Heliyon 10 (2024) e37422

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

MBCN: A novel reference database for Effcient Metagenomic analysis of human gut microbiome

Bo Zheng [a,b,1], Junming Xu [c,1], Yijie Zhang [b,1], Junjie Qin [c], Decai Yuan [a], Tingting Fan [a], Weibin Wu [d], Yan Chen [b,*], Yuyang Jiang [a,b,e]

[a] State Key Laboratory of Chemical Oncogenomics, Tsinghua Shenzhen International Graduate School, Shenzhen, 518055, PR China
[b] School of Pharmacy, Shenzhen University Medical School, Shenzhen University, Shenzhen, 518055, PR China
[c] Department of Human Microbiome, Promegene Institute, Shenzhen, 518000, PR China
[d] Shenzhen Bay Biotechnology Co., Ltd. Shenzhen, 518110, PR China
[e] School of Pharmaceutical Sciences, Tsinghua University, Beijing, 100084, PR China

ARTICLE INFO

Keywords:
Human gut microbiome
Metagenome-assembly genomes
Genome catalog
Profile database
Metagenomic analysis

ABSTRACT

Metagenomic shotgun sequencing data can identify microbes and their proportions. But metagenomic shotgun data profiling results obtained from multiple projects using different reference databases are difficult to compare and apply meta-analysis. Our work aims to create a novel collection of human gut prokaryotic genomes, named Microbiome Collection Navigator (MBCN). 2379 human gut metagenomic samples are screened, and 16,785 metagenome-assembled genomes (MAGs) are assembled using a standardized pipeline. In addition, MAGs are combined with the representative genomes from public prokaryotic genomes collections to cluster, and pan-genomes for each cluster's genomes are constructed to build Kraken2 and Bracken databases. The databases built by MBCN are more comprehensive and accurate for profiling metagenomic reads comparing with other collections on simulated reads and virtual bio-projects. We profile 1082 human gut metagenomic samples with MBCN database and organize profiles and metadata on the web program. Meanwhile, using MBCN as a reference database, we also develop a unified, standardized, and systematic metagenomic analysis pipeline and platform, named MicrobiotaCN (http://www.microbiota.cn) and common statistical and visualization tools for microbiome research are integrated into the web program. Taken together, MBCN and MicrobiotaCN can be a valuable resource and a powerful tool that allows researchers to perform metagenomic analysis by a unified pipeline efficiently.

1. Introduction

The human gut microbiome, which is now recognized as a complex ecosystem that plays a critical role in human health and disease, has been the subject of extensive research in recent years [1–5]. Prokaryotic microorganisms perform a variety of functions, such as breaking down complex carbohydrates, producing vitamins, and modulating the immune system, so much so that most of the functions are unclear [6]. However, it is difficult to study these microorganisms because most cannot be cultured by traditional methods [7].

* Corresponding author.
E-mail address: chenyan@szu.edu.cn (Y. Chen).
[1] These authors contribute equally to this work and share first authorship.

https://doi.org/10.1016/j.heliyon.2024.e37422
Received 21 February 2024; Received in revised form 9 July 2024; Accepted 3 September 2024
Available online 6 September 2024

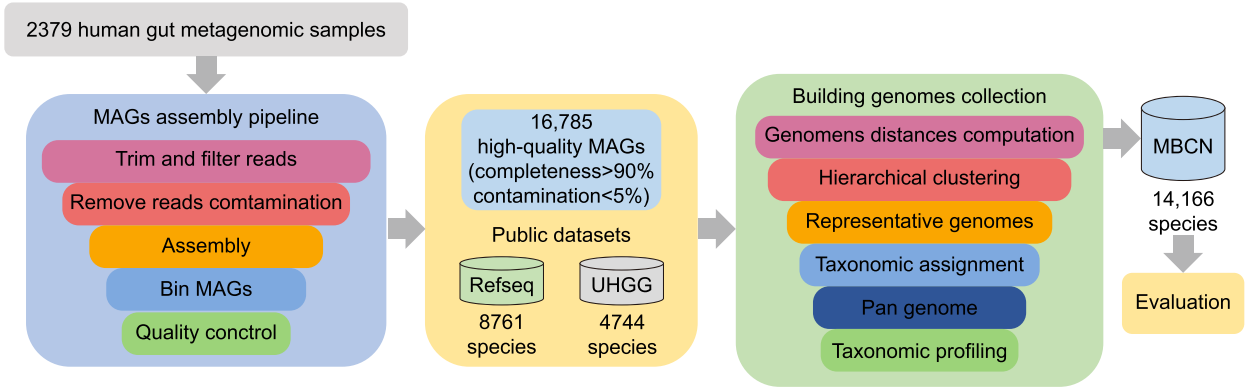


Fig. 1. Overview of workflow.
We first collect the latest metagenomic sequencing data and assembly high-quality reference genomes to create an updated human gut microbial genome collection called Microbiome Collection Navigator (MBCN). Using this collection, we built Kraken2 and Bracken reference databases and evaluate their performance. We additionally develop a metagenomic analysis pipeline using MBCN as reference and integrated common statistical and visualization tools into the MicrobiotaCN platform.

Recent advances in DNA sequencing technology have enabled the sequencing of the entire genomes of these prokaryotic microorganisms, which has led to a wealth of information on the genetic diversity and functional potential of the gut microbiome [8].

Metagenomics approaches have been driven by exponential increases in sequencing throughput and associated declines in costs, leading to its widespread adoption in clinical research [9]. However, accurate classification of the composition of microbial community from metagenomic data remains a challenge due to their complexity, the short-read lengths, and incomplete genome reference database [10,11]. The reference database is crucial in metagenomic analysis as the quality of the data analysis using metagenomic classifiers relies heavily on the quality of the reference database. Despite their importance, most "standard" reference databases of metagenomic classifiers may not be tailored to specific research questions, such as the gut microbiome [12]. To address this, initiatives like Metagenomics of the Human Intestinal Tract (MetaHIT) [13], the integrated gene catalog (IGC) [14] have obtained prokaryotic microorganisms genomes from gut, enabling their use in metagenomic studies. Recent approaches that recover high-quality metagenomic assembled genomes (MAGs) from metagenomic datasets are addressing this limitation, providing draft genomes of uncultured taxa [5,7,15]. Genome collections of human gut microbes have been built based on MAGs, such as Unified Human Gut Genome (UHGG) [16] and Humgut [17], but these collections often use only one representative genome for a species, leading to a large number of gene deletions. Furthermore, even for the same microorganism, genomes from different environments usually have significant differences, highlighting the need for a comprehensive and tailored reference database for accurate metagenomic analysis [18].

The ability to combine samples from multiple studies using new bioinformatics tools has transformed our understanding of microbial diversity and its relationship with human health, offering insights into the physiological states involved [19]. While incremental data can provide a more comprehensive understanding of the functional composition of the microbial world [20,21], the inconsistent methodologies used in different studies make it difficult for non-specialist researchers to apply meta-analysis to metagenomics research. To address this, various online analysis platforms, including Qiita [22], KBase [23], Galaxy [24], have been developed to provide easy, systematic, and comprehensive statistical analysis, interactive visualization, and meta-analysis of microbiome data for researchers without specialized programming skills. However, some of these platforms require complex registration and data uploading processes, which can be inconvenient for researchers who only need to perform simple analyses on small amounts of data. Other platforms, such as MicrobiomeAnalyst [25] and MG-RAST [26], allow direct analysis without registration, but have limitations in terms of the types of analysis they can perform, or only collect amplicon sequencing data for meta-analysis. Overall, the availability of user-friendly platforms for meta-analysis has facilitated access to valuable tools and insights for non-specialist researchers in the field of metagenomics.

We develop a method for analyzing human gut metagenomic data that involves assembling a high-quality microbial reference database from recent human gut metagenomic datasets, as well as genomes from UHGG and Refseq. This reference database, called the Microbiome Collection Navigator (MBCN), uses the Genome Taxonomy Database (GTDB) [27] for taxonomy classification, which provides a quantitative and convenient operational species definition compared to the NCBI taxonomy [28,29]. To demonstrate the effectiveness of MBCN, we compare it to several widely used human gut microbial genome collections and find that it had superior recall and precision, as well as a higher proportion of assigned reads. Additionally, we have developed a web-based program called to meet the requirements of current metagenomic shotgun data analysis. This web allows researchers to easily obtain generated profiles and metadata of human gut shotgun metagenomic data, and integrates common statistical and visualization tools for exploratory analysis of metagenomic species abundance profiles and taxonomic features.

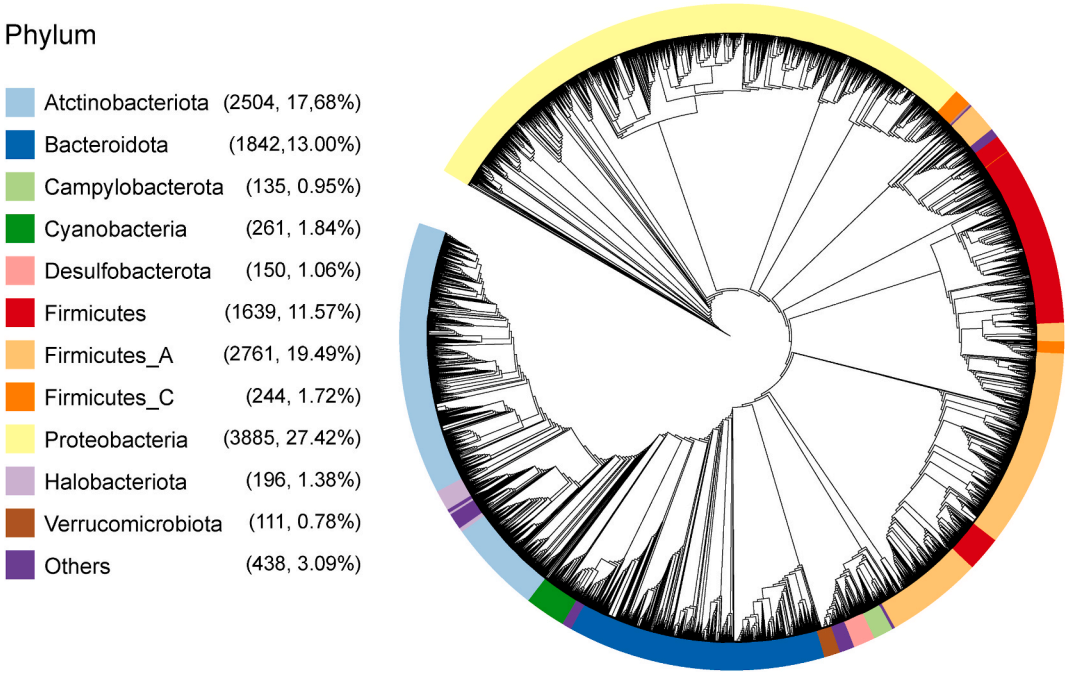


Fig. 2. Phylogenetic tree of representative genomes of MBCN established by PhyloPhlAn.
The circular tree is drawn by using iTOL (https://itol.embl.de/) with option "Ignore branch lengths". Clades and outer circles are colored by Genome Taxonomy Database (GTDB) phylum annotation.

2. Results

2.1. Human gut microbiome genome collection

In this study, human gut metagenomic sequencing data from 2379 samples collected worldwide from individuals of various ages are obtained from 14 different bio-projects (Supplementary Table S1). We develop a workflow to construct a collection of human gut microbiome genomes and facilitate downstream analysis (Fig. 1): The latest human gut metagenomic sequencing data from NCBI is collected and a workflow is constructed for metagenomic assembly, binning, and quality assessment, which is then used to assemble high-quality microbial reference MAGs. Fastp is used for quality control. Megahit assembled contigs, and BWA-MEM mapped reads back. MetaWRAP with three binning algorithms (metaBAT2, MaxBin2, CONCOCT) refined samples into MAGs. CheckM assessed MAGs, retaining 16,785 high-quality MAGs, contain 14,166 clusters, with 8673 representative genomes from Refseq, 4374 from UHGG, and 1119 from our own assembled MAGs. Subsequently, we perform clustering by dRep and taxonomic assignment by GTDB-Tk on the obtained MAGs. The majority of clusters (9995 out of 14,166) comprising a single MAG. The most diverse MBCN cluster is *Escherichia coli*, which consist of 8314 different MAGs, followed by 7511 MAGs related to *Agathobacter rectalis*. Pan-genomes are constructed for each species using Prokka and Roary, and utilized for building Kraken2 and Bracken databases. These MAGs are integrated with publicly available human gut microbial genomes to create the most up-to-date human gut microbial genome collection, named MBCN.

Both GTDB and NCBI naming systems have limitations in accurately identifying species-level clusters, leading to ambiguous taxonomic annotations. This results in a disparity between the number of species-level clusters (14,166) and the total cluster names (12,450 in GTDB and 6318 in NCBI). Notably, many clusters share names, including 2838 clusters share GTDB names with other clusters, and 8587 clusters share NCBI names. This problem is particularly pronounced in the case of *Collinsella* clusters, with 160 clusters named after GTDB *Collinsella* species and 471 clusters named after NCBI *Collinsella* species.

Fig. 2 depicts the distribution of 13,853 bacterial and 313 archaeal species across the phylogenetic tree. Additionally, 21 % of *Firmicutes A*'s diversity comprises unassigned species-level GTDB clusters, while other main gut phyla contribute less than 10 %. Our assembly identified 191 representative genomes of unassigned GTDB clusters, primarily from *Firmicutes A* (88 clusters) and *Actinobacteriota* (42 clusters). Some phyla have few genomes, but they are closely associated with health in our assembled MAGs. *Akkermansia* is the only genus of *Verrucomicrobiota* in the human gut, which has received extensive attention for its role in health and disease [4]. Only one unassigned representative genome of a *Pyramidobacter* cluster assembled by our workflow belonging to the *Synergistota* phylum is reported to be related to disease [3].

2.2. Evaluating Kraken2 database performance by simulated reads

The performance of Kraken2 and Bracken databases built by MBCN, UHGG and Humgut is evaluated using data containing simulated Illumina reads from 100 genomes. These genomes come from 84 representative species that are previously used to compare metagenomic assembly algorithms and are available at http://www.bork.embl.de/~mende/simulated_data [30].

Kraken2 is initially used to assign reads, and the unassigned reads accounted for 9.03 % using the MBCN database, 61.85 % using the UHGG database, and 76.81 % using the Humgut database. All metagenomic classifiers report false positives for species that are not present, resulting in thousands of low-abundance predictions that researchers need to filter. Therefore, Kraken2 profiling results are adjusted using Bracken at the species and genus levels and are compared with different minimum species abundance thresholds. The performance of Kraken2 and Bracken databases built by MBCN, UHGG, and Humgut with different minimum species abundance thresholds (Fig. 3) show that MBCN has significantly fewer false positives than UHGG and Humgut at both species and genus levels. Moreover, the false discovery rate (FDR) of MBCN is much smaller than that of UHGG and Humgut. By setting the detection limit at 0.025 %, most false positives are filtered out.

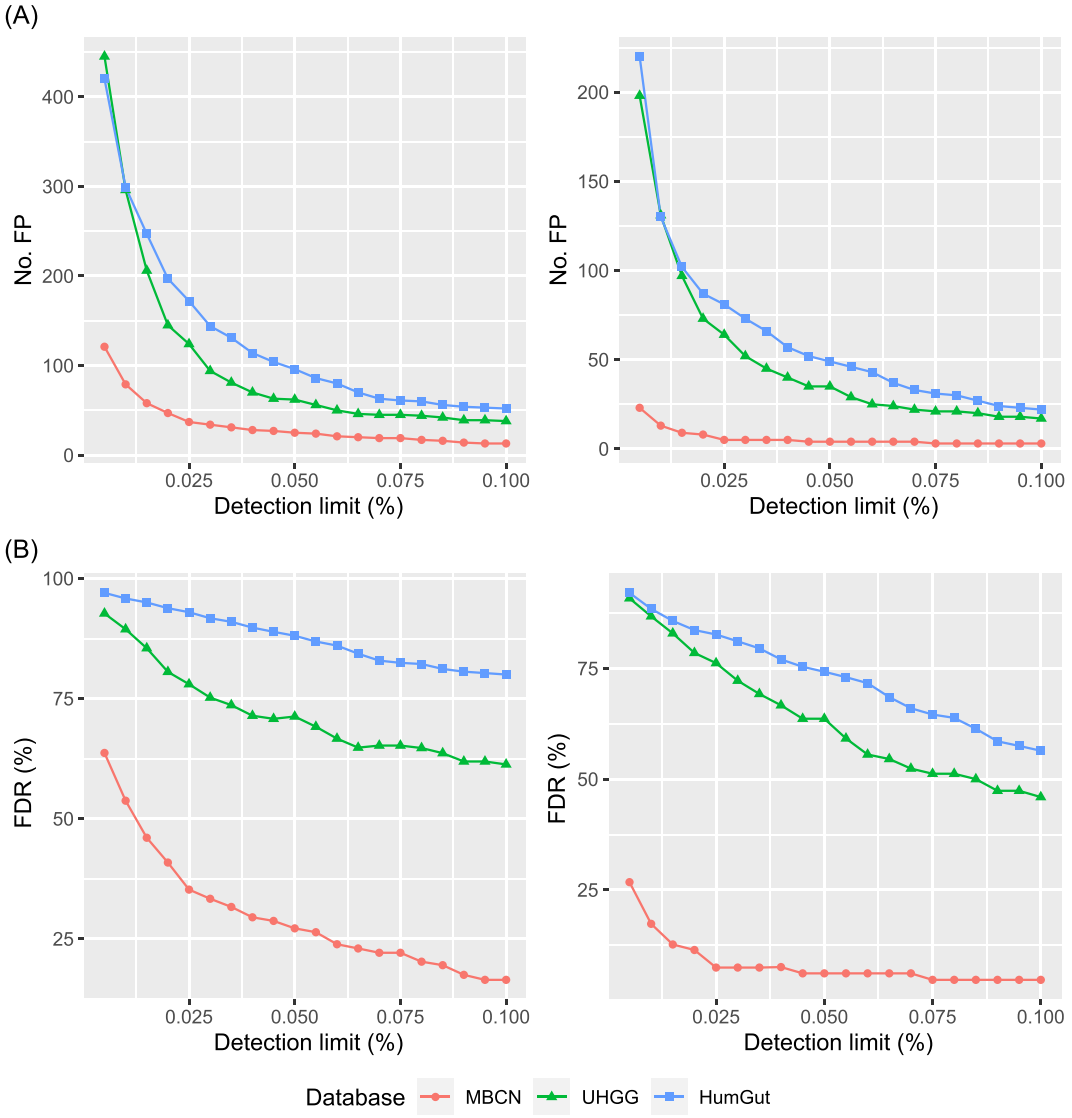


Fig. 3. The comparison of performance of Kraken2 database built by MBCN, Unified Human Gut Genome (UHGG) and Humgut with different minimum species abundance thresholds.

A. Number of false positives (FPs) of profiles obtain using Kraken2 database built by MBCN, UHGG and Humgut at species (left figure) and genus (right figure) level with different minimum species abundance thresholds (detection limits).

B. False discovery rate (FDR) of profiles obtain using Kraken2 database built by MBCN, UHGG and Humgut at species (left figure) and genus (right figure) level with different minimum species abundance thresholds (detection limits).

To evaluate the performance of databases further quantitatively in predicting relative abundance of species the database, we utilize real reads from a mock microbial community (https://github.com/LomanLab/mockcommunity) to assess the database. This microbial community contain 8 bacteria and 2 yeasts, and here we focus only on the relative abundances of the bacteria. The performance of the MBCN, UHGG, and Humgut databases in predicting relative abundance of species is evaluated and compared with the results of MetaPhlAn, as shown in Table 1. MBCN, UHGG, and Humgut performs well in avoiding false positive species, while MetaPhlAn erroneously identifies more false positives. In terms of recall, MBCN, UHGG, and MetaPhlAn all identified the 8 bacterial species, while Humgut only identify 5 species. For predicting relative abundances, MBCN performs the best in terms of both L1 and L2 distances. Overall, MBCN demonstrate superior performance in quantitative evaluation using real reads from the mock microbial community, excelling at avoiding false positives, achieving full recall of species, and most accurately predicting relative abundances based on L1 and L2 distance metrics. UHGG performs well at avoiding false positives and recall respectively, but are surpassed by MBCN for overall quantitative assessment. In comparison, Humgut has difficulty achieving full species recall, and MetaPhlAn has worse performance avoiding false positives, which lead in worse results in predicting abundances.

2.3. Comparison of the different databases on human gut metagenomic samples

To further investigate the performance of different databases in practical applications, we utilize data from PRJNA453965, a study that employs shotgun metagenomic analysis to compare gut microbial communities between breast cancer patients and healthy controls [31]. The study obtains 133 stool samples from premenopausal breast cancer patients (n = 18), healthy premenopausal controls (n = 25), postmenopausal breast cancer patients (n = 44), and healthy postmenopausal controls (n = 46), which are previously aligned with a reference catalog of the human gut microbiome (IGC) to obtain a relative abundance profile. We compare the community profiles generated using Kraken2 databases built based on MBCN with those obtained using the IGC, UHGG, and Humgut databases. To mitigate potential confounding factors, all breast cancer patients include in the study are diagnosed by pathological examination at the Affiliated Tumor Hospital. The healthy control group is recruited from individuals visiting the Medical Examination Center of the First Affiliated Hospital of Guangxi Medical University who are confirmed to be free of breast cancer upon examination. Exclusionary criteria for all study participants are: presence of diarrhea, diabetes, ulcerative colitis, Crohn's disease, or other infectious diseases; and use of antibiotics, steroid hormones, Chinese herbal medicines (oral, intramuscular, or intravenous), or probiotics such as yogurt during the 3 months prior to fecal sample collection. Additionally, breast cancer patients do not receive chemotherapy, radiation therapy, or surgery prior to the collection of fecal samples for analysis.

Since the community composition of the fecal samples is unknown, to assess the community profiles produced by each metagenomic classifier, other measurable aspects are evaluated. The percentage of reads assigned to a species by MBCN account for 89.77 $\pm$ 2.21 %, which is significantly higher than that of IGC (50.59 $\pm$ 13.45 %), UHGG (72.39 $\pm$ 4.29 %) and Humgut (79.03 $\pm$ 4.68 %). Even though Humgut has introduced genomes from Refseq, it shows little improvement compared to UHGG. MBCN's ability to identify more sequences than other databases can be attributed to the addition of many genomes from RefSeq and MAGs assembled that are not present in other reference databases. Furthermore, we assess the difference in gut microbes (with a mean relative abundance $\geq$0.025 %) between breast cancer patients and healthy controls compare with the results of mOTUs and MetaPhlAn. Base on the results of the simulated communities, species with an estimated mean relative abundance of <0.025 % are removed. The results of IGC, UHGG, Humgut, mOTUs and MetaPhlAn do not demonstrate any significant differential species between premenopausal breast cancer patients and healthy premenopausal controls (Wilcoxon rank-sum test, FDR-adjusted p-value (q-value) < 0.05). However, the results obtained using MBCN show that six species are significantly different in premenopausal samples, all of which are enriched in healthy controls (Table 2). Notably, *Sutterella parvirubra* [32] and *Parabacteroides gordonii* [33] are reported to be enriched in healthy individuals.

MBCN has the potential to identify more differential species in postmenopausal breast cancer patients and healthy postmenopausal controls compared to IGC, UHGG, Humgut, mOTUs and MetaPhlAn databases (Fig. 4). Supplementary Table S2 shows the comparison of the differential species reported by these databases. Supplementary Table S3-S8 include the details of differential species reported by each database. Specifically, MBCN identifies 35 differential species, which is substantially higher than Humgut (16 species), UHGG (15 species), IGC (7 species), mOTUs (7 species), and MetaPhlAn (7 species). It is important to acknowledge that due to inherent differences in the underlying algorithms and methodologies employed by these databases and tools, direct comparisons of absolute numbers should be interpreted with caution. Nevertheless, MBCN's ability to detect a higher number of differential species suggests a potential increase in sensitivity or a broader taxonomic coverage. All databases consistently reported a higher abundance of *Escherichia coli* in postmenopausal breast cancer patients, while *Lachnospira eligens* was reported to be lower. MBCN stands out by containing a larger number of genes to identify species not present in other databases, with 3449 species for IGC, 4744 species for UHGG, and 5170 species

Table 1
Performance of databases in predicting the relative abundance.

Database	MBCN	UHGG	Humgut	MetaPhlAn
Relative abundance of FP	0 %	0 %	0.61 %	12.93 %
Recall	100 %	100 %	62.5 %	100 %
L1 distance	0.3751	0.4582	0.8177	0.7257
L2 distance	0.1381	0.1693	0.3018	0.2998

MBCN: Microbiome Collection Navigator; UHGG: Unified Human Gut Genome; FP: false positive.

Table 2
Relative abundance of different species in premenopausal breast cancer patients and premenopausal healthy controls reported by MBCN database.

Names	*p* value	q value	Control mean	Control sd	Case mean	Case sd
Sutterella parvirubra	0.000193	0.013974	0.001692	0.006598	9.65E-08	4.21E-07
CAG-882 sp 000435595	0.000094	0.013586	0.001221	0.005996	2.20E-07	6.60E-07
Butyricimonas sp. 900184685	0.000573	0.025207	0.001123	0.002927	3.15E-05	7.52E-05
Parabacteroides gordonii	0.000828	0.030025	0.000114	0.000162	1.92E-05	5.69E-05
Prevotella sp. 000434975	0.001655	0.044013	0.010765	0.053139	3.04E-05	1.26E-04
Prevotella sp. 900313215	0.001001	0.032244	0.009081	0.034932	3.77E-05	1.35E-04

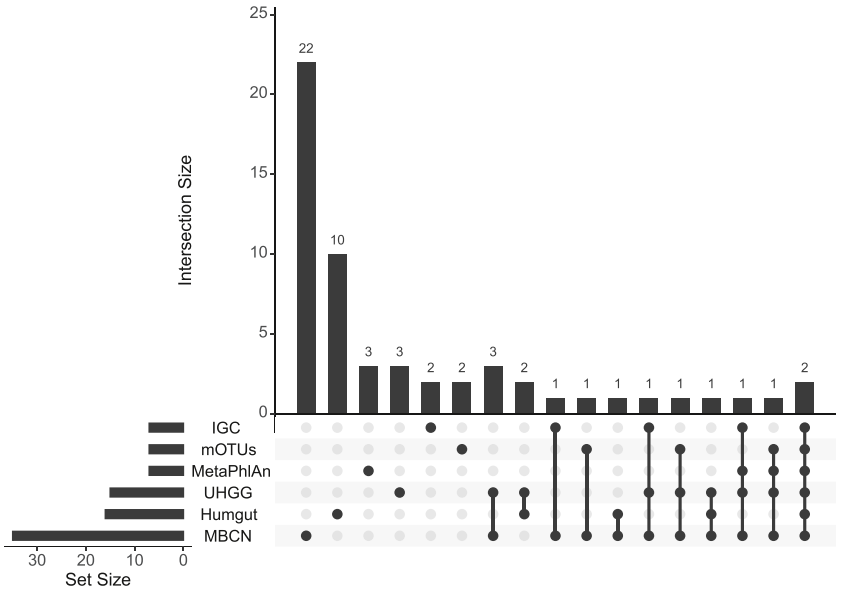


Fig. 4. Significantly differential species between postmenopausal breast cancer patients and healthy postmenopausal controls in PRNAJ453965 identified by 6 different reference databases.
The sets display individually on the left vertical axis show the number of differential species identified by 6 different reference databases. Each set is represented by a horizontal bar. The horizontal axis shows the intersections of the different sets. Each intersection is a combination of one or more sets. The intersections are sorted from largest to smallest. The upset plot provides an overview of how the different sets overlap and combine with each other.

for Humgut. This expanded gene content enables MBCN to report several novel differential species in postmenopausal breast cancer patients and healthy controls, such as *Enterobacter hormaechei* [34], *Anaerostipes hadrus* [35], *Phocaeicola* species [36], and *Alistipes* [37] are associated with breast cancer or human health. The higher number of reads assigned to most species in MBCN compared to other databases results in the identification of more differential species whose average relative abundance in other databases is below 0.025 %, such as *F23-B02 sp000431075*, *CAG-349 sp003539515*, *Prevotella* sp*000436915*. After applying FDR correction, some species, including *Megamonas funiformis*, *CAG-41 sp900066215*, and *CAG-882 sp003486385*, show significant differences in the Wilcoxon test (p-value <0.05), whereas others, such as *Eubacterium_R sp000434995*, *Lachnospira* sp003537285, and *Odoribacte splanchnicus*, do not reach statistical significance. It is worth noting that not all species exhibit an increase in the number of assigned reads in MBCN results. Increased reference sequences and variations in taxonomic assignments across species and genera can lead to the reassignment of reads. For instance, some reads initially assigned to *Escherichia fergusonii* in UHGG were reassigned to *Escherichia coli* in MBCN, resulting in a reduction of *Escherichia fergusonii*'s relative abundance below 0.025 % in MBCN's output. This underscores the importance of considering both the completeness and the specific taxonomic resolution of each database when interpreting results.

2.4. Web program

To ensure consistency in the metagenomic abundance profiles provided on our web program, we collect metadata from 1082 human gut metagenomic sequencing samples and generate profiles using the Kraken2 reference database built with our proprietary MBCN reference database. We organize the metadata and profiles to create a profile database on our web program. Meanwhile, using MBCN as a reference database, we also develop a unified, standardized, and systematic metagenomic analysis pipeline and platform,

named MicrobiotaCN (http://www.microbiota.cn) and common statistical and visualization tools for microbiome research are integrated into the web program. Within the interface of our web program, users can filter samples based on metadata and obtain a table of profiling results that can be downloaded or imported into subsequent statistical and visualization analyses.

As an illustration of the online analysis capabilities of web program, we utilize the analysis tool of web program to analyze the data from the aforementioned breast cancer study (PRJNA453965). The previously obtained profile and metadata are uploaded to the analysis module of web program to showcase the utility of commonly used statistical and visualization tools for microbiome research on our web program.

2.4.1. Data import

In the data input module, users can submit four tables to perform online visual analysis. The taxon or OTU abundance profile table and the metadata file containing the grouping information are necessary. If the row names of the submitted abundance profile table meet the requirement for the seven-level taxonomy (e.g., "*k_Bacteria*; *p_Actinobacteria*; *c_Actinobacteria*; *o Bifidobacteriales*; *f_Bifidobacteriaceae*; *g_Bifidobacterium*; *s_Bifidobacterium_longum*"), there is no need to submit the corresponding table (OUT to tax table) containing correspondence between the row names of the submitted abundance profile table and the seven-level taxonomy. Metadata must be submitted containing the sample number in the first column consistent with the column names of the abundance profile table and the corresponding grouping table. Factor metadata containing categorical metadata information is necessary, while numeric metadata information is required only for environmental factor analysis. Files can be uploaded as tab-delimited text (.txt) or comma-separated values (.csv). Users can refer to the related FAQs and tutorials for more details or try our test examples.

2.4.2. Species composition display

After data filtering, this module can use histograms (Fig. 5-A), box plots (Fig. 5-B), pie charts (Fig. 5-C), and heat maps (Fig. 5-D) to display the species abundance of each classification level, and groups are displayed separately. These four kinds of plots are used to display the phylum-level relative abundance of the PRJNA453965 project profiled using the Kraken2 database built by MBCN for postmenopausal breast cancer patients (Post + Case), healthy postmenopausal controls (Post + Control), premenopausal breast cancer patients (Pre + Case), and healthy premenopausal controls (Pre + Control) separately (Fig. 5A–D), which indicate *Bacteroidota* dominates gut microbiota in all groups, followed by *Firmicutes_A*.

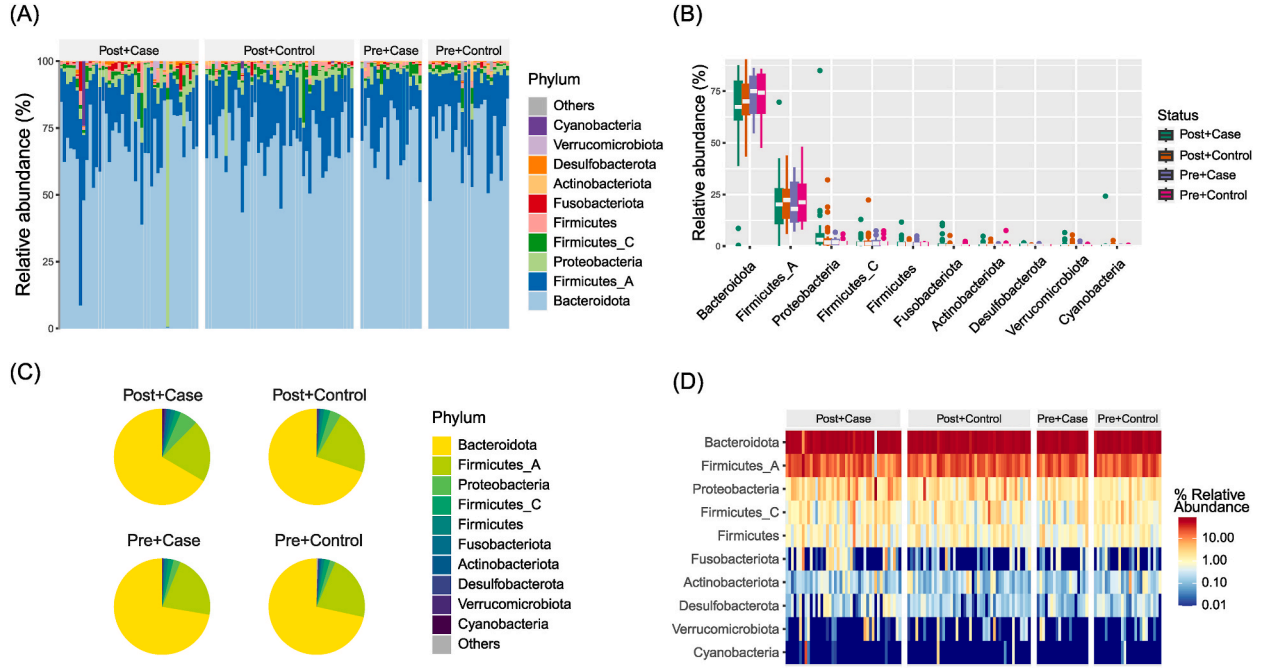


Fig. 5. Sample figures of the phylum-level relative abundance in four groups of PRJNA453965.

A. Histograms show the phylum-level relative abundance in four groups. The length of the color bar indicates the relative abundance of corresponding phylum.

B. Pie charts show the phylum-level relative abundance in four groups. The area of section indicates the relative abundance of corresponding phylum, and the color from light to dark indicates the mean relative abundance of phylum gradually decreases.

C. Box plots show the relative abundance of top 10 phylum in four groups. The mean relative abundance of phylum gradually decreases from left to right. The color of box indicates the group.

D. Heat map shows the relative abundance of top 10 phylum in four groups. The mean relative abundance of phylum gradually decreases from top to bottom. And the color from red to blue indicates the mean relative abundance of phylum gradually decreases.

2.4.3. Community diversity analysis

Community diversity analysis is mainly implemented based on the R vegan packages and is performed at different classification levels depending on the available annotations (Fig. 6-A). The alpha-diversity analysis function currently supports five common diversity measures and can automatically estimate the corresponding statistical significance. Beta-diversity analysis supports three common distance measures, and results are presented based on principal component analysis (PCA) (Fig. 6-B), principal coordinate analysis (PCoA), or non-metric multidimensional scaling (NMDS). We demonstrate the analysis of alpha-diversity and beta-diversity for the PRJNA453965 project (Fig. 6A and B), which indicate no significant difference of alpha-diversity or beta-diversity between each two groups.

2.4.4. Differential analysis and biomarker identification

The module offers four methods for differential analysis to distinguish important features between different groups, including Metastats, Wilcoxon rank-sum test, LEfSe, and random forest software. Metastats [38] is the first statistical method that specifically targets questions arising in clinical research. It analyzes metagenomic samples and identifies features that differentiate the two populations statistically (Fig. 7-A), which indicates relative abundance of 10 species with the most significant differences between postmenopausal breast cancer patients (Post + Case) and postmenopausal healthy controls (Post + Control) from Metastats analysis. Wilcoxon rank-sum test uses a volcano plot to show the results, where statistical significance is combined with the magnitude of change to identify data with large and statistically significant changes point (Fig. 7-B), which indicates fold difference and p-value of species enriched in two groups, such as *Sutterella parvirubra* is enriched in postmenopausal breast cancer patients (Post + Case), and *Prevottela rara* is enriched in postmenopausal healthy controls (Post + Control). LEfSe [39] uses the Kruskal-Wallis rank-sum test to detect features with significantly different abundances in different groups, and then it performs linear discriminant analysis to determine the most likely characteristics that explain the differences between classes (Fig. 7-C), which indicates LDA scores of significantly different clades between two group. We analyze the relative abundance of significantly different clades in different populations and plotted

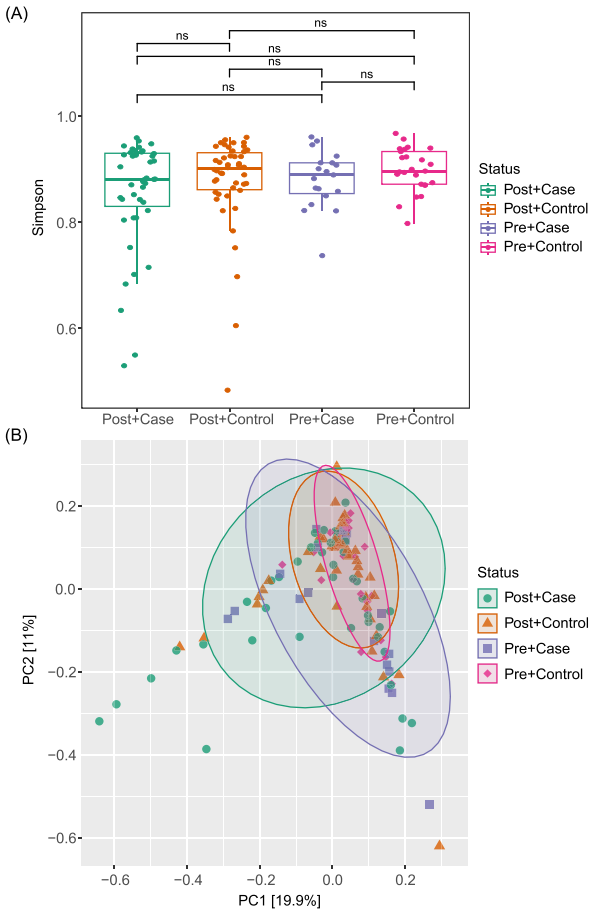
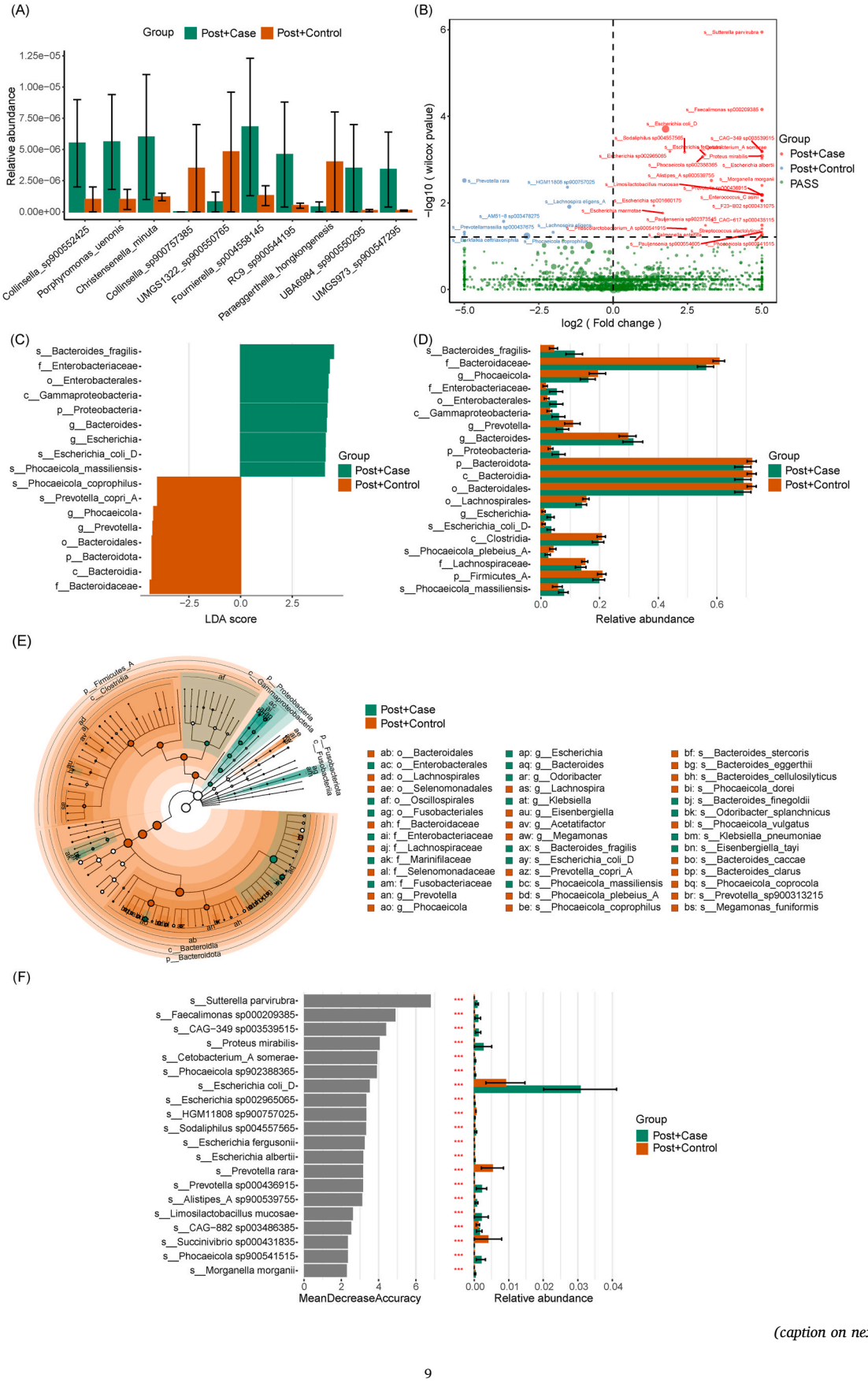


Fig. 6. Sample figures of community diversity analysis in four groups of PRJNA453965.
A. Shannon's Diversity Index. The upper line between the boxes indicates the significance of difference, and "ns" indicates no significant difference.
B. Principal component analysis with Bray-Curtis distance. The color and shape of dots indicate the group. The ellipses represent the 95 % confidence interval for groups.



(caption on next page)

Fig. 7. Sample figures of differential analysis of PRJNA453965.
A. Relative abundance of 10 species with the largest differences between postmenopausal breast cancer patients (Post + Case) and postmenopausal healthy controls (Post + Control) from Metastats analysis.
B. Wilcoxon rank-sum test in postmenopausal breast cancer patients (Post + Case) and postmenopausal healthy controls (Post + Control). Red dots indicate species enriched in postmenopausal breast cancer patients (Post + Case), blue dots indicate species enriched in postmenopausal healthy controls (Post + Control), species with p-values less than 0.05 are represented as green dots (PASS), and the size of the dots represents the relative abundance of the species.
C. Linear discriminant analysis (LDA) scores of significantly different clades (LDA scores >4) between postmenopausal breast cancer patients (Post + Case) and postmenopausal healthy controls (Post + Control) enriched species in each group.
D. Relative abundance of significantly different clades in different groups.
E. Taxonomic cladogram to demonstrate the taxonomic hierarchical distribution of marker species that are significantly enriched in each group of community samples.
F. MeanDecreaseAccuracy of random forest analysis of the importance of features in postmenopausal breast cancer patients (Post + Case) and postmenopausal healthy controls (Post + Control) and corresponding species abundance box plot of different groups.

taxonomic cladograms to demonstrate the taxonomic hierarchy distribution of significantly enriched marker species in each group of community samples (Fig. 7D and E). For example, *Bacteroides fragilis* is enriched in postmenopausal healthy controls (Post + Control), while *Bacteroidota, Bacteroidia*, *Bacteroidales* are enriched in postmenopausal breast cancer patients (Post + Case). Random forest is a non-parametric machine learning algorithm that predicts class based on the majority vote of the ensemble. It weighs the importance of each feature and provides unbiased estimates of classification error. MeanDecreaseAccuracy of random forest analysis indicates the importance of features on the left side of Fig. 7F, corresponding relative abundance of species on the right side, which indicates importance species in two groups, such as *Sutterella parvirubra* in postmenopausal breast cancer patients (Post + Case), and *Prevottela rara* in postmenopausal healthy controls (Post + Control).

2.4.5. Other analysis

Additionally, the web program offers other analysis modules to cater to various analysis requirements such as network analysis, cluster analysis, and environmental analysis. The network analysis module enables users to calculate the correlation of different species and generate network diagrams that can be saved in HTML format and interacted with using a mouse. This module provides two different graph types: chord graph and NetworkD3 dynamic network graph. Users can display the taxonomic classification level and set the Spearman or Pearson correlation coefficient threshold to draw a network relationship diagram of the eligible taxonomic classification.

The cluster analysis module allows users to choose the clustering method and distance for performing unsupervised clustering of samples. The cluster index tab offers a function for selecting the number of clusters to use in the next cluster tab.

The environmental analysis module performs Redundancy analysis (RDA) and Distance based redundancy analysis (db-RDA) on continuous environmental factors to analyze their impact on taxonomic abundance. Permission must be obtained for use of copyrighted material from other sources (including the web). Please note that it is compulsory to follow figure instructions.

2.4.6. Comparison with other web programs

Several web-based tools have been developed for microbiome data analysis. Here we compare MicrobiotaCN with MicrobiomeAnalyst, MicrobioSee and ImageGP. Table 3 summarizes the main features of each tool.

In terms of specific analysis functions, all four platforms support species composition, diversity, and biomarker analysis. Additionally, MicrobiotaCN and MicrobiomeAnalyst enables clustering and correlation analysis. For statistical comparisons and discovery, MicrobiotaCN, MicrobiomeAnalyst and MicrobioSee provide robust capabilities, while ImageGP does not offer these functionalities. Another strength of MicrobiotaCN is its integration with public data resources to enable meta-analysis, a feature only matched by MicrobioSee among the other tools. Overall, MicrobiotaCN demonstrates the most comprehensive feature set. A key advantage of MicrobiotaCN is its one-time data upload requirement for all downstream visualizations, which greatly improves analysis efficiency compared to other tools like MicrobiomeAnalyst, MicrobioSee and ImageGP that mandate repetitive data uploads.

Table 3
Comparison of MicrobiotaCN with other web tools.

Tools	MicrobiotaCN	MicrobiomeAnalyst	MicrobioSee	ImageGP
Input upload requirement	once for all visualizations	required for each new visualization	required for each new visualization	required for each new visualization
Species composition	+	+	+	+
Clustering	+	+	–	–
Diversity	+	+	+	+
Correlation analysis	+	+	–	–
Comparison	+	+	+	–
Biomarker	+	+	+	+
Integration with public data	+	+	–	–

Symbols used for feature evaluations with '+' for support and '-' for absent.

In summary, MicrobiotaCN provides a comprehensive set of analysis and visualization capabilities for microbial community studies. Its one-time data upload design significantly improves analysis efficiency. These advantages make MicrobiotaCN a uniquely powerful platform for microbiome research.

3. Discussion

MBCN is a genome collection used for profiling reference database construction and can serve as a global reference for bacteria in the intestinal tract of healthy individuals. The profiling reference database built by MBCN is available for download in web program and is more accurate and comprehensive than other intestinal microbial genome collections. It utilizes an improved standardized assembly process to obtain high-quality MAGs, and selects high-quality genomes from public databases to build a reference database, significantly reducing genome contamination. In addition, MBCN innovatively uses pan-genomes to build a reference database instead of a single representative genome, covering as much genetic information as possible for the species. By introducing high-quality genomes from Refseq, the variety of species included is significantly increased, helping to reduce unclassified sequences. In comparison to other studies, our study has constructed a reference database and downstream metagenomic analysis, and innovatively developed a web program for the entire metagenomic analysis pipeline based on the constructed genome collection, providing a directly available reference database for profiling. The web program provides online interactive analysis using the profiles obtained from a unified process for meta-analysis, allowing for the comparison and analysis of profiles from different studies. Moreover, it provides a more comprehensive and innovative analysis module.

The wealth of genetic information also poses some challenges, with the size and resources required for the database being significantly increased. A marker-based mOTUs database is built alongside the Kraken2 database for users with low computing resources. However, the results of different databases cannot be combined for analysis. One challenge that remains is the naming of species in our genome collection, as there is a profound inconsistency between the total number of species-level clusters and the number of annotated names. To address this issue, a higher-order taxonomy is used for clusters without precise taxonomy, and the abundance of these clusters is not analyzed at the species level. Files are prepared to build a custom Kraken2 database, in which all MBCN clusters received artificial TAXIDs, classified as clusters instead of classifications.

MBCN database will be updated 1–2 times per year to maintain currency and incorporate new data. Each update will include newly available gut metagenomic datasets that meet our criteria for quality, completeness, and relevance. Priority will be given to large-scale datasets from diverse human populations. New reference genome sequences relevant to the human gut microbiome will be added based on publication in public databases. Newly developed bioinformatic pipelines and algorithms validated by the research community will be implemented to enhance the analytic capabilities of MBCN. This regular update schedule will ensure MBCN remains a comprehensive, cutting-edge resource for human gut metagenomic research and discovery. User feedback will help guide enhancements incorporated through each update.

4. Conclusion

In summary, we create a novel collection of human gut prokaryotic genomes named Microbiome Collection Navigator (MBCN) and build reference databases for human gut metagenome research. Using MBCN as a reference database, we also develop a unified, standardized, and systematic metagenomic analysis pipeline and platform, named MicrobiotaCN (http://www.microbiota.cn) and common statistical and visualization tools for microbiome research are integrated into the web program. MBCN and MicrobiotaCN can be a valuable resource and a powerful tool that allows researchers to perform metagenomic analysis by a unified pipeline efficiently. Our database primarily focuses on human gut microbiota to enhance the accuracy of metagenomic sequencing results and facilitate the exploration of the relationship between gut microbiota and diseases. In the future, we plan to expand our database to include other microorganisms. Specifically, as research reveals the colonization of oral microorganisms in the intestine, we intend to incorporate oral microbiota genomes into our database to enable more comprehensive analysis of human microbiome data.

5. Star methods

5.1. Data collection

A text search using the keywords “human gut microbiome metagenomics” is performed within BioProject database of NCBI to retrieve all relevant BioProjects. Following this initial retrieval, a manual screening process is meticulously performed to exclude unrelated BioProjects. To further refine our selection, the identified BioProjects are compared against those utilized in the construction by UHGG. Through this rigorous comparison and elimination process, 14 BioProjects containing 2379 samples that haven’t previously been assembled or utilized for such purposes are identified and selected. All samples are downloaded as compressed express files using the Aspera download system (https://www.ibm.com/products/aspera).

The main source comes from the recently published Unified Human Gut Genome (UHGG) collection (http://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/human-gut/v2.0/all_genomes/), which contains genomes of 4744 species related to human gut. Another source is RefSeq, which can be obtained from ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/and ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/archaea/. We select some high-quality genomes from RefSeq according to a series of criteria to construct a reference database: 1) Completeness greater than 90 % and contamination less than 5 %. 2) N50 more than 20 kb. 3) Number of contigs less than 500. 4) Number of undetermined bases less than 10,000. 5) Has a species name in GTDB R07-RS207

classification. A total of 8761 high-quality genomes are obtained from RefSeq.

5.2. Metagenomic assembly, binning, and quality assessment

The quality control of sequencing reads belonging to 14 different BioProjects is performed using fastp [40] (https://github.com/OpenGene/fastp) with the following options: '-q 20 -u 30 -n 5 -y -Y 30 -l 90 –trim_poly_g'. For metagenomic assembly, Megahit [41] (https://github.com/voutcn/megahit) with the following options: '–min-contig-len 200 –k-min 21 –k-max 121 –k-step 20' is used. The reads are mapped back to each component using BWA-MEM [42] (https://github.com/lh3/bwa) with default parameters. The continuous depth is calculated by using jgi_summarize_bam_contig_depths. Subsequently, MetaWRAP [43] (https://github.com/bxlab/metaWRAP) is used with default parameters and three different binning algorithms (metaBAT2 [44], MaxBin2 [45], CONCOCT [46]) for metagenomic binning and refinement of each sample individually to obtain MAGs. After quality assessment of MAGs using the CheckM [47] (https://github.com/Ecogenomics/CheckM) lineage_wf workflow, 16,785 high-quality MAGs with completeness >90 % and contamination <5 % are retained.

We have made the code repository for the complete assembly workflow publicly available on the website (https://gitee.com/zjhzxjm/mbcn-flow). This workflow enables users to assemble contigs and optionally further binning contigs into bins from metagenomic data and leverage the database for taxonomic profiling and abundance calculation. The reproducible workflow can be utilized on new datasets, customized to suit specific needs, or built upon for novel approaches to microbiome characterization.

5.3. Genome clustering and taxonomic assignment

Genome clustering is carried out using dRep [48] (https://github.com/MrOlm/drep) with a 95 % genome distance threshold, which is considered a rough estimate of species delineation [49]. Only genomes with completeness >90 % and contamination <5 % are retained and assigned a genome quality score (completeness - 5 × contamination + 0.5 × log (N50)). The genome with the highest quality score in each cluster is designated as the representative genome for taxonomic assignment.

Taxonomic assignment of genomes is performed using the GTDB toolkit [50] (GTDB-Tk, version 07-RS207, https://github.com/Ecogenomics/GTDBTk) with default parameters. The GTDB classification system differs from that of the NCBI taxonomic database. The NCBI taxonomic names corresponding to the GTDB taxonomic names of all genomes are obtained using the script.

5.4. Pan-genome construction and build reference database

Although representative genomes provide species-level resolution and have a small size, they do not capture the intraspecific diversity present in microbial communities, which contains a significant portion of genetic information. To address this, we combine the genes predicted by Prokka [51] (https://github.com/tseemann/prokka) with default parameters of each cluster and use the pan-genome pipeline Roary [52] (https://sanger-pathogens.github.io/Roary) to generate a non-redundant pan-genome based on clusters, which integrate sequence information from all strains of the same species. Roary is configured with the following options: '-i 90 -cd 90 -s' (a minimum amino acid identity of 90 % for a positive match, a core gene defined at 90 % presence, and no paralog splitting).

Microbial identification and abundance estimation from complex organisms or environmental samples have been a challenge in microbiology for a long time. Various profiling methods have been proposed for this purpose, based on sequence similarity to a reference database of previously characterized sequence data. These methods can be divided into four groups based on how they establish sequence similarity: (i) nucleotide or protein alignment methods such as Centrifuge [53], Kaiju [54] and DIAMOND [55], (ii) marker gene methods such as MetaPhlAn [56] and mOTUs [57], and (iii) k-mer based methods such as Kraken2 [58] and Bracken [59]. With the improvement of computing devices, k-mer based methods are gradually becoming the mainstream method due to their high computational efficiency, despite their high memory requirements.

As all profiling methods can report false positives, manual filtering by researchers is required. We build Kraken2 and its derivatives Bracken database with MBCN to profile metagenomic data, providing good performance on a server with a large amount of memory. Using the taxonomic name and tree information from GTDB, we build the classification tree files and assign the GTDB taxonomy ID to each sequence of all pan-genomes. All pan-genomes are combined to build the Kraken2 and Bracken database.

We perform statistics on all assembled bins and select frequently occurring species to generate a condensed database MBCN_bac120 containing only 120 representative core gut microbial genomes. A more compact database will be preferable in certain scenarios, for example: For quick preliminary analysis where comprehensive characterization is not needed like validation and benchmarking of new analytical methodologies, or conducting experiments on low-memory machines or portable devices where storage space and computing capacity is limited. The condensed database retains genomes reflecting the core functional diversity of the gut microbiome while dramatically reducing the overall size. By judiciously selecting representative genomes, the key capabilities of the full database can be preserved in a streamlined package suitable for particular applications or accessibility constraints.

5.5. Development of web program

The web interface is implemented in Shiny (version 1.6.0) in R (version 4.0.5) and designed in a TabLayout. The "Home" tab provides a general introduction to the web program and guides the user through the web. The "Quick Search" tab and "Advanced Search" tab contain the relative abundance profile of species using Kraken2 database built by MBCN with corresponding metadata,

including simple or advanced filter options of metadata for users to select samples of interest. The "Tool" tab contains the download of the reference database built in this study. Users can download the database built in this study to profile metagenomic sequencing data from the web to obtain the relative abundance profile of species. The "Analyze" tab provides users with online analysis capabilities. The "Help" tab provides assistance. The application can be accessed at http://www.microbiota.cn.

A set of data analysis and visualization tools is provided through a web interface to perform data analysis in a simple, code-free manner. Users can upload the abundance profile obtained using our database or any other tool and the corresponding sample metadata for visual online analysis, including abundance display, Venn diagram analysis, alpha diversity, beta diversity, difference analysis, environmental factor analysis, network analysis. Most analysis modules are developed based on the R package microeco (version 0.3.3) [60].

Data availability Statement

The original sources presented in the study are publicly available. This data can be found in NCBI. The data underlying this study can be found in the http://www.microbiota.cn.

CRediT authorship contribution statement

Bo Zheng: Software, Methodology. **Junming Xu:** Methodology. **Yijie Zhang:** Writing – review & editing, Project administration, Data curation. **Junjie Qin:** Validation, Resources. **Decai Yuan:** Writing – review & editing, Visualization. **Tingting Fan:** Software, Investigation. **Weibin Wu:** Writing – review & editing, Visualization, Methodology. **Yan Chen:** Writing – review & editing, Writing – original draft, Validation, Supervision, Investigation, Funding acquisition. **Yuyang Jiang:** Writing – original draft, Supervision.

Declaration of competing interest

The authors have declared no competing interests.

Acknowledgments

This work is supported by Science, Technology and Innovation Commission of Shenzhen Municipality (No.2021293 and No.202231).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e37422.

References

[1] Junjie Qin, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, et al., A metagenome-wide association study of gut microbiota in type 2 diabetes, Nature 490 (2012) 55–60, https://doi.org/10.1038/nature11450.
[2] Qiang Feng, Suisha Liang, Huijue Jia, Andreas Stadlmayr, Longqing Tang, Zhou Lan, Dongya Zhang, et al., Gut microbiome development along the colorectal adenoma-carcinoma sequence, Nat. Commun. 6 (2015) 6528, https://doi.org/10.1038/ncomms7528.
[3] Zahra Amirkhanzadeh Barandouzi, Angela R. Starkweather, Wendy A. Henderson, Adwoa Gyamfi, Xiaomei S. Cong, Altered composition of gut microbiota in depression: a systematic review, Front. Psychiatr. 11 (2020) 541, https://doi.org/10.3389/fpsyt.2020.00541.
[4] I.G. Macchione, L.R. Lopetuso, G. Ianiro, M. Napoli, G. Gibiino, G. Rizzatti, V. Petito, A. Gasbarrini, F. Scaldaferri, Akkermansia muciniphila: key player in metabolic and gastrointestinal disorders, Eur. Rev. Med. Pharmacol. Sci. 23 (2019) 8075–8083, https://doi.org/10.26355/eurrev_201909_19024.
[5] Alexandre Almeida, Alex L. Mitchell, Miguel Boland, Samuel C. Forster, Gregory B. Gloor, Aleksandra Tarkowska, D Lawley Trevor, Robert D. Finn, A new genomic blueprint of the human gut microbiota, Nature 568 (2019) 499–504, https://doi.org/10.1038/s41586-019-0965-1.
[6] Andrew Maltez Thomas, Nicola Segata, Multiple levels of the unknown in microbiome research, BMC Biol. 17 (2019) 48, https://doi.org/10.1186/s12915-019-0667-z.
[7] Stephen Nayfach, Zhou Jason Shi, Rekha Seshadri, Katherine S. Pollard, Nikos C. Kyrpides, New insights from uncultivated genomes of the global human gut microbiome, Nature 568 (2019) 505–510, https://doi.org/10.1038/s41586-019-1058-x.
[8] Philip Hugenholtz, Gene W. Tyson, Microbiology: metagenomics, Nature 455 (2008) 481–483, https://doi.org/10.1038/455481a.
[9] J Gregory Caporaso, Christian L. Lauber, William A. Walters, Donna Berg-Lyons, James Huntley, Noah Fierer, Sarah M. Owens, et al., Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms, ISME J. 6 (2012) 1621–1624, https://doi.org/10.1038/ismej.2012.8.
[10] Alexander Sczyrba, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Droge, Ivan Gregor, et al., Critical assessment of metagenome interpretation-a benchmark of metagenomics software, Nat. Methods 14 (2017) 1063–1071, https://doi.org/10.1038/nmeth.4458.
[11] Simon H. Ye, J Siddle Katherine, Daniel J. Park, Pardis C. Sabeti, Benchmarking metagenomics tools for taxonomic classification, Cell 178 (2019) 779–794, https://doi.org/10.1016/j.cell.2019.07.010.
[12] Nicholas D. Youngblut, Ruth E. Ley, Struo2: efficient metagenome profiling database construction for ever-expanding microbial genome datasets, PeerJ 9 (2021) e12198, https://doi.org/10.7717/peerj.12198.
[13] Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, et al., A human gut microbial gene catalogue established by metagenomic sequencing, Nature 464 (2010) 59–65, https://doi.org/10.1038/nature08821.
[14] Junhua Li, Huijue Jia, Xianghang Cai, Huanzi Zhong, Qiang Feng, Shinichi Sunagawa, Manimozhiyan Arumugam, et al., An integrated catalog of reference genes in the human gut microbiome, Nat. Biotechnol. 32 (2014) 834–841, https://doi.org/10.1038/nbt.2942.

[15] Edoardo Pasolli, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, et al., Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle, Cell 176 (2019) 649–662 e620, https://doi.org/10.1016/j.cell.2019.01.001.
[16] Alexandre Almeida, Stephen Nayfach, Miguel Boland, Francesco Strozzi, Martin Beracochea, Zhou Jason Shi, Katherine S. Pollard, et al., A unified catalog of 204,938 reference genomes from the human gut microbiome, Nat. Biotechnol. 39 (2021) 105–114, https://doi.org/10.1038/s41587-020-0603-3.
[17] Pranvera Hiseni, Knut Rudi, Robert C. Wilson, Finn Terje Hegge, Lars Snipen, HumGut: a comprehensive human gut prokaryotic genomes collection filtered by metagenome data, Microbiome 9 (2021) 165, https://doi.org/10.1186/s40168-021-01114-w.
[18] Herve Tettelin, Vega Masignani, Michael J. Cieslewicz, Claudio Donati, Duccio Medini, Naomi L. Ward, Samuel V. Angiuoli, et al., Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome, Proc. Natl. Acad. Sci. U.S.A. 102 (2005) 13950–13955, https://doi.org/10.1073/pnas.0506758102.
[19] Luke R. Thompson, Jon G. Sanders, Daniel McDonald, Amnon Amir, Joshua Ladau, Kenneth J. Locey, Robert J. Prill, et al., A communal catalogue reveals Earth's multiscale microbial diversity, Nature 551 (2017) 457–463, https://doi.org/10.1038/nature24621.
[20] Jonas Halfvarson, Colin J. Brislawn, Regina Lamendella, Yoshiki Vazquez-Baeza, William A. Walters, Lisa M. Bramer, Mauro D'Amato, et al., Dynamics of the human gut microbiome in inflammatory bowel disease, Nat Microbiol 2 (2017) 17004, https://doi.org/10.1038/nmicrobiol.2017.4.
[21] Olivia U. Mason, Nicole M. Scott, Antonio Gonzalez, Adam Robbins-Pianka, Baelum Jacob, Jeffrey Kimbrel, Nicholas J. Bouskill, et al., Metagenomics reveals sediment microbial community response to Deepwater Horizon oil spill, ISME J. 8 (2014) 1464–1475, https://doi.org/10.1038/ismej.2013.254.
[22] Antonio Gonzalez, Jose A. Navas-Molina, Tomasz Kosciolek, Daniel McDonald, Yoshiki Vazquez-Baeza, Gail Ackermann, Jeff DeReus, et al., Qiita: rapid, web-enabled microbiome meta-analysis, Nat. Methods 15 (2018) 796–798, https://doi.org/10.1038/s41592-018-0141-9.
[23] Adam P. Arkin, Robert W. Cottingham, Christopher S. Henry, Nomi L. Harris, Rick L. Stevens, Sergei Maslov, Paramvir Dehal, et al., KBase: the United States department of energy systems biology knowledgebase, Nat. Biotechnol. 36 (2018) 566–569, https://doi.org/10.1038/nbt.4163.
[24] Vahid Jalili, Enis Afgan, Qiang Gu, Dave Clements, Daniel Blankenberg, Jeremy Goecks, James Taylor, Anton Nekrutenko, The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update, Nucleic Acids Res. 48 (2020) W395–W402, https://doi.org/10.1093/nar/gkaa434.
[25] Achal Dhariwal, Jasmine Chong, Salam Habib, Irah L. King, Luis B. Agellon, Jianguo Xia, MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data, Nucleic Acids Res. 45 (2017) W180–W188, https://doi.org/10.1093/nar/gkx295.
[26] Kevin P. Keegan, M Glass Elizabeth, Folker Meyer, MG-RAST, a metagenomics service for analysis of microbial community structure and function, Methods Mol. Biol. 1399 (2016) 207–233, https://doi.org/10.1007/978-1-4939-3369-3_13.
[27] Donovan H. Parks, Maria Chuvochina, Christian Rinke, Aaron J. Mussig, Pierre-Alain Chaumeil, Philip Hugenholtz, GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy, Nucleic Acids Res. 50 (2022) D785–D794, https://doi.org/10.1093/nar/gkab776.
[28] Scott Federhen, Type material in the NCBI taxonomy database, Nucleic Acids Res. 43 (2015) D1086–D1098, https://doi.org/10.1093/nar/gku1127.
[29] Donovan H. Parks, Maria Chuvochina, Pierre-Alain Chaumeil, Christian Rinke, Aaron J. Mussig, Philip Hugenholtz, A complete domain-to-species taxonomy for Bacteria and Archaea, Nat. Biotechnol. 38 (2020) 1079–1086, https://doi.org/10.1038/s41587-020-0501-8.
[30] Daniel R. Mende, Alison S. Waller, Shinichi Sunagawa, Aino I. Jarvelin, Michelle M. Chan, Manimozhiyan Arumugam, Jeroen Raes, Bork Peer, Assessment of metagenomic assembly using simulated next generation sequencing data, PLoS One 7 (2012) e31386, https://doi.org/10.1371/journal.pone.0031386.
[31] Jia Zhu, Ming Liao, Ziting Yao, Wenying Liang, Qibin Li, Jianlun Liu, Huawei Yang, et al., Breast cancer in postmenopausal women is associated with an altered gut metagenome, Microbiome 6 (2018) 136, https://doi.org/10.1186/s40168-018-0515-3.
[32] Kaisa Hiippala, Veera Kainulainen, Marko Kalliomaki, Perttu Arkkila, Reetta Satokari, Mucosal prevalence and interactions with the epithelium indicate commensalism of Sutterella spp, Front. Microbiol. 7 (2016) 1706, https://doi.org/10.3389/fmicb.2016.01706.
[33] Jessica C. Ezeji, Daven K. Sarikonda, Austin Hopperton, Hailey L. Erkkila, Daniel E. Cohen, Sandra P. Martinez, Fabio Cominelli, et al., Parabacteroides distasonis: intriguing aerotolerant gut anaerobe with emerging antimicrobial resistance and pathogenic and probiotic roles in human health, Gut Microb. 13 (2021) 1922241, https://doi.org/10.1080/19490976.2021.1922241.
[34] Jia V. Li, Hutan Ashrafian, Marco Bueter, James Kinross, Caroline Sands, Carel W. le Roux, Stephen R. Bloom, et al., Metabolic surgery profoundly influences gut microbial-host metabolic cross-talk, Gut 60 (2011) 1214–1223, https://doi.org/10.1136/gut.2010.234708.
[35] David Zeevi, Tal Korem, Anastasia Godneva, Noam Bar, Alexander Kurilshikov, Maya Lotan-Pompan, Adina Weinberger, et al., Structural variation in the gut microbiome associates with host health, Nature 568 (2019) 43–48, https://doi.org/10.1038/s41586-019-1065-y.
[36] Brianna S. Chrisman, Kelley M. Paskov, Nate Stockham, Jae-Yoon Jung, Maya Varma, Peter Y. Washington, Christine Tataru, et al., Improved detection of disease-associated gut microbes using 16S sequence-based biomarkers, BMC Bioinf. 22 (2021) 509, https://doi.org/10.1186/s12859-021-04427-7.
[37] Jilei Zhang, Rong Lu, Yongguo Zhang, Zaneta Matuszek, Zhang Wen, Yinglin Xia, Tao Pan, Jun Sun, tRNA queuosine modification enzyme modulates the growth and microbiome recruitment to breast tumors, Cancers 12 (2020) 628, https://doi.org/10.3390/cancers12030628.
[38] James Robert White, Niranjan Nagarajan, Mihai Pop, Statistical methods for detecting differentially abundant features in clinical metagenomic samples, PLoS Comput. Biol. 5 (2009) e1000352, https://doi.org/10.1371/journal.pcbi.1000352.
[39] Paul B. Eckburg, M Bik Elisabeth, Charles N. Bernstein, Elizabeth Purdom, Les Dethlefsen, Michael Sargent, Steven R. Gill, E Nelson Karen, David A. Relman, Diversity of the human intestinal microbial flora, Science 308 (2005) 1635–1638, https://doi.org/10.1126/science.1110591.
[40] Shifu Chen, Yanqing Zhou, Yaru Chen, Gu Jia, fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics 34 (2018) i884–i890, https://doi.org/10.1093/bioinformatics/bty560.
[41] Dinghua Li, Ruibang Luo, Chi-Man Liu, Chi-Ming Leung, Hing-Fung Ting, Kunihiko Sadakane, Hiroshi Yamashita, Tak-Wah Lam, MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices, Methods 102 (2016) 3–11, https://doi.org/10.1016/j.ymeth.2016.02.020.
[42] Heng Li, Richard Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform, Bioinformatics 26 (2010) 589–595, https://doi.org/10.1093/bioinformatics/btp698.
[43] Gherman V. Uritskiy, Jocelyne DiRuggiero, James Taylor, MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis, Microbiome 6 (2018) 158, https://doi.org/10.1186/s40168-018-0541-1.
[44] Dongwan D. Kang, Li Feng, Edward Kirton, Thomas Ashleigh, Egan Rob, Hong An, Wang Zhong, MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies, PeerJ 7 (2019) e7359, https://doi.org/10.7717/peerj.7359.
[45] Yu-Wei Wu, Blake A. Simmons, Steven W. Singer, MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets, Bioinformatics 32 (2016) 605–607, https://doi.org/10.1093/bioinformatics/btv638.
[46] Johannes Alneberg, Brynjar Smari Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z. Ijaz, Leo Lahti, Nicholas J. Loman, Anders F. Andersson, Christopher Quince, Binning metagenomic contigs by coverage and composition, Nat. Methods 11 (2014) 1144–1146, https://doi.org/10.1038/nmeth.3103.
[47] Donovan H. Parks, Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, Gene W. Tyson, CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes, Genome Res. 25 (2015) 1043–1055, https://doi.org/10.1101/gr.186072.114.
[48] Matthew R. Olm, Christopher T. Brown, Brandon Brooks, Jillian F. Banfield, dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication, ISME J. 11 (2017) 2864–2868, https://doi.org/10.1038/ismej.2017.126.
[49] Chirag Jain, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, Srinivas Aluru, High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries, Nat. Commun. 9 (2018) 5114, https://doi.org/10.1038/s41467-018-07641-9.
[50] Pierre-Alain Chaumeil, Aaron J. Mussig, Philip Hugenholtz, Donovan H. Parks, GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database, Bioinformatics 36 (2019) 1925–1927, https://doi.org/10.1093/bioinformatics/btz848.
[51] Torsten Seemann, Prokka: rapid prokaryotic genome annotation, Bioinformatics 30 (2014) 2068–2069, https://doi.org/10.1093/bioinformatics/btu153.

[52] Andrew J. Page, Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T.G. Holden, Maria Fookes, Daniel Falush, A Keane Jacqueline, Julian Parkhill, Roary: rapid large-scale prokaryote pan genome analysis, Bioinformatics 31 (2015) 3691–3693, https://doi.org/10.1093/bioinformatics/btv421.
[53] Daehwan Kim, Li Song, P Breitwieser Florian, Steven L. Salzberg, Centrifuge: rapid and sensitive classification of metagenomic sequences, Genome Res. 26 (2016) 1721–1729, https://doi.org/10.1101/gr.210641.116.
[54] Peter Menzel, Kim Lee Ng, Anders Krogh, Fast and sensitive taxonomic classification for metagenomics with Kaiju, Nat. Commun. 7 (2016) 11257, https://doi.org/10.1038/ncomms11257.
[55] Benjamin Buchfink, Chao Xie, Daniel H. Huson, Fast and sensitive protein alignment using DIAMOND, Nat. Methods 12 (2015) 59–60, https://doi.org/10.1038/nmeth.3176.
[56] Francesco Beghini, Lauren J. McIver, Aitor Blanco-Miguez, Leonard Dubois, Francesco Asnicar, Sagun Maharjan, Ana Mailyan, et al., Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3, Elife 10 (2021) e65088, https://doi.org/10.7554/eLife.65088.
[57] Alessio Milanese, Daniel R. Mende, Lucas Paoli, Guillem Salazar, Hans-Joachim Ruscheweyh, Miguelangel Cuenca, Pascal Hingamp, et al., Microbial abundance, activity and population genomic profiling with mOTUs2, Nat. Commun. 10 (2019) 1014, https://doi.org/10.1038/s41467-019-08844-4.
[58] Derrick E. Wood, Jennifer Lu, Langmead Ben, Improved metagenomic analysis with Kraken 2, Genome Biol. 20 (2019) 257, https://doi.org/10.1186/s13059-019-1891-0.
[59] Jennifer Lu, Florian P. Breitwieser, Peter Thielen, Steven L. Salzberg, Bracken: estimating species abundance in metagenomics data, PeerJ Computer Science 3 (2017) e104, https://doi.org/10.7717/peerj-cs.104.
[60] Chi Liu, Yaoming Cui, Xiangzhen Li, Minjie Yao, microeco: an R package for data mining in microbial community ecology, FEMS Microbiol. Ecol. 97 (2021) e65088, https://doi.org/10.1093/femsec/fiaa255.