Enlightening the Taxonomy Darkness of Human Gut Microbiomes With a Cultured Biobank

Chang Liu
Institute of Microbiology Chinese Academy of Sciences
Mengxuan 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, novel taxa, hGMB

Posted Date: February 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-168279/v1

Abstract

Background

The cultivated gut microbial resource plays essential roles in gut microbiome studies such as unraveling gut microbial functions and host-microbe interactions. Though several major studies have been performed to understand the cultured human gut microbiota, up to 70% of the Unified Human Gastrointestinal Genome species remain uncultivated. Large-scale gut microbial isolation and identification and their access to public are imperative for gut microbial studies and further understanding of the human gut microbial functions.

Results

Here, we report the construction of a human Gut Microbial Biobank (hGMB) (homepage: hgmb.nmdc.cn) by cultivation of 10,558 isolates from 239 feces samples of healthy Chinese volunteers, and deposited 1,170 strains representing 400 different species in culture collections of International Depository Authority for long-term preservation and public access worldwide. The hGMB enriched the existing cultivable gut microbial repository and represented over 80% of the common and dominant human gut microbial genera and species of global human gut 16S rRNA gene amplicon data (n=11,647).Moreover, 102 new species were characterized and denominated and 28 new genera and 3 new families were proposed, following the rules of International Code of Nomenclature of Prokaryotes. The hGMB uncovered 24 “most-wanted” and “medium priority” taxa proposed by the Human Microbiome Project, while the novel-taxon genomes represented 22 previously-uncultured species in Unified Human Gastrointestinal Genome (UHGG) and contributed 24 potentially “dark-taxon” representatives that were not discovered by UHGG. The 115 newly-sequenced hGMB genomes covered over 50% of the known genes (KEGG Orthologs) in the global human gut gene catalogs and over 10% of the “most-wanted” functionally unknown proteins in FUnkFams database.

Conclusions

A publicly accessible human Gut Microbial Biobank (hGMB) is established and contains 1,170 strains and represents 400 human gut microbial species. The hGMB expands gut microbial resources and genomic repository by adding 102 novel species, 28 new genera and 3 new families, and 115 new genomes of human gut microbes.

Introduction

The gut microbiome (GM) is recognized to be crucial to host’s physical and mental health [1]. When GM dysbiosis happens, it often triggers host’s immune dysfunction [2], metabolic disorder [3], and impaired cognitive and physiological developments [4]. Both culture-dependent and -independent studies have delivered unprecedent 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 [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 develop bioinformatic tools and databases [13-16]. However, functional characterization and verification at biological and molecular levels still rely on culture-based

experiments. Cultured microbial resources that harbor unknown genes of interest and/or produce specific metabolites are indispensable. Furthermore, 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]. Enlightening the “dark matters” of GMs needs extensive efforts on microbial cultivation as well as physiological and genetic characterizations.

During the past few years, several large-scale cultivations have been made [20, 21, 23-28], and there were in total 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, 29], with the majority of gut microbes remained uncultured. The validly taxonomic description and nomination of newly-cultured gut microorganisms, on the other hand, even lag behind [30]. For example, the Culturomics reported 247 novel taxa in 2016 [25], while 117 out of the 247 novel taxa remain unclassified till the time of this writing, as their taxonomic descriptions and nomenclatures did not fulfill the rules of International Code of Nomenclature of Prokaryotes (ICNP) [31]. In some other studies [20, 23, 24, 26], the taxonomic characterization and nomenclature were absent. For those newly-isolated taxa without a valid description, their taxonomic names could not be validly approved by International Committee on Systematics of Prokaryotes (ICSP) even though they were effectively published [30, 31], and their taxonomic information together with the 16S rRNA gene sequences could not be included by authoritative 16S rRNA gene sequence databases widely used for valid taxonomy classification as the EZBioCloud [32], NCBI [33] and SILVA All-species Living Tree Project (LTP) [16]. Consequently, some microbes were repeatedly claimed to be novel in different cultivation-based studies. Example is that 54 microbial taxa firstly cultivated in 2016 [23] were still considered as novel taxa in works of 2019 [20, 21] (Table S1). The lack of valid taxonomic description and nomination of the newly-cultivated taxa complicated the scientific discourse of new microbes among researchers, and impeded the accession and exchange of bacterial materials among scientific communities worldwide [34, 35]. The timely characterization and nomenclature of new bacterial isolates is imperative and of practical meanings.

In this study, we cultivated 10,558 bacterial isolates that represented 400 gut microbial species from 239 fecal samples of health donors by large-scale cultivation, and deposited 1,170 representative strains to culture collections of International Depository Authority (IDA) for public access globally. The hGMB largely represented the taxonomic composition of human gut microbial community. We sequenced 115 new bacterial genomes and denominated 102 new bacterial taxa. Data analysis revealed that the newly-identified taxa are prevalent among global human gut microbiome and illuminated number of taxonomic “dark matters”.

Results

The construction of hGMB by large-scale bacterial cultivation and characterization

In total, 239 fresh feces samples of healthy Chinese volunteers (For gender and age, see Table S2) were collected and used for large-scale gut microbial isolation and cultivation, by following a previously established workflow [36] and using 11 pretreatment methods and 67 different culture conditions (Table S3 and S4). Single colonies on agar plates were collected and sequenced for 16S rRNA genes (>1.4 kb). We in total harvested over 18,560 colonies, and 10,558 pure cultures were obtained (culture IDs and full-length 16S rRNA gene sequences are available in Table S5). The taxonomy of these cultures was 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 400 potential taxa at species level by clustering with 16S rRNA gene sequence identity threshold of 98.7%, and 108 taxa were novel taxon candidates (Figure S1). We sequenced 115 genomes representing the 108 novel-taxon candidates and other 7 previously described taxa but no corresponding genomes found in NCBI database. The 115 new genomes, including 10 complete genomes and 105 draft genomes, were achieved and they are publicly-accessible via public databases as NCBI and China National Microbiology Data Center (NMDC) (see Data Availability). The assembly quality of each genome was evaluated and exhibited in Table S6. All 115 genomes had >50% completeness, <5% contamination and an estimated quality score (completeness – 5 × contamination) > 50. Then, 108 novel taxon candidates were taxonomically characterized according to (1) phylogenetic analysis, (2) morphology observation, (3) BIOLOG tests, and (4) genome features. Their taxonomic status was finally determined based on above four-aspect results as described in Methods. As a result, 102 out of the 108 taxa were new species, from which 28 novel genera and 3 novel families were further proposed (Table S7 and hGMB homepage [37]). The other 6 novel-taxon candidates were new strains of known species as they did not fully meet the criteria of new species (Grey text in Figure S1). The 102 new species were denominated and their protologues are provided in Table 1. More detailed descriptions of 102 new species are documented in Supplementary Data 1. Representative strains of 400 taxa were selected as described in Methods. With these efforts, we constructed the human Gut Microbial Biobank (hGMB), which comprises of 1,170 strains (Table S7) representing 400 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) for public access (hGMB homepage at CGMCC website: http://www.cgmcc.net/english/hgmb), and the type strains of novel 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 also available at hGMB homepage [37] and eLMSG [38].

The hGMB expands the existing human gut microbial collections and provides “most wanted” gut microbial resources

To better understand the cultivated bacterial diversity of human gut microbiota as well as to demonstrate the expansion of the existing public-available human gut bacterial repository by hGMB, 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 by comparison of 16S rRNA gene sequence distance as described in Methods. 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 1,519 cultivable bacterial species from human gut. Figure 1c shows the shared and unique bacterial species among the 6 studies. The hGMB provides 138 unique gut microbial species to the large-scale-cultivation-based gut microbial repository. Notably, 76 of the 138 hGMB unique species were taxonomically-undescribed novel taxa. We also calculated the numbers of novel taxa identified by each study (Figure 1d). The six studies claimed in total 416 nonredundant novel taxon candidates, of which 102 were well described and denominated under the rules of ICNP by hGMB, accounting for 24.5% of the total novel taxa.

By BLAST analysis, we further identified that 24 hGMB species were on the list of “most-wanted” or “medium priority” taxa proposed by the Human Microbiome Project [39] (Table S7). One “most-wanted” taxon-the

Eubacterium difficile sp. nov. (Taxon_69) and 9 “medium priority” taxa including three novel genera (*Simiaoa* gen. nov., *Jutongia* gen. nov. and *Wansuia* gen. nov.) were novel taxa firstly described in this work (Table 1).

The hGMB largely represents the taxonomic diversity of human gut microbiota

To further evaluate the taxonomic representativeness of hGMB to the main taxonomic composition of 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 had specimen numbers ranging from 102 to 3,538, represented human gut microbiota from donors of diverse genetic and environmental backgrounds (see Table S8 for accessions of the studies). The 26 datasets were separately processed, quality-controlled and weighted by a standard USEARCH-based analysis pipeline as described in Methods section. Results showed that the 26 datasets contained totally 11,647 quality-controlled samples (n=11,647) and each had 228±85 OTUs. The taxonomy status of each OTU was annotated using LTP_vhGMB customized by supplementation of LTP database v132 with the 102 novel taxa. 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 2a and 2b, the accumulative curves were almost saturated after sampling 24 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 taxonomically-defined human gut microbiota composition at genus and species levels. 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 2c). 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 2d). 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.

Novel taxa are prevalent in global human gut microbiome and illuminate the taxonomic “dark matters”

The 102 out of the 400 hGMB species were the first time being reported and they represent new taxa. To display the distribution and abundance of the new taxa in human gut microbiomes, we retrieved the open-access metagenomes (n=1,129) representing the health human GMs globally for combined analysis. The metagenomic datasets (Table S9) was selected by searching in GMrepo[40] with defined filter conditions as described in Methods. The distribution of 102 new taxa among the 1,129 metagenomes were investigated by kraken2-based annotation of each sample with customized taxonomically-defined GTDB database supplemented with 102 hGMB new-species genomes, and the relative abundance of novel taxa in each sample was estimated by Bracken (See Methods for details). In average 72.4±14.3% of the total reads of the 1,129 metagenomes were taxonomically classified and the hGMB novel taxa covered 15.4±7.4% of the classified reads. The results shown in Table S10 and Figure 3 revealed that 101 out of the 102 novel taxa were annotated in at least one metagenome, and 31 of the 101 novel taxa had average RAs>0.1% (box-and-whiskers plot in Figure 3). Notably,

the new hGMB taxa were widely distributed among global human gut metagenomes, as 95, 82 and 17 of the hGMB novel taxa were found in >50%, >90% and 100% of the investigated metagenomic samples (n=1,129), respectively, accounting for 93.1%, 80.4% and 16.7%. of all the novel taxa described in this study (bar chart in Figure 3).

Most recently, researchers identified 4,644 inferred prokaryotic species by construction of the largest-to-date Unified Human Gastrointestinal Genome (UHGG) database, and 70% of the UHGG species was assigned based on metagenome-assembled genomes (MAGs) but lacking cultured representatives [[8]]. To assess the possible contribution of new genomes in hGMB to the improvement of cultured representatives of UHGG species as well as to the illumination of potential “dark taxonomic taxa” that were not identified from culture-independent metagenomic studies, the Mash distance between the 102 novel taxon genomes and 4,644 UHGG representatives were calculated, and the genome pairs maintaining a distance<0.05 were identified as from the same species. As shown in Table S10, 78 out of the 102 new genomes got matched to the UHGG species, and 22 of them are uncultured species having only MAG representatives in UHGG. Thus, the hGMB species made the 22 UHGG genomes cultivated. Additionally, 6 UHGG species matched by hGMB genomes had only cultured genomes from unknown environments, demonstrating their representatives occurred in human gut. Notably, 24 hGMB new genomes did not get hit on any UHGG species-level genomes, indicating they were “dark species” in human GMs that had not been identified by previous cultivation-based or metagenomic studies.

New hGMB genomes enrich the global human gut gene catalogs and recover cultivated “dark” gene repository

Gene cataloging outlines human GM functionality potentials, and several gene catalogs have been established [8, 41]. We created nonredundant gene catalogs containing 341,876 nonredundant genes with 115 newly-sequenced hGMB genomes (named hGMB.catalog) and compared them with the largest-to-date human GM catalogs, the Unified Human Gastrointestinal Protein (UHGP) catalogue and the Integrated Gene Catalog (IGC) by BLAST analysis. Though the majority (79-90%) of the nonredundant genes in hGMB catalogs were represented by IGC and UHGP (Table S11), the hGMB further enriched human GM gene catalogs. With threshold value of 60% of amino acid sequence identity (for functional conservation), the hGMB contributed 43,202 and 70,123 of new 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 31,042 and 38,967, separately. As shown in Figure 4a, the hGMB.catalog covered 16.0% and 22.9% of IGC genes under the threshold identities of 60% and 40%, respectively. For UHGP, the coverages by hGMB.catalog were 14.0% and 20.6% at functional and structural level, respectively.

We then investigated the representativeness of hGMB genomes to the characterized functions of human GMs. For this purpose, the UHGP, IGC and all the 115 hGMB genomes were annotated with eggNOG [15]. A cumulative analysis of the KO and GO profiles were conducted to unravel the coverages of IGC and UHGP by random incremental selection of the hGMB genomes, and the results were shown in the rarefaction curves (Figures 4b and 4c). It revealed that the hGMB genomes covered 55.5% and 56.2% of the KO genes from IGC and UHGP catalogs, respectively (purple lines in Figure 4b and 4c). Similarly, the hGMB genomes represented 47.1% and 49.2% of the known GO functions of IGC and UHGP catalogs, respectively (blue lines in Figures 4b and 4c).

In addition to the representativeness of functionally-known genes of human GMs, the hGMB provided also a cultivated repository of functionally-unknown genes within the global gene catalogs, and the recovery of these “dark” genes by cultivated hGMB members would facilitate the culture-based experimental studies to bring more

human gut “dark” functions to light. The eggNOG annotation results of IGC and UHGP catalogs revealed that 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 5.2% (grey line in Figure 4b) and 4.1% (grey line in Figure 4c) of the unannotated genes in IGC and UHGP, respectively. Those functionally unknown genes matched for IGC and UHGP are listed in Tables S12 and S13. We also plotted by hGMB.catalog and hGMB genomes the coverage of the Function Unknown Families of homologous proteins (FUnkFams), a “most wanted” list of conserved microbial protein families with no known domains and prioritized for functional characterization [42]. The results revealed that, with a threshold value of 40% sequence identity, the hGMB covered 6,635 out of 61,970 (10.7%) of the functionally unknown proteins in FUnkFams (Figures 4a and 4d). The profiles of the matched FUnkFams sequences to the hGMB genomes are summarized in Table S14 facilitating further culture-based study of them.

Discussion

By implications of previous experiences on cultivation and understandings of gut microbial physiology and ecology [23, 25, 26, 36], 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 S4). For example, based on our previous study [36], 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, the 8-carbohydrates mixture was supplemented to media to improve human gut microbial cultivability (Table S4). The results indicated that this mixture improved the growth of quite a few gut bacterial isolates, especially members of *Clostridiales* and *Erysipelotrichales*. According to our statistics, the *Eubacterium hominis* sp. nov., *Eubacterium segne* 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. As shown in Table S7, the 102 new species identified in hGMB belonged to 24 different families (including 3 novel families), and *Lachnospiraceae* was the most abundant family including 29 new species and 7 new genera (*Wujia* gen. nov., *Simiaoa* gen. nov., *Jutongia* gen. nov., *Qiania* gen. nov., *Zhenhengia* gen. nov., *Jingyaoa* gen. nov., *Wansuia* gen. nov.). Similarly, the *Lachnospiraceae* is one of the most dominant families in the GM of healthy adults, accounting for 10-45% of total bacteria in feces [43], and is considered playing diverse but controversial roles in the maintenance of host gut homeostasis [27, 44]. 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 [45-47], protection of hosts from pathogen infections [46, 48, 49] and from stress-induced visceral hypersensitivity [45]. On the other hand, researches displayed positive correlations between *Lachnospiraceae* and diseases such as non-alcoholic fatty liver disease (NAFLD) [50] and chronic kidney disease (CKD) [51]. Animal experiments demonstrated that the gavage of *Lachnospiraceae* accelerated the development of diabetes in obese mice [52] and aggravated the inflammation of intestinal epithelial cells in TLR5$^{-/-}$ mouse [53]. The contradictory conclusions signified that the function(s) of *Lachnospiraceae*, a predominant gut microbial family in humans, are 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.

The hGMB provides also members of *Christensenellaceae*, including *Christensenella minuta*, *Christensenella tenuis* and 3 new genera (*Guopingia* gen. nov., *Luoshenia* gen. nov. and *Gehongia* gen. nov.). The *Christensenellaceae* is a recently identified gut commensal bacterial family containing limited cultivated representatives [54], and has been considered as a promising probiotic candidate for intervention of obesity and other metabolic syndromes [55, 56]. Particularly, the *Christensenella minuta* was experimentally verified to reduce weight gain in recipient mice [57]. To explore and evaluate *Christensenellaceae*'s therapeutic potential, more studies are necessary. The hGMB provides resources serving further studies. Notably, the *Guopingia* and its type species *Guopingia 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 S7, Figures 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* [58], *Faecalibacterium prausnitzii* [59], *Roseburia intestinalis[60]*, *Lactobacillus* and *Bifidobacterium* members [61, 62], while some others, as *Enterococcus faecium* [63], *Ruminococcus gnavus* [64],*Clostridioides difficile* [65] and *Klebsiella* species [66], 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 [67, 68]. An example is the *Bacteroides fragilis*, as both pathogenic and probiotic strains were identified from this species [68, 69]. 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 [70]. 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 obtained. These bacterial isolates represent 400 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 400 human gut microbial species was established. The hGMB expands gut microbial resources and genomic repository by adding 102 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 novel taxa were described and denominated following the rules of ICNP for later valid approval of nomenclatures. Further analysis revealed that the hGMB represented over 80% of the prevalent microbial genera and species in human guts, and covered 50% of KEGG Orthology functions and 10% of the functionally-unknown genes in FUnkFams. By integrative analysis of hGMB genomes with UHGG database and 1,129 global health human gut metagenomes, we profiled the taxonomic prevalence, distribution and genetic features of the 102 new hGMB species among 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 whole project was approved by the Research Ethics Committee of Institute of Microbiology, Chinese Academy of Science, and the assigned number authority of the ethical approval is APIMCAS2017049. We inquired each donor candidates about the health conditions, history of clinical visits for the last half year and

history of antibiotic treatments for the last two months in person before a consent form was signed for donation of feces, and the ones without any clearly diagnosed chronic and malignant disease were considered as healthy donors. 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 areas of China (Beijing, Henan, Heibei, Xinjiang, Guangdong, Inner Mongolia), and the age and gender information was listed in Table S2. The samples collected in Beijing were kept fresh and transferred into anaerobic workstation (AW500, Electrotek, UK) for sample pretreatment within 2 hours, while the feces from the other areas were froze 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 location were merged together for pretreatment and subsequent isolation steps. The 11 pretreatment conditions were given in Table S3 and the alcohol pretreatment strategies were derived from Browne et al. [23]. The gas flow 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 serially diluted into 10^{-1}-10^{-8} folds. Then, 100 µl of each dilution were spread onto different agar plates for either aerobic or anaerobic incubations at 37 °C. We applied 67 different culture conditions for bacterial cultivation and isolation as shown in Table S5. The detailed recipes of 21 base media 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 [36] : 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 centrifuged 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 dilution were used as template 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 sequenced 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 medium and streaking on agar plates for further purification, preservation and characterization either anaerobically or aerobically. During the taxonomic characterization and preparation of strain transferred to IDAs, strains were serially inoculated into new media and cultivated and transferred for several generations. In each inoculation and cultivation step, the 16S rRNA gene sequences of the new culture was sequenced and checked. The taxonomy 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 [71]. The isolates with 16S rRNA gene sequence identities ≤ 98.7% to any known species in both databases were considered as candidates of novel taxa [71]. All the isolates potentially representing novel 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 novel taxon, 1 strain was designed as type strain for later 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 different 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 [36]: 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 S6 and hGMB special page on CGMCC (http://www.cgmcc.net/english/hgmb). To meet the rules of International Code of Nomenclature of Prokaryotes (ICNP), the 102 type strains of new species in hGMB were also preserved in a second IDA as KCTC or NBRC, and the majority of accessions could be found in Table 1 and hGMB homepage.

Polyphasic characterization and nomenclature of novel taxa

The delineations of novel taxa were based on the analysis of each type strain in terms of phylogenetic, genomic, physiological and morphological characteristics as described in previous work [36, 72] 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 [73] under Neighbour-Joining method to depict the phylogenetic distribution and taxonomic relation of each novel taxa and its closely-related taxa (Figure SD-1a to Figure SD-108a in Supplementary Data 1). Additionally, the genome-based phylogenomic tree for each new species was also constructed using gtdb-tk with classify_wf command under default parameters [74] (Figure SD-1c to Figure SD-108c in Supplementary Data 1). The closely-related taxa on phylogenetic and phylogenomic trees were used for further genome-based analysis. The genome-based analysis of novel taxa included the calculation of the Average Nucleotide Identity (ANI), digital DNA:DNA hybridization (dDDH) and the percentage of conserved proteins (POCP). The ANI values and the heatmaps (Figure SD-1d to Figure SD-108d in Supplementary Data 1) were generated using OrthoANI OTA software [75]. The dDDH value between draft genome of new species and its phylogenetically and phylogenomically closest genomes were calculated using the Genome-to-Genome Distance Calculator 2.1 (GGDC) [76]. 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 [72, 77]. The physiological and biochemical features of type strains of novel 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 novel taxa was proposed according to the rules of ICNP. After comprehensive consideration of several main works [36, 71, 72, 77-79], the following criteria were used for proposing novel taxa: 1, Taxon meeting the following three criteria simultaneously was defined as new species: (1) the 16S rRNA sequence identity<98.7%, (2) dDDH value<70%, (3) ANI<95%, or ANI between 95%~96% but the morphology and physiology feature of novel taxon was distinct from that of its closely-related species. 2, If the new species simultaneously had (1) a 16S rRNA gene sequence

identity<95% to any known species, (2) a POCP value <50% to its closely-related taxon, (3) any significant difference in morphology and physiology with neighbor genera, and (4) location at an independent clade on the phylogenetic tree, it would be further defined as new genus. 3, If the type species in the new genus (1) had a 16S rRNA gene sequence identity<90% to any known species, (2) was clustered on a separate clade distant from any known genera on the phylogenetic tree and its closest neighbor genera were from at least two different families, and (3) maintained significant difference in morphology and physiology to the neighbor families, the taxon would be further defined as new family.

Genome sequencing and analysis

The genomes of all 108 novel-taxon candidates and 7 known species with no genome available in NCBI were sequenced. The genomic DNA were extracted using either the DNeasy Blood & Tissue Kit (Qiagen, Germany) or the Wizard Genomic DNA Purification kit (Promega, USA). 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 [80] into a number of contigs (k-mer sizes of 59, 79, 99 and 119), and the contigs longer than 500 nt were retained as final splicing. The assembled contigs were then BLASTed against NCBI nt database using blastn with e-value of 1e-5 to remove potential contamination contigs not hitting to the target taxonomic classification. Above library preparation, sequencing and assembly steps were performed by commercial company (Guangdong Magigene Biotechnology Co.,Ltd., China). The quality and assembly information of the genomes from commercial company were further assessed in lab. The numbers and N50 of contigs in each genome, the contamination and the completeness were estimated using CheckM v1.0.12 (lineage_wf function) [81], and were listed in Table S6. The Estimated quality score of each assembly was calculated by "completeness – 5 × contamination" [8]. Genomes with contamination>5% were further decontaminated using MAGpurify v2.1.2 [29]. If any quality-controlled genome had <50% completeness, or >5% contamination or an estimated quality score<50 would be re-sequenced. For the estimation of average coverage depth of assembly to the sequencing short reads, bwa v0.7.17 (mem function) [82]were used for reads mapping, and samtools v 1.9

(view -F 4 -bS and depth commands)[83] were used for depth estimation. A one-line script (less samtools.depth.output.file|awk '{sum+=$3; sumsq+=$3*$3} END { print "Average = ",sum/NR}') were used for extraction of average depth from the output of samtools and the results.

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]. The function annotations of all genomes were performed with eggNOG database v4.5 by local emapper v1.0.3 (-m diamond) [15]. The comparison of novel-taxon genomes with 4,644 species-level genomes in UHGG were performed using Mash v2.2.2 (dist function) [87], and the genome pairs maintaining a mash distance <0.05 (equal to ANI>95%) were identified to represent the same species. The MAGnify accession and culturing status of UHGG genomes hit by hGMB new genomes were extracted from the Table S3 of the UHGG publication [8]. Default parameters were used for each software unless otherwise specified.

Human gut metagenome collection and analysis

The publically available metagenomic data representing the global health human GMs were selected by search with defined filter conditions (experiment_type = 'Metagenomics' AND QCStatus = ‘Good runs’ AND host age > 5 AND country=is not null AND Recent Antibiotics Use = 'No' AND Phenotype = ‘Health’) in GMrepo [40]. In total 1,168 entries obtained from the above query from 6 different studies including male and female donors from 5 countries worldwide (Canada, United Republic of Tansania, Italy, China and USA), and 1,129 packages of the qualified raw data were successfully downloaded from NCBI using sra toolkit v2.10.8 [88] and used for further analysis. The accession information of the 1,129 samples were listed in Table S9. The distribution of novel taxa among metagenomes were estimated by Kraken 2 v2.0.9-beta [89]. A customized Kraken 2 database was constructed for taxonomic annotation by collection of all the representative genomes (n=8,377) with defined species designation from GTDB release 95 [90] and combination of them with 102 novel taxon genomes from hGMB to generate the GTDB-species_vhGMB database. Then, the 1,129 metagenomes were taxonomically annotated. The abundance of assigned species in each metagenomic sample was estimated using Bracken [91]. Default parameters were used for each software unless otherwise specified.

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 novel 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 novel taxa in each study. There were 68, 100, 141 and 22 novel 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 novel taxa among studies. The 16S rRNA gene sequences representing novel 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 [73]. 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 novel taxa. To display the hGMB coverage of Human Microbiome Project's Most Wanted taxa [39], 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+ [92]. If the 16S rRNA gene sequences of hGMB members had sequence identities >98.7% 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 S7 (Column named as "Most wanted" taxa). All the taxa included by the hGMB were exhibited as taxonomic cladogram using GraPhlAn v1.1.3 [93], 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 [94]. Default parameters were used for each software unless otherwise specified.

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 S8. To enable an equally-weighted representation of human GMs, the 26 studies were separately processed and quality-controlled by 64-bit Usearch v11 [95] 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 102 novel 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 [36]. The equally-weighted average values (RA and FO) were further calculated by averaging the mean values of each study. All the mean values of RAs and FOs relating to the 26 studies were presented as the equally-weighted average values ± standard deviation (SD) unless otherwise specified. 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 [96] and displayed as accumulating curves. The distribution of dominant taxa in global human GMs were displayed as box-and-whiskers plots 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) [97]. The nonredundant gene catalog hGMB.catalog was constructed using 115 genomes sequenced in this study by CD-HIT software v4.5.8 [98] (-o out.file -c 0.95 -aS 0.9 -n 5 -M 64000 -T 48). The hGMB.catalog containing 341,876 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. It revealed 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 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 [96] to generate data used for the construction of cumulative curves. The BLAST analysis of single genomes in hGMB and hGMB gene catalogs 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.catalog 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 [99-101].

To profile the coverage of functionally-unknown genes of IGC, UHGP and FUnkFams by hGMB genomes, DIAMOND-based BLAST analysis [97] 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 115 hGMB genomes were profiled as Table S12-14. The ratios of unannotated genes of 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.

Statistical analysis

All statistical analyses were performed with IBM SPSS Statistics 20. All the box-and–whiskers plots, bar charts and accumulative curves were generated using Graphpad Prism v6 [102] 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 coverage 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 400 hGMB species is available at hGMB homepage (http://hgmb.nmdc.cn) [37]. The 1,170 strains and their 16S rRNA gene sequences were accessible via hGMB special page on CGMCC official website (http://www.cgmcc.net/english/hgmb). The taxonomic descriptions of all novel taxa are also accessible at eLMSG under accessions from MSG071057 to MSG071268, MSG071857 and MSG071858 (link type: https://www.biosino.org/elmsg/record/MSG071057) [38]. All the genomes obtained in this study are available at NCBI under Bioproject PRJNA656402 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA656402) [103], NODE with the project accession OEP001106 (https://www.biosino.org/node/project/detail/OEP001106) [104], 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) [103], and in NMDC under accessions NMDC10014003 (http://hgmb.nmdc.cn/subject/hgmb) [105]. The gene catalog hGMB.catalog is deposited at hGMB homepage [37].

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). We thank NMDC (China National Microbiology Data Center) team for the data preseravation, hGMB homepage construction and maintenance, and thank Wan Liu for the submission of taxonomic information and genomes to eLMSG (https://www.biosino.org/elmsg/index).

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. CL, CYL. HWL 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. Nat Biotech. 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 Biotech. 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; 533:543-46.

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 Biotech. 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. 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.
27. 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.
28. Lau JT, Whelan FJ, Herath I, Lee CH, Collins SM, Bercik P, et al. Capturing the diversity of the human gut microbiota through culture-enriched molecular profiling. Genome Med. 2016; 8.
29. 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.
30. Oren A, Garrity GM, Parte AC. Why are so many effectively published names of prokaryotic taxa never validated? Int J Syst Evol Micr. 2018; 68:2125-9.
31. Parker CT, Tindall BJ, Garrity GM. International Code of Nomenclature of Prokaryotes Prokaryotic Code (2008 Revision). Int J Syst Evol Micr. 2019; 69:7-111.
32. Yoon SH, Ha SM, Kwon S, Lim J, Kim Y, Seo H, et al. Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. Int J Syst Evol Micr. 2017; 67:1613-7.
33. Federhen S. The NCBI Taxonomy database. Nucleic Acids Res. 2012; 40:136-43.
34. 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.
35. 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.
36. 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.
37. This_study. hGMB. hgmb.nmdc.cn. Accessed 04 Sep 2020.
38. eLMSG. www.biosino.org/elmsg. Accessed 4 Sep 2020.
39. 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.
40. Wu SC, Sun CQ, Li YZ, Wang T, Jia LH, Lai SY, et al. GMrepo: a database of curated and consistently annotated human gut metagenomes. Nucleic Acids Res. 2020; 48:545-53.
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. Nature Biotechnology. 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. Ishiguro E, Haskey N, Campbell K. Gut microbiota : interactive effects on nutrition and health. London, United Kingdom: Academic Press; 2018.

44. Vacca M, Celano G, Calabrese FM, Portincasa P, Gobbetti M, De Angelis M. The Controversial Role of Human Gut Lachnospiraceae. Microorganisms. 2020; 8.
45. 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.
46. 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; 517:205-7.
47. 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.
48. 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.
49. 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.
50. 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.
51. 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.
52. 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.
53. 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.
54. 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.
55. Waters JL, Ley RE. The human gut bacteria Christensenellaceae are widespread, heritable, and associated with health. Bmc Biol. 2019; 17.
56. 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.
57. 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.
58. 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; 25:1096-103.
59. 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.
60. 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.

61. 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.
62. 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.
63. 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.
64. Titecat M, Wallet F, Vieillard MH, Courcol RJ, Loiez C. Ruminococcus gnavus: An unusual pathogen in septic arthritis. Anaerobe. 2014; 30:159-60.
65. 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.
66. 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.
67. Wexler HM. Bacteroides: the good, the bad, and the nitty-gritty. Clin Microbiol Rev. 2007; 20:593-621.
68. Sears CL, Geis AL, Housseau F. Bacteroides fragilis subverts mucosal biology: from symbiont to colon carcinogenesis. J Clin Invest. 2014; 124:4166-72.
69. 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.
70. 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.
71. Stackebrandt E. Taxonomic parameters revisited: tarnished gold standards. Microbiol Today. 2006; 33:152-5.
72. 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 (vol 1, 16131, 2016). Nat Microbiol. 2016; 1.
73. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016; 33:1870-4.
74. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. 2020; 36:1925-7.
75. 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.
76. 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.
77. 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.
78. Kim M, Oh HS, Park SC, Chun J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes (vol 64, pg 346, 2014). Int J Syst Evol Micr. 2014; 64:1825-.
79. Chun J, Oren A, Ventosa A, Christensen H, Arahal DR, da Costa MS, et al. Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. Int J Syst Evol Micr. 2018; 68:461-6.

80. 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.
81. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015; 25:1043-55.
82. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997. 2013.
83. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078-9.
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. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016; 17.
88. Sherry S, Xiao C, Durbrow K, Kimelman M, Rodarmer K, Shumway M, et al. Ncbi sra toolkit technology for next generation sequence data. In Plant and Animal Genome XX Conference (January 14-18, 2012) Plant and Animal Genome. 2012
89. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019; 20.
90. Zhou CF, Xu QW, He S, Ye W, Cao RF, Wang PY, et al. GTDB: an integrated resource for glycosyltransferase sequences and annotations. Database-Oxford. 2020, ARTN baaa04710.1093/database/baaa047.
91. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data. Peerj Comput Sci. 2017; 3:e104.
92. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST plus : architecture and applications. Bmc Bioinformatics. 2009; 10.
93. Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. Peerj. 2015; 3.
94. Bardou P, Mariette J, Escudie F, Djemiel C, Klopp C. jvenn: an interactive Venn diagram viewer. Bmc Bioinformatics. 2014; 15.
95. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010; 26:2460-1.
96. Dixon P. VEGAN, a package of R functions for community ecology. J Veg Sci. 2003; 14:927-30.
97. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015; 12:59-60.
98. 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.
99. Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol. 2001; 307:1113-43.
100. 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.
101. Tian WD, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol. 2003; 333:863-82.
102. Basham B. Graphpad Prism. Biotechnol Softw I J. 1997; 14:14-7.
103. 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.
104. NODE. The hGMB data deposited in NODE under Project OEP001106. https://www.biosino.org/node/project/detail/OEP001106. Accessed 4 Sep 2020.
105. NMDC. The hGMB data deposited in NMDC under Bioproject NMDC10014003. http://hgmb.nmdc.cn/subject/hgmb. Accessed 4 Sep 2020.
106. 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.
107. 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.
108. 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.
109. Togo A, Valero R, Delerce J, Raoult D, Million M. “Anaerotruncus massiliensis,” a new species identified from human stool from an obese patient after bariatric surgery. New microbes and new infections. 2016; 14:56.
110. 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.
111. 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.
112. 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.

Tables

Table 1. The protologues of 102 novel taxa in hGMB (Rank*: “nom. rev.” indicated the nomenclatures that were published effectively but not validly, the original publication of the nomenclature was cited after the proposed name; Phenotypic Description*: more detailed descriptions of new taxa in terms of phylogenetic, phylogenomic, genomic, physiological and morphological properties were available in Supplementary Data 1.)

Taxonomy	Rank*	Etymology	Type Designation	Phenotypic Description*	GMCC / KCTC / NBRC Accessions
Yeguiaceae	fam. nov.	Ye.gui'a'ce.ae. N. L. neut. n. *Yeguia*, type genus of the family. -aceae, ending to denote a family,N. L. fem. pl. n. *Yeguiaceae*, family of the genus *Yeguia*	Type genus: *Yeguia*		
Yeguia	gen. nov.	Ye.gui'a N.L. fem. n. *Yeguia,* named in honour of the Chinese medical scientist Gui Ye	Type species: *Yeguia hominis*		
Yeguia 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	Cells are ovoid with peaked ends, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 5-10 days.	CGMCC 1.32813
Luoshenia	gen. nov.	Luo.shen'i.a. N.L. fem. n. *Luoshenia*, named after the Chinese Goddess Luoshen	Type species: *Luoshenia tenuis*		
Luoshenia tenuis	sp. nov.	te'nu.is. L. fem. adj. *tenuis,* thin, slim, referring to the predicted potential function of the strain in weight-loss	NSJ-44^T from human feces	Cells are ovoid with spiky ends, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32817
Feifaniaceae	fam. nov.	Fei.fa.ni.a.ce'ae. 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	Type species: *Feifania hominis*		

		after the Chinese microbiologist Feifan Tang			
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	Cells are rod-shaped, motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-12 days.	CGMCC 1.32862
Bianqueaceae	fam. nov.	Bian.que.ace'ae. N.L. fem. 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*		
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. fem. adj. *renquensis*, pertaining Renqiu county of China, the birthplace of Chinese medical scientist QueBian	NSJ-32^T from human feces	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32805
Gehongia	gen. nov.	Ge.hong'i.a. 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. fem. adj. *tenuis*, thin, slim, referring to the predicted potential function of the strain in weight-loss	NSJ-53^T from human feces	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32829 /KCTC 25141
Guopingia	gen. nov.	Guo.ping'i.a. N.L. fem. n. *Guopingia*,	Type species: *Guopingia tenuis*		

		named after the Chinese microbiologist Guoping Zhao			
Guopingia tenuis	sp. nov.	te'nu.is. L. fem. adj. *tenuis*, thin, slim, referring to the predicted potential function of the strain in weight-loss	NSJ-63^T from human feces	Cells are spherical, motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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. fem. adj. *zhengdingensis*, referring to Zhengding county of China, the birthplace of Li Gao	NSJ-31^T from human feces	Cells are spherical, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32804
Congzhengia	gen. nov.	Cong.zheng'i.a. N.L. fem. n. *Congzhengia*, named after the Chinese medical scientist Congzheng Zhang	Type species: *Congzhengia minquanensis*		
Congzhengia minquanensis	sp. nov.	min.quan.en'sis. N.L. fem. adj. *minquanensis*, referring to Minquan county of China, the birthplace of Congzheng Zhang	H8^T from human feces	Cells are spherical, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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 xinanensis*		
Fumia xinanensis	sp. nov.	xin.an.en'sis. N.L. fem. adj. *xinanensis*, referring to Xin'an	NSJ-33^T from human feces	Cells are rod-shaped or ovoid, non-motile.	CGMCC 1.32806

		county where Fumi Huang was born		Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	
Wujia	gen. nov.	Wu.ji'a. N.L. fem. n. *Wujia*, named after the Chinese medical scientist Wuji	Type species: *Wujia chipingensis*		
Wujia chipingensis	sp. nov.	chi.ping'en.sis. N.L. fem. adj. *chipingensis*, referring to Chiping county of China, the birthplace of the Chinsese medical scientist Wuji Cheng	NSJ-4^T from human feces	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are rod-shaped with blunt ends, non-motile. Growth in modified MGAM	CGMCC 1.32876

				medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	
Jutongia	gen. nov.	Ju.tong'ia. N.L. fem. n. *Jutongia*, in honor of the Chinsese medical scientist Jutong Wu	Type species: *Jutongia huaianensis*		
Jutongia huaianensis	sp. nov.	huai.an'en.sis. N.L. fem. adj. *huaianensis*, referring to huai'an county of China, the birthplace of the Chinsese medical scientist Jutong Wu	NSJ-37^T from human feces	Cells are straight rod-shaped, motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32810
Qiania	gen. nov.	Qian'i.a. N.L. fem. n. *Qiania*, named after the Chinese medical scientist Yi Qian	Type species: *Qiania dongpingensis*		
Qiania dongpingensis	sp. nov.	dong.ping'en'sis. N.L. fem. adj. *dongpingensis*, referring to Dongping county of China, the birthplace of Yi Qian	NSJ-38^T from human feces	Cells are ovoid to rod-shaped with tapered ends, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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. fem. adj. *yiwuensis*, referring to Yiwu city of China, where Zhenheng Zhu was born	NSJ-12^T from human feces	Cells are straight rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32465 /KCTC 15954

Jingyaoa	gen. nov.	Jing.yao'a. N.L. fem. n. *Jingyaoa*, named after the Chinese medical scientist Jingyao Zhang.	Type species: *Jingyaoa shaoxingensis*		
Jingyaoa shaoxingensis	sp. nov.	shao.xing'en.sis. N.L. fem. adj. *shaoxingensis*, referring to Shaoxing city of China, where Jingyao Zhang was born	NSJ-46^T from human feces	Cells are spherical or ovoid or short rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32819
Wansuia	gen. nov.	Wan.su'i.a. N.L. adj. fem., *Wansuia*, in honor of the Chinese medical scientist Wansu Liu	Type species: *Wansuia hejianensis*		
Wansuia hejianensis	sp. nov.	he.jian.en'sis, N.L. fem. adj. *hejianesis*, referring to Hejian county of China, the birthplace of the Chinese medical scientist Wansu Liu	NSJ-29^T from human feces	Cells are ovoid to rod-shaped with spiky ends, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32802 /KCTC 25078
Zhenpiania	gen. nov.	Zhen.pian'i.a. 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	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32877
Lentihominibacter	gen. nov.	Len.ti.ho.mi.ni.bac'ter. L. masc. n. *lentus*, slow. L. masc. n. bacter, a rod. N.L.	Type species: *Lentihominibacter hominis*		

		masc. n. *Lentihominibacter*, slowly growing rod-shaped bacterium			
Lentihominibacter faecis	sp. nov.	L. gen. fem. n. *faecis*, of faeces, from which the organism was isolated	BX16^T from human feces	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32874
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	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are rod-shaped, motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32879
Shuzhengia	gen. nov.	Shu.zheng'i.a. N.L. fem. n. *Shuzhengia*, named after the 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	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32880
Anaerofilum hominis	sp. nov.	ho'mi.nis. L. gen. masc. n. *hominis*, of a human being,	BX8^T from human feces	Cells are rod-shaped, motile. Growth in	CGMCC 1.32881

		referring to the human gut habitat		modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	
Zongyangia	gen. nov.	Zong.yang'i.a. 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	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32830 /KCTC 25132
Youxingia	gen. nov.	You.xing'i.a. N.L. fem. n. *Youxingia*, named after the Chinese medical scientist Youxing Wu	Type species: *Youxingia wuxianesis*		
Youxingia wuxianensis	sp. nov.	wu.xian.en'sis. N.L. fem. adj. *wuxianensis*, referring to the Wuxian county of China, where Youxing Wu was born	NSJ-64^T from human feces	Cells are ovoid to rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32840 /KCTC 25128
Qingrenia	gen. nov.	Qing.re'ni.a. N.L. fem. n. *Qingrenia*, named after the Chinese medical scientist Qingren Wang	Type species: *Qingrenia yutianensis*		
Qingrenia yutianensis	sp. nov.	yu.tian.en'sis. N.L. fem. adj. *yutianensis*, referring to Yutian county of China, where Qingren Wang was born	NSJ-50^T from human feces	Cells are ovoid with spiky ends, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32823

Jilunia	gen. nov.	Ji.lun'i.a. N.L. fem. n. *Jilunia*, named after the Chinese microbiologist Jilun Li	Type species: *Jilunia laotingensis*		
Jilunia laotingensis	sp. nov.	lao.ting.en'sis. N.L. fem. adj. *laotingensis*, referring to the Laoting county where Jilun Li was born	N12^T from human feces	Cells are spherical or ovoid or short rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32860
Paratissierella	gen. nov.	Pa.ra.tis.sier.el'la. Gr. prep. para, beside. N.L. fem. dim. n. *Tissierella*, a genus name. N.L. masc. n. *Paratissierella*, resembling the genus *Tissierella*	Type species: *Paratissierella segnis*		
Paratissierella segnis	sp. nov.	L. fem. adj. *segnis*, slow, inactive, lazy, referring the slow growth of the strain	BX21^T from human feces	Cells are rod-shaped, motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32882
Bittarella (ex Durand et. al, 2017)	nom. rev.	N.L. fem. dim. n. *Bittarella*, in honour of Dr Bittar, a French microbiologist *[106]*	Type species: *Bittarella massiliensis*		
Bittarella massiliensis (ex Durand et. al, 2017)	nom. rev.	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 *[106]*	NSJ-19^T from human feces	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32824 /KCTC 25133
Eggerthella hominis	sp. nov.	ho'mi.nis. L. gen. masc. n. *hominis*, of a human being,	NSJ-70^T from human feces	Cells are straight rod-shaped, non-motile.	CGMCC 1.32846 /KCTC 25139

		referring to the human gut habitat		Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	
Gordonibacter massiliensis (ex Ngom et. al, 2020)	nom. rev.	mas.si.li.en'sis. L. adj. masc. *massiliensis*, of Massilia, Marseilli, where the bacterium was for the first time isolated *[107]*	NSJ-58^T from human feces	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32834 /KCTC 25146
Bacteroides multiformis	sp. nov.	mul.ti.for'mis. L. masc. adj. *multiformis*, many-shaped, multiform, referring to the various size and shape of the strain	L5^T from human feces	Cells are spherical or ovoid or rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are rod-shaped in various sizes, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 1-3 days.	CGMCC 1.32853 /KCTC 25155
Bacteroides difficilis	sp. nov.	dif.fi'ci.lis. L. masc. adj. *difficilis*, difficult, referring the difficulty of culturing the strain	NSJ-74^T from human feces	Cells are ovoid or short rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32850
Bacteroides hominis	sp.	ho'mi.nis. L. gen.	NSJ-2^T from	Cells are	CGMCC

	nov.	masc. n. *hominis*, of a human being, referring to the human gut habitat	human feces	spherical or ovoid or short rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are ovoid to rod-shaped with round or blunt ends, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are straight or slightly curved rod-shaped, non-motile. Growth in modified MGAM medium occurs at 38 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are ovoid or short rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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.	426-9^T from human feces	Cells are rod-shaped, non-motile. Growth in modified	CGMCC 1.13558 /NBRC 113433

		adj. *acidifaciens*, acid-producing		MGAM medium occurs at 38 °C, pH 7.0-7.5, in1-3 days.	
Parabacteroides segnis	sp. nov.	seg'nis. L. masc. adj. *segnis*, slow, inactive, lazy, referring the slow growth of the strain	BX2^T from human feces	Cells are rod-shaped, motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are rod-shaped, motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are ovoid to short rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are rod-shaped, motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32885
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	Cells are ovoid to short rod-shaped, non-motile. Growth in modified	CGMCC 1.32470 /KCTC 15949

				MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	
Christensenella tenuis	sp. nov.	te'nu.is. L. fem. adj. *tenuis*, thin, slim, referring to the predicted potential function of the strain in weight-loss	NSJ-35^T from human feces	Cells are rod-shaped with spiky ends, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are ovoid to rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32461 /KCTC 15960
Clostridium lentum	sp. nov.	len'tum. L. neut. adj. *lentum*, slow, referring to the slow growth of the type strain	NSJ-42^T from human feces	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32815
Clostridium facile	sp. nov.	fa'ci.le. L. neut. adj. *facile*, easy, without difficulty, referring that the type strain is easily cultured	NSJ-27^T from human feces	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 1-3 days.	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	Cells are rod-shaped, non-motile. Growth in modified MGAM	CGMCC 1.32844 /KCTC 25150

				medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	
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	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32462 /KCTC 15959
Blautia massiliensis (ex Durand et. al, 2017)	nom. rev.	mas.si.li.en'sis. L. fem. adj. *massiliensis*, of Massilia, the Latin name of Marseill, where the bacterium was for the first time isolated [108]	4-46^T from human feces	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.52830 /NBRC 113773
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	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.52850 /NBRC 113774
Blautia segnis	sp. nov.	seg'nis. L. fem. adj. *segnis*, slow, inactive, lazy, referring the slow growth of the strain	BX17^T from human feces	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32886
Blautia tarda	sp. nov.	tar'da. L. fem. adj. *tarda*, slow, inactive, lazy, referring the slow growth of the strain	BX19^T from human feces	Cells are rod-shaped with tapered ends, non-motile. Growth in modified MGAM medium occurs at 37	CGMCC 1.32887

				°C, pH 7.0-7.5, in 3-10 days.	
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	Cells are rod-shaped with tapered ends, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 1-3 days.	CGMCC 1.32807
Blautia lenta	sp. nov.	len'ta. L. fem. adj. *lenta*, slow, referring to the slow growth of the type strain	M16^T from human feces	Cells are curved or straight rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32888
Blautia difficilis	sp. nov.	dif.fi'ci.lis. L. fem. adj. *difficilis*, difficult, referring the difficulty of culturing the strain	M29^T from human feces	Cells are ovoid to short rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32889
Clostridium segne	sp. nov.	seg'ne. L. neut. adj. *segne*, slow, inactive, lazy, referring the slow growth of the strain	BX14^T from human feces	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are ovoid, non-motile. Growth in modified MGAM	CGMCC 1.32463

				medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	
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	Cells are straight rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are straight rod-shaped with peaked ends, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32891
Eubacterium segne	sp. nov.	seg'ne. L. neut. adj. *segne*, slow, inactive, lazy, referring the slow growth of the strain	BX4^T from human feces	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32892
Eubacterium difficile	sp. nov.	dif.fi'ci.le. L. neut. adj. *difficile*, difficult, referring the difficulty of culturing the strain	M5^T from human feces	Cells are curved or straight rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32893
Hungatella hominis	sp. nov.	ho'mi.nis. L. gen. masc. n. *hominis*, of a human being,	NSJ-66^T from human feces	Cells are fusiform rod-shaped, non-motile.	CGMCC 1.32842 /KCTC 25127

		referring to the human gut habitat		Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	
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	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are spiral or vibrio or rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32831
Ruminococcus difficilis	sp. nov.	dif.fi'ci.lis. L. masc. adj. *difficilis*, difficult, referring the difficulty of culturing the strain	M6^T from human feces	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32867
Roseburia lenta	sp. nov.	len'ta. L. fem. adj. *lenta*, slow, referring to the slow growth of the type strain	NSJ-9^T from human feces	Cells are short comma-shaped or long, thin	CGMCC 1.32469

				rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	
Roseburia yibonii	sp. nov.	yi.bon'i.i N.L. gen. masc. n. *yibonii*, referring to Chinese actor Wang Yibon, whose series inspired the researcher during the bacterial identification	BX0805^T from human feces	Cells are comma-shaped with spiky ends or clavate ends, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32827
Roseburia zhanii	sp. nov.	zha'ni.i N.L. gen. masc. n. *zhanii*, of Zhan, referring to Zhan Xiao, a Chinese actor whose series inspired the researcher during the bacterial identification	BX1005^T from human feces	Cells are rod-shaped or comma-shaped with spiky ends, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32828
Roseburia rectibacter	sp. nov.	rec.ti.bac'ter. L. masc. 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	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32845
Roseburia difficilis	sp. nov.	dif.fi'ci.lis. L. fem. adj. *difficilis*, difficult, referring the difficulty of culturing the strain	NSJ-67^T from human feces	Cells are spherical, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-	CGMCC 1.32843

				7.5, in 3-10 days.	
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	Cells are ovoid to rod-shaped spiky ends, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32866
Agathobaculum faecis	sp. nov.	fae'cis. L. gen. fem. n. *faecis*, of faeces, from which the organism was isolated	NSJ-28^T from human feces	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32801
Anaerotruncus massiliensis (ex Togo et. al, 2016)	nom. rev.	mas.si.li.en'sis. L. masc./fem. adj. *massiliensis*, pertaining to Marseille, France, where the organism was for the first time isolated [109]	22A2-44^T from human feces	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.52380 /NBRC 113434
Dysosmobacter segnis	sp. nov.	seg'nis. L. masc. adj. *segnis*, slow, inactive, lazy, referring the slow growth of the strain	BX15^T from human feces	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	4P-15^T from human feces	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.52500 /NBRC 113913
Flintibacter faecis	sp. nov.	fae'cis. L. gen. fem. n. *faecis*, of faeces, from which the organism was isolated	BX5^T from human feces	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are rod-shaped with spiky ends, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.31644 /KCTC 15861
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	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32825 /KCTC 25134
Lawsonibacter faecis	sp. nov.	fae'cis. L. gen. fem. n. *faecis*, of faeces, from which the organism was isolated	NSJ-52^T from human feces	Cells are club-shaped rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32826 /KCTC 25135
Lawsonibacter celer	sp.	ce'ler. L. masc. adj.	NSJ-47^T from	Cells are	CGMCC

	nov.	*celer*, rapid, pertaining to fast growth of the strain	human feces	straight rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 1-3 days.	1.32820
Neobittarella (ex Bilen et.al, 2018)	nom. rev.	Neo.bit.ta.rel'la N.L. fem. n. *Neobittarella*, in honor of microbiologist Fadi Bittar [110]	Type species: *Neobittarella massiliensis*		
Neobittarella massiliensis (ex Bilen et.al, 2018)	nom. rev.	mas.si.li.en'sis L. masc./fem. adj. *massiliensis*, referring to Marseille, where the organism was isolated [110]	NSJ-65^T from human feces	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are rod-shaped with straight spiky ends, motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are ovoid or rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.31611 /KCTC 15862
Ruminococcus lentus	sp.nov	len'tus. L. masc. adj. *lentus*, slow, referring to the slow growth of the type strain; Nevertheless, *Ruminococcus*	NSJ-14^T from human feces	Cells are spherical, non-motile. Growth in modified MGAM	CGMCC 1.52640 /KCTC 15952

		bicirculans [111] strain 80/3, which shares 16S rRNA identity of 99.77% with the strain NSJ-14, was effectively but not validly published, and the nomenclature was incorrect, thus we propose *Ruminococcus lentus* as the valid species name.		medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	
Ruminococcus intestinalis	sp. nov.	in.tes.ti.na'lis. N.L. masc. adj. *intestinalis*, pertaining to the intestine habitat	NSJ-71^T from human feces	Cells are spherical or ovoid, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are straight rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32818
Romboutsia faecis	*sp. nov.*	*fae'cis L. gen. fem. n. faecis, referring to faecal origin*	NSJ-18^T from human feces	Cells are curved or straight rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.31399
Intestinimonas massiliensis (ex Durand et. al, 2017)	nom. rev.	mas.si.li.en'sis. L. fem. adj. *massiliensis*, of Massilia, the Latin name of Marseill, where the bacteria was for the first time isolated [112]	NSJ-30^T from human feces	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-	CGMCC 1.32803 /KCTC 25082

				7.5, in 3-10 days.	
Hydrogeniiclostridium 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	Cells are curved or straight rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	CGMCC 1.32814 /KCTC 25093
Catenibacterium faecis	sp. nov.	fae'cis. L. gen. fem. n. *faecis*, of faeces, from which the organism was isolated	NSJ-22^T from human feces	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are rod-shaped, motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are ovoid or rod-shaped with spiky ends, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are ovoid, motile. Growth in modified MGAM medium occurs at 37	CGMCC 1.32835 /KCTC 25147

				°C, pH 7.0-7.5, in 3-10 days.	
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	Cells are spherical (tetracocci), non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 3-10 days.	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	Cells are spherical, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in 2-3 days.	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	Cells are straight rod-shaped, non-motile. Growth in modified MGAM medium occurs at 37 °C, pH 7.0-7.5, in1-3 days.	CGMCC 1.32849

Figures

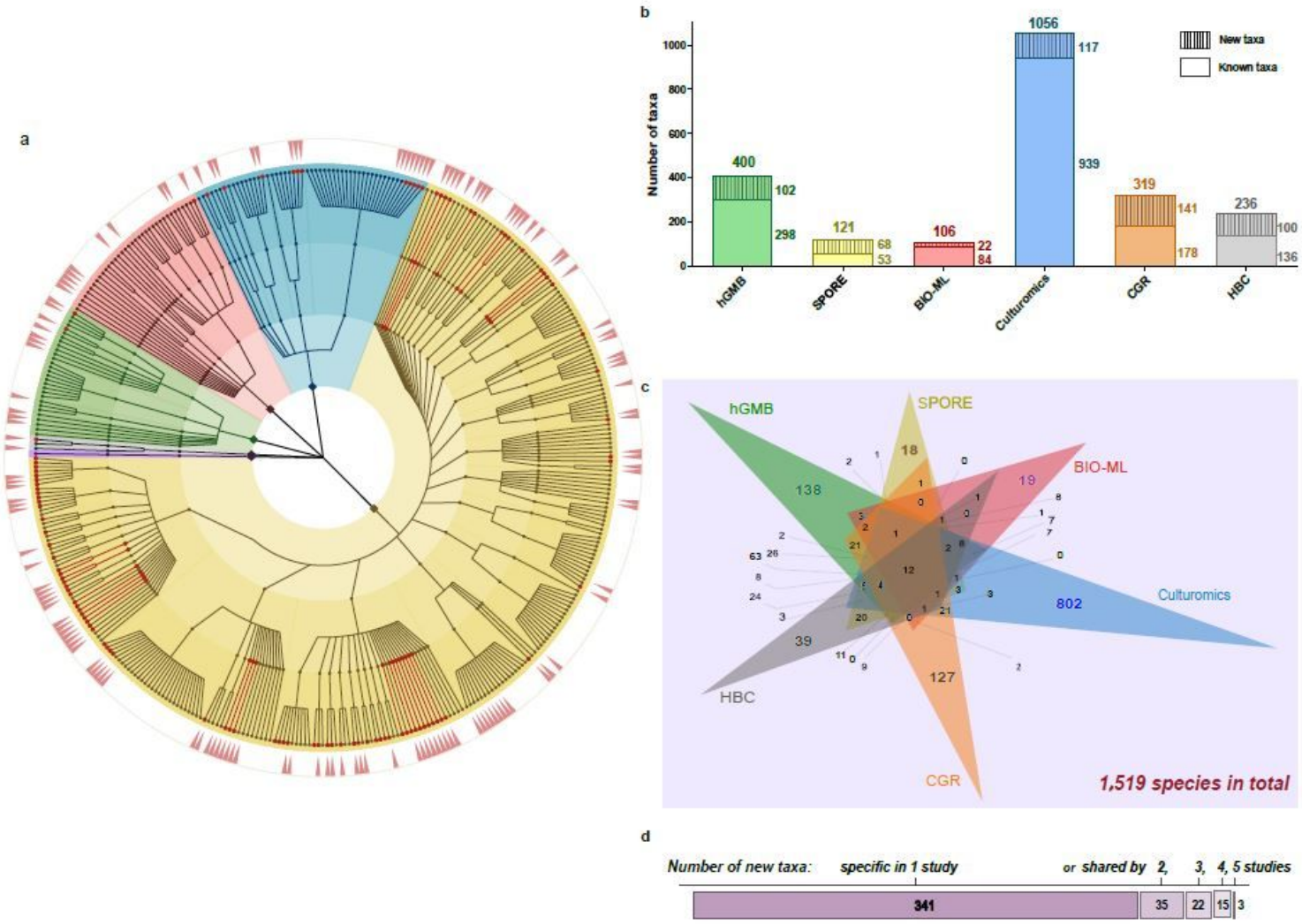

Figure 1

The taxonomic diversity and specificity of hGMB. (a) The taxonomic cladogram displaying the taxonomic diversity of hGMB. The nodes of 102 newly-characterized species, 28 novel genus and 3 novel family are indicated in red. 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 138 species that are solely covered by hGMB. (b) The taxonomic diversity of gut microbes from different gut microbial collections. hGMB (this study): a human gut microbial culture collection constructed in this study contains 400 species with 102 novel taxa; SPORE [23]: a human gut microbial culture collection constructed in 2016 comprises 121 species including 68 novel-taxon candidates; BIO-ML [21]: a human gut microbial culture collection constructed in 2019 comprises 106 species with 20 novel-taxon candidates; Culturomics [25]: the culturomics study of human gut microbes in 2016 reveal the discovery of 1,056 species including 247 novel taxa, of which 117 were still new by the time of manuscript preparation; CGR [20]: a human gut microbial culture collection constructed in 2019 comprises 319 species determined based on the 16S rRNA gene sequence clustering at identity of 98.7%, of which 141 taxa are potentially novel; HBC [24]: a human gut microbial culture collection constructed in 2019 contains 236 species with 100 potentially-novel taxa. (c) The Venn diagram displaying the unique and shared taxa in each study. The numbers of taxa uniquely in one

collection or shared by different studies are labeled in the panel. (d) Summary of novel taxa claimed by 1 or more than 1 study. Numbers in the bar represent the number of novel taxa.

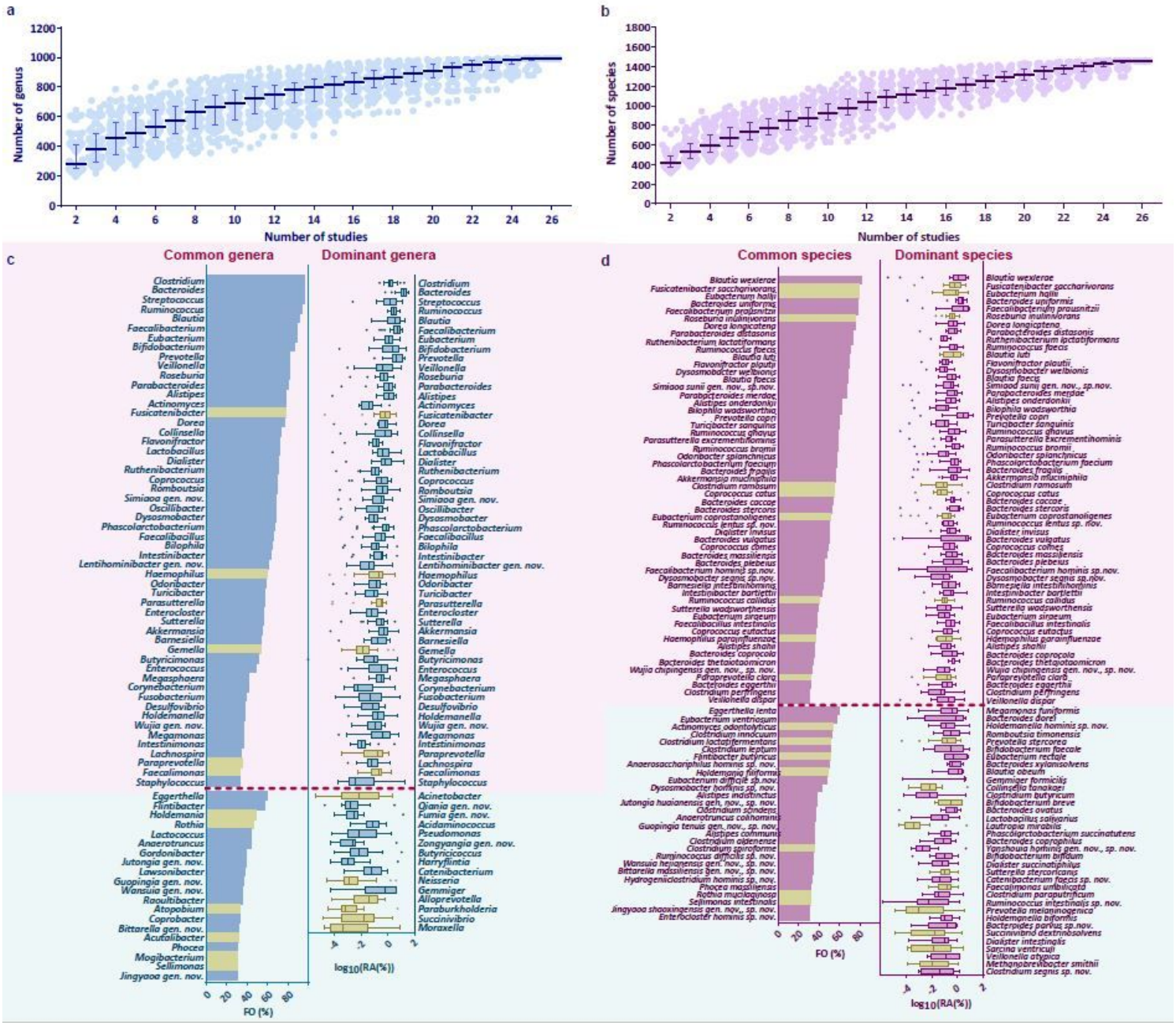

Figure 2

The cultivable recovery of major composition of human gut microbiota by hGMB at genus and species levels. (a) and (b) The rarefaction curves displaying the increase trend of the numbers of assigned genera (a) and species (b) as 1 to 26 16S rRNA gene amplicon datasets (Table S8) were sampled for combined analysis. (c) The coverage of human gut common (bar chart) and dominant (box-and-whiskers plot) genera by hGMB. All the genera covered by hGMB were colored in blue while the genera absent in hGMB were color in olive brown. (d) The coverage of human gut common (bar chart), dominant (box and whiskers plot) species by hGMB. All the genera covered by hGMB were colored in purple while the genera absent in hGMB were color in olive brown. Common genera/species: genera/species with equally-weighted average frequency of occurrence (FO) >30% (definition: 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 the mean value of the average FOs of the 26 analyzed studies); Dominant genera/species: genera/species with equally-weighted average relative abundance (RA)>0.1% (log 10 (RA(%)) >−1) (definition: The equally-weighted average RA is the mean value of the average RAs of the 26 analyzed studies). The light-pink background in panel c and d highlighted the core genera/species shared by both dominant and common genera/species, while the light-blue background marked out the taxa presenting uniquely in either dominant or common genera/species. The bar chart in panel c and d shows the mean values of the 26 FO averages (%), while the box-and-whiskers plot shows the Log 10 of average Ras (%) of each taxon in each study, center line: median, bounds of box: quartile, whiskers: Tukey extreme.

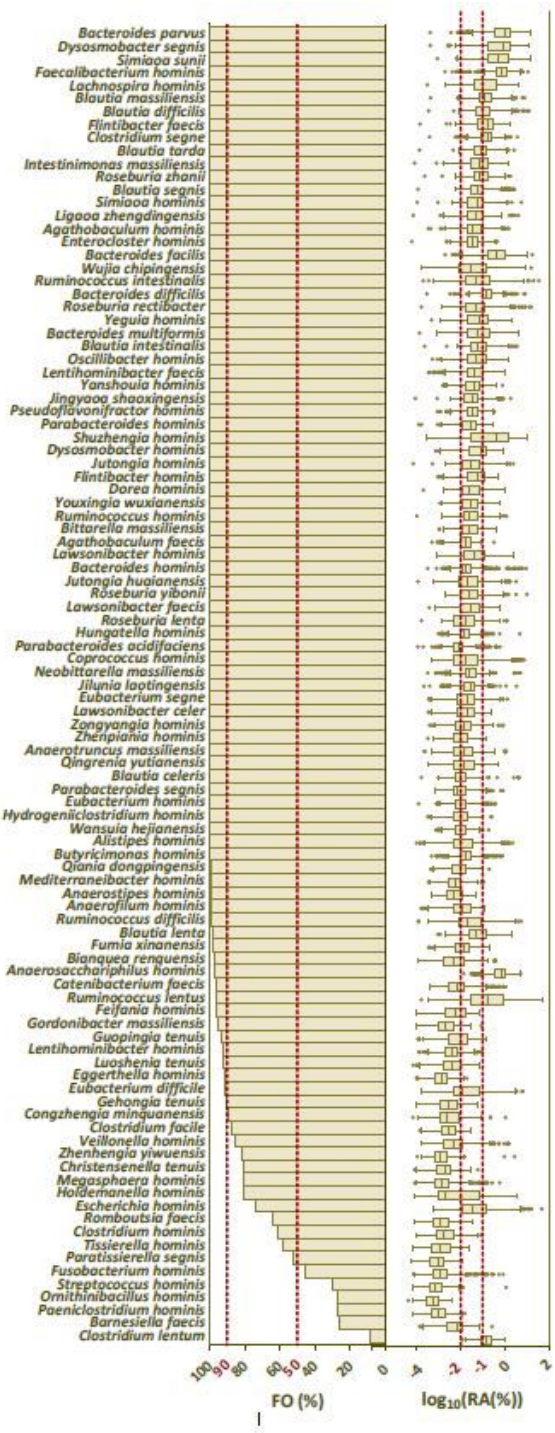

Figure 3

The prevalence of novel taxa in hGMB among global health human gut metagenomes (n=1,129). The bar charts demonstrated the frequency of occurrence (FO) of each novel taxa among 1,129 analyzed health human gut metagenomes (Table S9) (definition: 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 box-and-whiskers plot displayed the relative abundance (RA) of each novel taxa among all samples in Log 10 format. center line: median, bounds of box: quartile, whiskers: Tukey extreme.

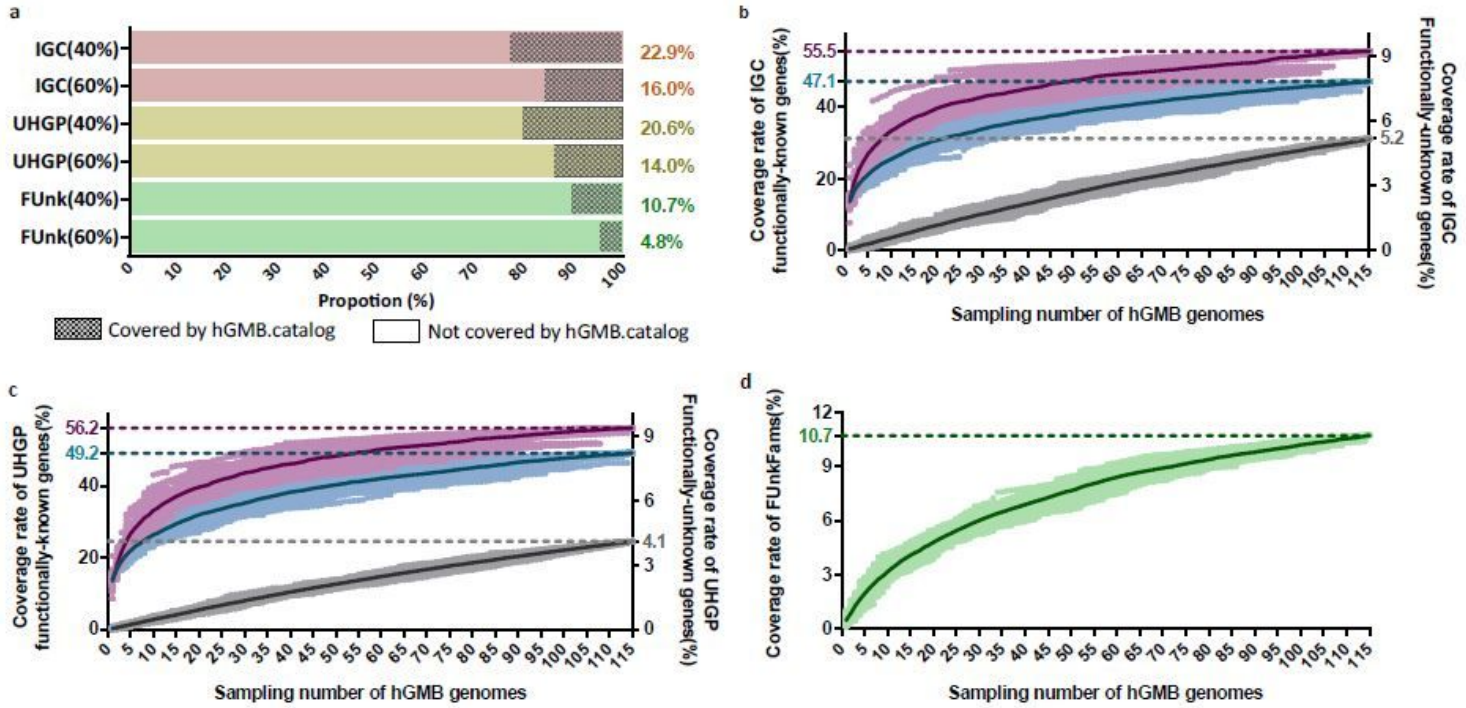

Figure 4

The functional coverage of global human gut gene catalogs by new hGMB genomes. (a) The coverage of IGC [41], UHGP [8] and FUnkFams [42] by hGMB.catalog. The hGMB.catalog was constructed by extraction of 341,876 nonredundant genes from 115 newly-sequenced genomes in hGMB and was BLAST analyzed against subject gene catalog IGC (pink bars), UHGP (yellow bars) and FUnkFams (green bars) with cut-off sequence identities of 40% and 60%, respectively. The y-axis names indicated the names of subject gene catalog and the sequence identities used for BLAST (in bracket). The coverage rates were listed in panel on the right side of each bar. (b),(c) The rarefaction curves displaying the accumulatively-increased coverage of the 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 coverage rates of KO functions of IGC or UHGP gene catalogs when specified numbers of genomes were randomly sampled from 115 hGMB genomes; Dark purple line: the mean coverage rate of KO functions; Light blue dot: the coverage rates of GO functions of IGC or UHGP gene catalogs; Dark blue line: the mean coverage rate of GO functions; Grey dot: the coverage rates of unannotated genes of IGC or UHGP; Black line: the mean coverage rate of unannotated genes of IGC or UHGP. (d) The rarefaction curves displaying the accumulatively-increased coverage of conserved functionally-unknown proteins in FUnkFams. The sampling was repeated for 50 times at each x-axis point; Light green dot: the coverage rates of FUnkFams proteins when sampled randomly; Dark green line: the mean value of the coverage rates.

Supplementary Files

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

- FigureS1final.pdf
- SupplementaryData1.final.doc
- TableS10new.xlsx
- TableS11new.xlsx
- TableS12new.txt
- TableS13new.txt
- TableS14new.txt
- TableS1new.xlsx
- TableS2new.xlsx
- TableS3new.xlsx
- TableS4final.xlsx
- TableS5new.xlsx
- TableS6new.xlsx
- TableS7new.xlsx
- TableS8new.xlsx
- TableS9new.xlsx