Biomarker discovery: LC-MS Proteomics

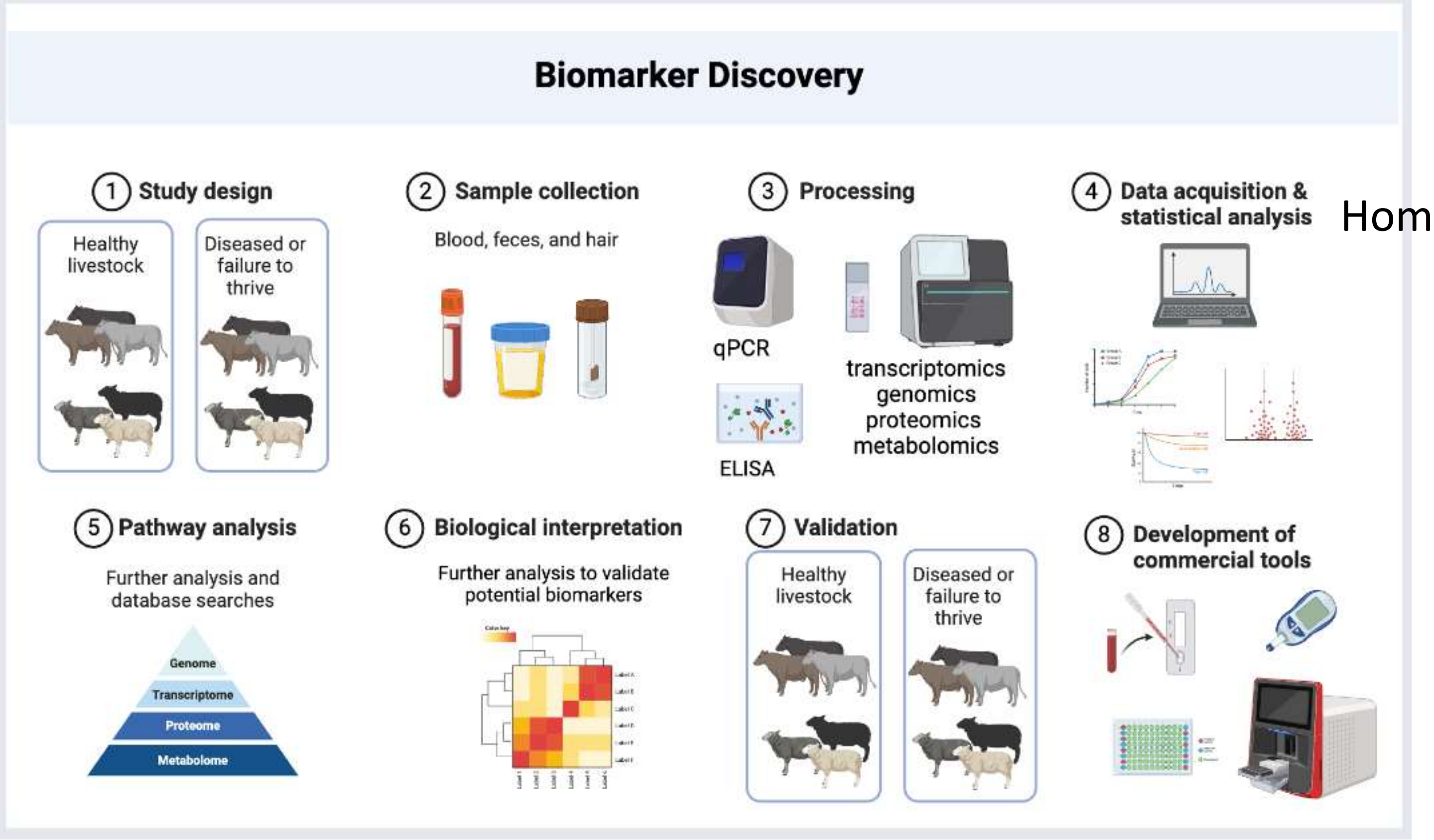
Nanocourse: Data Science using R
September 6th, 2024

Jeon Lee

Agenda

1. Overview of biomarker discovery steps
2. Batch correction/data harmonization
3. Harmonization of proteomics data with missing values
4. Introduction to LC-MS proteomics
    - MS for metabolomics/proteomics
    - LC; LC-MS/MS
    - Peak annotation
    - Typical proteomics data & analysis steps
5. Demo: Proteomics analysis
6. Hands-on practice

Overview of biomarker discovery steps


Biomarker Discovery
1 Study design
Healthy livestock
Diseased or failure to thrive
2 Sample collection
Blood, feces, and hair
3 Processing
qPCR
ELISA
transcriptomics
genomics
proteomics
metabolomics
4 Data acquisition & statistical analysis
5 Pathway analysis
Further analysis and database searches
Genome
Transcriptome
Proteome
Metabolome
6 Biological interpretation
Further analysis to validate potential biomarkers
7 Validation
Healthy livestock
Diseased or failure to thrive
8 Development of commercial tools

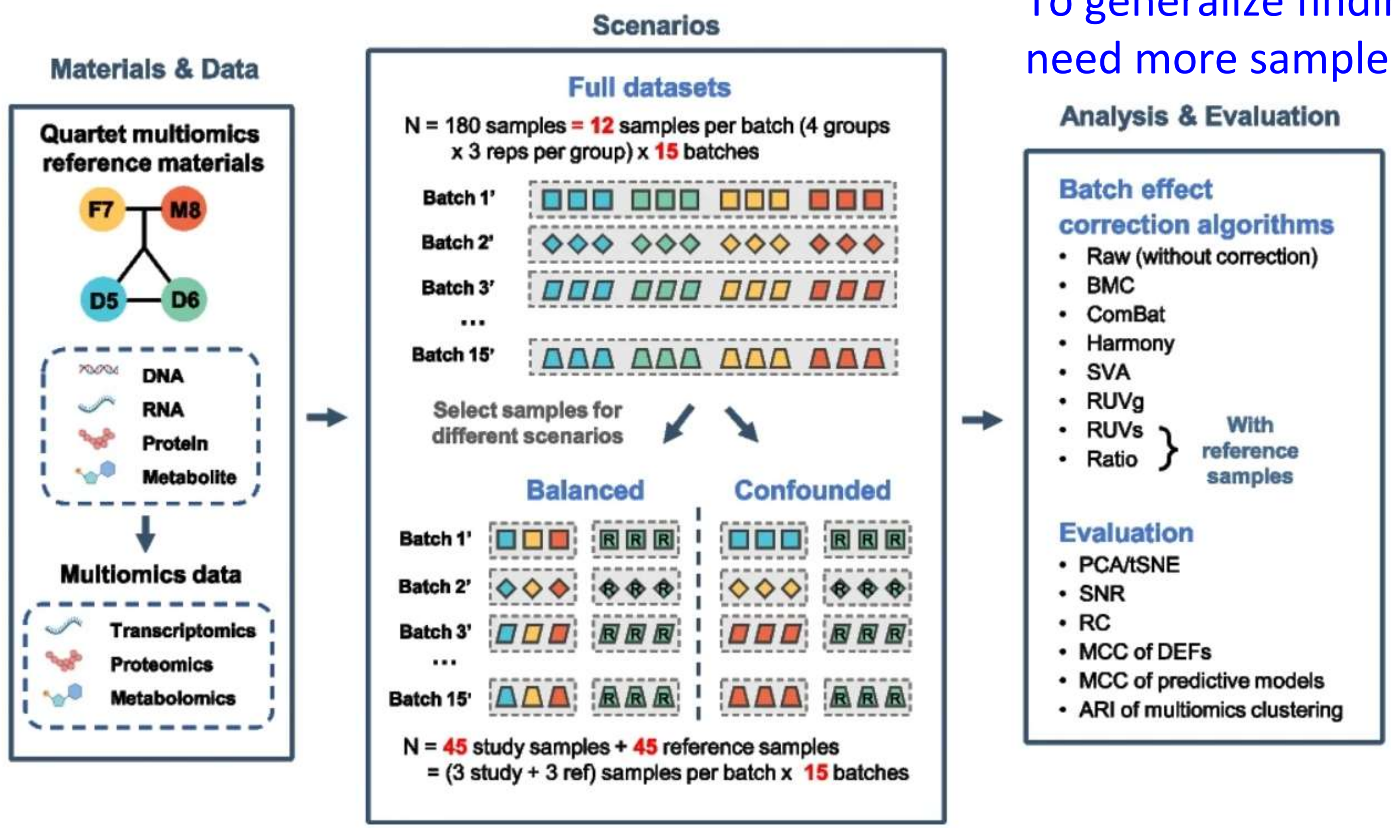
Homogenize data

<Adopted from *https://app.biorender.com/profile/auriol_purdie/templates/6635bb7572a44e4ad29ce313* >

For RNAseq, have all the data

Batch correction/data harmonization (1)

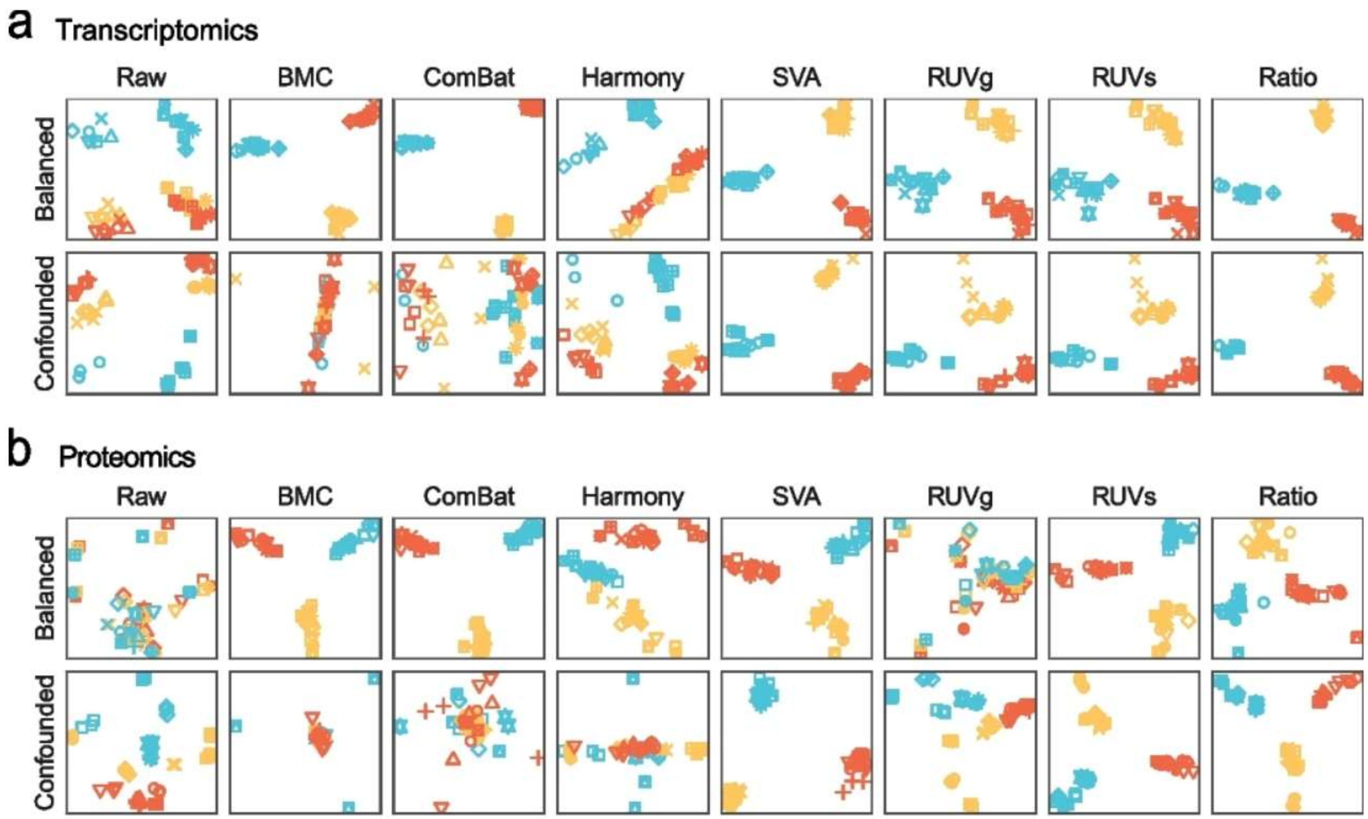
To generalize findings with high confidence, need more samples.



Batch effect- Comparing two samples is not enough due to batch differences

Yu, Y., Zhang, N., Mai, Y. *et al.* Correcting batch effects in large-scale multiomics studies using a reference-material-based ratio method. *Genome Biol* **24**, 201 (2023).

Batch correction/data harmonization (2)

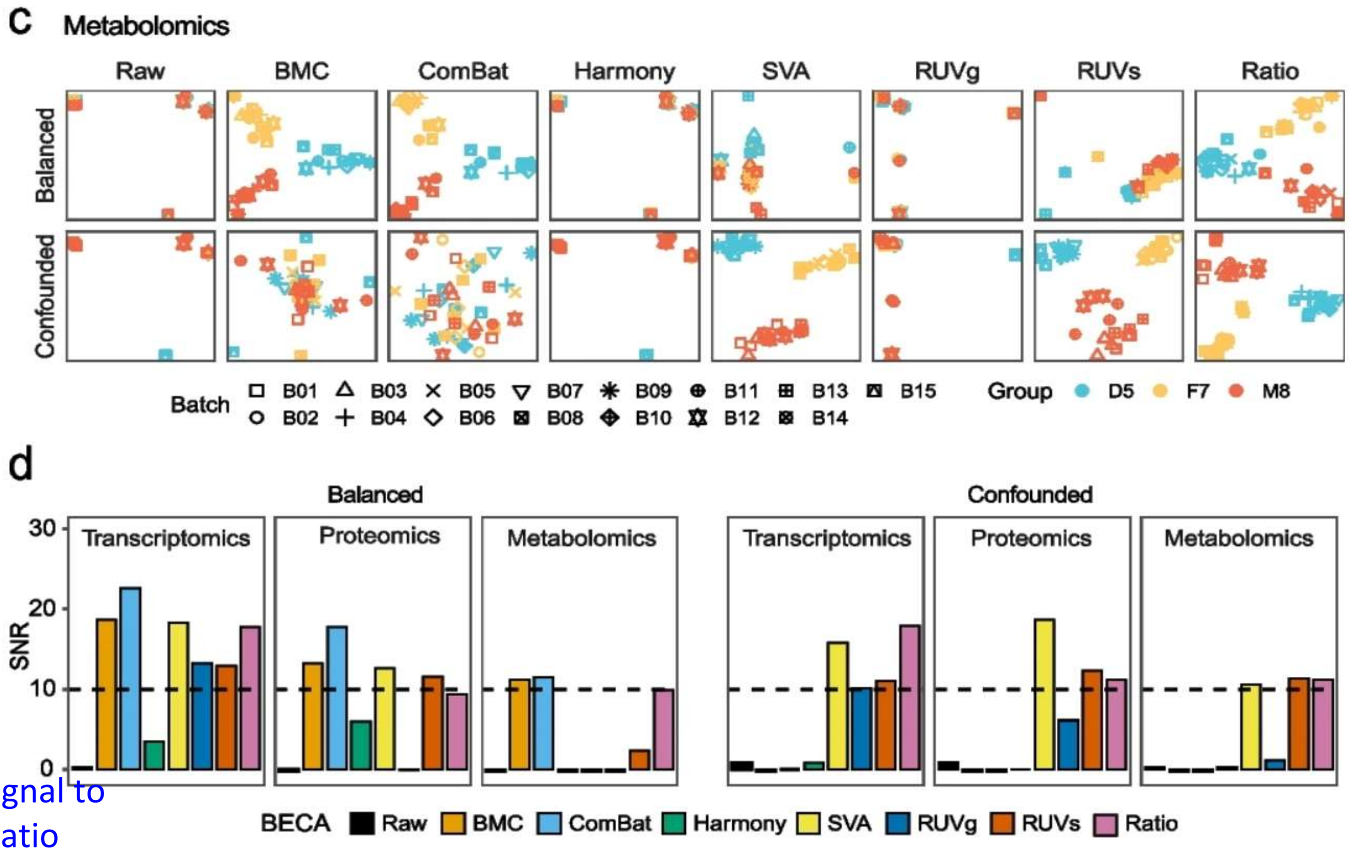


Has slightly better ability to remove batch effect (could be bc has more features/data points than proteomics (10000 vs 1000))

Use similar technique, but how batch effect influences each level depends

Yu, Y., Zhang, N., Mai, Y. *et al.* Correcting batch effects in large-scale multiomics studies using a reference-material-based ratio method. *Genome Biol* **24**, 201 (2023).

Batch correction/data harmonization (3)



SNR=signal to noise ratio

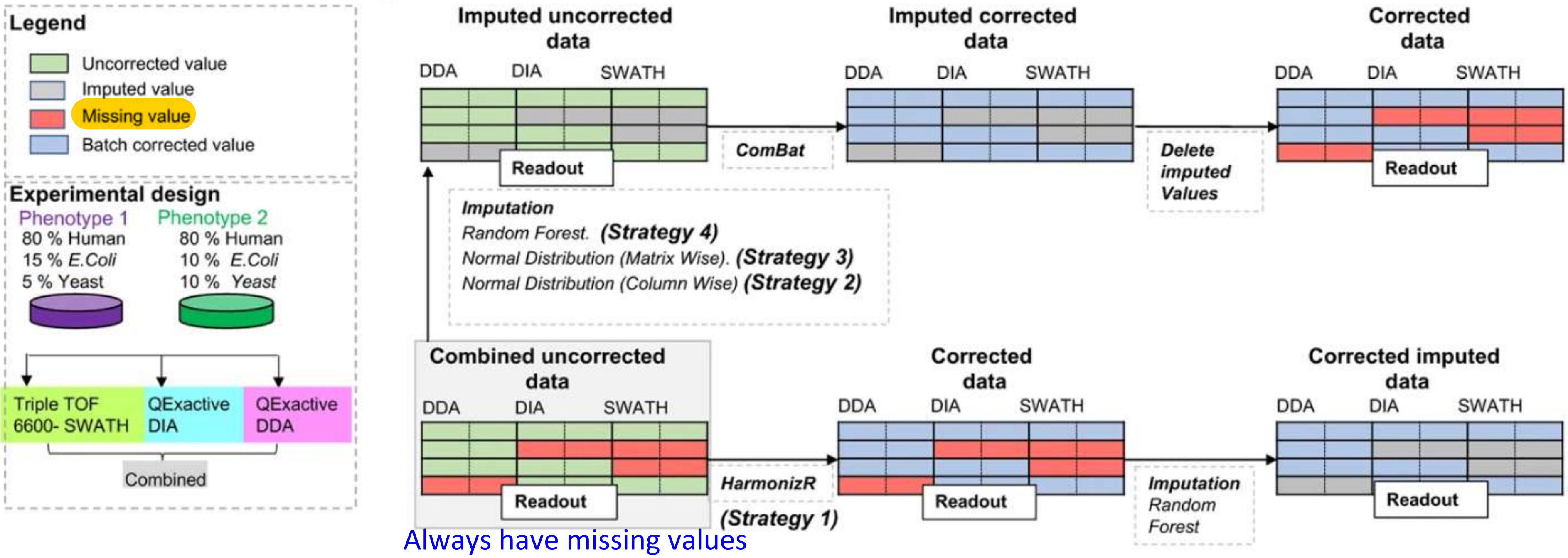
Yu, Y., Zhang, N., Mai, Y. *et al.* Correcting batch effects in large-scale multiomics studies using a reference-material-based ratio method. *Genome Biol* **24**, 201 (2023).

Usually, lots of missing data points

Harmonization of proteomics data with missing values (1)

For single cell and LC-MS data:
Have missing values

Can't batch correct imputed data bc need complete data to batch correct. So imput first

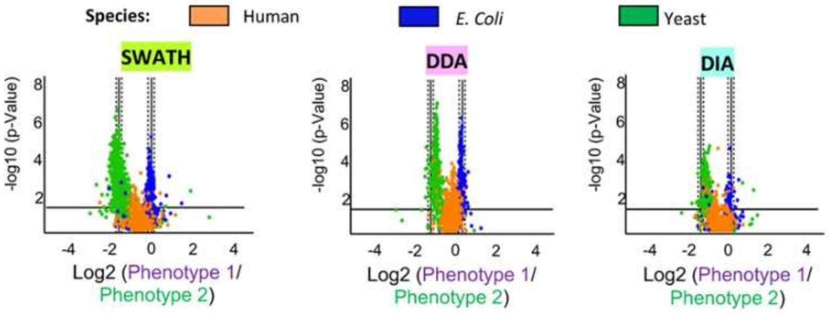


Always have missing values

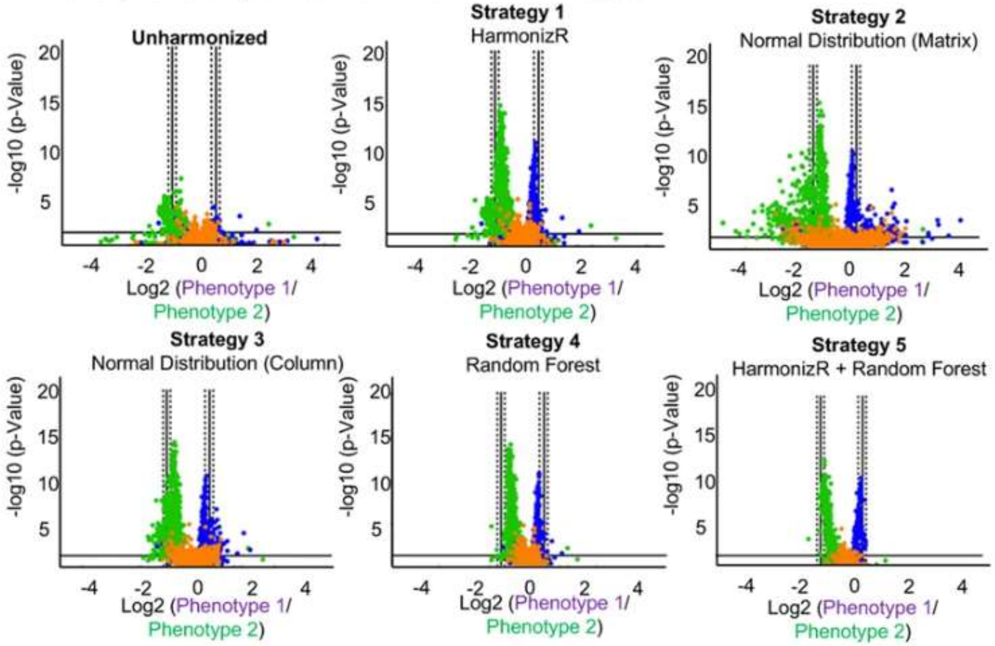
Voß, H., Schlumbohm, S., Barwikowski, P. *et al.* HarmonizR enables data harmonization across independent proteomic datasets with appropriate handling of missing values. *Nat Commun* **13**, 3523 (2022)

Harmonization of proteomics data with missing values (2)

Volcano plot visualization of individual experimental setups



Volcano plot comparison of different strategies



Voß, H., Schlumbohm, S., Barwikowski, P. *et al.* HarmonizR enables data harmonization across independent proteomic datasets with appropriate handling of missing values. *Nat Commun* **13**, 3523 (2022)

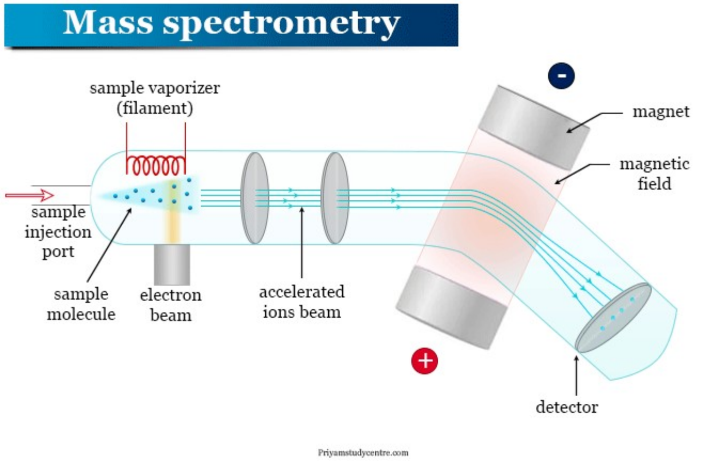
Introduction to LC-MS proteomics

MS for metabolomics/proteomics

- MS is an analytical technique that ionizes chemical species and sorts the ions based on their mass-to-charge ratio (m/z).

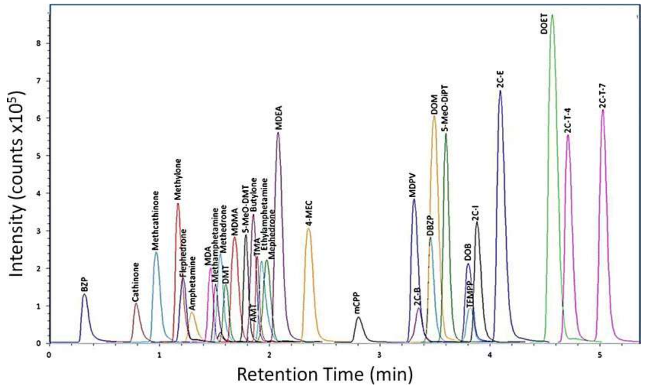
MS LC

- Mass spectrometers are comprised of an ionization source and a mass detector, for example,
  - MALDI-TOF: matrix assisted laser desorption ionization, time-of-flight detection
  - ESI-trap: electrospray ionization, ion trap detection

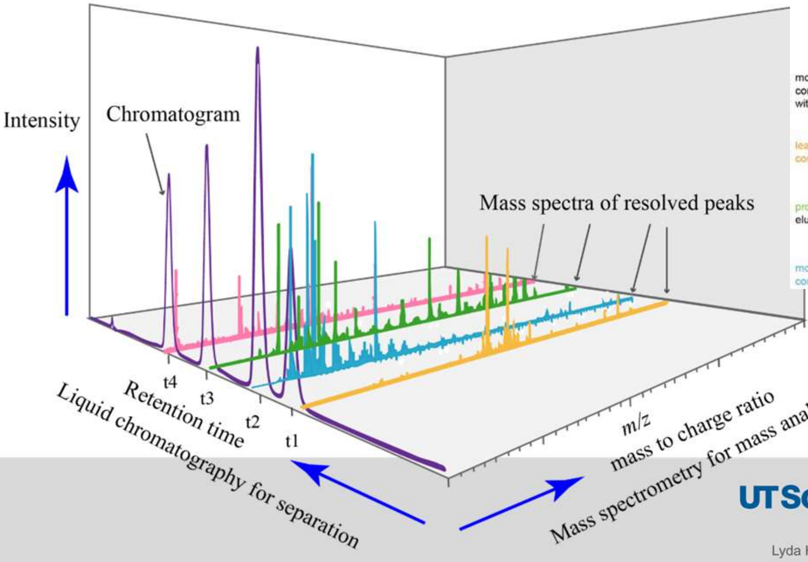


Liquid chromatography

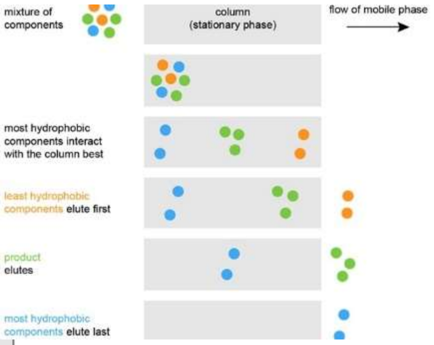
- LC is the separation technique of choice for larger and non-volatile molecules such as proteins and complex peptides
- LC is also an ideal method for separating isomers, which have the same mass and will otherwise not be differentiated by a mass spectrometer
- LC-MS offers broad sample coverage because different column chemistries, such as reversed phase liquid chromatography, can be used



Ex. LC-MS chromatogram for 32 targeted analytes
Swortwood, M. et al. (2012), Anal. Bioanal. Chem.


Intensity
Chromatogram
Mass spectra of resolved peaks
t4
t3
t2
t1
Retention time
Liquid chromatography for separation
m/z
mass to charge ratio
Mass spectrometry for mass analysis

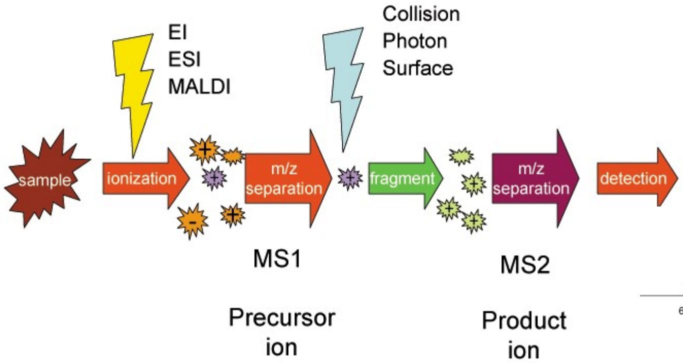
Liquid chromatography–mass spectrometry, Wikipedia


mixture of components
column (stationary phase)
flow of mobile phase
most hydrophobic components interact with the column best
least hydrophobic components elute first
product elutes
most hydrophobic components elute last

Reversed-Phase Chromatography

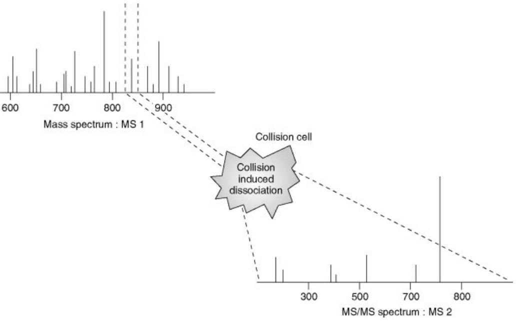
Tandem mass spectrometry (LC-MS/MS)

- Combination of LC and two mass analyzers in mass spectrometry (MS/MS)
- Once samples are ionized to generate a mixture of ions, precursor ions of a specific mass-to-charge ratio (*m/z*) are selected (MS1) and then fragmented (MS2) to generate a product ions for detection.
- The fragments then reveal aspects of the chemical structure of the precursor ion.

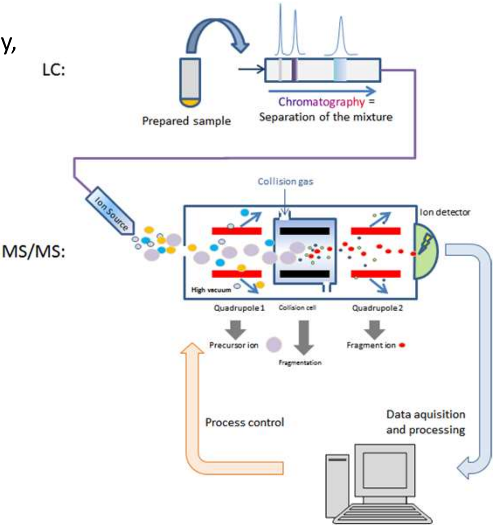


To get better res, ionizes twice

Tandem_mass_spectrometry, WikiPedia



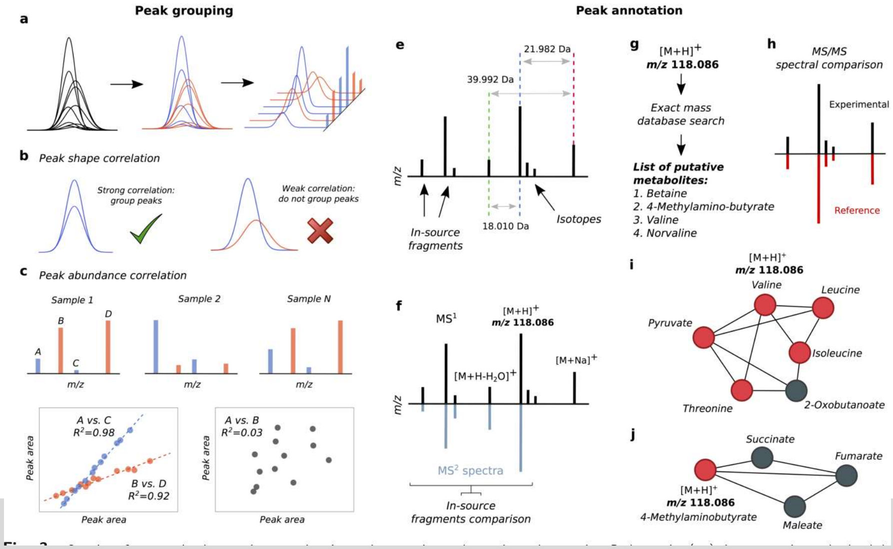
A brief history of mass spectrometry, healthcare-in-europe.com



F.A. Mellon, Encyclopedia of Food Sciences and Nutrition (2nd Edition, 2003)

Peak annotation

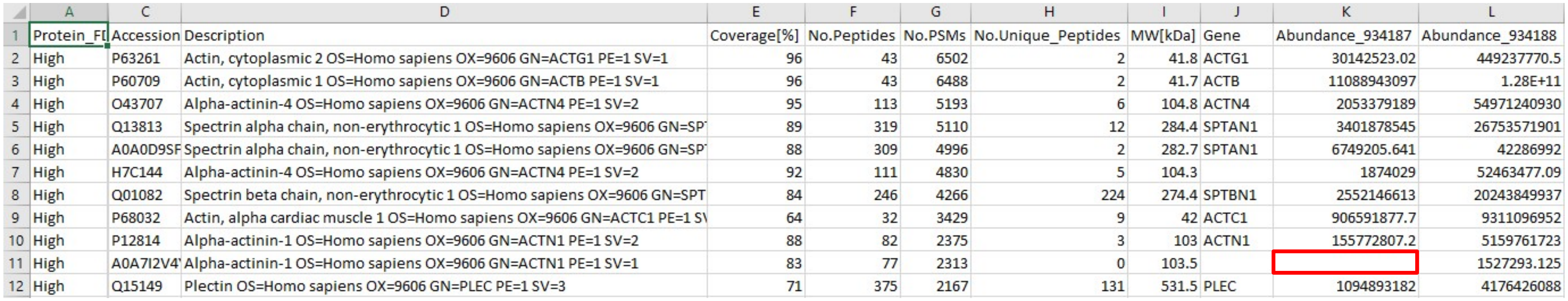
- Peak grouping (a-c) aims at grouping peaks that belong to each metabolite/protein.
- In feature annotation (e, f), expected theoretical distances between known ion adduct masses are compared with experimental distances found among peaks (e).
- After peak annotation, putative identification can be achieved by accurate mass search (g) or by comparison with MS/MS data (h).



Xavier, D. et al (2018) Analytical chemistry

Typical proteomics data

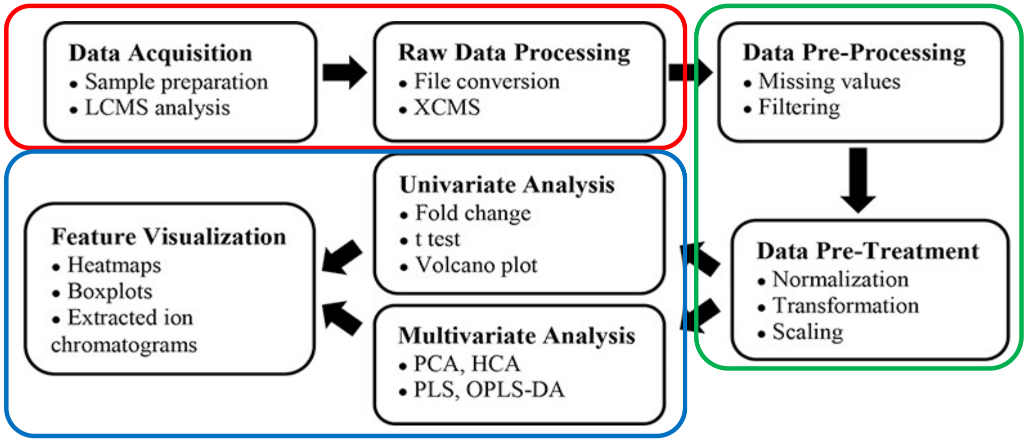
- Each row includes an accession id, of which protein has been detected, other meta data, and its abundance measured across the samples.
  - Coverage[%] = no. amino acids in all found peptides / total no. amino acids in the entire protein sequence
  - No.Peptides: No. distinct peptide sequences in the protein group
  - No.Unique_Peptides: No. peptide sequences unique to a protein group
  - No.PSMs: Total no. identified peptide sequences for the protein, including those redundantly identified
  - MW[kDa]: Molecular weight without considering post-translational modifications
- Some proteins are detected in some samples but not in the other samples



	A	C	D	E	F	G	H	I	J	K	L
1	Protein_FD	Accession	Description	Coverage[%]	No.Peptides	No.PSMs	No.Unique_Peptides	MW[kDa]	Gene	Abundance_934187	Abundance_934188
2	High	P63261	Actin, cytoplasmic 2 OS=Homo sapiens OX=9606 GN=ACTG1 PE=1 SV=1	96	43	6502	2	41.8	ACTG1	30142523.02	449237770.5
3	High	P60709	Actin, cytoplasmic 1 OS=Homo sapiens OX=9606 GN=ACTB PE=1 SV=1	96	43	6488	2	41.7	ACTB	11088943097	1.28E+11
4	High	O43707	Alpha-actinin-4 OS=Homo sapiens OX=9606 GN=ACTN4 PE=1 SV=2	95	113	5193	6	104.8	ACTN4	2053379189	54971240930
5	High	Q13813	Spectrin alpha chain, non-erythrocytic 1 OS=Homo sapiens OX=9606 GN=SPT	89	319	5110	12	284.4	SPTAN1	3401878545	26753571901
6	High	A0A0D9SF	Spectrin alpha chain, non-erythrocytic 1 OS=Homo sapiens OX=9606 GN=SPT	88	309	4996	2	282.7	SPTAN1	6749205.641	42286992
7	High	H7C144	Alpha-actinin-4 OS=Homo sapiens OX=9606 GN=ACTN4 PE=1 SV=2	92	111	4830	5	104.3		1874029	52463477.09
8	High	Q01082	Spectrin beta chain, non-erythrocytic 1 OS=Homo sapiens OX=9606 GN=SPT	84	246	4266	224	274.4	SPTBN1	2552146613	20243849937
9	High	P68032	Actin, alpha cardiac muscle 1 OS=Homo sapiens OX=9606 GN=ACTC1 PE=1 SV	64	32	3429	9	42	ACTC1	906591877.7	9311096952
10	High	P12814	Alpha-actinin-1 OS=Homo sapiens OX=9606 GN=ACTN1 PE=1 SV=2	88	82	2375	3	103	ACTN1	155772807.2	5159761723
11	High	A0A7I2V4	Alpha-actinin-1 OS=Homo sapiens OX=9606 GN=ACTN1 PE=1 SV=1	83	77	2313	0	103.5			1527293.125
12	High	Q15149	Plectin OS=Homo sapiens OX=9606 GN=PLEC PE=1 SV=3	71	375	2167	131	531.5	PLEC	1094893182	4176426088

Proteomics analysis steps

- Three steps in proteomics analysis
  - Data acquisition/raw data processing
  - Data pre-processing
  - (main/downstream) data analysis
- Typical data pre-processing steps include: 1) missing value imputation, 2) transformation, (3) scaling, and (4) normalization



Demo: Proteomics analysis

"Demo_Proteomics_Analysis.html"