Enlightening the Taxonomy Darkness of Human Gut Microbiomes With Cultured Biobank

Chang Liu
Institute of Microbiology Chinese Academy of Sciences
Meng-Xuan Du
Institute of Microbiology Chinese Academy of Sciences
Rexiding Abuduaini
Institute of Microbiology Chinese Academy of Sciences
Hai-Ying Yu
Institute of Microbiology Chinese Academy of Sciences
Dan-Hua Li
Institute of Microbiology Chinese Academy of Sciences
Yu-Jing Wang
Institute of Microbiology Chinese Academy of Sciences
Nan Zhou
Institute of Microbiology Chinese Academy of Sciences
Min-Zhi Jiang
Institute of Microbiology Chinese Academy of Sciences
Peng-Xia Niu
Institute of Microbiology Chinese Academy of Sciences
Shan-Shan Han
Institute of Microbiology Chinese Academy of Sciences
Hong-He Chen
Institute of Microbiology Chinese Academy of Sciences
Wen-Yu Shi
Institute of Microbiology Chinese Academy of Sciences
Linhuan Wu
Institute of Microbiology Chinese Academy of Sciences
Yu-Hua Xin
Institute of Microbiology Chinese Academy of Sciences
Juncai Ma
Institute of Microbiology Chinese Academy of Sciences
Yuguang Zhou
Institute of Microbiology Chinese Academy of Sciences
Cheng-Ying Jiang

Institute of Microbiology Chinese Academy of Sciences
Hong-Wei Liu
Institute of Microbiology Chinese Academy of Sciences
Shuang-Jiang Liu (liusj@im.ac.cn)
Institute of Microbiology Chinese Academy of Sciences https://orcid.org/0000-0002-7585-310X

Research

Keywords: human gut microbiomes, cultivation, biobank, new taxa, hGMB

Posted Date: September 10th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-74101/v1

Abstract

Background The cultivated gut microbial resource plays essential role in gut microbiome studies such as gut microbial function and their interactions with host. Though several major studies had been performed to understand the cultured human gut microbiota, up to 70% of the Unified Human Gastrointestinal Genome species remain uncultivated and their taxonomy is not clear. Large-scale gut microbial isolation and identification and their access to pubic are imperative for gut microbial studies and for understanding of the human gut microbial functions.

Results Here, we report the construction of an human Gut Microbial Biobank (hGMB) (homepage: hgmb.nmdc.cn) by large-scale cultivation of 10,558 isolates from 239 feces of healthy Chinese volunteers, and deposited 1,170 strains representing 404 different species in International Depository Authority for long-term preservation and public access worldwidely. We discovered and denominated 107 new species, and proposed 28 new genera and 3 new families. The new species and their newly sequenced genomes uncovered 16 “most-wanted” or “medium priority” taxa proposed by the Human Microbiome Project and 42 previously-uncultured MAGs in IGGdb, respectively. The hGMB represented over 80% of the common and dominant human gut microbial genera or species of global human gut 16S rRNA gene amplicon data (n=11,647), and covered 70% of the known genes (KEGG Orthologys) and 10% of the functionally-unknown genes in the global human gut gene catalogs.

Conclusions A publically accessible human Gut Microbial Biobank (hGMB) that contains 1,170 strains and represents 404 human gut microbial speces is estabolished. The hGMB expands the currently known, taxonomically-characterized gut microbial resources and genomic repository by adding 107 new species and 115 new genomes of human gut microbes. Based on the newly discovered species in this study, 28 new genera and 3 new families of human gut microbes were identified and proposed.

Introduction

The human gut microbiome (GM) is defined as a new organ and correlates host’s physical and mental health [1]. When GM dysbiosis happened, it often triggered host’s immune dysfunction [2], metabolic disorder [3], and impaired developments of body and recognition [4]. Both culture-dependent and -independent studies have delivered unprecedentedly knowledge of GM diversities and functions [5–7]. Still, our understanding of human GMs is very limited. According to the most recent work of unified human gastrointestinal genome (UHGG) [8], more than 70% of gut microbial species have not been cultivated and 40% of the protein-coding sequences have no functional annotations (see also references [9, 10]). Those unknown microorganisms and their genetic elements are called “dark matters” of GMs and they hide secrets of GM functions and GM-host interactions [9, 11, 12].

In order to disclose the identity and function of those “dark matters”, great efforts have been made to improve bioinformatic tools and to expand database capacity [13–16]. However, functional characterization and verification at biological and molecular level are still needed to be done

experimentally. The cultured microbial resources that harbor unknown genes of interest and/or produce specific metabolites are indispensable. Further, previous works showed that cultivated gut microbial resources played fundamental roles not only in culture-dependent causative studies of host-GM interactions [17–19] but also in the cultivation-independent omics studies [20–22]. Clearly, enlightening the dark matters of GMs needs extensive efforts on microbial cultivation and physiological and genetic characterizations. During the past few years, several large-scale cultivations have been made [20, 21, 23–25], and there were totally over 1,500 microbial species cultivated from those works. According to our and previous analyses, the cultivated human gut microbes accounted for 30–50% of the detected human gut microbial species from metagenomic and 16S rRNA gene amplicon datasets [8, 9, 22, 26]. The taxonomy and nomination of gut microorganisms, on the other hand, even lag behind. For example, the Culturomics [25] reported 247 new taxa in 2016, there are 117 out of the 247 new taxa remained unclassified till the time of this writing. This resulted in that some microbes were repeatedly claimed to be novel in different studies. Examples are that 54 microbial taxa firstly cultivated in 2016 [23] were still considered as new taxa in two works of 2019 [20, 21]. Taxonomic description and valid nomination for cultivated taxa promote researches and facilitate the accession and exchange of bacterial materials among scientific communities worldwide [27, 28].

In this study, we cultivated 10,558 bacterial isolates that represented 404 gut microbial species from 239 fecal samples of health donors by large-scale cultivation, and deposited 1,170 representative strains to International Depositttory Authority (IDA) for public access globally. We sequenced 115 new bacterial genomes and denominated 107 new bacterial taxa. Data analysis revealed that the hGMB largely represented the taxonomic composition of human gut microbial community and covered the major functions of human GMs.

Results

The large-scale bacterial cultivation expands human gut microbial collections

Totally 239 fresh feces samples of healthy Chinese volunteers (For gender and age, see Table S1) were collected and used for large-scale gut microbial isolation and cultivation, by following a previously established workflow [29] and using 11 pretreatment methods and 67 different culture conditions (Tables S2 and S3). Single colonies on agar plates were collected and sequenced for 16S rRNA genes (>1.4 kb). We totally harvested over 18,560 colonies, and 10,558 pure cultures were obtained (culture IDs and their full-length 16S rRNA gene sequences are available in Table S4). The phylogenies of those cultures were determined with BLAST analyses of their 16S rRNA genes against both the EZBioCloud and the NCBI 16S ribosomal RNA sequence databases. The 10,558 cultures were phylogenetically grouped into 404 taxa at species level, and 107 taxa were potentially new species (Table S5 and hGMB homepage [30]). We sequenced 115 taxa, including the 107 new taxa and 8 other taxa that were previously cultivated but were not sequenced. As a result, 115 new genomes including 21 complete genomes and 94 draft genomes were achieved and they are publicly-accessible via public databases as NODE, NMDC and NCBI (see Data Availability). The 107 new taxa were taxonomically characterized according to (1) phylogenetic analysis,

(2) morphology observation, (3) BIOLOG tests, and (4) genome features. Results demonstrated that the 107 taxa were all new species, and 28 novel genera and 3 novel families were proposed. These novel taxa were denominated and their protologues are provided in Table 1. More detailed phenotypic features of these novel taxa are documented in Supplementary Data 1. With these efforts, we constructed the human Gut Microbial Biobank (hGMB), which comprises of 1,170 strains (Table S5) representing 404 bacterial species from 159 genera, 53 families and 6 phyla (Figure 1a). All 1,170 strains in hGMB have been deposited in China General Microbiological Culture Collection Center (CGMCC), and the type strains of new taxa were also deposited in Korean Collection for Type Cultures (KCTC) or NITE Biological Resource Center (NBRC) (Table 1). The strain accession numbers, their genome data and phenotypical features are available at hGMB homepage [30] and eLMSG [31].

To better understand the cultivated bacterial diversity of human gut microbiota, we compared the hGMB with recent major works on large-scale collections of human gut microbes, as of SPORE [23], CGR [20], BIO-ML [21], Culturomics [25], and HBC [24]. By revisiting the data and extracting the taxonomic information from those studies, we (1) mined the taxonomy status of all known bacterial taxa, and (2) identified any taxa without validly published names. The results are displayed in Figures 1b-1d. The 6 studies (SPORE [23], CGR [20], BIO-ML [21], Culturomics [25], and HBC [24] and this hGMB) collected in total 1,523 bacterial species from human gut. Figure 1c shows the shared and unique bacterial species among the 6 studies. The hGMB provides 142 unique gut microbial species. Notably, 80 of the 142 hGMB unique species had not been cultured previously. We also calculated the numbers of new taxa at each study (Figure 1d). The six studies claimed in total 420 nonredundant novel taxon candidates, and the hGMB contributed 107, accounting for 25.7% of the total new taxa.

New taxa reveal previously-uncultured MAGs, and are prevalent in global human gut microbiome

Of the 107 new taxa, 16 taxa were on the list of “most-wanted” or “medium priority” taxa proposed by the Human Microbiome Project [40] (Table S5). The “Most Wanted” taxon *Eubacterium difficilis* sp. nov. (Taxon_69) was firstly cultivated in this study, and 3 new genera *Simiaoa* gen. nov., *Jutongia* gen. nov. and *Wansuia* gen. nov. of the “medium priority” taxa were characterized in this study (Table 1). To identify the representativeness of the 107 new taxa in the uncultured metagenome-assembled genomes (MAGs) as well as their prevalence in global human GMs (n=3,810), we performed genome-wide analysis on 23,790 genomic OTUs from the integrated genome database for GM and other environments (IGGdb) [26]. Results revealed that 75 genomic OTUs from IGGdb matched with our new taxa genomes. In addition, 42 of the matched genomic OTUs were previously-uncultured MAGs (Table S6). Thus, the 42 previously-uncultured MAGs have representatives of cultivated bacterial species and genomes in the hGMB. Further analysis manifested that 39 of the 75 hit genomic OTUs were prevalent in global human GM (ie: prevalence ≥5% among 3,810 human gut metagenomes defined by IGGdb work [26]) (Table S6).

To further evaluate the prevalence of the 107 new taxa in global human gut microbiota, we collected publicly-available 16S rRNA gene amplicon datasets of 26 studies (N=26) from NCBI SRA database (date: 2020-02-22). The 26 datasets, each had specimen numbers ranging from 102 to 3,538, represented

human gut microbiotas from donors of diverse genetic and environmental backgrounds (see Table S7 for accessions of each study). The 26 datasets were separately processed, quality-controlled and weighted by a standard USEARCH-based analysis pipeline as described in Methods section. Microbial taxonomy information was extracted from the 26 datasets using LTP_vhGMB customized by supplementation of LTP database v132 with the 107 new taxa. Results showed that the 26 datasets contained totally 11,647 quality-controlled specimens (n=11,647) and that each specimen had 228±85 OTUs. Our analyses showed that the 107 new taxa covered in average 36.2±11.0% of the total reads of 16S rRNA gene sequences from the 26 studies (Figure 2a). Figure 2b shows that 92 of the 107 new taxa were found in at least one of the 26 studies, 42 new taxa in over 50%, and 3 species (*Eubacterium difficilis* sp. nov., *Changea tenuis* gen. nov., sp. nov. and *Neobitarella massiliensis* gen. nov., sp. nov.) in all 26 studies. At genus level, 25 out of the 28 novel genera were detected in the 26 studies, and 3 (*Neobitarella* gen. nov., *Changea* gen. nov. *and Lentihominibacter* gen. nov.) were found in all 26 studies (Figure 2c). At family level (Figure 2d), the Family *Bianqueaceae* fam. nov. was detected in all the 26 studies, and the Families *Feifaniaceae* fam. nov. and *Tianshiaceae* fam. nov. were detected in 25 and one of the 26 studies, respectively.

The hGMB largely represents the taxonomic diversity of human gut microbiota

With the customized LTP_vhGMB database, we re-annotated the total 16S rRNA gene amplicon reads from the 26 studies (Table S7). The equally-weighted average relative abundance (RA) and frequency of occurrence (FO) for each annotated species or genus was calculated as described in Methods. Results showed that 76.3±8.0% and 53.7±11.8% of the total reads were assigned into 990 genera and 1461 species, respectively. As shown in Figures 3a and 3b, the accumulation curves were almost saturated after 24 sampling datasets from the 26 studies, at either genus or species level. It manifested that the taxonomic composition of the 26 studies could largely represent the human gut microbiota diversity at genus and species levels. At 16S rRNA gene amplicon level, the hGMB members (404 species) covered 82.0±12.3% of the total reads of the 26 studies. At genus level, 140 out of the 159 genera in hGMB were detected in the 26 studies. We identified that 386 genera appeared in over 1% (equally-weighted average FO>1%) of the 26 study samples (n=11,647), and hGMB covered 129 genera of them. If we defined the genera with equally-weighted average RAs> 0.1% as “dominant genera”, and those genera with equally-weighted average FOs>30% as “common genera”, 69 and 74 genera were recognized as dominant and common genera, respectively (Figure 3c). The 69 dominant genera represented 94.7±4.7%, while the 74 common genera represented 91.3±11.3%, of the total annotated 16S amplicon reads. The hGMB covered 85.1% and 84.1% of the common and dominant genera, respectively. If the same criteria were used to define “dominant species” (equally-weighted average RAs> 0.1%) and “common species” (equally-weighted average FOs>30%), 91 dominant and 84 common species were recognized from the 26 studies (Figure 3d). The hGMB covered 79.1% of the dominant species and 80.9% of the common species. There were 12 and 16 newly described species of hGMB belonging to the dominant and common species, respectively. Noticeably, 5 novel species from this study (*Simiaoa sunii* gen. nov. sp. nov., *Ruminococcus bicirculans* sp. nov., *Faecalibacterium hominis* sp. nov., *Dysosmobacter segnis* sp. nov., and *Wujia*

chipingensis gen. nov. sp. nov.*)* were both dominant and common species in human gut microbial communities, which fell into the category of “core taxa” in GMs (Figure 3c and 3d).

The hGMB covers and enriches the global human gut gene catalog

Gene cataloging outlines human GM functionality potentials, and several gene catalogs have been established [8, 41]. We created nonredundant gene catalogs with hGMB genomes of the 107 new taxa (named hGMB.new) and of the 404 taxa (named hGMB.all), and compared them with the largest-to-date human GM catalogs, i.e., the Integrated Gene Catalog (IGC) that contains over 9.3 million nonredundant genes and the Unified Human Gastrointestinal Protein (UHGP) catalogue that contains 13 million nonredundant protein sequences. Of the 404 genomes, 115 were newly sequenced in this study, while the left 289 of known taxa were retrieved from NCBI Assembly database. The hGMB.new and hGMB.all catalogs contained 322,792 and 1,074,954 nonredundant genes, respectively. Though the majority (76-95%) of the nonredundant genes in hGMB catalogs were covered by IGC and UHGP (Table S8), hGMB further enriched human GM gene catalogs. With threshold value of 60% of amino acid sequence identity (for functional conservation), the hGMB contributed 85,424 and 250,167 of nonredundant sequences to the UHGP and IGC, respectively. When the identity value was decreased to 40% (for structural conservation), the numbers of new genes added to the UHGP and IGC were 48,741 and 119,955, separately. As shown in Figure 4a, the hGMB covered 26.9% and 35.9% of IGC catalog under the threshold identities of 60% and 40%, respectively. For UHGP, the coverages of UHGP genes by hGMB.all were 26.8% and 36.8%, at functional and structural level, respectively.

We then investigated the hGMB representativeness of the characterized functions of human GM. For this purpose, the UHGP, IGC, hGMB gene catalogs as well as all the hGMB genomes were annotated with eggNOG4 [15]. Results showed that 69.0% of genes in IGC catalog were annotated into seed eggNOG orthologs, 59.7% into COGs, 38.4% into KOs, and 19.3% into GOs (Gene Orthologs). For all proteins of UHGP catalog, 69.4%, 60.2%, 39.5% and 20.9% of the UHGP-90 sequences (sequences clustered at 90% identity) were annotated into seed eggNOG orthologs, COGs, KOs and GOs, respectively. The identities of KOs and GOs in ICG and UHGP catalogs and in hGMB genomes were extracted for generating profiles displaying the presence/absence binary code (0/1) of genes from hGMB genomes in the two catalogs. A cumulative analysis of the KO and GO profiles were conducted to unravel the coverages of each catalogs by random incremental selection of the hGMB genomes, and the results were shown in the rarefaction curves (Figure 4b and 4c). It revealed that the hGMB genomes covered 71.3% and 74.1% of the KO genes from IGC and UHGP catalogs, respectively (purple lines in Figure 4b and 4c). Similarly, the hGMB genomes represented 62.0% and 65.6% of the known GO functions of IGC and UHGP catalogs, respectively (blue lines in Figure 4b and 4c). Moreover, the CRs of both KO and GO functions increased rapidly to over 50% for both IGC and UHGP as the sampled genome numbers reached up to 40. After that, the increasement of CRs slowed down, and approached saturation when over 200 genomes were sampled. It implied that the best-fitting 40 hGMB members showed great potential in representing the major known features of human gut microbiome.

The hGMB illuminates the functional “dark matter” in human gut microbiome

In addition to the well representativeness of functionally-known genes of human GMs, the hGMB provided also a cultivable repository of functionally unknown genes in the global gene catalogs. According to the eggNOG annotation results of IGC and UHGP catalogs, there were 30.9% and 30.6% genes/proteins were functionally unknown. BLAST analysis (amino acid sequence identity>40%) revealed that the hGMB genomes covered 11.1% (grey line in Figure 4b) and 9.6% of them (grey line in Figure 4c). Those functionally unknown genes matched for IGC and UHGP are listed in Tables S9 and S10. The presence frequency of each hit genes among 404 genomes were then calculated, and the top 100 frequently unknown-genes in IGC and UHGP were respectively displayed in the heatmaps of Figure S1 and S2.

The Function Unknown Families of homologous proteins (FUnkFams) is a “most wanted” list of conserved microbial protein families with no known domains that should be prioritized for functional characterization [42]. The FUnkFams comprised 61,970 amino acid sequences from 6,668 conserved protein families. We plotted the accumulative coverage of unannotated genes from hGMB genomes to FUnkFams, and the results revealed that, with a threshold value of 40% sequence identity, the hGMB.all catalog covered 28.1% of the functionally unknown proteins in FUnkFams (Figure 4d). The profiles of the matched FUnkFams sequences to the hGMB genomes are summarized in Table S11 facilitating further culture-based study and the top 100 frequent genes were exhibited in Figure S3. We deduced that newly identified bacterial taxa contained more functionally-unknown genes. Indeed, the 107 novel taxa carried significantly more unannotated genes than the known species did (Figure 4e), which implied that the cultivation and characterization of new taxa would facilitate the culture-based experimental studies to bring more human gut “dark” functions to light.

We observed that the distribution of the top 100 frequently hit unknown genes from three gene catalogs were mainly clustered by the taxonomic distance of hGMB taxa, and the functionally unidentified genes in Bacteroidetes was more conserved than those in Firmicutes and other phyla (Supplementary Figure S1-S3). To further confirm this observavtion, PCoA analyses were performed with the hit functionally-unknown genes of FUnkFams, IGC, UHGP, hGMB.new and hGMB.all gene catalogs. As manifested in Figure 5, the distribution of functionally-unknown genes in Bacteroidetes genomes were distinct from genomes of the other phyla. We then developed and illustrated an LDA Effect Size (LEfSe)-based analysis for target taxonomy-specific conserved functionally-unknown genes that should be prioritized for further culture-based functional characterization. The presence/absence profile of hGMB.all functionally-unknown gene hitting in hGMB members of Order Bacteroidales and Order Clostridiales were extracted, and the LEfSe of top 500 abundant hit genes were calculated. The heatmaps of the top 100 significantly differential genes in either Bacteroidales or Clostridiales groups shown in supplementary Figure S4, displayed clearly segmented distribution of functionally-unknown genes between two groups and genes in Bacteroidales were more conserved.

Discussion

Due to the rapid development of omics techniques such as high-throughput sequencing and data-mining tools, cultivation-independent studies have generated massive data and powerful predictions to advance the profiling and understanding of GM features [6, 43–46]. Still, researchers encounter challenges in interpretation of omics data, as considerable amounts of bacterial taxa and genetic elements are “dark matters” [9, 11]. According to our analysis in this study, about 30–40% of human gut microbial genes in IGC and UHGP catalogs were functionally-unknown and they are hiding GM secrets and interactions to human host. As shown in Fig. 4b-d, the hGMB covered 10% of the functionally-unknown genes in IGC and UHGP, and 28% of the “most wanted” domain-unknown proteins in FUnkFams [42]. We profiled the distributions of above mentioned hits among hGMB members and tabulated them into 0/1 tables (Table S9-S11) to enable a fast browsing of gene distribution and to target gene hosts. The displays of frequently-presenting genes among hGMB members (Figure S1-S4) revealed a taxonomy-associated clustering of functionally-unknown genes, and more conserved and functionally-unknown genes were observed in Bacteroidales. We annotated the top 100 significantly differed genes list in Figure S4 using Pfam database, and the results revealed that 42 of them got no hit on any protein/domain in Pfam, while 24 got hit on immunity 17 protein, a predicted immunity protein deduced to function in toxin defense in bacterial polymorphic toxin systems [47].

By implications of previous experiences on cultivation and understandings of gut microbial physiology and ecology [23, 25, 29, 48], we adopted 11 pretreatments and 67 culture conditions (including different media) and obtained 10,558 pure bacterial isolates in this study. Intensive efforts were made on modification of cultural media, particularly in diversifying the ingredients in culture media (Table 3). Based on our previous study [29], we found that mouse gut microbes preferred 8 carbon sources (D-mannose, D-fructose, Fructo-oligosaccharide, D-galactose, Palatinose, L-Rhamnose, D-(+)-Cellobiose and D-Trehalose) for growth. In this study, those 8 carbohydrates mixture was supplemented to media to improve human gut microbial cultivability (Table 3). The results indicated that this mixture improved the growth of quite a few of new gut bacterial isolates, especially members of *Clostridiales* and *Erysipelotrichales*. According to our statistics, the *Eubacterium hominis* sp. nov., *Eubacterium segnis* sp. nov., *Agathobaculum hominis* sp. nov., *Fusobacterium hominis* sp. nov., *Wujia chipingensis* gen. nov. sp. nov. and *Luoshenia tenuis* gen. nov. sp. nov. were all exclusively isolated from agar plates of modified mGAM supplemented with 8 carbohydrates mixture.

In this study, 107 new species, 28 new genera and 3 new families of human gut microbes were identified, characterized and denominated. According to our analysis, 80 of the 107 newly denominated hGMB species were cultivated exclusively in this study. The other 28 unclassified taxa were ever cultivated in previous studies [20, 21, 23–25], but were not systematically characterized neither denominated effectively. The genomes of all 107 newly-characterized species were sequenced. These new taxa as well as their genome sequence data enriched human gut microbial resources and would be valuable assets for mechanistic studies on host-microbiome interactions. Besides of 107 new species, we also sequenced 8 previous cultured species that had not genome data. For the other 289 species in hGMB, rather than re-sequencing of their genomes, we used publically available genomes to performed gene catalog analysis

in this study. We believed that the public genomes well represent the core genomic features of these species as genomic ANI values within the same species > 95%. Yet, in the future and if desired, re-sequencing of these strains in hGMB would be easily achieved.

Genome BLAST analysis of the new taxa revealed that 42 previously-uncultured MAGs from IGGdb have representatives in hGMB, and 22 out of the 42 newly cultured genomes represented the prevalent species-level OTUs in global human GMs (n = 3,810). Further taxonomic analysis revealed that 9 of prevalent new species (*Jutongia hominis* gen. nov. sp. nov., *Jutongia huaianensis* gen. nov. sp. nov., *Wujia chipingensis* gen. nov. sp. nov., *Dorea hominis* sp. nov., *Enterocloster hominis* sp. nov., *Lachnospira hominis* sp. nov., *Roseburia rectibacter* sp. nov., *Roseburia wangyiboensis* sp. nov. and *Roseburia xiaozhani*s sp. nov.) were from *Lachnospiraceae*. The *Lachnospiraceae*, as one of the most dominant families in the GM of healthy adults, accounted for 10–45% of total bacteria in feces [49], and were considered playing diverse but controversial roles in the maintenance of host gut homeostasis [50, 51]. On one hand, *Lachnospiraceae* members such as *Roseburia* species, were beneficial to hosts via production of short-chain fatty acids (SCFAs) and secondary bile acids [52–54], protection of hosts from pathogen infections [53, 55, 56] and from stress-induced visceral hypersensitivity [52]. Yet, on the other hand, researches displayed positive correlations between *Lachnospiraceae* and diseases such as non-alcoholic fatty liver disease (NAFLD) [57] and chronic kidney disease (CKD) [58]. Animal experiments demonstrated that the gavage of *Lachnospiraceae* accelerated the development of diabetes in obese mice [59] and aggravated the inflammation of intestinal epithelial cells in $TLR5^{-/-}$ mouse [60]. The contradictory conclusions signified that the function(s) of *Lachnospiraceae*, a predominant gut microbial family in humans, are very complicated. As a solution, the culture-based study of *Lachnospiraceae*-host interactions would enable a better understanding of their complex roles in health and disease, on condition that diverse cultivable *Lachnospiraceae* members are available. The hGMB contains 93 strains from 49 different *Lachnospiraceae* species, provides an accessible *Lachnospiraceae* repository for future study. As shown in Tables S5 and S6, we totally characterized 30 new *Lachnospiraceae* species of 17 genera, including 7 new genera (*Wujia* gen. nov., *Simiaoa* gen. nov., *Jutongia* gen. nov., *Qiania* gen. nov., *Zhenhengia* gen. nov., *Jingyaoa* gen. nov., *Wansuia* gen. nov.).

The hGMB provides also members of *Christensenellaceae*, including *Christensenella minuta*, *Christensenella tenuis* and 3 new genera (*Changea* gen. nov., *Luoshenia* gen. nov. and *Gehongia* gen. nov.). The *Christensenellaceae* is a recently identified gut commensal bacterial family containing limited cultivated representatives [61], and has been considered as a promising probiotic candidate for intervention of obesity and other metabolic syndromes [62, 63]. Particularly, the *Christensenella minuta* was experimentally verified to reduce weight gain in recipient mice [64]. To explore and evaluate *Christensenellaceae*'s therapeutic potential, more studies are needed. The hGMB provides resources and serve further studies. Notably, the *Changea* and its type species *Changea tenuis* was widely occurring in global human GMs as they were found in all investigated datasets, making it an interesting candidate for study. In addition to the contribution of previously-uncultured gut microbes to the public (Table S6, Figs. 2 and 3), the hGMB also includes considerable numbers of strains representing known species that were

research hotspots in human GM studies. Some of these “star species” were commonly recognized to have probiotic potentials, such as *Akkermansia muciniphila* [65], *Faecalibacterium prausnitzii* [66], *Roseburia intestinalis* [67], *Lactobacillus* and *Bifidobacterium* members [68, 69], while some others, as *Enterococcus faecium* [70], *Ruminococcus gnavus* [71], *Clostridioides difficile* (*Peptoclostridium difficile*) [72] and *Klebsiella* species [73], were revealed to play pathogenic role in hosts. There is a large group of gut microbial species that were reported to have strain-specific effects on hosts [74, 75]. An example is the *Bacteroides fragilis*, as both pathogenic and probiotic strains were identified from this species [75, 76]. Most recently, the *Bacteroides xylanisolvens* strain from hGMB has been demonstrated to function as probiotic in alleviation of nonalcoholic hepatic steatosis via Bacteroides-Folate-Liver Axis [77]. In summary, the hGMB improves the cultivated GM diversity and thus would facilitate in-depth and extensive studies of their functional features.

Conclusion

In this study, 10,558 bacterial isolates from 239 fecal samples of healthy Chinese volunteers were obstained. Based on 16S rRNA gene identity, those bacterial isolates were identified to be 404 species of 159 genera, belonging to 53 families and 6 phyla. A publically accessible human Gut Microbial Biobank (hGMB) that contains 1,170 representative bacterial strains and represent 404 human gut microbial speces was estabolished. The hGMB expands the currently known, taxonomically-characterized gut microbial resources and genomic repository by adding 107 new species and 115 new genomes of human gut microbes. Based on the newly discovered species in this study, 28 new genera and 3 new families of human gut microbes were identified and proposed. All new taxa were described and denominated following the rules of ICNP. Further analysis revealed that the hGMB represented over 80% of the prevalent microbial genera and species in human guts, and covered 70% of KEGG Orthology functions. The hGMB harbored about 10% of the functionally-unknown genes in global human gut gene catalogs, and observed taxonomy-associated distribution of functionally-unknown genes among hGMB members. By integrative analysis of global human GM datasets, we profile the taxonomic prevalence, distribution and genetic features of the 404 hGMB species among global human GMs, demonstrating that the hGMB has great potential in bringing more human gut microbial “dark matters” to light.

Methods

Sample collection and treatment

The feces samples (n=239) were collected from healthy volunteers who did not receive any medical treatment for the last 2 months before sampling. The sample donors were mainly from six different area of China (Beijing, Henan, Heibei, Xinjiang, Guangdong, Inner Mongolia), and the age and gender information of them was listed in Table S1. The samples collected in Beijing were kept in fresh and transferred into anaerobic workstation (AW500, Electrotek, UK) for sample pretreatment within 2 hours, while the feces from the other area were frozen on dry ice immediately after sampling and delivered to the Lab for pretreatment. To enable a better recovery of diversity, about 10 samples collected at the same

time and place were merged together for pretreatment and subsequent isolation steps. The 11 pretreatment conditions were given in Table S2 and the alcohol pretreatment strategies were derived from Lawley et al. [23]. The atmosphere composition in anaerobic workstation was 85% N_2, 5% CO_2, and 10% H_2.

Bacterial isolation and cultivation

The pretreated samples were filtered using cell strainer (BD Falcon, USA) to remove the large insoluble particles in suspension and diluted into 10^{-1}-10^{-8} folds. Then, 100 µl of each dilution were plated onto different agar plates for either aerobic or anaerobic incubations at 37 °C. We used 67 different culture conditions for bacterial cultivation and isolation as shown in Table S3. The detailed recipes of 21 base medium and Supplements used in this study are provided in Supplementary Methods. The supplementation of clarified rumen fluid and sheep blood in culture media was conducted by following Lagier et al. [25]. The colony isolation and identification were performed as described in our previous study [29] : All the single colonies appearing on the agar plates after incubation for 2 to 60 days were picked. The picked colonies were then inoculated into 48-well plates containing 700 µl of broth media in each well. The 96-well plates containing isolates were incubated at 37 °C under 2-30 days depending on the growth rate of isolates. Then, 50 µl of the media in each well were collected and centrifugated at 13,000 rpm for 1 min. The bacterial pellet was lysed with 2 µl of NaOH/SDS lysis buffer (Amresco, USA) and diluted with 100 µl deionized water. Two microliters of the templates were used for PCR-based amplification of 16S rRNA gene sequences with DreamTaq Green PCR Master Mix (Thermo Fisher Scientific, USA) (primers: 27 F: 5'-AGAGTTT GATCCTGGCTCAG-3'; 1492 R: 5'-GGTTACCTTGTTACGACTT-3'). The PCR products were identified using Sanger sequencing (TIANYI HUIYUAN Ltd., China). The wells containing a single 16S rRNA gene were further enlarged and cultured by inoculation in tubes containing 5 ml of liquid media and streaking on agar media plates for further preservation and characterization either anaerobically or aerobically. The phylogenies of all the cultured isolates were recognized by BLAST analysis of the 16S rRNA gene sequences against both the EZBioCloud and the NCBI 16S ribosomal RNA sequence database (Update date: 2020/08/08, number of sequences: 21,632). The isolates with 16S rRNA gene sequence identities>98.7% to any species (valid names only) in EZBioCloud were considered as known species. The isolates with 16S rRNA gene sequence identities ≤ 98.7% to any known species in both databases were considered as candidates of new taxa. All the isolates potentially representing new taxa were further grouped into different species-level clusters based on the 16S rRNA gene sequence identity (cut-off value 98.7% for different species) and for each species-level new taxa, 1 strain was designed as type strain for latter genomic sequencing and polyphasic characterization.

The preservation strategy of bacterial strains

We totally performed 16 batches of bacterial isolation and cultivation, and used different fecal samples in each batch of work. In each batch, we deposit 1 representative strain of every identified species for long-term cryopreservation in CGMCC for public use, no matter whether strains of these species had ever been preserved or not in previous batch of isolating work. We use such redundant-preservation strategy to

1) ensure that at least 1 strain for each species could be properly recovered after long-term storage, and 2) enable a better strain-level diversity in hGMB considering that the strains of the same species from different donors might differ in genomic or physiological features. The cryopreservation of selected strains were performed as described in previous work [29]: Pure cultures were inoculated onto agar plates and incubated until enough single colonies appearing on the plates. All the colonies on agar plates were collected using cell scraper, suspended in protective solution (15% glycerol and 85% bovine serum solution) and stored at −80 °C or in liquid nitrogen. The CGMCC accessions of 1,170 preserved strains were available in Table S5 and hGMB homepage. To meet the rules of International Code of Nomenclature of Prokaryotes (ICNP), the 107 type strains of new species in hGMB were also preserved in a second IDA as KCTC or NBRC, and the accessions could be found in Table 1 and hGMB homepage.

Polyphasic characterization and nomenclature of new taxa

The delineations of new taxa were based on the analysis of each type strain in terms of phylogenetic, genomic, physiological and morphological characteristics as described in previous work [29, 78] and documented in Supplementary Data 1. For each new species, the phylogenetic tree was constructed with the 16S rRNA gene sequences of the type strains from the phylogenetically close neighboring genus and species using MEGA7 [79] under Neighbour-Joining method to depict the phylogenetic distribution and taxonomic relation of each new taxa and its closely-related taxa (Figure SD-1a to Figure SD-108a in Supplementary Data 1). The genome-based analysis of new taxa included the calculation of the Average Nucleotide Identity (ANI), digital DNA:DNA hybridization (dDDH) and the percentage of conserved proteins (POCP). The ANI was calculated with OrthoANIu tool [80]. The dDDH value between draft genome of new species and its phylogenetically closest genome was calculated using the Genome-to-Genome Distance Calculator 2.1 (GGDC) [81]. The POCP between each genome and its phylogenetically closest genome was calculated using BLASTp v2.9.0+ and was used for taxonomy delineation at genus level [82]. The physiological and biochemical features of type strains of new taxa were profiled using ANI MicroPlates (BIOLOG, the USA) following the manufacturer's instruction. The bacterial cell morphology was observed using transmission electron microscope (TEM) JEM-1400 (JOEL, Japan) (Figure SD-1b to Figure SD-108b in Supplementary Data 1). The motility of bacteria was examined with the light microscopy Axiostar plus 156 (ZEISS, Germany). The nomenclature of each characterized new taxa was proposed according to the rules of ICNP. Taxon meeting the following three criteria simultaneously was defined as new species: 1) the 16S rRNA sequence identity<98%, 2) ANI<95% and 3) dDDH value<70%. The new species has a 1) 16S rRNA gene sequence identity<95% to any known species, 2) POCP value <50% to its closest genome, 3) significant difference in morphology and physiology with neighbor genera, and 4) an independent clade on the phylogenetic tree would be further defined as new genus. If the type species in the new genus 1) had a 16S rRNA gene sequence identity<90% to any known species, 2) is clustered on a separate clade distant from and known genera on the phylogenetic tree and its closest neighbor genera were from different families, and 3) maintained significant difference in morphology and physiology to the neighbor families would be further defined as new family.

Genome sequencing, collection and analysis

The genomes of all 107 new species and 8 known species with no genome available were sequenced. The genomic DNA were extracted using the DNeasy Blood & Tissue Kit (Qiagen, Germany). The DNA concentrations were measured using Qubit 4.0 (Thermo Fisher Scientific, USA). The degradation of purified DNA was checked by electrophoresis, and the DNA was considered as undegraded if no apparent smear was observed on agarose gel. The bacterial species having more than 5 mg undegraded DNA were sequenced using PacBio SMRT technique for achievement of complete genomes. The qualified genomic DNA was fragmented with G-tubes and end-repaired to prepare SMRTbell DNA template libraries (with fragment size of >10 Kb selected by bluepippin system) according to the manufacturer's specification (PacBio, USA). Library quality was detected by Qubit 3.0 Fluorometer (Life Technologies, USA) and average fragment size was estimated on an Agilent 4200 (Agilent, CA). SMRT sequencing was performed on the Pacific Biosciences RSII sequencer (PacBio, USA), according to standard protocols. The raw reads were filtered by the SMRT 2.3.0 to discard low quality reads and the filtered reads were assembled to generate one contig without gaps. The hierarchical genome-assembly process (HGAP) pipeline was used to correct for random errors in the long seed reads (seed length threshold 6 Kb) by aligning shorter reads from the same library against them. The corrected, preassembled reads were used for de novo assembly. For the genomic DNAs not qualified for SMRT sequencing were sequenced using Hiseq X-ten platform (Illumina, USA) to generate draft genomes. The sequencing libraries were generated with NEB Next® Ultra™ DNA Library Prep Kit for Illumina® (New England Biolabs, USA) following manufacturer's recommended procedures and the index codes were added. The library quality was evaluated by the Qubit 3.0 Fluorometer (Life Technologies, USA) and the average fragment size was estimated using Agilent 4200 (Agilent, CA). The DNA library was sequenced on an Illumina Novaseq platform and 1-2 GB 150 bp paired-end reads were generated. The raw data were quality controlled using company's own compiling pipeline. The filtered paired reads were assembled using the SPAdes software v3.9.0 [83] into a number of contigs, and the contigs longer than 500 nt were retained as final splicing. Above library preparation, sequencing and assembly steps were performed by commercial company (Guangdong Magigene Biotechnology Co.,Ltd., China). For the genome component prediction, the coding genes were predicted with glimmer3 [84] and Prodigal v2.6.3 [85], and the rRNA genes were retrieved by RNAmmer v1.2 [86]. For the 289 previously sequenced species, the nucleic acid and amino acid sequences of each genomes were downloaded from NCBI Genome and Assembly databases (All genome accessions were available in Table S5 and hGMB homepage [30]) for reannotation together with our new genomes. The function annotations of all genomes were performed with eggNOG database v4.5 by local emapper v1.0.3 (-m diamond) [15]. Default parameters were used for each software unless otherwise specified. The coverage of hGMB new genomes to the uncultured and prevalent species-level genomic OTUs in IGGdb were analyzed by mapping the filtered short reads of new genomes to the 6,198,663 marker genes identified for the 23,790 OTUs in IGGdb v1.0.0 with IGGsearch [26]. Only the genomes hitting on the OTU with either the percentage of makers detected >90% (percent_markers_detected parameter) or the species abundance >90% (species_abund parameter) were considered as high-confidence hit and exhibited in Table S6. The prevalence of each hit OTUs in global human microbiomes (n=3,810) were derived directly from the Table 17 of IGGdb work [26].

Bacterial diversities of different culture collections

We collected the taxonomic information of cultures from five representative large-scale cultivation-based studies (CGR [20], BIO-ML [21], SPORE [23],HBC [24] and Culturomics [25]) of human GM for diversity comparison and determination of the resource overlaps. The taxonomic information of all known species was directly mined from corresponding publications, and the taxonomic names of them were used for further comparison. For those unclassified new isolates without validly published names, their corresponding 16S rRNA gene sequences were used for bacterial diversity comparison. The 16S rRNA gene sequences were either retrieved from publication (Culturomics [25]) or extracted from genome data using RNAmmer v1.2 [86] (for CGR [20], BIO-ML [21], SPORE [23] and HBC [24]). The genome-derived 16S rRNA gene sequences > 1 kb were retained for further analysis. The 16S rRNA gene sequences of new taxa isolates/genomes from one study were clustered using Usearch11 (command: -cluster_fast query.fasta -id 0.987 -centroids clustered.16S.fasta -uc clusters.uc) to reveal the nonredundant 16S rRNA gene sequences of species-level new taxa in each study. There were 68, 100, 141 and 22 new taxa recovered from genomes based on a 16S rRNA gene identity<98.7% for study SPORE [23], CGR [20], HBC [24] and BIO-ML [21], respectively. With this method, we totally recovered 1,056 species for Culturomics, 106 for BIO-ML [21], 121 for SPORE [23], 236 for CGR [20] and 319 for HBC [24]. For SPORE [23], CGR [20] and HBC [24], the number of recovered species was a bit less than that was reported in original papers, which was due to the use of different criteria (genome-based ANI or 16S rRNA gene sequence identity) in species identification depending on each work. We then analyzed the overlaps of potentially new taxa among studies. The 16S rRNA gene sequences representing new taxa in each study were combined together, and the Kimura 2-parameter model based evolution distance between 16S rRNA gene sequences was calculated using MEGA7 [79]. If the new isolates from different studies had 16S rRNA gene sequence distance <0.013 to each other, they were regarded as the “shared” species by those studies, otherwise, the isolates were defined as study-unique new taxa. To display the hGMB coverage of Human Microbiome Project's Most Wanted taxa [40], the OTU sequences of the “Most Wanted” taxa analysis were collected, and used for BLAST analysis against the 16S rRNA gene sequences of hGMB members with Blastn v2.9.0+ [87]. If the 16S rRNA gene sequences of hGMB members had sequence identities >97% to the OTUs representing taxa of high and middle priority defined in previous work, then the corresponding hGMB members were considered as cultivable “most wanted” taxa and indicated in Table S5 (Column named as "Most wanted" taxa). All the taxa included by the hGMB were exhibited as taxonomic cladogram using GraPhlAn v1.1.3 [88], and the species presenting exclusively in hGMB were displayed as outer ring of the cladogram. The unique and shared bacteria within hGMB and five investigated collections were displayed using Venn and bar charts generated by Jvenn [89].

The 16S rRNA gene amplicon data collection and analysis

We collected 26 publicly-available 16S rRNA gene amplicon datasets from NCBI SRA database. The accessions, sample size, location, host phenotype and other basic information of the 26 NCBI Bioprojects are given in Table S7. To enable an equally-weighted representation of human GMs, the 26 studies were separately processed and quality-controlled by 64-bit Usearch v11 [90] following the recommended

uparse-based pipeline (https://drive5.com/usearch/manual/uparse_pipeline.html). The only modification of the procedure was that an additional chimera removal step was introduced after OTU sequences were generated with the command "-uchime2_ref" against SILVA v132 database. After generation of OTU table for each study, the samples maintaining <10,000 reads were removed. As a result, 11,647 out of the 13,055 samples from 26 studies were retained for further analysis, and each sample contained 228±85 OTUs. The OTU sequences of each study were then annotated using a customized database LTP_vhGMB developed by update of the LTP database v132 [16] with the taxonomic information of 107 new taxa in hGMB. The RA and FO of annotated species, genera and families for each separate study and for all the 26 studies together were calculated as described in our previous publication [29]. The equally-weighted average values (RA, FO and CR) were further calculated by averaging the mean values of each study. All the mean values of CRs, RAs, and FOs relating to the 26 studies exhibited in the Results were presented as the equally-weighted average values ± standard deviation (SD) unless otherwise specified. The presence/absence and the abundance of each hGMB new taxon among the 26 studies were displayed as heatmaps. BLAST analysis of the 16S rRNA gene sequences of hGMB members against the OTU sequences of each study was performed, and the CRs of total reads from 26 studies by hGMB taxa were calculated by summation of the RAs of OTUs having sequence identity >97% to 16S rRNA gene sequences of hGMB. The equally-weighted average RA> 0.1% was the criterion to define dominant species/genera, while the equally-weighted average FOs>30% was the criteria for definition of common species/genera in global human GMs. The saturability of sampled studies were calculated using specaccum function in vegan R package [91] and displayed as accumulating curves. The distribution of dominant taxa in global human GMs were displayed as box and whiskers plot while the common taxa were displayed as bar charts.

Gene catalog construction and analysis

The representative metagenome-based human gut Integrated Gene Catalog (IGC) [41] containing over 9.3 million nonredundant genes, the largest-to-date genome-based Unified Human Gastrointestinal Protein (UHGP) catalogue [8, 10] comprising 13 million nonredundant protein sequences and the Function Unknown Families of homologous proteins (FUnkFams) catalogs [42] comprising 61,970 amino acid sequences from 6,668 conserved protein families were downloaded and reannotated with eggNOG database v4.5 by emapper v1.0.3 (-m diamond) [15] and generated indexed databases for each gene catalogs with DIAMOND v0.9.24 (makedb command) [92]. The nonredundant gene catalog hGMB.new was constructed using 107 new genomes sequenced in this study and the hGMB.all was constructed using 115 genomes in this study and 297 representative genomes downloaded from NCBI Assembly database (The accessions of all mentioned genomes could be found in hgmb.nmdc.cn and Table S5) by CD-HIT software v4.5.8 [93] (-o out.file -c 0.95 -aS 0.9 -n 5 -M 64000 -T 48). The hGMB.new and hGMB.all respectively containing 322,792 and 1,074,954 nonredundant genes were then annotated with eggNOG database v4.5 by emapper v1.0.3 [15]. The eggNOG orthologs, COG categories, KOs, GOs and functionally unknown genes were summarized from the eggNOG annotation results. For the calculation of gene coverage (%), the profiles of annotated genes in different gene catalogs and single genomes were tabularized in the form of presence/absence binary code (0/1), which were further calculated using

specaccum function in vegan R package [91] to generate data used for the construction of cumulative curves. The BLAST analysis of single genomes in hGMB and hGMB gene catalogs (hGMB.all and hGMB.new) against the IGC, UHGP and FUnkFams catalogs were performed using DIAMOND blastp (–more-sensitive -f 6 qseqid sseqid pident length qlen slen qcovhsp evalue qseq full_sseq mismatch gapopen qstart qend sstart send). The coverage rates of hGMB catalogs to the global gene catalogs were calculated with two different cutoff values of the amino acid sequence identity 60% and 40%, respectively. The 40% was the threshold identity value of Structural Classification of Proteins (SCOP), while 60% was the minimum amino acid sequence identity for function conservation [94-96].

To profile the coverage of functionally-unknown genes of IGC, UHGP and FUnkFams by hGMB genomes, DIAMOND-based BLAST analysis [92] of single genomes in hGMB against three gene catalogs were performed as described in last paragraph with a sequence identity cut-off value of 40%. The presence of each covered unannotated genes in 404 hGMB members were profiled as Table S9-11. The ratios of unannotated genes of new and known genomes in hGMB were calculated based on the eggNOG annotations of single genomes. The unannotated rates between two groups were displayed as box and whiskers plots. The distribution of unannotated genes of different gene catalogs (IGC, UHGP, FUnkFams, hGMB.all and hGMB.new) in each hGMB member was visualized with PCoA. The hit genes with presenting times> 10 were used for plotting PCoA. The taxonomy-associated distribution of functionally-unknown genes among hGMB members were analyszed using the online version of LEfSe with default parameters (http://huttenhower.sph.harvard.edu/galaxy/), and displayed with heatmaps.

Statistical analysis

All statistical analyses were performed with IBM SPSS Statistics 20. All the heatmaps were constructed using pheatmap R package [97]. All the box-and–whisker plots, bar charts and accumulating curves were generated using Graphpad Prism v6 [98] unless indicated otherwise. Comparison of two groups of data was statistically assessed with Mann–Whitney U test, while comparison of multi groups (>2) of data was evaluated by Kruskal-Willis test . $P < 0.05$ was considered being statistically significant ($p < 0.05$: *, $p < 0.01$: **, $p < 0.001$: ***). The RA, FO and CR values relating to 26 amplicon studies were exhibited in the forms of equally-weighted average values ± SD. All the other calculations were expressed in the form of mean ± SD unless indicated otherwise. The boxplots showed the median values and whiskers extending to include all the valid data denoted by Turkey test. All figures showed data from at least three biological replicates.

Declarations

Availability of data and materials

The datasets generated and analyzed in this study are available as the following: Basically, all the descriptive information and data related to 404 hGMB species is available at hGMB homepage (hgmb.nmdc.cn) [30]. The taxonomic descriptions of all new taxa are also accessible at eLMSG under

accessions from MSG071057 to MSG071268 (link type: https://www.biosino.org/elmsg/record/MSG071057) [31]. All the genomes obtained in this study are available at NCBI under Bioproject PRJNA656402 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA656402) [99], NODE with the project accession OEP001106 (https://www.biosino.org/node/project/detail/OEP001106) [100] , and NMDC under Project NMDC10014003 (http://hgmb.nmdc.cn/subject/hgmb/download). The sequences of 16S rRNA genes of all taxa in hGMB are deposited in Genbank under Bioproject PRJNA656402 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA656402) [99], and in NMDC under accessions NMDC10014003 (http://hgmb.nmdc.cn/subject/hgmb) [101]. The other 16S rRNA gene amplicon datasets analyzed in this study were available at NCBI with accessions listed in Table S7. The gene catalog hGMB.all and hGMB.new are deposited at hGMB homepage [30] and NODE under accessions .

Acknowledgements

This work was financially supported by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDB38020300) and National Natural Science Foundation of China (Grant No. 2019YFA0905601).

Author information

Affiliations

State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, No.1 Beichenxi Road, Chaoyang District, Beijing 100101, P. R. China

Chang Liu, Meng-Xuan Du, Rexiding Abuduaini, Hai-Ying Yu, Dan-Hua Li, Yu-Jing Wang, Nan Zhou, Min-Zhi Jiang, Peng-Xia Niu, Shan-Shan Han , Hong-He Chen, Wen-Yu Shi, Linhuan Wu, Yu-Hua Xin, Juncai Ma, Yuguang Zhou, Cheng-Ying Jiang & Shuang-Jiang Liu

University of Chinese Academy of Sciences, Beijing, 100049, P. R. China

Rexiding Abuduaini, Yu-Jing Wang, Hong-Wei Liu & Shuang-Jiang Liu

China General Microorganism Culture Collection, Institute of Microbiology, Chinese Academy of Sciences, No.1 Beichenxi Road, Chaoyang District, Beijing 100101, P. R. China

Yu-Hua Xin & Yuguang Zhou

Microbial Resources and Big Data Center, Institute of Microbiology, Chinese Academy of Sciences, No.1 Beichenxi Road, Chaoyang District, Beijing 100101, P. R. China

Wen-Yu Shi, Linhuan Wu & Juncai Ma

Environmental Microbiology Research Center, Institute of Microbiology, Chinese Academy of Sciences, No.1 Beichenxi Road, Chaoyang District, Beijing 100101, P. R. China

Chang Liu, Dan-Hua Li, Nan Zhou, Peng-Xia Niu, Hong-He Chen, Cheng-Ying Jiang & Shuang-Jiang Liu

State Key Laboratory of Mycology, Institute of Microbiology, Chinese Academy of Sciences, No. 1 Beichenxi Road, Chaoyang District, Beijing 100101, P. R. China.

Hong-Wei Liu

Corresponding Authors

S-J Liu: liusj@im.ac.cn, Tel +86 10 64807423

C Liu: liuc@im.ac.cn, Tel +86 10 64807581

Author contributions

CL, MXD, HYY, DHL, PXN and HHC performed the microbe isolation, cultivation and genome sequencing. RA and YJW performed the characterization of new species. MXD and SSH performed the sample collection and preparation. NZ, WJW, YHX and YGZ conducted the microbial strain preservation. MZJ and RA performed the genome extraction. CL and CYJ conducted the bioinformatic analysis. WYS, LHW and JCM uploaded all the data and constructed the webpage. HWL analyzed the data. CL and SJL designed the studies, analyzed the data, and wrote the manuscript.

Ethics declarations

Ethics approval and consent to participate

Not applicable

Consent for publications

Not applicable

Competing interest

The authors declare no competing interests.

References

1. Sarkar A, Harty S, Lehto SM, Moeller AH, Dinan TG, Dunbar RIM, et al. The Microbiome in Psychology and Cognitive Neuroscience. Trends Cogn Sci. 2018; 22:611-36.
2. Rooks MG, Garrett WS. Gut microbiota, metabolites and host immunity. Nat Rev Immunol. 2016; 16:341-52.

3. Tilg H, Zmora N, Adolph TE, Elinav E. The intestinal microbiota fuelling metabolic inflammation. Nat Rev Immunol. 2020; 20:40-54.
4. Osadchiy V, Martin CR, Mayer EA. The Gut-Brain Axis and the Microbiome: Mechanisms and Clinical Implications. Clin Gastroenterol H. 2019; 17:322-32.
5. Lagier JC, Dubourg G, Million M, Cadoret F, Bilen M, Fenollar F, et al. Culturing the human microbiota and culturomics'. Nat Rev Microbiol. 2018; 16:540-50.
6. Heintz-Buschart A, Wilmes P. Human Gut Microbiome: Function Matters. Trends Microbiol. 2018; 26:563-74.
7. Devkota S. Big data and tiny proteins: shining a light on the dark corners of the gut microbiome. Nat Rev Gastro Hepat. 2020; 17:68-9.
8. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nature Biotechnol. 2020, 10.1038/s41587-020-0603-3.
9. Thomas AM, Segata N. Multiple levels of the unknown in microbiome research. Bmc Biol. 2019; 17.
10. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, et al. A unified sequence catalogue of over 280,000 genomes obtained from the human gut microbiome. bioRxiv. 2019, 10.1101/762682:762682.
11. Peisl BYL, Schymanski EL, Wilmes P. Dark matter in host-microbiome metabolomics: Tackling the unknowns-A review. Anal Chim Acta. 2018; 1037:13-27.
12. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature. 2013; 499:431-7.
13. Mallick H, Ma SY, Franzosa EA, Vatanen T, Morgan XC, Huttenhower C. Experimental design and quantitative analysis of microbial community multiomics. Genome Biol. 2017; 18.
14. Ugarte A, Vicedomini R, Bernardes J, Carbone A. A multi-source domain annotation pipeline for quantitative metagenomic and metatranscriptomic functional profiling. Microbiome. 2018; 6.
15. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 2016; 44:286-93.
16. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, et al. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Res. 2014; 42:643-8.
17. Tramontano M, Andrejev S, Pruteanu M, Klunemann M, Kuhn M, Galardini M, et al. Nutritional preferences of human gut bacteria reveal their metabolic idiosyncrasies. Nat Microbiol. 2018; 3:514-22.
18. Strandwitz P, Kim KH, Terekhova D, Liu JK, Sharma A, Levering J, et al. GABA-modulating bacteria of the human gut microbiota. Nat Microbiol. 2019; 4:396-403.
19. Li LY, Abou-Samra E, Ning ZB, Zhang X, Mayne J, Wang J, et al. An in vitro model maintaining taxon-specific functional activities of the gut microbiome. Nat Commun. 2019; 10.

20. Zou YQ, Xue WB, Luo GW, Deng ZQ, Qin PP, Guo RJ, et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nat Biotechnol. 2019; 37:179-85.
21. Poyet M, Groussin M, Gibbons SM, Avila-Pacheco J, Jiang X, Kearney SM, et al. A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. Nat Med. 2019; 25:1442-52.
22. Vilanova C, Porcar M. Are multi-omics enough? Nat Microbiol. 2016; 1.
23. Browne HP, Forster SC, Anonye BO, Kumar N, Neville BA, Stares MD, et al. Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. Nature. 2016; 533:543-6.
24. Forster SC, Kumar N, Anonye BO, Almeida A, Viciani E, Stares MD, et al. A human gut bacterial genome and culture collection for improved metagenomic analyses. Nat Biotechnol. 2019; 37:186-92.
25. Lagier JC, Khelaifia S, Alou MT, Ndongo S, Dione N, Hugon P, et al. Culture of previously uncultured members of the human gut microbiota by culturomics. Nat Microbiol. 2016; 1.
26. Nayfach S, Shi ZJ, Seshadri R, Pollard KS, Kyrpides NC. New insights from uncultivated genomes of the global human gut microbiome. Nature. 2019; 568:505-10.
27. Murray AE, Freudenstein J, Gribaldo S, Hatzenpichler R, Hugenholtz P, Kampfer P, et al. Roadmap for naming uncultivated Archaea and Bacteria. Nat Microbiol. 2020, 10.1038/s41564-020-0733-x.
28. Young JM. Legitimacy is an essential concept of the International Code of Nomenclature of Prokaryotes - a major revision of the Code is called for. Int J Syst Evol Micr. 2009; 59:1252-7.
29. Liu C, Zhou N, Du MX, Sun YT, Wang K, Wang YJ, et al. The Mouse Gut Microbial Biobank expands the coverage of cultured bacteria. Nat Commun. 2020; 11:79.
30. This_study. hGMB. hgmb.nmdc.cn. Accessed 04 Sep 2020.
31. eLMSG. www.biosino.org/elmsg. Accessed 4 Sep 2020.
32. Durand GA, Fournier PE, Raoult D, Edouard S. 'Bittarella massiliensis' gen. nov., sp. nov. isolated by culturomics from the gut of a healthy 28-year-old man. New Microbes New Infect. 2017; 16:28-9.
33. Afouda P, Traore SI, Dione N, Andrieu C, Tomei E, Richez M, et al. Description and genomic characterization of Massiliimalia massiliensis gen. nov., sp. nov., and Massiliimalia timonensis gen. nov., sp. nov., two new members of the family Ruminococcaceae isolated from the human gut. Anton Leeuw Int J G. 2019; 112:905-18.
34. Ngom, II, Hasni I, Lo CI, Traore SI, Fontanini A, Raoult D, et al. Taxono-genomics and description of Gordonibacter massiliensis sp. nov., a new bacterium isolated from stool of healthy patient. New Microbes New Infect. 2020; 33:100624.
35. Durand GA, Pham T, Ndongo S, Traore SI, Dubourg G, Lagier JC, et al. Blautia massiliensis sp. nov., isolated from a fresh human fecal sample and emended description of the genus Blautia. Anaerobe. 2017; 43:47-55.
36. Togo AH, Diop A, Dubourg G, Khelaifia S, Richez M, Armstrong N, et al. Anaerotruncus massiliensis sp. nov., a succinate-producing bacterium isolated from human stool from an obese patient after

bariatric surgery. New Microbes New Infect. 2019; 29:100508.

37. Bilen M, Founkou MDM, Cadoret F, Dubourg G, Daoud Z, Raoult D. Sanguibacter massiliensis sp. nov., Actinomyces minihominis sp. nov., Clostridium minihomine sp. nov., Neobittarella massiliensis gen. nov. and Miniphocibacter massiliensis gen. nov., new bacterial species isolated by culturomics from human stool samples. New Microbes New Infect. 2018; 24:21-5.
38. Wegmann U, Louis P, Goesmann A, Henrissat B, Duncan SH, Flint HJ. Complete genome of a new Firmicutes species belonging to the dominant human colonic microbiota ('Ruminococcus bicirculans') reveals two chromosomes and a selective capacity to utilize plant glucans. Environ Microbiol. 2014; 16:2879-90.
39. Durand G, Afouda P, Raoult D, Dubourg G. "Intestinimonas massiliensis" sp. nov, a new bacterium isolated from human gut. New Microbes New Infect. 2017; 15:1-2.
40. Fodor AA, DeSantis TZ, Wylie KM, Badger JH, Ye YZ, Hepburn T, et al. The "Most Wanted" Taxa from the Human Microbiome for Whole Genome Sequencing. Plos One. 2012; 7.
41. Li JH, Jia HJ, Cai XH, Zhong HZ, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol. 2014; 32:834-41.
42. Wyman SK, Avila-Herrera A, Nayfach S, Pollard KS. A most wanted list of conserved microbial protein families with no known domains. PloS one. 2018; 13:e0205749.
43. Liu YX, Qin Y, Chen T, Lu MP, Qian XB, Guo XX, et al. A practical guide to amplicon and metagenomic analysis of microbiome data. Protein Cell. 2020, 10.1007/s13238-020-00724-8.
44. Gonzalez-Riano C, Dudzik D, Garcia A, Gil-de-la-Fuente A, Gradillas A, Godzien J, et al. Recent Developments along the Analytical Process for Metabolomics Workflows. Anal Chem. 2020; 92:203-26.
45. Karczewski KJ, Snyder MP. Integrative omics for health and disease. Nat Rev Genet. 2018; 19:299-310.
46. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis (vol 35, pg 833, 2017). Nat Biotechnol. 2017; 35:1211-.
47. Zhang DP, de Souza RF, Anantharaman V, Iyer LM, Aravind L. Polymorphic toxin systems: Comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. Biol Direct. 2012; 7.
48. Goodman AL, Kallstrom G, Faith JJ, Reyes A, Moore A, Dantas G, et al. Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. P Natl Acad Sci USA. 2011; 108:6252-7.
49. Ishiguro E, Haskey N, Campbell K. Gut microbiota : interactive effects on nutrition and health. London, United Kingdom: Academic Press; 2018.
50. Sorbara MT, Littmann ER, Fontana E, Moody TU, Kohout CE, Gjonbalaj M, et al. Functional and Genomic Variation between Human-Derived Isolates of Lachnospiraceae Reveals Inter- and Intra-Species Diversity. Cell Host Microbe. 2020; 28:134-46.

51. Vacca M, Celano G, Calabrese FM, Portincasa P, Gobbetti M, De Angelis M. The Controversial Role of Human Gut Lachnospiraceae. Microorganisms. 2020; 8.
52. Zhang JD, Song LJ, Wang YJ, Liu C, Zhang L, Zhu SW, et al. Beneficial effect of butyrate-producing Lachnospiraceae on stress-induced visceral hypersensitivity in rats. J Gastroen Hepatol. 2019; 34:1368-76.
53. Buffie CG, Bucci V, Stein RR, McKenney PT, Ling LL, Gobourne A, et al. Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. Nature. 2015; 517:205-U7.
54. La Rosa SL, Leth ML, Michalak L, Hansen ME, Pudlo NA, Glowacki R, et al. The human gut Firmicute Roseburia intestinalis is a primary degrader of dietary beta-mannans. Nat Commun. 2019; 10.
55. Reeves AE, Koenigsknecht MJ, Bergin IL, Young VB. Suppression of Clostridium difficile in the Gastrointestinal Tracts of Germfree Mice Inoculated with a Murine Isolate from the Family Lachnospiraceae. Infect Immun. 2012; 80:3786-94.
56. Caballero S, Kim S, Carter RA, Leiner IM, Susac B, Miller L, et al. Cooperating Commensals Restore Colonization Resistance to Vancomycin-Resistant Enterococcus faecium. Cell Host Microbe. 2017; 21:592-602.
57. Shen F, Zheng RD, Sun XQ, Ding WJ, Wang XY, Fan JG. Gut microbiota dysbiosis in patients with non-alcoholic fatty liver disease. Hepatob Pancreat Dis. 2017; 16:375-81.
58. Vaziri ND, Wong J, Pahl M, Piceno YM, Yuan J, DeSantis TZ, et al. Chronic kidney disease alters intestinal microbial flora. Kidney Int. 2013; 83:308-15.
59. Kameyama K, Itoh K. Intestinal colonization by a Lachnospiraceae bacterium contributes to the development of diabetes in obese mice. Microbes Environ. 2014; 29:427-30.
60. Cullender TC, Chassaing B, Janzon A, Kumar K, Muller CE, Werner JJ, et al. Innate and adaptive immunity interact to quench microbiome flagellar motility in the gut. Cell Host Microbe. 2013; 14:571-81.
61. Morotomi M, Nagai F, Watanabe Y. Description of Christensenella minuta gen. nov., sp. nov., isolated from human faeces, which forms a distinct branch in the order Clostridiales, and proposal of Christensenellaceae fam. nov. Int J Syst Evol Microbiol. 2012; 62:144-9.
62. Waters JL, Ley RE. The human gut bacteria Christensenellaceae are widespread, heritable, and associated with health. Bmc Biol. 2019; 17.
63. Aleman JO, Bokulich NA, Swann JR, Walker JM, De Rosa JC, Battaglia T, et al. Fecal microbiota and bile acid interactions with systemic and adipose tissue metabolism in diet-induced weight loss of obese postmenopausal women. J Transl Med. 2018; 16.
64. Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, et al. Human Genetics Shape the Gut Microbiome. Cell. 2014; 159:789-99.
65. Depommier C, Everard A, Druart C, Plovier H, Van Hul M, Vieira-Silva S, et al. Supplementation with Akkermansia muciniphila in overweight and obese human volunteers: a proof-of-concept exploratory study. Nat Med. 2019; 25:1096-+.

66. Sokol H, Pigneur B, Watterlot L, Lakhdari O, Bermudez-Humaran LG, Gratadoux JJ, et al. Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. P Natl Acad Sci USA. 2008; 105:16731-6.
67. Kasahara K, Krautkramer KA, Org E, Romano KA, Kerby RL, Vivas EI, et al. Interactions between Roseburia intestinalis and diet modulate atherogenesis in a murine model. Nat Microbiol. 2018; 3:1461-71.
68. Hill D, Sugrue I, Tobin C, Hill C, Stanton C, Ross RP. The Lactobacillus casei Group: History and Health Related Applications. Front Microbiol. 2018; 9.
69. Nowak A, Paliwoda A, Blasiak J. Anti-proliferative, pro-apoptotic and anti-oxidative activity of Lactobacillus and Bifidobacterium strains: A review of mechanisms and therapeutic perspectives. Crit Rev Food Sci. 2019; 59:3456-67.
70. Seishima J, Iida N, Kitamura K, Yutani M, Wang ZY, Seki A, et al. Gut-derived Enterococcus faecium from ulcerative colitis patients promotes colitis in a genetically susceptible mouse host. Genome Biol. 2019; 20.
71. Titecat M, Wallet F, Vieillard MH, Courcol RJ, Loiez C. Ruminococcus gnavus: An unusual pathogen in septic arthritis. Anaerobe. 2014; 30:159-60.
72. Saha S, Kapoor S, Tariq R, Schuetz AN, Tosh PK, Pardi DS, et al. Increasing antibiotic resistance in Clostridioides difficile: A systematic review and meta-analysis. Anaerobe. 2019; 58:35-46.
73. Shimasaki T, Seekatz A, Bassis C, Rhee Y, Yelin RD, Fogg L, et al. Increased Relative Abundance of Klebsiella pneumoniae Carbapenemase-producing Klebsiella pneumoniae Within the Gut Microbiota Is Associated With Risk of Bloodstream Infection in Long-term Acute Care Hospital Patients. Clin Infect Dis. 2019; 68:2053-9.
74. Wexler HM. Bacteroides: the good, the bad, and the nitty-gritty. Clin Microbiol Rev. 2007; 20:593-621.
75. Sears CL, Geis AL, Housseau F. Bacteroides fragilis subverts mucosal biology: from symbiont to colon carcinogenesis. J Clin Invest. 2014; 124:4166-72.
76. Sun F, Zhang Q, Zhao J, Zhang H, Zhai Q, Chen W. A potential species of next-generation probiotics? The dark and light sides of Bacteroides fragilis in health. Food Res Int. 2019; 126:108590.
77. Qiao S, Bao L, Wang K, Sun S, Liao M, Liu C, et al. Activation of a Specific Gut Bacteroides-Folate-Liver Axis Benefits for the Alleviation of Nonalcoholic Hepatic Steatosis. Cell Rep. 2020; 32:108005.
78. Lagkouvardos I, Pukall R, Abt B, Foesel BU, Meier-Kolthoff JP, Kumar N, et al. The Mouse Intestinal Bacterial Collection (miBC) provides host-specific insight into cultured diversity and functional potential of the gut microbiota. Nat Microbiol. 2016; 1.
79. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016; 33:1870-4.
80. Lee I, Kim YO, Park SC, Chun J. OrthoANI: An improved algorithm and software for calculating average nucleotide identity. Int J Syst Evol Micr. 2016; 66:1100-3.

81. Meier-Kolthoff JP, Auch AF, Klenk HP, Goker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. Bmc Bioinformatics. 2013; 14.
82. Qin QL, Xie BB, Zhang XY, Chen XL, Zhou BC, Zhou JZ, et al. A Proposed Genus Boundary for the Prokaryotes Based on Genomic Insights. J Bacteriol. 2014; 196:2210-5.
83. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. 2012; 19:455-77.
84. Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics. 2007; 23:673-9.
85. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. Bmc Bioinformatics. 2010; 11.
86. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007; 35:3100-8.
87. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST plus : architecture and applications. Bmc Bioinformatics. 2009; 10.
88. Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. Peerj. 2015; 3.
89. Bardou P, Mariette J, Escudie F, Djemiel C, Klopp C. jvenn: an interactive Venn diagram viewer. Bmc Bioinformatics. 2014; 15.
90. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010; 26:2460-1.
91. Dixon P. VEGAN, a package of R functions for community ecology. J Veg Sci. 2003; 14:927-30.
92. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015; 12:59-60.
93. Fu LM, Niu BF, Zhu ZW, Wu ST, Li WZ. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012; 28:3150-2.
94. Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol. 2001; 307:1113-43.
95. Wilson CA, Kreychman J, Gerstein M. Assessing annotation transfer for genomics: Quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol. 2000; 297:233-49.
96. Tian WD, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol 2003; 333:863-82.
97. Kolde R, Kolde MR. Package 'pheatmap'. R Package. 2015; 1:790.
98. Basham B. Graphpad Prism. Biotechnol Softw I J. 1997; 14:14-7.
99. This_study. The data of hGMB deposited in NCBI database under Bioproject PRJNA656402 https://www.ncbi.nlm.nih.gov/bioproject/PRJNA656402. Accessed 4 Sep 2020.

100. NODE. The hGMB data deposited in NODE under Project OEP001106. https://www.biosino.org/node/project/detail/OEP001106. Accessed 4 Sep 2020.
101. NMDC. The hGMB data deposited in NMDC under Bioproject NMDC10014003. http://hgmb.nmdc.cn/subject/hgmb. Accessed 4 Sep 2020.

Tables

Table 1
The protologues of 107 new taxa in hGMB

Taxonomy	Rank	Etymology	Type Designation	CGMCC / KCTC / NBRC Accessions
Tianshiaceae	fam. nov.	Tian.shi'a'ce.ae. N. L. neut. n. *Tianshia*, type genus of the family. -aceae, ending to denote a family,N. L. fem. pl. n. *Tianshiaceae*, family of the genus *Tianshia*	Type genus: *Tianshia*	
Tianshia	gen. nov.	Tian.shi'a N.L. fem. n. *Tianshia*, named in honour of the Chinese medical scientist Tianshi Ye	Type species: *Tianshia hominis*	
Tianshia hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-40^T from human feces	CGMCC 1.32813
Luoshenia	gen. nov.	Luo.shen'ia N.L. fem. n. *Luoshenia*, named after the Chinese Goddess Luoshen	Type species: *Luoshenia tenuis*	
Luoshenia tenuis	sp. nov.	te'nu.is L. masc./fem. adj. *tenuis*, thin, slim, referring to the predicted potential function of the strain in weight-loss.	NSJ-44^T from human feces	CGMCC 1.32817
Feifaniaceae	fam. nov.	*Feifaniaceae* Fei.fa'ni.a N.L. fem. n. *Feifania*,, type genus of the family. -aceae, ending to denote a family. N. L. fem. pl. n. *Feifaniaceae*, family of the genus *Feifania*	Type genus: *Feifania*	
Feifania	gen. nov.	Fei.fa'ni.a N.L. fem. n. *Feifania*, named after Chinese microbiologist Feifan Tang	Type species: *Feifania hominis*	
Feifania hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	BX7^T from human feces	CGMCC 1.32862
Bianqueaceae	fam. nov.	Bian.qu.ea'ce.ae. N. L. neut. n. *Bianquea*, type genus of the family. -aceae, ending to denote a family. N. L. fem. pl. n. *Bianqueaceae*, family of the genus *Bianquea*	Type genus: *Bianquea*	

Taxonomy	Rank	Etymology	Type Designation	CGMCC / KCTC / NBRC Accessions
Bianquea	gen. nov.	Bian.que'a N.L. fem. n. *Bianquea*, named after the Chinese medical scientist Bian Que	Type species: *Bianquea renquensis*	
Bianquea renquensis	sp. nov.	ren.qu'en.sis N.L. masc./fem. adj. *renquensis*, pertaining Renqiu county of China, the birthplace of Chinese medical scientist QueBian	NSJ-32^T from human feces	CGMCC 1.32805
Gehongia	gen. nov.	Ge.hong'ia N.L. fem. n. *Gehongia*, named after Ge Hong (284–364 AD), a Chinese medical scientist	Type species: *Gehongia tenuis*	
Gehongia tenuis	sp. nov.	te'nu.is L. masc./fem. adj. *tenuis*, thin, slim, referring to the predicted potential function of the strain in weight-loss	NSJ-53^T from human feces	CGMCC 1.32829 /KCTC 25141
Changea	gen. nov.	Chang'e.a N.L. fem. n. *Changea*, named after the Chinese Goddess Chang'e	Type species: *Changea tenuis*	
Changea tenuis	sp. nov.	te'nu.is L. masc./fem. adj. *tenuis*, thin, slim, referring to the predicted potential function of the strain in weight-loss	NSJ-63^T from human feces	CGMCC 1.32839
Ligaoa	gen. nov.	Li.gao'a N.L. fem. n. *Ligaoa*, named in honour of the Chinese medical scientist Li Gao	Type species: *Ligaoa zhengdingensis*	
Ligaoa zhengdingensis	sp. nov.	zheng.ding.en'sis N.L. masc./fem. adj. *zhengdingensis*, referring to Zhengding county of China, the birthplace of Li Gao	NSJ-31^T from human feces	CGMCC 1.32804
Congzhengia	gen. nov.	Tsong. zheng'i.a N.L. fem. n. *Congzhengia*, named in honour of the Chinese medical scientist Congzheng Zhang	Type species: *Congzhengia minquanensis*	

Taxonomy	Rank	Etymology	Type Designation	CGMCC / KCTC / NBRC Accessions
Congzhengia minquanensis	sp. nov.	min.quan.en'sis N.L. masc./fem. adj. *minquanensis*, referring to Minquan county of China, the birthplace of Congzheng Zhang	H8^T from human feces	CGMCC 1.32875
Fumia	gen. nov.	Fu.mi'a N.L. fem. n. *Fumia*, named in honour of the Chinese medical scientist Fumi Huang	Type species: *Fumia xinanesis*	
Fumia xinanesis	sp. nov.	xin'an.en'sis N.L. masc./fem. adj. *xinanesis*, referring to Xin'an county where Fumi Huang was born	NSJ-33^T from human feces	CGMCC 1.32806
Wujia	gen. nov.	Wu.ji'a N.L. fem. n. *Wujia*, named after Chinese medical scientist Wuji	Type species: *Wujia chipingensis*	
Wujia chipingensis	sp. nov.	Cheng. chi.ping'en.sis N.L. masc./fem. adj. *chipingensis*, referring to Chiping county of China, the birthplace of the Chinsese medical scientist Wuji Cheng	NSJ-4^T from human feces	CGMCC 1.52560
Simiaoa	gen. nov.	Si.miao.'a. N.L. fem. n. *Simiaoa* named after Sun Simiao, a Chinese medical scientist	Type species: *Simiaoa sunii*	
Simiaoa sunii	sp. nov.	sun'i.i. N.L. gen. n. *sunii*, named after the family name of the Chinese medical scientist Simiao Sun	NSJ-8^T from human feces	CGMCC 1.52840
Simiaoa hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	H15^T from human feces	CGMCC 1.32863
Jutongia hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	BX3^T from human feces	CGMCC 1.32876

Taxonomy	Rank	Etymology	Type Designation	CGMCC / KCTC / NBRC Accessions
Jutongia	gen. nov.	Ju.tong'ia, L. adj. fem., Jutongia, in honor of the Chinsese medical scientist Jutong Wu	Type species: *Shizhenia lianus*	
Jutongia huaianensis	sp. nov.	huai.an'en.sis N.L. masc. adj. huaianensis, huai'an county of China, the birthplace of the Chinsese medical scientist Jutong Wu	NSJ-37^T from human feces	CGMCC 1.32810
Qiania	gen. nov.	qian'i.a N.L. fem. n. *Qiania*, named after Chinese medical scientist Yi Qian	Type species: *Qiania dongpingensis*	
Qiania dongpingensis	sp. nov.	dong.ping' en'sis N.L. masc./fem. adj. *dongpingensis*, referring to Dongping county of China, the birthplace of Yi Qian	NSJ-38^T from human feces	CGMCC 1.32811
Zhenhengia	gen. nov.	Zhen.heng'i.a N.L. fem. n. *Zhenhengia*, named after the Chinese medical scientist Zhenheng Zhu	Type species: *Zhenhengia yiwuensis*	
Zhenhengia yiwuensis	sp. nov.	yi.wu'en'sis N.L. masc./fem. adj. *yiwuensis*, referring to Yiwu city, where Zhenheng Zhu was born	NSJ-12^T from human feces	CGMCC 1.32465 /KCTC 15954
Jingyaoa	gen. nov.	Jing'yao'a N.L. fem. n. *Jingyaoa*, named after Chinese medical scientist Jingyao Zhang.	Type species: *Jingyaoa shaoxingensis*	
Jingyaoa shaoxingensis	sp. nov.	shao.xing'en.sis N.L. masc./fem. adj. *shaoxingensis*, referring to Shaoxing city of China, where Jingyao Zhang was born	NSJ-46^T from human feces	CGMCC 1.32819
Wansuia	gen. nov.	Wan.su'ia, L. adj. fem., *Wansuia*, in honor of the Chinese medical scientist Wansu Liu	Type species: *Wansuia hejianensis*	

Taxonomy	Rank	Etymology	Type Designation	CGMCC / KCTC / NBRC Accessions
Wansuia hejianensis	sp. nov.	he.jian'esis, N.L. masc./fem. adj. *hejianesis*, referring to Hejian county of China, the birthplace of the Chinese medical scientist Wansu Liu	NSJ-29^T from human feces	CGMCC 1.32802 /KCTC 25078
Zhenpiania	gen. nov.	Zhen.pian'ia N.L. fem. n. *Zhenpiania*, named after the Chinese medical scientist Zhenpian Li.	Type species: *Zhenpiania hominis*	
Zhenpiania hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	BX12^T from human feces	CGMCC 1.32877
Lentihominibacter	gen. nov.	Len.ti.homini.bac'ter A.pi.bac'ter L. fem. n. lentus, slow. L. gen. masc. n. bacter, a rod. N.L. masc. n. *Lentihominibacter*, slowly growing rod-shaped bacterium	Type species: *Lentihominibacter hominis*	
Lentihominibacter hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-24^T from human feces	CGMCC 1.32874
Lentihominibacter_faecis	sp. nov.	L. gen. fem. n. *faecis*, of faeces, from which the organism was isolated	BX16^T from human feces	CGMCC 1.32878
Yanshouia hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	BX1^T from human feces	CGMCC 1.32879
Shuzhengia	gen. nov.	Shu.zheng'ia N.L. fem. n. *Shuzhengia*, named after Chinese microbiologist Shuzheng Zhang	Type species: *Shuzhengia hominis*	
Shuzhengia hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	BX18^T from human feces	CGMCC 1.32880
Anaerofilum hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	BX8^T from human feces	CGMCC 1.32881

Taxonomy	Rank	Etymology	Type Designation	CGMCC / KCTC / NBRC Accessions
Zongyangia	gen. nov.	Zong.yang'ia. N.L. fem. n. *Zongyangia*, named after the Chinese medical scientist Zongyang Yang	Type species: *Zongyangia hominis*	
Zongyangia hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-54^T from human feces	CGMCC 1.32830 /KCTC 25132
Youxingia	gen. nov.	You.xing'ia N.L. fem. n. *Youxingia*, named after Chinese medical scientist Youxing Wu	Type species: *Youxingia wuxianesis*	
Youxingia wuxianesis	sp. nov.	wu.xian.en'sis N.L. masc./fem. adj. *wuxianesis*, referring to the Yutian county of China, where Youxing Wu was born	NSJ-64^T from human feces	CGMCC 1.32840 /KCTC 25128
Qingrenia	gen. nov.	Qing.re'ni.a N.L. fem. n. *Qingrenia*, named after Chinese medical scientist Qingren Wang	Type species: *Qingrenia yutianesis*	
Qingrenia yutianesis	sp. nov.	yu.tian.en'sis N.L. masc./fem. adj. *yutianesis*, Yutian county of China, where Qingren Wang was born	NSJ-50^T from human feces	CGMCC 1.32823
Jilunia	gen. nov.	Ji.lun'ia N.L. fem. n. *Jilunia*, named after Chinese microbiologist Jilun Li	Type species: *Jilunia laotingensis*	
Jilunia laotingensis	sp. nov.	lao.ting.en'sis N.L. masc./fem. adj. *laotingensis*, referring to the Laoting county where Jilun Li was born	N12^T from human feces	CGMCC 1.32860
Paratisierella	gen. nov.	Para'tissierella, Gr. prep. para, beside. N.L. fem. dim. n. *Tissierella*, a genus name. N.L. masc. n. *Paratissierella*, resembling the genus Tissierella	Type species: *Paratisierella segnis*	
Paratisierella segnis	sp. nov.	L. masc. adj. *segnis*, slow, inactive, lazy, referring the slow growth of the strain	BX21^T from human feces	CGMCC 1.32882

Taxonomy	Rank	Etymology	Type Designation	CGMCC / KCTC / NBRC Accessions
Bittarella	gen. nov.	N.L. fem. dim. n. *Bittarella*, in honour of Dr Bittar, a French microbiologist	Type species: *Bittarella massiliensis*	
Bittarella massiliensis	sp. nov.	N.L. fem. dim. n. Bittarella. mas.sil.i.en'sis L. masc./fem. adj. *massiliensis*, of Massilia, the Latin name of Marseille where the strain was for the first time isolated, and *Bittarella massiliensis* is the type species of the genus *Bittarella* [32]	NSJ-19^T from human feces	CGMCC 1.32824 /KCTC 25133
Massiliimalia timonensis	sp. nov.	ti.mo.nen'sis N.L. masc./fem. adj. *timonensis*, from 'Timone, ' the name of the main hospital of Marseille, France, where the type strain was isolated [33]	NSJ-15^T from human feces	CGMCC 1.32466 /KCTC 15951
Eggerthella hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-70^T from human feces	CGMCC 1.32846 /KCTC 25139
Gordonibacter massiliensis	sp. nov.	ma.si.li.e'n.sis. L. adj. masc. *massiliensis*, of Massilia, Marseilli, where the bacteria was for the first time isolated [34]	NSJ-58^T from human feces	CGMCC 1.32834 /KCTC 25146
Bacteroides brevis	sp. nov.	L. gen. neut. n. *brevis*, short, denoting the formation of short rods	NSJ-39^T from human feces	CGMCC 1.32812
Bacteroides multiformis	sp. nov.	mul.ti.for'mis L. masc./fem. adj. *multiformis*, many-shaped, multiform,referring to the various size and shape of the strain)	L5^T from human feces	CGMCC 1.32865
Bacteroides facilis	sp. nov.	L. masc. adj. *facilis*, easy, referring that the type strain is easily cultured	NSJ-77^T from human feces	CGMCC 1.32853 /KCTC 25155
Bacteroides celeris	sp. nov.	ce'le.ris L. fem. adj. *celeris*, rapid, pertaining to fast growth of the strain	NSJ-48^T from human feces	CGMCC 1.32821

Taxonomy	Rank	Etymology	Type Designation	CGMCC / KCTC / NBRC Accessions
Bacteroides difficilis	sp. nov.	diff'i.cil.is L. masc. adj. *difficilis*, difficult, referring the difficulty of culturing the strain	NSJ-74^T from human feces	CGMCC 1.32850
Bacteroides hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-2^T from human feces	CGMCC 1.31481 /KCTC 15964
Bacteroides parvus	sp. nov.	par' vus. L. masc. adj. *parvus*, small, referring that its colonies on MGAM agar media are significantly small.	NSJ-21^T from human feces	CGMCC 1.31612 /KCTC 25073
Barnesiella faecis	sp. nov.	L. gen. fem. n. *faecis*, of faeces, from which the organism was isolated	BX6^T from human feces	CGMCC 1.32883
Butyricimonas hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-56^T from human feces	CGMCC 1.32832
Parabacteroides acidifaciens	sp. nov.	a.ci.di.fa'ci.ens L. neut. n. acidum, acid; L. v. facio, to produce; N.L. part. adj. *acidifaciens*, acid-producing	426-9^T from human feces	CGMCC 1.13558 /NBRC 113433
Parabacteroides segnis	sp. nov.	L. masc. adj. *segnis*, slow, inactive, lazy, referring the slow growth of the strain	BX2^T from human feces	CGMCC 1.32884
Parabacteroides hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-79^T from human feces	CGMCC 1.32855 /KCTC 25129
Alistipes hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	New-7^T from human feces	CGMCC 1.31637 /KCTC 15866
Ornithinibacillus hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	BX22^T from human feces	CGMCC 1.32885

Taxonomy	Rank	Etymology	Type Designation	CGMCC / KCTC / NBRC Accessions
Streptococcus hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-17^T from human feces	CGMCC 1.32470 /KCTC 15949
Streptococcus lentus	sp. nov.	L. masc. adj. *lentus*, slow, referring the slow growth of the strain	NSJ-72^T from human feces	CGMCC 1.32848
Christensenella tenuis	sp. nov.	te'nu.is L. masc./fem. adj. *tenuis*, thin, slim, referring to the predicted potential function of the strain in weight-loss	NSJ-35^T from human feces	CGMCC 1.32808
Clostridium hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-6^T from human feces	CGMCC 1.32461 /KCTC 15960
Clostridium beijingense	sp. nov.	bei.jing.en'se N.L. neut. adj. *beijingense*, from Beijing, where the type strain was isolated	NSJ-49^T from human feces	CGMCC 1.32822
Clostridium lentus	sp. nov.	L. masc. adj. *lentus*, slow, referring to the slow growth of the type strain	NSJ-42^T from human feces	CGMCC 1.32815
Clostridium facilis	sp. nov.	L. masc. adj. *facilis*, easy, without difficulty, referring that the type strain is easily cultured	NSJ-27^T from human feces	CGMCC 1.32800
Anaerosacchariphilus hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-68^T from human feces	CGMCC 1.32844 /KCTC 25150
Anaerostipes hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-7^T from human feces	CGMCC 1.32462 /KCTC 15959
Blautia massiliensis	sp. nov.	ma.si.li.e'n.sis. L. adj. masc. *massiliensis*, of Massilia, the Latin name of Marseill, where the bacteria was for the first time isolated [35]	4-46^T from human feces	CGMCC 1.52830 /NBRC 113773

Taxonomy	Rank	Etymology	Type Designation	CGMCC / KCTC / NBRC Accessions
Blautia intestinalis	sp. nov.	in.tes.ti.na'lis N.L. fem. adj. *intestinalis*, pertaining to the intestines where the type strain inhabits	27-44^T from human feces	CGMCC 1.52850 /NBRC 113774
Blautia segnis	sp. nov.	L. masc. adj. *segnis*, slow, inactive, lazy, referring the slow growth of the strain	BX17^T from human feces	CGMCC 1.32886
Blautia tardus	sp. nov.	tar'dus L. masc. adj. *tardus*, slow, inactive, lazy, referring the slow growth of the strain	BX19^T from human feces	CGMCC 1.32887
Blautia celeris	sp. nov.	ce'le.ris L. fem. adj. *celeris*, rapid, pertaining to fast growth of the strain	NSJ-34^T from human feces	CGMCC 1.32807
Blautia lentus	sp. nov.	L. masc. adj. *lentus*, slow, referring to the slow growth of the type strain	M16^T from human feces	CGMCC 1.32888
Blautia difficilis	sp. nov.	diff'i.cil.is L. masc. adj. *difficilis*, difficult, referring the difficulty of culturing the strain	M29^T from human feces	CGMCC 1.32889
Clostridium segnis	sp. nov.	L. masc. adj. *segnis*, slow, inactive, lazy, referring the slow growth of the strain	BX14^T from human feces	CGMCC 1.32890
Coprococcus hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-10^T from human feces	CGMCC 1.32463
Dorea hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-36^T from human feces	CGMCC 1.32809
Enterocloster hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	BX10^T from human feces	CGMCC 1.32891
Eubacterium segnis	sp. nov.	L. masc. adj. *segnis*, slow, inactive, lazy, referring the slow growth of the strain	BX4^T from human feces	CGMCC 1.32892

Taxonomy	Rank	Etymology	Type Designation	CGMCC / KCTC / NBRC Accessions
Eubacterium difficilis	sp. nov.	diff'i.cil.is L. masc. adj. *difficilis*, difficult, referring the difficulty of culturing the strain	M5^T from human feces	CGMCC 1.32893
Hungatella hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-66^T from human feces	CGMCC 1.32842 /KCTC 25127
Hungatella faecis	sp. nov.	L. gen. fem. n. *faecis*, of faeces, from which the organism was isolated	L36^T from human feces	CGMCC 1.32864
Lachnospira hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-43^T from human feces	CGMCC 1.32816
Ruminococcus hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-13^T from human feces	CGMCC 1.52490
Mediterraneibacter hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-55^T from human feces	CGMCC 1.32831
Ruminococcus difficilis	sp. nov.	diff'i.cil.is L. masc. adj. difficilis, difficult, referring the difficulty of culturing the strain	M6^T from human feces	CGMCC 1.32867
Roseburia lentus	sp. nov.	L. masc. adj. lentus, slow, referring to the slow growth of the type strain	NSJ-9^T from human feces	CGMCC 1.32469
Roseburia wangyiboensis	sp. nov.	wang'yi'boensis, N.L.m Wang Yibo, a Chinese actor whose series inspired the researcher during the bacterial identification	BX0805^T from human feces	CGMCC 1.32827
Roseburia xiaozhanis	sp. nov.	xi'ao'zhan'is N.L. fem. n. Xiao Zhan, a chinese actor whose series inspired the researcher during the bacterial identifications	BX1005^T from human feces	CGMCC 1.32828

Taxonomy	Rank	Etymology	Type Designation	CGMCC / KCTC / NBRC Accessions
Roseburia rectibacter	sp. nov.	L. adj. rectus, straight; N.L. masc. n. bacter, rod; N.L. masc. n. rectibacter, straight rod shaped, referring to the cell shape of the strain	NSJ-69^T from human feces	CGMCC 1.32845
Roseburia difficilis	sp. nov.	diff'i.cil.is L. masc. adj. difficilis, difficult, referring the difficulty of culturing the strain	NSJ-67^T from human feces	CGMCC 1.32843
Agathobaculum hominis	sp. nov.	ho'mi.nis L. gen. masc. n. hominis, of a human being, referring to the human gut habitat	M2^T from human feces	CGMCC 1.32866
Agathobaculum faecis	sp. nov.	L. gen. fem. n. *faecis*, of faeces, from which the organism was isolated	NSJ-28^T from human feces	CGMCC 1.32801
Anaerotruncus massiliensis	sp. nov.	mas.si.li.en'sis L. masc./fem. adj. massiliensis, pertaining to Marseille, France, where the organism was for the first time isolated [36]	22A2-44^T from human feces	CGMCC 1.52380 /NBRC 113434
Dysosmobacter segnis	sp. nov.	L. masc. adj. segnis, slow, inactive, lazy, referring the slow growth of the strain	BX15^T from human feces	CGMCC 1.32894
Dysosmobacter hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-60^T from human feces	CGMCC 1.32836 /KCTC 25148
Faecalibacterium hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	4P15^T from human feces	CGMCC 1.52500 /NBRC 113913
Flintibacter faecis	sp. nov.	L. gen. fem. n. *faecis*, of faeces, from which the organism was isolated	BX5^T from human feces	CGMCC 1.32861
Flintibacter hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	New-19^T from human feces	CGMCC 1.31644 /KCTC 15861

Taxonomy	Rank	Etymology	Type Designation	CGMCC / KCTC / NBRC Accessions
lawsonibacter hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-51^T from human feces	CGMCC 1.32825 /KCTC 25134
lawsonibacter faecis	sp. nov.	L. gen. fem. n. *faecis*, of faeces, from which the organism was isolated	NSJ-52^T from human feces	CGMCC 1.32826 /KCTC 25135
Lawsonibacter celeris	sp. nov.	ce'le.ris L. fem. adj. *celeris*, rapid, pertaining to fast growth of the strain	NSJ-47^T from human feces	CGMCC 1.32820
Neobittarella	gen. nov	Neo.bita.rel'la, L. adj. fem., *Neobittarella*, in honor of microbiologist Fadi Bittar	Type species: *Neobittarella massiliensis*	
Neobittarella massiliensis	sp. nov.	mas.si.li.en'sis L. masc./fem. adj. massiliensis, referring to Marseille, where the organism was isolated [37]	NSJ-65^T from human feces	CGMCC 1.32841 /KCTC 25131
Oscillibacter hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-62^T from human feces	CGMCC 1.32838 /KCTC 25149
Pseudoflavonifractor hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	New-38^T from human feces	CGMCC 1.31611 /KCTC 15862
Ruminococcus bicirculans	sp. nov.	bai.circu'lans L. masc. adj. have two circles, referring the cell shapes of the type strain [38]	NSJ-14^T from human feces	CGMCC 1.52640 /KCTC 15952
Ruminococcus intestinalis	sp. nov.	in.tes.ti.na'lis N.L. fem. adj. intestinalis, pertaining to the intestine habitat	NSJ-71^T from human feces	CGMCC 1.32847
Paeniclostridium hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-45^T from human feces	CGMCC 1.32818
Romboutsia hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-18^T from human feces	CGMCC 1.31399

Taxonomy	Rank	Etymology	Type Designation	CGMCC / KCTC / NBRC Accessions
Intestinimonas massiliensis	sp. nov.	ma.si.li.e'n.sis. L. adj. masc. *massiliensis*, of Massilia, the Latin name of Marseill, where the bacteria was for the first time isolated [39]	NSJ-30^T from human feces	CGMCC 1.32803 /KCTC 25082
Hydrogeniiclostidium hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-41^T from human feces	CGMCC 1.32814 /KCTC 25093
Catenibacterium faecis	sp. nov.	L. gen. fem. n. *faecis*, of faeces, from which the organism was isolated	NSJ-22^T from human feces	CGMCC 1.31663
Eubacterium hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	New-5^T from human feces	CGMCC 1.32837 /KCTC 15860
Holdemanella hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	L34^T from human feces	CGMCC 1.32895
Megasphaera hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-59^T from human feces	CGMCC 1.32835 /KCTC 25147
Veillonella hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-78^T from human feces	CGMCC 1.32854 /KCTC 25159
Tissierella hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-26^T from human feces	CGMCC 1.31394 /KCTC 25080
Fusobacterium hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-57^T from human feces	CGMCC 1.32833
Escherichia hominis	sp. nov.	ho'mi.nis L. gen. masc. n. *hominis*, of a human being, referring to the human gut habitat	NSJ-73^T from human feces	CGMCC 1.32849

Figures

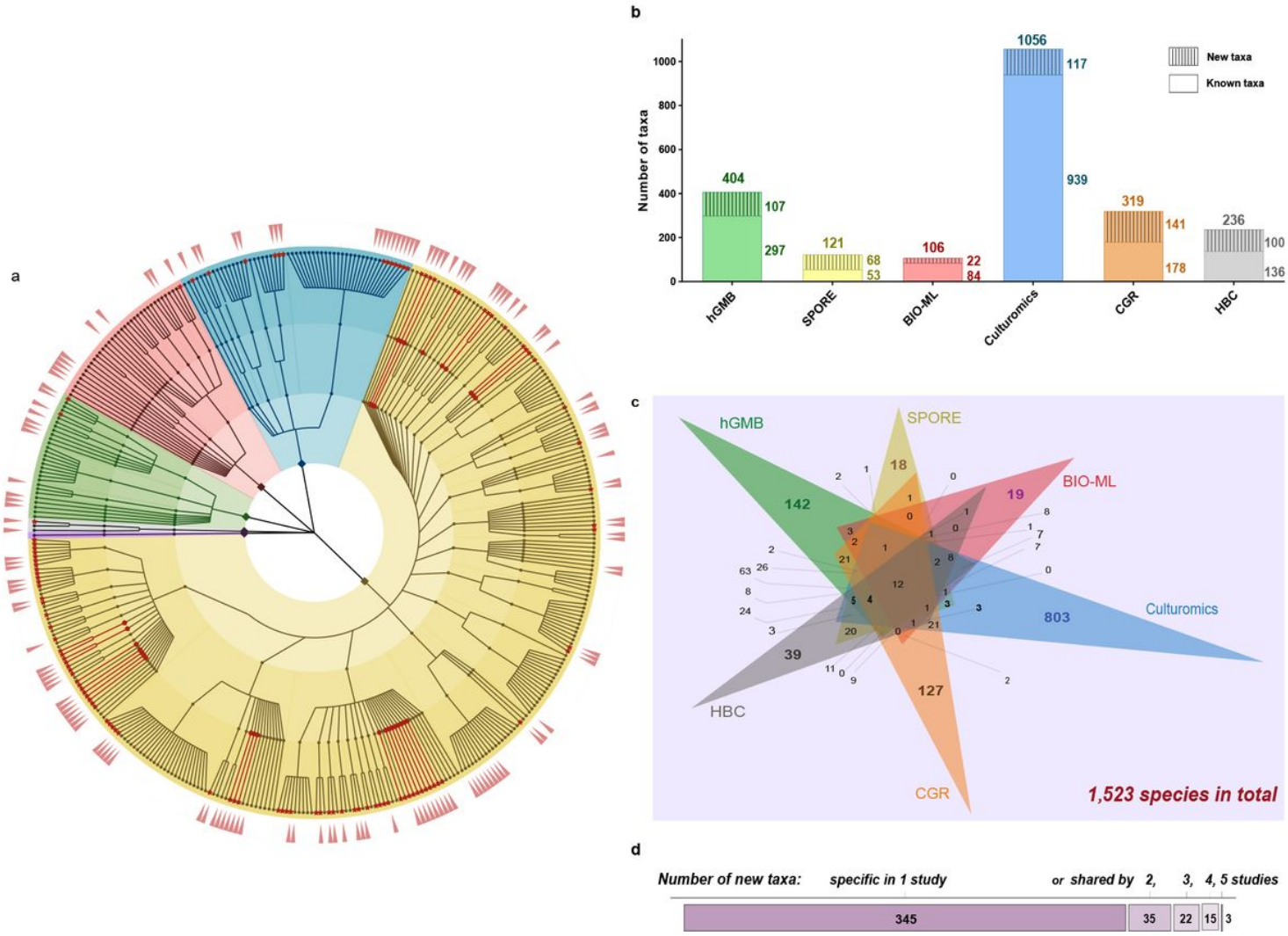

Figure 1

The taxonomic diversity and specificity of hGMB. (a) The taxonomic cladogram displaying the taxonomic diversity of hGMB. The nodes of 107 newly-characterized species are shown with red stars, and the nodes of 28 novel genus and 3 novel family are indicated with red circles. The background is color-coded according to 6 phyla, Yellow: Firmicutes, Blue: Bacteroides, Red: Proteobacteria, Green: Actinobacteria, Grey: Fusobacteria, Purple: Verrucomicrobia. The outer ring (the coral red pointers) shows the unique 142 species that are solely covered by hGMB. (b) The taxonomic diversity of gut microbes from different gut microbial collections. hGMB (this study): a culture collection constructed in this study contains 404 species with 107 new taxa; SPORE [23]: a culture collection constructed in 2016 comprises 121 species with 68 new-taxon candidates by now; BIO-ML [21]: a culture collection constructed in 2019 comprises 106 species with 20 new-taxon candidates; Culturomics [25]: the culturomics study in 2016 reveal the discovery of 1,056 species including 247 new taxa, of which 117 were still new by now; CGR [20]: a culture collection constructed in 2019 comprises 319 species based on the 16S rRNA gene sequence clustering, of which 141 taxa are potentially novel; HBC [24]: a culture collection constructed in 2019 contains 236 species with 100 potentially-new taxa. (c) The Venn diagram displaying the unique and shared taxa by each study. The numbers of taxa uniquely in one collection or shared by different studies

are labeled in the panel. (d) Summary of new taxa claimed by 1 or more than 1 study. Numbers in the bar represent the number of new taxa.

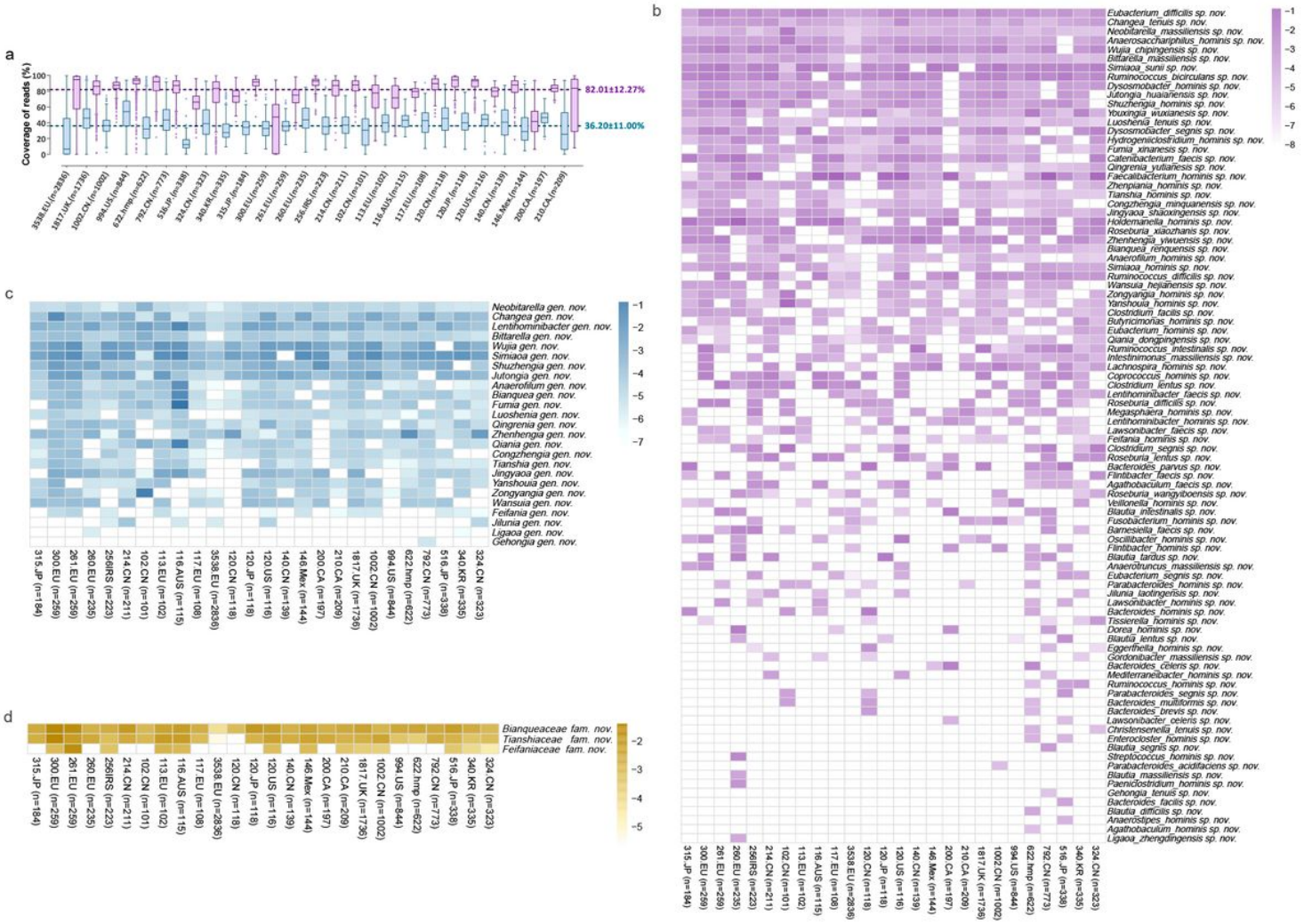

Figure 2

The prevalence of hGMB novel and total taxa in global human gut microbiota datasets. Each study was named in forms of “original sample size. nation abbreviation of study”. The quality-controlled sample size of each study used in the analysis are listed below the x-axis study name as “(n=)”. More information about each study is available at Table S7. (a) The coverage of 26 datasets by the 107 new taxa (blue bars) and 404 total taxa (purple bars) from hGMB at reads level (sequence identity >97%). The box-and-whiskers plot shows the average coverage rate (CR) of total reads for 26 studies, center line: median, bounds of box: quartile, whiskers: Tukey extreme; The equally-weighted average CR in the form of Mean± SD are listed in the panel. The heatmaps display the distributions and relative abundances of novel taxa at species (b), genus (c) and families (d) levels in the 26 studies. The depth of colors is positively correlated to the Log10 of the equally-weighted average RAs, the white-color grid indicated the absence of corresponding taxa.

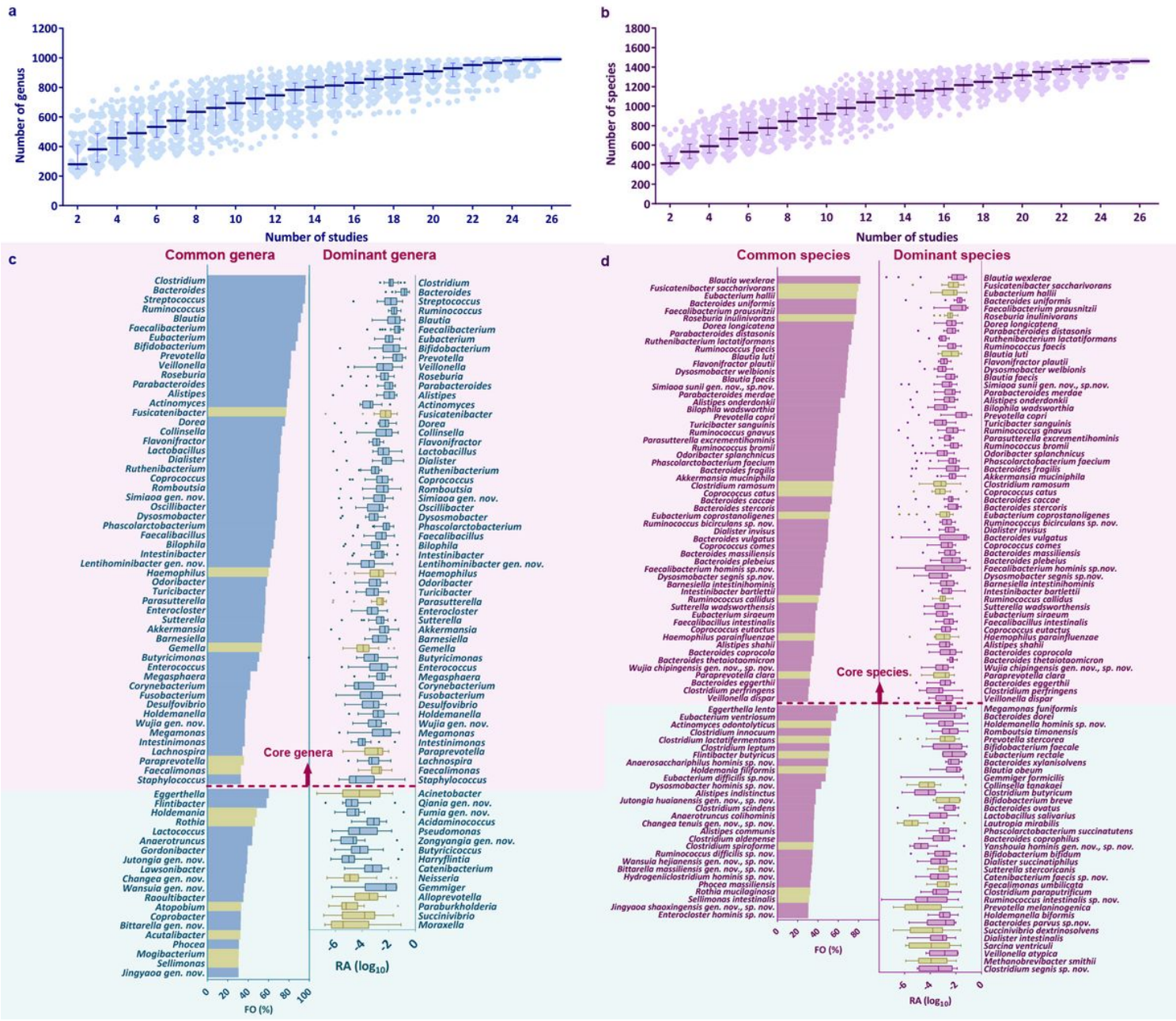

Figure 3

The hGMB largely represents human gut microbiota at genus and species levels. (a) and (b): The rarefaction curves at genus (a) and species (b) levels, as calculated from datasets of 26 studies (Table S7). (c): The coverage of human gut common and dominant genera by hGMB. The hGMB genera were colored in blue. (d) The coverage of human gut common and dominant species by hGMB. The hGMB genera are colored in purple. Notes: 1) FO=100% is defined when a taxon presents in all samples, while FO=0 is defined when a taxon presents in none of the samples; The equally-weighted average FO is calculated by averaging the 26 average FOs of each study; 2) The equally-weighted average RA is calculated by averaging the average RAs of the 26 studies). The light pink background in panel c and d highlight the genera/species shared by both common and dominant groups, while the light blue background marks out the taxa presenting uniquely in either group. The bar chart in panel b and c shows

the mean values of the 26 FO averages (%), while the box-and-whiskers plot shows the 26 average RAs of each taxon, center line: median, bounds of box: quartile, whiskers: Tukey extreme. “Common taxa” is defined as equally-weighted average FOs>30%; “Dominant taxa” is defined as equally-weighted average RAs> 0.1%; and “Core taxa” is defined equally-weighted average FOs>30% and equally-weighted average RAs> 0.1%.

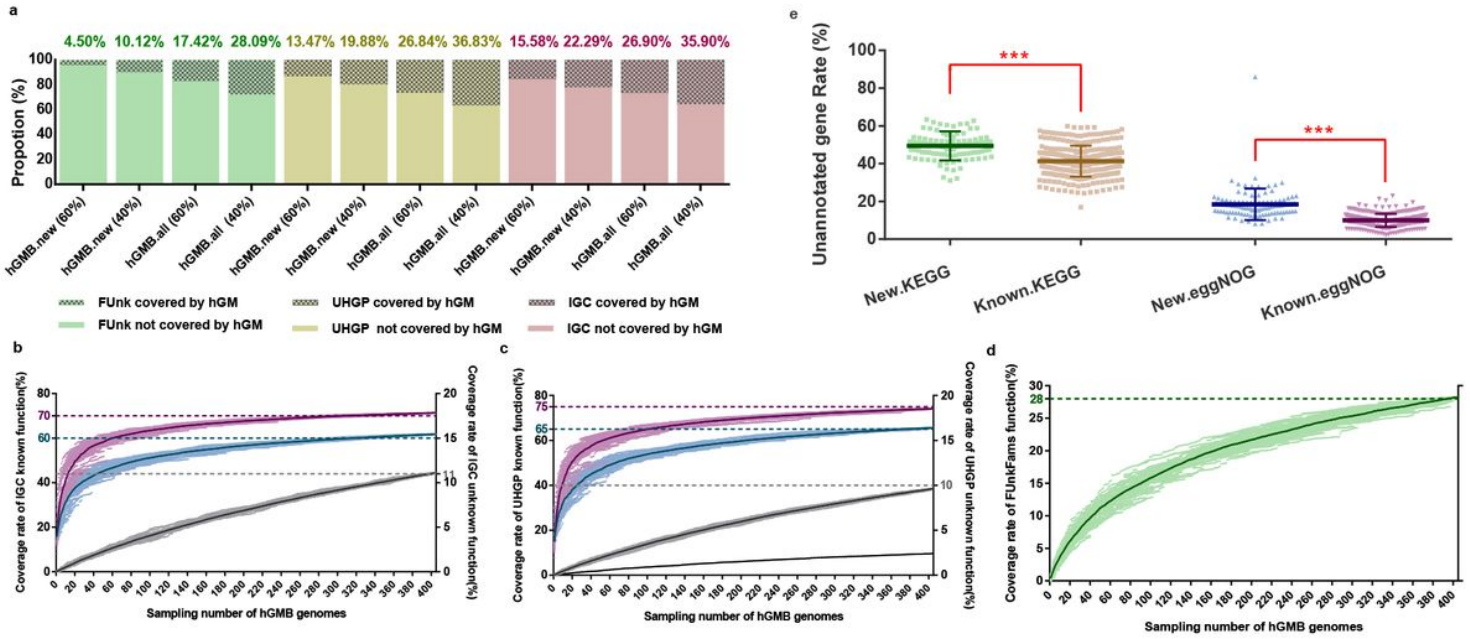

Figure 4

The functional coverage of global human gut microbiomes by hGMB and the un-annotated gene rates of new or known bacterial taxa. (a) The coverage of IGC, UHGP and FUnkFams by hGMB.all and hGMB.new. The cut-off values of sequence identities during BLAST are listed in the panel follwoing the names of the hGMB gene catalogs in x-axis. The coverage rates (CRs) were listed in panel on the top of each bar. (b),(c) The rarefaction curves displaying the accumulative coverage of the annotated KOs (purple), GOs (blue) and unannotated genes (grey) in IGC (b) and UHGP (c) catalogs. The sampling was repeated for 50 times at each x-axis point; Light purple dot: the CRs of KO functions of IGC or UHGP gene catalogs when specified numbers of genomes were randomly sampled from 404 hGMB genomes; Dark purple line: the mean CR of KO functions; Light blue dot: the CRs of GO functions KO functions of IGC or UHGP gene catalogs; Dark blue line: the mean CR of GO functions; Grey dot: the CRs of unannotated genes of IGC or UHGP; Black line: the mean CR of unannotated genes of IGC or UHGP. (d) The rarefaction curves displaying the accumulative coverage of conserved unknown proteins in FUnkFams. The sampling was repeated for 50 times at each x-axis point; Light green dot: the CRs of FUnkFams proteins when sampled randomly; Dark green line: the mean value of the CRs. (e) The unannotated gene rates of genomes from new taxa and known taxa in hGMB. New.KEGG: the genomes of 107 new taxa from hGMB annotated with KEGG; Known.KEGG: the genomes of 256 known hGMB taxa annotated with KEGG; New.eggNOG: the genomes of 107 new taxa from hGMB annotated with eggNOG; Known.eggNOG: the genomes of 256 known hGMB taxa annotated with eggNOG.

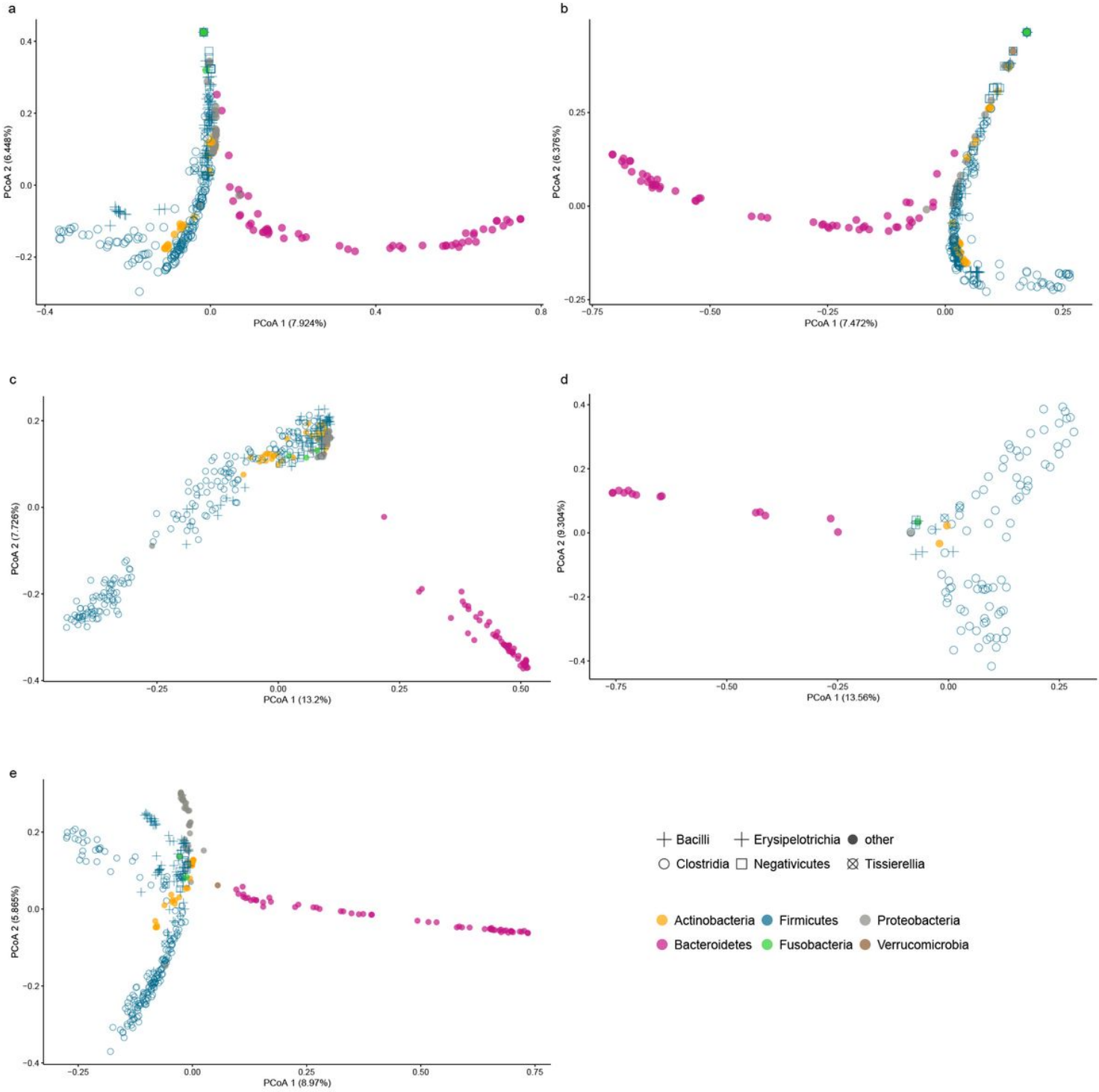

Figure 5

The PCoA analysis depicting the taxonomy-associated distribution of functionally-unknown genes in each hGMB member. (a)-(c) The PCoA displaying the distribution of functionally-unknown genes of IGC (a), UHGP (b) and FUnkFams (c) covered in hGMB members. (d) and (e) The PCoA displaying the distribution of functionally-unknown genes of hGMB.new (d) and hGMB.all (e) in each hGMB genome. As indicated in the panel, the phyla were distinguished by different colors, and the orders in Firmicutes were labeled as different symbol shapes.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMethods.docx
- SupplementaryData1.doc
- TableS11.txt
- TableS10.txt
- TableS9.txt
- TableS8.xlsx
- TableS7.xlsx
- TableS6.xlsx
- TableS5.xlsx
- TableS4.xlsx
- Supplementaryfiles.docx
- TableS3.xlsx
- TableS2.xlsx
- TableS1.xlsx
- FigureS4.pheatmap.BC.pdf
- FigureS3.pheatmap.FUnksort.pdf
- FigureS2.pheatmap.UHGPsort.pdf
- FigureS1.pheatmap.IGCsort.pdf