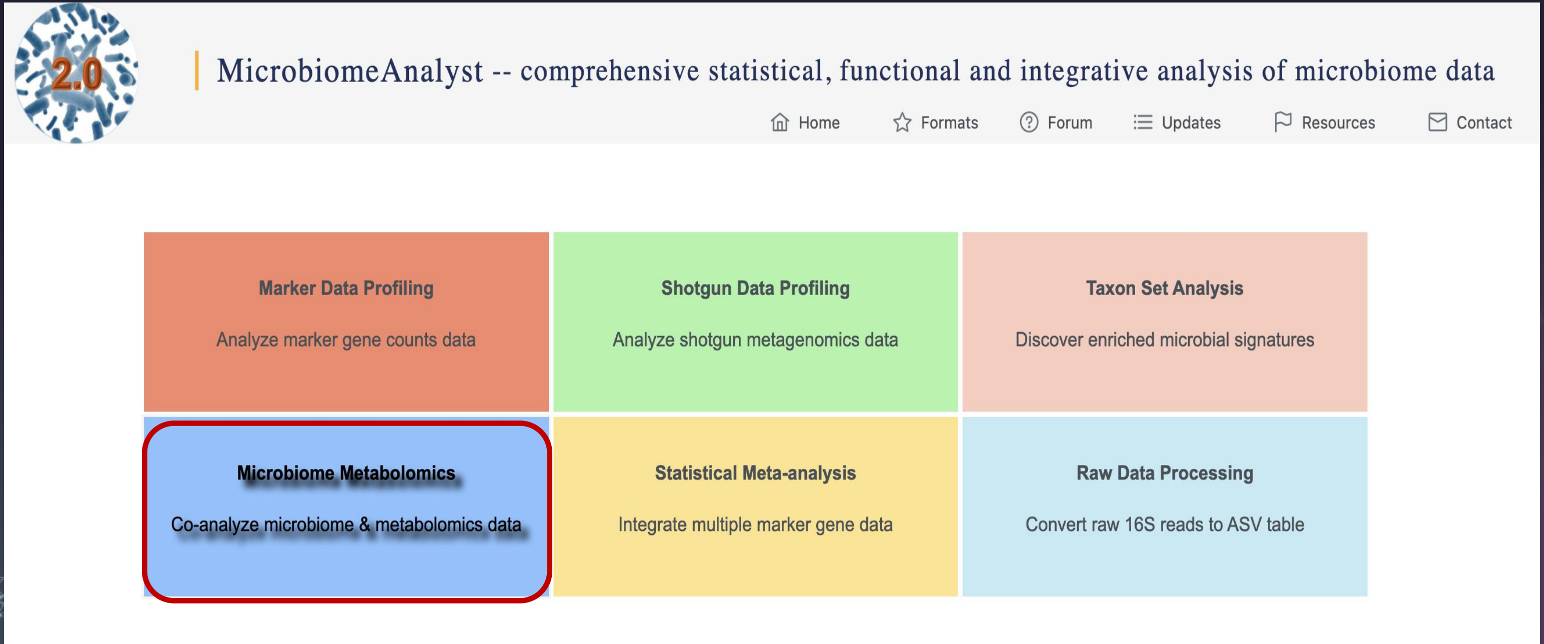
MicrobiomeAnalyst 2.0

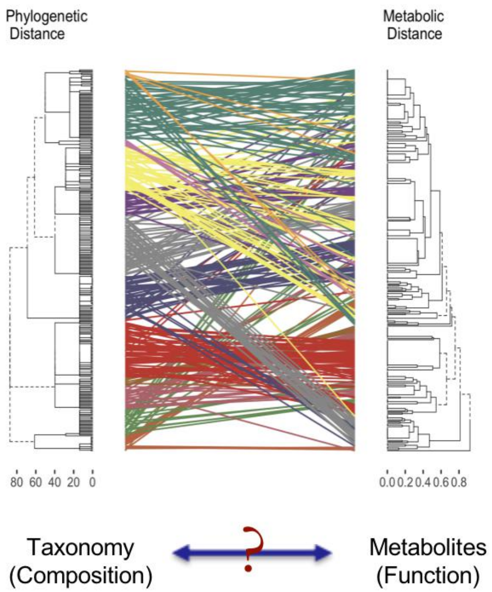
Comprehensive statistical, functional and integrative analysis of microbiome data

xialab@mcgill 2023-Mar-02

Tutorial for Microbiome Metabolomics



Motavition



- Metabolomics is critical to connecting microbial community composition and phenotypes at the level of altered metabolic processes in recent microbiome studies
- General statistical correlation analysis often leads to many false positives, making biological interpretation difficult.
- Integrating high dimensional microbiome and their corresponding metabolomics data remains a significant challenge.

Overview

Goal: to support intergrative analysis for microbiome and metabolomics data in both statistically and biologically meaningful perspective

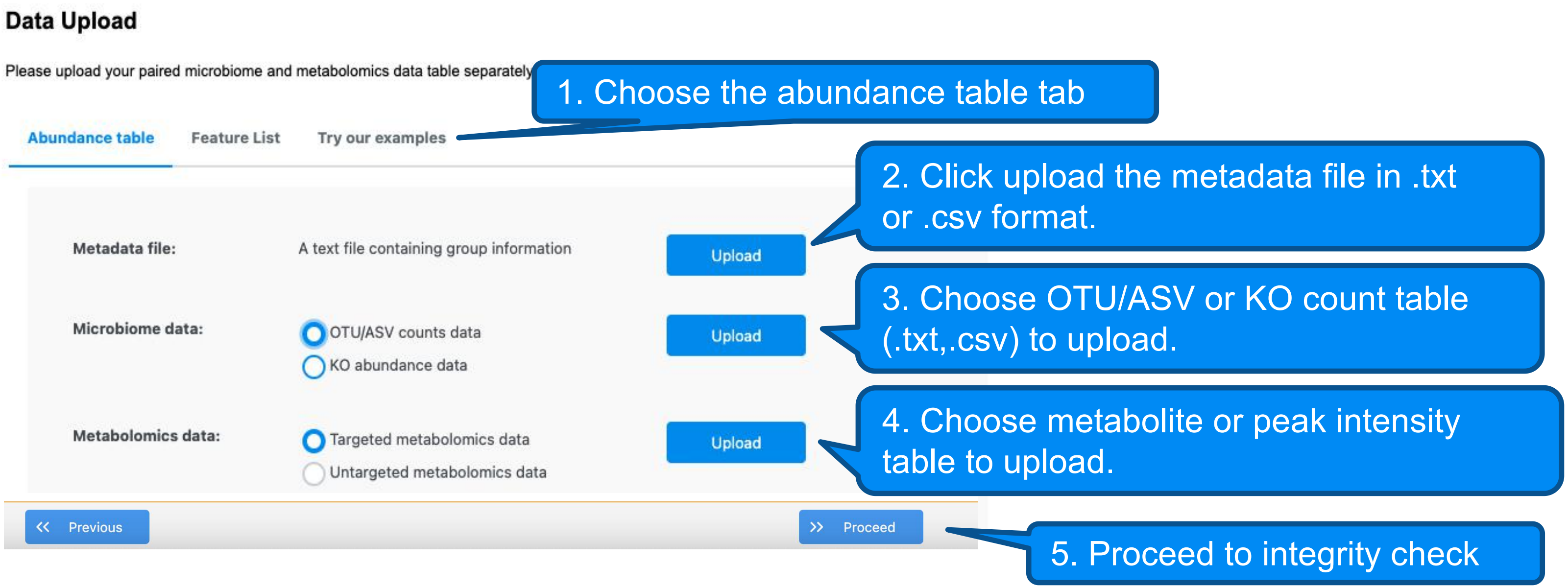
Approches (refer to method selection page for more details)**:**

- **Dimensionality reduction**: Procrustes analysis (PA) and data integration analysis for biomarker discovery using latent components (DIABLO) are implemented to reveal the overall pattern of the paired microbiome and metabolomics datasets.
- **Metabolic network and pathway analysis:** Support pathway enrichment analysis for KO, metabolites and peaks against customized metabolic space based on the taxonomy input and visualization in an interactive network.
- **Microbiome-metabolome correlation analysis:** Provide statistical correlation, model-based correlation analysis as well as integrated corelation analysis

Data type:

- Microbiome data: ASV/OTU/KO count table, taxonomy list
- Metabolomics data: Targeted (metabolite) / Untargeted (Peak) as intensity table or list

I. Start from paired abundance tables:

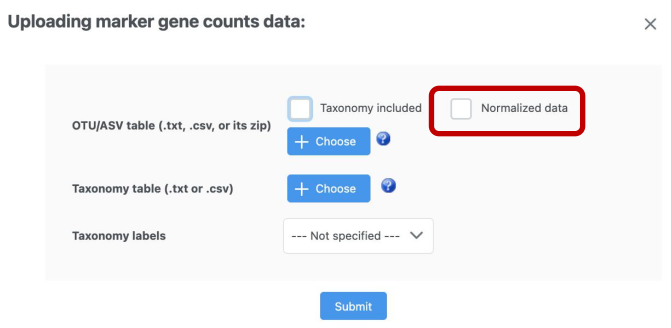


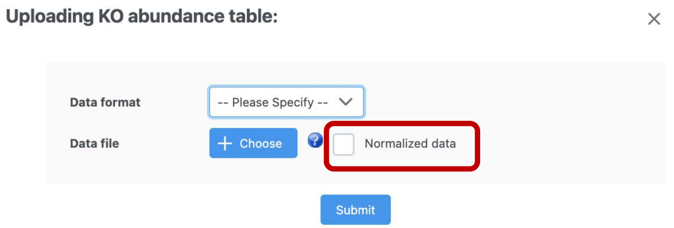
- Metadata file should be consistent between the two data types.
- Taxonomy annotation is required for the marker gene data which can be provided as a table or as IDs in OTU/ASV table.
- For metabolites, users need to select the ID type. Names, KEGG IDs and HMDB IDs are accepeted. For peaks, general formats (mz, mz__rt and mz@rt) are accepted.

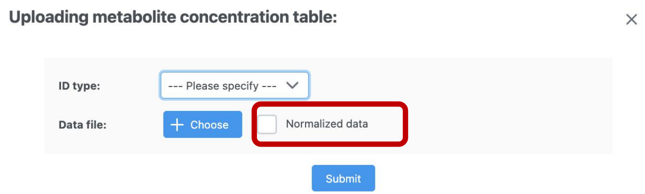
Notes for normalized data upload:

It is highly advised to upload your microbiome abundance table containing raw counts to benefit the best practices for data analysis. However, if your data **has already been normalized**:

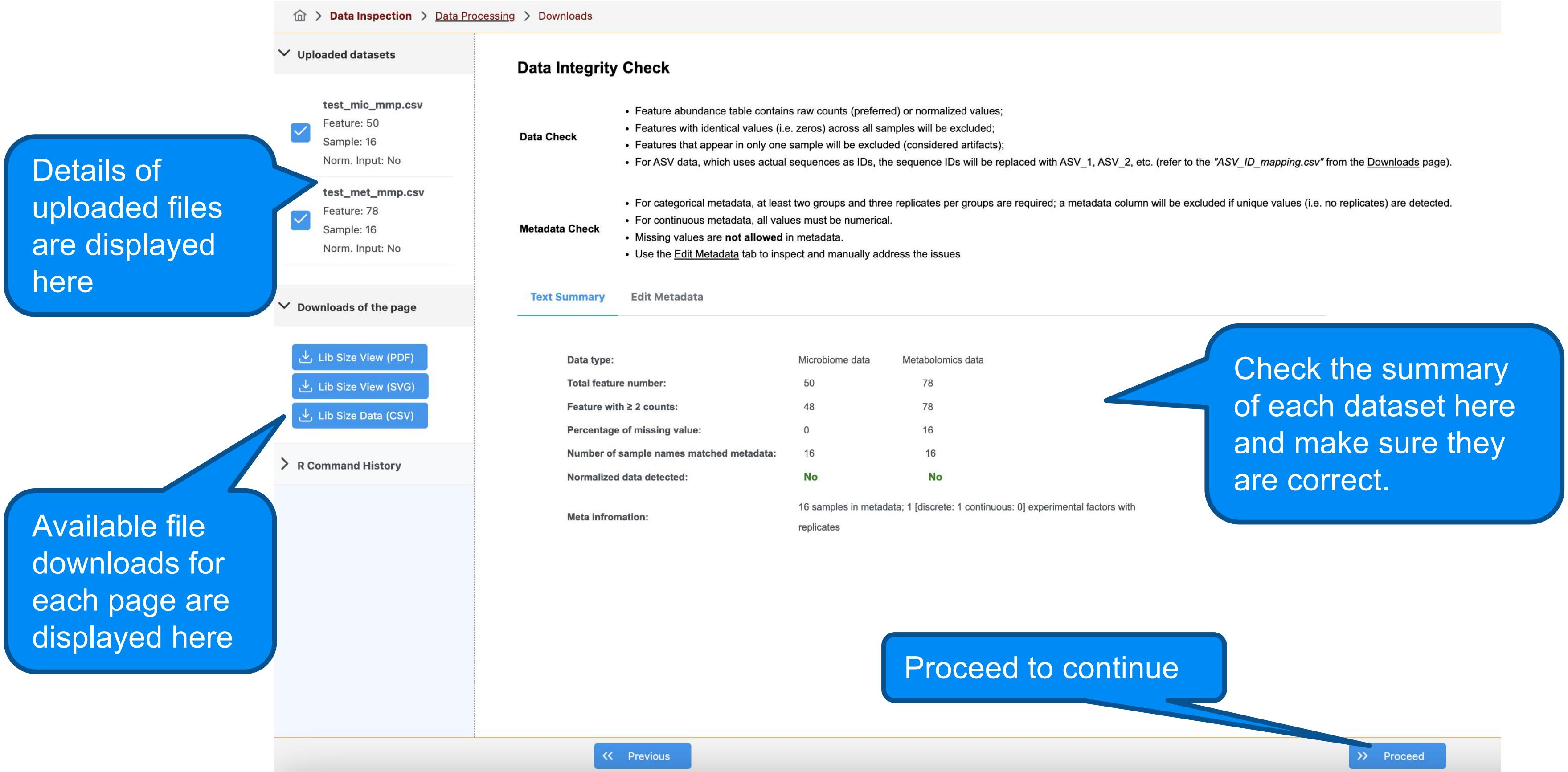
- Indicate it is Normalized data using the check box (shown in the red box to the different data types)
- Bypass data filtering and normalization (Optional)
- The function for predicting metabolites from different taxonomy will become inappropriate during data analysis.





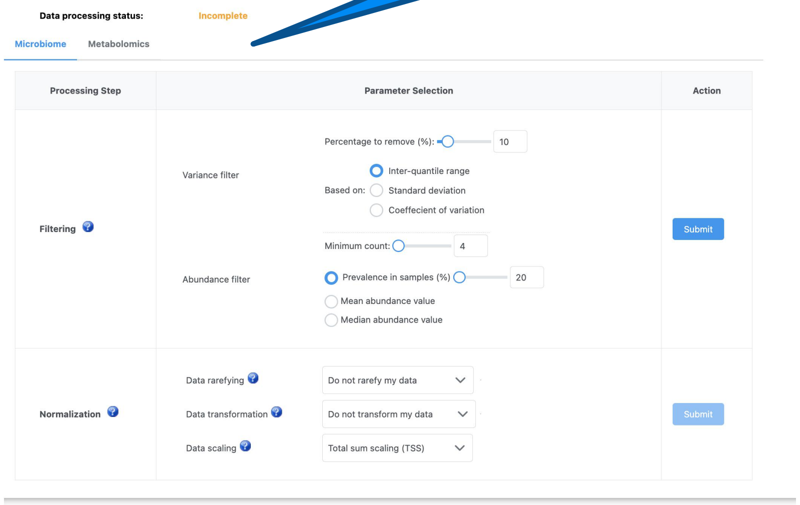


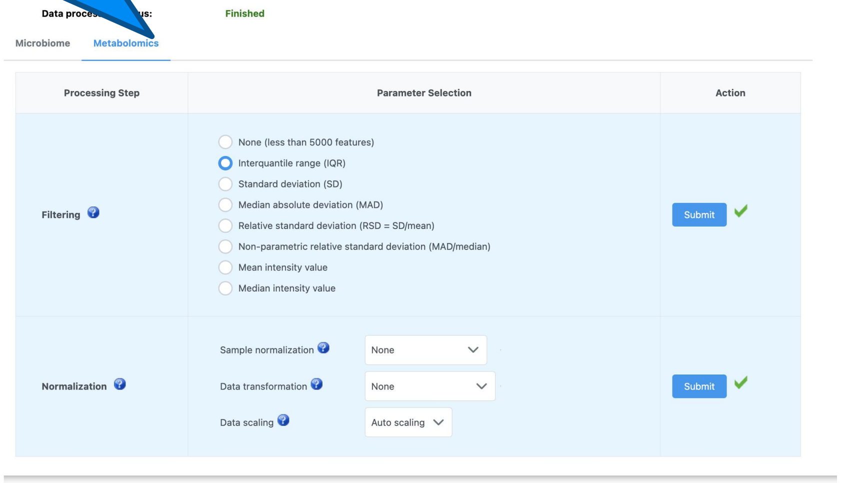
Data integrity check:



Data processing:

Click here to change the omics type for processing



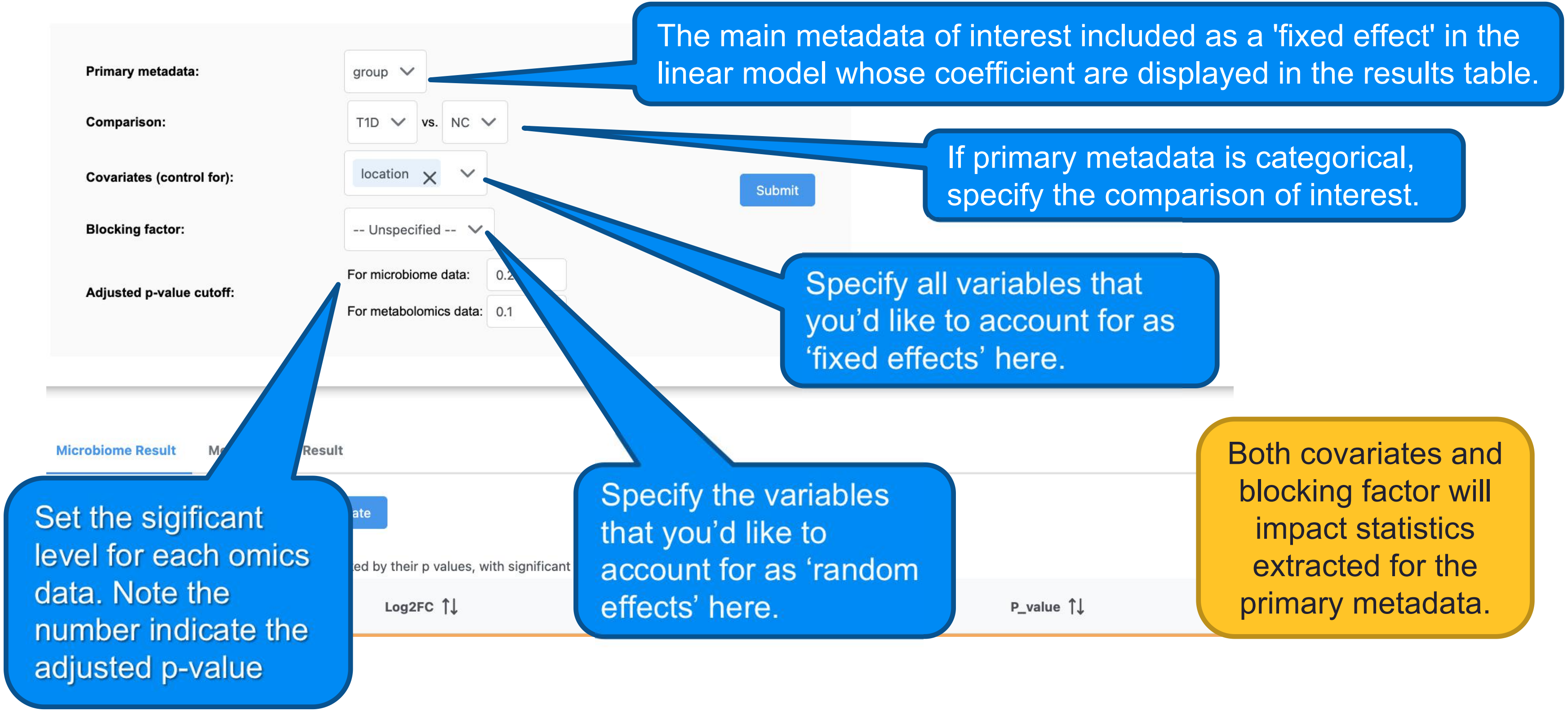


- Different approaches are applied to different omics type
- Background color and ticks suggest the completeness of each processing step , as exemplified by the figure on the right.

Data processing:

Comparision analysis:

- Aim to identify significant features from individual omics.
- MaAsLin2 is used for microbiome and Limma is used for metabolomics data. Both methods are based on general linear models.



Comparision analysis:

- Comparision results are displayed as tables including fold change, raw p-value and adjusted p-value. Feature details can be viewed as box plot by clicking the picture icon.

Check the comparision results for each omics type by selecting the corresponding tab

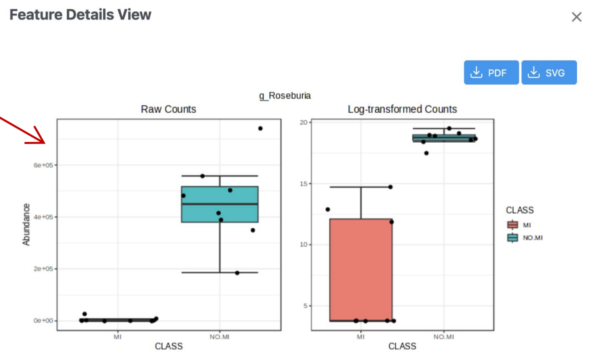
Microbiome Result | Metabolomics Result

Taxonomy level: Genus ▾ [Update]

The result for different taxonomy can be updated and checked here.

The table below shows at most 500 features ranked by their p v[illegible]th significant features highlighted in orange.

Name	Log2FC	[illegible]	P_value	FDR	View
g_Roseburia	6.99	0.[illegible]	3.98E-9	1.47E-7	
g_Mucispirillum	6.17	[illegible]	[illegible]	[illegible]	
g_Enterococcus	-4.84	[illegible]	[illegible]	[illegible]	
g_Acetatifactor	2.53	[illegible]	[illegible]	[illegible]	
g_Oscillibacter	2.98	0.464	1.58E-5	1.16E-4	
g_unclassified_Lachnospiraceae	2.94	0.465	1.89E-5	1.16E-4	
g_Desulfovibrio	-1.52	0.264	4.93E-5	2.61E-4	
g_unclassified_Tyzzerella	3.67	0.657	6.79E-5	3.14E-4	
g_Akkermansia	-8.03	1.46	7.87E-5	3.23E-4	
g_Prevotella	-5.28	1.04	1.69E-4	6.25E-4	



Methods Selection:

Explore Overall Patterns via Dimensionality Reduction

Procrustes analysis (PA) is a simple visualization technique that superimposes the principal components of two datasets at a low-dimensional space. Procrustes essentially computes reduced dimensions for each data set using a method similar to PCA. Then, one of the reduced dimension matrices is rotated until it has maximum similarity with the other. Scores from both matrices are plotted at the same time, with pairs belonging to the same sample connected by a line. Procrustes is asymmetric, therefore the order that the 'omics datasets are uploaded will impact the results. (more details ....)

Data Integration Analysis for Biomarker discovery using Latent cOmponents (DIABLO) is a supervised method for multi-omics biomarker exploration. It is based on a generalized version of PLS (multi block PLS-DA) that seeks to find related multi-dimensional components that maximally separate sample labels. DIABLO is symmetric with respect to the 'omics data, therefore the order that 'omics datasets are uploaded will not impact the results. (more details ...)

Method selection: Procrustes

- Procrustes
- DIABLO

Submit

Obtain Functional Insights via Pathways & Network

Microbiome data and metabolomics data are projected into KEGG metabolic network for visual exploration as well as enrichment analysis. The integration strategies are based on microbiome data types.

- Marker genes data will be used to customize the metabolic network for enrichment analysis of metabolomics data. Users can click a node to view the most correlated microbes of metabolites
- For the metagenomics data, both KOs and metabolomic features will be projected to the selected network for integration analysis.

Compound origin: All uploaded taxa/ko

- All uploaded taxa/ko
- Significant Taxa
- Bacteria
- Bacteria and Human
- Whole metabolic map

Submit

Tune the KEGG metabolic netowork for enrichment analysis. Significant taxa is only applicable for marker gene input.

Discover Metabolite-microbe Correlation

The relationships between the paired microb ntuitively presented using an interactive heatmap.Two types of heatmaps are provided including:

- Correlation heatmap is based on the statistical correlation.
- Prediction heatmap is based on GEM-based predictive models. The predictive models are logistic regression models trained based on high-quality genome-scale metabolic models (GEMs) that can predict the potential of each metabolite production across different taxonomy levels. Prediction heatmap is only suitable for integrating different taxonomy and metabolites.

Heatmap type: Correlation heatmap

- Correlation heatmap
- Prediction heatmap

Submit

Choose the heatmap mode here. Prediction heatmap is **not** suitable for KO and peaks input.

Procruste analysis summary:

Procrustes analysis summary

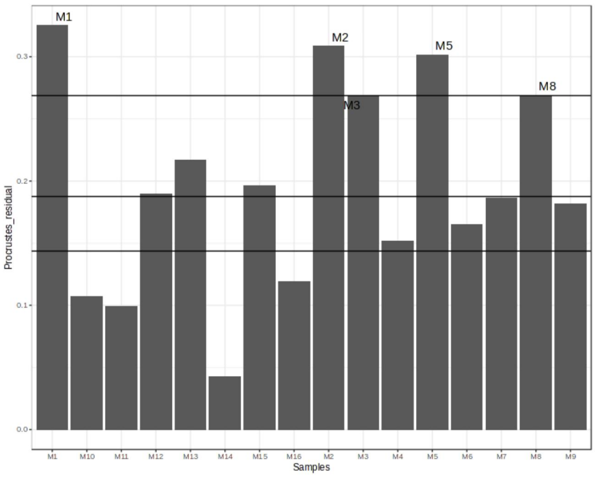
The resulting plot of procrustes analysis along with key statistics is shown below.

Summary statistics of Procrustes analysis are displayed below. Procrustes sum of squares repres[...] The shorter the distance, the better

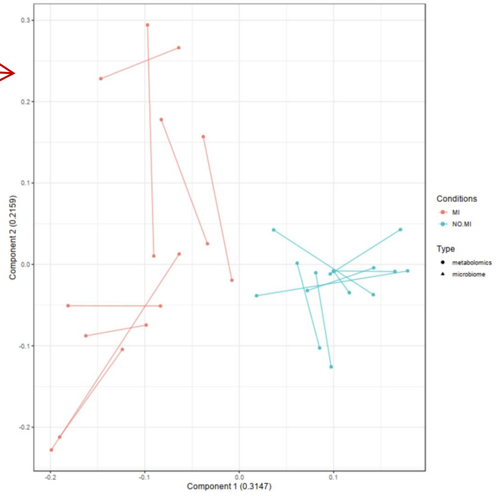
Procrustes Sum of Squares:	0.4793
Correlation in a symmetric Procrustes rotation:	0.7216
Significance:	0.001

Check statistical summary here

Diagnostic | Sample space



View plot summary

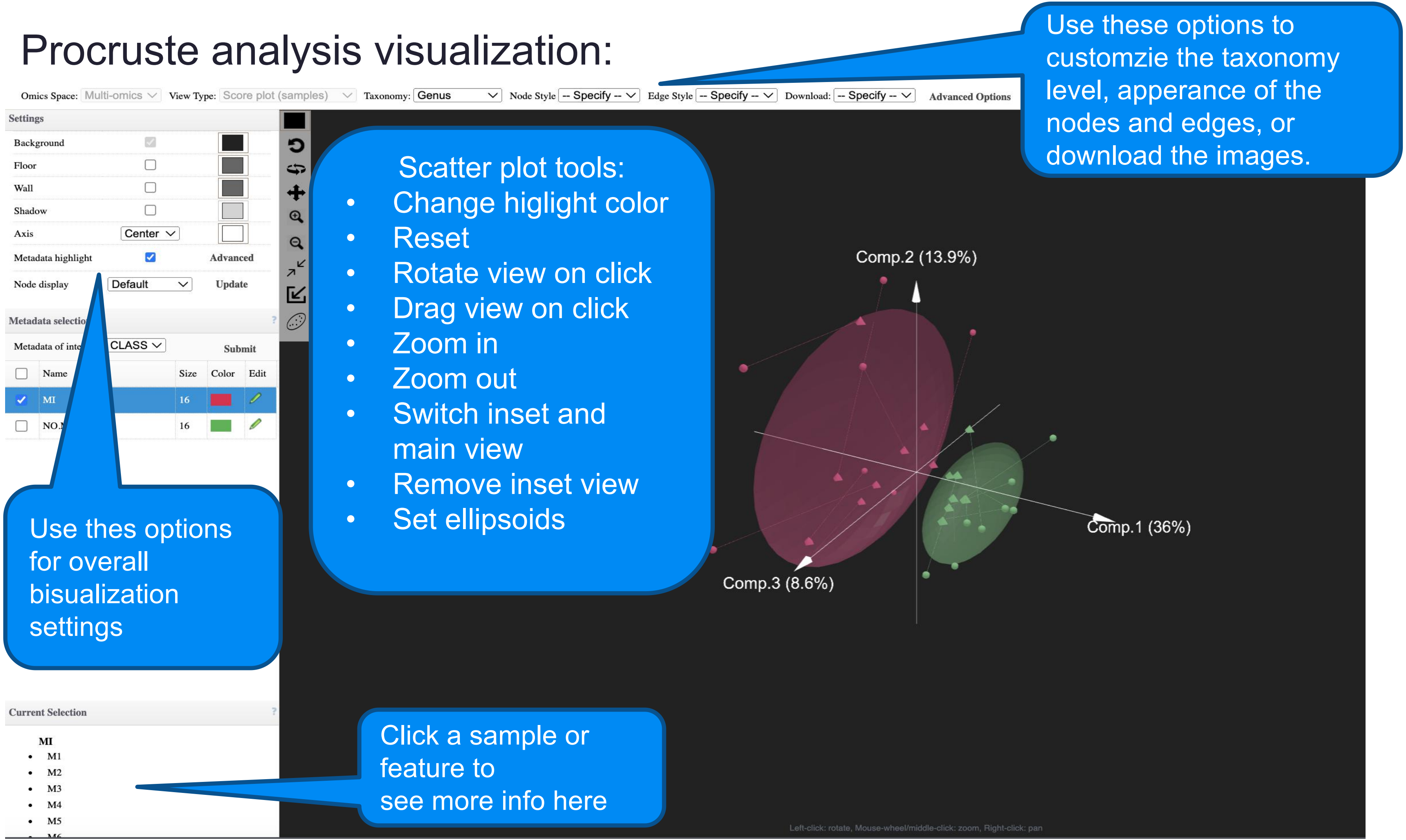


Plot of procrustes residual differences which is based on the sum of squared deviations. The higher the residual difference, the more different is the paired sample between both omics. The three horizontal lines correspond to 75, median and 25 percentile of all residuals. Please refer to **vegan** for more details.

Previous | Proceed

Proceed to visualization

Procruste analysis visualization:



DIABLO analysis summary:

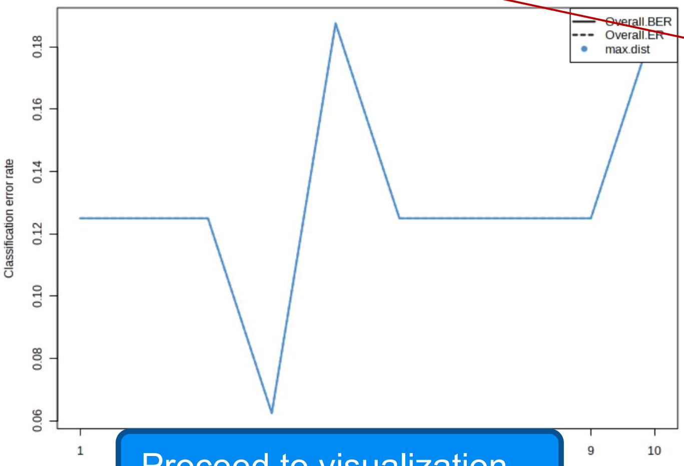
DIABLO analysis summary

The summary plots displayed below can be used to visualize the effectiveness of the dimension reduction before proceeding to the visual analytics page.

- **Diagnostic** tab shows how well the top components explain the variability in the integrated data set.
- **Sample space** tab visualizes overall distribution of samples in 2D space.
- **Loading** tab focuses on the contributions of features to the top components.

Diagnostic | Sample space | Loading

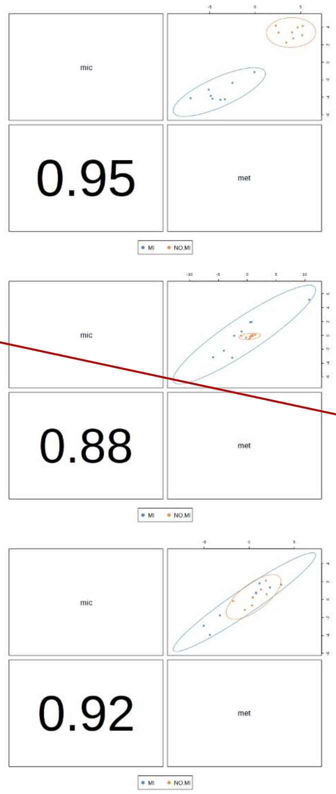
View plot summary

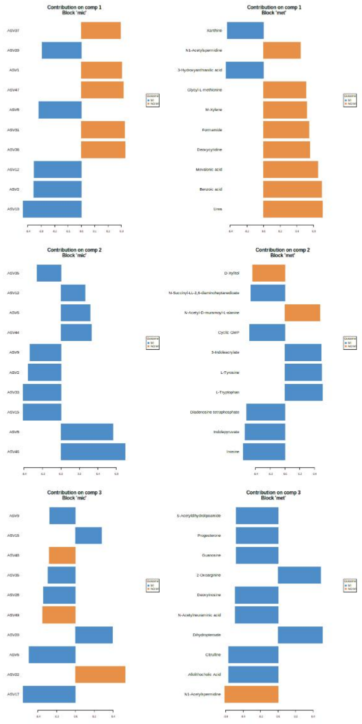


Proceed to visualization

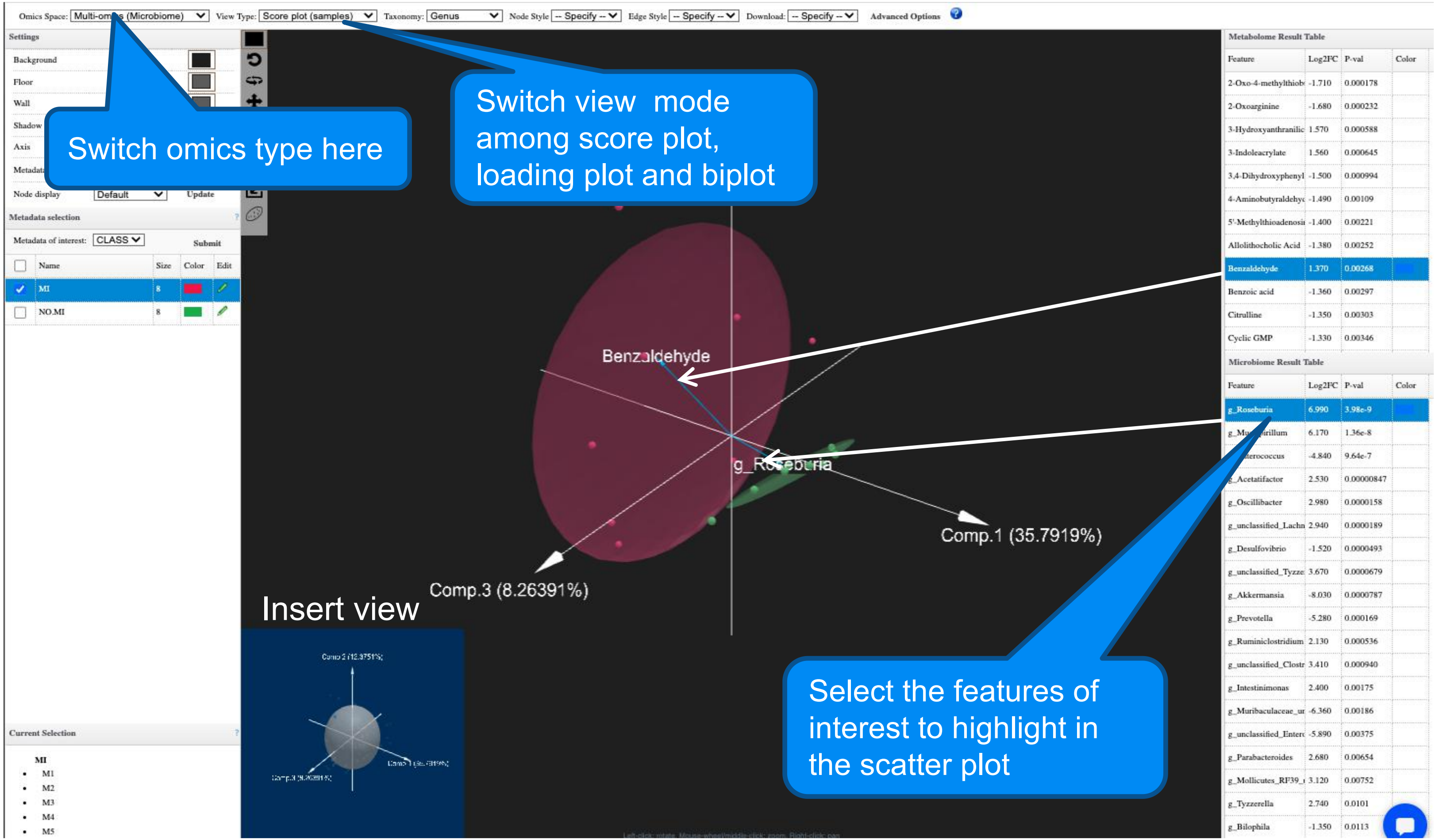
<< Previous

>> Proceed

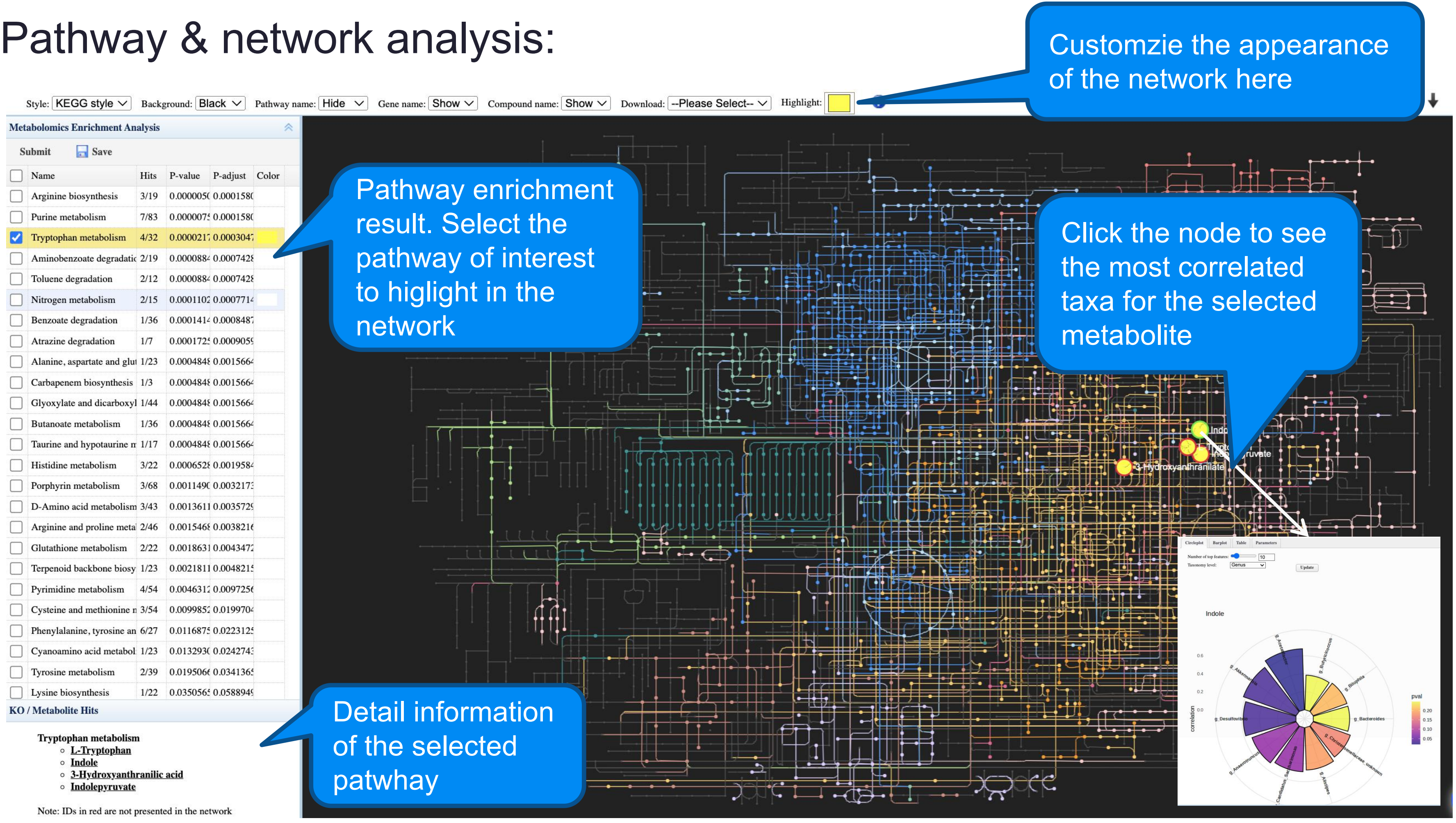




DIABLO analysis visualization:

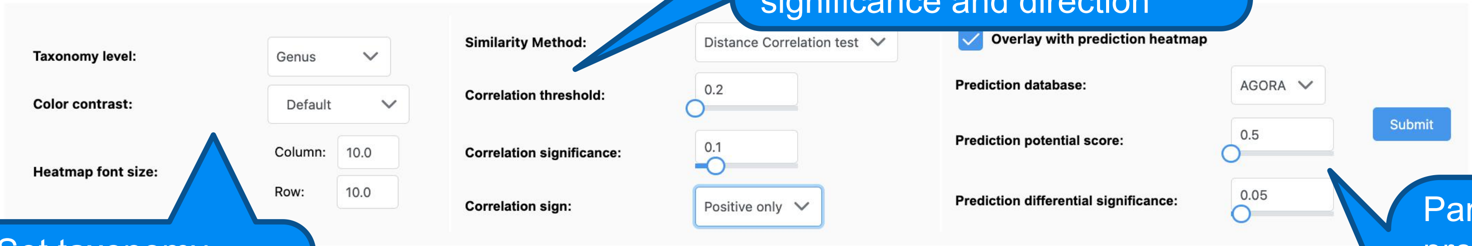


Pathway & network analysis:



Correlation analysis:

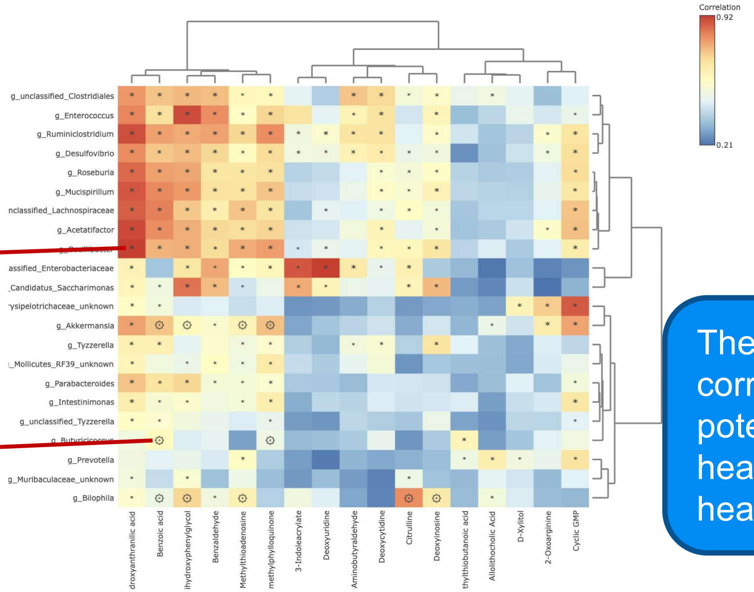
Heatmap Visualization & Correlation:



Parameters for correlation heatmap including correlation method, threshhold, significance and direction

Set taxonomy level, color and font size for heatmap

Parameters for prediction heatmap including GEM database used for prediction model, potential score and significance

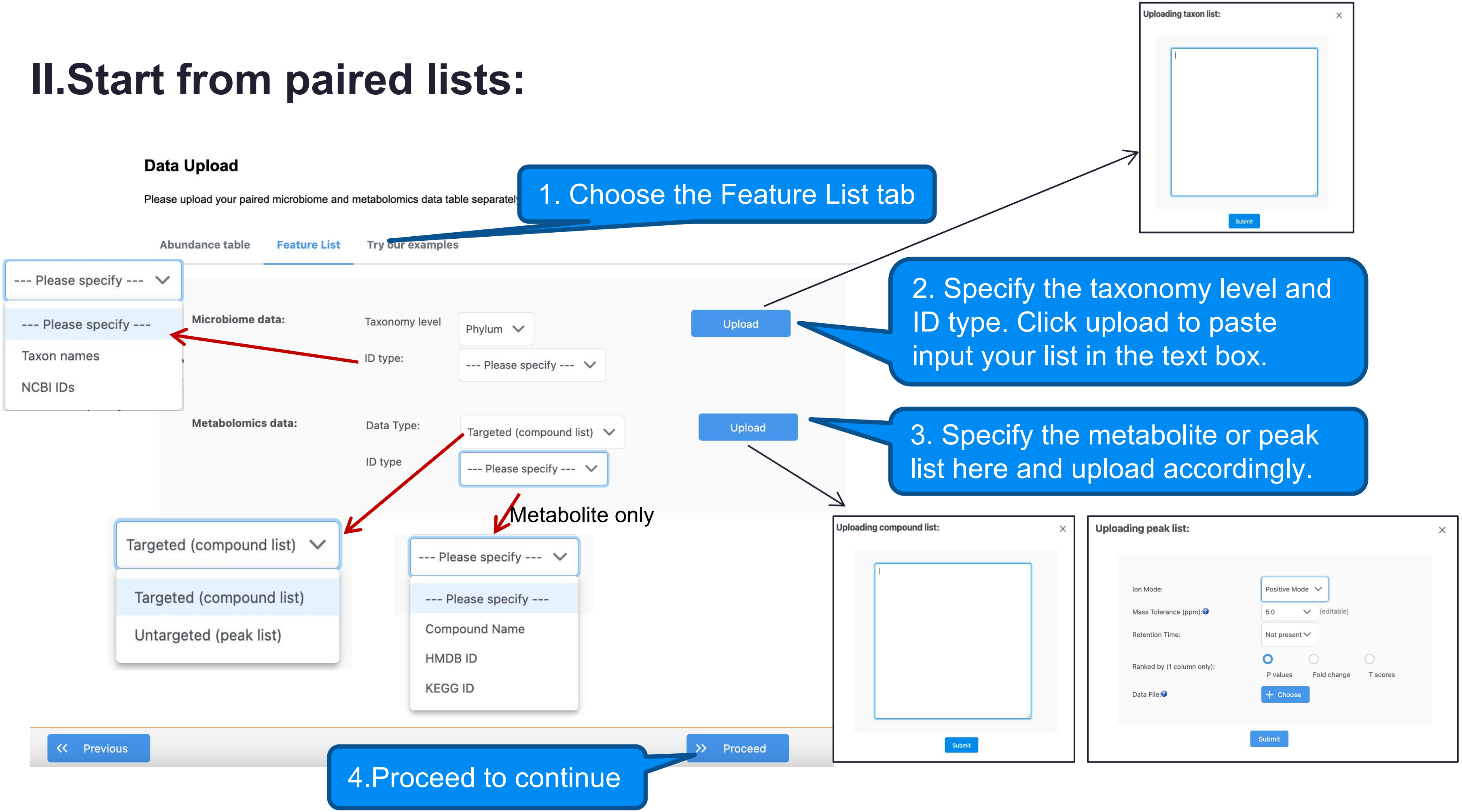


Asterisk means the given correlation passes the setting significance

Diamond indicate the given correlation is identified as significant by both statistical correlation result and prediction result

The color indicates the level of correlation and average potential score in the correlation heatmap and prediction heatmap, respectively.

II.Start from paired lists:



Name Mapping:

Input feature Mapping

All the input microbiome and metabolomics features need to . GEMs are genome-scale metabolic models which can be used for metabolic potential prediction. Two databases, Agora and EMBL GE ound. The tables below show the matched taxon and compounds. To **remove** a taxon or compound from further analysis, use the **De**

Check the name mapping results for each omics type by selecting the corresponding tab.

Taxon Name Mapping | Compound Name Mapping

For microbial taxonomy, KEGG, GEM (Agora and EMBL) and NCBI ID are provided.

For metabolites, KEGG and GEMs ID are provided.

Query	KEGG Hits	Agora Hits	EMBL Hits	NCBI Taxonomy	
Haemophilus_D	Haemophilus_D	Haemophilus_D	Haemophilus_D	724	Del
Bacteroides	Bacteroides	Bacteroides	Bacteroides	816	Del
Gemella	Gemella	Gemella	Gemella	1378	Del
Faecalibacterium	Faecalibacterium	Faecalibacterium	Faecalibacterium	216851	Del
Barnesiella	Barnesiella	Barnesiella	Barnesiella	397864	Delet
Pseudoruminococcus					Delete
Slackia_A	Slackia_A	Slackia_A	Slackia_A	84108	Delete
Granulicatella	Granulicatella	Granulicatella	Granulicatella	117563	Delete
Fusobacterium_C	Fusobacterium_C	Fusobacterium_C	Fusobacterium_C	848	Delete
Streptococcus	Streptococcus	Streptococcus	Streptococcus	1301	Delete
Acidaminococcus	Acidaminococcus	Acidaminococcus	Acidaminococcus	904	Delete

Unwanted features can be deleted here.

Methods Selection:

Please choose an integration method to proceed

Explore Overall Patterns via Dimensionality Reduction

Procrustes analysis (PA) is a simple visualization technique that superimposes the principal components of two datasets at a low-dimensional space. Procrustes essentially computes reduced dimensions for each data set using a method similar to PCA. Then, one of the reduced dimension matrices is rotated until it has maximum similarity with the other. Scores from both matrices are plotted at the same time, with pairs belonging to the same sample connected by a line. Procrustes is asymmetric, therefore the order that the 'omics datasets are uploaded will impact the results. (more details ....)

Data Integration Analysis for Biomarker discovery using Latent cOmponents (DIABLO) is a supervised method for multi-omics biomarker exploration. It is based on a generalized version of PLS (multi block PLS-DA) that seeks to find related multi-dimensional components that maximally separate sample labels. DIABLO is symmetric with respect to the 'omics data, therefore the order that 'omics datasets are uploaded will not impact the results. (more details ...)

Method selection: Procrustes Submit

Not appliable for list input

Obtain Functional Insights via Pathways & Network

Microbiome data and metabolomics data are projected into KEGG metabolic network for visual exploration as well as enrichment analysis. The integration strategies are based on microbiome data types.

- Marker genes data will be used to customize the metabolic network for enrichment analysis of metabolomics data. Users can click a node to view the most correlated microbes of metabolites
- For the metagenomics data, both KOs and metabolomic features will be projected to the selected network for integration analysis.

Compound origin: All uploaded taxa/ko Submit

Metabolic network and Heatmap visualization is same as the table input

Discover Metabolite-microbe Correlations via Statistics & GEMs

The relationships between the paired microbiome and metabolomics are intuitively presented using an interactive heatmap.Two

- Correlation heatmap is based on the statistical correlation.
- Prediction heatmap is based on GEM-based predictive models. The predictive models are logistic regression models tr of each metabolite production across different taxonomy levels. Prediction heatmap is only suitable for integratin

Heatmap type: Prediction heatmap Submit

Only prediction heatmap is appliable for list input. It shows the potential of each taxon to produce each metabolite

The End

For more information, visit Tutorials, Resources and Contact pages on www.microbiomeanalyst.ca
Also visit our forum for FAQs on www.omicsforum.ca