nature protocols

PROTOCOL

https://doi.org/10.1038/s41596-019-0264-1

Using MicrobiomeAnalyst for comprehensive statistical, functional, and meta-analysis of microbiome data

Jasmine Chong[1], Peng Liu[1], Guangyan Zhou[1] and Jianguo Xia[1,2,3,4]*

MicrobiomeAnalyst is an easy-to-use, web-based platform for comprehensive analysis of common data outputs generated from current microbiome studies. It enables researchers and clinicians with little or no bioinformatics training to explore a wide variety of well-established methods for microbiome data processing, statistical analysis, functional profiling and comparison with public datasets or known microbial signatures. MicrobiomeAnalyst currently contains four modules: Marker-gene Data Profiling (MDP), Shotgun Data Profiling (SDP), Projection with Public Data (PPD), and Taxon Set Enrichment Analysis (TSEA). This protocol will first introduce the MDP module by providing a step-wise description of how to prepare, process and normalize data; perform community profiling; identify important features; and conduct correlation and classification analysis. We will then demonstrate how to perform predictive functional profiling and introduce several unique features of the SDP module for functional analysis. The last two sections will describe the key steps involved in using the PPD and TSEA modules for meta-analysis and visual exploration of the results. In summary, MicrobiomeAnalyst offers a one-stop shop that enables microbiome researchers to thoroughly explore their preprocessed microbiome data via intuitive web interfaces. The complete protocol can be executed in ~70 min.

Introduction

Rapid advances in high-throughput sequencing technologies have profoundly changed the study of microbiomes across diverse environments[1–3]. Here, the term 'microbiome' refers to the set of microorganisms inhabiting a specific biological niche, including their genomic content and metabolic products[4]. Microbiomes are either host associated—in which case microorganisms inhabit higher organisms such as humans, animals, and plants—or free living such as microbial assemblages found in water and soil. It is now widely accepted that microbial communities are critical components of their ecosystems, and disruption of these communities can be detrimental. For instance, the human microbiome has been shown to greatly influence development, immunity, and even behavior of their hosts[5]. As a result of their biomedical importance and translational potential, the past decade has witnessed a tremendous growth in the number and scale of microbiome studies. Three approaches have been commonly used to study microbiomes: (i) marker gene surveys to gain an overview of community structure, (ii) shotgun metagenomics to understand a microbiome's functional potential, and (ii) metatranscriptomics to measure its functional activities through gene expression profiling. Several powerful pipelines—such as quantitative insights into microbial ecology (QIIME)[6], mothur[7], UPARSE[8], divisive amplicon denoising algorithm 2 (DADA2)[9], One Codex[10], Kraken[11], and metagenomic phylogenetic analysis (MetaPhlAn)[12]—can preprocess raw sequencing reads into feature abundance tables. These tables, together with associated sample information (i.e., metadata), are the main inputs for downstream statistical analysis and functional interpretation.

Microbiome data present several key analytical challenges. First, differences in the number of sequencing reads per sample (i.e., library size) are often very large, requiring proper data normalization before meaningful statistical analysis can be applied. Second, abundance tables at the lowest taxonomic levels are often very sparse. This sparsity may arise from either under-sampling or true absence of taxa. Third, microbiome data is compositional[13]. If a dominant feature increases, the relative abundance (proportion) of all other features will decrease, even though their absolute

[1]Institute of Parasitology, McGill University, Ste-Anne-de-Bellevue, Quebec, Canada. [2]Department of Animal Science, McGill University, Ste-Anne-de-Bellevue, Quebec, Canada. [3]Department of Microbiology & Immunology, Montreal, Quebec, Canada. [4]Department of Human Genetics, Montreal, Quebec, Canada. *e-mail: jeff.xia@mcgill.ca

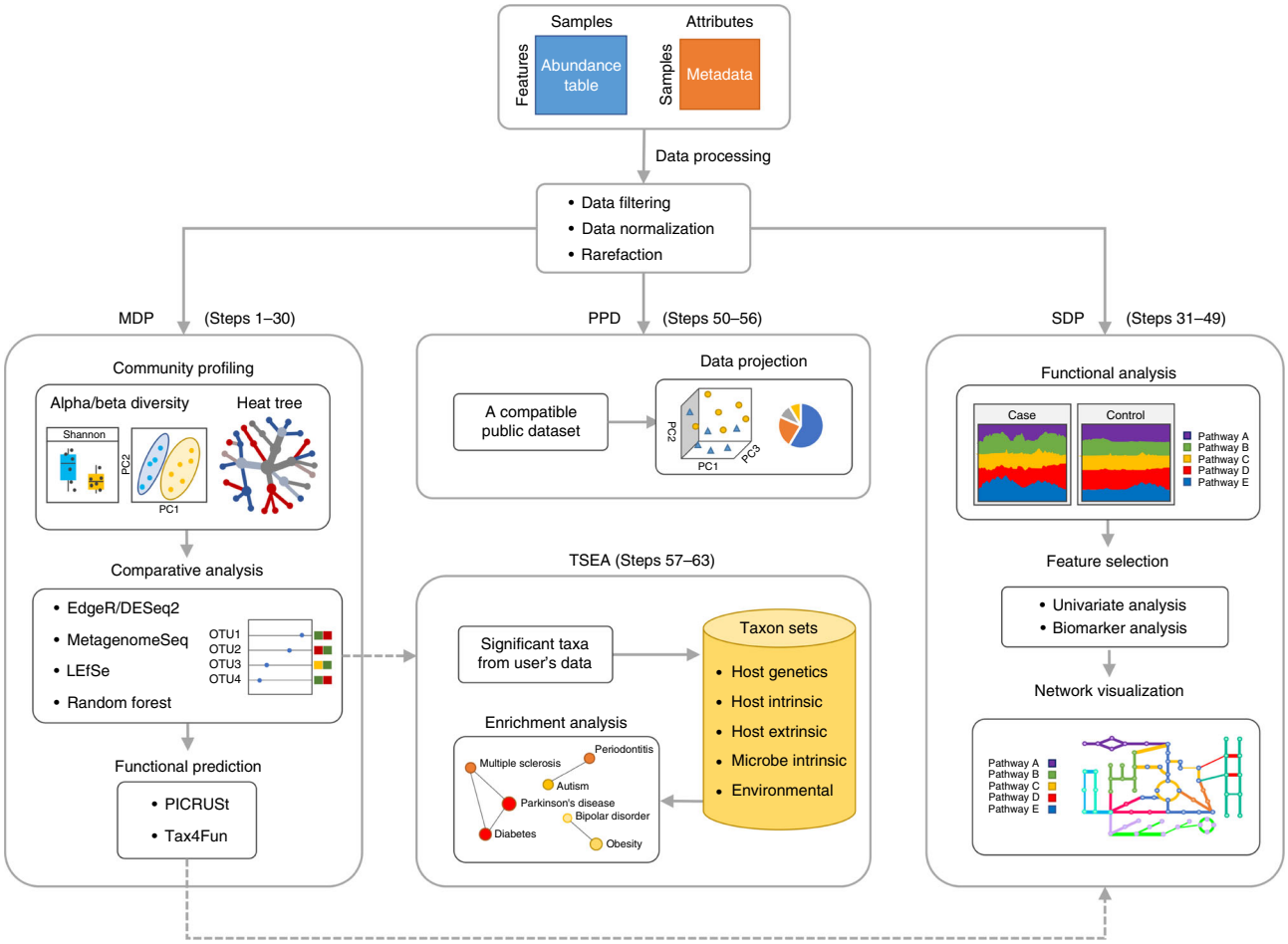


Fig. 1 | Overview of the MicrobiomeAnalyst workflow. MicrobiomeAnalyst comprises four modules: Marker-gene Data Profiling (MDP), Shotgun Data Profiling (SDP), Projection with Public Data (PPD), and Taxon Set Enrichment Analysis (TSEA). The key functions of each module are illustrated in their respective boxes. PC, principal coordinate.

abundances remain constant. These inherent characteristics of microbiome data must be considered in downstream statistical analyses. Novel statistical methods that account for such characteristics are required for a proper analysis of microbiome data. Most newly developed methods are available in the R programming language and software environment. In particular, the R package phyloseq[14] has provided a wealth of functions for manipulating feature tables, taxonomic trees, and sample metadata. Although R is incredibly powerful and flexible, learning R can be challenging for clinicians and bench researchers.

MicrobiomeAnalyst[15] was developed as a web-based tool to enable microbiome researchers to effortlessly perform comprehensive statistical analysis, interactive visualization, and meta-analysis of microbiome data without prior coding expertise. Users can choose from a wide array of well-established methods and explore the results in real time to gain a better understanding of their data. Since its initial publication in 2017[15], MicrobiomeAnalyst has gradually become popular among microbiome researchers. Over the past 12 months, the web server has processed >70,000 data analysis jobs submitted from >20,000 users worldwide. We have been actively improving the current features and adding new functions based on users' feedback and developments in the field. To meet the growing user traffic and computational demand, the server has been recently migrated to a high-performance Google Cloud platform.

Overview of the analysis workflow and the interface design

The overall workflow of MicrobiomeAnalyst is depicted in Fig. 1. There are four modules: (i) Marker-gene Data Profiling (MDP), which is dedicated to the analysis of marker-gene survey data; (ii) Shotgun Data Profiling (SDP), for the analysis of shotgun metagenomics or metatranscriptomics data; (iii) Projection to Public Data (PPD), for visual comparison of users' marker-gene data with a

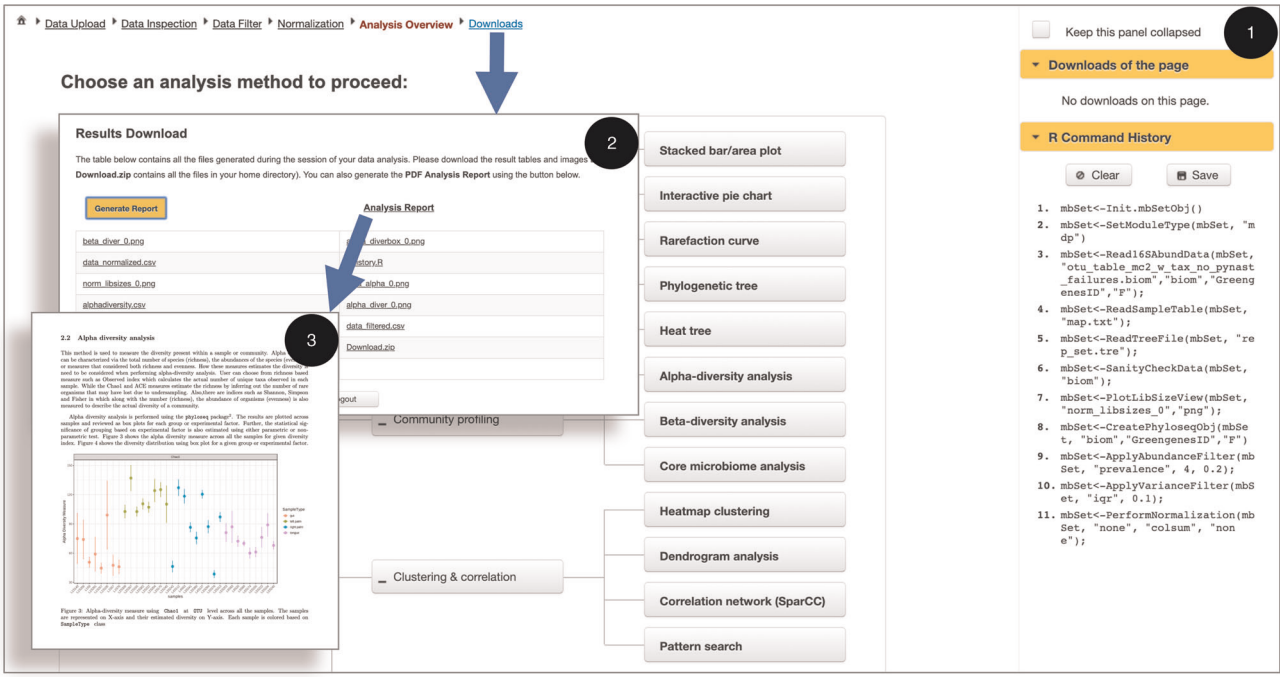


Fig. 2 | Comprehensive data analysis and report generation. A screenshot of the MDP 'Analysis Overview' page (1) to illustrate the comprehensive set of analysis methods available. The top left corner shows the navigation track with the current page highlighted in red. The 'R Command History' panel to the right of the page displays all underlying R commands. The 'Downloads of the page' panel displays the results generated from the current page. Users can also click the 'Downloads' link from the top navigation track to enter the 'Results Download' page (2) and batch-download all results as well as to generate a comprehensive analysis report (3).

compatible public dataset available in MicrobiomeAnalyst; and (iv) Taxon Set Enrichment Analysis (TSEA), which is used to test whether certain predefined groups of taxa (taxon sets) are statistically overrepresented in a list of taxa submitted by users. The four modules share the same general workflow—data preparation, followed by data analysis and visual exploration. In the data preparation stage, the user's data are uploaded for filtering and normalization. Following this, a wide variety of statistical and visualization methods can be performed on the processed data to detect, for example, overall patterns, significant features, potential interactions, and functional insights. In the case of the MDP module, a total of 19 carefully selected methods are provided (Fig. 2 (1)). The web interface for each method allows users to adjust the key parameters for interactive analysis and visual exploration of the results.

MicrobiomeAnalyst uses a dynamic navigation track and real-time system messages to guide users through each step of their data preparation and analysis. As users proceed, the completed steps will be added to the navigation track at the top of the page (Fig. 2 (1)). Users can click on any hyperlink on this track to return to the specified page. Upon completion of their analysis, users can click the 'Downloads' link at the end of the navigation track to enter the 'Results Download' page (Fig. 2 (2)) and batch-download all results and images generated during the analysis. This page also allows users to generate a comprehensive analysis report describing all the steps performed together with detailed introductions to the corresponding methods and their associated outputs (Fig. 2 (3)). The system messages provide real-time feedback and recommendations if errors occur. On the right-hand side is the 'R Command History' panel, which displays the underlying R commands as they occur in real time. Users can install the underlying R package (MicrobiomeAnalystR) from GitHub (https://github.com/xia-lab/MicrobiomeAnalystR) and use these R commands to reproduce their results locally. This feature was recently added to help improve the transparency, flexibility, and reproducibility of microbiome data analysis following the same concept as our MetaboAnalyst web server[16] and its companion MetaboAnalystR package[17].

Comparison with other web-based tools

Metagenomics rapid annotations using subsystems technology (MG-RAST)[18], visualization analysis of microbial population structures (VAMPS)[19], and Calypso[20] are three popular web-based platforms

Table 1 | Comparisons of MicrobiomeAnalyst with other web-based tools for microbiome data analysis

	MicrobiomeAnalyst	MG-RAST	VAMPS	Calypso
URL	https://www.microbiomeanalyst.ca/	http://www.mg-rast.org	https://vamps2.mbl.edu/	http://cgenome.net/calypso
Data input	16S rRNA, metagenomics	16S rRNA, metagenomics	16S rRNA	16S rRNA, metagenomics
File format	Count tables, BIOM, mothur	Sequences	Sequences	Count tables, BIOM, mothur, QIIME format
Real-time interaction	+++++	++	+++	++++
Data filtering and normalization	+++++	+++	++++	++++
Diversity profiling	Yes	Yes	Yes	Yes
Comparative analysis	+++++	+++	+++	+++++
Correlation analysis	+++++	–	–	++++
Functional prediction	PICRUSt, Tax4Fun	–	–	–
Functional annotation	KEGG, COG	SEED, KEGG, COG	–	–
Pathway visualization	Interactive KEGG metabolic network	KEGG mapper	–	–
Co-analysis with public data	37 public datasets	>44, 000 public datasets	–	–
Taxon set enrichment analysis	>2,000 taxon sets	–	–	–
Features for reproducibility	Analysis report, R command history and R package	Repository for public projects and R package	Repository for public projects	–

Some features are assessed using the symbols '–' for absent and '+' for present, with more '+' symbols indicating better support.

for microbiome data analysis. MG-RAST is a public resource for annotation and storage of raw metagenomics data. The web interface also provides some basic support for statistical analysis and visualization. More advanced analysis can be achieved by using its associated matR package (https://mg-rast.github.io/matR). VAMPS is dedicated to the visualization of microbial communities through various approaches such as heatmaps, pie charts and principal coordinates analysis (PCoA) plots. Calypso is an easy-to-use tool that supports data processing as well as diversity, comparative, and network analyses of microbiome data. In comparison to these tools, MicrobiomeAnalyst, through its modern interface design and high-performance implementation, offers a real-time visual analytics experience that enables researchers to easily navigate the complex tasks of microbiome data processing, analysis and interpretation. For instance, the MDP module currently offers 19 carefully selected methods for statistical analysis and visualization. The taxon set enrichment analysis is also a feature unique to MicrobiomeAnalyst. Another feature that is highly appreciated by the users of MicrobiomeAnalyst is its publication-ready graphical outputs created throughout their data analysis. MicrobiomeAnalyst strives for transparency and reproducibility by providing a comprehensive analysis report and R command history, as well as the release of its companion R package. Detailed comparisons between MicrobiomeAnalyst and these three web-based tools are shown in Table 1.

Limitations

Preprocessing of raw sequencing data is not supported by the MicrobiomeAnalyst web server, which focuses on real-time interactive data analysis. Owing to Internet bandwidth and server memory constraints, it would be impractical for users to preprocess their raw sequencing data on MicrobiomeAnalyst in real time. To partially address this limitation, we have also developed the MicrobiomeAnalystR package, which integrates DADA2 (ref. [9]) to enable users to preprocess raw sequencing reads into amplicon sequence variant (ASV)[21] abundance tables suitable for downstream statistical analysis. More details surrounding raw data preprocessing and commonly used pipelines are available in Box 1. MicrobiomeAnalyst was primarily developed for analysis of cross-sectional microbiome data and lacks functionality for time-series data analysis. We aim to fill this gap by adding new methods to evaluate the temporal stability of microbial communities to identify core, persistent, and transient groups[22]. Finally, users are required to re-upload and re-perform data processing steps each time they open a new session of MicrobiomeAnalyst. This could affect the reproducibility of some analysis results, such as classification results from 'Random Forests' or empirical *P* values from Sparse Correlations for Compositional data (SparCC)[23] analysis. We are developing a new component that will allow registered users to save their work and resume analyses at a later time[24].

Box 1 | Preprocessing of raw 16S rRNA amplicon sequencing data

This box describes the general steps and available tools for raw sequence data preprocessing.
Amplicon sequencing of marker genes is a widely used method for taxonomic profiling of microbial communities across different hosts and environments. After raw reads are obtained from sequencing platforms, bioinformatics pipelines are needed to translate raw reads into taxonomic information. Traditionally, raw reads are converted into OTUs, which are clusters of reads that meet a 97% similarity threshold. It is now generally recommended to convert raw reads to high-resolution ASVs, which can be identified on the basis of their unique biological sequences to facilitate meta-analysis across studies[21]. The main preprocessing steps of all bioinformatics pipelines are (i) quality control of sequencing reads, (ii) clustering of reads, and (iii) taxonomic assignment. The commonly used pipelines include QIIME[6], mothur[7], UPARSE[8], and, more recently, DADA2 (ref. [9]). DADA2 works by generating a parametric error model that is trained on all raw sequencing data and applies the model to correct and collapse sequencing errors into ASVs. The MicrobiomeAnalystR package integrates DADA2 for raw 16S rRNA amplicon sequencing data.

Experimental design

The protocol below is organized into four sections to showcase all four modules in MicrobiomeAnalyst: (i) a comprehensive analysis of 16S rRNA marker-gene abundance data (Steps 1–30); (ii) predictive functional profiling, followed by pathway enrichment analysis and network visualization of a Kyoto Encyclopedia of Genes and Genomes (KEGG) ortholog (KO) abundance table (Steps 31–49); (iii) visual data exploration with a public dataset (Steps 50–56); and (iv) taxon set enrichment analysis (Steps 57–63). A detailed step-by-step tutorial is available in the Procedure below.

Comprehensive analysis of 16S rRNA abundance data

The MDP module is the most heavily used module, containing more than half of all methods currently available in MicrobiomeAnalyst. Typically, the first question of microbiome data analysis is to determine whether there are any patterns within the data. Such exploratory analyses are conducted via commonly used ecological methods, including alpha- and beta-diversity analysis. Multivariate statistics can then be used to assess the robustness of such patterns. The next logical step is to identify which taxa are responsible for the observed differences. Identification of important taxa and their correlations or co-occurrence patterns can be accomplished using different univariate statistical methods or more complex multivariate procedures[23,25,26]. For well-studied microbial communities such as the human gut microbiomes, it is also possible to predict their functional potential[27,28]. The resulting gene abundance data can offer important functional insights without the need to perform shotgun metagenomics.

Functional profiling and network visualization of gene abundance data

The SDP module offers a similar set of methods for pattern discovery and comparative analysis of gene abundance data produced from either predictive functional profiling or metagenomics/metatranscriptomics. A unique feature of the SDP is its functional annotations based on modules, pathways and metabolic networks. MicrobiomeAnalyst enables users to easily visualize the distribution of these functions across samples and study conditions. It also supports explicit statistical testing to identify enriched functions[29]. Users can interactively explore the results within a metabolic network environment for further functional insights[30].

Visual comparison with a public dataset

With the increasing number of public datasets, meta-analysis has become a powerful approach for both comparison and hypothesis generation[31–33]. The PPD module is intended to enable users to visually explore their own 16S rRNA data within the context of a compatible public dataset. These public datasets are obtained mainly from Qiita[34]. Datasets selected by users for meta-analysis must share at least 20% taxonomic features for meaningful comparisons. In this module, users' and the public data are co-processed and then co-projected into an interactive 3D PCoA plot for visual comparison. Users can compare the taxonomic compositions of samples to find out which taxa are driving group separations. This enables users to contextualize their data to gain a global perspective in order to, for example, identify compositional differences across different environments[35] or populations[36].

Enrichment analysis of a list of taxa

Following comparative analysis, users will produce a list of taxa that are significantly associated with a phenotype of interest. However, such a list often lacks context for developing hypotheses or obtaining mechanistic insights. Enrichment analysis, a popular method already used for the interpretation of lists of genes[37] and metabolites[38], can be applied to gain deeper understandings from a list of taxa. However, a key obstacle is the need to create a comprehensive and meaningful collection of taxon sets, similar to gene sets or metabolite sets. To address this gap, we have manually curated 2,393 taxon sets from high-impact journals (impact factor >3) across different fields of microbiome research. These taxon sets can be downloaded from the 'Resources' page at the MicrobiomeAnalyst website. These taxon sets are further categorized into five categories: taxon sets associated with (i) host single-nucleotide polymorphisms (SNPs), (ii) host-intrinsic factors (e.g., diseases), (iii) host-extrinsic factors (e.g., diet and lifestyle), (iv) environmental factors (e.g., chemical exposures), and (v) microbe-intrinsic factors (e.g., mobility and shape).

Materials

Equipment

Computer requirements

- Browser requirements: MicrobiomeAnalyst runs on all modern major web browsers. For the best experience, we recommend Google Chrome v.75+, Firefox v.67+, Safari v.12+, or Microsoft Internet Explorer v.11+. JavaScript must be enabled in your browser.
- Internet connection requirements: a fast connection is highly recommended.
- Hardware requirements: >2 GB of RAM and screen resolution of 1,200 × 800 is preferred.

Data files

- *Input files*. The main input files for MicrobiomeAnalyst are three tab-delimited plain-text files: a feature abundance table containing read counts of features (operational taxonomic units (OTUs)/ASVs/genes) across multiple samples, a taxonomy file for those features (OTUs/ASVs), and a metadata file describing group information for those samples. MicrobiomeAnalyst also accepts BIOM files generated from the QIIME pipeline, as well as outputs from the mothur pipeline. In addition, if users would like to perform phylogenetic tree analysis or UniFrac distance-based analysis, a tree file generated from any commonly used algorithm is required. See Box 2 for more details about these file formats.
- *Example datasets*. MicrobiomeAnalyst provides multiple example datasets for testing purposes. From the data upload page of each module, users can directly use our example data from the 'Example datasets for testing' panel. Three example datasets are used in this protocol. The first dataset consists of

Box 2 | Data formatting and upload

This box describes how to prepare processed microbiome data for MicrobiomeAnalyst.
MicrobiomeAnalyst accepts abundance data generated from several commonly used bioinformatic pipelines. These files can be uploaded in plain-text format (.txt or.csv) or directly as .biom or .shared files. Users must also provide a metadata file describing group information for the same samples. The following are short descriptors of how to format the abundance, taxonomy, and metadata files for MicrobiomeAnalyst.

Abundance files (.txt/.csv)
The abundance table should be formatted so that features are in rows and samples are in columns. The first line should start with '#NAME'. If the feature names contain taxon names, ensure that the taxa levels are separated by semicolons (e.g., `Bacteria; Firmicutes; Clostridia`). If the features do not contain specific taxon names (e.g., OTU000001), a taxonomy mapping file must also be provided (see below).

Taxonomy files (.txt/.csv)
The taxonomy file should be formatted so that feature names are in the first column, beginning with '#TAXONOMY'. Each row should contain the taxonomic classification of all the features under the column subheadings 'Phylum', 'Class', 'Order', 'Family', 'Genus', and 'Species'. The feature names must match those that appear in the abundance file.

Metadata files (.txt/.csv)
The metadata file should be formatted so that the first column contains the sample names, starting with '#NAME'. Subsequent columns contain information for each sample with regard to group assignment or other experimental factors. The sample names must match those that appear in the abundance file.

43 stool samples from pediatric inflammatory bowel disease (IBD) patients and healthy controls obtained from the Integrated Human Microbiome Project (iHMP)[3]. These data were preprocessed using the DADA2 pipeline integrated within the MicrobiomeAnalystR package. These data will be used for the MDP and TSEA modules to explore microbial differences between the two groups. The second dataset consists of 21 fecal microbiome samples from a study of aging mice[39]. These data will be used first by the MDP module to generate a predicted gene abundance table that will then be used as the input for the SDP module. The third dataset consists of 26 environmental microbiome samples from arable soil across North and South America[40]. This dataset is intended to be used with the PPD module to perform meta-analysis with other microbiome datasets.

Equipment setup

Download the example data

Go to the MicrobiomeAnalyst homepage (https://www.microbiomeanalyst.ca) and click "Resources" from the top menu bar. From the "Example Datasets" tab, click on each zipped folder to save it on your computer. After they are downloaded, unzip each folder so that all files will be accessible for uploading to MicrobiomeAnalyst.

Procedure

Stage 1: Comprehensive analysis of 16S abundance data ● Timing ~30 min, depending on the size of the dataset

1 *Startup*. Go to the MicrobiomeAnalyst home page (https://www.microbiomeanalyst.ca) and click the 'Marker Data Profiling (MDP)' circle to enter the MDP module.
? TROUBLESHOOTING

2 *Uploading data*. Detailed instructions for preparing input files can be found in Box 2. Upon entering the MDP module, click 'Example data sets for testing' to expand the panel containing all available example datasets. Select the 'Pediatric IBD' dataset listed under 'Data Type'. Click the 'Submit' button to upload the data. Alternatively, choose the 'Plain text table format' panel. Click the 'Choose File' button next to 'OTU/ASV table' and locate the 'ibd_asv_table.csv' file. Repeat this step for the 'ibd_meta.csv', 'ibd_taxa.txt', and 'ibd_tree.tre' files. From the 'Taxonomy labels' dropdown menu, click 'Greengenes Taxonomy'. Click the 'Submit' button to upload the data.
? TROUBLESHOOTING

3 *Data integrity check*. This page consists of two tabs. The first tab, 'Text Summary', provides a text summary of the uploaded files. The second tab, 'Library Size Overview', graphically describes the read counts for all uploaded samples, which is informative for downstream data filtering and normalization. Click 'Proceed' at the bottom of the page to move forward.
? TROUBLESHOOTING

4 *Data filtering*. Filtering is generally recommended to remove low-quality features, thereby improving downstream statistical analysis. Keep the default selections for the 'Low count filter' and 'Low variance filter' sliders and click 'Submit' to perform data filtering. A message will appear in the upper-right corner, indicating the results of the data filtering step. Note that the filtered data will not be used for alpha-diversity analysis and users can turn off the filters by dragging the corresponding slider to zero value. More details on data filtering can be found in Box 3. Click 'Proceed' at the bottom right of the page to navigate to the next page.

5 *Data normalization*. On the 'Data Normalization' page, users can perform data rarefying, scaling, and transformation. The aim of data normalization is to standardize the data to enable accurate comparisons. More details can be found in Box 3. Keep the default selections for the options (only 'Data scaling' set to 'Total sum scaling') and click 'Submit', followed by 'Proceed' to move to the 'Analysis Overview' page.

6 *Community profiling*. Users can evaluate microbial community diversity profiles using the 'Alpha-diversity' and 'Beta-diversity' analysis options (refer to Box 4 for further details). To start, click 'Alpha-diversity analysis' from the 'Analysis Overview' page.

7 *Alpha diversity*. At the top of the page are several drop-down menus where users can explore different alpha-diversity measures or choose a taxonomic level to evaluate diversity differences. By default, alpha diversity is evaluated at the feature (OTU/ASV) level using Chao1, and significant differences are evaluated using *t*-tests. The bottom half of the page contains two graphical summaries of the results. To the left is a dot plot displaying the alpha-diversity measures across samples, and to the right is a box plot summarizing the alpha-diversity measures across groups.

Box 3 | Data filtering and normalization

This box describes the different approaches available in MicrobiomeAnalyst for data filtering and normalization. Microbiome data are affected by various sources of systematic variation arising from sample preparation to sequencing. Filtering and normalization aim to remove or reduce such systematic variability[52]. The merits and pitfalls of the most commonly used methods are further discussed below. The choice of method is dependent on the type of analyses to be performed[53].

Data filtering

The purpose of data filtering is to remove low-quality and/or uninformative features to improve downstream statistical analysis. MicrobiomeAnalyst offers three data-filtering procedures (i) minimal data filtering (applied to all analyses), which removes features containing all zeros or appearing in only one sample (considered artifacts); (ii) low-count filtering, which removes features that may exist due to sequencing errors or low-level contamination; and (iii) low-variance filtering, which removes features unlikely to be associated with the conditions under study. The last two options are not used for within-sample profiling (alpha diversity) but are highly recommended for comparative analysis.

Data rarefying

Rarefying is commonly used to account for uneven library sizes. This method works by randomly subsampling without replacement to the size of the smallest library that is not considered defective. It has been criticized because of potential loss of useful information[54]. However, the procedure has been shown to be useful for very small (<1,000 reads/sample) or very uneven library sizes between groups (>10×)[53], as well as important for comparing ecological communities (beta diversity)[55].

Data scaling

Scaling involves multiplying feature counts by a sample-specific factor to account for uneven sequencing depth, transforming raw reads into relative abundances. The most commonly used method is total sum scaling (TSS), whereby count data are divided by the total number of reads in each sample. This method has been criticized because the total number of reads can be dominated by a few most abundant features, which biases resulting relative abundances[56]. Moreover, TSS does for not account for heteroskedasticity of feature variance across measured values[53,57]. Other scaling factors, such as upper quantile (UQ)[58] and cumulative sum scaling (CSS)[59], have been proposed to address such issues. In particular, when performing differential abundance analysis, CSS has been recommended for controlling the FDR in data with large group sizes[52]. However, when performing community-level comparisons, such as estimation of beta diversity, TSS is recommended because it most accurately captures the composition of the original communities, whereas UQ and CSS distort communities[53,55].

Data transformation

The aim of data transformation is to stabilize the variance of the data. The centered log ratio (CLR) is commonly used and is recommended because of the compositionality of microbiome data[13]. Furthermore, its variants, relative log expression (RLE) and trimmed mean of *M* (mean) values (TMM), have consistently demonstrated high performance in identifying differentially abundant features[54–56].

From these results, it can be seen that the within-sample diversities of the pediatric IBD patients and the healthy controls are significantly different: the alpha-diversity measures are significantly lower in the IBD patients compared to the values in the controls.

8 (Optional) Explore different alpha-diversity measures; each one makes different assumptions about the community structure and will therefore reveal different aspects of the community structure (refer to Box 4 for further details). Also try different taxonomic levels to see whether the same trend can be observed across higher taxonomic levels.

9 *Beta diversity*. Click the 'Analysis Overview' link on the navigation track at the top of the page. Next, click 'Beta-diversity analysis'. The top half of this page contains parameters for beta-diversity analysis (refer to Box 4 for further details). The two tabs on the bottom of the page show 2D and 3D PCoA plots, respectively. By default, the difference in diversity between pediatric IBD patients and controls is assessed using the Bray–Curtis index. The permutational multivariate analysis of variance (PERMANOVA) suggests that the clusters for the two groups are significantly different (P value < 0.001).

10 *3D PCoA exploration*. Click the 'Interactive PCoA 3D' tab to further explore the PCoA results in an interactive 3D scatter plot based on the first three components (Fig. 3). Use your mouse to rotate and zoom in and out of the plot. Again, there is a clear separation between the two groups.

11 Double-click on several data points (representing samples) from the IBD group (red) to view the corresponding pie charts of that samples' taxonomic abundance. Users can change the taxonomic level of the pie chart as well as merge small (below a user-specified cutoff) taxa. Change the taxonomic level to 'Genus' and click 'Update'.

12 Double-click on several control samples and view their corresponding pie charts. Notice the taxonomic differences between the healthy controls and IBD patients. For instance, it seems that the

Box 4 | Alpha and beta diversity

This box describes the alpha- and beta-diversity analyses available in MicrobiomeAnalyst for community profiling.

Alpha diversity is a measure of within-sample diversity, whereas beta diversity is a measure of between-sample diversity. Alpha-diversity measures can be considered summary statistics of the diversity of single samples, whereas beta-diversity estimates can be considered dissimilarity scores between pairs of samples[60]. For the latter, these measures permit further analyses via clustering or dimensionality reduction techniques. Various statistical tests can be applied to evaluate whether the differences are significant. More details are available below.

Alpha diversity

Alpha diversity summarizes both the species richness (total number of species) and/or evenness (abundance distribution across species) within a sample[60]. Six alpha-diversity measures are currently supported in MicrobiomeAnalyst, each assessing different aspects of the community. 'Observed' calculates the total number of features per sample, whereas 'ACE' and 'Chao1' estimate taxa richness by accounting for features that are undetected because of low abundance. 'Shannon' and 'Simpson' take both species richness and evenness into account, with varying weight given to evenness. Finally, 'Fisher' models the community abundance structure as a logarithmic series distribution.

Beta diversity

Beta diversity evaluates differences in the community composition between samples. Resulting beta-diversity estimates can be combined into a distance matrix and used for ordination to visualize patterns. Samples close to each other are more similar in their microbial community profiles. MicrobiomeAnalyst supports the five most commonly used beta-diversity measures. 'Jaccard distance' uses just the presence or absence of features to calculate differences in microbial composition; 'Bray-Curtis dissimilarity' uses abundance data and calculates differences in feature abundance; 'Jensen-Shannon divergence' assesses the distance between two probability distributions that account for the presence and abundance of microbial features; 'Unweighted UniFrac' and 'weighted UniFrac' use the phylogenetic distance between features – the former is based purely on phylogenetic distance, whereas the latter is further weighted by the relative abundance of features.

Beta-diversity measures can be visualized using either PCoA or nonmetric multidimensional scaling (NMDS). Both methods take the distance matrix as input; PCoA maximizes the linear correlation between samples, whereas NMDS maximizes the rank-order correlation between samples[61]. Users should use PCoA if distances between samples are so close that a linear transformation would suffice. NMDS is suggested if users wish to highlight the gradient structure within their data[61,62]. NMDS is iterative and may return different results for the same dataset. Furthermore, MicrobiomeAnalyst calculates a stress value for the NMDS plot, which is a measure of goodness of fit. Generally, values >0.2 suggest a poor fit, whereas values <0.1 indicate a good fit.

Ordination measures between the groups are assessed for their statistical significance using either PERMANOVA, analysis of group similarities (ANOSIM) or homogeneity of group dispersions (PERMDISP). These tests evaluate global differences in microbiome composition between groups. PERMANOVA tests whether the centroids of all groups are equivalent. It uses the distances (or dissimilarity) between samples of the same group and compares them to the distances between groups[63,64]. This method is sensitive to multivariate dispersions; therefore, PERMDISP should also be used to evaluate whether the dispersion (or variation) between samples differs from the dispersion between groups[63,64]. ANOSIM tests whether within-group distances are greater or equal to between-group distances, using the ranks of all pair-wise sample distances.

samples from IBD patients are dominated by *Escherichia*, whereas healthy controls are dominated by *Bacteroides* (Fig. 3).

13 (Optional) By default, beta-diversity analysis is visualized using PCoA on a Bray–Curtis dissimilarity index and assessed using PERMANOVA. To gain a different perspective, change the ordination method to 'Nonmetric Multidimensional Scaling (NMDS)' and the statistical method to 'Analysis of Similarities (ANOSIM)', which are both rank-based approaches. Then change the distance method to 'Unweighted UniFrac', which uses the phylogenetic distance between features, rather than their abundance information (refer to Box 4 for more details). Click the 'Update' and explore the results.

14 *Heat tree analysis*. Return to the 'Analysis Overview' page and click 'Heat tree'. The heat tree analysis uses the hierarchical structure of taxonomic classifications to depict group-wise relative abundance for microbial communities[41]. The upper part of the page contains key parameters for creating and customizing a heat tree. Set 'Genus' as the current taxonomy level, specify 'Reingold-Tilford' for heat tree layout, keep 'Comparison' as the current view mode, and then select 'CD vs Control' for the comparison of interest (Crohn's disease versus control). Click 'Submit' to generate the corresponding the heat tree (Fig. 4). The layout of the tree may vary slightly because of the random nature of the algorithm. The difference table, which contains between-group comparisons using the non-parametric Wilcoxon test at different taxonomic levels, can be downloaded from the upper-right panel of the page.

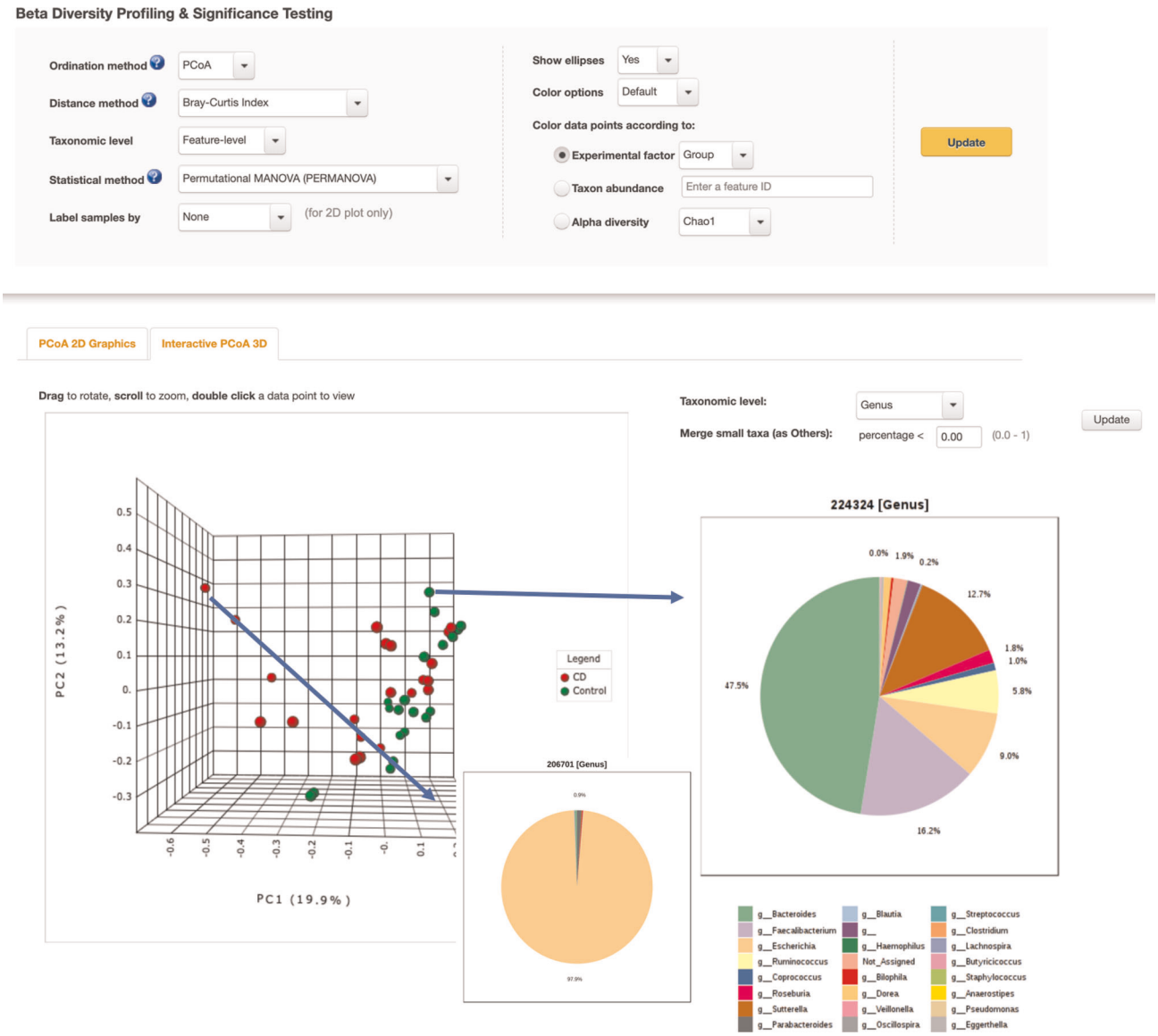
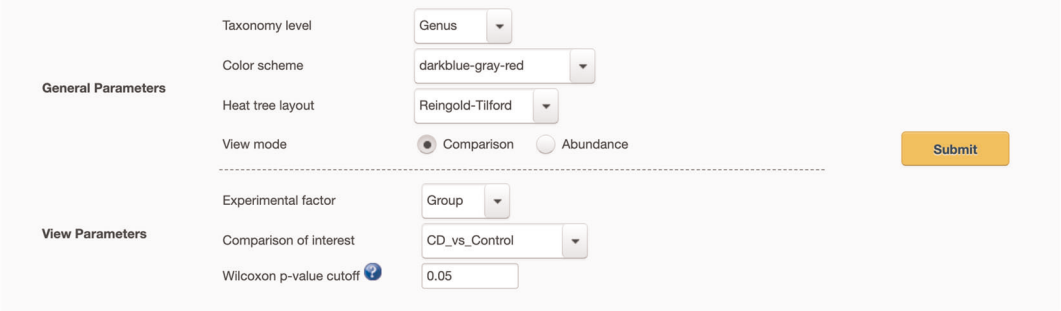


Fig. 3 | Interactive 3D PCoA plot for beta-diversity analysis. A screenshot of the 3D PCoA plot and pie charts generated by the beta-diversity analysis. Users can rotate the graph or double-click any sample to view a pie-chart summary of its microbial abundances at a selected taxonomic level. Two pie charts, one from a control sample and one from a Crohn's disease (CD) sample, are shown. The control sample is dominated by *Bacteroides*, whereas the CD sample is dominated by *Escherichia*.

15 *Correlation network analysis*. Click on 'Analysis Overview' in the top navigation track and then select 'Correlation network (SparCC)'. The correlation network analysis uses four methods to calculate pairwise correlations between taxonomic features: SparCC[23], Pearson's correlation, Spearman's rank correlation, and Kendall's tau correlation. SparCC, in particular, was designed to address the issue of spurious correlations due to the compositional nature of microbiome data (refer to Box 5 for more details). The top part of the page contains all the necessary parameters for performing correlation analysis and generating the network. To begin, make sure that 'SparCC' is selected from the 'Algorithm' dropdown menu. Select 'Genus' from the 'Taxonomy level' dropdown menu and select 'CD vs Control' from the 'Comparison of interest' dropdown menu. Keep the default thresholds for *P* value and correlation. Click 'Submit' to generate the network (Fig. 5). This step can be time intensive when there are many features and samples in the dataset.

16 *Exploring the correlation network*. In the resulting correlation network, nodes represent taxonomic features and edges represent correlations greater than the correlation threshold between pairs of taxa. By default, nodes are colored by their abundance, and the sizes of the edges reflect the strength of the correlations between taxa. To update the node coloring, select 'by Taxonomy', keep the



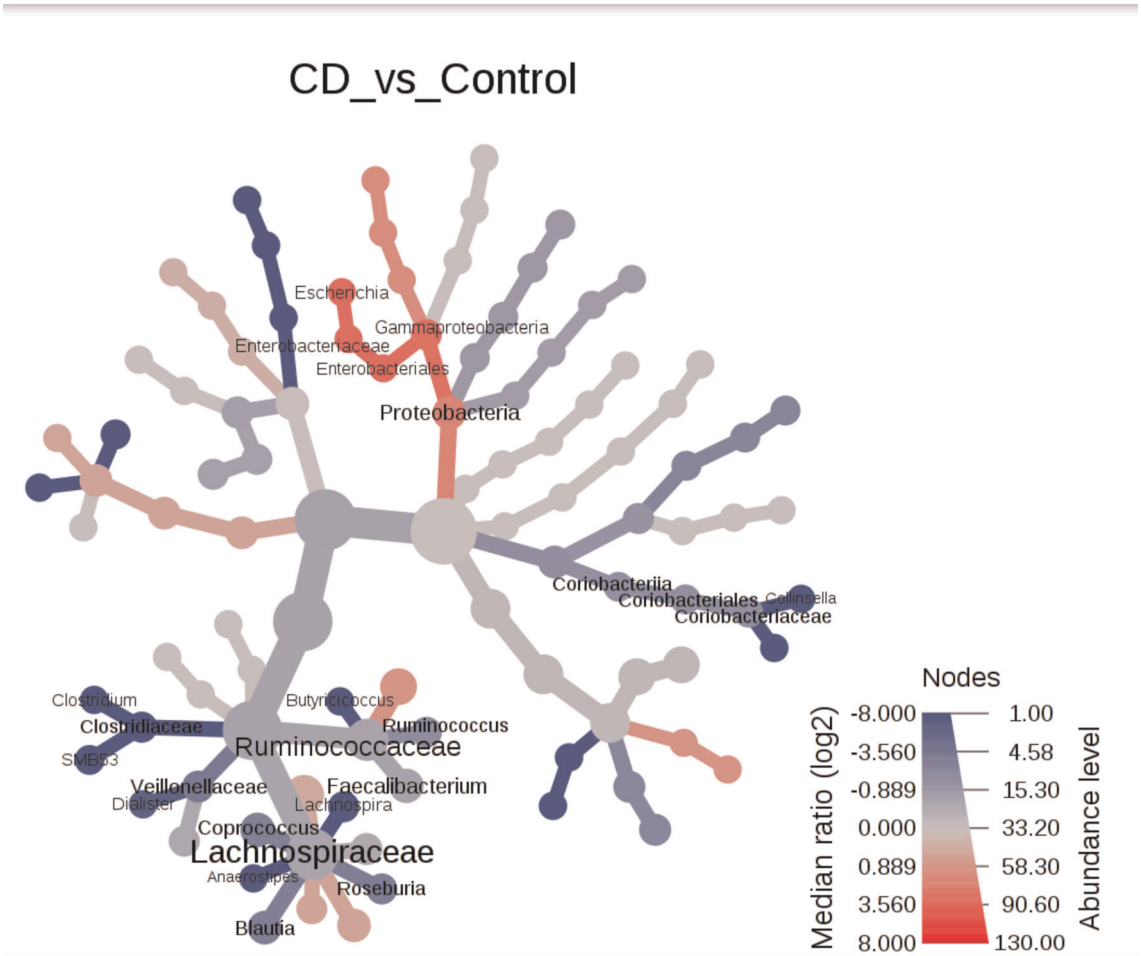


Fig. 4 | Heat tree visualization of taxonomic differences. A screenshot of a heat tree to illustrate the taxonomic differences between the two selected groups. The top of the page shows the key parameters. The color gradient and the size of node, edge, and label are based on the log2 ratio of median abundance. In this case, blue and red indicate that corresponding taxa are lower and higher, respectively, in Crohn's disease patients as compared with controls.

taxonomy level as 'Phylum', and click 'Submit' (Fig. 5). The nodes are now colored based on their phyla, with the color legend on the left of the network. The network is also interactive. For instance, double-click the yellow node *Bifidobacterium* (on the left side). A box plot comparing the abundances of this taxon between the two groups will appear to the right of the network, with numerical values below showing the correlation coefficients between the node and its closest neighbors. Finally, at the top of the network is the MD-index (microbial dysbiosis index), which is an empirical estimation of the degree of dysbiosis within the microbiome[42]. Here the MD-index is −0.67 (the value could be slightly different in different runs), which suggests an overall decrease in taxa abundance in CD patients as compared to controls.

17 (Optional) Compare the results of the correlation analysis using the other correlation algorithms (Spearman's, Pearson's or Kendall's; Box 5).

Box 5 | Correlation, comparison and classification

This box describes methods for correlation (SparCC), comparison (LEfSe) and classification (RF) analyses available in MicrobiomeAnalyst.

Correlations

The aim of correlation networks is to identify potential interactions between microbes that could represent mutualistic, commensal, parasitic or even competitive relationships[65]. Uncovering such interactions could hold important therapeutic implications for the health of the microbial community and ultimately lead to understanding microbiome function[66]. Several simple methods for computing correlation networks exist, such as Pearson's correlation, which determines whether linear relationships exist between two taxa, and Spearman's and Kendall's rank correlations, which measure rank relationships between pairs. However, these naïve methods often fail to address the compositional nature of microbiome data and can be unreliable because of the identification of spurious correlations[67]. Alternatively, compositionally robust methods such as SparCC[23] and sparse inverse covariance estimation for ecological association and statistical inference (SPIEC-EASI)[68] have been introduced, both of which make a strong assumption of a sparse correlation network[13]. SparCC uses a log ratio transformation and performs multiple iterations to identify taxa pairs that are outliers to background correlations. SPIEC-EASI uses graphical network models to infer the entire correlation network at once. Both methods are computationally intensive, although an efficient implementation of the SparCC algorithm, named FastSpar, was recently introduced[69]. MicrobiomeAnalyst implements FastSpar as well as Pearson's, Spearman's and Kendall's methods for correlation analysis.

LEfSe

LEfSe is a non-parametric statistical method developed to identify microbial taxa that are significantly different between groups[26]. LEfSe first uses the Kruskal–Wallis test to identify taxa whose relative abundances are significantly different between groups. LDA is then applied to taxa that meet the significance threshold to estimate their effect size. This approach outputs a ranked list of taxa based on their LDA scores. A significance level of $P < 0.05$ and an LDA score of 2 are often used to determine taxa that best characterize each phenotype. The original LEfSe implementation, which is available on the Huttenhower Galaxy (https://huttenhower.sph.harvard.edu/galaxy), considers the entire set of taxa (all taxonomic ranks) when performing LEfSe. In comparison, the MicrobiomeAnalyst implementation performs LEfSe only at the user's specified taxonomic level. In addition, the original LEfSe implementation uses raw *P* values when determining significant taxa. The MicrobiomeAnalyst implementation provides users the option to use either raw or FDR-adjusted *P* value cutoffs.

Random Forests

RF is a supervised machine-learning algorithm that has been applied to microbiome data for classification as well as to identify microbial taxa that differentiate between phenotypes[25,45]. RF is well suited for large and noisy data such as those from the microbiome because it is able to identify non-linear relationships, deal with variable interactions, and is robust to overfitting[70]. RF works by constructing multiple decision trees using a randomly selected subset of the training data. Each tree is formed by selecting at random, at each node, a small group of input features to split on. The class prediction is achieved via the majority vote from all trees. To evaluate the classification accuracy, 1/3 of samples are omitted during tree construction and are subsequently classified using the models to compute the out-of-bag or OOB error rates. The importance of a variable is calculated as the mean decrease in accuracy across all trees when the variable is shuffled.

18 *Classical univariate analysis*. Return to the 'Analysis Overview' page and click 'Classical univariate analysis'. MicrobiomeAnalyst offers *t*-tests/ANOVA and their nonparametric counterparts. The results for all differential-abundance analyses follow the same layout. The top half of the page contains parameters with which users can customize their analyses, such as the taxonomic level, statistical method, and significance cutoff. The bottom half of the page contains the result table for the analysis. Features in the table are ranked by their false-discovery rate (FDR)-adjusted *P* values, and those below the cutoff are highlighted in orange.

19 Click the 'Details' link under the 'View' column of the result table. A box plot will appear in a pop-up dialog, showing the abundances of the selected feature across different groups.

20 (Optional) Explore the significant features identified at different taxonomic levels.

21 *Identification of significant features using methods developed for RNA-seq data analysis*. Click on the 'Analysis Overview' from the navigation track to return to the 'Analysis Options' page.

22 Click 'RNA-seq methods' from the 'Comparison & classification' options. By default, edgeR is performed at the feature level. Compared to classical univariate analysis, edgeR identifies 54 significant features. Change the taxonomy level to 'Species' and click 'Submit'. A total of 14 species are identified as significant.

23 One of the top features is 's_coli', which stands for *Escherichia coli*. Visualize the box plot using the 'Details' hyperlink. The box plot shows a trend of *E. coli* being more abundant in CD patients as compared with healthy controls.

24 Next, select 'DESeq2' from the 'Algorithm' dropdown menu and click 'Submit'. Compared to edgeR, DESeq2 is a more conservative algorithm, and it identifies three species as significantly

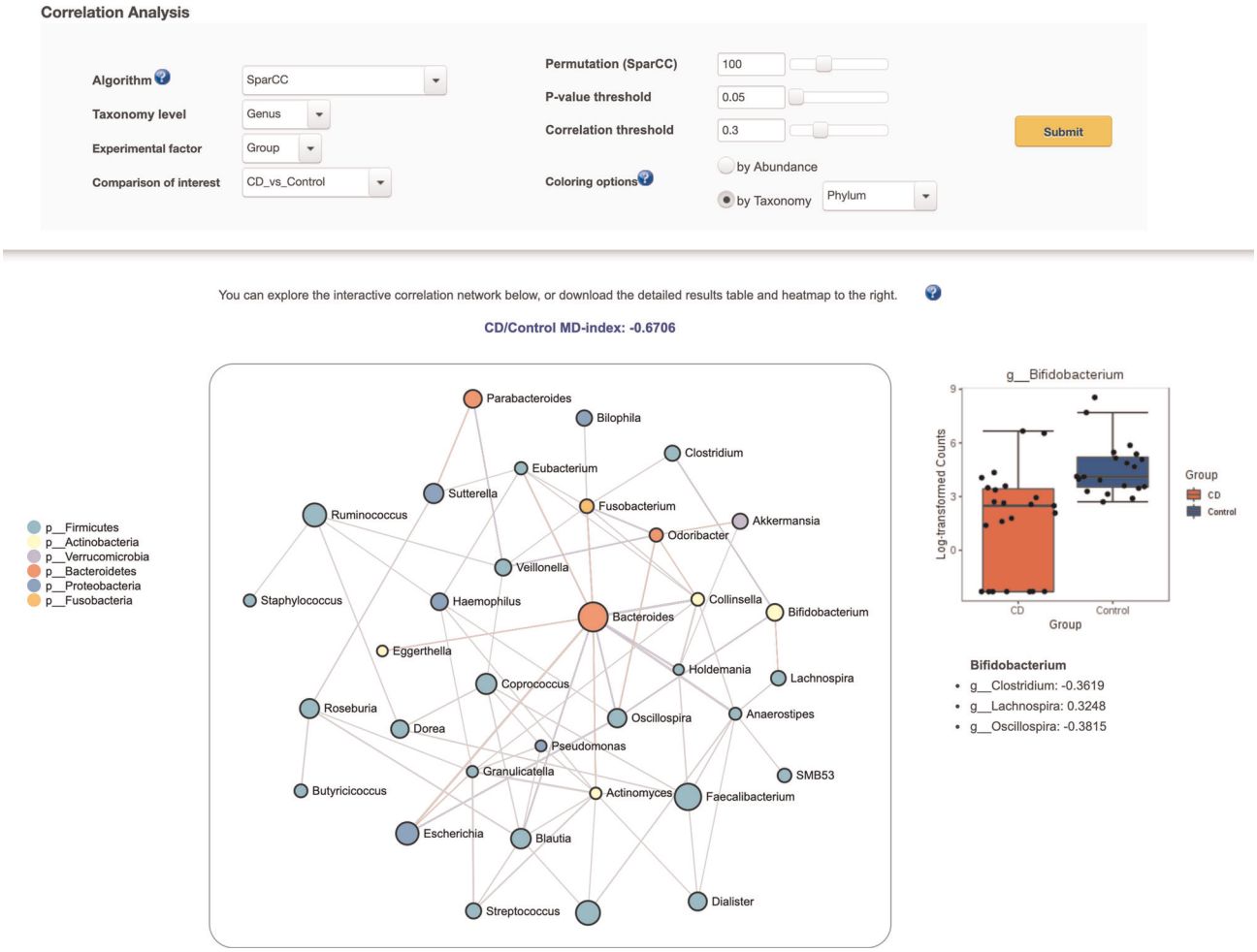


Fig. 5 | Correlation network analysis. A screenshot of the correlation network generated using the SparCC algorithm. In the center of the image is the correlation network, with nodes representing taxa at the genus level, and edges representing correlations between taxa pairs. The nodes are colored on the basis of phylum. To the right is a box plot of *Bifidobacterium* showing reduced abundance in CD patients versus healthy controls.

different between CD and control groups. All three were also identified with edgeR. For instance, *E. coli* is already implicated in IBD pathogenesis[43,44] and *Haemophilus parainfluenzae* has been shown to be increased in IBD[42].

25 (Optional) Try different taxonomic levels for further exploratory analysis. Return to the 'Analysis View' page and explore the 'metagenomeSeq' approach which was specifically designed for differential abundance analysis of marker-gene data.

26 *Biomarker discovery with linear discriminant analysis effect size (LEfSe).* Next, we will identify robust biomarkers of CD using the LEfSe approach (Box 5). Return to the 'Analysis Overview' page and click 'LEfSe'. The top half contains analysis parameters, whereas the bottom contains two tabs. The first tab is a graphical summary of the LEfSe results, whereas the second tab displays the results table. From the parameter panel, change the taxonomic level to 'Genus', the significance cutoff to '0.1' (FDR-adjusted or *q* value) and click 'Submit'. Eleven taxa are identified as significant when the following cutoffs are used: *q* value <0.1 and linear discriminant analysis (LDA score) >2.0 (Fig. 6).

27 By default, the graphical output shows a dot plot containing at most the top 15 features ranked by their LDA scores. Change the number of features included in the graphic by entering '11' in the text box next to 'Top features' and click 'Update' (Fig. 6). From the updated graphical summary, the mini heatmap to the right indicates the abundance of the microbial features across the groups. Ten taxa at genus level are decreased in CD patients as compared to healthy controls, whereas *Escherichia* is the only genus that is increased in CD patients. For a multiclass dataset, the interpretation of the plot would essentially be the same. The mini heatmap will indicate which taxa are most abundant in which groups. Users can also choose to view a bar plot summary (under the

Fig. 6 | Graphical summary of LEfSe analysis. Significant taxa are ranked in decreasing order by their LDA scores (*x* axis). The mini heatmap to the right of the plot indicates whether the taxa are higher (red) or lower (blue) in each group.

'Graphical output' drop-down menu), which uses different colors to indicate the most positively associated taxa for each phenotype.

28 *Classification using 'Random Forests'* From the 'Analysis Overview' page, click 'Random Forests'. The random forests (RF) algorithm is a powerful machine-learning method that can be applied to microbiome data for classification and selection of important features[45] (Box 5). By default, the RF model was created using 500 trees. Use the drop-down menu to set this to '5000', set the 'Taxonomic level' to 'Genus', and click 'Submit'. From the 'Classification Performance' tab, the out-of-bag (OOB) error using 5,000 trees is 0.14. This value may be different for some users because of the randomness of the algorithm (Fig. 7). The plot shows the performance of the RF model, trained on genus-level data, in predicting the sample classification to either CD or control. RF can naturally deal with multiclass datasets and will compute OOB errors and classification performance for each group.

Fig. 7 | Visualization of the 'Random Forests' results. A screenshot of the 'Random Forests' analysis results. The classification performance for each group is shown in the table to the right. Users can click the 'Important Features' tab to view those features with large impact on the accuracy of the model.

29 *Identification of significant features with RF*. Click the 'Important Features' tab to view the graphical result. The plot layout is identical to that of the LEfSe plot (Steps 26 and 27), except that features are ranked by their mean decrease accuracy. For multiclass datasets, the mini heatmap helps visualize the patterns of change across different groups. Using LEfSe and RF, *Roseburia* and *Anaerostipes caccae* are consistently identified as showing the most important differences between pediatric CD and healthy controls, marked by a decreased abundance in CD patients. Both microbes are producers of butyrate, a metabolite with known anti-inflammatory effects, and their depletion has been linked to IBD[46–48].

30 *Analysis report generation and results download*. Following the analysis, click the 'Downloads' hyperlink from the top navigation track. The 'Results Download' page will appear, showing all figures, result tables and the 'R Command History' file. Click the 'Generate Report' button to create a PDF report detailing all analyses performed and embedded with the results (Fig. 2 (3)). Click the 'Analysis Report' link to download the report. Click the 'Download.zip' link to download a zipped file containing all results generated in the analysis session.

Box 6 | Functional prediction

This box describes methods available in MicrobiomeAnalyst for predictive functional profiling.
Despite their cost effectiveness for taxonomic surveys, marker-gene data do not directly provide any functional information. Inferring potential functions from 16S rRNA sequencing data is thus greatly appealing. Two well-established methods for predictive functional profiling are available in MicrobiomeAnalyst, PICRUSt[27] and Tax4Fun[28]. PICRUSt was the first tool that popularized the method of inferring microbiome functions from 16S rRNA data. It leverages the idea that phylogenetically related organisms are more likely to have similar gene contents. From 16S rRNA data, the PICRUSt algorithm searches for the most closely related organisms with annotated genomes and assumes that their functional information is also present in the data. On the other hand, Tax4Fun is an R package that combines precomputed functional profiles from KEGG prokaryotic organisms and normalized taxonomic abundances. To use Tax4Fun, the input 16S rRNA sequencing data must be annotated using the SILVA reference database[71], whereas for PICRUSt, the Greengenes database[49] must be used. Both methods rely on available genome annotations to make inferences, and are suitable for predictive functional profiling of microbiomes from well-studied environments such as the human gut.

Stage 2: Predictive functional profiling and analysis of gene abundance data ● Timing ~20 min, depending on the dataset size.

31 *Startup*. Return to the MicrobiomeAnalyst home page and click 'Marker Data Profiling (MDP)' to enter the module.

32 *Example data upload*. From the example datasets, select 'Aging Mouse Gut' and click 'Submit'. Repeat Steps 3–5 to perform data processing.

33 *Prediction of functional potential*. Phylogenetic investigation of communities by reconstruction of unobserved states (PICRUSt) is a computational approach that can predict gene abundance from properly annotated OTU abundance tables (see Box 6 for more details). Click the 'PICRUSt (Greengenes)' hyperlink from the 'Analysis Overview' page.
▲ **CRITICAL STEP** The MicrobiomeAnalyst implementation of PICRUSt is based on the Greengenes[49] reference OTUs (18May2012 version). If users have not annotated their data to this database, the analysis will fail.

34 From the PICRUSt page, click the 'Predict Functional Potential' button.
▲ **CRITICAL STEP** This step will take 1~2 min, depending on the server load at the time.

35 Upon completion, a box plot of KO counts across all samples will appear on the page. From the 'Downloads of the page' menu, click the 'KO Table' and the 'Metadata file' options to download these files. These files will be used as input for the SDP module. The procedures described below work equally well for shotgun metagenomics and metatranscriptomics data.

36 Return to the MicrobiomeAnalyst home page and click 'Shotgun Data Profiling (SDP)' to enter the module.

37 *Data upload*. To upload the PICRUSt data, first select the 'Upload a gene abundance table' panel. From the 'Gene ID type' drop-down menu, choose 'KEGG Ortholog (KO)'. Next, click the 'Choose File' button next to 'Abundance' file to locate the 'functionalprof_picrust.csv' file. Press 'Open' to select the file. Next click the 'Choose File' button next to 'Metadata file' and locate the 'metadata.csv' file. Click 'Submit' to upload all the data. To facilitate the testing process, we have also included these data in the example datasets. To use this feature, click 'Example data sets for testing' at the bottom of the page. Select the 'KO Mouse Dataset' and the click 'Submit' to upload the data.
▲ **CRITICAL STEP** The required format is a.txt or.csv file with genes in rows and samples in columns. The accepted gene identifiers include KOs, Enzyme Commissions (ECs), and Clusters of Orthologous Groups (COGs). The first row must contain sample names and begins with '#NAME'. The same metadata file used for MDP can be used, with sample names as the first column, followed by metadata variables. Click on the 'Data Format' page for further details.

38 *Data integrity check*. The 'Data Integrity Check' page summarizes the results of the data upload. Click 'Proceed' to continue.

39 *Data filtering*. Keep the default 'low count filter' and 'low variance filter' settings and click 'Submit'. Refer to Step 4 for more details. A message will appear on the top right indicating the number of remaining features. Click 'Procced' to move forward.

40 *Data normalization*. Keep the 'Data scaling' set to 'Cumulative sum scaling' and click 'Submit'. Refer to Box 3 for details of the available normalization methods. Click 'Proceed' to continue.

41 *Analysis overview*. The 'Analysis Overview' page provides several options for functional profiling, clustering analysis, differential-abundance analysis, and biomarker analysis. Differential-abundance

analysis and biomarker analysis are covered in Steps 18–29. Here, we will show how to obtain a functional overview. Click 'Diversity overview' to begin.

42 *Functional diversity profiling*. On the 'Functional Diversity Profiling' page, users can view a graphical summary of the potential functions of their gene abundance data by binning related genes into several functional categories, including KEGG metabolism, KEGG pathways, KEGG modules, and COG functional categories. The top of the page contains parameters with which to customize the plot, such as the functional category and color scheme. The default view is KEGG metabolism using the total number of hits. From the plot, we see minor variations across the young, middle-aged, and old mice at this functional level. To explore different functional levels, select 'KEGG pathways' and 'Total hits normalized by category size'. Click 'Submit' to update the plot. From the new stacked-area plot we can see the distributions of different pathways across samples and conditions.

43 *Enrichment analysis*. We can perform enrichment analysis to statistically assess whether certain pathways or modules are significantly associated with the age factor. The enrichment is calculated using the well-established globaltest algorithm[29], which is a robust test to identify whether particular gene sets (i.e., KEGG pathways) are significantly associated with the phenotype shifts on the basis of their abundance profiles. Return to the 'Analysis Overview' page and click 'Association analysis'. A pop-up will appear. Keep the experimental factor set to 'Age' and press 'Proceed'.

44 *Visualization using the KEGG global metabolic network*. On the 'Network Viewer' page, users can visually explore the enriched pathways within the KEGG global metabolic network (Fig. 8). The page consists of three sections: the top toolbar, the left panel containing the pathway analysis results, and the central area, which displays the metabolic network. To demonstrate the utility of this page, click the checkbox next to 'Phenylalanine metabolism'. Matched KOs from user's data will now be highlighted as edges on the network, with their colors based on the highlight colors specified by the users.

45 *Network exploration*. Further explore the results of the enrichment analysis. Use your mouse to zoom in and out, as well as to drag the network in any direction. Users can double-click any highlighted edge to view the associated reactions. The bottom left corner of the page lists all matched KOs in the selected pathway. If users click on any KO, they will be directly taken to the corresponding page on the KEGG website.

46 *Network customization*. The toolbar at the top of the page contains many useful options with which users can customize their network. These include changing the background of the network, showing or hiding pathway names, and switching the overall network styles. Adjust these settings to customize the network.

47 *Further network customization*. Users can also highlight their specified pathways in different colors. For instance, click the colored box next to 'Highlight'. A color palette will appear. Directly click on a region showing the color of interest and press 'Choose' to close the dialog. Next, click on the 'Geraniol degradation' pathway from the left-hand side to highlight all matching edges.

48 *Network download*. Following the network exploration, click the drop-down menu next to 'Download' and select 'PNG Image'. A 'Download Dialog' will pop up on your screen with the created network. Right-click the PNG image and save it under your preferred name. Alternatively, users can export the KEGG network in SVG format.

49 (Optional) To further explore the gene abundance data (e.g., differential abundance analysis and biomarker analysis), follow Steps 18–29.

Stage 3: Visual data exploration with a compatible public dataset ● Timing ~10 min, depending on the dataset size.

50 *Startup*. Return to the MicrobiomeAnalyst home page and click 'Projection with Public Data (PPD)' to enter the module.

51 *Data upload*. The PPD upload page is similar to the MDP upload page. Click 'Example data sets for testing' to show all available example datasets. Select the 'Arable soil' dataset. Click the 'Submit' button to upload the data. Alternatively, click 'Choose File' next to 'ASV/OTU table' and locate 'soil_test_otu.txt'. Click 'Choose File' next to 'Metadata file' and locate 'soil_sample.txt', and click 'Choose File' next to Taxonomy table and locate 'soil_test_taxa.txt'. Specify the 'Taxonomy labels' as 'Greengenes OTU Ids'. Click 'Submit' to upload the data.

52 *Data integrity check*. The 'Data Integrity Check' page summarizes the results of the data upload. Click 'Proceed' to continue.

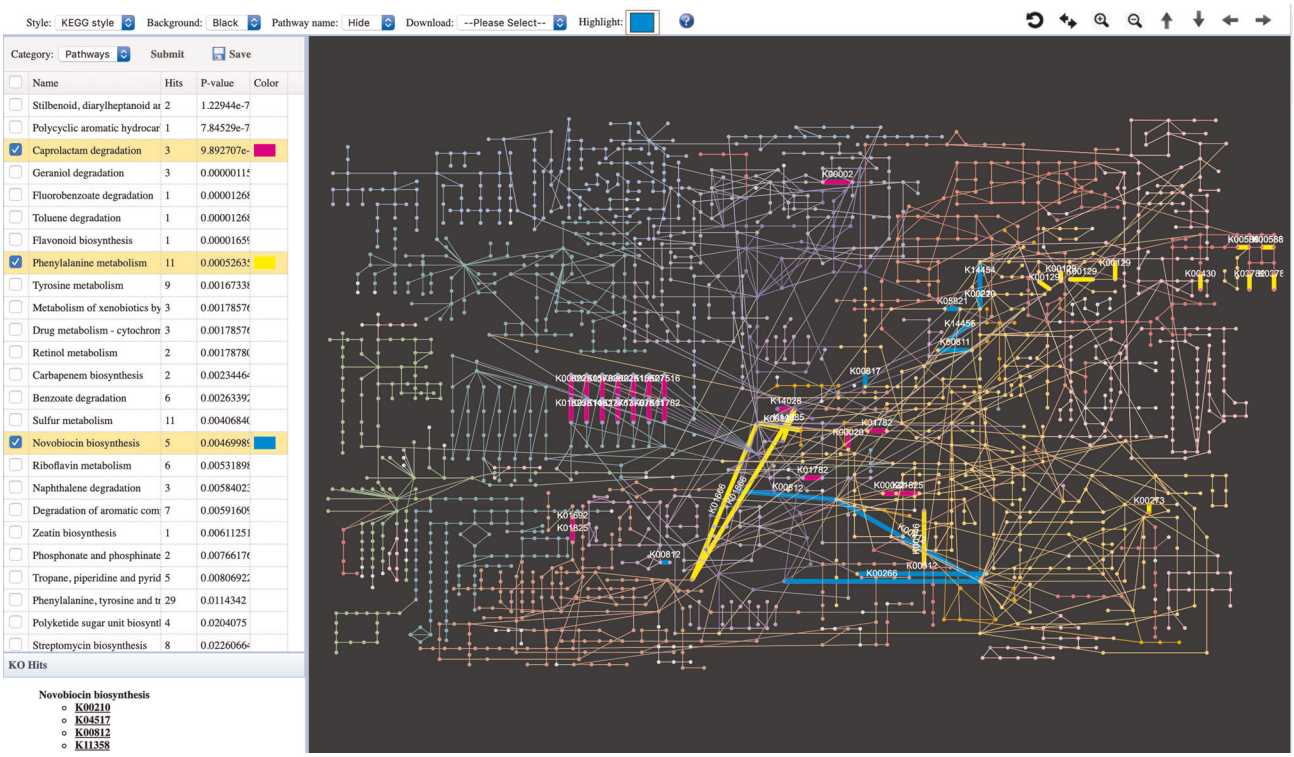


Fig. 8 | Visualization of enriched pathways in the KEGG global metabolic network. A screenshot of the KEGG global metabolic network. The top toolbar contains all options for network customization, such as background color, highlight color, and whether to show pathway names. The left panel contains the results of the enrichment analysis, and the bottom of the panel provide links to the KEGG website for all matched KOs. Selected pathways are highlighted in different colors within the network.

53 *Data selection*. The 'Data Selection' page contains all datasets available in MicrobiomeAnalyst for users to co-project with their data. The datasets are organized by body sites (for human samples), organism (samples from other mammals), and environmental samples. As the example data originate from arable soil, click the 'Environmental' tab to view all available options. Select 'Global soil' and click 'Submit'.

▲ **CRITICAL STEP** At least 20% of taxa must be shared between the user's data and the selected public dataset.

54 *Interactive data visualization*. The 3D PCoA should look similar to Fig. 3. Refer to Steps 10–12 for instructions on navigating the plot. The user's data are represented as circles, whereas the public data are squares. Use your mouse to rotate or zoom in and out of the graph. It is clear that the samples from the user's data separate into three clusters, with one cluster close to samples from dry soil and surface soil groups, while the other clusters of samples are far from all reference data.

55 (Optional) *Comparison of taxa abundances of different samples*. Double-click a data point (i.e., a sample) to view a pie chart summary of its taxa abundances (Steps 11 and 12). Note that all generated pie charts will appear in the right-hand 'View History' panel. Visually compare these pie charts to get a sense of how the samples are different across conditions at various taxonomic levels.

56 *Analysis download*. Click on the 'Downloads' link from the top navigation track to download the results.

Stage 4: Enrichment analysis of a list of taxa ● Timing ~10 min

57 *Startup*. Return to the MicrobiomeAnalyst home page and click 'Taxon Set Enrichment Analysis (TSEA)' to enter the module.

58 *Data upload*. The required format is a list of taxa. Use the example taxa list file ('ibd_taxa.txt') described in the 'Equipment setup' section. Open the 'ibd_taxa.txt' file in your preferred text editor (e.g., Notepad). Select all taxa names and copy and paste the contents into the text area of MicrobiomeAnalyst. Keep the 'Input type' as 'Mixed-level Taxon Names'. Click the 'Submit' button.

▲ **CRITICAL STEP** Users must upload their taxa list as a single column of taxon names or IDs and correctly specify the input type in order to proceed.

Taxon Set	Total	Hits	Expect	P value	FDR	Details
Type I Diabetes (increased)	13	6	0.15	1.44E-9	1.77E-7	View
Anorexic (decrease)	7	5	0.0807	2.23E-9	1.77E-7	View
Pediatric Crohn's Disease	7	5	0.0807	2.23E-9	1.77E-7	View
Liver Cirrhosis (China, decrease)	52	8	0.6	2.64E-8	1.57E-6	View
Alzheimer's (decrease)	7	4	0.0807	4.16E-7	1.98E-5	View

Fig. 9 | TSEA results. At the top of the page is an enrichment network. Users can click any node to view more details about the underlying taxon set via the 'Taxon Set View' option on the right. The result table with detailed statistical information is shown at the bottom of the page.

59 *Name mapping*. The next page shows the results of the 'Taxonomic Name/ID Mapping' function. The purpose of this page is to match taxon names from user's data to the underlying taxon set libraries of MicrobiomeAnalyst. Taxon names without hits will be highlighted in yellow and will be excluded from further analysis. Click the 'Submit' button at the bottom of the page to continue.

60 The 'Taxon Set Library' page shows all available taxon sets for enrichment analysis. There are three levels of taxon sets: 'Mixed-level' (including phylum to species), 'Species-level', and 'Strain-level'. In this use case, the taxa are a mix of genus and species names. Under the 'Mixed-level taxon sets' heading, click the 'Host-intrinsic taxon sets' and click 'Submit' to proceed to the next step.

? TROUBLESHOOTING

61 *Network exploration*. The TSEA result is shown as an enrichment network (Fig. 9). In the network, each node represents a taxon set, with its color corresponding to the *P* value and its size corresponding to the number of hits. Two nodes are connected if the number of shared taxa is >20%. On the basis of the network, 'Pediatric Crohn's Disease' receives the largest number of hits and is highly interconnected with other taxon sets such as 'Type 1 Diabetes', 'Colorectal carcinogenesis', 'Crohn's Disease', and 'Anorexic (decrease)'. Drag nodes around or use the mouse scroll to zoom in or out. Click any taxon set to view its details in the 'Taxon Set View' in the right-side panel. All matching taxa will be highlighted in red. The link to the corresponding publication is provided as a hyperlink to PubMed, as well as to where the evidence was gathered within the publication.

62 *Exploration of the TSEA results table*. Scroll down the page to view the results table. Ten taxon sets have FDR-adjusted *P* values <0.05. 'Anorexic' is one of the most enriched taxon sets. This is not unexpected, because malnutrition is a common complication of pediatric IBD, potentially stemming from anorexia[50,51]. Spend some time exploring the rest of the TSEA results.

63 *Download of results*. Click the 'Downloads' link from the top navigation track to enter the 'Results Download' page. Generate the corresponding analysis report and download the results. Click 'Logout' to exit the session.

Troubleshooting

Troubleshooting advice can be found in Table 2.

Table 2 | Troubleshooting table

Step	Problem	Possible reason	Solution
1	The home page does not display properly.	JavaScript is not enabled in your browser.	Check the documentation for your specific browser on how to enable JavaScript. For Google Chrome, click the three vertical dots in the upper-right corner of your browser and go to 'Settings'. Scroll down the page and click 'Advanced'. Under 'Privacy and security', click 'Site Settings' and then click 'JavaScript'. Turn on 'Allowed'.
2	Data upload failed.	There are issues with the uploaded files, such as missing files or incorrect formatting.	The MicrobiomeAnalyst system messages will indicate possible reasons for failed uploads. Check for the following issues and reformat and re-upload your files: (i) incorrect tab-delimited formatting (if uploading plain-text tables), (ii) selection of incorrect taxonomy labels, (iii) not using semicolons to separate taxonomic ranks in the taxonomy table, and (iv) uploading data in formats that are not currently supported.
3	Data integrity check failed.	Sample names do not match between the metadata and abundance tables.	Ensure that sample names are consistent among all uploaded files.
		There are duplicated taxonomy names in your taxonomy table.	Ensure that taxonomy names for your uploaded count table and taxonomy table match.
60	No matches are available for user's uploaded list of taxa.	Despite the large size of taxon set libraries, not all existing microbes are covered.	We will keep adding new taxon sets to expand our coverage of the microbiome.
	After some time, the server fails to respond.	The user's session has timed out (default = 45 min).	Refresh the page and re-upload your files. We are implementing a user account management system so that registered users can save and resume their analyses.

Timing

Steps 1–30, Stage 1, comprehensive analysis of 16S abundance data: ~30 min, depending on the dataset size
Steps 31–49, Stage 2, predictive functional profiling and analysis of gene abundance data: ~20 min, depending on the dataset size
Steps 50–56, Stage 3, visual data exploration with a compatible public dataset: ~10 min, depending on the dataset size
Steps 57–63, Stage 4, enrichment analysis of a list of taxa: ~10 min

Anticipated Results

This protocol enables users to perform a comprehensive analysis of their microbiome data. Three example datasets are provided: one each for pediatric IBD samples, aging mouse samples, and arable soil samples. The major graphical outputs produced during the analysis are shown in Figs. 3–9. Users are not only able to profile their microbial communities and identify important features, they can also gain functional insights through enrichment analysis and metabolic network-based visualizations. The PPD and TSEA modules further permit users to perform meta-analysis by comparing their data with either a compatible public dataset or known microbial signatures for potential validation or novel insights.

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All example datasets used in the protocol are integrated as example datasets in their respective modules and are also available for download from the 'Resources' page of MicrobiomeAnalyst

(https://www.microbiomeanalyst.ca/MicrobiomeAnalyst/docs/Resources.xhtml). There are no restrictions on their use.

Code availability

MicrobiomeAnalyst is freely accessible as a web-based application. The underlying R code is freely available at GitHub (https://github.com/xia-lab/MicrobiomeAnalystR) under a GNU General Public License v.2 or later. The code in this protocol has been peer-reviewed.

References

1. Gilbert, J. A., Jansson, J. K. & Knight, R. The Earth Microbiome project: successes and aspirations. *BMC Biol.* **12**, 69 (2014).
2. Gevers, D. et al. The Human Microbiome Project: a community resource for the healthy human microbiome. *PLoS Biol.* **10**, e1001377 (2012).
3. iHMP Research Network Consortium. 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 (2014).
4. Marchesi, J. R. & Ravel, J. The vocabulary of microbiome research: a proposal. *Microbiome* **3**, 31 (2015).
5. Gilbert, J. A. et al. Current understanding of the human microbiome. *Nat. Med.* **24**, 392–400 (2018).
6. Bolyen, E. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).
7. Schloss, P. D. et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75**, 7537–7541 (2009).
8. Edgar, R. C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* **10**, 996–998 (2013).
9. Callahan, B. J. et al. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
10. Minot, S. S., Krumm, N. & Greenfield, N. B. One Codex: a sensitive and accurate data platform for genomic microbial identification. Preprint at *bioRxiv*, https://doi.org/10.1101/027607 (2015).
11. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
12. Segata, N. et al. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9**, 811–814 (2012).
13. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* **8**, 2224 (2017).
14. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* **8**, e61217 (2013).
15. Dhariwal, A. et al. MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Res.* **45**, W180–W188 (2017).
16. Chong, J. et al. MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. *Nucleic Acids Res.* **46**, W486–W494 (2018).
17. Chong, J., Yamamoto, M. & Xia, J. MetaboAnalystR 2.0: from raw spectra to biological insights. *Metabolites* **9**, E57 (2019).
18. Wilke, A. et al. The MG-RAST metagenomics database and portal in 2015. *Nucleic Acids Res.* **44**, D590–D594 (2016).
19. Huse, S. M. et al. VAMPS: a website for visualization and analysis of microbial population structures. *BMC Bioinforma.* **15**, 41 (2014).
20. Zakrzewski, M. et al. Calypso: a user-friendly web-server for mining and visualizing microbiome–environment interactions. *Bioinformatics* 33, (782–783 (2016).
21. Callahan, B. J., McMurdie, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* **11**, 2639 (2017).
22. Baksi, K. D., Kuntal, B. K. & Mande, S. S. 'TIME': a web application for obtaining insights into microbial ecology using longitudinal microbiome data. *Front. Microbiol.* **9**, 36 (2018).
23. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8**, e1002687 (2012).
24. Zhou, G. et al. NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucleic Acids Res.* **47**, W234–W241 (2019).
25. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
26. Segata, N. et al. Metagenomic biomarker discovery and explanation. *Genome Biol.* **12**, R60 (2011).
27. Langille, M. G. et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* **31**, 814–821 (2013).
28. Aßhauer, K. P., Wemheuer, B., Daniel, R. & Meinicke, P. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics* **31**, 2882–2884 (2015).
29. Goeman, J. J., van de Geer, S. A., de Kort, F. & van Houwelingen, H. C. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* **20**, 93–99 (2004).

30. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109–114 (2012).
31. Rocca, J. D. *et al.* The Microbiome Stress Project: toward a global meta-analysis of environmental stressors and their effects on microbial communities. *Front. Microbiol.* **9**, 3272 (2018).
32. Wirbel, J. et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med* **25**, 679–689 (2019).
33. Sze, M. A. & Schloss, P. D. Looking for a signal in the noise: revisiting obesity and the Mmcrobiome. *MBio* **7**, e01018-16 (2016).
34. Gonzalez, A. et al. Qiita: rapid, web-enabled microbiome meta-analysis. *Nat. Methods* **15**, 796–798 (2018).
35. Ley, R. E., Lozupone, C. A., Hamady, M., Knight, R. & Gordon, J. I. Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat. Rev. Microbiol.* **6**, 776–788 (2008).
36. Lozupone, C. A. et al. Meta-analyses of studies of the human microbiota. *Genome Res.* **23**, 1704–1714 (2013).
37. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
38. Xia, J. & Wishart, D. S. MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res.* **38**, W71–W77 (2010).
39. Langille, M. G. et al. Microbial shifts in the aging mouse gut. *Microbiome* **2**, 50 (2014).
40. Rousk, J. et al. Soil bacterial and fungal communities across a pH gradient in an arable soil. *ISME J.* **4**, 1340–1351 (2010).
41. Foster, Z. S., Sharpton, T. J. & Grunwald, N. J. Metacoder: an R package for visualization and manipulation of community taxonomic diversity data. *PLoS Comput. Biol.* **13**, e1005404 (2017).
42. Gevers, D. et al. The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* **15**, 382–392 (2014).
43. Palmela, C. et al. Adherent-invasive *Escherichia coli* in inflammatory bowel disease. *Gut* **67**, 574–587 (2018).
44. Fang, X. et al. *Escherichia coli B2* strains prevalent in inflammatory bowel disease patients have distinct metabolic capabilities that enable colonization of intestinal mucosa. *BMC Syst. Biol.* **12**, 66 (2018).
45. Knights, D., Costello, E. K. & Knight, R. Supervised classification of human microbiota. *FEMS Microbiol. Rev.* **35**, 343–359 (2011).
46. Zhu, C. et al. *Roseburia intestinalis* inhibits interleukin−17 excretion and promotes regulatory T cells differentiation in colitis. *Mol. Med. Rep.* **17**, 7567–7574 (2018).
47. Morgan, X. C. et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* **13**, R79 (2012).
48. Riviere, A., Selak, M., Lantin, D., Leroy, F. & De Vuyst, L. *Bifidobacteria* and butyrate-producing colon bacteria: importance and strategies for their stimulation in the human gut. *Front. Microbiol.* **7**, 979 (2016).
49. DeSantis, T. Z. et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069–5072 (2006).
50. Collins, A., Nolan, E., Hurley, M., D'Alton, A. & Hussey, S. Anorexia nervosa complicating pediatric Crohn disease—case report and literature review. *Front. Pediatr.* **6**, 283 (2018).
51. Gerasimidis, K., McGrogan, P. & Edwards, C. A. The aetiology and impact of malnutrition in paediatric inflammatory bowel disease. *J. Hum. Nutr. Diet.* **24**, 313–326 (2011).
52. Pereira, M. B., Wallroth, M., Jonsson, V. & Kristiansson, E. Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC Genomics* **19**, 274 (2018).
53. Weiss, S. et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, 27 (2017).
54. McMurdie, P. J. & Holmes, S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* **10**, e1003531 (2014).
55. McKnight, D. T. et al. Methods for normalizing microbiome data: an ecological perspective. *Methods Ecol. Evol.* **10**, 389–400 (2019).
56. Dillies, M. A. et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* **14**, 671–683 (2013).
57. Hugerth, L. W. & Andersson, A. F. Analysing microbial community composition through amplicon sequencing: from sampling to hypothesis testing. *Front. Microbiol.* **8**, 1561 (2017).
58. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinforma.* **11**, 94 (2010).
59. Joseph, N., Paulson, C., Corrada Bravo, H. & Pop, M. Robust methods for differential abundance analysis in marker gene surveys. *Nat. Methods* **10**, 1200–1202 (2013).
60. Morgan, X. C. & Huttenhower, C. Chapter 12: human microbiome analysis. *PLoS Comput. Biol.* **8**, e1002808 (2012).
61. Ramette, A. Multivariate analyses in microbial ecology. *FEMS Microbiol. Ecol.* **62**, 142–160 (2007).
62. Kuczynski, J. et al. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat. Methods* **7**, 813–819 (2010).
63. Anderson, M. J. & Walsh, D. C. PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: what null hypothesis are you testing? *Ecol. Monog.* **83**, 557–574 (2013).
64. Anderson, M. J. Permutational multivariate analysis of variance (PERMANOVA). *Wiley StatsRef: Statistics Reference Online*, https://doi.org/10.1002/9781118445112.stat07841(2014).

65. Faust, K. et al. Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* **8**, e1002606 (2012).
66. Layeghifard, M., Hwang, D. M. & Guttman, D. S. Disentangling interactions in the microbiome: a network perspective. *Trends Microbiol.* **25**, 217–228 (2017).
67. Pearson, K. Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proc. R. Soc. Lond.* **60**, 489–498 (1897).
68. Kurtz, Z. D. et al. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* **11**, e1004226 (2015).
69. Watts, S. C., Ritchie, S. C., Inouye, M. & Holt, K. E. FastSpar: rapid and scalable correlation estimation for compositional data. *Bioinformatics* **35**, 1064–1066 (2018).
70. Touw, W. G. et al. Data mining in the life sciences with Random Forest: a walk in the park or lost in the jungle? *Brief. Bioinform.* **14**, 315–326 (2013).
71. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–596 (2013).

Acknowledgements

The authors thank Genome Canada, Génome Québec, the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Canada Research Chairs (CRC) Program for funding support.

Author Contributions

J.C. and J.X. prepared the manuscript. J.C., P.L., G.Z., and J.X. contributed to the development of MicrobiomeAnalyst. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41596-019-0264-1.

Correspondence and requests for materials should be addressed to J.X.

Peer review information *Nature Protocols* thanks Tiffany Weir and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

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

Received: 17 June 2019; Accepted: 29 October 2019;
Published online: 15 January 2020

Related links
Key references using this protocol
Khan, N. et al. *Mucosal Immunol.* **12**, 772–783 (2019): https://doi.org/10.1038/s41385-019-0147-3
Stinson, L. F., Boyce, M. C., Payne, M. S. & Keelan, J. A. *Front. Microbiol.* **10**, 1124 (2019): https://doi.org/10.3389/fmicb.2019.01124
Amrane, S. et al. *Sci. Rep.* **9**, 12807 (2019): https://doi.org/10.1038/s41598-019-49189-8

nature research

Corresponding author(s): Jianguo Xia

Last updated by author(s): Jun 18, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☒	☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	MicrobiomeAnalyst and MicrobiomeAnalystR
Data analysis	MicrobiomeAnalyst

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All example datasets used in the protocol are available from the MicrobiomeAnalyst "Tutorials" page.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	NA
Data exclusions	NA
Replication	NA
Randomization	NA
Blinding	NA

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging