ARTICLE

www.the-innovation.org

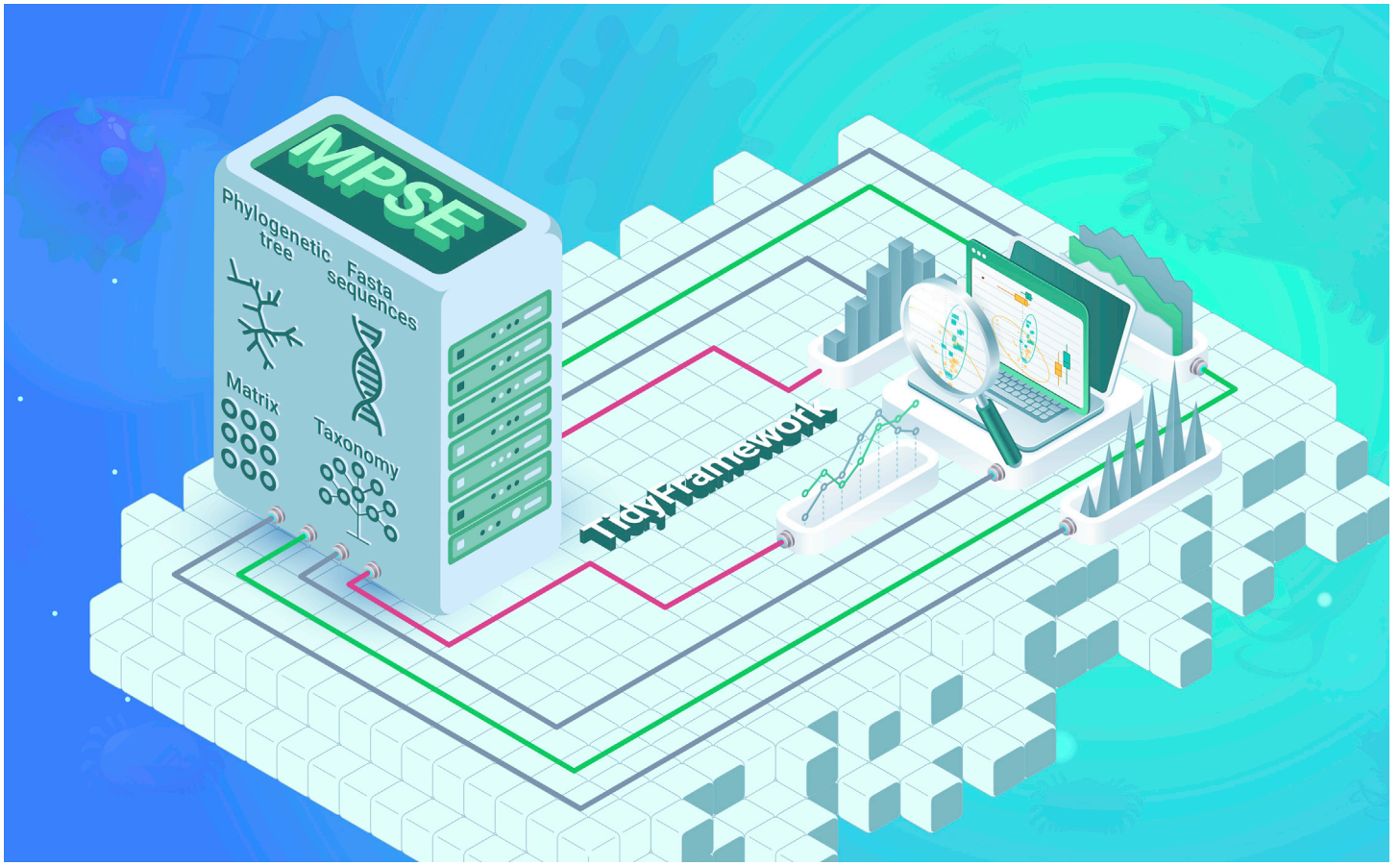
MicrobiotaProcess: A comprehensive R package for deep mining microbiome

Shuangbin Xu,[1,2] Li Zhan,[2] Wenli Tang,[2] Qianwen Wang,[2] Zehan Dai,[2] Lang Zhou,[1,2] Tingze Feng,[2] Meijun Chen,[2] Tianzhi Wu,[2] Erqiang Hu,[2] and Guangchuang Yu[1,2,*]

*Correspondence: gcyu1@smu.edu.cn

Received: August 19, 2022; Accepted: January 30, 2023; Published Online: February 2, 2023; https://doi.org/10.1016/j.xinn.2023.100388

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- *MicrobiotaProcess* is a bioinformatics tool for microbiome profiling.
- *MicrobiotaProcess* defines an *MPSE* structure to better integrate both primary and intermediate microbiome datasets.
- *MicrobiotaProcess* provides a set of functions under a unified tidy framework, which helps users to explore related datasets more efficiently.
- *MicrobiotaProcess* improves the integration and exploration of downstream data analysis.
- *MicrobiotaProcess* offers many visual methods to quickly render clear and comprehensive visualizations that reveal meaningful insights.

ARTICLE

MicrobiotaProcess: A comprehensive R package for deep mining microbiome

Shuangbin Xu,[1,2] Li Zhan,[2] Wenli Tang,[2] Qianwen Wang,[2] Zehan Dai,[2] Lang Zhou,[1,2] Tingze Feng,[2] Meijun Chen,[2] Tianzhi Wu,[2] Erqiang Hu,[2] and Guangchuang Yu[1,2,*]

[1]Division of Laboratory Medicine, Microbiome Medicine Center, Zhujiang Hospital, Southern Medical University, Guangzhou 510515, China

[2]Department of Bioinformatics, School of Basic Medical Sciences, Southern Medical University, Guangzhou 510515, China

*Correspondence: gcyu1@smu.edu.cn

Received: August 19, 2022; Accepted: January 30, 2023; Published Online: February 2, 2023; https://doi.org/10.1016/j.xinn.2023.100388

Citation: Xu S., Zhan L., Tang W., et al., (2023). *MicrobiotaProcess:* A comprehensive R package for deep mining microbiome. The Innovation **4(2)**, 100388.

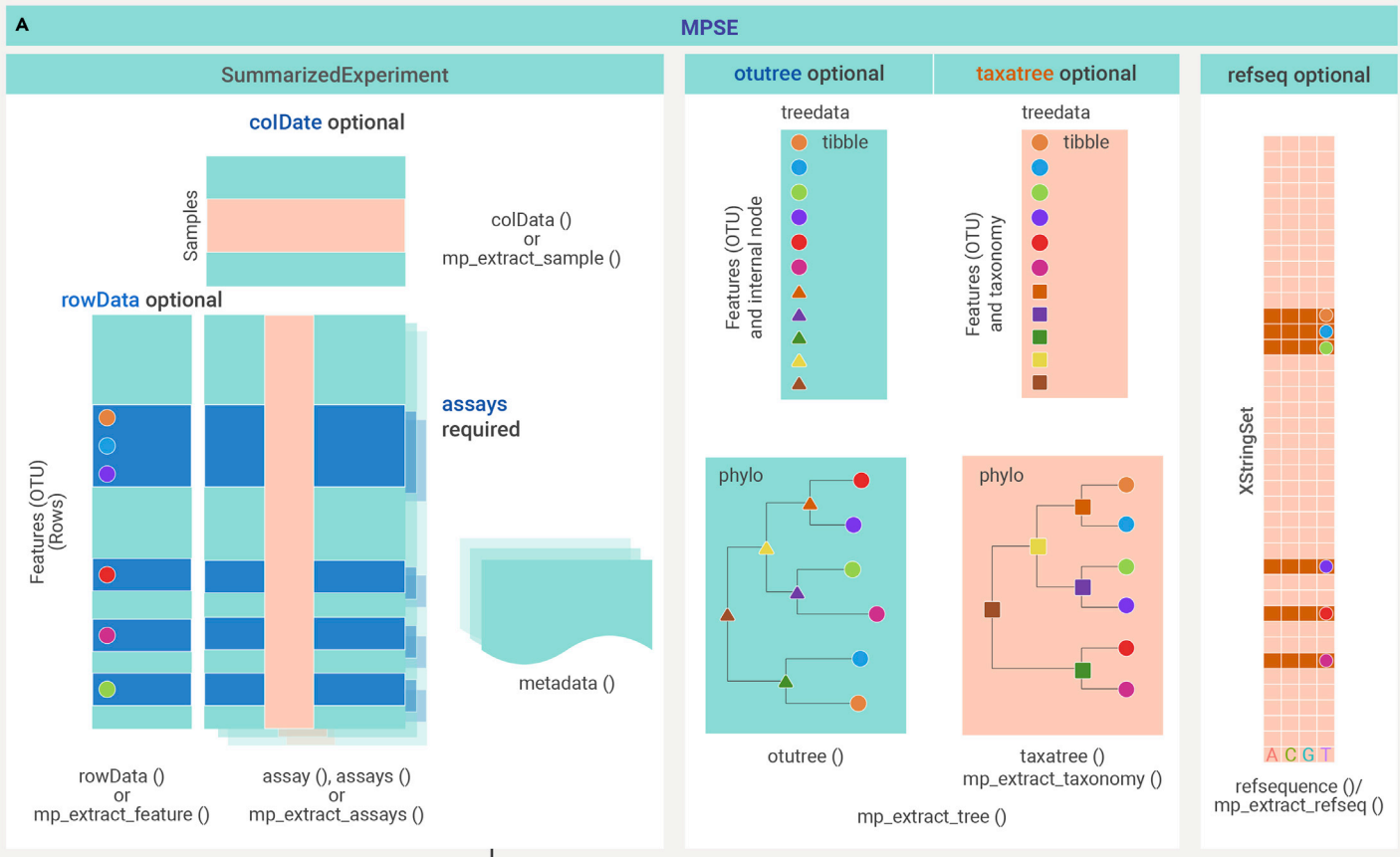
The data output from microbiome research is growing at an accelerating rate, yet mining the data quickly and efficiently remains difficult. There is still a lack of an effective data structure to represent and manage data, as well as flexible and composable analysis methods. In response to these two issues, we designed and developed the *MicrobiotaProcess* package. It provides a comprehensive data structure, *MPSE*, to better integrate the primary and intermediate data, which improves the integration and exploration of the downstream data. Around this data structure, the downstream analysis tasks are decomposed and a set of functions are designed under a tidy framework. These functions independently perform simple tasks and can be combined to perform complex tasks. This gives users the ability to explore data, conduct personalized analyses, and develop analysis workflows. Moreover, *MicrobiotaProcess* can interoperate with other packages in the R community, which further expands its analytical capabilities. This article demonstrates the *MicrobiotaProcess* for analyzing microbiome data as well as other ecological data through several examples. It connects upstream data, provides flexible downstream analysis components, and provides visualization methods to assist in presenting and interpreting results.

INTRODUCTION

A wide array of important roles of the microbiota in diverse environments have been investigated and explored substantially,[1,2] largely because of the development of high-throughput sequencing technologies and bioinformatics. During the last decades, many bioinformatics algorithms and tools for the exploration and analysis of microbiome data have been built in the scientific community, such as *qiime2*,[3] *dada2*,[4] *usearch*,[5] *mothur*,[6] *MetaPhlAn*.[7,8] These pipelines or tools conduct initial bioinformatics analysis of microbiome data, but the relevant data to a microbiome experiment became heterogeneous consisting of feature-oriented (operational taxonomic unit [OTU] or amplicon sequence variants [ASV]) data frames, sample-oriented data frames, the representative sequences, and the phylogenetic tree of the sequences after the initial analysis, which brings new challenges for the downstream data analysis and reproducibility. There are often many important links (such as the features of the feature table and the nodes of the phylogenetic tree, feature table, metadata, etc.) among the diverse data that need to be preserved throughout an analysis. If a researcher failed to integrate these data comprehensively, the information might be lost and thus generate errors. In addition, many intermediate steps that may confuse may be repeatedly performed in the downstream statistical analysis. For instance, the dissimilarity indices (such as *Bray-Curtis*, *(Un)Weighted Unifrac*, *Jaccard*, etc.) for the communities of microbes might need to be calculated and reused in hierarchical cluster analysis, principal coordinate analysis (*PCoA*), and permutational multivariate analysis of variance (*PERMANOVA*), and so on. If the intermediate data can be effectively integrated and stored, it will improve the efficiency of analysis, enhance reproducibility and avoid errors.[9] The R programming language has become one of the most popular tools for biomedical data analysis.[10] Some efforts have been made to build common representations and infrastructures for complex, highly interdependent datasets in R.[10,11] For example, *SummarizedExperiment*[12] class is widely used to integrate matrix-like objects of feature abundance, the normalized feature abundance, a sample- and feature-oriented metadata data frames as a standardized data structure across many *Bioconductor*[13] packages. To integrate the phylogenetic tree structure and heterogeneous associated data, we defined the *treedata*[14,15] class, which has also been widely used in several packages, such as *tidytree*,[14] *treeio*,[15] and *ggtree*.[16] However, these data structures did not cover all the heterogeneous data of a microbiome experiment, and the existing tools[17–22] for the downstream statistical analysis of the microbiome also did not well integrate the primary heterogeneous data and the useful intermediate data. For instance, the *phyloseq*[17] class is used in the *phyloseq*,[17] *microViz*,[19] and *MicrobiomeAnalyst*[22] packages, but the data structure can only store the primary input datasets; it cannot integrate the normalized data and the intermediate data such as the alpha diversity, dis-similarity indices, the result of differential analysis, and so on. The *animalcules* package was developed using the *MultiAssayExperiment*[23] class, and it cannot integrate intermediate data and the phylogenetic tree, which is often needed in the calculation of specified dissimilarity ((*Un*)*Weighed UniFrac*). These data structures for the microbiome cannot take into account the diversified needs of downstream analysis, which makes it more convenient for some specific needs, while other needs may be troublesome. Moreover, the differences in these data structures are too large, which restricts downstream integrated analysis.

Aside from the data structure defects, downstream statistical, visual, and functional analyses of microbiome data are complex because the appropriate analysis workflow often needs to be explored and adjusted according to the research design (such as different sequencing methods: 16S, metagenome, or metatranscriptome) and the different statistical methods, which often are developed in different platforms or packages across various programming syntax and environments. There remains a lack of flexible and comprehensive packages that can streamline the personalized analysis of microbiome data with a unified and user-friendly syntax. Recently, the tidy concept was proposed and has gained popularity in the R data analysis community with tidy data tools,[9,10] such as *dplyr*,[24] *tidyr*,[24] and *ggplot2*,[25] to allow users to freely and easily explore data and focus more on the special problems. This tidy concept has been applied to different disciplines, including genomic,[26,27] transcriptomic,[28] functional enrichment analysis,[29] and phylogenetic data analyses.[16,30] Through the human-readable data structure and analysis grammar, these tidy data tools make it easier for the community to develop modular manipulation, visualization, and analysis methods, decrease the learning curve for users, and facilitate reproducibility for the related studies.[9,24,31] However, this principle of tidiness has not been implemented in the existing tools[17–22] for microbiome data analysis. This factor greatly hinders the flexibility and ease of use of downstream data analysis in this discipline and also limits the possibilities for researchers to explore data and develop personalized analysis pipelines.

To fill these gaps, we developed the *MicrobiotaProcess* package. We defined the *MPSE* class for storing the microbiome experiment or related ecological dataset and the related intermediate data for downstream analysis. The *MPSE* class defined in *MicrobiotaProcess* inherits both the *SummarizedExperiment*[12] and the *treedata*[14,15] classes, which are popular and widely used in the *Bioconductor*[13] ecosystem. It takes merit from *Bioconductor* packages that are based on these two data structures. The importance of data structure is that it is the foundation for downstream analysis. A good data structure helps to unify the downstream analysis. When operating on certain data, the associated data in the data structure can be updated synchronously, which can decrease many errors caused by improper operations. To make downstream data analysis modular and easy to use, we introduce a user-friendly grammar (i.e., tidy interface) to process, analyze, and visualize microbiome data stored in the *MPSE* data structure. Additionally, we developed a differential abundance analysis method for finding prospective biomarkers with a better false-positive rate based on analysis results of real and simulated datasets (Supplemental file B). All the functions are developed under the tidy framework, which allows users to build human-readable and flexible analysis workflows. We believe it can remove a major obstacle for scientists to



B

```
##  # A MPSE-tibble (MPSE object) abstraction: 4,142 x 11
##  # OTU=218 | Samples=19 | Assays=Abundance | Taxonomy=Kingdom, Phylum, Class, Order, Family, Genus, Species
##     OTU    Sample Abundance time  Kingdom      Phylum           Class  Order   Family  Genus  Species
##     <chr>  <chr>      <int> <chr> <chr>        <chr>            <chr>  <chr>   <chr>   <chr>  <chr>
##  1 OTU-1       F3D0     579 Early k__Bacteria  p__Bacteroidetes c__Ba~ o__Bac~ f__Mur~ g__un~ s__un_f~
##  2 OTU-2       F3D0     345 Early k__Bacteria  p__Bacteroidetes c__Ba~ o__Bac~ f__Mur~ g__un~ s__un_f~
##  3 OTU-3       F3D0     449 Early k__Bacteria  p__Bacteroidetes c__Ba~ o__Bac~ f__Mur~ g__un~ s__un_f~
##  4 OTU-4       F3D0     430 Early k__Bacteria  p__Bacteroidetes c__Ba~ o__Bac~ f__Mur~ g__un~ s__un_f~
##  5 OTU-5       F3D0     154 Early k__Bacteria  p__Bacteroidetes c__Ba~ o__Bac~ f__Bac~ g__Ba~ s__un_g~
##  6 OTU-6       F3D0     470 Early k__Bacteria  p__Bacteroidetes c__Ba~ o__Bac~ f__Mur~ g__un~ s__un_f~
##  7 OTU-7       F3D0     282 Early k__Bacteria  p__Bacteroidetes c__Ba~ o__Bac~ f__Mur~ g__un~ s__un_f~
##  8 OTU-8       F3D0     184 Early k__Bacteria  p__Bacteroidetes c__Ba~ o__Bac~ f__Rik~ g__Al~ s__un_g~
##  9 OTU-9       F3D0      45 Early k__Bacteria  p__Bacteroidetes c__Ba~ o__Bac~ f__Mur~ g__un~ s__un_f~
## 10 OTU-10      F3D0     158 Early k__Bacteria  p__Bacteroidetes c__Ba~ o__Bac~ f__Mur~ g__un~ s__un_f~
## # ...with 4,132 more rows
```

Figure 1. The structure and formatted output of the *MPSE* class (A) *MPSE* class instantiates an object capable of storing various data types generated from a microbiome study. It is built on top of the *SummarizedExperiment* class. In the assays component, the rows represent features such as OTUs, species, or genes (horizontal), and the columns represent samples (vertical). The *rowData* and *colData* components can hold information (such as metadata) about those features and samples, respectively. It also incorporates the *treedata* class, the *otutree* and *taxatree* components can store the phylogenetic tree of the features and the species tree of the features, respectively. The *refseq* component can store the sequences of features by incorporating *XstringSet* class. The component can also be extracted via the corresponding extraction functions (composed of '*mp_extract_*' and a specific explanatory term). (B) A tidy-like output of the *MPSE* class (via 'options(MPSE_show_tbl = TRUE)'). The rows represent the values (abundance, annotation, or other metadata) of each feature in each sample.

explore and analyze the microbiome (16S, metagenome, and metatranscriptome) and other ecological data.

RESULTS

A container improving the exploration of the downstream data of the microbiome

MicrobiotaProcess introduces the *MPSE* class, which is built on top of the *SummarizedExepriment*, and also incorporates the *treedata* and the *XStringSet*[32] classes, for storing microbiome or other related assay data, metadata, and phylogenetic tree data (Figure 1A). The primary matrices data, such as count matrices and standardized matrices, are stored in the *assays* defined in *SummarizedExperiment*, where rows represent features (such as ASVs/OTUs and genes, etc.) and columns represent samples. The sample characteristics are stored in the *colData* defined in *SummarizedExperiment*. To store the phylogenetic tree and associated data, we defined the *otutree* slot, which inherits a *treedata* class defined in *tidytree* and *treeio*. The phylogenetic tree file that contained the evolutionary statistics inferences and other associated data can be parsed using *treeio* and stored in the *otutree* slot. We also defined a *taxatree* slot, which is also a *treedata* class to store hierarchical relationships among the taxa and their lineages. The *taxatree* and *otutree* components can be processed and visualized directly via the in-house developed *ggtree* package suite (*tidytree*, *treeio*, *ggtree*, and *ggtreeExtra*). Moreover, the results of the downstream analysis can also be stored in the *MPSE* class, which is also different from existing related packages. For example, we used *mp_rrarefy* to rarefy the primary matrices data and added the results into the *assays* slot, then used *mp_cal_alpha* to calculate the alpha diversity of the samples. The results of alpha diversity were added into the *colData* slot, which can be visualized further using *mp_plot_alpha*.

```
library(SummarizedExperiment)
library(MicrobiotaProcess)
data(mouse.time.mpse)
assays(mouse.time.mpse)
## List of length 1
## names(1): Abundance
mouse.time.mpse %<>% mp_rrarefy()
assays(mouse.time.mpse)
## List of length 2
## names(2): Abundance RareAbundance
# the alpha index will add into colData when action = 'add'
mouse.time.mpse %<>% mp_cal_alpha(.abundance = RareAbundance, action = "add")
print(mouse.time.mpse, n=4)
## # A MPSE-tibble (MPSE object) abstraction: 4,142 × 18
## # OTU=218 | Samples=19 | Assays=Abundance, RareAbundance | Taxanomy=Kingdom, Phylum, Class, Order, Family, Genus, Species
## OTU Sample Abundance RareAbundance time Observe Chao1
## <chr> <chr> <int> <int> <chr> <dbl> <dbl>
## 1 OTU_1 F3D0 579 214 Early 104 104.
## 2 OTU_2 F3D0 345 116 Early 104 104.
## 3 OTU_3 F3D0 449 179 Early 104 104.
## 4 OTU_4 F3D0 430 167 Early 104 104.
## # ... with 4,138 more rows, and 11 more variables: ACE <dbl>,
## # Shannon <dbl>, Simpson <dbl>, Pielou <dbl>,
## # Kingdom <chr>, Phylum <chr>, Class <chr>, Order <chr>,
## # Family <chr>, Genus <chr>, Species <chr>
```

Each component of the *MPSE* class can be extracted by the self-explanatory function names, prefixed with 'mp_extract_' and followed by a specific explanatory term (see also *Accessors to fetch internal data* in the Methods session) (Figure 1A). In the following example, we used *mp_extract_assays* to extract the *assays* component and *mp_extract_sample* to extract the *colData* component that contained sample metadata.

```
rare.tb <- mouse.time.mpse %>% mp_extract_assays(.abundance = RareAbundance)
sample.da <- mouse.time.mpse %>% mp_extract_sample()
```

Through the *MPSE* class, all related data and results relevant to a microbiome experiment can be stored in a single instance, which enables improved exploration of the downstream data, facilitates data sharing, and enhances reproducibility.

Bridging the upstream analysis tools and MPSE constructor

The output files of the upstream analysis of the microbiome are various and usually not human friendly. The first challenge that researchers should encounter is frequently learning how to parse these output files. The *phyloseq* provided some functions to parse the output of common upstream analysis tools for the 16S rRNA dataset, but not for metagenomics or the output of commonly used tools (*qiime2*, *dada2*). To address the need for multiple types of microbiome data (16S rRNA, metagenomics, and other related ecological data) as well as output obtained from commonly used tools, *MicrobiotaProcess* provides several functions to parse the output of the upstream analysis of the microbiome (Figure 2A). To enhance the interoperability with other R packages, *MicrobiotaProcess* also provides *as.MPSE* to convert common *S4* objects defined in the *Bioconductor* ecosystem, such as *phyloseq*, *biom*,[33] *SummarizedExperiement*, and *TreeSummarizedExperiment*. All of the parse functions and the convert functions output the same object type, *MPSE*, making them consistent and robust for downstream manipulation. In addition, an *MPSE* class can also be constructed by calling the function of the same name with the required parameters (see also the *parser functions and the MPSE constructor* in the Methods session). Here, we used *as.MPSE* to convert a *biom* object, which is the output of *read_biom* of *biomformat*,[33] to an *MPSE* object and used the *MPSE* function to build an *MPSE* object from scratch.

```
library(MicrobiotaProcess)
library(biomformat)
biom_file <- system.file("extdata", "rich_sparse_otu_table.biom", package = "biomformat")
xx <- read_biom(biom_file)
# convert to an MPSE object using as.MPSE function
mpse2 <- xx %>% as.MPSE()
library(MicrobiotaProcess)
library(vegan)
data(varespec, varechem)
# building an MPSE object using MPSE function
mpse1 <- MPSE(assays=list(Abundance=t(varespec)), colData=varechem)
```

Through these functions, *MicrobiotaProcess* can bridge the upstream analysis tools and address the need for downstream analysis based on the *MPSE* class (Figures 2C1 and C2). It supports multiple types of data and we will describe some case studies using *MicrobiotaProcess* to analyze 16S rRNA, metagenomics, Kyoto Encyclopedia of Genes and Genomes (KEGG) gene datasets, and related ecological data in the following sections.

The unified tidy framework analysis grammar

To facilitate data manipulation and exploration of the microbiome or other ecological data based on the *MPSE*, *MicrobiotaProcess* defined a tidy-like format output for the *MPSE* class (Figure 1B). The *assays* of MPSE are converted to a tidy format, in which three columns (feature identifiers, sample, and feature abundance) are fixed, and each row represents the abundance of each feature in each sample. Optional columns include the taxonomy annotation and sample metadata information or newly calculated information. *MicrobiotaProcess* provides a unified analysis grammar to express all the steps of a typical microbiome community analysis workflow using self-explanatory function names, which are composed of '*mp_*', a specific verb, and one or two explanatory terms (Figures 2C1 and C2). Moreover, *MicrobiotaProcess* extends the *dplyr* verbs to support processing data stored in an *MPSE* object (Figure 2B). Following the concept of tidiness, these verbs provide robust and standardized operations for data analysis and transformation and can be assembled into a workflow using the pipe operator (%>% or |>). This enables users to effectively analyze ecological data, such as metagenomic data, and easily build reproducible and human-readable pipelines.

Differential abundance analysis

Potential disease microbiota biomarkers or probiotic microbiota can be detected by differential abundance analysis. Although there are many tools available, it remains challenging to obtain a good false-positive control.[34] To identify the more conservative differential taxa, we developed *mp_diff_analysis*. We found our method can limit the false positive rate with a better balance of type I and type II errors when compared with *metagenomeSeq*,[35] *LEfSe*,[36] *ANCOMBC*,[37] *LinDA*,[38] and *ZicoSeq*[39] in different datasets simulated from log-normal and normal distributions with different means and standard deviations (Figures SB.3–SB.12). We additionally evaluated the *mp_diff_analysis* and other tools using 10 real 16S rRNA datasets.[40–48] We found that our method tends to identify a relatively small number of significant ASV (Figures SB.13–SB.15). The significant ASV detected by our methods tended to also be detected by other tools (Figures SB.14 and SB.15). Because of the design of the *LEfSe*, we were unable to obtain the p values of all features to be corrected using a false discovery rate (FDR). To maintain consistency, we applied the Bonferroni method to the results of each tool by dividing the p value threshold by the total number of features examined as a new threshold. We found the results of our method (using both Bonferroni and p value modes) are similar to the results obtained by other tools using Bonferroni mode (Figure SB.13). The results are similarly consistent with the results using simulated data, which suggests that, although our method may lose some sensitivity, it is more conservative and accurate.

Interface to integrate external data

To improve interoperability between the *MicrobiotaProcess* and other tools, we implemented *left_join*, which was inherited from the tidyverse

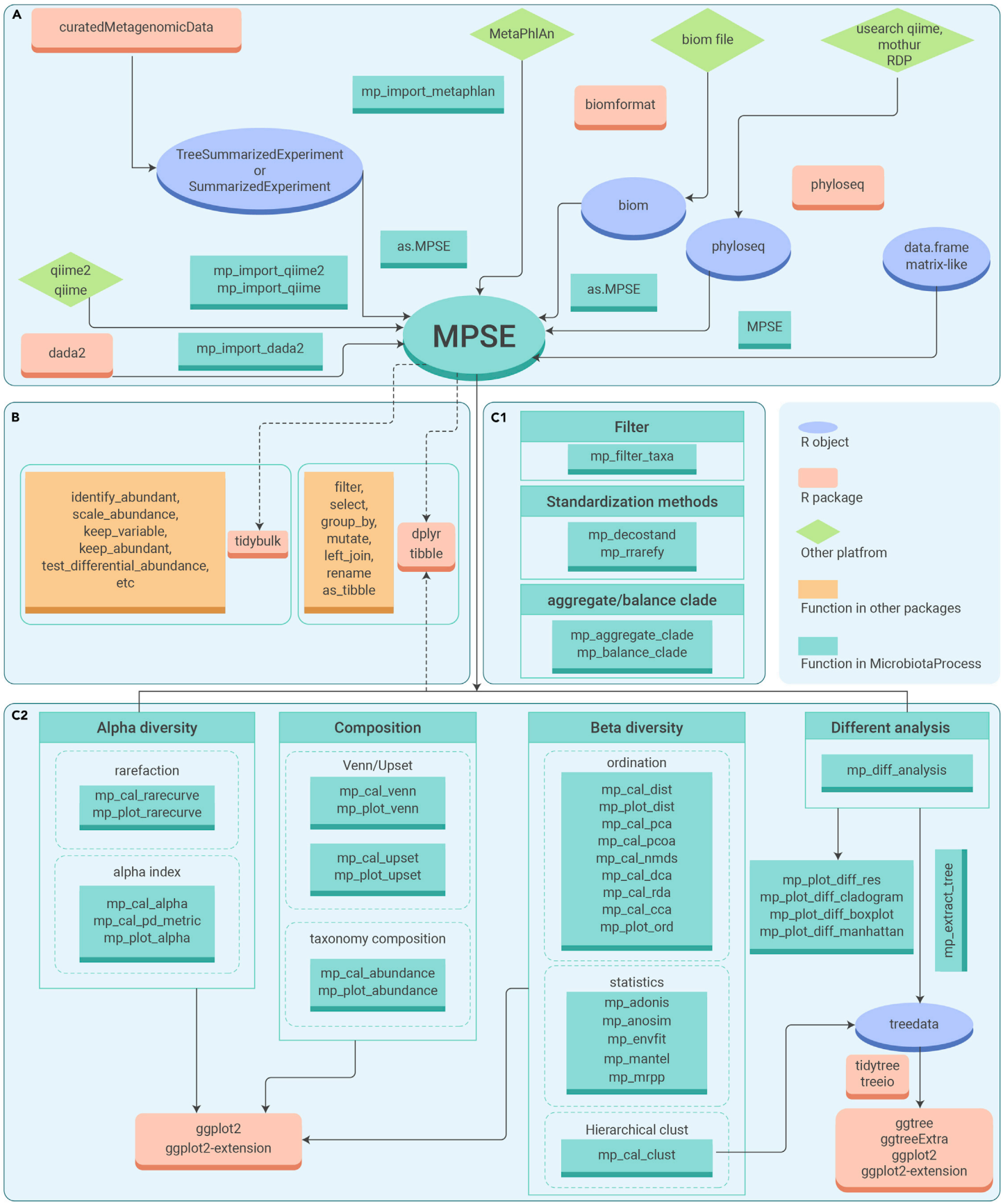


Figure 2. Overview of the design of the *MicrobiotaProcess* package (A) The package provides several functions to import upstream analysis results (*mp_import_qiime2*, *mp_import_metaphlan*, and *mp_import_dada2*, etc), several converters (*as.MPSE* methods) to import data from other objects, and a method (*MPSE*) for building an *MPSE* object from scratch. Unifying downstream data processing and data analysis through *MPSE* objects. (B) Some functions provided in the *dplyr* and *tidybulk* packages are compatible with *MPSE* objects. (C1 and C2) *MicrobiotaProcess* provides a wide variety of functions for routine microbiome analysis. The functions were designed to follow the tidy principle.

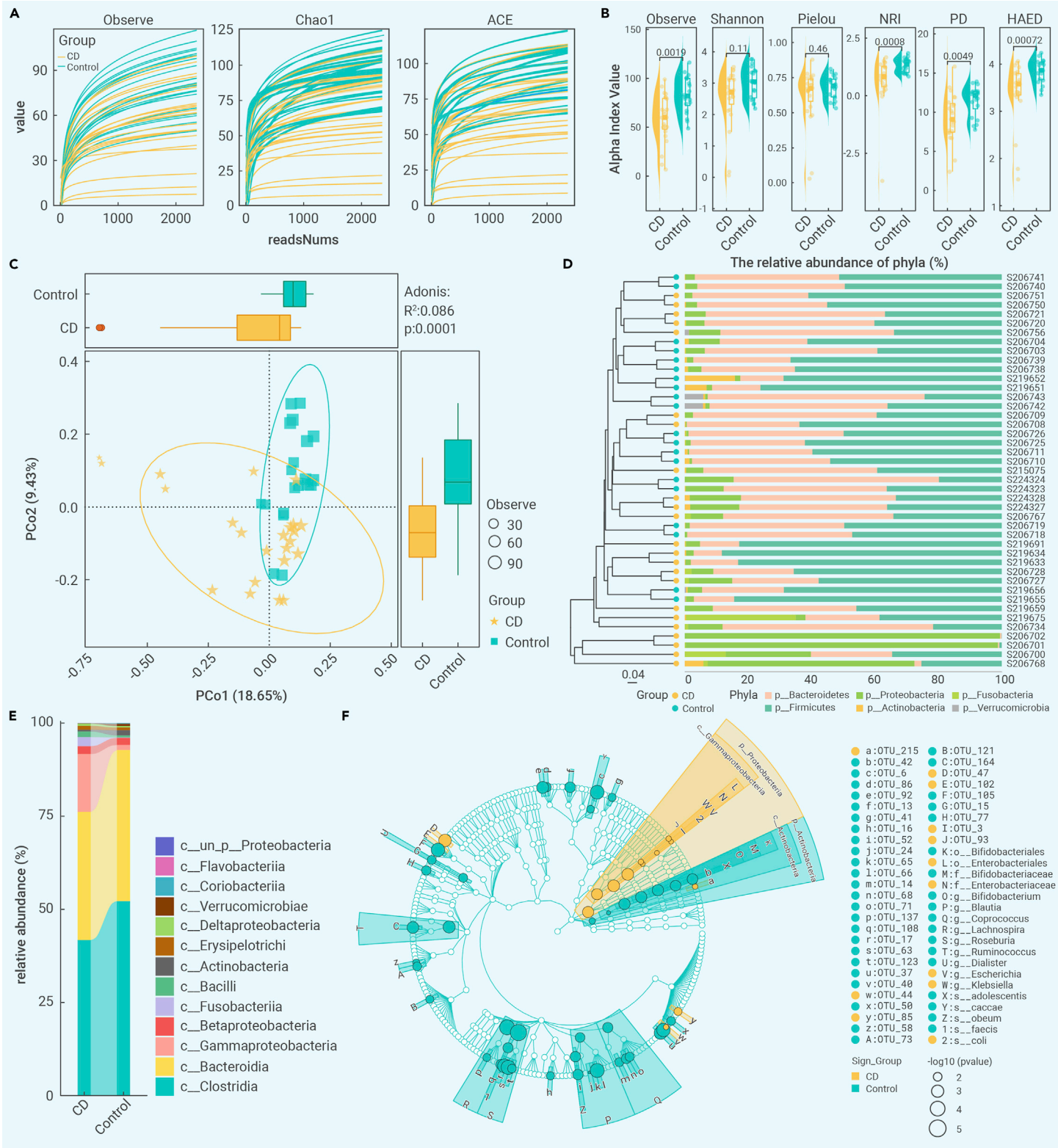


Figure 3. Characterizing gut microbiota in patients with CD (A) The rarefaction curves of the sample from CD and control groups (visualized by *mp_plot_rarecurve*). (B) The alpha diversity boxplot between the CD and control groups, and labeled with the p value calculated by Mann-Whitney U test (visualized by *mp_plot_alpha*). (C) The *PCoA* plot and the result of Adonis with *PERMANOVA* (visualized by *mp_plot_ord*). (D) A cladogram displayed the hierarchical clustering result of the samples based on the OTU abundance, and a bar chart displayed the relative abundance at the phyla level (visualized by *ggtree* and *ggtreeExtra*). (E) The relative abundance bar chart at the class level for each group. (F) Cladogram of the significant differential taxa, with the highlighted clades representing the differential species, was enriched in the corresponding group. The colored points represent the differential taxa, and the sizes were scaled by the FDR (*kruskal.test*) (visualized by *mp_plot_diff_res*).

ecosystem,[24] to integrate the external data. Any data frame that contains the same sample or OTU column as in the *MPSE* class or *dist* class, where the sample distance metrics that are stored can be integrated using *left_join*, and then the following analysis can be done with the functions provided by *MicrobiotaProcess*. So that users can integrate sample dis-similarity indices calculated by user-defined distance metrics to an *MPSE* object, *PCoA* and the differential analysis of the dissimilarity indices can be performed using *mp_cal_pcoa*, *mp_plot_ord*, and

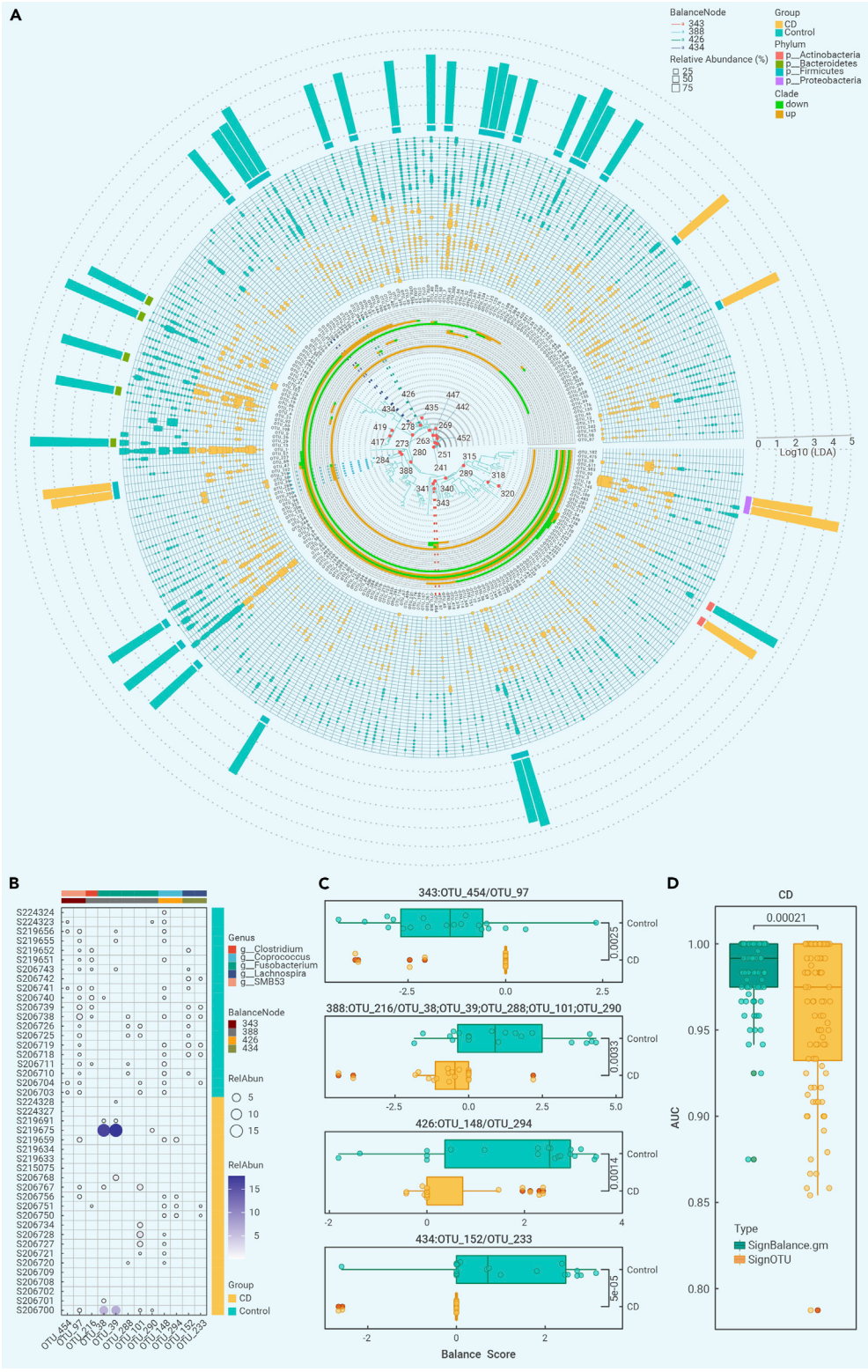


Figure 4. The results of differential clades at a broader resolution on the phylogenetic tree (A) The differential clades at the phylogenetic tree of the community species, the red points in the phylogenetic tree pinpoint the differential clades, the external point layer represented the relative abundance of each OTU on each sample, and the external bar plot represented the mean LDA of the differential OTUs. (B) The relative abundance of the non-significantly different OTUs before phylogenetic transform. (C) The balance scores of significantly differential clades. (D) Comparison of model performances based on the significantly differential clades and the significantly differential OTUs.

Data visualization

Visualization is an important aspect of the exploration and interpretation of microbiome data. To help users quickly render clear, meaningful, and comprehensive visualizations that give useful microbiome insight, especially through the integration of different data, we developed many visualization methods based on *ggplot2*, *ggtree*, and *ggtreeExtra*, which are composed of 'mp_plot_', and one or two explanatory terms (Figure 2C2). These approaches have the advantage of allowing the visualization results to be a proper *ggplot* or *ggtree* object[49] (*mp_plot_diff_cladogram* and *mp_plot_diff_res*), which can be modified, inspected, and reused with the universal and concise *ggplot2* or *ggtree* syntax (Figures SA.10, SA.13 and SA.17).

Case studies

To illustrate the flexibility and versatility of *MicrobiotaProcess*, we reanalyzed three public datasets as examples. Each example addresses a different problem and has a different focus on demonstrating the features of *MicrobiotaProcess*. The details of the analysis and the corresponding code are presented in Supplementary file A.

Example 1: Characterization of the gut microbiota using 16S rRNA data. The 16S rRNA dataset of pediatric stool was obtained from the Integrative Human Microbiome Project Consortium[50] and downloaded from the *MicrobiomeAnalyst*[22] website, which contained 23 Crohn disease (CD) and 20 control samples. We performed the calculation of the alpha and beta diversity indices, the analysis of the diversity measures with the related test method (subsequent Mann-Whitney U test and *PERMANOVA*), and the visualization of patterns in the microbiota community using *MicrobiotaProcess*. The result of alpha diversity showed that the richness of the CD's microbiota community was significantly lower than the control's (Figures 3A and 3B); the evenness of the microbiota species between the CD and control group did not have a significant difference, but the median of the diversity of the CD group was higher than the control group (Figure 3B). We also found the nearest relative index (NRI) of most samples was larger than zero, and the NRI of the CD group was significantly lower than the control's using the *mp_cal_pd_metric* and *mp_plot_alpha* of *MicrobiotaProcess* (Figure 3B). This means most communities whose species are more closely related than under random assembly.[51] But the control group's is more phylogenetically clustered than the CD group's, which suggests the bacterial composition of the CD group might have been more influenced by the competitive exclusion than the control group.[52] Then the

mp_plot_dist (Figure SA.20). More important, the results of differential abundance analysis performed by other external software tools (e.g., *ACNOMBC*, *ZicoSeq*, and *LinDA*) can be stored in a data frame containing a column of OTU, a column of enriched groups, and other statistical results such as p values or FDR values and integrated into an *MPSE* object. After that, we can use this information in a downstream analysis or use *MicrobiotaProcess* to visualize the data to facilitate the interpretation of the results (Figures SA.21–SA.24). Through the interface, *MicrobiotaProcess* can better interact with other metagenomic bioinformatics tools.

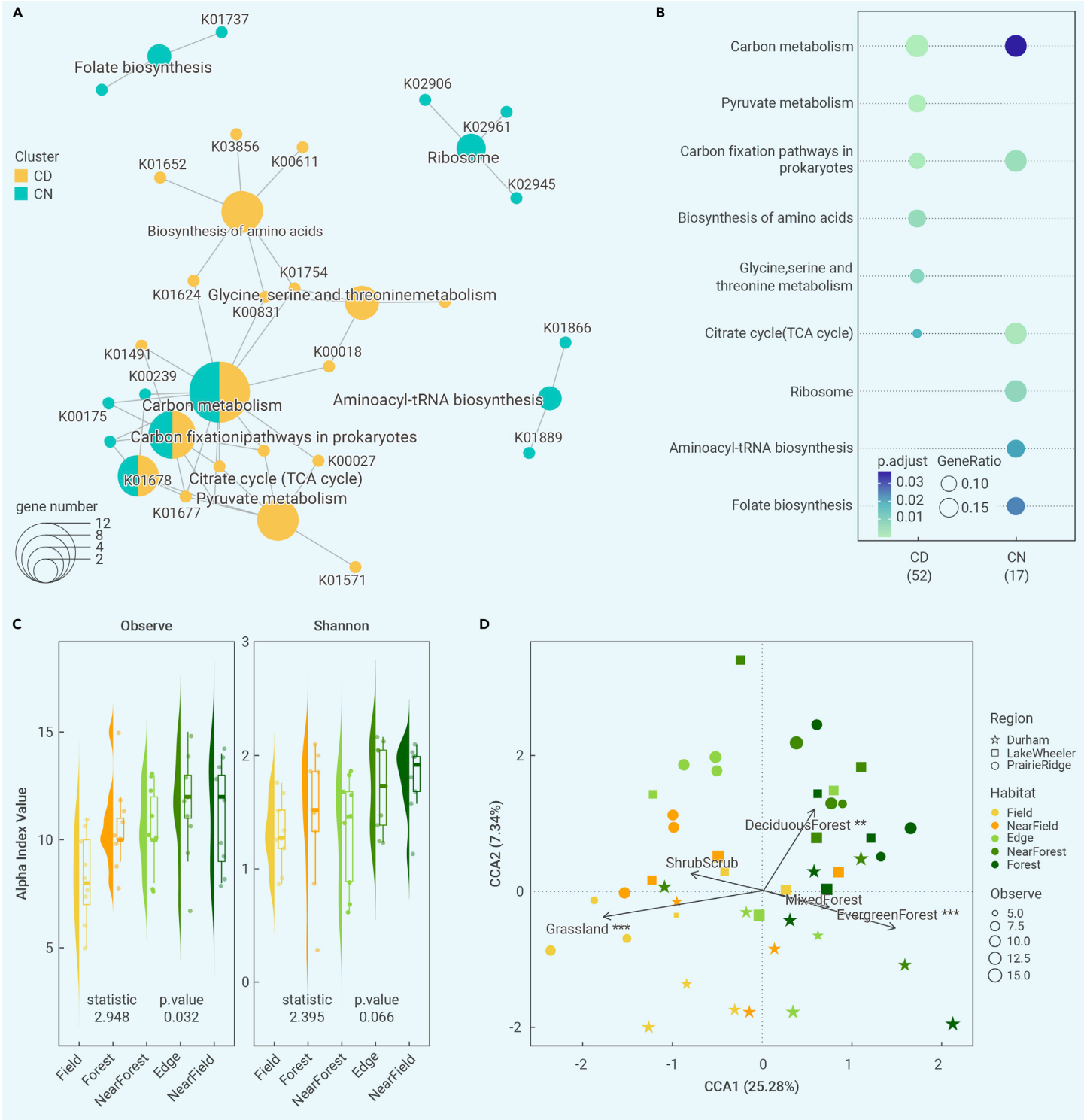


Figure 5. Comparing functional profiles between the CD and control groups, and characterizing mosquito community structure with the raincloud plot of the alpha diversity and the *pCCA* plot of the mosquito communities (A) The Category-Gene Network was used to visualize KEGG enrichment results between the CD and control groups. (B) The dot plot of the KEGG enrichment results. (Visualized by *cnetplot* and *dotplot* provided by the *enrichplot* package). (C) The alpha diversity result showed that the mosquito species richness gradually increases from field to forest (visualized by *mp_plot_alpha*). (D) The *pCCA* result showed that the mosquito communities were significantly associated with some environment landscape variables, such as grassland (Grassland), evergreen tree canopy (EvergreenForest), and deciduous tree canopy (DeciduousForest). Each point represents one sample and the corresponding data of the samples are encoded as visual characteristics of the points (observe species as to size, habitat as to color, and the region as to shape). The arrows represent the environmental factors and the asterisks indicate significant levels related to the Mosquito communities in the study (*0.05, **0.01, ***0.001) (visualized by *mp_plot_ord*).

PCoA and the *adonis* with *PERMANOVA* results revealed that the microbiota communities between the CD and control group had significant differences (Figure 3C). We also found that Proteobacteria might be more abundant in CD by combining the hierarchical relationships and the phyla abundance presented as the horizontally stacked bar charts (Figure 3D).

To further investigate the relationship between the microbial communities and CD, we used *mp_cal_abundance* and *mp_plot_abundance* to display the abundance of class level, this result showed that the abundance of Gammaproteobacteria belonged to Proteobacteria might be enriched in the CD group (Figures 3E and SA.6). Next, we used *mp_diff_analysis* to identify the different OTUs, the result also revealed that the OTUs significantly enriched in the CD group mainly belong to Proteobacteria, whereas the OTUs significantly enriched in the control group mainly belonged to Firmicutes and Actinobacteria (Figures 3F and SA.13). To identify the clades of closely related species at a

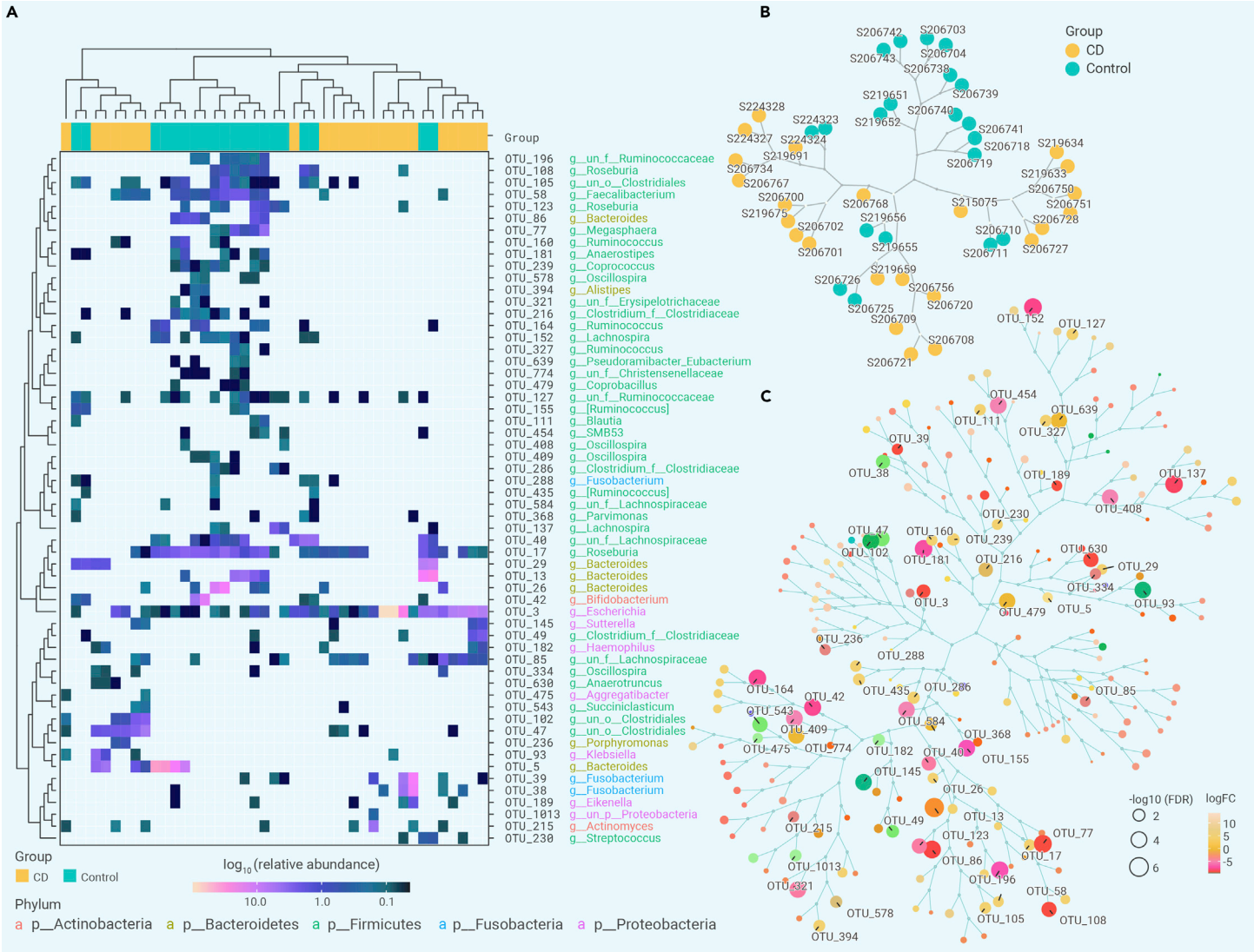


Figure 6. The results of differential OTUs are based on the *edgeR_quasi_likelihood* method using *tidybulk* (A) The relative abundance heatmap of the different OTUs (visualized by the *mp_plot_abundance* function). (B) The hierarchical cluster of samples is based on the relative abundance of the different OTUs (visualized by *ggtree*). (C) The hierarchical cluster of OTUs is based on their relative abundance across different samples, the differential OTUs were labeled with their names (visualized by *ggtree*).

broader resolution on the phylogenetic tree, we calculated the balance score of internal nodes of the phylogenetic tree using *mp_balance_clade* and identified the differential clades using *mp_diff_analysis*. Most of the differential clades near the tips of the tree were significantly different species, such as OTU_44/OTU_25 (both belong to Lachnospiraceae), and OTU_47/OTU_102 (both belong to Clostridiales). We also found some differential clades contain closely related species that were not detected in the previous differential analysis, such as OTU_454/OTU_97 (both belong to Clostridiaceae *SMB53*), OTU_152/OTU_233 (both belong to *Lachnospira*) (Figure 4A), and so on. These results suggested that the balance transform can improve the detection of differential signals by accumulating the small consilient differences at a broader resolution (internal nodes of different depths on the phylogenetic tree) and thus enhances the ability to discover biological associations and interpret the results at different resolutions (Figures 4B and 4C). In addition, compared to the differential OTUs, the significantly differential clades can effectively improve the supervised classification accuracy in the CD microbiota datasets (Figure 4D). Similar results can also be observed in other inflammatory bowel disease studies (Figure SB.1), which indicates that the identification of differential clades after phylogenetic transformation can improve the performance of classification models to determine associations with clinical outcomes. Through the tidy-like unified grammar of *MicrobiotaProcess*, the alpha and beta diversity of the stool microbial communities between CD patients and controls can be explored rapidly. The differential microbial organisms and clades at a broader resolution on the phylogenetic tree between the CD patients and the control groups can also be identified and visualized intuitively.

Example 2: Differential abundance analysis and functional characterization of metagenomic data. The metagenomics taxa abundance and KEGG gene abundance datasets were obtained from another pediatric CD study.[48] We used the *mp_diff_analysis* function to identify the differential species and genes, followed by KEGG pathway enrichment analysis of differential genes using *clusterProfiler* and *MicrobiomeProfiler*.[53] We found the *Clostridium symbiosum* and *Faecalibacterium prausnitzii* were nominally (uncorrected $p \leq 0.05$, FDR > 0.05) significant-enriched in the CD group compared with the control group (Figures SA.28 and SA.29). Interestingly, the KEGG enrichment results showed that the KEGG pathways of the CD stool group were significantly enriched in the biosynthesis of amino acids and glycine, serine, and threonine metabolism, and pyruvate metabolism (Figures 5A and 5B). Although these results are not revealed in the original paper,[48] it is consistent with other related studies. These studies have found that CD microbiomes have the potential to enhance synthetic amino acids and pyruvate metabolism.[46,54,55] Interestingly, enrichment analysis using differential KEGG genes identified by other differential abundance methods cannot identify these two pathways simultaneously (Figure SB.2). In this example, we use *MicrobiotaProcess* for a differential abundance analysis, and then use the in-house developed packages, *clusterProfiler* and *MicrobiomeProfiler*, for functional enrichment analysis. The combination of these packages

facilitates the functional interpretation of metagenomic data and the discovery of functional signatures that leads to disease progression.

Example 3: Analyzing environmental-trait interactions in ecological communities. The dataset is from a study on the landscape ecology of mosquito communities in North Carolina for determining the risk for vector-borne disease.[56] Three agricultural/mixed-use landscapes (regions) were selected. Within each region, three 200-m transects that spanned a field-forest habitat gradient were identified, and five traps were set for each transect to collect mosquitoes.[56] Here, we used *MicrobiotaProcess* to calculate alpha diversity and visualize the patterns of the community composition to evaluate whether the mosquito community structure was influenced by spatial variability.

We found that the richness of the mosquito species was significantly associated with the habitat with mean richness increasing from field to forest (Figure 5C). The partial constrained correspondence analysis (*pCCA*) result showed that the mosquito communities were significantly associated with some environmental landscape variables, such as grassland (Grassland), evergreen tree canopy (EvergreenForest), and deciduous tree canopy (DeciduousForest) (Figure 5D). The mosquito species showed a clear preference for forested or field habitats (Figure SA.31). These results were consistent with the findings of the original study.[56] This example shows that *MicrobiotaProcess* can be used not only for analyzing microecological data but also for general ecological data, helping us to discover the impact of environmental factors on community structure and functions.

Example 4: Interact with existing tools. Because the *MPSE* class inherits *SummarizedExperiment*, the methods developed for *SummarizedExperiment* can also be applied to an *MPSE* object. For example, the *tidybulk* package provides a *test_differential_abundance* to perform differential transcriptome analysis using *edgeR*[57] quasi-likelihood, *edgeR* likelihood-ratio, limma-voom, limma-voom-with-quality-weights, or DESeq2.[58] These methods can be directly applied to an *MPSE* object within the tidy framework for differential abundance analysis and indeed they have been reported for analyzing microbiome data.[34,59] Here, we re-analyze example 1 by using the *test_differential_abundance* function to perform the differential analysis based on the *edgeR quasi-likelihood*. Then the result was extracted by the *mp_extract_tree* function and can be manipulated and visualized using the *ggtree* package suite.

Compared with the results of example 1, the number of the differential OTUs identified by edgeR is higher (Figure SA.18) and the main conclusions are consistent. We found the abundance of OTUs belonging to *Bifidobacterium*, *Faecalibacterium*, *Roseburia*, and *Coprobacillus* was also significantly decreased in the CD group compared with the control group, and the abundance of several OTUs belonging to *Escherichia*, *Klebsiella*, and *Haemophilus*, which belonged to Gammaproteobacteria, was significantly enriched in the CD group (Figure 6). *MicrobiotaProcess* is able to interoperate with existing tools, including *tidybulk*, *dplyr*, *ggtree* package suite, *clusterProfiler*, and *MicrobiomeProfiler*, which we have demonstrated earlier. Moreover, the *left_join* method allows the integration of external data, including the sample distance and differential analysis results mentioned above. This enables users to quickly integrate new methods as they emerge in the future. These features significantly enhance *MicrobiotaProcess*' capabilities. It can connect upstream data, unify downstream analysis, and easily integrate into existing analysis pipelines.

DISCUSSION AND CONCLUSION

Microbiomics technologies have become increasingly popular methods for exploring the relationship between microbial communities and the host or the environment (e.g., intestine, skin, soil, and the ocean).[1,50,60–64] Data analysis is still one of the bottlenecks in this field, especially in downstream analysis, the integration of heterogeneous data and the need for personalized analysis have brought new challenges. *MicrobiotaProcess* provides a comprehensive data structure, the *MPSE* class, to store heterogeneous microbiome data, including feature (e.g., species, OTUs) abundance, sample data (e.g., clinical information), and feature data (e.g., taxonomic relationships, functional profiles). Moreover, simultaneously with the analysis, intermediate results can be stored in an *MPSE* object. For example, the normalized or rarefied data, the diverse dissimilarity metrics, and the results of the differential analysis can be integrated with the microbiome data into an *MPSE* object. These intermediate results can be further processed, visualized, and reused. This can prevent repeated calculation, facilitate data sharing, and enhance analytic reproducibility. *MicrobiotaProcess* implements a set of functions within the tidy principles to unveil the characteristics and biological function of microbial communities in a diverse environment. The returned values of these functions are predictable and consistent. Each function is designed to accomplish a simple task and these functions can be combined to accomplish complex tasks. In this way, it has better flexibility, and the development of workflow through function series can meet most of the personalized analysis needs. Considering the interoperation between *MicriobiotaProcess* and other tools, some functions developed by other software can be applied to *MPSE* objects, making the downstream analysis function more comprehensive and having wider application scenarios. For example, differential analysis using *tidybulk* and functional analysis using *clusterProfiler* can be integrated into the analysis of the microbiome and other ecological data through *MPSE* objects.

MicrobiotaProcess provides several functions to parse outputs obtained from upstream tools and allows converting commonly used objects to *MPSE* objects (Figure 2A). This enables the functionalities provided by *MicrobiotaProcess* to be well connected to upstream analysis. The downstream analysis results need to be interpreted through visualization. *MicrobiotaProcess* provides several functions for visualization. More important, the tidy interface extracts relevant data and connects to the visualization functions developed by the R community, allowing users to have more options for visual exploration and presentation of data. For example, differential analysis results can be extracted by *mp_extract_tree*, followed by a visual exploration of the result via the *ggtree* package suit (Figures 4A and 6). Although *MicrobiotaProcess* was mainly designed for analyzing microbiome data, some of its methods, such as alpha, beta diversity, (partial) constrained correspondence analysis (*pCCA* or *CCA*), and redundancy analysis (*RDA*),[65] are also applicable to other ecological data (Figures 5C and 5D). In addition, *MicrobiotaProcess* implements several methods from scratch, including the *mp_cal_pd_metric* for calculating several phylogenetic diversity metrics by combining the phylogenetic tree and the species abundance of a community (Figure 3B), the *mp_diff_analysis* for identifying potential biomarkers with better control of the false-positive rate (Figures SB.3–SB.6), and a phylogenetic transform function (*mp_balance_clade*) for identifying differential clades of related bacteria at different resolutions on the phylogenetic tree (Figures 4A and 4C). Using differential clades can improve the performance of the supervised classification model (Figures 4D and SB.1), which can be used for disease diagnosis.

Methods of microbiome data analysis are rapidly evolving. Some new methods not in the *MicrobiotaProcess* package will be developed by other software, such as the generalized Lotka-Volterra model for the analysis of microbial interactions,[66] *LinDA*, and *ZicoSeq* for the differential abundance analysis of microbiome data. Through the *left_join* method, users can integrate the analysis results of these new methods into an MPSE object, and the results can be explored, visualized, and further analyzed using the functions implemented in *MicrobiotaProcess* (Figures SA.20–SA.24). Compared with other related tools, *MicrobiotaProcess* has many unique advantages (Figure SA.33), and we will develop more functions as needed in the future. In summary, we believe that *MicrobiotaProcess* will be a valuable resource for analyzing microbiomes and other ecological data.

MATERIALS AND METHODS

More information is available in the supplemental information file.

Data availability

The MicrobiotaProcess R package is open source and freely available on Bioconductor (https://bioconductor.org/packages/MicrobiotaProcess) and GitHub (https://github.com/YuLab-SMU/MicrobiotaProcess). R markdown files and datasets that are used to generate the supplemental files are available on GitHub (https://github.com/YuLab-SMU/MP_supplementary_file).

REFERENCES

1. Turnbaugh, P.J., Ley, R.E., Hamady, M., et al. (2007). The human microbiome Project. Nature ***449***, 804–810.

2. Lloyd-Price, J., Abu-Ali, G., and Huttenhower, C. (2016). The healthy human microbiome. Genome Med. ***8***, 51.
3. Bolyen, E., Rideout, J.R., Dillon, M.R., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. ***37***, 852–857.
4. Callahan, B.J., McMurdie, P.J., Rosen, M.J., et al. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. Nat. Methods ***13***, 581–583.
5. Edgar, R.C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat. Methods ***10***, 996–998.
6. Schloss, P.D., Westcott, S.L., Ryabin, T., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. ***75***, 7537–7541.
7. Segata, N., Waldron, L., Ballarini, A., et al. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. Nat. Methods ***9***, 811–814.
8. Truong, D.T., Franzosa, E.A., Tickle, T.L., et al. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat. Methods ***12***, 902–903.
9. Lowndes, J.S.S., Best, B.D., Scarborough, C., et al. (2017). Our path to better science in less time using open data science tools. Nat. Ecol. Evol. ***1***, 0160.
10. Ihaka, R., and Gentleman, R. (1996). R: a language for data analysis and graphics. J. Comput. Graph Stat. ***5***, 299–314.
11. Huber, W., Carey, V.J., Gentleman, R., et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Methods ***12***, 115–121.
12. Morgan, M., Obenchain, V., Hester, J., and Pagès, H. (2021). SummarizedExperiment: SummarizedExperiment Container.R package version 1.28.0, https://bioconductor.org/packages/SummarizedExperiment.
13. Gentleman, R.C., Carey, V.J., Bates, D.M., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. ***5***, R80.
14. Yu, G. (2021). tidytree: A Tidy Tool for Phylogenetic Tree Data Manipulation.R package version 0.4.2. https://github.com/YuLab-SMU/tidytree.
15. Wang, L.-G., Lam, T.T.-Y., Xu, S., et al. (2020). treeio: an R package for phylogenetic tree input and output with richly annotated and associated data. Mol. Biol. Evol. ***37***, 599–603.
16. Yu, G., Smith, D.K., Zhu, H., et al. (2017). ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. ***8***, 28–36.
17. McMurdie, P.J., and Holmes, S. (2013). Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One ***8***, e61217.
18. Huang, R., Soneson, C., Ernst, F.G.M., et al. (2020). TreeSummarizedExperiment: a S4 class for data with hierarchical structure. F1000Res. ***9***, 1246.
19. Barnett, D., Arts, I., and Penders, J. (2021). microViz: an R package for microbiome data visualization and statistics. J. Open Source Softw. ***6***, 3201.
20. Liu, C., Cui, Y., Li, X., and Yao, M. (2021). microeco: an R package for data mining in microbial community ecology. FEMS Microbiol. Ecol. ***97***. fiaa255.
21. Zhao, Y., Federico, A., Faits, T., et al. (2021). animalcules: interactive microbiome analytics and visualization in R. Microbiome ***9***, 76.
22. Dhariwal, A., Chong, J., Habib, S., et al. (2017). MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. Nucleic Acids Res. ***45***, W180–W188.
23. Ramos, M., Schiffer, L., Re, A., et al. (2017). Software for the integration of multiomics experiments in bioconductor. Cancer Res. ***77***, e39–e42.
24. Wickham, H., Averick, M., Bryan, J., et al. (2019). Welcome to the tidyverse. J. Open Source Softw. ***4***, 1686.
25. Wickham, H. (2011). ggplot2. WIREs. Comp. Stat. ***3***, 180–185.
26. Lee, S., Cook, D., and Lawrence, M. (2019). plyranges: a grammar of genomic data transformation. Genome Biol. ***20***, 4.
27. Lawrence, M., Huber, W., Pagès, H., et al. (2013). Software for computing and annotating genomic ranges. PLoS Comput. Biol. ***9***, e1003118.
28. Mangiola, S., Molania, R., Dong, R., et al. (2021). tidybulk: an R tidy framework for modular transcriptomic data analysis. Genome Biol. ***22***, 42.
29. Wu, T., Hu, E., Xu, S., et al. (2021). clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innovation ***2***, 100141.
30. Xu, S., Dai, Z., Guo, P., et al. (2021). ggtreeExtra: compact visualization of richly annotated phylogenetic data. Mol. Biol. Evol. ***38***, 4039–4042.
31. Wickham, H. (2014). Tidy data. J. Stat. Softw. ***59***, 1–23.
32. Pagès, H., Aboyoun, P., Gentleman, R., and DebRoy, S. (2021). Biostrings: Efficient Manipulation of Biological Strings.R package version 2.66.0. https://bioconductor.org/packages/Biostrings.
33. McMurdie, P.J., and Paulson, J.N. (2021). Biomformat: An Interface Package for the BIOM File Format.R package version 1.26.0. https://github.com/joey711/biomformat/.
34. Weiss, S., Xu, Z.Z., Peddada, S., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome ***5***, 27.
35. Paulson, J.N., Stine, O.C., Bravo, H.C., et al. (2013). Differential abundance analysis for microbial marker-gene surveys. Nat. Methods ***10***, 1200–1202.
36. Segata, N., Izard, J., Waldron, L., et al. (2011). Metagenomic biomarker discovery and explanation. Genome Biol. ***12***, R60.
37. Lin, H., and Peddada, S.D. (2020). Analysis of compositions of microbiomes with bias correction. Nat. Commun. ***11***, 3514.
38. Zhou, H., He, K., Chen, J., and Zhang, X. (2022). LinDA: linear models for differential abundance analysis of microbiome compositional data. Genome Biol. ***23***, 95.
39. Yang, L., and Chen, J. (2022). A comprehensive evaluation of microbial differential abundance analysis methods: current status and potential solutions. Microbiome ***10***, 130.
40. Schubert, A.M., Rogers, M.A.M., Ring, C., et al. (2014). Microbiome data distinguish patients with clostridium difficile infection and non-C. Difficile-associated diarrhea from healthy controls. mBio ***5***. e01021-14.
41. Yurgel, S.N., Douglas, G.M., Comeau, A.M., et al. (2017). Variation in bacterial and eukaryotic communities associated with natural and managed wild blueberry habitats. Phytobiomes J. ***1***, 102–113.
42. Scher, J.U., Sczesnak, A., Longman, R.S., et al. (2013). Expansion of intestinal *Prevotella copri* correlates with enhanced susceptibility to arthritis. Elife ***2***, e01202.
43. Charlson, E.S., Chen, J., Custers-Allen, R., et al. (2010). Disordered microbial communities in the upper respiratory tract of cigarette smokers. PLoS One ***5***, e15216.
44. Kesy, K., Oberbeckmann, S., Kreikemeyer, B., and Labrenz, M. (2019). Spatial environmental heterogeneity determines young biofilm assemblages on microplastics in baltic sea mesocosms. Front. Microbiol. ***10***, 1665.
45. Turnbaugh, P.J., Hamady, M., Yatsunenko, T., et al. (2009). A core gut microbiome in obese and lean twins. Nature ***457***, 480–484.
46. Heinken, A., Hertel, J., and Thiele, I. (2021). Metabolic modelling reveals broad changes in gut microbial metabolism in inflammatory bowel disease patients with dysbiosis. NPJ Syst. Biol. Appl. ***7***, 19.
47. Baxter, N.T., Ruffin, M.T., Rogers, M.A.M., and Schloss, P.D. (2016). Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. Genome Med. ***8***, 37.
48. Douglas, G.M., Hansen, R., Jones, C.M.A., et al. (2018). Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn's disease. Microbiome ***6***, 13.
49. Xu, S., Li, L., Luo, X., et al. (2022). Ggtree: a serialized data object for visualization of a phylogenetic tree and annotation data. iMeta ***1***, e56.
50. Integrative HMP iHMP Research Network Consortium (2014). The integrative human microbiome project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. Cell Host Microbe ***16***, 276–289.
51. Webb, C.O. (2000). Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. Am. Nat. ***156***, 145–155.
52. Pearse, W.D., Purvis, A., Cavender-Bares, J., and Helmus, M.R. (2014). Metrics and models of community phylogenetics. In Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology (Springer), pp. 451–464.
53. Chen, M., and Yu, G. (2021). MicrobiomeProfiler: An R/shiny Package for Microbiome Functional Enrichment Analysis. R package version 1.4.0. https://github.com/YuLab-SMU/MicrobiomeProfiler/.
54. Bjerrum, J.T., Steenholdt, C., Ainsworth, M., et al. (2017). Metabonomics uncovers a reversible proatherogenic lipid profile during infliximab therapy of inflammatory bowel disease. BMC Med. ***15***, 184.
55. Polunin, G., Sedakov, I., Borota, O., et al. (2013). The influence of sodium lactate on carbohydrates metabolism in patients with inflammatory bowel diseases and colon cancer. Annal. Oncol. ***24***, iv116.
56. Reiskind, M.H., Griffin, R.H., Janairo, M.S., and Hopperstad, K.A. (2017). Mosquitoes of field and forest: the scale of habitat segregation in a diverse mosquito assemblage. Med. Vet. Entomol. ***31***, 44–54.
57. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics ***26***, 139–140.
58. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. ***15***, 550.
59. Ventero, M.P., Espinosa, N., Jover, R., et al. (2021). Evolution of intestinal microbiome in a process of faecal microbiota transplantation (FMT) in a patient with Clostridioides difficile infection: NGS analysis with different bioinformatics software programs. Enferm. Infecc. Microbiol. Clín. ***39***, 184–187.
60. Shreiner, A.B., Kao, J.Y., and Young, V.B. (2015). The gut microbiome in health and in disease. Curr. Opin. Gastroenterol. ***31***, 69–75.
61. He, J., Xu, S., Zhang, B., et al. (2020). Gut microbiota and metabolite alterations associated with reduced bone mineral density or bone metabolic indexes in postmenopausal osteoporosis. Aging ***12***, 8583–8604.
62. Sunagawa, S., Coelho, L.P., Chaffron, S., et al. (2015). Structure and function of the global ocean microbiome. Science ***348***, 1261359.
63. Grice, E.A., and Segre, J.A. (2011). The skin microbiome. Nat. Rev. Microbiol. ***9***, 244–253.
64. Fierer, N. (2017). Embracing the unknown: disentangling the complexities of the soil microbiome. Nat. Rev. Microbiol. ***15***, 579–590.
65. Oksanen, J., Blanchet, F.G., Friendly, M., et al. (2020). Vegan: Community Ecology Package.R package version 2.6-4. https://cran.r-project.org/package=vegan.
66. Joseph, T.A., Shenhav, L., Xavier, J.B., et al. (2020). Compositional Lotka-Volterra describes microbial dynamics in the simplex. PLoS Comput. Biol. ***16***, e1007917.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (32270677).

AUTHOR CONTRIBUTIONS

G.Y. and S.X. planned the study and drafted the manuscript. S.X. analyzed and interpreted the data, and revised the manuscript. L.Z, W.T., and Q.W. were responsible for data collection, data analysis, and revising the manuscript. Z.D., L.Z., and T.F. contributed to data analysis and interpretation. M.C., T.W, and E.H. participated in manuscript revision. All authors

have given final approval for the manuscript to be published and have agreed to be responsible for all aspects of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

SUPPLEMENTAL INFORMATION

It can be found online at https://doi.org/10.1016/j.xinn.2023.100388.

LEAD CONTACT WEBSITE

http://yulab-smu.top/.