WELCOME!

Summer School will begin at 8:30 a.m. PDT

8:30-8:35 a.m.	Introduction	Kelly Stratton
8:35-9:25	Types of Proteomics	Paul Piehowski & David Degnan
9:25-9:35	Networking Break	
9:35-10:40	Typical Statistical Processing	Kelly Stratton
10:40-10:50	Networking Break	
10:50-11:40	Biological Interpretation	David Degnan & Tyler Sagendorf
11:40-11:45	Closing Remarks	David Degnan

EMSL

Summer School
Day 3: Proteomics

David Degnan & Kelly Stratton
Biostatistics & Data Science
07.26.2023

Instructor Intro

David Degnan

Biological Data Scientist

- Multi-omics Statistics, Top-Down Proteomics, Metabolomics, R Package and Shiny Application Development, Containerization
- Day 1: Data Science for ‘Omics Data
- Day 3: Proteomics
- david.degnan@pnnl.gov
- linkedin.com/in/David-Degnan

Instructor Intro

Kelly Stratton

Biostatistician

- Data Scientist, Data Transformations IRP Lead
- Statistics, R, visualization, analysis of ‘omics data
- Day 1: Data Science for ‘Omics Data
- Day 3: Proteomics
- kelly.stratton@pnnl.gov

Instructor Intro

Paul Piehowski

Chemist

- Functional and Systems Biology Team Lead
- Mass spectrometry, proteomics, nanoPOTS platform
- Day 3: Proteomics
- paul.piehowski@pnnl.gov

Instructor Intro

Tyler Sagendorf

Data Scientist

- Data visualization, data wrangling, R, statistics, proteomics
- Day 3: Proteomics
- tyler.sagendorf@pnnl.gov

8:30-8:35 a.m.	Introduction	Kelly Stratton
8:35-9:25	**Types of Proteomics**	**Paul Piehowski & David Degnan**
9:25-9:35	Networking Break	
9:35-10:40	Typical Statistical Processing	Kelly Stratton
10:40-10:50	Networking Break	
10:50-11:40	Biological Interpretation	David Degnan & Tyler Sagendorf
11:40-11:45	Closing Remarks	David Degnan

EMSL

Introduction to MS-Based Proteomics

Paul Piehowski
Scientist IV, Team Lead
Biomolecular Pathways

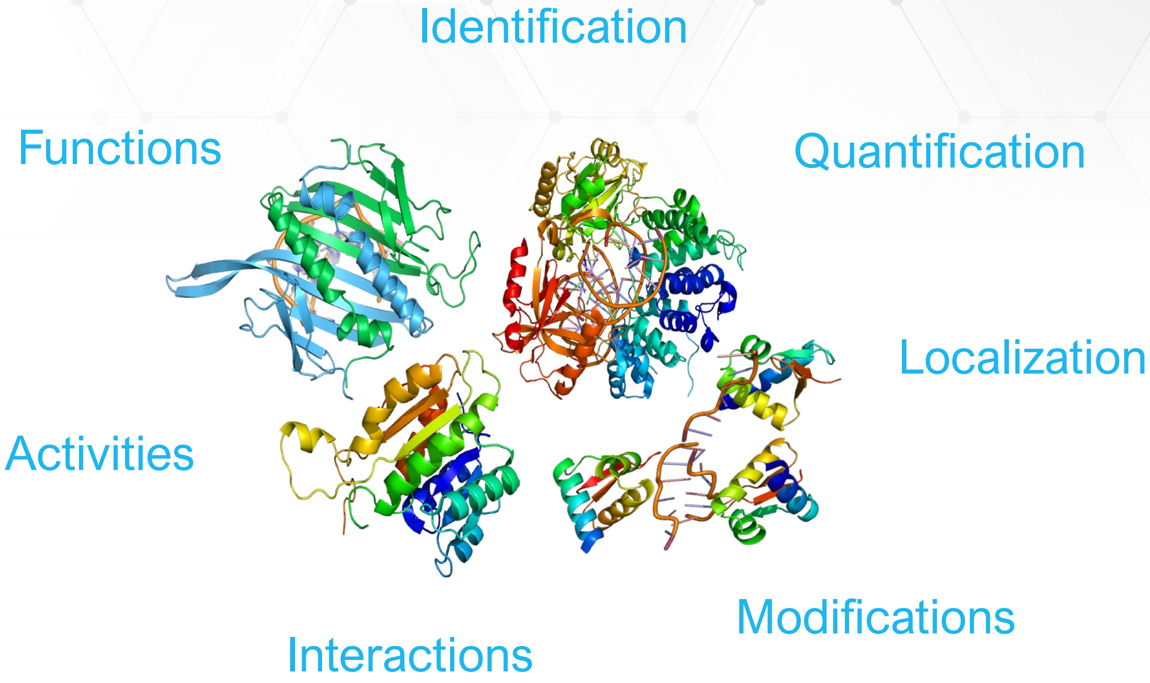
What are we going to talk about today?

- Primer on proteomics and mass spectrometry
- Bottom-up proteomics
  - Understanding bottom-up proteomics
  - Quantification
  - Discovery approaches
    - Global quantification
    - PTM’s
    - Spatial and Single Cell
    - Metabolic Labeling
  - Targeted approaches
- Top-down proteomics
  - Intact
  - Native

What is Proteomics?

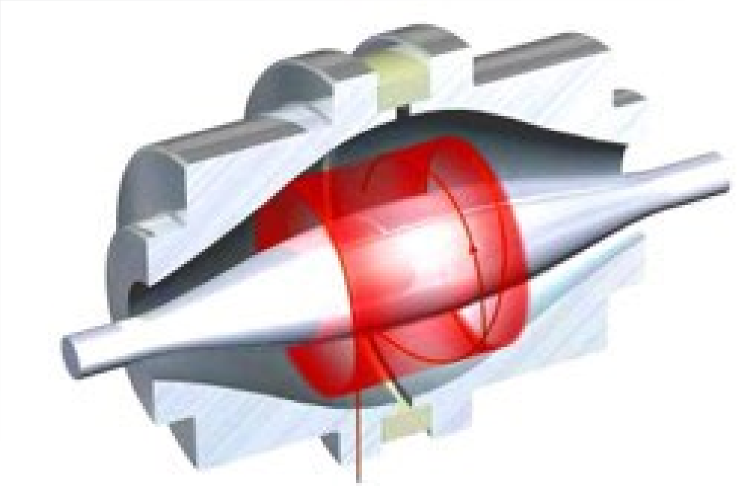
Proteomics is the large-scale study of proteins that are, or can be, expressed by a genome, cell, tissue, or organism at a certain time.

- Techniques for proteomics include:
  - ➢ Mass spectrometry (MS)
  - ➢ Nuclear magnetic resonance (NMR)
  - ➢ Light and electron microscopy
  - ➢ Fourier transform infrared spectroscopy
  - ➢ Others



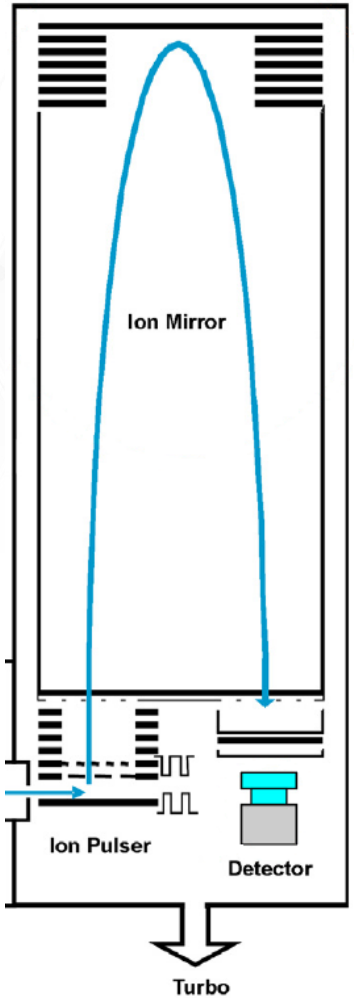
What is Mass Spectrometry – Overview

Orbitrap



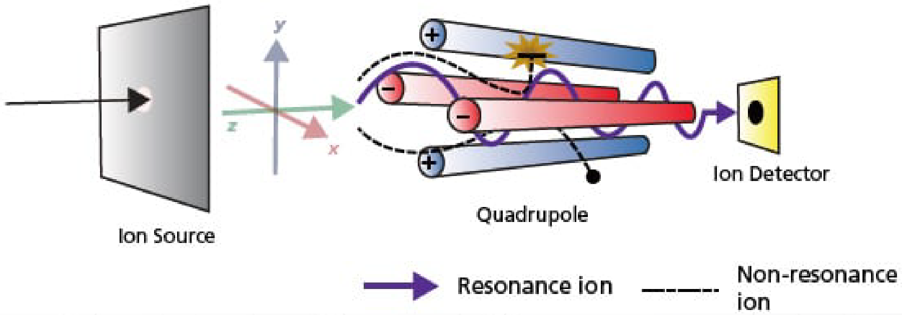
https://analyteguru-prod.s3.amazonaws.com/uploads/2013/10/intact-monoclonal-antibody-characterization-using-a-bench-top-orbitrap-mass-spectrometer.jpg

Time-of-Flight

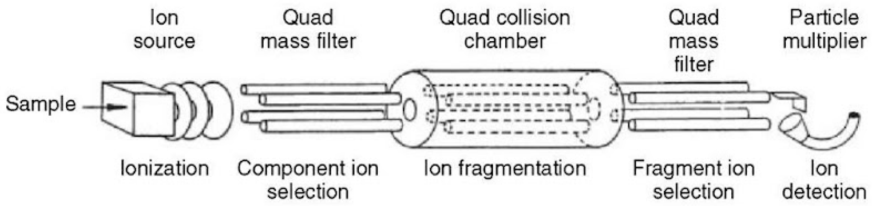


https://www.creative-proteomics.com/images/Agilent-6540-UHD-Quadrupole-Time-of-Flight-Accurate-Mass-Mass-Spectrometer-2.png

Quadrupole(s)



https://www.shimadzu.com/an/service-support/technical-support/analysis-basics/fundamental/mass_analyzers.html

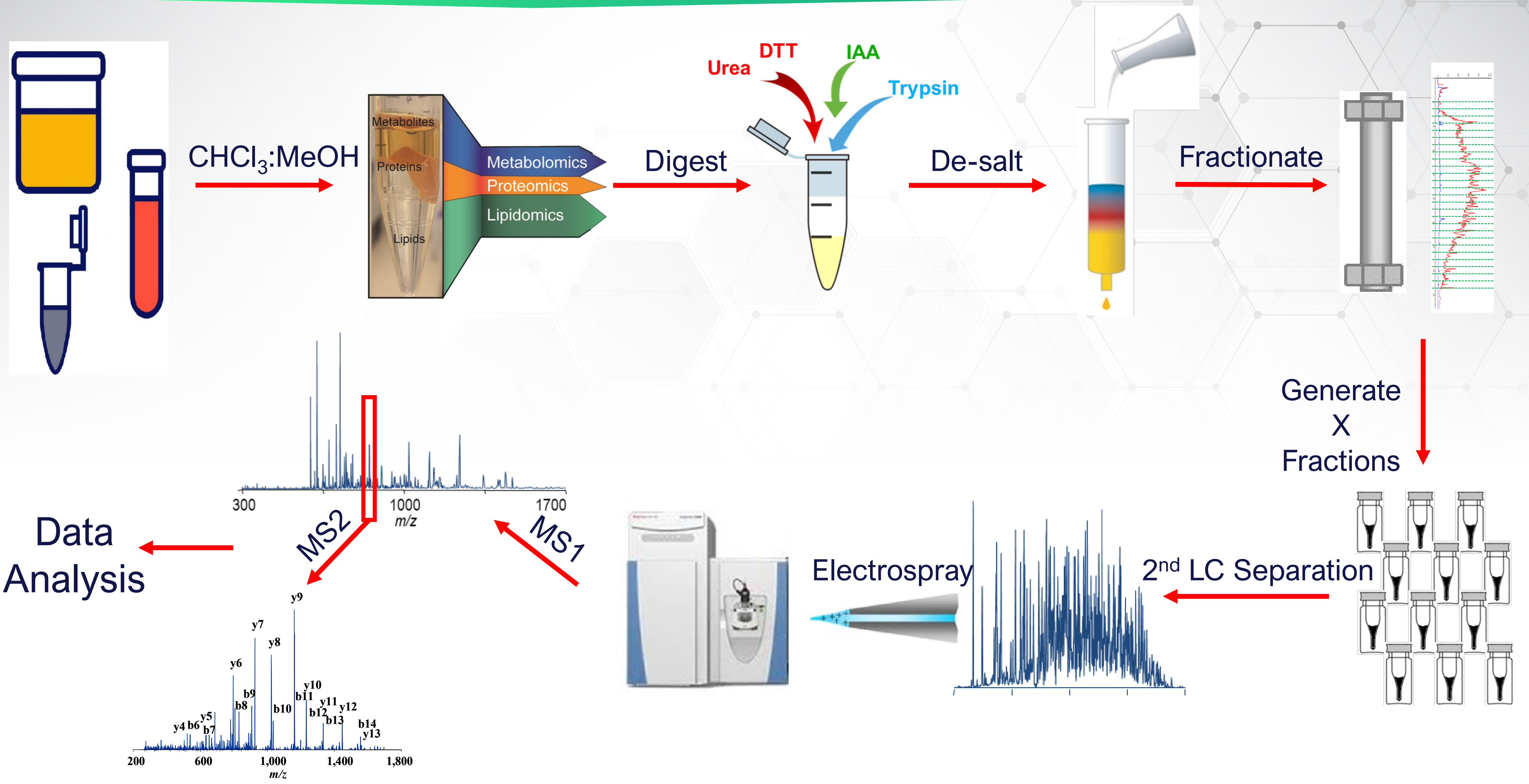


https://what-when-how.com/proteomics/quadrupole-mass-analyzers-theoretical-and-practical-considerations-proteomics/

What are we going to talk about today?

- Primer on proteomics and mass spectrometry
- Bottom-up proteomics
  - Understanding bottom-up proteomics
  - Quantification
  - Discovery approaches
    - Global quantification
    - PTM’s
    - Spatial and Single Cell
    - Metabolic Labeling
  - Targeted approaches
- Top-down proteomics
  - Intact
  - Native

The Making of Bottom-up Proteomics Data



$CHCl_3$:MeOH

Metabolomics

Proteomics

Lipidomics

Digest

Urea

DTT

IAA

Trypsin

De-salt

Fractionate

Generate
X
Fractions

2^{nd} LC Separation

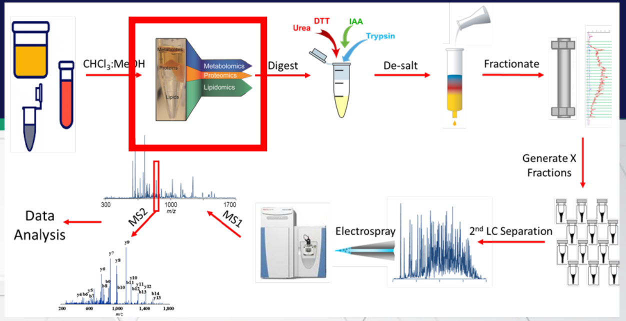
Electrospray

MS1

MS2

Data Analysis

MPLEx-Extracts and Partitions Biomolecules



Extract

- Lyse the sample (if necessary) in water
- Add cold (-20°C) chloroform/methanol (2:1 ,v/v) to sample in 5:1 ratio over sample volume.

Isolate

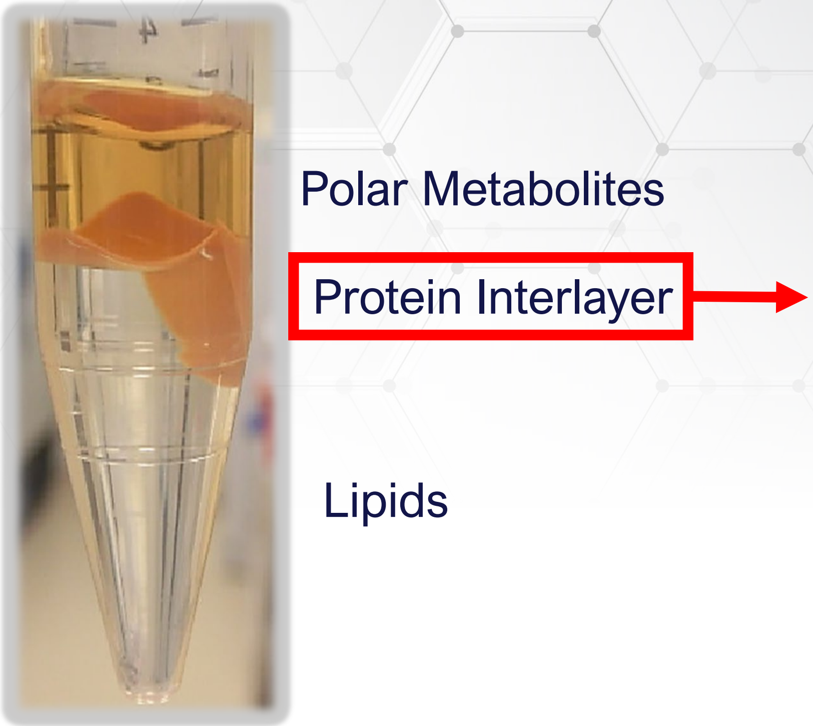
- Let stand on ice for 5 min, vortex
- Centrifuge at 12,000 rpm for 10 min at 4°C

Collect

- Collect the upper layer (metabolites)
- Collect the lower layer (lipids)
- Dry the protein interlayer

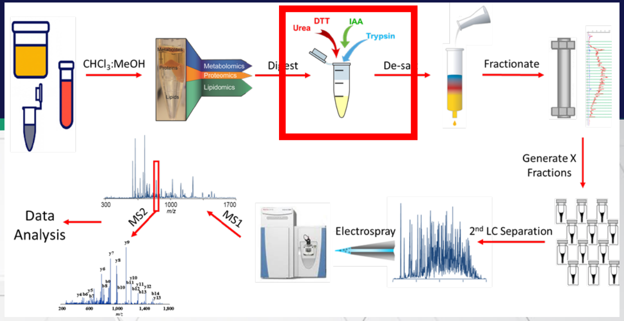
Digest

- Re-solubilize the protein pellet in 8M urea with sonication (add powder urea for typical global digestion)
- BCA assay, add 10mM DTT, incubate 60°C for 30 min
- Digest with trypsin
- C-18 SPE clean-up



Nakayasu *et al*. mSystems. 2016;1(3) & Burnum-Johnson *et al*. Analyst, 2017, 142, 442-448

Protein-to-Peptide

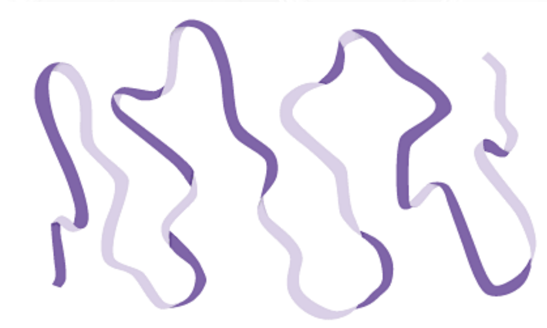


Folded Protein

Denaturation

Disrupt structure and reduce di-sulfide bonds

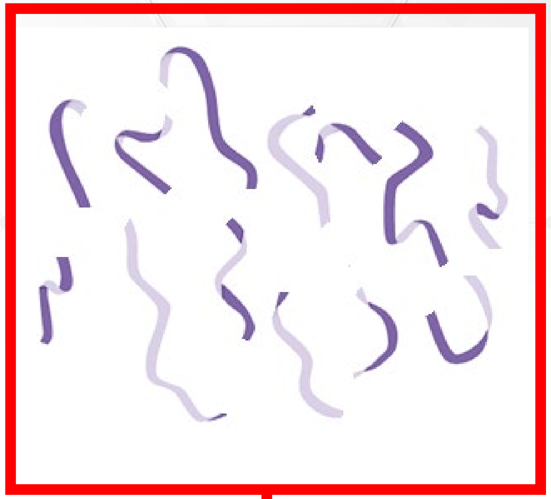
Unfolded Protein



Trypsin Digestion

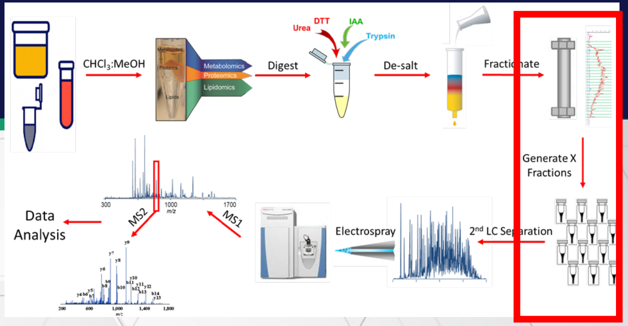
Cleaves at K and R

Peptides

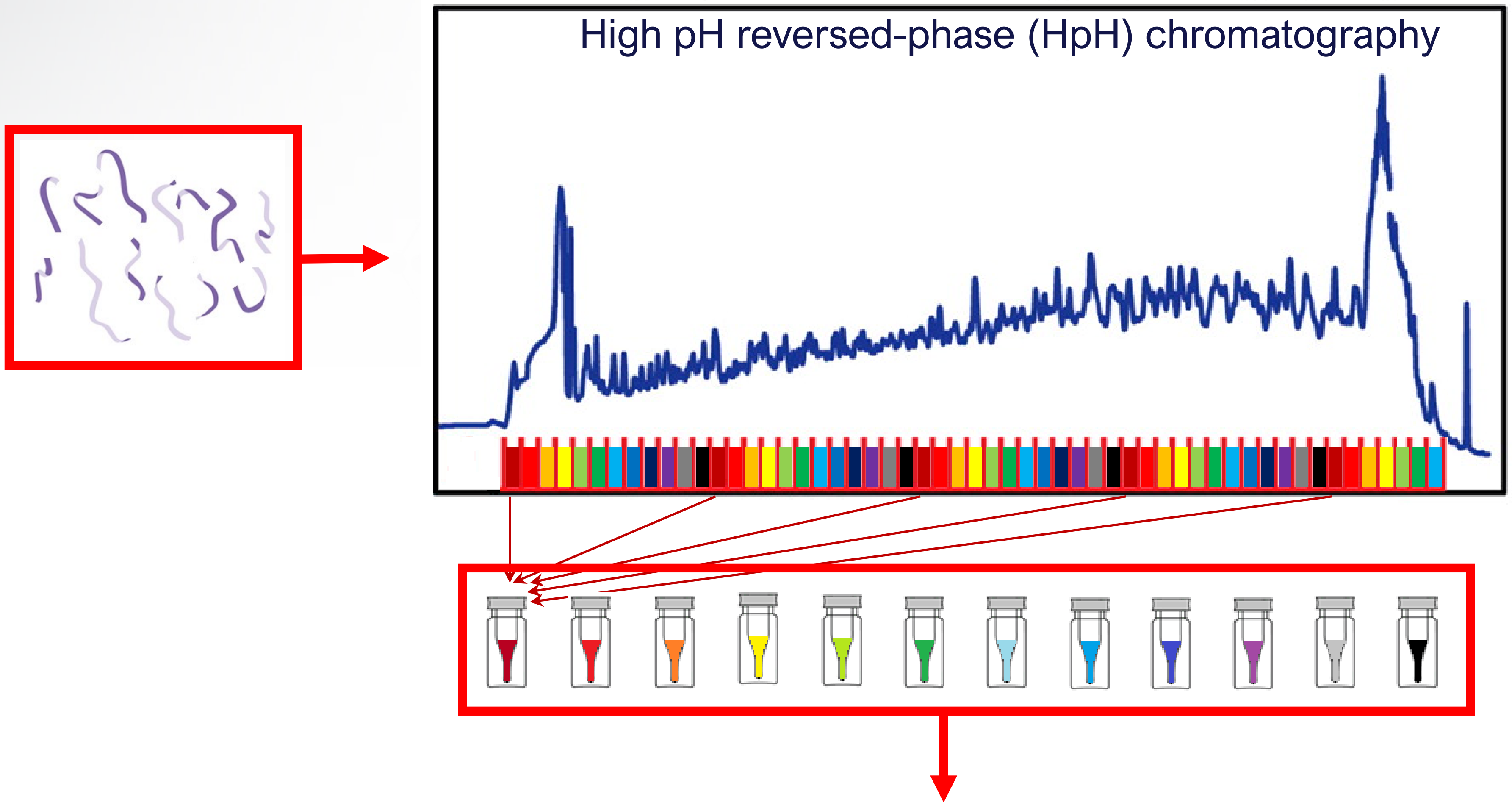


http://commons.wikimedia.org/wiki/User:Debstar

Fractionation Reduces Complexity

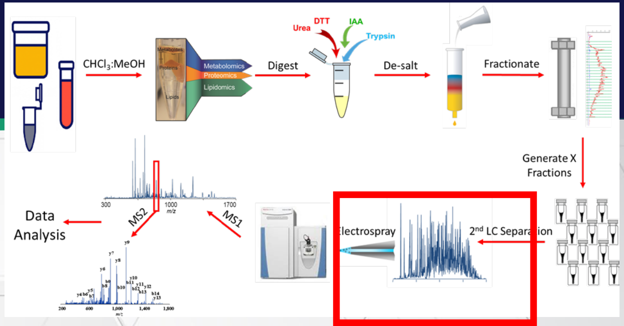


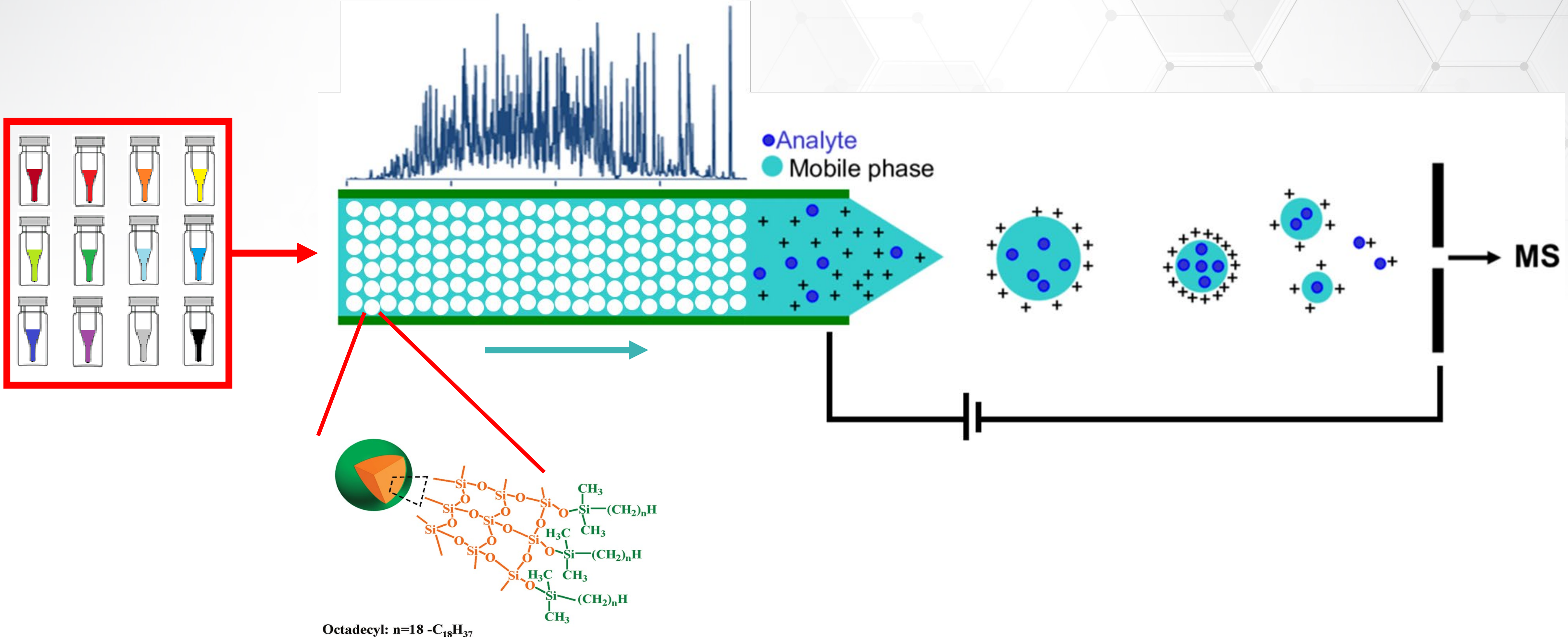
Many Species x Many Proteins x Tryptic Digestion = ***Complexity***



High pH reversed-phase (HpH) chromatography

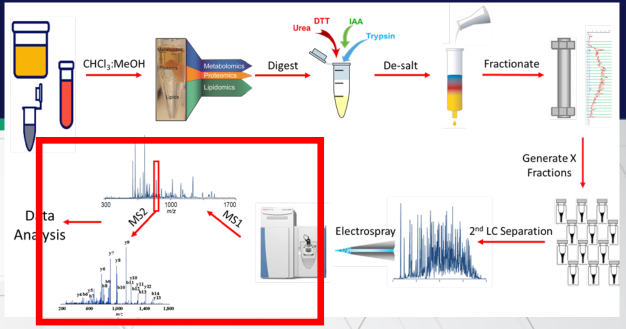
Liquid Chromatography Electrospray Ionization (LC-ESI)

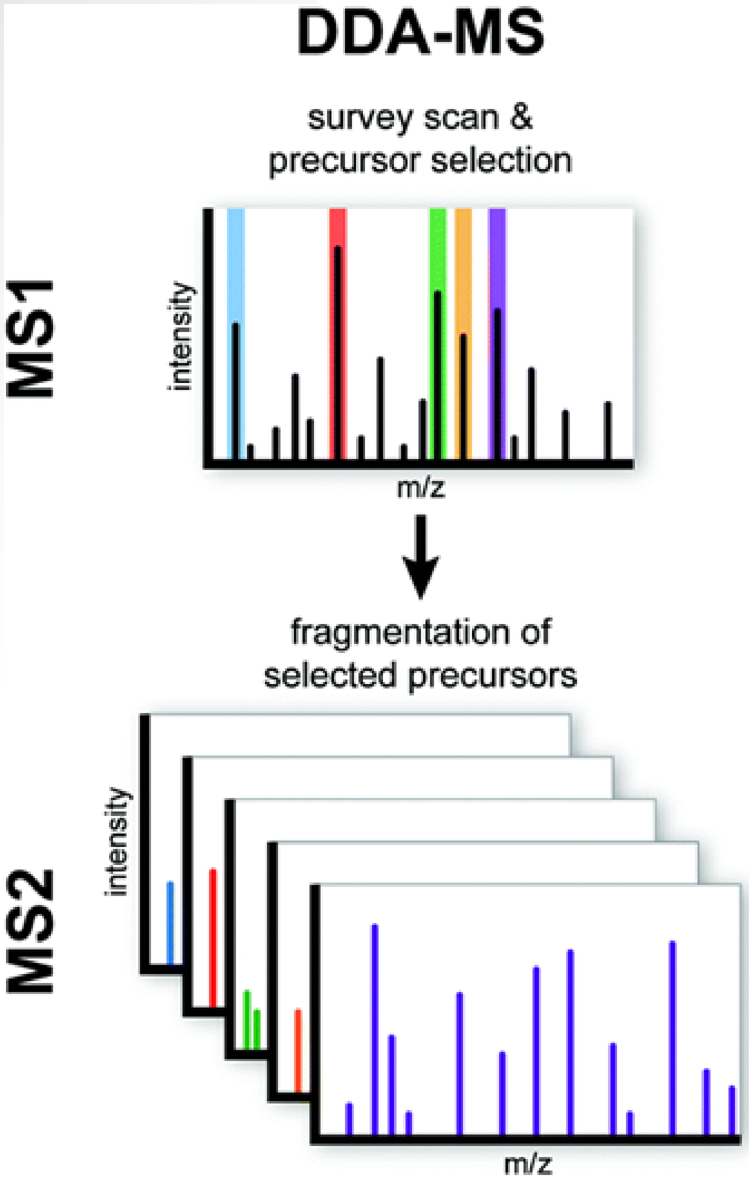




Octadecyl: n=18 -$C_{18}H_{37}$

Data Dependent Acquisition (DDA)

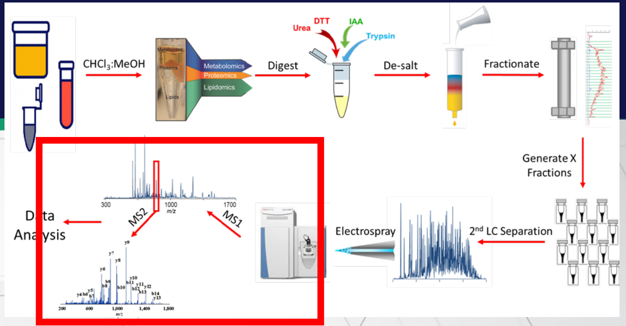


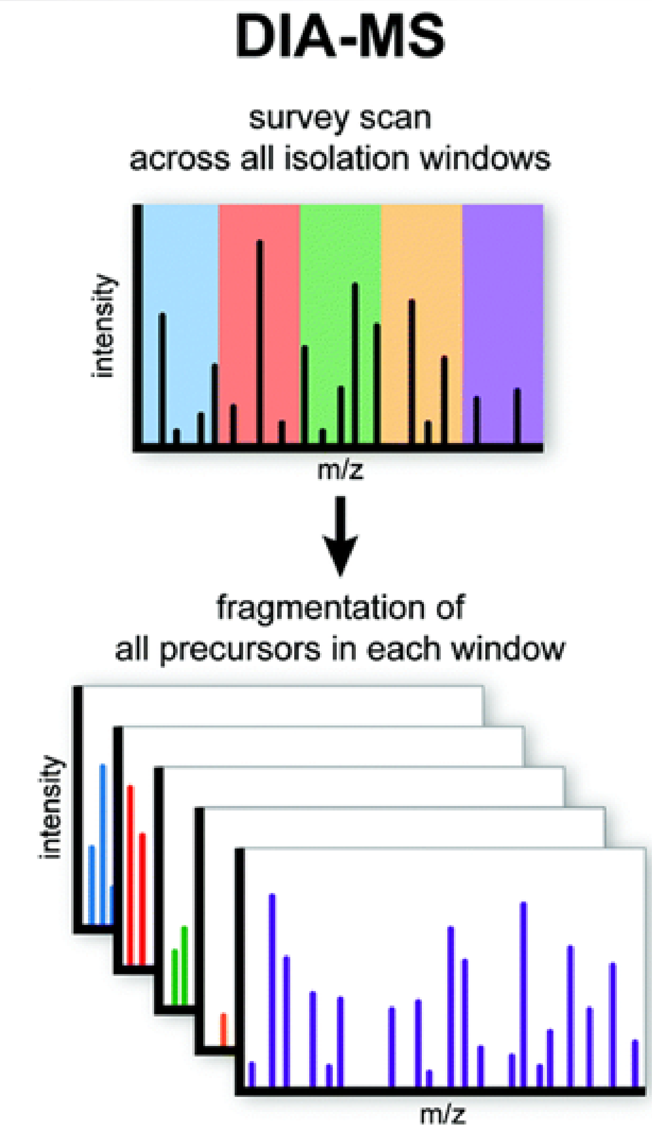

DDA-MS
survey scan & precursor selection
MS1
intensity
m/z
fragmentation of selected precursors
MS2
intensity
m/z

- Called data dependent because the ions/peptides selected for fragmentation is **dependent** on the MS1 **data**
- The most abundant ions from the MS1 are isolated for fragmentation serially to produce an MS2 fragmentation spectra of a single peptide
- This process is repeated n number of times, and then another MS1 spectrum is taken and the process repeats

DOI: 10.1039/D0MO00072H (Review Article) *Mol. Omics*, 2021, **17**, 29-42

Data Independent Acquisition (DIA)

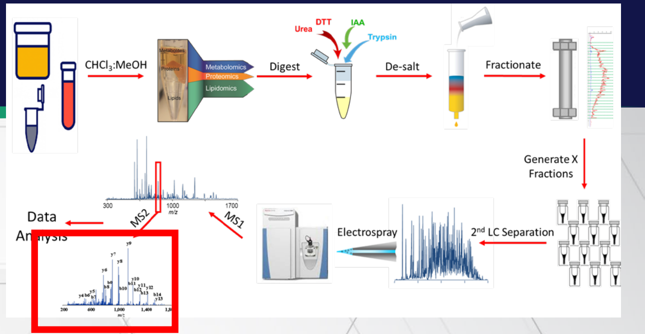




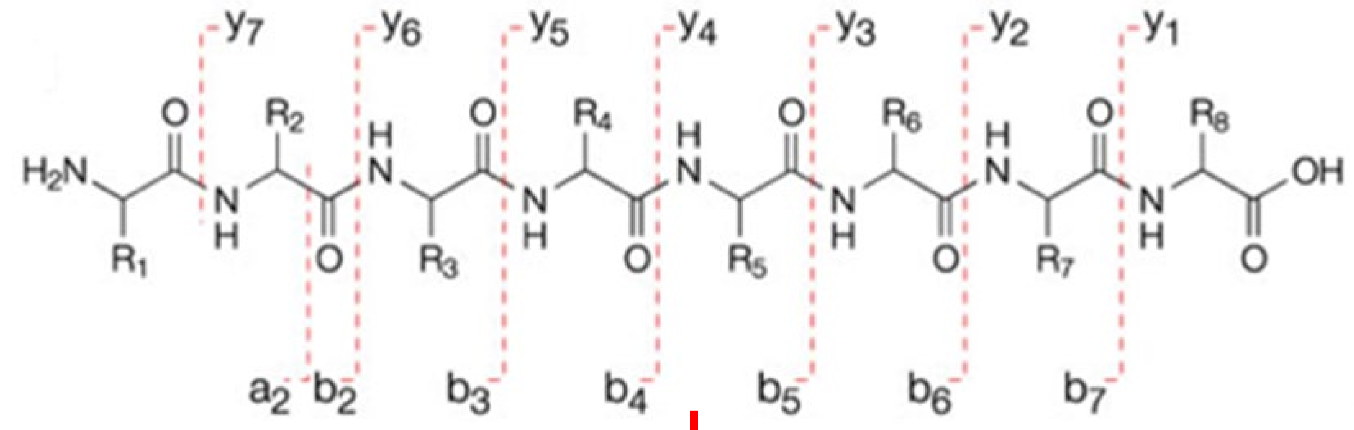
- Called data independent because the ions/peptides selected for fragmentation is **independent** on the MS1 **data**
- Instead, **regions** of the MS1 spectrum are isolated for fragmentation serially following a user defined pattern producing MS2 fragmentation of multiple peptides
- This process is repeated n number of times to cover the desired m/z space

DOI: 10.1039/D0MO00072H (Review Article) *Mol. Omics*, 2021, **17**, 29-42

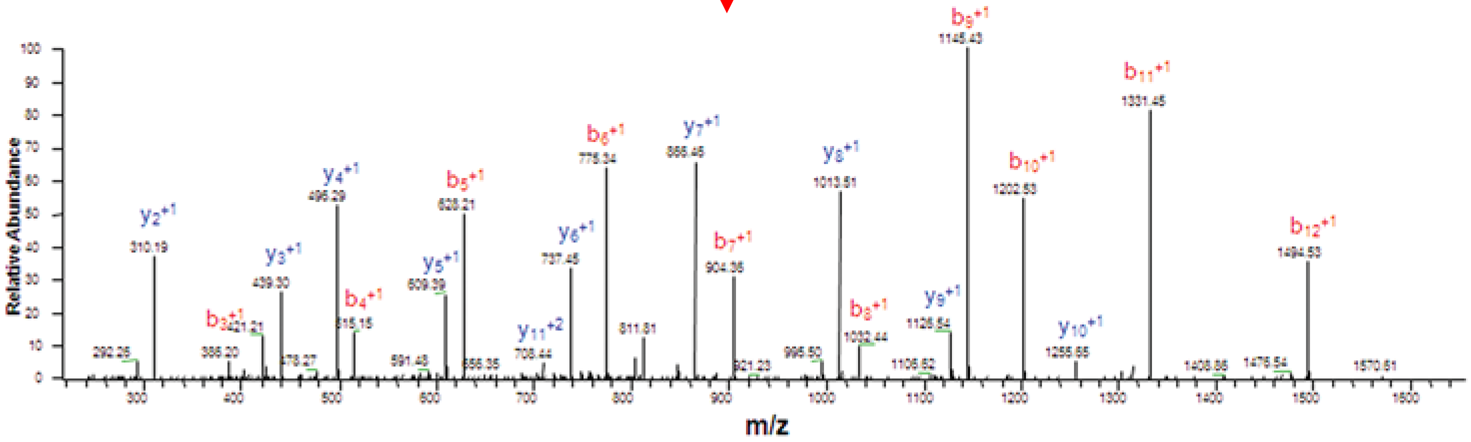
Peptide Fragmentation (MS2)



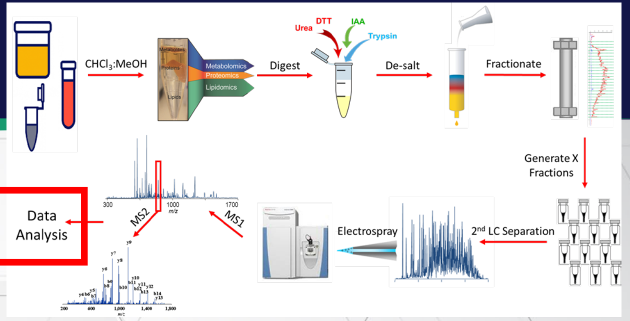
N-Terminus



C-Terminus



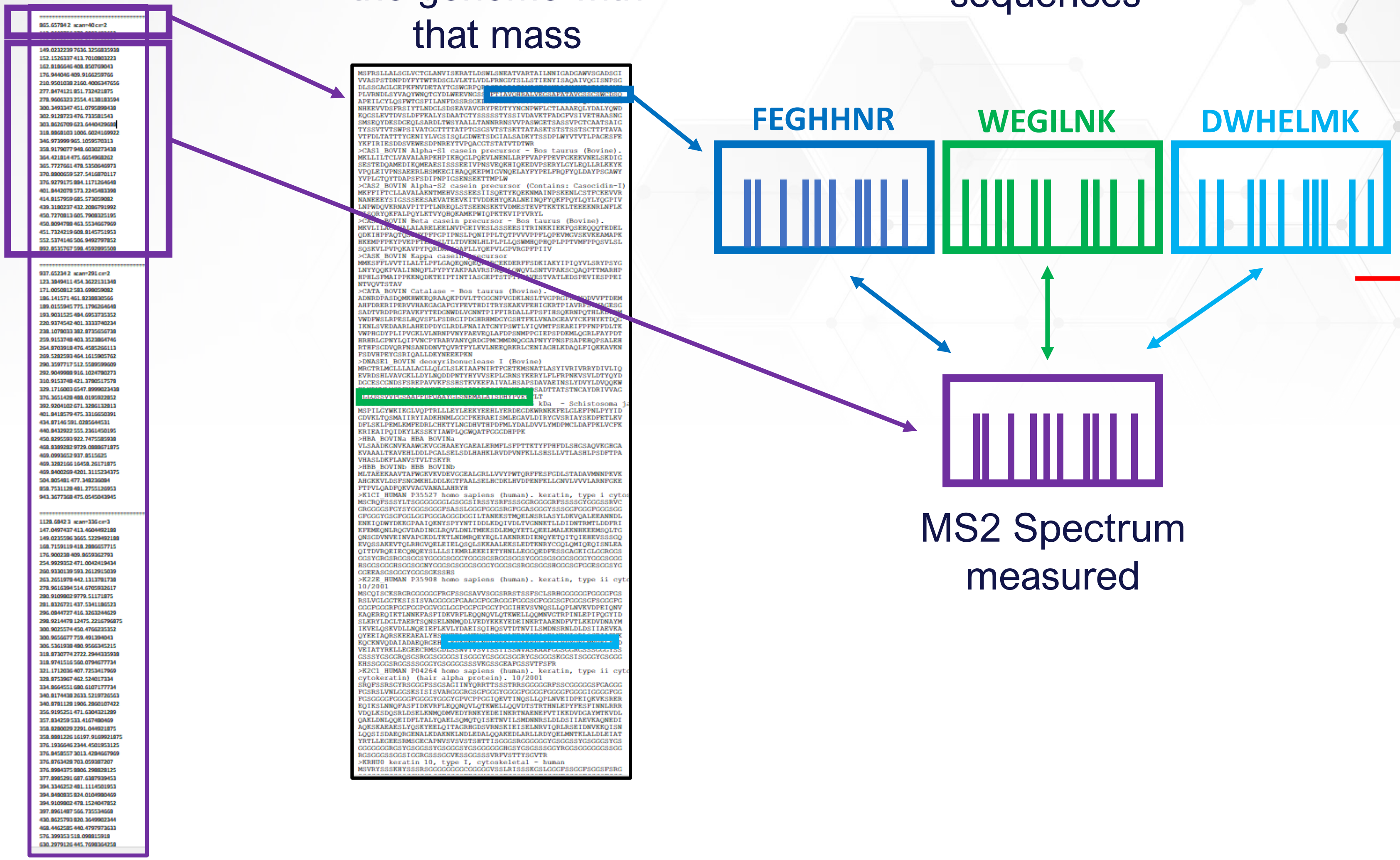
Identifying Peptides From MS2 Spectra



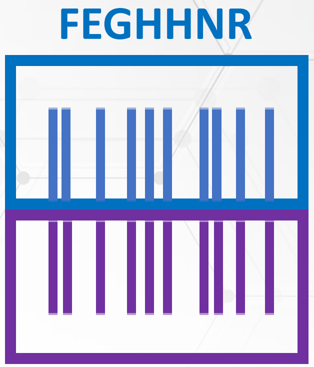
List of masses selected for MS2

Peptide sequences from the genome with that mass

Theoretical spectra for predicted peptide sequences



MS2 Spectrum measured



Output table of matches

Mass	PeptideSequence	ConfidenceScores
Mass1	Sequence1	CS1
Mass2	Sequence2	CS2
Mass3	Sequence3	CS3
Mass4	Sequence4	CS4
Mass5	Sequence5	CS5
Mass6	Sequence6	CS6
Mass7	Sequence7	CS7
…..	……	……

What are we going to talk about today?

- Primer on proteomics and mass spectrometry
- Bottom-up proteomics
  - Understanding bottom-up proteomics
  - Quantification
  - Discovery approaches
    - Global quantification
    - PTM's
    - Spatial and Single Cell
    - Metabolic Labeling
  - Targeted approaches
- Top-down proteomics
  - Intact
  - Native

Bottom-up Proteomics - Approaches

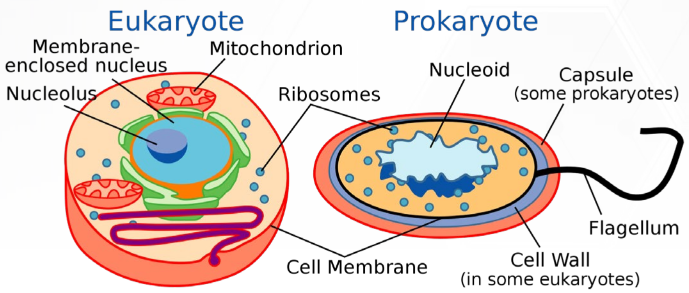
There are two main bottom-up proteomics approaches for mass spectrometry

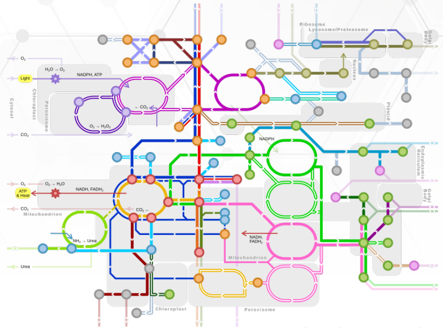
Discovery Proteomics

- Whole proteomes
- High to moderate abundance proteins
- Identification and quantification

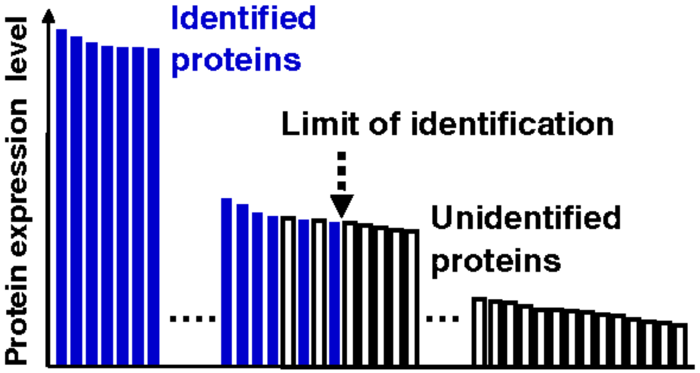
Targeted Proteomics

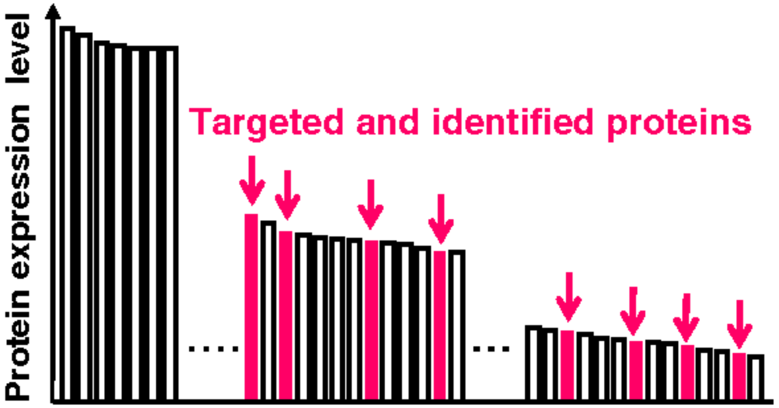
- Selected proteins
- High to low abundance proteins
- Quantification


Eukaryote
Prokaryote
Membrane-enclosed nucleus
Mitochondrion
Nucleolus
Ribosomes
Nucleoid
Capsule (some prokaryotes)
Flagellum
Cell Membrane
Cell Wall (in some eukaryotes)



Metabolic pathway





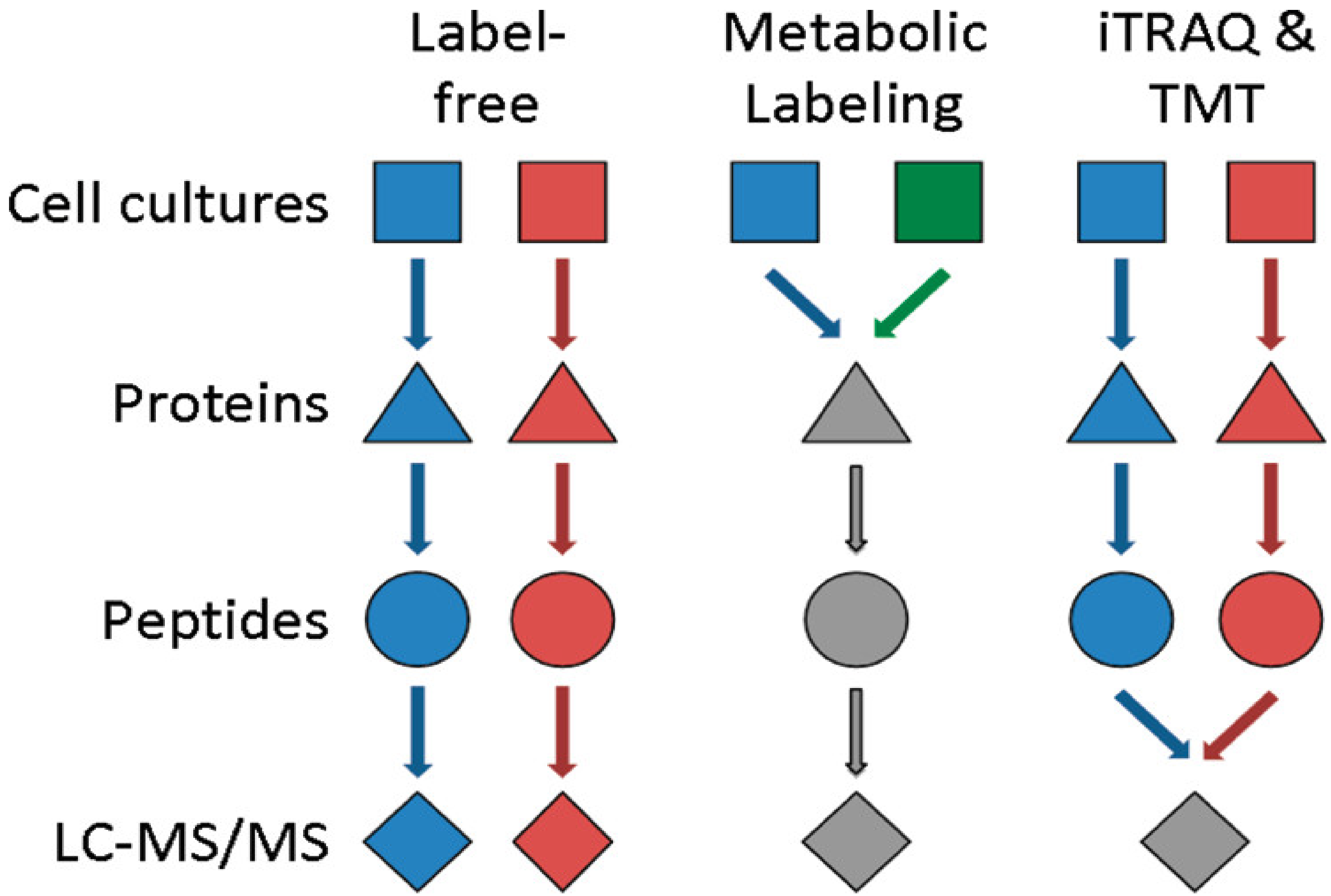
Middle Figures are from Wikipedia

Yasuo Uchida, Masanori Tachikawa, et al. Fluids and Barriers of the CNS. 2013 DOI:10.1186/2045-8118-10-21

What are we going to talk about today?

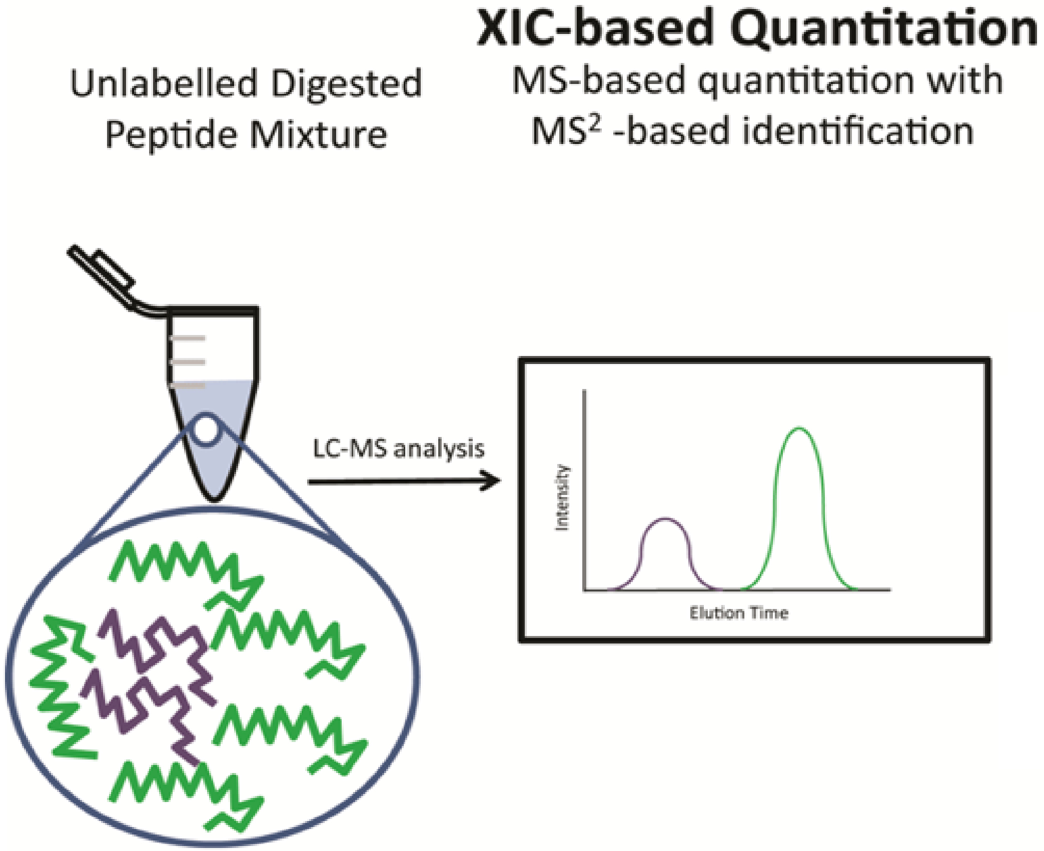
- Primer on proteomics and mass spectrometry
- Bottom-up proteomics
  - Understanding bottom-up proteomics
  - Quantification
  - Discovery approaches
    - Global quantification
    - PTM’s
    - Spatial and Single Cell
    - Metabolic Labeling
  - Targeted proteomics
- Top-down proteomics
  - Intact
  - Native

Quantification types in Proteomics



https://pubs.acs.org/doi/10.1021/pr200748h

Label-free quantification



XIC-based Quantitation

MS-based quantitation with MS^2 -based identification

Unlabelled Digested Peptide Mixture

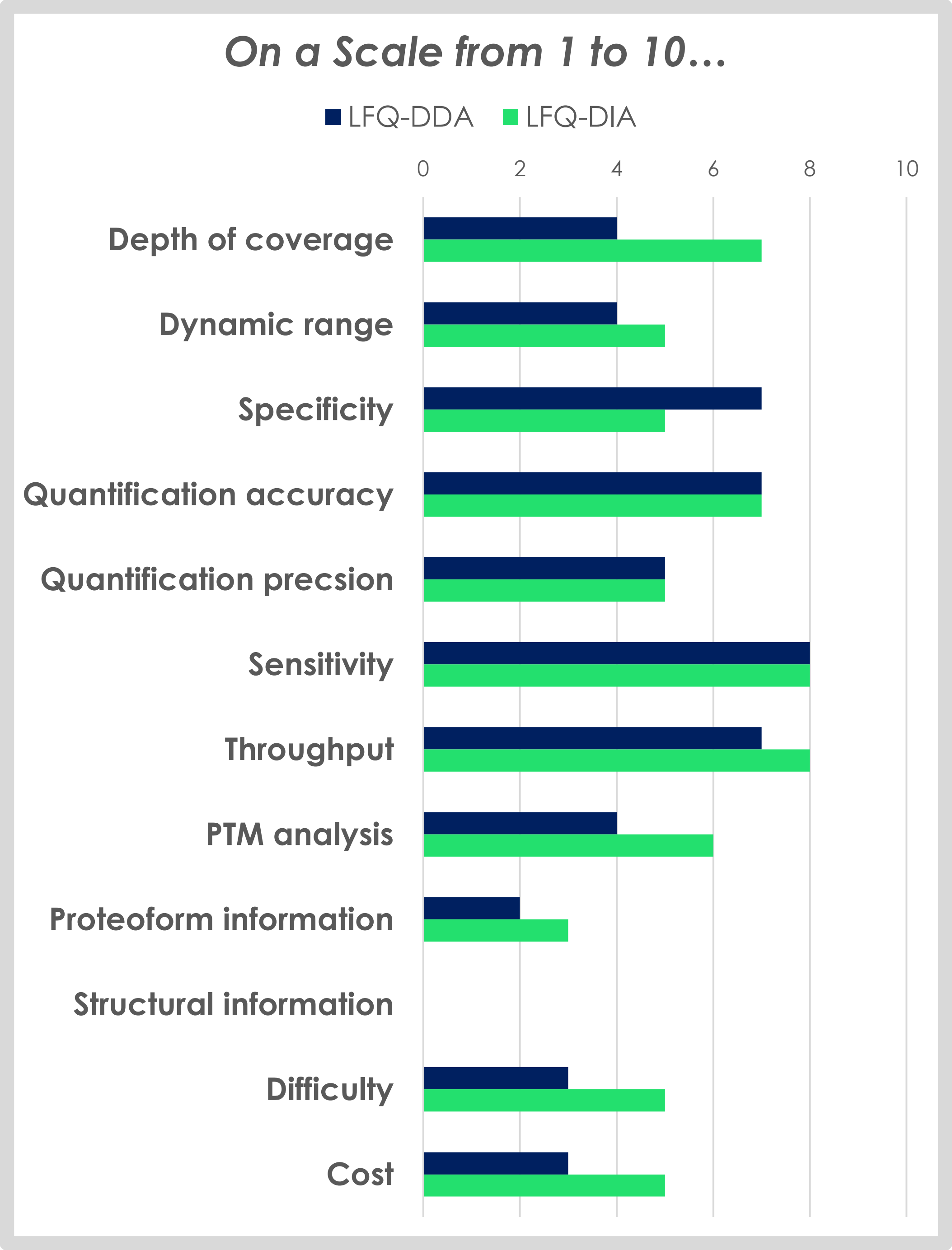
https://www.bioanalysis-zone.com/chapter-5-modern-techniques-in-quantitative-proteomics/

LFQ
DDA VS DIA

LFQ- Label-Free Quantification

DDA- Data Dependent acquisition

DIA- Data Independent Acquisition



On a Scale from 1 to 10…

DDA Strengths

- Lower complexity samples
- Tighter control over FDR
- High confidence
- Easiest

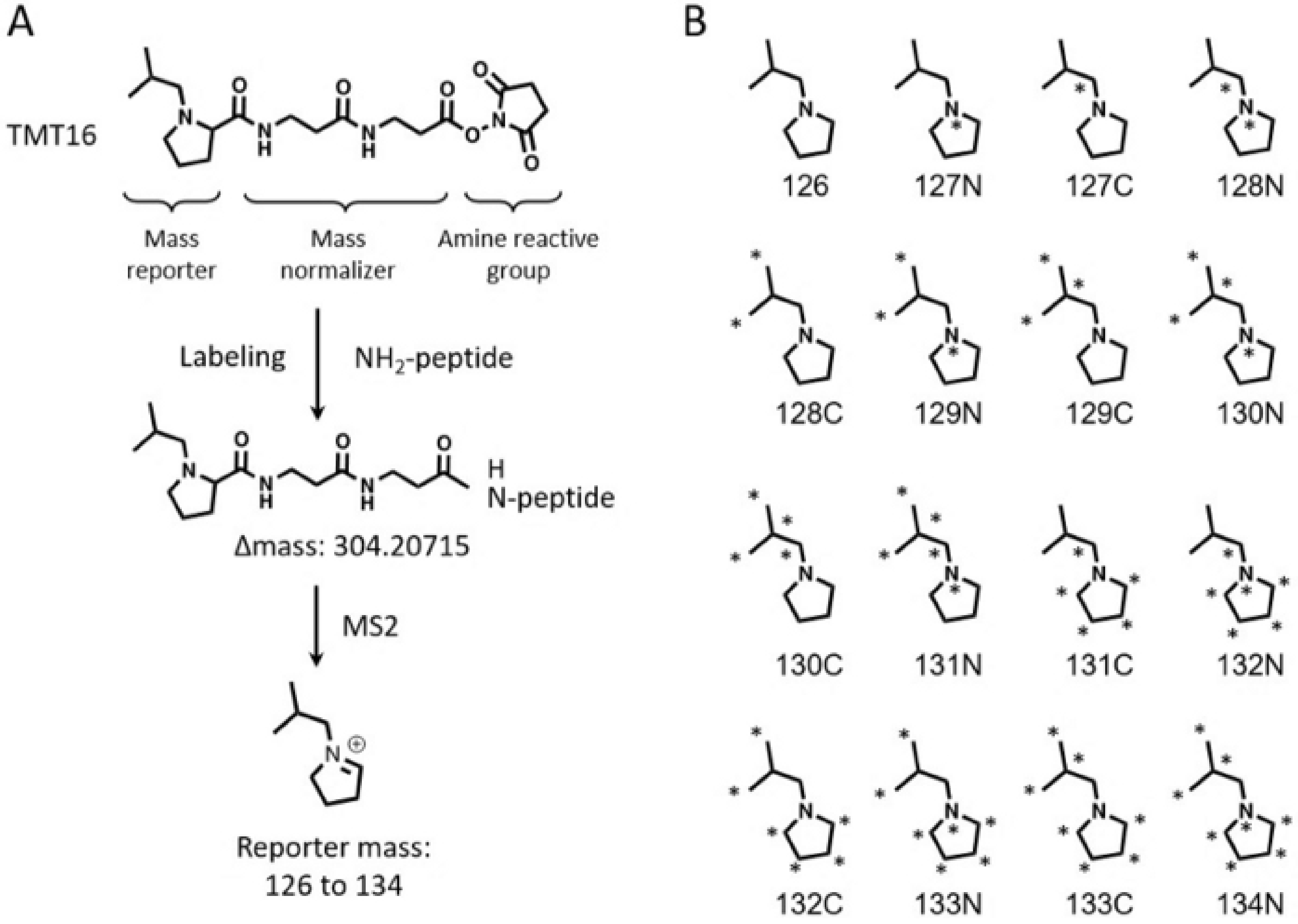
DIA Strengths

- More complex samples
- Larger sample sets

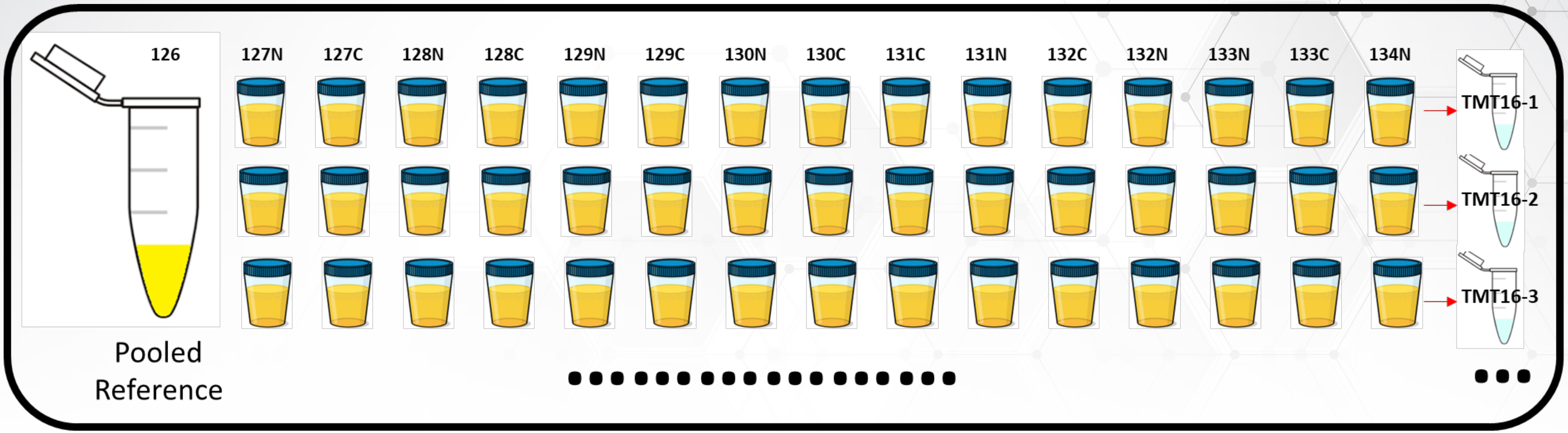
Isobaric (same mass) Labeling

TMT- Tandem Mass Tags

iTRAQ- Isobaric Tag for Relative and Absolute Quantification



Isobaric (same mass) Labeling



126

127N

127C

128N

128C

129N

129C

130N

130C

131C

131N

132C

132N

133N

133C

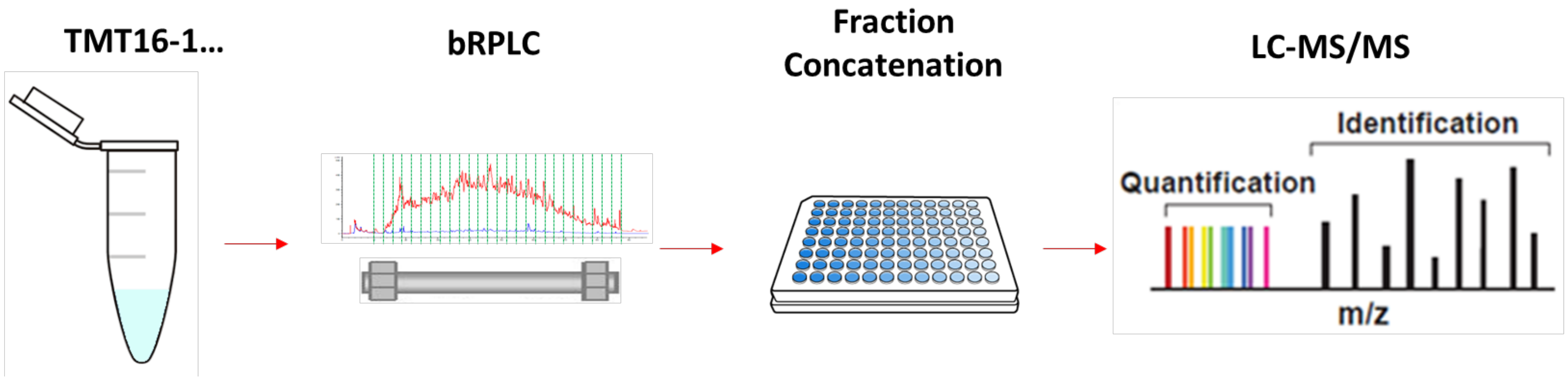
134N

Pooled Reference

TMT16-1

TMT16-2

TMT16-3



TMT16-1...

bRPLC

Fraction Concatenation

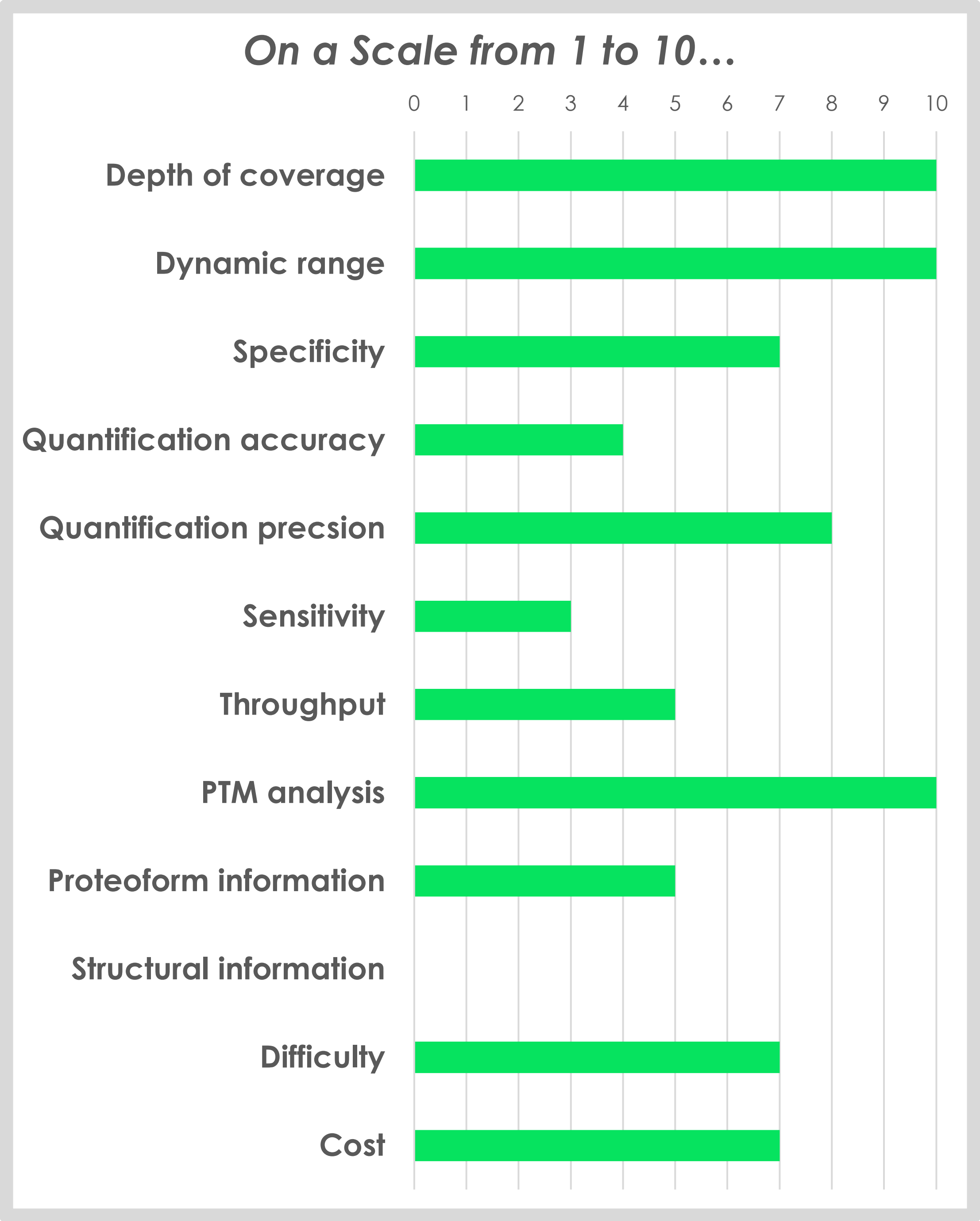
LC-MS/MS

Identification

Quantification

m/z

TMT



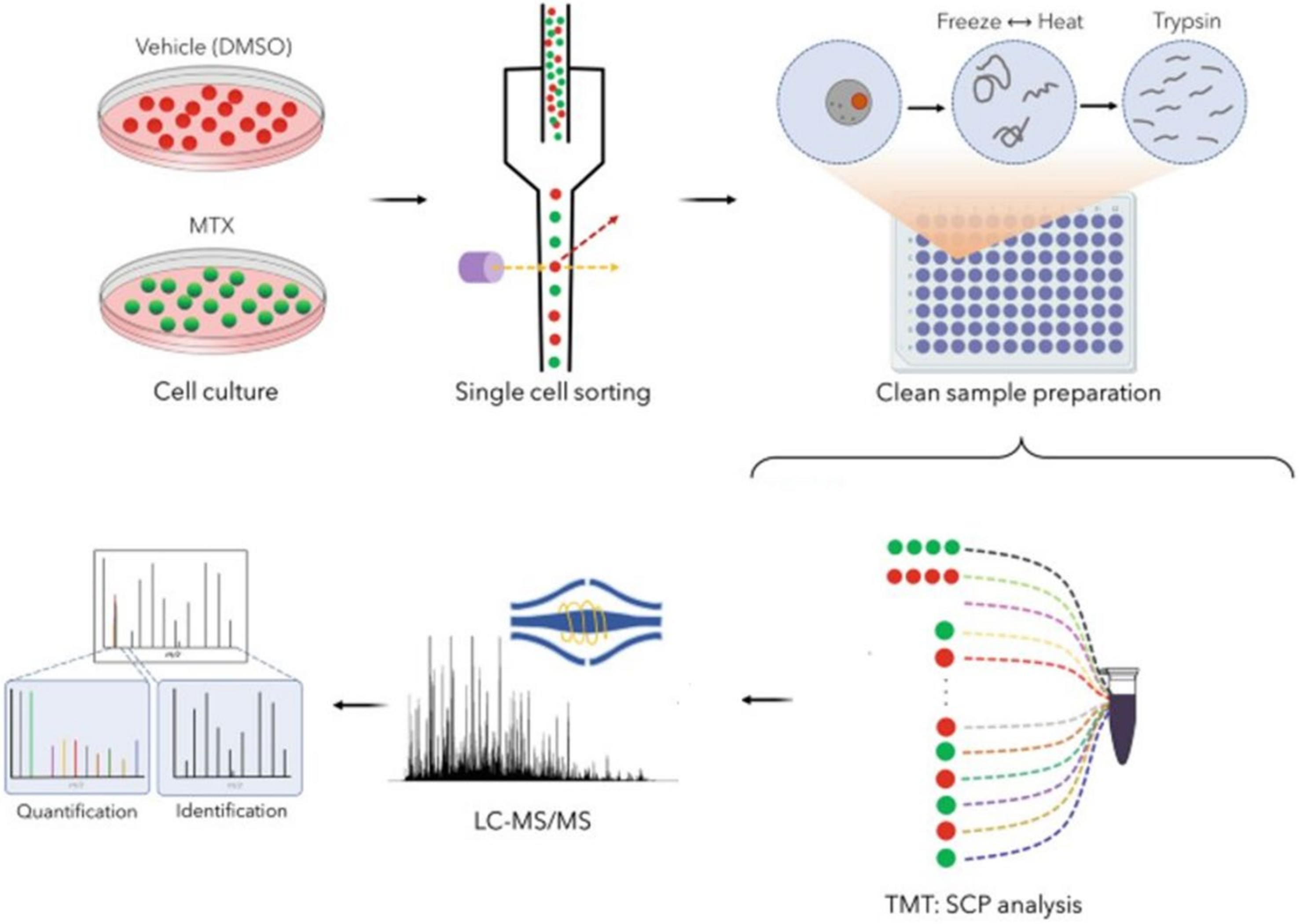
TMT Strengths

- Highest available protein coverage
- Throughput from multiplexing
- Extra Material due to multiplexing

TMT Weaknesses

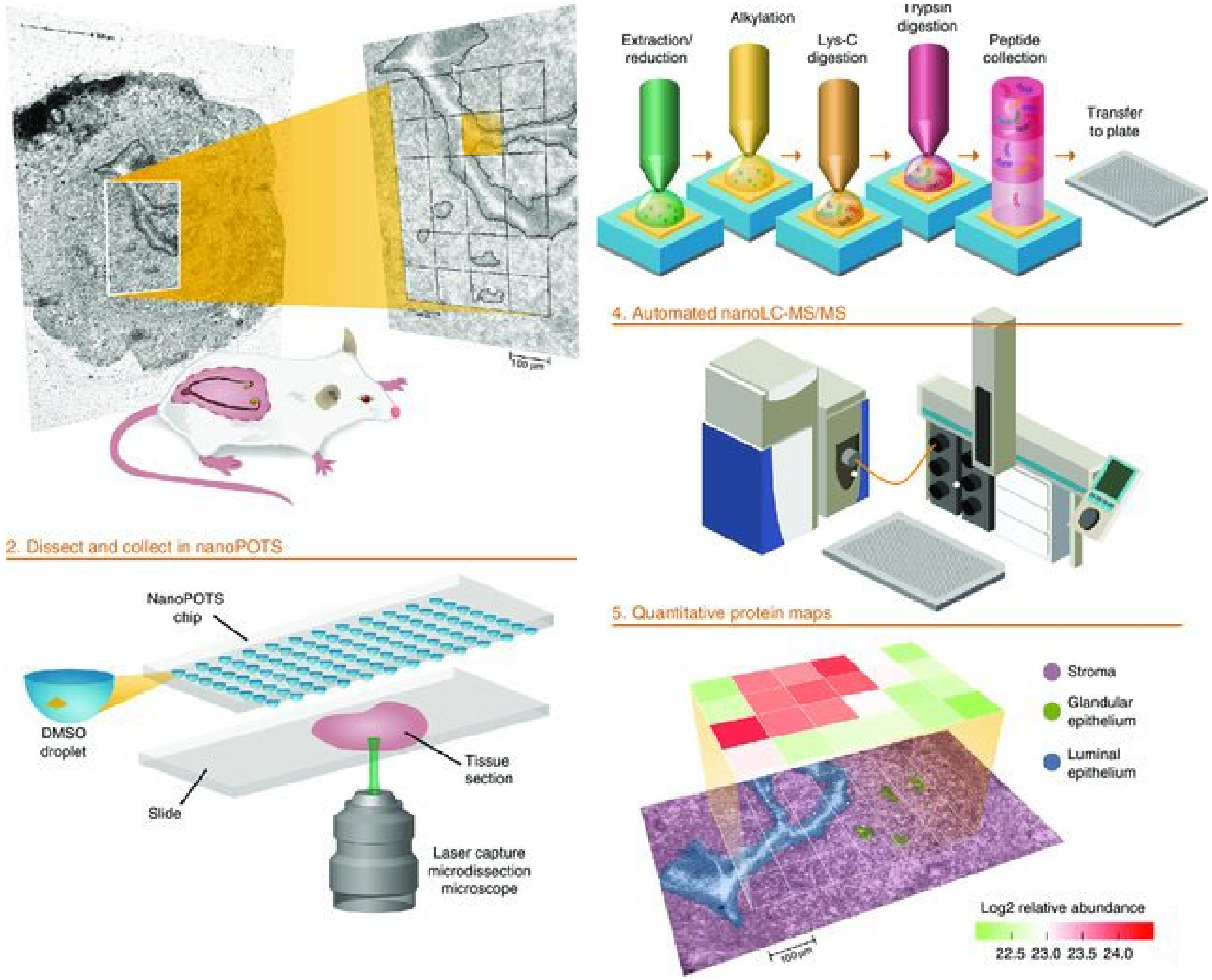
- Laborious sample prep
- Expensive reagents
- Ratio Compression

Single cell proteomics



https://link.springer.com/protocol/10.1007/978-1-0716-1771-7_8

Spatial Proteomics

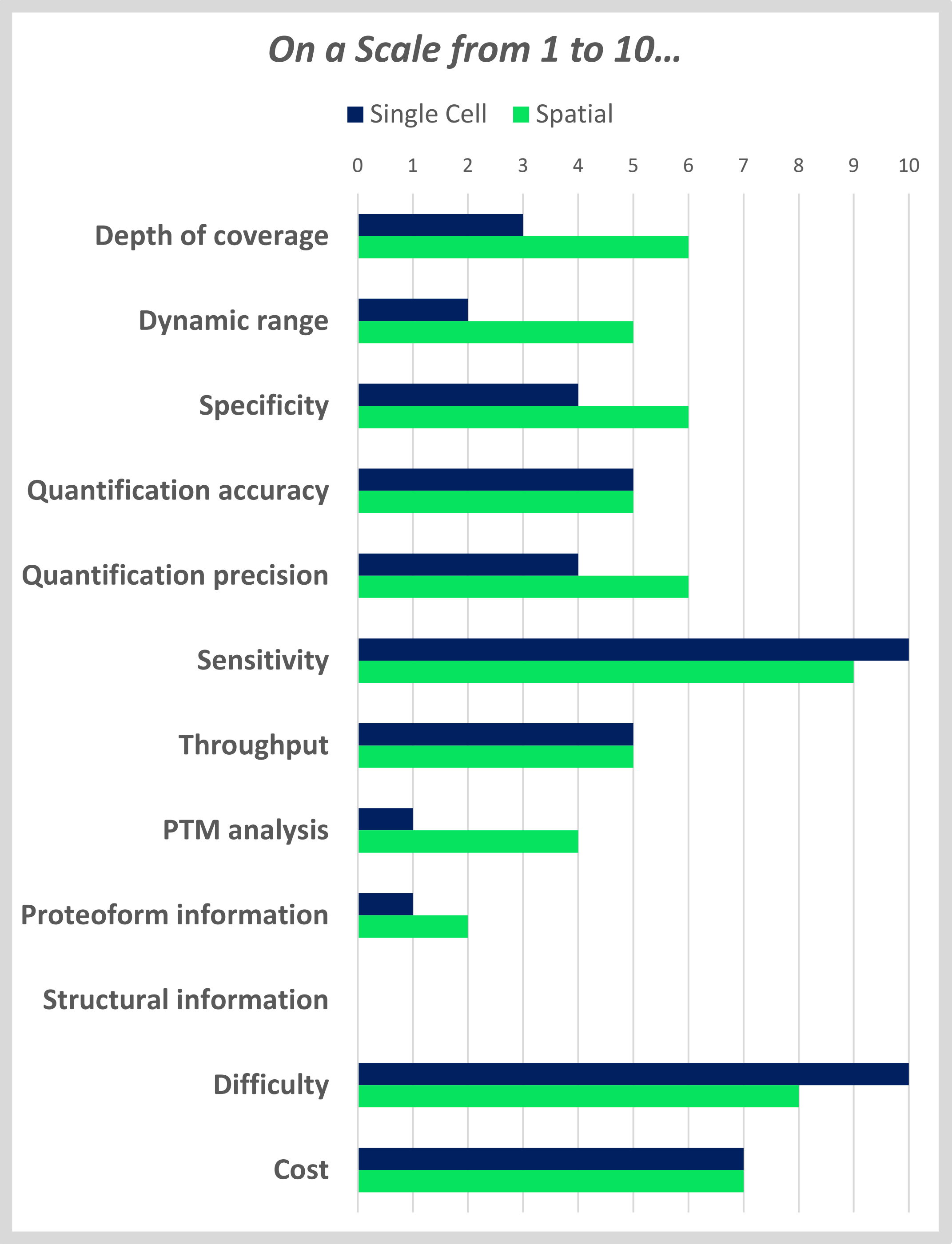


High Sensitivity Proteomics

nanoPOTS- nanodroplet processing in one pot for trace samples

microPOTS- same but microdoplet

SCoPE-MS- Single Cell ProtEomics by Mass Spectrometry



Single Cell Strengths

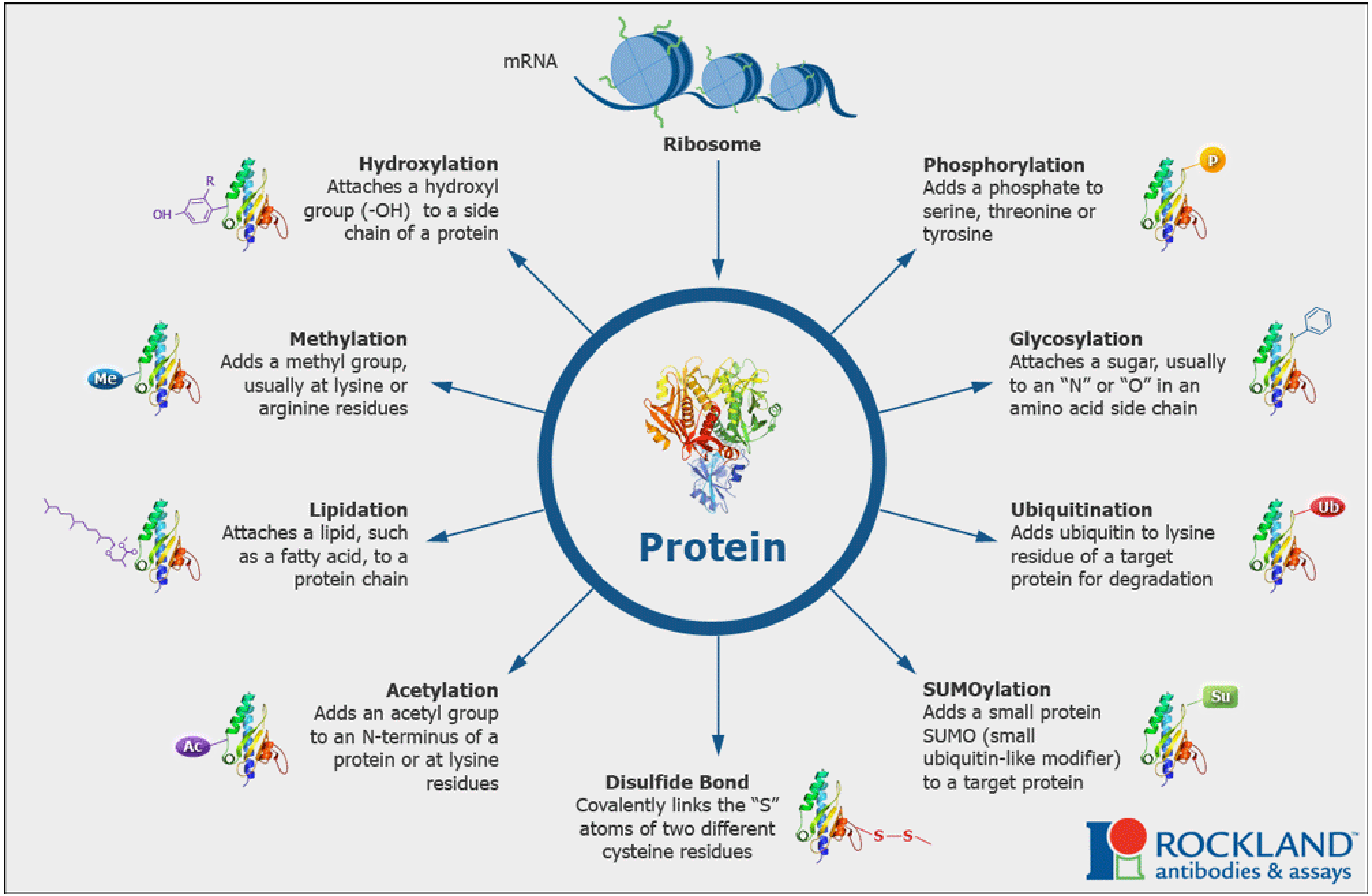
- Single cell information

Spatial Strengths

- Spatial information

Post-Translational Modifications (PTMs)

Post-translational modifications is the chemical modification of a protein after its translations.



Hydroxylation
Attaches a hydroxyl group (-OH) to a side chain of a protein

Phosphorylation
Adds a phosphate to serine, threonine or tyrosine

Methylation
Adds a methyl group, usually at lysine or arginine residues

Glycosylation
Attaches a sugar, usually to an "N" or "O" in an amino acid side chain

Lipidation
Attaches a lipid, such as a fatty acid, to a protein chain

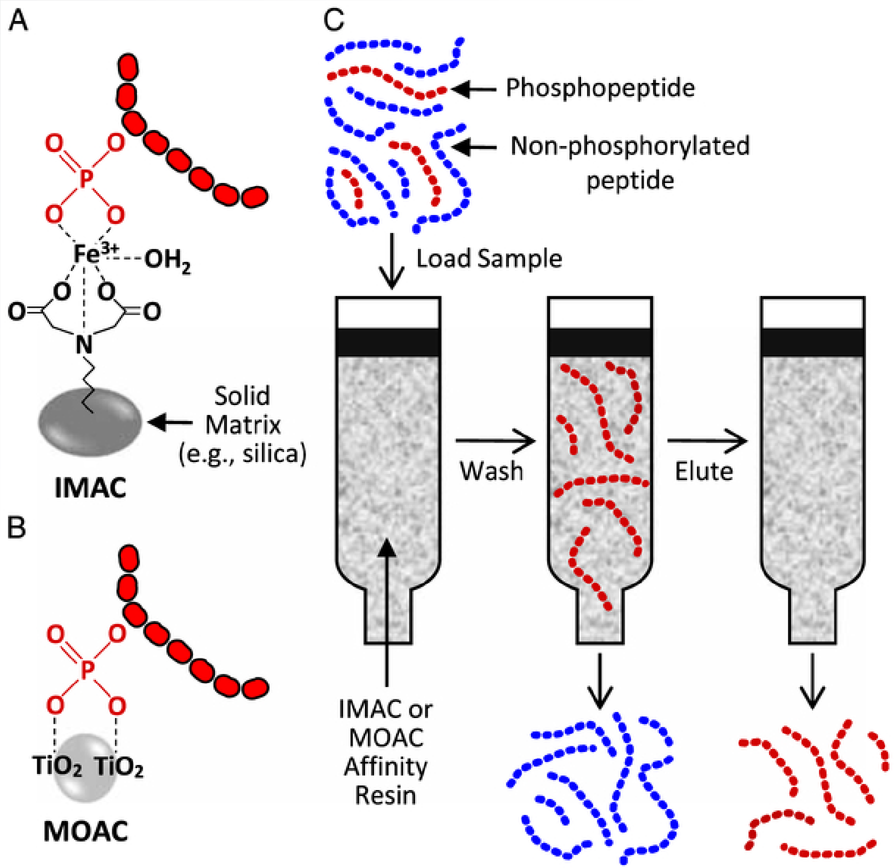
Ubiquitination
Adds ubiquitin to lysine residue of a target protein for degradation

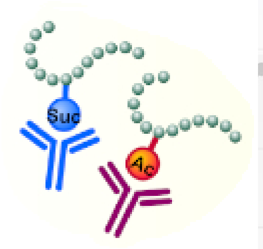
Acetylation
Adds an acetyl group to an N-terminus of a protein or at lysine residues

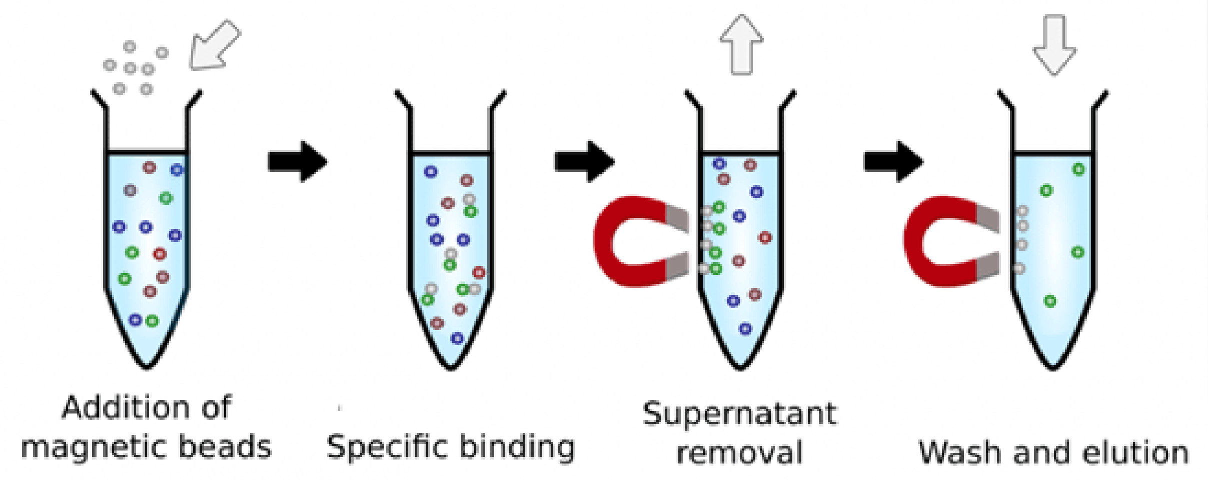
Disulfide Bond
Covalently links the "S" atoms of two different cysteine residues

SUMOylation
Adds a small protein SUMO (small ubiquitin-like modifier) to a target protein

Profiling PTM’s requires an enrichment step





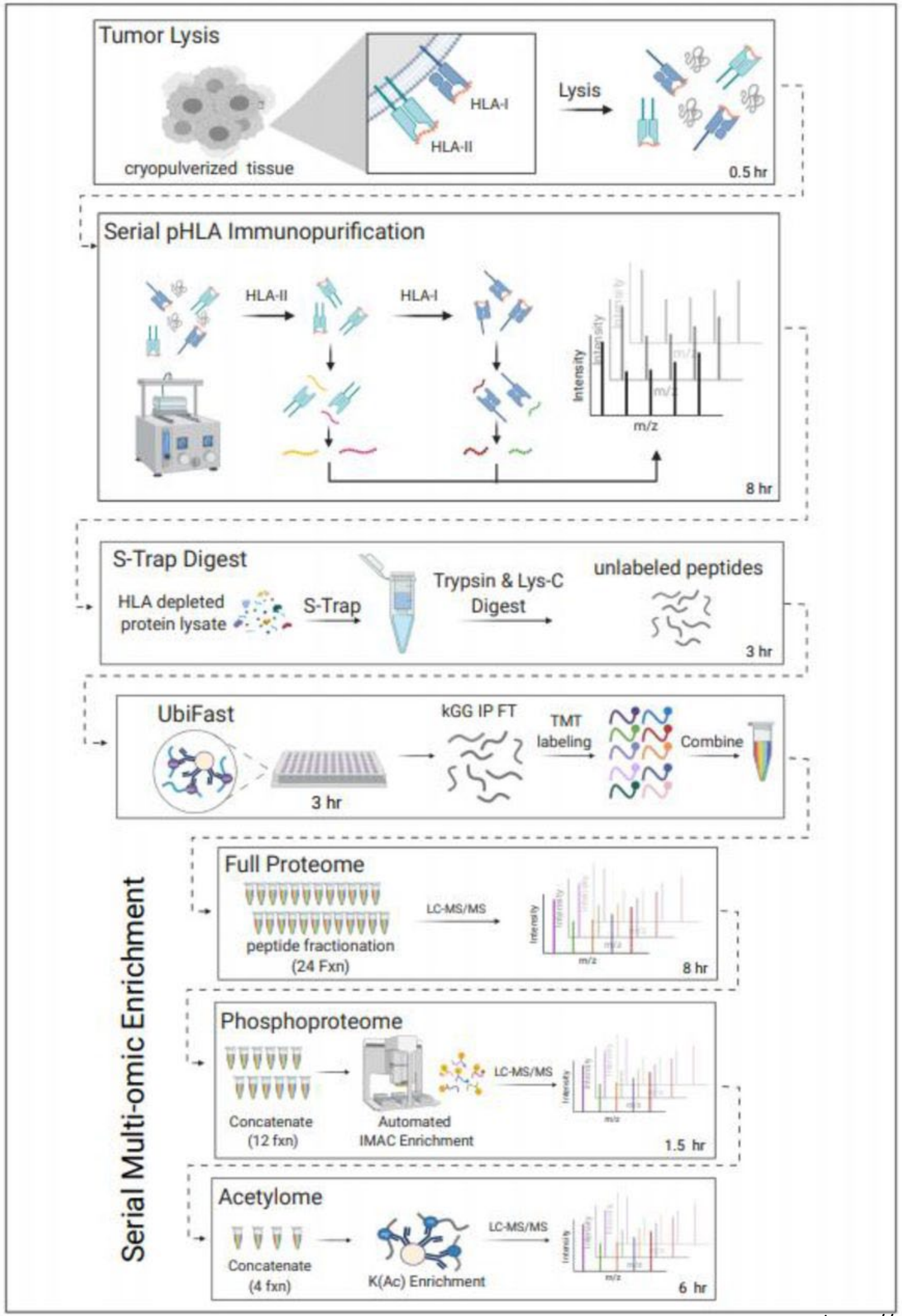


PTMs are generally present in low abundance, for this reason TMT is our method of choice

DOI: 10.1249/jes.0000000000000146

https://www.acrobiosystems.com/A1128-Pre-Coupling-Magnetic-Beads.html

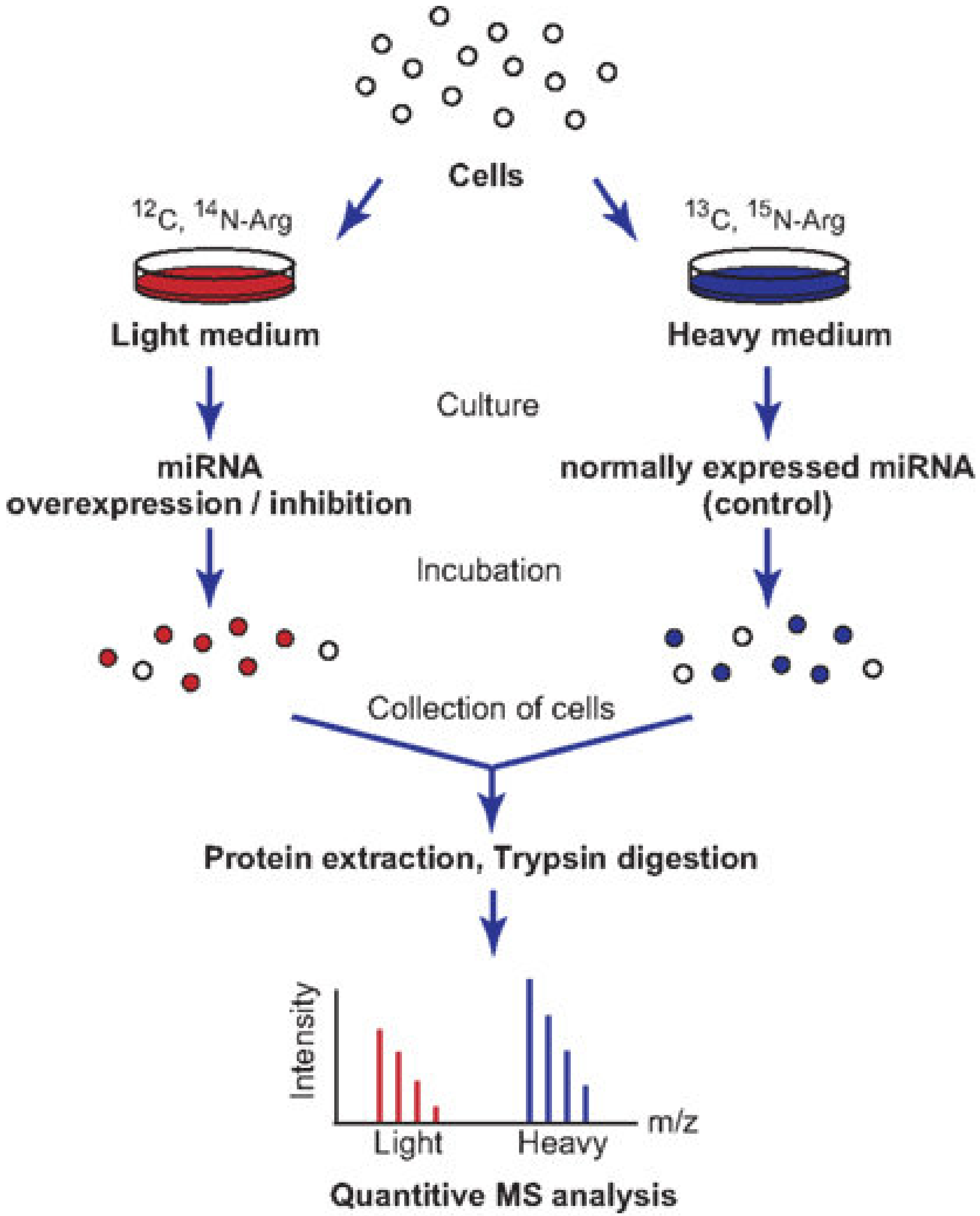
Multi-PTM workflows



https://www.biorxiv.org/content/10.1101/2021.06.22.449417v1.full

Metabolic Labeling

SILAC- Stable Isotope Labeling by Amino acids in cell Culture



SILAC Strengths

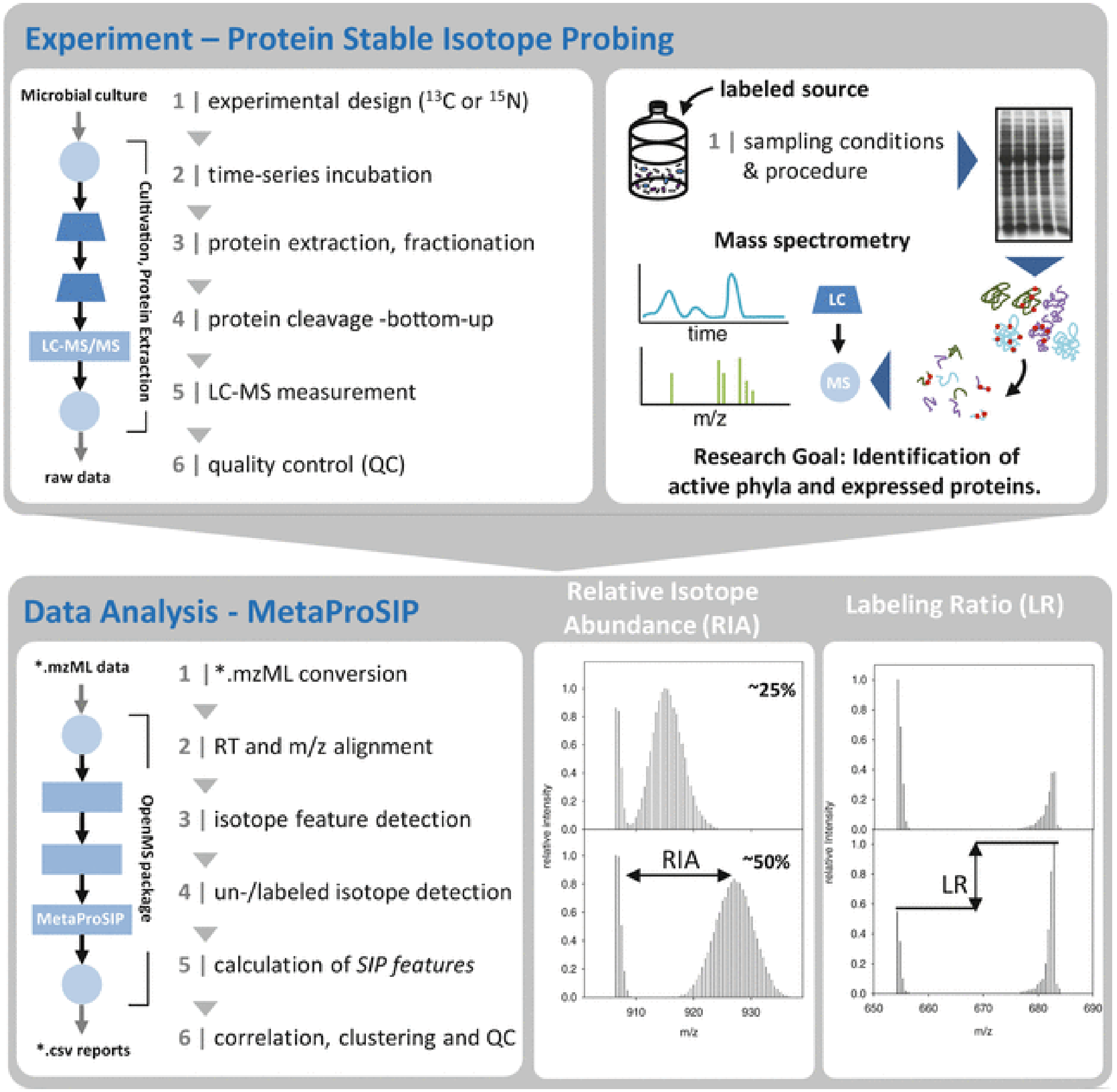
- Best quality 1 to 1 comparisons
- Protein turnover
- Nascent protein studies

SILAC Weaknesses

- Can’t study things that can’t be labeled
- Low throughput
- Heavy amino acids can get expensive for large experiments

https://doi.org/10.3389/fgene.2011.00029

Metabolic Labeling- SIP



Experiment – Protein Stable Isotope Probing

1 | experimental design (^{13}C or ^{15}N)

2 | time-series incubation

3 | protein extraction, fractionation

4 | protein cleavage -bottom-up

5 | LC-MS measurement

6 | quality control (QC)

Research Goal: Identification of active phyla and expressed proteins.

Data Analysis - MetaProSIP

1 | *.mzML conversion

2 | RT and m/z alignment

3 | isotope feature detection

4 | un-/labeled isotope detection

5 | calculation of *SIP features*

6 | correlation, clustering and QC

SIP Strengths

- Allows studies of complex metabolism

SIP Weaknesses

- Challenging data analysis

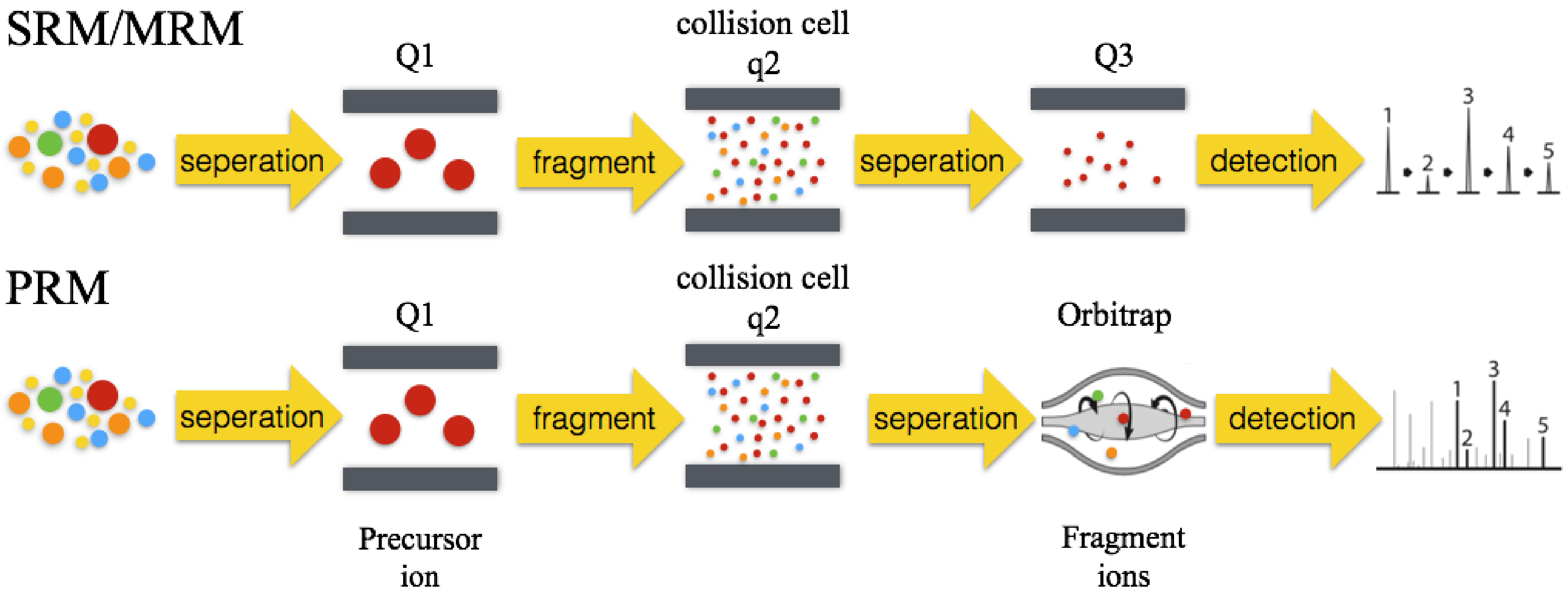
SIP- Stable Isotope Probing

https://doi.org/10.1016/j.copbio.2016.04.010

What are we going to talk about today?

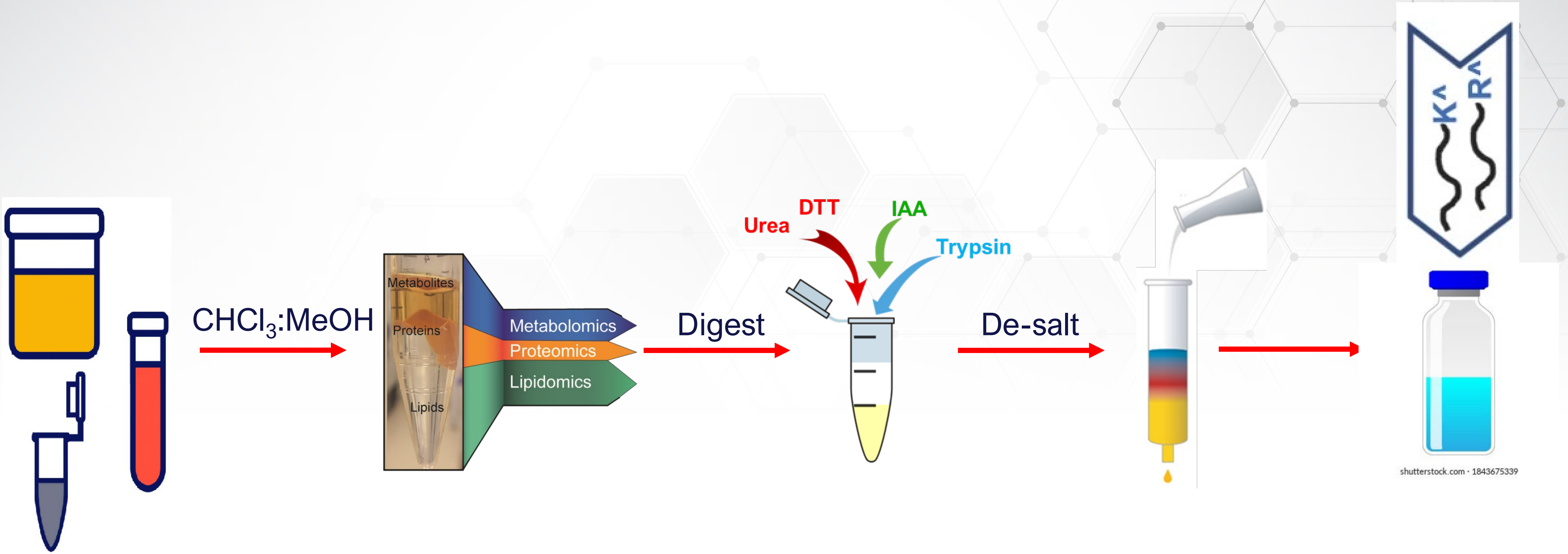
- Primer on proteomics and mass spectrometry
- Bottom-up proteomics
  - Understanding bottom-up proteomics
  - Quantification
  - Discovery approaches
    - Global quantification
    - PTM’s
    - Spatial and Single Cell
    - Metabolic Labeling
  - Targeted approaches
- Top-down proteomics
  - Intact
  - Native

Targeted proteomics



https://www.creative-proteomics.com/services/parallel-reaction-monitoring-prm.htm

Targeted Proteomics uses heavy labeled standards



$CHCl_3$:MeOH

Metabolomics

Proteomics

Lipidomics

Digest

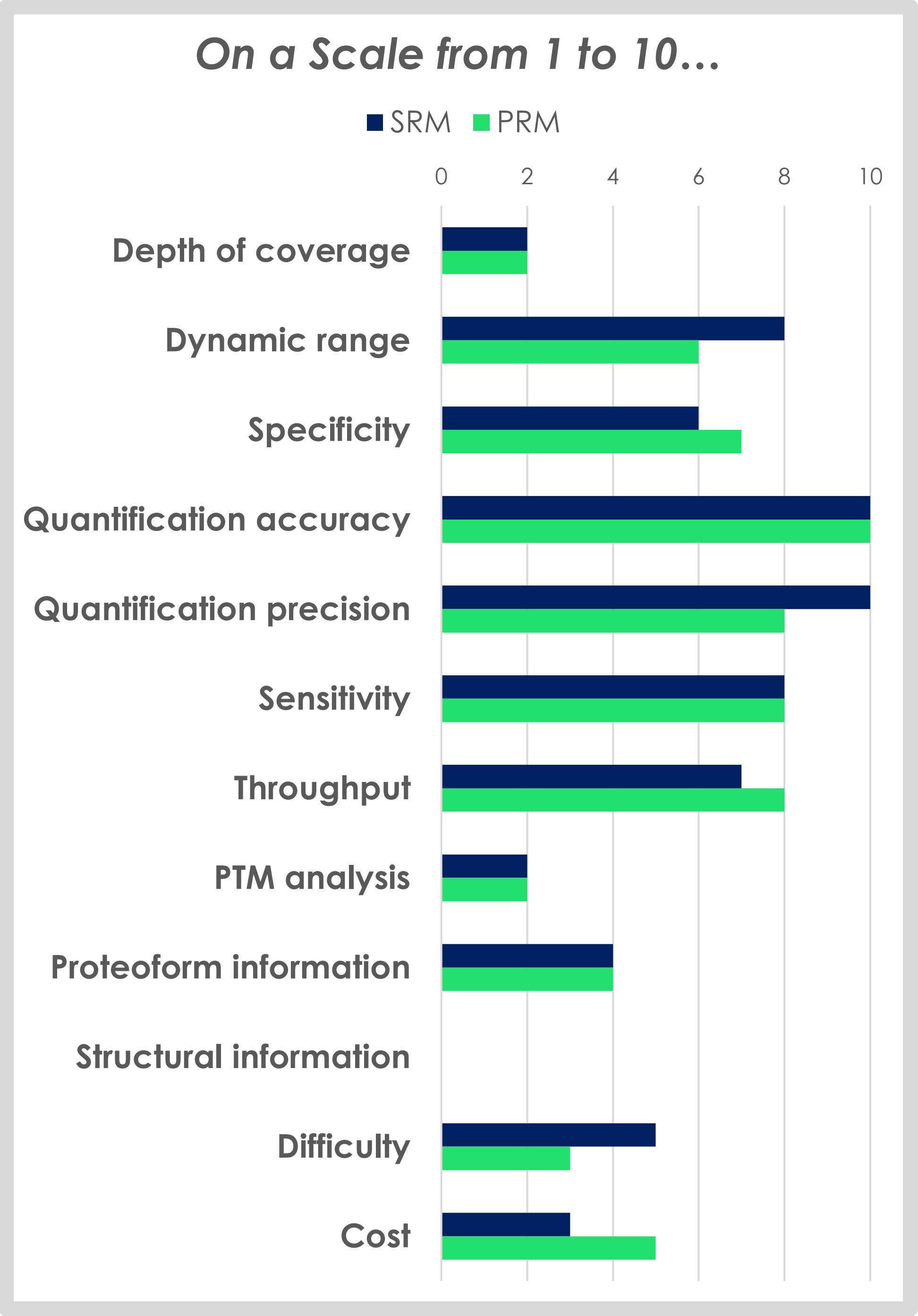
De-salt

shutterstock.com · 1843675339

Targeted Proteomics

SRM- Selected Reaction Monitoring

PRM- Parallel Reaction Monitoring



SRM Strengths

- Better dynamic range
- Better precision
- Cheap robust MS

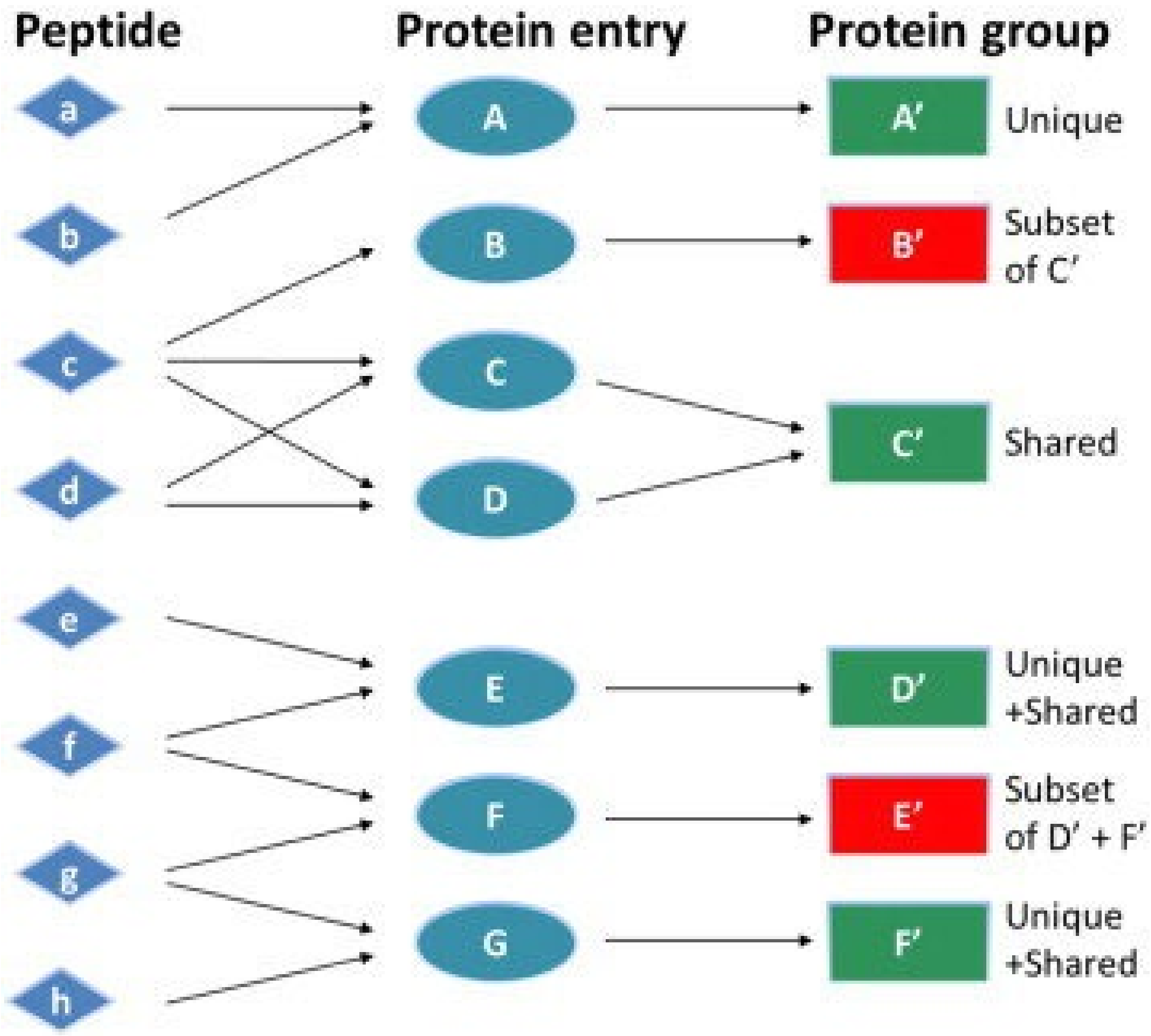
PRM Strengths

- Better specificity
- No upfront method development
- Easier to implement

What are we going to talk about today?

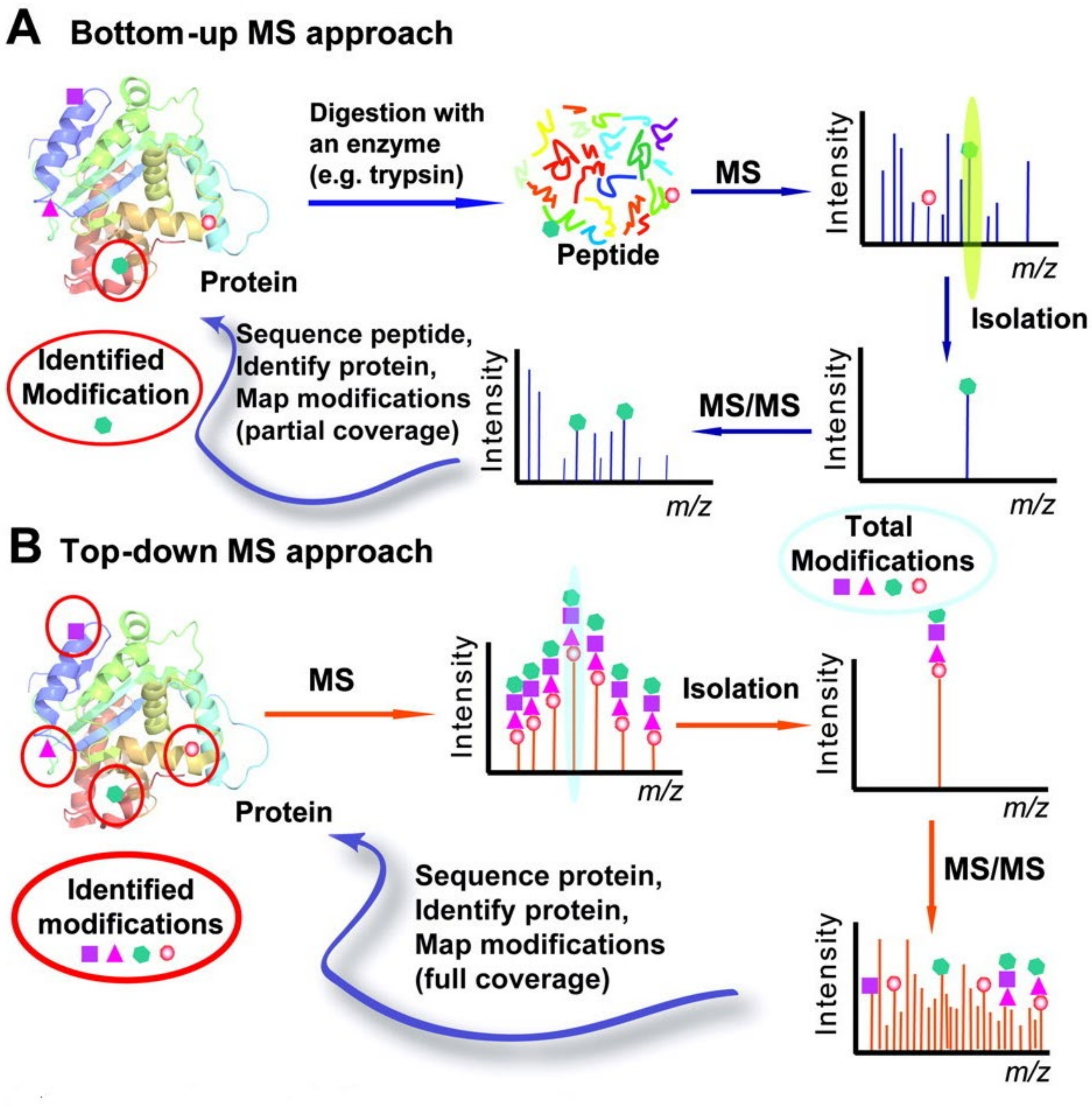
- Primer on proteomics and mass spectrometry
- Bottom-up proteomics
  - Understanding bottom-up proteomics
  - Quantification
  - Discovery approaches
    - Global quantification
    - PTM’s
    - Spatial and Single Cell
    - Metabolic Labeling
  - Targeted approaches
- Top-down proteomics
  - Intact
  - Native

The Protein Inference Problem



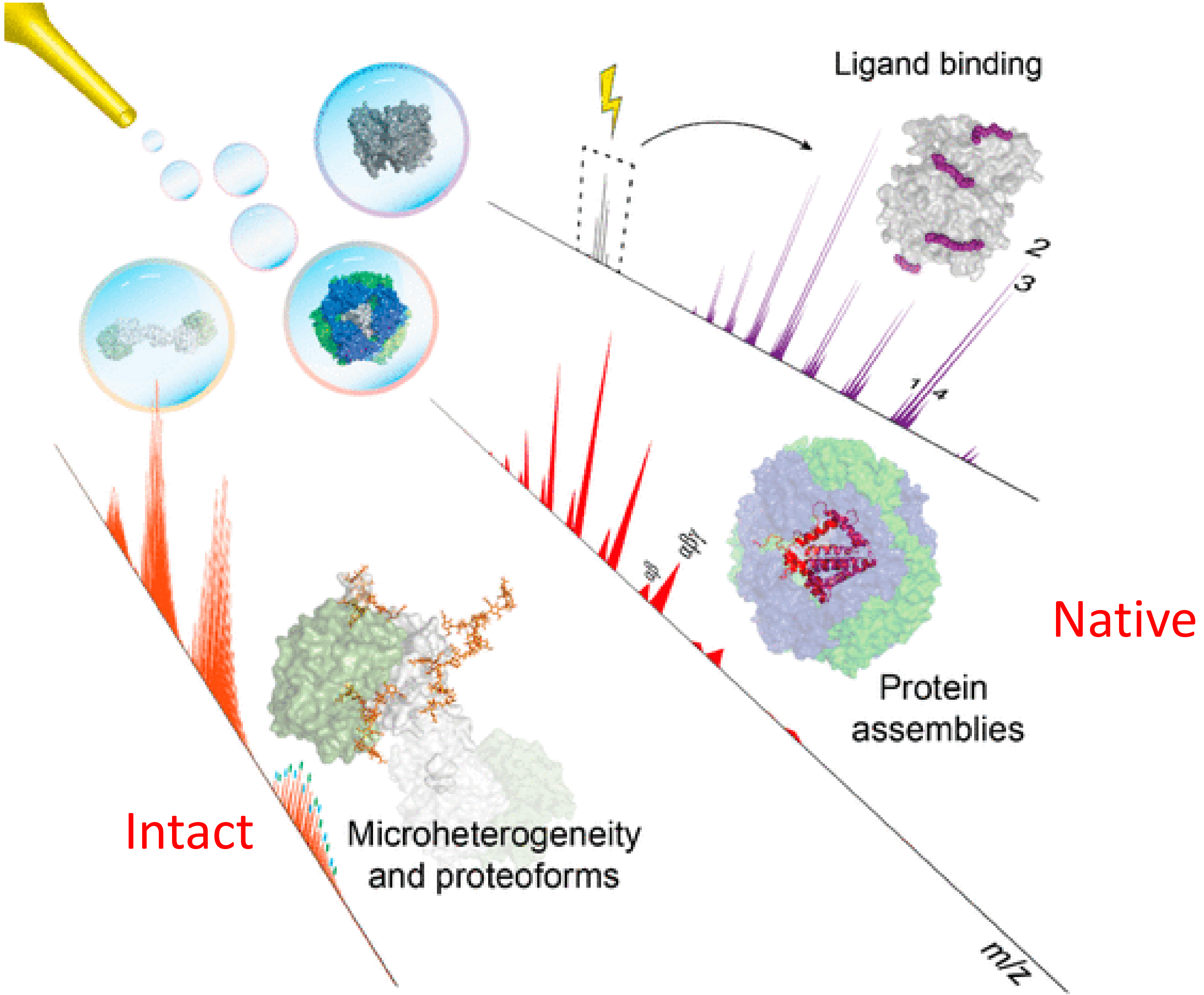
10.1016/j.jprot.2015.07.006

Top-Down Proteomics



https://doi.org/10.1161/CIRCGENETICS.110.957829

Native Proteomics

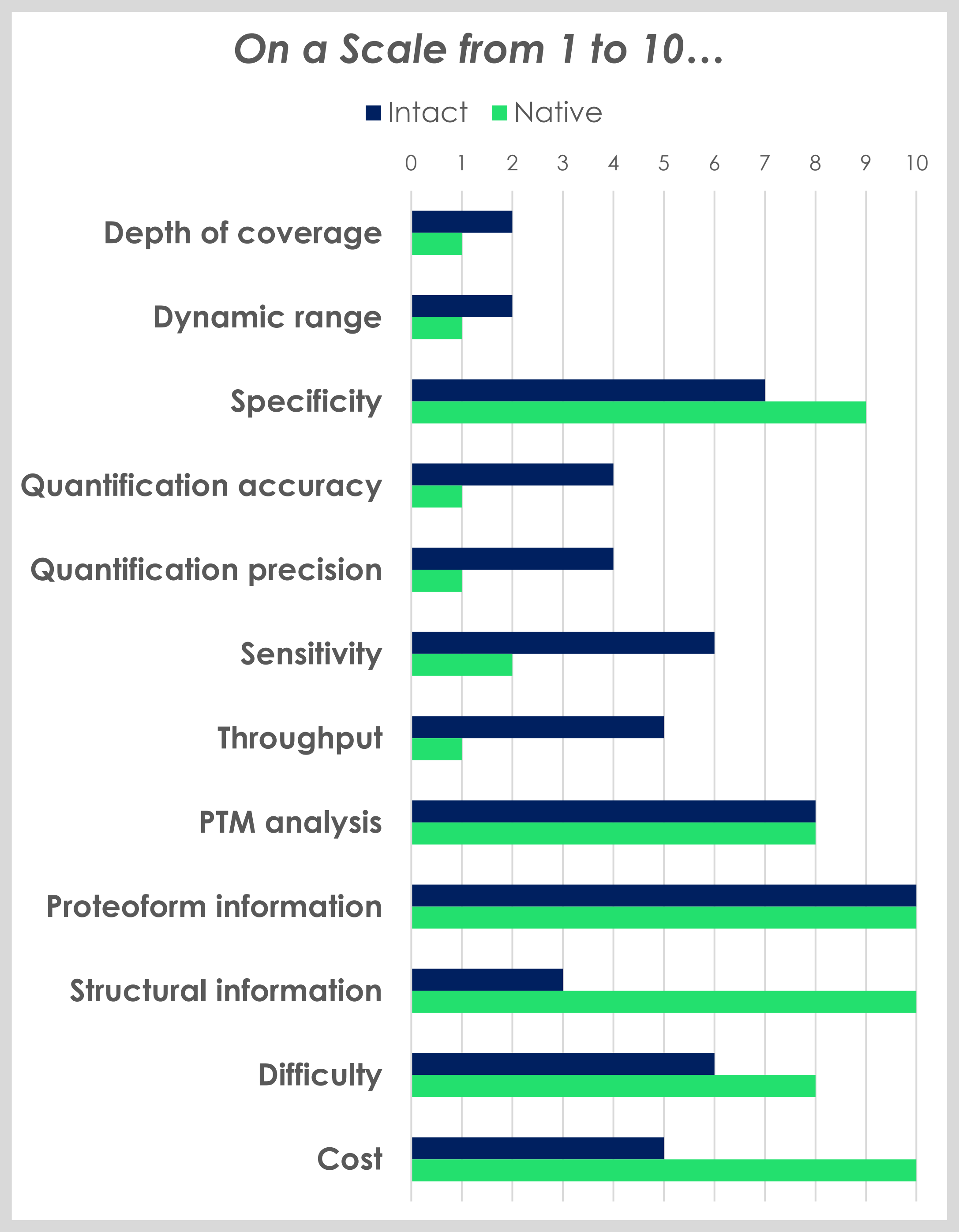


Native

Intact

https://pubs.acs.org/doi/full/10.1021/acs.chemrev.1c00212

Top Down Proteomics



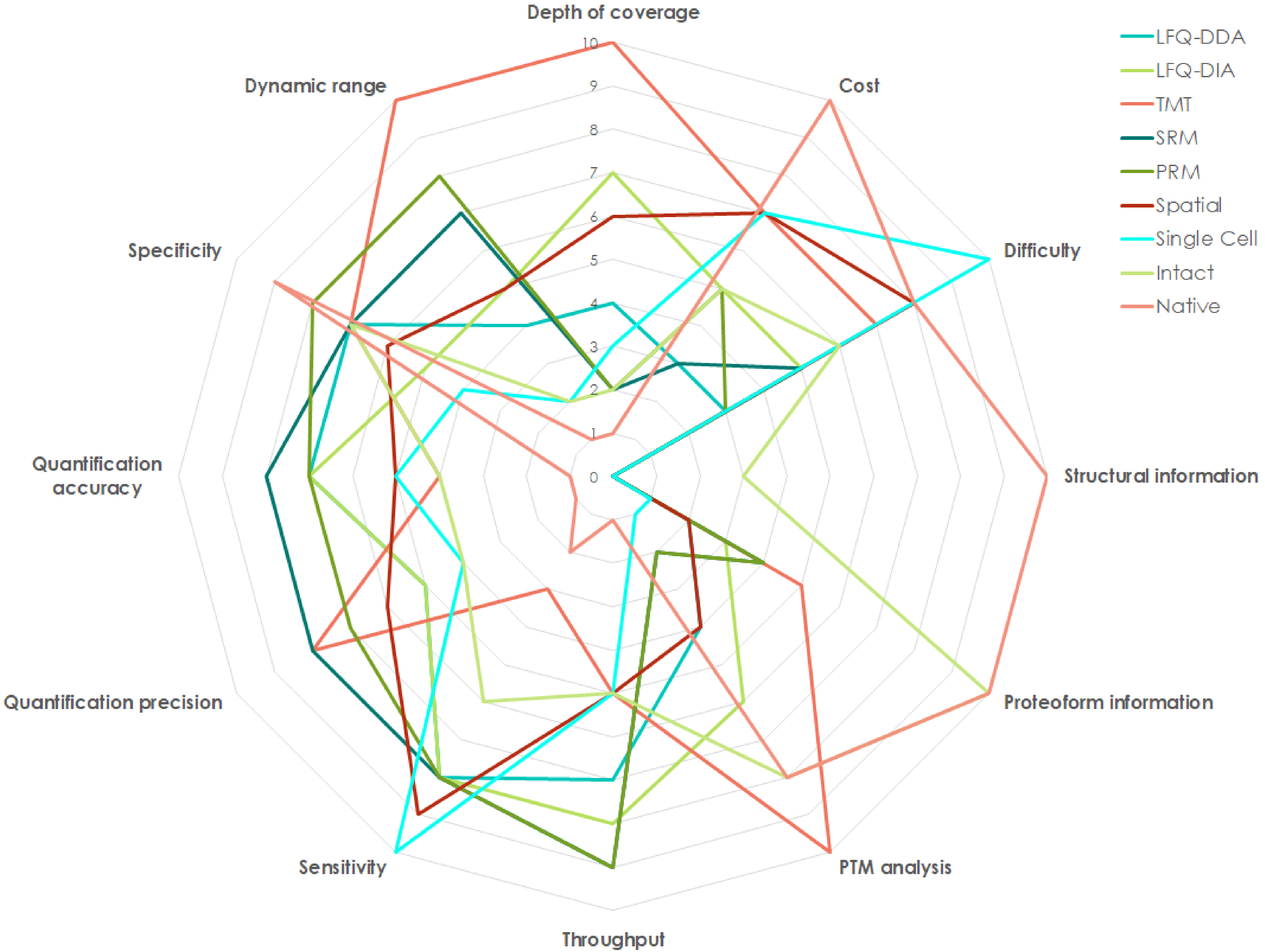
Intact Strengths

- Characterizing multiple PTMs
- Histones
- Plaque proteins (Tau, α-syn)

Native Strengths

- Protein complexes
- Ligand binding

In Summary


Depth of coverage
Cost
Difficulty
Structural information
Proteoform information
PTM analysis
Throughput
Sensitivity
Quantification precision
Quantification accuracy
Specificity
Dynamic range
10
9
8
7
6
5
4
3
2
1
0
LFQ-DDA
LFQ-DIA
TMT
SRM
PRM
Spatial
Single Cell
Intact
Native

From One Sample to an Expression Data Matrix

Expression Data, or “e_data”, is used in differential statistics, multi-omic integration, network generation, and supervised/unsupervised machine learning methods.

Peptide	Sample 1	Sample 2	Sample 3
Peptide 1	34636000	45342000	34534000
Peptide 2	2353000	NA	9345300
Peptide 3	NA	787453000	NA

Expression Matrix (E_Data)

Peptide	SARS-CoV-2-Delta_Control1	SARS-CoV-2-Delta_Control2	SARS-CoV-2-Delta_Treatment1	SARS-CoV-2-Delta_Treatment2
A.LHTEGDKAFVEFLTDEIKEEK.K	17953839	20071472	20745779	18206556
A.LIVYDDLSK.Q	109536335	115459820	106127139	74522014
A.LLAAHPNER.L	1752288782	1796561709	1703186182	2438218572
A.LLAGLGAVTLTK.E	2571804	4269824	4852871	2630414
A.LLDVNLPDM*EGYDVGR.A	110239193	82436688	100447189	102006001
A.LLDYDSELRPTLK.Q	18263322	17416268	15069260	25083207
A.LLHSADLLEEVK.E	15184670	18160176	15353092	6463005
A.LLILKPDAVQR.G	14581430	15764607	16009605	9502368
A.LLSLPNVEQVLR.G	294215486	266026856	292986771	328573619
A.LLTHDDVK.Q	6503093	6096751	6215913	7243116.5

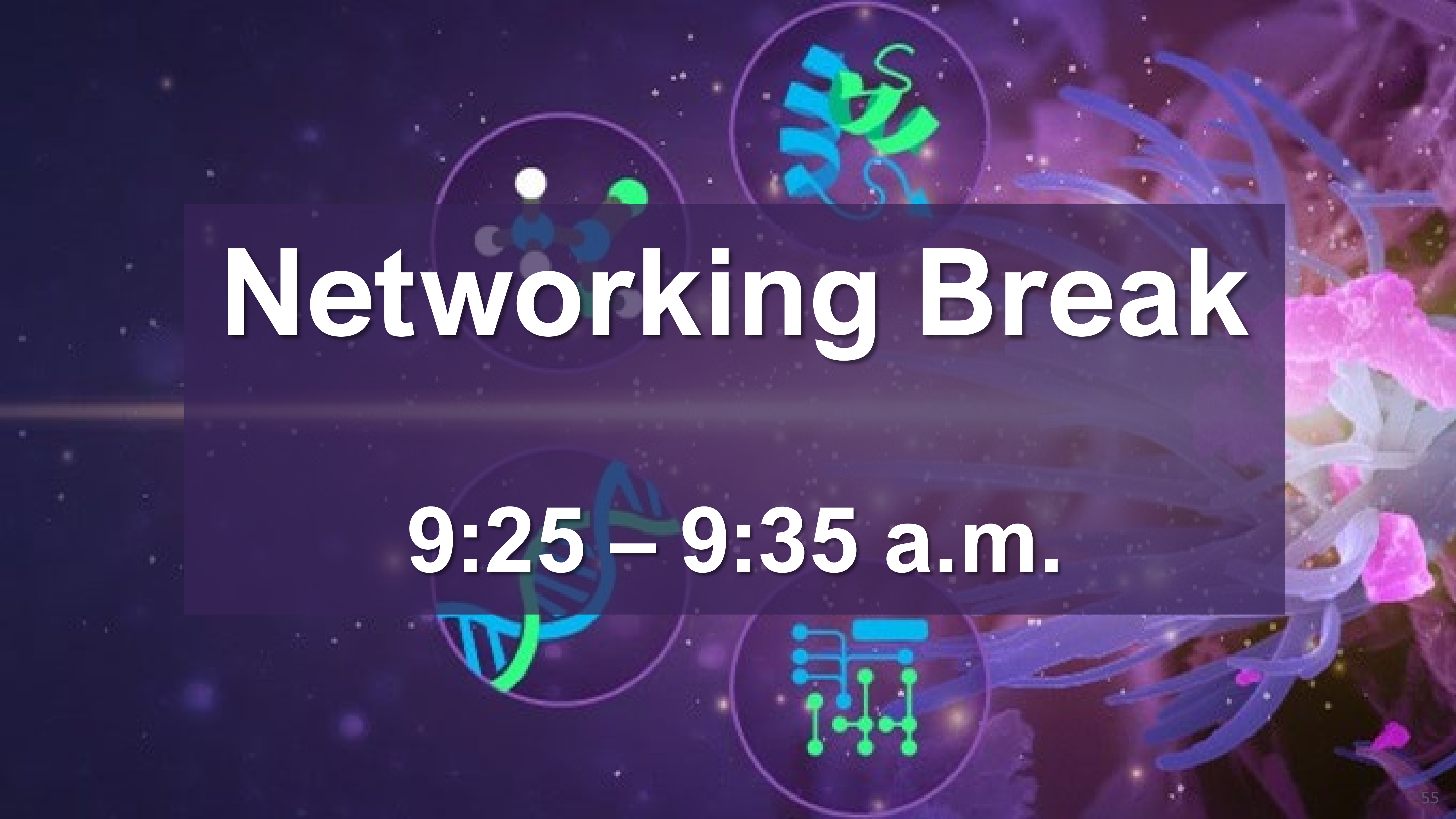
Sample Information (F_Data)

Sample	Group	Batch
SARS-CoV-2-Delta_Control1	Control	1
SARS-CoV-2-Delta_Control2	Control	2
SARS-CoV-2-Delta_Treatment1	Treatment	1
SARS-CoV-2-Delta_Treatment2	Treatment	2

Biomolecule Information (E_Meta)

Peptide	Protein	Contaminant
A.LHTEGDKAFVEFLTDEIKEEK.K	YP_009724389	No
A.LIVYDDLSK.Q	YP_009725389	No
A.LLAAHPNER.L	YP_009726297	No
A.LLAGLGAVTLTK.E	YP_009726298	No
A.LLDVNLPDM*EGYDVGR.A	YP_009726296	No
A.LLDYDSELRPTLK.Q	NA	Yes
A.LLHSADLLEEVK.E	YP_009625291	No
A.LLILKPDAVQR.G	YP_00971542	No
A.LLSLPNVEQVLR.G	YP_009724293	No
A.LLTHDDVK.Q	YP_009785674	No

Questions?



Networking Break

9:25 – 9:35 a.m.

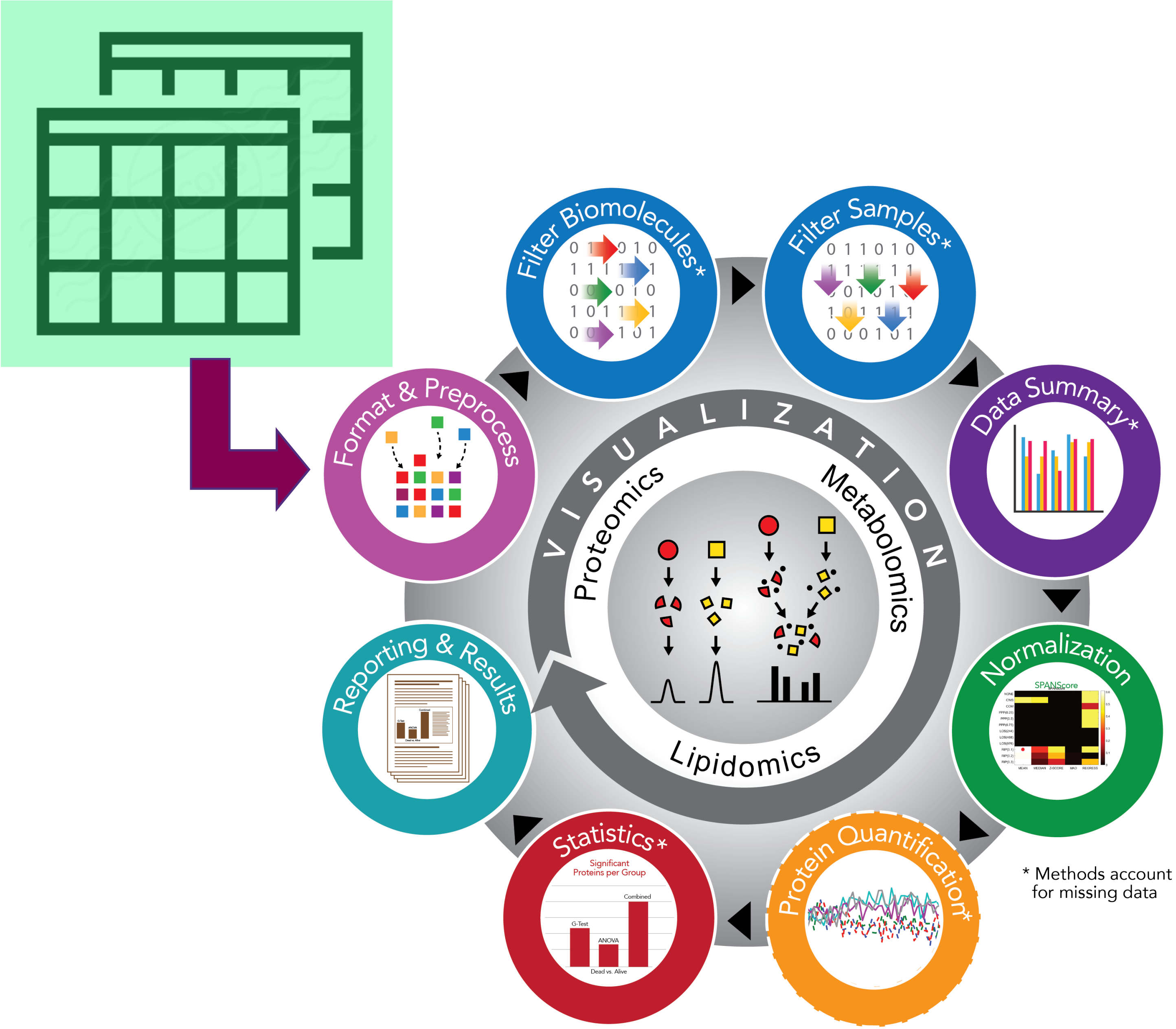
8:30-8:35 a.m.	Introduction	Kelly Stratton
8:35-9:25	Types of Proteomics	Paul Piehowski & David Degnan
9:25-9:35	Networking Break	
9:35-10:40	**Typical Statistical Processing**	**Kelly Stratton**
10:40-10:50	Networking Break	
10:50-11:40	Biological Interpretation	David Degnan & Tyler Sagendorf
11:40-11:45	Closing Remarks	David Degnan

EMSL

Typical Statistical Processing

Kelly Stratton

Typical Statistical Processing



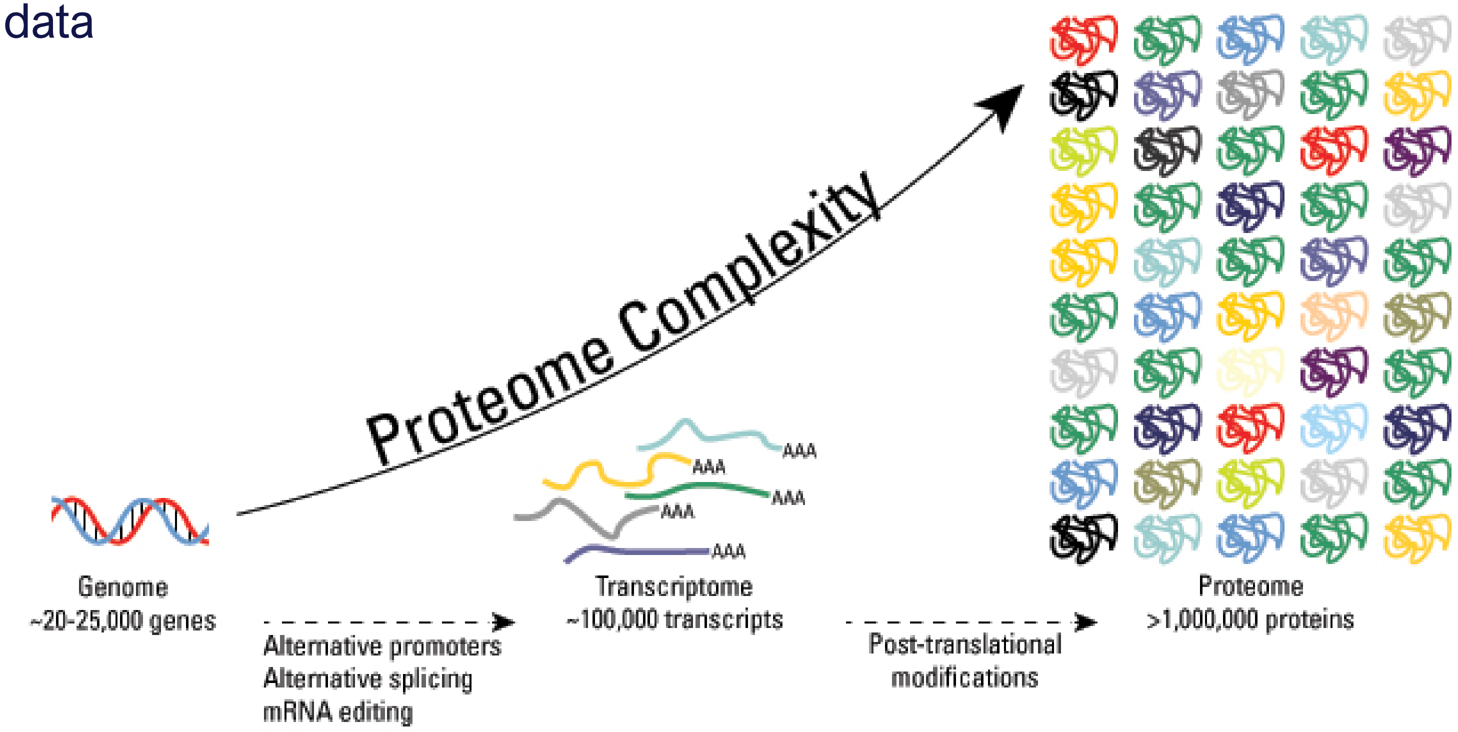
* Methods account for missing data

Typical Statistical Processing

Challenges with Proteomics Data

- Noisy Data
  - Misidentifications
  - Relative quantification
  - Unstable variance
  - Large amounts of missing data

- Biology
  - Unknown/complex interactions between proteins and other small molecules
  - Peptides map to more than one protein
  - Function changes



Tools

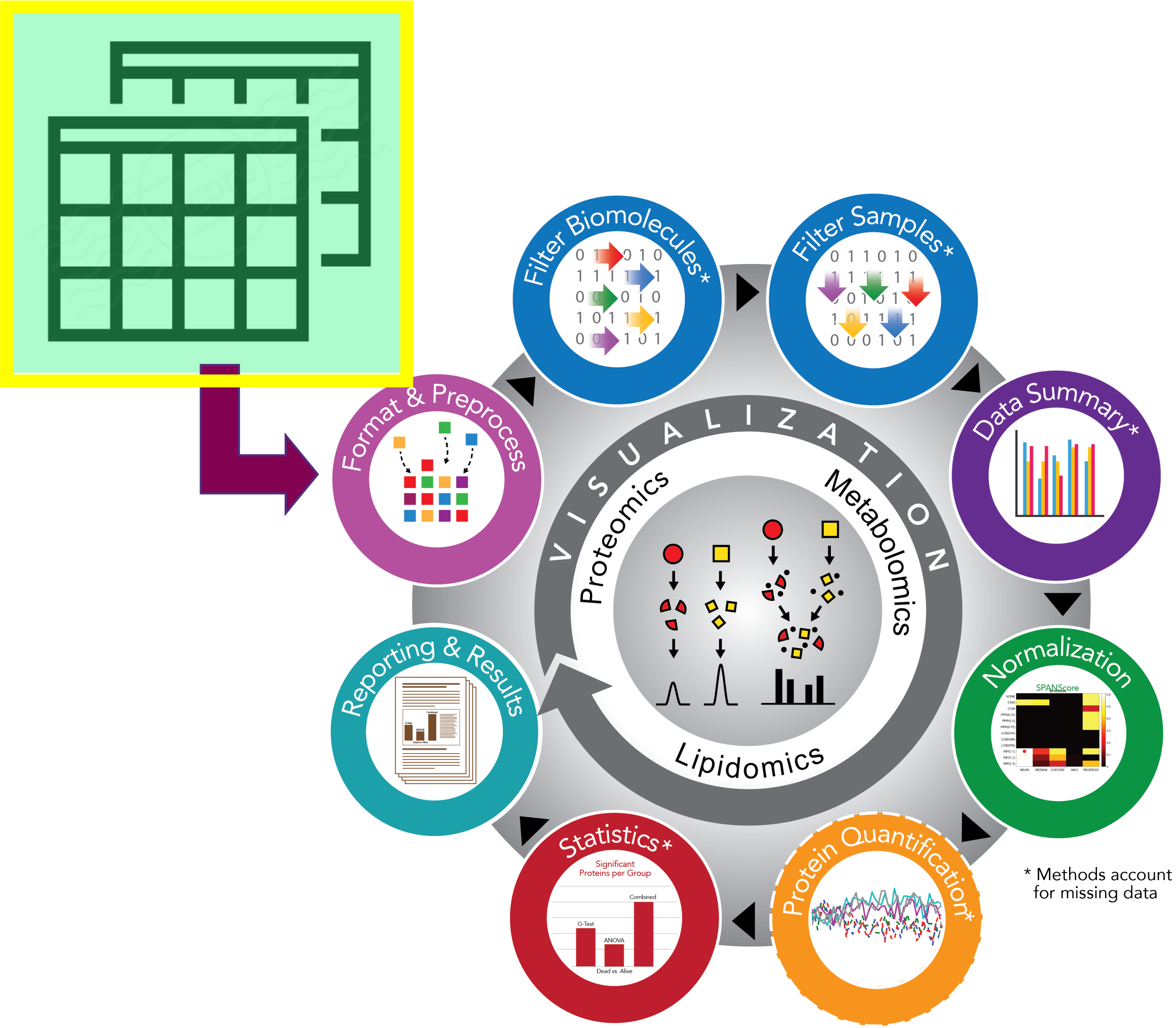
Open Source Tools

- *pmartR*
  - https://github.com/pmartR/pmartR
  - Streamlines data processing, exploration, QC, statistical analysis, and interactive visualization of biological data
  - Provides methods robust to missing data, which are ubiquitous in mass spectrometry data
  - Operates on **isobaric tag labeled and label-free proteomic**, metabolomic (NMR and GC-/LC-MS), lipidomic, RNA-seq count data
- Multiomics Analysis Portal (MAP) user interface
  - https://map.emsl.pnnl.gov/app/map
- Other options: MetaboAnalyst, Msstats
  - https://www.metaboanalyst.ca/
  - https://www.bioconductor.org/packages/release/bioc/html/MSstats.html

Stratton, K. G., Webb-Robertson, B. J. M., McCue, L. A., Stanfill, B., Claborne, D., Godinez, I., ... & Bramer, L. M. (2019). pmartR: Quality Control and Statistics for Mass Spectrometry-Based Biological Data. *Journal of proteome research*, *18*(3), 1418-1425.

pmartR: https://github.com/pmartR/pmartR; PMart: https://github.com/pmartR/PMart_ShinyApp; Web Application: release Aug 2021

Typical Statistical Processing

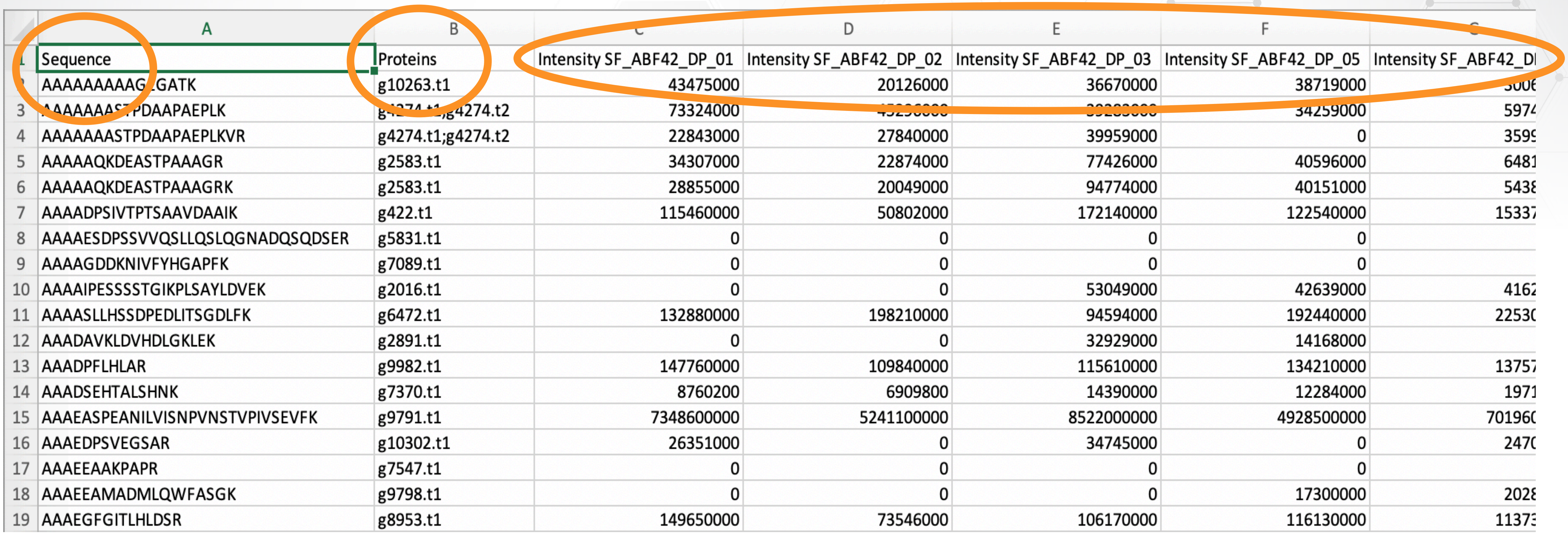


* Methods account for missing data

Typical Statistical Processing

Data Format

- MaxQuant protein level quantification
- We prefer to start at peptide level

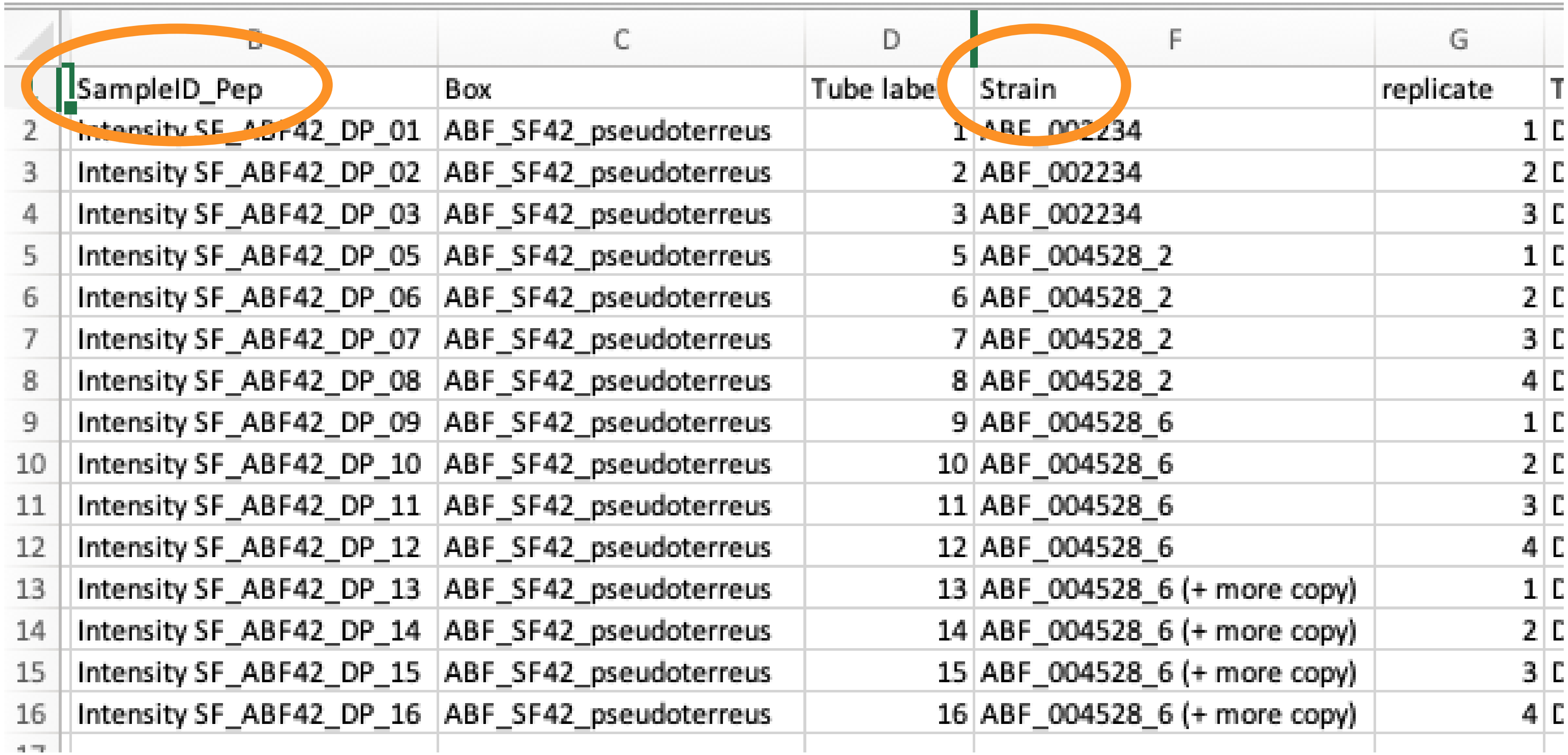


	A	B	C	D	E	F	G
1	Sequence	Proteins	Intensity SF_ABF42_DP_01	Intensity SF_ABF42_DP_02	Intensity SF_ABF42_DP_03	Intensity SF_ABF42_DP_05	Intensity SF_ABF42_DI
2	AAAAAAAAAGLGATK	g10263.t1	43475000	20126000	36670000	38719000	3006
3	AAAAAAASTPDAAPAEPLK	g4274.t1;g4274.t2	73324000	45296000	39283000	34259000	5974
4	AAAAAAASTPDAAPAEPLKVR	g4274.t1;g4274.t2	22843000	27840000	39959000	0	3599
5	AAAAAQKDEASTPAAAGR	g2583.t1	34307000	22874000	77426000	40596000	6481
6	AAAAAQKDEASTPAAAGRK	g2583.t1	28855000	20049000	94774000	40151000	5438
7	AAAADPSIVTPTSAAVDAAIK	g422.t1	115460000	50802000	172140000	122540000	15337
8	AAAAESDPSSVVQSLLQSLQGNADQSQDSER	g5831.t1	0	0	0	0	
9	AAAAGDDKNIVFYHGAPFK	g7089.t1	0	0	0	0	
10	AAAAIPESSSSTGIKPLSAYLDVEK	g2016.t1	0	0	53049000	42639000	4162
11	AAAASLLHSSDPEDLITSGDLFK	g6472.t1	132880000	198210000	94594000	192440000	22530
12	AAADAVKLDVHDLGKLEK	g2891.t1	0	0	32929000	14168000	
13	AAADPFLHLAR	g9982.t1	147760000	109840000	115610000	134210000	13757
14	AAADSEHTALSHNK	g7370.t1	8760200	6909800	14390000	12284000	1971
15	AAAEASPEANILVISNPVNSTVPIVSEVFK	g9791.t1	7348600000	5241100000	8522000000	4928500000	701960
16	AAAEDPSVEGSAR	g10302.t1	26351000	0	34745000	0	2470
17	AAAEEAAKPAPR	g7547.t1	0	0	0	0	
18	AAAEEAMADMLQWFASGK	g9798.t1	0	0	0	17300000	2028
19	AAAEGFGITLHLDSR	g8953.t1	149650000	73546000	106170000	116130000	11373

Typical Statistical Processing

Data Format

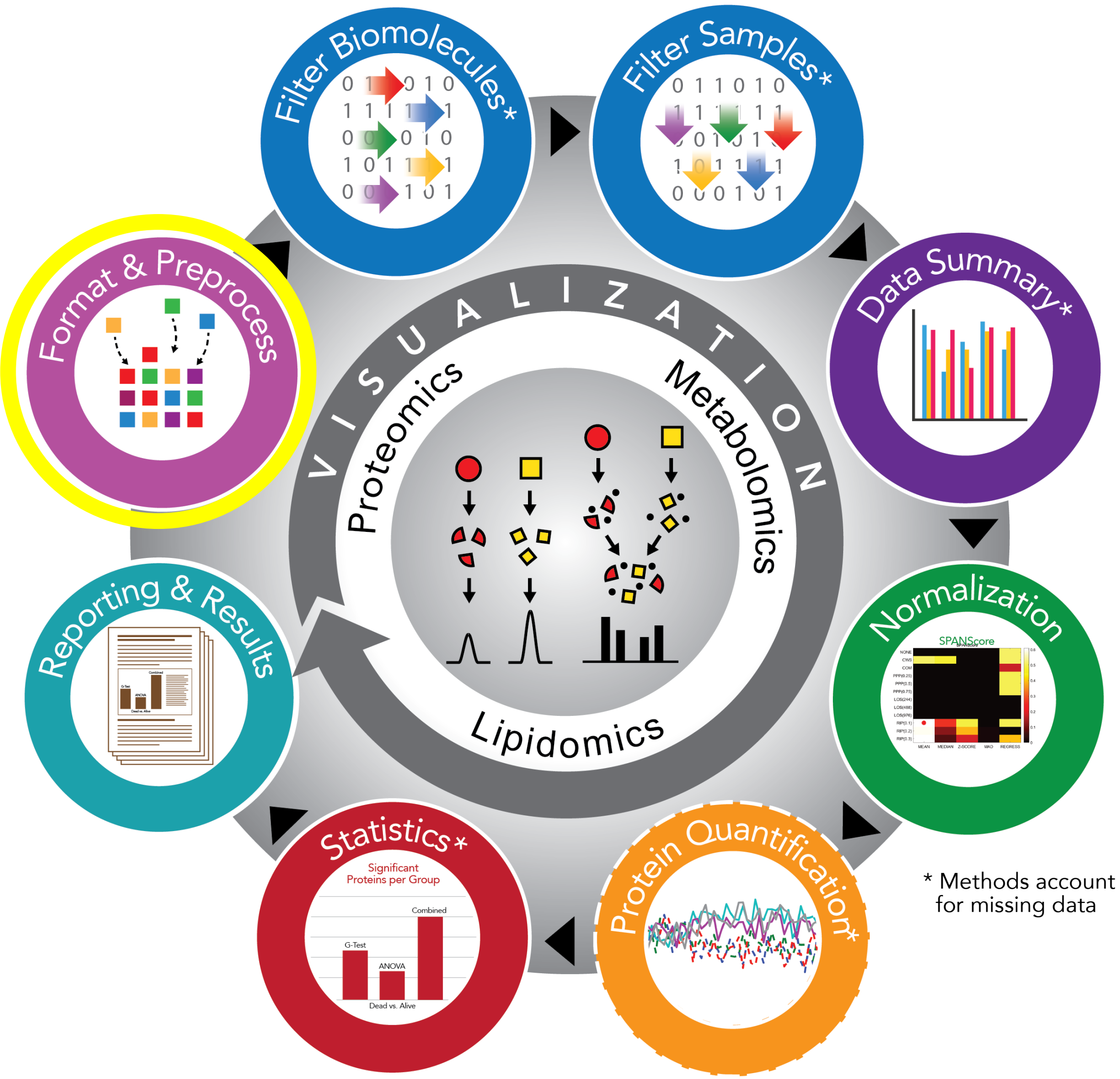
- Sample IDs, experimental groups, etc.



	B	C	D	F	G
1	SampleID_Pep	Box	Tube labe	Strain	replicate
2	Intensity SF_ABF42_DP_01	ABF_SF42_pseudoterreus	1	ABF_002234	1
3	Intensity SF_ABF42_DP_02	ABF_SF42_pseudoterreus	2	ABF_002234	2
4	Intensity SF_ABF42_DP_03	ABF_SF42_pseudoterreus	3	ABF_002234	3
5	Intensity SF_ABF42_DP_05	ABF_SF42_pseudoterreus	5	ABF_004528_2	1
6	Intensity SF_ABF42_DP_06	ABF_SF42_pseudoterreus	6	ABF_004528_2	2
7	Intensity SF_ABF42_DP_07	ABF_SF42_pseudoterreus	7	ABF_004528_2	3
8	Intensity SF_ABF42_DP_08	ABF_SF42_pseudoterreus	8	ABF_004528_2	4
9	Intensity SF_ABF42_DP_09	ABF_SF42_pseudoterreus	9	ABF_004528_6	1
10	Intensity SF_ABF42_DP_10	ABF_SF42_pseudoterreus	10	ABF_004528_6	2
11	Intensity SF_ABF42_DP_11	ABF_SF42_pseudoterreus	11	ABF_004528_6	3
12	Intensity SF_ABF42_DP_12	ABF_SF42_pseudoterreus	12	ABF_004528_6	4
13	Intensity SF_ABF42_DP_13	ABF_SF42_pseudoterreus	13	ABF_004528_6 (+ more copy)	1
14	Intensity SF_ABF42_DP_14	ABF_SF42_pseudoterreus	14	ABF_004528_6 (+ more copy)	2
15	Intensity SF_ABF42_DP_15	ABF_SF42_pseudoterreus	15	ABF_004528_6 (+ more copy)	3
16	Intensity SF_ABF42_DP_16	ABF_SF42_pseudoterreus	16	ABF_004528_6 (+ more copy)	4

Typical Statistical Processing

Preprocessing

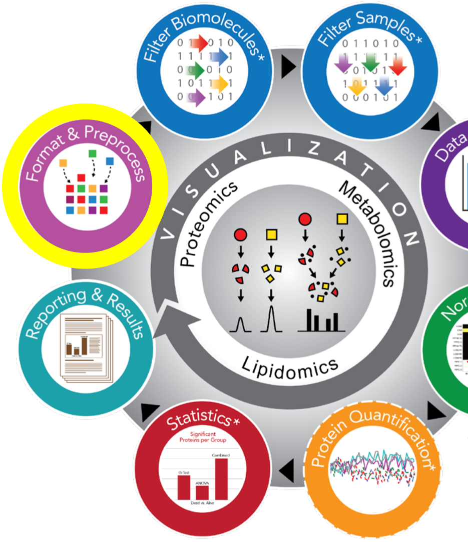


* Methods account for missing data

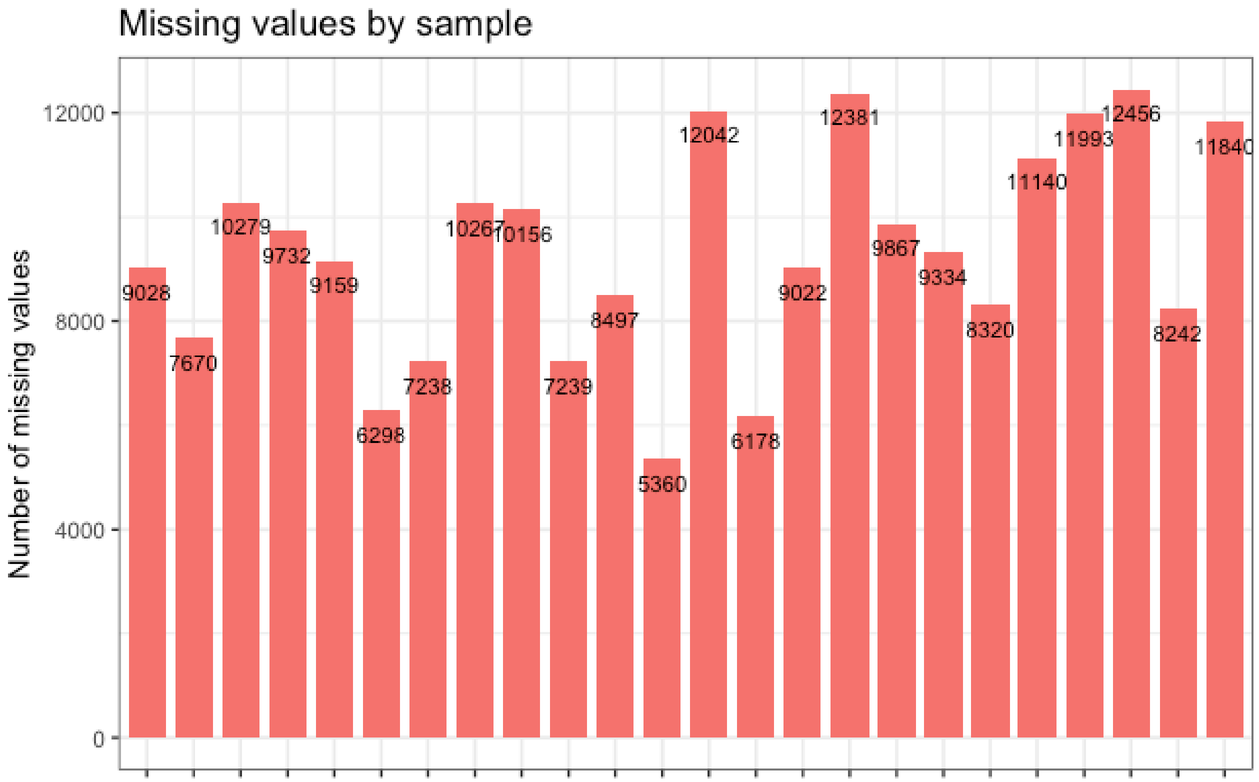
Typical Statistical Processing

Missing Values

- Proteomics data often contain >40% missing data
- Patterns of missingness vary

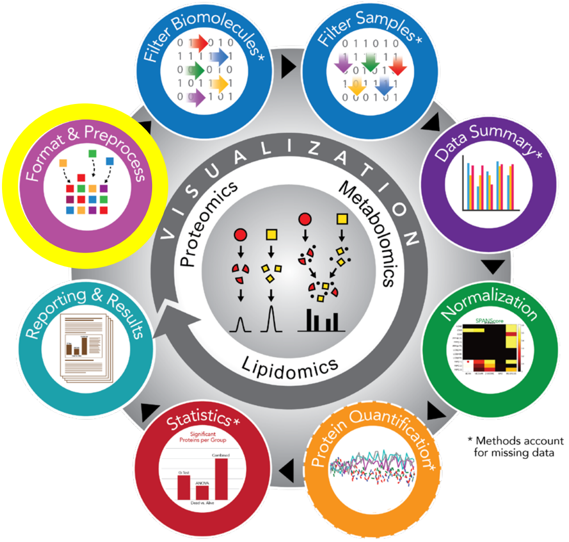


* Methods account for missing data



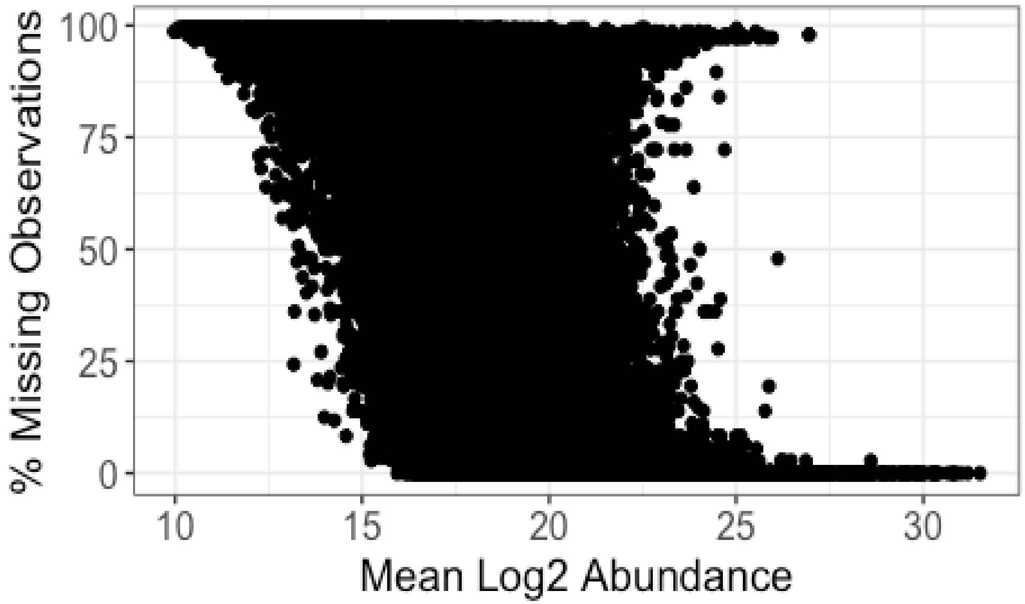
Typical Statistical Processing

Missing Values

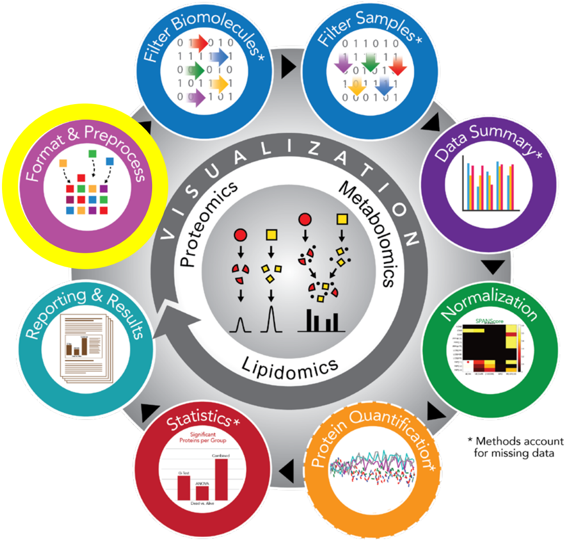


* Methods account for missing data

- If at all possible, 0’s ➔ NA values
- Imputation should be chosen & applied carefully when necessary
  - < 25% of values missing for a biomolecule
  - See JPR manuscripts with evaluation of imputation methods
    - Label-free proteomics in Webb-Robertson et al. (2015)
    - Isobaric-labelled proteomics in Bramer et al. (2020)



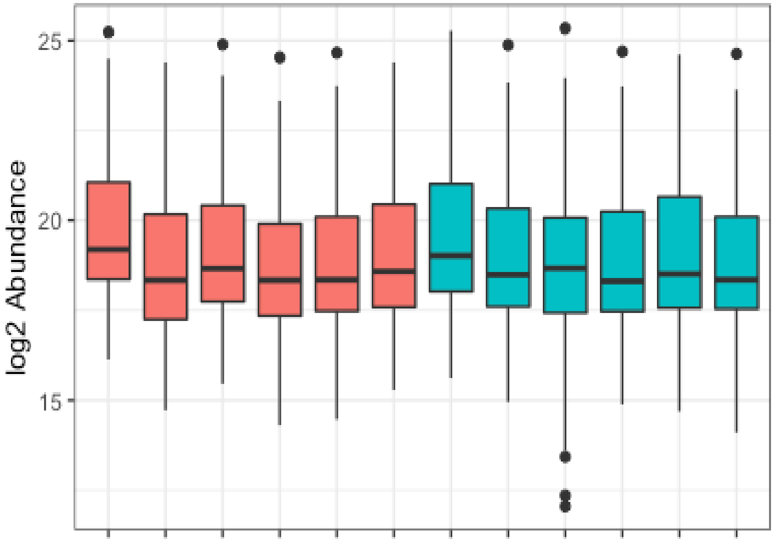
Typical Statistical Processing



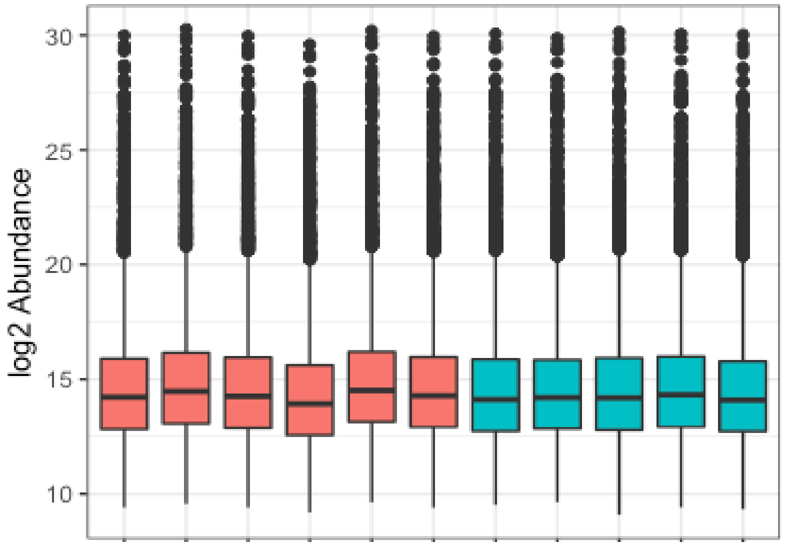
* Methods account for missing data

Missing Values

- Common imputation methods
  - Missing data = 0
  - ½ LOD or minimum
  - Can unintentionally change structure of data



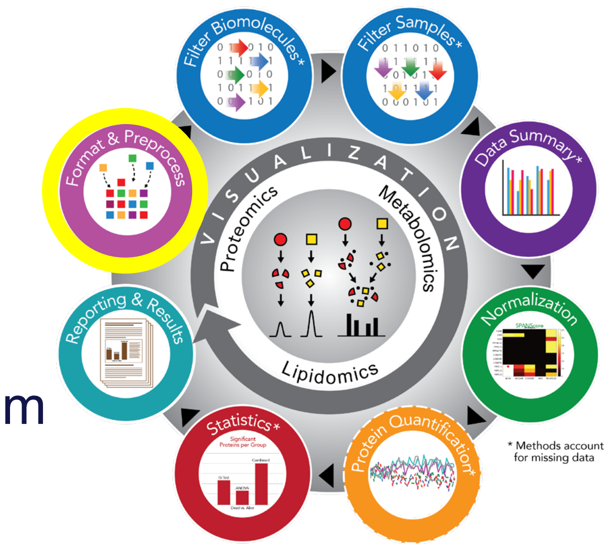
Before Imputation



After Imputation

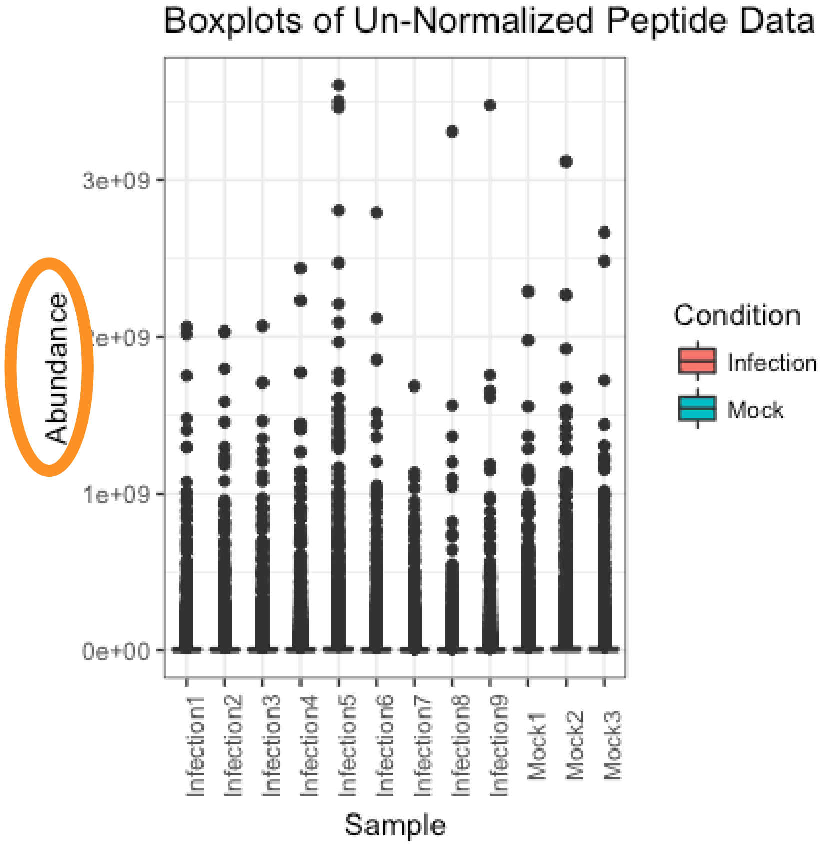
Typical Statistical Processing

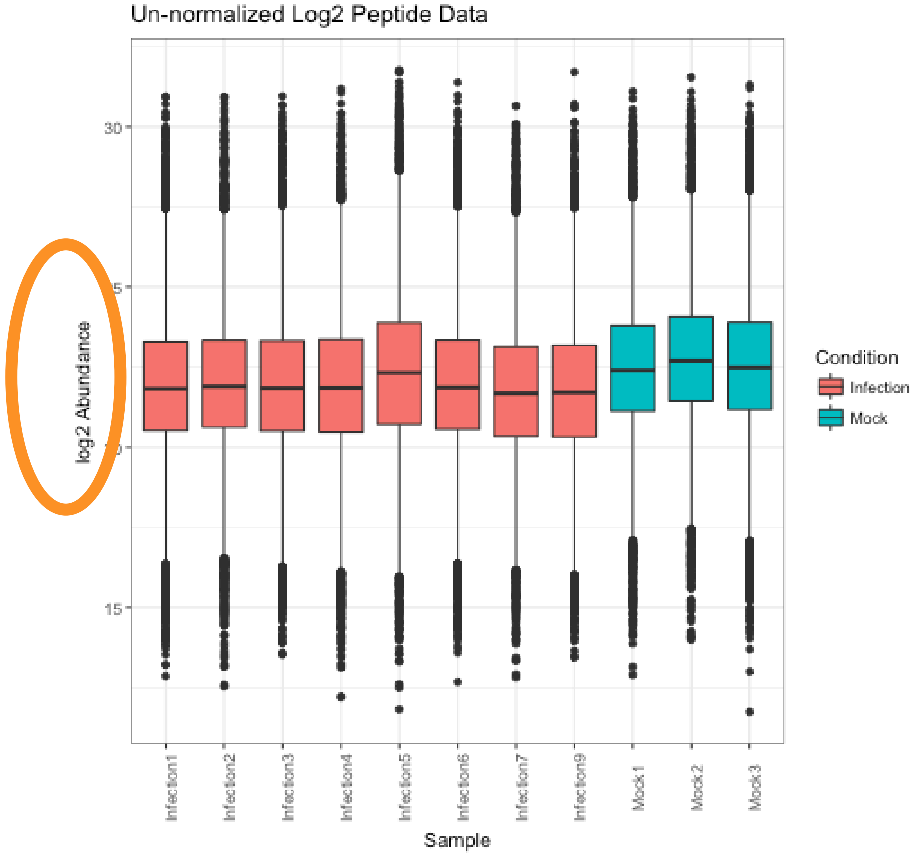
Data Transformation



* Methods account for missing data

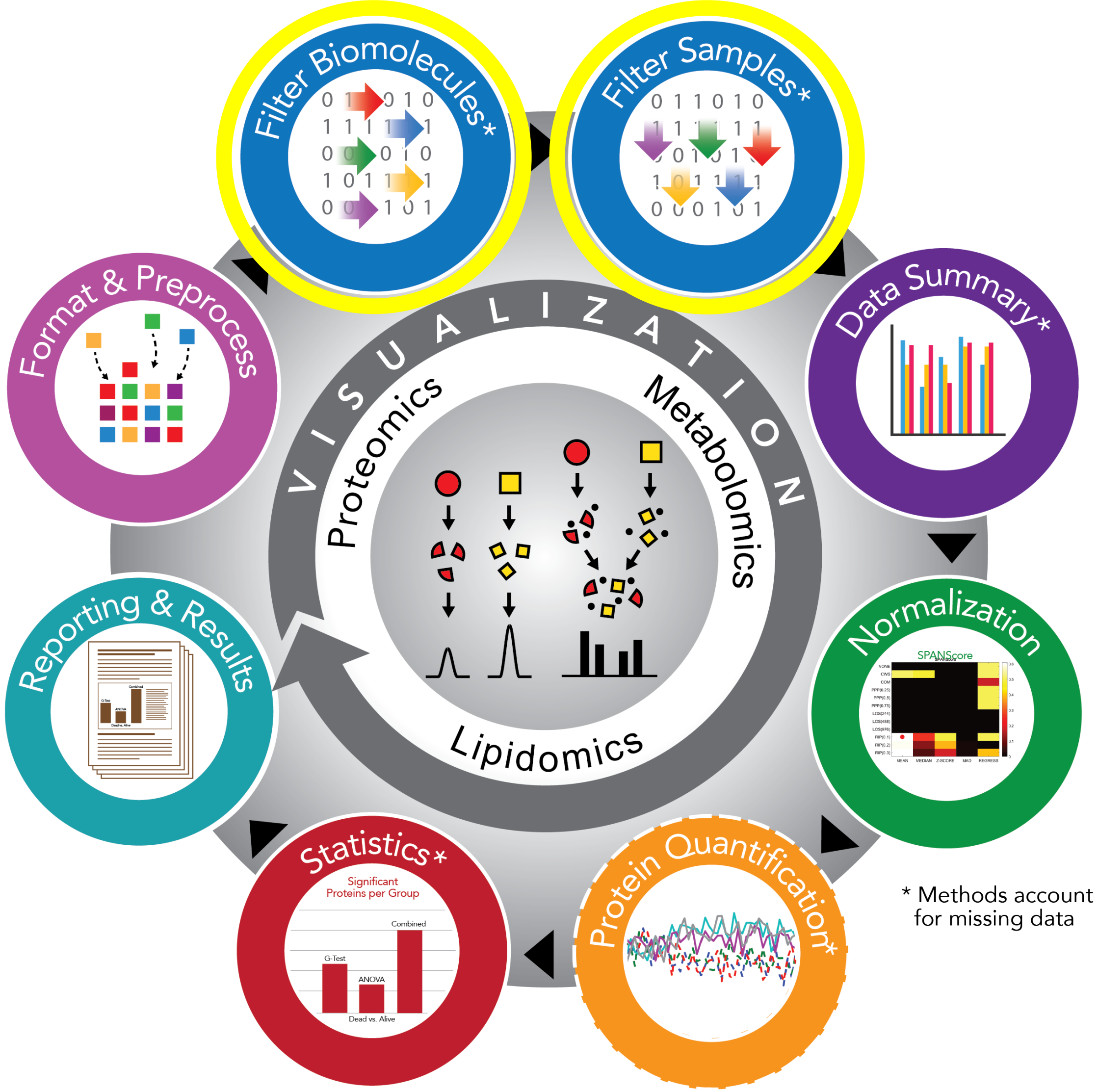
- Relative quantities w/highly skewed distributions
- log transform for Normal assumption in downstream analyses





Typical Statistical Processing

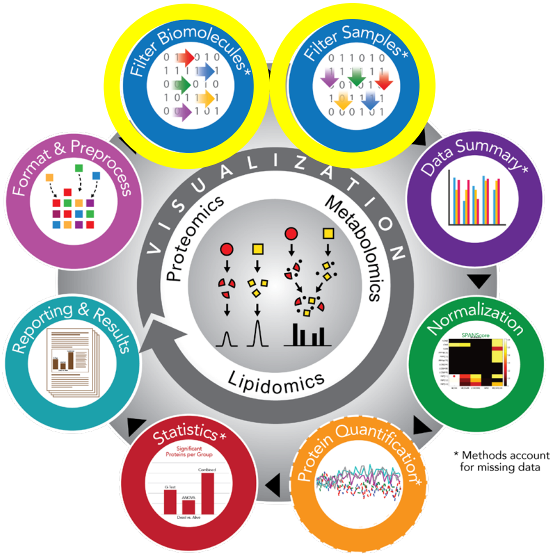
Filters



* Methods account for missing data

Typical Statistical Processing

Filters

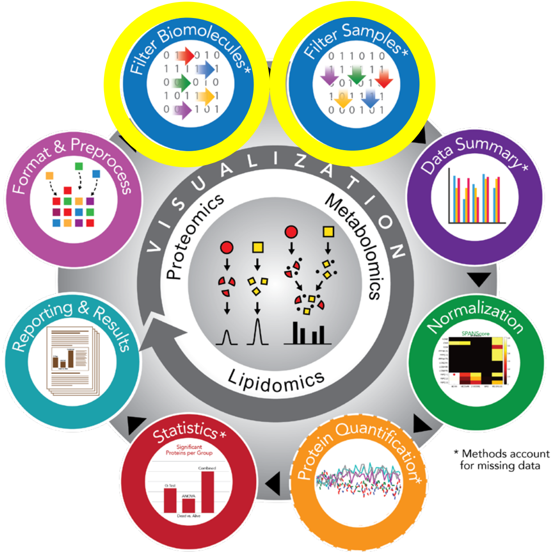


* Methods account for missing data

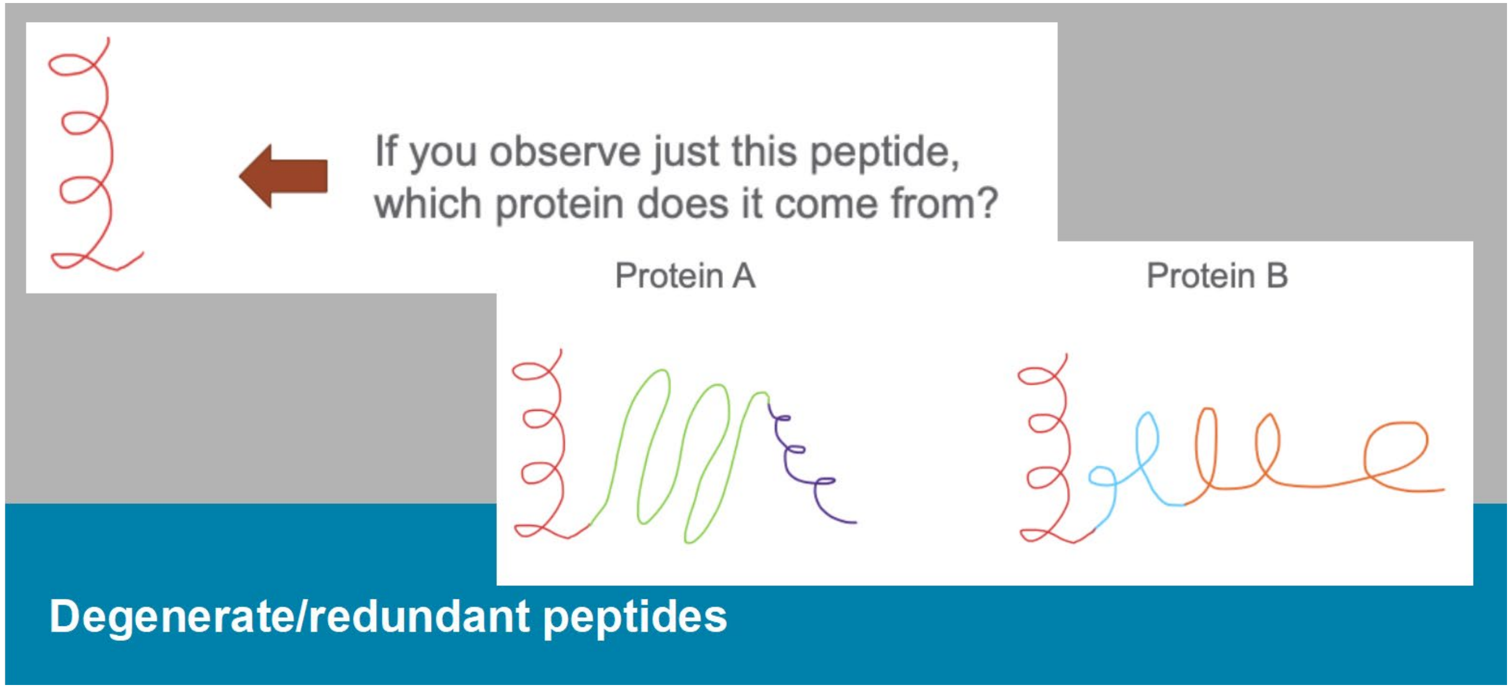
- Proteomics-specific filters
  - Degenerate / redundant peptides
  - One-hit-wonders
  - Reverse hit peptides
  - Contaminant proteins
- Other common filters
  - Molecule occurrence
  - Coefficient of variation
  - Sample outliers

Typical Statistical Processing

Proteomics-Specific



- Degenerate / redundant peptides
- One-hit-wonders – peptides observed just once
- Reverse hit peptides – for false discovery rate
- Contaminant proteins – from sample prep or accidental



If you observe just this peptide, which protein does it come from?

Protein A

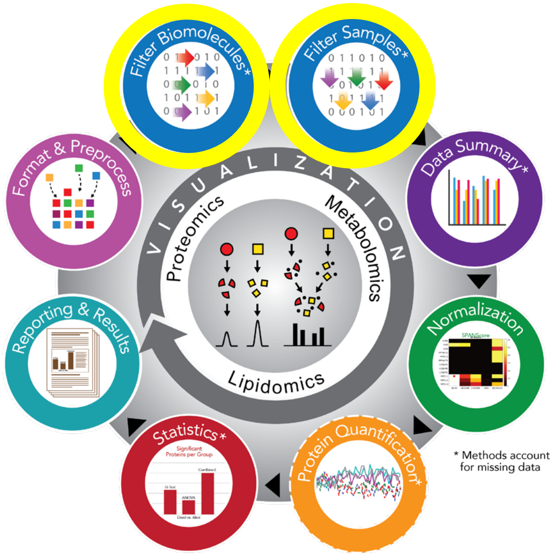
Protein B

Degenerate/redundant peptides

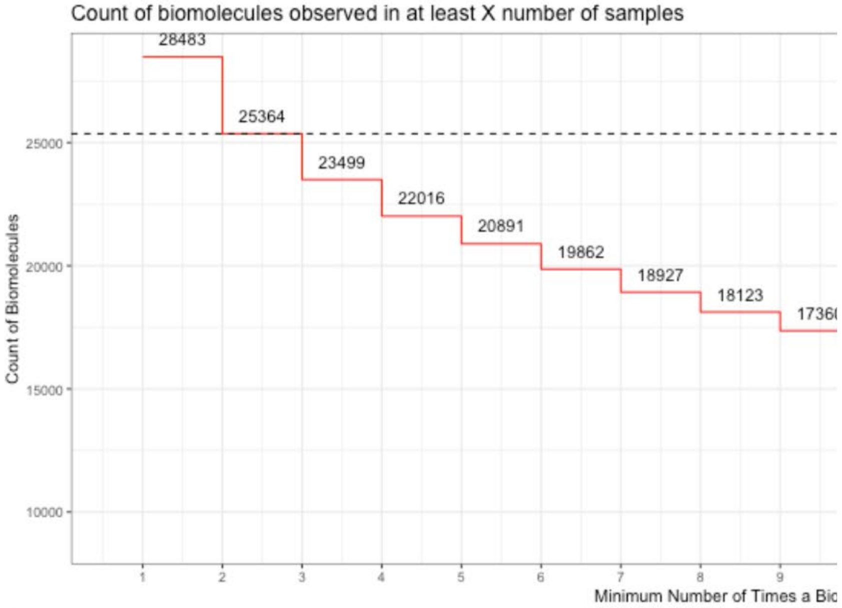
Typical Statistical Processing

Other Common Filters

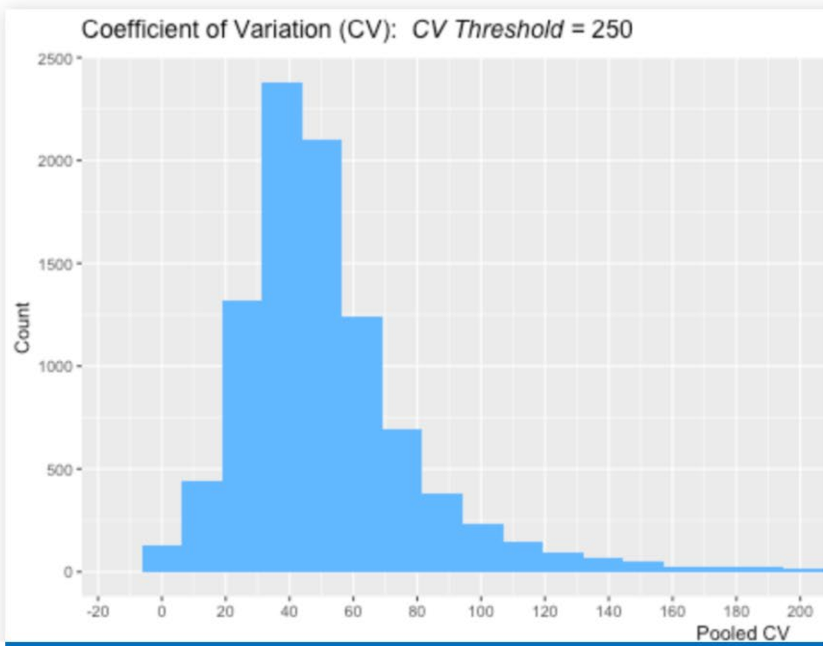
- Molecule filter
- Coefficient of variation filter



* Methods account for missing data



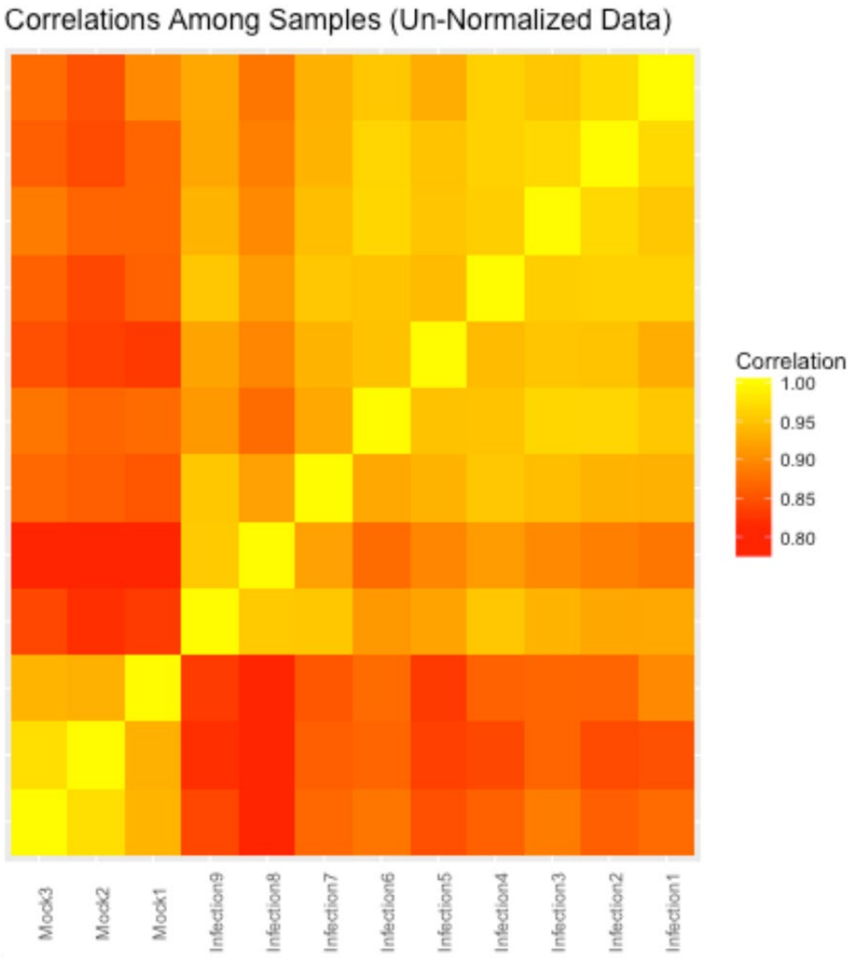
Molecule Filters



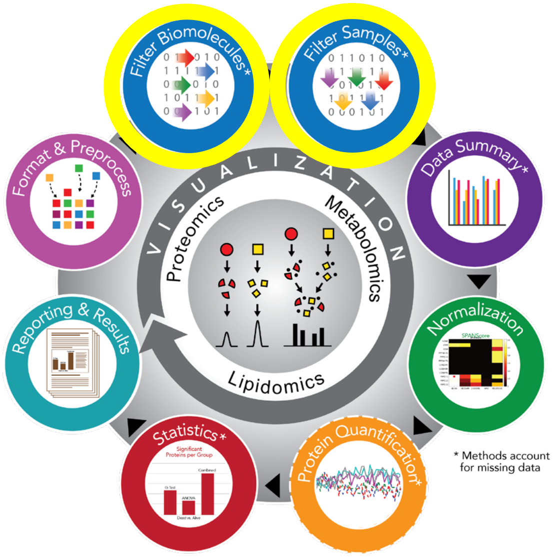
Coefficient of Variation

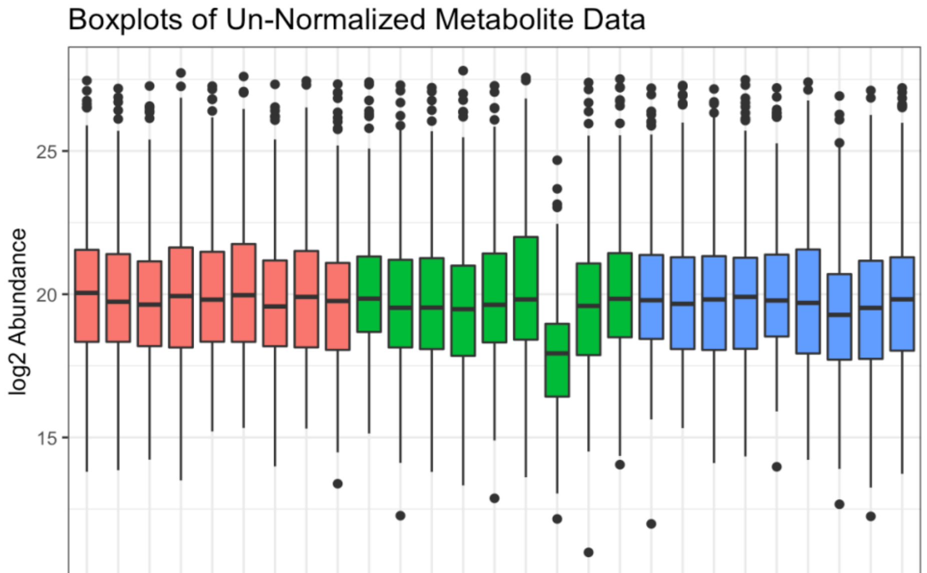
Typical Statistical Processing

Sample Filters



Correlation Heatmap



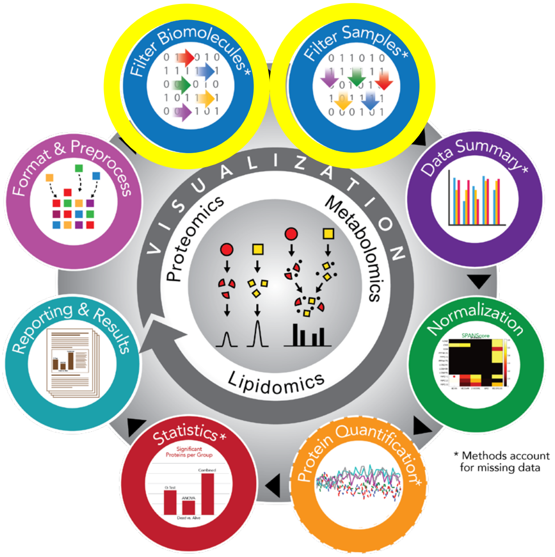


Boxplots/Visualization

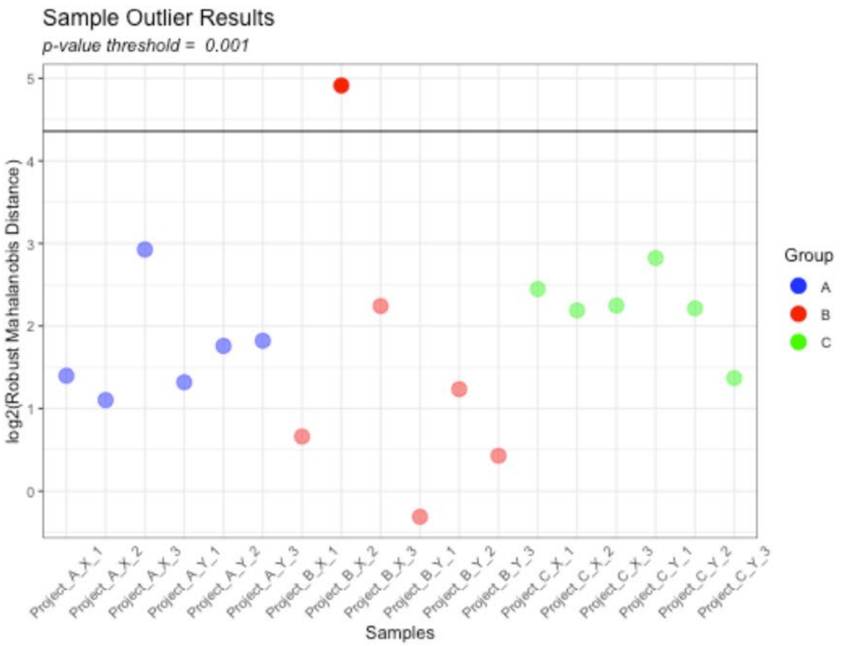
Typical Statistical Processing

Sample Filters

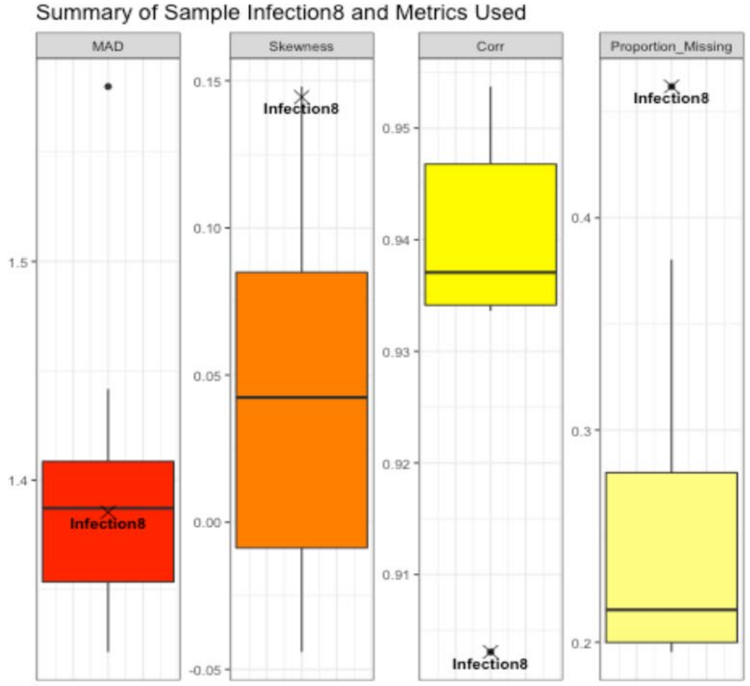
- Objective identification
  - MAD
  - Skewness
  - Kurtosis
  - Correlation
  - Percent Missing



* Methods account for missing data



Outlier Metric

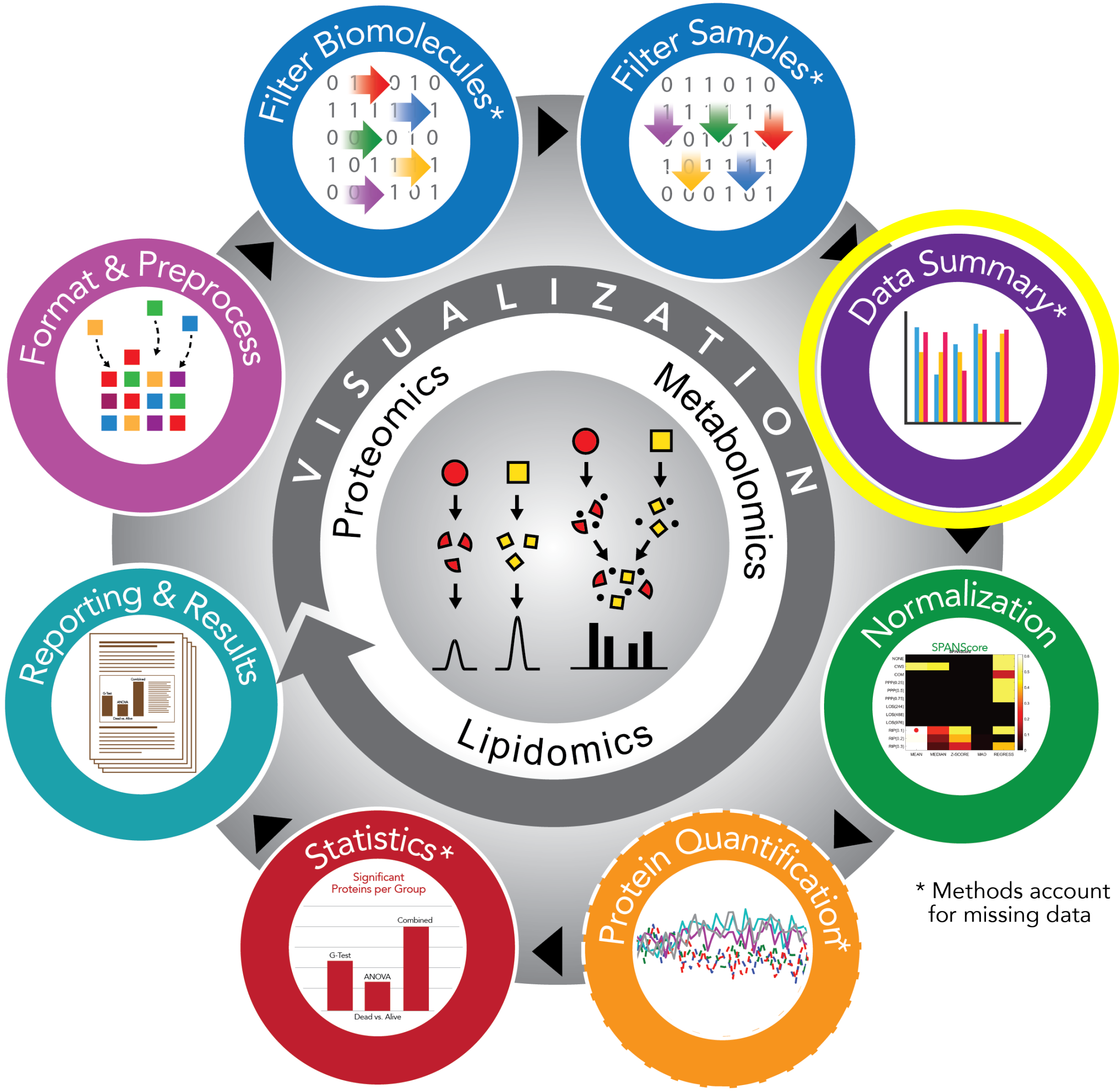


Outlier Assessment

Matzke et al. 2011 – Bioinformatics

Typical Statistical Processing

Data Summary / Exploratory Data Analysis

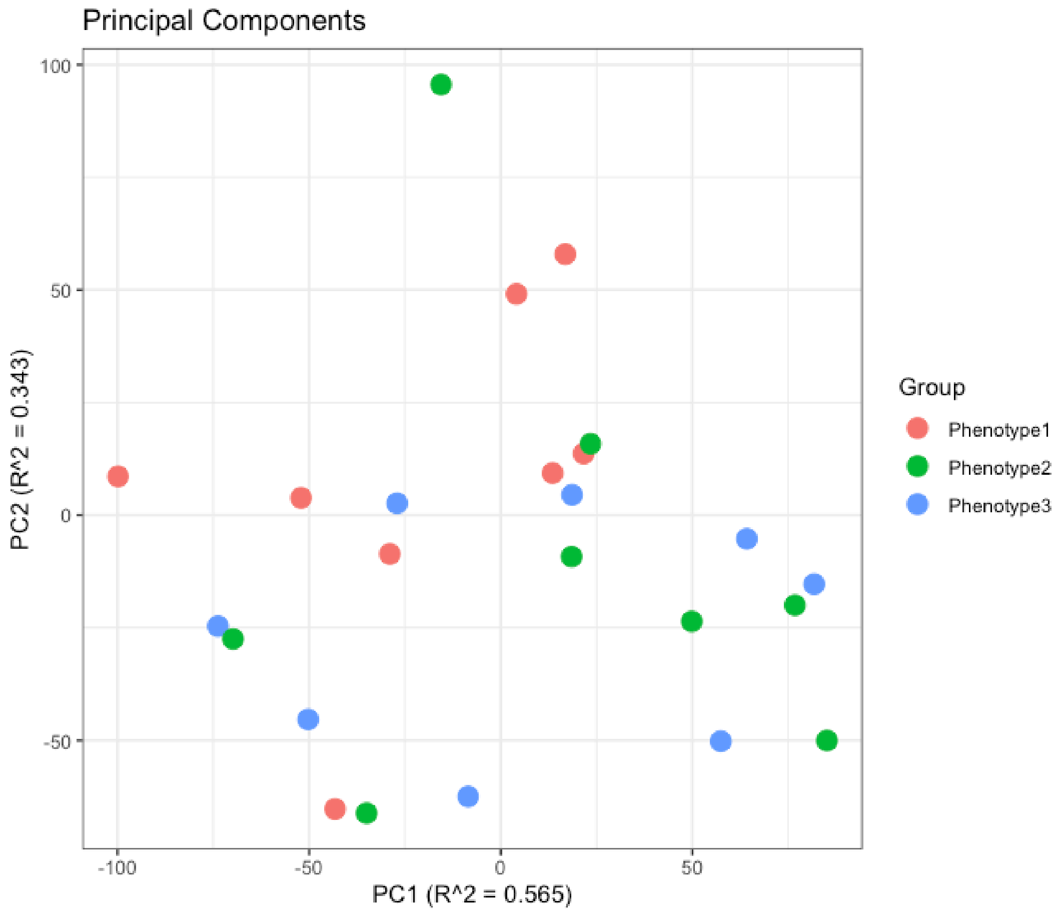


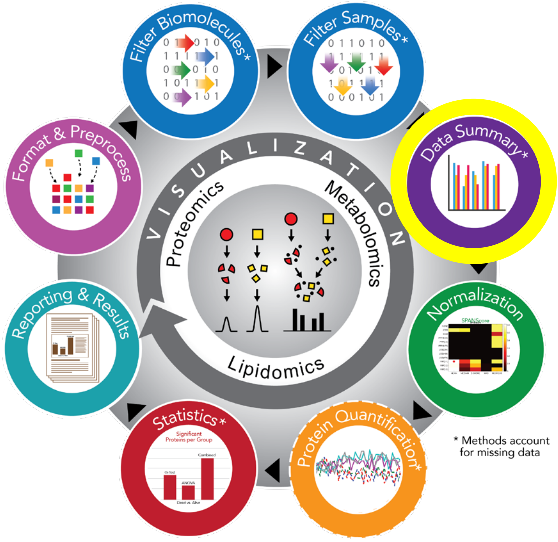
* Methods account for missing data

Typical Statistical Processing

EDA

- PCA plot




VISUALIZATION

Proteomics

Metabolomics

Lipidomics

Filter Biomolecules*

Filter Samples*

Data Summary*

Normalization

Protein Quantification*

Statistics*

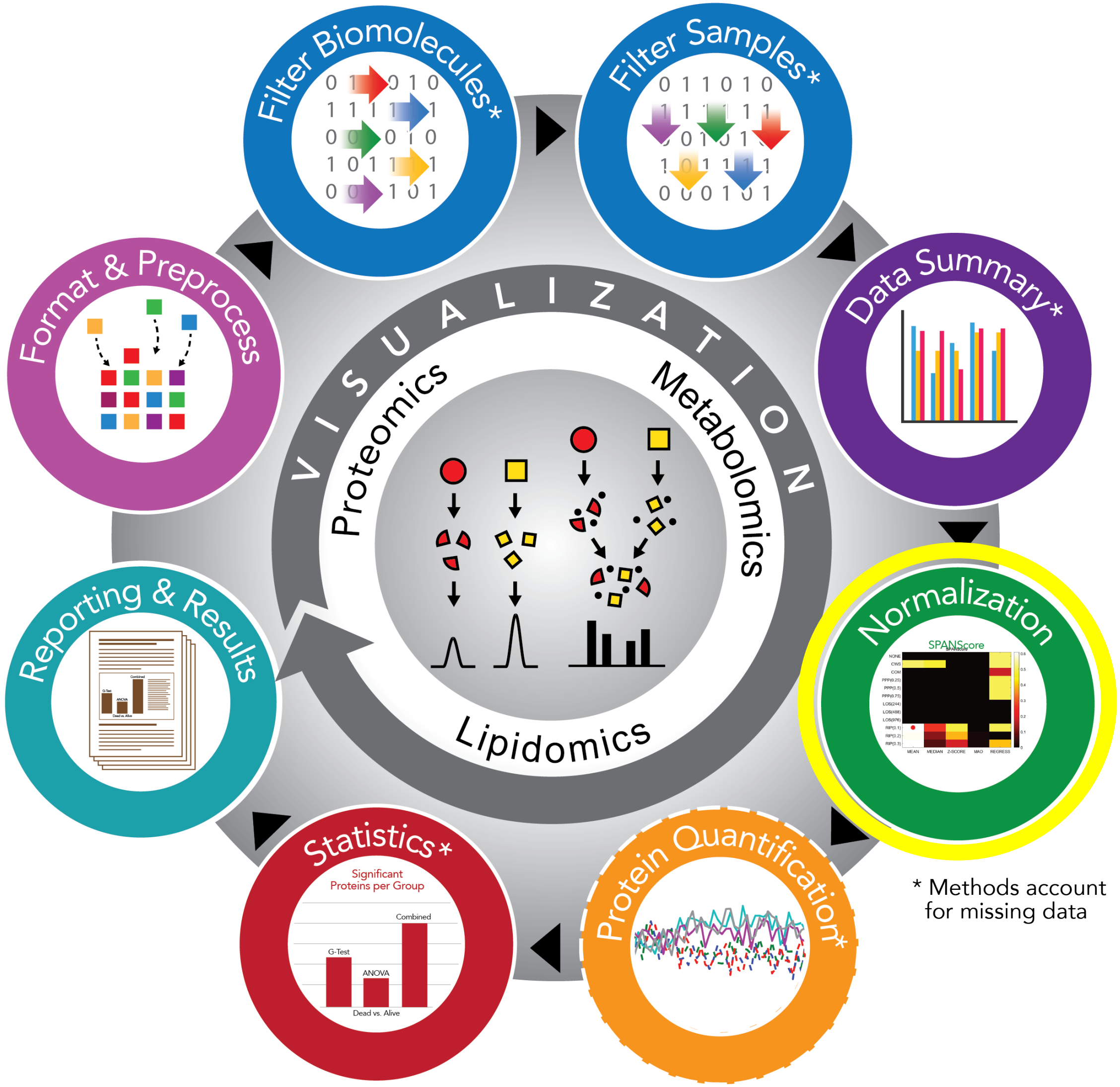
Reporting & Results

Format & Preprocess

* Methods account for missing data

Typical Statistical Processing

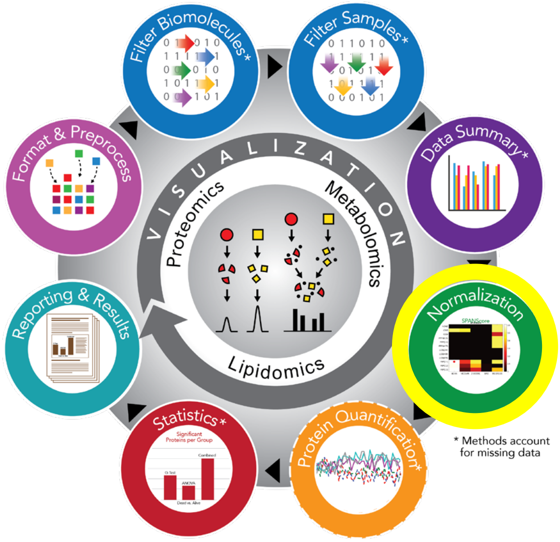
Normalization



* Methods account for missing data

Typical Statistical Processing

Normalization

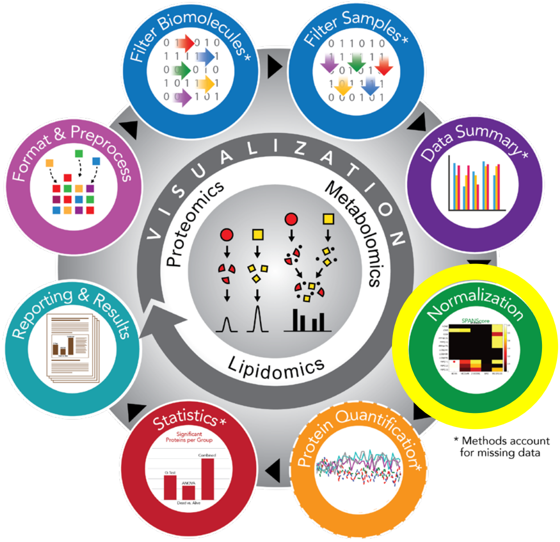


- **Put relative quantities on comparable scale from sample to sample**
- Normalization aims to remove unwanted variability due to
  - Sample prep & handling
  - Sample storage
  - Instrumentation – run to run variability, same machine over time, different machines, etc.

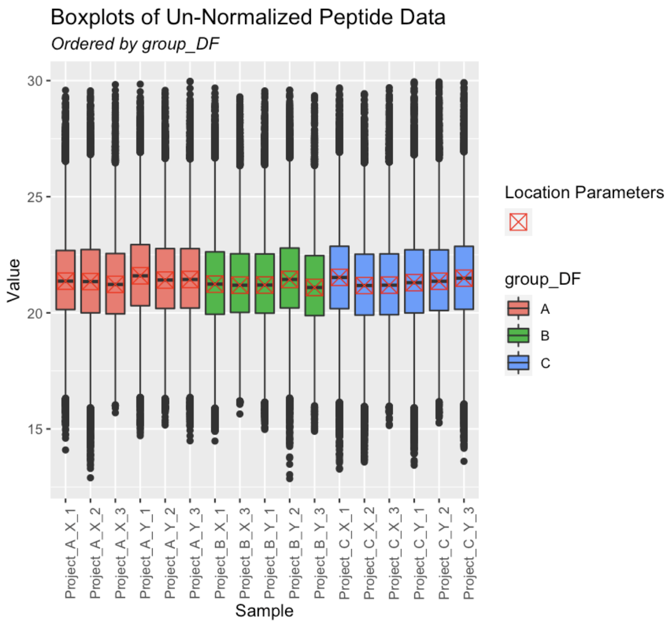
Typical Statistical Processing

Normalization

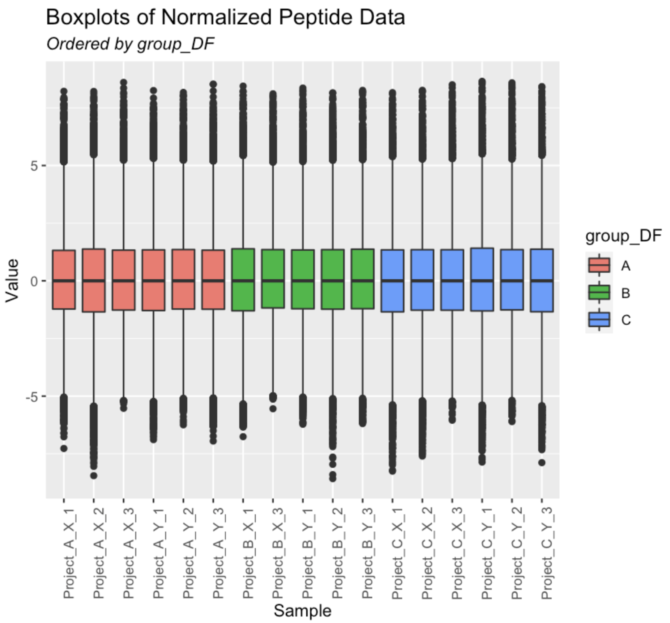
- **No single method works for all data types or experiments**
  - Scaling factors may/may not be affected by biological differences
  - Significance can be introduced or removed by normalization
  - Assure normalization method does not introduce bias (Webb-Robertson et al. 2011 – Proteomics)



* Methods account for missing data


Boxplots of Un-Normalized Peptide Data

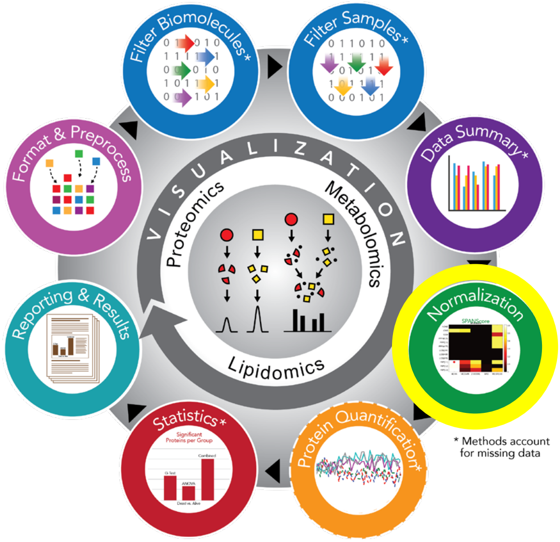
Ordered by group_DF


Boxplots of Normalized Peptide Data

Ordered by group_DF

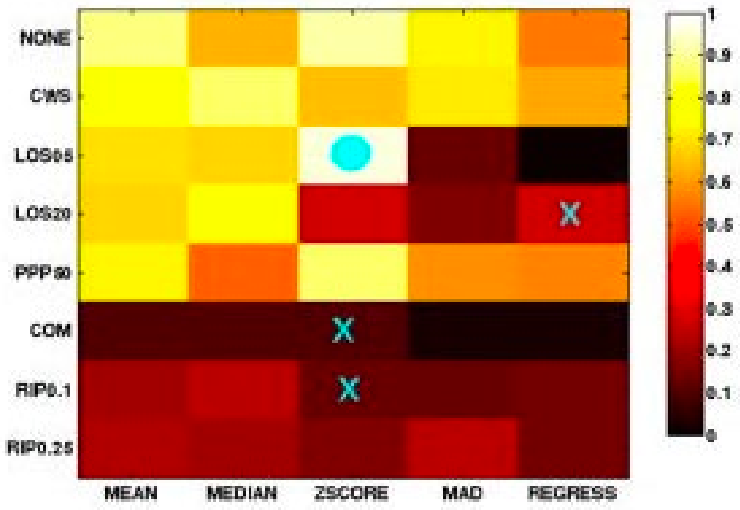
Typical Statistical Processing

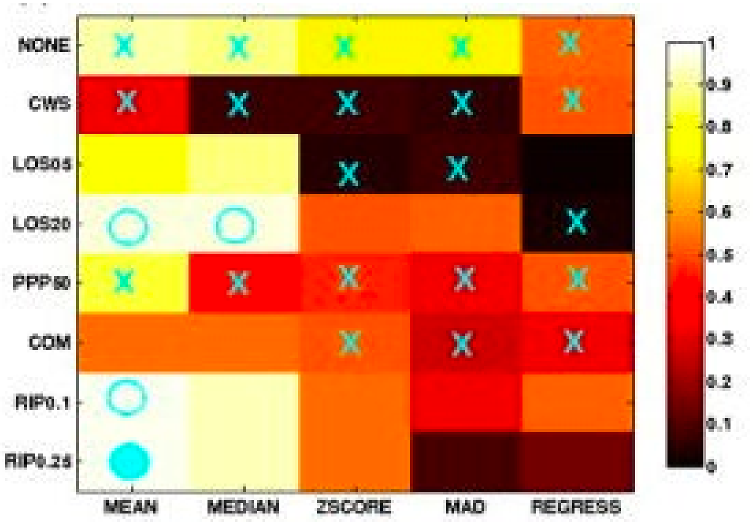
Normalization



* Methods account for missing data

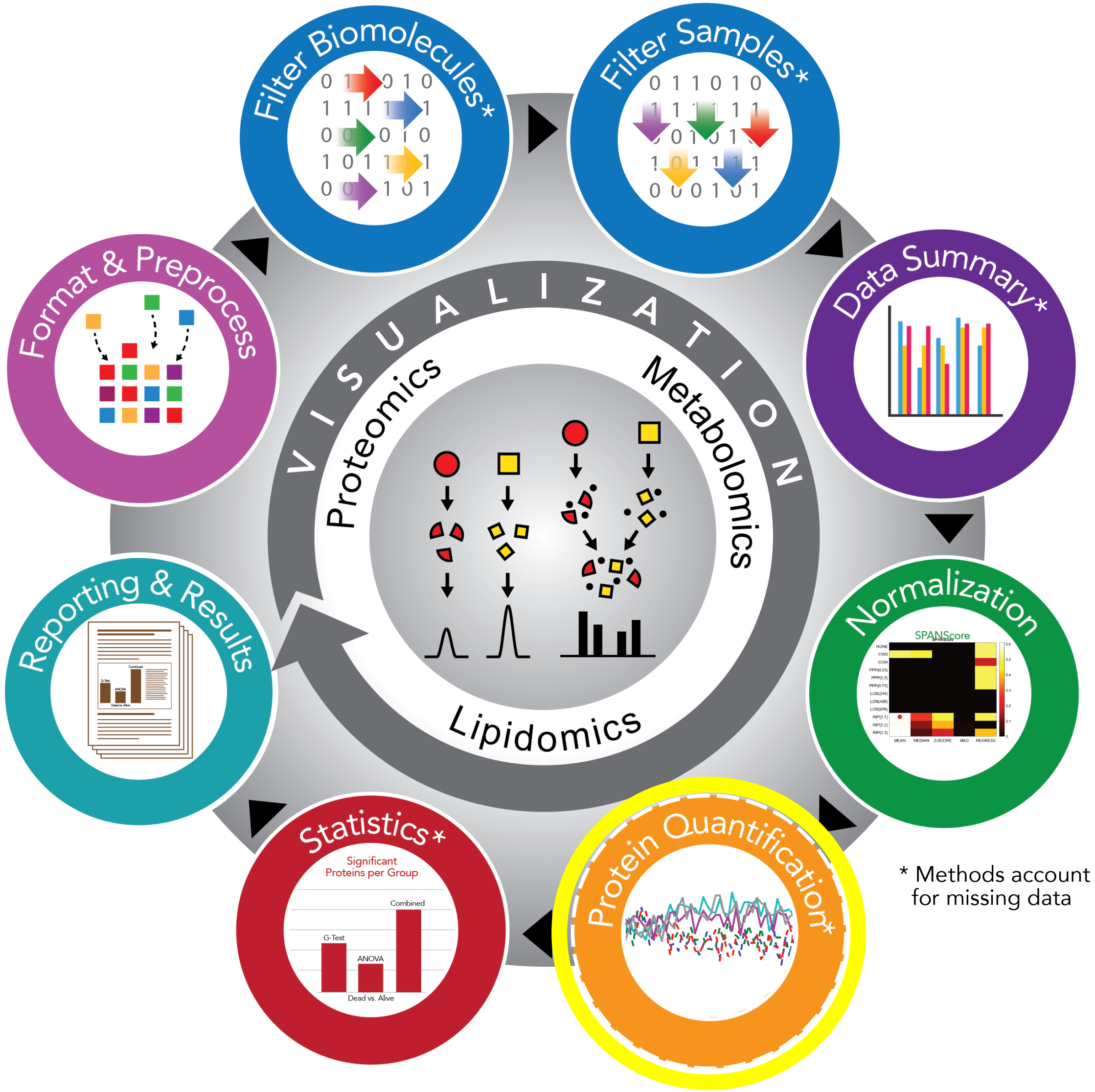
- Common methods
  - Median or mean centering
  - Based on all data or subset of data
- **SPANS** (Webb-Robertson et al. 2011 – Proteomics)
  - With sufficient # of biomolecules (like peptides), we can utilize more sophisticated, data driven techniques to identify an appropriate normalization method





Typical Statistical Processing

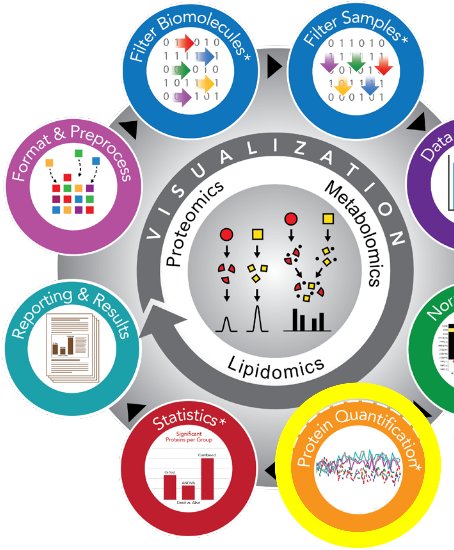
Protein Quantification



* Methods account for missing data

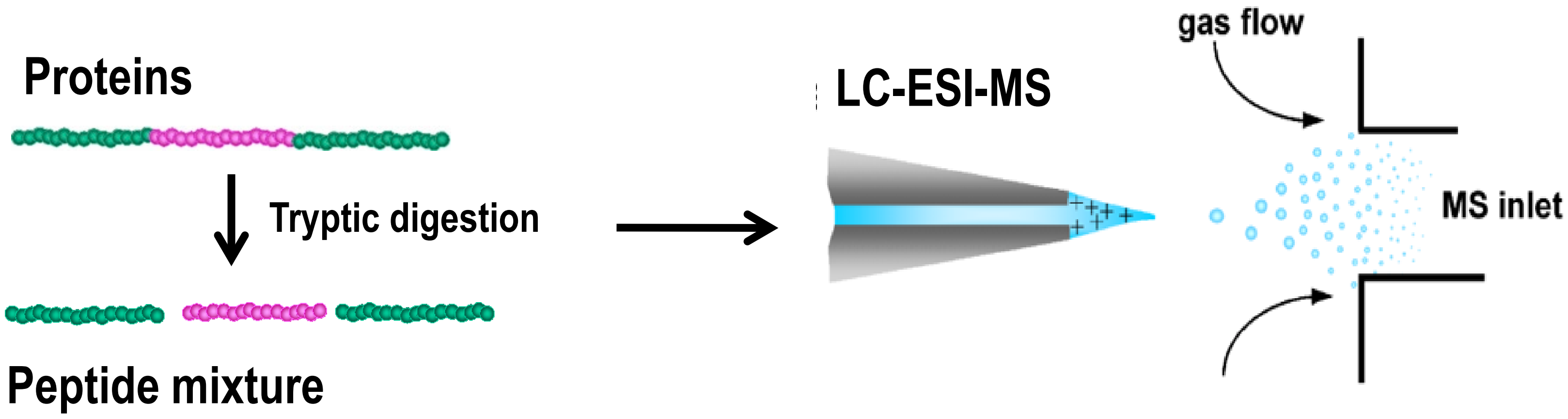
Typical Statistical Processing

Protein Rollup / Quantification



* Methods account for missing data

- We have peptide level measurements
- We are interested in proteins, which map to genes and pathways

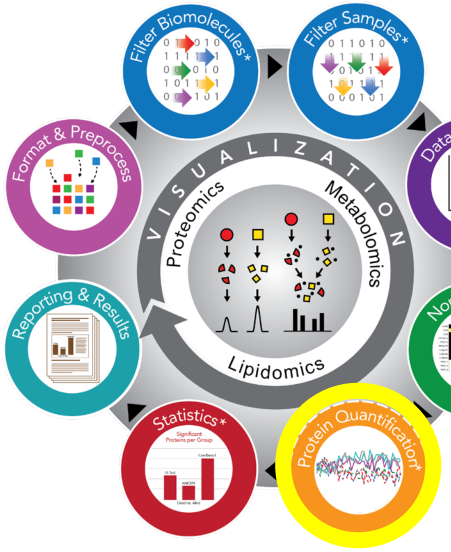


- Protein quantification is the process of assigning each peptide to one protein and summarizing the observed abundances at the protein level

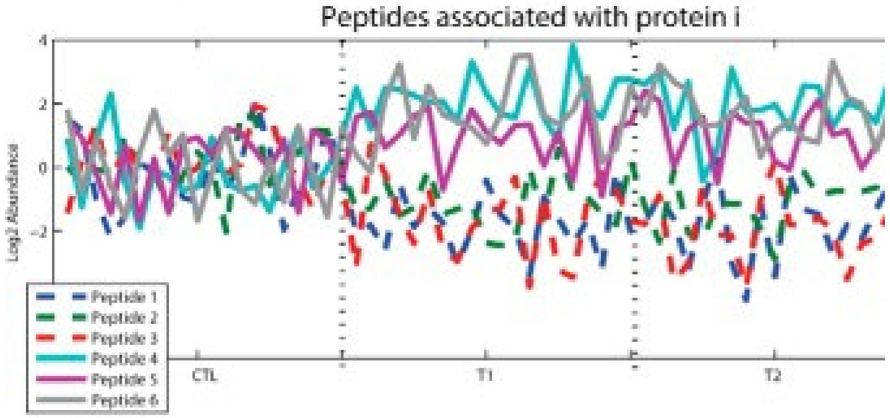
Typical Statistical Processing

Protein Rollup / Quantification

- Common Methods
  - rollup
  - r-rollup – reference peptide selected & used to scale other peptides
  - q-rollup – quantile-based threshold filters peptides
  - z-rollup – peptides scaled by computing z-score
- Account for Isoforms
  - BP-Quant (Webb-Robertson et al. 2014-Mol Cell Proteomics)
  - PQPQ (Forshed 2013-Methods Mol Biol)



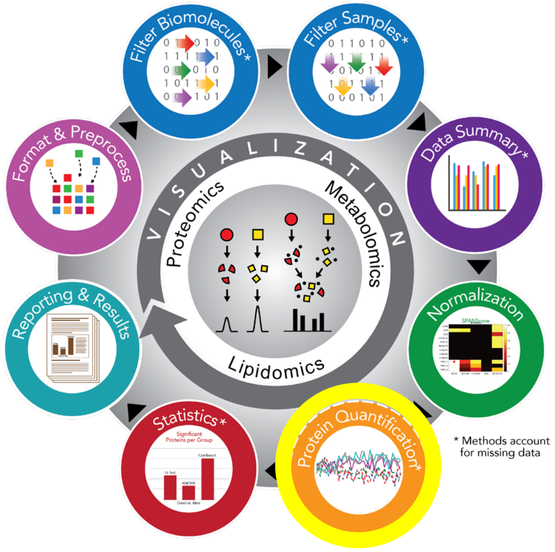
* Methods account for missing data



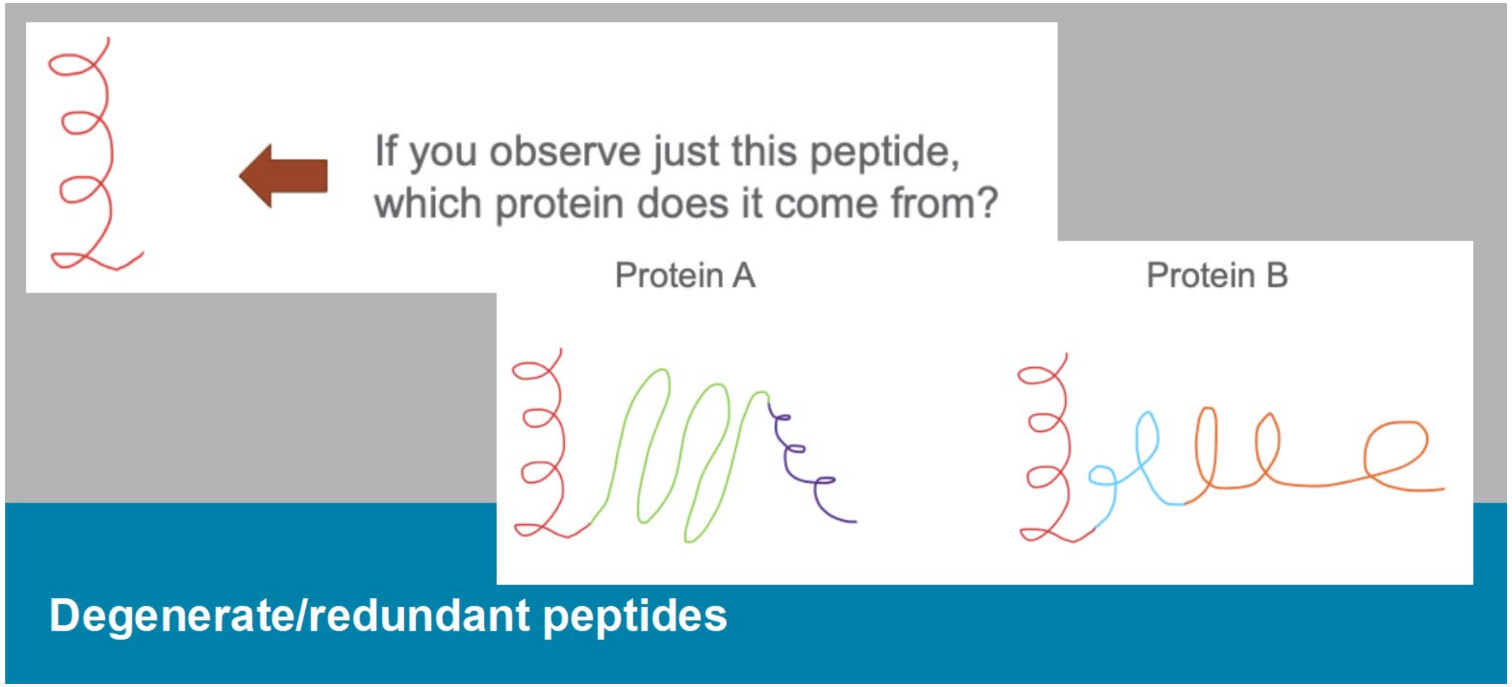
Typical Statistical Processing

Protein Rollup / Quantification

- For metaproteomics, redundant peptides must be handled with care – an ongoing research area



* Methods account for missing data



If you observe just this peptide, which protein does it come from?

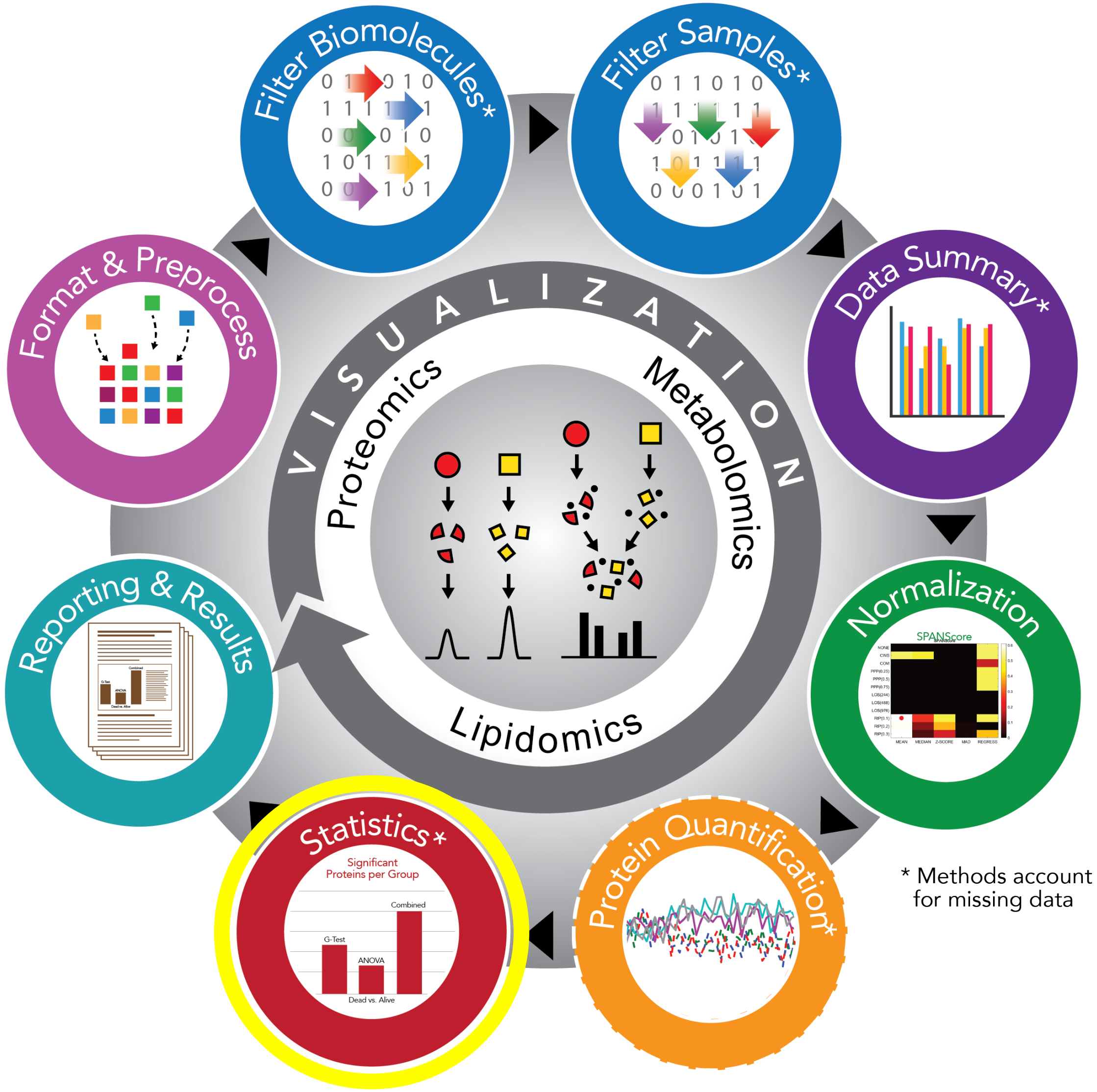
Protein A

Protein B

Degenerate/redundant peptides

Typical Statistical Processing

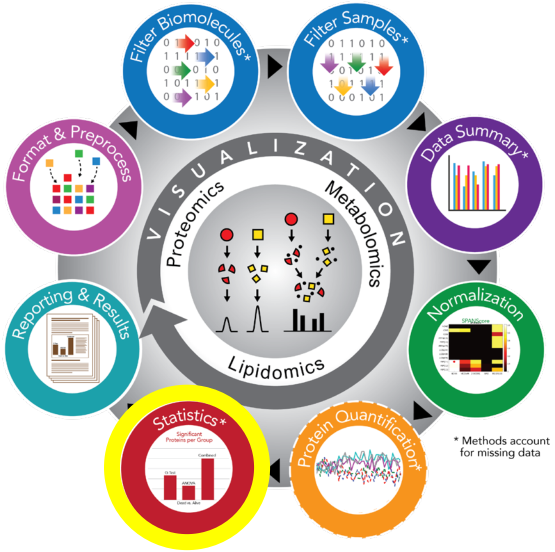
Statistical Comparisons



* Methods account for missing data

Typical Statistical Processing

Stats for Standard Experimental Designs*



* Methods account for missing data

- Quantitative test: are there differences in the mean abundances of each biomolecule between the treatments/groups?
- Qualitative test: are patterns of presence/absence for each biomolecule associated with treatment group?
- Using quantitative and qualitative statistical tests shown to improve identification of significant peptides/proteins (JPR Webb-Robertson et al. 2010)

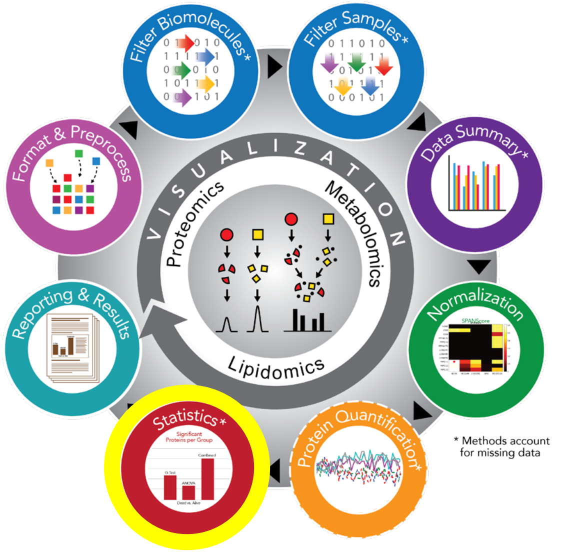
Protein	Abundance Data							
	Group1_1	Group1_2	Group1_3	Group1_4	Group2_1	Group2_2	Group2_3	Group2_4
A	16.4	16.9	16.2	16.7	16.9	17.2	17.5	17.9
B	NA	NA	NA	NA	17.5	16.9	17.3	17.1
C	16.5	NA	NA	16.3	17.0	16.8	NA	17.2

*single point in time experiments, comparisons between groups

Typical Statistical Processing

Quantitative Test: ANOVA

- Filter out unreliable biomolecules – i.e., not enough data for statistics



* Methods account for missing data

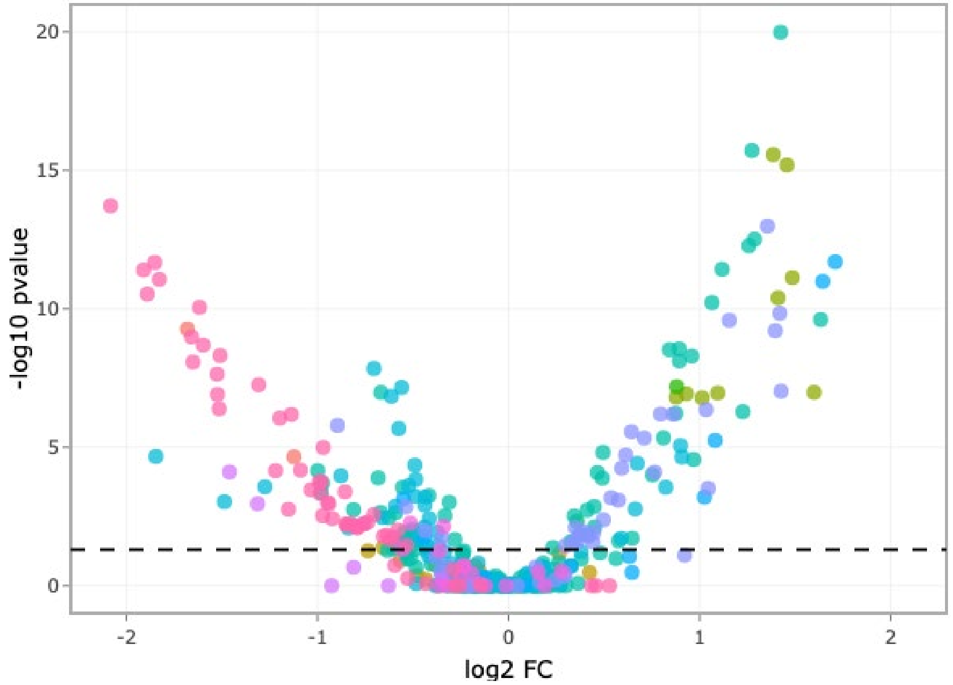
Protein	Abundance Data								
	Group1_1	Group1_2	Group1_3	Group1_4	Group2_1	Group2_2	Group2_3	Group2_4	
A	16.4	16.9	16.2	16.7	16.9	17.2	17.5	17.9	keep
B	NA	NA	NA	NA	17.5	16.9	17.3	17.1	remove
C	16.5	NA	NA	16.3	17.0	16.8	NA	17.2	keep

- Typically filter so there are at least 2 observations per biomolecule per group

Quantitative Test: ANOVA

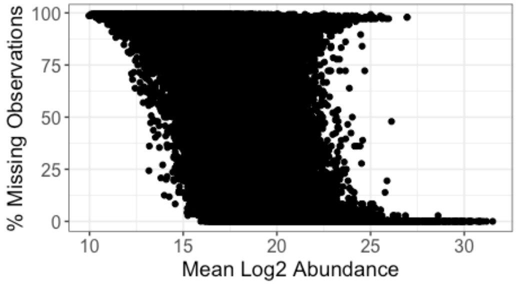
ANOVA (F test)

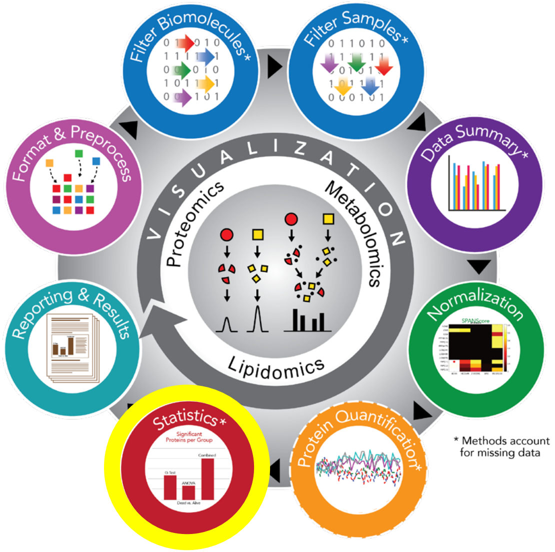
- Estimate fold change between groups for each molecule
- Volcano plot
  - Fold changes and threshold above/below (threshold depends on context)
  - Typically use to subset down to manageable list of biomolecules of interest



Typical Statistical Processing

Qualitative Test



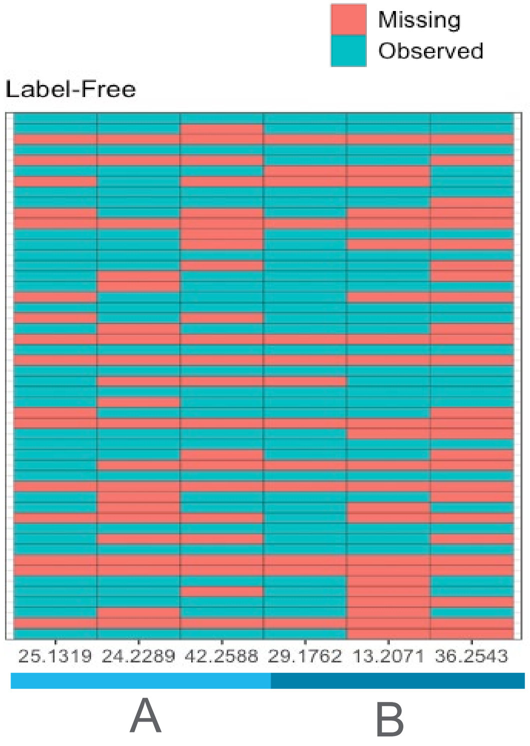


* Methods account for missing data

G test

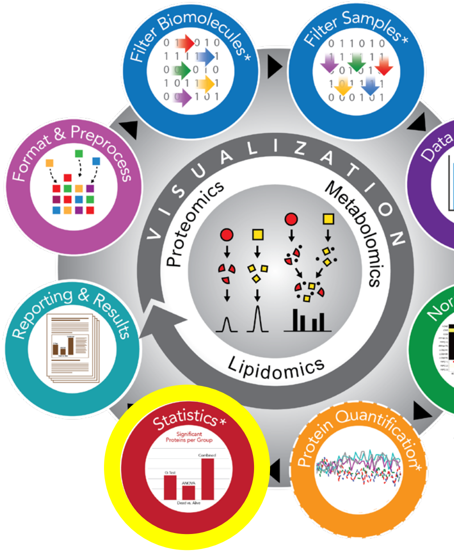
- Determine if proportion of missing values are associated with treatment group, compared to random chance
- Fisher’s test of independence with correction for small sample size
- Helpful when data don’t have enough quantitative info to do a quantitative test and/or in combination with quantitative test

Biomolecule 1	**Present**	**Absent**	**Total**
Treatment A	0	3	**3**
Treatment B	2	1	**3**
Total	**2**	**4**	



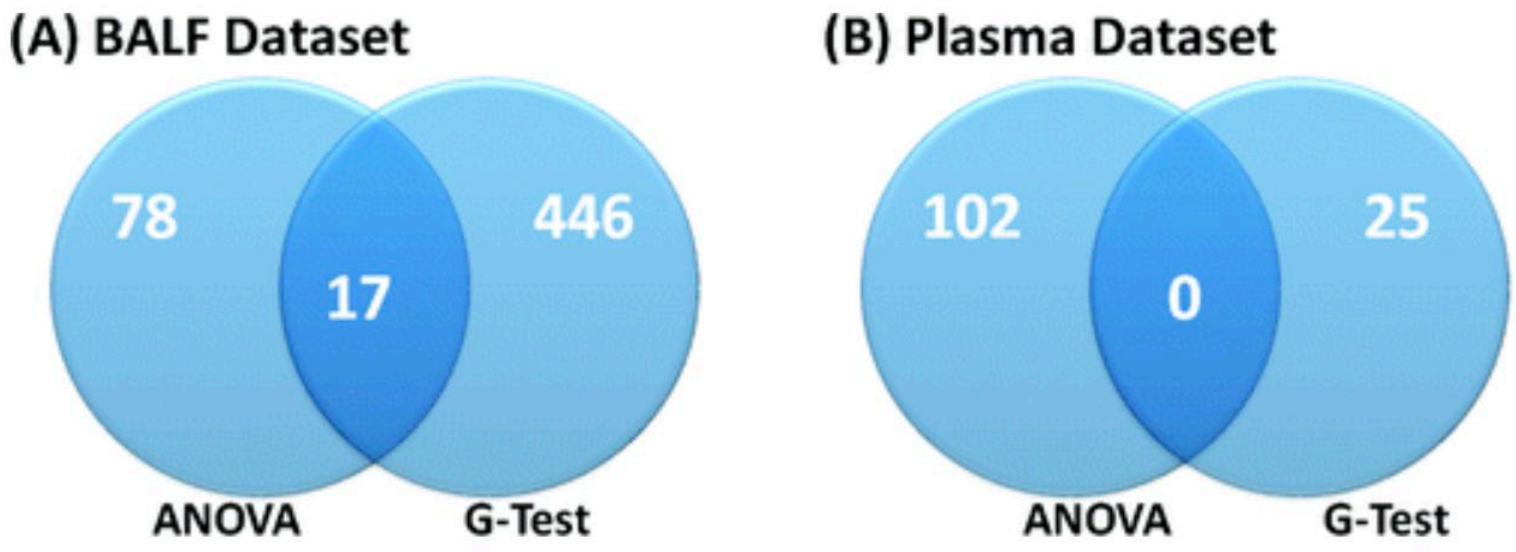
Typical Statistical Processing

Combined Quantitative & Qualitative Tests



* Methods account for missing data

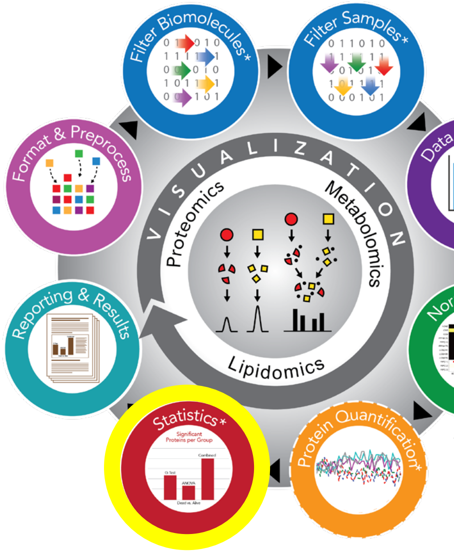
ANOVA and G-test



JPR Webb-Robertson et al (2010)

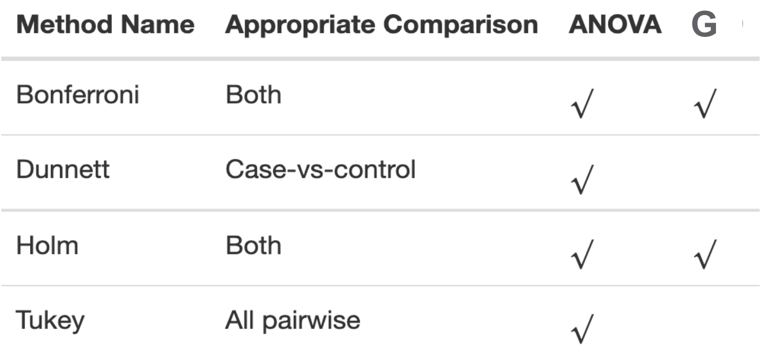
Typical Statistical Processing

Statistical Comparisons



* Methods account for missing data

- Multiple comparison adjustment methods
- (adjusted) p-values
  - Using p-value $\leq \alpha$ → $\leq \alpha$% Type 1 error rate
  - Multiple tests → error rate is inflated
  - Multiple tests for a biomolecule



Method Name	Appropriate Comparison	ANOVA	G
Bonferroni	Both	√	√
Dunnett	Case-vs-control	√	
Holm	Both	√	√
Tukey	All pairwise	√	

- Many tests – one for each biomolecule
  - Benjamini-Hochberg
  - Benjamini & Yekutieli

Questions?

Networking Break

10:40 – 10:50 a.m.

8:30-8:35 a.m.	Introduction	Kelly Stratton
8:35-9:25	Types of Proteomics	Paul Piehowski & David Degnan
9:25-9:35	Networking Break	
9:35-10:40	Typical Statistical Processing	Kelly Stratton
10:40-10:50	Networking Break	
10:50-11:40	**Biological Interpretation**	**David Degnan & Tyler Sagendorf**
11:40-11:45	Closing Remarks	David Degnan

EMSL

Biological Interpretation

David Degnan & Tyler Sagendorf

From Statistical Significance to Biological Stories

What are we trying to *understand* about the system?

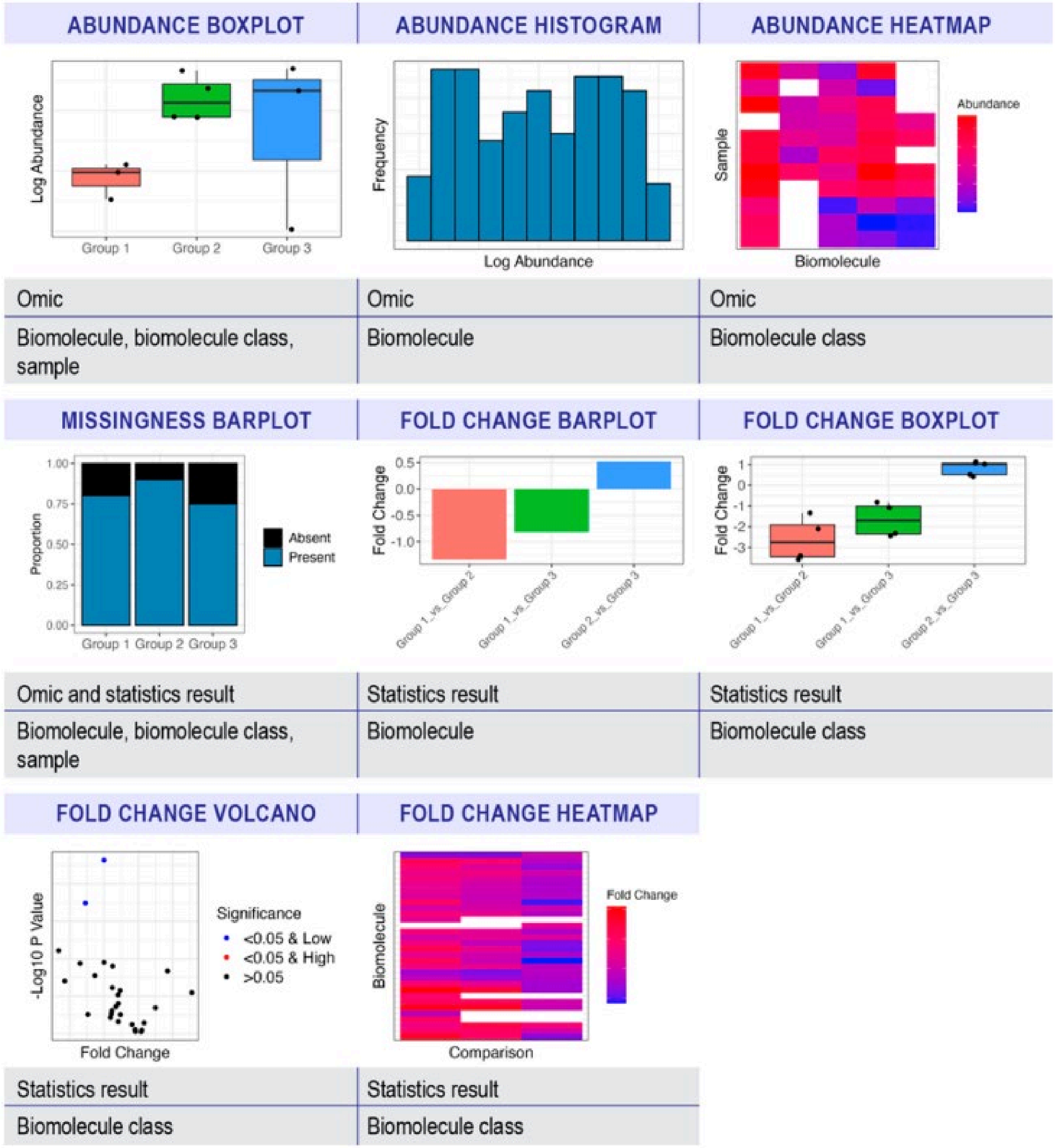
- …trends in peptide/protein abundances or fold-changes?
- …predictive power of peptides/proteins?
- …interrelationships of peptides/proteins, especially significant subsets?
- …enrichment analyses of all or significant peptides/proteins?

From Statistical Significance to Biological Stories

Understanding *Biomolecule Abundances / Fold Changes*

Understanding Biomolecule Abundances / Fold Changes

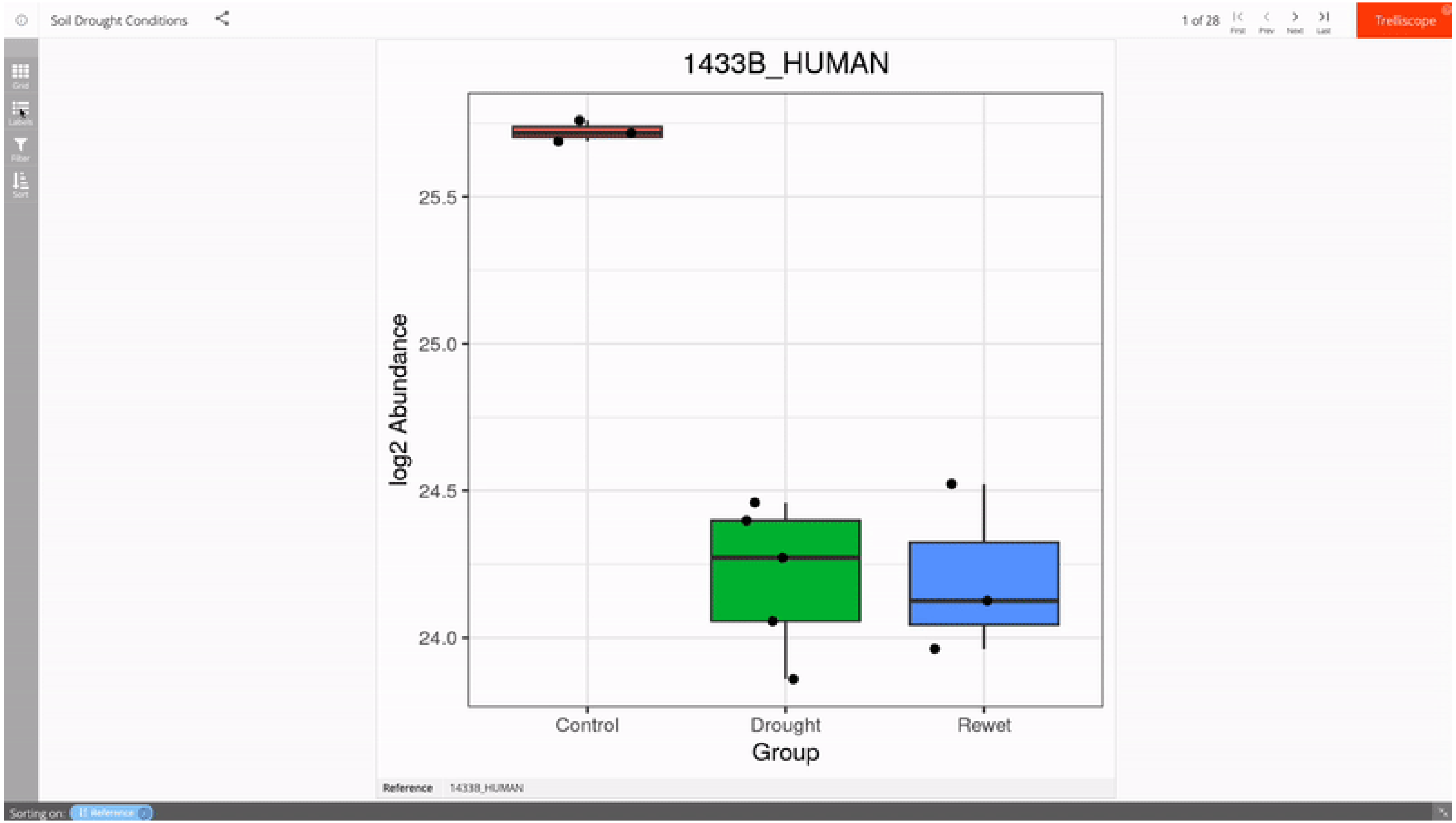
There are many ways to visualize data



	ABUNDANCE BOXPLOT	ABUNDANCE HISTOGRAM	ABUNDANCE HEATMAP
pmartR object	Omic	Omic	Omic
Panel by variable	Biomolecule, biomolecule class, sample	Biomolecule	Biomolecule class
	MISSINGNESS BARPLOT	**FOLD CHANGE BARPLOT**	**FOLD CHANGE BOXPLOT**
pmartR object	Omic and statistics result	Statistics result	Statistics result
Panel by variable	Biomolecule, biomolecule class, sample	Biomolecule	Biomolecule class
	FOLD CHANGE VOLCANO	**FOLD CHANGE HEATMAP**	
pmartR object	Statistics result	Statistics result	
Panel by variable	Biomolecule class	Biomolecule class	

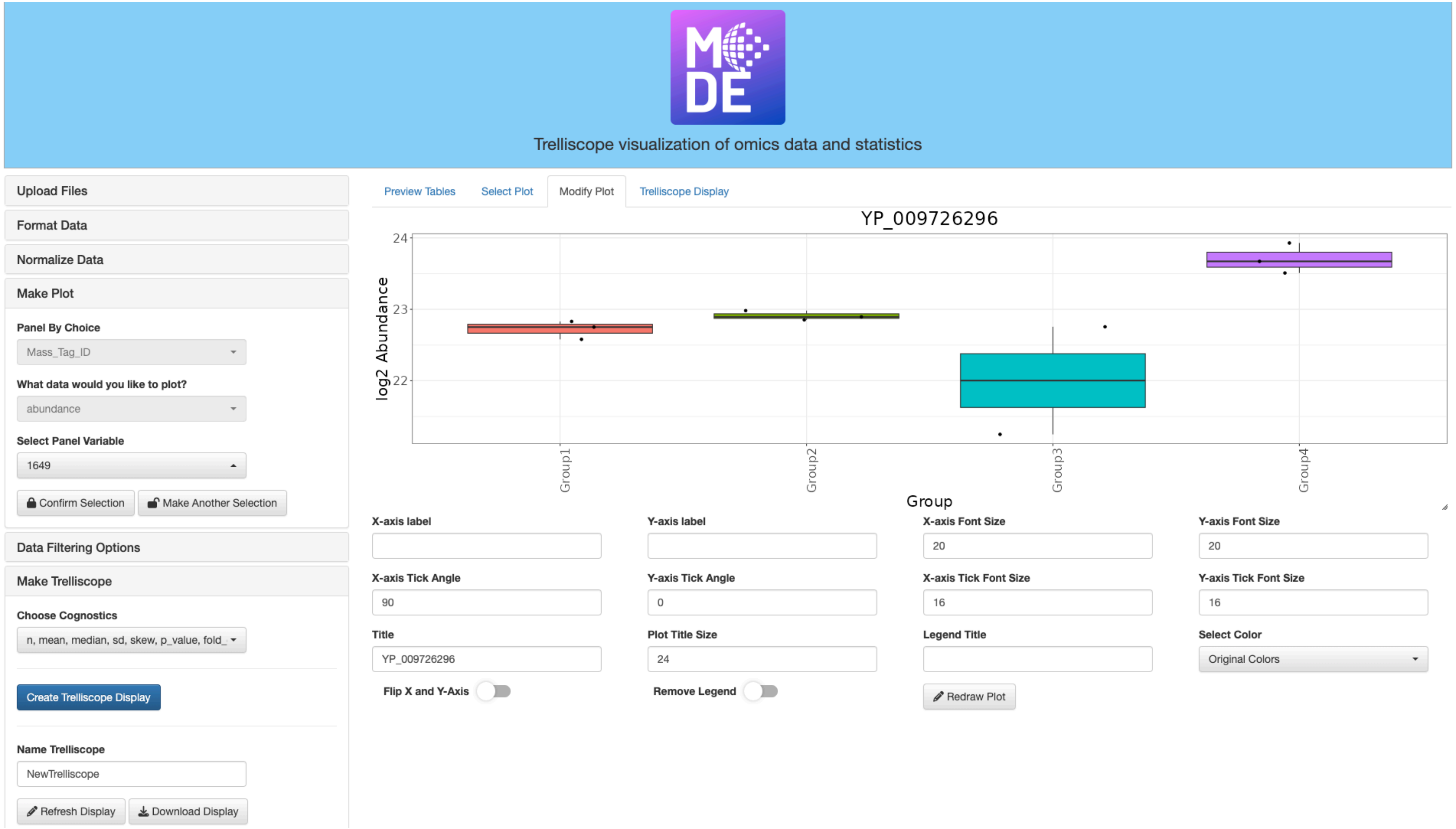
Understanding Biomolecule Abundances / Fold Changes

The Advantages of Trelliscope



Understanding Biomolecule Abundances / Fold Changes

The power of MODE



Panel By Choice: Mass_Tag_ID

What data would you like to plot?: abundance

Select Panel Variable: 1649

Choose Cognostics: n, mean, median, sd, skew, p_value, fold_

Name Trelliscope: NewTrelliscope

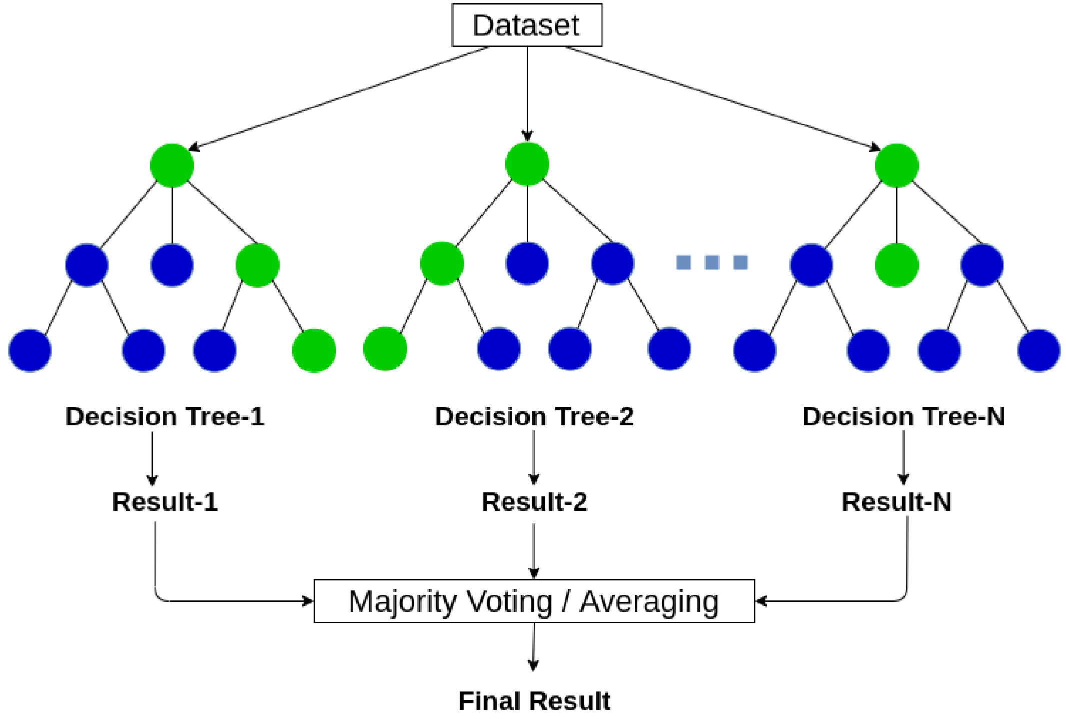
X-axis label	Y-axis label	X-axis Font Size	Y-axis Font Size
		20	20
X-axis Tick Angle	Y-axis Tick Angle	X-axis Tick Font Size	Y-axis Tick Font Size
90	0	16	16
Title	Plot Title Size	Legend Title	Select Color
YP_009726296	24		Original Colors

MODE (multi-omics data exploration)

From Statistical Significance to Biological Stories

Understanding *Predictors (Target Proteins)*

Understanding Predictors



Supervised ML (Prediction)

Understanding Predictors

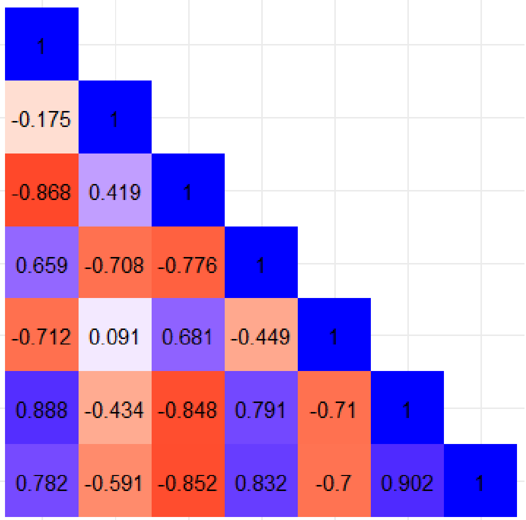
Algorithm	Outcome Type	Closed Equation	Variable Importance
Logistic Regression	Categorical – Binary	Yes	If variables are standardized
Random Forest	Categorical – Multiclass	No	Yes
Linear Regression	Continuous	Yes	If variables are standardized
K-nearest neighbors	Categorical	No	No
Naïve Bayes Classification	Continuous	Yes	No
Support Vector Machines	Categorical – Binary	Yes	No

From Statistical Significance to Biological Stories

Understanding *Biomolecular Relationships*

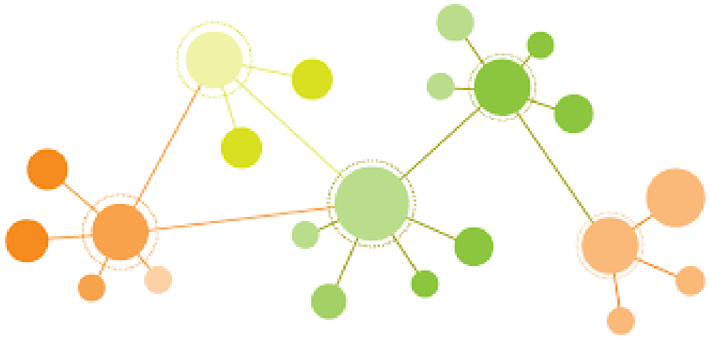
Biomolecular Relationships

Understanding *Biomolecular Relationships*

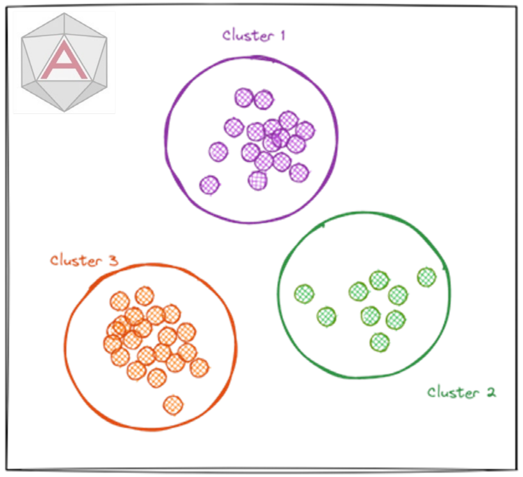


1						
-0.175	1					
-0.868	0.419	1				
0.659	-0.708	-0.776	1			
-0.712	0.091	0.681	-0.449	1		
0.888	-0.434	-0.848	0.791	-0.71	1	
0.782	-0.591	-0.852	0.832	-0.7	0.902	1

Correlation Matrices



Interaction Networks



Unsupervised ML (Clustering)

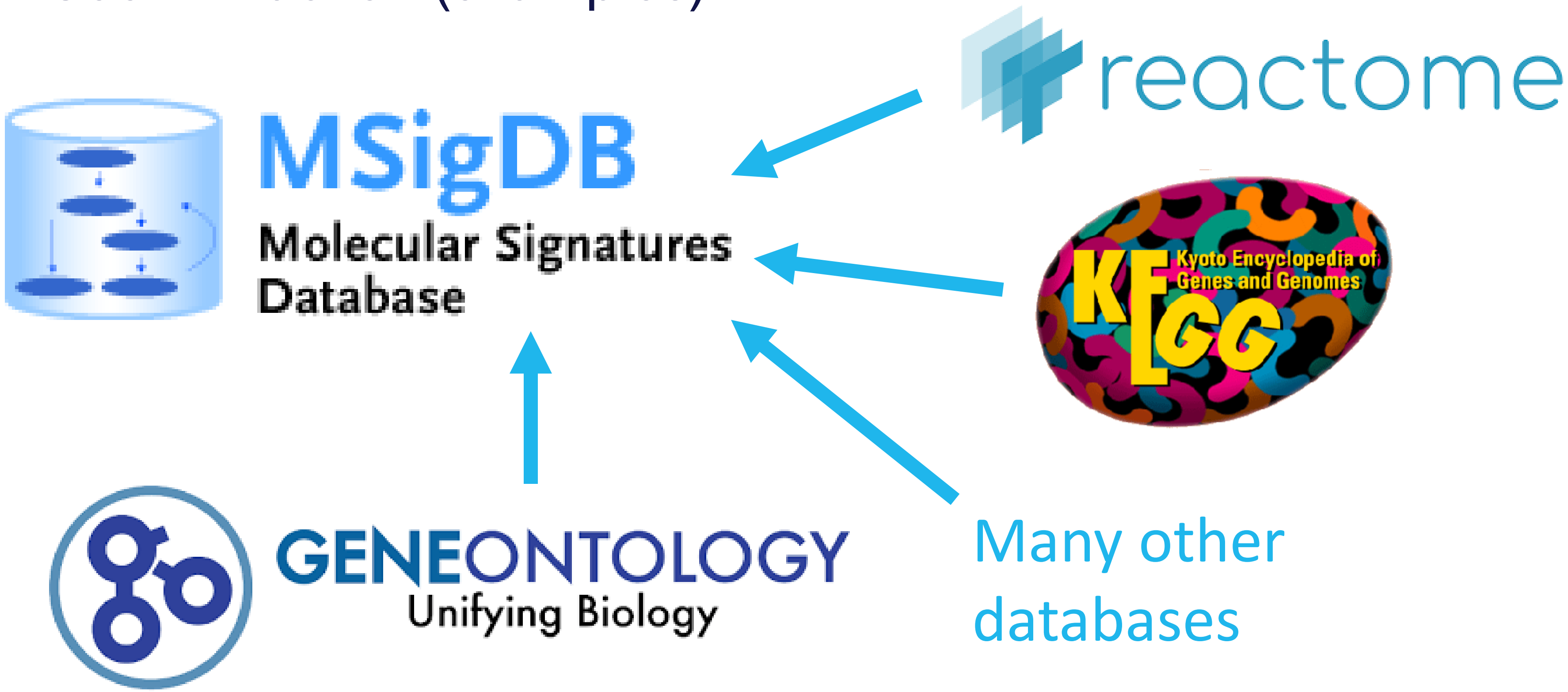
From Statistical Significance to Biological Stories

Understanding *Enrichment Analysis*

Overview: Enrichment Analysis and Set Databases

Enrichment Analysis + MSigDB

- How does enrichment analysis work?
- Distinction between enrichment analysis and over-representation analysis
- What are the advantages of enrichment analysis?
- See it in action (examples)


MSigDB
Molecular Signatures Database
reactome
Kyoto Encyclopedia of Genes and Genomes
KEGG
GENEONTOLOGY
Unifying Biology

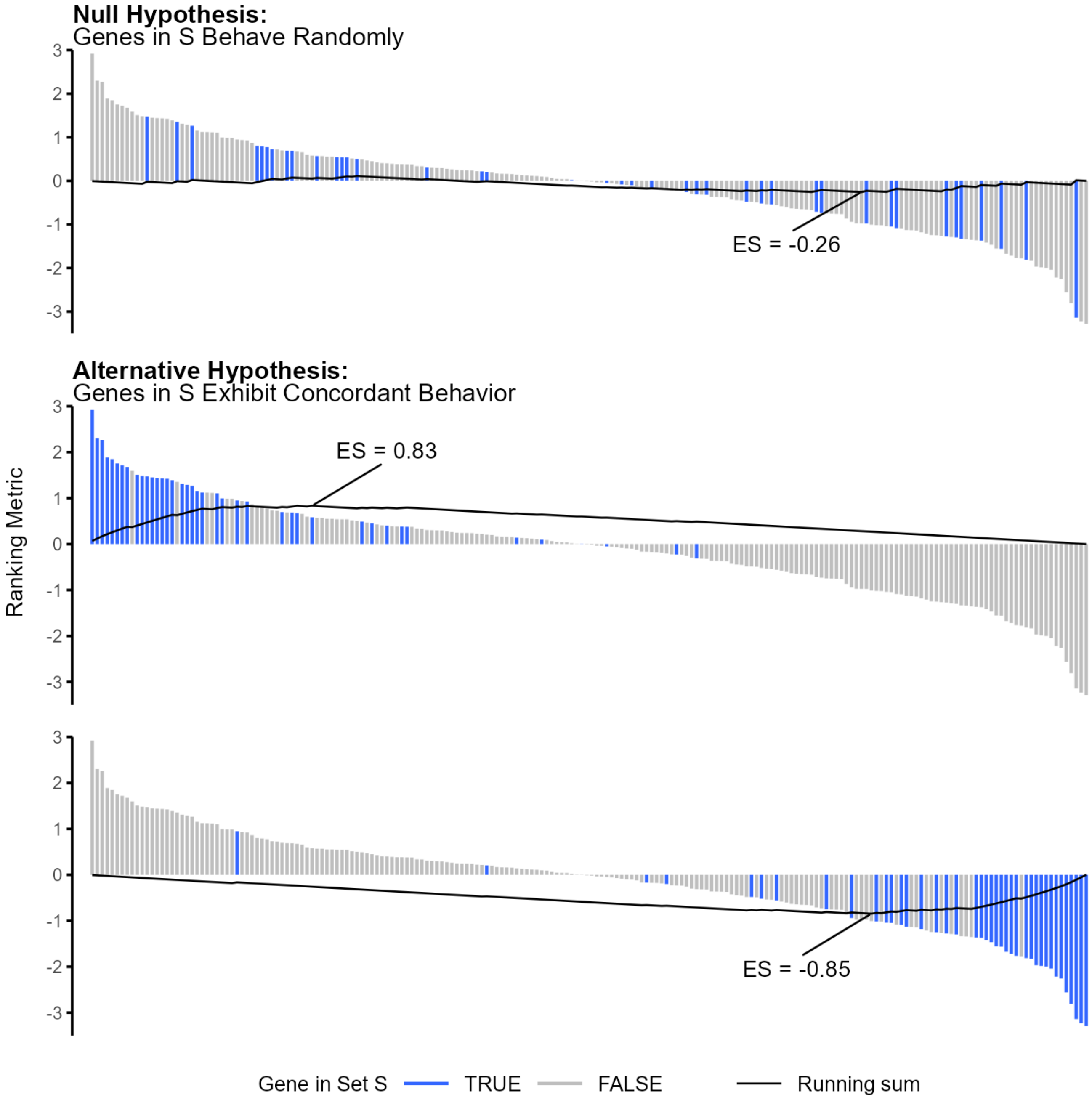
Many other databases

How does (pre-ranked) enrichment analysis work?

Main Idea: biomolecule-level statistics → set-level statistics

1. Obtain a pre-defined list of biomolecule sets to test.
2. Sort **all biomolecules** in the experiment in descending order by some statistic (“ranking metric”).
3. For each set from #1, determine if its members are primarily located in the top or bottom of the sorted vector from #2.
   1. Calculate a set-level Enrichment Score (ES) statistic
   2. Permute the biomolecule labels and calculate the ES again. Repeat a large number of times to obtain an empirical distribution of permutation ES, pES
   3. Define the normalized enrichment score (NES) as the ES divided by the absolute mean of the pES that have the same sign
   4. Define the enrichment p-value as the proportion of pES (with the same sign as the ES) that are at least as extreme as the ES
4. Adjust p-values to account for multiple hypothesis testing.

Calculating ES



Enrichment analysis
≠
Over-representation analysis!

Over-representation analysis

1. P-value: Hypergeometric test
2. Sensitive to the approach used to classify biomolecules as “interesting” (e.g., FDR cutoff)
3. Does not consider direction of change
4. Language: Biomolecule sets are **over-represented** in the subset

Enrichment analysis

1. P-value: Permutation-based
2. Uses all biomolecules in the experiment
3. Biomolecules may be sorted according to some directional statistic
4. Language: Biomolecule sets are positively or negatively **enriched**

Note: *ORA > enrichment analysis, in some cases.*
Ex: separating the biomolecules into clusters and then performing ORA on each cluster.

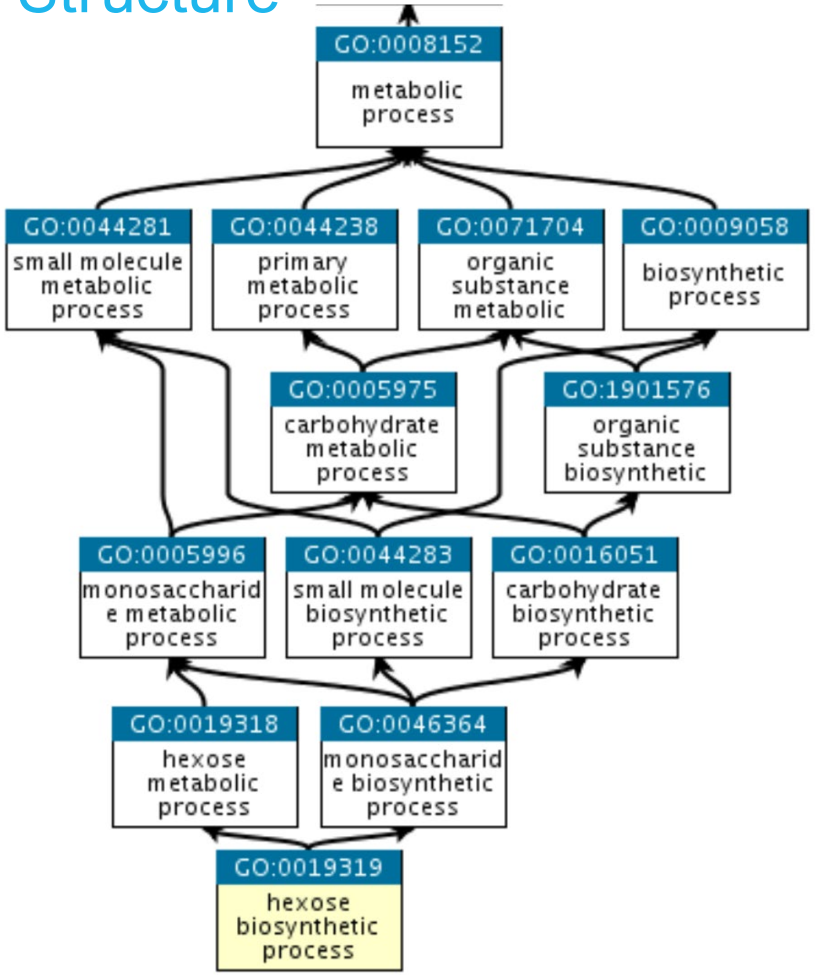
Advantages of Enrichment Analysis

1. Does not rely on arbitrary cutoffs
2. Makes better use of all available data
3. No biological knowledge needed → reduces bias
4. Detects small, concordant changes in related biomolecules
5. ***Generalizable to any biomolecule sets!** *Ex:* substrates grouped by kinases or metabolites grouped by chemical subclasses

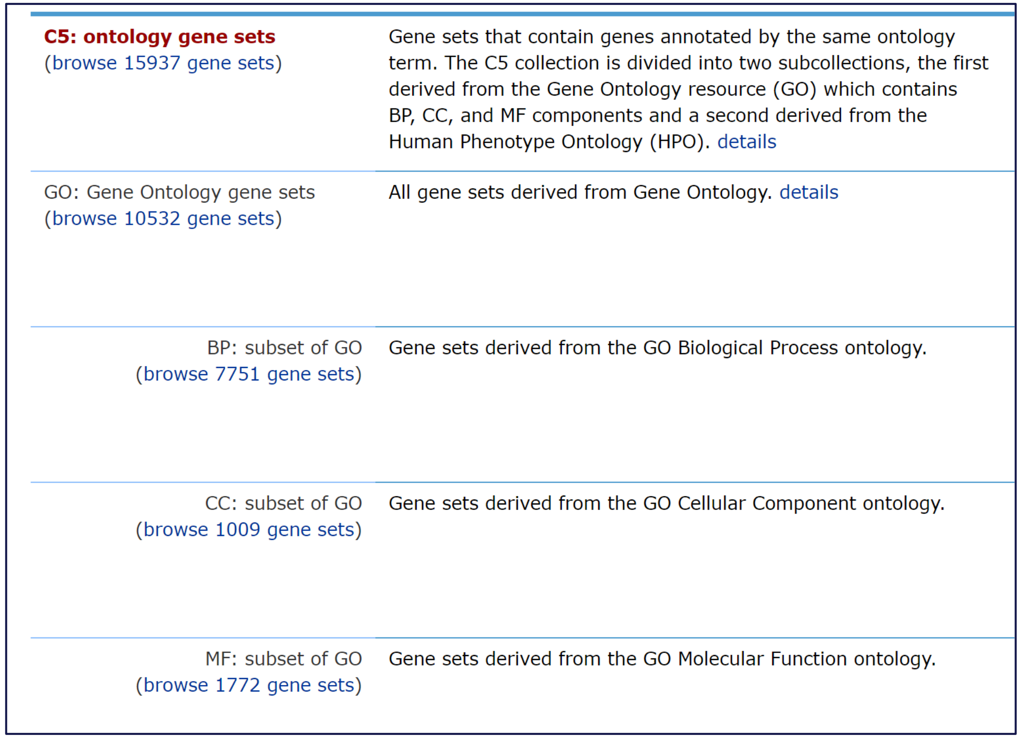
**No reason why the algorithm should be limited to biological data, either!*

Database Examples

Gene Ontology Structure


GO:0008152
metabolic process
GO:0044281
small molecule metabolic process
GO:0044238
primary metabolic process
GO:0071704
organic substance metabolic
GO:0009058
biosynthetic process
GO:0005975
carbohydrate metabolic process
GO:1901576
organic substance biosynthetic
GO:0005996
monosaccharide metabolic process
GO:0044283
small molecule biosynthetic process
GO:0016051
carbohydrate biosynthetic process
GO:0019318
hexose metabolic process
GO:0046364
monosaccharide biosynthetic process
GO:0019319
hexose biosynthetic process

MSigDB - C5 subcollection description:

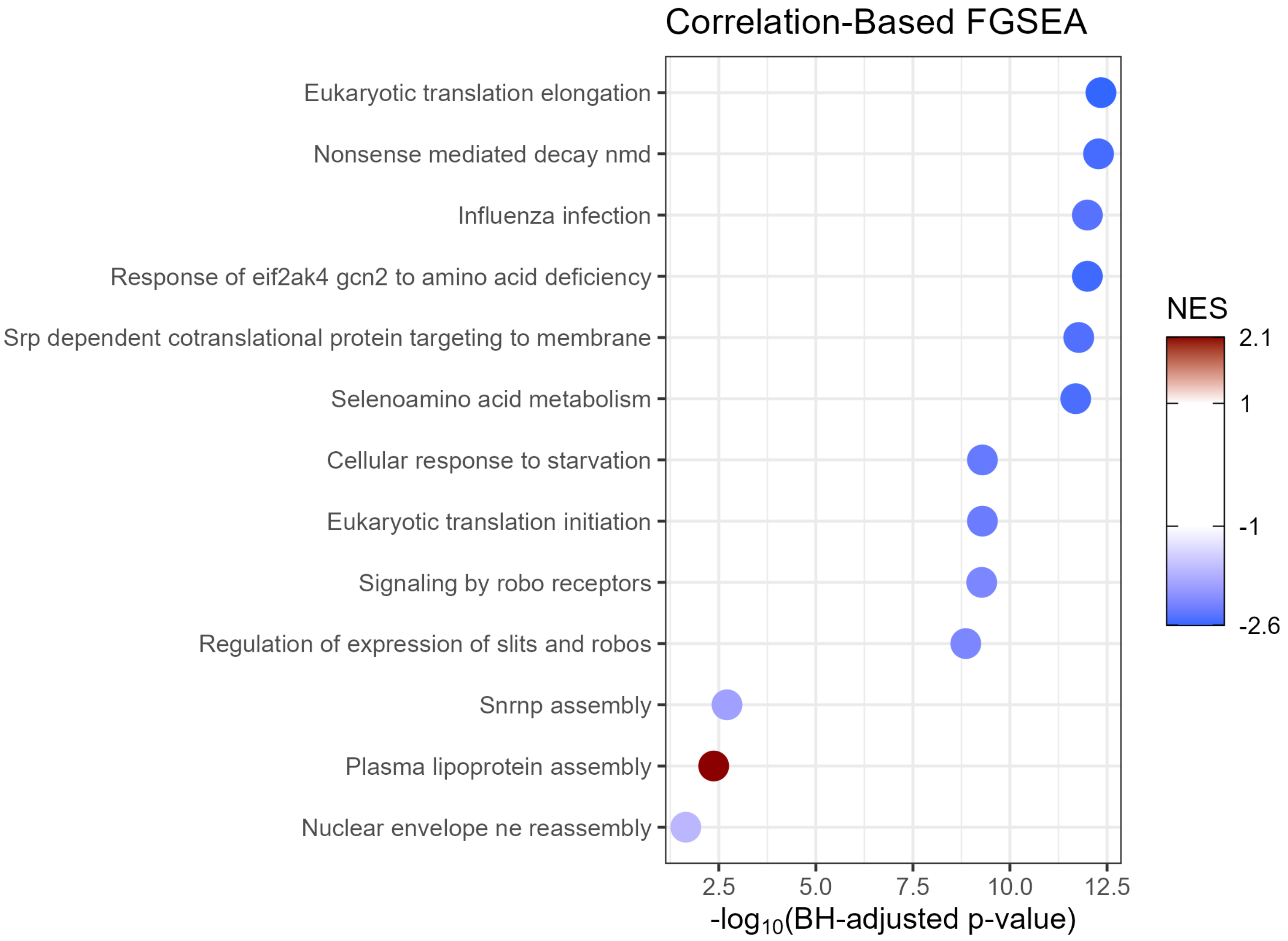


C5: ontology gene sets (browse 15937 gene sets)	Gene sets that contain genes annotated by the same ontology term. The C5 collection is divided into two subcollections, the first derived from the Gene Ontology resource (GO) which contains BP, CC, and MF components and a second derived from the Human Phenotype Ontology (HPO). details
GO: Gene Ontology gene sets (browse 10532 gene sets)	All gene sets derived from Gene Ontology. details
BP: subset of GO (browse 7751 gene sets)	Gene sets derived from the GO Biological Process ontology.
CC: subset of GO (browse 1009 gene sets)	Gene sets derived from the GO Cellular Component ontology.
MF: subset of GO (browse 1772 gene sets)	Gene sets derived from the GO Molecular Function ontology.

Sources:
http://geneontology.org/docs/ontology-documentation/
https://www.gsea-msigdb.org/gsea/msigdb/collections.jsp

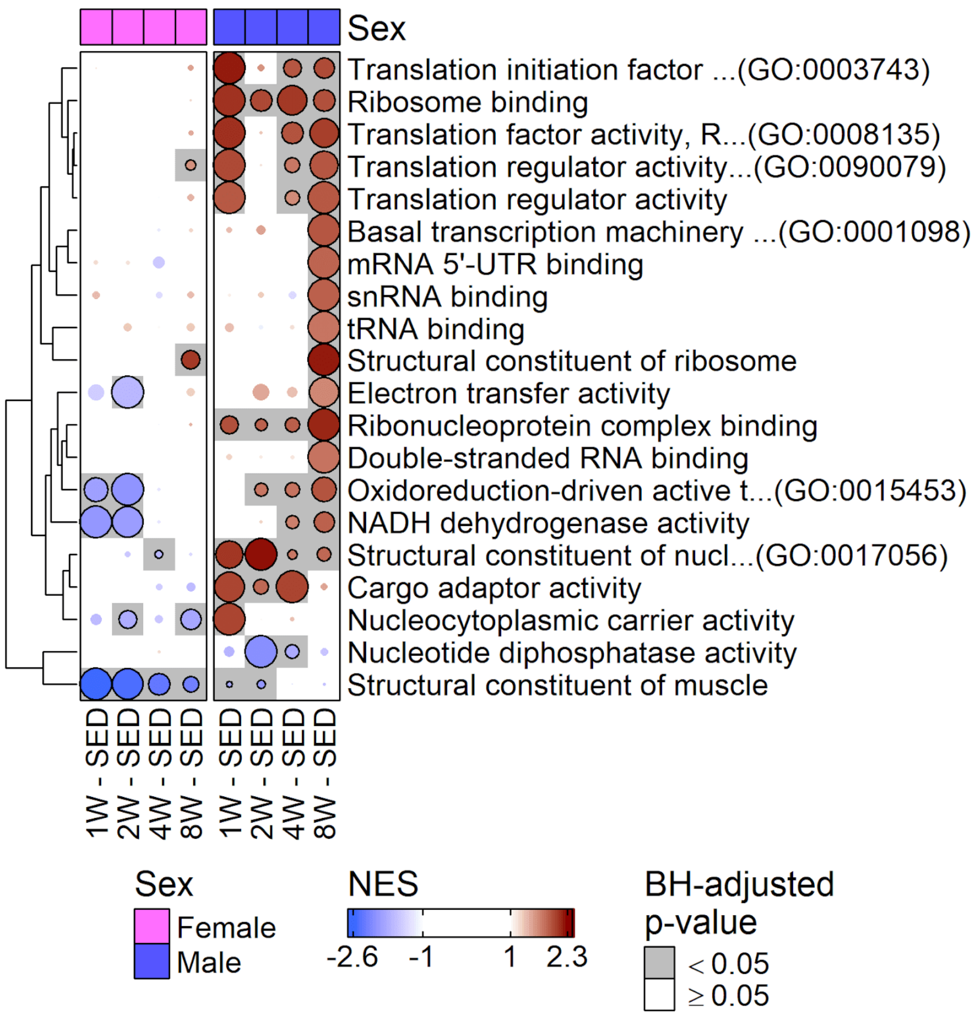
Examples

- *fgsea* + *msigdbr* R packages applied to data from pmartRdata

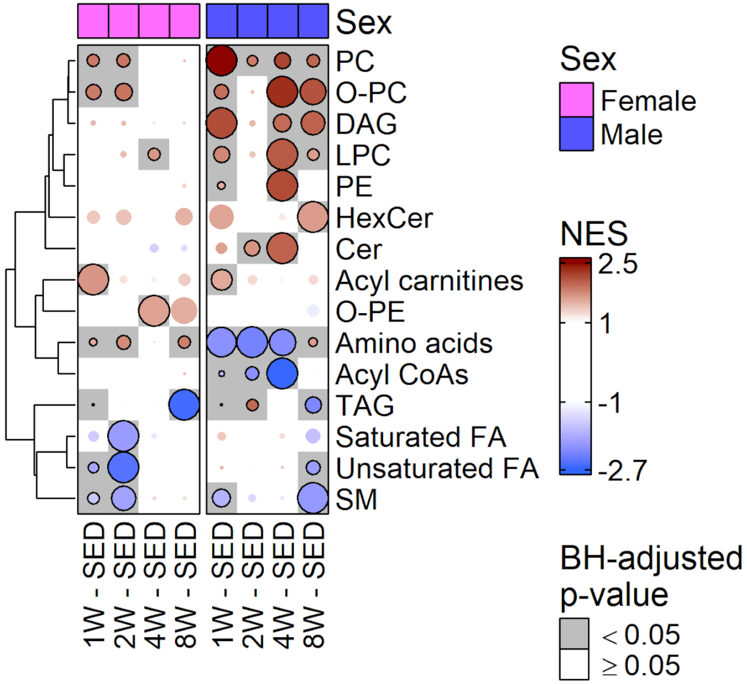


Examples (Advanced)

Proteomics GO-MF



Metabolomics/Lipidomics



Source: *Sexual dimorphism and the multi-omic response to exercise training in rat subcutaneous white adipose tissue (https://doi.org/10.1101/2023.02.03.527012)*

Important Considerations

1. Set redundancy, relevance, size (reliability vs. specificity)
2. Mapping between organisms and/or biomolecule identifiers
3. Choice of ranking metric

Resources / References

- *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles* (https://doi.org/10.1073/pnas.0506580102)
- *Fast gene set enrichment analysis* (FGSEA; https://doi.org/10.1101/060012)
- *Pathway Analysis: State of the Art* (https://doi.org/10.3389/fphys.2015.00383)
- *Functional Analysis for RNA-Seq* (https://hbctraining.github.io/Training-modules/DGE-functional-analysis/lessons/02_functional_analysis.html)—ORA overview
- Molecular Signatures Database (MSigDB; https://www.gsea-msigdb.org/gsea/msigdb/collections.jsp)
- *Sexual dimorphism and the multi-omic response to exercise training in rat subcutaneous white adipose tissue* (https://doi.org/10.1101/2023.02.03.527012)—utilizes FGSEA and ORA with a novel p-value correction method, and extends the FGSEA framework to perform Kinase–Substrate Enrichment Analysis (KSEA) and to summarise the behavior of metabolite subclasses (see Methods)

R packages:

- msigdbr (https://cran.r-project.org/web/packages/msigdbr/index.html)
- fgsea (https://bioconductor.org/packages/release/bioc/html/fgsea.html)—functions for FGSEA and ORA
- MotrpacRatTraining6moWAT (https://pnnl-comp-mass-spec.github.io/MotrpacRatTraining6moWAT/index.html)—enrichmat function, fgsea and msigdbr wrappers

Biological Interpretation

Here, we covered 4 ways to conduct biological interpretation:

1. Visualizing -omics scaled abundances / fold-changes
2. Predictive modeling
3. Investigating relationships
4. Enrichment Analysis and Over-Representation Analysis

This is a growing area of research.

Questions?

Closing Remarks

Afternoon Session

1:15-2:30 p.m.	Multi-Omics Analysis Portal	David Degnan
2:30-2:45	Networking Break	
2:45-4:00	pmartR Statistics and Visualization	Kelly Stratton