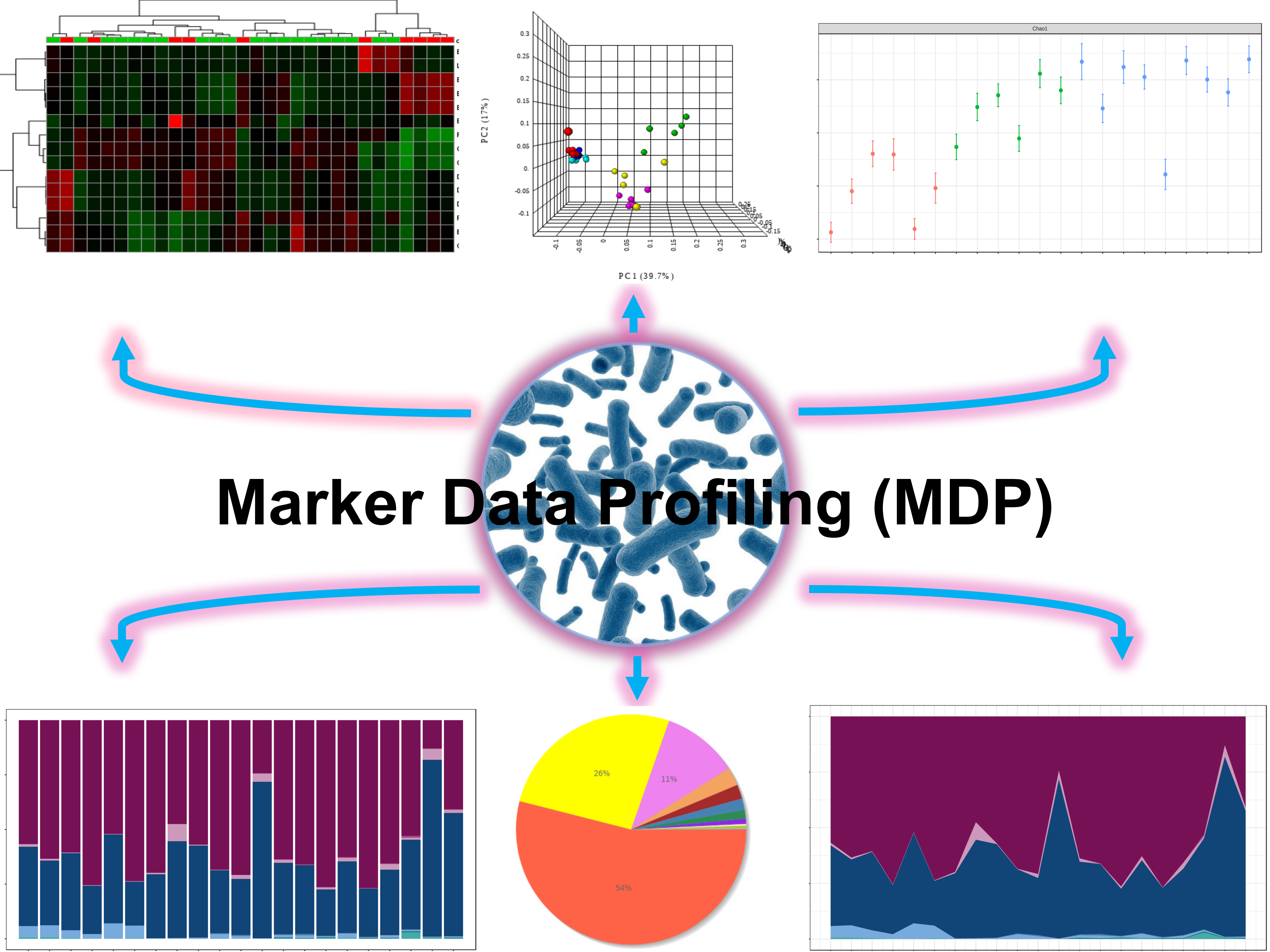
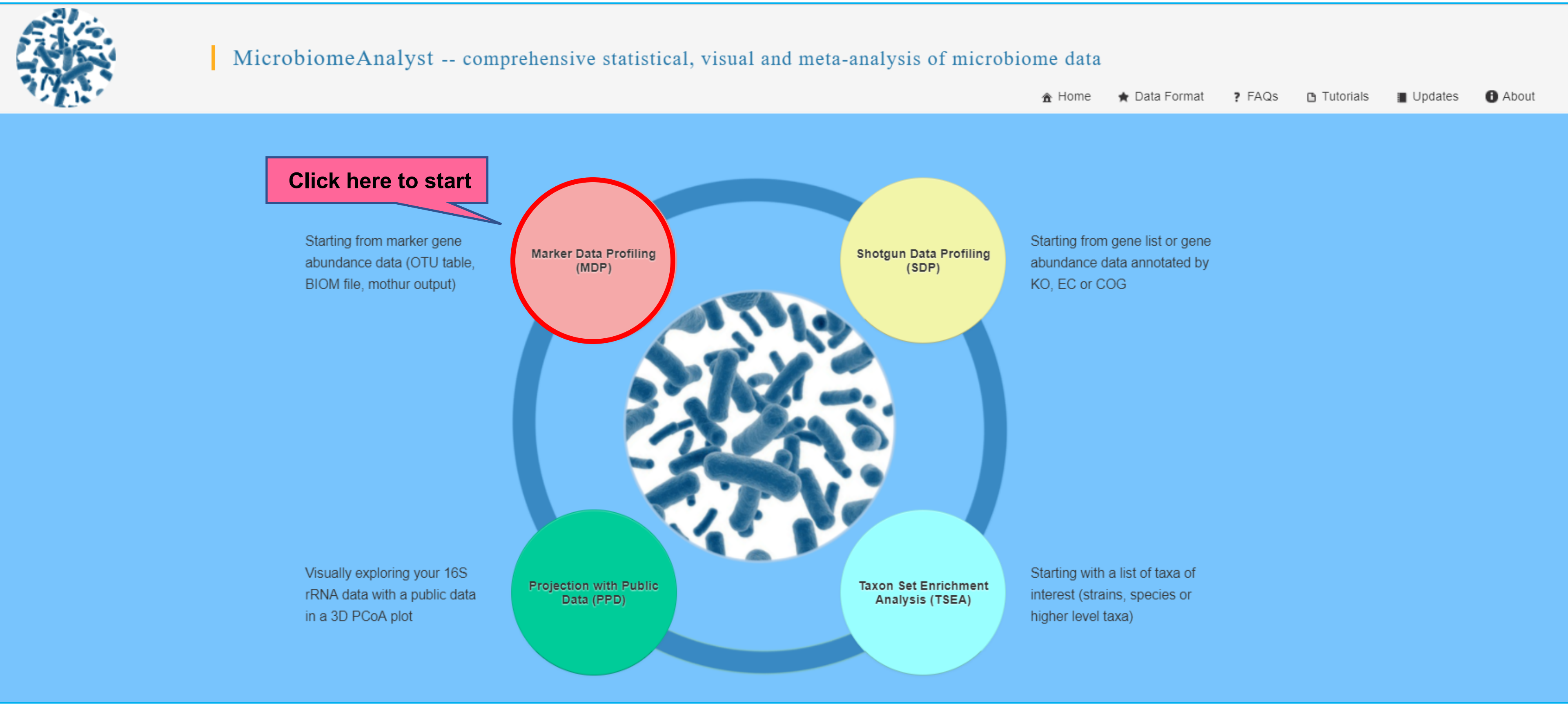


Marker Data Profiling (MDP)

Goal for this tutorial

- To perform a comprehensive analysis on a OTU table from 16S rRNA sequencing data, including:
  - ❖ **Diversity and compositional analysis**
  - ❖ **Comparative analysis**
  - ❖ **Predictions of metabolic potentials**


Home
Data Format
FAQs
Tutorials
Updates
About
Marker Data Profiling (MDP)
Shotgun Data Profiling (SDP)
Projection with Public Data (PPD)
Taxon Set Enrichment Analysis (TSEA)

MicrobiomeAnalyst -- comprehensive statistical, visual and meta-analysis of microbiome data

Click here to start

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

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

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

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.

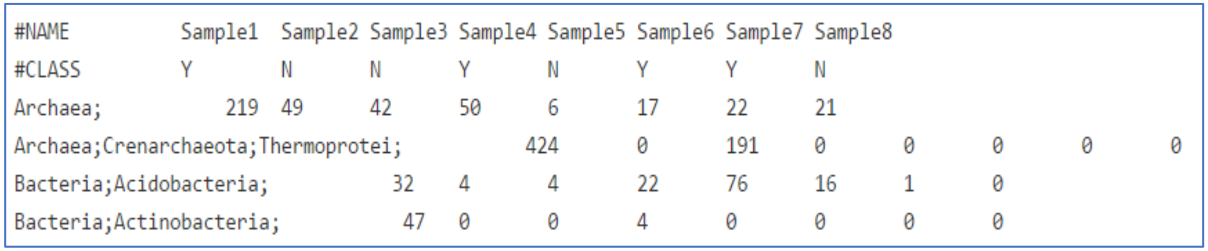
Details about each format are in the next few slides.

Data Formatting

1. **Tab-delimited text file**

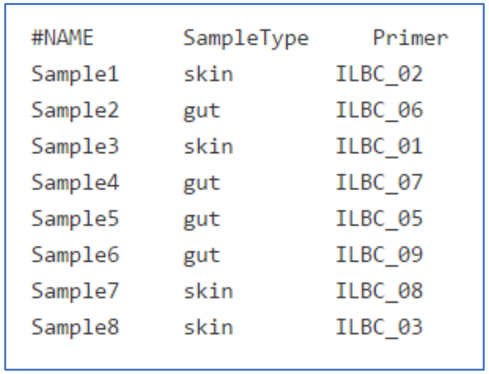
- Manipulate data headings in a spreadsheet program like MS Excel
- Save as a **tab delimited (.txt) or comma-separated (.csv) file**
- The headings **#NAME** (all capital letters) must be used
    - ❖ #NAME is for sample names (first column in abundance; first row in metadata file)
    - ❖ 2nd Column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

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**

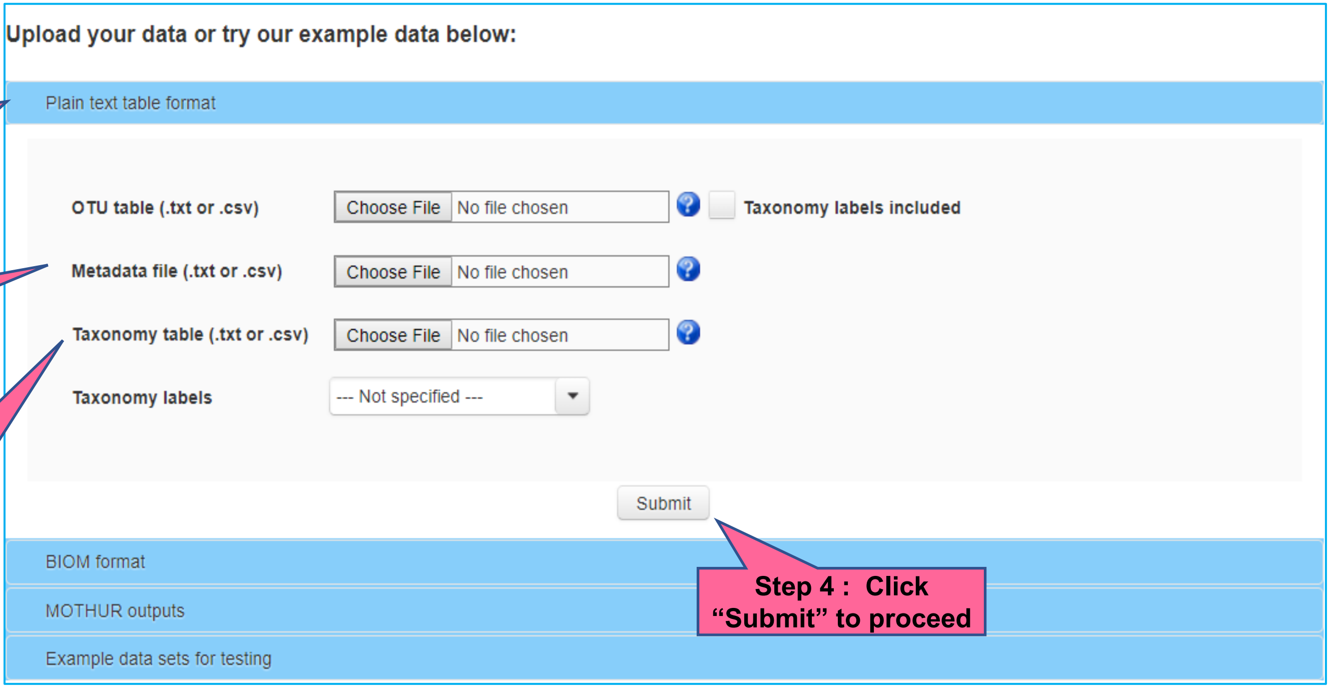
- Two files needed: a **consensus taxonomy** (taxonomy) file and a **.shared** (abundance) file.
- Metadata file can be uploaded separately.
  - For details, please visit the mothur home page (https://mothur.org/wiki/Main_Page).

1. Data Upload

Step 1: Upload your taxonomic profile data (supporting three formats)

Step 2: Upload your metadata file

Step 3: Upload your taxonomy table separately (if not present) and also specify the annotated taxonomic labels



Upload your data or try our example data below:

Plain text table format

OTU table (.txt or .csv)	Choose File No file chosen	□ Taxonomy labels included
Metadata file (.txt or .csv)	Choose File No file chosen	
Taxonomy table (.txt or .csv)	Choose File No file chosen	
Taxonomy labels	--- Not specified ---	

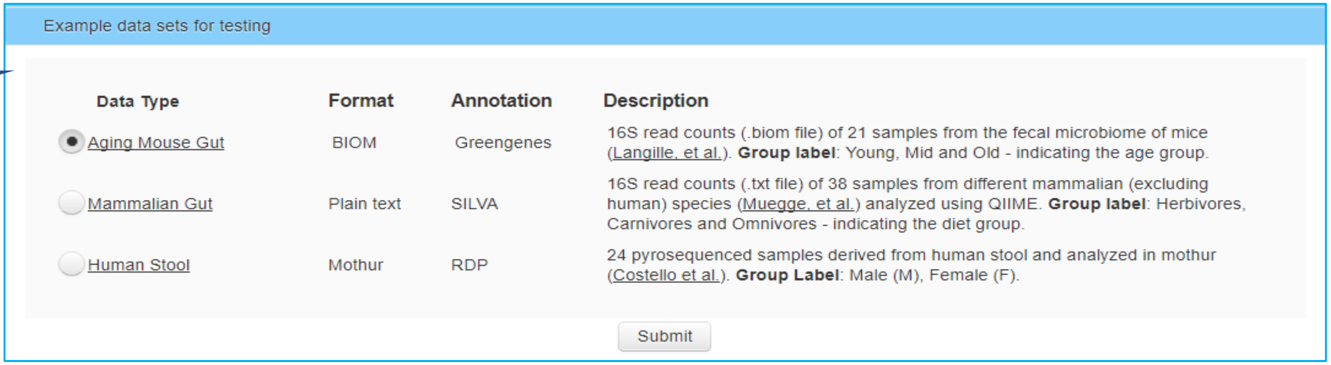
BIOM format

MOTHUR outputs

Example data sets for testing

Step 4 : Click “Submit” to proceed

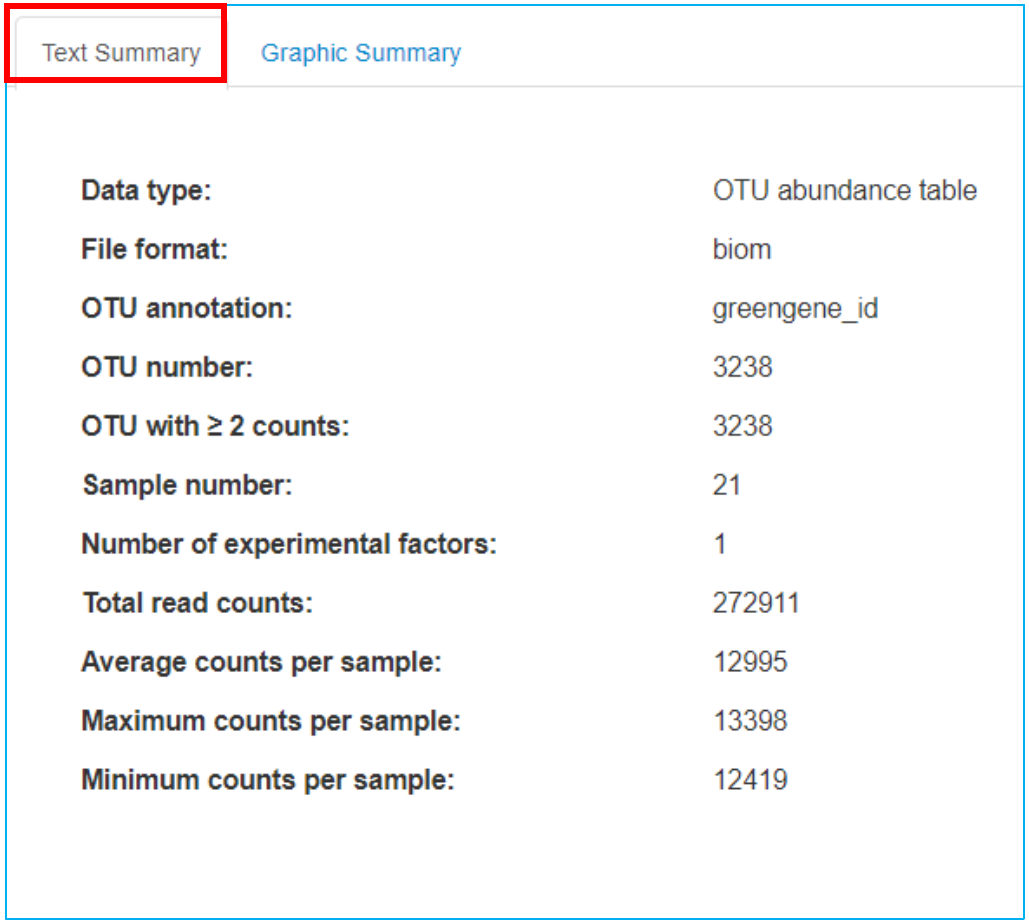
You can try our example also



Example data sets for testing

Data Type	Format	Annotation	Description
● Aging Mouse Gut	BIOM	Greengenes	16S read counts (.biom file) of 21 samples from the fecal microbiome of mice (Langille, et al.). **Group label**: Young, Mid and Old - indicating the age group.
○ Mammalian Gut	Plain text	SILVA	16S read counts (.txt file) of 38 samples from different mammalian (excluding human) species (Muegge, et al.) analyzed using QIIME. **Group label**: Herbivores, Carnivores and Omnivores - indicating the diet group.
○ Human Stool	Mothur	RDP	24 pyrosequenced samples derived from human stool and analyzed in mothur (Costello et al.). **Group Label**: Male (M), Female (F).

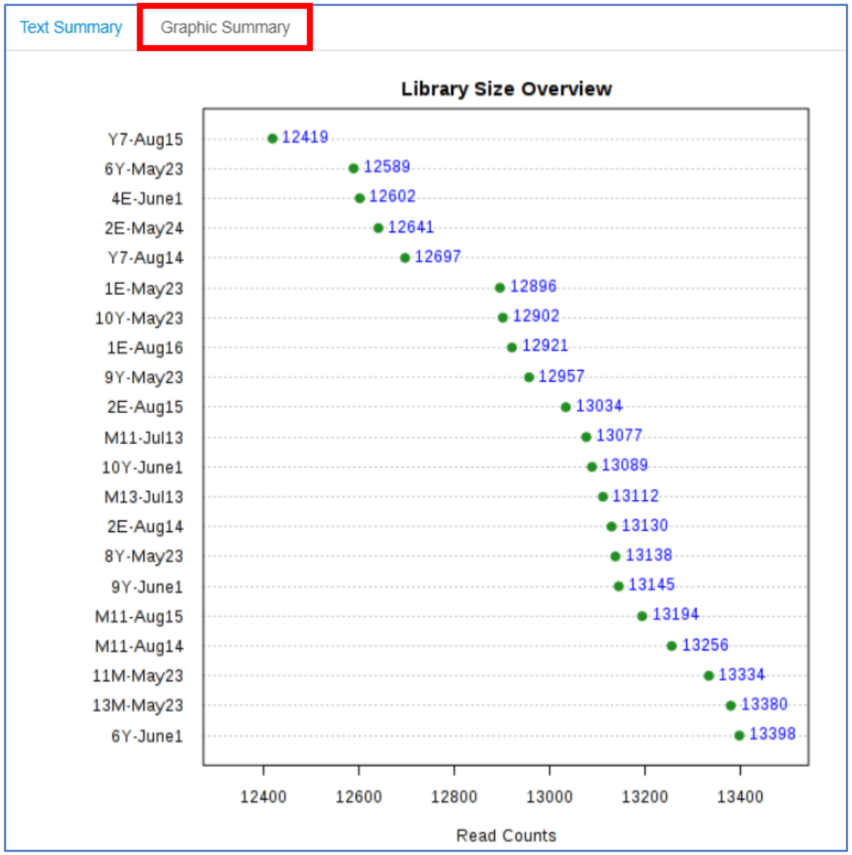
2. a) Data Integrity Check



Data type:	OTU abundance table
File format:	biom
OTU annotation:	greengene_id
OTU number:	3238
OTU with ≥ 2 counts:	3238
Sample number:	21
Number of experimental factors:	1
Total read counts:	272911
Average counts per sample:	12995
Maximum counts per sample:	13398
Minimum counts per sample:	12419

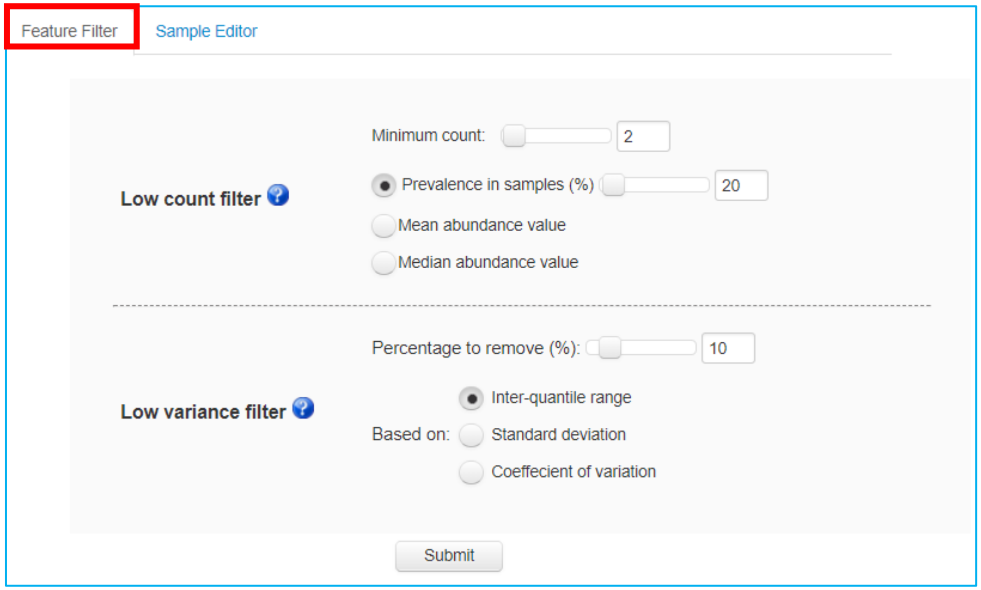
- Provides processing and summary information for user uploaded data.

2. b) Graphic Summary



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

3. a) Data Filtering (Features)



Feature Filter | Sample Editor

Low count filter

Minimum count: 2

● Prevalence in samples (%) 20

○ Mean abundance value

○ Median abundance value

Low variance filter

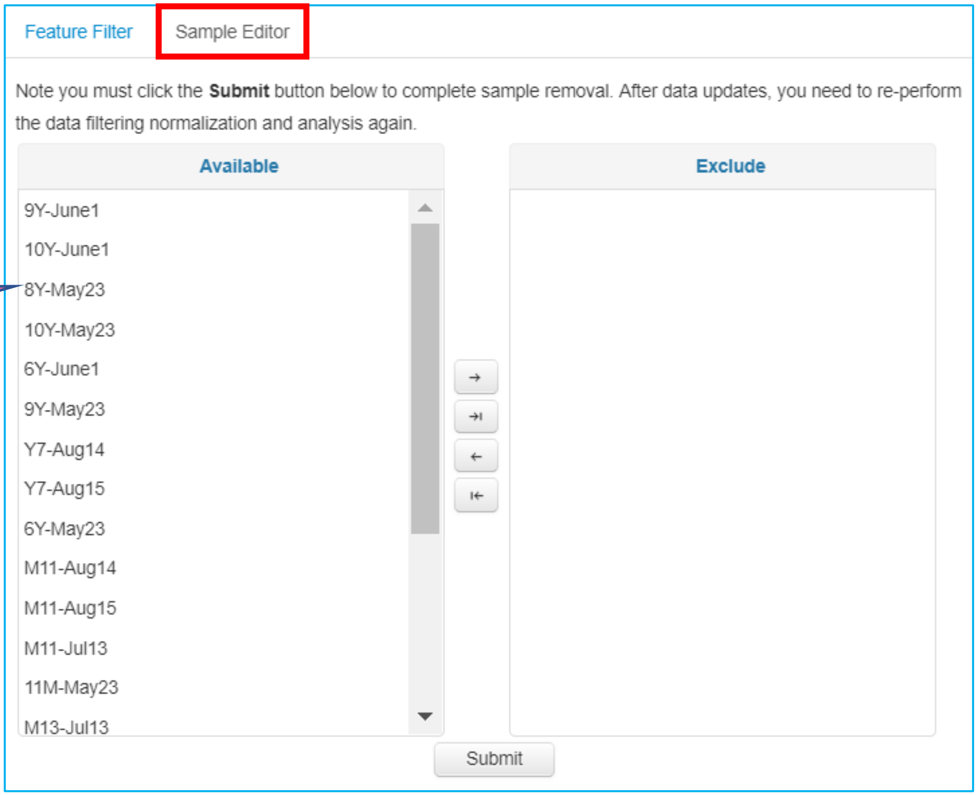
Percentage to remove (%): 10

Based on:
● Inter-quantile range
○ Standard deviation
○ Coeffecient of variation

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 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. b) Sample Filtering (Editor)

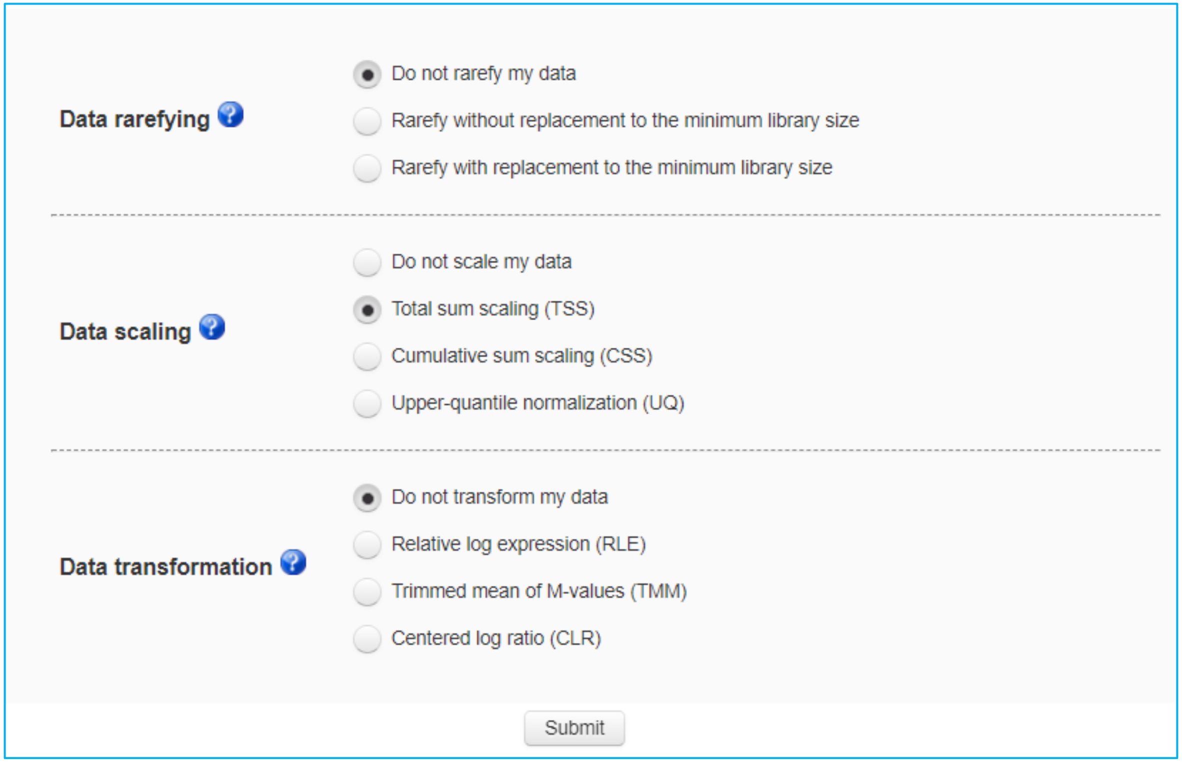


Note you must click the **Submit** button below to complete sample removal. After data updates, you need to re-perform the data filtering normalization and analysis again.

User can select samples to remove from downstream analysis

- Users can remove samples that are detected as outlier via graphical summary result or downstream analysis. (e.g. Beta-diversity analysis)

4. Data Normalization



Data rarefying

● Do not rarefy my data
○ Rarefy without replacement to the minimum library size
○ Rarefy with replacement to the minimum library size

Data scaling

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

Data transformation

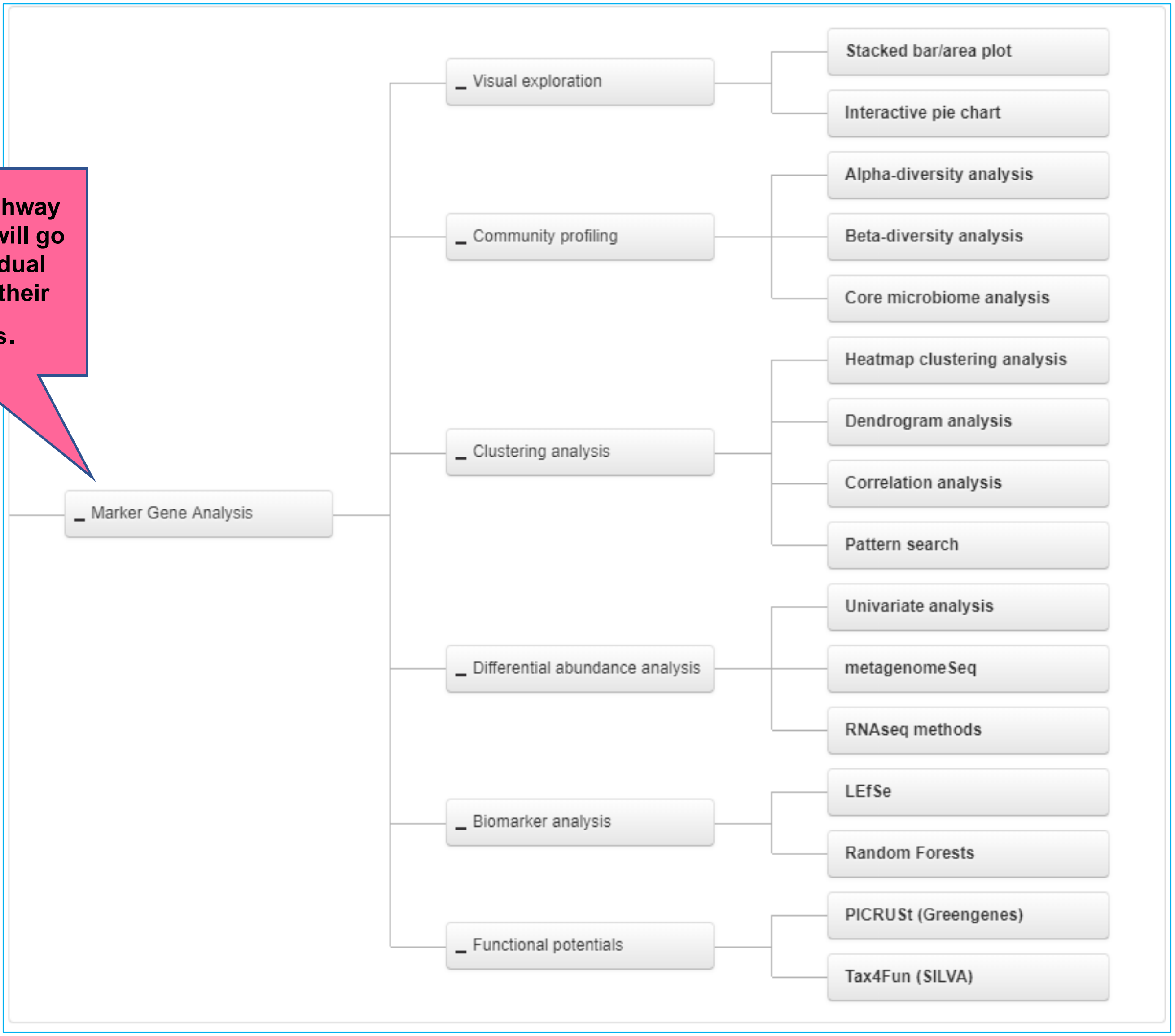
● Do not transform my data
○ Relative log expression (RLE)
○ Trimmed mean of M-values (TMM)
○ Centered log ratio (CLR)

- Normalizing is required to account for **uneven sequencing depth**, **under-sampling** 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)**

5. Data Analysis

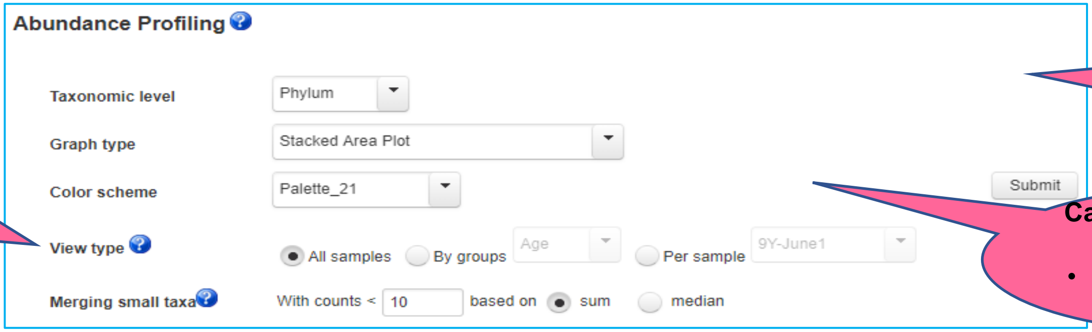
Six analysis pathway supported. We will go through individual pathways and their components.



- Marker Gene Analysis
  - Visual exploration
    - Stacked bar/area plot
    - Interactive pie chart
  - Community profiling
    - Alpha-diversity analysis
    - Beta-diversity analysis
    - Core microbiome analysis
  - Clustering analysis
    - Heatmap clustering analysis
    - Dendrogram analysis
    - Correlation analysis
    - Pattern search
  - Differential abundance analysis
    - Univariate analysis
    - metagenomeSeq
    - RNAseq methods
  - Biomarker analysis
    - LEfSe
    - Random Forests
  - Functional potentials
    - PICRUSt (Greengenes)
    - Tax4Fun (SILVA)

A. Visual Exploration

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



Abundance Profiling

Taxonomic level: Phylum

Graph type: Stacked Area Plot

Color scheme: Palette_21

Submit

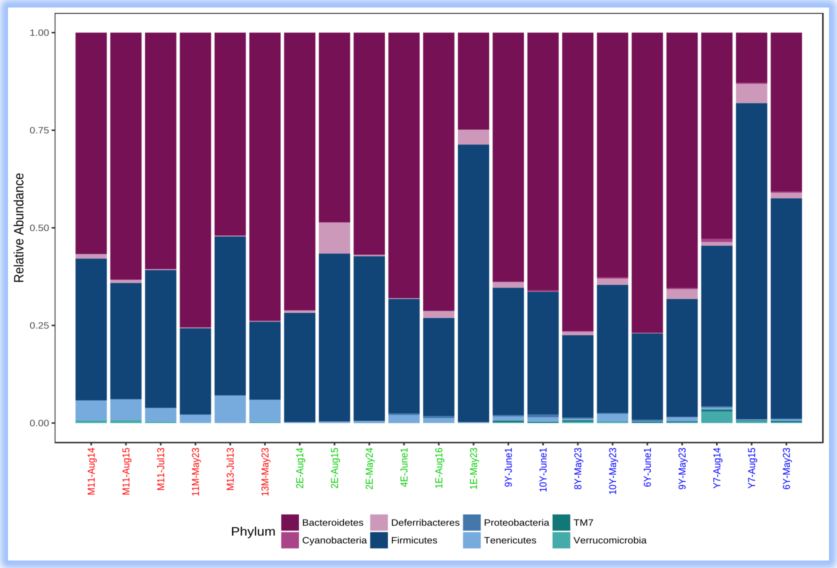
View type: ● All samples ○ By groups Age ○ Per sample 9Y-June1

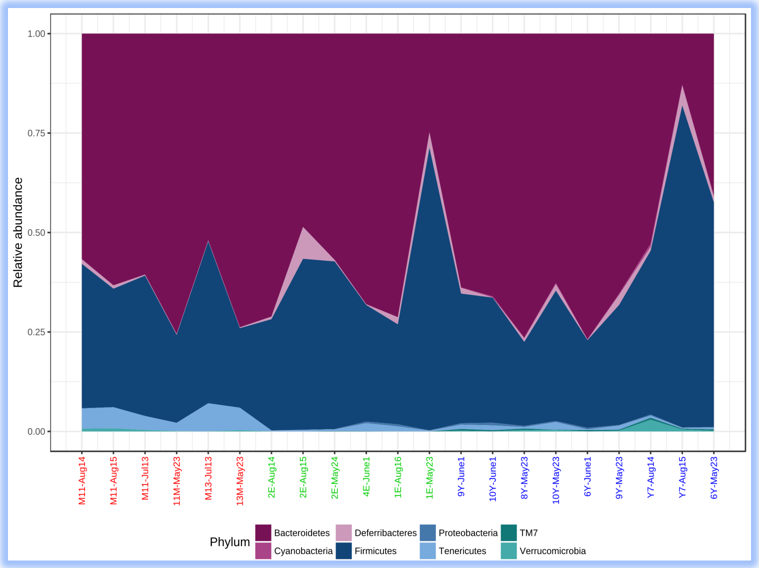
Merging small taxa: With counts < 10 based on ● sum ○ median

Chose different taxonomy level for plotting

Can be viewed in 3 ways:

- Bar graph
- Normalized Bar graph
- Stack Area plot





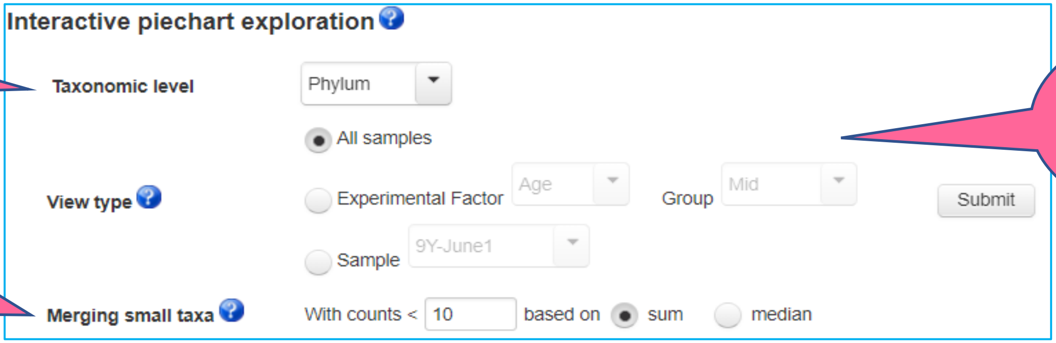
1. Stacked Bar/Area plot

- 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

Chose different taxonomy level for plotting

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



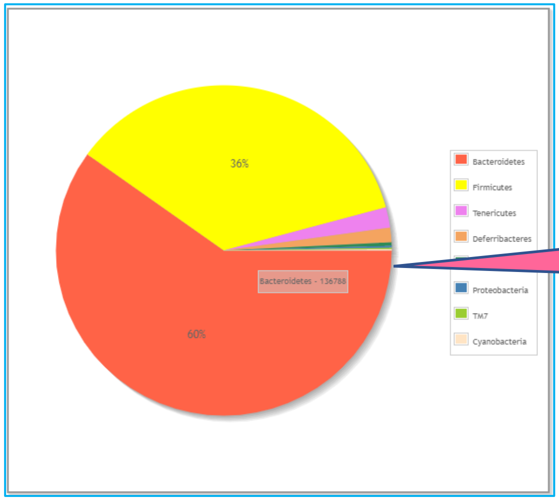
Interactive piechart exploration

Taxonomic level: Phylum

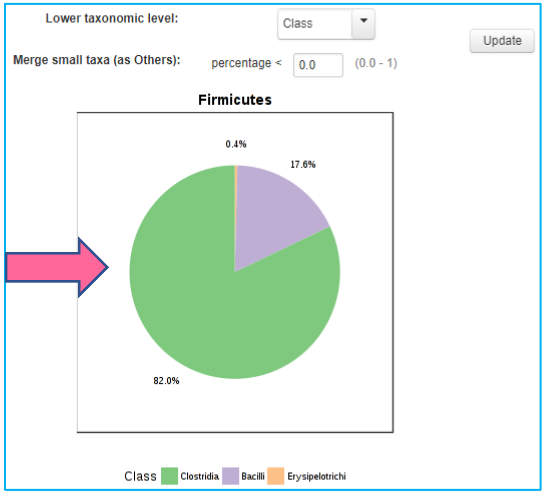
View type:
- ● All samples
- ○ Experimental Factor: Age — Group: Mid
- ○ Sample: 9Y-June1

Merging small taxa: With counts < 10 based on ● sum ○ median

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



Click on it for projection to lower taxonomic level



Lower taxonomic level: Class

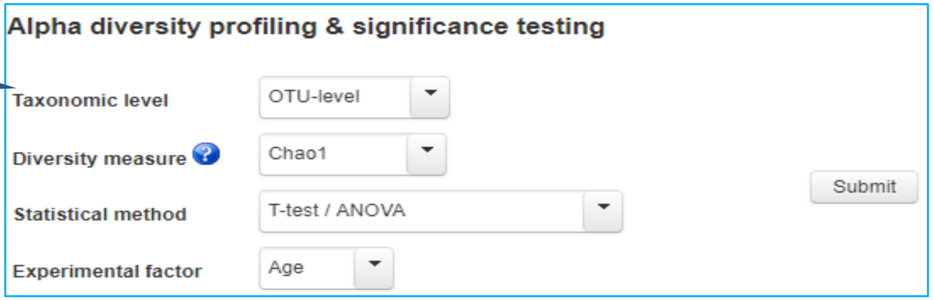
Merge small taxa (as Others): percentage < 0.0 (0.0 - 1)

2. 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)

B. Community Profiling

Chose different taxonomy level for analysis



Alpha diversity profiling & significance testing

Taxonomic level: OTU-level

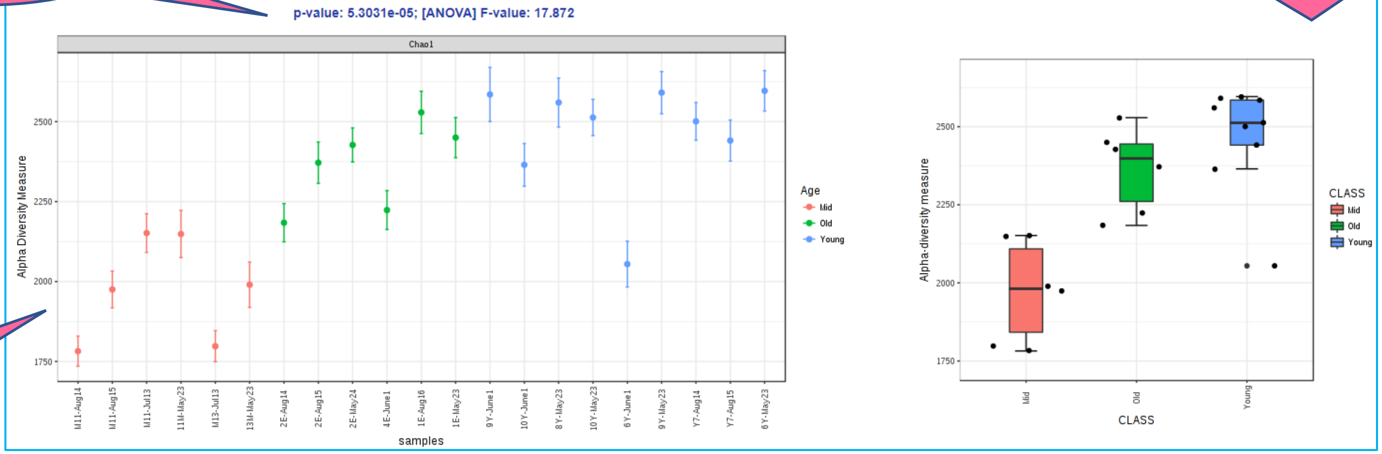
Diversity measure: Chao1

Statistical method: T-test / ANOVA

Experimental factor: Age

significance testing result

Sample group-wise diversity measure



p-value: 5.3031e-05; [ANOVA] F-value: 17.872

Sample-wise diversity measure

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** (evenness), **Observed** (richness), **Shannon** (account for both evenness and richness).
- Statistical significance testing between groups using parametric and non-parametric tests.

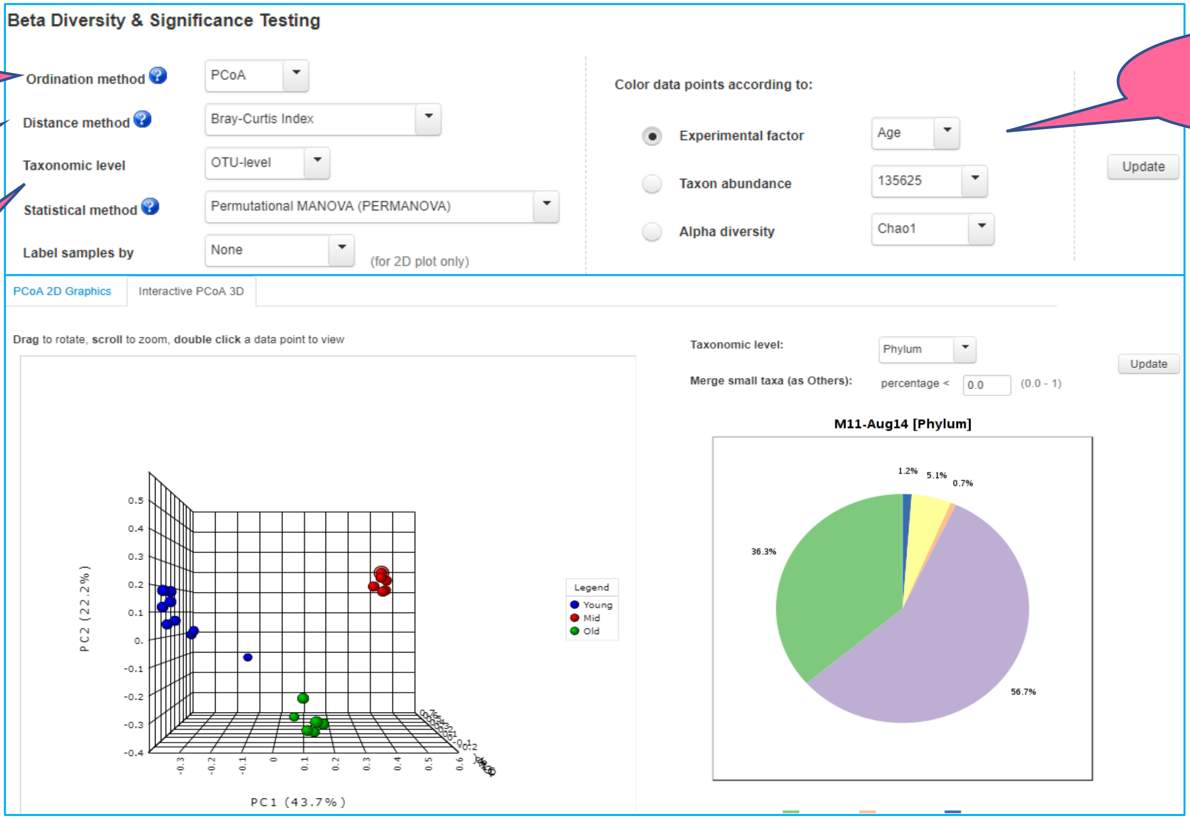
B. Community Profiling

2 ordination method supported: PCoA and NMDS

Chose from different distance methods

Chose different taxonomy level for analysis

Chose different



Beta Diversity & Significance Testing

Ordination method: PCoA

Distance method: Bray-Curtis Index

Taxonomic level: OTU-level

Statistical method: Permutational MANOVA (PERMANOVA)

Label samples by: None (for 2D plot only)

Color data points according to:

- ● Experimental factor: Age
- ○ Taxon abundance: 135625
- ○ Alpha diversity: Chao1

Update

PCoA 2D Graphics | Interactive PCoA 3D

Drag to rotate, scroll to zoom, double click a data point to view

Taxonomic level: Phylum

Merge small taxa (as Others): percentage < 0.0 (0.0 - 1)

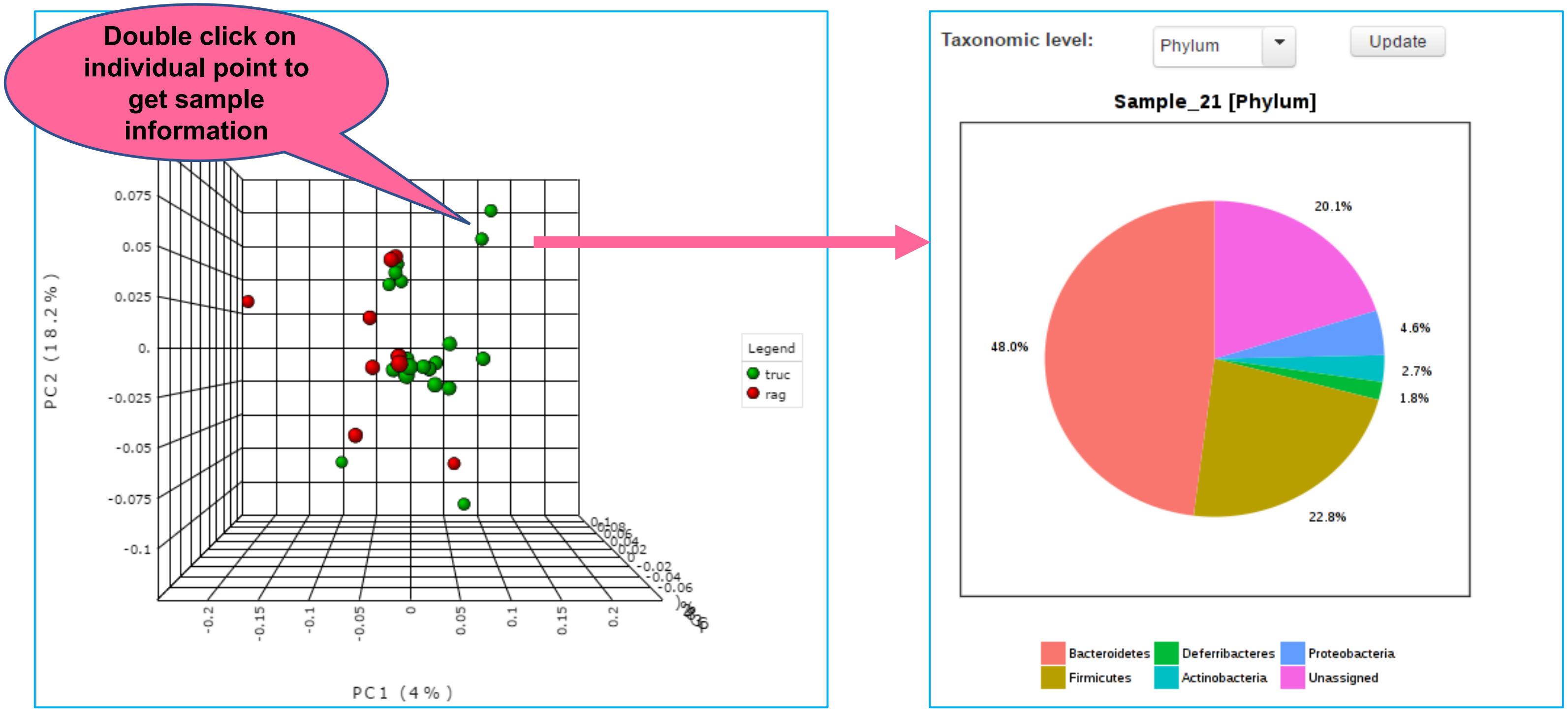
Update

M11-Aug14 [Phylum]

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.

B. Community Profiling



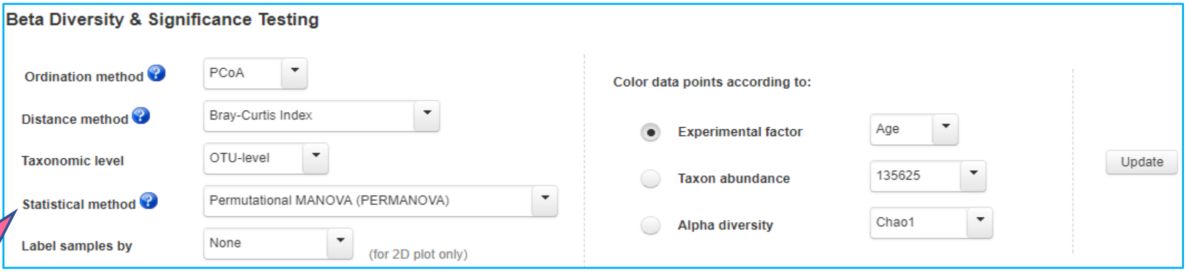
Double click on individual point to get sample information

Taxonomic level: Phylum ▾ Update

2. Beta diversity analysis & significance testing

- Results of PCoA/NMDS analysis can be visualized in **3D** using **ordination-based** distances supported.

B. Community Profiling



Beta Diversity & Significance Testing

Ordination method: PCoA

Distance method: Bray-Curtis Index

Taxonomic level: OTU-level

Statistical method: Permutational MANOVA (PERMANOVA)

Label samples by: None (for 2D plot only)

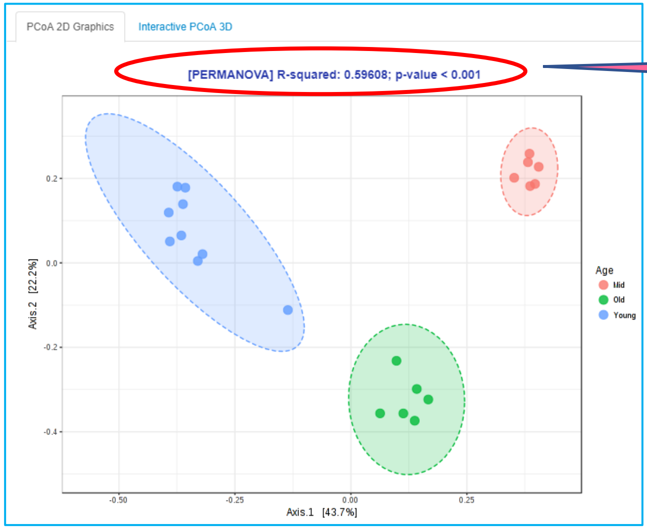
Color data points according to:

● Experimental factor: Age

○ Taxon abundance: 135625

○ Alpha diversity: Chao1

Chose from different statistical methods for significance testing (3 supported)



[PERMANOVA] R-squared: 0.59608; p-value < 0.001

Significance testing results

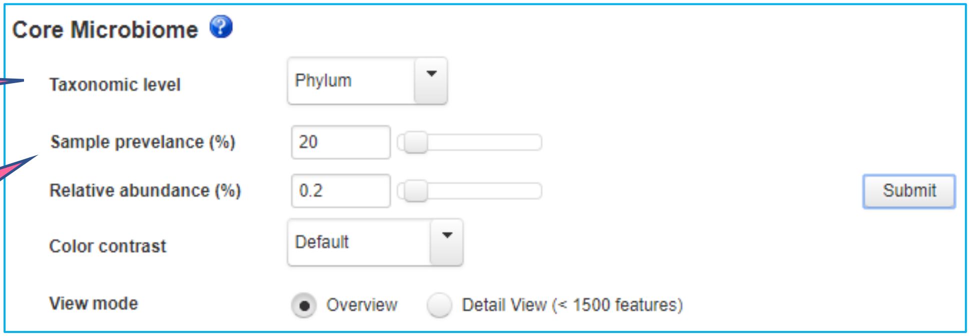
2. Beta diversity analysis & significance testing

- 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

Chose different taxonomy level for analysis

User can chose their own sample prevalence (%) as well as relative abundance for classification of core taxa



Core Microbiome

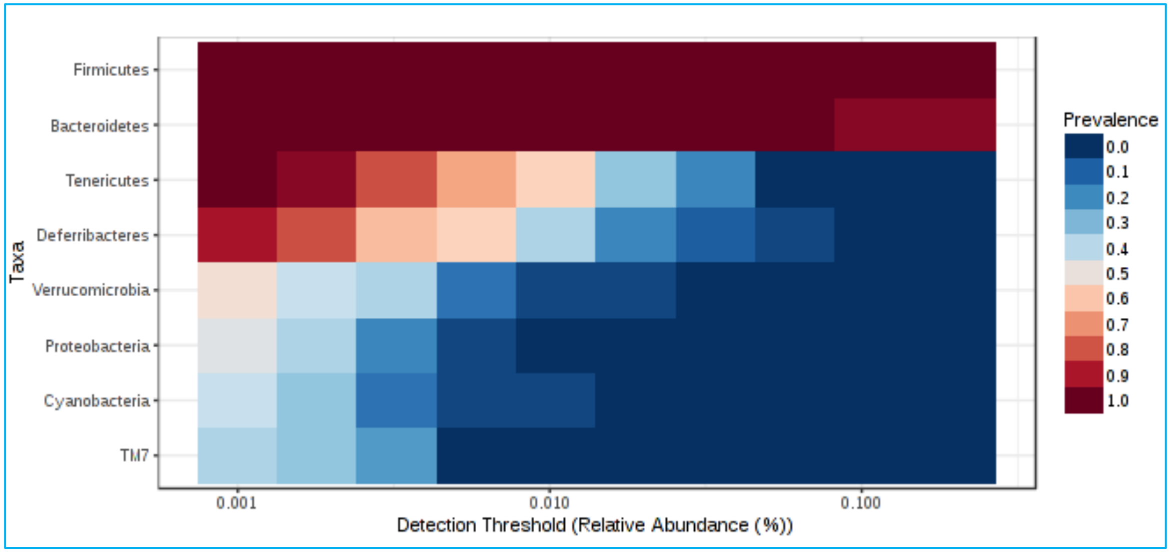
Taxonomic level: Phylum

Sample prevelance (%): 20

Relative abundance (%): 0.2

Color contrast: Default

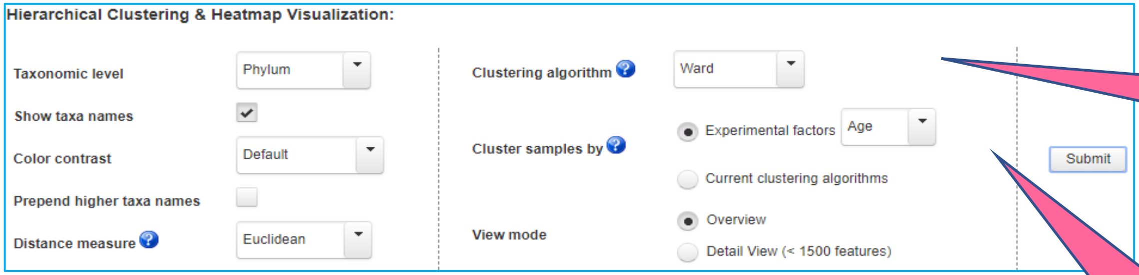
View mode: ● Overview ○ Detail View (< 1500 features)



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



Hierarchical Clustering & Heatmap Visualization:

Taxonomic level: Phylum

Show taxa names: ☑

Color contrast: Default

Prepend higher taxa names: □

Distance measure: Euclidean

Clustering algorithm: Ward

Cluster samples by: ● Experimental factors: Age ○ Current clustering algorithms

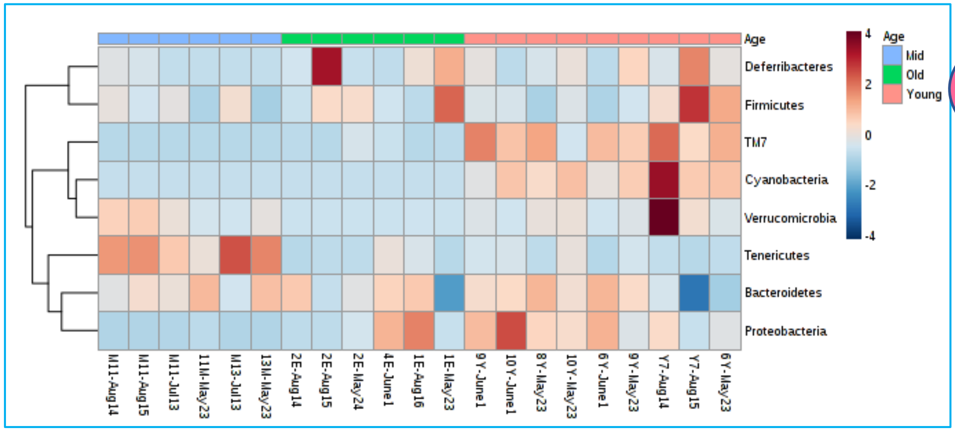
View mode: ● Overview ○ Detail View (< 1500 features)

Chose different taxonomy levels.

Chose from different distance measure.

Chose from different clustering algorithm.

Samples can be clustered based on either clustering algorithm or selected experimental factor



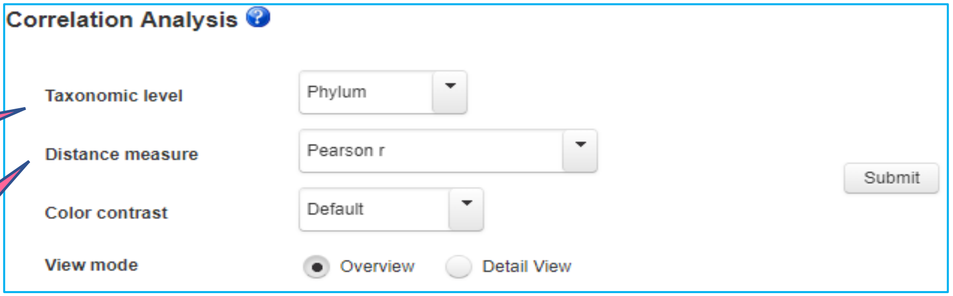
1. Heatmap and clustering analysis

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

Chose different taxonomy levels.

3 most common method supported for performing correlation analysis



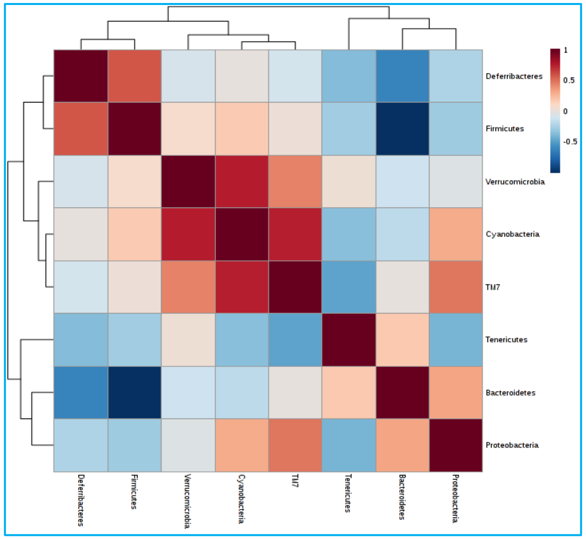
Correlation Analysis

Taxonomic level: Phylum

Distance measure: Pearson r

Color contrast: Default

View mode: ● Overview ○ Detail View



2. Correlation analysis

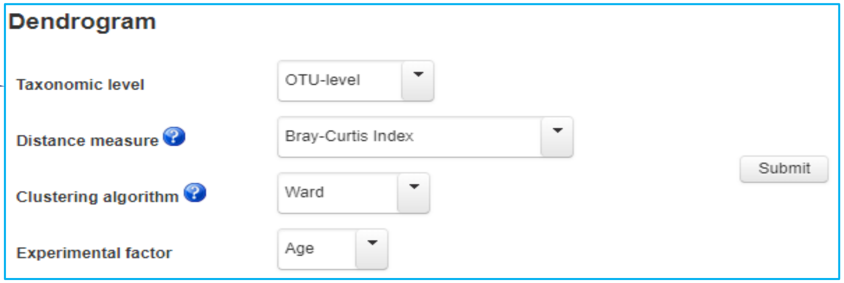
- Helps in identifying biologically or biochemically meaningful relationship or associations between taxa or features.
- Can be analyzed at various level (Phylum to OTU) by merging data based on taxonomic rank.

C. Clustering analysis

Data can be merged at different taxonomy levels.

Chose from different distance measure.

Chose from different clustering algorithm.



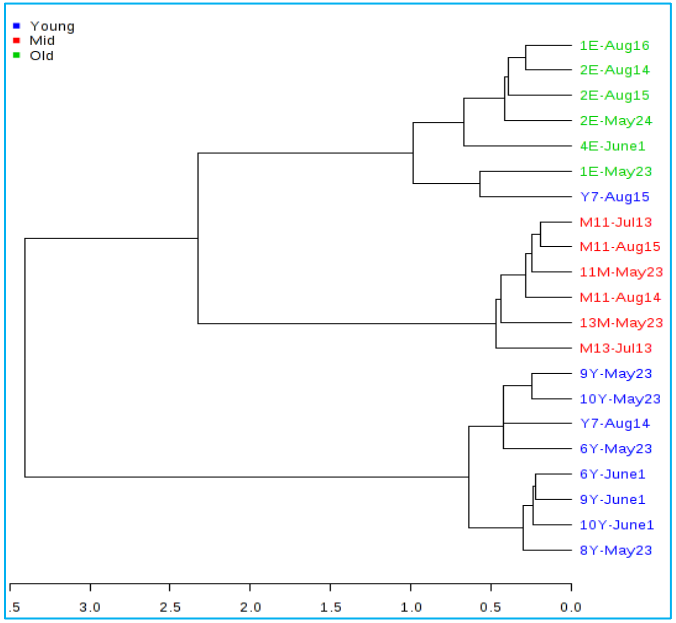
Dendrogram

Taxonomic level: OTU-level

Distance measure: Bray-Curtis Index

Clustering algorithm: Ward

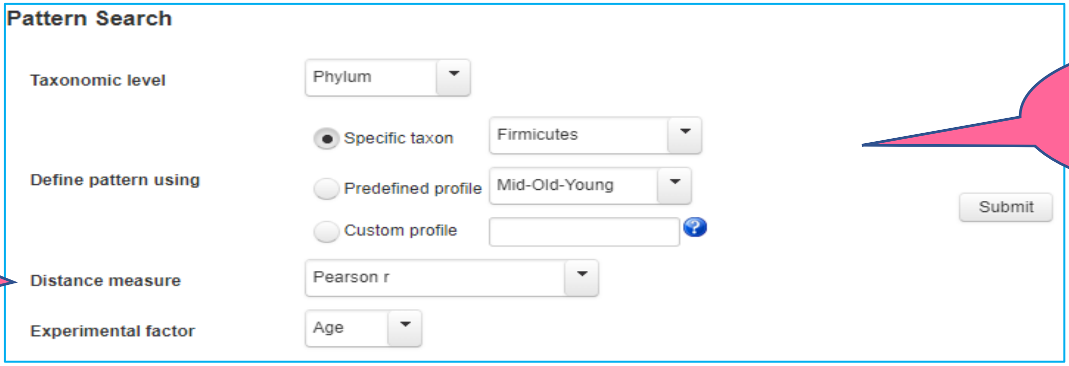
Experimental factor: Age



3. Dendrogram and clustering analysis

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

C. Clustering analysis



Pattern Search

Taxonomic level: Phylum

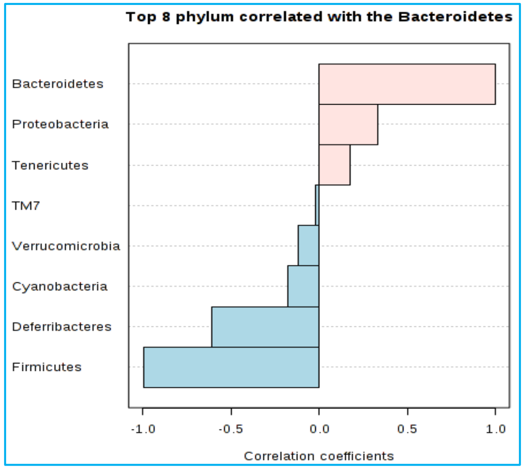
Define pattern using:
- ● Specific taxon: Firmicutes
- ○ Predefined profile: Mid-Old-Young
- ○ Custom profile:

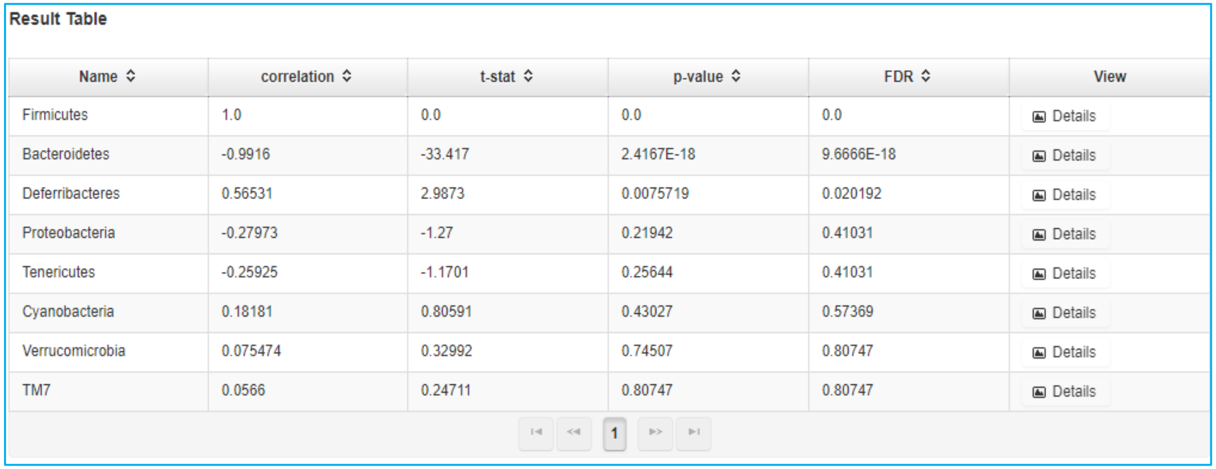
Distance measure: Pearson r

Experimental factor: Age

User can define their own pattern based on their interest

3 most common method supported for performing correlation analysis





Result Table

Name	correlation	t-stat	p-value	FDR	View
Firmicutes	1.0	0.0	0.0	0.0	Details
Bacteroidetes	-0.9916	-33.417	2.4167E-18	9.6666E-18	Details
Deferribacteres	0.56531	2.9873	0.0075719	0.020192	Details
Proteobacteria	-0.27973	-1.27	0.21942	0.41031	Details
Tenericutes	-0.25925	-1.1701	0.25644	0.41031	Details
Cyanobacteria	0.18181	0.80591	0.43027	0.57369	Details
Verrucomicrobia	0.075474	0.32992	0.74507	0.80747	Details
TM7	0.0566	0.24711	0.80747	0.80747	Details

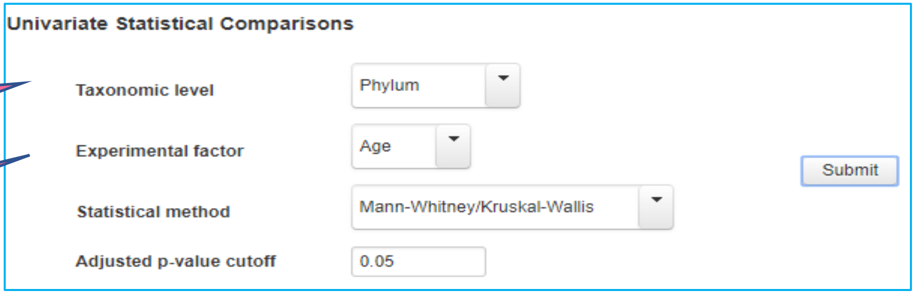
4. Pattern Search

- 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

Features can be merged at different taxonomic level

Chose from different Experimental factors



Univariate Statistical Comparisons

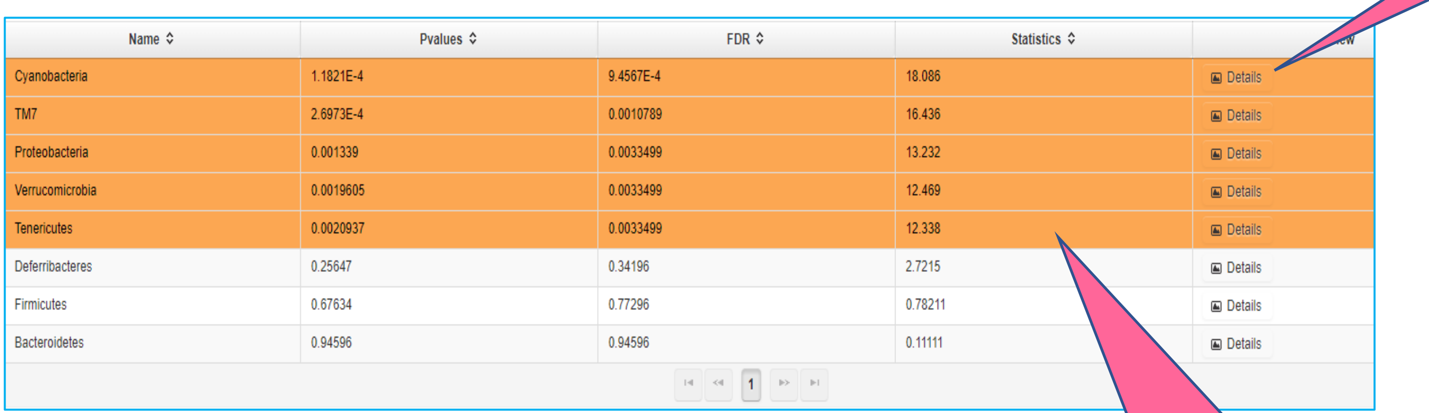
Taxonomic level: Phylum

Experimental factor: Age

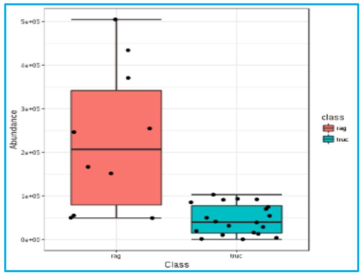
Statistical method: Mann-Whitney/Kruskal-Wallis

Adjusted p-value cutoff: 0.05

Click on "Details" to see group-wise data distribution for each individual feature



Name	Pvalues	FDR	Statistics	ew
Cyanobacteria	1.1821E-4	9.4567E-4	18.086	Details
TM7	2.6973E-4	0.0010789	16.436	Details
Proteobacteria	0.001339	0.0033499	13.232	Details
Verrucomicrobia	0.0019605	0.0033499	12.469	Details
Tenericutes	0.0020937	0.0033499	12.338	Details
Deferribacteres	0.25647	0.34196	2.7215	Details
Firmicutes	0.67634	0.77296	0.78211	Details
Bacteroidetes	0.94596	0.94596	0.11111	Details



Differential abundant taxa are highlighted in orange color

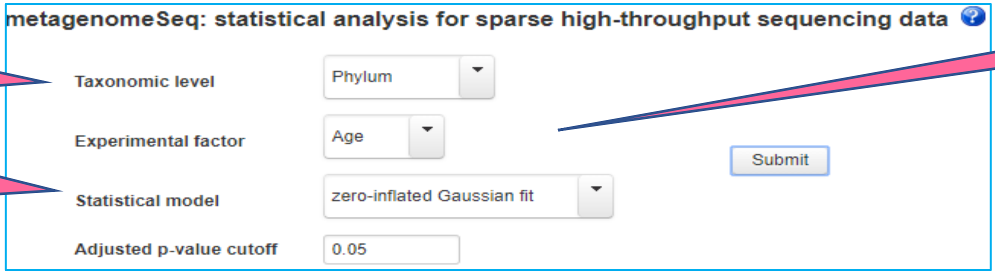
1. Univariate Statistical Comparisons

- 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 non parametric test options.
- P-values adjusted using **FDR** method.

D. Differential Abundance analysis

Features can be merged at different taxonomic level

Chose from 2 statistical models based on number of groups



metagenomeSeq: statistical analysis for sparse high-throughput sequencing data

Taxonomic level: Phylum

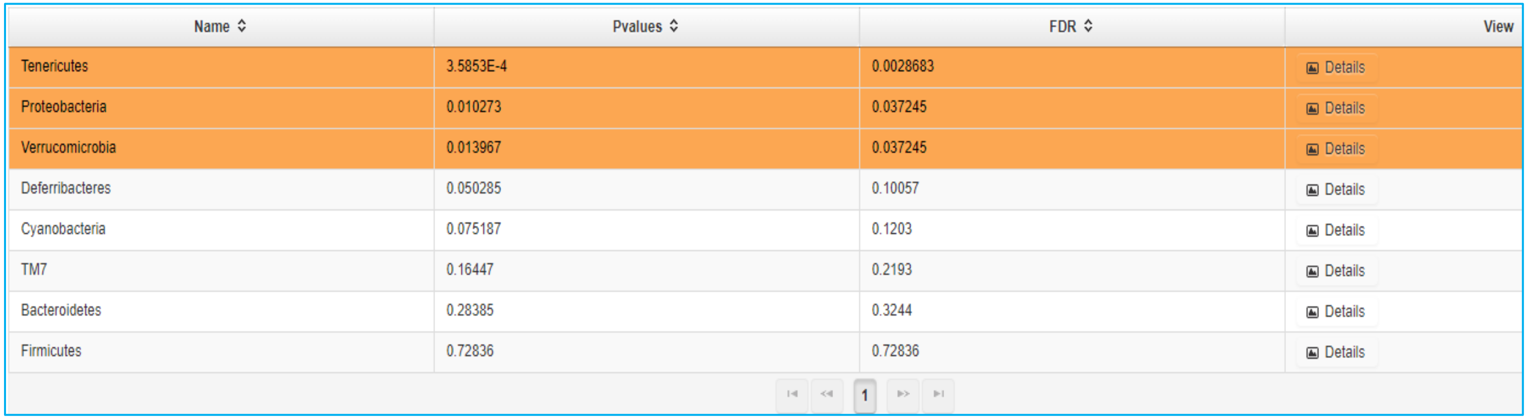
Experimental factor: Age

Statistical model: zero-inflated Gaussian fit

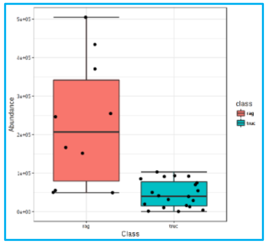
Adjusted p-value cutoff: 0.05

Chose from different Experimental factors

Click on “Details” to see group-wise data distribution for each individual feature



Name	Pvalues	FDR	View
Tenericutes	3.5853E-4	0.0028683	Details
Proteobacteria	0.010273	0.037245	Details
Verrucomicrobia	0.013967	0.037245	Details
Deferribacteres	0.050285	0.10057	Details
Cyanobacteria	0.075187	0.1203	Details
TM7	0.16447	0.2193	Details
Bacteroidetes	0.28385	0.3244	Details
Firmicutes	0.72836	0.72836	Details



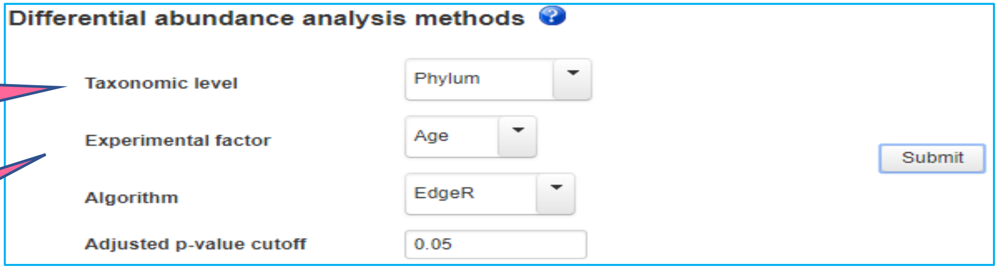
2. metagenomeSeq

- Detect differential abundant features in microbiome experiments with an explicit design.
- Accounts for **under-sampling** and **sparsity** in such data.
- Performs zero-inflated Gaussian fit (**fitZIG**) or fit-Feature (**fitFeature**) on data after normalizing the data through **cumulative sum scaling** (CSS) method (novel approach)
- fitFeature model is recommended over fitZIG for two groups comparison.
- Very sensitive and specific in nature (fails with very low sample size)

D. Differential Abundance analysis

Features can be merged at different taxonomic level

Chose from different Experimental factors



Differential abundance analysis methods

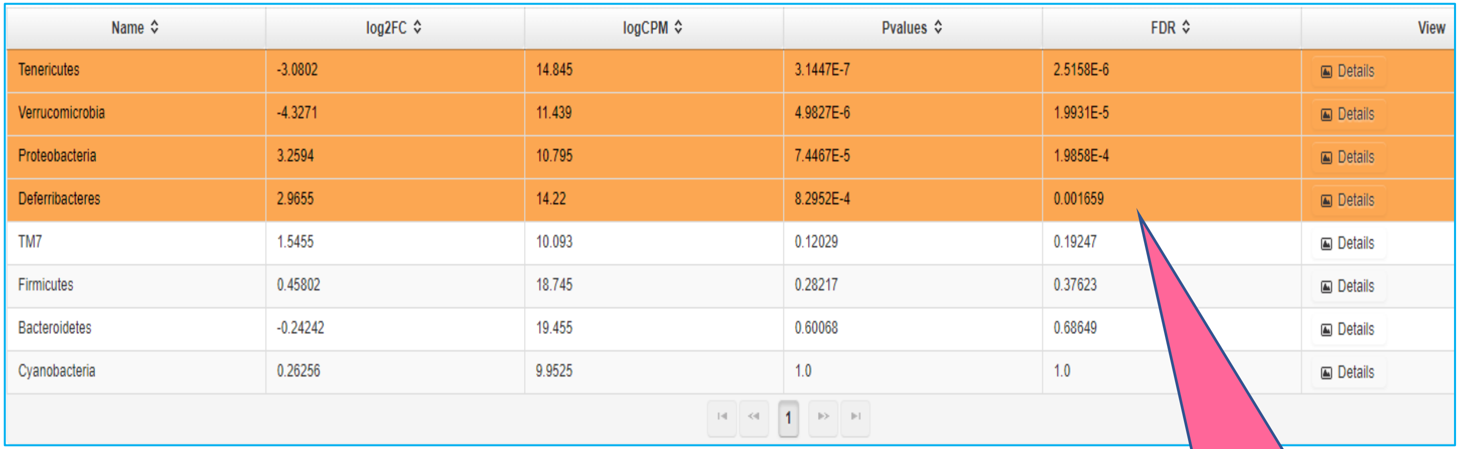
Taxonomic level: Phylum

Experimental factor: Age

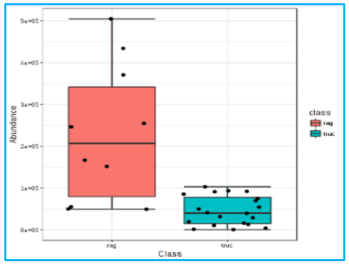
Algorithm: EdgeR

Adjusted p-value cutoff: 0.05

Click on “Details” to see group-wise data distribution for each individual feature



Name	log2FC	logCPM	Pvalues	FDR	View
Tenericutes	-3.0802	14.845	3.1447E-7	2.5158E-6	Details
Verrucomicrobia	-4.3271	11.439	4.9827E-6	1.9931E-5	Details
Proteobacteria	3.2594	10.795	7.4467E-5	1.9858E-4	Details
Deferribacteres	2.9655	14.22	8.2952E-4	0.001659	Details
TM7	1.5455	10.093	0.12029	0.19247	Details
Firmicutes	0.45802	18.745	0.28217	0.37623	Details
Bacteroidetes	-0.24242	19.455	0.60068	0.68649	Details
Cyanobacteria	0.26256	9.9525	1.0	1.0	Details



Differential abundant taxa are highlighted in orange color

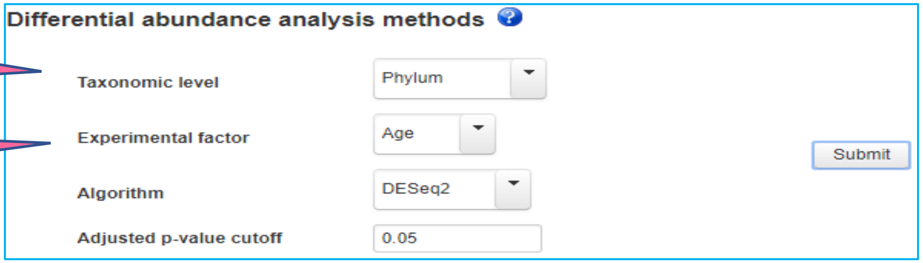
3. EdgeR

- 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.

D. Differential Abundance analysis

Features can be merged at different taxonomic level

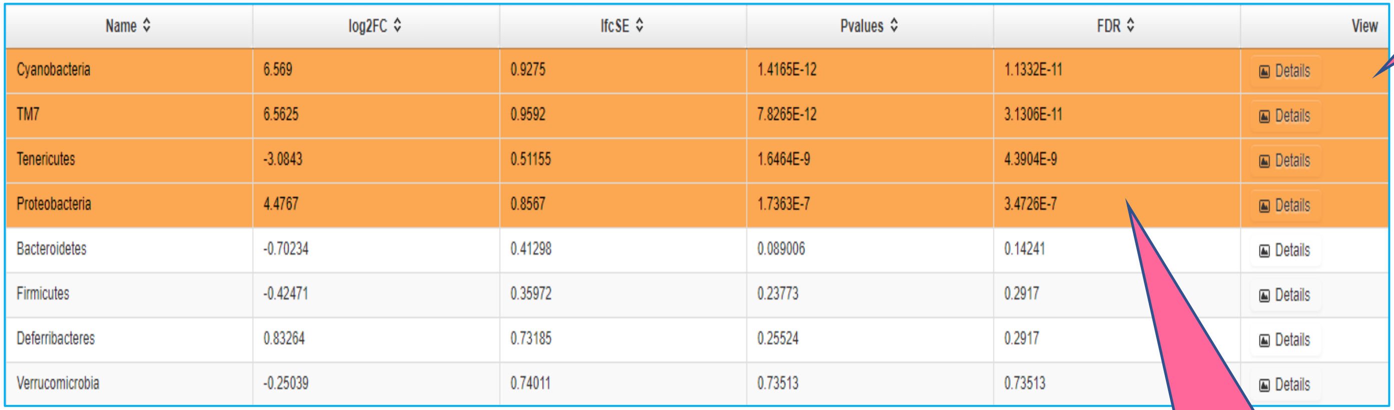
Chose from different Experimental factors



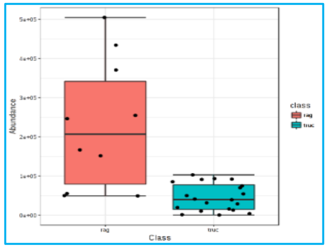
Differential abundance analysis methods

Taxonomic level	Phylum
Experimental factor	Age
Algorithm	DESeq2
Adjusted p-value cutoff	0.05

Click on “View Data” to see group-wise data distribution for each individual feature



Name	log2FC	lfcSE	Pvalues	FDR	View
Cyanobacteria	6.569	0.9275	1.4165E-12	1.1332E-11	Details
TM7	6.5625	0.9592	7.8265E-12	3.1306E-11	Details
Tenericutes	-3.0843	0.51155	1.6464E-9	4.3904E-9	Details
Proteobacteria	4.4767	0.8567	1.7363E-7	3.4726E-7	Details
Bacteroidetes	-0.70234	0.41298	0.089006	0.14241	Details
Firmicutes	-0.42471	0.35972	0.23773	0.2917	Details
Deferribacteres	0.83264	0.73185	0.25524	0.2917	Details
Verrucomicrobia	-0.25039	0.74011	0.73513	0.73513	Details

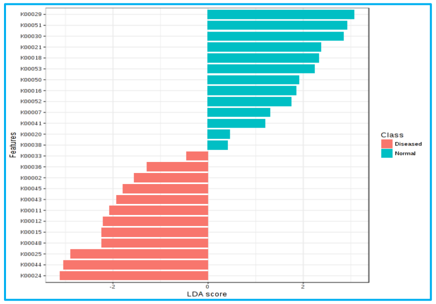


Differential abundant taxa are highlighted in orange color

4. DESeq2

- Developed for RNAseq data analysis.
- Uses negative binomial generalized linear models to estimate **dispersion** and **logarithmic fold changes**.

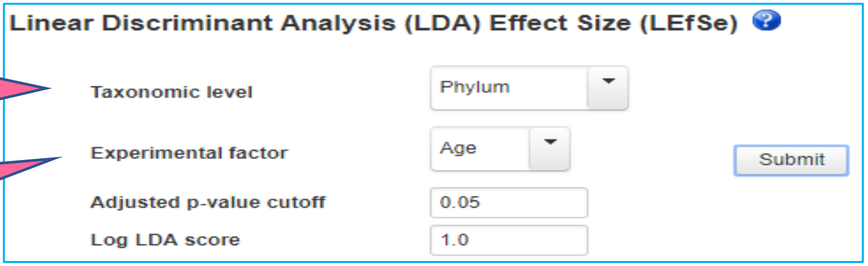
E. Biomarker Analysis



Click here to view Effect size of differential features

Features can be merged at different taxonomic level

Chose from different Experimental factors



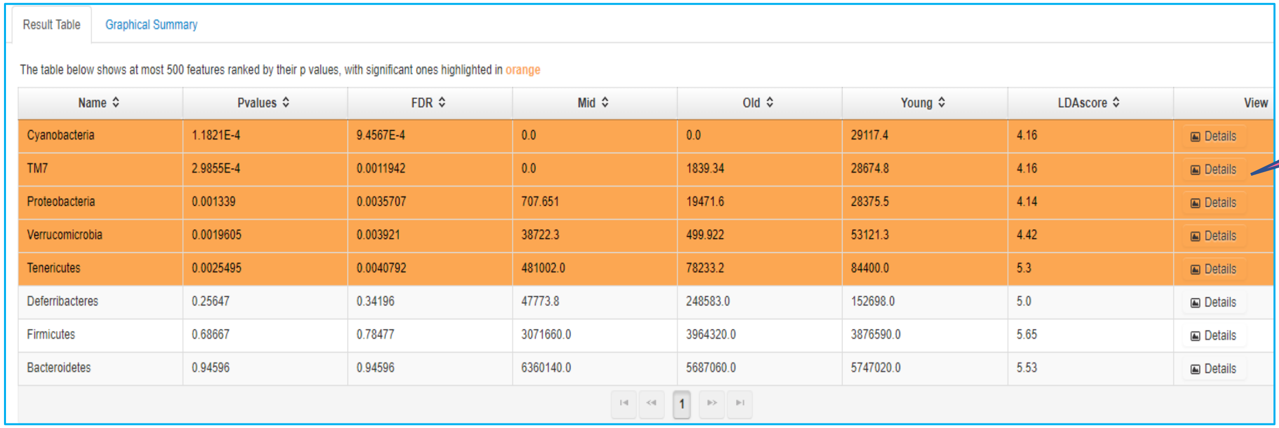
Linear Discriminant Analysis (LDA) Effect Size (LEfSe)

Taxonomic level: Phylum

Experimental factor: Age

Adjusted p-value cutoff: 0.05

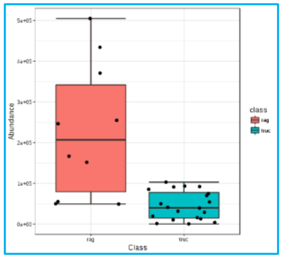
Log LDA score: 1.0



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

Name	Pvalues	FDR	Mid	Old	Young	LDAscore	View
Cyanobacteria	1.1821E-4	9.4567E-4	0.0	0.0	29117.4	4.16	Details
TM7	2.9855E-4	0.0011942	0.0	1839.34	28674.8	4.16	Details
Proteobacteria	0.001339	0.0035707	707.651	19471.6	28375.5	4.14	Details
Verrucomicrobia	0.0019605	0.003921	38722.3	499.922	53121.3	4.42	Details
Tenericutes	0.0025495	0.0040792	481002.0	78233.2	84400.0	5.3	Details
Deferribacteres	0.25647	0.34196	47773.8	248583.0	152698.0	5.0	Details
Firmicutes	0.68667	0.78477	3071660.0	3964320.0	3876590.0	5.65	Details
Bacteroidetes	0.94596	0.94596	6360140.0	5687060.0	5747020.0	5.53	Details

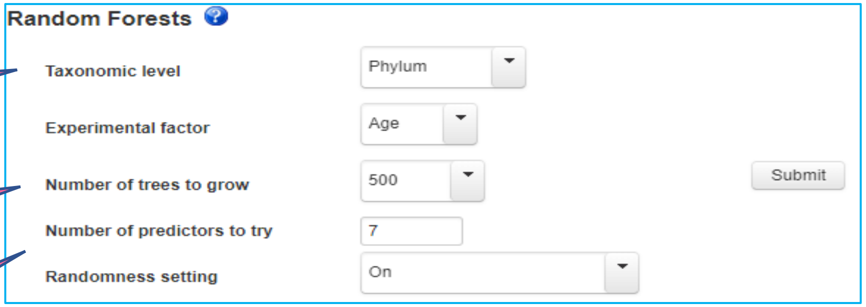
Click on “Details” to see group-wise data distribution for each individual feature



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 sum-rank test:** statistical significance) and biomarker discovery.(**Linear Discriminant analysis:** Effect Size)

E. Biomarker Analysis



Random Forests

Taxonomic level: Phylum

Experimental factor: Age

Number of trees to grow: 500

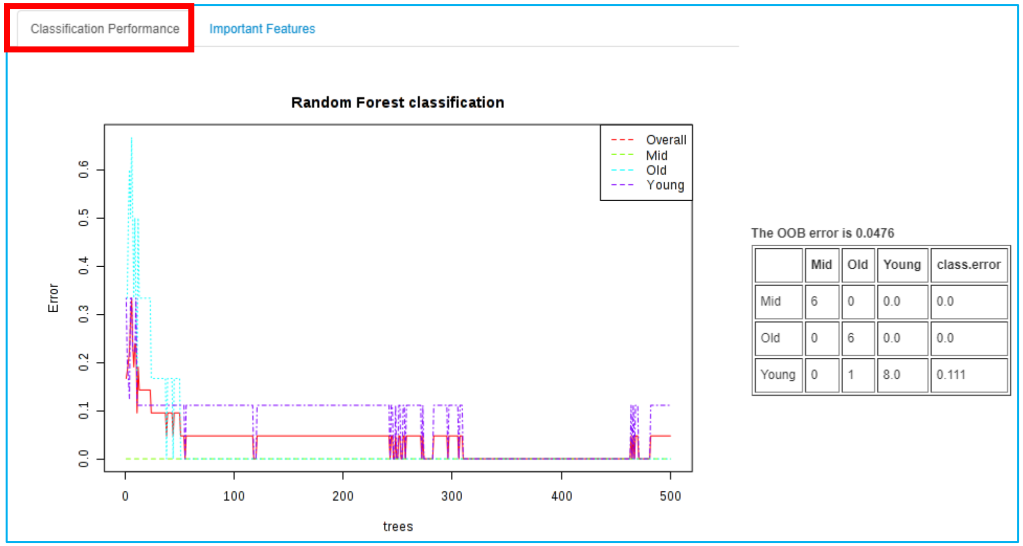
Number of predictors to try: 7

Randomness setting: On

Features can be merged at different taxonomic level

User can choose from no. of trees to be used for classification

No. of predictors for each node



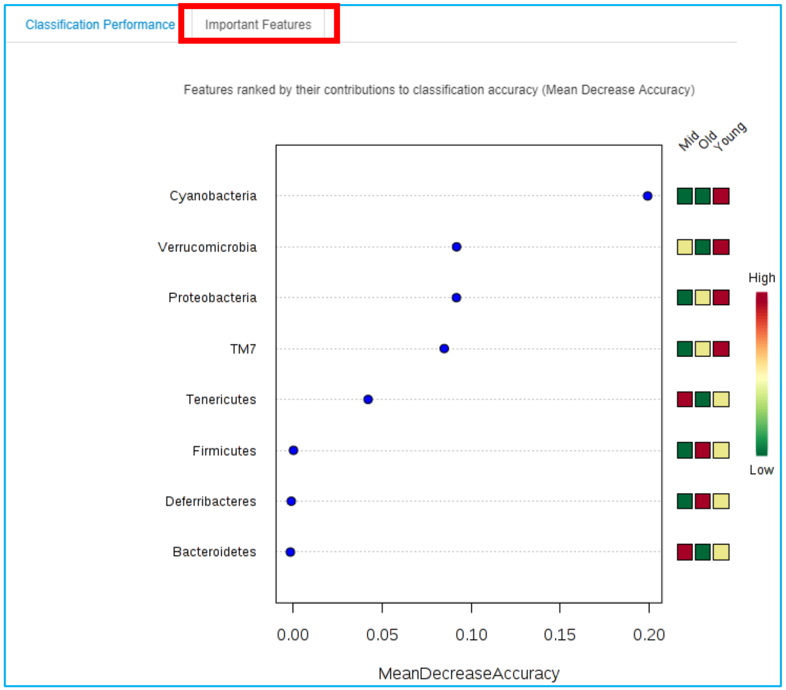
The OOB error is 0.0476

	Mid	Old	Young	class.error
Mid	6	0	0.0	0.0
Old	0	6	0.0	0.0
Young	0	1	8.0	0.111

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.

E. Biomarker Analysis

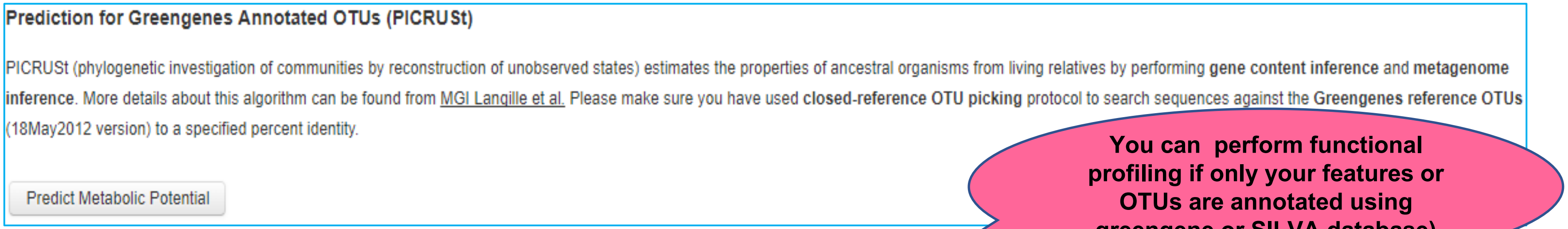


Most important features for classification of data into provided class groups

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.

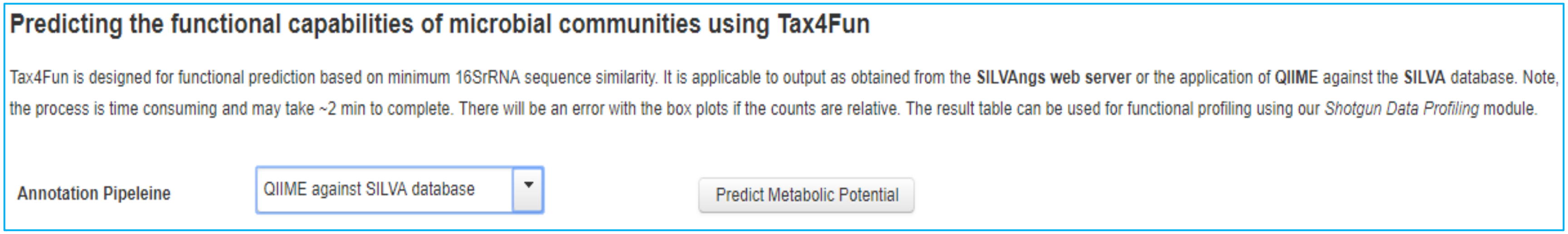
F. Functional potential



Prediction for Greengenes Annotated OTUs (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** (18May2012 version) to a specified percent identity.

You can perform functional profiling if only your features or OTUs are annotated using greengene or SILVA database)



Predicting the functional capabilities of microbial communities using Tax4Fun

Tax4Fun is designed for functional prediction based on minimum 16SrRNA sequence similarity. It is applicable to output as 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 min 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.

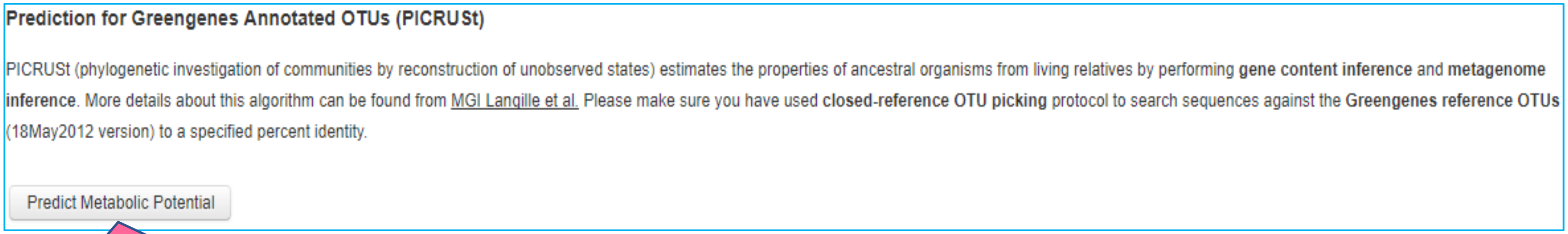
Annotation Pipeleine: QIIME against SILVA database

Functional potential prediction: inferring functional (metabolic) profile from taxonomic profile.

- 2 methods available:

  - **PICRUSt:** It's an **evolutionary modeling algorithm**. Its predictions based on **topology** of the tree and phylogenetic **distance** to next sequenced organism. It is based on **Greengenes** annotated OTUs.
  - **Tax4Fun:** Prediction based on minimum **16SrRNA sequence similarity** using **SILVA** annotated OTUs.

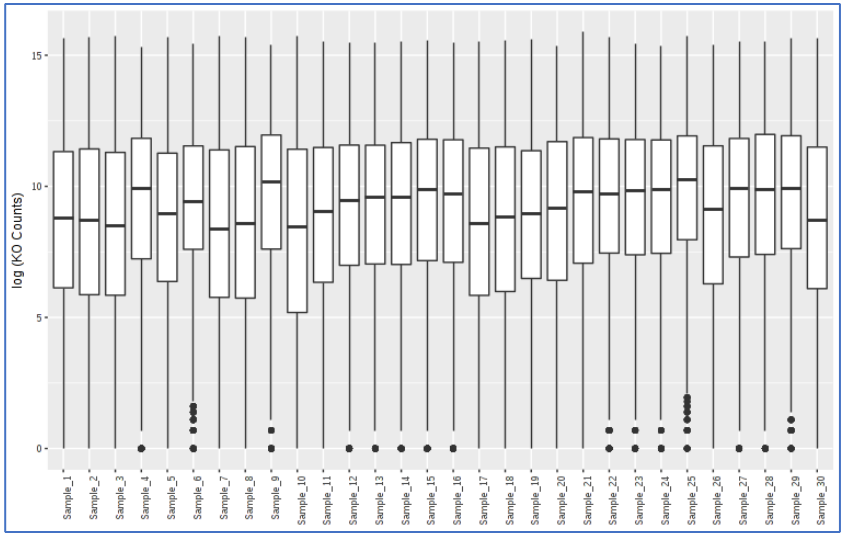
F. Functional potential



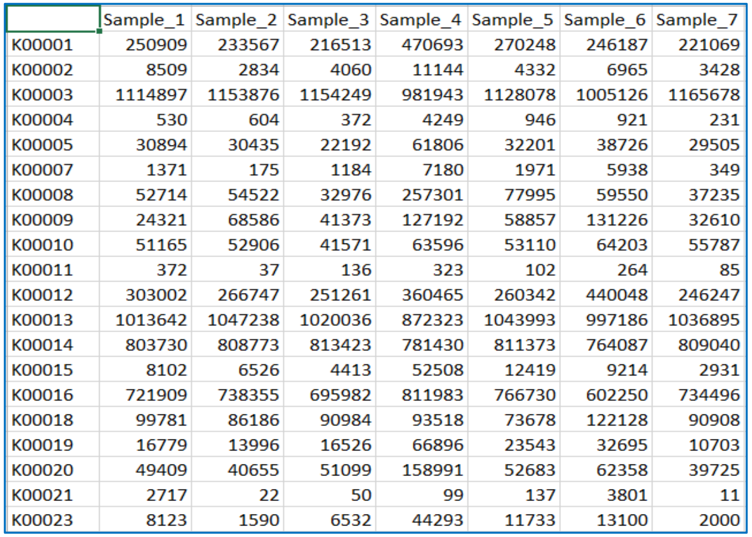
Prediction for Greengenes Annotated OTUs (PICRUSt)

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** (18May2012 version) to a specified percent identity.

Click on “Predict” for profiling



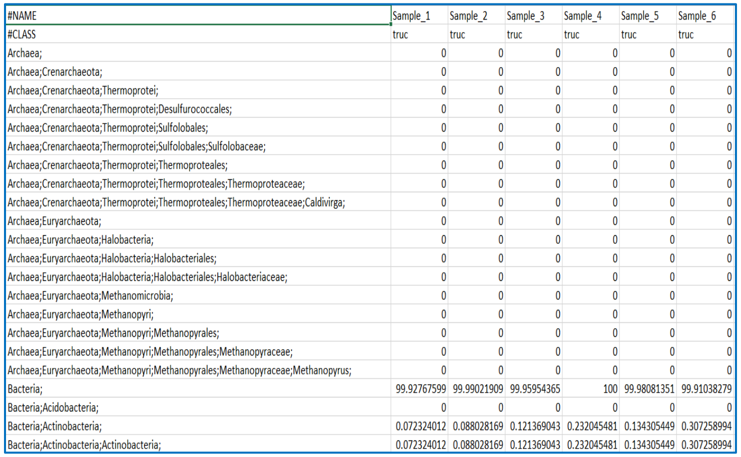
Count distribution od predicted metagenomic abundance data (KO counts) [log-scale]



	Sample_1	Sample_2	Sample_3	Sample_4	Sample_5	Sample_6	Sample_7
K00001	250909	233567	216513	470693	270248	246187	221069
K00002	8509	2834	4060	11144	4332	6965	3428
K00003	1114897	1153876	1154249	981943	1128078	1005126	1165678
K00004	530	604	372	4249	946	921	231
K00005	30894	30435	22192	61806	32201	38726	29505
K00007	1371	175	1184	7180	1971	5938	349
K00008	52714	54522	32976	257301	77995	59550	37235
K00009	24321	68586	41373	127192	58857	131226	32610
K00010	51165	52906	41571	63596	53110	64203	55787
K00011	372	37	136	323	102	264	85
K00012	303002	266747	251261	360465	260342	440048	246247
K00013	1013642	1047238	1020036	872323	1043993	997186	1036895
K00014	803730	808773	813423	781430	811373	764087	809040
K00015	8102	6526	4413	52508	12419	9214	2931
K00016	721909	738355	695982	811983	766730	602250	734496
K00018	99781	86186	90984	93518	73678	122128	90908
K00019	16779	13996	16526	66896	23543	32695	10703
K00020	49409	40655	51099	158991	52683	62358	39725
K00021	2717	22	50	99	137	3801	11
K00023	8123	1590	6532	44293	11733	13100	2000

Result KO table

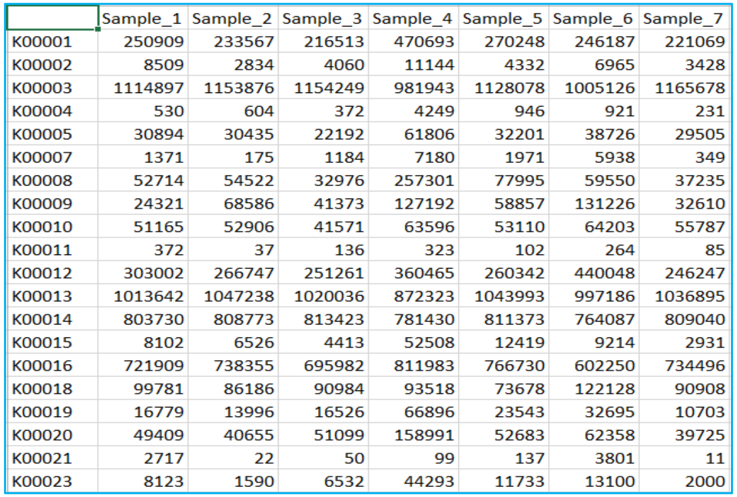
F. Functional potential



#NAME	Sample_1	Sample_2	Sample_3	Sample_4	Sample_5	Sample_6
#CLASS	truc	truc	truc	truc	truc	truc
Archaea;	0	0	0	0	0	0
Archaea;Crenarchaeota;	0	0	0	0	0	0
Archaea;Crenarchaeota;Thermoprotei;	0	0	0	0	0	0
Archaea;Crenarchaeota;Thermoprotei;Desulfurococcales;	0	0	0	0	0	0
Archaea;Crenarchaeota;Thermoprotei;Sulfolobales;	0	0	0	0	0	0
Archaea;Crenarchaeota;Thermoprotei;Sulfolobales;Sulfolobaceae;	0	0	0	0	0	0
Archaea;Crenarchaeota;Thermoprotei;Thermoproteales;	0	0	0	0	0	0
Archaea;Crenarchaeota;Thermoprotei;Thermoproteales;Thermoproteaceae;	0	0	0	0	0	0
Archaea;Crenarchaeota;Thermoprotei;Thermoproteales;Thermoproteaceae;Caldivirga;	0	0	0	0	0	0
Archaea;Euryarchaeota;	0	0	0	0	0	0
Archaea;Euryarchaeota;Halobacteria;	0	0	0	0	0	0
Archaea;Euryarchaeota;Halobacteria;Halobacteriales;	0	0	0	0	0	0
Archaea;Euryarchaeota;Halobacteria;Halobacteriales;Halobacteriaceae;	0	0	0	0	0	0
Archaea;Euryarchaeota;Methanomicrobia;	0	0	0	0	0	0
Archaea;Euryarchaeota;Methanopyri;	0	0	0	0	0	0
Archaea;Euryarchaeota;Methanopyri;Methanopyrales;	0	0	0	0	0	0
Archaea;Euryarchaeota;Methanopyri;Methanopyrales;Methanopyraceae;	0	0	0	0	0	0
Archaea;Euryarchaeota;Methanopyri;Methanopyrales;Methanopyraceae;Methanopyrus;	0	0	0	0	0	0
Bacteria;	99.92767599	99.99021909	99.95954365	100	99.98081351	99.91038279
Bacteria;Acidobacteria;	0	0	0	0	0	0
Bacteria;Actinobacteria;	0.072324012	0.088028169	0.121369043	0.232045481	0.134305449	0.307258994
Bacteria;Actinobacteria;Actinobacteria;	0.072324012	0.088028169	0.121369043	0.232045481	0.134305449	0.307258994

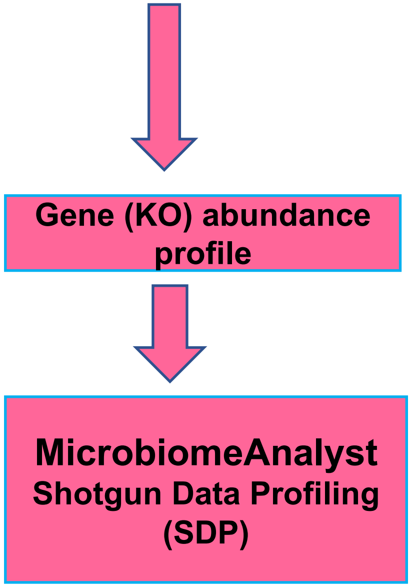
OTU table

Functional profiling



	Sample_1	Sample_2	Sample_3	Sample_4	Sample_5	Sample_6	Sample_7
K00001	250909	233567	216513	470693	270248	246187	221069
K00002	8509	2834	4060	11144	4332	6965	3428
K00003	1114897	1153876	1154249	981943	1128078	1005126	1165678
K00004	530	604	372	4249	946	921	231
K00005	30894	30435	22192	61806	32201	38726	29505
K00007	1371	175	1184	7180	1971	5938	349
K00008	52714	54522	32976	257301	77995	59550	37235
K00009	24321	68586	41373	127192	58857	131226	32610
K00010	51165	52906	41571	63596	53110	64203	55787
K00011	372	37	136	323	102	264	85
K00012	303002	266747	251261	360465	260342	440048	246247
K00013	1013642	1047238	1020036	872323	1043993	997186	1036895
K00014	803730	808773	813423	781430	811373	764087	809040
K00015	8102	6526	4413	52508	12419	9214	2931
K00016	721909	738355	695982	811983	766730	602250	734496
K00018	99781	86186	90984	93518	73678	122128	90908
K00019	16779	13996	16526	66896	23543	32695	10703
K00020	49409	40655	51099	158991	52683	62358	39725
K00021	2717	22	50	99	137	3801	11
K00023	8123	1590	6532	44293	11733	13100	2000

KO table

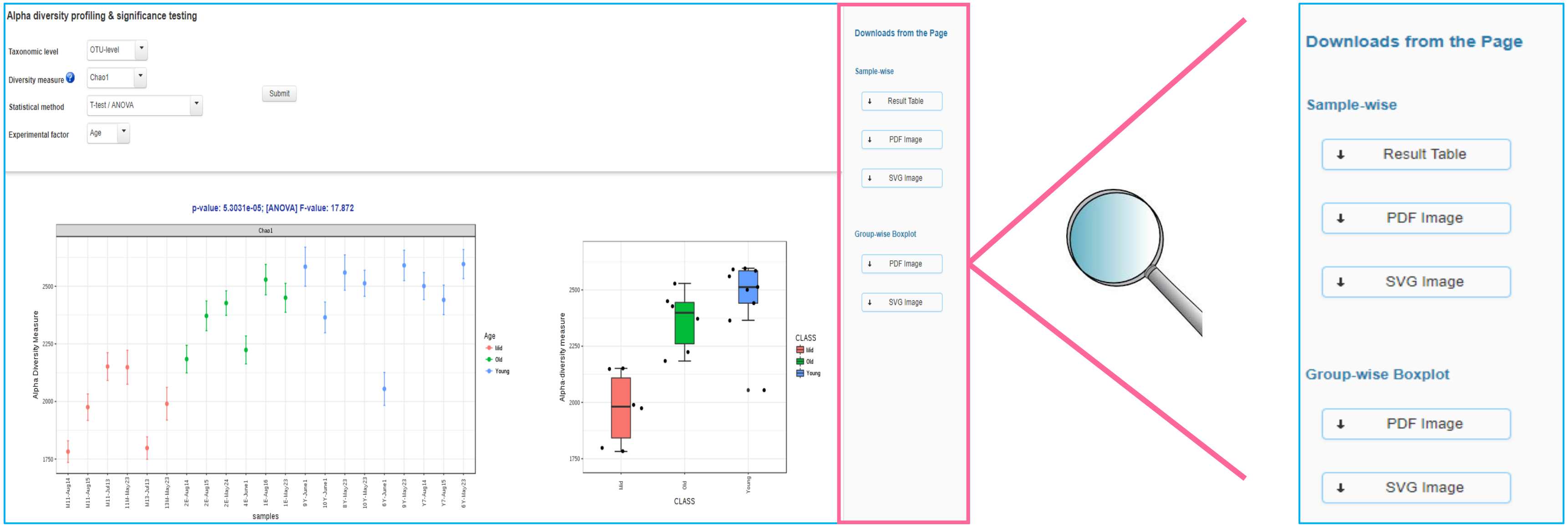


Gene (KO) abundance profile

MicrobiomeAnalyst Shotgun Data Profiling (SDP)

- 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.

Download Results



Alpha diversity profiling & significance testing

Taxonomic level: OTU-level

Diversity measure: Chao1

Statistical method: T-test / ANOVA

Experimental factor: Age

p-value: 5.3031e-05; [ANOVA] F-value: 17.872

- The analysis results (images and tables) can be downloaded from east panel present at every individual analysis page.
- Images can be downloaded in SVG and PDF format.
- Tables are available in CSV format to download.

==END==