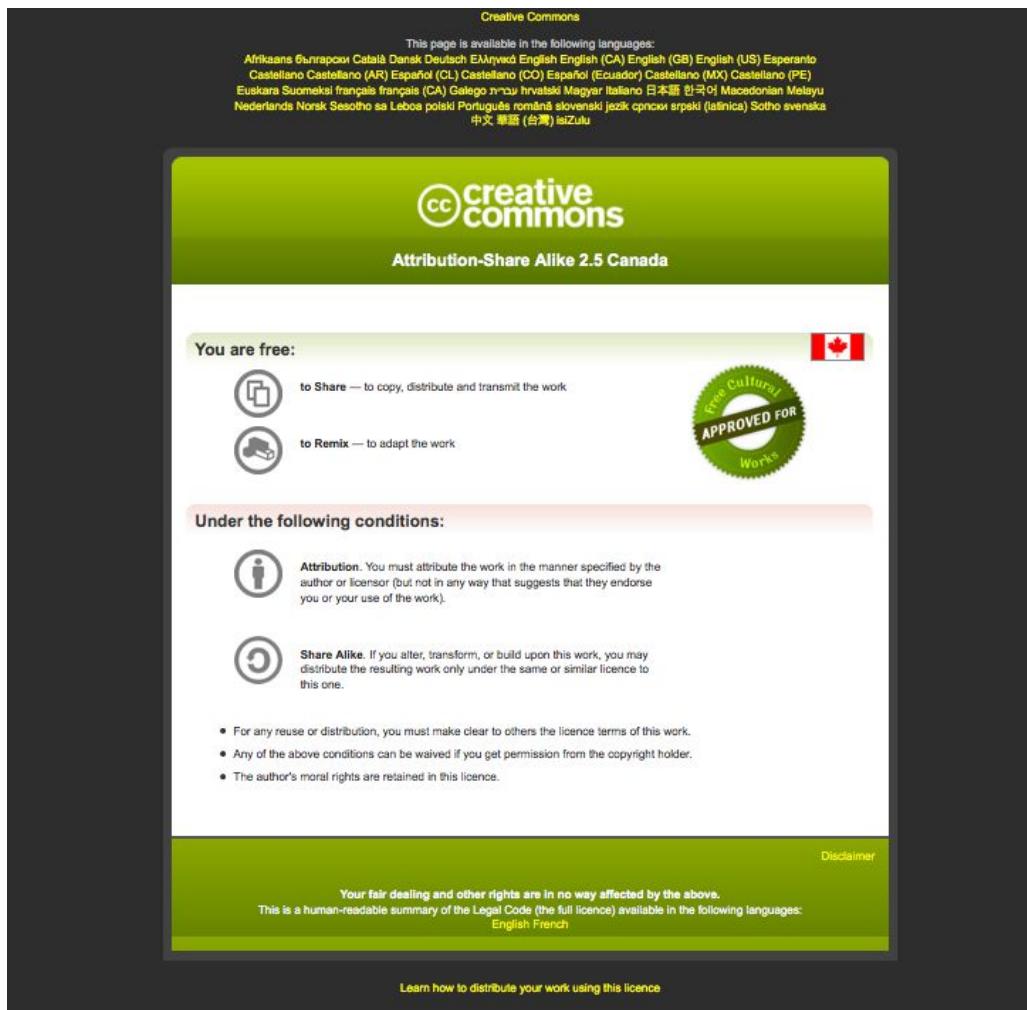




# Canadian Bioinformatics Workshops

[www.bioinformatics.ca](http://www.bioinformatics.ca)

[bioinformaticsdotca.github.io](https://bioinformaticsdotca.github.io)



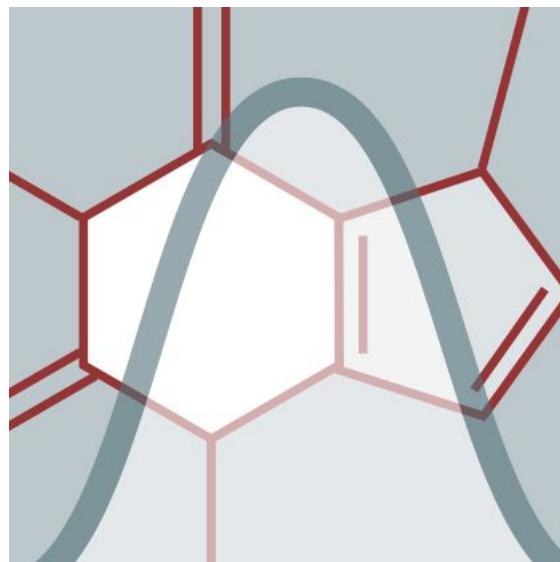
# Metabolomics Data Analysis (Lab)



Jeff Xia

Informatics and Statistics for Metabolomics

July 6-7, 2023



McGill



# Schedule For July 7, 2023

Time	Module
9:00 (MST)/11:00 (EST)	Module 5: Backgrounder in Omics Data Science (Jeff Xia)
10:30 (MST)/12:30 (EST)	Break/Lunch (45 min)
11:15 (MST)/13:15 (EST)	Module 6: Data Analytics for Untargeted Metabolomics (Jeff Xia)
12:15 (MST)/14:15 (EST)	Lunch/Break (45 min)
13:00 (MST)/15:00 (EST)	Module 7 (Lab): Metabolomics Data Analysis using MetaboAnalyst 5.0 (Jeff Xia)
15:00 (MST)/17:00 (EST)	Break (30 min)
15:30 (MST)/17:30 (EST)	Module 8: Integrating Metabolomics with other Omics (Jeff Xia)
17:00 (MST)/19:00 (EST)	Finish

# Learning Objectives

1. To practice raw LC-MS metabolomics data processing;
2. To practice functional analysis on LC-MS global metabolomics data;
3. To practice how to use MetaboAnalyst 5.0 to facilitate one-factor statistical analysis;
4. To practice metadata based statistical analysis on global metabolomics data sets;

# **Raw LC-MS Spectral Data Processing Demos**

# How can we measure these things with LC-MS?

Liquid Chromatography (LC)

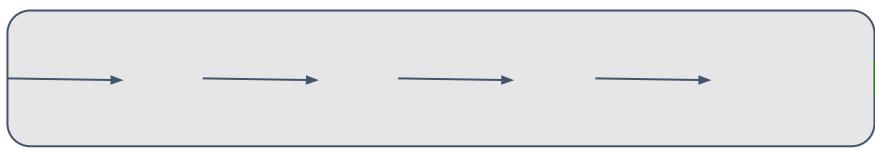
+

Mass Spectrometry (MS)

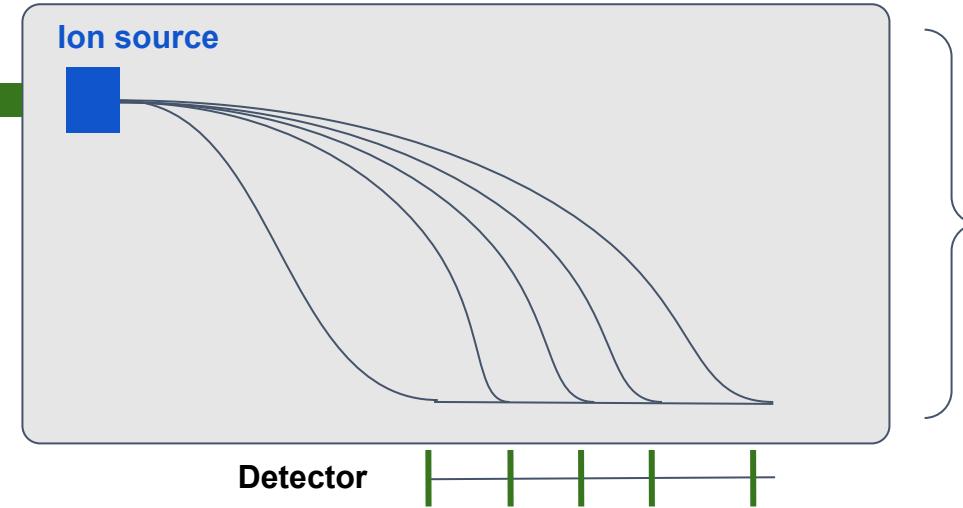
=

LC-MS

The column is filled with 'stuff'



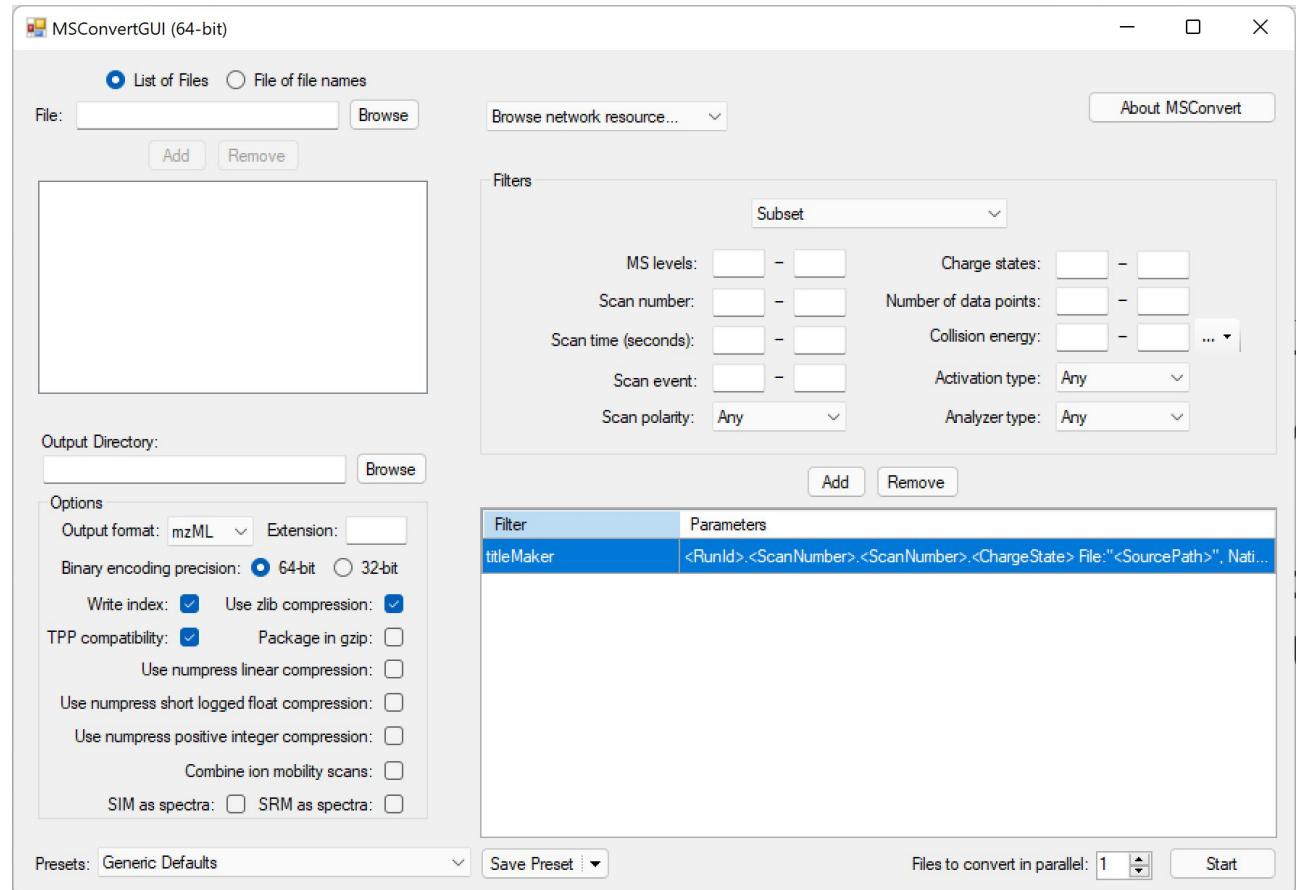
The **speed** that a molecule moves through the 'column' depends on its **chemical characteristics (polarity, etc)** and **mobile phase**



Therefore, the **location** that the ionized molecule collides with the detector is associated with the **mass** of the molecule

# Use Centroid and open-source data format

- We need to convert the MS data into centroid mode to condense the Gaussian Profile peaks into centroids.
- Open-source formats (.mzML/ etc.)..



# MetaboAnalyst: LC-MS Spectral Processing Module

<https://www.metaboanalyst.ca/MetaboAnalyst/ModuleView.xhtml>

## Module Overview

Input Data Type	Available Modules (click on a module to proceed, or scroll down for more details)					
Raw Spectra (mzML, mzXML or mzData)	<a href="#">LC-MS Spectra Processing</a>					
MS Peaks (peak list or intensity table)			Functional Analysis	Functional Meta-analysis		
Annotated Features (compound list or table)		Enrichment Analysis	Pathway Analysis	Joint-Pathway Analysis	Network Analysis	
Generic Format (.csv or .txt table files)	Statistical Analysis [one factor]	Statistical Analysis [metadata table]	Biomarker Analysis	Statistical Meta-analysis	Power Analysis	Other Utilities

Please use [OmicsForum](#) for support & troubleshooting request

# User registration and log-in (Optional)

MetaboAnalyst 5.0 - user-friendly, streamlined metabolomics data analysis

Module Overview

Input Data Type	Available Modules (click on a module to proceed, or scroll down for more details)					
Raw Spectra (mzML, mzXML or mzData)	LC-MS Spectra Processing					
MS Peaks (peak list or intensity table)			Functional Analysis	Functional Meta-analysis		
Annotated Features (compound list or table)		Enrichment Analysis	Pathway Analysis	Joint-Pathway Analysis	Network Analysis	
Generic Format (.csv or .txt table files)	Statistical Analysis [one factor]	Statistical Analysis [metadata table]	Biomarker Analysis	Statistical Meta-analysis	Power Analysis	Other Utilities

Click here to start

LC-MS Spectra Upload

MetaboAnalyst currently supports mzML, mzXML, CDF or mzData formats in centroid mode. Quality control (QC) spectra are not required but recommended. QC should start with "QC\_" or marked as "QC" in meta data. BLANK should be marked as "BLANK" in meta data for subtraction. The following two data types are allowed:

Spectrum (max: 200 spectra).  
Columns - spectral names and group labels (example)

Click login to register/login

Please Select all files, then click Upload to start. Once the upload has completed, click Proceed to continue.

+ Select

Reset Proceed

**NOTE:** Register or Login is optional. You can upload your files directly, but the jobs for registered users will be kept for 180 days.

MetaboAnalyst 5.0 - user-friendly, streamlined metabolomics data analysis

Use the panel below to add an account.

Start by clicking Add New to add a new project. In order to purge a project, select it and click Delete.

Once you start a project, you will be prompted to log in.

Projects Available

Project ID	Title	Description	Type	Date created	Action
1	bug_removal	test	raw	2022/02/03	<button>Load</button> <button>Delete</button>
2	test demo	this is test project	raw	2022/06/12	<button>Start</button> <button>Delete</button>

Click buttons to operate your projects

+ Add New

Click to create new projects

Module 7

MetaboAnalyst 5.0 - user-friendly, streamlined metabolomics data analysis

Log in to start a new analysis or resume your previous analysis

Email \*  Password \*  Login

Forgot password?

Create account

Login here

Register here

# Data Uploading

## LC-MS Spectra Upload

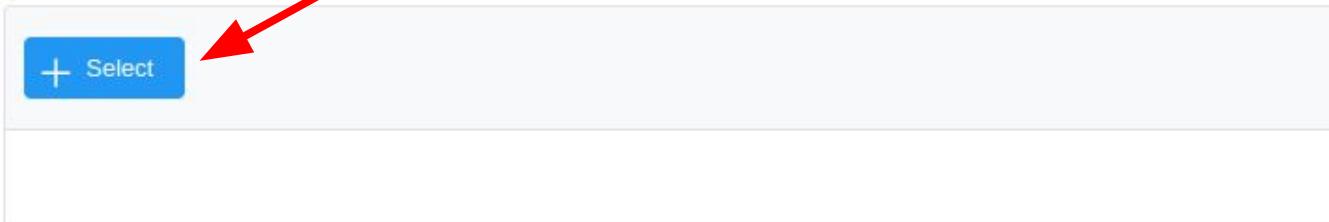
MetaboAnalyst currently supports [mzML](#), [mzXML](#), [CDF](#) or [mzData](#) formats in centroid mode. Quality control (QC) spectra are not required but recommended. QC should start with "QC\_" or marked as "QC" in meta data. BLANK should be marked as "BLANK" in meta data for subtraction. The following two data types are allowed:

1. [Required] Spectra uploaded as individual zip files - one zip (.zip) per spectrum [max: 200 spectra].
2. [Optional] Meta data uploaded as a plain text (.txt) file containing two columns - spectral names and group labels [\[example\]](#)

Spectra processing can take a long time to complete, to avoid waiting:

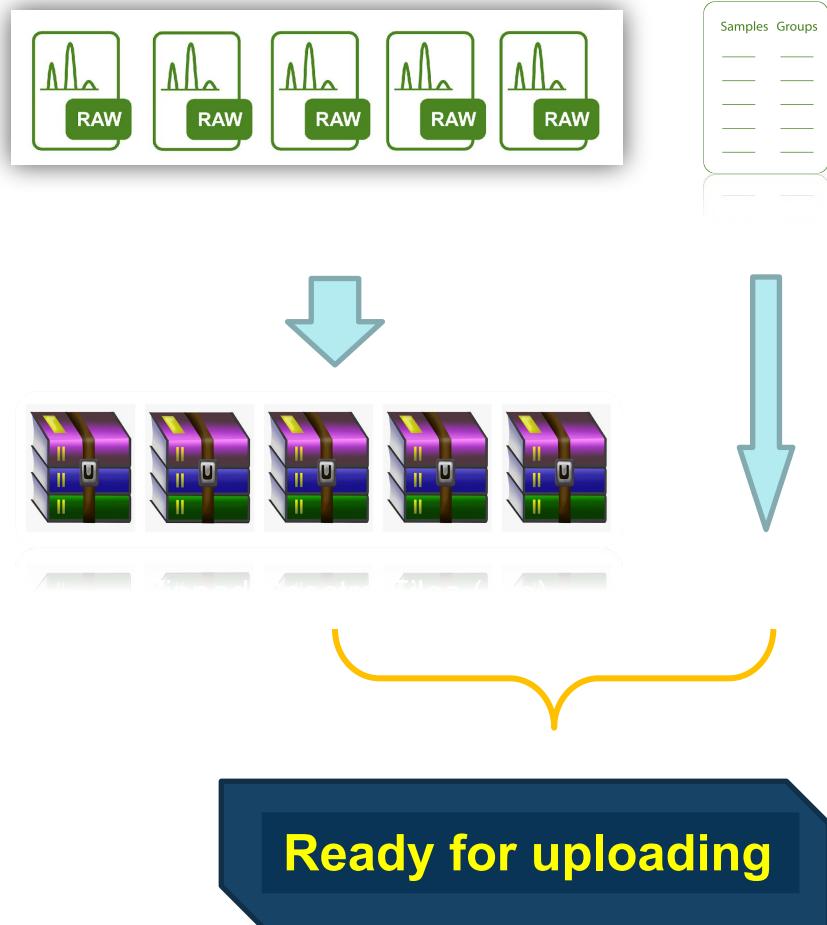
1. For guest users (default), after job submission, click **Create Bookmark URL** and save the URL so you can return later to check your job status.
2. For registered users, use the buttons on the left panel to manage your projects.

Please **Select** all files, then click **Upload** to start. Once the upload has completed, click **Proceed** to continue.



The screenshot shows a file upload interface. At the top left is a blue button labeled '+ Select'. Below it is a large, empty rectangular input field for file selection. At the bottom of the interface are two blue buttons: 'Reset' on the left and 'Proceed' on the right.

# Raw data organization



trimmed_metadata.txt	219 bytes	30 Jul 2020	☆
CD-9WOBP.zip	1.9 MB	15 Jun 2020	☆
CD-6KUCT.zip	2.0 MB	15 Jun 2020	☆
CD-77FXR.zip	2.1 MB	15 Jun 2020	☆
HC-AMR37.zip	2.1 MB	15 Jun 2020	☆
HC-9X47O.zip	2.1 MB	15 Jun 2020	☆
QC_PREFB02.zip	2.2 MB	15 Jun 2020	☆
QC_PREFA02.zip	2.2 MB	15 Jun 2020	☆
CD-9OS5Y.zip	2.2 MB	15 Jun 2020	☆
HC-AUP8B.zip	2.3 MB	15 Jun 2020	☆
HC-9SNJ4.zip	2.4 MB	15 Jun 2020	☆

# Submit to upload your data

Home

Upload

Spectra check

Spectra processing

Job status

Spectra result

Download

Exit

Log in

## LC-MS Spectra Upload

MetaboAnalyst currently supports [mzML](#), [mzXML](#), [CDF](#) or [mzData](#) formats in centroid mode. Quality control (QC) spectra are not required but recommended. QC should start with "QC\_" or marked as "QC" in meta data. BLANK should be marked as "BLANK" in meta data for subtraction.

The following two data types are allowed:

1. [Required] Spectra uploaded as individual zip files - one zip (.zip) per spectrum [max: 200 spectra].
2. [Optional] Meta data uploaded as a plain text (.txt) file containing two columns - spectral names and group labels [\[example\]](#)

Spectra processing can take a long time to complete, to avoid waiting:

1. For guest users (default), after job submission, click **Create Bookmark URL** and save the URL so you can return later to check your job status.
2. For registered users, use the buttons on the left panel to manage your projects.

Please **Select** all files, then click **Upload** to start. Once the upload has completed, click **Proceed** to continue.

+ Select

Reset

Proceed

Try our example data

# Data Integrity Check

The screenshot shows the 'Data Integrity Check' interface. On the left, a sidebar menu includes 'Upload', 'Spectra check' (which is highlighted in blue), 'Spectra processing', 'Job status', 'Spectra result', 'Download', and 'Exit'. A large orange box contains the text: 'Results of the Data Integrity Check are shown here.' An arrow points from this box to the main content area. The main area has a header 'Data Integrity Check:' with a 'Show R Commands' link. Below the header is a list of instructions:

1. Spectral Format - only mzML, mzXML, mzData and netCDF formats are currently supported;
2. MS Mode - only spectra in **centroid mode** are supported in the online platform. Click **Convert** to centroid your profile data online. **This conversion process will take some time, please be patient...**
3. If a meta data file is provided:
  - o The first column (spectral names) must match the sample names in the meta-data file;
  - o The second column (group labels) must contain at least two groups (not including QC), each containing  $\geq 3$  replicates.

The main content is a table with columns: Spectra, Centroid, Size (MB), Group, Convert, and Include. The table lists various mzML files and their characteristics. Some files are grouped under 'QC'. The 'Convert' and 'Include' columns contain blue wrench icons with checked boxes. The 'Next' button at the bottom is also highlighted with an orange box. A note in the bottom right corner says: 'If your data is not in centroid mode, click Convert wrench button to convert it online.'

Spectra	Centroid	Size (MB)	Group	Convert	Include
Semi_025.mzML	True	15.7	Semi_immue		<input checked="" type="checkbox"/>
Semi_091.mzML	True	15.3	Semi_immue		<input checked="" type="checkbox"/>
Semi_157.mzML	True	16.0	Semi_immue		<input checked="" type="checkbox"/>
Semi_061.mzML	True	15.6	Semi_immue		<input checked="" type="checkbox"/>
Semi_143.mzML	True	15.7	Semi_immue		<input checked="" type="checkbox"/>
Semi_045.mzML	True	15.6	Semi_immue		<input checked="" type="checkbox"/>
QC_005.mzML	True	15.8	QC		<input checked="" type="checkbox"/>
QC_001.mzML	True	16.1	QC		<input checked="" type="checkbox"/>
QC_003.mzML	True	15.9	QC		<input checked="" type="checkbox"/>
Naive_109.mzML			Naive		<input checked="" type="checkbox"/>
Naive_127.mzML			Naive		<input checked="" type="checkbox"/>
Naive_139.mzML			Naive		<input checked="" type="checkbox"/>
Naive_007.mzML			Naive		<input checked="" type="checkbox"/>
Naive_027.mzML			Naive		<input checked="" type="checkbox"/>
Naive_071.mzML			Naive		<input checked="" type="checkbox"/>

Click **Next** to move on to the Parameters Selection page (At least 3 samples included for next)

Xia Lab @ McGill (last updated 2022-06-14)

# Parameter Selection

**TIP1:** Default Parameters setting option is 'customized'. If you are not a parameter expert, please try to use the automated optimization pipeline.

The screenshot shows the 'LC-MS Spectra Processing' interface. On the left, a sidebar menu includes 'Upload', 'Spectra check' (which is highlighted in blue), 'Spectra processing', 'Job status', 'Spectra result', 'Download', and 'Exit'. At the top right, there is a link 'Show R Commands'. The main content area is titled 'LC-MS Spectra Processing' and contains a brief description about parameter optimization. Below this, there are two options: 'Default/manual' (selected) and 'Auto-optimized'. A note states: 'Default/manual option will use the parameters in the current display. You can manually overwrite these settings; Auto-optimized will automatically select the best parameter combination (for the centWave only)'.

The 'Parameter Setting' section is divided into five steps: 1. Peak Picking, 2. Peak Alignment, 3. Peak Annotation, 4. Contaminant Removal, and 5. Blank Subtraction. Step 1 is currently active. In the 'Peak Picking' step, the 'Method' is set to 'centWave'. The parameters are: min\_peakwidth: 5.0, max\_peakwidth: 30.0, ppm: 5.0, and mzdiff: 0.01. An orange arrow points from the text '1. Adjust the following parameters according to the LC-MS instrument/extraction methods used.' to the 'ppm' input field. Step 2 (Peak Alignment) shows Method: loess, Bandwidth: 10.0, minFraction: 0.8. Step 3 (Peak Annotation) shows Polarity: positive, Adducts: View, More options: View. Step 4 (Contaminant Removal) has a checked checkbox. Step 5 (Blank Subtraction) has an unchecked checkbox. At the bottom, a large blue button labeled 'Submit Job' is highlighted with an orange arrow pointing to it from the text '2. Click Submit Job to perform the spectra processing.'

**1.** Adjust the following parameters according to the LC-MS instrument/extraction methods used.

**2.** Click **Submit Job** to perform the spectra processing.

# Job Status View

The screenshot shows the 'Job Status View' interface. On the left is a sidebar with buttons: Upload, Spectra check, Spectra processing, Job status (which is selected), Spectra result, Download, and Exit. A callout box points to the 'Job status' button with the text: 'The status of the job will update here in real-time.'

The main area has a title 'Job Status View'. It says: 'Depending on the current server load and the size of your data, it can take a few hours up to several days to complete your job.' Below this are two bullet points:

- If you have not logged in, please click **Create Job URL** and save the job link. You can then close the current page and come back later using this link.
- At any time during data analysis, keep only **one active web page open** (except FAQs or other static web pages) as multiple tabs/windows will interfere with each other leading to unpredictable results.

A 'Job Status' box contains the following information:

Job ID:	10702
Bookmark Link:	<a href="#">Create Job URL</a>
Current Status:	Running
Priority:	Level 1
Parameters:	Save
Job Progress:	<div style="width: 10%;">10%</div>

The 'Text Output' section shows a list of log messages:

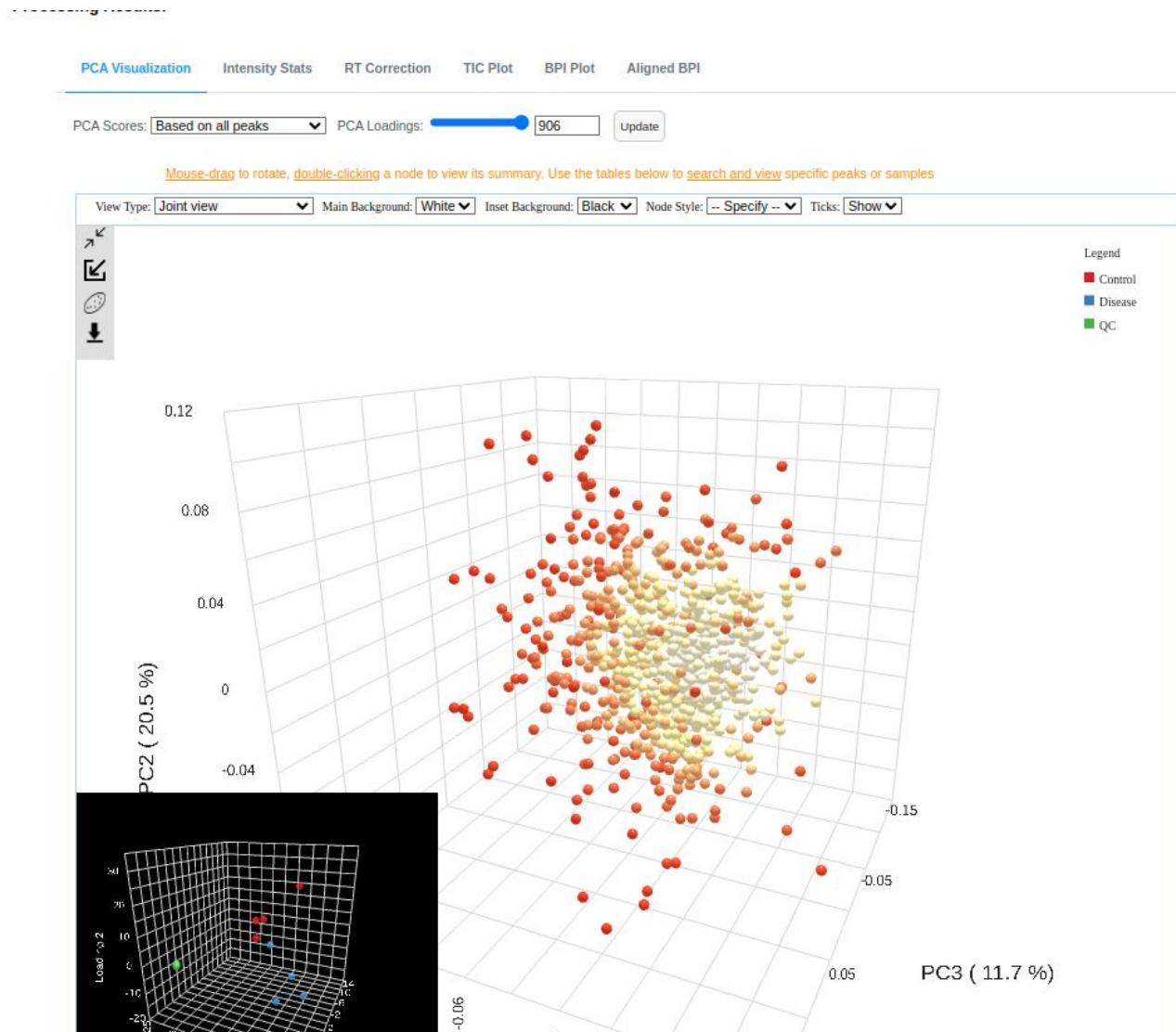
- This step will take a short time...
- Raw file import begin...
- Naive\_007.mzML import done!
- Naive\_027.mzML import done!
- Naive\_071.mzML import done!
- Naive\_109.mzML import done!
- Naive\_127.mzML import done!
- Naive\_139.mzML import done!
- QC\_001.mzML import done!
- QC\_003.mzML import done!
- QC\_005.mzML import done!
- Semi\_025.mzML import done!

The 'Output File' section shows 'Status Text' and the date '2022-07-11 11:27:06'.

At the bottom are three buttons: Refresh Status, Cancel Job, and Proceed. A callout box points to the 'Proceed' button with the text: 'Once the job is complete (Job Progress 100%), click **Proceed** to view the results.'

# Results Visualization

You could interactively view the data results with 3D PCA  
and check the main loadings for the distribution  
(Double click the nodes to view TIC/EIC)



# Result Summary

[Result Summary](#)

[Spectra / Sample Table](#)

[Feature / Peak Table](#)

## Raw Spectra Processing Result Summary:

MetaboAnalyst has finished raw spectra processing with OptiLCMS (1.0.5):

There are 10 samples of 3 groups (Control, Disease, QC) included for processing!

Total of 906 features have been detected and aligned across the whole sample list.

The mass deviation of this study was estimated/set as 5 ppm.

55 features (6.06%) have been annotated as isotopes.

30 features (3.31%) have been annotated as adducts.

346 unique formulas have been matched to HMDB database.

791 potential compounds have been matched to HMDB database.

# Exploring the Results -2

Result Summary    **Spectra / Sample Table**    Feature / Peak Table

Spectra ↑	Group ↑	Peaks No. ↑↓	Missing (%) ↑↓	RT Range	m/z Range	View
Naive_007	Naive	3510	13.87	9.15~292.71	85.065~1273.52	
Naive_027	Naive	3517	13.69	9.15~292.71	85.065~1273.52	
Naive_071	Naive	3433	15.75	9.15~292.71	85.065~1273.52	
Naive_109	Naive	3167	22.28	9.15~292.71	85.065~1273.52	
Naive_127	Naive	3438	15.63	9.15~292.71	85.065~1273.52	
Naive_139	Naive	3450	15.34	9.15~292.71	85.065~1273.52	
QC_001	QC	3349	17.82	9.15~292.71	85.084~1264.131	
QC_003	QC	3385	16.93	9.15~292.71	85.065~1264.131	
QC_005	QC	3397	16.64	9.15~292.71	85.065~1266.511	
Semi_025	Semi_immue	3588	11.95	9.15~292.71	85.065~1264.131	
Semi_045	Semi_immue	3667	10.01	9.15~292.71	85.065~1273.52	
Semi_061	Semi_immue	3631	10.9	9.15~292.71	85.065~1264.131	
Semi_091	Semi_immue	3573	12.32	9.15~292.71	85.065~1264.131	
Semi_143	Semi_immue	3596	11.75	9.15~292.71	85.065~1264.131	
Semi_157	Semi_immue	3620	11.17	9.15~292.71	85.065~1264.131	

Click the **View** button to see TIC of the corresponding spectra.

Total Ion Chromatogram

CD-77FXR

Intensity

Retention Time

The labels marked in the TIC is the corresponding m/z value of the base ion in the peak.

<< < 1 > >> 20 ▾

> Download Page

# Exploring the Results - 3

Result Summary   Spectra / Sample Table   Feature / Peak Table

For isotopes/adducts annotation, the matching is based on the m/z value of its corresponding parent ion. Otherwise, it is considered as in the format of the primary ion.  
All compounds/formulas are matched to [HMDB](#) (v5) based on the mass error (ppm value) for raw spectra processing.  
Intensity is average of all samples. Coefficient of variation (CV) is also the summarized based on all samples.  
When group information is provided, p values will be calculated with t-test/ANOVA based on log transformed data.

m/z ↑↓	RT/s ↑↓	Intensity ↑↓	CV (%) ↑↓	P values ↑↓	FDR	Annotations	Putative IDs	View
1190.7142	113.32	192755.1	54.28	2.3205802E-15	0.0			
768.4118	71.34	298137.8	43.58	2.5673008E-15	0.0	[M+Na+NaCOOH]+ 677.43 [M+H-CH2]+ 781.419		
759.723	60.93	124919.2	52.69	3.624273E-15	0.0			
438.6319	72.01	4262468.0	48.17	5.093338E-15	0.0	[2M+Na]+ 207.821		
1008.5172	116.55	378720.3	48.79	6.2724633E-15	0.0			
913.7906	108.08	128345.3	52.76	2.3088964E-14	0.0			
1200.3439	71.5	92447.1	51.53	2.6098027E-14	0.0			

This table is showing all MS features. Click the button of Putative IDs show the potential Chemical IDs of the features towards HMDB.

Putative IDs

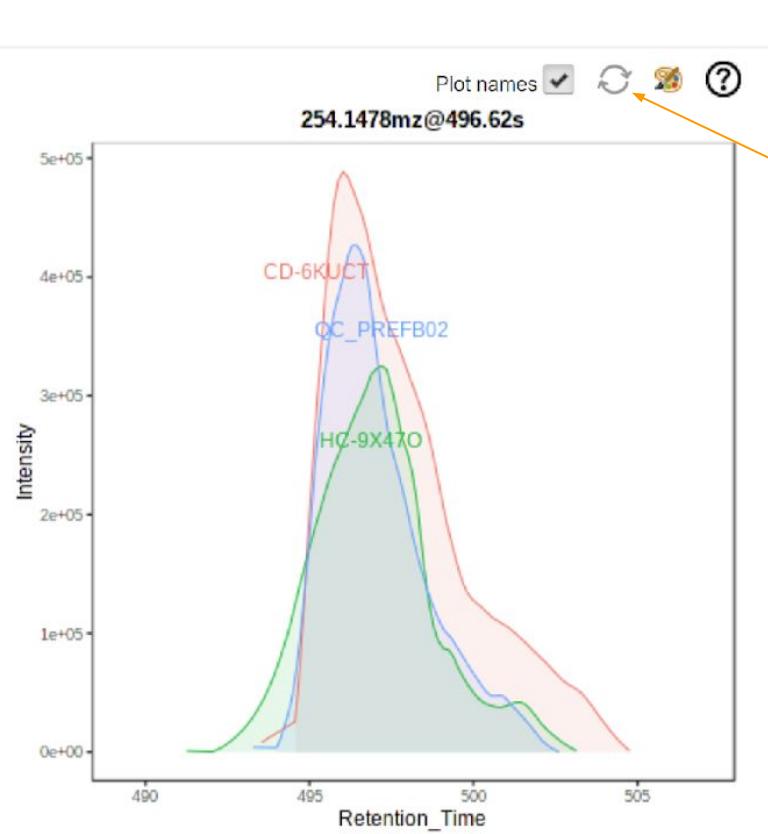
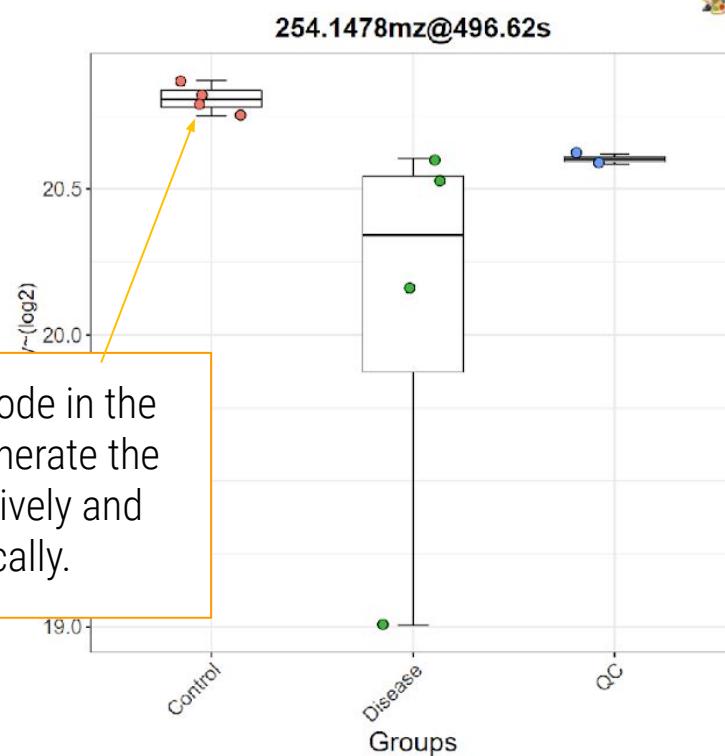
X

Formulas	Compounds
<a href="#">C34H64NO10P</a>	<a href="#">PS(14:0/14:1(9Z))</a> ; <a href="#">PS(14:1(9Z)/14:0)</a>

2. Click the button under **View** to see a dynamic Extracted Ion Chromatogram for the selected feature (see next page).

# Exploring the Results -4

## Boxplots and EICs



**TIP1:** If the plotting failed, please clean the cache of your browser or use another browser.

2. Click this 'reset' icon to restart the generation of EIC plot.

3. Scroll down to the bottom of page and click "**Proceed**" to view the Downloads page.

Mouse over a data point on a boxplot to view its sample name. Double click to show its EIC. Clicking different data points will stack their EICs. Click the Reset icon to restart.

# **Functional Analysis of Untargeted Metabolomics Demos**

# Start Functional Analysis



**MetaboAnalyst 5.0** - user-friendly, streamlined metabolomics data analysis

[Home](#)

[Data Formats](#)

[Tutorials](#)

[OmicsForum](#)

[APIs](#)

[Update History](#)

[MetaboAnalystR](#)

[Contact](#)

[User Stats](#)

[Publications](#)

[COVID-19 Data](#)

[About](#)

## Module Overview

Input Data Type	Available Modules (click on a module to start)					
Raw Spectra (mzML, mzXML or mzData)				LC-MS Spectra Processing		
MS Peaks (peak list or intensity table)			Functional Analysis	Functional Meta-analysis		
Annotated Features (compound list or table)		Enrichment Analysis	Pathway Analysis	Joint-Pathway Analysis	Network Analysis	
Generic Format (.csv or .txt table files)	Statistical Analysis [one factor]	Statistical Analysis [metadata table]	Biomarker Analysis	Statistical Meta-analysis	Power Analysis	Other Utilities

Click here to start

roll down for more details

# Peak List Uploading

The screenshot shows the 'Peak List Uploading' module interface. The left sidebar includes icons for home, upload, processing, set parameter, view result, download, and exit. The main area has a header 'Please upload your data' and instructions about putative annotation at the individual compound level. It features two tabs: 'A peak list profile' (selected) and 'A peak intensity table'. A red box highlights the tab selection. Below is a form for 'Upload a peak list file' with fields for Ion Mode (Negative Mode), Mass Tolerance (5.0 ppm), Retention Time (Not present), Ranking (P values selected), and Data File (Choose button). A blue 'Submit' button is at the bottom. An orange box labeled '1. Switch the different uploading data type from here (peak list or table)' points to the tab selection. Another orange box labeled '2. Set parameters (mass error and ion mode) based your instrument' points to the mass tolerance field. A third orange box labeled '3. Click submit to continue' points to the 'Submit' button.

1. Switch the different uploading data type from here (peak list or table)

Please upload your data

This module supports functional analysis of untargeted metabolomics data generated from high-resolution mass spectrometry (HRMS). The basic assumption is that putative annotation at individual compound level can collectively predict changes at functional levels as defined by metabolite sets or pathways. This is because changes at group level rely on "collective behavior" which is more tolerant to random errors in compound annotation as demonstrated by [Li et al.](#) To use this approach,

- The input peak list or peak table must contain the complete data, not just significant data - we need the complete data to estimate the null model (background);
- [Required] Feature or peak names must be their numeric mass (m/z) values for putative annotation;
- [Optional] You can also provide retention time (RT) to further improve peak annotation

A peak list profile    A peak intensity table

Upload a peak list file

Ion Mode: Negative Mode (editable)

Mass Tolerance (ppm): 5.0 (editable)

Retention Time: Not present (editable)

Ranked by (1 column only): P values (selected) T scores

Enforce Primary Ions (V2 only):

Data File:

Submit

Try our example datasets

Data	Format
<input checked="" type="radio"/> IBD	Three columns (m.z, p.value, t.score)
<input type="radio"/> IBD 2	Four columns (m.z, p.value, t.score, rt)

using a Q-Exactive Plus Orbitrap (negative ion mode) from the Integrative Human Microbiome Project ([iHMP](#)).

2. Set parameters (mass error and ion mode) based your instrument

3. Click submit to continue

# Data Integrity Check

**Data Integrity Check:**

1. Checking the class labels - at least three replicates are required in each class.
2. If the samples are paired, the pair labels must conform to the specified format.
3. The data (except class labels) must not contain non-numeric values.
4. The presence of missing values or features with constant values (i.e. all zeros).

**Data processing information:**

Checking data content ...passed.

A total of 4187 m/z features were found in your uploaded data.

The instrument's mass accuracy is **5 ppm**.

The instrument's analytical mode is **negative**.

The uploaded data contains **3** columns.

The column headers of uploaded data are **m.z, p.value, t.score**.

The range of m/z peaks is trimmed to 50-2000. **0** features have been trimmed.

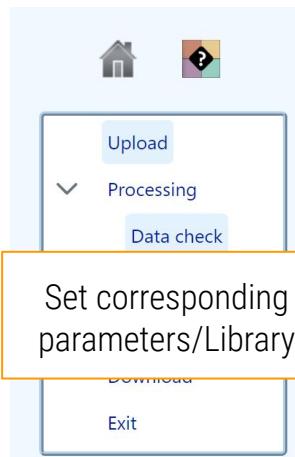
A total of 4187 input mz features were retained for further analysis.

**Edit Groups**    **Missing Values**    **► Proceed**

1. Check Data Integrity Result to make sure correct

2. Click Proceed to continue

# Set Parameters



## Specify analysis parameters:

### Algorithms

- Mummichog P-value cutoff:  (default top 10% peaks)  
 GSEA (using the overall rank based on t.score)

### Visual analytics:

- Scatter plot - test significant peaks  
 Heatmaps - test peaks in a visual pattern (good for multiple groups)

### Advanced options

[Edit Currency Metabolites](#)  
[Edit Adducts](#)

Customize the Metabolites

## Select a pathway library: (KEGG pathway info were obtained in Oct. 2019)

### Mammals

- Homo sapiens (human) [MFN] ?

### Birds

- Gallus gallus (chicken) [KEGG]

### Fish

- Danio rerio (zebrafish) [KEGG]  
 Danio rerio (zebrafish) [MTF] ?

Customize the Adducts

## Currency Metabolite Customization

Use the panels below to select metabolites to include as currency:

Available	Include
Acetoacetyl CoA (C00332)	
Acetyl coenzyme A (C00024)	
Adenosine diphosphate (C00008)	
Adenosine monophosphate (C00020)	
Carbon monoxide (C00237)	
Coenzyme A (C00010)	
Flavin adenine dinucleotide (C00016)	
FADH2 (C00016)	
Guanosine triphosphate (C00044)	
Guanosine diphosphate (C00035)	
Guanosine monophosphate (C00144)	
Hydrogen (C00282)	
Hydrogen peroxide (C00027)	
Carbonic acid (C01353)	
Pyrophosphate (C00013)	
Phosphate (C00009)	
Carbon dioxide (C00011)	

Submit

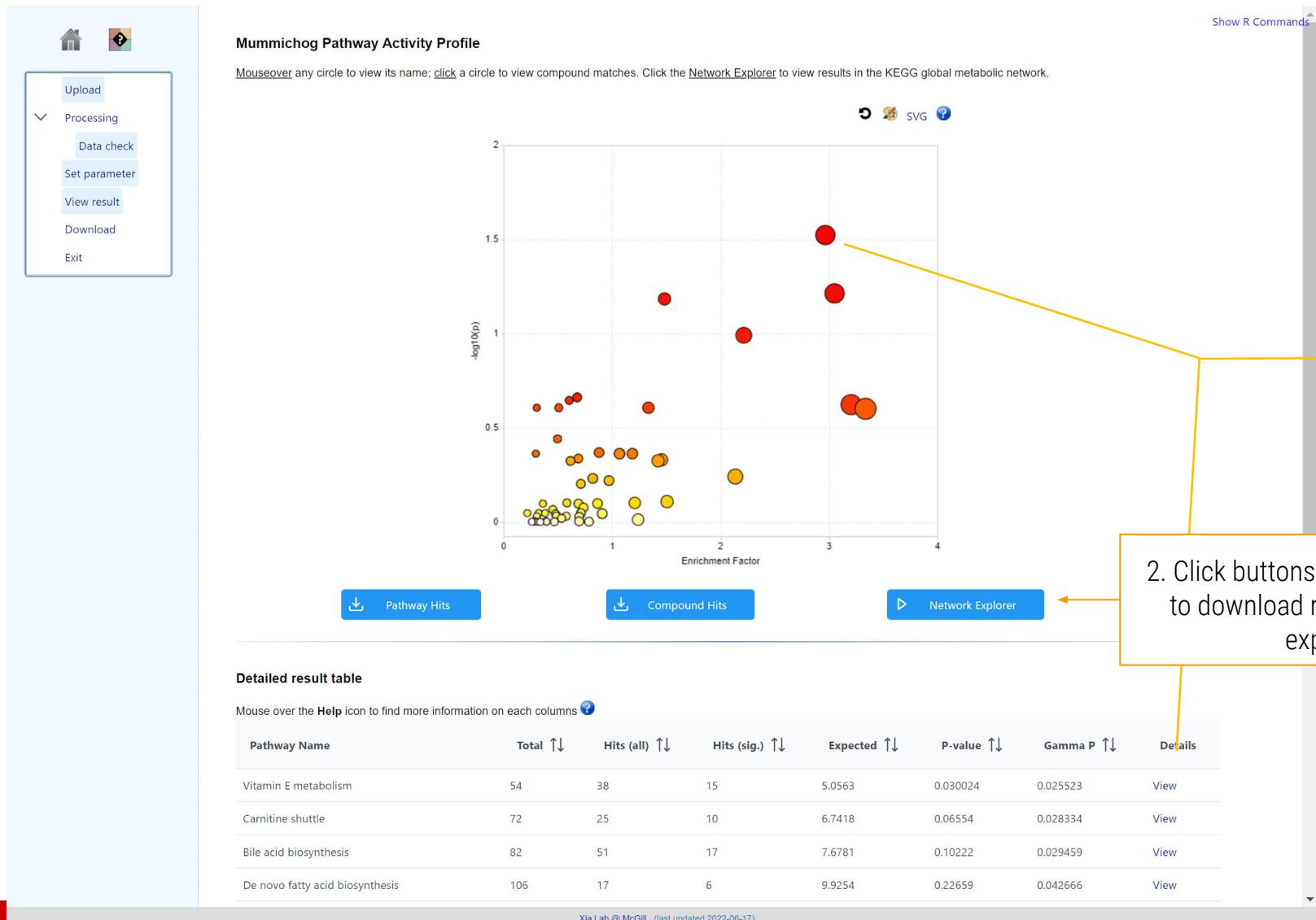
## Adduct Customization

Use the panels below to select adducts to consider:

Available	Include
M-3H [3-]	
M+FA-H [1-]	
M+Hac-H [1-]	
M+TFA-H [1-]	
2M-H [1-]	
2M+FA-H [1-]	
2M+Hac-H [1-]	
3M-H [1-]	
M-H [1-]	
M-2H [2-]	
M-H2O-H [1-]	
M+H-O [1-]	
M+K-2H [1-]	
M+Na-2H [1-]	
M+Cl [1-]	
M+Cl37 [1-]	
M+Br [1-]	
M+Br81 [1-]	
M+ACN-H [1-]	
M+HCOO [1-]	
M+CH3COO [1-]	
M(C13)-H [1-]	

Submit

# Pathway analysis results



1. The compounds/empirical compounds hits in this pathway

The colored compounds/empirical compounds indicate potential matches from the user's input, with red colors indicating significant hits and blue colors indicating non-significant hits.

Pathway	Metabolites
Vitamin E metabolism	CE5849; C00024; C00027; CE0812; CE5643; C00020; C00100; CE5848; CE5841; CE5840; CE5843; CE5842; CE5845; CE5721; CE5847; CE5723; CE6000; CE6219; CE5022; C11378; CE5021; CE5844; C02477; CE5838; CE7144; CE7145; C00601; C14153; CE7047; C00010; CE5986; CE5835; CE5837; CE5719; CE5718; CE5850; CE5851; CE5856; CE5855; CE1926; CE5899; CE1924; CE1925; CE5017; CE5846; CE4898; CE1928; CE7101; CE5655; C00088; CE7072; CE7073; CE7074; CE5839

2. Click buttons at the bottom of this page to download results or go the network exploration page

# Peak uploading – peak table

Please upload your data

This module supports functional analysis of untargeted metabolomics data generated from high-resolution mass spectrometry (HRMS). The basic assumption is that putative annotation at individual compound level can collectively predict changes at functional levels as defined by metabolite sets or pathways. This is because changes at group level rely on "collective behavior" which is more tolerant to random errors in compound annotation as demonstrated by [Li et al.](#) To use this approach,

- The input peak list or peak table must contain the complete data, not just significant data - we need the complete data to estimate the null model (background);
- [Required] Feature or peak names must be their numeric mass (m/z) values for putative annotation;
- [Optional] You can also provide retention time (RT) to further improve peak annotation

[A peak list profile](#)   [A peak intensity table](#)

**Upload a peak intensity table**

Ion Mode: Negative Mode  
Mass Tolerance (ppm): 5.0 (editable)  
Retention Time: Not present  
Data Source: Generic  
Data Format: Samples in columns  
Data File: Choose File | No file chosen

**Submit**

Try our example datasets

Data	Format	Description
<input checked="" type="radio"/> <a href="#">Immune Cell</a>	Generic peak intensity table (no retention time)	Example peak intensity table from KO experiment of dendritic cells and epithelial cells treated in DSS.
<input type="radio"/> <a href="#">Covid-19</a>	Peak intensity table with retention time	Peak intensity table of COVID-19 global metabolomics <a href="#">study</a> with over 9,000 peaks.

**1. Switch to uploading peak intensity table tab**

**2. Set parameters (mass error and ion mode) based your instrument**

# Peak uploading – Preprocessing

## Data Integrity Check:

1. Checking the class labels - at least three replicates are required in each class.
2. If the samples are paired, the pair labels must conform to the specified format.
3. The data (except class labels) must not contain non-numeric values.
4. The presence of missing values or features with constant values (i.e. all zeros).

Data processing information:

Checking data content ...passed.  
Samples are in columns and features in rows.  
The uploaded file is in comma separated values (.csv) format.  
The uploaded data file contains 12 (samples) by 4353 (peaks(mz/rt)) data matrix.  
Samples are not paired.  
4 groups were detected in samples.  
Only English letters, numbers, underscore, hyphen and forward slash (/) are allowed.  
**Other special characters or punctuations (if any) will be stripped off.**  
All data values are numeric.  
**404 features with a constant or single value across samples were found and deleted.**  
A total of 1869 (3.9%) missing values were detected.  
**By default, missing values will be replaced by 1/5 of min positive values of their corresponding variables.**  
Click the **Skip** button if you accept the default practice;  
Or click the **Missing value imputation** to use other methods.

## Data Filtering:

The purpose of the data filtering is to identify and remove variables that are unlikely to be of use when modeling the data. No phenotype information are used in the filtering process, so the result can be used with any downstream analysis. This step is strongly recommended for untargeted metabolomics datasets (i.e. spectral binning data, peak lists) with large number of variables, many of them are from baseline noises. Filtering can usually improve the results. For details, please refer to the paper by [Hacklstaedt, et al.](#)

Non-informative variables can be characterized in three groups: 1) variables of very small values (close to baseline or detection limit) - these variables can be detected using mean or median; 2) variables that are near-constant values throughout the experiment conditions (housekeeping or homeostasis) - these variables can be detected using standard deviation (SD) or the robust estimate such as interquartile range (IQR); and 3) variables that show low repeatability - this can be measured using QC samples using the relative standard deviation(RSD = SD/mean). Features with high percent RSD should be removed from the subsequent analysis (the suggested threshold is 20% for LC-MS and 30% for GC-MS). For data filtering based on the first two categories, the following empirical rules are applied during data filtering:

- Less than 250 variables: 5% will be filtered;
- Between 250 - 500 variables: 10% will be filtered;
- Between 500 - 1000 variables: 25% will be filtered;
- Over 1000 variables: 40% will be filtered;

Please note, in order to reduce the computational burden to the server, the **None** option is only for less than 5000 features. The maximum allowed number of variables is 5000. [For power analysis, the max number is 2500](#) to improve power and to control computing time. Over that, the IQR filter will still be applied to keep only top maximum features, even if you choose **None** option.

Filtering features if their RSDs are >  % in QC samples

None (less than 5000 features)  
 Interquartile range (IQR)  
 Standard deviation (SD)  
 Median absolute deviation (MAD)  
 Relative standard deviation (RSD = SD/mean)  
 Non-parametric relative standard deviation (MAD/median)  
 Mean intensity value  
 Median intensity value

## Normalization overview:

The normalization procedures are grouped into three categories. The sample normalization allows general-purpose adjustment for differences among your sample, data transformation and scaling are two different approaches to make individual features more comparable. You can use one or combine them to achieve better results.

### Sample Normalization

- None  
 Sample-specific normalization (i.e. weight, volume) [Specify](#)  
 Normalization by sum  
 Normalization by median  
 Normalization by reference sample (PQN) [Specify](#)  
 Normalization by a pooled sample from group [Specify](#)  
 Normalization by reference feature [Specify](#)  
 Quantile normalization

### Data transformation

- None  
 Log transformation (generalized logarithm transformation or glog)  
 Cube root transformation (takes the cube root of data values)

### Data scaling

- None  
 Mean centering (mean-centered only)  
 Auto scaling (mean-centered and divided by the standard deviation of each variable)  
 Pareto scaling (mean-centered and divided by the square root of the standard deviation of each variable)  
 Range scaling (mean-centered and divided by the range of each variable)

1. Perform Data Integrity Check

2. Perform Data Filtering

3. Perform Data Normalization

# Set parameters

The screenshot shows the MetaboAnalyst 5.0 web interface. On the left, a sidebar menu includes 'Upload', 'Processing' (with sub-options 'Data check', 'Missing value', 'Data filter', 'Data editor', 'Normalization', 'Set parameter' - which is highlighted in blue), 'View result', 'Metabolic network', 'Heatmap viewer', 'Download', and 'Exit'. The main area is titled 'MetaboAnalyst 5.0 - user-friendly, end-to-end metabolomics data analysis'. It features two main sections: 'Specify analysis parameters:' and 'Select a pathway library:'. The 'Specify analysis parameters:' section contains fields for 'Algorithms' (checkboxes for 'Mummichog' checked and 'GSEA' uncheckable), 'View options:' (radio buttons for 'Scatter plot (Test significant features)' checked and 'Heatmaps (Test manually selected patterns)'), and 'Advanced options' (links to 'Edit Currency Metabolites' and 'Edit Adducts'). The 'Select a pathway library:' section lists pathway libraries categorized by organism: Mammals, Birds, Fish, Insects, Nematodes, and Fungi. Under 'Mammals', 'Homo sapiens (human) [MFN]' is selected (radio button checked). Other options include 'Homo sapiens (human) [BioCyc]', 'Homo sapiens (human) [KEGG]', 'Mus musculus (mouse) [BioCyc]', 'Mus musculus (mouse) [KEGG]', 'Rattus norvegicus (rat) [KEGG]', and 'Bos taurus (cow) [KEGG]'. A large orange bracket on the right side groups the 'Set parameter' menu item and the 'Specify analysis parameters:' section, with the text 'Set corresponding parameters/Library' positioned next to it.

MetaboAnalyst 5.0 - user-friendly, end-to-end metabolomics data analysis

Specify analysis parameters:

Algorithms	<input checked="" type="checkbox"/> Mummichog <input type="checkbox"/> GSEA	P-value cutoff: <input type="text" value="1.0E-5"/> (default top 10% peaks)
View options:	<input checked="" type="radio"/> Scatter plot (Test significant features) <input type="radio"/> Heatmaps (Test manually selected patterns)	
Advanced options	<a href="#">Edit Currency Metabolites</a> <a href="#">Edit Adducts</a>	

Select a pathway library: (KEGG pathway info were obtained in Oct. 2019)

Mammals	<input checked="" type="radio"/> Homo sapiens (human) [MFN] ? <input type="radio"/> Homo sapiens (human) [BioCyc] <input type="radio"/> Homo sapiens (human) [KEGG] <input type="radio"/> Mus musculus (mouse) [BioCyc] <input type="radio"/> Mus musculus (mouse) [KEGG] <input type="radio"/> Rattus norvegicus (rat) [KEGG] <input type="radio"/> Bos taurus (cow) [KEGG]
Birds	<input type="radio"/> Gallus gallus (chicken) [KEGG]
Fish	<input type="radio"/> Danio rerio (zebrafish) [KEGG] <input type="radio"/> Danio rerio (zebrafish) [MTF] ?
Insects	<input type="radio"/> Drosophila melanogaster (fruit fly) [KEGG] <input type="radio"/> Drosophila melanogaster (fruit fly) [BioCyc]
Nematodes	<input type="radio"/> Caenorhabditis elegans (nematode) [KEGG]
Fungi	<input type="radio"/> Saccharomyces cerevisiae (yeast) [KEGG] <input type="radio"/> Saccharomyces cerevisiae (yeast) [BioCyc]

# Heatmap based pattern specific analysis

The screenshot shows the MetaboAnalyst 5.0 web interface. At the top, there is a logo and the text "MetaboAnalyst 5.0 - user-friendly, end-to-end metabolomics data analysis". On the left, a sidebar menu includes "Upload", "Processing" (with sub-options: Data check, Missing value, Data filter, Data editor), "Normalization", "Set parameter", "View result", "Metabolic network", "Heatmap viewer", "Download", and "Exit". The main area is titled "Specify analysis parameters". It contains two tabs: "Algorithms" (with "Mummichog" checked and "GSEA" unselected) and "View options" (with "Scatter plot (Test significant features)" and "Heatmaps (Test manually selected patterns)" radio buttons; the "Heatmaps" button is highlighted with a red box and an orange arrow points to it from a callout box). Below these tabs are "Advanced options" (with links to "Edit Currency Metabolites" and "Edit Adducts"). A section titled "Select a pathway library" follows, listing various organisms categorized by taxonomic group: Mammals, Birds, Fish, Insects, Nematodes, and Fungi. The "Homo sapiens (human) [MFN]" option under Mammals is selected. A callout box on the right says "1. Select Heatmaps radio to start!". At the bottom, there is footer text: "Xia Lab @ McGill (last updated 2021-01-09)" and "bioinformatics.ca".



Controller panel to adjust parameters for clustering.

Overview of the peak across the spectrum. Default is clustered based on p value.

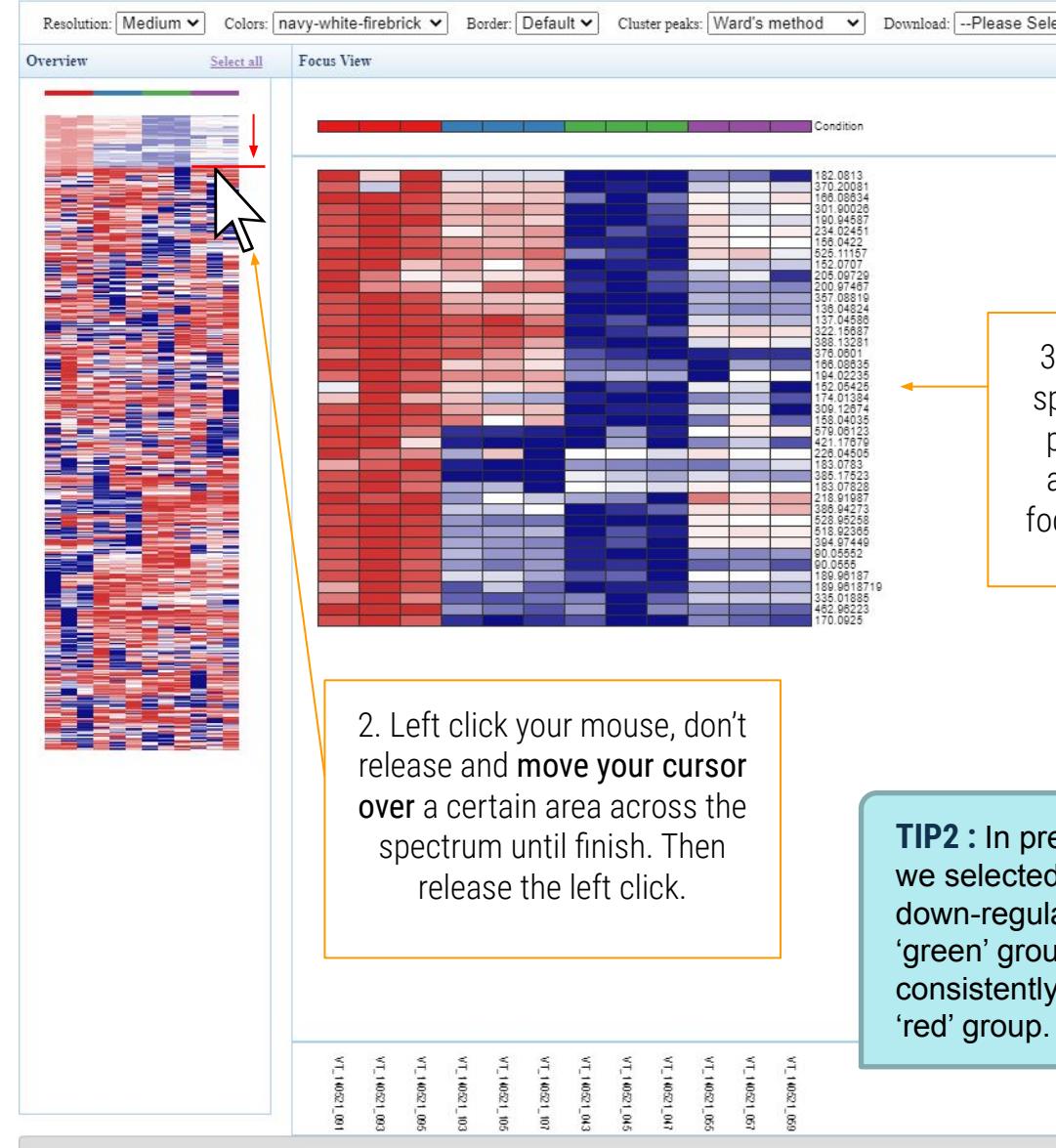
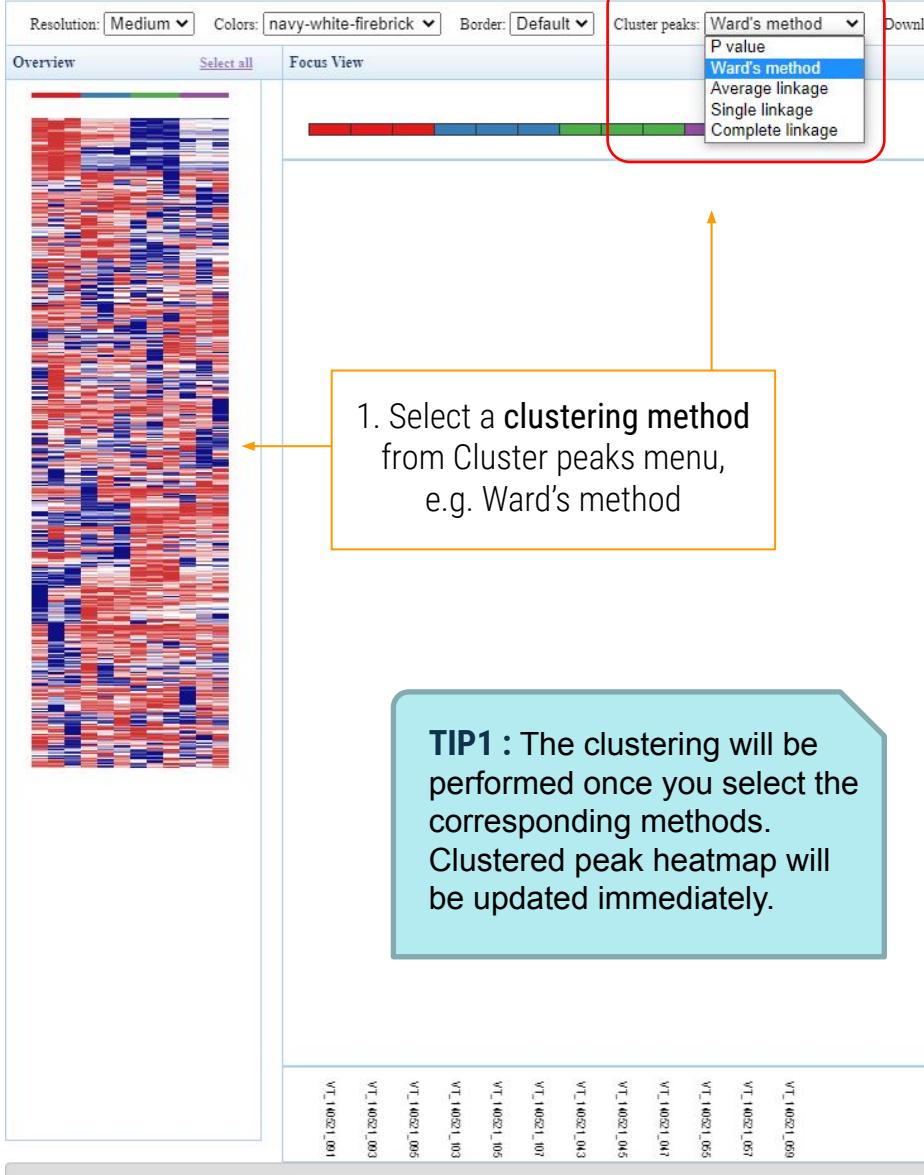
Sample Names View panel.

Focus view of the specific peak pattern from the whole spectrum. Default is the top 50 peaks.

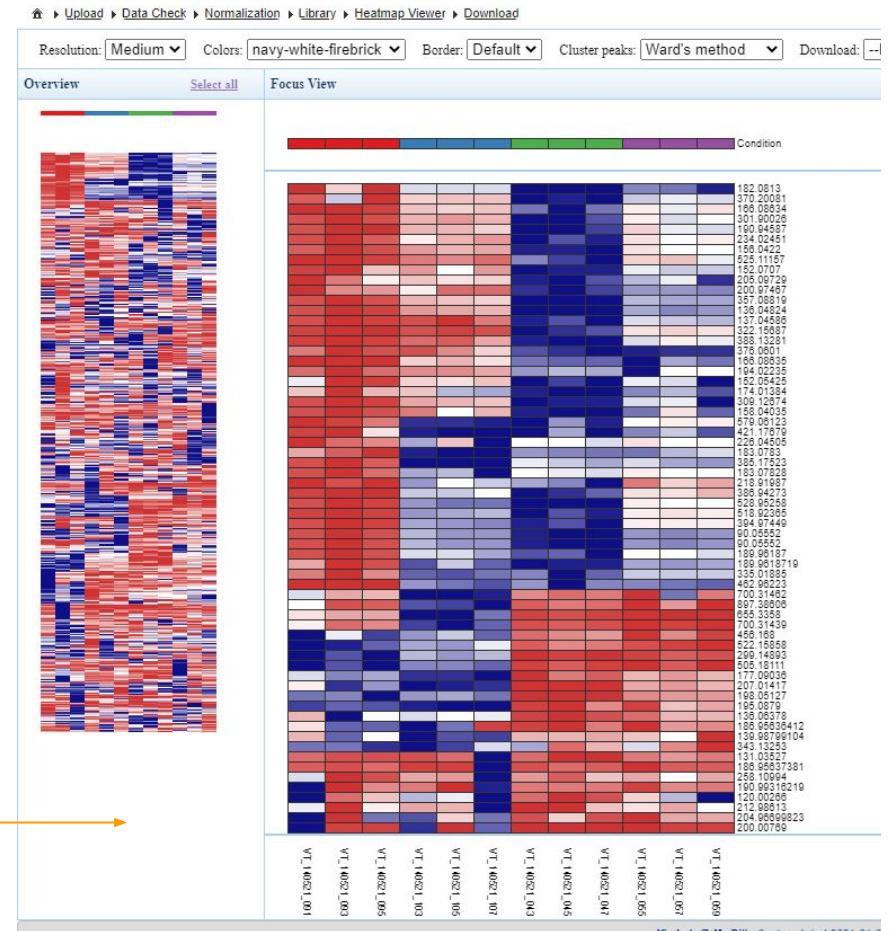
- Sample: VT\_140521\_045
- Metadata: Condition
- Label: DSS\_WT\_dDC

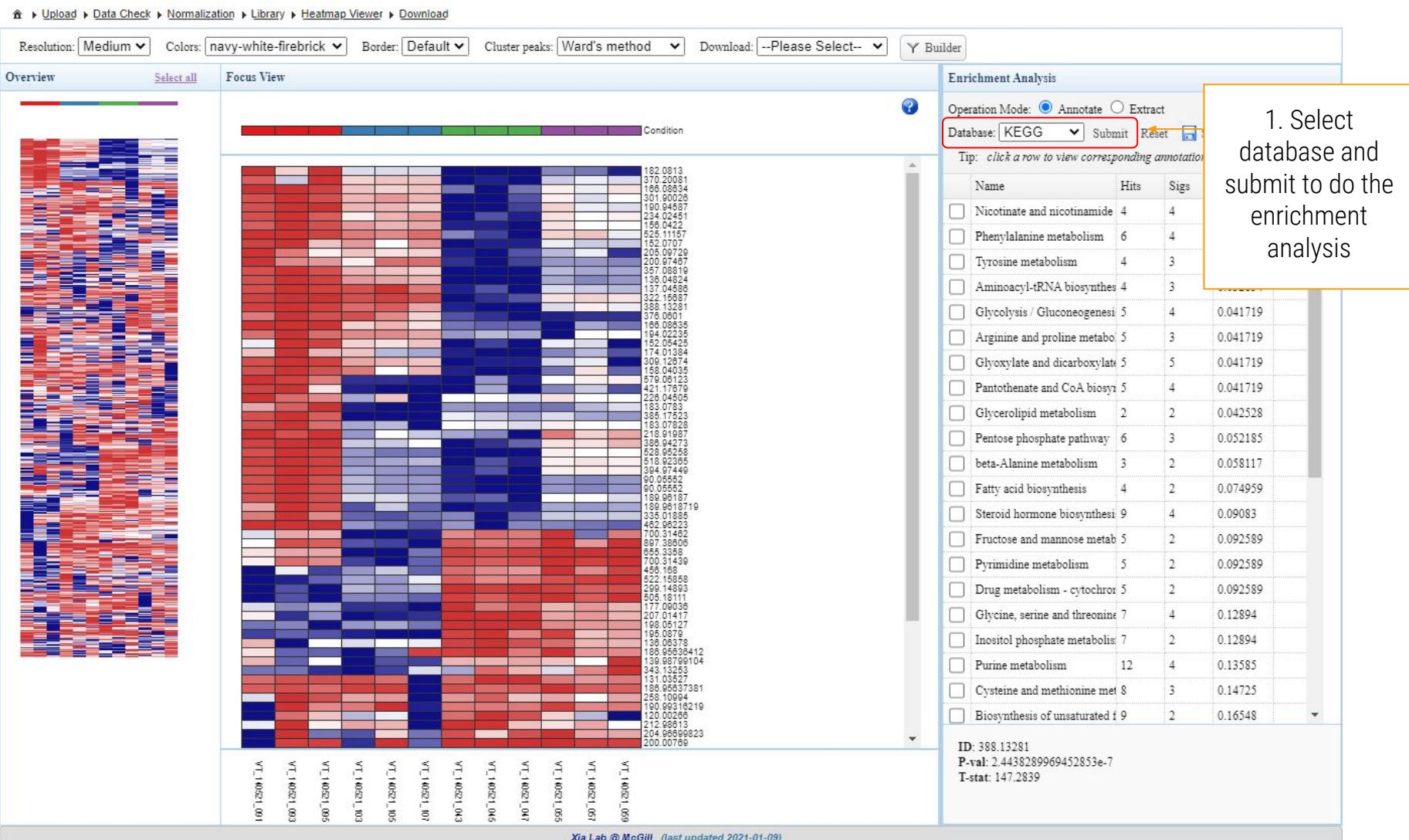
Pattern based Enrichment analysis panel.

Dynamic Display panel, used to show the peak/sample information dynamically.









1. Select database and submit to do the enrichment analysis

Resolution: Medium Colors: navy-white-firebrick Border: Default Cluster peaks: Ward's method Download: --Please Select-- Builder

**Overview** [Select all](#) **Focus View**

Condition

182.0813  
370.20081  
186.08834  
186.08835  
190.04587  
234.02451  
186.04221  
186.01587  
182.0707  
205.09729  
200.97467  
186.08836  
186.04824  
137.04588  
322.15887  
388.15821  
186.08837  
186.08835  
194.02235  
186.04425  
186.01584  
300.12874  
186.04035  
579.09123  
186.01585  
226.04505  
183.0783  
388.17523  
186.01586  
388.04273  
528.95258  
518.92385  
186.01587  
90.05852  
90.05852  
189.98187  
189.98171  
335.01885  
462.98223  
170.0925  
170.31482  
867.38808  
655.3358  
186.01588  
145.187  
522.15888  
299.14893  
505.18111  
186.01589  
207.01417  
198.05127  
195.0579  
186.01587  
186.08834812  
139.9879104  
343.13253  
186.01587  
186.08837381  
258.10994  
190.99315219  
120.09813  
186.08833  
204.98699823  
200.00769  
334.09183  
186.08833  
64.0458376  
184.98583457 M-H<sup>+</sup> | C00631  
140.01083899 M+Cl<sup>-</sup> | C00065  
186.08834808  
218.98381  
481.38272  
325.98218

M-H<sup>+</sup> | C00631  
140.01083899 M+Cl<sup>-</sup> | C00065

Tip: click a row to view corresponding annotations or to extract

Name	Hits	Sigs	Gamma-p	Color
Nicotinate and nicotinamide	4	4	0.019178	P0
Phenylalanine metabolism	6	4	0.027558	P1
Tyrosine metabolism	4	3	0.031897	P2
Aminoacyl-tRNA biosynthes	4	3	0.031897	P3
Glycolysis / Gluconeogenesi	5	4	0.041229	P4
Arginine and proline metabo	5	3	0.041229	P5
Glyoxylate and dicarboxylat	5	5	0.041229	P6
Pantothenate and CoA biosy	5	4	0.041229	
Glycerolipid metabolism	2	2	0.04151	
Pentose phosphate pathway	6	3	0.052271	
beta-Alanine metabolism	3	2	0.057704	
Alanine, aspartate and glutamine synthesis	4	2	0.075283	
Propanoate metabolism	4	3	0.075283	
Urea biosynthesi	9	4	0.093129	
mannose metab	5	2	0.093719	
Metabolism	5	2	0.093719	
Glutathione - cytochro	5	2	0.093719	
Aspartate and threonine	7	4	0.13172	
Glutamate metaboli	7	2	0.13172	
methionine me	8	3	0.15084	
Biosynthesis of unsaturated	9	2	0.16985	
Pentose and glucuronate inte	10	5	0.18869	
Tryptophan metabolism	10	3	0.18869	
Citrate cycle (TCA cycle)	4	2	1	
Primary bile acid biosynthet	7	5	1	

Sample: VT\_140521\_059  
Metadata: Condition  
Label: DSS\_WT\_Epi

2. Select the pathways you are interested in, then the corresponding hits will appear at the right side of the Focus view panel.

# **Statistical Analysis (one-factor) Demos**

# MetaboAnalyst: Statistical Analysis (one-factor)

<https://www.metaboanalyst.ca/MetaboAnalyst/ModuleView.xhtml>

## Module Overview

Input Data Type	Available Modules (click on a module to proceed, or scroll down for more details)					
Raw Spectra (mzML, mzXML or mzData)			LC-MS Spectra Processing			
MS Peaks (peak list or intensity table)			Functional Analysis	Functional Meta-analysis		
Annotated Features (compound list or table)			Enrichment Analysis	Pathway Analysis	Joint-Pathway Analysis	Network Analysis
Generic Format (.csv or .txt table files)	Statistical Analysis [one factor]	Statistical Analysis [metadata table]	Biomarker Analysis	Statistical Meta-analysis	Power Analysis	Other Utilities

Please use [OmicsForum](#) for support & troubleshooting request

# Data Upload

## Try our test data

Data Type	Description
<input checked="" type="radio"/> <a href="#">Concentrations</a>	Metabolite concentrations of 77 urine samples from cancer patients measured by 1H NMR ( <a href="#">Eisner R, et al.</a> ). Group 1- cachexic; group 2 - control
<input type="radio"/> <a href="#">Concentrations</a>	Metabolite concentrations of 39 rumen samples measured by proton NMR from dairy cows fed with different proportions of barley grain ( <a href="#">Ametaj BN, et al.</a> ). Group label - 0, 15, 30, or 45 - indicating the percentage of grain in diet.
<input type="radio"/> <a href="#">NMR spectral bins</a>	Binned 1H NMR spectra of 50 urine samples using 0.04 ppm constant width ( <a href="#">Psihogios NG, et al.</a> ) Group 1- control; group 2 - severe kidney disease.
<input type="radio"/> <a href="#">NMR peak lists</a>	Peak lists and intensity files for 50 urine samples measured by 1H NMR ( <a href="#">Psihogios NG, et al.</a> ). Group 1- control; group 2 - severe kidney disease.
<input type="radio"/> <a href="#">Concentrations (paired)</a>	Compound concentrations of 14 urine samples collected from 7 cows at two time points using 1H NMR (unpublished data). Group 1- day 1, group 2- day 4.
<input type="radio"/> <a href="#">MS peak intensities</a>	LC-MS peak intensity table for 12 mice spinal cord samples ( <a href="#">Saghatelyan et al.</a> ). Group 1- wild-type; group 2 - knock-out.
<input type="radio"/> <a href="#">MS peak lists</a>	Three-column LC-MS peak list files for 12 mice spinal cord samples ( <a href="#">Saghatelyan et al.</a> ). Group 1- wild-type; group 2 - knock-out.
<input type="radio"/> <a href="#">LC-MS mzTab</a>	LC-MS mzTab file of 15 mouse liver samples collected using LTQ Orbitrap Velos by ( <a href="#">Hartler et al.</a> ) Group 1 - mouse liver 1; group 2 - mouse liver 2; group 3 - mouse liver 3.
<input type="radio"/> <a href="#">GC-MS mzTab</a>	GC-MS mzTab file of 6 <i>Arabidopsis</i> samples obtained using ( <a href="#">MS-DIAL</a> ). Group 1 - cont; group 2 - MeKo.

Click Submit

Submit

# Data Check

The screenshot shows the Data Check software interface. On the left is a sidebar with icons for Home and Help, and a list of options: Upload, Processing (expanded), Data check (selected and highlighted in blue), Missing value, Data filter, Data editor, Normalization, Statistics, Download, and Exit. The main area is titled "Data Integrity Check:" and contains the following text:

**Data Integrity Check:**

1. Checking the class labels - at least three replicates are required in each class.
2. If the samples are paired, the pair labels must conform to the specified format.
3. The data (except class labels) must not contain non-numeric values.
4. The presence of missing values or features with constant values (i.e. all zeros).

**Data processing information:**

Checking data content ...passed.  
Samples are in rows and features in columns  
The uploaded file is in comma separated values (.csv) format.  
The uploaded data file contains 77 (samples) by 63 (compounds) data matrix.  
Samples are not paired.  
2 groups were detected in samples.  
Only English letters, numbers, underscore, hyphen and forward slash (/) are allowed.  
Other special characters or punctuations (if any) will be stripped off.  
All data values are numeric.  
A total of 0 (0%) missing values were detected.  
[By default, missing values will be replaced by 1/5 of min positive values of their corresponding variables](#)  
Click the **Proceed** button if you accept the default practice;  
Or click the **Missing Values** button to use other methods.

**Edit Groups**    **Missing Values**    **► Proceed**

Click Proceed

# Normalization

Select "Normalization by median", "Log transformation", and "None". Click "Normalize" and then "Proceed".

The screenshot shows the MetaboAnalyst software interface. On the left, a sidebar menu is open under the 'Processing' section, with 'Normalization' selected. On the right, the main configuration area is displayed, divided into three sections: Sample normalization, Data transformation, and Data scaling.

- Sample normalization:**
  - None
  - Sample-specific normalization (i.e. weight, volume) [Specify](#)
  - Normalization by sum
  - Normalization by median
  - Normalization by a reference sample (PQN) [Specify](#)
  - Normalization by a pooled sample from group (group PQN) [Specify](#)
  - Normalization by reference feature [Specify](#)
  - Quantile normalization (suggested only for > 1000 features)
- Data transformation:**
  - None
  - Log transformation (base 10)
  - Square root transformation (square root of data values)
  - Cube root transformation (cube root of data values)
- Data scaling:**
  - None
  - Mean centering (mean-centered only)
  - Auto scaling (mean-centered and divided by the standard deviation of each variable)
  - Pareto scaling (mean-centered and divided by the square root of the standard deviation of each variable)
  - Range scaling (mean-centered and divided by the range of each variable)

At the bottom of the configuration area, there are three buttons: 'Normalize', 'View Result', and 'Proceed'.

**Sample normalization**

- None
- Sample-specific normalization (i.e. weight, volume) [Specify](#)
- Normalization by sum
- Normalization by median
- Normalization by a reference sample (PQN) [Specify](#)
- Normalization by a pooled sample from group (group PQN) [Specify](#)
- Normalization by reference feature [Specify](#)
- Quantile normalization (suggested only for > 1000 features)

**Data transformation**

- None
- Log transformation (base 10)
- Square root transformation (square root of data values)
- Cube root transformation (cube root of data values)

**Data scaling**

- None
- Mean centering (mean-centered only)
- Auto scaling (mean-centered and divided by the standard deviation of each variable)
- Pareto scaling (mean-centered and divided by the square root of the standard deviation of each variable)
- Range scaling (mean-centered and divided by the range of each variable)

# Analysis Overview

There are many analysis options. We will explore:

- T-test
- Correlation Heatmaps
- PCA
- PLS-DA
- Dendrogram

The screenshot shows a software interface for data analysis. On the left, a vertical menu bar includes icons for Home and Help, followed by sections for Upload, Processing (with sub-options like Data check, Missing value, Data filter, Data editor), Normalization, Statistics (selected), Download, and Exit. To the right, a large panel titled "Select an analysis path to explore:" lists various statistical and chemometric methods. An orange box highlights the "T-tests" link under the "Univariate Analysis" section, with an arrow pointing from the text "Click T-tests" to it.

Select an analysis path to explore :

**Univariate Analysis**

[Fold Change Analysis](#) [T-tests](#) [Volcano plot](#)  
One-way Analysis of Variance (ANOVA)  
[Correlation Heatmaps](#) [Pattern Search](#) [Correlation Networks \(DSPC\)](#)

**Advanced Significance Analysis**

[Significance Analysis of Microarray \(and Metabolites\) \(SAM\)](#)  
[Empirical Bayesian Analysis of Microarray \(and Metabolites\) \(EBAM\)](#)

**Chemometrics Analysis**

[Principal Component Analysis \(PCA\)](#)  
[Partial Least Squares - Discriminant Analysis \(PLS-DA\)](#)  
[Sparse Partial Least Squares - Discriminant Analysis \(sPLS-DA\)](#)  
[Orthogonal Partial Least Squares - Discriminant Analysis \(orthoPLS-DA\)](#)

**Cluster Analysis**

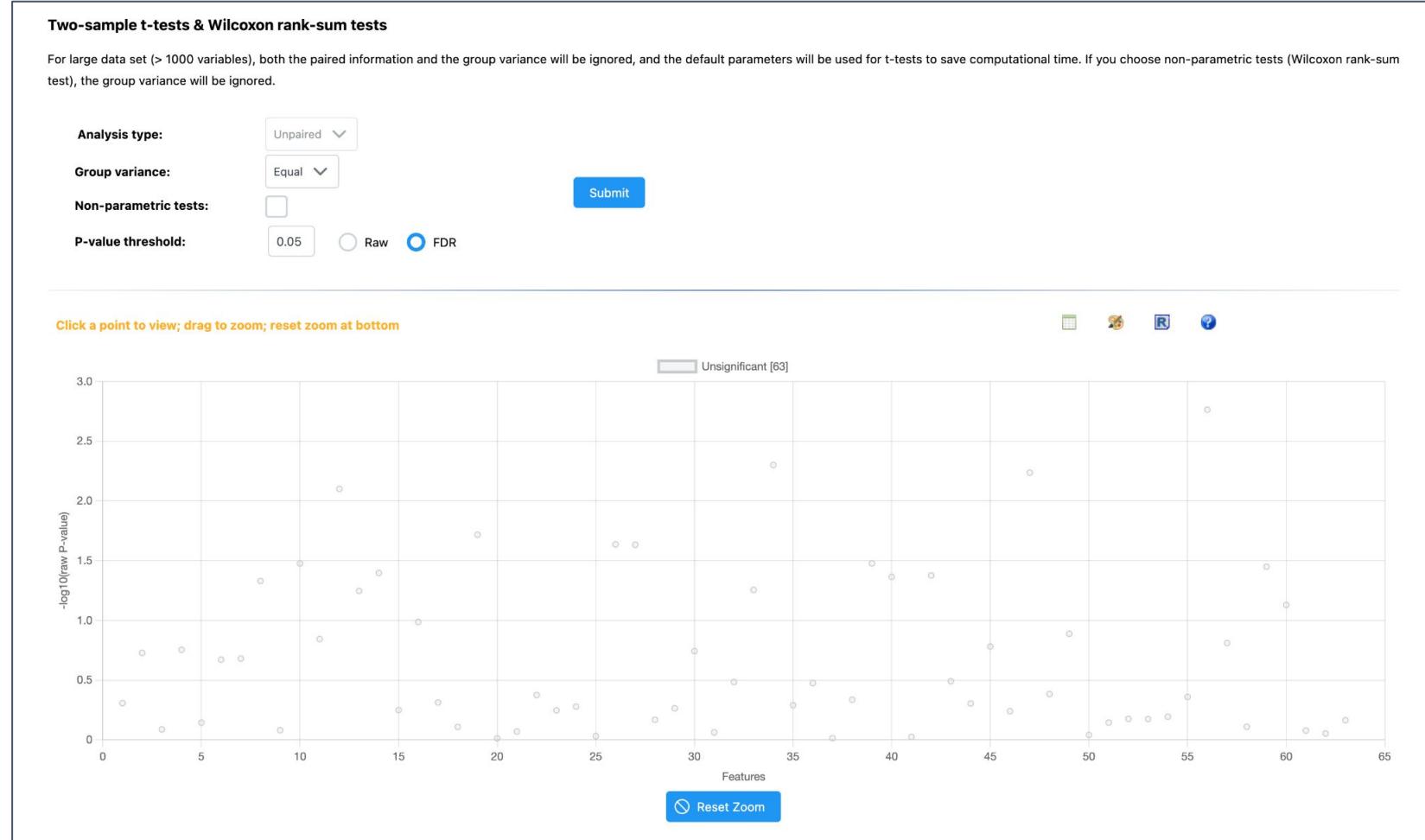
Hierarchical Clustering: [Dendrogram](#) [Heatmaps](#)  
Partitional Clustering: [K-means](#) [Self Organizing Map \(SOM\)](#)

**Classification & Feature Selection**

[Random Forest](#)  
[Support Vector Machine \(SVM\)](#)

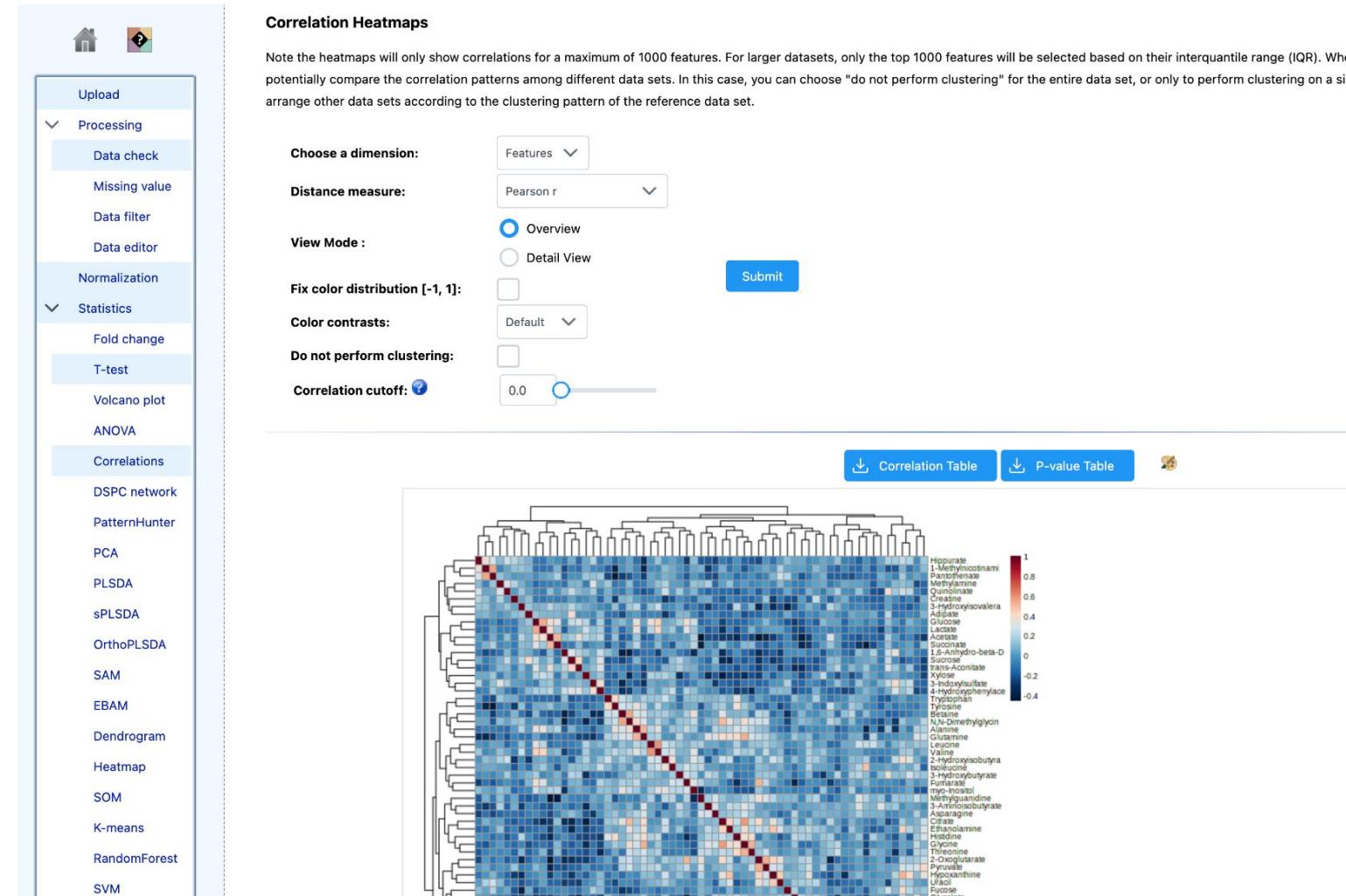
# T-tests

No metabolites have statistically significant differences after multiple hypothesis correction. Navigate back to the Analysis Overview page and click "Correlation Heatmaps".



# Correlation Heatmap

The heatmap shows pairwise correlations between features. Groups of features with similar patterns form clusters in the heatmap. Adjust the parameters and click "Submit" to explore the settings. When done, navigate back to the analysis overview and click "PCA".



# PCA

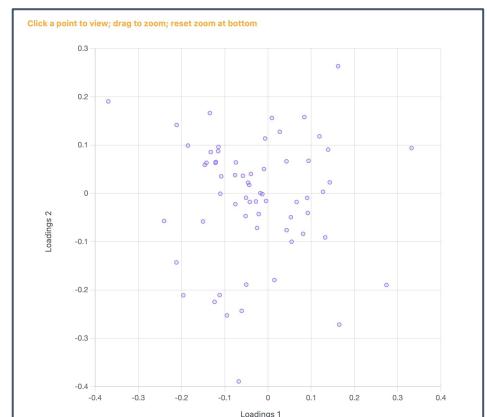
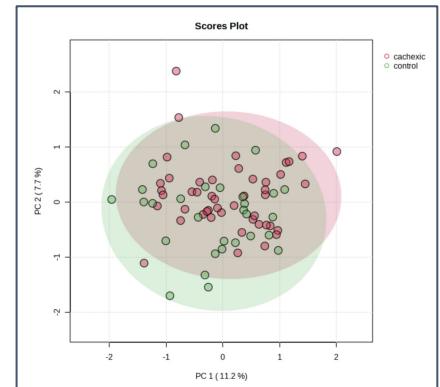
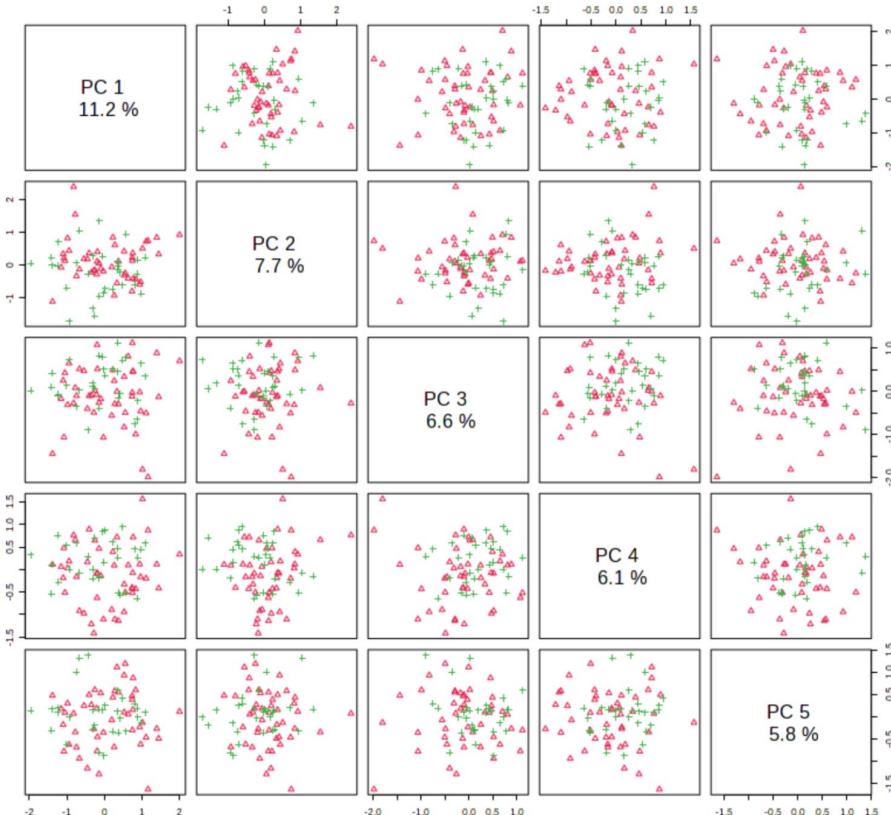
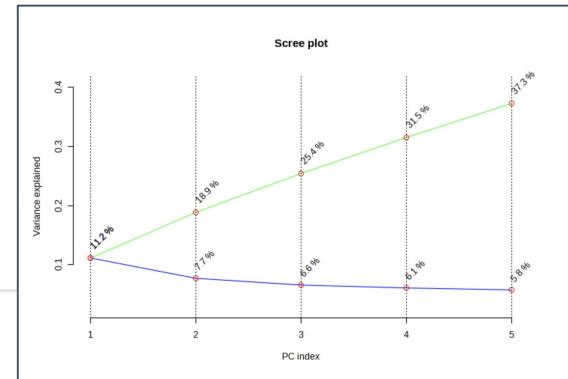
There are many views of the PCA results in the different tabs. Click some to explore the results. When finished, navigate back to the analysis overview and click "PLS-DA".

## Principal Component Analysis (PCA)

Overview Scree Plot 2D Scores Plot Loadings Plot Synchronized 3D Plots Biplot

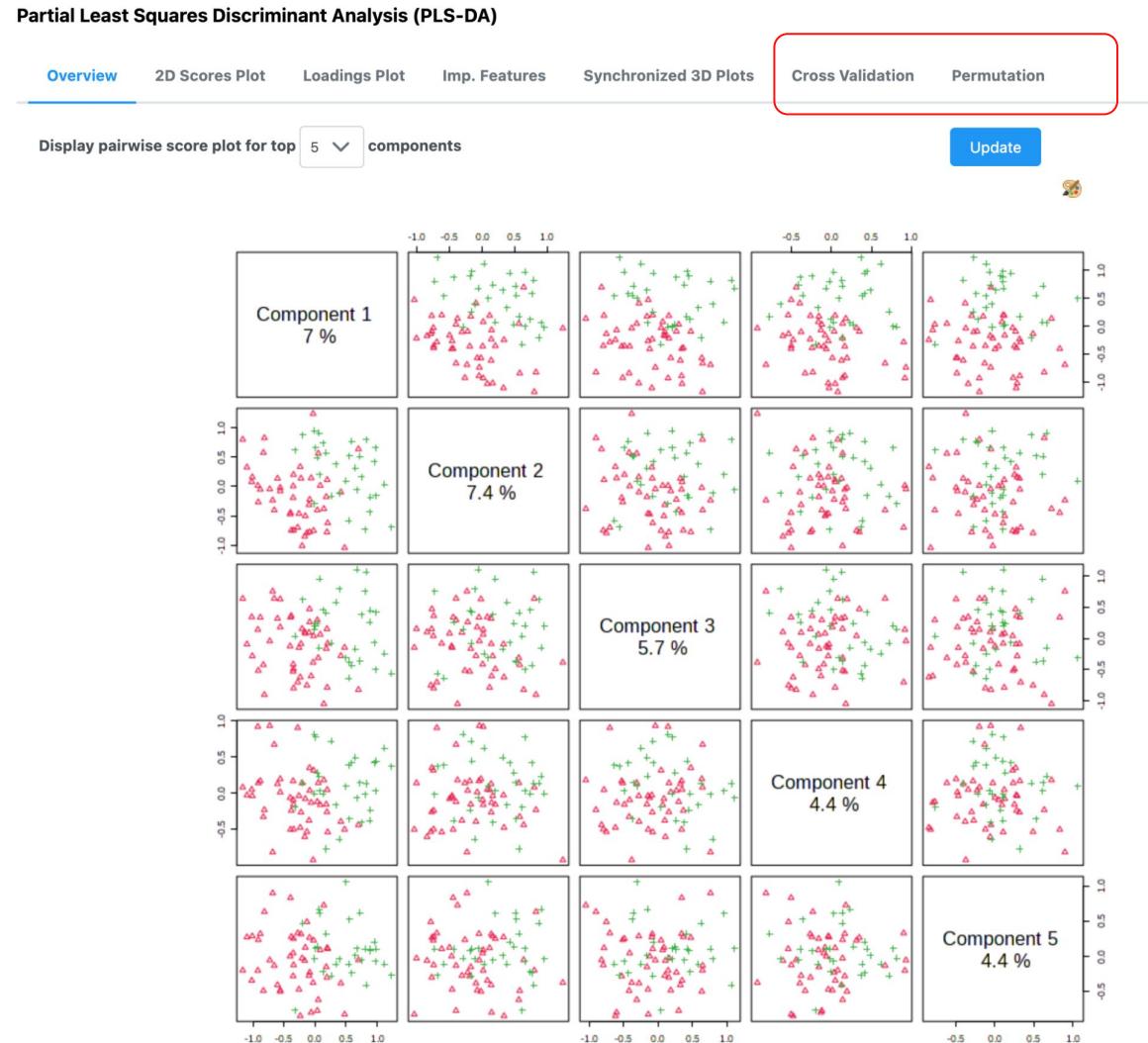
Display pairwise score plot for top 5 PCs

Update



# PLS-DA

The PLS-DA has many similar tabs compared to the PCA tool. Crucially, there is a tab for cross validation and permutation analysis. Make sure to try out this functionality, and read all the help tips to learn how to use it. When finished, navigate back and click "Dendrogram".



# Dendrogram

Quickly generate dendograms using different clustering algorithms.

Hierarchical Clustering Dendrogram

Distance Measure:

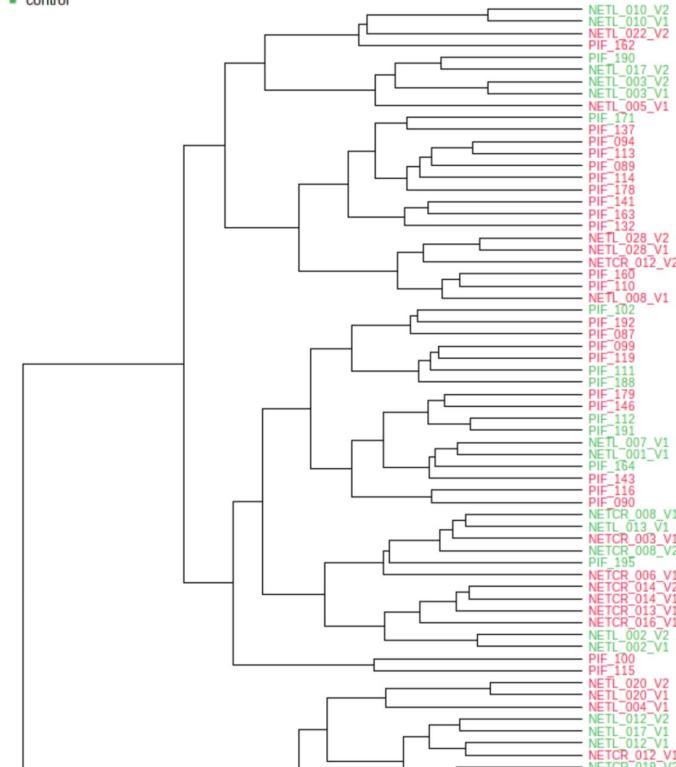
Euclidean

Submit

Clustering Algorithm:

Ward

■ cachexic  
■ control



# **Statistical Analysis (meta-data) Demos**

# MetaboAnalyst: Statistical Analysis (metadata table)

<https://www.metaboanalyst.ca/MetaboAnalyst/ModuleView.xhtml>

## Module Overview

Input Data Type	Available Modules (click on a module to proceed, or scroll down for more details)					
Raw Spectra (mzML, mzXML or mzData)			LC-MS Spectra Processing			
MS Peaks (peak list or intensity table)			Functional Analysis	Functional Meta-analysis		
Annotated Features (compound list or table)			Enrichment Analysis	Pathway Analysis	Joint-Pathway Analysis	Network Analysis
Generic Format (.csv or .txt table files)	Statistical Analysis [one factor]	Statistical Analysis [metadata table]	Biomarker Analysis	Statistical Meta-analysis	Power Analysis	Other Utilities

Please use [OmicsForum](#) for support & troubleshooting request

# Data Upload

Select the **second** example dataset and click "Submit".

## Try our test data

Data	Study Design	Description
<input type="radio"/> <a href="#">Data</a> <a href="#">Metadata</a>	Multiple factors / covariates	LC-MS peak intensity data table and meta-data of 20 healthy and 39 COVID-19 patient samples; <b>Four metadata</b> - 3 categorical and 1 numeric.
<input checked="" type="radio"/> <a href="#">Data</a> <a href="#">Metadata</a>	Multiple factors / covariates	LC-MS peak intensity data from plasma samples of 175 individuals to study trichloroethylene (TCE) exposure. <b>Nine metadata</b> - 6 categorical and 3 numeric. Please refer to <a href="#">Walker D. et al</a> for more details.
<input type="radio"/> <a href="#">Data</a> <a href="#">Metadata</a>	Time series + one condition	LC-MS peak intensity data collected from <i>Arabidopsis thaliana</i> during a wounding time course (four time points). <b>WT</b> - wild type; <b>MT</b> - <i>dde2-2</i> mutant. Please refer to <a href="#">(Meinicke P. et al)</a> for more information
<input type="radio"/> <a href="#">Data</a> <a href="#">Metadata</a>	Time series only data	LC-MS peak intensity data collected from only <b>wild type</b> <i>Arabidopsis thaliana</i> during a wounding time course (four time points). Please refer to <a href="#">(Meinicke P. et al)</a> for more information
<input type="radio"/> <a href="#">Data</a> <a href="#">Metadata</a>	Multiple factors / covariates	Lipidomics data collected from plasma of patients along the spectrum of diabetes progression. <b>Five metadata</b> - 1 categorical and 4 numeric. Please refer to <a href="#">Wigger L. et al.</a> for more information.

Submit

# Data Check (Missing Values)

## Data Integrity Check:

1. Checking the class labels - at least three replicates are required in each class.
2. If the samples are paired, the pair labels must conform to the specified format.
3. The data (except class labels) must not contain non-numeric values.
4. The presence of missing values or features with constant values (i.e. all zeros).

**Data processing information:**

Checking data content ...passed.  
Samples are in columns and features in rows.  
The uploaded data file contains 175 (samples) by 7830 (peaks(mz/rt)) data matrix.  
The data is not time-series data.  
3 groups were detected from primary meta-data factor: TCE\_Exp\_Category.  
Only English letters, numbers, underscore, hyphen and forward slash (/) are allowed.  
**Other special characters or punctuations (if any) will be stripped off.**  
All data values are numeric.  
A total of 193204 (14.1%) missing values were detected.  
[By default, missing values will be replaced by 1/5 of min positive values of their corresponding variables](#)  
Click the **Proceed** button if you accept the default practice;  
Or click the **Missing Values** button to use other methods.

1 - Click "Missing Values"

Missing Values

▷ Proceed

## Step 1. Remove features with too many missing values

Remove features with >  % missing values

## Step 2. Estimate the remaining missing values

Replace by LoDs (1/5 of the minimum positive value of each variable)

Exclude variables with missing values

Replace by column (feature)

Estimate missing values using

▷ Process

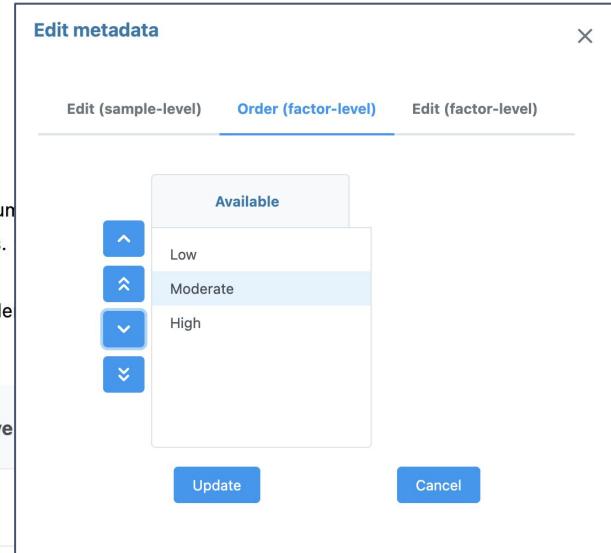
2 - Click "Submit"

# Metadata Check

Update your metadata using the table below

- Update metadata type: categorical option for experimental groups (i.e. control vs diseased), continuous for numerical variables.
- Edit metadata content: click Edit to modify underlying groups to address those that do not meet requirements.
- Modify metadata name: click on corresponding cell on the main table to modify name
- Specify group order of categorical metadata: click Edit and go to Order tab to specify the order which the underlying groups appear.
- Exclude metadata that do not pass sanity check.

Name	Status	Type	Edit	Remove
TCE_Exp_Category	OK	Categorical ▾	Edit	✖
TCE_Exp_Conc	OK	Continuous ▾	Edit	✖
Age	OK	Continuous ▾	Edit	✖
Sex	OK	Categorical ▾	Edit	✖
Smoking_Status	OK	Categorical ▾	Edit	✖
Alcohol_Use	OK	Categorical ▾	Edit	✖
BMI	OK	Continuous ▾	Edit	✖
Batch	OK	Categorical ▾	Edit	✖



1 - Verify that the variable types are correct

2 - Click "Edit" in the "TCE\_Exp\_Category" row and re-order the groups to Low-Medium-High. Click "Update"

▷ Skip to analysis

▷ Proceed

3 - Click "Proceed"

# Data Filtering

## Data Filtering:

The purpose of the data filtering is to identify and remove variables that are unlikely to be of use when modeling the data. No phenotype information is lost during downstream analysis. This step is strongly recommended for untargeted metabolomics datasets (i.e. spectral binning data, peak lists) with little prior knowledge. For details, please refer to the paper by [Hackstadt, et al.](#).

Non-informative variables can be characterized in three groups: 1) variables that show **low repeatability** - this can be measured using QC samples; 2) variables that can be detected using standard deviation (SD); or the robust estimate such as interquartile range (IQR); and 3) variables of **very small values** (e.g. median).

For data filtering based on the last two categories, the default parameters follow the empirical rules: 1) Less than 250 variables: 5% will be filtered; 2) 250-500 variables: 25% will be filtered; and 3) Over 500 variables: 40% will be filtered. You can turn off data filtering by dragging the slider to adjust the percentage (e.g. 2500 for power analysis) to control computing time on our server.

The screenshot shows a user interface for data filtering. At the top, there is a button labeled "Filter based on QC". Below it, a checkbox is checked, and a slider is set to 25%, with the text "Filtering features if their RSDs are > 25% in QC samples". Under the heading "Statistical Filters", several options are listed: "Interquartile range (IQR)" (selected), "Standard deviation (SD)", "Median absolute deviation (MAD)", "Relative standard deviation (RSD = SD/mean)", "Non-parametric relative standard deviation (MAD/median)", "Mean intensity value", and "Median intensity value". To the right of these filters, a slider is set to 40%, with the text "Percentage to filter out: 40%". At the bottom, there are two buttons: "Submit" and "Proceed".

Click "Submit" and "Proceed"

# Normalization

Select “Normalization by median”, “Log transformation”, and “None”. Click “Normalize” and “Proceed”.

**Sample normalization**

- None
- Sample-specific normalization (i.e. weight, volume) [Specify](#)
- Normalization by sum
- Normalization by median
- Normalization by a reference sample (PQN) [Specify](#)
- Normalization by a pooled sample from group (group PQN) [Specify](#)
- Normalization by reference feature [Specify](#)
- Quantile normalization (suggested only for > 1000 features)

**Data transformation**

- None
- Log transformation (base 10)
- Square root transformation (square root of data values)
- Cube root transformation (cube root of data values)

**Data scaling**

- None
- Mean centering (mean-centered only)
- Auto scaling (mean-centered and divided by the standard deviation of each variable)
- Pareto scaling (mean-centered and divided by the square root of the standard deviation of each variable)
- Range scaling (mean-centered and divided by the range of each variable)

Normalize

View Result

Proceed

# Analysis Overview

There are many options for analyzing the data.

Here, we will first use the “Metadata Visualization” and “Interactive PCA Visualization” to understand general trends with respect to the metadata. Then, we will use “Linear Models with Covariate Adjustment” and “Random Forest” to analyze the data. Click “Metadata Visualization”.

The screenshot shows a software interface for data analysis. On the left, a sidebar menu includes: Home, Upload, Processing (with sub-options: Data check, Missing value, Metadata check, Data filter, Data editor), Normalization, Multi-factors (selected), Download, and Exit. The main content area is titled "Select an analysis path to explore :". It lists several analysis methods:

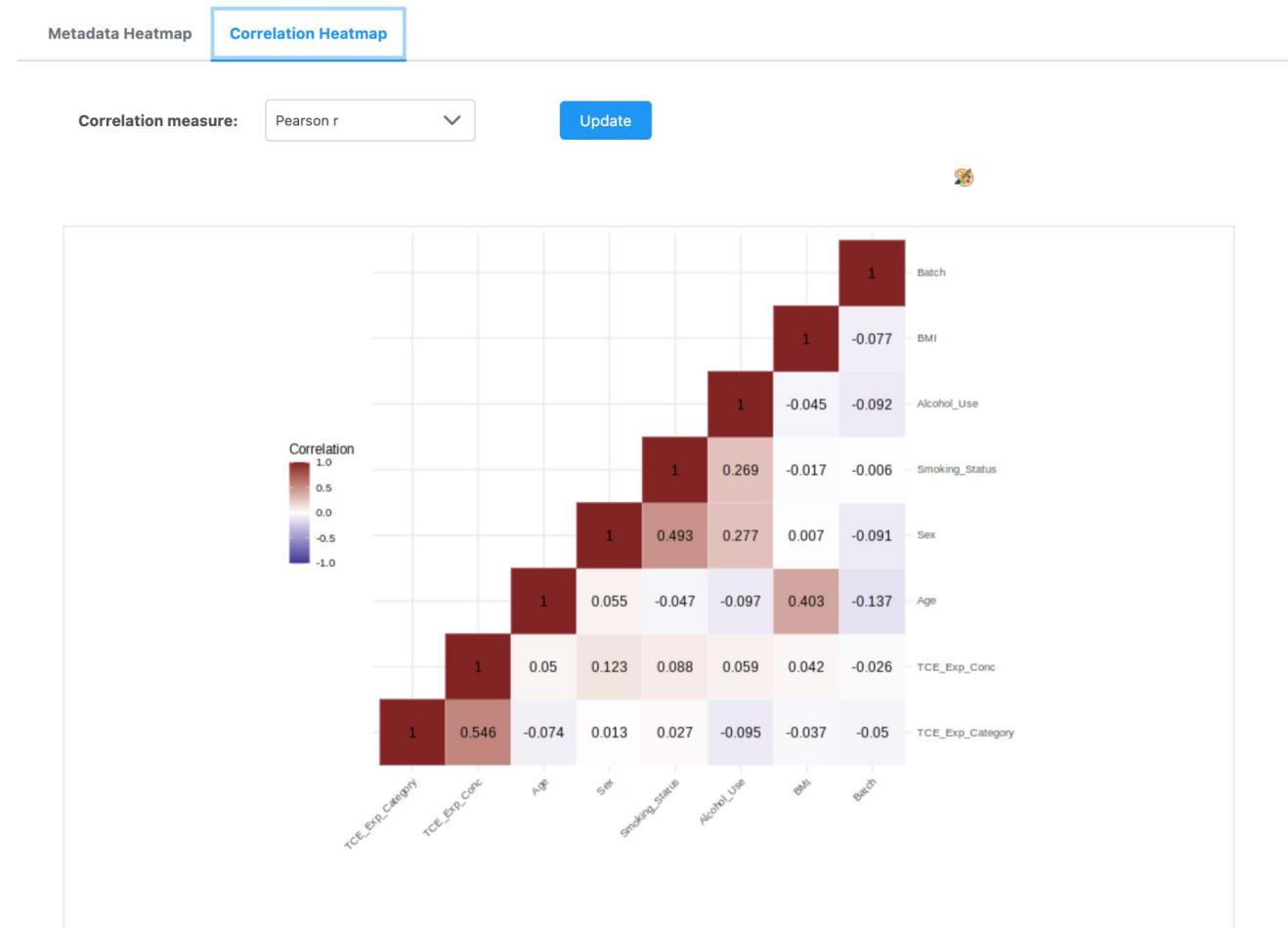
- Data and Metadata Overview**
  - [Metadata Visualization](#)  
Users can explore the metadata patterns and correlations through intuitive graphics. It is very useful for users to identify highly dependent metadata and quickly assess the overall patterns of the metadata.
  - [Interactive PCA Visualization](#)  
Users can visualize data using different colors or shapes based on selected metadata in an 2D and 3D (interactive) PCA plots. It is very useful to detect overall patterns of data with regard to different metadata.
  - [Hierarchical Clustering and Heatmap Visualization](#)  
This method displays data and metadata in the form of colored cells. It provides direct visualization of feature abundances across different samples and metadata.
- Univariate Analysis**
  - [Linear Models with Covariate Adjustment](#)  
This approach uses linear models (limma or lm) to perform significance testing with covariate adjustments. Users can choose different metadata to be included in the analysis.
  - [Correlation and Partial Correlation Analysis](#)  
This approach allows users to explore the correlations or partial correlations (with covariate adjustments) between metabolomics features and different metadata of interest.
  - [Multi-factor ANOVA](#)  
Depending on experimental design, either two-way ANOVA (multiple factors/covariates), two-way repeated ANOVA or two-way mixed ANOVA (Time-series + one factor; depends whether factor is within or between subjects), or one-way repeated ANOVA (Time-series only).
- Multivariate Analysis**
  - [ANOVA Simultaneous Component Analysis \(ASCA\)](#)  
This approach is designed to identify major patterns with regard to the two given factors and their interaction. The implementation was based on the algorithm described by [AK Smilde, et al.](#) with additional improvements on feature selection and model validation.
  - [Multivariate Empirical Bayes Analysis of Variance \(MEBA\) for Time Series](#)  
This approach is designed to compare temporal profiles across different biological conditions. It is based on the timecourse

# Metadata Visualization

This shows us which metadata variables are related. **TCE\_Exp\_Conc** and **TCE\_Exp\_Category** are highly correlated. **Age** and **BMI** are related, as are **Smoking\_Status**, **Alcohol\_Use**, and **Sex**. Navigate back and click "Interactive PCA".

## Metadata Overview

The metadata heatmaps displays patterns and correlations between different meta-data. Pay attention to those metadata that are highly correlated (i.e. collinear) or redundant metadata. It is advised to keep only one of the correlated meta-data, as more variables will lead to reduced statistical power and difficulties in interpretations.



# PCA

## Principal Component Analysis (PCA)

Color based on:

Batch

Update

Shape based on:

Sex

Pairwise Score Plots

Synchronized 3D Plots

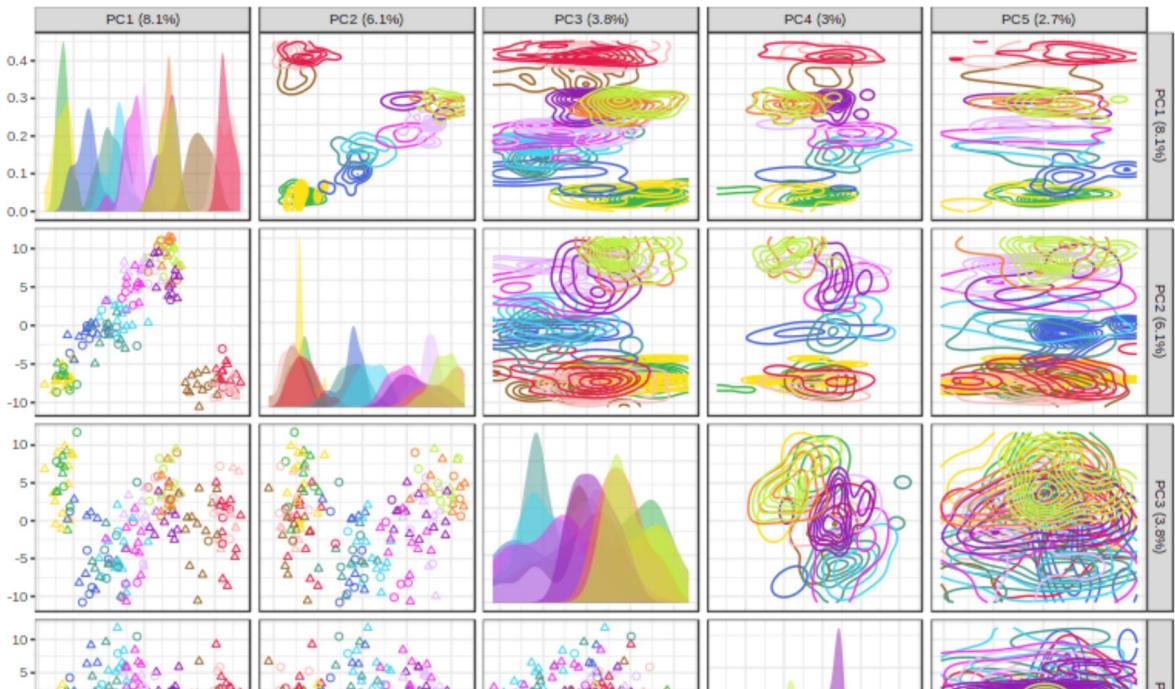
Display score plots for top

5

PCs

Update

Change the “Color based on” parameter and update the view to annotate the sample scores plot with different metadata. Notice that “**Batch**” explains many patterns in the data. Navigate back and select “Linear Models with Covariate Adjustment”.



# Linear Models

1 - Remember the relationships between metadata in the "Metadata Visualization" part? We don't want to include multiple correlated variables in our analysis. Put "TCE\_Exp\_Conc" as the primary metadata, add "Age", "Sex", and "Batch" as covariates, and click "Submit".

## Linear models with covariate adjustments

The underlying method is based on [limma](#) for its high-performance implementation. Some data may include some form of blocking in the study design, which can be modeled as either fixed or random effects. Please note that although you can specify a blocking factor (to be modeled as random effects), we in general recommend **keeping this option unspecified** (the default). Using fixed effect model not only is computationally more efficient, but also gives results that are more consistent with the interpretation of differences. Please refer to the excellent book by [Paul D. Allison \(2009\)](#) for more technical discussions.

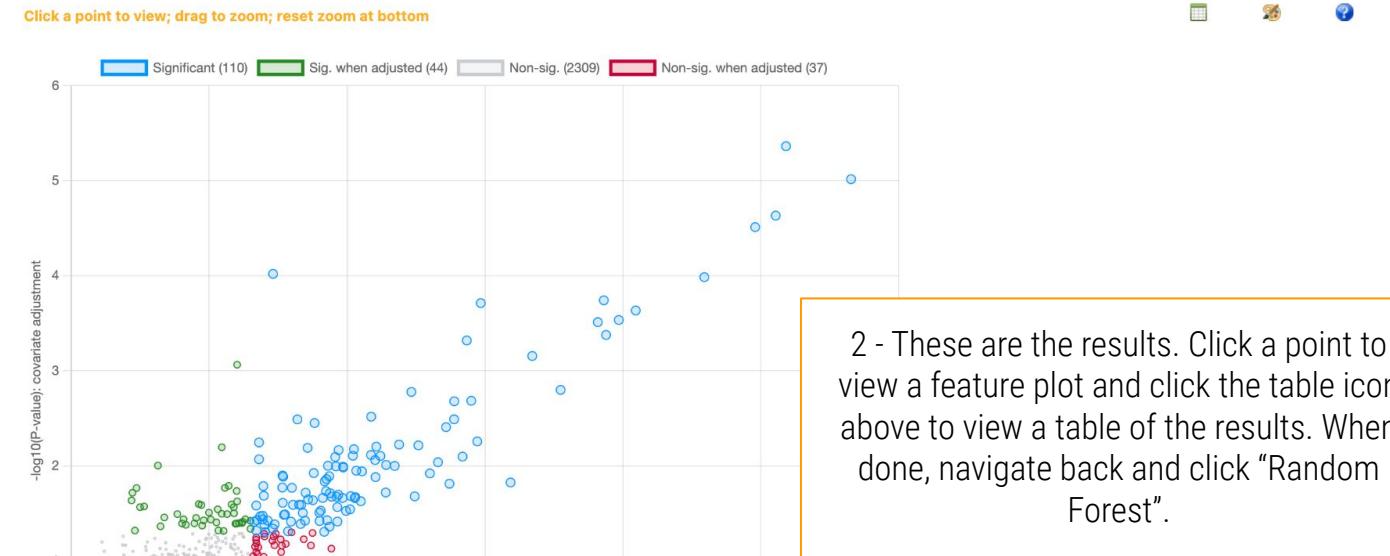
Primary metadata: TCE\_Exp\_Conc

Covariates (control for): Age X Sex X Batch X

Blocking factor: -- Unspecified --

P-value cutoff: 0.05

Submit



2 - These are the results. Click a point to view a feature plot and click the table icon above to view a table of the results. When done, navigate back and click "Random Forest".

# Random Forest

1 - Leave "TCE\_Exp\_Category" as the primary metadata and choose "Age", "Sex", and "Batch" as metadata predictors. Click "Submit".

## Random Forest

Classification Var. Importance Outlier Detection

Primary metadata:

TCE\_Exp\_Category

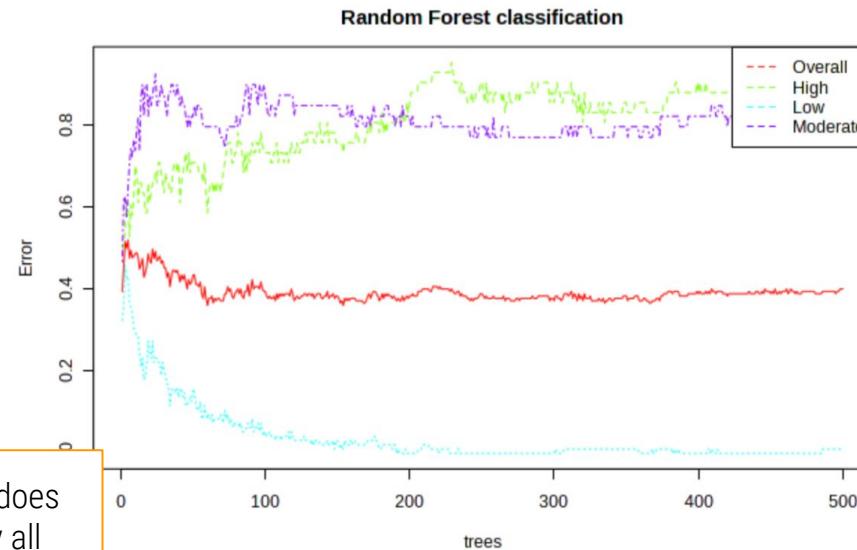
Choose metadata for predictors:

Age X Sex X Batch X

Update

Randomness:

On



The OOB error is 0.4

	High	Low	Moderate	class.error
High	4	35	2.0	0.902
Low	0	94	1.0	0.0105
Moderate	1	31	7.0	0.821

2 - These are the results. The model does not perform well. It classifies nearly all samples as "Low".

# We are on a Coffee Break & Networking Session

## Workshop Sponsors:



Canadian Centre for  
Computational  
Genomics



HPC4Health



OICR  
Ontario Institute  
for Cancer Research



GenomeCanada

