

MicrobiomeAnalyst 2.0

Comprehensive statistical, functional and integrative analysis of microbiome data



Tutorial for Marker Data Profiling



MicrobiomeAnalyst -- comprehensive statistical, functional and integrative analysis of microbiome data

[Home](#)

[Formats](#)

[Forum](#)

[Updates](#)

[Resources](#)

[Contact](#)

Marker Data Profiling

Analyze marker gene counts data

Shotgun Data Profiling

Analyze shotgun metagenomics data

Taxon Set Analysis

Discover enriched microbial signatures

Microbiome Metabolomics

Co-analyze microbiome & metabolomics data

Statistical Meta-analysis

Integrate multiple marker gene data

Raw Data Processing

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 marker 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

IDs can be OTU/ASV IDs, taxonomy labels or sequences

Headings are mandatory

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 | Enterobacteriales | Enterobacteriaceae | ... | NA |
| ... | ... | ... | ... | ... | ... | ... | ... |
| ID | Bacteria | Bacteroidota | Bacteroidia | Bacteroidales | Bacteroidaceae | ... | NA |

Data Upload

Please upload your data based on their format or try our

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



+ Choose



+ Choose



+ Choose



Metadata file (.txt or .csv)

Taxonomy table (.txt or .csv)

(Optional) phylogenetic tree (.tre, .nwk)

Taxonomy labels

--- Not specified ---

Submit

Select the format of your data to upload

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.

Check here if taxonomy labels are included in the count table

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

Specify the taxonomy label for parsing and function prediction accordingly.

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

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.mbSetObj()
2. mbSet<-SetModuleType(mbSet, "m
  dg")
3. mbSet<-Read16SAbundData(mbSet,
  "otu_table_mc2_w_tax_no_pynast
  _failures.biom", "biom", "Greeng
  enesID", "F", "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
  dg");
6. mbSet<-SanityCheckData(mbSet,
  "biom");
7. mbSet<-SanityCheckSampleData(m
  bSet);
8. mbSet<-SetMetaAttributes(mbSe
  t, "1");
9. mbSet<-PlotLibSizeView(mbSet,
  "norm_libsizes_0", "png");
10. mbSet<-CreatePhyloseqObj(mbSe
  t, "biom", "GreengenesID", "F",
  "1");
11. mbSet<-ApplyAbundanceFilter(mb
  Set, "abundance", 4, 0.2);
12. mbSet<-ApplyVarianceFilter(mbS
  et, "lgr", 0.1);
13. mbSet<-PerformNormalization(mb
  Set, "none", "column", "none",
  "true");
```

R commands are shown here

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, ...

Metadata Check

- For categorical metadata, at least two groups and three replicates per group are required; a metadata column with only one value is detected.
- 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 Size 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 |



Please note that only name matched samples will be processed

If your data is normalized, you can use these buttons 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 | |
| DaysSinceExperimentStar | 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

Filtering result will present here after submit

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** (you can also click the **Submit** button to apply the filter).

- **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. A 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.

OK
A total of 437 low abundance features were removed based on prevalence. A total of 4 low variance features were removed based on iqr. The number of features remains after the data filtering step: 28

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).

The interface shows two filter sections. The 'Low count filter' section has a 'Minimum count' slider set to 4, a 'Prevalence in samples (%)' slider set to 20, and radio buttons for 'Mean abundance value' and 'Median abundance value'. The 'Low variance filter' section has a 'Percentage to remove (%)' slider set to 10, and radio buttons for 'Inter-quartile range', 'Standard deviation', and 'Coefficient of variation'. A blue 'Submit' button is located to the right of the sliders.

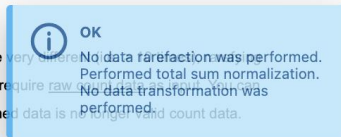
The 'Sample Editor' window has a title bar with a close button. Below the title is a note: 'Note you must click the **Submit** button below to complete sample removal. After the data updates, you need to re-perform the data filtering and normalization steps again.' The main area is divided into two columns: 'Available' and 'Exclude'. The 'Available' column contains a list of sample IDs: Urial2, Okapi1, Okapi2, BlackLemur, BigHornW3, Gazelle3, BlackRhino1, BaboonW, Chimp1, SpgbkW, and BushDog1. The 'Exclude' column is empty. Between the columns are four navigation buttons: a right arrow (>), a double right arrow (>>), a left arrow (<), and a double left arrow (<<). A blue 'Submit' button is at the bottom right of the window.

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

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 different, rarefaction was performed. Normalization 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. No data transformation was performed. 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 not original count data.



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.

Mouse over to see the detail information

Abundance Profiling

Data options

Organize samples by then by

Merge samples to groups then by

View an individual sample

Taxonomy level prepend higher taxa

Taxa resolution

Merging small taxa with counts < based on

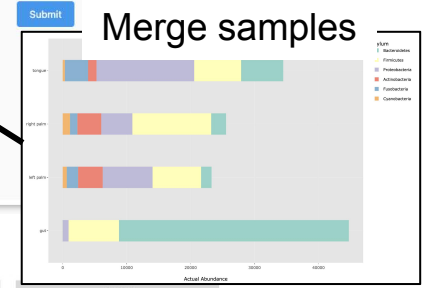
Showing top n taxa, with n =

Graph options

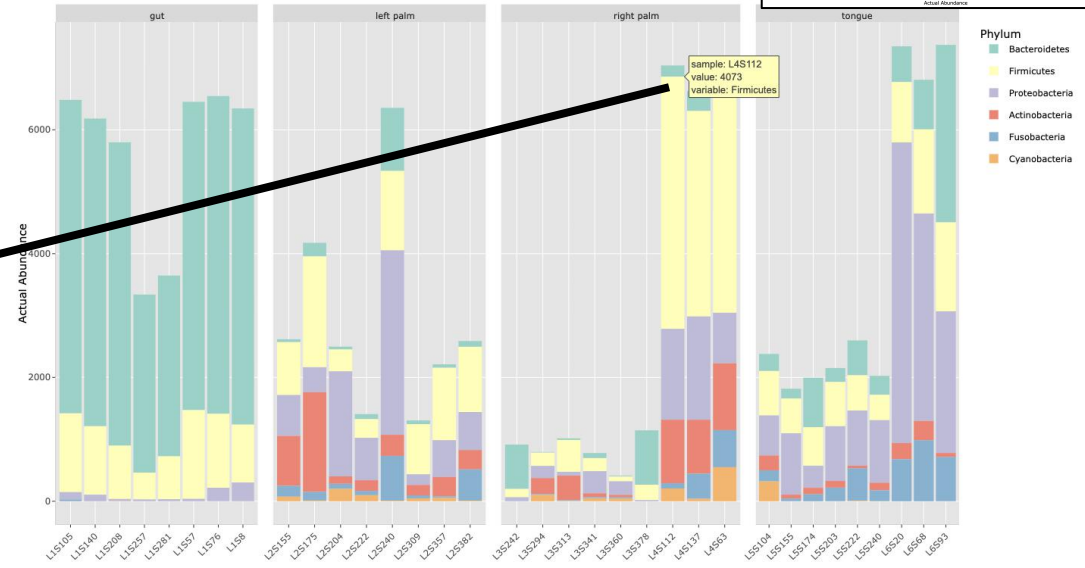
Graph type:

Color scheme:

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



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



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.

Data options ⓘ

All samples (sum)
 An experimental factor
 A specific sample

SampleType group

L1S140

Taxonomy level ⓘ

Phylum

Taxa options ⓘ

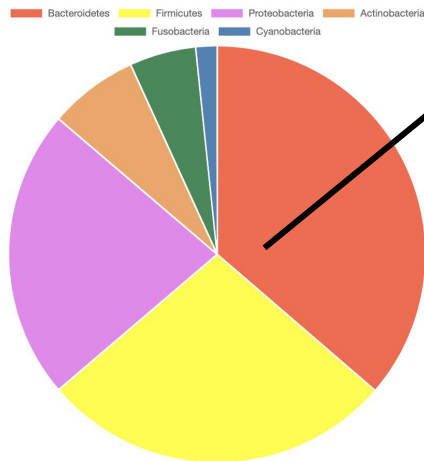
Merging small taxa with counts <
 Showing top n taxa, with n =

based on

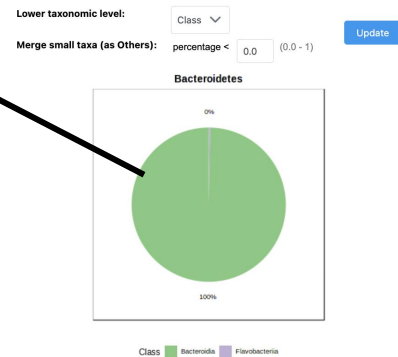
Did You Know?

If there are too many small taxa, use **Merging small taxa** with a high threshold for major pattern and clean legend. Or, you can use **Showing top taxa** to only show top number of taxa.

Click a section to view its lower-level compositions (except those Not_Assigned and Others taxa);



Click on it for projection to lower taxonomic level



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

Data source: Original Filtered

Group by: SampleType

Steps: 5 10 20

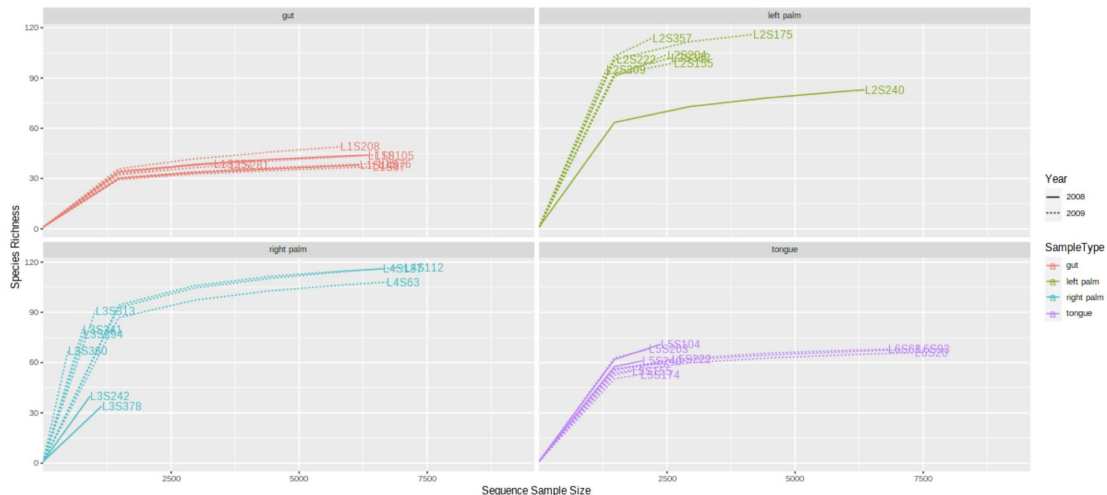
Color by: SampleType

Type by: Year

Update

Step determines the number of subsamples for generating rarefaction curve

Separate by multiple metadata variables



A. Visual Exploration

Phylogenetic tree:

Helps in determining evolutionary 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

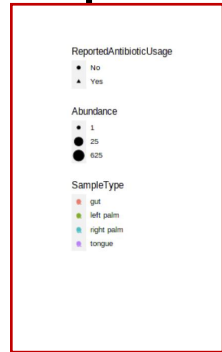
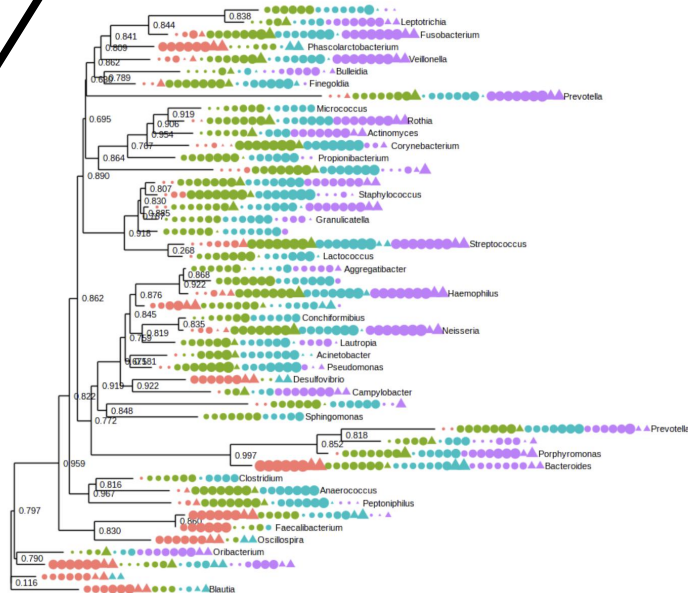
Color: SampleType

Shape: ReportedAntibioticUsage

Taxonomy level: Genus

Tree shape: Rectangular

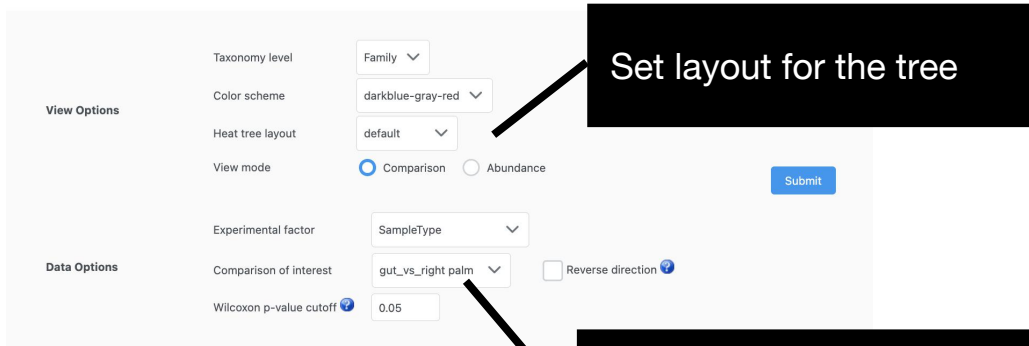
Submit



A. Visual Exploration

Heat tree:

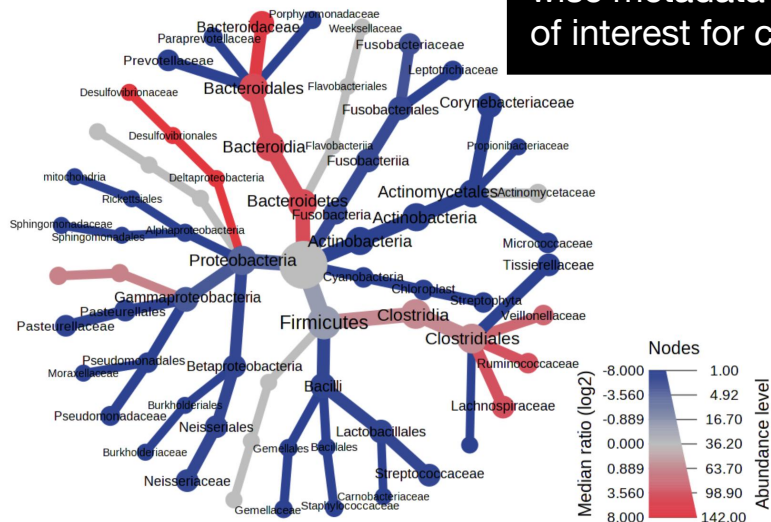
- A hierarchical tree of taxonomic levels with abundance indicated by colors.
- It presents abundance ratios of two groups at each taxonomic level
- It can compare every pair of factors in each metadata variable



Set layout for the tree

gut_vs_right palm

Need to specify the pairwise metadata variables of interest for comparison



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.

Data source Original Filtered

Taxonomy level Feature-level

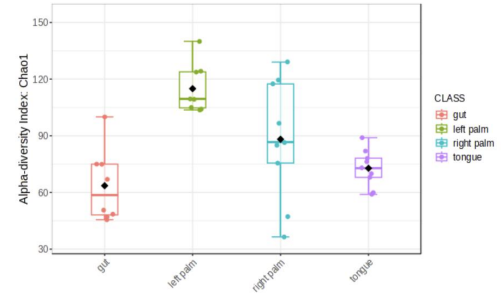
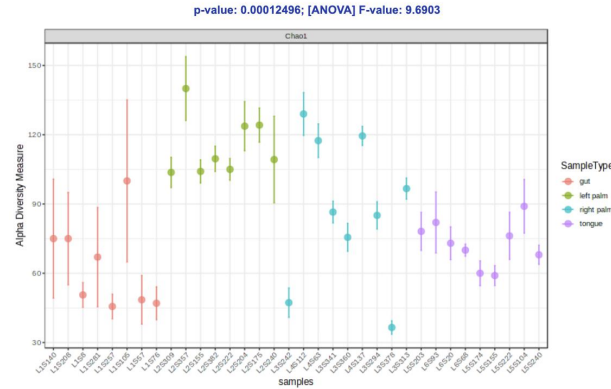
Experimental factor SampleType

Diversity measure Chao1

Statistical method T-test / ANOVA

Color options Default

Submit



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 unfrac distances

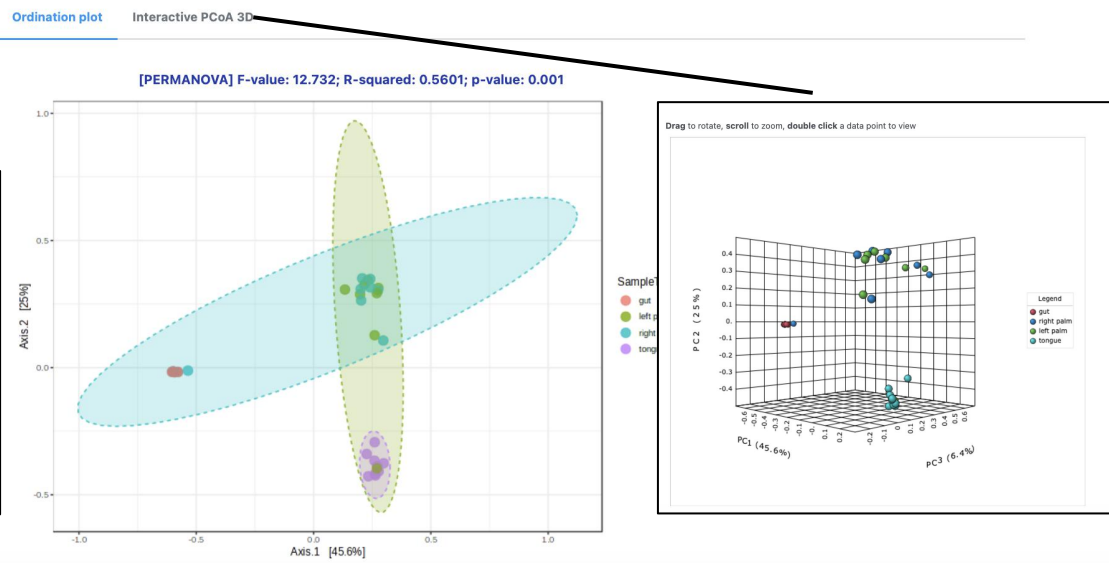
- Bray-Curtis Index
- Bray-Curtis Index
- Jensen-Shannon Divergence
- Jaccard Index
- Unweighted UniFrac Distance
- Weighted UniFrac Distance

Statistical methods to test the strength and statistical significance of sample groupings based on ordination based distances, including: ANOSIM, PERMANOVA, PERMDISP and MiRKAT.

Ordnation method: PCoA
Distance method: Bray-Curtis Index
Taxonomic level: Feature-level
Statistical method: PERMANOVA
Label samples by: None (2D plot only)

Color options: Default
Color by: Experimental factor (SampleType), Taxon abundance (Enter a feature ID), Alpha diversity (Chao1)

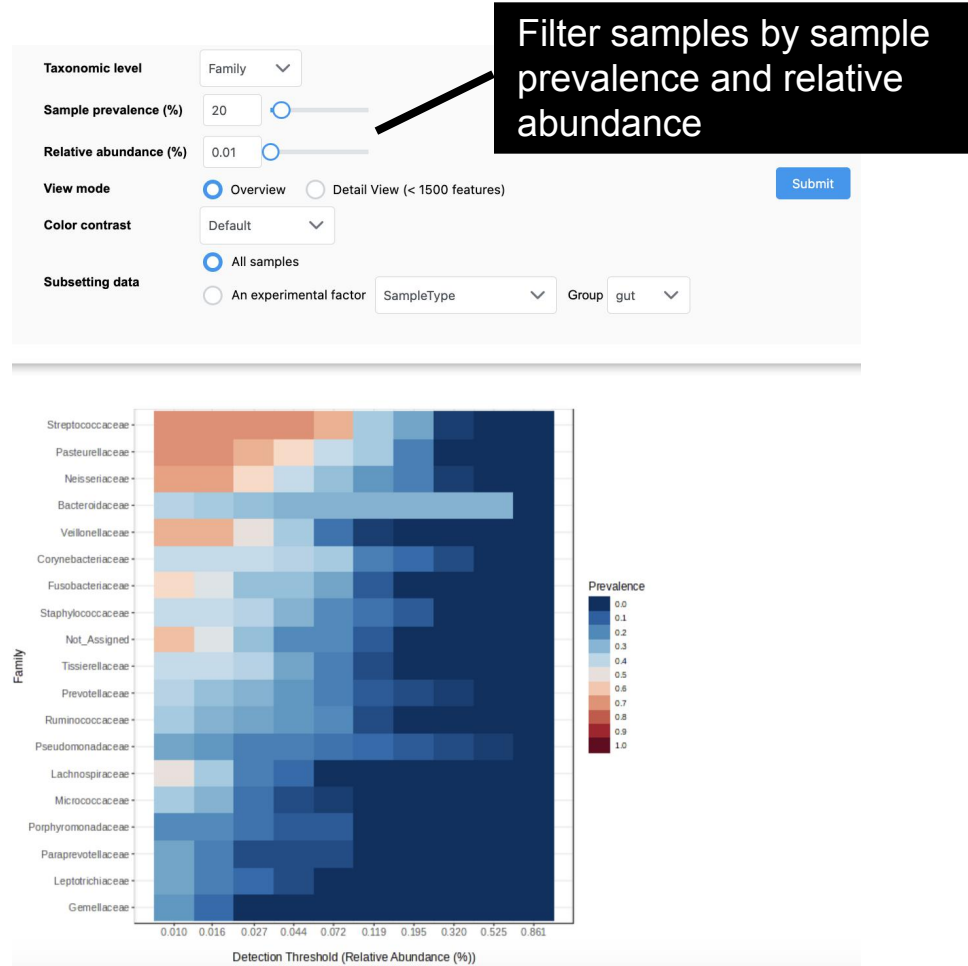
Update



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)

Taxonomy level Genus Prepend higher taxa

Data source: Normalized data

Standardization: Autoscale features

View mode Overview Detail View (< 1500 features)

Color contrast Default

Show feature names Column names Font size: 13
 Row names Font size: 8.8

Annotation bar Height: 3 % Font size: 11.8

Distance measure Euclidean

Clustering algorithm Ward

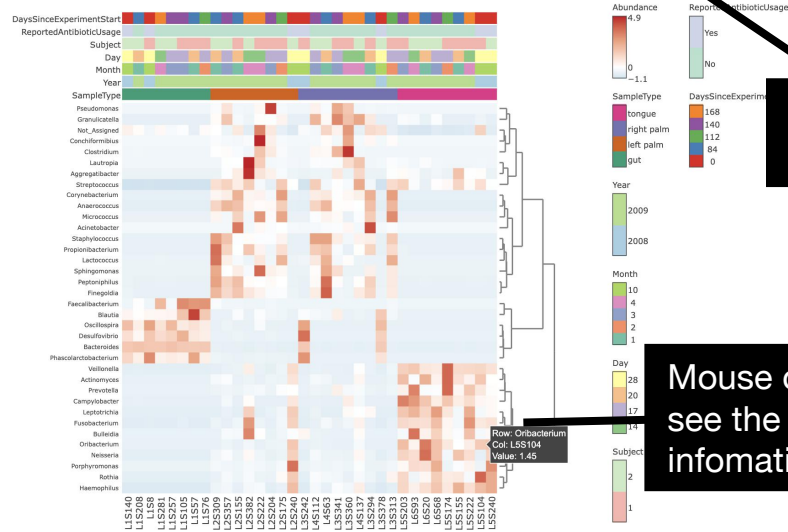
Cluster samples by Current clustering algorithm An experimental factor SampleType

Submit

See the heatmap in a new tab in detail view

Reset the heatmap

Mouse over to see the detail information



C. Clustering & Correlation Network

Dendrogram Analysis

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

Taxonomic level: Genus

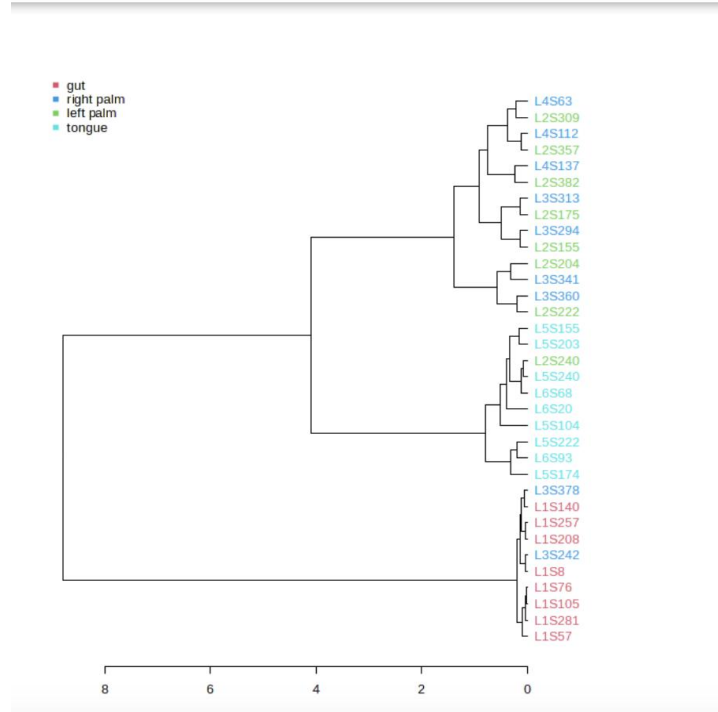
Distance measure: Bray-Curtis Index

Clustering algorithm: Ward

Experimental factor: SampleType

Color options: Default

Submit



C. Clustering & Correlation Network

Correlation Analysis

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

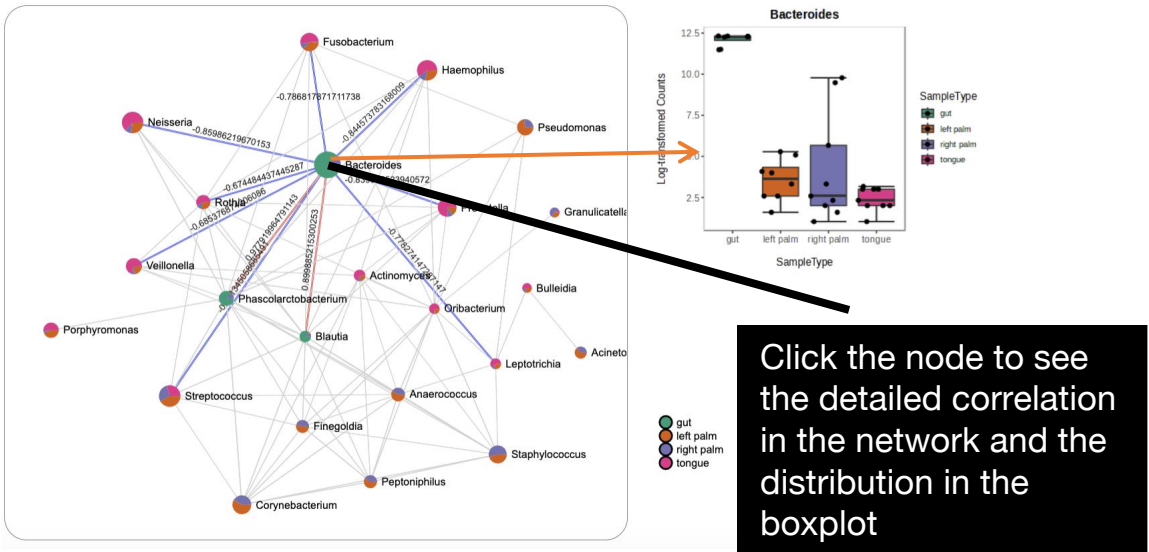
Seven statistical method are provided to calculate the correlation including SECOM (Pearson1), SECOM (Pearson2), SECOM (Distance), SparCC, Pearson, Spearman and Kendall.

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
Submit

You can zoom, drag or double click a node to get more details. The network, result table, summary plot and heatmap can be downloaded from the right panel.



Click the node to see the detailed correlation in the network and the distribution in the boxplot

C. Clustering & Correlation Network

Define your own pattern of interest

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.

Taxonomy level: Genus prepend higher taxa

Distance measure: Pearson r

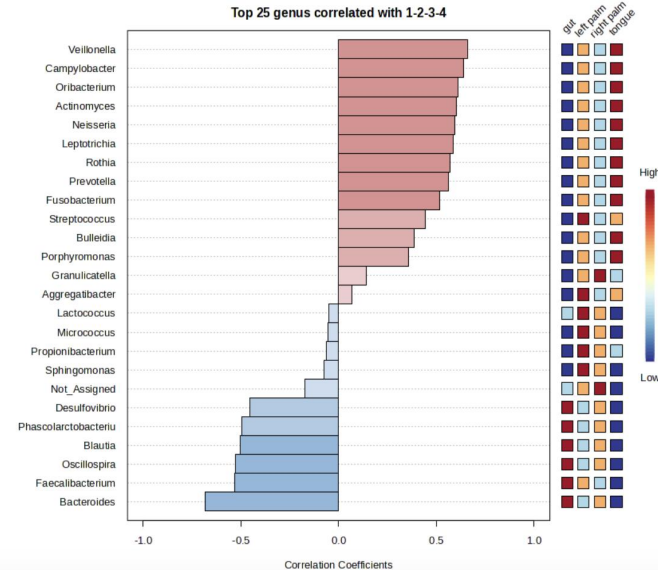
Experimental factor: SampleType

Pattern based on: Predefined 1-2-3-4 A feature Custom profile

Submit

Graphical Summary Result Table

Check the statistical result here.



D. Comparison & Classification

Single-factor analysis

Select the metadata of interest

Select statistical methods

Taxonomy level: Genus

Experimental factor: SampleType

Statistical method: T-test/ANOVA

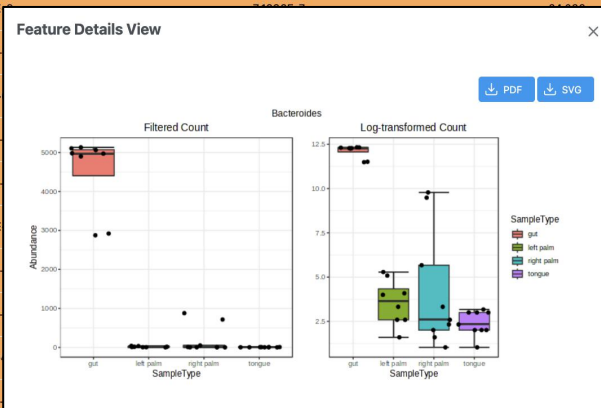
Adjusted p-value cutoff: 0.05

Submit

- T-test/ANOVA
- Mann-Whitney/Kruskal-Wallis
- T-test/ANOVA
- metagenomeSeq (0-inflated)
- metagenomeSeq (fitFeature)
- EdgeR
- DESeq2

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

| Name | Pvalues | FDR | Statistics | View |
|------------------|------------|-----------|------------|------|
| Bacteroides | 2.3456E-10 | 8.4443E-9 | 38.112 | |
| Haemophilus | 3.9998E-10 | | | |
| Veillonella | 1.5777E-10 | | | |
| Campylobacter | 1.7651E-10 | | | |
| Oribacterium | 2.1652E-10 | | | |
| Actinomyces | 2.7252E-10 | | | |
| Neisseria | 4.8344E-10 | | | |
| Anaerococcus | 8.0901E-10 | | | |
| Leptotrichia | 1.0547E-09 | | | |
| Oscillospira | 1.1851E-09 | | | |
| Faecalibacterium | 3.1344E-09 | | | |
| Fusobacterium | 3.7243E-04 | 0.0011046 | 8.2627 | |



D. Comparison & Classification

Multi-factor analysis: Model Parameters

The screenshot shows a web interface for a statistical tool. The breadcrumb navigation at the top reads: Home > Data Upload > Data Inspection > Data Filter > Normalization > Analysis Overview > Multiple Regression > Downloads. A 'Navigate to:' dropdown is in the top right, and a 'Show Info Pane' link is below it.

Multiple Linear Regression with Covariate Adjustment

This tool uses general linear models to find associations between microbial features and metadata. The model includes the primary metadata, covariate, and blocking factor variables.

- Primary metadata:** included as a 'fixed effect' in the model. Statistical tests are performed for each primary metadata variable. If the primary metadata variable has more than two 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.

Form Fields:

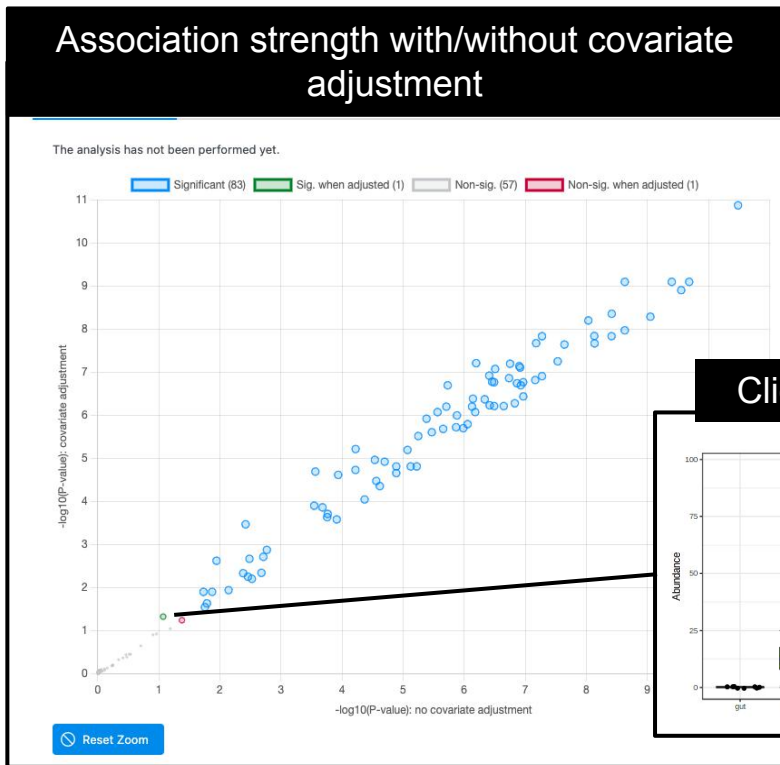
- Taxonomy level:** Feature-level (dropdown)
- Primary metadata:** SampleType (dropdown)
- Comparison:** tongue vs. gut (dropdowns)
- Covariates (control for):** Subject, ReportedAntibioticUsage (tags with dropdown)
- Blocking factor:** -- Unspecified -- (dropdown)
- Adjusted p-value cutoff:** 0.05 (input field)
- Submit** (button)

Callout Boxes:

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

D. Comparison & Classification

Multi-factor analysis: Results

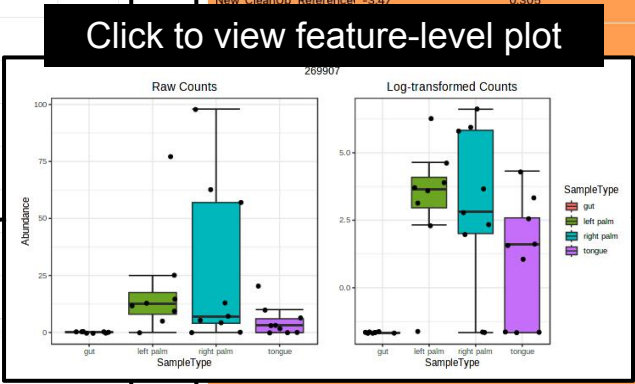


Statistical results sorted by p-value

Graphical Summary [Results Table](#)

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

| Name ↑↓ | Log2FC ↑↓ | St.Error ↑↓ | P-value ↑↓ | FDR ↑↓ | View |
|-----------------------|-----------|-------------|------------|----------|----------------------|
| 1078207 | 8.62 | 0.617 | 3.82E-14 | 1.36E-11 | View |
| 365628 | -3.84 | 0.341 | 6.84E-12 | 8.1E-10 | View |
| 968675 | 9.64 | 0.851 | 5.78E-12 | 8.1E-10 | View |
| New_CleanUp_Reference | -3.47 | 0.305 | 5.21E-12 | 8.1E-10 | View |
| | | | 1.42E-11 | 1.26E-9 | View |
| | | | 6.28E-11 | 4.46E-9 | View |
| | | | 8.08E-11 | 5.21E-9 | View |
| | | | 1.08E-10 | 6.38E-9 | View |
| | | | 2.12E-10 | 1.08E-8 | View |
| | | | 3.09E-10 | 1.46E-8 | View |
| | | | 3.4E-10 | 1.48E-8 | View |
| | | | 3.55E-10 | 1.48E-8 | View |



D. Comparison & Classification

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)

Taxonomy level: Genus

Experimental factor: SampleType

P-value cutoff: 0.1 Original FDR-adjusted

Log LDA score: 2.0

Submit

Graphical Summary

Result Table

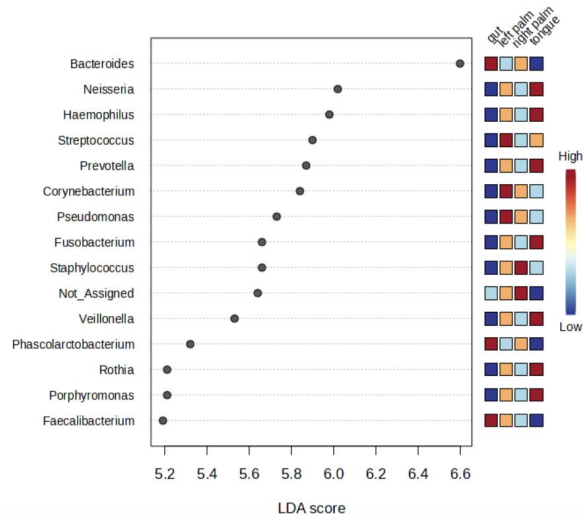
Graphical output

Dot Plot

Update

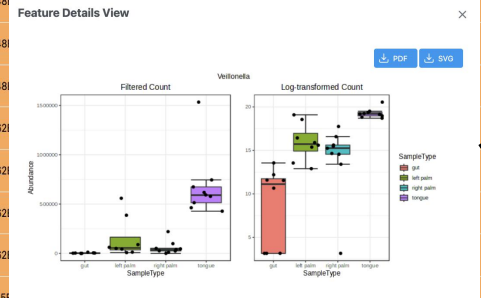
Number of top features

15



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 | 5.06 | |
| Haemophilus | 1.5584E-5 | 1.6548 | | | | 1907000.0 | 5.98 | |
| Prevotella | 1.8387E-5 | 1.6548 | | | | 1492000.0 | 5.87 | |
| Oribacterium | 2.7817E-5 | 1.6662 | | | | 55621.0 | 4.43 | |
| Actinomyces | 2.9274E-5 | 1.6662 | | | | 115620.0 | 4.76 | |
| Fingoldia | 3.3527E-5 | 1.6662 | | | | 151.19 | 4.92 | |
| Fusobacterium | 3.7026E-5 | 1.6662 | | | | 926560.0 | 5.66 | |
| Blautia | 4.7001E-5 | 1.7255E-4 | | | | 0.0 | 4.62 | |
| Lactococcus | 4.7932E-5 | 1.7255E-4 | 342.84 | 152360.0 | 108150.0 | 0.0 | 4.88 | |



D. Comparison & Classification

Taxonomy level: Genus

Experimental factor: SampleType

Choose metadata for predictors:

Number of trees to grow: 500

Number of predictors to try: 7

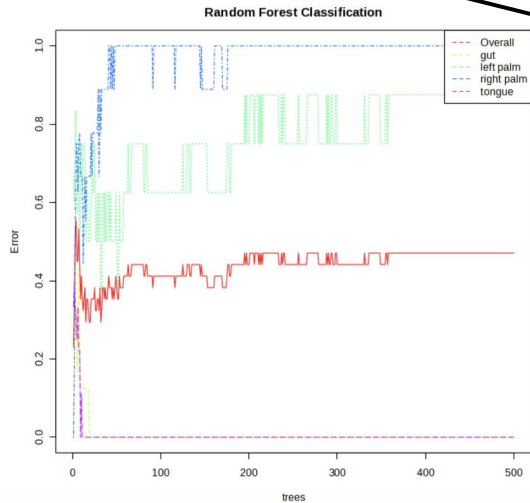
Randomness setting: On

Submit

No. of trees to be used for classification

No. of predictors for each node

Classification Performance



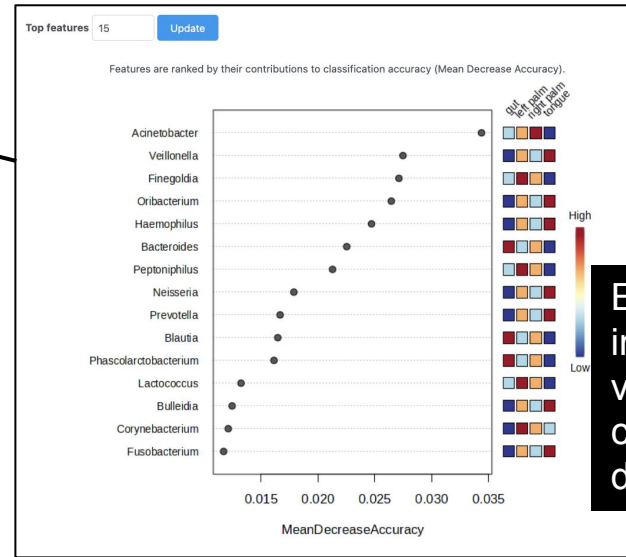
The OOB error is 0.471

| | gut | left palm | right palm | tongue | class_error |
|------------|-----|-----------|------------|--------|-------------|
| gut | 8 | 0 | 0.0 | 0.0 | 0.0 |
| left palm | 0 | 1 | 6.0 | 1.0 | 0.875 |
| right palm | 2 | 7 | 0.0 | 0.0 | 1.0 |
| tongue | 0 | 0 | 0.0 | 9.0 | 0.0 |

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.



Estimates of important variables in the classification of data.

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 16SrRNA 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

QIIME against SILVA database

SILVAngs

Predict Functional Potential

Tax4Fun need SILVA taxonomy annotation and specify the pipeline here

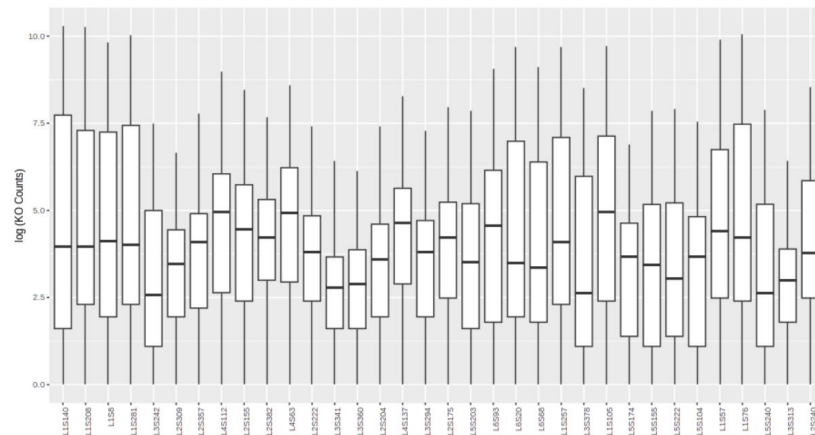
Predicting functional capabilities of microbial communities using Tax4Fun2

[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 sequence is included in the OTU/ASV table.

Predict Functional Potential

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

Result figure:



The KO table and figures can be downloaded 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