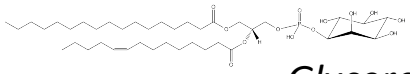
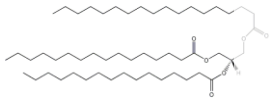
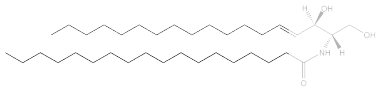


Glycerophospholipids



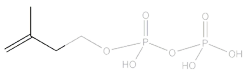
Glycerolipids

Fatty acyls

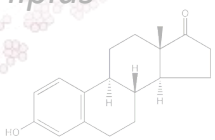


Sphingolipids

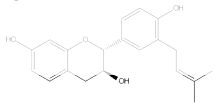
Prenol lipids



Sterol lipids



Polyketides



Computational Tools for Metabolomics

Scott Walmsley, Ph.D.

Mass Spectrometry Facility

DOPS-Skaggs SOP

UC Anschutz

What is metabolomics?

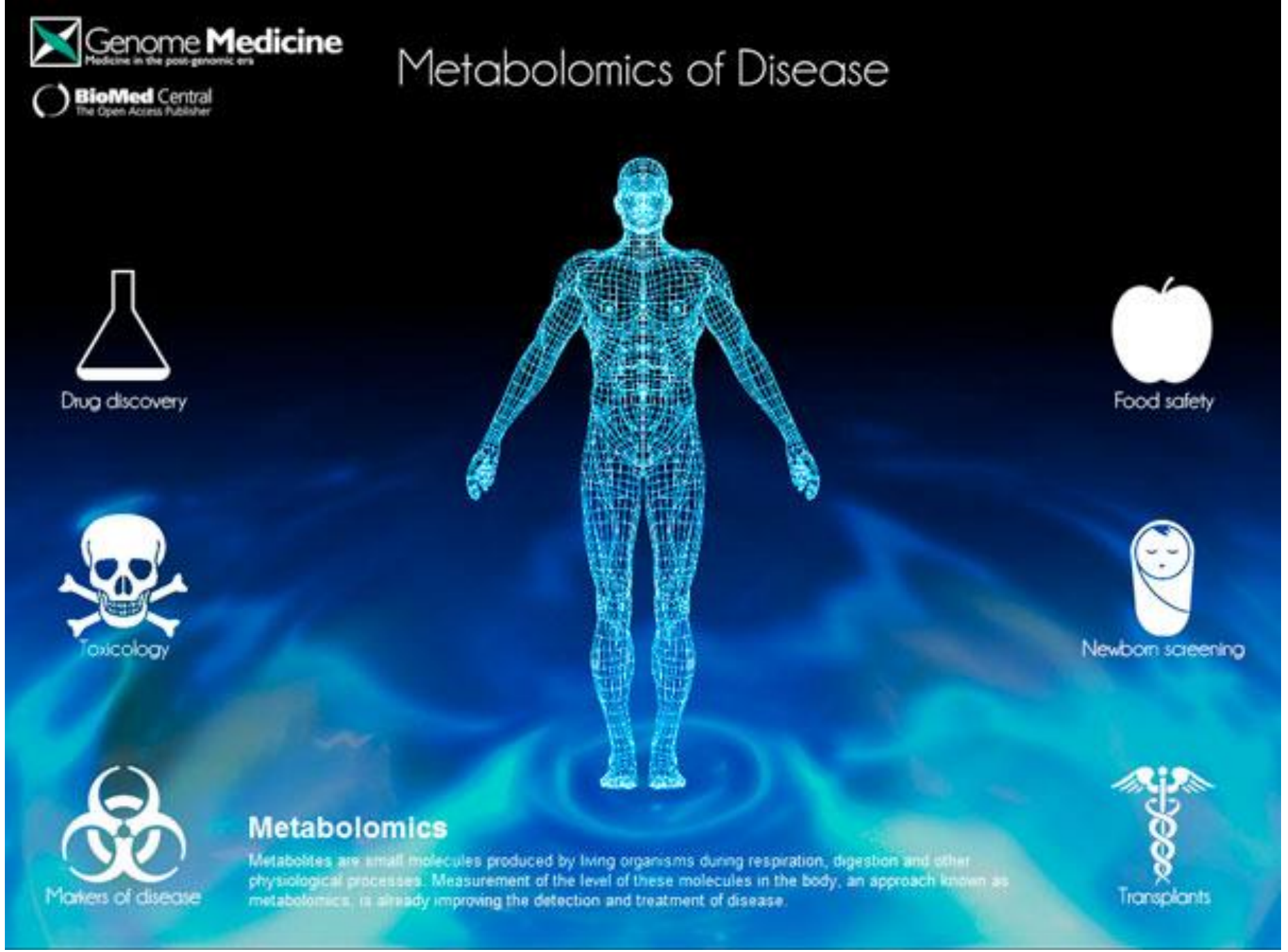
- "Systemic study of the unique chemical fingerprints that specific cellular processes leave behind" *Bennett Daviss, The Scientist 19(8) April 2005*

- But…is not limited to cellular processes…

  toxins, drugs, endogenous versus non endogenous compounds

- …..and has great utility for……

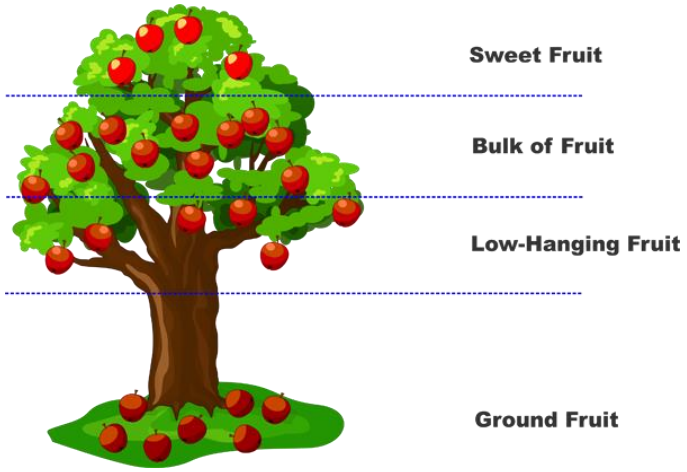
  discovery research, translational science, clinical profiling, personalized medicine


Genome Medicine
Medicine in the post-genomic era
BioMed Central
The Open Access Publisher
Metabolomics of Disease
Drug discovery
Food safety
Toxicology
Newborn screening
Markers of disease
Transplants
Metabolomics
Metabolites are small molecules produced by living organisms during respiration, digestion and other physiological processes. Measurement of the level of these molecules in the body, an approach known as metabolomics, is already improving the detection and treatment of disease.

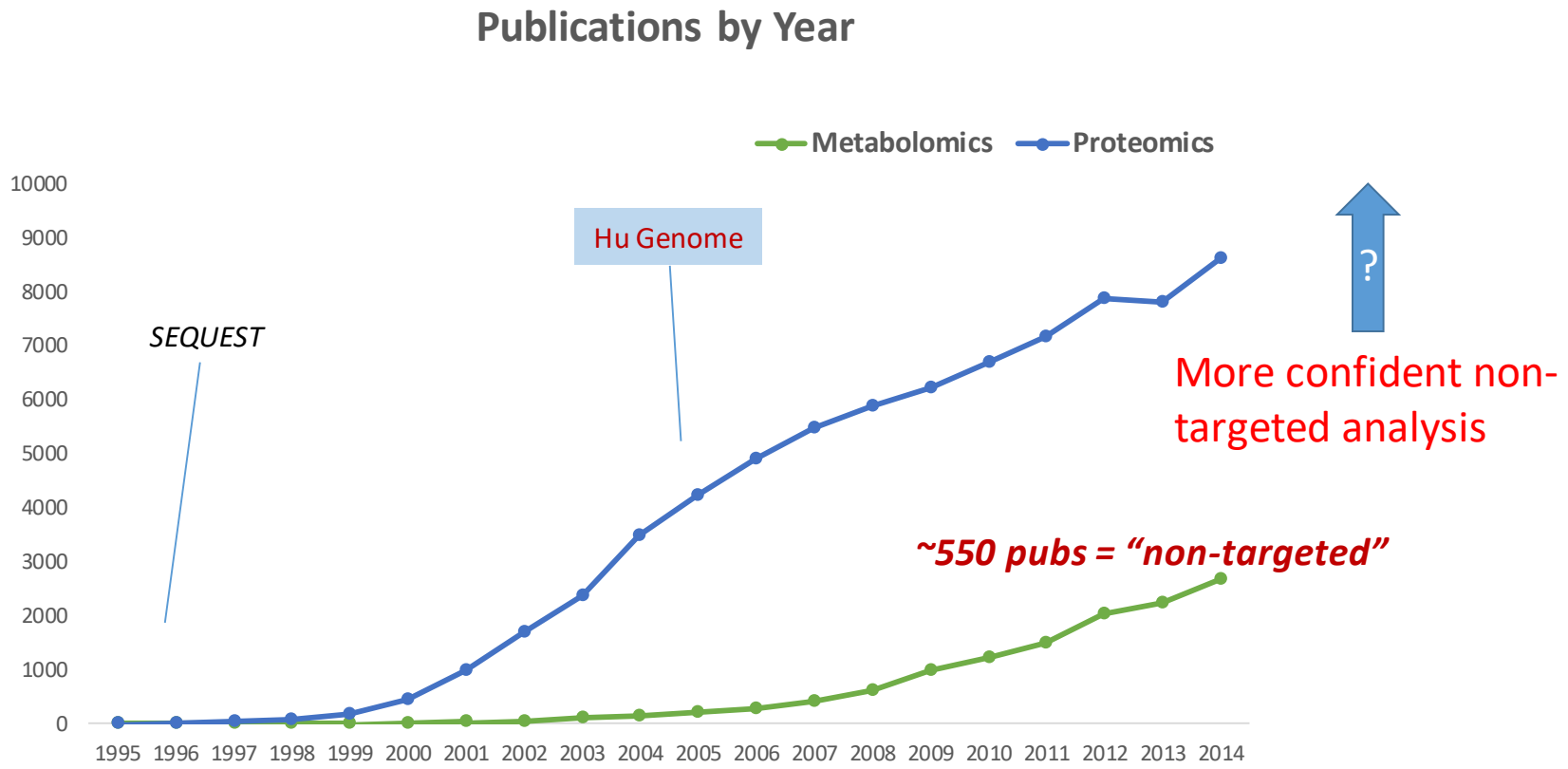
Relevance

- Rapid and accurate results from large sample sets
- Improve the knowledge of what's in the sample

- Produce a set of tools available to the discipline which aid the researcher's analysis of metabolomics data



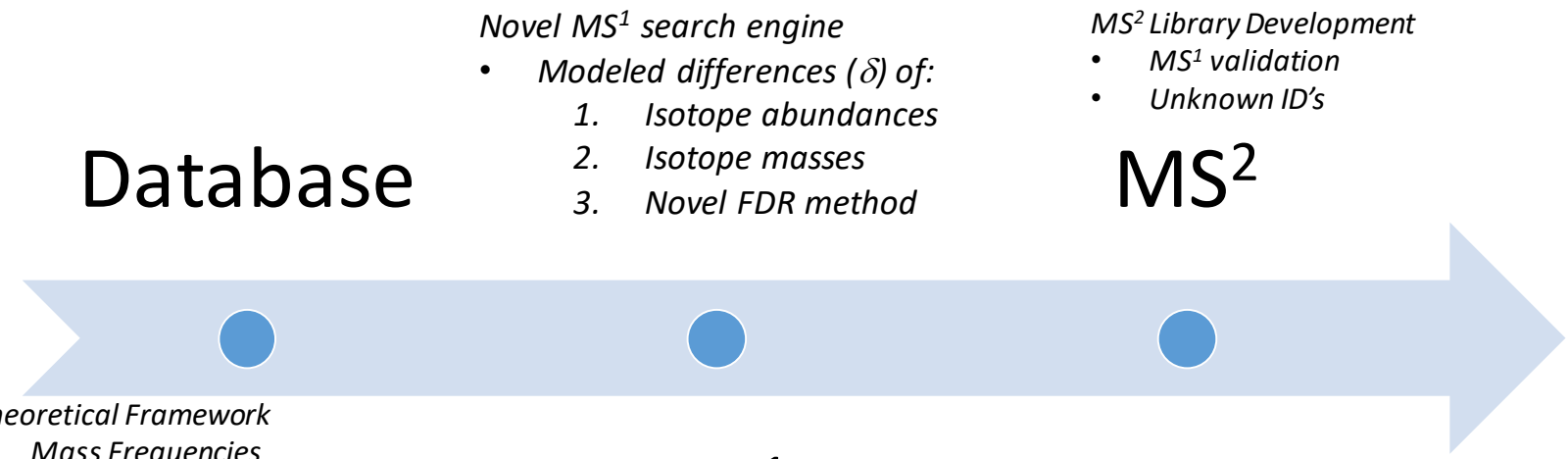
Proteomics vs metabolomics key publications and dates



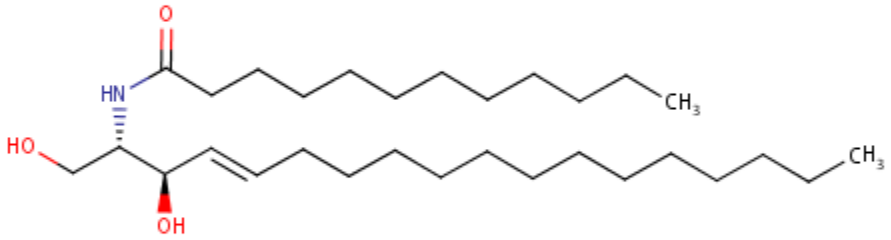
Metabolomics

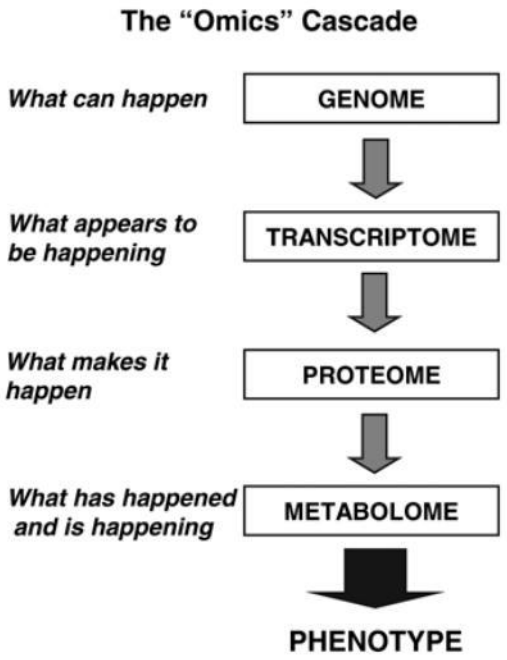
Challenges:

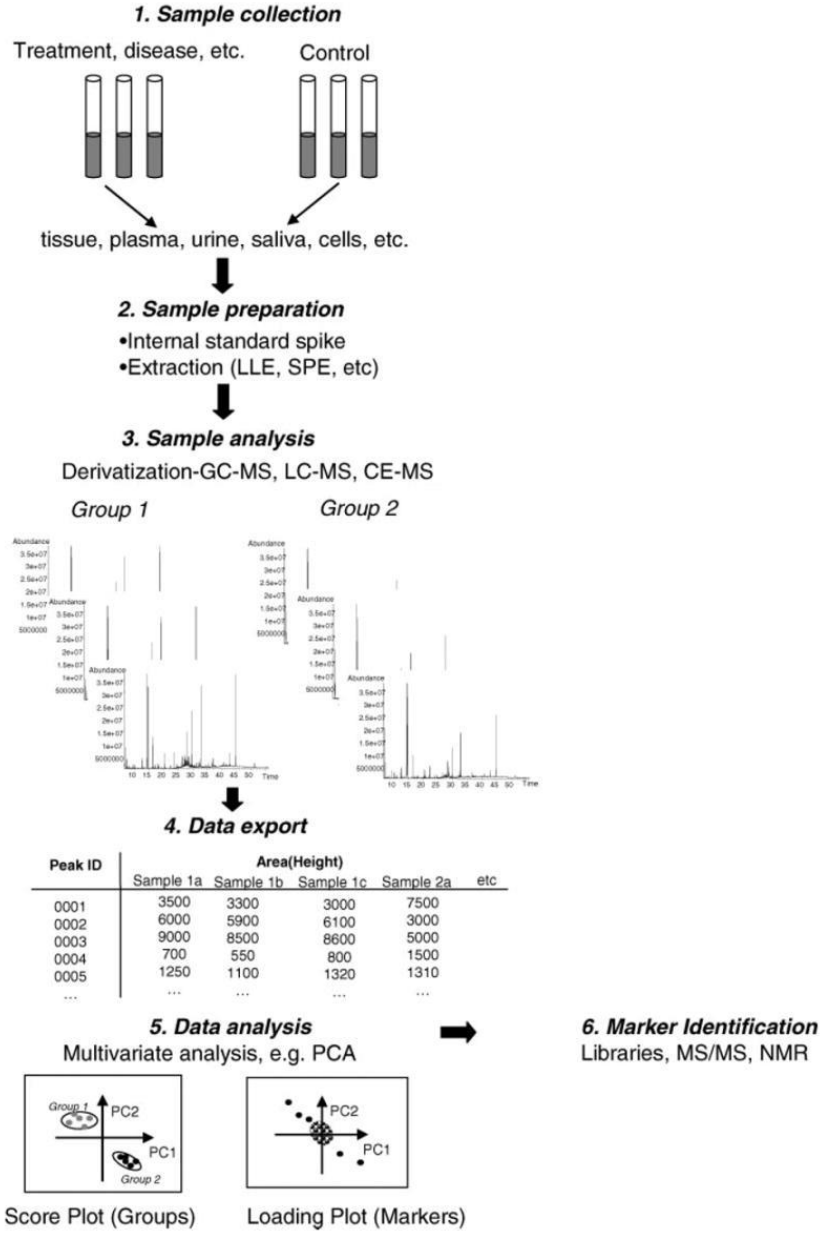
- Learning from proteomics
- Development of tools to enable rapid and accurate outcomes



Mass Spec and Metabolomics



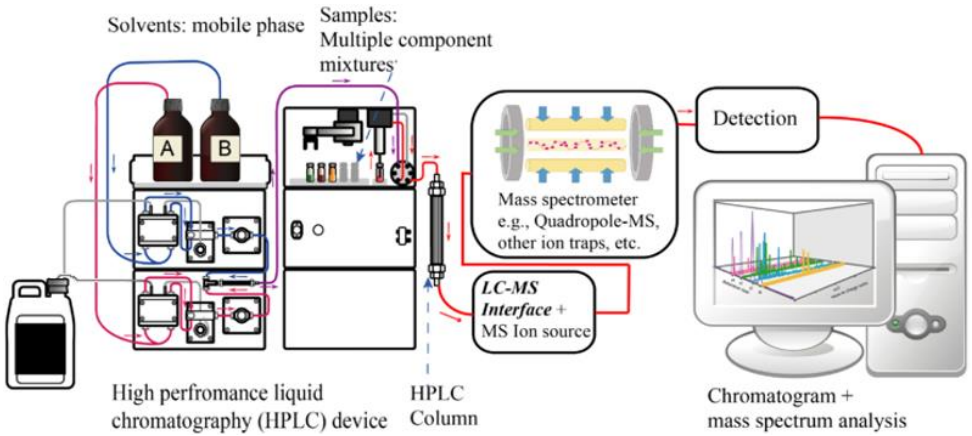


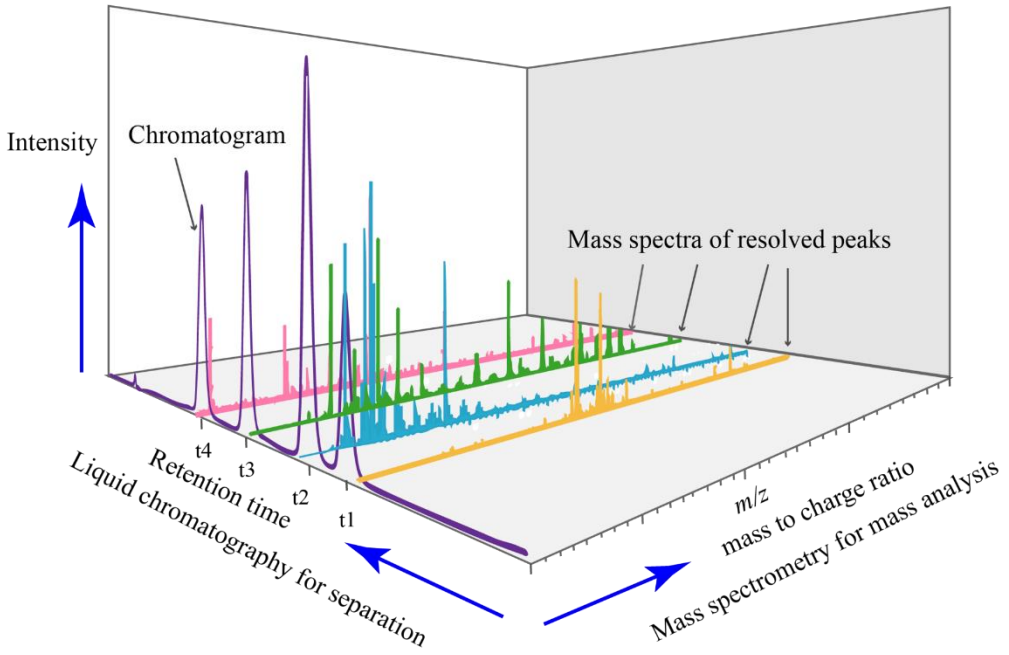


Dettmer K, Aronov PA, Hammock BD. Mass spectrometry-based metabolomics. Mass Spectrom Rev. 2007 Jan-Feb;26(1):51-78. Review.

Instrumentation

- HPLC
  - Separate molecules
  - Somewhat reliable
- Mass spectrometer
  - Desolvate molecules
  - Ionization
  - Resolve Ions my mass



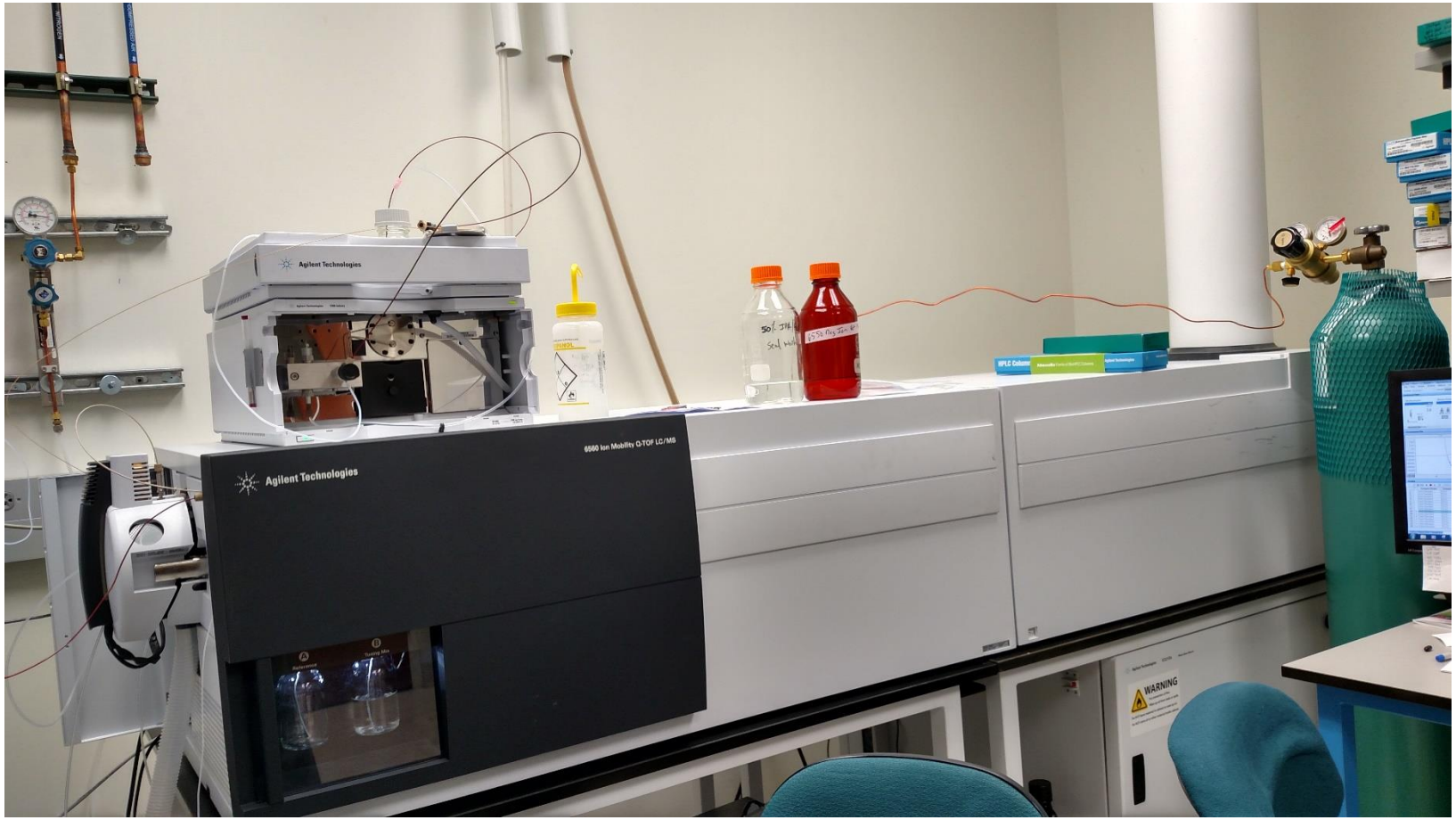


https://en.wikipedia.org/wiki/Liquid_chromatography%E2%80%93mass_spectrometry#/media/File:Liquid_chromatography_MS_spectrum_3D_analysis.png

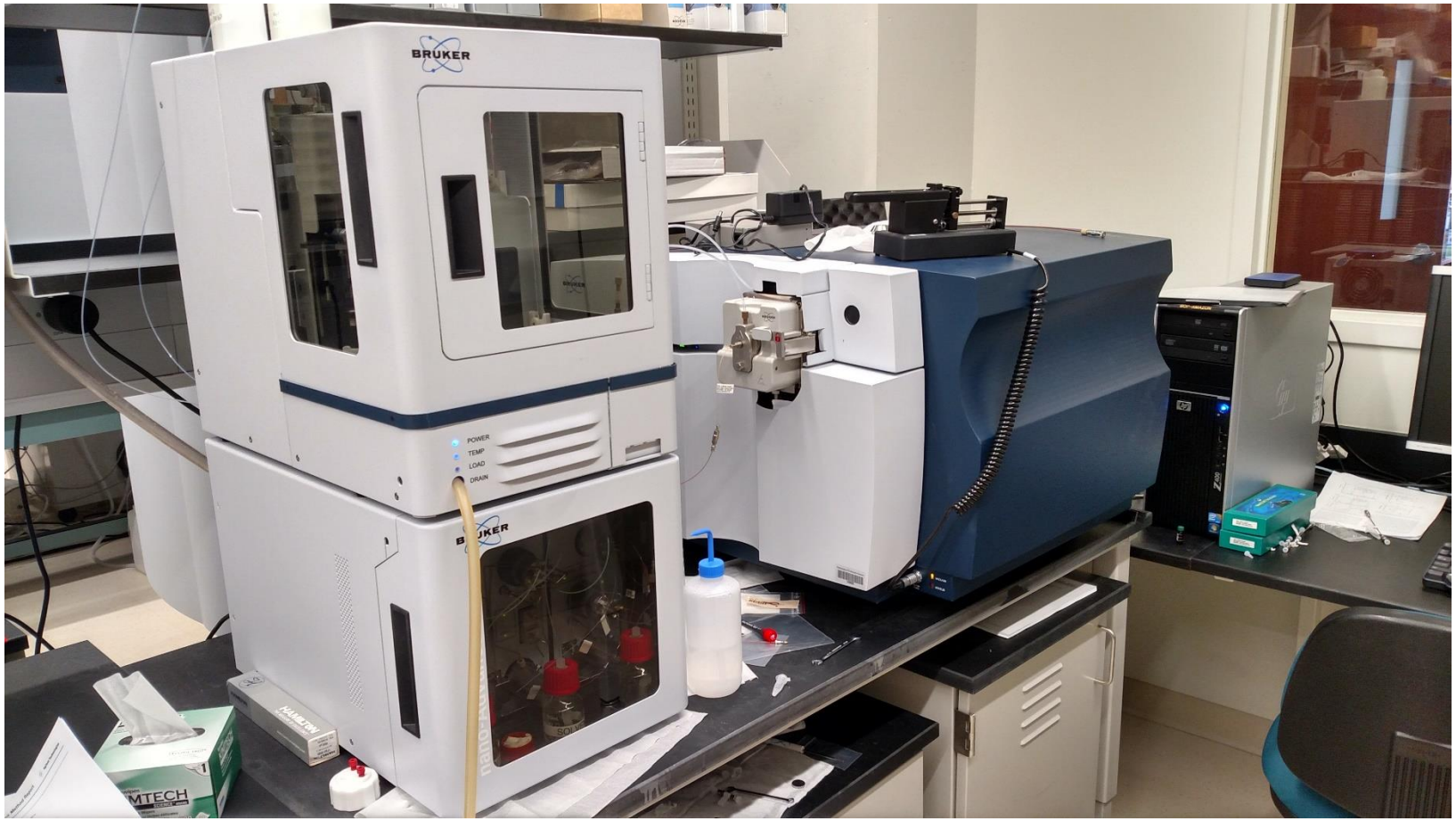
Mass Spectrometers (Mostly Untargeted Data)

- Quadrupole Time of Flight
- Ion Trap
- Ion Mobility MS
  - Isobars
- Triple- Quadrupole (Targeted MS)
- Gas Chromatography-MS

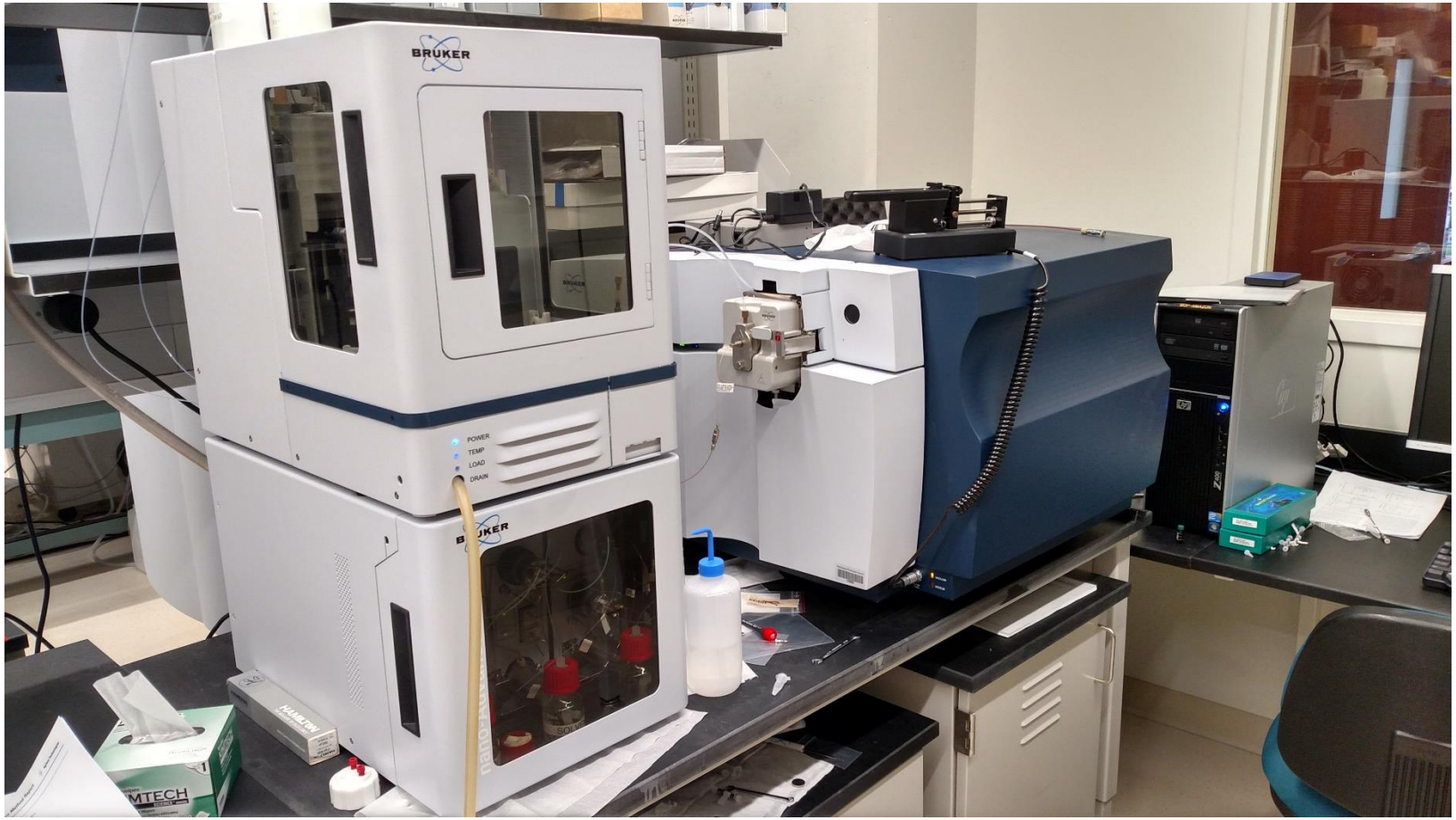
Ion Mobility-QTOF



Ion Trap



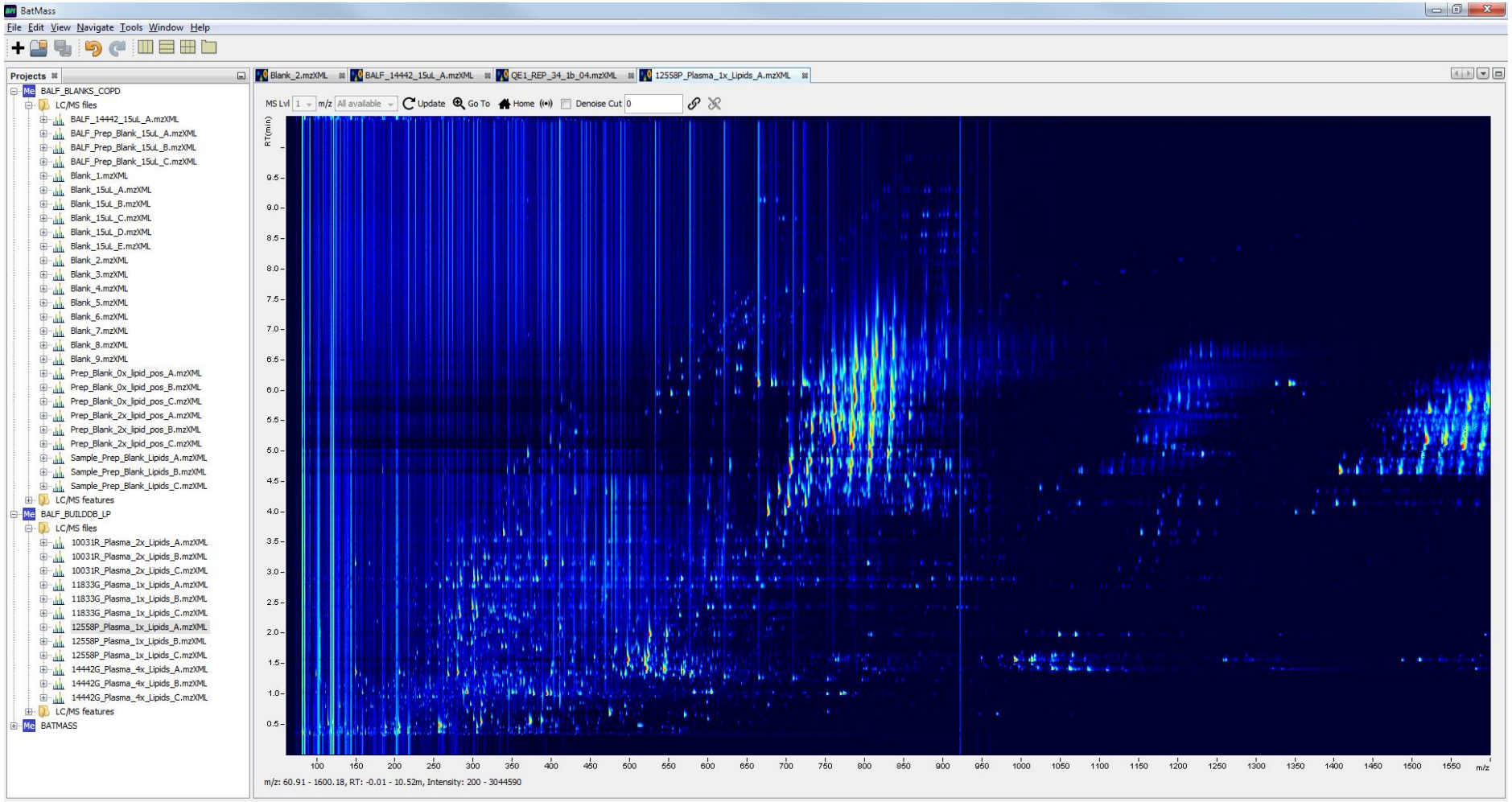
GC-MS



Piece-meal Data center

Mass Spectrometry

- Many forms of instrument
- All produce similar forms of data.



Background

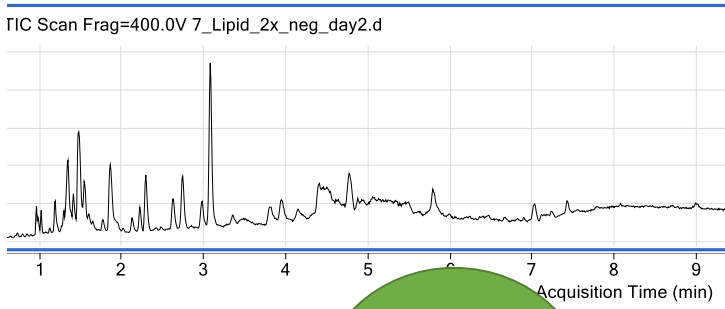
Where metabolomics data comes from

Cells, Biofluids

Sample → Extract

Lipid, Aqueous Fractions

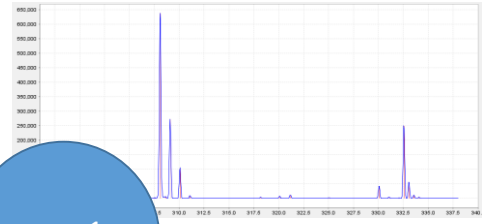
QTOF Mass Spectrometer



Raw Data

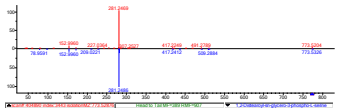
Mass and Abundance Signatures of Compounds

Raw Data → Chromatogram



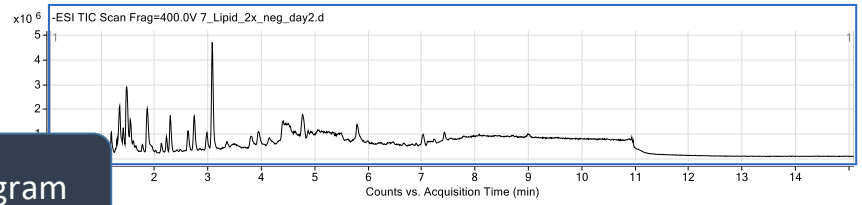
Raw Data → MS^1

Raw Data → MS^2



Background

Where metabolomics data comes from



Raw Data

Chromatogram

MS^1

Targeted

Non-Targeted

Pursue a few masses from the sample

Pursue as many masses from the sample

Attempt to capture as much information from the sample

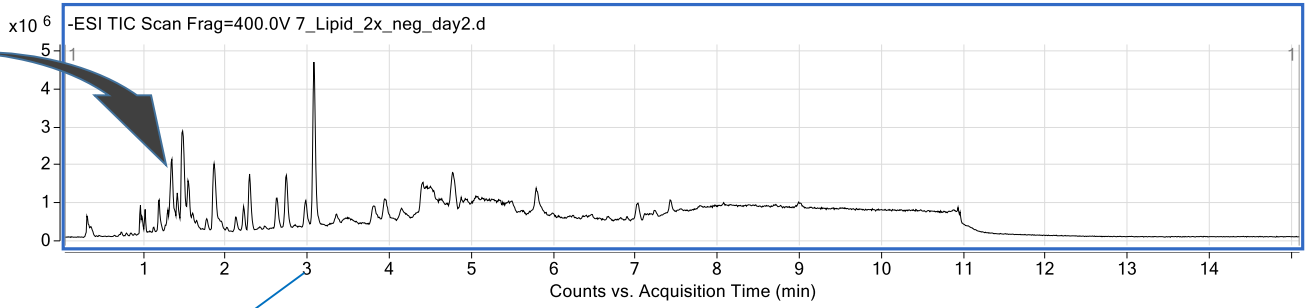
Background

Where metabolomics data comes from.

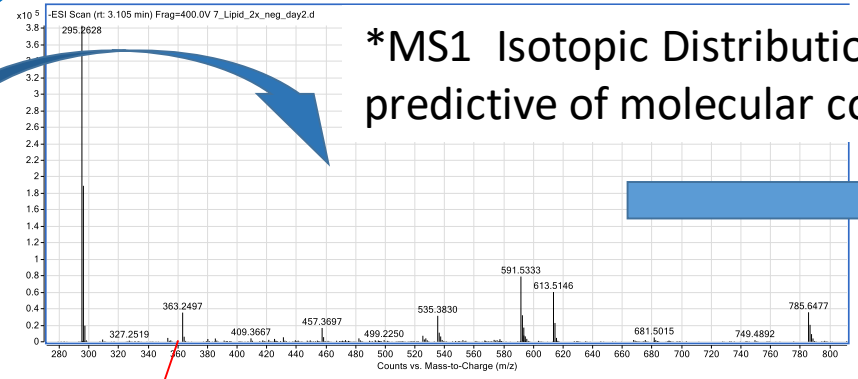
Raw Data

HPLC

Chromatogram



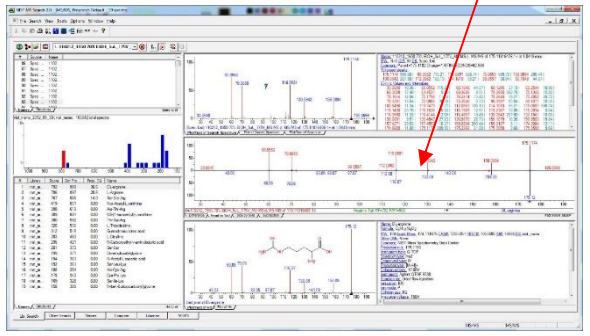
MS^1



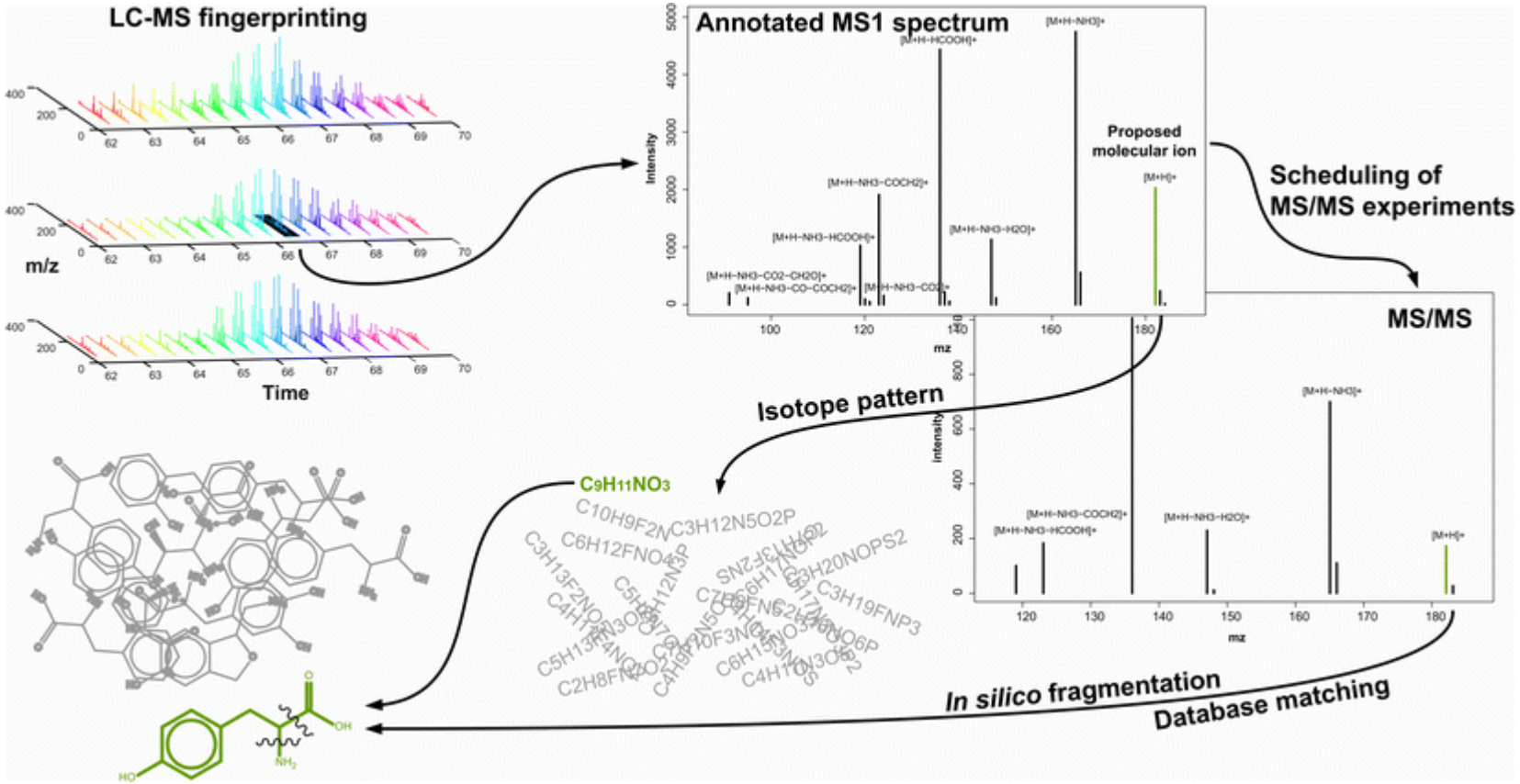
*MS1 Isotopic Distributions are highly predictive of molecular composition

C6H12O6

MS^2



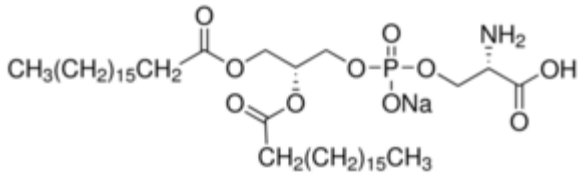
*MS2 Fragmentation patterns are highly predictive of molecular structure

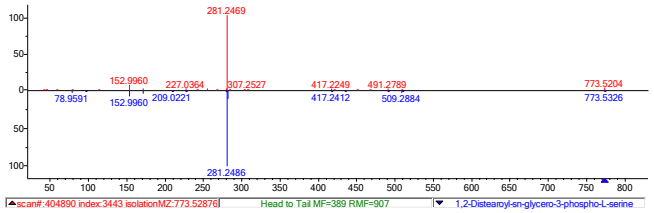


1: Stanstrup J, Gerlich M, Dragsted LO, Neumann S. Metabolite profiling and beyond: approaches for the rapid processing and a nnotation of human blood serum mass spectrometry data. Anal Bioanal Chem. 2013 Jun;405(15):5037-48. doi: 10.1007/s00216-013-6954-6. Epub 2013 Apr 25. PubMed PMID: 23615935.

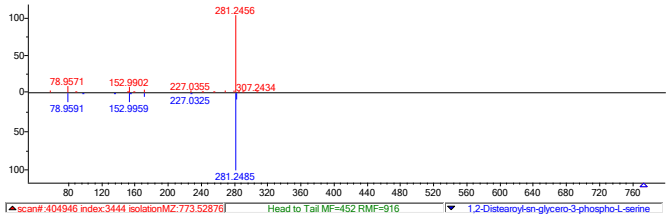
MS^2 Small Molecules

PS(18:0/18:0)
1,2-Distearoyl-sn-glycero-3-phospho-L-serine
$C_{42}H_{82}NO_{10}P$

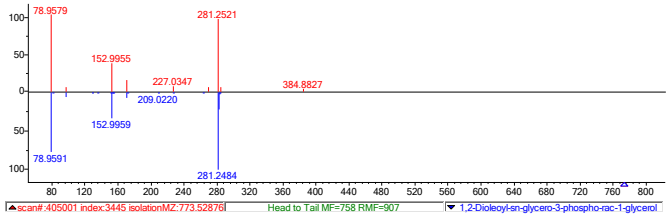




40V [M-H-]- : 773.52876



60V [M-H-NH3]- : 791.567635

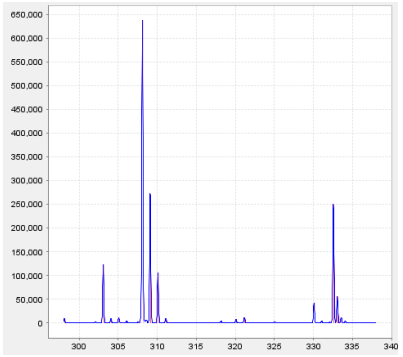


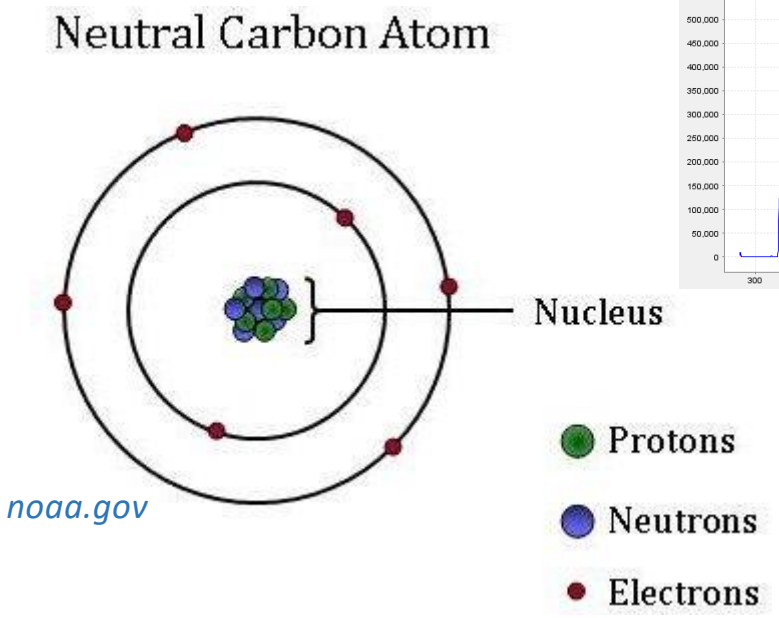
80V [M-H-]- : 773.52876

MS1 Isotope pattern recognition

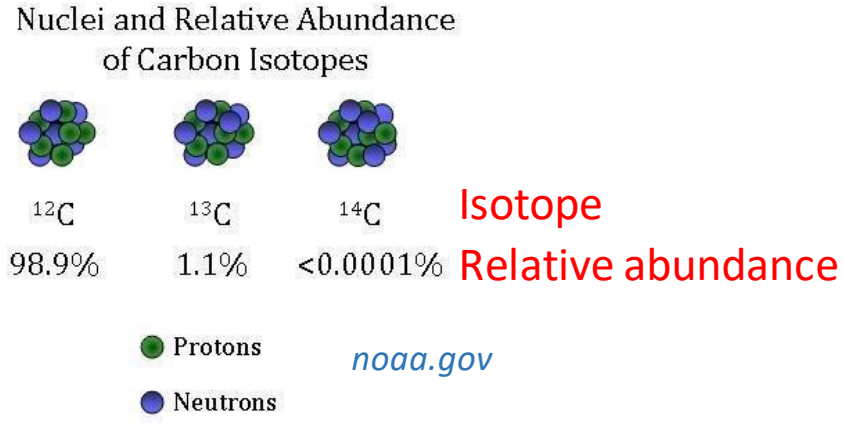
- Isotope ratios are used to compute relative isotopic distributions and molecular formulae
- Molecular formulae are used to say what compound the molecule likely is
- Sample specific databases are used to reduce the ambiguity of the compound lists.

Stable Isotopes





noaa.gov



Isotope

Relative abundance

noaa.gov

Typical Metabolites Contain:

Element	Isotope	mass	Mass difference	Abundance (%)
Hydrogen	^{1}H	1.007825		99.985
	^{2}H	2.014102	+1.006277	0.015
Carbon	^{12}C	12.0		98.890
	^{13}C	13.003355	+1.003355	1.110
Nitrogen	^{14}N	14.003074		99.634
	^{15}N	15.000109	+0.997035	0.366
Oxygen	^{16}O	15.994915		99.762
	^{17}O	16.999132	+1.004217	0.038
	^{18}O	17.999161	+2.004246	0.200
Phosphor	^{31}P	30.973762		100
Sulfur	^{32}S	31.972071		95.020
	^{33}S	32.971459	+0.999388	0.750
	^{34}S	33.967867	+1.995796	4.210
	^{36}S	35.967081	+3.995010	0.02

Tools can predict the exact center masses of each isotope

Existing Tools: Isotopic abundance prediction and matching

Prediction: use to build theoretical isotopic spectra

- BRAIN
- EMASS

Matching : use to generate formulae based on closest mass and isotopes (isotopes and or monoisotopic mass only)

- SIRIUS
- RDISop

NO estimate of error of match!

NO Uniform score for match!

Computationally intensive dependent on mass.

MS1: Compound – Formula matching

- Computationally intensive
- No constrained search space
- No estimate of quality of sample

Existing Tools

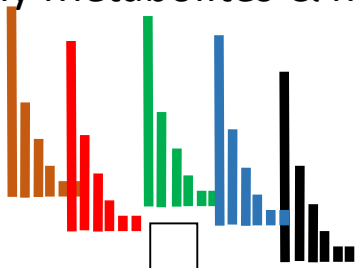
*** NO limited database**

Single Metabolite & MS1 Spectrum

Single Formula Spectrum Match (FSM)

Proposed Algorithm

Many Metabolites & MS1 Spectra

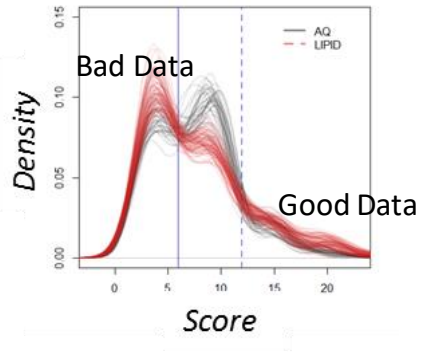


*** Well defined database**

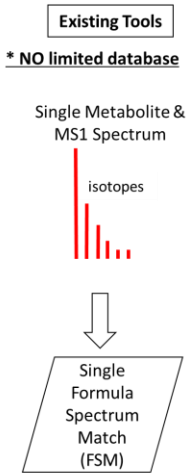
Lists of all FSMs and Scores

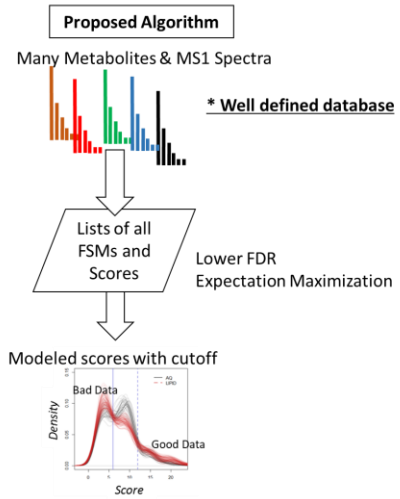
Lower FDR
Expectation Maximization

Modeled scores with cutoff



Novel MS1 Search Engine





No limit on elemental composition. High FPR.

Limited by database size. Low-Medium FPR.

Huge potential for errors within limits of tolerance for mass and isotope intensities.

Tunable results by global modelling sensing instrument error and by repeated measures.

Figure 1.

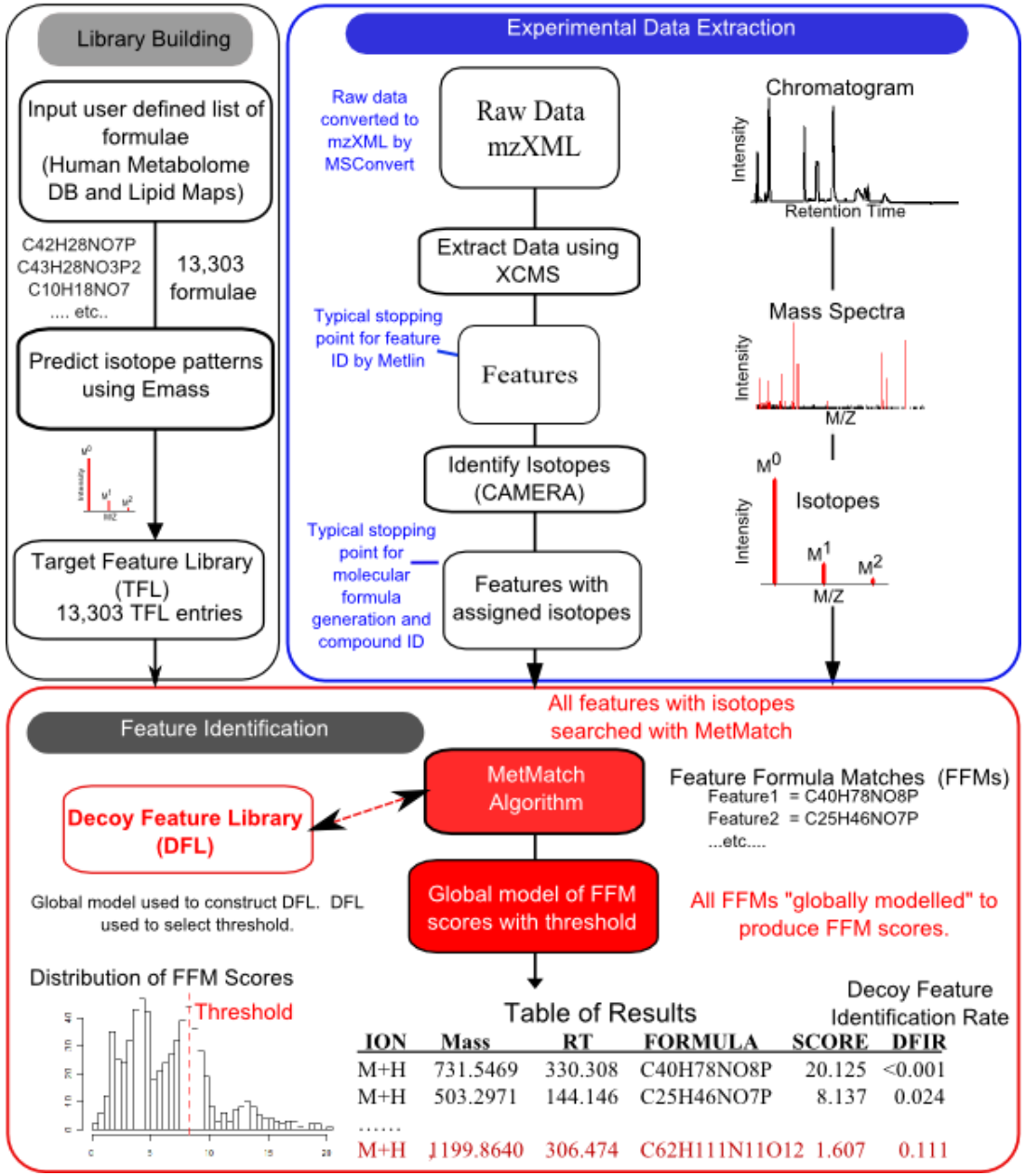
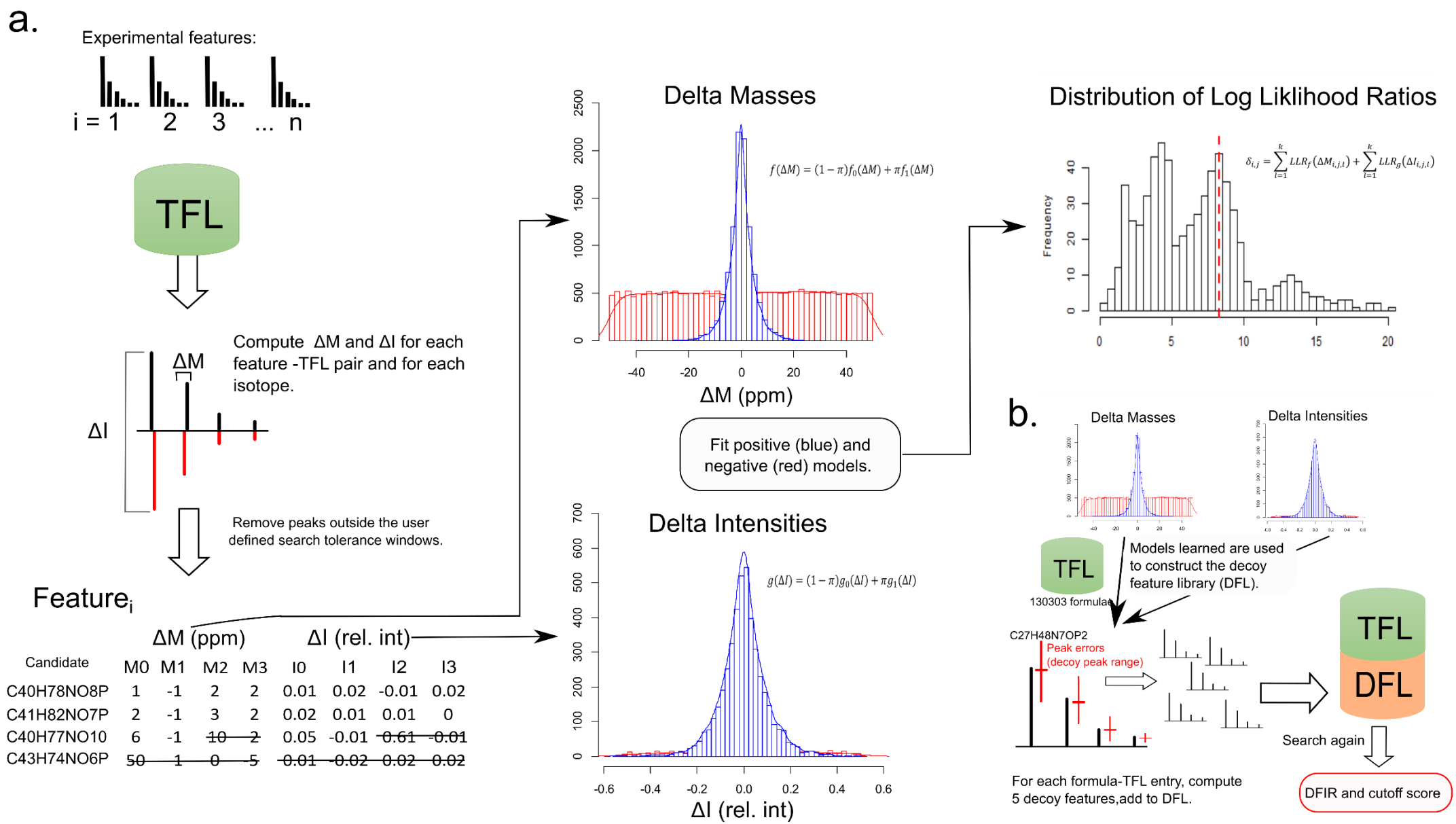


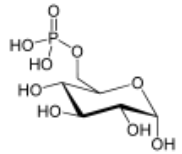
Figure 2.


a.
Experimental features:
i = 1 2 3 ... n
TFL
ΔM
ΔI
Compute ΔM and ΔI for each feature -TFL pair and for each isotope.
Remove peaks outside the user defined search tolerance windows.
Feature$_i$
ΔM (ppm) ΔI (rel. int)
Candidate M0 M1 M2 M3 I0 I1 I2 I3
C40H78NO8P 1 -1 2 2 0.01 0.02 -0.01 0.02
C41H82NO7P 2 -1 3 2 0.02 0.01 0.01 0
C40H77NO10 6 -1 10 2 0.05 -0.01 0.61 -0.01
C43H74NO6P 50 1 0 5 0.01 -0.02 0.02 0.02
Delta Masses
$f(\Delta M) = (1-\pi)f_0(\Delta M) + \pi f_1(\Delta M)$
ΔM (ppm)
Fit positive (blue) and negative (red) models.
Delta Intensities
$g(\Delta I) = (1-\pi)g_0(\Delta I) + \pi g_1(\Delta I)$
ΔI (rel. int)
Distribution of Log Liklihood Ratios
$\delta_{i,j} = \sum_{l=1}^{k} LLR_f(\Delta M_{i,j,l}) + \sum_{l=1}^{k} LLR_g(\Delta I_{i,j,l})$
Frequency
b.
Delta Masses
Delta Intensities
TFL
130303 formulas
Models learned are used to construct the decoy feature library (DFL).
C27H48N7OP2
Peak errors (decoy peak range)
TFL
DFL
Search again
For each formula-TFL entry, compute 5 decoy features,add to DFL.
DFIR and cutoff score

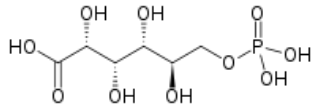
Database Problem: creating sample specific databases can help.

Intermediate #	Molecule	Exact Mass	Molecular Formula	# Path Cmpd	n mass	# Path Mass	P(m)	P(m?)	P(I)	p(cmpd) = P(m)*P(I)	p(path)
entry	Glucose 6-phosphate	260.029719	C6H13O9P	4	19	7	0.053	0.008	0.500	0.026	0.004
1	6-Phosphocluconolactone	258.014068	C6H11O9P	1	2	1	0.500	0.500	1.000	***0.500***	***0.500***
2	6-Phosphogluconic acid	276.024633	C6H13O10P	1	2	1	0.500	0.500	0.500	0.250	0.250
3	D-Ribulose 5-phosphate	230.019154	C5H11O8P	1	7	2	0.143	0.071	1.000	0.143	**0.071**
4	Ribose 5-phosphate	230.019154	C5H11O8P	2	7	2	0.143	0.071	1.000	0.143	**0.071**
5	Xylulose-5-phosphate	230.019154	C5H11O8P	1	7	2	0.143	0.071	1.000	0.143	**0.071**
6	D-Sedoheptulose 7-phosphate	290.040283	C7H15O10P	1	2	1	0.500	0.500	1.000	***0.500***	***0.500***
7	D-Erythrose 4-phosphate	200.008589	C4H9O7P	1	1	1	1.000	1.000	1.000	***1.000***	***1.000***

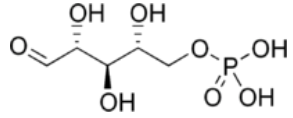
0.



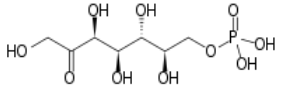
2.



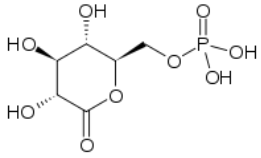
4.



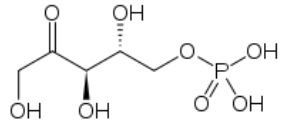
6.



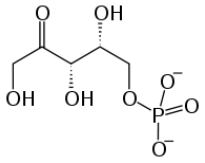
1.



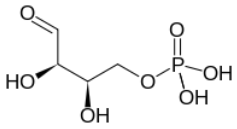
3.



5.



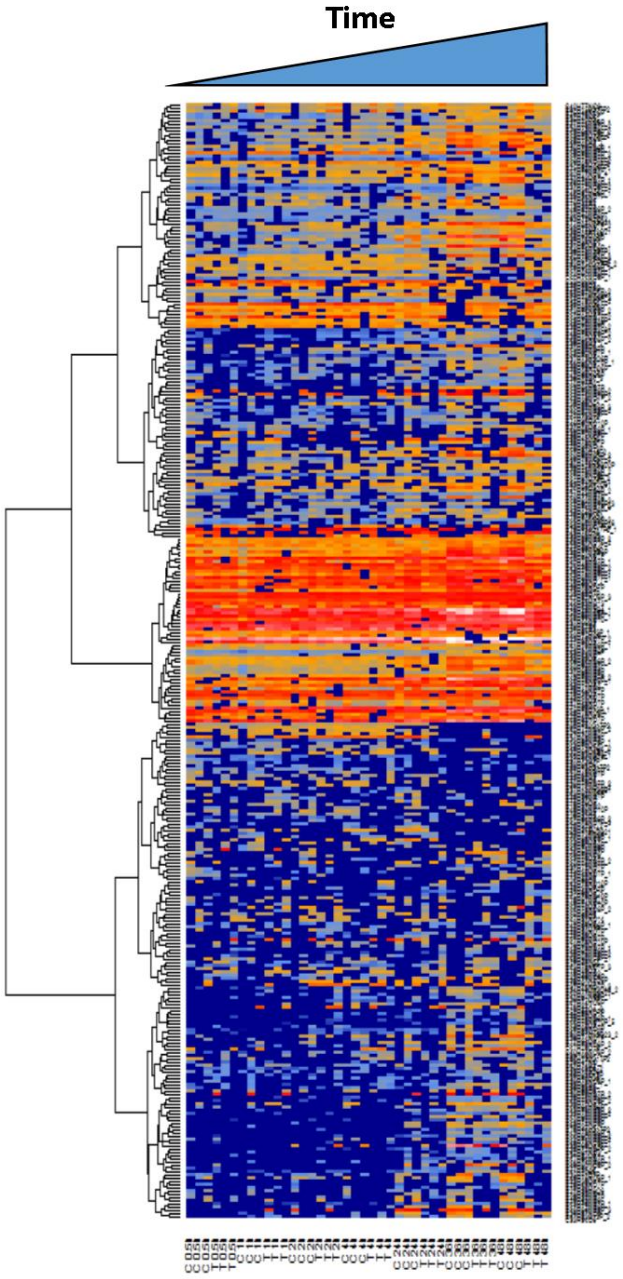
7.

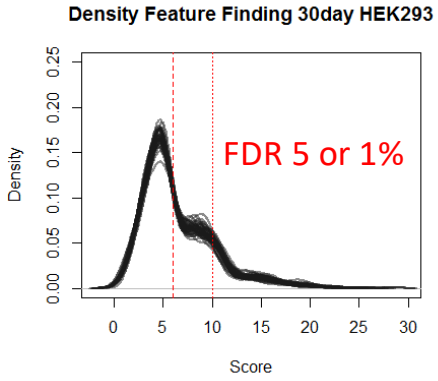


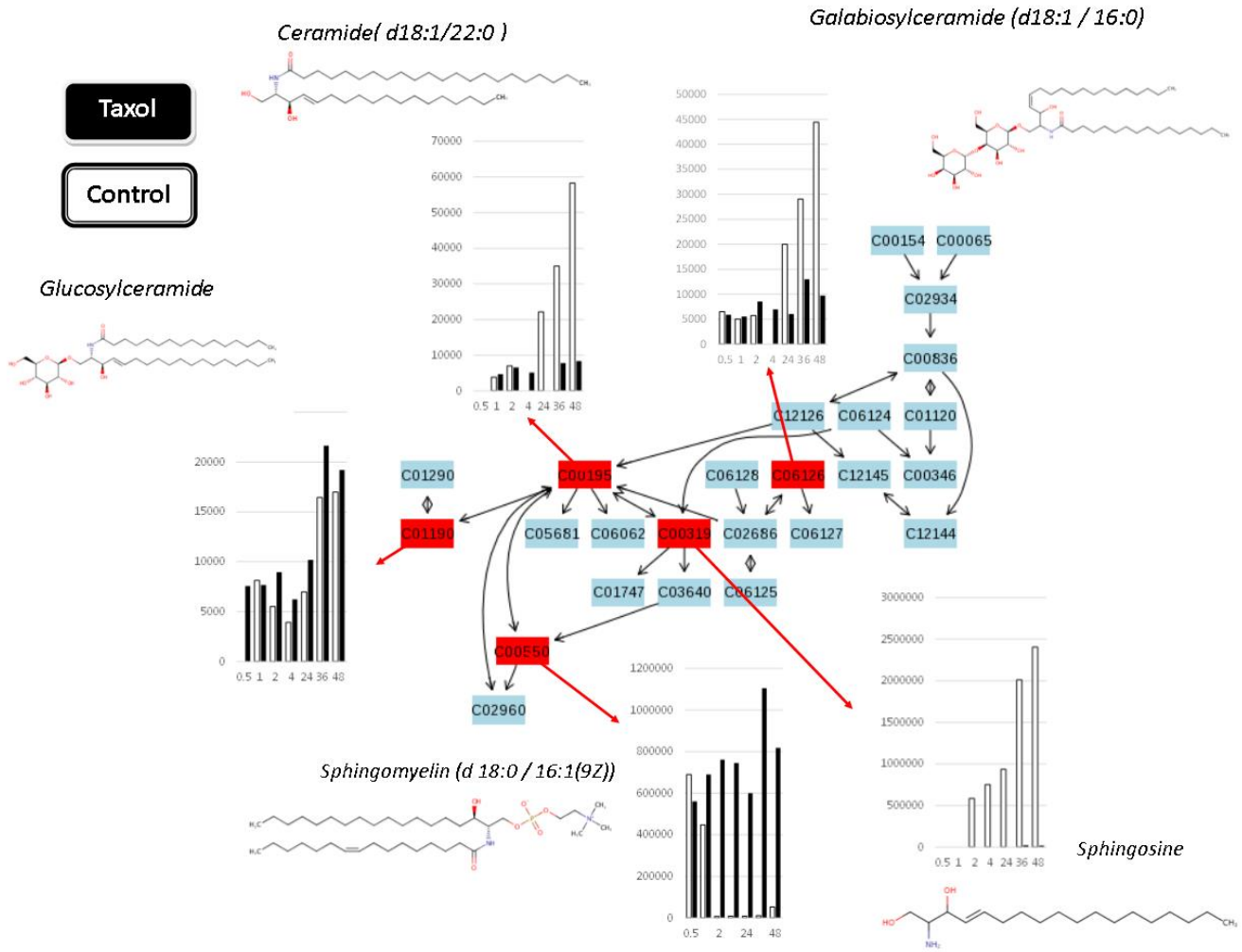
Application: Biological Response, Toxicity

- HEK293 Cells : Lipid Fraction
- LCMS (QTOF)
- Time series, n=3 * 7 time points, Taxol Responsive

Biological Response, Taxol Responsive ->Match expected biological outcome

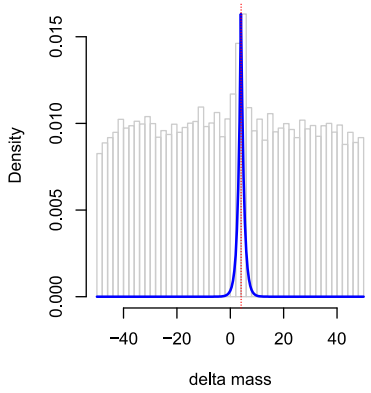


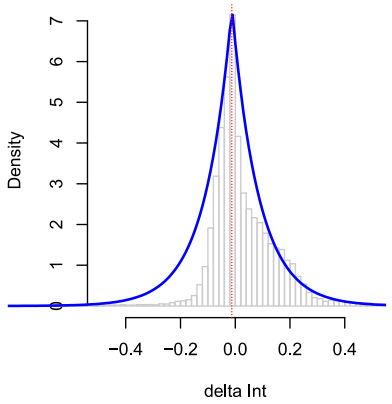




Sphingolipid Metabolism

Pluskal Data Method Compared to Metabmatch: Rank order Hits





	Annotated Formula (MS2)	Type of ion	m/z	Retention time (min)	Rank (Pluskal)	Rank Metabmatch
Acetyl-CoA	$C_{23}H_{38}N_7O_{17}P_3S$	$[M+H]^+$	810.1331	10.7	96	6
GDP-glucose	$C_{16}H_{25}N_5O_{16}P_2$	$[M+H]^+$	606.0851	15.6	6	1
Glutathione (GSSG)	$C_{20}H_{32}N_6O_{12}S_2$	$[M+H]^+$	613.1601	15.2	6	1
ADP	$C_{10}H_{15}N_5O_{10}P_2$	$[M+H]^+$	428.0371	12.9	5	1
Coenzyme B	$C_{11}H_{22}NO_7PS$	$[M+H]^+$	344.0916	10.5	3	1
UMP	$C_9H_{13}N_2O_9P$	$[M+H]^+$	325.0435	12.8	3	1
Glutamyl-cysteine	$C_8H_{14}N_2O_5S$	$[M+H]^+$	251.0700	11.8	2	1
Adenosine	$C_{10}H_{13}N_5O_4$	$[M+H]^+$	268.1045	6.6	1	1
Arginine	$C_6H_{14}N_4O_2$	$[M+H]^+$	175.1191	25.3	1	1
Arginino-succinate	$C_{10}H_{18}N_4O_6$	$[M+H]^+$	291.1304	14.3	1	1
ATP	$C_{10}H_{16}N_5O_{13}P_3$	$[M+H]^+$	508.0033	14	1	1
Biotin	$C_{10}H_{16}N_2O_3S$	$[M+H]^+$	245.0957	6.2	1	1
CTP	$C_9H_{16}N_3O_{14}P_3$	$[M+H]^+$	483.9919	15.7	1	1

Global modelling tends corrections due to instrument performance (mass accuracy)

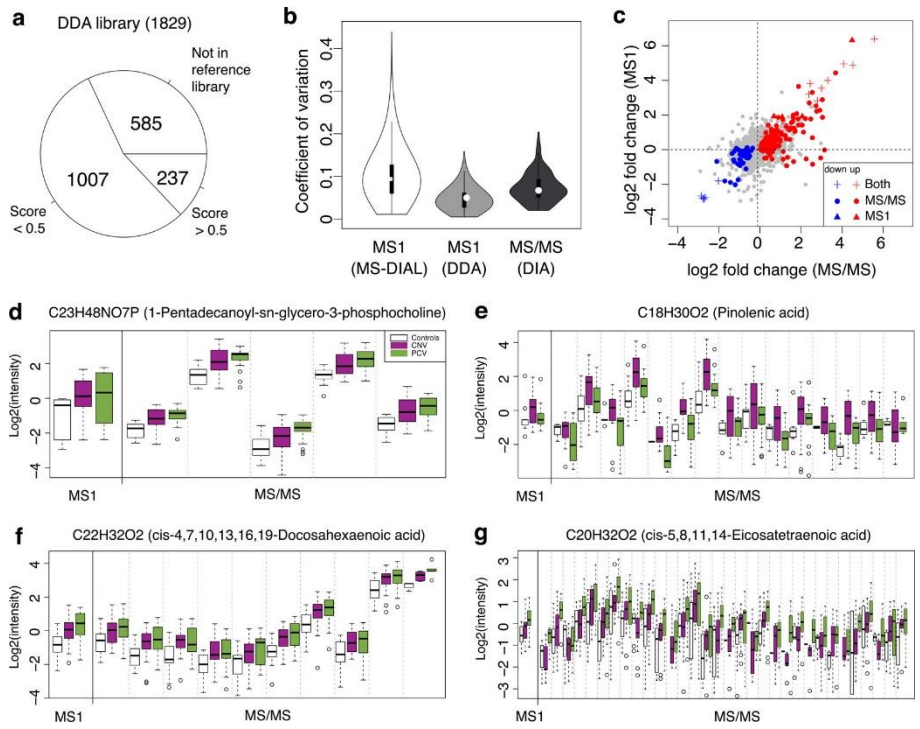
Conclusions: MS1 Global Formula matching

- Specific databases provide increased accuracy
- MS1 Search engine proves to be rapid
  - Proving useful in other technologies (SWATH MS)
  - MetaboDIA implementation (MS1 -> MS2 corrections

- Used in MetaboDIA

Chen G, Walmsley S, Cheung GCM, Chen L, Cheng CY, Beuerman RW, Wong TY, Zhou L, Choi H. ***Customized Consensus Spectral Library Building for Untargeted Quantitative Metabolomics Analysis with Data Independent Acquisition Mass Spectrometry and MetaboDIA Workflow***. Anal Chem. 2017 May 2;89(9):4897-4906. doi: 10.1021/acs.analchem.6b05006. Epub 2017 Apr 18. PubMed PMID: 28391692.

Application: Experiment specific Spectral Library Building



- MS2 fragmentation is considered better than using MS1 isotope ratios for identification.
- But MS2 fragmentation in metabolomics can be less reproducible.
- Collecting repeated measures is not common place.
- We collect repeated measured of MS2 spectra and build consensus spectra.
- We then use MS1 algorithm to correct MS2 spectra matches to a library when chemical structures are similar.
- It works because MS1 algorithm ‘tunes’ the resultant masses by globally modelling the mass errors.

Chen G, Walmsley S, Cheung GCM, Chen L, Cheng CY, Beuerman RW, Wong TY, Zhou L, Choi H. ***Customized Consensus Spectral Library Building for Untargeted Quantitative Metabolomics Analysis with Data Independent Acquisition Mass Spectrometry and MetaboDIA Workflow.*** Anal Chem. 2017 May 2;89(9):4897-4906. doi: 10.1021/acs.analchem.6b05006. Epub 2017 Apr 18. PubMed PMID: 28391692.

Small Molecule Databases (Sample specific databases)

- *A Prototypic Small Molecule Database for Bronchoalveolar Lavage-Based Metabolomics, Scientific Data, (Accepted Feb 8, 2018)*

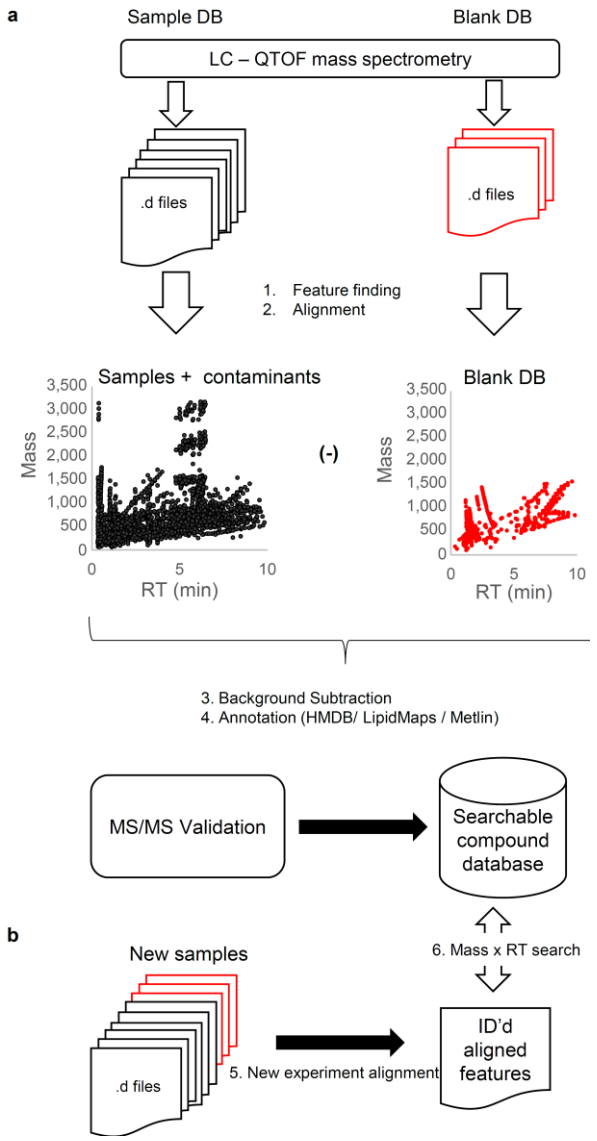
- *Compound library:*
  - *Informatics support system*
  - *MS1 database*
  - *MS2 linkages*
  - *HMDB , LipidMaps linkages*

A Prototypic Small Molecule Database for Bronchoalveolar Lavage-Based Metabolomics, Scientific Data, (Accepted Feb 8, 2018)

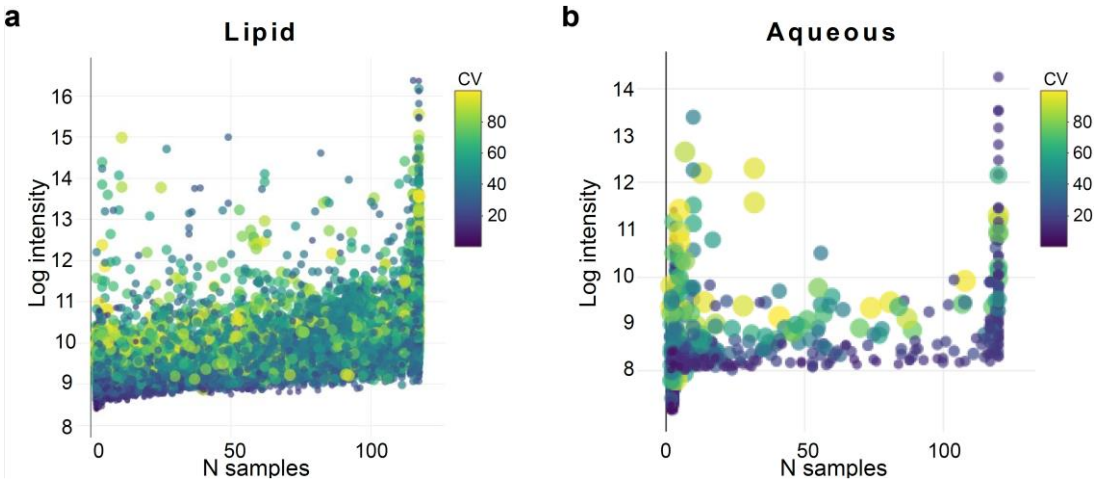
- Analysis of bronchoalveolar lavage fluid (BALF) using mass spectrometry-based metabolomics can provide insight into lung diseases, such as asthma.
- The important step of compound identification is hindered by the lack of a small molecule database that is specific for BALF.
- We assembled prototypic, small molecule databases derived from human BALF samples (n=117).
  - What is normally observed?

- *What is observable in a sample versus what is theoretically in a sample?*
  - *Current practice uses theoretically maximum permutable combinations per isotope ratio within tolerance to produce a potentially large list of formulae, and a much larger list of compound names.*

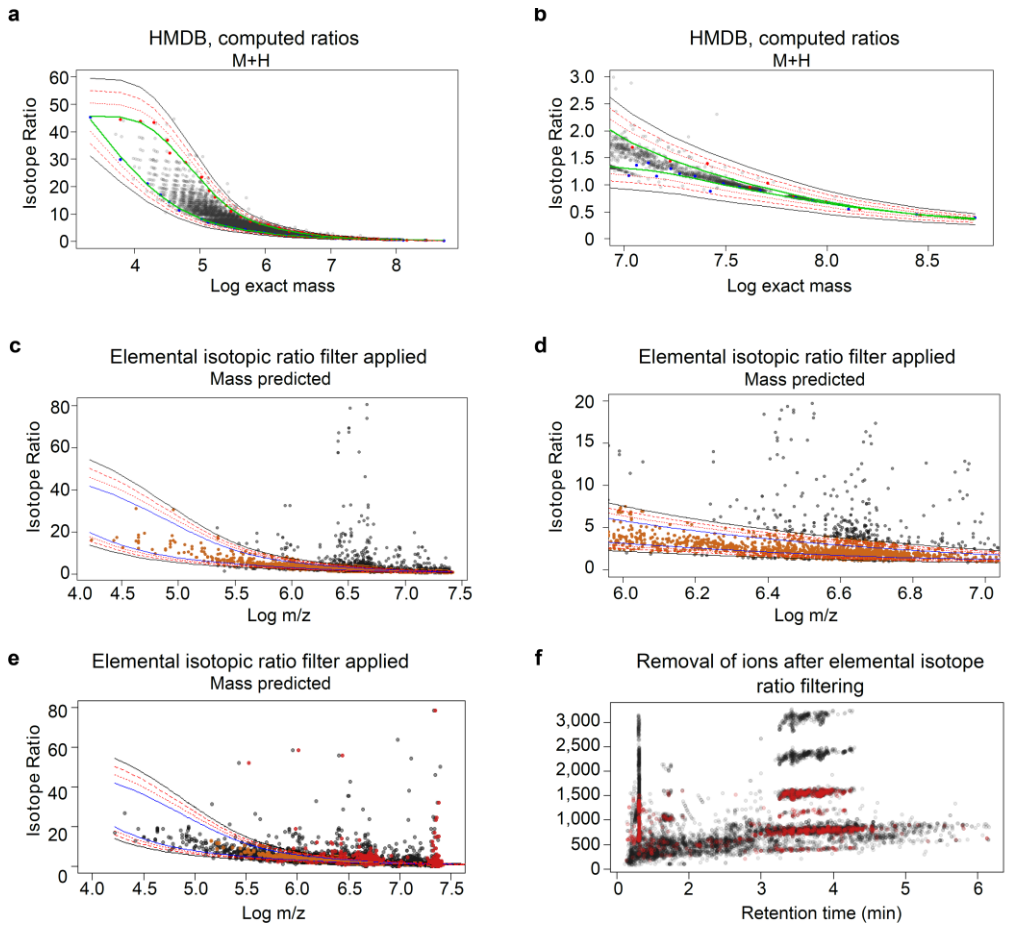
Sample specific databases



- Common practice: identify molecules every experiment. Focus on those molecules relevant to the disease.
- New paradigm: identify all that you can in a reliable system.
  - Reduce false IDs by increasing specificity in the technological system.
  - Update to annotations of higher quality as encountered in the future.



Filtering errors in isotopic cluster assignments

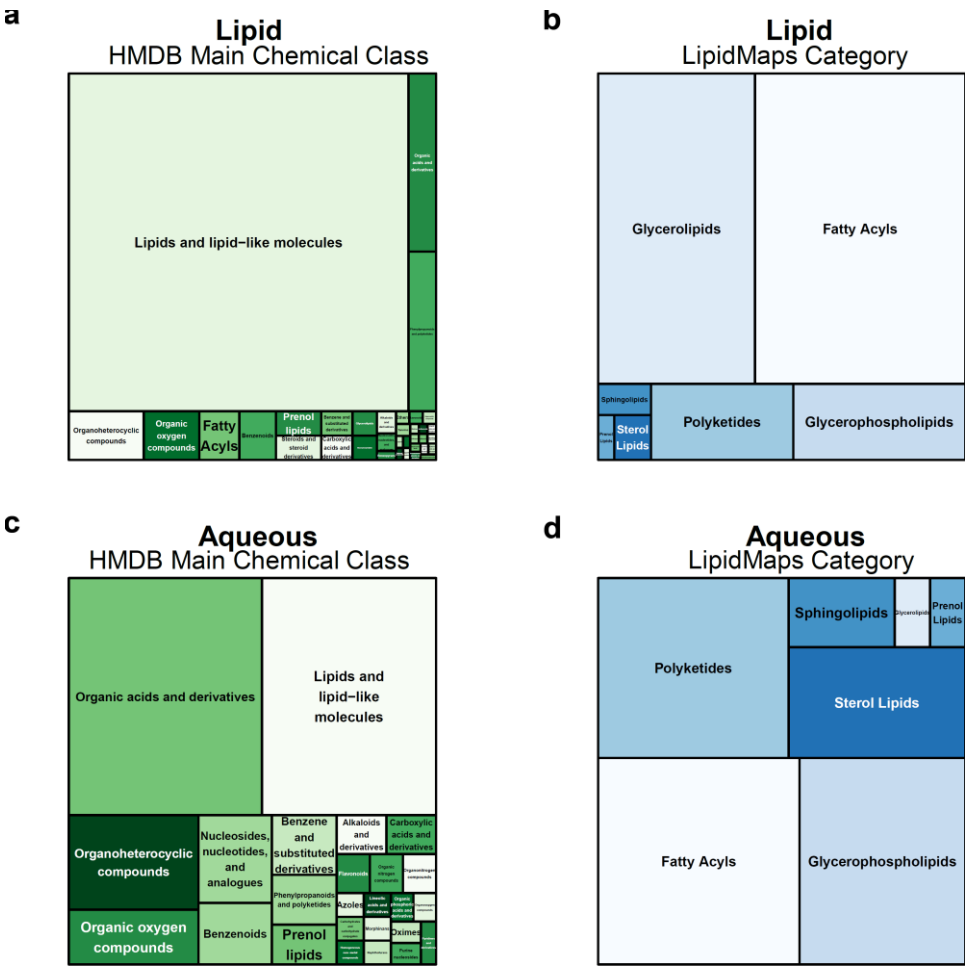


Artifact removal in metabolomics data in not common place.

Methods that aid artifact removal will be highly valued.

Expected Enrichments of Ontologies

- Expected ontology profiles?
- After collection from >100 samples, profiles are still incomplete.
- Use of database in withheld dataset produced a 90% match rate.
- Use of database in external dataset (same biofluid) produced ~45% match rate.
  - Technology (HPLC) imprecision
  - Sample variability



Acknowledgements

UC Denver- AMC Department of Pharmaceutical Sciences

Mass Spec Lab, Nichole Reisdorph, Director

MetaboDIA / MetMatch

Hyungwon Choi, Ph.D.

National university of Singapore, Saw See Hock School of Public Health ASTAR, Singapore