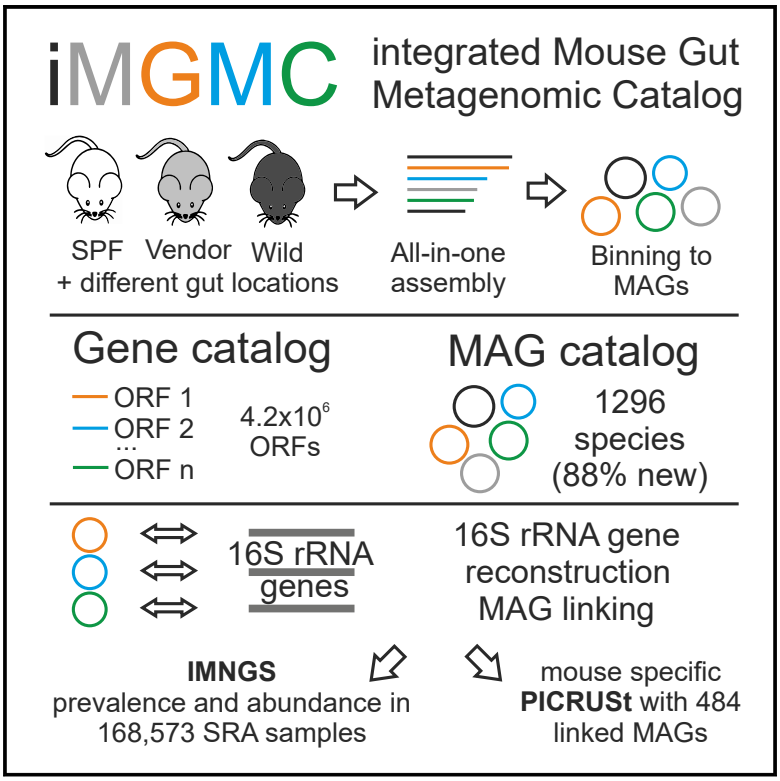
Article

Cell Reports

An Integrated Metagenome Catalog Reveals New Insights into the Murine Gut Microbiome

Graphical Abstract



Highlights

- Large-scale metagenomic assembly uncovers hundreds of mouse microbiome species
- Most microbes found in the mouse gut are unique to the ecosystem
- Integration of gene catalog and microbial genomes creates a comprehensive resource
- Linking of 16S rRNA genes with metagenome-assembled genomes allows new applications

Authors

Till R. Lesker, Abilash C. Durairaj, Eric J.C. Gálvez, ..., Alexander Sczyrba, Alice C. McHardy, Till Strowig

Correspondence

till.strowig@helmholtz-hzi.de

In Brief

Gene catalogs and genome references facilitate taxonomic and functional annotation of sequencing data. Through the combination of data from laboratory and wild mice, Lesker et al. create a comprehensive resource, the integrated mouse gut metagenome catalog, to characterize the microbial ecosystem in the murine gut.

Lesker et al., 2020, Cell Reports *30*, 2909–2922
March 3, 2020
https://doi.org/10.1016/j.celrep.2020.02.036

CellPress

Cell Reports
Article

An Integrated Metagenome Catalog Reveals New Insights into the Murine Gut Microbiome

Till R. Lesker,[1,7] Abilash C. Durairaj,[1] Eric J.C. Gálvez,[1] Ilias Lagkouvardos,[2] John F. Baines,[3,4] Thomas Clavel,[2,5] Alexander Sczyrba,[6,7] Alice C. McHardy,[7,8] and Till Strowig[1,9,10,11,*]

[1]Department of Microbial Immune Regulation, Helmholtz Centre for Infection Research, 38124 Braunschweig, Germany
[2]ZIEL Institute for Food and Health, Technical University of Munich, 85354 Freising, Germany
[3]Max Planck Institute for Evolutionary Biology, 24306 Plön, Germany
[4]Institute for Experimental Medicine, Kiel University, 24118 Kiel, Germany
[5]Functional Microbiome Research Group, Institute of Medical Microbiology, RWTH University Hospital, 52074 Aachen, Germany
[6]Faculty of Technology and Center for Biotechnology, Bielefeld University, 33501 Bielefeld, Germany
[7]Department of Computational Biology of Infection Research, Helmholtz Centre for Infection Research, 38124 Braunschweig, Germany
[8]Braunschweig Integrated Centre of Systems Biology, 38106 Braunschweig, Germany
[9]Hanover Medical School, 30625 Hannover, Germany
[10]RESIST, Cluster of Excellence 2155, Hanover Medical School, 30625 Hanover, Germany
[11]Lead Contact
*Correspondence: till.strowig@helmholtz-hzi.de
https://doi.org/10.1016/j.celrep.2020.02.036

SUMMARY

The complexity of host-associated microbial ecosystems requires host-specific reference catalogs to survey the functions and diversity of these communities. We generate a comprehensive resource, the integrated mouse gut metagenome catalog (iMGMC), comprising 4.6 million unique genes and 660 metagenome-assembled genomes (MAGs), many (485 MAGs, 73%) of which are linked to reconstructed full-length 16S rRNA gene sequences. iMGMC enables unprecedented coverage and taxonomic resolution of the mouse gut microbiota; i.e., more than 92% of MAGs lack species-level representatives in public repositories (<95% ANI match). The integration of MAGs and 16S rRNA gene data allows more accurate prediction of functional profiles of communities than predictions based on 16S rRNA amplicons alone. Accompanying iMGMC, we provide a set of MAGs representing 1,296 gut bacteria obtained through complementary assembly strategies. We envision that integrated resources such as iMGMC, together with MAG collections, will enhance the resolution of numerous existing and future sequencing-based studies.

INTRODUCTION

The gut microbiota is a dynamic and highly diverse microbial ecosystem that affects many aspects of the host's physiology (Kamada et al., 2013). Culture-independent methods such as shotgun metagenome sequencing have revolutionized experimental approaches to characterizing and investigating these communities. Gene catalogs and collections of metagenome-assembled genomes (MAGs) facilitate taxonomic and functional annotation of sequencing data, thereby maximizing insights gained from short reads (Li et al., 2014; Sunagawa et al., 2015; Xiao et al., 2015, 2016; Almeida et al., 2019; Pasolli et al., 2019). Typically, generation of reference gene catalogs involves sample-specific assembly, prediction of genes, and dataset-wide clustering of gene entries to reduce redundancy. However, this approach results in reduced taxonomic resolution of gene entries. This is due to the clustering of highly related but distinct genes and the lack of high-resolution taxonomic information, which can be best obtained from marker genes such as the 16S rRNA gene, for which large reference collections exist. Yet a specific challenge of large-scale metagenomic approaches is the linking of specific 16S rRNA gene sequences to MAGs, which results typically in low linking rates (Parks et al., 2017; Pasolli et al., 2019). Here we present a comprehensive approach and corresponding computational workflow to construct integrated gene catalogs, resulting in a significant improvement of the taxonomic resolution of gene entries, together with linking genes to MAGs and reconstructed full-length 16S rRNA genes. The integrated mouse gut metagenome catalog (iMGMC) was constructed from 298 publicly available and newly sequenced metagenome samples. Moreover, we present a set of additional MAGs obtained from separate single-sample assembly of 898 metagenomic sequencing samples that, together with MAGs integrated into iMGMC, comprise 1,296 species-level bacterial genomes.

RESULTS

Construction of iMGMC

Pioneering work resulted in the construction of several gene catalogs, including a microbiome gene catalog from the mouse gut (hereafter called MGCv1) comprising 2.6 million non-redundant genes (Qin et al., 2010; Li et al., 2014; Xiao et al., 2015, 2016). To advance this resource, we developed a bioinformatic workflow that combines a global assembly (all in one) with binning

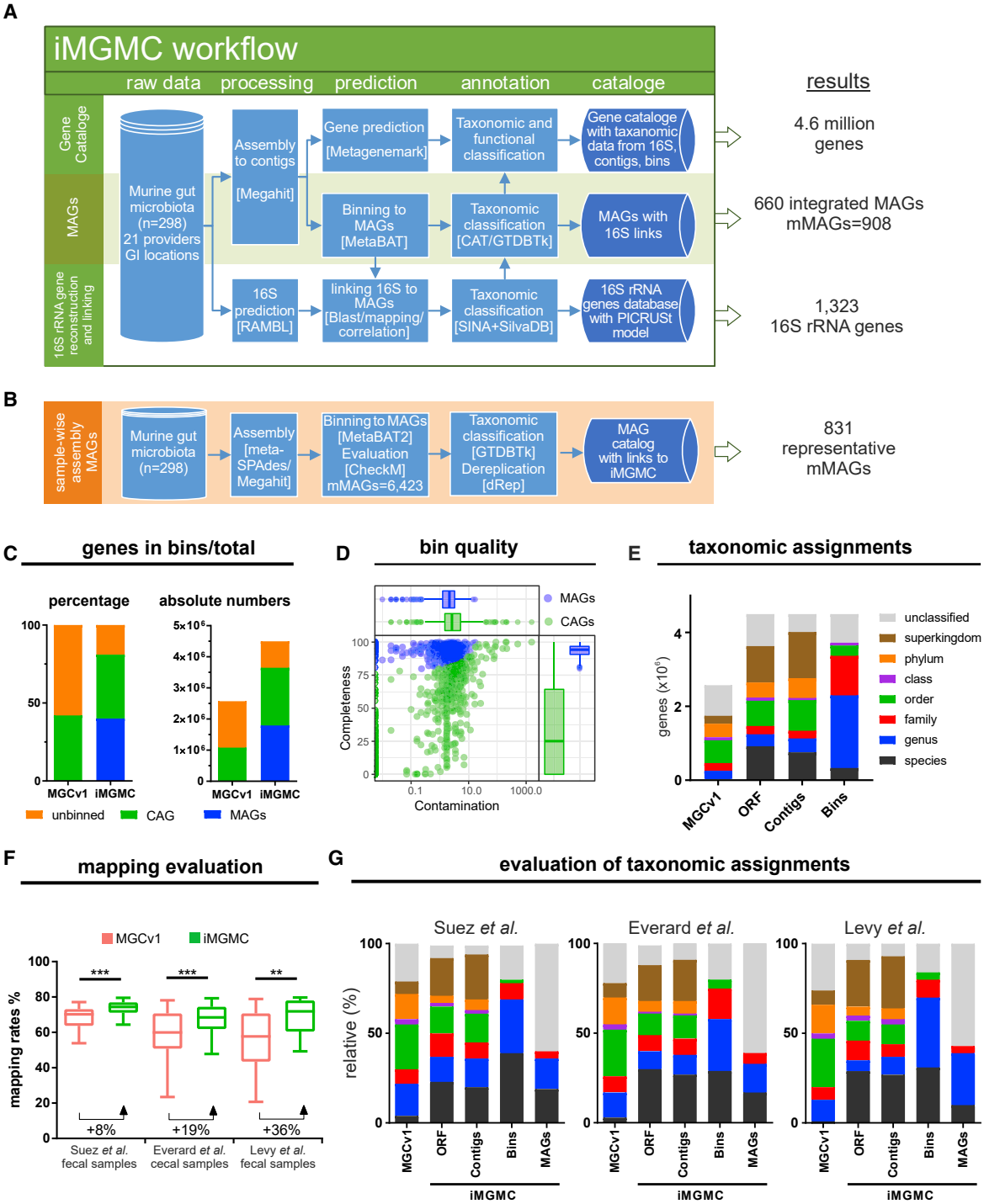


Figure 1. Generation and Evaluation of iMGMC

(A) Flowchart displaying the steps and bioinformatics tools (names in brackets) used for the generation of iMGMC. This resource includes genes, MAGs, 16S rRNA gene sequences, and MAG-16S rRNA gene links.

(B) Flowchart displaying the steps and bioinformatics tools (names in brackets) used for the single-sample assembly approach.

(C) Comparison of relative and total numbers of gene entries and their association to bins of different levels of completeness between a previous mouse gut gene catalog (MGCv1) (Xiao et al., 2015) and iMGMC. Bins were defined as (1) co-abundance genomes (CAG) if they were ≥200 kbp lengths and contained ≥700 ORFs or (2) MAGs if their quality (marker gene completeness − contamination) as determined by CheckM was ≥80%.

(legend continued on next page)

of contigs to MAGs and with an innovative linkage of reconstructed 16S rRNA gene sequences to these MAGs (Figures 1A and 1B; see STAR Methods for details). This methodology enables maintaining complex information such as the distribution of distinct contigs and bins over a large number of samples. We applied this approach to newly sequenced samples from commercial mouse providers and wild mice (n = 108) combined with a previously published set of data included in MGCv1 (n = 190) (see Table S1). In total, 1.3 Tbp from 298 metagenomic libraries were assembled (Li et al., 2016), resulting in 1.2 million contigs with a total assembly size of 4.5 Gbp and comprising 4.6 million open reading frames (ORFs) compared with 2.6 million ORFs in the MGCv1, an increase of 77% as a result of both increased sample numbers and methodological changes (Figure 1C). We tested the redundancy of ORFs by clustering them (Fu et al., 2012), which resulted in a reduction of 2% of ORFs (n = 99,670) (data not shown). Therefore, we maintained all ORF in iMGMC without clustering. Subsequently, contigs were binned (Kang et al., 2015), resulting in 1,462 bins (>200 kbp) containing 87% of iMGMC entries. Hence, only 13% of entries remained in contigs/bins shorter than 200 kbp. We then defined 660 bins as iMAGs (integrated MAGs, containing 40% of iMGMC ORFs), based on the presence of established sets of bacterial marker genes using CheckM (completeness − contamination $\geq$80%) (Figures 1C and 1D; Table S2; Parks et al., 2015). According to recently established guidelines (Bowers et al., 2017) iMGMC contains 908 MAGs of medium quality (mMAG, completeness >50%, contamination <10%) (Tables S2 and S3). MGCv1 did not provide MAGs but rather provided co-abundance groups (CAGs), containing at least 700 genes. Comparison of the numbers of CAGs and genes in CAGs revealed large increases in iMGMC compared with MGCv1 (1,217 versus 541 CAGs and 81% versus 40% of genes, respectively) (Figure 1C).

Several large-scale studies reconstructing microbial genomes from thousands of metagenomes employed single-sample assembly rather than the all-in-one approach described here (Crits-Christoph et al., 2018; Almeida et al., 2019; Nayfach et al., 2019; Pasolli et al., 2019). For single-sample assembly, metaSPAdes has been routinely demonstrated to outperform Megahit (Pasolli et al., 2019); however, metaSPAdes is not able to handle large datasets for all-in-one assembly. Hence, to conduct a comparison of the two approaches, we compared the results obtained from the aforementioned dataset (n = 298 libraries) using a approach-specific optimal assembler. Strikingly, the all-in-one approach outperforms the single-sample assembly for this dataset when assessed by number of distinct MAGs obtained after dereplication, i.e., clustering with ANI (average nucleotide identity) >95% (725 versus 568 MAGs, dRep default filtering setting) without compromising the quality (completeness and contamination) (Figure S1; Table S3). A significant fraction of mMAGs (45%) was recovered by either of the two methods, and all-in-one assembly-specific bins had a lower relative abundance, suggesting the approach is potentially suited to recover low-abundant microbes (Figure S1). To compare the quality of MAG assembly by the two methods, previously known related bacterial genomes (ANI >99%) (n = 26 MAGs; see STAR Methods for details) were identified and the comparison revealed equal MAG quality (Table S2). Thus, although an influence of the different assemblers, supposedly optimized for each approach, cannot be excluded from the comparison, we considered the all-in-one assembly to be as good as single-sample assembly for the analyzed dataset while offering distinct advantages for iMGMC construction.

Because assessing genome reconstruction solely based on marker gene presence potentially overestimates genome completeness, further evaluation was conducted using known bacterial genomes that were present within the assembly (see STAR Methods for details). This analysis demonstrated that large fractions of previously known bacterial genomes identified within the MAGs (n = 57) were located within distinct MAGs (78% ± 19% of genomes) (Table S2), providing additional evidence for the efficacy of this approach.

The taxonomic assignment of entries in classical gene catalogs, specifically after sample-specific assembly and clustering of ORFs by similarity (Xiao et al., 2015), is limited by the ability of algorithms to predict the taxonomic placement based on relatively short ORFs (Sczyrba et al., 2017). Taking advantage of the clustering-free approach, we annotated each iMGMC entry using the taxonomic information obtained from the respective gene and contig, as well as from the bin (Figure 1E). As a result, the relative taxonomic assignment rate improved between 28% and 1,021% at different taxonomic levels (Figure 1E). To independently evaluate the performance of iMGMC data from three external studies (Everard et al., 2014; Suez et al., 2014; Levy et al., 2015), which were excluded from the construction of iMGMC or MGCv1, were mapped against both catalogs. This revealed an increased number of reads (up to 36%) mapping to iMGMC (Figures 1F and 1G). Hence, through the combination of additional samples and an optimized assembly strategy, an improved gene catalog has been generated.

(D) Quality determination of individual binned contigs by CheckM by analyzing marker gene completeness and contamination. Boxplots display marker gene completeness and contamination of 660 iMAGs and 802 CAGs, respectively. Data are displayed as a box-whisker plot representing 10%, first quartile, median, third quartile, and 90%.

(E) Absolute numbers of gene entries colored according to the lowest-possible taxonomic annotation of the ORF, contig, or bin. Different taxonomic profilers were employed for classification: ORF used DIAMOND-BlastP, contigs used CAT (Contig annotation tool), and bins used GTDBTk (Genome Taxonomy Database Toolkit).

(F and G) Comparison of mapping rates (F) and taxonomic assignment (G) of previously published mouse gut metagenome datasets based on the original mouse gut catalog (MGCv1, red) and iMGMC (green). Relative improvements in mapping rates are indicated in (F). Two-tailed paired t test was performed to analyze the differences in (F); ***$p < 0.001$, **$p < 0.01$. Suez et al. (2014) (n = 40 fecal samples), Everard et al. (2014) (n = 34 fecal samples), Levy et al. (2015) (n = 10 fecal samples). Data are displayed as a box-whisker plot representing minimum, first quartile, median, third quartile, and maximum.

See also Figure S1 and Tables S1, S2, and S3.

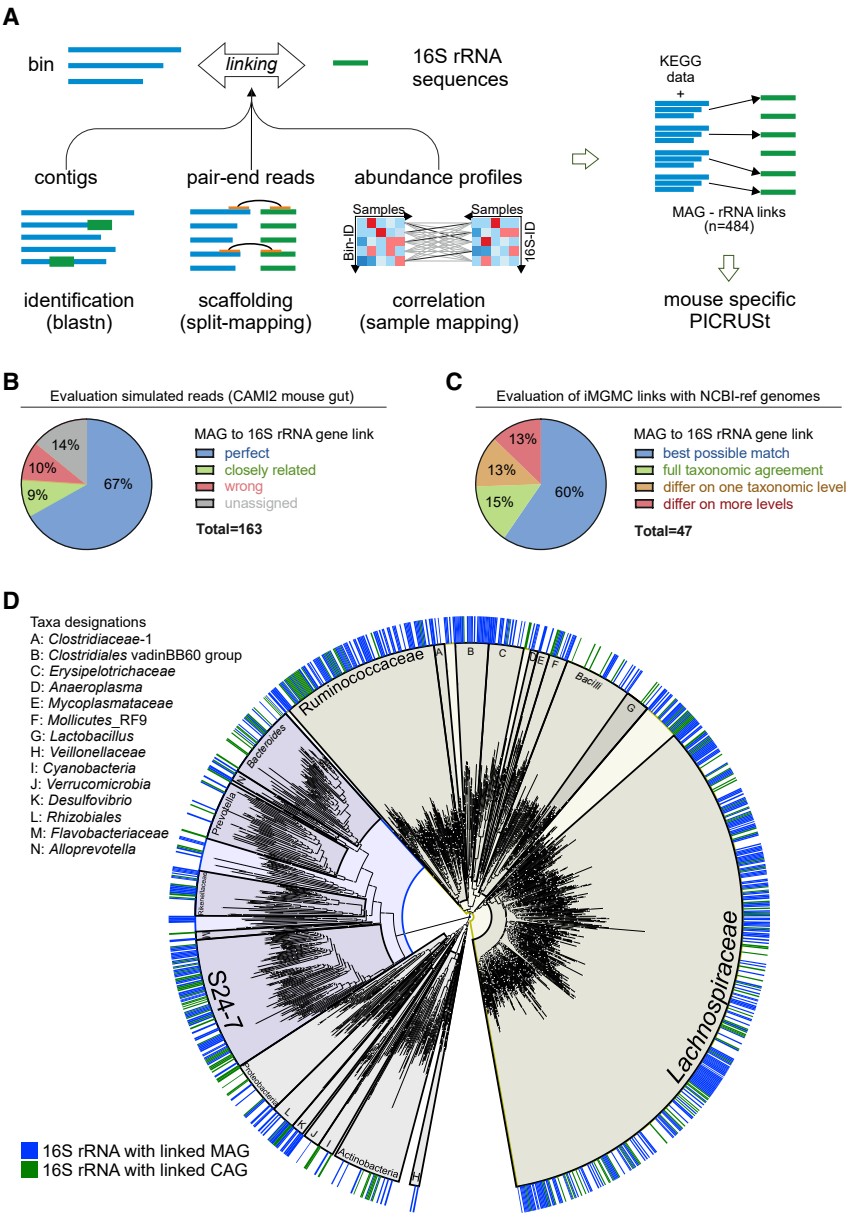


Figure 2. Linking Approach of Reconstructed 16S rRNA Genes to MAGs

(A) Overview of the methodology to link MAGs to 16S rRNA gene sequences by combining mapping-based and statistical approaches. Resulting linked pairs of MAGs and reconstructed 16S rRNA gene sequences were used, together with KEGG annotations, for construction of mouse-gut-specific PICRUSt prediction.

(B) Evaluation of linking method with simulated data (CAMI2 mouse gut). See STAR Methods for details.

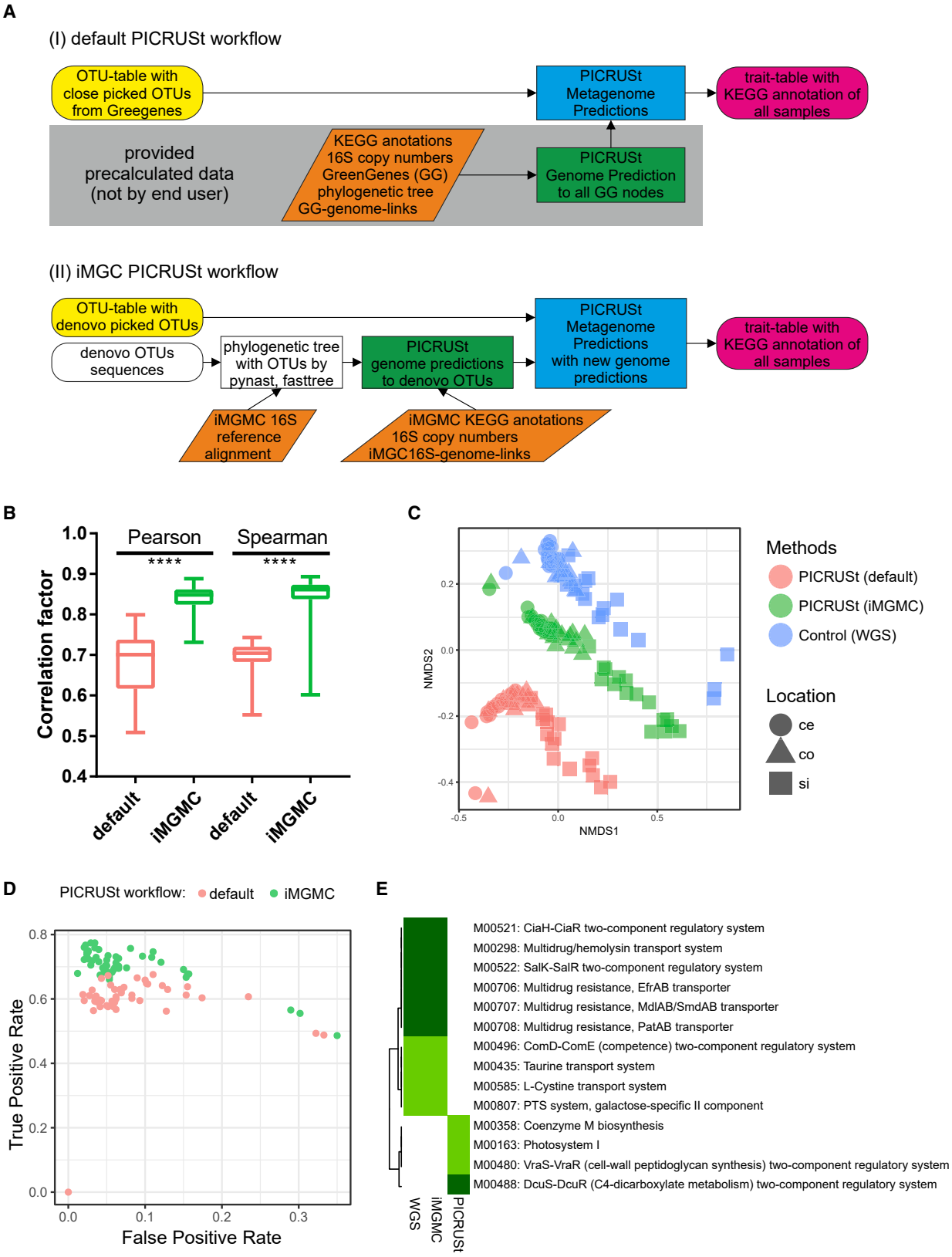
(C) Evaluation of linking in iMGMC with NCBI reference genomes. See STAR Methods and Table S2 for details.

(D) Phylogenetic tree containing the reconstructed 16S rRNA gene sequences. Taxonomic groups are highlighted. The color in the outer ring indicates the presence of a linked MAG (blue) or CAG (green).

See also Figure S2 and Tables S2 and S4.

Reconstruction and Linking of 16S rRNA Gene Sequences to MAGs

Because 16S rRNA genes are typically not efficiently recovered in standard assemblies obtained from short-read sequencing data due to their highly conserved regions (Miller et al., 2011), dedicated tools for 16S rRNA gene reconstruction from metagenomic sequencing data were developed (Miller et al., 2011; Zeng et al., 2017). From the iMGMC dataset, 1,323 full-length, unique 16S rRNA gene sequences were assembled using RAMBL that included only a minor fraction of chimeric sequences, as determined using Uchime2 (4.3%) (see STAR Methods for details). We postulated that linking 16S rRNA genes to bins and iMAGs would allow efficient integration of functional and taxonomic information. However, no high-throughput method exists for creating such links. Hence, we developed an integrated score combining mapping- and correlation-based associations to assign a 16S rRNA gene sequence to each bin (Figures 2A and S2; see STAR Methods for details). To evaluate this approach, we assessed it using a synthetic dataset generated using 791 known genomes (Fritz et al., 2019). The dataset contained 64 distinct samples that were assembled using the all-in-one approach, resulting in 438 mMAGs. As for the iMGMC dataset, RAMBL was used to reconstruct 16S rRNA gene sequences (n = 460). Of the 438 mMAGs, 204 reached the iMGMC quality criteria (CheckM: completeness − contamination ≥ 80%), and for 163 of these MAGs (79%), it was possible to assign a reconstructed 16S rRNA sequence. MAGs were mapped to reference genomes with FastANI, allowing the identification of the gold standard 16S rRNA gene for each MAG. Strikingly, our linking approach predicted for 103 (63.2%) MAGs the best possible reconstructed 16S rRNA gene sequence (agreement of linked to gold standard 16S rRNA gene sequences) (Figure 2B). From the remaining 60 linked sequences, 29 could be filtered out by taxonomic disagreement of the 16S rRNA gene and MAG on at least the family level. For 15 of the remaining 31 links, the linked 16S rRNA gene sequences was closely related to the gold standard (same genus), and the other 16 connections (9.8%) were distinct from the gold standard. Encouraged by the performance of the linking approach, an automated approach without curation was first performed for iMGMC. Then, the predicted MAG-16S rRNA gene pairs were evaluated in iMGMC using MAGs with linked 16S rRNA gene sequences for which reference genomes exist (see STAR Methods for details). Of the 47 identified genomes and respective bins, 28 agreed perfectly



(legend on next page)

(100% sequence identity) between known and linked 16S rRNA genes, with an additional 7 matching taxonomic assignment down to the genus level (Figure 2C). The remaining 12 genomes and bins disagreed at varying taxonomic levels (Figure 2C; Data S1), demonstrating a similar performance as for the synthetic dataset. To improve the quality, manual curation was then performed to exclude MAG-16S rRNA gene links with stark differences between MAG and 16S rRNA gene taxonomy (larger than family level) and those in which 16S rRNA genes were associated multiple times to different MAGSs/bins. Finally, in iMGMC, 485 of the 660 iMAGs (73%) were assigned to a unique 16S rRNA gene sequence (Figure 2D). Altogether, this shows that the proposed scoring scheme is able to link MAGs and bins to corresponding reconstructed 16S rRNA genes in a largely improved manner, though not in an error-free manner, enabling novel applications.

Improved Functional Prediction via MAG-16S rRNA Gene Links in iMGMC

The establishment of databases of microbial reference genomes has spurred the development of approaches to simulate functional profiles of metagenomes based on marker gene datasets, e.g., 16S rRNA amplicon profiles (Langille et al., 2013; Aßhauer et al., 2015). Because numerous bacteria within murine gut communities lack reference genomes, we hypothesized that the default phylogenetic investigation of communities by reconstruction of unobserved states (PICRUSt)-based predictions of mouse-associated metagenome functions are limited (Langille et al., 2013). Enabled by the linking of MAGs to 16S rRNA gene sequences, we constructed a mouse-optimized PICRUSt version (PICRUSt-iMGMC), employing the original PICRUSt algorithm in conjunction with the iMGMC data (Figure 3A; see STAR Methods for details). Comparison of KEGG (Kyoto Encyclopedia of Genes and Genomes) ortholog (KO) profiles predicted by default PICRUSt and PICRUSt-iMGMC to the corresponding shotgun metagenomic libraries (whole-genome sequencing, WGS) demonstrated a higher correlation to the WGS-based KO profiles for PICRUSt-iMGMC than for default PICRUSt (Pearson: 0.84 versus 0.68, +23%; Spearman: 0.84 versus 0.70, 21%) (Figures 3B and 3C). The highest correlations were observed for colon samples (Figure S3). Similar improvements were obtained with distinct datasets not used for the construction of the catalog, including samples from wild mice (Figure S3; Rosshart et al., 2017; Fabbiano et al., 2018). The improved correlation of PICRUSt-iMGMC largely derived from increased sensitivity, i.e., true-positive rates, rather than decreased false-positive rates, enabling the prediction of functionalities otherwise lost (Figures 3D and 3E). Even mapping WGS data to the KEGG database with DIAMOND (Buchfink et al., 2015) or combining iMGMC information with KEGG did not improve the prediction (Figure S3). PICRUSt-iMGMC/KEGG even decreased the correlation, suggesting that inclusion of related but divergent genomes reduces prediction accuracy (Figure S3). Hence, our resource enabled the development of mouse-specific PICRUSt models with substantial improvement in the prediction of metagenomic functional profiles.

iMGMC Reveals High Prevalence of Previously Unknown Taxa in the Mouse Gut Microbiota

Both metagenomic and cultivation-based studies showed that the gut microbiome of mice is composed of distinct bacterial species compared with human gut microbiome, many of which are still uncultured and lack genomic information (Xiao et al., 2015; Lagkouvardos et al., 2016b). Analysis of the 660 iMAGs corroborates this notion, revealing that only 52 of them are known species (Genome Taxonomy Database [GTDB], ANI >95%) (Tables S2 and S4; Parks et al., 2018).

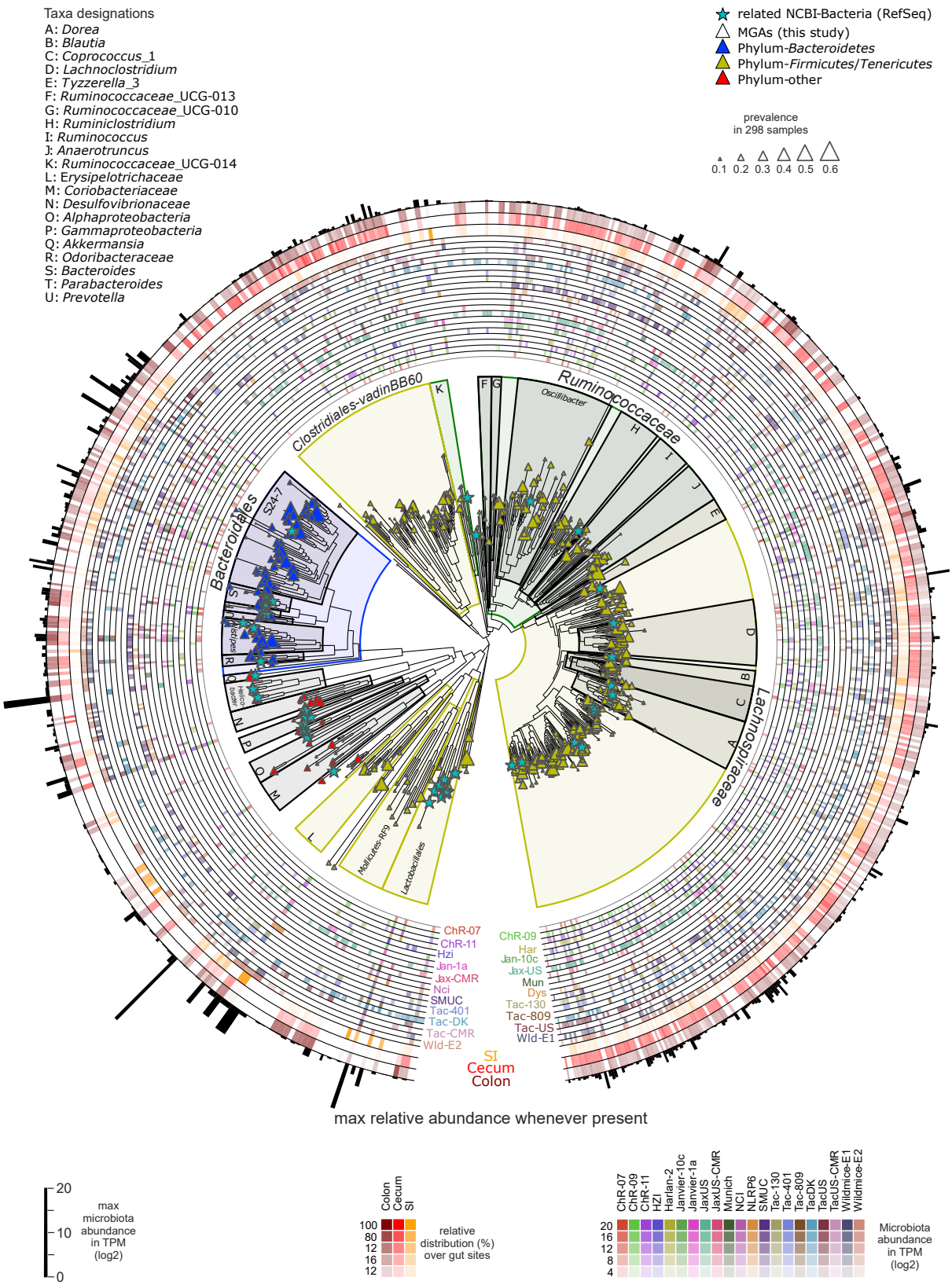
To construct a comprehensive phylogenetic tree of the mouse gut microbiota, we assigned iMAGs and closely related, previously sequenced genomes (n = 64) into clusters (Figure 4). In line with previous reports (Clavel et al., 2016; Lagkouvardos et al., 2016b), our analysis corroborates that the murine gut microbiome is dominated overall by two main phyla: Firmicutes (77% of MAGs and 73% of 16S rRNA gene sequences) and Bacteroidetes (14%/18%) (Figure 4). Bacteroidetes included the second-largest MAG cluster, namely, the family Muribaculaceae (64%/49%), which is recognized as abundant in the mouse gut (Lagkouvardos et al., 2016b, 2019). Strikingly, $\geq$13% of MAGs were from phylogenetic groups (e.g., family or order) that lacked reference genomes in NCBI RefSeq, such as the Clostridiales-vadinBB60 group (n = 70) and Mollicutes RF9 (n = 14) (Figure 4). None of these 84 MAGs had representative mMAGs (ANI >95%,

Figure 3. Mouse Gut Microbiota Optimized PICRUSt-iMGMC Model

(A) PICRUSt workflows used in this study: (I) Default workflow for an end user starting from close-reference-picked operational taxonomic units (OTUs) against the GreenGenes database relying on functional metagenome prediction using precalculated genome predictions files. (II) Novel PICRUSt workflow starting from *de novo*-picked OTUs and using iMAGs with 16S rRNA gene links to create ecosystem-specific functional metagenome predictions.

(B–E) For comparison of PICRUSt-KO profiles generated using default PICRUSt and PICRUSt-iMGMC from 16S rRNA gene amplicon sequencing with real KO profiles determined by shotgun metagenome sequencing (WGS), samples from different anatomical locations (n = 50) were analyzed. (B) Correlation between KO profiles of metagenomes determined by WGS and default PICRUSt (red) or by WGS and PICRUSt-iMGMC (green) using Pearson and Spearman correlation coefficients. Two-tailed paired t test was performed to analyze the differences; ****$p < 0.0001$. Data represent n = 50 samples and are displayed as a box-whisker plot representing minimum, first quartile, median, third quartile, and maximum. (C) Comparison of KO profiles generated using default PICRUSt (red), PICRUSt-iMGMC (green), and WGS (blue) from different anatomical locations. Non-metric multidimensional scaling (NMDS) was performed to visualize similarities. (D) False-positive rates and true-positive rates were obtained by comparing the default PICRUSt (red) and PICRUSt-iMGMC (green) KEGG module predictions against WGS results. The true-positive rate reflects the fraction of KEGG modules commonly predicted by both WGS and default PICRUSt/PICRUSt-iMGMC, and the false-positive rate reflects the fraction of KEGG modules predicted by default PICRUSt/PICRUSt-iMGMC but absent from WGS data. (E) KEGG module predictions that differ between default PICRUSt and PICRUSt-iMGMC predictions. KEGG module prediction by default PICRUSt and PICRUSt-iMGMC was compared with WGS for all samples, and significant differences in completeness were identified using a Wilcoxon test (false discovery rate [FDR] corrected). The heatmap displays select KEGG modules with highly similar completeness between PICRUSt-iMGMC and WGS but divergent completeness between default PICRUSt and WGS. See STAR Methods for details.

See also Figure S3.



(legend on next page)

GTDB), although related mMAGs were identified (Parks et al., 2017).

To increase recovery of MAGs from the mouse gut microbiome, we applied the scalable single-sample assembly approach to hundreds of additional samples (n = 576) (Table S1) from 36 more recent studies, resulting in the recovery of 13,619 mMAGs. For some additional libraries (n = 31), the single-sample assembly did not work properly, e.g., because of size or potential complexity; hence, we used Megahit for these. After joint dereplication (ANI < 95%) of mMAGs from all 874 samples, we obtained a set of 1,296 mMAGs representing a diverse collection of bacteria from the mouse gut, of which only 134 had a representative genome/mMAG (ANI >95%, GTDB) (Figure S4; Tables S2, S3, and S4). The mMAGs were also compared against the recently established Integrated Gut Genomes (IGG) database, which comprises a dereplicated collection of microbial genomes from the human gut recovered by metagenomics and sequencing of isolated strains (Nayfach et al., 2019). Of the 1,296 mMAGs, only 118 had a match in IGG (dRep, ANI >95%), of which only 19 did not have a match in GTDB. Hence, more than 88% of the species represent potentially novel species. Some of these mMAGs (388 of 1,296) are not directly contained in iMGMC, because they only resulted from single-sample assemblies, not the all-in one assembly, but they will be a resource for future studies.

As for the iMAGs, the comparison of reconstructed 16S rRNA gene sequences to several databases indicated a high fraction of them represents previously unknown sequences. For instance, only 164 of 1,323 (12%) had at least a 97% identical match in NCBI RefSeq (Table S6). Several taxonomic groups were represented by 16S rRNA gene sequences but underrepresented by MAGs, such as the family Prevotellaceae (49 16S rRNA gene sequences and 3 MAGs), the class Bacilli (81/10), and the phyla Proteobacteria (67/24) and Actinobacteria (78/22) (Figure S6). Thus, our analysis identified taxonomic groups that are interesting targets for future studies to extend our understanding of microbiome-modulated phenotypes in mouse models.

Provider-Specific Microbial and Functional Diversification of the Mouse Microbiota

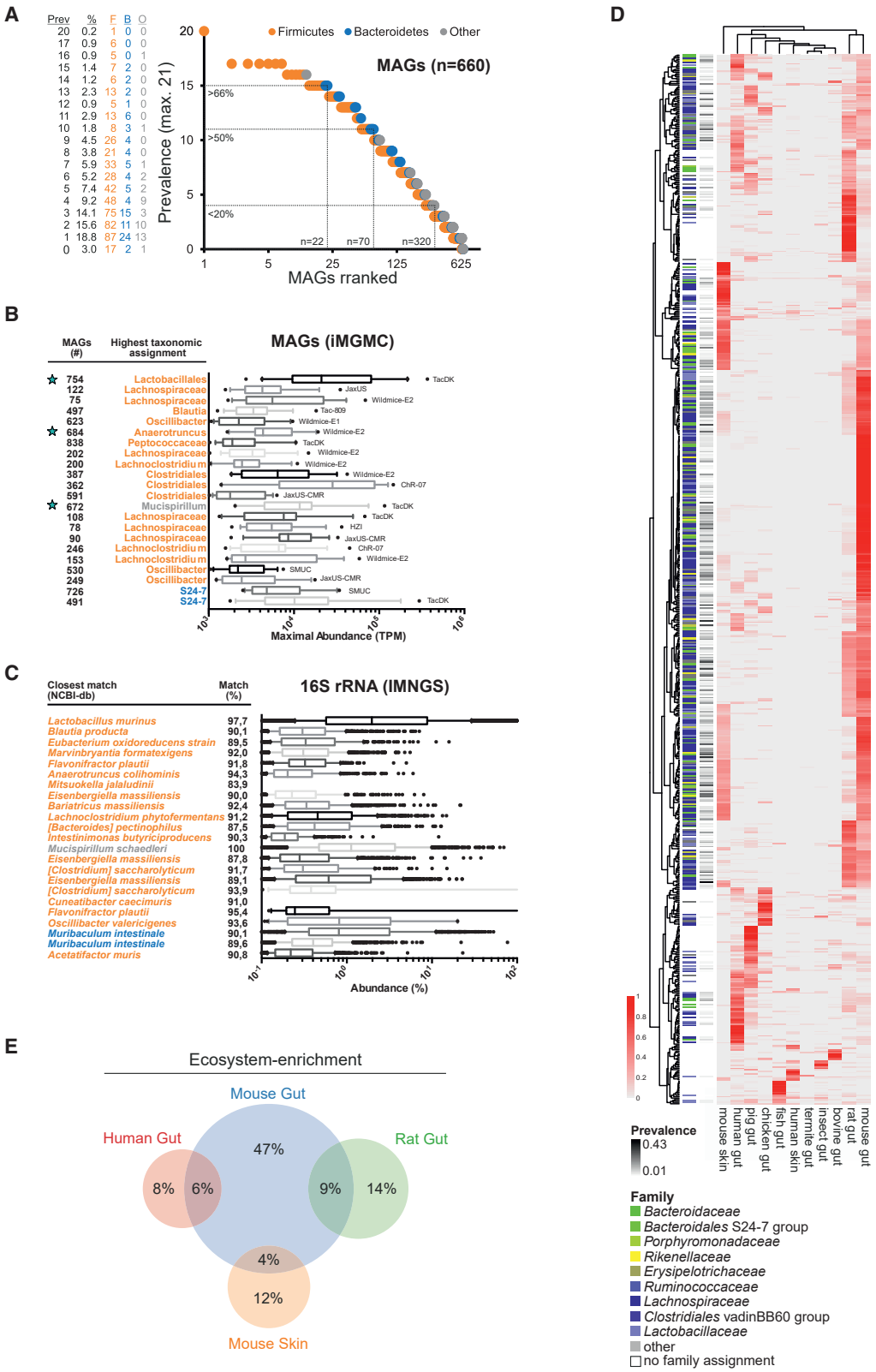
Studies have demonstrated mostly via 16S rRNA amplicon sequence analysis that the composition of murine microbiomes varies among providers (Rausch et al., 2016). Yet the presence of a core set of bacteria, based on the detection of 26 CAGs in >95% of mice, was proposed previously (Xiao et al., 2015). Hence, we analyzed the abundance of each iMAG in all 298 samples (see STAR Methods for details). Strikingly, each mouse line featured a unique combination of MAGs (Figure 4; Table S3). Around 10% of MAGs (70/660) were shared by at least half of the providers (Figure 5A). The most prevalent MAG, matching *Lactobacillus murinus* ASF361, was detected in almost all providers (20/21). Three additional members of the altered Schaedler flora (ASF) community, a gut microbiota model, and only four other previously sequenced bacteria were found in at least 50% of providers, while the remaining 62 (88%) represent uncultured bacteria. We next analyzed the MAGs shared by at least two-thirds of the providers (n = 21 MAGs), from which most belonged taxonomically to Firmicutes (n = 18), two belonged to the family Muribaculaceae (phylum Bacteroidetes), and one was identical to *Mucispirillum schaedleri* (phylum Deferribacteres) (Figure 5B). The relative abundance of these MAGs revealed differences among providers (up to 100-fold), suggesting that their respective abundance within each community is influenced by environmental factors.

Taking advantage of the link between MAG and 16S rRNA gene sequences, the global prevalence and relative abundance of the corresponding 16S rRNA gene sequences were assessed across all 16S rRNA amplicon datasets deposited in Sequence Read Archive (SRA) using IMNGS (integrated microbial NGS platform) (Figure 5C; Lagkouvardos et al., 2016a). The most prevalent MAG, *Lactobacillus murinus,* is present in 36% of all samples derived from the mouse gut (n = 9,496) but largely absent from the human gut and rat gut microbiota samples (1.4% positive) (Table S6). To assess whether the newly reconstructed 16S rRNA gene sequences represented taxa commonly found in mice, we employed IMNGS and queried all 1,323 16S rRNA gene sequences to assess their relative abundance in SRA samples derived from diverse ecosystems (Figures 5D and 5E). Of 569 sequences enriched in the mouse gut, mouse skin, rat gut, or human gut (see STAR Methods for details of selection), 44% were most prevalent in the mouse gut, and an additional 6% were shared with the mouse skin. Other sequences were shared with the rat microbiome (12%) and the human gut microbiome (7%) (Figure 5E), corroborating the high host-specific speciation.

To assess the potential functional consequences of differences in microbiota composition, unsupervised clustering of all iMAG according to their functional potential was performed, demonstrating that distinct variable-abundant taxonomic clusters, such as the Clostridiales-vadinBB60 group or the family Muribaculaceae, represent functionally distinct microbes within the mouse microbiome (Figures 6A–6C; Lagkouvardos et al., 2019). These variations in MAGs contributed to substantial differences in the functional potential of the microbiome within each mouse line (Figure 6D; Table S5). Despite the differences in composition and functionality between mouse lines, the analysis of datasets from mice subjected to experimental diets allowed the retrospective identification of MAG networks, rather than gene clusters that show conserved changes in their relative

Figure 4. Phylogenetic Tree of the 660 iMAGs Included in iMGMC

MAGs are shown as triangles, and 64 closely related, previously sequenced bacteria used for comparison are shown as stars (genomes from NCBI RefSeq with a mapping rate of >50% coverage). The color of the triangles indicates their taxonomic association to different phyla, and the size of the triangles indicates the prevalence in all iMGMC samples. The phylogenetic tree was built based on CheckM marker genes. The names of some taxonomic clusters are displayed in full or abbreviated in the tree. For this taxonomic identification, we used the SILVA database. The inner rings show the relative abundance of the 660 iMAGs in the 21 investigated mouse providers (threshold: 0.1%). The last three rings visualize the relative abundance of 469 of 660 iMAGs at different anatomical sites (threshold: 0.1%; SI, small intestine). The outer bar plots show their respective maximal relative abundance.
See also Figure S4 and Table S4.



(legend on next page)

abundance induced by these diets in mice from different providers (Figure S5). In summary, our analysis revealed the presence of specific bacteria commonly found in mouse lines yet a high species level and consequently functional variability within the murine gut microbiome.

DISCUSSION

Gene catalogs, 16S rRNA gene databases, and more recently MAG collections commonly represent separate references for shotgun metagenome and 16S rRNA amplicon sequencing analyses (Li et al., 2014; Sunagawa et al., 2015; Xiao et al., 2015, 2016; Almeida et al., 2019; Pasolli et al., 2019). To overcome this separation, a resource that can serve as (1) a reference for the mouse gut microbiota and (2) a blueprint to generate integrated metagenome catalogs for less characterized microbial ecosystems was developed. The combination of iMGMC and a comprehensive MAG collection comprising predominantly novel taxa (<95% ANI) will allow scientists to analyze next-generation sequencing (NGS) data by mapping against iMGMC containing bacterial and non-bacterial genes or directly against the MAGs. The iMAG-16S rRNA gene pairs enabled the development of an ecosystem-optimized version of PICRUSt. We anticipate this to be widely adapted to predict metagenome profiles based on 16S rRNA amplicon sequencing data and suggest that ecosystem-optimized versions of PICRUSt will be resources.

For the establishment of the integrated gene catalog, methods identified to yield optimal results by the CAMI (Critical Assessment of Metagenome Interpretation) challenge (Sczyrba et al., 2017), e.g., for assembly of MAGs or binning when dealing with large datasets, were used and complemented with a novel approach linking MAGs and 16S rRNA sequences. The complementation of MAG reconstruction with a novel approach linking MAGs and 16S rRNA sequences, which was manually curated in iMGMC, builds on developments in the metagenomics field (Nawrocki et al., 2015; Parks et al., 2017; Zeng et al., 2017). The evaluation of the linking pipeline using a synthetic dataset supports the performance of the current approach, but future refinements will be required for application in large-scale studies processing thousands of samples, i.e., for the human microbiome.

For the construction, a distinct assembly strategy was used compared with large-scale metagenomics studies (Crits-Christoph et al., 2018; Almeida et al., 2019; Nayfach et al., 2019; Pasolli et al., 2019). The all-in-one approach generated, for our dataset, MAGs with quality comparable to that of the single-sample approach, but the number of obtained MAGs, as well as the strain heterogeneity, was higher. In line with recent observations (Pasolli et al., 2019), we believe the all-in-one approach is promising for studies that contain multiple samples from connected ecosystems, e.g., longitudinal sampling of individuals or sampling from cohabitated animals, allowing the reconstruction of lower-abundant MAGs. We also evaluated the utility of the all-in-one assembly approach for another large dataset by processing metagenomic sequencing data from the pig microbiome. From 287 fecal samples (1,758 Gbp) used to construct a previous reference gene catalog (Xiao et al., 2016), we obtained 12.2 million ORFs and 1,050 MAGs, representing a 58% and 45% increase, respectively, compared with the original work (unpublished data, Till R. Lesker).

However, two caveats of the all-in-one approach are (1) the potential collapse of different strains on MAGs, which thus have species-level representation, and (2) the limited scalability of the all-in-one approach for thousands of samples. Therefore, we provide for the mouse gut microbiome an additional set of dereplicated mMAGs that complements iMGMC for strictly genome-based analysis, as well as almost 20,000 non-dereplicated mMAGs that can be explored to analyze bacterial strain diversity to an extent similar to that of the human gut microbiome. Moreover, the availability of this separate set of MAGs allows the streamlined expansion of the MAG collection with metagenomic sequencing data from additional mouse lines and providers and sample-wise assembly, which is likely to increase the known diversity in the mouse microbiome, because the rarefaction curves for MAGs per sample indicate further growth (Figure S4).

Using the iMGMC resource, we were able to demonstrate that the mouse gut microbiome predominantly contains bacteria that

Figure 5. Identification of MAGs Shared between Laboratory Mice

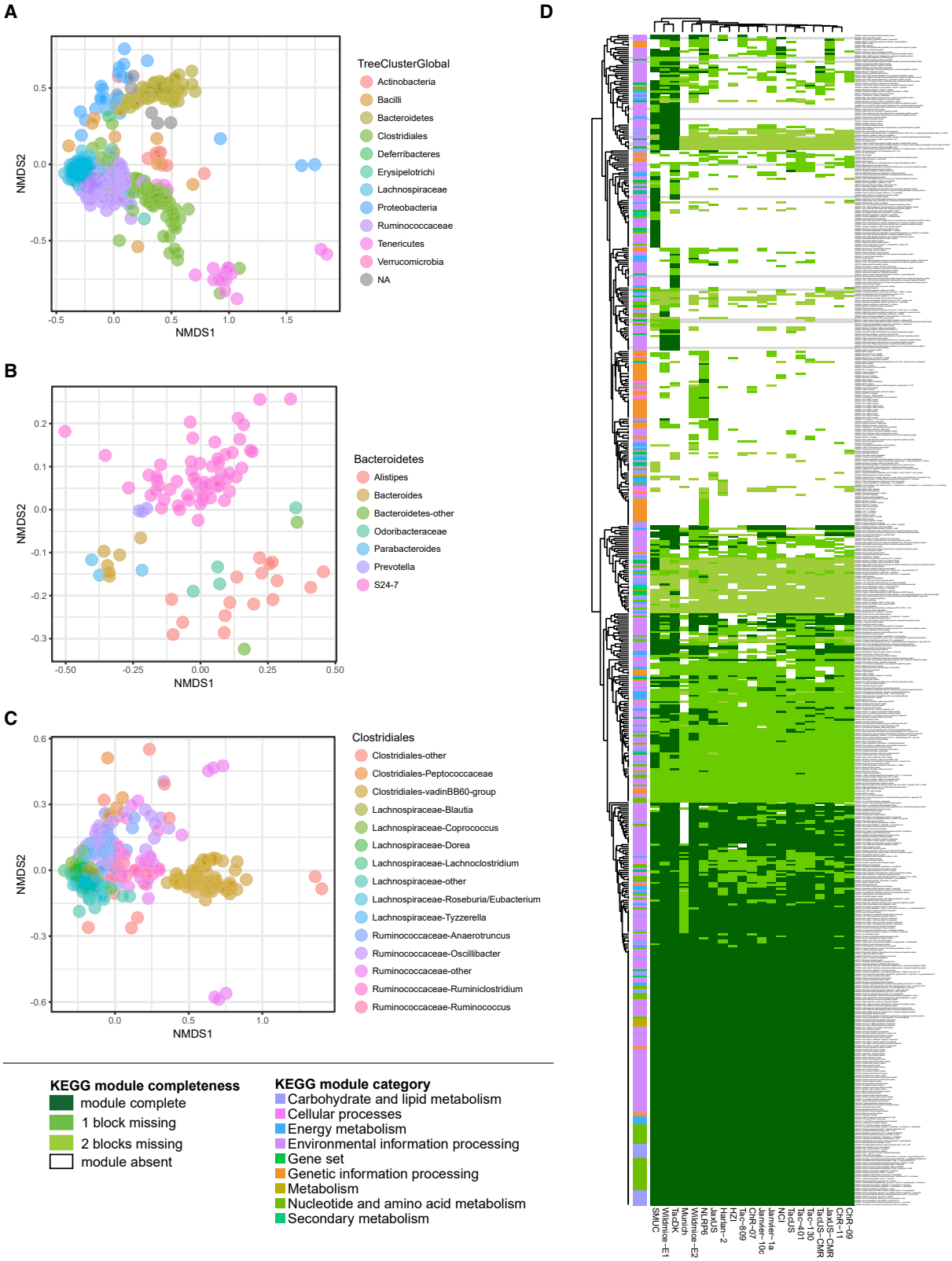
(A) Prevalence of iMAGs (n = 660) in samples from 21 mouse providers. iMAGs were considered present in a provider if its relative abundance reached at least 0.1% in one sample of the provider. Numbers on the left indicate the fraction (%) and taxonomic grouping (F, Firmicutes; B, Bacteroidetes; O, other phyla) of iMAGs with an indicated prevalence (Prev). In the right panel, iMAGs were ranked by prevalence, and dashed lines indicate the number of iMAGs present in >66%, >50%, and >20% of providers, respectively.

(B) Comparison of maximal abundance among providers for each iMAG (n = 22) present in at least two-thirds of providers. For each MAG, the bin number, the highest taxonomic assignment based on the manually curated phylogenetic tree, and the provider with the highest abundance are listed. Stars indicate iMAGs with matches in NCBI RefSeq. Data are displayed as a box-whisker plot representing 10%, first quartile, median, third quartile, and 90%.

(C) Comparison of the relative abundance of 16S rRNA gene sequences linked to MAGs in the IMNGS database. For each 16S rRNA gene, the closest named relative 16S rRNA gene sequence was determined and blasted to the NCBI-16S rRNA gene database. The color of the dots and names indicate their taxonomic association to different phyla (F, Firmicutes; B, Bacteroidetes; O, other phyla). Data are displayed as a box-whisker plot representing 10%, first quartile, median, third quartile, and 90%.

(D and E) IMNGS was used to determine the prevalence of iMGMC 16S rRNA gene sequences (n = 1,323) in distinct hosts and ecosystems. Of these, 1,113 reached at least a prevalence threshold of 1% prevalence within one of the evaluated environment (0.1% sample-depth cutoff of presence). Resulting sequences (n = 1,113) were filtered further to have at least 1% relative mean abundance in at least one environment. (D) Heatmap displaying the mean relative abundance within an ecosystem (row normalized) of those 16S rRNA gene sequences, which have at least 1% relative mean abundance in at least one environment (n = 739). (E) Venn diagram visualizing the distribution of 16S rRNA gene sequences subsampled to be enriched (>50% relative abundance normalized over the ecosystems in Figure 4D) in the mouse gut, mouse skin, rat gut, and human gut microbiome (n = 569). Numbers indicate the fraction of 16S rRNA gene sequences enriched or shared between indicated ecosystems.

See also Figure S5.



(legend on next page)

were neither cultured nor identified in other high-throughput sequencing studies (Parks et al., 2017). Our resource then allows, for instance, the identification of bacteria widely shared among mouse lines or the identification of bacterial networks that are concomitantly altered by dietary interventions in different mouse lines. Another utility of iMGMC is the availability of linked MAG-16S rRNA gene pairs, which enables the incorporation of data from large 16S rRNA gene databases such as the IMNGS database, encompassing 168,573 short-read datasets (build 1711), thereby allowing large-scale screening for identified MAGs, such as the evaluation of a core microbiome in the mouse gut. Finally, the MAG-16S rRNA gene pairs also enabled the development of an ecosystem-optimized version of PICRUSt, which produced gene profiles more closely resembling WGS data. We anticipate this to be widely adapted to predict metagenome profiles based on 16S rRNA amplicon sequencing data and suggest that ecosystem-optimized versions of PICRUSt will be resources.

Altogether, the clustering-free construction of gene catalogs with the reconstruction of numerous MAGs through complementary approaches and the linking of 16S rRNA gene sequences to iMAGs provide a highly integrated resource for sequencing-based work and will enable future studies to explore the taxonomy, functionality, and community structure of the mouse gut and other ecosystems in more depth. Strikingly, only 9% of the identified MAGs were shared with humans, corroborating the need for host-specific dedicated references.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.celrep.2020.02.036.

ACKNOWLEDGMENTS

T.S. was funded by the Helmholtz Association (project VH-NG-933), by the Ministry for Science and Culture of Lower Saxony (research consortium COALITION), by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, projects STR-1343/1 and STR-1343/2, as well as under Germany's Excellence Strategy EXC 2155 "RESIST," project 39087428), and the European Union (project StG337251). J.F.B. was funded by the DFG under Germany's Excellence Strategy EXC 22167 (project 390884018) and by the DFG Collaborative Research Center (CRC) 1182. T.C. received funding from the DFG (project CL481/2-1 and Project-ID 403224013 – SFB 1382, Gut-liver axis). The authors acknowledge the support by the Genome Analysis Platform of the Helmholtz Centre for Infection Research.

AUTHOR CONTRIBUTIONS

Conceptualization, T.R.L., A.C.M., and T.S.; Methodology, T.R.L., A.C.D., A.S., and A.C.M.; Investigation, T.R.L., A.C.D., E.J.C.G., and I.L.; Writing – Original Draft, T.R.L., A.C.D., and T.S.; Writing – Review & Editing, all authors; Funding Acquisition, T.C., J.F.B., and T.S.; Resources, T.C. and J.F.B.; Supervision, A.C.M. and T.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 25, 2019
Revised: October 2, 2019
Accepted: February 7, 2020
Published: March 3, 2020

REFERENCES

Almeida, A., Mitchell, A.L., Boland, M., Forster, S.C., Gloor, G.B., Tarkowska, A., Lawley, T.D., and Finn, R.D. (2019). A new genomic blueprint of the human gut microbiota. Nature *568*, 499–504.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403–410.

Figure 6. Analysis of Functional Diversity within Bacterial Members of the Mouse Gut Microbiota and between Mouse Providers Using iMGMC

(A–C) Ordination analysis of functional profiles of MAGs contained in iMGMC based on the presence of KOs. Comparison of all iMAGs (A, n = 660), as well as those with taxonomic assignment to the orders Bacteroidales (B, n = 94) and Clostridiales (C, n = 482). The distances reflect the differences in the functional capabilities of the MAGs according to the presence of KOs. Colors represent different taxonomic clusters according to the manually curated phylogenetic MAG tree (see Figure 4).

(D) To characterize the functional potential of each provider's microbiome, individual libraries (n = 299) were mapped to iMGMC. The mapped reads were used to quantify KOs present in each library. This information was translated to KEGG module completeness scores using KEGG's "Reconstruct Module" function and summarized per provider. The completeness of each KEGG module was expressed using a color code from dark green (module complete) to white (module absent).

See also Tables S5 and S6.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Asnicar, F., Weingart, G., Tickle, T.L., Huttenhower, C., and Segata, N. (2015). Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ *3*, e1029.

Aßhauer, K.P., Wemheuer, B., Daniel, R., and Meinicke, P. (2015). Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. Bioinformatics *31*, 2882–2884.

Blondel, V.D., Guillaume, J.L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. *2008*, P10008.

Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., Schulz, F., Jarett, J., Rivers, A.R., Eloe-Fadrosh, E.A., et al.; Genome Standards Consortium (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat. Biotechnol. *35*, 725–731.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60.

Cambuy, D.D., Coutinho, F.H., and Dutilh, B.E. (2016). Contig annotation tool CAT robustly classifies assembled metagenomic contigs and long sequences. bioRxiv. https://doi.org/10.1101/072868.

Caporaso, J.G., Bittinger, K., Bushman, F.D., DeSantis, T.Z., Andersen, G.L., and Knight, R. (2010a). PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics *26*, 266–267.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., et al. (2010b). QIIME allows analysis of high-throughput community sequencing data. Nat. Methods *7*, 335–336.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., Fierer, N., Knight, R., et al. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc. Natl. Acad. Sci. USA *108* (*Suppl 1*), 4516–4522.

Clavel, T., Lagkouvardos, I., Blaut, M., and Stecher, B. (2016). The mouse gut microbiome revisited: From complex diversity to model ecosystems. Int. J. Med. Microbiol. *306*, 316–327.

Crits-Christoph, A., Diamond, S., Butterfield, C.N., Thomas, B.C., and Banfield, J.F. (2018). Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. Nature *558*, 440–444.

Dröge, J., Gregor, I., and McHardy, A.C. (2014). Taxator-tk: Fast and Precise Taxonomic Assignment of Metagenomes by Approximating Evolutionary Neighborhoods. arXiv, arXiv:1404.1029.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460–2461.

Everard, A., Lazarevic, V., Gaïa, N., Johansson, M., Ståhlman, M., Backhed, F., Delzenne, N.M., Schrenzel, J., François, P., and Cani, P.D. (2014). Microbiome of prebiotic-treated mice reveals novel targets involved in host response during obesity. ISME J. *8*, 2116–2130.

Fabbiano, S., Suárez-Zamorano, N., Chevalier, C., Lazarević, V., Kieser, S., Rigo, D., Leo, S., Veyrat-Durebex, C., Gaïa, N., Maresca, M., et al. (2018). Functional Gut Microbiota Remodeling Contributes to the Caloric Restriction-Induced Metabolic Improvements. Cell Metab. *28*, 907–921.e7.

Fritz, A., Hofmann, P., Majda, S., Dahms, E., Dröge, J., Fiedler, J., Lesker, T.R., Belmann, P., DeMaere, M.Z., Darling, A.E., et al. (2019). CAMISIM: simulating metagenomes and microbial communities. Microbiome *7*, 17.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152.

Gregor, I., Dröge, J., Schirmer, M., Quince, C., and McHardy, A.C. (2016). PhyloPythiaS+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. PeerJ *4*, e1603.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics *29*, 1072–1075.

Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat. Commun. *9*, 5114.

Kamada, N., Seo, S.U., Chen, G.Y., and Núñez, G. (2013). Role of the gut microbiota in immunity and inflammatory disease. Nat. Rev. Immunol. *13*, 321–335.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. *45* (*D1*), D353–D361.

Kang, D.D., Froula, J., Egan, R., and Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ *3*, e1165.

Kang, D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ *7*, e7359.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol. Biol. Evol. *33*, 1870–1874.

Lagkouvardos, I., Joseph, D., Kapfhammer, M., Giritli, S., Horn, M., Haller, D., and Clavel, T. (2016a). IMNGS: A comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies. Sci. Rep. *6*, 33721.

Lagkouvardos, I., Pukall, R., Abt, B., Foesel, B.U., Meier-Kolthoff, J.P., Kumar, N., Bresciani, A., Martínez, I., Just, S., Ziegler, C., et al. (2016b). The Mouse Intestinal Bacterial Collection (miBC) provides host-specific insight into cultured diversity and functional potential of the gut microbiota. Nat. Microbiol. *1*, 16131.

Lagkouvardos, I., Lesker, T.R., Hitch, T.C.A., Gálvez, E.J.C., Smit, N., Neuhaus, K., Wang, J., Baines, J.F., Abt, B., Stecher, B., et al. (2019). Sequence and cultivation study of Muribaculaceae reveals novel species, host preference, and functional potential of this yet undescribed family. Microbiome *7*, 28.

Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.A., Clemente, J.C., Burkepile, D.E., Vega Thurber, R.L., Knight, R., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat. Biotechnol. *31*, 814–821.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Levy, M., Thaiss, C.A., Zeevi, D., Dohnalová, L., Zilberman-Schapira, G., Mahdi, J.A., David, E., Savidor, A., Korem, T., Herzig, Y., et al. (2015). Microbiota-Modulated Metabolites Shape the Intestinal Microenvironment by Regulating NLRP6 Inflammasome Signaling. Cell *163*, 1428–1443.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., Arumugam, M., Kultima, J.R., Prifti, E., Nielsen, T., et al.; MetaHIT Consortium; MetaHIT Consortium (2014). An integrated catalog of reference genes in the human gut microbiome. Nat. Biotechnol. *32*, 834–841.

Li, D., Luo, R., Liu, C.M., Leung, C.M., Ting, H.F., Sadakane, K., Yamashita, H., and Lam, T.W. (2016). MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods *102*, 3–11.

Miller, C.S., Baker, B.J., Thomas, B.C., Singer, S.W., and Banfield, J.F. (2011). EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. Genome Biol. *12*, R44.

Nawrocki, E.P., Burge, S.W., Bateman, A., Daub, J., Eberhardt, R.Y., Eddy, S.R., Floden, E.W., Gardner, P.P., Jones, T.A., Tate, J., and Finn, R.D. (2015). Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. *43* (*D1*), D130–D137.

Nayfach, S., Shi, Z.J., Seshadri, R., Pollard, K.S., and Kyrpides, N.C. (2019). New insights from uncultivated genomes of the global human gut microbiome. Nature *568*, 505–510.

Nelson, M.C., Morrison, H.G., Benjamino, J., Grim, S.L., and Graf, J. (2014). Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. PloS One *9*, e94249.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). metaSPAdes: a new versatile metagenomic assembler. Genome Res. *27*, 824–834.

Olm, M.R., Brown, C.T., Brooks, B., and Banfield, J.F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. *11*, 2864–2868.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. *25*, 1043–1055.

Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., and Tyson, G.W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat. Microbiol. *2*, 1533–1542.

Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.A., and Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. *36*, 996–1004.

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., et al. (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell *176*, 649–662.e20.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2—Approximately Maximum-Likelihood Trees for Large Alignments. PLoS One *5*, e9490.

Pruesse, E., Peplies, J., and Glöckner, F.O. (2012). SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics *28*, 1823–1829.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al.; MetaHIT Consortium (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59–65.

R Core Team (2016). R: A language and environment for statistical computing (R Foundation for Statistical Computing). https://www.R-project.org/.

Rausch, P., Basic, M., Batra, A., Bischoff, S.C., Blaut, M., Clavel, T., Gläsner, J., Gopalakrishnan, S., Grassl, G.A., Günther, C., et al. (2016). Analysis of factors contributing to variation in the C57BL/6J fecal microbiota across German animal facilities. Int. J. Med. Microbiol. *306*, 343–355.

Rosshart, S.P., Vassallo, B.G., Angeletti, D., Hutchinson, D.S., Morgan, A.P., Takeda, K., Hickman, H.D., McCulloch, J.A., Badger, J.H., Ajami, N.J., et al. (2017). Wild Mouse Gut Microbiota Promotes Host Fitness and Improves Disease Resistance. Cell *171*, 1015–1028.e13.

Schäfer, J., and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat. Appl. Genet. Mol. Biol. *4*, Article32.

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., et al. (2017). Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. Nat. Methods *14*, 1063–1071.

Suez, J., Korem, T., Zeevi, D., Zilberman-Schapira, G., Thaiss, C.A., Maza, O., Israeli, D., Zmora, N., Gilad, S., Weinberger, A., et al. (2014). Artificial sweeteners induce glucose intolerance by altering the gut microbiota. Nature *514*, 181–186.

Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., et al.; Tara Oceans coordinators (2015). Ocean plankton. Structure and function of the global ocean microbiome. Science *348*, 1261359.

Thiemann, S., Smit, N., Roy, U., Lesker, T.R., Gálvez, E.J.C., Helmecke, J., Basic, M., Bleich, A., Goodman, A.L., Kalinke, U., et al. (2017). Enhancement of IFNγ Production by Distinct Commensals Ameliorates *Salmonella*-Induced Disease. Cell Host Microbe *21*, 682–694.e5.

Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.A., Affourtit, J.P., et al. (2009). A core gut microbiome in obese and lean twins. Nature *457*, 480–484.

Wang, Q., Garrity, G.M., Tiedje, J.M., and Cole, J.R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. *73*, 5261–5267.

Xiao, L., Feng, Q., Liang, S., Sonne, S.B., Xia, Z., Qiu, X., Li, X., Long, H., Zhang, J., Zhang, D., et al. (2015). A catalog of the mouse gut metagenome. Nat. Biotechnol. *33*, 1103–1108.

Xiao, L., Estellé, J., Kiilerich, P., Ramayo-Caldas, Y., Xia, Z., Feng, Q., Liang, S., Pedersen, A.Ø., Kjeldsen, N.J., Liu, C., et al. (2016). A reference gene catalogue of the pig gut microbiome. Nat. Microbiol. *1*, 16161.

Zeng, F., Wang, Z., Wang, Y., Zhou, J., and Chen, T. (2017). Large-scale 16S gene assembly using metagenomics shotgun sequences. Bioinformatics *33*, 1447–1456.

Zhu, W., Lomsadze, A., and Borodovsky, M. (2010). *Ab initio* gene identification in metagenomic sequences. Nucleic Acids Res. *38*, e132.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
Stool samples from mouse lines	This study	PRJEB32890
Stool samples from mouse lines (MGMCv1)	http://doi.org/10.1038/nbt.3353	PRJEB7759
Stool samples from mouse lines (Wild mice)	This study	PRJEB32890
Stool samples from mouse lines (Munich)	http://doi.org/10.1038/nmicrobiol.2016.131	PRJEB10572
Deposited Data		
Metagenomic sequencing data	This study	PRJEB32890
Assembled sequencing data	This study	https://zenodo.org/record/3631711
Software and Algorithms		
BBMap	https://sourceforge.net/projects/bbmap/	N/A
bowtie	(Langmead et al., 2009)	https://github.com/BenLangmead/bowtie2
bwa	(Li and Durbin, 2009)	https://github.com/lh3/bwa
CD-HIT	(Fu et al., 2012)	https://github.com/weizhongli/cdhit
CheckM	(Parks et al., 2015)	https://github.com/Ecogenomics/CheckM
FastTree	(Price et al., 2010)	https://github.com/PavelTorgashov/FastTree
GeneMark.hmm	(Zhu et al., 2010)	https://github.com/aghozlane/spasm/tree/master/MetaGeneMark
GraphPad Prism	GraphPad Software, Inc.	https://www.graphpad.com/scientific-software/prism/
MegaHit	(Li et al., 2016)	https://github.com/voutcn/megahit
MetaBAT	(Kang et al., 2015)	https://bitbucket.org/berkeleylab/metabat/src/master/
MetaBAT2	(Kang et al., 2019)	https://bitbucket.org/berkeleylab/metabat
metaSPAdes	(Nurk et al., 2017)	https://github.com/ablab/spades/releases
MUSCLE	(Edgar, 2004)	https://www.drive5.com/muscle/
NCBI blast	(Altschul et al., 1990)	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast
PICRUSt	(Langille et al., 2013)	https://github.com/picrust/picrust
R version 3.X	R Core Team	https://www.r-project.org/
RAMBL	(Zeng et al., 2017)	https://github.com/homopolymer/RAMBL/
Usearch	(Edgar, 2010)	https://drive5.com/usearch/
Code for 16S/MGS linking	This study	https://github.com/strowig-lab/iMGMC
Other		
Database: IMNGS	(Lagkouvardos et al., 2016a)	https://www.imngs.org/

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and software should be directed to and will be fulfilled by the Lead Contact: Till Strowig (till.strowig@helmholtz-hzi.de). This study did not generate new unique reagents.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Experimental animals were obtained from vendors or the animal facility of the Helmholtz Centre for Infection Research, which operate in accordance to European and German law. Specifically, 8 - 16 week-old male and female (approximate ratio 50:50%) WT C57Bl6/N and *Nlrp6*$^{-/-}$ mice were housed in individually-ventilated cages with free access to autoclaved water and chow. Before removal of intestinal contents mice were humanly euthanized using CO_2 inhalation with a low flowrate.

METHOD DETAILS

Sample Collection and DNA Extraction

For the *de novo* generation of metagenomic sequencing data, luminal fecal content was collected from different gastrointestinal GI sites (Ileum: SI, Cecum: Cec and Colon: Col) of mice obtained from different vendors and stored at −20°C until processing. DNA was isolated using an established protocol (Turnbaugh et al., 2009). Briefly, each sample was treated with 500 μl of extraction buffer (200 mM Tris, 20 mM EDTA, 200 mM NaCl, pH 8.0), 200 μl of 20% SDS, 500 μl of phenol:chloroform:isoamyl alcohol (24:24:1) and 100 μl of zirconia/silica beads (0.1 mm diameter). Samples were homogenized with a bead beater (BioSpec) for 2 min. DNA was precipitated with absolute isopropanol and finally washed with 70% ethanol. DNA extracts were resuspended in TE Buffer with 100 μg/ml RNase I and finally, column purified to remove traces of PCR inhibitors.

Metagenomic Sequencing

Total DNA was quantified and diluted to 25 ng/μl. 60 μl of total DNA was used for shearing by sonication (Covaris). Fragmentation was performed as follow: Processing time = 150 s, Fragment size = 200 bp, Intensity = 5, duty cycle = 10. Illumina library preparation was performed using the NEBNext Ultra DNA library prep kit (New England Biolabs). The library preparation was performed according to the manufacturer's instructions. We use as input a total of 500 ng of DNA, the size selection was performed using AMPure XP beads (First bead selection = 55 μl, and second = 25 μl). Adaptor enrichment was performed using seven cycles of PCR using the NEBNext Multiplex oligos for Illumina (Set 1 and Set 2)(New England Biolabs) and then subjected to Illumina HiSeq2000 PE100 sequencing. Source and sequencing depth for each sample are listed in Table S1.

16S rRNA Gene Amplification, Sequencing, and Data analysis

Amplification of the V4 region (F515/R806) of the 16S rRNA gene was performed according to previously described protocols (Caporaso et al., 2011; Thiemann et al., 2017). Briefly, for DNA-based amplicon sequencing 25 ng of DNA were used per PCR reaction (30 μl). The PCR conditions consisted of initial denaturation for 30 s at 98°C, followed by 25 cycles (10 s at 98°C, 20 s at 55°C, and 20 s at 72°C. Each sample was amplified in triplicates and subsequently pooled. After normalization PCR amplicons were sequenced on an Illumina MiSeq platform (PE250). Obtained reads were assembled, quality controlled, and clustered using the QIIME v1.8.0 (Quantitative Insights Into Microbial Ecology) analysis pipeline (Caporaso et al., 2010b). In short, quality filtering was set to -q 30, minimum read length 200 bp and a minimum number of sequences per sample = 1000. The OTU clusters and representative sequences were determined using open-reference OTU picking (Nelson et al., 2014) using UCLUST (Edgar, 2010) at 97% identity, followed by taxonomy assignment using the RDP Classifier (Wang et al., 2007) with a bootstrap confidence cutoff of 80%. The OTU absolute abundance table and mapping file were used for statistical analyses and data visualization in the R statistical programming environment (R Core Team, 2016).

Construction of the iMGMC

i) Assembly and prediction of ORFs

Demultiplexed libraries were filtered to remove host reads using BBMap (parameters see code) using the Ensembl masked mouse genome GRCm38.75 and phiX. All mouse filtered metagenomic libraries were used in single "all in one" assembly approach using Megahit (Li et al., 2016) with specific parameters (-kmin 5 -k 27,37,47,57,67,77,87,97) using a SGI-UV2000 cluster with 256 cores and 2 TB shared memory. Resulting contigs were filtered to minimum 1000bp lengths and renamed with numbers from largest to smallest contig. For protein prediction, we used Metagenemark (Zhu et al., 2010) (parameters see code). ORFs were filtered to remove ORF shorter than 100bp after which they were reordered and renamed according to their length (Figures 1A and 1D).

ii) Binning and evaluation of binning using CheckM

All libraries were mapped with BWA (Li and Durbin, 2009) (default parameters) to the contigs. The mapping results were transformed and indexed to bam-format using sambawa. Metagenome binning was performed with MetaBAT (Kang et al., 2015)b(version 0.32) using the following parameters -verysensitive -pB 20 -B 100 -minclustersize 200000. The resulting clusters were evaluated with CheckM (Parks et al., 2015). To assign a bin to an integrated MAGs (metagenome assembled genome, iMAG) we used a threshold of marker gene completeness - contamination > = 80%. All other bins of contigs with at least 200kbp length were defined as co-abundance groups (CAGs). We used the marker gene alignment from CheckM derived from all 660 MAGs and 64 selected genomes from NCBI RefSeq to construct a phylogenetic tree using a nearest neighbor joining approach with 1000 bootstraps in MEGA7 (Kumar et al., 2016). The tree was plotted using GraPhlAn (Asnicar et al., 2015).

iii) Taxonomic classification

The taxonomic classification for all gene entries in the catalog was performed based on different levels, i.e., ORF, contig and bin/CAG/MAGs. For ORFs, assignments were performed using CAT (Cambuy et al., 2016) and DIAMOND (Buchfink et al., 2015) against the NCBI NR protein database. For all contigs and bins the classification was performed by taxator (Dröge et al., 2014), PhyloPythiaS+ (Gregor et al., 2016), GTDB-Tk (Parks et al., 2018) and CAT using default parameters.

Furthermore, we additionally included for the MAGs taxonomic information from the placement of MAGs within the phylogenetic tree as well as the information from the linked 16 s rRNA gene sequences.

iv) Functional annotation of gene catalog proteins

All proteins were annotated using blastp (Altschul et al., 1997) against the KEGG gene database (01/2018) (Kanehisa et al., 2017) following the best hit approach (e-value 0.001). KEGG Ortholog annotation was used to reconstruct KEGG module completeness. For the annotation of ORFs within MAGs, we used the respective linked annotation data for each ORF from iMGMC. The R statistical programming environment was used for statistical analyses and data visualization.

v) Full-length 16S rRNA sequence reconstruction, annotation and phylogeny

We used RAMBL (Zeng et al., 2017) to reconstruct full-length 16S rRNA gene sequences from all libraries in one batch. Resulting sequences (n = 1,323) were classified with SINA (Pruesse et al., 2012) using the SILVA NRref database (version 123). The phylogenetic tree was built using the nearest neighbor joining method (maximum likelihood) with 1000 bootstraps using MEGA7 (Kumar et al., 2016).

vi) MAG to 16S rRNA gene connections via multi-scale linkage

The linking pipeline incorporates three different approaches: First, we searched for integrated 16S rRNA sequences in the assembled contigs of all clustered bins including CAGs and MAGs. Therefore, we mapped with BlastN all contigs to all reconstructed 16S rRNA gene sequences. We removed alignments of less than 100bp and lower than 95% of identity, resulting in a matrix of blast scores of each bin to each 16S rRNA gene sequence.

Second, we used in parallel a scaffolding approach utilizing the information from paired-end read sequencing. Therefore, the reads from all libraries were partitioned into new libraries by mapping against all bins (n = 1462) with BBSplit. Then, the new libraries were mapped against all 1,323 reconstructed 16S rRNA gene sequences. Unambiguous and ambiguous mapped reads were counted separately into two matrices of all bins x 16S rRNA gene sequences.

The third method uses the abundance profiles across all samples to correlate 16S rRNA gene sequences and bins. To determine abundance profiles over all 298 samples, we mapped all libraries individually to all bins including CAGs and MAGs as well as against unbinned contigs and in parallel to all reconstructed full- length 16S rRNA gene sequences. Read counts of bins were transformed to TPM (transcripts per million) and stored in an abundance matrix, as well as the unambiguous 16S rRNA gene sequence counts. Pearson and Spearman correlation were calculated for both abundance matrices to obtains scores for all bins to all 16S rRNA gene sequences.

Finally, data of all three approaches were weighed in an integration scoring to aim in associating the bins of iMGMC to the corresponding 16S rRNA gene sequences and their respective annotations, in different steps:

(I) *Indirect association:* We used the normalized abundance values of the bins and 16S rRNA gene sequences to obtain their corresponding correlation (both Pearson and Spearman). We estimated consensus interdependence scores from both the correlation methods between any 16S rRNA gene and bin pair by integrating correlation values between the bins and 16S rRNA genes by taking the geometric mean of both the correlation values between each bin and 16S rRNA gene and assigning a negative sign if either of these correlation values was negative.

$$V(x,y) = value[I(x,y)] = \sqrt{abs(P(x,y)) * abs(S(x,y))}$$

$$Sg(x,y) = sign[(I(x,y))] = \begin{cases} -, & any\ [P(x,y), S(x,y)] < 0 \\ +, & else \end{cases}$$

$$I(x,y) = V(x,y) * Sg(x,y)$$

Where

$P(x,y)$ = Pearson correlation between a 16S rRNA gene 'x' and metagenome bin 'y'

$S(x,y)$ = Spearman correlation between a 16S rRNA gene 'x' and metagenome bin 'y'

$I(x,y)$ = Integrated correlation between a 16S rRNA gene 'x' and metagenome bin 'y'

For the highly correlated bin / 16S rRNA gene pairs, the Pearson and Spearman correlation values have very small difference. The normalization described above widened the distance between the true positives and false positives.

(II) Direct association:

1. Mapping bins to 16S rRNA gene sequences [$M(x,y)$]: These quantify the fraction of reads in a bin 'y' containing matching reads from 16S rRNA gene 'x' by mapping the reads in bin 'y' to the 16S rRNA gene 'x'. We normalized the number of uniquely mapped reads in bin 'y' to a 16S rRNA gene 'x' $m(x,y)$ by the total number of 16S reads mapped to the bin 'y' [$\sum_{i=1}^{n} m(x,i)$].

$$\mathrm{M}(x,y) = \frac{m(x,y)}{\sum_{i=1}^{n} m(x,i)}$$

Where n is the number of bins

2. BLAST bins to 16S rRNA gene sequences [$B(x,y)$]: These quantify the fraction of reads in a bin 'y' containing reads in 16S rRNA gene 'x' aligning of the reads of 16S rRNA gene 'x' to the bin 'y' using BLAST. We normalized the number of uniquely mapped reads in bin 'y' to a 16S rRNA gene 'x' $b[(x,y)]$ by the maximum of reads from 16S rRNA genes mapped to the bin 'y' $[\max_{0<i\leq n} b(x,i)]$.

$$\mathrm{B}(x,y) = \frac{b(x,y)}{\max\limits_{0<i\leq n} b(x,i)}$$

Where n is the number of bins.

(III) Integrating the direct and indirect associations between bin and 16S rRNA gene sequences:

The direct associations are sparse, i.e., there are very few 16S rRNA gene sequences reads present in each bin, while the indirect associations are not sparse. Hence, we integrated the scores in a way that does not allow the indirect associations to dominate over the direct associations. For this, we integrated the three scores [I(x,y), $M(x,y), B[(x,y)]$, as done in the STRING database. The only difference between the STRING database and the scoring scheme employed here is that for combining scores we took a geometric mean of the dissimilarity scores while combining them instead of simply multiplying the different scores (as done in STRING database).

$$F = 1 - \sqrt[3]{(1-I)*(1-B)*(1-M)}$$

Where F is the combined score for the bins – 16S rRNA gene sequences relationship.

We observed that the integrated correlation scores [$I(x,y)$] representing the indirect association tended to dominate over the direct association scores in several instances. Hence, we regularized the indirect association score by multiplying the Pearson and Spearman correlation values, instead of calculating their geometric mean:

$$V_{reg}(x,y) = value\left[I_{reg}(x,y)\right] = abs(P(x,y)) * abs(S(x,y))$$

$$I_{reg}(x,y) = V_{reg}(x,y) * Sg(x,y)$$

$$F_{reg} = 1 - \sqrt[3]{\left(1-I_{reg}\right)*(1-B)*(1-M)}$$

Where

V_{reg}: Regularized integrated correlation value.

I_{reg}: Regularized integrated correlation score.

F_{reg}: Regularized combined score for the bin / 16S rRNA genes relationship.

The negative values are turned zeros. The closer the F_{reg} value to 1, the higher the confidence of the bin – 16S rRNA gene sequences relationship. However, the 16S rRNA gene sequence 'x' might have the highest confidence score to the metagenome bin 'y', but the metagenome bin 'y' need not have the highest confidence score to rRNA gene 'x'. To address this issue, we enriched these relationships by normalizing these scores by the highest confidence scores of the corresponding metagenome bin 'y' and rRNA gene 'x'.

(IV) Enriching bin to 16S rRNA gene relationship:

We estimated the probability of a metagenome bin 'y' to rRNA gene 'x' relationship:

$$\Pr(x,y) = \frac{F_{reg}(x,y)}{\max\limits_{0<i\leq m} F_{reg}(i,y)} * \frac{F_{reg}(x,y)}{\max\limits_{0<j\leq n} F_{reg}(x,j)}$$

Where n is the number of bins and m is the number of 16S rRNA genes.

The obtained normalized confidence score or the estimated probability is the statistical likelihood of the confidence scores, adjusted for the background distribution of the confidence scores for all possible 16S rRNA gene / bin pair relationships.

Evaluation of iMGMC

i) A reference-based evaluation using QUAST

To identify genomes present in the MAGs GTDB was searched for microbial genomes with an ANI of at least 99% to a cluster which consists of MAGs in the three different assembly and binning approaches (see Table S2). We selected genomes labeled as NCBI RefSeq-assemblies to guarantee the quality of the reference, 26 genomes fulfilled these criteria. QUAST was used to evaluate the quality of the assembly approaches (Gurevich et al., 2013) (see Figure S2).

ii) Assessing binning efficiency using know NCBI reference genomes recovered as MAGs

To evaluate the binning of contigs to a higher order, those reference genomes being contained within the assembly were identified. Therefore, synthetic reads (100 bp) from all 9,748 bacterial genomes available in the NCBI Assembly database (Version January 2017) were generated with BBMap and mapped against all contigs using bowtie (Langmead et al., 2009). Genomes that were contained to at least 50% within the contigs were selected for evaluation (n = 57). Specifically, the binning efficiency for each reference genome was evaluated by quantifying the distribution of the synthetic reads over the bins and unbinned contigs. The analysis contained both the total proportion of reads mapped to contigs (= total recovered genome fraction) as well as the fraction of reads contained within contigs and mapping to one or more bins (= binned genome fraction).

ii) Assessing the bins to 16S rRNA gene linking approach with reference genomes

To evaluate the linking approach, those NCBI reference genomes which constitute part of MAGs were identified. Therefore, synthetic reads (100 bp) from all 9,748 bacterial genomes available in the NCBI Assembly database (Version January 2017) were generated with BBMap and mapped against all MAGs. Those mapping to at least 50% to a single MAGs, were used for evaluating the MAG / 16S rRNA gene links. First, the 16S rRNA gene sequences of the NCBI genomes were matched to the best-reconstructed 16S rRNA gene sequences via BlastN and the identity was calculated. Then, these reference sequences were compared to the predicted 16S rRNA gene sequence from the linking approach and the taxonomic agreement between these sequences was scored. Optimally, an identical match of reference 16S rRNA gene sequence to the linked 16S rRNA gene sequence would be obtained by the scoring scheme.

iii) Assessing the bins to 16S rRNA gene linking approach with simulated dataset

We evaluate the linking approach on simulated data. We use the 2nd CAMI Toy Mouse Gut Dataset (Fritz et al., 2019, PMID: 30736849) following our original pipeline. The dataset was created with 64 abundances profiles and 791 reference genomes. In brief we perform a pooled assembly of all 64 samples with MegaHIT (Li et al., 2016), followed by a binning of contigs with MetaBAT2 (Kang et al., 2019) to mMAGs (n = 438). RAMBL (Zeng et al., 2017) was using to reconstruct 16S rRNA gene sequences (n = 460). Gold standard (MAG - > RABL-16S-sequence) were created by mapping MAGs to reference genomes with FastANI. Reference genomes were assigned to reconstructed 16S rRNA sequences using BlastN (min ident 97% and min coverage 100bp) and used as a gold standard mapping. 204 MAGs reaching the quality criteria (CheckM completeness -contamination $\leq$ 80%) and of 163 it was possible to assign a reconstructed 16S rRNA sequence. Our linking approach predicted for 103 (63.2%) MAGs the best possible reconstructed 16S rRNA gene sequence (gold standard agreement). From the remaining 60 connections, 29 were filtered out by agreement of the taxonomic classification (disagreement on family level). These connections were replaced by predictions of our alternative linking approach (RAMBL-16S to MAGs). This step map 6 additional MAGs to the correct 16S rRNA gene sequence. Resulting final correct mappings are 109 of 163 (66.9%), for 23 (14.1%) no predictions were possible, 31 (19%) are incorrect links. For 15 of these 31 links represent closely related hit to the gold standard, the remaining 16 connections (9.8%) are distinct to the gold standard. Closely related match defined as first to third next 16S-rRNA sequence in list ordered by similarity created by an alignment with muscle (Edgar, 2004).

"Single-Sample" Assembly and Binning

Sample-wise assembly was conducted essentially as recently described (Pasolli et al., 2019). Briefly, the metagenomic assembly was performed using metaSPAdes (Nurk et al., 2017) or Megahit (Li et al., 2016) in default mode. Then sample-specific contig binning was performed using MetaBAT2 (Kang et al., 2019) followed by controlling genome completeness and contamination using CheckM (Parks et al., 2015).

Species-Level Clustering of MAGs

To dereplicate MAGs, we used dRep (Olm et al., 2017) into species-level OTUs estimation using Mash on the basis of 95% whole-genome nucleotide similarity, which is consistent with the definition of known species (Jain et al., 2018). All mMAGs were filtered by the MIMAG (Bowers et al., 2017) medium-quality standard (Completeness > = 50%, Contamination < 10%) or similar to MIMAG high-quality standard (Completeness > 90%, Contamination < 5%) to hqMAGs using CheckM metrics (Parks et al., 2015). By default dRep use a completeness filtering of 75% to ensure a suitable aligned fraction, therefore we kept this setting for the evaluation of the "single-sample" versus "All-in-one" assembly and binning approach (Figure S2). Results of the different MAG subsets: "All-in-one ", "single-sample," "single-sample extra studies" can be found in Table S3.

Abundance Estimation of Genomes (TPM)

Libraries were mapped against all MAGs using BBMap. For normalization, the read counts were divided by genome length in kilobases minus 50 bp. The resulting reads per kilobase (RPK) were counted up and divided by 1,000,000 (PMSF: per million scaling factor). TPM = RPK/PMFS of each genome bin.

PICRUSt

To test if the MAGs linked with reconstructed 16S rRNA sequences represent a large part of the mouse gut catalog we created an extended genomic reference PICRUSt prediction model: 484 MAGs with unique linked 16S sequences were used according to the PICRUSt "Genome Prediction Tutorial": 1) Determination of 16S copy numbers was performed by rrnDB Estimate (version 5.2.), 2)

KEGG Orthology (KO) profiles of the MAGs were extracted from iMGMC. 3) A tree (Edgar, 2004; Price et al., 2010) of RAMBL reconstructed 16S sequences was used to build the models. Furthermore, to verify the prediction power of the model we added the KO profiles of 3772 KEGG genomes to iMGMC model. Moreover, use the sequences of the GreenGenes database (Version 13.5 OTU-RepSet 97) together with the iMGMC reconstructed 16S sequences and the 16S of the KEGG genomes to build an integrated prediction model.

To make our PICRUSt models accessible for de-novo clustered OTUs, we modified our pipeline in a way that for each dataset, a new PICRUSt model is created from scratch: Sequences of the OTUs where include by pynast (Caporaso et al., 2010a) into a pre-calculated alignment of the reconstructed 16S rRNA genes. A tree generated by FastTree (Price et al., 2010) is be used for the PICRUSt genome predictions to create pre-calculated PICRUSt models that include the denovo OTUs for the mapping into the following standard workflow.

Global Distribution of iMGMC 16S rRNA Gene Sequences in NCBI-SRA (IMNGS Analysis)

We checked all 1,323 reconstructed full-length 16S rRNA gene sequences using the IMNGS pipeline (Lagkouvardos et al., 2016a) for their prevalence and relative abundance in 168,573 SRA samples (build 1711). To evaluate the specific environment of an OTU, we looked also at its relative abundance in related mouse gut environments such as the skin (mouse, human) and different gut sites (human, rat, bovine, chicken, fish, insect, pig, termite) where at least 100 samples were available. We filter for 16S rRNA less than 0.1% abundance in one of the selected environments. Furthermore, we checked for a specific preference of an OTU in the mouse gut, human gut, and rat gut, by using a relative abundance normalized over the environments of at least 50%. Other OTUs where moreover checked to be dominant in combination with the mouse gut site together with mouse skin, human gut or rat gut.

Identification of Sub-communities in the Intestinal Bacterial Community

We obtained the co-abundances between all metagenome bin-pairs using a shrinkage approach to the correlation estimation, as described in Schäfer and Strimmer (2005) (and available of cor.shrink function in corpcor library in R). We removed the co-abundance values less than 0.5 and inferred the sub-communities using a modularity optimization algorithm described in Blondel et al. (2008) (and available as cluster_louvain function in igraph library in R). Among all the sub-communities obtained, we were interested in those sub-communities of more than 5 members.

Association between sub-communities and diet across multiple vendor mouse gut-microbial communities:

We performed multi-factor ANOVA address relations of difference in dietary supplements and the difference in mouse strains over the abundance of the metagenome bins. For this, we modeled the metagenome bins abundance data against the interaction effect of the metagenome bin and dietary supplements, the dietary supplements and the mouse strains of samples from every vendor separately. Any sub-community which showed a significant response to difference in diet, metagenome bin, and the interaction effect of diet and the metagenome bin (p value less than 1%), but not a significant response to difference in mouse strains (p value more than 10%) was considered to have an association between diet interventions and the microbial sub-community.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical significance was verified through "Two-tailed paired t-test" using GraphPad Prism software and is reported in figure legends including exact value of n, what n represents, as well as definition of center, and dispersion and precision measures. In the figures p values are indicated as explained in the figure legends. All other computational quantitative analyses were performed with the open source software tools referenced in the STAR Methods along with the described procedures.

DATA AND CODE AVAILABILITY

The new raw sequencing data are available in Bioprojects PRJEB32890 (this study). The datasets (e.g., "All-in-one" assembly) generated during and/or analyzed during the current study are available at https://zenodo.org/record/3631711. Additional previously published datasets are listed in Table S1.

The code generated in this study is available in the GitHub repository, https://github.com/strowig-lab/iMGMC-1.