Spectra processing, functional integration and covariate adjustment of global metabolomics data using MetaboAnalyst 5.0

Section I: Metabolomics Raw Spectra data processing & functional analysis

TA: Zhiqiang Pang

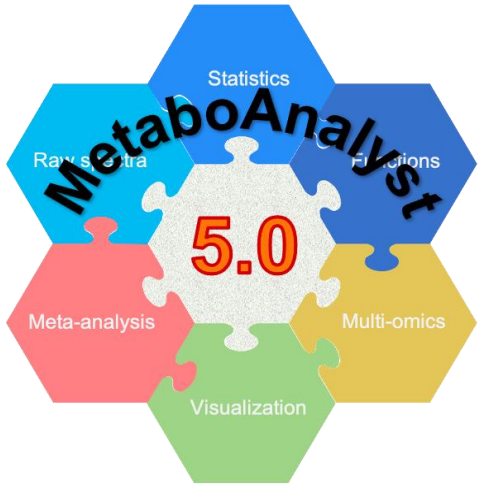
(zhiqiang.pang@mail.mcgill.ca)

18th Annual Conference of the Metabolomics Society

METABOLOMICS 2022

Valencia, Spain | JUNE 19-23

Pre-Conference Workshops

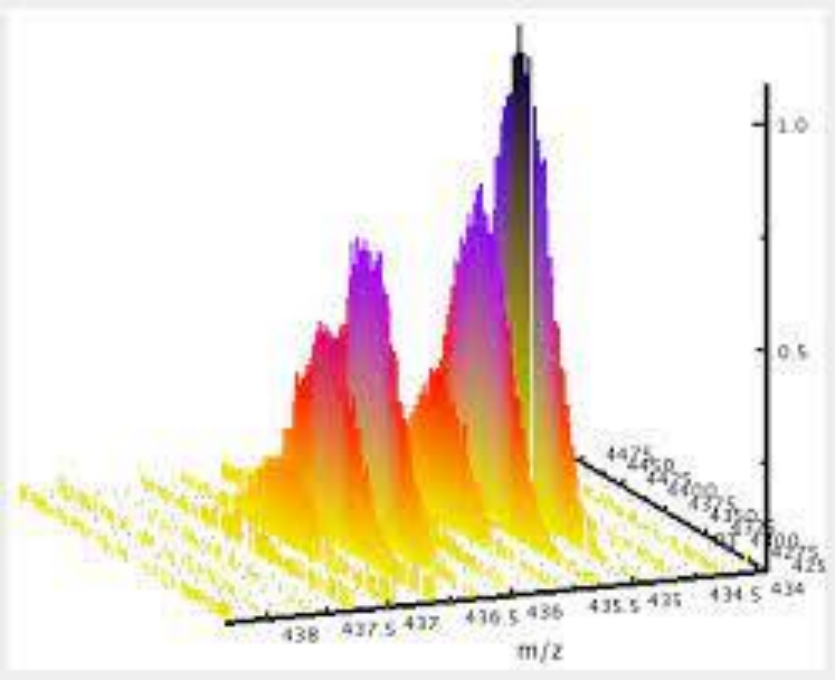


Learning Questions..

- What the algorithm MetaboAnalyst is using for raw spectra processing?
- How many data files we are supporting for processing online at the maximum?

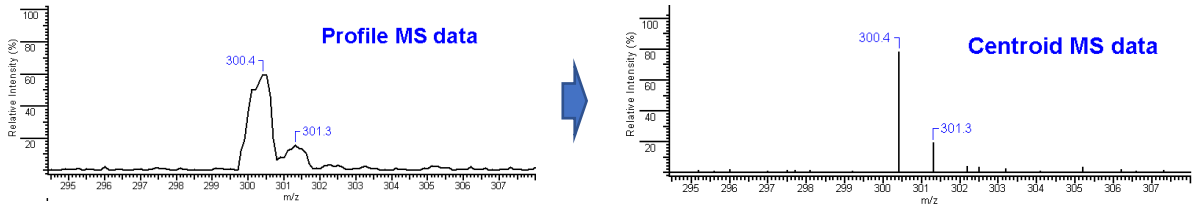
What is raw spectra (pre-)processing?

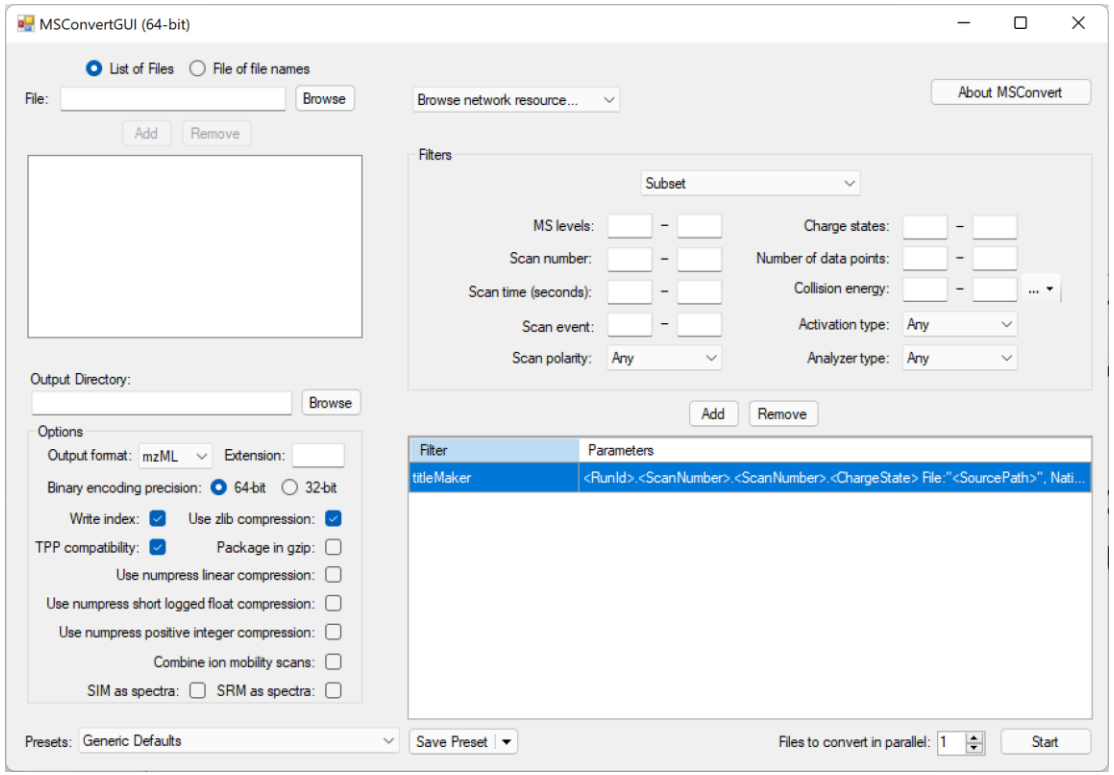
- Convert the raw spectra data from MS instrument into metabolic features (MS peaks);
- Usually contains 2~3 dimensions. For DI-MS, the raw spectra includes m/z and intensity (2D); while for the LC-MS, the raw spectra data includes m/z, retention time and intensity (3D);
- The most common vendor raw data file formats are .raw/.RAW/.wiff/.d/.D/ etc. They need to be converted into open-source format for further processing with open-source software.



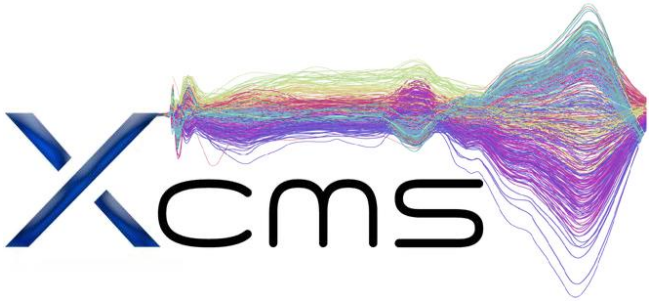
Profile or Centroid?

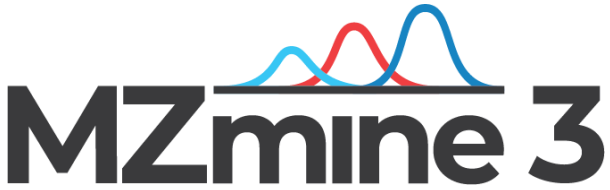
- The vendor raw spectra data is usually in profile format, which is redundant for regular LC-MS based metabolomics analysis;
- We need to convert the MS data into centroid mode to condense the Gaussian Profile peaks into centroids.
- Open-source formats (.mzML/ etc.)..

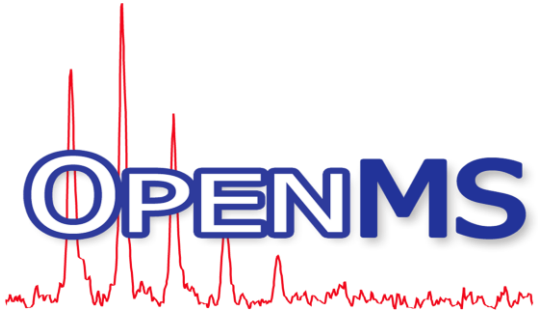




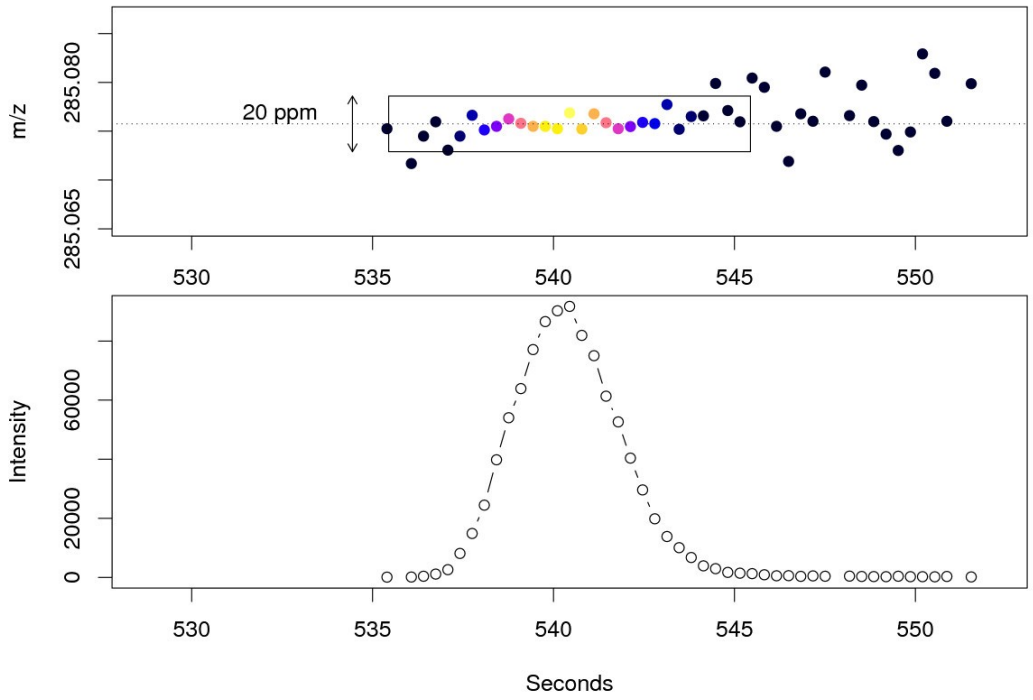
Open-source Software for raw spectra processing..

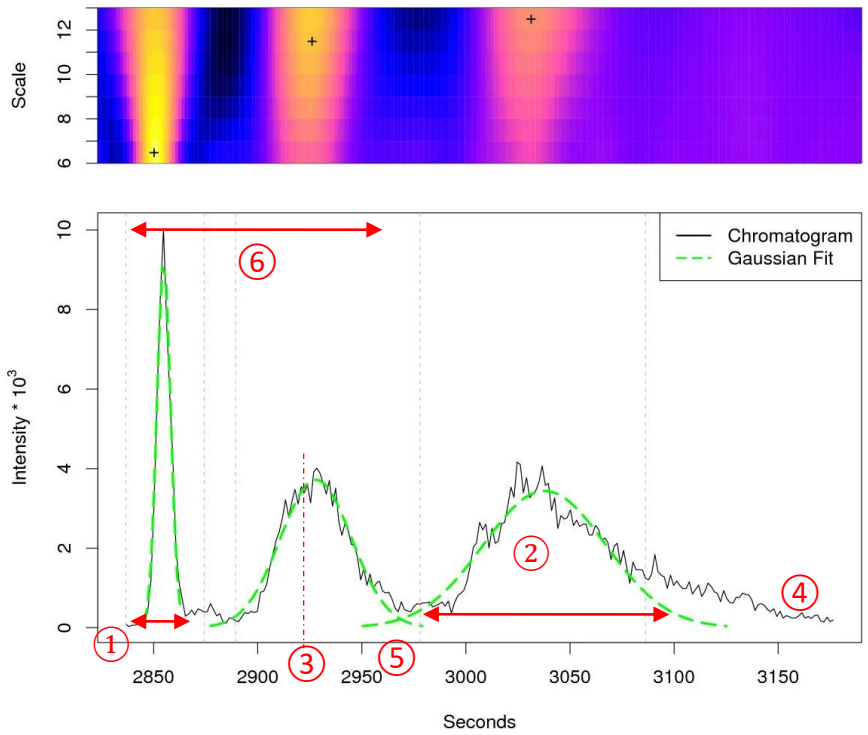




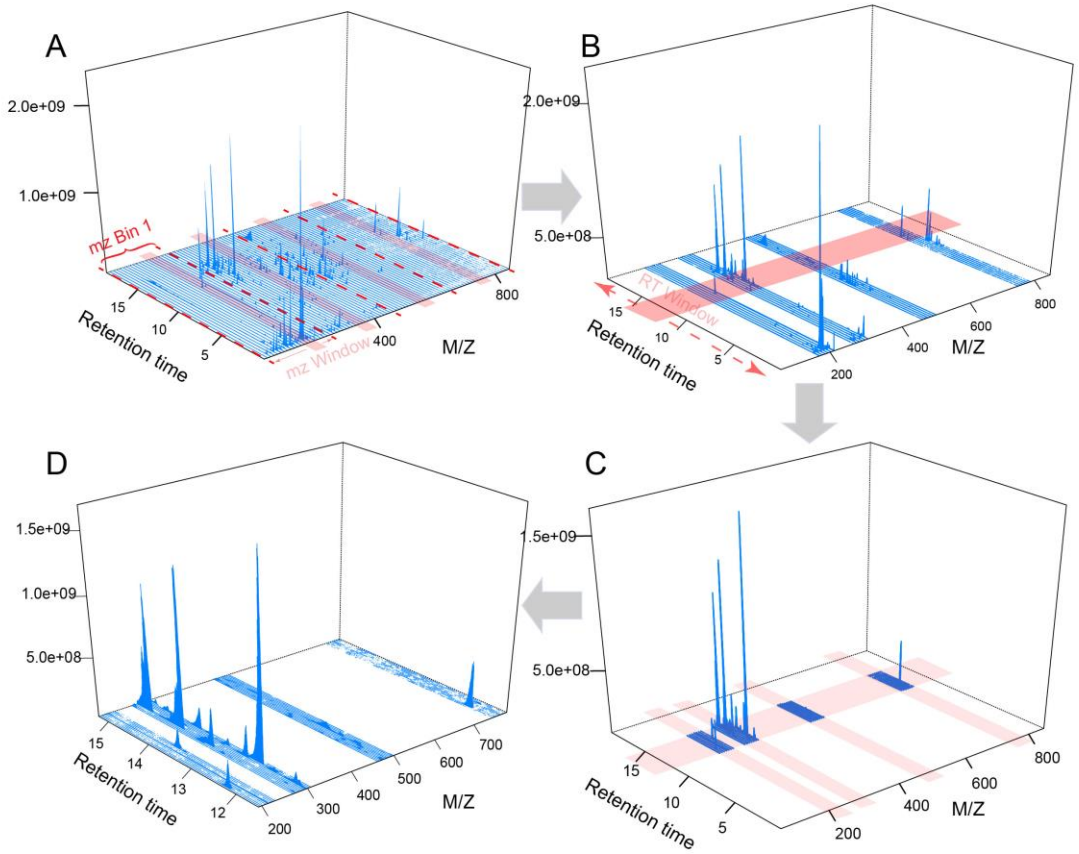


centWave


m/z
285.080
285.065
20 ppm
530
535
540
545
550
Intensity
60000
20000
0
Seconds


Scale
12
10
8
6
Chromatogram
Gaussian Fit
Intensity * 10^3
10
8
6
4
2
0
2850
2900
2950
3000
3050
3100
3150
Seconds
1
2
3
4
5
6

ROI Extraction



- Data-driven ROI extraction;
- Regions with high abundance of MS signals;
- Both low intensity peaks as well as high intensity peaks will be retained;

DoE-based Parameter Optimization

DoE -- Central composite design

44 runs

Order	Peakwith_min	Peakwith_max	mzdiff	snthresh	bw
1	-1	-1	-1	-1	-1
2	1	-1	-1	-1	-1
3	-1	1	-1	-1	-1
...	...	...	...	...	...
43	0	0	0	0	1
44	0	0	0	0	0

3 level for every parameters (-1, 0, 1)

Relative reliable peaks ratio (identified by their isotopes)

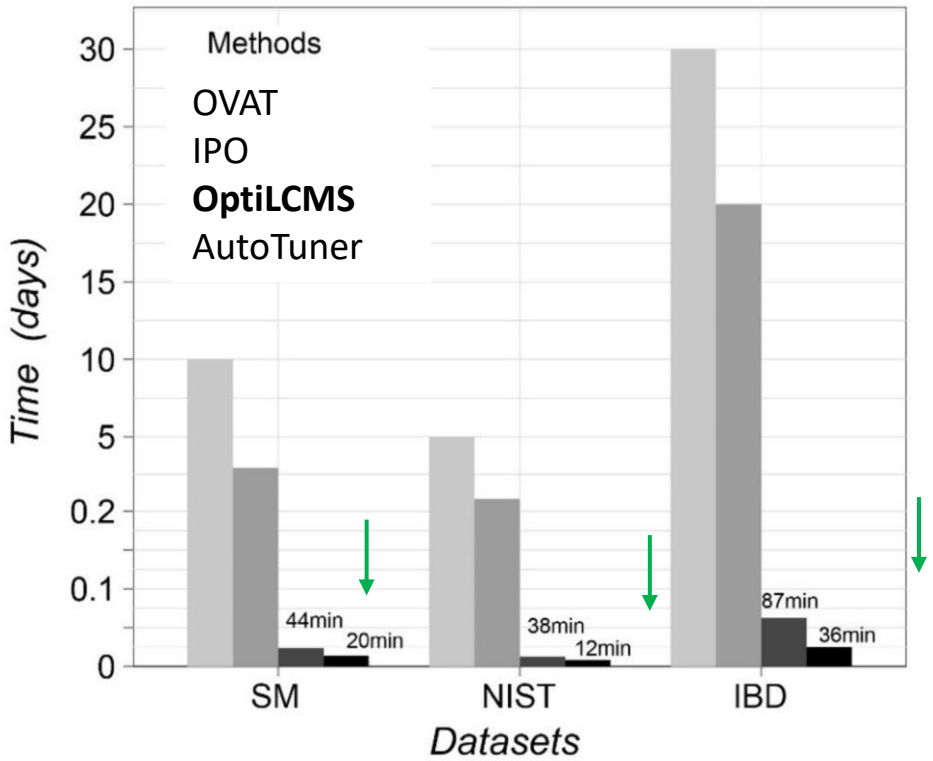
A co-efficient describing the stability of a grouped feature

$$QS = \frac{RP^{3/2}}{'all\ peaks' - LIP} * GR^2 * QcoE$$

Gaussian peaks ratio

- The most important parameters are evaluated with 44 DoE runs
- Instead of $3^8 = 6561$ one-variable-at-a-time runs.

Performance Evaluation - Speed



+ 3 datasets: Standard Mixture (SM), NIST-SRM 1950 and IBD data from iHMP2.

Pang, Z.; Chong, J.; Li, S.; Xia, J. MetaboAnalystR 3.0: Toward an Optimized Workflow for Global Metabolomics. *Metabolites* **2020**, *10*, 186

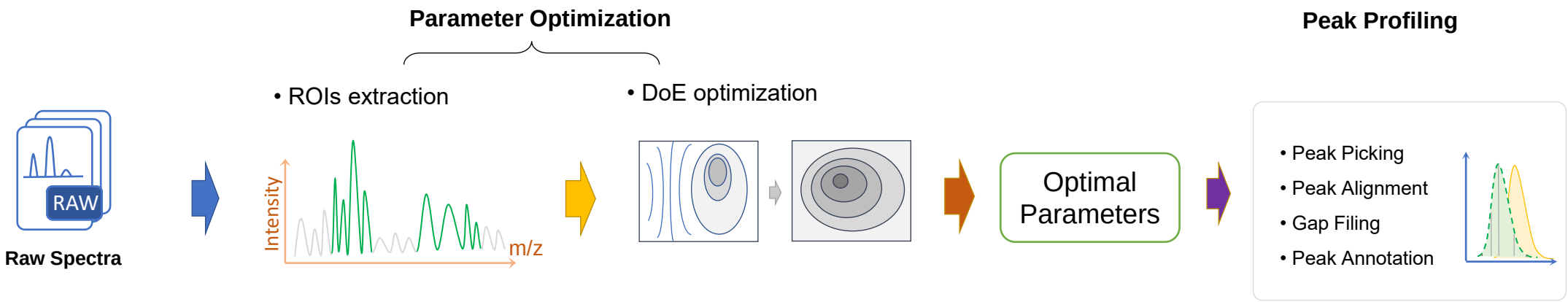
Bench markings

	Default	Optimized
Total peaks	2,492	2,423
Isotopes / Adducts	667 (26.8%)	1,112 (45.9%)
Formula Assigned	663	762
Potential compounds	1,085	1,692
Variance (PC1 + PC2)	37%	50%
Significant peaks	855	1,091

	Default	Optimized
Total peaks	4,344	5,113 (+ 17.7%)
Isotopes	760	1,274 (+ 67.6%)
Adducts	927	1,132 (+ 22.1%)
Formulas assigned	632	687 (+ 8.7%)
Potential compound matches	1,587	1,803 (+ 13.6%)
Variance explained (PC1 + PC2)	76.5%	81.3% (+ 4.8%)

Workflow Overview

- Our optimization approach is designed to extract a region abundant with MS signals for a design of experiment (DoE)-based optimization.



metaboanalyst.ca/MetaboAnalyst/ModuleView.xhtml

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

GenomeCanada

GenomeQuébec

NIH

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

Statistical Analysis [one factor]

This module offers various commonly used statistical and machine learning methods including t-tests, ANOVA, PCA, PLS-DA and Orthogonal PLS-DA. It also provides clustering and visualization tools to create dendrograms and heatmaps as well as to classify data based on random forests and SVM.

Statistical Analysis [metadata table]

This module aims to detect associations between phenotypes and metabolomics features with considerations of other experimental factors / covariates based on general linear models coupled with PCA and heatmaps for visualization. More options are available for two-factors / time-series data.

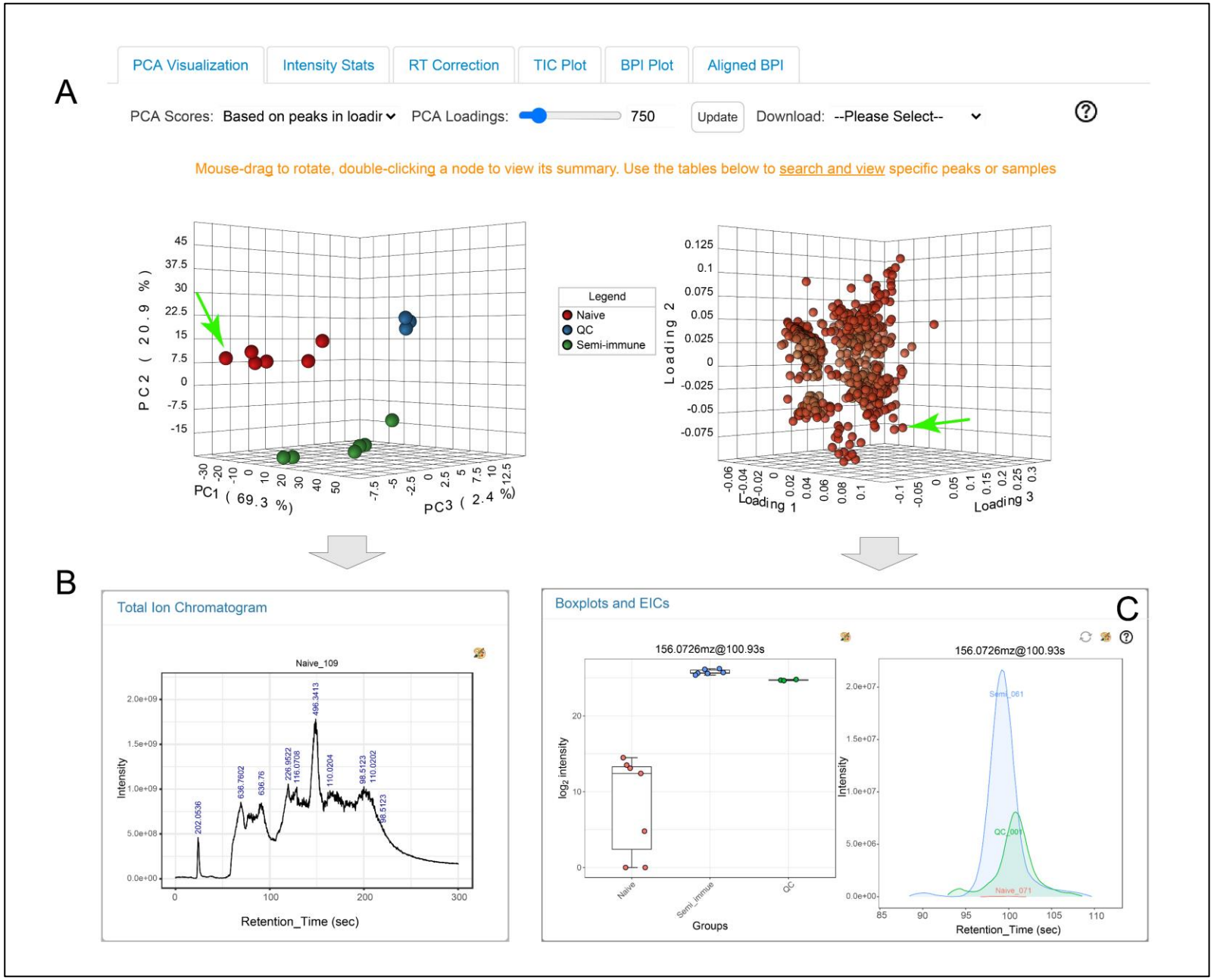
Biomarker Analysis

This module performs various biomarker analyses based on receiver operating characteristic (ROC) curves for a single or multiple biomarkers using well-established methods. It also allows users to manually specify biomarker models and perform new sample prediction.

Functional Utilities

- Raw Data Uploading (.mzML/.mzXML/.mzData/.cdf);
- Centroiding on the fly;
- Parameters Optimization (automatically);
- Peak Profiling (Peak Picking/Alignment/Gap filling);
- Peak Annotation (Adducts + Isotopes);
- Putative Compound Mapping;
- Result visualization…

Result Demo..

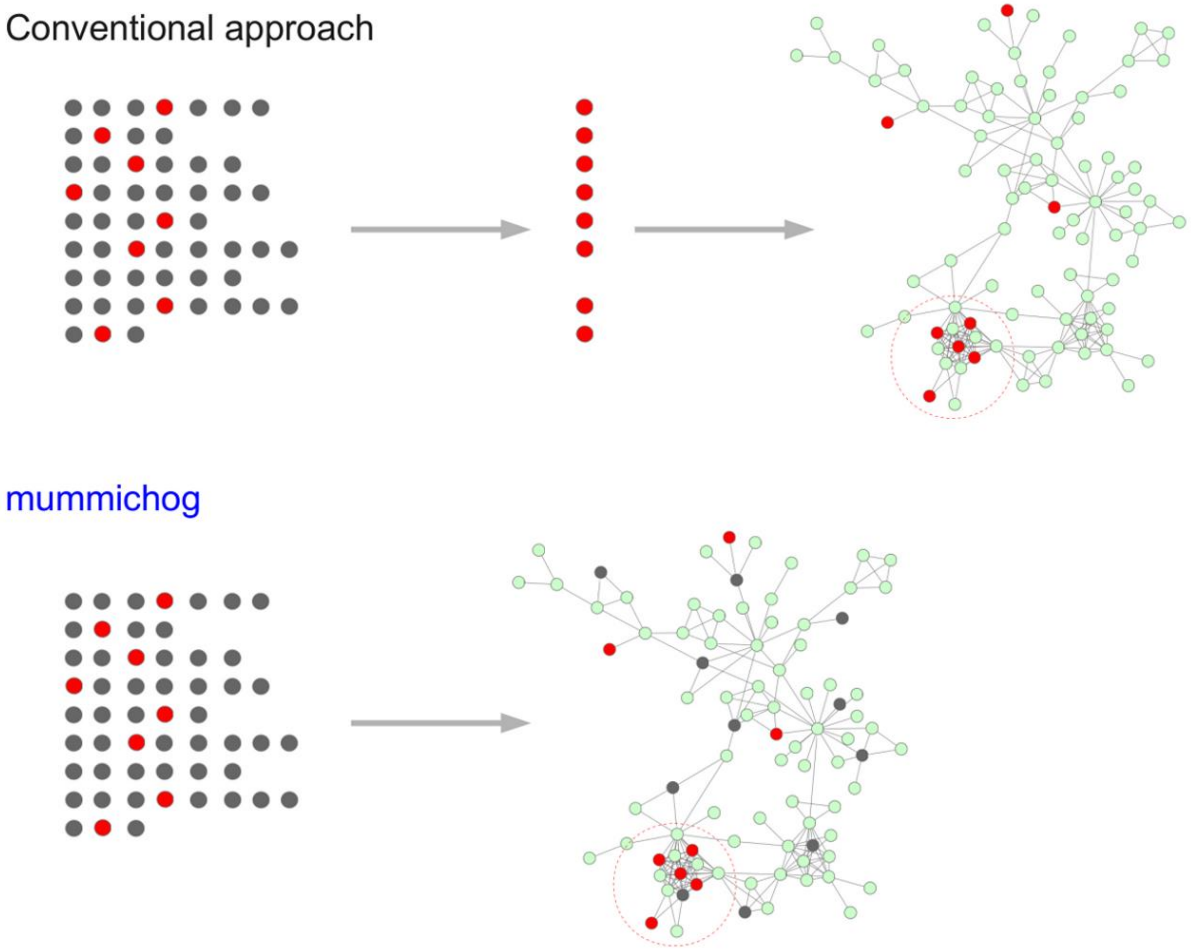

A
PCA Visualization
Intensity Stats
RT Correction
TIC Plot
BPI Plot
Aligned BPI
PCA Scores: Based on peaks in loadir
PCA Loadings: 750
Update
Download: --Please Select--
Mouse-drag to rotate, double-clicking a node to view its summary. Use the tables below to search and view specific peaks or samples
PC2 (20.9 %)
PC1 (69.3 %)
PC3 (2.4 %)
Legend
Naive
QC
Semi-immune
Loading 2
Loading 1
Loading 3
B
Total Ion Chromatogram
Naive_109
Intensity
Retention_Time (sec)
Boxplots and EICs
C
156.0726mz@100.93s
log2 intensity
Groups
Naive
Semi_immue
QC
Semi_061
QC_001
Naive_071

Functional Analysis

Functions describe the collective behavior of groups of molecules acting together to complete a task (e.g., lysine degradation).

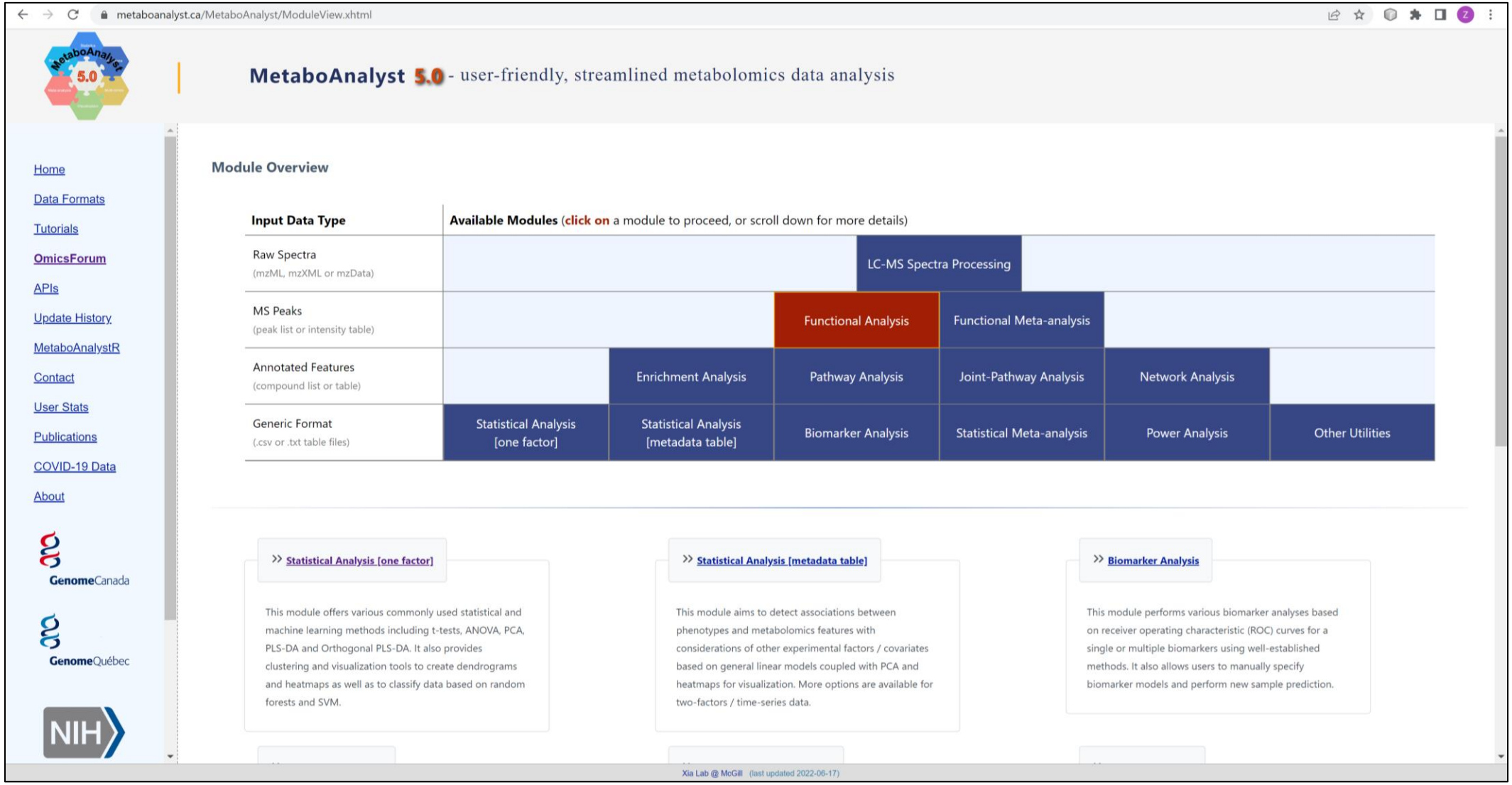
The goal of enrichment analysis is to evaluate whether the members involved in a particular tasks how more consistent behaviors (e.g.,more changes larger than normal) compared with random variations.

Mummichog provides a practical solution of one-step functional analysis, bypassing the bottleneck of upfront metabolite identification.



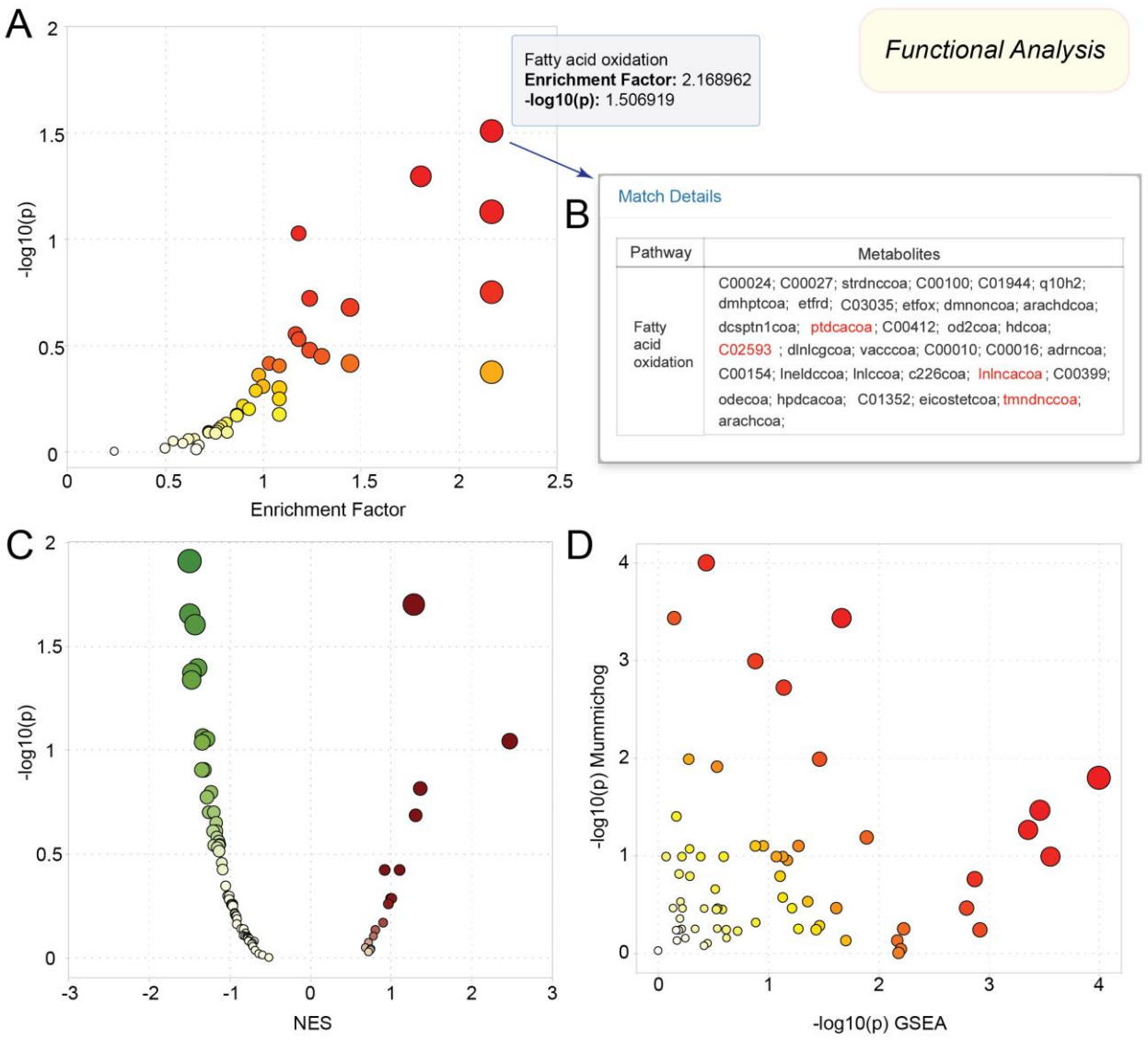
Li S, Park Y, Duraisingham S, Strobel FH, Khan N, Soltow QA, et al. (2013) Predicting Network Activity from High Throughput Metabolomics. PLoS Comput Biol 9(7): e1003123.

Implementation in MetaboAnalyst

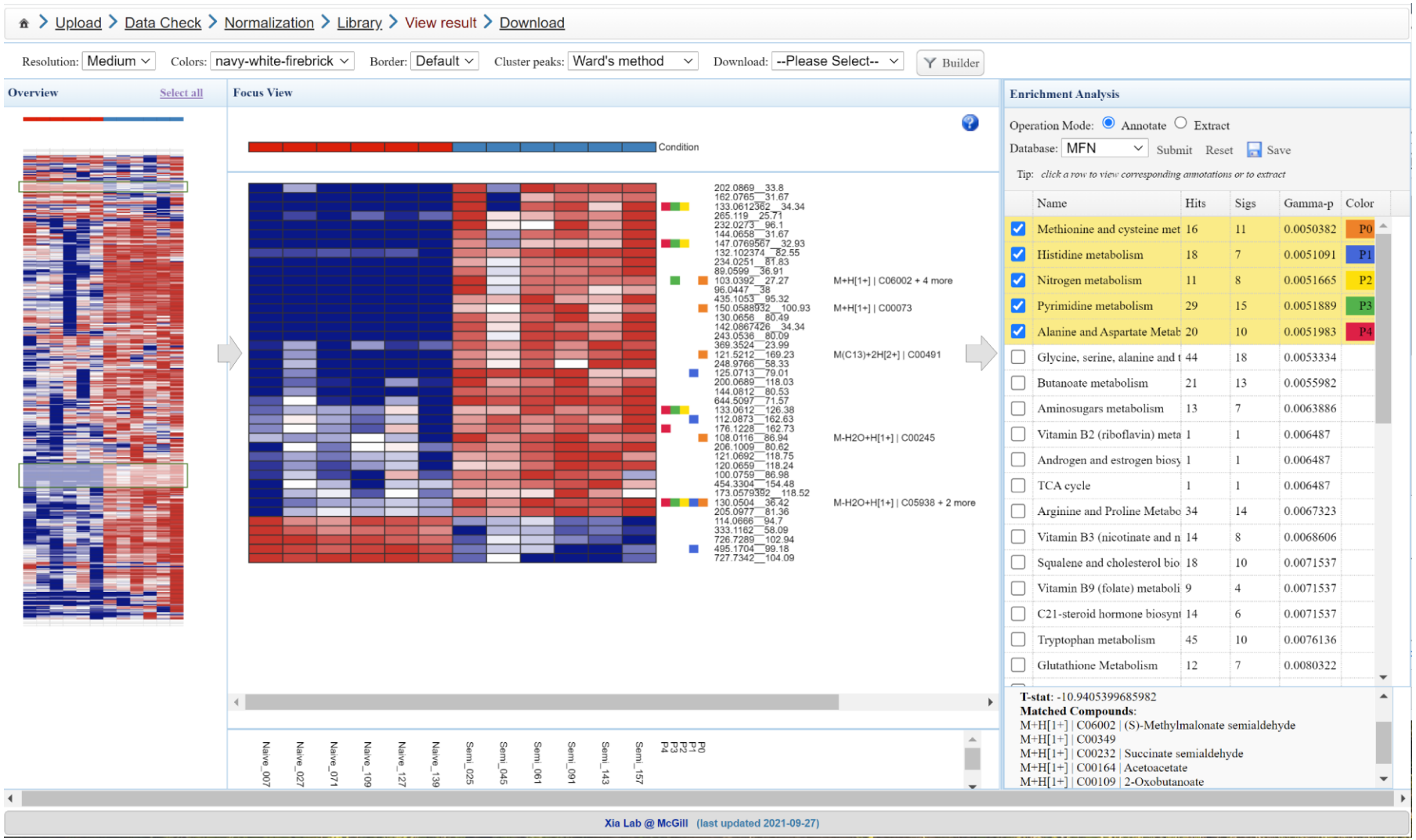


Functional Analysis Results

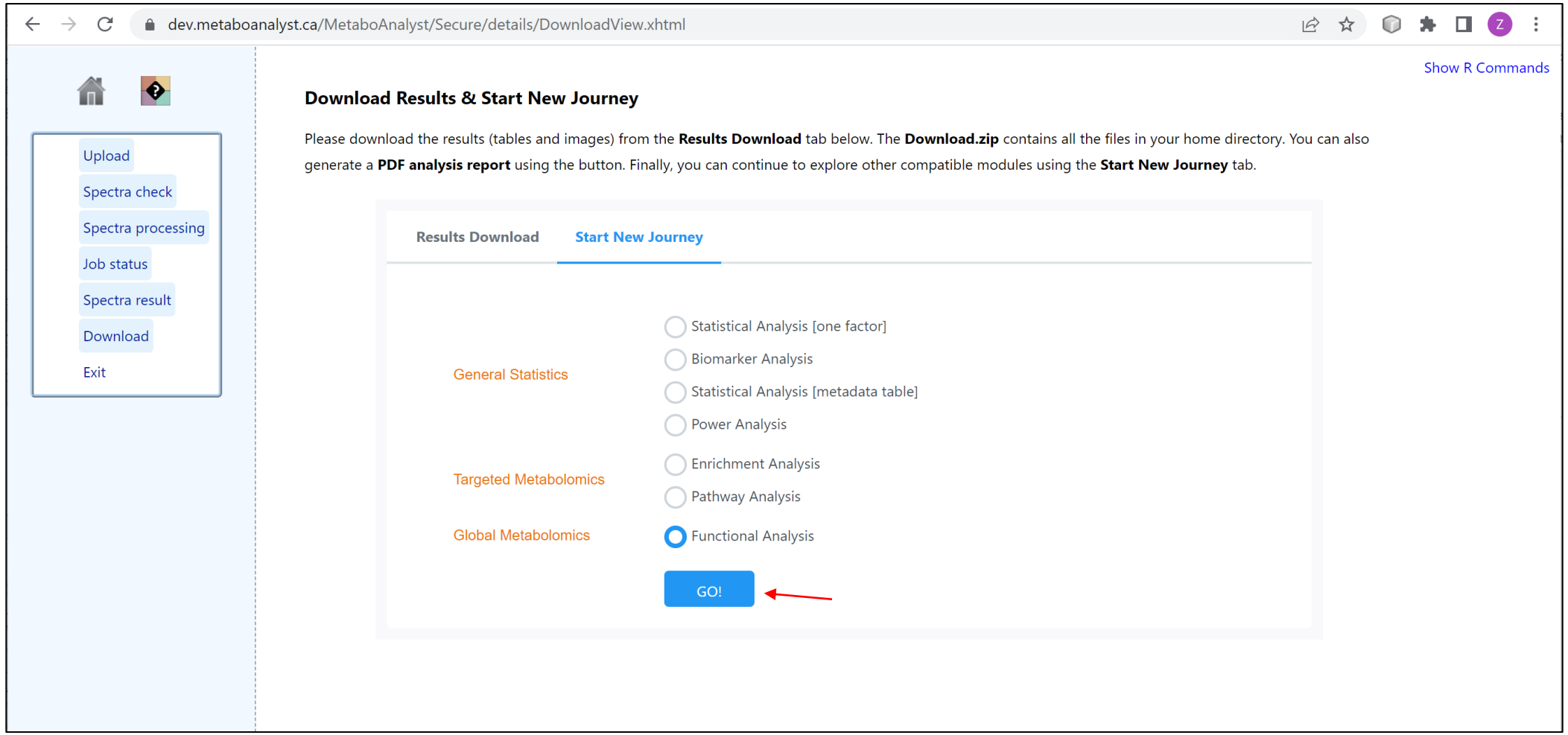
- **Mummichog**: implements an over-representation analysis (ORA) method to evaluate pathway-level enrichment based on significant peaks. Users need to specify a pre-defined cutoff.
- **GSEA**: considers the overall ranks of features without using a significance cutoff, and is claimed to be able to detect subtle and consistent changes.
- **Integrated**: combines both Mummichog and GSEA together with Fisher's Method.



Metabolic Pattern-based functional analysis



From raw spectra into biological insight



Tutorials

- **Publication: https://www.nature.com/articles/s41596-022-00710-w**
- **Or our manuscript: https://www.dropbox.com/s/7184c4dheeiiz2p/NP-MetaboAnalyst-2022.pdf?dl=0**
  - ➔ **Stage 1: LC–HRMS raw spectra processing**
  - ➔ **Stage 2: functional analysis of LC–HRMS peaks**

Questions?

- **https://www.omicsforum.ca/**
- **If your question is not covered, please create a new topic – we will try to answer them in the coming days.**

Caution:

1. For raw spectra processing, you are strongly encouraged to use 1st example rather than the 2nd one to avoid waiting in queue for learning purpose;
2. Avoid downloading and uploading any example raw spectra data due the limited bandwidth.
3. Default MetaboAnalyst includes **www.metaboanalyst.ca** and **dev.metaboanalyst.ca**, please use backup node **genap.metaboanalyst.ca** ONLY if the defaults are not accessible.