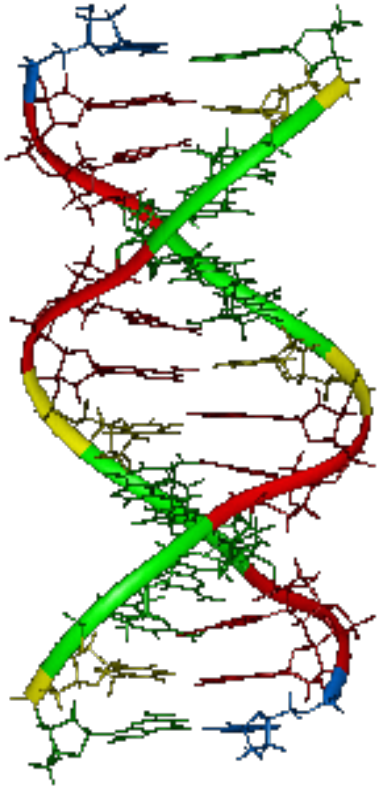
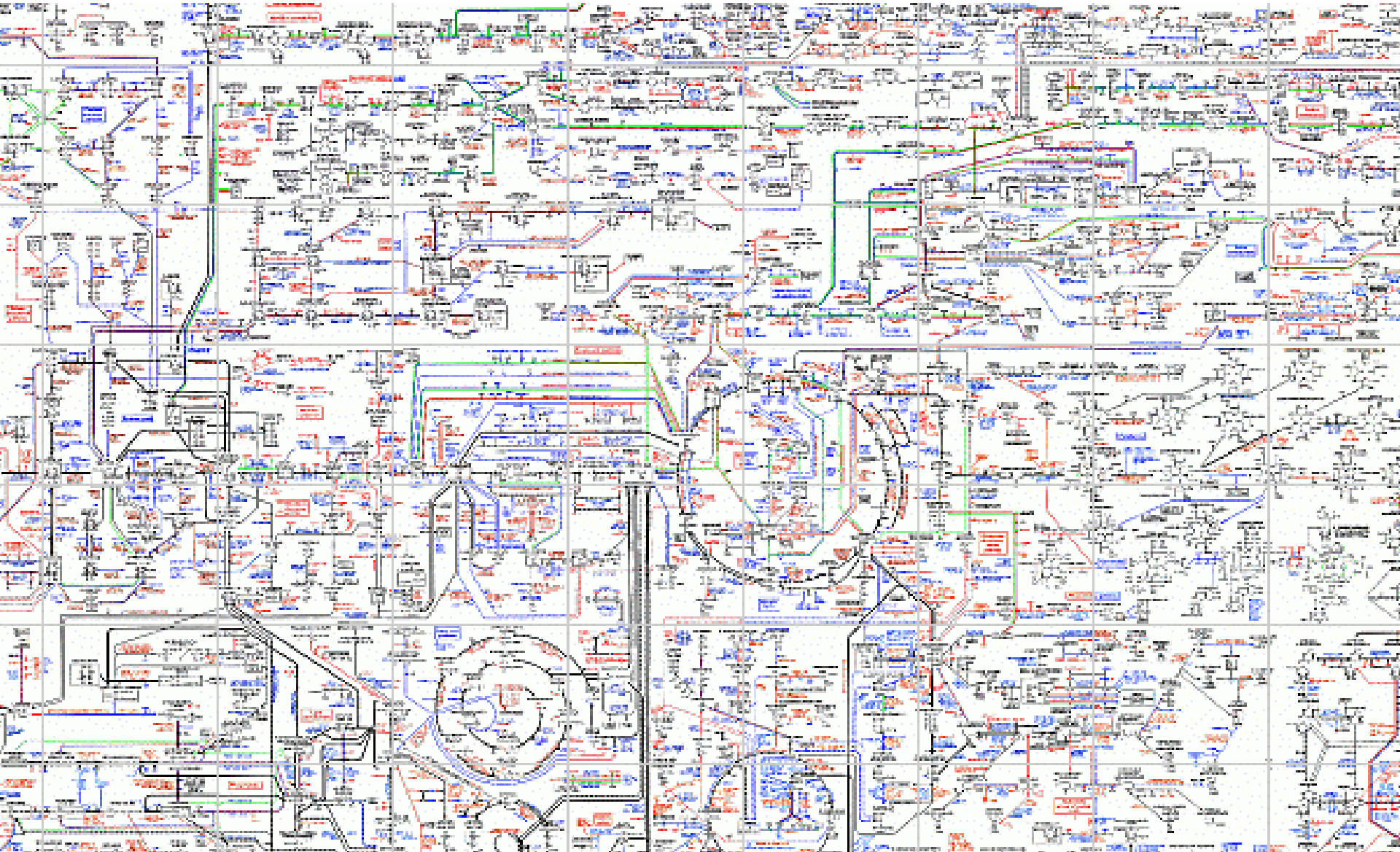
Metabolite Analysis and Mass Spectrometry

Metabolites Are the Canaries of the Genome



A single base change can lead to a 10,000X change in metabolite levels

Metabolism is “Understood”



Metabolite

- Metabolites are the intermediates and products of metabolism
- Primary metabolites are synthesized by the cell because they are indispensable for their growth
- Secondary metabolites are compounds produced by an organism that are not required for primary metabolic processes
- A by-product of the breakdown of either food

  or medication by the body
- Any compound detected in the body <1500 Da

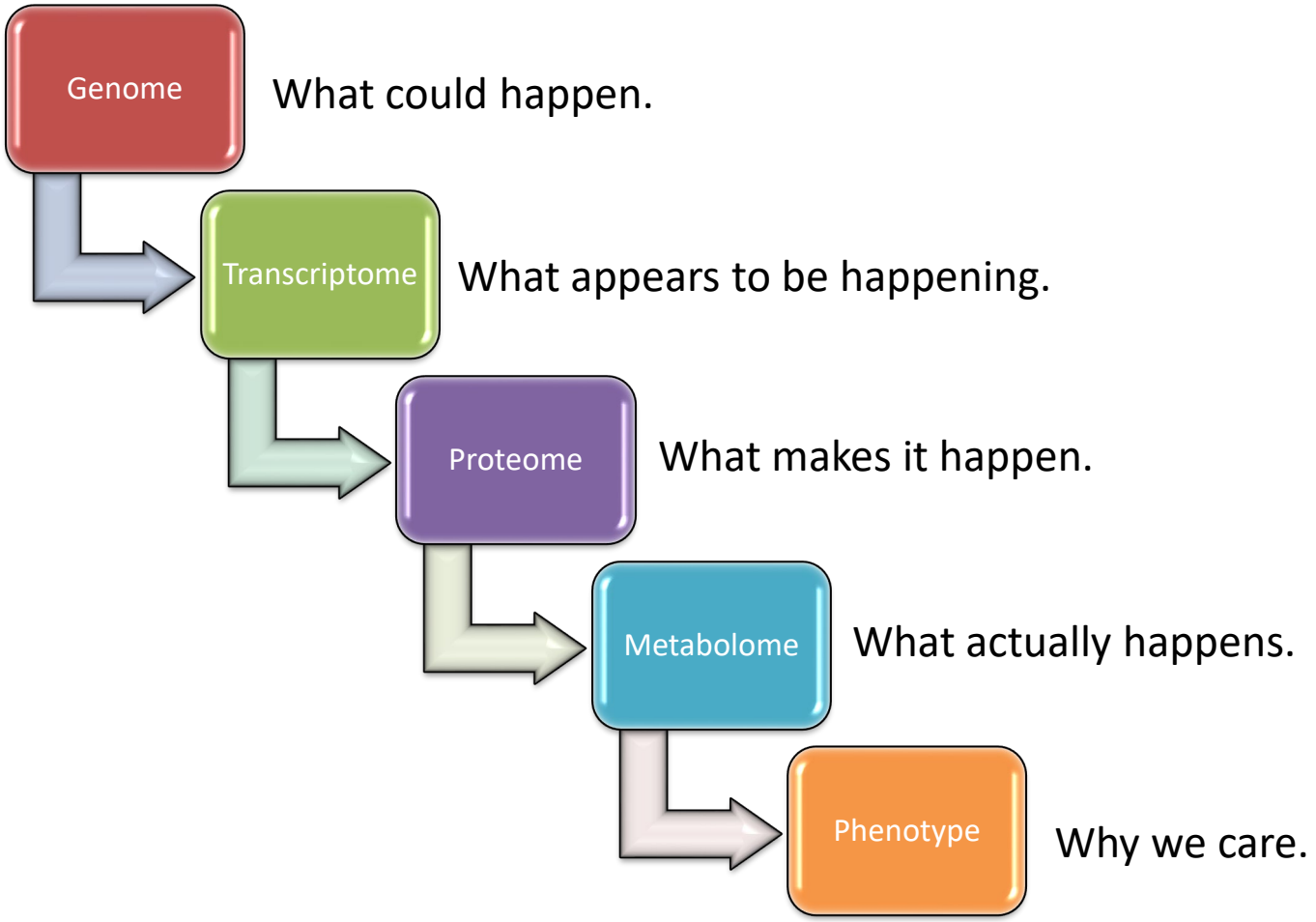
Metabolites

Primary/central metabolites		Secondary/remote metabolites	
Nucleosides	Monosaccharides	Indoles	Prostanoids
Nucleotides	Aminoglycosides	Alkaloids	Steroids
Purines	Sugar phosphates	Aminoketones	Sulfolipids
Pyrimidines	Alcohol phosphates	Cyclic amines	Leukotrienes
Coenzyme As	Acyl phosphates	Aromatic amines	Lipoamides
Pyridoxals	Sugar acids	Cobalamines	Polyketides
Biotins	Sugar lactones	Polyamines	Terpenes
Quinones	Deoxysugars	Polyenes	Macrolides
Nucleoside sugars	Phospholipids	Polyethers	Bile acids
Flavins	Glycerolipids	Polyphenols	
Pteridines	Sphingolipids	Dideoxysugars	
Tetrapyrrols	Glycolipids	Sugar sulfates	
Thiazoles	Carotenoids	Disaccharides	
Amino acids	Fatty acids	Trisaccharaides	
Acyl amino acids	Hydroxy acids	Tetrasaccharides	
Amino acid phosphates	Keto acids	Glycopeptides	
Aromatic acids	Dicarboxylic acids	Glycosylglycerides	
Alcohols and polyols	Tricarboxylic acids	Glycosylated natural products	

The Metabolome is Connected to All Other "Omes"

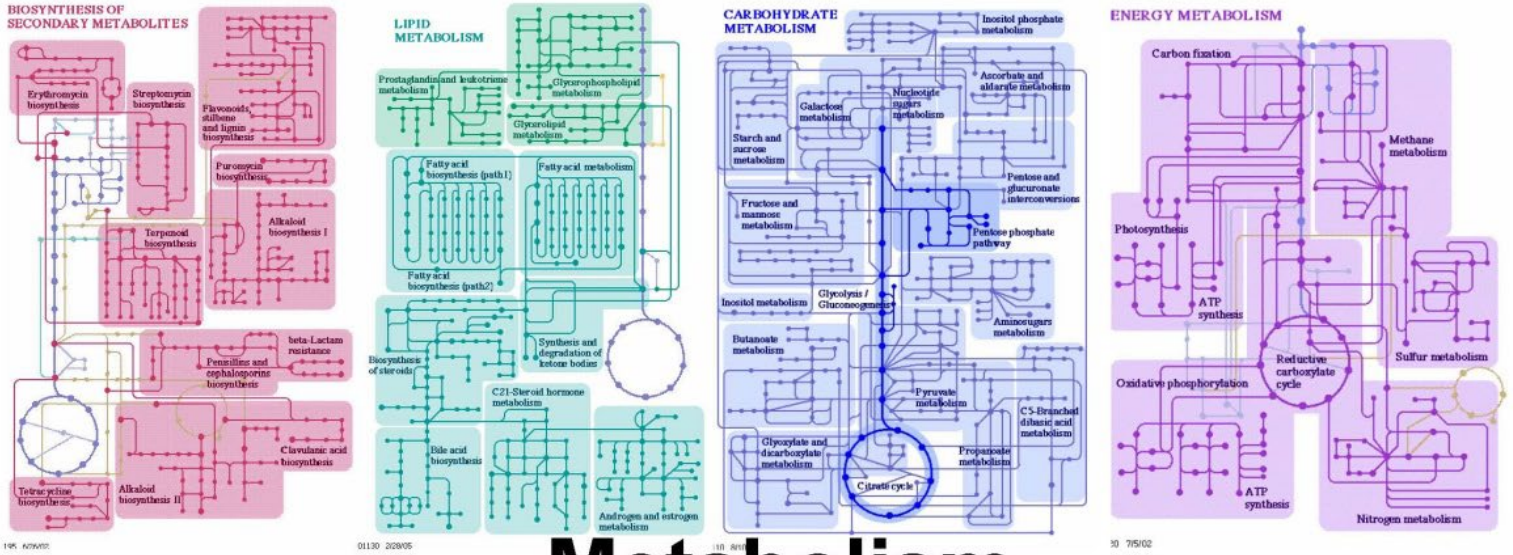
- Small molecules (i.e. AMP, CMP, GMP, TMP) are the primary constituents of the genome & transcriptome
- Small molecules (i.e. the 20 amino acids) are the primary constituents of the proteome
- Small molecules (i.e. lipids) give cells their shape, form, integrity and structure
- Small molecules (sugars, lipids, AAs, ATP) are the source of all cellular energy
- Small molecules serve as cofactors and signaling molecules for both the proteome and the genome
- *The genome & proteome largely evolved to catalyze the chemistry of small molecules*

Metabolomics to Phenotype

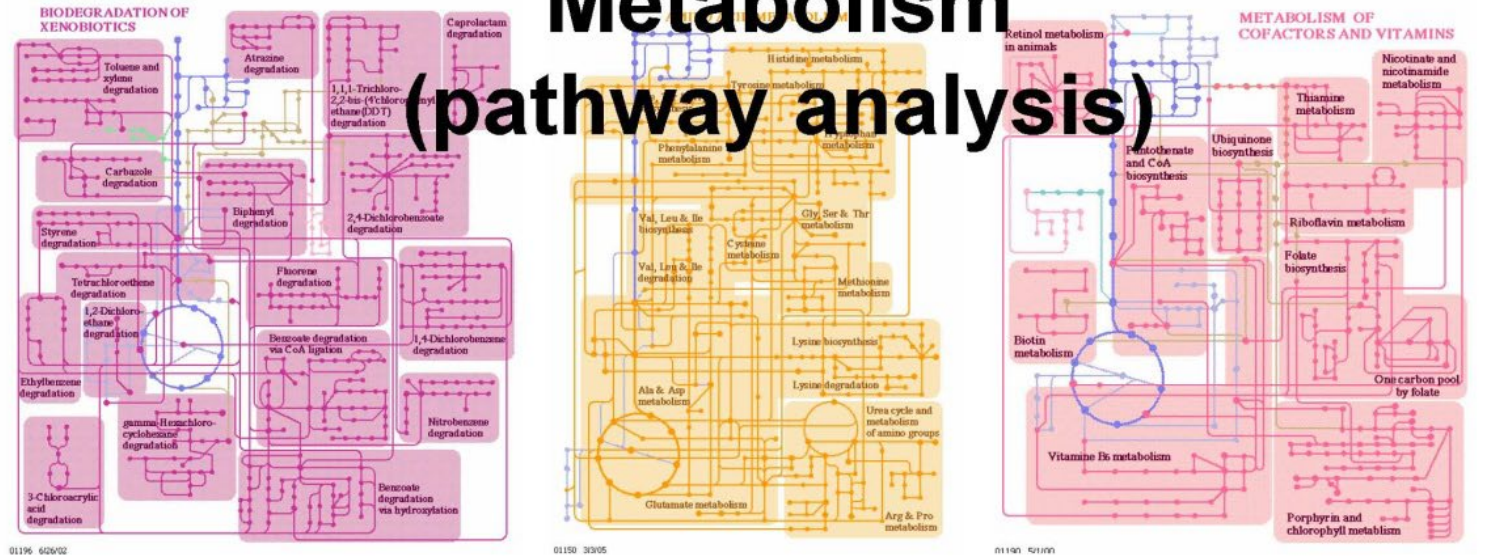


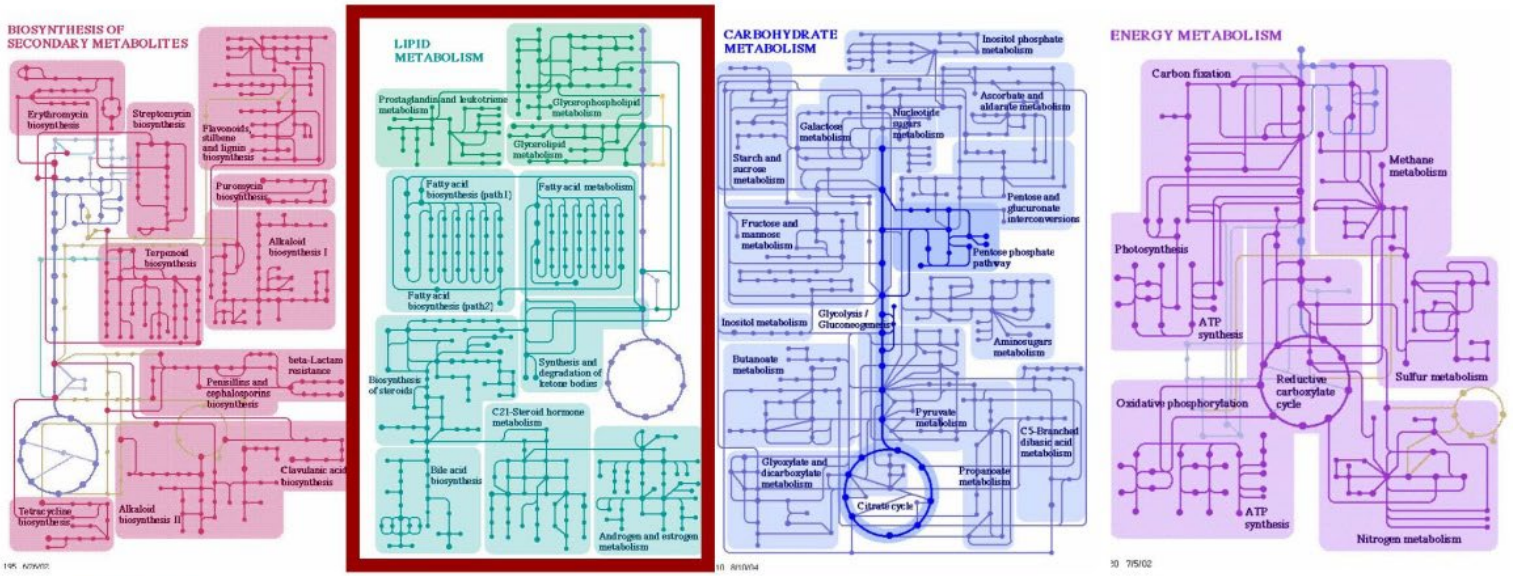
Metabolite Analysis

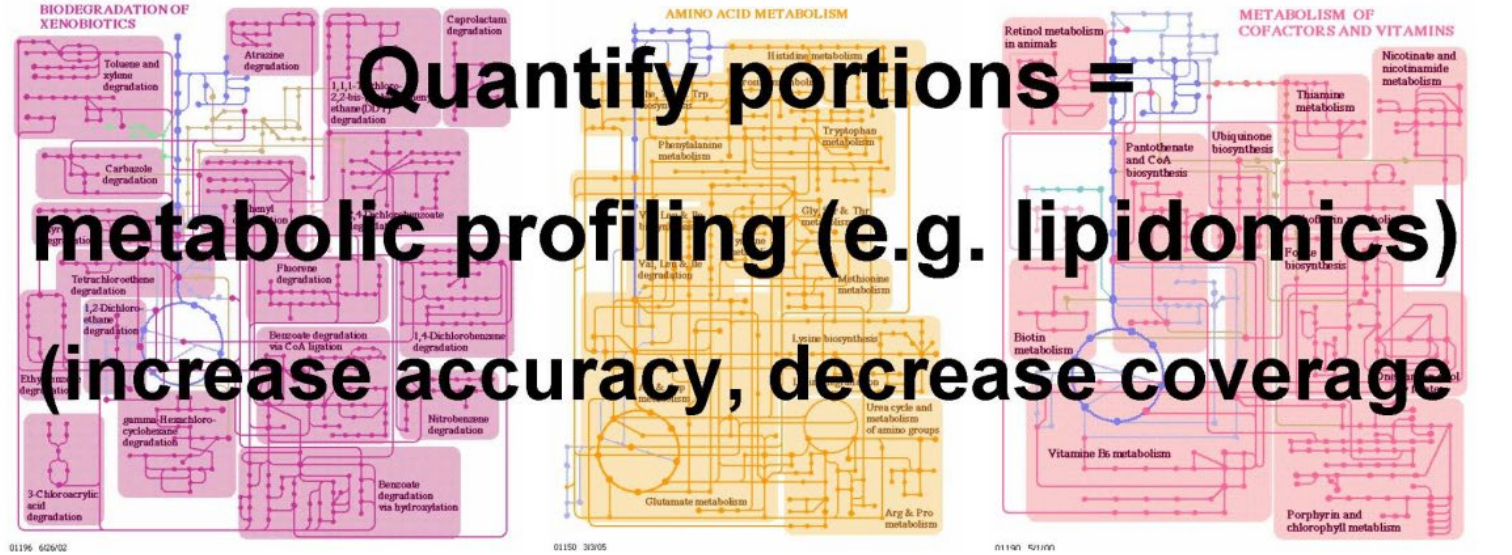
- Metabolomics: quantitative measurement of all low molecular weight metabolites in a given sample, cell or tissue.
  - Quantitation of specific metabolite.
  - Profiling a group of related compounds.
  - Quant/qual of all metabolite.
  - fingerprinting: rapid global analysis of a class of compounds.



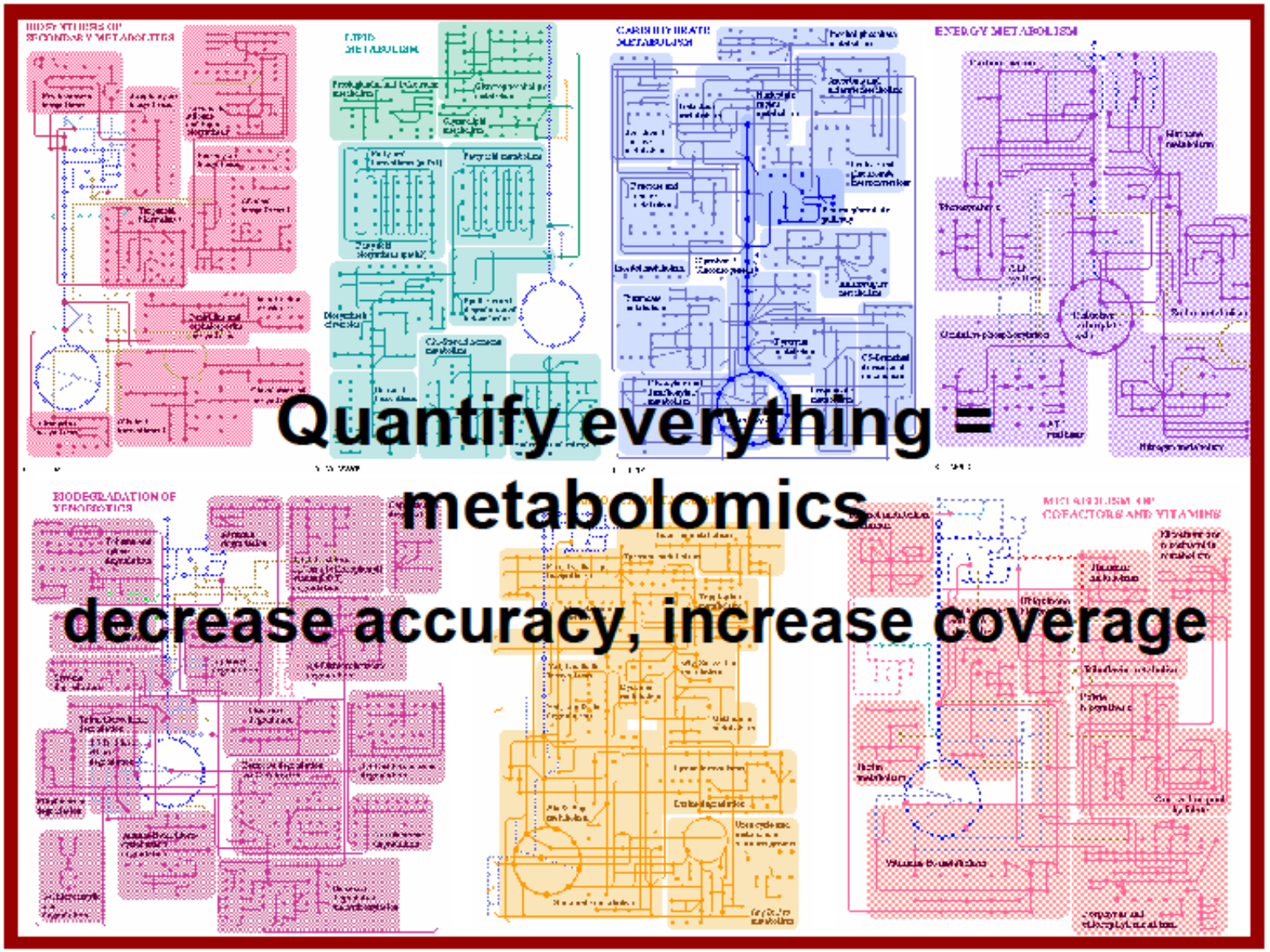
Metabolism (pathway analysis)

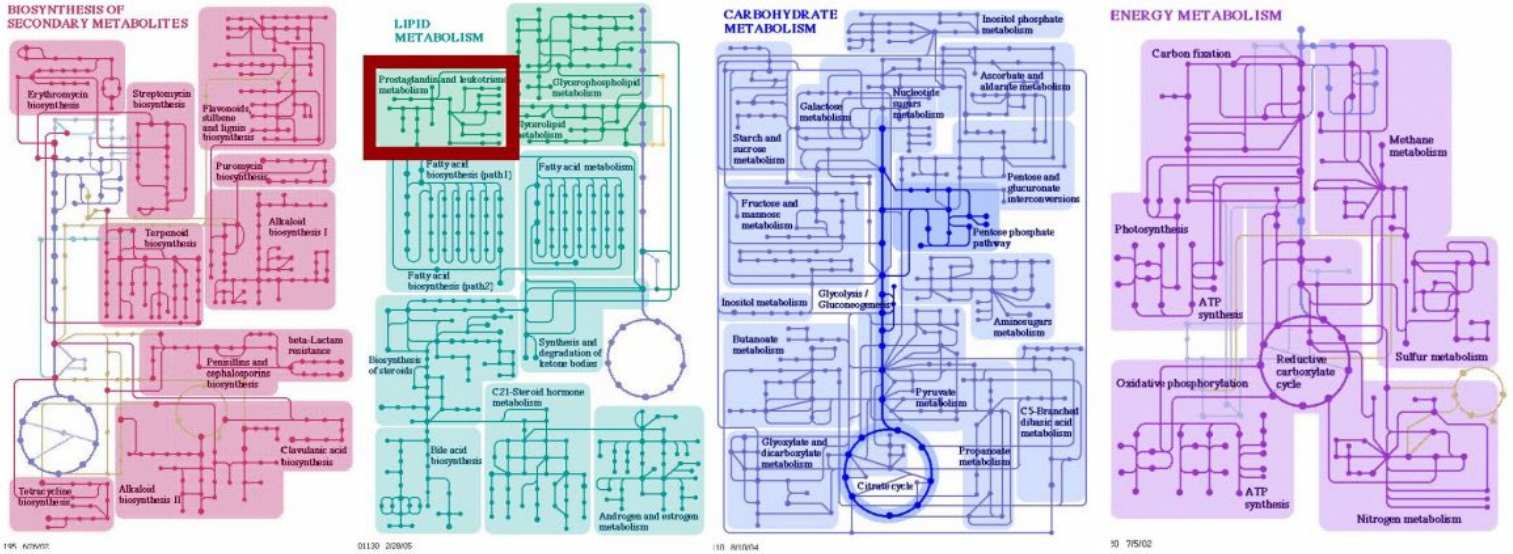


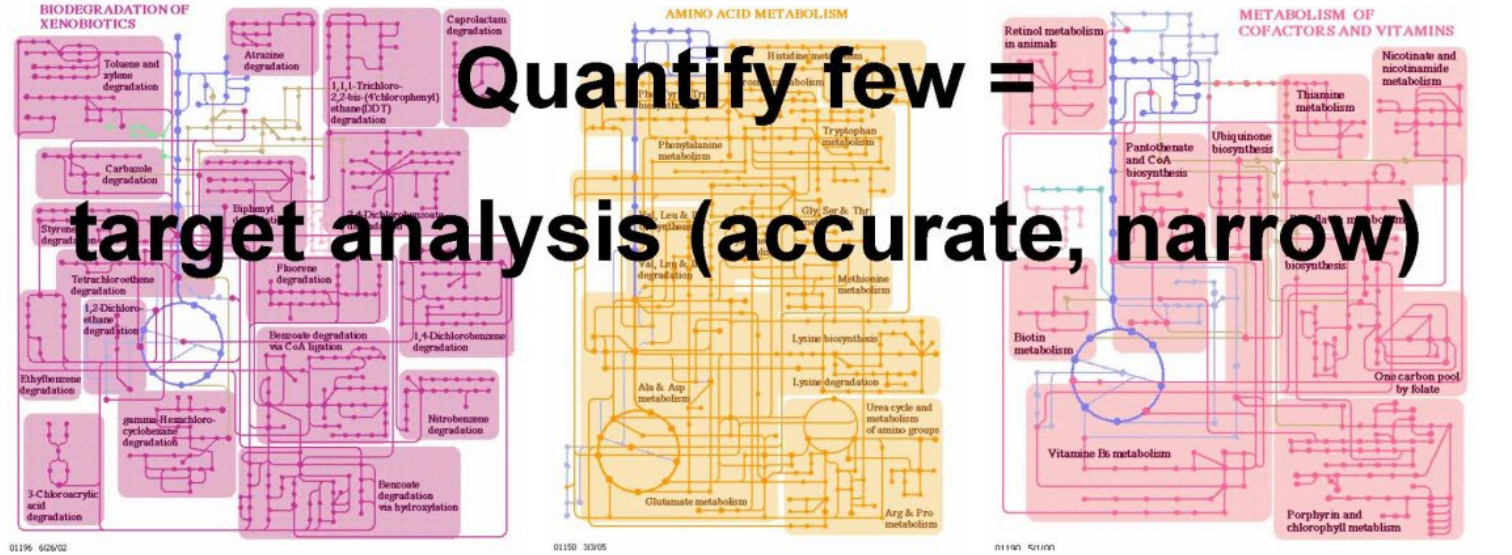




Quantify portions = metabolic profiling (e.g. lipidomics) (increase accuracy, decrease coverage


Quantify everything =
metabolomics
decrease accuracy, increase coverage





Quantify few = target analysis (accurate, narrow)

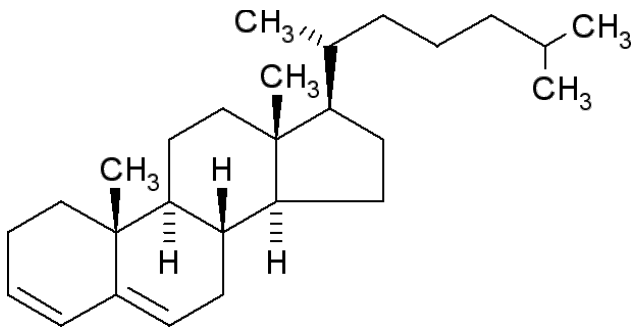
Challenges in Metabolomics

- Metabolites have a wide range of molecular weights and *large variations* in concentration

- The metabolome is much *more dynamic* than proteome and genome, which makes the metabolome more ***time sensitive***

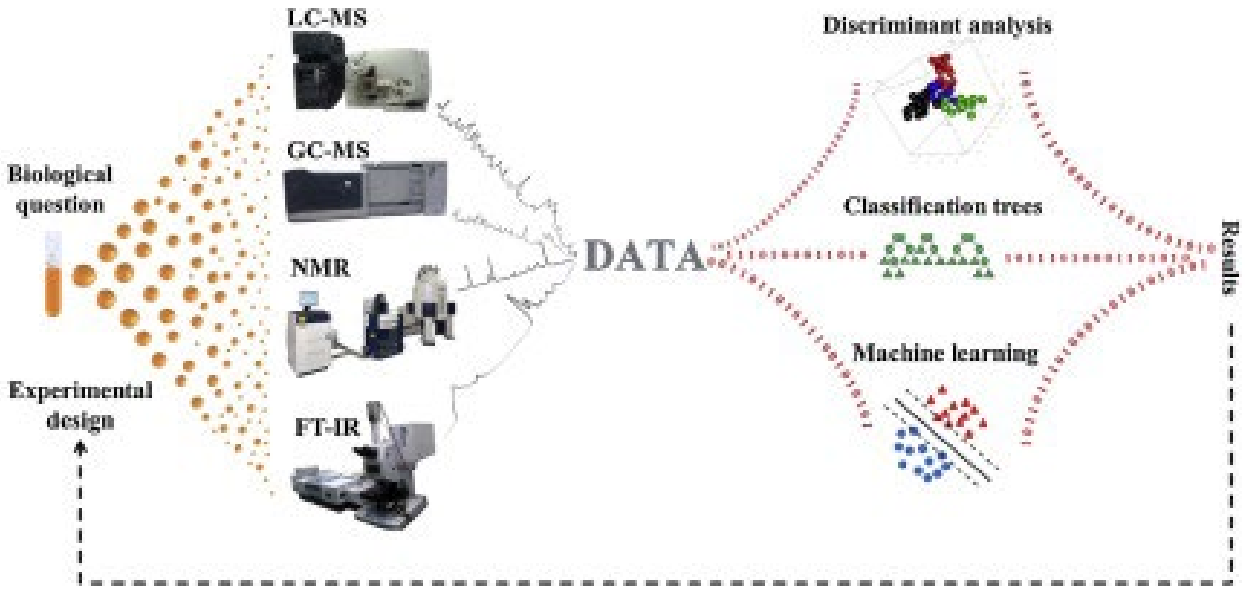
- Metabolites can be either polar or nonpolar, as well as organic or inorganic molecules. This makes the ***chemical separation*** a key step in metabolomics

- Metabolites have *chemical structures*, which makes the ***identification*** using MS an extreme challenge



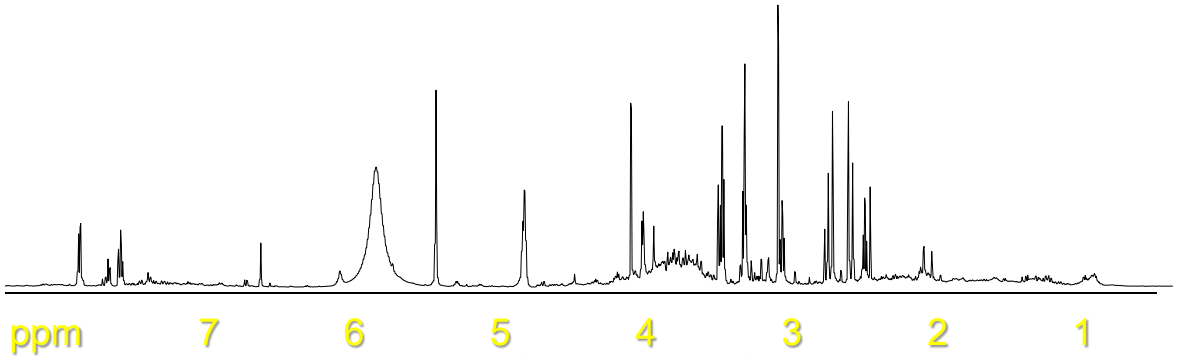
cholesta-3,5-diene

Metabolomics Flow Chart



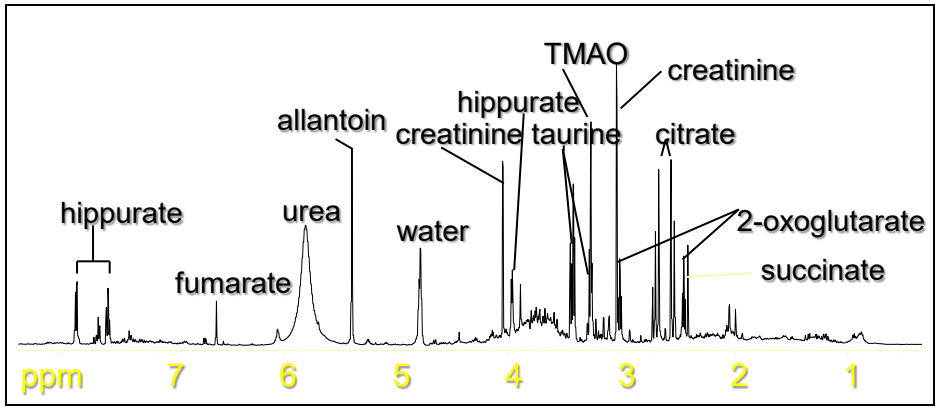
R. Gooadacre *et* al Analytica Chimica Acta 879 (2015) 10–23

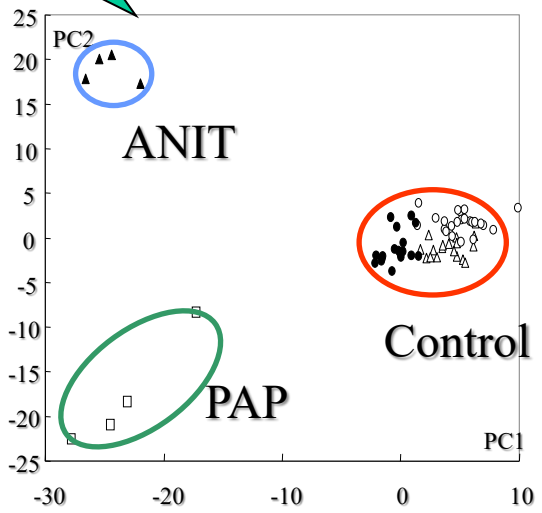
2 Routes to Metabolomics



Quantitative (Targeted) Methods

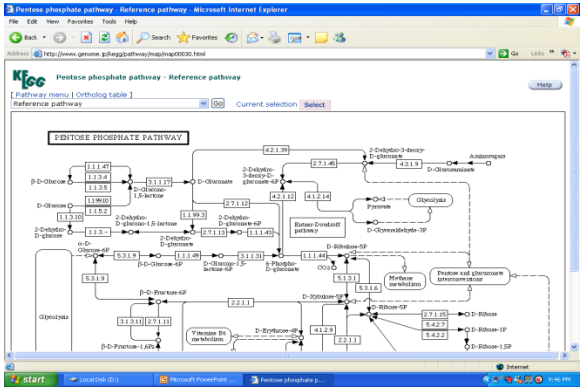
Chemometric (Profiling) Methods



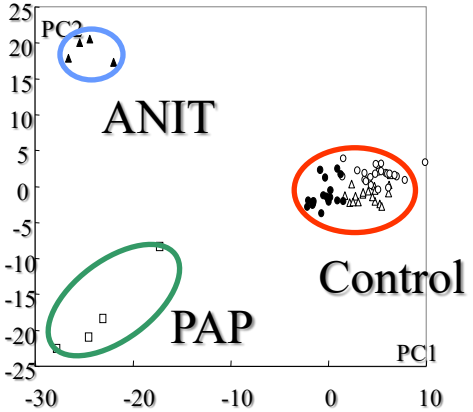


Quantitative (Targeted)

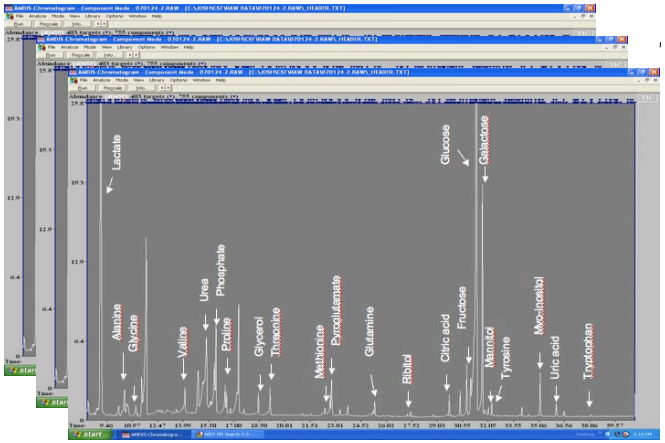
Sample Prep



Biological Interpretation



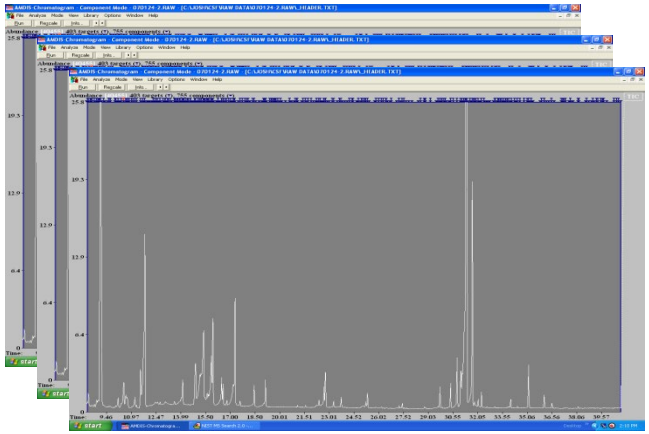
Data Reduction



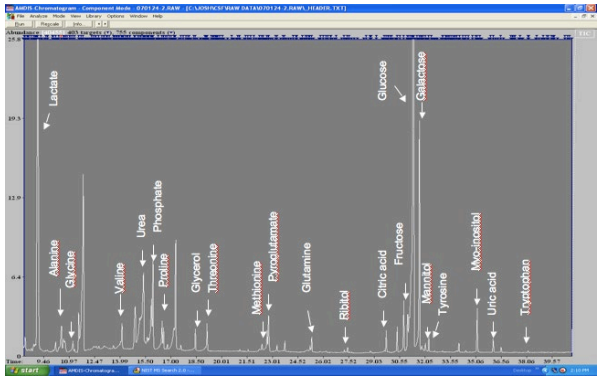
Metabolite Identification & Quantification

Profiling (Untargeted)

Sample Prep

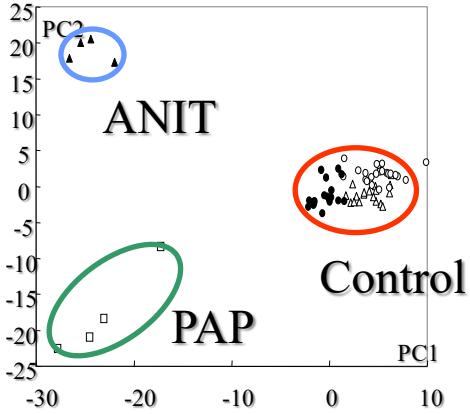


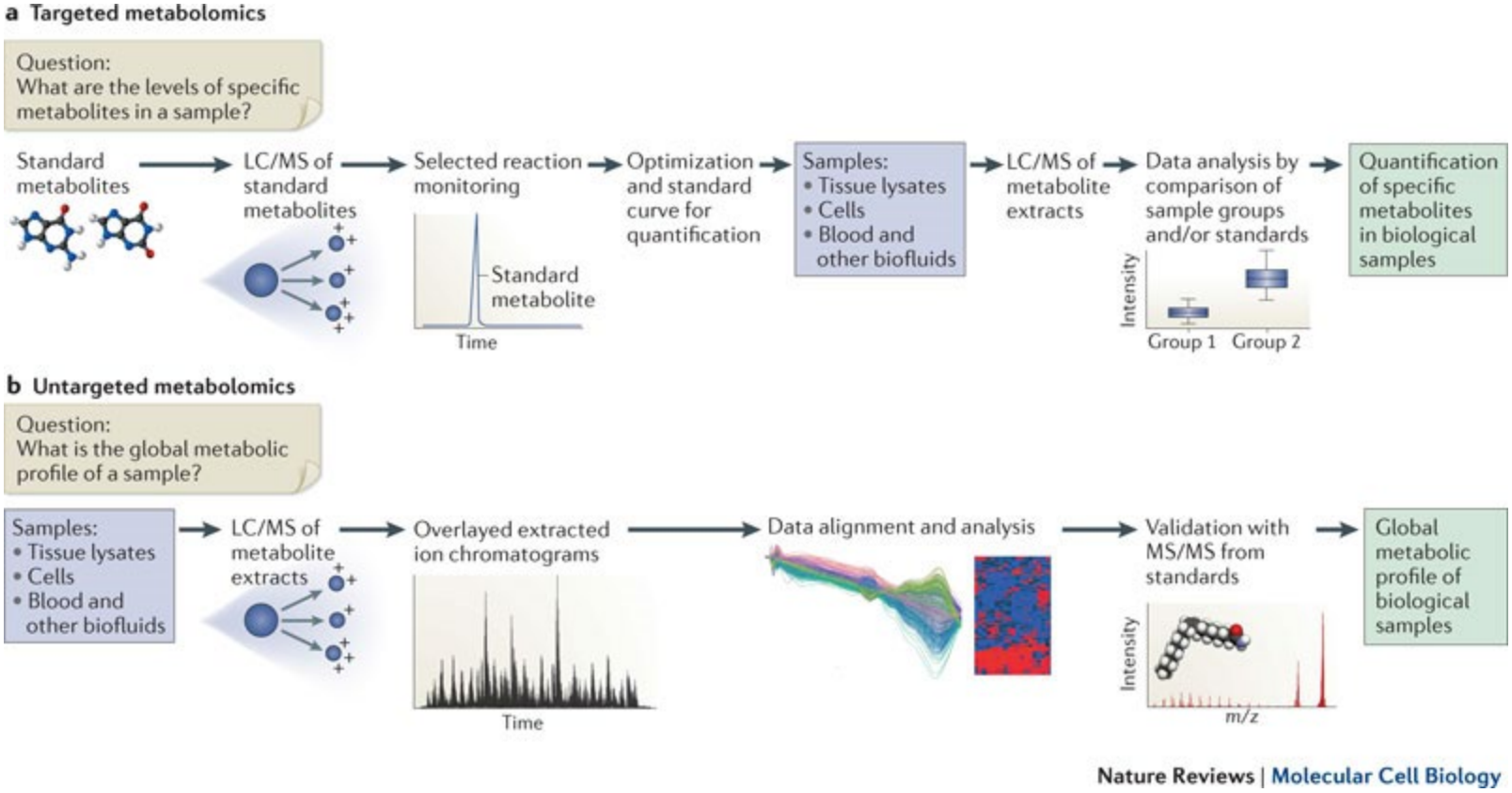
Data Collection



Metabolite Identification

Data Reduction





Nature Reviews Molecular Cell Biology 13, 263-269 (April 2012)

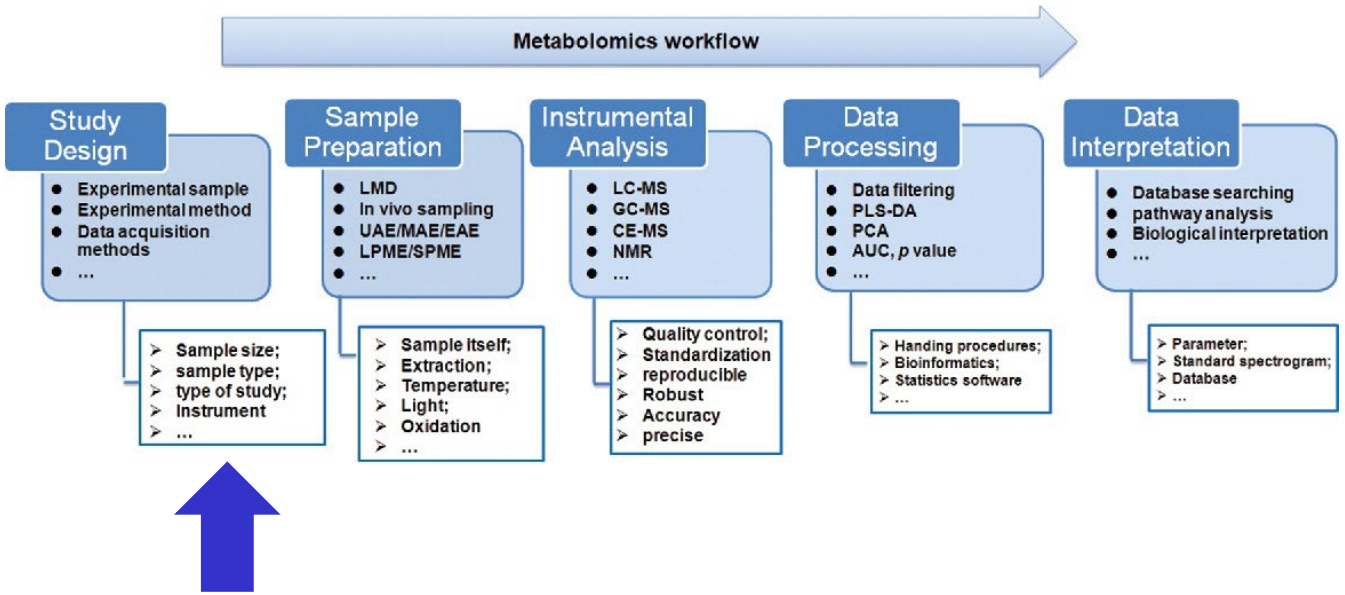
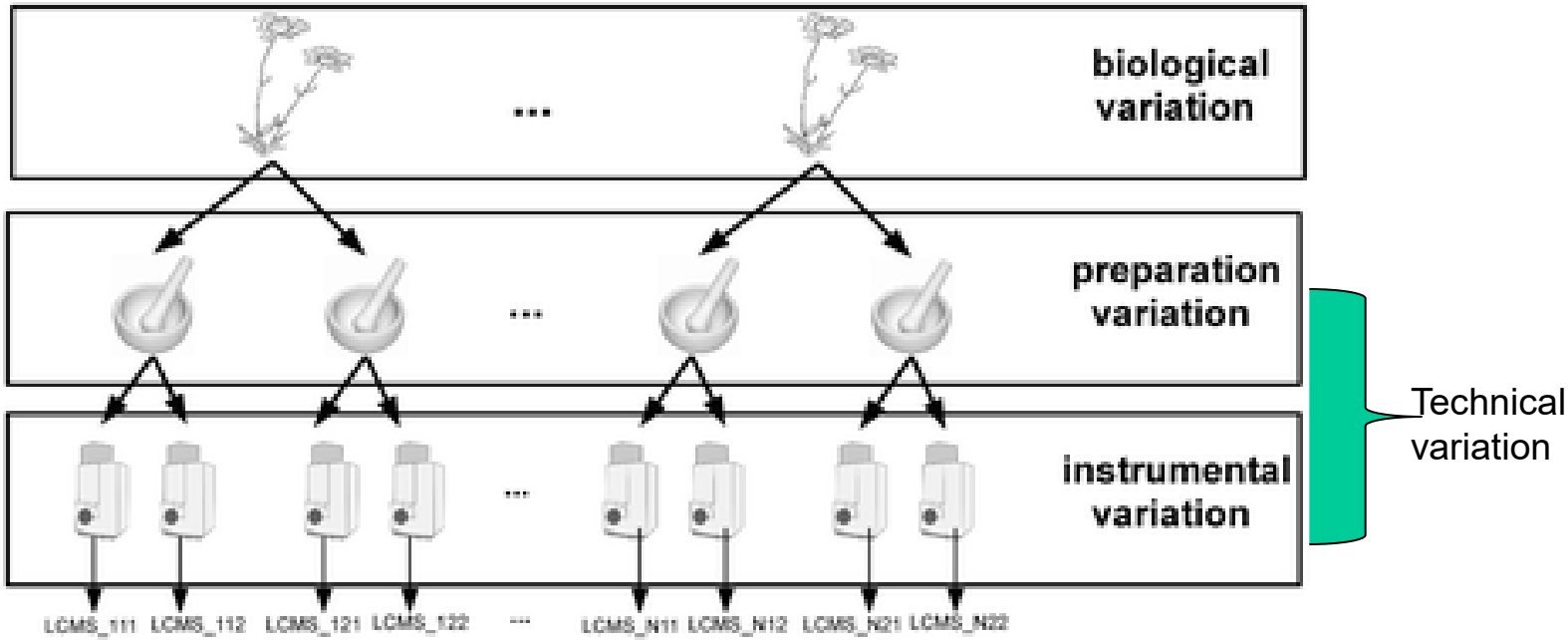


Figure 1. General workflows of metabolomics study, including study design, sample preparation, data acquirement and data mining.

Published in: Zhi-Gang Gong; Jing Hu; Xi Wu; Yong-Jiang Xu; *Critical Reviews in Analytical Chemistry* Ahead of Print
DOI: 10.1080/10408347.2017.1289836

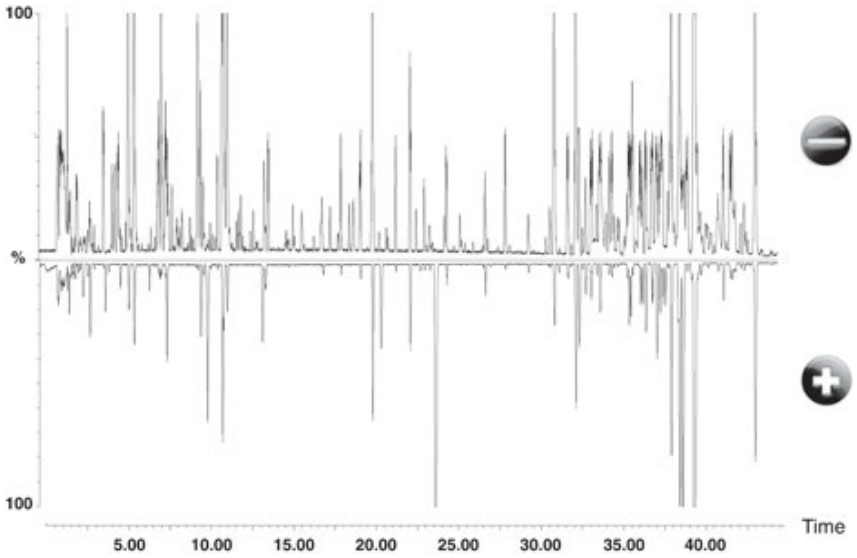
Study Design

Hierarchical Pilot Study to determine variation



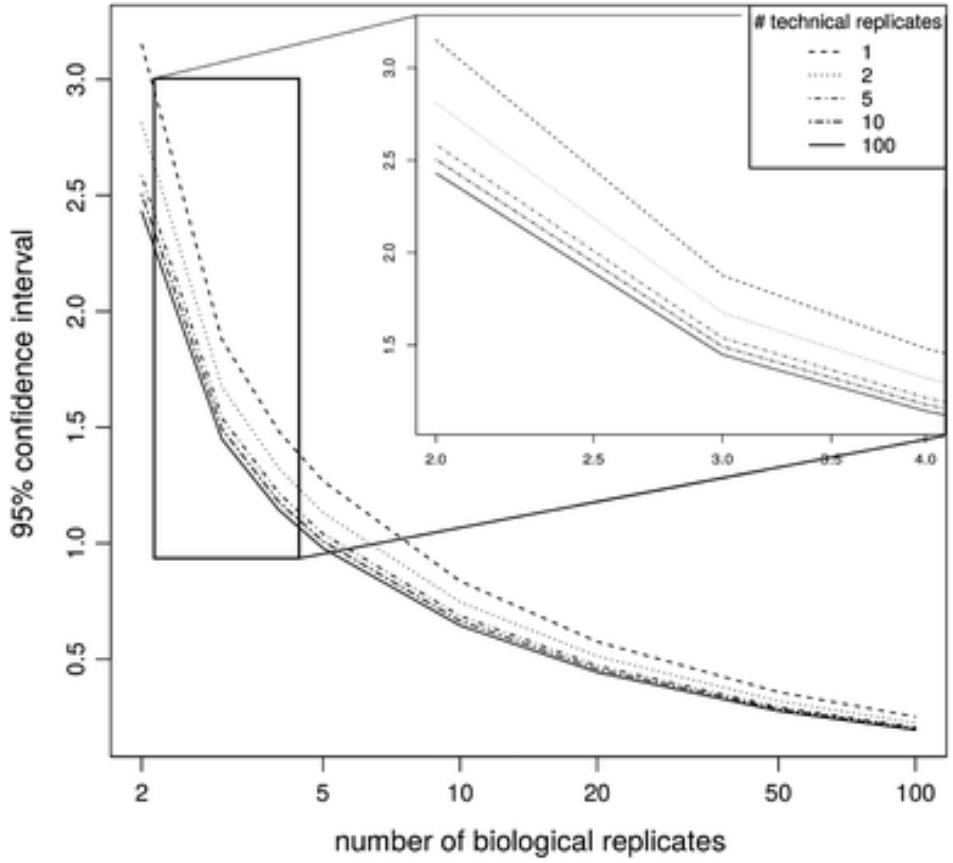
Trutschel, D., Schmidt, S., Grosse, I. et al. Metabolomics (2015) 11: 851.

Knowledge discovery in metabolomics: An overview of MS data handling



Journal of Separation Science
Volume 33, Issue 3, pages 290-304, 19 JAN 2010 DOI: 10.1002/jssc.200900609
http://onlinelibrary.wiley.com/doi/10.1002/jssc.200900609/full#fig1

Study Design



Trutschel, D., Schmidt, S., Grosse, I. et al. Metabolomics (2015) 11: 851.

cost, sample availability and the accuracy of the tests

Study Design

- Internal Standards?
  - Recovery standards to estimate the % recovery from sample preparation method
  - Analytical standards for normalization in data analysis. Must be immune from ionization/detection variation and not affect sample features. Easier for metabolite subsets than for untargted.

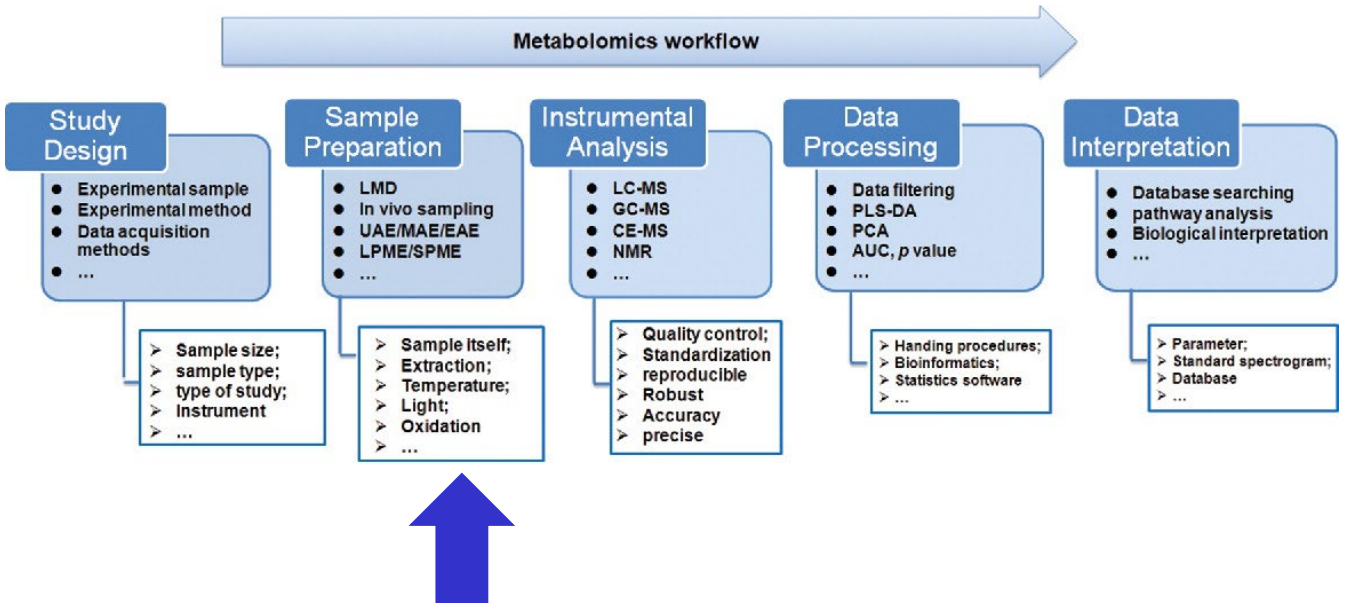
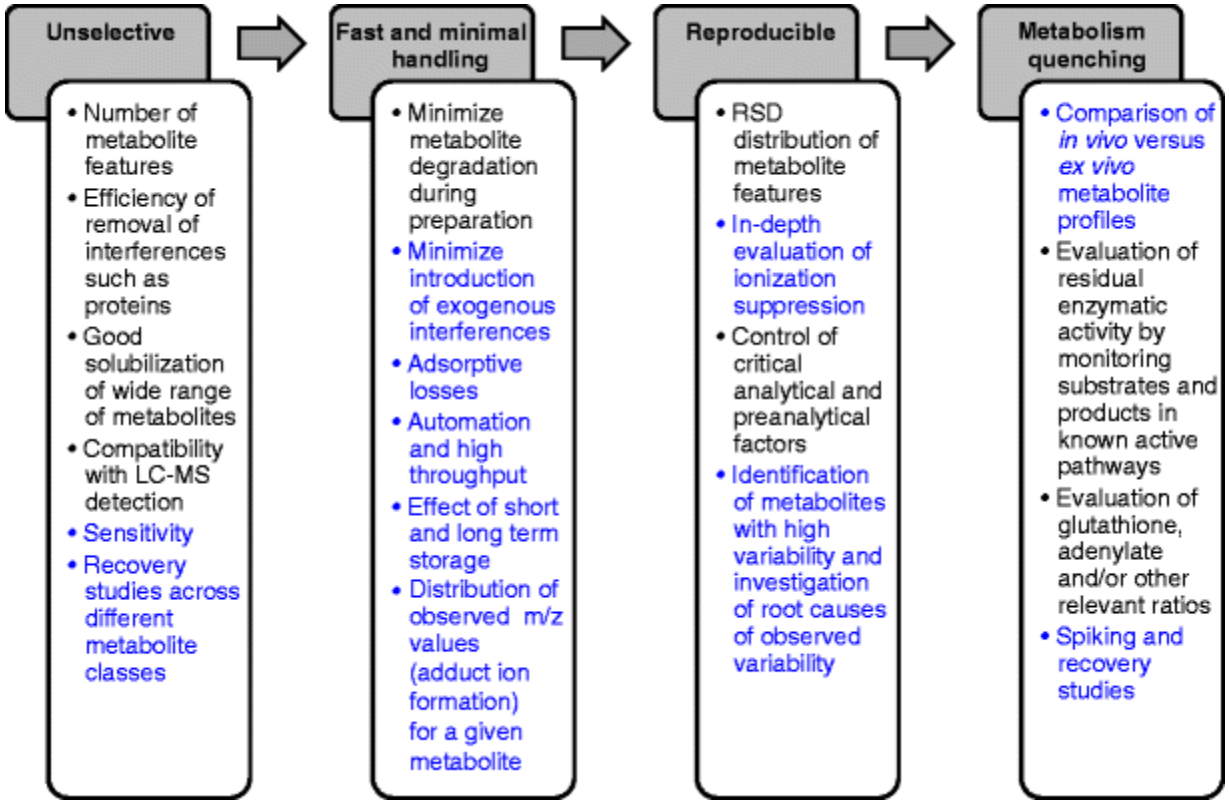


Figure 1. General workflows of metabolomics study, including study design, sample preparation, data acquirement and data mining.

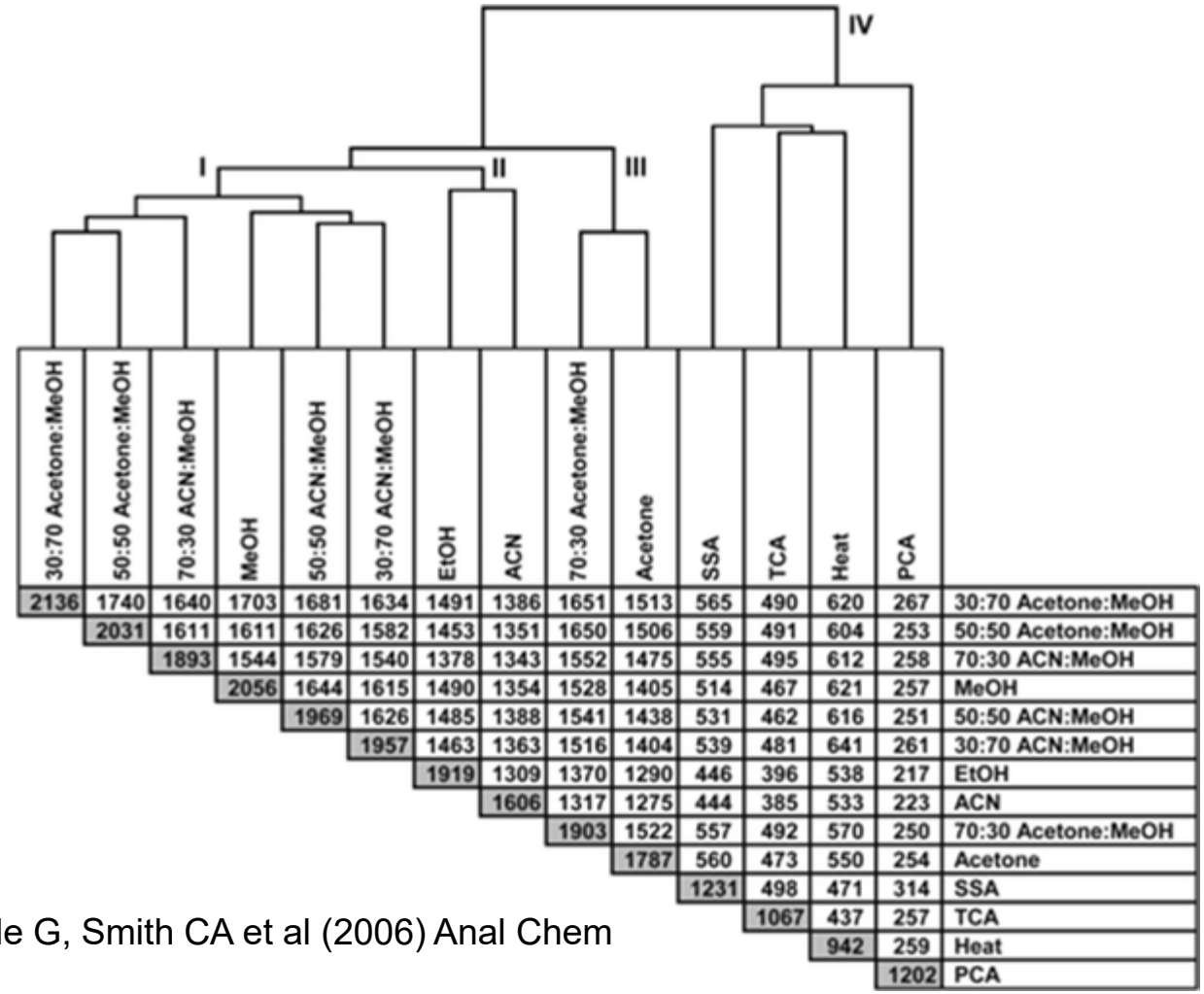
Published in: Zhi-Gang Gong; Jing Hu; Xi Wu; Yong-Jiang Xu; *Critical Reviews in Analytical Chemistry* Ahead of Print
DOI: 10.1080/10408347.2017.1289836

Sample Preparation Method Criteria



Vuckovic, D. Anal Bioanal Chem (2012) 403: 1523.

Sample Preparation Method Comparison



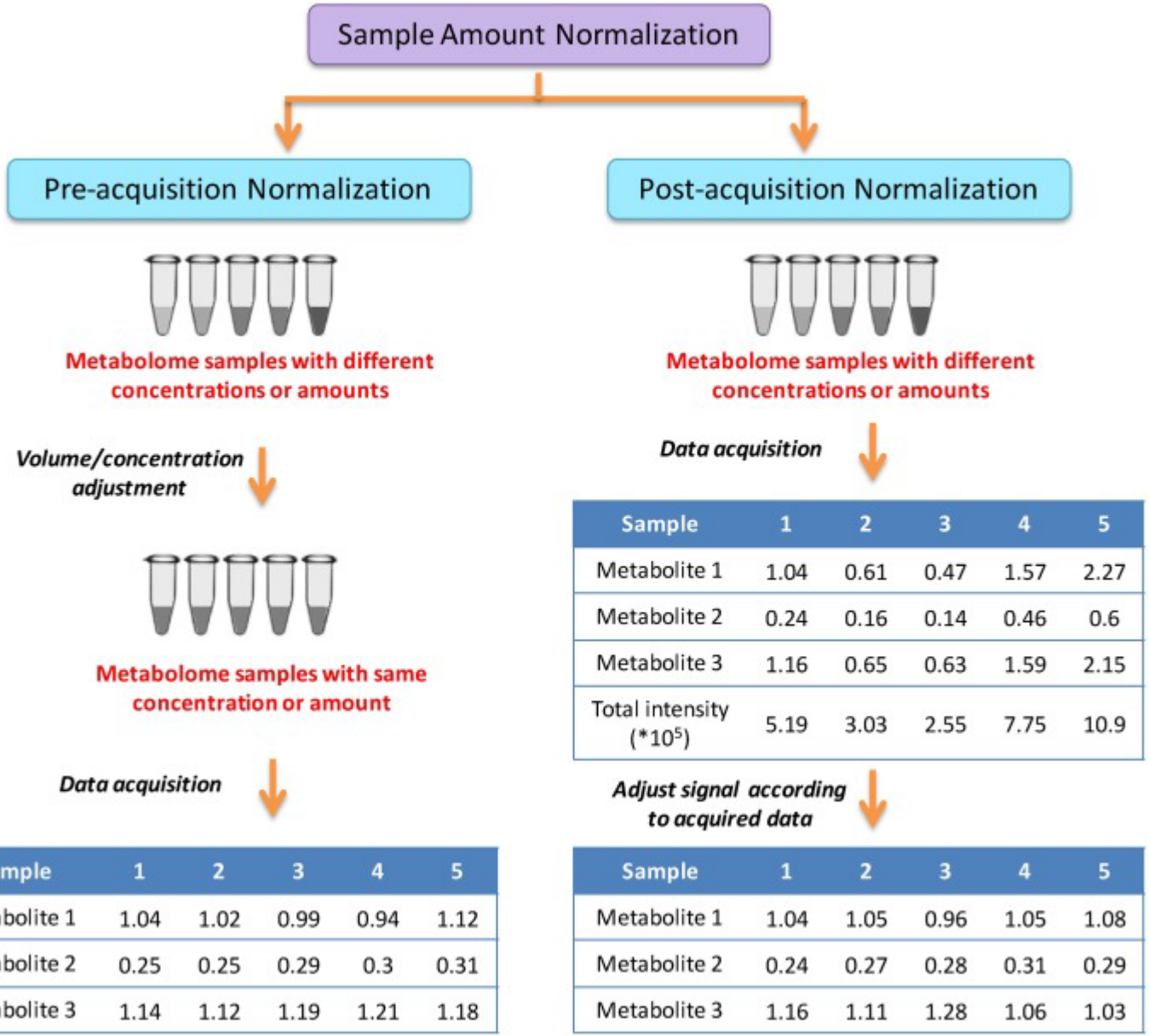
30:70 Acetone:MeOH	50:50 Acetone:MeOH	70:30 ACN:MeOH	MeOH	50:50 ACN:MeOH	30:70 ACN:MeOH	EtOH	ACN	70:30 Acetone:MeOH	Acetone	SSA	TCA	Heat	PCA	
2136	1740	1640	1703	1681	1634	1491	1386	1651	1513	565	490	620	267	30:70 Acetone:MeOH
	2031	1611	1611	1626	1582	1453	1351	1650	1506	559	491	604	253	50:50 Acetone:MeOH
		1893	1544	1579	1540	1378	1343	1552	1475	555	495	612	258	70:30 ACN:MeOH
			2056	1644	1615	1490	1354	1528	1405	514	467	621	257	MeOH
				1969	1626	1485	1388	1541	1438	531	462	616	251	50:50 ACN:MeOH
					1957	1463	1363	1516	1404	539	481	641	261	30:70 ACN:MeOH
						1919	1309	1370	1290	446	396	538	217	EtOH
							1606	1317	1275	444	385	533	223	ACN
								1903	1522	557	492	570	250	70:30 Acetone:MeOH
									1787	560	473	550	254	Acetone
										1231	498	471	314	SSA
											1067	437	257	TCA
												942	259	Heat
													1202	PCA

Want EJ, O'Maille G, Smith CA et al (2006) Anal Chem 78:743

	Metabolome coverage	Free concentration	Metabolism quenching step	Repeatability	In vivo sampling	Ionization suppression	Protein removal efficiency	Small volume samples
Solvent precipitation or extraction	Excellent but depends on solvent	No	Limited and only if rapid after collection	Excellent	No	High	Low	No
Microdialysis	No data	Yes	Yes	No data	Yes	High	High	Yes—no sample is withdrawn
Ultrafiltration	Excellent for polar metabolites; poor for hydrophobic metabolites	Yes	Limited and only if rapid after collection	Good	No	High	High	No
SPE	Good	Depends on sample pre-treatment	No	Good	No	Medium	Not evaluated to date	No
DBS	Good	No	No	Medium	No	Not evaluated to date	Low	Yes
TFC	Good	Not known	No	Medium	No	Medium	Not evaluated to date	Yes
SPME	Good but lower sensitivity	Yes	Yes (in vivo)	Good	Yes	Low	High	Yes—no sample is withdrawn for in vivo SPME

Vuckovic, D. Anal Bioanal Chem (2012) 403: 1523.

Sample Normalization



Sample	1	2	3	4	5
Metabolite 1	1.04	1.02	0.99	0.94	1.12
Metabolite 2	0.25	0.25	0.29	0.3	0.31
Metabolite 3	1.14	1.12	1.19	1.21	1.18

Sample	1	2	3	4	5
Metabolite 1	1.04	0.61	0.47	1.57	2.27
Metabolite 2	0.24	0.16	0.14	0.46	0.6
Metabolite 3	1.16	0.65	0.63	1.59	2.15
Total intensity ($*10^5$)	5.19	3.03	2.55	7.75	10.9

Sample	1	2	3	4	5
Metabolite 1	1.04	1.05	0.96	1.05	1.08
Metabolite 2	0.24	0.27	0.28	0.31	0.29
Metabolite 3	1.16	1.11	1.28	1.06	1.03

Sample Normalization

- Pre-acquisition
  - Adjust volume or weight (can input for data processing.
  - Measure concentration of known unaffected metabolite (creatine in urine).
  - Normalize to some measure osmolality, UV…

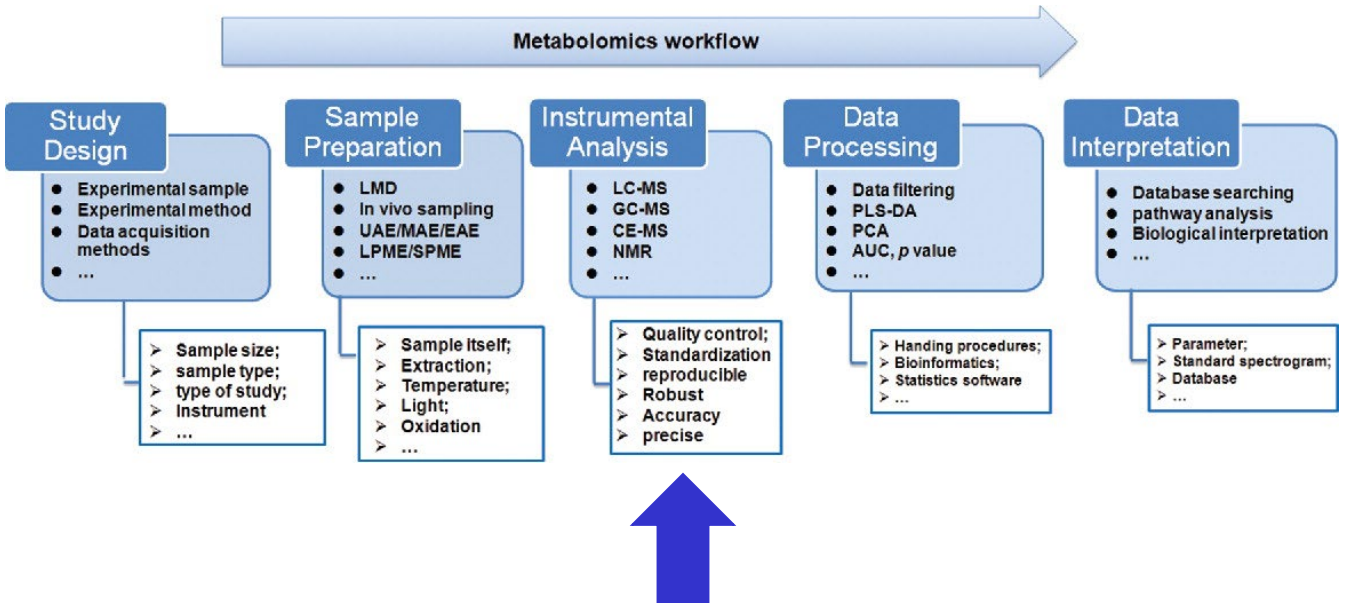


Figure 1. General workflows of metabolomics study, including study design, sample preparation, data acquirement and data mining.

Published in: Zhi-Gang Gong; Jing Hu; Xi Wu; Yong-Jiang Xu; *Critical Reviews in Analytical Chemistry* Ahead of Print
DOI: 10.1080/10408347.2017.1289836

Metabolomics Methods

Technique	Advantages	Disadvantages
GC–MS	•High resolving power and accuracy•Reproducible retention time•Applicable to volatile and semi-volatile compounds (through derivatization)•low cost compared to LC–MS and NMR•Comprehensive databases available	•Volatility can be a restriction•Heat sensitive compounds cannot generally be analysed•Derivatization may complicate sample preparation and identification (due to additives and multiple derivative products)
LC–MS	•High accuracy<comma> resolving power<comma> sensitivity and specificity•Sample preparation is minimal compared to GC–MS (generally no derivatization required)•Applicable to complex mixtures<comma> polar and non-polar compounds	•Without comprehensive MS–MS or MSnstructural information is limited•Matrix effects•Formation of multiple adducts
NMR	•Robust and highly reproducible•Provides very specific structural information•Non-destructive (samples can be recovered)•Minimal sample preparation•Highly quantitative	•Very expensive•Spectral interpretation is time consuming•Low sensitivity (micromolar range) compared to MS (picomolar range)
FT-IR	•Rapid and high throughput•Relatively inexpensive•Provides information rich data	•Water is an issue in mid IR (samples must be dried)<comma> although wet samples can be analysed using attenuated total reflectance (ATR)•Mixtures may complicate data interpretation•Not all compounds can be detected•Low chemical specificity

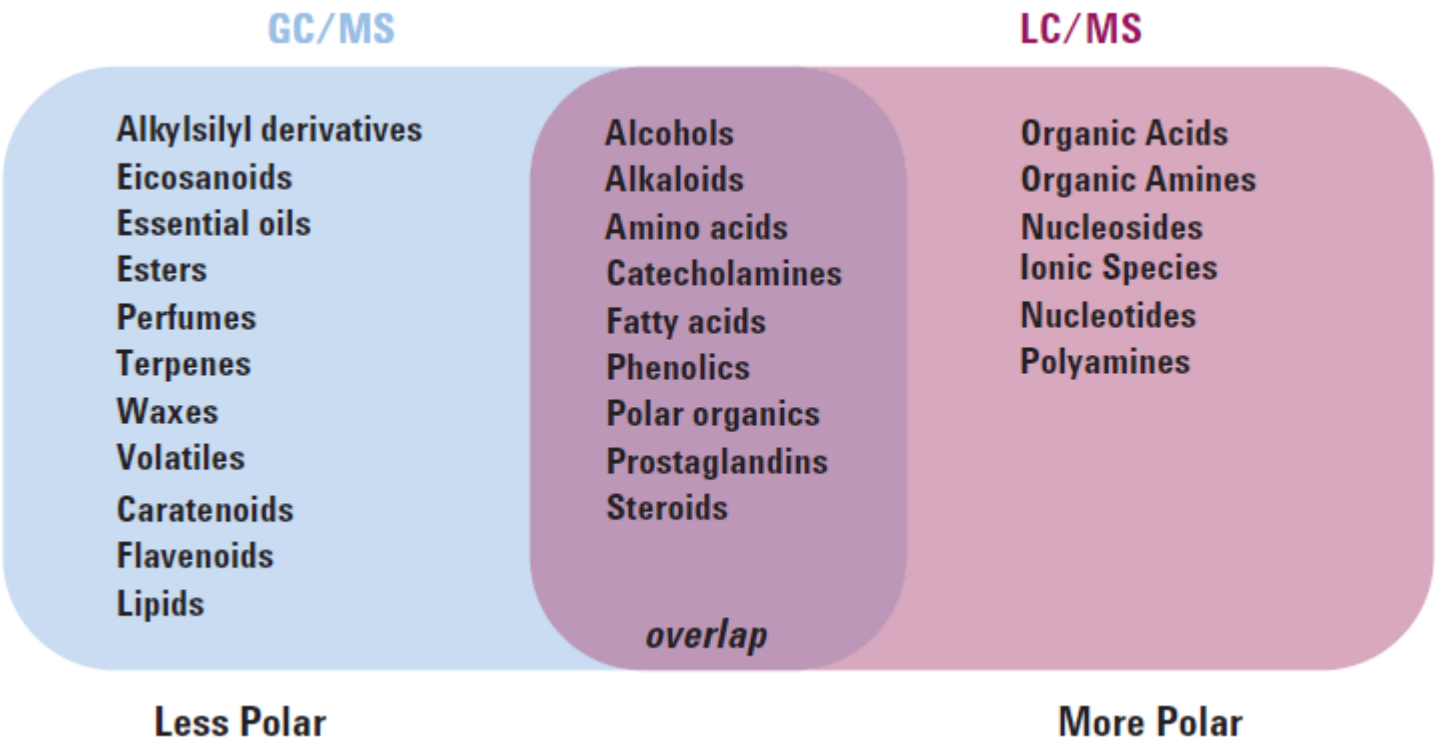
R. Gooadacre *et* al Analytica Chimica Acta 879 (2015) 10–23

NMR or MS Profiling	Very Robust	High Throughput	500-10,000 variables	5-50 identified compounds
GC-MS	Robust	High Throughput	500-1000 primary metabolites <550 Da	50-200 identified compounds
LC-MS	Not as Robust	Low Throughput	50-70 secondary metabolites <2500 Da	20-100 identified compounds

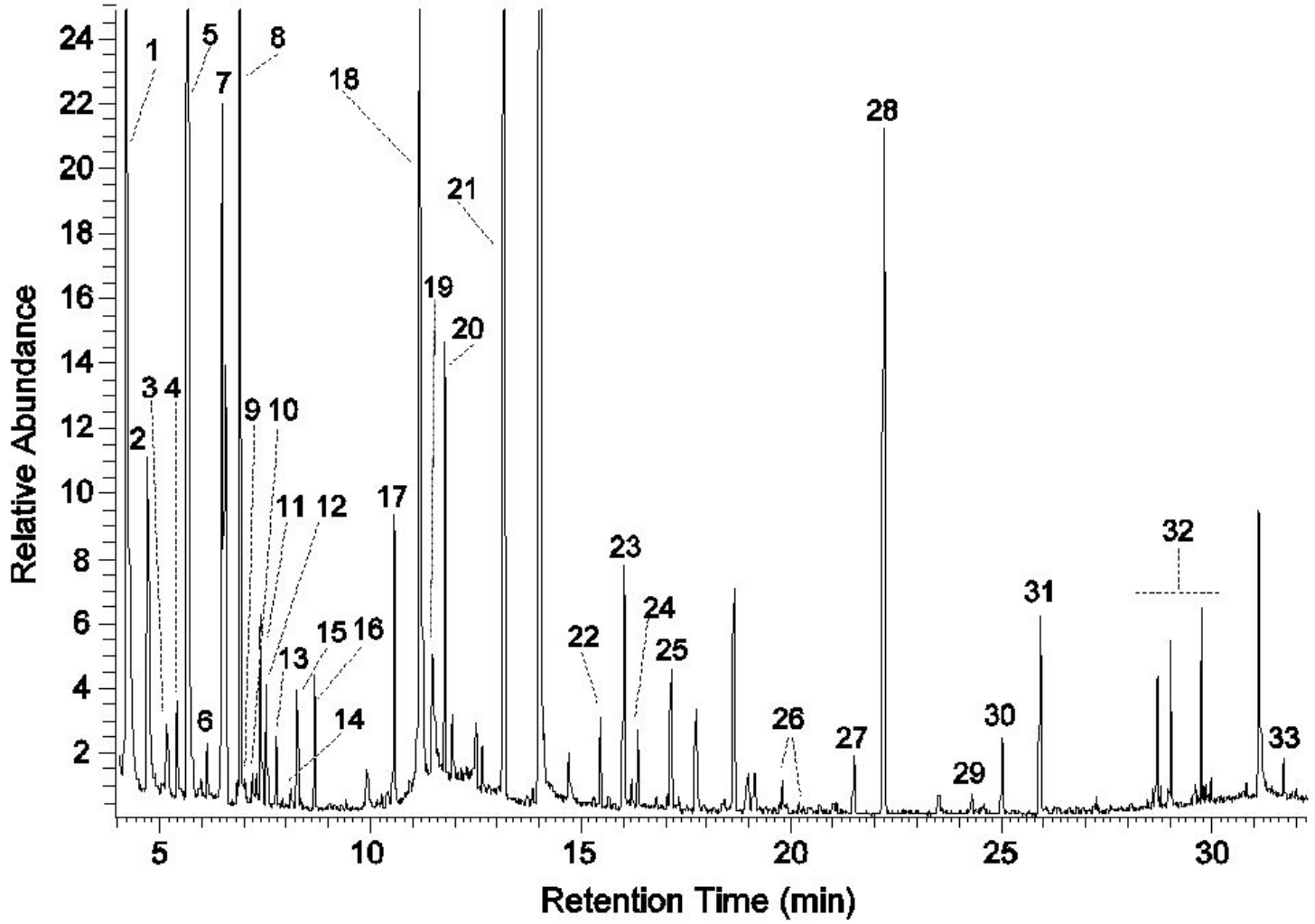
Metabolomics Separation methods

- Gas chromatography (GC)
  – one of the most widely used and powerful methods
  – high chromatographic resolution
  – compounds must be volatile (or derivatized)
- High performance liquid chromatography (HPLC)
  – lower chromatographic resolution
  – wider range of analytes can be analyzed (polar)
- Capillary electrophoresis (CE)
  – higher theoretical separation efficiency than HPLC
  – suitable for wider range of metabolites than GC
  – most appropriate for charged analytes (electrophoretic technique)

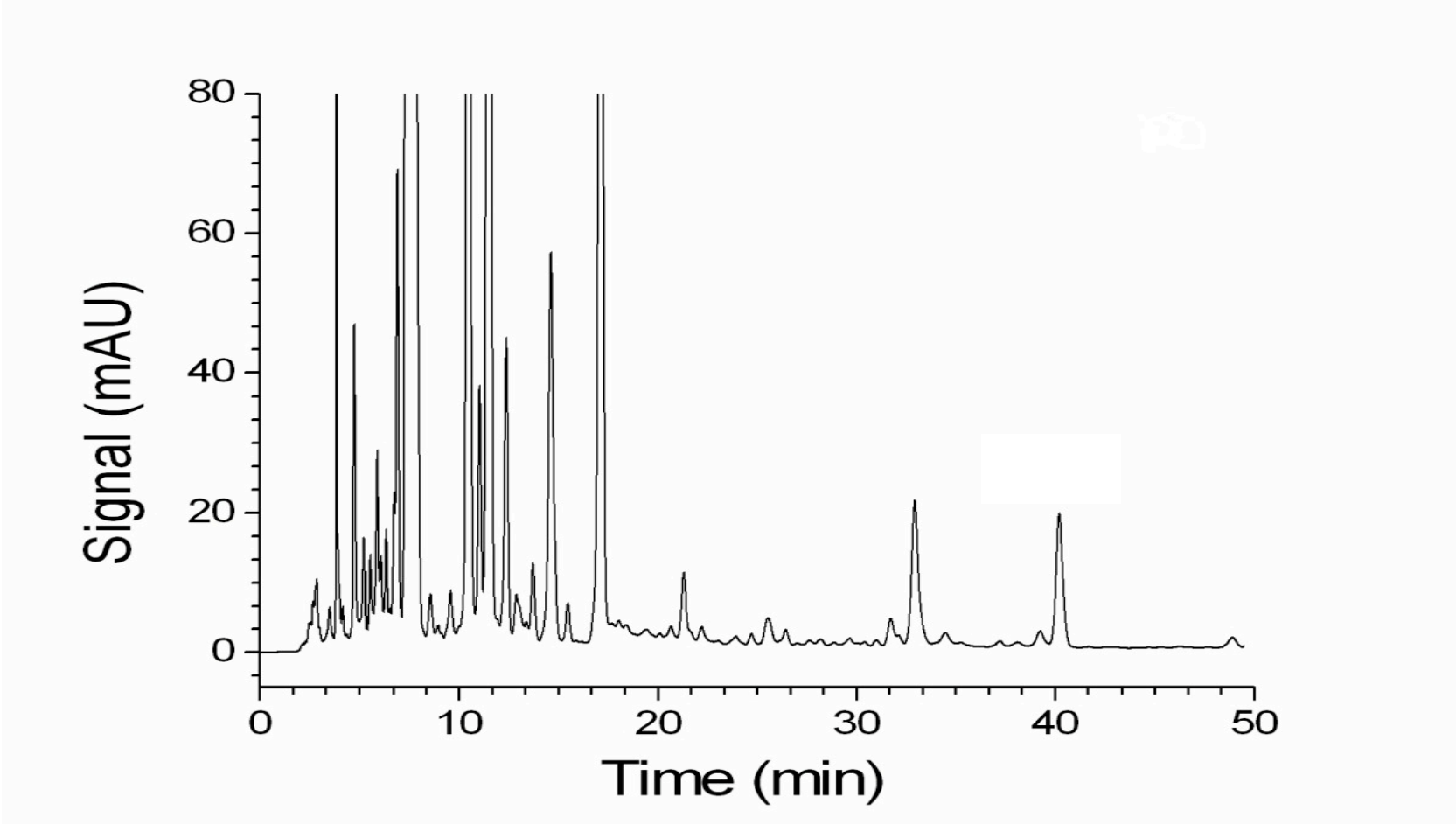
Metabolite Analysis



GC-MS Chromatogram of a Biological Mixture



HPLC of a Biological Mixture



NMR

- Advantages
  - No sample separation necessary
  - Essentially universal detector
  - Non-destructive
- Disadvantages
  - Low sensitivity
  - Results difficult to interpret
  - Decreased quantification
  - Expensive

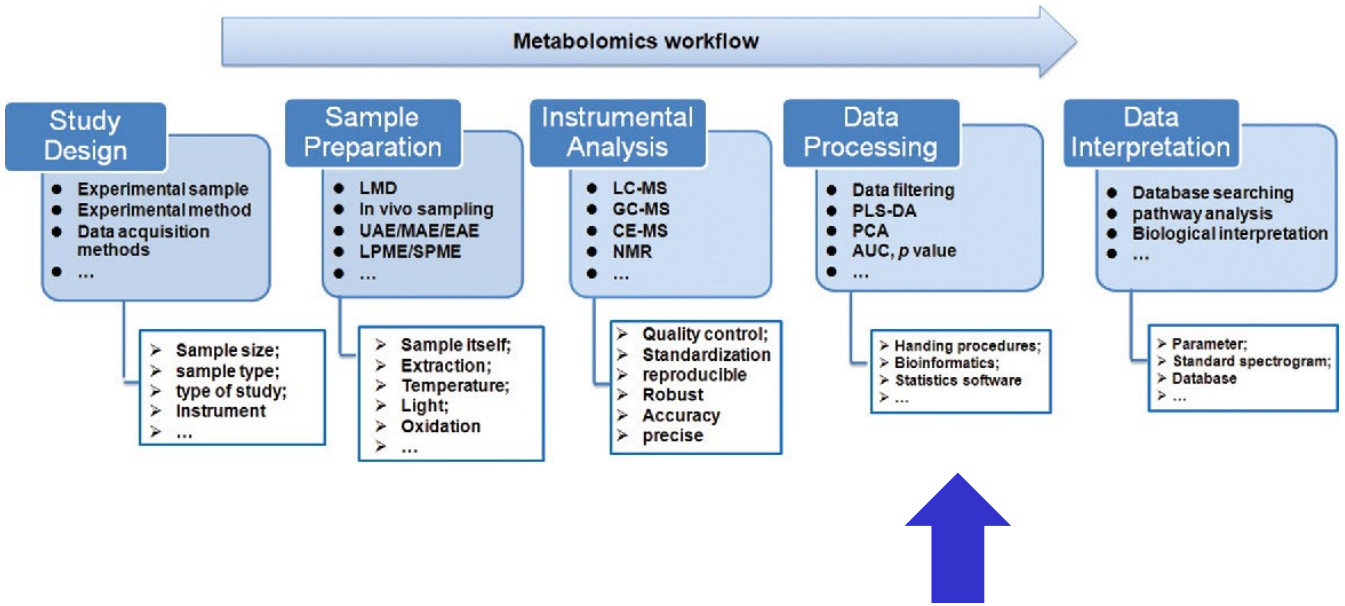
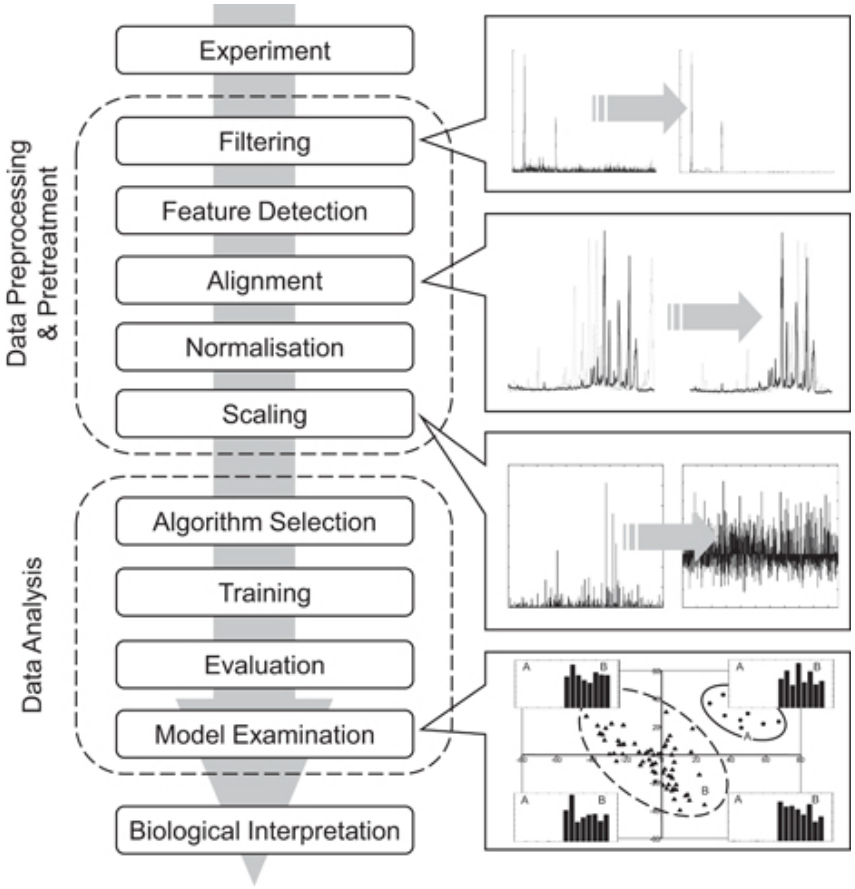


Figure 1. General workflows of metabolomics study, including study design, sample preparation, data acquirement and data mining.

Published in: Zhi-Gang Gong; Jing Hu; Xi Wu; Yong-Jiang Xu; *Critical Reviews in Analytical Chemistry* Ahead of Print
DOI: 10.1080/10408347.2017.1289836

Knowledge discovery in metabolomics: An overview of MS data handling

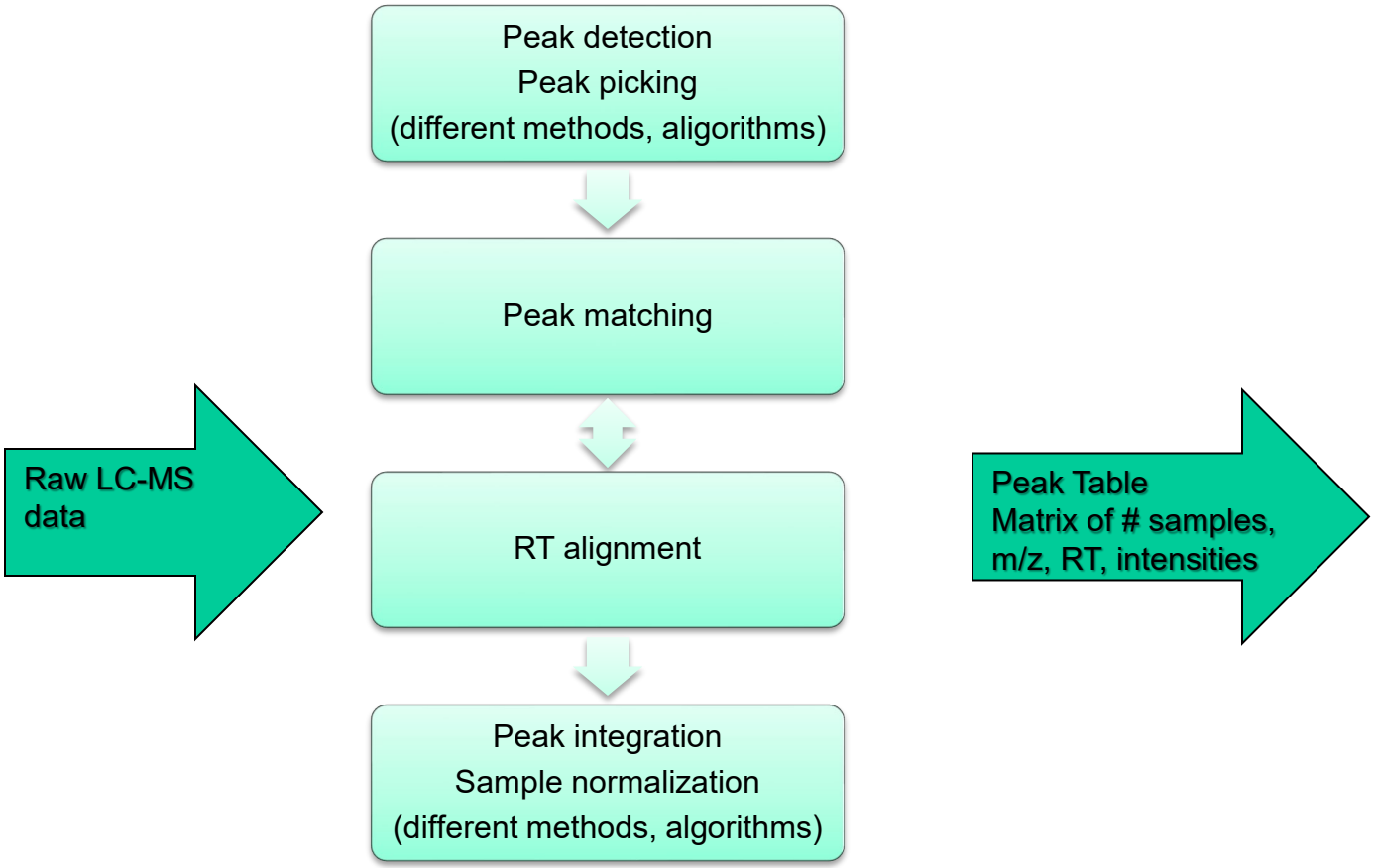


Journal of Separation Science
Volume 33, Issue 3, pages 290-304, 19 JAN 2010 DOI: 10.1002/jssc.200900609
http://onlinelibrary.wiley.com/doi/10.1002/jssc.200900609/full#fig4

Data Pre-Processing

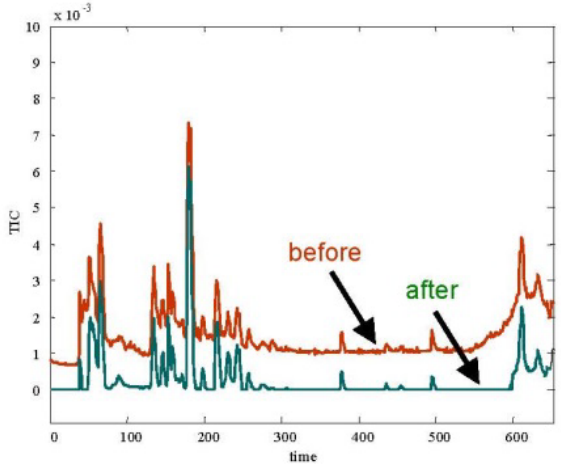
- Noise reduction
- RT correction
- Feature finding
- Integration
- Chromatogram alignment (between all samples)

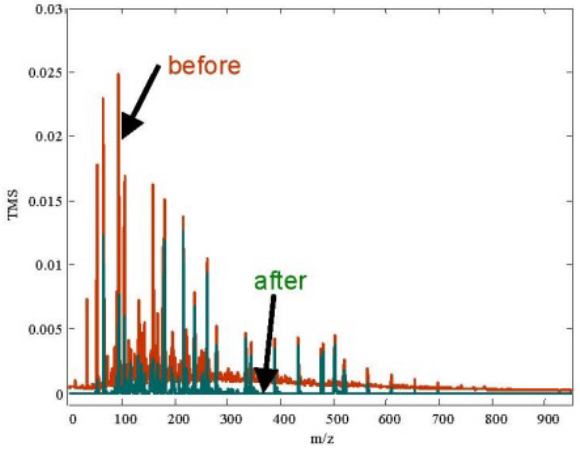
Data Pre-Processing



Enhancement of signal to noise ratio

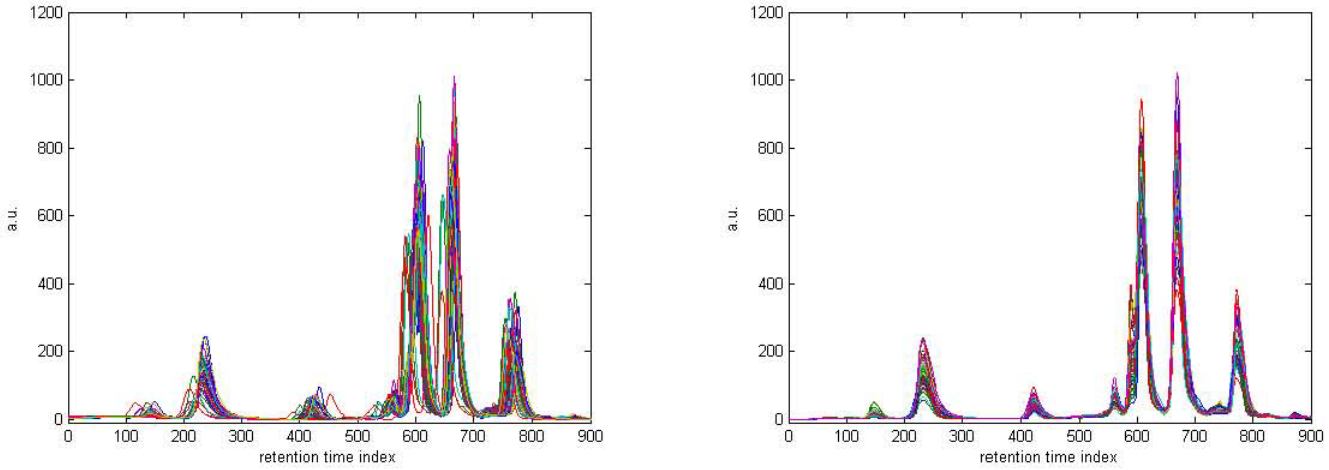
- elimination of noise (wavelets)
- background removal (e.g. PALS)

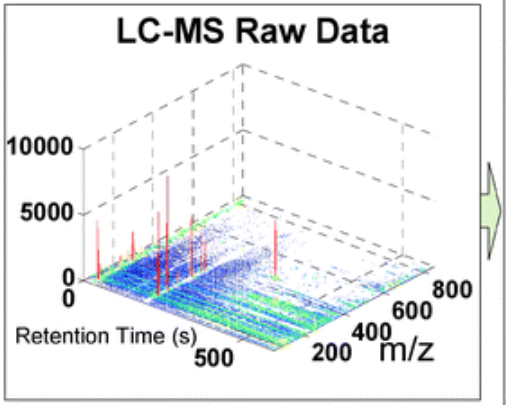


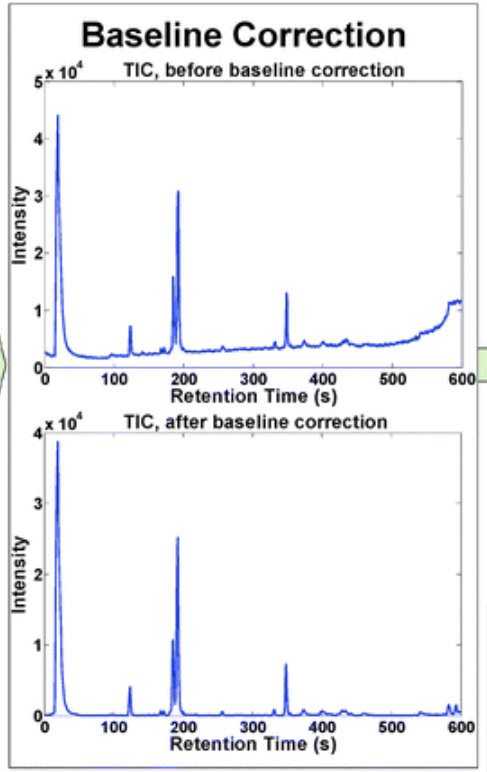


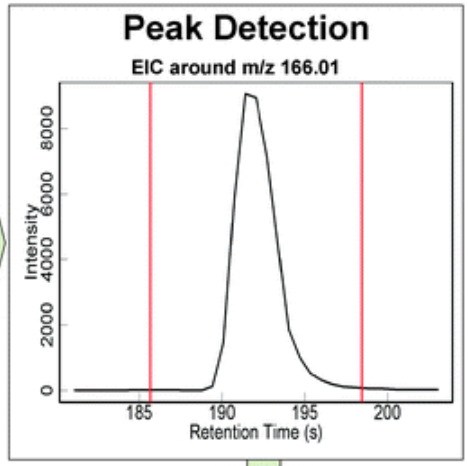
Synchronization of time axes

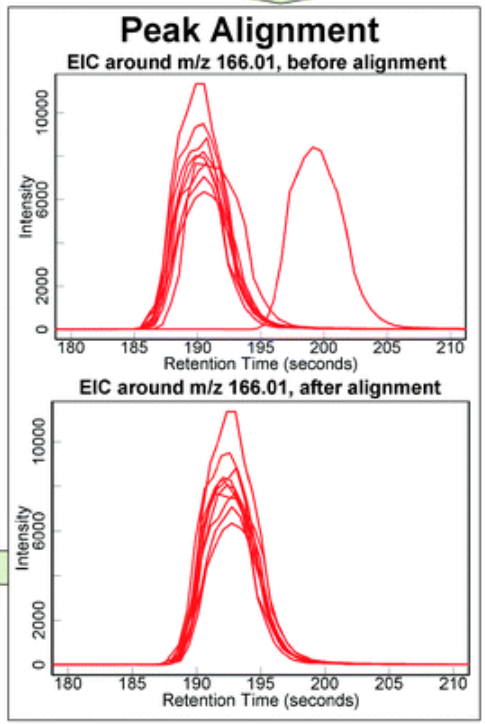
- HPLC chromatograms of green tea extracts











Ion Annotation

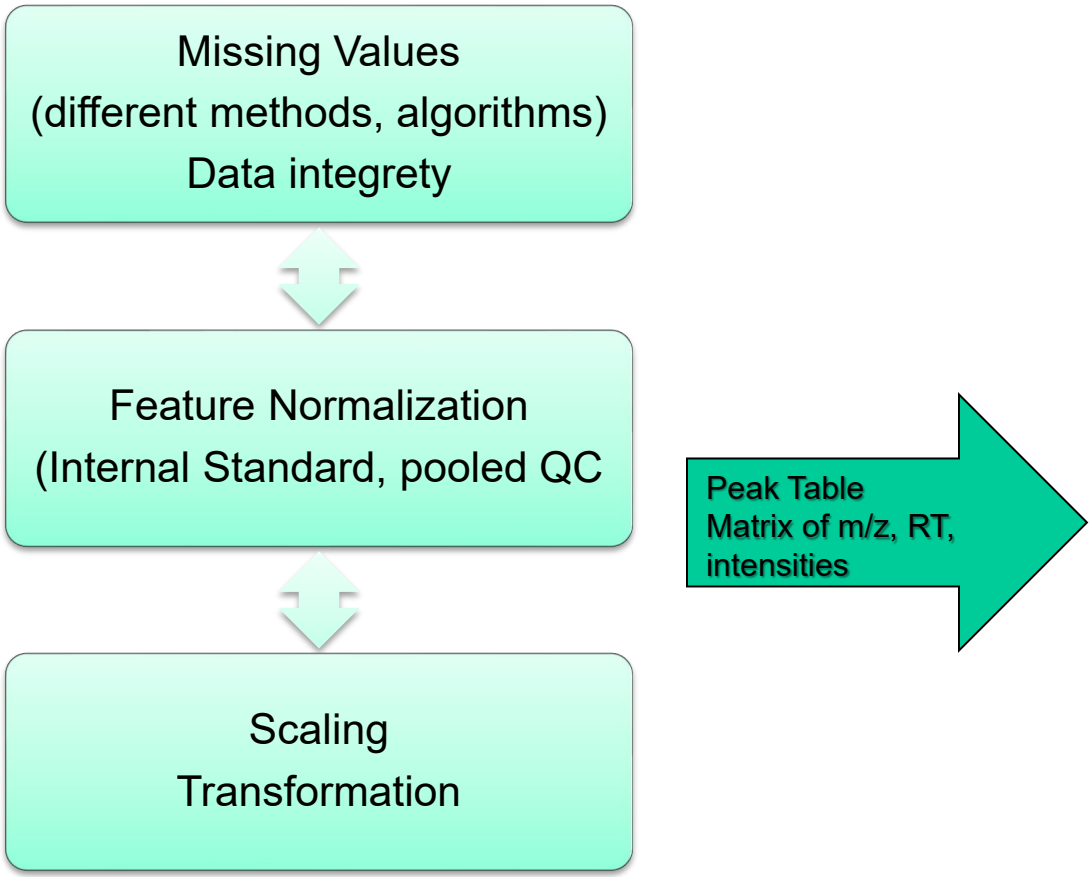
Annotated ions

m/z	Retention Time	Isotopes	Adducts
297.13	184.44	[10][M]+	[M-H]- 298.15
298.14	184.43	[10][M+1]+	
299.14	184.42	[10][M+2]+	
335.11	184.40		[M-2H+K]- 298.15
343.14	184.48	[11][M]+	
344.14	184.48	[11][M+1]+	
345.15	184.45	[11][M+2]+	

Peak list

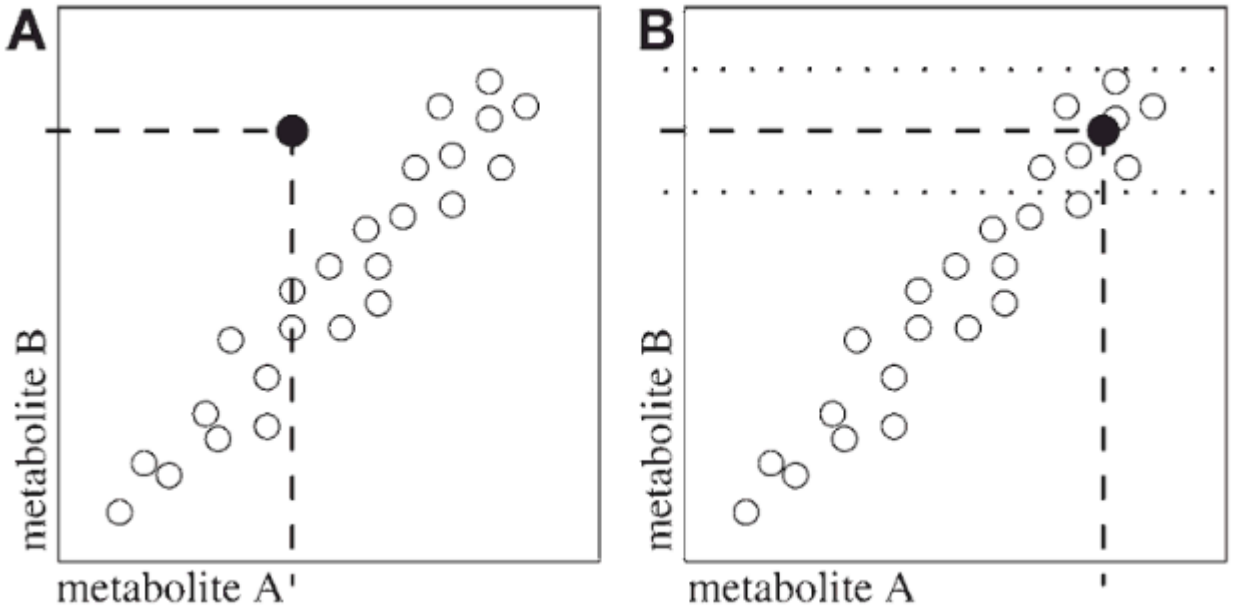
m/z	Retention Time	Sample 1	...	Sample N
297.13	184.44	10927.7		8409.6
298.14	184.43	2281.8		1453.5
299.14	184.42	521.0		312.0
335.11	184.40	308.7	...	296.3
343.14	184.48	7166.5		6056.8
344.14	184.48	1454.2		1200.9
345.15	184.45	303.6		244.1

Data Pre-Treatment



Treatment Missing Values

- Discard all
- Replace by average value (A)
- Replace by average of nearest neighbors (B)



Normalization

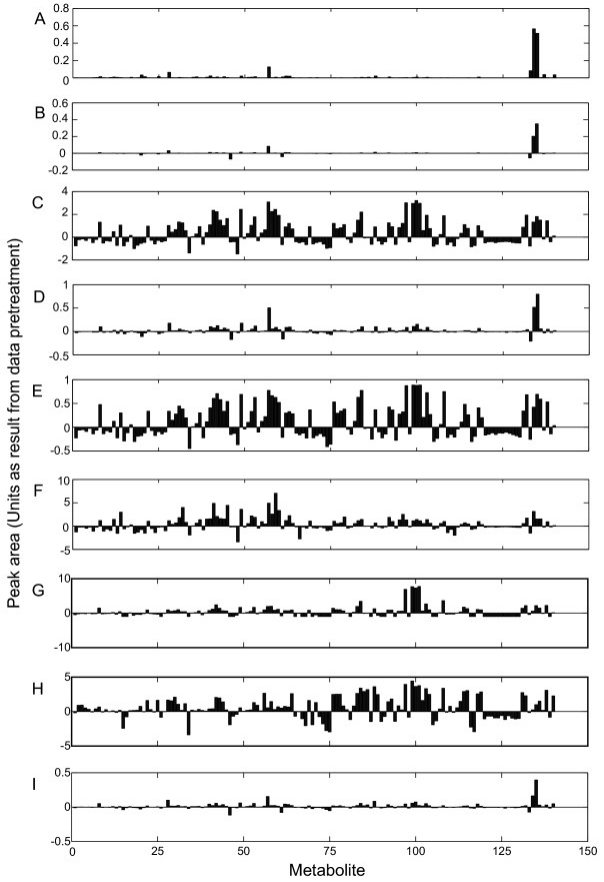
- Scalar
  - Assumes total signal per sample is =
- Internal Standard
  - Ratio response metabolite/ISTD
  - Optimal combination of ISTDs, endogenous
- Pooled QC based
  - Batch/analyte ratio (mean, median)

Centering, scaling, transformation

Class	Method	Formula	Unit	Goal	Advantages	Disadvantages
I	Centering	$\tilde{x}_{ij} = x_{ij} - \bar{x}_i$	O	Focus on the differences and not the similarities in the data	Remove the offset from the data	When data is heteroscedastic, the effect of this pretreatment method is not always sufficient
II	Autoscaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$	(-)	Compare metabolites based on correlations	All metabolites become equally important	Inflation of the measurement errors
	Range scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\left(x_{i_{max}} - x_{i_{min}}\right)}$	(-)	Compare metabolites relative to the biological response range	All metabolites become equally important. Scaling is related to biology	Inflation of the measurement errors and sensitive to outliers
	Pareto scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$	O	Reduce the relative importance of large values, but keep data structure partially intact	Stays closer to the original measurement than autoscaling	Sensitive to large fold changes
	Vast scaling	$\tilde{x}_{ij} = \frac{\left(x_{ij} - \bar{x}_i\right)}{s_i} \cdot \frac{\bar{x}_i}{s_i}$	(-)	Focus on the metabolites that show small fluctuations	Aims for robustness, can use prior group knowledge	Not suited for large induced variation without group structure
	Level scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\bar{x}_i}$	(-)	Focus on relative response	Suited for identification of e.g. biomarkers	Inflation of the measurement errors
III	Log transformation	$\tilde{x}_{ij} = {}^{10}\log\left(x_{ij}\right)$ $\tilde{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	Log O	Correct for heteroscedasticity, pseudo scaling. Make multiplicative models additive	Reduce heteroscedasticity, multiplicative effects become additive	Difficulties with values with large relative standard deviation and zeros
	Power transformation	$\tilde{x}_{ij} = \sqrt{\left(x_{ij}\right)}$ $\tilde{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	$\sqrt{O}$	Correct for heteroscedasticity, pseudo scaling	Reduce heteroscedasticity, no problems with small values	Choice for square root is arbitrary

Centering, Scaling and Transforming

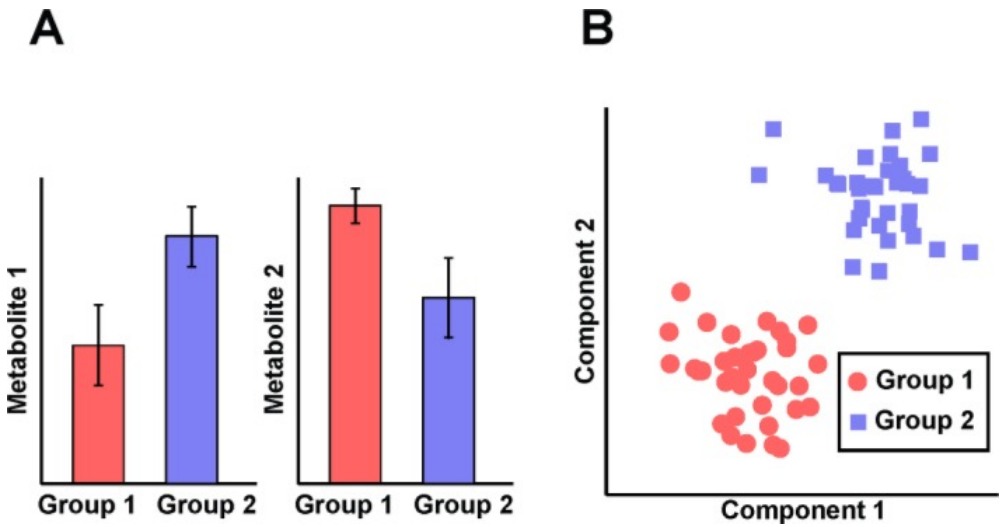
A. Raw
B. Centering
C. Auto
D. Pareto
E. Range
F. Vast
G. Level
H. Log
I. Power



All Feature-wise normalization methods

Statistical Analysis

- Univariate
  - Preliminary overview of data
  - T-test, ANOVA
  - Group-wise differences parameter by parameter
  - Reduces to analytes which show the strongest responses (+/-) to tested conditions
- Multivariate
  - In addition to differences from conditions also looks at relationships among analytes
  - PCA, PLS

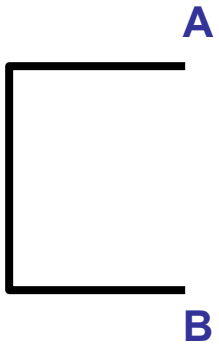


Comput Struct Biotechnol J. 2013; 4

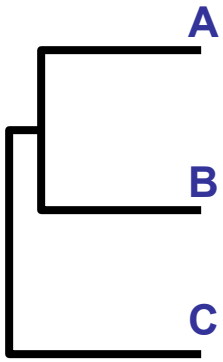
Statistical Analysis

- Clustering
  - Group and visualize samples according to intrinsic similarities in their measurements, irrespective of sample groupings
  - Multivariate
  - Hierarchical (HCA) or partial clustering algorithms (k-means

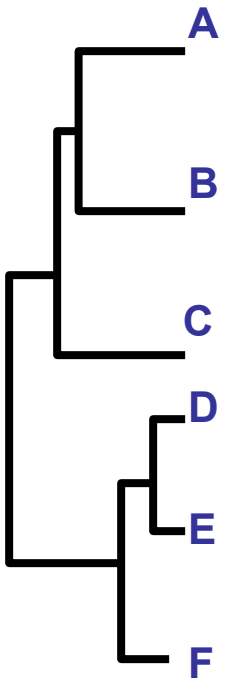
Hierarchical Clustering



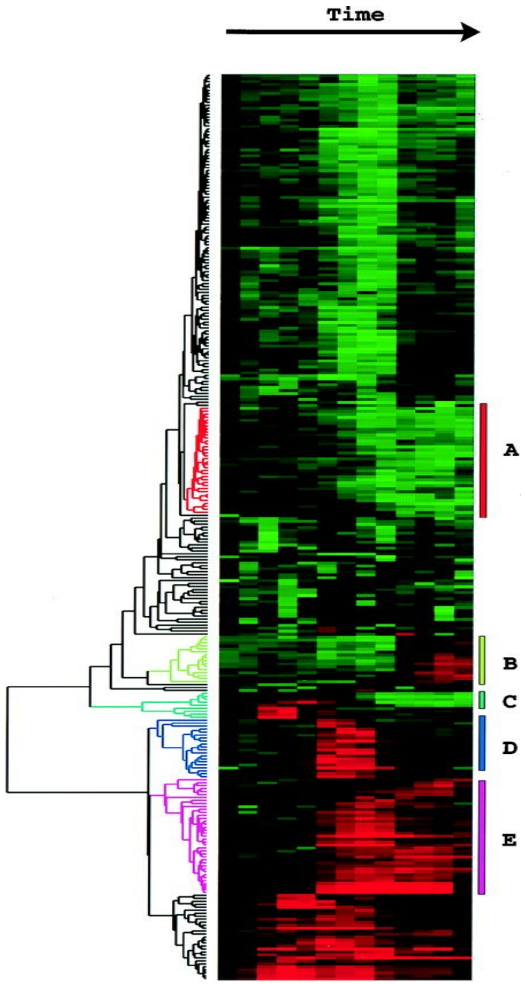
Find 2 most similar metabolite expression levels or curves



Find the next closest pair of levels or curves



Iterate



Heat map

Statistical Analysis

- Classification
  - Based on training set create model of data
  - Assigns a new sample to a category and its metabolites to known group or pathway
  - Need to validate model
    - ROC, Crossvalidation

Data Analysis Progression

- Unsupervised Methods
  - PCA or cluster to see if natural clusters form or if data separates well
  - Data is “unlabeled” (no prior knowledge)
- Supervised Methods/Machine Learning
  - Data is labeled (prior knowledge)
  - Used to see if data can be classified
  - Helps separate less obvious clusters or features
- Statistical Significance
  - Supervised methods always generate clusters -- this can be very misleading
  - Check if clusters are real by label permutation

Name	Main Application	Specific Features	License	User Interface
metaXCMS	Importing XCMS output	Post processing of XCMS for comparison of multiple (≥3) classes and visualizing statistical analyses.	Free	R language*) and GTK
XCMS2	Importing tandem mass spectrometry (MS/MS) raw data	Processing of tandem mass spectrometry data for metabolite identification and structural characterization	Free	Plug-in of R language*)
MetAlign	Importing many common formats, including Masslynx, Xcalibur, netCDF, and the old-style HP/Agilent format of GC-MS / LC-MS data	Interface-driven data processing program. Includes baseline correction, smoothing, feature detection and alignment	Free	Local application (GUI)
MAVEN	Data processing of LC-MS and pathway visualization	Tools for all aspects of data analysis, from feature extraction to pathway-based graphical data display	Free	Local application (GUI)
metaP-server	Statistical analysis, database searching, pathway visualization	A web-based metabolomics data analysis tool	Free	Web
MetDAT	Statistical analysis, database searching, pathway visualization	A modular and workflow-based free online pipeline for mass spectrometry data processing, analysis and interpretation	Free	Web
Pathway projector	Pathway visualization	A Web-based zoomable pathway browser that uses KEGG atlas and Google Maps API	Free	Web
		A web-based		

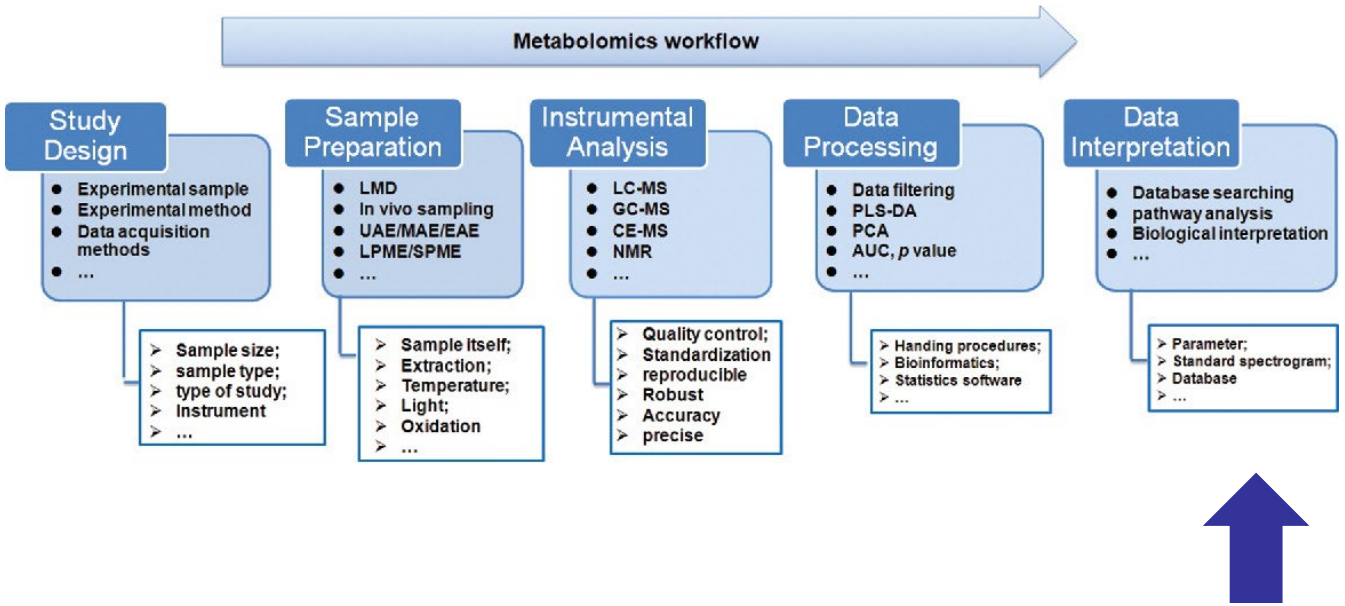


Figure 1. General workflows of metabolomics study, including study design, sample preparation, data acquirement and data mining.

Published in: Zhi-Gang Gong; Jing Hu; Xi Wu; Yong-Jiang Xu; *Critical Reviews in Analytical Chemistry* Ahead of Print
DOI: 10.1080/10408347.2017.1289836

Data Interpretation

- Identification of Significant Metabolites
- 4 levels of ID
  - Positively identified compounds
    - Confirmed by match to known standard
  - Putatively identified compounds
    - Match to MS + RT or MS/MS + RT
  - Compounds putatively identified in a compound class
  - Unknowns

Databases for Metabolomics

- NMR spectral databases
  - Primarily small molecule spectra, not all metabolites
- MS or MS/MS spectral databases
  - Primarily small molecule spectra, not all metabolites
- Compound databases
  - Mostly compound names, structures, IDs, physprops
- Pathway databases
  - Mix of metabolite, drug, protein, signaling pathways
- Comprehensive metabolomic databases
  - Combines most/all of the above, focus on metabolites

Databases for Metabolomics

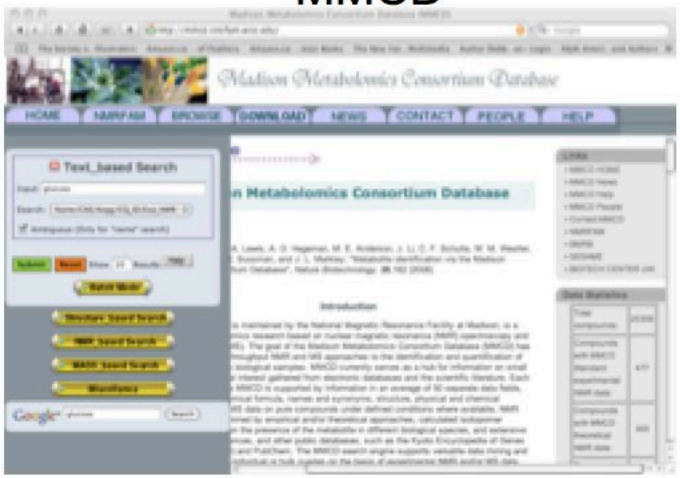
- NMR spectral databases
  - Primarily small molecule spectra, not all metabolites
- MS or MS/MS spectral databases
  - Primarily small molecule spectra, not all metabolites
- Compound databases
  - Mostly compound names, structures, IDs, physprops
- Pathway databases
  - Mix of metabolite, drug, protein, signaling pathways
- Comprehensive metabolomic databases
  - Combines most/all of the above, focus on metabolites

NMR Spectral DBs

SBDS

NMRShiftDB

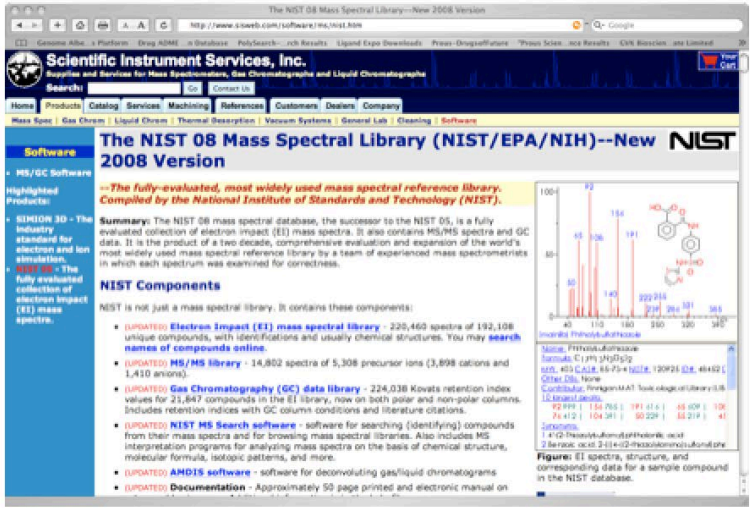
MMCD



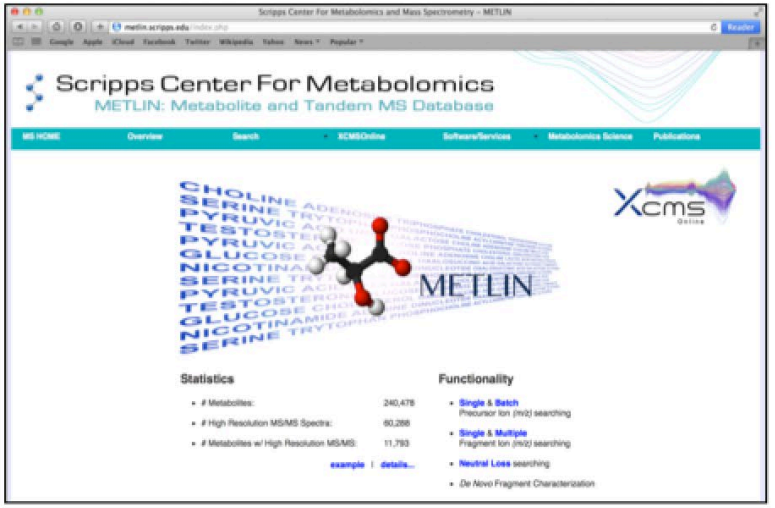
BMRB

MS Spectral DBs

NIST/AMDIS

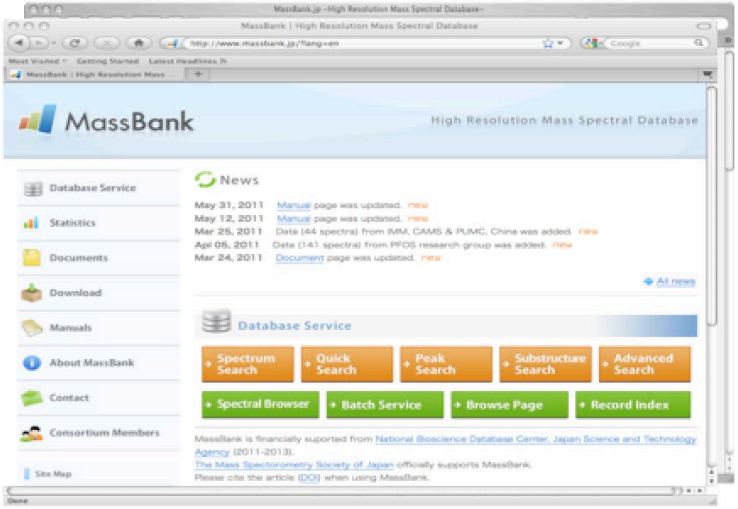


Metlin



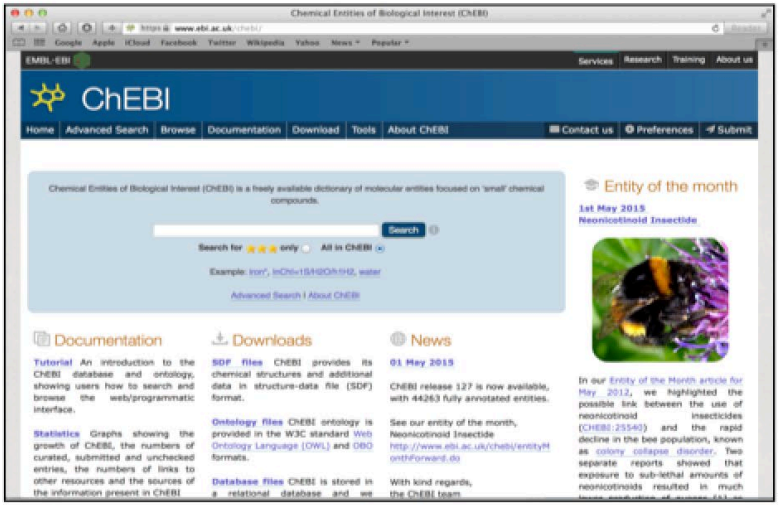
GolmDB

MassBank



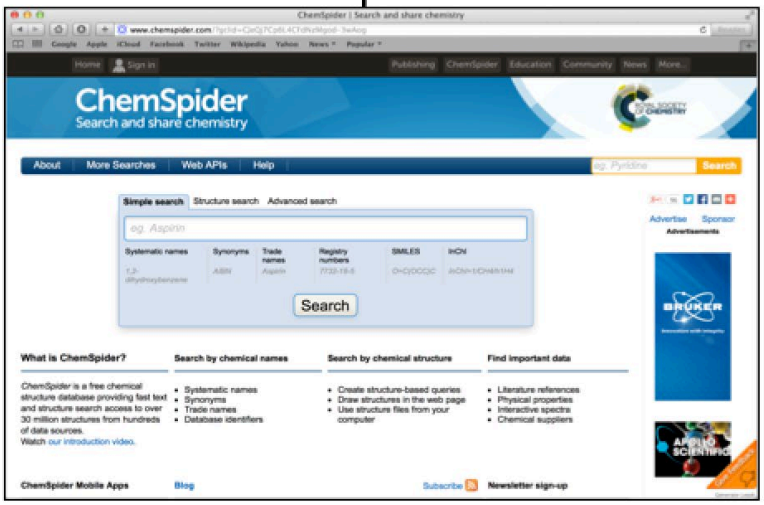
Compound DBs

ChEBI

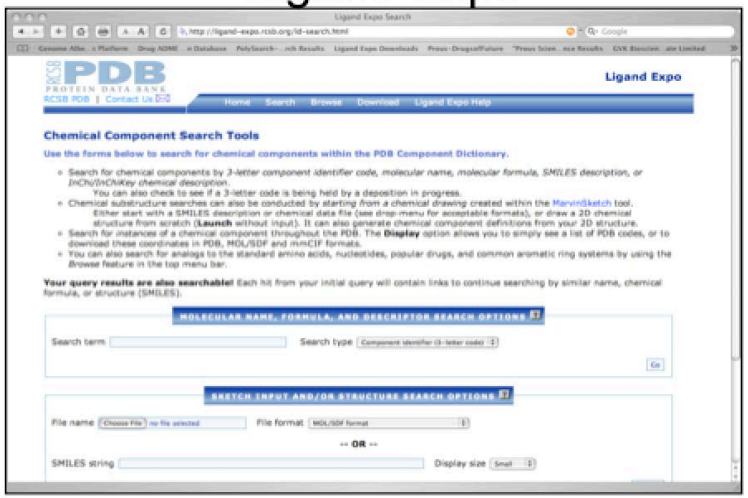


PubChem

ChemSpider

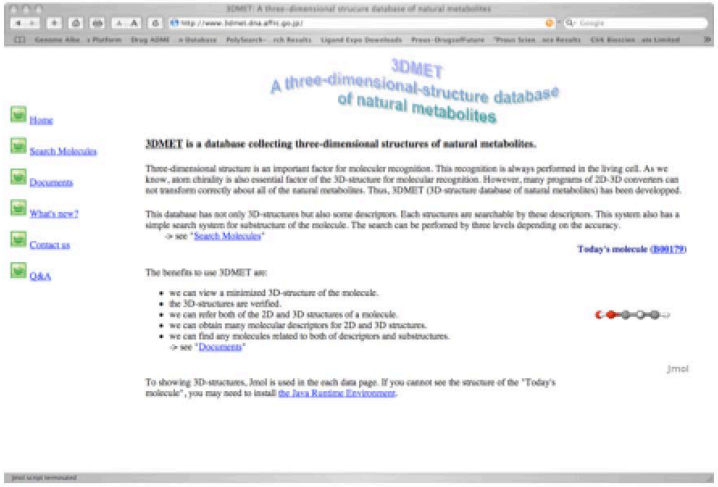


Ligand Expo

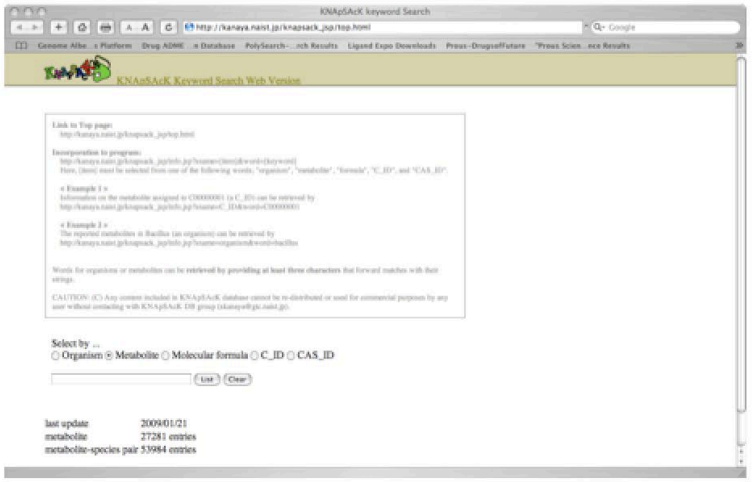


Other Compound DBs

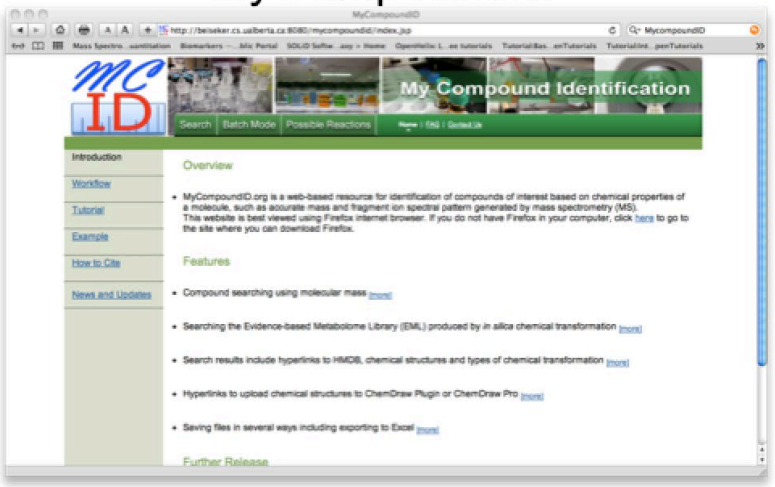
3DMet



KNApSAcK



MyCompoundID



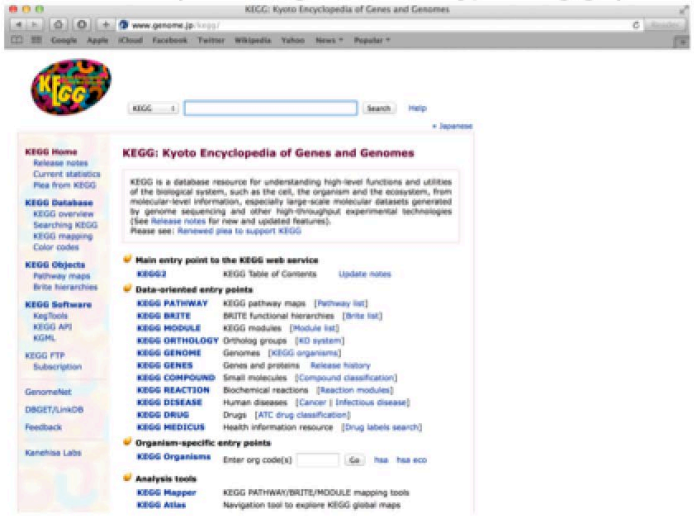
LipidMaps

Other DBs

- 3DMet
  - 3D structure database of 8581 natural metabolites
- KNApSAcK
  - Database of 50,000 plant metabolites linked to species information
- MyCompoundID
  - Database of 11 million metabolically transformed metabolites
- LipidMaps
  - Database with 30,000 lipids (Fatty acyls, glycerolipids, glycerophospholipids, sphingolipids, sterols, prenols, saccharolipids, polyketides)

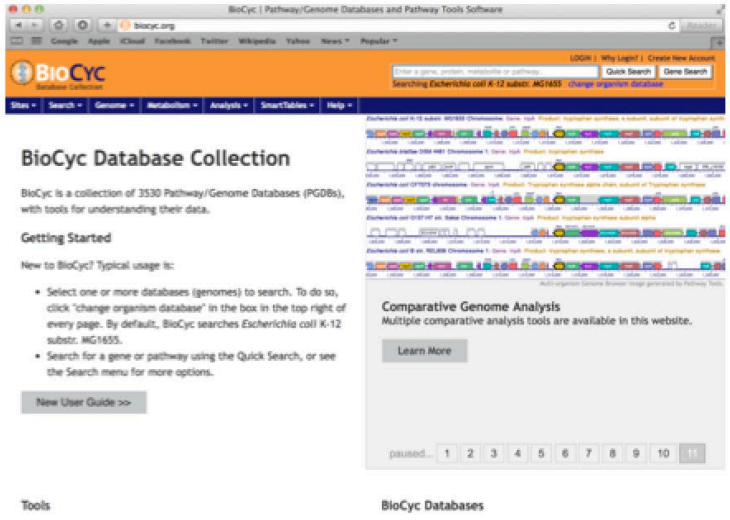
Pathway DBs

KEGG (www.genome.jp/kegg/)



SMPDB (www.smpdb.ca)

BioCyc/MetaCyc



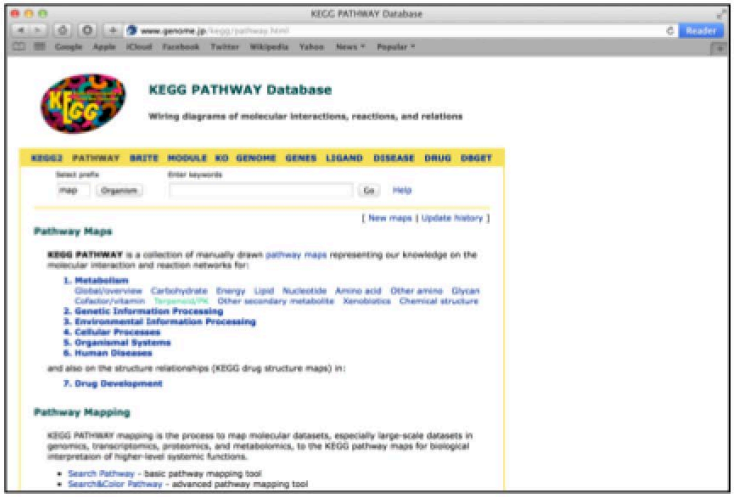
Reactome (www.reactome.org)

Pathway DBs

- Rich source of biological data that relates metabolites to genes, proteins, diseases, signaling events and processes
- Provide various tools to permit visualization and gene/metabolite mapping
- Often cover multiple species

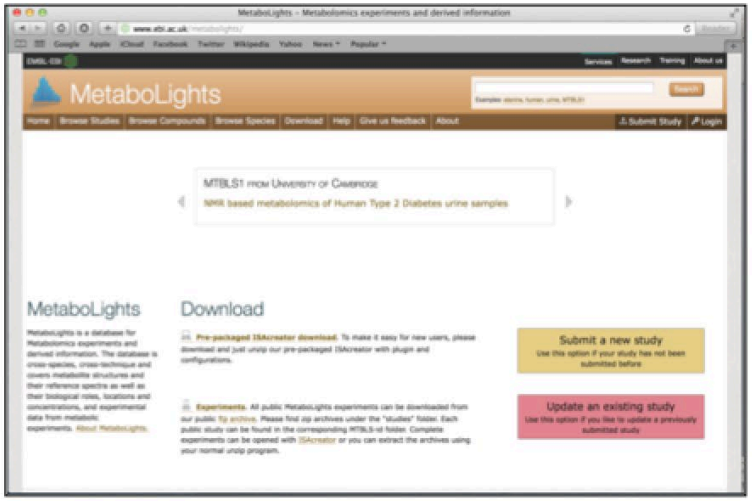
Comprehensive MetDBs

KEGG

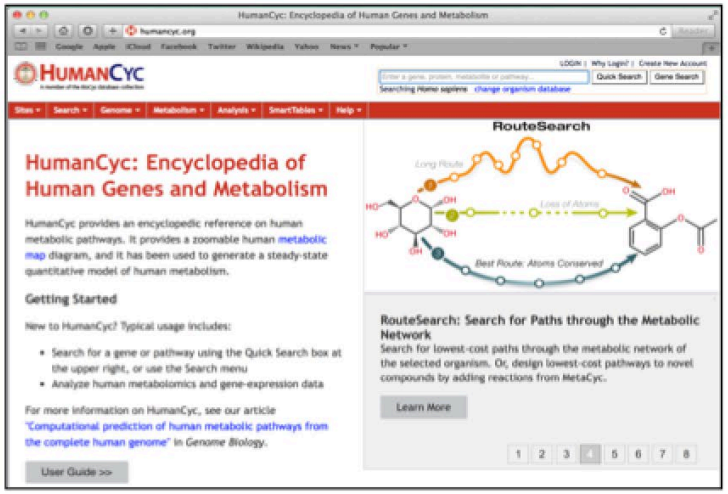


EcoCyc

MetaboLights



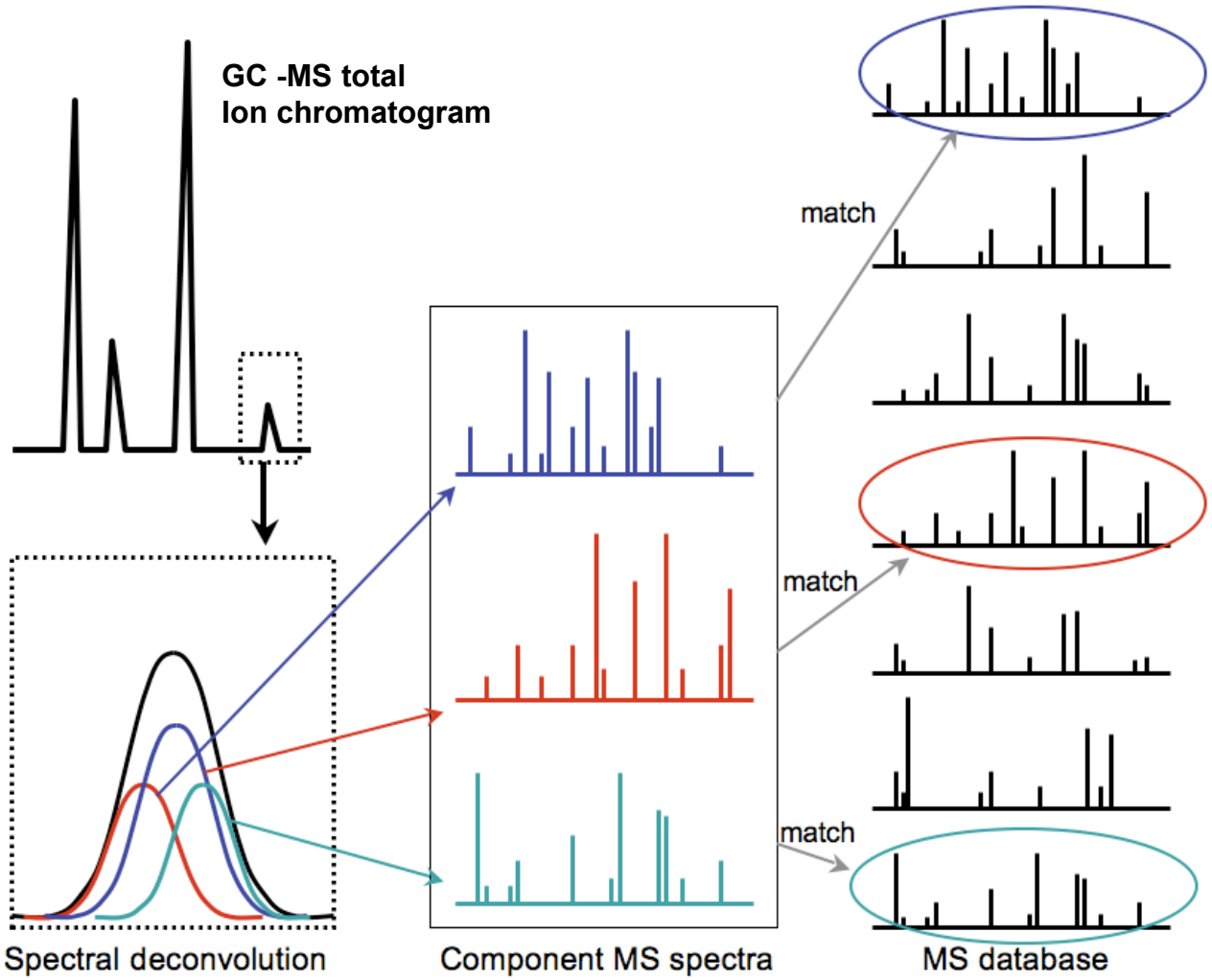
HumanCyc



Comprehensive MetDBs

- Must contain >1000 metabolites
- Usually are organism specific
- Continuously updated
- Must contain
  - chemical + pathway data
  - or chemical + spectral + biological data
  - or chemical + pathway + spectral data
  - or chemical + pathway + spectral + biological data

Metabolite ID by GC-MS

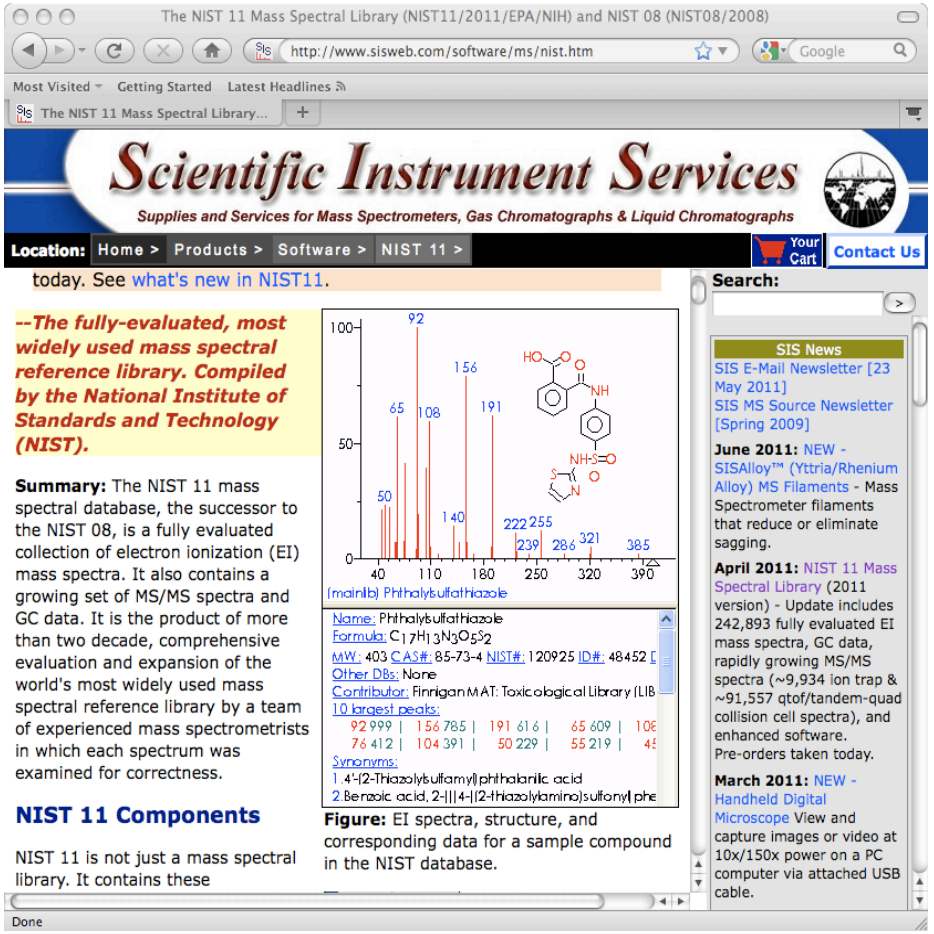


Metabolite ID by GC-MS

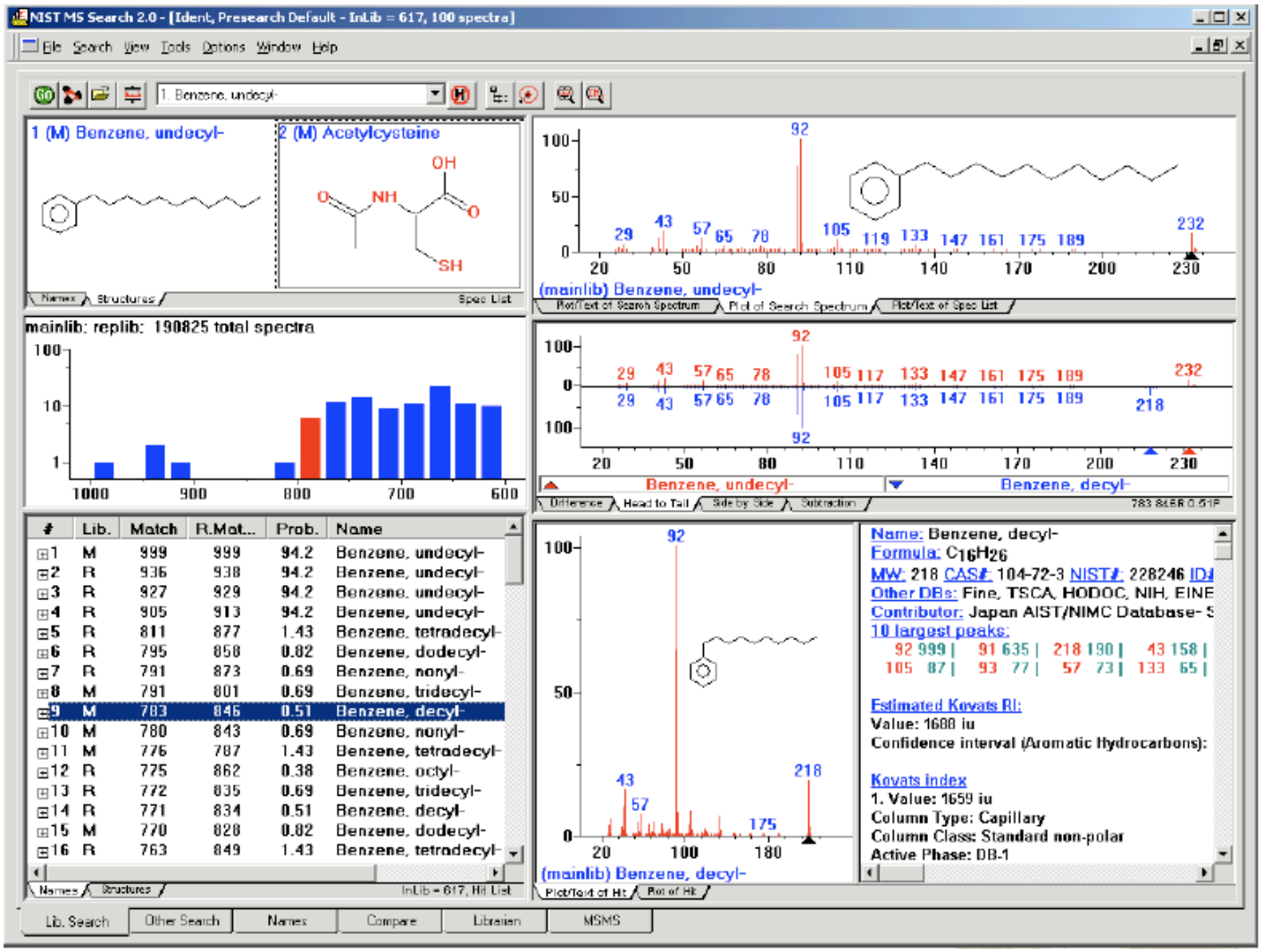
- GC-MS is often best for identification of amino acids, organic acids, sugars, fatty acids and molecules with MW<500
- GC has higher resolution and better reproducibility than LC
- EI-MS is more standardized than soft ionization methods, so EI spectra are more comparable
- Most common route is to use AMDIS + NIST database

NIST 11 MS Database

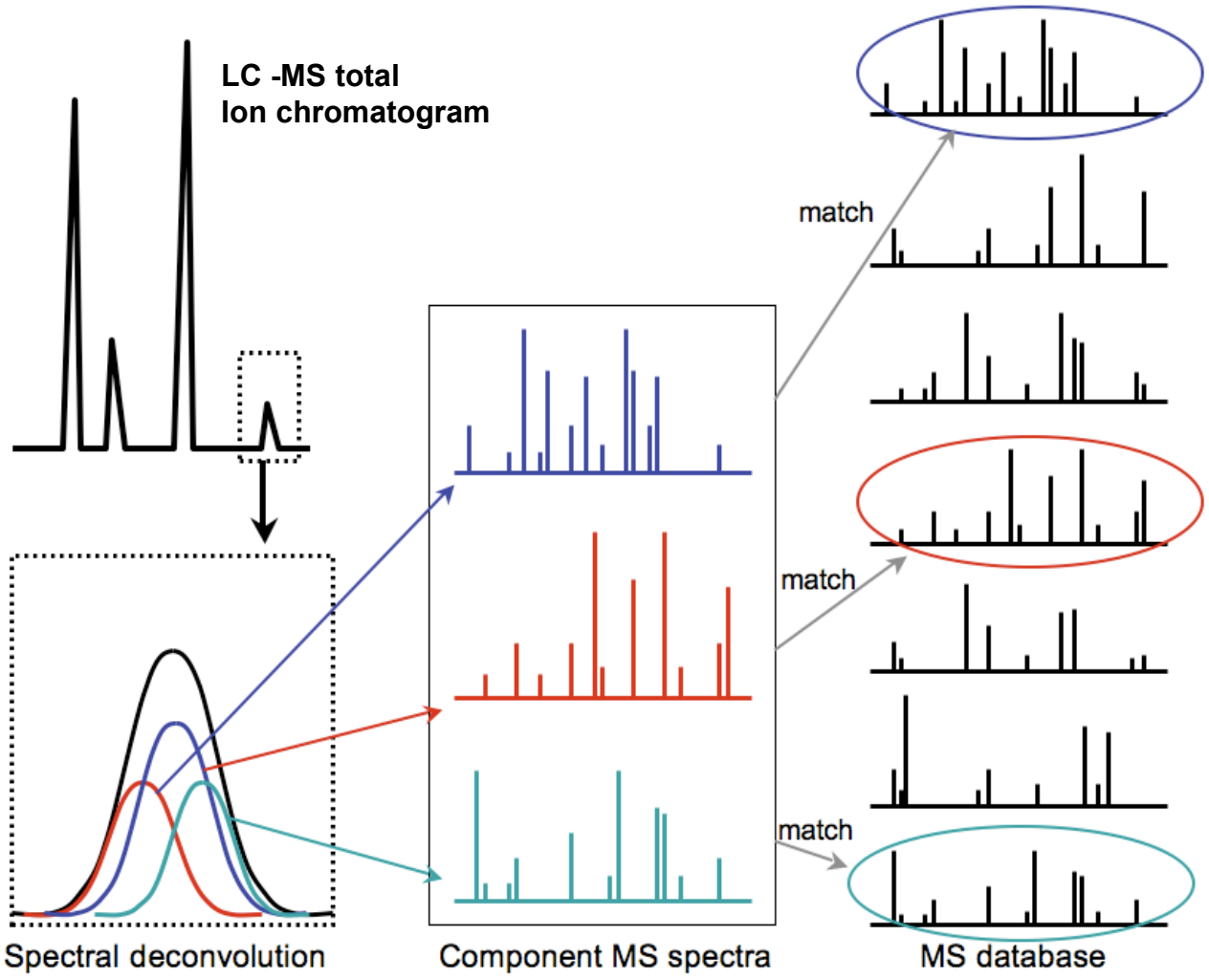
- 243,893 EI spectra of 212,961 cmpds
- 9934 ion trap MS for 4649 cmpds
- 91,557 Qtof & QqQ spectra for 3774 compounds
- 224,038 RI values for 21,847 cmpds



NIST MS Search Software



Metabolite ID by LC-MS

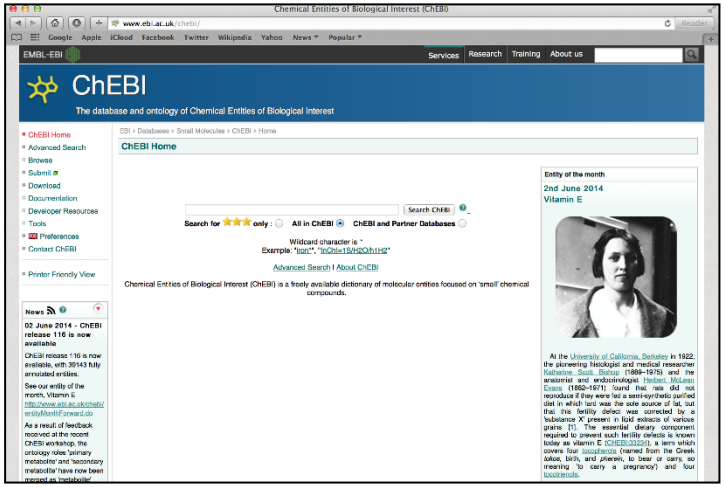


Metabolite ID by LC-MS

- LC-MS is often best for identification of lipids, bases, amino acids, organic acids, fatty acids and other somewhat hydrophobic molecules
- Metabolite ID typically requires both MS and MS/MS data (along with retention time information) and internal standards
- Compound ID can be done by high accuracy mass matching and/or by MS/MS matching to spectral databases

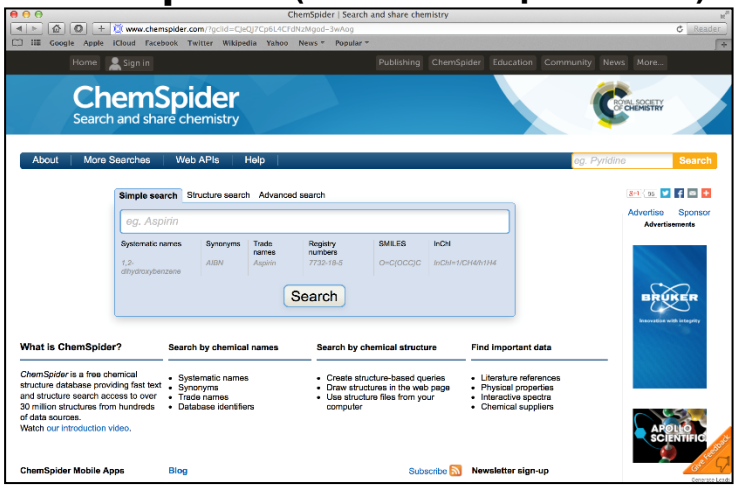
Simple MW Search DBs

ChEBI (www.ebi.ac.uk/chebi/)

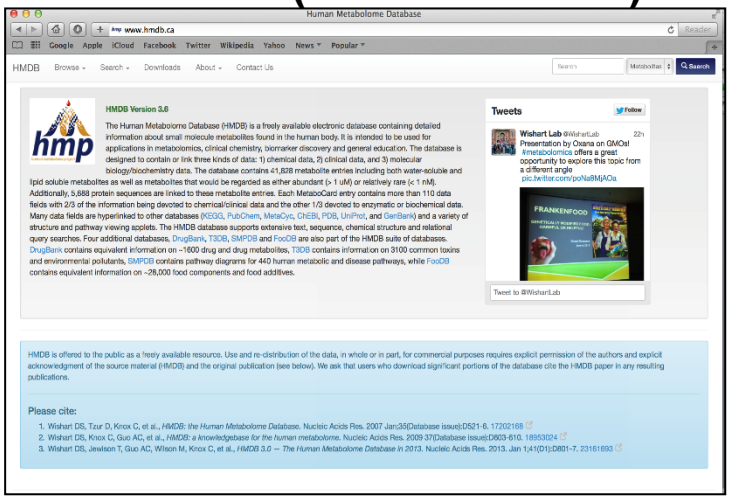


PubChem (http://pubchem.ncbi.nlm.nih.gov/)

ChemSpider (www.chemspider.com)



HMDB (www.hmdb.ca)

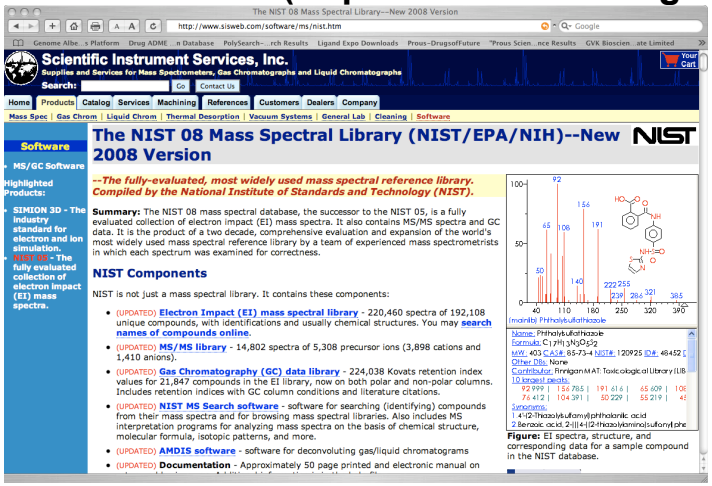


Advanced MS Search DBs

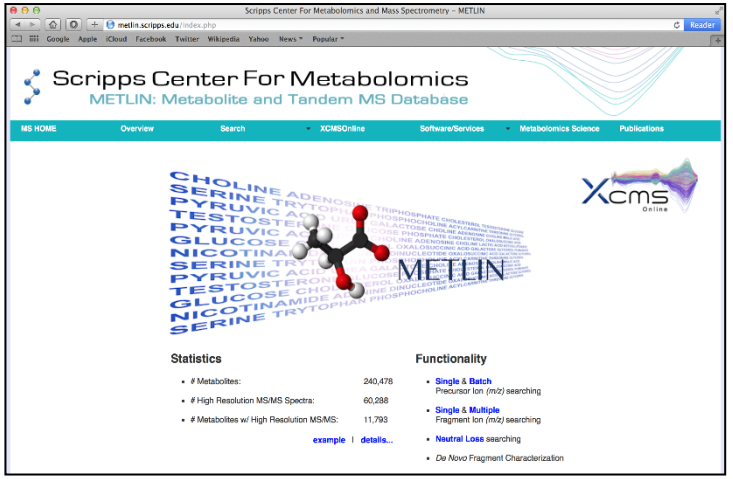
- These databases support not only MW or MW range searches, but also support parent ion searches (positive, negative, neutral), peak list searches (from MS or MS/MS data) as well as MS/MS spectral matching
- These DBs are intended more for MS-based metabolomics and compound ID than the simple MW search tools

Advanced MS Search DBs

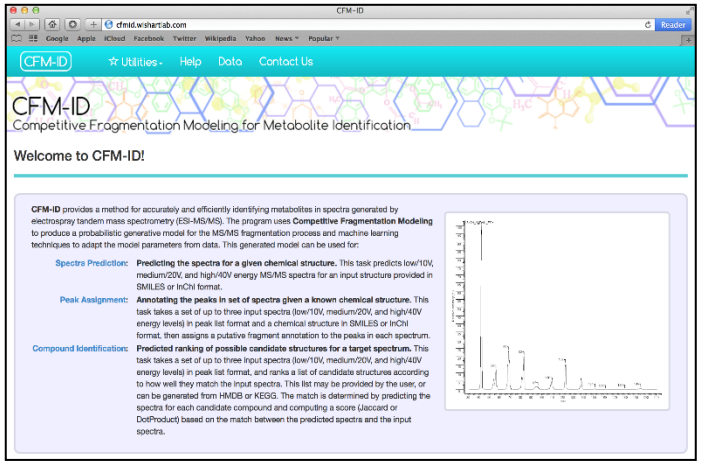
NIST/AMDIS (http://chemdata.nist.gov)



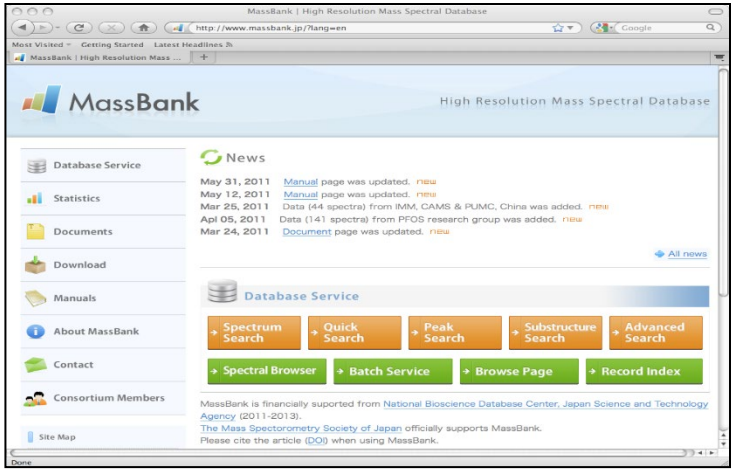
Metlin (http://metlin.scripps.edu/)



CFM-ID (http://cfmid.wishartlab.com)



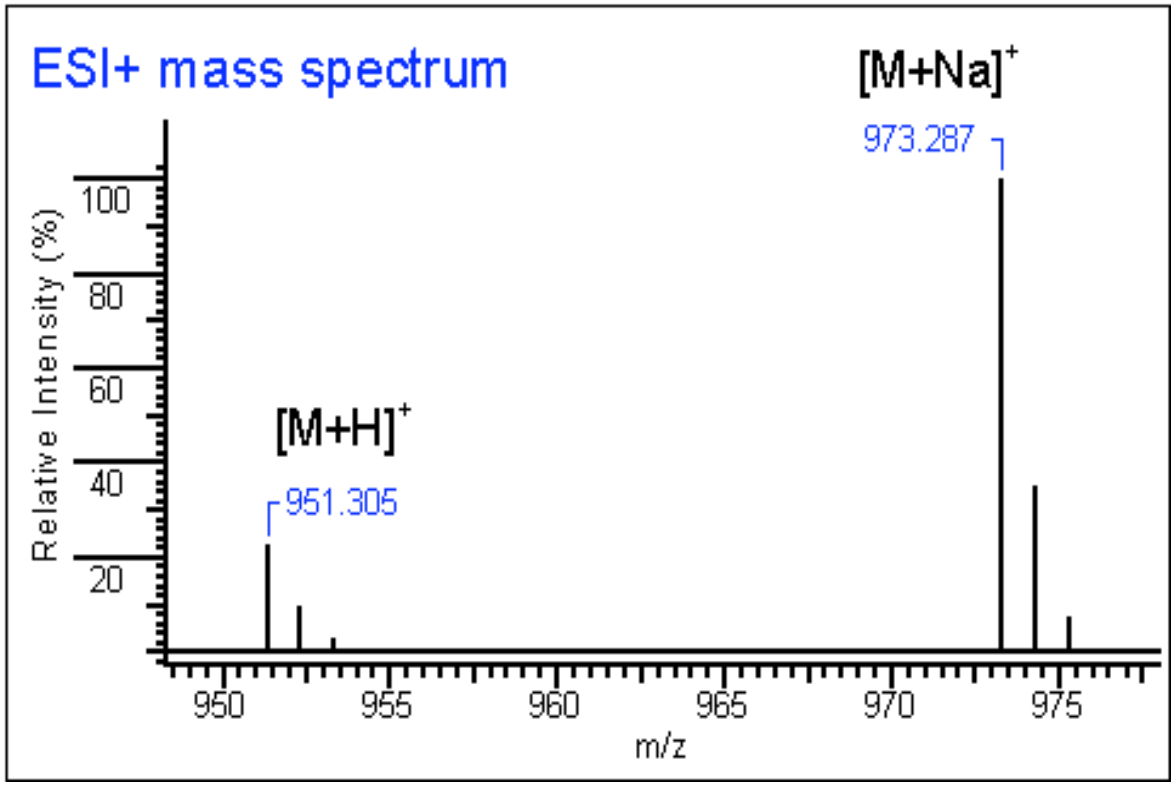
MassBank (www.massbank.jp)



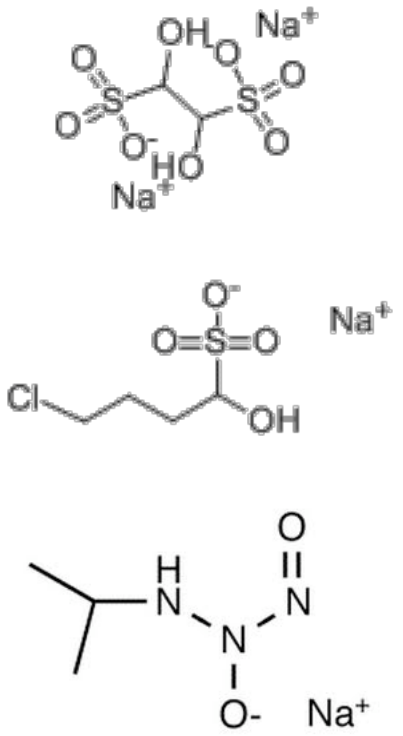
Metabolite ID by MS alone - Complications

- LC-ESI-MS often leads to the production of salt adducts, neutral loss species and multiply charged species
- Up to 80% of LC-MS signals arise from these “noise” sources
- Key challenge is to distinguish adducts or multiply charged species from parent ions or to group adducts or multiply charged species with parent ions

Adduct Formation



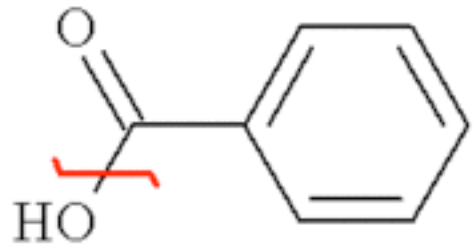
Effect on ESI Mass Spectrum



Sample Na Adducts

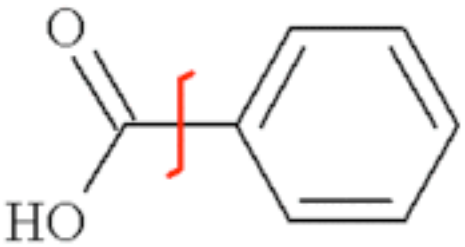
Neutral Loss Fragments

benzoic acid
Nominal Mass = 122 Da



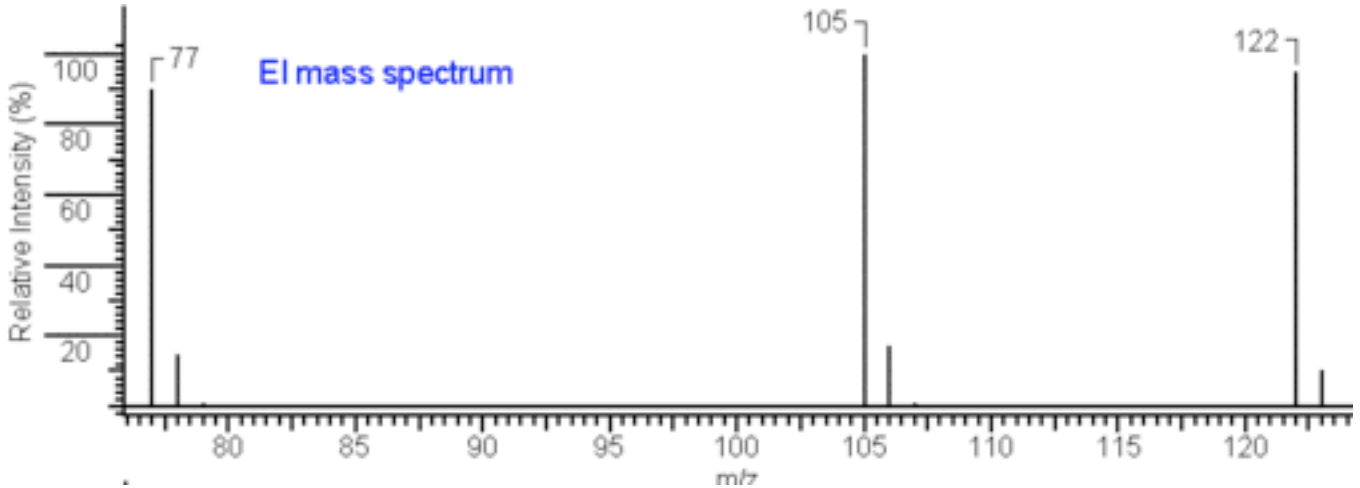
HO
17 Da

C_7H_5O
105 Da



CHO_2
45 Da

C_6H_5
77 Da



Handling MS Complications

- MZedDB, Metlin and HMDB are able to handle or predict adducts
- Metlin and MZedDB are able to handle or predict ion pairs or multiply charged species
- Metlin can potentially handle or predict neutral loss species
- Searching by MS or MS ranges can lead to lots of hits (high FP rate)

Resolving LC-MS Complications

- Identify, remove (or consolidate) adducts and multiply charged species
- Identify, remove (or consolidate) fragments (neutral losses, breakdown products, rearrangements)
- Identify, remove (or consolidate) isotope peaks
- Remove noise peaks (from sample blanks or peaks that do not appear in >2/3 technical replicates or peaks that do not show dilution trends in 4 dilution replicates)

Resolving LC-MS Complications

- Raw +ve mode spectrum
- Remove adducts
- Remove multiple charges
- Remove neutral losses
- Remove isotope peaks
- Remove noise peaks
- Final spectrum
- Repeat for –ve mode

- 15,000 features
- 12,000 features
- 10,000 features
- 8,000 features
- 3,000 features
- 2,500 features
- 2,500 M+H peaks
- 1,500 M-H peaks

Mzmine **MetFusion** **MagMA**

Exploiting High Mass Accuracy to ID Compounds

Type	Mass Accuracy
FT-ICR-MS	0.1 - 1 ppm
Orbitrap	0.5 - 1 ppm
Magnetic Sector	1 - 2 ppm
TOF-MS	3 - 5 ppm
Q-TOF	3 - 5 ppm
Triple Quad	3 - 5 ppm
Linear IonTrap	50-200 ppm (10 ppm in Ultra-Zoom)

$$ppm = \left(\frac{m_{exp} - m_{calc}}{m_{exp}}\right) * 1E+6$$

Formula Filters

- Use additional MS information (isotopic abundance) as well as chemical bonding restrictions (Lewis & Senior rules), known or presumed atomic compositional data and matches to known or hypothesized structures to reduce the possible # of structures/formulas that are generated

Impact of Mass Accuracy on Formula Numbers

Table 3

Number of possible molecular formulas at different levels of mass accuracy and the impact of isotopic abundance accuracy. A mass spectrometer capable of 3 ppm but with 2% correct isotopic pattern outperforms even a (non-existing) mass spectrometer with 0.1 ppm mass accuracy! The results are computed for randomly selected targets, so single results vary but the trend remains. LEWIS and SENIOR check was applied. Candidates with unrelated high element counts were already excluded

	without isotope abundance information					2% isotopic abundance accuracy	5% isotopic abundance accuracy
molecular mass [Da]	10 ppm	5 ppm	3 ppm	1 ppm	0.1 ppm	3 ppm	5 ppm
150	2	1	1	1	1	1	1
200	3	2	2	1	1	1	1
300	24	11	7	2	1	1	6
400	78	37	23	7	1	2	13
500	266	115	64	21	2	3	33
600	505	257	155	50	5	4	36
700	1046	538	321	108	10	10	97
800	1964	973	599	200	20	13	111
900	3447	1712	1045	345	32	18	196

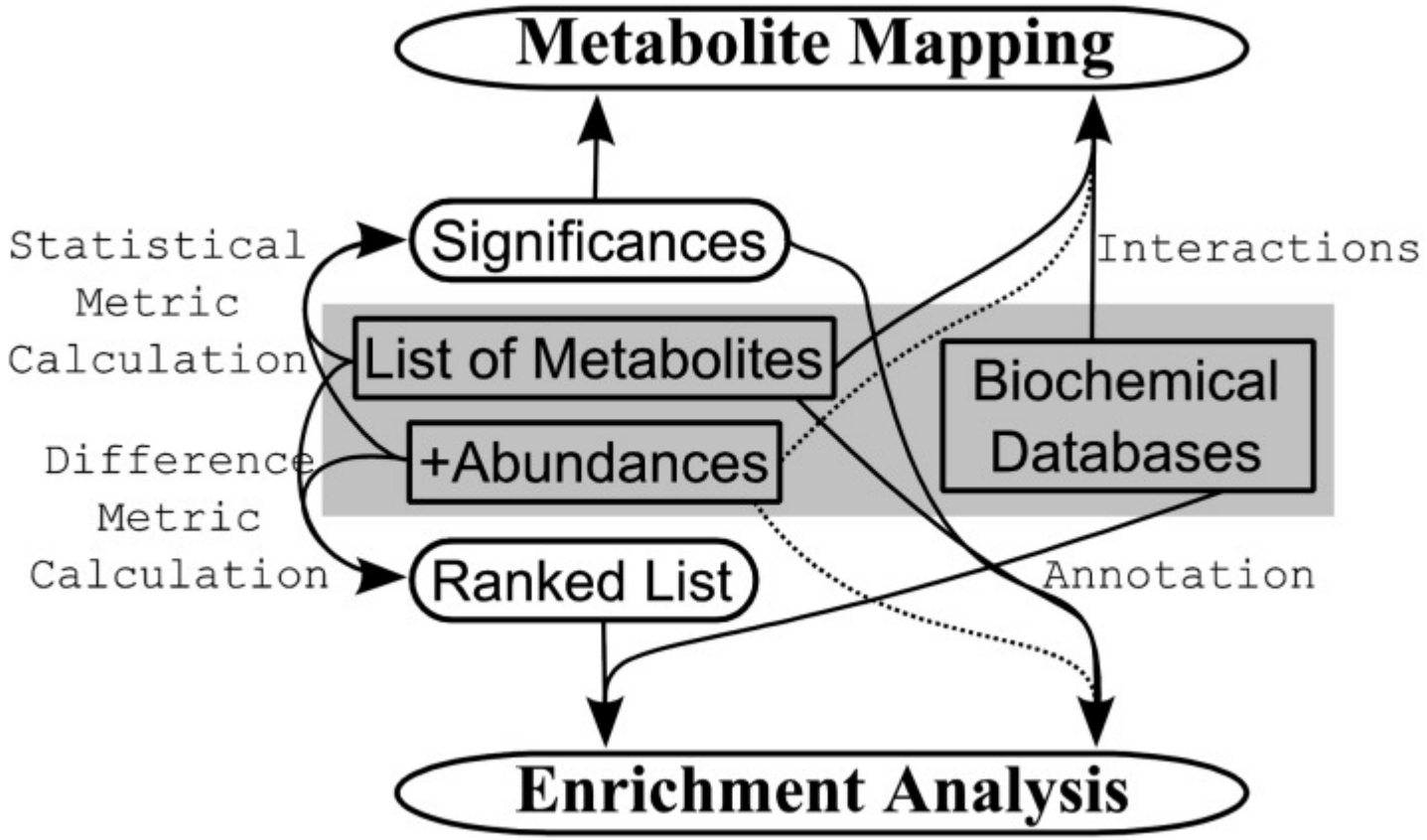
Kind and Fiehn *BMC Bioinformatics* 2006 **7**:234 doi:10.1186/1471-2105-7-234

Some Points of Caution

- Many databases (PubChem, ChEBI, Metlin, FiehnLib, NIST) mix non-metabolites with metabolites or plant metabolites with animal and/or microbial metabolites or drugs/buffer reagents with metabolites
- This leads to many “silly” hits
- If you know the source organism use this information to limit the search or use organism-specific metabolome databases (HMDB, FooDB, DrugBank, KnapSack, etc.)

Pathway Analysis

Putting metabolomics data into a biological context

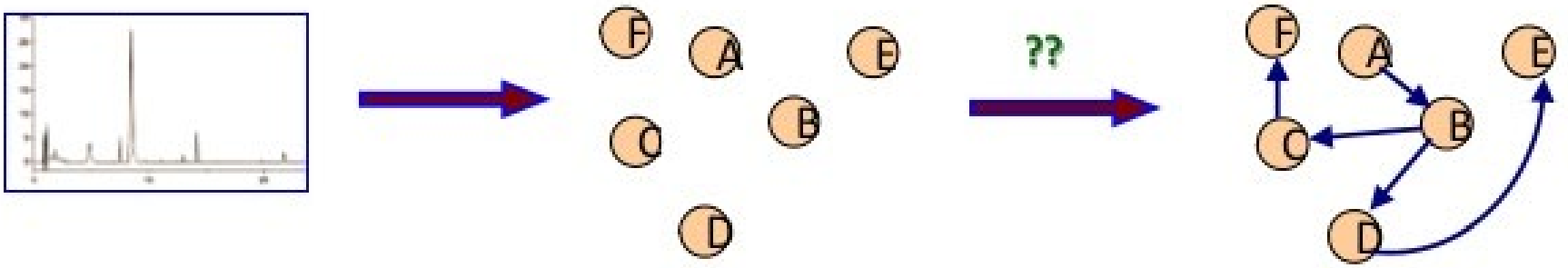


Biological Databases for Pathway Analysis

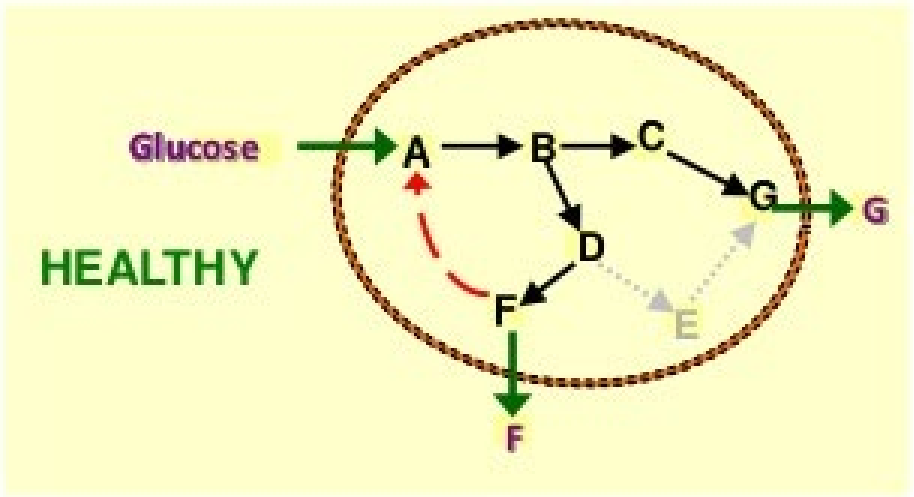
Database	Description	Website	Reference
Kyoto Encyclopedia of Genes and Genomes (KEGG)	466 pathways, 17,333 metabolites, and 9,764 biochemical reactions	http://www.genome.jp/kegg/	Kanehisa et al. (2012)
MetaCyc	2260 pathways from 2600 different organisms	http://metacyc.org/	Caspi et al. (2008)
The small molecule pathway database (SMPDB)	1,594 metabolites mapping 727 small molecule pathways found in humans	http://www.smpdb.ca/	Jewison et al. (2014)
WikiPathways	1,910 pathways	http://wikipathways.org/	Kelder et al. (2012)
Plant metabolic network (PMN/PlantCyc)	Multi-species pathway database for plant metabolomics	http://www.plantcyc.org/	Chae et al. (2014)

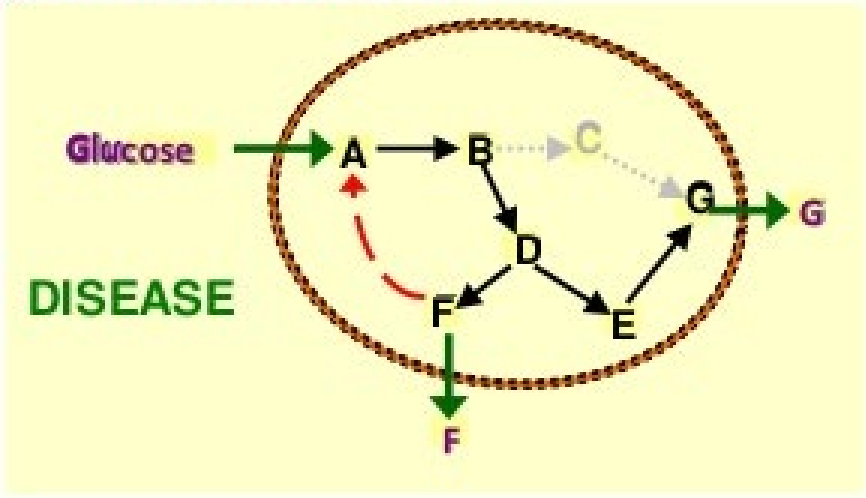
Metabolic Network Inference

✸ *Search for the link between* ***metabolome data*** *and* ***underlying metabolic networks.***

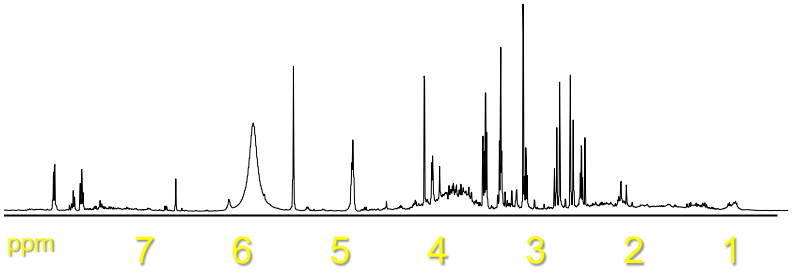


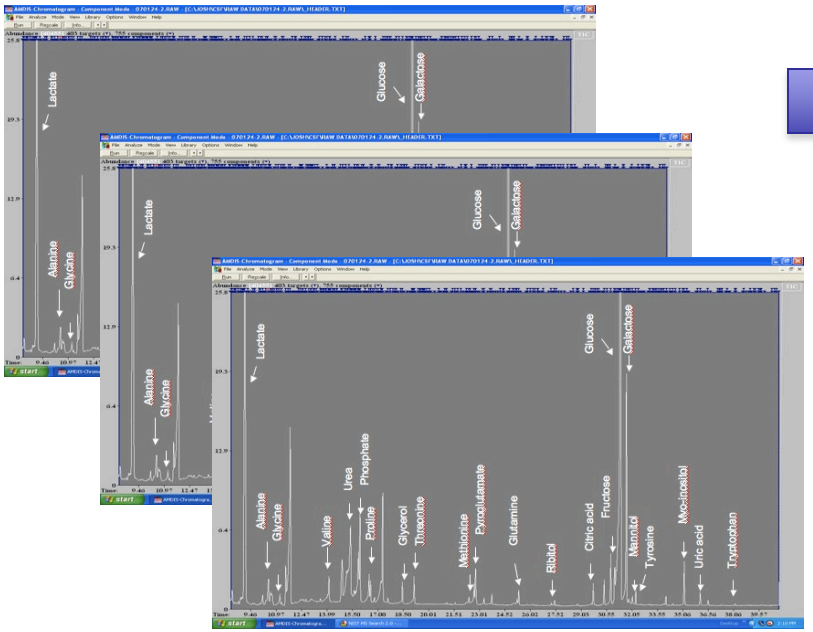
✸ *As an example: can we distinguish healthy from diseased networks:*





From Spectra to Lists

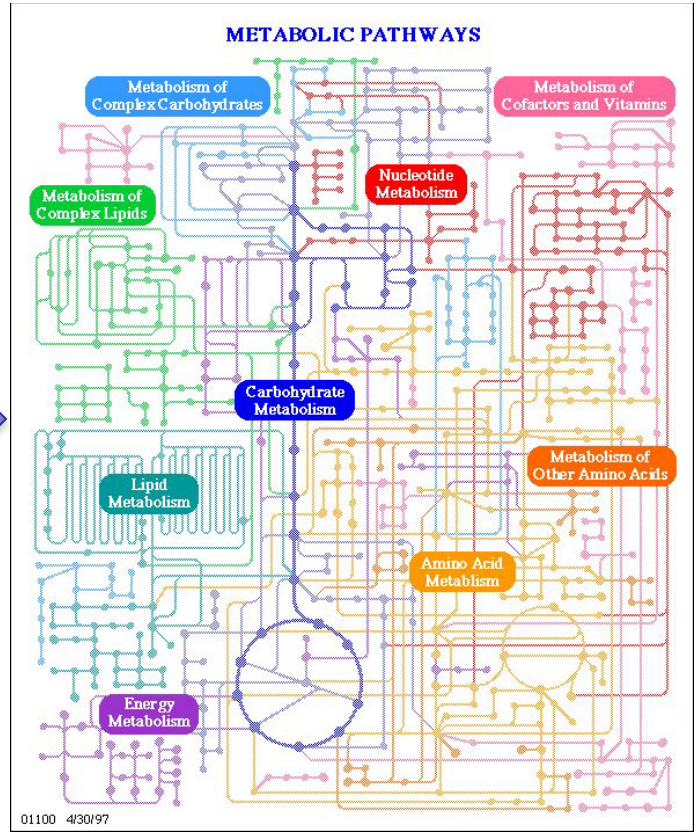




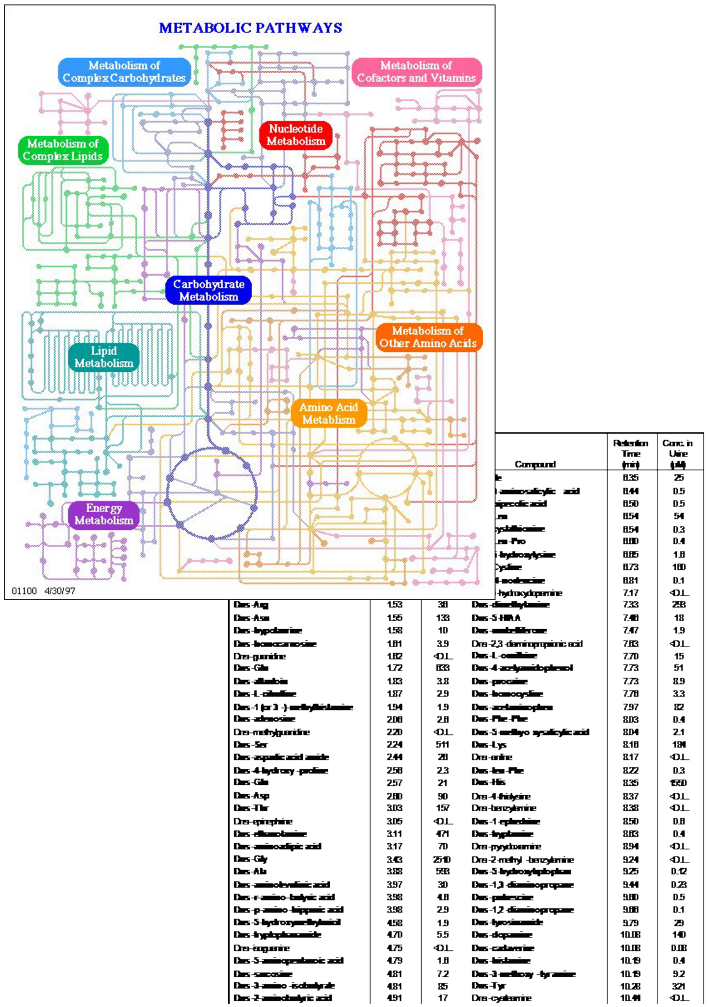
Compound	Retention Time (min)	Conc. in Urine (µM)	Compound	Retention Time (min)	Conc. in Urine (µM)
Dns-o-phospho -L-serine	0.92	<D.L.*	Dns-Ile	6.35	25
Dns-o-phospho -L-tyrosine	0.95	<D.L.	Dns-3-aminosalicylic acid	6.44	0.5
Dns-adnosine monophosphate	0.99	<D.L.	Dns-pipecolic acid	6.50	0.5
Dns-o-phosphoethanolamine	1.06	16	Dns-Leu	6.54	54
Dns-glucosamine	1.06	22	Dns-cystathionine	6.54	0.3
Dns-o-phospho -L-threonine	1.09	<D.L.	Dns-Leu-Pro	6.60	0.4
Dns-6-dimet hylamine purine	1.20	<D.L.	Dns-5-hydroxylysine	6.65	1.6
Dns-3-methyl -histidine	1.22	80	Dns-Cystine	6.73	160
Dns-taurine	1.25	834	Dns-N-norleucine	6.81	0.1
Dns-carnosine	1.34	28	Dns-5-hydoxydopamine	7.17	<D.L.
Dns-Arg	1.53	36	Dns-dimethylamine	7.33	293
Dns-Asn	1.55	133	Dns-5-HIAA	7.46	18
Dns-hypotaurine	1.58	10	Dns-umbelliferone	7.47	1.9
Dns-homocarnosine	1.61	3.9	Dns-2,3-diaminopropionic acid	7.63	<D.L.
Dns-guanidine	1.62	<D.L.	Dns-L-ornithine	7.70	15
Dns-Gln	1.72	633	Dns-4-acetylamidophenol	7.73	51
Dns-allantoin	1.83	3.8	Dns-procaine	7.73	8.9
Dns-L-citrulline	1.87	2.9	Dns-homocystine	7.76	3.3
Dns-1 (or 3 -)-methylhistamine	1.94	1.9	Dns-acetaminophen	7.97	82
Dns-adenosine	2.06	2.6	Dns-Phe-Phe	8.03	0.4
Dns-methylguanidine	2.20	<D.L.	Dns-5-methyo xysalicylic acid	8.04	2.1
Dns-Ser	2.24	511	Dns-Lys	8.16	184
Dns-asparlic acid amide	2.44	26	Dns-amine	8.17	<D.L.
Dns-4-hydroxy -proline	2.56	2.3	Dns-leu-Phe	8.22	0.3
Dns-Glu	2.57	21	Dns-His	8.35	1550
Dns-Asp	2.60	90	Dns-4-thidysine	8.37	<D.L.
Dns-Thr	3.03	157	Dns-benzylamine	8.38	<D.L.
Dns-epinephrine	3.05	<D.L.	Dns-1-ephedrine	8.50	0.6
Dns-ethanolamine	3.11	471	Dns-tryptamine	8.63	0.4
Dns-aminoadipic acid	3.17	70	Dns-pyryduramine	8.94	<D.L.
Dns-Gly	3.43	2510	Dns-2-methyl -benzylamine	9.24	<D.L.
Dns-Ala	3.88	593	Dns-5-hydroxyltiptophan	9.25	0.12
Dns-aminolevulinic acid	3.97	30	Dns-1,3-diaminopropane	9.44	0.23
Dns-r-amino -butyric acid	3.98	4.6	Dns-putrescine	9.60	0.5
Dns-p-amino -hippuric acid	3.98	2.9	Dns-1,2-diaminopropane	9.66	0.1
Dns-5-hydro xymethyluricil	4.58	1.9	Dns-tyrosinamide	9.79	29
Dns-tryptophanamide	4.70	5.5	Dns-dopamine	10.08	140
Dns-isoguanine	4.75	<D.L.	Dns-cadaverine	10.08	0.08
Dns-5-aminopentanoic acid	4.79	1.6	Dns-histamine	10.19	0.4
Dns-sarcosine	4.81	7.2	Dns-3-methoxy -tyramine	10.19	9.2
Dns-3-amino -isobutyrate	4.81	85	Dns-Tyr	10.28	321
Dns-2-aminobutyric acid	4.91	17	Dns-cysteamine	10.44	<D.L.

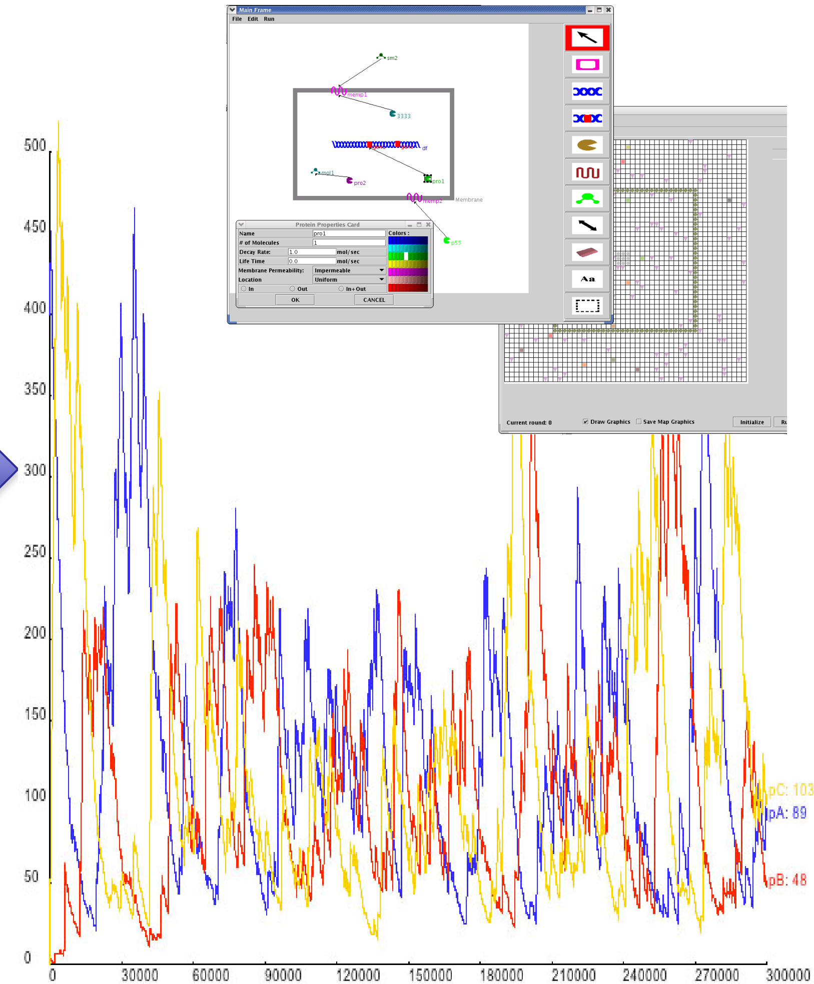
From Lists to Pathways

Compound	Retention Time (min)	Conc. in Urine (µM)	Compound	Retention Time (min)	Conc. in Urine (µM)
Dns-o-phospho -L-serine	0.92	<D.L. *	**Dns-Ile**	6.35	25
Dns-o-phospho -L-tyrosine	0.95	<D.L.	**Dns-3-aminosalicylic acid**	6.44	0.5
Dns-adnosine monophosphate	0.99	<D.L.	**Dns-pipecolic acid**	6.50	0.5
Dns-o-phosphoethanolamine	1.06	16	**Dns-Leu**	6.54	54
Dns-glucosamine	1.06	22	**Dns-cystalhionine**	6.54	0.3
Dns-o-phospho -L-threonine	1.09	<D.L.	**Dns-Leu-Pro**	6.60	0.4
Dns-6-dimet hylamine purine	1.20	<D.L.	**Dns-5-hydroxylysine**	6.65	1.6
Dns-3-methyl -histidine	1.22	80	**Dns-Cystine**	6.73	160
Dns-taurine	1.25	834	**Dns-N-norleucine**	6.81	0.1
Dns-carnosine	1.34	28	Dns-5-hydoxydopamine	7.17	<D.L.
Dns-Arg	1.53	36	**Dns-dimethylamine**	7.33	293
Dns-Asn	1.55	133	**Dns-5-HIAA**	7.46	18
Dns-hypotaurine	1.58	10	**Dns-umbelliferone**	7.47	1.9
Dns-homocarnosine	1.61	3.9	Dns-2,3-diaminopropanic acid	7.63	<D.L.
Dns-guanidine	1.62	<D.L.	**Dns-L-ornithine**	7.70	15
Dns-Gln	1.72	633	**Dns-4-acetyamidophenol**	7.73	51
Dns-allantoin	1.83	3.8	**Dns-procaine**	7.73	8.9
Dns-L-citrulline	1.87	2.9	**Dns-homocystine**	7.76	3.3
Dns-1 (or 3 -)-methylhistamine	1.94	1.9	**Dns-acetaminophen**	7.97	82
Dns-adenosine	2.06	2.6	**Dns-Phe-Phe**	8.03	0.4
Dns-methylguanidine	2.20	<D.L.	**Dns-5-methyo xysalicylic acid**	8.04	2.1
Dns-Ser	2.24	511	**Dns-Lys**	8.16	184
Dns-aspartic acid amide	2.44	26	Dns-aniline	8.17	<D.L.
Dns-4-hydroxy -proline	2.56	2.3	**Dns-leu-Phe**	8.22	0.3
Dns-Glu	2.57	21	**Dns-His**	8.35	1550
Dns-Asp	2.60	90	Dns-4 thialysine	8.37	<D.L.
Dns-Thr	3.03	157	Dns-benzylamine	8.38	<D.L.
Dns-epinephrine	3.05	<D.L.	**Dns-1-ephedrine**	8.50	0.6
Dns-ethanolamine	3.11	471	**Dns-tryptamine**	8.63	0.4
Dns-aminoadipic acid	3.17	70	Dns-pyrydoxamine	8.94	<D.L.
Dns-Gly	3.43	2510	Dns-2-methyl -benzylamine	9.24	<D.L.
Dns-Ala	3.88	593	**Dns-5-hydroxytriptophan**	9.25	0.12
Dns-aminolevulinic acid	3.97	30	**Dns-1,3-diaminopropane**	9.44	0.23
Dns-r-amino -butyric acid	3.98	4.6	**Dns-putrescine**	9.60	0.5
Dns-p-amino -hippuric acid	3.98	2.9	**Dns-1,2-diaminopropane**	9.66	0.1
Dns-5-hydro xymethyluracil	4.58	1.9	**Dns-tyrosinamide**	9.79	29
Dns-tryptophanamide	4.70	5.5	**Dns-dopamine**	10.08	140
Dns-isoguanine	4.75	<D.L.	**Dns-cadaverine**	10.08	0.08
Dns-5-aminopentanoic acid	4.79	1.6	**Dns-histamine**	10.19	0.4
Dns-sarcosine	4.81	7.2	**Dns-3-methoxy -tyramine**	10.19	9.2
Dns-3-amino -isobutyrate	4.81	85	**Dns-Tyr**	10.28	321
Dns-2-aminobutyric acid	4.91	17	Dns-cysteamine	10.44	<D.L.



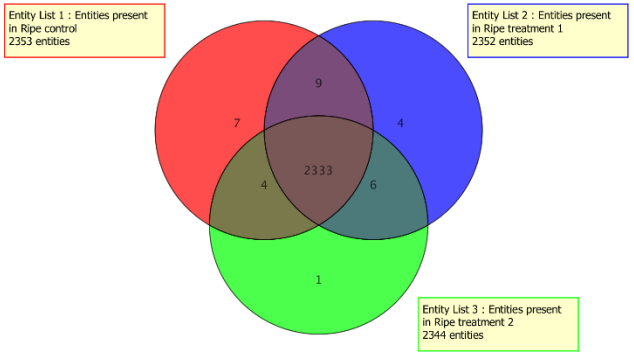
From Pathways & Lists to Models & Biomarkers

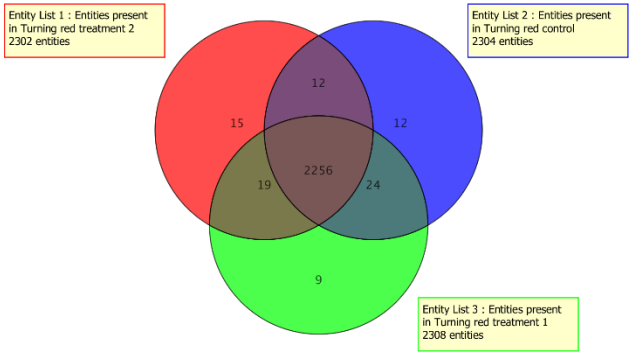




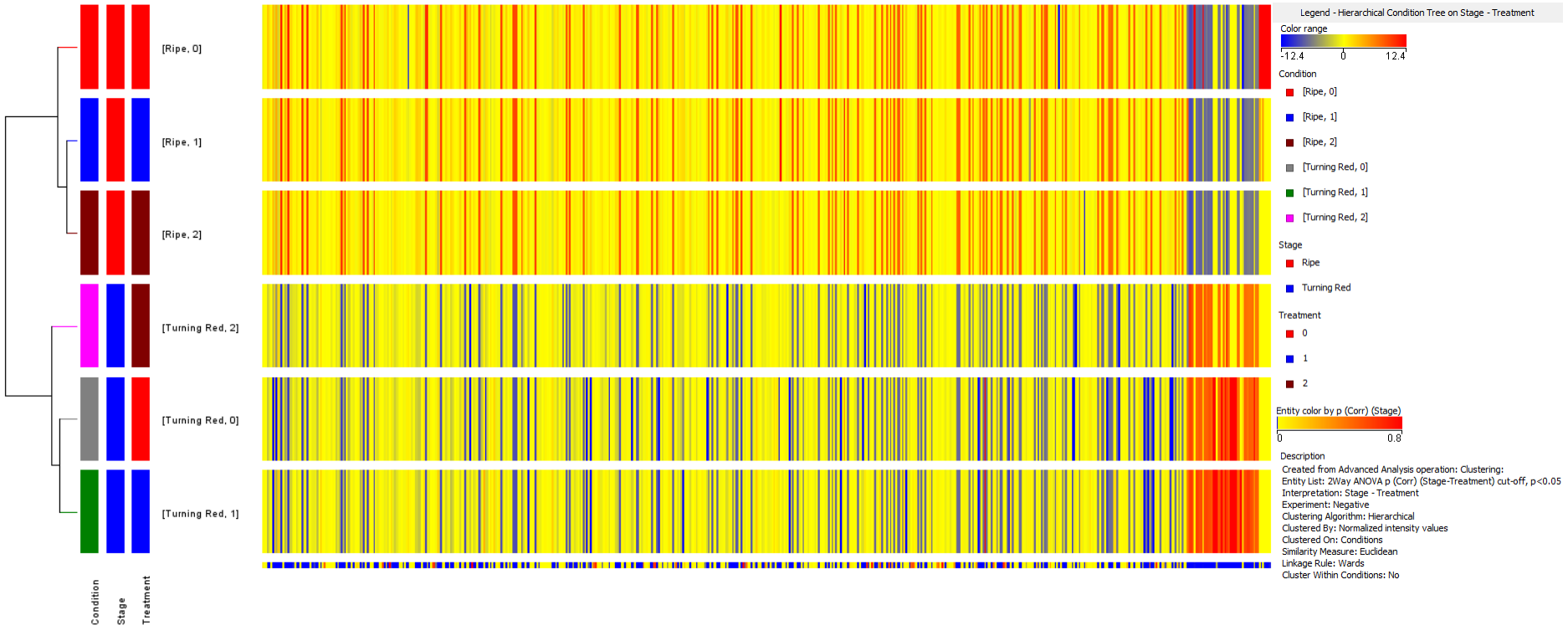
Example LC-MS

- Blueberry Treatment 1 and 2
- Ripe vs Turning red



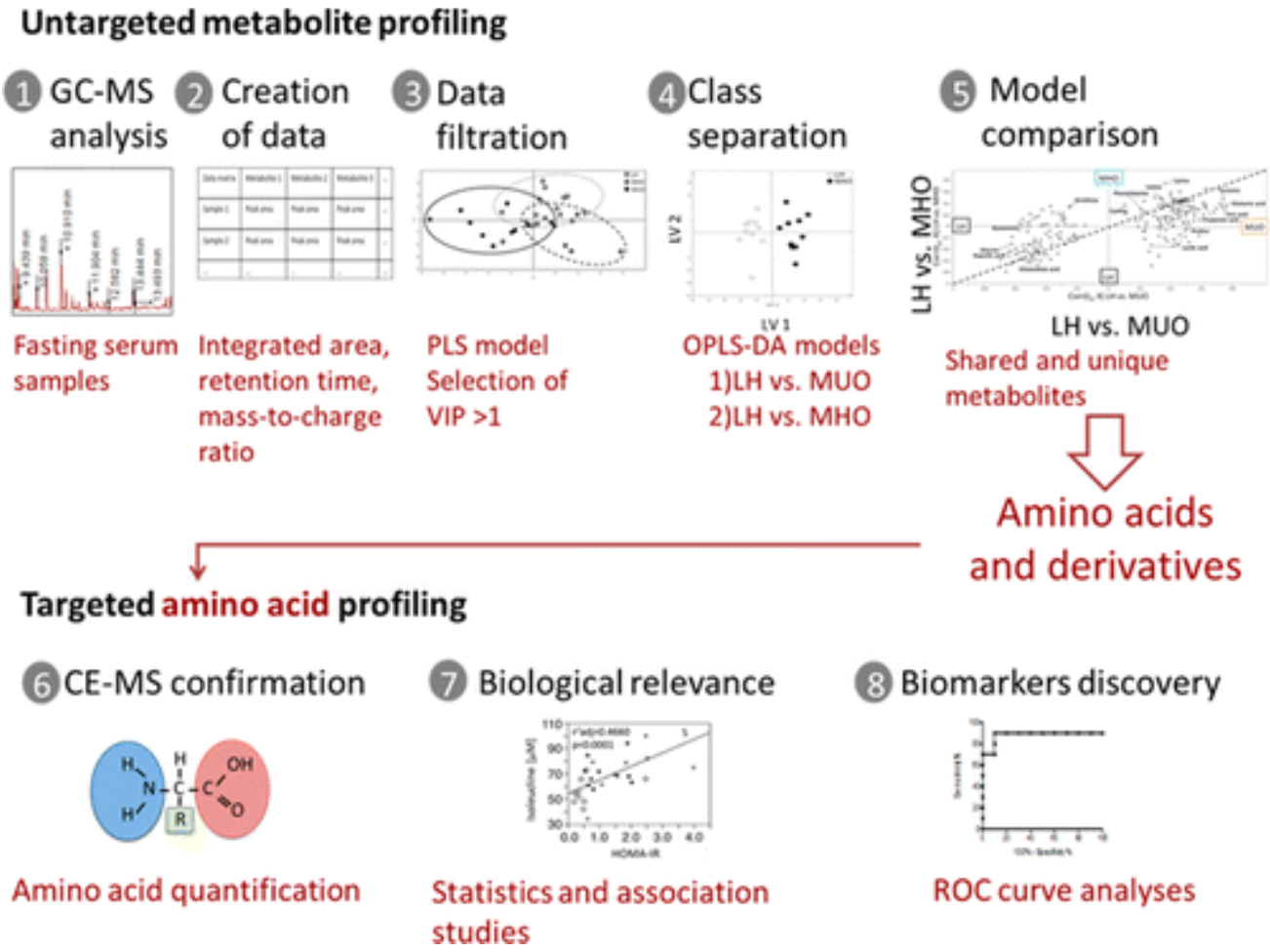


Example LC-MS



Treatment 2 has more of an effect on blueberries turning red

Serum and Adipose Tissue Amino Acid Metabolite Quantitation

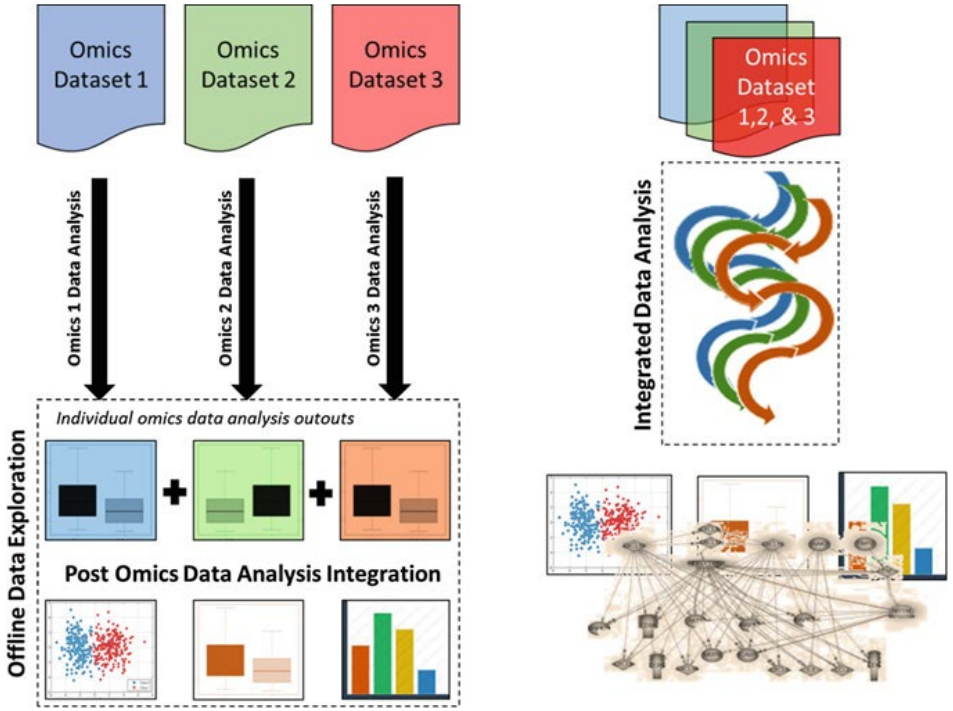


J. Proteome Res., 2014, 13 (7), pp 3455–3466

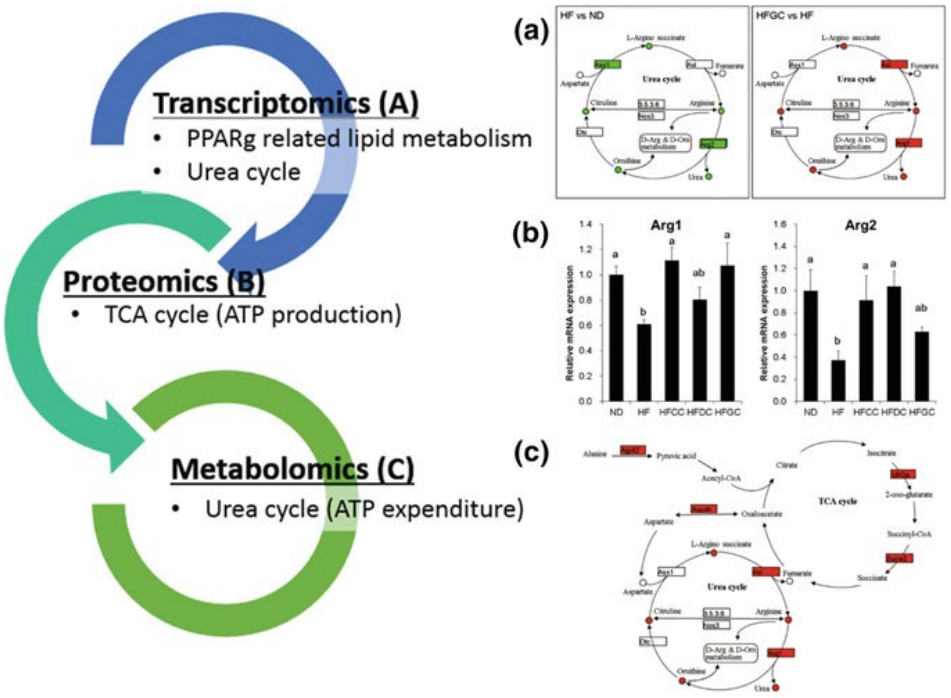
Key Informatics Challenges in Metabolomics

- Spectra -> Lists
  - Data integrity and quality
  - Data alignment and normalization
  - Data reduction and classification
  - Assessment of significance
  - Metabolite identification/quantification
- Lists -> Pathways & Biomarkers
  - Pathway mapping and identification
  - Biological interpretation

Multi-Omics

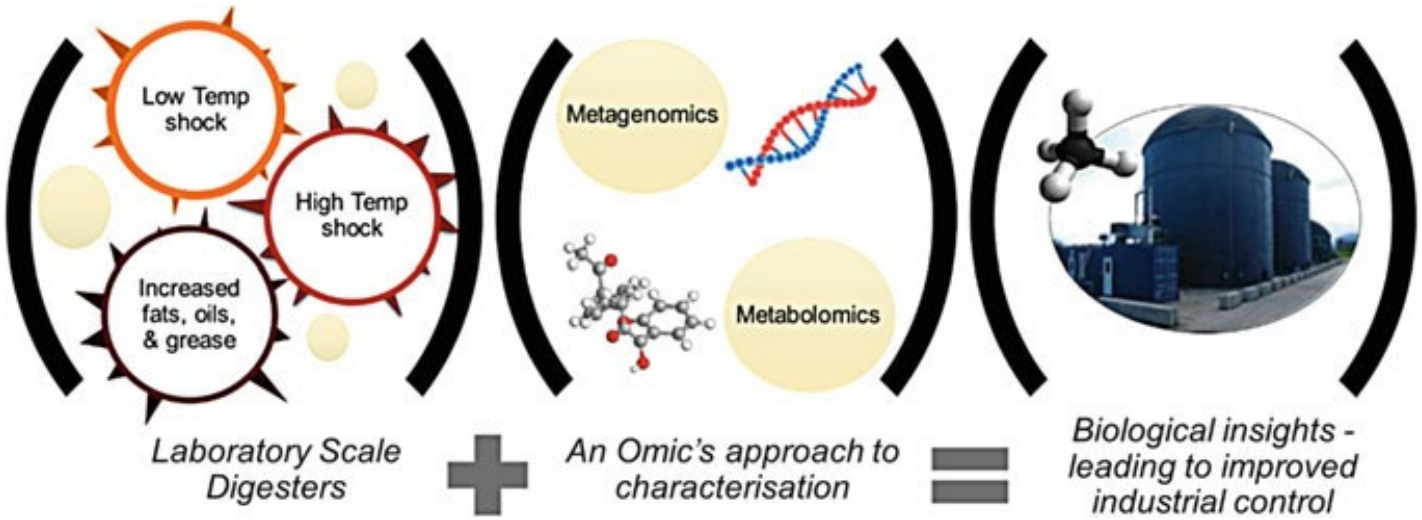

Omics Dataset 1
Omics Dataset 2
Omics Dataset 3
Omics 1 Data Analysis
Omics 2 Data Analysis
Omics 3 Data Analysis
Individual omics data analysis outouts
Offline Data Exploration
Post Omics Data Analysis Integration
Omics Dataset 1,2, & 3
Integrated Data Analysis

Multi-Omics



Takahashi S, Saito K, Jia H, Kato H (2014)
PLoS One 9:e91134

Multi-Omics

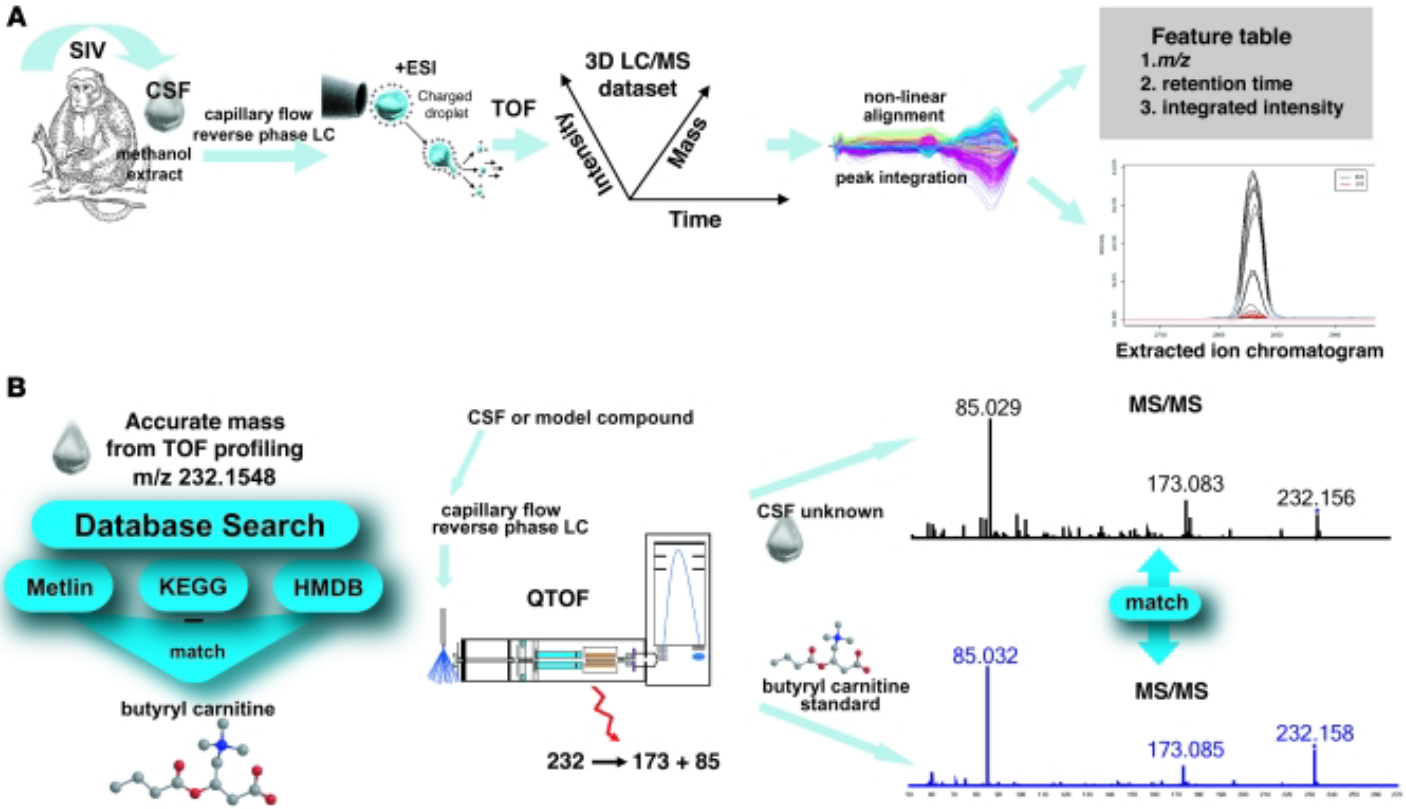


Beale DJ, Barratt R, Marlow DR, Dunn MS, Palombo EA, Morrison PD, Key C (2013) Biofouling 29:283–294

References

- Goodacre *et al* Analytica Chimica Acta **879** 2015 1-23.
- Julia *et al* Front Bioeng Biotechnol. 2015 **3** 1-20.
- Suizdak *et al* Nat Rev Mol Cell Bio 2016 **13**, 263-269.
- Suizdak *et al* Nat Rev Mol Cell Bio 2016 **17**, 451-459.
- Methods Mol Biol. 2014; **1198**: 333–353.

Metabolomic analysis of the cerebrospinal fluid reveals changes in phospholipase expression in the CNS of SIV-infected macaques



Fox et al, *J. Clin Invest.* 2008, 118, 2661.

Metabolomics

Table 3
Selected metabolites

Name	Molecular formula	*m/z* observed	*m/z* calculated	Relative error (ppm)	Retention time (min)	*P* value	Fold change
Carnitine	$C_7H_{15}NO_3$	162.1124	162.1125	−0.62	2.9	0.010 (0.52)	9.1 (1.1)
Isovaleryl carnitine	$C_{12}H_{23}NO_4$	246.1694	246.1699	−2.44	12.5	0.007 (0.35)	9.9 (1.2)
Octanoyl carnitine	$C_{15}H_{29}NO_4$	288.2173	288.2169	1.04	23.9	0.005 (0.29)	3.7 (1.3)
Butyryl or isobutyryl carnitine	$C_{11}H_{21}NO_4$	232.1548	232.1544	1.72	7.1	<0.001 (0.36)	10.4 (1.8)
Palmitic acid (C16:0)	$C_{16}H_{32}O_2$	257.2478	257.2475	1.17	47.2	0.011 (0.58)	35.5 (1.2)
Oleic acid (C18:1)	$C_{18}H_{34}O_2$	283.2634	283.2632	0.71	39.5	0.029 (0.97)	2.5 (1.0)
Linoleic acid (C18:2)	$C_{18}H_{32}O_2$	281.2472	281.2475	−1.07	40.7	0.004 (nd)	3.2 (nd)
Stearic acid (C18:0)	$C_{18}H_{36}O_2$	285.2793	285.2788	1.75	50.9	0.005 (0.37)	14.4 (0.7)
Myristic acid (C14:0)	$C_{14}H_{28}O_2$	229.2164	229.2162	0.87	43.0	0.009 (0.20)	5.5 (1.2)
LPC (16:0)	$C_{24}H_{50}NO_7P$	496.3373	496.3398	−5.04	38.6	0.039 (0.59)	26.3 (1.2)
LPC (18:2)	$C_{26}H_{52}NO_7P$	522.3542	522.3554	−2.30	39.5	0.042 (0.61)	27.5 (1.2)
LPC (18:0)	$C_{26}H_{54}NO_7P$	524.3720	524.3711	1.72	43.5	0.038 (0.44)	73.9 (1.4)

Twelve metabolites were identified from the features that changed following SIV infection, based on mass accuracy and retention time. The observed and calculated *m/z* are given, with the differences shown in parts per million (ppm). The fold change reflects the average increase in samples taken after infection. *P* values (obtained using *t* test) and fold changes for animals that did not progress to encephalitis are shown in parentheses. nd, not determined.