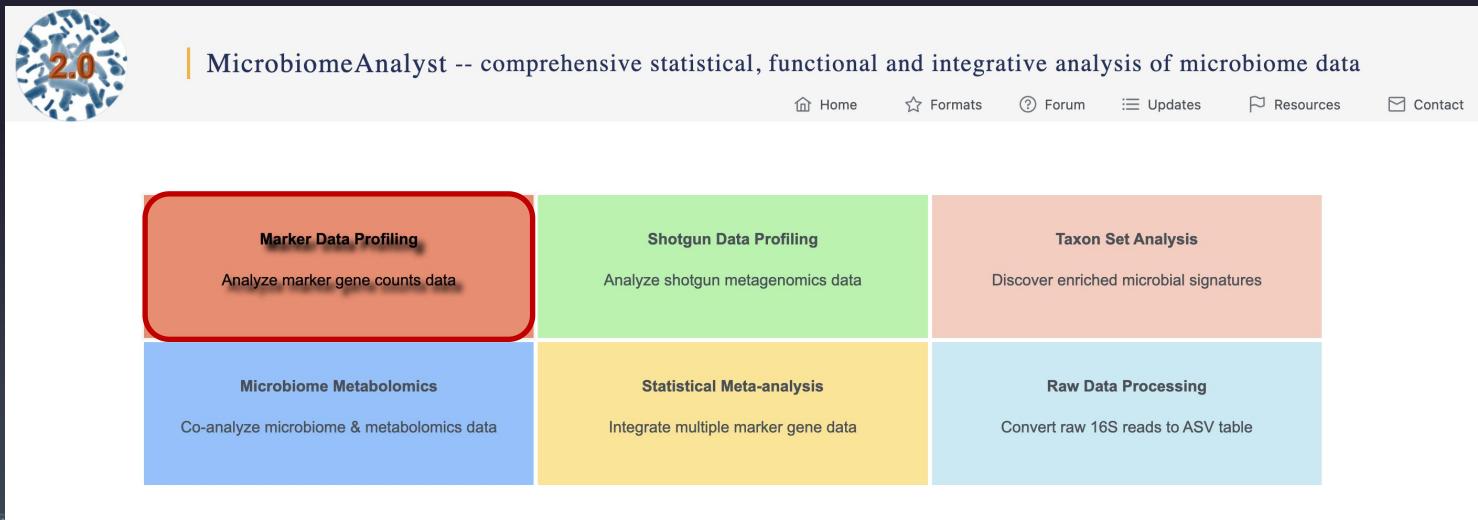


MicrobiomeAnalyst 2.0

Comprehensive statistical, functional and
integrative analysis of microbiome data

Tutorial for Marker Data Profiling



The image shows the MicrobiomeAnalyst homepage. At the top left is a blue circular logo with '2.0' in red. To its right is the header text 'MicrobiomeAnalyst -- comprehensive statistical, functional and integrative analysis of microbiome data'. Below the header are six navigation links: Home, Formats, Forum, Updates, Resources, and Contact. The main content area is divided into six colored boxes arranged in two rows of three. The top row contains 'Marker Data Profiling' (red), 'Shotgun Data Profiling' (green), and 'Taxon Set Analysis' (orange). The bottom row contains 'Microbiome Metabolomics' (blue), 'Statistical Meta-analysis' (yellow), and 'Raw Data Processing' (light blue). Each box has a brief description below it. The 'Marker Data Profiling' box is highlighted with a red border.

Marker Data Profiling	Shotgun Data Profiling	Taxon Set Analysis
Analyze marker gene counts data	Analyze shotgun metagenomics data	Discover enriched microbial signatures
Microbiome Metabolomics	Statistical Meta-analysis	Raw Data Processing
Co-analyze microbiome & metabolomics data	Integrate multiple marker gene data	Convert raw 16S reads to ASV table

Overview

Motivation: The previous version of MicrobiomeAnalyst provided a user-friendly web-based platform that helped users to perform comprehensive exploratory analysis on marker gene data. However, the fast-evolving methods, knowledge and datasets arising from current microbiome data analysis call for up-to-date tools.

Goal: To provide a real-time platform for maker gene data analysis that allows users to easily explore and understand their data using updated methods and knowledge databases.

Enhanced Features in Version 2.0

- ❖ Editable metadata and multi-factor comparison analysis
- ❖ Deal with the normalized input data
- ❖ Update the methods for correlation analysis
- ❖ Update Statistical methods for significance testing in beta-diversity profiling
- ❖ Add Tax4Fun2 for function prediction and update the background database
- ❖ Enhanced visualization: interactive barplot and heatmap

Analysis Strategies

Visual Exploration

- Interactive stack
- bar/area plot
- Interactive pie chart
- Rarefaction curve
- Phylogenetic tree
- Heat tree

Community Profiling

- Alpha diversity
- Beta diversity
- Core microbiome

Clustering & Correlation

- Interactive Heatmap
- Dendrogram
- Correlation network
- Pattern search

Comparison & Classification

- Single-factor analysis
- Multi-factor analysis
- LEfSe
- Random Forest

Functional Prediction

- PICRUSt (Greengenes)
- Tax4Fun (SILVA)
- Tax4Fun2

Data Formatting

- Text file:
tab delimited (.txt) /
comma-separated (.csv) file
- BIOM format
- Mothur output:
.shared (abundance) file
.taxonomy file

Count table

#NAME	Sample1	Sample2	...	Sample
ID1	10035	2204	...	0
ID2	214	0	...	26
...
ID	0	89	...	0

Metadata

#NAME	study_group	gender	...
Sample1	Control	M	...
Sample2	Case	M	...
...
Sample	Case	F	...

Taxonomy table(optional)

#TAXONOMY	Kingdom	Phylum	Class	Order	Family	Genus	Species
ID1	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Bacteroidaceae	...	plebeius
ID2	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae	...	NA
...
ID	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Bacteroidaceae	...	NA

Data Upload

Please upload your data based on their formats or try our examples.

Text table format BIOM format MOTHUR outputs Try our examples

OTU/ASV table (.txt, .csv, or its zip) Taxonomy included Sequences included Normalized data
+ Choose ?

Metadata file (.txt or .csv)
+ Choose ?

Taxonomy table (.txt or .csv)
+ Choose ?

(Optional) phylogenetic tree (.tre, .nwk)
+ Choose ?

Taxonomy labels
--- Not specified --- ▾

--- Not specified --- ▾
Greengenes Taxonomy
SILVA Taxonomy
Greengenes OTU ID
QIIME
Not Specific / Other

Submit

Select the format of your data to upload

If you would like use Tax4Fun2 for function prediction, make sure to have the asv sequences in the #NAME column and check here

Check here if taxonomy labels are included in the count table

Specify the taxonomy label for parsing and function prediction accordingly.

If your data has already been normalized, read the notes below and check here.

Did you know?

It is highly advised to upload your OTU/ASV abundance table containing **raw counts** to benefit the best practices for data analysis. If your data has already been normalized:

- Indicate it is Normalized data during data upload (this page);
- (Optional) Bypass data filtering and normalization during data inspection (next page);
- Some methods (abundance profiling, alpha diversity, function prediction, etc) will become inappropriate during data analysis.

Data Integrity Check

Available file downloads for each page are displayed here

Downloads of the page

- [Lib Size View \(PDF\)](#)
- [Lib Size View \(SVG\)](#)
- [Lib Size Data \(CSV\)](#)

R Command History

[Clear](#) [Save](#)

```
1. mbSet<-Init.mbsSetObj()
2. mbSet<-SetModuleType(mbSet, "m
dp")
3. mbSet<-Read16SAbundanceData(mbSet,
"otu_table_mc2_w_tax_no_pynast
_failedbiom","biom","GreengenesD",
"true", "false");
4. mbSet<-ReadSampleTable(mbSet,
"map.txt");
5. mbSet<-ReadTreeFile(mbSet, "re
p_set.tre", "otu_table_mc2_w_t
ax_no_pynast_failures.biom", "m
dp");
6. mbSet<-SanityCheckData(mbSet,
"biom");
7. mbSet<-SanityCheckSampleData(m
bSet);
8. mbSet<-SetMetaAttributes(mbSe
t, "1");
9. mbSet<-PlotLibSizeView(mbSet,
"noLibsizes", 0, "png");
10. mbSet<-CreatePhyloseqObj(mbSe
t, "GreengenesID", "F", "T",
"false");
11. mbSet<-ApplyAbundanceFilter(mb
Set, "Prevalence", 4, 0.2);
12. mbSet<-ApplyVarianceFilter(mbS
et, "Iqr", 0.1);
13. mbSet<-PerformNormalization(mb
Set, "none", "colsum", "none",
"true");
```

Data Integrity Check

Data Check

- Feature abundance table contains raw counts (preferred) or normalized values;
- Features with identical values (i.e. zeros) across all samples will be excluded;
- Features that appear in only one sample will be excluded (considered artifacts);
- For ASV data, which uses actual sequences as IDs, the sequence IDs will be replaced with ASV_1, ASV_2, etc.

Metadata Check

- For categorical metadata, at least two groups and three replicates per group are required; a metadata column must have at least 3 distinct values.
- For continuous metadata, all values must be numerical.
- Missing values are not allowed in metadata.
- Use the [Edit Metadata](#) tab to inspect and manually address the issues

Text Summary

Library Overview Edit Metadata

Data type:

OTU abundance table

File format:

biom

Sample names match (metadata vs. OTU table):

Yes

Normalized counts detected:

No

OTU annotation:

GreengenesID

OTU number:

3426

OTUs with ≥ 2 counts:

2920

Number of experimental factors:

7

Number of experimental factors with replicates:

7 [discrete: 7 continuous: 0]

Total read counts:

180573

Average counts per sample:

5310

Maximum counts per sample:

11313

Minimum counts per sample:

1114

Phylogenetic tree uploaded:

Yes

Number of samples in metadata:

34

Number of samples in OTU table:

34

Number of sample names matched (metadata vs. OTU table):

34

Number of samples that will be processed:

34



loads page).

are detected.

Please note that only name matched samples will be processed

R commands are shown here

If your data is normalized, you can use these button to skip the filtration and normalization steps

<< Previous

>> Analysis View

>> Proceed

Edit Metadata

Text Summary Library Size Overview **Edit Metadata**

• Update metadata type: categorical option for experimental groups (i.e. control vs diseased), continuous for numerical measures;
• Edit metadata content: click [Edit](#) to modify underlying groups to address those that do not meet requirements.
• Modify metadata name: click on corresponding cell on the main table to modify name
• Specify group order of categorical metadata: click [Edit](#) and go to [Order](#) tab to specify the order (low, medium, high). By default, they are ordered by alphabetical order.
• Exclude metadata that do not pass sanity check.

Currently selected data: --- Not available ---

Name	Status	Type	Edit	Remove
SampleType	OK	Categorical	Edit	
Year	OK	Categorical	Edit	
Month	OK	Categorical	Edit	
Day	OK	Categorical	Edit	
Subject	OK	Categorical	Edit	
ReportedAntibioticUsage	OK	Categorical	Edit	
DaysSinceExperimentStart	OK	Categorical	Edit	

Make sure all variable types were inferred correctly

For categorical, adjust 'Order' to control order in downstream plots and analysis

Edit metadata

[Edit \(sample-level\)](#) **Order (factor-level)** [Edit \(factor-level\)](#)

Available

Yes
No

[Update](#) [Cancel](#)

Data Filtering

Data Filtering

Data filtering aims to remove low quality or uninformative features to improve downstream statistical analysis. You can disable any data filter by dragging the slider to the left until it is at the minimum value.

- Low count filter - features with very small counts in very few samples are likely due to sequencing errors or low-level contaminations. You need to first specify a minimum count prevalence filter means at least 20% of its values should contain at least 4 counts. You can also filter based on their *mean* or *median* values.
- Low variance filter - features that are close to constant throughout the experiment conditions are unlikely to be associated with the conditions under study. Their variances can be measured using *inter-quartile range (IQR)*, *standard deviation* or *coefficient of variation (CV)*. The lowest percentage based on the cutoff will be excluded.

By default, all downstream data analysis will be based on filtered data. You can choose to use the original unfiltered data for some analyses (i.e. alpha diversity).

Low count filter

Minimum count: 4

Prevalence in samples (%) 20

Mean abundance value

Median abundance value

Submit Edit Samples

Low variance filter

Percentage to remove (%): 10

Inter-quartile range

Based on: Standard deviation

Coeffecient of variation



Sample Editor

Note you must click the Submit button below to complete sample removal. After the data updates, you need to re-perform the data filtering and normalization steps again.

Available	Exclude
Urial2	
Okapi1	
Okapi2	
BlackLemur	
BigHornW3	
Gazelle3	
BlackRhino1	
BaboonW	
Chimp1	
SpgbkW	
BushDog1	

Submit

Users can remove samples that are detected as outlier via results from graphical summary or rarefaction curve analysis.

<< Previous

>> Proceed

Data Normalization

Data Normalization

Normalization aims to address the variability in sampling depth and the sparsity of the data to enable more biologically meaningful comparisons. When the library sizes are very similar, rarefaction is also recommended (see [Weiss, S et al.](#)). Note, rarefying is mainly used for 16S marker gene data and is disabled for shotgun metagenomics data. All of these methods require raw data. You can either rarefy your data followed by either data scaling or data transformation. However, you cannot apply **both** data scaling and data transformation, because scaled or transformed data is no longer raw count data.



OK

No data rarefaction was performed.
Performed total sum normalization.
No data transformation was performed.

Data rarefying

Do not rarefy my data
 Rarefy to the minimum library size

Data scaling

Do not scale my data
 Total sum scaling (TSS)
 Cumulative sum scaling (CSS)
 Upper-quartile normalization (UQ)

Data transformation

Do not transform my data
 Relative log expression (RLE)
 Trimmed mean of M-values (TMM)
 Centered log ratio (CLR)

Submit

- Normalization is required to account for uneven sequencing depth, undersampling and sparsity present in such data. (useful before any meaningful comparison)
- Several commonly used methods are present. (3 categories: rarefaction, data scaling and data transformation)
- Check rarefaction curve to get the minimum sequence depth of your libraries. If the minimum library size is too small, you can either resequence your samples or exclude them from downstream analysis.

Analysis approaches selection

Analysis Overview

A

Visual Exploration

[Stacked bar/area plot](#) [Interactive pie chart](#) [Rarefaction curve](#) [Phylogenetic tree](#) [Heat tree](#)

Data overview and general pattern discovery through intuitive visualization techniques

B

Community Profiling

[Alpha diversity](#) [Beta diversity](#) [Core microbiome](#)

Quantitative analysis of community profiles using multiple well-established statistical methods

C

Clustering & Correlation Network

[Interactive Heatmap](#) [Dendrogram](#) [Correlation network](#) [Pattern search](#)

Identifications of inherent patterns and correlations within your data (unsupervised)

D

Comparison & Classification

[Single-factor analysis](#) [Multi-factor analysis](#) [LEfSe](#) [Random Forest](#)

Identification of significant features or potential biomarkers via statistical and machine learning methods (supervised)

E

Functional Prediction

[PICRUSt \(Greengenes\)](#) [Tax4Fun \(SILVA\)](#) [Tax4Fun2](#)

Prediction of metagenome functional profiles from 16S marker gene data

A. Visual Exploration

Stacked Bar/Area plot:

- Provides exact composition of each community through direct quantitative comparison of abundances.
- It can be created for all samples, sample-group wise or individual sample-wise at multiple taxonomic level present in data.

Abundance Profiling

Data options Organize samples by SampleType then by None
Merge samples to groups SampleType then by None
View an individual sample L1S140

Taxa resolution Phylum prepend higher taxa
Merging small taxa with counts < 10 based on Total
Showing top n taxa, with n = 10

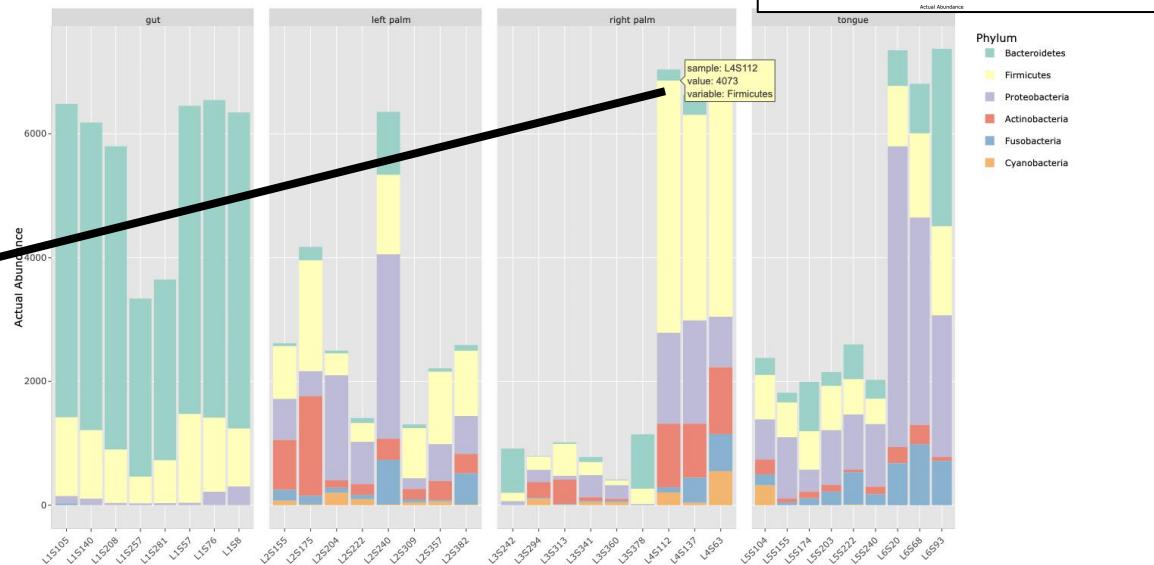
Graph options Graph type Stacked Bar (Actual Abundance)
Color scheme Set3

Mouse over to see the labels; click and drag to zoom-in and double-click to zoom-out completely

Submit Merge samples

Phylum

- Bacteroidetes
- Firmicutes
- Proteobacteria
- Actinobacteria
- Fusobacteria
- Cyanobacteria



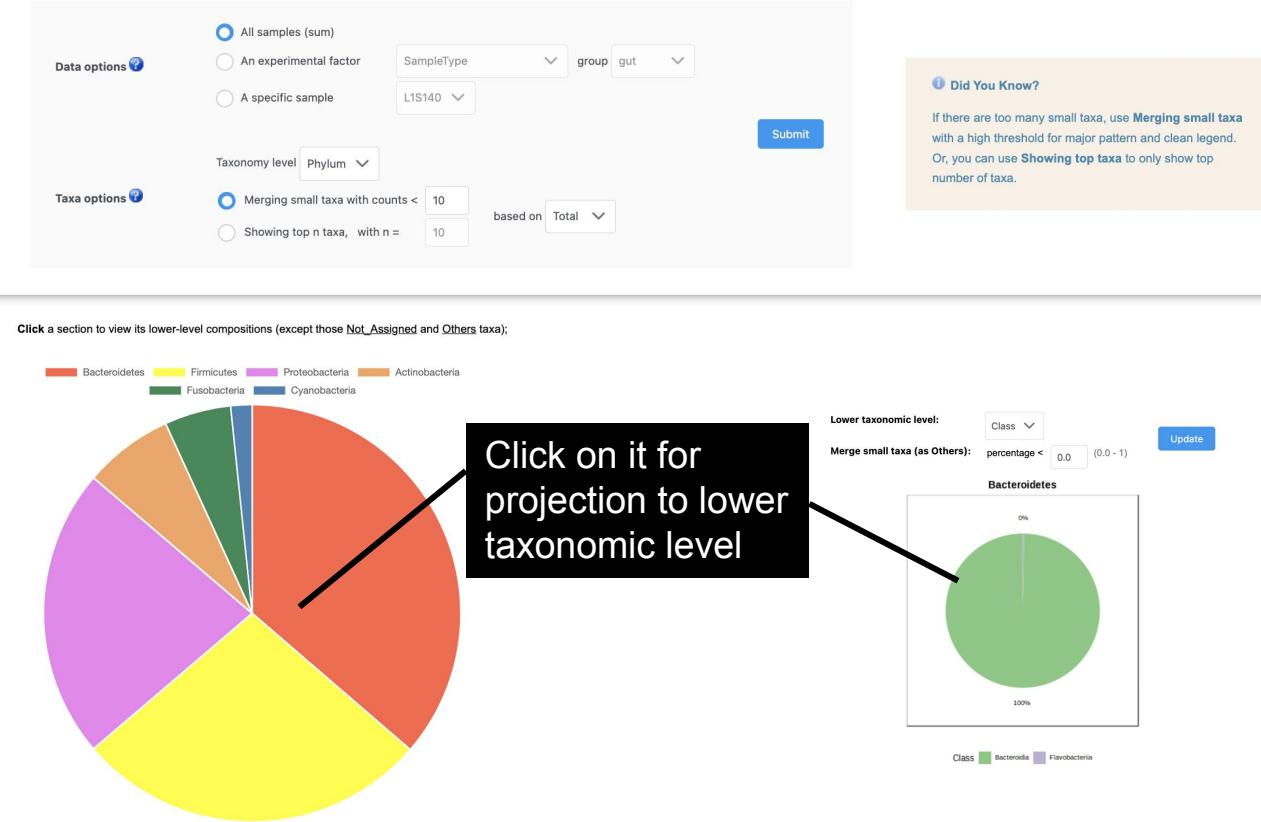
Mouse over to see the detail information

Use these options to adjust data groups, taxonomy included as well as heatmap view.

A. Visual Exploration

Pie Chart:

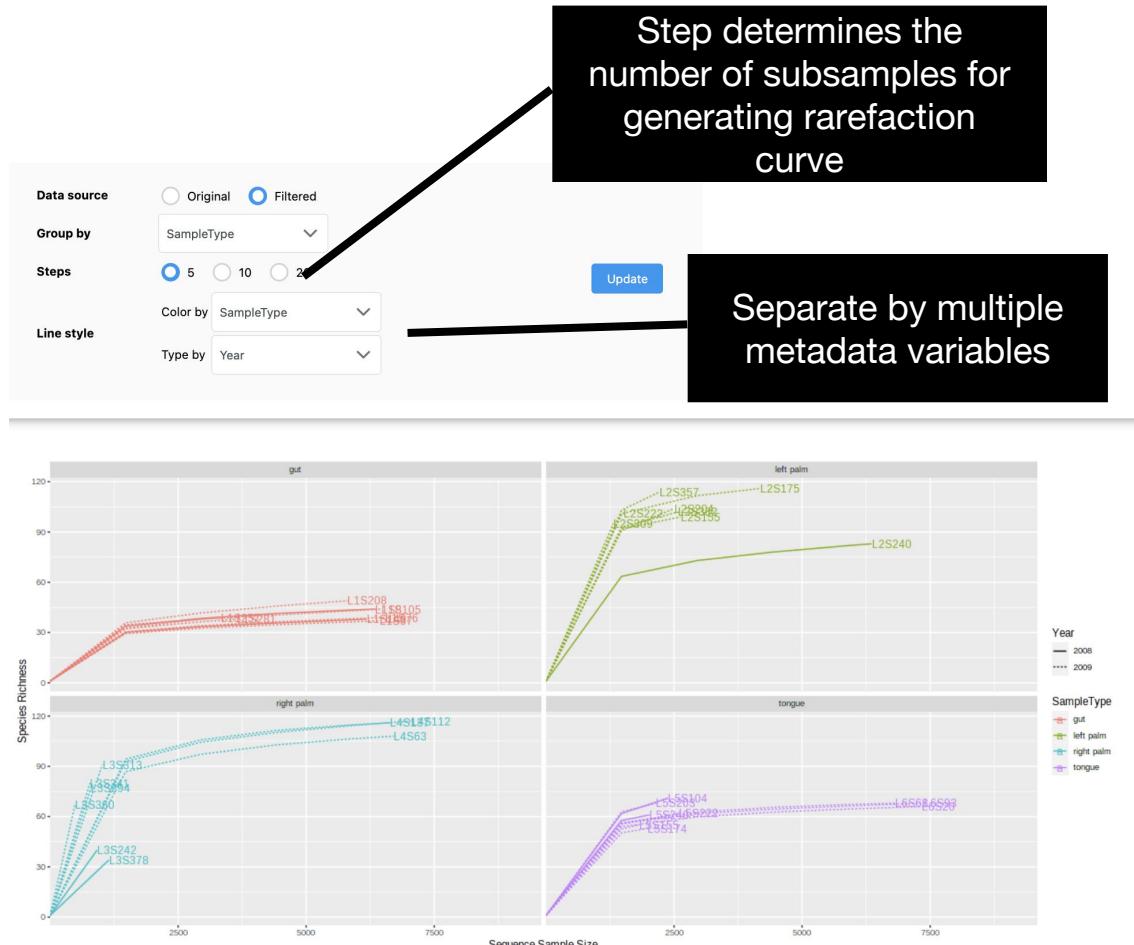
- Visualize the taxonomic compositions of microbial community.
- It can be created for all samples, sample-group wise or individual sample-wise at multiple taxonomic level present in data.



A. Visual Exploration

Rarefaction curve:

- Helps in determining number of observed OTUs (alpha diversity)
- Determining sequence depth of each sample
- Determining if sample reaches sequencing plateau (number of recovered OTUs increase with increasing sequence depth)
- If sequence depth is not enough to reach plateau, you can consider to resequence these samples to increase sequence depth
- Helps in deciding if the dataset should be rarefied or excluding samples (not enough reads and have not reach plateau) from downstream analysis



A. Visual Exploration

Phylogenetic tree:

Helps in determining evolutional relations among different taxonomic groups at different levels.

Two types of tree shapes are provided:
Rectangular and
Radial

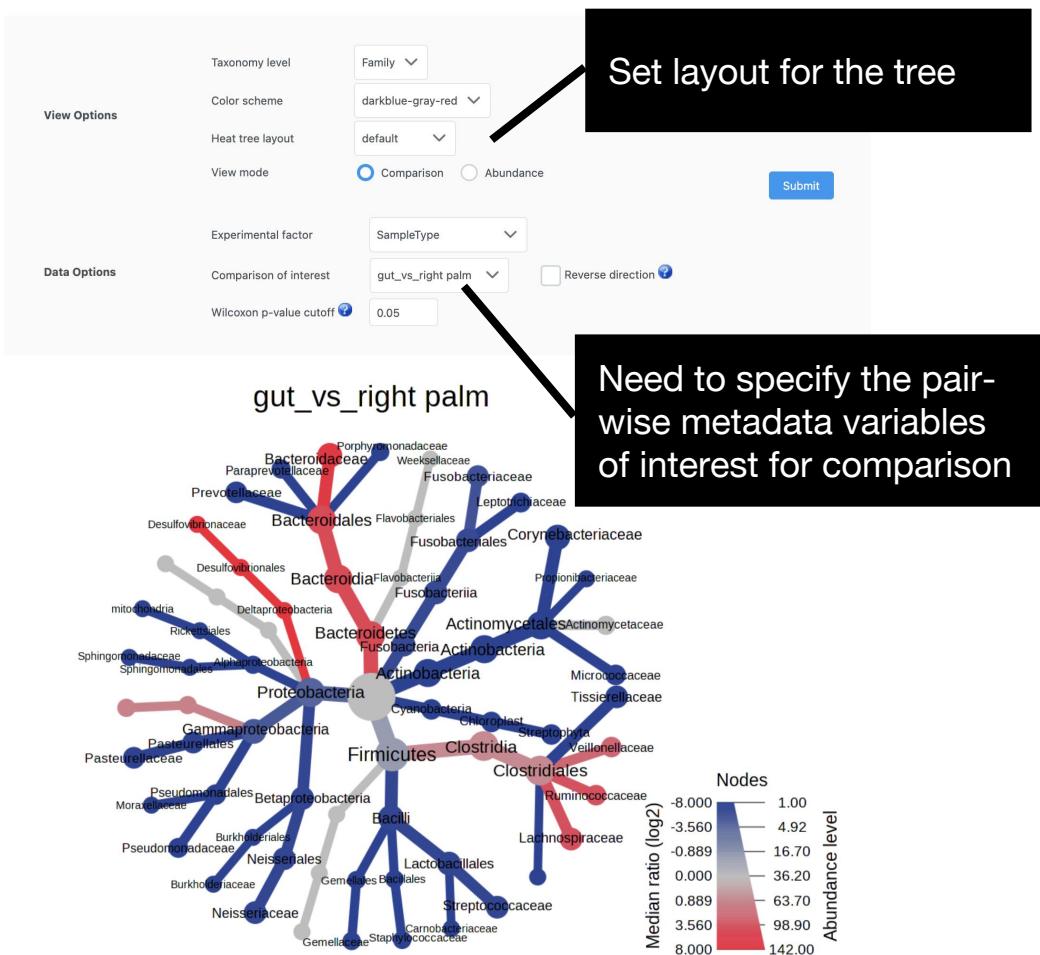


Select multiple metadata variables to customize tree

A. Visual Exploration

Heat tree:

- A hierarchical tree of taxonomic levels with abundance indicated by colors.
- It presents abundance ratios of two groups at each taxonomic level
- It can compare every pair of factors in each metadata variable



B. Community Profiling

Alpha diversity profiling:

- Supporting 6 widely used metrics to calculate the alpha diversity: Chao1 and ACE (estimated number of OTUs), Observed number of OTUs for richness, Shannon and Simpson take account for both evenness and richness.
- Statistical significance testing between groups using parametric and non-parametric tests.

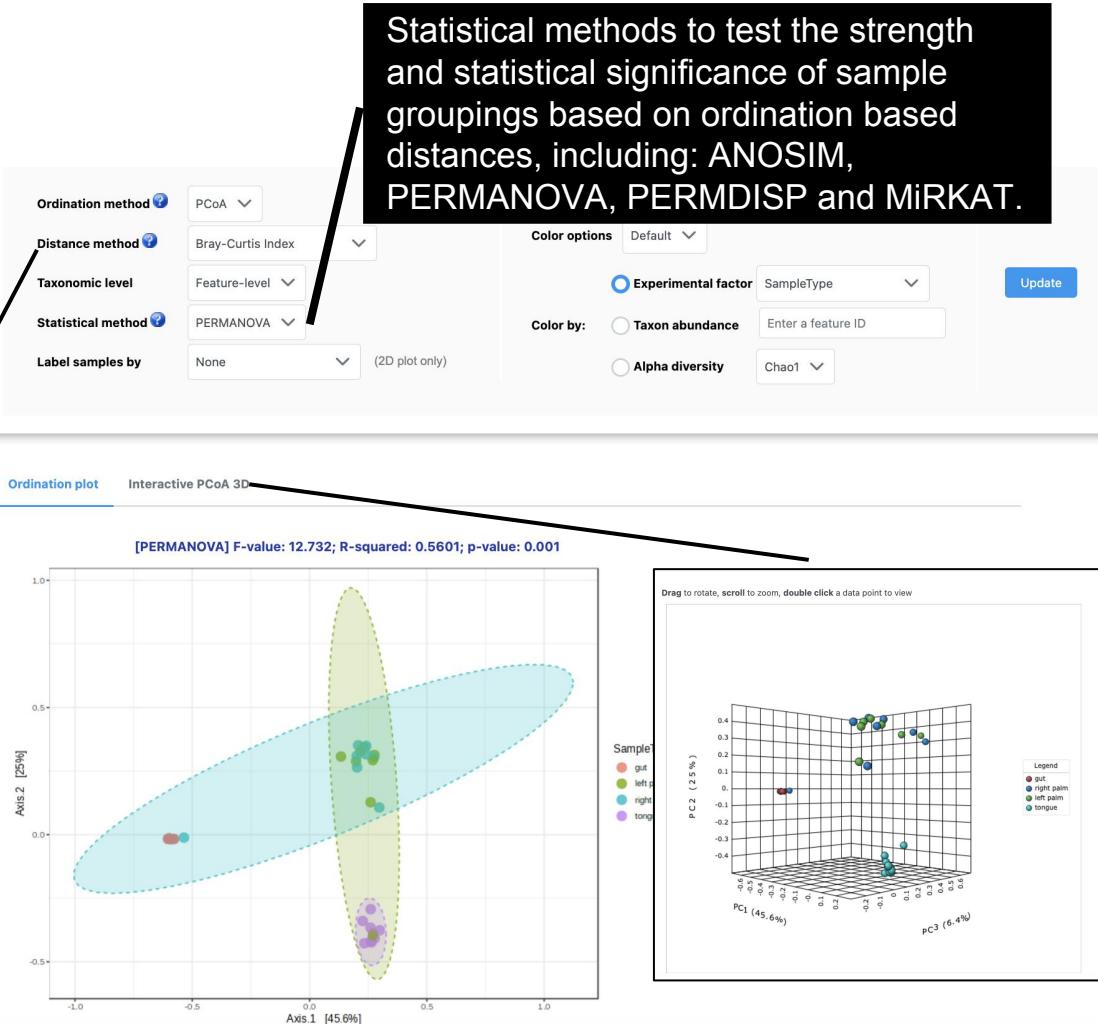
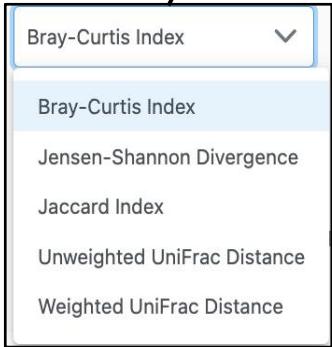


B. Community Profiling

Beta diversity profiling:

- Assess the differences between microbial communities(or samples)
- Visualize using PCoA (Principal Coordinate Analysis) or NMDS (Nonmetric Multidimensional Scaling)

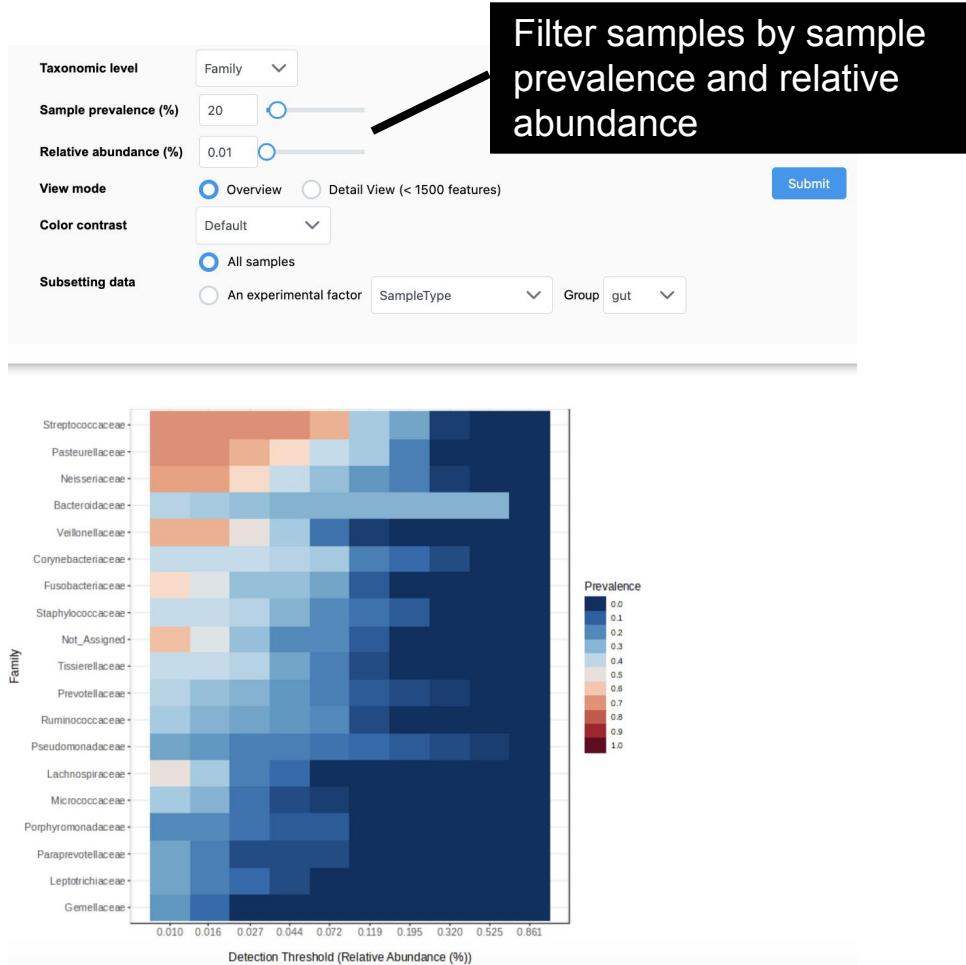
Phylogenetic tree
need to be provided
for unweight- and
weight unifrac
distances



B. Community Profiling

Core microbiome analysis:

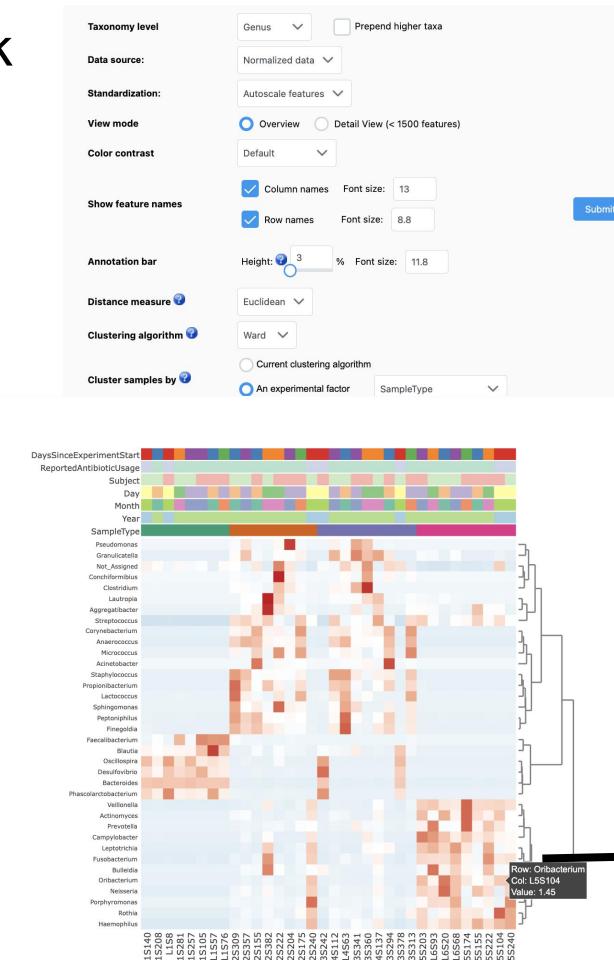
Helps in identifying core taxa or features that remain unchanged in their composition across different sample groups based on sample prevalence and relative abundance.



C. Clustering & Correlation Network

Clustering Heatmap Visualization:

- Visualize the relative patterns of high-abundance features against a background of features that are mostly low-abundance or absent.
 - Identify abundance patterns, clusters
 - Various distance and clustering methods supported.(both sample and feature-wise)



See the heatmap
in a new tab in
detail view

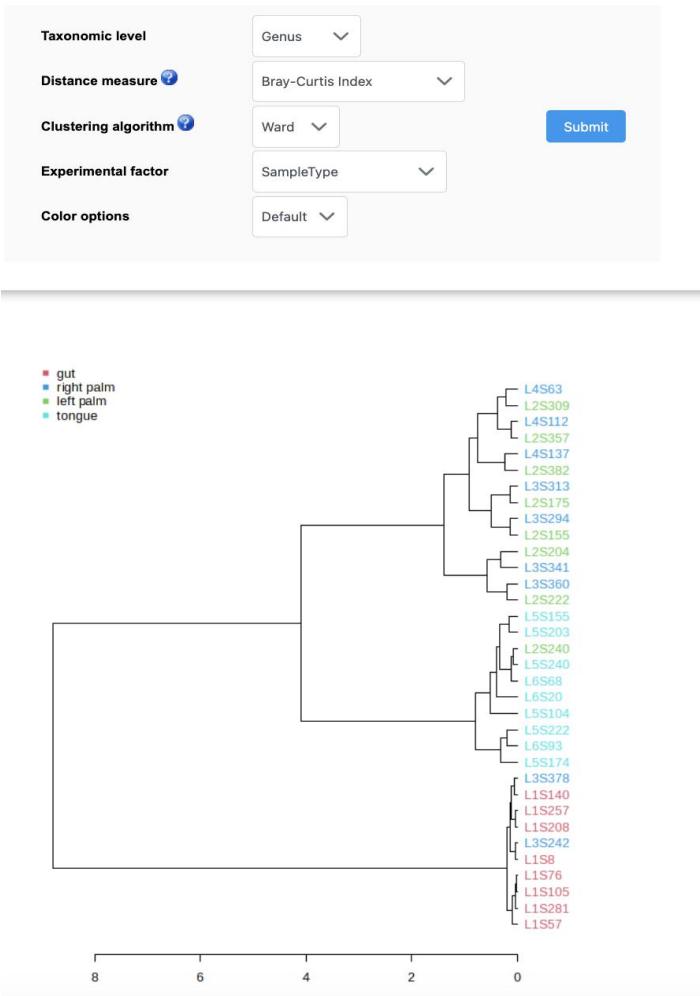
Reset the heatmap

Mouse over to
see the detail
infomation

C. Clustering & Correlation Network

Dendrogram Analysis

- Performs phylogenetic analysis on samples using either various phylogenetic or nonphylogenetic distance measures.
- Unweighted and weighted unifrac distances are based on phylogenetic tree, therefore, phylogenetic tree must be provided to calculate these distances.



C. Clustering & Correlation Network

Correlation Analysis

To identify biologically meaningful relationship or associations between taxa or features.

Seven statistical method are provided to calculate the correlation including SECOM (Pearson1), SECOM (Pearson2), SECOM (Distance), SparCC, Pearson, Spearman and Kendall.

Set the comparison of interest here

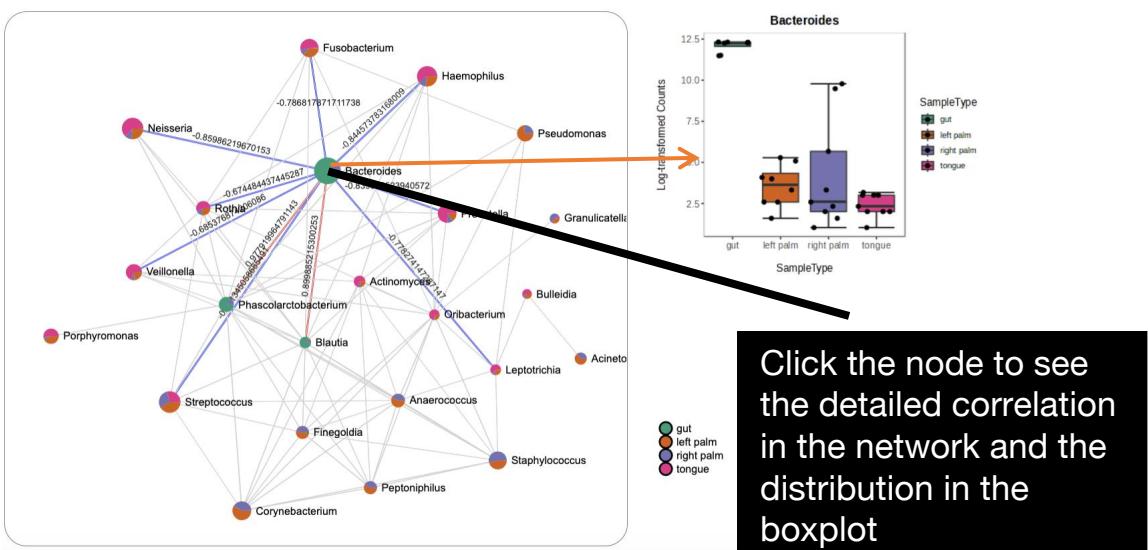
Change the piechart style of the node here

Algorithm: SECOM (Pearson1)
Taxonomy level: Genus
Experimental factor: SampleType
Analysis mode: All groups
Comparison of interest: Specify

Permutation (SparCC): 100
P-value threshold: 0.05
Correlation threshold: 0.3

Node style: Piechart (relative abundance) High-level taxonomy Phylum

Submit



C. Clustering & Correlation Network

Pattern Search

- Helps in identifying or search for a pattern based on correlation analysis on defined pattern.
- Pattern can be defined based on either feature of interest or based on predefined or custom profile of experimental factors.

Define your own pattern of interest

Taxonomy level: Genus prepend higher taxa

Distance measure: Pearson r

Experimental factor: SampleType

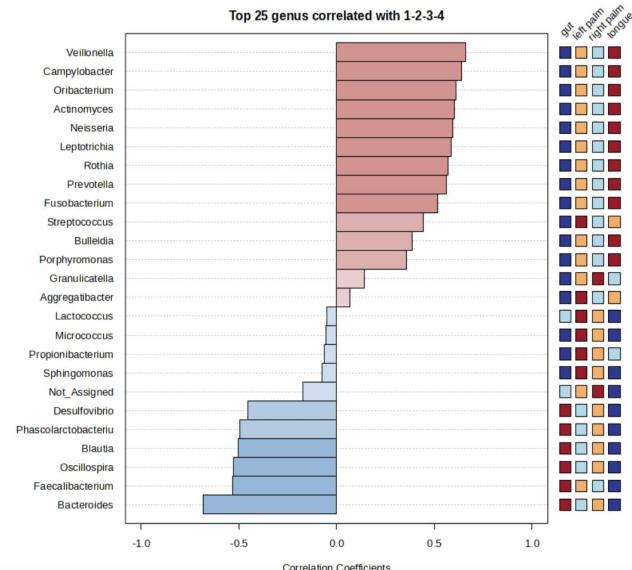
Pattern based on: Predefined A feature Custom profile

1-2-3-4 Metabacter

Submit

Graphical Summary Result Table

Check the statistical result here.



D. Comparison & Classification

Single-factor analysis

T-test/ANOVA

Mann-Whitney/Kruskal-Wallis

T-test/ANOVA

metagenomeSeq (0-inflated)

metagenomeSeq (fitFeature)

EdgeR

DESeq2

Taxonomy level: Genus

Experimental factor: SampleType

Statistical method: T-test/ANOVA

Adjusted p-value cutoff: 0.05

Submit

Select the metadata of interest

Select statistical methods



D. Comparison & Classification

Multi-factor analysis: Model Parameters

The screenshot shows a web-based analytical tool for performing multiple regression analysis. The top navigation bar includes links for Data Upload, Data Inspection, Data Filter, Normalization, Analysis Overview, Multiple Regression, and Downloads. A 'Navigate to:' dropdown and a 'Show Info Pane' button are also present.

The main content area is titled "Multiple Linear Regression with Covariate Adjustment". It describes the tool's purpose: finding associations between microbial features and primary metadata. It lists three types of variables:

- Primary metadata:** included as a 'fixed effect' in the model. Statistical groups, you must specify the comparison of interest.
- Covariates (control for):** included as 'fixed effects' in the model. These variables are accounted for in the statistics extracted for the primary metadata.
- Blocking factor:** included as 'random effects' in the model. These variables are accounted for in the statistics extracted for the primary metadata. Note that the blocking factor must have a reasonably balanced design with adequate sample size or the contrast matrix will be rank deficient.

The interface includes several input fields and dropdown menus:

- Taxonomy level: Feature-level
- Primary metadata: SampleType
- Comparison: tongue vs. gut
- Covariates (control for): Subject, ReportedAntibioticUsage
- Blocking factor: -- Unspecified --
- Adjusted p-value cutoff: 0.05

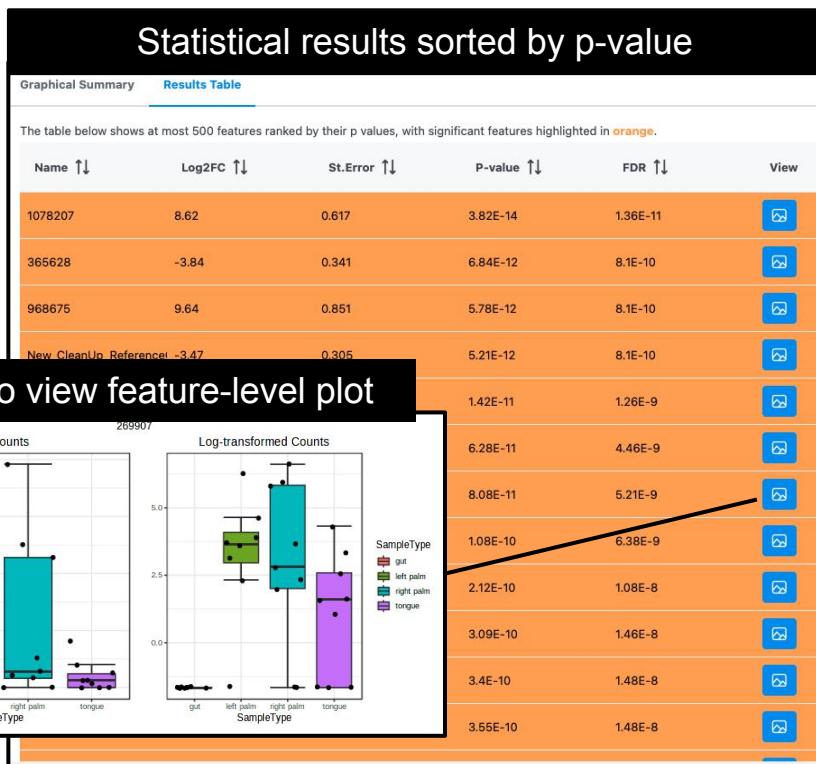
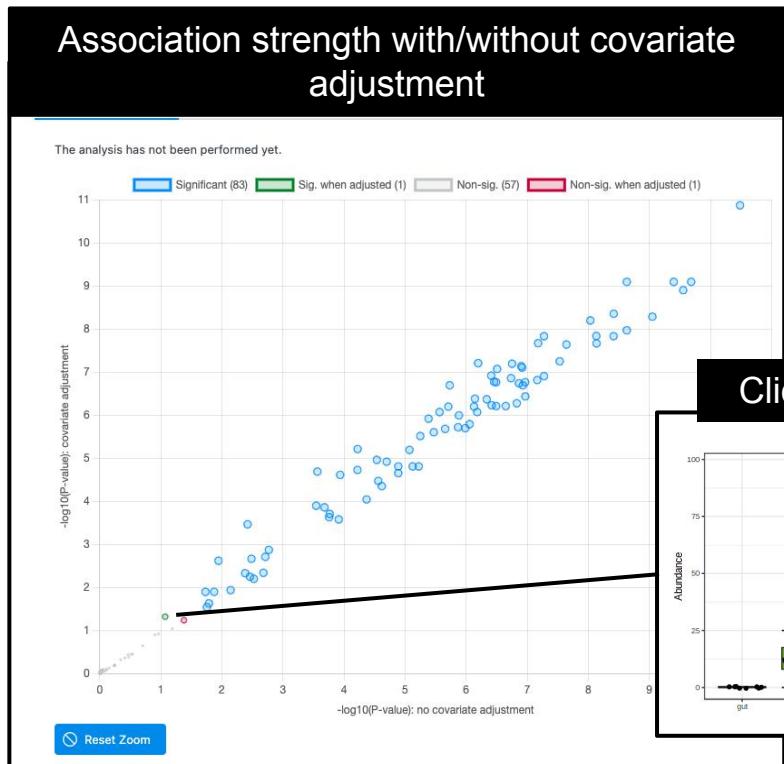
A blue "Submit" button is located at the bottom right of the form.

Annotations provide instructions for using the tool:

- A black callout box points to the "Comparison" dropdown: "Find associations between microbial features and this metadata".
- A black callout box points to the "Primary metadata" dropdown: "If primary metadata is categorical, specify the comparison of interest".
- An orange callout box points to the "Submit" button: "Click 'Submit' after specifying the model parameters".
- A black callout box points to the "Covariates (control for)" field: "Specify all variables that you'd like to account for as 'fixed effects' here".

D. Comparison & Classification

Multi-factor analysis: Results



D. Comparison & Classification

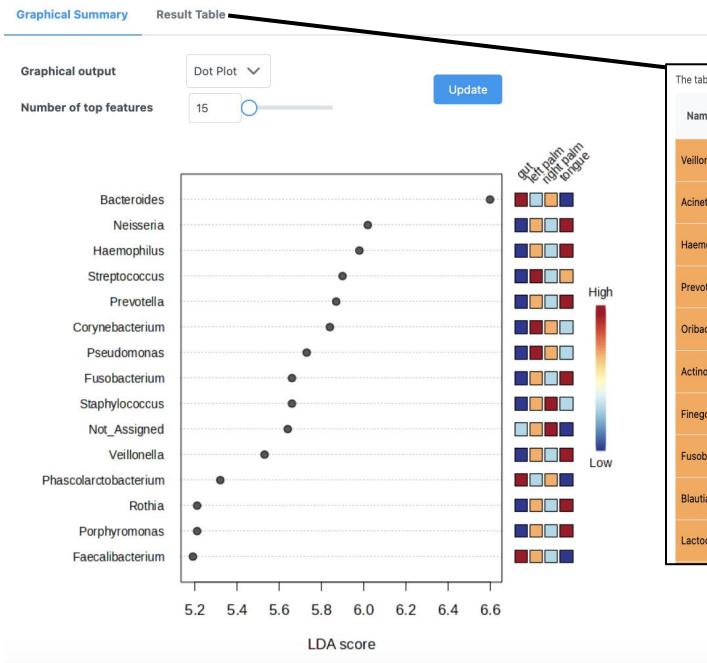
Taxonomy level: Genus

Experimental factor: SampleType

P-value cutoff: 0.1 (Original) FDR-adjusted

Log LDA score: 2.0

Submit

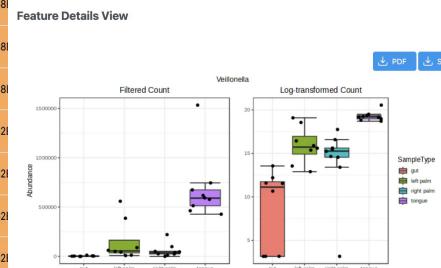


Linear Discriminant Analysis Effect Size (LEfSe):

Performs a set of statistical tests for detecting differentially abundant features (KW sumrank test: statistical significance) and biomarker discovery.(Linear Discriminant analysis: Effect Size)

The table below shows at most 500 features ranked by their p values, with significant features highlighted in orange.

Name ↑↓	Pvalues ↑↓	FDR ↑↓	gut ↑↓	left palm ↑↓	right palm ↑↓	tongue ↑↓	LDAscore ↑↓	View
Veillonella	1.3148E-5	1.6548E-4	3062.3	150850.0	57132.0	682550.0	5.53	
Acinetobacter	1.3987E-5	1.6548					0.0	
Haemophilus	1.5584E-5	1.6548					1907000.0	
Prevotella	1.8387E-5	1.6548					1492000.0	
Oribacterium	2.7817E-5	1.6662					1492000.0	
Actinomyces	2.9274E-5	1.6662					1492000.0	
Finegoldia	3.3527E-5	1.6662					1492000.0	
Fusobacterium	3.7026E-5	1.6662					1492000.0	
Blautia	4.7001E-5	1.7255E-4					1492000.0	
Lactococcus	4.7932E-5	1.7255E-4					1492000.0	



D. Comparison & Classification

Taxonomy level: Genus

Experimental factor: SampleType

Choose metadata for predictors:

No. of trees to grow: 500

No. of predictors to try: 7

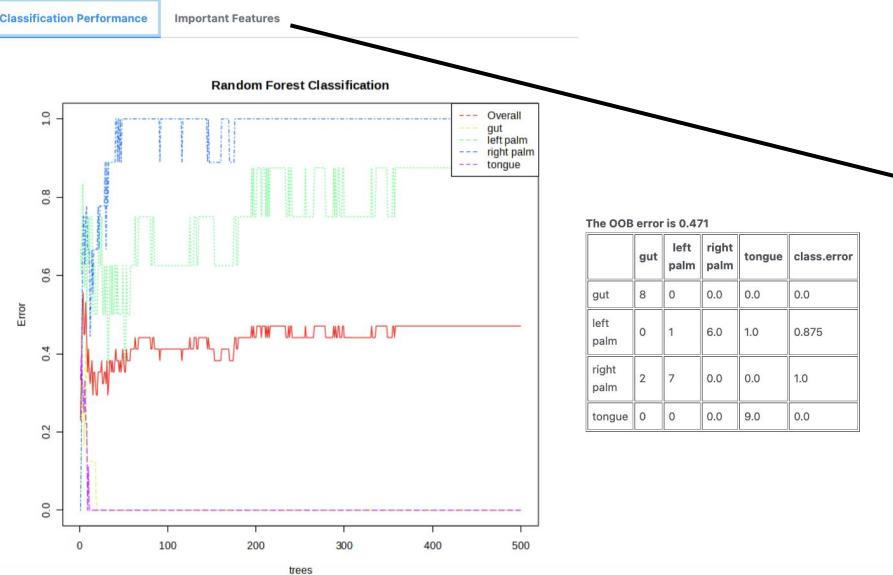
Randomness setting: On

Submit

No. of trees to be used for classification

No. of predictors for each node

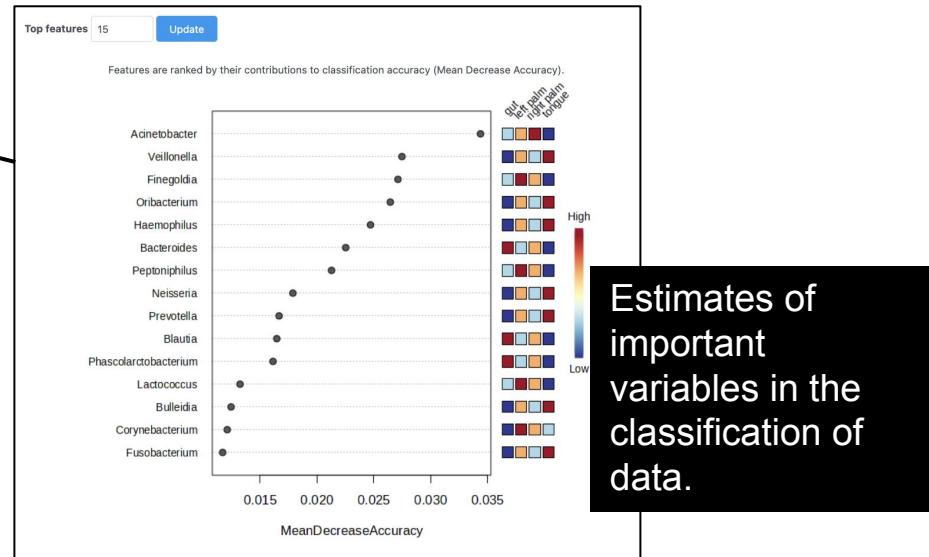
This screenshot shows a user interface for configuring a machine learning model. It includes dropdown menus for 'Taxonomy level' (Genus), 'Experimental factor' (SampleType), and 'Choose metadata for predictors'. Below these are input fields for 'No. of trees to grow' (set to 500) and 'No. of predictors to try' (set to 7). A 'Randomness setting' checkbox is checked. A large black box highlights the 'No. of trees to be used for classification' field. Another black box highlights the 'No. of predictors for each node' field.



Random forests:

Ensemble learning method used for classification, regression and other tasks.

- It operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) of the individual trees.
- Random forests correct for decision trees habit of overfitting to their training set.



E. Function prediction

Predicting functional capabilities of microbial communities using PICRUSt

PICRUSt (phylogenetic investigation of communities by reconstruction of unobserved states) estimates the properties of ancestral organisms from living relatives by performing gene content inference and metagenome inference. More details about this algorithm can be found from [MGI Langille et al.](#). Please make sure you have used **closed-reference OTU picking protocol** to search sequences against the [Greengenes reference OTUs](#) (May2012 version and May2013 version) to a specified percent identity.

Greengenes reference OTUs May2013 version ▾ Predict Functional Potential

PICRUSt need greengenes taxonomy annotation and specify the database here

Predicting functional capabilities of microbial communities using Tax4Fun

Tax4Fun is designed for functional prediction based on minimum 16S rRNA sequence similarity. It is applicable to outputs obtained from the [SILVAngs web server](#) or the application of [QIIME](#) against the [SILVA](#) database. Note, the process is time consuming and may take ~2 mins to complete. There will be an error with the box plots if the counts are relative. The result table can be used for functional profiling using our [Shotgun Data Profiling](#) module.

Annotation Pipeline QIIME against SILVA database ▾ Predict Functional Potential

QIIME against SILVA database SILVAngs

Predicting functional capabilities of microbial communities using Tax4Fun2

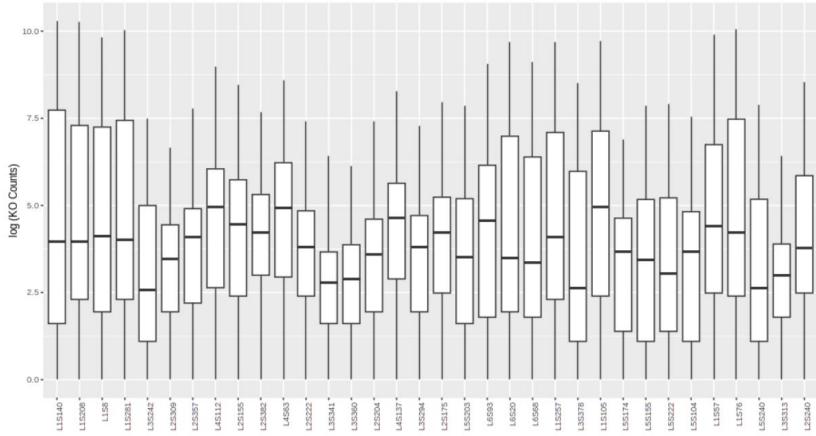
[Tax4Fun2](#) is used to predict functional profiles of prokaryotic communities based on 16S rRNA gene sequencing data. The prediction is based on the Ref99NR database. Note, Tax4Fun2 needs 16S rRNA gene sequences for prediction. Please make sure the sequences are included in the count tabel for Tax4Fun2.

Predict Functional Potential

Tax4Fun need SILVA taxonomy annotation and specify the pipeline here

ASV Sequences need to be included in the count tabel for Tax4Fun2

Result figure:



The KO table and figures can be dowloaded in the left panel.

The End



For more information, visit Tutorials, Resources
and Contact pages on www.microbiomeanalyst.ca
Also visit our forum for FAQs on www.omicsforum.ca