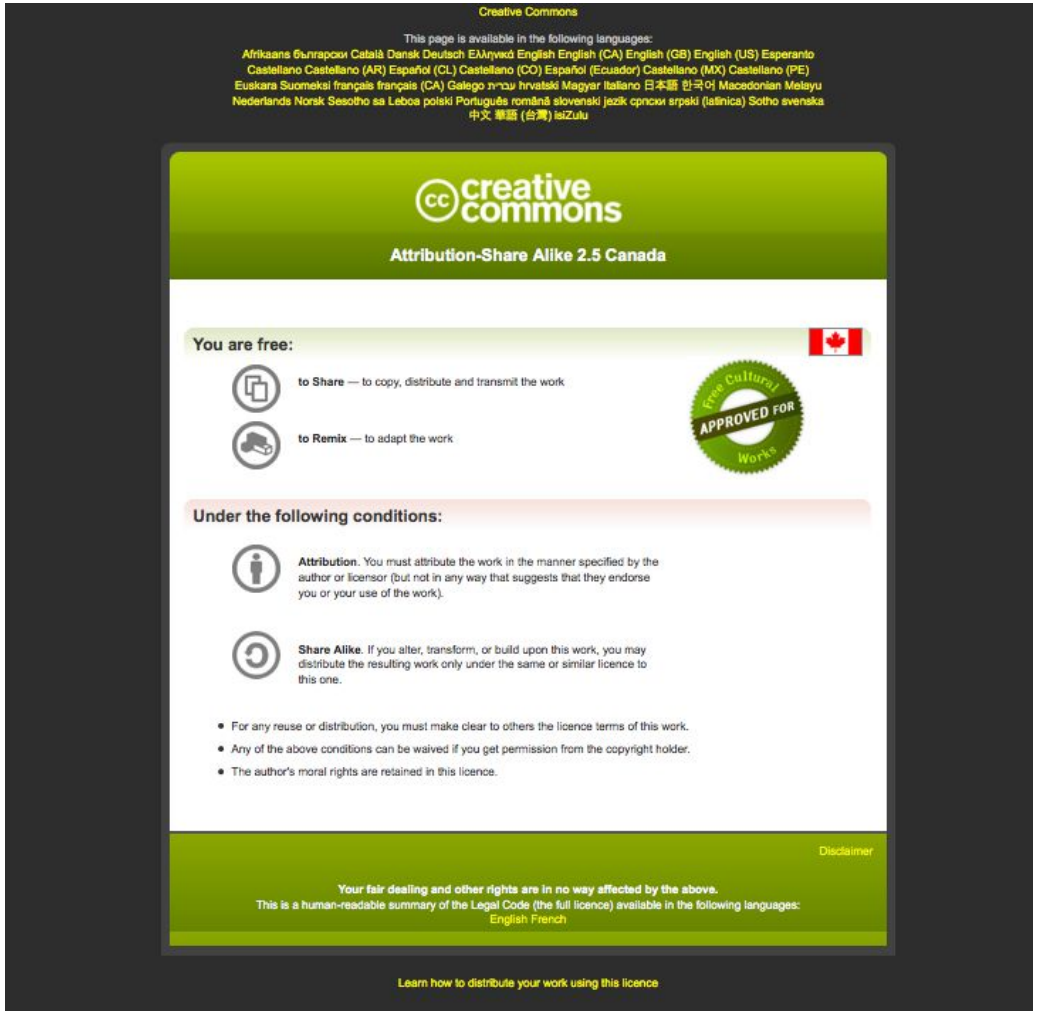
bioinformatics.ca

Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io


Creative Commons
This page is available in the following languages:
Afrikaans български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto
Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
Euskara Suomeksi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu
Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska
中文 華語 (台灣) isiZulu
creative commons
Attribution-Share Alike 2.5 Canada
You are free:
to Share — to copy, distribute and transmit the work
to Remix — to adapt the work
Free Cultural
APPROVED FOR
Works
Under the following conditions:
Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.
For any reuse or distribution, you must make clear to others the licence terms of this work.
Any of the above conditions can be waived if you get permission from the copyright holder.
The author's moral rights are retained in this licence.
Disclaimer
Your fair dealing and other rights are in no way affected by the above.
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
English French
Learn how to distribute your work using this licence

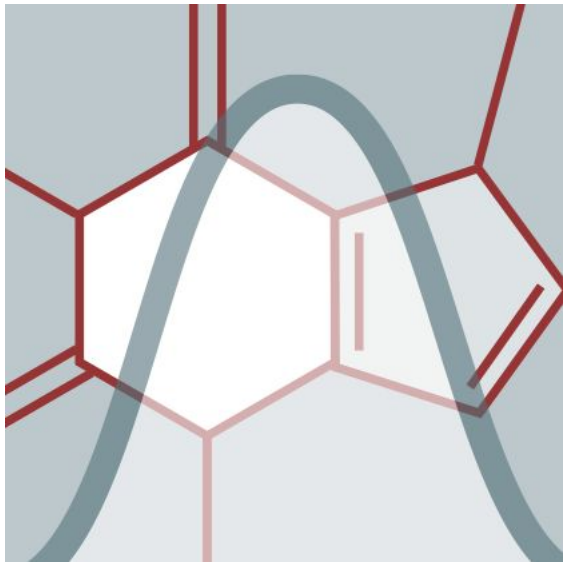
Metabolomics Data Analysis (Lab)

Jeff Xia

Informatics and Statistics for Metabolomics

July 6-7, 2023

bioinformatics.ca



McGill

TMIC

The Metabolomics Innovation Centre

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

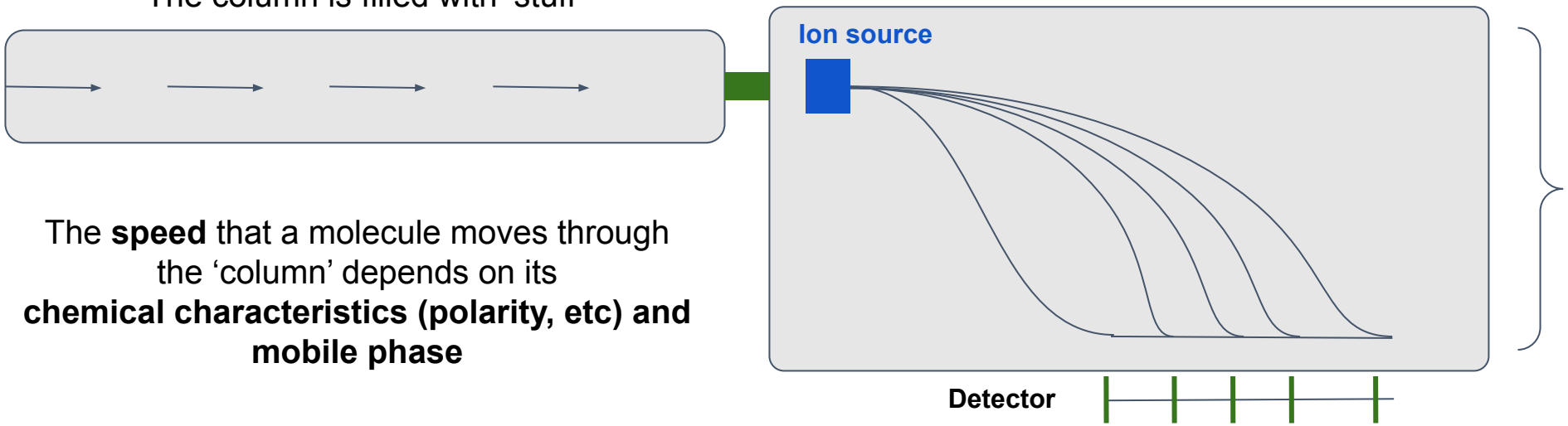
1. To practice raw LC-MS metabolomics data processing;
2. To practice functional analysis on LC-MS global metabolomics data;
3. To practice how to use MetaboAnalyst 5.0 to facilitate one-factor statistical analysis;
4. To practice metadata based statistical analysis on global metabolomics data sets;

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’



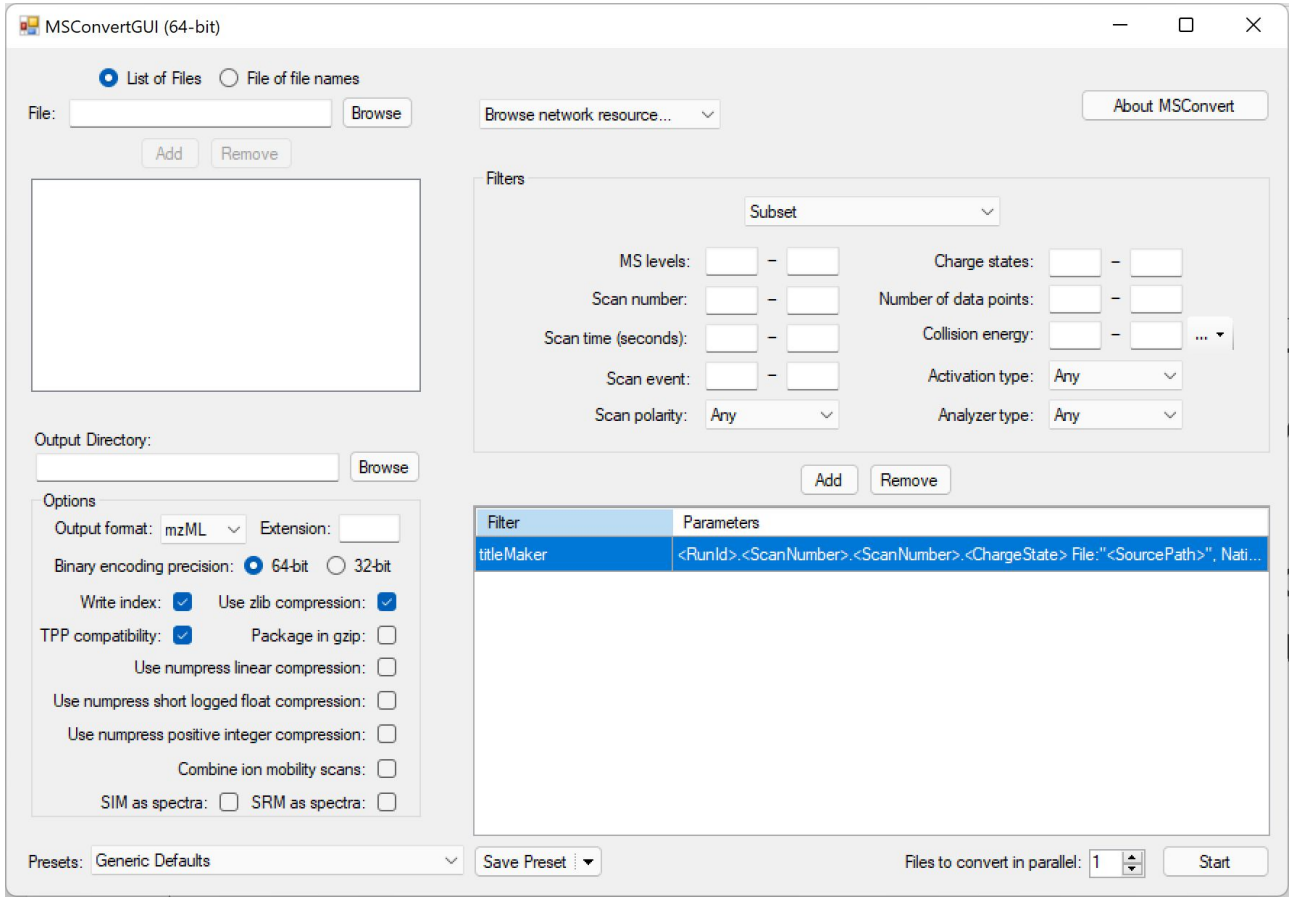
The **speed** that a molecule moves through the ‘column’ depends on its **chemical characteristics (polarity, etc) and mobile phase**

The **path** of the ionized molecule depends on its **mass:charge ratio** (*m/z*)

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

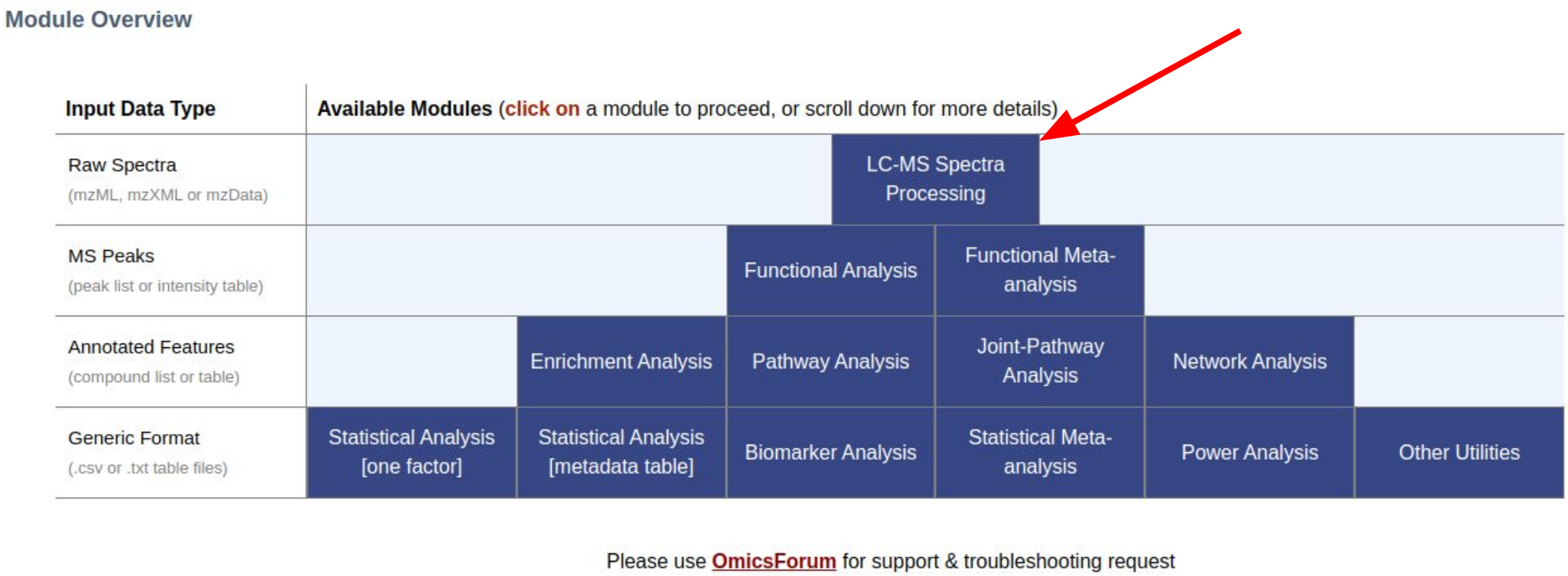
Use Centroid and open-source data format

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

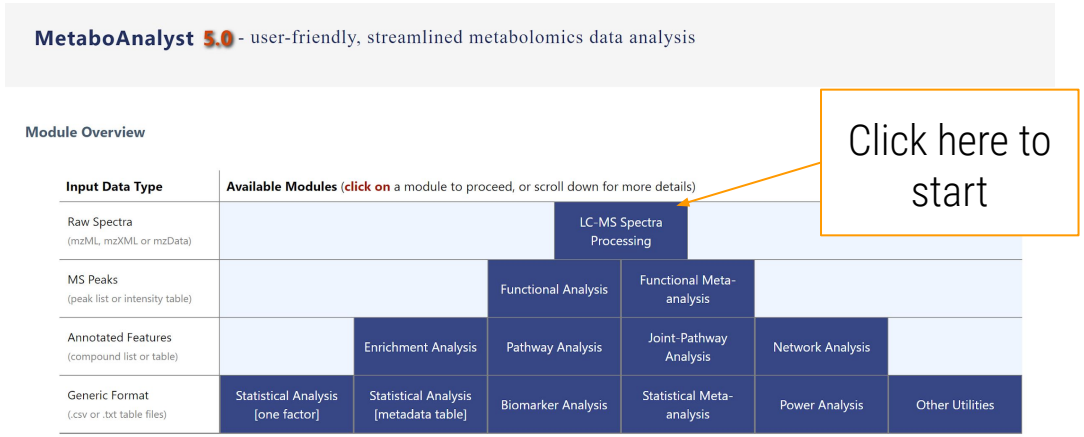


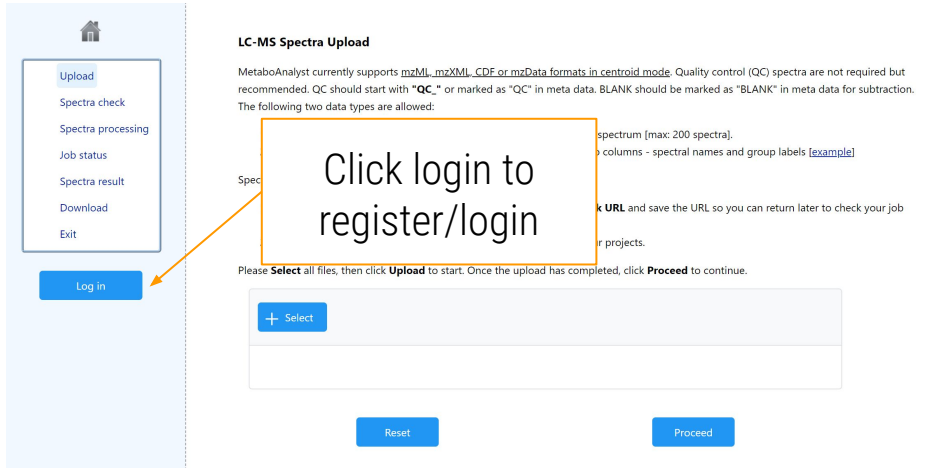
MetaboAnalyst: LC-MS Spectral Processing Module

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

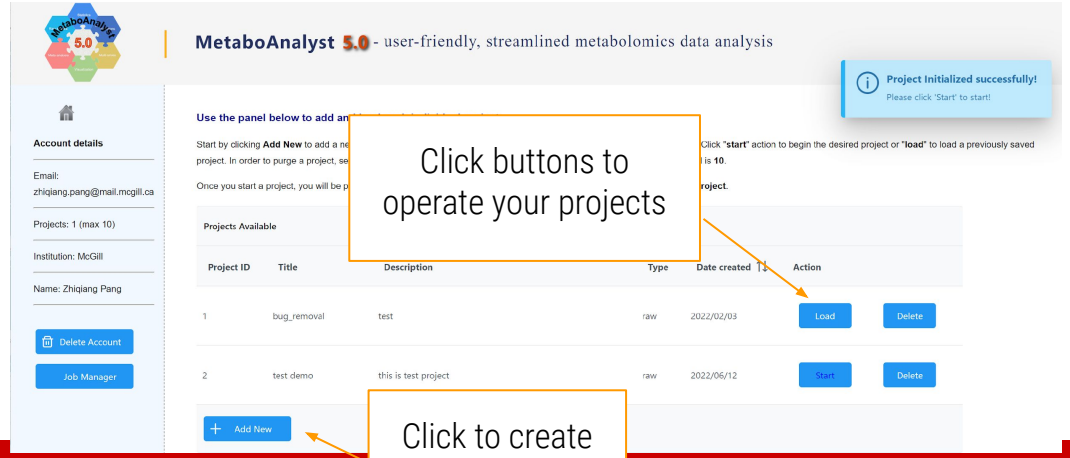


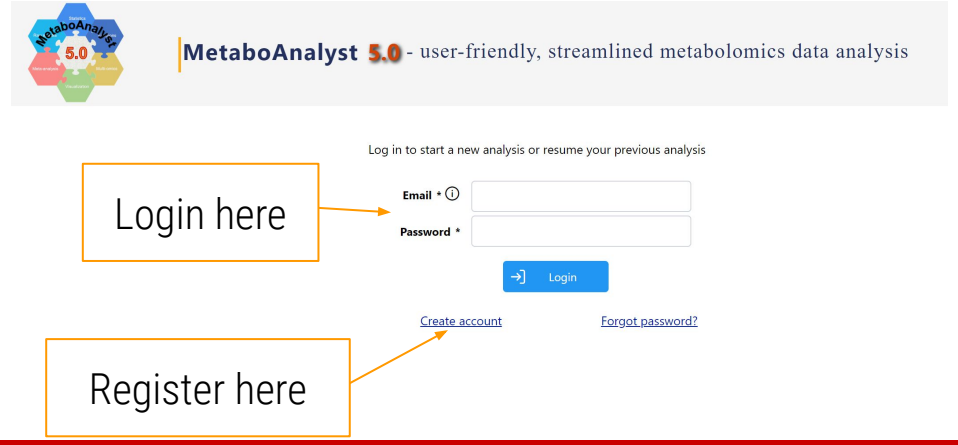
User registration and log-in (Optional)





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





Data Uploading

LC-MS Spectra Upload

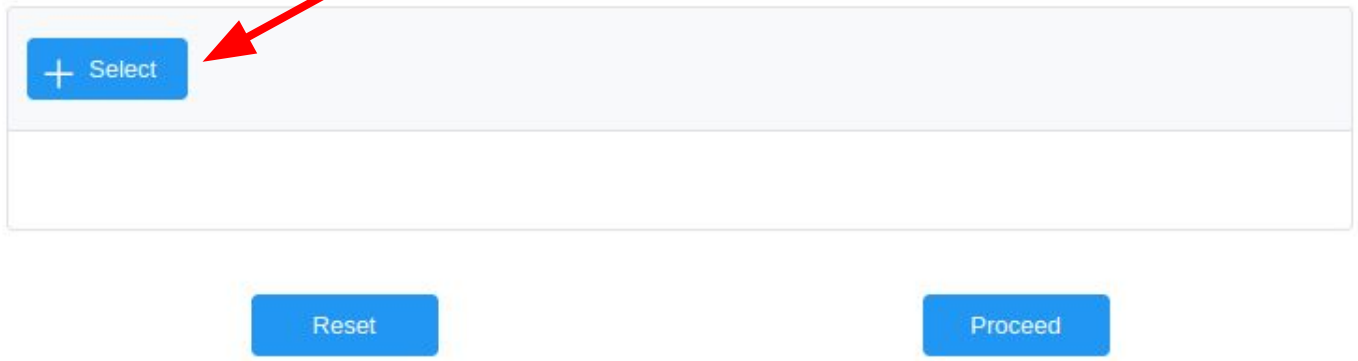
MetaboAnalyst currently supports mzML, mzXML, CDF or mzData formats in centroid mode. Quality control (QC) spectra are not required but recommended. QC should start with **"QC_"** or marked as "QC" in meta data. BLANK should be marked as "BLANK" in meta data for subtraction. The following two data types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

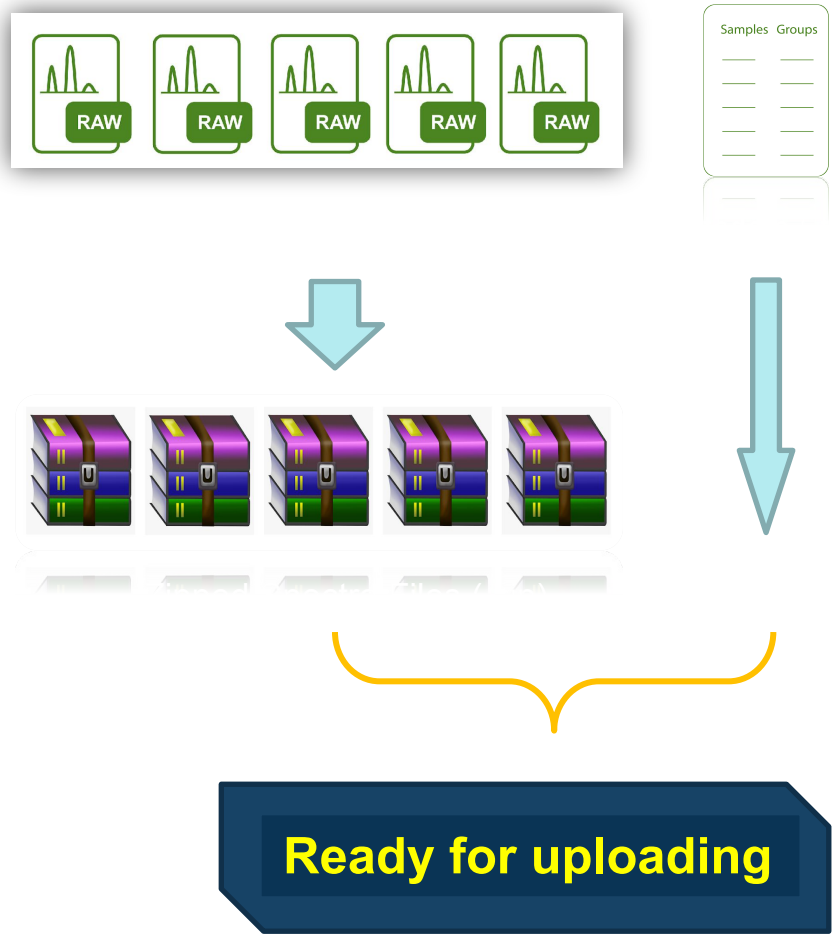
Spectra processing can take a long time to complete, to avoid waiting:

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.

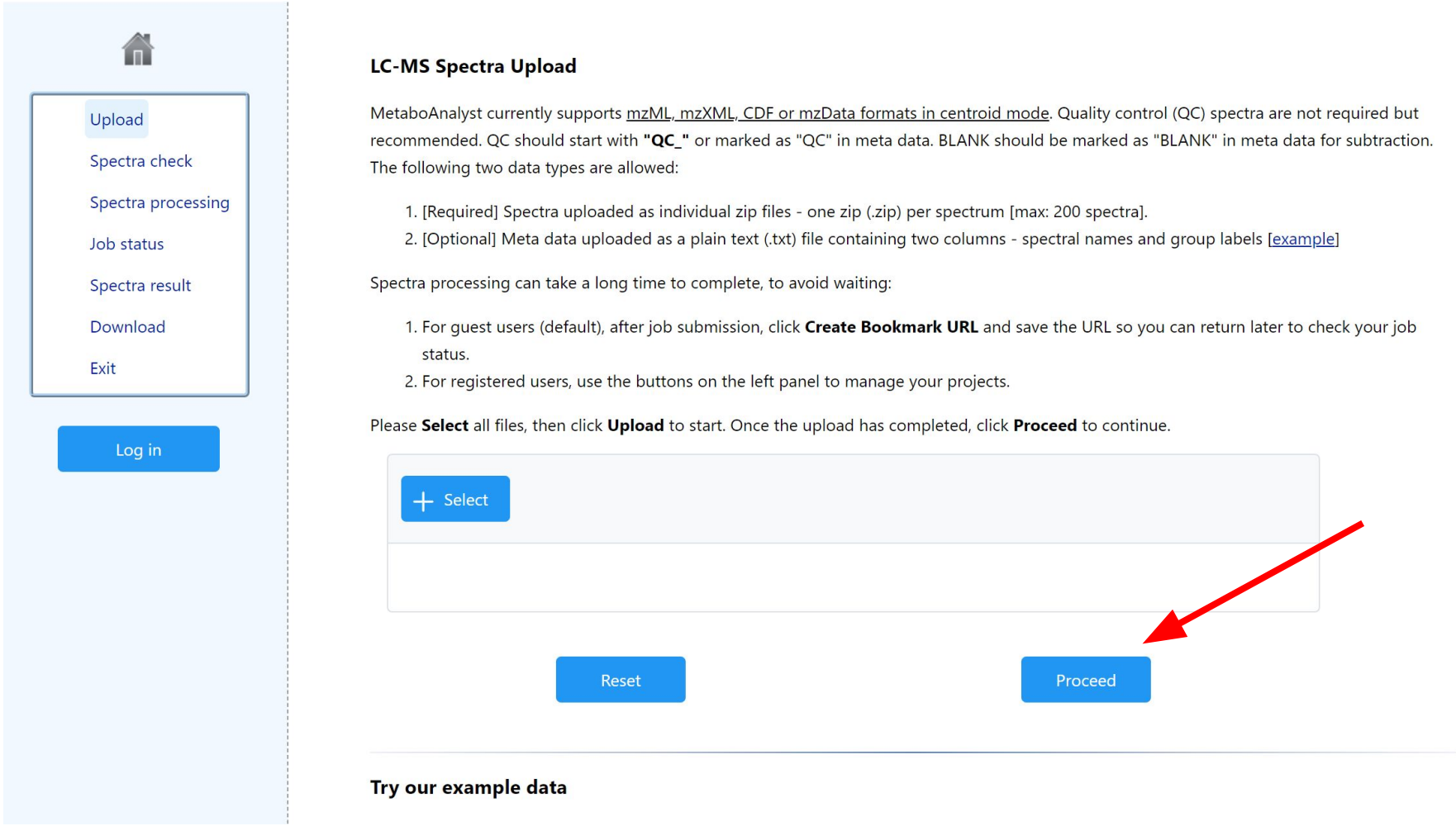


Raw data organization

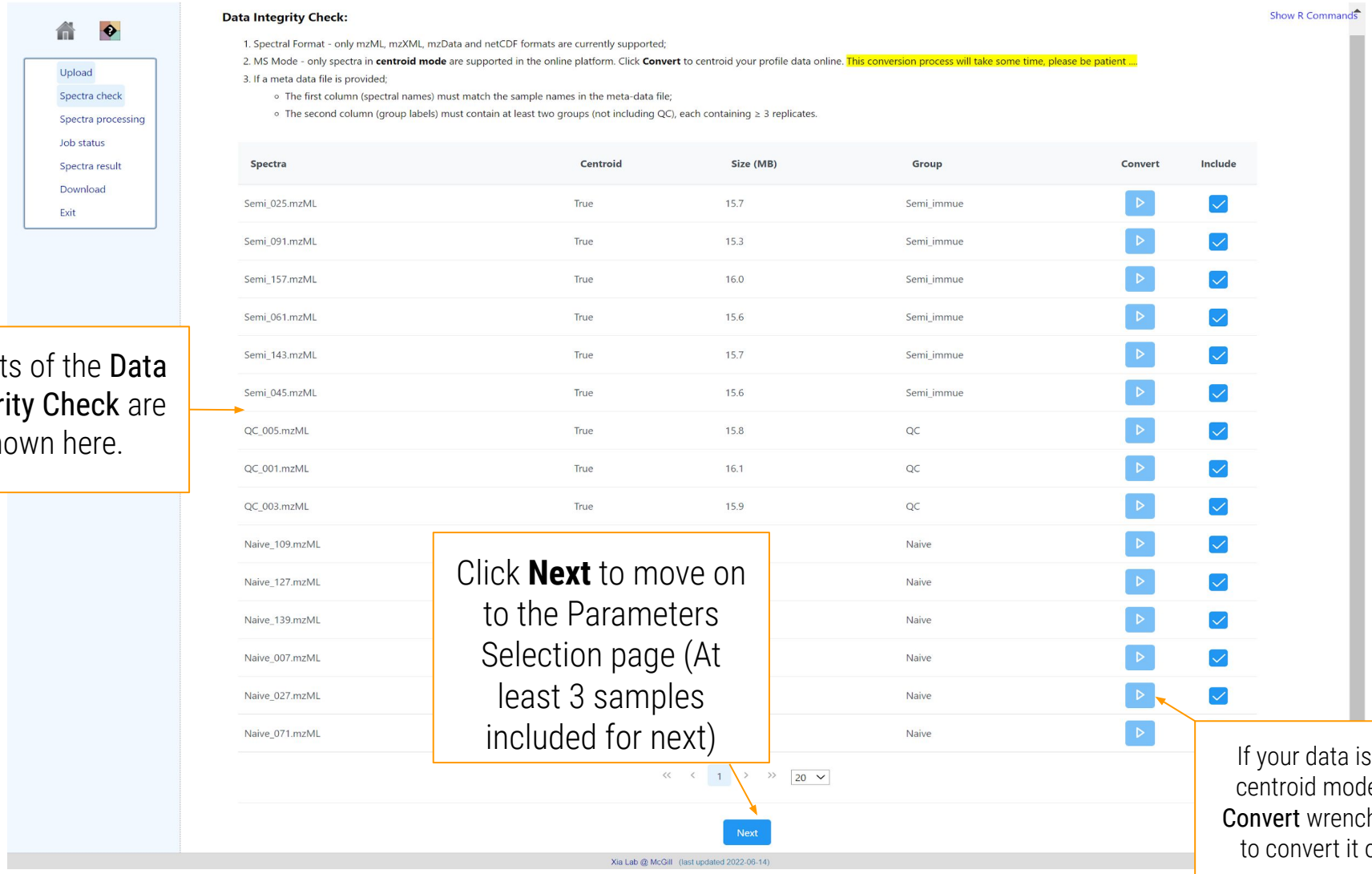


trimmed_metadata.txt	219 bytes	30 Jul 2020
CD-9WOBP.zip	1.9 MB	15 Jun 2020
CD-6KUCT.zip	2.0 MB	15 Jun 2020
CD-77FXR.zip	2.1 MB	15 Jun 2020
HC-AMR37.zip	2.1 MB	15 Jun 2020
HC-9X47O.zip	2.1 MB	15 Jun 2020
QC_PREFB02.zip	2.2 MB	15 Jun 2020
QC_PREFA02.zip	2.2 MB	15 Jun 2020
CD-9OS5Y.zip	2.2 MB	15 Jun 2020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

Submit to upload your data



Data Integrity Check



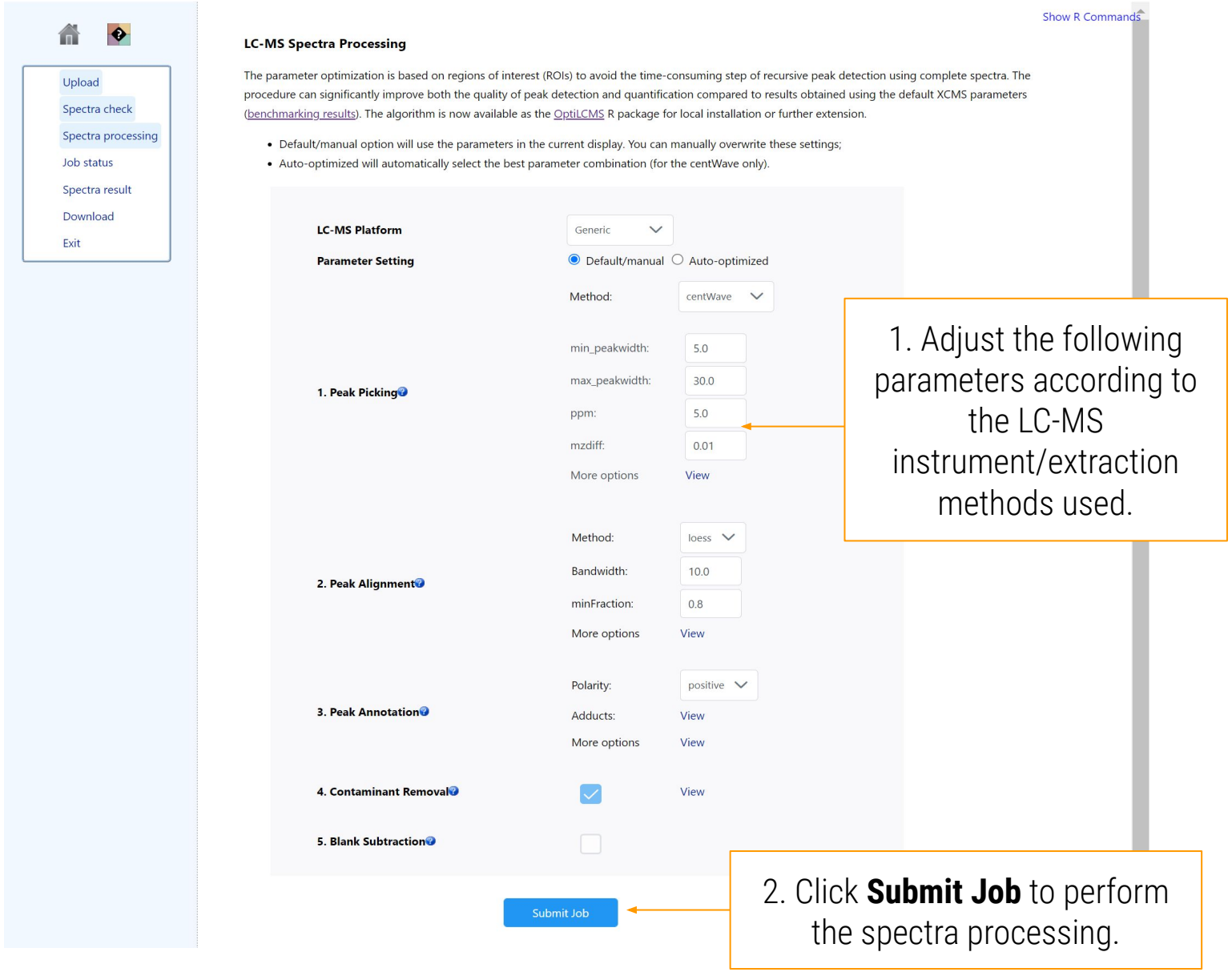
Results of the **Data Integrity Check** are shown here.

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

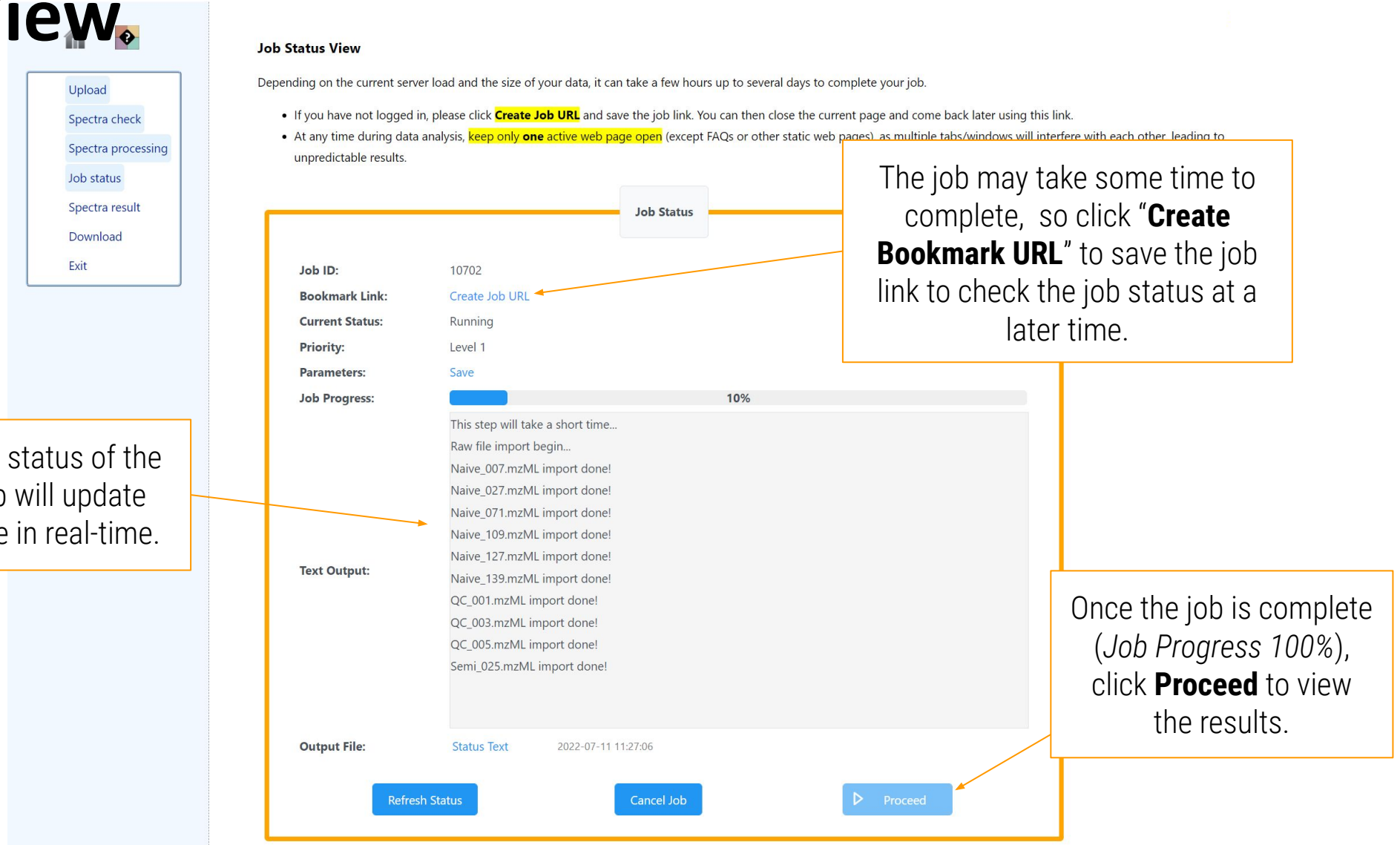
If your data is not in centroid mode, click **Convert** wrench button to convert it online.

Parameter Selection

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



Job Status View



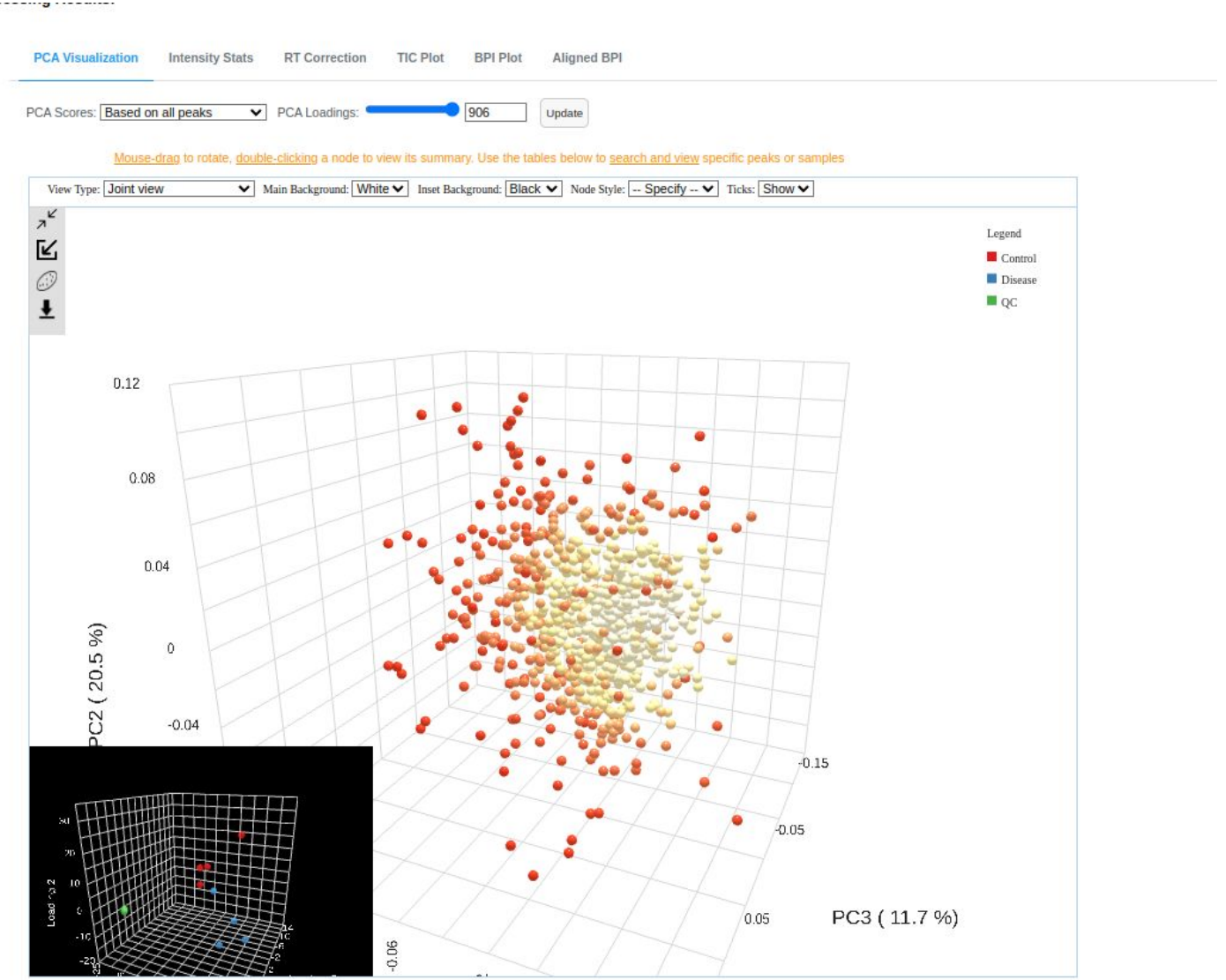
The job may take some time to complete, so click "**Create Bookmark URL**" to save the job link to check the job status at a later time.

The status of the job will update here in real-time.

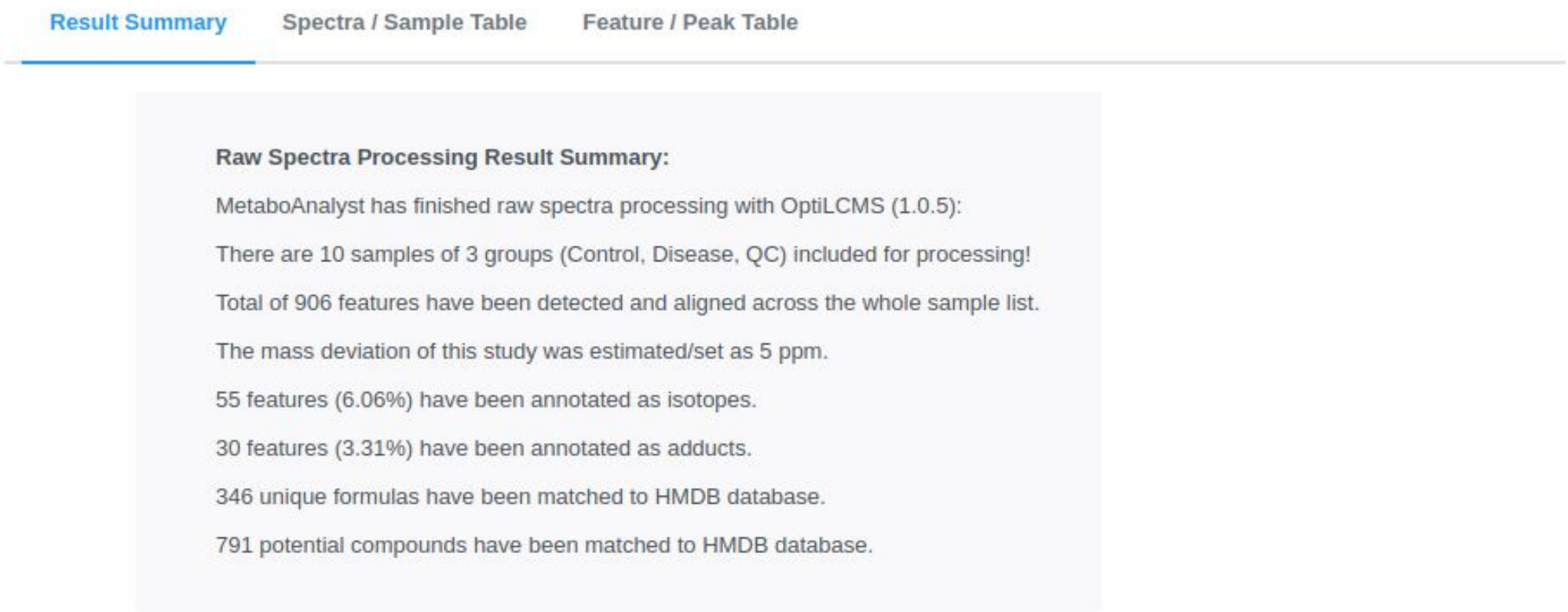
Once the job is complete (*Job Progress 100%*), click **Proceed** to view the results.

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



Exploring the Results -2

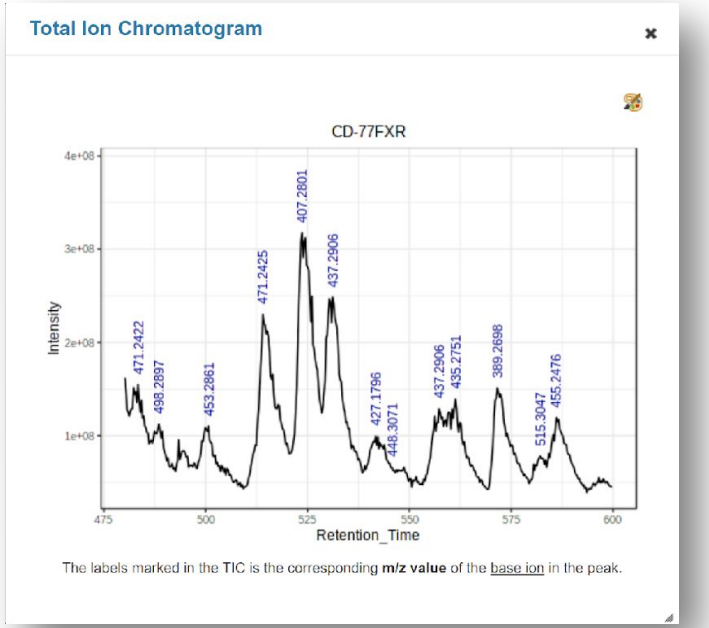
Result Summary | Spectra / Sample Table | Feature / Peak Table

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	

« ‹ 1 › » 20

Download Page

Click the ***View*** button to seeTIC of the corresponding spectra.



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 wiil 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			

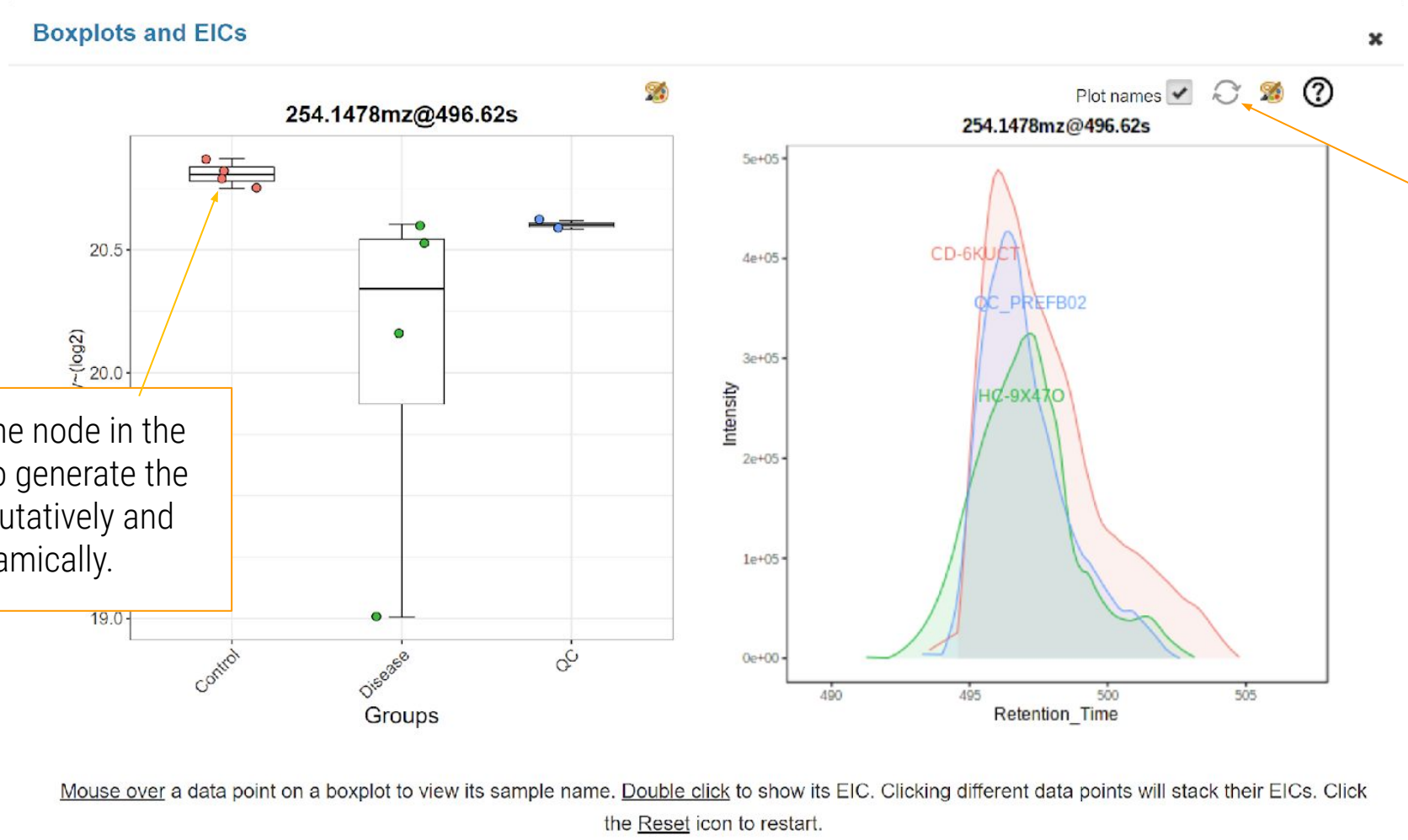
This table is showing all MS features. Click the button of Putative IDs show the potential Chemical IDs of the features towards HMDB.

Putative IDs

Formulas	Compounds
C34H64NO10P;	PS(14:0/14:1(9Z)); PS(14:1(9Z)/14:0);

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

Exploring the Results -4



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

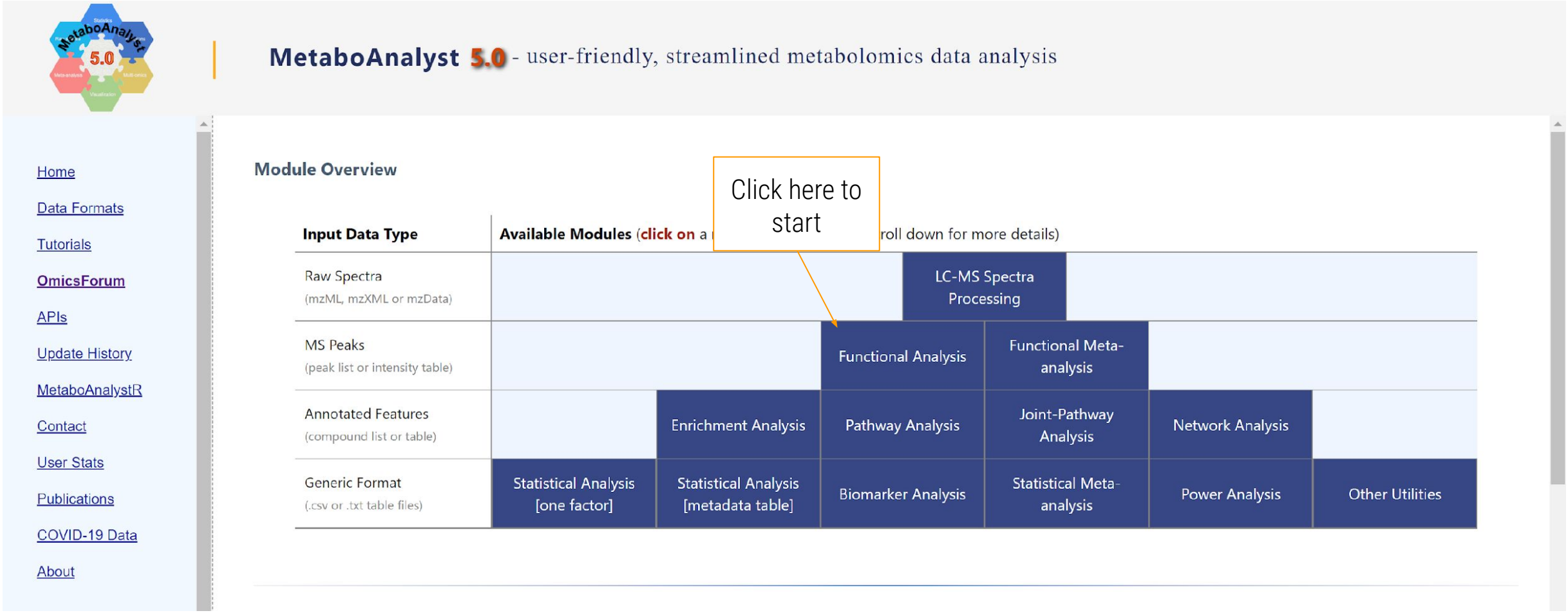
1. Click the node in the boxplot to generate the EIC cumutatively and dynamically.

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

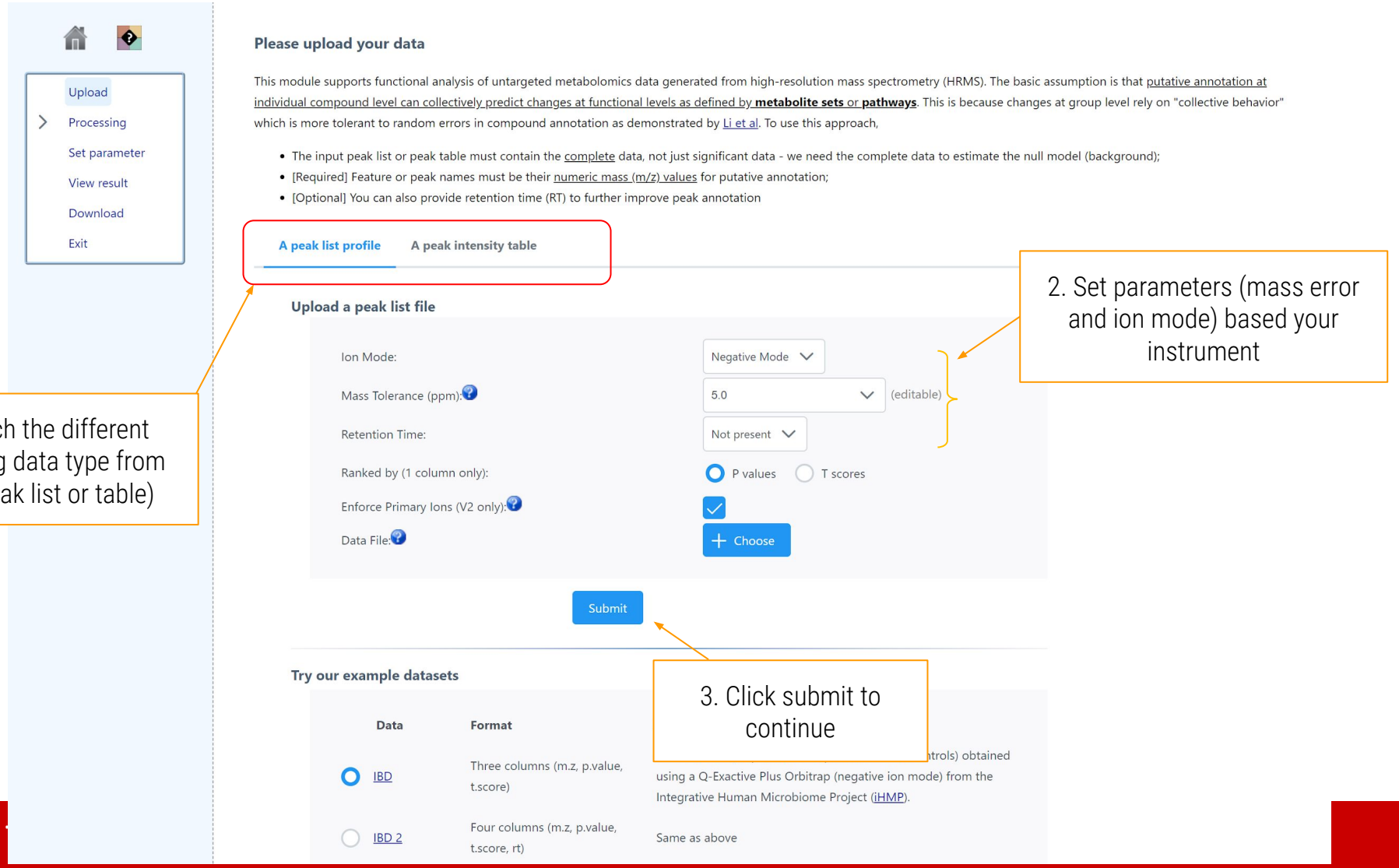
3. Scroll down to the bottom of page and click "**Proceed**" to view the Downloads page.

Functional Analysis of Untargeted Metabolomics Demos

Start Functional Analysis



Peak List Uploading

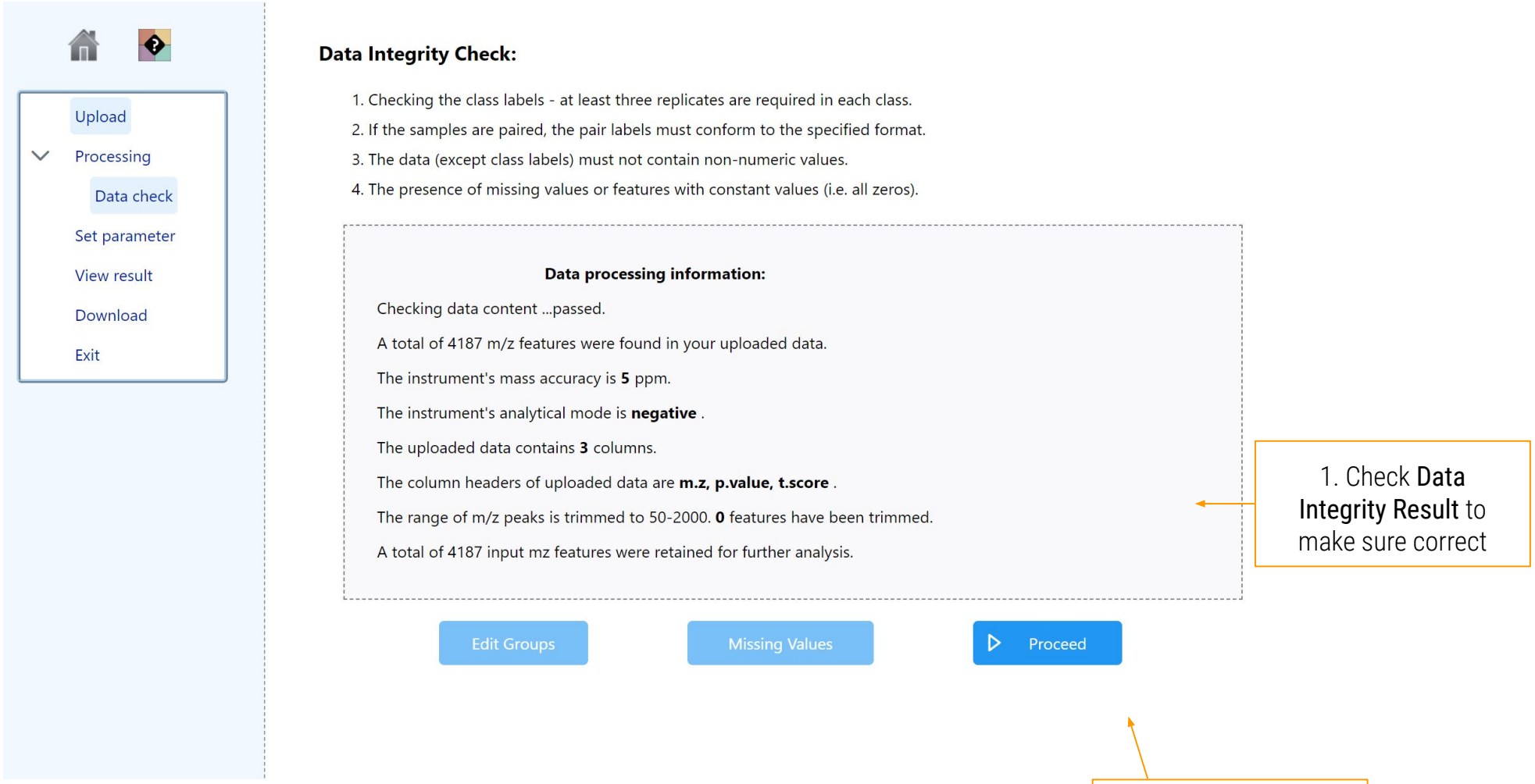


1. Switch the different uploading data type from here (peak list or table)

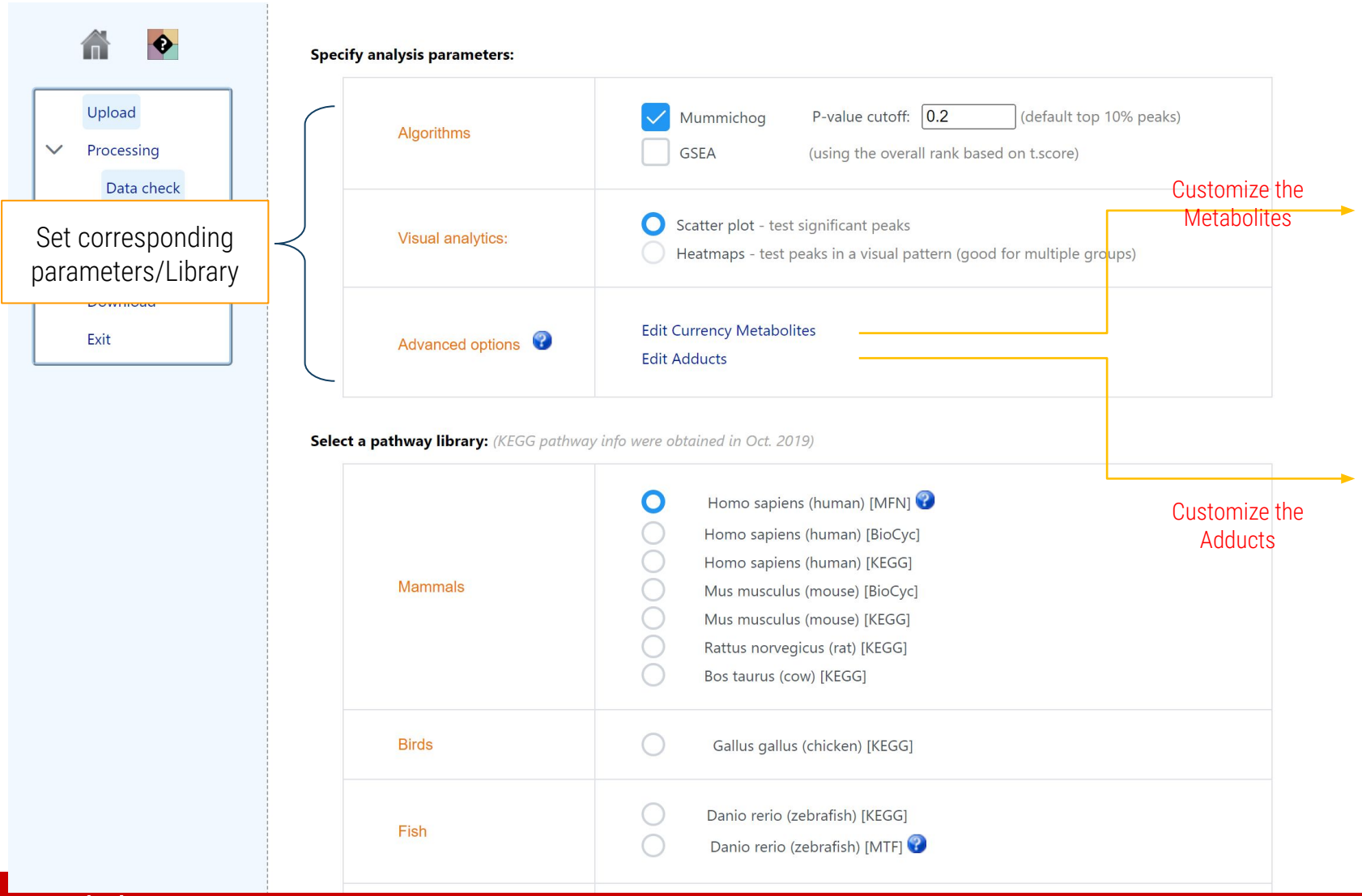
2. Set parameters (mass error and ion mode) based your instrument

3. Click submit to continue

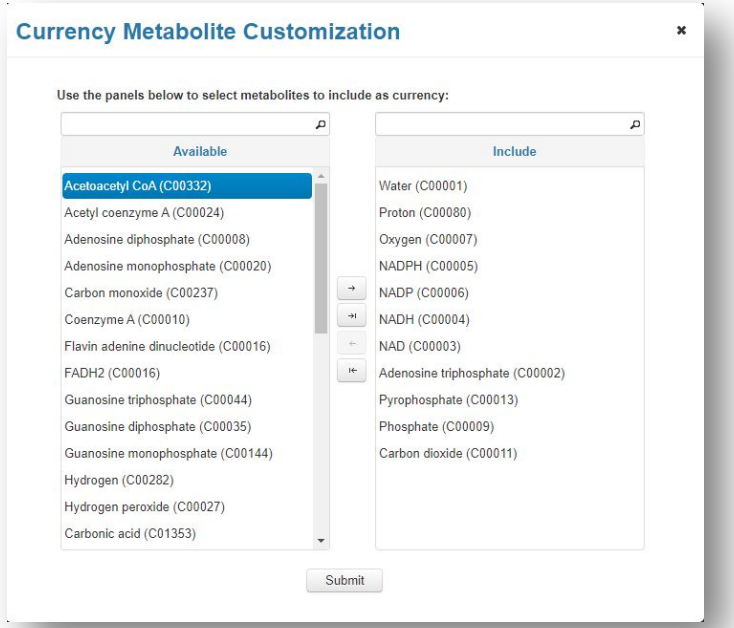
Data Integrity Check



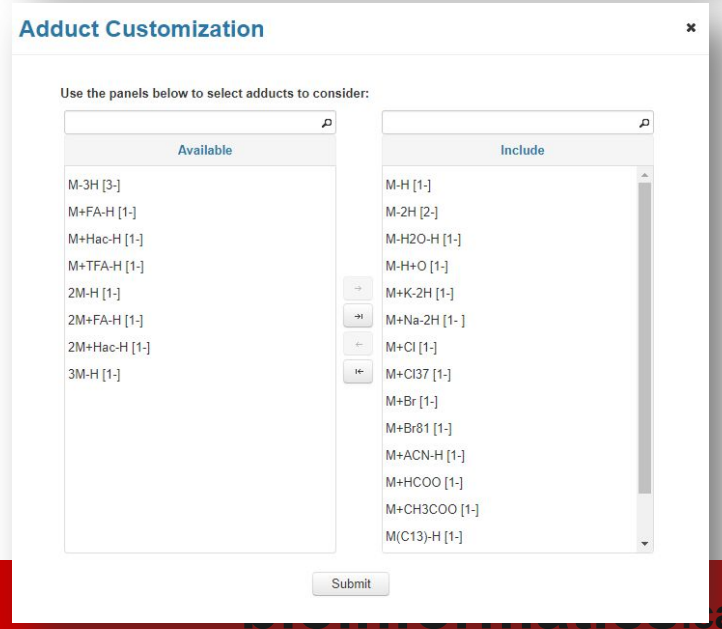
Set Parameters



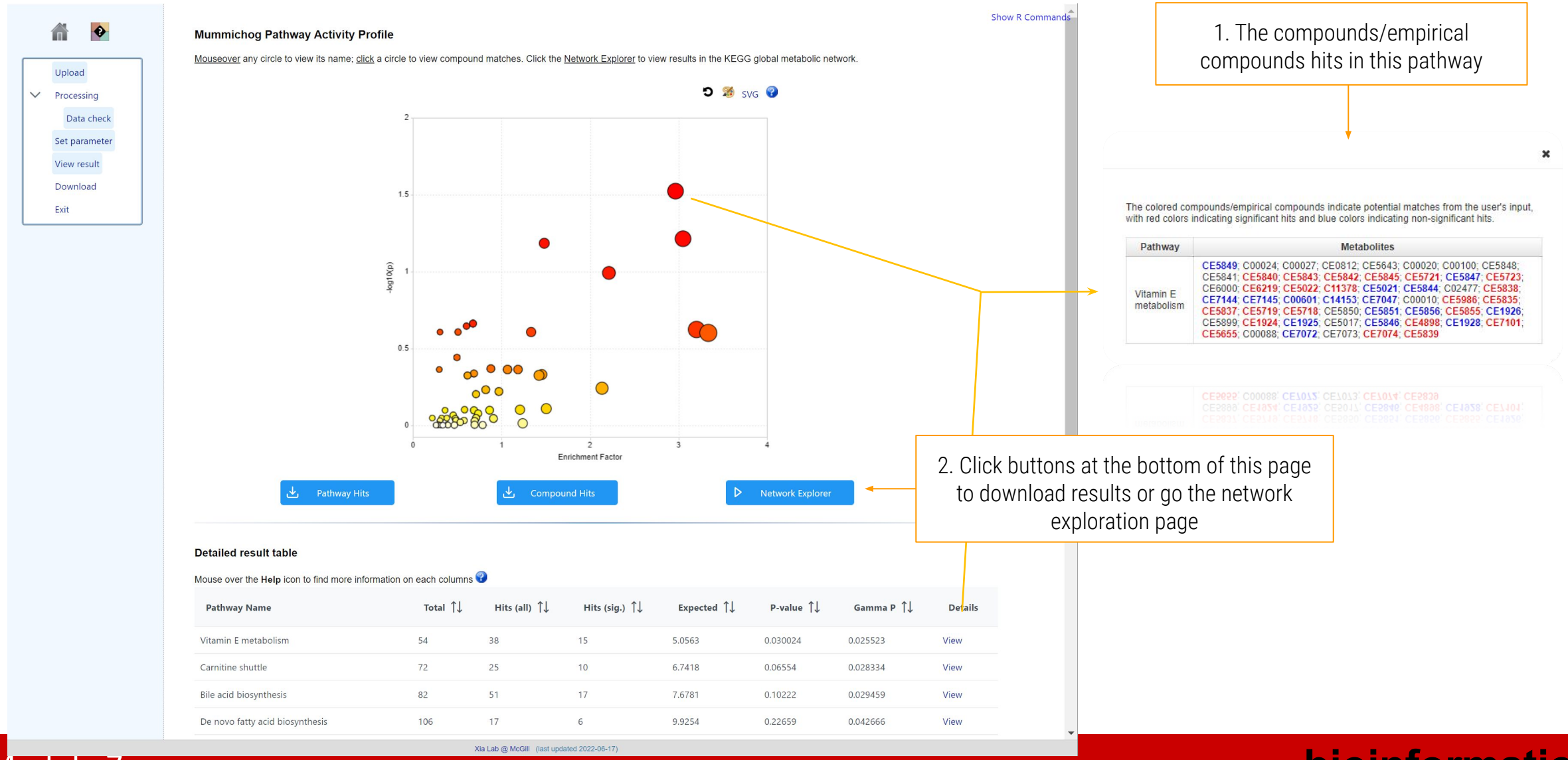
Customize the Metabolites



Customize the Adducts



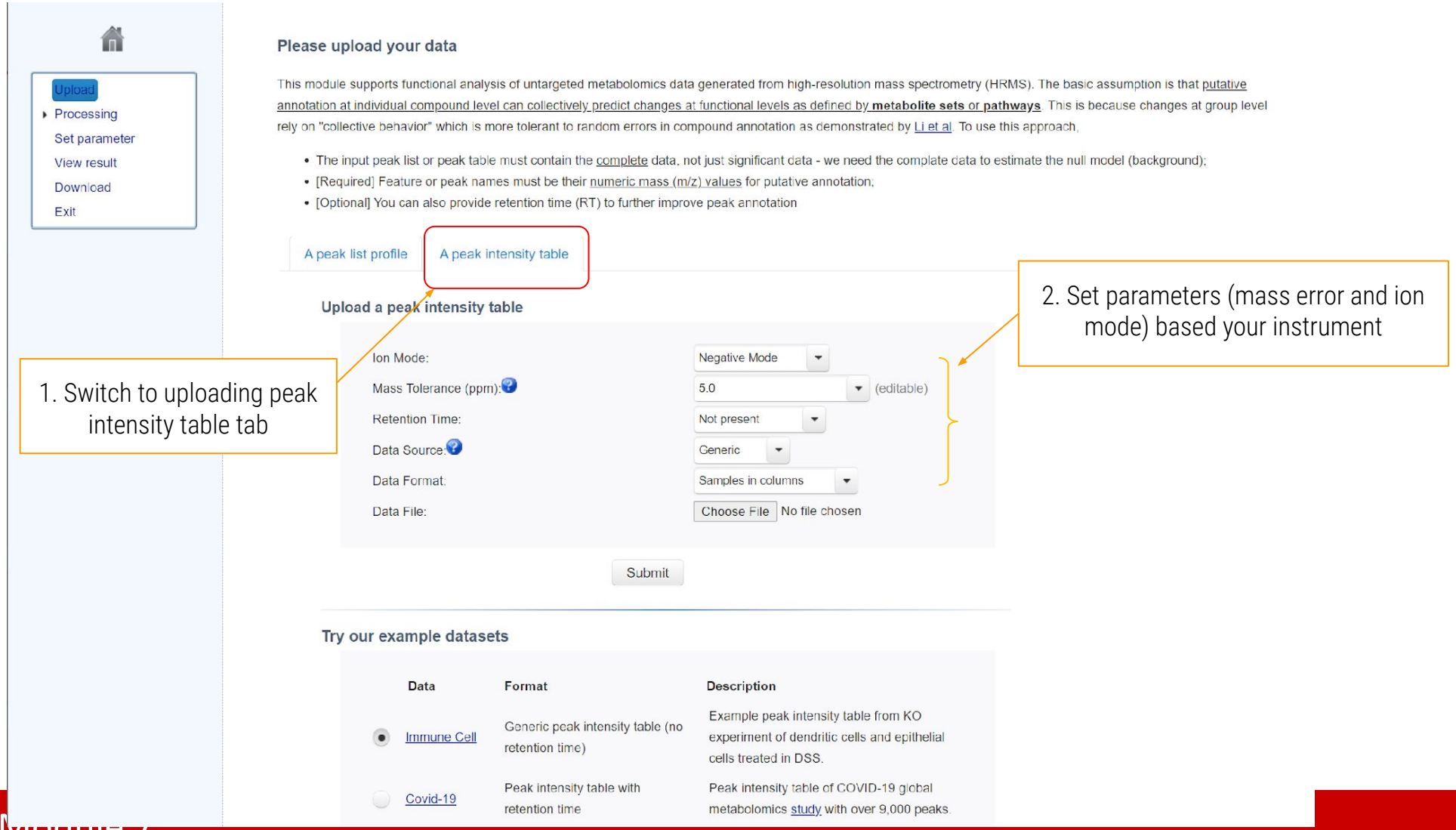
Pathway analysis results



1. The compounds/empirical compounds hits in this pathway

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

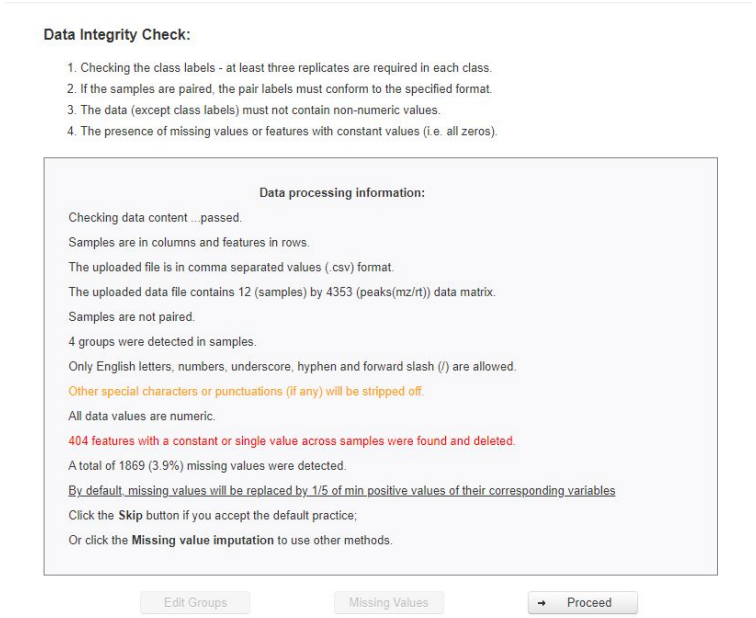
Peak uploading – peak table

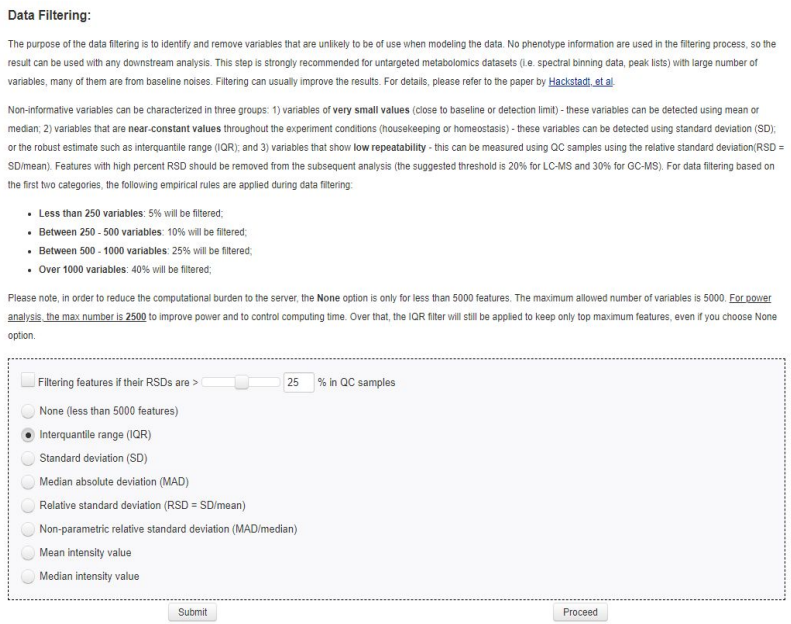


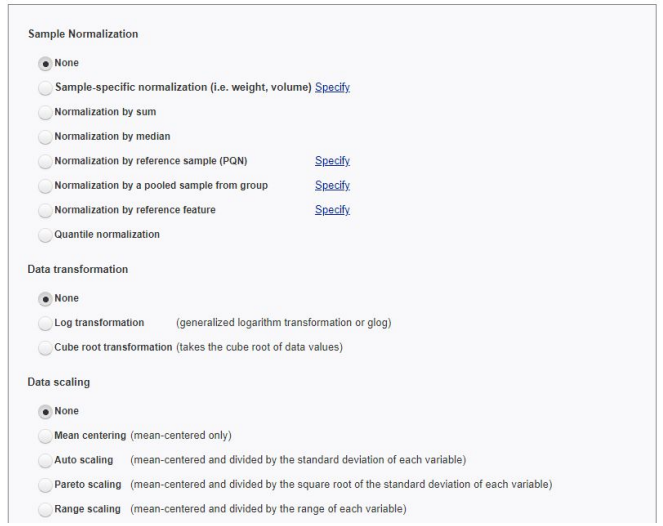
1. Switch to uploading peak intensity table tab

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

Peak uploading – Preprocessing





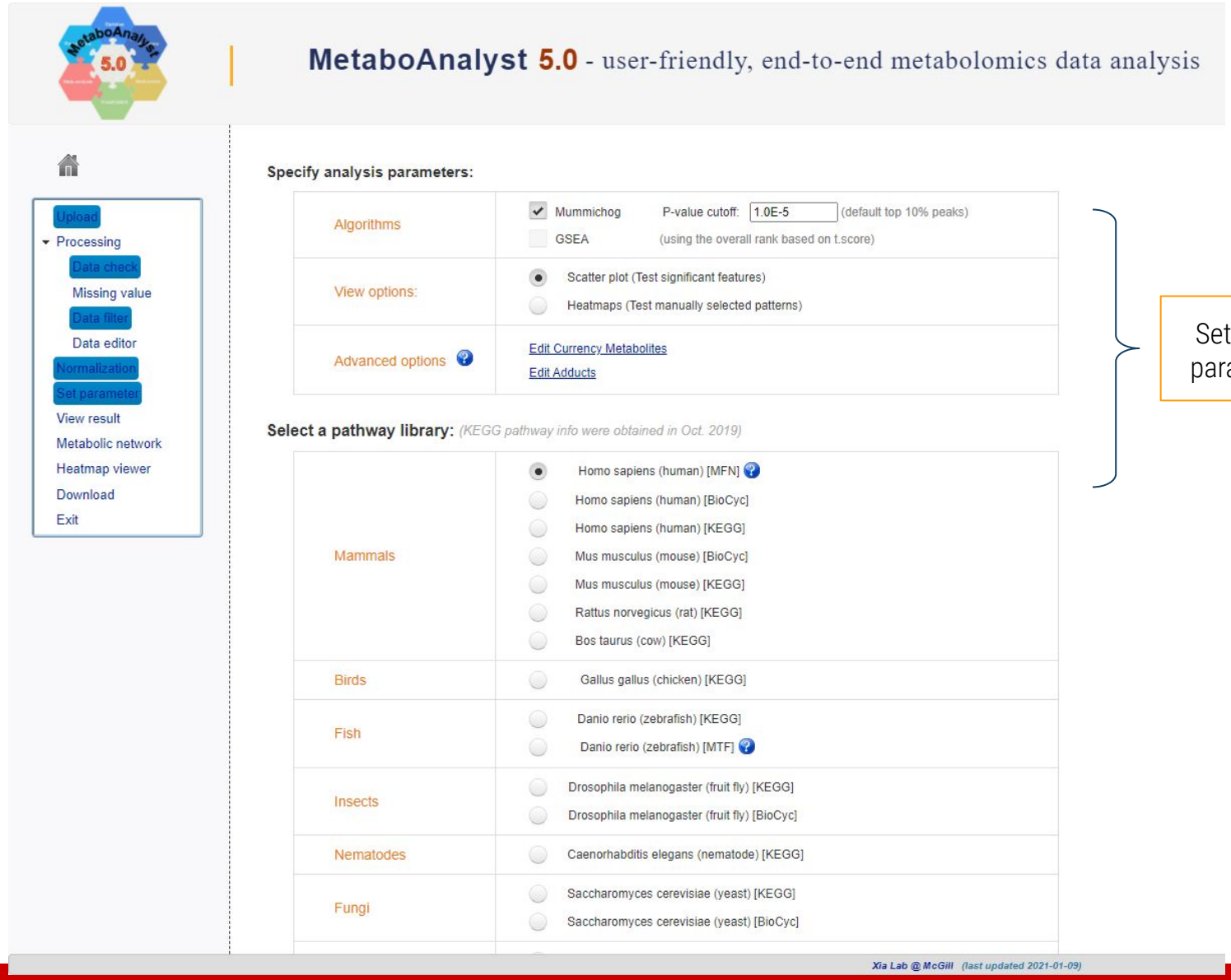


1. Perform Data Integrity Check

2. Perform Data Filtering

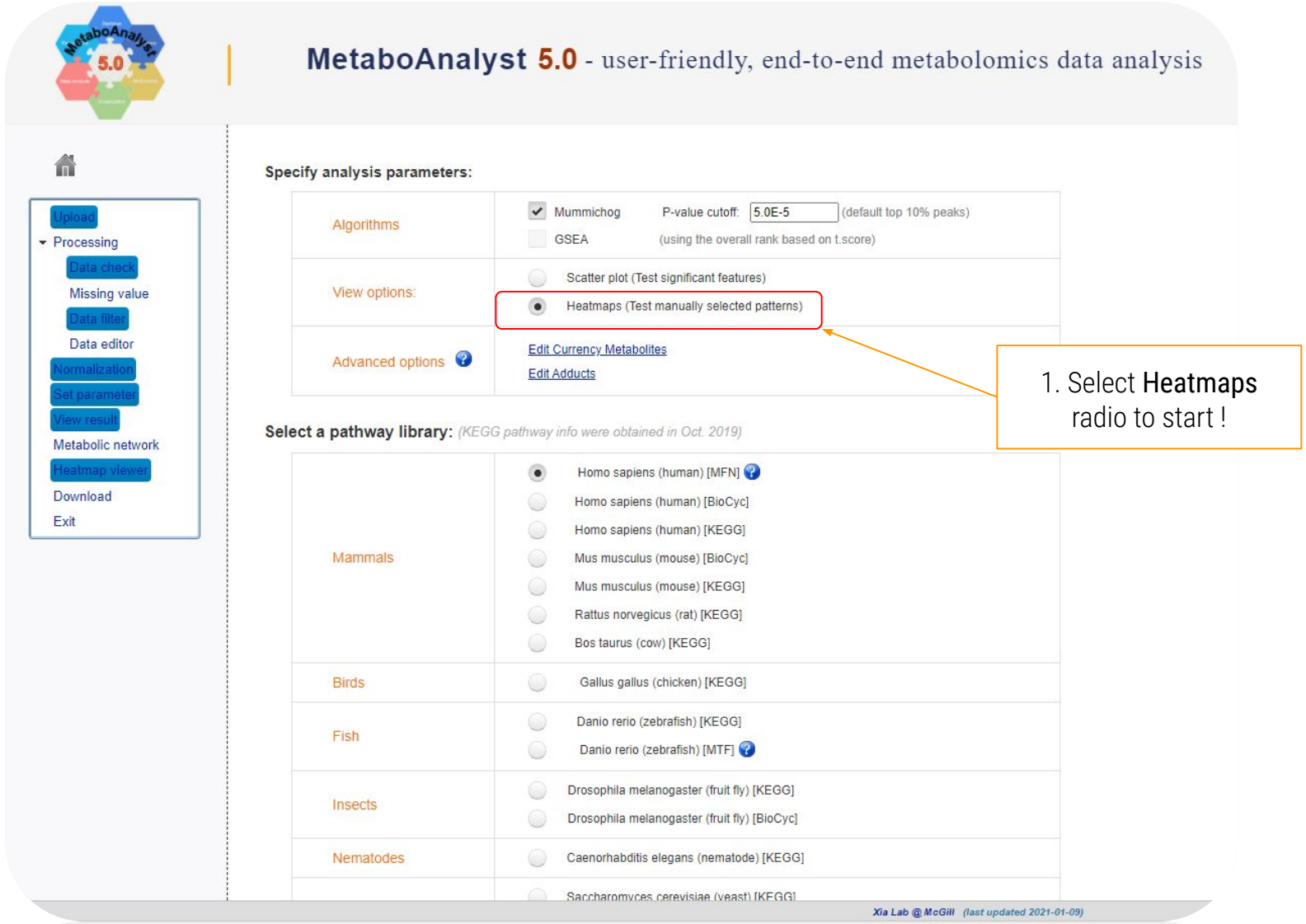
3. Perform Data Normalization

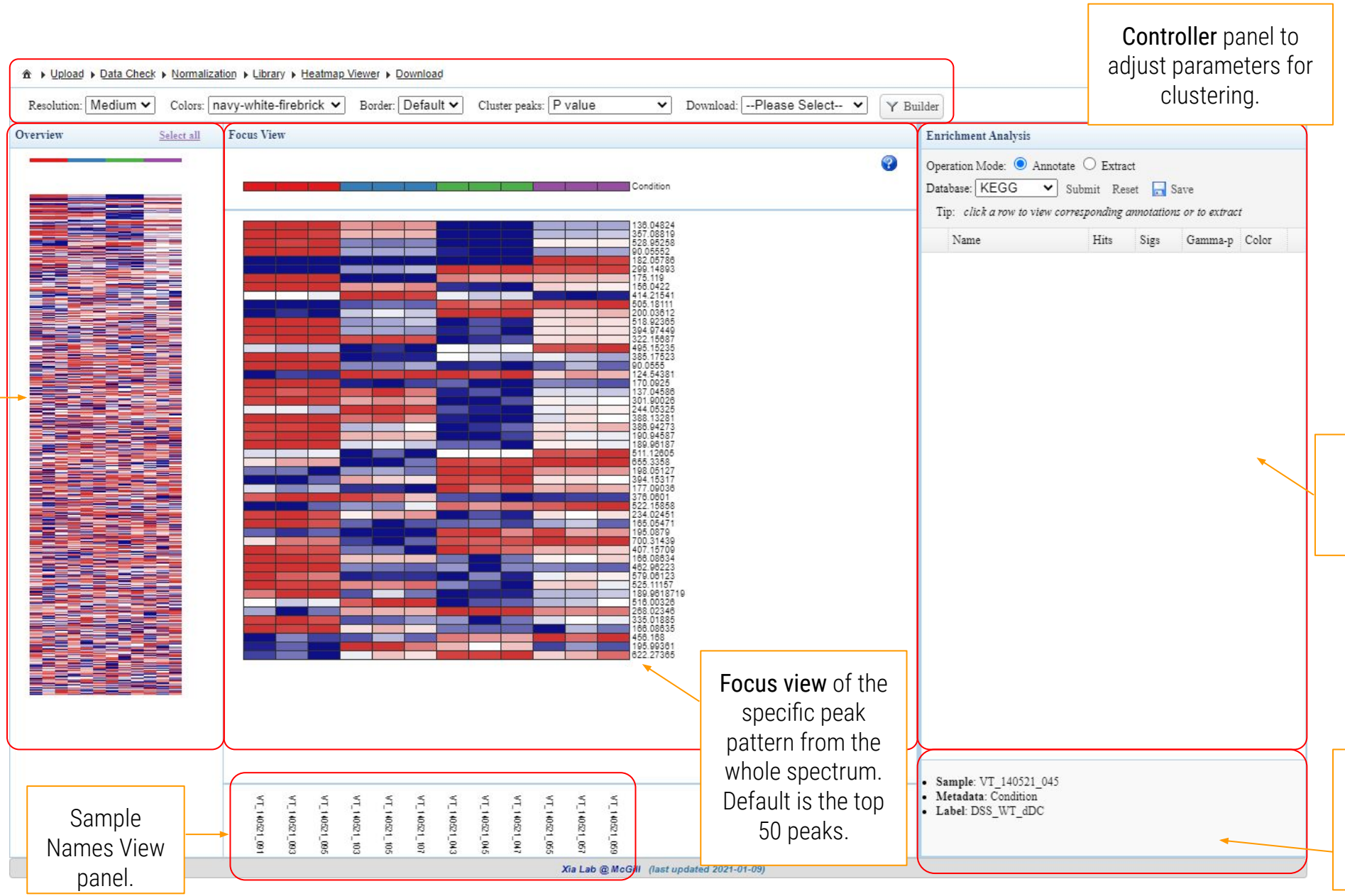
Set parameters



Set corresponding parameters/Library

Heatmap based pattern specific analysis





Controller panel to adjust parameters for clustering.

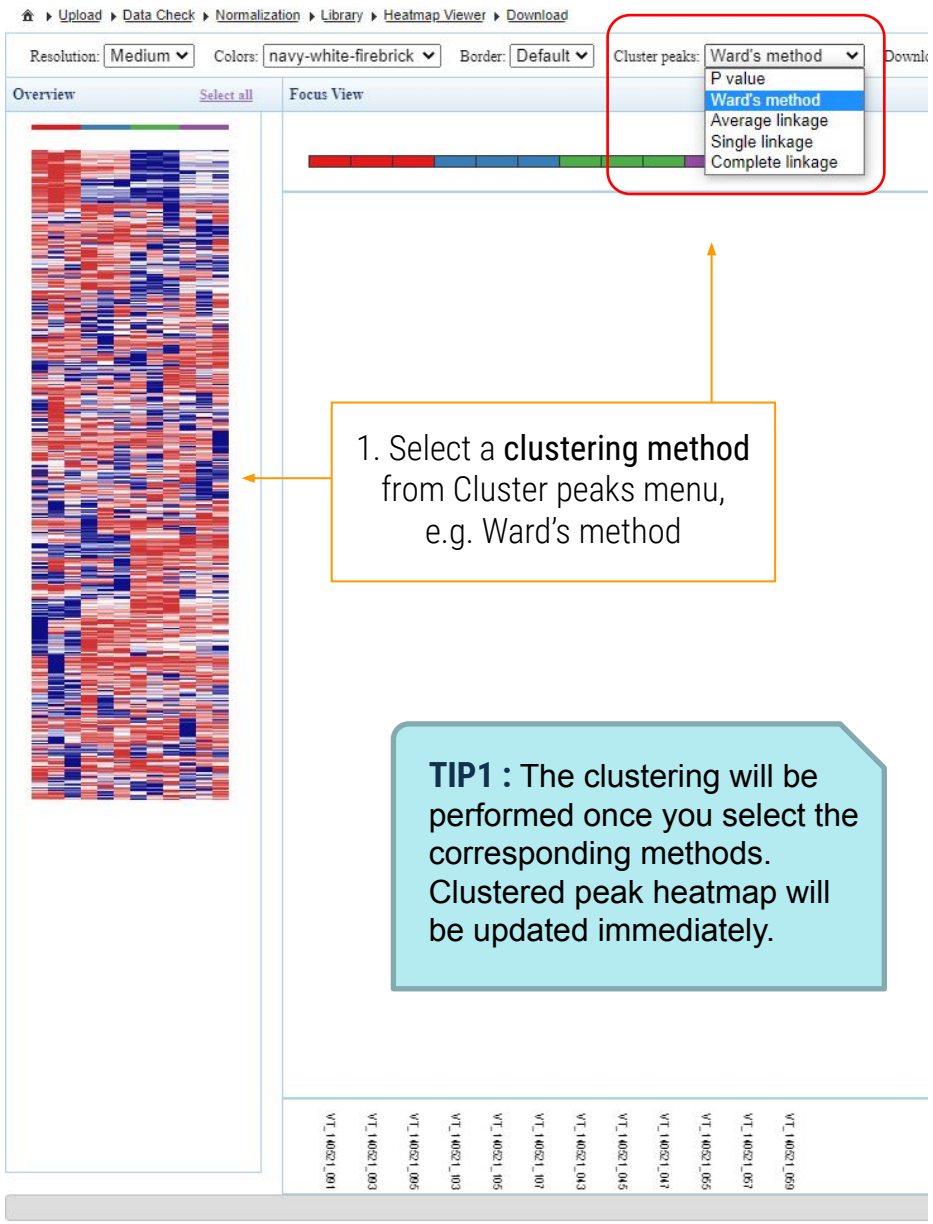
Overview of the peak across the spectrum. Default is clustered based on p value.

Pattern based **Enrichment analysis** panel.

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

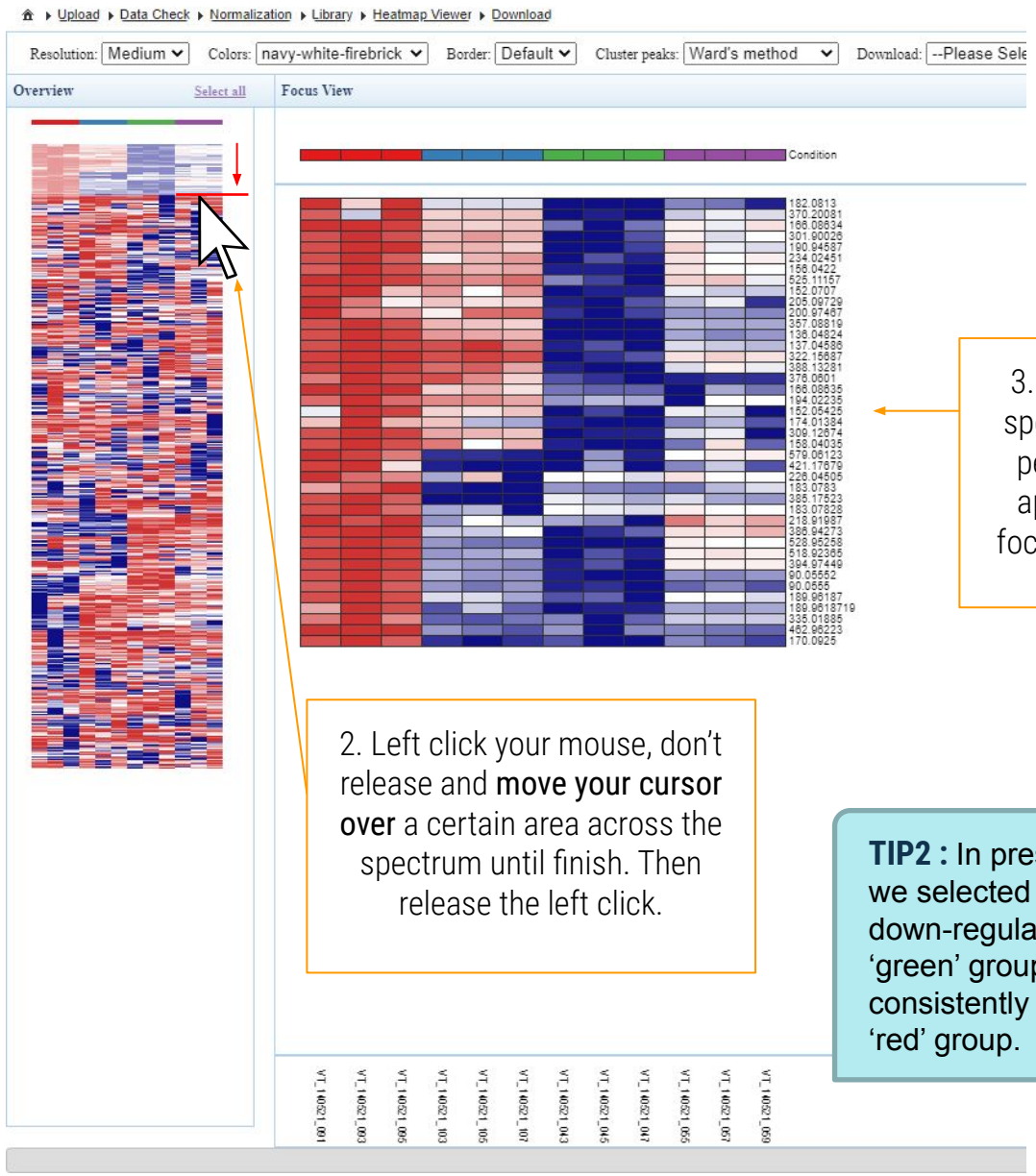
Sample Names View panel.

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



1. Select a **clustering method** from Cluster peaks menu, e.g. Ward's method

TIP1 : The clustering will be performed once you select the corresponding methods. Clustered peak heatmap will be updated immediately.

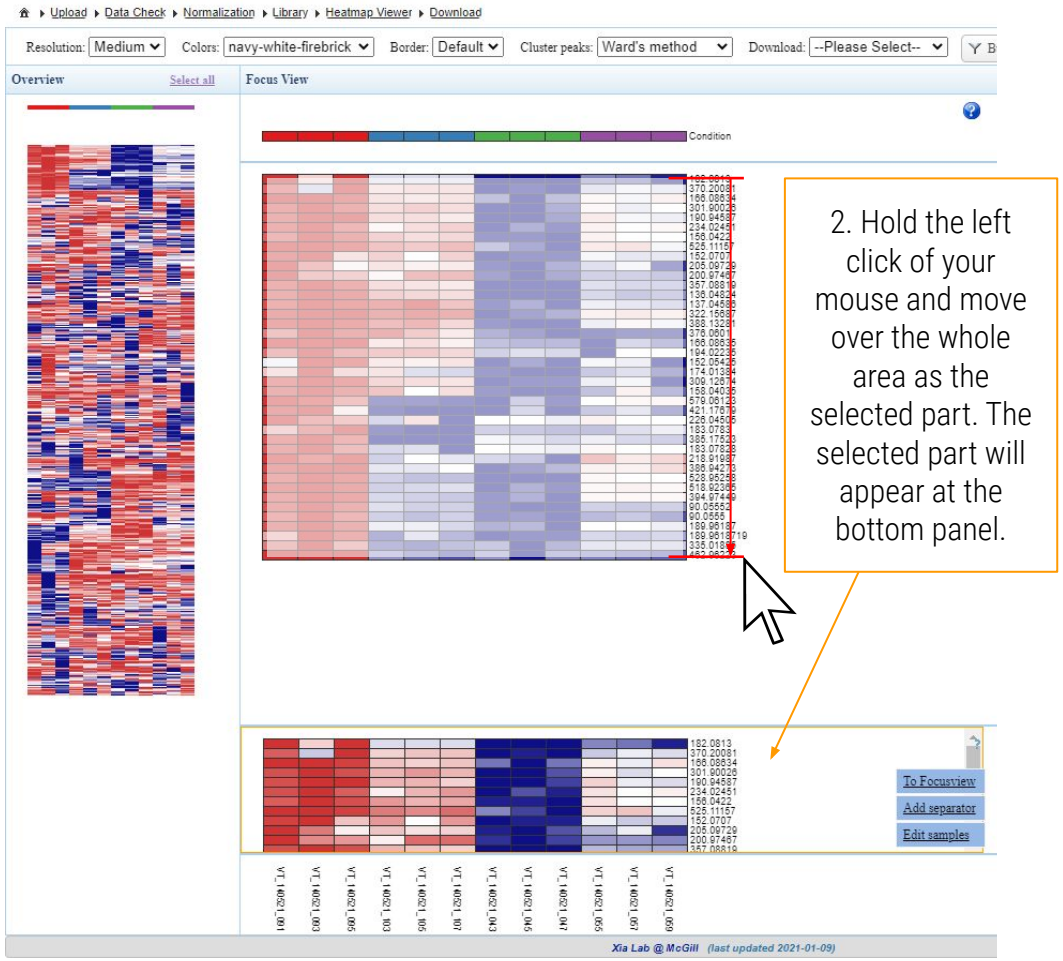


2. Left click your mouse, don't release and **move your cursor over** a certain area across the spectrum until finish. Then release the left click.

3. The selected specific spectral peaks' **pattern** appears in the focus view panel.

TIP2 : In present example, we selected a consistently down-regulated pattern in 'green' group, and consistently upregulated in 'red' group.

1. If you want to stitch the different peak patterns. Click the **Builder Button** and the stitch tools appears.



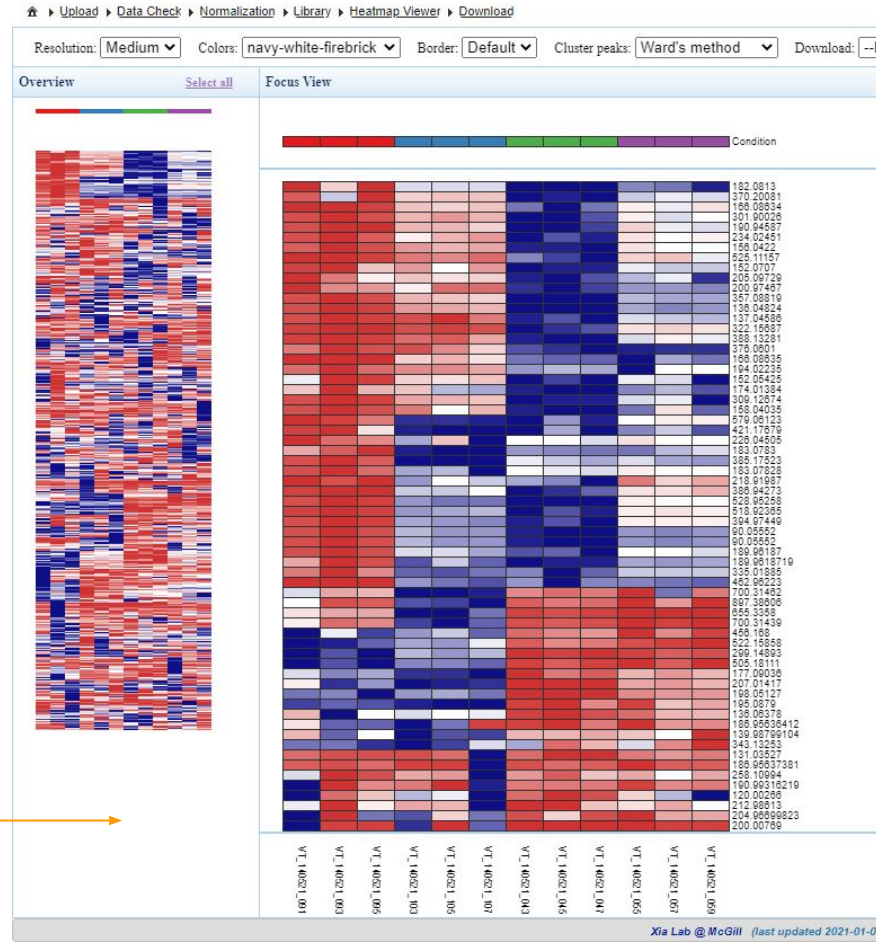
2. Hold the left click of your mouse and move over the whole area as the selected part. The selected part will appear at the bottom panel.

3. Left click and move your mouse over a new pattern from the overview panel.

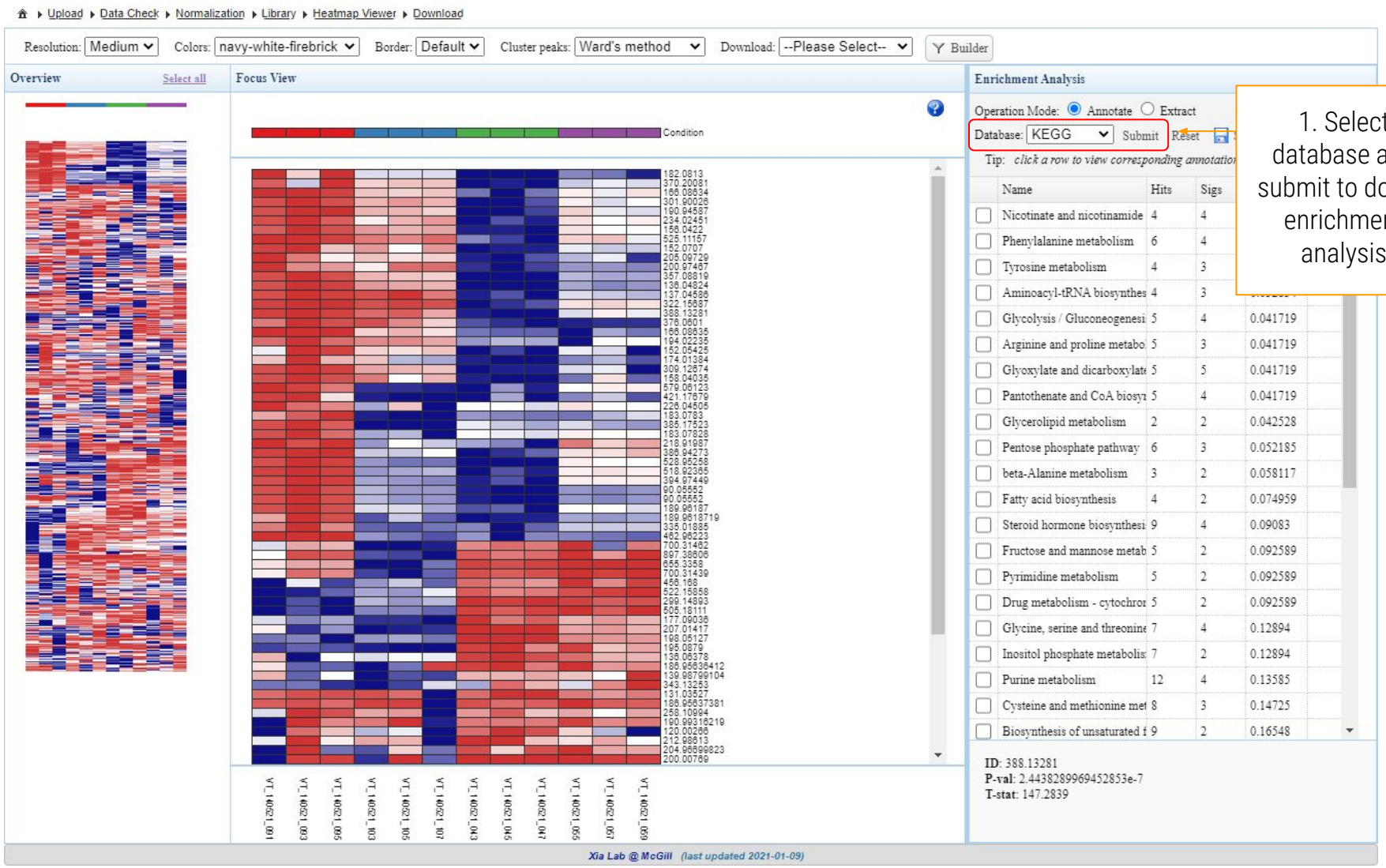
4. New pattern appears immediately at the top focus view panel. Hold your left click mouse and move over from this focus view to confirm the area.

5. Newly selected area will be stitched with the previously selected pattern at the bottom focus view.

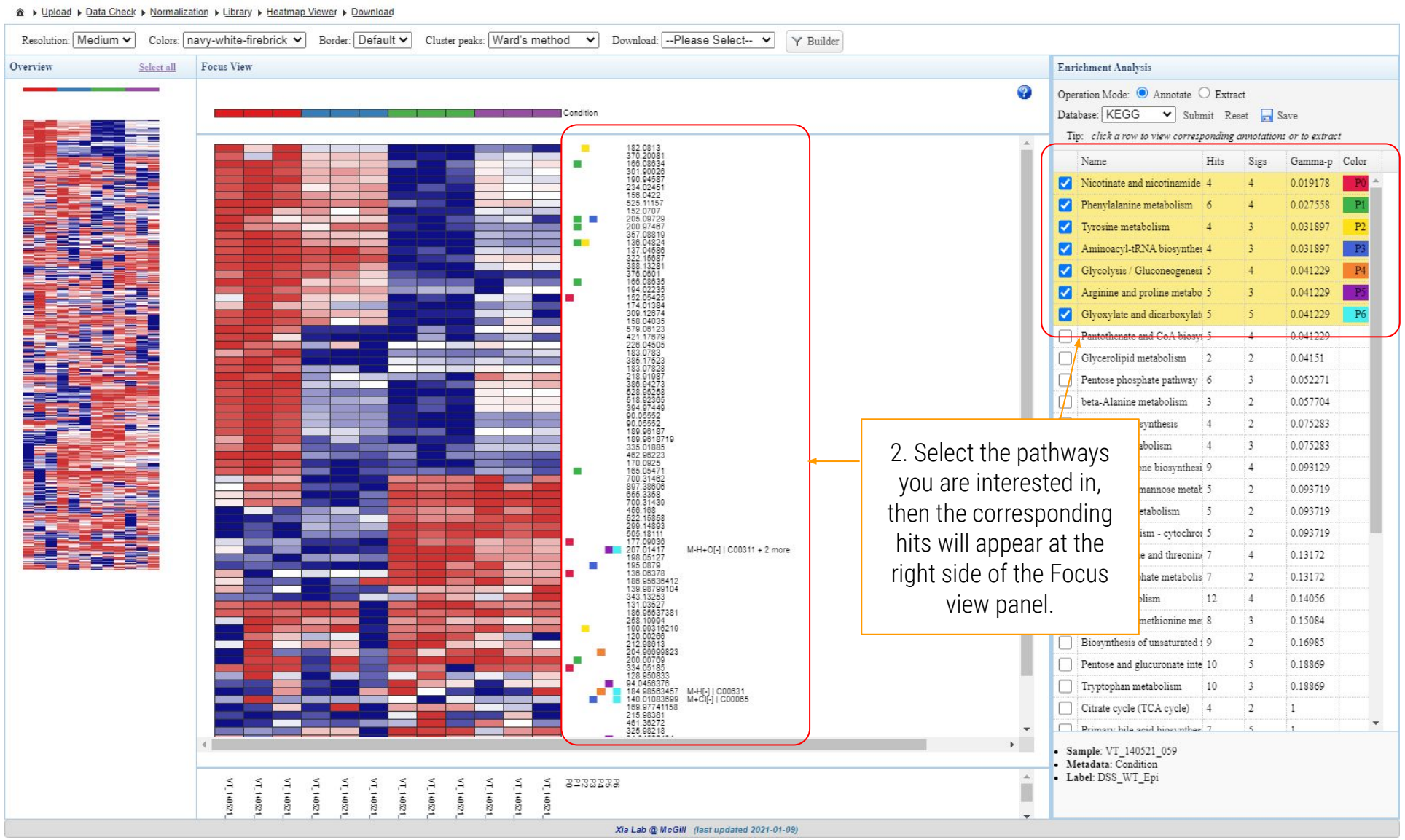
6. Click To Focus view.



7. The stitched peaks' pattern will be presented at the Focus view.



1. Select database and submit to do the enrichment analysis

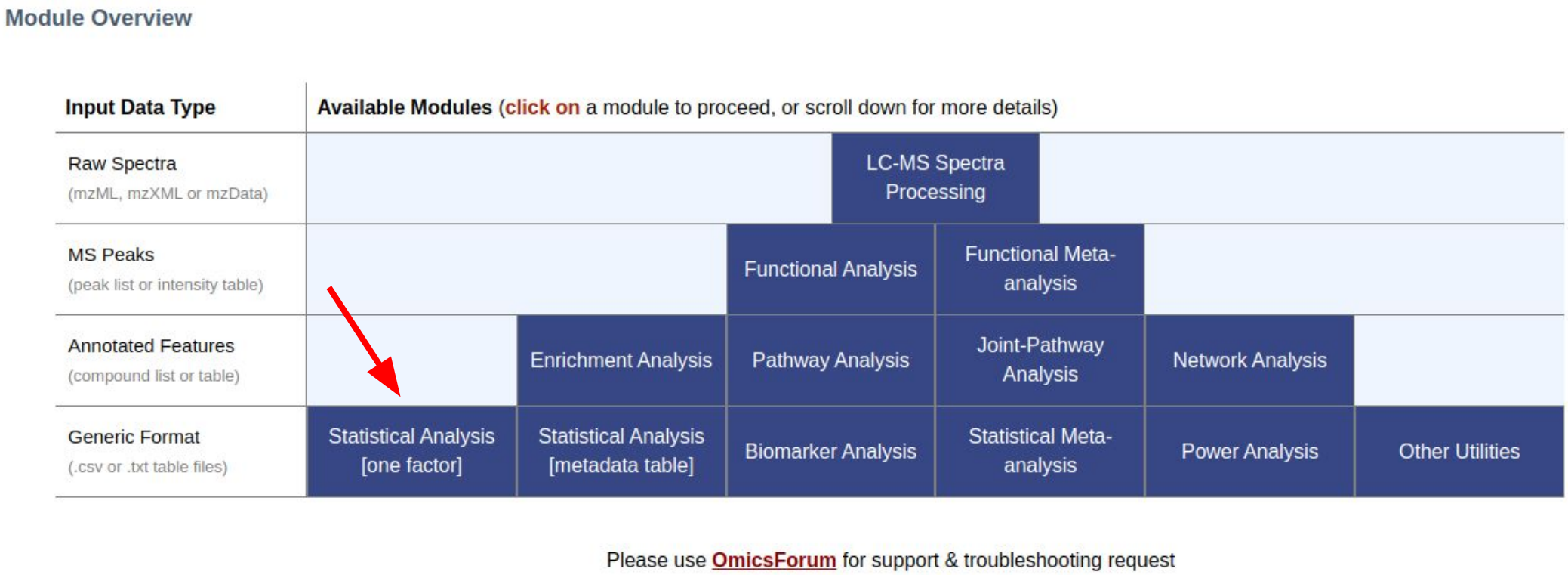


2. Select the pathways you are interested in, then the corresponding hits will appear at the right side of the Focus view panel.

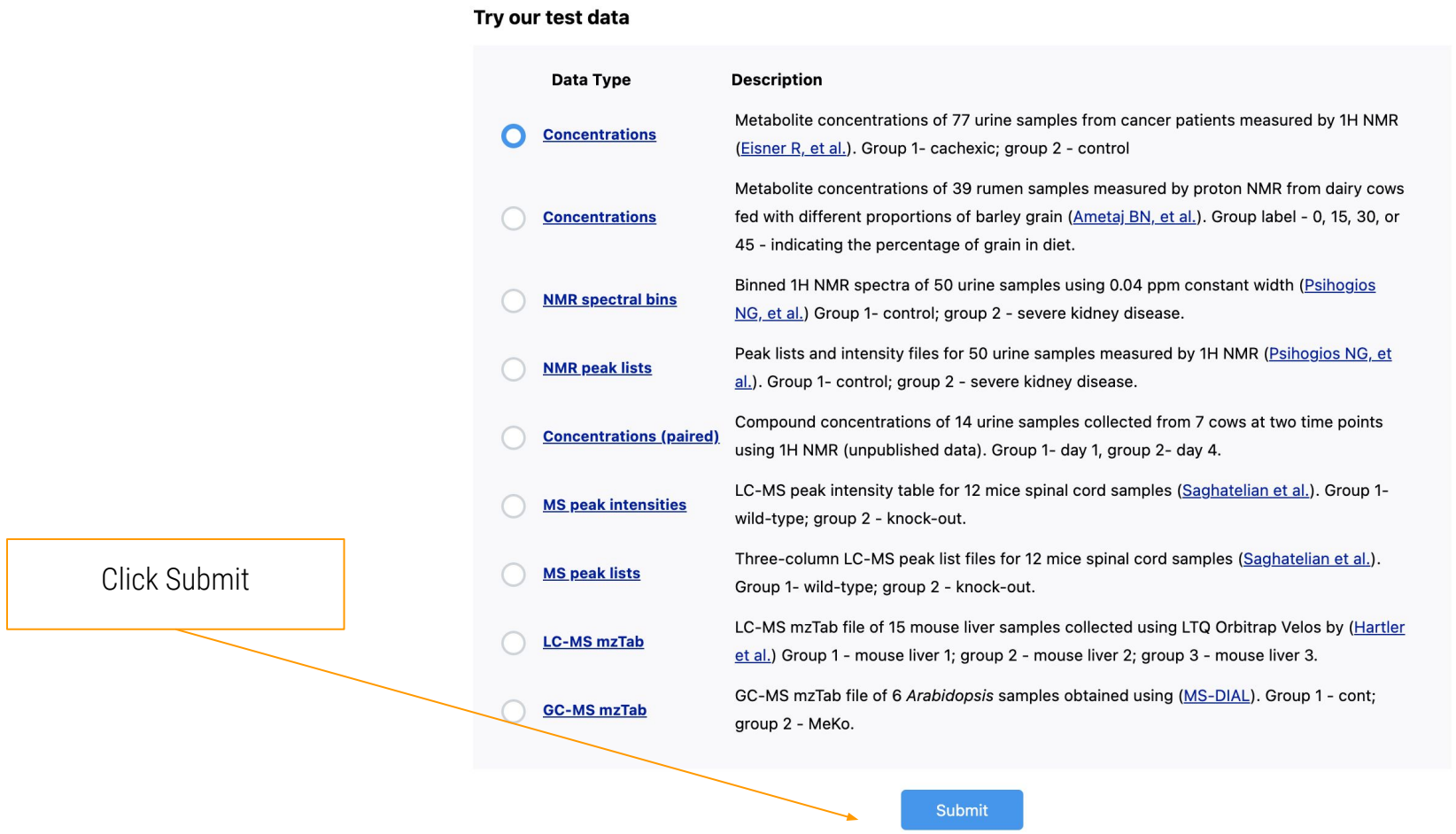
Statistical Analysis (one-factor) Demos

MetaboAnalyst: Statistical Analysis (one-factor)

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

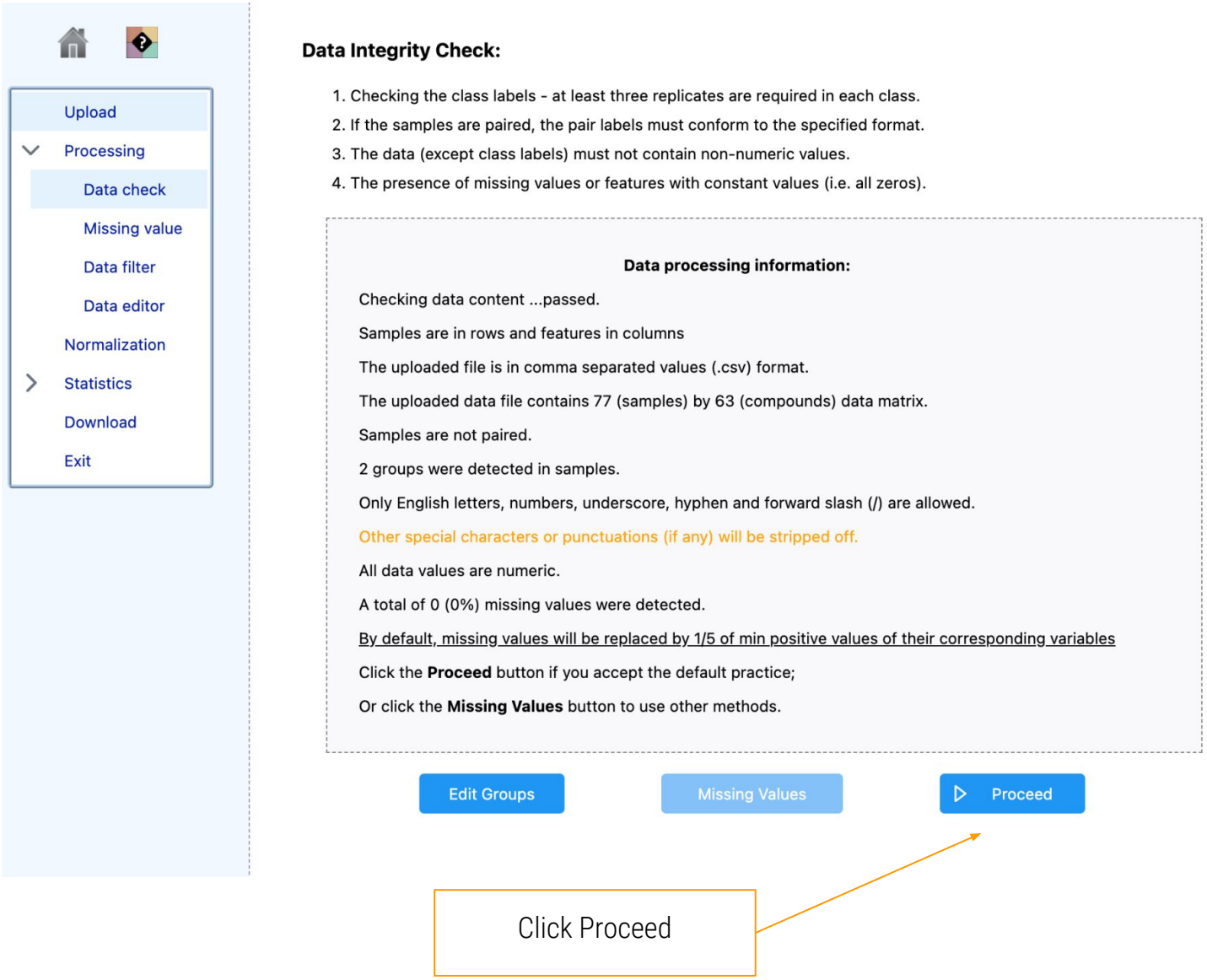


Data Upload



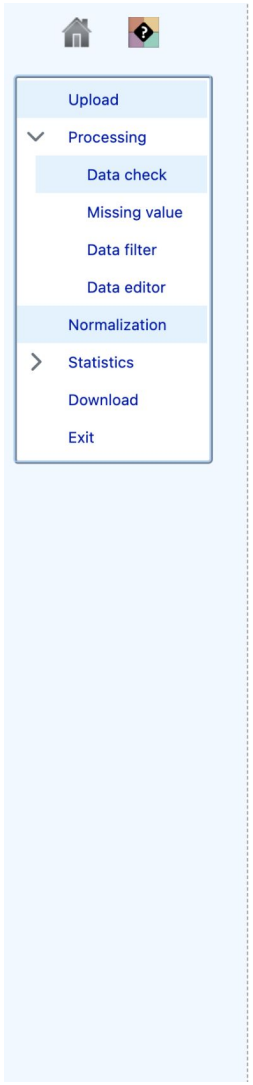
Click Submit

Data Check

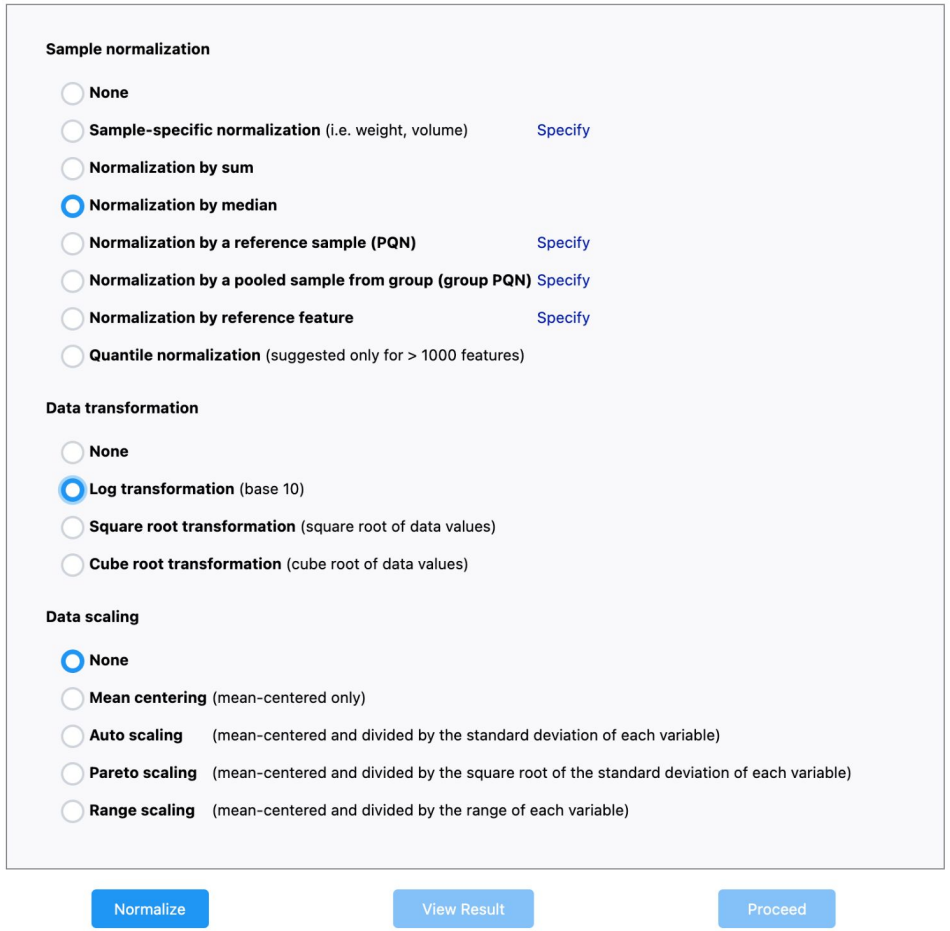

Upload
Processing
Data check
Missing value
Data filter
Data editor
Normalization
Statistics
Download
Exit
Data Integrity Check:
1. Checking the class labels - at least three replicates are required in each class.
2. If the samples are paired, the pair labels must conform to the specified format.
3. The 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”.



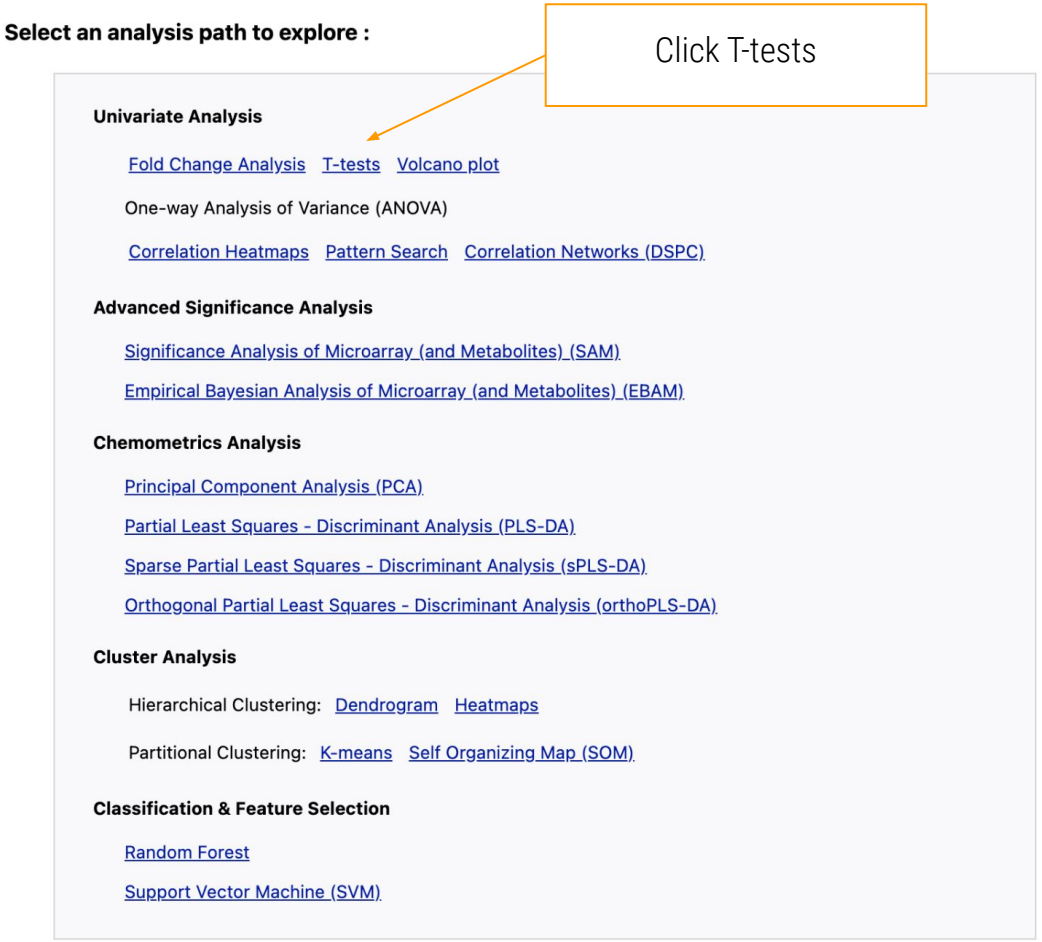
- Sample normalization is for general-purpose adjustment for systematic differences among samples;
- Data transformation applies a mathematical transformation on individual values themselves. A simple mathematical approach is used to "normalization #metaboanalyst" to find more information.
- Data scaling adjusts each variable/feature by a scaling factor computed based on the dispersion of the variable.



Analysis Overview

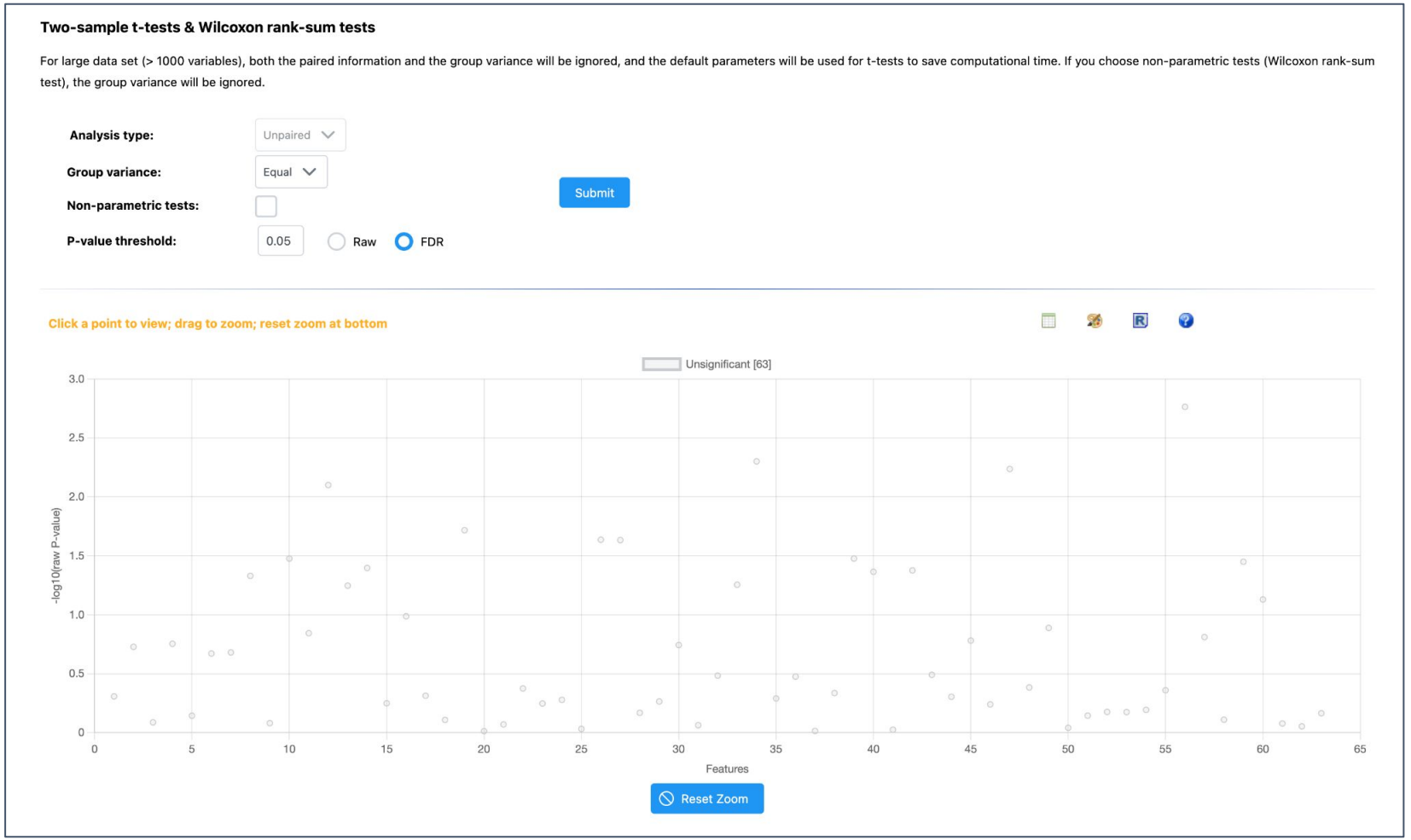
There are many analysis options. We will explore:

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



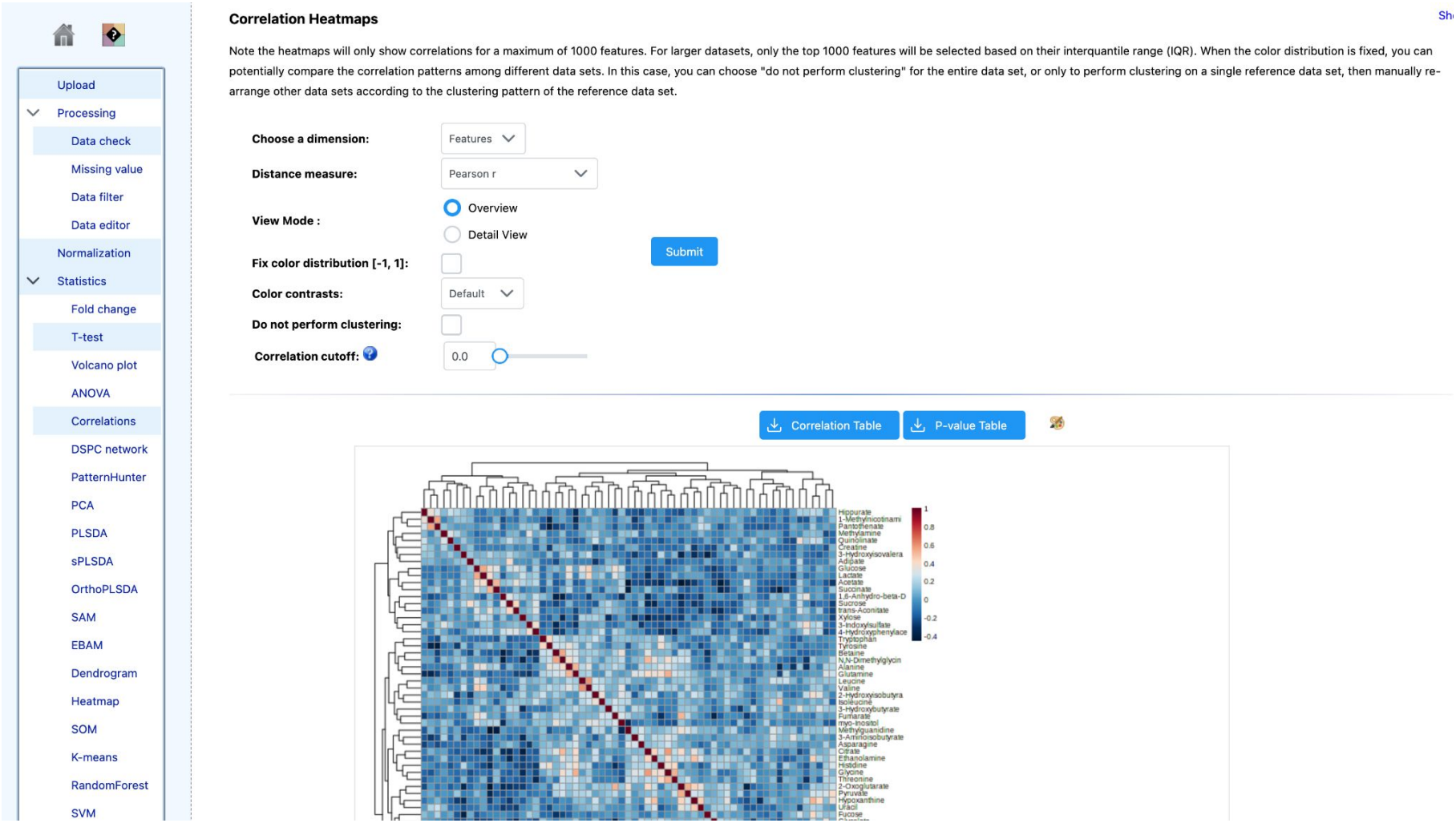
T-tests

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



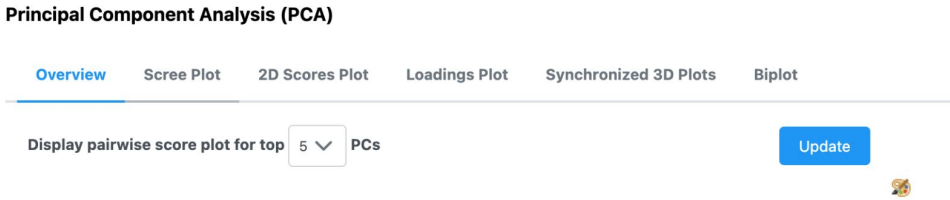
Correlation Heatmap

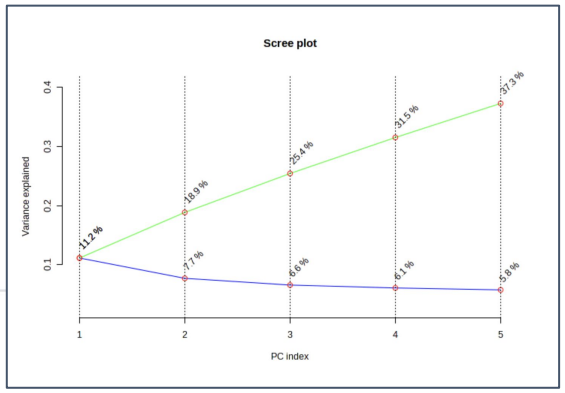
The heatmap shows pairwise correlations between features. Groups of features with similar patterns form clusters in the heatmap. Adjust the parameters and click “Submit” to explore the settings. When done, navigate back to the analysis overview and click “PCA”.

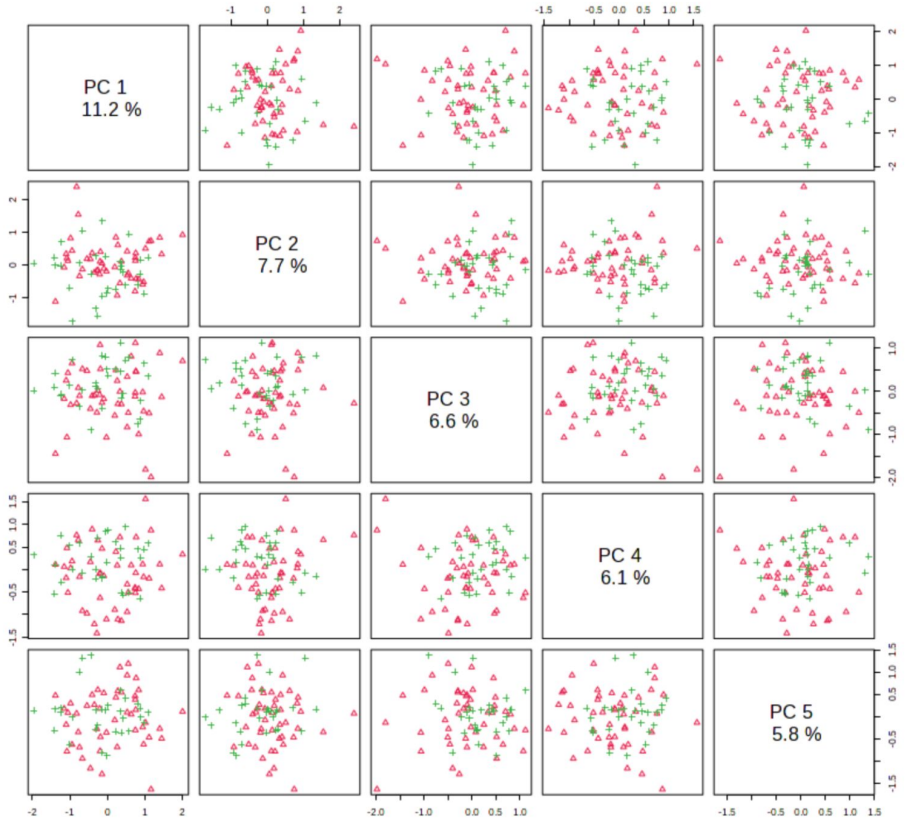


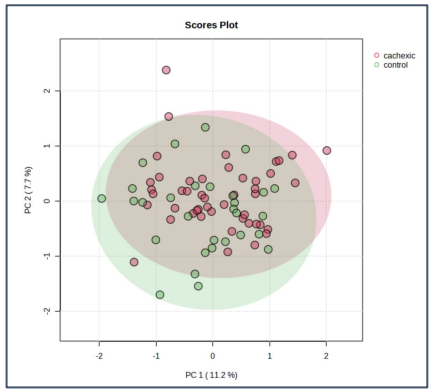
PCA

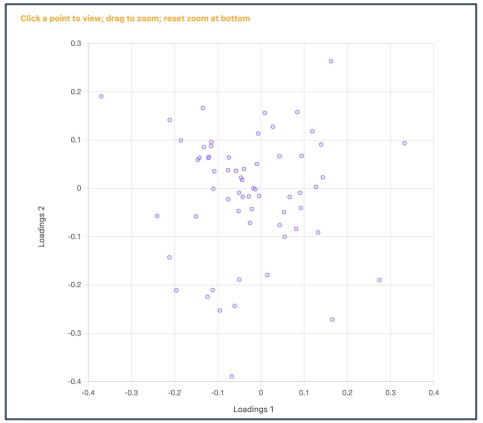
There are many views of the PCA results in the different tabs. Click some to explore the results. When finished, navigate back to the analysis overview and click “PLS-DA”.









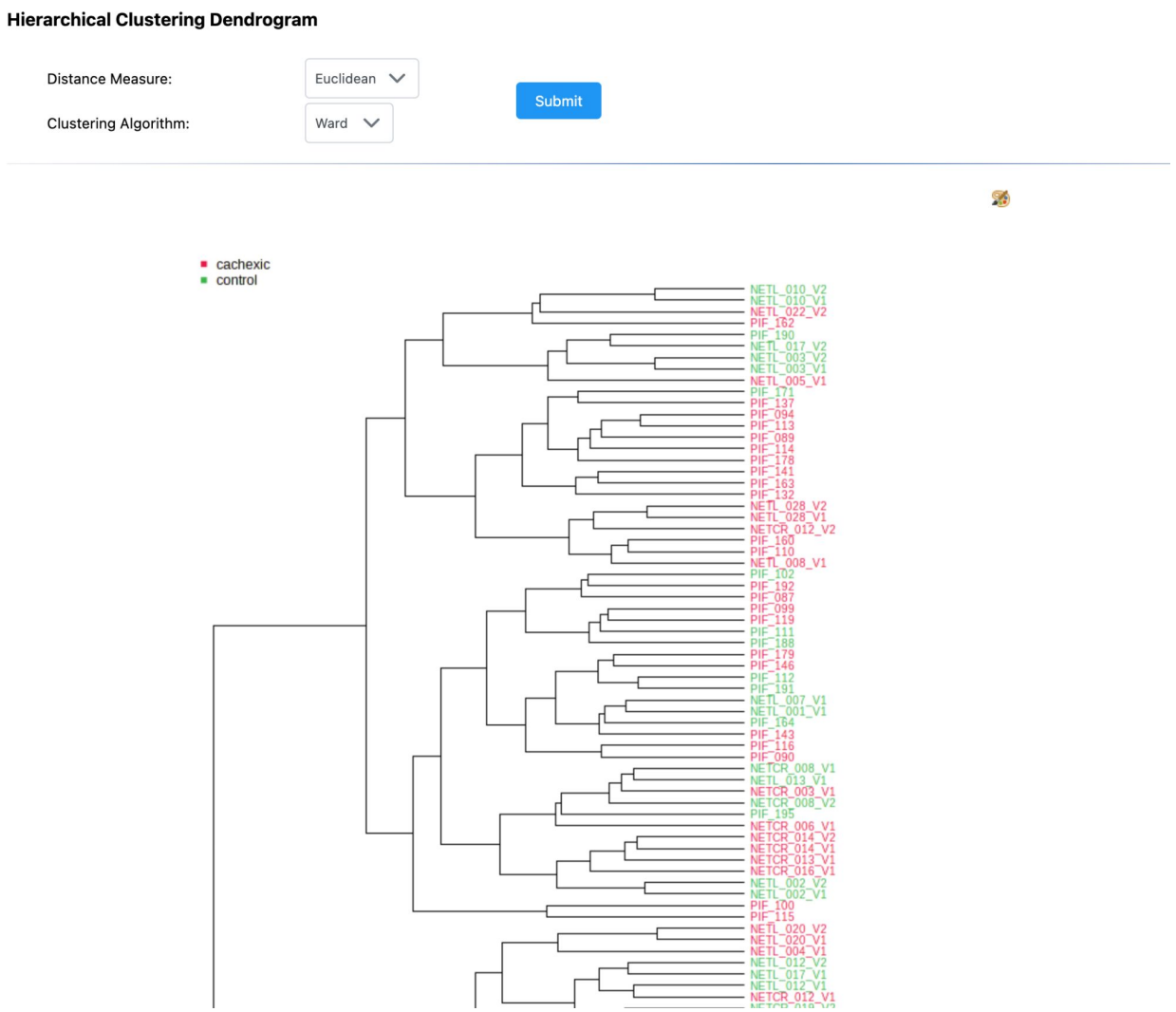


PLS-DA

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

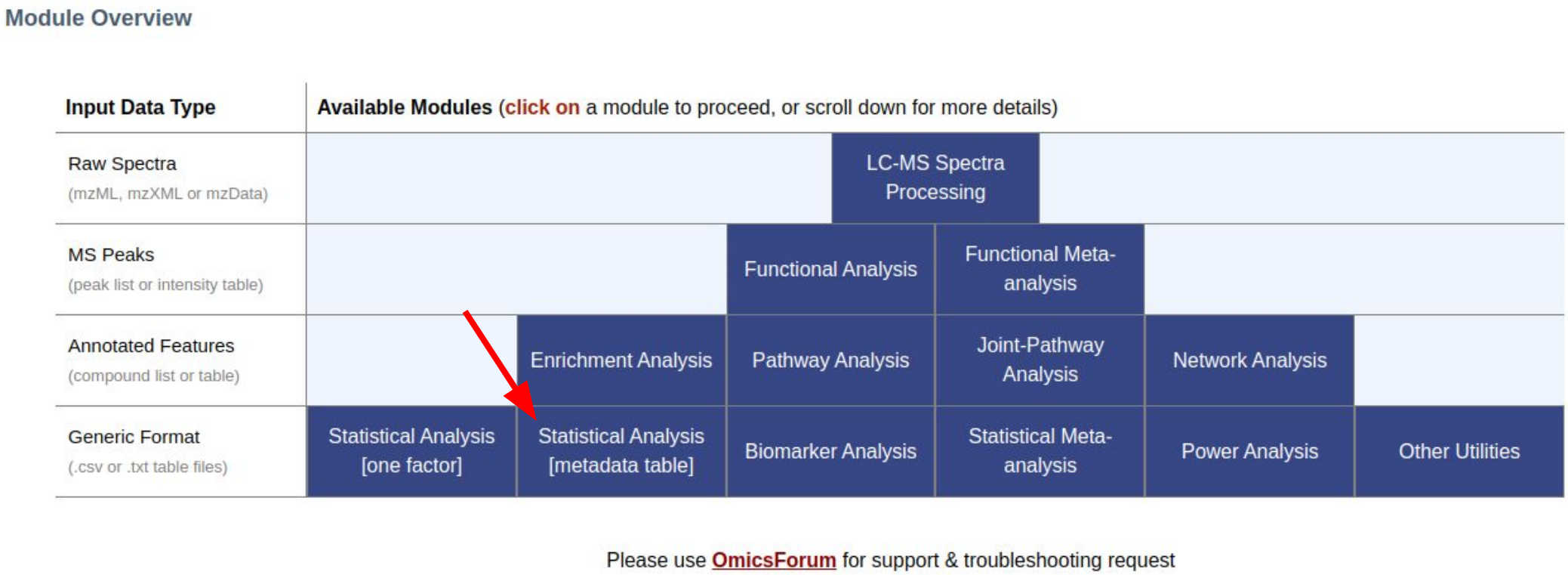
Quickly generate dendrograms using different clustering algorithms.



Statistical Analysis (meta-data) Demos

MetaboAnalyst: Statistical Analysis (metadata table)

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



Data Upload

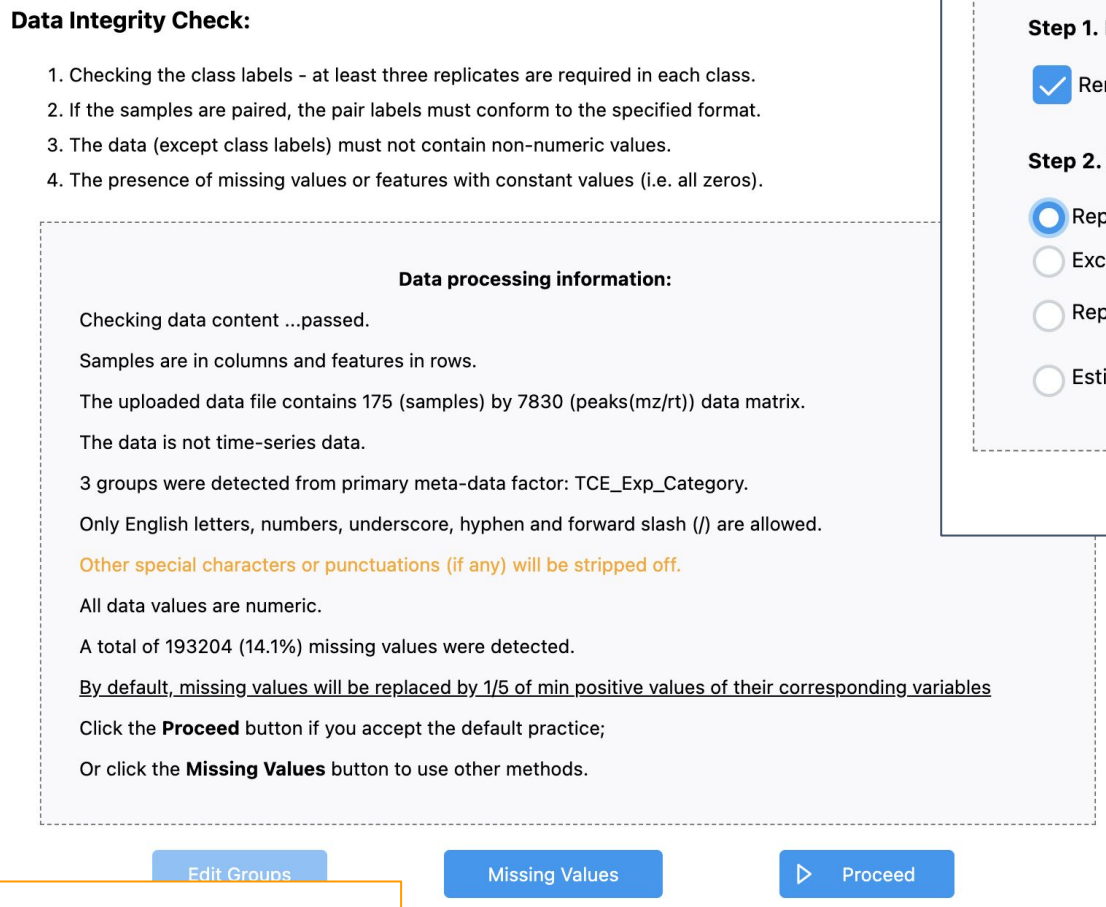
Select the **second** example dataset and click “Submit”.

Try our test data

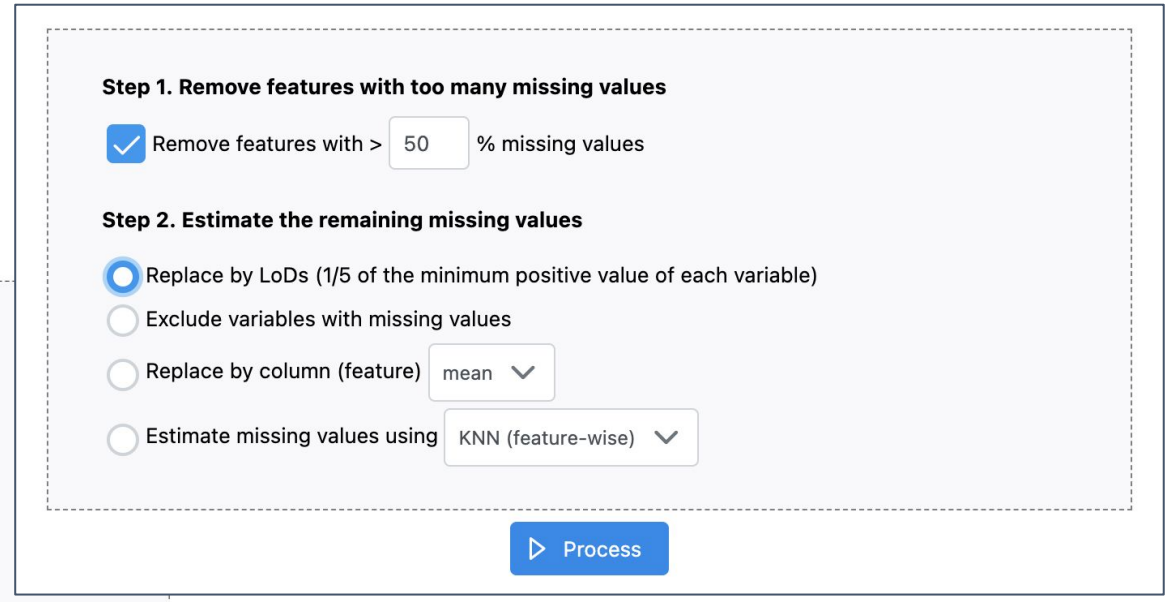
	Data	Study Design	Description
○	Data Metadata	Multiple factors / covariates	LC-MS peak intensity data table and meta-data of 20 healthy and 39 COVID-19 patient samples; **Four metadata** - 3 categorical and 1 numeric.
◉	Data Metadata	Multiple factors / covariates	LC-MS peak intensity data from plasma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

Submit

Data Check (Missing Values)



1 - Click “Missing Values”



2 - Click “Submit”

Metadata Check

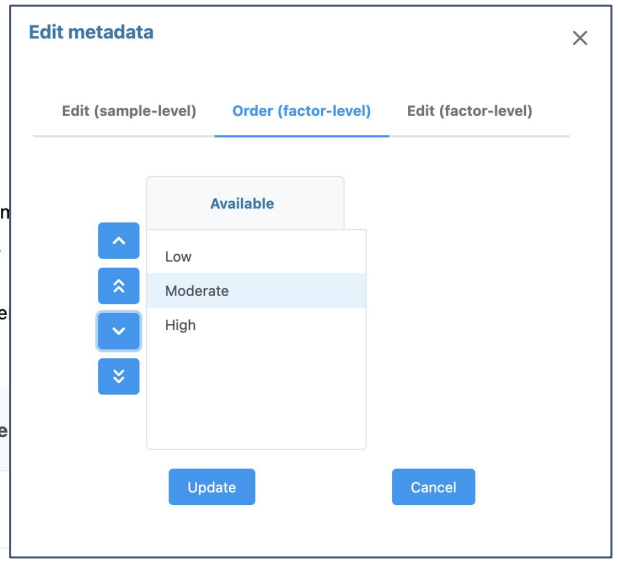
Update your metadata using the table below

- Update metadata type: categorical option for experimental groups (i.e. control vs diseased), continuous for nun
- 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 unde
- 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	

Skip to analysis

Proceed



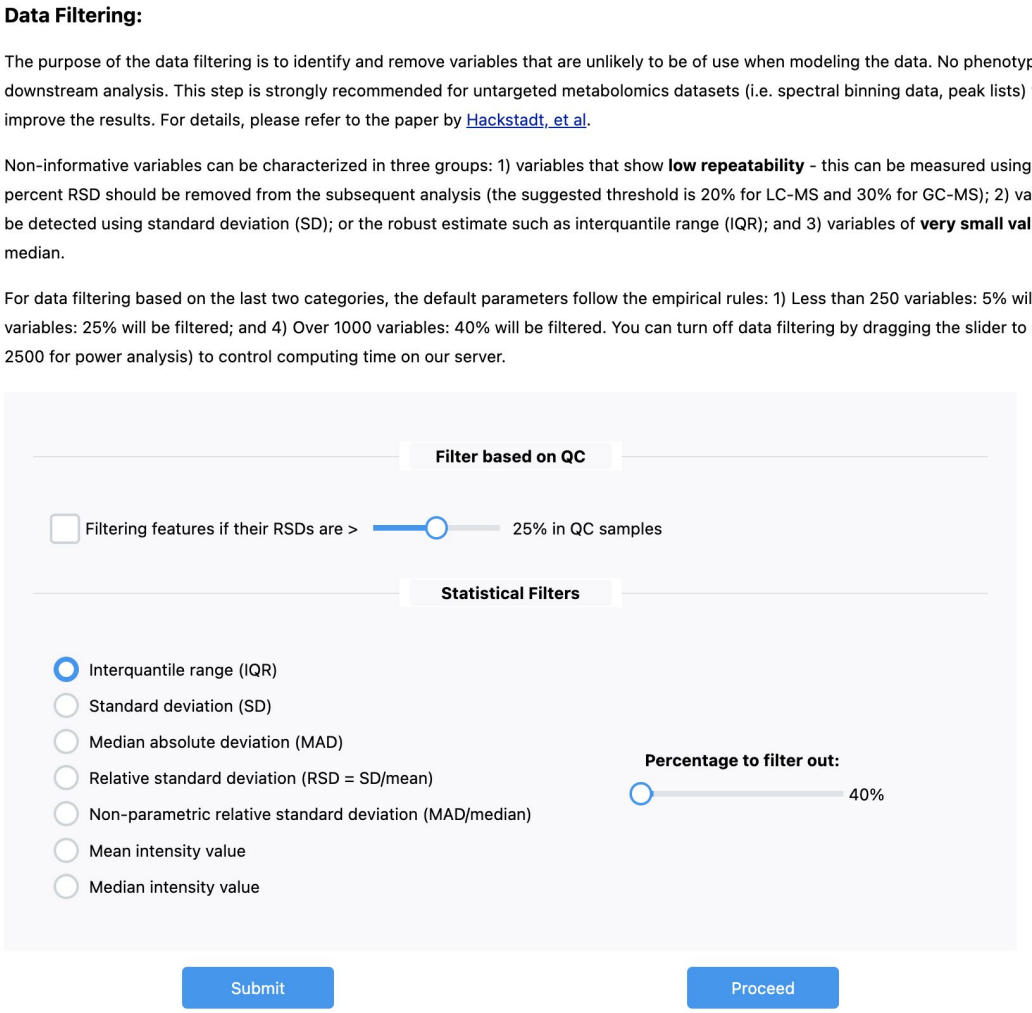
1 - Verify that the variable types are correct

2 - Click “Edit” in the “TCE_Exp_Category” row and re-order the groups to Low-Medium-High. Click “Update”

3 - Click “Proceed”

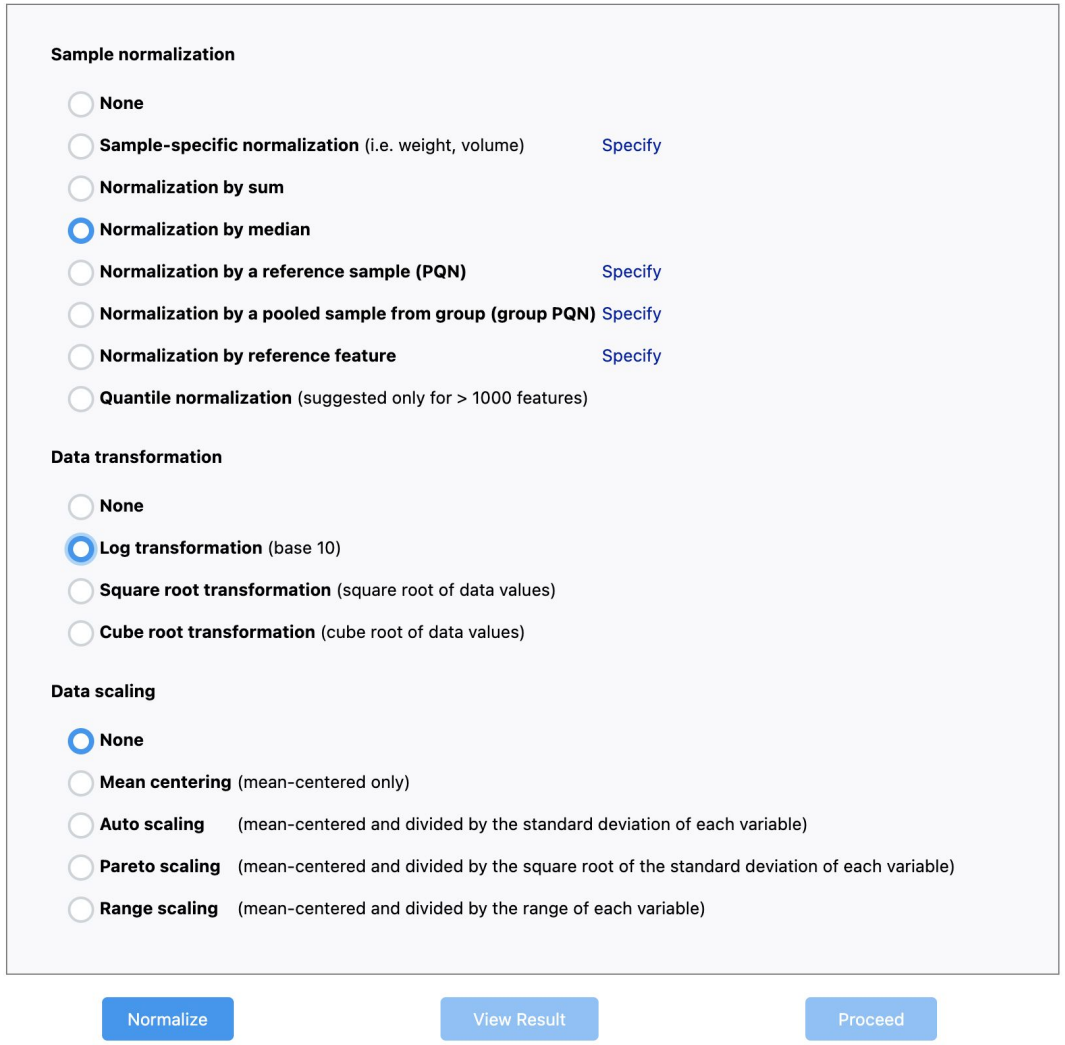
Data Filtering

Click “Submit” and “Proceed”



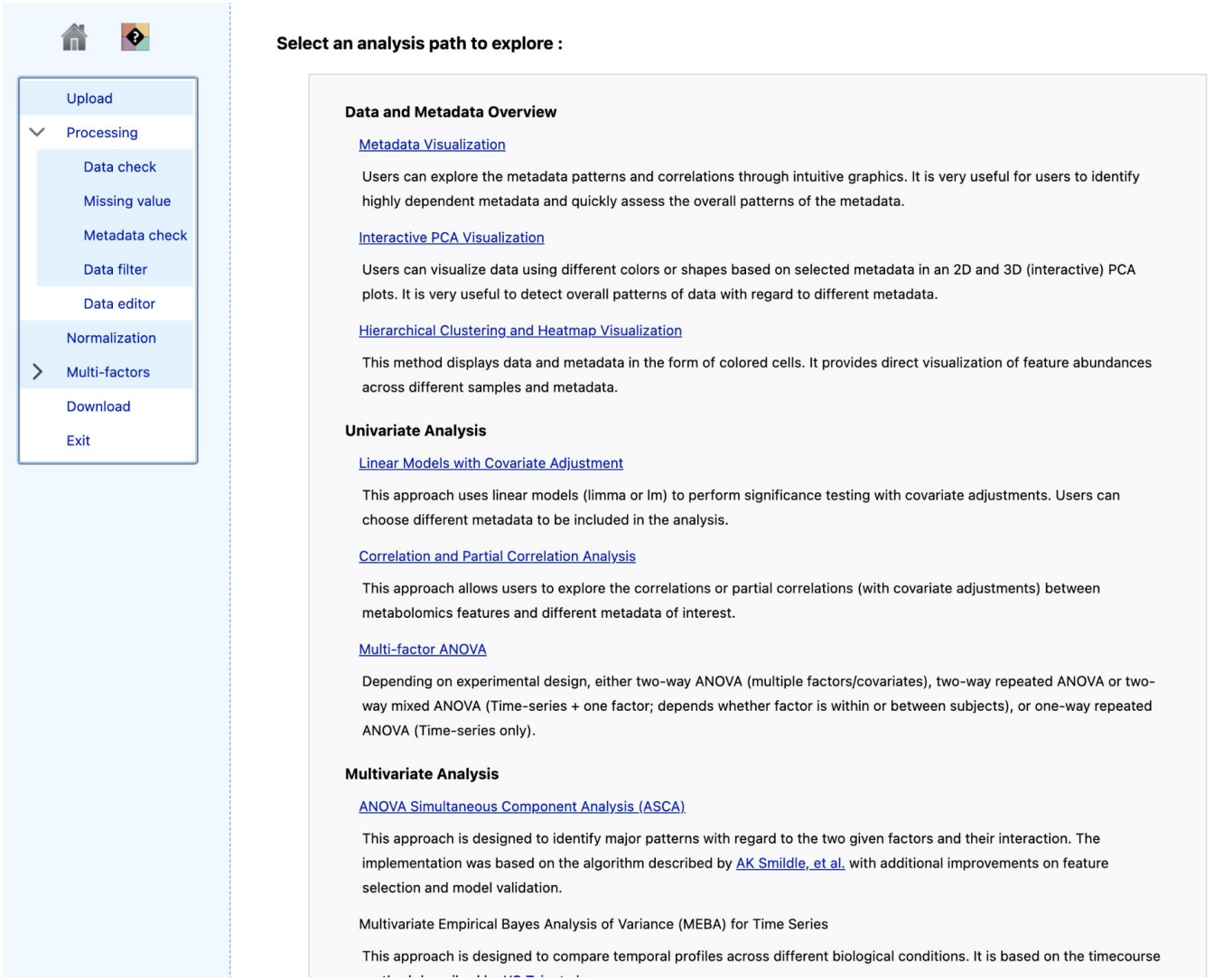
Normalization

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



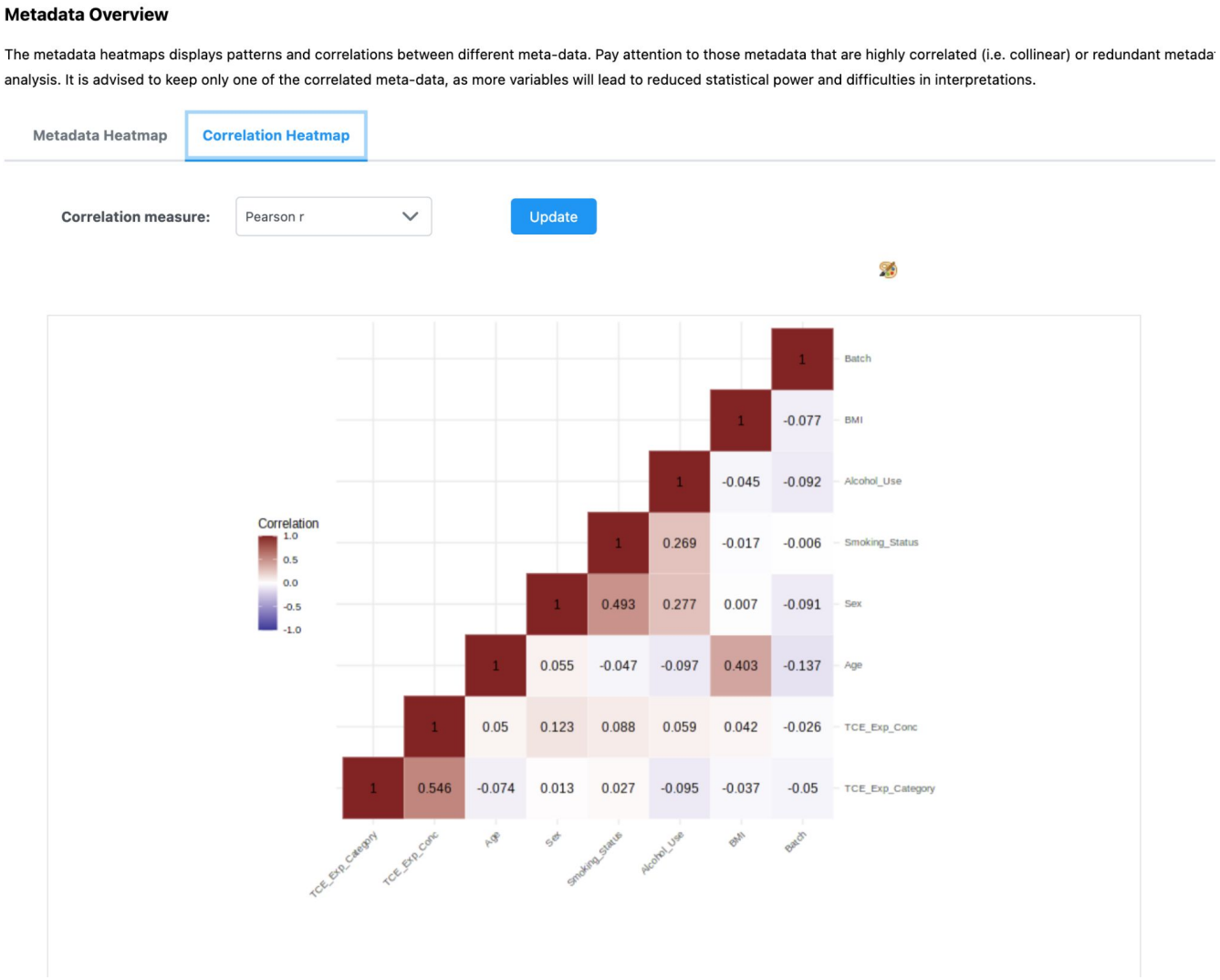
Analysis Overview

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”.



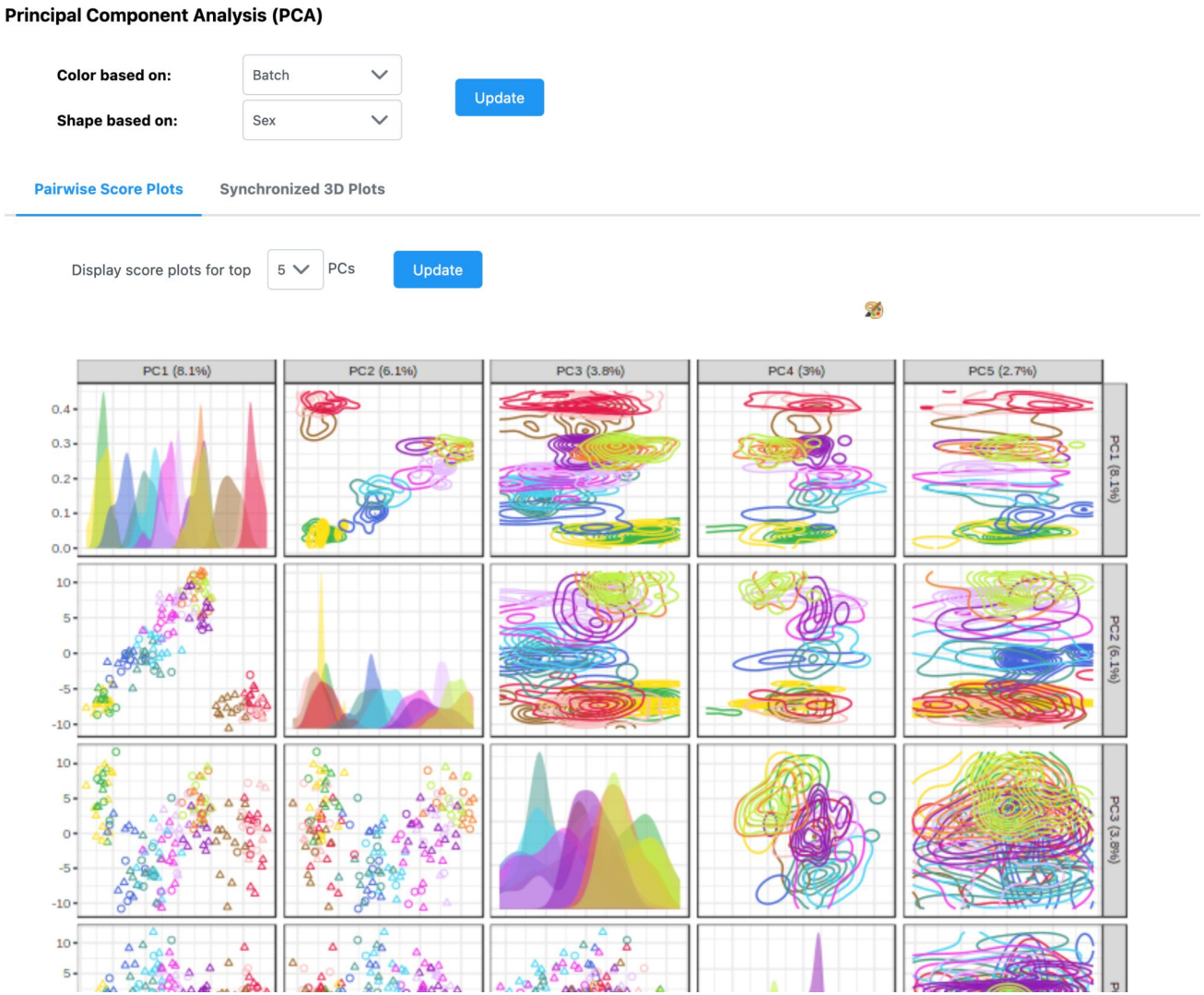
Metadata Visualization

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”.



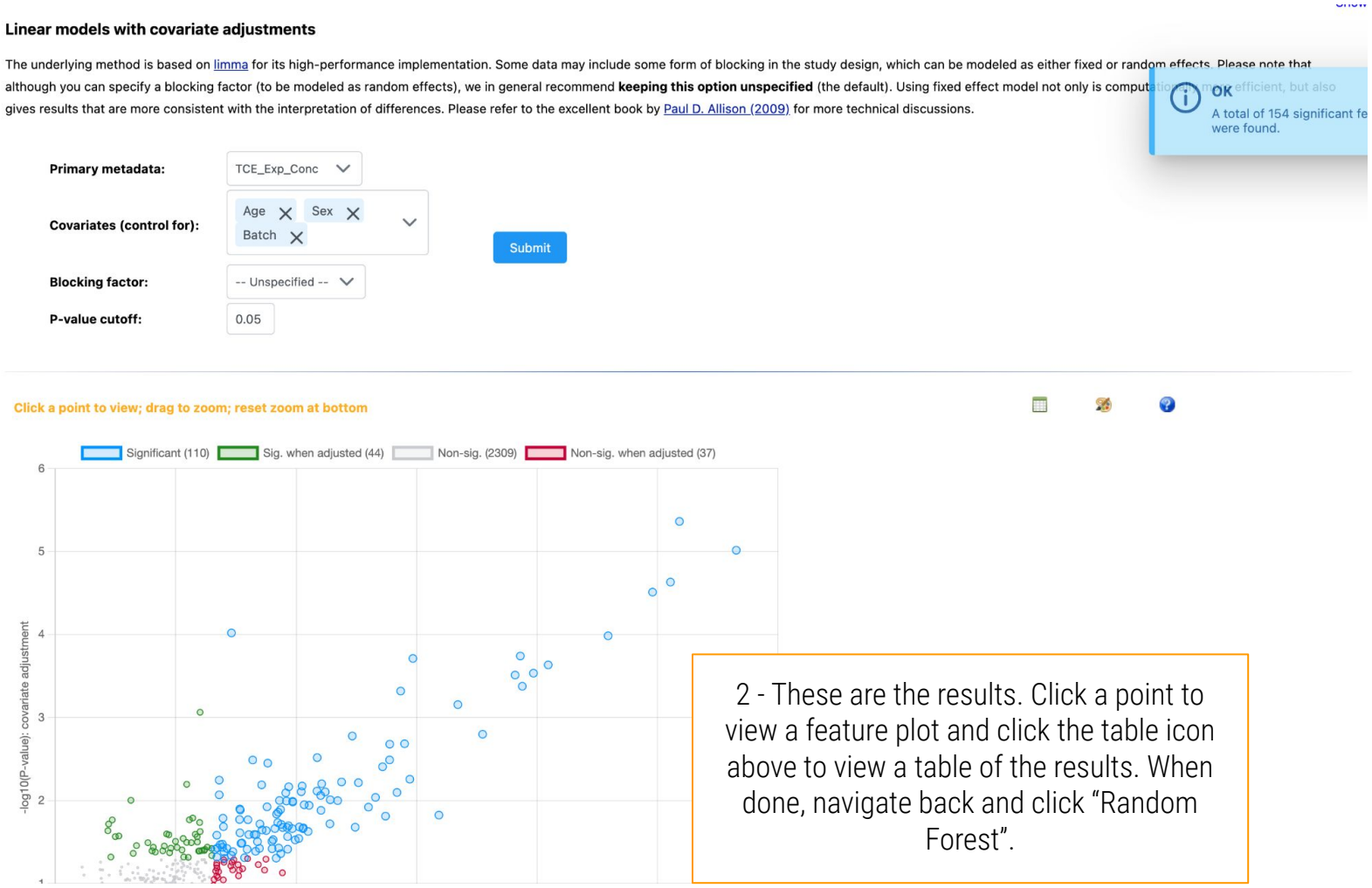
PCA

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

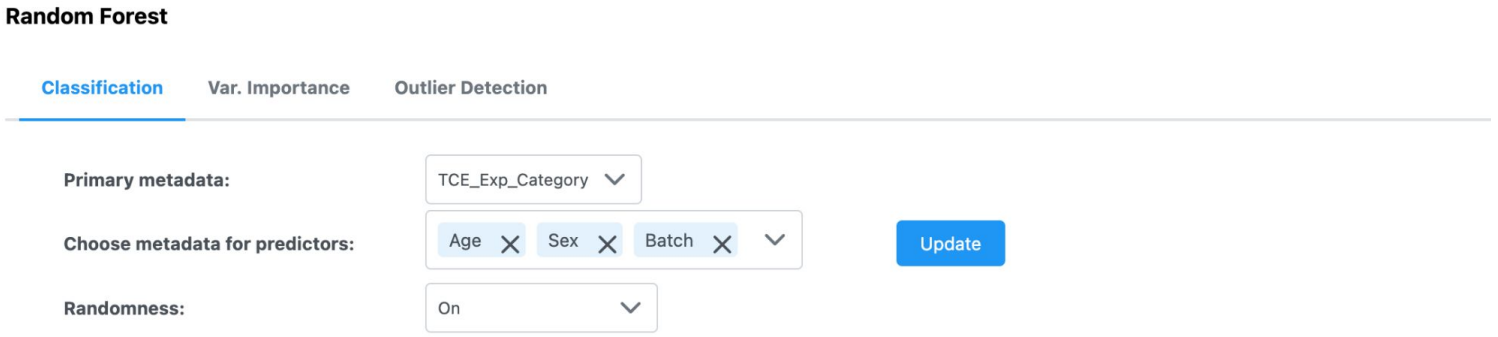
1 - Remember the relationships between metadata in the “Metadata Visualization” part? We don’t want to include multiple correlated variables in our analysis. Put “TCE_Exp_Conc” as the primary metadata, add “Age”, “Sex”, and “Batch” as covariates, and click “Submit”.

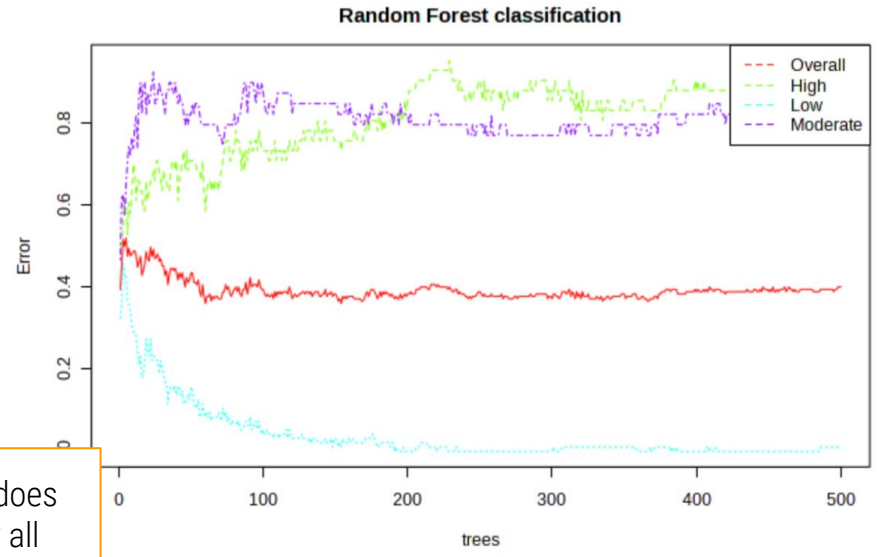


2 - These are the results. Click a point to view a feature plot and click the table icon above to view a table of the results. When done, navigate back and click “Random Forest”.

Random Forest

1 - Leave “TCE_Exp_Category” as the primary metadata and choose “Age”, “Sex”, and “Batch” as metadata predictors. Click “Submit”.





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

GenomeCanada