Microbiome

SOFTWARE ARTICLE Open Access

animalcules: interactive microbiome analytics and visualization in R

Yue Zhao[1,2†], Anthony Federico[1,2†], Tyler Faits[1,2], Solaiappan Manimaran[1,3], Daniel Segrè[2,4], Stefano Monti[1,2] and W. Evan Johnson[1,2,3*]

Abstract

Background: Microbial communities that live in and on the human body play a vital role in health and disease. Recent advances in sequencing technologies have enabled the study of microbial communities at unprecedented resolution. However, these advances in data generation have presented novel challenges to researchers attempting to analyze and visualize these data.

Results: To address some of these challenges, we have developed *animalcules*, an easy-to-use interactive microbiome analysis toolkit for 16S rRNA sequencing data, shotgun DNA metagenomics data, and RNA-based metatranscriptomics profiling data. This toolkit combines novel and existing analytics, visualization methods, and machine learning models. For example, the toolkit features traditional microbiome analyses such as alpha/beta diversity and differential abundance analysis, combined with new methods for biomarker identification are. In addition, *animalcules* provides interactive and dynamic figures that enable users to understand their data and discover new insights. *animalcules* can be used as a standalone command-line R package or users can explore their data with the accompanying interactive R Shiny interface.

Conclusions: We present *animalcules*, an R package for interactive microbiome analysis through either an interactive interface facilitated by R Shiny or various command-line functions. It is the first microbiome analysis toolkit that supports the analysis of all 16S rRNA, DNA-based shotgun metagenomics, and RNA-sequencing based metatranscriptomics datasets. *animalcules* can be freely downloaded from GitHub at https://github.com/compbiomed/animalcules or installed through Bioconductor at https://www.bioconductor.org/packages/release/bioc/html/animalcules.html.

Keywords: Microbiome analysis, Visualization, Interactive toolkit, Biomarker identification

* Correspondence: wej@bu.edu
†Yue Zhao and Anthony Federico have equal contributions.
[1]Section of Computational Biomedicine, Boston University School of Medicine, Boston, MA, USA
[2]Bioinformatics Program, College of Engineering, Boston University, Boston, MA, USA
Full list of author information is available at the end of the article

Background

The complex role of the gut microbiota in shaping human health and disease has been intensely investigated and explored in recent years, largely due to the availability of culture-independent molecular-based high-throughput sequencing technologies. It is estimated that every human host coexists with an average of 500–1000 different bacterial species [1–3] and research has discovered that the microbiome is associated with host lifestyle and diet [4, 5] as well as many diseases such as obesity, type 2 diabetes [6] and cancer [7]. New sequencing technology brings not only more data and capacity for microbiome research but also new challenges for data analytics and interpretation. Improved tools and methods for microbiome data analytics can enhance our ability to understand the roles of microbes in diverse environments, particularly understanding how they interact with each other as well as their human hosts.

Current microbiome analysis typically consists of two important components: upstream community profiling (e.g., what is the abundance of all microbes in each sample?) and downstream high-level analysis (e.g., alpha/beta diversity analysis, differential abundance analysis) [8]. In recent years, evolving data analytics, visualization, and machine learning methods have been gradually applied to the development of many software tools and web servers for microbiome data analysis covering these two components [9–13]. However, new techniques and sequencing technologies have steepened the learning curve for scientific researchers applying new methods for microbiome data analysis and interpretation [14]. Furthermore, existing tools are mostly dedicated to one aspect of analysis and/or are restricted to analyzing one type of microbiome data. For example, while there are many tools and workflows for analyzing 16S rRNA data, there are no existing tools and pipelines tailored for comprehensively addressing the analytical needs of RNA-based metatranscriptomics.

Table 1 gives a summary of the functions of these tools with respect to the analysis needs of microbiome data. For marker gene-based data such as 16S rRNA, QIIME II [15] and Mothur [16] provide a user interface and a plethora of analytic and visualization tools, but do not provide support for metagenomic and metatranscriptomic data. Vegan [17] provides a wide variety of functions for metagenomic data visualization, but lacks a user-interface, and tools for host and microbial read alignment, differential expression, etc. BioBakery [18] provides a comprehensive suite of tools for most metagenomic analysis needs for microbial communities, but relies on a small set of markers to identify species, and does not address host or microbial expression. Microbiome helper [19] is a collection of scripts in multiple languages to facilitate interaction and interoperability microbiome and metagenomic tools, but does not provide interactive visualizations or a graphical user interface. The microbiome package in R [20] provides command-line workflows for a wide variety of the metagenomic data analysis tasks. Phyloseq [21] has a Shiny interface with tools for annotation, visualization, and diversity analysis, but does not provide abundance analysis, and is no longer actively maintained by its developers. Metavizr [22] provides an interface and suite of functions for specific metagenomic visualizations. None of these methods are comprehensive or specifically address the needs for multiple types of 16S rRNA, metagenomic or metatranscriptomic data. Therefore, there are no existing toolkits that contain a complete workflow for microbiome data analysis and interpretation (with or without a graphical user interface).

Here, we present *animalcules*, an interactive analysis and visualization toolkit for microbiome data. *animalcules* supports the importing of microbiome profiles in multiple formats such as a species count table, an organizational taxonomic unit (OTU) or amplicon sequence variants (ASV) counts table, or Biological

Table 1 Comparison of *animalcules* and other popular microbiome analysis tools

	Biobakery	Vegan	Mothur	Microbiome	Metavizr	Microbiome helper	Qiime2	Phyloseq	animalcules
Filtering and data summary	✓		✓	✓		✓	✓	✓	✓
Interactive visualization					✓		✓		✓
Dimension reduction			✓	✓	✓	✓	✓	✓	✓
Differential abundance analysis	✓		✓	✓	✓	✓	✓	✓	✓
Diversity analysis		✓	✓	✓	✓	✓	✓	✓	✓
Support for 16S rRNA data	✓	✓	✓	✓	✓	✓	✓	✓	✓
Support for total RNA-seq data	✓	✓							✓
Biomarker identification							✓		✓
Interface and command-line						✓	✓		✓
Language/platform	Python	R	R/Web	R	R	R/Python/Perl	Python	R	R

Observation Matrix (BIOM) format [23]. These formats could be generated from common microbiome data sources and analytical tools including 16S rRNA, metagenomics, and metatranscriptomic data. Once data is uploaded, *animalcules* provides a useful data summary and filtering function where users can view and filter their dataset using sample metadata, microbial prevalence, or relative abundance. Filtering the data in this way can significantly reduce the time spent performing preprocessing and downstream analysis tasks. For data visualizations, such as relative abundance bar charts and 3D dimension reduction plots (PCA/PCoA/tSNE/UMAP), *animalcules* supports interactive operations where users can check the sample/microbe information on each data point and adjust the figure format as needed, which is helpful for recognizing elements or data patterns when the sample size or number of microbes is large. Aside from common diversity analysis, differential abundance analysis, and dimension reduction, *animalcules* supports biomarker identification by training a logistic regression or random forest model with cross-validated biomarker performance evaluation. *animalcules* provides a graphical user interface (GUI) through R/Shiny, which can be used even by users without prior programming knowledge, while experienced programmers can choose the command-line based R package or a combination of both.

Implementation

Data structures and software design

All data handling tasks and functions in *animalcules* are based upon and work with the MultiAssayExperiment (MAE) data structure [24]. The MAE class is a standard data structure for multi-omics experiments with efficient data retrieval and manipulation methods that support the linkage of samples across multiple assays. The MAE object has three key components: colData (contains subject or cell line level metadata), ExperimentList (stores data for one or more assays), and sampleMap (relates experiments and samples). In *animalcules*, three tables (sample metadata table, microbe count table, and taxonomy table) as well as the mapping relationship between them are stored in the MAE class. It ensures correct alignment of assays and subjects, and provides coordinated subsetting of samples and features. Additionally, it is easy to convert to or from a MAE object from the SummarizedExperiment class, which has been applied in many Bioconductor packages, enabling smooth interaction between other tools [24]. One important advantage of applying the MAE class in the microbiome research field is its extensible design supporting many multi-omics layers of data. Multi-omics is becoming a trend in the field, e.g., studying host-microbe interactions by combining host gene expression data and microbial abundance data. *animalcules* is the first software tool for microbiome analysis to integrate the MAE object and takes advantage of its unique properties by allowing the user to store microbial data, host transcriptomics, metabolomics, as well as taxonomy information within the same object (currently *animalcules* only supports microbiome analysis, but the MAE structure enables future development that can address these data types. Additionally, the MAE enables integration with other tools that do manage these data types, e.g. host transcriptomics). The MAE object can also store processed versions of various assays (e.g., dimension-reduced data) which allows for efficient manipulation and analysis downstream. This approach advances standard microbiome analysis and data sharing by efficiently integrating the various multi-omics datasets required.

Lastly, because all of the data is integrated within a single R object, users can serialize the data to a single file which can be used for further analysis or share with other researchers. For example, after processing and analyzing their data through the Shiny application, users can export their datasets in the form of a serialized MAE object file, which can be later uploaded to Shiny or imported in R for further exploration through the *animalcules* command line functions or other methods. Integrating the MAE object brings efficiency, scalability, and reproducibility to microbiome analysis through *animalcules*.

Installation and usage

animalcules requires $R \geq 4.0.0$ and can be installed through Github or Bioconductor. After loading the *animalcules* library in R, users can choose between launching the R Shiny GUI (via the run_animalcules() function), or using the available command-line functions directly. In the GUI, users can choose from the following tabs: Upload (select an example dataset, upload a new dataset, or load a previously uploaded dataset), Summary and Filter (understand the data distribution and filter the data by microbial features or sample phenotypes), Abundance (relative abundance bar charts, heatmaps, and individual microbes boxplots), Diversity (statistical tests and boxplots for alpha diversity and beta diversity), Dimension Reduction (PCA, PCoA, tSNE, and UMAP), Differential Abundance (microbial differential abundance between sample groups), and Biomarker (identify predictive microbial biomarkers). Common R functions in the package are summarized in Table 2. A detailed tutorial on how to use the command-line version of *animalcules* for microbiome data analysis can be found at https://compbiomed.github.io/animalcules-docs/articles/animalcules.html. In addition, we have released a Docker container that comprises our *animalcules* package and

Table 2 Table of exported functions and their descriptions available through the *animalcules* R package

Data and interface	
run_animalcules()	Initiates a local instance of the *animalcules* Shiny application
Data summary and manipulation	
filter_summary_bar_density()	Visualize sample/microbe data with a bar plot (categorical) or density plot (continuous)
filter_summary_pie_box()	Visualize sample/microbe data with a pie chart (categorical) or box plot (continuous)
filter_categorize()	Convert continuous variables into a various number of factors
counts_to_logcpm()	Covert counts table to a log counts per million table
counts_to_relabu()	Covert counts table to a relative abundances table
upsample_counts()	Up-sample counts table to a higher taxon level
find_taxonomy()	Find taxonomy for unlimited ids
find_taxon_mat()	Find taxonomy information matrix for unlimited ids
mae_pick_samples()	Isolate or discard samples from a multi-assay experiment object
mae_pick_organisms()	Isolate or discard microbes from a multi-assay experiment object
Sample level visualization	
relabu_barplot()	Generate stacked bar plots of sample and group level microbe relative abundances
relabu_boxplot()	Generate box plots comparing organism prevalence across groups of samples
relabu_heatmap()	Generate a sample by microbe heatmap of counts
dimred_pca()	Return a 2D/3D scatter plot for dimensionality reduction through PCA
dimred_pcoa()	Return a 2D/3D scatter plot for dimensionality reduction through PCoA
dimred_umap()	Return a 2D/3D scatter plot for dimensionality reduction through UMAP
dimred_tsne()	Return a 2D/3D scatter plot for dimensionality reduction through t-SNE
Alpha and beta diversity	
diversities()	Return alpha diversity
do_alpha_div_test()	Compute various statistical tests for alpha diversity
alpha_div_boxplot()	Generate box plots comparing alpha diversity across groups of samples
diversity_beta_test()	Compute various statistical tests for beta diversity
diversity_beta_boxplot()	Generate box plots comparing beta diversity across groups of samples
diversity_beta_heatmap()	Generate a heatmap comparing beta diversity across groups of samples
Differential abundance analysis	
differential_abundance()	Performs differential abundance analysis across groups of samples
Biomarker discovery	
find_biomarker()	Identifies microbes as potential biomarkers for groups of samples

all dependencies, which is available at: https://github.com/compbiomed/animalcules.

Data upload and output

animalcules offers multiple options for importing data into the GUI or working with the MAE object for command line analysis. These include simple tab-delimited OTU, ASV, or other count matrices, typically generated by other tools such as QIIME II [15] or PathoScope [25], or using a MAE object available in the user's session or in a file from a previous session of *animalcules.* Regardless of how the data is imported, the assay/OTU data will be available in the "Assay Viewer" section of the Upload tab.

Six of the data importing options are described below:

1. *Count Table or OTU File* (*without taxonomy*): This is the simplest option that enables the upload of an out, ASV, or count table that has genomes/OTUs in the rows and samples in the columns. All functions and tools can be used for filtering, visualization, analysis of the data, except the individual microbiomes or OTUs cannot be aggregated at different levels.
2. *Count Table or OTU File* (*with taxonomy*): This option provides an extension of the previous but allows for associating the OTUs/ASVs with taxonomy information and the aggregation of

microbes at different levels (e.g., species, genus, phylum, etc.). This information can be provided as a separate table, with a row for each OTU in the table. In addition, users can provide NCBI taxonomy IDs or NCBI accession numbers [26] and *animalcules* will automatically generate the taxonomy table using the tools available in the *taxize* R package [27]. The taxonomy table will be stored as a separate assay in the MAE object, but will be linked to the rows of the OTU table through internal functions. The taxonomy table will be available in the “Assay Viewer” section in the Upload tab.

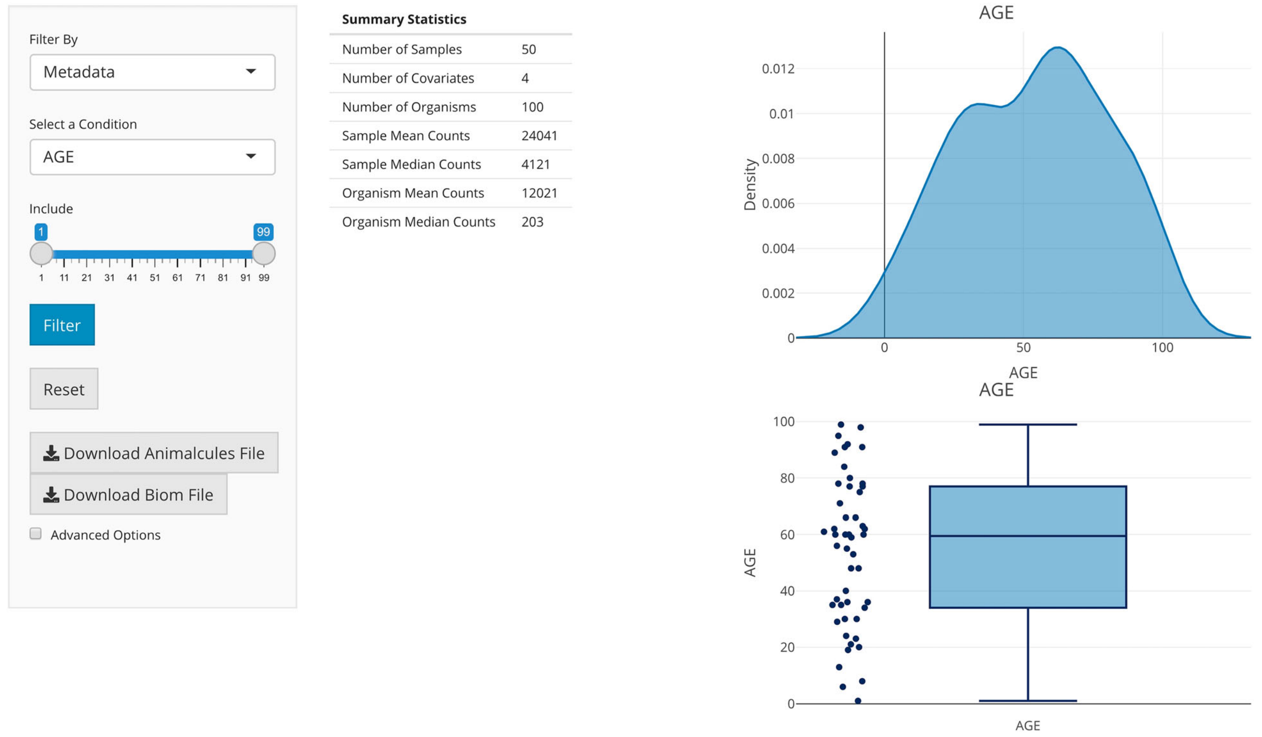
3. *animalcules Object File*: Users can also directly upload a MAE object into the toolkit or workflow. A MAE object could be generated from a previous *animalcules* session (stored as an .rds file), converted from the output of any pre-processing pipeline, or generated from some other source. This option allows for the efficient storage and re-upload of data from a previous session, or enables the interaction between the command-line version and the GUI version of animalcules. For example, users can conduct part of the analysis in the GUI, save the results, and continue their analysis using command-line tools (inside and outside of *animalcules*), and then re-upload the data to the GUI for further analysis or visualization. This feature enables compatibility and interactivity that is not available in other microbiome GUI or command-line toolkits.
4. *Pathoscope Output Files*: *animalcules* enables the direct upload of files generated from the *PathoScope* pipeline [25]. These files are generally single tab-delimited tables for each sample in the dataset, and contain NCBI taxonomy IDs for individual microbes. *animalcules* combines and converts these files into a MAE object, and uses the *taxize* package to generate the taxonomy table.
5. *BIOM and other formats*: The standard BIological Observation Matrix (BIOM) format is a commonly used format for representing samples by observation contingency tables [23]. The BIOM format is commonly used by QIIME II pipeline tools. We used the *biomformat* R package [28] for uploading a BIOM file into *animalcules* as well as outputting a BIOM file from *animalcules*. This enables interactivity between *animalcules* and other microbiome analysis tools such as QIIME II. Furthermore, we have included instructions in our help documentation on how to convert between *animalcules* and Phyloseq objects [21] to enhance interoperability between these tools.
6. *Example Data*: In *animalcules*, we have three pre-defined example datasets, including a simulated dataset, a Tuberculosis 16S rRNA profiling dataset, and an Asthma metatranscriptomic dataset. These example datasets allow users to try all the features and functions in animalcules before users upload their own data, making it easy to learn how to use *animalcules* and understand what analyses they can perform.

Data filtering and summary

The *animalcules* Shiny interface provides summary statistics to help users efficiently and effectively assess data quality and filter low-quality microbes and samples. Users can visualize the total number of reads for each organism through a scatter and density plot and filter organisms based on average read number, relative abundance, or prevalence. Additionally, users can visualize sample covariates through a pie and bar plot for categorical covariates or a scatter and density plot for continuous covariates (Fig. 1). Samples can be filtered based on one or more covariates. Finally, users have the option to discard specific samples and/or organisms. As samples and organisms are removed through any of the filtering methods, summary statistics and plots are automatically refreshed to display any changes that may occur. If changes have been made, users may download the modified data for later use. Visualizations of sample and microbe data before and after filtering are generated with animalcules::filter_summary_bar_density() and animalcules::filter_summary_pie_box() functions. For users who wish to inspect their data before or after filtering, *animalcules* enables users to view and download five types of assays generated including a count table, relative abundance table, logCPM table, taxonomy table, and annotation table. In addition, these tables can also be accessed directly from the MAE object through standard R command line tools.

Data visualization

A typical analysis involves visualization of microbe abundances across samples or groups of samples. *animalcules* implements three common types of visualization plots including stacked bar plots, heatmaps, and box plots. The stacked bar plots, generated with animalcules::relabu_barplot() are used to visualize the relative abundance of microbes at a given taxonomic level in each sample, represented as a single bar (Fig. 2). Bars can be color-labeled by one or more sample attributes and samples can also be aggregated by these attributes via summing microbe abundances within groups. This is an efficient way for researchers to identify sample- or group-level patterns at various taxonomic levels. Users also have the option to sort the bars by sample attributes



Filter By

Metadata

Select a Condition

AGE

Include

1 – 99

Filter

Reset

Download Animalcules File

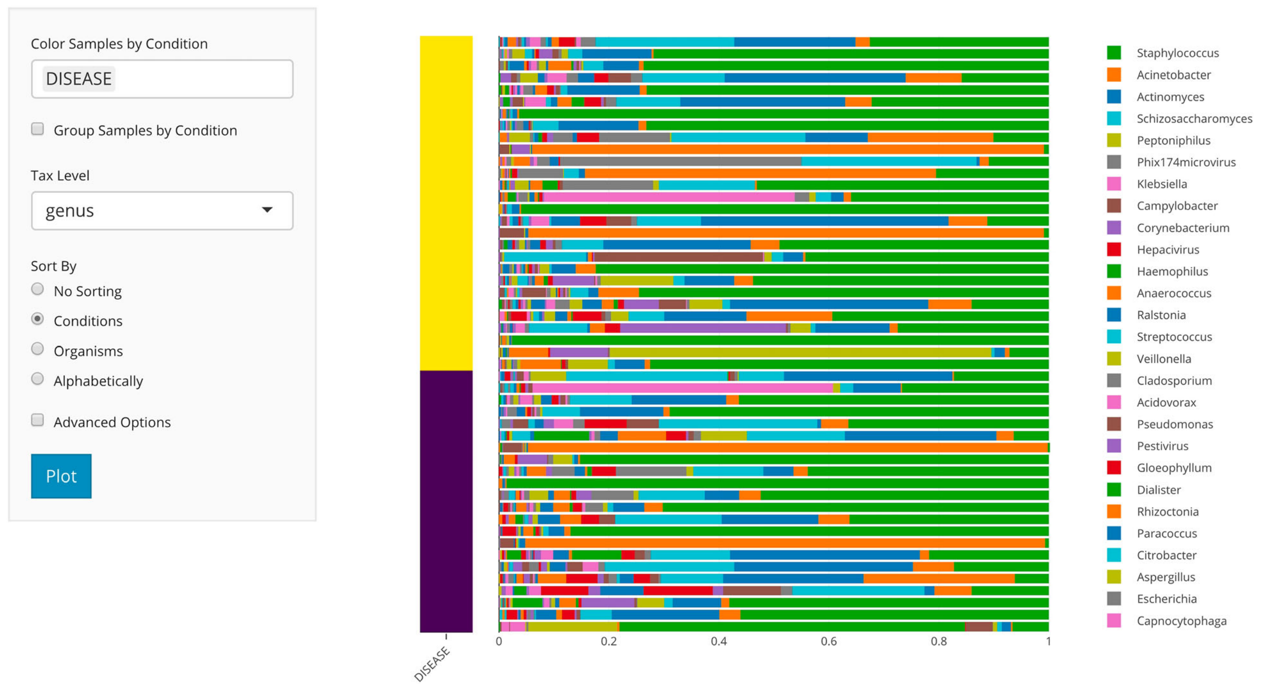
Download Biom File

□ Advanced Options

Summary Statistics

Number of Samples	50
Number of Covariates	4
Number of Organisms	100
Sample Mean Counts	24041
Sample Median Counts	4121
Organism Mean Counts	12021
Organism Median Counts	203

Fig. 1 *animalcules* Data filtering and summary tab. In the right panel, a table of data summary metrics, a scatter/boxplot, and a density plot are displayed for continuous variables. For categorical variables, *animalcules* will automatically identify and show the pie and bar plots instead



Color Samples by Condition

DISEASE

□ Group Samples by Condition

Tax Level

genus

Sort By

○ No Sorting

● Conditions

○ Organisms

○ Alphabetically

□ Advanced Options

Plot

Fig. 2 *animalcules* Abundance tab. In the subtab panel, users can select between a bar plot, heatmap, or box plot. In the bar plot setting, in the left panel, users can select the color by variable, taxonomy level, and sort by option. In the right panel, *animalcules* will show an interactive plot where users can mouse-hover to check the identity of any color bar shown in the plot or by the abundance of one or more organisms. There is also a convenient method for isolating or removing samples. With this tool, users can quickly scan through different combinations of sample attributes and taxon levels for differential abundance in one or more groups, outliers in terms of community profile, as well as sample clusters not represented by known attributes.

Alternatively, users can investigate these questions through the heatmap visualization, which represents a sample-by-organisms matrix that can be visualized at different taxonomic levels. Many of the previously mentioned options are also compatible with the heatmap such as color-labeling samples, sorting matrix rows by attributes or organisms, and isolating or discarding organisms and samples. After identifying potential differentially abundant microbes, users can use the boxplot visualization to directly compare the abundance of one or more organisms between categorical attributes. Organisms can be chosen from a given taxonomic level and abundance can be represented as either counts, logCPM, or relative abundance. This plot can also be generated in the command line using the animalcules::relabu_heatmap() function.

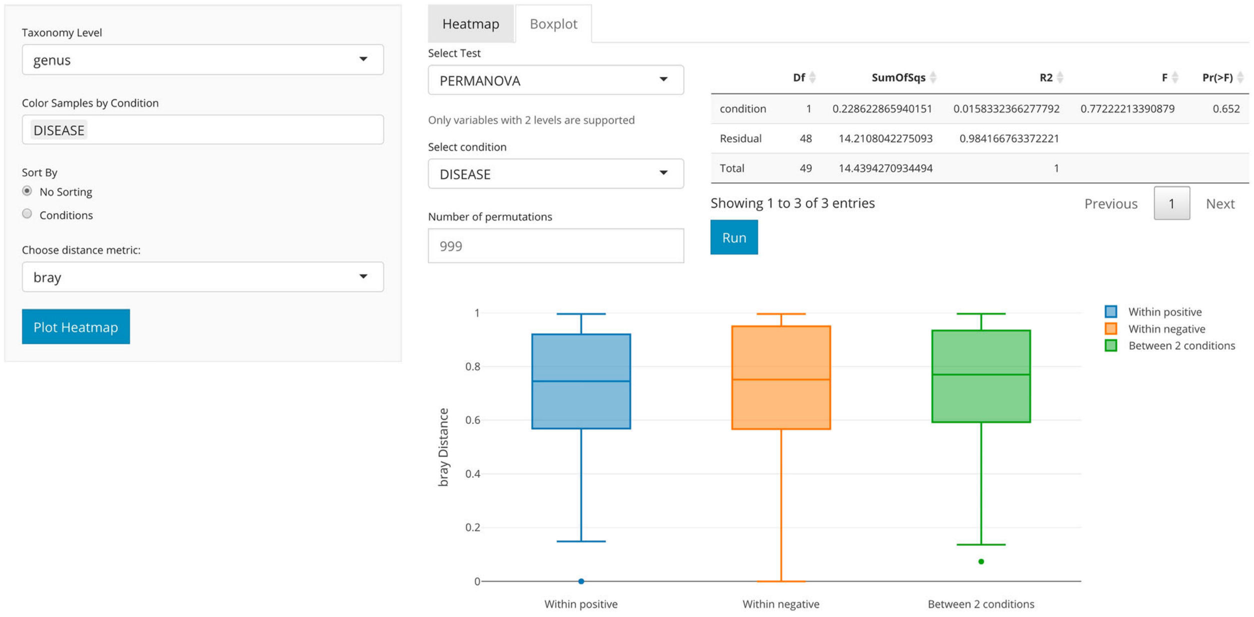
Diversity analysis

Alpha diversity, which describes the richness and evenness of a microbial community, is a vital indicator and measurement in microbiome analysis [29]. *animalcules* provides an interactive box plot comparison of alpha diversity between selected groups of samples. Both taxonomy levels and alpha diversity metrics (e.g., Shannon, Gini Simpson, Inverse Simpson, unique organisms) can be changed and diversity can be calculated at multiple taxonomic levels [30, 31]. Alpha diversity values for each sample can be output into the MAE object or as separate tables or files. Users can also conduct alpha diversity statistical tests including Wilcoxon rank-sum test, *T* test, and Kruskal-Wallis test [32, 33]. The alpha diversity boxplot as well as the statistical tests could be generated in the command line using the animalcules::alpha_div_boxplot() function and animalcules::do_alpha_div_test() function.

On the other hand, one can use distances between each microbial community sample, or so-called beta diversity, as another key metric to consider for each analysis. Users can plot the beta diversity heatmap by selecting different beta diversity dissimilarity metrics including Bray-Curtis [34], Jaccard index [35], or weighted/unweighted unifrac. Users can also conduct beta diversity statistical testing between groups including PERMANOVA [36], Wilcoxon rank-sum test, or Kruskal-Wallis test (Fig. 3). The beta diversity comparison boxplot as well as the statistical tests can be generated in the command line using the animalcules::diversity_beta_boxplot() function and animalcules::diversity_beta_test() function.

Dimension reduction

A crucial step in any data analysis workflow is to visualize and summarize highly variable data in a lower-



Taxonomy Level: genus

Color Samples by Condition: DISEASE

Sort By
● No Sorting
○ Conditions

Choose distance metric: bray

Select Test: PERMANOVA

Only variables with 2 levels are supported

Select condition: DISEASE

Number of permutations: 999

	Df	SumOfSqs	R2	F	Pr(>F)
condition	1	0.228622865940151	0.0158332366277792	0.77222213390879	0.652
Residual	48	14.2108042275093	0.984166763372221		
Total	49	14.4394270934494	1		

Showing 1 to 3 of 3 entries

Fig. 3 *animalcules* Diversity tab. In the subtab panel, the user could select between alpha diversity analysis and beta diversity analysis. Here in the beta diversity analysis, the right panel controls what statistical test to use, which condition to test on, and show statistical test results in a table as well as a boxplot dimensional space (Fig. 4). In *animalcules*, we implement four commonly used dimensionality reduction techniques including principal components analysis (PCA), principal coordinates analysis (PCoA), t-distributed stochastic neighbor Eebedding (t-SNE), and uniform manifold approximation and projection (UMAP) [37–40]. Both PCA and PCoA project samples onto a new set of axes whereby a maximum amount of variation is explained by the first, second, and third axes while t-SNE and UMAP are non-linear methods for mapping data to a lower-dimensional embedding. Dimension reduction values for the dataset can be output into the MAE object or as separate tables or files.

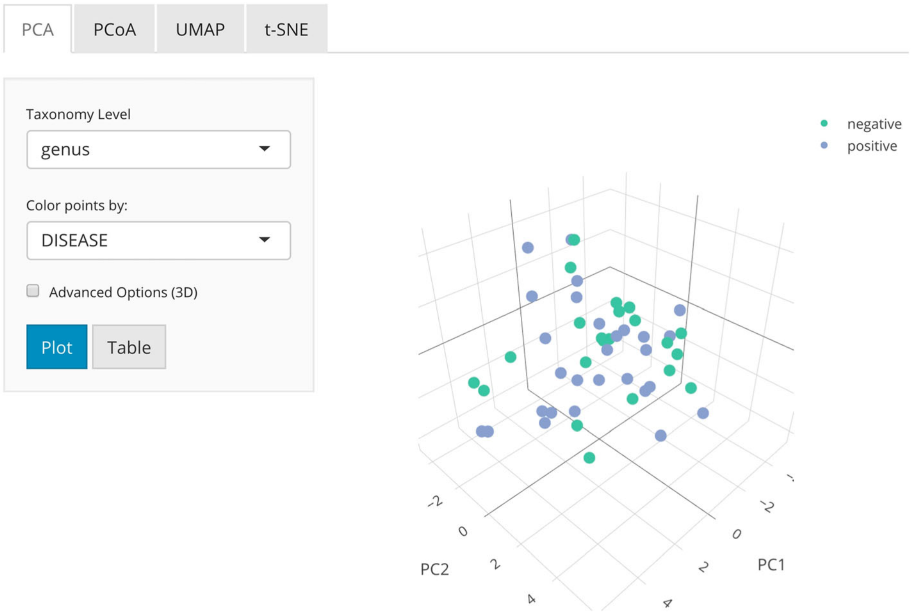
The original data used in each dimensionality reduction method can be either counts, logCPM, or relative abundance, and can be visualized using a 2D or 3D (if two dimensions of explained variance are inadequate) scatter plot. Data points can be colored by continuous sample attributes and shaped by categorical attributes. With multiple dimensionality reduction techniques and methods for data normalization, users can rapidly visualize the global and local structure of their data, identify clustering patterns across one or more conditions, as well as detect sample outliers. Dimensionality reduction can also be carried out in the command line using the animalcules::dimred_pca() function for PCA, animalcules::dimred_pcoa() function for PCoA, animalcules::dimred_tsne() function for t-SNE, and animalcules::dimred_umap() function for UMAP.

Differential abundance analysis

There are many available tools for differential abundance estimation and inference. For example, generalized linear model (GLM)-based methods including DESeq2 [41], edgeR [42], and limma [43] model count based microbiome data or gene expression data by a negative binomial distribution (DESeq2 and edgeR) or using log-counts (per million) and a Gaussian distribution (limma) assumption. Core microbes that have different abundance in different groups could be identified. Here in *animalcules*, we provide a DESeq2-based differential abundance analysis (Fig. 5). With the command-line function animalcules::differential_abundance(), which by default uses the "DESeq2" method. Users can choose the target variable, covariate variable, taxonomy level, minimum count cut-off, and an adjusted *p* value threshold. The analysis report will output not only the adjusted *p* value and log2-fold-change of the microbes but also the percentage, prevalence, and the group size-adjusted fold change. Besides using DESeq2, in *animalcules* we also support differential abundance analysis with limma, which requires users to specify in the command-line function as: animalcules::differential_abundance (method='limma').

Biomarker identification

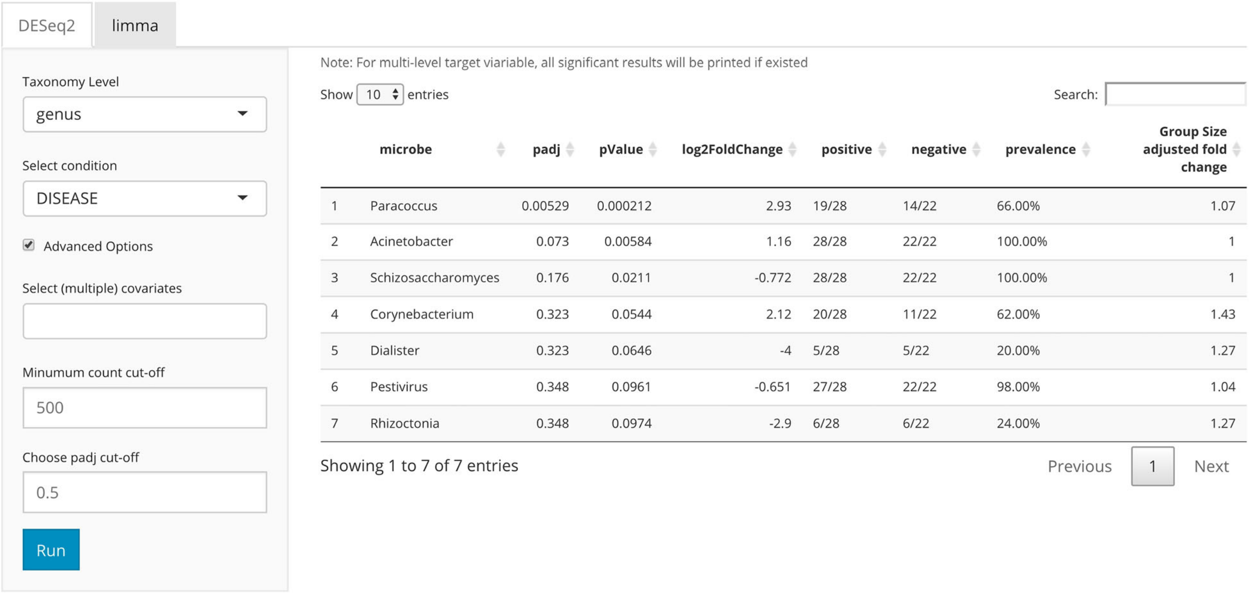
One unique feature of *animalcules* is the biomarker identification module. Users can choose either a logistic regression [44] or a random forest [45] classification model



Taxonomy Level: genus

Color points by: DISEASE

Fig. 4 *animalcules* Dimension Reduction tab. In the subtab panel, the user could select between PCA, PCoA, t-SNE, and UMAP. Here in the PCA subtab, the user could choose the taxonomy level, color by variable, and in advanced options, the user could also specify up to three PCs for visualization, shape by variable, and which data type to use



Taxonomy Level
genus

Select condition
DISEASE

☑ Advanced Options

Select (multiple) covariates

Minumum count cut-off
500

Choose padj cut-off
0.5

Note: For multi-level target viariable, all significant results will be printed if existed

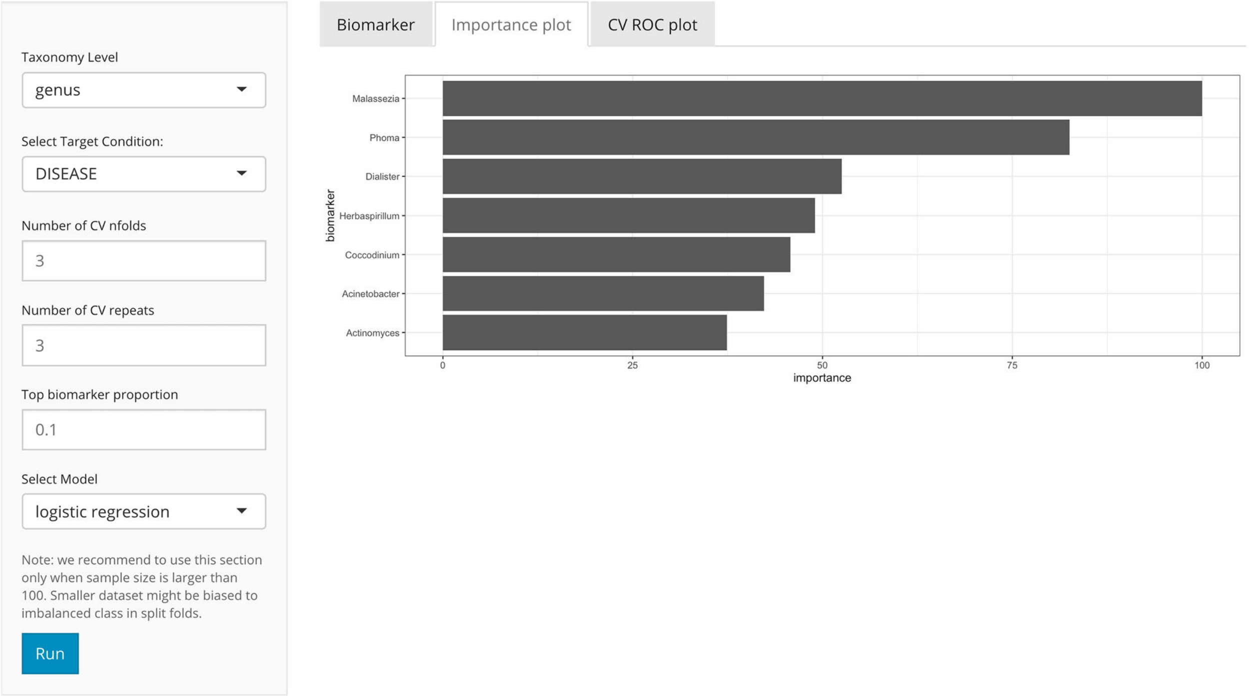
Show 10 entries

Search:

	microbe	padj	pValue	log2FoldChange	positive	negative	prevalence	Group Size adjusted fold change
1	Paracoccus	0.00529	0.000212	2.93	19/28	14/22	66.00%	1.07
2	Acinetobacter	0.073	0.00584	1.16	28/28	22/22	100.00%	1
3	Schizosaccharomyces	0.176	0.0211	-0.772	28/28	22/22	100.00%	1
4	Corynebacterium	0.323	0.0544	2.12	20/28	11/22	62.00%	1.43
5	Dialister	0.323	0.0646	-4	5/28	5/22	20.00%	1.27
6	Pestivirus	0.348	0.0961	-0.651	27/28	22/22	98.00%	1.04
7	Rhizoctonia	0.348	0.0974	-2.9	6/28	6/22	24.00%	1.27

Showing 1 to 7 of 7 entries

Fig. 5 *animalcules* Differential Abundance tab. In the subtab panel, users select between DESeq2 and limma. In the left panel, users specify taxonomy level, target condition, covariate variables, count cut-off, and adjusted p-value threshold. In the right panel, a detailed differential abundance result table is shown



Taxonomy Level
genus

Select Target Condition:
DISEASE

Number of CV nfolds
3

Number of CV repeats
3

Top biomarker proportion
0.1

Select Model
logistic regression

Note: we recommend to use this section only when sample size is larger than 100. Smaller dataset might be biased to imbalanced class in split folds.

Fig. 6 *animalcules* Biomarker tab. In the left panel, users select taxonomy level, and target condition. In the advanced options: number of cross-validations folds, number of cross-validations repeats, biomarker proportion, and classification model. In the right panel, *animalcules* will show the biomarker list, importance plot, and ROC plot to identify a microbe biomarker. The feature importance score for each microbe will be provided (Fig. 6), in addition to AUC values and average cross-validation ROC curves for evaluating biomarker prediction performance. The biomarker identification can also be conducted by the command-line function animalcules::find_biomarker().

Biomarker identification is often a relatively difficult task in microbiome research projects. By default, we've selected the logistic regression model, which has better model interpretability and lower model complexity compared to the random forest approach. However, when logistic regression underfits the dataset, then users are able to easily switch to the random forest method. For parameter selection, most of the hyper-parameters are selected heuristically and automatically by a grid-search based hyperparameter tuning embedded in the N-time K-fold cross-validation, using $K = 3$ as a default. Smaller values for K are beneficial for small to moderate sample sizes, as it will decrease the possibility that any fold has an imbalanced class distribution. When a user has a larger dataset, we recommend a larger K, such as 5 or 10. The method produces the proportion of candidate biomarkers (form the cross-validation) that include a particular microbe as an "importance score" to aid in the selection of microbes for the final biomarker. The user-supplied parameter P (default $P = 0.20$) controls the importance score cutoff for inclusion of a microbe in the final model and evaluation. There is no strict rule or intuition selecting P, but most often it is beneficial to try a few different values of P and compare the performance of the final biomarker predictions in the testing set.

We emphasize a concern that for small datasets (e.g., with less than 100 samples), the identified biomarker can be biased due to model overfitting. Our current strategy to avoid overfitting in larger datasets relies on repeated k-fold cross-validation. However, when sample size is small, it can still be biased to a subset of the samples because one fold of the cross-validation might randomly have an unbalanced class distribution (e.g., 1 positive and 10 negative samples). Therefore, we advise users to use extreme caution when using the biomarker identification module when they have a smaller dataset.

Results

To illustrate the utility of *animalcules*, we include two example analyses using the pre-loaded datasets packaged within *animalcules*; the first being an asthma metatranscriptomic dataset, and the second a TB 16S rRNA dataset. For brevity, we do not explore all *animalcules* functions in each analysis, but focus on and expand the relevant analyses for the scientific questions for each example. Both analyses could be reproduced within the *animalcules* Rshiny app by using the corresponding example datasets.

Example 1: asthma nasal swabs metatranscriptomic dataset

The asthma metagenomic shotgun RNA sequencing dataset was generated from participants of the AsthMap (Asthma Severity Modifying Polymorphisms) project and originally reported in a research article characterizing asthma-associated microbial communities [46]. It contains 14 total samples of nasal epithelial cells collected from 8 children and adolescents with asthma and 6 healthy controls. The goal of this study was to further understand the relationship between the microbiome and host inflammatory processes in asthmatic children.

To characterize the relationship between microbial communities and asthma, species-level abundances were visualized by plotting the group-wise relative abundance of microbes across asthma and control subjects. This plot can be generated with the animalcules::relabu_barplot() function as well as under the *Abundance* tab of the Shiny application. It is clear that *Moraxella catarrhalis* is overrepresented in asthmatics versus controls, which was a major discovery in the original publication. This microbe—which is known to cause infections in the respiratory system—could serve as a biomarker for early disease detection, severity of disease, or potential for exacerbation. In addition, other dysbiosis to the airway microbiome included differences in other genera such as *Corynebacterium aurimucosum*, which is underrepresented in asthmatics versus controls (Fig. 7).

To further investigate the overrepresentation and underrepresentation of *M. catarrhalis* and *C. aurimucosum* respectively in asthmatics versus controls, we use boxplots, generated with the animalcules::boxplot() function, to visualize the relative abundance in each group and to get a better sense of the mean and variance of the distribution across samples. These plots confirm the previous results by showing a drastic difference in abundance (Fig. 8). Furthermore, we employed DESeq2 to conduct a differential abundance analysis of microbe species for asthmatics versus controls. This analysis shows that *M. catarrhalis* is significantly (q = 1.78e-3) overrepresented ($\text{Log}_2\text{FC} = 5.9$) in asthmatics. It also shows that *C. aurimucosum* is overrepresented ($\text{Log}_2\text{FC} = 2.66$) in controls, however not at a statistically significant level ($q = 0.236$). This table was generated with the animalcules::differential_abundance() feature.

Through the *animalcules* interface, we were able to rapidly visualize sample- and group-level microbial communities between asthmatic and control samples and test for over- and underrepresented organisms in asthmatics, identifying *M. catarrhalis* and *C. aurimucosum* respectively.

Example 2: tuberculosis 16S rRNA profiling dataset

This 16S rRNA TB dataset comes from a pilot TB study containing 12 subjects, 30 respiratory tract

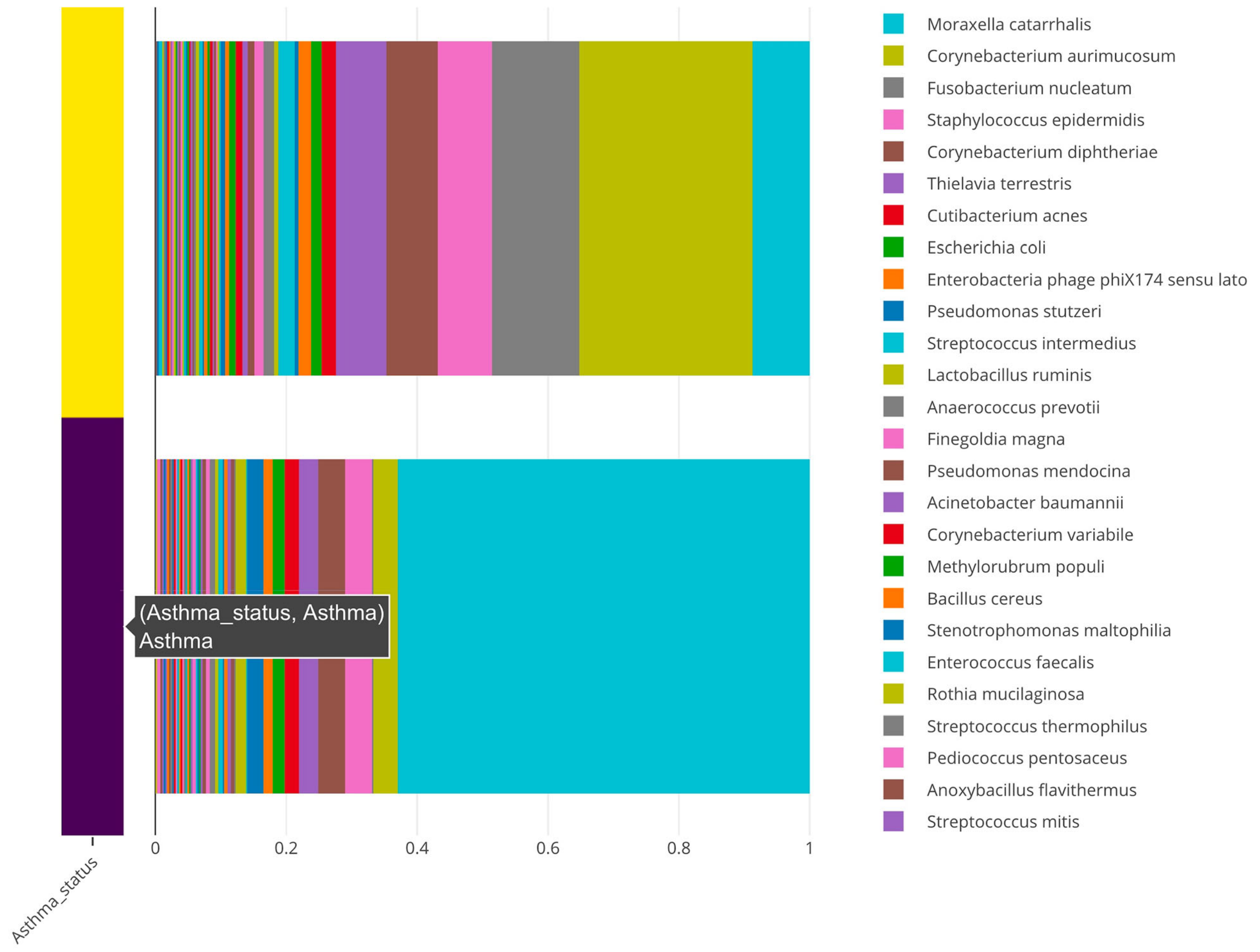


Fig. 7 Relative abundance of microbial species bat plot. A stacked bar plot representing the group-wise relative abundance of microbial species in asthmatics (purple) and healthy controls (yellow)

samples and 417 species of microbe [47] Among the 12 subjects, there are 6 patients with pulmonary tuberculosis and 6 healthy control individuals. Sample tissue type includes sputum, oropharynx, and nasal respiratory tract. The goal of this study is to learn the microbial community differences in the respiratory tract between healthy and TB patients, and to evaluate the sample/tissue types that were most effective for exploring differences between the microbiome of TB samples vs. controls.

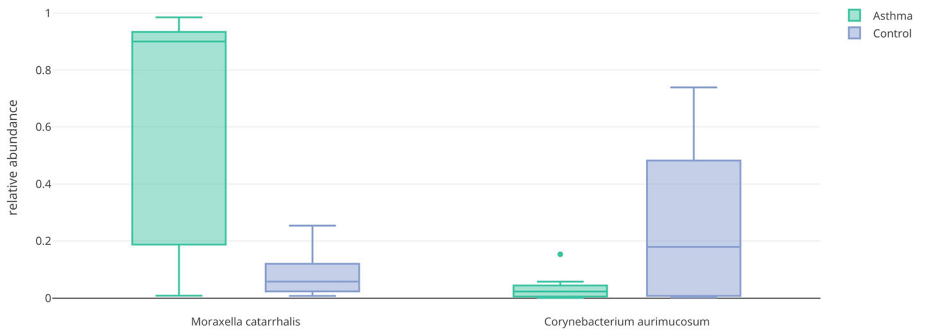


Fig. 8 Relative abundance boxplot for differentially abundant species. *Left*. A boxplot of relative abundance of *M. catarrhalis* in asthmatics (green) and healthy controls (blue). *Right*. A boxplot of relative abundance of *C. aurimucosum* in asthmatics (green) and healthy controls (blue)

We first conducted an overall assessment of the data, focusing on how the microbial taxonomy affects sample variables such as disease status. We used the barplot function in *animalcules* to visualize the taxonomic profile for each sample, colored by any annotation variable (here we used disease information, where dark blue represents control and yellow represents TB samples). In Fig. 9, we display the genus and phylum level abundances.

From the taxonomy barplot, we find different patterns that exist in TB vs. control samples. At the genus level (Fig. 9a), *Streptococcus* appears to have a higher relative abundance in TB samples compared to the control samples. In the phylum level (Fig. 9b), we found *Firmicutes* to be more abundant in TB samples. Both figures were generated using command-line function animalcules::rebalu_barplot().

To obtain a quantitative understanding of the ecological diversity difference between TB and control samples, we compared the alpha and beta diversity of our samples. For alpha diversity, we compared the Shannon index in TB vs. control samples (see Fig. 10a). *animalcules* automatically conducted a non-parametric Wilcoxon rank-sum test and a parametric Welch two-sample *T* test on these diversity measures. Here, the Wilcoxon rank-sum test gives a *p* value of 0.0060, while Welch two-sample *T* test gives a *p* value 0.0077, thus showing a significant difference in diversity between TB and control groups. From the boxplot, we observe that the alpha diversity is higher in the control group. The alpha diversity boxplot was generated by animalcules::alpha_div_boxplot(), and the statistical test was generated by animalcules::do_alpha_div_test().

As for comparing beta diversity, we plotted the Bray-Curtis distance to compare: within the TB group, within the control group, and between the TB/control (Fig. 10b). The average distance between the two groups is higher than two separate within-group distances, meaning both TB samples and control samples are more similar to themselves. Furthermore, we conducted a PERM ANOVA test between the two groups, and it shows a significant difference with a *p* value of 0.003. The beta diversity comparison boxplot was generated by animalcules::diversity_beta_boxplot(), and the PERMANOVA test was generated by animalcules::diversity_beta_test().

After exploring this TB dataset in terms of relative abundance and diversity analysis, we were certain that there is a significant difference between TB and control groups in the microbiome. Here, with the biomarker function in *animalcules*, we were able to build a microbial biomarker that could help us predict TB status. Using a logistic regression model, 3-fold cross-validation (CV), the number of CV repeats as 3, and top biomarker proportion as 0.05, we identified an 8-genus biomarker for TB classification. Then we tested the biomarker performance by using only the 8-genus biomarker for cross-validation, and the prediction performance ROC is displayed in (Fig. 11**)**. We used animalcules::find_biomarker() to identify the biomarker, plot the feature importance score barplot and the ROC curve. Here, we have a very high AUC = 0.913, thus providing evidence that the microbiome could serve as a biomarker for TB prediction, and our biomarker has a differentiating power between TB vs. healthy controls. This result suggests that further evaluation of microbial biomarkers for TB is warranted. Previously, people have been using transcriptomic biomarkers for TB diagnosis [48], our

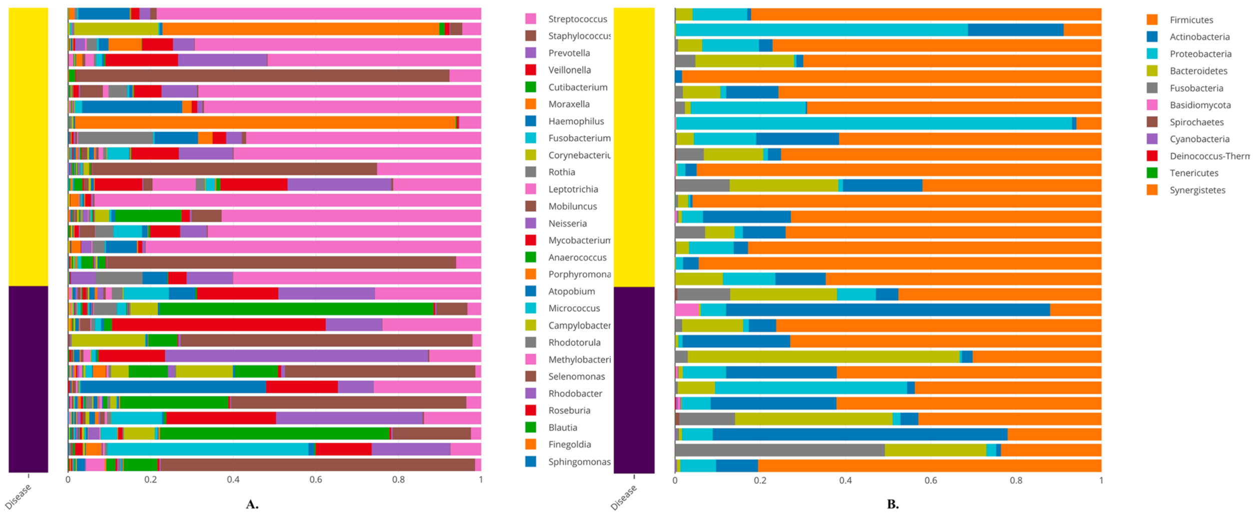


Fig. 9 Sample-wise relative abundance bar plot. A stacked bar plot representing the sample-wise relative abundance of microbial species in TB (yellow) and healthy controls (blue). Panel **a** is the genus level and panel **b** is the phylum level

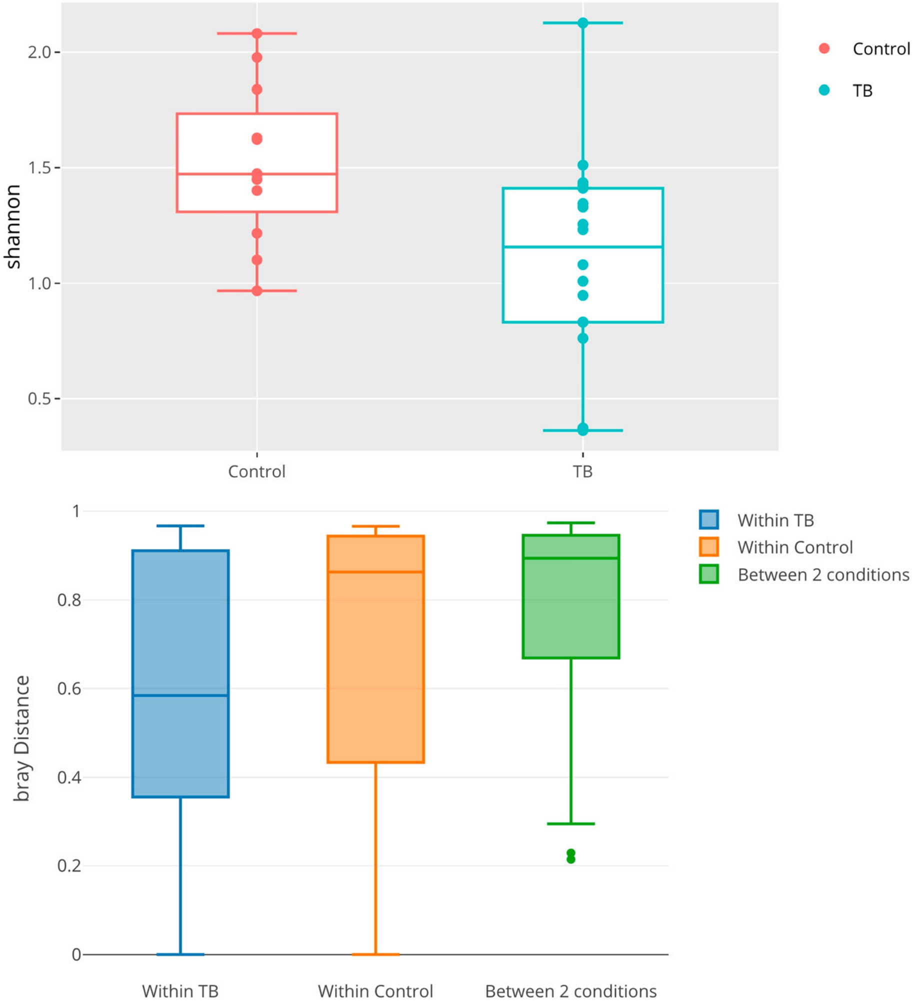


Fig. 10 TB example dataset diversity analysis. **a** Alpha diversity boxplot between control (red) group and TB (blue) group. **b** Beta diversity boxplot within the TB (blue), within the control(orange), and between TB/control group (green)

new finding of using microbes as diagnosis biomarker can lead to a potential total RNA-seq directed TB disease biomarker that involves both host transcriptome gene expression as well as microbial abundance, which has the potential of higher accuracy for TB diagnosis because it considers both host and microbial side, or even the host-microbe interaction in TB.

To summarize, with the help of *animalcules*, we explored and compared the microbial community difference between TB and control samples. Our analysis shows that the microbial community structure in the control group is more diverse and evenly distributed compared to the one in the TB group. Also, the TB group as well as control group each has a specific microbial composition that is shared within the group. Finally, we identified a subset of microbes that indicate its differentiating power between TB vs. control samples, which can be used as a new TB disease biomarker.

Discussion

A fundamental characteristic of *animalcules* is its seamless interaction with the user through dynamic visualization tools. This design logic is rooted in the fact that researchers in microbiology must analyze their data at multiple levels (taxonomy) and multiple scales (normalization), thus data visualization and analysis become complicated without an organized analysis framework and workflow. *animalcules* solves this problem by providing a platform for interactively exploring large datasets, making it easier for users to identify patterns inherent in the dataset through appropriate analysis methods. Key analysis methods allow users to investigate differences in grouped relative abundance patterns between multiple sample groups in the phylum level, check the top abundant species in one specific sample group, or to check the individual sample-wise microbiome composition at different taxon levels. Patterns identified can be further tested

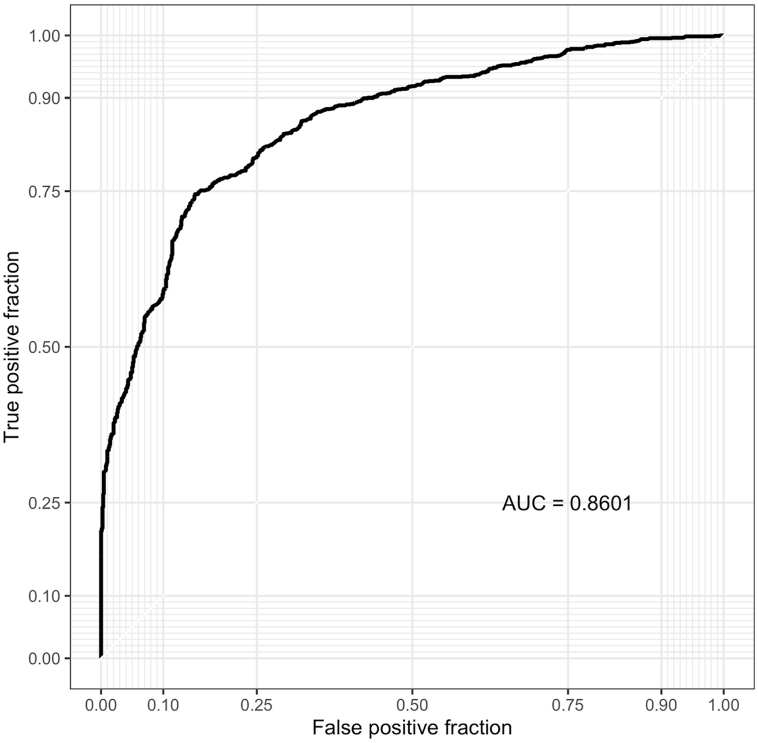


Fig. 11 Biomarker ROC curve. ROC shows AUC and cross-validation prediction performance of the identified biomarker

through alpha/beta diversity statistical tests, differential abundance analysis, as well as biomarker identification.

Furthermore, *animalcules* utilizes the MAE object, an efficient data structure for multi-omic sequencing data, which could be extended in the future to incorporate host sequencing assays, and enable compact methods for analyzing host-microbe interactions. The flexibility of *animalcules* makes it also a helpful tool for future integration across different data sources and downstream analyses: in addition to helping analyze 16S rRNA and shotgun sequencing microbiome data, *animalcules* could be used to visualize the outcomes of simulated microbiome studies [49] and of microcosm experiments involving synthetic microbial consortia [50]. The taxonomic abundance information processed through *animalcules* could further be used for the systematic estimation of microbial co-occurrence networks [51], expanding the possibilities for informative visualization of complex datasets, and the comparison with mechanistic models of microbe-microbe interdependencies [52].

On the reproducibility front, one unfortunate weakness in R/Shiny at the time of this publication is the lack of a systematic structure for generating histories or tracking analyses. More specifically, because each minute change in the interface (e.g., filtering numerical scales) triggers an action in the server-side of the application, logging the R History may become too large too quickly to be useful. Rather, for the sake of reproducibility, we recommend using GUI for preliminary and exploratory analyses, and then follow up with command line functions and code within R Markdown documents for creating reproducibility reports that can be shared.

Conclusion

In this report, we present *animalcules*, an open-source R package and Shiny application dedicated to microbiome analysis for both 16S rRNA and shotgun sequencing (metagenomics and metatranscriptomics) data. We incorporate leading and novel methods in an efficient framework for researchers to characterize and understand the microbial community structure in their data, leading to valuable insights into the connection between the microbial community and phenotypes of interest.

Availability and requirements

Project name: *animalcules*

Project home page: https://github.com/compbiomed/animalcules

Official/formal distribution: https://www.bioconductor.org/packages/release/bioc/html/animalcules.html

Help documentation: https://compbiomed.github.io/animalcules-docs/

Operating system(s): Linux, OS X, Windows

Programming language: R

License: GNU GPLv3

Acknowledgements
This research was supported by grants from the NIH R01GM127430 and U01CA220413. We also thank the Evans Center for Interdisciplinary Biomedical Research at the Boston University School of Medicine for their generous support of this project through the Affinity Research Collaborative on Systems Biology Approaches to Microbiome Research. The authors thank Lucas Schiffer (Boston University) as well as other members of the Johnson lab for helpful comments and advice.

Authors' contributions
YZ, AF, and SM contributed to the software development. YZ, AF, TF, SM, and WEJ contributed to the writing and review of the manuscript before submission for publication. All authors read and approved the final manuscript.

Availability of data and materials
Animalcules is freely available on GitHub at https://github.com/compbiomed/animalcules or Bioconductor at https://bioconductor.org/packages/release/bioc/html/animalcules.html and is accompanied by comprehensive documentation and tutorials at https://compbiomed.github.io/animalcules-docs/.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Section of Computational Biomedicine, Boston University School of Medicine, Boston, MA, USA. [2]Bioinformatics Program, College of Engineering, Boston University, Boston, MA, USA. [3]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA. [4]Departments of Biology, Biomedical Engineering, and Physics, Boston University, Boston, MA, USA.

Received: 20 May 2020 Accepted: 2 February 2021
Published online: 28 March 2021

References

1. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. Nat Med. 2018;24(4):392–400. https://doi.org/10.1038/nm.4517.
2. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. Nature. 2007;449:804–10.
3. Lloyd-Price J, Abu-Ali G, Huttenhower C. The healthy human microbiome. Genome Med. 2016;8:1–11.
4. David LA, Materna AC, Friedman J, Campos-Baptista MI, Blackburn MC, Perrotta A, et al. Host lifestyle affects human microbiota on daily timescales. Genome Biol. 2014;15:R89.
5. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature. 2014;505:559–63.
6. Hartstra AV, Bouter KEC, Bäckhed F, Nieuwdorp M. Insights into the role of the microbiome in obesity and type 2 diabetes. Diabetes Care. 2015;38:159–65.
7. Schwabe RF, Jobin C. The microbiome and cancer. Nat Rev Cancer. 2013;13:800–12.
8. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best practices for analysing microbiomes. Nat Rev Microbiol. 2018;16:410–22.
9. Saulnier DM, Riehle K, Mistretta TA, Diaz MA, Mandal D, Raza S, et al. Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome. Gastroenterology. 2011;141:1782–91.
10. Qu K, Guo F, Liu X, Lin Y, Zou Q. Application of Machine Learning in Microbiology. Front Microbiol. 2019;10:827. https://doi.org/10.3389/fmicb.2019.00827.
11. Zhou YH, Gallins P. A review and tutorial of machine learning methods for microbiome host trait prediction. Front Genet. 2019;10:579.
12. Arango-Argoty G, Garner E, Pruden A, Heath LS, Vikesland P, Zhang L. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome. 2018;6(1):23. https://doi.org/10.1186/s40168-018-0401-z.
13. Reiman D, Metwally A, Yang Dai. Using convolutional neural networks to explore the microbiome. Annu Int Conf IEEE Eng Med Biol Soc. 2017;2017:4269–72. https://doi.org/10.1109/EMBC.2017.8037799.
14. Allaband C, McDonald D, Vázquez-Baeza Y, Minich JJ, Tripathi A, Brenner DA, et al. Microbiome 101: studying, analyzing, and interpreting gut microbiome data for clinicians. Clin Gastroenterol Hepatol. 2019;17:218–30.
15. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019;37:852–7.
16. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75:7537–41.
17. Dixon P. VEGAN, a package of R functions for community ecology. J Veg Sci. 2003;14:927–30.
18. McIver LJ, Abu-Ali G, Franzosa EA, Schwager R, Morgan XC, Waldron L, Segata N, Huttenhower C. bioBakery: a meta'omic analysis environment. Bioinformatics. 2018;34(7):1235–7. https://doi.org/10.1093/bioinformatics/btx754.
19. Comeau AM, Douglas GM, Langille MGI. Microbiome helper: a custom and streamlined workflow for microbiome research. mSystems. 2017;2:e00127–16. https://doi.org/10.1128/mSystems.00127-16.
20. Leo Lahti et al. microbiome R package. http://microbiome.github.io.
21. Mcmurdie PJ, Holmes S. phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One. 2013;8:e61217 journals.plos.org.
22. Bravo H, Chelaru F, Wagner J, Kancherla J, Paulson J. metavizr: R Interface to the metaviz web app for interactive metagenomics data analysis and visualization; 2017.
23. McDonald D, Clemente JC, Kuczynski J, Rideout JR, Stombaugh J, Wendel D, Wilke A, Huse S, Hufnagle J, Meyer F, Knight R, Caporaso JG. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. Gigascience. 2012;1(1):7. https://doi.org/10.1186/2047-217X-1-7.
24. Ramos M, Schiffer L, Re A, Azhar R, Basunia A, Rodriguez C, et al. Software for the integration of multiomics experiments in bioconductor. Cancer Res. 2017;77:e39–42.
25. Hong C, Manimaran S, Shen Y, Perez-Rogers JF, Byrd AL, Castro-Nallar E, et al. PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. Microbiome. 2014;2:33. https://doi.org/10.1186/2049-2618-2-33.
26. Federhen S. The NCBI Taxonomy database. Nucleic Acids Res. 2012; 40(Database issue):D136-43. https://doi.org/10.1093/nar/gkr1178.
27. Chamberlain SA, Szöcs E. taxize: taxonomic search and retrieval in R. F1000Res. 2013;2:191. https://doi.org/10.12688/f1000research.2-191.v2.
28. McMurdie PJ, Paulson JN. biomformat: An interface package for the BIOM file format. 2020. https://github.com/joey711/biomformat/.
29. Whittaker RH. Evolution and measurement of species diversity. Taxon. 1972; 21(2/3):213. https://doi.org/10.2307/1218190.
30. Spellerberg IF, Fedor PJ. A tribute to Claude-Shannon (1916-2001) and a plea for more rigorous use of species richness, species diversity and the "Shannon-Wiener" Index. Glob Ecol Biogeogr. 2003;12:177–9.
31. Jost L. *Entropy and diversity*. Oikos. 2006;113:363–75.
32. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Stat Assoc. 1952;47:583–621.
33. Mann HB, Whitney DR. On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other. Annals of Mathematical Statistics. 1947;18:50–60. https://doi.org/10.1214/aoms/1177730491.
34. Bray JR, Curtis JT. An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecol Monogr. 1957;27:325–49.
35. Jaccard P. The distribution of the flora in the alpine zone. New Phytol. 1912; 11:37–50.
36. Anderson MJ. Permutational Multivariate Analysis of Variance (PERMANOVA). In: Wiley StatsRef: Statistics Reference Online. Wiley; 2017. p. 1–15. https://onlinelibrary.wiley.com/doi/full/10.1002/9781118445112.stat07841.

37. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. London, Edinburgh, Dublin Philos Mag J Sci. 1901;2:559–72.
38. Borg I, Groenen P. Modern multidimensional scaling: Theory and applications; 2005.
39. van der Maaten L, Hinton G, Van Der Maaten L, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res. 2008;9:2579–605.
40. McInnes L, Healy J, Melville J. UMAP: Uniform manifold approximation and projection for dimension reduction. 2018. http://arxiv.org/abs/1802.03426. Accessed 21 Sept 2019.
41. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
42. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40. https://doi.org/10.1093/bioinformatics/btp616.
43. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3:Article3. https://doi.org/10.2202/1544-6115.1027.
44. Jr DH, Lemeshow S, Sturdivant R. Applied logistic regression. 2013.
45. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
46. Castro-Nallar E, Bendall ML, Pérez-Losada M, Sabuncyan S, Severance EG, Dickerson FB, et al. Composition, taxonomy and functional diversity of the oropharynx microbiome in individuals with schizophrenia and controls. PeerJ. 2015;3:e1140.
47. Botero LE, Delgado-Serrano L, Cepeda ML, Bustos JR, Anzola JM, Del Portillo P, et al. Respiratory tract clinical sample selection for microbiota analysis in patients with pulmonary tuberculosis. Microbiome. 2014;2:29. https://doi.org/10.1186/2049-2618-2-29.
48. Leong S, Zhao Y, Joseph NMNM, Hochberg NSNS, Sarkar S, Pleskunas J, et al. Existing blood transcriptional classifiers accurately discriminate active tuberculosis from latent infection in individuals from south India. Tuberculosis. 2018;109:41–51. https://doi.org/10.1016/j.tube.2018.01.002.
49. Baldini F, Heinken A, Heirendt L, Magnusdottir S, Fleming RMT, Thiele I. The microbiome modeling toolbox: From microbial interactions to personalized microbial communities. Bioinformatics. 2019;35:2332–4. https://doi.org/10.1093/bioinformatics/bty941.
50. Venturelli OS, Carr AV, Fisher G, Hsu RH, Lau R, Bowen BP, et al. Deciphering microbial interactions in synthetic human gut microbiome communities. Mol Syst Biol. 2018;14. https://doi.org/10.15252/msb.20178157.
51. Kishore D, Birzu G, Hu Z, Delisi C, Korolev KS, Segrè D. Inferring microbial co-occurrence networks from amplicon data: a systematic evaluation. bioRxiv. 2020:2020.09.23.309781. https://doi.org/10.1101/2020.09.23.309781.
52. Bernstein DB, Dewhirst FE, Segrè D. Metabolic network percolation quantifies biosynthetic capabilities across the human oral microbiome. Elife. 2019;8:e39733. https://doi.org/10.7554/eLife.39733.

Publisher's Note

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