filtering out human-derived reads by aligning against the human reference genome GRCh37 (hg19) with KneadData (v.0.10.0)[54] integrated Bowtie2 (v.2.1.0)[55]. The clean paired-end reads were then assembled with MEGAHIT (v.1.2.9)[56]. Contigs longer than 1,000 bp were used for genome binning with MetaBAT2 (v.2.12.1)[57]. Briefly, raw reads

were mapped back to the filtered contigs using Bowtie2 (v.2.1.0)[55], and coverage was calculated using the *jgi_summarize_bam_contig_depths* script. Genome binning was subsequently performed with '*metabat*' script. A total of 500,154 genome bins were reconstructed from 6,657 metagenomic samples.

Quality control of genome bins

For newly reconstructed genome bins, genome quality (completeness and contamination) was assessed using CheckM (v.1.1.3)[58] with the '*lineage_wf* ' workflow. We selected only those genomes that met the UHGG criteria: > 50% genome completeness, < 5% contamination, and an estimated quality score (completeness – 5 × contamination) > 50. Including UHGG, there are 478,588 genomes that meet these criteria and were kept for downstream analysis and included in the GMR.

Cluster metagenomic assemblies into species-level genome bins

To address the computational burden of clustering all 478,588 genomes, we employed a two-step approach. First, we used GTDB (R214)[21, 22] to annotate all genomes at the species level. In this process, 1,731, 529, 57, 61, 69, 31, and 0 genomes from CGMR, Zheng, IMGG, PRJEB65297, PRJNA772518, CGR2, and hGMB, could not be annotated to the species level respectively. As UHGG annotations were based on the R202 version, 451 and 2,525 genomes from Chinese and non-Chinese samples, respectively, could not be assigned a species-level annotation. We re-annotated these genomes using the newer R214 version. However, 101 and 899 genomes from Chinese and non-Chinese samples, respectively, still lacked species-level annotations. Finally, we employed dRep[23] to cluster the remaining 3,478 unannotated genomes at a 95% ANI threshold with the following options: '*-pa 0.9*' (primary cluster at 90%), '*-sa 0.95*' (secondary cluster at 95%), '*-cm larger*' (coverage method: larger), '*-comp 50*' larger (completeness threshold of 50%), and '*-con 5*' (contamination threshold of 5%), resulting in 1,762 distinct clusters. Combining these clusters with the 4,902 genomes that received species-level annotations through re-annotation, we obtained a final collection of 6,664 clusters.

The best quality genome from each species cluster was selected as its representative on the basis of genome completeness, minimal contamination and assembly N50 based on the following formula[9]:

$$\text{Score} = \text{completeness} - 5 \times \text{contamination} + 0.5 \times \log(\text{N50})$$

where N50 is the assembly contiguity characterized by the minimum contig size in which half of the total genome sequence is contained. The genome with the highest score was chosen as the species representative, and the final 6,664 genomes was used to generate the GMR representative catalog.

Functional annotation of GMR genomes

For functional annotation, protein-coding sequences (CDS) of each of the 478,588 genomes were predicted and annotated using Prodigal (v2.6.3)[59] and Prokka (v1.13.3)[24], respectively. Prodigal applied the following options: '-c' to predict proteins with closed ends only, '-m' to prevent genes from being built across stretches of sequence marked as Ns, and '-p single' to indicate single mode for genome assemblies containing a single species. Prokka was used to functionally annotate archaeal and bacterial genomes separately, using the '*--kingdom Archaea*' and '*--kingdom Bacteria*' parameters, respectively.

Pangenome generation

Pan-genome analyses were carried out using Roary (v3.12.0)[60]. Roary utilizes a MCL (Markov Clustering) algorithm to cluster orthologous genes based on BLASTP results and identify core and accessory genes within the genomes of interest. For all genomes in each species cluster, we constructed a pan-genome with a minimum amino acid identity for a positive match at 90% (*'-i 90'*), a core gene defined at 90% presence (*'-cd 90'*) and no paralog splitting (*'-s'*) as UHGG criteria. For species with more than 1,000 genomes, we set the parameter *-g* 50000.

Single nucleotide variant (SNV) calling

For SNV analysis, we employed Snippy (v.4.6.0, freely available on GitHub,

https://github.com/tseemann/snippy), a rapid and accurate tool designed for the analysis of single nucleotide polymorphisms (SNPs) in bacterial genomes. Snippy utilizes a mapping-based approach to identify SNPs and insertion-deletion events (indels) in draft or complete genomes. For each species cluster, we employed the *pan_genome_reference.fa* file generated by Roary as the reference genome, considering it a comprehensive and reliable reference for that particular species. We set the minimum number of reads covering a site to be considered as 5 ('*--mincov 5*').

Genomic dissimilarity

For each genome, we focused exclusively on the '*snp*' variants in the Snippy output file, temporarily disregarding '*ins*', '*del*', and '*complex*' classifications. Since the reference genome used was the same for all genomes within a species, the variant sites were consistent across genomes, allowing us to merge the variant sites for all genomes of a species into a single matrix. In this matrix, each row represented a variant site, and each column represented the base type of a specific genome at that site. Each column (sample) was treated as an individual sequence, and these sequences were aligned to generate a phylogenetic distance matrix. The pairwise genetic distance between samples were calculated using the Kimura 2-parameter method from the EMBOSS package[61]. In the case of *NA* values representing distances, which indicates distances between genomes that are sufficiently large to be un-computable, we fill these with the maximum value from the distance matrix plus one.

Phylogenetic analyses

A total of 2,145 species are shared between Chinese and non-Chinese populations, encompassing 456,636 genomes. To enhance the robustness of our analysis, we focused only on species with more than 100 genomes. Furthermore, species for which the proportion of genomes from either population exceeded 30% were selected, thereby minimizing potential bias. This resulted in the identification of 304 species, comprising 306,525 genomes, for phylogenetic analysis. For each microbial genome, we identified SNVs and calculated genetic distances using the aforementioned methods. Hierarchical clustering was performed using the *hclust()* function in R, and

phylogenetic trees were constructed with the *as.phylo()* function in package *ape* (v.5.8.0). Trees were visualized and annotated with Interactive Tree Of Life (iTOL)[62] (https://itol.embl.de/itol.cgi) and ggtree (v.3.12.0)[63]. The *cutree()* function was used to divide the trees into distinct groups.

The GMR unique species-level marker genes

For species with fewer than 3,000 genomes, pan-genomes were constructed using all available genomes with Roary. For the 29 species with more than 3,000 genomes, the top 3,000 genomes, ranked by score (completeness – 5 × contamination + 0.5 × log(N50)), were selected for pan-genome construction. Core genes were identified from Roary's *gene_presence_absence.Rtab* files, with coreness thresholds based on the number of genomes: 80% for species with fewer than 10 genomes, 60% for species with 10 to 100 genomes, and 50% for species with over 100 genomes. Species with fewer than 200 core genes were excluded from further analysis.

To identify marker genes, we employed the criteria established by MetaPhlAn4[16] for marker gene evaluation. Core genes for each species were aligned against the genomes of other species using Bowtie with the '*--sensitive*' parameter. For computational efficiency, we selected up to 100 of the highest-quality genomes per species for alignment. Core genes were split into 150-nt fragments to simulate metagenomic reads, which were then mapped against representative genomes. A core gene fragment was considered a hit if it mapped to its species' genomes above the coreness threshold and mapped to fewer than 1% of genomes from other species (quasi-markers) or not at all (perfectly unique markers). A maximum of 200 marker genes per species were selected, prioritized by uniqueness and length (longer genes first). Species with fewer than 10 markers were excluded.

This process produced a marker gene database comprising 2,777 species. We compared these species against the MetaPhlAn4 *mpa_vJun23_CHOCOPhlAnSGB_202403* database using the *mpa_vJun23_CHOCOPhlAnSGB_202403_SGB2GTDB.tsv* file. This comparison identified 579 species missing from the MetaPhlAn4 database. To create a more

comprehensive resource, we integrated our marker genes with those from MetaPhlAn4 for subsequent analyses.

Profiling microbiome composition

To assess the performance of our customized database in profiling gut microbial composition, we analyzed 1,076 publicly available fecal metagenomic samples from an independent Chinese cohort[26]. Taxonomic profiling was performed using MetaPhlAn4[16] (v.4.1.0) with default parameters and our custom database. The same analysis was conducted for the CGMR cohort.

In vitro cultivation of *B. fragilis_A* isolates

Three *B. fragilis_A* isolates (FSXLVLS8E, FSHXHZ1E1, FSXLVWS8E1, 1×10^8 CFU) were inoculated into anaerobic GMM medium and cultured in 250mL bottles containing 150 mL of medium. The cultures were incubated at 37□ in anaerobic environment. After reaching the logarithmic growth phase, the bacterial cultures were harvested. To collect the bacterial fluid, the cultures were centrifuged at 12,000 g for 5 min. The resulting supernatant, which contains the secreted components, was carefully separated and stored at -80□ for further analysis.

Metabolome analysis

We have collected fermentation products from *B. fragilis_A* isolates for metabolome analysis. For culture supernatants, 20mL samples were mixed with 180mL of 80% methanol and vortexed for 15s. All mixtures were then incubated at 4□ for 1h to precipitate proteins and subsequently centrifuged at 14,000 g for 15 min. The resulting supernatants were collected and stored at -80□ for metabolome analysis.

The untargeted metabolome analysis was performed using an LC-MS/MS system comprising a ThermoFisher Vanquish UHPLC system coupled with an Orbitrap Q ExactiveTMHF mass spectrometer. Samples were injected onto a Hypesil Gold column (100 × 2.1 mm, 1.9μm) using a 17-min linear gradient at a flow rate of 0.2mL/min. The eluents for the positive polarity mode were eluent A (0.1% formic acid in water) and eluent B (methanol). The eluents for the negative polarity mode

were eluent A (5 mM ammonium acetate, pH 9.0) and eluent B (methanol). The solvent gradient was set as follows: 2% B, 1.5 min; 2–85% B, 3 min; 100% B, 10 min; 100-2% B, 10.1 min; 2% B, 12 min. Q ExactiveTMHF mass spectrometer was operated in positive/negative polarity mode with spray voltage of 3.2 kV, capillary temperature of 320□, sheath gas flow rate of 40 arb and aux gas flow rate of 10 arb, Funnel RF level of 40, Aux gas heater temperature of 350□. The raw data files generated by UHPLC-MS/MS were processed using Compound Discoverer 3.1 (CD3.1, ThermoFisher) software for peak alignment, peak picking, and quantitation of each metabolite. The normalized data was used to predict the molecular formula based on additive ions, molecular ion peaks, and fragment ions. The identified peaks were matched with databases such as mzCloud, mzVault, and MassList to obtain accurate qualitative and relative quantitative results. Metabolite annotations were performed using databases such as KEGG, HMDB, and LIPIDMaps.

Protein structure and function prediction

To gain insights into the three-dimensional structures and potential functions of proteins of interest, we employed a combination of state-of-the-art computational tools, namely AlphaFold3[36] and DeepFRI[64]. AlphaFold3 is a powerful protein structure prediction tool developed by DeepMind, which utilizes deep learning algorithms to predict protein structures with high accuracy. Once the protein structures were predicted using AlphaFold3, we utilized deepFRI to infer potential functions for these proteins. DeepFRI takes the protein sequence and structure as input and predicts the functions using a pre-trained graph-based neural network model. The output from deepFRI includes a list of predicted functions with confidence scores for each protein. Protein structures were visualized using PyMOL(https://pymol.org/). The physicochemical properties of proteins can be predicted using the ProtParam tool (https://web.expasy.org/protparam/).

Statistical analysis

Comparison of microbial genomes from Chinese and non-Chinese populations

The GMR dataset contains 478,588 microbial genomes, including 247,134 from the

Chinese population and 231,454 from the non-Chinese population. The median genome sizes for the two groups were 2.27Mb and 2.21Mb, respectively, with N50 lengths of 30Kb and 27Kb. Comparison between the two groups were assessed using the Wilcoxon test.

Comparison of reads mapping

After integrating GMR-specific marker genes into the MetaPhlAn4 database, an independent Chinese cohort[26] was used for taxonomic profiling. Paired Wilcoxon tests were employed to compare the reads mapping rates between groups.

Microbial abundance associations to human phenotypic traits

We first performed a centered log-ratio transformation on the taxonomic data and performed Spearman correlation analysis with the phenotypic data. P values were adjusted using the Benjamini–Hochberg method, and associations with an FDR-adjusted P value < 0.05 were considered significant. For species associated with multiple phenotypes, we used MaAsLin2 for multivariable association analysis with default settings, treating the phenotypes as fixed effects. MaAsLin2 was also employed to identify significant associations while adjusting for potential confounders.

Replication of microbiome–phenotype associations

To replicate newly detected microbiome composition–phenotype associations in CGMR, we used data from 551 patients from the Qi *et al.*[19] as a replication cohort. We identified 318 associations across 20 overlapping phenotypes. To assess the variability of these associations, Cochran's Q tests were performed to evaluate heterogeneity in effect sizes and directions between the two cohorts, using the *metagen()* function from the meta package (v.7.0.0) in R. Associations with an FDR-adjusted Cochran's Q test P value < 0.05 were considered heterogeneous.

The correlation between microbial genomic and phenotypic dissimilarities

We first computed the genetic distances between all pairs of genomes within each

species using the aforementioned methods, and then analyzed the phenotypes of the hosts from which each genome originated. Specifically, we represented the differences between host phenotypes using the Manhattan distance. Subsequently, we calculated the Spearman's correlation coefficient between the two matrices. The Benjamini-Hochberg procedure was applied to control FDR.

The correlation between genetic distance and abundance variations

We computed genetic distances between all genome pairs within each species using the methods described above. Abundance variation was calculated using Manhattan distance. Next, we computed Spearman's correlation coefficient between the genetic distance and abundance variation matrices. The Benjamini-Hochberg procedure was used to control the FDR.

Genetic dissimilarity comparison

To assess population stratification for each species, we divided the genetic distance matrix into three groups: within the Chinese population, within the non-Chinese population, and between populations. We then performed a Kruskal-Wallis test to compare the genetic distances across these groups. The Benjamini-Hochberg procedure was used to control the FDR.

t-SNE analysis

Dimensionality reduction of the genetic distance matrix for each species was performed using the R *Rstne()* function, resulting in two vectors. Each SNV site was then subjected to Spearman correlation analysis with tSNE1 and tSNE2. The Benjamini-Hochberg procedure was applied to control FDR.

Enrichment of species showed significant population stratification

We analyzed genetic variation in 304 species, identifying population stratification in 75 of them. To determine whether these 75 species were enriched at a specific taxonomic level, we performed a hypergeometric test at the family level. The GMR dataset contains 6,664 species across 285 families, with the 75 stratified species

distributed among 19 families. We used the R *phyper()* function to assess the enrichment of species within each family.

Data availability

The reference genomes of 6,664 representative species from GMR have been uploaded to the National Genomics Data Center (NGDC) with the accession number OMIX006649. The metabolome data of isolates are deposited in the National Genomics Data Center (NGDC) with accession number OMIX004554. The custom MetaPhlAn4 database can be download from https://zenodo.org/records/13946308.

Code availability

The workflow used to generate the MAGs from metagenomic data and statistical analysis codes are available via https://github.com/MicrobiomeCardioMetaLab/GMR_project.

Figure Legends

Fig. 1 | Establishment of an integrated catalog of 478,588 microbial reference genomes in the human gut.

a. Workflow and bioinformatic pipeline for constructing the GMR database. We assembled 180,719 high quality MAGs from 6,657 Chinese metagenomic samples with a total amount of 76.8 TB clean sequences. By integrating these MAGs with existing databases, we constructed the GMR, comprising 478,588 microbial genomes from 6,664 species, 26.4% of which are newly identified.

b. Summary of the microbial genomes in the GMR. The inner pie chart shows the proportion of genomes derived from Chinese (247,134 genomes, 51.6%) and non-Chinese populations. The outer pie chart highlights the distribution of genomes from different data sources.

c. Distribution of microbial genomes quality scores in the GMR. The X-axis represents the completeness of genomes, while the Y-axis represents contamination levels. Genomes from Chinese samples are shown in red, and those from non-Chinese samples are in blue. Color intensity reflects the number of genomes.

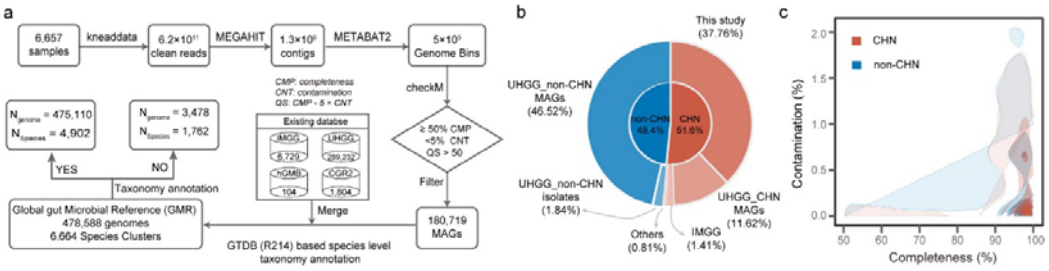


Fig. 2 | Taxonomic classification of the GMR reveals 6,664 species clusters with 1,762 identified as potential novel species.

a. Pie chart illustrating the proportion of classified and novel species in the GMR. Taxonomically classified species account for 73.56%, while the remaining represent newly identified or novel species.

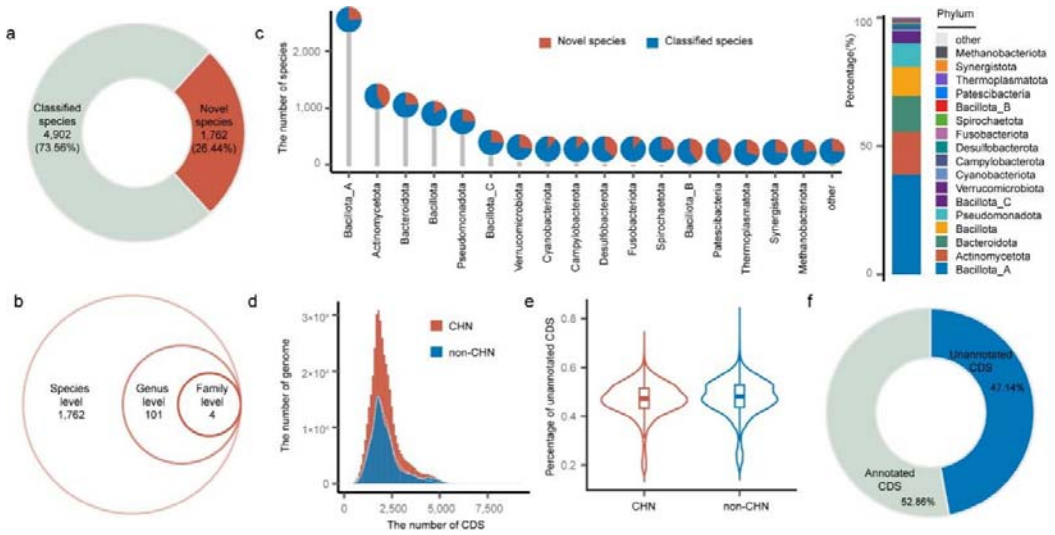
b. Breakdown of novel species at the genus and family levels. Of the 1,762 species unassigned to known species, 101 could not be classified at the genus level, and 4 remain unclassified at the family level.

c. Phylum-level classification of species in the GMR. The left bar chart shows the number of classified and novel species within each phylum. Each bar is topped with a pie chart showing the proportion of novel species and classified species within that phylum. Red represents novel species, while blue represents classified species. The right bar chart presents the proportion of species within each phylum, with phyla containing fewer than 10 species grouped under other.

d. Distribution of the total number of CDSs (coding sequences) per microbial genome from Chinese and non-Chinese populations. On average, genomes from Chinese and non-Chinese populations contained 2,126 and 2,172 CDSs per genome, respectively.

e. Comparison of un-annotatable CDS percentages between Chinese and non-Chinese populations. Boxplots display the median values (center lines), interquartile ranges (IQRs, represented by the box edges), and whiskers, which extend to the most extreme data points within 1.5 times the IQR from the lower and upper quartiles.

f. Pie chart showing the proportion of functionally annotated versus unannotated CDSs. Based on the UniProt, RefSeq, Pfam and TIGRFAMs databases, 47.14% of CDSs remain unannotated.


a
Classified species 4,902 (73.56%)
Novel species 1,762 (26.44%)
b
Species level 1,762
Genus level 101
Family level 4
c
Novel species
Classified species
The number of species
Bacillota_A
Actinomycetota
Bacteroidota
Bacillota
Pseudomonadota
Bacillota_C
Verrucomicrobiota
Cyanobacteriota
Campylobacterota
Desulfobacterota
Fusobacteriota
Spirochaetota
Bacillota_B
Patescibacteria
Thermoplasmatota
Synergistota
Methanobacteriota
other
Percentage(%)
Phylum
other
Methanobacteriota
Synergistota
Thermoplasmatota
Patescibacteria
Bacillota_B
Spirochaetota
Fusobacteriota
Desulfobacterota
Campylobacterota
Cyanobacteriota
Verrucomicrobiota
Bacillota_C
Pseudomonadota
Bacillota
Bacteroidota
Actinomycetota
Bacillota_A
d
CHN
non-CHN
The number of genome
The number of CDS
e
Percentage of unannotated CDS
CHN
non-CHN
f
Unannotated CDS 47.14%
Annotated CDS 52.86%

819

820

Fig. 3 | Integrating microbial genomes from newly classified species into MetaPhlAn4 enhances taxonomic profiling.

a. Distribution of microbial genomes across species-level clusters. The inner circle represents the proportion of novel and classified species within each level. A total of 3,270 species clusters contains at least 3 genomes.

b. Summary of core and marker genes identified in each species. On average, each species contains 1,626 core genes and 676 marker genes.

c. Overlap between GMR species (with at least 10 species-specific marker genes) and species in the MetaPhlAn4 database. The GMR contributes additional species-specific marker genes for 579 species, complementing the MetaPhlAn4 database.

d. Taxonomic profiling at the species level using the newly constructed database in an independent Chinese cohort with diverse filtering cutoffs. A total of 1,076 publicly available fecal metagenomic samples were analyzed. The X-axis shows different filtering cutoffs from 0% to 10%. The left Y-axis (bar chart) represents the number of detected species, with the red portion highlighting the contribution of newly identified species in the GMR. The right Y-axis (blue line) represents the improvement in species detection achieved by including the newly identified species from the GMR.

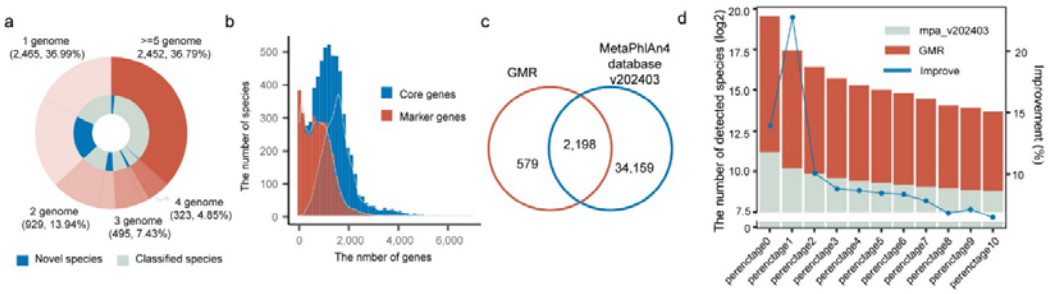


Fig. 4 | Abundance of newly classified species is widely associated with host intrinsic traits and dietary habits across independent cohorts.

a. Summary of microbial species profiled in a cohort of 3,234 participants. The blue bars represent the total number of detected species, while the red bars indicate the number of newly classified species in the GMR. On average, 201 species were detected per sample, with 11 belonging to newly classified species in the GMR. The pie chart on the right illustrates the proportion of samples with at least one newly detected species, showing that 97.8% of the samples contained at least one new species.

b. Host phenotypic associations with the abundance of newly profiled species. Bars represent different phenotypic traits, sorted in descending order by the number of significant associations. The direction of the associations is color-coded: red indicates positive associations, while blue represents negative associations. The length of each bar reflects the number of significant associations (FDR < 0.05).

c. Replication of 318 significant associations in an independent cohort of 551 participants. The Y-axis shows P-values from the Cochran-Q test for phenotypic associations with microbial species, while the X-axis displays the 318 associations ordered by heterogeneity test P-values. The pie chart in the upper right shows the proportion of replicable associations, with blue indicating P-values < 0.05 and red indicating P-values > 0.05 in the Cochran-Q test.

d. The relative abundance of *Pseudoflavonifractor_A sp018374635* was negatively associated with age. The X-axis shows participant ages, and the Y-axis represents species abundance. The solid line represents linear regression, with gray shading indicating the 95% confidence interval. Each dot represents a participant.

e. The relative abundance of *Streptococcus infantis_I* was significantly higher with increased defecation frequency. The box plot compares species abundance across different defecation frequencies. The boxes represent the median, first, and third quartiles (25th and 75th percentiles), with whiskers extending to the most extreme data points within 1.5 times the interquartile range. Each dot represents a participant.

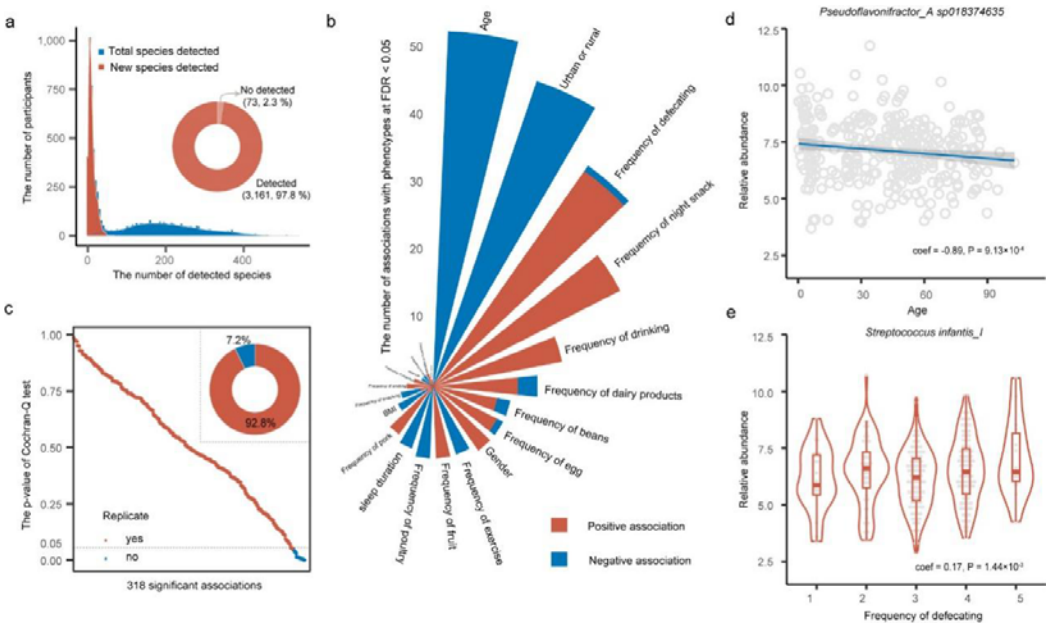

a
Total species detected
New species detected
No detected (73, 2.3 %)
Detected (3,161, 97.8 %)
The number of participants
The number of detected species
b
The number of associations with phenotypes at FDR < 0.05
Age
Urban or rural
Frequency of defecating
Frequency of night snack
Frequency of drinking
Frequency of dairy products
Frequency of beans
Frequency of egg
Gender
Frequency of exercise
Frequency of fruit
Frequency of poultry
sleep duration
Positive association
Negative association
c
7.2%
92.8%
The p-value of Cochran-Q test
Replicate
yes
no
318 significant associations
d
Pseudoflavonifractor_A sp018374635
Relative abundance
Age
e
Streptococcus infantis_I
Relative abundance
Frequency of defecating

Fig. 5 | Commonly presented species between populations showed genetic stratification in relation to antibiotic resistance and virulence.

a. Overlap of microbial species clusters between Chinese and non-Chinese populations. In general, 2,145 species have been found to be presented in both Chinese and non-Chinese populations.

b. Summary of MAGs per species clusters from Chinese and non-Chinese populations. The bar plot on the left depicts the 304 species clusters with at least 100 MAGs and a proportion of at least 30% MAGs in either Chinese and non-Chinese populations. The scatter plot on the right shows the proportion of MAGs per species cluster from Chinese and non-Chinese populations. The darkness of color gradient and the size of the dots indicate proportions.

c. Summary of microbial species with significant genetic stratification between populations at family level. The X-axis represents the total number of species within each family, while the Y-axis shows the number of species showed significant genetic stratification between Chinese and non-Chinese populations. Each dot represents one family with the size of the dots indicates P-value from Hypergeometric test, and the darkness of color represents the proportion of species showed significant stratification between Chinese and non-Chinese populations within each family. Among the 23 families examined, only two families exhibited a significant correlation at an FDR threshold of less than 0.05 and was subsequently labeled with its corresponding family name.

d. Within and between genomic dissimilarities in 74 species based on microbial SNVs. Red indicates genetic distance within the Chinese population, while blue indicates genetic distance within the non-Chinese population. Grey indicates genetic distance between the two populations. The boxes represent the median, first, and third quartiles (25th and 75th percentiles), with whiskers extending to the most extreme data points within 1.5 times the interquartile range.

e. Phylogenetic tree of *Acetatifactor intestinalis* in GMR. *A. intestinalis* MAGs from Chinese and non-Chinese populations showed genetic stratification based on SNV

level dissimilarity matrix. Each leaf represents a MAG and is colored according to population.

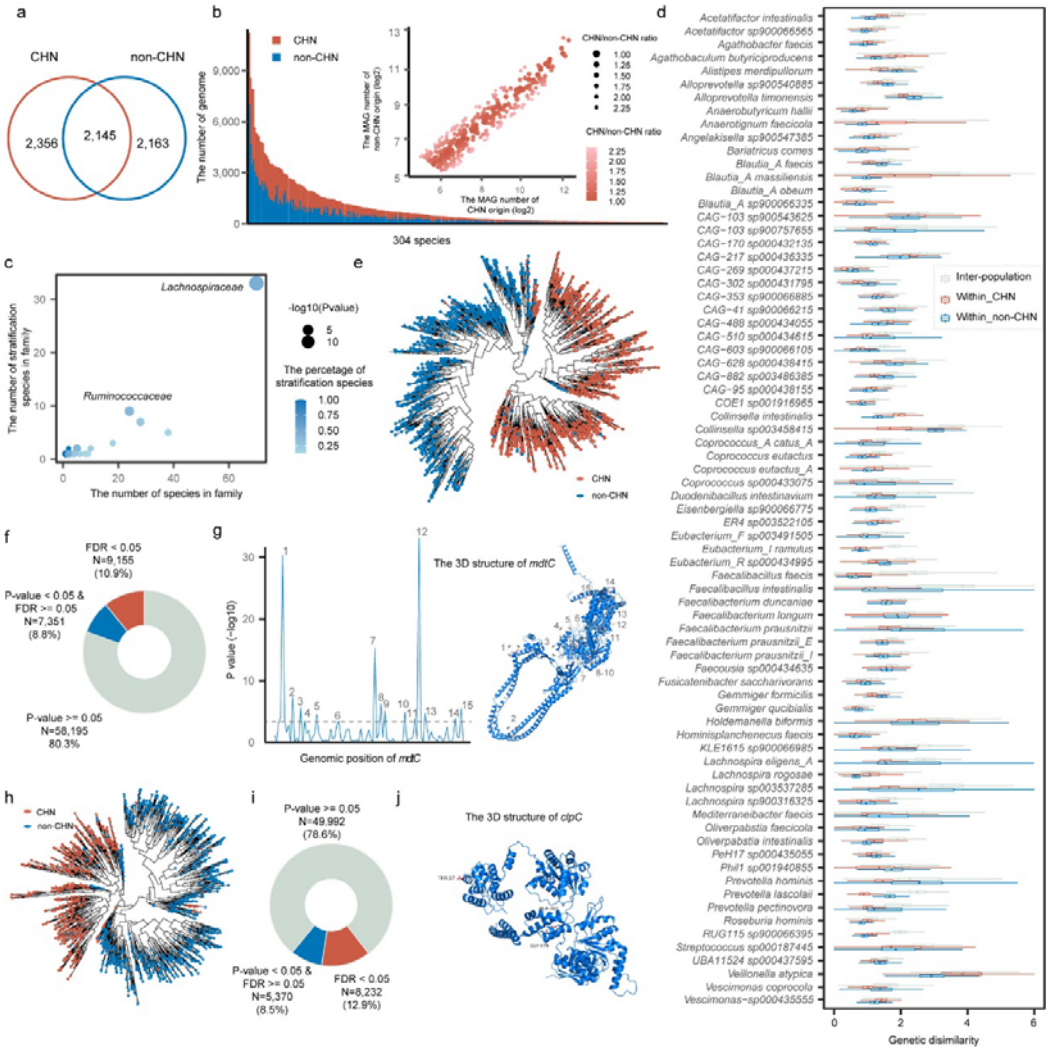
f. Summary of SNVs associations to the two t-SNE components of *A. intestinalis*. A total of 16,506 associations with P-values less than 0.05 were observed, of which 9,155 with FDR less than 0.05.

g. SNV sites on the *mdtC* gene that were significantly associated with the two t-SNE coordinates based on the species-level genomic dissimilarity matrix. The left panel presents a line plot of P-values (Y-axis) across these loci (X-axis). The right panel depicts the 3D structure of the *mdtC* protein, highlighting the positions of the 15 SNVs that potentially impact the corresponding amino acid residues.

h. Phylogenetic tree of *RUG115 sp900066395* in GMR. *RUG115 sp900066395* MAGs from Chinese and non-Chinese populations showed genetic stratification based on SNV level dissimilarity matrix. Each leaf represents a MAG and is colored according to population.

i. Summary of SNVs associations to the two t-SNE components of *RUG115 sp900066395*. A total of 13,502 associations with P-values less than 0.05 were observed, of which 8,232 with FDR less than 0.05.

j. The 3 SNVs located in a virulence gene *clpC* that associated with the t-SNE components of *RUG115 sp900066395*.


a
CHN
non-CHN
2,356
2,145
2,163
b
CHN
non-CHN
The number of genome
304 species
The MAG number of non-CHN origin (log2)
The MAG number of CHN origin (log2)
CHN/non-CHN ratio
c
Lachnospiraceae
Ruminococcaceae
The number of stratification species in family
The number of species in family
-log10(Pvalue)
The percetage of stratification species
e
CHN
non-CHN
d
Inter-population
Within_CHN
Within_non-CHN
Genetic disimilarity
f
FDR < 0.05
N=9,155
(10.9%)
P-value < 0.05 &
FDR >= 0.05
N=7,351
(8.8%)
P-value >= 0.05
N=58,195
80.3%
g
P value (-log10)
The 3D structure of mdtC
Genomic position of mdtC
h
CHN
non-CHN
i
P-value >= 0.05
N=49,992
(78.6%)
P-value < 0.05 &
FDR >= 0.05
N=5,370
(8.5%)
FDR < 0.05
N=8,232
(12.9%)
j
The 3D structure of clpC

919

920

Fig. 6 | Within species genetic diversity is associated with host physiological traits and lifestyles.

a. Summary of the within-species genetic dissimilarities between pairs of individuals and their phenotypic differences in the CGMR cohort. The heatmap depicts the correlation between within species genetic diversity and host phenotypic traits. The darkness of each cell indicates the Spearman correlation coefficient and asterisks indicate significant correlations with FDR < 0.05.

b. More differences in the frequency of probiotic consumption were associated with more genetic variations in *Ruminococcus_B gnavus*. The boxes represent the median, first, and third quartiles (25th and 75th percentiles), with whiskers extending to the most extreme data points within 1.5 times the interquartile range. Each dot represents a pair of MAGs that derived from different individuals.

c. There is a significant difference in genetic distance between the overweight group and the non-overweight group in *Bacteroides fragilis_A*. Individuals with a BMI greater than 25 are classified as overweight, while those with a BMI of 25 or less are classified as non-overweight. The boxes represent the median, first, and third quartiles (25th and 75th percentiles), with whiskers extending to the most extreme data points within 1.5 times the interquartile range. Each dot represents a pair of MAGs that derived from different individuals.

d. The SNV-based phylogenetic tree reveals that *Faecalibacterium prausnitzii_D* strains cluster into two discrete clades, with clade 1 colored in red and clade 2 in blue. Each leaf represents a MAG derived from an individual.

e. Association between *F. prausnitzii_D* clades and host age. The age of hosts with *F. prausnitzii_D* clade 2 was significantly higher than those with clade 1. Beta values represent the standardized effect sizes estimated by linear regression, adjusted for gender and BMI. The corresponding P-values are shown.

f. Correlation between base type at position 150 of *cbiA* gene and host age. At position 150 of the *cbiA* gene, clade 1 is enriched in the G nucleotide. The P-value from Fisher's exact test is provided. Hosts with the G nucleotide have significantly

lower ages compared to those with the C nucleotide. The boxplots show the median, first quartile (25th percentile), and third quartile (75th percentile), with whiskers extending to the most extreme data points within 1.5 times the interquartile range. Each dot represents an individual.

g. Correlation between base type at position 1,254 of *cbiA* gene and host age. At position 1,254 of the *cbiA* gene, clade 1 is enriched in the C nucleotide, while clade 2 is enriched in the A nucleotide. The P-value from Fisher's exact test is provided. Hosts with the A nucleotide type have significantly higher ages compared to those with the C type. The boxplots show the median, first quartile (25th percentile), and third quartile (75th percentile), with whiskers extending to the most extreme data points within 1.5 times the interquartile range. Each dot represents an individual.

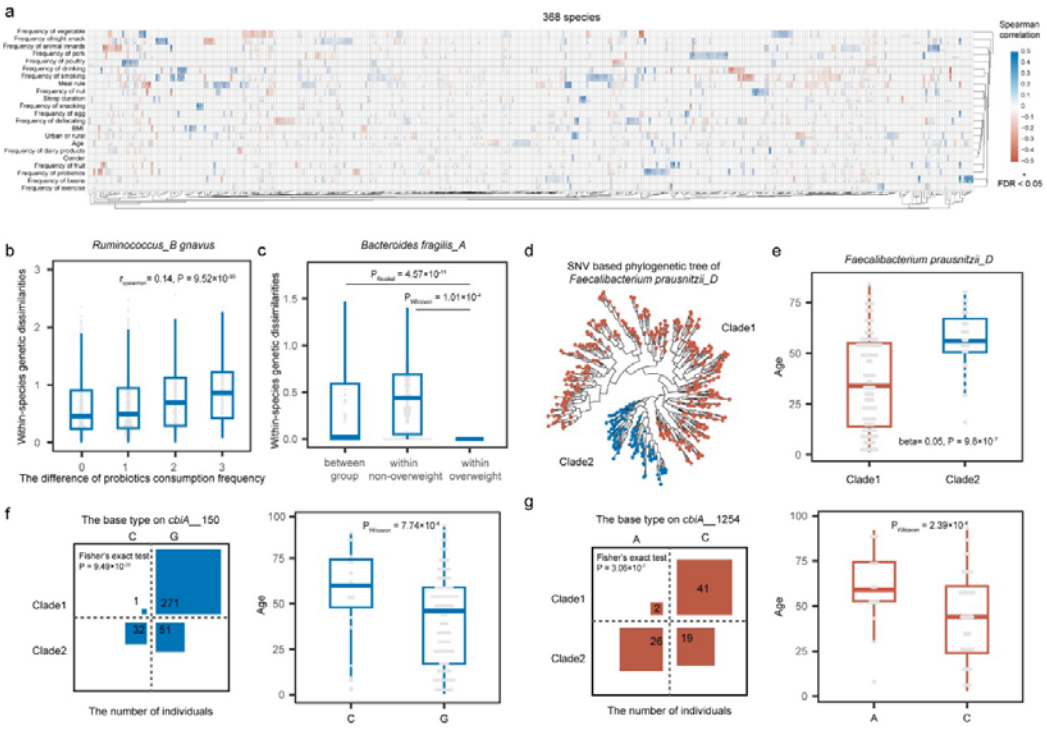


Fig. 7 | Limited association between microbial genetic diversity and abundance variation.

a. Overlap of species detected by MetaPhlAn4 and MAGs assembled from individual metagenomes in the CGMR cohort. A total of 945 species with both information available in our CGMR cohort.

b. Summary of MAGs per species detected by both methods. The bar depicts the 232 species with at least 10 participants with both abundance and MAGs detected.

c. Summary of correlation analysis between species abundance and genome similarity. For these species, 74.14% (172 out of 232) of them with their genomic dissimilarity not significantly associated with abundance difference (FDR > 0.05).

d. Comparison of microbial species differences between abundance and genomic dissimilarity. Each dot represents a microbial species. The X-axis represents the difference in genetic distance, and the Y-axis represents the difference in abundance among all MAGs of a species. Cross bars represent the standard deviations of abundance difference and genomic dissimilarity. Spearman correlation was applied to assess the correlation between the difference of abundance and genomic dissimilarity for each species. 172 species with not a significant correlation (FDR > 0.05) are highlighted as red dots. The size of each dot reflects the FDR value.

e. Genetic difference of *Bacteroides caccae* between individuals from the same different regions. The boxes represent the median, first, and third quartiles (25th and 75th percentiles), with whiskers extending to the most extreme data points within 1.5 times the interquartile range. Each dot represents a pair of individuals.

f. The relative abundance of *B. caccae* was associated with residential area. The box plot compares species abundance across different defecation frequencies. The boxes represent the median, first, and third quartiles (25th and 75th percentiles), with whiskers extending to the most extreme data points within 1.5 times the interquartile range. Each dot represents a participant.

g. The genetic distance of *Roseburia hominis* increases with increasing frequency of defecation. The boxes represent the median, first, and third quartiles (25th and 75th

percentiles), with whiskers extending to the most extreme data points within 1.5 times the interquartile range. Each dot represents a pair of individuals.

h. The relative abundance of *R. hominis* was associated with defecation frequency. The boxes represent the median, first, and third quartiles (25th and 75th percentiles), with whiskers extending to the most extreme data points within 1.5 times the interquartile range. Each dot represents a participant.

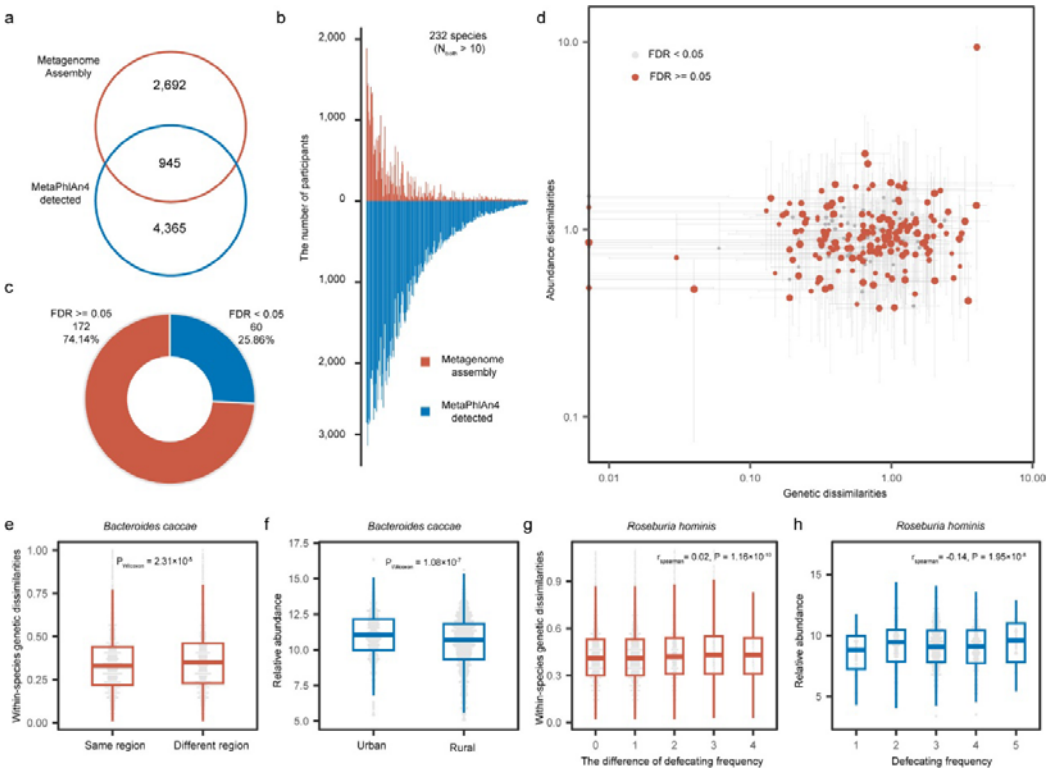


Fig. 8 | Microbial SNVs may regulate differential metabolic capacity in *Bacteroides fragilis_A* isolates.

a. Phylogenetic tree constructed from 95 *B. fragilis_A* MAGs and 3 live isolates. Each leaf represents a MAG derived from an individual or isolates. Different clades are shown in different colors.

b. Summary of detected SNV sites across three *B. fragilis_A* isolates. A total of 129 SNVs were identified among the three isolates. The horizontal bars on the left represent the total number of SNV sites detected in each isolate, while the vertical bars indicate the number of prevalent SNVs shared between isolates.

c. Summary of gene functions enriched at these 129 SNVs sites, including Beta-galactosidase and Glucose-6-phosphate 1-dehydrogenase.

d. Summary of SNV associations with 623 metabolites profiled by untargeted metabolomic analysis of bacterial cultures. A total of 23,580 associations with P-values < 0.05 were identified, of which 11,549 had an FDR < 0.05.

e. The base type at site 159 of the *BHNJOMMA_04139* gene is correlated with levels of prostaglandin F3α. The boxes represent the median, first, and third quartiles (25th and 75th percentiles), with whiskers extending to the most extreme data points within 1.5 times the interquartile range. Each dot represents a strain.

f. The base type at site 407 of the *BHNJOMMA_04139* gene is correlated with levels of prostaglandin F3α. The boxes represent the median, first, and third quartiles (25th and 75th percentiles), with whiskers extending to the most extreme data points within 1.5 times the interquartile range. Each dot represents a strain.

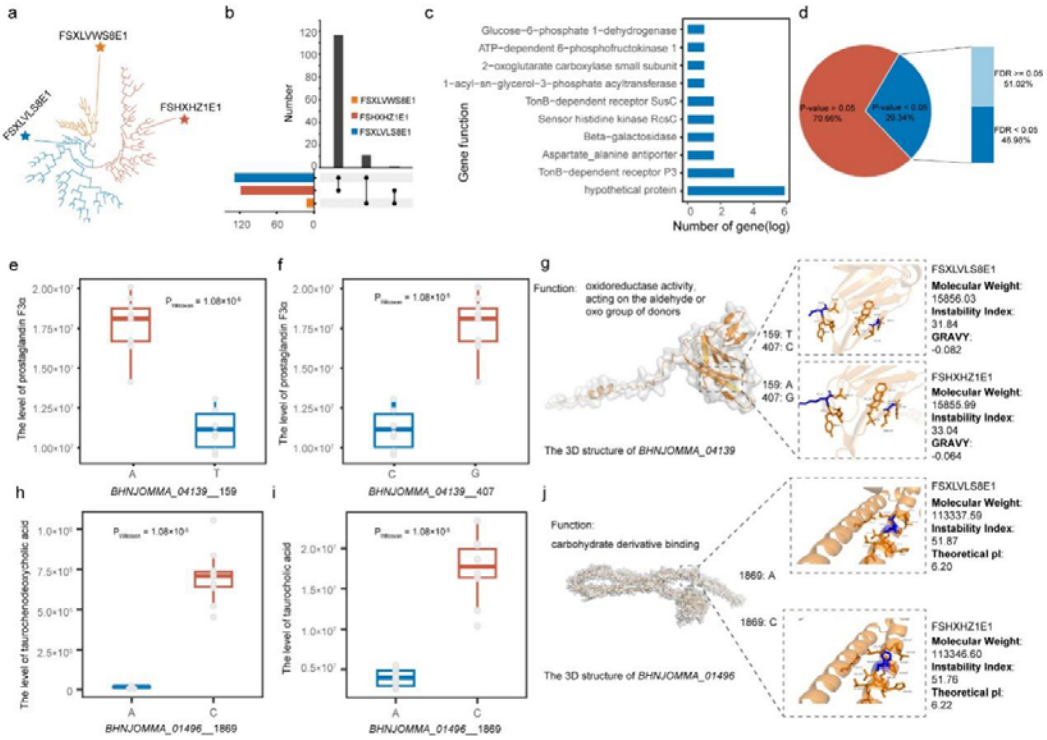
g. 3D structure of *BHNJOMMA_04139* predicted by AlphaFold3. At position 159, a T to C mutation results in an Asparagine to Lysine substitution; at position 407, an A to G mutation results in an Alanine to Glycine substitution. These changes may affect the molecular weight, instability index and grand average of hydropathicity (GRAVY) of this protein.

h. The base type at site 1,869 of the *BHNJOMMA_01496* gene are associated with differences in the levels of Taurochenodeoxycholic acid. The boxes represent the

median, first, and third quartiles (25th and 75th percentiles), with whiskers extending to the most extreme data points within 1.5 times the interquartile range. Each dot represents a strain.

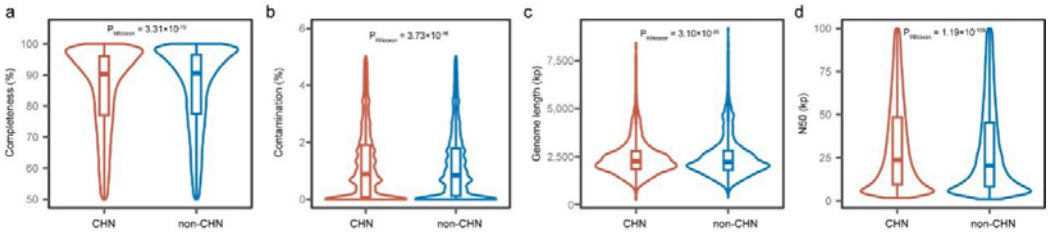
i. The base type at site 1,869 of the *BHNJOMMA_01496* gene are associated with differences in the levels of Taurocholic acid. The boxes represent the median, first, and third quartiles (25th and 75th percentiles), with whiskers extending to the most extreme data points within 1.5 times the interquartile range. Each dot represents a strain.

j. 3D structure of *BHNJOMMA_01496* predicted by AlphaFold3. At position 1,869, an A to C mutation results in a Glutamine to Histidine substitution. These changes disrupted a hydrogen bond with Glutamine 443 and may affect the molecular weight, instability index and grand average of hydropathicity (GRAVY) of this protein.

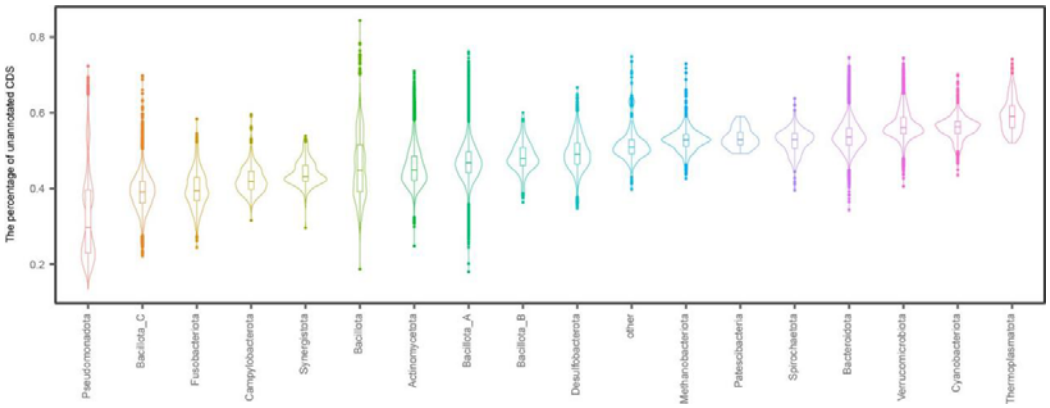


Supplementary Fig.1 | Summary of microbial genomes in the GMR from Chinese and non-Chinesepopulations.

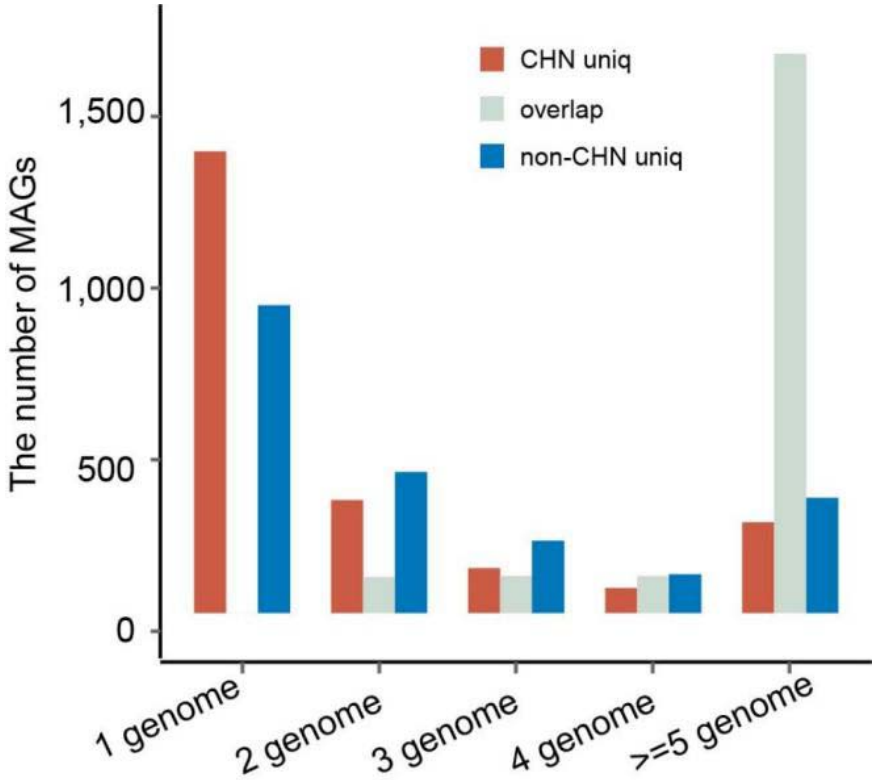
Panels show the distribution of completeness (**a**), contamination (**b**), genome length (**c**), and N50 length (**d**) across the 478,588 genomes included in the GMR. The boxplots display the median values (center lines), interquartile ranges (IQRs, represented by the box edges), and whiskers, which extend to the most extreme data points within 1.5 times the IQR from the lower and upper quartiles.



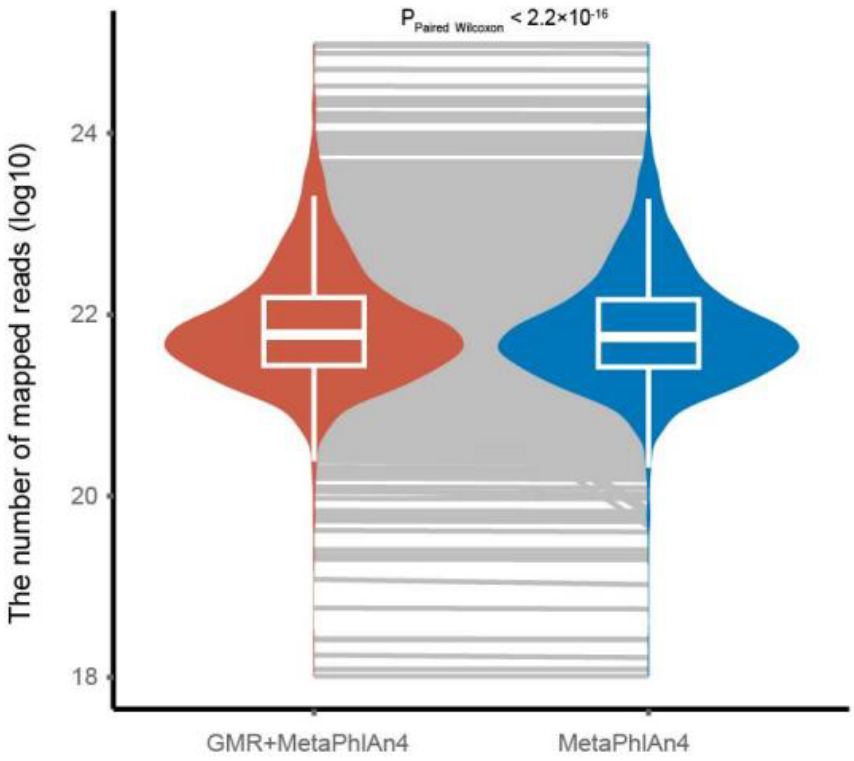
Supplementary Fig. 2 | Phylum-level proportion of functionally annotated versus unannotated CDSs. The bar plot shows the percentage of CDS with unassigned function within each phylum. The phyla comprising < 10 of each genome set was collapsed into other. The boxplots display the median values (center lines), interquartile ranges (IQRs, represented by the box edges), and whiskers, which extend to the most extreme data points within 1.5 times the IQR from the lower and upper quartiles.



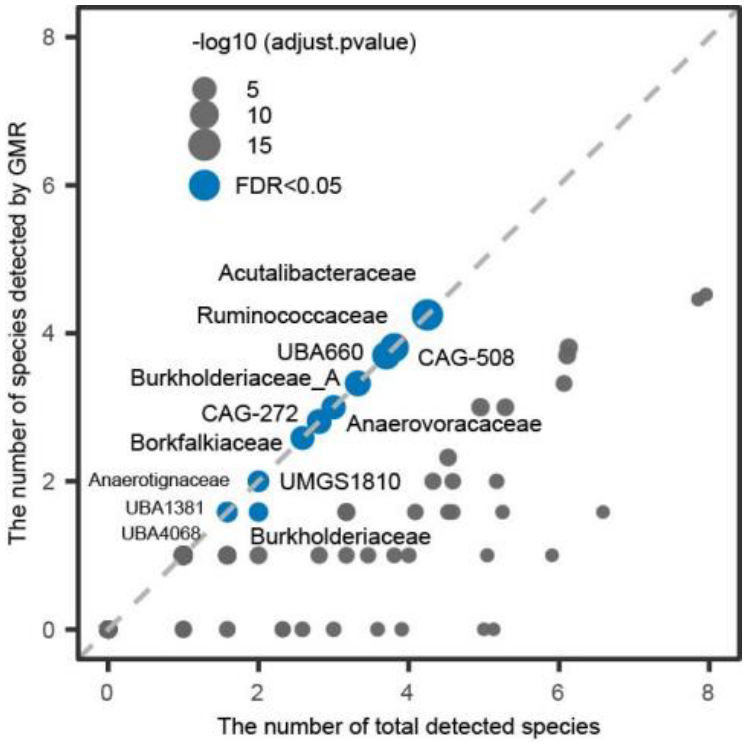
Supplementary Fig. 3 | Distribution of microbial genomes across species-level clusters. The bar plot shows the number of genomes unique to Chinese and non-Chinese populations in each species cluster.



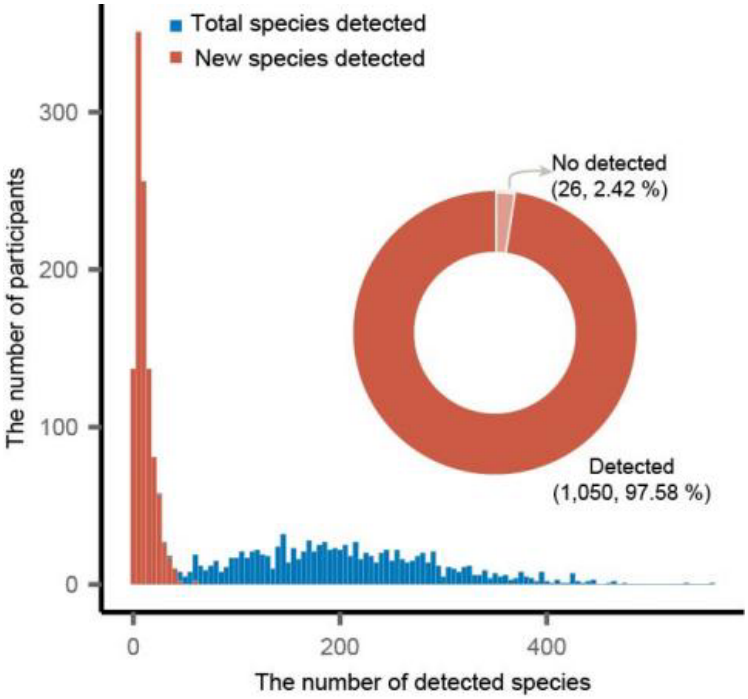
Supplementary Fig. 4 | Enhancement of the reads mapping rate for an independent Chinese-derived dataset using the new database. After integrating the GMR into the MetaPhlAn4 database for species identification, an improvement in the reads mapping rate was observed ($P_{\text{Paired wilcoxon}} < 2.2\times10^{-16}$).



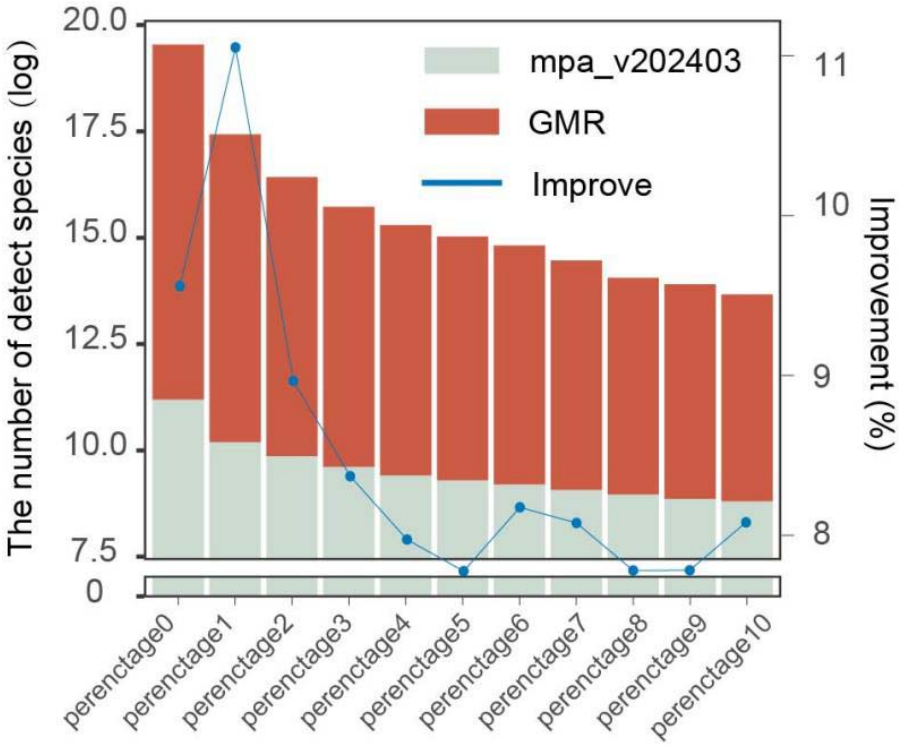
Supplementary Fig. 5 | Summary of microbial species showing significant improvement at the family level using the new database. A total of 326 novel species were detected after integrating the GMR into the MetaPhlAn4 database. The X-axis represents the total number of species detected within each family, while the Y-axis shows the number of newly identified species. Each dot represents a family, with the size of the dot indicating the FDR value from the hypergeometric test. Blue dots represent families with an FDR < 0.05. Among the 92 families containing newly detected species, 13 showed a significant improvement at an FDR threshold below 0.05, and these families are labeled accordingly.



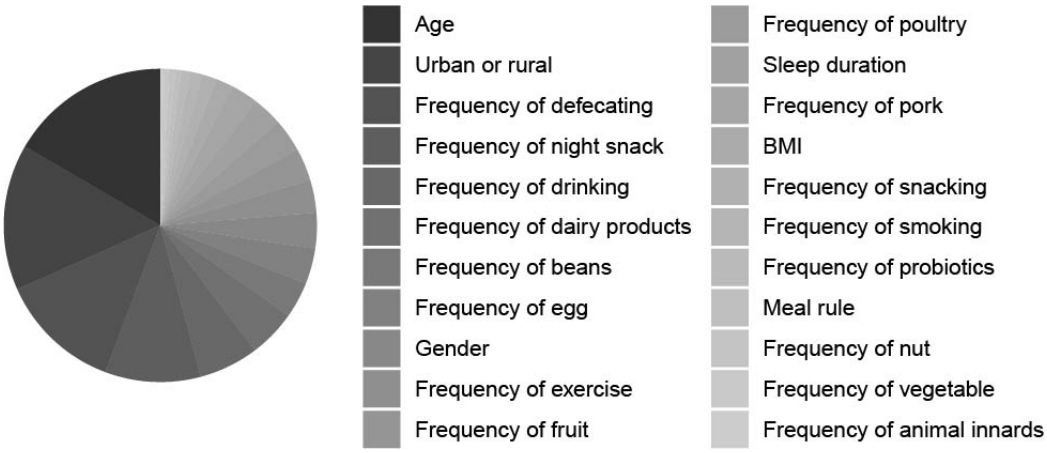
Supplementary Fig. 6 | Summary of microbial species profiled in a cohort of 1,076 participants. The blue bars represent the total number of detected species, while the red bars indicate the number of newly classified species in the GMR. On average, 205 species were detected per sample, with 11 belonging to newly classified species in the GMR. The pie chart on the right illustrates the proportion of samples with at least one newly detected species, showing that 97.58% of the samples contained at least one new species.



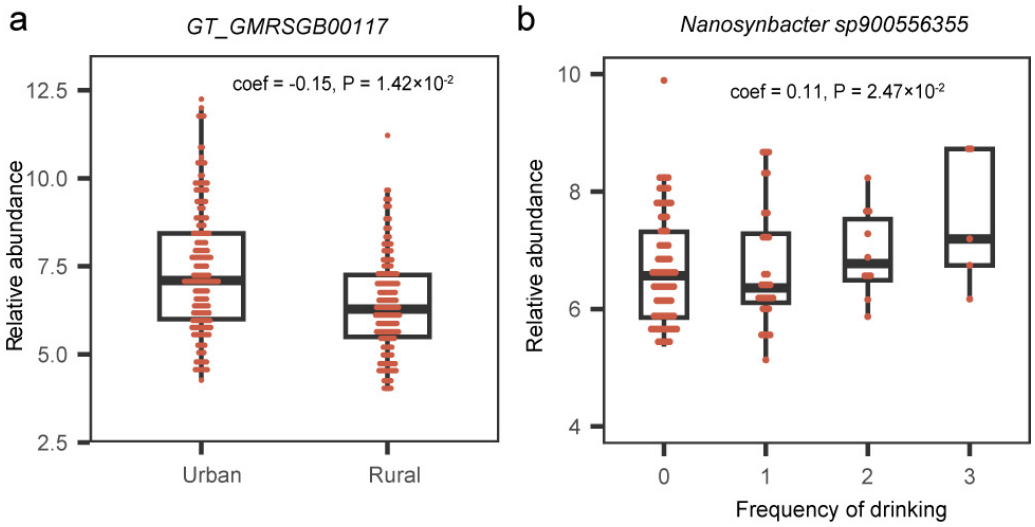
Supplementary Fig. 7 | Taxonomic profiling at the species level using the newly constructed database in the CGMR cohort with diverse filtering cutoffs. A total of 3,234 fecal metagenomic samples were analyzed. The X-axis shows different filtering cutoffs, from 0% to 10%. The left Y-axis (bar chart) represents the number of detected species, with the red portion highlighting the contribution of newly identified species in the GMR. The right Y-axis (blue line) represents the improvement in species detection achieved by including the newly identified species from the GMR.



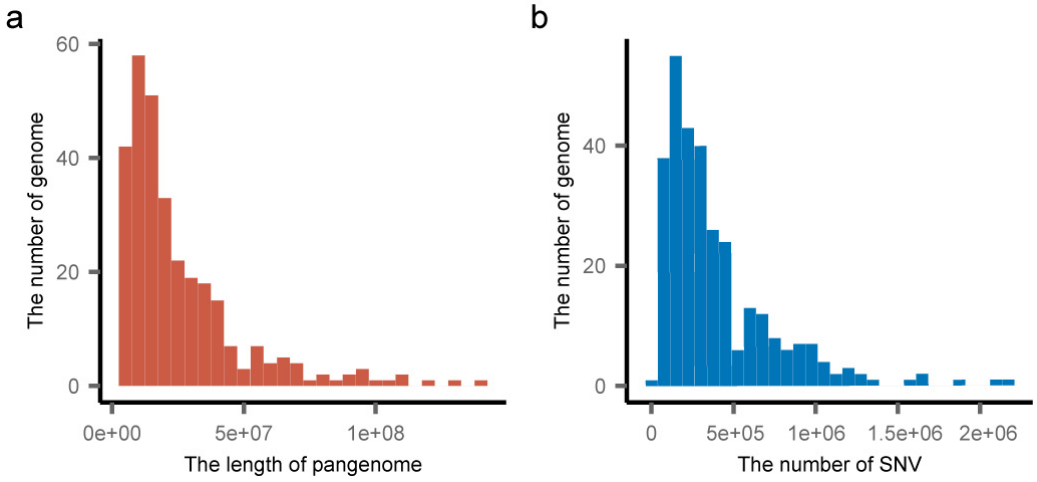
Supplementary Fig. 8 | Summary of host phenotypic associations with the abundance of newly profiled species. Pie chart illustrates the percentage of phenotypes that each bacterial species is significantly associated with. Different colors represent different phenotypes.



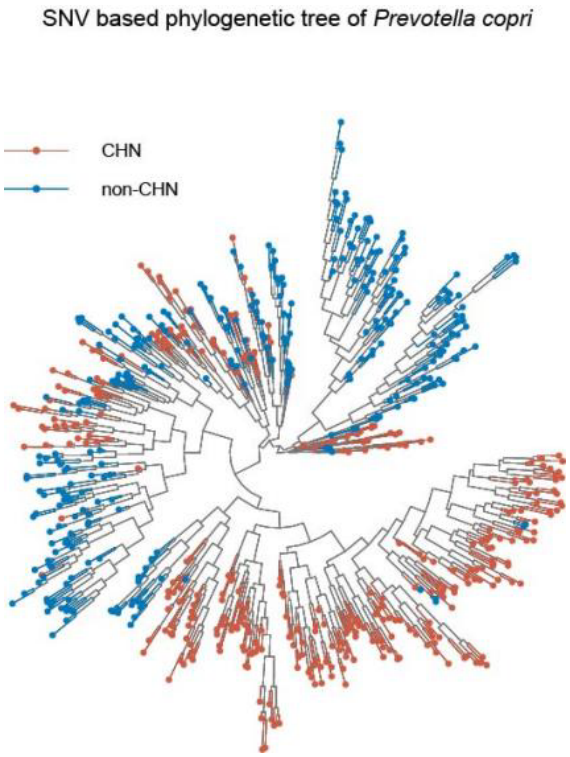
Supplementary Fig. 9 | Associations between species abundance and phenotypes in the CGMR cohort. a. The relative abundance of *GT_GMRSGB00117* is observed to be lower in rural areas**. b.** The relative abundance of *Nanosynbacter sp900556355* decreases with increasing beef consumption frequency. The box plot compares species abundance across different defecation frequencies. The boxes represent the median, first, and third quartiles (25th and 75th percentiles), with whiskers extending to the most extreme data points within 1.5 times the interquartile range. Each dot represents a participant.



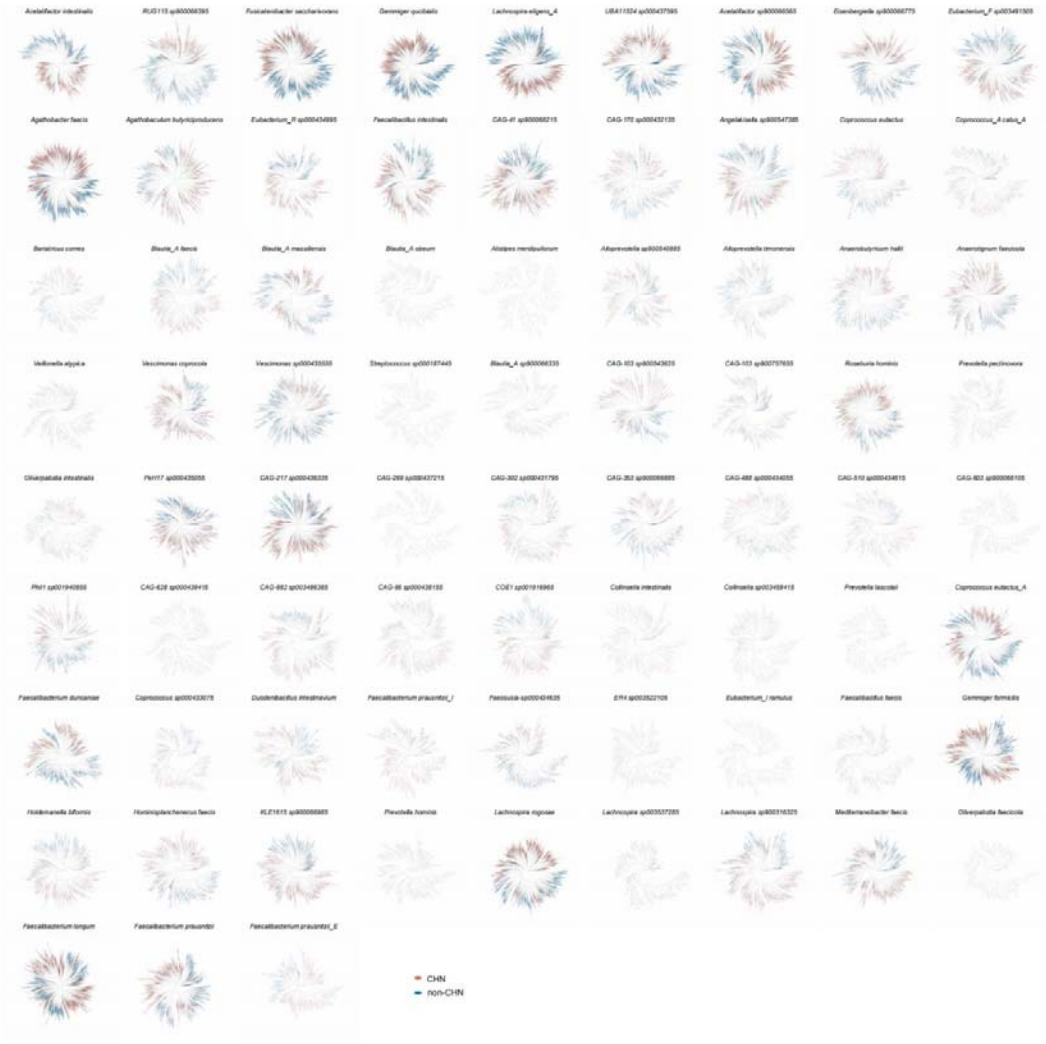
Supplementary Fig. 10 | Summary of genetic variation in 304 species common to Chinese and non-Chinese. a-b. The red bars represent the length of pangenome, while the blue bars indicate the number of detected SNVs. On average, the pangenome size of these species was 25.96M bp, with an average of 399,764 SNVs detected.



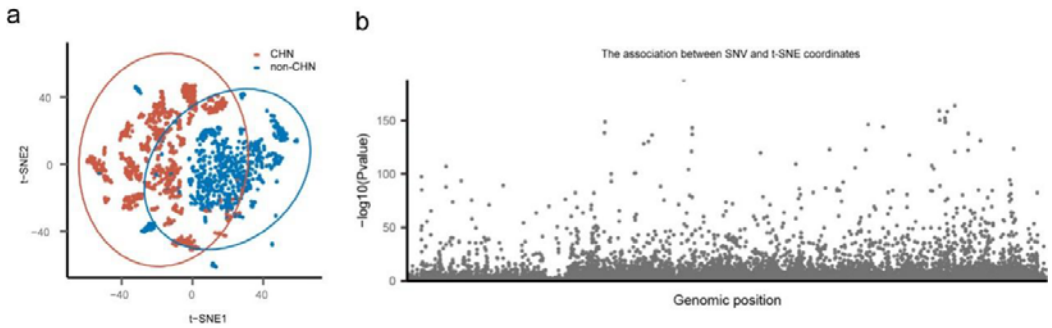
Supplementary Fig. 11 | The SNV based phylogenetic tree of *Prevotella copri* in GMR. *Prevotella copri* MAGs from Chinese populations showed clear clustering pattern based on SNV level dissimilarity matrix. Each leaf represents a MAG and is colored according to population.



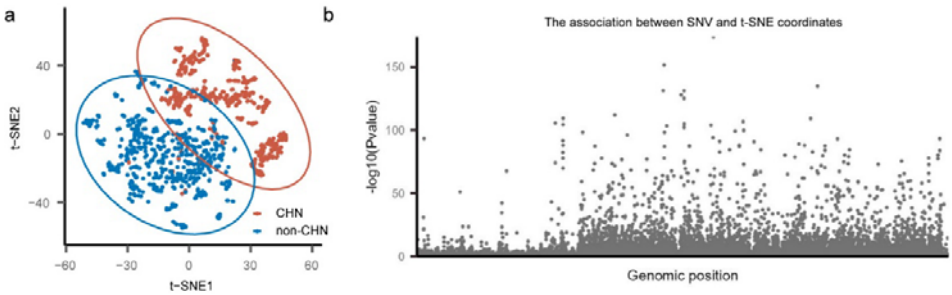
Supplementary Fig. 12 | The difference of genetic makeup of species between Chinese and non-Chinese individuals. 75 out of the 304 species (24.7%) exhibited significant population stratification.



Supplementary Fig. 13 | Genetic distance between *Acetatifactor intestinalis* genomes from Chinese and non-Chinese individuals. a. *Acetatifactor intestinalis* MAGs from Chinese and non-Chinese populations showed genetic stratification in the two t-SNE components based on SNV level dissimilarity matrix. Each dot represents a participant. **b.** Loci significantly associated with genetic distance. The X-axis represents genomic position of SNVs, and the Y-axis represents the -log10(pvalue) of the correlation between SNVs and tSNE.



Supplementary Fig. 14 | Genetic distance between *RUG115 sp900066395* genomes from Chinese and non-Chinese individuals. a. *RUG115 sp900066395* MAGs from Chinese and non-Chinese populations showed genetic stratification in the two t-SNE components based on SNV level dissimilarity matrix. Each dot represents a participant. **b.** Loci significantly associated with genetic distance. The X-axis represents genomic position of SNVs, and the Y-axis represents the -log10(pvalue) of the correlation between SNVs and tSNE.



Supplementary Fig. 15 | Three SNV sites in the *clpC* gene were significantly associated with the two t-SNE coordinates of *RUG115 sp900066395*. a-d. Correlation between tSNE and the three variant nucleotide sites on the *clpC* gene. Red represents the Chinese population and blue represents the non-Chinese population. The boxplots show the median, first quartile (25th percentile), and third quartile (75th percentile), with whiskers extending to the most extreme data points within 1.5 times the interquartile range. Each dot represents a MAG derived from an individual.

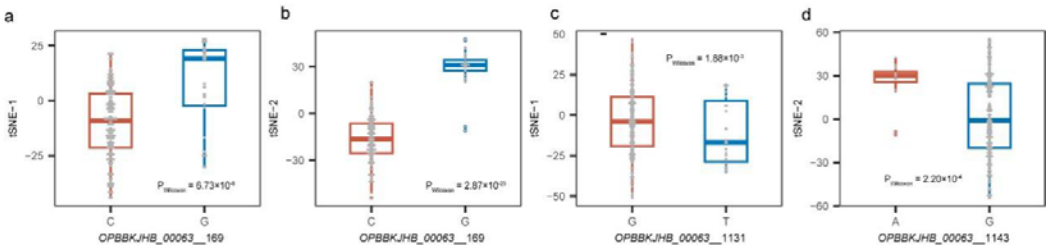


Table Legends

Table S1. The summary of assembly statistics of the 478,588 human gut microbial genomes used to generate the Global Microbiome Reference (GMR) collection.

Table S2. The metadata of 6,664 representative species of GMR and their novelty status.

Table S3. The summary of CDS identified per genome in the GMR and the proportion of CDS with unknown function.

Table S4. Number of genomes per species in the GMR database.

Table S5. Number of the core CDS and marker genes identified for species with genome greater than or equal to 3.

Table S6. Species information ultimately incorporated into the MetaPhlAn4 database.

Table S7. Families with the highest species enrichment detected after GMR Integration.

Table S8. The summary of species abundance profiling of the CGMR database using customized database.

Table S9. Sum of abundances of unique species detected in each sample by CGMR.

Table S10. The association between the abundance of novel species detected using our personalized database and phenotypes of CGMR (FDR < 0.05).

Table S11. The correlations of species abundance and phenotype found were validated in another independent cohort.

Table S12. 2145 species shared between the Chinese and non-Chinese populations

Table S13. Association between single nucleotide polymorphism (SNP) loci and distance matrix on *Acetatifactor intestinalis*.

Table S14. Association between genetic distance of species and phenotypes.

Table S15. Comparative analyses revealed 66 significant correlations between conspecific subspecies and host phenotypes.

Table S16. Association between SNV sites on the *Faecalibacterium prausnitzii_D* and age of host.

Table S17. Association between genetic distance of species and their relative abundance.

Table S18. For 60 species with independent genetic distances and abundances, the correlations between their genetic distances, relative abundances, and phenotypes were analyzed.

Table S19. For 172 species with independent genetic distances and abundances, the correlations between their genetic distances, relative abundances, and phenotypes were analyzed.

Table S20. Metabolites of three strains of *Bacteroides fragilis_A* culture medium.

Table S21. SNVs detected in three *Bacteroides fragilis_A* isolates.

Table S22. Association between SNVs in *Bacteroides fragilis_A* and their metabolites.

References

1. Fan, Y. & Pedersen, O. Gut microbiota in human metabolic health and disease. *Nat Rev Microbiol* **19**, 55-71 (2021).
2. Wang, Y. et al. Decoding microbial genomes to understand their functional roles in human complex diseases. *Imeta* **1**, e14 (2022).
3. Dong, Q. et al. Temporal colonization and metabolic regulation of the gut microbiome in neonatal oxen at single nucleotide resolution. *ISME J* **18** (2024).
4. Zhernakova, D.V. et al. Host genetic regulation of human gut microbial structural variation. *Nature* **625**, 813-821 (2024).
5. Chen, L. et al. The long-term genetic stability and individual specificity of the human gut microbiome. *Cell* **184**, 2302-2315 e2312 (2021).
6. Zahavi, L. et al. Bacterial SNPs in the human gut microbiome associate with host BMI. *Nat Med* **29**, 2785-2792 (2023).
7. Gunjur, A. et al. A gut microbial signature for combination immune checkpoint blockade across cancer types. *Nat Med* **30**, 797-809 (2024).
8. Suzuki, T.A. et al. Codiversification of gut microbiota with humans. *Science* **377**, 1328-1332 (2022).
9. Almeida, A. et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol* **39**, 105-114 (2021).
10. Huang, P. et al. Gut microbial genomes with paired isolates from China illustrate probiotic and cardiometabolic effects. *Cell Genom*, 100559 (2024).
11. Jin, H. et al. A high-quality genome compendium of the human gut microbiome of Inner Mongolians. *Nat Microbiol* **8**, 150-161 (2023).
12. Blanco-Miguez, A. et al. Extension of the Segatella copri complex to 13 species with distinct large extrachromosomal elements and associations with host conditions. *Cell Host Microbe* **31**, 1804-1819 e1809 (2023).
13. Karcher, N. et al. Genomic diversity and ecology of human-associated Akkermansia species in the gut microbiome revealed by extensive metagenomic assembly. *Genome Biol* **22**, 209 (2021).
14. Liu, C. et al. Enlightening the taxonomy darkness of human gut microbiomes with a cultured biobank. *Microbiome* **9**, 119 (2021).
15. Lin, X. et al. The genomic landscape of reference genomes of cultivated human gut bacteria. *Nat Commun* **14**, 1663 (2023).
16. Blanco-Miguez, A. et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat Biotechnol* **41**, 1633-1644 (2023).
17. Zhang, P. et al. Metagenome-wide analysis uncovers gut microbial signatures and implicates taxon-specific functions in end-stage renal disease. *Genome Biol* **24**, 226 (2023).
18. Li, C. et al. Deep insights into the gut microbial community of extreme longevity in south Chinese centenarians by ultra-deep metagenomics and large-scale culturomics. *NPJ Biofilms Microbiomes* **8**, 28 (2022).
19. Zheng, W.Q.M.K.Z.S.Z.M.Y.P.X.Z.Q.W.C.Y.J.S.J.X.W.H.K.Y. Air pollution, gut microbiota, and cognitive function alteration in older adults. *ISEE Conference Abstracts* **2024** (2024).
20. 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).

21. Parks, D.H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* **36**, 996-1004 (2018).

22. 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).

23. 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).

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

25. Qin, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59-65 (2010).

26. Lei; Chao Zeng, J.W.Z.Y.J.L.Y.Z.W.Z.M.D.T.Y.Y.Y.H.L.Y.W.Z.W.C.L.G. Association between gut microbiome-related metabolites and symptomatic hand osteoarthritis in two independent cohorts. *EBioMedicine* **98** (2023).

27. Mallick, H. et al. Multivariable association discovery in population-scale meta-omics studies. *PLoS Comput Biol* **17**, e1009442 (2021).

28. Tett, A. et al. The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations. *Cell Host Microbe* **26**, 666-679 e667 (2019).

29. Nagakubo, S., Nishino, K., Hirata, T. & Yamaguchi, A. The Putative Response Regulator BaeR Stimulates Multidrug Resistance of Escherichia coli via a Novel Multidrug Exporter System, MdtABC. *Journal of Bacteriology* **184**, 4161-4167 (2002).

30. Nair, S., Milohanic, E. & Berche, P. ClpC ATPase is required for cell adhesion and invasion of Listeria monocytogenes. *Infect Immun* **68**, 7061-7068 (2000).

31. Osman, D. et al. The requirement for cobalt in vitamin B(12): A paradigm for protein metalation. *Biochim Biophys Acta Mol Cell Res* **1868**, 118896 (2021).

32. Smith, A.D. & Refsum, H. Vitamin B-12 and cognition in the elderly. *Am J Clin Nutr* **89**, 707S-711S (2009).

33. Stabler, S.P., Lindenbaum, J. & Allen, R.H. Vitamin B-12 deficiency in the elderly: current dilemmas. *Am J Clin Nutr* **66**, 741-749 (1997).

34. Karim, S.M. & Hillier, K. Prostaglandins: pharmacology and clinical application. *Drugs* **8**, 176-207 (1974).

35. Zhu, L. et al. The cyclooxygenase-1/mPGES-1/endothelial prostaglandin EP4 receptor pathway constrains myocardial ischemia-reperfusion injury. *Nat Commun* **10**, 1888 (2019).

36. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* (2024).

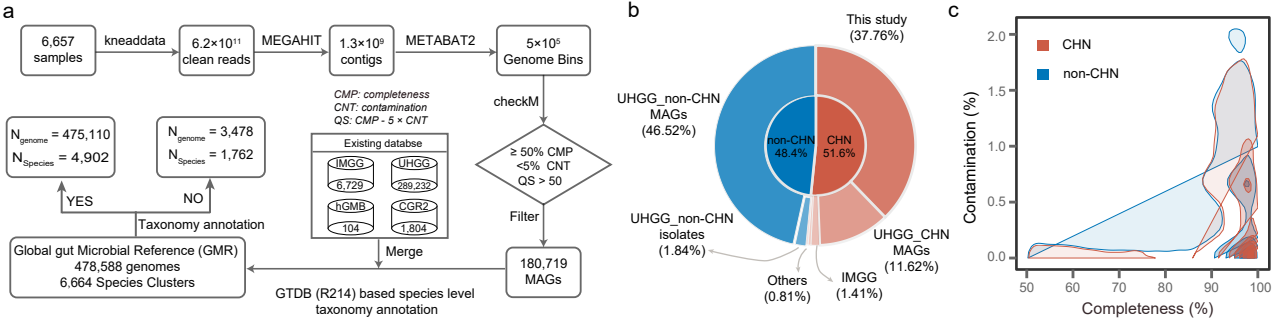
37. Nakanishi, M. & Rosenberg, D.W. Multifaceted roles of PGE2 in inflammation and cancer. *Semin Immunopathol* **35**, 123-137 (2013).

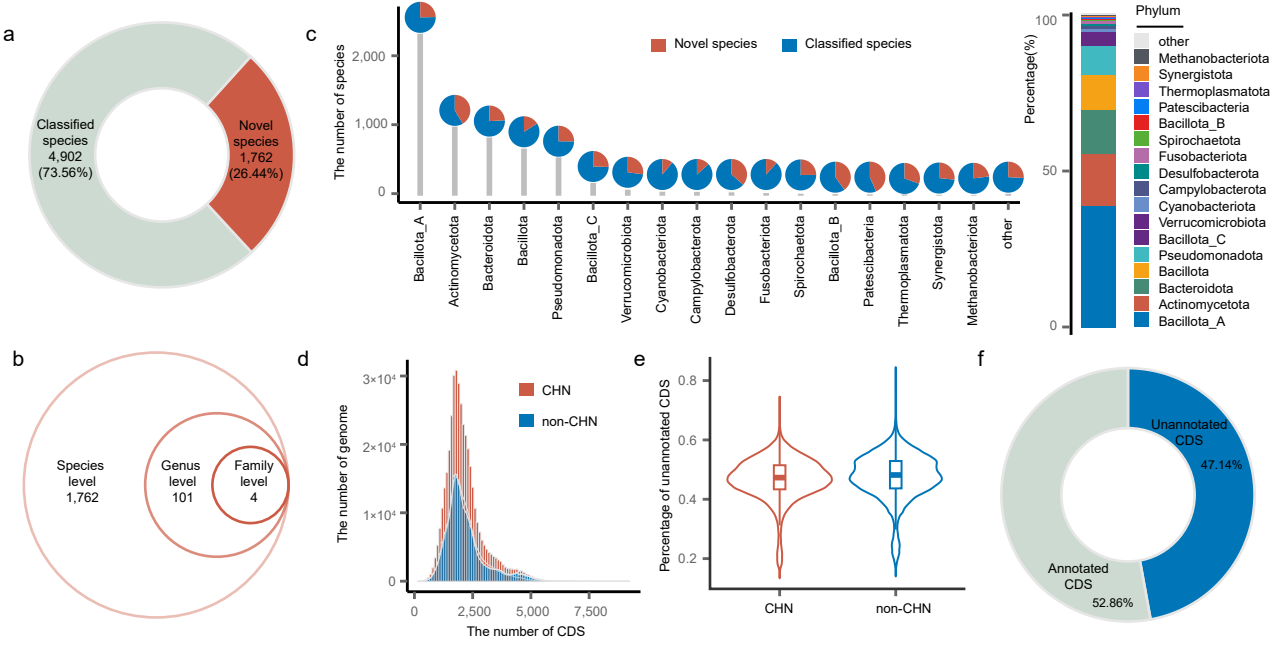
38. Lun, W. et al. Mechanism of action of the bile acid receptor TGR5 in obesity. *Acta Pharm Sin B* **14**, 468-491 (2024).

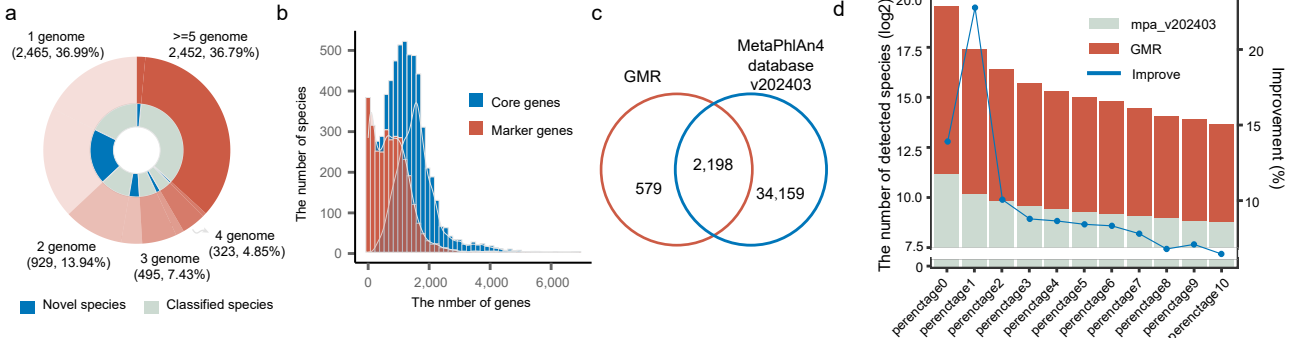
39. Chavez-Talavera, O., Haas, J., Grzych, G., Tailleux, A. & Staels, B. Bile acid alterations in nonalcoholic fatty liver disease, obesity, insulin resistance and type 2 diabetes: what do the human studies tell? *Curr Opin Lipidol* **30**, 244-254 (2019).

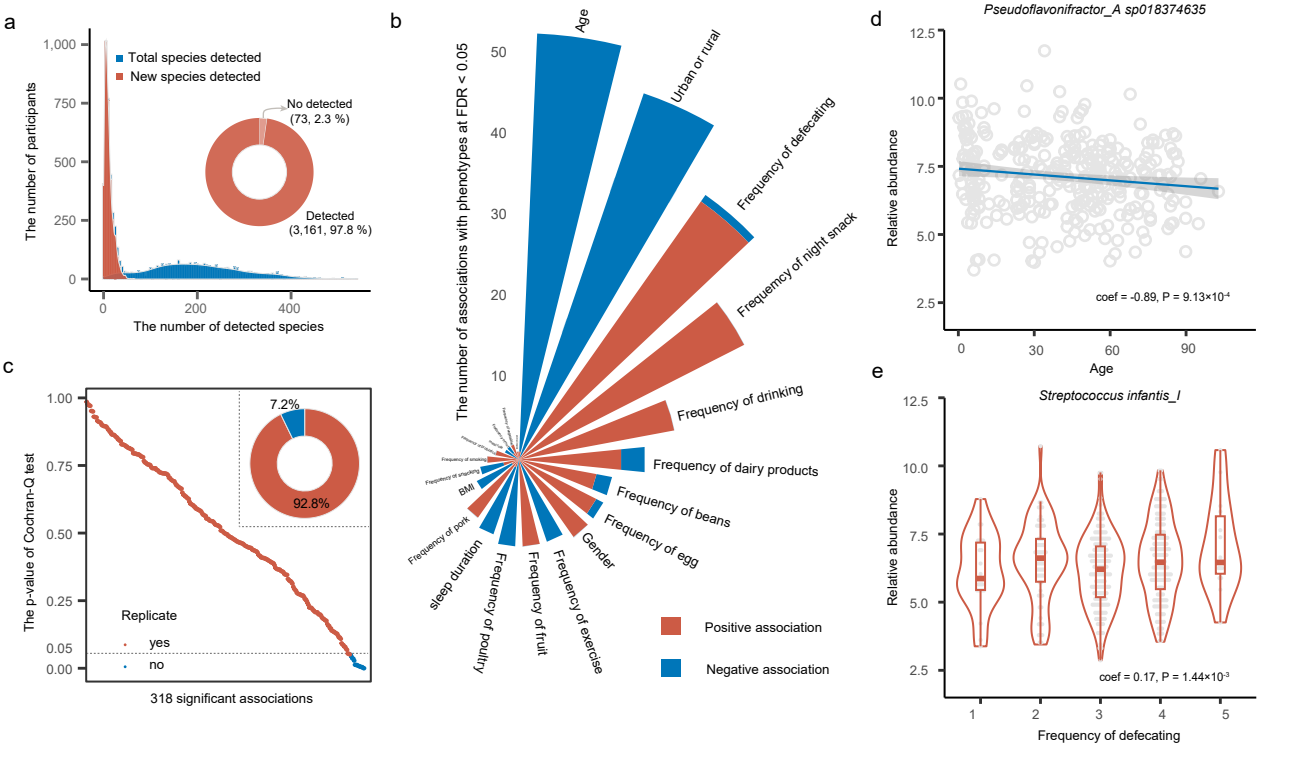
40. Chen, L. et al. Genetic and Microbial Associations to Plasma and Fecal Bile Acids in Obesity Relate to Plasma Lipids and Liver Fat Content. *Cell Rep* **33**, 108212 (2020).
41. Collins, S.L., Stine, J.G., Bisanz, J.E., Okafor, C.D. & Patterson, A.D. Bile acids and the gut microbiota: metabolic interactions and impacts on disease. *Nat Rev Microbiol* **21**, 236-247 (2023).
42. Wei, J. et al. Association Between Gut Microbiota and Elevated Serum Urate in Two Independent Cohorts. *Arthritis Rheumatol* **74**, 682-691 (2022).
43. Viana, D. et al. A single natural nucleotide mutation alters bacterial pathogen host tropism. *Nat Genet* **47**, 361-366 (2015).
44. Olm, M.R. et al. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat Biotechnol* **39**, 727-736 (2021).
45. Schloissnig, S. et al. Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45-50 (2013).
46. Yoshida, H., Bogaki, M., Nakamura, M. & Nakamura, S. Quinolone resistance-determining region in the DNA gyrase gyrA gene of Escherichia coli. *Antimicrob Agents Chemother* **34**, 1271-1272 (1990).
47. Wang, J. & Jia, H. Metagenome-wide association studies: fine-mining the microbiome. *Nat Rev Microbiol* **14**, 508-522 (2016).
48. Maini Rekdal, V., Bess, E.N., Bisanz, J.E., Turnbaugh, P.J. & Balskus, E.P. Discovery and inhibition of an interspecies gut bacterial pathway for Levodopa metabolism. *Science* **364** (2019).
49. Castelle, C.J. & Banfield, J.F. Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. *Cell* **172**, 1181-1197 (2018).
50. Nayfach, S., Shi, Z.J., Seshadri, R., Pollard, K.S. & Kyrpides, N.C. New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505-510 (2019).
51. Pasolli, E. et al. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649-662 e620 (2019).
52. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
53. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-580 (1999).
54. Langmead, B., Wilks, C., Antonescu, V. & Charles, R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics* **35**, 421-432 (2019).
55. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).
56. 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).
57. Kang, D.D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
58. 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).

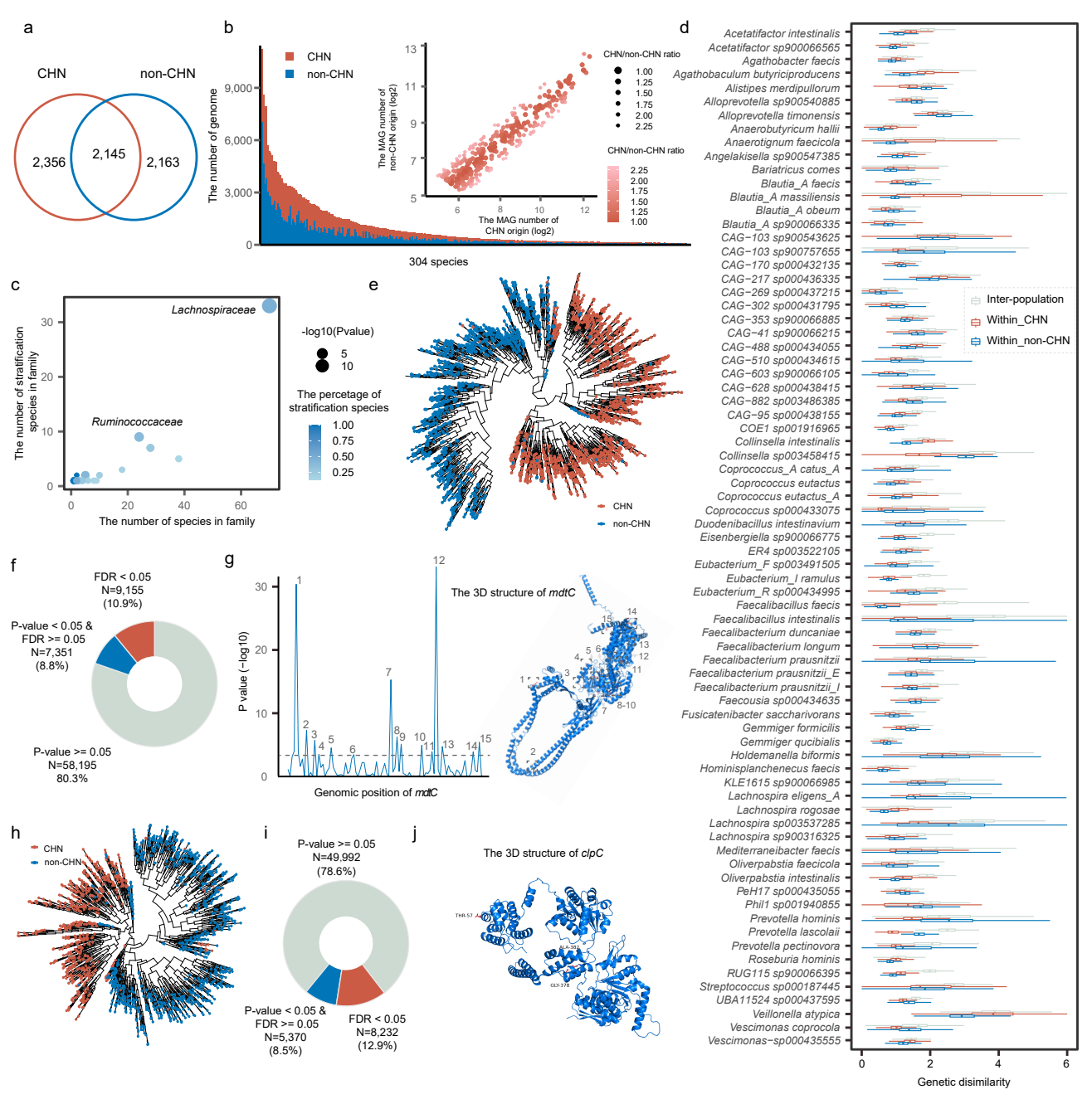
59. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
60. Page, A.J. et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691-3693 (2015).
61. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276-277 (2000).
62. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* **47**, W256-W259 (2019).
63. Xu, S. et al. Ggtree: A serialized data object for visualization of a phylogenetic tree and annotation data. *Imeta* **1**, e56 (2022).
64. Gligorijevic, V. et al. Structure-based protein function prediction using graph convolutional networks. *Nat Commun* **12**, 3168 (2021).

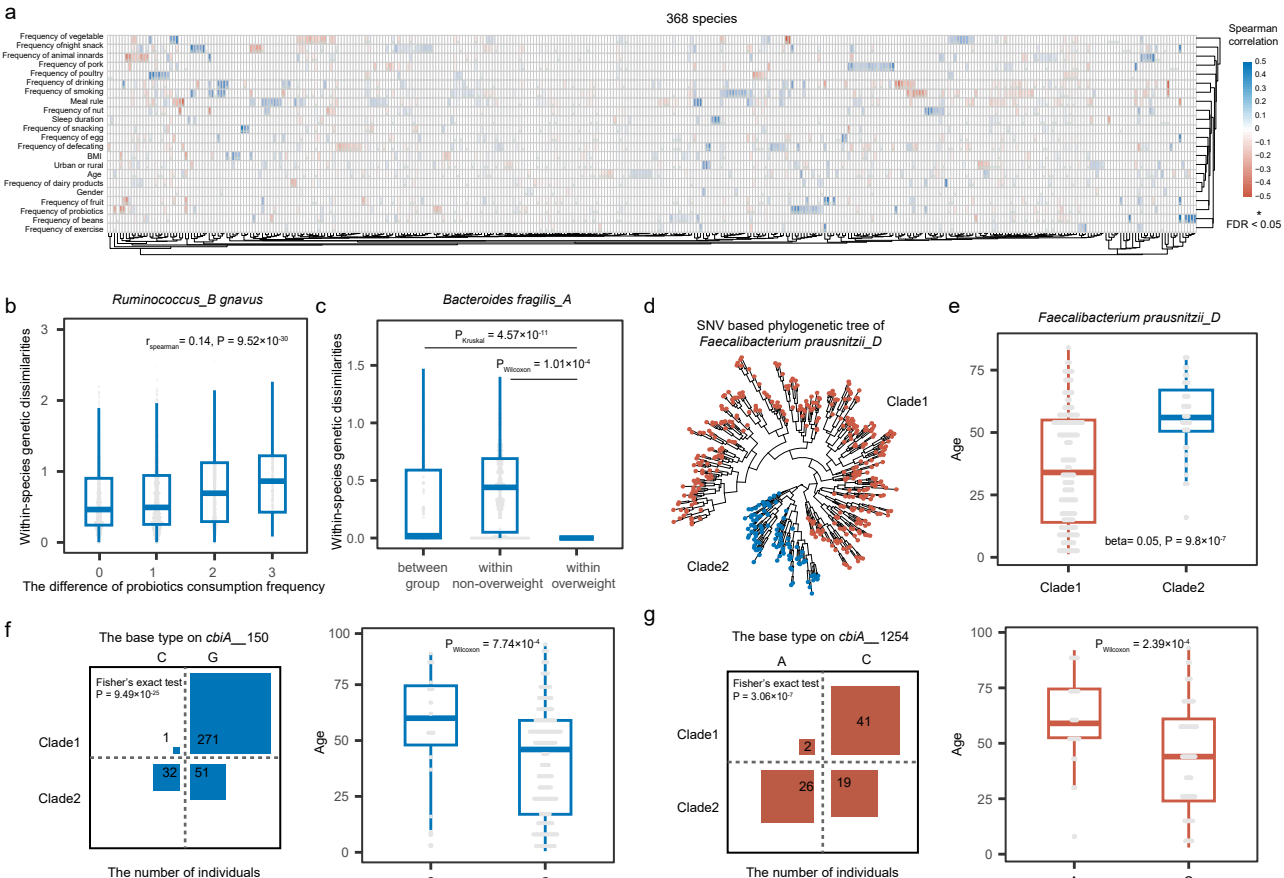

a
6,657 samples
kneaddata
6.2×10^11 clean reads
MEGAHIT
1.3×10^9 contigs
METABAT2
5×10^5 Genome Bins
checkM
CMP: completeness
CNT: contamination
QS: CMP - 5 × CNT
≥ 50% CMP
<5% CNT
QS > 50
Filter
180,719 MAGs
GTDB (R214) based species level taxonomy annotation
Existing databse
IMGG
6,729
UHGG
289,232
hGMB
104
CGR2
1,804
Merge
Global gut Microbial Reference (GMR)
478,588 genomes
6,664 Species Clusters
Taxonomy annotation
YES
NO
N genome = 475,110
N Species = 4,902
N genome = 3,478
N Species = 1,762
b
This study (37.76%)
UHGG_non-CHN MAGs (46.52%)
non-CHN 48.4%
CHN 51.6%
UHGG_non-CHN isolates (1.84%)
Others (0.81%)
IMGG (1.41%)
UHGG_CHN MAGs (11.62%)
c
CHN
non-CHN
Contamination (%)
Completeness (%)
0.0
0.5
1.0
1.5
2.0
50
60
70
80
90
100

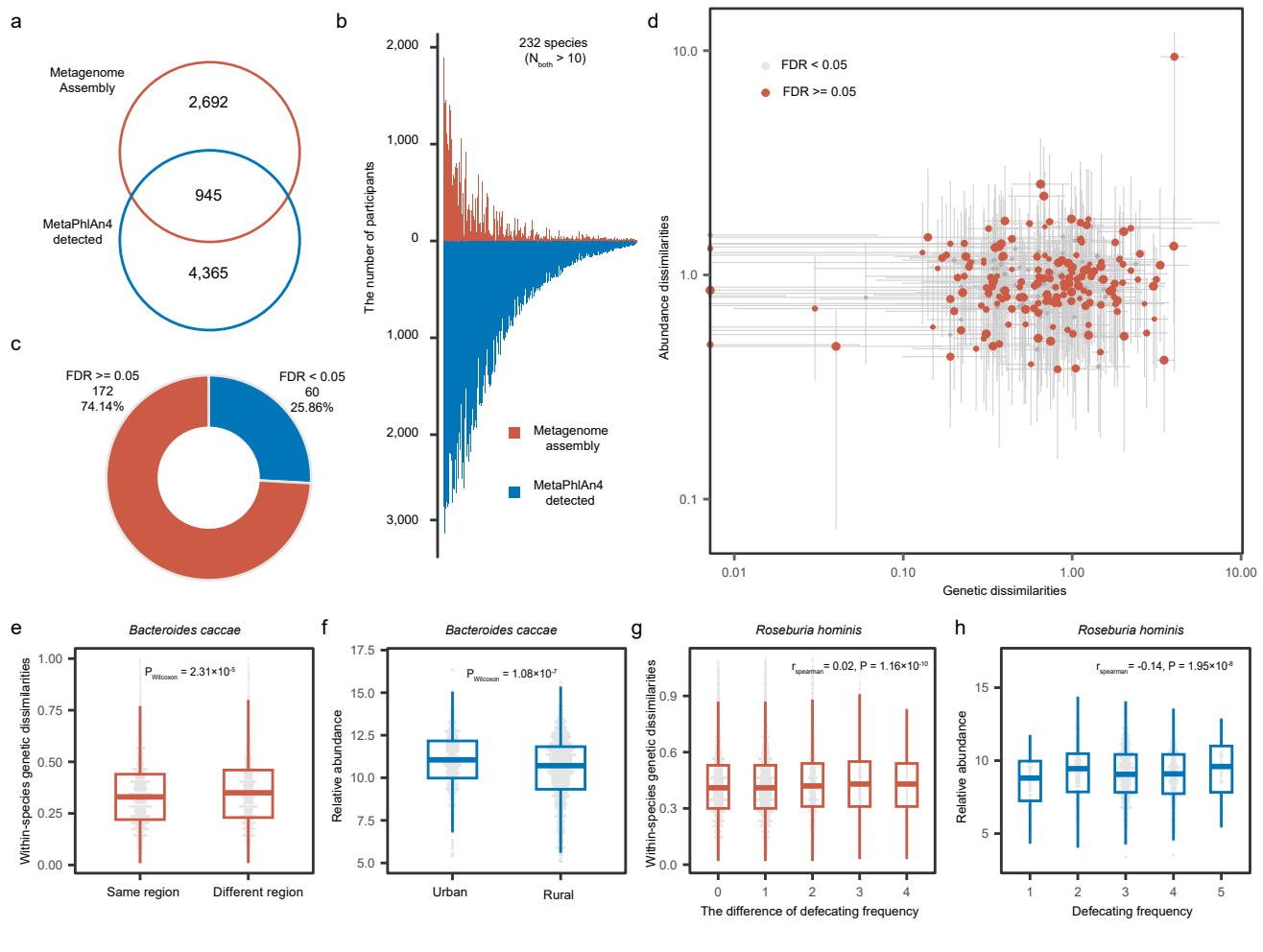

a
Classified species 4,902 (73.56%)
Novel species 1,762 (26.44%)
b
Species level 1,762
Genus level 101
Family level 4
c
Novel species
Classified species
The number of species
2,000
1,000
0
Bacillota_A
Actinomycetota
Bacteroidota
Bacillota
Pseudomonadota
Bacillota_C
Verrucomicrobiota
Cyanobacteriota
Campylobacterota
Desulfobacterota
Fusobacteriota
Spirochaetota
Bacillota_B
Patescibacteria
Thermoplasmatota
Synergistota
Methanobacteriota
other
Percentage(%)
100
50
0
Phylum
other
Methanobacteriota
Synergistota
Thermoplasmatota
Patescibacteria
Bacillota_B
Spirochaetota
Fusobacteriota
Desulfobacterota
Campylobacterota
Cyanobacteriota
Verrucomicrobiota
Bacillota_C
Pseudomonadota
Bacillota
Bacteroidota
Actinomycetota
Bacillota_A
d
CHN
non-CHN
The number of genome
3×10⁴
2×10⁴
1×10⁴
0
0
2,500
5,000
7,500
The number of CDS
e
Percentage of unannotated CDS
0.8
0.6
0.4
0.2
CHN
non-CHN
f
Unannotated CDS
47.14%
Annotated CDS
52.86%

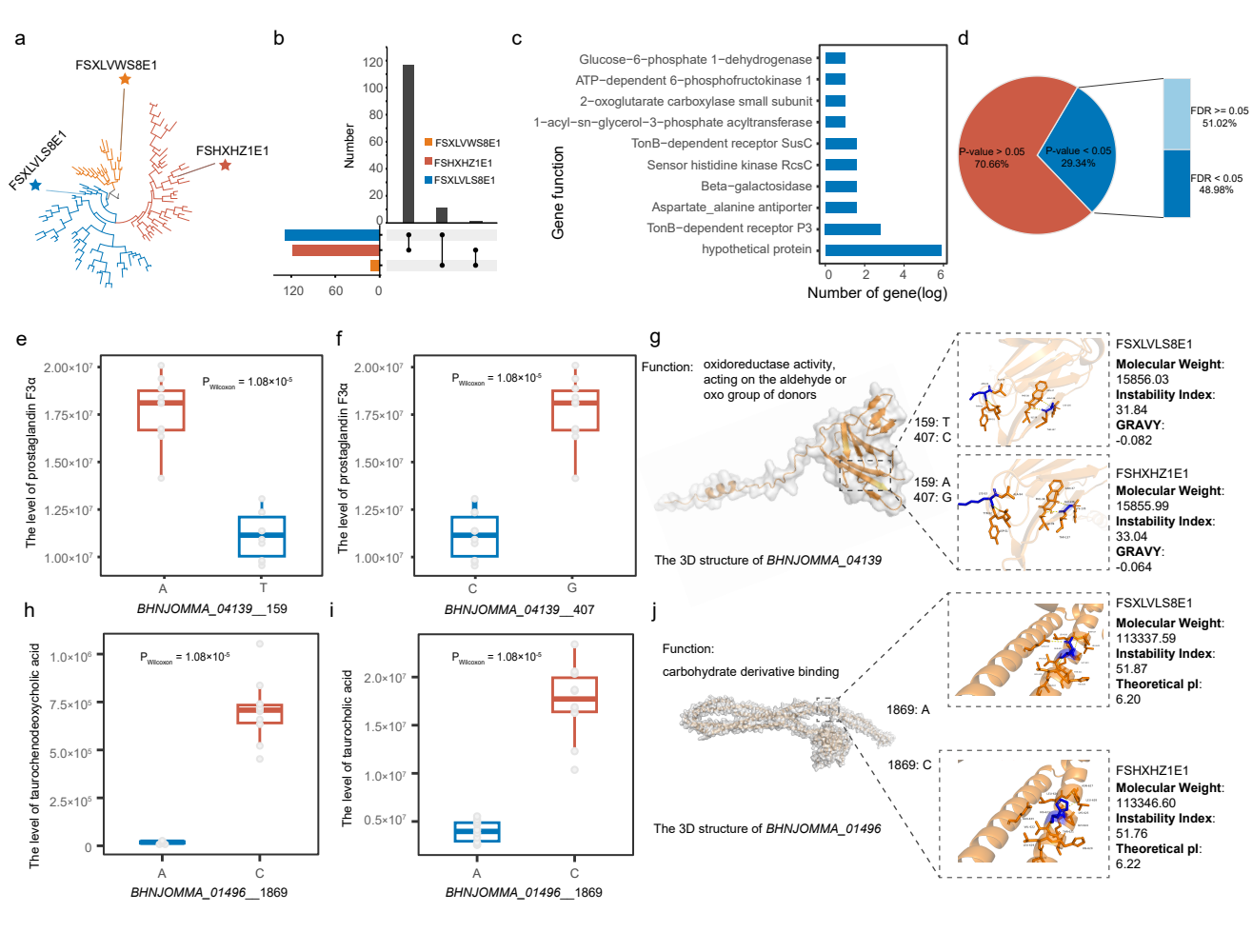

a
1 genome
(2,465, 36.99%)
>=5 genome
2,452, 36.79%)
2 genome
(929, 13.94%)
3 genome
(495, 7.43%)
4 genome
(323, 4.85%)
Novel species
Classified species
b
The number of species
The nmber of genes
Core genes
Marker genes
c
GMR
MetaPhlAn4
database
v202403
579
2,198
34,159
d
The number of detected species (log2)
Improvement (%)
mpa_v202403
GMR
Improve
perentage0
perentage1
perentage2
perentage3
perentage4
perentage5
perentage6
perentage7
perentage8
perentage9
perentage10


a
Total species detected
New species detected
No detected
(73, 2.3 %)
Detected
(3,161, 97.8 %)
The number of participants
The number of detected species
b
The number of associations with phenotypes at FDR < 0.05
Age
Urban or rural
Frequency of defecating
Frequency of night snack
Frequency of drinking
Frequency of dairy products
Frequency of beans
Frequency of egg
Gender
Frequency of exercise
Frequency of fruit
Frequency of poultry
sleep duration
Frequency of pork
BMI
Frequency of snacking
Frequency of smoking
Positive association
Negative association
c
The p-value of Cochran-Q test
7.2%
92.8%
Replicate
yes
no
318 significant associations
d
Pseudoflavonifractor_A sp018374635
Relative abundance
coef = -0.89, P = 9.13×10-4
Age
e
Streptococcus infantis_I
Relative abundance
coef = 0.17, P = 1.44×10-3
Frequency of defecating


a
CHN
non-CHN
2,356
2,145
2,163
b
CHN
non-CHN
The number of genome
9,000
6,000
3,000
0
304 species
The MAG number of non-CHN origin (log2)
The MAG number of CHN origin (log2)
CHN/non-CHN ratio
1.00
1.25
1.50
1.75
2.00
2.25
c
Lachnospiraceae
Ruminococcaceae
The number of stratification species in family
The number of species in family
-log10(Pvalue)
5
10
The percetage of stratification species
1.00
0.75
0.50
0.25
e
CHN
non-CHN
d
Inter-population
Within_CHN
Within_non-CHN
Genetic disimilarity
f
FDR < 0.05
N=9,155
(10.9%)
P-value < 0.05 &
FDR >= 0.05
N=7,351
(8.8%)
P-value >= 0.05
N=58,195
80.3%
g
P value (−log10)
Genomic position of mdtC
The 3D structure of mdtC
h
CHN
non-CHN
i
P-value >= 0.05
N=49.992
(78.6%)
P-value < 0.05 &
FDR >= 0.05
N=5,370
(8.5%)
FDR < 0.05
N=8,232
(12.9%)
j
The 3D structure of clpC


a
368 species
Frequency of vegetable
Frequency ofnight snack
Frequency of animal innards
Frequency of pork
Frequency of poultry
Frequency of drinking
Frequency of smoking
Meal rule
Frequency of nut
Sleep duration
Frequency of snacking
Frequency of egg
Frequency of defecating
BMI
Urban or rural
Age
Frequency of dairy products
Gender
Frequency of fruit
Frequency of probiotics
Frequency of beans
Frequency of exercise
Spearman correlation
0.5
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.5
* FDR < 0.05
b
Ruminococcus_B gnavus
r_spearman = 0.14, P = 9.52×10^-30
Within-species genetic dissimilarities
The difference of probiotics consumption frequency
c
Bacteroides fragilis_A
P_Kruskal = 4.57×10^-11
P_Wilcoxon = 1.01×10^-4
Within-species genetic dissimilarities
between group
within non-overweight
within overweight
d
SNV based phylogenetic tree of Faecalibacterium prausnitzii_D
Clade1
Clade2
e
Faecalibacterium prausnitzii_D
Age
beta= 0.05, P = 9.8×10^-7
Clade1
Clade2
f
The base type on cbiA__150
C
G
Fisher's exact test
P = 9.49×10^-25
Clade1
Clade2
1
271
32
51
The number of individuals
P_Wilcoxon = 7.74×10^-4
Age
C
G
g
The base type on cbiA__1254
A
C
Fisher's exact test
P = 3.06×10^-7
Clade1
Clade2
2
41
26
19
The number of individuals
P_Wilcoxon = 2.39×10^-4
Age
A
C


a
Metagenome Assembly
2,692
945
MetaPhlAn4 detected
4,365
b
232 species
(N_{both} > 10)
The number of participants
2,000
1,000
0
1,000
2,000
3,000
Metagenome assembly
MetaPhlAn4 detected
c
FDR >= 0.05
172
74.14%
FDR < 0.05
60
25.86%
d
FDR < 0.05
FDR >= 0.05
Abundance dissimilarities
10.0
1.0
0.1
0.01
0.10
1.00
10.00
Genetic dissimilarities
e
Bacteroides caccae
Within-species genetic dissimilarities
$P_{Wilcoxon}$ = 2.31×10^{-5}
1.00
0.75
0.50
0.25
0.00
Same region
Different region
f
Bacteroides caccae
Relative abundance
$P_{Wilcoxon}$ = 1.08×10^{-7}
17.5
15.0
12.5
10.0
7.5
5.0
Urban
Rural
g
Roseburia hominis
Within-species genetic dissimilarities
$r_{spearman}$ = 0.02, P = 1.16×10^{-10}
0.9
0.6
0.3
0.0
0
1
2
3
4
The difference of defecating frequency
h
Roseburia hominis
Relative abundance
$r_{spearman}$ = -0.14, P = 1.95×10^{-8}
15
10
5
1
2
3
4
5
Defecating frequency


a
FSXLVWS8E1
FSXLVLS8E1
FSHXHZ1E1
b
Number
FSXLVWS8E1
FSHXHZ1E1
FSXLVLS8E1
c
Glucose−6−phosphate 1−dehydrogenase
ATP−dependent 6−phosphofructokinase 1
2−oxoglutarate carboxylase small subunit
1−acyl−sn−glycerol−3−phosphate acyltransferase
TonB−dependent receptor SusC
Sensor histidine kinase RcsC
Beta−galactosidase
Aspartate_alanine antiporter
TonB−dependent receptor P3
hypothetical protein
Gene function
Number of gene(log)
d
P-value > 0.05
70.66%
P-value < 0.05
29.34%
FDR >= 0.05
51.02%
FDR < 0.05
48.98%
e
The level of prostaglandin F3α
$P_{Wilcoxon}$ = 1.08×10-5
A
T
BHNJOMMA_04139__159
f
The level of prostaglandin F3α
$P_{Wilcoxon}$ = 1.08×10-5
C
G
BHNJOMMA_04139__407
g
Function: oxidoreductase activity, acting on the aldehyde or oxo group of donors
159: T
407: C
159: A
407: G
The 3D structure of BHNJOMMA_04139
FSXLVLS8E1
Molecular Weight: 15856.03
Instability Index: 31.84
GRAVY: -0.082
FSHXHZ1E1
Molecular Weight: 15855.99
Instability Index: 33.04
GRAVY: -0.064
h
The level of taurochenodeoxycholic acid
$P_{Wilcoxon}$ = 1.08×10-5
A
C
BHNJOMMA_01496__1869
i
The level of taurocholic acid
$P_{Wilcoxon}$ = 1.08×10-5
A
C
BHNJOMMA_01496__1869
j
Function: carbohydrate derivative binding
1869: A
1869: C
The 3D structure of BHNJOMMA_01496
FSXLVLS8E1
Molecular Weight: 113337.59
Instability Index: 51.87
Theoretical pI: 6.20
FSHXHZ1E1
Molecular Weight: 113346.60
Instability Index: 51.76
Theoretical pI: 6.22