AMERICAN SOCIETY FOR MICROBIOLOGY | mSphere®

RESEARCH ARTICLE
Host-Microbe Biology

An Expanded Gene Catalog of Mouse Gut Metagenomes

Jiahui Zhu,[a,b] Huahui Ren,[b,c] Huanzi Zhong,[b,c] Xiaoping Li,[b,c] Yuanqiang Zou,[b,c] Mo Han,[b,c] Minli Li,[a] Lise Madsen,[b,c,d] Karsten Kristiansen,[b,c,e] Liang Xiao[b,e,f,g]

[a]State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, China
[b]BGI-Shenzhen, Shenzhen, China
[c]Laboratory of Genomics and Molecular Biomedicine, Department of Biology, University of Copenhagen, Copenhagen, Denmark
[d]Institute of Marine Research, Bergen, Norway
[e]Qingdao-Europe Advanced Institute for Life Sciences, BGI-Shenzhen, Qingdao, China
[f]Shenzhen Engineering Laboratory of Detection and Intervention of Human Intestinal Microbiome, BGI-Shenzhen, Shenzhen, China
[g]BGI College & Henan Institute of Medical and Pharmaceutical Science, Zhengzhou University, Zhengzhou, China

Jiahui Zhu and Huahui Ren have contributed equally to this work. Author order was determined on the basis of seniority.

ABSTRACT High-quality and comprehensive reference gene catalogs are essential for metagenomic research. The rather low diversity of samples used to construct existing catalogs of the mouse gut metagenome limits the numbers of identified genes in existing catalogs. We therefore established an expanded catalog of genes in the mouse gut metagenome (EMGC) containing >5.8 million genes by integrating 88 newly sequenced samples, 86 mouse gut-related bacterial genomes, and 3 existing gene catalogs. EMGC increases the number of nonredundant genes by more than 1 million genes compared to the so-far most extensive catalog. More than 60% of the genes in EMGC were assigned to *Bacteria*, with 54.20% being assigned to a phylum and 35.33% to a genus, while 30.39% were annotated at the KEGG orthology level. Nine hundred two metagenomic species (MGS) assigned to 122 taxa are identified based on the EMGC. The EMGC-based analysis of samples from groups of mice originating from different animal providers, housing laboratories, and genetic strains substantiated that diet is a major contributor to differences in composition and functional potential of the gut microbiota irrespective of differences in environment and genetic background. We envisage that EMGC will serve as a valuable reference data set for future metagenomic studies in mice.

IMPORTANCE We established an expanded gene catalog of the mouse gut metagenome not only to increase the sample size compared to that in existing catalogs but also to provide a more comprehensive reference data set of the mouse gut microbiome for bioinformatic analysis. The expanded gene catalog comprises more than 5.8 million unique genes, as well as a wide range of taxonomic and functional information. Particularly, the analysis of metagenomic species with the expanded gene catalog reveals a great novelty of mouse gut-inhabiting microbial species. We envisage that the expanded gene catalog of the mouse gut metagenome will serve as a valuable bioinformatic resource for future gut metagenomic studies in mice.

KEYWORDS diet, gene catalog, metagenomic species, mouse gut metagenome

Mice are among the most widely used animal models for biomedical studies to decipher the complex interplay between the gut microbiota and host phenotypes (1–4). Amplicon sequencing of the 16S rRNA gene has been widely used for analyses of the gut microbiota due to low costs and short analysis cycles. However, the taxonomic information is, in most cases, limited to the genus level, and amplicon sequencing generally provides limited information on function (5, 6). A key to the use of mouse models

Citation Zhu J, Ren H, Zhong H, Li X, Zou Y, Han M, Li M, Madsen L, Kristiansen K, Xiao L. 2021. An expanded gene catalog of mouse gut metagenomes. mSphere 6:e01119-20. https://doi.org/10.1128/mSphere.01119-20.

Editor Maria L. Marco, University of California, Davis

Address correspondence to Liang Xiao, xiaoliang@genomics.cn.

Received 10 November 2020
Accepted 31 January 2021
Published 24 February 2021

for detailed functional analyses of the gut microbiota is the availability of comprehensive catalogs of microbial genes and derived metagenomic species (MGSs)/metagenome-assembled genomes (MAGs). The first catalog of genes in the mouse gut microbiome included 2.6 million nonredundant genes from fecal samples of 184 mice (7). Subsequent studies further explored the diversity and functional potential of the mouse gut microbiota by isolating and sequencing an increasing number of bacterial strains from the mouse gut (8–11) and establishing a mouse intestinal bacterial collection (miBC), depositing bacterial strains and associated genomes from the mouse gut (9). Recently, Lesker et al. generated an integrated mouse gut metagenome catalog (iMGMC), comprising 4.6 million unique genes and 830 high-quality MAGs, and by linking MAGs to reconstructed 16S rRNA gene sequences, they provided a pipeline enabling improved prediction of functional potentials based on 16S rRNA gene amplicon sequencing (12).

Here, we constructed an expanded mouse gut metagenome catalog (EMGC) by integrating 3 published gene catalogs, including the gene catalog of the mouse gut metagenome (MGGC) released in 2015 comprising 2,571,074 genes (7), a feed and diet gene catalog for mice (FDGC) (13), the integrated mouse gut metagenome catalog (iMGMC) (12), 72 available sequenced mouse gut-related bacterial genomes (8–11), 14 high-quality genomes assembled from published sequencing data of isolates (9), and 88 newly shotgun-sequenced samples. Our new nonredundant reference gene catalog comprises 5,862,027 genes and was annotated by NR (released on 5 January 2019) and KEGG (release 87) databases (14). Finally, we generated 902 MGSs from the gene abundance profiles for 326 laboratory mice of EMGC and compared these MGSs with the high-quality MAG collection (12) and the recent collection of bacteria isolated from the mouse gut (11). By combining these individual data sets, we increased the number of sequenced bacterial genes of the mouse gut microbiome by more than 1 million and significantly increased the mapping ratio of reads obtained by shotgun sequencing of samples from the mouse gut and fecal samples, providing a resource for future studies on the mouse gut microbiota.

(This article was submitted to an online preprint archive [15].)

RESULTS

Construction and evaluation of EMGC. Fecal samples from 88 C57BL/6J male mice were sequenced using the BGISEQ-500 platform providing 1,098-Gb high-quality host-free data with an average of 12.47 Gb per sample (see Table S1A in the supplemental material) and a catalog comprising 2,602,584 nonredundant genes (PMGC). We next used 72 mouse gut-related bacterial genomes (8–11) from IMG and NCBI RefSeq and 14 high-quality genomes (completeness >90% and contamination <5%) assembled from reads accessible from PRJEB10572 (9) (Table S1B) to generate a mouse gut cultured bacterial gene set (MiCB). All gene catalogs, including MGGC (7) together with FDGC (13) and iMGMC (12), downloaded from GigaDB and the Zenodo repository (Table S1C), respectively, were integrated to construct an expanded nonredundant mouse gut bacterial gene catalog (EMGC) (Fig. 1). The expanded catalog comprises 5,862,027 genes, which is more than twice the number of genes in the MGGC (7) and 1 million genes more than the iMGMC catalog (12) (Table 1). Thus, 18.93% of the genes in EMGC are not represented in either the iMGMC or MGGC (see Fig. S1).

To compare the performance of EMGC with that of MGGC and iMGMC, we mapped sequencing reads from the FDGC, MGGC, and PMGC studies to the three catalogs. Of the sequencing reads from PMGC, which is part of EMGC, 55.72% were mapped to MGGC and 56.56% to the iMGMC. In contrast, the EMGC allowed mapping of 79.52% of the reads (Fig. 2A), close to the maximum achievable mapping rate in prokaryotes (16).

A comparison of mapping rates of reads from 326 fecal samples obtained from different mouse strains and providers (Table S1D) to those from EMGC demonstrated that mice from the laboratory animal center at Sun Yat-Sen University exhibited a lower mapping rate than samples from the other providers, where the median mapping rates

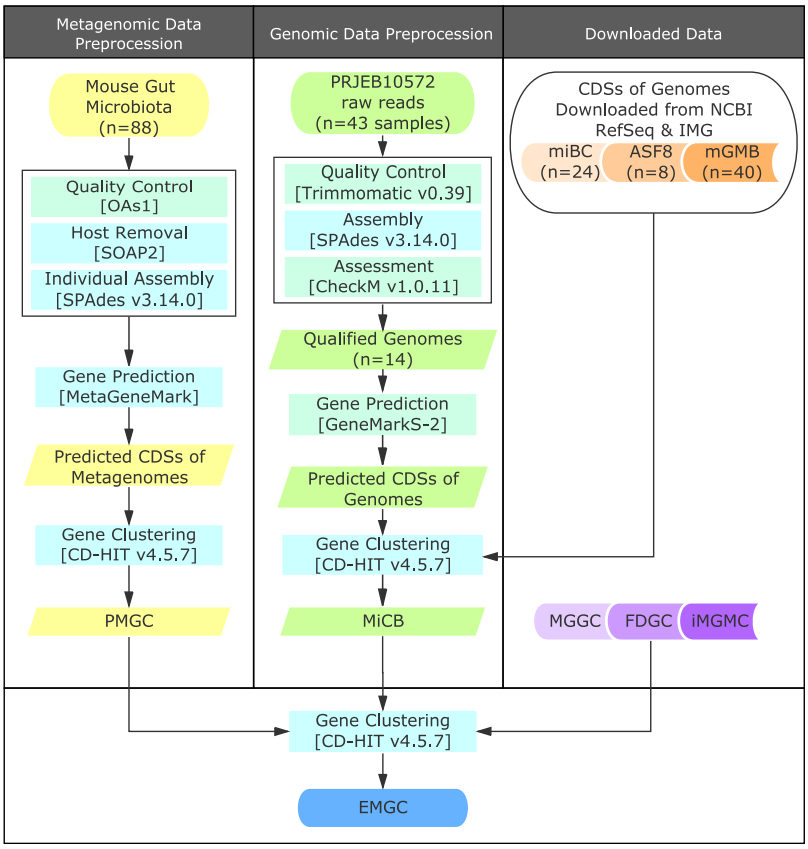

FIG 1 Construction of the EMGC. Metagenomic sequencing data of 88 mouse gut metagenomes were processed by the pipeline as displayed to generate nonredundant genes for PMGC. Unassembled strains of miBC (under BioProject PRJEB10572) were assembled and filtered by genome quality (completeness, >90%; contamination, <5%) of assembled genomes. Qualified genomes were used for gene prediction. CDSs from assembled genomes and downloaded genomes were gathered and clustered to MiCB. PMGC and MiCB along with 3 downloaded gene sets, FDGC, MGGC, and iMGMC, were merged to generate EMGC.

were higher than 80% (Fig. 2B). Median mapping rates of reads obtained from all mouse strains were also higher than 80% (see Fig. S2A). Richness estimated by Chao2 indicated that our EMGC covered 98.24% of the genes in the 326 fecal samples (Fig. S2B), whereas the incidence-based coverage estimator (ICE) suggested that 97.49% of the genes were covered.

TABLE 1 General features of gene catalogs

Catalog	Sample size (*n*)	Total no. of ORFs[a]	Length (bp)					
			Total	Avg	Max	Min	N_{50}	N_{90}
PMGC[b]	88	2,602,584	1,920,079,578	737.76	120,489	102	981	396
MGGC[c]	184	2,572,074	1,959,483,705	761.83	120,489	102	1,014	408
FDGC[d]	54	793,847	585,096,360	737.04	23,610	102	978	396
MiCB[e]	NA	267,801	251,087,538	937.59	79,287	100	1,206	504
iMGMC[f]	292	4,499,720	3,505,479,714	779.04	120,399	102	1,107	390
EMGC[g]	434	5,862,027	4,542,473,508	774.90	120,489	100	1,104	393

[a]ORFs, open reading frames.
[b]PMGC, a sub-mouse gut gene catalog from 88 mouse gut metagenomes.
[c]MGGC, gene catalog of mouse gut metagenome released in 2015.
[d]FDGC, feed and diet gene catalog for mice.
[e]MiCB, mouse intestinal cultured bacteria gene set.
[f]iMGMC, integrated mouse gut metagenome catalog.
[g]EMGC, an expanded gene catalog of mouse gut metagenomes.

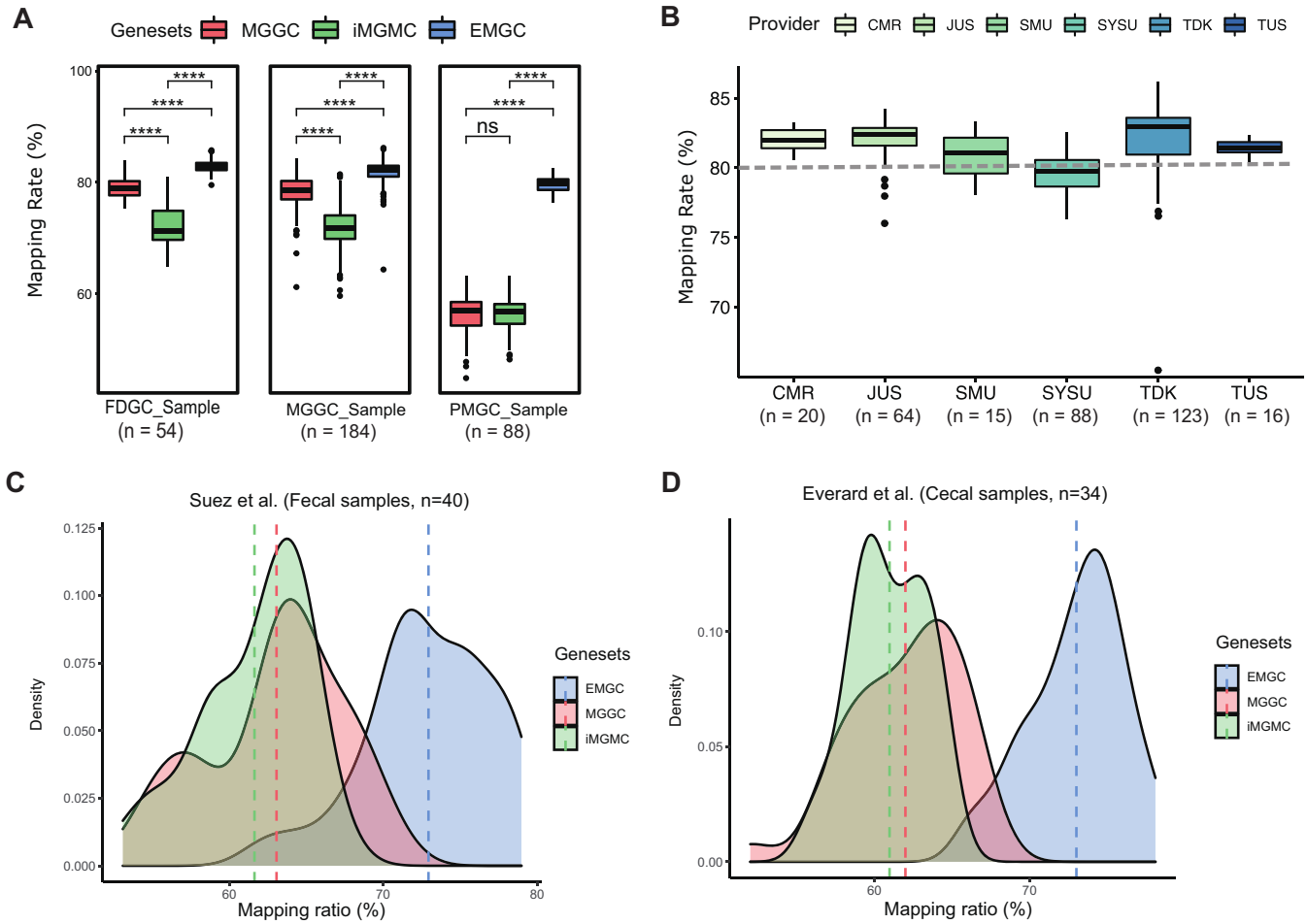

FIG 2 Performance of the EMGC. (A) Comparison of mapping rates between MGGC, iMGMC, and EMGC. ****, BH-adjusted *P* value of <0.0001 by Wilcox rank sum test. (B) Display of mapping rates among samples' providers, including the Wallenberg Laboratory for Cardiovascular and Metabolic Research (CMR), the Jackson Laboratory in the United States (JUS), the laboratory animal center of Southern Medical University (SMU), the laboratory animal center of Sun Yat-Sen University (SYSU), and Taconic in Denmark (TDK) and in the United States (TUS). Dashed line represents a mapping rate of 80%. (C and D) Density curves for the mapping rates of fecal metagenomes and cecal metagenomes which were not included in gene catalog construction. Dashed lines represented the average values of the mapping rate of each gene catalog.

To further evaluate the quality of the EMGC, we mapped the metagenomic data obtained from 40 fecal samples from control mice and mice that had consumed noncaloric artificial sweeteners (17) and metagenomic data obtained from 34 cecal samples from control mice and mice treated with prebiotic (18) (Table S1C). For the reads obtained in the study of Suez et al. (17), 63.07% mapped to the MGGC and 61.04% to iMGMC, whereas 72.94% mapped to the EMGC (Fig. 2C; Table S1E). For the reads obtained from the study by Everard et al. (18), 61.99% mapped to MGGC and 60.97% mapped to iMGMC, but 73.01% of the reads mapped to the EMGC (Fig. 2D; Table S1F). Together, these results demonstrate a significantly increased mapping rate of reads using the EMGC as a reference.

Taxonomic and functional characteristics of EMGC. We taxonomically annotated the genes of EMGC using Kaiju (19) and the NCBI NR database to provide an overview of the taxonomical composition visualized by a Krona plot (20). This plot revealed that 67% of the genes were able to be annotated (see Fig. S3). We assigned 54.20% of the genes to the phylum level and 44.30% of the genes to the family level (see Fig. S4A and B). We next annotated the genes in the EMGC to the KEGG (release 87) database (14) and identified 6,704 KEGG functional orthologs (KO) and 290 KEGG pathways (see Fig. S5).

To further examine the quality of the EMGC, we calculated the occurrence frequency and average abundance of the 1,109,381 genes not present in the previous

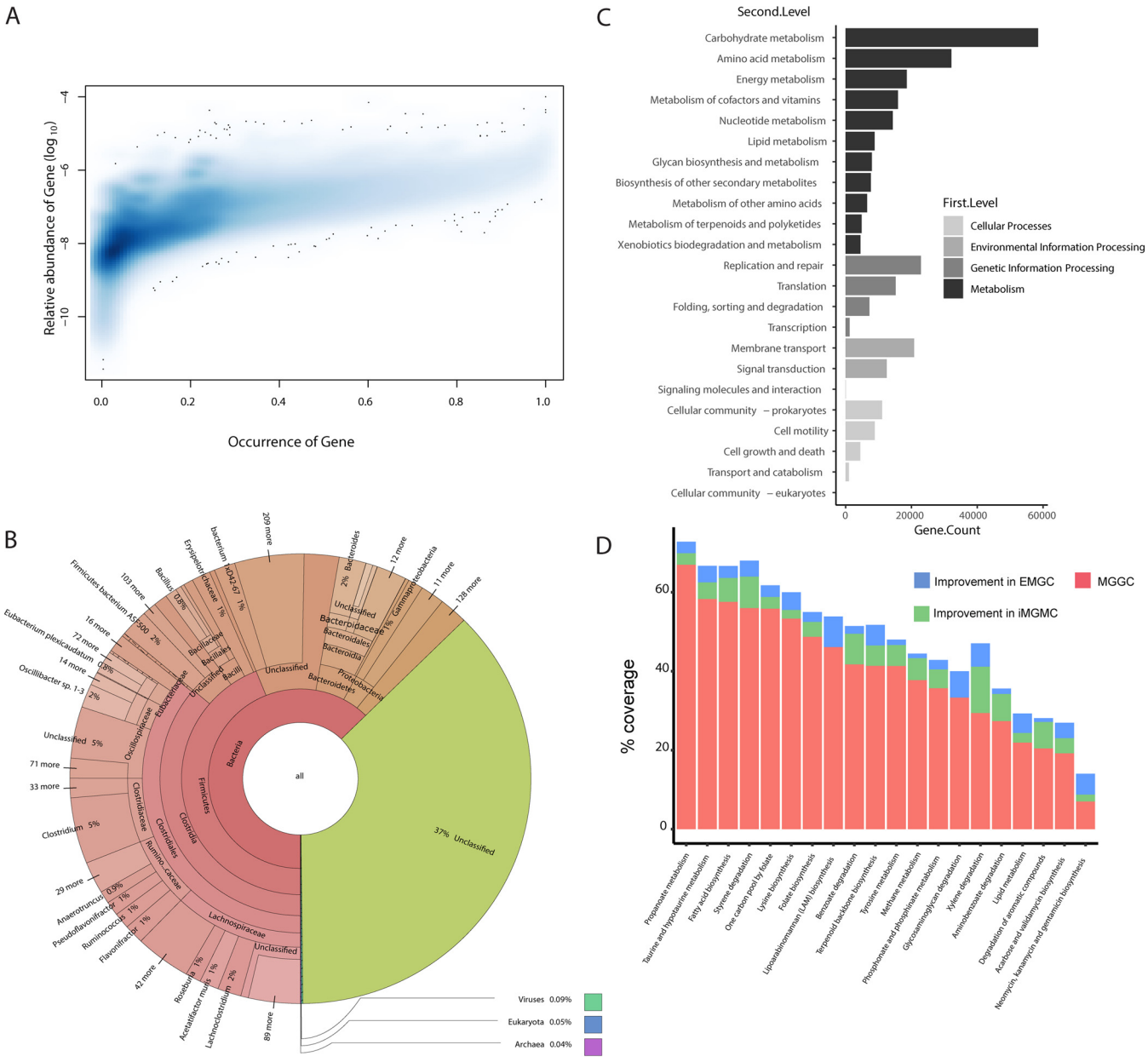

FIG 3 Description of new genes included in EMGC. (A) Two-dimensional (2D) density histogram showing the distribution of occurrences and mean relative abundances of new genes. (B) General display of taxonomic composition of new genes by Krona. (C) Frequency of functional pathways associated with the new genes. (D) Stacked histogram of KO coverage of functional pathways improved in EMGC compared to that in MGGC and iMGMC. Coverage is calculated as [(annotated KO numbers)/(total KO numbers)] × 100 in a given pathways.

MGGC and iMGMC. As shown in Fig. 3A, 40.41% of these genes exhibited an occurrence frequency and mean abundance higher than 0.1 and 10^{-8}, respectively. We also extracted taxonomic and functional information of these new genes. Annotation of the new genes at the species level revealed that the top 5 species could be assigned to *Oscillibacter* sp. 1-3, *Firmicutes* bacterium ASF500, *Acetatifactor muris*, bacterium 1xD42-67, and *Eubacterium plexicaudatum* (Fig. 3B), all isolated from the mouse gut based on information from the NCBI BioSample database. In relation to functions, the general distribution of KEGG pathways in these additional genes is similar to the overall distribution in EMGC (Fig. 3C). We identified 189 KOs in EMGC which are not present in either MGGC or iMGMC. Furthermore, 42 KEGG pathways are covered by additional KOs (Table S1G) in EMGC. For 5 KEGG pathways, including lipoarabinomannan (LAM) biosynthesis, glycosaminoglycan degradation, xylene degradation, neomycin, kanamycin

and gentamicin biosynthesis, and terpenoid backbone biosynthesis, we found that more than 5% of the additional KOs are only represented in EMGC compared to that in iMGMC (Fig. 3D; Table S1G).

Changes in the microbiota composition and functional potential. We reported earlier that the mouse gut metagenome is affected by animal providers and housing as well as strain and diet (7). To investigate to what extent diet affected the gut metagenomes independently of strain and providers, we selected 7 groups (G1 to G7) of samples from different strains from different providers fed a low-fat (LF) diet or a high-fat (HF) diet and housed in the same facility (see details in Table S1D). We estimated the impact of diet on the variation of gut microbiota based on the relative abundance profiles of genera and KOs using a permutational multivariate analysis of variance (PERMANOVA). The analyses indicated that diet explained at least 33.9% (P value = 0.003) of the total variation at the genus level and 47.3% (P value = 0.006) at the KO level (Fig. 4A; see also Fig. S6A). Compared to that for mice fed an LF diet, mice fed an HF diet exhibited an increase in alpha diversity at the genus level independent of housing laboratories, strains, and providers (Fig. 4B). In contrast, at the KO level, alpha diversity in mice fed an LF diet generally, except for group 7, exhibited an increased diversity compared to that for mice fed an HF diet (Fig. S6B). Principal-coordinate analysis (PCoA) similarly confirmed that the diet strongly influenced the genus profile (Fig. 4C) and the KO profile (Fig. S6C).

To further examine diet-induced changes, we examined genera and KOs enriched in samples from either HF- or LF-diet-fed mice by Wilcoxon rank sum test. As shown in Fig. 4D, in all 7 groups, the 36 genera found at higher abundance in samples from HF-diet-fed mice belong to *Firmicutes*, whereas four genera within the *Bacteroidetes* phylum and one genus within the *Fibrobacteres* phylum were found at higher abundance in samples from LF-diet-fed mice (Fig. 4D; Table S1H). However, whereas 363 KOs were enriched in LF-diet-fed mice, only 270 KOs were detected at higher abundance in HF-diet-fed than in LF-diet-fed mice (Table S1I). To investigate which taxa contributed to the disparate response to LF and HF diet at the taxonomy and the functional levels, we identified the taxa at the phylum level that contributed to the enrichment of KOs. Whereas genera within the *Bacteroidetes* phylum were the main contributor accounting for 6.02% of the KOs enriched in LF-diet-fed mice, genera within the *Firmicutes* phylum accounted for 3.96% of the KOs enriched in HF-diet-fed mice (see Fig. S7).

Construction of metagenomic species. We identified 902 metagenomic species (MGSs; >700 genes) using the relative gene abundances based on 326 fecal samples obtained from different mouse strains and providers using MGS canopy clustering and taxonomic annotation as described previously (21). The 902 MGSs were assigned to 122 taxa (see Fig. S8; Table S1J). We also generated MGS profiles for the 7 groups of mice fed an LF or an HF diet. The Shannon indices and PCoA plot revealed a clear effect of diet, independent of mouse strain and provider (see Fig. S9A and B).

We next compared the 902 MGSs with the 830 high-quality nonredundant MAGs generated in the iMGMC project (12) and 115 bacterial genomes from the mouse gut microbial biobank (mGMB) project (11). Five hundred fifty-nine (61.97%) MGSs were classified as the same species as the MAGs from the iMGMC project (maximal unique match index [MUMi] value [22–24] >0.54) (Table S1J). As shown in Fig. 5, *Firmicutes* and Bacteroides were the most prevalent phyla among all MGSs and MAGs. We also identified 56 MGSs representing genomes of species from the mGMB project, and of these, 8 MGSs could be identified as mGMB genomes, but not as MAGs (Table S1J). Of note, for more than one-third of the MGSs, we were unable to identify corresponding entities in the MAG collection or in the cultured genomes collection.

DISCUSSION

The EMGC represents the most comprehensive catalog of genes in the mouse gut microbiome. It covers samples from feces and cecum from different mouse strains fed different diets, obtained from different providers, and housed in different laboratories.

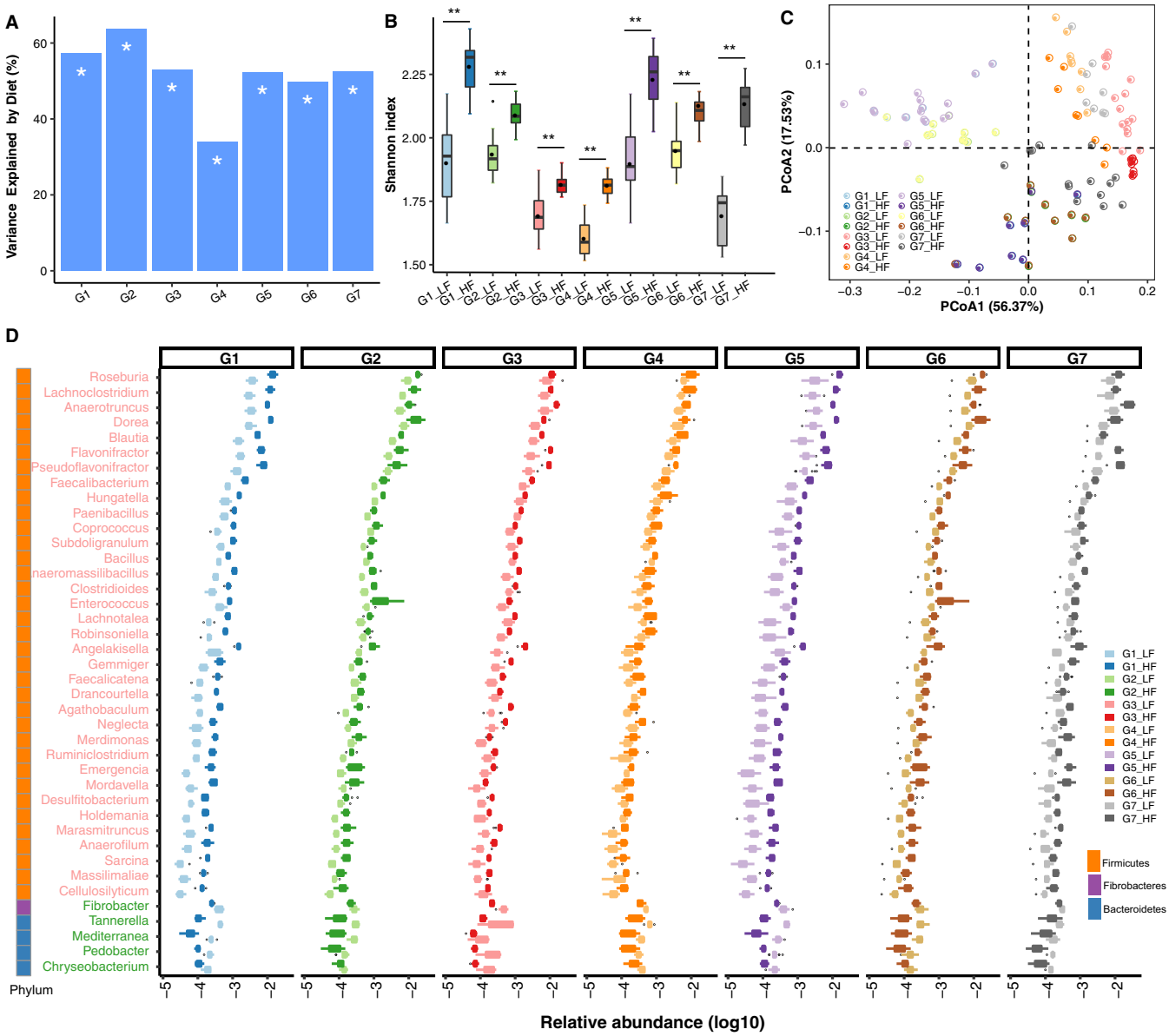

FIG 4 Influence of diet on the composition of the microbiota. (A) PERMANOVA to estimate the influence of diet on the composition of gut metagenomes among all 7 sample groups. G1, C57BL/6 mice provided by Taconic in Denmark (TDK) and hosted in National Institute of Nutrition and Seafood Research of Norway (NIFES); G2, Sv129 mice provided by TDK and hosted in NIFES; G3, C57BL/6 mice provided by the Jackson Laboratory in the United States (JUS) and hosted by Pfizer-I; G4, C57BL/6 mice provided by Taconic in the United States (TUS) and hosted in Pfizer-I; G5, C57BL/6 mice provided by TDK and hosted in the University of Copenhagen (KU); G6, Sv129 mice provided by TDK and hosted in KU; G7, C57BL/6 mice provided by the laboratory animal center of Sun Yat-Sen University (SYSU). *, $P < 0.05$. Shannon index (**, BH-adjusted $P < 0.01$, Wilcox rank sum test) (B) and PCoA based on genus profile (C) for 7 groups fed the HF and LF diets. (D) Genera differently enriched (BH-adjusted $P < 0.05$, Wilcox rank sum test; relative abundance, >1e−5) in mice fed the HF and LF diets among all 7 groups. Genera in light red represent genera enriched in HF-diet-fed mice, while genera in light green represent genera enriched in LF-diet-fed mice.

The majority of the genes identified in this study were assigned to known species, which might improve the coverage of known species and the detection of low-abundant taxa. The improvement in KO coverages of a number of pathways will enhance the functional characterization of the mouse gut microbiota. In addition, the analysis of samples from different mouse strains from different animal providers and different housing laboratories confirms the pronounced effect of diets on the taxonomic and functional composition of the gut microbiota.

In spite of the increased number of genes in EMGC, there are still some limitations. The sample size and variation of sample types in the EMGC are still small. The

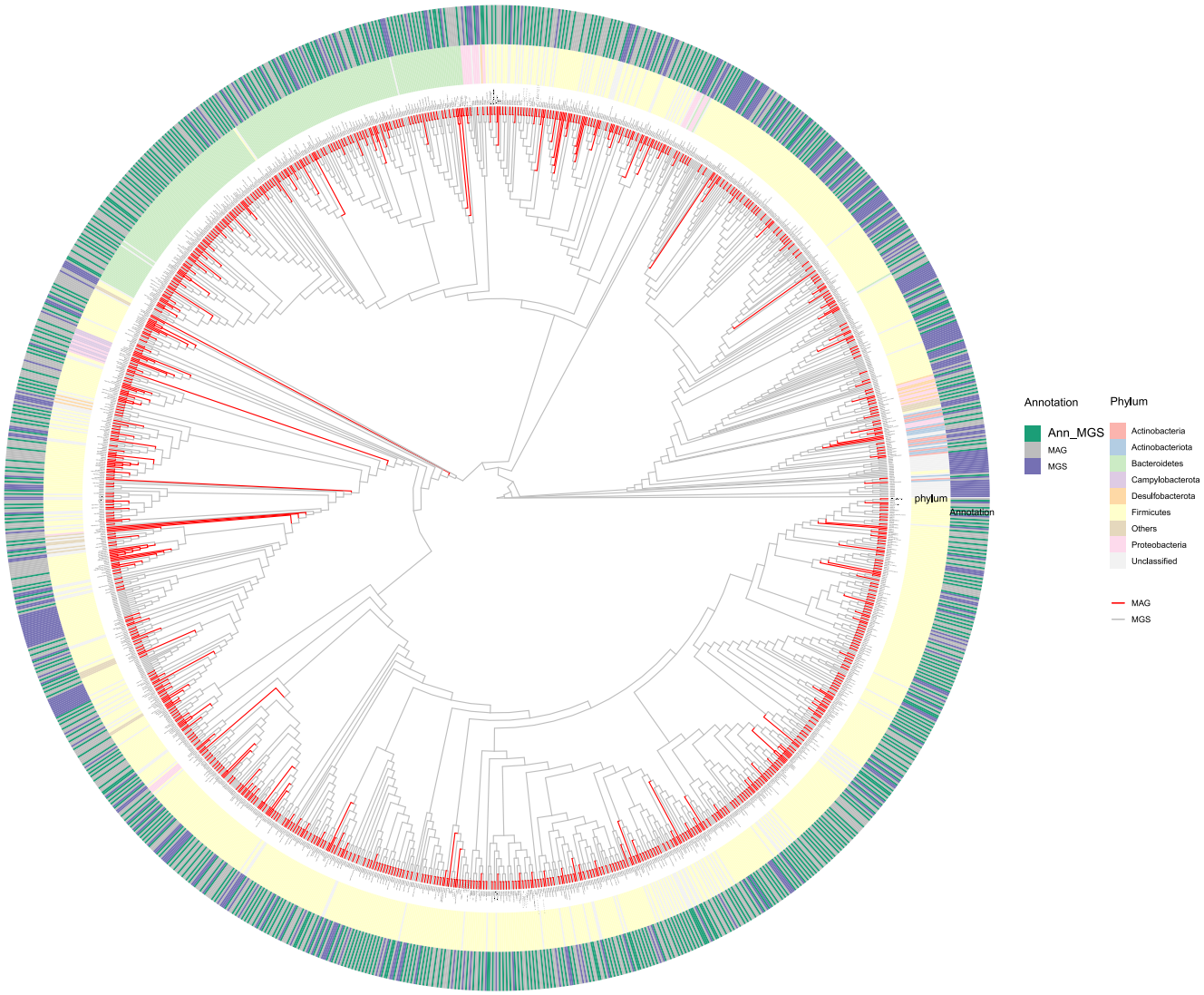

FIG 5 Phylogenetic tree of the 902 MGSs and 830 high-quality iMGMC MAGs. MUMi distances for MGSs and MAGs were used to construct the phylogenetic tree using hierarchical clustering. MAGs are shown as red branches and MGSs as gray branches. The outer ring shows the relation between MGSs and MAGs. MGSs which have a MUMi value of >0.54 are marked as "Ann_MGS" in green blocks, otherwise, in purple blocks. MAGs are all in gray blocks. Colored blocks in the inner cycle indicate phyla assigned to MGSs and MAGs.

majority of samples included in EMGC were collected from C57BL/6 mice, which might affect the applicability in studies on other laboratory mouse strains and wild-caught mice. Many confounding factors in addition to those addressed in the present study will most probably impinge on the gut microbiota (1, 3, 4, 25). It therefore seems important to include more samples from other mouse strains to gain further insights into the effect of confounding factors, which may lead to pronounced variability in the mouse gut microbiota, which again might limit the reproducibility of biomedical research using mouse models (25, 26). Although both culture-independent and culture-dependent studies on the mouse gut microbiota have been carried out to improve the understanding of host-microbe interaction in mouse models, the majority of the mouse gut metagenome members still remain relatively uncharacterized (7, 9, 11, 12). Besides, we also noticed the limitations of the construction approach of EMGC. EMGC is based on metagenome assembly and focused on the gene-level characterization of the mouse gut microbiome. The applications of protein-level metagenome assembler (27), annotation (28), and binning tools

(12, 29–31) are needed in future studies to enhance our understanding of the composition and functions of the mouse gut microbiota.

Even though more and more metagenomic analysis methods are being developed at an increasing speed, gene catalogs are still necessary resources, not only to provide reliable and consistent taxonomic annotation but also to reduce the gap between phylogenetic and functional biases (16, 32). The high-quality reference gene catalog, EMGC, together with the 902 metagenomic species, is able to support and improve accurate metagenome-wide association analyses using mouse models, which may assist in functional characterization of observed correlations between the microbiota composition and the functional potential in relation to host phenotypes.

MATERIALS AND METHODS

Data acquisition. DNA from 88 stool samples of C57BL/6J wild-type male mice, collected from the laboratory of BGI-Wuhan, was extracted and shotgun sequenced using the BGISEQ-500 platform and paired-end 100-bp (PE100) sequencing as described previously (33). An optimized sequencing quality filter for the cPAS-based BGISEQ platform, OAs1 (33), was applied in the quality control step, followed by host removal with SOAP2 (v2.21, parameters: -m 0 -x 1000 -c 0.9) (34) using GRCm38 (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001635.20/) as the reference mouse genome. Individual assembly of metagenomic reads was performed using metaSPAdes v3.14.0 (parameters: -k 49; other parameters were set to the default) (35, 36).

Genes were predicted by MetaGeneMark (v2.7) (37) from metagenome-assembled contigs with a length of >500 bp and filtered by length of >100 bp. Redundant predicted genes were removed by CD-HIT (v4.5.7, parameters: -G 0 -n 8 -aS 0.9 -c 0.95 -d 0 -r 1 -g 1) (38) in order to generate a sub-mouse gut gene catalog (PMGC).

Raw reads of 43 unassembled bacterial genomes from the miBC (EBI project identifier [ID] PRJEB10572) were downloaded from EBI and filtered by Trimmomatic (v 0.39) (39). Draft genomes were assembled separately by SPAdes (-k 29,39,49,69 –careful) (40, 41) and filtered by CheckM (v1.0.13) (42). After assessment using the criteria of completeness of >90% and contamination of <5%, the remaining 14 genomes were used for gene prediction by GeneMarkS-2 (v1.07) (43).

A total of 72 genomes, including 24 sequenced strains of miBC (9), 8 genomes of the altered Schaedler flora (8) (PRJNA175999 to PRJNA176003, PRJNA213740, PRJNA213743), and 40 genomes of mGMB (11) (released before 26 February 2019, PRJNA486904), as well as their coding sequences (CDSs) and translated CDSs were all downloaded from the NCBI RefSeq database and the Integrated Microbial Genomes (IMG) database (44). We gathered CDSs from 86 bacterial genomes and filtered out genes smaller than 100 bp. We clustered CDSs using CD-HIT (v4.5.7, parameters: -G 0 -n 8 -aS 0.9 -c 0.95 -d 0 -r 1 -g 1) (38), establishing a gene catalog termed mouse intestinal cultured bacteria gene set (MiCB). Detailed information on the included genomes is provided in Table S1B in the supplemental material.

All public mouse-related microbial metagenomic data sets used in this study are listed in Table S1C, including (i) 184 host-free sequenced mouse gut microbiomes and the gene catalog of mouse gut metagenome (MGGC) (7), (ii) 54 mouse gut microbiomes and the related gene catalog (FDGC) (13), (iii) 830 high-quality dereplicated MAGs and the iMGMC gene catalog (12), (iv) 40 mouse fecal metagenomes (17), and (v) 34 mouse cecum metagenomes (18).

Construction of EMGC and selection of new genes. All downloaded genes were filtered by length of >100 bp and integrated to construct the EMGC using CD-HIT (v4.5.7, parameters: -G 0 -n 8 -aS 0.9 -c 0.95 -d 0 -r 1 -g 1) (38). The output of EMGC clusters was analyzed to generate a list for new genes in EMGC that are not present in iMGMC and MGGC. Metagenomes were mapped to gene catalogs by SOAP2 (v2.21, parameters: -m 0 -x 1000 -c 0.95) (34). Mapping rates between groups and catalogs were compared by Wilcoxon rank sum test (R ggpubr package). *P* values were adjusted by using the Benjamini-Hochberg method. The profile of relative gene abundances for the 326 laboratory mice (see Table S1D for an overview of these mice) was calculated based on the method of Qin et al. (45) using EMGC. Richness estimation by the Chao2 index and incidence-based coverage estimator (ICE) was calculated based on the gene abundance profiles (16). The occurrence and average abundance of new genes were calculated using the relative gene abundance profiles.

Taxonomic and functional annotation. Genes predicted from metagenome assemblies were taxonomically annotated by Kaiju (v1.6.3) (19) using the NCBI-NR database (released on 5 January 2019) and the parameters of the program were set to "-a greedy -e 5 -E 0.01 -v -z 4 -s 65." For genes from mouse gut-related bacterial genomes, we kept the original taxonomic information of the genomes and assigned them to the corresponding genes. All genes were searched against KEGG (version 87) (14) by DIAMOND blastx mode (v2.0.6.144, parameter: –evalue 0.001) (46) for functional annotation. The filtering parameters of DIAMOND were set with a query coverage threshold of 80% and a minimum score of 60 (16, 47). The best hits which met the above-described criteria were retained. DIAMOND results were turned into functional annotation based on the information provided by KEGG to create a gene KO list for the generation of KO relative abundance profiles. The taxonomic and functional information of new genes of EMGC was extracted for further analysis.

Evaluation of the effect of diet on the gut microbiota. To evaluate the consistent effect of diet among providers and mouse strains on taxonomic and functional composition of gut metagenomes, samples from 7 groups fed high-fat (HF) or low-fat (LF) diets and representing different providers and

mouse strains were selected (Table S1D). The calculation of relative abundance profiles for taxa and KO was according to Qin et al. (45). Permutational multivariate analysis of variance (PERMANOVA) (R vegan package) based on Bray-Curtis dissimilarity was applied to determine the influence of diet on gut metagenomes within different groups. The Shannon index of the relative abundance profiles was used to estimate alpha diversity of the samples. Principal-coordinates analysis (PCoA) of selected samples was performed based on the relative abundance profiles using Bray-Curtis dissimilarity (R ape4 package) to visualize the effect of diet on the bacterial composition of the gut microbiota. The Wilcox rank sum test was used for analysis of differences of genera and KO relative abundance profiles. *P* value adjustment was applied for multiple hypothesis testing using the Benjamini-Hochberg (BH) method. A BH-adjusted *P* value of <0.05 was considered statistically significant.

Metagenomic species clustering. The gene relative abundance profiles of 326 laboratory mice were clustered using the coabundance canopy algorithm (21). Coabundance genomes (CAGs) that were present in >90% samples were chosen, and CAGs with >700 genes were considered metagenomic species (MGSs) (21). CAGs and MGSs were assigned to a given taxon when >50% genes belonged to that specific taxon (21). The taxonomic distribution of MGSs was calculated using the R package phytool (48). The Shannon index was calculated based on the MGS profiles of samples from the 7 selected mouse groups, and PCoA was based on the same profile.

All MGSs were searched against the 830 high-quality metagenome assembly genomes (MAGs) and the 115 mGMB genomes by MUMmer3 (v3.23) (49) for calculation of MUMi values (22, 23). If the MUMi value for two items was >0.54, then these two items were recognized as the same species (23). The result of the comparison between MGSs and MAGs was hierarchically clustered by R package hclust with weighted pair group method with averaging (WPGMA) and then visualized in a cladogram with annotation by a R package ggtree (50, 51). The result of the comparison between MGSs and mGMB genomes is presented in Table S1J.

Data availability. The host-free sequenced data and assembled metagenomes of 88 mice in this study have been deposited in the China National GenBank Sequence Archive with project ID CNP0000619. EMGC can be reached by link http://ftp.cngb.org/pub/CNSA/data2/CNP0000619/Other/. The public data sets presented in this study can be found in online repositories. The names of the repositories and accession numbers can be found in Table S1B and S1C.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, PDF file, 0.1 MB.

FIG S2, PDF file, 0.3 MB.

FIG S3, PDF file, 0.1 MB.

FIG S4, PDF file, 0.1 MB.

FIG S5, PDF file, 0.01 MB.

FIG S6, PDF file, 0.3 MB.

FIG S7, PDF file, 0.1 MB.

FIG S8, PDF file, 0.1 MB.

FIG S9, PDF file, 0.5 MB.

TABLE S1, XLSX file, 0.3 MB.

ACKNOWLEDGMENTS

This work is funded by the National Natural Science Foundation of China (grant no. 81670606), the National Key Research and Development Program of China (no. 2018YFC1313800), and the Development and Reform Commission of Shenzhen Municipality (no. DRC-SZ [2015]162). This work was supported by China National GenBank (CNGB) and CNGB Sequence Archive.

We thank the sequencing and bioinformatic staff of BGI-Shenzhen, especially Dan Wang, Ying Xu, Fangming Yang, Jie Zhu, Zhongkui Xia, Suisha Liang, Jinghong Yu, and Guangwen Luo for help and advice in our work.

REFERENCES

1. Nguyen TLA, Vieira-Silva S, Liston A, Raes J. 2015. How informative is the mouse for human gut microbiota research? Dis Model Mech 8:1–16. https://doi.org/10.1242/dmm.017400.
2. Justice MJ, Dhillon P. 2016. Using the mouse to model human disease: increasing validity and reproducibility. Dis Model Mech 9:101–103. https://doi.org/10.1242/dmm.024547.
3. Perlman RL. 2016. Mouse models of human disease: an evolutionary perspective. Evol Med Public Health 2016:170–176. https://doi.org/10.1093/emph/eow014.
4. Hugenholtz F, de Vos WM. 2018. Mouse models for human intestinal microbiota research: a critical evaluation. Cell Mol Life Sci 75:149–160. https://doi.org/10.1007/s00018-017-2693-8.
5. Morgan XC, Huttenhower C. 2014. Meta'omic analytic techniques for studying the intestinal microbiome. Gastroenterology 146:1437.e1–1448.e1. https://doi.org/10.1053/j.gastro.2014.01.049.
6. Wang J, Jia H. 2016. Metagenome-wide association studies: fine-mining the microbiome. Nat Rev Microbiol 14:508–522. https://doi.org/10.1038/nrmicro.2016.83.

7. Xiao L, Feng Q, Liang S, Sonne SB, Xia Z, Qiu X, Li X, Long H, Zhang J, Zhang D, Liu C, Fang Z, Chou J, Glanville J, Hao Q, Kotowska D, Colding C, Licht TR, Wu D, Yu J, Sung JJY, Liang Q, Li J, Jia H, Lan Z, Tremaroli V, Dworzynski P, Nielsen HB, Bäckhed F, Doré J, Le Chatelier E, Ehrlich SD, Lin JC, Arumugam M, Wang J, Madsen L, Kristiansen K. 2015. A catalog of the mouse gut metagenome. Nat Biotechnol 33:1103–1108. https://doi.org/10.1038/nbt.3353.
8. Wannemuehler MJ, Overstreet A-M, Ward DV, Phillips GJ. 2014. Draft genome sequences of the altered Schaedler flora, a defined bacterial community from gnotobiotic mice. Genome Announc 2:e00287-14. https://doi.org/10.1128/genomeA.00287-14.
9. Lagkouvardos I, Pukall R, Abt B, Foesel BU, Meier-Kolthoff JP, Kumar N, Bresciani A, Martínez I, Just S, Ziegler C, Brugiroux S, Garzetti D, Wenning M, Bui TPN, Wang J, Hugenholtz F, Plugge CM, Peterson DA, Hornef MW, Baines JF, Smidt H, Walter J, Kristiansen K, Nielsen HB, Haller D, Overmann J, Stecher B, Clavel T. 2016. The mouse intestinal bacterial collection (miBC) provides host-specific insight into cultured diversity and functional potential of the gut microbiota. Nat Microbiol 1:16131. https://doi.org/10.1038/nmicrobiol.2016.131.
10. Brugiroux S, Beutler M, Pfann C, Garzetti D, Ruscheweyh H-J, Ring D, Diehl M, Herp S, Lötscher Y, Hussain S, Bunk B, Pukall R, Huson DH, Münch PC, McHardy AC, McCoy KD, Macpherson AJ, Loy A, Clavel T, Berry D, Stecher B. 2016. Genome-guided design of a defined mouse microbiota that confers colonization resistance against *Salmonella enterica* serovar Typhimurium. Nat Microbiol 2:16215. https://doi.org/10.1038/nmicrobiol.2016.215.
11. Liu C, Zhou N, Du M-X, Sun Y-T, Wang K, Wang Y-J, Li D-H, Yu H-Y, Song Y, Bai B-B, Xin Y, Wu L, Jiang C-Y, Feng J, Xiang H, Zhou Y, Ma J, Wang J, Liu H-W, Liu S-J. 2020. The mouse gut microbial biobank expands the coverage of cultured bacteria. Nat Commun 11:79. https://doi.org/10.1038/s41467-019-13836-5.
12. Lesker TR, Durairaj AC, Gálvez EJC, Lagkouvardos I, Baines JF, Clavel T, Sczyrba A, McHardy AC, Strowig T. 2020. An integrated metagenome catalog reveals new insights into the murine gut microbiome. Cell Rep 30:2909.e6–2922.e6. https://doi.org/10.1016/j.celrep.2020.02.036.
13. Xiao L, Sonne SB, Feng Q, Chen N, Xia Z, Li X, Fang Z, Zhang D, Fjære E, Midtbø LK, Derrien M, Hugenholtz F, Tang L, Li J, Zhang J, Liu C, Hao Q, Vogel UB, Mortensen A, Kleerebezem M, Licht TR, Yang H, Wang J, Li Y, Arumugam M, Wang J, Madsen L, Kristiansen K. 2017. High-fat feeding rather than obesity drives taxonomical and functional changes in the gut microbiota in mice. Microbiome 5:43. https://doi.org/10.1186/s40168-017-0258-6.
14. Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28:27–30. https://doi.org/10.1093/nar/28.1.27.
15. Zhu J, Ren H, Zhong H, Li X, Zou Y, Han M, Li M, Madsen L, Kristiansen K, Xiao L. 16 September 2020. An expanded gene catalog of mouse gut metagenomes. BioRxiv https://doi.org/10.1101/2020.09.16.299339.
16. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T, Juncker AS, Manichanh C, Chen B, Zhang W, Levenez F, Wang J, Xu X, Xiao L, Liang S, Zhang D, Zhang Z, Chen W, Zhao H, Al-Aama JY, Edris S, Yang H, Wang J, Hansen T, Nielsen HB, Brunak S, Kristiansen K, Guarner F, Pedersen O, Doré J, Ehrlich SD, Bork P, Wang J, MetaHIT Consortium. 2014. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol 32:834–841. https://doi.org/10.1038/nbt.2942.
17. Suez J, Korem T, Zeevi D, Zilberman-Schapira G, Thaiss CA, Maza O, Israeli D, Zmora N, Gilad S, Weinberger A, Kuperman Y, Harmelin A, Kolodkin-Gal I, Shapiro H, Halpern Z, Segal E, Elinav E. 2014. Artificial sweeteners induce glucose intolerance by altering the gut microbiota. Nature 514:181–186. https://doi.org/10.1038/nature13793.
18. Everard A, Lazarevic V, Gaïa N, Johansson M, Ståhlman M, Backhed F, Delzenne NM, Schrenzel J, François P, Cani PD. 2014. Microbiome of prebiotic-treated mice reveals novel targets involved in host response during obesity. ISME J 8:2116–2130. https://doi.org/10.1038/ismej.2014.45.
19. Menzel P, Ng KL, Krogh A. 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun 7:11257. https://doi.org/10.1038/ncomms11257.
20. Ondov BD, Bergman NH, Phillippy AM. 2011. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics 12:385. https://doi.org/10.1186/1471-2105-12-385.
21. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, Gautier L, Pedersen AG, Le Chatelier E, Pelletier E, Bonde I, Nielsen T, Manichanh C, Arumugam M, Batto J-M, Quintanilha dos Santos MB, Blom N, Borruel N, Burgdorf KS, Boumezbeur F, Casellas F, Doré J, Dworzynski P, Guarner F, Hansen T, Hildebrand F, Kaas RS, Kennedy S, Kristiansen K, Kultima JR, Léonard P, Levenez F, Lund O, Moumen B, Le Paslier D, Pons N, Pedersen O, Prifti E, Qin J, Raes J, Sørensen S, Tap J, Tims S, Ussery DW, Yamada T, MetaHIT Consortium, Renault P, Sicheritz-Ponten T, Bork P, Wang J, et al. 2014. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol 32:822–828. https://doi.org/10.1038/nbt.2939.
22. Deloger M, El Karoui M, Petit M-A. 2009. A Genomic distance based on MUM indicates discontinuity between most bacterial species and genera. J Bacteriol 191:91–99. https://doi.org/10.1128/JB.01202-08.
23. Bäckhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, Li Y, Xia Y, Xie H, Zhong H, Khan MT, Zhang J, Li J, Xiao L, Al-Aama J, Zhang D, Lee YS, Kotowska D, Colding C, Tremaroli V, Yin Y, Bergman S, Xu X, Madsen L, Kristiansen K, Dahlgren J, Wang J, Jun W. 2015. Dynamics and stabilization of the human gut microbiome during the first year of life. Cell Host Microbe 17:690–703. https://doi.org/10.1016/j.chom.2015.04.004.
24. Li J, Zhong H, Ramayo-Caldas Y, Terrapon N, Lombard V, Potocki-Veronese G, Estellé J, Popova M, Yang Z, Zhang H, Li F, Tang S, Yang F, Chen W, Chen B, Li J, Guo J, Martin C, Maguin E, Xu X, Yang H, Wang J, Madsen L, Kristiansen K, Henrissat B, Ehrlich SD, Morgavi DP. 2020. A catalog of microbial genes from the bovine rumen unveils a specialized and diverse biomass-degrading environment. Gigascience 9:giaa057. https://doi.org/10.1093/gigascience/giaa057.
25. Laukens D, Brinkman BM, Raes J, De Vos M, Vandenabeele P. 2016. Heterogeneity of the gut microbiome in mice: guidelines for optimizing experimental design. FEMS Microbiol Rev 40:117–132. https://doi.org/10.1093/femsre/fuv036.
26. Stappenbeck TS, Virgin HW. 2016. Accounting for reciprocal host–microbiome interactions in experimental science. Nature 534:191–199. https://doi.org/10.1038/nature18285.
27. Steinegger M, Mirdita M, Söding J. 2019. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. Nat Methods 16:603–606. https://doi.org/10.1038/s41592-019-0437-4.
28. Hauser M, Steinegger M, Söding J. 2016. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. Bioinformatics 32:1323–1330. https://doi.org/10.1093/bioinformatics/btw006.
29. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 7:e7359. https://doi.org/10.7717/peerj.7359.
30. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, Banfield JF. 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat Microbiol 3:836–843. https://doi.org/10.1038/s41564-018-0171-1.
31. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol 2:1533–1542. https://doi.org/10.1038/s41564-017-0012-7.
32. Nayfach S, Pollard KS. 2016. Toward accurate and quantitative comparative metagenomics. Cell 166:1103–1116. https://doi.org/10.1016/j.cell.2016.08.007.
33. Fang C, Zhong H, Lin Y, Chen B, Han M, Ren H, Lu H, Luber JM, Xia M, Li W, Stein S, Xu X, Zhang W, Drmanac R, Wang J, Yang H, Hammarström L, Kostic AD, Kristiansen K, Li J. 2018. Assessment of the cPAS-based BGISEQ-500 platform for metagenomic sequencing. Gigascience 7:1–8. https://doi.org/10.1093/gigascience/gix133.
34. Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, Wang J. 2009. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25:1966–1967. https://doi.org/10.1093/bioinformatics/btp336.
35. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Res 27:824–834. https://doi.org/10.1101/gr.213959.116.
36. Kultima JR, Sunagawa S, Li J, Chen W, Chen H, Mende DR, Arumugam M, Pan Q, Liu B, Qin J, Wang J, Bork P. 2012. MOCAT: a metagenomics assembly and gene prediction toolkit. PLoS One 7:e47656. https://doi.org/10.1371/journal.pone.0047656.
37. Zhu W, Lomsadze A, Borodovsky M. 2010. *Ab initio* gene identification in metagenomic sequences. Nucleic Acids Res 38:e132. https://doi.org/10.1093/nar/gkq275.

38. Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152. https://doi.org/10.1093/bioinformatics/bts565.
39. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
40. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.
41. Nurk S, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, Lapidus A, Prjibelsky A, Pyshkin A, Sirotkin A, Sirotkin Y, Stepanauskas R, McLean J, Lasken R, Clingenpeel SR, Woyke T, Tesler G, Alekseyev MA, Pevzner PA. 2013. Assembling genomes and mini-metagenomes from highly chimeric reads, p 158–170. *In* Deng M, Jiang R, Sun F, Zhang X (ed), Research in computational molecular biology. Springer, Berlin, Germany.
42. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055. https://doi.org/10.1101/gr.186072.114.
43. Lomsadze A, Gemayel K, Tang S, Borodovsky M. 2018. Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. Genome Res 28:1079–1089. https://doi.org/10.1101/gr.230615.117.
44. Chen I-MA, Chu K, Palaniappan K, Pillay M, Ratner A, Huang J, Huntemann M, Varghese N, White JR, Seshadri R, Smirnova T, Kirton E, Jungbluth SP, Woyke T, Eloe-Fadrosh EA, Ivanova NN, Kyrpides NC. 2019. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. Nucleic Acids Res 47:D666–D677. https://doi.org/10.1093/nar/gky901.
45. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D, Peng Y, Zhang D, Jie Z, Wu W, Qin Y, Xue W, Li J, Han L, Lu D, Wu P, Dai Y, Sun X, Li Z, Tang A, Zhong S, Li X, Chen W, Xu R, Wang M, Feng Q, Gong M, Yu J, Zhang Y, Zhang M, Hansen T, Sanchez G, Raes J, Falony G, Okuda S, Almeida M, LeChatelier E, Renault P, Pons N, Batto J-M, Zhang Z, Chen H, Yang R, Zheng W, Li S, Yang H, et al. 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490:55–60. https://doi.org/10.1038/nature11450.
46. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60. https://doi.org/10.1038/nmeth.3176.
47. Xiao L, Estellé J, Kiilerich P, Ramayo-Caldas Y, Xia Z, Feng Q, Liang S, Pedersen AØ, Kjeldsen NJ, Liu C, Maguin E, Doré J, Pons N, Le Chatelier E, Prifti E, Li J, Jia H, Liu X, Xu X, Ehrlich SD, Madsen L, Kristiansen K, Rogel-Gaillard C, Wang J. 2016. A reference gene catalogue of the pig gut microbiome. Nat Microbiol 1:16161. https://doi.org/10.1038/nmicrobiol.2016.161.
48. Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol 3:217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x.
49. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. Genome Biol 5:R12. https://doi.org/10.1186/gb-2004-5-2-r12.
50. Yu G, Smith DK, Zhu H, Guan Y, Lam TT. 2017. GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol 8:28–36. https://doi.org/10.1111/2041-210X.12628.
51. Yu G. 2020. Using ggtree to visualize data on tree-like structures. Curr Protoc Bioinformatics 69:e96. https://doi.org/10.1002/cpbi.96.