

DATABASE

Open Access



MAGdb: a comprehensive high quality MAGs repository for exploring microbial metagenome-assemble genomes

Guo Ye^{1,2}, Hao Hong², Ting Li², Jin Li², Jia-Qi Wu², Shuai Jiang², Zhi-Tong Meng³, He-Tian Yuan², Wen Xue², Ai-Ling Li^{1,2}, Tao Zhou^{1,2}, Ting-Ting Li^{2*} and Tao Li^{1,2*}

*Correspondence:
tingtingli@mail.ncba.ac.cn;
tli@ncba.ac.cn

¹ Institute of Translational Medicine, Zhejiang University School of Medicine, Zhejiang, Hangzhou 310029, China
² Nanhu Laboratory, State Key Laboratory of Biomedical Analysis (SKLBA), Beijing 100850, China

³ School of Computer and Artificial Intelligence, Zhengzhou University, Henan, Zhengzhou 450001, China

Abstract

Metagenomic analyses of microbial communities have unveiled a substantial level of interspecies and intraspecies genetic diversity by reconstructing metagenome-assembled genomes (MAGs). The MAG database (MAGdb) boasts an impressive collection of 74 representative research papers, spanning clinical, environmental, and animal categories and comprising 13,702 paired-end run accessions of metagenomic sequencing and 99,672 high quality MAGs with manually curated metadata. MAGdb provides a user-friendly interface that users can browse, search, and download MAGs and their corresponding metadata information. It represents a valuable resource for researchers in discovering potential novel microbial lineages and understanding their ecological roles. MAGdb is publicly available at <https://magdb.nanhulab.ac.cn/>.

Keywords: Microbiome, Metagenome, Genome sequencing, Metagenome-assembled genomes, Database

Background

Microorganisms are the most abundant and widely distributed life forms on Earth, playing a crucial role in biogeochemical cycles and maintaining ecological balance [1, 2]. They inhabit diverse environments, from air, soil, and water to extreme habitats like deep-sea hydrothermal vents and glaciers [3, 4]. The importance of microorganisms has spurred extensive research, especially with the advancement of high-throughput sequencing technology and reduced sequencing costs. Genome-resolved metagenomic analyses, such as those for Earth's [5], glacier [6], aquatic [7], and human [8] microbiome genome catalogs, have been conducted using two main methodologies: the gene amplicon/marker genes sequencing (e.g., 16S rRNA) and shotgun metagenomics [9]. While 16S rRNA sequencing provides general insights into microbiota, it lacks the resolution to distinguish closely related taxa and unable to accurately identify viruses [10]. Therefore, shotgun metagenomics, which



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

sequences genomic DNA without targeting specific genes, has become the primary tool for studying microorganisms [11].

Genomic analyses are yielding unprecedented insights into microbial evolution and diversity and are elucidating the extent and complexity of the genetic variation in both hosts and pathogens that underlie diseases [12]. However, more than 99% of prokaryotes in the environment cannot be traditionally cultured in the laboratory. In parallel to culturing, *de novo* assembly of shotgun metagenomic reads and binning into MAGs, a culturing-independent and reference-free approach, is thought to be a useful strategy to efficiently discover the potential microbial diversity that is recalcitrant to the current culturing approaches in the laboratory [13–16]. Using MAGs has provided massive expansion of the tree of life from different environmental niches, enabling the discovery of unknown species and the exploration of microbial source tracking [17, 18]. In the last few years, thousands of MAGs have been reported [5, 19–22]. Building and mining MAGs sequences are becoming central processes and common tasks in inferring the functional capabilities of bacteria, as they provide genome-level insights into the functional potential of individual microbial entities. However, MAG sequences vary in quality and may contain omissions and contamination due to the inherent complexities of metagenomic data and the challenges associated with assembly processes [13, 23]. Consequently, recovering high-quality MAGs (more than 90% completeness and less than 5% contamination; hereafter referred to as HMAGs) based on the “minimum information about a metagenome-assembled genome” (MIMAG) standard [24] from shotgun metagenomic sequence data is a crucial process for future analysis. Moreover, metagenomic assemblies and binning are both time-consuming and resource-intensive processes. Therefore, there is an urgent need for an “all-in-one” database that contains high-quality MAGs data from a variety of environments and host-associated microbiota with rapidly growing metagenomic data. High-quality MAGs reference databases are required to confidently investigate the structure and function of complex microbial communities in natural or engineered ecosystems. But as far as we know, there are currently lacking comprehensive databases to provide permanent storage and public access for high-quality MAGs data based on representative metagenomic studies.

To address these limitations and facilitate the reusability and accessibility of MAGs data, we established MAGdb, a curated database that particularly focuses on high-quality assembled microbiome sequences. Overall, we collect 13,702 paired-end sequencing runs from shotgun metagenomic sequencing across 74 papers, spanning clinical, environmental, and animal research areas. The main features of MAGdb include: (i) manually curated paper information for each collected runs and all consistent format metadata, (ii) consistent taxonomic assignments of HMAGs and precomputed genome information, (iii) easily accessible categorized HMAGs with complete traceability to source raw data. MAGdb enables researchers to quickly acquire MAG sequences on microbiota of interest and provides sequence download for exploring the composition and roles of the microbiome in different areas.

Results

Design of MAGdb database

The construction scheme of the MAGdb database is illustrated in Fig. 1. In brief, the metagenomic raw data, MAGs (if provided) and the metadata were collected and

manually curated based on the unit of representative research papers (see method section). Subsequently, we employed a combination analyze pipeline of metagenomic assembly and binning and recovered MAGs from related publications that do not provide MAGs (Fig. 1). The MAGs were produced by three different binning tools and then integrated and refined to remove duplicates and improve the quality of assembled genomes with metaWRAP [3]. In order to provide strict genome quality control, we only selected those MAGs that met or exceeded the high-quality standard of >90% completeness and <5% contamination for subsequent analyses, which we referred to as the MAGdb catalog. Finally, all curated data were assembled into the database system, and the web platform was implemented.

Overview of the MAGdb content and statistics

To date, the MAGdb has successfully collected 13,702 microbial metagenomic sequence samples from 74 research publications, covering 66 countries across 5 continents, and these samples were classified into clinical, environmental, and animal categories (Fig. 2a). Clinical samples occupy the largest proportion (76.2%), followed by environmental samples (12.04%), with the lowest proportion of animal samples (11.4%) (Fig. 2a). Specifically, the clinical category included 29 publications and 10,439 run accessions, the

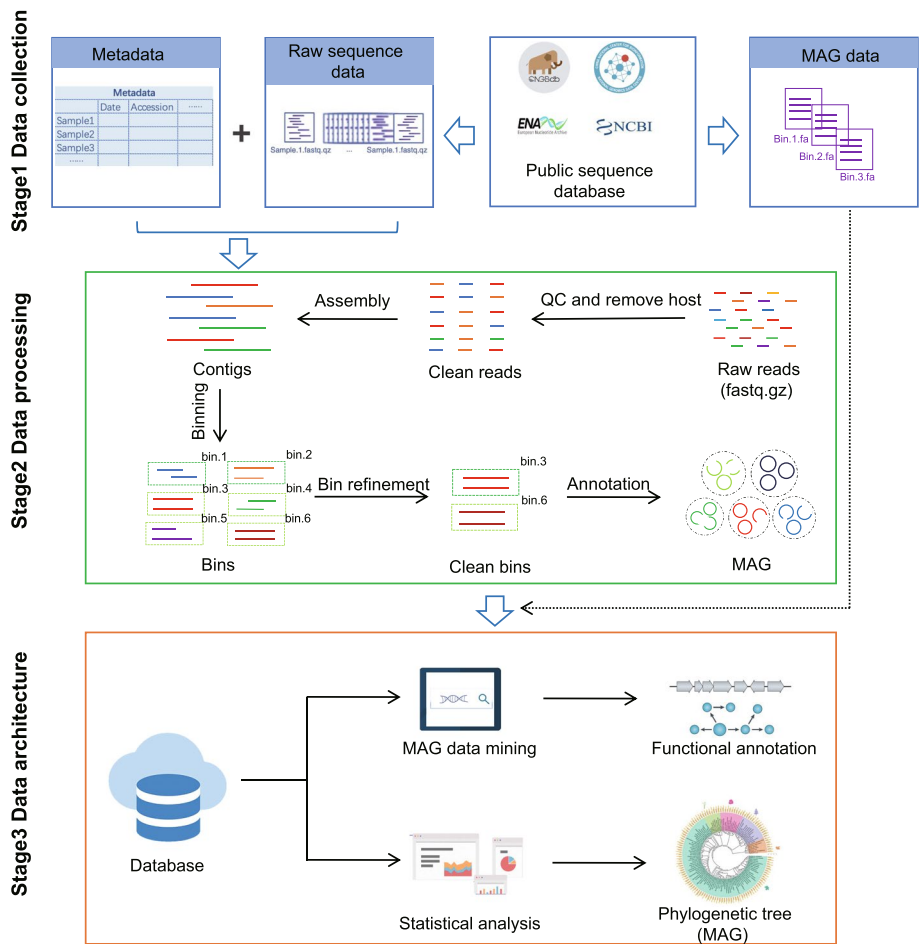


Fig. 1 Workflow for constructing the global MAG database

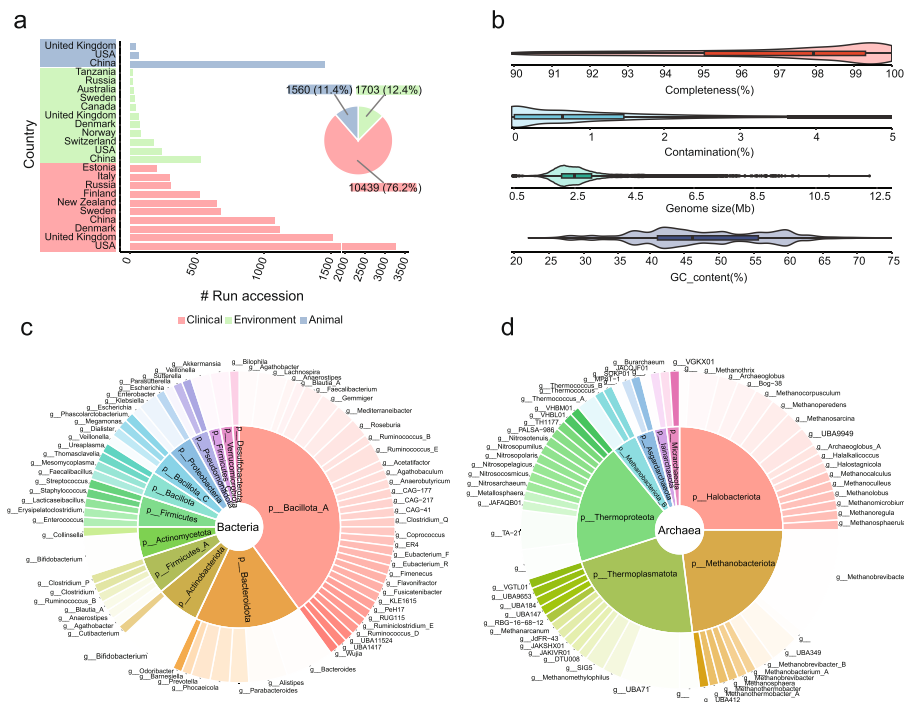


Fig. 2 Summary of the data statistics in MAGdb database. **a** The bar and pie chart depict the sample size distribution across top ten different countries and all collected run accessions in three categories. **b** Distribution of quality metrics for the HMAGs ($n = 99,672$), showing the interquartile range between the first and third quartiles and the line inside represents the median, respectively. **c** The distribution of quantitative events in different classification levels (domain, phylum, genus) in bacteria and **d** in archaea

Table 1 Detailed data statistics of MAGdb

Category	#Papers	#Paired-end runs	#HMAGs	#Species	#Genera
Clinical	29	10,439	81,933	2926	1014
Environmental	30	1703	6439	1019	1438
Animal	15	1560	11,300	1845	766
Total	74	13,702	99,672	5381*	2753*

#Papers: the number of papers collected in different categories of MAGdb database.
#Paired runs: the number of runs whose data possess paired-end read, with the corresponding metadata available in the public database.
#HMAGs: the number of HMAGs recovered from paired-end runs.
#Species: the number of species annotated with HMAGs.
#Genera: the number of genera annotated with HMAGs.
*The number of unique species/genera after removing duplicates.

environmental category included 30 publications and 1703 run accessions, and the animal category included 15 publications and 1560 run accessions (Table 1). These extensive datasets serve as the fundamental resources of the MAGdb database, providing an expansive landscape for exploring MAGs.

The MAGdb now contains a total of 99,672 HMAGs in three categories (Table 1). The HMAGs all meet or exceed the high-quality level of the MIMAG criteria (completion > 90%, contamination < 5%), exhibiting a mean completeness of 96.84% ($\pm 2.81\%$) and a mean contamination rate of 1.02% ($\pm 1.09\%$), with genome sizes ranging from 0.52

to 12.26 Mb and GC% content varying from 22.4% to 75% (Fig. 2b). We further taxonomically annotated these HMAGs using the GTDB-Tk based on the GTDB database. The MAGdb catalog covered 90 known phyla (82 for bacteria and 8 for archaea), 196 known classes (177 for bacteria and 19 for archaea), 501 known orders (474 for bacteria and 27 for archaea), and 2753 known genera (2687 for bacteria and 66 for archaea). The most frequently occurring genera and corresponding phyla in the bacterial domain and archaea domain are shown in Fig. 2c–d.

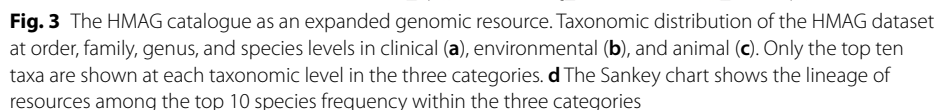
We also analyzed the correlation between sequencing read counts and HMAG completeness or the number of recovered MAGs. The relationship between per-sample sequencing read counts and HMAG mean completeness showed divergent trends across sample types in the clinical (a), environmental (c) and animal (e) categories: while human gut samples and animal-derived (e.g., sus scofa lung, fecal) exhibit progressive completeness declines, environmental samples (e.g., soil, water) show either gradual increases or stable plateaus (Additional file 1: Fig. S2 a, c, e). However, the number of recovered HMAGs increases with sequencing read counts in clinical (b), environmental (d), and animal (f) category, especially in fecal samples (Additional file 1: Fig. S2 b, d, f). These findings show that sequencing depth influences both MAG completeness and yield, and that optimal sequencing strategies should consider microbial community complexity.

The HMAG catalog exhibits a highly diverse and complex range of microbial species. We totally annotated 5381 species and 2753 genera from the 99,672 HMAGs, while there were still 6316 HMAGs remaining unclassified at the species level. The top 10 classification levels (order, family, genus and species) for each category are depicted in Fig. 3a–c. Taxonomic analysis revealed *Escherichia coli* as the dominant species in clinical samples. However, most HMAGs derived from environmental and animal specimens remained unclassified at the species level. The large proportion of unclassified HMAGs suggests extensive undiscovered microbial diversity in these ecosystems. We also analyzed the top 10 species of the three categories, as well as their higher taxonomic ranks shown in Fig. 3d.

Database web interface and modules

We designed a user-friendly interface (Fig. 4a) that allows users to effectively browse and query MAGs and related information. In short, MAGdb can be divided into three main modules, namely “Rawdata,” “MAG,” and “Download.” These modules provide detailed and convenient publication links, raw data metadata, MAGs sequences, and sequence information for users to browse, search, and download in various aspects of the microbiome (Fig. 4b, c). By employing these modules, users can easily mine the sequence information of MAGs, gaining valuable insights into microbial diversity, functional potential, and genetic characteristics.

The “Rawdata” module provides the list of the publication items in each category, labeled with the journal name and publication date (Fig. 4b, upper panel). Each publication item was provided with the number of collected run accessions, quality control reports, summary of the publication, official journal website link, and metadata. All the above characteristics contain additional links for users to explore detailed information. For example, the “quality control report” link provides multiple visualizations depicting the QC results of sequence data, such as reads length, GC content, adapter content, and



The “MAG” module primarily offers a comprehensive resource for users to browse and explore MAG sequences in each publication item (Fig. 4b, middle panel). Users can click the publication items to access a browsing page containing sequence information of all MAGs generated in the corresponding study. In addition, the “HMAG” link allows

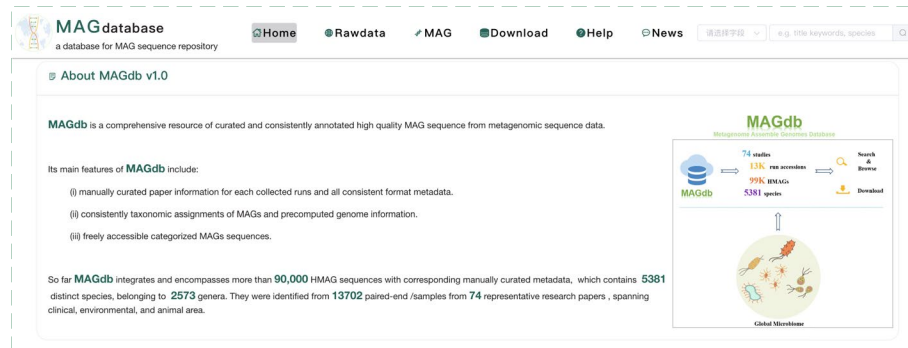
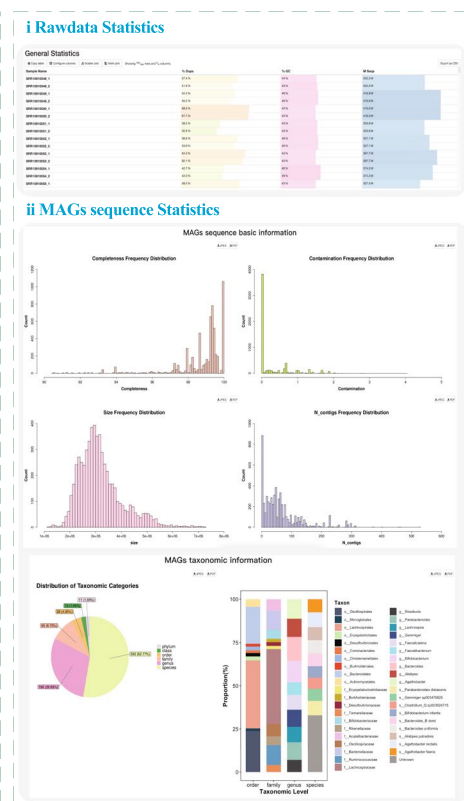
a web design and interface**b data browsing****c detailed information interface**

Fig. 4 Features of MAGdb database. **a** Website design and interface of the MAGdb. **b** Data browsing interface. **c** Detailed information interface

users to swiftly navigate to the global summary page that provides the MAG statistical plots from this study, including completeness, contamination, genome size, number of contigs, N50, and taxonomic classification information (Fig. 4c, lower panel). Users can freely download the plot as a portable graphic file as well as the MAGs statistic information.

The “Download” module provides the page for downloading the MAG sequences in terms of the publications in each category (Fig. 4b, lower panel). Each publication item directs to a link for downloading a compressed file. Additionally, users can download data of multiple publications at once through batch selection. In addition, the “Help” page offers comprehensive guidance and step-by-step procedures for users to navigate

and effectively utilize the MAGdb system. The “News” page displays information about versions, timestamps, and changes made in each version.

Phylogenetic and functional characterization of MAGdb

To investigate the evolutionary relationships and functional divergence within MAGdb, we first established a non-redundant genome set through dereplication at 95% ANI. This yielded 7303 non-redundant representative HMAGs, which were subsequently subjected to further analyses. To determine the phylogenetic relationships of these representative genomes, a maximum-likelihood phylogenetic tree of the representative HMAGs was generated with FastTree [25] based on 120 bacterial marker genes identified by GTDB-tk [26]. The HMAGs belonging to p__Bacillota_A exhibited the most widespread distribution in the phylogenetic tree, indicating high phylogenetic diversity. To better elucidate the functional diversity of the HMAGs in MAGdb, we annotated gene functions of the representative HMAGs with eggNOG databases [27], including COGs, KEGG pathway, and level-4 ECs. We found that a total of 94% of genes from the MAGdb had a match to at least one of the databases of COGs ($n=13,598,009$ genes across 24 functional categories), ECs ($n=4,311,408$ genes matching 3610 enzymes), KEGG ($n=4,913,916$ genes from 872 pathway), GOs ($n=1,172,159$ genes from 20,792 GOs), and CAZy ($n=194,221$ from 141 CAZys) (Fig. 5b). To gain further insights into the relationships between microbial phyla, we constructed co-occurrence networks based on the frequency of each HMAG phylum across the three categories. The co-occurrence network analysis revealed distinct patterns of microbial interactions across different sample types. In animal-associated samples, we observed strong positive correlations between microbial phyla such as p__Bacillota_C, p__Spirochaetota, and P__Bacteroidota. These phyla exhibited a high degree of co-occurrence, suggesting potential ecological relationships and functional associations within the animal microbiota. In contrast, environmental samples showed strong co-occurrence between p__Latescibacterota, p__Marinisomatota, and p__Bacillota_A, forming multiple network modules that indicate significant interactions among these phyla in the environmental microbiome. These phyla may play a crucial role in the microbial communities under specific environmental conditions. However, clinical samples exhibited predominantly negative co-occurrence patterns among major phyla, including p__Bacillota_A, p__Firmicutes_A, and p__Actinobacteriota (Fig. 5c). These antagonistic interactions likely reflect niche competition due to the variability introduced by factors such as host health status, treatments, or environmental conditions.

Discussion and future direction

Technological advancements in assembly and binning tools have led to a significant increase in the assembled fraction of the average metagenome, coupled with an exponential increase in the number of MAGs [28]. In this study, we constructed the MAGdb, an online database of curated and consistently annotated HMAGs. With 13,702 samples collected from 74 metagenomic research papers, 99,672 HMAGs were obtained in the current version of MAGdb, linked to 5381 species. Notably, 6316 HMAGs (6.3%) have not been annotated to the species level. These likely represent novel taxa, indicating that many new species remain to be discovered. With the continued accumulation

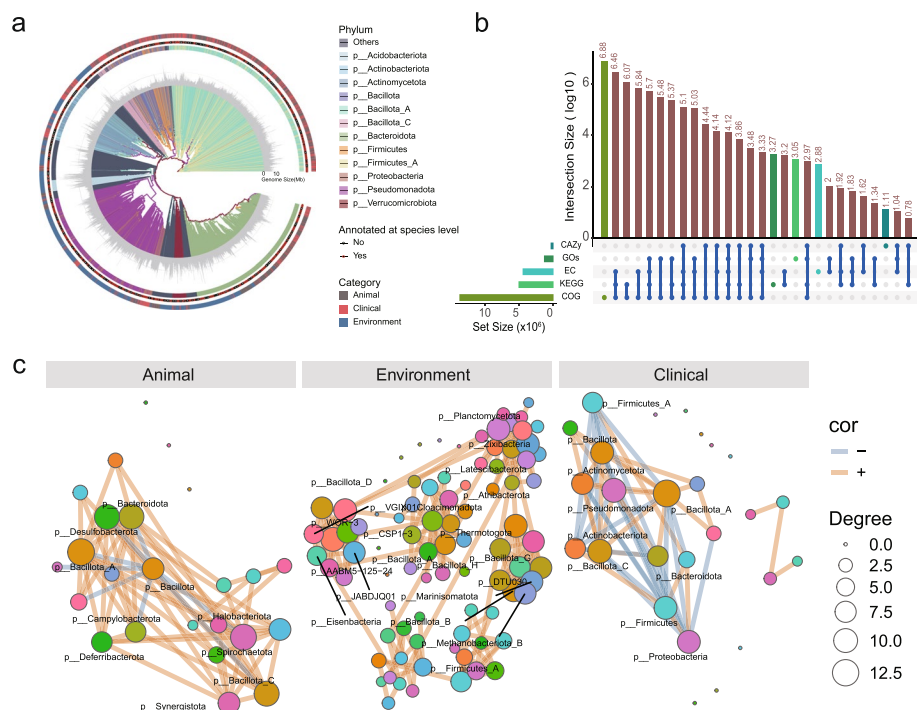


Fig. 5 Exploratory Analysis of MAGdb. **a** Phylogenetic tree of non-redundant HMAGs constructed in this study. From the inner to outer circles, the first circle indicates the genome size of HMAGs, the second circle indicates the corresponding phyla which MAGs belonged to, the third circle shows whether the MAGs could be annotated to a species or not, and the fourth circle indicates the source category of each HMAG's original sample category. **b** Number of proteins with functional annotations across the five functional categories and their degree of overlap. Vertical bars represent the number of proteins unique to each functional category or shared between the specific functional categories. Horizontal bars in the lower panel indicate the total number of proteins with functional annotation in each functional category. **c** Co-occurrence network of differential phylum. Interactions between phylum (nodes) are represented by connecting lines (edges), and each node is colored according to the phylum to which it belongs. The colors of the edge lines represent positive (light red) or negative (dark blue) interactions

of sequencing data and improvements in annotation methods, currently unidentified HMAGs may eventually be classified as novel microbial species, further enriching our understanding of microbial diversity. MAGdb exhibits outstanding comprehensive coverage over previous metagenomic databases in the microbial sequence data analysis area. Extensive comparisons with existing metagenomic databases indicated that MAGdb contains more MAGs with higher quality [29, 30], thus enabling more accurate and reliable microbial genome analyses.

We believe that MAGdb will have positive impacts on various aspects of microbiological and metagenomics researches. Specifically, MAGdb serves as a comprehensive resource for the identification, characterization, and functional annotation of metagenome-assembled genomes, enhancing our understanding of microbial functions and interactions. For example, inferring the functional capabilities of microorganisms from mining MAGs is becoming a central process in microbiology in many studies [31], such as CRISPR/Cas loci [32, 33], antimicrobial resistance genes [34] and mobile genetic elements [35]. Moreover, MAG sequences can offer invaluable insights into the identification of previously undescribed organisms, often referred to as the microbial dark matter, along with a comprehensive understanding of their genetic composition [20, 36, 37]. The

advantage of MAGs exploration of unknown species becomes crucial in the context of emerging and novel infectious diseases, providing essential clues for the early detection of potential pathogens [18]. MAGs also provide information beyond the reference genome. Unlike reference genomes, MAGs facilitate the identification of unique genetic features, such as specific mutation sites and genetic variations, and the identification of key genetic features associated with environmental adaptability and ecological roles. Additionally, subspecies variations, such as strain diversity, mobile gene composition, and copy number variations, have been demonstrated to be associated with host traits and lifestyle habits [38]. The insights derived from MAGs not only extend our understanding beyond the limitations of existing reference genomes but also play a critical role in advancing diverse fields, ranging from environmental microbiology to research in human health.

The MAGdb database will continue to be upgraded and improved in the future. MAGdb is in its release version 1.0, and it will regularly get updates per year, due to the exponential number of novel metagenomes sequence added to public repositories. In addition to continuously collecting new metagenomic data over the next few years, we plan to add new contents to MAGdb, including (but not limited to) metavirome data for viral operational taxonomic units (vOTU) and deploy the analysis modules for mining functional profiles and evolutionary relationships. Additionally, we observed a gradual increase of third-generation sequencing data (long-read) within metagenome at the time when this manuscript was prepared. We therefore anticipate providing MAGs derived from third-generation sequencing assemblies in future versions. With these advancements, we believe that our database improves the reusability and exploration of metagenomic data further and helps users better understand the relationship between the microbial populations and their interactions with the environment where they live.

Conclusion

MAGdb provides a comprehensive platform for global integration and standardization of MAGs, facilitating microbial diversity exploration and cross-study comparisons for users in downstream analysis. The platform also leverages advanced visualization to illustrate the sequence data and MAGs patterns within each publication. Additionally, the website provides detailed annotations of each HMAG record including sequence characteristics and taxonomy affiliation. By providing a centralized and consistent resource, MAGdb enhances the reproducibility and comparability of studies, allowing researchers to explore microbial diversity and function on a large scale. Furthermore, the detailed annotations of HMAG records, including sequence characteristics and taxonomy affiliations, offer a robust framework for downstream analysis, facilitating the discovery of novel species, functional genes, and their roles in various ecosystems.

Construction and content

Data collection and curation

We collected the literature containing metagenome studies from the PubMed database with relevant keywords, such as “microbiome AND shotgun metagenomics” and “de-novo assembly AND microbial AND genome,” providing a list of more than 2000 publications since 2015. In order to include only high-quality studies, we firstly excluded

reviews, letters, editorials, and other publications without disclosing the original sequencing data. We then refined this list by reading the abstracts and results of each publication, keeping only those studies that meet the following criteria: (1) Research quality: Studies published in peer-reviewed, high-impact journals, known for the field of microbiome metagenomics; (2) Data accessibility: Studies that provided raw data, metadata, or detailed experimental protocols to ensure reproducibility and independent verification; (3) Focus on relevant topics: Studies directly related to microbiome metagenomics in clinical (e.g., human gut, skin samples), environmental (e.g., soil, marine samples), or animal (e.g., cow, goat samples) contexts. Only those publications that aligned with the above criteria were retained. Subsequently, we manually reviewed the experiment section and only extracted shotgun metagenomics validated samples. Through multiple rounds of manual curation, the information from selected studies was collected into an Excel sheet with pre-defined fields. We did not include any studies which required additional ethics committee approvals or authorizations for access. As a result, we finally acquired a total of 74 metagenome-associated studies for subsequent analysis (Additional file 1: Fig. S1).

Paired-end raw metagenome shotgun sequencing runs together with MAGs data (if provided) were downloaded from online repositories (EBI-ENA [39], NCBI-SRA [40], CNGB-CNSA [41] and NGDC-GSA [42]) whose links can be found in the related publications. Meanwhile, for each run or sample, we also collected relevant metadata, including technical metadata, such as the run accession, sequencing platform, number of reads, and base length, as well as biological metadata, such as sample source, sample name, country, and continent. Finally, the meticulously curated metadata of raw metagenome sequencing data were compiled into an Excel spreadsheet with the above pre-defined fields.

Data processing

Our systematic analysis included 74 published metagenomic studies. Among these, only 5 studies provided both single-sample assembled genomes and complete assembly provenance metadata. For these studies containing MAGs, we assessed genome quality using CheckM (v1.2.2) [43], retaining only HMAGs that met the criteria (completeness > 90%, contamination < 5%). For the remaining 69 studies lacking MAGs, we performed de novo analysis of 6,692 metagenomic samples using a standardized bioinformatic pipeline based on the EasyMetagenome framework [44] (<https://github.com/YongxinLiu/EasyMetagenome>). Additionally, fastqc (v0.12.1, <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) was employed to evaluate the overall quality of the raw metagenomic data, followed by the use of the MultiQC (v1.21) [45] tool to aggregate results from the fastqc report across many samples into a single self-contained HTML report before subsequent analysis.

Next, in order to obtain the MAG sequences from each publication, we processed and analyzed the paired-end metagenomic data from articles that did not provide MAG sequences [46]. In brief, the fastp (v0.23.2) [47] was used for short paired-end read quality filtering and adapter removal with the specified parameter “-dedup -q 20”, followed by host DNA removal via bowtie2 (v2.5.1) [48] alignment against the host reference genome with setting parameters “-end-to-end -no-mixed -no-discordant -no-unal

–very-sensitive” if necessary (raw metagenomic data obtained from clinical and animal samples undergoes a rigorous process to remove host genome contamination before assembly analysis). Subsequently, the remaining paired-end sequence reads were assembled into contiguous sequences for each of the sequenced samples separately by MEGAHIT (v1.2.9) [49] with default parameters. Thereafter, the metawrap binning module [3], integrating MaxBin2 (v2.2.7) [50], metaBAT2 (v2.12.1) [51], and CONCOCT (v1.1.0) [52] three binning tools, were used to bin the assemblies with the “–metabat2 –max-bin2 –concoct” option. The default of the minimum length of contigs used for constructing bins with MaxBin2 and CONCOCT were 1000 bp, and metaBAT2 was defaulted to 1500 bp. To reconcile and dereplicate the three generated binner outputs, refinement of MAGs was performed by the bin_refinement module of metaWRAP [3], and CheckM (v1.2.2) [43] was used to estimate the completeness and contamination of the bins with parameters “–x 10 –c 50”, corresponding to the minimum completion and maximum contamination were 50% and 10%, respectively. Finally, we only kept the MAGs exhibiting > 90% completeness and < 5% contamination as the HMAG sequence for further analysis. All HMAGs were taxonomically annotated using the GTDB-Tk (v2.3.0) [26] (reference database version R214) with “classify_wf” workflow with default parameters [53], which produced standardized taxonomic labels that were used for user reference. The final results of each study were compiled in a single matrix-like table containing information for all generated HMAGs in this process (Fig. 1).

Exploratory analysis of HMAGs

To systematically explore all recovered HMAGs, we performed phylogenetic diversity and functional potential analysis. In order to reduce redundancy of the HMAGs and to determine the representative genomes, the resulting HMAGs across all samples and assemblies were dereplicated based on 95% ANI with the following options: “–S_algorithm skani –clusterAlg centroid –pa 0.9 –sa 0.95 –nc 0.10 –cm larger –multi-round_primary_clustering” using dRep (v3.5.0) [54]. These non-redundant HMAGs were subsequently subjected to comprehensive phylogenetic analysis and functional annotation. The taxonomy annotation of the non-redundant HMAGs was performed using the module “classify_wf” of GTDB-Tk (v2.3.0) [26] against the GTDB release R214 with default parameters. The phylogenetic tree of bacterial non-redundant representative HMAG was built using FastTree (v2.1.11) [25] with the protein sequence alignments generated by the GTDB-Tk tool with parameter setting “–wag –boot 1000”, while all other parameters were set to their default values. Tree visualization and annotation were performed using an R package ggtree (v3.14.0) [55]. The putative protein-coding sequences (CDSs) of representative genomes were predicted using Prodigal (v2.6.3) [56] with the “–p meta” parameter. The predicted CDSs were then dereplicated by cd-hit-est (v4.8.1) [57] with the options “–c 0.95 –aS 0.9”. Subsequently, the representative, non-redundant CDSs were annotated with EggNOG-mapper (v2.1.12) [58], employing DIAMOND search mode against the EggNOG v5.0 database with default parameters [27]. The KEGG (Kyoto Encyclopedia of Genes and Genome) pathway, Clusters of Orthologous Groups of proteins (COGs) functional annotations, level-4 Enzyme Commission categories (Ecs), Gene Ontologies (GOs) and carbohydrate-active enzymes (CAZy) were derived from the EggNOG-mapper results. To construct co-occurrence networks of HMAGs

across the three categories, we calculated the frequency of each HMAG phylum in every sample. This frequency matrix served as the input for network construction. Microbial co-occurrence networks at the phylum level were constructed using the ggClusterNet 2 R package (v2.00) [59] with parameters “ $N=0$, $r=0.2$, $p=0.05$, method = spearman”.

Web implementation

All HAMGs and metadata were stored in a MongoDB database (<https://docs.mongodb.com/>). The MAGdb web-interface (the frontend webpages) was implemented using JavaScript and HTML for frontend development. The used core JavaScript libraries include Vue.js (<https://vuejs.org/>) as the main frontend framework. The backend was mostly implemented in Node.js (<https://nodejs.org/>) as the framework for the application. MAGdb is available online without registration and is optimized for Chrome (recommended), Firefox, Windows Edge, and macOS Safari.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-025-03711-6>.

Additional file 1: This file contains two supplementary figures. Fig. S1 Screening flow of the included studies and samples. Fig. S2 Relationship between per-sample metagenomic sequencing read count and mean HMAG completeness or the number of recovered MAGs count in different sample types

Acknowledgements

We would like to thank Professor Hebing Chen's team for the fruitful discussion. Additionally, we also would like to express our deep gratitude to the NCBI and all available projects for their valuable resources for MAGdb.

Authors' contributions

Tao Li (conceptualization, funding acquisition, supervision), Ting-ting Li (conceptualization, supervision, investigation, methodology, writing review and editing), Guo Ye (data curation, formal analysis, investigation, methodology, visualization, writing original draft, writing review and editing), Hao Hong (formal analysis, investigation, methodology, visualization), Ting Li (data curation, investigation), Jin Li (data curation, investigation), Jia-Qi Wu (data curation, visualization), Shuai Jiang (funding acquisition, language polishing, grammar checking), Zhi-Tong Meng (website maintenance, content update), He-Tian Yuan (literature search, data collection), Wen Xue (language polishing, grammar checking), Ai-Ling Li (project administration, supervision), Tao Zhou (funding acquisition, project administration, supervision). All the authors have read and approved the final manuscript.

Funding

This work was supported by grants from China National Natural Science Foundation (No. 82341098 to Tao Zhou, No. 82130052 to Tao Li, and No. 32100421 to Shuai Jiang), Nanhu Laboratory (No. NSS2021CI05002) and the Central Government Guides Local Science and Technology Development Fund Projects (No. 2024ZYYDSA400333).

Data availability

The MAGdb database and its content are freely accessible to all academic users at <https://magdb.nanhulab.ac.cn>. Users can download HMAG sequence via the 'Download' page. The specific or selected samples can be exported on the 'Rawdata' or 'MAG' page. The curated metadata for all projects and the analysis codes are available on GitHub at <https://github.com/YeGuoZJU/MAGdbV2> [60]. The analysis codes were also uploaded to Zenodo at <https://zenodo.org/records/15955387> [61].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Peer review information

Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Received: 15 August 2024 Accepted: 24 July 2025

Published online: 10 September 2025

References

- Shoemaker WR, Locey KJ, Lennon JT. A macroecological theory of microbial biodiversity. *Nat Ecol Evol*. 2017;1:1–6.
- Martiny JBH, Jones SE, Lennon JT, Martiny AC. Microbiomes in light of traits: a phylogenetic perspective. *Science*. 2015. <https://doi.org/10.1126/science.aac9323>.
- Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome*. 2018;6:1–13.
- Gilbert JA, Stephens B. Microbiology of the built environment. *Nat Rev Microbiol*. 2018;16:661–70.
- Nayfach S, Roux S, Seshadri R, Udway D, Varghese N, Schulz F, et al. A genomic catalog of Earth's microbiomes. *Nat Biotechnol*. 2021;39:499–509.
- Liu Y, Ji M, Yu T, Zaugg J, Anesio AM, Zhang Z, et al. A genome and gene catalog of glacier microbiomes. *Nat Biotechnol*. 2022. <https://doi.org/10.1038/s41587-022-01367-2>.
- Cheng M, Luo S, Zhang P, Xiong G, Chen K, Jiang C, et al. A genome and gene catalog of the aquatic microbiomes of the Tibetan Plateau. *Nat Commun*. 2024. <https://doi.org/10.1038/s41467-024-45895-8>.
- Human T, Jumpstart M, Strains R, Institutes TN, Institutes N, Project HM, et al. A catalog of reference genomes from the human microbiome. *Genome*. 2010;328:994–9.
- Bharti R, Grimm DG. Current challenges and best-practice protocols for microbiome analysis. *Brief Bioinform*. 2021;22:178–93.
- Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat Commun*. 2019;10:1–11.
- Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol*. 2017;35:833–44.
- Lasken RS, McLean JS. Recent advances in genomic DNA sequencing of microbial species from single cells. *Nat Rev Genet*. 2014;15:577–84.
- Forouzan E, Shariati P, Mousavi Maleki MS, Karkhane AA, Yakhchali B. Practical evaluation of 11 de novo assemblers in metagenome assembly. *J Microbiol Methods*. 2018;151:99–105.
- Sharon I, Banfield JF. Genomes from metagenomics. *Science*. 2013;342:1057–8.
- Zhou Y, Liu M, Yang J. Recovering metagenome-assembled genomes from shotgun metagenomic sequencing data: methods, applications, challenges, and opportunities. *Microbiol Res*. 2022;260: 127023.
- Sangwan N, Xia F, Gilbert JA. Recovering complete and draft population genomes from metagenome datasets. *Microbiome*. 2016;4:1–11.
- Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. *Nat Microbiol*. 2016;1:1–6.
- Ko KKK, Chng KR, Nagarajan N. Metagenomics-enabled microbial surveillance. *Nat Microbiol*. 2022;7:486–96.
- Zeng S, Patangia D, Almeida A, Zhou Z, Mu D, Paul Ross R, et al. A compendium of 32,277 metagenome-assembled genomes and over 80 million genes from the early-life human gut microbiome. *Nat Commun*. 2022;13:1–15.
- Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*. 2019;176:649–662.e20.
- Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol*. 2021;39:105–14.
- Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, et al. A new genomic blueprint of the human gut microbiota. *Nature*. 2019;568:499–504.
- Zhang Z, Yang C, Veldsman WP, Fang X, Zhang L. Benchmarking genome assembly methods on metagenomic sequencing data. *Brief Bioinform*. 2023;24:1–17.
- Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol*. 2017;35:725–31.
- Price MN, Dehal PS, Arkin AP. FastTree 2 - approximately maximum-likelihood trees for large alignments. *PLoS One*. 2010. <https://doi.org/10.1371/journal.pone.0009490>.
- Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics*. 2020;36:1925–7.
- Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. EggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res*. 2019;47:D309–14.
- Ayling M, Clark MD, Leggett RM. New approaches for metagenome assembly with short reads. *Brief Bioinform*. 2020;21:584–94.
- Gurbich TA, Almeida A, Beracochea M, Burdett T, Burgin J, Cochrane G, et al. MGnify genomes: a resource for biome-specific microbial genome catalogues: MGnify biome-specific genome catalogues. *J Mol Biol*. 2023;435: 168016.
- Shi W, Qi H, Sun Q, Fan G, Liu S, Wang J, et al. GcMeta: A Global Catalogue of Metagenomics platform to support the archiving, standardization and analysis of microbiome data. *Nucleic Acids Res*. 2019;47:D637–48.
- Eisenhofer R, Odriozola I, Alberdi A. Impact of microbial genome completeness on metagenomic functional inference. *ISME Commun*. 2023;3:1–5.
- Münch PC, Franzosa EA, Stecher B, McHardy AC, Huttenhower C. Identification of natural CRISPR systems and targets in the human microbiome. *Cell Host Microbe*. 2021;29:94–106.e4.

33. Ciciani M, Demozzi M, Pedrazzoli E, Visentin E, Pezzè L, Signorini LF, et al. Automated identification of sequence-tailored Cas9 proteins using massive metagenomic data. *Nat Commun*. 2022;13:1–8.
34. Lee K, Raguideau S, Sirén K, Asnicar F, Cumbo F, Hildebrand F, et al. Population-level impacts of antibiotic usage on the human gut microbiome. *Nat Commun*. 2023. <https://doi.org/10.1038/s41467-023-36633-7>.
35. Vatanen T, Jabbar KS, Ruotula T, Honkanen J, Avila-Pacheco J, Siljander H, et al. Mobile genetic elements from the maternal microbiome shape infant gut microbial assembly and metabolism. *Cell*. 2022;185:4921–4936.e15.
36. Pavlopoulos GA, Baltoumas FA, Liu S, Selvitopi O, Camargo AP, Nayfach S, et al. Unraveling the functional dark matter through global metagenomics. *Nature*. 2023. <https://doi.org/10.1038/s41586-023-06583-7>.
37. Nayfach S, Pérez-Espino D, Call L, Low SJ, Sberro H, Ivanova NN, et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat Microbiol*. 2021;6:960–70.
38. Zahavi L, Lavon A, Reicher L, Shoer S, Godneva A, Leviatan S, et al. Bacterial SNPs in the human gut microbiome associate with host BMI. *Nat Med*. 2023. <https://doi.org/10.1038/s41591-023-02599-8>.
39. Yuan D, Ahamed A, Burgin J, Cummins C, Devraj R, Gueye K, et al. The European Nucleotide Archive in 2023. *Nucleic Acids Res*. 2023;1–6.
40. Katz K, Shutov O, Lapoint R, Kimelman M, Rodney Brister J, O'Sullivan C. The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Res*. 2022;50:D387–90.
41. Guo X, Chen F, Gao F, Li L, Liu K, You L, et al. CNSA: a data repository for archiving omics data. *Database*. 2020;2020:1–6.
42. Chen T, Chen X, Zhang S, Zhu J, Tang B, Wang A, et al. The Genome Sequence Archive family: toward explosive data growth and diverse data types. *Genomics, Proteomics Bioinforma*. 2021;19:578–83.
43. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*. 2015;25:1043–55.
44. Bai D, Chen T, Xun J, Ma C, Luo H, Yang H, et al. EasyMetagenome: a user-friendly and flexible pipeline for shotgun metagenomic analysis in microbiome research. *iMeta*. 2025;4:1–23. Available from: <https://onlinelibrary.wiley.com/doi/https://doi.org/10.1002/imt2.70001>
45. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016;32:3047–8.
46. Saheb Kashaf S, Almeida A, Segre JA, Finn RD. Recovering prokaryotic genomes from host-associated, short-read shotgun metagenomic sequencing data. *Nat Protoc*. 2021;16:2520–41.
47. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018;34:i884–90.
48. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9:357–9.
49. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*. 2015;31:1674–6.
50. Wu YW, Simmons BA, Singer SW. Maxbin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*. 2016;32:605–7.
51. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*. 2019;2019:1–13.
52. Alneberg J, Bjarnason BS, De Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic contigs by coverage and composition. *Nat Methods*. 2014;11:1144–6.
53. Parks DH, Chuvpochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol*. 2018;36:996.
54. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: A tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J*. 2017;11:2864–8. Available from: <https://doi.org/10.1038/ismej.2017.126>
55. Chen M, Luo X, Xu S, Li L, Li J, Xie Z, et al. Scalable method for exploring phylogenetic placement uncertainty with custom visualizations using treeio and ggtree. *iMeta*. 2025;1–8.
56. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. 2010;11: 119. <https://doi.org/10.1186/1471-2105-11-119>.
57. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012;28:150–2.
58. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol Biol Evol*. 2021;38:5825–9.
59. Wen T, Liu YX, Liu L, Niu G, Ding Z, Teng X, et al. ggClusterNet 2: an R package for microbial co-occurrence networks and associated indicator correlation patterns. *iMeta*. 2025;1–12.
60. Guo Y, Hao H, Ting L, Jin L, Jia-Qi W, Shuai J, et al. MAGdb: a comprehensive high quality MAGs repository for exploring microbial metagenome-assemble genomes. *Github*. 2025. Available from: <https://github.com/YeGuoZJU/MAGdbV2>
61. Guo Y, Hao H, Ting L, Jin L, Jia-Qi W, Shuai J, et al. MAGdb: a comprehensive high quality MAGs repository for exploring microbial metagenome-assemble genomes. *Zenodo*. 2025; Available from: <https://zenodo.org/records/15955387>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.