MicrobiotaCN: an on-line, standard, convenient and comprehensive microbiome data analysis platform based on self-built gut prokaryotic genome collection

Bo Zheng
Tsinghua Shenzhen International Graduate School
Junming Xu
Promegene Institute
Junjie Qin
Promegene Institute
Tingting Fan
Shenzhen Bay Laboratory
Lulu Li
Tsinghua Shenzhen International Graduate School
Yan Chen
Tsinghua Shenzhen International Graduate School
Yuyang Jiang (jiangyy@sz.tsinghua.edu.cn)
Tsinghua Shenzhen International Graduate School

Research Article

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

Posted Date: May 20th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1669983/v1

Abstract

Background: The quality of the metagenomic data analysis depends on the quality of the reference database; however, the current database has some shortcomings. Different studies often use different reference databases for metagenomic analysis, resulting in inconsistent results, which can only be analyzed in isolation, and the results obtained from multiple projects are difficult to compare. Our work aimed to create a novel collection of human gut prokaryotic genomes, MBCN, as a reference database in a standardized metagenomic analysis platform named MicrobiotaCN that allows researchers to perform metagenomic analysis by the same standard pipeline efficiently.

Results: About 2,477 human gut metagenomic samples were screened, and 16,785 MAGs (metagenomic assembled genomes) were assembled using a standardized pipeline. In addition, MAGs were combined with the representative genomes from the RefSeq and UHGG collections to cluster with 95% ANI clusters, and pan-genome for each cluster's genomes were constructed. MBCN collection contained 14,166 genomic species-level clusters. Kraken2 database was built with pan-genomes of MBCN and mOTUs database with the representative genomes of each cluster of MBCN. Comparing the Kraken database built by MBCN with other collections like UHGG on simulated reads and virtual bio-projects, the database built by MBCN was found to have a better assignment rate and more accurate profiling. In virtual bio-project and practical applications, MBCN had the potential to discover more biomarkers than other databases. We profiled 1,082 human gut metagenomic samples with MBCN Kraken2 database and organized the profiles and metadata on the platform, allowing users to get metagenomic profiles by the same standard pipeline. Simultaneously, common statistical and visualization tools for microbiome research were integrated into the on-line analysis platform.

Conclusions: The reference database built based on MBCN was more comprehensive and accurate for profiling metagenomic reads, which integrates the use of the MicrobiotaCN online analysis platform will obtain a unified, one-stop metagenomic data analysis result. Thus, this could be a valuable resource for researchers to obtain profiles by a unified comprehensive reference database from different studies for meta-analysis. All data are available for free at http://www.microbiota.cn.

Introduction

Gut microbes are closely related to many human diseases[1–4], but many gut microbes cannot be cultured *in-vitro*. However, most of the functions of the microbial genome and its effects on health are unclear[5]. It is difficult to study these microorganisms because most cannot be isolated and cultivated by traditional culture methods [6]. Thus, fully exploiting and utilizing microorganisms is one of the important topics of current microbial research. The development of metagenomic sequencing technology has facilitated the study of gut microbes, with alignment using metagenomic classifiers becoming the method of choice for metagenomic data analysis. Here, the information derived from this method is determined by the database to which it is aligned.

The quality of the metagenomic data analysis using metagenomic classifiers directly depends on the quality of the reference database. Different databases cause deviations of many species and abundance results, significantly impacting the subsequent analysis. Nevertheless, the databases still have some flaws; a significant problem is matching. When analyzing human gut metagenomic sequencing samples, researchers often want to use databases containing only human gut microbes rather than all-natural microbes, significantly increasing computational resource requirements. However, previous human gut microbes collections of microorganisms are not explicitly designed for profiling and are often used with insufficient comprehensiveness.

Furthermore, conspecific microbial genomes from different isolates often exhibit considerable genomic heterogeneity, which may be due to clonal bias, environmental adaptation, or human error likely to occur during the cultivation process. Previous collections of human gut microbes tend to adopt only one representative genome for a species, resulting in many gene deletions in one species. A comprehensive and high-quality reference genome is essential for the functional characterization and taxonomy of the human gut microbiota[7]. Although there are many statistical analysis platforms for metagenomic data like MicrobiomeAnalyst[8], various studies often use different reference databases for metagenomic analysis due to the lack of standardization of the previous step; that is, there are few profile results profiled by consistent reference database. Thus, statistical analysis is done in isolation for each project, and the results obtained cannot be compared and traced across multiple projects.

In order to solve the problems that the quality of reference databases used in human gut metagenome research is not enough and taxonomic profiles obtained from different studies cannot be comprehensively analyzed, and to meet the new requirements arising from the current microbiome data analysis, this study constructed a collection of human gut prokaryotic genomes named MBCN and developed a unified, standardized, and systematic metagenomic analysis pipeline and platform, named MicrobiotaCN. This data analysis platform integrates common statistical and visualization tools for microbiome research. Thus, researchers can easily perform exploratory analysis on metagenomic abundance profiles and taxonomic features generated according to the same standard process.

Methods

Human gut metagenomic sample

In the Sequence Reading Archive (NCBI/SRA, https://www.ncbi.nlm.nih.gov/sra), a text search was conducted on all human gut metagenome samples. Recently published human gut metagenome samples not included in other collections were collected. All samples are downloaded as compressed express files using the Aspera download system (https://www.ibm.com/products/aspera). About 2,477 samples covering 13 bio-projects were collected.

Metagenomic assembly, binning and quality assessment

For sequencing reads belonging to 13 different BioProjects, fastp[9](https://github.com/OpenGene/fastp) was used for quality control to remove low-quality sequences, host sequences, and other sequences that need to be filtered. Megahit[10] (https://github.com/voutcn/megahit) was used for metagenomic assembly. After obtaining the genome contigs of each sample, BWA-MEM[11] (https://github.com/lh3/bwa) was used to map the reads back to each component with default parameters. The continuous depth was calculated using jgi_summarize_bam_contig_depths. And MetaWRAP[12] (https://github.com/bxlab/metaWRAP) was used by three different binning algorithms (metaBAT2[13], MaxBin2[14], CONCOCT[15]) to perform metagenomic binning and refinement for each sample individually to obtain MAGs (Metagenome-assembled Genomes). After MAGs quality assessment through CheckM[16] (https://github.com/Ecogenomics/CheckM) lineage_wf workflow, 16,785 high-quality MAGs with completeness > 90% and contamination < 5% were retained.

SNV analysis

About 1,018 species and at least three homogenous genomes generated the SNV catalog. For each species, bowtie2[17] (https://github.com/BenLangmead/bowtie2) was used to map all homogenous genomes to representative genomes. Only when alternative alleles were detected in at least two homogenous genomes, each SNV locus was included in the catalog, and the final SNV catalog was generated by unifying the SNV coordinates according to the position of the SNV coordinates in the representative genome of the species.

Genome clustering

Genome clustering was performed using dRep[18] (https://github.com/MrOlm/drep) with 95% genome distance, which is regarded as a rough estimate of a species[19]. All genomes were checked for completeness and contamination using CheckM. The genomes with completeness > 90% and contamination < 5% were retained and had a genome quality score (completeness – 5×contamination + 0.5×log(N50)). The genome with the highest quality score in each cluster were regarded as the representative genome for taxonomic assignment.

Taxonomic assignment

Genome taxonomic assignment was performed using the GTDB toolkit [20] (GTDB-Tk, version 06-RS202, https://github.com/Ecogenomics/GTDBTk), having its classification system different from the NCBI taxonomic database. The NCBI taxonomic names corresponding to the GTDB taxonomic names of all genomes were obtained using the script.

Pan-genome construction

Although representative genomes have the smallest size and maintain species-level resolution, it does not capture intraspecific diversity, which usually contains a large part of the genetic information in the microbial community. The genes predicted by Prokka[21] (https://github.com/tseemann/prokka) of each cluster were combined, and a pan-genome was generated based on clusters using the pan-genome pipeline Roary[22] (https://sanger-pathogens.github.io/Roary/) to integrate sequence information from all strains of the same species in a non-redundant manner.

Build reference database for profiling

Identifying microbial taxa present in complex organisms or environmental samples is one of the oldest and most common challenges in microbiology. Recently, many metagenomic classifiers have been developed to classify metagenomic data and estimate taxa abundance profiles.

When a server with a large amount of memory (> 100 Gb) is available, Kraken2[23] (http://ccb.jhu.edu/software/kraken/) and its derivatives Bracken[24] (https://ccb.jhu.edu/software/bracken/) provide good performance indicators; as long as the database load time is amortized, it can be used on a large number of samples very fast and allows the creation and use of custom databases. According the taxonomic

name and tree information from GTDB, we build the classification tree files (names.dmp and nodes.dmp), and assign the GTDB taxonomy ID to each sequence of all pan-genomes. All pan-genomes were combined to build Kraken2 and Bracken database.

For a low memory server, marker-based methods like mOTUs were better than Kraken2. The marker-based methods identify a clade-specific single-copy gene set, so identifying one of these genes can prove the associated clade members' presence. The tool fetchMG extracts marker genes from genomes and metagenomes quickly and accurately. We built an empty mOTUs reference database first, and marker genes extracted by fetchMG from the MBCN genome were added to the empty mOTUs reference database using the script for adding new genomes provided by mOTUs.

Online Platform

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 platform and guides the user through the platform. 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 https://www.microbiota.cn.

A set of data analysis and visualization tools was 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 were developed based on the R package microeco (version 0.3.3)[25].

Results

Overview of platform

The overall design and the flowchart of our platform consisted of four parts (Fig. 1):

1) The latest human gut metagenomic sequencing data from NCBI were collected regularly to construct MAGs.

2) Human gut microbial genomes were collected from public datasets and integrated with the assembled MAGs into the latest human gut microbial genome collection.

3) Reference databases of two metagenomic classifiers were built based on the human gut microbial genome collection for profiling: DNA-to-DNA tool Kraken2, and DNA-to-marker tool mOTUs. And the profiling results of 1,082 human gut metagenomic sequencing data with metadata obtained by Kraken2 reference database built based on MBCN were sorted out to establish the profile database.

4) An interactive online statistical analysis platform was established, which allows users to upload the relative abundance profile of species obtained using our database and the corresponding sample metadata for online visual analysis and graphing. Users can also combine the profile results from the profile database on the platform for meta-analysis.

Assemble genomes from metagenomic sequencing data

Human gut metagenomic sequencing data of 2,477 samples collected from people of different ages worldwide were downloaded, which belong to 13 different bio-projects. A standardized workflow was constructed, and 16,785 MAGs were assembled from downloaded human gut metagenomic sequencing data.

In order to record the single nucleotide variation (SNV) information in the assembled MAGs, the assembled genomes were aligned using bowtie2. A catalog containing 12,979,130 SNVs from 1,018 species with three or more homogenous genomes was generated. The *Actinobacteriota* phylum had the highest SNVs rate (1.11 SNV per kbp), followed by *Bacteroidota* with 0.813 SNV per kpb. Other main human gut phyla with a high rate of SNVs were *Proteobacteria* (0.649 SNV per kpb), *Firmicutes C* (0.622 SNV per kpb), *Firmicutes A* (0.517 SNV per kpb), and *Firmicutes* (0.441 SNV per kpb). Next, the detected SNVs were assigned to the continent of origin of each genome (Table 1). It is worth noting that the Asian genomes contributed to the most SNVs. However, the average contribution from the European genomes was about twice that of the Asian or North American genomes. Our results demonstrated high strain variability among continents, and a considerable diversity is yet to be discovered.

Table 1
SNVs detected in genomes from each continent.

Continent	Asia	Europe	North America
Number of SNVs	11457436	9475744	9665289
Number of SNVs per kbp	0.5653	1.0560	0.6464

Human Gut Microbiome Genome Collection

This study's human guts microbial genome collection, named MBCN (Metagenomic binning clustered nuclein), consisted of 14,166 clusters. Among the 14,166 representative MAGs, 8,673 were from Refseq, 4,374 from UHGG, and 1,119 from our assembled MAG.

The taxonomic classification of each cluster was assigned using GTDB-Tk, and most MBCN clusters (9,995 out of 14,166 clusters) consisted of a single MAG. On the other hand, the most diverse MBCN cluster comprises 8,314 different MAGs named *Escherichia coli_D*, followed by 7,511 MAGs related to *Agathobacter rectalis*.

GTDB did not name many genomes but named them after their genus, family, order, or class. Likewise, NCBI taxonomic annotations for many genomes lead to ambiguous names not specific to species, which greatly exacerbated the difference between the total species-level clusters (14,166 clusters in MBCN) and the total cluster names (12,450 GTDB names, 6,318 NCBI names). Many clusters share the same taxonomic annotations, such as 2,838 clusters share GTDB names with other clusters, and 8,587 clusters share NCBI names. This is especially true for the various *Collinsella* clusters, where 160 different clusters are named after GTDB *Collinsella* species while 471 clusters are named after NCBI *Collinsella* species.

Figure 2 demonstrates the phylogenetic distribution of the 13,853 bacterial and 313 archaeal species. *Proteobacteria* accounted for the largest number of branches (3,885 clusters). Other MAGs were distributed in 23 phyla, including the prominent human gut phyla, such as *Firmicutes A* (2,761 clusters), *Actinobacteriota* (2,504 clusters), *Bacteroidota* (1,842 clusters), and *Firmicutes* (1,639 clusters). Unassigned species-level GTDB names clusters were distributed in 21 phyla, such as *Firmicutes A* (587 clusters), *Actinobacteriota* (214 clusters), and *Bacteroidota* (138 clusters), *Firmicutes* (92 clusters), and *Proteobacteria* (69 clusters). In the main human gut phyla, more than 21% of the diversity of *Firmicutes A* was contributed by unassigned species-level GTDB names clusters and less than 10% in other main human gut phyla. Our assembly workflow assembled representative genomes of 191 unassigned species-level GTDB names cluster, distributed in 11 phyla. Most newly assembled unassigned species-level GTDB names representative genomes of clusters belonged to *Firmicutes A* (88 clusters), followed by *Actinobacteriota* (42 clusters). In our assembled MAGs, some phyla had few genomes but were closely associated with health. *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 was reported to be related to disease[3].

Evaluating Kraken2 Database Performance by Simulated Reads

Data containing simulated Illumina reads from 100 genomes was used to evaluate the performance of the two reference databases. This data contains 100 genomes from 84 representative species, previously used to compare metagenomic assembly algorithms, and is freely available at http://www.bork.embl.de/~mende/simulated_data[26].

First, Kraken2 was used to assign reads. The unassigned reads using the MBCN database accounted for 9.03%, while the unassigned reads using the UHGG database accounted for 61.85%. Many metagenomic classifiers, including Kraken2, will report species not present, which can result in thousands of low-abundance FP (False Positive) species predictions and require researchers to filter low-abundance predictions. Kraken2 profiling results were adjusted at the species and genus levels using Bracken and compared the two databases' profiling results with different minimum species abundance thresholds (detection limits). Figure 3 demonstrates the performance of the Kraken2 database built by MBCN compared to the UHGG collection at species and genus levels. At both species and genus levels, the number of FPs found by MBCN was much smaller than that of UHGG (Fig. 3A). The false discovery rate FDR = FP/(TP + FP) is the proportion of FP in the detection results. Likewise, the FDR of MBCN was much smaller than that of UHGG (Fig. 3B). And we found that setting the detection limit at 0.025% could filter out most false positives.

The performance of both databases in predicting the relative abundance of species was evaluated (Table 2). The sum of the relative abundances of false positive and negative species of MBCN was much smaller than that of UHGG, with a higher recall rate (Recall = TP/(TP + TN)). The L1 (Manhattan) distance is the sum of the absolute errors of all ground truth and predicted species, and it provides a measure that incorporates FP predictions. The L1 distance was chosen because it does not give extra weight to high abundance species, and precision and recall are reported independently. MBCN possessed a smaller L1 distance and demonstrated that MBCN outperforms UHGG in estimating relative abundance, concordant to previous results.

Table 2
Performance of databases in predicting the relative abundance

Taxonomic level	Genus		Species	
Database	MBCN	UHGG	MBCN	UHGG
Relative abundance of FP	2.74%	8.33%	9.80%	23.76%
Relative abundance of FN	4.67%	54.13%	14.11%	66.93%
Recall	92.65%	29.41%	82.14%	41.67%
L1 distance	0.4928	1.0820	0.5567	1.2422

Comparison of the different databases on human gut metagenomic samples

To further investigate the performance of different databases in practical applications, data from PRJNA453965, a study that used shotgun metagenomic analysis to compare gut microbial communities between breast cancer patients and healthy controls, was used [27]. A total of 133 stool samples were obtained from premenopausal breast cancer patients (n = 18), healthy premenopausal controls (n = 25), postmenopausal breast cancer patients (n = 44), and healthy postmenopausal controls (n = 46), previously aligned with a reference catalog of the human gut microbiome (IGC), yielding a relative abundance profile. The community profiles generated using Kraken2 databases built based on UHGG and MBCN were compared with profile obtained using IGC database.

Since the community composition of the fecal samples was unknown, other measurable aspects of the community profiles produced by each metagenomic classifier were assessed. The percentage of reads assigned to a species by MBCN was much higher than UHGG and IGC, attributed to the addition of a large number of genomes from RefSeq in MBCN that do not exist in the other reference databases. Table 3 demonstrates the performance of three metagenomic reference databases. As expected, several species were reported by MBCN and UHGG, more than those reported by IGC, the vast majority of which may be low-abundance FPs. Species with an estimated mean relative abundance < 0.025% were removed as these are expected to be predominantly FP predictions based on the results of the simulated communities. Although UHGG reported several species, the reported abundance of ≥ 0.025% was lower than that of MBCN, while IGC still reported the least number.

Table 3
Comparison of metagenomic reference databases on 133 stool samples.

Database	IGC	MBCN	UHGG
Assigned reads	50.59 ± 13.45%	89.77 ± 2.21%	72.39 ± 4.29%
Number of reported species(≥ 0.025%)	57.8 ± 8.6	91.1 ± 25.5	88.2 ± 22.9

Further, the difference in gut microbes between breast cancer patients and healthy controls was assessed (mean relative abundance ≥ 0.025%). The results reported by UHGG and IGC demonstrated no significant differential species between premenopausal breast cancer patients and healthy premenopausal controls (p-value > 0.05, Wilcoxon rank-sum test). In contrast, the results reported using MBCN demonstrated that six species were significantly different in premenopausal samples, all enriched in healthy controls(Table 4). *Sutterella parvirubra*[28] and *Parabacteroides gordonii*[29] were reported to be enriched in health.

Table 4
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 sp000435595	0.000094	0.013586	0.001221	0.005996	2.20E-07	6.60E-07
Butyricimonas sp900184685	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 sp000434975	0.001655	0.044013	0.010765	0.053139	3.04E-05	1.26E-04
Prevotella sp900313215	0.001001	0.032244	0.009081	0.034932	3.77E-05	1.35E-04

Figure 4 demonstrates that MBCN has the potential to report more differential species in postmenopausal breast cancer patients and healthy postmenopausal controls. MBCN reported 35 differential species, much higher than UHGG (15 species) and IGC (7 species). MBCN included 71% species reported by IGC and 67% species reported by UHGG. **Table S1** demonstrates the details of differential species reported by these database. MBCN contains more genes to identify species not included in other databases than UHGG (4744 species) and IGC (3449 species). MBCN reported differential species to be associated with human health between postmenopausal breast cancer patients and healthy postmenopausal controls, such as *Enterobacter hormaechei*[30], *Anaerostipes hadrus*[31], *Phocaeicola* species[32]. Alistipes were markedly altered in mice post-injected with breast cancer cells[33]. The number of reads assigned to most species was higher by MBCN than other databases, resulting in more differential species, whose average relative abundance in other databases was < 0.025%, such as *F23-B02 sp000431075*, *CAG-349 sp003539515*, *Prevotella sp000436915*. After FDR correction, some species like *Megamonas funiformis*, *CAG-41 sp900066215*, and *CAG-882 sp003486385* showed significant differences in the Wilcoxon test (p-value < 0.05), and there were also species whose differences were insignificant, such as *Eubacterium_R sp000434995*, *Lachnospira* sp003537285, *Odoribacte splanchnicus*. Not all species identified an increase in the number of reads in the MBCN results. Increased reference sequences and variations in some species and genera lead to the reassignment of reads. For example, some reads assigned initially to *Escherichia fergusonii* in UHGG were assigned to *Escherichia coli* in MBCN. Thus, reducing the relative abundance of *Escherichia fergusonii* to below 0.025% in the MBCN results.

Profile database

In order to provide users of our platform with metagenomic abundance profiles generated according to the same standard process, we collected metadata of 1,082 human gut metagenomic sequencing samples, and obtained profiles using the Kraken2 reference database built based on MBCN. The metadata and profiles were sorted to establish a profile database on our platform. Users can filter samples by metadata in the web of platform, and obtain a table of profiling results, which can be downloaded or directly imported into the analysis steps for subsequent statistical and visualization analysis.

Online Analysis

To illustrate the on-line analysis utility of MicrobiotaCN, we use the analysis tool of the platform to analyze the data of the above breast cancer study (PRJNA453965) as an example. The profile obtained earlier and metadata were uploaded to the analysis module of MicrobiotaCN to display the utility of common statistical and visualization tools for microbiome research on the MicrobiotaCN platform.

Data Import

In the data input module, users can submit four tables to perform online visual analysis, in which 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, like "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 namse of the abundance profile table and the corresponding grouping table. Factor metadata containing categorical metadata information is necessary, while the numeric metadata information (Numeric metadata) is required only for environmental factor analysis. Files can be uploaded as tab-delimited text (.txt) or comma-separated values (.csv). Users can go to the related FAQs and tutorials for more details or try our test examples.

Species composition display

After data filtering, this module can use histograms, box plots, pie charts, and heat maps to display the species abundance of each classification level, and groups were displayed separately.

Figure 5 demonstrates the phylum-level relative abundance of PRJNA453965 project profiled using the Kraken2 database built by MBCN using histograms, pie charts, box plots, and heat maps, respectively, and according to postmenopausal Breast cancer patients (Post + Case), healthy postmenopausal controls (Post + Control), premenopausal breast cancer patients (Pre + Case) and healthy premenopausal controls (Pre + Control) demonstrated separately.

Community diversity analysis

Community diversity analysis was mainly implemented based on the R vegan packages and performed at different classification levels depending on the available annotations. The alpha-diversity analysis function currently supports five common diversity measures. It can automatically estimate the corresponding statistical significance, while beta-diversity analysis supports three common distance measures, and results are presented based on principal component analysis (PCA), principal coordinate analysis (PCoA), or non-metric multidimensional scaling (NMDS). Figure 6 demonstrates the analysis of alpha-diversity and beta-diversity for the PRJNA453965 project.

Differential analysis and biomarker identification

The module allows users to perform differential analysis to identify features that differ significantly between groups and provides four methods: Metastats, Wilcoxon rank-sum test, LEfSe, and random forest.

Metastats[34] are the first statistical methods explicitly developed to address questions posed in clinical research. Metastats compare metagenomic samples (expressed as counts of individual characteristics, such as species, genes, and functional groups) from two different populations (e.g., disease patients versus healthy controls) and identify features that distinguish the two populations statistically. It integrates non-parametric multiple testing and p-value correction, and thus, only two groups can be analyzed. Figure 7A demonstrates a boxplot of the ten species with a large difference between postmenopausal breast cancer patients (Post + Case) and healthy postmenopausal controls (Post + Control) using Metastats analysis.

The results of the Wilcoxon rank-sum test are displayed using a volcano plot. A volcano plot is a scatter plot that combines a measure of statistical significance (such as a p-value) in a statistical test with the magnitude of change, which can quickly and intuitively identify data with large and statistically significant changes point. Figure 7B demonstrates the results of the Wilcoxon rank-sum test in postmenopausal breast cancer patients (Post + Case) and healthy postmenopausal controls (Post + Control). The vertical axis represents the p-value of the Wilcoxon rank-sum test (transformed by-Log 10), and the horizontal dotted line indicates the dividing line with a 0.05 p-value. The horizontal axis represents the fold change of the difference between groups (transformed by Log 2), and if the upper limit of the fold change of difference exceeds, it is adjusted to the upper limit of the fold of difference. Users can set the fold change threshold (vertical dotted line) to exclude species with small fold change from significantly different species.

LEfSe[35] (Linear discriminant analysis Effect Size) uses the non-parametric Kruskal-Wallis rank-sum test to detect features with significantly different abundances in different groups. Next, it performs linear discriminant analysis to assess the effect size of these significant features to determine the most likely characteristics (organisms, clades, OTUs, genes, or functions) that explain differences between classes. A combination of p-values and effect sizes can be used to select important features. Compared with Metastats, LEfSe can analyze more than two groups using the Kruskal-Wallis rank-sum test and directly perform statistical tests and difference analysis on different classification levels simultaneously. The results of LEfSe consist of three parts, as demonstrated in Fig. 7C-E: a histogram of the distribution of LDA scores of significantly different clades, used to display the significantly enriched species in each group and their importance(Fig. 7C); the box plot of the relative abundance of significantly different clades in different groups(Fig. 7D); and the taxonomic cladogram to demonstrate the hierarchical taxonomic distribution of marker species significantly enriched in each group of community samples(Fig. 7E).

Random forest is a general-purpose non-parametric machine learning algorithm that performed well in many recent analyses and classifications of microbiome data. It uses an ensemble of classification trees (forests) and makes class predictions based on the majority vote of the ensemble. Building a forest provides an unbiased estimate of classification error by aggregating cross-validation results using bootstrap samples. The algorithm also weighs the importance of each feature in terms of the increase in classification error upon permutation. Graphical output was generated to summarize its classification performance in terms of tree increase (Fig. 7F). The left side of the figure demonstrates the importance of features in postmenopausal breast cancer patients (Post + Case) and healthy postmenopausal controls (Post + Control). MeanDecreaseAccuracy means changing the value of a variable into a random number, a random forest. The larger the degree of decrease in the prediction accuracy, the greater the importance of the variable. In addition, on the right side is the corresponding species abundance box plot of different groups.

Other Analysis

In addition, the platform also provides other analysis modules for other analysis requirements, sush as network analysis, cluster analysis, and environmental analysis.

The network analysis module allows users to calculate the correlation of different species and draw network diagrams that can be saved in HTML format and interacted with a mouse. This module can draw two different graphs: chord graph and NetworkD3 dynamic network graph. The user 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 used to perform unsupervised clustering of samples. The cluster index tab provides a function for selecting the number of clusters for use in the next cluster tab.

The environmental analysis module performs RDA and db-RDA analysis on continuous environmental factors to analyze their impact on taxonomic abundance.

Discussion

MicrobiotaCN is a platform with a standardized process to build a profiling reference database, organize profiles, and provide online interactive analysis to use the profiles obtained by a unified process for meta-analysis. Although there are many metagenomic analysis platforms like MicrobiomeAnalyst[8], the functions of these platforms are limited to providing downstream data analysis. MicrobiotaCN provides a unified reference database and uses it to obtain profiles so that profiles from different studies can be compared and analyzed through this platform. In addition, it provides a more comprehensive analysis and a brand new analysis module. For example, it only provides PCoA and NMDS methods for beta-diversity analysis and LEfSe and random forests for differential analysis and biomarker identification. It provides additional PCA methods for beta-diversity analysis, while metastats and Wilcoxon rank-sum test for differential analysis. MicrobiotaCN provides an environmental analysis module with RDA and db-RDA to analyze the impact of continuous environmental factors on taxonomic abundance, which is unavailable on MicrobiomeAnalyst.

MBCN is a collection of the genomes for profiling reference database construction and can be used as a global reference for bacteria in the intestinal tract of healthy people. The profiling reference database is available for download at MicrobiotaCN. The profiling reference database built by MBCN is more accurate and comprehensive than other intestinal microbial genome collections. It uses 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 used pan-genomes to build a reference database instead of a single representative genome, covering as much genetic information as possible for the species. The introduction of high-quality genomes from Refseq dramatically increases the variety of species included, helping to reduce unclassified sequences. Some studies constructed genome collections based on Refseq and UHGG genomes, such as HumGut[36]. HumGut collected healthy human gut metagenomes with 5,170 species at 95% of ANI, while MBCN collected the gut genomes of non-healthy people not included in HumGut. HumGut only provides the FASTA files for all HumGut genomes rather than the reference database for profiling, which means that users need to build the reference database. Compared to other studies, our study constructed a reference database and downstream metagenomic analysis and innovatively developed a platform for the entire metagenomic analysis pipeline based on the constructed genome collection, providing a directly available reference database for profiling.

The plethora of genetic information also poses some problems, the foremost being that the size and resources required of the database are increased significantly. A marker-based mOTUs database was built beside the Kraken2 database for users with low computing resources. Nevertheless, the results of different databases cannot combine to analyze. One challenge that remained was the naming of species in our genome collection. There was a profound inconsistency between the total number of species-level clusters and the number of names

annotated. A higher-order taxonomy it could identify was used for clusters without precise taxonomy. Thus, the abundance of this cluster was not analyzed at the species level. Files were prepared to build a custom Kraken2 database, in which all MBCN clusters received artificial TAXIDs, classified as clusters instead of classifications.

In future work, we will also expand our method to more disease-related genomes and metagenomes and continue collecting human gut metagenomic samples and profiles with our database. As for the analysis module, more analysis functions and plot interaction options will be provided.

Conclusion

As the frontier in biomedical research, there are still problems in current metagenomic research and data analysis. Although many new software and algorithms have been developed in recent years, there are no uniform standards. The microbiota is complex and dynamic, and to fully understand its behavior as a system and its interactions with the host requires the collection, analysis, and integration of data from different sources. A unified reference database can obtain metagenomic data pairs; thus, it is vital for researchers in related fields. This platform meets the research requirements by building a unified reference database, corresponding profile database, and online interactive analysis functions. The metagenomic analysis platform constructed in this study will provide a much-needed standardized approach to human gut microbiome analysis, filling a critical gap in current microbiome research.

Declarations

Authors' contributions

BZ, JQ , YC and YJ conceived the study. BZ is the principle investigator. BZ, JX and JQ worked out the technical aspects of the paper. TF and LL conducted data management and bioinformatics analysis. BZ was a major contributor in writing the manuscript. YC and YJ revised the manuscript critically. All authors read and approved the final manuscript.

Funding

This work was supported by Shenzhen Progression and Reform Committee (No. 2019156) and Department of Science and Technology of Guangdong Province (No. 2017B030314083).

Availability of data and material

The reference database built based on MBCN genome collection can be found at http://www.microbiota.cn/. This also includes metadata and profiles obtained by Kraken2 databases built based on MBCN genome collection.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

Not applicable

References

1. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 2012;490(7418):55–60.
2. Feng Q, Liang S, Jia H, Stadlmayr A, Tang L, Lan Z, et al. Gut microbiome development along the colorectal adenoma-carcinoma sequence. Nat Commun. 2015;6:6528.
3. Barandouzi ZA, Starkweather AR, Henderson WA, Gyamfi A, Cong XS. Altered Composition of Gut Microbiota in Depression: A Systematic Review. Front Psychiatry. 2020;11:541.

4. Macchione IG, Lopetuso LR, Ianiro G, Napoli M, Gibiino G, Rizzatti G, et al. Akkermansia muciniphila: key player in metabolic and gastrointestinal disorders. Eur Rev Med Pharmacol Sci. 2019;23(18):8075–83.
5. Thomas AM, Segata N. Multiple levels of the unknown in microbiome research. BMC Biol. 2019;17(1):48.
6. Nayfach S, Shi ZJ, Seshadri R, Pollard KS, Kyrpides NC. New insights from uncultivated genomes of the global human gut microbiome. Nature. 2019;568(7753):505–10.
7. Parks DH, Rigato F, Vera-Wolf P, Krause L, Hugenholtz P, Tyson GW, et al. Evaluation of the Microba Community Profiler for Taxonomic Profiling of Metagenomic Datasets From the Human Gut Microbiome. Front Microbiol. 2021;12:643682.
8. Dhariwal A, Chong J, Habib S, King IL, Agellon LB, Xia J. MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. Nucleic Acids Res. 2017;45(W1):W180-W8.
9. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884-i90.
10. Li D, Luo R, Liu CM, Leung CM, Ting HF, Sadakane K, et al. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods. 2016;102:3–11.
11. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589–95.
12. Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome. 2018;6(1):158.
13. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ. 2019;7:e7359.
14. Wu YW, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics. 2016;32(4):605–7.
15. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic contigs by coverage and composition. Nat Methods. 2014;11(11):1144–6.
16. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25(7):1043–55.
17. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
18. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. 2017;11(12):2864–8.
19. Jain C, Rodriguez RL, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9(1):5114.
20. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. 2019.
21. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068–9.
22. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31(22):3691–3.
23. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20(1):257.
24. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data. PeerJ Computer Science. 2017;3.
25. Liu C, Cui Y, Li X, Yao M. microeco: an R package for data mining in microbial community ecology. FEMS Microbiol Ecol. 2021;97(2).
26. Mende DR, Waller AS, Sunagawa S, Jarvelin AI, Chan MM, Arumugam M, et al. Assessment of metagenomic assembly using simulated next generation sequencing data. PLoS One. 2012;7(2):e31386.
27. Zhu J, Liao M, Yao Z, Liang W, Li Q, Liu J, et al. Breast cancer in postmenopausal women is associated with an altered gut metagenome. Microbiome. 2018;6(1).
28. Hiippala K, Kainulainen V, Kalliomaki M, Arkkila P, Satokari R. Mucosal Prevalence and Interactions with the Epithelium Indicate Commensalism of Sutterella spp. Front Microbiol. 2016;7:1706.
29. Ezeji JC, Sarikonda DK, Hopperton A, Erkkila HL, Cohen DE, Martinez SP, et al. Parabacteroides distasonis: intriguing aerotolerant gut anaerobe with emerging antimicrobial resistance and pathogenic and probiotic roles in human health. Gut Microbes. 2021;13(1):1922241.
30. Li JV, Ashrafian H, Bueter M, Kinross J, Sands C, le Roux CW, et al. Metabolic surgery profoundly influences gut microbial-host metabolic cross-talk. Gut. 2011;60(9):1214–23.

31. Zeevi D, Korem T, Godneva A, Bar N, Kurilshikov A, Lotan-Pompan M, et al. Structural variation in the gut microbiome associates with host health. Nature. 2019;568(7750):43–8.
32. Chrisman BS, Paskov KM, Stockham N, Jung JY, Varma M, Washington PY, et al. Improved detection of disease-associated gut microbes using 16S sequence-based biomarkers. BMC Bioinformatics. 2021;22(1):509.
33. Zhang J, Lu R, Zhang Y, Matuszek Z, Zhang W, Xia Y, et al. tRNA Queuosine Modification Enzyme Modulates the Growth and Microbiome Recruitment to Breast Tumors. Cancers (Basel). 2020;12(3).
34. White JR, Nagarajan N, Pop M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput Biol. 2009;5(4):e1000352.
35. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, et al. Diversity of the human intestinal microbial flora. Science. 2005;308(5728):1635–8.
36. Hiseni P, Rudi K, Wilson RC, Hegge FT, Snipen L. HumGut: a comprehensive human gut prokaryotic genomes collection filtered by metagenome data. Microbiome. 2021;9(1):165.

Figures

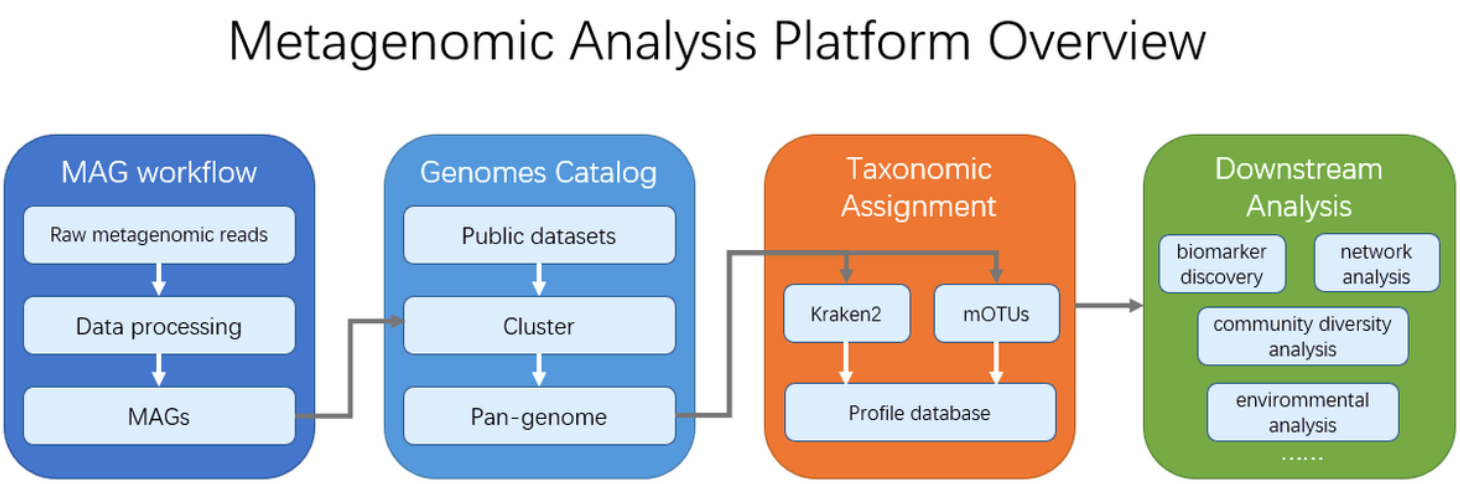

Figure 1

Overview of metagenomic dat a platform.

Our platform consisted of four parts: 1) The workflow to obtain MAGs; 2) The construction of genome collection; 3) Building the reference database of taxonomic classification tools and profile database; 4) Establishment of an interactive online statistical analysis platform.

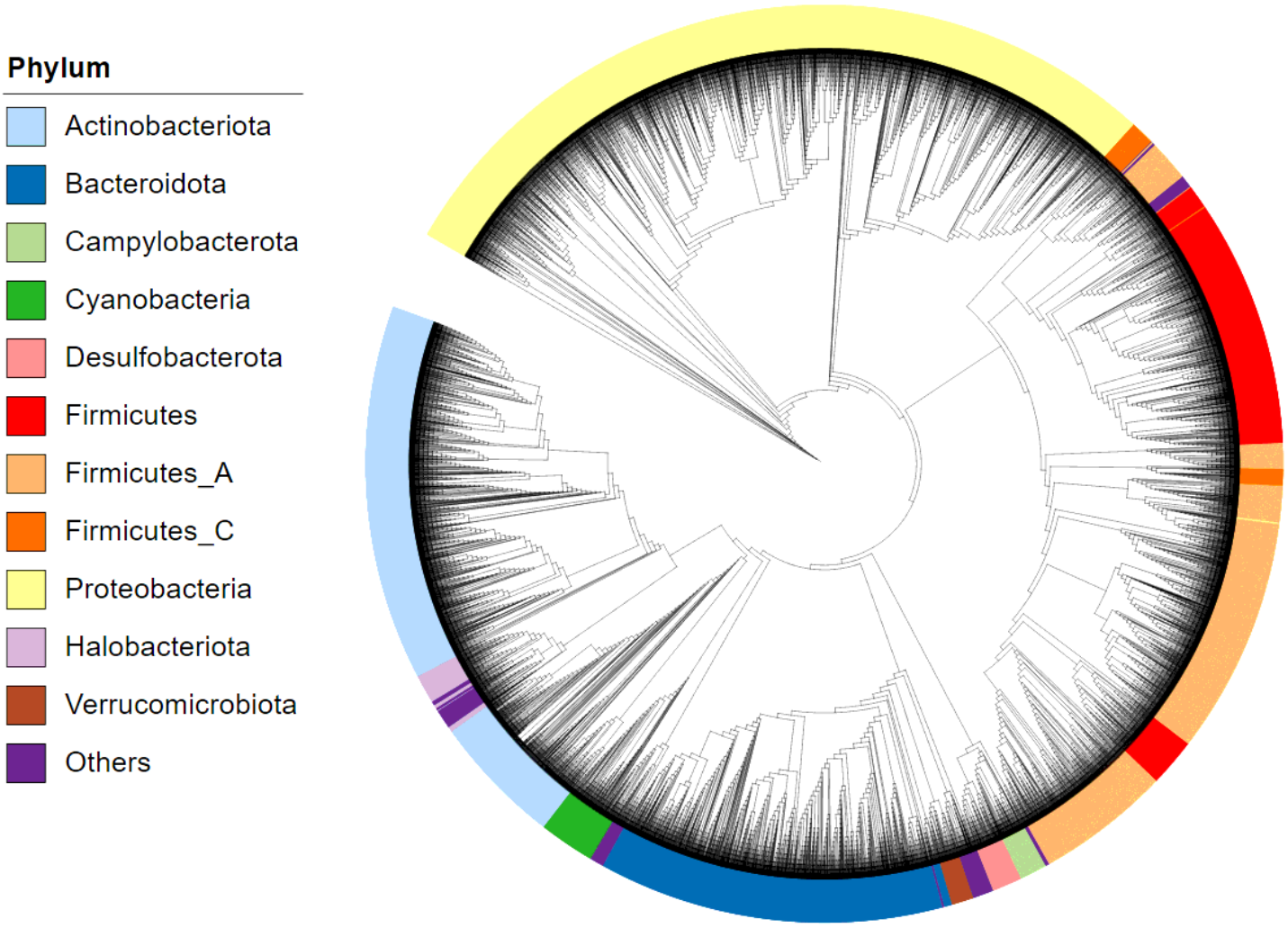

Figure 2

Phylogenetic tree of representative genomes of MBCN established by PhyloPhlAn.

The circular tree is drawed using iTOL (https://itol.embl.de/) with option “Ignore branch lengths”. Clades and outer circles are colored by GTDB phylum annotation.

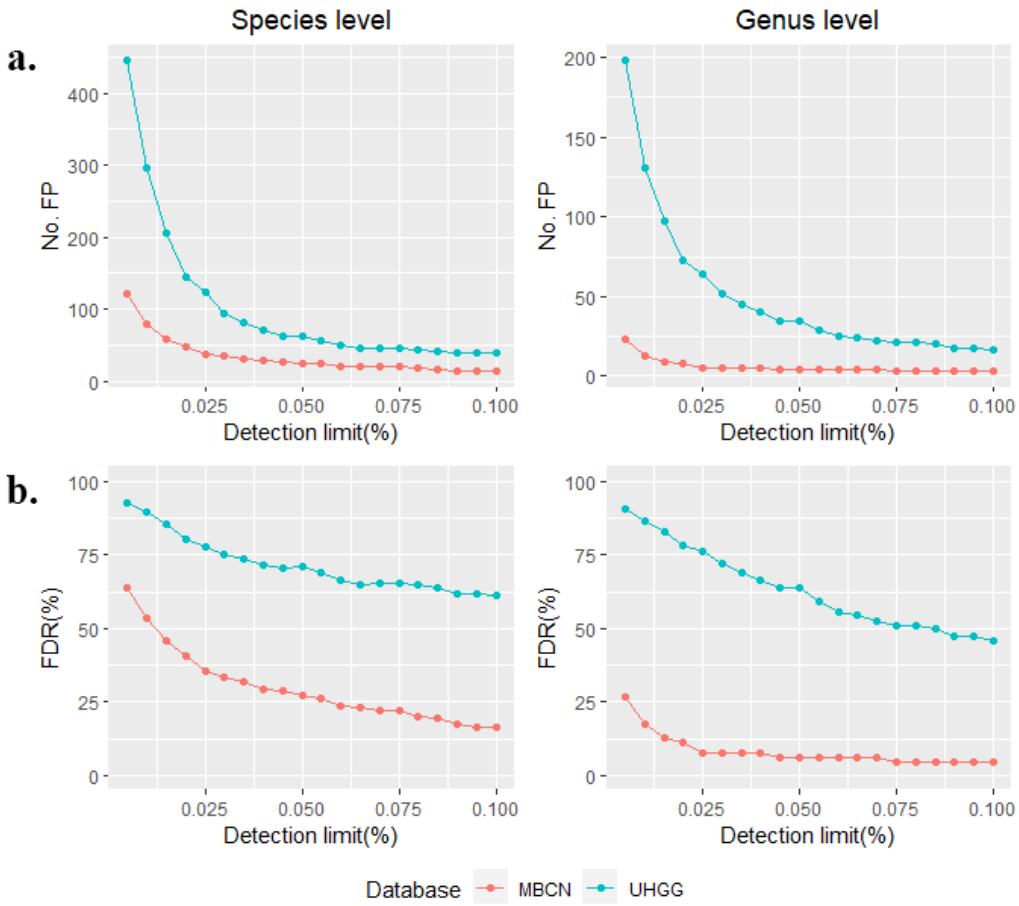

Figure 3

The performance of Kraken2 database built by MBCN in comparison to UHGG collection.

a. Number of FPs of profiles obtained using Kraken2 database built by MBCN and UHGG at species (left figure) and genus (right figure) level with different minimum species abundance thresholds (detection limits).

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

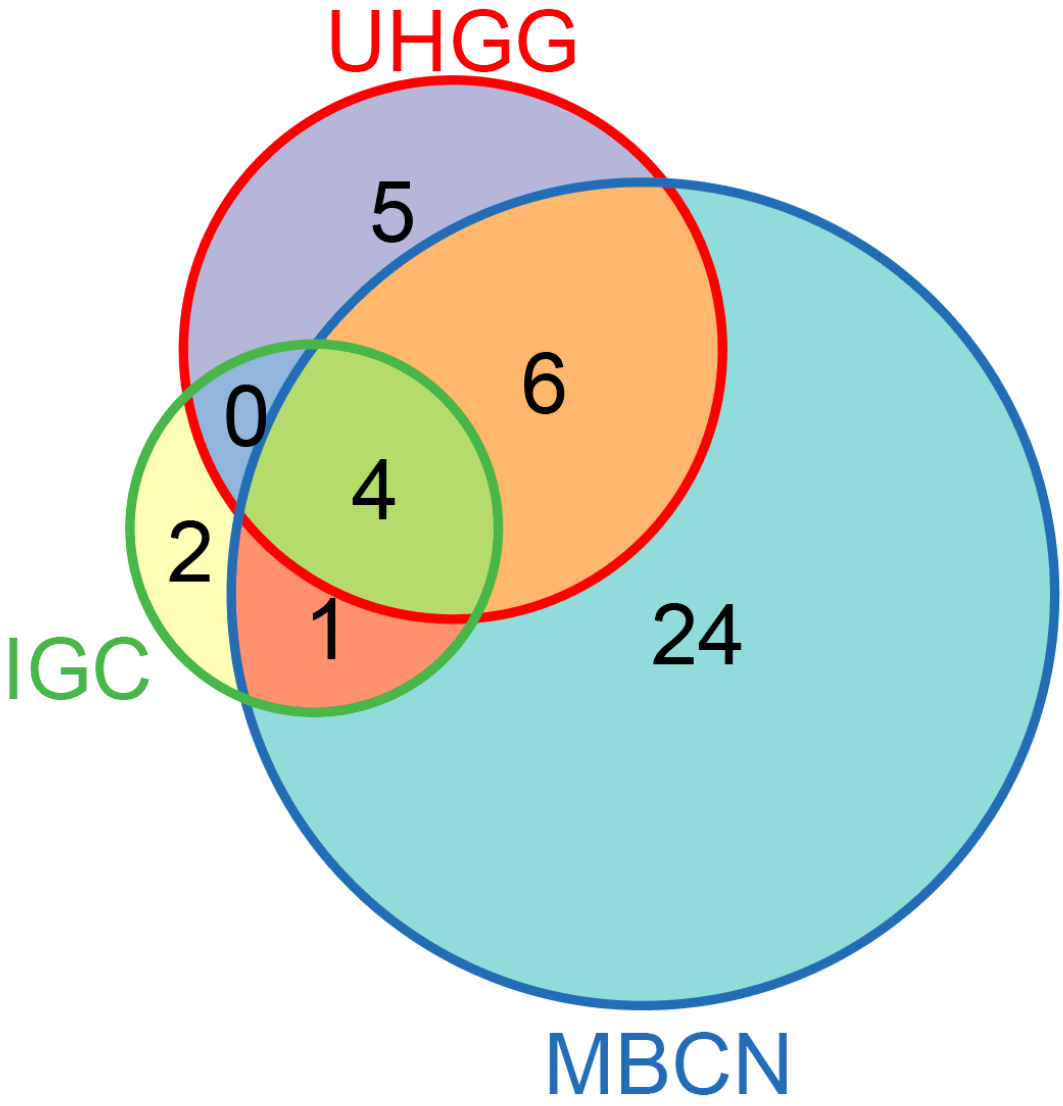

Figure 4

Significantly differential species between postmenopausal breast cancer patients and healthy postmenopausal controls in PRNAJ453965.

Venn diagram shows the overlaps of significantly differential species between postmenopausal breast cancer patients and healthy postmenopausal controls among the profile results using IGC, MBCN and UHGG.

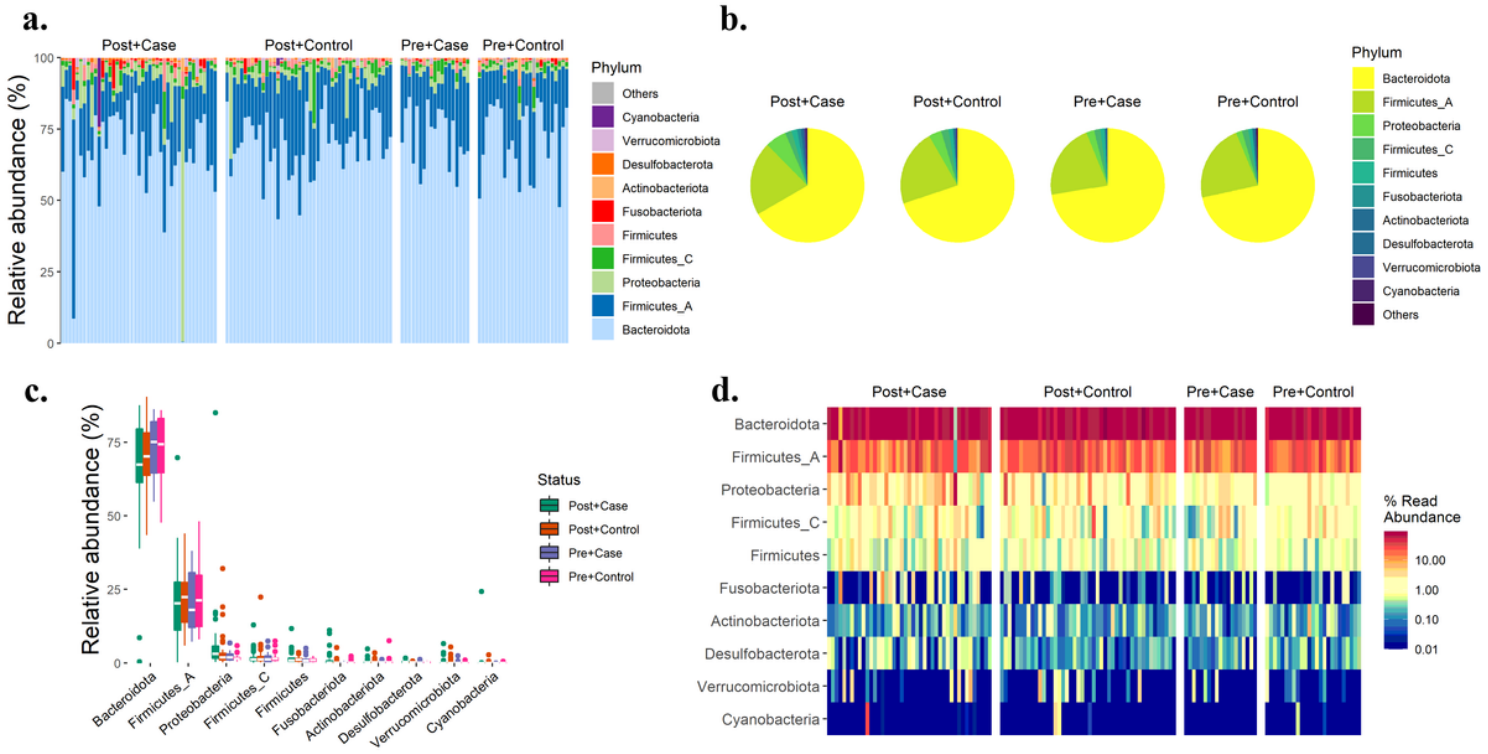

Figure 5

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.

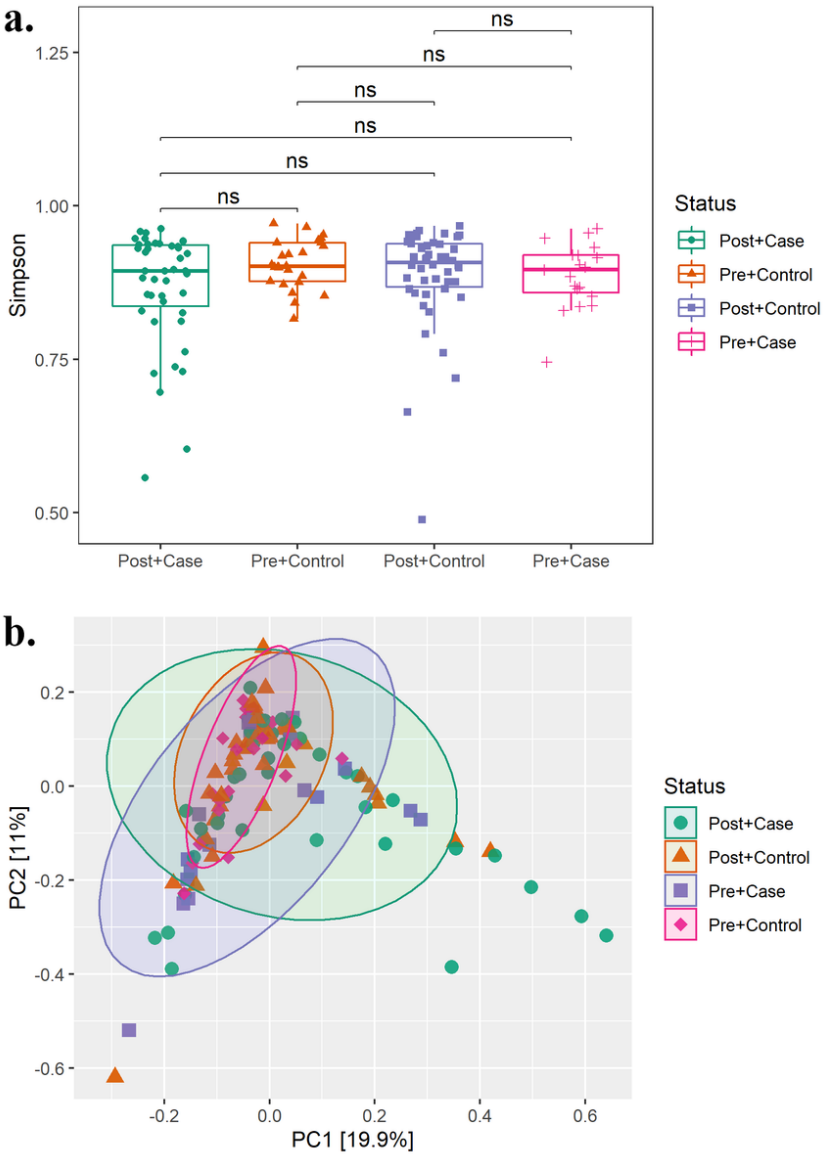

Figure 6

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. PCA with Bray-Curtis distance. The color and shape of dots indicate the group. The ellipses represent the 95% confidence interval for groups.

Figure 7

Sample figures of differential analysis of PRJNA453965.

a. Relative abaundace 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. 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.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1.xlsx