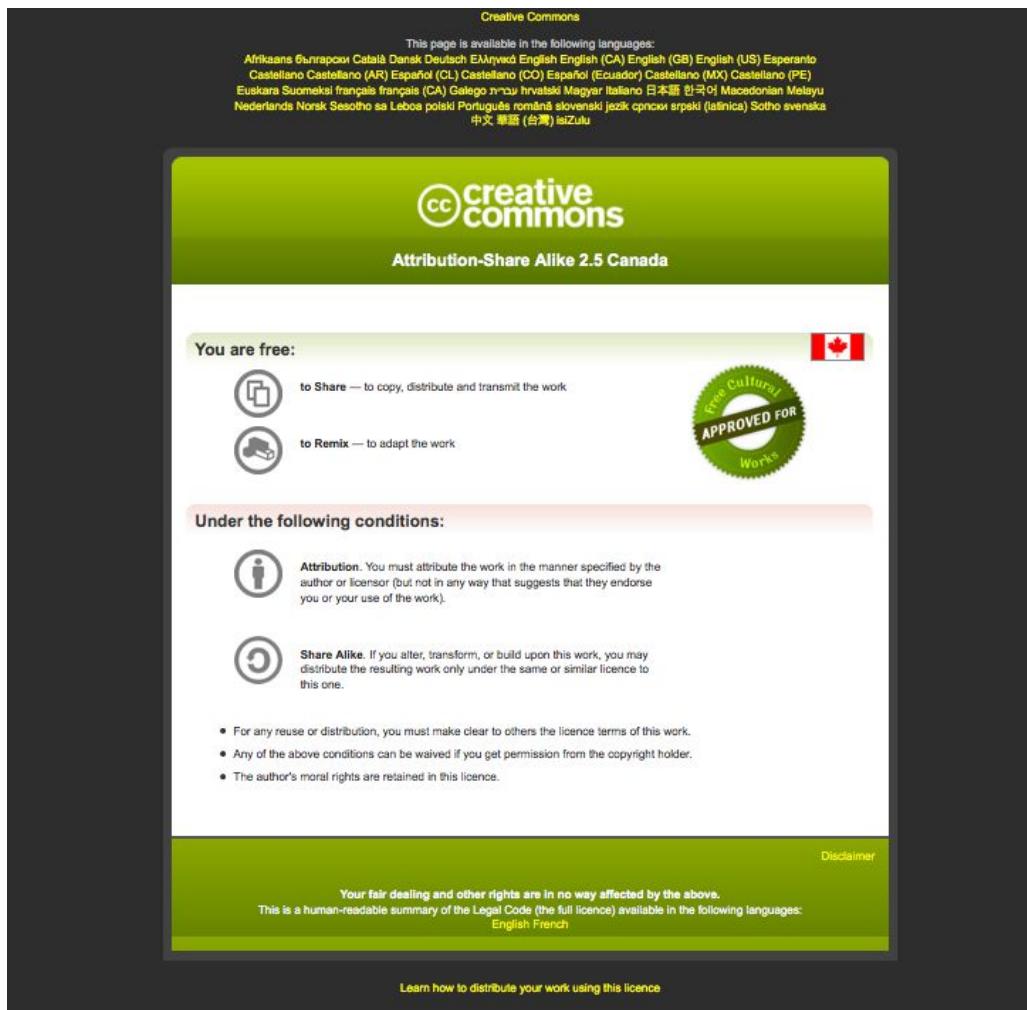




Canadian Bioinformatics Workshops

www.bioinformatics.ca

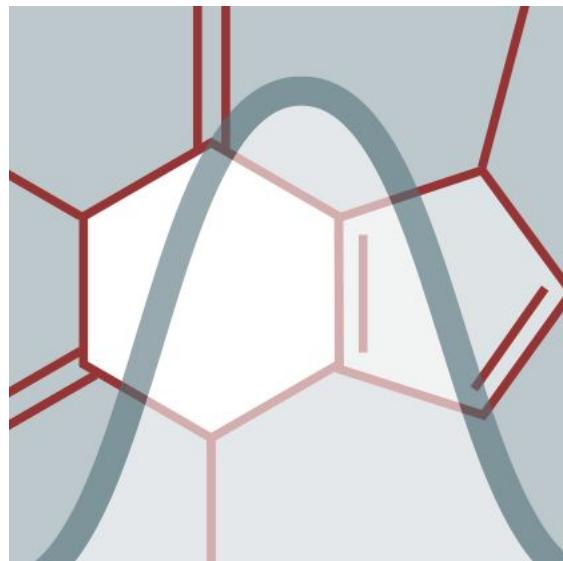
bioinformaticsdotca.github.io



Data analytics for untargeted metabolomics



Jianguo (Jeff) Xia
Metabolomics Analysis
July 6-7, 2023



McGill

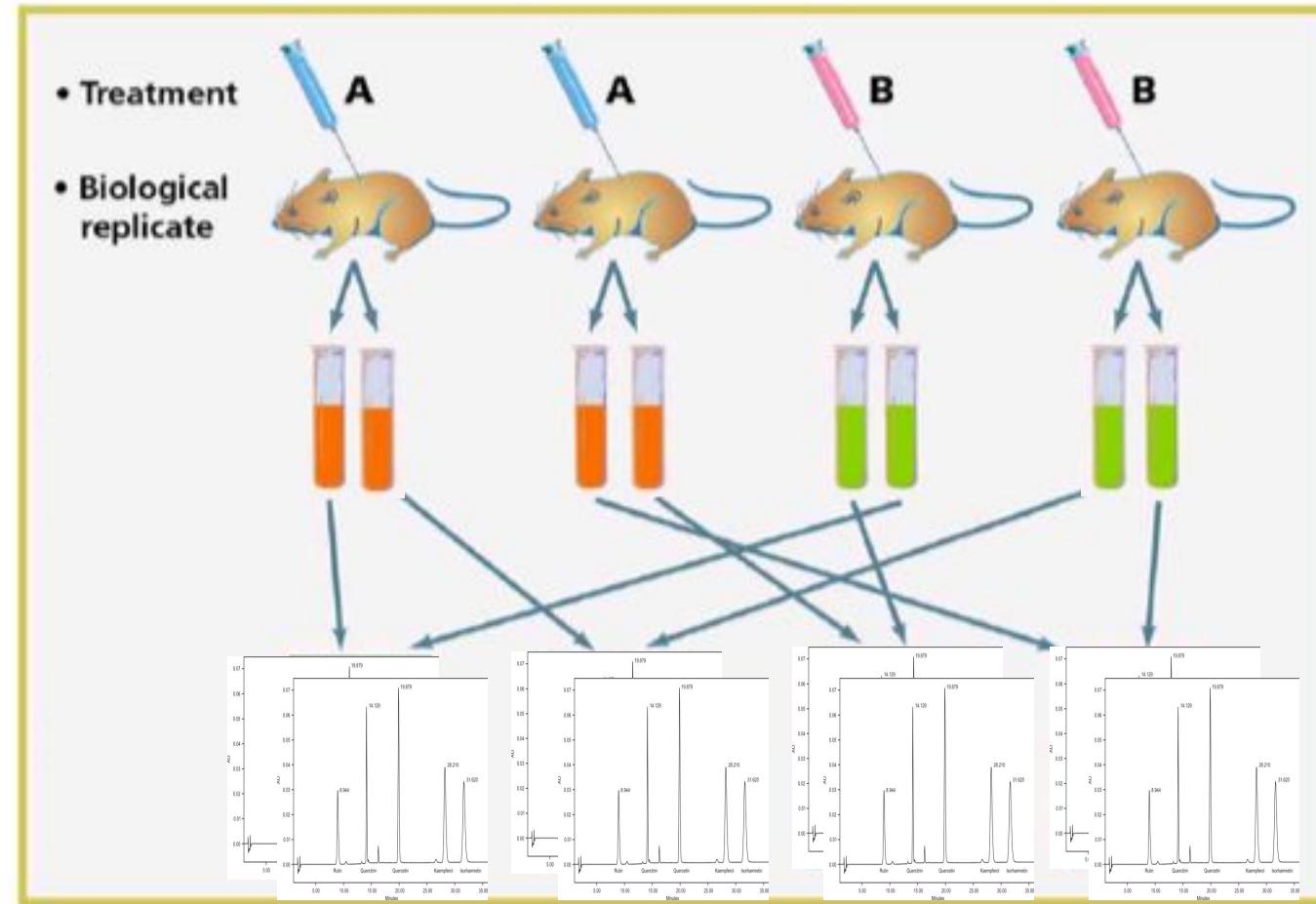
Learning Objectives

1. Become familiar with basic steps in LC-MS spectral processing
2. Become familiar with enrichment analysis for quantitative metabolomics data
3. Become familiar with enrichment analysis from LC-MS untargeted data

A Typical Metabolomics Experiment

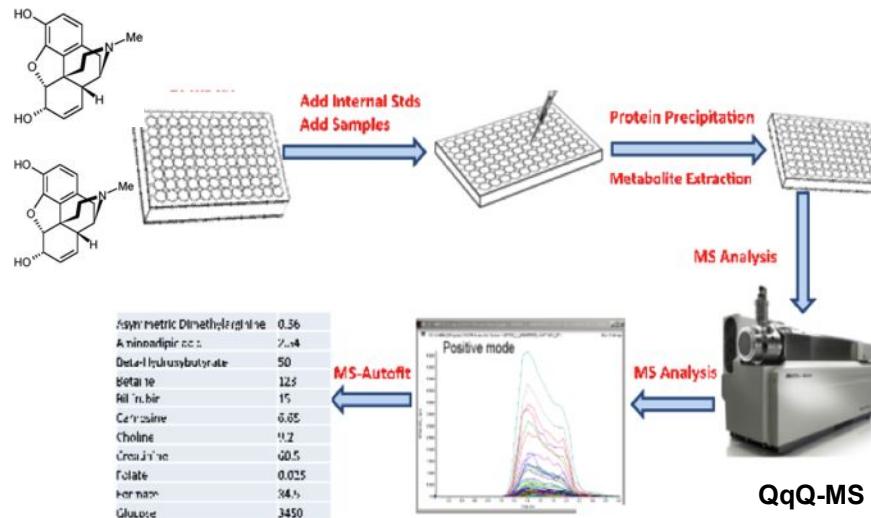
Biological
replicates

Technical
replicates

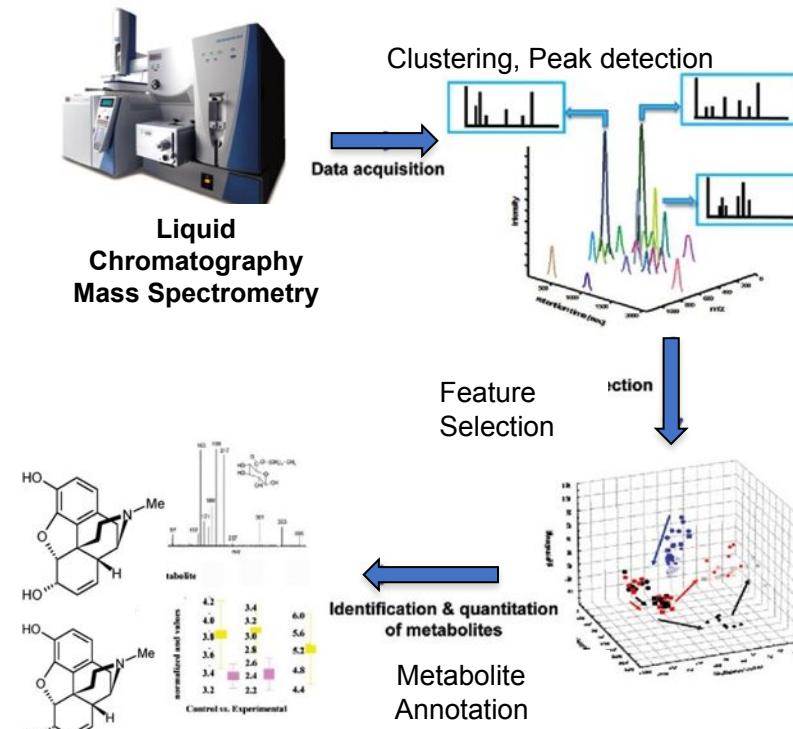


Data generation: two routes to metabolomics

Targeted / Quantitative



Untargeted / Global

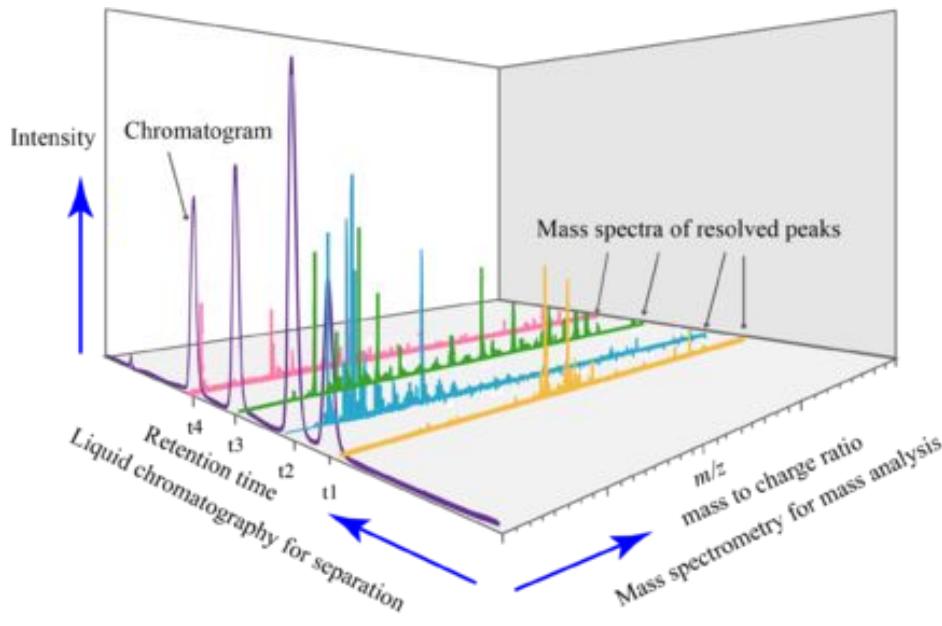


Common Tasks

- Purpose: to convert various raw data forms into data matrices suitable for statistical / functional analysis
- Supported data formats
 - Concentration tables (Targeted Analysis)
 - Spectral bins (Untargeted)
 - Peak lists (Untargeted)
 - Raw spectra (Untargeted, in mzML/mzXML/mzData/CDF)

LC-MS Raw Spectral Data Processing

MS Features



Raw spectral data is large and noisy

Features: LC-MS signals (m/z , RT, intensity) produced by the same molecular ion

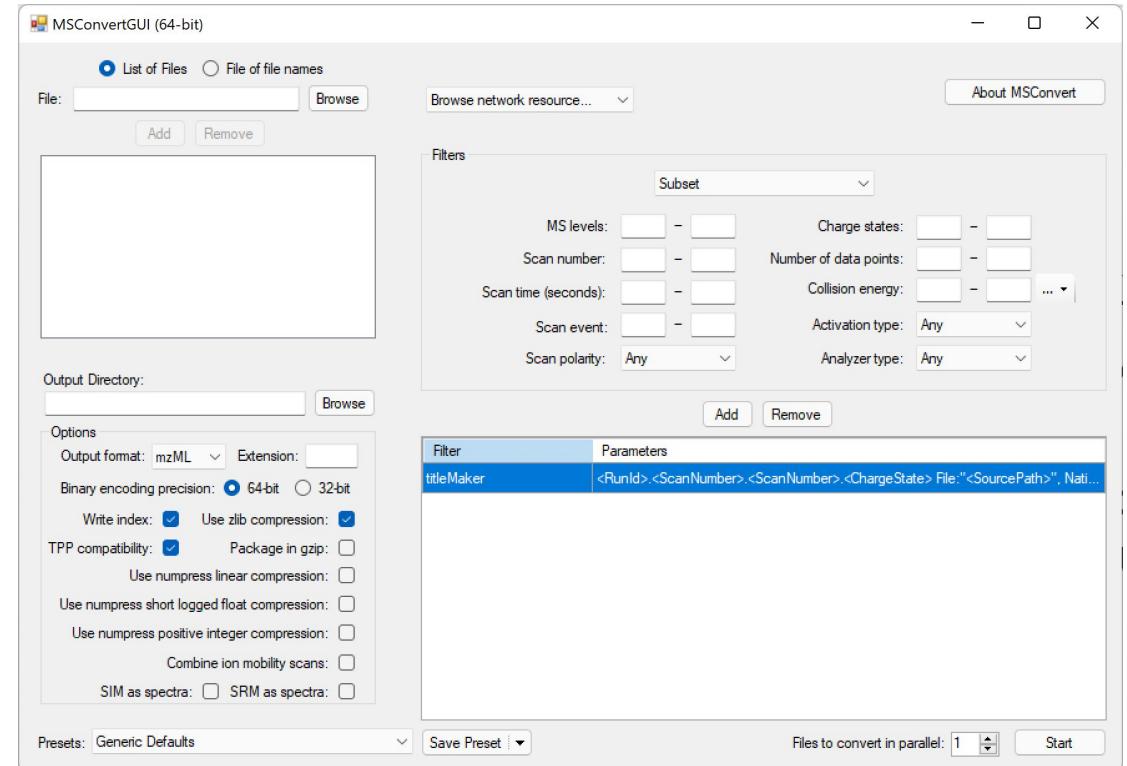
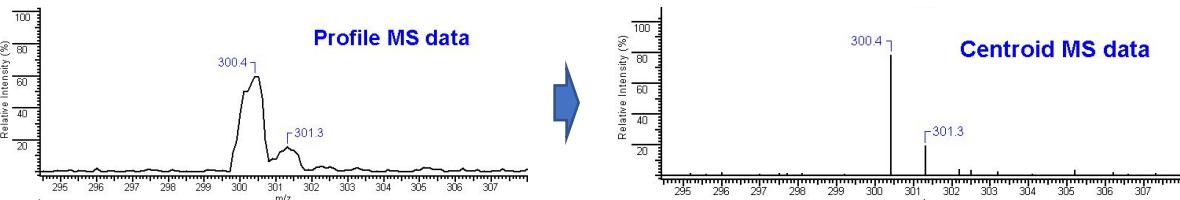
Single compound can give rise to multiple of mass signals (adducts, fragments, isotopes)

Processing MS data

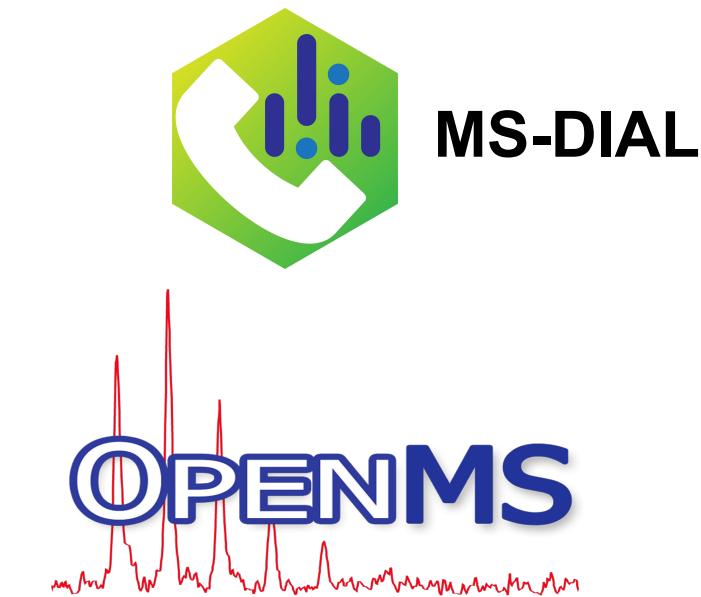
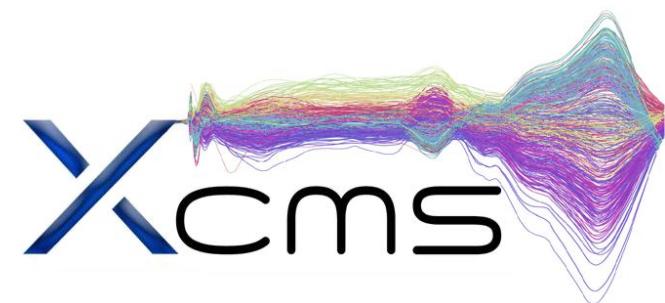
- Identify, quantify, and align all possible features (peaks) across samples
- Output: a table of features (RT, m/z) with their quantitative information for subsequent statistical analysis

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.)..

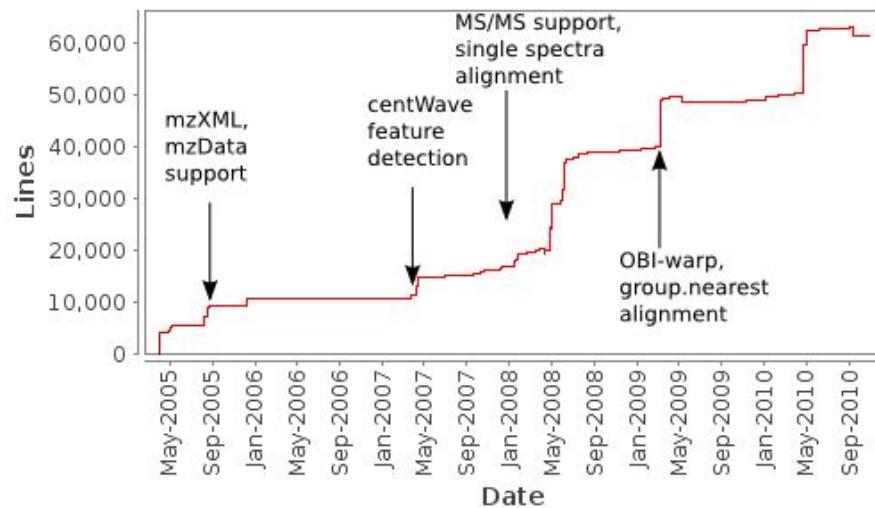


Open-source software for spectra processing..



Why XCMS?

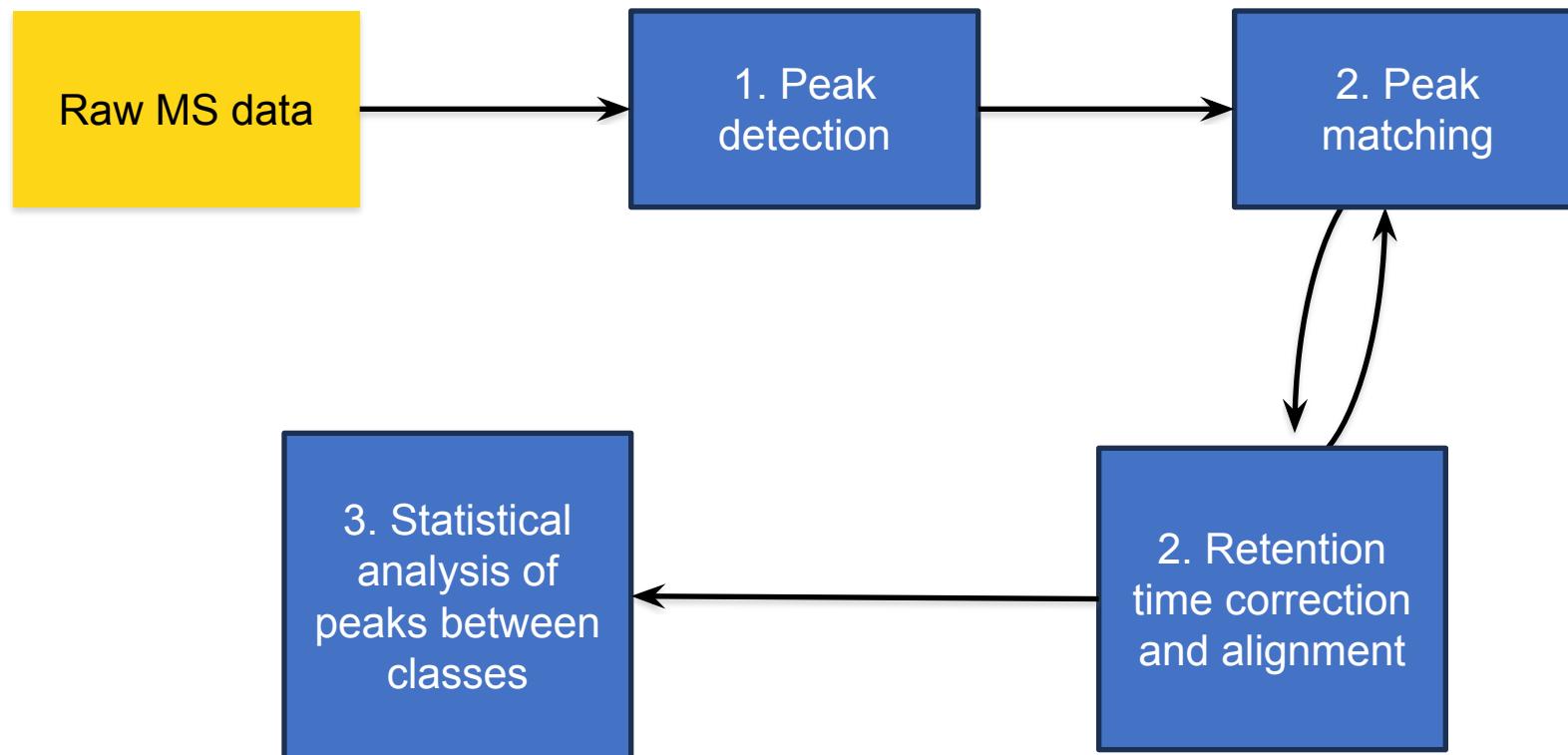
- First open-source software (R package) developed in 2006 by Siuzdak lab to process LC/GC-MS spectra
- Permits batch processing
- Widely used by metabolomics community for untargeted LC-MS based metabolomics



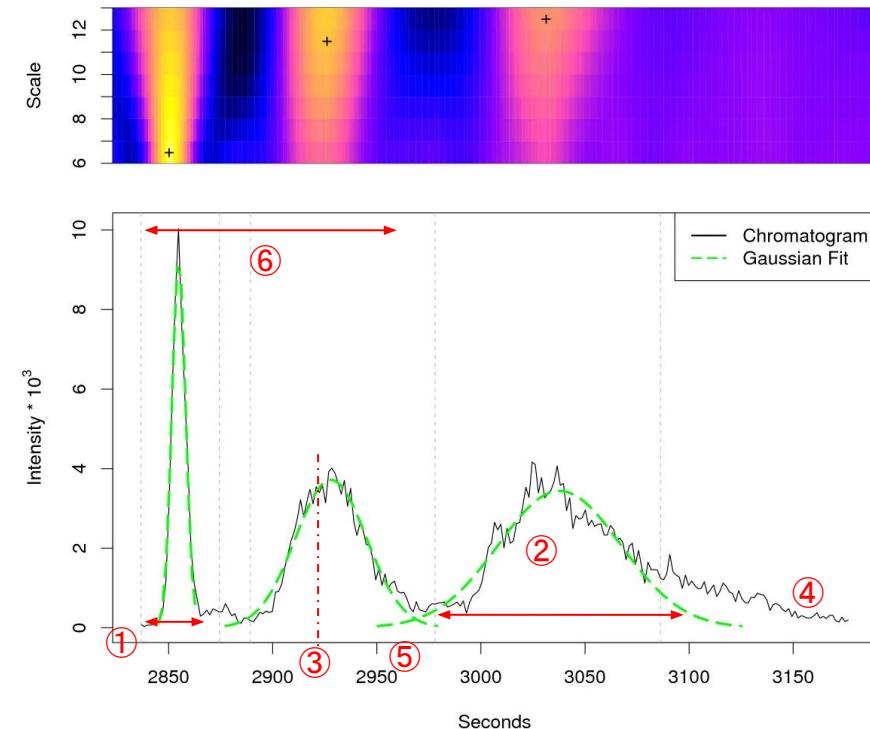
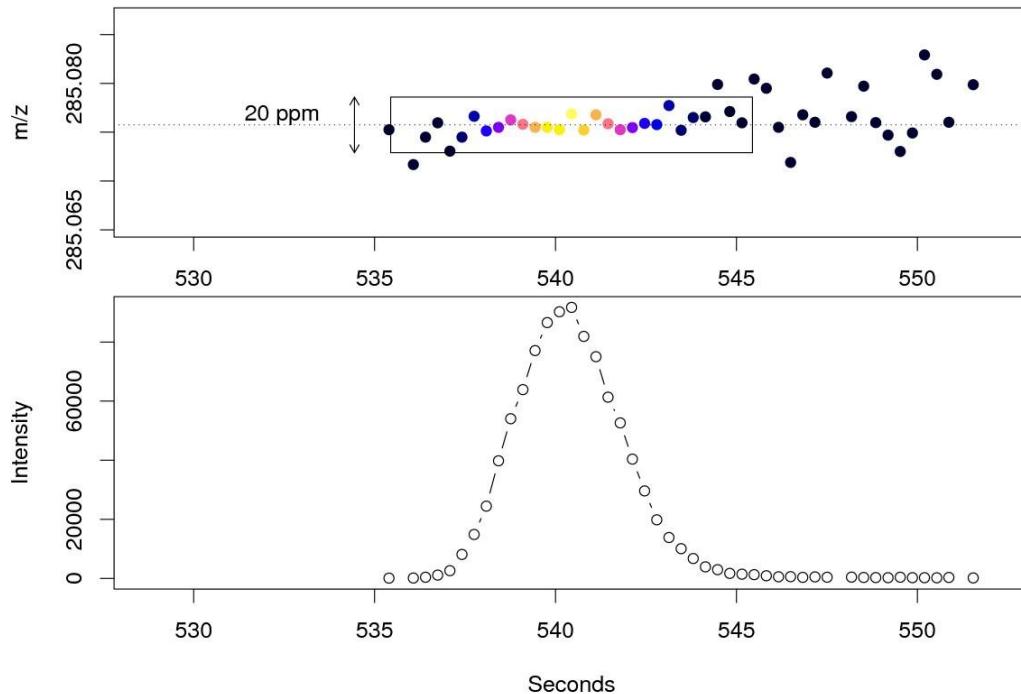
- LC-MS, MS/MS, MRM

<http://metablogomics.blogspot.com/2010/11/short-history-of-xcms.html>

XCMS Flowchart



The centWave algorithm

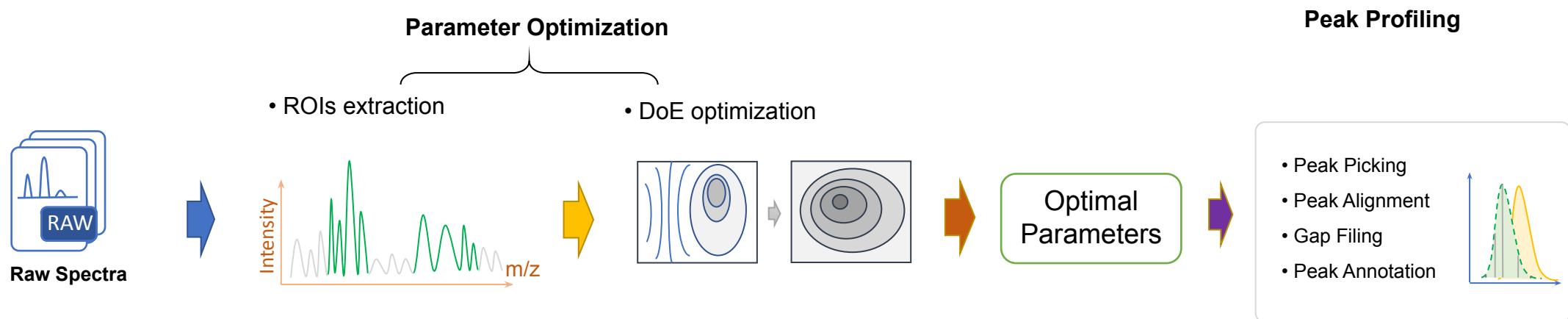


Many parameters need to be manually tuned to work well for your data
Need to be familiar with instrument and analytics

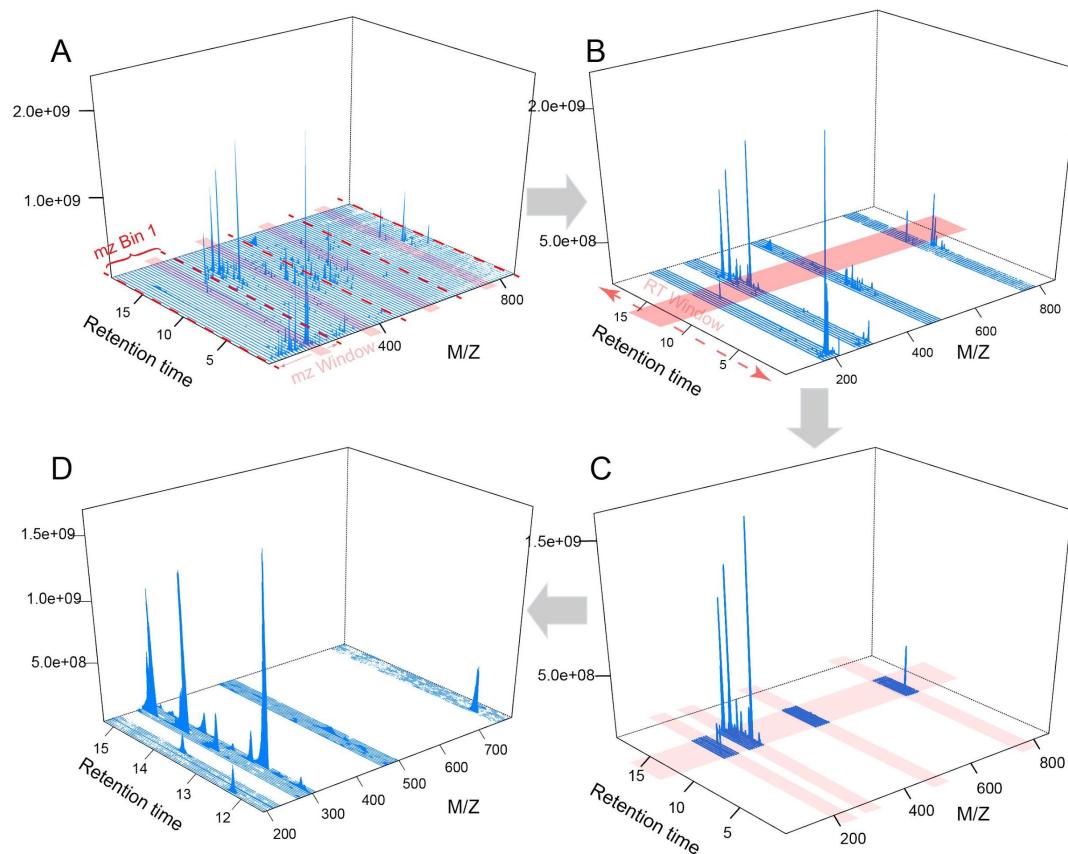
Can we make this automatic?

Auto-optimization workflow

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



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

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

44 runs

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

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

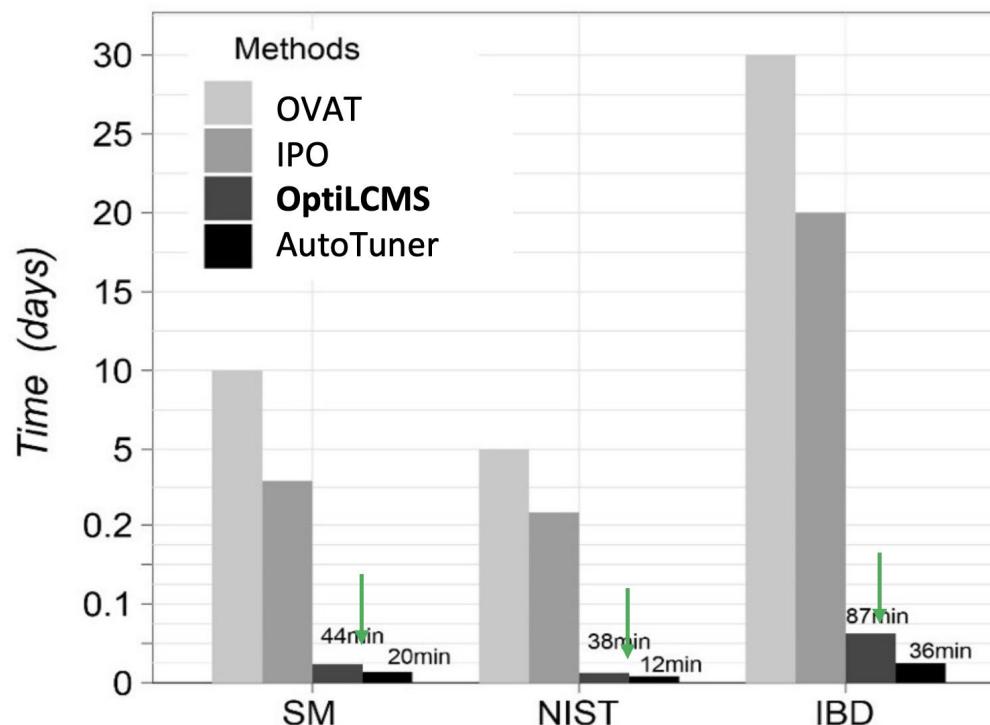
Relative reliable peaks ratio
(identified by their isotopes)

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

A co-efficient describing the stability
of a grouped feature

Gaussian peaks ratio

Performance Evaluation – fast & better results



	Default	Optimized
Total peaks	4,344	5,113 (+18%)
Isotopes	760	1,274 (+68%)
Adducts	927	1,132 (+22%)
Formulas assigned	632	687 (+8.7%)
Potential matches	1,587	1,803 (+14%)
Variance explained	76.5%	81.3% (+5%)

Three 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

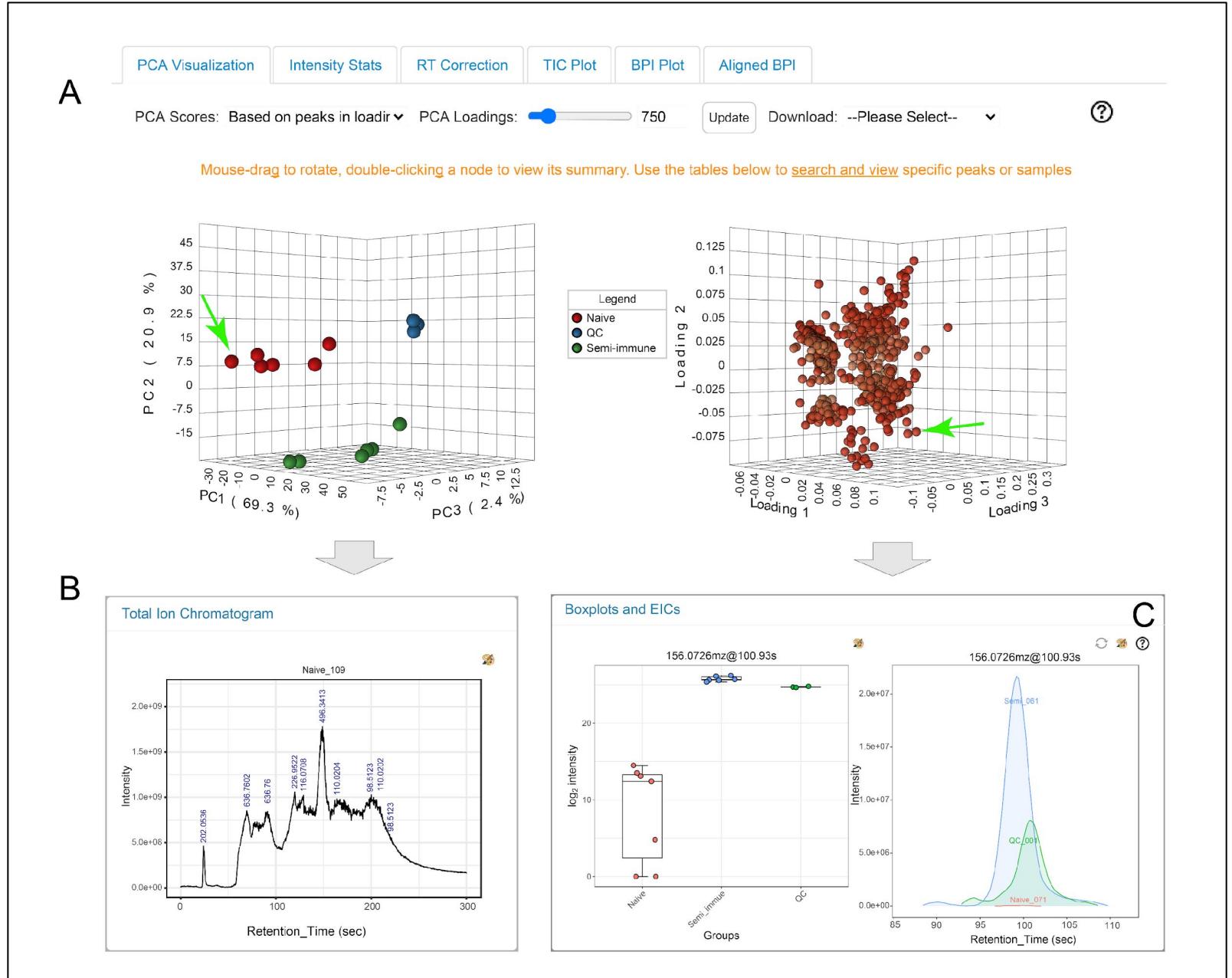
Raw LC-MS spectra processing

Input Data Type	Available Modules (click on a module to proceed. Scroll down for more details)					
Raw Spectra (mzML, mzXML or mzData)					LC-MS Spectral 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	Biomarker Analysis	Time-series/Two-factor Analysis	Statistical Meta-analysis	Power Analysis	Other Utilities



- From raw LC-MS spectra to peak abundance table
- Accept mzML, mzXML, CDF or mzData formats

Results



TIC: Total Ion Chromatography; EIC, Extracted Ion Chromatography.

After spectral processing, you can now ...

Download Results & Start New Journey

Please download the results (tables and images) from the **Results Download** tab below. The **Download.zip** contains all the files in your home directory. You can also generate a **PDF analysis report** using the button. Finally, you can continue to explore other compatible modules using the **Start New Journey** tab.

The screenshot shows a user interface for post-spectral processing analysis. At the top, there are two tabs: "Results Download" and "Start New Journey", with "Start New Journey" being highlighted with a blue border. Below the tabs, there are three main sections: "General Statistics", "Targeted Metabolomics", and "Global Metabolomics". Each section lists several analysis options, each preceded by a radio button. The "General Statistics" section includes: Statistical Analysis [one factor] (selected), Biomarker Analysis, Statistical Analysis [metadata table], and Power Analysis. The "Targeted Metabolomics" section includes: Enrichment Analysis and Pathway Analysis. The "Global Metabolomics" section includes: Functional Analysis. At the bottom of the interface is a large blue "GO!" button.

Category	Analysis Options
General Statistics	<input checked="" type="radio"/> Statistical Analysis [one factor] <input type="radio"/> Biomarker Analysis <input type="radio"/> Statistical Analysis [metadata table] <input type="radio"/> Power Analysis
Targeted Metabolomics	<input type="radio"/> Enrichment Analysis <input type="radio"/> Pathway Analysis
Global Metabolomics	<input type="radio"/> Functional Analysis

GO!

New-generation spectra processing: **asari**

1 **Trackable and scalable LC-MS metabolomics data processing using asari**

2

3 Shuzhao Li*, Amnah Siddiq, Maheshwor Thapa, Shujian Zheng

4 Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA

5

6 *Corresponding author, E-mail: shuzhao.li@jax.org

7

8 **Metabolomics holds the promise to study biological systems**

9 **comprehensively in biological systems.** However, the use of mass spectrometry) has been limited by the complexity of the samples.

10 **challenges still exist in the analysis of metabolomics data from complex experiments into metabolite identification.**

11 **current software tools. We present Asari, a new tool for LC-MS**

12 **metabolomics data processing. Asari is designed with a set of new algorithmic**

13 **framework and data structures, and all steps are explicitly trackable. It offers substantial**

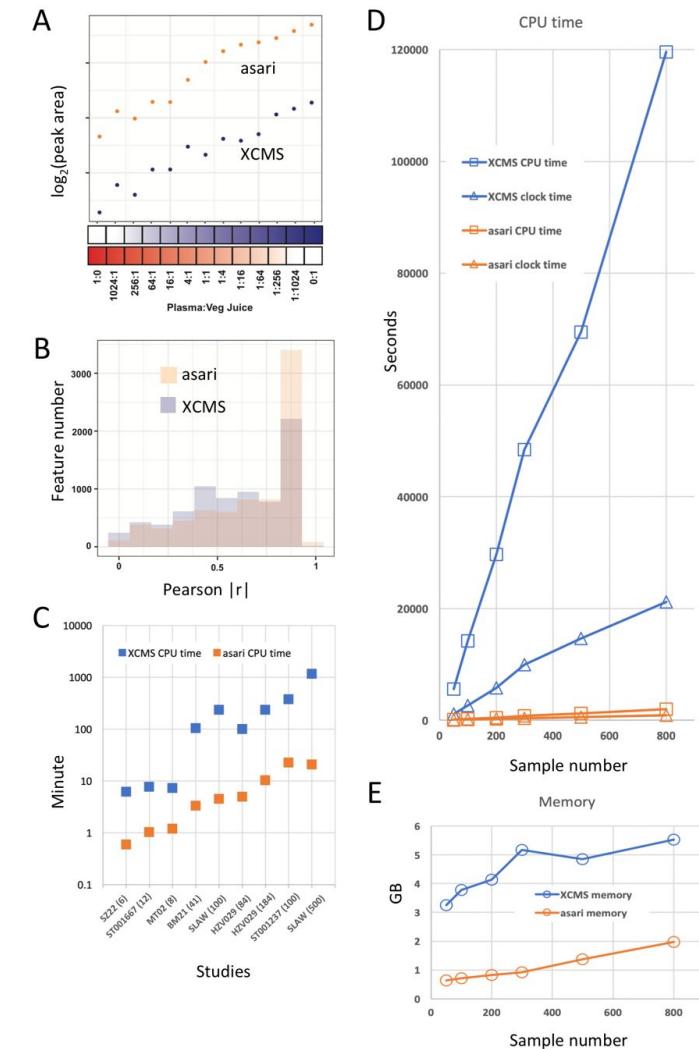
14 **improvement of computational performance over current tools, and is highly scalable.**

BioRxiv <https://doi.org/10.1101/2022.06.10.495665>

asari vs. XCMS

=

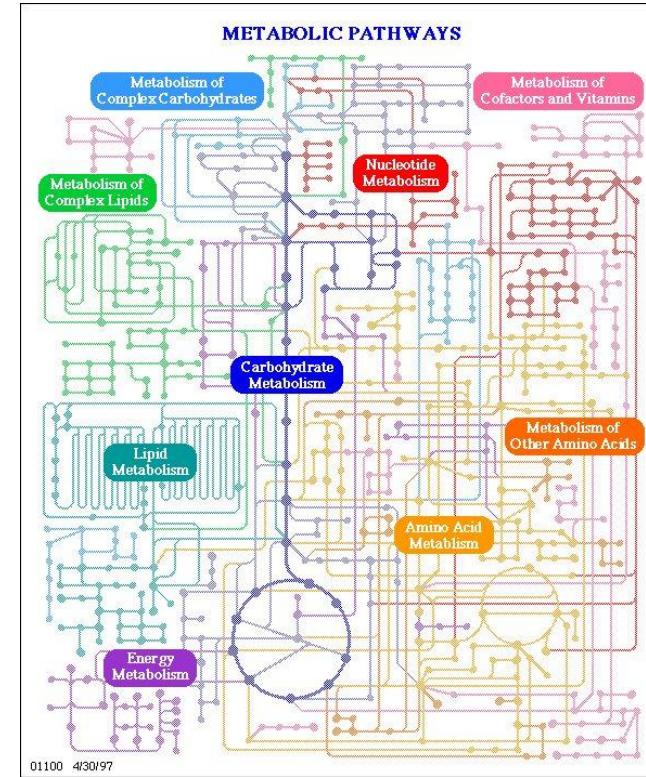
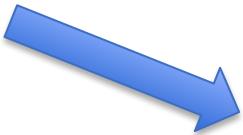
kallisto vs. HISAT



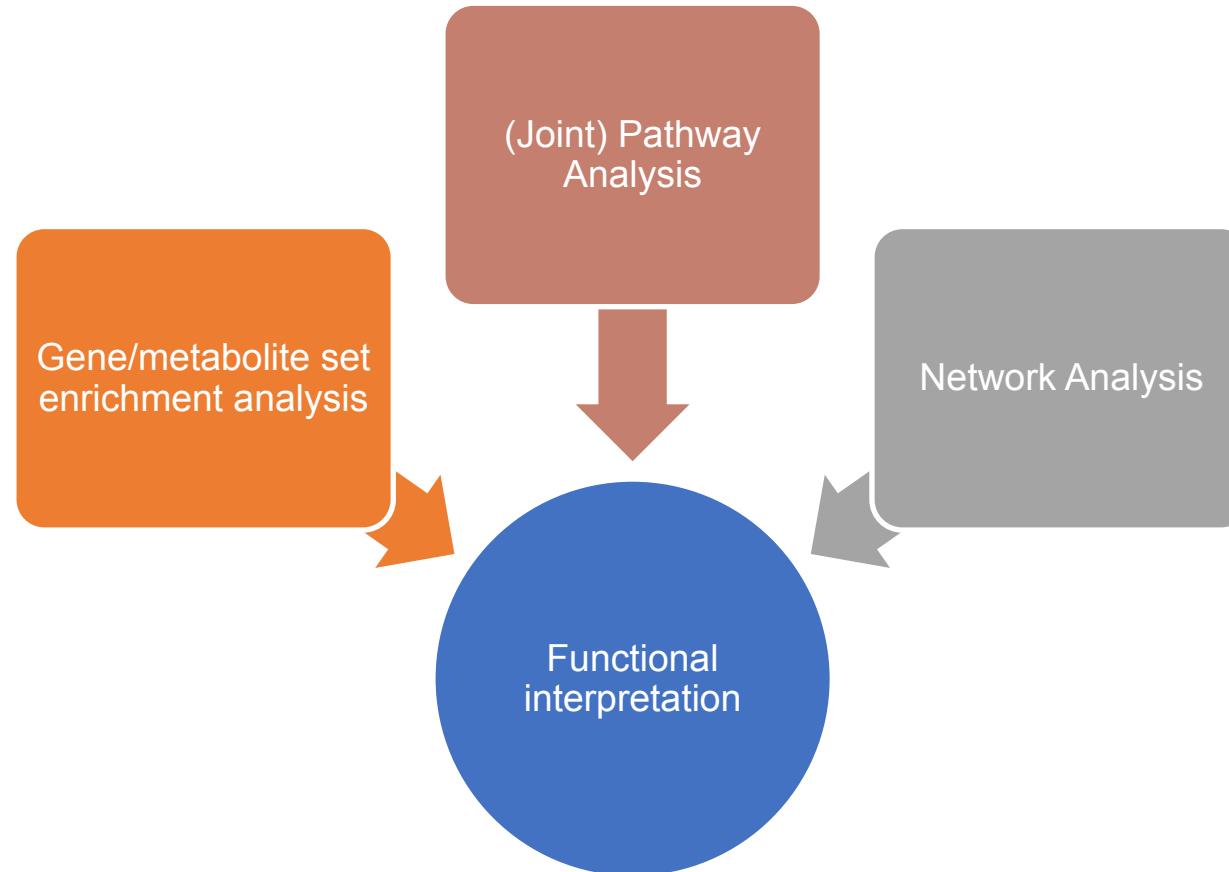
From compounds / peaks to functions

Compound	Retention Time (min)	Conc. in Urine (μ M)
Dns-o-phospho-L-serine	0.92	<DL *
Dns-o-phospho-L-tyrosine	0.95	<DL
Dns-adenosine monophosphate	0.99	<DL
Dns-o-phosphoethanolamine	1.06	16
Dns-glucosamine	1.06	22
Dns-o-phospho-L-threonine	1.09	<DL
Dns-6-dimethylamine purine	1.20	<DL
Dns-3-methyl-histidine	1.22	80
Dns-taurine	1.25	834
Dns-carnosine	1.34	28
Dns-Arg	1.53	36
Dns-Asn	1.55	133
Dns-hypotaurine	1.58	10
Dns-homocarnosine	1.61	3.9
Dns-guanidine	1.62	<DL
Dns-Gln	1.72	633
Dns-allantoin	1.83	3.8
Dns-L-citrulline	1.87	2.9

m.z	p.value	t.score
304.2979	1.02E-10	14.7179316
177.1024	1.62E-10	14.2666
345.0277	1.72E-10	-14.209195
491.0325	1.83E-10	-14.146348
258.0048	2.17E-10	-13.987636
483.1205	2.22E-10	-13.967634
694.9937	2.81E-10	-13.745172
270.9767	3.27E-10	13.6060705
371.604	3.53E-10	-13.534483
316.5773	3.71E-10	13.4893333
451.0505	4.04E-10	-13.412347
257.0543	4.09E-10	-13.401887
762.9787	4.71E-10	-13.274141
231.0422	5.14E-10	-13.195677
614.0797	6.11E-10	13.0407288
213.0066	6.79E-10	12.9471714
416.2122	7.92E-10	12.8119438



Strategies for omics data interpretation



Enrichment Analysis

- Purpose: To test if there are **biologically meaningful** groups of metabolites that are significantly enriched in your data
- Biological meaningful in terms of:
 - Pathways
 - Disease associated metabolite signatures
 - Localization
 - SNP associated metabolites
 -

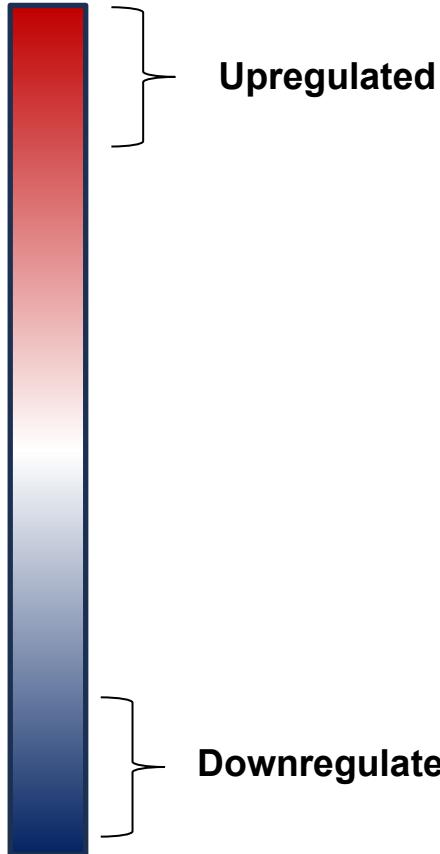
We need to define functions first

Please select a metabolite set library

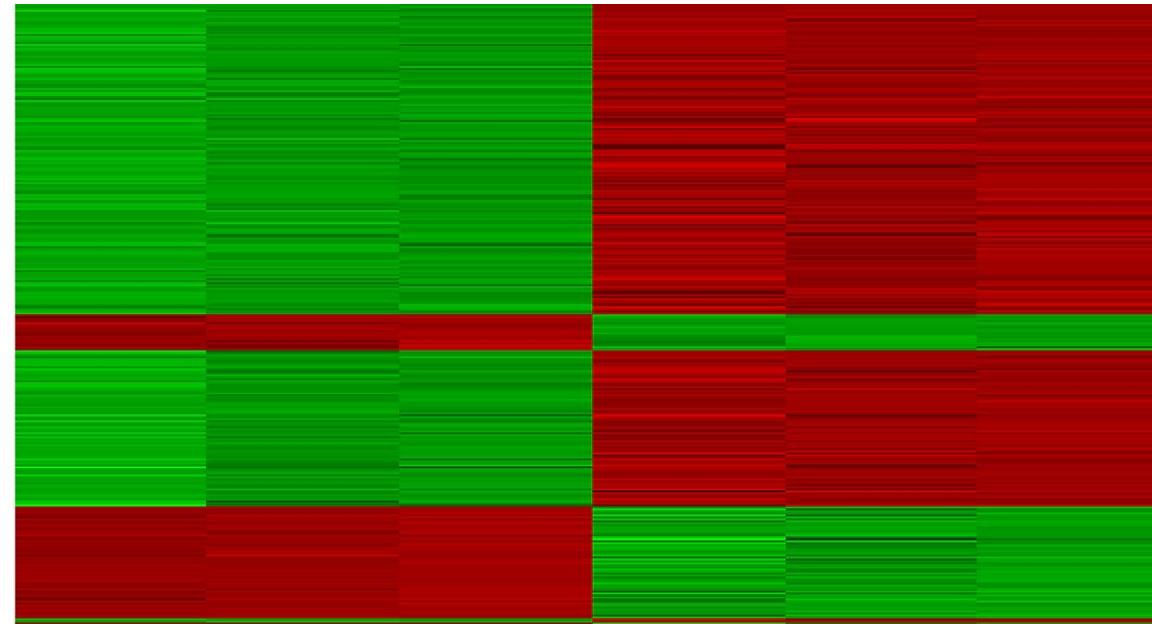
Pathway based	<input checked="" type="radio"/> SMPDB 99 metabolite sets based on normal human metabolic pathways. <input type="radio"/> KEGG 84 metabolite sets based on KEGG human metabolic pathways (Oct. 2019). <input type="radio"/> Drug related 461 metabolite sets based on drug pathways from SMPDB.
Disease signatures	<input type="radio"/> Blood 344 metabolite sets reported in human blood. <input type="radio"/> Urine 384 metabolite sets reported in human urine. <input type="radio"/> CSF 166 metabolite sets reported in human cerebral spinal fluid (CSF). <input type="radio"/> Feces 44 metabolite sets reported in human feces.
Chemical structures	<input type="radio"/> Super-class 35 super chemical class metabolite sets or lipid sets <input type="radio"/> Main-class 464 main chemical class metabolite sets or lipid sets <input type="radio"/> Sub-class 1072 sub chemical class metabolite sets or lipid sets
Other types	<input type="radio"/> SNPs 4,598 metabolite sets based on their associations with SNPs loci. <input type="radio"/> Predicted 912 metabolic sets predicted to change in the case of dysfunctional enzymes. <input type="radio"/> Locations 73 metabolite sets based on organ, tissue, and subcellular localizations.
Self defined	<input type="radio"/> Upload here define your own customized metabolite sets

Only use metabolite sets containing at least 2 entries ▾

Functions are coordinated changes



- Multiple molecules work together to finish a task



Clustered (co-regulated) compounds

Detecting functional changes

- To test whether members involved in a function have more consistent changes (i.e. most move in one direction) compared to random changes
- Enrichment analysis
 - Over representation analysis (ORA)
 - Starting from significant compounds
 - Different cutoff could get different results
 - Gene Set Enrichment Analysis (GSEA)
 - A complete ranked compound list
 - Cutoff-free

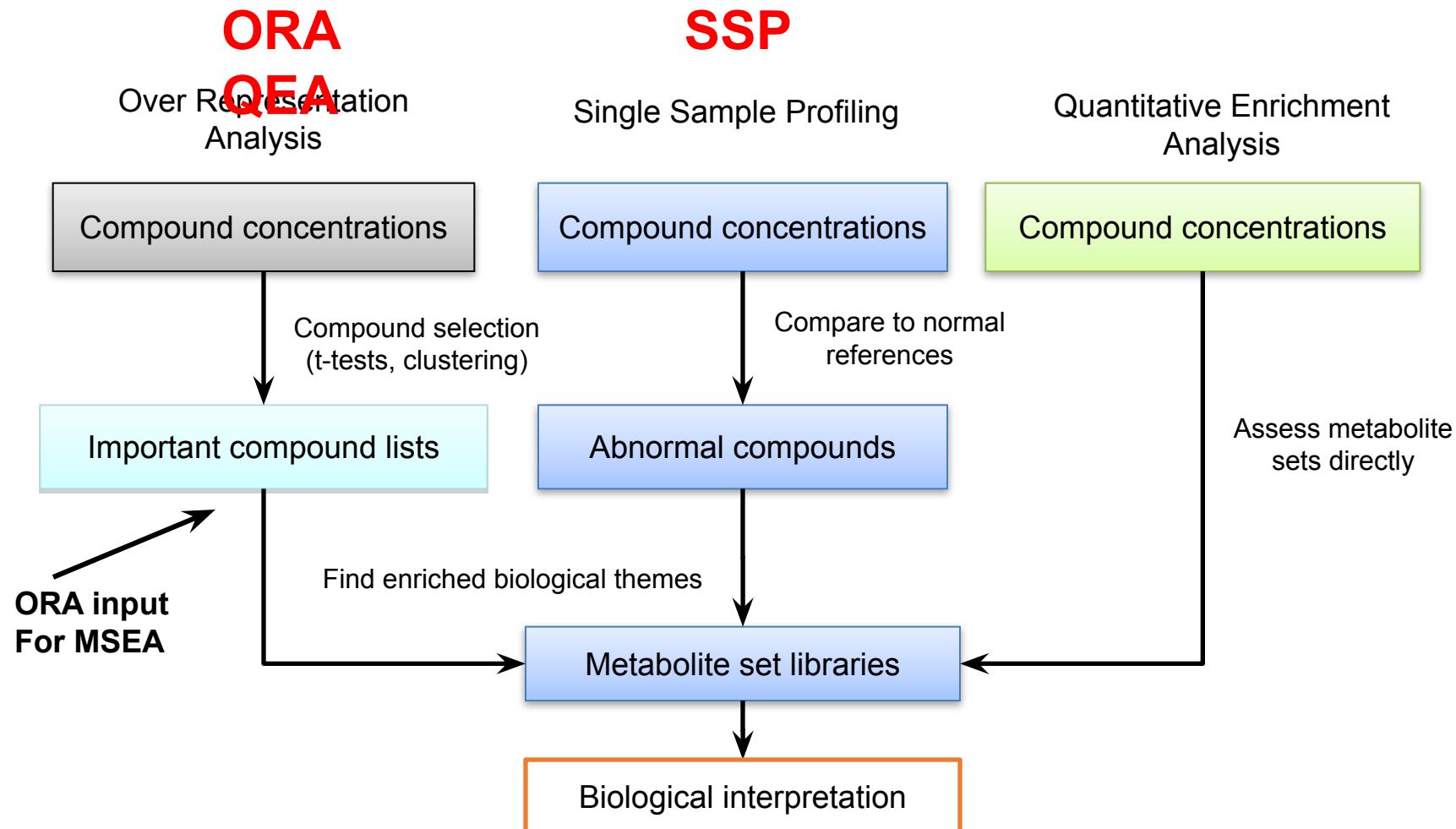
Functional analysis in metabolomics

- Targeted metabolomics is similar to gene expression profiling.
 - The compound concentration table can be used for enrichment analysis against pathways to identify which pathways or biological processes are changed significantly under the conditions.
- It is more challenging to perform functional analysis for untargeted metabolomics, as peaks cannot be directly mapped to pathways
 - Can we predict functions from peaks in a similar manner?

Metabolite Set Enrichment Analysis

- **Similar to Gene Set Enrichment Analysis (GSEA)**
- **Accepts 3 kinds of input files**
 - list of metabolite names only (ORA – over representation analysis)
 - list of metabolite names + concentration data from a single sample (SSP – single sample profiling)
 - a concentration table with a list of metabolite names + concentrations for multiple samples/patients (QEA – quantitative enrichment analysis)

The MSEA Approach



Upload a compound list, or sample concentration

Over Representation Analysis Single Sample Profiling Quantitative Enrichment Analysis

Please enter a one-column compound list:

Acetoacetic acid
Beta-Alanine
Creatine
Dimethylglycine
Fumaric acid
Glycine
Homocysteine
L-Cysteine
L-Isolucine
L-Phenylalanine
L-Serine
L-Threonine
L-Tyrosine
L-Valine
Phenylpyruvic acid
Propionic acid
Pyruvic acid

Input Type: Compound names ▾

Feature Type: Metabolites ▾

Try Example: None List 1 (metabolites) List 2 (lipids)

Submit

Over Representation Analysis Single Sample Profiling Quantitative Enrichment Analysis

Enter your data below (two-column data):

L-Isolecine 0.34
Fumaric acid 0.47
Acetone 0.58
Succinic acid 9.4
1-Methylhistidine 9.6
L-Asparagine 19.62
3-Methylhistidine 9.7
L-Threonine 93.19
Creatine 720
cis-Aconitic acid 14.39
L-Tryptophan 35.78
L-Carnitine 16.01
L-Serine 17.32
L-Tyrosine 67.51
L-Alanine 219.02
L-Fucose 20.37
D-Glucose 22.02

Input Type: Compound names ▾

Feature Type: Metabolites ▾

Biofluid (unit): Urine (umol/mmol_creatinine) ▾

Use the example data
- urine sample (umol/mmol_creatinine)

Submit

Concentration Comparison (Single Sample Profiling)

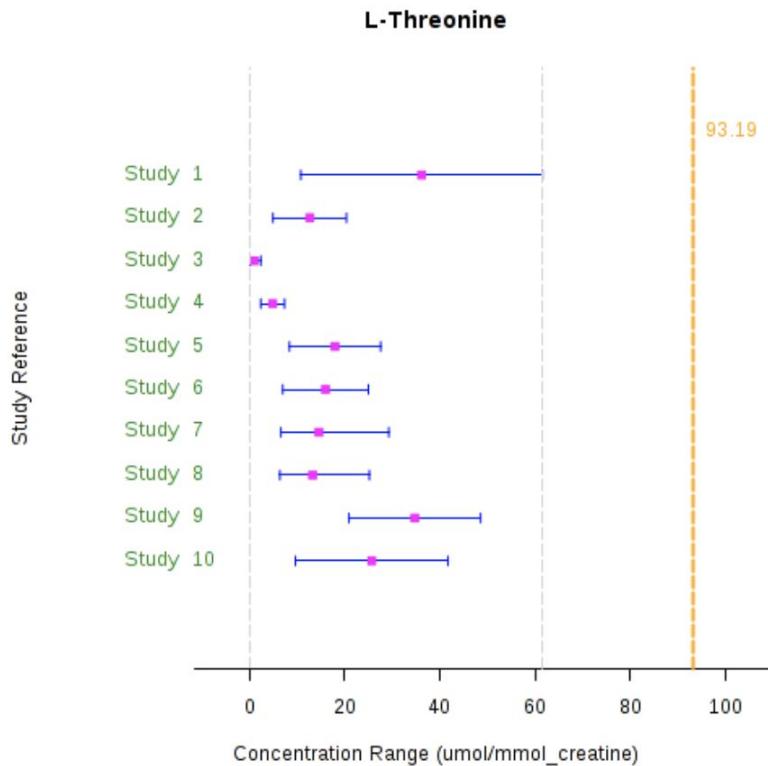
Comparison with Reference Concentration

Note: reference concentrations are in the form of **mean(min - max)** format. In cases where the ranges were not reported in the original literature, the min and max were calculated using the 95% confidence intervals. In the **Comparison** column, **H**, **M**, **L** means **higher**, **medium (within range)**, **lower** compared to the reference concentrations. Click the **Image Icon** link to see a graphical summary for the comparisons.

Compound	Concentration	Reference concentrations	Comparison	Detail	Include
Fumaric acid	0.47	0.95 (0.02 - 1.88); 10.4 (2.8 - 53.7); 0.5 (0.1 - 1.7); 10.7 (0.1 - 28.2); 0.4 (0.2 - 0.8); 0.7 (0.2 - 1.7); 1.342 (0.025 - 2.659)	M	View	<input type="checkbox"/>
Acetone	0.58	4.2 (0.98 - 15.3); 0.92 (0.2 - 2.8); 3.9 (0.8 - 17.6); 2.24 (0 - 6.37); 3.914 (0 - 7.843)	M	View	<input type="checkbox"/>
Succinic acid	9.4	7.7 (1.9 - 20); 197.2 (29.4 - 486.2); 185.4 (6 - 342.6); 11.6 (4 - 27.3); 8.25 (0.5 - 16); 9.9 (4.9 - 14.9); 14.4 (9.5 - 19.3); 12.6 (0.47 - 24.73); 3.8 (1.25 - 6.7); 14.48 (11.28 - 17.68); 7.5 (0.5 - 16); 5.6 (1.8 - 9.4); 6.2 (2.5 - 13.5); 4.7 (1.1 - 14.5); 6 (0.3 - 33.3); 14.581 (4.918 - 24.244); 163.491 (0 - 355.25)	M	View	<input type="checkbox"/>
1-Methylhistidine	9.6	4.6 (1.9 - 7.3); 2.3 (0 - 7.4); 46.1 (0 - 99.6); 15.9 (0 - 35.4); 28.1 (0 - 59.9); 1.3 (0 - 4.06); 45.5 (3.9 - 87.1); 33.6 (0 - 70); 30 (0 - 73); 8.3 (2.4 - 28.4); 15.9 (0 - 35.4); 21.329 (8.358 - 34.3); 46.0833 (0 - 99.5153); 45.473 (3.856 - 87.09); 15.919 (0 - 35.422); 33.624 (0 - 69.984); 28.0954 (0 - 59.9214); 2.34 (0 - 7.4164); 1.334 (0 - 4.093)	M	View	<input type="checkbox"/>
L-Asparagine	19.62	0.96 (0.31 - 1.61); 9.211 (3.289 - 15.1); 10 (4.6 - 16.32); 10.52 (6.67 - 14.37); 8.8 (4.6 - 17.7); 9.5 (3 - 26); 10.1 (4.6 - 17.8); 35.7098 (23.0844 - 48.3352); 25.158 (1.696 - 48.62)	M	View	<input type="checkbox"/>
3-Methylhistidine	9.7	12.5 (8.3 - 16.7); 42.76 (19.92 - 65.6); 16.5 (2.8 - 59.8); 44.944 (3.626 - 86.262); 17.977 (12.075 - 23.879); 18.191 (11.1587 - 25.2233); 15.806 (10.402 - 21.21); 13.556 (11.069 - 16.043); 18.203 (11.691 - 24.715); 17.705 (11.70151 - 23.70849); 16.0206 (11.1256 - 20.9156); 17.0268 (15.4548 - 18.5988)	M	View	<input type="checkbox"/>
L-Threonine	93.19	36.2 (10.82 - 61.58); 12.7 (4.934 - 20.4); 1 (0.16 - 2.4); 4.9 (2.4 - 7.4); 18 (8.4 - 27.6); 16 (7 - 25); 14.6 (6.6 - 29.3); 13.3 (6.4 - 25.2); 34.7611 (20.9314 - 48.5908); 25.712 (9.672 - 41.752)	H	View	<input checked="" type="checkbox"/>



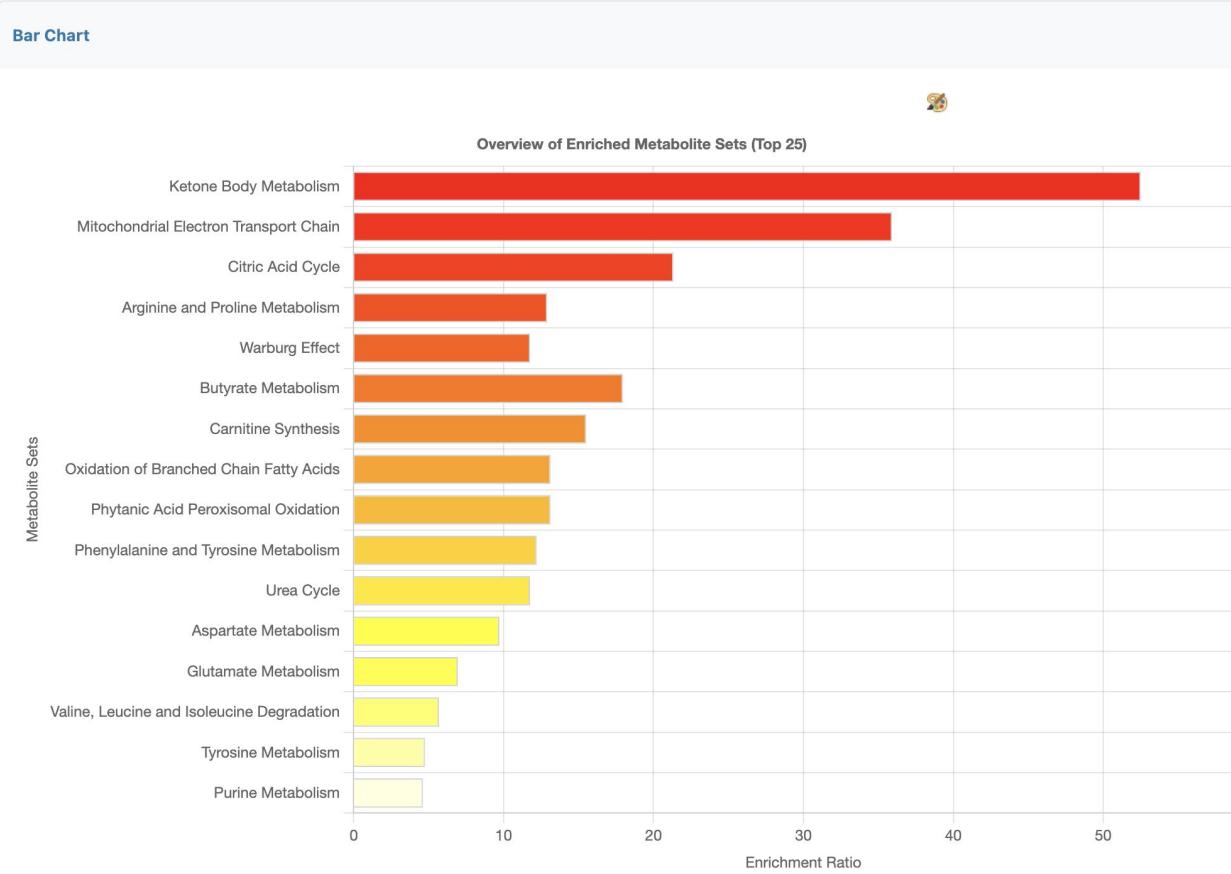
Concentration comparison (cont.)



Normal ranges based
on reports from
literature from HMDB

Study	Concentration	Reference	Note
Study 1	36.2 (10.82 - 61.58)	(Pubmed)	Both
Study 2	12.7 (4.934 - 20.4)	(Pubmed)	Both
Study 3	1 (0.16 - 2.4)	Geigy Scientific Tables, 8th Rev edition, pp. 165-177. Edited by Cornelius Lentner.; West Caldwell, N.J. : Medical education Div., Ciba-Geigy Corp.; Basel, Switzerland c1981-1992. (Pubmed)	Both
Study 4	4.9 (2.4 - 7.4)	Geigy Scientific Tables, 8th Rev edition, pp. 165-177. Edited by Cornelius Lentner.; West Caldwell, N.J. : Medical education Div., Ciba-Geigy Corp.; Basel	Female

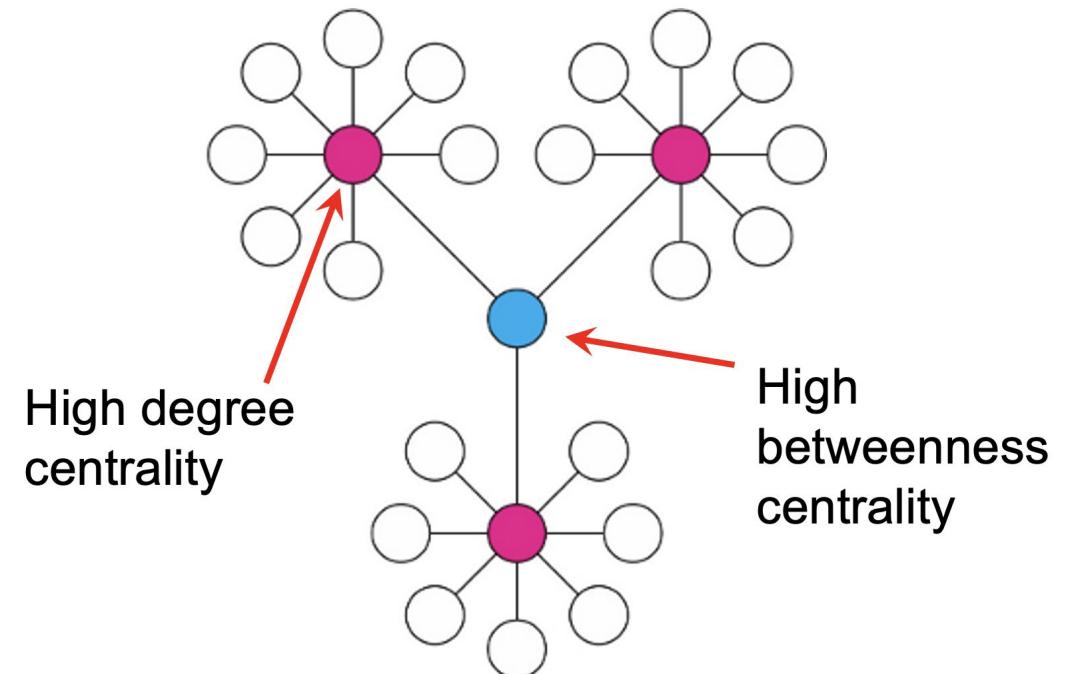
Results from Enrichment Analysis (targeted)



Pathway analysis – consider structures

- Which positions are important?

- Hubs
 - Nodes that are highly connected (red ones)
 - Bottlenecks
 - Nodes on many shortest paths between other nodes (blue ones)

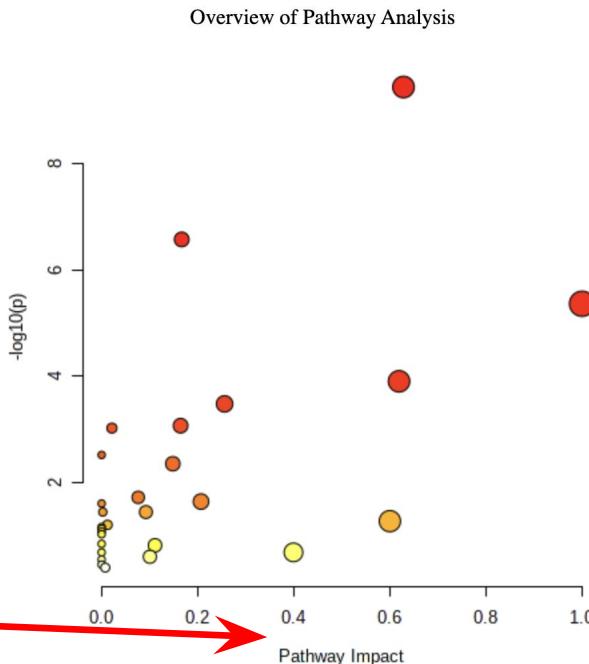


- Graph theory
 - Degree centrality
 - Betweenness centrality

Junker et al. BMC Bioinformatics 2006

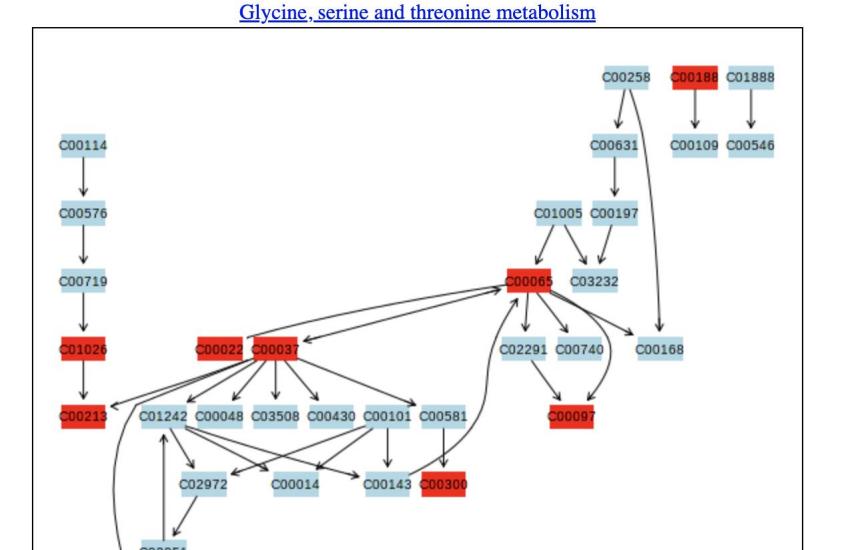
Pathway Results

Pathway Impact:
normalized sum of
degree / betweenness
values of matched nodes
within each pathway



Click the corresponding Pathway Name to view its graphical presentation;

Pathway Name	Match
Glycine, serine and threonine metabolism	8/33
Aminoacyl-tRNA biosynthesis	7/48
Phenylalanine, tyrosine and tryptophan biosynthesis	3/4



Matched metabolites:

Pathway	Metabolites
Glycine, serine and threonine metabolism	L-Serine ; Choline; Betaine aldehyde; Betaine; Guanidinoacetate; 3-Phospho-D-glycerate; N,N-Dimethylglycine ; L-Cystathionine; Glycine ; O-Phospho-L-serine; Sarcosine ; 5,10-Methylenetetrahydrofolate; L-Threonine ; Lipoylprotein; Aminoacetone; D-Glycerate; [Protein]-S8-aminomethylidihydrolipoyllysine; Tetrahydrofolate; Dihydrolipoylprotein; 2-Phospho-D-glycerate; D-Serine; Hydroxypyruvate; Creatine ; 3-Phosphonooxypyruvate; L-Cysteine ; 2-Oxobutanoate; Glyoxylate; L-2-Amino-3-oxobutanoic acid; Pyruvate ; CO ₂ ; 5-Aminolevulinate; Methylglyoxal; Ammonia

OK

Generate report & download results

Metabolomic Data Analysis with MetaboAnalyst 4.0

Name: guest7188993447605746741

June 11, 2020

1 Background

The Pathway Analysis module combines results from powerful pathway enrichment analysis with pathway topology analysis to help researchers identify the most relevant pathways involved in the conditions under study.

There are many commercial pathway analysis software tools such as Pathway Studio, MetaCore, or Ingenuity Pathway Analysis (IPA), etc. Compared to these commercial tools, the pathway analysis module was specifically developed for metabolomics studies. It uses high-quality KEGG metabolic pathways as the backend knowledgebase. This module integrates many well-established (i.e. univariate analysis, over-representation analysis) methods, as well as novel algorithms and concepts (i.e. Global Test, GlobalAnova, network topology analysis) into pathway analysis. Another feature is a Google-Map style interactive visualization system to deliver the analysis results in an intuitive manner.

2 Data Input

The Pathway Analysis module accepts either a list of compound labels (common names, HMDB IDs or KEGG IDs) with one compound per row, or a compound concentration table with samples in rows and compounds in columns. The second column must be phenotype labels (binary, multi-group, or continuous). The table is uploaded as comma separated values (.csv).

3 Compound Name Matching

The first step is to standardize the compound labels used in user uploaded data. This is a necessary step since these compounds will be subsequently compared with compounds contained in the pathway library. There are three outcomes from the step - exact match, approximate match (for common names only), and no match. Users should click the *textbf{View}* button from the approximate matched results to manually select the correct one. Compounds without match will be excluded from the subsequently pathway analysis.

Table 1 shows the conversion results. Note: 1 indicates exact match, 2 indicates approximate match, and 0 indicates no match. A text file contain the result can be found the downloaded file *name.map.csv*

Query	Match	HMDB	PubChem	KEGG	SMILES
1 1,6-Anhydro-beta-D-glucose	Levoglucosan	HMDB0000640	2724705	C1(C=O)C[C@H]2C(O)[C@H]([C@H](C=C2C)C(=O)O)O	
2 1-Methylnicotinamide	1-Methylnicotinamide	HMDB0000693	457	C02918	C[N+](=O)=CC=C(C=C1)C(=O)N
3 2-Aminobutyrate	L-Alpha-amino-butyric acid	HMDB0000452	80283	C02356	CC[C@H](C(=O)O)N
4 2-Hydroxyisobutyrate	Alpha-Hydroxyisobutyric acid	HMDB0000729	11671	C00026	CC(C)(C(=O)O)O
5 2-Oxoglutarate	Oxoglutaric acid	HMDB0000208	51	C00026	C(CC(=O)O)C(=O)C(=O)O

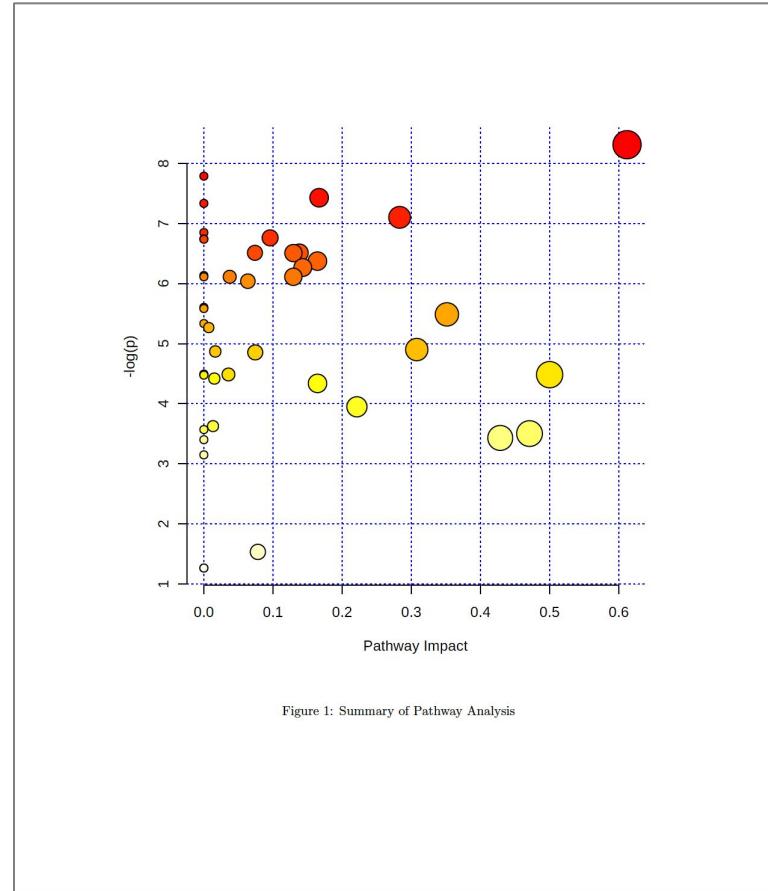
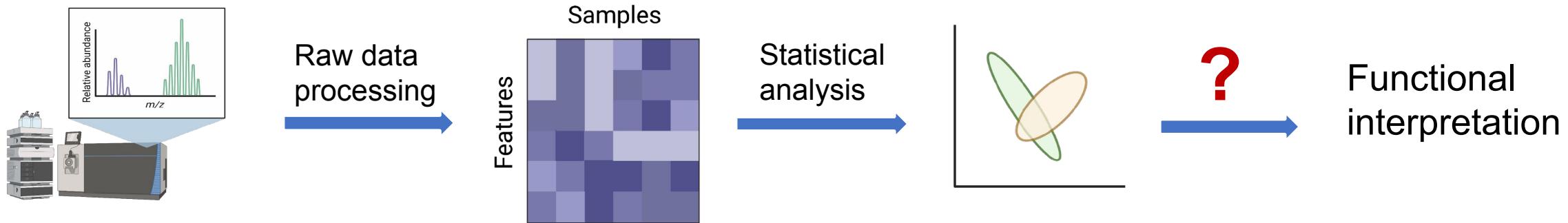
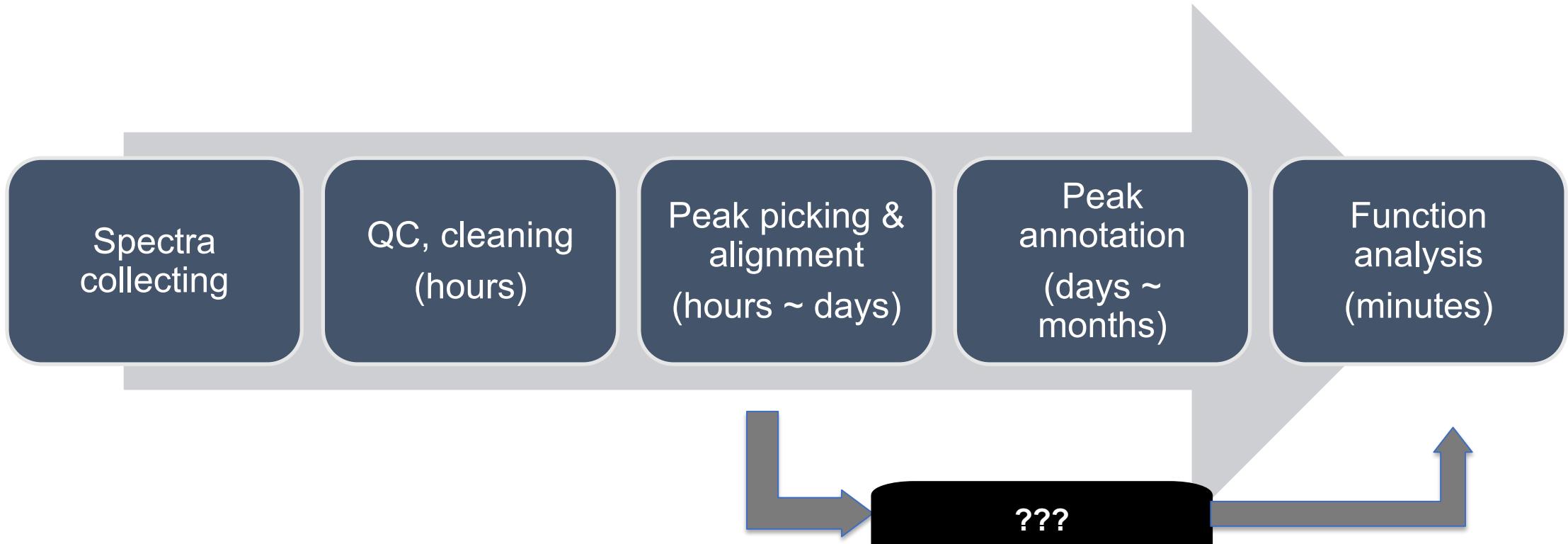


Figure 1: Summary of Pathway Analysis

Can we do the same to untargeted metabolomics?



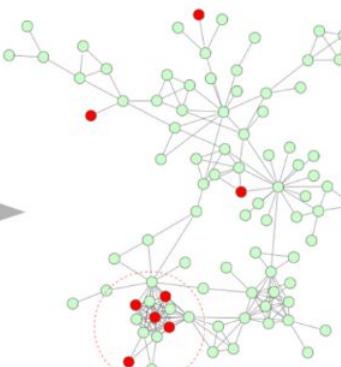
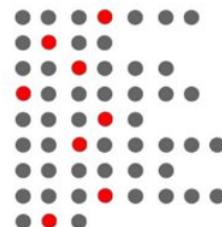
Conventional approaches requires compound ID known



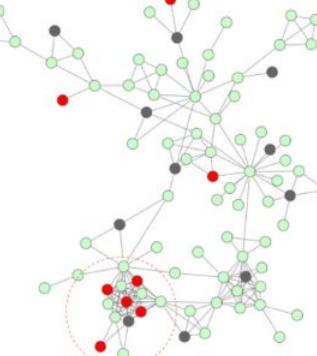
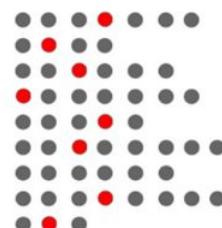
Can we perform enrichment analysis direct from peaks?

Mummichog

Conventional approach



mummichog



- We have a general idea what each peak is
- We have an idea of how each compound in a pathway could look like as peaks
- Mummichog: map many possible matches to a background pathway and look for consistent functional patterns
- Pathway-level results will be robust

S. Li et al., *Plos Computational Biology* (2013).

How does mummichog work?

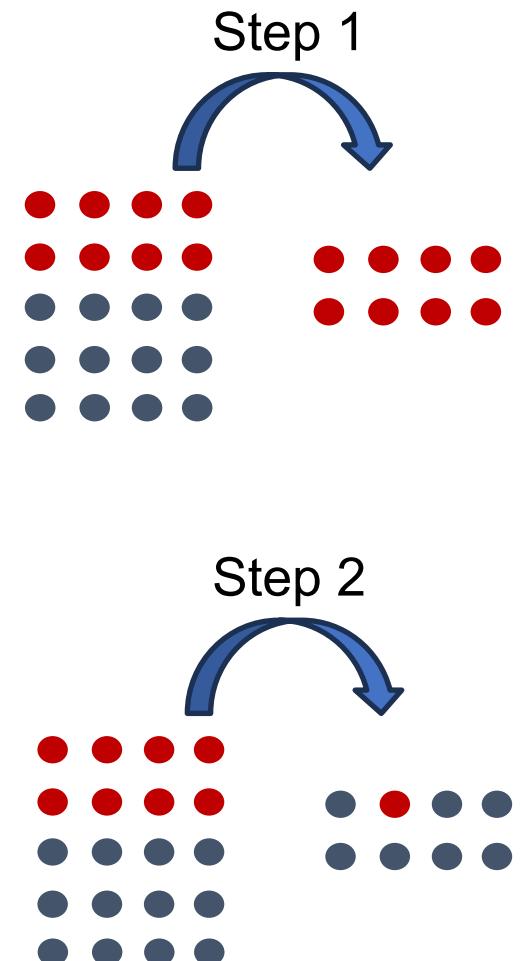
Input requirement:

- Significant list: selected by t-test or fold change
- Reference list: all features detected

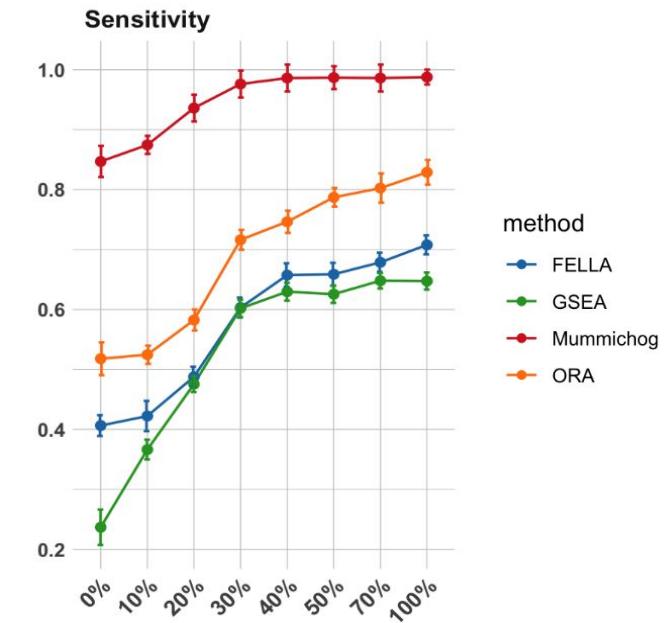
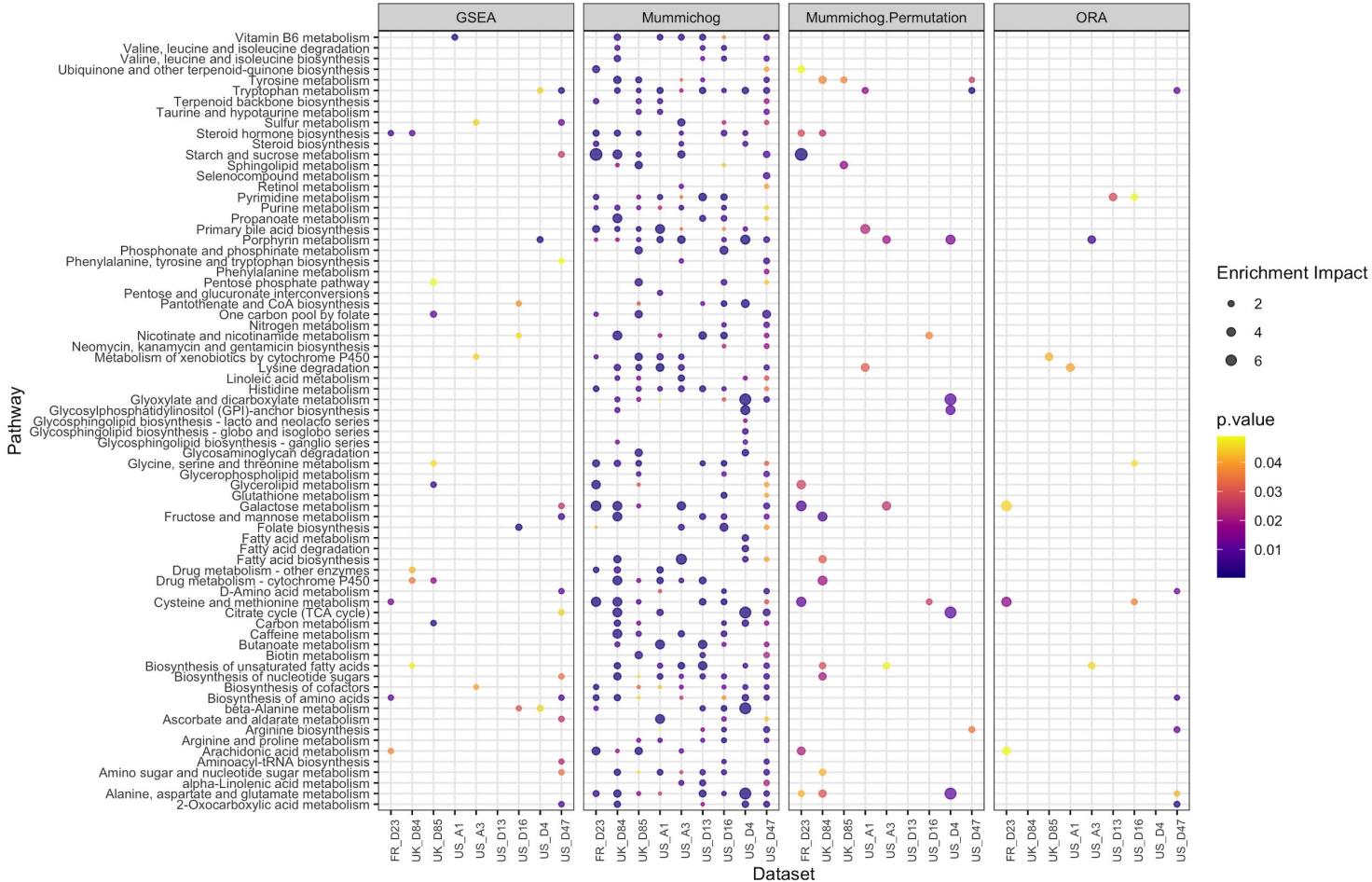
Step 1: Match the peaks to tentative metabolites. Looked up all the significant metabolites in each pathway and calculate the p-value using Fisher's exact test

Step 2: Randomly pull features with the same length as the significant ones, and repeat Step 1 for 100 ~ 1000 times

Step3: Test if certain pathways are enriched in the significant peaks as compared to null models (Gamma distribution)



Mummichog – sensitive, robust enrichment analysis for untargeted metabolomics



Implementation in MetaboAnalyst

The screenshot shows the MetaboAnalyst 5.0 interface. On the left is a vertical sidebar with links to Home, Data Formats, Tutorials, OmicsForum, APIs, Update History, MetaboAnalystR, Contact, User Stats, Publications, COVID-19 Data, and About. Logos for GenomeCanada and GenomeQuébec are also present. The main content area has a header "MetaboAnalyst 5.0 - user-friendly, streamlined metabolomics data analysis". Below it is a "Module Overview" section with a table showing available modules based on input data type. The table has four columns: Input Data Type, Available Modules (with a note to click on a module), LC-MS Spectra Processing, and other processing modules. Rows correspond to Raw Spectra, MS Peaks, Annotated Features, and Generic Format. Under Generic Format, Statistical Analysis [one factor] is highlighted in blue, while others like Biomarker Analysis and Power Analysis are in grey. Below the table are three callout boxes with detailed descriptions of the Statistical Analysis [one factor], Statistical Analysis [metadata table], and Biomarker Analysis modules.

Input Data Type	Available Modules (click on a module to proceed, or scroll down for more details)	LC-MS Spectra Processing	
Raw Spectra (mzML, mzXML or mzData)			
MS Peaks (peak list or intensity table)		Functional Analysis	Functional Meta-analysis
Annotated Features (compound list or table)		Enrichment Analysis	Pathway 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 dendograms 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.

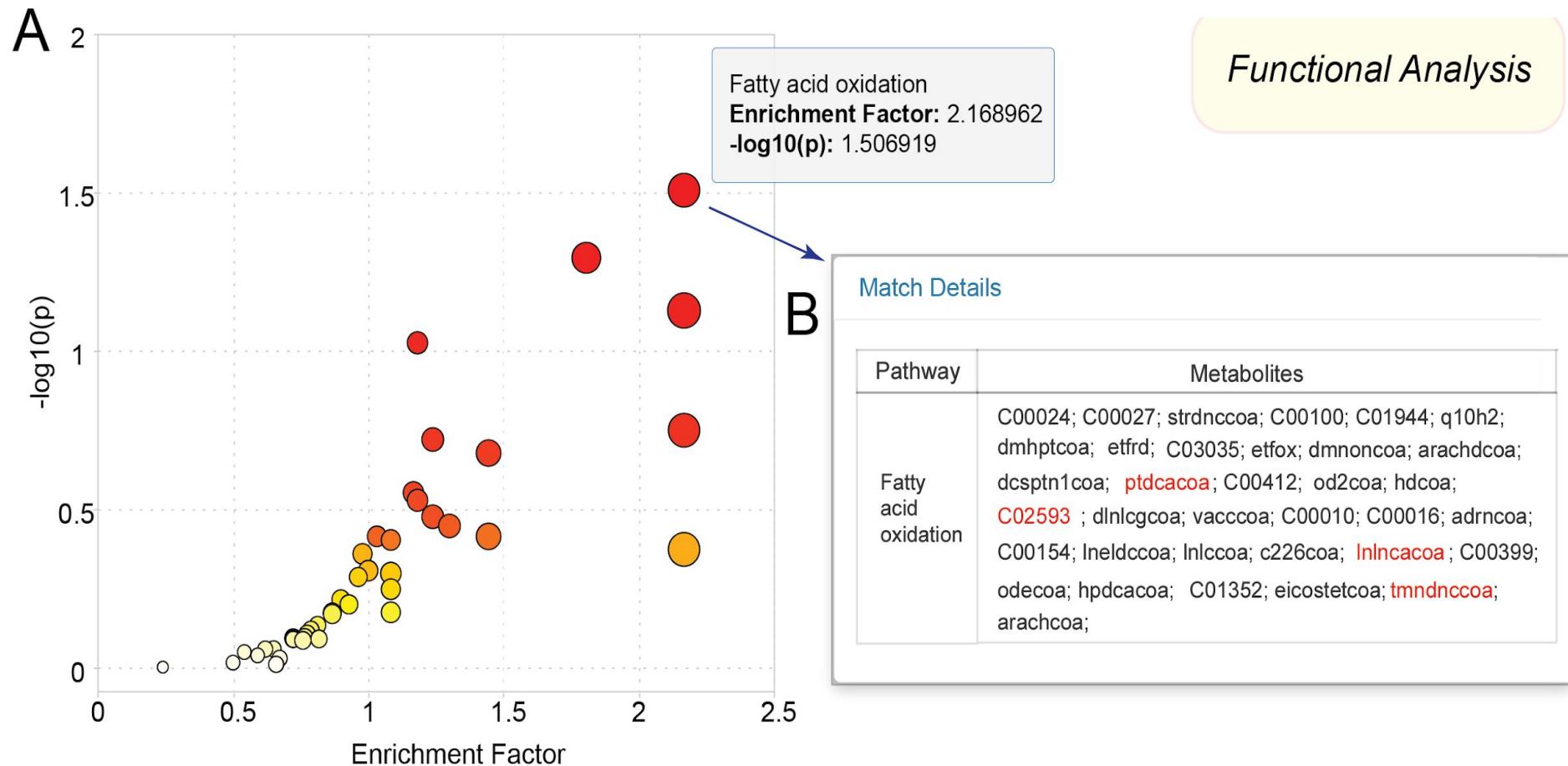
>> [Biomarker Analysis](#)

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.

Critical: input preparation

- LC - **high-resolution** MS (HR-MS)
 - Orbitrap, Q-TOF
 - Reason: putative annotation needs to be approximately correct (better guess leads to more accurate functional analysis)
- Needs to be **complete** peak list or peak intensity table
 - Not just significant peaks
 - Reason: mummichog using permutation to estimate the null/background distribution
- In general, the algorithm works well for **> 3000** peaks (assuming human plasma samples).

From ranked peak lists to functions

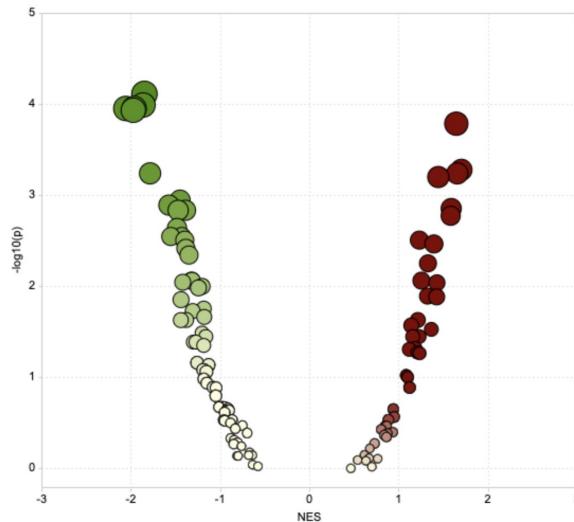


How to interpret result table (key parameters)

Pathway Name	Total ↑↓	Hits (all) ↑↓	Hits (sig.) ↑↓	Expected ↑↓	P(Fisher) ↑↓	P(Gamma) ↑↓	Details
Vitamin E metabolism	54	38	15	5.0563	0.030024	0.025523	View
Carnitine shuttle	72	25	10	6.7418	0.06554	0.028334	View

- **Total:** the total number of the given pathway
- **Hits (all):** all the peaks mapped to the pathway
- **Hits (sig):** all the significant peaks mapped to the pathway
- **Expected:** The expected number of metabolite hits in the pathway.
- **P(Fischer):** The Fisher's exact p-value for the pathway
- **P(Gamma):** P-values derived from Gamma distribution based on permutation tests for the pathway.

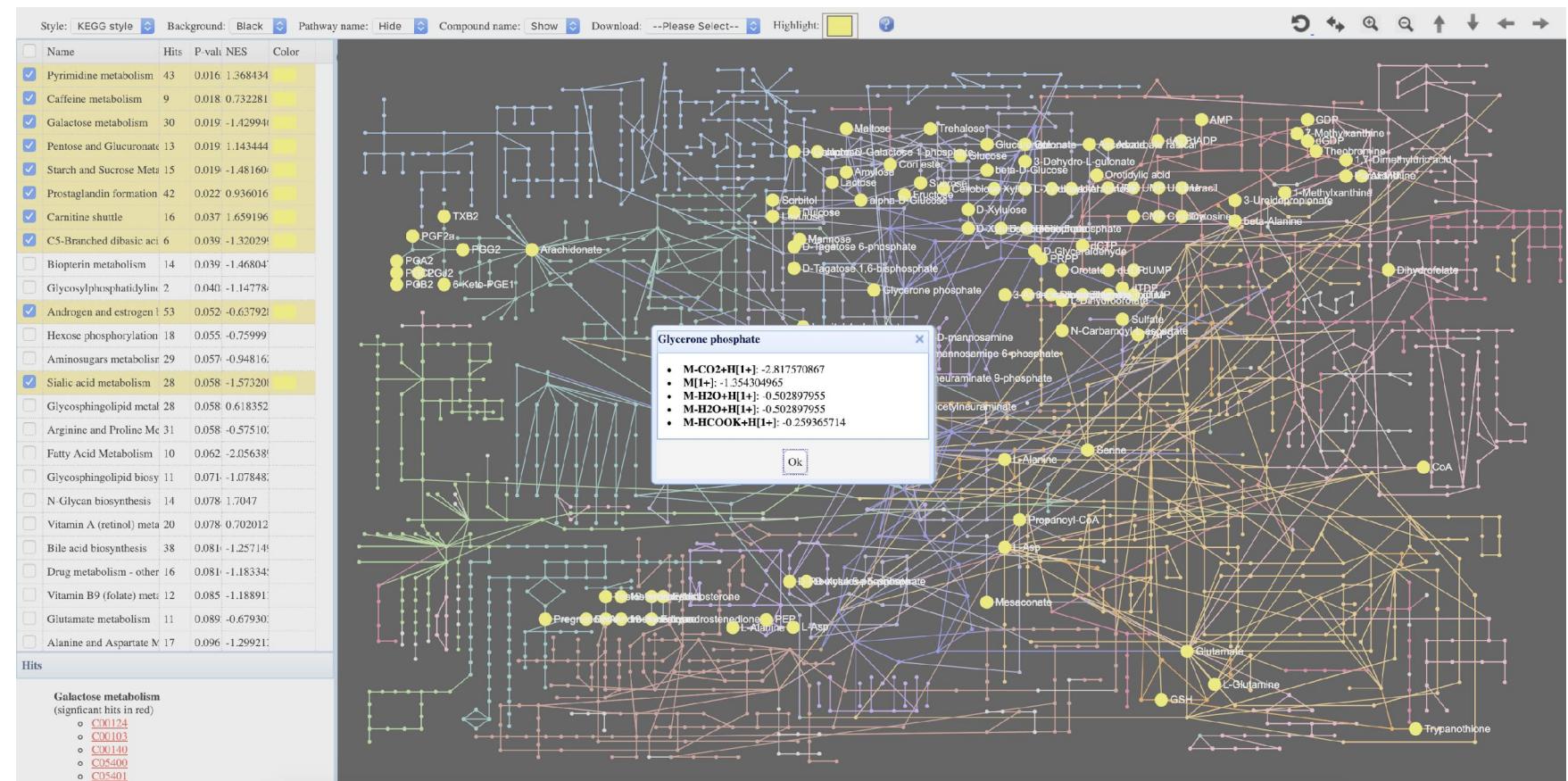
From pathways to “candidate” compounds



The red compounds indicate all potential matches from the user's input to the selected pathway.

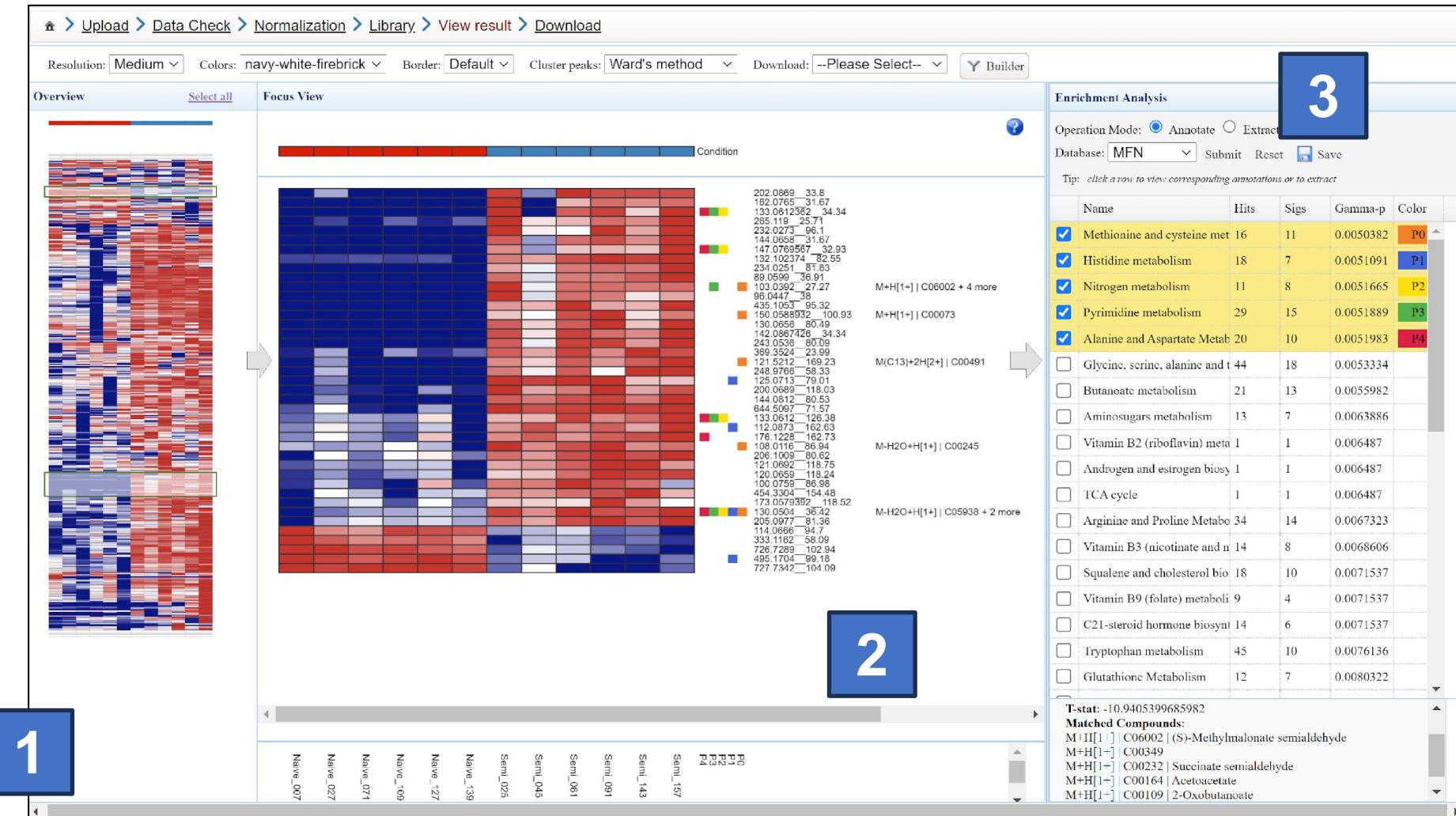
Pathway	Metabolites
Prostaglandin formation from arachidonate	CE5703; CE5700; CE6238; C00425 ; CE5707; CE5704 ; CE5705 ; CE5724; CE5708; CE1243; CE6236; C00584; CE5535 ; C00959 ; CE5537 ; CE5925 ; CE5533 ; C13856; CE0955 ; CE4980 ; C0132; CE5304; C04686 ; CE4938; C05963; C04741 ; CE5931 ; CE0737 ; C01041 ; CE4878; CE5926 ; C00051 ; CE5924 ; C00219 ; C00704; CE5928 ; CE5929 ; C00020 ; CE4876; C0010 ; C00954; C00024; C05962 ; C00030; C00072 ; CE7107; CE7105; C0028; C00696; CE1447 ; C04685 ; C11695 ; CE6244; CE6245; CE6242; CE6243; CE6240 ; CE6241; C05953 ; C00027; C05956 ; C05957 ; C00427; C05955 ; CE4877; CE6235; C05959 ; C00639 ; CE5730; CE5534 ; CE5930 ; C00282; CE3481 ; C00189; CE7054; C11304; C02198; CE5828

OK



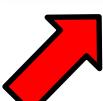
Both mummichog and GSEA are available in MetaboAnalyst 5.0

From patterns to functions



MetaboAnalyst 5.0 Lab

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	



We are on a Break

Workshop Sponsors:



Canadian Centre for
Computational
Genomics

