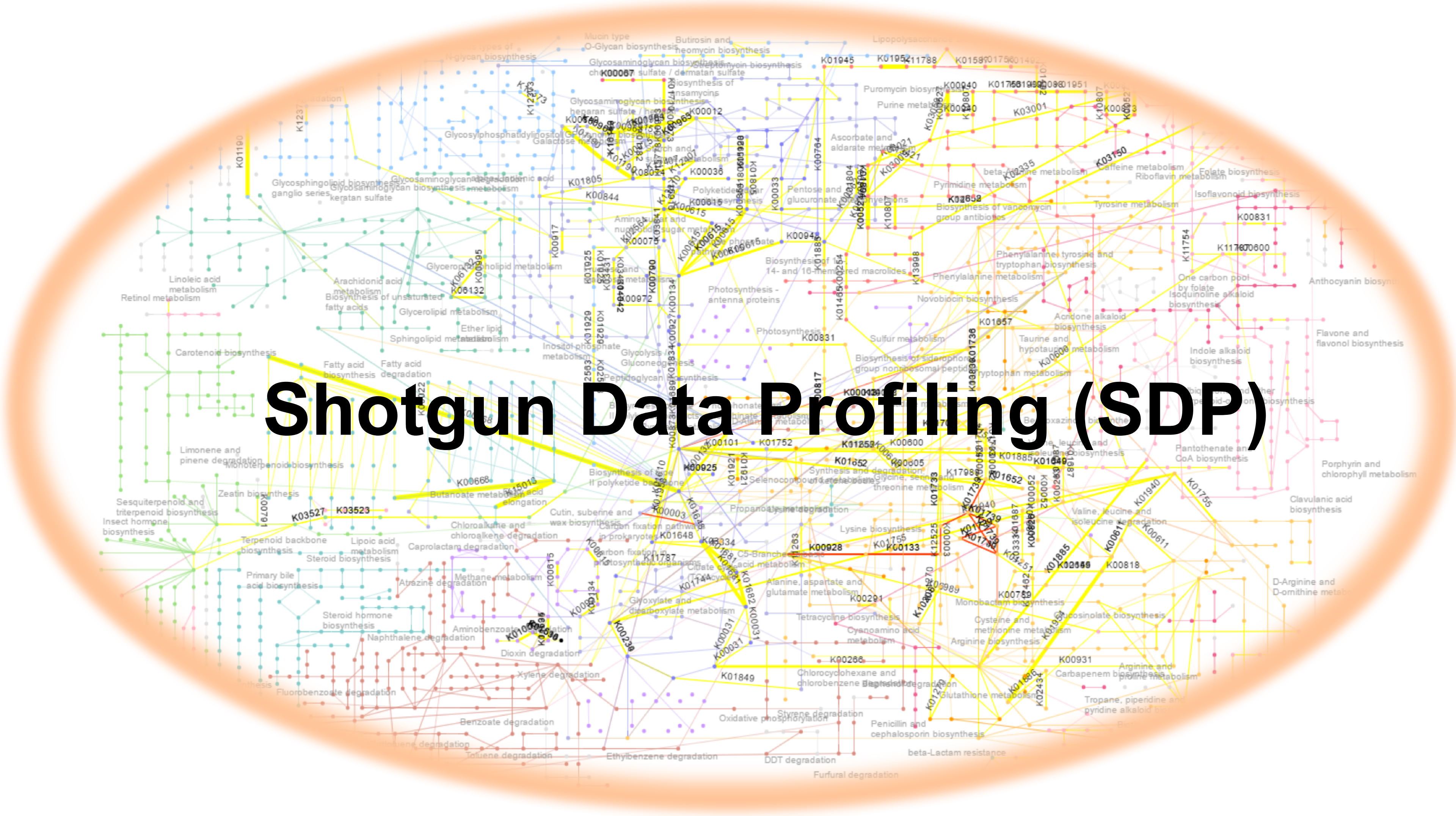
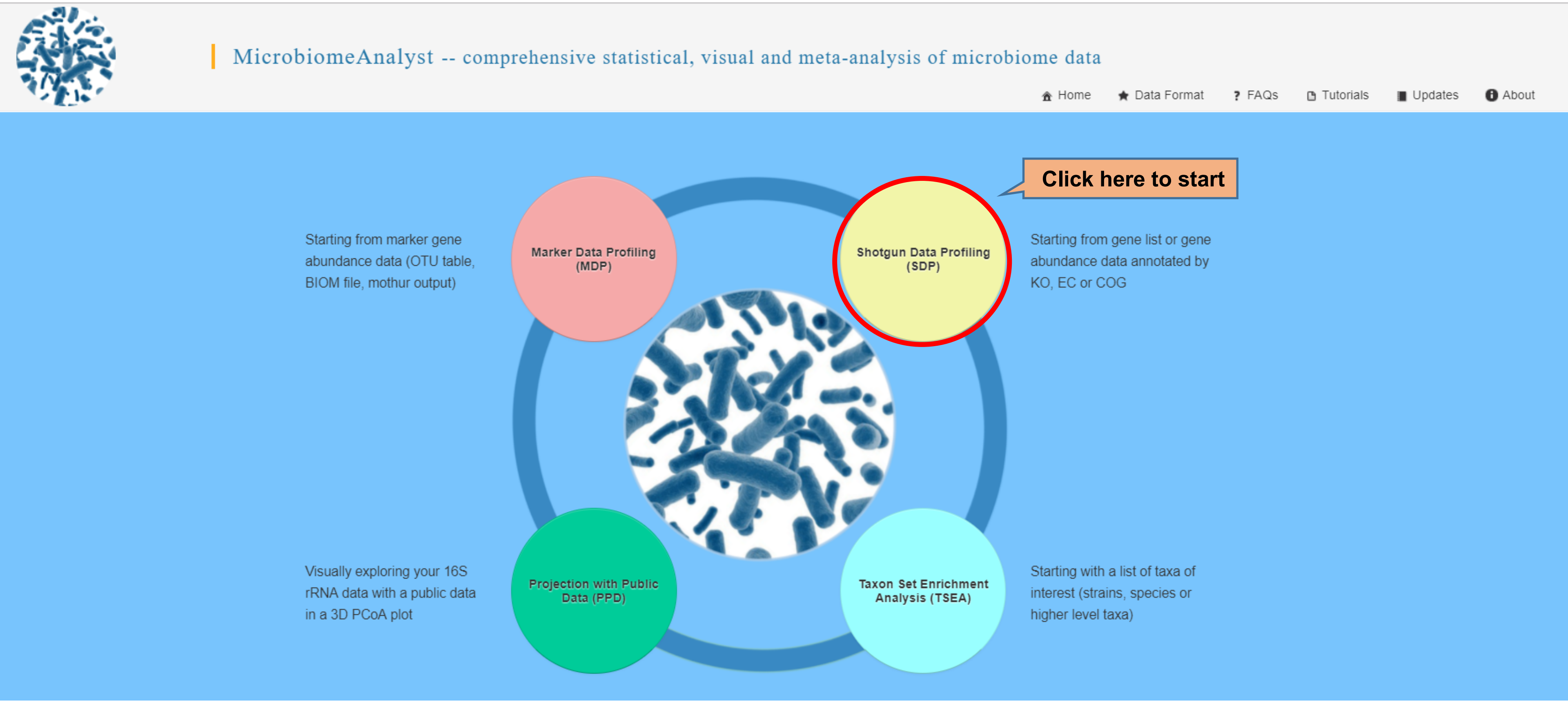
Shotgun Data Profiling (SDP)


Mucin type O-Glycan biosynthesis
Butirosin and neomycin biosynthesis
Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate
Streptomycin biosynthesis
Biosynthesis of ansamycins
Glycosaminoglycan biosynthesis - heparan sulfate / heparin
Glycosylphosphatidylinositol (GPI)-anchor biosynthesis
Galactose metabolism
Starch and sucrose metabolism
Puromycin biosynthesis
Purine metabolism
Ascorbate and aldarate metabolism
beta-Alanine metabolism
Caffeine metabolism
Riboflavin metabolism
Folate biosynthesis
Isoflavonoid biosynthesis
Pyrimidine metabolism
Glycosphingolipid biosynthesis - ganglio series
Glycosaminoglycan biosynthesis - keratan sulfate
Glycosaminoglycan degradation
Polyketide sugar unit biosynthesis
Pentose and glucuronate interconversions
Biosynthesis of vancomycin group antibiotics
Tyrosine metabolism
Amino sugar and nucleotide sugar metabolism
Pentose phosphate pathway
Biosynthesis of 12-, 14- and 16-membered macrolides
Phenylalanine, tyrosine and tryptophan biosynthesis
Phenylalanine metabolism
One carbon pool by folate
Isoquinoline alkaloid biosynthesis
Anthocyanin biosynthesis
Linoleic acid metabolism
Retinol metabolism
Arachidonic acid metabolism
Biosynthesis of unsaturated fatty acids
Glycerophospholipid metabolism
Fructose and mannose metabolism
Photosynthesis - antenna proteins
Novobiocin biosynthesis
Glycerolipid metabolism
Ether lipid metabolism
Sphingolipid metabolism
Inositol phosphate metabolism
Photosynthesis
Sulfur metabolism
Acridone alkaloid biosynthesis
Taurine and hypotaurine metabolism
Flavone and flavonol biosynthesis
Indole alkaloid biosynthesis
Carotenoid biosynthesis
Fatty acid biosynthesis
Fatty acid degradation
Glycolysis / Gluconeogenesis
Peptidoglycan biosynthesis
Biosynthesis of siderophore group nonribosomal peptides
Tryptophan metabolism
Benzoxazinoid biosynthesis
Limonene and pinene degradation
Monoterpenoid biosynthesis
Pantothenate and CoA biosynthesis
Porphyrin and chlorophyll metabolism
Biosynthesis of type II polyketide backbone
Selenocompound metabolism
Glycine, serine and threonine metabolism
Butanoate metabolism
Fatty acid elongation
Zeatin biosynthesis
Sesquiterpenoid and triterpenoid biosynthesis
Insect hormone biosynthesis
Propanoate metabolism
Lysine degradation
Cutin, suberine and wax biosynthesis
Chloroalkane and chloroalkene degradation
Carbon fixation pathways in prokaryotes
Valine, leucine and isoleucine degradation
Clavulanic acid biosynthesis
Terpenoid backbone biosynthesis
Lipoic acid metabolism
Caprolactam degradation
Steroid biosynthesis
Carbon fixation in photosynthetic organisms
C5-Branched dibasic acid metabolism
Lysine biosynthesis
Primary bile acid biosynthesis
Atrazine degradation
Methane metabolism
Citrate cycle (TCA cycle)
Alanine, aspartate and glutamate metabolism
Monobactam biosynthesis
Glucosinolate biosynthesis
D-Arginine and D-ornithine metabolism
Steroid hormone biosynthesis
Glyoxylate and dicarboxylate metabolism
Tetracycline biosynthesis
Cysteine and methionine metabolism
Aminobenzoate degradation
Cyanoamino acid metabolism
Arginine biosynthesis
Naphthalene degradation
Dioxin degradation
Xylene degradation
Chlorocyclohexane and chlorobenzene degradation
Carbapenem biosynthesis
Arginine and proline metabolism
Fluorobenzoate degradation
Glutathione metabolism
Tropane, piperidine and pyridine alkaloid biosynthesis
Benzoate degradation
Oxidative phosphorylation
Styrene degradation
Penicillin and cephalosporin biosynthesis
Toluene degradation
Ethylbenzene degradation
DDT degradation
beta-Lactam resistance
Furfural degradation
K01945
K11788
K00067
K00012
K01805
K00844
K00036
K00615
K00075
K00942
K06132
K00972
K00831
K01657
K00600
K00101
K01752
K01652
K00605
K00925
K00668
K03527
K03523
K00003
K01648
K01787
K00928
K00133
K00291
K00789
K00818
K00266
K00931
K01849
K03001
K03150
K02335
K11754
K01190

Goal for this tutorial

- To perform an exploratory and biomarker analysis on shotgun metagenomics data and visualize the results within KEGG metabolic networks along with pathway analysis.



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

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)

Click here to start

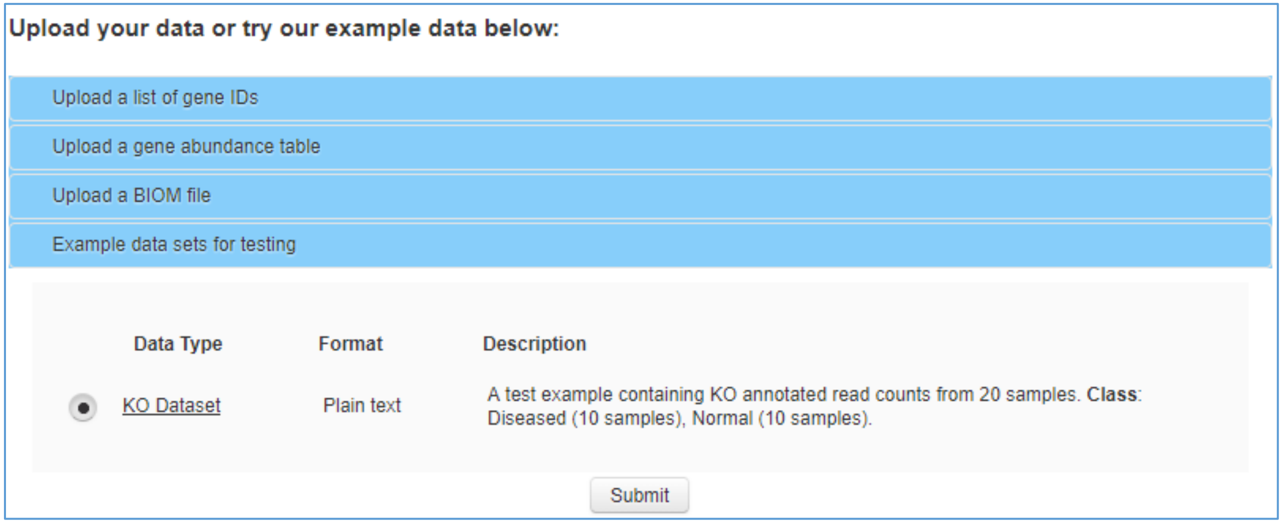
Starting from marker gene abundance data (OTU table, BIOM file, mothur output)

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)

Shotgun Data Profiling (SDP)



Upload your data or try our example data below:

Upload a list of gene IDs

Upload a gene abundance table

Upload a BIOM file

Example data sets for testing

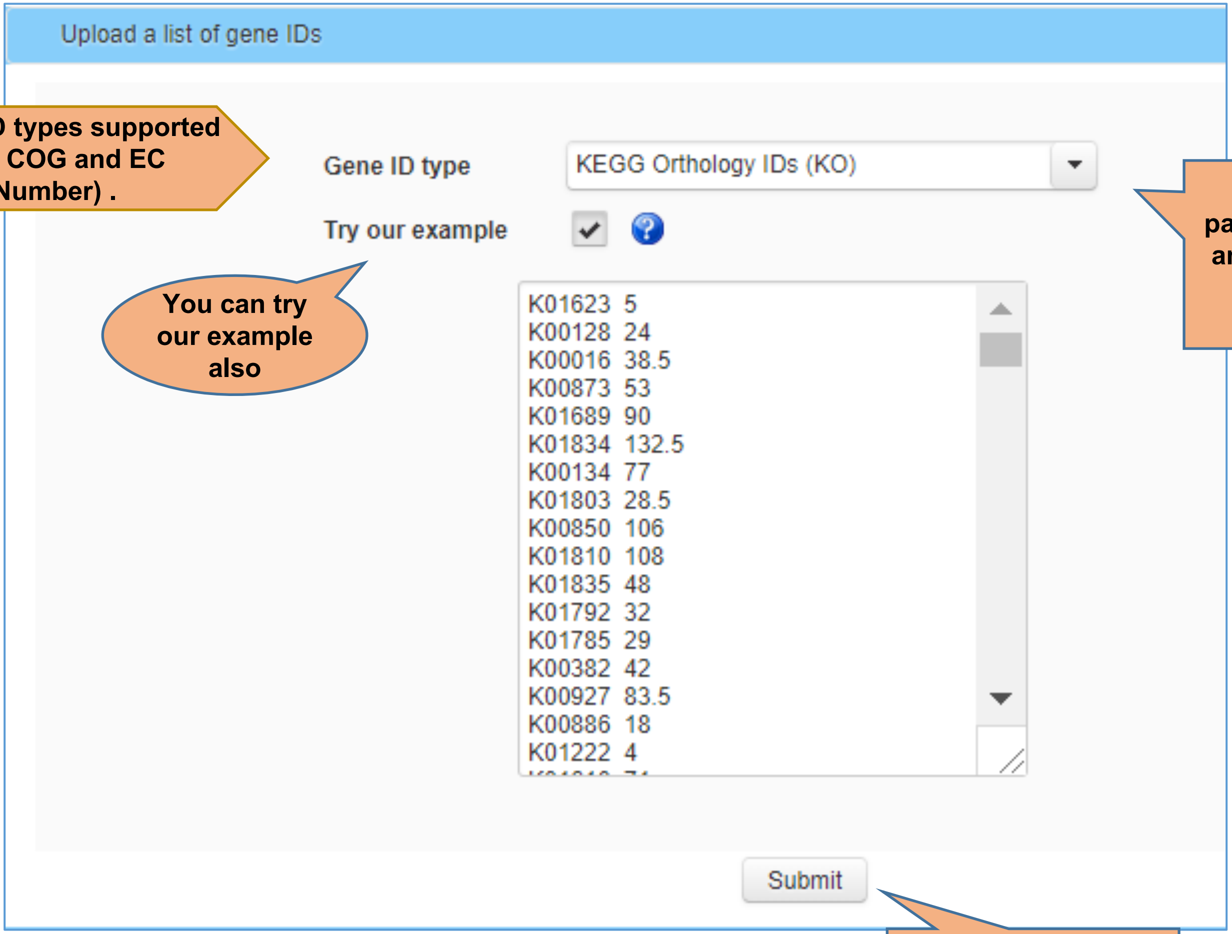
	Data Type	Format	Description
●	KO Dataset	Plain text	A test example containing KO annotated read counts from 20 samples. **Class**: Diseased (10 samples), Normal (10 samples).

Two types of user inputs:

- ❖ A list of gene IDs.
- ❖ Abundance table (in text or BIOM format)

Note genes need to be annotated in KO, EC, or COG for functional analysis,

A) 1. Upload a list



Upload a list of gene IDs

Gene ID type: KEGG Orthology IDs (KO)

Try our example ☑

```
K01623  5
K00128  24
K00016  38.5
K00873  53
K01689  90
K01834  132.5
K00134  77
K01803  28.5
K00850  106
K01810  108
K01835  48
K01792  32
K01785  29
K00382  42
K00927  83.5
K00886  18
K01222  4
```

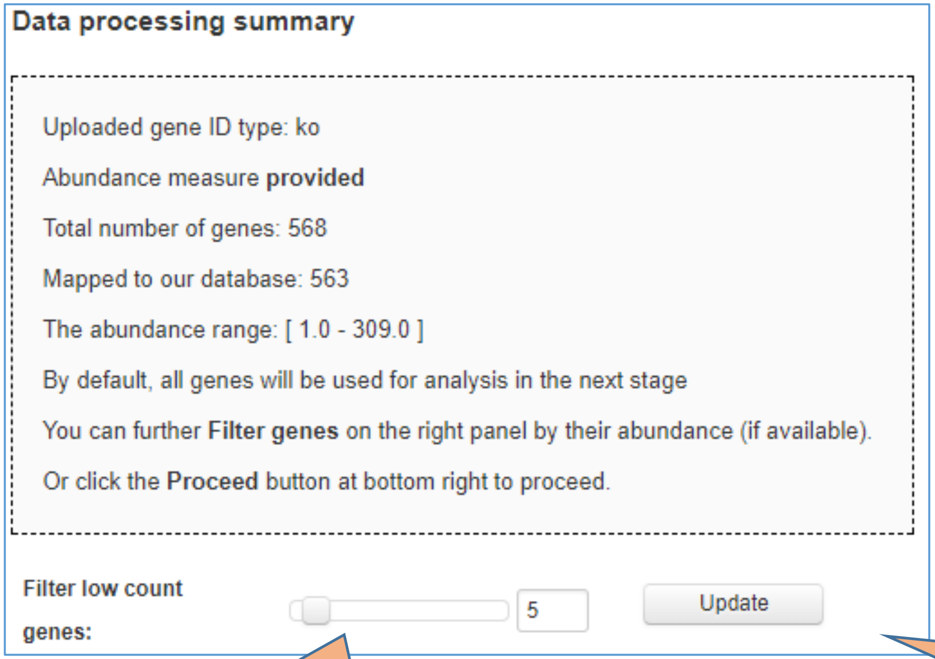
3 gene ID types supported (KO, COG and EC Number) .

You can try our example also

Step 1 : Choose the parameters above. Copy and paste a list of gene IDs along with their expression value

Step 2 : Click "Submit" to proceed.

2. Data Integrity Check



Data processing summary

Uploaded gene ID type: ko

Abundance measure **provided**

Total number of genes: 568

Mapped to our database: 563

The abundance range: [1.0 - 309.0]

By default, all genes will be used for analysis in the next stage

You can further **Filter genes** on the right panel by their abundance (if available).

Or click the **Proceed** button at bottom right to proceed.

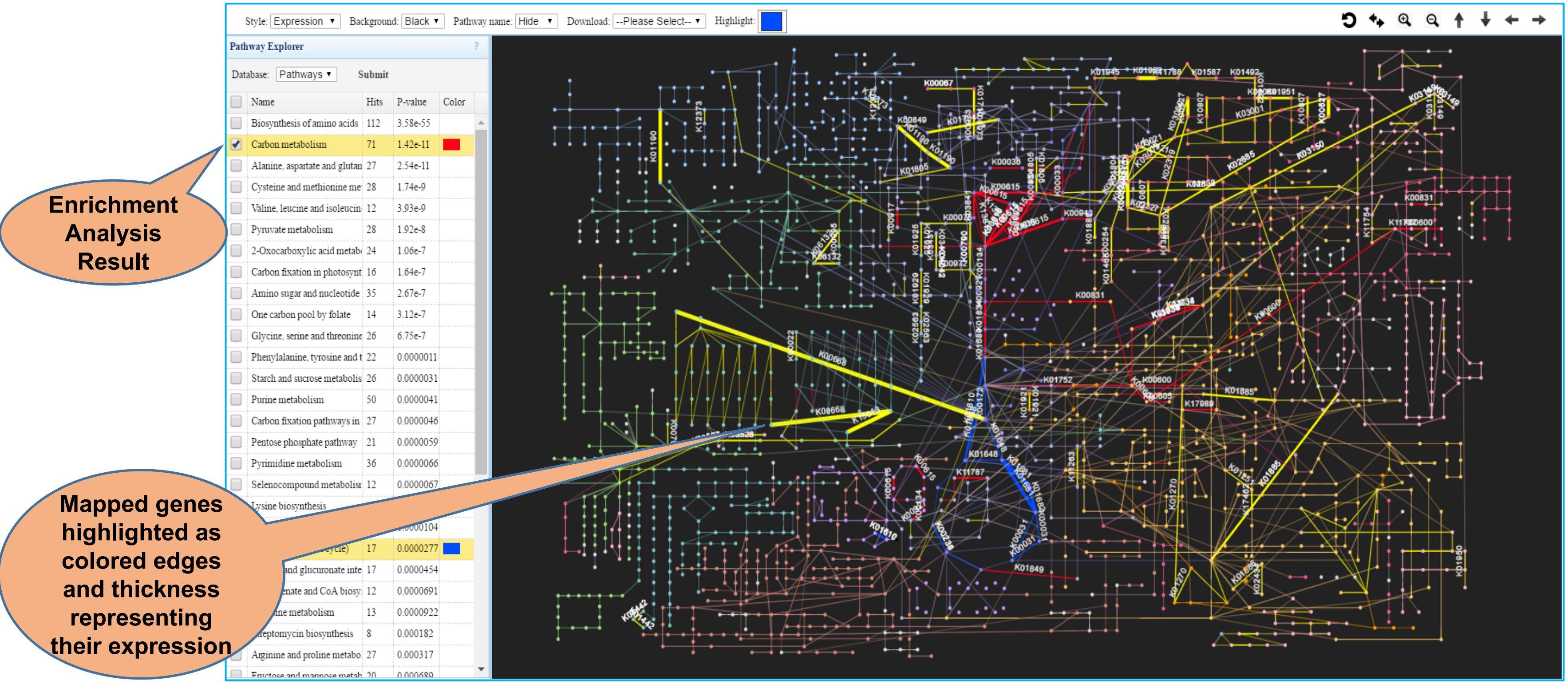
Filter low count genes: 5

genes with low count can be filtered out

Click "Proceed" to visualize the result within KEGG metabolic network

Provides processing and summary information for user uploaded gene list.

3. KEGG Metabolic Networks (I)



Style: Expression ▾ Background: Black ▾ Pathway name: Hide ▾ Download: --Please Select-- ▾ Highlight: 🟦

Pathway Explorer

Database: Pathways ▾ Submit

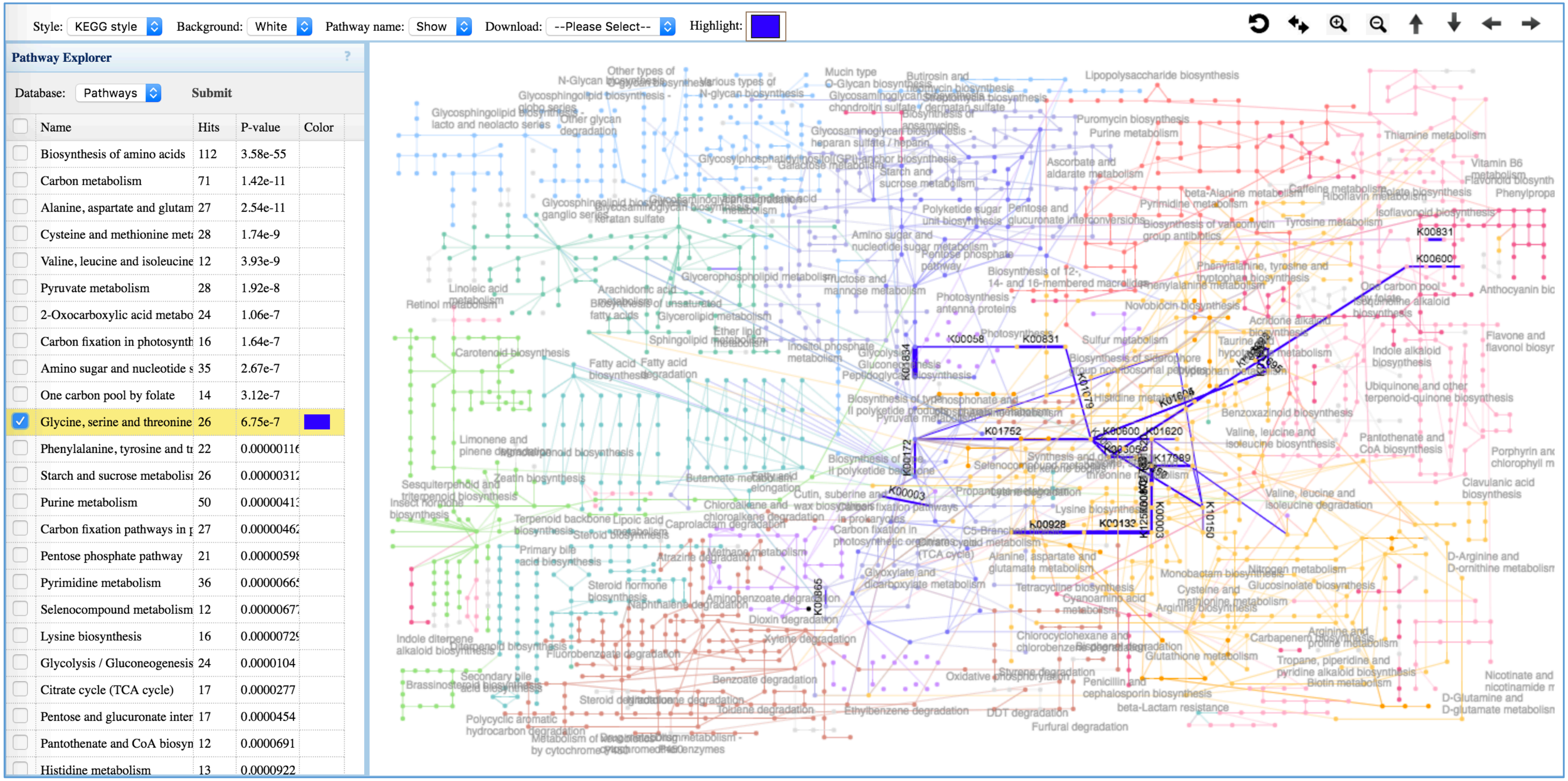
☐	Name	Hits	P-value	Color
☐	Biosynthesis of amino acids	112	3.58e-55	
☑	Carbon metabolism	71	1.42e-11	🟥
☐	Alanine, aspartate and glutan	27	2.54e-11	
☐	Cysteine and methionine me	28	1.74e-9	
☐	Valine, leucine and isoleucin	12	3.93e-9	
☐	Pyruvate metabolism	28	1.92e-8	
☐	2-Oxocarboxylic acid metab	24	1.06e-7	
☐	Carbon fixation in photosynt	16	1.64e-7	
☐	Amino sugar and nucleotide	35	2.67e-7	
☐	One carbon pool by folate	14	3.12e-7	
☐	Glycine, serine and threonine	26	6.75e-7	
☐	Phenylalanine, tyrosine and t	22	0.0000011	
☐	Starch and sucrose metabolis	26	0.0000031	
☐	Purine metabolism	50	0.0000041	
☐	Carbon fixation pathways in	27	0.0000046	
☐	Pentose phosphate pathway	21	0.0000059	
☐	Pyrimidine metabolism	36	0.0000066	
☐	Selenocompound metabolisr	12	0.0000067	
	Lysine biosynthesis			
			.0000104	
	cycle)	17	0.0000277	🟦
	and glucuronate inte	17	0.0000454	
	enate and CoA biosy	12	0.0000691	
	ine metabolism	13	0.0000922	
	reptomycin biosynthesis	8	0.000182	
	Arginine and proline metabo	27	0.000317	
☐	Fructose and mannose metab	20	0.000689	

Enrichment Analysis Result

Mapped genes highlighted as colored edges and thickness representing their expression

1. Click “Submit” on the Pathway Explorer to perform pathways enrichment analysis.
2. Select a highlight color (default orange)
3. Click on a pathway name (a table row) to highlight the corresponding pathways
    - Gene IDs (KO) are represented as edge (reaction linking two metabolites) in the network and its thickness are based on their expression levels.

3. KEGG Metabolic Networks (II)



Style: KEGG style Background: White Pathway name: Show Download: --Please Select-- Highlight: 🟦

Pathway Explorer

Database: Pathways **Submit**

☐	Name	Hits	P-value	Color
☐	Biosynthesis of amino acids	112	3.58e-55	
☐	Carbon metabolism	71	1.42e-11	
☐	Alanine, aspartate and glutam	27	2.54e-11	
☐	Cysteine and methionine met	28	1.74e-9	
☐	Valine, leucine and isoleucine	12	3.93e-9	
☐	Pyruvate metabolism	28	1.92e-8	
☐	2-Oxocarboxylic acid metabo	24	1.06e-7	
☐	Carbon fixation in photosynth	16	1.64e-7	
☐	Amino sugar and nucleotide s	35	2.67e-7	
☐	One carbon pool by folate	14	3.12e-7	
☑	Glycine, serine and threonine	26	6.75e-7	🟦
☐	Phenylalanine, tyrosine and tr	22	0.0000011	
☐	Starch and sucrose metabolis	26	0.0000031	
☐	Purine metabolism	50	0.0000041	
☐	Carbon fixation pathways in p	27	0.0000046	
☐	Pentose phosphate pathway	21	0.0000059	
☐	Pyrimidine metabolism	36	0.0000066	
☐	Selenocompound metabolism	12	0.0000067	
☐	Lysine biosynthesis	16	0.0000072	
☐	Glycolysis / Gluconeogenesis	24	0.0000104	
☐	Citrate cycle (TCA cycle)	17	0.0000277	
☐	Pentose and glucuronate inter	17	0.0000454	
☐	Pantothenate and CoA biosyn	12	0.0000691	
☐	Histidine metabolism	13	0.0000922	

Lipopolysaccharide biosynthesis
Puromycin biosynthesis
Purine metabolism
Thiamine metabolism
Vitamin B6 metabolism
Ascorbate and aldarate metabolism
Starch and sucrose metabolism
Pyrimidine metabolism
Caffeine metabolism
Riboflavin metabolism
Tyrosine metabolism
Isoflavonoid biosynthesis
K00831
K00600
Anthocyanin bio
Phenylalanine metabolism
Novobiocin biosynthesis
One carbon pool by folate
Flavone and flavonol biosyn
Indole alkaloid biosynthesis
Ubiquinone and other terpenoid-quinone biosynthesis
Benzoxazinoid biosynthesis
Pantothenate and CoA biosynthesis
Valine, leucine and isoleucine biosynthesis
Porphyrin and chlorophyll m
Clavulanic acid biosynthesis
Valine, leucine and isoleucine degradation
D-Arginine and D-ornithine metabolism
Nitrogen metabolism
Glucosinolate biosynthesis
Cysteine and methionine metabolism
Arginine biosynthesis
Carbapenem biosynthesis
Glutathione metabolism
Tropane, piperidine and pyridine alkaloid biosynthesis
Biotin metabolism
Nicotinate and nicotinamide m
D-Glutamine and D-glutamate metabolism
Penicillin and cephalosporin biosynthesis
beta-Lactam resistance
Furfural degradation
DDT degradation
Ethylbenzene degradation
Toluene degradation
Benzoate degradation
Dioxin degradation
Xylene degradation
Oxidative phosphorylation
Glyoxylate and dicarboxylate metabolism
Alanine, aspartate and glutamate metabolism
Citrate cycle (TCA cycle)
Carbon fixation pathways in prokaryotes
Carbon fixation in photosynthetic organisms
Lysine biosynthesis
Sulfur metabolism
Histidine metabolism
Photosynthesis
Photosynthesis - antenna proteins
Inositol phosphate metabolism
Fatty acid degradation
Fatty acid biosynthesis
Butanoate metabolism
Carotenoid biosynthesis
Limonene and pinene degradation
Zeatin biosynthesis
Sesquiterpenoid and triterpenoid biosynthesis
Insect hormone biosynthesis
Terpenoid backbone biosynthesis
Steroid biosynthesis
Primary bile acid biosynthesis
Steroid hormone biosynthesis
Indole diterpene alkaloid biosynthesis
Secondary bile acid biosynthesis
Brassinosteroid biosynthesis
Polycyclic aromatic hydrocarbon degradation
Metabolism of xenobiotics by cytochrome P450
Retinol metabolism
Linoleic acid metabolism
Arachidonic acid metabolism
Glycerophospholipid metabolism
Glycerolipid metabolism
Sphingolipid metabolism
Ether lipid metabolism
Fructose and mannose metabolism
Amino sugar and nucleotide sugar metabolism
Pentose phosphate pathway
Glycosaminoglycan biosynthesis - heparan sulfate / heparin
Glycosphingolipid biosynthesis - globo series
Glycosphingolipid biosynthesis - lacto and neolacto series
Glycosphingolipid biosynthesis - ganglio series
N-Glycan biosynthesis
Mucin type O-Glycan biosynthesis
Various types of N-glycan biosynthesis
Other glycan degradation
Galactose metabolism
K01834
K00058
K01079
K01752
K00600
K01620
K17989
K00003
K00928
K00133
K10150
K00865
K00172

Customizing the styles using the menus on the top too bar, for example:

- Switching background from black to white;
- Showing the pathway names.

B) Analyzing shotgun gene count data

Data Formatting

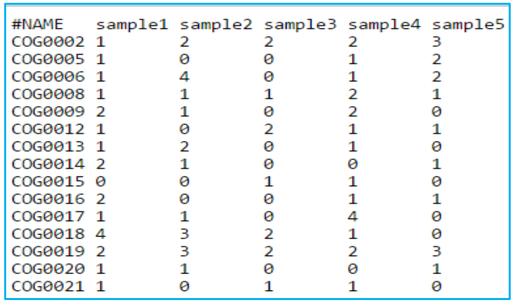
1. Tab-delimited text file

- User have to upload both gene abundance table and metadata file separately.
- 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.

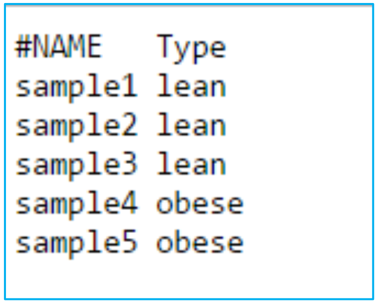
2. BIOM format

- Standard format for storing gene abundance information (metadata file separately in .txt file).

For Example:



#NAME	sample1	sample2	sample3	sample4	sample5
COG0002	1	2	2	2	3
COG0005	1	0	0	1	2
COG0006	1	4	0	1	2
COG0008	1	1	1	2	1
COG0009	2	1	0	2	0
COG0012	1	0	2	1	1
COG0013	1	2	0	1	0
COG0014	2	1	0	0	1
COG0015	0	0	1	1	0
COG0016	2	0	0	1	1
COG0017	1	1	0	4	0
COG0018	4	3	2	1	0
COG0019	2	3	2	2	3
COG0020	1	1	0	0	1
COG0021	1	0	1	1	0



#NAME	Type
sample1	lean
sample2	lean
sample3	lean
sample4	obese
sample5	obese

Abundance table and Metadata file in tab-delimited (.txt) format

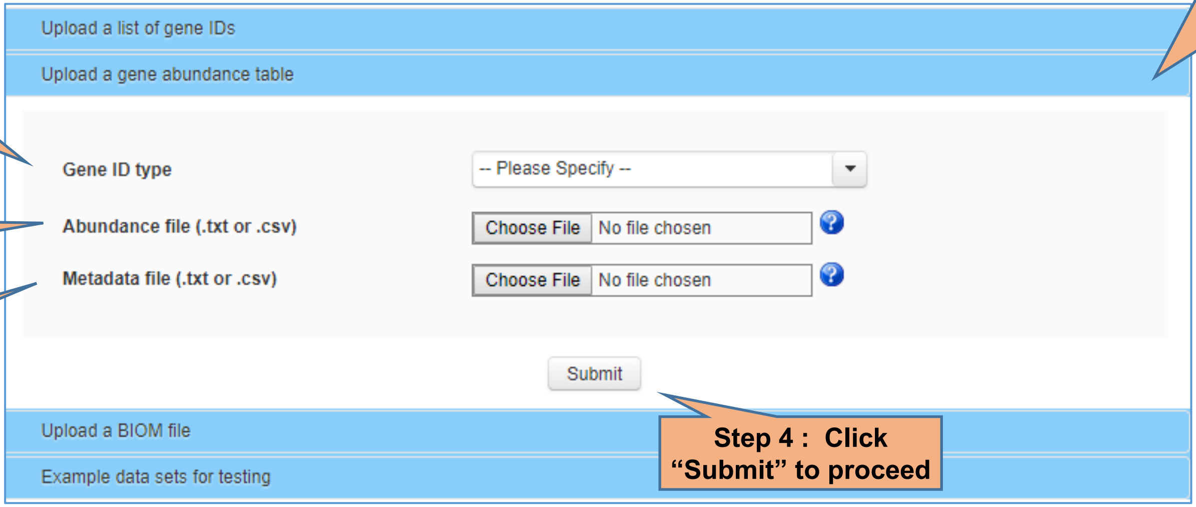
1. Data Upload

Step 1: Upload your gene abundance profile data in table or BIOM format

Step 2: Chose a gene ID type
3 IDs supported (KO, COG and EC numbers)

Step 3: Upload your abundance data file

Step 4: Upload your metadata file



Upload a list of gene IDs

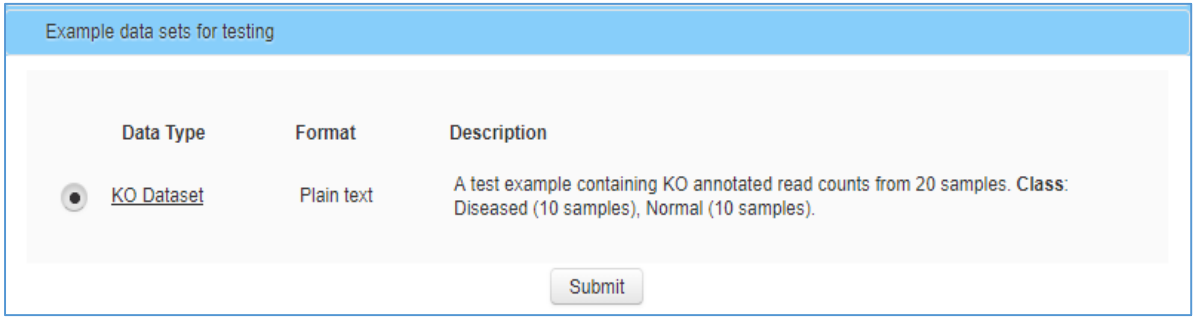
Upload a gene abundance table

Gene ID type	-- Please Specify --
Abundance file (.txt or .csv)	Choose File No file chosen
Metadata file (.txt or .csv)	Choose File No file chosen

Upload a BIOM file

Example data sets for testing

Step 4 : Click "Submit" to proceed

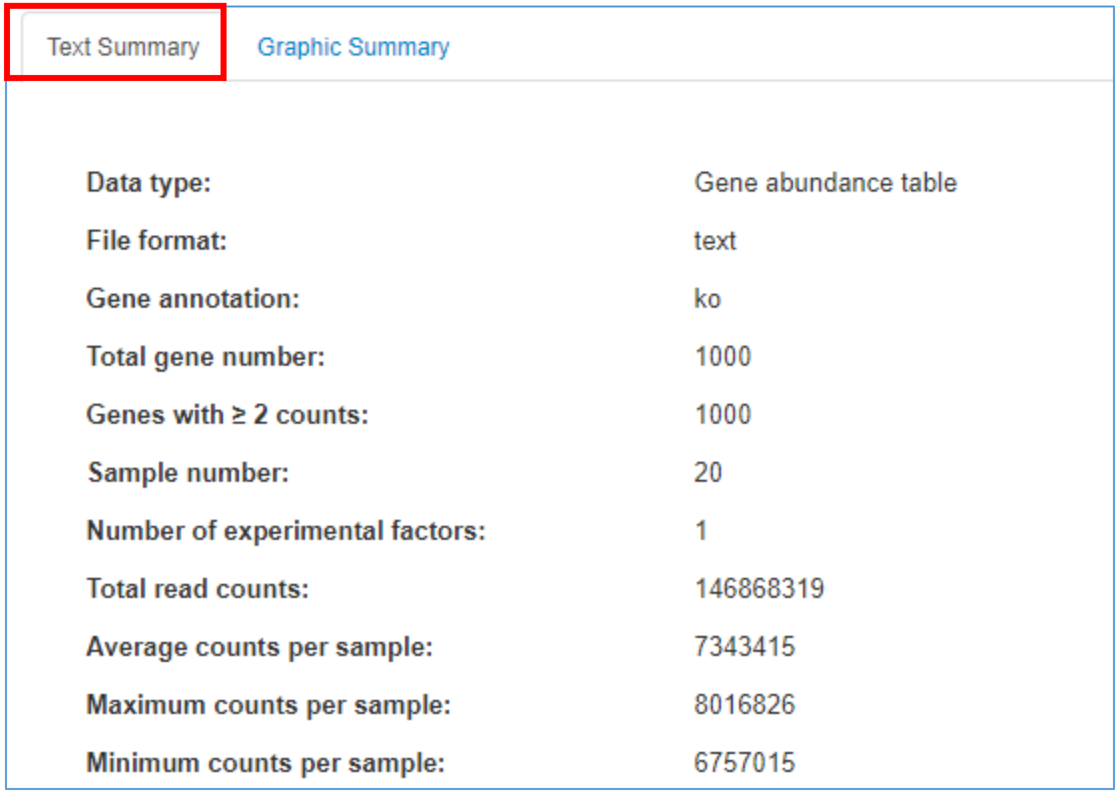


Example data sets for testing

	Data Type	Format	Description
●	KO Dataset	Plain text	A test example containing KO annotated read counts from 20 samples. **Class**: Diseased (10 samples), Normal (10 samples).

You can try our example also

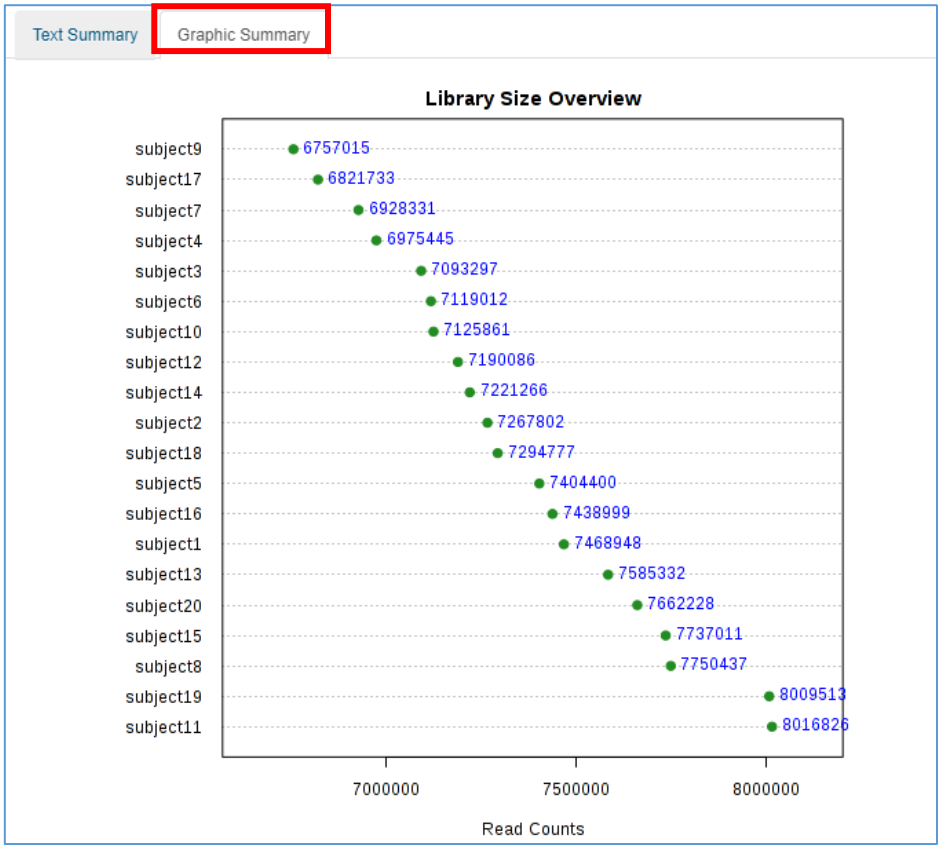
2. a) Data Integrity Check



Data type:	Gene abundance table
File format:	text
Gene annotation:	ko
Total gene number:	1000
Genes with ≥ 2 counts:	1000
Sample number:	20
Number of experimental factors:	1
Total read counts:	146868319
Average counts per sample:	7343415
Maximum counts per sample:	8016826
Minimum counts per sample:	6757015

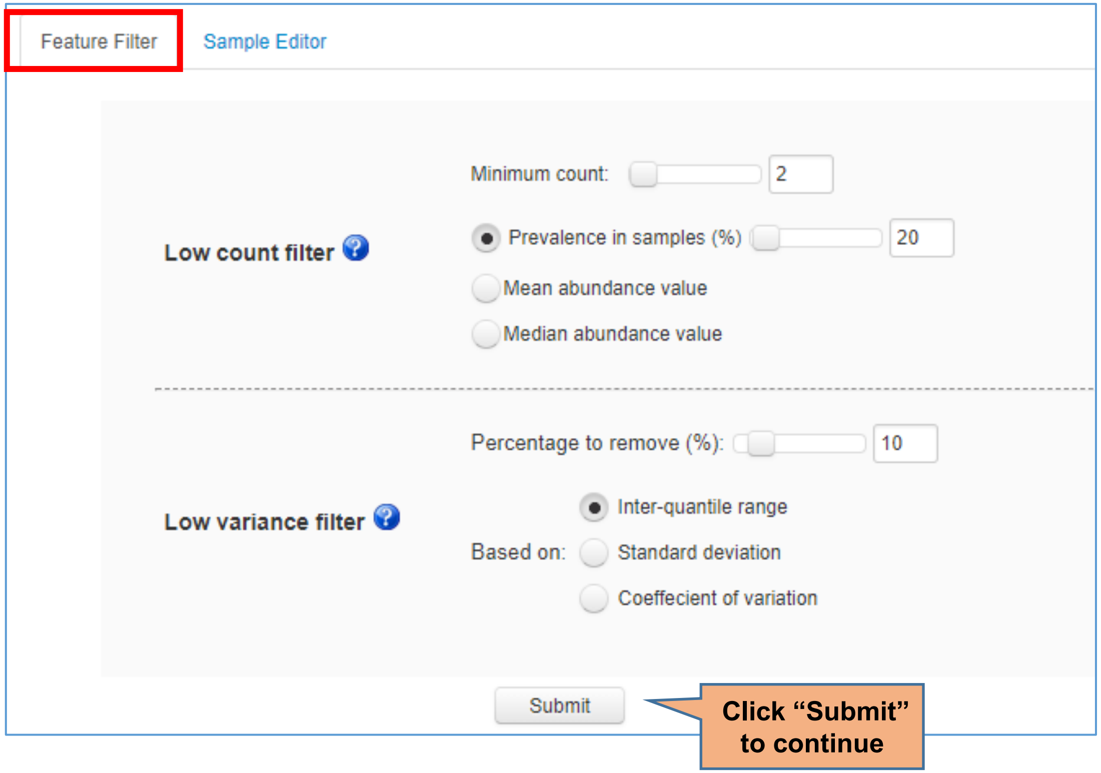
Provides processing and summary information for user uploaded data.

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)



Low count filter

Minimum count: 2

● Prevalence in samples (%) 20

○ Mean abundance value

○ Median abundance value

Low variance filter

Percentage to remove (%): 10

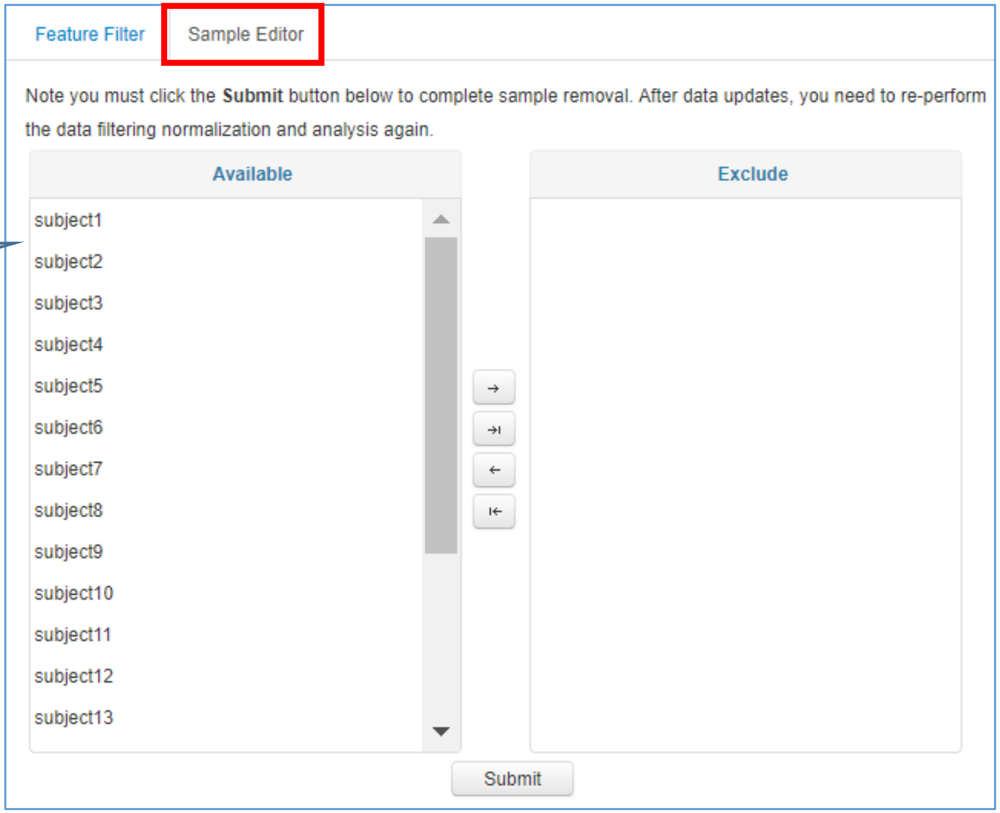
Based on:
● Inter-quantile range
○ Standard deviation
○ Coeffecient of variation

Click "Submit" to continue

- Identifying and removing variables or features that are unlikely to be of use when modeling the data. (e.g., features containing all zeros or constant across all the samples)

- **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.

Available

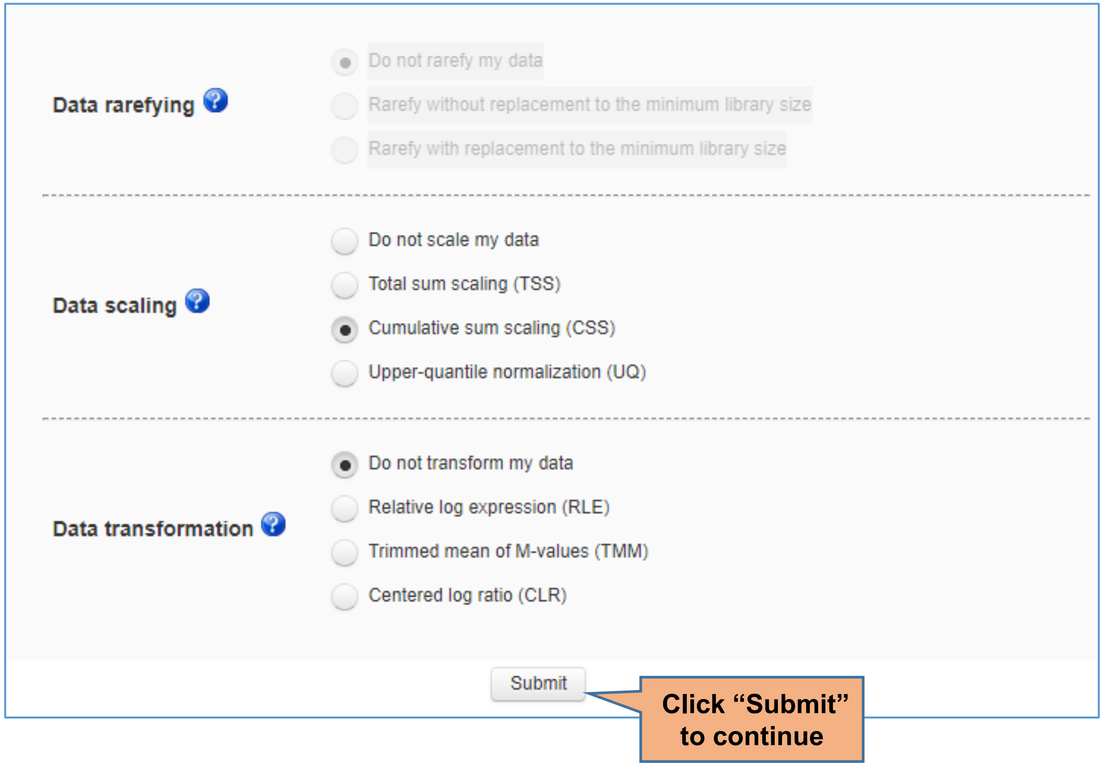
subject1
subject2
subject3
subject4
subject5
subject6
subject7
subject8
subject9
subject10
subject11
subject12
subject13

Exclude

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)

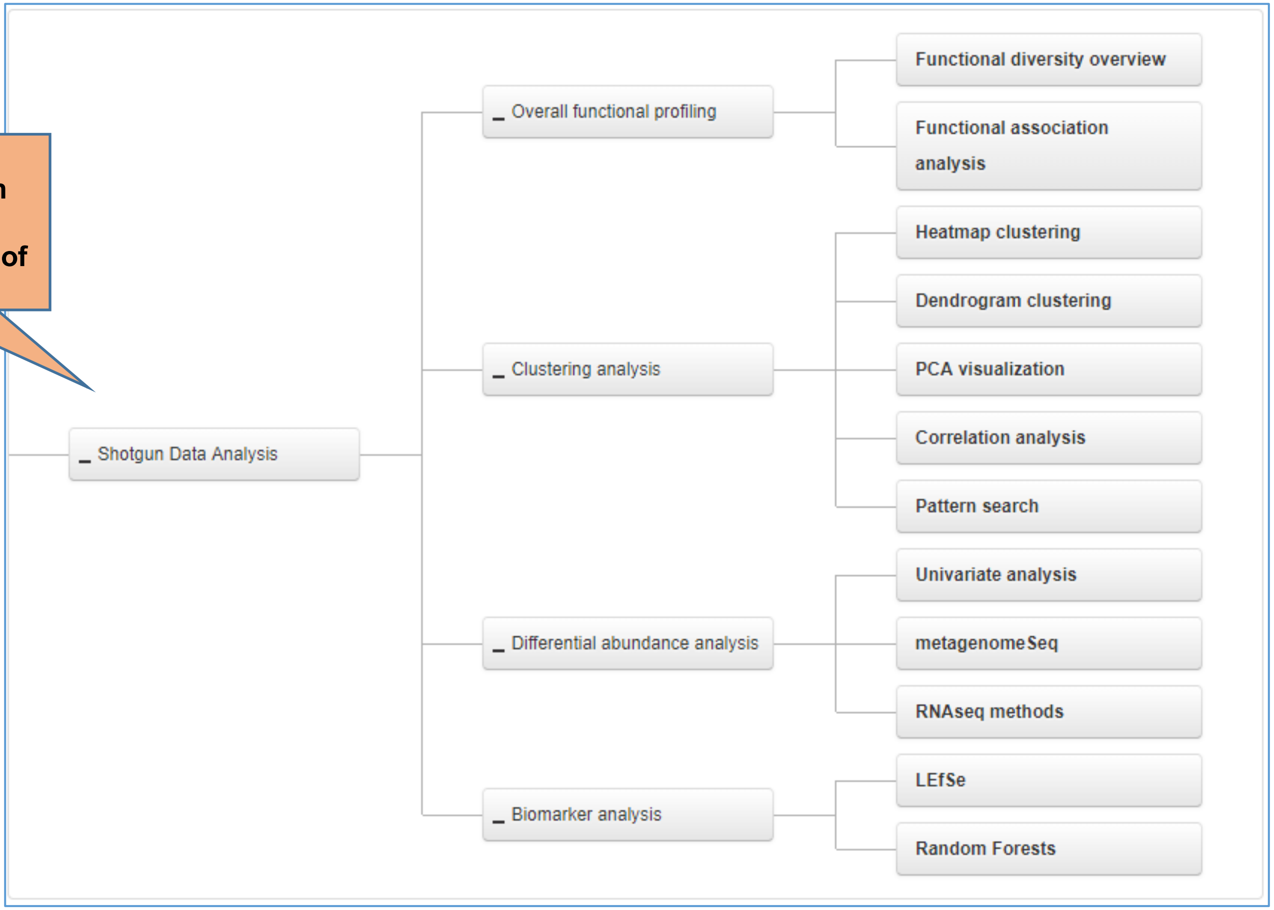
Click “Submit” to continue

- 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. (2 categories: **data scaling and data transformation)**

5. Data analysis

User can get an overview along with comparative and functional analysis of shotgun data.

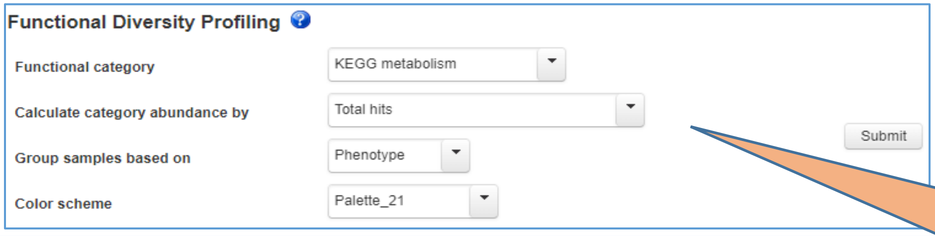


- Shotgun Data Analysis
  - Overall functional profiling
    - Functional diversity overview
    - Functional association analysis
  - Clustering analysis
    - Heatmap clustering
    - Dendrogram clustering
    - PCA visualization
    - Correlation analysis
    - Pattern search
  - Differential abundance analysis
    - Univariate analysis
    - metagenomeSeq
    - RNAseq methods
  - Biomarker analysis
    - LEfSe
    - Random Forests

A. Functional Profiling

User can select from different categories based on input gene id type :

- KEGG metabolism, pathways, modules or COG or EC functional category



Functional Diversity Profiling

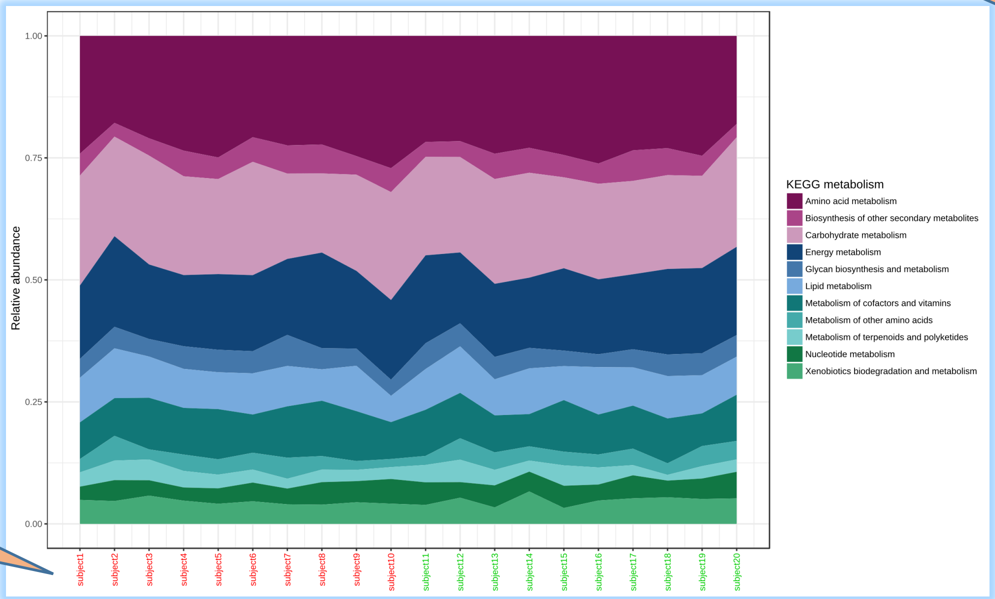
Functional category: KEGG metabolism

Calculate category abundance by: Total hits

Group samples based on: Phenotype

Color scheme: Palette_21

The abundance of functional categories can be estimated by 3 different method to account for one to many gene mapping issue



Samples colored on the basis of selected experimental factor

1. Functional Diversity Profiling

- Samples have been compared to provide a coarser view of the data by collapsing related genes (KO, COG or EC) to observations of functions. (rather than observations of specific genes)
- **5** main functional categories present to collapse within based on **gene ID type** : **KEGG metabolism, pathways, modules and COG or EC functions**.

A. Functional Profiling

2. Functional Association analysis and Metabolic Network Exploration: associations between any functional categories with the experimental factor or sample groups is calculated by integrating the abundance changes of all members within each functional group to evaluate the strength of association

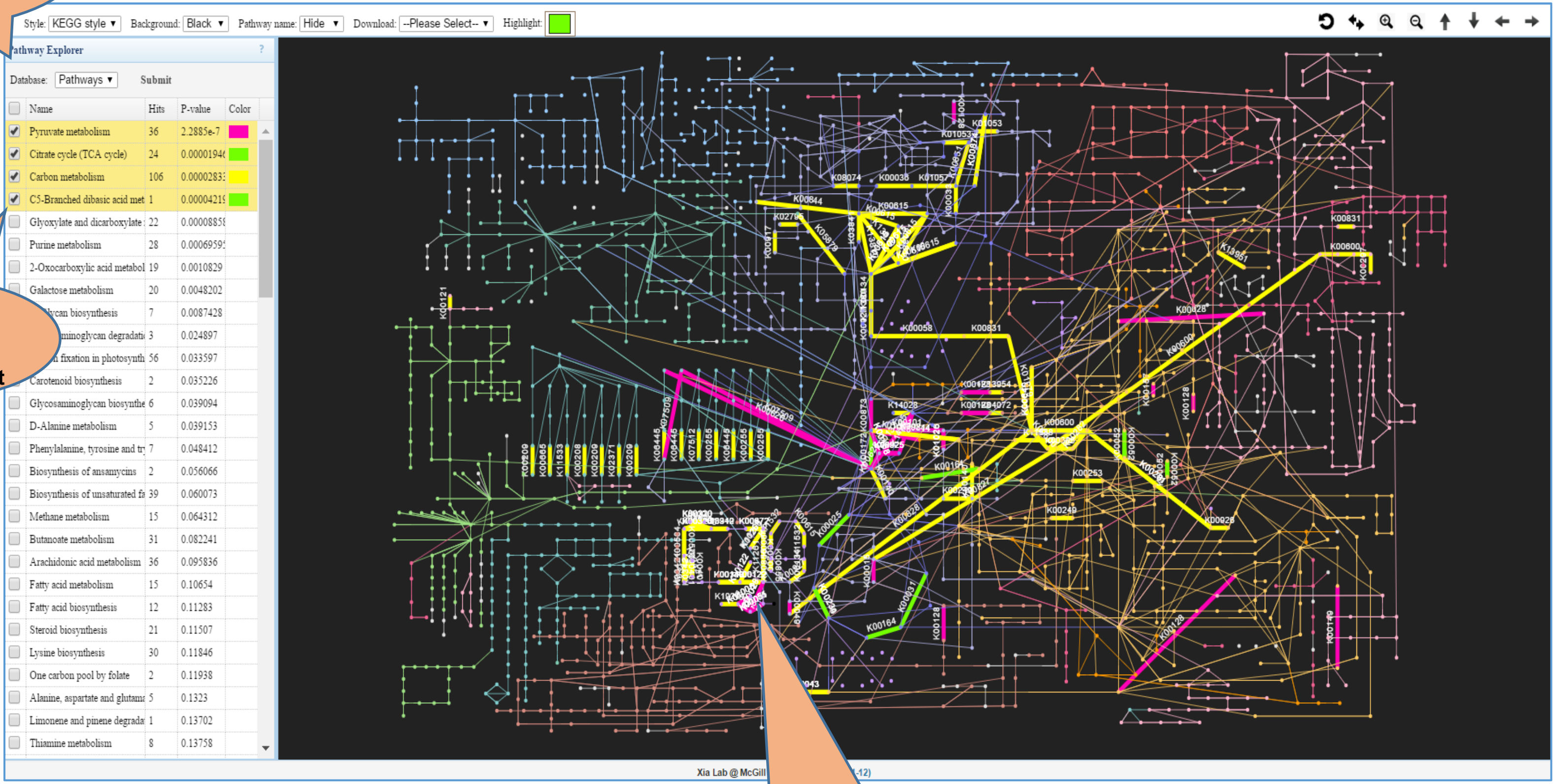
- It is based on the globaltest algorithm. For details:
    - "A global test for groups of genes: testing association with a clinical outcome". Bioinformatics 2004 Jan 1;20(1):93-9.

- Significant functional categories (pathways and modules) can be visualized within Metabolic networks.

A. Functional Profiling

2. Functional Association Testing and Metabolic Network Exploration:

User can chose from 2 functional categories : pathways or modules



Style: KEGG style ▾ Background: Black ▾ Pathway name: Hide ▾ Download: --Please Select-- ▾ Highlight: 🟩

Pathway Explorer

Database: Pathways ▾ Submit

□	Name	Hits	P-value	Color
☑	Pyruvate metabolism	36	2.2885e-7	🟪
☑	Citrate cycle (TCA cycle)	24	0.0001946	🟩
☑	Carbon metabolism	106	0.0002833	🟨
☑	C5-Branched dibasic acid met	1	0.0004219	🟩
□	Glyoxylate and dicarboxylate :	22	0.0008858	
□	Purine metabolism	28	0.0006959	
□	2-Oxocarboxylic acid metabol	19	0.0010829	
□	Galactose metabolism	20	0.0048202	
	lycan biosynthesis	7	0.0087428	
	minoglycan degradati	3	0.024897	
	n fixation in photosynth	56	0.033597	
	Carotenoid biosynthesis	2	0.035226	
□	Glycosaminoglycan biosynthe	6	0.039094	
□	D-Alanine metabolism	5	0.039153	
□	Phenylalanine, tyrosine and try	7	0.048412	
□	Biosynthesis of ansamycins	2	0.056066	
□	Biosynthesis of unsaturated fa	39	0.060073	
□	Methane metabolism	15	0.064312	
□	Butanoate metabolism	31	0.082241	
□	Arachidonic acid metabolism	36	0.095836	
□	Fatty acid metabolism	15	0.10654	
□	Fatty acid biosynthesis	12	0.11283	
□	Steroid biosynthesis	21	0.11507	
□	Lysine biosynthesis	30	0.11846	
□	One carbon pool by folate	2	0.11938	
□	Alanine, aspartate and glutama	5	0.1323	
□	Limonene and pinene degrada	1	0.13702	
□	Thiamine metabolism	8	0.13758	

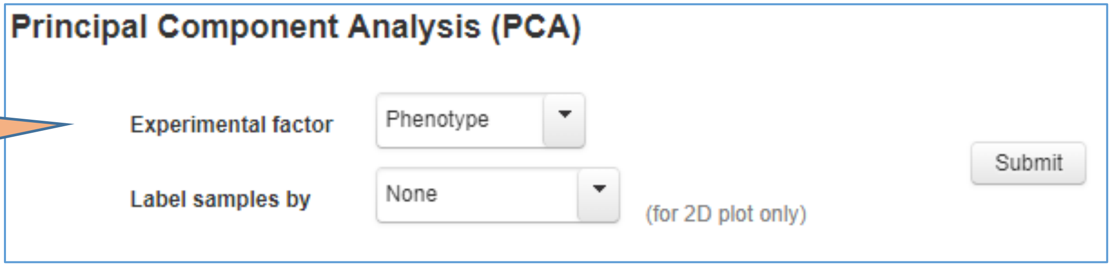
Xia Lab @ McGill -12)

Functional categories association analysis Result

Significant functional categories (pathways or modules) can be highlighted with different colors

B. Clustering Analysis

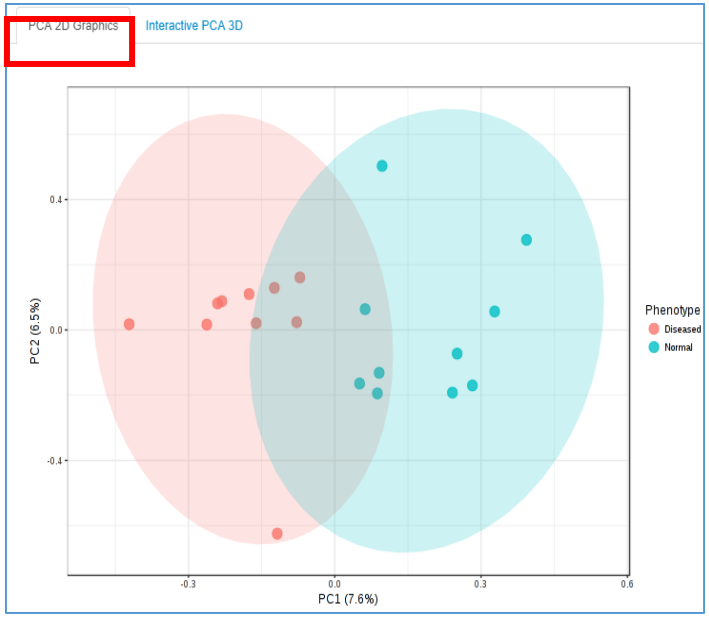
Chose from different sample groups or experimental factors

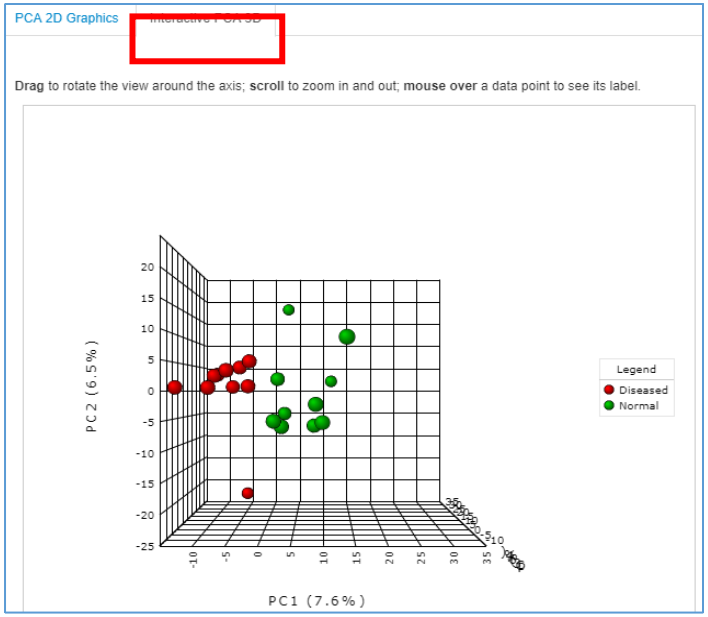


Principal Component Analysis (PCA)

Experimental factor: Phenotype

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



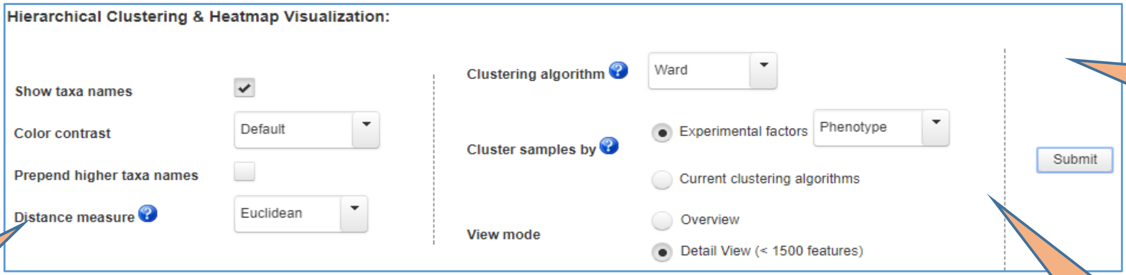


Drag to rotate the view around the axis; **scroll** to zoom in and out; **mouse over** a data point to see its label.

1. Principal Component Analysis (PCA)

- Data reduction technique that can be used to visualize the high-dimensional and complex metagenomic data into 2-3D.
- It emphasizes on variation and shows strong patterns in a dataset. (w.r.t experimental factors)

B. Clustering Analysis



Hierarchical Clustering & Heatmap Visualization:

Show taxa names ☑

Color contrast: Default

Prepend higher taxa names □

Distance measure: Euclidean

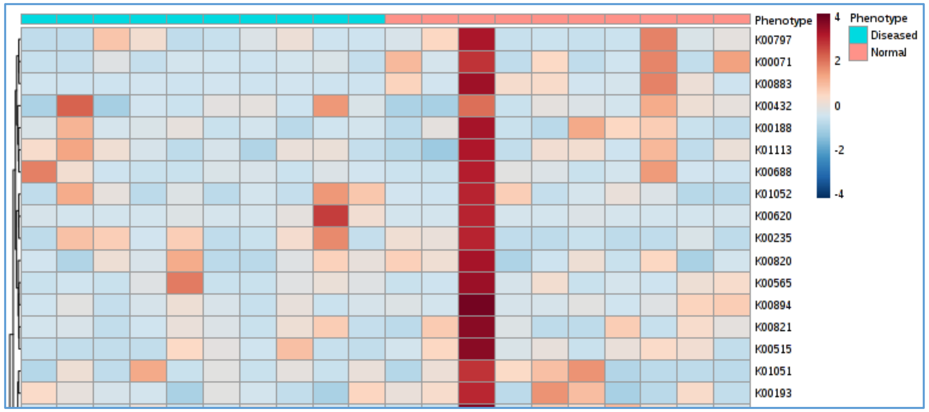
Clustering algorithm: Ward

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

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

Chose from different clustering algorithm.

Chose from different distance measure.



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

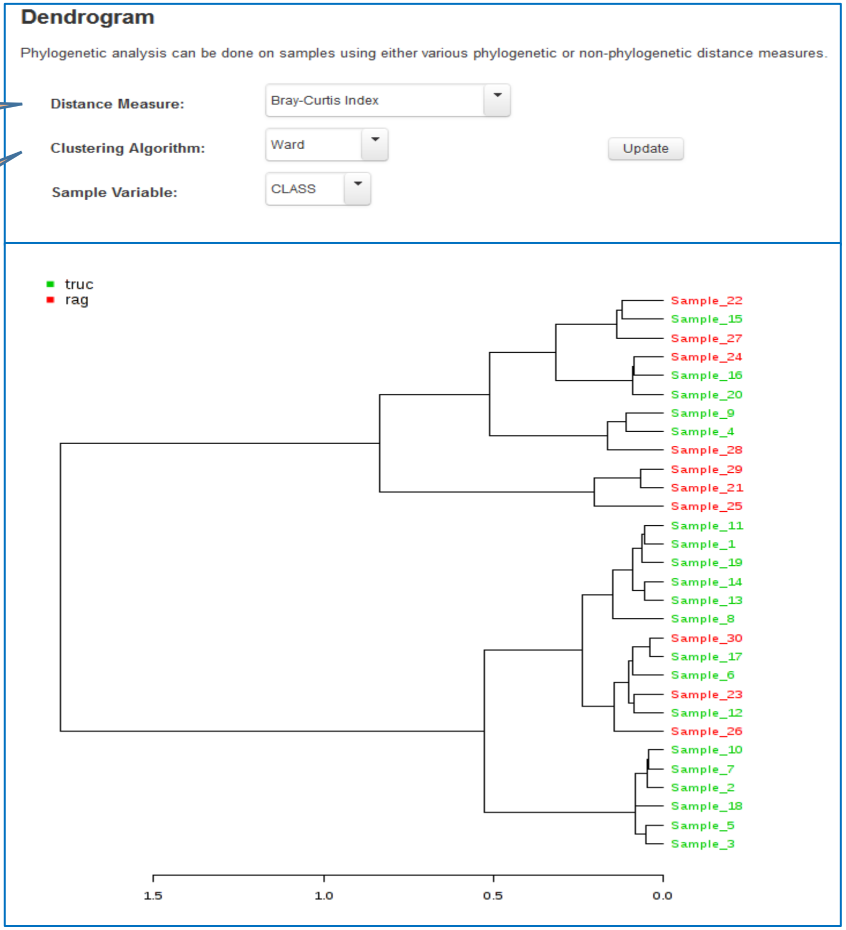
2. Heatmap

- 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)
- Provides a summary of normalized user’s data.

B. Clustering Analysis

Chose from different distance measure.

Chose from different clustering algorithm.



Dendrogram

Phylogenetic analysis can be done on samples using either various phylogenetic or non-phylogenetic distance measures.

Distance Measure: Bray-Curtis Index

Clustering Algorithm: Ward

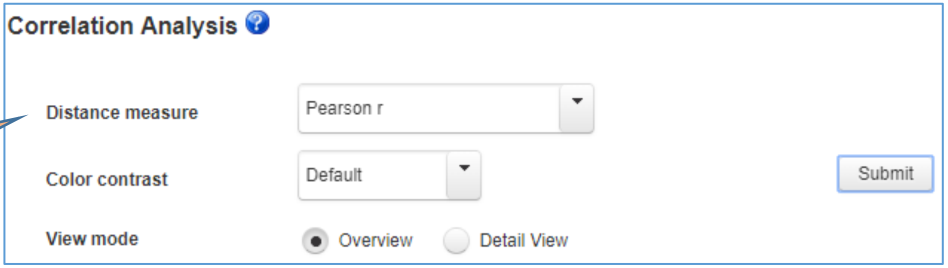
Sample Variable: CLASS

3. Dendrogram

- Performs phylogenetic analysis on samples using ordination based distance measures. (support for 5 most widely used)

B. Clustering analysis

3 most common method supported for performing correlation analysis

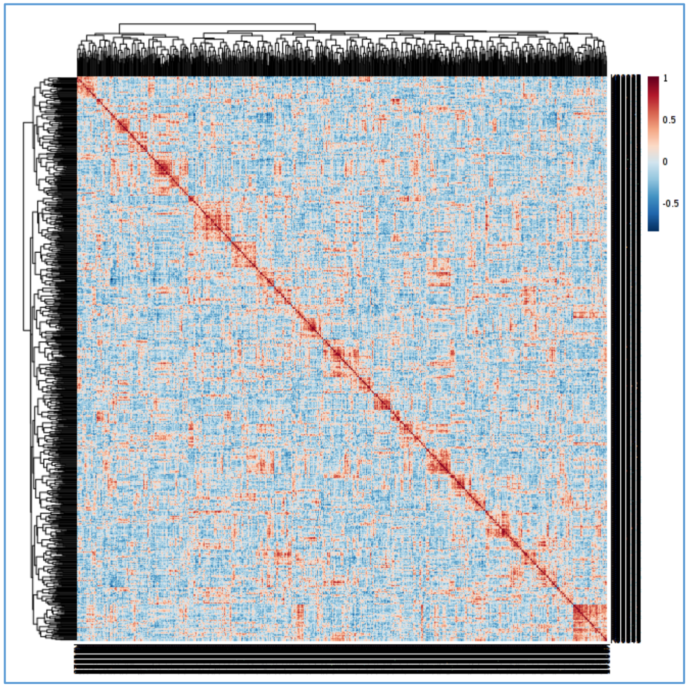


Correlation Analysis

Distance measure: Pearson r

Color contrast: Default

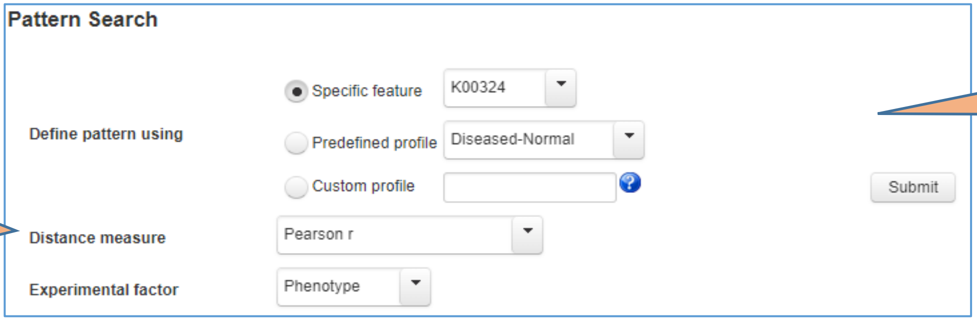
View mode: ● Overview ○ Detail View



4. Correlation analysis

- Helps in identifying biologically or biochemically meaningful relationship between features. (genes)

B. Clustering analysis



Pattern Search

Define pattern using

● Specific feature: K00324

○ Predefined profile: Diseased-Normal

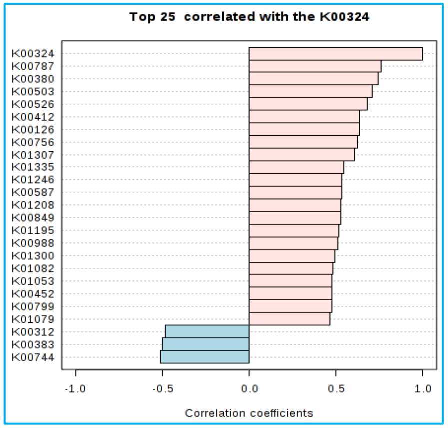
○ Custom profile:

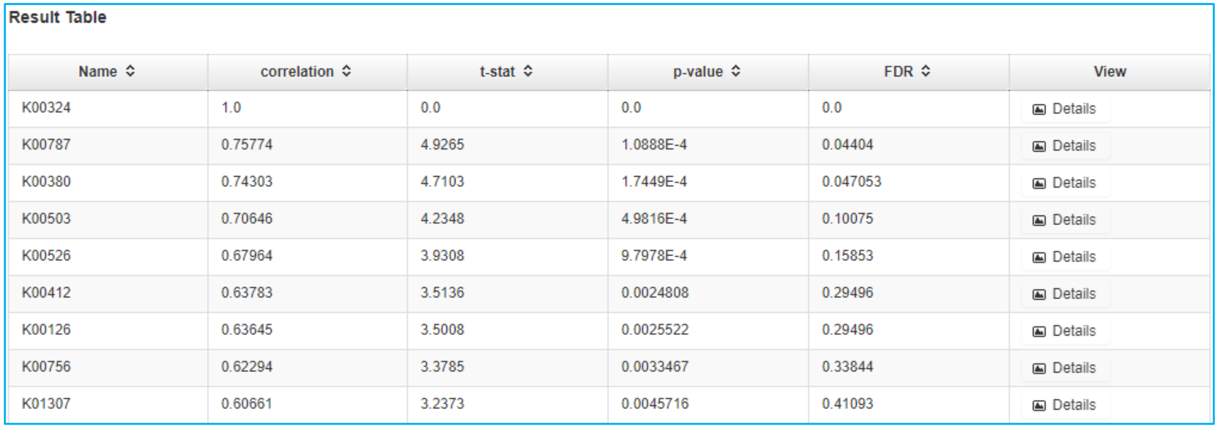
Distance measure: Pearson r

Experimental factor: Phenotype

User can define their own pattern based on their interest

3 most common method supported for performing correlation analysis





Result Table

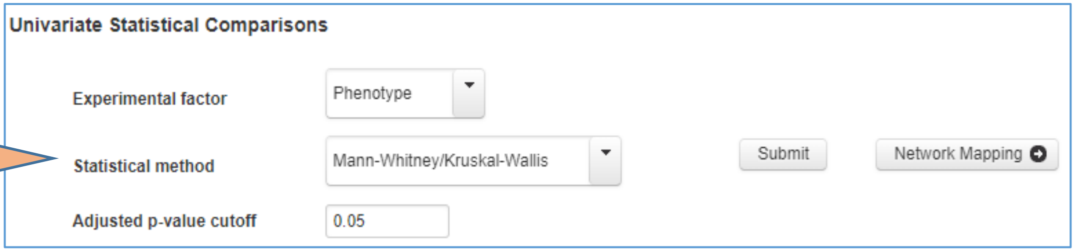
Name	correlation	t-stat	p-value	FDR	View
K00324	1.0	0.0	0.0	0.0	Details
K00787	0.75774	4.9265	1.0888E-4	0.04404	Details
K00380	0.74303	4.7103	1.7449E-4	0.047053	Details
K00503	0.70646	4.2348	4.9816E-4	0.10075	Details
K00526	0.67964	3.9308	9.7978E-4	0.15853	Details
K00412	0.63783	3.5136	0.0024808	0.29496	Details
K00126	0.63645	3.5008	0.0025522	0.29496	Details
K00756	0.62294	3.3785	0.0033467	0.33844	Details
K01307	0.60661	3.2373	0.0045716	0.41093	Details

5. 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.

C. (a) Differential abundance analysis

Chose from parametric or non-parametric statistical tests



Univariate Statistical Comparisons

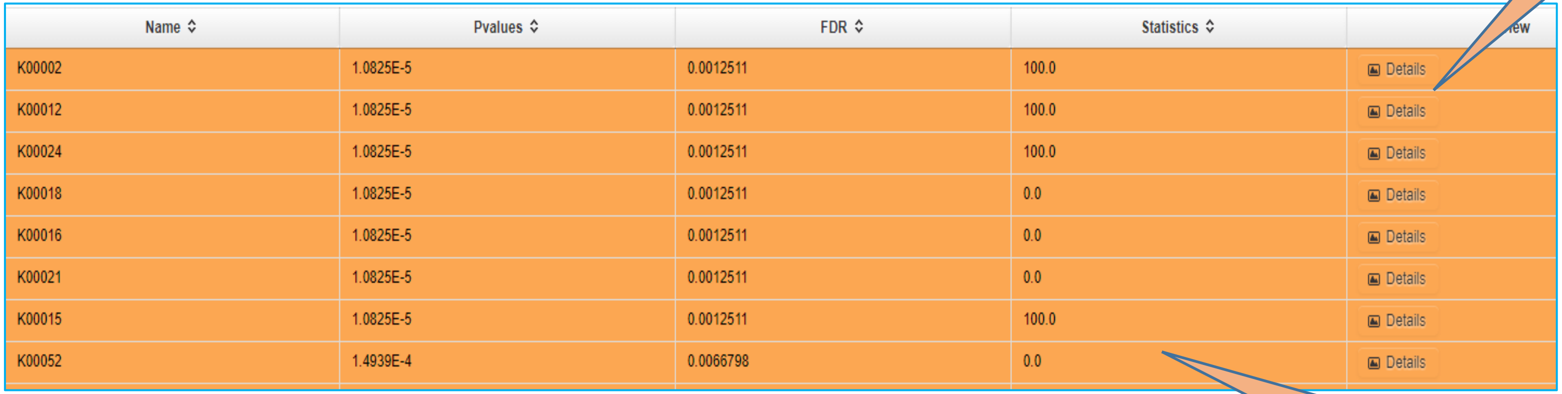
Experimental factor: Phenotype

Statistical method: Mann-Whitney/Kruskal-Wallis

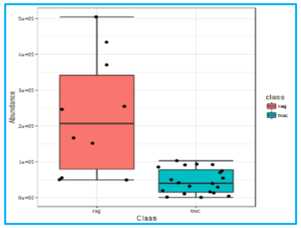
Adjusted p-value cutoff: 0.05

Click here to visualize the differential genes in metabolic networks

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



Name	Pvalues	FDR	Statistics	iew
K00002	1.0825E-5	0.0012511	100.0	Details
K00012	1.0825E-5	0.0012511	100.0	Details
K00024	1.0825E-5	0.0012511	100.0	Details
K00018	1.0825E-5	0.0012511	0.0	Details
K00016	1.0825E-5	0.0012511	0.0	Details
K00021	1.0825E-5	0.0012511	0.0	Details
K00015	1.0825E-5	0.0012511	100.0	Details
K00052	1.4939E-4	0.0066798	0.0	Details



Differential abundant genes (KO) are highlighted in orange color

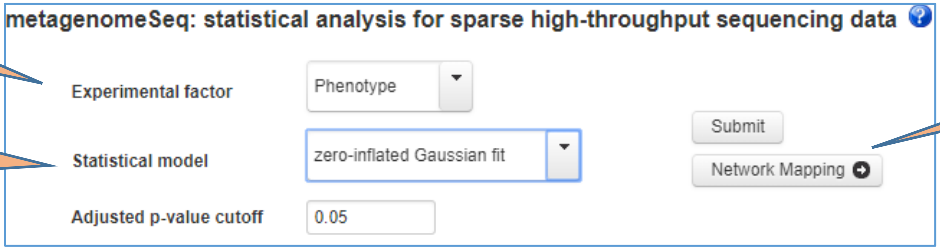
1. Univariate Statistical Comparisons

- 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.

C. (a) Differential Abundance Analysis

Chose from different Experimental factors

Chose from 2 statistical models based on number of groups



metagenomeSeq: statistical analysis for sparse high-throughput sequencing data

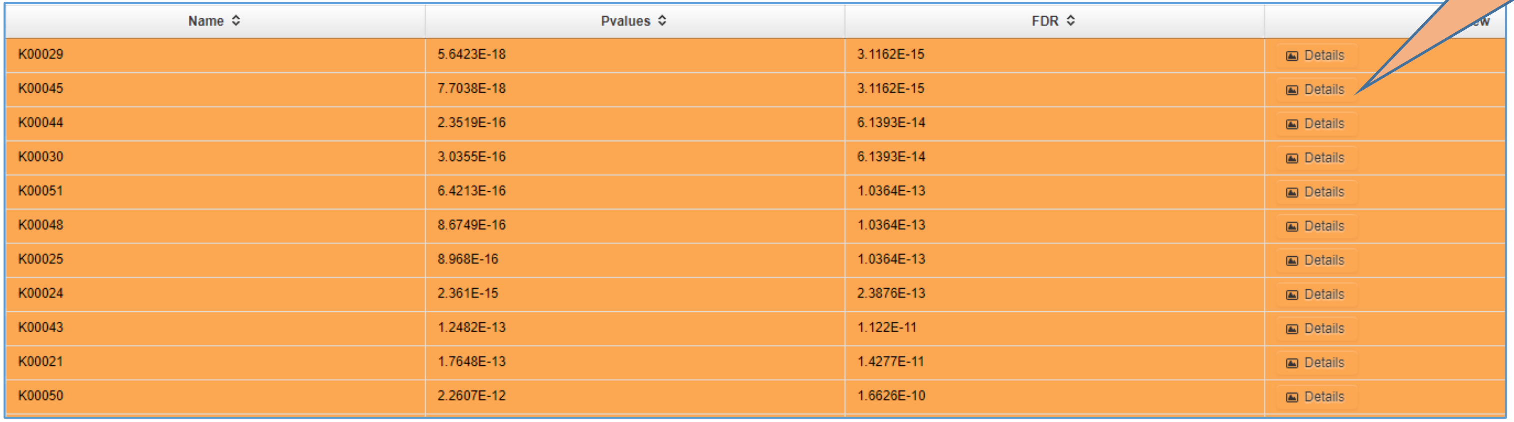
Experimental factor: Phenotype

Statistical model: zero-inflated Gaussian fit

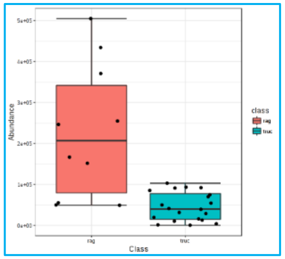
Adjusted p-value cutoff: 0.05

Click to perform metabolic network Mapping

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



Name	Pvalues	FDR	ew
K00029	5.6423E-18	3.1162E-15	Details
K00045	7.7038E-18	3.1162E-15	Details
K00044	2.3519E-16	6.1393E-14	Details
K00030	3.0355E-16	6.1393E-14	Details
K00051	6.4213E-16	1.0364E-13	Details
K00048	8.6749E-16	1.0364E-13	Details
K00025	8.968E-16	1.0364E-13	Details
K00024	2.361E-15	2.3876E-13	Details
K00043	1.2482E-13	1.122E-11	Details
K00021	1.7648E-13	1.4277E-11	Details
K00050	2.2607E-12	1.6626E-10	Details

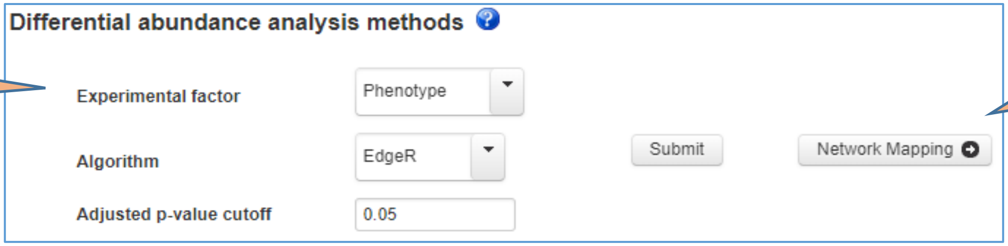


2. metagenomeSeq

- R package which aims to 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)

C. (a) Differential Abundance Analysis

Chose from different Experimental factors



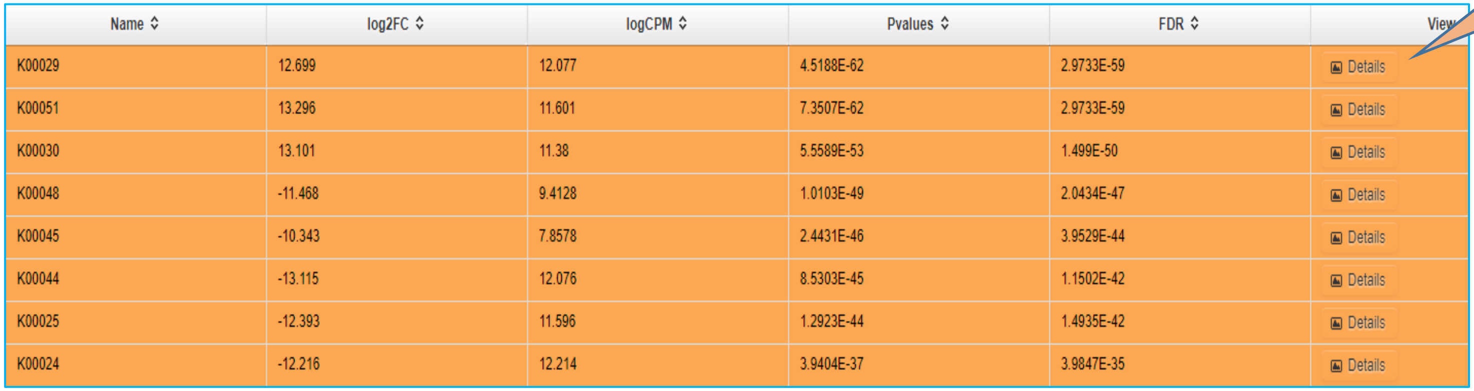
Differential abundance analysis methods

Experimental factor: Phenotype

Algorithm: EdgeR

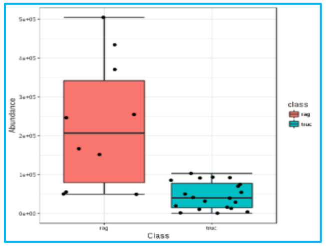
Adjusted p-value cutoff: 0.05

Click to perform Functional Enrichment Analysis on differentially abundant features



Name	log2FC	logCPM	Pvalues	FDR	View
K00029	12.699	12.077	4.5188E-62	2.9733E-59	Details
K00051	13.296	11.601	7.3507E-62	2.9733E-59	Details
K00030	13.101	11.38	5.5589E-53	1.499E-50	Details
K00048	-11.468	9.4128	1.0103E-49	2.0434E-47	Details
K00045	-10.343	7.8578	2.4431E-46	3.9529E-44	Details
K00044	-13.115	12.076	8.5303E-45	1.1502E-42	Details
K00025	-12.393	11.596	1.2923E-44	1.4935E-42	Details
K00024	-12.216	12.214	3.9404E-37	3.9847E-35	Details

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



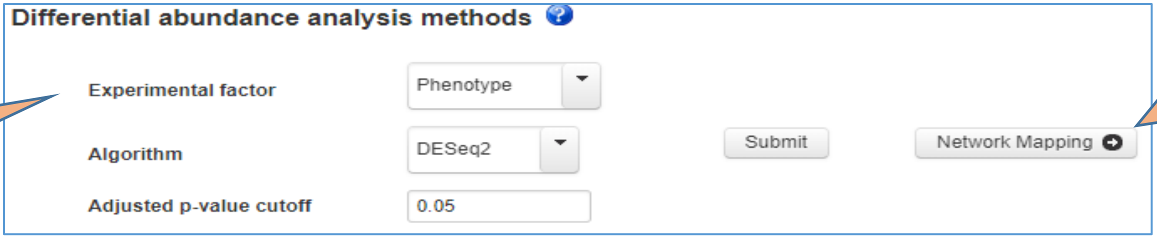
3. EdgeR

- 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.

Note: If no significant gene will be identified using p-value cut-off, then top 500 genes based on their p-values will be used for network analysis.

C. (a) Differential Abundance Analysis

Chose from different Experimental factors



Differential abundance analysis methods

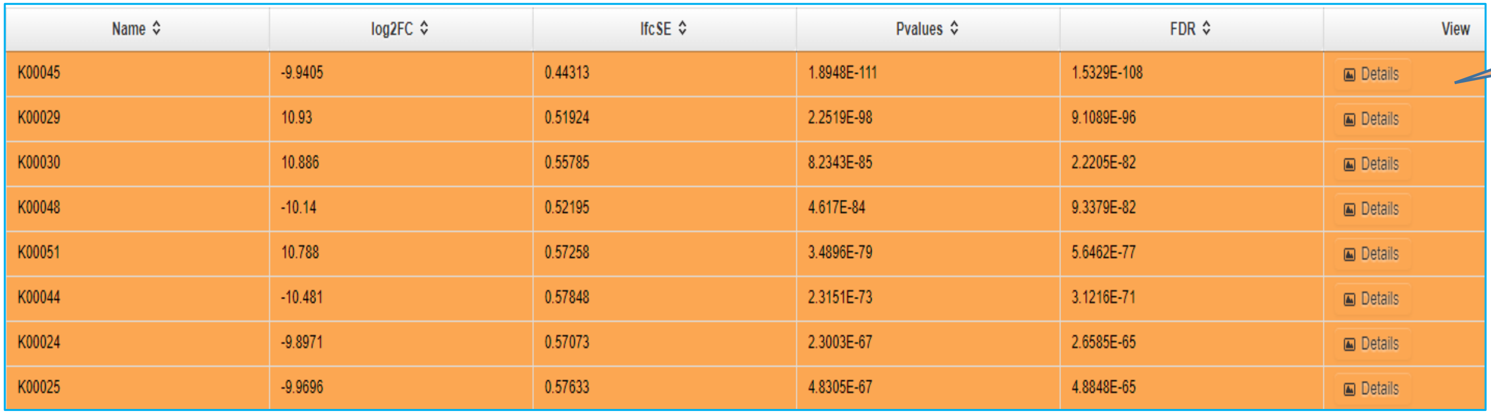
Experimental factor: Phenotype

Algorithm: DESeq2

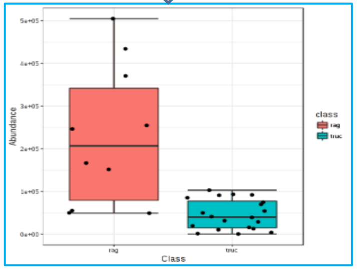
Adjusted p-value cutoff: 0.05

Click to perform Functional Enrichment Analysis on differentially abundant features

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



Name	log2FC	lfcSE	Pvalues	FDR	View
K00045	-9.9405	0.44313	1.8948E-111	1.5329E-108	Details
K00029	10.93	0.51924	2.2519E-98	9.1089E-96	Details
K00030	10.886	0.55785	8.2343E-85	2.2205E-82	Details
K00048	-10.14	0.52195	4.617E-84	9.3379E-82	Details
K00051	10.788	0.57258	3.4896E-79	5.6462E-77	Details
K00044	-10.481	0.57848	2.3151E-73	3.1216E-71	Details
K00024	-9.8971	0.57073	2.3003E-67	2.6585E-65	Details
K00025	-9.9696	0.57633	4.8305E-67	4.8848E-65	Details

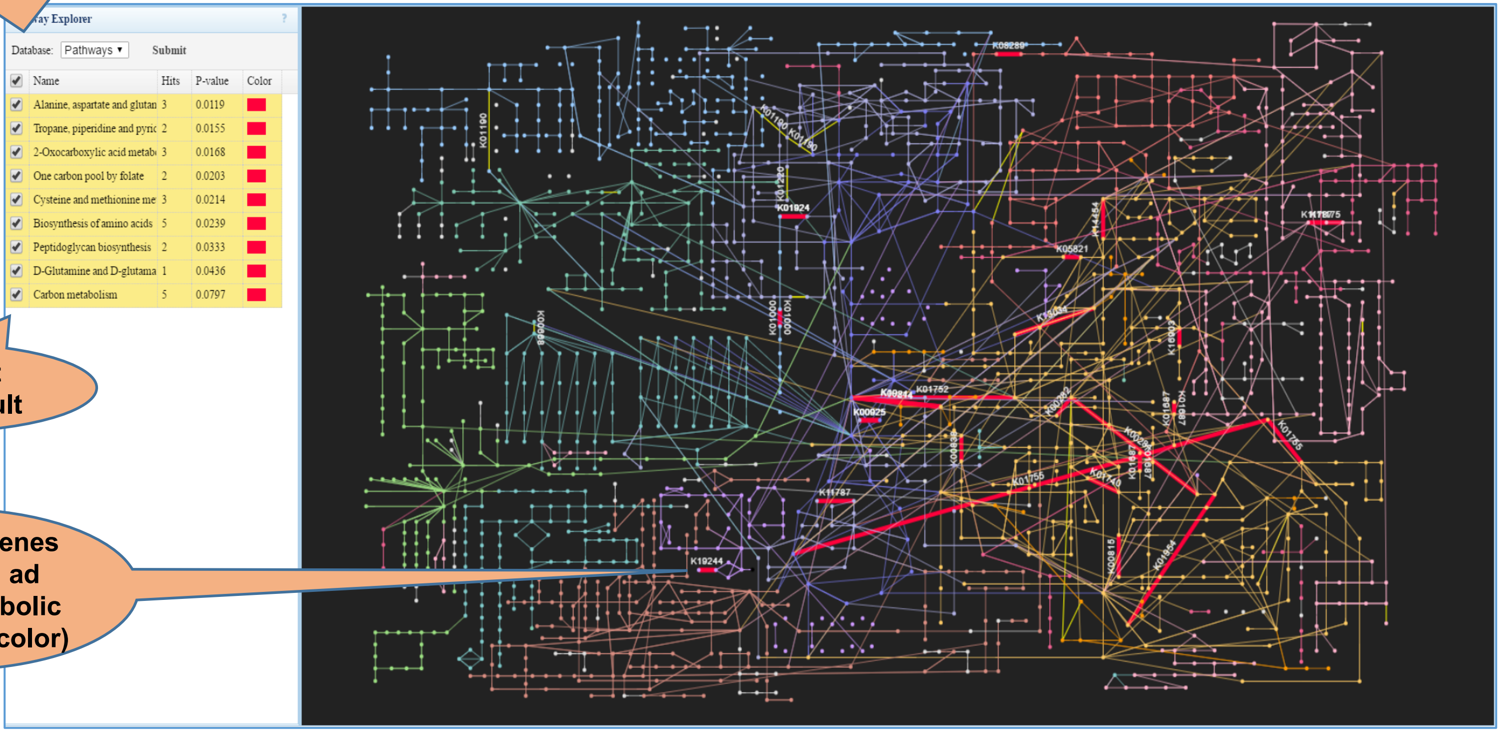


4. DESeq2

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

Note: If no significant gene will be identified using p-value cut-off, then top 500 genes based on their p-values will be used for network analysis.

C. (b) Network and Functional Enrichment Analysis



way Explorer

Database: Pathways ▾ Submit

☑	Name	Hits	P-value	Color
☑	Alanine, aspartate and glutan	3	0.0119	🟥
☑	Tropane, piperidine and pyric	2	0.0155	🟥
☑	2-Oxocarboxylic acid metab	3	0.0168	🟥
☑	One carbon pool by folate	2	0.0203	🟥
☑	Cysteine and methionine me	3	0.0214	🟥
☑	Biosynthesis of amino acids	5	0.0239	🟥
☑	Peptidoglycan biosynthesis	2	0.0333	🟥
☑	D-Glutamine and D-glutama	1	0.0436	🟥
☑	Carbon metabolism	5	0.0797	🟥

User can choose from either KEGG metabolic pathways or modules.

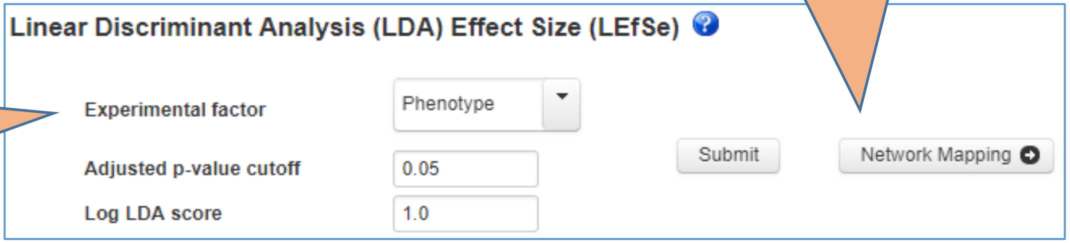
Enrichment Analysis result

Significant genes highlighted ad edge in metabolic network (red color)

- Significant genes from differential analysis are mapped to KO IDs;
- Functional enrichment analysis is performed;(KEGG modules or pathways)
- The enriched pathways or modules can be interactively visualized within the metabolic networks.

D. Biomarker analysis

Click here to visualize the differential genes in metabolic networks



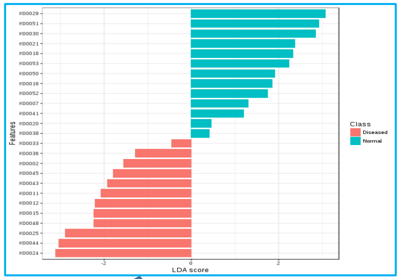
Linear Discriminant Analysis (LDA) Effect Size (LEfSe)

Experimental factor: Phenotype

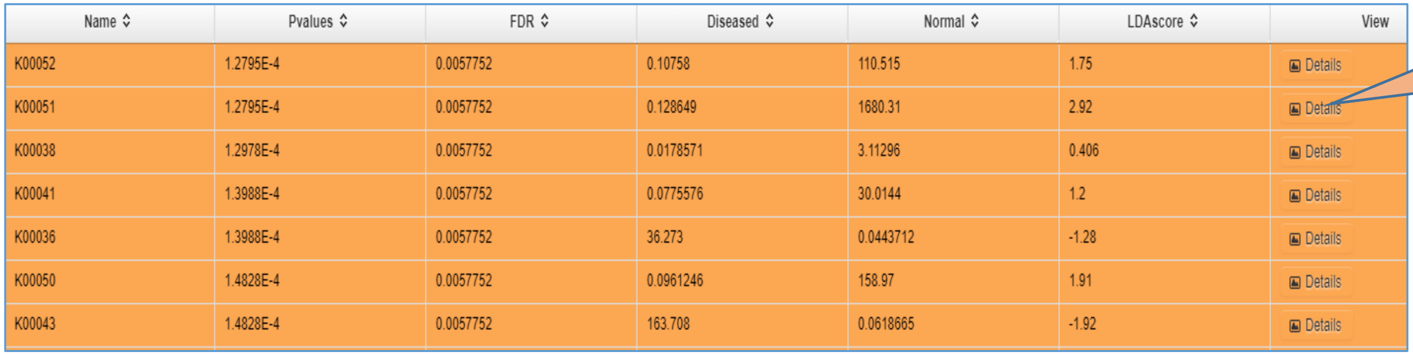
Adjusted p-value cutoff: 0.05

Log LDA score: 1.0

Chose from different Experimental factors

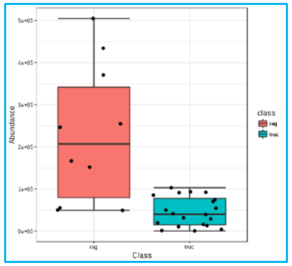


Effect size (LDA score) of differential features



Name	Pvalues	FDR	Diseased	Normal	LDAscore	View
K00052	1.2795E-4	0.0057752	0.10758	110.515	1.75	Details
K00051	1.2795E-4	0.0057752	0.128649	1680.31	2.92	Details
K00038	1.2978E-4	0.0057752	0.0178571	3.11296	0.406	Details
K00041	1.3988E-4	0.0057752	0.0775576	30.0144	1.2	Details
K00036	1.3988E-4	0.0057752	36.273	0.0443712	-1.28	Details
K00050	1.4828E-4	0.0057752	0.0961246	158.97	1.91	Details
K00043	1.4828E-4	0.0057752	163.708	0.0618665	-1.92	Details

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



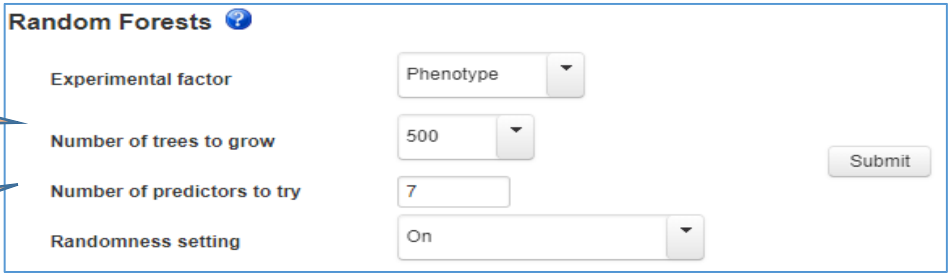
1. LEfSe

- 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)
- Network and functional enrichment analysis can also be performed on DE genes.

D. Biomarker analysis

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

No. of predictors for each node



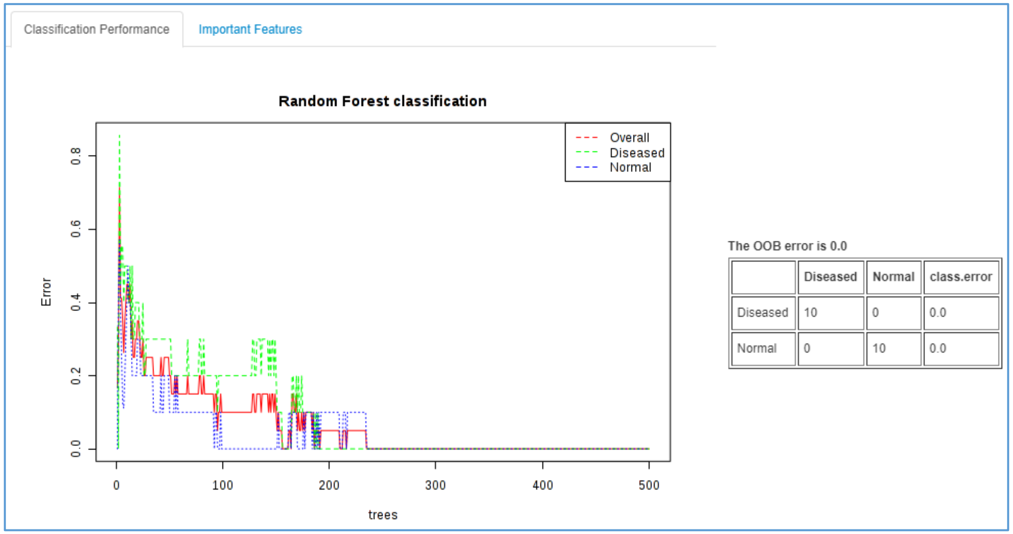
Random Forests

Experimental factor: Phenotype

Number of trees to grow: 500

Number of predictors to try: 7

Randomness setting: On



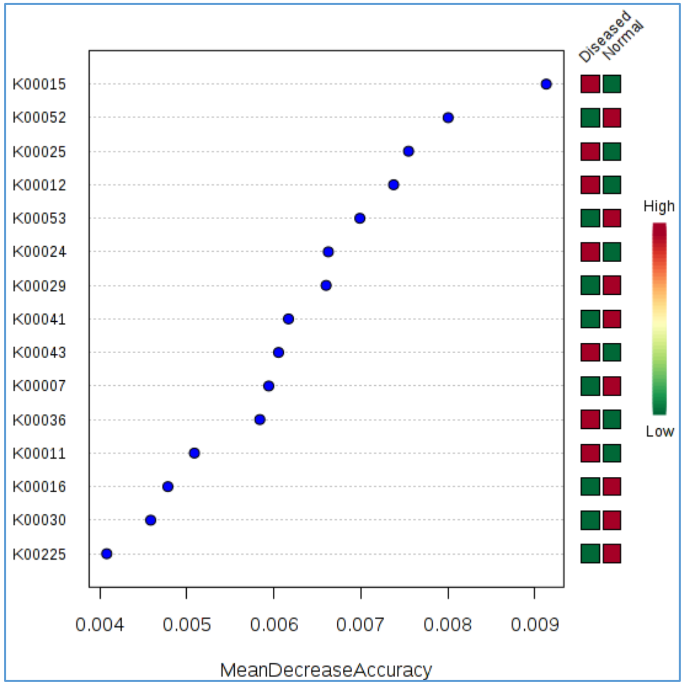
The OOB error is 0.0

	Diseased	Normal	class.error
Diseased	10	0	0.0
Normal	0	10	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.

D. 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

==END==