Introduction to Metabolomics Platforms and Data Analysis

Dr. Katerina Kechris

Professor
Department of Biostatistics and Informatics

University of Colorado Anschutz Medical Campus

Outline

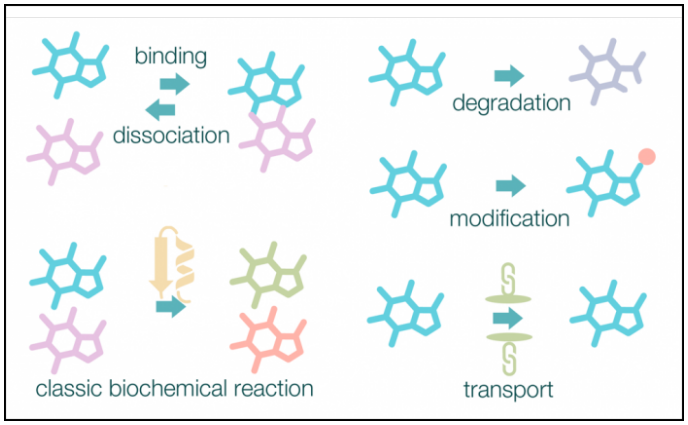
- Introduction
- Technology Types
- Mass Spectrometry
- Analysis Methods

Introduction

- Metabolomics is the study of small molecules in biological or ecological samples
- Molecules can be identified utilizing analytical strategies
- Methods can be targeted or un-targeted
- Primary goal is to accurately identify and quantify levels of small molecules in samples

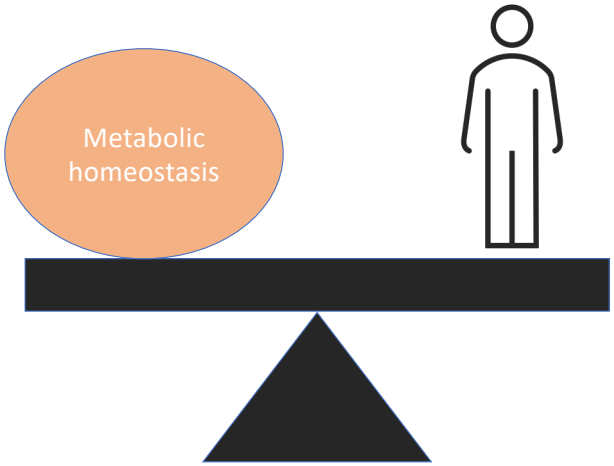
Metabolism

- Series of chemical reactions
  - Enzymes
  - Catabolism/Anabolism



https://www.ebi.ac.uk/

- Homeostasis
  - Metabolites maintain balance



Metabolism

Microbiota

Exposures, Diet
Drugs, Lifestyle

Genetics

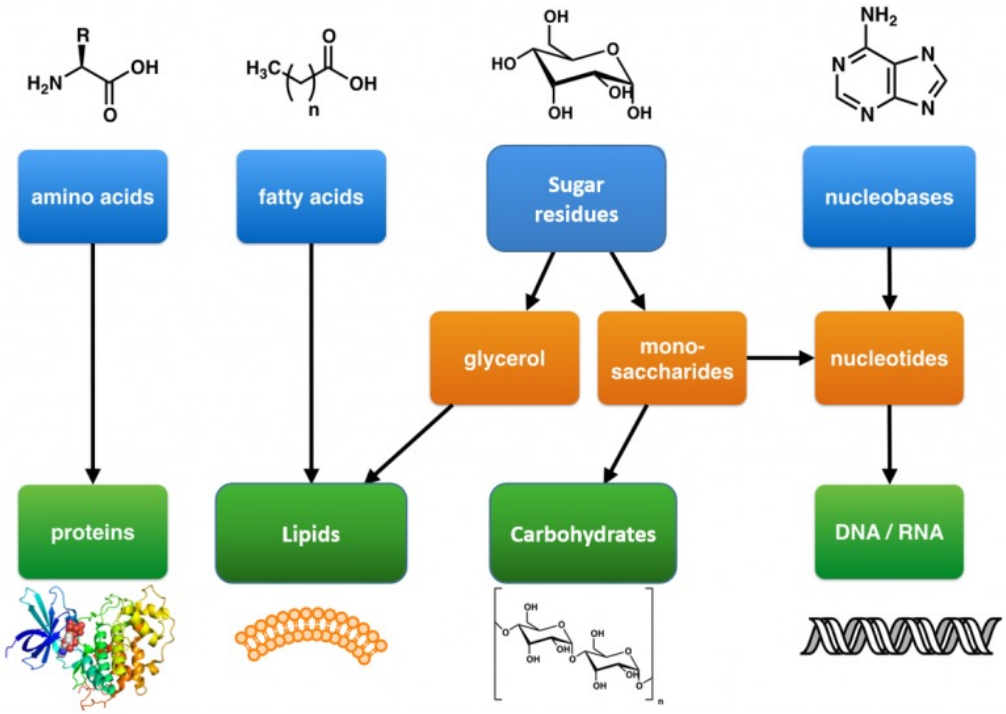
Metabolism

Phenotype

https://www.newscientist.com/ https://www.peoplespharmacy.com/
http://www.tricountycessation.org/ http://californiaagriculture.ucanr.edu
http://foodmetabolome.org/ http://www.metabolon.com/ http://www.gettyimages.ca/

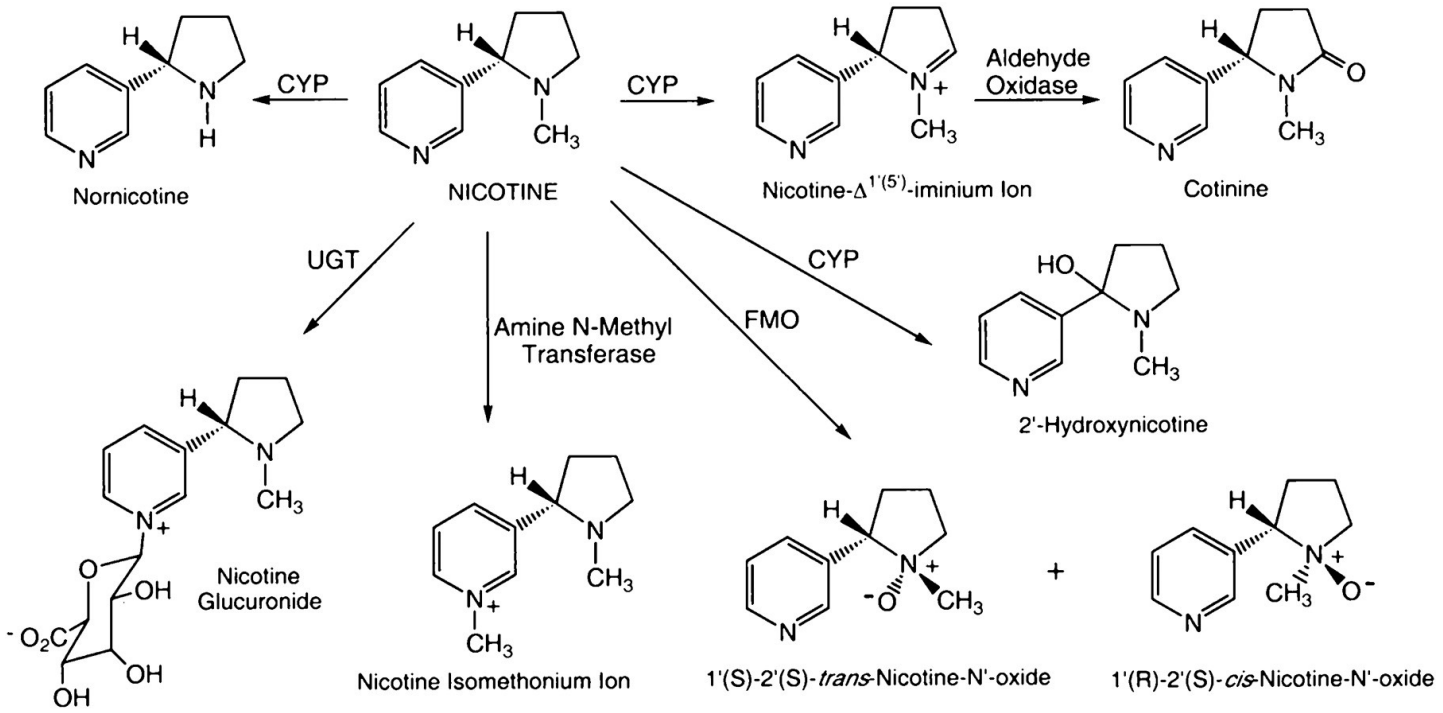
Classes

Small Molecules/Compounds/Metabolites



https://wou.edu/chemistry/chapter-11-introduction-major-macromolecules/

Example: Nicotine Metabolism



Nakajima & Yokoi (2005) Drug Metab Pharmacokinet

Early History

1968-1978

- Linus Pauling and Noah Robinson
- Profiled 50-200 metabolites early GC-MS
- Profile disease, gender, and more (15K subjects)
- 1000 ft stainless steel column

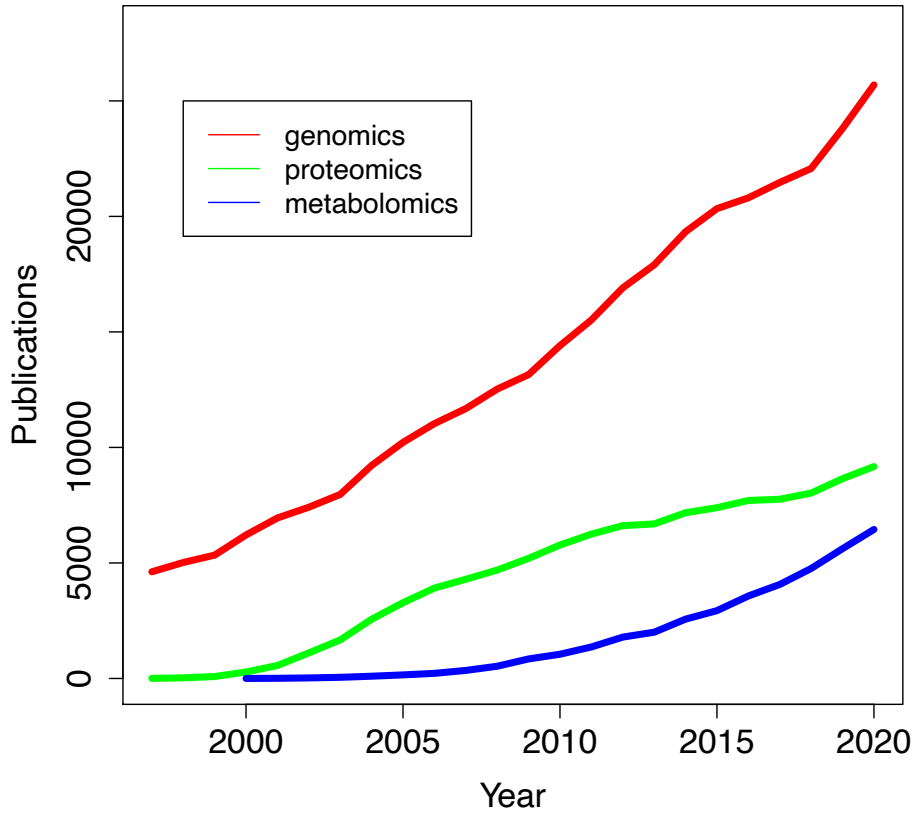
1998

- Term “metabolomics” coined

2007

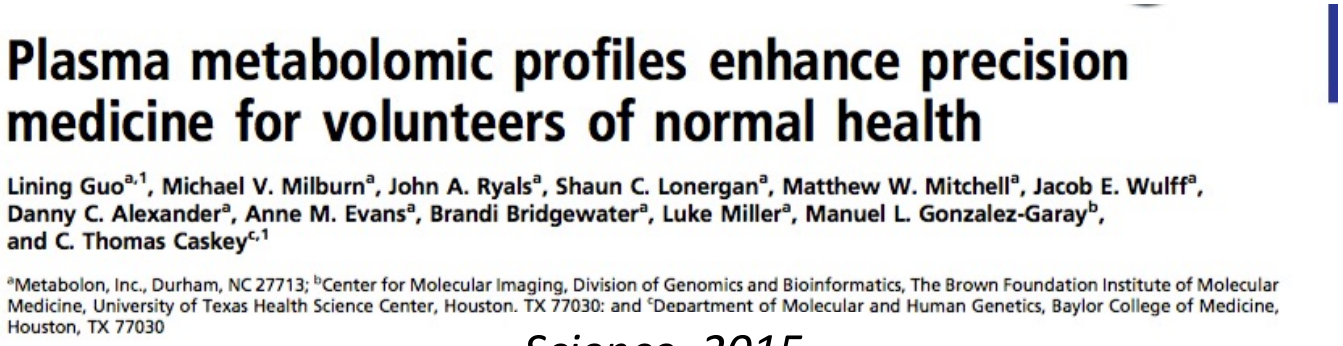
- Wishart group (U of Alberta)
- Human Metabolome Project
- 2500 human, 1200 drug, and 3500 food metabolites

‘Omics Publications



Source: PubMed 03/12/21, searched -omic* in “TextWord”

Translational Research



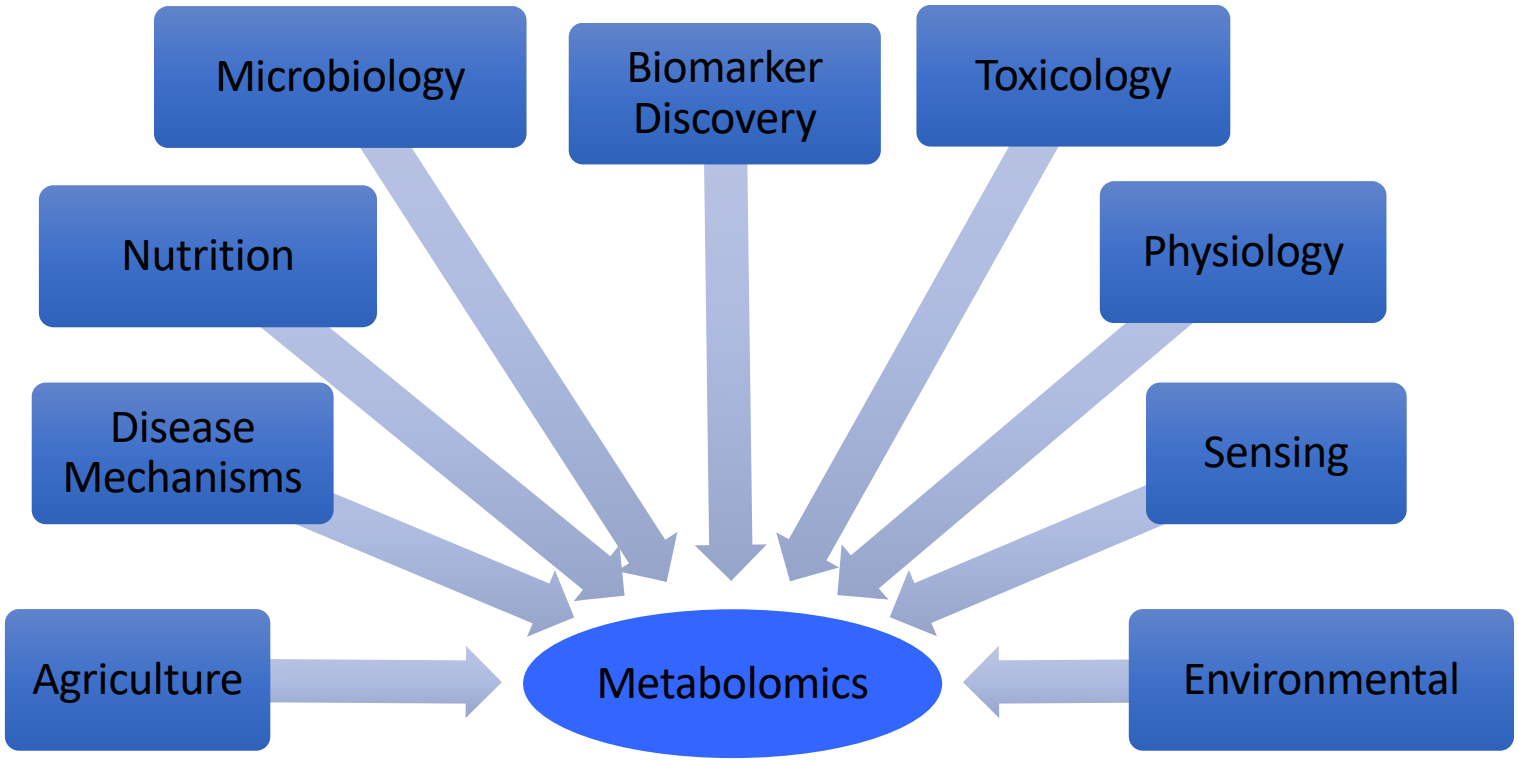
Plasma metabolomic profiles enhance precision medicine for volunteers of normal health

Lining Guo[a,1], Michael V. Milburn[a], John A. Ryals[a], Shaun C. Lonergan[a], Matthew W. Mitchell[a], Jacob E. Wulff[a], Danny C. Alexander[a], Anne M. Evans[a], Brandi Bridgewater[a], Luke Miller[a], Manuel L. Gonzalez-Garay[b], and C. Thomas Caskey[c,1]

[a]Metabolon, Inc., Durham, NC 27713; [b]Center for Molecular Imaging, Division of Genomics and Bioinformatics, The Brown Foundation Institute of Molecular Medicine, University of Texas Health Science Center, Houston. TX 77030: and [c]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030

Science, 2015

Key Applications



Types of Questions

- **Identification:** What small molecules are present?
- **Association:** Can metabolites be used as biomarkers? Are metabolite levels associated with disease status/spectrum? physiological state? diet? environmental exposures?
- **Prediction:** Do small molecules have predictive power for diagnosis? prognosis? precision medicine?
- **Mechanisms:** What are the relevant metabolic mechanisms?
- **Genetics**: Is there variability due to genetic background?
- **Microbial effects:** What is the role of the microbiome on metabolism?
- **Metabolic flux:** What is the rate of turnover of molecules through a metabolic pathway?

PLOS ONE

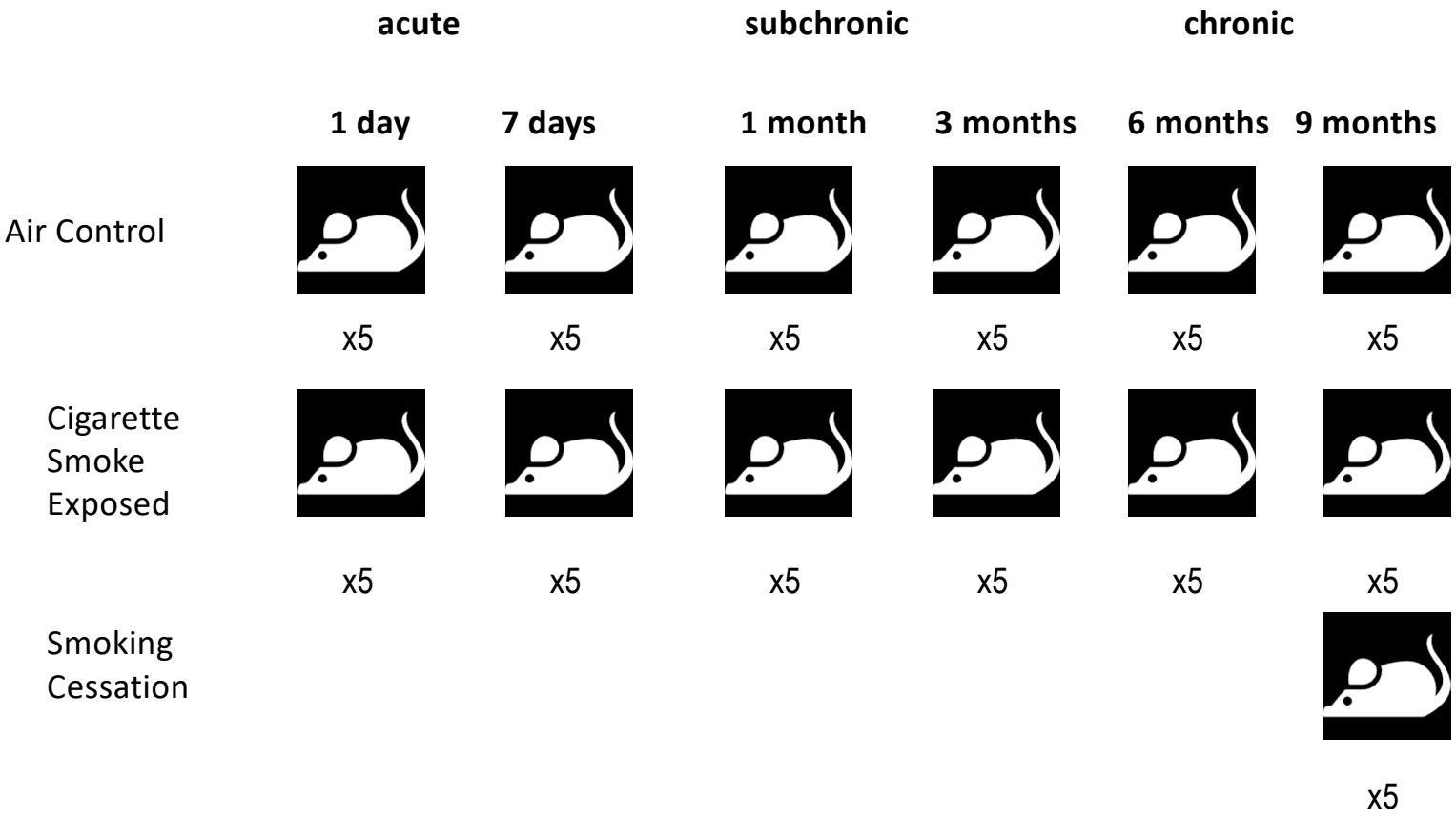
PLOS One. 2017; 12(6): e0178281.
Published online 2017 Jun 2. doi: 10.1371/journal.pone.0178281

PMCID: PMC5456044
PMID: 28575117

Gene and metabolite time-course response to cigarette smoking in mouse lung and plasma

Mikaela A. Miller,[1] Thomas Danhorn,[2] Charmion I. Cruickshank-Quinn,[3] Sonia M. Leach,[2] Sean Jacobson,[4] Matthew J. Strand,[5] Nichole A. Reisdorph,[3] Russell P. Bowler,[4,*] Irina Petrache,[4,6,*] and Katerina Kechris[1,*]

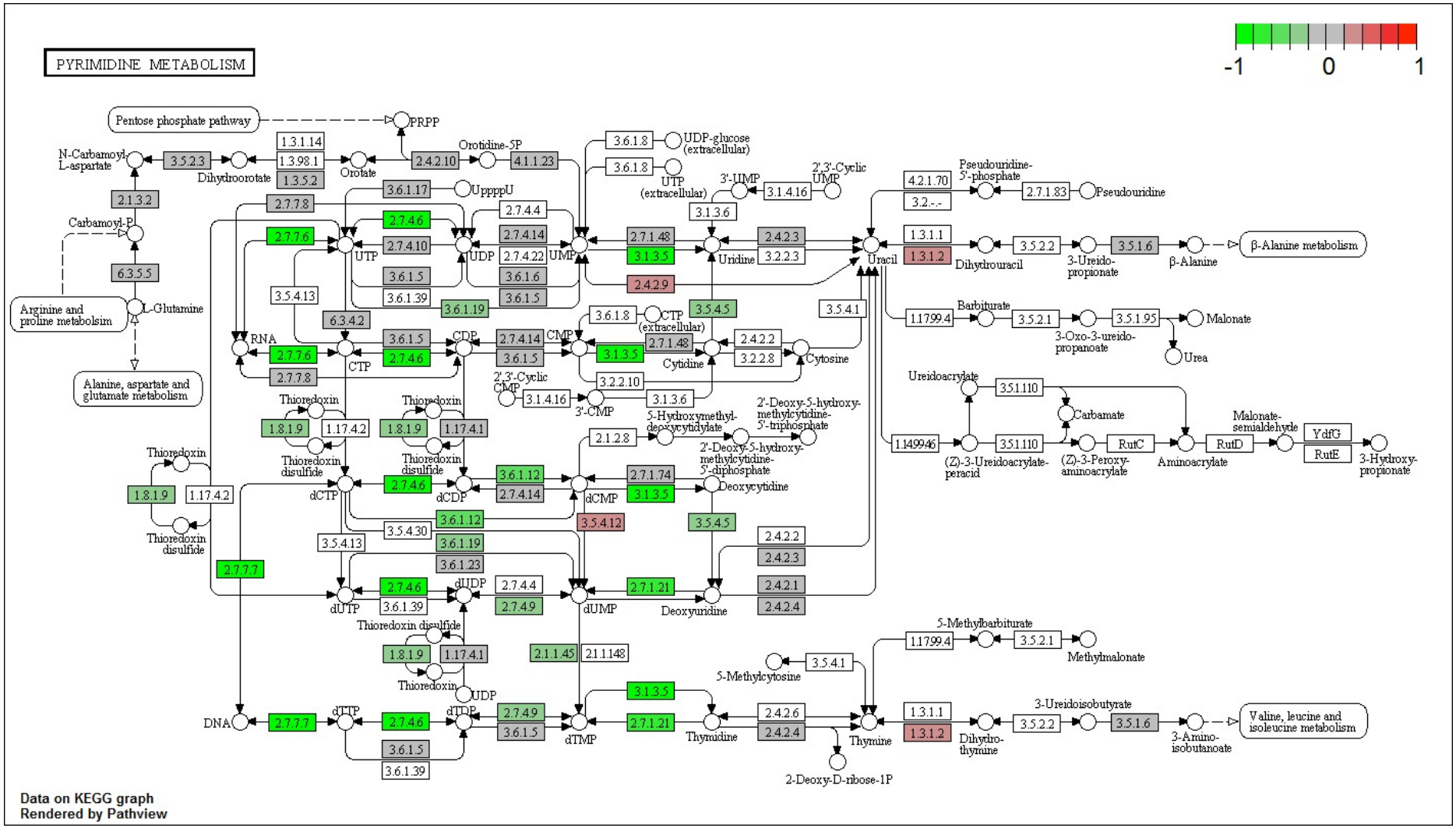
Experimental Design



Altered Pathways

	1 day	7 days	3 months	6 months	9 months	Cessation
Oxidative phosphorylation				✓		
Parkinson's disease						✓
Pyrimidine metabolism					✓	✓
Leishmaniasis		✓	✓	✓		✓
Fc gamma R-mediated phagocytosis	✓	✓	✓	✓		✓
Inositol phosphate metabolism					✓	
Gap junction	✓	✓	✓			
Pathways in cancer	✓	✓	✓	✓		✓
Phosphatidylinositol (PI) signaling system	✓	✓	✓	✓	✓	✓

Checkmarks - perturbation in metabolomic data
Pink - enrichment for upregulated genes in CS-exposed mice (FDR≤0.10)
Blue - enrichment for downregulation in CS-exposed mice (FDR≤0.10)
Shaded fields under the “Cessation” column persisted following cessation when compared to AC mice


PYRIMIDINE METABOLISM
-1
0
1
Pentose phosphate pathway
PRPP
N-Carbamoyl-L-aspartate
Dihydroorotate
Orotate
Orotidine-5P
UDP-glucose (extracellular)
UTP (extracellular)
3'-UMP
2',3'-Cyclic UMP
UppppU
Carbamoyl-P
Arginine and proline metabolsim
L-Glutamine
Alanine, aspartate and glutamate metabolism
UTP
UDP
UMP
Uridine
Uracil
Pseudouridine-5'-phosphate
Pseudouridine
Dihydrouracil
3-Ureido-propionate
β-Alanine
β-Alanine metabolism
Barbiturate
3-Oxo-3-ureido-propanoate
Malonate
Urea
RNA
CTP
CDP
CMP
CTP (extracellular)
Cytidine
Cytosine
2',3'-Cyclic CMP
3'-CMP
Ureidoacrylate
Carbamate
(Z)-3-Ureidoacrylate-peracid
(Z)-3-Peroxy-aminoacrylate
Aminoacrylate
Malonate-semialdehyde
YdfG
RutE
RutC
RutD
3-Hydroxy-propionate
Thioredoxin
Thioredoxin disulfide
5-Hydroxymethyl-deoxycytidylate
2'-Deoxy-5-hydroxy-methylcytidine-5'-triphosphate
2'-Deoxy-5-hydroxy-methylcytidine-5'-diphosphate
dCTP
dCDP
dCMP
Deoxycytidine
dUTP
dUDP
dUMP
Deoxyuridine
Thioredoxin disulfide
Thioredoxin
UDP
DNA
dTTP
dTDP
dTMP
Thymidine
Thymine
2-Deoxy-D-ribose-1P
5-Methylcytosine
5-Methylbarbiturate
Methylmalonate
Dihydro-thymine
3-Ureidoisobutyrate
3-Amino-isobutanoate
Valine, leucine and isoleucine metabolism
Data on KEGG graph
Rendered by Pathview

Outline

- Introduction
- Technology Types
- Mass Spectrometry
- Analysis Methods

Most Common Sample Types

- Plasma
  - contains fibrinogen (normal clotting of blood) & albumin (keeps fluid in bloodstream)
- Serum
  - plasma minus clotting factors and blood cells
- Urine
- But anything possible (saliva, stool, tissues, etc.)
- Considerations
  - fasting
  - time of day
  - hemolysis
  - freeze/thaw

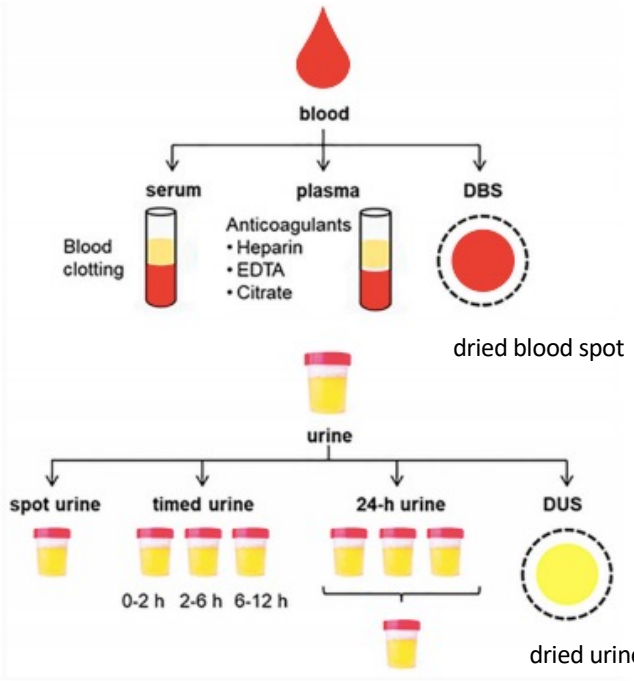
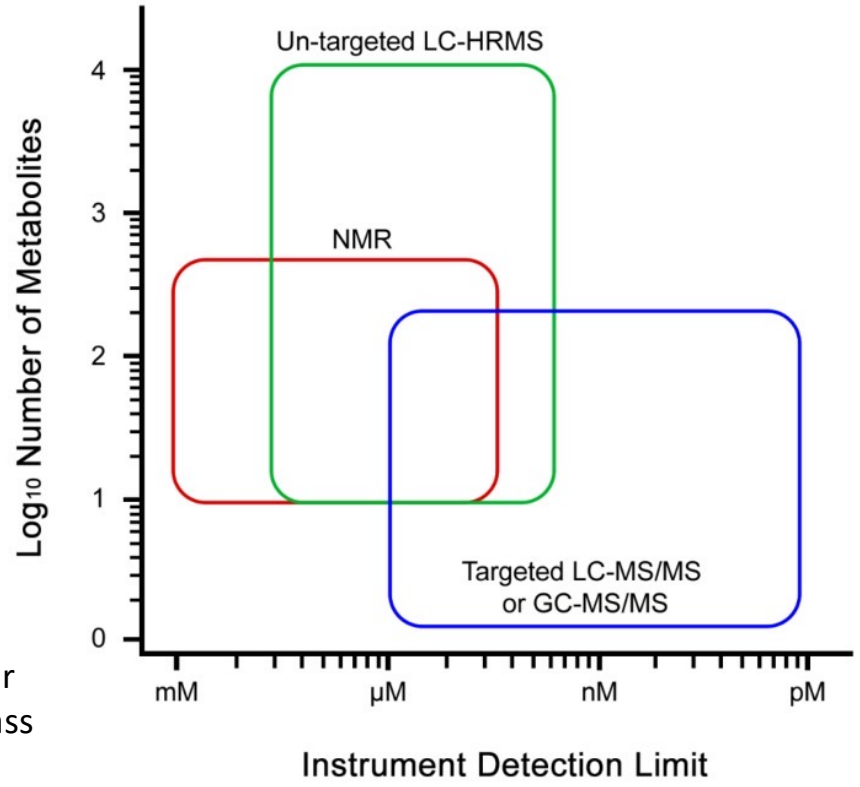
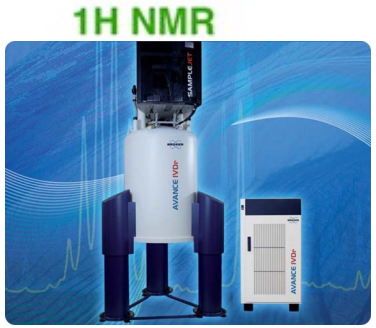


Figure 2. Summary of sampling alternatives for blood and urine collection.

González-Domínguez et al (2020) Metabolites

Technologies



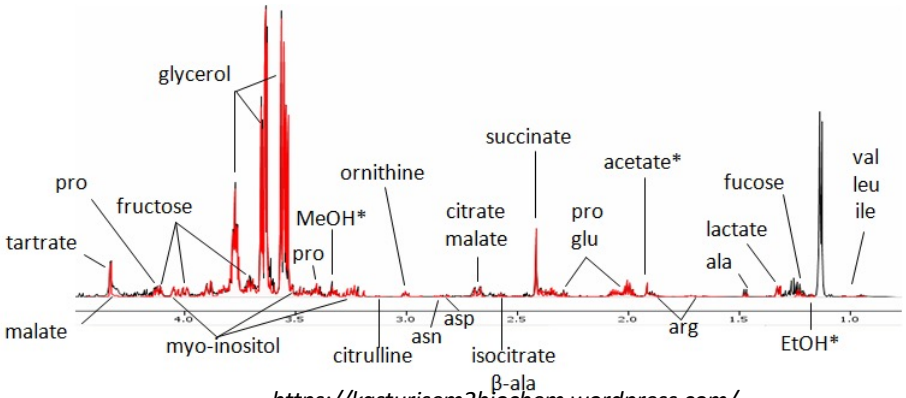


Liquid-Chromatography (LC) or Gas-Chromatography (GC) Mass Spectrometry (MS); Nuclear Magnetic Resonance (NMR)

Pinu et al., (2019) Metabolites, agilent.com, bruker.com

NMR

- Sometimes referred to as “metabonomics”
- No separation needed & can handle intact tissues, non-destructive
- High analytical reproducibility, simple sample prep
- Absolute, quantitative data
- Also supports imaging & real-time metabolite profiling of living cells, metabolic flux analysis
- But less sensitive, and for a limited number of metabolites



https://kasturisem2biochem.wordpress.com/

Mass Spectrometry Technologies

- Gas-Chromatography (GC)-MS
  - Uses fragmentation to identify molecules
- Liquid-Chromatography (LC)-MS
  - Can detect intact compounds and fragments but is destructive (sample is used)
  - Most common - ultra-high pressure (UHPLC)

Types

Targeted

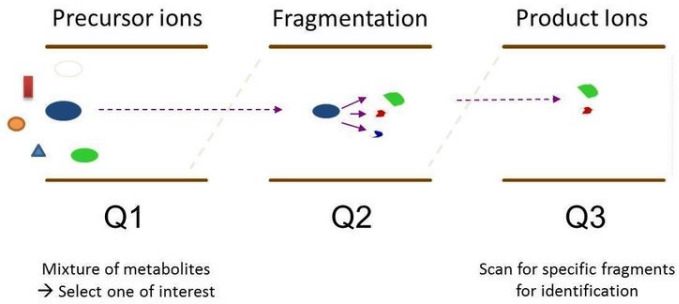
- Groups of characterized and annotated metabolites
- Hypothesis testing
- Known identities
- Absolute quantitation

Untargeted

- Comprehensive coverage (limited only by techniques)
- Hypothesis generation
- Predicted identification
- Relative quantitation

Targeted Analysis

- Multiple reaction monitoring (MRM)
  - Parent mass of a metabolite of interest selected

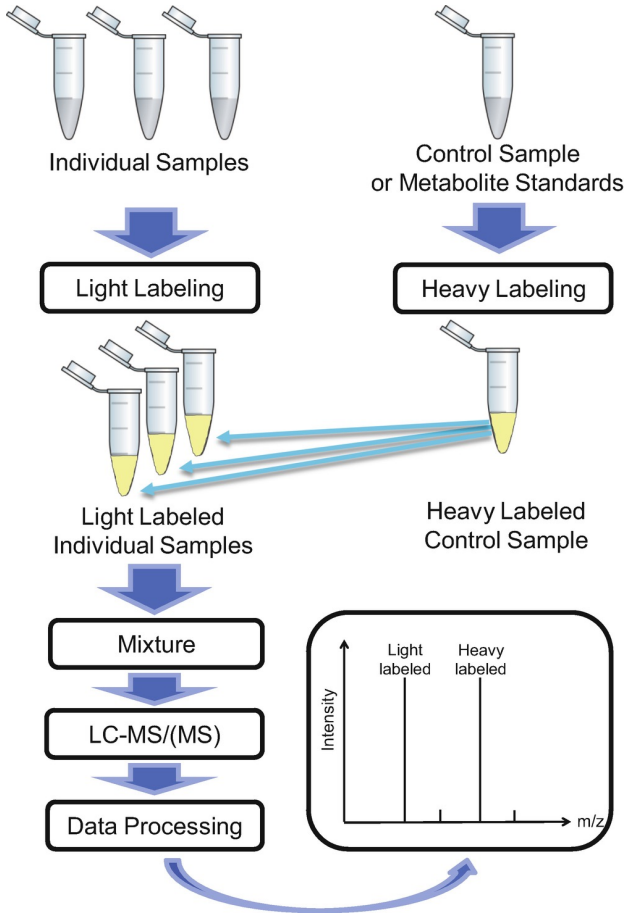


  - Use internal standards for identification and quantitation

https://www.molgen.mpg.de/101162/Metabolomics

Internal Standards

- Light labeling = endogenous metabolites
- Use stable isotopes as heavy standards
  - No decay over time
  - No safety concerns (radioactivity)
- Spike in heavy standards at known concentration
- Measure relative intensities (peak areas) between light and heavy, then can back calculate concentration of endogenous metabolites



Zhao &, Li (2021) In: Hu S. (eds) Cancer Metabolomics., vol 1280. Springer

Untargeted Analysis

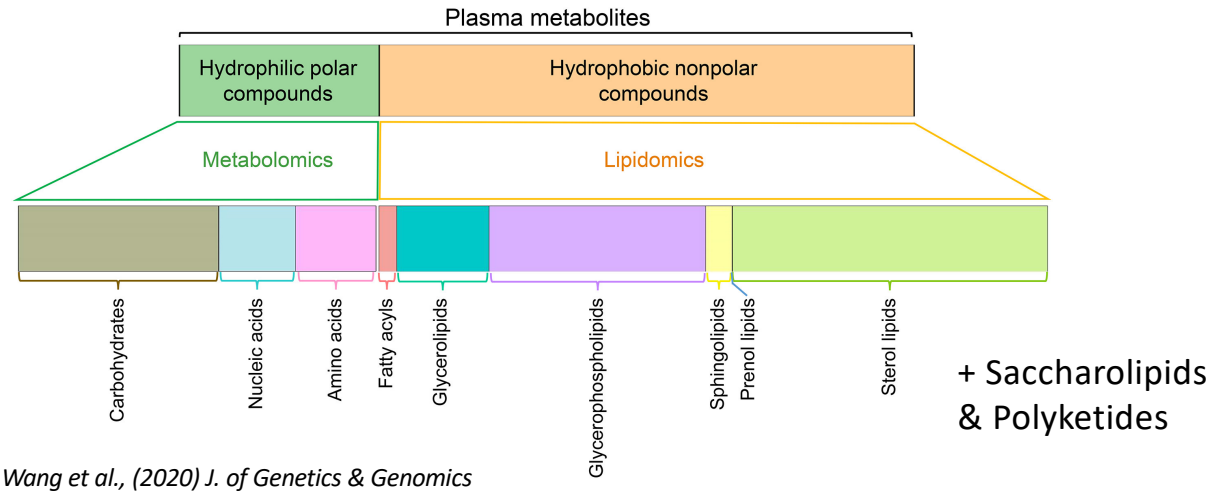
+ Allows for novel discovery and comprehensive coverage

- Challenges include
- time to process extensive amounts of raw data
- difficulties in identifying unknown small molecules
- coverage of the platform
- bias towards detection of high-abundance molecules

Can be followed by targeted analysis to validate top candidates

Lipidomics

- Profiling lipids requires own sample preparation, analytical protocol & data processing
- Within each lipid class, many variations of different length of carbon chains & number of double bonds
- Leads to particular patterns in mass spectra data, leveraged in data processing



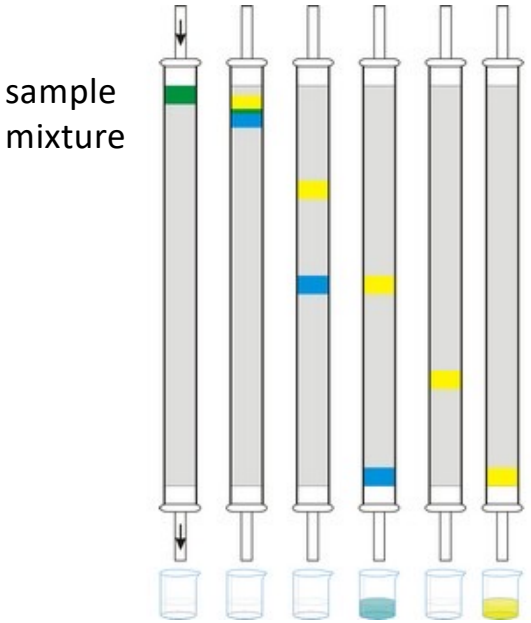
Wang et al., (2020) J. of Genetics & Genomics

Outline

- Introduction
- Technology Types
- Mass Spectrometry
- Analysis Methods

Sample Preparation - Chromatography

- No digestion as in proteomics (to obtain peptides)
- Chromatography
  - Spreads out analytes over time
  - Less likely that ionization capacity overcome by large quantities of analyte
  - Two phases
    - Stationary Phase – analyte separation
    - Mobile Phase – carries analytes through chromatographic column to the mass spectrometer
  - Mobile phase: gas or liquid

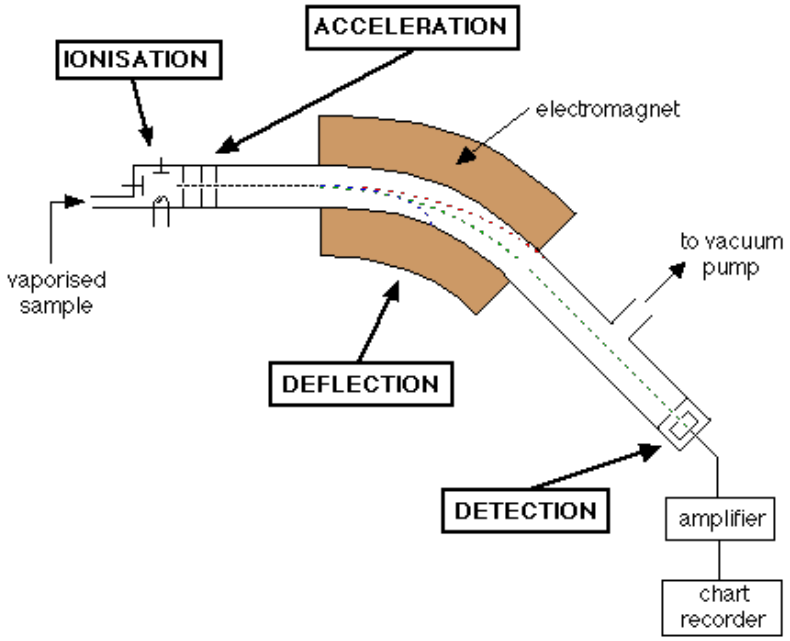


https://glossary.periodni.com

Mass Spectrometry Basic Principles

Atoms and molecules can be deflected by magnetic fields if ionized (electrically charged)

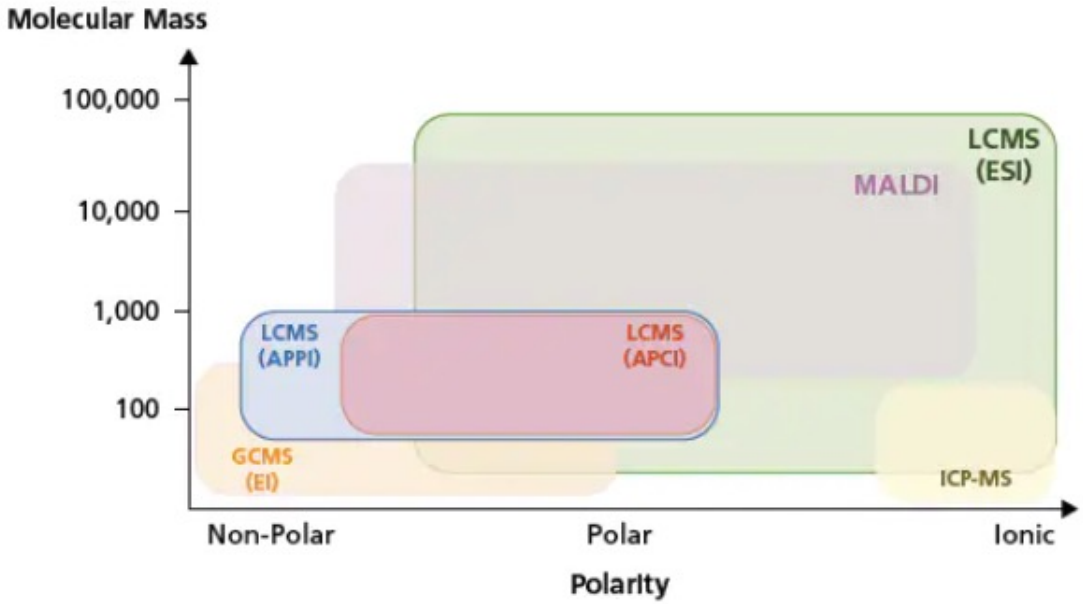
1. Ionization – atom/molecule acquiring negative/positive charge
   Targeted: Only ions of preselected interest sent to the detector
   Untargeted: Reads all ions across a wide m/z range
2. Acceleration – ions are accelerated
3. Deflection – ions are deflected by magnetic field
4. Detection – beam of ions are detected electrically
   Types : Time of Flight (TOF), orbitrap
   High Resolution: 4 decimal places offers molecular formula determination



https://www.chemguide.co.uk/analysis/masspec/howitworks.html

Ionization Step

During MS, analytes are ionized:



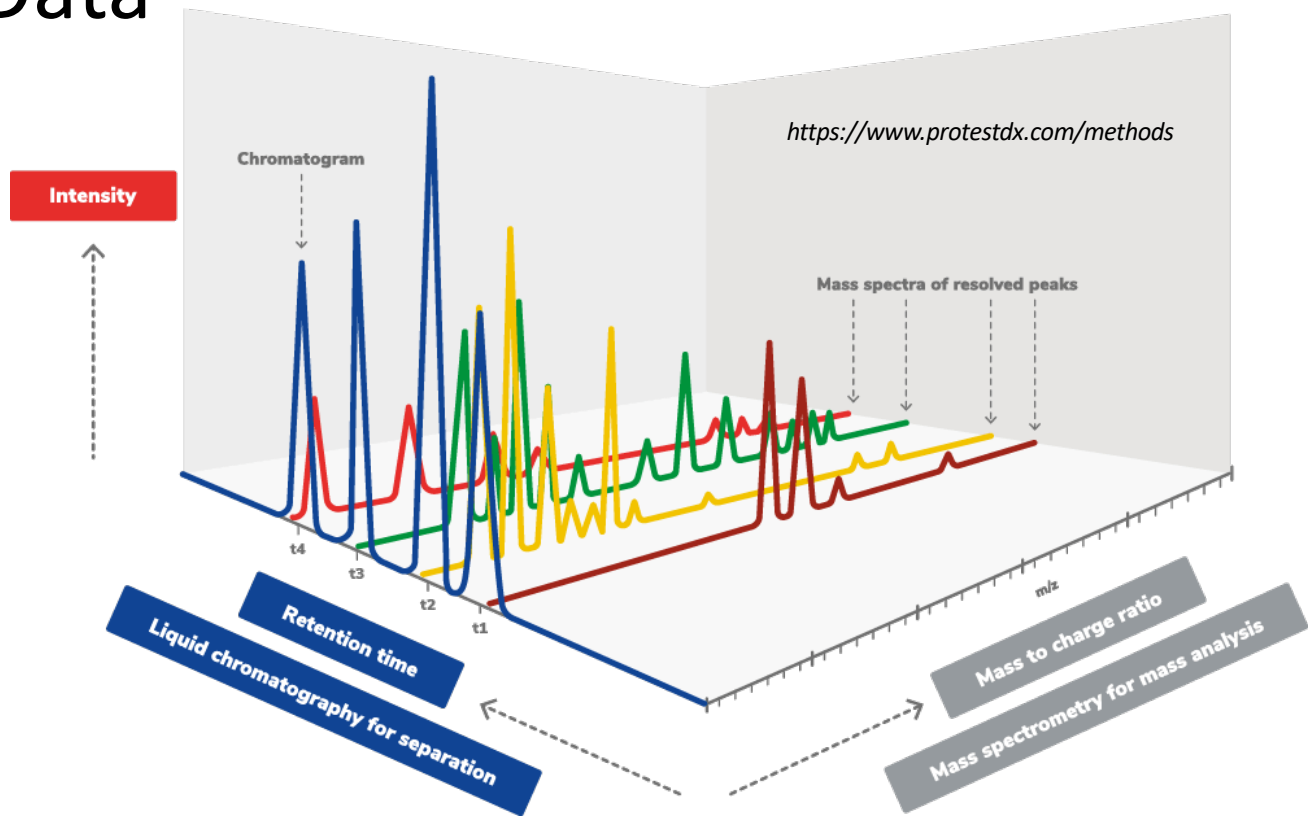
https://www.shimadzu.com/

Electrospray ionization (ESI) (most popular)

Matrix-Assisted Laser Desorption Ionization (MALDI)

Atmospheric Pressure Chemical Ionization (APCI)

Raw Data



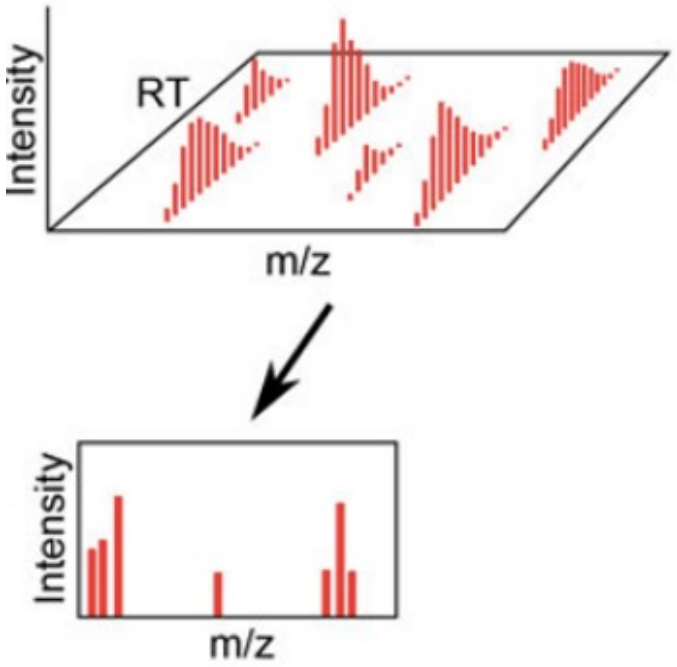
Retention Time (RT): time for solute to pass through chromatography column, time from injection to detection.

MZ: mass to charge ratio of compound (m/z)

Raw Data

Compound identified based on RT and MZ

For fixed RT: m/z ratio (x-axis) by relative current (y-axis: relative abundance or intensity)



Smith et al. (2014) BMC Bioinformatics
https://www.chemguide.co.uk/analysis/masspec/howitworks.html

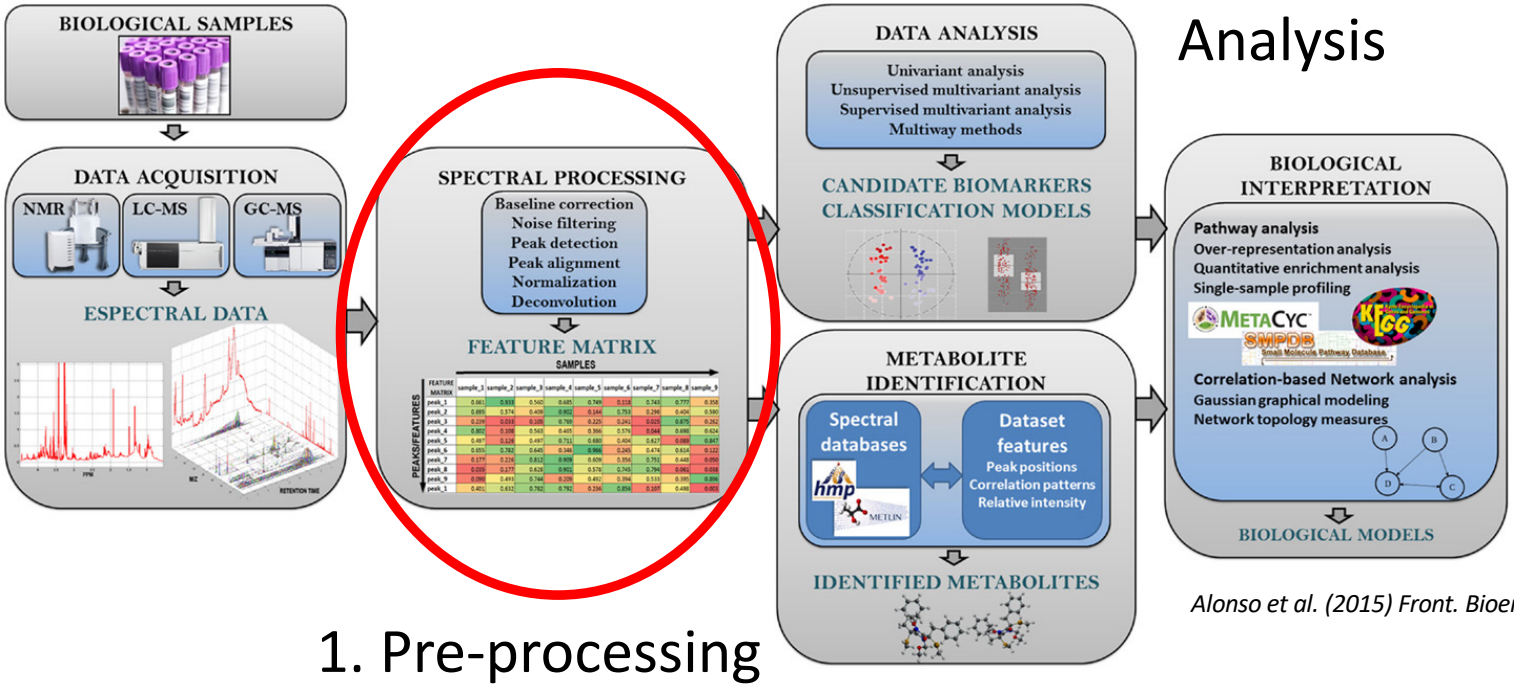
Raw Data

- May be very large files (& binary)
- Often proprietary (only with commercial software)
- Open source formats
  - ANDI NetCDF, mzXML, mzData
  - Converters available (msConvert, FileConverter, etc)

Outline

- Introduction
- Technology Types
- Mass Spectrometry
- Analysis Methods

Pipeline



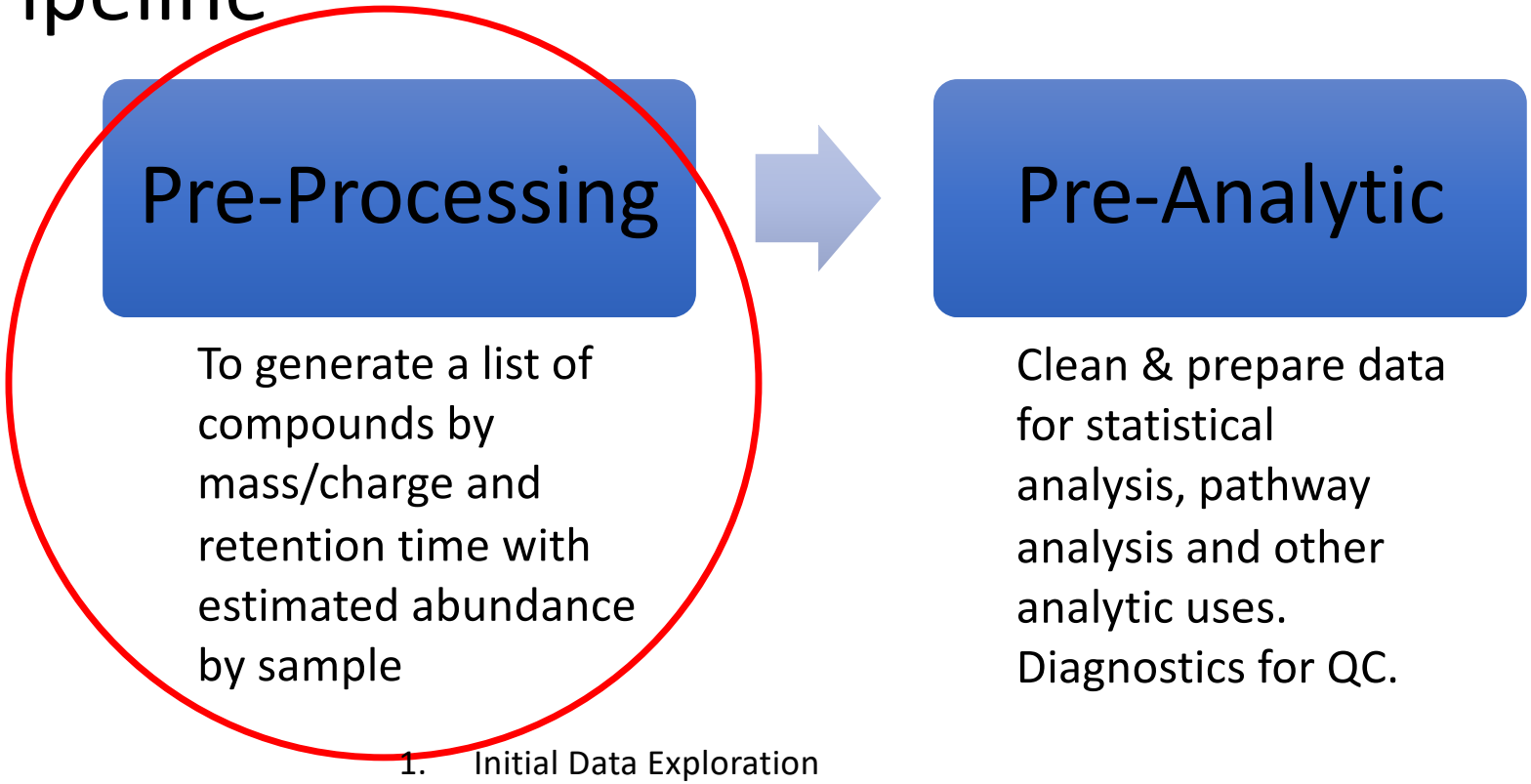
4. Downstream Analysis

1. Pre-processing
2. Pre-analytic

3. Identification

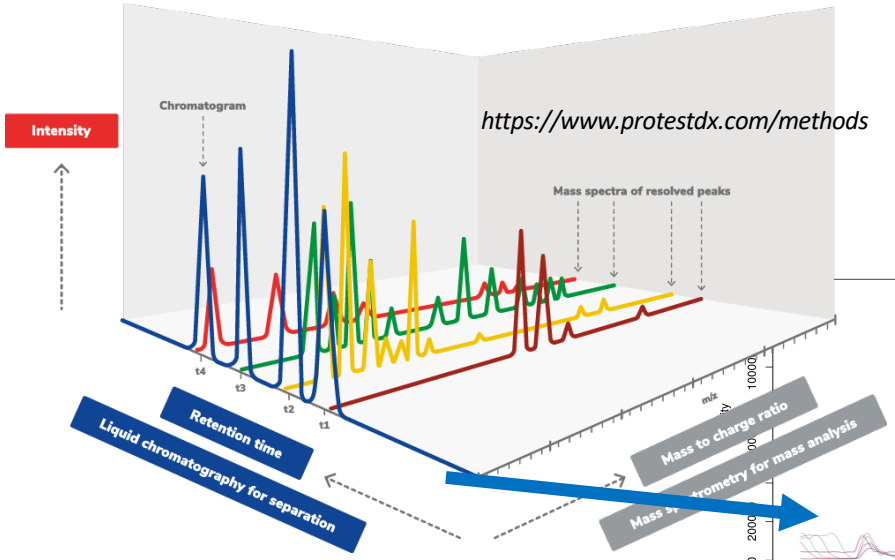
Alonso et al. (2015) Front. Bioeng. Biotechnol.

Pipeline


Pre-Processing
Pre-Analytic
To generate a list of compounds by mass/charge and retention time with estimated abundance by sample
Clean & prepare data for statistical analysis, pathway analysis and other analytic uses. Diagnostics for QC.
1. Initial Data Exploration
2. Baseline Correction
3. Peak Detection/Alignment
4. Drift and Normalization

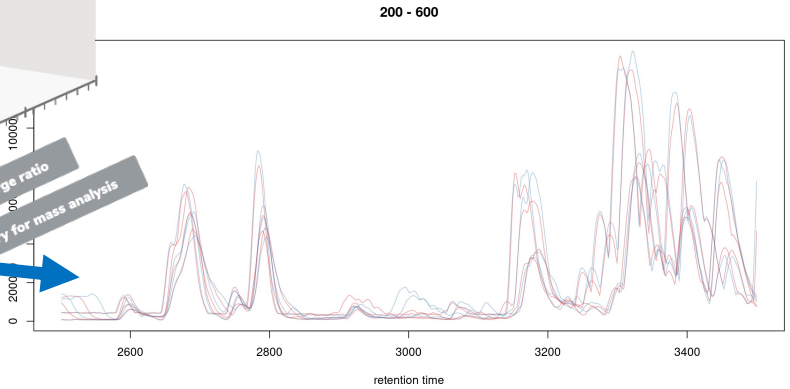
1. Initial Data Exploration

- To look at overall patterns & identify problematic samples
- Chromatogram building



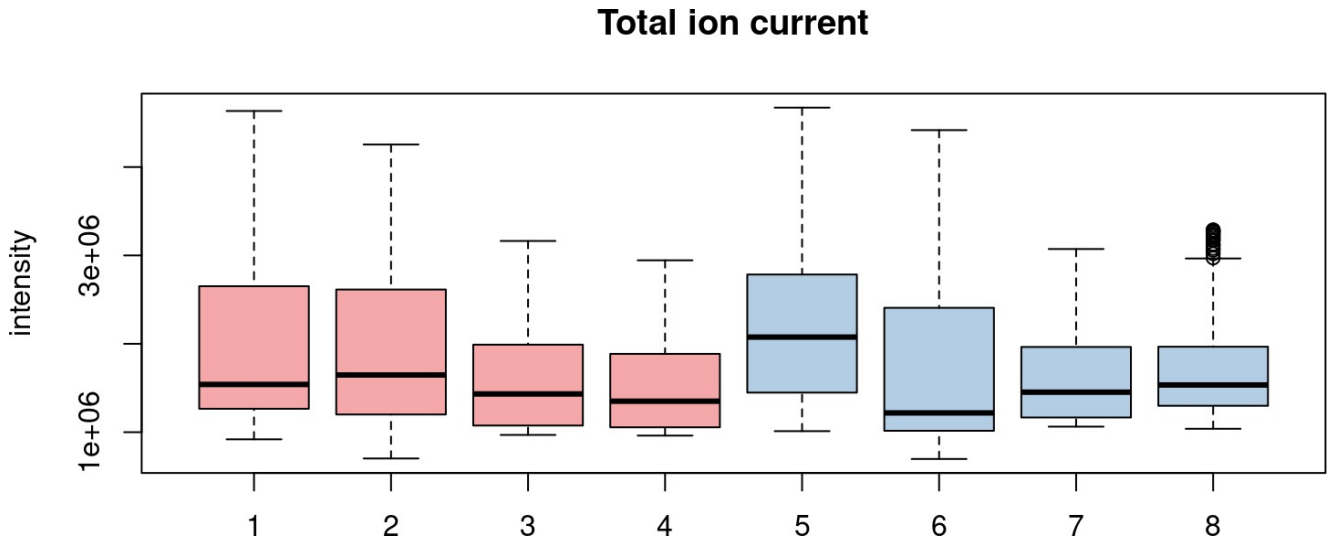
https://www.protestdx.com/methods

https://www.bioconductor.org/ xcms



1. Initial Data Exploration

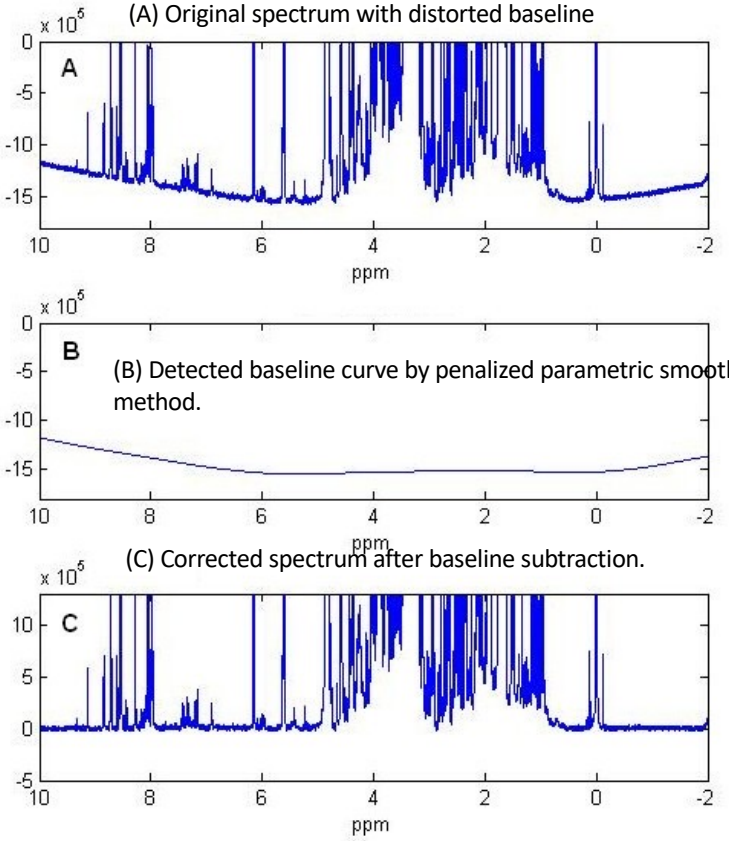
- Overall Intensity



https://www.bioconductor.org/ xcms

2. Baseline Correction

(A) Original spectrum with distorted baseline



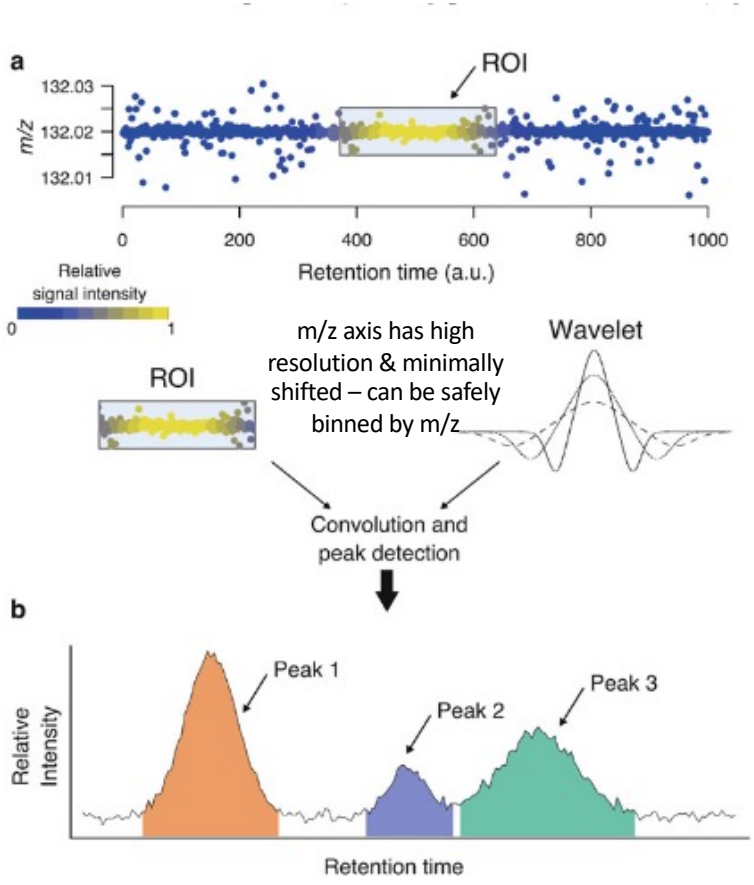
(B) Detected baseline curve by penalized parametric smoothing method.

(C) Corrected spectrum after baseline subtraction.

Xi & Rocke (2008) BMC Bioinformatics

3. Peak Detection

- By Sample
- Determine peaks above specified noise level that are true compounds
- Returns RT, m/z and integrated intensity for each peak

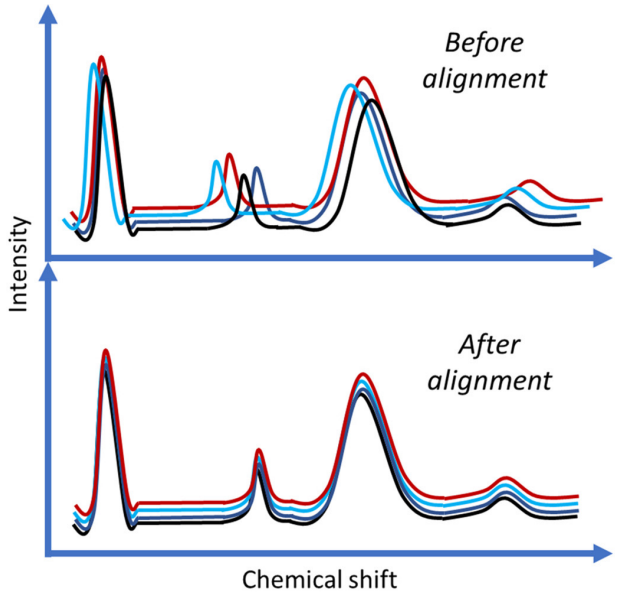


S. Li Editor (2020) Computational Methods and Data Analysis for Metabolomics

Fig. 3 The centWave peak detection overview. First, regions of interest (ROI) are detected. For each detected ROI, a wavelet filter is applied and peaks within the ROI are detected

3. Peak Alignment

- Time of elution can vary across samples
- Alignment refers to find corresponding signals for the same compound across samples
- Often non-linear shifts



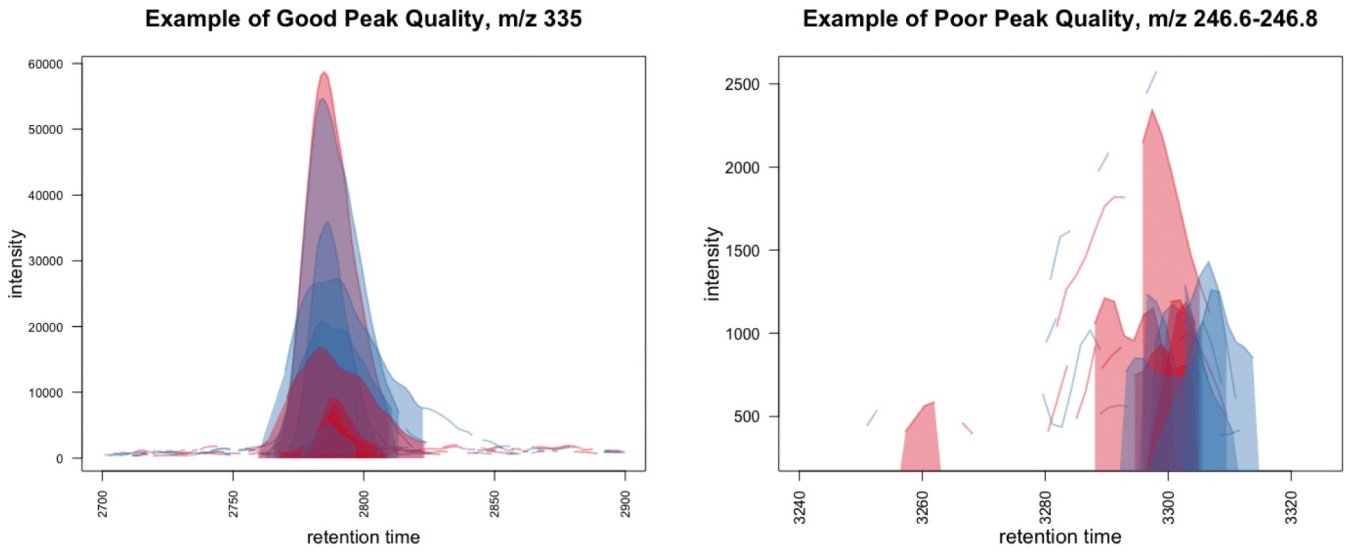
Karaman et al. (2018) Comprehensive Analytical Chemistry

3. Peak Alignment

- Before peak detection – spectral alignment
  - Applying ‘warping’ methods (stretch/shrink), non-linear transformations to the RT to maximize correlation between spectra; Fast-Fourier transform methods
- After peak detection – peak alignment
  - Applied to peak coordinates, identify RT boundaries of peaks; use kernel density estimators

Alonso et al. (2015) Frontiers in Bioeng & Biotech.

3. Peak Quality



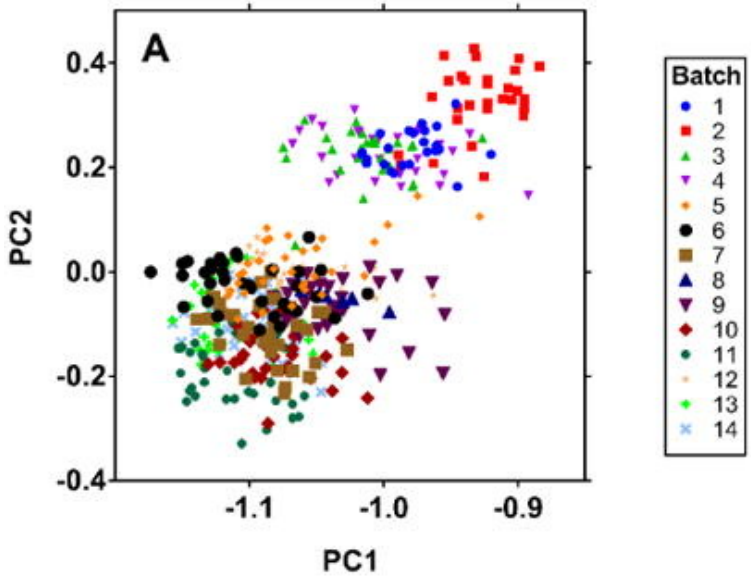
(a) High Quality Feature

(b) Low Quality Feature

Schiffman et al., (2019) BMC Bioinformatics

4. Intensity drift

- Instrument response changes with time
- “Drift” can occur when samples are collected over multiple batches or long period of times



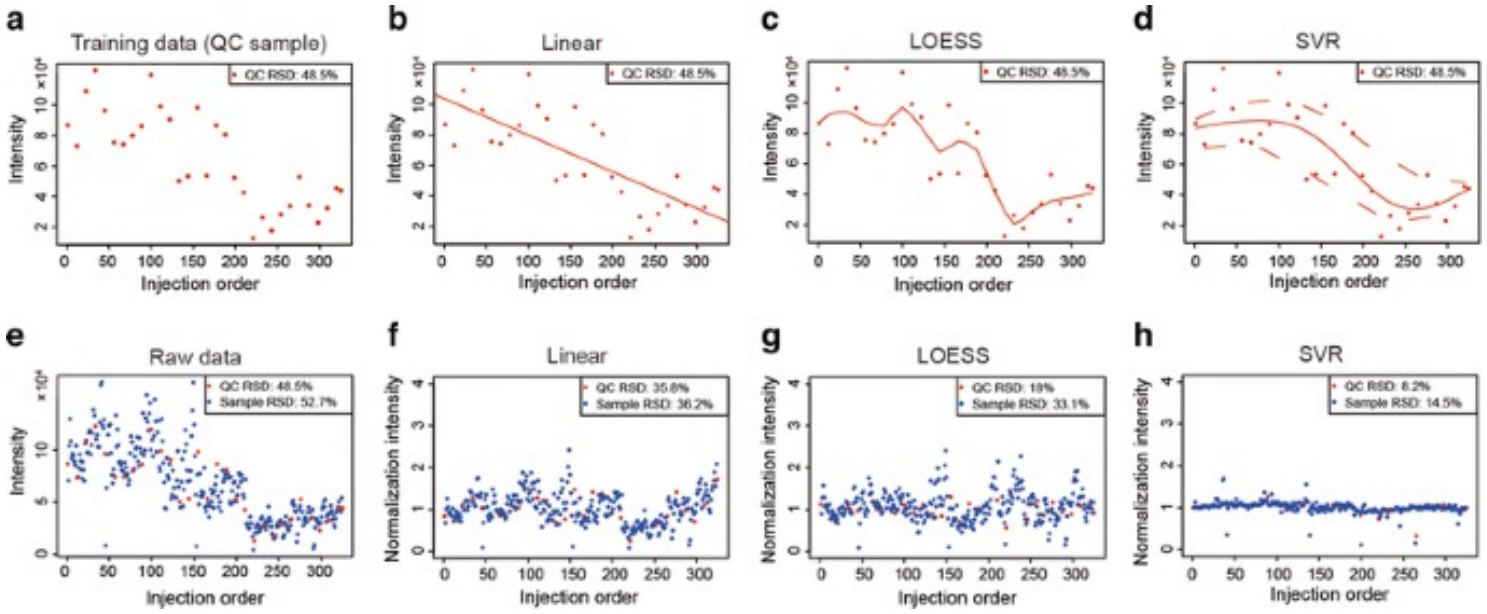
Thonusin et al., (2017) J. Chromatogr A

4. Strategies for Normalization

- Sum (or median) metabolite signal intensity normalization
  - assumes all metabolites experience same pattern of drift and that and that sum/median is approximately same in all samples
- Use Internal Standards (IS)
  - Not practical to include an isotope-labeled IS for all metabolites of interest.

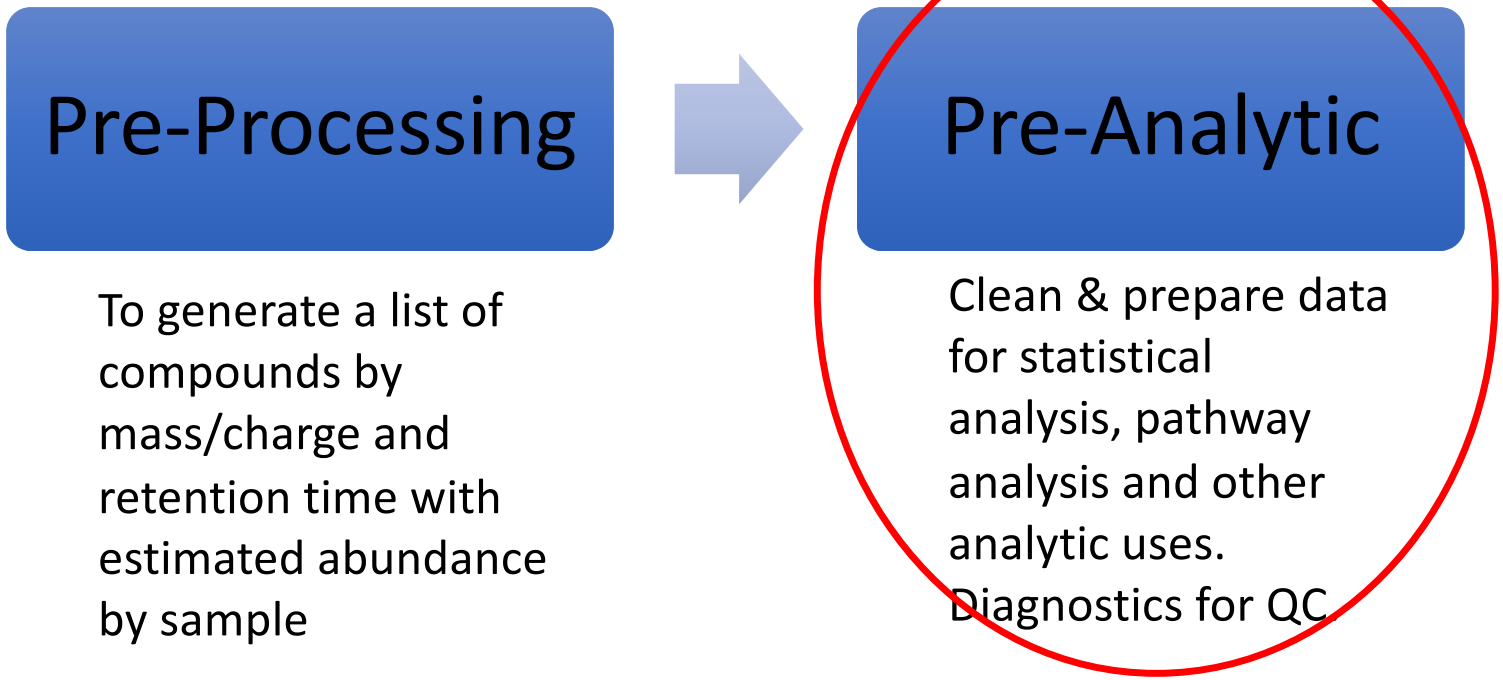
4. Strategies for Normalization

- Quality Control Samples
  - Included in each batch, can monitor and adjust for drift



Shen et al. (2016) Metabolites

Pipeline

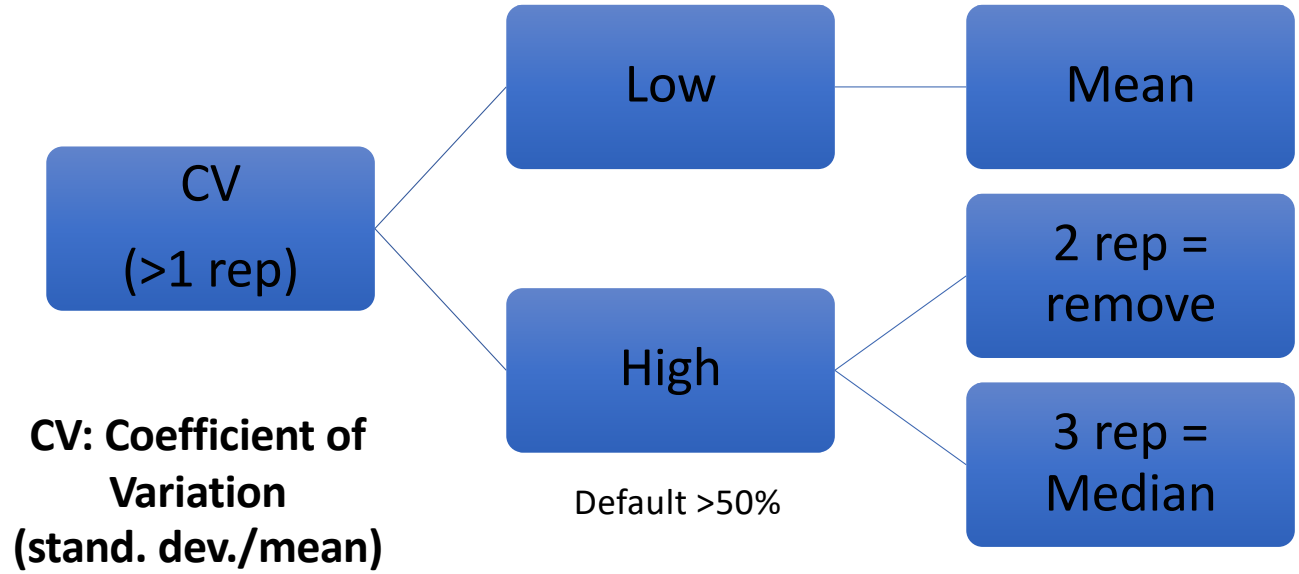


To generate a list of compounds by mass/charge and retention time with estimated abundance by sample

Clean & prepare data for statistical analysis, pathway analysis and other analytic uses. Diagnostics for QC.

1. Summarization of Replicates
2. Filter Compounds
3. Imputation
4. Normalization/Batch Adjustment

1. Summarization of Replicates



Summarize technical replicates to result in one measure / subject / compound

Hughes et al. (2014) Bioinformatics MSPrep

1. Summarization of Replicates

Most compounds with high CV have two consistent and one extreme observation.

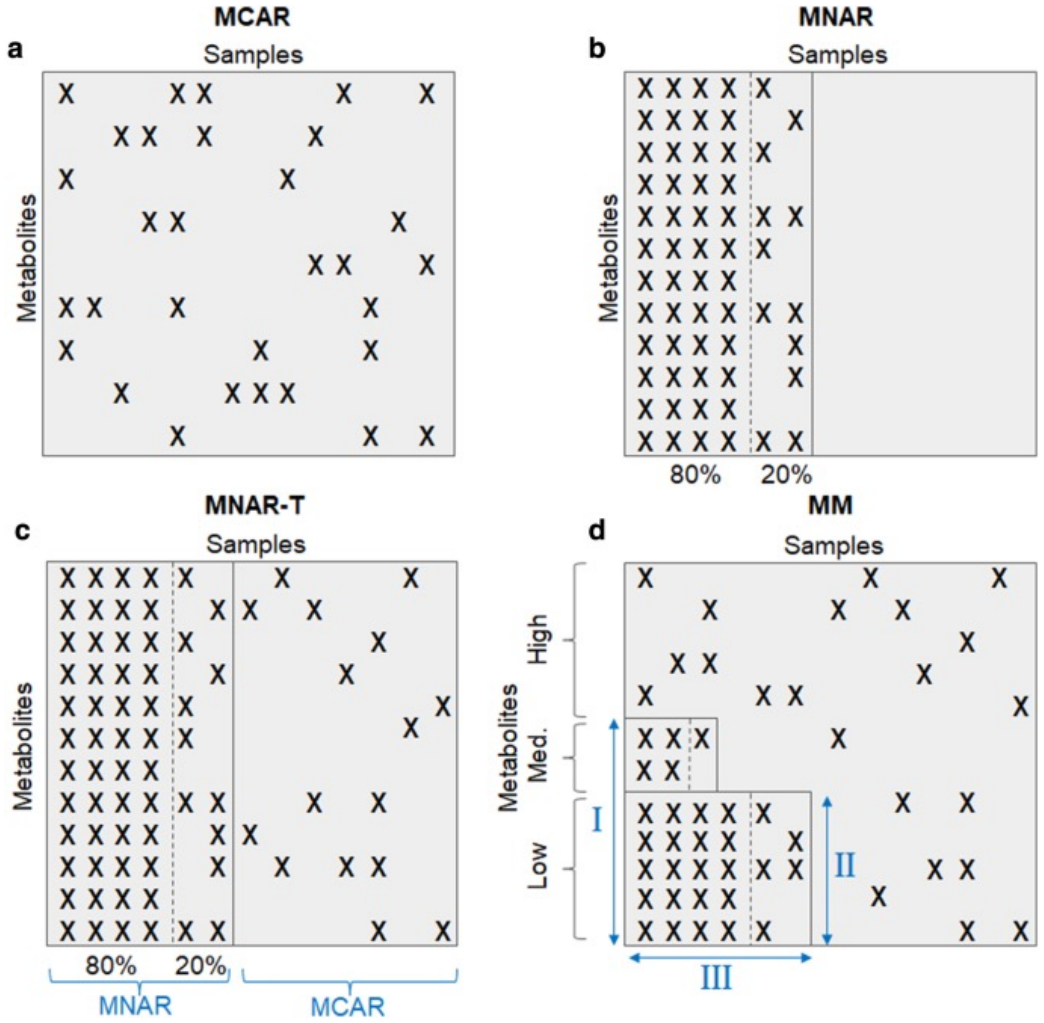
Replicates

Low	*Med*	*High*	*Mean*	*SD*	*CV*	*Median*
20839	**30624**	251768	101077	130594	129.20%	30624
101193	**101761**	448747	217234	200497	92.30%	101761
9734	**263033**	**264957**	179241	146801	81.90%	263033
98592	**467745**	**468810**	345049	213439	61.86%	467745

Hughes et al. (2014) Bioinformatics MSPrep

2. Missing Data

- Three modes of missing data
  - Compound not present in sample
  - Compound present, but below detectable limit or specified noise level
  - Compound failed to be identified
    - Mechanical error (lab prep, LC column, coelution, etc.)
    - Errors in pre-processing (misalignment, retention time shift, mass accuracy, etc.)
- Missing Completely at Random (MCAR)
- Missing Not at Random (MNAR)



Lee & Styczynski, (2018) Metabolomics

2. Missing Data: Filtering Compounds

Goal: Remove compounds with too many missing values.

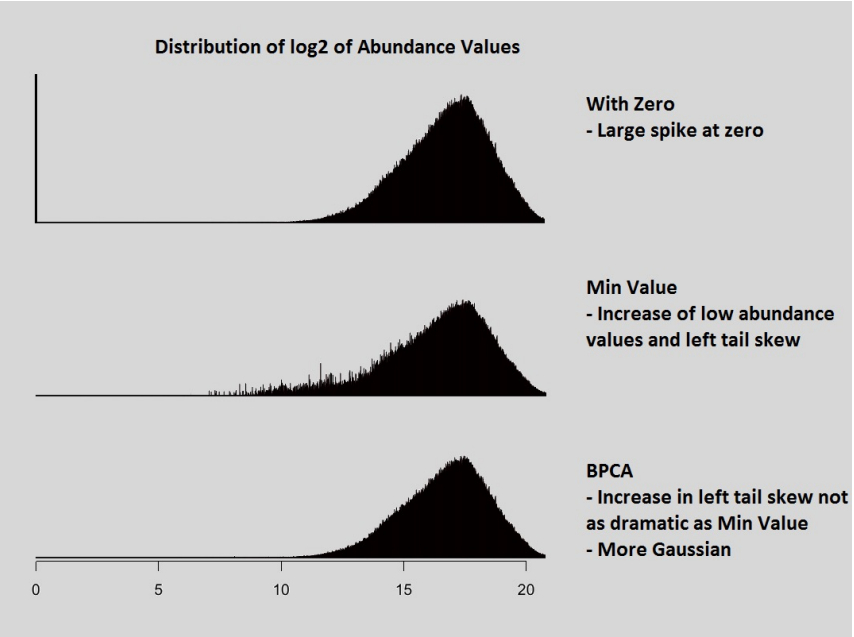
- Include compounds found in > X% of subjects (e.g., 80%)
  - Data driven approaches (Schiffman et al., (2019) BMC Bioinformatics)
- Removes noise detected as compounds, uncommon food or drug metabolites
- Simplifies downstream analysis
- Impact on downstream analysis
  - Too stringent reduces power and may miss important compounds
  - Too lenient increases false positives and drug/food metabolites, etc.

3. Imputation

- Downstream analyses may need complete data (e.g., machine learning methods)
- Strategies for Imputation
  I. With Zero – Zeros remain in dataset.
     - Normalization & some statistical analyses no longer valid.
     - Present/absent modeling.
  II. Minimum Value – All missing values are replaced with ½ the minimum observed value for that compound.

3. Imputation

III. Random Forests
IV. K-Nearest Neighbors
V. Bayesian Principal Component Analysis (BPCA)



Kokla et al (2019) BMC Bioinformatics

4. Normalization/Batch Effects

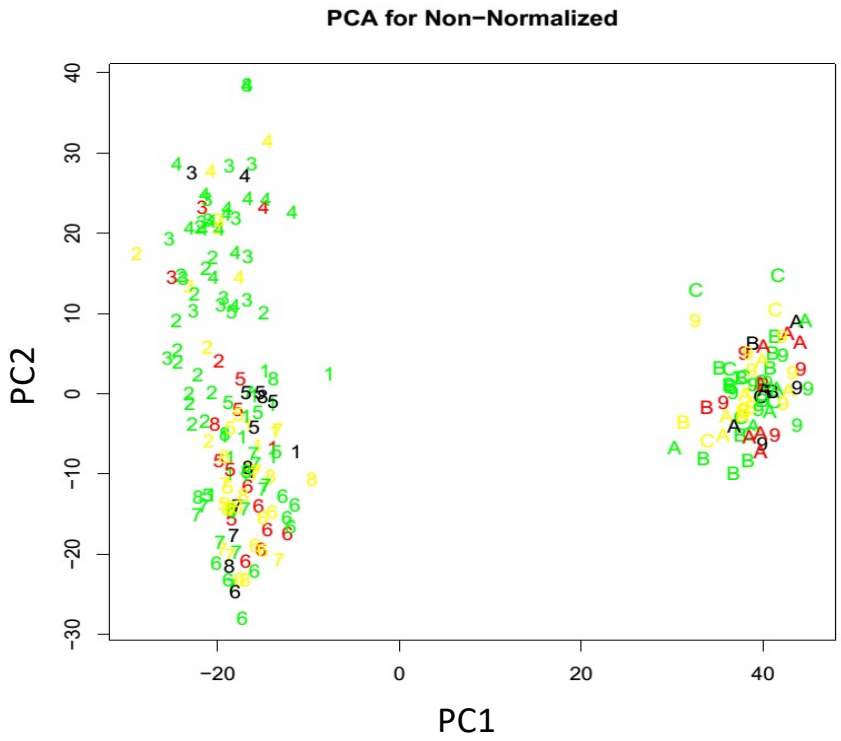
- Although QC samples can adjust for drift (see earlier slides), may have additional variation due to measured (e.g., batch effects) and unmeasured factors
- Utilize internal controls to measure variation.
  - Spiked controls or biological controls
- Data driven strategies
- Batch Effects
  - E.g., ComBat: Empirical Bayes Method to directly estimate and remove batch effect, can be applied post-hoc to any adjusted data

4. Normalization: Diagnostic Plots

- Principal Component Analysis (PCA)
- Boxplots of abundance by subject

Example: Color = Disease stage,

Symbol = Batch



4. Normalization

Strategies for Normalization

A. No Normalization
   - Data prepared by one technician, run in one batch, and/or specific compounds (small count) potentially may not need normalization

B. Median Normalization
   - Simplistic, but commonly used method
   - Location shift accomplished by centering each subjects abundance on the median
   - Q75 often selected instead of median if raw data is skewed or large quantity missing data

4. Normalization

C. Quantile Normalization

D. Surrogate Variable Analysis (SVA)

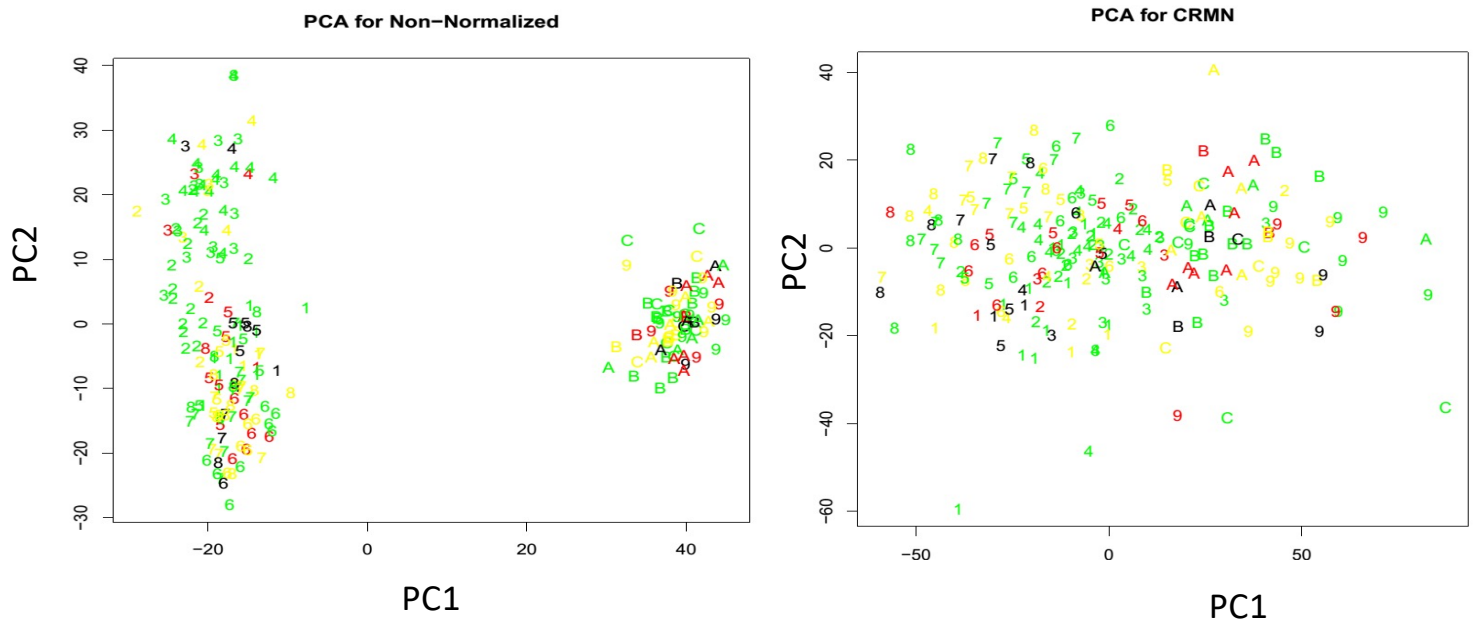
E. Remove Unwanted Variation – metabolomics (RUVm)

G. Cross Contribution Normalization (CRMN)

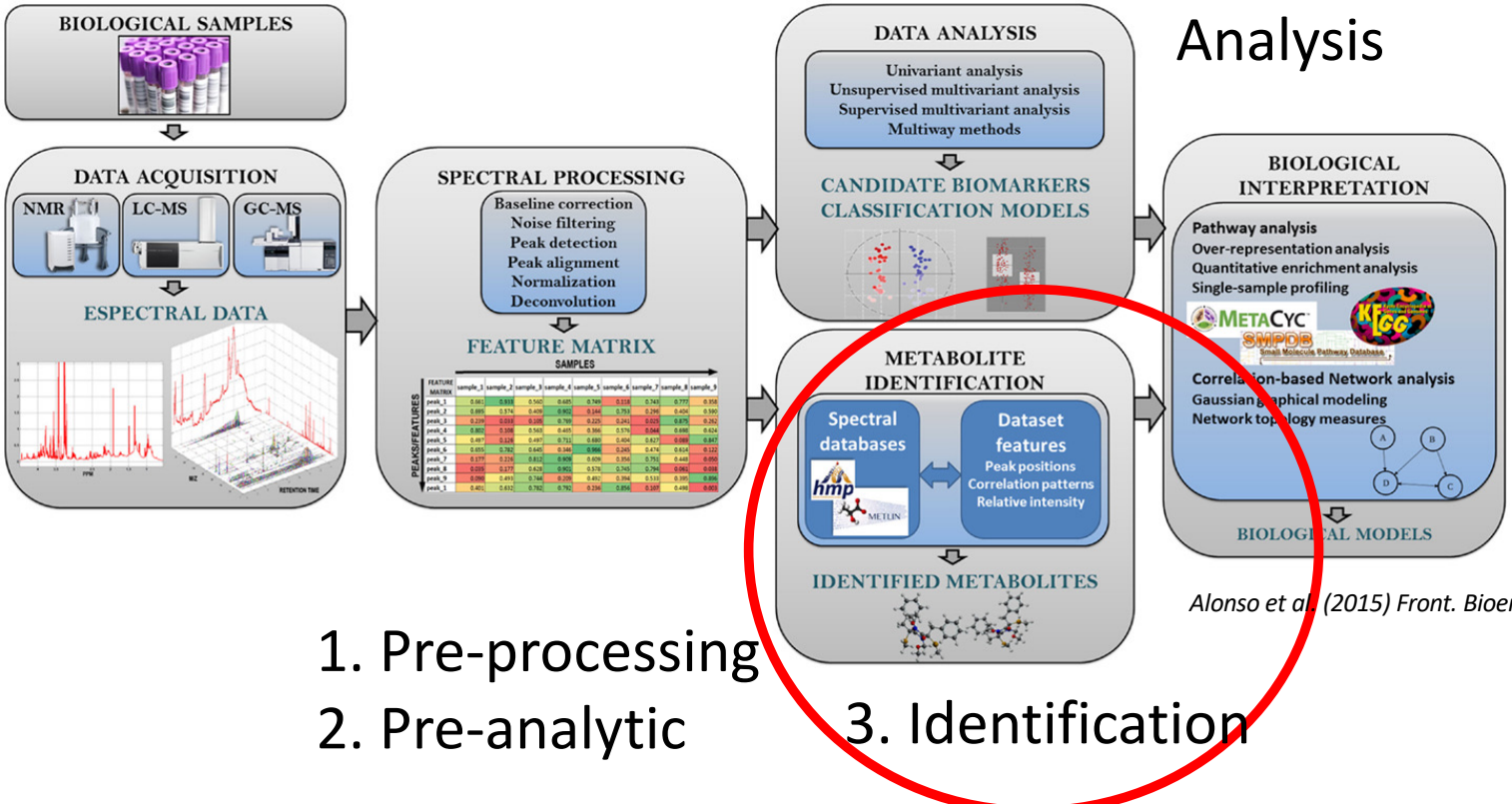
- Changing abundance of compounds is known to have potential to effect on other compounds
- So some variation in internal standards could be due to variation in other compounds
- Estimates the variance due to unmeasured sources similar to RUV, and estimates any significant cross contributing variance that comes from non-spiked compounds

Example

Color = Disease Stage, Symbol = Batch



Pipeline



4. Downstream Analysis

1. Pre-processing
2. Pre-analytic
3. Identification

Alonso et al. (2015) Front. Bioeng. Biotechnol.

Identification

- **Goal:** identify most likely compound appearing at particular mass & retention time
- Anytime prior to pathway analysis

- Most based strictly off mass detected
- Often multiple potential matches and scoring algorithm can determine best match
- Major challenge in untargeted metabolomics

Databases

- Facilitate metabolite identification
  - metabolite information - chemical formula, names and synonyms, physical and chemical properties
  - classification systems
  - structures and images
  - links to other databases (e.g., KEGG, PDB, CHeBI)
  - experimental data, mass spectrometry data

Identification - Databases

- **HMDB** – Human Metabolome Database
  - http://www.hmdb.ca/
- **Metlin**
  - https://metlin.scripps.edu/
- **mzCloud**
  - https://www.mzcloud.org
- **MassBank/MoNA**
  - http://www.massbank.jp
- **NIST 17**
  - https://chemdata.nist.gov
- **Lipid MAPS** – Lipid Metabolites And Pathways Strategy
  - http://www.lipidmaps.org/

Database	Compounds with spectra	Number of Spectra
NIST 17	13,808	574,826
METLIN	>200,000	NA
MoNA	~75,000	261,917
mzCloud	> 8000	~2,000,000[a]
HMDB	2265	22,247[b]
LipidBlast	119,200	212,516

[a]The number of recalibrated spectra in mzCloud
[b]Total experimental LC-MS/MS spectra in HMDB

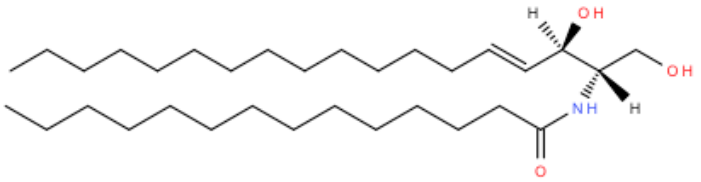
S. Li Editor (2020) Computational Methods and Data Analysis for Metabolomics

Example

Ceramide
Cer(d18:1/14:0)
from LIPID Maps

Structure database (LMSD)

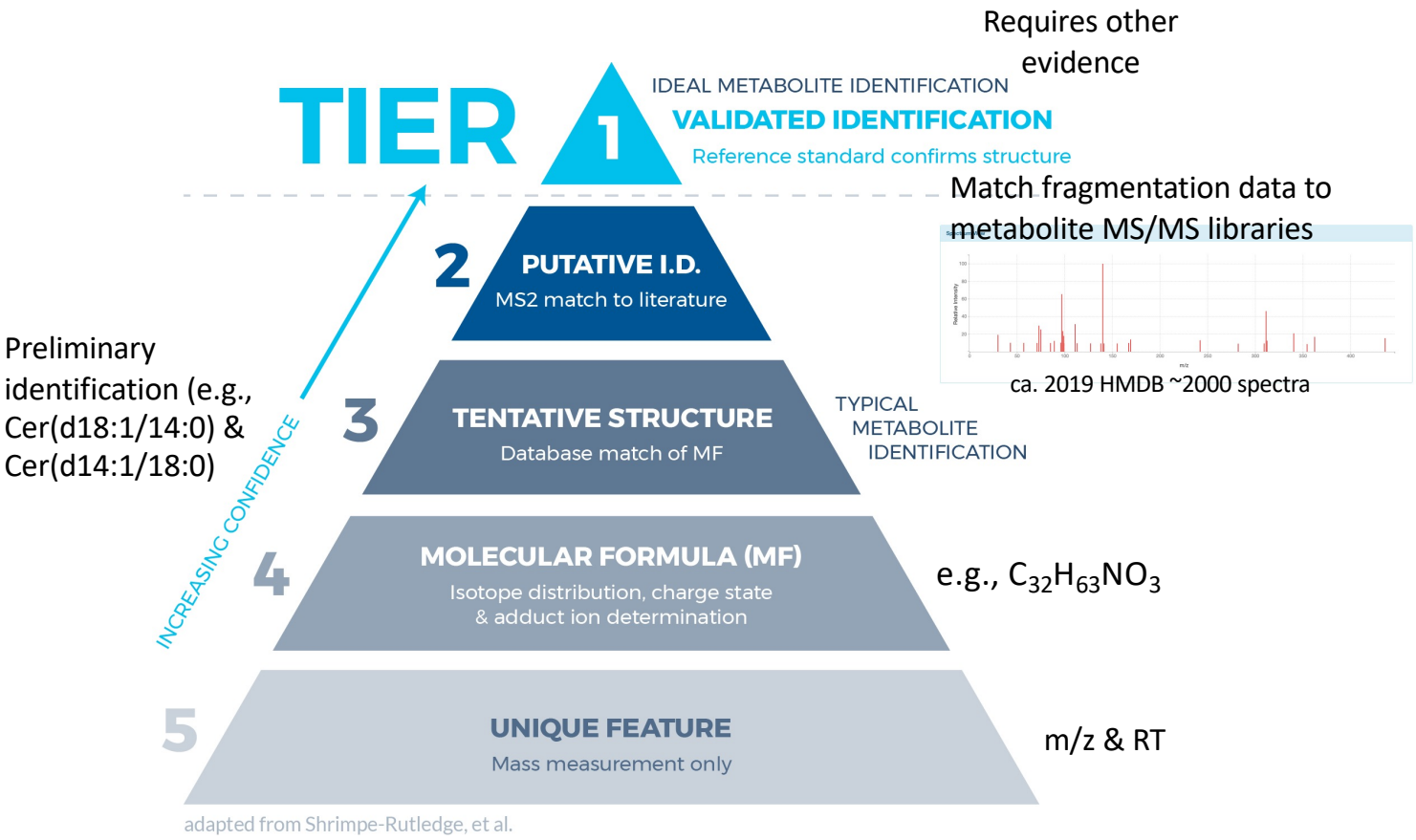
Download file | MDLMOL | MDLMOL files can be opened in various drawing programs. In ChemDr open as filetype "MDL Molfile".

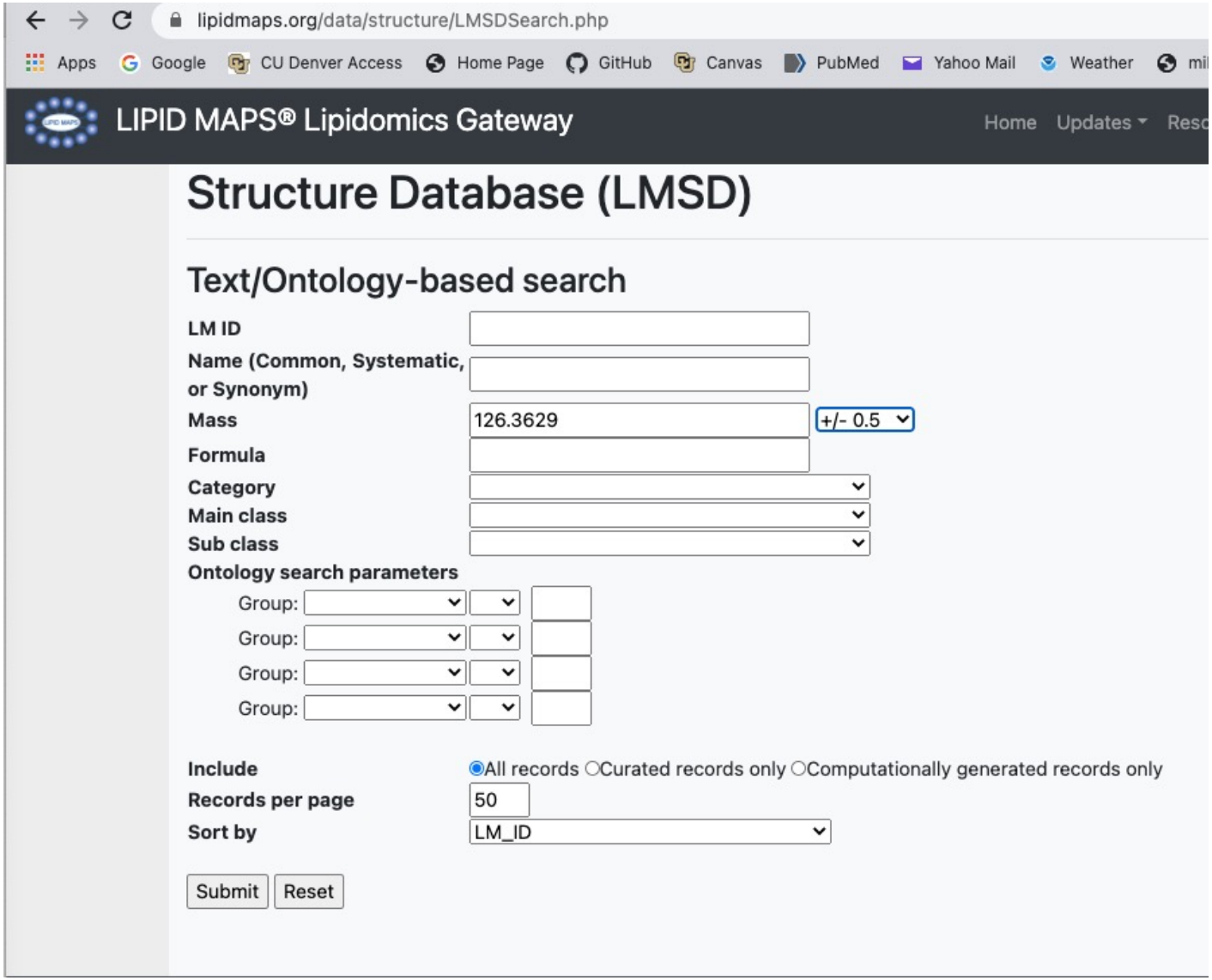


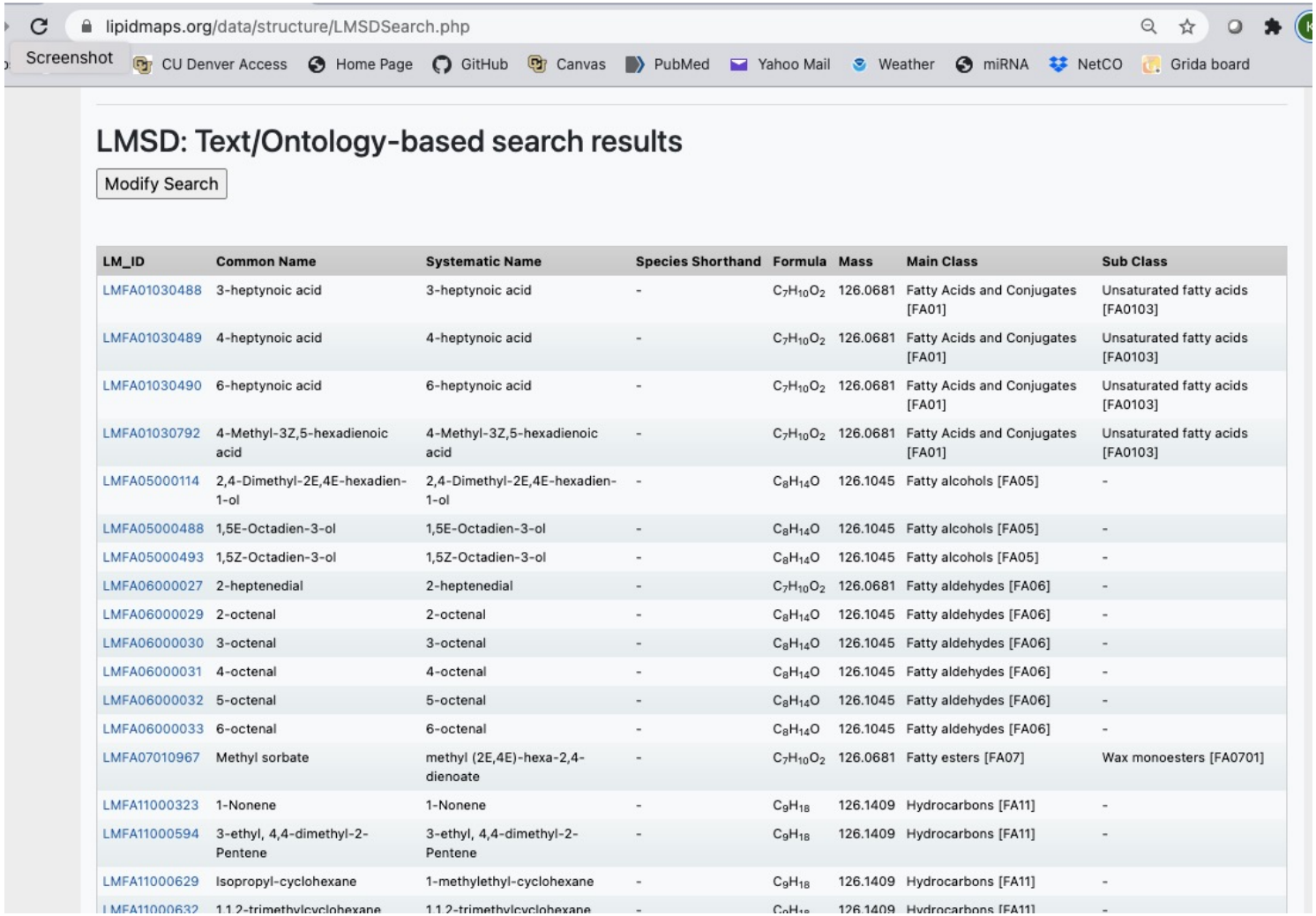
LM ID	LMSP02010001
Common Name	Cer(d18:1/14:0)
Systematic Name	N-(tetradecanoyl)-sphing-4-enine
Synonyms	C14 Cer; N-(tetradecanoyl)-ceramide; N-(myristoyl)-ceramide; Cer[NS]
Exact Mass	509.4808 (neutral) **Calculate m/z:** (Select m/z)
Formula	$C_{32}H_{63}NO_3$
Category	Sphingolipids [SP]
Main Class	Ceramides [SP02]
Sub Class	N-acylsphingosines (ceramides) [SP0201]
PubChem Compound ID (CID)	5282310
METABOLOMICS ID	-
KEGG ID	-
HMDB ID	HMDB11773

http://www.lipidmaps.org/data/LMSDRecord.php?LMID=LMSP02010001

Levels of Identification




lipidmaps.org/data/structure/LMSDSearch.php
Apps
Google
CU Denver Access
Home Page
GitHub
Canvas
PubMed
Yahoo Mail
Weather
LIPID MAPS® Lipidomics Gateway
Home
Updates
Structure Database (LMSD)
Text/Ontology-based search
LM ID
Name (Common, Systematic, or Synonym)
Mass
126.3629
+/- 0.5
Formula
Category
Main class
Sub class
Ontology search parameters
Group:
Group:
Group:
Group:
Include
All records
Curated records only
Computationally generated records only
Records per page
50
Sort by
LM_ID
Submit
Reset



LMSD: Text/Ontology-based search results

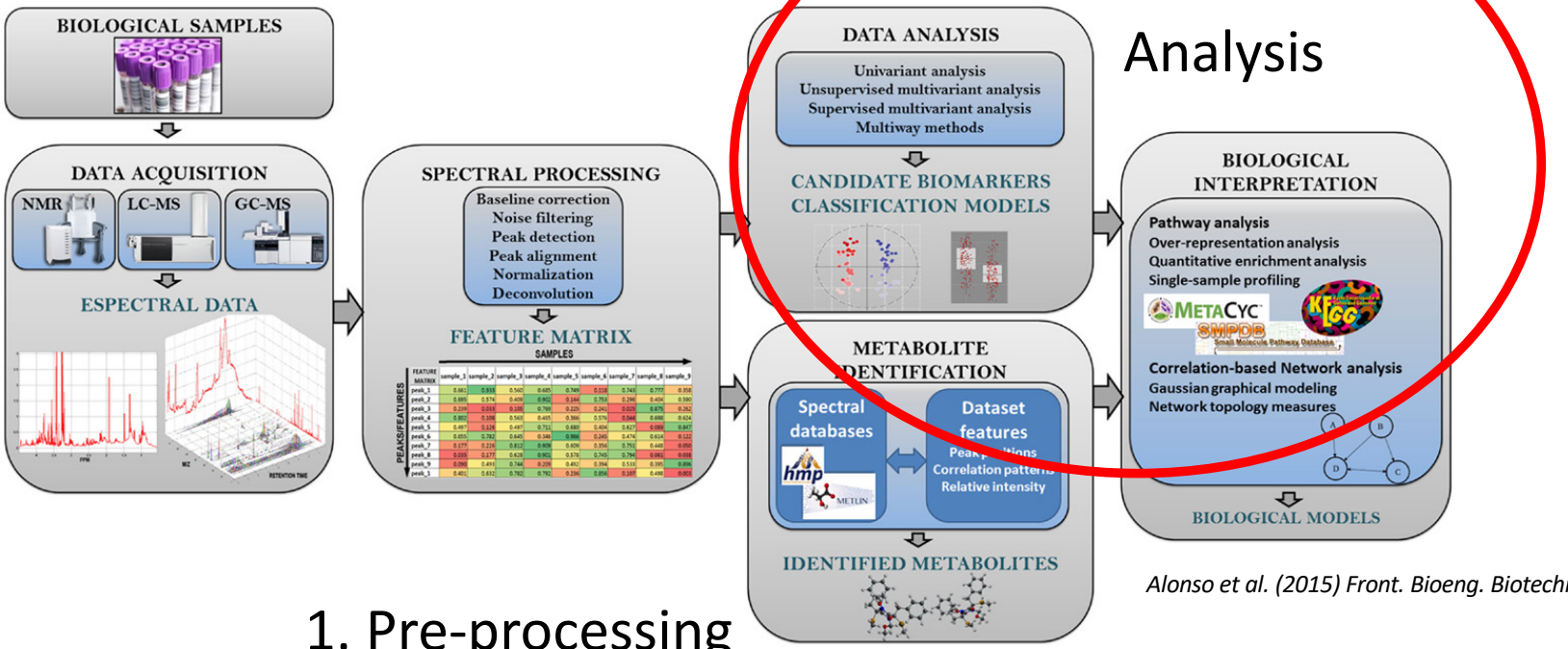
Modify Search

LM_ID	Common Name	Systematic Name	Species Shorthand	Formula	Mass	Main Class	Sub Class
LMFA01030488	3-heptynoic acid	3-heptynoic acid	-	$C_7H_{10}O_2$	126.0681	Fatty Acids and Conjugates [FA01]	Unsaturated fatty acids [FA0103]
LMFA01030489	4-heptynoic acid	4-heptynoic acid	-	$C_7H_{10}O_2$	126.0681	Fatty Acids and Conjugates [FA01]	Unsaturated fatty acids [FA0103]
LMFA01030490	6-heptynoic acid	6-heptynoic acid	-	$C_7H_{10}O_2$	126.0681	Fatty Acids and Conjugates [FA01]	Unsaturated fatty acids [FA0103]
LMFA01030792	4-Methyl-3Z,5-hexadienoic acid	4-Methyl-3Z,5-hexadienoic acid	-	$C_7H_{10}O_2$	126.0681	Fatty Acids and Conjugates [FA01]	Unsaturated fatty acids [FA0103]
LMFA05000114	2,4-Dimethyl-2E,4E-hexadien-1-ol	2,4-Dimethyl-2E,4E-hexadien-1-ol	-	$C_8H_{14}O$	126.1045	Fatty alcohols [FA05]	-
LMFA05000488	1,5E-Octadien-3-ol	1,5E-Octadien-3-ol	-	$C_8H_{14}O$	126.1045	Fatty alcohols [FA05]	-
LMFA05000493	1,5Z-Octadien-3-ol	1,5Z-Octadien-3-ol	-	$C_8H_{14}O$	126.1045	Fatty alcohols [FA05]	-
LMFA06000027	2-heptenedial	2-heptenedial	-	$C_7H_{10}O_2$	126.0681	Fatty aldehydes [FA06]	-
LMFA06000029	2-octenal	2-octenal	-	$C_8H_{14}O$	126.1045	Fatty aldehydes [FA06]	-
LMFA06000030	3-octenal	3-octenal	-	$C_8H_{14}O$	126.1045	Fatty aldehydes [FA06]	-
LMFA06000031	4-octenal	4-octenal	-	$C_8H_{14}O$	126.1045	Fatty aldehydes [FA06]	-
LMFA06000032	5-octenal	5-octenal	-	$C_8H_{14}O$	126.1045	Fatty aldehydes [FA06]	-
LMFA06000033	6-octenal	6-octenal	-	$C_8H_{14}O$	126.1045	Fatty aldehydes [FA06]	-
LMFA07010967	Methyl sorbate	methyl (2E,4E)-hexa-2,4-dienoate	-	$C_7H_{10}O_2$	126.0681	Fatty esters [FA07]	Wax monoesters [FA0701]
LMFA11000323	1-Nonene	1-Nonene	-	C_9H_{18}	126.1409	Hydrocarbons [FA11]	-
LMFA11000594	3-ethyl, 4,4-dimethyl-2-Pentene	3-ethyl, 4,4-dimethyl-2-Pentene	-	C_9H_{18}	126.1409	Hydrocarbons [FA11]	-
LMFA11000629	Isopropyl-cyclohexane	1-methylethyl-cyclohexane	-	C_9H_{18}	126.1409	Hydrocarbons [FA11]	-
LMFA11000632	1,1,2-trimethylcyclohexane	1,1,2-trimethylcyclohexane	-	C_9H_{18}	126.1409	Hydrocarbons [FA11]	-

Search using $C_{16}H_{26}O_3$

LM_ID	Common Name	Systematic Name	Species Shorthand	Formula
LMFA01020279	10,11-epoxy-3,7,11-trimethyl-2E,6E-tridecadienoic acid	10,11-epoxy-3,7,11-trimethyl-2E,6E-tridecadienoic acid	-	$C_{16}H_{26}O_3$
LMFA01031023	7-Hydroxy-10E-Hexadecen-8-ynoic acid	7-Hydroxy-10E-Hexadecen-8-ynoic acid	-	$C_{16}H_{26}O_3$
LMFA01050143	Tetranor-12R-HETE	8R-hydroxy-4Z,6E,10Z-hexadecatrienoic acid	-	$C_{16}H_{26}O_3$
LMFA01050562	9S-HHTrE	9S-hydroxy-6Z,10E,12Z-hexadecatrienoic acid	-	$C_{16}H_{26}O_3$
LMFA01060229	-	4-oxo-2E,15-hexadecendioic acid	-	$C_{16}H_{26}O_3$
LMFA01070019	(2E,6E,10R,11S)-10,11-epoxy-3,7,11-trimethyltrideca-2,6-dienoic acid	10R,11S-epoxy-3,7,11-trimethyl-2E,6E-tridecadienoic acid	-	$C_{16}H_{26}O_3$
LMFA01070023	13,14-epoxy-7Z,10Z-Hexadecenoic acid	13,14-epoxy-7Z,10Z-Hexadecenoic acid	-	$C_{16}H_{26}O_3$
LMFA01150024	5-(3,4-dimethyl-5-pentylfuran-2-yl)-pentanoic acid	5-(5-pentyl-3,4-dimethylfuran-2-yl)-pentanoic acid	-	$C_{16}H_{26}O_3$
LMFA02010008	(9R,13R)-1a,1b-dinor-10,11-dihydro-12-oxo-15-phytoenoic acid	(1R,2R)-3-oxo-2-(2'Z-pentenyl) cyclopentanehexanoic acid	-	$C_{16}H_{26}O_3$
LMFA02010009	(9S,13S)-1a,1b-dinor-10,11-dihydro-12-oxo-15-phytoenoic acid	(1S,2S)-3-oxo-2-(2'Z-pentenyl) cyclopentanehexanoic acid	-	$C_{16}H_{26}O_3$
LMPR0103010004	Juvenile hormone III (W)	-	-	$C_{16}H_{26}O_3$

Pipeline



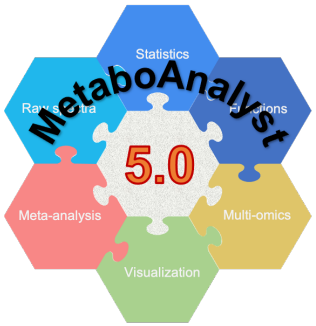
4. Downstream Analysis

1. Pre-processing
2. Pre-analytic
3. Identification

Alonso et al. (2015) Front. Bioeng. Biotechnol.

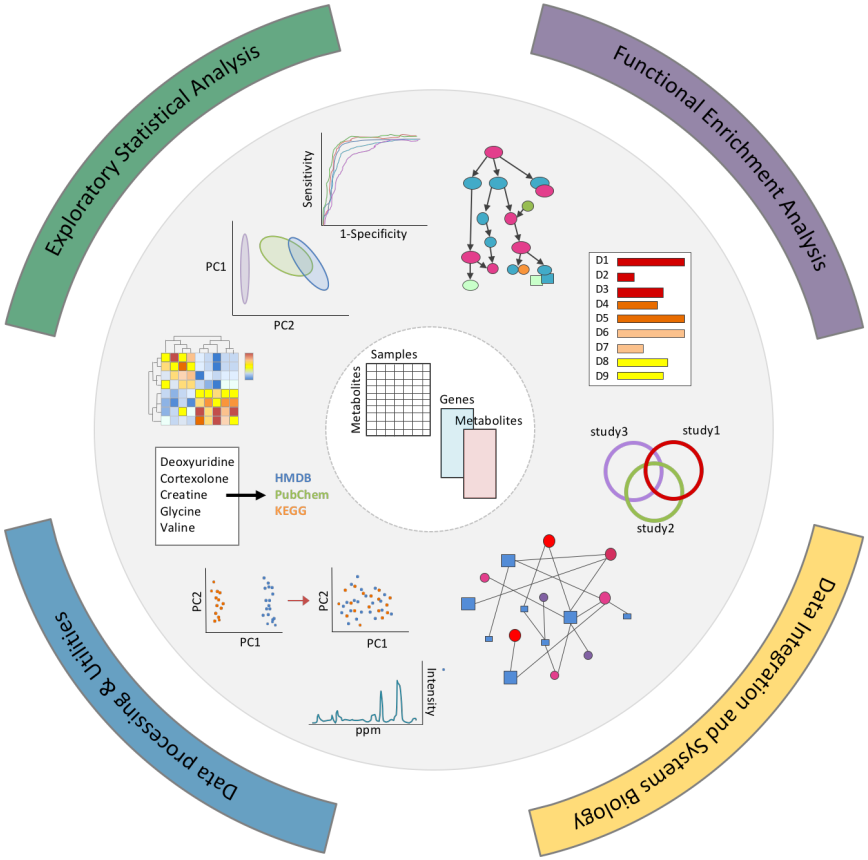
Downstream Analysis

- Statistical analysis
  - t-test, regression, ANOVA
  - multiple testing correction (e.g., Bonferroni, false discovery rate)
- Machine learning
  - dimension reduction (principal component analysis), hierarchical clustering
  - logistic regression, random forests, support vector machines, neural networks, deep learning
  - cross-validation



MetaboAnalyst 5.0

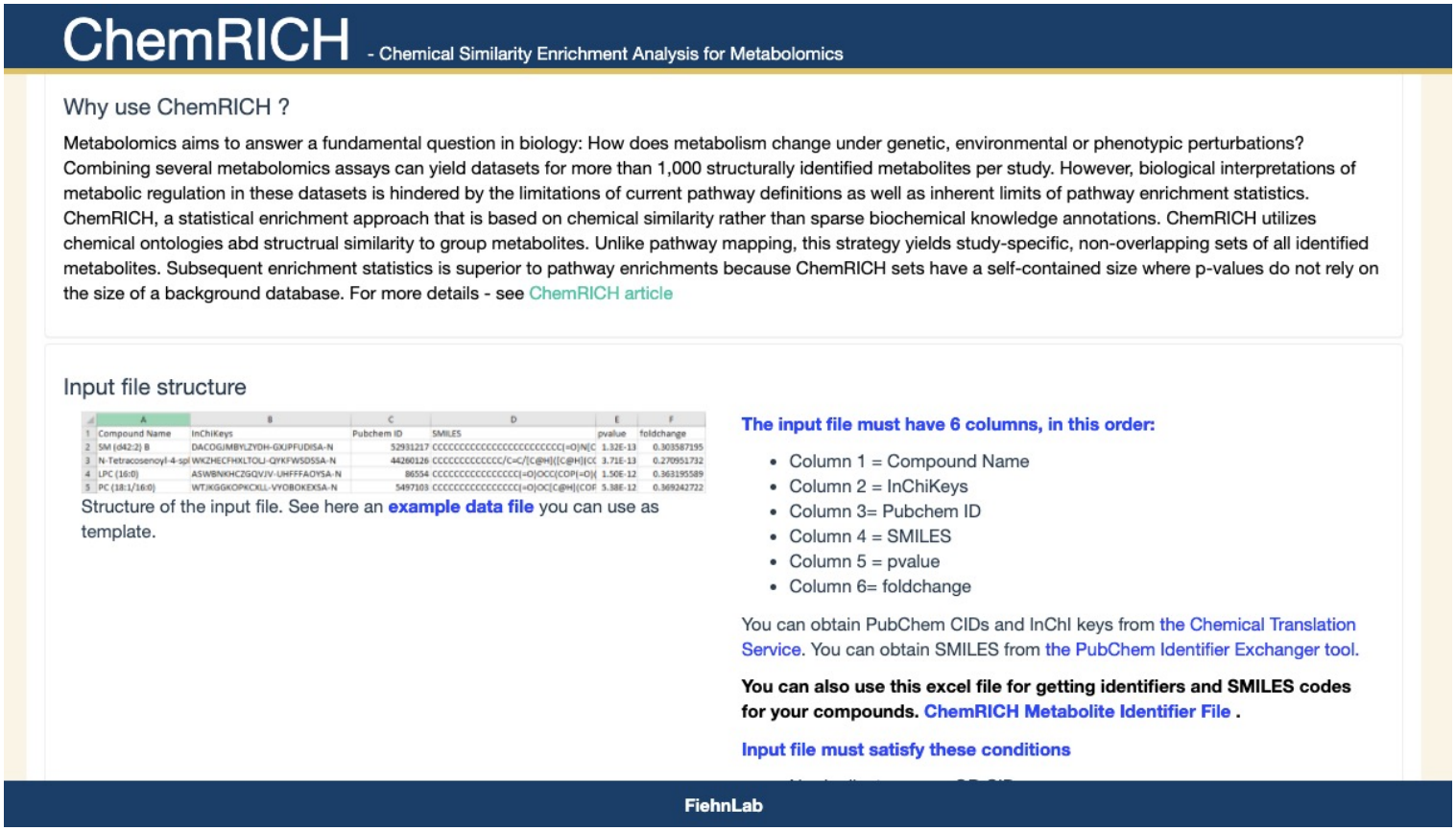
user-friendly, streamlined metabolomics data analysis



Pathway Databases

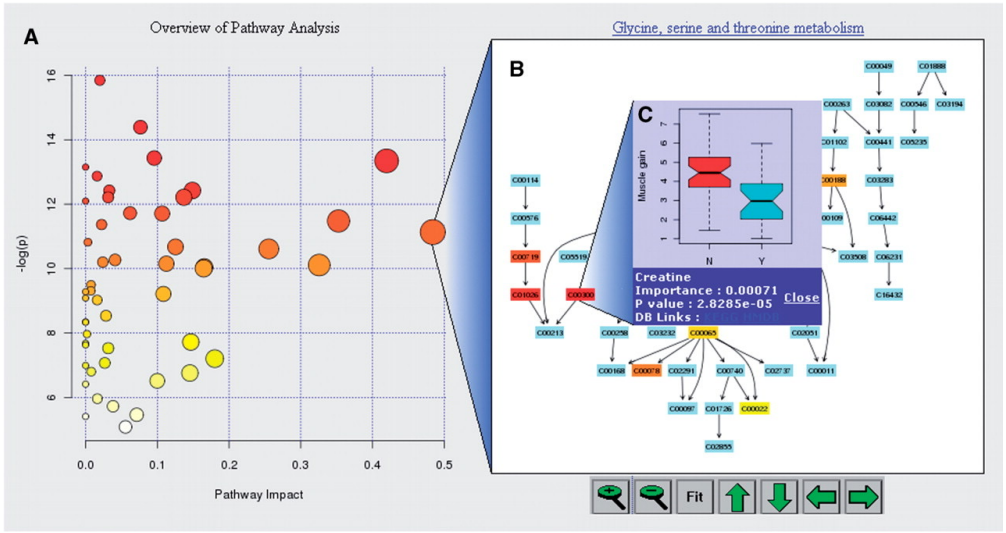
- **KEGG** - Kyoto Encyclopedia of Genes and Genomes
  - http://www.genome.jp/kegg/
  - ~530 pathways
- **MetaCyc**
  - https://MetaCyc.org
  - ~2600 pathways
- **Reactome**
  - https://reactome.org/
  - ~20,000 pathways

Enrichment Analysis



Pathway Analysis

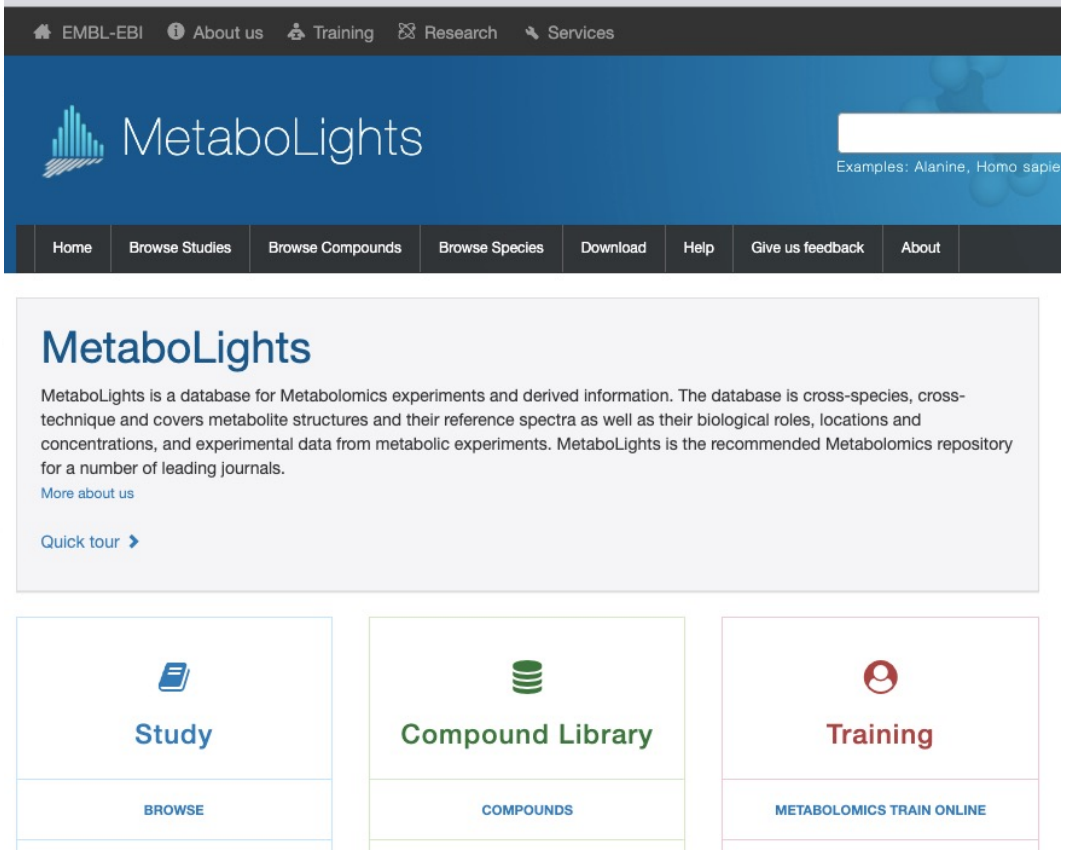
metPA in MetaboAnalyst



Pathway Impact:
Examines compounds' position in pathway, centrality/hub

https://www.metabolomicsworkbench.org/

https://www.ebi.ac.uk/metabolights/



https://mscat.metabolomicsworkbench.org/

MSCAT: Metabolomics Software CATalog

Metabolomics Software Database | Database Metrics | Workflow generator

Tool Name	Website	Publication	Last Updated	Operating System	User Interface	Molecule Type
filter data...						
CluMSID	Link	Link	2020-10-16	Windows, Mac, Linux	CLI	metabolomics
GNPS-MassIVE	Link	Link	2020-10-12	Windows, Mac, Linux	web	metabolomics
mineXpert	Link	Link	2020-10-12	Windows, Mac, Linux	GUI	metabolomics, protec
nPYc-Toolbox	Link	Link	2020-10-07	Mac, Linux, Windows	CLI	metabolomics
El-MAVEN	Link	Link	2020-09-28	Windows, Mac	GUI	metabolomics
VSClust	Link	Link	2020-08-12	Windows, Mac, Linux	web	metabolomics, genon
mixOmics	Link	Link	2020-08-09			genomics, transcriptc

“Metabolomics: the Superglue of Omics”

Giera, Spilker, Siuzdak (2018)