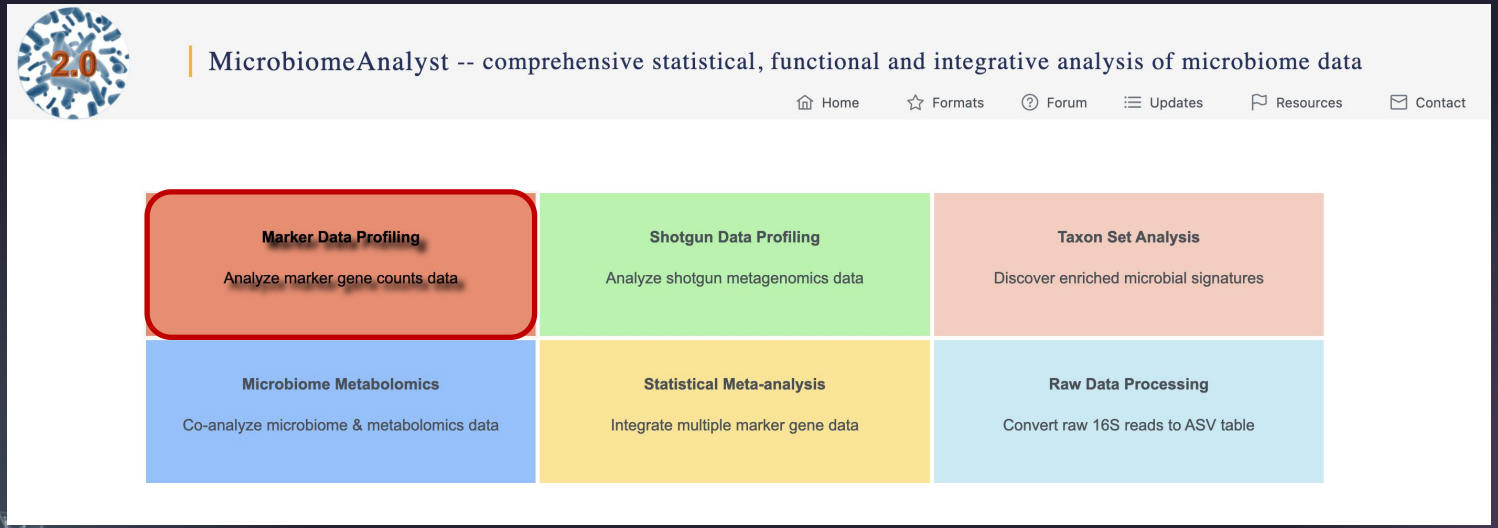
MicrobiomeAnalyst 2.0

Comprehensive statistical, functional and integrative analysis of microbiome data

xialab@mcgill 2023-Feb-28

Tutorial for Marker Data Profiling



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.

Ehanced Features in Version 2.0

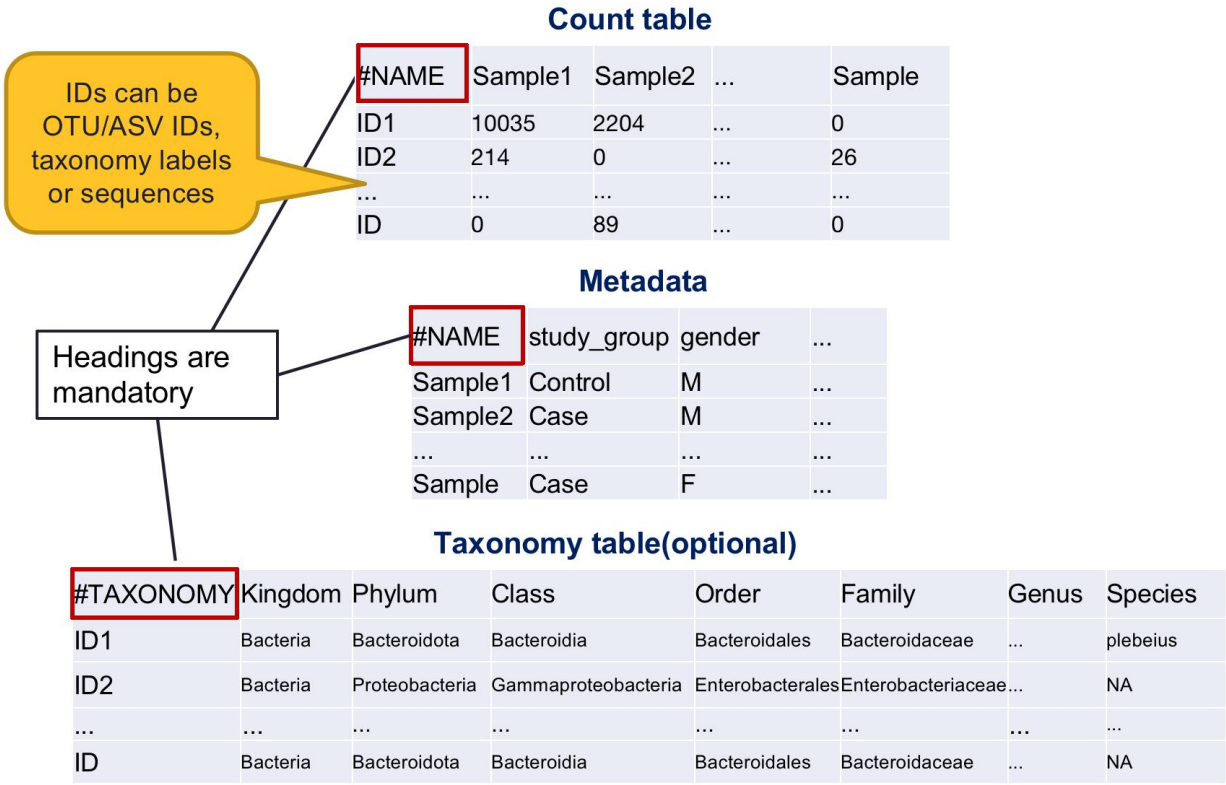
- Editable metadata and multi-factor comparison analysis
- Deal with the normalized input data
- Update the methods for correlation analysis
- Update Statistical methods for significance testing in beta-diversity profiling
- Add Tax4Fun2 for function prediction and update the background database
- Enhanced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

Analysis Strategies

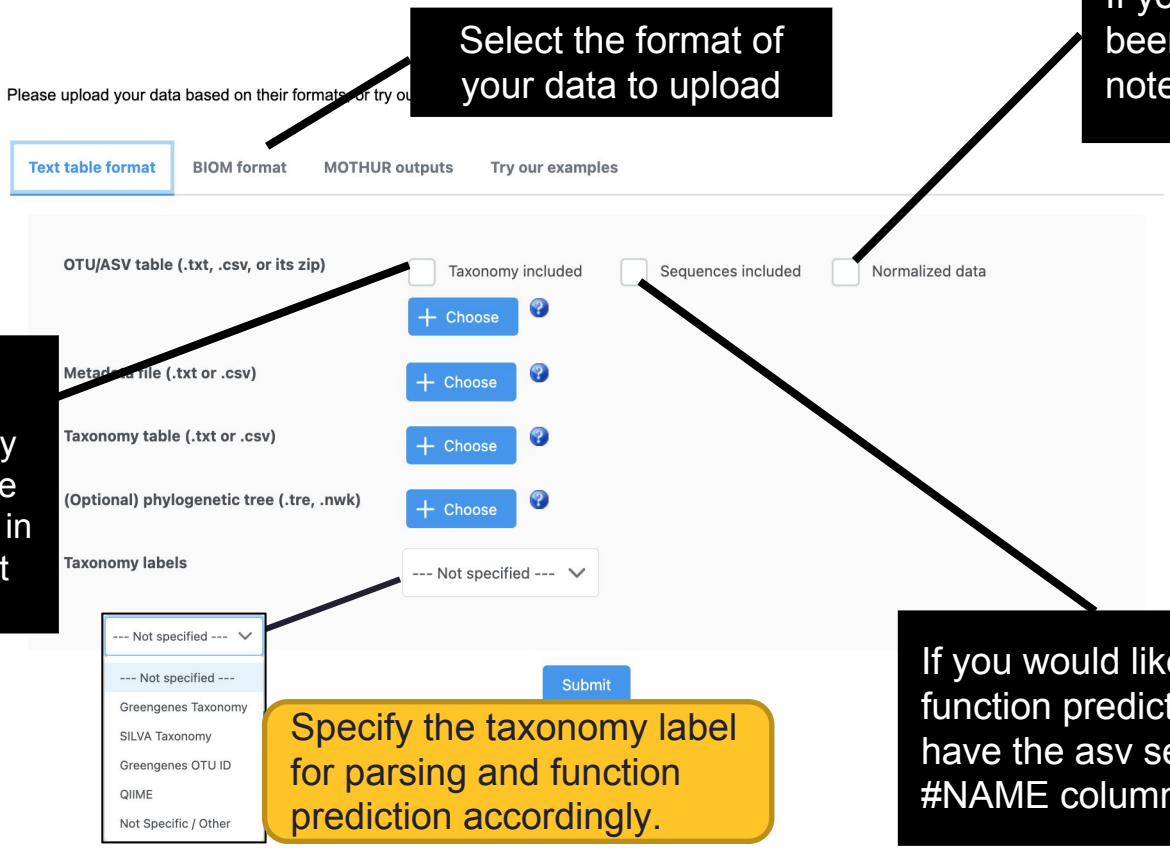
Visual Exploration	Community Profiling	Clustering & Correlation	Comparison & Classification	Functional Prediction
• Interactive stack bar/area plot • Interactive pie chart • Rarefaction curve • Phylogenetic tree • Heat tree	• Alpha diversity • Beta diversity • Core microbiome	• Interactive Heatmap • Dendrogram • Correlation network • Pattern search	• Single-factor analysis • Multi-factor analysis • LEfSe • Random Forest	• 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



Data Upload



Select the format of your data to upload

If your data has already been normalized, read the notes below 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 you would like use Tax4Fun2 for function prediction, make sure to have the asv sequences in the #NAME column 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.

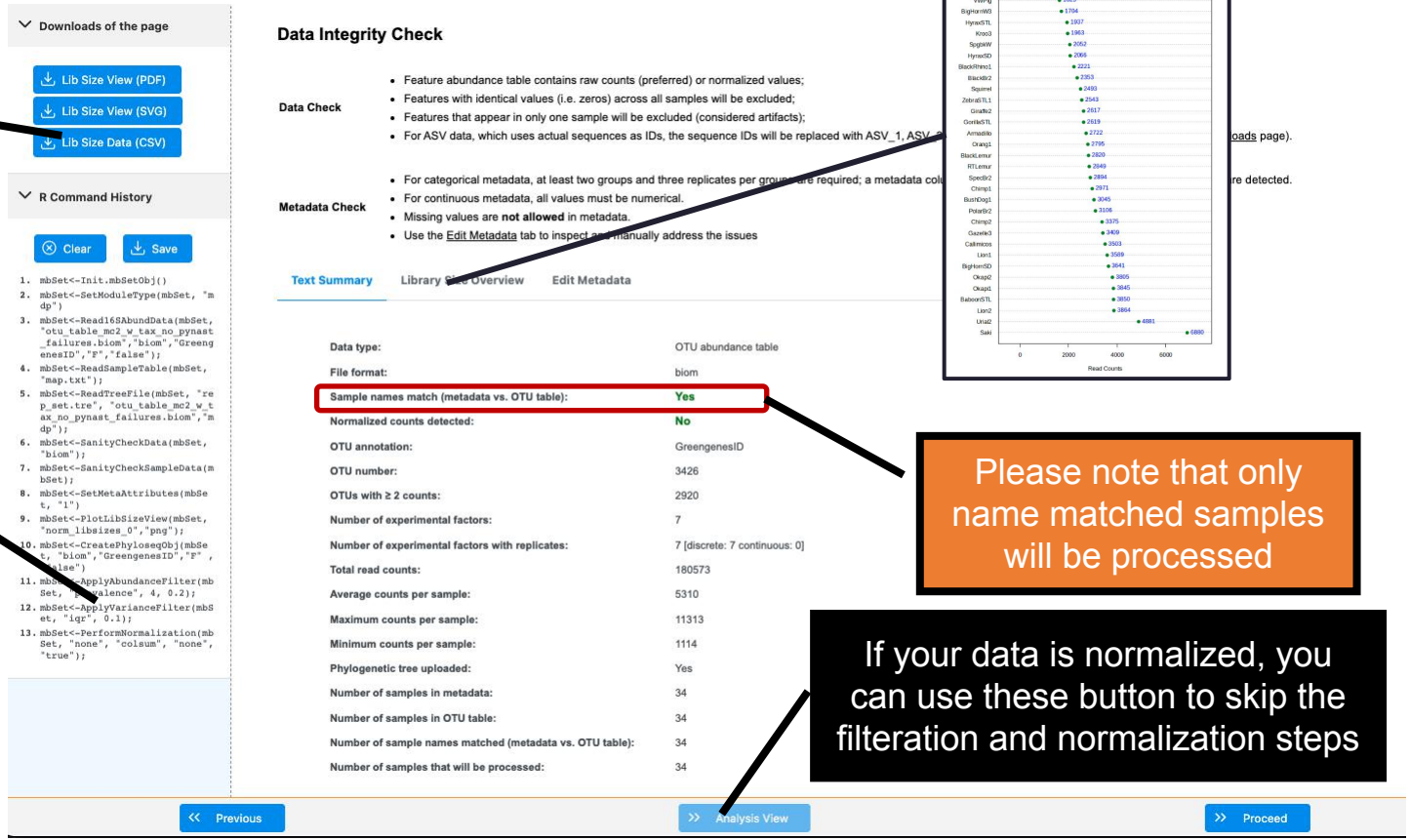
Data Integrity Check

Available file downloads for each page are displayed here

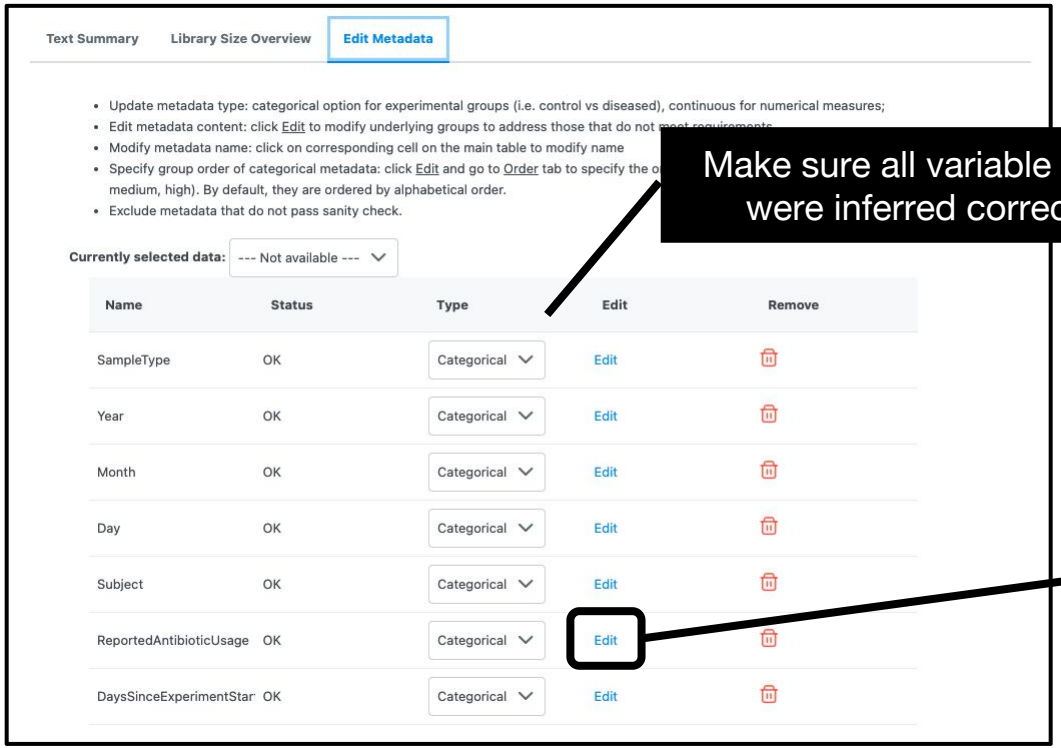
R commands are shown here

Please note that only name matched samples will be processed

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

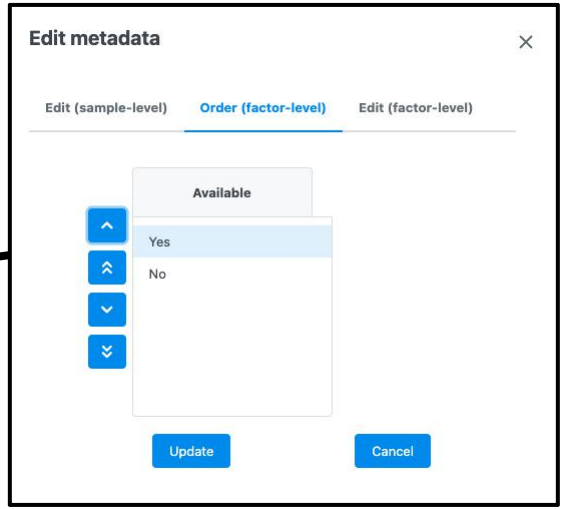


Edit Metadata

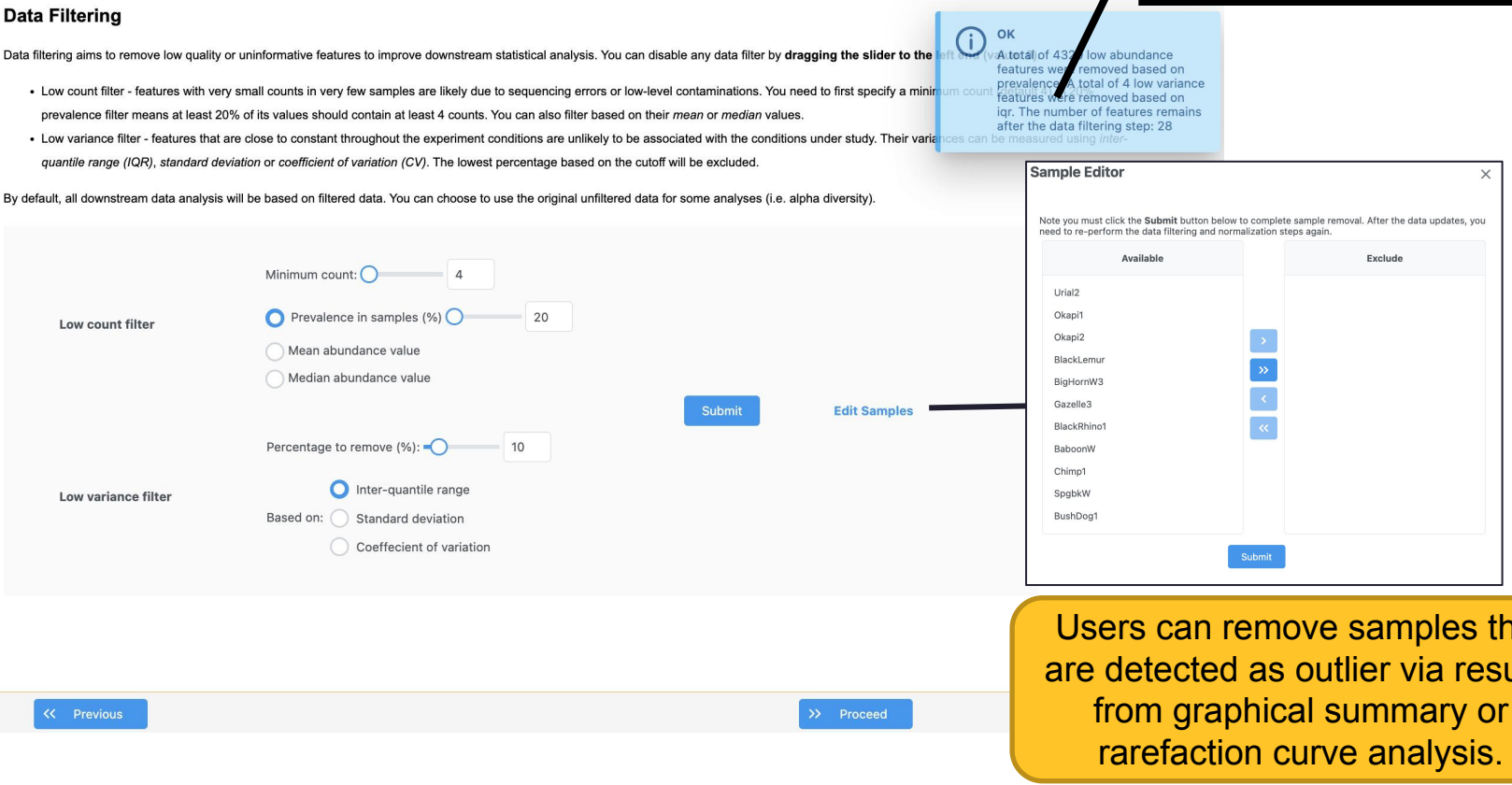


Make sure all variable types were inferred correctly

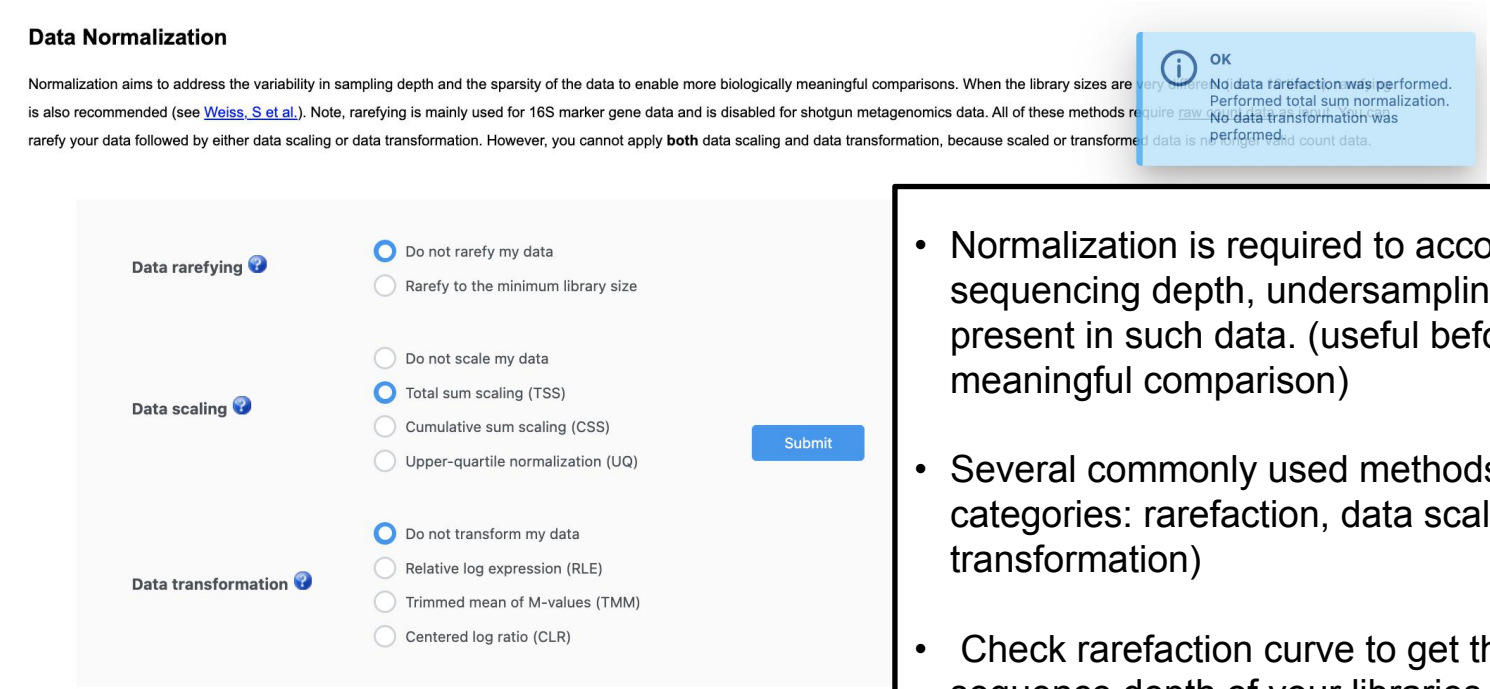
For categorical, adjust ‘Order’ to control order in downstream plots and analysis



Data Filtering

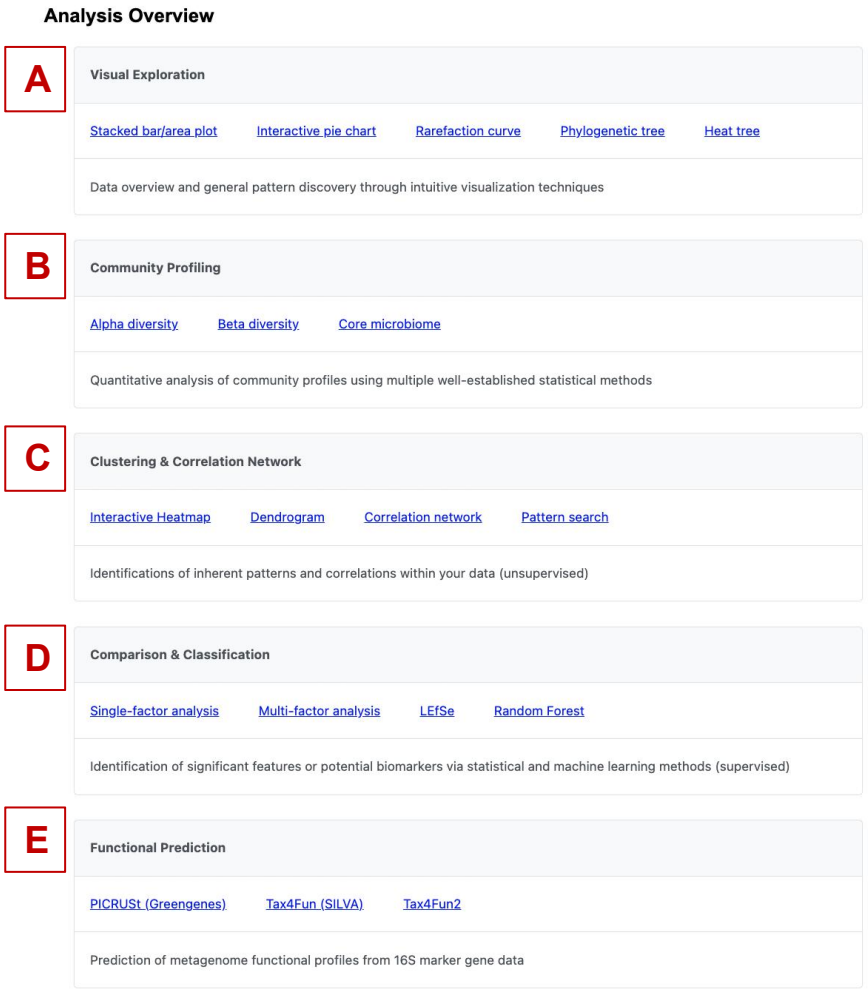


Data Normalization



- 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

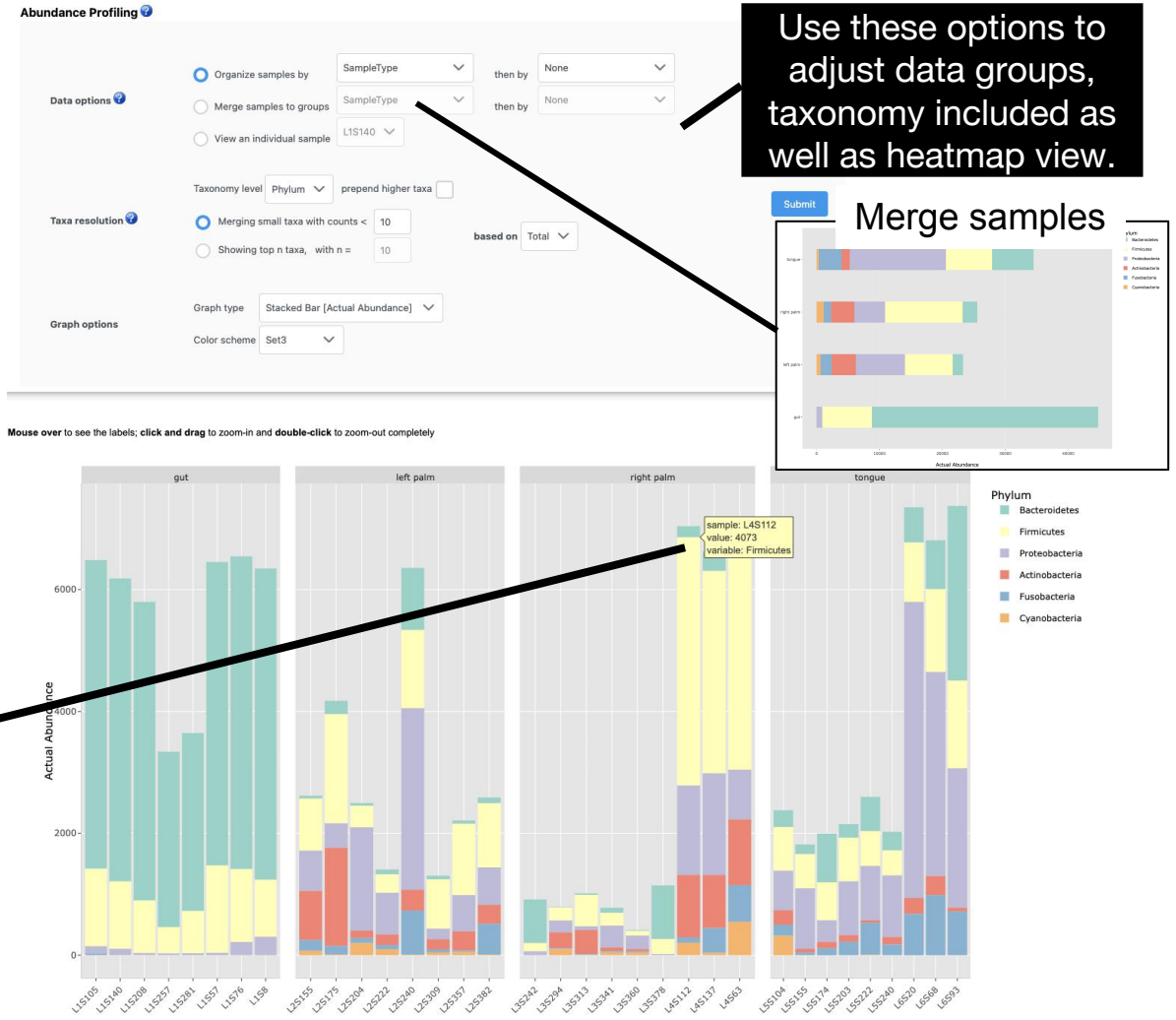


A. Visual Exploration

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.

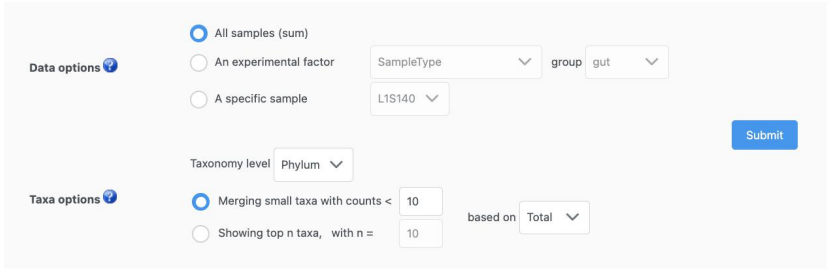
Mouse over to see the detail infomation

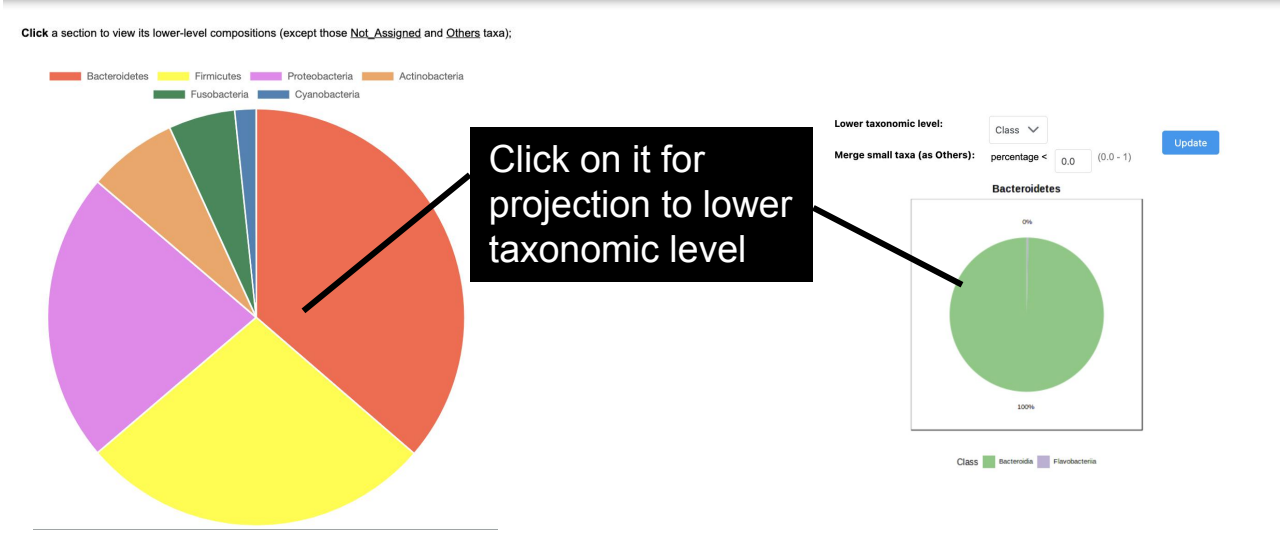


A. Visual Exploration

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

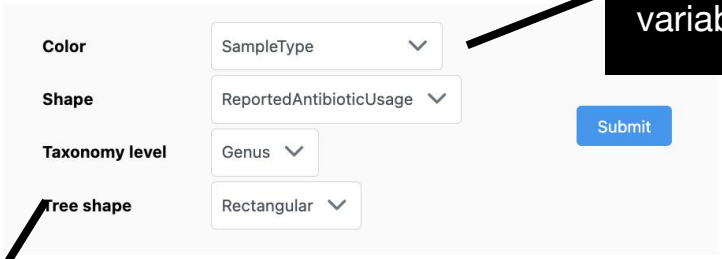
A. Visual Exploration

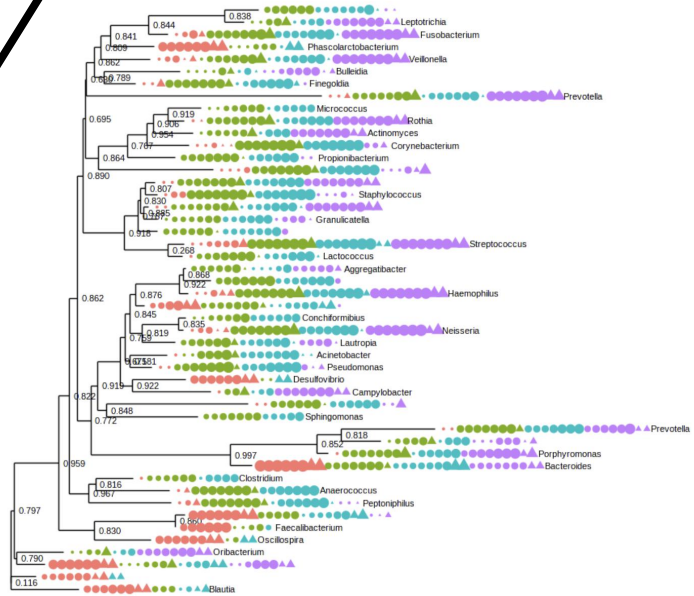
Phylogenetic tree:

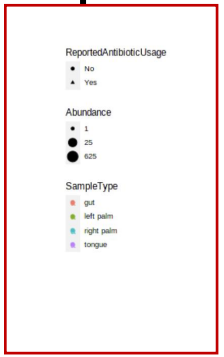
Helps in determining evolutional relations among different taxonomic groups at different levels.

Two types of tree shapes are provided: Rectangular and Redial

Select multiple metadata variables to customize tree



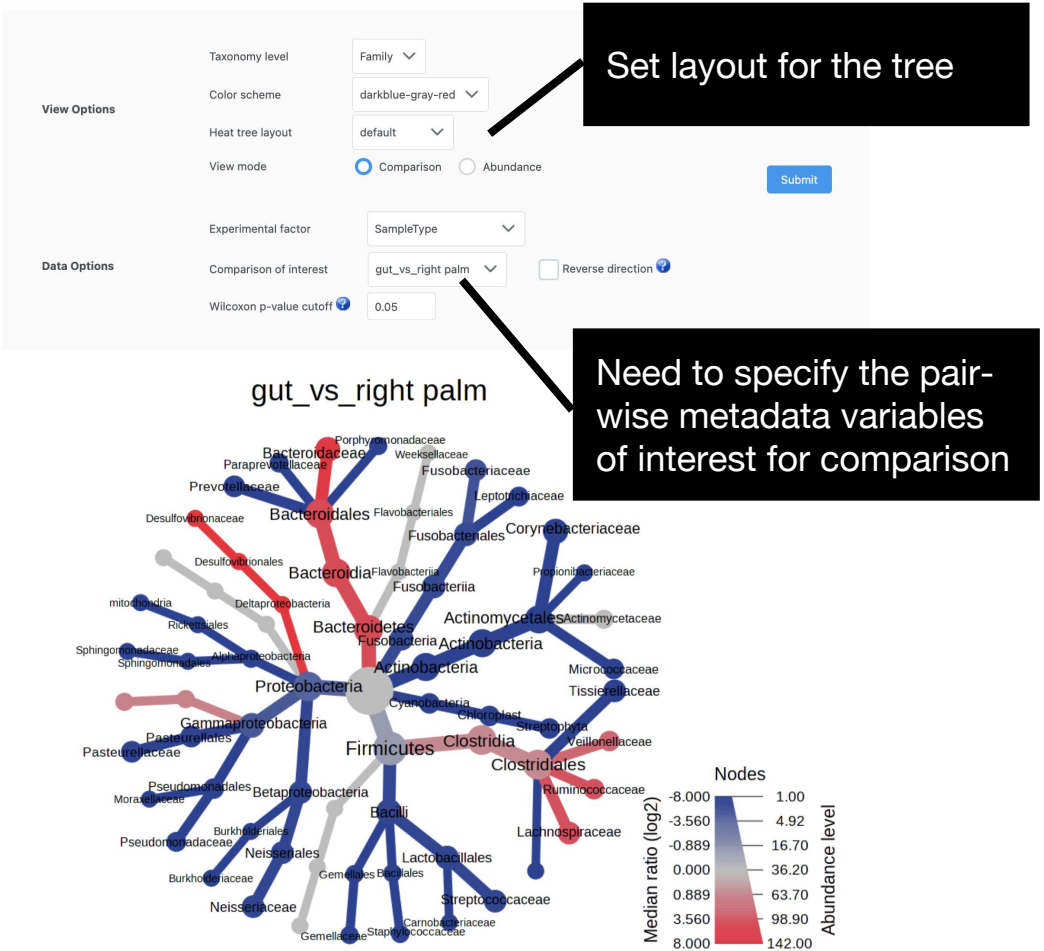




A. Visual Exploration

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

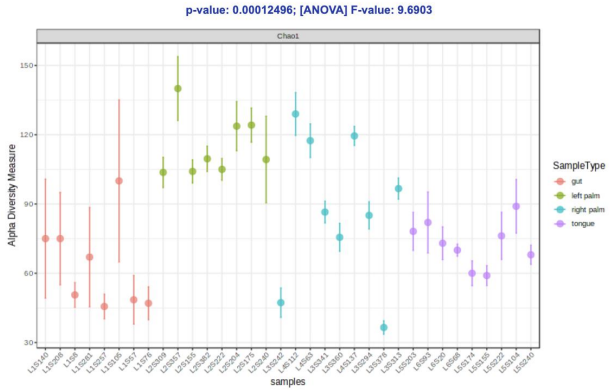


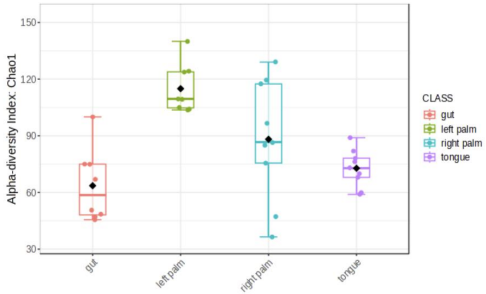
B. Community Profiling

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.





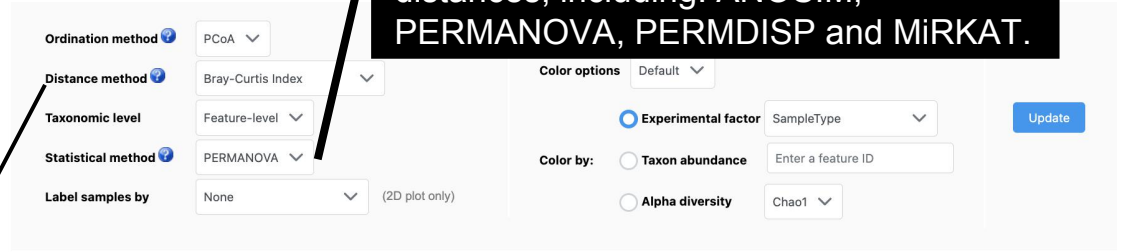
B. Community Profiling

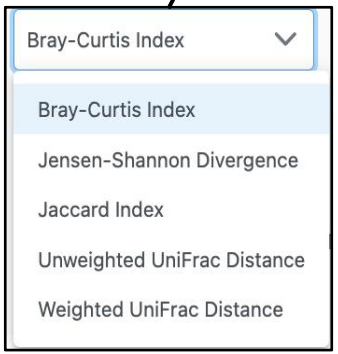
Beta diversity profiling:

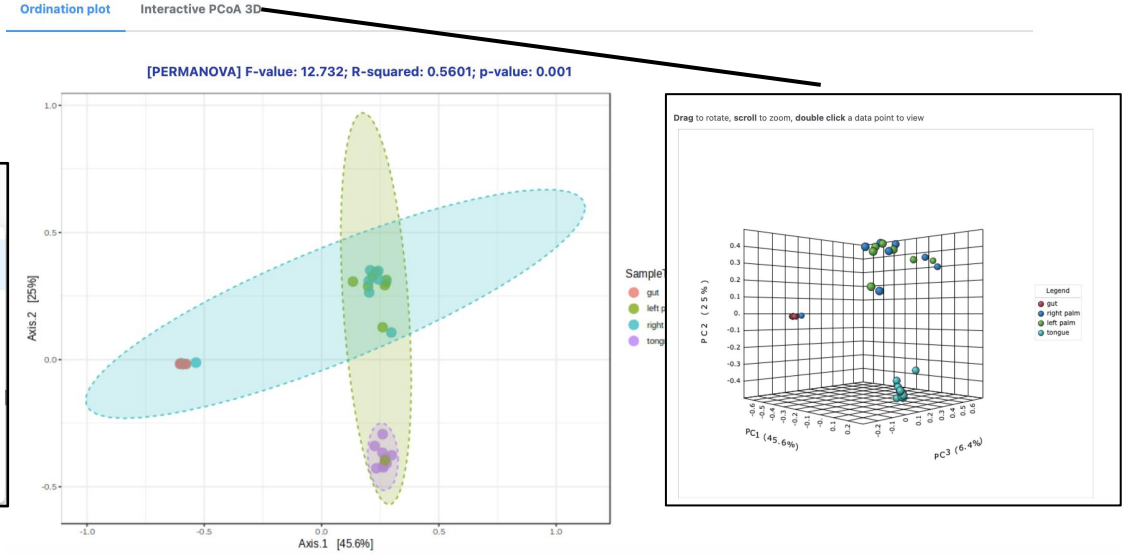
- Assess the differences between microbial communities(or samples)
- Visualize using PCoA (Principal Coordinate Analysis) or NMDS (Nonmetric Multidimensional Scaling)

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

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



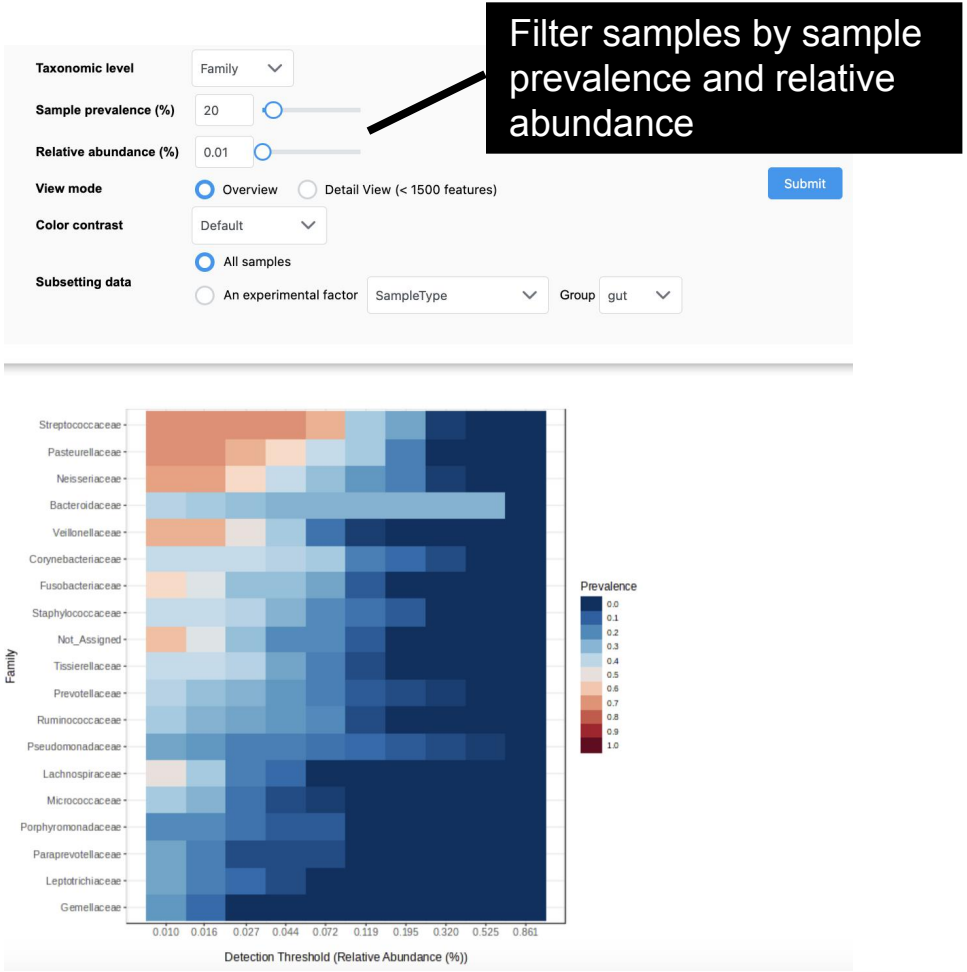




B. Community Profiling

Core microbiome analysis:

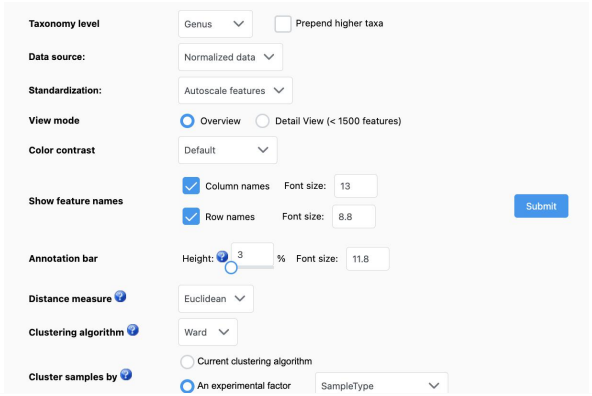
Helps in identifying core taxa or features that remain unchanged in their composition across different sample groups based on sample prevalence and relative abundance.

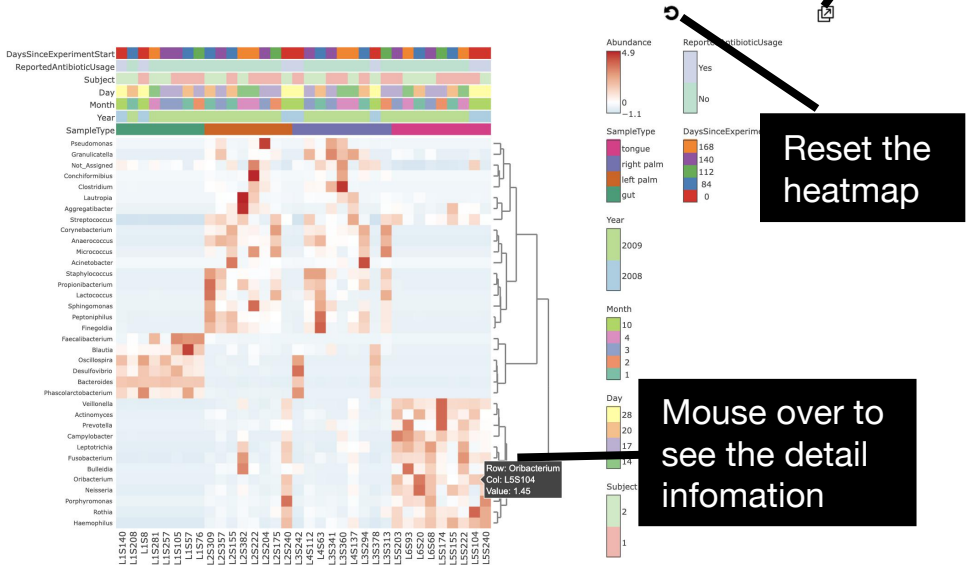


C. Clustering & Correlation Network

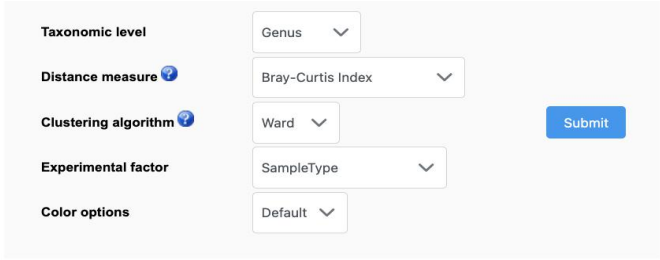
Clustering Heatmap Visualization:

- 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)



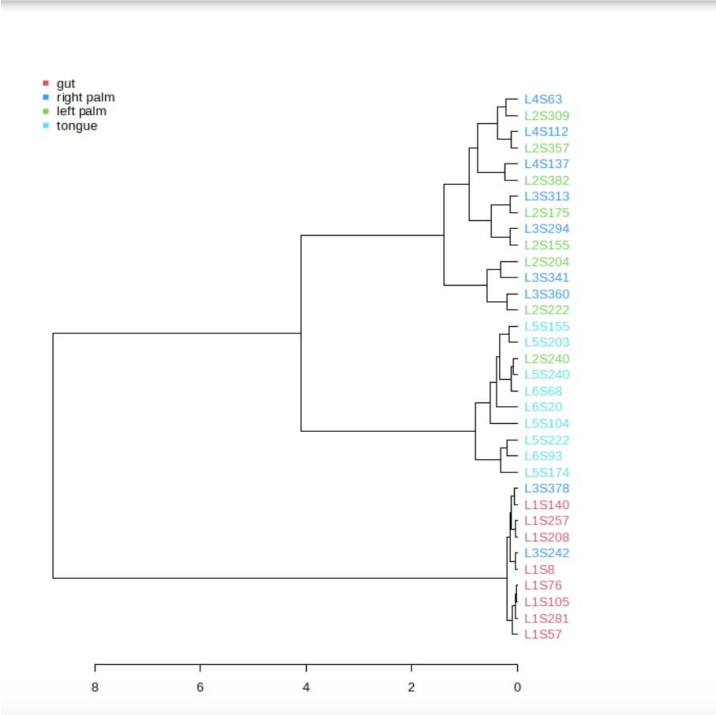


C. Clustering & Correlation Network



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

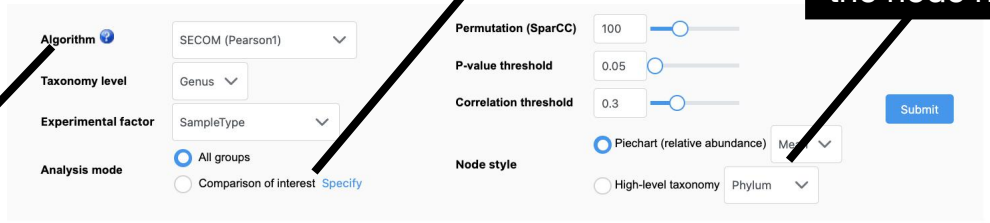
Correlation Analysis

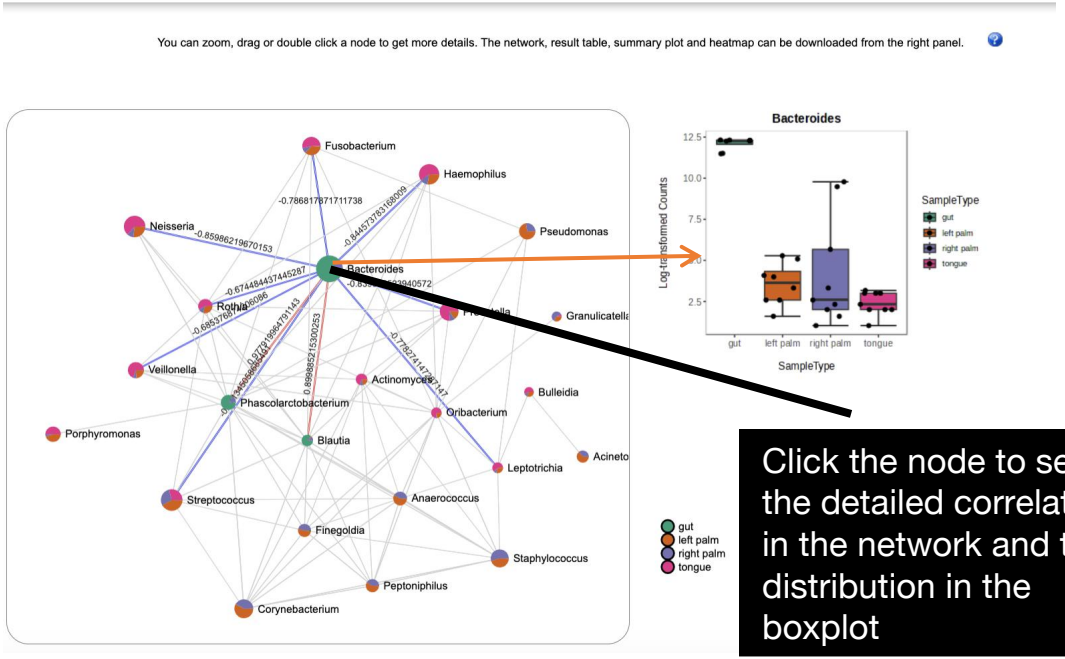
To identify biologically meaningful relationship or associations between taxa or features.

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



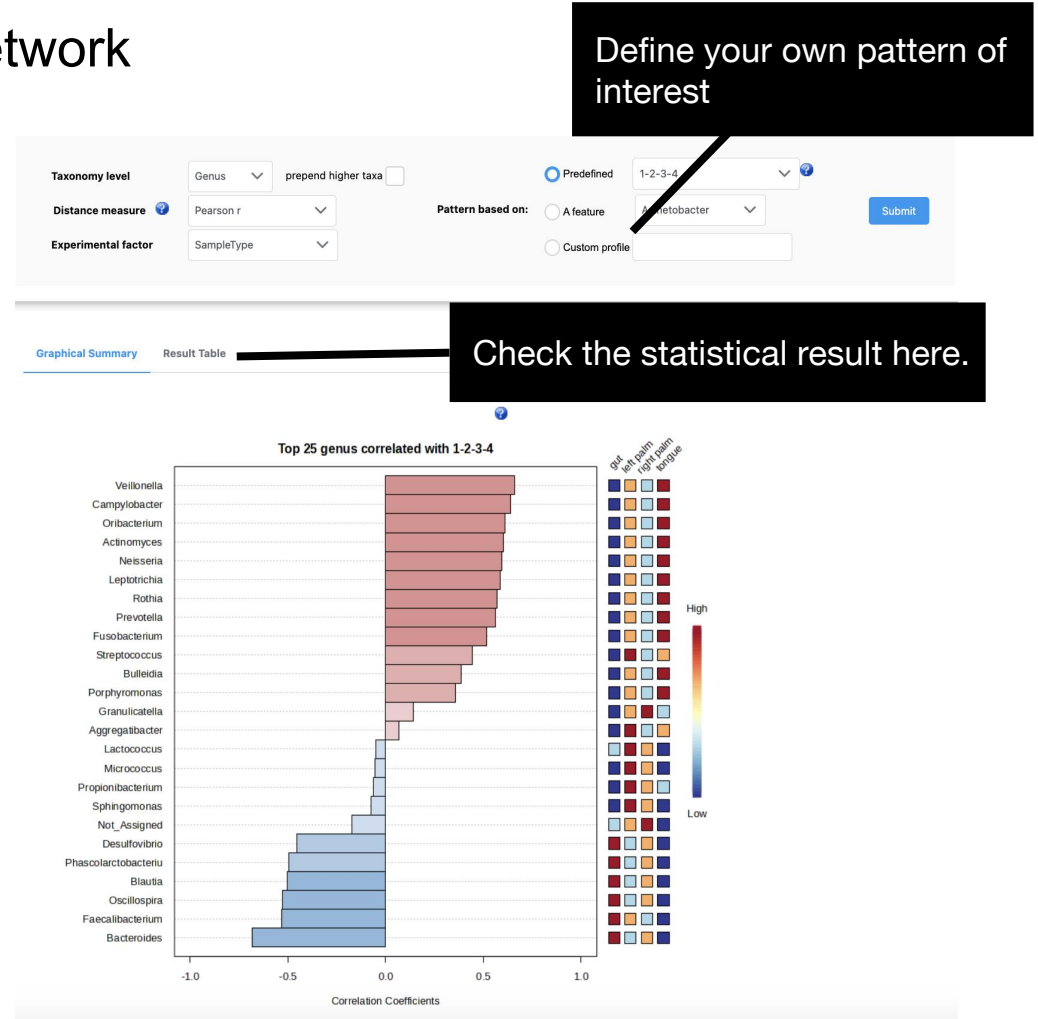


Click the node to see the detailed correlation in the network and the distribution in the boxplot

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.

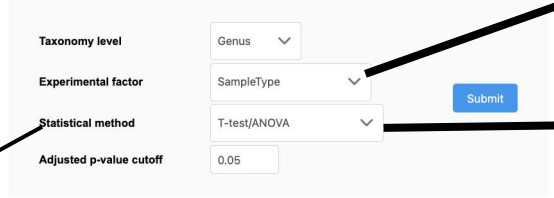


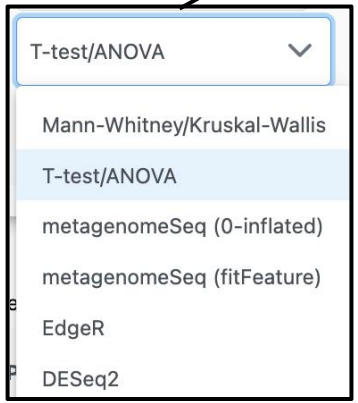
D. Comparison & Classification

Single-factor analysis

Select the metadata of interest

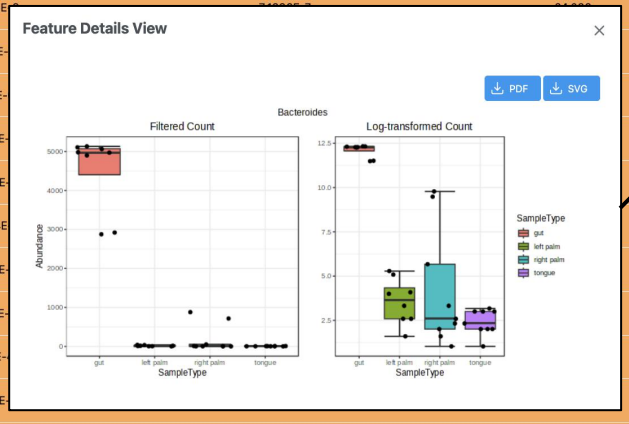
Select statistical methods





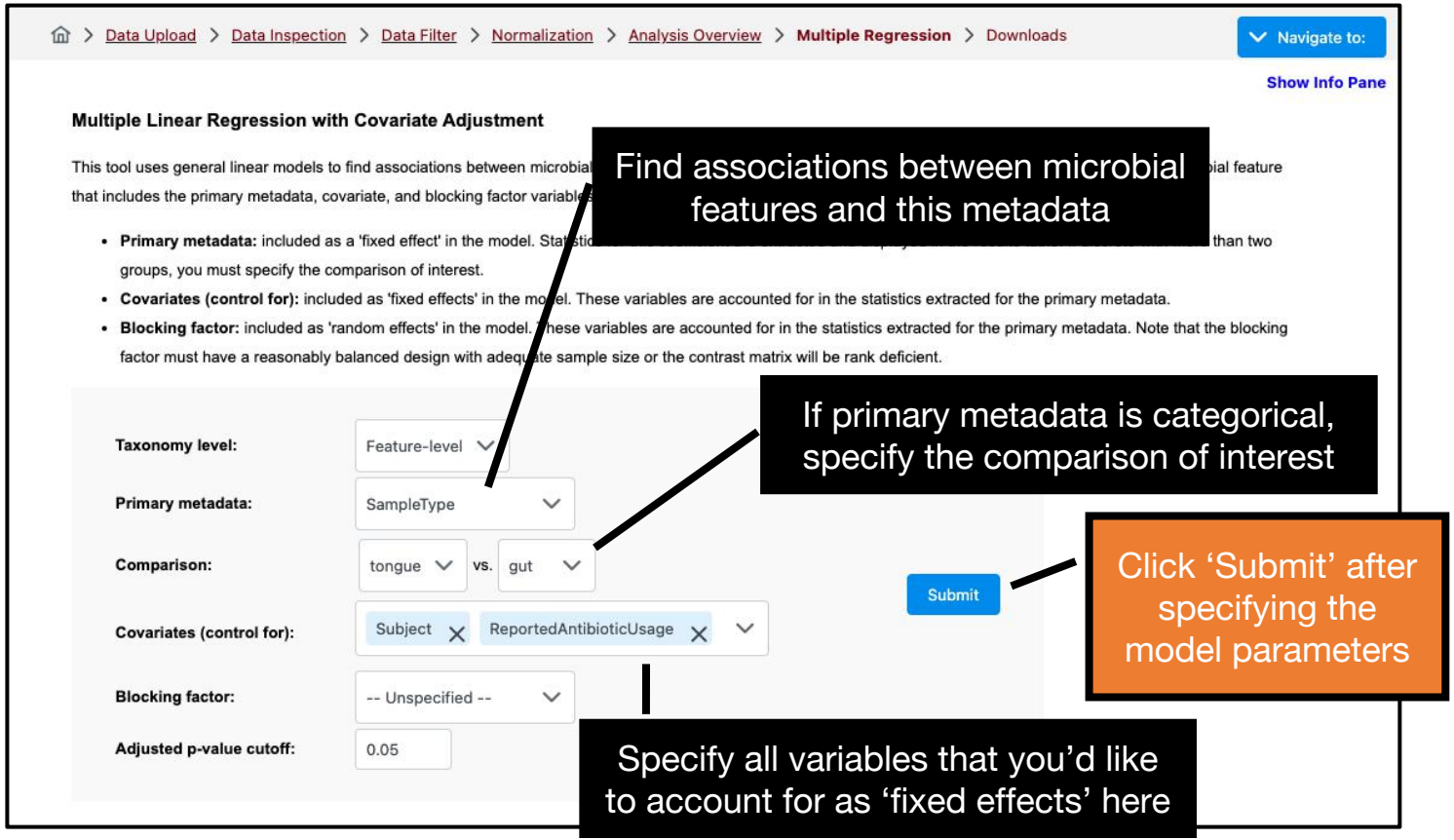
The table below shows at most 500 features ranked by their p values, with significant features highlighted in orange.

Name	Pvalues	FDR	Statistics	View
Bacteroides	2.3456E-10	8.4443E-9	38.112	
Haemophilus	3.9998E-			
Veillonella	1.5777E-			
Campylobacter	1.7651E-			
Oribacterium	2.1652E-			
Actinomyces	2.7252E-			
Neisseria	4.8344E			
Anaerococcus	8.0901E-			
Leptotrichia	1.0547E-			
Oscillospira	1.1851E-			
Faecalibacterium	3.1344E-			
Fusobacterium	3.7243E-4	0.0011046	8.2627	



D. Comparison & Classification

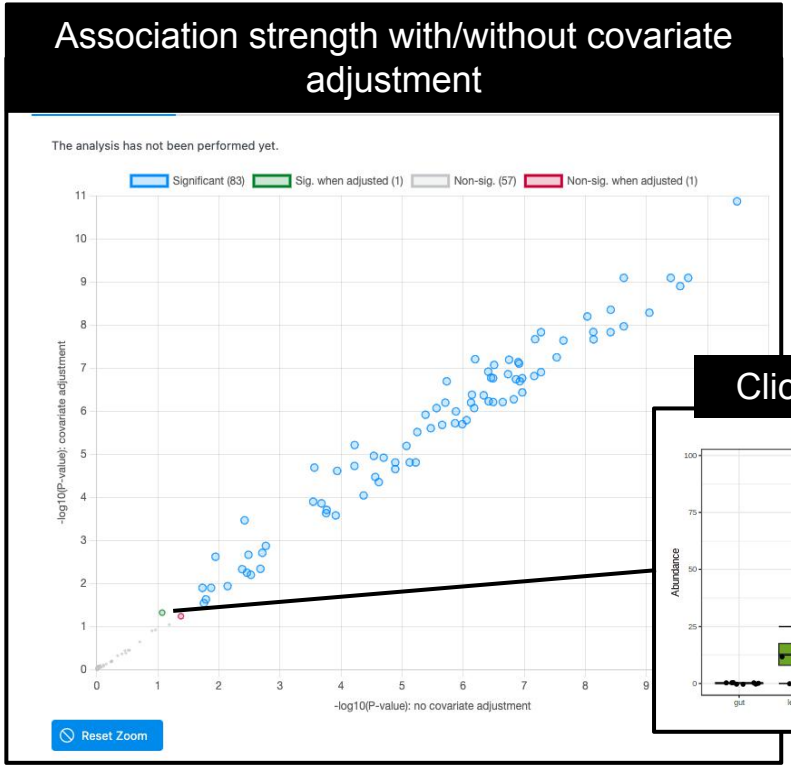
Multi-factor analysis:
Model Parameters



D. Comparison & Classification

Multi-factor analysis: Results

Association strength with/without covariate adjustment



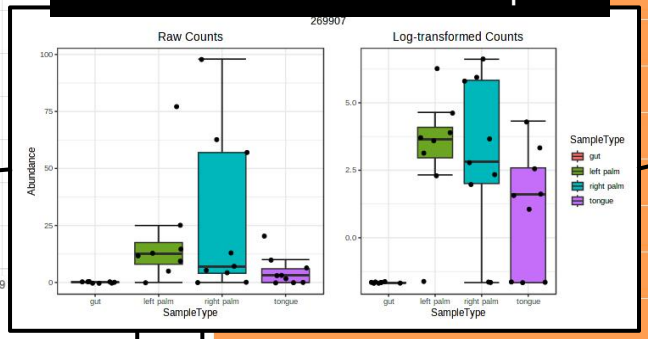
Statistical results sorted by p-value

Graphical Summary | Results Table

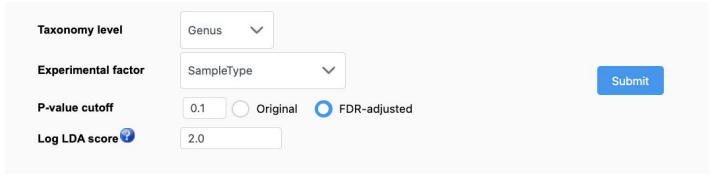
The table below shows at most 500 features ranked by their p values, with significant features highlighted in orange.

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

Click to view feature-level plot

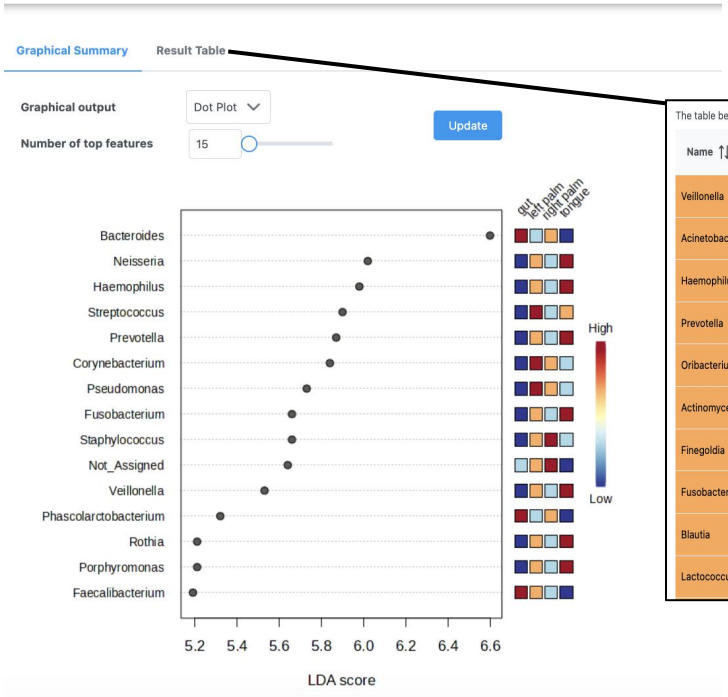


D. Comparison & Classification

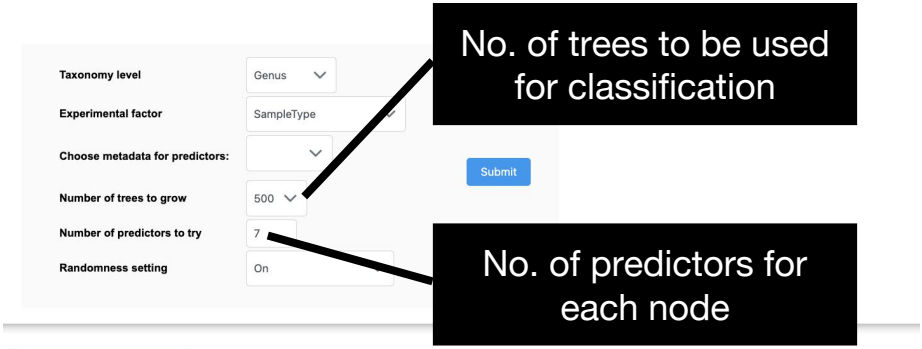


Linear Discriminant Analysis Effect Size (LEfSe):

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



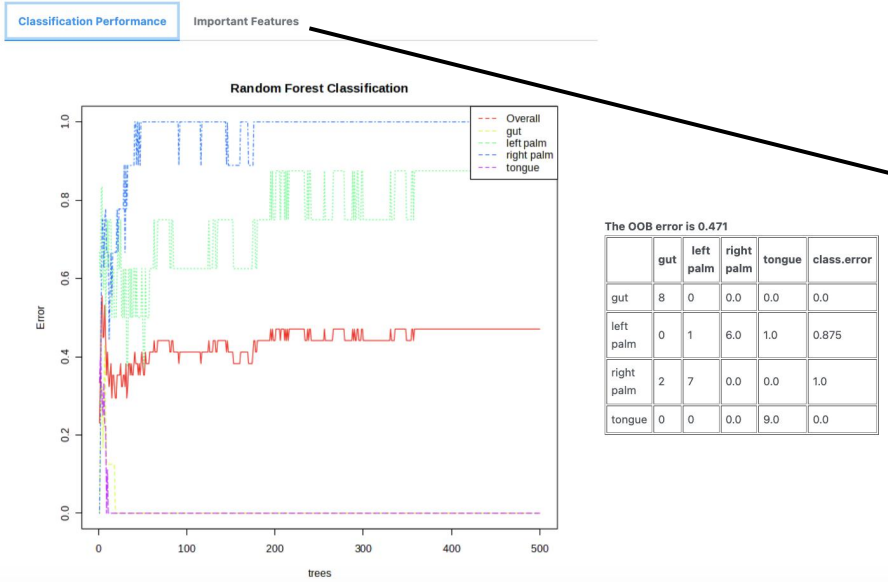
D. Comparison & Classification



Random forests:

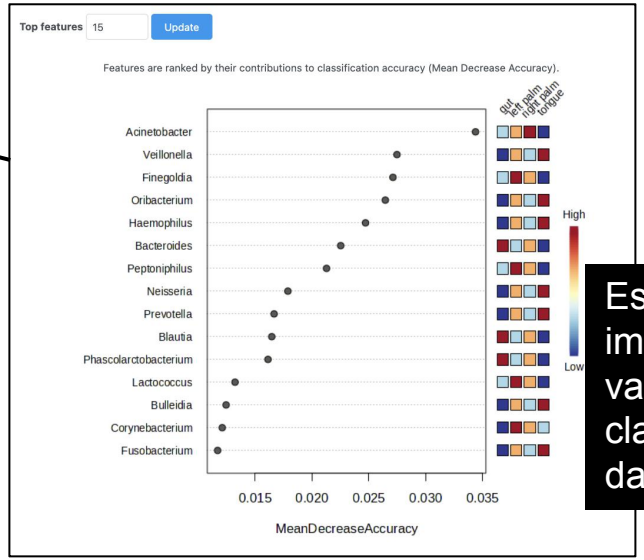
Ensemble learning method used for classification, regression and other tasks.

- 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.



The OOB error is 0.471

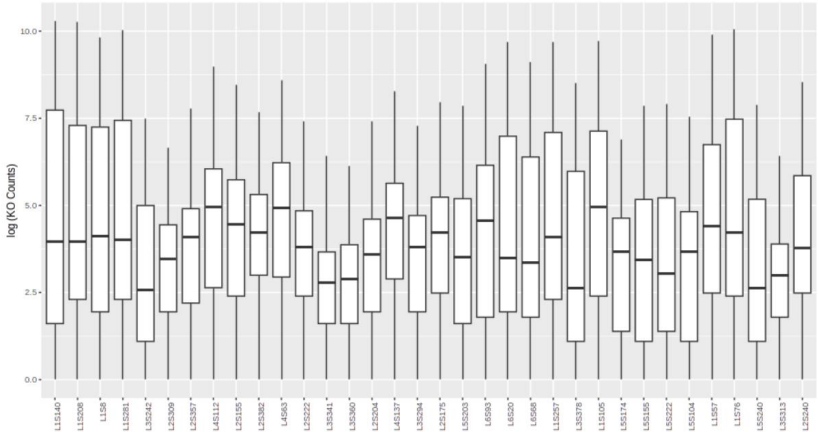
	gut	left palm	right palm	tongue	class.error
gut	8	0	0.0	0.0	0.0
left palm	0	1	6.0	1.0	0.875
right palm	2	7	0.0	0.0	1.0
tongue	0	0	0.0	9.0	0.0



Estimates of important variables in the classification of data.

E. Function prediction

Result figure:



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 greengnes 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 Pipeleine QIIME against SILVA database Predict Functional Potential

QIIME against SILVA database
SILVAngs

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 predi
based on the Ref99NR database. Note, Tax4Fun2 needs 16S rRNA gene sequences for prediction. Please make sure the seque
OTU/ASV table.

Predict Functional Potential

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

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