Comprehensive mouse gut metagenome catalogue reveals major difference to the human counterpart

Silas Kieser[1,2,3], Evgeny M. Zdobnov[3,4,5,*] and Mirko Trajkovski[1,2,5,*]

[1]Department of Cell Physiology and Metabolism, Faculty of Medicine, Centre Medical Universitaire, Geneva, Switzerland.

[2]Diabetes Center, Faculty of Medicine, Centre Medical Universitaire, Geneva, Switzerland.

[3]Swiss Institute of Bioinformatics, Geneva, Switzerland.

[4]Department of Genetic Medicine and Development, University of Geneva, Geneva, Switzerland.

[5]Institute of Genetics and Genomics in Geneva (iGE3), University of Geneva, Geneva, Switzerland.

*e-mail: Evgeny.Zdobnov@unige.ch , Mirko.Trajkovski@unige.ch

Keywords: Mouse, Human, Microbiome, Metagenome, Genome, Gut, Gene catalogue, Species, Sub-species, Plasmids, Viruses

Abstract

Mouse is the most used model for studying the impact of microbiota on its host, but the repertoire of species and subspecies from the mouse gut microbiome remains largely unknown. Here, we constructed a Comprehensive Mouse Gut Metagenome (CMGM) catalogue by assembling all currently available mouse gut metagenomes. We recovered 33'109 metagenome-assembled genomes (MAGs) from bacteria, 3470 plasmids, and over 120'000 viral contigs, together encoding 78 million proteins. We integrated all MAGs into 1449 species, of which 71.7% are newly identified, and 4007 subspecies. Rarefaction analysis indicates a comprehensive sampling of species and subspecies. 300 species represent newly identified genera, and we discovered 8 new families. CMGM enables an unprecedented coverage of mouse faecal and cecum metagenomes reaching 94%. Comparing CMGM to the human gut microbiota shows an overlap of only 18% at species, and 11% at the gene level, demonstrating that human and mouse gut microbiota are largely distinct.

Introduction

Mouse is the most used model for studying the microbiota importance due to several factors: availability of samples from different parts of the gastrointestinal tract, treatment options, controlled housing environment and diet, defined genetic background, and ethical considerations. However, the mouse gut microbiota has been poorly characterized, and only a fraction of the diversity observed by 16S rDNA sequencing is represented by genomes in public databases[1]. The majority of the studies on the mouse microbiome are performed by sequencing variable regions of the 16S, sometimes mislabelled as metagenomics. While this technique has allowed a general overview into the microbiota and information down to the genus level, it is not suited for identifying species for most of the organisms[2]. Different species from the same genus and even subspecies from the same species can exert contrasting functions[3], stressing the importance of annotating the gene content at a low taxonomic level.

Shotgun metagenomics allows studying the full microbiota diversity of an environment, including uncultured microorganisms, viruses, and plasmids. However, its interpretation is limited by the availability of reference genomes. Previous efforts led to the creation of a gene catalogue of the mouse metagenome (MGC v1)[4], by sequencing faecal samples from mice with different genotypes and housed in different conditions. This catalogue enables the functional annotation of genes and allows a 50% mapping rate of faecal sequences. However, the mapping rate of sequences from cecum samples is only 37%, and the catalogue does not contain genomic references. Recently developed algorithms enable assembly of genomes from metagenomes, leading to a recovery of new species from the human gut and other environments[5–9]. The integrated mouse gut metagenomic catalogue (iMGMC)[10] increased the fraction of reads mapped to genes compared to the MGC v1, however, mapping to the recovered metagenome-assembled genomes (MAGs) remains at about 40%[10]. Accordingly, many mouse genomes remain unclassified with the current state-of-the-art, and none of the approaches so far provide information of the microbiota on a subspecies level.

Here we report a Comprehensive Mouse Gut Metagenome (CMGM) collection that contains genomes generated by assembling gut microbiomes sequenced by us and all publicly available mouse metagenomes. This resource improves the mapping rate of genomic reads from mouse faecal and cecum metagenomes to over 94% and provides full classification on the level of subspecies, viruses, and plasmids. This nearly complete catalogue of the mouse gut bacterial species allows comparison between the newly assembled mouse gut microbiomes and the human counterpart, highlighting major differences between human and mouse in both species' composition and their abundance.

Results

Assembly of high-quality genomes from mouse gut metagenomes

We selected all metagenomic datasets associated with the mouse intestinal tract that are sequenced with a paired-end layout from the NCBI sequence read archive (accessed December 2019). To these, we added samples generated by our lab resulting in 1464 datasets (Extended Data Table 1). Each sample was processed using metagenome-atlas[11], which handles pre-processing, assembly, and binning of the metagenome datasets. The resulting MAGs were filtered based on fragmentation (N50>5000) and a quality score calculated from the output of checkM[12] as 'completeness minus 5 times contamination'. Bins with a quality score of <50 were excluded, resulting in 33'109 MAGs from which 11'373 (34%) had high quality (Quality score >90, Fig. 1b, Extended Data Fig. 1a). We included 776 complete mouse-associated bacterial genomes retrieved from RefSeq belonging to 331 species (Extended Data Table 2), which also includes genomes from mouse specific culture collections: Oligo-mouse-microbiota[13] (12 genomes), and Mouse Gut Microbial Biobank (mGMB, 41 genomes)[14]. As the genomes of the mouse Intestinal Bacterial Collection (miBC, 53 genomes)[1] were not available, we assembled them from the raw reads. Surprisingly, some reference genomes had contamination values of 100%, suggesting that the sequenced genomes consist of multiple strains. In total, 13 reference genomes did not pass the quality filtering,

and we included 816 reference genomes in the CMGM collection, resulting in a total of 33'925 genomes.

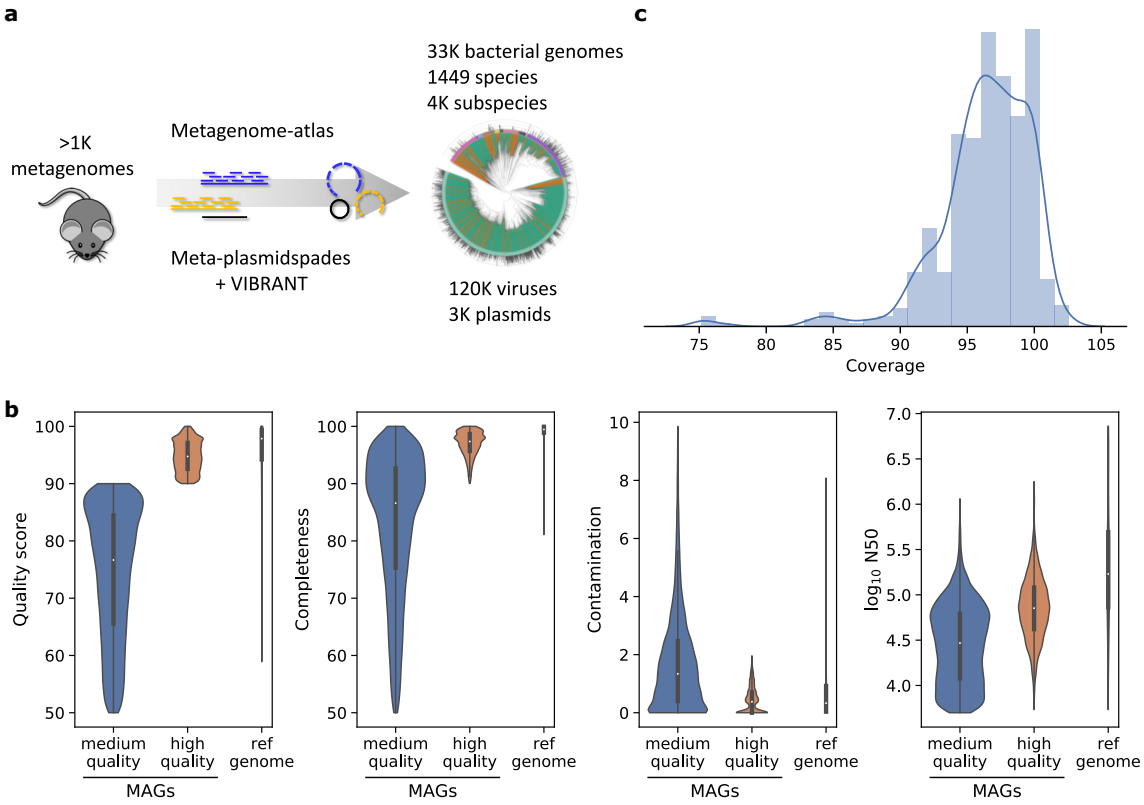


Fig. 1| Many metagenome-assembled genomes have comparable quality to reference genomes
a, Scheme of the workflow. **b,** Violin plots showing the quality score, completeness, contamination estimated using checkM and the $\log_{10}$ N50 from the assembly for the reference genomes and MAGs present in CMGM. **c,** Coverage of reference genomes by MAGs (n=494).

While MAGs were more fragmented and had a lower median quality score than the reference genomes, the quality score and N50 of the high-quality MAGs were comparable to the values for the references (Fig. 1b). For 60% of the reference genomes, we recovered MAGs that align to them with high coverage and identity (ANI >95%, IQR 94-99%, Fig. 1c). This validates our metagenome assembly approach to recover "reference quality" genomes *de novo*. Some of the remaining differences might be attributed to strain variation, as the coverage is higher for more similar genomes (Extended Data Fig. 1b).

Since we assembled genomes from individual samples, the same strain could have been recovered multiple times, especially if different gut locations of the same

mouse were sampled. To remove this potential redundancy, we clustered the genomes based on the average nucleotide identity (ANI) calculated using bindash[15]. 95% ANI was used as threshold to delineate genomes from the same species[16,17]. For each species cluster, the genome with the highest quality and

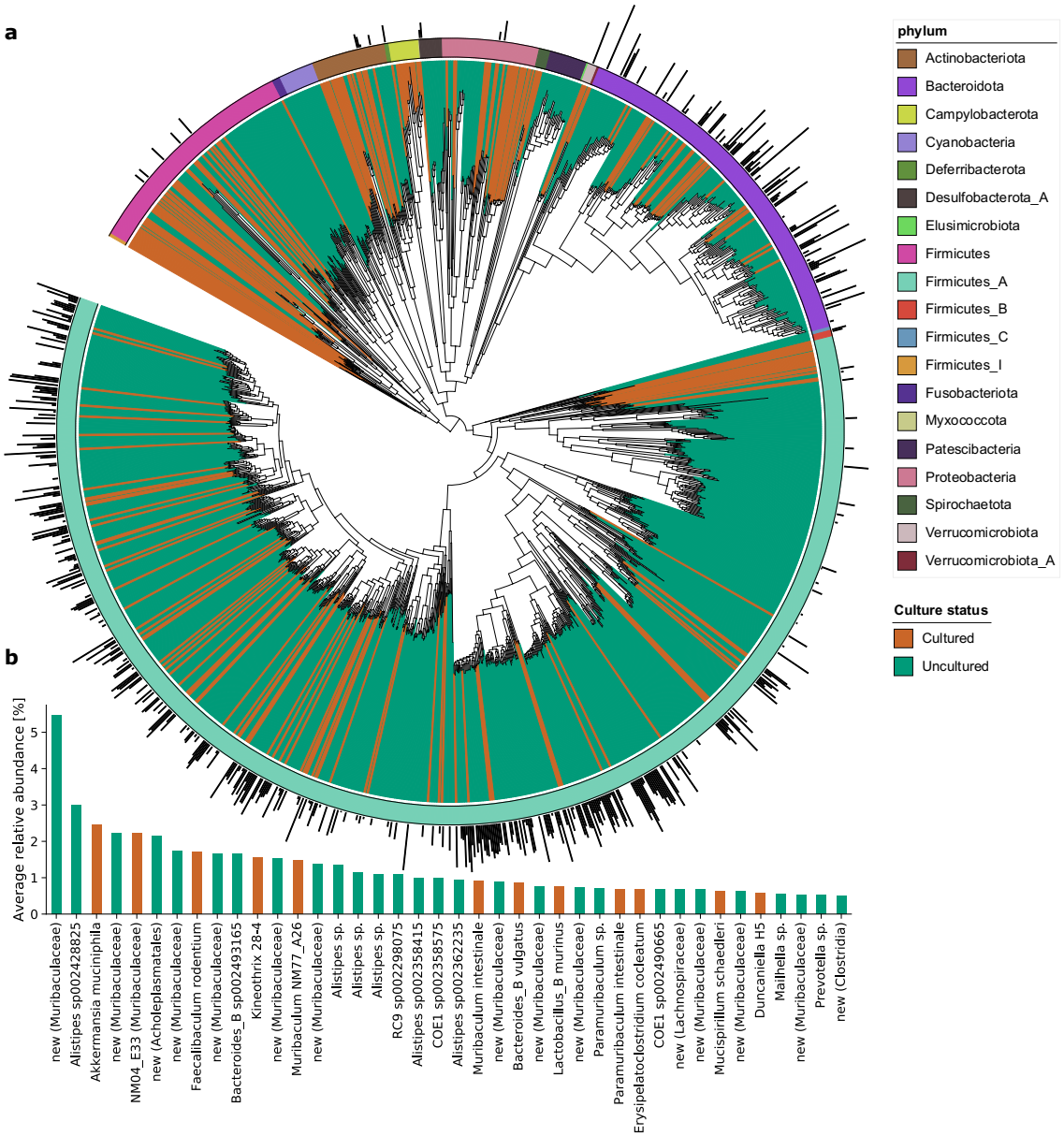


Fig. 2| The mouse gut microbiome is predominantly uncultured. **a,** Maximum-likelihood phylogenetic tree of the 1449 bacterial species detected in the mouse gut. Clades are colored by culture status. The color ring indicates the phylum attribution and the bar in the outer ring indicates the median abundance in mouse gut microbiome (centered log ratio). Values < 0 are omitted. **b**, Bar plot of the 40 most abundant species in the mouse gut microbiome colored by cultured status.

lowest fragmentation was selected as representative, but reference genomes were preferred over MAGs. The species representatives were annotated using the genomic taxonomy database (GTDB[18,19]). Species that contain a reference genome of an isolate were counted as cultured, even when they might not be available from official culture collections. Similarly, species named after an isolated strain in GTDB were annotated as cultured.

Majority of the species from the mouse gut are uncultured

The CMGM genome collection represents 1449 species (Fig. 2a), of which 71.7% have not been previously identified. 76.4% of the CMGM species are uncultured, with 17.5% having a mouse-specific cultured strain. 300 represent the first species for their genus, and we discovered 8 new families. 218 species do not have a cultured species at the order level. Since many of the most abundant species are uncultured (Fig. 2b), the sum of cultured species accounts on average for less than 20% of the mouse metagenome. 4607 genomes contain one or multiple full-length 16S gene sequences, which allowed us to link 51% of the 1449 species in the CMGM catalogue to a 16S sequence. This represents over 50% advance over the latest reports linking 484 genomes from the mouse gut to 16S sequences[10].

We used StrainDrep[20] to further split the 33'925 genomes into 4007 subspecies. We observed clear clusters, for which the genome with the highest quality was chosen as representative (Extended Data Fig. 2). Subspecies contain a specific subset of genes (Extended Data Fig. 2). In datasets with faeces and cecum samples of the same mouse, genomes from the same species belong also to the same subspecies/strain. This indicates that our approach consistently recovers dominant strains/subspecies.

To assemble viruses and plasmids from metagenomic and genomic datasets, we used a pipeline based on metaplasmid-spades[21] and VIBRANT[22]. The assembled plasmids and viral contigs were dereplicated into non-redundant catalogues. We recovered 120'983 viral contigs, 1128 of which are complete circular and 3106 are of high or medium quality as estimated by VIBRANT. 88% of the contigs were classified as lytic, 8.6% as lysogenic, and 3.7% are integrated prophages. The

Plasmid catalogue consists of 3470 circular plasmids including 48 plasmids which were recovered from the 53 assembled genomes of the miBC[1].

Evaluation of the CMGM catalogues

Rarefaction analysis shows that the number of species reached a saturation point at 1036 when considering species with at least two conspecific genomes (Fig. 3a). This indicates that the CMGM catalogue contains all species commonly living in the mouse gut. However, more rare species remain to be discovered, as the rarefaction curves with singletons (species which were recovered only in one sample) did not converge. Strikingly, rarefaction analysis reached a saturation

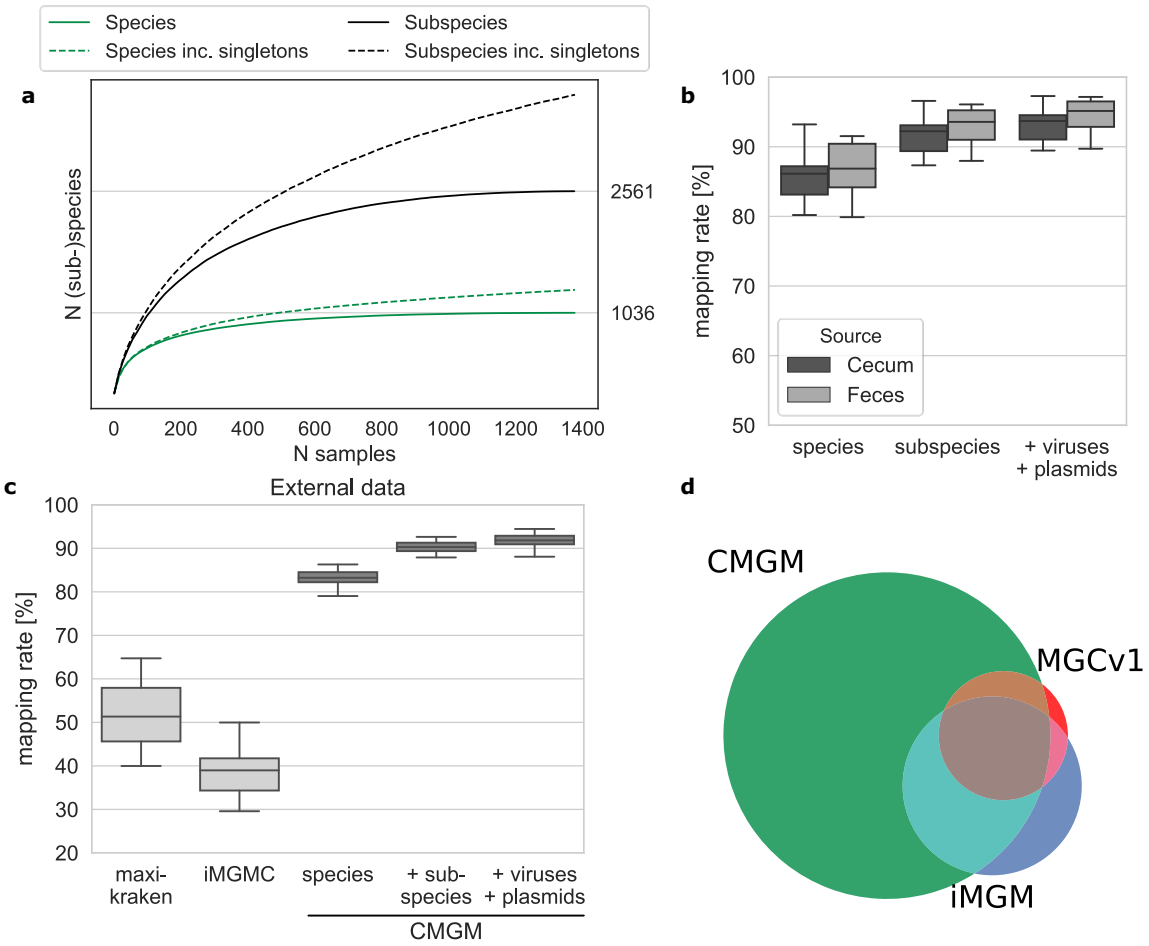


Fig. 3| CMGM catalog provides close-to-complete coverage of the mouse microbiome

a, Rarefaction curves of species and subspecies. **b,c**, Comparison of mapping rates on assembled genomes, plasmids and viruses of the mouse gut metagenome on internal (b) and external data (c). **d,** Overlap of different gene catalogs from the mouse metagenome.
CMGM: this study, iMGM: Lesker et al. 2020, MGCv1: Xiao et al 2015

point at 2561 subspecies (Fig. 3a), indicating that the CMGM catalogue also contains all subspecies commonly found in the mouse gut.

CMGM achieves a mapping rate of the mouse metagenome of 95.1% for faecal and 93.7% for cecum samples. Specifically, the mouse microbiome diversity captured by the CMGM species covers 86% of reads from both faecal and cecum samples (Fig. 3b). Microbiome profiling using the subspecies representatives increased the mapping rate by 6% compared to species alone. Viruses and plasmids added 1.5% of mapped reads (Fig. 3b).

To independently evaluate the mapping rate of the CMGM catalogue, we used an external dataset of cecum samples, which was explicitly left out from this catalogue, and not contained in previous ones. The CMGM species covered 83% of the metagenomic reads, twice as many compared to the previous genome collection from the mouse gut[9] (Fig. 3c). The addition of subspecies and extra-chromosomal elements increased the mapping rate by 7% and 1.5%, respectively. In total, 91.8% of reads from an external sample were mapped by CMGM, which is over 40% increase compared to the maxi-kraken database that contains all RefSeq genomes from bacteria, archaea, protist, fungi, and viruses.

We predicted over 260 million genes from the assembled contigs and clustered them to generate the CMGM protein catalogue. This non-redundant protein catalogue contains 78 million proteins, over 10 times more than the previous mouse gene collections[4,10] (Fig. 3d). 83.1% of our gene catalogue could be annotated, and 49.7% of genes are linked to 8077 Kegg annotations. To facilitate further comparisons, we produced the CMGM gene catalogue clustered at 90 and 50% amino acid identity.

To test the applicability of the CMGM, and propose how this catalogue allows discovering compelling biological insights, we compared mice from three different providers fed a high fat-fed diet (HFD) for 7-8 weeks to control mice on chow diet[4] (N=67). We used aldex2 with a linear model to account for the different mouse providers. Many *Becteroidota* species were significantly decreased and species from the phylum *Firmicutes* increased (Extended Data Fig. 3A). Interestingly, Shannon diversity did not decrease with HFD (Extended Data Fig.

3B). This example suggests that using the CMGM catalogue as reference for metagenomic studies enables discovering precise and comprehensive changes of species induced by a treatment or a disease. It also sets the ground for reanalysis of the existing datasets for uncovering species that are involved or altered by the condition of interest.

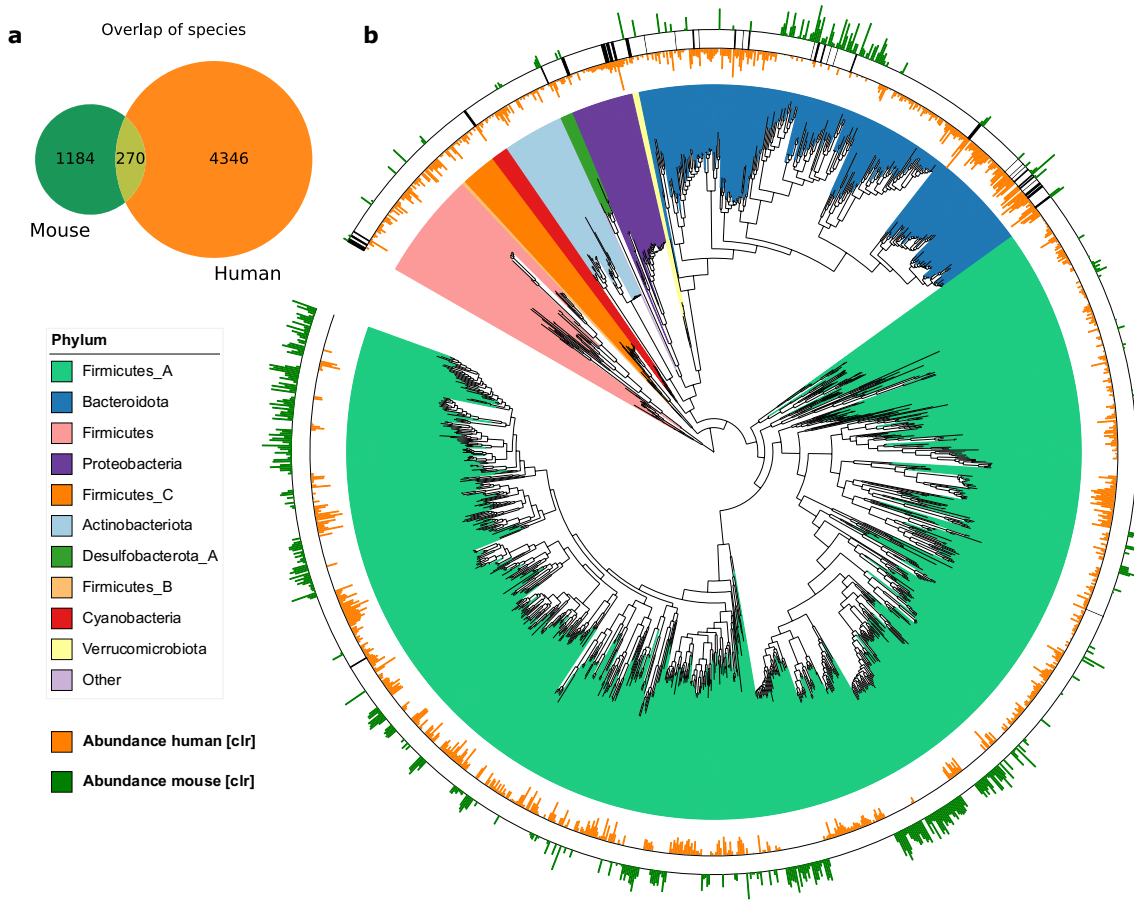


Fig. 4| Human and mouse guts harbor distinct bacterial species.
a, Venn diagrams of the overlap between mouse and human gut microbiota at the species level. **b,** Phylogenetic tree of abundant species (centered log ratio, clr >0) in either human or mouse microbiome. Clades are colored by phylum attribution. The black bars in the middle ring indicate shared species between human and mouse (ANI > 95%). The bar plot in the inner ring indicates the median abundance in human microbiome (inverted axis), and the bar plot in the outermost ring the abundance in the mouse microbiome (values clr < 0 are omitted).

Comparison between human and mouse gut microbiomes

Studying mice microbiota and its impact on the host as a proxy for humans implies their similarities. However, 16S rDNA profiling and gene catalogues don't allow a

comprehensive analysis of the analogy between human and mouse microbiota down to species level. Also, much fewer species from the mouse gut are sequenced than from the human gut[23]. The CMGM catalogue, together with the recent creation of genome collections from the human gut[24], renders this comparison possible. Here, we compared the species representatives from CMGM to the ones from the unified human gut genomes[24] and applied the same criteria as for clustering (ANI > 95%). From the 1449 CMGM species, 18.29% (270) were identified in the human gut microbiota (Fig. 4a, Extended Data Fig. 5). The shared species account on average for 15.6% of the mouse and 29.6% of the human gut microbiome.

When low abundant species (centred log-ratio (CLR) <0) are left out from this comparison only 45 species were shared (Fig. 4b, Extended Data Fig. 4, Extended Data Table 3) that corresponds to an overlap of 10% of abundant species. Curiously, 27 of the 45 shared species belong to the phylum *Bacteroidota*, whereas only two of them are from the phylum *Firmicutes_A*. Only 12 species are abundant in both microbiomes (CLR >0), 10 of which belong to the phylum *Bacteroidota*. 10 out of the 12 most abundant species have cultured representatives. These data reveal major differences between human and mouse microbiota at the species level.

To investigate the functional repertoire of the human and mouse microbiome, we compared the CMGM protein catalogue to the unified human gut protein (UHGP)[24] at 90% amino acid identity. Similar as in MGCv1[4], only a fraction of genes could be annotated with Kegg orthologs (50% in CMGM and 13% in UHGP). Comparison based on this set of 8077 functional annotations indicates an overlap of 99% between the human and mouse microbiome. However, if all 78mio genes are taken into account, the overlap drops to 10.8%, corroborating the limited overlap of species between these two microbiomes.

We compared CMGM and UHGG[24] at higher taxonomic resolutions based on the GTDB taxonomy. In GTDB, if a taxonomic group is polyphyletic, it is split into several sister clades based on relative evolutionary divergence[18], which permits more robust comparisons and correlations. The new clades are usually named

with an alphabetical suffix, for example, the phylum Firmicutes is split into 12 sister phyla. More than half of the species in both microbiomes belong to the phyla Firmicutes_A. *Firmicutes_A* and *Bacteroidota* (*Bacteroidetes*) are the most abundant phyla in both human and mouse microbiomes (Fig. 5a). Overall, 17 phyla have representatives in both human and mouse microbiome. 5 phyla including *Synergistota* and *Eremiobacterota* are only found in human and not in mice microbiota. In contrast, the phylum *Deferribacterota* and the two species *Chlamydia muridarum* and *Chlamydophila psittaci* which represent an own phylum, are specific to mice. No archaea were reconstructed from the mouse gut metagenome, whereas 0.4 % of the genomes in the UHGG belong to this domain.

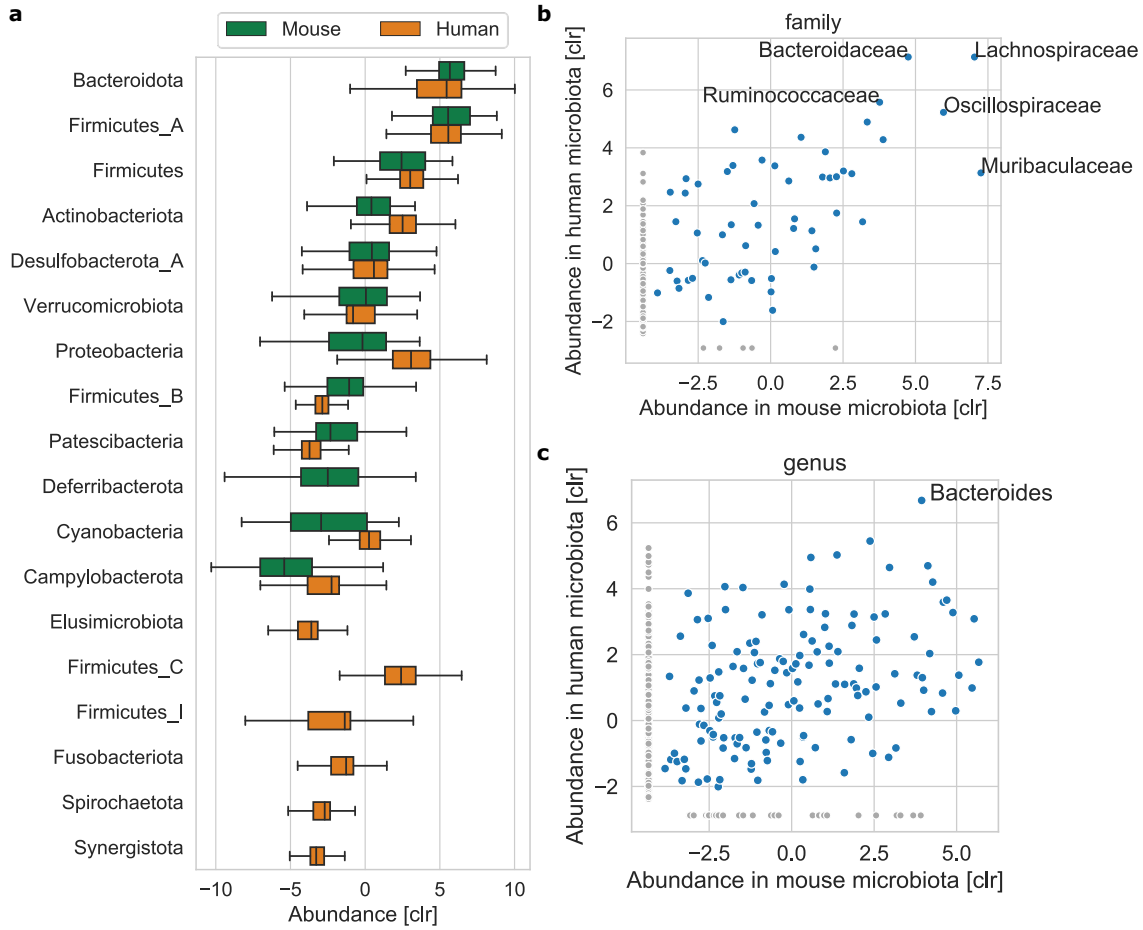


Fig. 5| Human and mouse microbiomes are similar at higher taxonomic level.

a, Phylum abundances in human and mice microbiota. **b,c,** Correlation of average abundance of families (b) and genera (c) in human and mice microbiotas. CLR = centred log ratio.

At the family level, humans and mice share 91 of the 105 defined taxa, whose

average abundance in human and mouse microbiota are correlated (r=0.63, Fig. 5b). The families *Lachnospiraceae*, *Oscillospiraceae*, and *Ruminococcaceae*, have high abundance in both human and mice. The family *Muribaculaceae* is 60 times more abundant in mice than in humans, whereas *Bacteroidaceae* is 10 times less. While at the genus level, 227 of 273 of taxa are shared (83% overlap), the abundance of the genera showed only a very limited correlation (r=0.37, Fig. 5c), in line with the results based on 16S rDNA seqiencing[25]. These data show that even when at higher taxonomic levels (phylum to family) the mouse and human microbiome show similarities, there are major differences of the genera abundances and a very limited overlap at the species level. This is further supported by comparing the influence of age and obesity on the microbiome of mouse and human, which shows that host-adaption is the strongest difference (Extended Data Figure 5).

Discussion

We generated a comprehensive catalogue of the mouse gut metagenome: 33'925 genomes, 78 million protein sequences, over 120'000 viral contigs, and 3470 plasmids. This resource now enables mapping of over 95% of faecal and 93% of cecum samples. From the 1449 genomic defined species, 71% are newly identified and 51% could be linked to a full-length 16S sequence. Integrated into databases of 16S genes, these sequences can help to link the functional repertoire of the genome with the 16S gene, therefore leveraging the use of amplicon sequencing. Three-quarters of the species are uncultured, and some do not have a representative at the order level. Hence, the CMGM catalogue is a valuable basis for targeted culturing of these missing strains.

The CMGM is the first collection containing plasmids and viruses. Expectedly most of the viral contigs represent fragments of viruses as they are recovered from unfiltered metagenomic reads. Higher diversity might be recovered in filtered virome samples from the mouse gut. The CMGM catalogue is also the first that contains in-depth information down to the subspecies level. Although it is possible to use single nucleotide polymorphisms to detect genotype-diversity in metagenome samples, such approaches make it hard to link the genotype to the

functional repertoire. On the other hand, the CMGM subspecies are naturally linked to a specific subset of genes.

Saturation in the rarefaction analysis shows that the CMGM catalogue contains all main species and subspecies commonly living in the mouse gut. Nevertheless, we cannot exclude that new samples may contain diversity that is not part of the CMGM, for example, species present in single samples or wild mice. However, CMGM is built by assembling all publicly available data from the mouse strains that are most experimentally used, thus comprehensively representing the microbiome of laboratory mice.

Comparing the mouse microbiota to the human counterpart reveals overlap and correlation of the average abundance from phylum down to family level. As suggested by amplicon sequencing[25], the genera are qualitatively the same but quantitatively different. We observed no correlation between their average abundances in human and mouse microbiota, despite identifying 83% of shared genera. Whereas a comprehensive and precise comparison at species level between the two microbiomes was not previously feasible[2,26], the comparison of CMGM with the UHGG collection reveals an overlap of only 10% of the abundant species. These findings effectively challenge our view on the analogy between human and mouse microbiota and may impact the experimental designs and approaches for studying the gut microbiota. Different ways can be envisaged to overcome these challenges. For example, advanced transplanting human gut microbiota into germ-free mice to create 'humanized' mouse models that would be kept in gnotobiotic conditions, or complementing the work by exploring additional animal models [27]. To leverage data produced using conventional mice, it will be worth finding functional homologues between the species adapted to mouse and human microbiota e.g. by identifying 'guilds'[28], groups of species that use the same type of resources in a similar way. The functionally annotated species in the CMGM collection lays the basis for such work. The knowledge of the genomes and a nearly complete mapping rate is a basis for precise analysis on higher taxonomic levels and function. Also, the studies included in the CMGM might contain biological insights that were not accessible previously, for example

because they relate to previously unknown species. The integration with human genome catalogues allows easy comparison at higher taxonomic level.

In summary, CMGM increases our knowledge of the mouse microbiota gene repertoire by ten-fold and is the first to identify the subspecies present in the mouse gut microbiota, which together with the majority of newly identified species allows comprehensive analysis of the mouse gut microbiome at an unprecedented depth. This work uncovers major differences between the mouse and human gut microbiome identities.

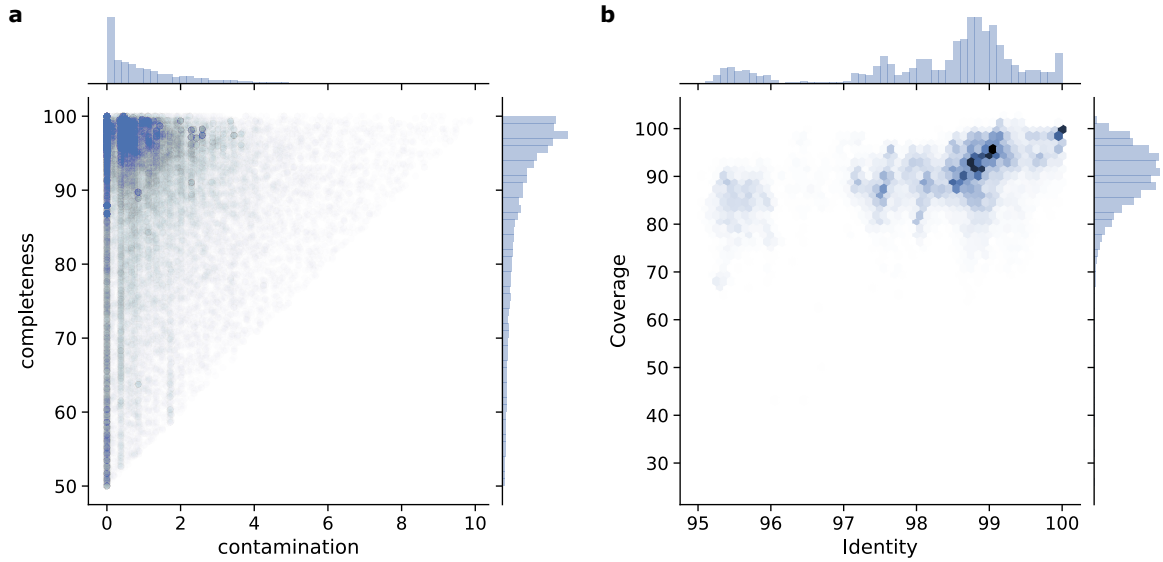
Author contributions

S.K. wrote the code, analysed, and interpreted the data, and generated the figures. E.Z. and M.T. guided the project, interpreted the data and supervised the work. All authors conceptualized the study and wrote the paper.

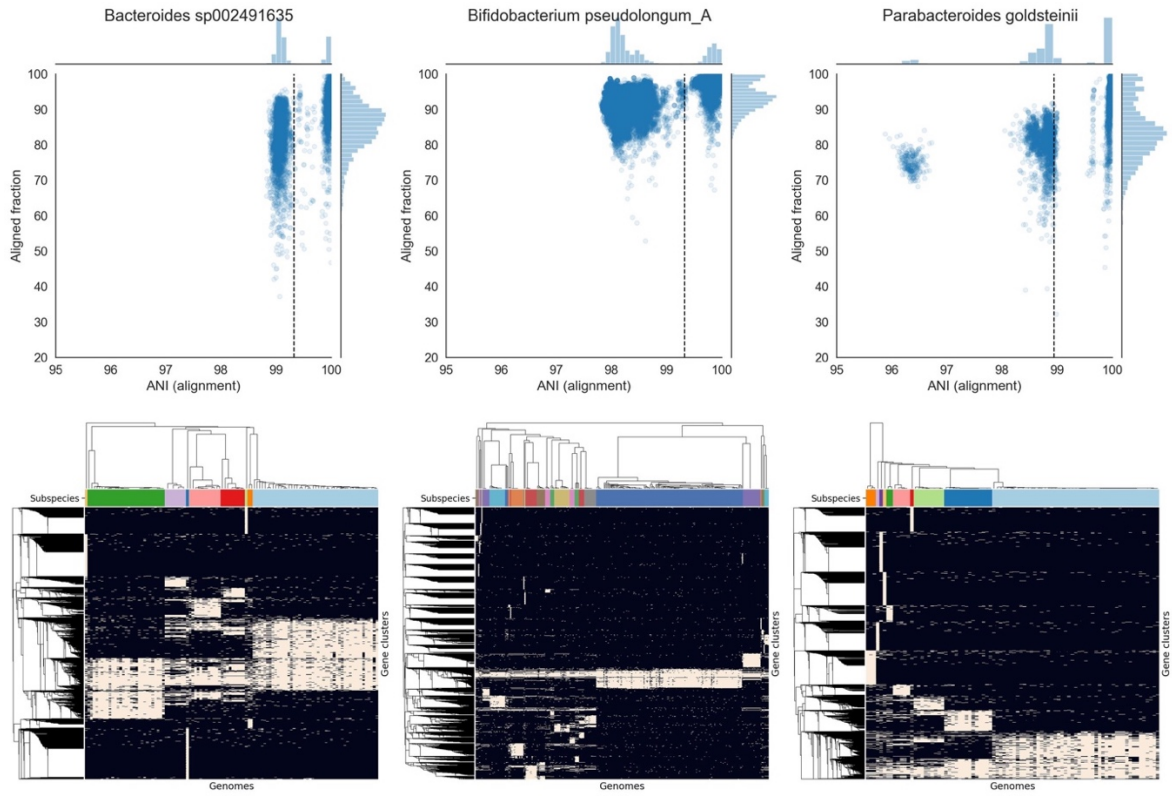
Acknowledgements

We are grateful to Christopher Rands for critical reading of the manuscript, and to all members from our labs for discussions. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ERC Consolidator Grant agreement No. 815962, Healthybiota) to M.T.

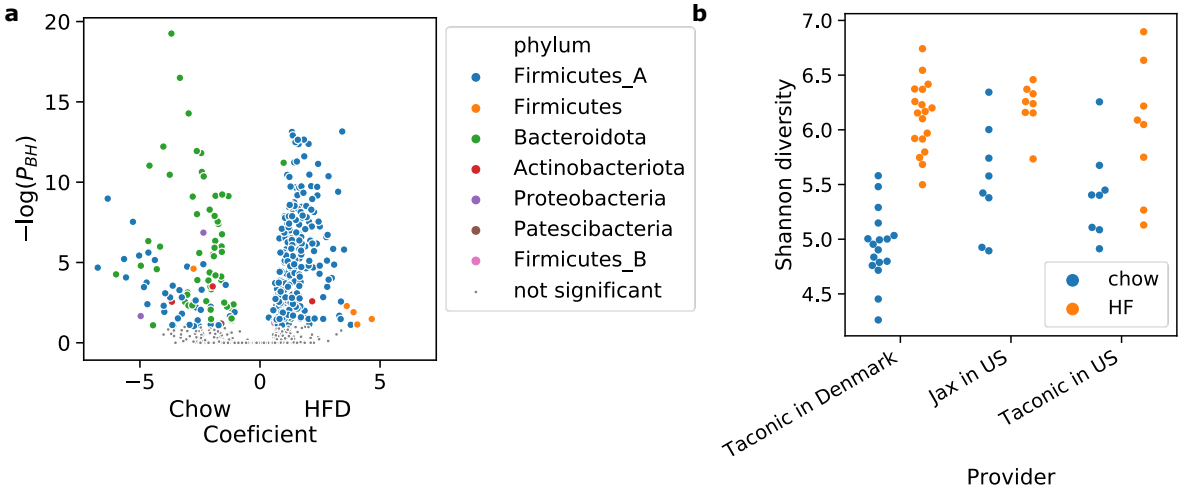
Extended Data



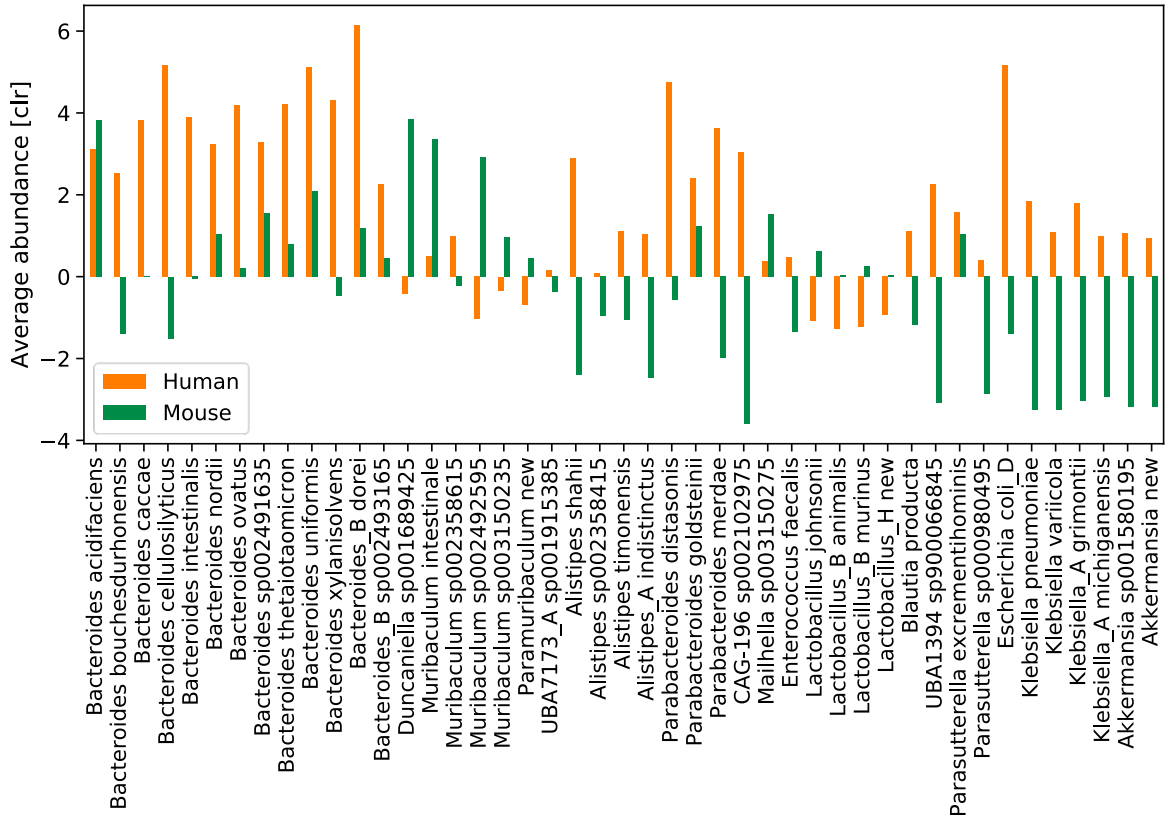
Extended Data Fig. 1| Quality estimates of MAGs at a genome level. **a,** Distribution of the MAGs included in the CMGM collection according to their completeness and contamination estimated with checkM. MAGs with ‘completeness -5×contamination`< 50 were excluded. **b,** Density plot of the coverage vs. identity of the MAGs alignments to 494 reference genomes.



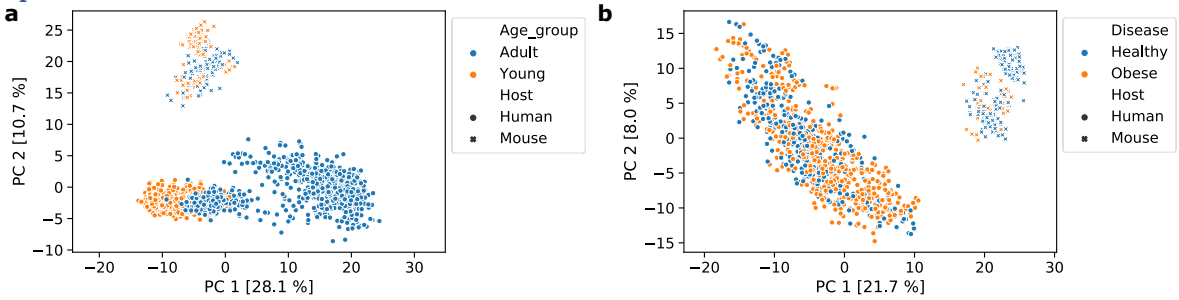
Extended Data Fig. 2| Subspecies clustering.
Example of subspecies clustering of three species with > 50 genomes. Top panels: Distribution of intra-species pairwise alignments with respect to average nucleotide identity (ANI) and aligned fraction. The dashed lines indicate the automatically selected threshold for sub-species clustering. Lower panels: presence/absence matrix of subspecies-specific gene clusters (GC50). Genomes are arranged by the dendrograms build from the pairwise ANI values. Gene clusters are ordered by dendrograms based on Jaccard distance.



Extended Data Fig. 3| Bacterial changes induced by a high fat diet.
a, Volcanoplot of coefficients associated with high-fat diet of the generalized linear model and the associated p-values corrected using Benjamini-Hochberg-correction. **b**, Shannon diversity of the same samples stratified by Provider.



Extended Data Fig. 4| Species present in both human and mouse microbiomes.
Bar plot of the 45 species shared between human and mouse microbiome with a minimal abundance of centred log-ratio (CLR) >0 in either of the two microbiomes. The species are ordered according to their taxonomy. Negative CLR values indicate abundance that is lower than the geometrical mean of species.



Extended Data Fig. 5| PCA of human and mouse metagenome samples.
Principal component analysis based on the robust Atchison distance based on the family abundance of mouse and human samples. **a,** Human samples come from healthy European adults or infants. Young` for mice signifies less than 12 weeks of age. **b,** Human samples come from adults that are either healthy or obese defined as Body Mass Index (BMI) > 30 kg m^{-2}. In mice, obesity is induced by high-fat diet.

Extended Data Table 1 | Metagenome samples used to construct CMGM.
The table shows the metagenome samples used for the generation of CMGM. The CMGM_Id corresponds to the SRA read id, except for the samples sequenced by our lab. The table contains information retrieved from NCBI that was available for most of the samples: Name, description, Link to bioproject, collection data, country, and submission centre. The column 'Source' specifies the organ from which the sample was taken. If the information was available in any of the metadata. Samples under the bioproject accession PRJNA646351 were sequenced for this study.

Link:
https://ezmeta.unige.ch/CMGM/v1/input_data/curated_metadata_SRAruns.xlsx

Extended Data Table 2 | Reference genomes associated with the mouse gut.
The table shows the assembly information of reference genomes associated with the mouse gut. These genomes were filtered for completeness and contamination before integration into CMGM. The columns `Isolated` and `Cultured` label if the genome is Isolated and cultured. The `coaction` describes if the genome is part of a mouse-specific culture collection. The genomes of the miBC collection are assembled for this study.

Link:

https://ezmeta.unige.ch/CMGM/v1/input_data/all_mouse_associated_ReferenceGenomes.tsv

Extended Data Table 3 | Core shared species between the human and the mouse microbiota.
The table shows detailed information about the 45 species shared between human and mouse microbiota with a minimal abundance of cantered log-ratio (CLR) >0 in either of the microbiomes, related to Extended Data Figure 4. The table contains the taxonomy and the cultivation status as well as the abundance in

relative and CLR. The number of subspecies in CMGM are indicated in the last column.

Link:

https://ezmeta.unige.ch/CMGM/v1/Comparison_with_Human_Microbiome/Shared_core_Species.xlsx

Online Methods

Sequencing of metagenomic data of mice

The sample collection and metagenomic sequencing were approved by the Swiss federal and Geneva cantonal authorities for animal experimentation (Office Vétérinaire Fédéral and Commission Cantonale pour les Expériences sur les animaux de Genève). Animals were on C57Bl/6J background, commercially available through Charles River, France. The mice experiment is detailed in[29]. Paired-end metagenomic libraries were prepared from 100 ngDNA using TruSeq Nano DNA Library Prep Kit (Illumina) and size selected at about 350 bp. The pooled indexed library was sequenced in a HiSeq4000 instrument at the iGE3 facility (University of Geneva).

Collection of public metagenome and genomic data

We searched the sequence read archive (SRA) of the National Center for Biotechnology Information (NCBI) for all publicly available paired-end metagenome runs from the mouse microbiome. We specifically excluded samples from human origin and amplicon sequences and different body parts than the gut. We extracted 1414 metagenome runs belonging to 43 projects. Metadata was retrieved using BioServices[30] and curated (Extended Data Table 1). We retrieved 776 assemblies from RefSeq who were linked to a biosample collected from mouse (Extended Data Table 2). We excluded reference genomes collected from other body parts than the gut or faeces.

Metagenome assembly and binning

Metagenomics and genomic reads were processed using the metagenome-atlas v2.3[11] pipeline with the command 'atlas run genomes'. The configuration file is available at the 'Code availability' section. In short, using tools from the BBmap suite v37.78[31], reads were quality trimmed, and contaminations from the mouse genome were filtered out. Reads were error corrected and merged before they were assembled with metaSpades v3.13[32]. Contigs were binned using metabat2 v 2.14[33] and maxbin2 v2.2[34], and their predictions were combined using DAS Tool v 1.1.1[35]. For the assembly of the 53 genomes of the mouse intestinal bacterial collection, we used the assembly workflow of metagenome-atlas and set 'spades_preset: normal' which uses the basic spades as assembler. The quality of the genomes was estimated using checkM v1.1[12].

Genome filtering and species clustering

We used StrainDrep v0.1[20] to filter and cluster genomes into species. For the configuration file see the 'Code availability' section. In short, genomes with an estimated quality of 'completeness-5*contamination' <50 or N50<=5000 were excluded. All Pair-wise ANI above 0.8 were calculated using bindash[15] and missing values were filled with the minimum value observed. Hierarchical clustering was performed with average linkage and a threshold of 95% using scipy[36]. For each species cluster, the genome with the highest score based on the following formula was selected as the representative.

Score= Completeness - 5×Contamination + 0.5×log(N50) + 100×isIsolate

Where Completeness and Contamination are estimated using checkM v1.1[12], N50 is the N50 score of the assembly contiguity, and 'isIsolate' is 1 for isolates and 0 for MAGs, to ensure that isolated genomes are preferred over MAGs even if they have lower quality.

Phylogenetic and taxonomic analysis

The species representatives were annotated using the genomic taxonomy database toolkit (GTDB-tk v1.2[18]). A maximum-likelihood tree based on the 120 bacterial marker genes from GTDB was built using fasttree v2.1 [37] and rooted at the midpoint. The phylogenetic trees are visualized with iTOL v5[38] and the

annotations were prepared using table2itol (https://github.com/mgoeker/table2itol). The Pearson correlation between the abundance of taxonomic groups in the human and mouse microbiota was performed with scipy v1.4.1[36].

Inferring cultured status

Species that contain a reference genome included in the CMGM catalogue are counted as cultured from a mouse origin. If GTDB-tk[18] was able to annotate the species to a reference with ANI >95%, we counted the species as cultured from a non-murine source. In both cases, if the reference genome was excluded from RefSeq (i.e. metagenome-assembled genomes) or labelled as uncultured we counted the species as isolated but not cultured.

Quantification

We used bbsplit[31] with the parameters 'ambiguous2=best minid=0.9' to map metagenomic reads to the references with 90 identity. The mapping rates were calculated as a fraction of the reads mapped to the reads used from the bbsplit log file. For most quantification, the mapped reads per genome were summed and the centred log ratio (CLR) was calculated using the sci-kit bio package (http://scikit-bio.org/) after imputing zeros using a multiplicative replacement approach. When relative abundance was used as a measure, we estimated the genome coverage as the median of blocks of 1000bp. For viruses and plasmids, the coverage over the whole contig was used. For the quantification, we used 31 cecum and 28 faecal samples from mice from our lab[29,39] as well as 184 faecal samples from the MGC v1[4]. For comparison, we quantified reads using kraken2[40] with the maxikraken2 database (lomanlab.github.io/mockcommunity/mc_databases.html, March 2019). The abundance estimation of species in the human microbiome is based on the quantification in 13132 samples[8]. We used aldex2 v1.18 for differential abundance analysis using the default parameters. Shannon diversity was calculated using the package scikit-bio (http://scikit-bio.org/) based on the relative abundance.

Gene prediction and clustering

Genes were predicted using prodigal v2.6[41] on all the assembled contigs. For the metagenome assemblies, we used the anonymous mode and for genome assemblies the parameters '–p normal --closed -m'. For the reference genomes, we downloaded the gene predictions from RefSeq. All predicted gene products were clustered using linclust[42] at 100% average amino acid identity (AAI), 0.8 coverage, and the parameters ' --kmer-per-seq 80 --cov-mode 1'. Genes were linked to the contigs and to the genomes they belong to and annotated using EggNOGmapper v2[43], which uses DIAMOND[44] to map genes to the EggNOG DB v5[45]. The catalogue was further subclustered using the same parameters as above but 90, and 50% AAI. We calculated the overlap between the 90% clustered catalogue with previous mouse gene catalogues, or the UHGP-90[24] using [42] at 90 AAI. Code availability: https://github.com/metagenome-atlas/genecatalog_atlas

Subspecies clustering

Subspecies were identified using StrainDrep v0.1[20]. In short, we calculated all intra-species pairwise genome alignments using minimap2v 2.17[46] and computed the average nucleotide identity (ANI). We used hierarchical clustering based on average linkage using scipy v1.4.1[36]. The optimal number of subspecies was automatically selected based on the maximal silhouette score. As the silhouette score can only be calculated for two or more groups, we classified species having over 95% of pair-wise comparisons with ANI > 99.5% as a single subspecies. Similar to the species clustering, for each subspecies cluster, the genome with the highest score was selected as representative genome.

The genes present in the genomes from a given species were mapped to gene clusters outlined above at 90% AAI. Gene clusters present in more than 80% of all the genomes of a subspecies were considered as this subspecies' core genes. Gene clusters that are part of the core genes in more than 80% of the subspecies are correspondingly considered as this species core genes and the genes present in a subspecies that are not part of the species core genes are subspecies specific.

Assembly of viruses and plasmids

We assembled circular contigs from our metagenome and genome datasets using (meta-)plasmid-spades v3.13[21]. The circular contigs from all samples were de-replicated using dedupe[31] and filtered for specificity to virus or plasmids using viralverify (github.com/ablab/viralVerify). In addition, we used VIBRANT v 1.2.0[22] to scan for viral fragments in our metagenome-assemblies. Viral fragments were dereplicated using bbsketch[31] based on Average amino acid identity >=99%. VIBRANT estimates the quality of vial contigs and classifies them as lytic, lysogenic. Prophages were also detected using VIBRANT.

Data Availability

The metagenomic samples sequenced for this study are available from the NCBI sequence read archive under the project id PRJNA646351. The assemblies generated in this study are deposited under study accession PRJNA646353. Reference genomes, MAGs, viruses, and plasmids used in this study together with their annotations are available at https://ezmeta.unige.ch/CMGM/v1. The configuration files and the metadata of the samples used for the construction of CMGM are available through the same link.

References

1. Lagkouvardos, I. *et al.* The Mouse Intestinal Bacterial Collection (miBC) provides host-specific insight into cultured diversity and functional potential of the gut microbiota. *Nat. Microbiol.* **1**, 16131 (2016).

2. Johnson, J. S. *et al.* Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* **10**, 1–11 (2019).

3. Costea, P. I. *et al.* Subspecies in the global human gut microbiome. *Mol. Syst. Biol.* **13**, 960 (2017).

4. Xiao, L. *et al.* A catalog of the mouse gut metagenome. *Nat. Biotechnol.* **33**, 1103–1108 (2015).

5. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).

6. Stewart, R. D. *et al.* Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat. Commun.* **9**, 870 (2018).

7. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649-662.e20

(2019).

8. Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).

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

10. Lesker, T. R. *et al.* An Integrated Metagenome Catalog Reveals New Insights into the Murine Gut Microbiome. *Cell Rep.* **30**, 2909-2922.e6 (2020).

11. Kieser, S., Brown, J., Zdobnov, E. M., Trajkovski, M. & McCue, L. A. ATLAS: a Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data. *BMC Bioinformatics* **21**, 257 (2020).

12. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–55 (2015).

13. Garzetti, D. *et al.* High-Quality Whole-Genome Sequences of the Oligo-Mouse-Microbiota Bacterial Community. *Genome Announc.* **5**, e00758-17 (2017).

14. Liu, C. *et al.* The Mouse Gut Microbial Biobank expands the coverage of cultured bacteria. *Nat. Commun.* **11**, (2020).

15. Zhao, X. BinDash, software for fast genome distance estimation on a typical personal laptop. *Bioinformatics* **35**, 671–673 (2019).

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

17. Olm, M. R. *et al.* Consistent metagenome-derived metrics verify and define bacterial species boundaries. *bioRxiv* 647511 (2019). doi:10.1101/647511

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

19. Parks, D. H. *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* 1–8 (2020). doi:10.1038/s41587-020-0501-8

20. Kieser, S. StrainDrep github.com/SilasK/StrainDrep.

21. Antipov, D., Raiko, M., Lapidus, A. & Pevzner, P. A. Plasmid detection and assembly in genomic and metagenomic datasets. *Genome Res.* gr.241299.118 (2019). doi:10.1101/gr.241299.118

22. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: Automated recovery, annotation and curation of microbial viruses, and evaluation of virome function from genomic sequences. *bioRxiv* 855387 (2019). doi:10.1101/855387

23. Hugenholtz, F. & de Vos, W. M. Mouse models for human intestinal microbiota research: a critical evaluation. *Cellular and Molecular Life Sciences* **75**, 149–160 (2018).

24. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* 762682 (2020). doi:10.1038/s41587-020-0603-3

25. Krych, L., Hansen, C. H. F., Hansen, A. K., van den Berg, F. W. J. & Nielsen, D. S. Quantitatively

Different, yet Qualitatively Alike: A Meta-Analysis of the Mouse Core Gut Microbiome with a View towards the Human Gut Microbiome. *PLoS One* **8**, e62578 (2013).

26. Hugenholtz, F. & de Vos, W. M. Mouse models for human intestinal microbiota research: a critical evaluation. *Cell. Mol. Life Sci.* **75**, 149–160 (2018).

27. Nguyen, T. L. A., Vieira-Silva, S., Liston, A. & Raes, J. How informative is the mouse for human gut microbiota research? *DMM Dis. Model. Mech.* **8**, 1–16 (2015).

28. Root, R. B. The Niche Exploitation Pattern of the Blue-Gray Gnatcatcher. *Ecol. Monogr.* **37**, 317–350 (1967).

29. Chevalier, C. *et al.* Gut Microbiota Orchestrates Energy Homeostasis during Cold. *Cell* **163**, 1360–1374 (2015).

30. Cokelaer, T., Pultz, D., Harder, L. M., Serra-Musach, J. & Saez-Rodriguez, J. BioServices: a common Python package to access biological Web Services programmatically. *Bioinformatics* **29**, 3241–3242 (2013).

31. Bushnell, B. BBmap. Available at: https://sourceforge.net/projects/bbmap/. (Accessed: 10th January 2018)

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

33. Kang, D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. 0–10 (2019). doi:10.7287/peerj.preprints.27522v1

34. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).

35. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).

36. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

37. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).

38. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).

39. Fabbiano, S. *et al.* Functional Gut Microbiota Remodeling Contributes to the Caloric Restriction-Induced Metabolic Improvements. *Cell Metab.* **28**, 907-921.e7 (2018).

40. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).

41. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

42. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nat. Commun.* **9**, 2542 (2018).

43. Huerta-Cepas, J. *et al.* Fast Genome-Wide Functional Annotation through Orthology

Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).

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

45. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).

46. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).