

RESEARCH ARTICLE

EasyMetagenome: A user-friendly and flexible pipeline for shotgun metagenomic analysis in microbiome research

Defeng Bai¹  | Tong Chen²  | Jiani Xun¹  | Chuang Ma^{1,3}  |
Hao Luo¹  | Haifei Yang^{1,4}  | Chen Cao⁵ | Xiaofeng Cao⁶  |
Jianzhou Cui⁷ | Yuan-Ping Deng⁸ | Zhaochao Deng⁹ | Wenxin Dong¹⁰ |
Wenxue Dong¹¹ | Juan Du¹² | Qunkai Fang¹³ | Wei Fang¹⁴ | Yue Fang¹⁵ |
Fangtian Fu¹⁶ | Min Fu¹⁷ | Yi-Tian Fu¹⁸ | He Gao¹⁹ | Jingping Ge²⁰ |
Qinglong Gong²¹ | Lunda Gu²² | Peng Guo²³ | Yuhao Guo²⁰ | Tang Hai²⁴ |
Hao Liu²⁵ | Jieqiang He²⁶ | Zi-Yang He²⁷ | Huiyu Hou¹ | Can Huang²⁸ |
Shuai Ji²⁹ | ChangHai Jiang³⁰ | Gui-Lai Jiang³¹  | Lingjuan Jiang³² |
Ling N. Jin³³  | Yuhe Kan³⁴ | Da Kang³⁵ | Jin Kou³⁶ | Ka-Lung Lam³⁷ |
Changchao Li³³ | Chong Li³⁸  | Fuyi Li³⁹ | Liwei Li⁴⁰ | Miao Li⁴¹ |
Xin Li⁴² | Ye Li⁴³ | Zheng-Tao Li⁴⁴ | Jing Liang⁴⁵ | Yongxin Lin⁴⁶ |
Changzhen Liu⁴⁷ | Danni Liu⁴⁸ | Fengqin Liu⁴⁹ | Jia Liu⁵⁰ | Tianrui Liu⁵¹ |
Tingting Liu⁵² | Xinyuan Liu⁵³ | Yaqun Liu⁵⁴ | Bangyan Liu⁵⁵ |
Minghao Liu⁵⁶ | Wenbo Lou²¹ | Yaning Luan¹⁵ | Yuanyuan Luo⁵³ |
Hujie Lv^{1,57} | Tengfei Ma⁵⁸ | Zongjiong Mai⁵⁹ | Jiayuan Mo⁴⁵ |
Dongze Niu⁶⁰ | Zhuo Pan⁶¹ | Heyuan Qi⁶² | Zhanyao Shi⁶³ |
Chunjiao Song⁶⁴ | Fuxiang Sun⁶⁵ | Yan Sun⁶⁶ | Sihui Tian⁶⁷ | Xiulin Wan¹ |
Guoliang Wang⁶⁸ | Hongyang Wang⁶⁹  | Hongyu Wang⁷⁰ |
Huanhuan Wang⁷¹ | Jing Wang⁷² | Jun Wang⁷³ | Kang Wang⁷⁴ |
Leli Wang⁷⁵ | Shao-kun Wang⁷⁶ | Xinlong Wang⁵² | Yao Wang¹ |
Zufei Xiao⁷⁷ | Huichun Xing⁷⁸ | Yifan Xu⁷⁸ | Shu-yan Yan⁷⁹ | Li Yang²² |
Song Yang⁷⁸ | Yuanming Yang⁸⁰ | Xiaofang Yao⁷⁵ | Salsabeel Yousuf¹ |
Hao Yu⁹ | Yu Lei⁸¹ | Zhengrong Yuan⁸²  | Meiyin Zeng¹ |
Chunfang Zhang⁹ | Chungge Zhang⁸³ | Huimin Zhang⁸⁴ | Jing Zhang⁴⁸ |
Na Zhang⁸⁵ | Tianyuan Zhang¹ | Yi-Bo Zhang⁷⁹ | Yupeng Zhang⁸⁶ |
Zheng Zhang⁸⁷ | Mingda Zhou⁸⁸ | Yuanping Zhou⁸⁹ | Chengshuai Zhu⁴⁴ |
Lin Zhu⁹⁰ | Yue Zhu⁹¹ | Zhihao Zhu⁸⁹ | Hongqin Zou⁹² | Anna Zuo⁹³ |

For affiliations refer to page 17.

Defeng Bai, Tong Chen, Jiani Xun, Chuang Ma, Hao Luo, and Haifei Yang contributed equally to this study.

This is an open access article under the terms of the [Creative Commons Attribution](#) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). *iMeta* published by John Wiley & Sons Australia, Ltd on behalf of *iMeta* Science.

Wenxuan Dong⁹⁴ | Tao Wen⁹⁵  | Shifu Chen^{48,96,97}  | Guoliang Li⁹⁸  |
Yunyun Gao¹  | Yong-Xin Liu¹ 

Correspondence

Tong Chen, State Key Laboratory for Quality Ensurance and Sustainable Use of Dao-di Herbs, National Resource Center for Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, China.
Email: chentong_biology@163.com

Tao Wen, College of Resource and Environmental Sciences, Nanjing Agricultural University, Nanjing, Jiangsu, China.
Email: taowen@njau.edu.cn

Shifu Chen, LifeX Institute, School of Medical Technology, Gannan Medical University, Ganzhou, China.
Email: chen@haplox.com

Guoliang Li, Jiangxi Provincial Key Laboratory of Conservation Biology, College of Forestry, Jiangxi Agricultural University, Nanchang, Jiangxi, China.
Email: guoliangli2016@gmail.com

Yunyun Gao and Yong-Xin Liu, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, Guangdong, China.
Email: gaoyunyun@caas.cn and liuyongxin@caas.cn

Funding information

National Natural Science Foundation of China, Grant/Award Numbers: 32470055, U23A20148; Agricultural Science and Technology Innovation Program, Grant/Award Number: CAAS-ZDRW202308; China Postdoctoral Science Foundation, Grant/Award Number: 2024M753580; The Fundamental Research Funds for the Central Public Welfare Research Institutes, Grant/Award Number: ZZ13-YQ-095; The Scientific and Technology Innovation Project of China Academy of Chinese Medical Sciences, Grant/Award Number: CI2021A04115; Scientific and technological innovation project of China Academy of Chinese Medical Sciences, Grant/Award Number: CI2023E002; The Key Project at Central Government Level: The Ability Establishment of Sustainable Use for Valuable Chinese Medicine Resources, Grant/Award Number: 2060302

Abstract

Shotgun metagenomics has become a pivotal technology in microbiome research, enabling in-depth analysis of microbial communities at both the high-resolution taxonomic and functional levels. This approach provides valuable insights of microbial diversity, interactions, and their roles in health and disease. However, the complexity of data processing and the need for reproducibility pose significant challenges to researchers. To address these challenges, we developed EasyMetagenome, a user-friendly pipeline that supports multiple analysis methods, including quality control and host removal, read-based, assembly-based, and binning, along with advanced genome analysis. The pipeline also features customizable settings, comprehensive data visualizations, and detailed parameter explanations, ensuring its adaptability across a wide range of data scenarios. Looking forward, we aim to refine the pipeline by addressing host contamination issues, optimizing workflows for third-generation sequencing data, and integrating emerging technologies like deep learning and network analysis, to further enhance microbiome insights and data accuracy. EasyMetagenome is freely available at <https://github.com/YongxinLiu/EasyMetagenome>.

KEYWORDS

metagenome, microbiome, microbiota, pipeline, visualization

Highlights

- EasyMetagenome offers multiple installation options and detailed question and answer explanations.
- EasyMetagenome includes detailed parameter explanations to accommodate various data scenarios, ensuring flexibility and adaptability.
- EasyMetagenome supports comprehensive downstream analysis from raw data, including read-based, assembly-based, binning, and genome analysis, with the ability to generate publication-ready visualizations.
- EasyMetagenome is developed as an open-source project, encouraging collaboration from the community. It's available for access and contribution at <https://github.com/YongxinLiu/EasyMetagenome>.

INTRODUCTION

Microorganisms, including bacteria, archaea, fungi, viruses, and protists, are ubiquitous in natural environments [1–3]. The term “microbiome” broadly refers to the community of microorganisms and their genomes within a particular host or environment [4]. Microbiome research has significantly expanded our understanding across various fields, including agriculture, food science, biotechnology, bio-economics, animal and plant nutrition and health, and particularly in human medicine [5, 6]. Since the release of the first shotgun metagenome from environmental microbiota samples [7], microbiome research has grown explosively in both data generation and methodological advancements.

Unlike amplicon sequencing, which typically targets specific genetic regions (such as 16S rRNA genes), shotgun metagenomics (also known as short-read, second-generation, next-generation, or high-throughput sequencing) captures all DNA in a sample. It can provide a relatively unbiased view of the complete taxonomic composition and functional potential of microbial communities [8, 9]. Additionally, it enables the recovery of metagenomic assembled genomes (MAGs), which offer insights into microbial diversity with high taxonomic resolution—down to the species level and, in some cases, even the strain level [10]. Therefore, shotgun metagenomic sequencing has become a transformative tool in microbiome research, greatly enhancing our understanding of microbial communities in humans and other environments [11, 12]. Although the development of third-generation long-read sequencing has helped address some limitations of short-read sequencing, its higher error rate and lower throughput mean that most studies still rely on hybrid approaches (combining long-read and short-read data) for error correction [13–16]. Thus, second-generation metagenomic sequencing remains crucial in microbiome research.

After generating shotgun metagenomic data from a microbial community of interest, the typical analysis begins with quality control and removal of host DNA, followed by profiling taxonomic, functional and genomic features of the microbiome [8, 17], to gain insights into the underlying biology. Various tools, R packages, and pipelines have been developed to accomplish each stage of the analysis. For example, fastp [18] or Trimmomatic [19] are used for basic quality control, Bowtie2 [20] or KneadData (<https://huttenhower.sph.harvard.edu/kneaddata>) are employed for removing reads that map to the host reference database, Kraken2 [21] and Bracken [22] are used for taxonomic profiling, HUMAnN3 [23] are utilized for generating gene families and MetaCyc [24] pathways, MEGAHIT [25] or MetaSPAdes [26] serve as metagenome assemblers, and MetaWRAP [27] is used for extracting draft metagenome-assembled genomes (MAGs) from metagenomes, among

others. Also, plenty of R packages for exploring microbial community diversity, structure, and potential functions [10, 28]. Each of these tools or R packages has multiple parameters or reference databases that can influence the final output and downstream analysis, presenting a challenge to comparability and reproducibility across studies [29, 30]. This underscores the importance of developing data workflows to maintain standardized data, ensure consistent results, and facilitate the reproducibility of findings across different studies.

Recently, several pipelines and workflows have been developed for microbiome research, although most are tailored to specific analyses and lack a holistic approach for broader data exploration. Additionally, many of these workflows do not offer detailed reports, flexible parameter adjustments, or meaningful downstream visualization options, which are essential for interpretation and reproducibility. For instance, Sunbeam focuses on some basic analyses, from data preprocessing and read-based analysis to assembly-based analysis [30]. The MetaSAMS workflow is primarily designed for taxonomic and functional analysis (read-based analysis) [31], while MEGAnnotator2 [15], SqueezeMeta [32], and ATLAS [33] workflows were sequentially developed to reconstruct MAGs from metagenomic sequences (assembly-based and binning). METABOLIC specifically profiles the functional traits of MAGs [34] and related workflows. In addition, online platforms, such as MicrobiomeAnalyst 2.0 [35] and Wekemo Bioincloud [36], offer useful tools for statistical meta-analysis and shotgun data matrix profiling. However, these platforms struggle to process large datasets online, which is especially limiting for shotgun metagenomics, where data sizes often range from hundreds of gigabytes (GB) to several terabytes (TB).

Here, we introduce EasyMetagenome, an easily deployable and configurable pipeline that supports read-based, assembly-based, and binning analyses, as well as in-depth prokaryote genome analysis, generating publication-ready visualizations. EasyMetagenome includes four main components: software and database installation, data analysis pipeline, statistics and visualization, and a dedicated questions and answers section. Nearly all steps are configurable, with reasonable default settings that allow for rapid deployment without extensive parameter tuning. Additionally, detailed parameter explanations are provided, making it adaptable to various data scenarios.

RESULT

Overview of EasyMetagenome

EasyMetagenome consists of four main components, software and database installation, the data analysis

pipeline, statistics and visualization, and a section with dedicated questions and answers. Each component is a separate shell script, with a built-in question and answer feature to help users understand and modify the process (Figure 1). It currently offers a portable workflow for data preprocessing, read-based analysis, assembly-based analysis, and binning, utilizing a range of state-of-the-art software and data resources in the field.

After following our installation instructions for the necessary software and databases, users can input paired-end metagenomic sequencing files to start the analysis. EasyMetagenome is a fully integrated, end-to-end pipeline that processes from raw reads to final data tables and publication-ready figures (Figure 1). First, EasyMetagenome

produces quality control and host genome decontamination of raw sequences/reads. Then, the clean reads can be analyzed for taxonomic composition and functional annotation through read-based analysis, or undergo assembly, gene prediction, functional gene annotation, and taxonomic classification in assembly-based analysis. In addition, the assembly is subsequently binned and refined using meta-WRAP, followed by dereplication and evaluation of recovered MAGs. For the statistics and visualization part, we have also provided various corresponding R configuration files, making it easy for users to generate and customize charts. As of November 27, 2024, EasyMetagenome has received 351 stars and 225 forks on Github (<https://github.com/YongxinLiu/EasyMetagenome>).

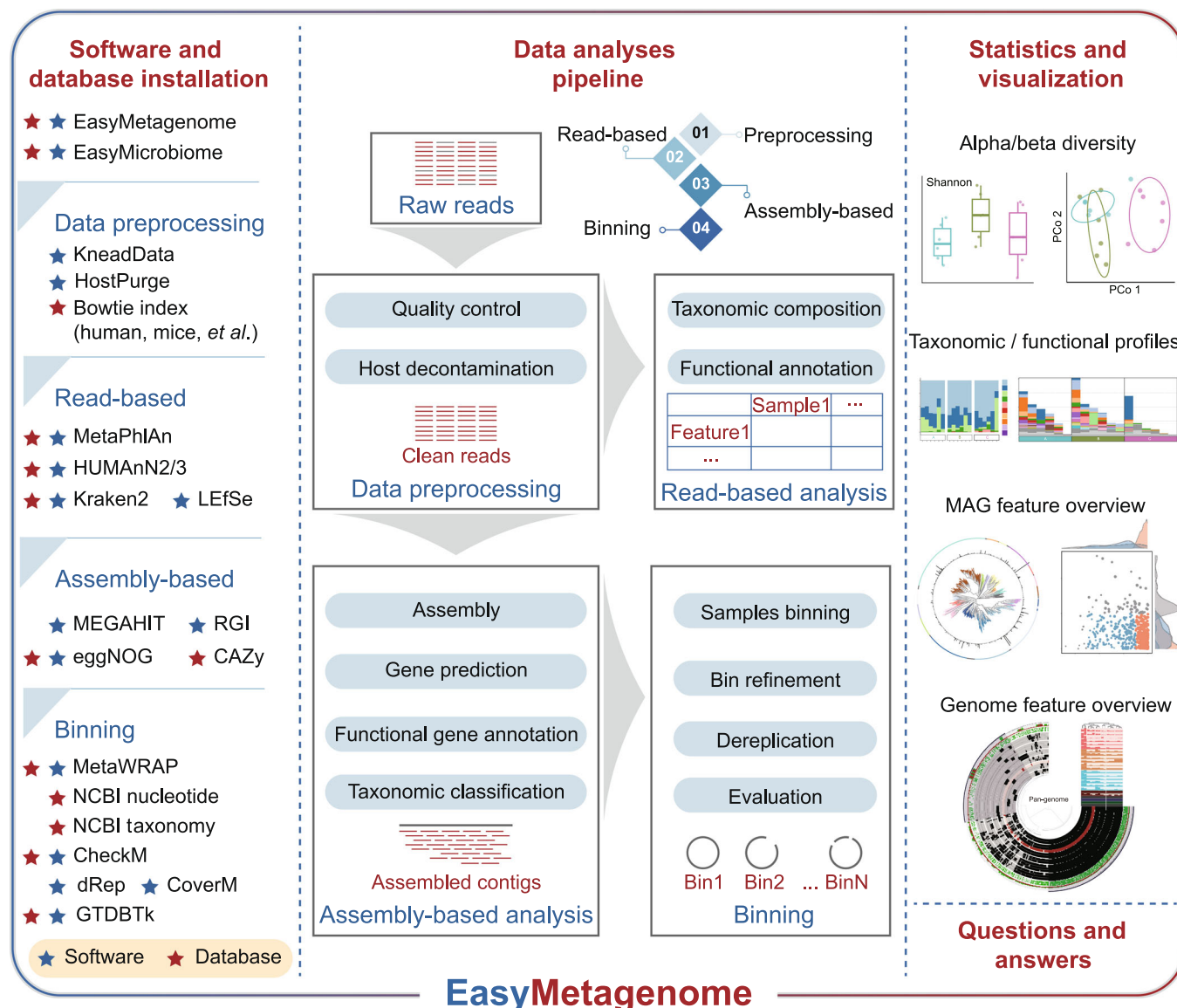


FIGURE 1 EasyMetagenome workflow. The workflow contains four main components, including software and database installation, data analysis pipeline, statistics and visualization, and the section for questions and answers.

FIGURE 2 Software, databases, and R packages included in EasyMetagenome. (A) Software used in the EasyMetagenome pipeline, covering data pre-processing, read-based, assembly-based, and binning analyses. (B) Standards or custom databases associated with each software. (C) R packages are utilized for differential analysis, community comparison, biomarker identification, diversity analysis, network analysis, and others.

TABLE 1 Open access tools that can be applied at each stage of metagenomic data.

Tool	Task	Repository
fastp v0.24.0	Short read quality control, filtering, and adapter removal	https://github.com/OpenGene/fastp
KneadData v0.12.0	Removal of host contamination	https://huttenhower.sph.harvard.edu/kneaddata
FastQC v0.12.1	Quality assessment of short reads	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
MetaPhlAn4 v4.0.6	Taxonomic profiling and relative abundance estimation	https://github.com/biobakery/MetaPhlAn
HUMAnN3 v3.7	Functional potential profiling of microbial communities	https://github.com/biobakery/humann
Kraken2 v2.1.3	Taxonomic classification	https://github.com/DerrickWood/kraken2
MEGAHIT v1.2.9	Short read assembly	https://github.com/voutcn/megahit
QUAST v5.0.2	Assembly quality assessment	https://github.com/ablab/quast
MetaProdigal v2.6.3	Gene prediction from assembled contigs	https://github.com/hyattpd/Prodigal
CD-HIT v4.8.1	Clustering of assembled contigs	https://github.com/weizhongli/cdhit
Salmon v1.8.0	Contig coverage estimation	https://github.com/samtools/samtools
eggNOG v5.0	Protein functional annotation	https://github.com/eggnogdb/eggno-mapper
MetaWRAP v1.3.2	Bin refinement and reassembly	https://github.com/bxlab/metaWRAP
dRep v2.6.2	Genome redundancy and representative genome selection	https://github.com/MrOlm/drep
CoverM v0.6.1	Coverage analysis of metagenomic assemblies	https://github.com/wwood/CoverM
GTDB-tk v2.4.0	Taxonomic classification based on the genome taxonomy database	https://github.com/Ecogenomics/GTDBTk
CheckM2 v1.0.1	Quality control of microbial genome bins	https://github.com/chklovski/CheckM2
eggno-mapper v2.1.6	Functional annotation of genes using eggNOG database	https://github.com/eggnogdb/eggno-mapper
DIAMOND v2.0.13	Fast protein sequence alignment	https://github.com/bbuchfink/diamond
RGI v5.2.0	Antibiotic resistance gene identification	https://github.com/arpcard/rgi
Bakta v1.9.4	Bacterial genome annotation	https://github.com/oschwengers/bakta
anvi'o v8	Pan-genomic analysis of microbial genomes	https://github.com/merenlab/anvio
fastANI v1.34	Fast pairwise genome sequence comparison	https://github.com/ParBLISS/FastANI

version of standard database plus with protozoa and fungi (PlusPF16G), a 69 GB version of standard database plus with protozoa and fungi (PlusPF), and a 144 GB version of standard database plus with protozoa, fungi, and plant (PlusPFP). We also provide the guidance on building custom microbial databases (Figure 2B).

After that, we offer a range of R scripts for generating publication-ready figures across various domains (Figure 2C). For difference analysis, we use 21 R packages to detect significant differences and visualize results. Diversity analysis is supported by 23 R packages, addressing alpha diversity, beta diversity, rarefaction, and compositional analysis. Community comparison employs eight key packages for ordination, clustering, and diversity metrics. Biomarker identification involves 22 packages utilizing machine learning and statistical modeling methods.

Network analysis relies on 11 packages to identify and visualize relationships within datasets. Additionally, 18 packages assist with data manipulation, visualization, and annotation, ensuring a comprehensive analytical approach.

Data analysis pipeline design

To guide the analysis process, we provide a clear working directory and file structure within our pipeline (Figure 3). The first step is to preprocess raw sequence data, which is essential for removing low-quality reads and contamination from host-associated sequences, ensuring cleaner and more reliable data for downstream analysis. The second step involves profiling the taxonomy and metabolic potential using a read-based approach, which is crucial for

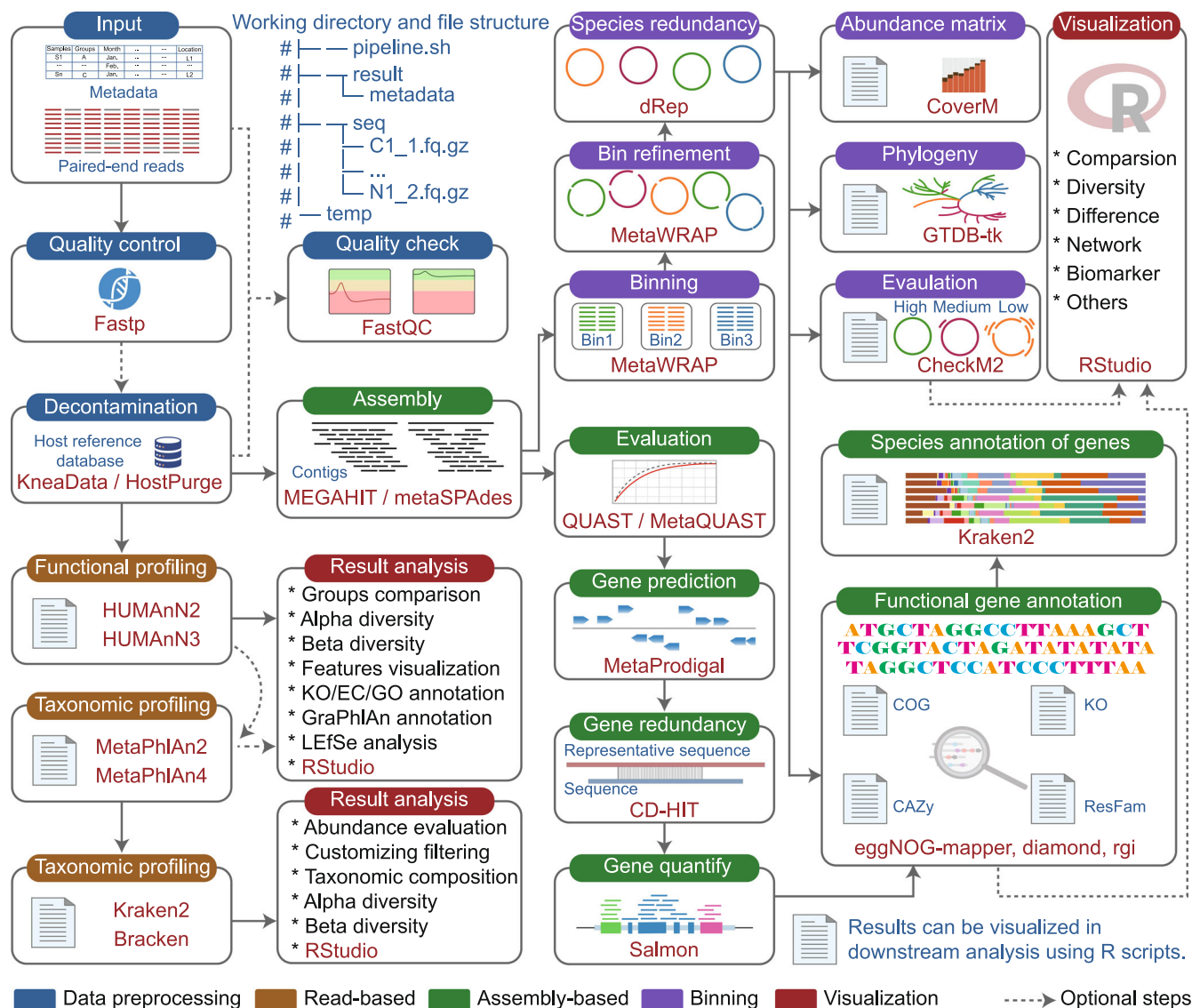


FIGURE 3 Summary of the EasyMetagenome analysis pipeline. EasyMetagenome pipeline includes data preprocessing for read-based, assembly-based, and binning analyses, with files available for data visualization. Dotted lines denote optional steps, and software are highlighted in red.

identifying the microbial composition and functional capabilities, providing valuable insights into the ecosystem being studied. In this step, Kraken2 and HUMAnN2 or HUMAnN3 are used for microbial composition and functional analysis, respectively. This method is straightforward and widely used, while it may overlook many previously unknown genes that are not annotated in current databases.

The assembly step is a critical part of the process, where clean reads are assembled into contigs using tools like MEGAHIT or metaSPAdes. MEGAHIT is optimized for quickly assembling large, complex metagenomic datasets with minimal computational resources, while metaSPAdes produces longer and strain-level contigs but requires more computational resources. Once the contigs

are assembled, they can undergo assembly evaluation, gene prediction, removal of redundant genes, and gene abundance quantification. Metagenomic datasets typically contain millions of genes. Then these genes can be combined into functional categories, such as KEGG Orthology (KO), Clusters of Orthologous Genes (COG), Carbohydrate Active Enzyme (CAZy), Antibiotic Resistance Genes (ARG).

In the metagenome binning step, contigs are clustered into genome bins. This can be done individually for each set of assembled contigs or by pooling contigs from multiple samples and mapping each sample back to a shared catalog of contigs. Tools like metaWRAP facilitate this process, generating MAGs through contigs binning. In metagenomic studies, binning enables the reconstruction of both

known and unknown genomes, providing a comprehensive description of the microbial community and serving as a foundation for further organismal analysis.

Taxonomic and functional composition of read-based analysis

Here, we present the results with read-based analysis, showing the diversity of taxonomic and functional

composition using data from centenarian's gut microbiome studies. Taxonomic composition is assessed using Kraken2 (Figure S1) and MetaPhlAn4 (Figure 4), followed by alpha and beta diversity analyses and functional composition assessment using HUMAnN3 (Figure 4). We provide multiple visualization styles to clarify and enrich the interpretation of these results, offering a detailed, multidimensional perspective on community structure and functional dynamics (Figure 4).

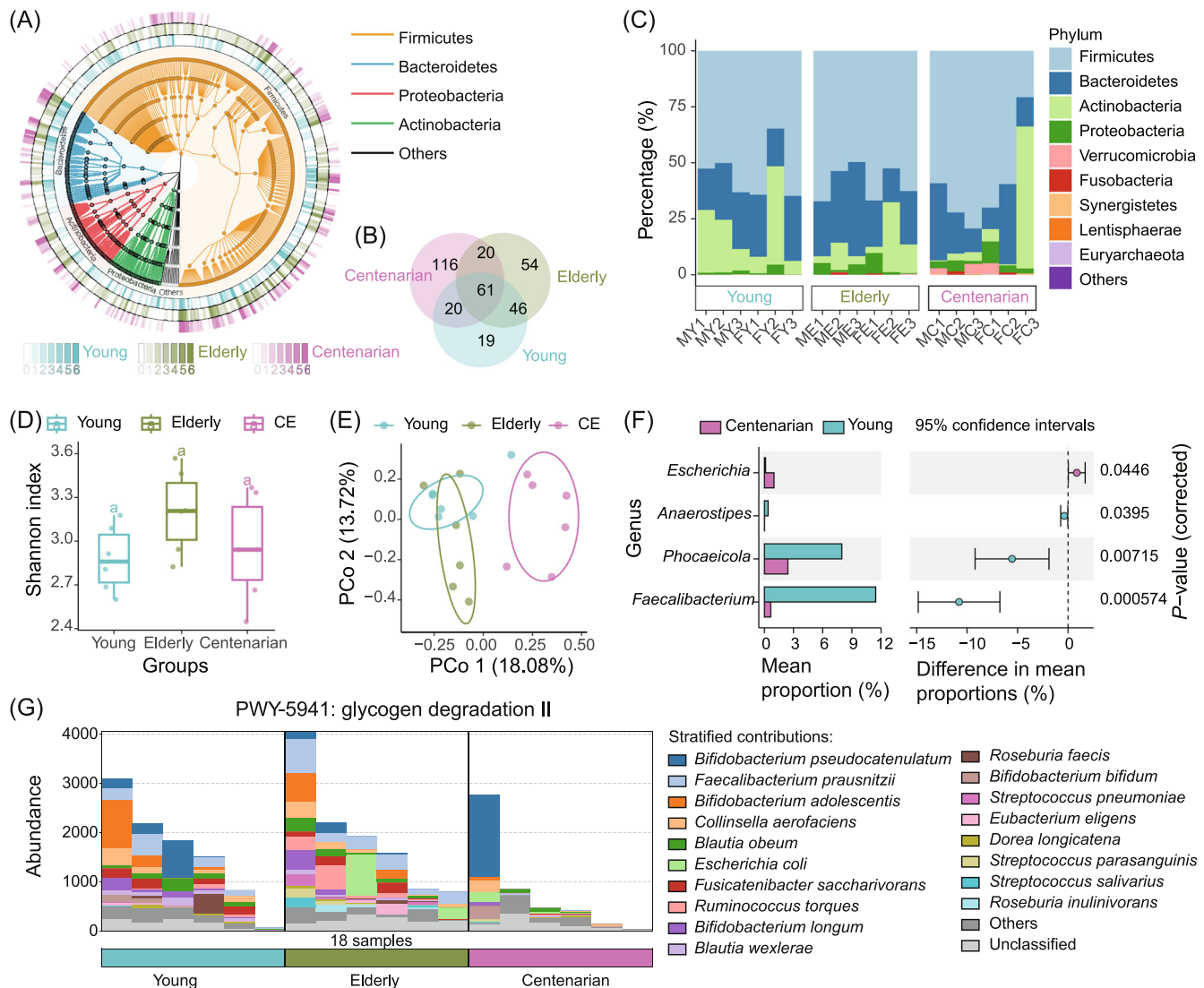


FIGURE 4 Taxonomic and functional composition from read-based analysis using EasyMetagenome in human gut microbiome samples. (A) Microbial taxonomic composition overview across groups. The inner line color represents the Phylum. The outer bar shading represents taxa presence or absence across six representative samples per group. (B) Venn diagram showing the unique and shared taxa across three groups. (C) Taxonomic composition across groups and samples in phylum-level. (D) Shannon diversity indices across groups, with letters indicating significant group differences ($p \leq 0.05$, ANOVA, Tukey's HSD). (E) Principal coordinates analysis (PCoA) based on Bray-Curtis dissimilarity ($p < 0.001$, PERMANOVA with ADONIS test). (F) STAMP plot showing genus-level differences between young and centenarian groups. (G) Abundance of the functional pathway PWY-5941 across groups and samples, including stratified contributions from specific microbial species. (A–F) are derived from species annotations with MetaPhlAn4. (G) represents functional potential analysis via HUMAnN3. The raw data from the Bioproject (PRJNA675598) of NCBI with accession number in Table S1.

For taxonomic annotation of metagenomic data, both Kraken2 and MetaPhlAn4 offer distinct yet complementary methods. Kraken2 classifies taxa using a *k*-mer-based approach compatible with databases such as PlusPF16G, PlusPF, and PlusPFP (Figure 2B). To refine taxonomic abundance estimates, Bracken builds on Kraken2's results by redistributing reads probabilistically. MetaPhlAn4, alternatively, achieves taxonomic profiling using a clade-specific marker gene catalog (e.g., mpa, mpa_bt2), offering a focused, gene-centered approach to taxonomic annotation. For functional potential analysis, HUMAnN3 maps functional genes and pathways using resources like UniRef, ChocoPhlAn, and a set of utility mapping files.

To visualize species and functional composition, we profile bacterial distributions at the phylum level across centenarian, elderly, and young groups, highlighting species presence and relative abundance within each sample (Figure 4A). Venn diagrams illustrate shared and unique taxa among the groups (Figure 4B), while stacked bar charts compare taxa distributions between and within groups (Figure 4C). We calculate alpha diversity metrics, including alpha diversity indices (Chao1, ACE, Shannon, Simpson, Inverse Simpson), and beta diversity metrics (Bray-Curtis, Euclidean, Jaccard, Manhattan) across samples and groups (Figure 4D,E). To analyze species differences between groups, STAMP plot is used, allowing groupwise comparisons with bar plots to illustrate the abundance of taxonomic unit or functional pathways (Figure 4F). Stratified contributions of these pathways by microbial species further illuminate community composition and function (Figure 4G). And the result of Kraken2 has been also displayed (Figure S1A,B).

Functional annotation of genes from assembled contigs and comparative analysis of MAGs

Our assembly-based analysis enables the functional annotation of gene sequences within assembled contigs using a range of comprehensive databases, including Gene Ontologies (GOs), level-4 Enzyme Commission categories (ECs), Clusters of Orthologous Genes (COGs), carbohydrate-active enzymes (CAZy), and Kyoto Encyclopedia of Genes and Genomes (KEGG) modules (Figure 5A, Figure S2A). This pipeline provides several key functionalities. First, it displays the distribution of protein functional annotations across categories and visualizes their overlap (Figure 5A). It also allows for the comparison of functional module abundance and diversity, including COGs, KOs, and CAZy, across distinct groups to explore microbiome functional variations (Figure S2A). Additionally, annotated functional

genes are categorized into hierarchical levels (like KEGG), to show a structured distribution of functional roles across these levels (Figure 5B).

Following the binning pipeline, we can generate quality evaluations, species annotations, and functional profiles of MAGs. The binned MAGs can then be analyzed for mean coverage, phylogenetic inference, and species distribution across various taxonomic levels, displaying mean coverage across all samples (Figure S2B) or within specific groups (Figure 5C). The phyla information of MAGs is summarized with GTDB-tk (Figure 5D, Figure S2C), and the completeness and contamination of MAGs are assessed in the CheckM (Figure S3). The high-quality MAGs, with completeness $\geq 90\%$ and contamination $\leq 5\%$, can be selected for further analysis to assess taxonomy, functional profiles, and the presence of CAZy (Figure 5E) or antibiotics resistance genes (Figure S2D).

Characterization of *Alistipes putredinis* and pan-genome analysis on MAGs

Our pipeline provides detailed functional annotation and analysis of isolates or MAGs. In this study, we focused on *Alistipes putredinis*, a prevalent species in the human gut microbiota, to gain in-depth functional insights into this organism. The MAG of *Alistipes putredinis* was primarily identified in the centenarian group, showing 96.23% completeness, 0.67% contamination, a genome size of 2.08 Mb, and a GC content of 55% (Figure 6A).

To examine the shared and unique functional gene content of *Alistipes putredinis* MAG in comparison with 14 related *Alistipes putredinis* genomes, we conducted a pan-genomic analysis (Figure 6B). This analysis identified a total of 3958 genes across the *Alistipes putredinis* pan-genome, encompassing 14 publicly available *Alistipes putredinis* genomes and *Alistipes putredinis* MAG annotated from our 18 metagenome samples. Based on gene cluster frequency, these 15 genomes were divided into 6 clades, revealing 1322 core genes (33.4%) and 1028 singleton genes (26.0%) unique to individual genomes. Clustering the 15 genomes according to gene cluster frequency (both number and identity) further segregated them into two major groups and 6 clades. Each genome contained between 1821 and 2323 genes (top right bar chart), with core genes comprising 57%–73% of the total genes in each genome (Figure 6B).

Then we calculated the average nucleotide identity (ANI) among the 15 *Alistipes putredinis*, and found that *Alistipes putredinis* MAG shares a 97.22%–99.11% similarity with other published genomes (Figure 6C). Since shared gene content between genomes effectively predicts phylogenetic relationships, we used the distribution of genes to infer the relationships among these genomes.

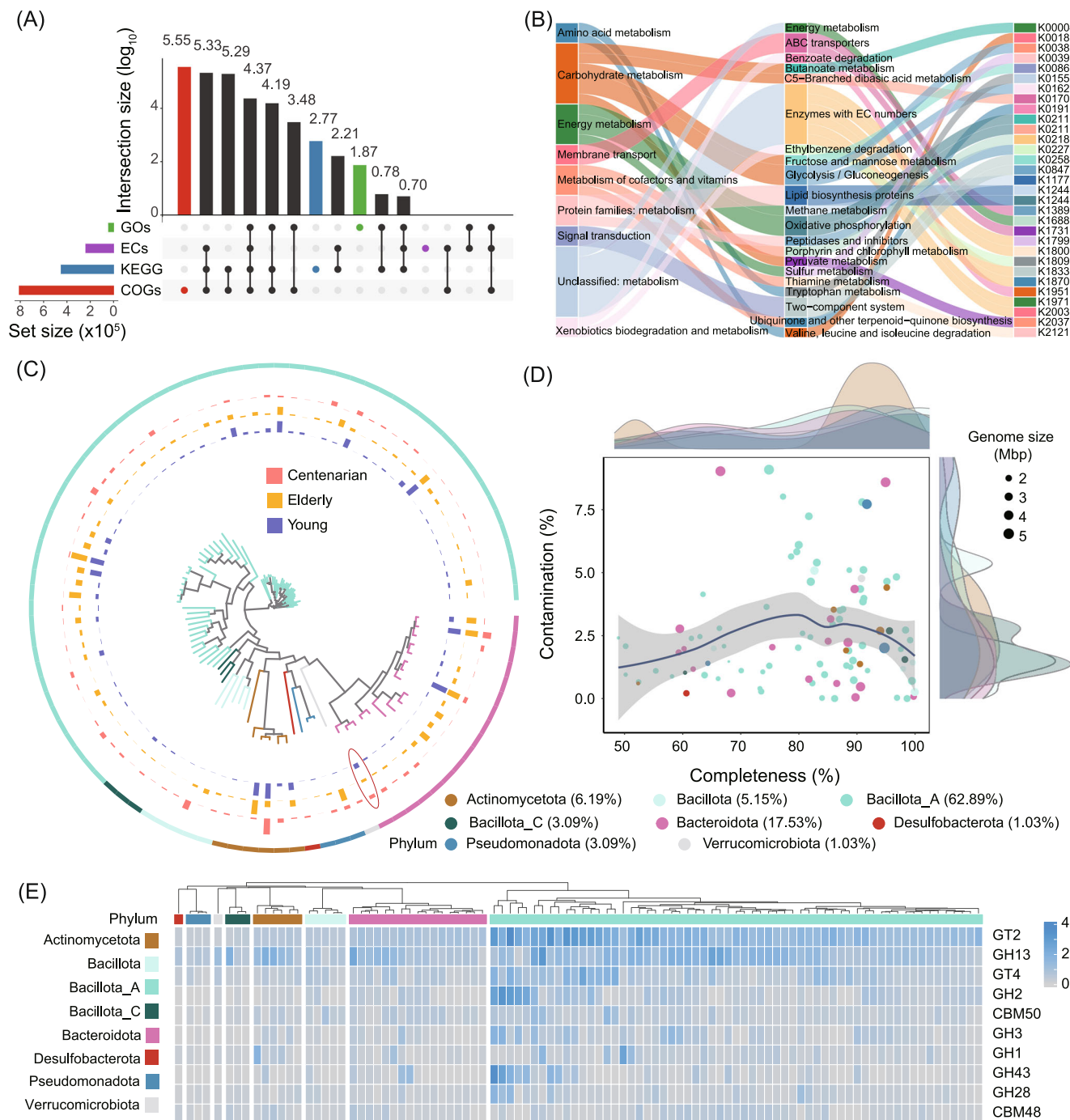


FIGURE 5 Functional annotation of assembled contigs and comparative analysis of MAGs in human gut microbiome samples.

(A) Functional annotations of assembled contigs across four categories, showing the number of annotated proteins per category and their overlaps. Vertical bars represent unique (colored) or shared (black) proteins between categories; horizontal bars display the total protein count with functional annotations for each category. (B) KEGG classification of annotated functional genes at primary, secondary, and tertiary levels, visualized in a flowchart of functional distribution. (C) Phylogenetic analysis and distribution of metagenome-assembled genomes (MAGs) across different groups, highlighting evolutionary relationships of representative MAGs. The tree colored by phylum. The bar shows mean relative abundance of MAGs in different groups calculated by CoverM. (D) The distribution of completeness and contamination on MAGs and the color of point represents phylum. (E) CAZy functional gene annotations within MAGs, illustrating the distribution of key carbohydrate-active genes across MAGs.

Compared with dendrograms generated using 40 core genes, the clustering of clades showed a notable difference (Figure 6D), suggesting that incorporating the whole genomic content, rather than only core genes, may enhance accuracy in determining relationships among different genomes. Our results may also suggest that

using the whole genomic content instead of part of core genes might be advantageous to infer the relationship between different genomes. Finally, we also annotated the *Alistipes putredinis* genomes with functional data from the COG database. We selected 24 COG functions with high AG rates to demonstrate functional diversity

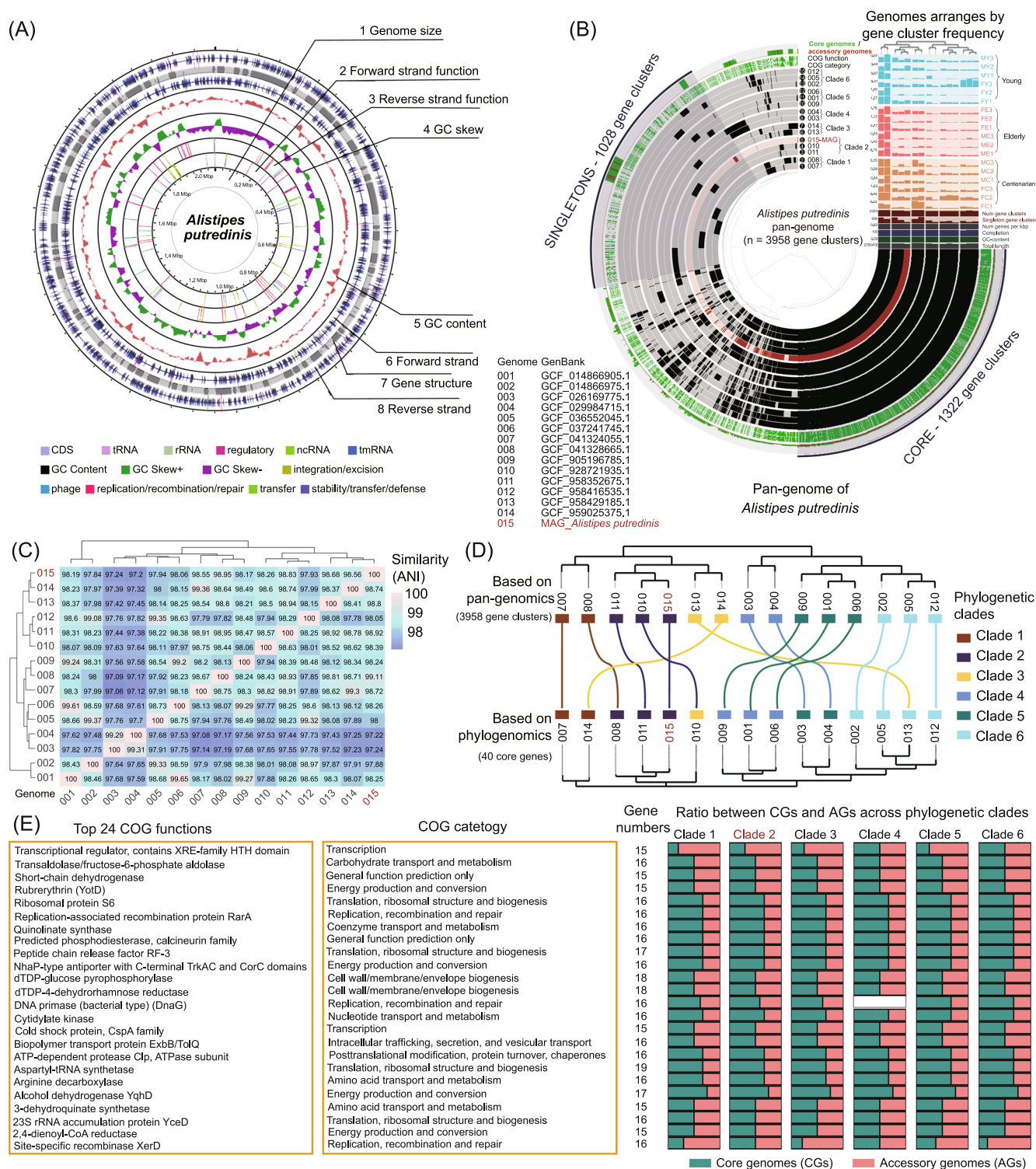


FIGURE 6 (See caption on next page.)

within the pan-genome. This allowed us to observe variations in CG/AG ratios across different clades for each selected COG function, providing insights into functional divergence among clades (Figure 6E).

DISCUSSION

In EasyMetagenome, we provide several options for handling diverse shotgun metagenomic data, suitable for various scenarios. Facing the challenge of host contamination in metagenomic data [37–39], we developed HostPurge (<https://github.com/HaoLuo-leo/HostPurge>), a new tool that integrates popular alignment and *k*-mer based methods, designed to enhance the efficiency, reliability, and comparability of large-scale data mining. However, existing tools are limited when handling short-read high-throughput sequencing data without specific host reference genomes [39]. Artificial intelligence algorithms and third-generation long-read sequencing may help overcome this challenge, serving as a key to potentially establishing high-quality MAGs databases. This highlights the urgent need to incorporate new algorithms and workflows tailored to diverse metagenomic data into our pipeline.

EasyMetagenome is primarily designed to run on Linux-based systems, leveraging its robust support for high-performance computational tools commonly used in metagenomic analysis. However, this focus may limit accessibility for users reliant on other operating systems, such as Windows or macOS. While compatibility can be

addressed through virtualization or containerized solutions (e.g., Docker), these approaches may introduce additional complexity for nonspecialist users. Also, the scalability of the pipeline may be challenged when applied to datasets of exceptionally large size or complexity. Additionally, computational compatibility could vary depending on the specific hardware and software environments, particularly for users with limited access to high-performance computing resources. Future iterations of EasyMetagenome could explore enhanced cross-platform support to broaden its usability and accessibility.

The sequencing data produced by these third-generation platforms differ qualitatively from second-generation sequencing, necessitating customized analysis tools [40]. Although some bioinformatics software and methods have been developed for long-read sequencing data, such as error correction methods [41] or hybrid assembly correction methods [42], the metagenomic analysis on third-generation sequencing data is still in its early stages. Short-read-based algorithms for third-generation metagenomic analysis are often not suitable for long-read datasets under default parameters [43], which may lead to higher error rates when interpreting long sequencing reads. Therefore, there is an urgent need to establish a “gold standard” pipeline for third-generation metagenomic data analysis and further enhance the comparability and reproducibility of microbiome exploration.

In the future, EasyMetagenome will focus on downstream data mining, using new methods to deepen our understanding of metagenomic data. For instance, network

FIGURE 6 Metagenomic assembly genome digging in example human gut microbiome. (A) Genome map of *Alistipes putredinis* (ncRNA, COG annotation, GC content, and GC skew). From inner to outer, the 1st circle shows the genome size (1 Genome size); the 2nd circle shows the COG function of the forward strand gene, and each color represents a function classification (2 Forward strand function); the 3rd circle shows the function of the reverse strand gene (3 Reverse strand function); the 4th circle shows the GC skew $((G - C)/(G + C))$, green > 0 , purple < 0 (4 GC skew); the 5th circle shows the GC content (5 GC content); the 6th circle shows the coding regions and structural RNA genes of the forward strand (6 Forward strand); the 7th circle shows the gene structure (7 Gene structure); the 8th circle shows the coding regions and structural RNA genes of the reverse strand (8 Reverse strand); (B) Pan-genomic analysis of *Alistipes putredinis* genomes. Pan-genome graph generated with Anvi'o, where the gene clusters (radial bars) are ordered by gene cluster frequency according to the organization of genes in the genome of *Alistipes putredinis*. And we divided 15 genomes of *Alistipes putredinis* used here into six clades. The circles show, from inner to outer, for the detected (black areas) or undetected (grey areas) of genes in each genome (circles 1–15 with 6 clades arranged); we also mapping the gene clusters to the cluster of COG database, and illustrated COG category (circle 16) and COG functions annotated (circle 17) for gene clusters in which at least one gene was functionally annotated. In the outermost circle describes the ratio of core genomes and accessory genomes. Extending off the pan-genome are bar plots showing the relevant properties for each genome. Above the genome property summaries, genome's median coverage across all centenarian (color in light brown), elderly (color in red), and young (color in cyan) pan-genome samples are shown in bar graph separately. (C) Average nucleotide identity (ANI) mean heatmap generated using the fastANI tool depicts the genomic relatedness resulting from a multi-genome comparison among strains of *Alistipes putredinis*. (D) Organization of *Alistipes putredinis* genomes based on gene clusters compared to phylogenomics. The cluster tree on the top showed the clustering of 15 genomes of *Alistipes putredinis* based on the 3958 gene clusters recovered from the pan-genomic analysis. The bottom tree clustering the same genomes based on phylogenomics using 40 core genes. Colors showed six phylogenetic clades based on pan-genomics. (E) The 24 most accessory functions identified in *Alistipes putredinis* isolates defined by high ratio of accessory genomes. The left two panels showed the COG functions and COG categories. The stack bar on the right showed an example ratio between core genomes (green) and accessory genomes (red) in different phylogenetic clades.

analysis has proven effective for exploring microbial interactions [44, 45], and identifying health-relevant gut microbiota by revealing stable and functional genome relationships. A recent model, “two competing guilds,” uses abundance network analysis to identify core microbiome features as health indicators [44]. Another approach, generalized microbe phenotype triangulation, hypothesizes that in most paired expressions, microbial abundance strongly correlates (positively or negatively) with disease severity, which can assist in species and functional predictions. This method goes beyond traditional differential abundance approaches by targeting core microbiota directly associated with disease mechanisms and has been successfully applied in diversity research [46, 47]. With these advances, EasyMetagenome aims to enhance our understanding of microbial functions, interactions, and coevolution with hosts.

CONCLUSION

As the demand for reproducibility in computational analysis of shotgun metagenomic data continues to increase, establishing a standardized bioinformatics workflow has become essential. To address this, we developed EasyMetagenome, a user-friendly, flexible pipeline that supports comprehensive data exploration and provides clear parameter explanations at each step, generating publication-ready results. In this study, we specifically present two examples to demonstrate the flexibility, usability, and reliability of our pipeline. This holistic pipeline to shotgun metagenomic analysis in microbiome research allows researchers to focus on resolving key questions in biological science rather than on navigating complex computational challenges.

Looking forward, we aim to address issues related to host contamination that impact data accuracy and processing efficiency. We will also integrate workflows optimized for third-generation sequencing data and incorporate emerging algorithms, such as deep learning for cleaner microbiome reads, network analysis for interaction studies, and GMPT for key biomarker identification. Our goal is to enhance explainability and improve efficiency in microbiome data analysis.

METHODS

Workflow overview and software of EasyMetagenome

Our pipeline starts with Illumina raw sequencing reads, followed by data preprocessing, read-based analysis,

assembly-based analysis, and binning. The resulting data is then statistically analyzed and visualized using R scripts.

Briefly, quality assessment can be performed using FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). Illumina raw paired-end sequences are quality-controlled using the fastp [18] software. The host sequence is then removed using KneadData (<https://huttenhower.sph.harvard.edu/kneaddata>) to generate the clean microbiome reads. After that, clean reads are rechecked for quality with FastQC. Taxonomic classification is carried out using Kraken2 [21], and species abundance is estimated. Gene family and pathway relative abundances are determined with HUMAnN2 or HUMAnN3 [23].

Concurrently, clean reads are assembled with either MEGAHIT [25] or metaSPAdes [26]. Subsequently, the assembled contigs are analyzed in two main streams. For assembly-based analysis, contig quality is evaluated using QUAST [48]. Gene prediction is performed with Meta-Prodigal [49], gene redundancy is assessed using CD-HIT [50], and gene quantification is carried out with Salmon [51]. The Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthology (KO) profiles [52], Carbohydrate Active Enzyme (CAZy) [53], and Clusters of Orthologous Groups of proteins (COG) [54] database are used for microbiome functional annotation using eggNOG [55]. The antibiotic resistance genes were also detected using the ResFams database [56] with RGI [57]. And Kraken2 [21] is used for species annotation of genes.

For binning, MAGs are retrieved using the “binning” and “bin_refinement” modules in MetaWRAP [25]. Species redundancy is addressed with dRep [58], abundance matrices are created with CoverM, phylogeny is determined using GTDB-tk [59], and completeness and contamination of the bins are estimated with CheckM2 [60]. Functional gene annotation of MAGs is again performed using eggnog-mapper [61], DIAMOND [62], and RGI [57].

Data set selection and quality control for EasyMetagenome analysis

To showcase the functionality of EasyMetagenome, we analyzed datasets from previously published data, including a set of human gut microbiome samples [63] and a set of environmental microbiome samples [64], highlighting the pipeline's versatility and visualization capabilities across diverse sample types. For human gut samples, it contains three groups of Young (23–47 years old), Elderly (85–89 years old), and Centenarian (100+ years old), each represented by six samples with an equal number of male and female people (Table S1). For environmental samples, we selected four composting

phases (Cooling, Mature, Mesophilic, and Thermophilic), each with four samples for replication (Table S2). Detailed accession numbers for these datasets are available in Tables S1 and S2 on NCBI.

All raw data were quality-controlled with fastp v0.24.0 [18], followed by KneadData v0.12.0 to isolate high-quality microbial reads [65], with human reads mapped to GRCh37 human reference genome and removed using Bowtie2 [20] (v2.5.1; parameters: --very-sensitive --dovetail). The cleaned reads were then assessed for quality using FastQC v0.12.1. The clean microbial reads were analyzed using three complementary approaches—read-based, assembly-based, and binning analysis—demonstrating the EasyMetagenome pipeline's ability to generate publication-quality visualizations.

Read-based analysis of EasyMetagenome in published metagenomic data

For read-based analysis, high-quality microbial reads were used for species-level community profiling, with relative abundances determined using MetaPhlAn4 v4.0.6 [66], applying against the mpa_vOct22_CHOCOPhlanSGB_202212 database under default parameters to capture all taxonomic levels. Additionally, functional potential profiling is performed using HUMAnN3 v3.7 [23]. Various EasyMetagenome scripts and dependent MicrobiomeStatPlot [67] were then applied to visualize these results: “metaphlan_hclust_heatmap.R” generated heatmaps showing taxonomic distribution at different levels across groups, while “graphlan_plot54.r” illustrated the overall taxonomic structures within groups. Additionally, “otu_mean.R” calculated the mean taxonomy between groups, “sp_vennDiagram.sh” created a Venn diagram, and “metaphlan_boxplot.R” produced boxplots displaying taxonomic levels across groups. For alpha diversity analysis, we used “otutab_rare.R” to calculate indices such as Richness, Chao1, ACE, Shannon, Simpson, and Inverse Simpson, and “alpha_boxplot.R” to generate boxplots for these metrics from EasyAmplicon [68, 69] ANOVA followed by Tukey's HSD test was applied to assess the significance of differences and determine the *p*-value. For beta diversity analysis, “usearch10” was used to calculate Bray–Curtis, Euclidean, Jaccard, and Manhattan metrics, and the results were visualized as PCoA plots using “beta_pcoa.R”. PERMANOVA with the ADONIS test was employed to evaluate the statistical significance and compute the *p*-value. The “tax_stackplot.R” script was used to display a stackplot of taxonomic composition across groups, while “compare_stamp.R” allowed for taxonomic comparisons between two groups. Additionally, “sp_pheatmap.sh” calculated pathway abundance differences, and

“humann_barplot” of HUMAnN3 v3.7 [23] was used to visualize taxonomic composition within a specific pathway across groups.

Taxonomic classification was performed using Kraken2 v2.1.3 [21] with default settings and a pre-built PlusPF database, while species abundance was estimated with Bracken v2.7 [22] to refine species-level abundance estimates. Both Kraken2 and MetaPhlAn4 can be used for taxonomic profiling, but they employ distinct, yet complementary, methods and reference databases. Kraken2 utilizes a *k*-mer-based classification approach, comparing reads against a standard or user-constructed database. In contrast, MetaPhlAn performs classification by aligning query reads to a database of clade-specific marker genes [22]. To leverage the strengths of each, EasyMetagenome includes visualizations generated from both methods. For Kraken2 data, we converted its output format to match MetaPhlAn's structure, applying the same visualization techniques described above to produce taxonomic annotation images based on Kraken2 results.

Assembly-based analysis of EasyMetagenome in published metagenomic data

For assembly-based analysis, high-quality microbial reads were *de novo* assembled separately using MEGAHIT v1.2.9 [25], with assembly metrics subsequently assessed using QUAST v5.0.2 [48]. MetaProdigal v2.6.3 [49] was used for gene prediction, and assembled contigs were clustered using CD-HIT v4.8.1 [48] with 95% identity and 90% coverage (-aS 0.9 -c 0.95). Then Salmon v1.8.0 [51] was employed to measure the abundance of contigs in each sample. Protein functions, including KEGG Ortholog (KO) [52], CAZy [53], and COGs [54] were annotated using eggNOG-mapper based on eggNOG v5.0 clusters [55]. To generate a summarized abundance table, “format_dbcan2list.pl” was used in combination with “summarizeAbundance.py”. Subsequently, R was used to visualize the functional annotations of genes across four categories, classify annotated functional genes at various hierarchical levels, and compare functional module levels of the microbiome across different groups.

Binning the metagenome-assembled genomes in published metagenomic data with the EasyMetagenome pipeline

For binning, MAGs were binned from contigs assembled with MEGAHIT v1.2.9 [25], using the “binning” and

“bin_refinement” modules in MetaWRAP v1.3.2 [25]. In detail, contigs were clustered into metagenomic bins using the “binning” module (--metabat2 --maxbin2), followed by refinement with the “bin_refinement” module (-c 50 -x 10). After that, the resulting refined bins were dereplicated using dRep v2.6.2 [58], with the parameters “-sa 0.95 -nc 0.3 -p 16 -comp 50 -con 10.” The abundance matrices of the dereplicated genomes across various samples were then evaluated using CoverM v0.6.1.

Taxonomic classification of the refined bins was performed using the “gtdbtk classify_wf” module from the GTDB-tk v2.4.0 (r214) database, and their phylogenetic relationships were inferred with the “gydbtk infer” module in GTDB-tk v2.4.0 [59]. The completeness and contamination of the MAGs were estimated with CheckM2 v1.0.1 [60], with high-quality MAGs selected based on the following thresholds: High-quality: $\geq 90\%$ completeness and $\leq 5\%$ contamination (for some environmental samples, adjustments may be required based on actual conditions); Medium-quality: $> 50\%$ completeness and $\leq 5\%$ contamination; Low-quality: $< 50\%$ completeness. Subsequently, functional annotations of the high-quality MAGs, including KEGG Ortholog (KO), CAZy, and COGs, were performed using eggno-mapper v2.1.6 [61], DIAMOND v2.0.13 [62], and RGI v5.2.0 [57]. Finally, R was used to visualize the genome quality of the MAGs, perform phylogenetic analysis, and display the distribution of MAGs, as well as the functional gene annotations within them.

Genome structures and pan-genomic analysis of *Alistipes putredinis*

After obtaining the MAG annotation results for all samples, we selected *Alistipes putredinis* from human gut metagenomic data (*Alistipes putredinis* MAG) for single bacterial genome annotation and pan-genomic analysis due to its high completeness (96.23%) and low contamination rate (0.67%). For single bacterial genome annotation, we first used Bakta v1.9.4 to annotate the *Alistipes putredinis* MAG [70]. Once the genome annotation results were obtained, we imported the annotated files in “embl” format into Proksee software for further analysis and visualization [71]. We also calculated GC skew to assess the relative amounts of guanine (G) and cytosine (C) in the selected genome, and performed additional genome annotation of *Alistipes putredinis* using mobileOG-db to identify mobile orthologous groups [72].

For pan-genomic analysis, we used anvi'o v8 to analyze the assembled genomes of *Alistipes putredinis* [73], incorporating both the *Alistipes putredinis* MAG and 14 publicly available *Alistipes putredinis* isolates from NCBI

(Table S3). These 15 *Alistipes putredinis* genomes were used to construct a contig database with the “anvi-gen-contigs-database” function in anvi'o v8. Bacterial single-copy gene information was then retrieved using the “anvi-run-hmms” function, and NCBI COG function annotations were obtained using the “anvi-run-ncbi-cogs” function. We mapped short reads from human gut microbiome samples to the constructed contig database and sorted the recruited reads using Bowtie2 [20] and samtools [74], separately. After recruiting metagenomic short reads from different age groups, we used the “anvi-profile” function to process the recruitment files and generate profile databases containing gene coverage and detection information for contig database. These profiles were merged into a single database for each sample, linking the contigs to their respective genomes using the “anvi-import-collection” function. Finally, we computed the pan-genome of *Alistipes putredinis* isolates using the “anvi-gen-genomes-storage” function for genome storage and the “anvi-pan-pangenome” function to generate the *Alistipes putredinis* pan-genome. The pan-genome was then visualized using the “anvi-display-pan” function in the anvi'o interactive interface [75].

After obtaining the pan-genomes of *Alistipes putredinis*, we compared the clustering of *Alistipes putredinis* genomes based on all gene clusters detected in the pan-genomic analysis with clustering based on phylogenomics, using 40 core genes selected from 1,322 core genes identified through the pan-genomic analysis. To further assess genomic similarity, we used fastANI v1.34 [76] to calculate the average nucleotide identity (ANI) between the *Alistipes putredinis* MAG and the other 14 published *Alistipes putredinis* genomes. Additionally, to examine the functional composition of the *Alistipes putredinis* pan-genomes, we displayed the most accessory functions (with unusually high AG ratios) identified across all *Alistipes putredinis* genomes using the COG database [54].

AUTHOR CONTRIBUTIONS

Defeng Bai: Methodology; software; validation; writing—original draft; writing—review and editing. **Tong Chen:** Conceptualization; funding acquisition; methodology; software; validation; writing—original draft; writing—review and editing. **Jiani Xun:** Methodology; software; validation; writing—original draft; writing—review and editing. **Chuang Ma:** Methodology; software; validation; writing—original draft; writing—review and editing. **Hao Luo:** Methodology; software; validation; writing—original draft; writing—review and editing. **Haifei Yang:** Methodology; software; validation; writing—original draft; writing—review and editing. **Chen Cao:** Software; validation; writing—review and editing.

Xiaofeng Cao: Software; validation; writing—review and editing. **Jianzhou Cui:** Software; validation; writing—review and editing. **Yuan-Ping Deng:** Software; validation; writing—review and editing. **Zhaochao Deng:** Software; validation; writing—review and editing. **Wenxin Dong:** Software; validation; writing—review and editing. **Wenxue Dong:** Software; validation; writing—review and editing. **Juan Du:** Software; validation; writing—review and editing. **Qunkai Fang:** Software; validation; writing—review and editing. **Wei Fang:** Software; validation; writing—review and editing. **Yue Fang:** Software; validation; writing—review and editing. **Fangtian Fu:** Software; validation; writing—review and editing. **Min Fu:** Software; validation; writing—review and editing. **Yi-Tian Fu:** Software; validation; writing—review and editing. **He Gao:** Software; validation; writing—review and editing. **Jingping Ge:** Software; validation; writing—review and editing. **Qinglong Gong:** Software; validation; writing—review and editing. **Lunda Gu:** Software; validation; writing—review and editing. **Peng Guo:** Software; validation; writing—review and editing. **Yuhao Guo:** Software; validation; writing—review and editing. **Tang Hai:** Software; validation; writing—review and editing. **Hao Liu:** Software; validation; writing—original draft. **Jie-qiang He:** Software; validation; writing—original draft. **Zi-Yang He:** Software; validation; writing—review and editing. **Huiyu Hou:** Software; validation; writing—review and editing. **Can Huang:** Software; validation; writing—review and editing. **Shuai Ji:** Software; validation; writing—original draft. **ChangHai Jiang:** Software; validation; writing—review and editing. **Gui-Lai Jiang:** Software; validation; writing—review and editing. **Lingjuan Jiang:** Software; validation; writing—review and editing. **Ling N. Jin:** Software; validation; writing—review and editing. **Yuhe Kan:** Software; validation; writing—review and editing. **Da Kang:** Software; validation; writing—review and editing. **Jin Kou:** Software; validation; writing—review and editing. **Ka-Lung LAM:** Software; validation; writing—original draft. **Changchao Li:** Software; validation; writing—original draft. **Chong Li:** Software; validation; writing—review and editing. **Fuyi Li:** Software; validation; writing—review and editing. **Liwei Li:** Software; validation; writing—review and editing. **Miao Li:** Software; validation; writing—review and editing. **Xin Li:** Software; validation; writing—review and editing. **Ye Li:** Software; validation; writing—review and editing. **Zheng-Tao Li:** Software; validation; writing—review and editing. **Jing Liang:** Resources; validation; writing—review and editing. **Yongxin Lin:** Software; validation; writing—review and editing. **Changzhen Liu:** Software; validation; writing—review and editing. **Danni Liu:** Software; validation; writing—

review and editing. **Fengqin Liu:** Software; validation; writing—review and editing. **Jia Liu:** Software; validation; writing—review and editing. **Tianrui Liu:** Software; validation; writing—review and editing. **Tingting Liu:** Software; validation; writing—review and editing. **Xinyuan Liu:** Software; validation; writing—review and editing. **Yaqun Liu:** Software; validation; writing—review and editing. **Bangyan Liu:** Software; validation; writing—original draft. **Minghao Liu:** Software; validation; writing—review and editing. **Wenbo Lou:** Software; validation; writing—review and editing. **Yaning Luan:** Software; validation; writing—review and editing. **Yuanyuan Luo:** Software; validation; writing—review and editing. **Hujie Lv:** Software; validation; writing—review and editing. **Tengfei Ma:** Software; validation; writing—review and editing. **Zongjiong Mai:** Software; validation; writing—review and editing. **Jiayuan Mo:** Software; validation; writing—review and editing. **Dongze Niu:** Software; validation; writing—review and editing. **Zhuo Pan:** Software; validation; writing—review and editing. **Heyuan Qi:** Software; validation; writing—review and editing. **Zhanyao Shi:** Software; validation; writing—review and editing. **Chunjiao Song:** Software; validation; writing—review and editing. **Fuxiang Sun:** Software; validation; writing—review and editing. **Yan Sun:** Software; validation; writing—review and editing. **Sihui Tian:** Software; validation; writing—review and editing. **Xiulin Wan:** Software; validation; writing—review and editing. **Guoliang Wang:** Software; validation; writing—review and editing. **Hongyang Wang:** Software; validation; writing—review and editing. **Hongyu Wang:** Software; validation; writing—review and editing. **Huanhuan Wang:** Software; validation; writing—review and editing. **Jing Wang:** Software; validation; writing—review and editing. **Jun Wang:** Software; validation; writing—review and editing. **Kang Wang:** Software; validation; writing—review and editing. **Leli Wang:** Software; validation; writing—review and editing. **Shao-kun Wang:** Software; validation; writing—review and editing. **Xinlong Wang:** Software; validation; writing—review and editing. **Yao Wang:** Software; validation; writing—review and editing. **Zufei Xiao:** Software; validation; writing—review and editing. **Huichun Xing:** Software; validation; writing—review and editing. **Yifan Xu:** Software; validation; writing—review and editing. **Shu-yan Yan:** Software; validation; writing—review and editing. **Li Yang:** Software; validation; writing—review and editing. **Song Yang:** Software; validation; writing—review and editing. **Yuanming Yang:** Software; validation; writing—review and editing. **Xiaofang Yao:** Software; validation; writing—review and editing. **Salsabeel Yousuf:** Software; validation; writing—review and editing. **Hao Yu:** Software; validation; writing—review and editing. **Yu Lei:** Software; validation; writing—review and editing. **Zhengrong Yuan:** Software; validation; writing—review

and editing. **Meiyin Zeng**: Software; validation; writing—review and editing. **Chunfang Zhang**: Software; validation; writing—review and editing. **Chunge Zhang**: Software; validation; writing—review and editing. **Huimin Zhang**: Software; validation; writing—review and editing. **Jing Zhang**: Software; validation; writing—review and editing. **Na Zhang**: Software; validation; writing—review and editing. **Tianyuan Zhang**: Software; validation; writing—review and editing. **Yi-Bo Zhang**: Software; validation; writing—review and editing. **Yupeng Zhang**: Software; validation; writing—review and editing. **Zheng Zhang**: Software; validation; writing—review and editing. **Mingda Zhou**: Software; validation; writing—review and editing. **Yuanping Zhou**: Software; validation; writing—review and editing. **Chengshuai Zhu**: Software; validation; writing—review and editing. **Lin Zhu**: Software; validation; writing—review and editing. **Yue Zhu**: Software; validation; writing—review and editing. **Zhihao Zhu**: Software; validation; writing—review and editing. **Hongqin Zou**: Software; validation; writing—review and editing. **Anna Zuo**: Software; validation; writing—review and editing. **Wenxuan Dong**: Software; validation; writing—review and editing. **Tao Wen**: Conceptualization; funding acquisition; validation; visualization; writing—original draft; writing—review and editing. **Shifu Chen**: Conceptualization; funding acquisition; validation; visualization; writing—original draft; writing—review and editing. **Guoliang Li**: Conceptualization; funding acquisition; validation; visualization; writing—original draft; writing—review and editing. **Yunyun Gao**: Conceptualization; funding acquisition; validation; visualization; writing—original draft; writing—review and editing. **Yong-Xin Liu**: Conceptualization; funding acquisition; validation; visualization; writing—original draft; writing—review and editing.

AFFILIATIONS

¹Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, Guangdong, China

²State Key Laboratory for Quality Ensurance and Sustainable Use of Dao-di Herbs, National Resource Center for Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, China

³School of Horticulture, Anhui Agricultural University, Hefei, China

⁴College of Life Sciences, Qingdao Agricultural University, Qingdao, China

⁵Key Laboratory for Bio-Electromagnetic Environment and Advanced Medical Theranostics, School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, Jiangsu, China

⁶Center for Water and Ecology, State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing, China

⁷Immunology Translational Research Programme, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

⁸Research Center for Parasites and Vectors, College of Veterinary Medicine, Hunan Agricultural University, Changsha, Hunan, China

⁹Institute of Marine Biology and Pharmacology, Ocean College, Zhejiang University, Zhoushan, Zhejiang, China

¹⁰Agro-Environmental Protection Institute, Ministry of Agriculture and Rural Affairs, Tianjin, China

¹¹Key Laboratory for Molecular Genetic Mechanisms and Intervention Research on High Altitude Disease of Tibet Autonomous Region, School of Medicine, Xizang Minzu University, Xianyang, China

¹²Karolinska Institutet, Department of Microbiology, Tumor and Cell Biology, Stockholm, Sweden

¹³College of Environment, Zhejiang University of Technology, Hangzhou, China

¹⁴College of Environmental and Resource Sciences, Zhejiang Agriculture and Forestry University, Hangzhou, China

¹⁵The College of Forestry, Beijing Forestry University, Beijing, China

¹⁶Department of Bioinformatics, Hangzhou VicrobX Biotech Co., Ltd, Hangzhou, Zhejiang, China

¹⁷Anhui Province Key Laboratory of Integrated Pest Management on Crops, College of Plant Protection, Anhui Agricultural University, Hefei, China

¹⁸Xiangya School of Basic Medicine, Central South University, Changsha, Hunan, China

¹⁹Institute of Microbiology, Guangdong Academy of Sciences, Guangzhou, Guangdong, China

²⁰Engineering Research Center of Agricultural Microbiology Technology, Ministry of Education, School of Life Sciences, Heilongjiang University, Harbin, China

²¹College of Animal Science and Technology, Jilin Agricultural University, Changchun, Jilin, China

²²Sansure Biotech Incorporation, Changsha, Hunan, China

²³School of Food Science and Biology, Hebei University of Science and Technology, Shijiazhuang, Hebei, China

²⁴School of Life Sciences, Shanxi Datong University, Datong, China

²⁵Department of Health & Environmental Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, China

²⁶College of Horticulture, Northwest A&F University, Yangling, Shaanxi, China

²⁷School of Agriculture, Food and Ecosystem Sciences, Faculty of Science, The University of Melbourne, VIC, Australia

²⁸Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa-shi, Chiba, Japan

²⁹Institute of Biotechnology, Helsinki Institute of Life Science, University of Helsinki, Helsinki, Finland

³⁰Shanghai Maritime University, Shanghai, China

³¹Suzhou Medical College, Soochow University, Suzhou, Jiangsu, China

³²Biomarker Discovery and Validation Facility, Institute of Clinical Medicine, Peking Union Medical College Hospital, Beijing, China

³³Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong, China

³⁴College of Biology and Oceanography, Weifang University, Weifang, Shandong, China

³⁵College of Environmental Science and Engineering, Beijing University of Technology, Beijing, China

- ³⁶College of Environmental and Municipal Engineering, Lanzhou Jiaotong University, Lanzhou, China
- ³⁷School of Life Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong, China
- ³⁸Department of Renewable Resources, University of Alberta, Edmonton, Alberta, Canada
- ³⁹School of Geographical Sciences, Northeast Normal University, Changchun, Jilin, China
- ⁴⁰Department of Gastroenterology, The Second Affiliated Hospital of Guangxi Medical University, Nanning, Guangxi, China
- ⁴¹Synaura Biotechnology (Shanghai) Co., Ltd., Shanghai, China
- ⁴²School of Public Health, University of Michigan, Ann Arbor, Michigan, USA
- ⁴³Institute of Soil Science, Chinese Academy of Sciences, Nanjing, Jiangsu, China
- ⁴⁴School of Art and Archaeology of Zhejiang University, Zhejiang, China
- ⁴⁵College of Animal Science and Technology, Guangxi University, Nanning, China
- ⁴⁶Fujian Provincial Key Laboratory for Subtropical Resources and Environment, Fujian Normal University, Fuzhou, China
- ⁴⁷College of Energy and Environmental Engineering, Hebei University of Engineering, Handan, Hebei, China
- ⁴⁸HaploX Biotechnology, Shenzhen, China
- ⁴⁹College of Life Sciences, Henan Agricultural University, Zhengzhou, China
- ⁵⁰College of Life Science, Nankai University, Tianjin, China
- ⁵¹Jiangxi Province Key Laboratory of Sustainable Utilization of Traditional Chinese Medicine Resources, Institute of Traditional Chinese Medicine Health Industry, China Academy of Chinese Medical Sciences, Jiangxi, China
- ⁵²Beijing Key Laboratory of Emerging Infectious Diseases, Institute of Infectious Diseases, Beijing Ditan Hospital, Capital Medical University, Beijing, China
- ⁵³State Key Laboratory of Tea Plant Biology and Utilization, Anhui Agricultural University, Hefei, Anhui, China
- ⁵⁴School of Life Sciences and Food Technology, Hanshan Normal University, Chaozhou, China
- ⁵⁵Tsinghua University, Beijing, China
- ⁵⁶State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China
- ⁵⁷Department of Life Sciences, Imperial College of London, London, UK
- ⁵⁸State Key Laboratory of Herbage Improvement and Grassland Agro-Ecosystems, Centre for Grassland Microbiome, College of Pastoral Agriculture Science and Technology, Lanzhou University, Lanzhou, Gansu, China
- ⁵⁹Department of Oncology, The Fifth Affiliated Hospital of Sun Yat-sen University, Zhuhai, Guangdong, China
- ⁶⁰National-Local Joint Engineering Research Center of Biomass Refining and High-Quality Utilization, Institute of Urban and Rural Mining, Changzhou University, Changzhou, Jiangsu, China
- ⁶¹Department of Pathology, Affiliated Cancer Hospital of Zhengzhou University, Zhengzhou, China
- ⁶²Institute of Microbiology, Chinese Academy of Sciences, Beijing, China
- ⁶³College of Water Sciences, Beijing Normal University, Beijing, China
- ⁶⁴Deyang People's Hospital, Deyang, Sichuan, China
- ⁶⁵New Direction Biotechnology (Tianjin) Co., Ltd, Tianjin, China
- ⁶⁶College of Energy and Environmental Engineering, Hebei Key Laboratory of Air Pollution Cause and Impact, Hebei University of Engineering, Handan, China
- ⁶⁷Institute of Botany, Chinese Academy of Sciences, Beijing, China
- ⁶⁸Institute of Biotechnology, Beijing Academy of Agriculture and Forestry Sciences, Beijing, China
- ⁶⁹National Resource Center for Chinese Materia Medica, China Academy of Chinese Medical Sciences, Jiangsu, China
- ⁷⁰College of Animal Science, Anhui Science and Technology University, Chuzhou, China
- ⁷¹State Key Laboratory for Biology of Plant Diseases and Insect Pests, Institute of Plant Protection, Chinese Academy of Agricultural Sciences, Beijing, China
- ⁷²State Key Laboratory of Environmental Criteria and Risk Assessment, Chinese Research Academy of Environmental Sciences, Beijing, China
- ⁷³China CDC Key Laboratory of Environment and Population Health, National Institute of Environmental Health, Chinese Center for Disease Control and Prevention, Beijing, China
- ⁷⁴College of Animal Science and Technology, Yangzhou University, Yangzhou, Jiangsu, China
- ⁷⁵Key Laboratory of Agro-Ecological Processes in Subtropical Region, Institute of Subtropical Agriculture, Chinese Academy of Sciences, Changsha, China
- ⁷⁶Institute of Ecological Conservation and Restoration, Chinese Academy of Forestry, Beijing, China
- ⁷⁷State Key Laboratory for Ecological Security of Regions and Cities, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen, China
- ⁷⁸Center of Liver Diseases Division 3, Beijing Ditan Hospital, Capital Medical University, Beijing, China
- ⁷⁹State Key Laboratory for Biology of Plant Diseases and Insect Pests, Key Laboratory of Invasive Alien Species Control of Ministry of Agriculture and Rural Affairs, Institute of Plant Protection, Chinese Academy of Agricultural Sciences, Beijing, China
- ⁸⁰Guangzhou University of Chinese Medicine, Guangzhou, China
- ⁸¹Key Laboratory of Livestock Biology, Northwest A&F University, Yangling, Shaanxi, China
- ⁸²College of Biological Sciences and Technology, Beijing Forestry University, Beijing, China
- ⁸³CAS Key Laboratory of Pathogenic Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China
- ⁸⁴School of Food Science and Technology, Shihezi University, Shihezi, Xinjiang, China
- ⁸⁵College of Biochemical Engineering, Beijing Union University, Beijing, China
- ⁸⁶College of Resources and Environmental Sciences, Henan Agricultural University, Zhengzhou, China
- ⁸⁷Tea Research Institute, Chinese Academy of Agricultural Sciences, Hangzhou, Zhejiang, China
- ⁸⁸College of Environmental Science and Engineering, Tongji University, Shanghai, China

⁸⁹Zhanjiang Key Laboratory of Human Microecology and Clinical Translation Research, the Marine Biomedical Research Institute, College of Basic Medicine, Guangdong Medical University, Zhanjiang, Guangdong, China

⁹⁰State Key Laboratory of Urban Water Resource and Environment, School of Environment, Harbin Institute of Technology, Harbin, China

⁹¹School of Ecology, Environment and Resources, Guangdong University of Technology, Guangzhou, Guangdong, China

⁹²Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences, Beijing, China

⁹³School of Traditional Chinese Medicine, Southern Medical University, Guangzhou, Guangdong, China

⁹⁴Department of Animal Sciences, Purdue University, West Lafayette, Indiana, USA

⁹⁵College of Resource and Environmental Sciences, Nanjing Agricultural University, Nanjing, Jiangsu, China

⁹⁶LifeX Institute, School of Medical Technology, Gannan Medical University, Ganzhou, China

⁹⁷Faculty of Data Science, City University of Macau, Macau, China

⁹⁸Jiangxi Provincial Key Laboratory of Conservation Biology, College of Forestry, Jiangxi Agricultural University, Nanchang, Jiangxi, China

ACKNOWLEDGMENTS

This study was financially supported by the National Natural Science Foundation of China (U23A20148, 32470055), the China Postdoctoral Science Foundation (2024M753580), the Agricultural Science and Technology Innovation Program (CAAS-ZDRW202308), the Fundamental Research Funds for the Central Public Welfare Research Institutes (ZZ13-YQ-095), the Scientific and Technology Innovation Project of China Academy of Chinese Medical Sciences (CI2021A04115), Scientific and technological innovation project of China Academy of Chinese Medical Sciences (CI2023E002), and the Key Project at Central Government Level: The Ability Establishment of Sustainable Use for Valuable Chinese Medicine Resources (2060302).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

EasyMetagenome is freely available, implemented in Shell and R, and is easy to install. Step-by-step protocols are provided on GitHub <https://github.com/YongxinLiu/EasyMetagenome>. All figure data and supplementary tables can be accessed via GitHub <https://github.com/YongxinLiu/EasyMetagenome/tree/master/appendix>. Supplementary materials (figures, tables, graphical abstract, slides, videos, Chinese translated version and update materials) may be found in the online DOI or iMeta Science <http://www.imeta.science/>.

ETHICS STATEMENT

No ethics approval was required for this study because all samples used were publicly available and obtained from open-access sources.

ORCID

Defeng Bai  <http://orcid.org/0009-0005-7762-7199>

Tong Chen  <http://orcid.org/0000-0003-3134-3113>

Jiani Xun  <http://orcid.org/0009-0005-7406-1513>

Chuang Ma  <http://orcid.org/0009-0009-0035-0733>

Hao Luo  <http://orcid.org/0009-0005-3391-8576>

Haifei Yang  <http://orcid.org/0009-0003-8204-5850>

Xiaofeng Cao  <https://orcid.org/0000-0003-2036-9748>

Gui-Lai Jiang  <https://orcid.org/0009-0000-5441-9438>

Ling N. Jin  <https://orcid.org/0000-0003-1267-7396>

Chong Li  <https://orcid.org/0000-0001-9330-5396>

Hongyang Wang  <http://orcid.org/0009-0009-2271-4411>

Zhengrong Yuan  <http://orcid.org/0000-0002-5175-0675>

Tao Wen  <http://orcid.org/0000-0003-0102-0487>

Shifu Chen  <http://orcid.org/0000-0001-5799-653X>

Guoliang Li  <http://orcid.org/0000-0002-0599-5356>

Yunyun Gao  <http://orcid.org/0000-0003-3389-9385>

Yong-Xin Liu  <http://orcid.org/0000-0003-1832-9835>

REFERENCES

- Berg, Gabriele, Daria Rybakova, Doreen Fischer, Tomislav Cernava, Marie-Christine Champomier Vergès, Trevor Charles, Xiaoyulong Chen, et al. 2020. "Microbiome Definition Re-Visited: Old Concepts and New Challenges." *Microbiome* 8: 1–22. <https://doi.org/10.1186/s40168-020-00875-0>
- Gao, Yunyun, Danyi Li, and Yong-Xin Liu. 2023. "Microbiome Research Outlook: Past, Present, and Future." *Protein & Cell* 14: 709–12. <https://doi.org/10.1093/procel/pwad031>
- Shi, Chun-Lin, Tong Chen, Canhui Lan, Ren-You Gan, Jun Yu, Fangqing Zhao, and Yong-Xin Liu. 2024. "iMetaOmics: Advancing Human and Environmental Health Through Integrated Meta-Omics." *iMetaOmics* 1: e21. <https://doi.org/10.1002/imo2.21>
- Scher, Jose U., and Steven B. Abramson. 2011. "The Microbiome and Rheumatoid Arthritis." *Nature Reviews Rheumatology* 7: 569–78. <https://doi.org/10.1038/nrrheum.2011.121>
- Xia, Han, Zhe Zhang, Chen Luo, Kangfei Wei, Xuming Li, Xiyu Mu, Meilin Duan, et al. 2023. "MultiPrime: A Reliable and Efficient Tool for Targeted Next-Generation Sequencing." *iMeta* 2: e143. <https://doi.org/10.1002/imt2.143>
- Li, Yahui, Huijie Lyu, Haifei Yang, Zhicheng Ju, Chuang Ma, Huiyu Hou, Yao Wang, et al. 2024. "FoodMicroDB: A Microbiome Database for Composition and Time-Series Research in Food." *iMetaOmics* 1: e40. <https://doi.org/10.1002/imo2.40>
- Venter, J. Craig, Karin Remington, John F. Heidelberg, Aaron L. Halpern, Doug Rusch, Jonathan A. Eisen, Dongying Wu, et al. 2004. "Environmental Genome Shotgun Sequencing of the Sargasso Sea." *Science* 304: 66–74. <https://doi.org/10.1126/science.1093857>
- Quince, Christopher, Alan W. Walker, Jared T. Simpson, Nicholas J. Loman, and Nicola Segata. 2017. "Shotgun

- Metagenomics, from Sampling to Analysis.” *Nature Biotechnology* 35: 833–44. <https://doi.org/10.1038/nbt.3935>
9. Wang, Danrui, Jiangtao Li, Lei Su, Wenli Shen, Kai Feng, Xi Peng, Zhujun Wang, et al. 2023. “Phylogenetic Diversity of Functional Genes in Deep-Sea Cold Seeps: A Novel Perspective on Metagenomics.” *Microbiome* 11: 276. <https://doi.org/10.1186/s40168-023-01723-7>
 10. Liu, Yong-Xin, Yuan Qin, Tong Chen, Meiping Lu, Xubo Qian, Xiaoxuan Guo, and Yang Bai. 2021. “A Practical Guide to Amplicon and Metagenomic Analysis of Microbiome Data.” *Protein & Cell* 12: 315–30. <https://doi.org/10.1007/s13238-020-00724-8>
 11. Nayfach, Stephen, and Katherine S. Pollard. 2016. “Toward Accurate and Quantitative Comparative Metagenomics.” *Cell* 166: 1103–16. <https://doi.org/10.1016/j.cell.2016.08.007>
 12. Gao, Yunyun, Kai Peng, Defeng Bai, Xiao-Ye Bai, Yujing Bi, Anqi Chen, Baodong Chen, et al. 2024. “The Microbiome Protocols Ebook Initiative: Building a Bridge to Microbiome Research.” *iMeta* 3: e182. <https://doi.org/10.1002/imt2.182>
 13. Utturkar, Sagar M., Dawn M. Klingeman, José M. Bruno-Barcena, Mari S. Chinn, Amy M. Grunden, Michael Köpke, and Steven D. Brown. 2015. “Sequence Data for *Clostridium Autoethanogenum* Using Three Generations of Sequencing Technologies.” *Scientific Data* 2: 150014. <https://doi.org/10.1038/sdata.2015.14>
 14. Bertrand, Denis, Jim Shaw, Manesh Kalathiyappan, Amanda Hui Qi Ng, M. Senthil Kumar, Chenhao Li, Mirta Dvornicic, et al. 2019. “Hybrid Metagenomic Assembly Enables High-Resolution Analysis of Resistance Determinants and Mobile Elements in Human Microbiomes.” *Nature Biotechnology* 37: 937–44. <https://doi.org/10.1038/s41587-019-0191-2>
 15. Lugli, Gabriele Andrea, Federico Fontana, Chiara Tarracchini, Christian Milani, Leonardo Mancabelli, Francesca Turroni, and Marco Ventura. 2023. “MEGAnnotator2: A Pipeline for the Assembly and Annotation of Microbial Genomes.” *Microbiome Research Reports* 2: 15. <https://doi.org/10.20517/mrr.2022.21>
 16. Zhang, Tianyuan, Hanzhou Li, Mian Jiang, Huiyu Hou, Yunyun Gao, Yali Li, Fuhao Wang, et al. 2024. “Nanopore Sequencing: Flourishing in Its Teenage Years.” *Journal of Genetics and Genomics* 51: 1361–74. <https://doi.org/10.1016/j.jgg.2024.09.007>
 17. Sharpton, Thomas J. 2014. “An Introduction to the Analysis of Shotgun Metagenomic Data.” *Frontiers in Plant Science* 5: 209. <https://doi.org/10.3389/fpls.2014.00209>
 18. Chen, Shifu. 2023. “Ultrafast One-Pass FASTQ Data Pre-processing, Quality Control, and Deduplication Using Fastp.” *iMeta* 2: e107. <https://doi.org/10.1002/imt2.107>
 19. Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics* 30: 2114–20. <https://doi.org/10.1093/bioinformatics/btu170>
 20. Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9: 357–9. <https://doi.org/10.1038/nmeth.1923>
 21. Lu, Jennifer, Natalia Rincon, Derrick E. Wood, Florian P. Breitwieser, Christopher Pockrandt, Ben Langmead, Steven L. Salzberg, and Martin Steinegger. 2022. “Metagenome Analysis Using the Kraken Software Suite.” *Nature Protocols* 17: 2815–39. <https://doi.org/10.1038/s41596-022-00738-y>
 22. Lu, Jennifer, Florian P. Breitwieser, Peter Thielen, and Steven L. Salzberg. 2017. “Bracken: Estimating Species Abundance in Metagenomics Data.” *PeerJ Computer Science* 3: e104. <https://doi.org/10.7717/peerj-cs.104>
 23. Beghini, Francesco, Lauren J. McIver, Aitor Blanco-Míguez, Leonard Dubois, Francesco Asnicar, Sagun Maharjan, Ana Mailyan, et al. 2021. “Integrating Taxonomic, Functional, and Strain-Level Profiling of Diverse Microbial Communities with Biobakery 3.” *elife* 10: e65088. <https://doi.org/10.7554/eLife.65088>
 24. Caspi, Ron, Richard Billington, Carol A. Fulcher, Ingrid M. Keseler, Anamika Kothari, Markus Krummenacker, Mario Latendresse, et al. 2018. “The MetaCyc Database of Metabolic Pathways and Enzymes.” *Nucleic Acids Research* 46: D633–9. <https://doi.org/10.1093/nar/gkx935>
 25. Li, Dinghua, Ruibang Luo, Chi-Man Liu, Chi-Ming Leung, Hing-Fung Ting, Kunihiko Sadakane, Hiroshi Yamashita, and Tak-Wah Lam. 2016. “MEGAHIT v1.0: A Fast and Scalable Metagenome Assembler Driven By Advanced Methodologies and Community Practices.” *Methods* 102: 3–11. <https://doi.org/10.1016/j.ymeth.2016.02.020>
 26. Prjibelski, Andrey, Dmitry Antipov, Dmitry Meleshko, Alla Lapidus, and Anton Korobeynikov. 2020. “Using SPAdes De Novo Assembler.” *Current Protocols in Bioinformatics* 70: e102. <https://doi.org/10.1002/cpbi.102>
 27. Uritskiy, Gherman V., Jocelyne DiRuggiero, and James Taylor. 2018. “MetaWRAP—A Flexible Pipeline for Genome-Resolved Metagenomic Data Analysis.” *Microbiome* 6: 158. <https://doi.org/10.1186/s40168-018-0541-1>
 28. Wen, Tao, Guoqing Niu, Tong Chen, Qingrong Shen, Jun Yuan, and Yong-Xin Liu. 2023. “The Best Practice for Microbiome Analysis Using R.” *Protein & Cell* 14: 713–25. <https://doi.org/10.1093/procel/pwad024>
 29. Nasko, Daniel J., Sergey Koren, Adam M. Phillippy, and Todd J. Treangen. 2018. “RefSeq Database Growth Influences the Accuracy of K-Mer-Based Lowest Common Ancestor Species Identification.” *Genome Biology* 19: 165. <https://doi.org/10.1186/s13059-018-1554-6>
 30. Clarke, Erik L., Louis J. Taylor, Chunyu Zhao, Andrew Connell, Jung-Jin Lee, Bryton Fett, Frederic D. Bushman, and Kyle Bittinger. 2019. “Sunbeam: An Extensible Pipeline for Analyzing Metagenomic Sequencing Experiments.” *Microbiome* 7: 46. <https://doi.org/10.1186/s40168-019-0658-x>
 31. Zakrzewski, Martha, Thomas Bekel, Christina Ander, Alfred Pühler, Oliver Rupp, Jens Stoye, Andreas Schlüter, and Alexander Goesmann. 2013. “MetaSAMS—A Novel Software Platform for Taxonomic Classification, Functional Annotation and Comparative Analysis of Metagenome Datasets.” *Journal of Biotechnology* 167: 156–65. <https://doi.org/10.1016/j.jbiotec.2012.09.013>
 32. Tamames, Javier, and Fernando Puente-Sánchez. 2019. “SqueezeMeta, a Highly Portable, Fully Automatic Metagenomic Analysis Pipeline.” *Frontiers in Microbiology* 9: 3349. <https://doi.org/10.3389/fmicb.2018.03349>
 33. Kieser, Silas, Joseph Brown, Evgeny M. Zdobnov, Mirko Trajkovski, and Lee Ann McCue. 2020. “ATLAS: A Snakemake Workflow for Assembly, Annotation, and Genomic Binning of Metagenome Sequence Data.” *BMC Bioinformatics* 21: 257. <https://doi.org/10.1186/s12859-020-03585-4>
 34. Zhou, Zhichao, Patricia Q. Tran, Adam M. Breister, Yang Liu, Kristopher Kieft, Elise S. Cowley, Ulas Karaoz, and Karthik Anantharaman. 2022. “METABOLIC: High-Throughput Profiling

- of Microbial Genomes for Functional Traits, Metabolism, Biogeochemistry, and Community-Scale Functional Networks.” *Microbiome* 10: 33. <https://doi.org/10.1186/s40168-021-01213-8>
35. Lu, Yao, Guangyan Zhou, Jessica Ewald, Zhiqiang Pang, Tanisha Shiri, and Jianguo Xia. 2023. “MicrobiomeAnalyst 2.0: Comprehensive Statistical, Functional and Integrative Analysis of Microbiome Data.” *Nucleic Acids Research* 51: W310–8. <https://doi.org/10.1093/nar/gkad407>
 36. Gao, Yunyun, Guoxing Zhang, Shun Yao Jiang, and Yong-Xin Liu. 2024. “Wekemo Bioincloud: A User-Friendly Platform for Meta-Omics Data Analyses.” *iMeta* 3: e175. <https://doi.org/10.1002/imt2.175>
 37. Saheb Kashaf, Sara, Alexandre Almeida, Julia A. Segre, and Robert D. Finn. 2021. “Recovering Prokaryotic Genomes from Host-Associated, Short-Read Shotgun Metagenomic Sequencing Data.” *Nature Protocols* 16: 2520–41. <https://doi.org/10.1038/s41596-021-00508-2>
 38. Wu-Woods, Natalie J., Jacob T. Barlow, Florian Trigodet, Dustin G. Shaw, Anna E. Romano, Bana Jabri, A Murat Eren, and Rustem F. Ismagilov. 2023. “Microbial-Enrichment Method Enables High-Throughput Metagenomic Characterization From Host-Rich Samples.” *Nature Methods* 20: 1672–82. <https://doi.org/10.1038/s41592-023-02025-4>
 39. Gihawi, Abraham, Yuchen Ge, Jennifer Lu, Daniela Puiu, Amanda Xu, Colin S. Cooper, Daniel S. Brewer, Mihaela Pertea, and Steven L. Salzberg. 2023. “Major Data Analysis Errors Invalidate Cancer Microbiome Findings.” *MBio* 14: e01607–23. <https://doi.org/10.1128/mbio.01607-23>
 40. Amarasinghe, Shanika L., Shian Su, Xueyi Dong, Luke Zappia, Matthew E. Ritchie, and Quentin Gouil. 2020. “Opportunities and Challenges in Long-Read Sequencing Data Analysis.” *Genome Biology* 21: 30. <https://doi.org/10.1186/s13059-020-1935-5>
 41. Zhang, Haowen, Chirag Jain, and Srinivas Aluru. 2020. “A Comprehensive Evaluation of Long Read Error Correction Methods.” *BMC Genomics* 21: 889. <https://doi.org/10.1186/s12864-020-07227-0>
 42. Fu, Shuhua, Anqi Wang, and Kin Fai Au. 2019. “A Comparative Evaluation of Hybrid Error Correction Methods for Error-Prone Long Reads.” *Genome Biology* 20: 26. <https://doi.org/10.1186/s13059-018-1605-z>
 43. Ciuffreda, Laura, Héctor Rodríguez-Pérez, and Carlos Flores. 2021. “Nanopore Sequencing and Its Application to the Study of Microbial Communities.” *Computational and Structural Biotechnology Journal* 19: 1497–511. <https://doi.org/10.1016/j.csbj.2021.02.020>
 44. Wu, Guojun, Ting Xu, Naisi Zhao, Yan Y. Lam, Xiaoying Ding, Dongqin Wei, Jian Fan, et al. 2024. “A Core Microbiome Signature as an Indicator of Health.” *Cell* 187: 6550–65.e11. <https://doi.org/10.1016/j.cell.2024.09.019>
 45. Wen, Tao, Penghao Xie, Shengdie Yang, Guoqing Niu, Xiaoyu Liu, Zhexu Ding, Chao Xue, et al. 2022. “ggClusterNet: An R Package for Microbiome Network Analysis and Modularity-Based Multiple Network Layouts.” *iMeta* 1: e32. <https://doi.org/10.1002/imt2.32>
 46. Ke, Shanlin, Scott T. Weiss, and Yang-Yu Liu. 2022. “Dissecting the Role of the Human Microbiome in COVID-19 via Metagenome-Assembled Genomes.” *Nature Communications* 13: 5235. <https://doi.org/10.1038/s41467-022-32991-w>
 47. Ke, Shanlin, Xu-Wen Wang, Andrew Ratanatharathorn, Tianyi Huang, Andrea L. Roberts, Francine Grodstein, Laura D. Kubzansky, Karestan C. Koenen, and Yang-Yu Liu. 2023. “Association of Probable Post-Traumatic Stress Disorder With Dietary Pattern and Gut Microbiome in a Cohort of Women.” *Nature Mental Health* 1: 900–13. <https://doi.org/10.1038/s44220-023-00145-6>
 48. Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. “QUAST: Quality Assessment Tool for Genome Assemblies.” *Bioinformatics* 29: 1072–5. <https://doi.org/10.1093/bioinformatics/btt086>
 49. Hyatt, Doug, Gwo-Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. “Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification.” *BMC Bioinformatics* 11: 119. <https://doi.org/10.1186/1471-2105-11-119>
 50. Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. “CD-HIT: Accelerated for Clustering the Next-Generation Sequencing Data.” *Bioinformatics* 28: 3150–2. <https://doi.org/10.1093/bioinformatics/bts565>
 51. Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017. “Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression.” *Nature Methods* 14: 417–9. <https://doi.org/10.1038/nmeth.4197>
 52. Kanehisa, M. 2000. “KEGG: Kyoto Encyclopedia of Genes and Genomes.” *Nucleic Acids Research* 28: 27–30. <https://doi.org/10.1093/nar/28.1.27>
 53. Yin, Yanbin, Xizeng Mao, Jincai Yang, Xin Chen, Fenglou Mao, and Ying Xu. 2012. “dbCAN: A Web Resource for Automated Carbohydrate-Active Enzyme Annotation.” *Nucleic Acids Research* 40: W445–51. <https://doi.org/10.1093/nar/gks479>
 54. Tatusov, R. L. 2000. “The COG Database: A Tool for Genome-Scale Analysis of Protein Functions and Evolution.” *Nucleic Acids Research* 28: 33–6. <https://doi.org/10.1093/nar/28.1.33>
 55. Hernández-Plaza, Ana, Damian Szklarczyk, Jorge Botas, Carlos P. Cantalapiedra, Joaquín Giner-Lamia, Daniel R. Mende, Rebecca Kirsch, et al. 2023. “eggNOG 6.0: Enabling Comparative Genomics Across 12 535 Organisms.” *Nucleic Acids Research* 51: D389–94. <https://doi.org/10.1093/nar/gkac1022>
 56. Gibson, Molly K., Kevin J. Forsberg, and Gautam Dantas. 2015. “Improved Annotation of Antibiotic Resistance Determinants Reveals Microbial Resistomes Cluster by Ecology.” *The ISME journal* 9: 207–16. <https://doi.org/10.1038/ismej.2014.106>
 57. Alcock, Brian P., William Huynh, Romeo Chalil, Keaton W. Smith, Amogelang R. Raphenya, Mateusz A. Wlodarski, Arman Edalatmand, et al. 2023. “CARD 2023: Expanded Curation, Support for Machine Learning, and Resistome Prediction at the Comprehensive Antibiotic Resistance Database.” *Nucleic Acids Research* 51: D690–9. <https://doi.org/10.1093/nar/gkac920>
 58. Olm, Matthew R., Christopher T. Brown, Brandon Brooks, and Jillian F. Banfield. 2017. “Drep: A Tool for Fast and Accurate Genomic Comparisons That Enables Improved Genome Recovery from Metagenomes through De-Replication.” *The ISME Journal* 11: 2864–8. <https://doi.org/10.1038/ismej.2017.126>
 59. Parks, Donovan H., Maria Chuvochina, David W. Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. 2018. “A Standardized Bacterial

- Taxonomy Based on Genome Phylogeny Substantially Revises the Tree of Life." *Nature Biotechnology* 36: 996–1004. <https://doi.org/10.1038/nbt.4229>
60. Chklovski, Alex, Donovan H. Parks, Ben J. Woodcroft, and Gene W. Tyson. 2023. "CheckM2: A Rapid, Scalable and Accurate Tool for Assessing Microbial Genome Quality Using Machine Learning." *Nature Methods* 20: 1203–12. <https://doi.org/10.1038/s41592-023-01940-w>
 61. Cantalapiedra, Carlos P., Ana Hernández-Plaza, Ivica Letunic, Peer Bork, and Jaime Huerta-Cepas. 2021. "eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale." *Molecular Biology and Evolution* 38: 5825–9. <https://doi.org/10.1093/molbev/msab293>
 62. Buchfink, Benjamin, Klaus Reuter, and Hajk-Georg Drost. 2021. "Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND." *Nature Methods* 18: 366–8. <https://doi.org/10.1038/s41592-021-01101-x>
 63. Sato, Yuko, Koji Atarashi, Damian R. Plichta, Yasumichi Arai, Satoshi Sasajima, Sean M. Kearney, Wataru Suda, et al. 2021. "Novel Bile Acid Biosynthetic Pathways Are Enriched in the Microbiome of Centenarians." *Nature* 599: 458–64. <https://doi.org/10.1038/s41586-021-03832-5>
 64. Xu, Zhicheng, Xingzu Gao, Guoxue Li, Long D. Nghiem, Wenhai Luo, and Fusuo Zhang. 2024. "Microbial Sources and Sinks of Nitrous Oxide During Organic Waste Composting." *Environmental Science & Technology* 58: 7367–79. <https://doi.org/10.1021/acs.est.3c10341>
 65. Gao, Yunyun, Hao Luo, Hujie Lyu, Haifei Yang, Salsabeel Yousuf, Shi Huang, and Yong-Xin Liu. 2025. "Benchmarking Short-Read Metagenomics Tools for Removing Host Contamination." *GigaScience*, 14. <https://doi.org/10.1093/gigascience/giaf004>
 66. Blanco-Míguez, Aitor, Francesco Beghini, Fabio Cumbo, Lauren J. McIver, Kelsey N. Thompson, Moreno Zolfo, Paolo Manghi, et al. 2023. "Extending and Improving Metagenomic Taxonomic Profiling With Uncharacterized Species Using MetaPhlAn 4." *Nature biotechnology* 41: 1633–44. <https://doi.org/10.1038/s41587-023-01688-w>
 67. Bai, Defeng, Chuang Ma, Jiani Xun, Hao Luo, Haifei Yang, Hujie Lyu, Zhihao Zhu, et al. 2025. "MicrobiomeStatPlots: Microbiome Statistics Plotting Gallery for Meta-Omics and Bioinformatics." *iMeta* 4: e70002. <https://doi.org/10.1002/imt2.70002>
 68. Liu, Yong-Xin, Lei Chen, Tengfei Ma, Xiaofang Li, Maosheng Zheng, Xin Zhou, Liang Chen, et al. 2023. "Easy-Amplicon: An Easy-to-Use, Open-Source, Reproducible, and Community-Based Pipeline for Amplicon Data Analysis in Microbiome Research." *iMeta* 2: e83. <https://doi.org/10.1002/imt2.83>
 69. Yousuf, Salsabeel, Hao Luo, Meiyin Zeng, Lei Chen, Tengfei Ma, Xiaofang Li, Maosheng Zheng, et al. 2024. "Unveiling Microbial Communities With EasyAmplicon: A User-Centric Guide to Perform Amplicon Sequencing Data Analysis." *iMetaOmics* 1: e42. <https://doi.org/10.1002/imo2.42>
 70. Schwengers, Oliver, Lukas Jelonek, Marius Alfred Dieckmann, Sebastian Beyvers, Jochen Blom, and Alexander Goesmann. 2021. "Bakta: Rapid and Standardized Annotation of Bacterial Genomes via Alignment-Free Sequence Identification." *Microbial Genomics* 7: 000685. <https://doi.org/10.1099/mgen.0.000685>
 71. Grant, Jason R., Eric Enns, Eric Marinier, Arnab Mandal, Emily K. Herman, Chih-yu Chen, Morag Graham, Gary Van Domselaar, and Paul Stothard. 2023. "Proksee: In-Depth Characterization and Visualization of Bacterial Genomes." *Nucleic Acids Research* 51: W484–92. <https://doi.org/10.1093/nar/gkad326>
 72. Brown, Connor L., James Mullet, Fadi Hindi, James E. Stoll, Suraj Gupta, Minyoung Choi, Ishi Keenum, et al. 2022. "mobileOG-db: A Manually Curated Database of Protein Families Mediating the Life Cycle of Bacterial Mobile Genetic Elements." *Applied and Environmental Microbiology* 88: e00991–00922. <https://doi.org/10.1128/aem.00991-22>
 73. Eren, A Murat, Özcan C Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, Mitchell L. Sogin, and Tom O. Delmont. 2015. "Anvi'o: An Advanced Analysis and Visualization Platform for 'Omics Data'." *PeerJ* 3: e1319. <https://doi.org/10.7717/peerj.1319>
 74. Danecek, Petr, James K. Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O. Pollard, Andrew Whitwham, et al. 2021. "Twelve Years of SAMtools and BCFtools." *Gigascience* 10: giab008. <https://doi.org/10.1093/gigascience/giab008>
 75. Delmont, Tom O., and A Murat Eren. 2018. "Linking Pan-genomes and Metagenomes: The Prochlorococcus Metapan-genome." *PeerJ* 6: e4320. <https://doi.org/10.7717/peerj.4320>
 76. Jain, Chirag, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, and Srinivas Aluru. 2018. "High Throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries." *Nature Communications* 9: 5114. <https://doi.org/10.1038/s41467-018-07641-9>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Figure S1. Taxonomic profiling of environmental microbiome samples via EasyMetagenome read-based analysis. (A) Taxonomic composition across groups (left) and samples (right) at the phylum level. (B) Alpha and beta diversity among groups. Alpha diversity analysis, including Shannon (left) and richness (middle) indices, with letters indicating significant differences between groups ($P < 0.05$, ANOVA, Tukey's HSD). Beta diversity (right) diversity analysis using principal coordinates analysis (PCoA) based on Bray-Curtis dissimilarity ($P < 0.001$, PERMANOVA with ADONIS test).

Figure S2. Functional annotation of assembled contigs and comparative analysis of MAGs in environmental microbiome samples. (A) Comparison of the levels of functional modules (COGs, KOs, and CAZy) of the microbiome across different groups. The left panel shows sets included in the intersection and independent sites, and the right bar or pie charts show the categories of the functional modules in these sets. The major enriched categories are shown in the legend. (B) Phylogenetic

analysis and distribution of metagenome assembled genomes (MAGs) across different groups, highlighting evolutionary relationships of representative MAGs. The bar shows mean coverage of MAGs in all samples calculated by CoverM. (C) The distribution of completeness and contamination in MAGs, with the color of the points representing the phylum. (D) Antibiotics resistance genes annotations within MAGs. The raw data from the Bioproject (PRJNA918803 and PRJNA917055) of NCBI with accession number in Table S2.

Figure S3. Genome quality of MAGs in environmental microbiome samples. (A) Completeness and contamination scores for all MAGs, colored by their quality classification categories. (B) Contigs N50 distribution for all MAGs. (C) GC content for all MAGs. High: $\geq 90\%$ completeness, $\leq 5\%$ contamination. Medium: $> 50\%$ completeness, $\leq 5\%$ contamination. Low: $< 50\%$ completeness.

Table S1. Detailed information for human gut samples.

Table S2. Detailed information for environmental samples.

Table S3. Comparative average nucleotide identity (ANI) analysis between *Alistipes putredinis* MAG and diverse isolates of *Alistipes putredinis* from NCBI.

How to cite this article: Bai, Defeng, Tong Chen, Jiani Xun, Chuang Ma, Hao Luo, Haifei Yang, Chen Cao, et al. 2025. "EasyMetagenome: A User-Friendly and Flexible Pipeline for Shotgun Metagenomic Analysis in Microbiome Research." *iMeta* 4, e70001. <https://doi.org/10.1002/imt2.70001>