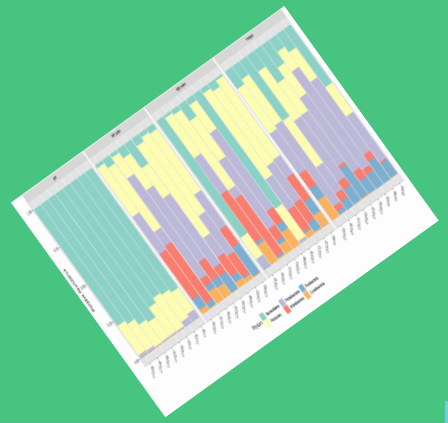
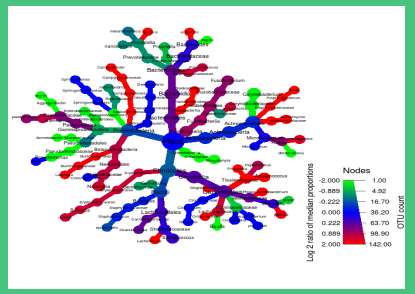
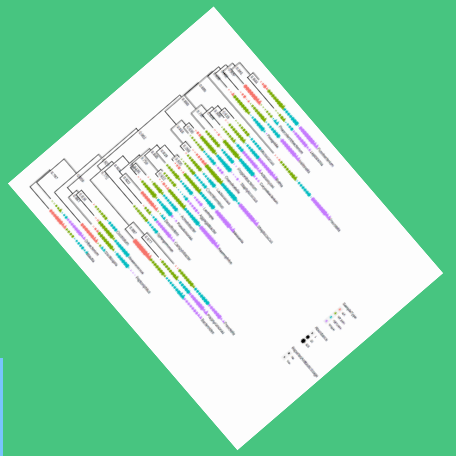
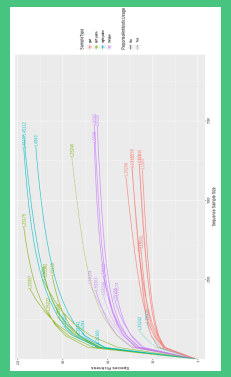








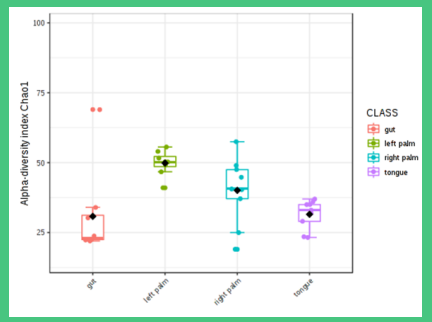
Marker Data Profiling

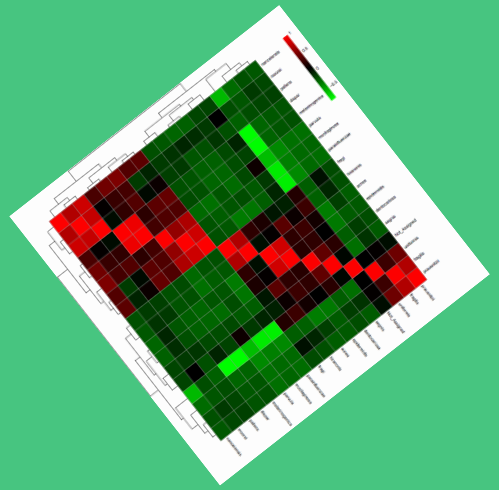
Marker Data Profiling (MDP)

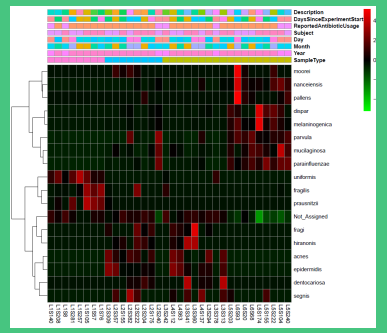
Shotgun Data Profiling (SDP)

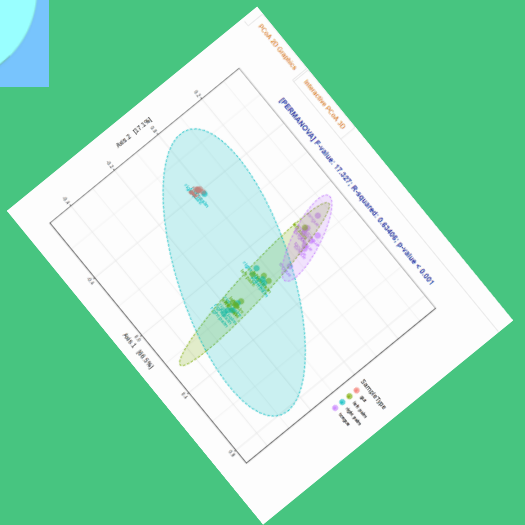
Projection with Public Data (PPD)

Taxon Set Enrichment Analysis (TSEA)









Goals of this tutorial

To perform a comprehensive analysis on data from 16S rRNA sequencing data, including:

- Compositional and structure analysis
- biodiversity (alpha and beta) analysis
- Comparative analysis
- Predictions of metabolic potentials

Starting from marker gene abundance data (OTU table, BIOM file, mothur output)

Marker Data Profiling (MDP)

Shotgun Data Profiling (SDP)

Starting from gene list or gene abundance data annotated by KO, EC or COG

Click here to start

Visually exploring your 16S rRNA data with a public data in a 3D PCoA plot

Projection with Public Data (PPD)

Taxon Set Enrichment Analysis (TSEA)

Starting with a list of taxa of interest (strains, species or higher level taxa)

Data Formatting

User can upload their 16S data in multiple formats :

- Tab-delimited text file (abundance, taxonomy and metadata file)
- BIOM format (containing at least abundance and taxonomy information)
- Mothur output files.
- Details about each format are in the next few slides.

Data Formatting

1. **Tab-delimited text file**

- Manipulate data headings in a spreadsheet program like MS Excel
- Save as a **tab delimited (.txt) or comma-separated (.csv) file**
- The headings **#NAME** (all capital letters) must be used
- #NAME is for sample names (first column in abundance; first row in metadata file)
- 2nd Column of metadata file is for the clinical metadata.
- Taxonomy information can be present within abundance table or uploaded

For Example:

#NAME	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8
#CLASS	Y	N	N	Y	N	Y	Y	N
Archaea;	219	49	42	50	6	17	22	21
Archaea;Crenarchaeota;Thermoprotei;				424	0	191	0	0	0	0	0
Bacteria;Acidobacteria;			32	4	4	22	76	16	1	0	
Bacteria;Actinobacteria;			47	0	0	4	0	0	0	0	

Taxonomic profiles with valid taxonomy identifier labelled names

#NAME	SampleType	Primer
Sample1	skin	ILBC_02
Sample2	gut	ILBC_06
Sample3	skin	ILBC_01
Sample4	gut	ILBC_07
Sample5	gut	ILBC_05
Sample6	gut	ILBC_09
Sample7	skin	ILBC_08
Sample8	skin	ILBC_03

Metadata file

Data Formatting

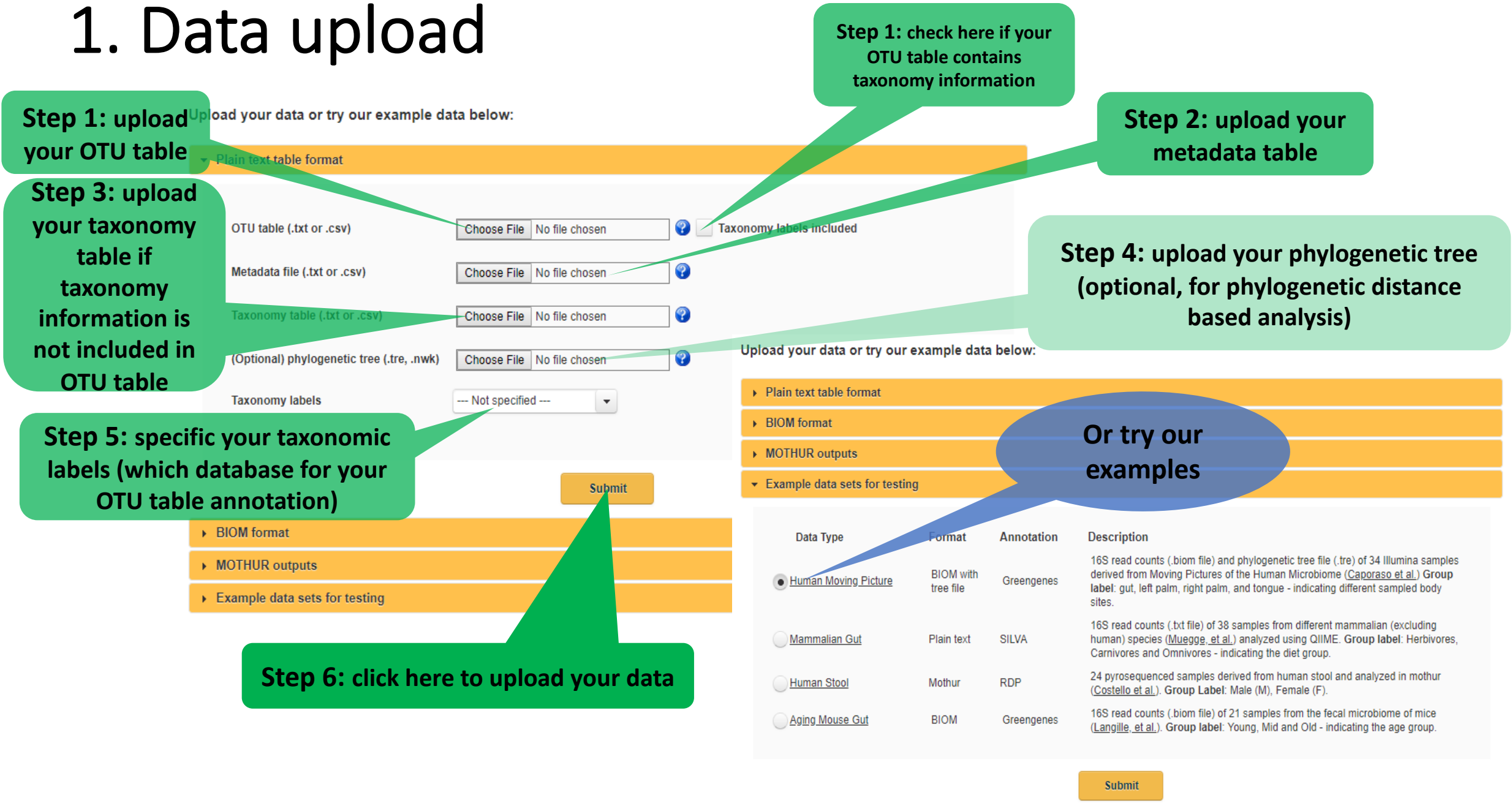
2. BIOM format

- General-use format (standard) for representing biological sample by observation contingency tables.
- For details, please check BIOM format page (http://biom-format.org/)
- QIIME and mothur can also generate output in this format.
- Must contain at least abundance and taxonomy information. (metadata file can be uploaded separately.)

3. Mothur output file

- Two files needed: a consensus taxonomy (taxonomy) file and a .shared (abundance) file.
- Metadata file can be uploaded separately.
- For details, please visit the mothur home page (https://mothur.org/wiki/Main_Page)

1. Data upload



2. Data Integrity Check

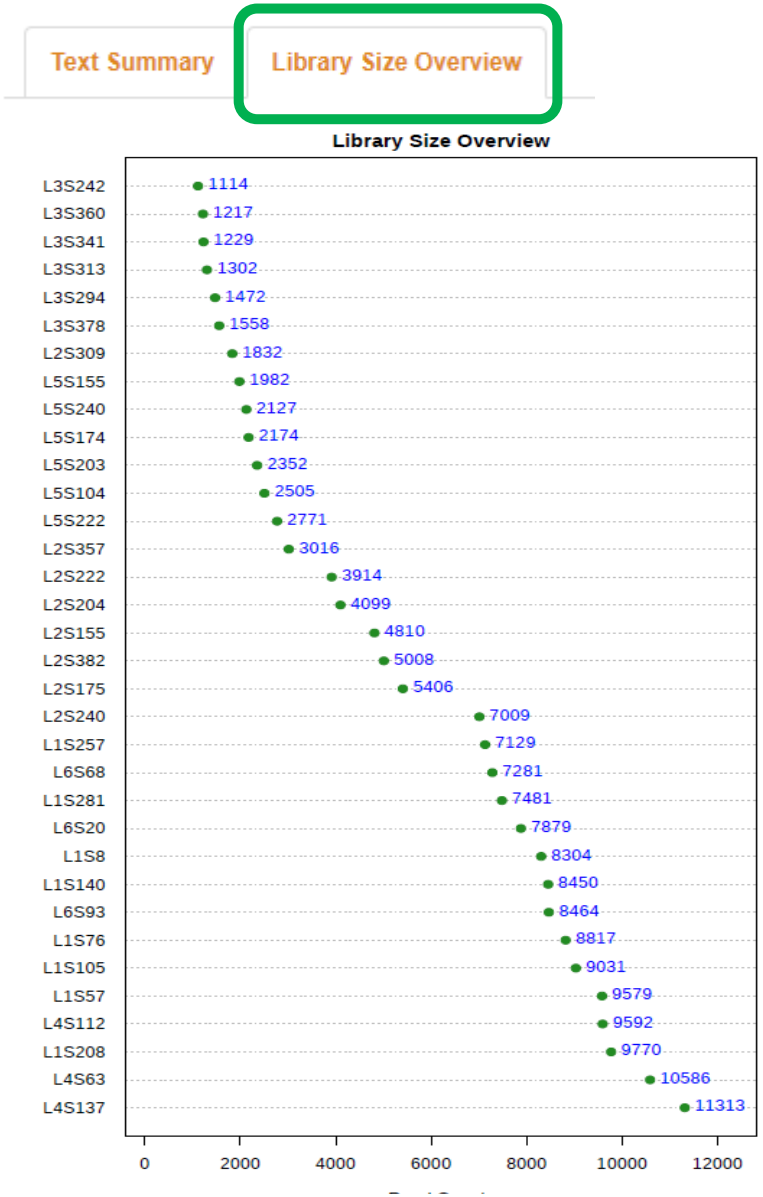
Text Summary | Library Size Overview

Data type:	OTU abundance table
File format:	biom
OTU annotation:	GreengenesID
OTU number:	3426
OTUs with ≥ 2 counts:	2920
Number of experimental factors:	8
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
Sample names match (metadata vs. OTU table):	Yes
Number of sample names matched (metadata vs. OTU table):	34
Number of samples that will be processed:	34

Pay attention to here, only name matched samples (metadata vs. OTU table) will be processed

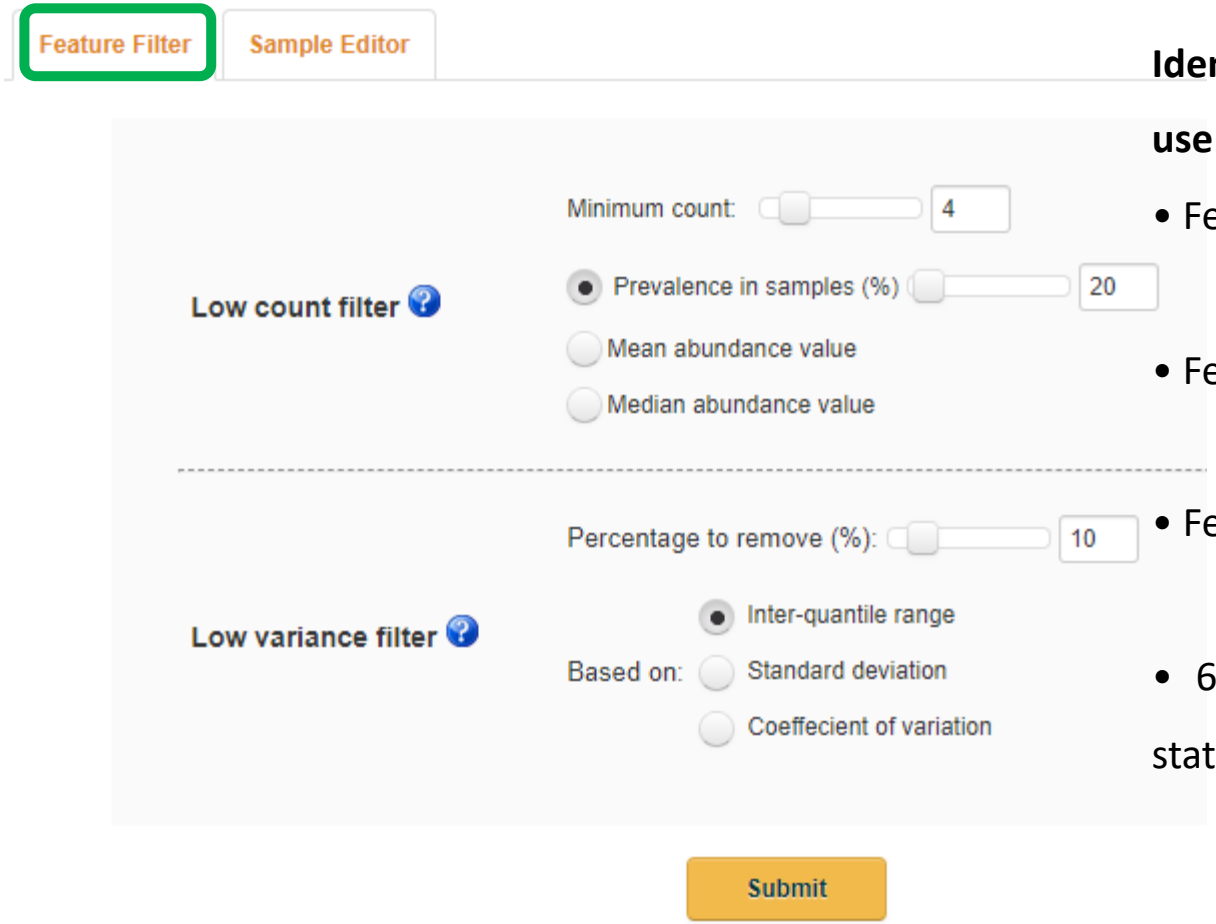
- Provides processing and summary information for user uploaded data.

2. Data Integrity Check



- Provides user the information about library size or total number of reads present in of each sample
- help in identifying the potential outliers due to undersampling or sequencing errors

3. Data Filtering (Features)



Identifying and removing variables or features that are unlikely to be of use when modeling the data.

- Features that are of low quality or low confidence

  All zeros, singleton or detected in only one sample

- Features that are of low abundance

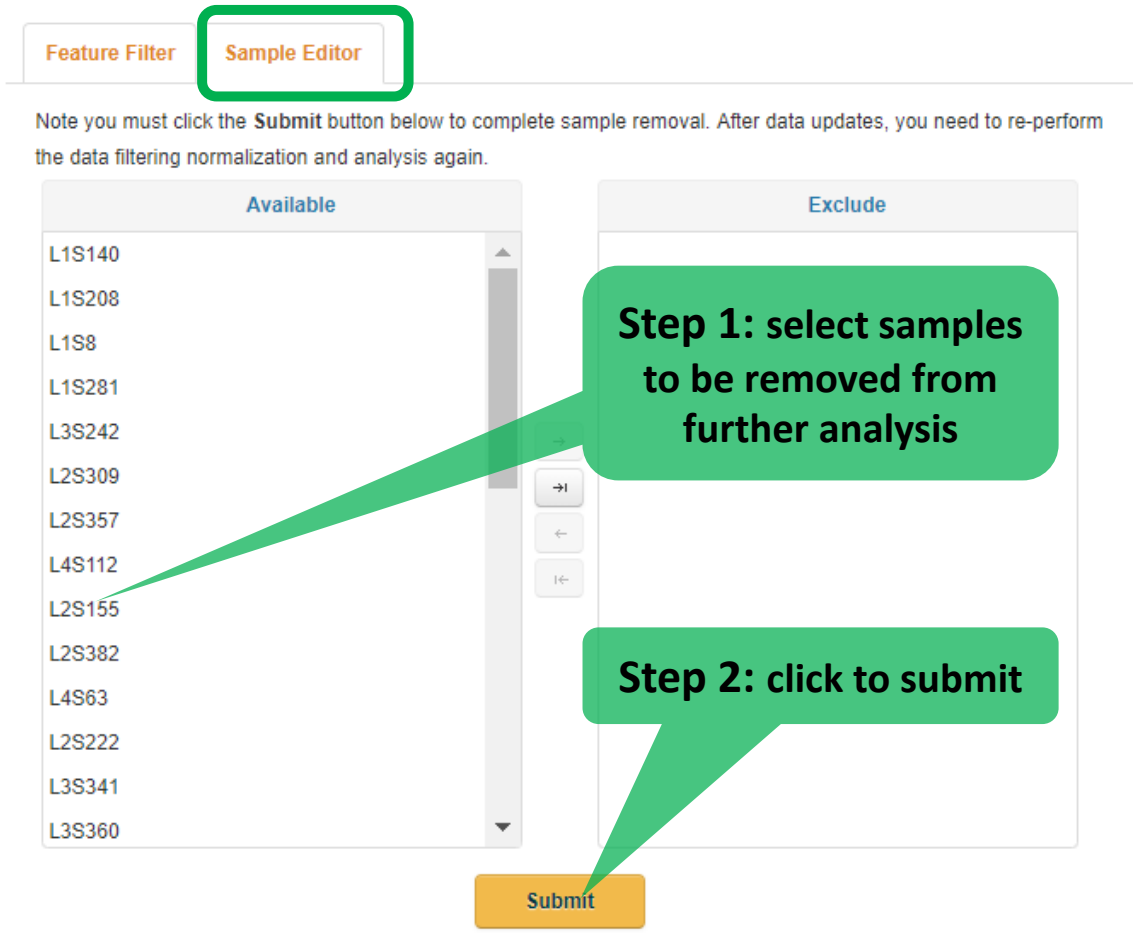
  May be less functionally important

- Features that are of low variance

  Less informative for comparative analysis

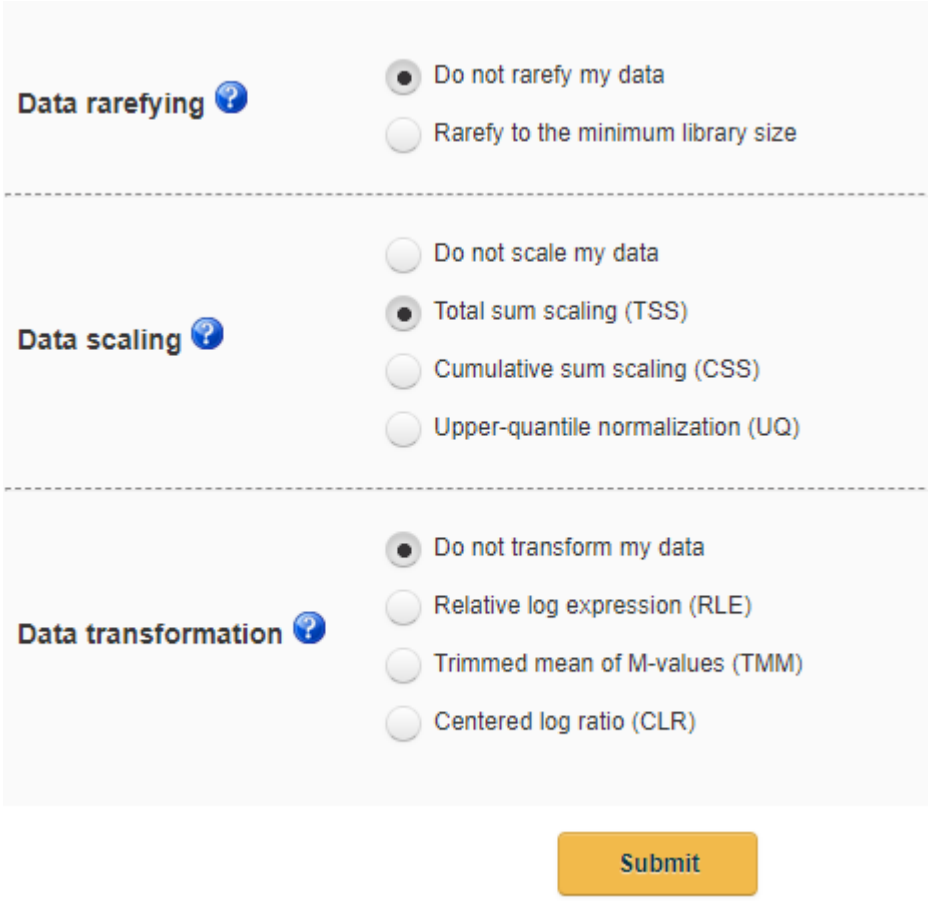
- 6 different approaches: on the basis of count (abundance) or using statistical approaches such as mean, median, IQR, standard deviation or C.V.

3. Data Filtering (Editor)



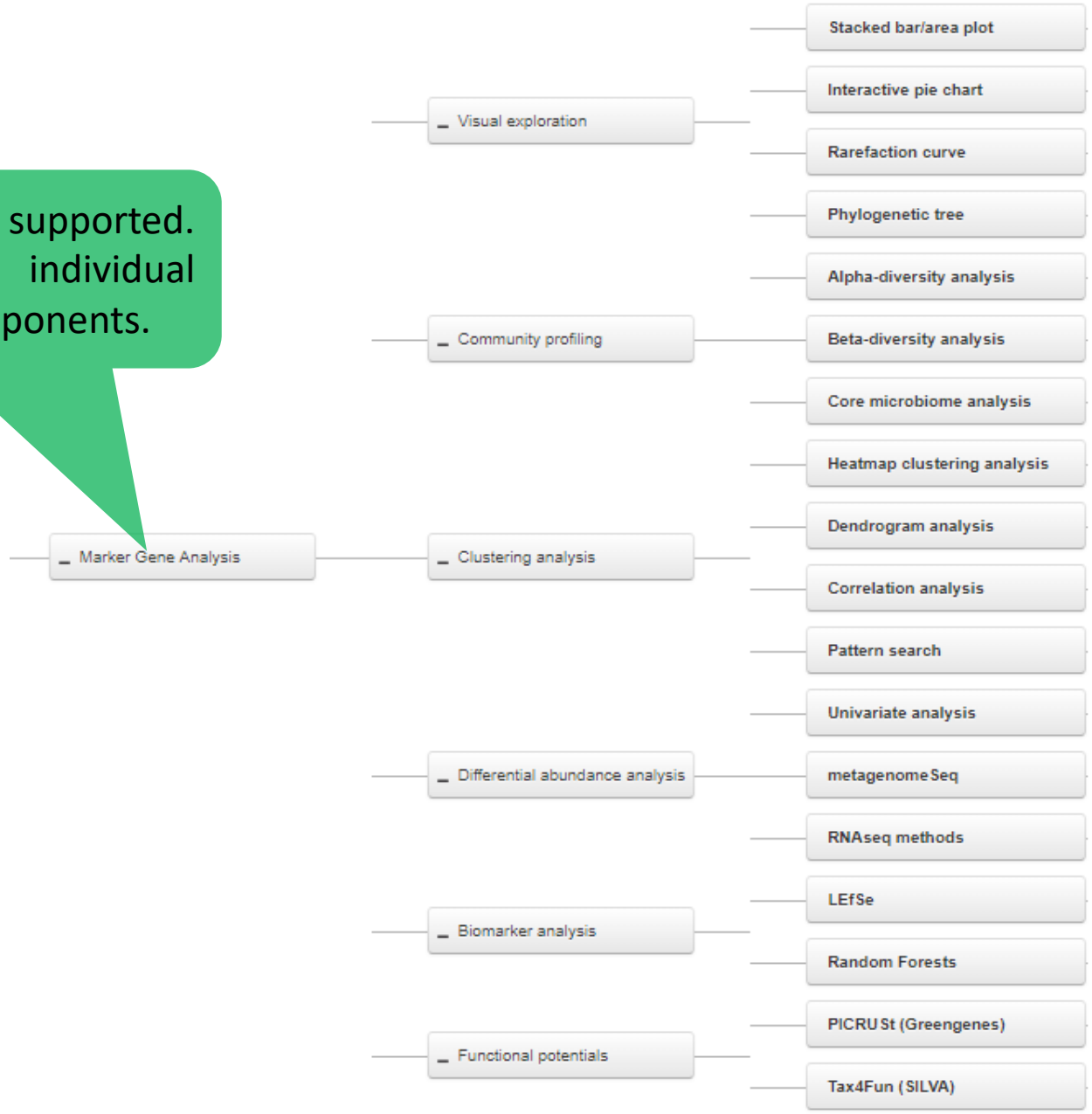
- Users can remove samples that are detected as outlier via results from **graphical summary** or **rarefaction curve analysis**.
- These samples will be excluded from downstream analysis (e.g. alpha- , beta- diversity analysis).

4. Data Normalization

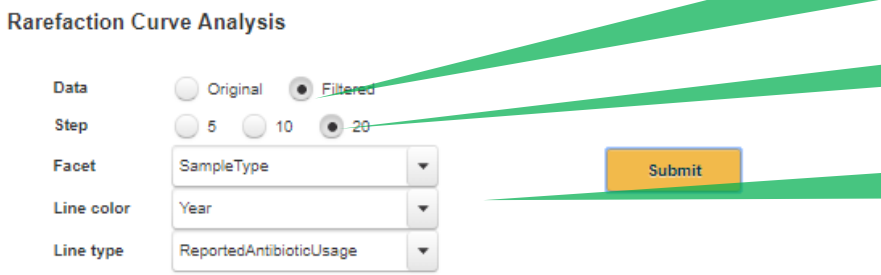


- Normalizing is required to account for **uneven sequencing depth**, **undersampling** and **sparsity present** in such data. (useful before any meaningful comparison)
- Several normalization methods which have been commonly used in the field are present. (3 categories: **rarefaction, data scaling and 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.

Six analysis pathway supported. We will go through individual pathways and their components.



A. Visual Exploration



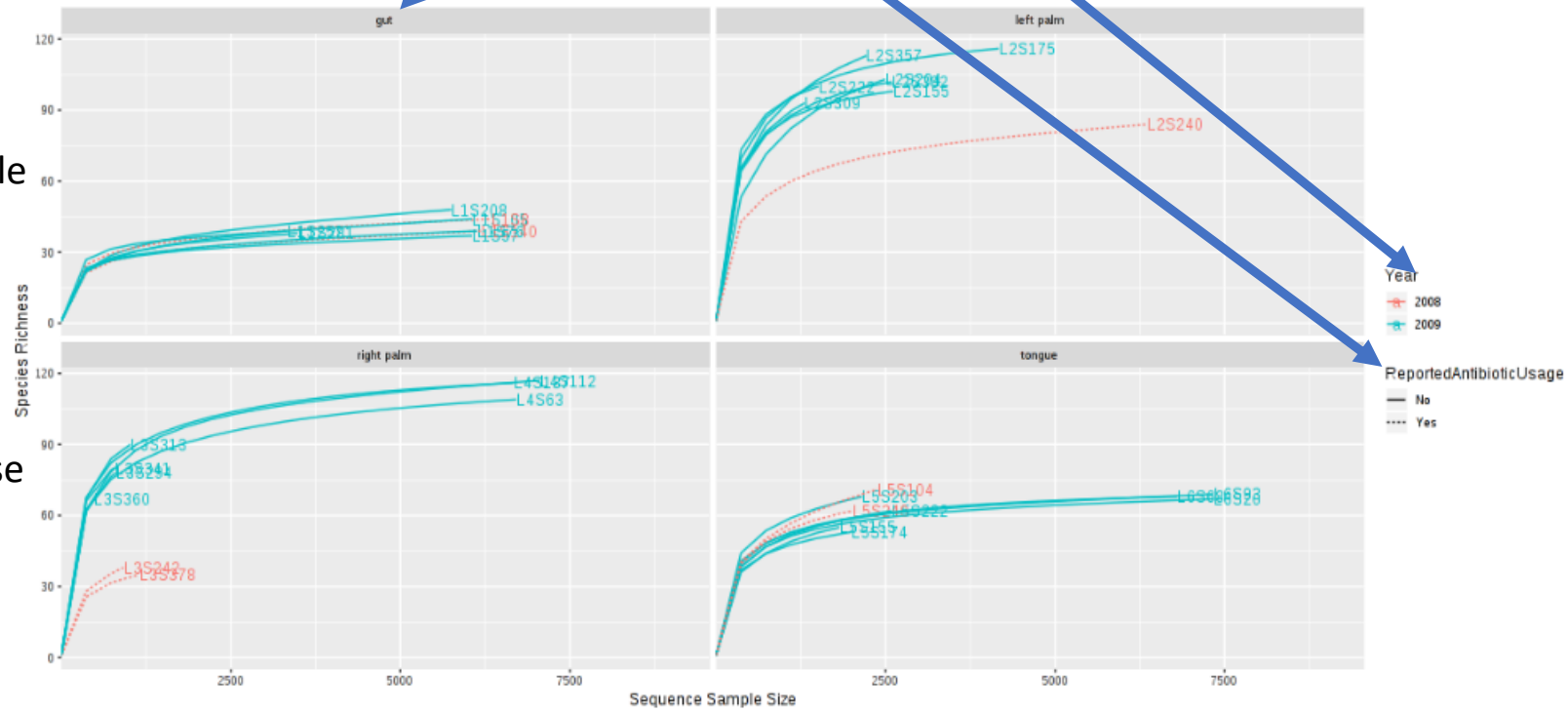
Select original or filtered data for rarefaction curve generation

Step determines the number of subsamples for generating rarefaction curve

Separate by multiple metadata variables

1. Rarefaction curve

- Helps in determining number of observed OTUs (alpha diversity)
- Determining **sequence depth** of each sample
- Determining if sample reaches sequencing **plateau** (with increasing sequence depth, number of recovered OTUs will not be increased)
- 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



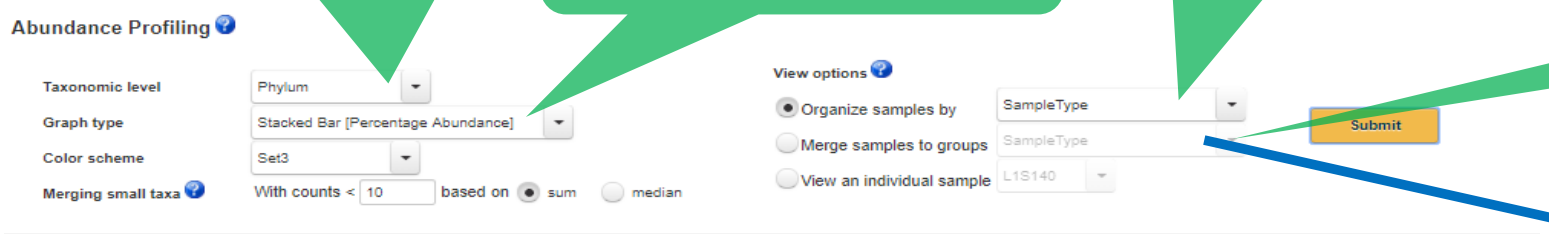
A. Visual Exploration

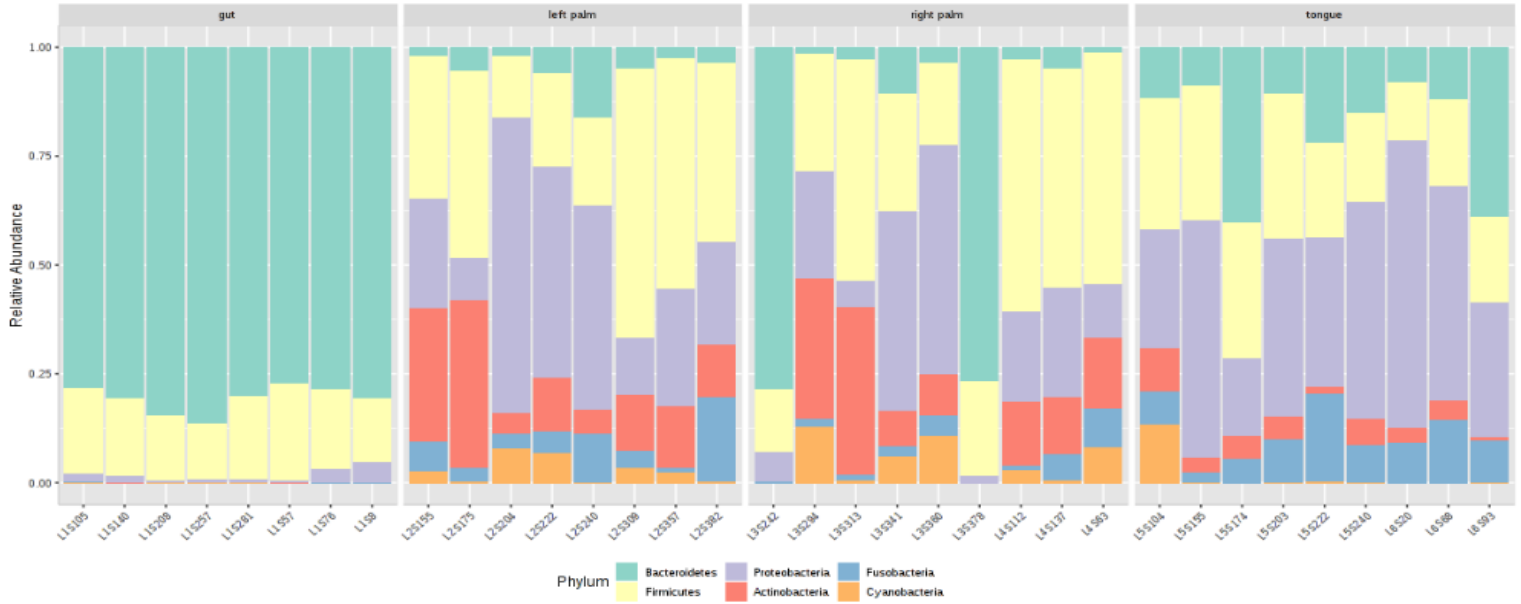
Select different taxonomic levels (eg. Phylum, genus)

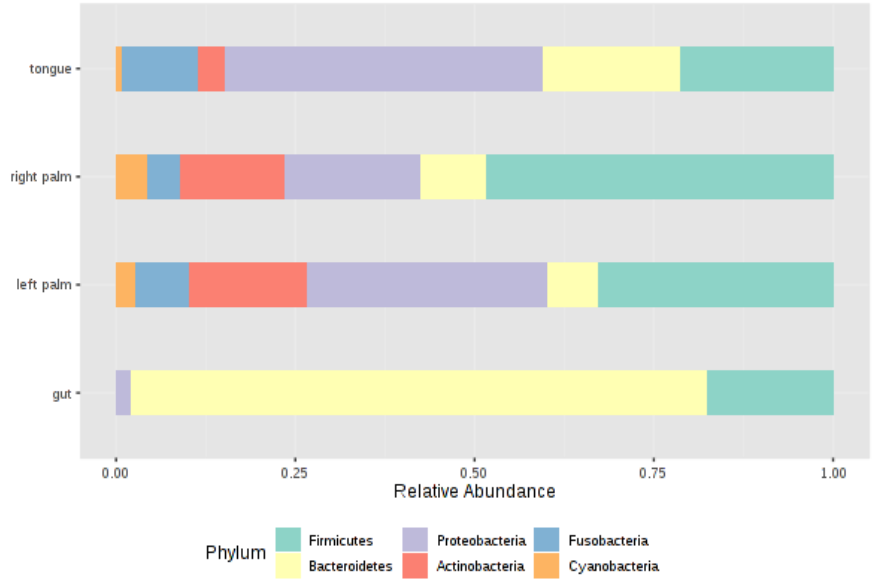
Can be viewed by actual or relative abundance

Group by metadata variables

Or merged to metadata variables







2. 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.(i.e. phylum to OTU)

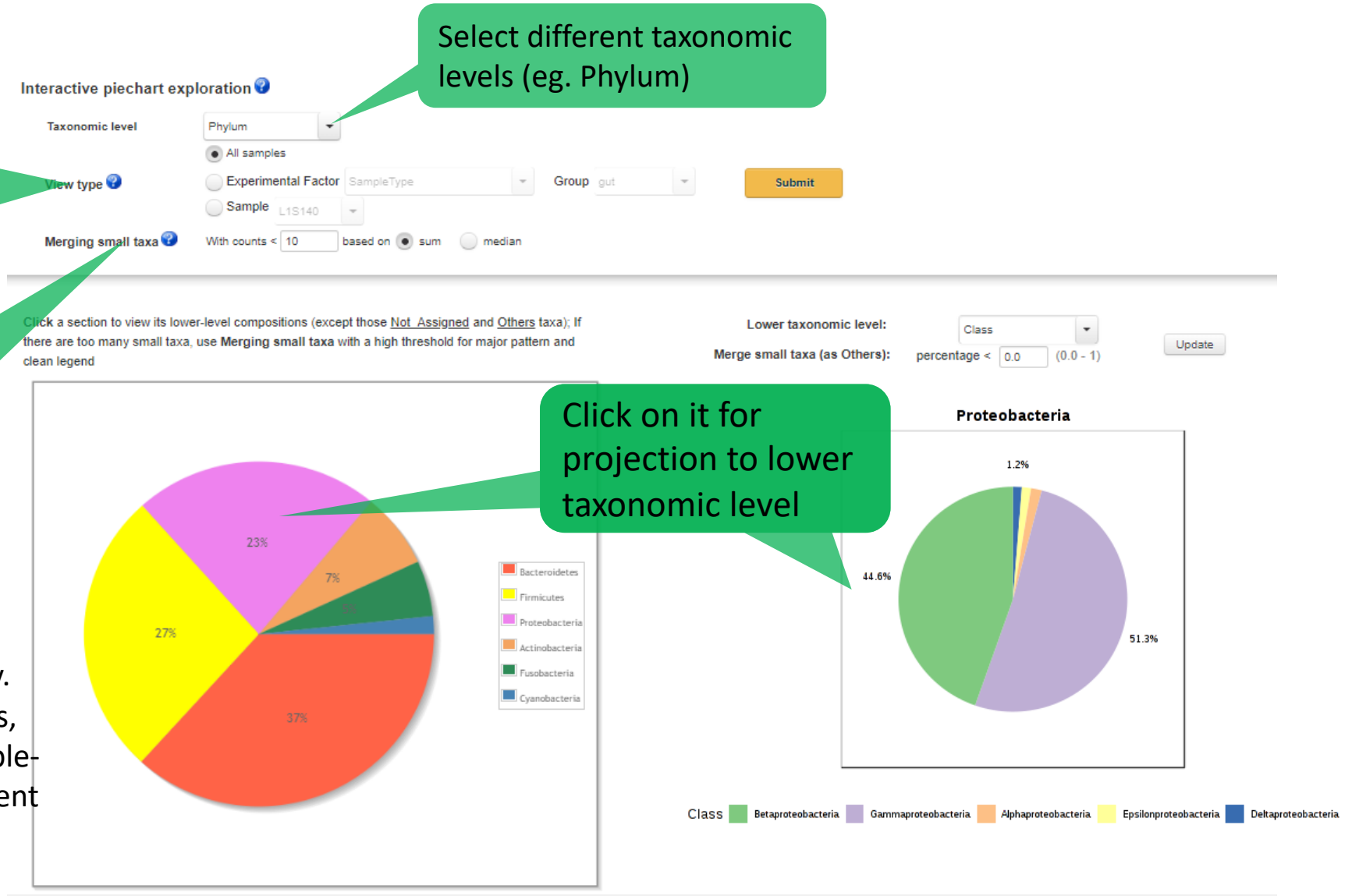
A. Visual Exploration

Can be viewed at 3 different levels: **Community-wise, sample-group wise and individual sample wise**

Select different taxonomic levels (eg. Phylum)

Less abundant taxa can be merged into “Others” category based on sum or median of their count

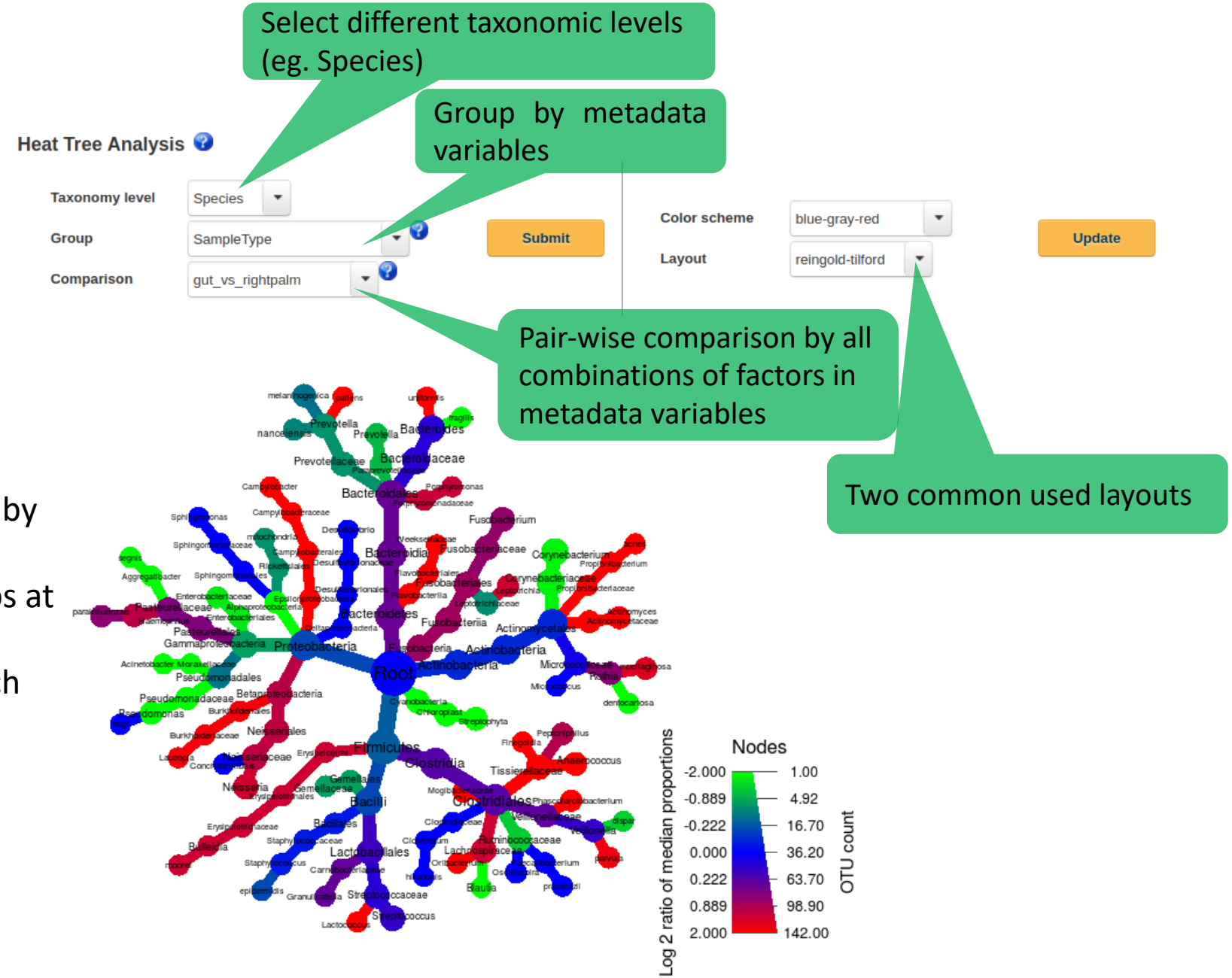
Click on it for projection to lower taxonomic level



3. Pie Chart

- Helps in visualizing the taxonomic compositions of microbial community.
- It can also be created for all samples, sample-group wise or individual sample-wise at multiple taxonomic level present in data.(i.e. phylum to OTU)

A. Visual Exploration



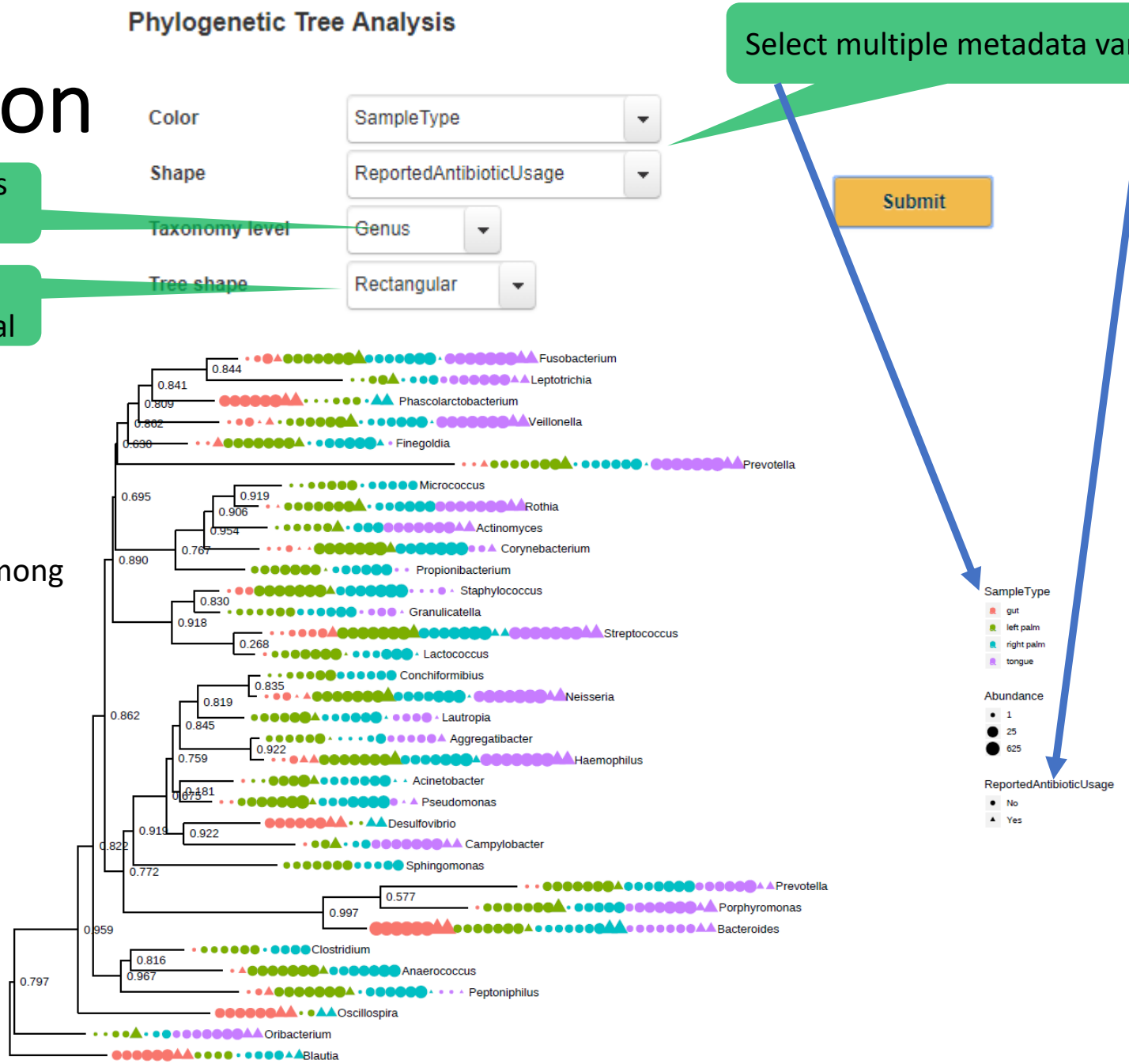
4. Heat tree

- Heat tree is actually 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

A. Visual Exploration

Select different taxonomic levels (eg. Genus)

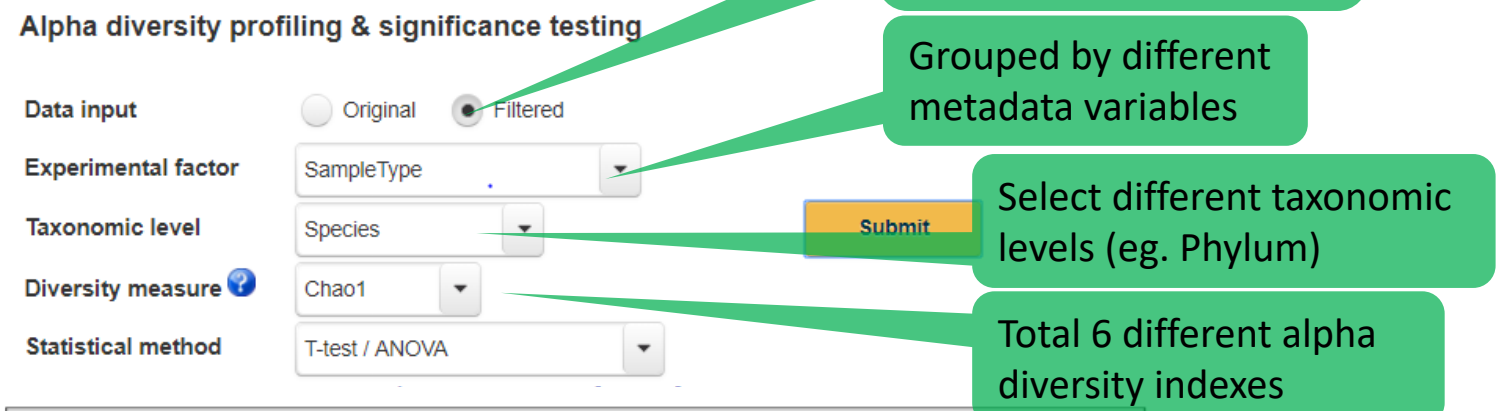
Two types of tree shapes are provided: Rectangular and Redial



5. Phylogenetic tree

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

B. Community Profiling

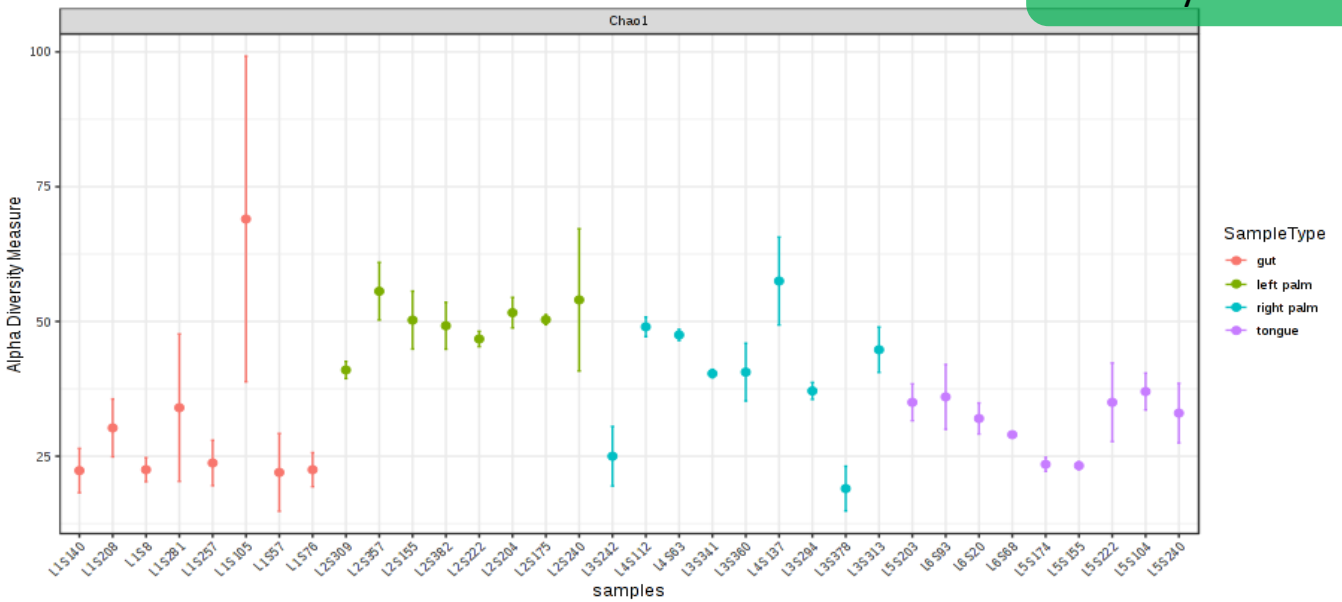


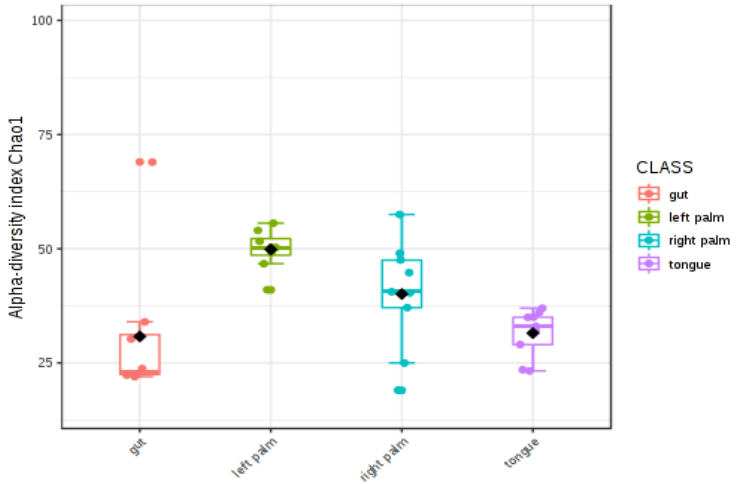
Select original or filtered data for rarefaction curve generation

Grouped by different metadata variables

Select different taxonomic levels (eg. Phylum)

Total 6 different alpha diversity indexes



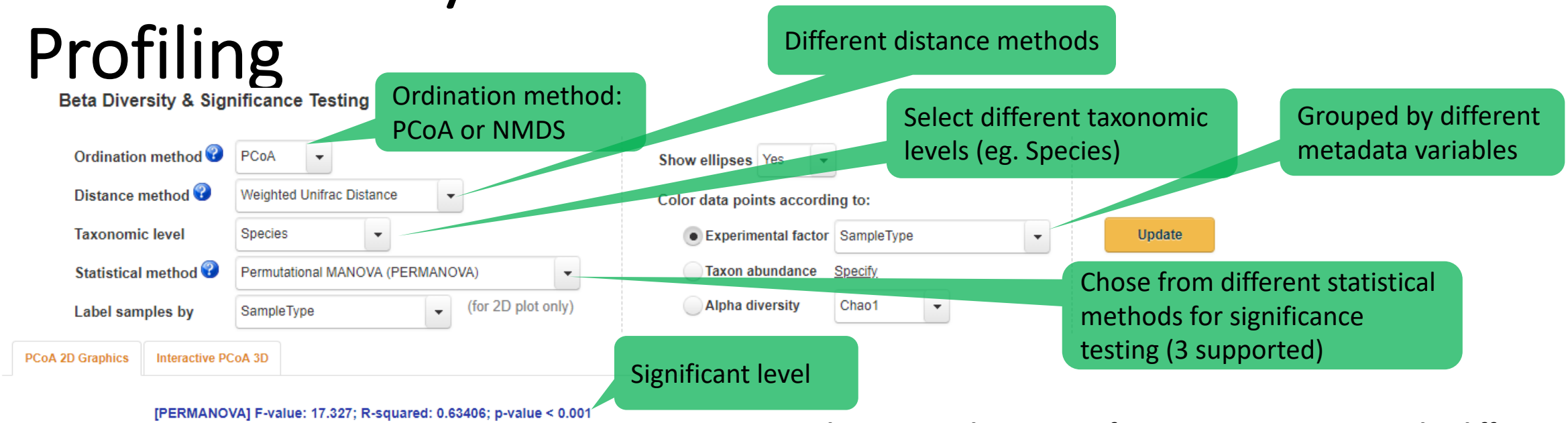


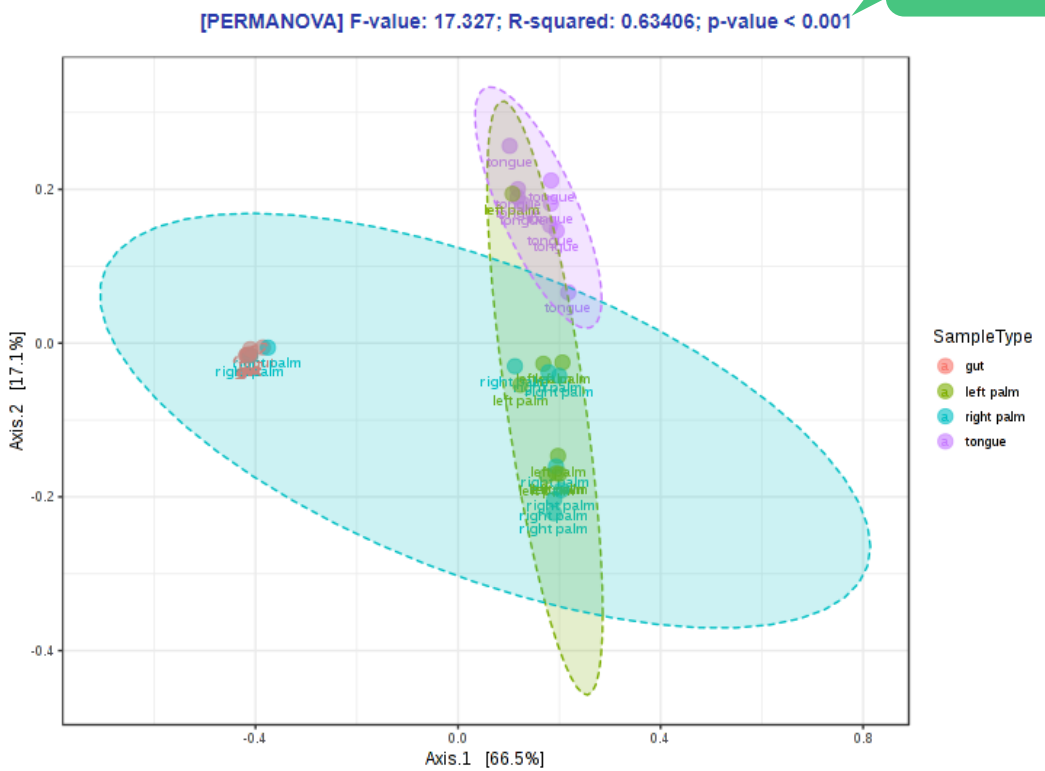
1. Alpha-diversity analysis & significance testing: assessing diversity within community or sample.

• Supporting 6 widely used metrics to calculate the alpha diversity supported such as Chao1 (estimated number of OTUs), and Observed number of OTUs for richness, while Shannon and Simpson take account for both evenness and richness.

• Statistical significance testing between groups using parametric and non-parametric tests.

B. Community Profiling

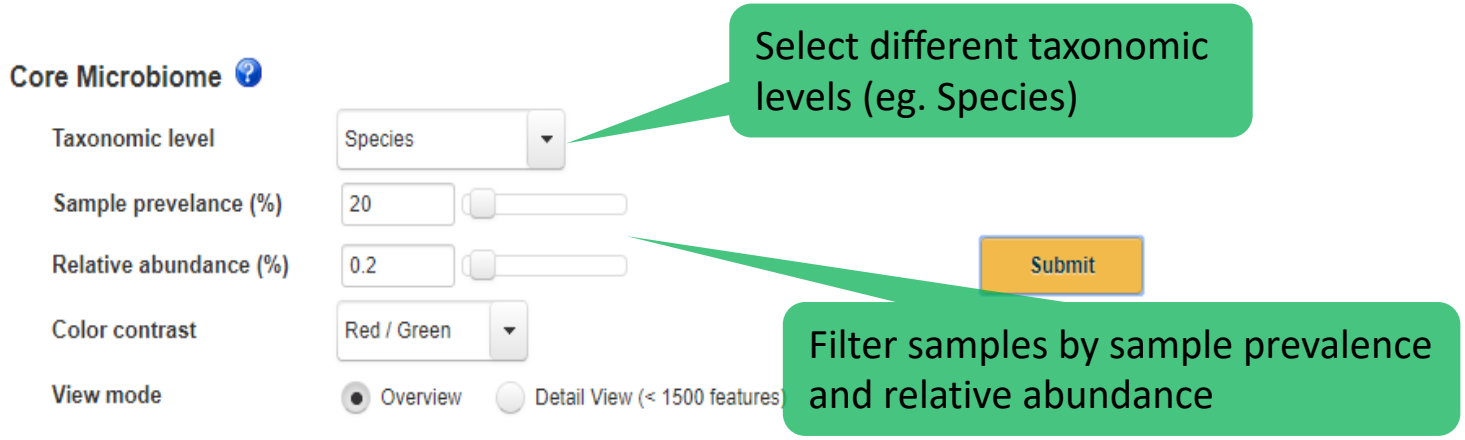




2. Beta diversity analysis & significance testing: assessing the differences between microbial communities (between samples).

- Dissimilarity matrix can be calculate via multiple distance method and can be visualized using PCoA (Principal Coordinate Analysis) or NMDS (Nonmetric Multidimensional Scaling)
- 5 widely used methods: compositional-based distance metrics such as Bray-Curtis or phylogenetic-based (Unweighted Unifrac) supported.
- Of these distance, unweight- and weight unifrac distances are based on phylogenetic tree, therefore, phylogenetic tree must be provided.
- 3 statistical methods supported to tests the strength and statistical significance of sample groupings based on ordination based distances.
- **ANOSIM/adonis, PERMANOVA and PERMDISP** supported.
- Helps in understanding the underlying reasons for pattern present in PCoA or NMDS plot.

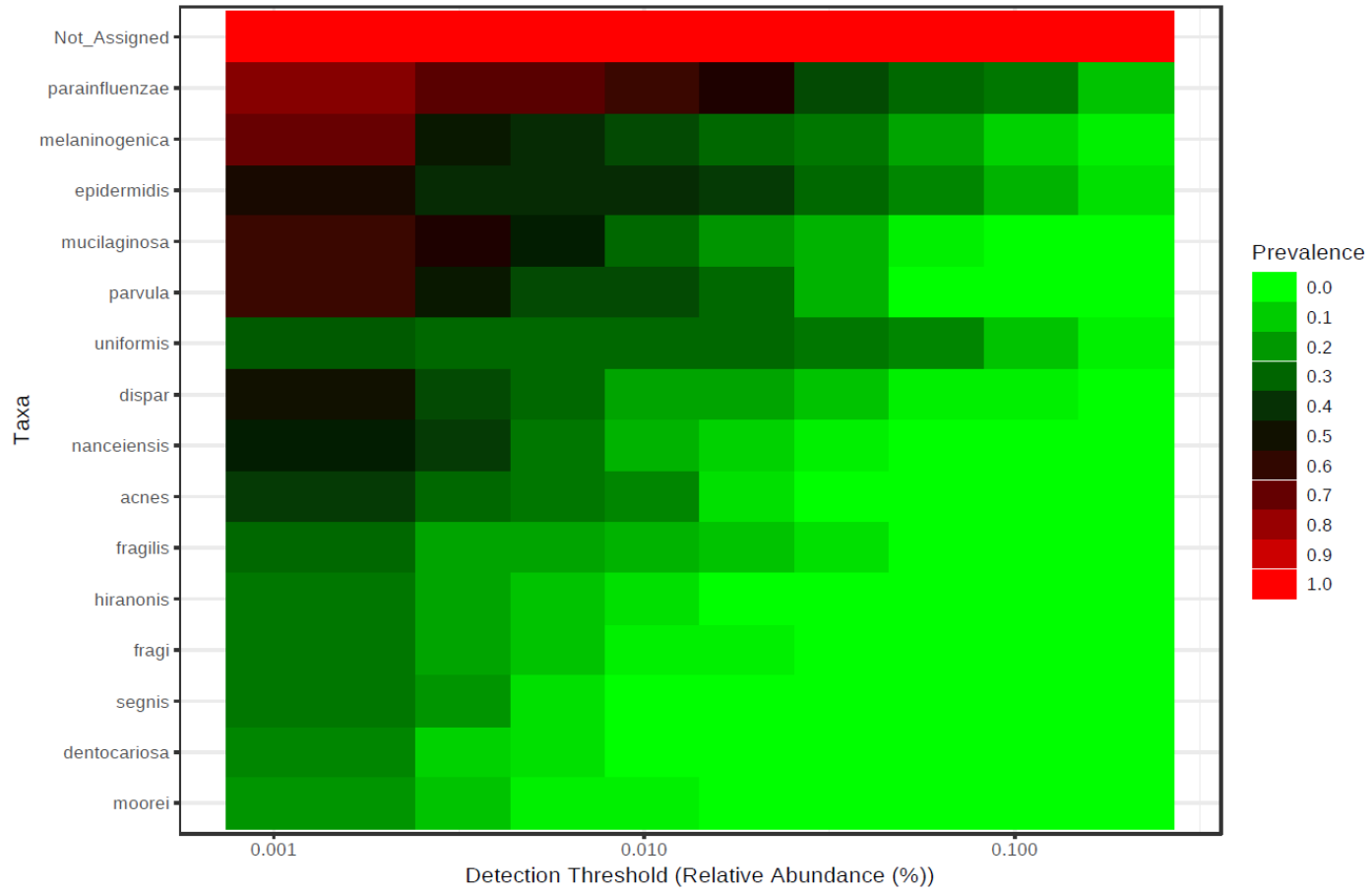
B. Community Profiling



3. 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.

- Can be performed at various taxonomical level. (Phylum to OTU)

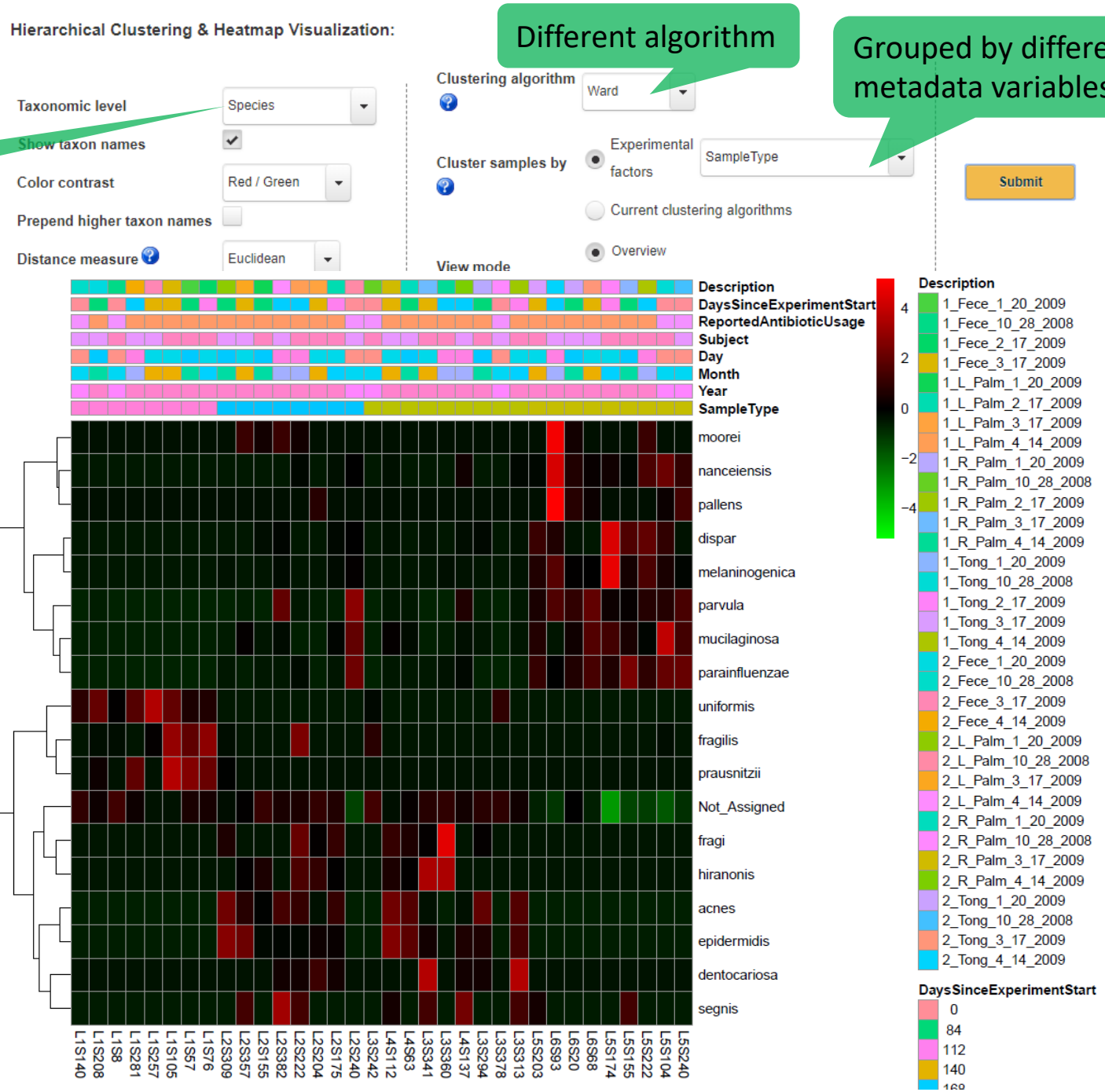


C. Clustering analysis

Select different taxonomic levels (eg. Species)

Different algorithm

Grouped by different metadata variables



1. Heatmap and clustering analysis

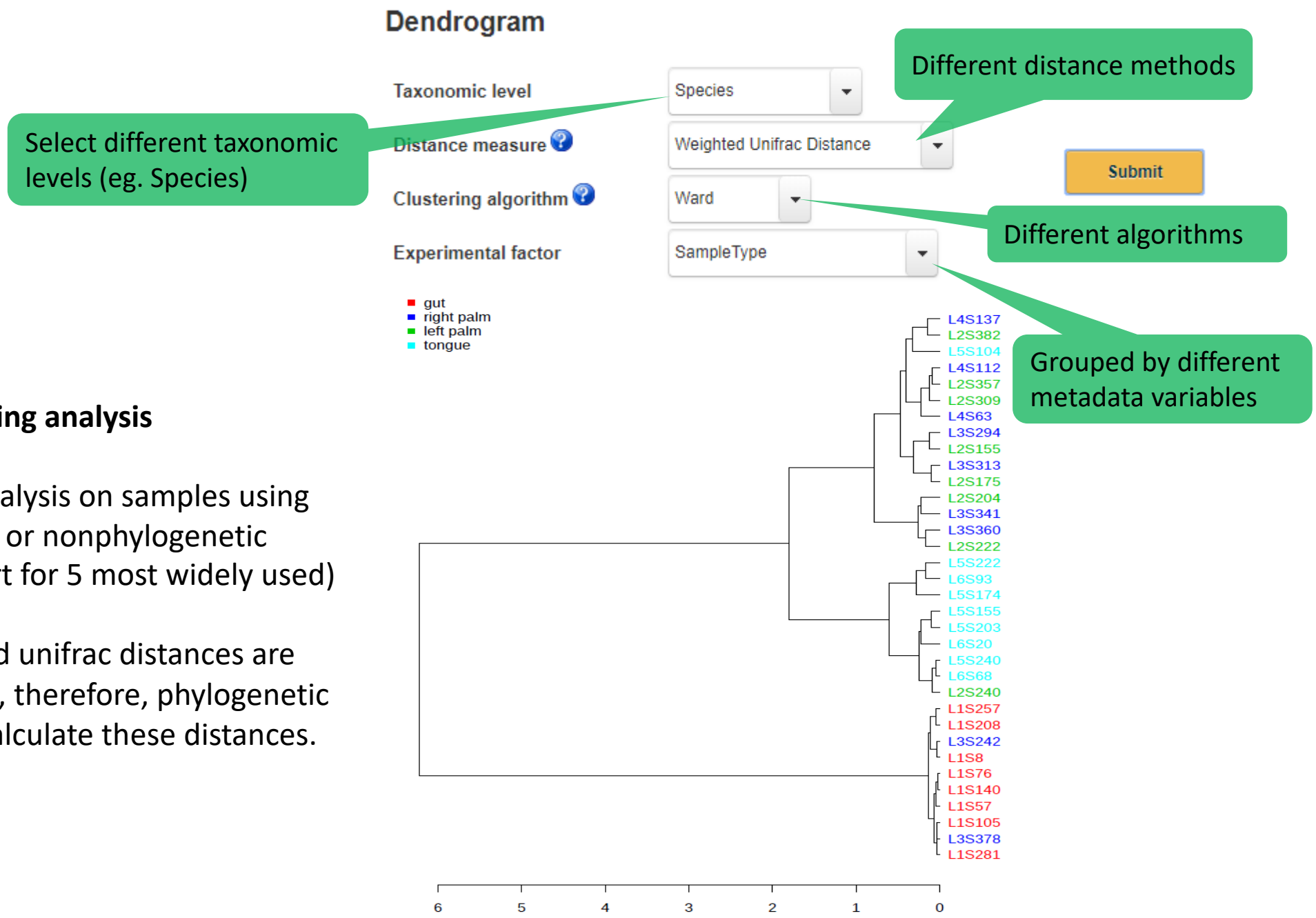
- Visualize the relative patterns of high-abundance features against a background of features that are mostly low-abundance or absent.

- Identify abundance patterns, clusters

- Various distance and clustering methods supported.(both sample and feature-wise)

- Features can be merged at multiple taxonomic levels also.(can also be visualized at individual OTU-level)

C. Clustering analysis

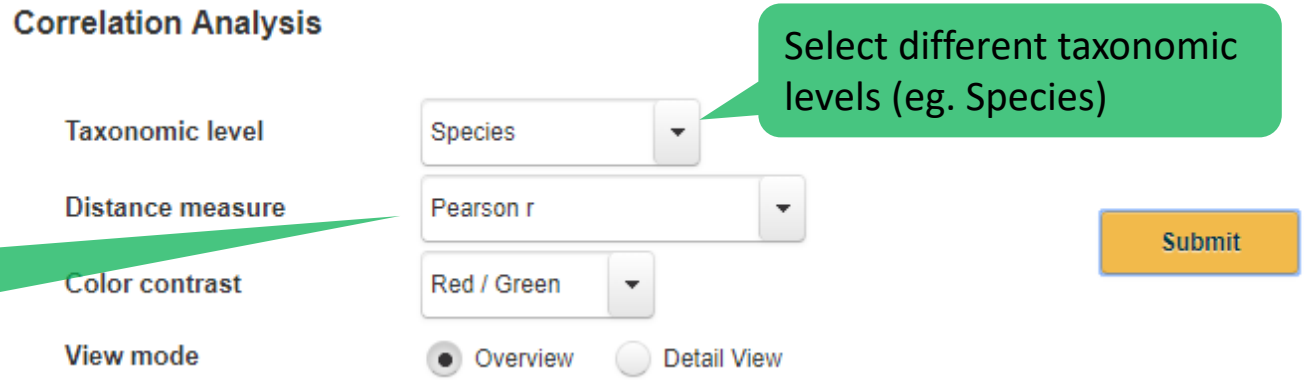


2. Dendrogram and clustering analysis

- Performs phylogenetic analysis on samples using either various phylogenetic or nonphylogenetic distance measures. (support for 5 most widely used)

- Unweighted and weighted unifrac distances are based on phylogenetic tree, therefore, phylogenetic tree must be provided to calculate these distances.

C. Clustering analysis



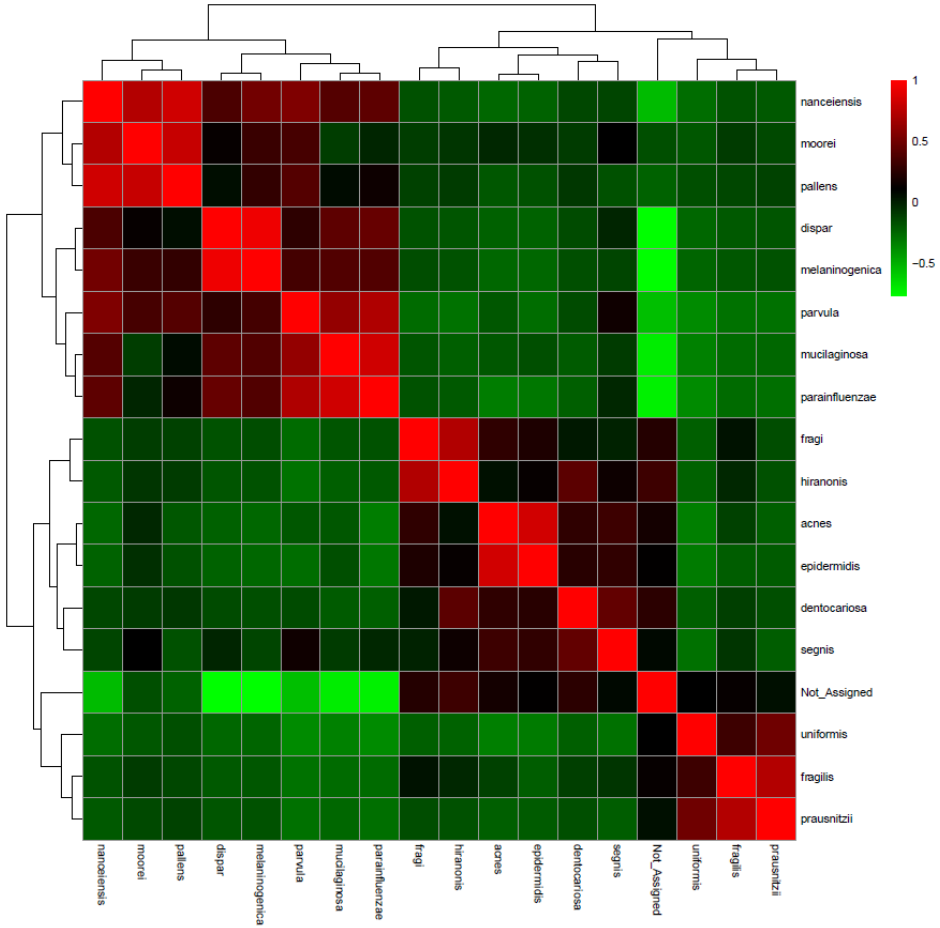
Select different taxonomic levels (eg. Species)

3 most common methods supported for performing Correlation analysis

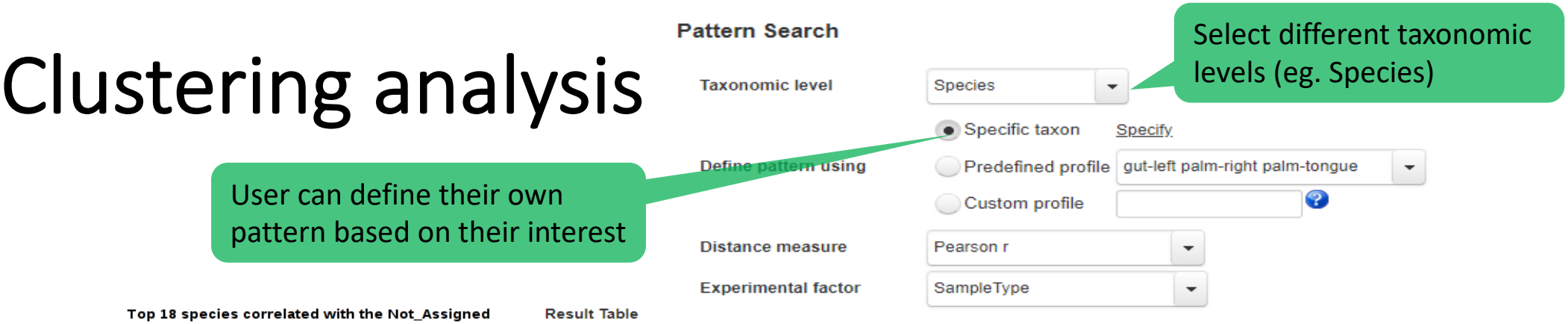
3. Correlation analysis

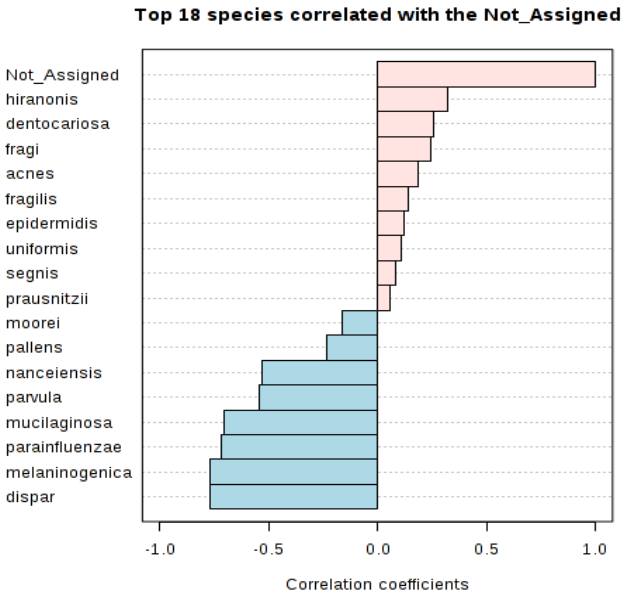
- Helps in identifying biologically or biochemically meaningful relationship or associations between taxa or features.

- Can be analyzed at various level (Phylum to OTU) by merging data based on taxonomic rank.



C. Clustering analysis





Result Table

Name	correlation	t-stat	p-value	FDR	View
Not_Assigned	1.0	0.0	0.0	0.0	Details
dispar	-0.77244	-6.8803	8.7202E-8	6.4311E-7	Details
melaninogenica	-0.76911	-6.8074	1.0718E-7	6.4311E-7	Details
parainfluenzae	-0.7188	-5.8488	1.6821E-6	7.5697E-6	Details
mucilaginosa	-0.70454	-5.616	3.3079E-6	1.1908E-5	Details
parvula	-0.54552	-3.6821	8.4775E-4	0.0025433	Details
nanceiensis	-0.53182	-3.5525	0.0012081	0.0031065	Details
hiranonis	0.32528	1.9459	0.060497	0.13612	Details
dentocariosa	0.25944	1.5196	0.13843	0.27685	Details
fragi	0.24493	1.4291	0.16267	0.2928	Details
pallens	-0.23017	-1.338	0.19034	0.31147	Details
acnes	0.1882	1.084	0.28648	0.42972	Details
moorei	-0.16123	-0.92414	0.36233	0.50169	Details
fragilis	0.13993	0.79944	0.42993	0.55277	Details
epidermidis	0.12296	0.7009	0.48843	0.58612	Details
uniformis	0.11011	0.62671	0.53529	0.60221	Details

4. 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 (gene) of interest or based on predefined or custom profile of experimental factors.

D. Differential abundance analysis

Univariate Statistical Comparisons

Taxonomic level: Species

Experimental factor: SampleType

Statistical method: Mann-Whitney/Kruskal-Wallis

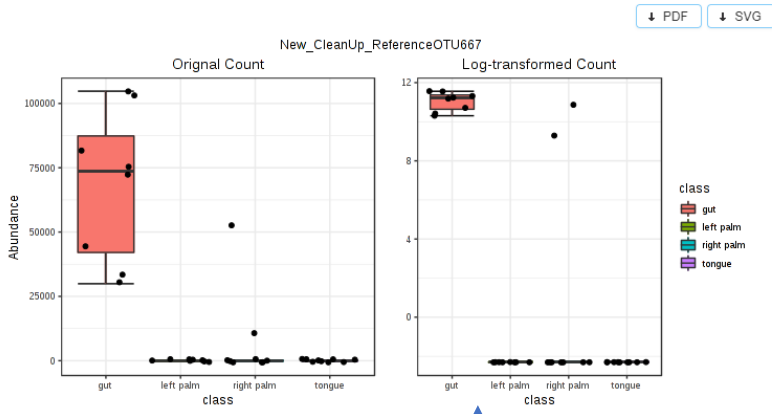
Adjusted p-value cutoff: 0.05

Submit

Select different taxonomic levels (eg. Species)

Select different metadata variables

Two common statistical methods



Click to view more details

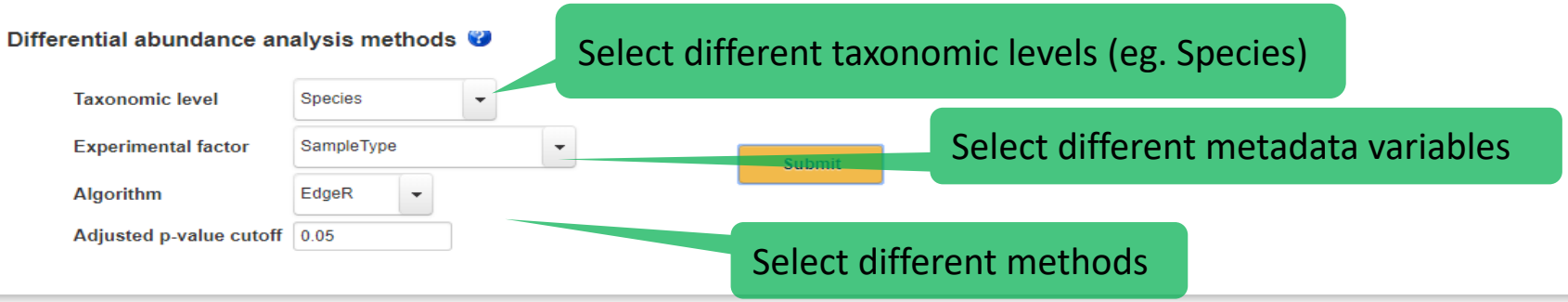
The table below shows at most 500 features ranked by their p values, with significant features highlighted in orange.

Name	Pvalues	FDR	Statistics	View
melaninogenica	7.8686E-6	1.1341E-4	26.399	Details
uniformis	1.8305E-5	1.1341E-4	24.646	Details
parainfluenzae	2.0014E-5	1.1341E-4	24.461	Details
nanceiensis	6.6475E-5	2.8252E-4	21.961	Details
mucilaginosa	1.2396E-4	4.2147E-4	20.658	Details
epidermidis	1.5219E-4	4.312E-4	20.229	Details
acnes	2.3552E-4	5.7198E-4	19.313	Details
parvula	3.5365E-4	7.5151E-4	18.459	Details
prausnitzii	0.0014611	0.0027598	15.463	Details
dispar	0.0016441	0.0027949	15.212	Details
pallens	0.0068093	0.010523	12.174	Details

1. Univariate Statistical Comparisons

- t-test/ANOVA (parametric) or Mann-Whitney/KW test (non-parametric) can be done.
- Depending upon no. of sample groups, statistical test is chosen from parametric or nonparametric test options.
- P-values adjusted using FDR method.

D. Differential abundance analysis



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

Name	log2FC	logCPM	Pvalues	FDR	View
uniformis	-6.1927	15.49	1.2706E-11	2.2872E-10	Details
epidermidis	4.6211	15.319	1.9513E-8	1.7561E-7	Details
parainfluenzae	3.9407	16.023	3.4941E-8	2.0964E-7	Details
prausnitzii	-3.811	13.736	1.0642E-7	4.7888E-7	Details
parvula	2.2585	14.031	0.0011635	0.0041884	Details
mucilaginosa	2.2333	14.108	0.0017577	0.0052731	Details
acnes	1.8684	13.247	0.0028033	0.0072086	Details
fragilis	-1.9564	13.35	0.0036537	0.0082209	Details
melaninogenica	2.1397	15.288	0.0041761	0.0083522	Details
Not_Assigned	-0.45305	19.512	0.050145	0.090261	Details
dispar	0.90192	13.955	0.28292	0.46295	Details

2. EdgeR

- Developed for RNAseq data analysis.
- Powerful statistical method (outperforms others methods with appropriate data filtration and normalization techniques)
- By default, **RLE** (Relative Log Expression) normalization is performed on the data.

E. Biomarker Analysis

Select different taxonomic levels (eg. Species)

Select different metadata variables

Linear Discriminant Analysis (LDA) Effect Size (LEfSe)

Taxonomic level: Species

Experimental factor: SampleType

Adjusted p-value cutoff: 0.05, 1.0

Submit

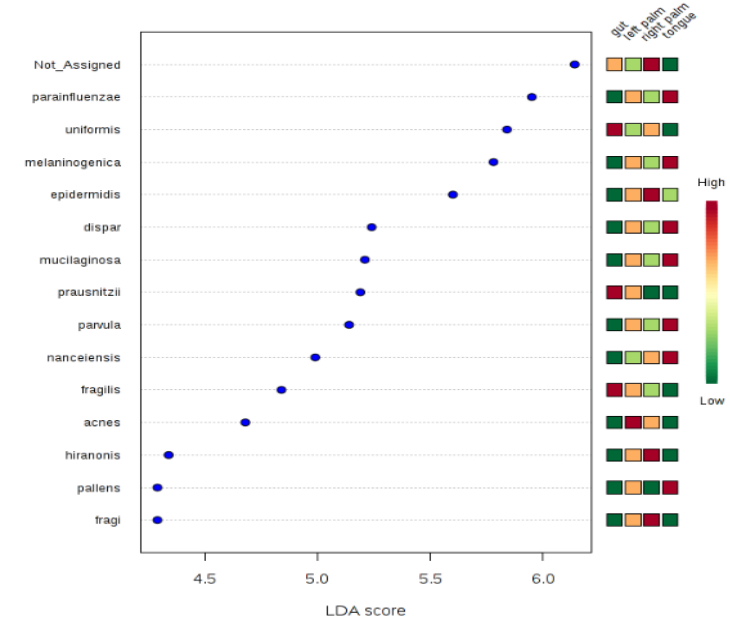
Result Table | Graphical Summary

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
melaninogenica	7.8686E-6	1.2009E-4	0.0	113170.0	34294.0	1209400.0	5.78	Details
uniformis	1.8305E-5	1.2009E-4	1368300.0	2572.0	139460.0	0.0	5.84	Details
parainfluenzae	2.0014E-5	1.2009E-4	5144.0	516980.0	173750.0	1792400.0	5.95	Details
nanceiensis	6.6475E-5	2.9914E-4	0.0	15432.0	18290.0	196620.0	4.99	Details
mucilaginosa	1.2396E-4	4.4627E-4	0.0	113170.0	43439.0	326930.0	5.21	Details
epidermidis	1.5219E-4	4.5656E-4	2572.0	740740.0	800180.0	4572.5	5.6	Details
acnes	2.3552E-4	6.0562E-4	0.0	95165.0	84591.0	0.0	4.68	Details
parvula	3.5365E-4	7.9571E-4	2572.0	144030.0	41152.0	278920.0	5.14	Details
Not_Assigned	8.6748E-4	0.001735	8171300.0	8060700.0	8502500.0	5772700.0	6.14	Details
prausnitzii	0.0014611	0.0026299	308640.0	7716.0	0.0	0.0	5.19	Details
dispar	0.0016441	0.0026903	0.0	36008.0	18290.0	347510.0	5.24	Details
pallens	0.0068093	0.010214	0.0	5144.0	0.0	38866.0	4.29	Details
hiranonis	0.007758	0.010742	0.0	28292.0	43439.0	0.0	4.34	Details
fragi	0.0090037	0.011576	0.0	28292.0	38866.0	0.0	4.29	Details
fragilis	0.010274	0.012329	141460.0	43724.0	16004.0	2286.2	4.84	Details
dentocariosa	0.018733	0.021075	0.0	12860.0	27435.0	0.0	4.14	Details
moorei	0.030546	0.032343	0.0	15432.0	0.0	22862.0	4.06	Details
segnis	0.16939	0.16939	0.0	20576.0	18290.0	6858.7	4.01	Details

Result Table | Graphical Summary

Top features 15 Update



1. LEfSe

- Compare the metagenomics (16S or shotgun) abundance profiles between samples in different state.
- Performs a set of statistical tests for detecting differentially abundant features (KW sumrank test: statistical significance) and biomarker discovery.(Linear Discriminant analysis: Effect Size)

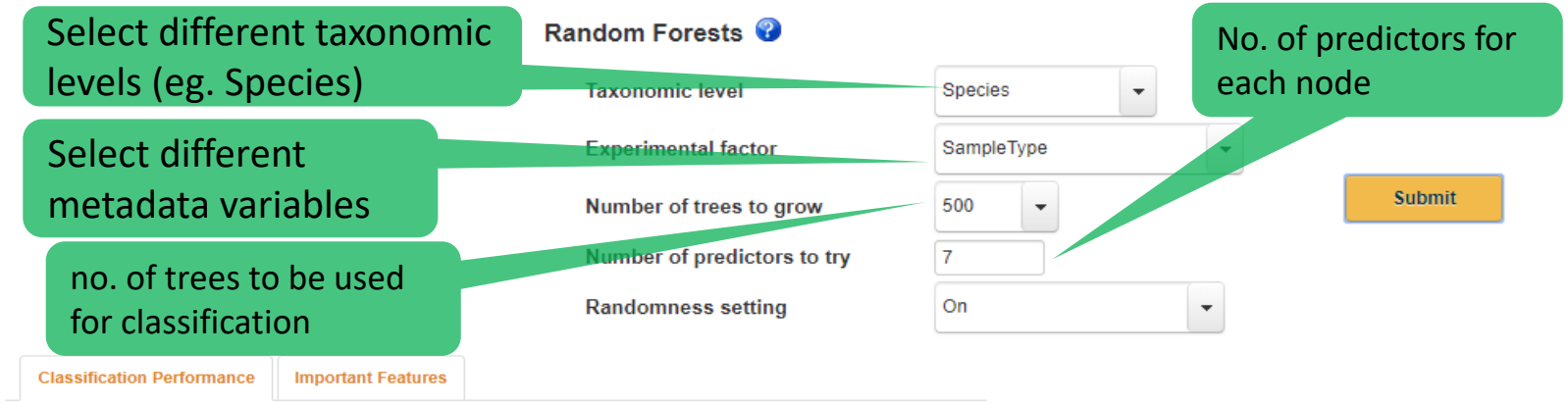
E. Biomarker Analysis

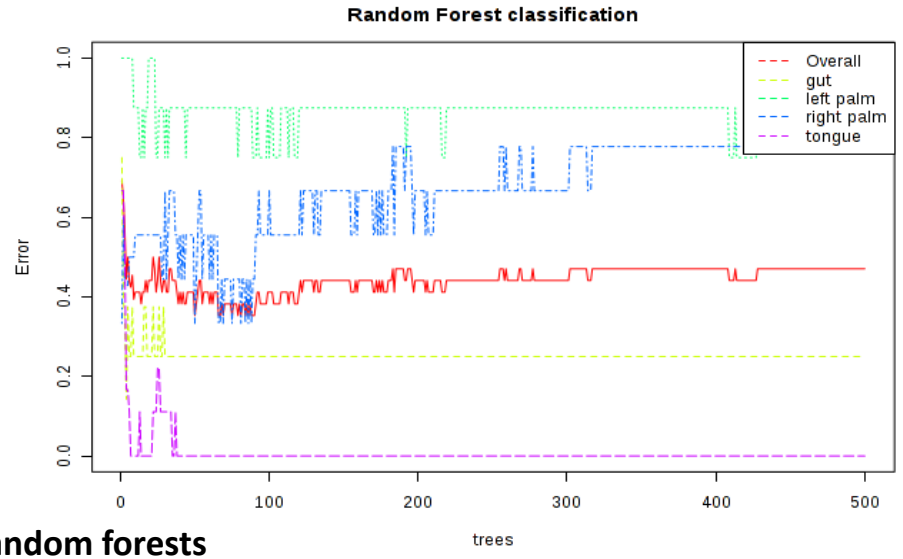
Select different taxonomic levels (eg. Species)

Select different metadata variables

no. of trees to be used for classification

No. of predictors for each node





The OOB error is 0.471

	gut	left palm	right palm	tongue	class.error
gut	6	0	2.0	0.0	0.25
left palm	0	1	6.0	1.0	0.875
right palm	2	5	2.0	0.0	0.778
tongue	0	0	0.0	9.0	0.0

2. 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.

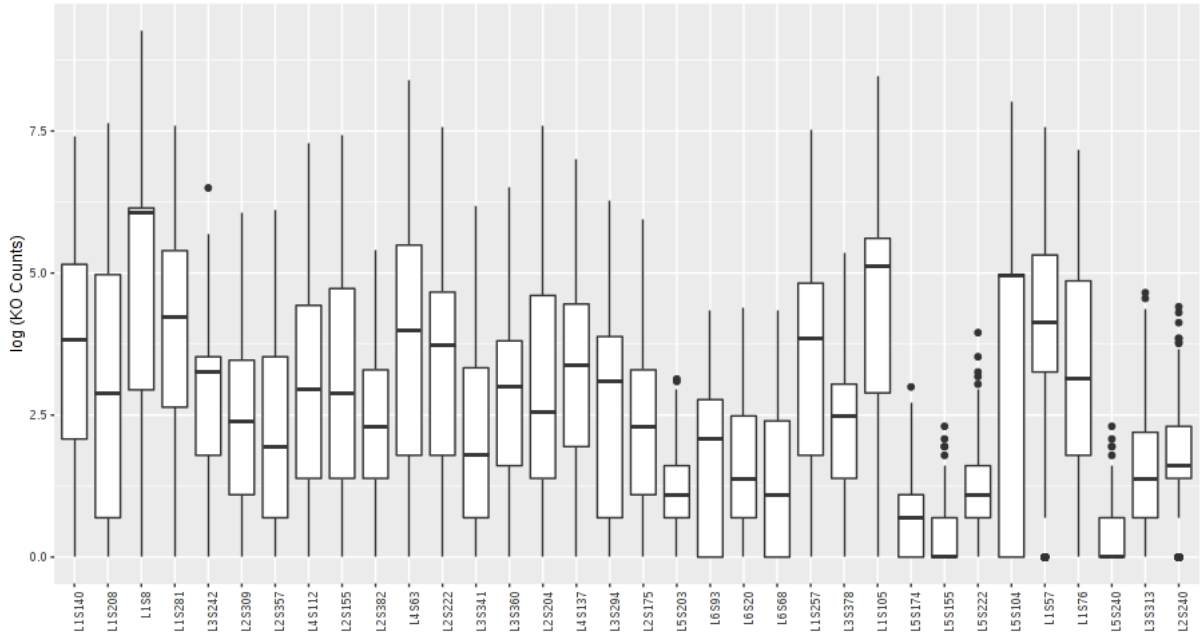
2. Random Forest

- It provides estimates of what variables are important in the classification of data.
- It computes proximities between pairs of cases that can be used in clustering, locating outliers, or give interesting views of the data.

F. Functional potential

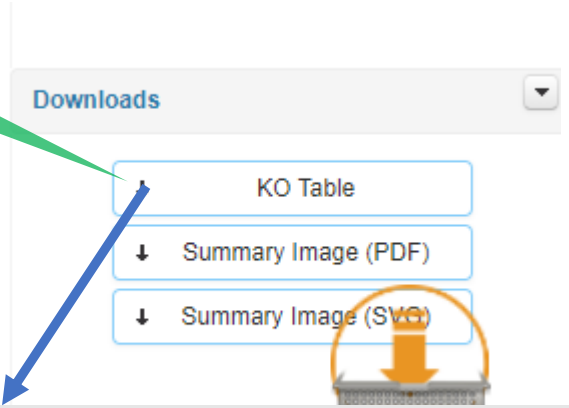
Predict Metabolic Potential

Click to predict potential functions



Count distribution od predicted metagenomic abundance data (KO counts) [log-scale]

Click to download KO table



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
	#NAME	L1S140	L1S208	L1S8	L1S281	L3S242	L2S309	L2S357	L4S112	L2S155	L2S382	L4S63	L2S222	L3S341
	K00001	183	231	1769	268	113	56	74	230	474	41	798	332	42
	K00002	15	33	0	13	0	12	28	45	7	7	169	40	2
	K00003	206	179	472	219	37	46	62	161	147	39	387	146	32
	K00004	0	0	0	0	0	1	2	11	95	2	20	43	1
	K00005	151	157	38	215	9	22	24	84	29	15	225	53	26
	K00007	0	2	0	0	0	0	1	4	5	0	7	2	0
	K00008	288	246	75	277	29	7	7	29	12	9	51	13	4
	K00009	13	16	23	18	8	1	2	3	6	0	6	2	2
	K00010	24	36	463	52	36	8	7	28	13	5	63	11	3
	K00011	44	8	434	6	26	0	0	2	1	0	4	2	0
	K00012	119	156	457	153	29	50	62	169	254	33	593	240	44
	K00013	176	173	33	194	8	50	60	157	148	34	494	179	33
	K00014	253	260	913	284	66	56	72	205	167	58	549	214	49
	K00015	0	2	0	0	0	19	4	23	18	2	109	26	6
	K00016	152	130	36	180	10	36	53	127	42	23	309	67	28
	K00018	68	32	27	59	4	4	6	24	8	12	20	37	14
	K00019	0	5	432	0	26	10	6	29	21	7	182	87	4
	K00020	131	111	437	106	28	60	63	193	263	14	789	204	40

After, prediction the result data is similar as shotgun metagenomic data.

- User have to go through the Shotgun Data Profiling module to perform comprehensive analysis.
- Please check, Tutorial II on (Shotgun data profiling) for stepwise detailed analysis on such data.

==THE END==