

# Introduction to Proteomics

Amanda Khoo

MBP Tech Talks

09 Nov 2019

# Introduction to Proteomics

1. Overview of shotgun proteomics
2. Searching raw data against protein databases
3. Protein grouping
4. Protein quantification
- 5. Tutorial 1:** Data analysis from single-shot label-free DDA data

## Time permitting:

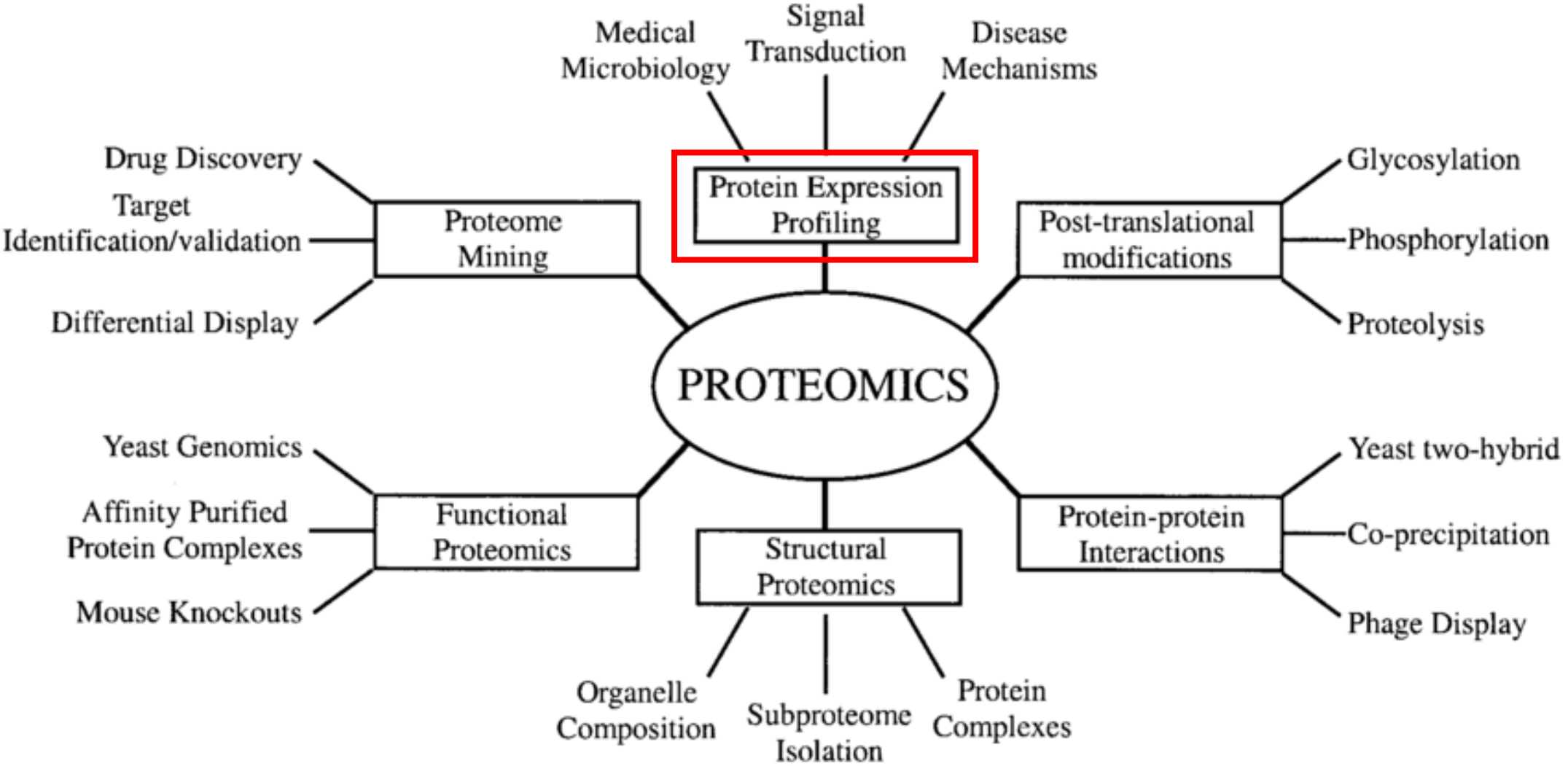
1. Other data types – fractionated, TMT, glycoproteomics, phosphoproteomics
- 2. Tutorial 2:** Data analysis from TMT data
- 3. Tutorial 3:** Data analysis from glycoproteomics data

**Nov 29: Intro to proteogenomics**

# Please download these files for the tutorial

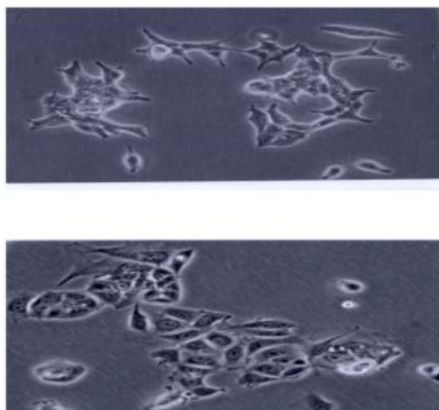
1. Install R and Rstudio
2. Have these packages installed: **ggplot2, reshape2, data.table**
  1. To install packages: `install.packages("ggplot2")`
3. Download these datafiles:
  1. **source\_file.R**
  2. **LFQ** – lfq\_script.R, parameters.txt, proteinGroups.txt, summary.txt, tables.pdf
  3. **TMT** – proteinGroups.txt, summary.txt, tables.pdf, tmt\_script.R
  4. **Glyco** – Asn-\_AspSites.txt, glyco\_script.R, tables.pdf

# Proteomics



# Discovery Proteomics: differential expression profiling by MS

## Biological Samples (case vs. control)



### Protein Mixtures

- Biofluids
- Tissue lysates



- digest to peptides
- fractionate peptides

## LC-MS/MS

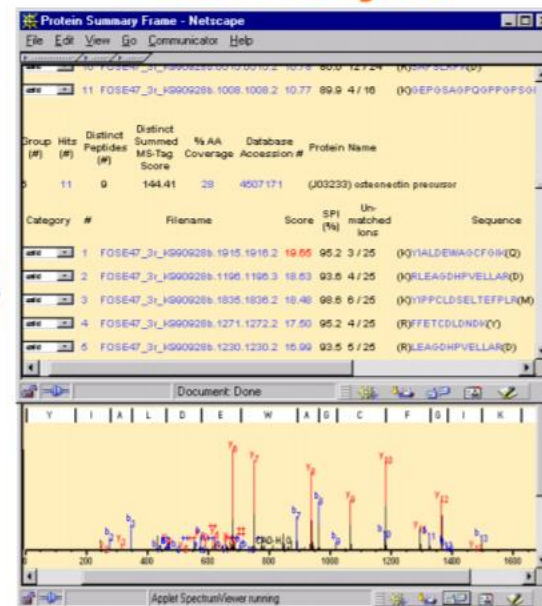


### Separate and Analyze Peptides by LC-MS/MS



- m/z and intensity of peptides
  - rich *pattern*
- Fragment ions for sequence

## Data Analysis

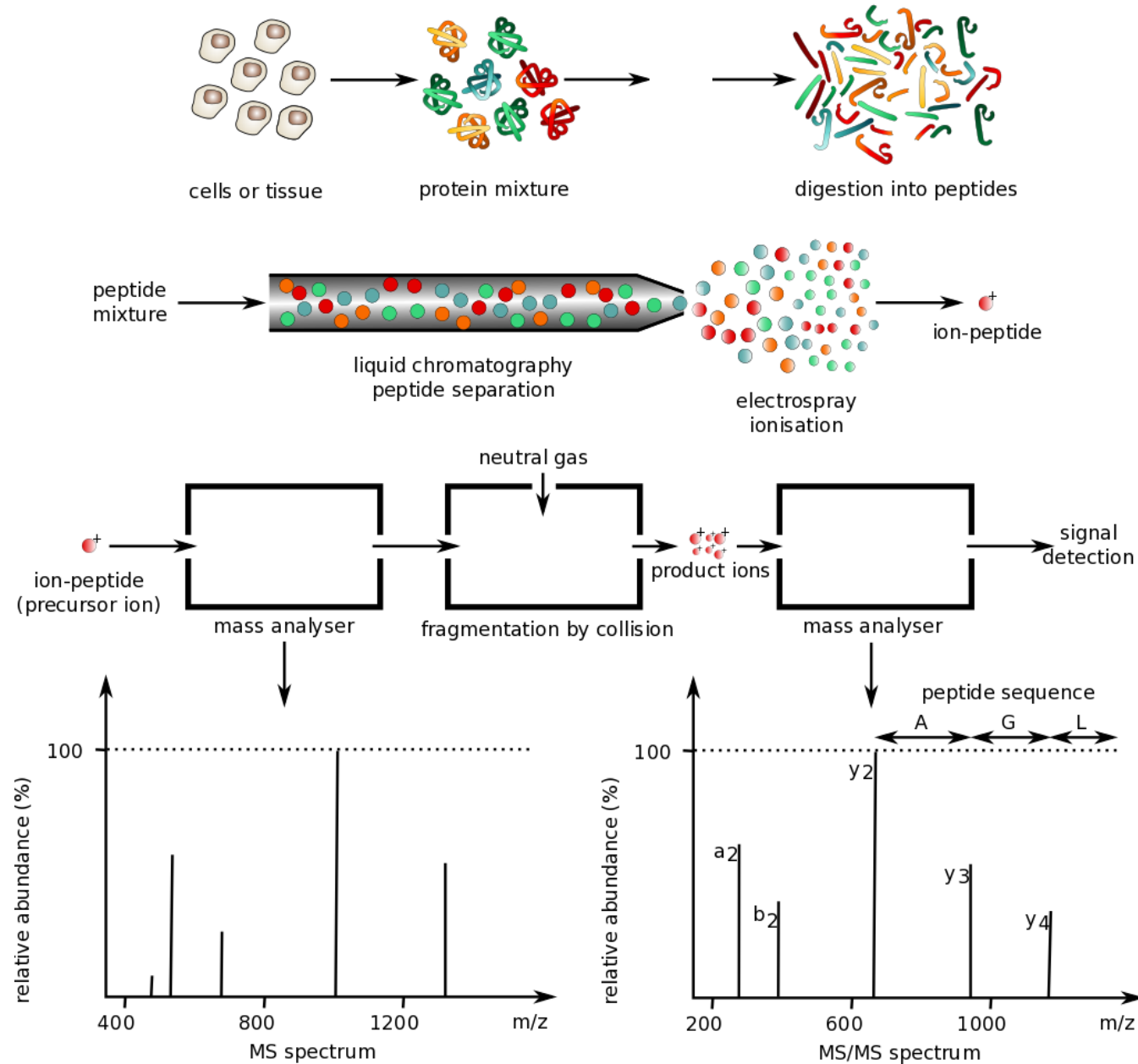


### Search DB using peptide m/z and sequence

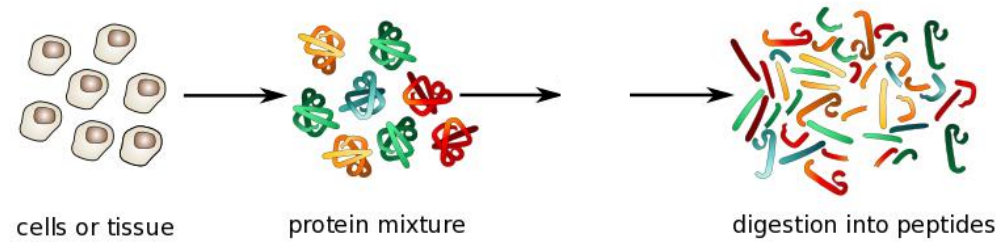


- Peptide **identity**
- Protein **identity**
- Relative abundance

# Sample Preparation



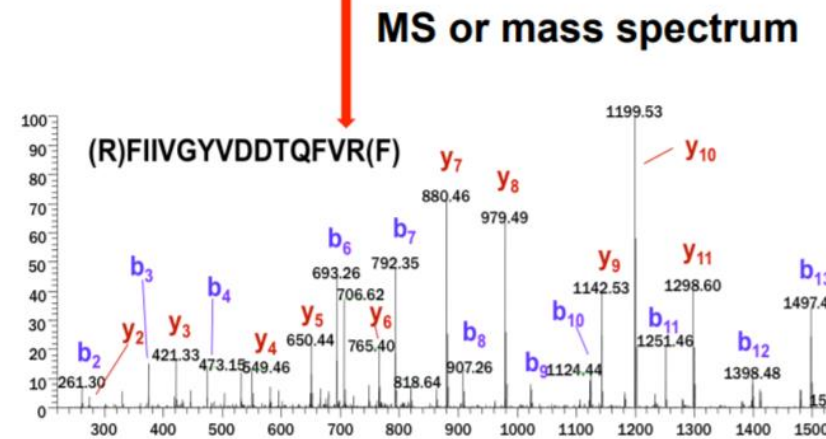
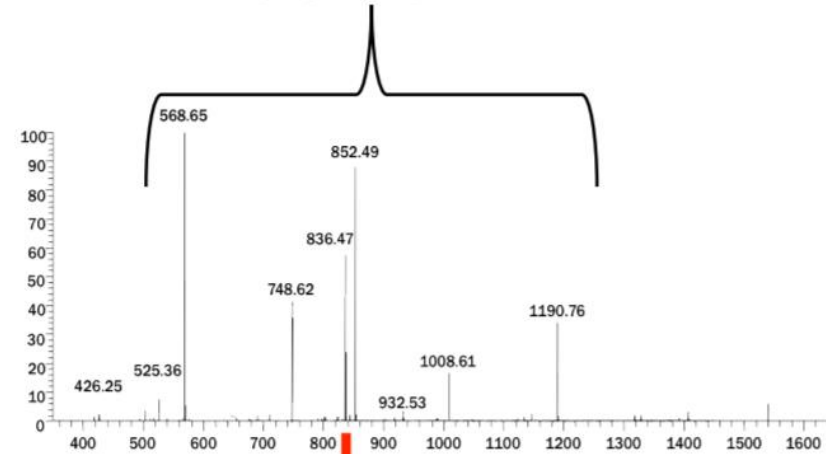
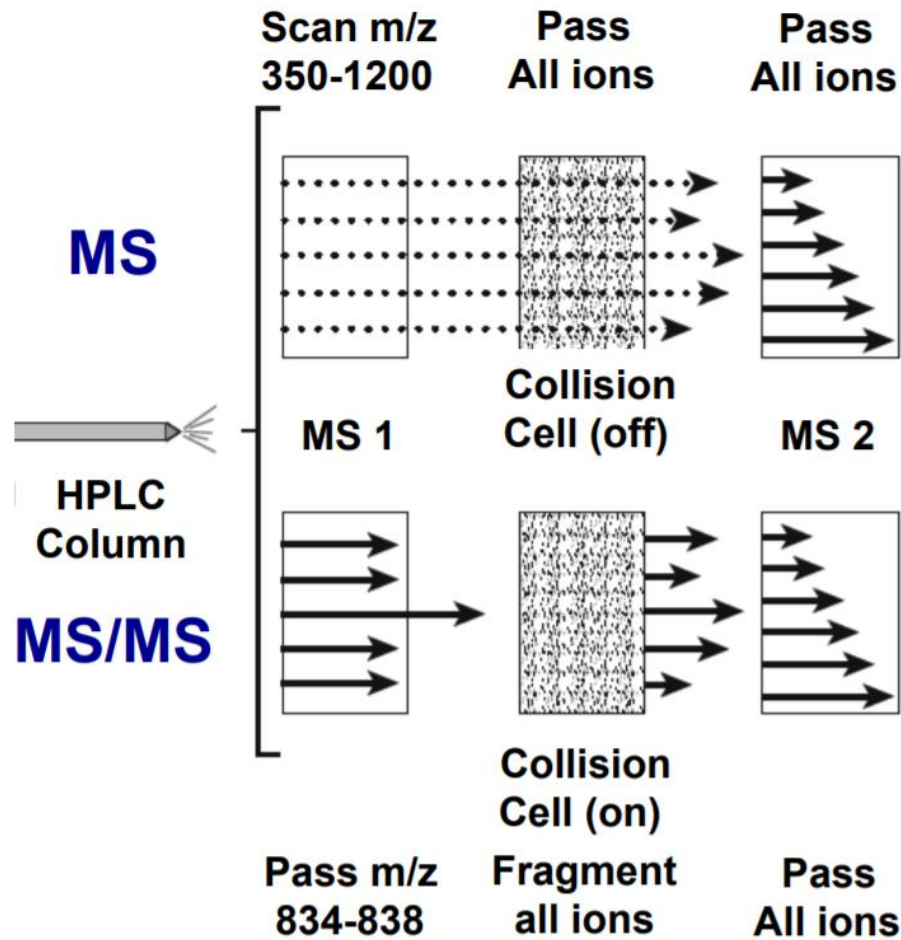
# Sample Preparation



**TABLE 2** | Proteolytic enzymes and digestion conditions that are recommended by the protocol presented here.

Protease	Specificity	Expected missed cleavages	pH	Enzyme/protein (wt/wt)	Temp. (°C)	Hours	Recommendations
<b>C-terminal cleavage</b>							
Chymotrypsin	F, Y, L, W, M	0–4	8	1/75	25	12	Dilute urea concentration to <2 M
→ LysC	K	0–2	8	1/75	37	12	
GluC	E (D) <sup>a</sup>	0–3 (0–4) <sup>b</sup>	8	1/75	25	12	Add 20 mM methylamine when applying urea. Dilute the urea concentration to <2 M
ArgC	R (K) <sup>c</sup>	0–2 (0–3) <sup>b</sup>	8	1/75	37	12	Add 8.5 mM CaCl <sub>2</sub> , 5 mM DTT and 0.5 mM EDTA. Add 20 mM methylamine when applying urea. Dilute urea to <2M
→ Trypsin	R, K	0–2	8	1/75	37	12	Dilute the urea concentration to <2 M
<b>N-terminal cleavage</b>							
AspN	D (E) <sup>d</sup>	0–3 (0–4) <sup>b</sup>	8	1/75	37	12	Add 20 mM methylamine when applying urea. Dilute the urea concentration to <2 M. Do not use metal chelators
LysN	K	0–2	8	1/75	37	12	Dilute the urea concentration to below 6 M. Do not use metal chelators

# How we sequence peptides: MS/MS intact peptide parent ions



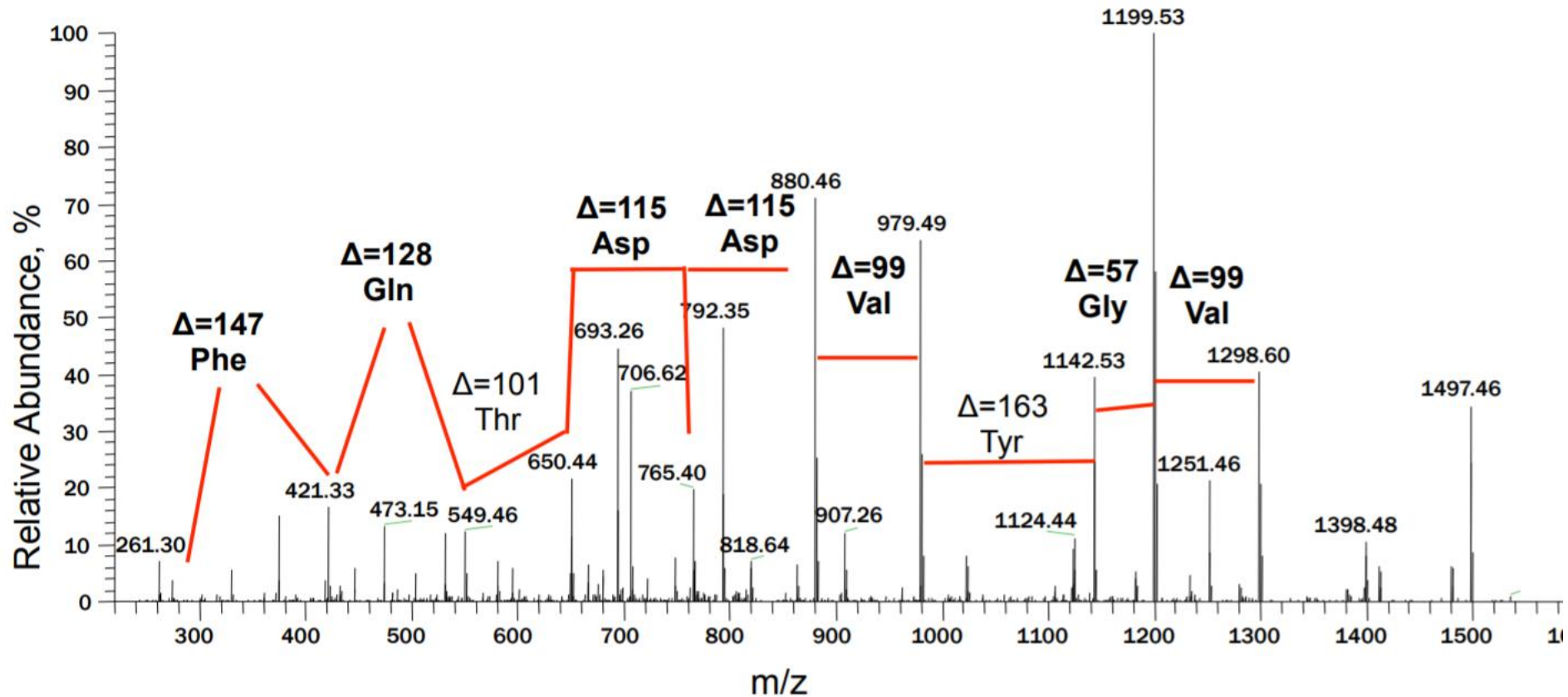
MS/MS means using two mass analyzers (combined in one instrument) to select an analyte (ion) from a mixture, then generate fragments from it to give structural information.



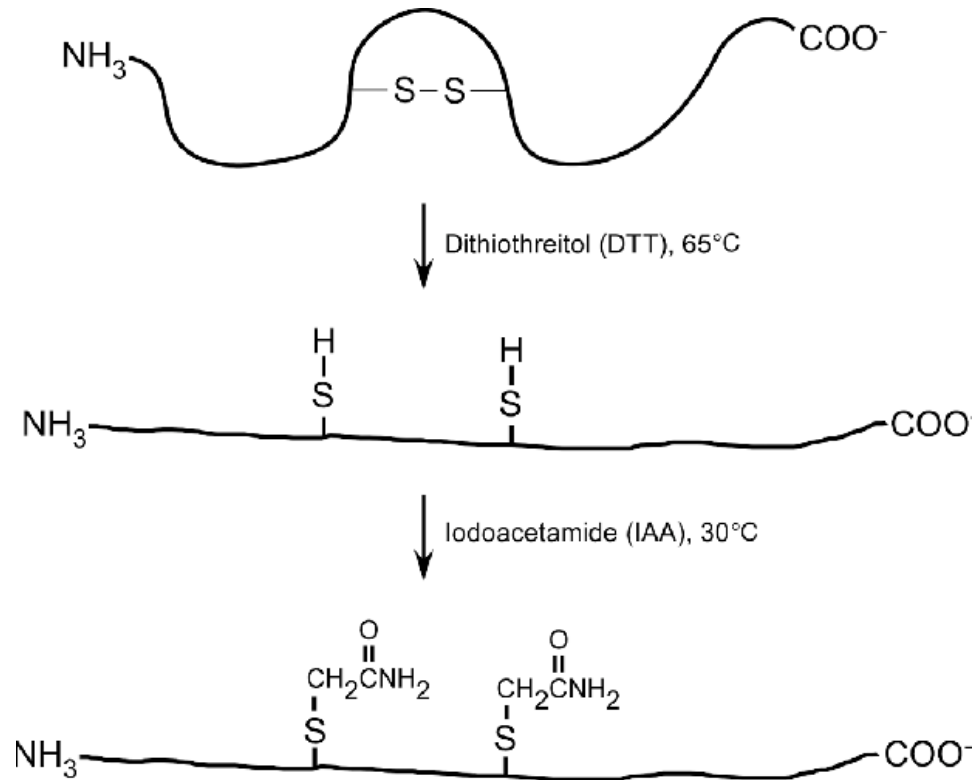
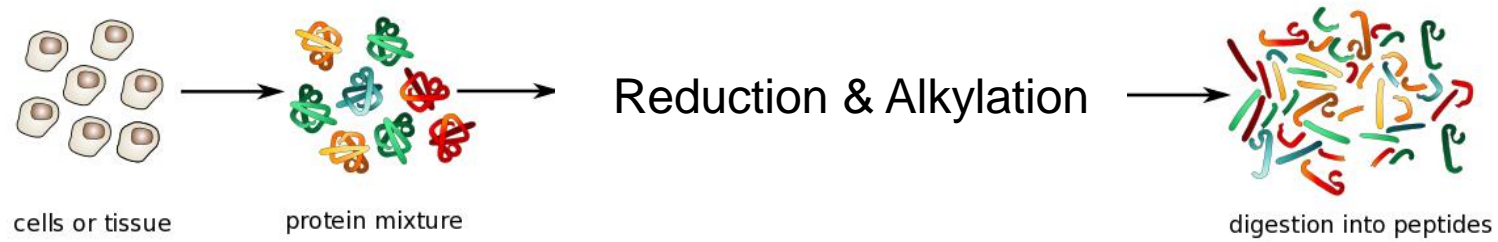
# Example electrospray MS/MS spectrum of a peptide



Filter for 2+, 3+ or 4+ charged peptides



# Sample Preparation



Carbamidomethylation (C)

## Most analyses of proteins are done by digestion of proteins to peptides (“bottom-up” proteomics)

---

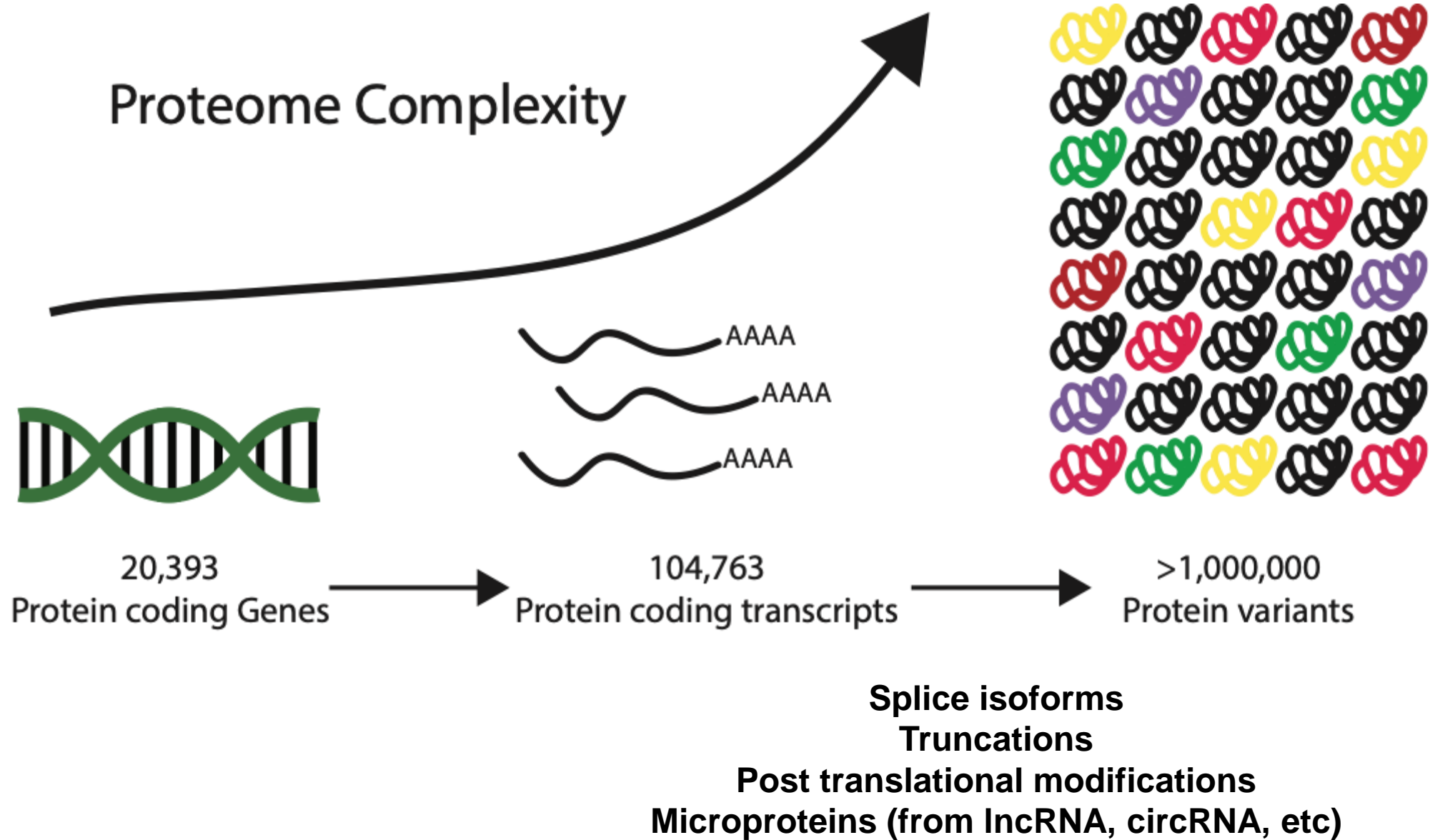
### Advantages:

- Data acquisition easily automated
- Fragmentation of tryptic peptides well understood
- Reliable software available for analysis
- Separation of peptides to create less complex subsets of the proteome for MS analysis is far easier than for proteins (relates to breadth and depth of coverage)

### Disadvantages:

- Simple relationship between peptide and protein lost
- Took highly complex mixture and made it 20-100x more complex
  - Puts high analytical demands on instrumentation

# Proteome is vastly more complex

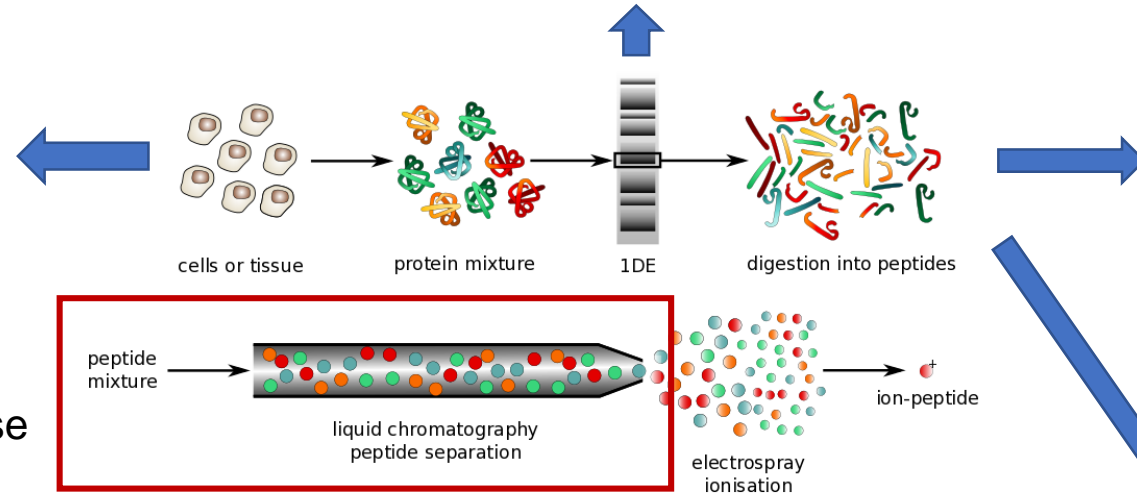


# Fractionation

**Protein separation** (Biophysical properties): SDS-PAGE, IEF, 2D gel

**Organelle fractionation** - nuclear, mitochondrial, cytosolic, etc  
**Fluids** – blood, urine, conditioned media

**Liquid Chromatography (LC)**  
 Hydrophobicity: C18 reversed-phase  
 Isoelectric point: capillary electrophoresis

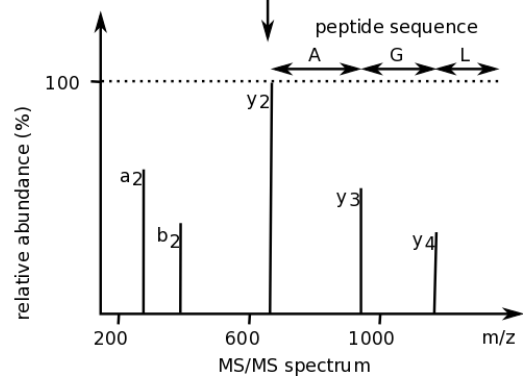
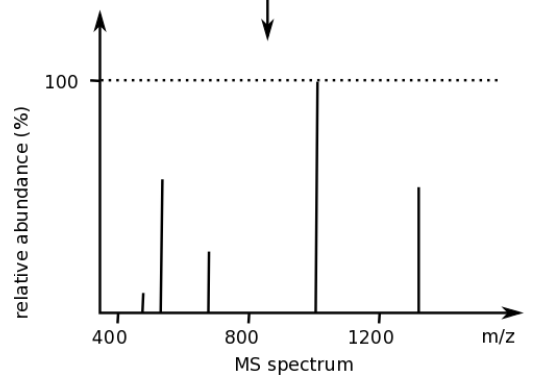
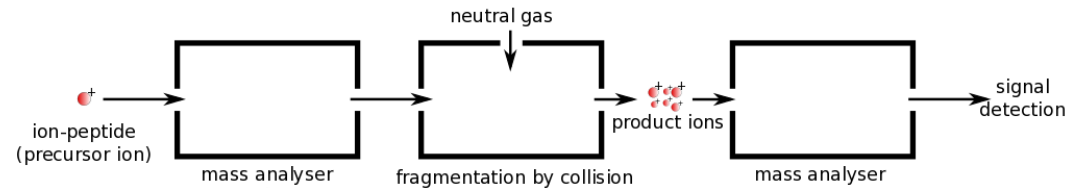


**Peptide fractionation**  
 (Biophysical properties)

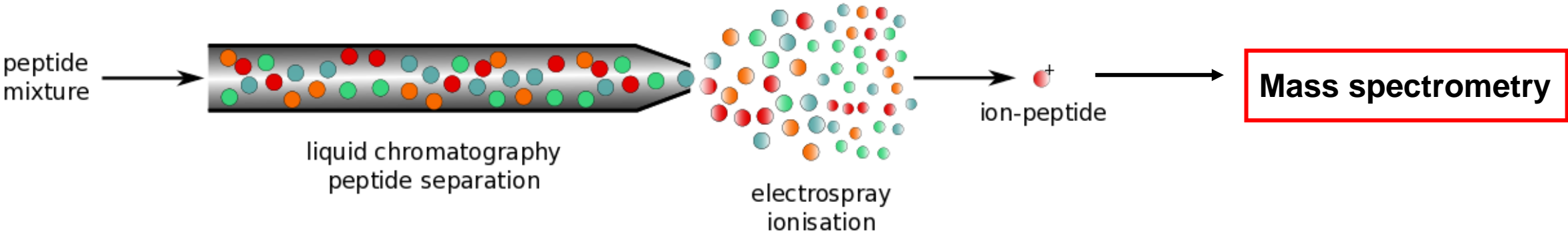
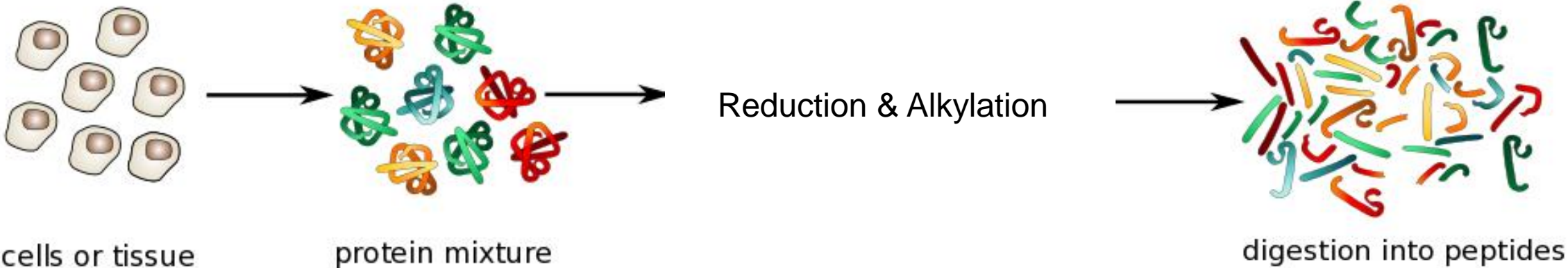
- Acidity/Basicity: SCX, SAX
- pH: high pH reversed-phase

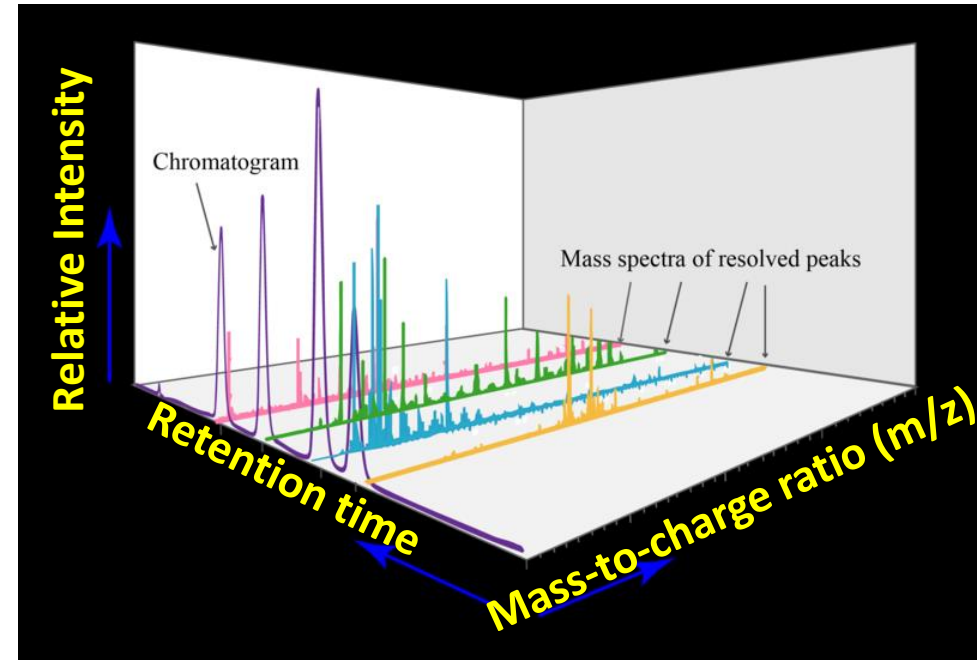
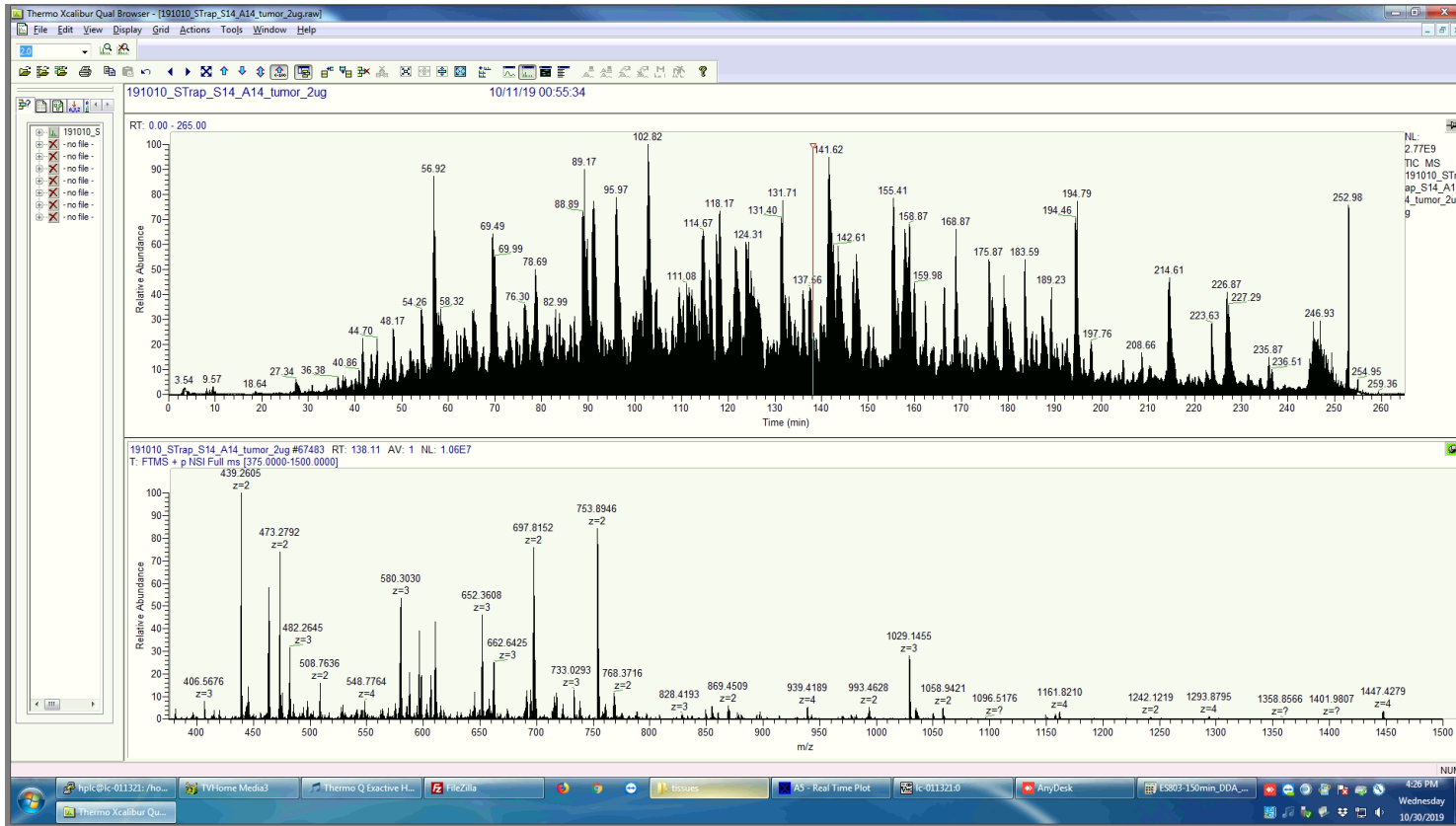
**PTM enrichment**

Cell surface: Glycocapture,  
 Cell-surface capture  
 Phosphoproteomics



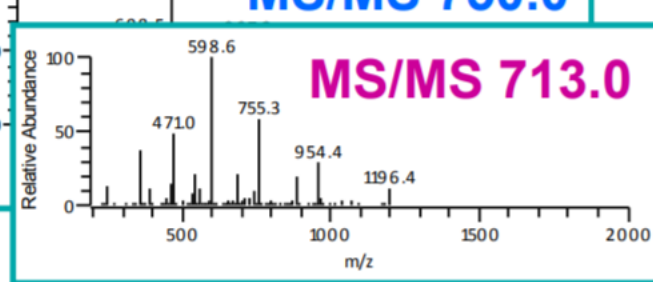
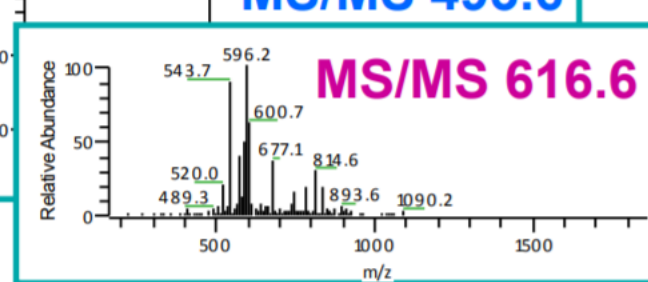
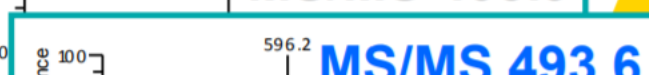
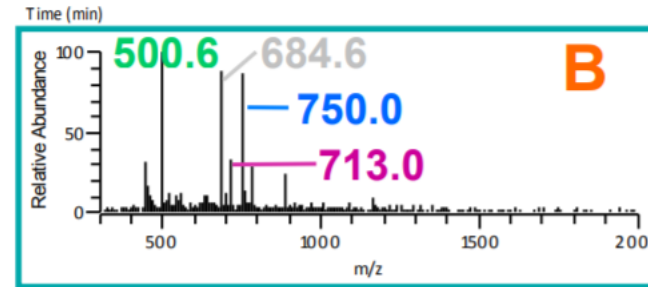
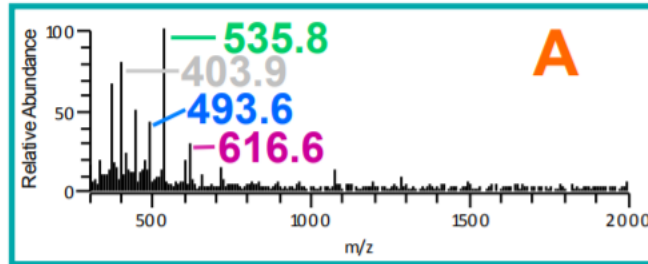
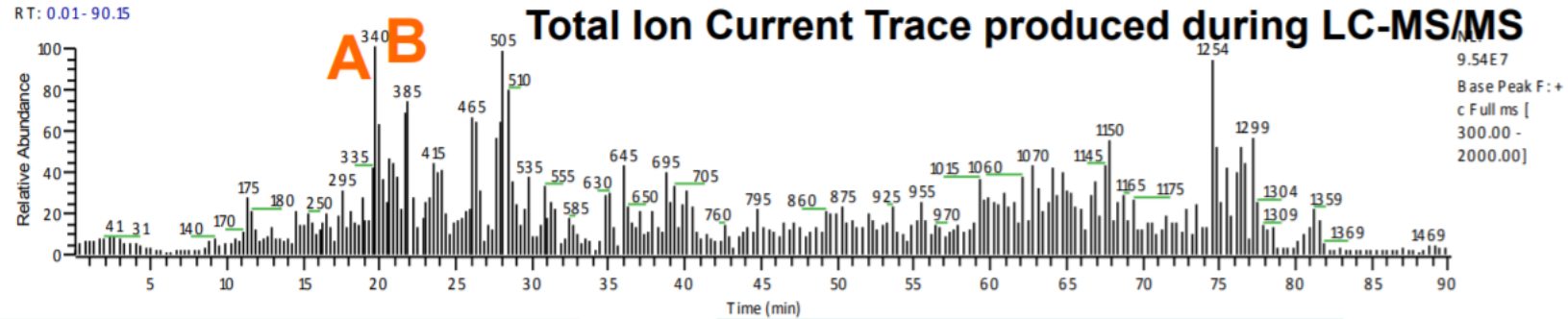
# Single-shot DDA workflow





$m/z = \text{mass of peptide} / \text{charge}$

# Automated Peptide Sequencing by LC/MS/MS (Data Dependent Acquisition)



1-2 sec  
cycle time

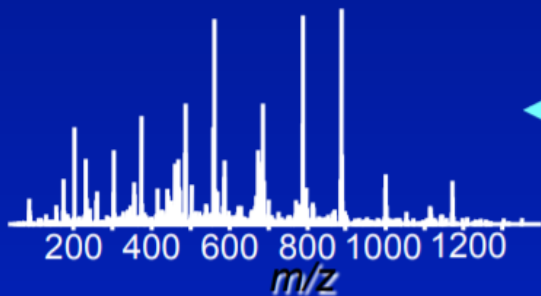
“Top 4 Method” (modern MS systems can do up to “top 20”)



# MS/MS Search Engines: looking up the answer in the back of the book

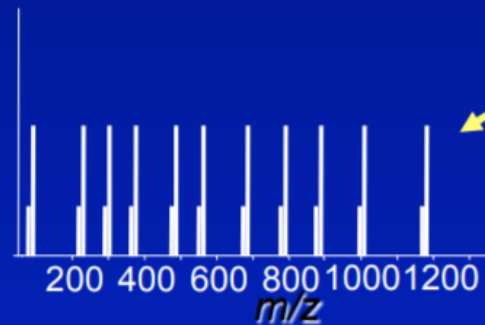
**Peptide Spectrum Match (PSM):** MS2 spectrum that matches to a peptide and passes peptide FDR

Acquired MS/MS spectrum



Sequence Database  
(translation of transcriptome)

Theoretical spectrum



correlate



similarity score

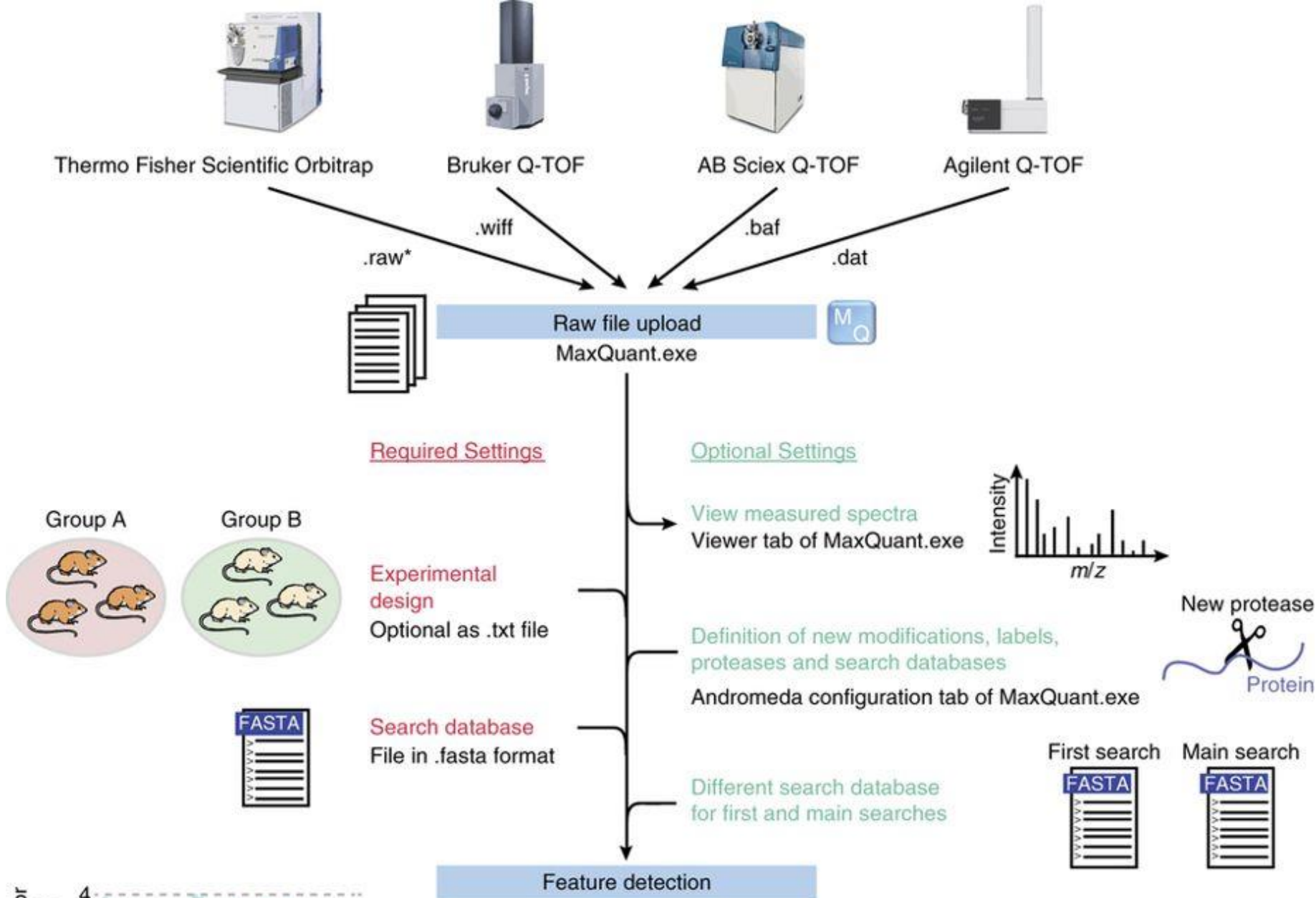
ISLLDAQSAPLR  
VVEELCPTPEGK  
DLLLQWCWENGK  
ECDVVSNTIIAEK  
GDAVFVIDALNR  
**VPTPNVSVVDLTNR**  
SYLFCMENSAEK  
PEQSDLRSWTAK

**Best matching database peptide**

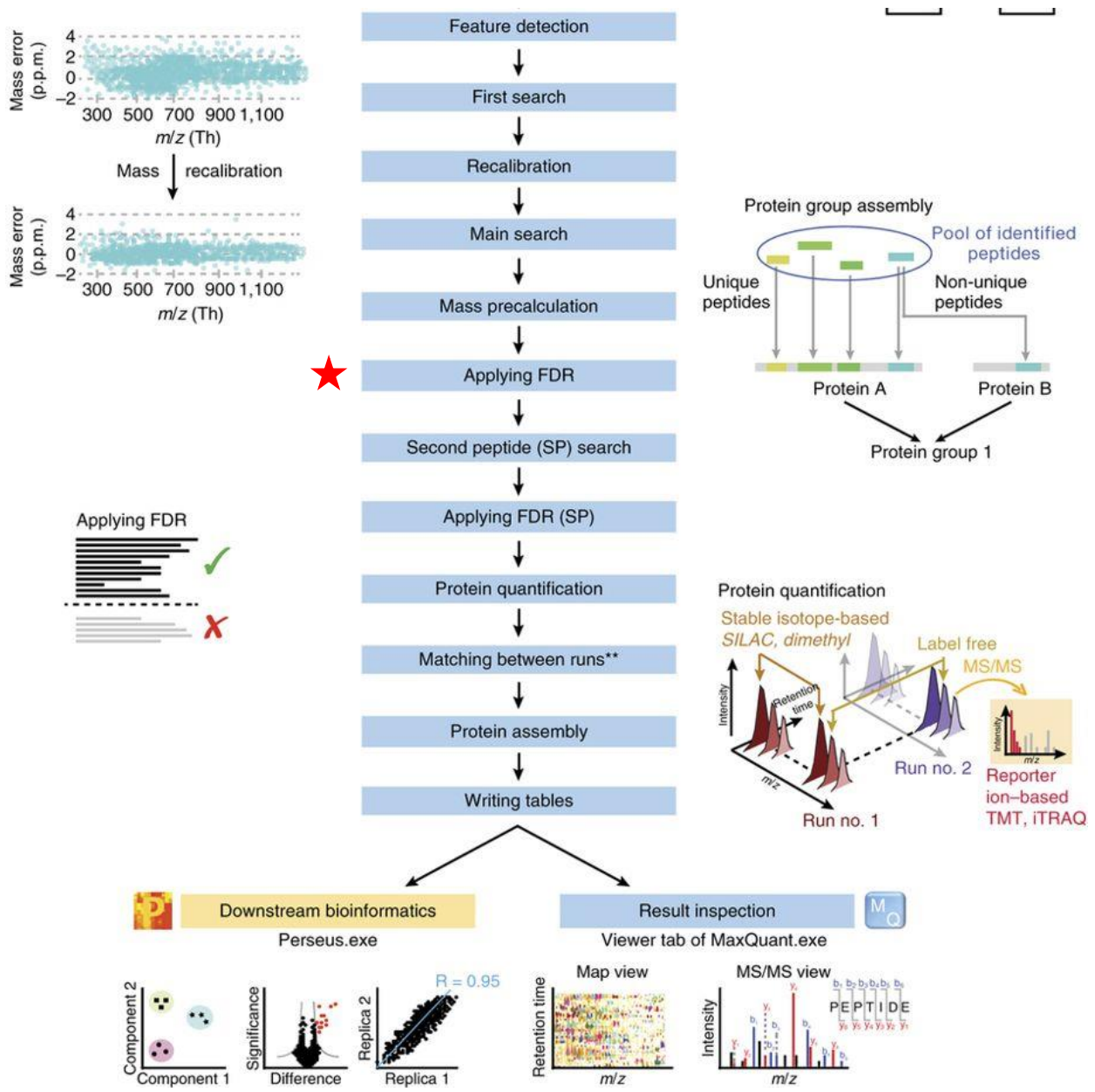
**Determine peptide FDR by searching reversed DB**

Algorithms: Mascot, MaxQuant, SpectrumMill, X-Tandem...

# MaxQuant



# MaxQuant



Documentation wiki:  
<http://www.coxdocs.org/doku.php?id=:maxquant:start>  
 Tutorial:  
<https://www.nature.com/articles/nprot.2016.136>

# Search parameters

The screenshot displays the MaxQuant software interface. At the top, the window title is "Session1 - MaxQuant". Below the title bar is a menu bar with "File", "Tools", "Window", and "Help". A secondary menu bar contains "Raw files", "Group-specific parameters", "Global parameters", "Performance", "Viewer", and "Configuration".

The main interface is divided into several sections:

- Input data:** Includes buttons for "Load", "Remove", "Load folder", and "Change folder".
- Exp. design file:** Includes a "Write template" button and a "Read from file" button.
- Edit exp. design:** Includes buttons for "Set experiment", "Set parameter group", "Set fractions", and "No fractions".

The central part of the interface is a table listing files. The table has the following columns: File, Exists, Size, Data format, Parameter group, Experiment, and Fraction. The data is as follows:

	File	Exists	Size	Data format	Parameter group	Experiment	Fraction
1	D:\FTPData\Ketlin\191009_ketlin_72.raw	True	2.5 GB	Thermo raw...	Group 0	k_72	1
2	D:\FTPData\Ketlin\191009_ketlin_93.raw	True	2.6 GB	Thermo raw...	Group 0	k_93	1
3	D:\FTPData\Ketlin\191009_ketlin_97.raw	True	2.6 GB	Thermo raw...	Group 0	k_97	1
4	D:\FTPData\Ketlin\191009_ketlin_113.raw	True	2.4 GB	Thermo raw...	Group 0	k_113	1
5	D:\FTPData\Ketlin\191030_ketlin_70_191002211216.raw	True	2.4 GB	Thermo raw...	Group 0	k_70	1
6	D:\FTPData\Ketlin\191030_ketlin_152_191003023234.raw	True	2.4 GB	Thermo raw...	Group 0	k_152	1
7	D:\FTPData\Ketlin\190919_ketlin_75kg_redo.raw	True	3.2 GB	Thermo raw...	Group 0	k_75	1

At the bottom of the interface, there is a status bar showing "7 items 1 selected" and a zoom level of "100%". Below this is a control panel with a "Number of threads" dropdown set to "7", and buttons for "Start", "Stop", "Partial processing", and "Details". A checkbox for "Send email when done" is also present. The version number "Version 1.5.8.3" is displayed in the bottom right corner.

# Search parameters

The screenshot displays the MaxQuant software interface for configuring search parameters. The window title is "Session1 - MaxQuant". The menu bar includes "File", "Tools", "Window", and "Help". The main navigation tabs are "Raw files", "Group-specific parameters", "Global parameters", "Performance", "Viewer", and "Configuration". The "Group-specific parameters" tab is active, showing "Group 0" selected. The "Type" sub-tab is "Modifications", with other options being "Instrument" and "First search". Under "Modifications", there are sub-sections for "Digestion", "Label-free quantification", and "Misc.". The "Variable modifications" section is expanded, showing a list of modifications on the left and a list of selected modifications on the right. The left list includes: Acetyl (K), Acetyl (N-term), Acetyl (Protein N-term), Amidated (C-term), Amidated (Protein C-term), Carbamidomethyl (C), Carbamyl (N-term), Cation:Na (DE), Cys-Cys, Deamidation (N), Deamidation (NQ), and Deamidation 18O (N). The right list includes: Oxidation (M) and Acetyl (Protein N-term). Below the lists, the "Max. number of modifications per peptide" is set to 5. At the bottom of the interface, there are controls for "Number of threads" (set to 7), "Start", "Stop", "Partial processing", "Send email when done" (checkbox), and "Details" buttons. The version number "Version 1.5.8.3" is displayed in the bottom right corner.

Session1 - MaxQuant

File Tools Window Help

Raw files Group-specific parameters Global parameters Performance Viewer Configuration

Group 0 Type Modifications Instrument First search

Digestion Label-free quantification Misc.

Parameter group Parameter section

Variable modifications

Acetyl (K)  
Acetyl (N-term)  
Acetyl (Protein N-term)  
Amidated (C-term)  
Amidated (Protein C-term)  
Carbamidomethyl (C)  
Carbamyl (N-term)  
Cation:Na (DE)  
Cys-Cys  
Deamidation (N)  
Deamidation (NQ)  
Deamidation 18O (N)

Oxidation (M)  
Acetyl (Protein N-term)

Max. number of modifications per peptide 5

Number of threads 7 Start Stop Partial processing Send email when done Details

Version 1.5.8.3

**Variable modifications: Includes modified and unmodified peptide in database search**

# Search parameters

Session1 - MaxQuant

File Tools Window Help

Raw files Group-specific parameters Global parameters Performance Viewer Configuration

Group 0 Type Modifications Instrument First search

Digestion Label-free quantification Misc.

Parameter group Parameter section

Digestion mode Specific

Enzyme

- ArgC
- AspC
- AspN
- Chymotrypsin
- Chymotrypsin+
- CnBR
- D.P
- GluC
- GluN
- LysC
- LysC/P
- LysN

Trypsin/P

Max. missed 2

Number of threads 7

Start Stop Partial processing Details

Send email when done

Version 1.5.8.3

**Enzymes: Trypsin/P (C-ter cleavage at K, R, even if followed by P)**

**Maximum 2 missed cleavages**

# Search parameters

Session1 - MaxQuant

File Tools Window Help

Raw files Group-specific parameters Global parameters Performance Viewer Configuration

Group 0

Type Modifications Instrument First search

Digestion Label-free quantification Misc.

Parameter group Parameter section

Digestion mode

Specific

Enzyme

ArgC  
AspC  
AspN  
Chymotrypsin  
Chymotrypsin+  
CnBR  
D.P  
GluC  
GluN  
LysC  
LysC/P  
LysN

> Trypsin/P

Max. missed

2

Number of threads: 7

Start Stop Partial processing

Send email when done:  Details

Version 1.5.8.3

**Enzymes: Trypsin/P (C-ter cleavage at K, R, even if followed by P)**

**Maximum 2 missed cleavages**

# Specify database

The screenshot shows the MaxQuant software interface with the 'Global parameters' tab selected. The 'Fasta files' section contains a text box with the path 'C:\Users\Kislinger Lab\Documents\Fasta files\Rattus\_norvegicus\_ensemble\_Lydia\_sept2019\_SUC2\_copy.fasta'. The 'Include contaminants' checkbox is checked. The 'Fixed modifications' list includes 'Acetyl (N-term)', 'Acetyl (Protein N-term)', 'Amidated (C-term)', 'Amidated (Protein C-term)', 'Carbamidomethyl (C)', 'Carbamyl (N-term)', 'Cation:Na (DE)', 'Cys-Cys', 'Deamidation (N)', 'Deamidation (NQ)', and 'Deamidation 18O (N)'. The 'Min. peptide length' is set to 7, 'Max. peptide mass [Da]' is 4600, 'Min. peptide length for unspecific search' is 8, and 'Max. peptide length for unspecific search' is 25. The 'Number of threads' is set to 7. At the bottom, there are buttons for 'Start', 'Stop', 'Partial processing', and 'Details', along with a 'Send email when done' checkbox.

**1. Specify species-specific database**  
If protein sequence is not in the database, you won't see it in your data!

**2. Include common contaminants database**

**3. Specify fixed modification**  
Will only match alkylated peptides

**4. Specify peptide length of 7-22 amino acids**

Version 1.5.8.3



# Target-Decoy search strategy

## Uniprot reference sequence (human)

P02768-1

KWVTFISLLFLFSSAYS**RGVFR**DAHKSEVAH

**RFK**DLGEENFKALVLIAFAQYLQQCPFEDHV**K**

↓ *in silico* digestion

(K)WVTFISLLFLFSSAYS**R**

(R)DLGEEN**FK**

(K)ALVLIAFAQYLQQCPFEDHV**K**

(K)SEVAH**RFK**

## Decoy database

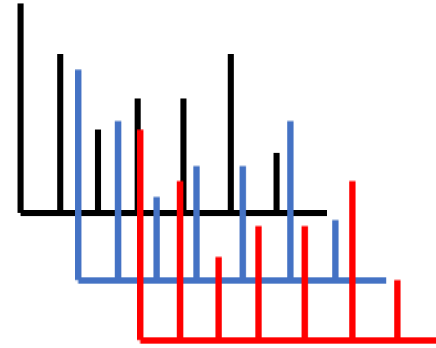
(K)SYASSFLFLLSIFTV**WR**

(R)F**NEEGLDK**

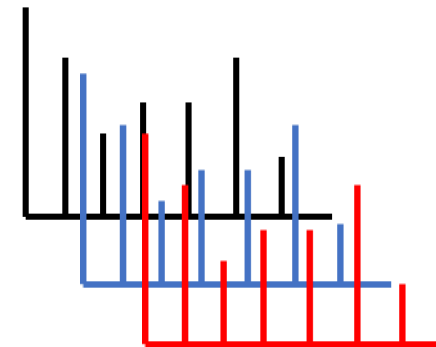
(K)VHDEFPCQQLYQAFAILVL**AK**

(K)F**RHAVESK**

## List of theoretical spectra



## List of experimental spectra



## Search engine

- Andromeda
- OMSSA
- Comet
- X!Tandem
- Mascot
- MSFragger
- MSGF+

Scoring  
1% FDR

**Protein Grouping**

# Set FDR = 1%

The screenshot shows the MaxQuant software interface. The title bar reads "Session1 - MaxQuant". The menu bar includes "File", "Tools", "Window", and "Help". Below the menu bar are several tabs: "Raw files", "Group-specific parameters", "Global parameters" (which is selected), "Performance", "Viewer", and "Configuration". Under the "Global parameters" tab, there are sub-tabs for "Sequences", "Adv. identification", "Label free quantification", "MS/MS - FTMS", "MS/MS - TOF", and "Advanced". The "Identification" sub-tab is active, showing a "Parameter section" with the following settings:

Parameter	Value
PSM FDR	0.01
Protein FDR	0.01
Site decoy fraction	0.01
Min. peptides	1
Min. razor + unique peptides	1
Min. unique peptides	0
Min. score for unmodified peptides	0
Min. score for modified peptides	40
Min. delta score for unmodified peptides	0
Min. delta score for modified peptides	6
Main search max. combinations	200
Base FDR calculations on delta score	<input type="checkbox"/>
Razor protein FDR	<input checked="" type="checkbox"/>

A red bracket highlights the PSM FDR, Protein FDR, and Site decoy fraction values, all of which are set to 0.01. At the bottom of the window, there is a "Number of threads" dropdown set to 7, and buttons for "Start", "Stop", "Partial processing", and "Details". A checkbox for "Send email when done" is also present. The version number "Version 1.5.8.3" is displayed in the bottom right corner.

# Protein Quantitation

Session1 - MaxQuant

File Tools Window Help

Raw files Group-specific parameters Global parameters Performance Viewer Configuration

Group 0 Type Modifications Instrument First search

Digestion Label-free quantification Misc.

Parameter group Parameter section

Label-free quantification

LFQ

LFQ min. ratio count 2

Fast LFQ

LFQ min. number of neighbors 3

LFQ average number of neighbors 6

Skip normalization

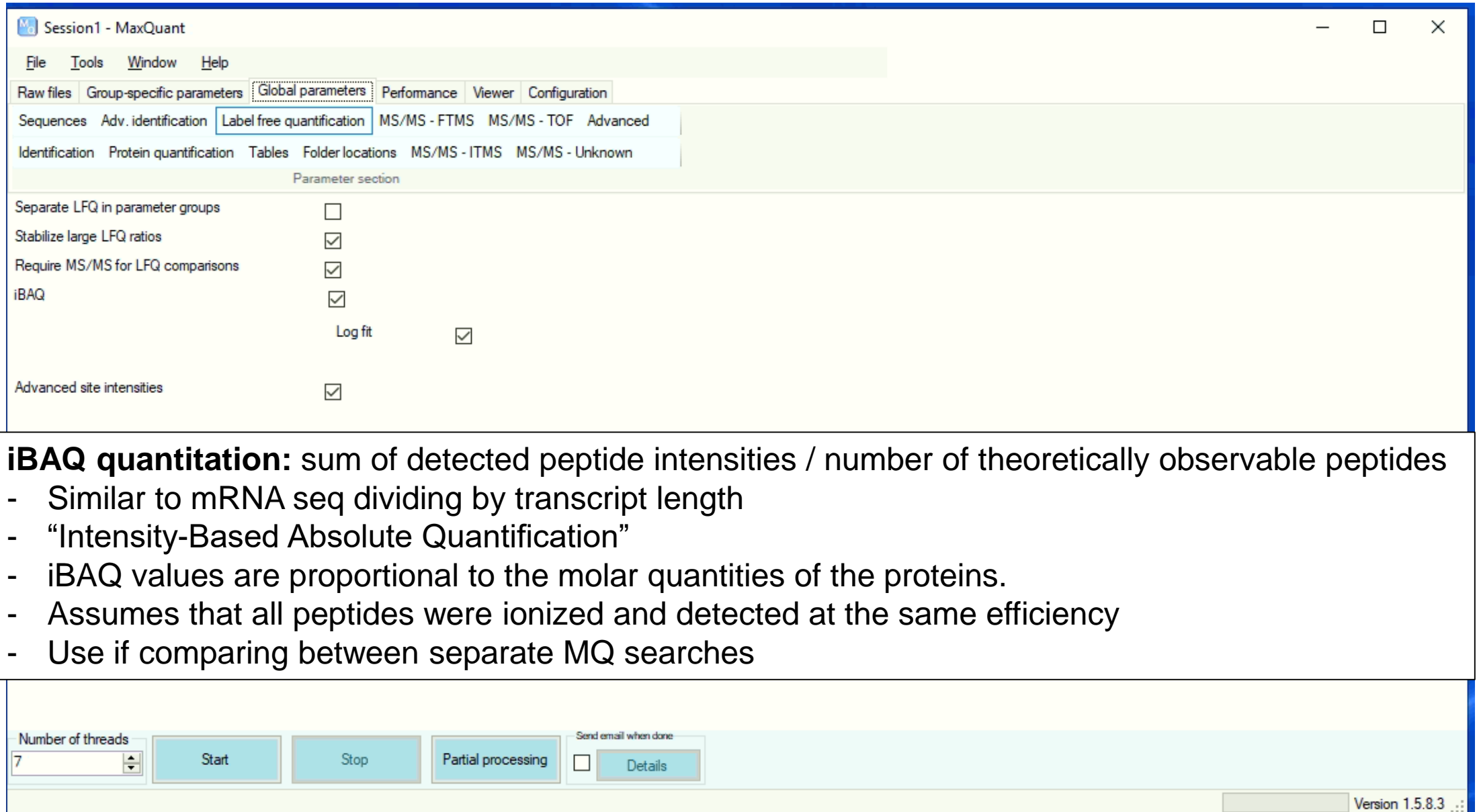
**Label-free quantitation (LFQ):** Applies normalization to raw intensities to exclude some “outliers”

- Actual normalization algorithm unknown but seems to work best compared to other normalization strategies e.g. median normalization of raw intensities
- Use this number for quantitation if comparing samples in the same search

Number of threads 7 Start Stop Partial processing  Send email when done Details

Version 1.5.8.3

# Protein Quantitation



The screenshot shows the MaxQuant software interface. The title bar reads 'Session1 - MaxQuant'. The menu bar includes 'File', 'Tools', 'Window', and 'Help'. The main menu is divided into several sections: 'Raw files', 'Group-specific parameters', 'Global parameters' (which is currently selected), 'Performance', 'Viewer', and 'Configuration'. Under 'Global parameters', there are sub-sections for 'Sequences', 'Adv. identification', 'Label free quantification' (highlighted), 'MS/MS - FTMS', 'MS/MS - TOF', and 'Advanced'. Below these are 'Identification', 'Protein quantification', 'Tables', 'Folder locations', 'MS/MS - ITMS', and 'MS/MS - Unknown'. The 'Parameter section' for 'Label free quantification' contains the following options:

- Separate LFQ in parameter groups:
- Stabilize large LFQ ratios:
- Require MS/MS for LFQ comparisons:
- iBAQ:
- Log fit:
- Advanced site intensities:

At the bottom of the interface, there is a control panel with a 'Number of threads' dropdown set to 7, and buttons for 'Start', 'Stop', 'Partial processing', and 'Details'. A checkbox for 'Send email when done' is also present. The version number 'Version 1.5.8.3' is displayed in the bottom right corner.

**iBAQ quantitation:** sum of detected peptide intensities / number of theoretically observable peptides

- Similar to mRNA seq dividing by transcript length
- “Intensity-Based Absolute Quantification”
- iBAQ values are proportional to the molar quantities of the proteins.
- Assumes that all peptides were ionized and detected at the same efficiency
- Use if comparing between separate MQ searches

# MaxQuant outputs

Name	Date modified	Type	Size
R_figures	2019-03-20 3:32 PM	File folder	
R_tableOutput	2019-02-28 4:57 PM	File folder	
aifMsms.txt	2019-01-17 12:12 PM	TXT File	0 KB
allPeptides.txt	2019-01-17 1:19 PM	TXT File	5,679,155 KB
evidence.txt	2019-01-17 12:20 PM	TXT File	1,255,960 KB
experimentalDesignTemplate.txt	2019-01-16 6:23 PM	TXT File	3 KB
libraryMatch.txt	2019-01-17 12:12 PM	TXT File	0 KB
matchedFeatures.txt	2019-01-17 12:13 PM	TXT File	0 KB
modificationSpecificPeptides.txt	2019-01-17 12:19 PM	TXT File	120,475 KB
ms3Scans.txt	2019-01-17 1:03 PM	TXT File	0 KB
msms.txt	2019-01-17 12:24 PM	TXT File	3,795,159 KB
msmsScans.txt	2019-01-17 1:15 PM	TXT File	1,459,984 KB
msScans.txt	2019-01-17 1:15 PM	TXT File	412,148 KB
mzRange.txt	2019-01-17 1:19 PM	TXT File	49,831 KB
Oxidation (M)Sites.txt	2019-01-17 12:25 PM	TXT File	11,481 KB
parameters.txt	2019-01-17 12:12 PM	TXT File	4 KB
peptides.txt	2019-01-17 12:21 PM	TXT File	150,987 KB
proteinGroups.txt	2019-01-17 12:24 PM	TXT File	66,217 KB
summary.txt	2019-01-17 12:19 PM	TXT File	45 KB
tables.pdf	2019-01-17 12:12 PM	Adobe Acrobat Docu...	179 KB

Glycoproteomics: Asn-\_AspSites.txt  
Phosphoproteomics: Phospho(STY).txt

Modified sites  $\neq$  modified peptides!

# **Tutorial 1: Filtering label-free single-shot DDA data**

# Filtering data

1. **Read in proteinGroups.txt file**
2. **Remove false hits (Reverse, Potential.contaminant, Only.identified.by.site)**
  - Reverse: False positives
  - Potential.contaminant: Proteins that match to contaminant database
  - Only.identified.by.site: Proteins identified based on only modified peptides
3. **Apply filter of minimum 2 unique peptides per protein group**
4. **(Optional) Filter out proteins detected in 2 or more replicates**
  - Note: Only do this if there are at least 3 replicates

# Get intensities

## 1. Get LFQ intensities (“^LFQ.intensity.”)

- Use this if comparing samples within the same search

## 2. Get iBAQ intensities (“^iBAQ.”)

- Use this if comparing samples in different searches
- May need additional normalization

## 3. Log-2 transform data -> to get normal distribution

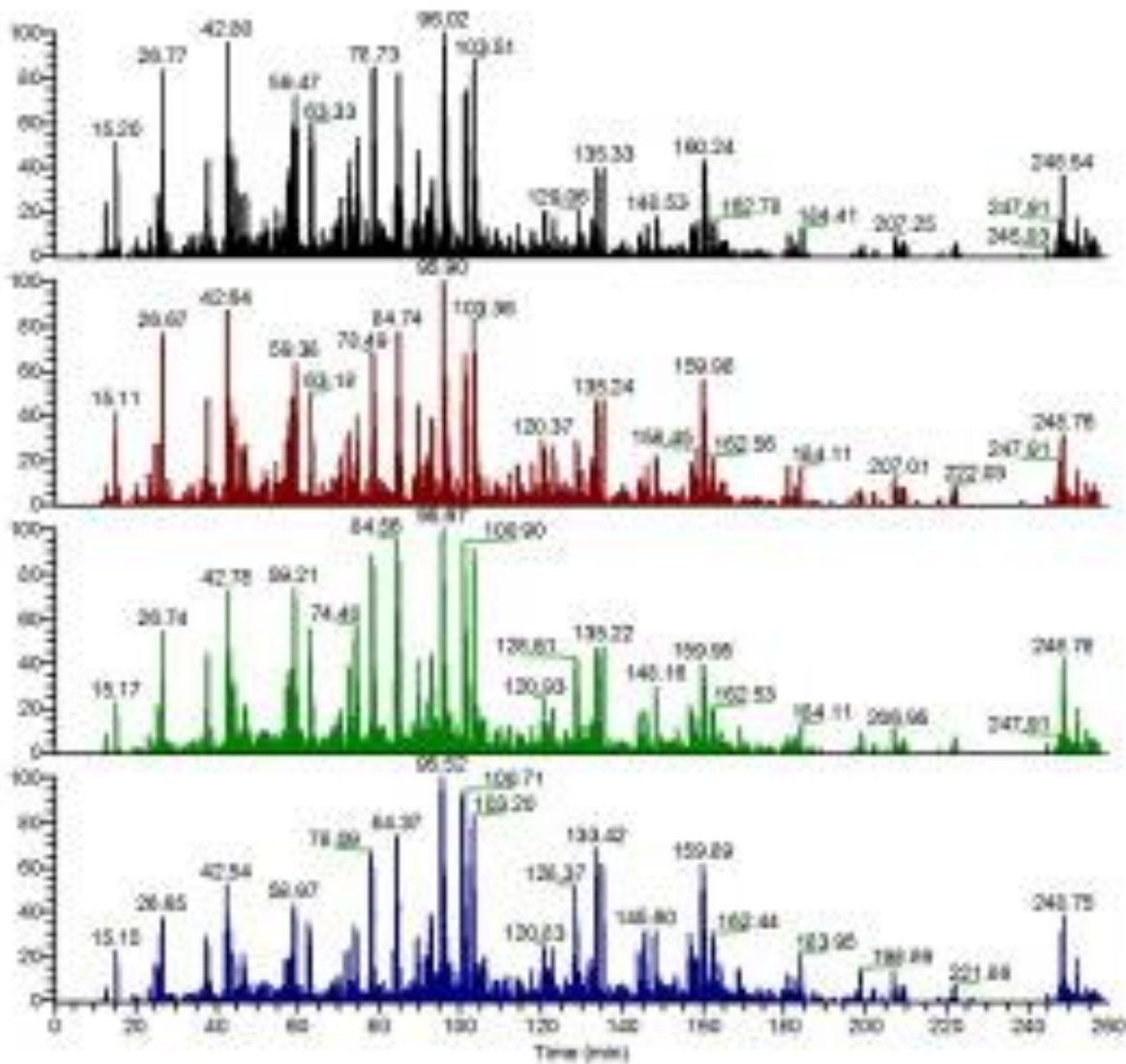
### Note on missing values:

- Missing peptides could either mean that (1) the peptide is present but not detected in that run, or (2) the peptide is absent.



# Checking data quality

Relative Abundance



Retention Time (min)

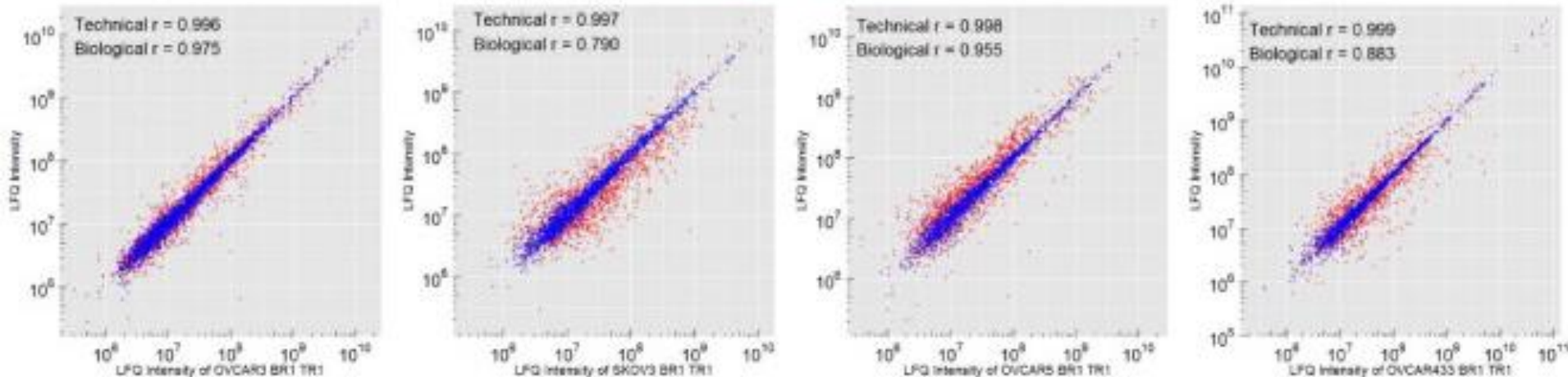
## Options for plotting chromatograms

1. XCalibur (paid software)
2. RforProteomics (R package on Bioconductor)
3. msScans.txt  
=> “Base.peak.intensity” vs “Retention.time”

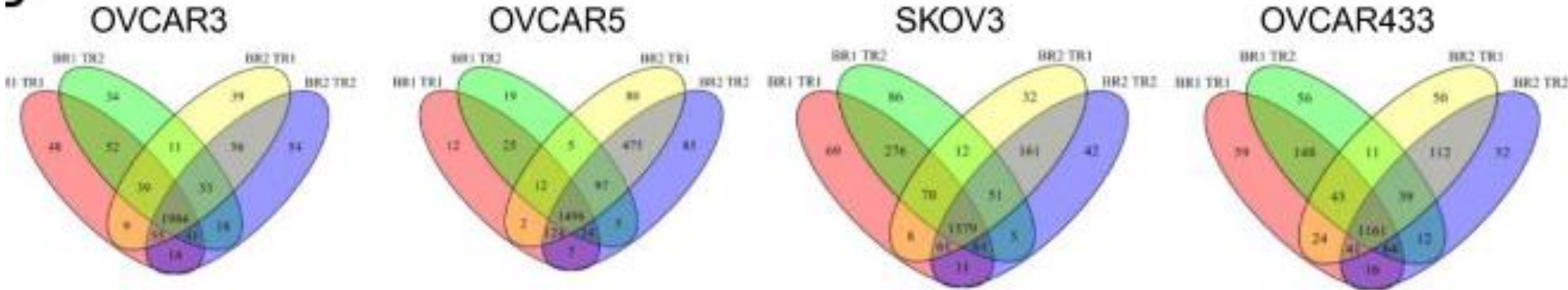
# Checking data quality

C

- Technical Replicate
- Biological Replicate

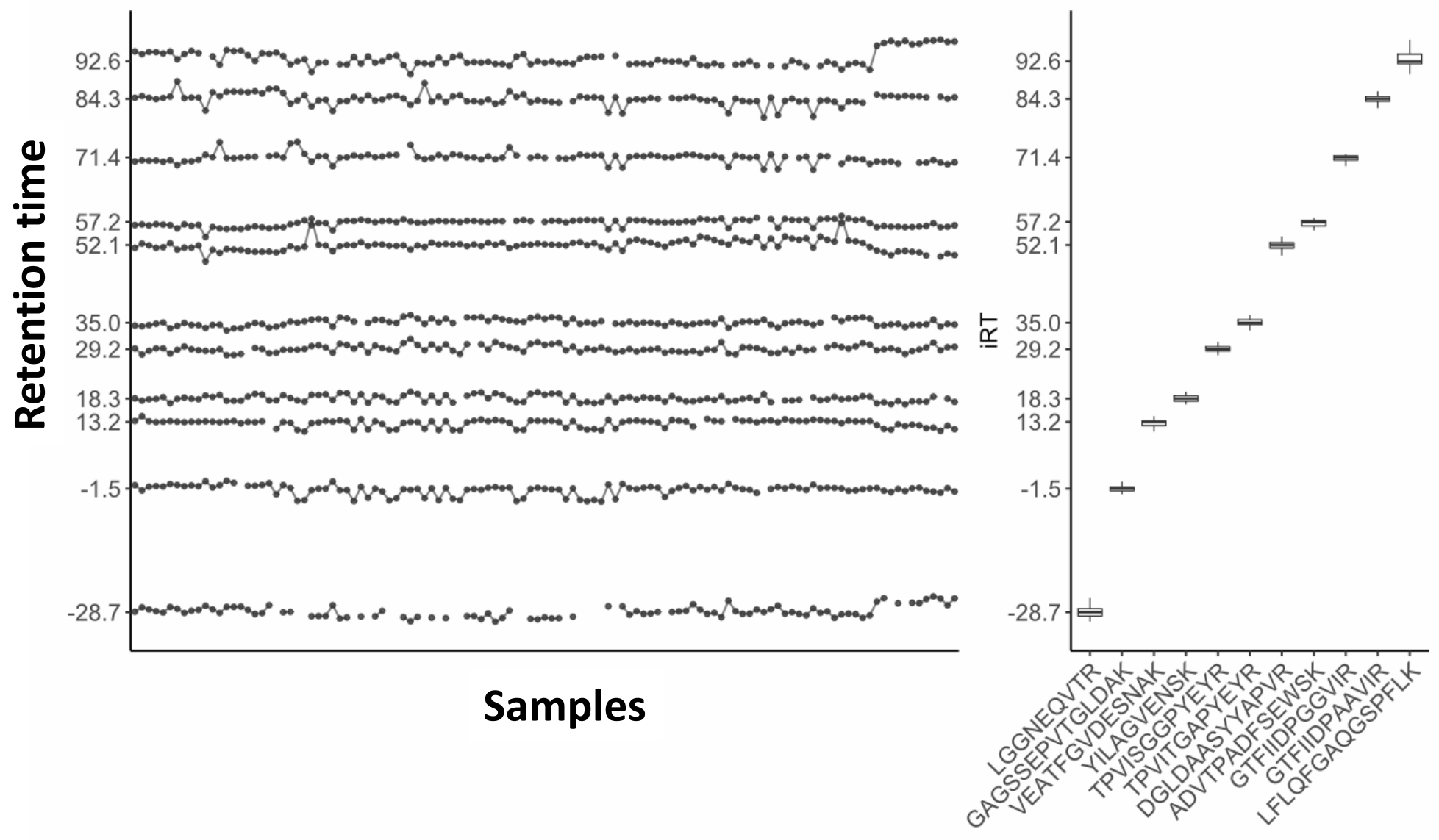


J



# Checking data quality

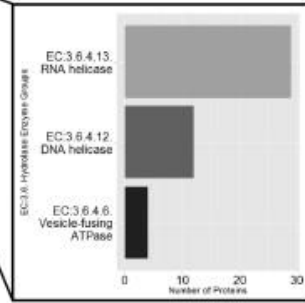
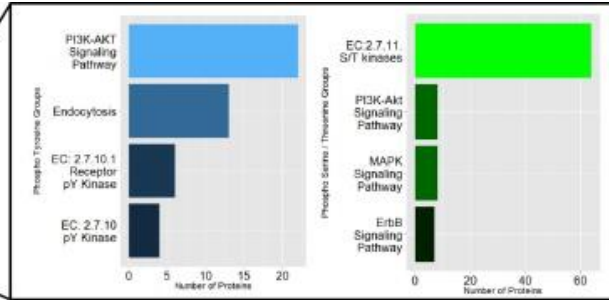
## Chromatographic performance



# Data analysis

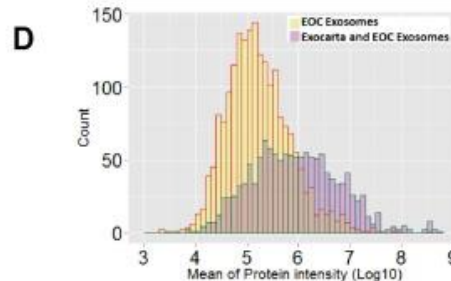
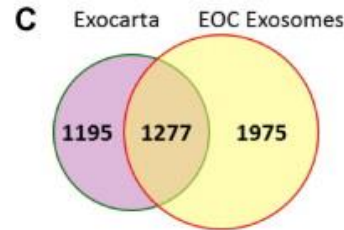
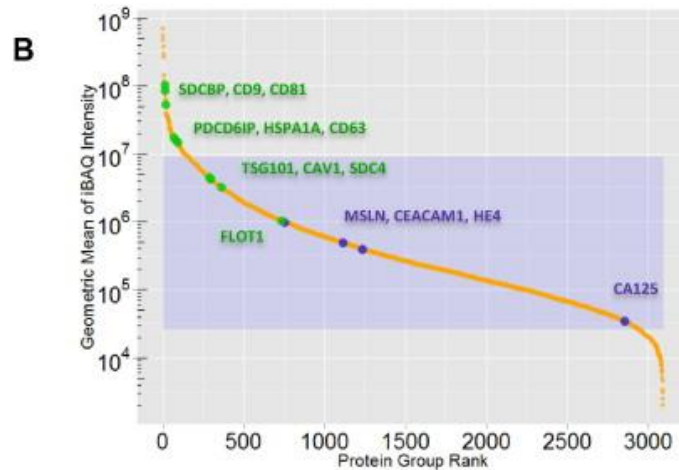
## A Gene enrichment

Description	Count in Exosomes
Acetylation	1169
<b>Phosphoprotein</b>	<b>1795</b>
Nucleotide-binding	491
<b>EC:3.6. Acid anhydride hydrolase</b>	<b>80</b>
EC:6.1.1. Aminoacyl-tRNA ligases	20



Others:  
Heatmaps, volcano plots, etc.

## Protein abundance vs rank



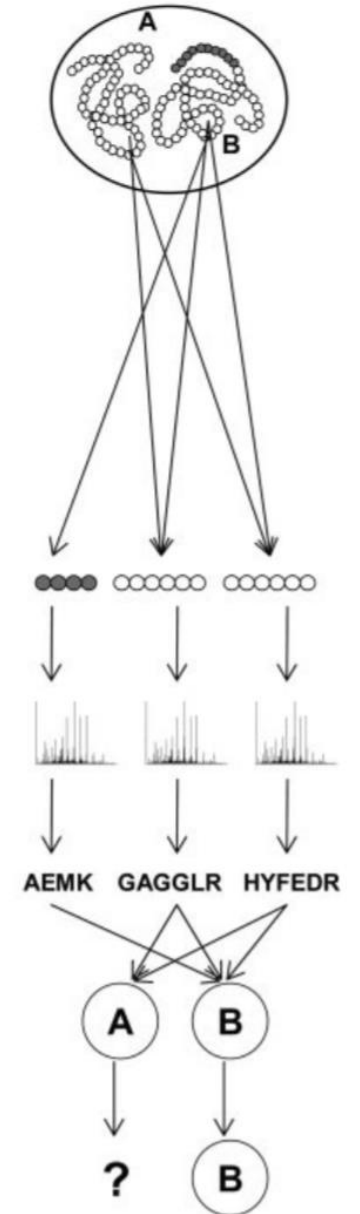
Comparing against other databases

Note: if comparing against RNA-seq data, make protein list **gene-centric**

# Protein inference problem

- Mass spec detects **peptides** (peptide-centric)
- The same peptide can be present in multiple different proteins -> **shared peptides**
- We're interested in knowing what **proteins** are present in the sample
- **Protein** detection based on **unique peptides**

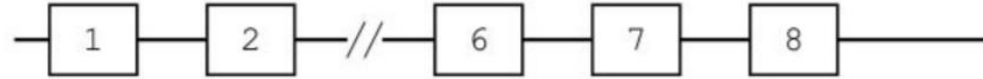
Shotgun Approach



# Protein inference problem: Case studies

## Gene CAPZB

>IPI00026185 IPI:IPI00026185.4|Swiss-Prot:P47756-1|ENSEMBL:ENSP00000264202  
Tax\_Id=9606 Splice isoform 1 of P47756 F-actin capping protein beta subunit



>IPI00218782 IPI:IPI00218782.1|Swiss-Prot:P47756-2|ENSEMBL:ENSP00000264203  
Tax\_Id=9606 Splice isoform 2 of F-actin capping protein beta subunit



P47756-1: MSDQQLDCALDLMRRLPPQ~~Q~~IEKNLSDLIDLVP~~S~~LCEDLLSSVDQPLKIARDKVVGKDYL 60  
MSDQQLDCALDLMRRLPPQ~~Q~~IEKNLSDLIDLVP~~S~~LCEDLLSSVDQPLKIARDKVVGKDYL

P47756-2: MSDQQLDCALDLMRRLPPQ~~Q~~IEKNLSDLIDLVP~~S~~LCEDLLSSVDQPLKIARDKVVGKDYL 60

P47756-1: LCDYNRDGDSYRSPWSNKYDPPLEDGAMP~~S~~ARLRKLEVEANNAFDQYRDLYFEGGVSSVY 120  
LCDYNRDGDSYRSPWSNKYDPPLEDGAMP~~S~~ARLRKLEVEANNAFDQYRDLYFEGGVSSVY

P47756-2: LCDYNRDGDSYRSPWSNKYDPPLEDGAMP~~S~~ARLRKLEVEANNAFDQYRDLYFEGGVSSVY 120

P47756-1: LWDLDHGFAGVILIKKAGDGSKKIKGCWDSIHVVEVQEKSSGRTAHYKLTSTVMLWLQTN 180  
LWDLDHGFAGVILIKKAGDGSKKIKGCWDSIHVVEVQEKSSGRTAHYKLTSTVMLWLQTN

P47756-2: LWDLDHGFAGVILIKKAGDGSKKIKGCWDSIHVVEVQEKSSGRTAHYKLTSTVMLWLQTN 180

P47756-1: KSGSGTMNLGGSLTRQMEKDET~~V~~SDCSPHIANIGRLVEDMENKIRSTLNEIYFGTKDIV 240  
KSGSGTMNLGGSLTRQMEKDET~~V~~SDCSPHIANIGRLVEDMENKIRSTLNEIYFGTKDIV

P47756-2: KSGSGTMNLGGSLTRQMEKDET~~V~~SDCSPHIANIGRLVEDMENKIRSTLNEIYFGTKDIV 240

P47756-1: NGLRSIDAIPDNQKFKQLQRELSQVLTQRQ 270  
NGLRS+ D K + L+ +L + L ++Q

P47756-2: NGLRSVQTFADKSKQEALKNDLVEALKRKQ 270

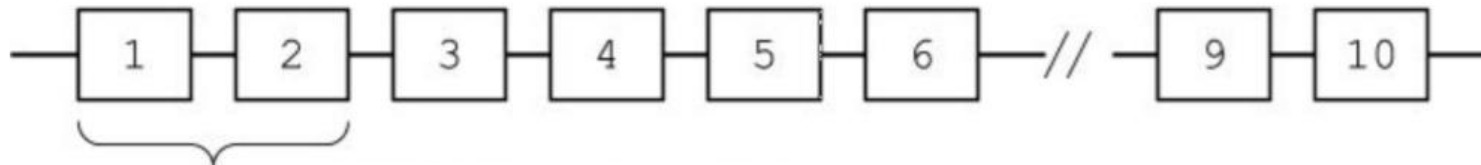
Not detected

**Conclusion:** Isoforms are indistinguishable from each other

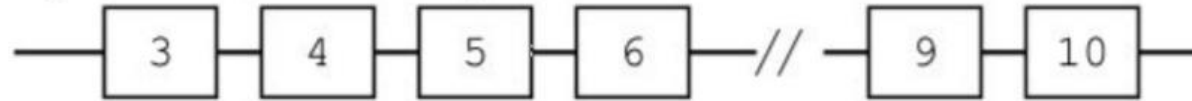
# Protein inference problem: Case studies

Gene: EPLIN

Q9UHB6-1 isoform Beta



Q9UHB6-2 isoform Alpha



Q9UHB6-3 (isoform 3)



> Splice isoform Beta of Q9UHB6 Epithelial protein lost in neoplasm

```
MESSPFNRRQWTSLSLRVTAKESLVNKNKSSAIVEIFSKYQKAAEETNMEKKRSNTENLSQHFRKGTLTVLKKKWENPG  
LGAESHTDSLRNSSTEIRHRADHPPAEVTSHAASGAKADQEEQIHPRSRRLRSPPEALVQGRYPHIKDGEDLKDHSTESKK  
MENCLGESRHEVEKSEISENTDASGKIEKYNVPLNRLKMMFEKGEPTQTKILRAQSRASGRKISENSYSLDDLEIGPGQ  
LSSSTFDSEKNESRRNLELPRLSETSIKDRMAKYQAAVSKQSSSTNYTNELKASGGEIKIHKMEQKENVPPGPEVCITHQ  
EGEKISANENSLAVRSTPAEDDSRDSQVKSEVQQPVHPKPLSPDSRASSLSESSPPKAMKKFQAPARETCVECQKTVPYPM  
ERLLANQQVFHISCFRCSYCNNKLSLGTYASLHGRIYCKPHFNQLFKSKGNYDEGFGHRPHKDLWASKNENEILERPAQ  
LANARETPHSPGVEDAPIAKVGVLAASMEAKASSQQEKEDKPAETKKLRIAWPPPELGSSGSALEEGIKMSKPKWPPED  
EISKPEVPEDVDLDLKKLRRSSSLKERSRPFTVAASFQSTSVKSPKTVSPPIRKGWSMSEQSEESVGGRVAERKQVENAK  
ASKKNGNVGKTTWQNKESKGETGKRSKEGHSLEMENENLVENGADSDEDDNSFLKQQSPQEPKSLNWSSFVDNTFAEEFT  
TQNQKSQDVELWEGEVVKELSVEEQIKRNRYYDEDEDEE
```

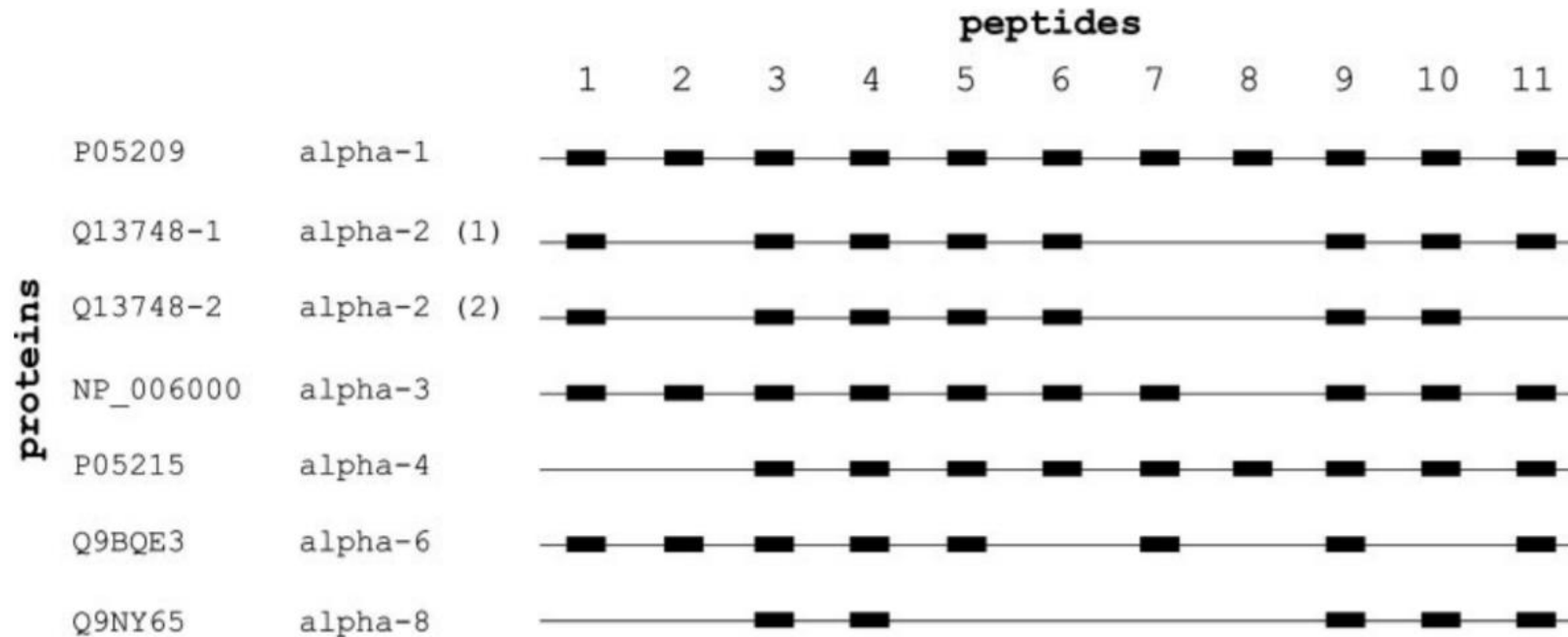
# Protein groups

None of the proteins was detected with a unique peptide

## Peptides identified:

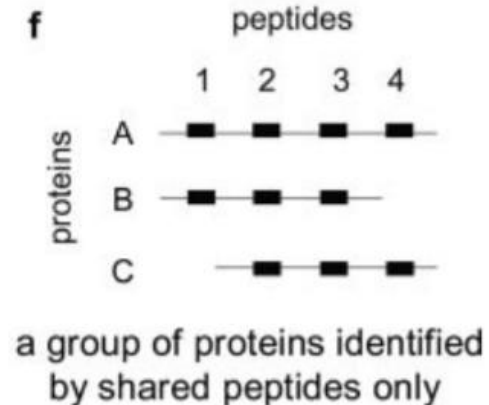
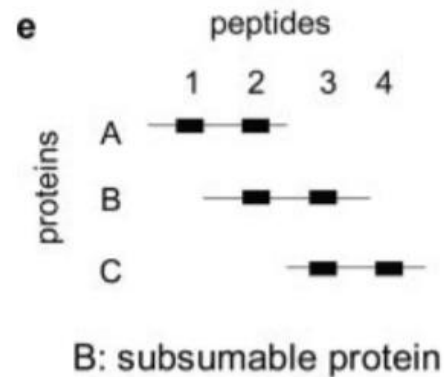
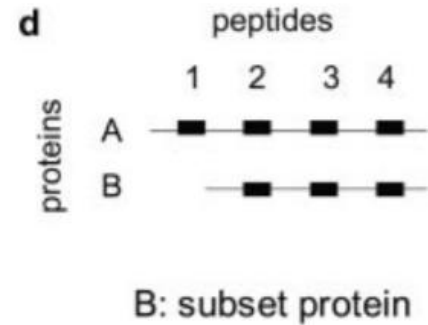
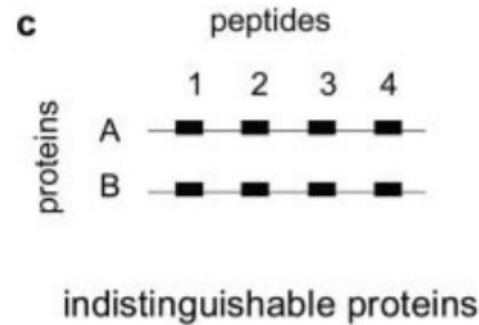
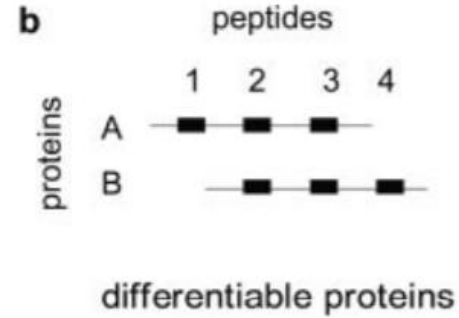
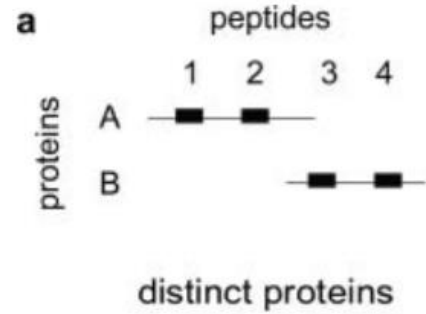
1	TIGGGDDSFNTFFSETGAGK	5	IHFPLATYAPVISA EK	9	VGINYQPPTVVPGGDLAK
2	AVFVDLEPTVIDEVR	6	AYHEQLSVAEITNACFEPANQMVK	10	AVCMLSNTTAAIAEAWAR
3	QLFHPEQLITGKEDAANNYAR	7	YMACCLLYR	11	LDHKFDLMYAK
4	NLDIERPTYTNLNR	8	SIQFVDWCPTGFK		

## Assignment of peptides to proteins:





# Peptide grouping scenarios



**Set of all detected proteins** = the minimum number of proteins sufficient to explain all observed peptides

- Includes distinct and differentiable proteins

- Situations c-f: presented as a **protein group**

# Razor vs unique peptides

**Unique peptide:** Peptide unique to one protein group

**Razor peptide:** Peptide shared between protein groups, but assigned to the protein group with more peptides

**Protein A**      Peptide A, Peptide B, Peptide C, Peptide D

**Protein B**      Peptide A, Peptide B, Peptide E

**Protein C**      Peptide A

# Razor vs unique peptides

**Unique peptide:** Peptide unique to one protein group

**Razor peptide:** Peptide shared between protein groups, but assigned to the protein group with more peptides

Razor peptides



**Protein A**     Peptide A, Peptide B, Peptide C, Peptide D

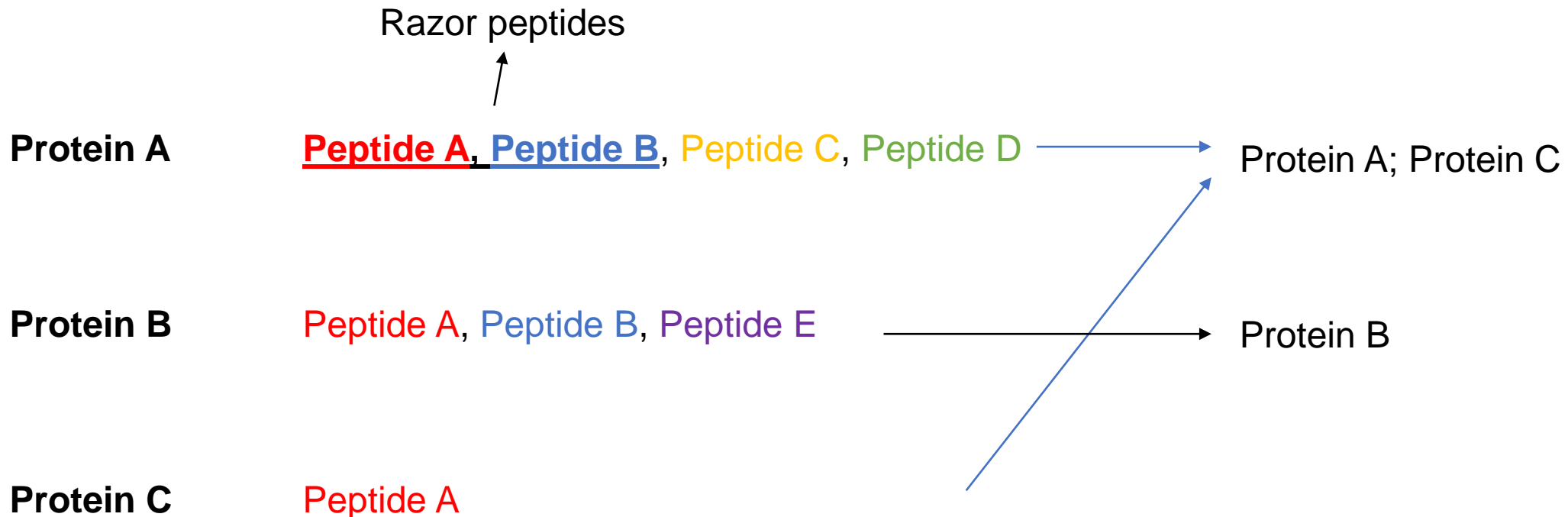
**Protein B**     Peptide A, Peptide B, Peptide E

**Protein C**     Peptide A

# Protein IDs

**Unique peptide:** Peptide unique to one protein group

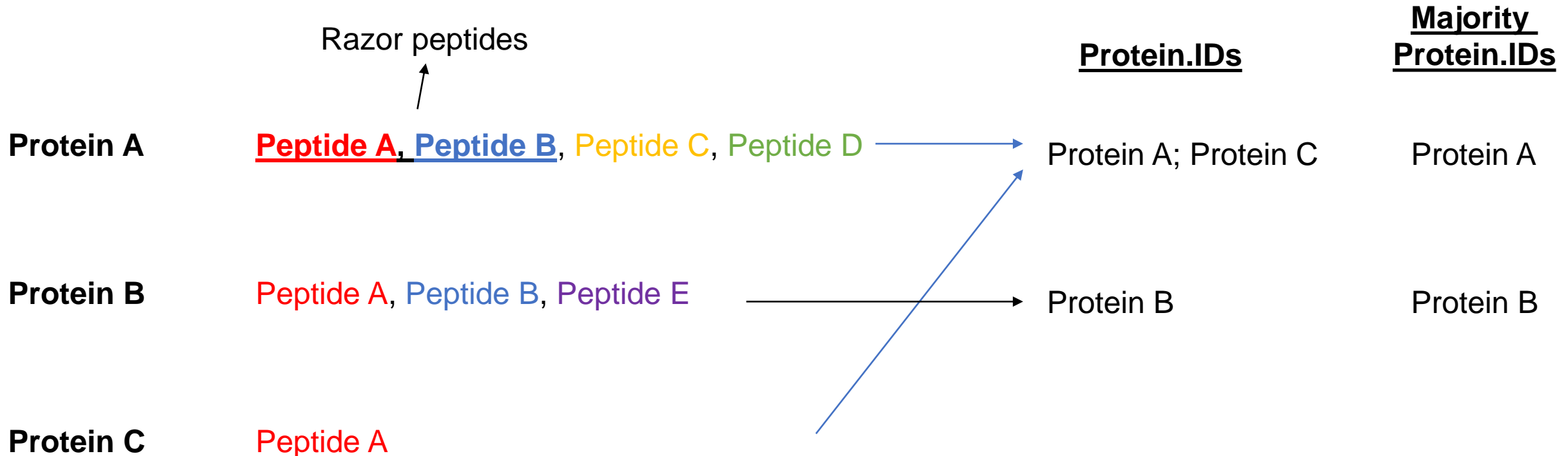
**Razor peptide:** Peptide shared between protein groups, but assigned to the protein group with more peptides



# Majority protein IDs

**Unique peptide:** Peptide unique to one protein group

**Razor peptide:** Peptide shared between protein groups, but assigned to the protein group with more peptides



# Factors affecting peptide detection

- Presence of tryptic sites – Arg (R) and Lys (K)
- Accessibility to enzyme - PTMs
- Length – 7-22 amino acids
- Low abundance
- Poor ionization
- Difficult to fragment

# Data repositories

**MassIVE** – **Mass** Spectrometry **I**nteractive **V**irtual **E**nvironment

<https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp>

- Raw files
- MaxQuant search output

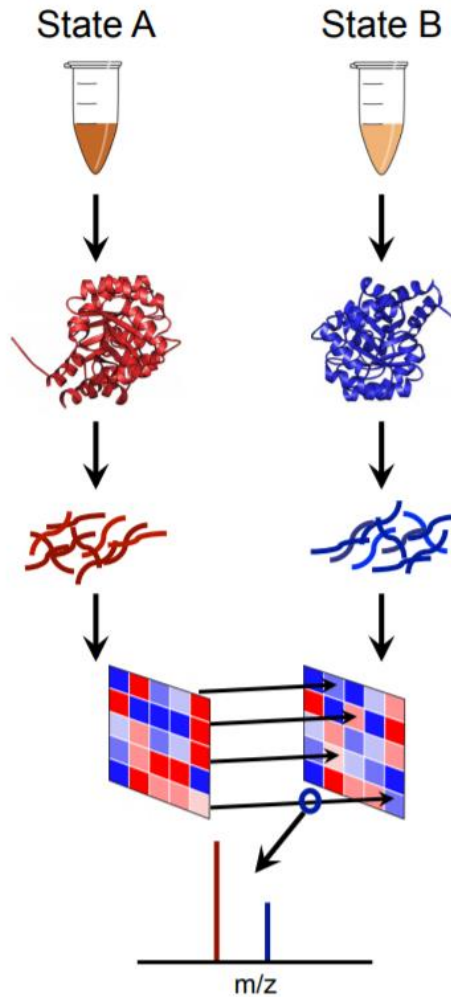
**ProteomeXchange**

<http://www.proteomexchange.org/>

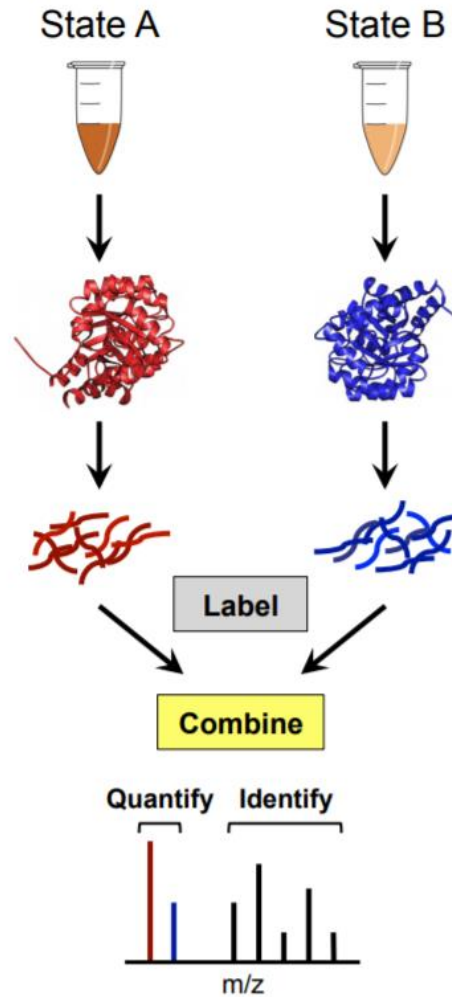
- PXDxxxxxx

# Relative Quantification Methods for Discovery Proteomics

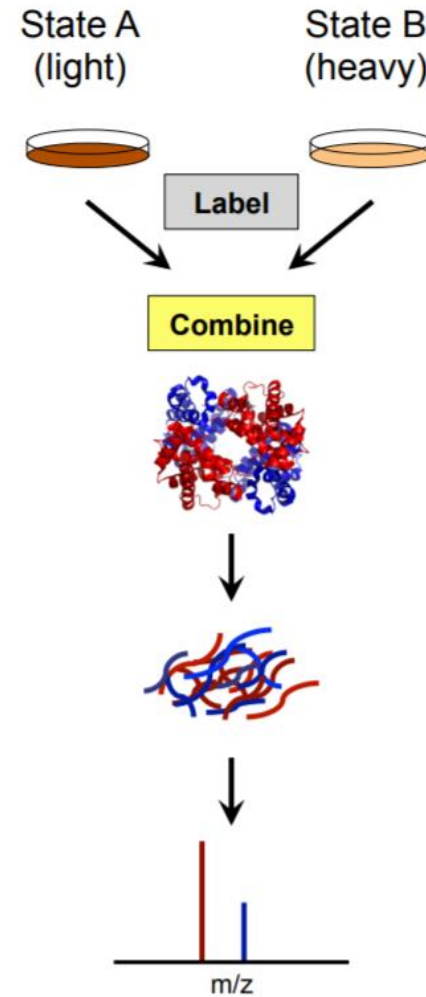
**Label-free quantification**  
(1 sample at a time)



**Chemical labeling**  
(up to 10 samples at a time)



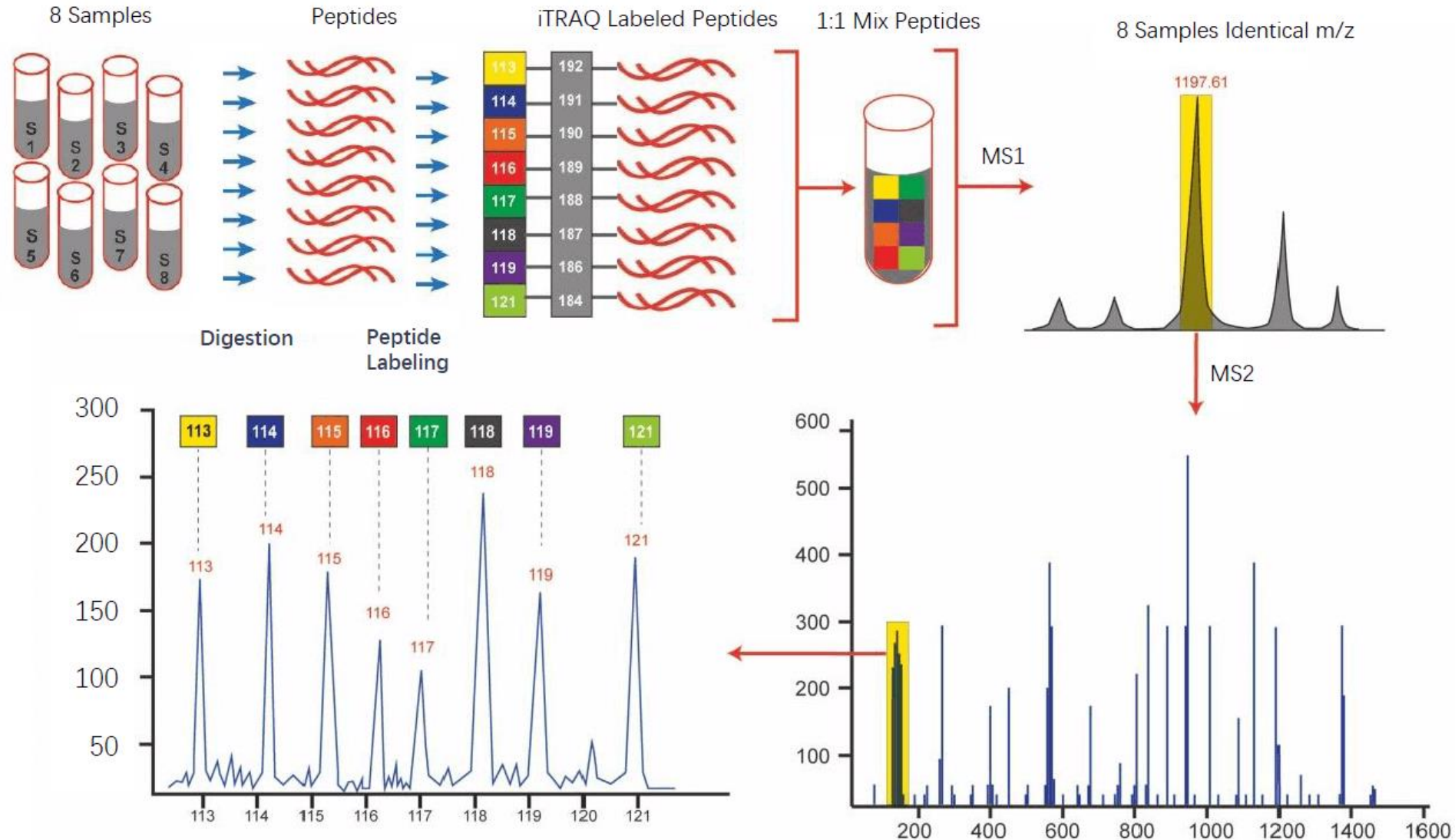
**Metabolic labeling (SILAC)**  
(up to 3 samples at a time)



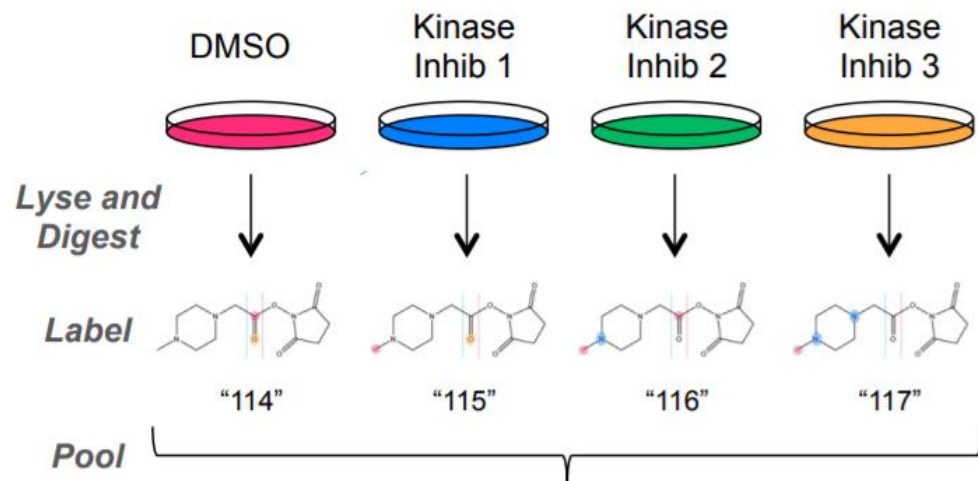
TMT  
iTRAQ



# Multiplexing (TMT, iTRAQ) – MS2 level quantitation



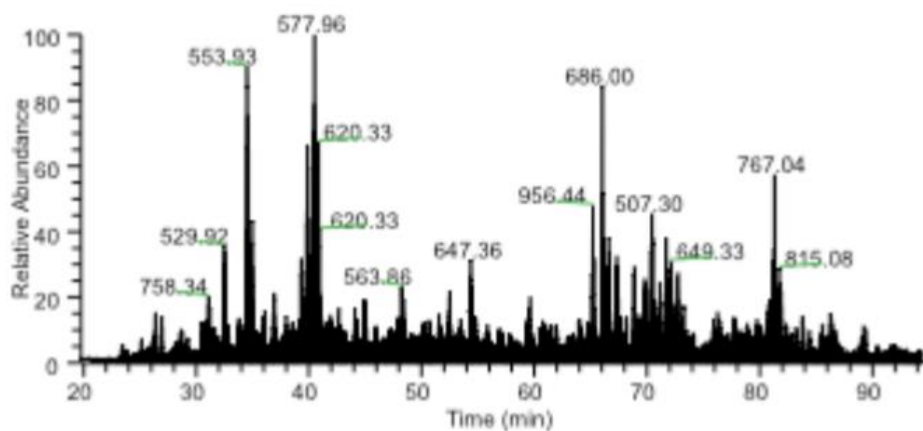
# iTRAQ Experimental Example



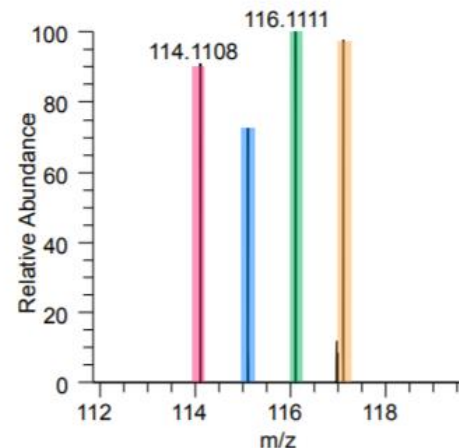
Phosphopeptide Enrichment



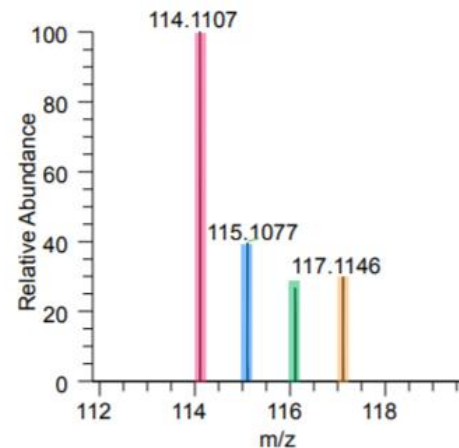
LCMS



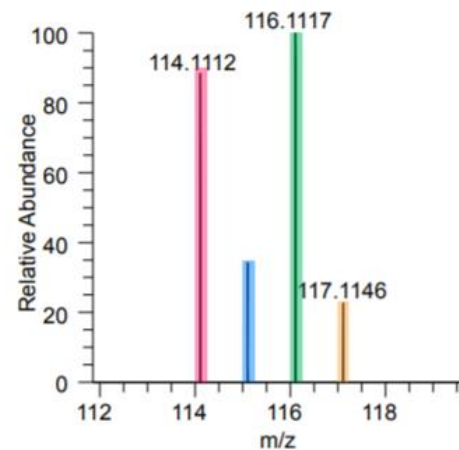
Peptide #1:  
No effect



Peptide #2:  
Sensitive to  
all  
inhibitors



Peptide #3:  
Sensitive to  
inhibitors 1 &  
3



# Multiplexing

## PROS

- Reduced run-to-run variation
- “High-throughput”: Up to 11 samples at once
- More robust quantitation
- Higher sensitivity for low abundance peptides

## CONS

- Requires fractionation
- Can only compare samples within a set
- Requires fixed study design i.e. if you want to run more samples later on, new samples might not be directly comparable to older samples

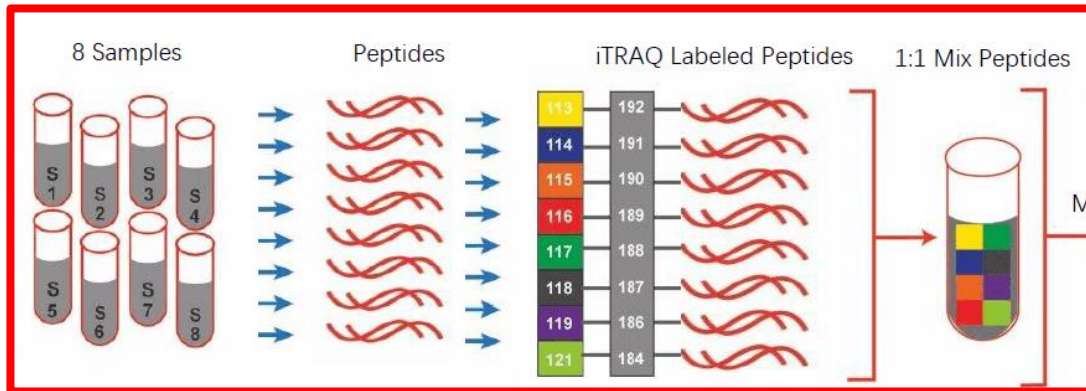
# Multiplexing

## PROS

- Reduced run-to-run variation
- “High-throughput”: Up to 11 samples at once
- More robust quantitation
- Higher sensitivity for low abundance peptides

## CONS

- Requires fractionation
- Can only compare samples within a set
- Requires fixed study design i.e. if you want to run more samples later on, new samples might not be directly comparable to older samples



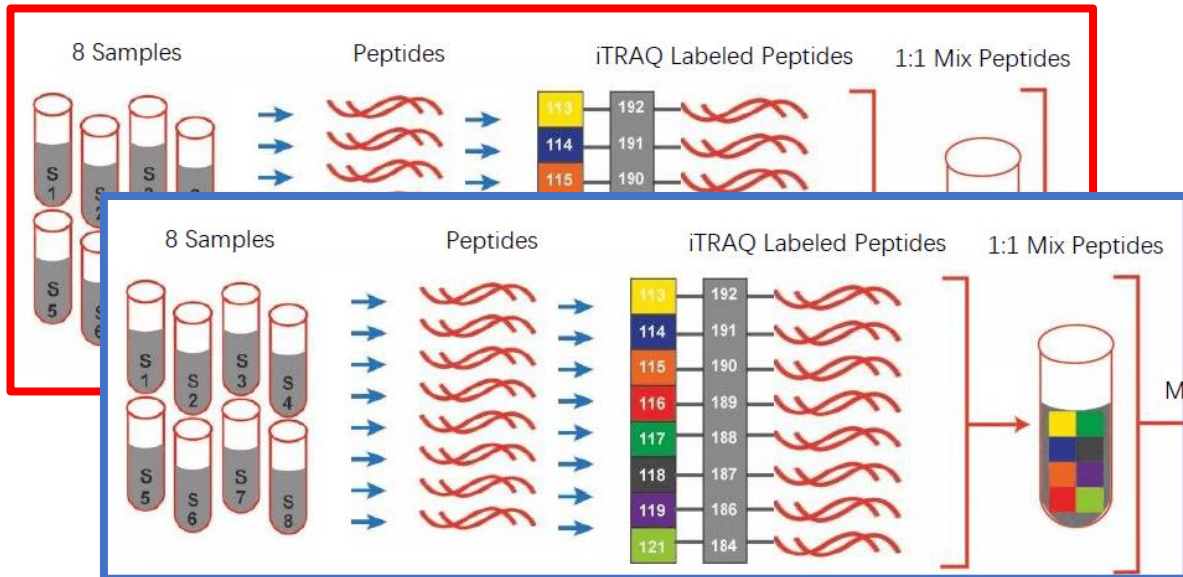
# Multiplexing

## PROS

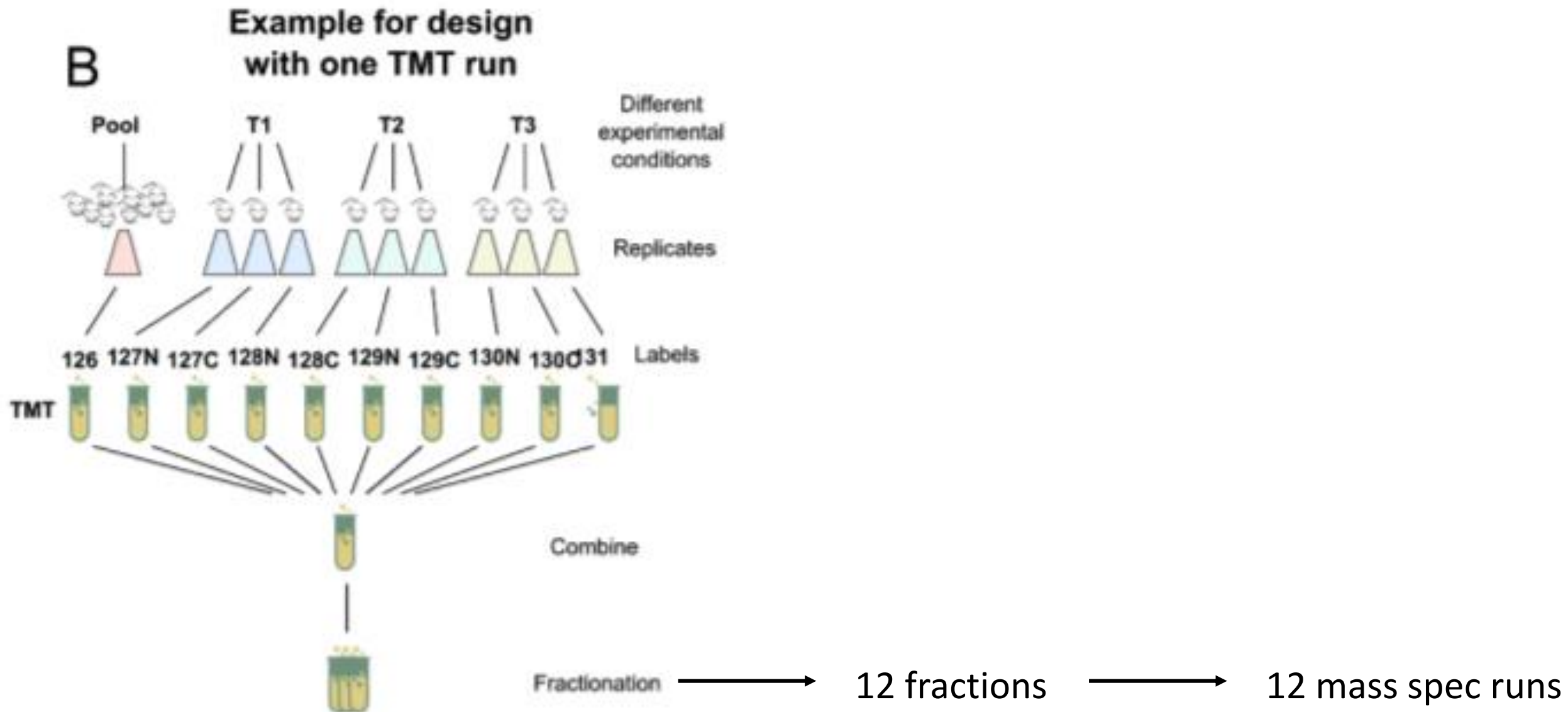
- Reduced run-to-run variation
- “High-throughput”: Up to 11 samples at once
- More robust quantitation
- Higher sensitivity for low abundance peptides

## CONS

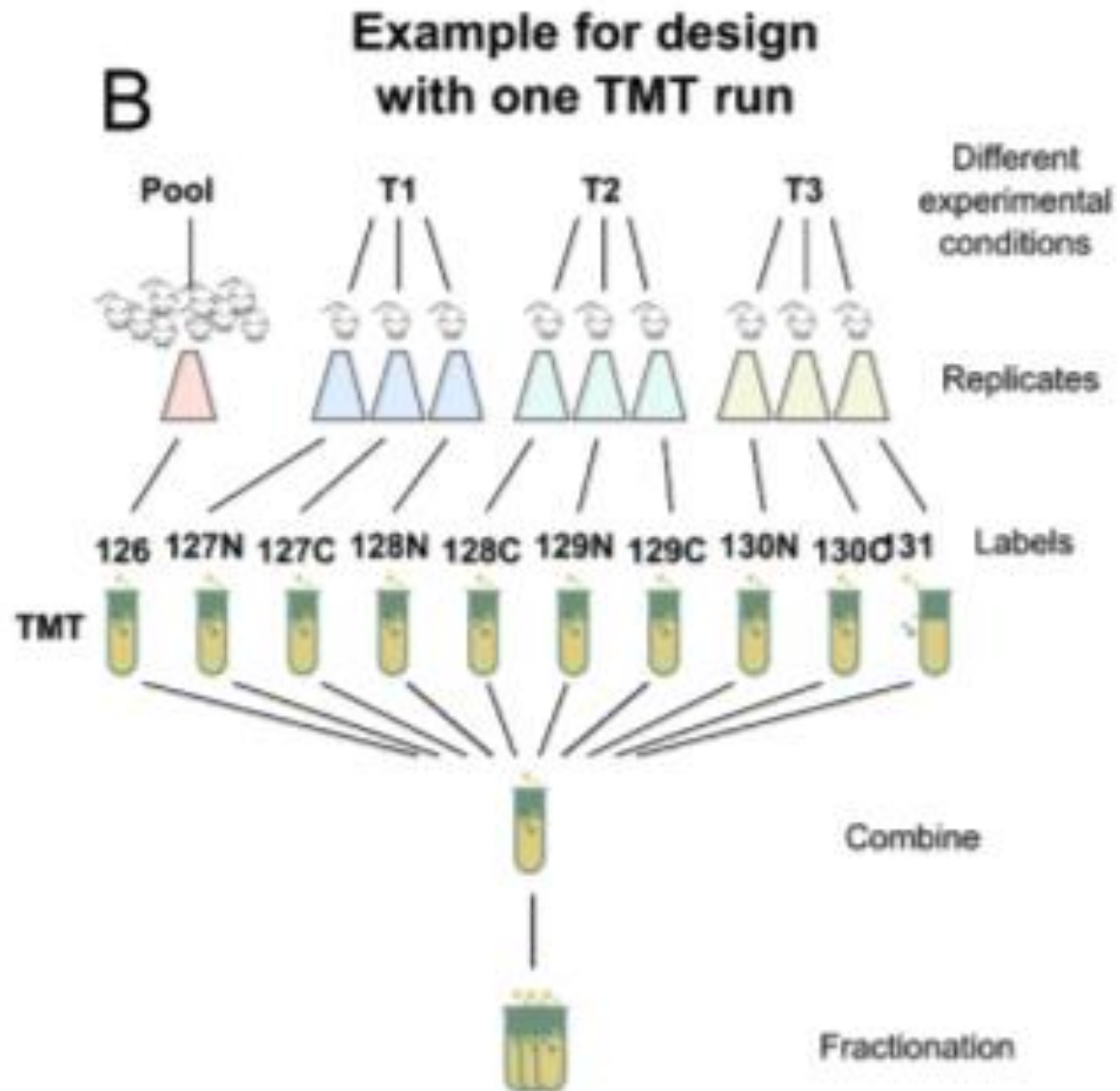
- Requires fractionation
- Can only compare samples within a set
- Requires fixed study design i.e. if you want to run more samples later on, new samples might not be directly comparable to older samples



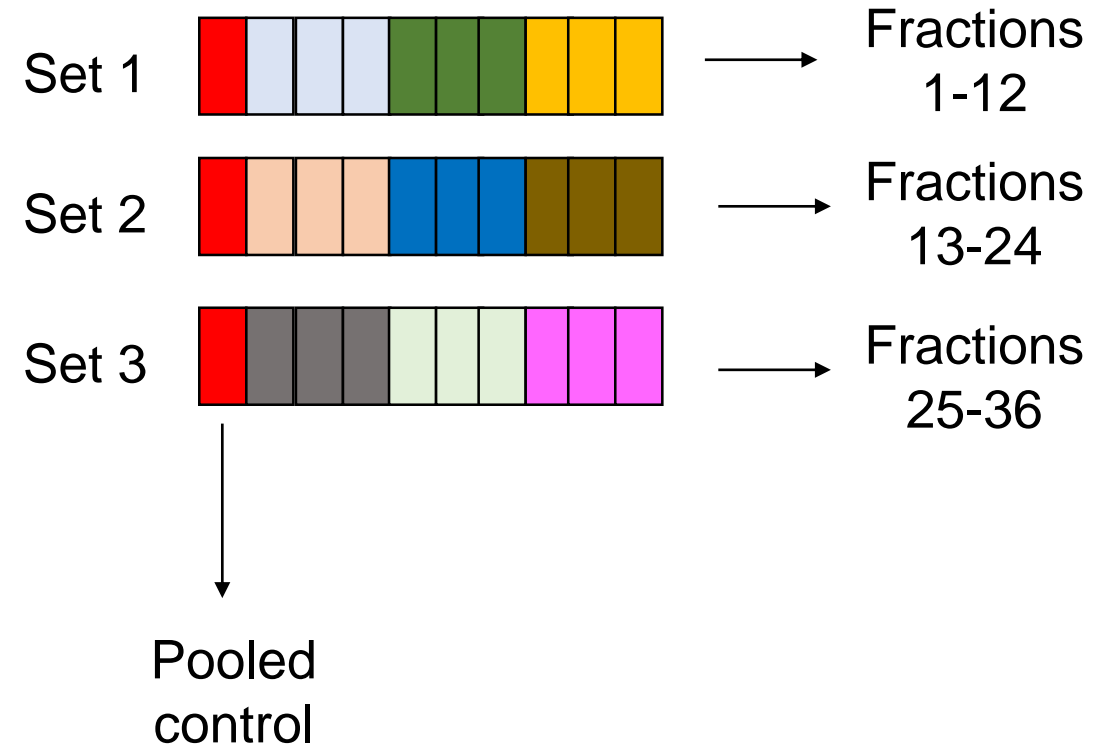
# TMT experimental design



# TMT experimental design



## More than one TMT run



# Setting up TMT database search

The screenshot shows the MaxQuant software interface. The window title is "Session1 - MaxQuant". The menu bar includes "File", "Tools", "Window", and "Help". Below the menu bar, there are tabs for "Raw files", "Group-specific parameters", "Global parameters", "Performance", "Viewer", and "Configuration". The "Group-specific parameters" tab is active, showing "Group 0" selected. The "Type" sub-tab is selected, displaying a list of search methods: "Standard", "Reporter ion MS2", "Reporter ion MS3", "NeuCode", and "Quantification only no calib". Below this list, there are two columns of checkboxes for various modifications: T8O, Arg10, Arg6, DimethLys0, DimethLys2, DimethLys4, DimethLys6, DimethLys8, DimethNter0, Dimeth, Dimeth, Dimeth, ICAT-0, ICAT-9, ICPL-L, ICPL-L, and ICPL-L. At the bottom of the window, there is a control panel with a "Number of threads" dropdown set to 1, buttons for "Start", "Stop", "Partial processing", and "Details", and a checkbox for "Send email when done". The version number "Version 1.5.8.3" is displayed in the bottom right corner.

Reporter ion **MS2** or **Reporter MS3** depending on acquisition method

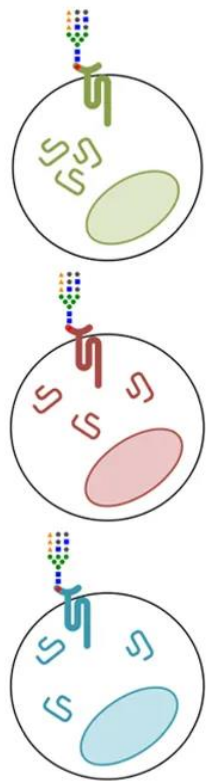
- > MS2: quantitation from MS2 level spectra
- > MS3: (SPS-MS3) quantitation from MS3 level spectra



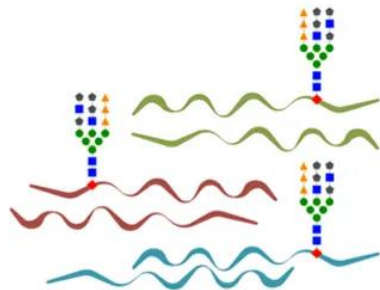
## **Tutorial 2: Filtering TMT DDA data**

# Glycoproteomics

b

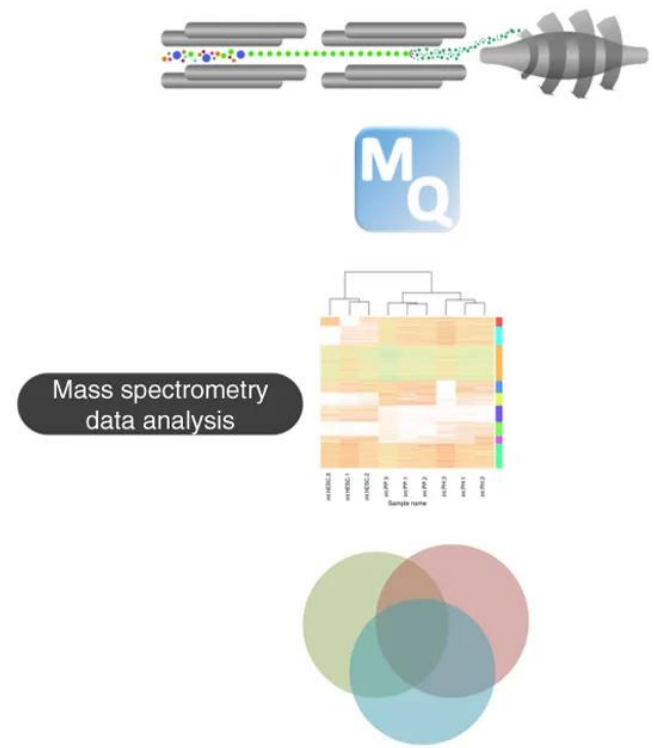
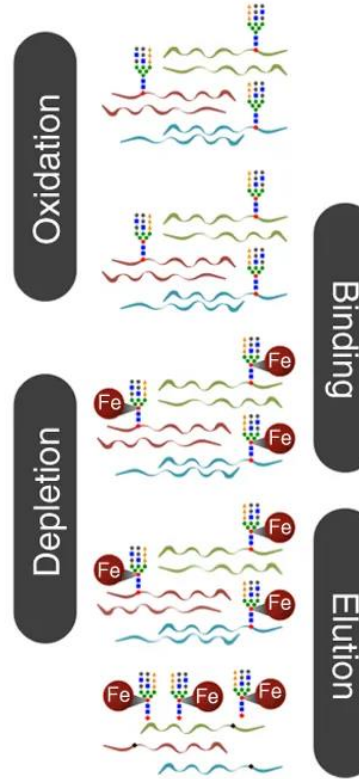


- Invertase
- Extraction
- Reduction
- Alkylation
- Digestion



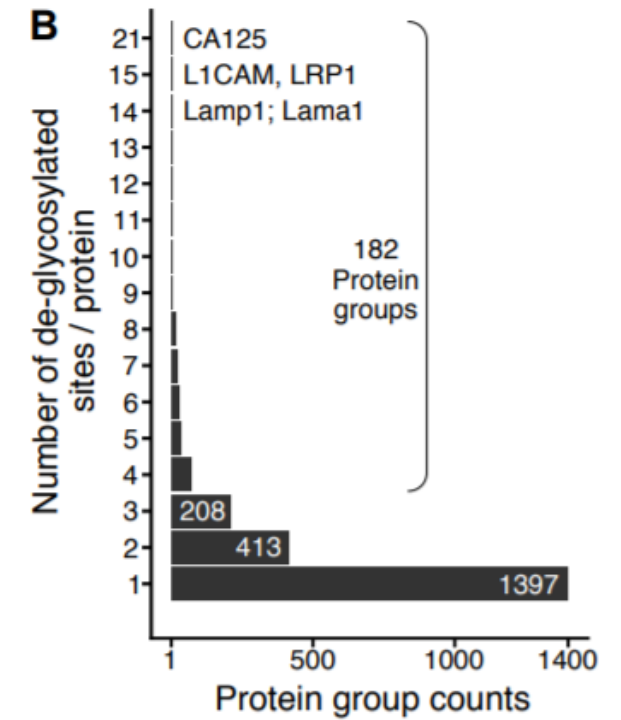
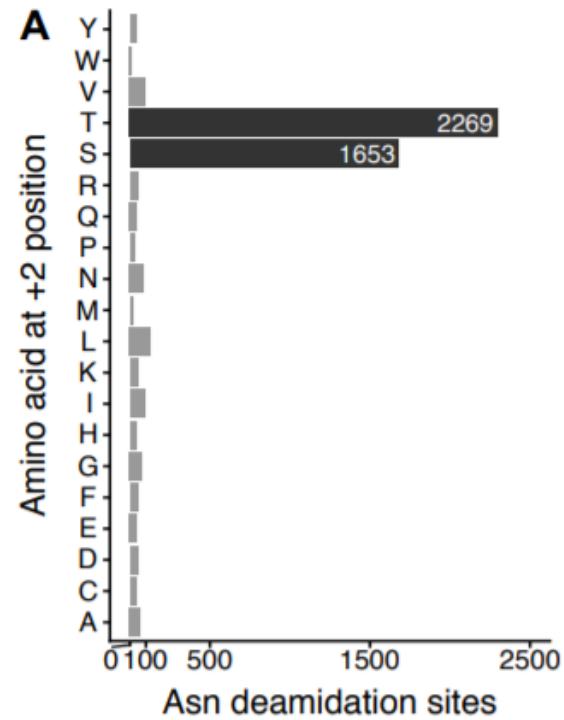
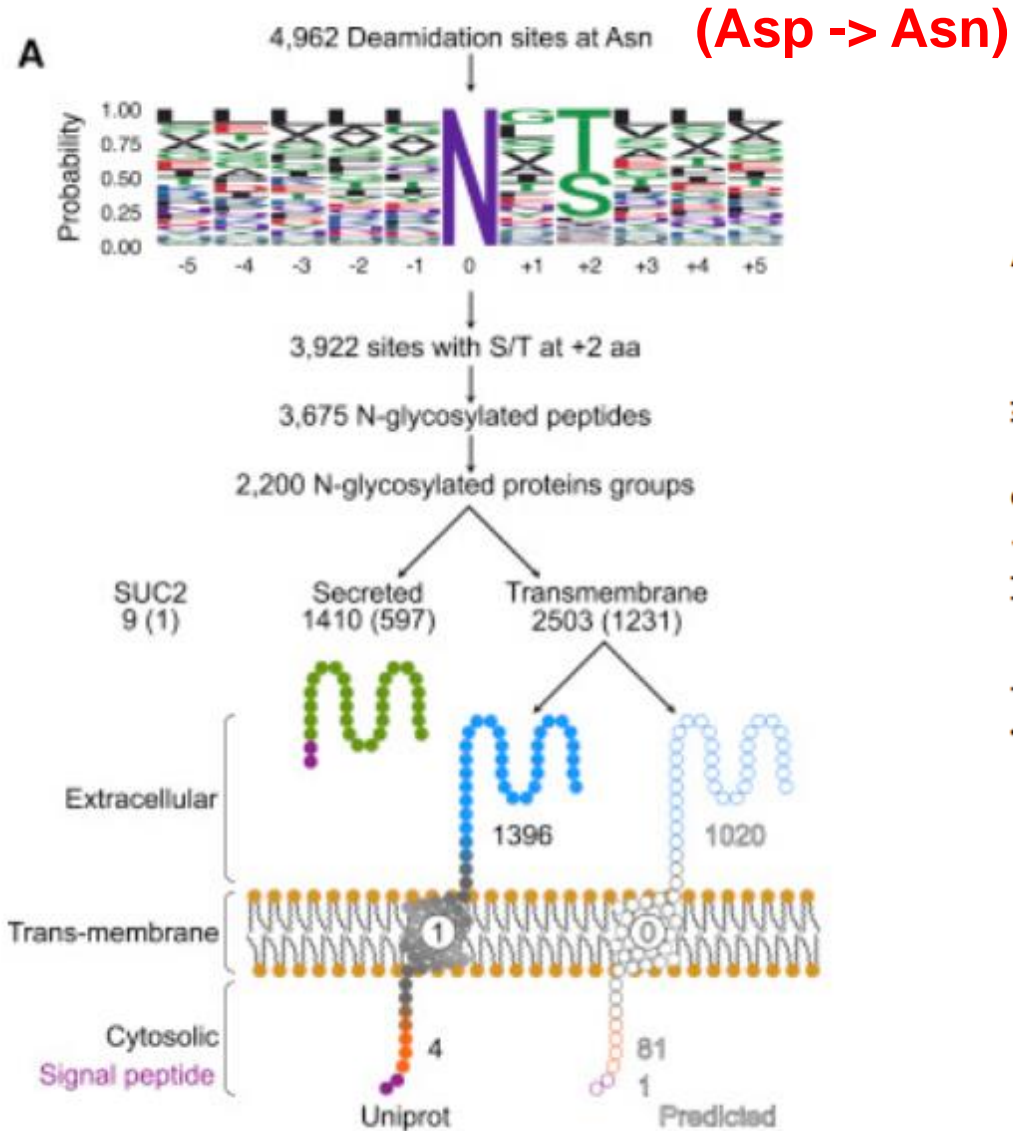
Glyco-peptide enrichment

(Asp -> Asn)



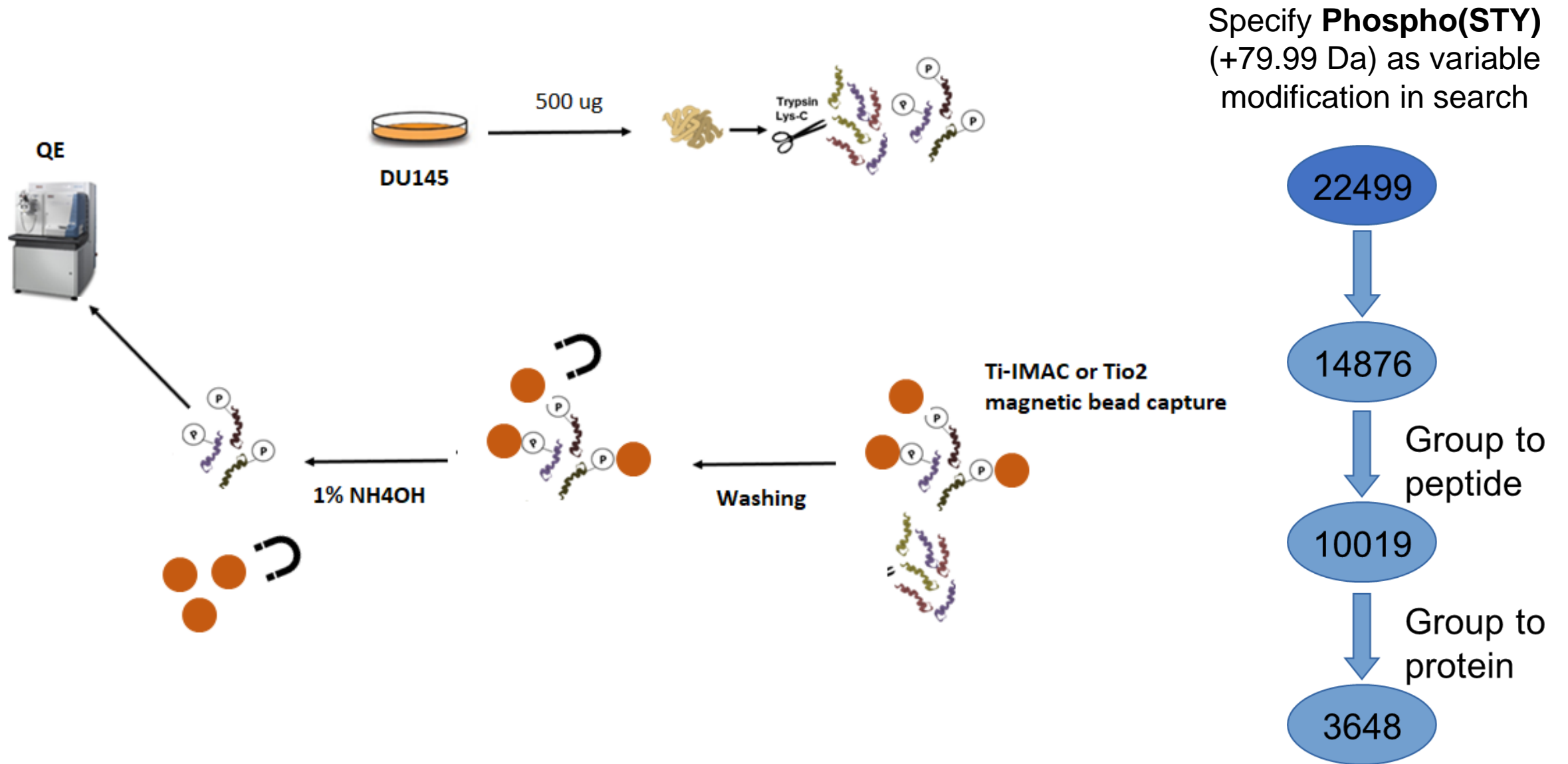
Specify **Asn -> Asp** (-1 Da)  
as variable modification in search

# Glycoproteomics



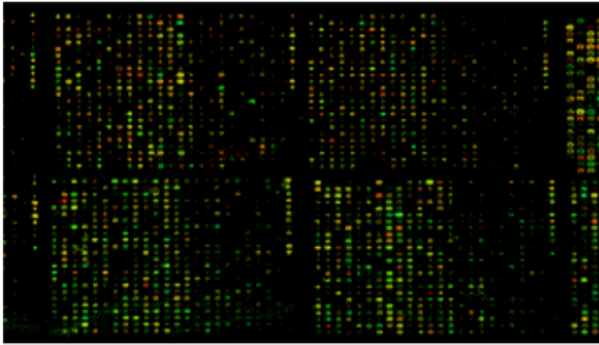
## **Tutorial 3: Filtering Glycoproteomics DDA data**

# Phosphoproteomics



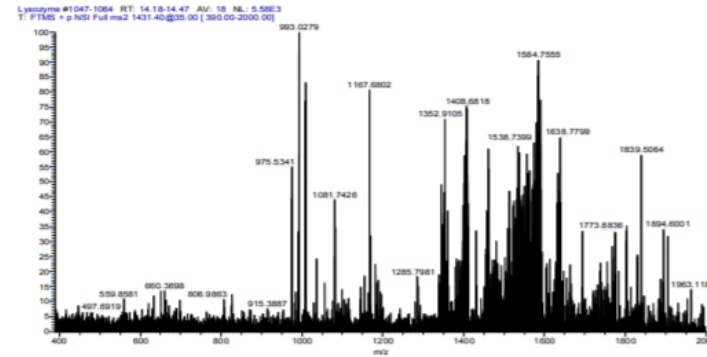
# Analytical challenges of proteomics differ in important ways from transcriptional analysis

## Transcriptional Profiling



- All possible features known
- Sample is static during analysis
- All features measured
- Robust means to amplify low numbers DNA or RNA (PCR)
- Signal not detected means feature not present

## MS-based Proteomics



- All possible features not known
- Sample is dynamic during analysis
- 20-50% of features measured
- No protein PCR (analytics have to deal with enormous dynamic range)
- Signal not detected means either that feature not present or feature present but not detected

# Proteomics part #2: Proteogenomics

- Data imputation
- Integrated “omics” analysis
- Protein identification from lncRNA, circRNA, etc

## **Supplementary slides**



## PEP Score

- Posterior error probability (PEP) is calculated using *Bayesian* statistics as a probability of false hit using the peptide identification score (s) and length of peptide(l).

$$p(X = \text{false}|s, L) = \frac{p(s, L|X = \text{false})p(X = \text{false})}{p(s, L)}$$

- The smaller the PEP, the more certain is the identification of a peptide.
- Longer peptides are automatically accepted with lower scores (based on their parent mass).

Longer peptides: less likely to be identified by chance

PEP score proteins: multiply peptide PEPs. Only peptides with distinct sequences and highest-scoring peptides are used.