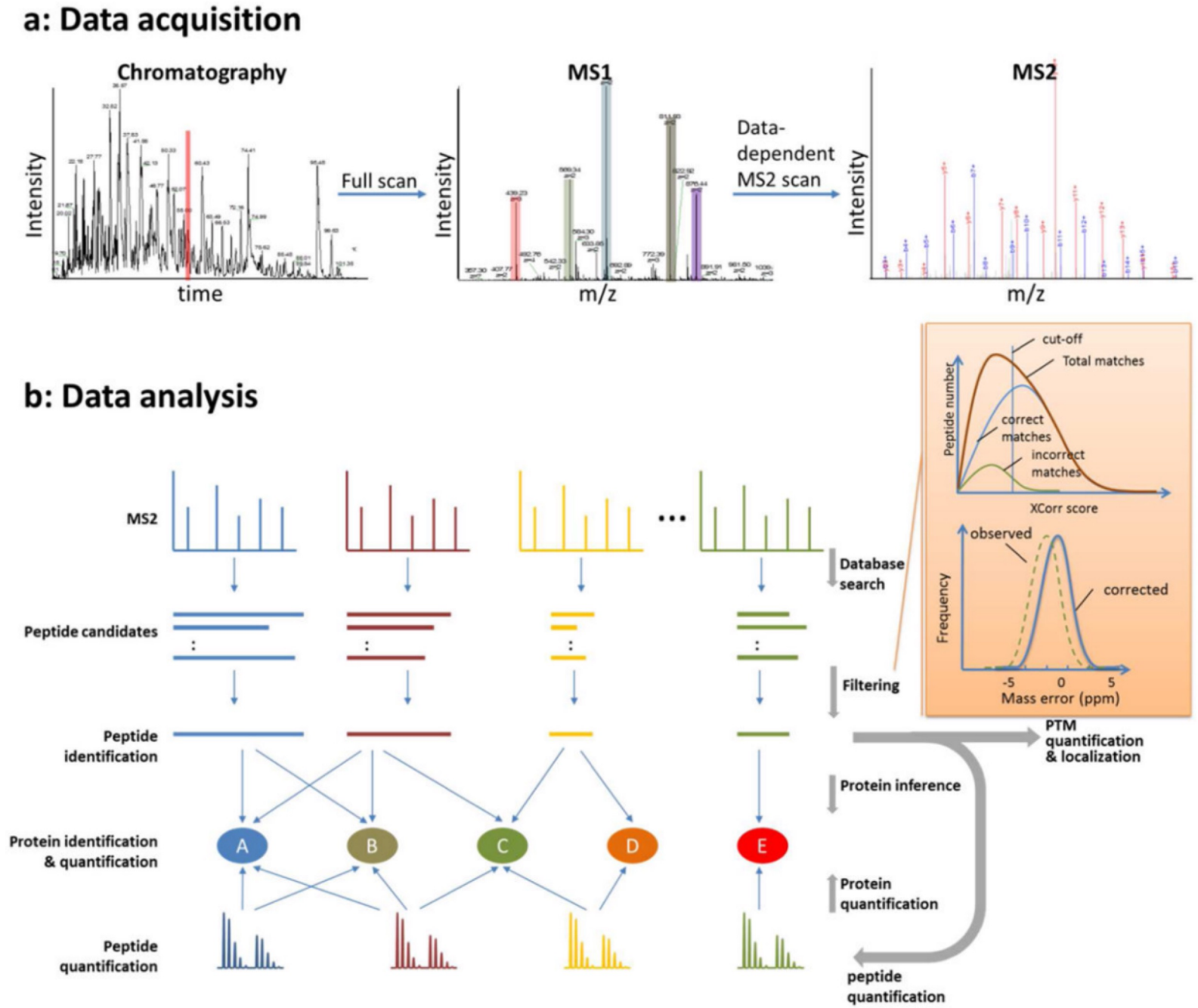
Shot Gun Proteomics

- “Bottom-up” protein analysis refers to the characterization of proteins by analysis of peptides released from the protein through proteolysis.
- When bottom-up is performed on a mixture of proteins it is called **shotgun proteomics.**
  - In a typical shotgun proteomics experiment, the peptide mixture is fractionated and subjected to LC-MS/MS analysis.
  - Peptide identification is achieved by comparing the tandem mass spectra derived from peptide fragmentation with theoretical tandem mass spectra generated from *in silico* digestion of a protein database.
  - Protein inference is accomplished by assigning peptide sequences to proteins.

Representative LC-MS/MS data and a generalized bioinformatic analysis pipeline for protein identification and quantification in shotgun proteomics.



Zhang, Yaoyang et al. “Protein analysis by shotgun/bottom-up proteomics.” *Chemical reviews* vol. 113,4 (2013): 2343-94. doi:10.1021/cr3003533

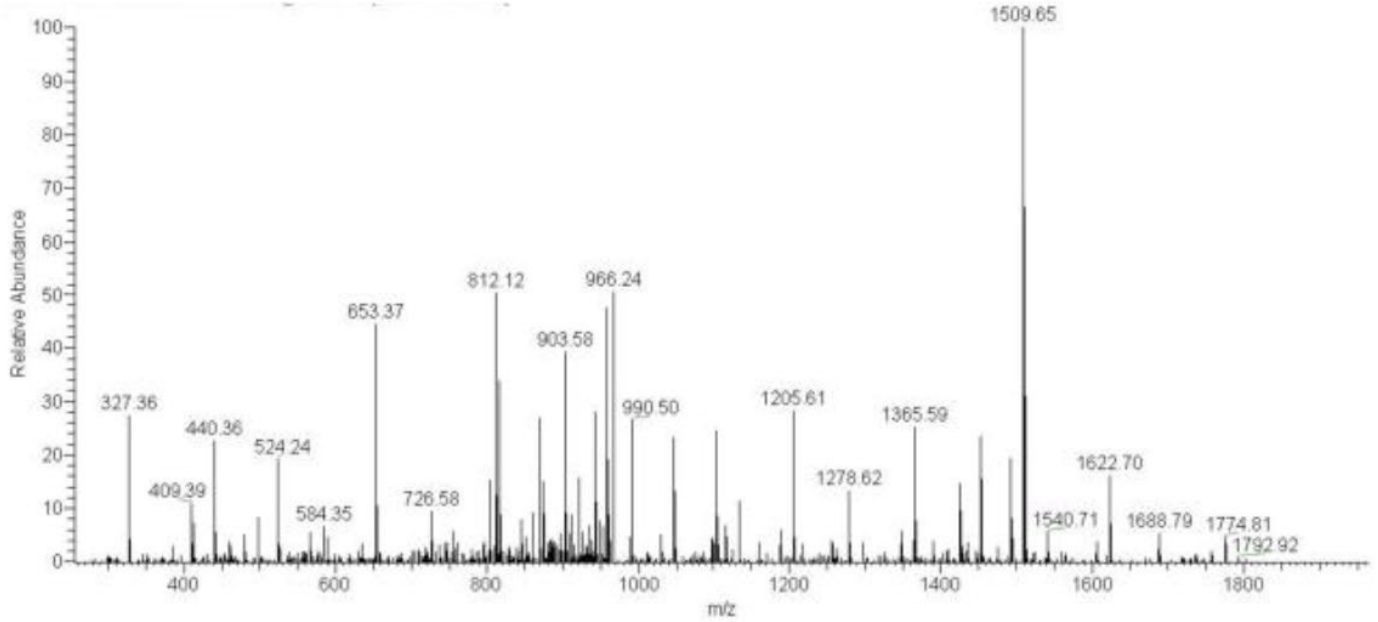
LECTURE 4: QUANTITATIVE MASS SPECTROMETRY-BASED PROTEOMICS

Quantitative Proteomics

- Quantitative proteomics aims at simultaneously quantitation of level differences between many proteins in different samples, not at measurement of their absolute concentrations.

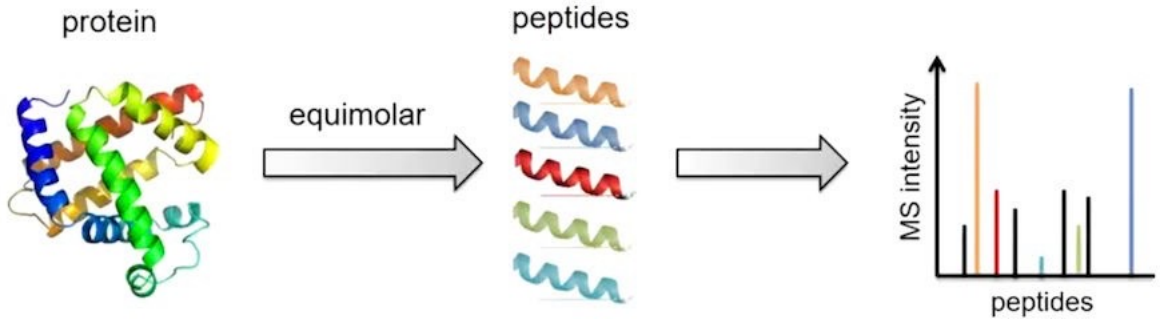
- Mass spectrum records a whole bunch of m/z

- BUT MS intensity does **NOT** tell us peptide abundance directly.



Why does MS^1 intensity not tell us peptide abundance directly?

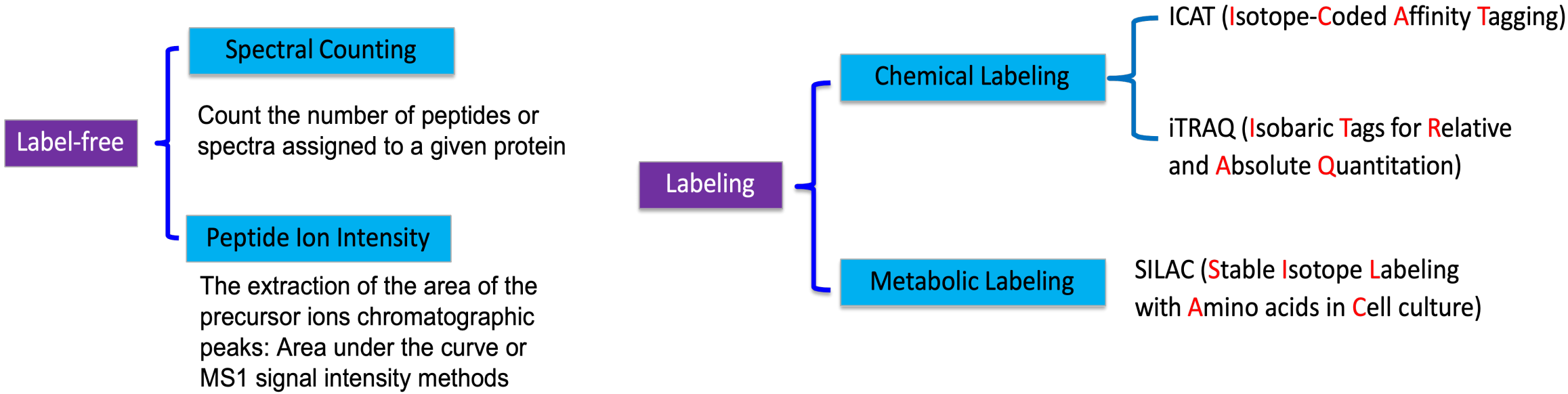
- There is a poor correlation between the amount of a peptide and the MS intensity in a single MS spectrum



- Many factors affect the MS intensity:
  - Peptide concentration
  - Day to day and long-term instrument reproducibility
  - Digestion efficiency
  - Recovery during sample preparation
  - Ionization efficiency
  - Instantaneous matrix effects

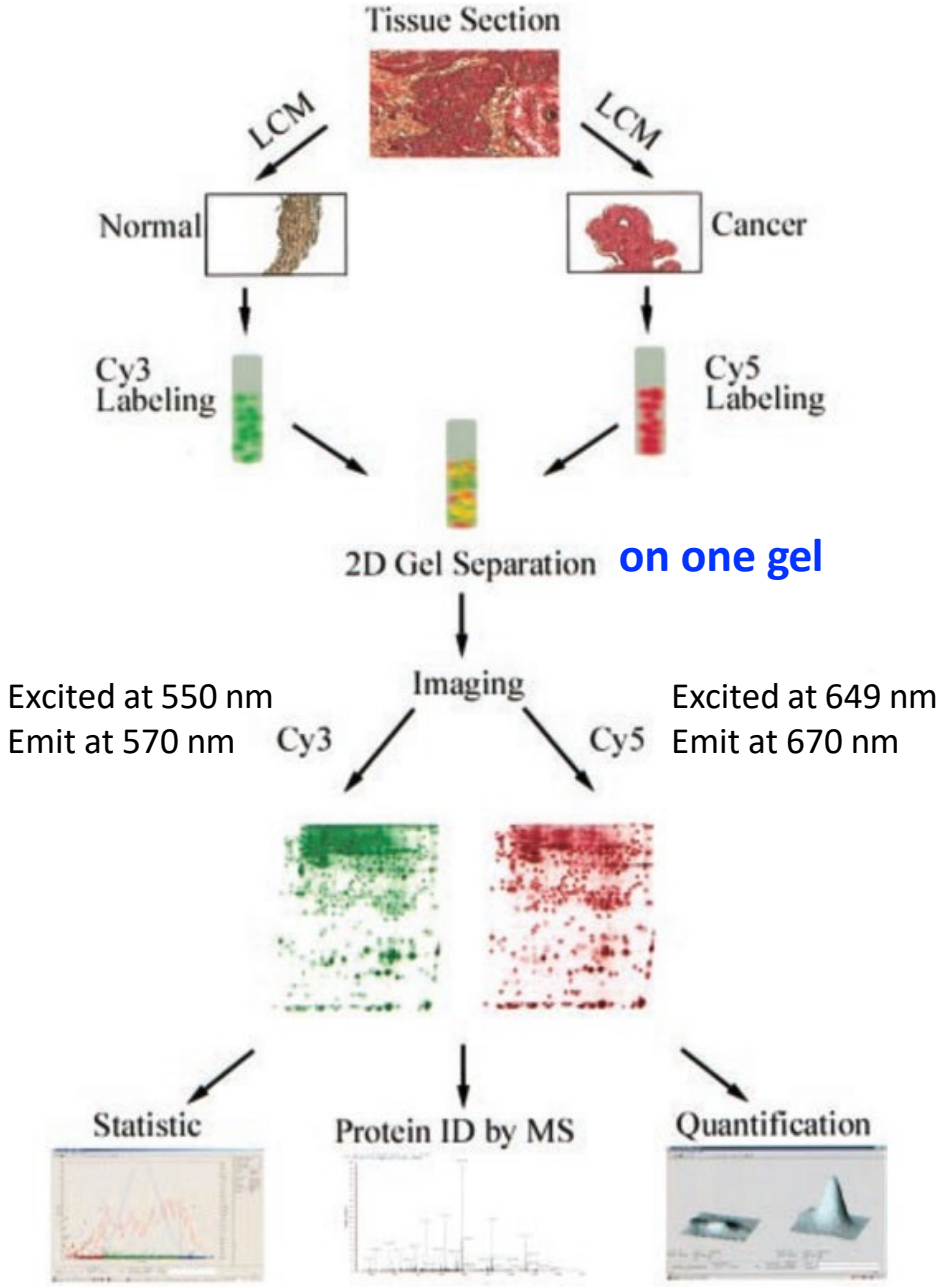
Quantitation Techniques in Proteomics

- Electrophoretic techniques
- Mass spectrometric techniques



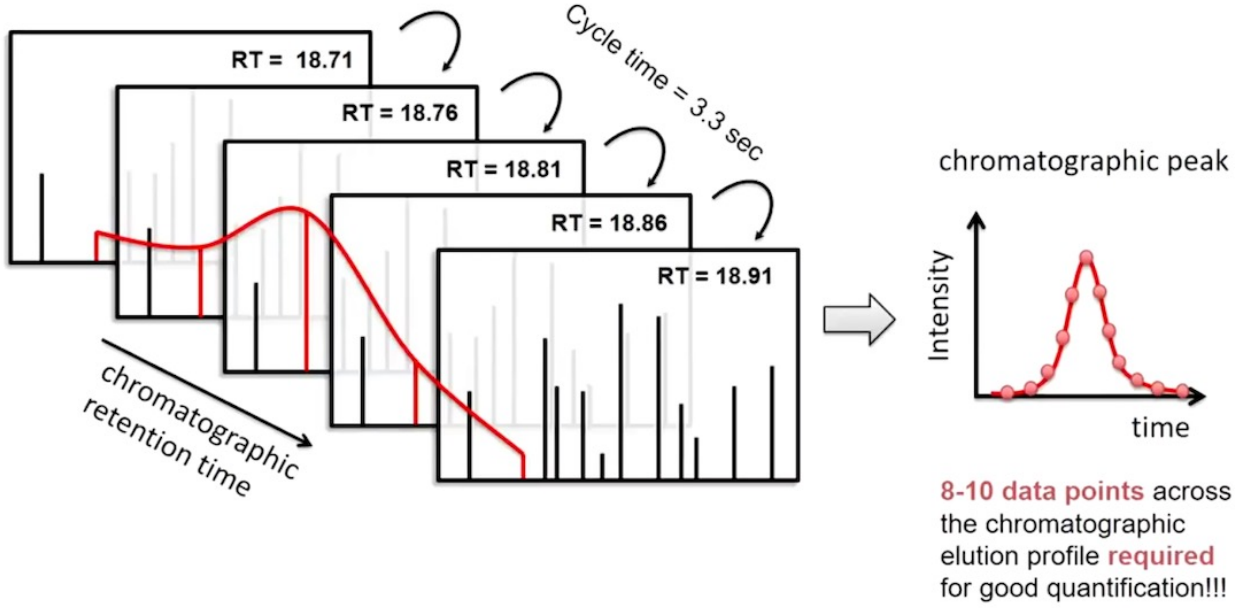
Electrophoretic Technique

- It is often difficult to obtain reproducible separations by standard 2D PAGE.
- Quantitative proteomics utilizes fluorescent dyes
  - Ease of use
  - High sensitivity (1 ng)
  - Low background (not staining the gel)
  - Compatibility with MS
  - Broad range of linearity
- Differential gel electrophoresis (DIGE) is used in quantitative proteomics.



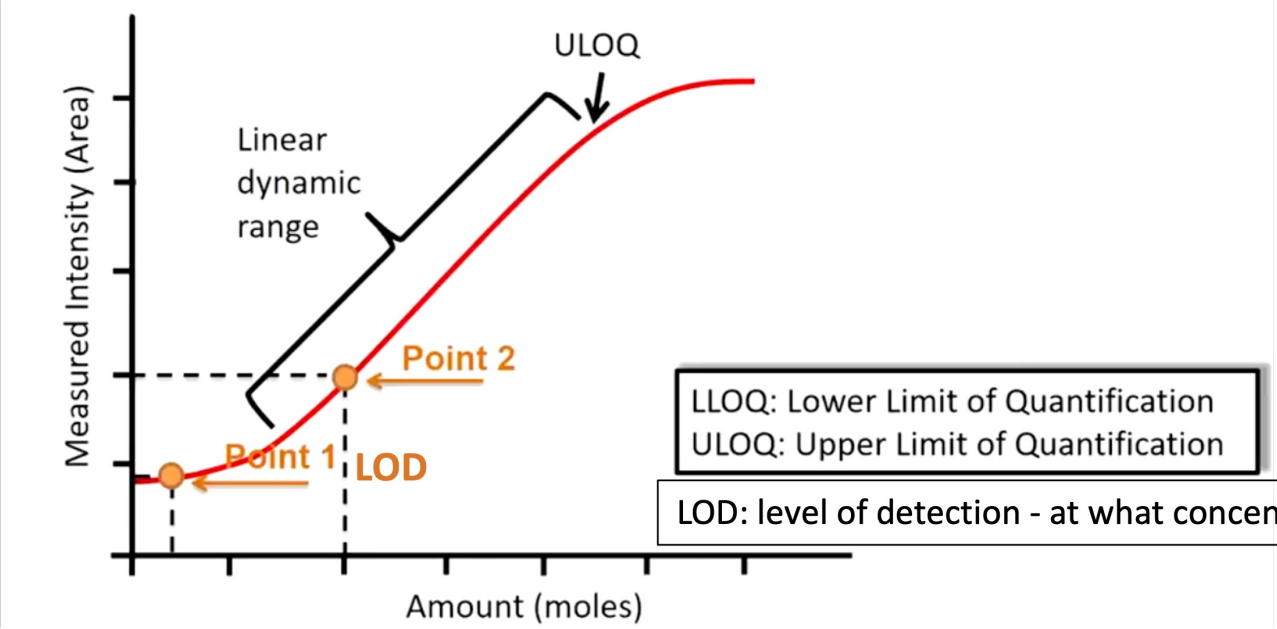
Quantitative LC-MS

- Fixed volume of the sample is injected
- Analyte spreads out, elutes over a certain timespan from the column (peaks)
- Only a fraction of the analyte really enters the MS



Cycle Time: A Crucial Parameter In Chromatography-based Quantification

Mass Spectrometry-based Quantification

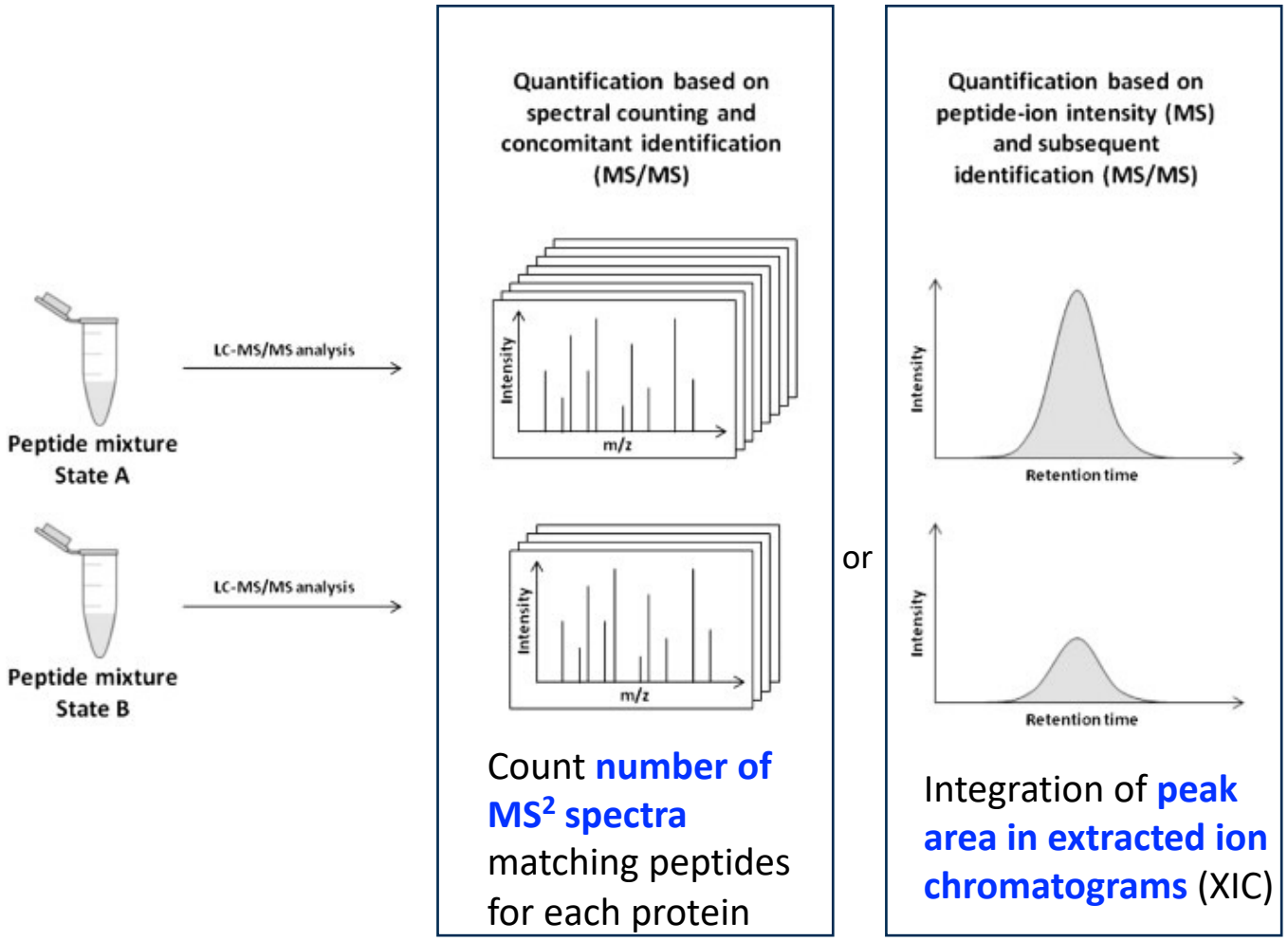


Chromatographic MS Signal Intensity Scales Linear With Peptide Abundance

- Accurate quantitative results can only be achieved when working within the linear dynamic range of every given peptide, respectively.
- The linear dynamic range and LLOQ and ULOQ are peptide and MS dependent

Label-free Methods: Spectral Counting or Peak Area

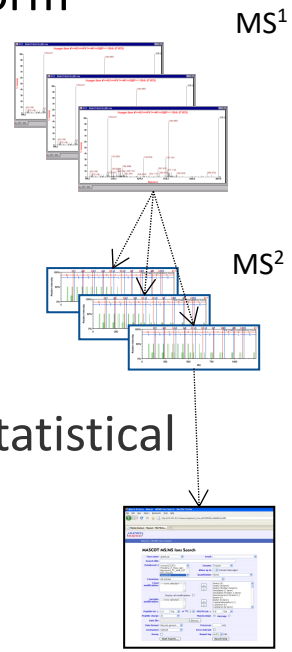
- Each sample is separately prepared and then subjected to individual LC-MS/MS runs.
- **Spectral Counting**: more abundant peptides are more likely to be observed and detected.
- **Peak Area**: signal intensities of ions after ESI correlate with ion concentrations.



https://ars.els-cdn.com/content/image/1-s2.0-S1570963913001556-gr1.jpg

Spectral Counting

- **Method summary:**
  - MS^2 spectra queried against a sequence database to make peptide-spectrum matches (PSMs)
  - Peptide ions from more abundant proteins trigger many more MS^2 scans
  - Count PSMs matched to a given protein => approximate protein abundance
- **Requirements:**
  - Not much – system setup for LC-MS/MS (most tandem mass spectrometers can perform spectral counting)
- **Advantages:**
  - It's really easy to do!
- **Disadvantages:**
  - Unclear how accurate the protein abundance values are. Normalization and careful statistical evaluation are still needed.
    - Origin of MS2 spectra…
    - proteins with only a few observable peptides
    - the quantitative changes between experiments are small
  - Need lots of replicates to get statistical significance (how many…?)
  - Lots of machine time needed



The Origin of Spectra Count

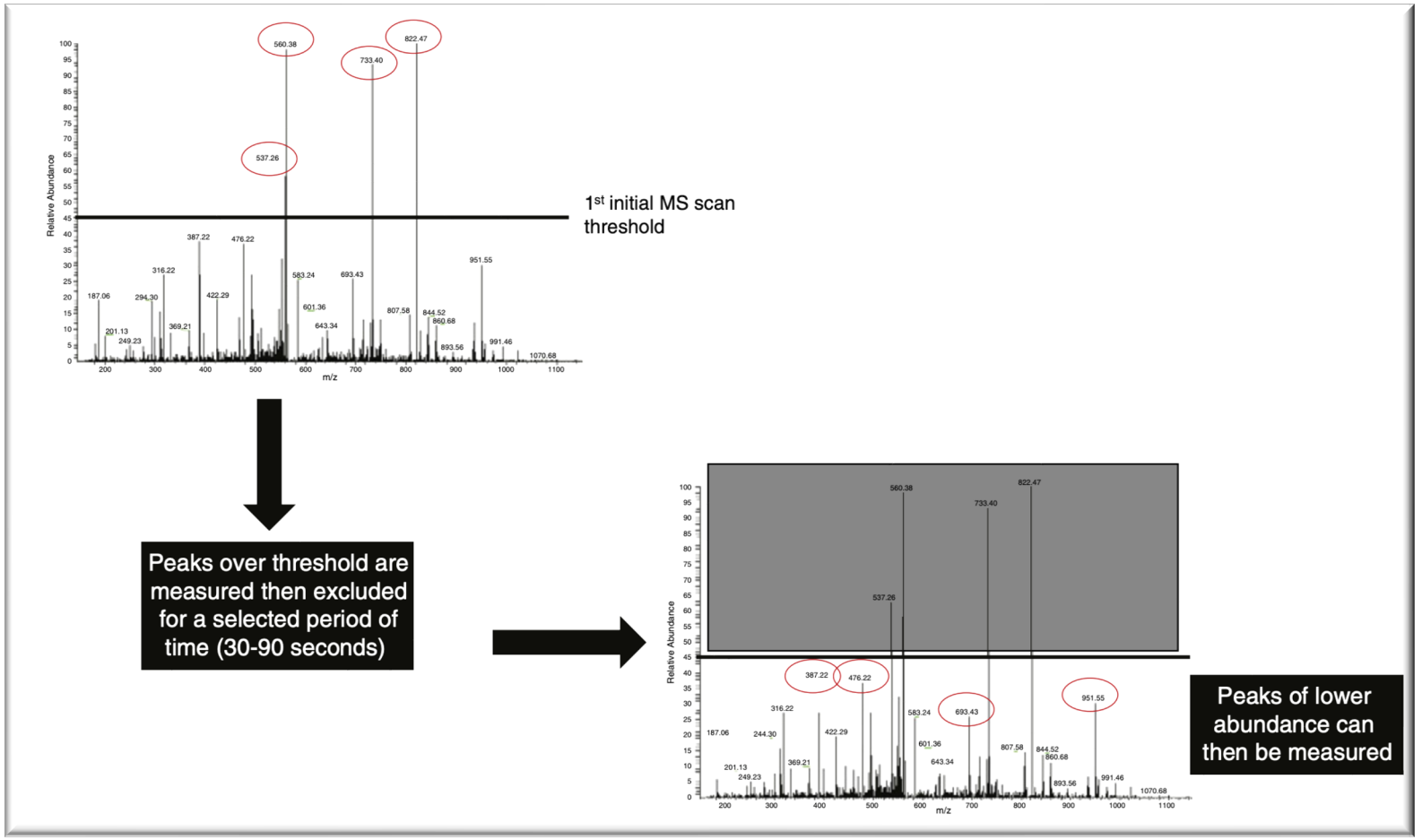
- Fully digested tryptic peptides can be positively charged at both *N*- and *C*-terminus upon electrospray ionization. Some tryptic peptides, containing histidine or extra arginine/lysine due to missed cleavage, can be 3^+ charged.
- The MS^2 spectra that matched to a protein is a combination of:
  - spectra from different partial tryptic peptides and full tryptic peptides,
  - spectra from the same peptide with different charges,
  - spectra of the same peptide with variable modifications,
  - repeated spectra from the same peptide due to expired dynamic exclusion.

Note: At some situations, spectra that are potentially matched to partial tryptic peptides or peptides with PTMs are not included during data analysis by spectral interpretation software.

Dynamic exclusion is a software technique, allowing the mass spectrometer to more efficiently identify peptides in a sample.

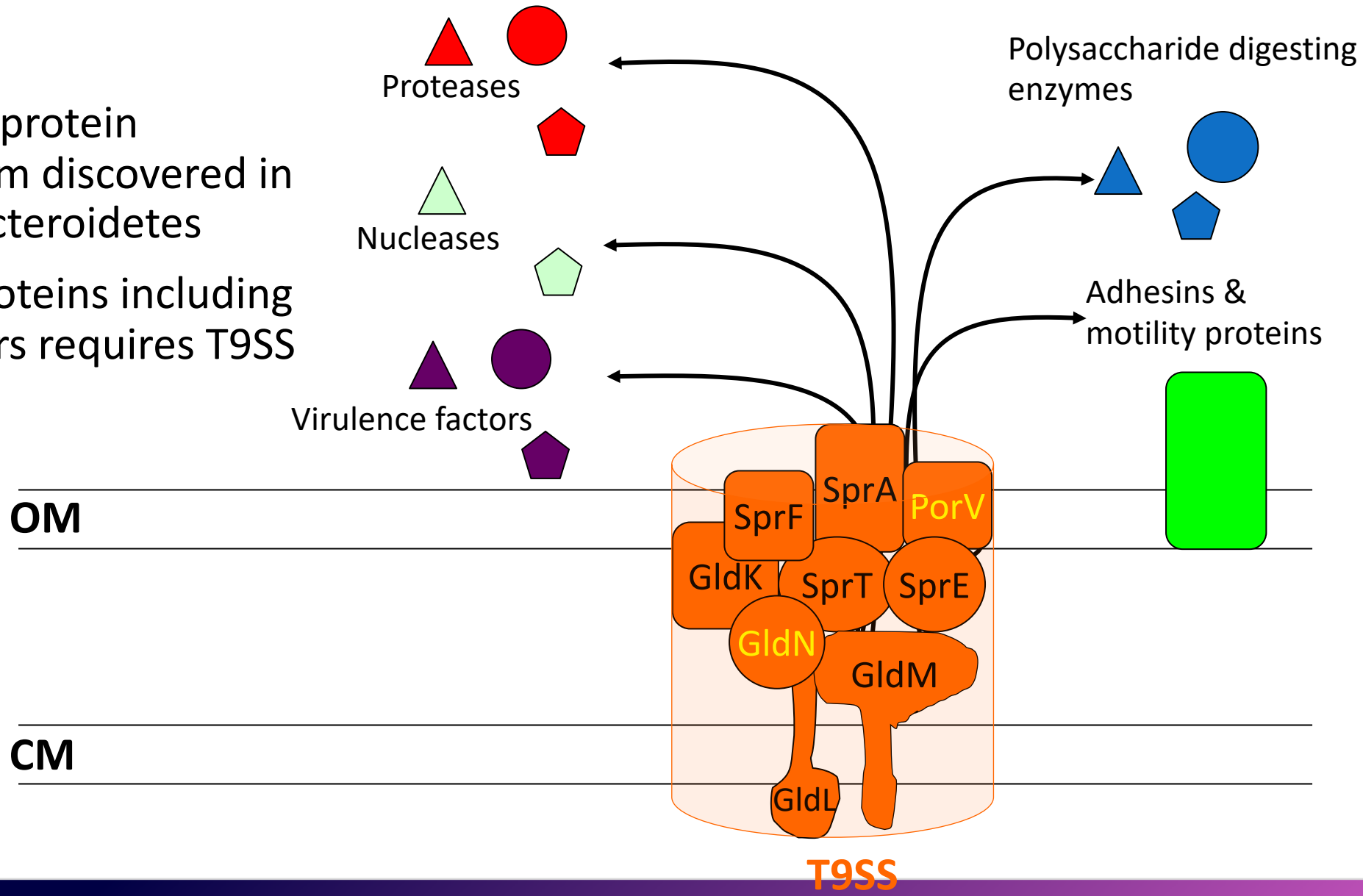
The first scan measures the ions with the highest intensity (most abundant). These masses are added to a temporary ‘exclusion’ list for a period

Once the high intensity peaks have been sequenced and excluded the MS can measure peaks under the threshold, thereby detecting less abundant peptides.



Case Study: Type IX Secretion System (T9SS)

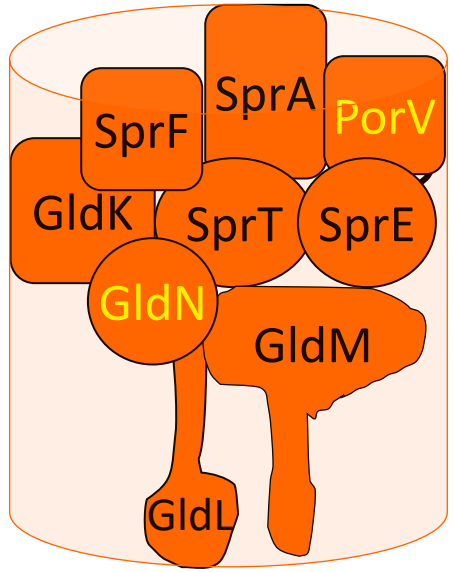
- T9SS is a novel protein secretion system discovered in the phylum Bacteroidetes
- Secretion of proteins including virulence factors requires T9SS



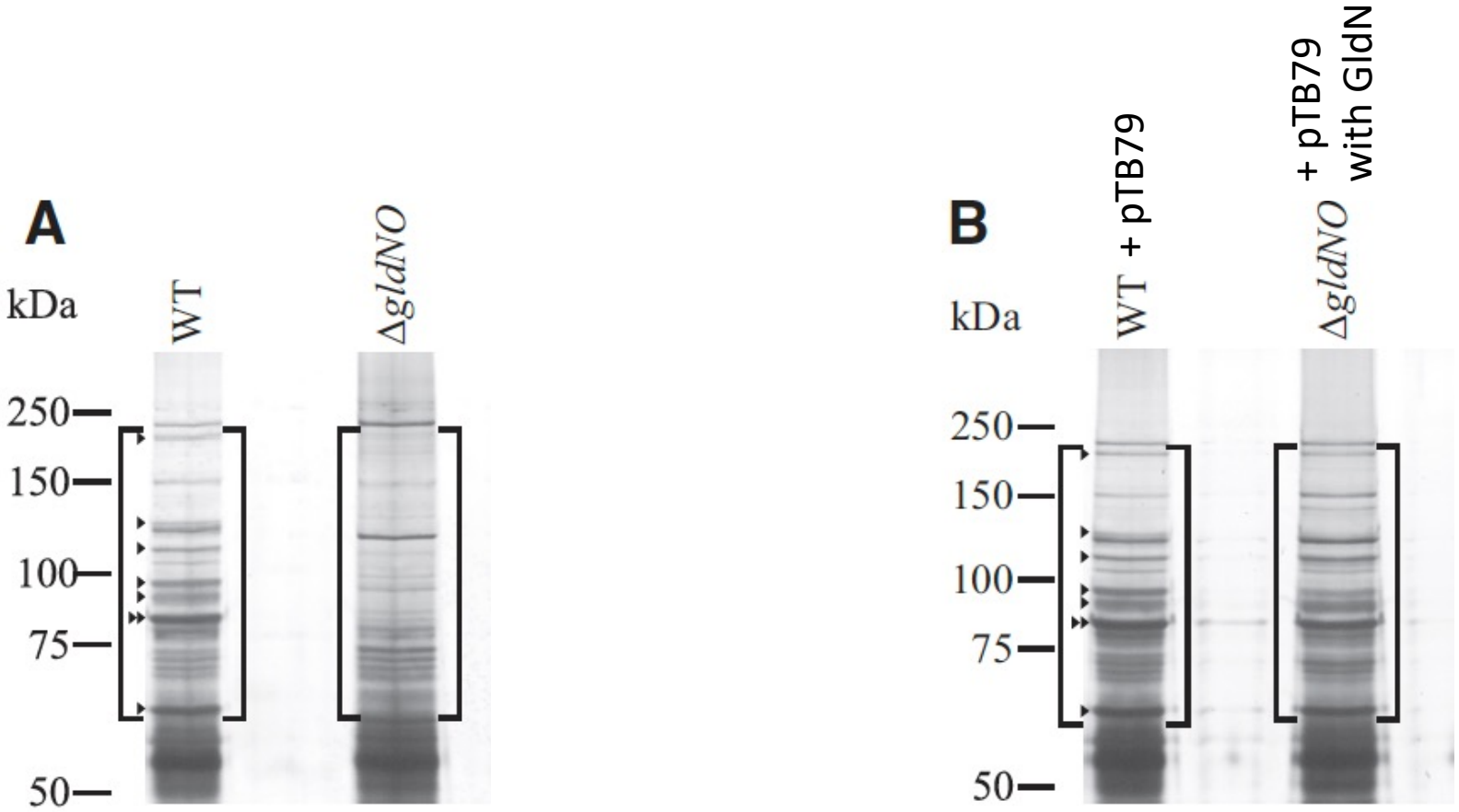
Purpose and Strategy

- **Purpose**: to identify which proteins are secreted by the bacterial type IX secretion system (T9SS) and their abundance

- **Strategy**: Create T9SS mutants (Δ*gldNO* and Δ*porV*) and complementation → Cell culturing → SDS-PAGE of secreted extracellular proteins → Enzymatic in-gel digestion → LC-MS/MS



Soluble extracellular proteins of wild-type and mutant cells



Kharade, S. S., & McBride, M. J. (2015). *197*(1), 147–158. https://doi.org/10.1128/JB.02085-14

TABLE 5 Candidate proteins secreted by the T9SS identified by LC-MS/MS analysis of cell-free culture fluid

Locus tag/protein name	Mol mass[b] (kDa)	Predicted localization[c]	CTD[d]	Predicted protein function[e]	Spectrum count for:		
					Wild type	Δ*gldNO* strain	Δ*gldNO* strain with pTB79
Fjoh_0074	123.1	OM, E	TIGR04183	Nuclease/phosphatase	42	3	108
Fjoh_0601	208.2	OM			115	0	84
Fjoh_0602	279.3	OM			68	0	38
Fjoh_0604	144.2	E			47	0	39
Fjoh_0606	409.5	OM			163	0	172
Fjoh_0808/RemA	154.0	E	TIGR04183	Motility adhesin	38	0	47
Fjoh_0886	99.1	E	TIGR04183	Peptidase	12	0	19
Fjoh_1022	51.1	E	TIGR04183	Licheninase	6	0	6
Fjoh_1123	121.9	E, OM	TIGR04131		34	0	10
Fjoh_1188	152.7	E, OM	TIGR04183		49	0	104
Fjoh_1189	181.4	E	TIGR04183	Lectin	74	0	112
Fjoh_1208	112.5	E	TIGR04183	α-Amylase	45	0	66
Fjoh_1231	97.8	E	TIGR04183	Pectate lyase	9	0	13
Fjoh_1269	94.3	E, OM	TIGR04183		27	4	43
Fjoh_1408[f]	106.0	E	TIGR04183	α-Amylase	2	0	4
Fjoh_1645[f]	258.1	E	TIGR04131		2	0	6
Fjoh_2150	39.0	E, OM	TIGR04183		6	0	6
Fjoh_2273	93.3	E	TIGR04131		4	0	5
Fjoh_2389[f]	57.7	E, OM	TIGR04183	Peptidase	2	0	7
Fjoh_2667	129.7	OM			28	0	7
Fjoh_2687	155.8	E			26	1	26
Fjoh_3108	30.9	OM, E, P			7	0	10
Fjoh_3246	299.4	OM, E	TIGR04183		12	0	77
Fjoh_3324	105.3	E	TIGR04183	Carbohydrate binding	16	1	40
Fjoh_3729	195.1	OM			46	0	32
Fjoh_3777	128.1	OM, E	TIGR04183	Deacylase	10	0	25
Fjoh_3952	330.6	E	TIGR04131		22	0	11
Fjoh_4174	102.5	E	TIGR04183	Carbohydrate binding	40	5	40
Fjoh_4176	95.4	E	TIGR04183	Carbohydrate binding	48	3	65
Fjoh_4177	144.9	E	TIGR04183	Glycoside hydrolase	22	0	35
Fjoh_4750	158.1	E	TIGR04131		13	0	3
Fjoh_4819	112.5	C, OM, P		Glycoside hydrolase	34	0	5
Fjoh_4934	84.8	E	TIGR04131		11	1	7

Kharade, S. S., & McBride, M. J. (2015). *197*(1), 147–158. https://doi.org/10.1128/JB.02085-14

Peak Area/Peptide Ion Intensity Based Protein Quantification

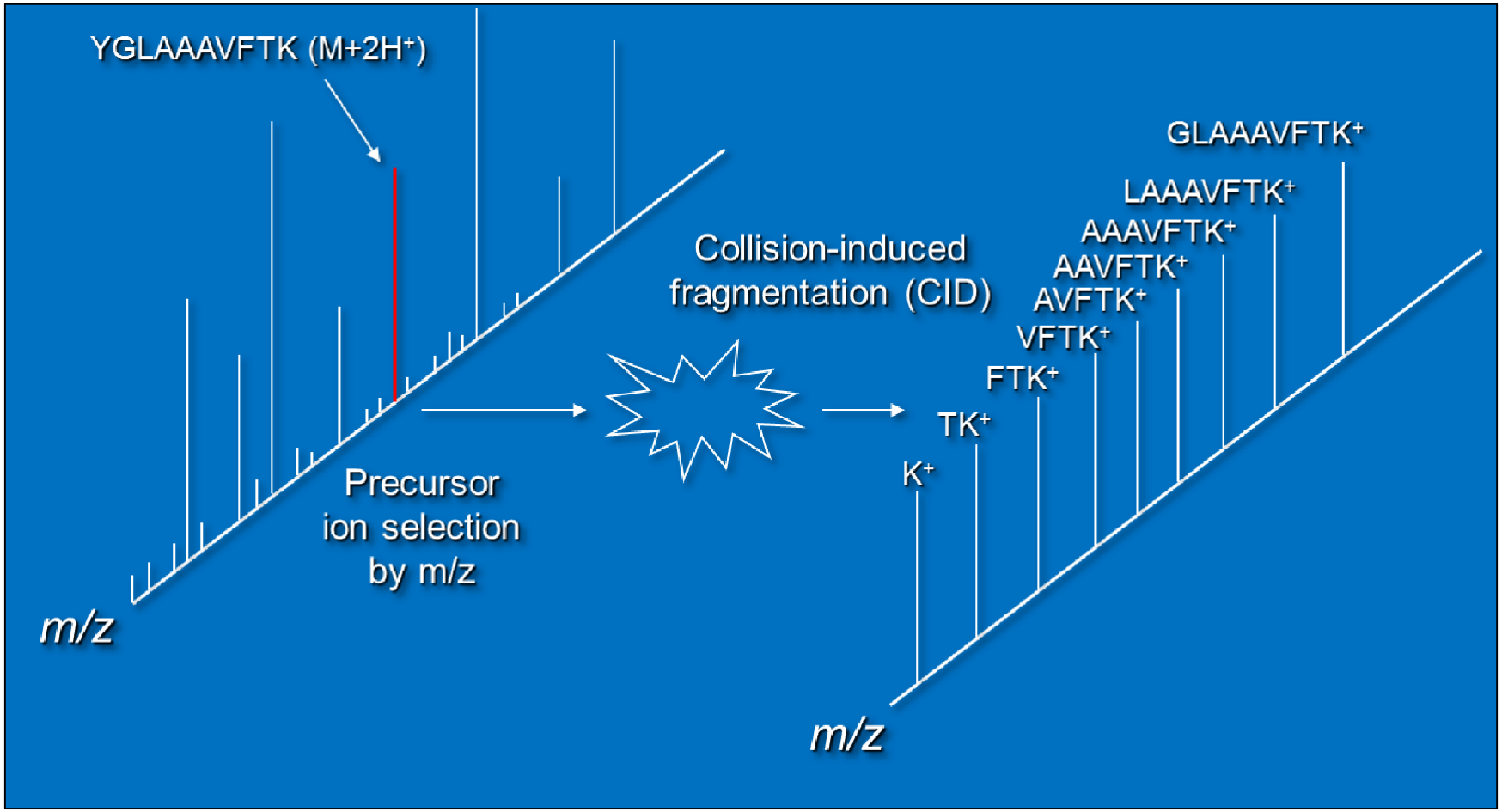
- MS signal intensity for the peptide at a certain time is proportional to the concentration eluting off the column
- The area under the chromatographic peak is proportional to the total amount of analyte eluting and thus to the amount in the sample. Hence, we want to integrate over time.

- **Method summary:**
  - Compare **MS^1 peptide ion abundance** across runs;
  - Specialist software must align different parallel LC-MS runs
  - Calculate ratios from aligned MS1 data

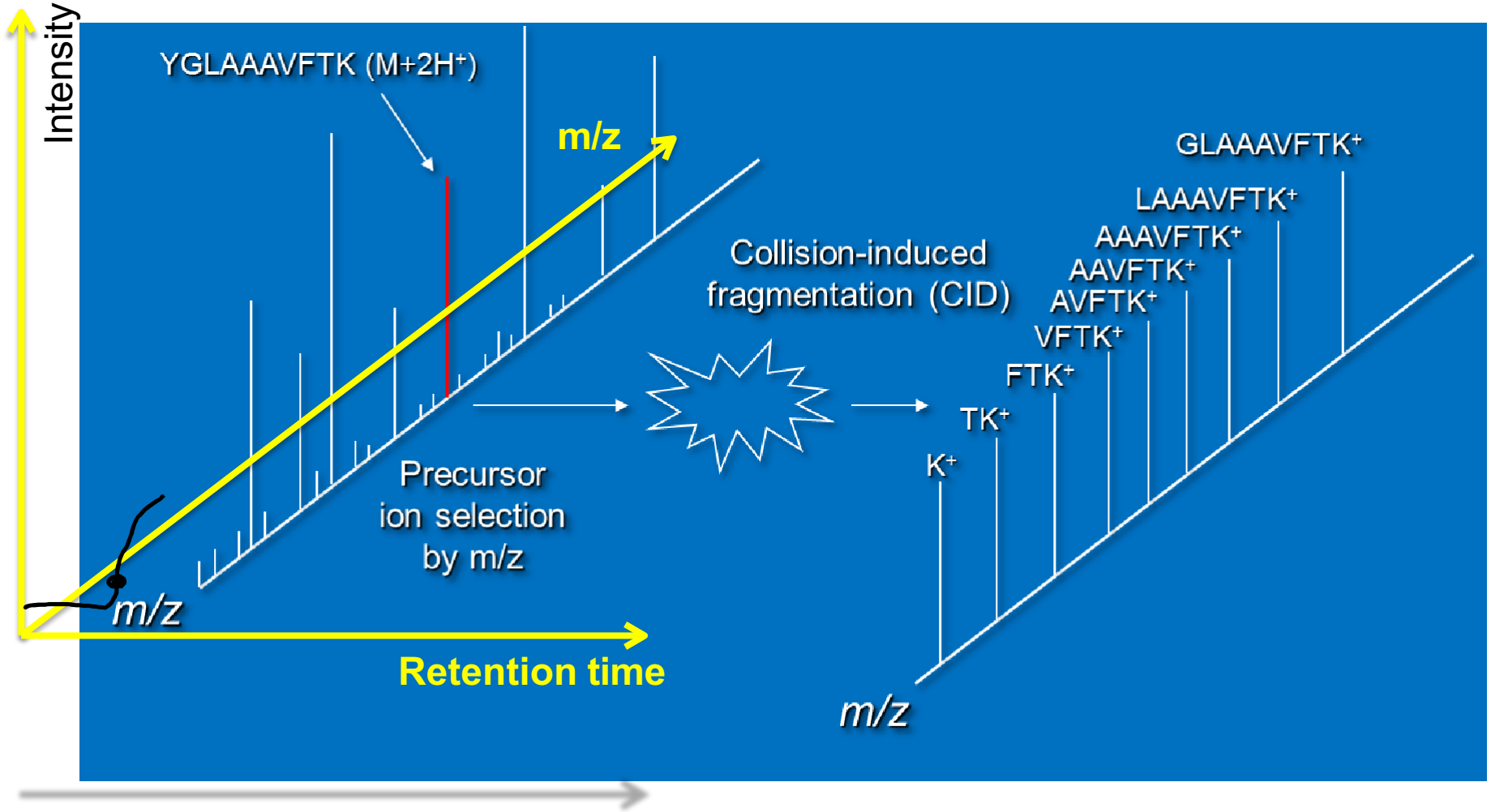
Peak Area/ Peptide Ion Intensity Based Protein Quantification

- **Advantages:**
  - In theory, should be more accurate than spectral counting – uses real intensity data
  - No complicated labelling protocols
- **Disadvantages:**
  - Data processing is fairly CPU intensive
  - Only works well if experimental system has high technical and biological reproducibility
- **Requirements:**
  - Very good LC delivery system (must be reproducible)
  - High resolution mass spec (Orbitrap etc...)
  - Good PC for running software

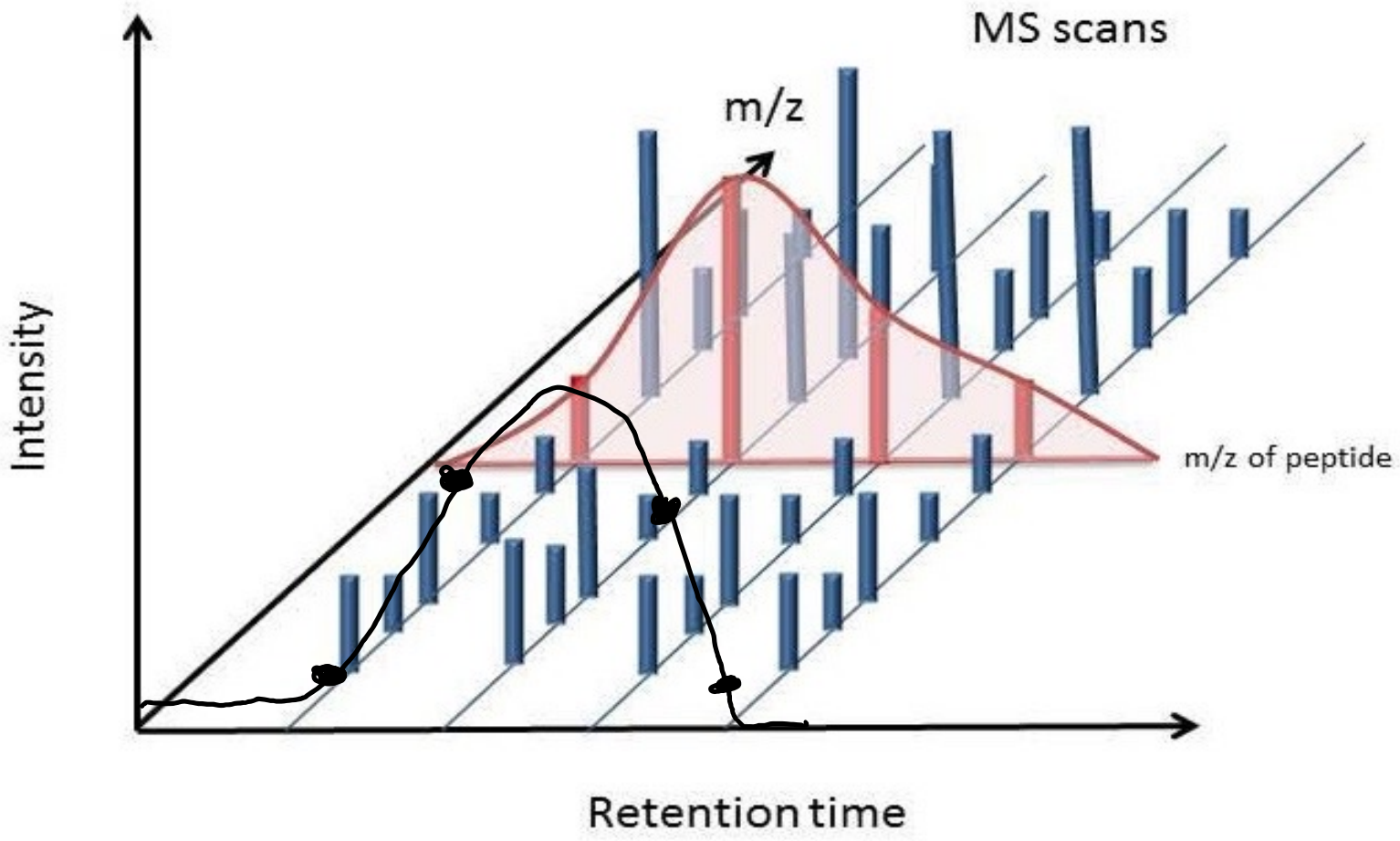
Acquired Data



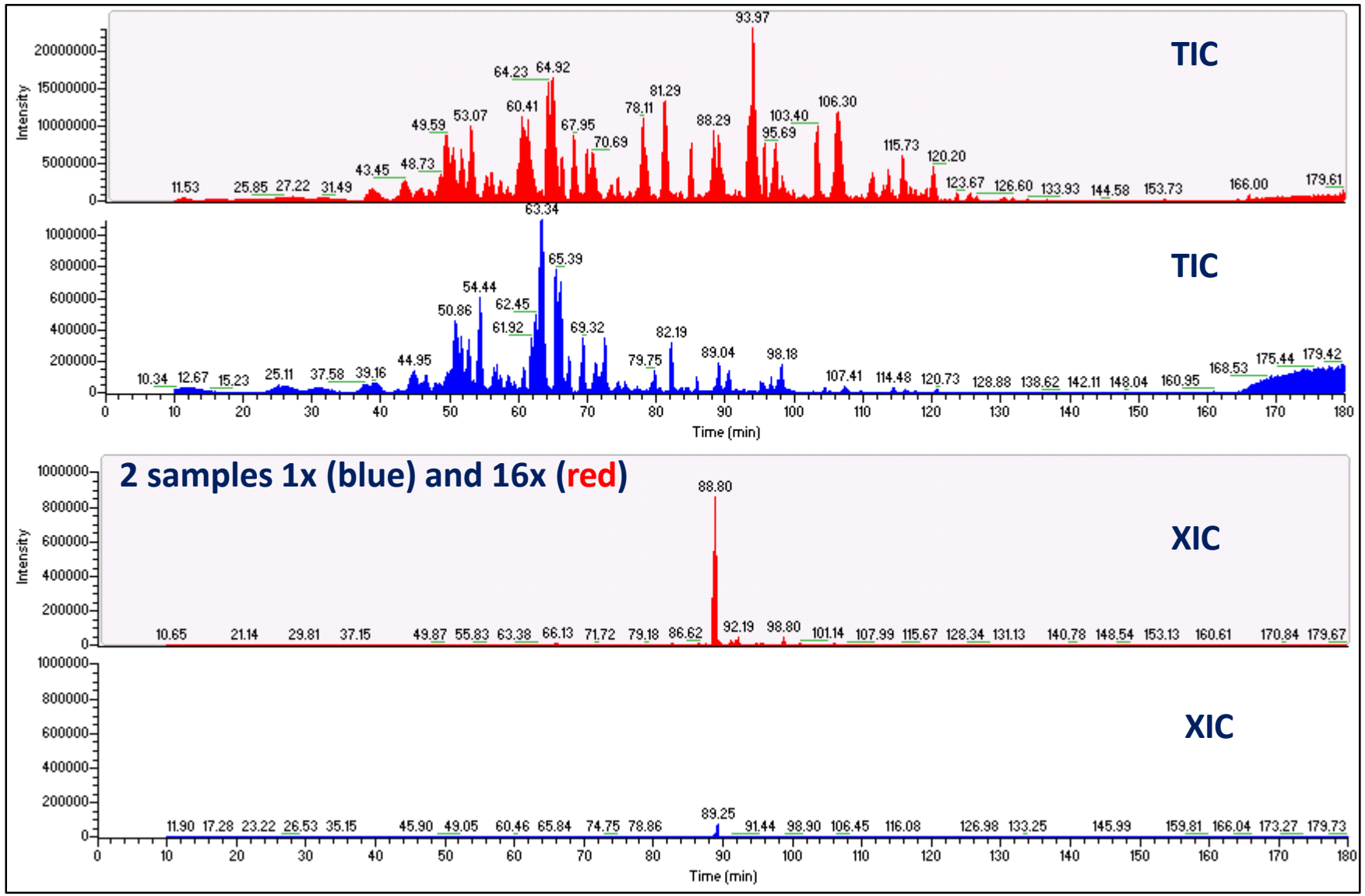
Acquired Data



Acquired Data ?



Peptide Ion Intensity



The total ion current (TIC) chromatogram represents the summed intensity across the entire range of masses being detected at every point in the analysis

In an **extracted-ion chromatogram (XIC** or EIC), also called a reconstructed-ion chromatogram (RIC), one or more m/z values representing one or more analytes of interest are extracted from the entire data set for a chromatographic run.

LC-MS Chromatographic Alignment

- To extract the peptide peak area, two basic parameters, m/z and retention time, must be determined.
- Typically, the m/z value is measured reproducibly in low resolution mass spectrometers such as the LTQ linear ion trap MS and extremely reproducibly in high resolution mass spectrometers such as LTQ-Orbitrap.
- Retention time of peptides can shift between experiments
- LC-MS maps can contain millions of peaks
- In label-free quantification, maps thus need to be aligned in order to find the corresponding peaks

What should be considered?

- **Fact: a single chromatographic condition, e.g., one specific column with a specific mobile phase and gradient, will not be optimal for each of the thousands of peptides in a single injection of a complex sample.**

- **Due to the fact, many peptides cannot be used for protein quantification with a single run.**

Protein Abundance Calculation

$$Ap = \sum_{i=0}^{n} (Ip/Fp)i$$

Ap = protein abundance, *Ip* = peptide intensity, and *Fp* = frequency of peptide sharing.

For a peptide shared by different proteins, the intensity of this peptide (Ip) was divided by sharing frequency (Fp). The aim of this strategy is to decrease the impact of shared peptides.

Normalization

- The aim of normalization is to remove systematic bias.
- Numerous normalization algorithms have been developed and applied in biological studies.
- Global normalization (central tendency), linear regression, local regression, and quantile techniques are the commonly used.

- In current LC-MS technology, no ideal normalization techniques exist.
- Using inappropriate or even flawed normalization will not improve the analysis and may introduce additional errors, thus it is better that no normalization is applied.
- However, filtration of unquantifiable peptides is absolutely necessary for an accurate analysis.

Label-free Methods: spectral counting or peak area

Pros:

- Simple workflow

  - No complicated (or expensive) labelling or tagging protocol

- Whole proteome analysis
- Comparison of multiple states (relative quantification)

Cons:

- Still can be expensive
  - Need plenty of replicates to get statistical power (machine time!)
- Reproducible sample prep, chromatography and MS performance is critical for this approach
- Not straightforward to validate results from big data sets
- Low abundance proteins hard to measure accurately

Applications of label-free Quantitative methods

- Identifying expression profiles in different biological processes
- Diagnosing certain diseases and cancer biomarkers
- Monitoring changes in certain biological process proteomes
- Studying protein interaction networks

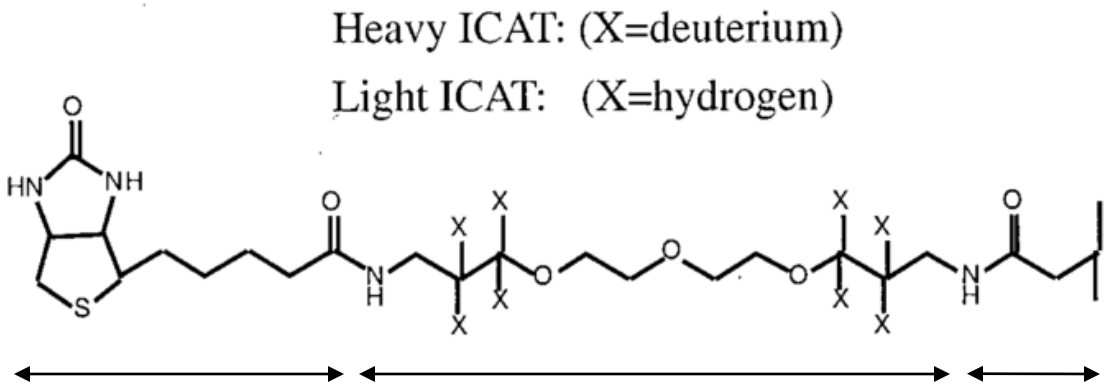
Stable-isotope Labeled Methods

- Provide a useful means of determining the relative expression level of individual proteins between samples with high precision (coefficients of variation less than 10%).

- Because two or more samples tagged with different numbers of stable isotopes can be mixed before any processing steps, sample-to-sample recovery differences are eliminated.

- Also allow post-translational modifications, splice variations and mutations (often unnoticed in immunoassays) to be detected and identified, increasing the clinical relevance of the assay and avoiding the issues of non-specific binding and cross-reactivity observed in immunoassays.

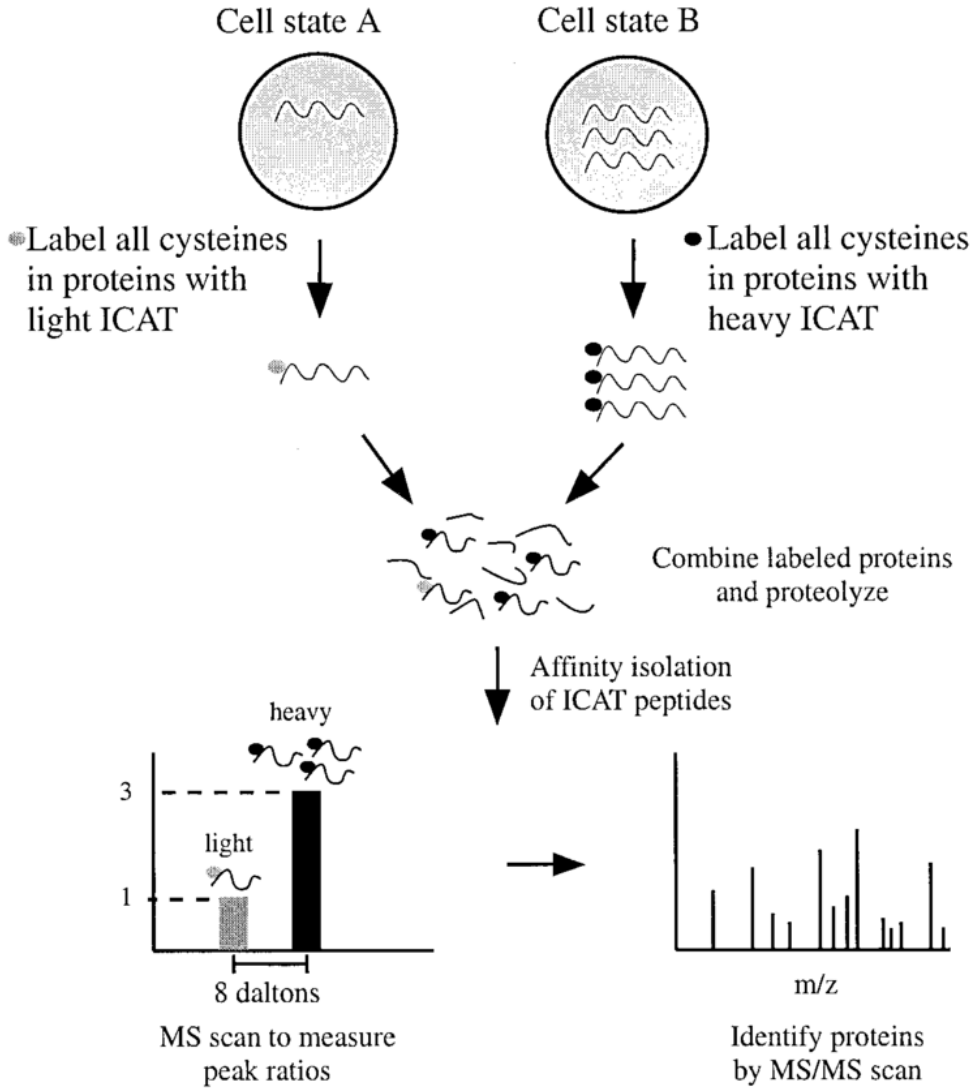
Labeling: ICAT (Isotope-Coded Affinity Tagging)



Biotin moiety, which permits isolation of the labeled molecules by affinity chromatography (interact with avidin immobilized)

Linker containing light (X= H) or heavy (X=D) isotopes

Iodoacetamide moiety, which binds to cysteine thiol group



Quantitate relative protein levels by measuring peak ratio

Identify peptide by sequence information

ICAT (Isotope-Coded Affinity Tagging)

Advantages

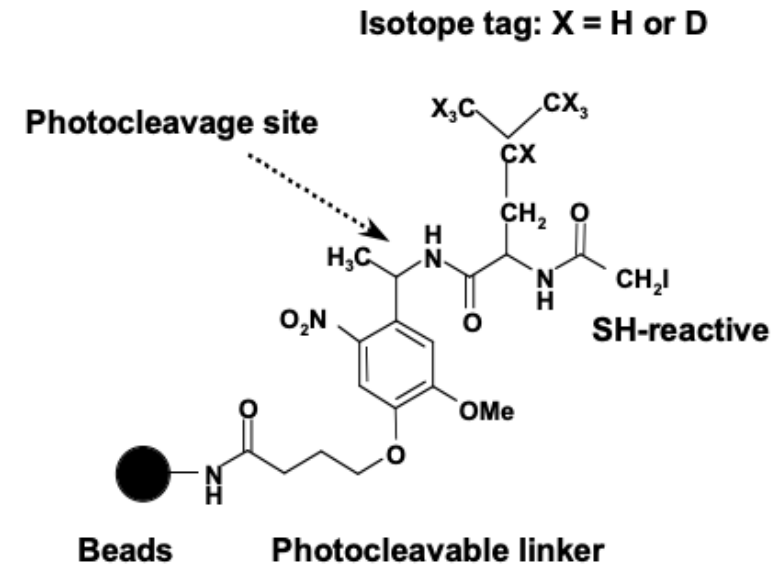
- High specificity
- High sensitivity
- Applicability to samples of different origin (cell, tissues, fluids)
- Effective labeling in the presence of guanidine, SDS, or urea

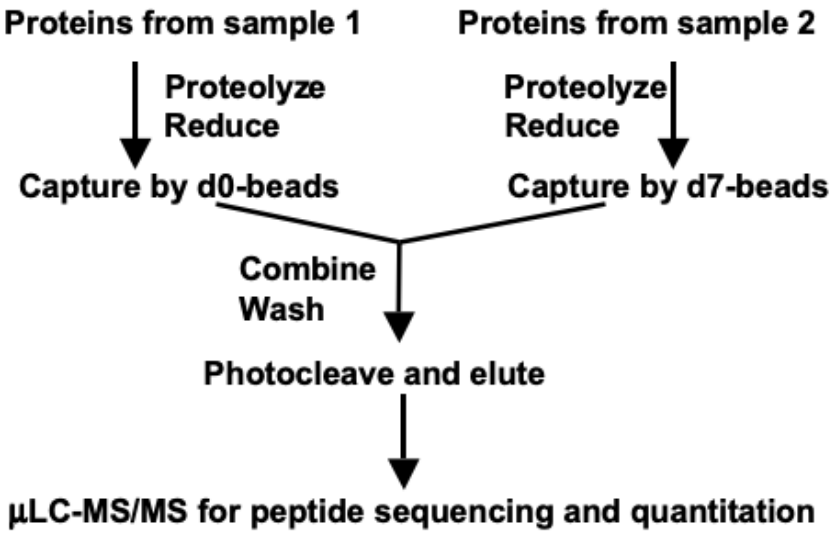
Disadvantages

- It doesn’t allow for quantification of proteins that do not contain Cys residues.

Solid-phase ICAT

- Faster and easier to conduct, because it doesn't require isolation of labeled peptides by chromatographic methods.
- Better, but it's expensive and light sensitive.

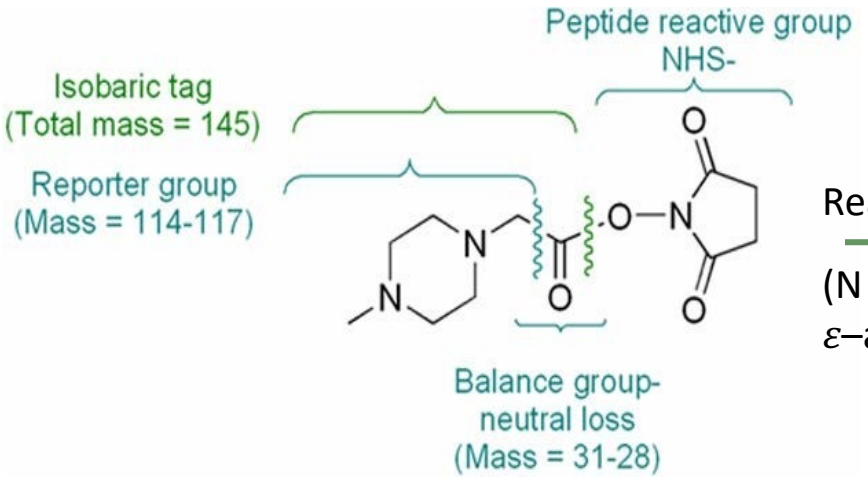




Zhou *et al.*, Nature Biotech. 2002, 19, 512

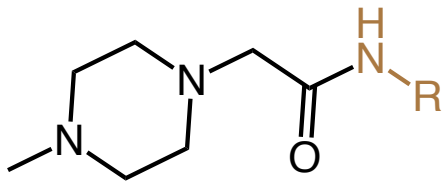
Labeling: iTRAQ (Isobaric Tags for Relative and Absolute Quantitation)

- iTRAQ permits simultaneous analysis of 2-8 samples.
  - TMT– 6-plex
  - iTRAQ-8 – 8-plex
- It is based on labeling of peptides with isobaric tags but produce different ions during fragmentation.

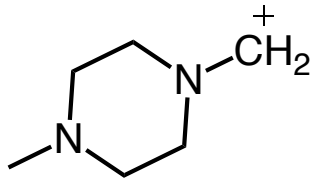


React with NH_2 groups

(N termini and
ε–amino groups of Lys)



MS/MS
fragmentation



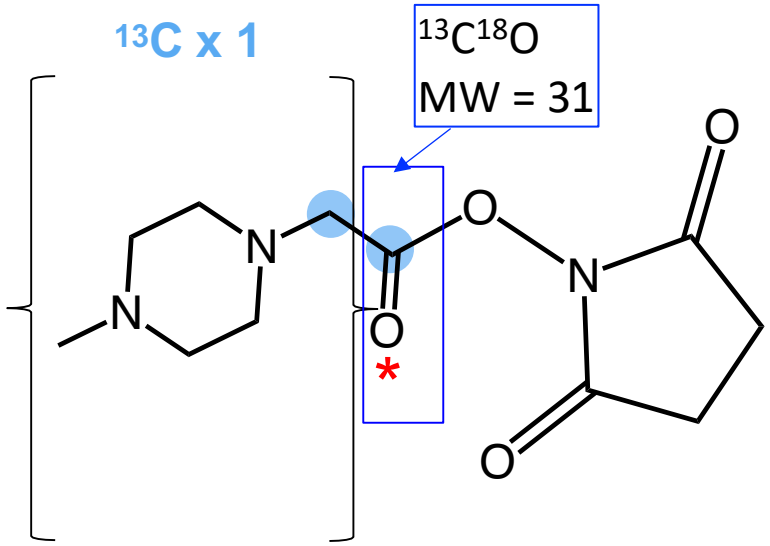
Reporter ion

+

Rest of molecule

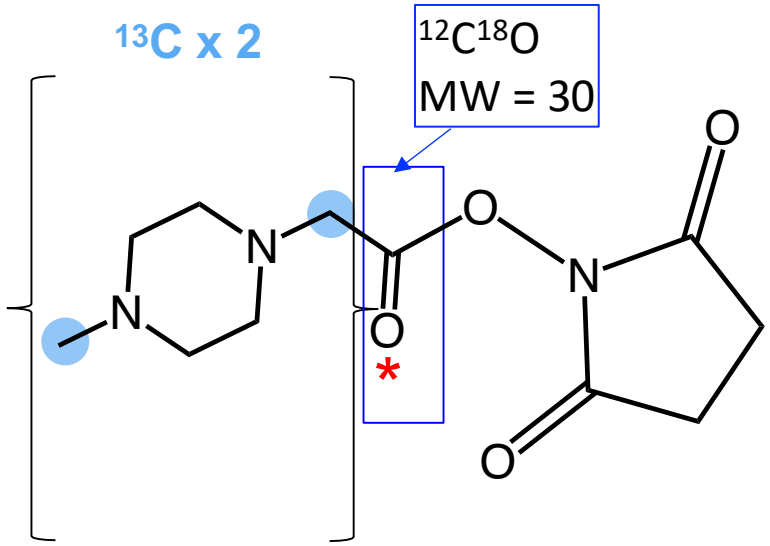
iTRAQ Reagents

^{13}C x 1



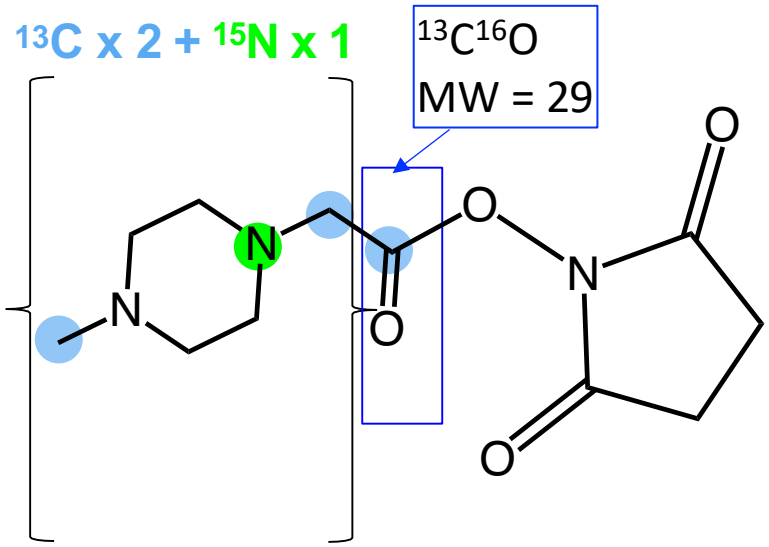
Produces an ion of MW = 114
after fragmentation

^{13}C x 2



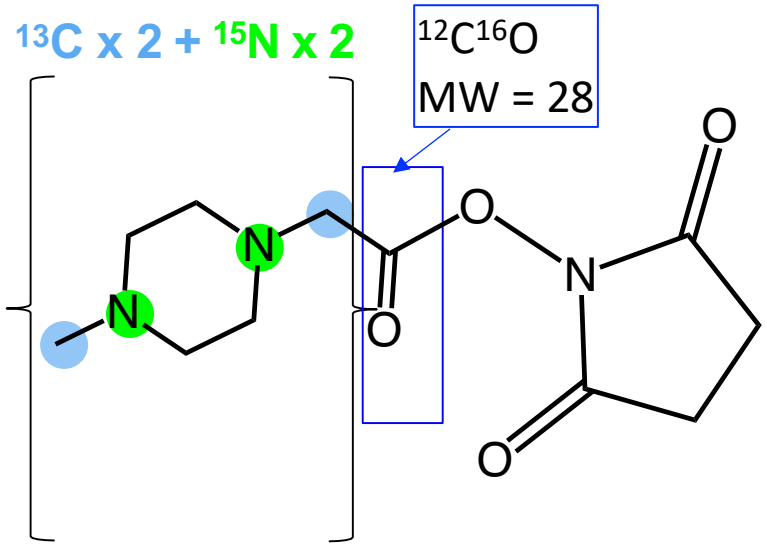
Produces an ion of MW = 115
after fragmentation

^{13}C x 2 + ^{15}N x 1



Produces an ion of MW = 116
after fragmentation

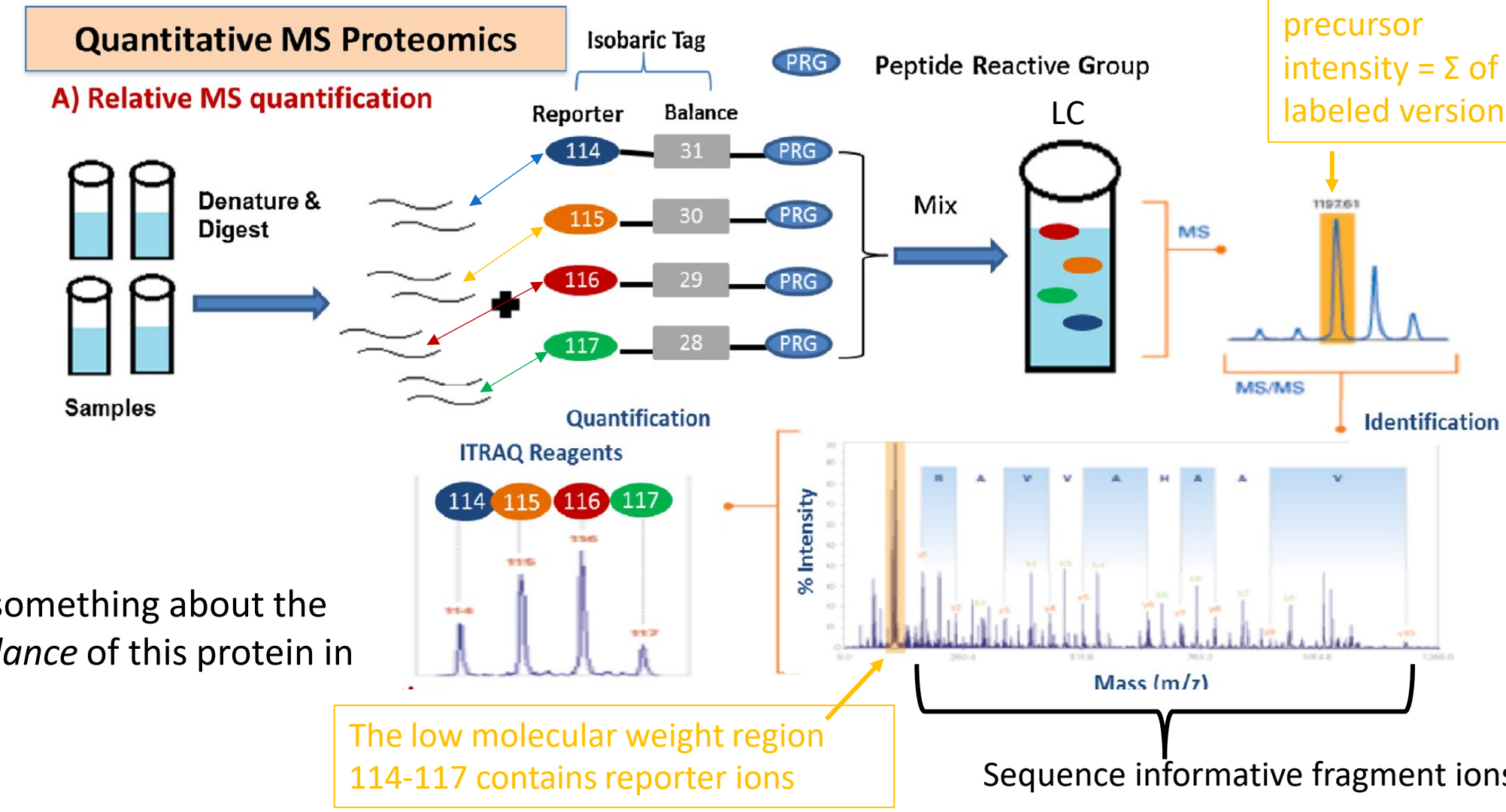
^{13}C x 2 + ^{15}N x 2



Produces an ion of MW = 117
after fragmentation

General Workflow with iTRAQ Method

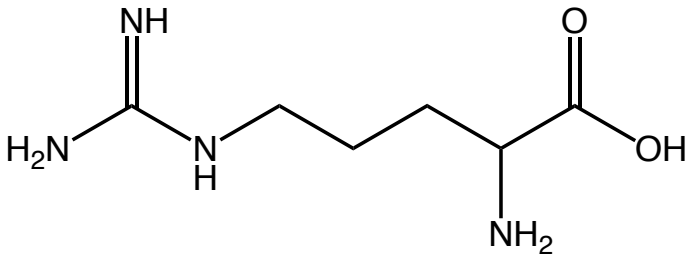
same peptide from 4 different samples: Observed precursor intensity = Σ of all labeled versions



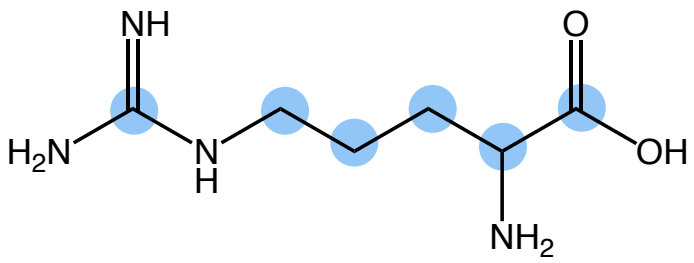
- **Ratio** tells us something about the *relative abundance* of this protein in the 4 samples

Labeling: SILAC (Stable Isotope Labeling with Amino Acids in Cell Culture)

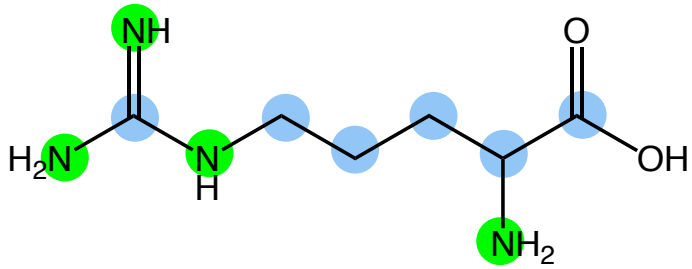
^{13}C ^{15}N



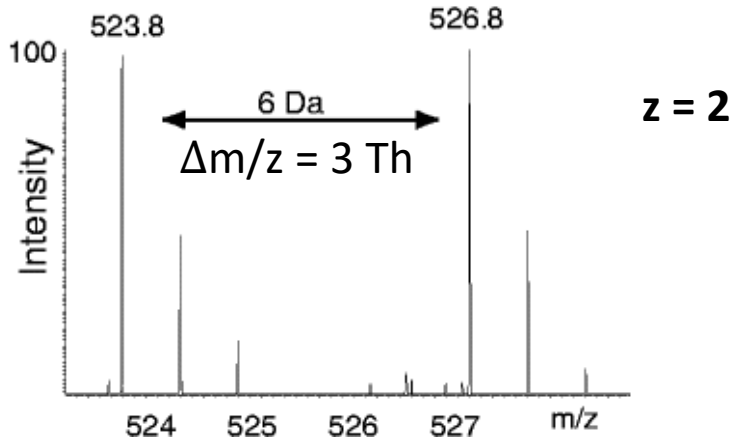
Arg_{+0} (Light)



Arg_{+6} (Medium)

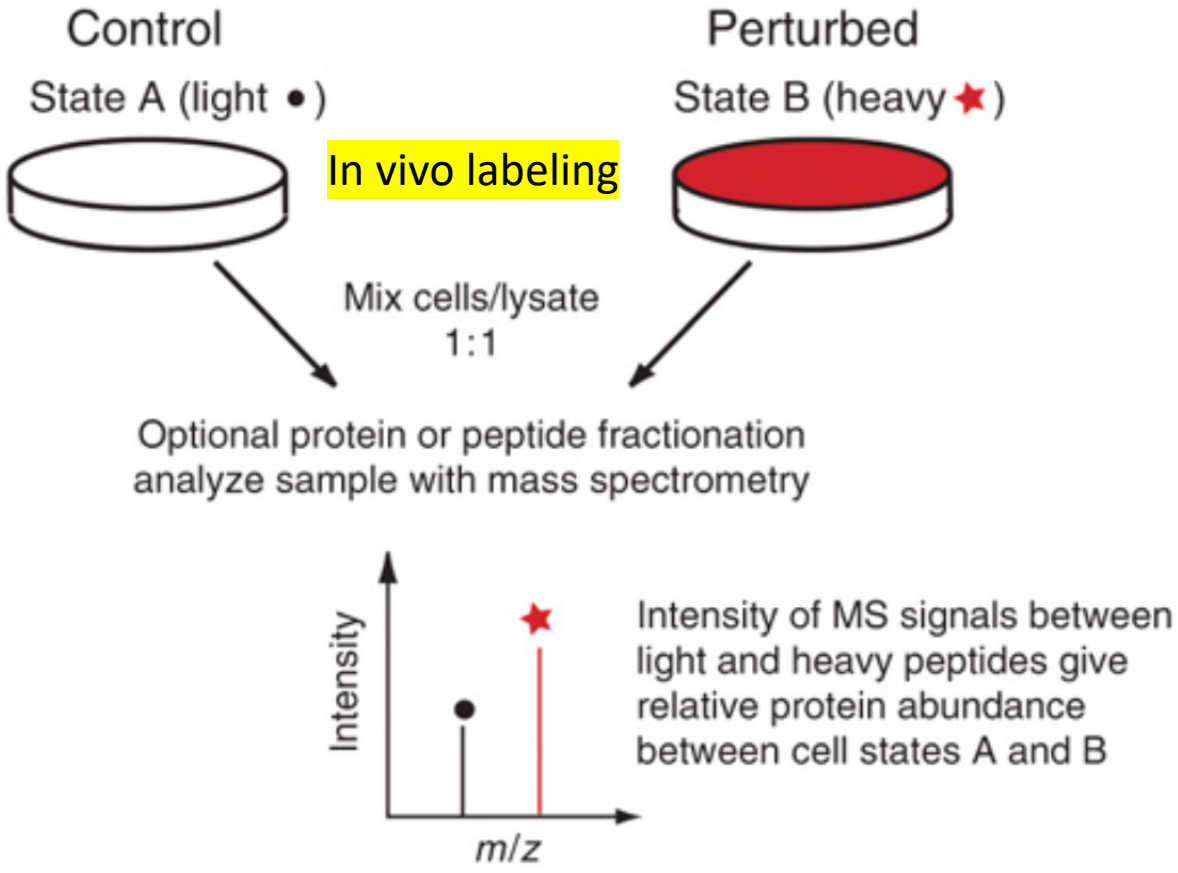


Arg_{+10} (Heavy)



z = 2

SILAC (Stable Isotope Labeling with Amino Acids in Cell Culture)

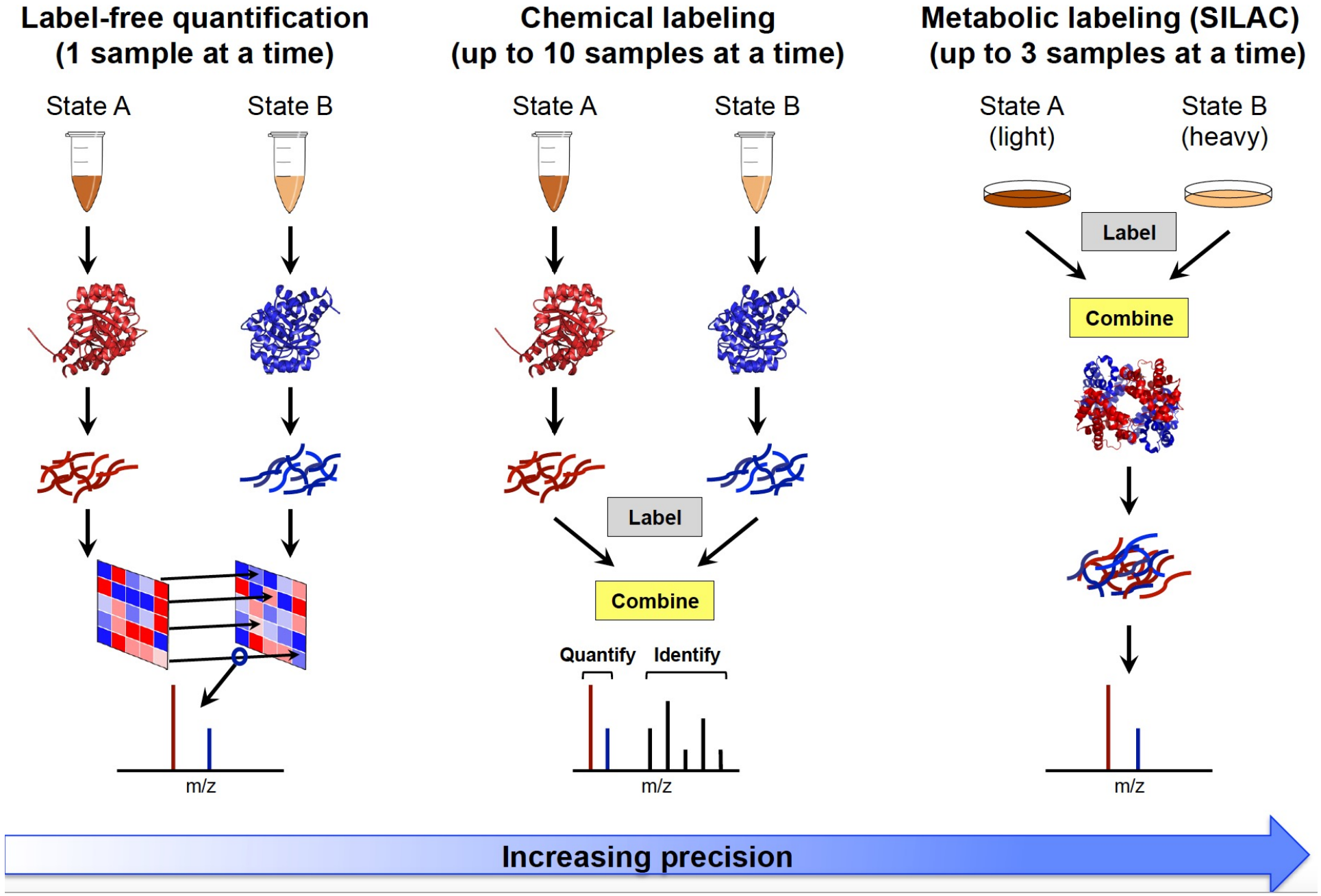


Drawbacks:

- Limited plex level (3 max)
- The method does not allow for the analysis of proteins directly from tissue.
- The stable-isotope enriched media are costly and may themselves affect cellular growth and protein production.
- The increase in nominal mass because of stable-isotope incorporation is not known until the sequence is determined.

Shao-En Ong & Matthias Mann. Nature Protocol. 2007

Summary


Label-free quantification
(1 sample at a time)
State A
State B
m/z
Chemical labeling
(up to 10 samples at a time)
State A
State B
Label
Combine
Quantify
Identify
m/z
Metabolic labeling (SILAC)
(up to 3 samples at a time)
State A
(light)
State B
(heavy)
Label
Combine
m/z
Increasing precision

Label versus label free

Label (pros and cons)

- Stable isotopes are expensive and not suitable for clinical samples
- But... less experimental variation if samples are mixed
- Even better the earlier workflows can be mixed...
- Improved quantitative precision and accuracy
- Improved confidence in peak identification

Label-free (pros and cons)

- Need lots of replicates to get statistical power
- So lots of time on MS instrument – therefore also can be expensive
- Fairly new technique therefore not enough high-quality published studies showing best practice
- No labelling needed
- No limit on the number of samples
- Applicable to any kind of samples

Definition of Absolute vs Relative Protein Quantification

Relative quantification

- Relative comparison of the same protein between samples
- ≥2 samples
- Output: protein ratio

Absolute quantification

- Comparison of the same protein between samples and different proteins within the same sample
- ≥1 sample
- Output: protein concentration (copies/cell, fmol/µg extract, ng/mL body fluid)

Challenges in LC-MS Platform

- Highly reproducible LC-MS analysis (retention time shift, fluctuations in MS signal intensity, peptide identification in separated MS/MS)
- Complex samples (overlapping signals, misaligned peptides)
- Large sample size (column degradation)

Limitations of LC-MS-based Approach to Large-scale Protein Profiling

- Sample size limited (ICAT, 2; iTRAQ, 8) for stable-isotope labeling approaches
- Difficult to trace protein abundance across a large number of samples
- Most peptides cannot be identified
- Difficult to identify & quantify low-abundance proteins

The multitude of quantitative MS-application

- Which type of quantitative mass spectrometric approach is most suited for my project?

  - What type of MS platform do I have access to?
  - How precise and accurate do my quantitative results have to be?
  - What type of sample am I working with?
  - How large is my project (number of samples)
  - Budget (costs and time)
  - Do I need relative or absolute quantitative data?

Application: Protein Biomarker Development

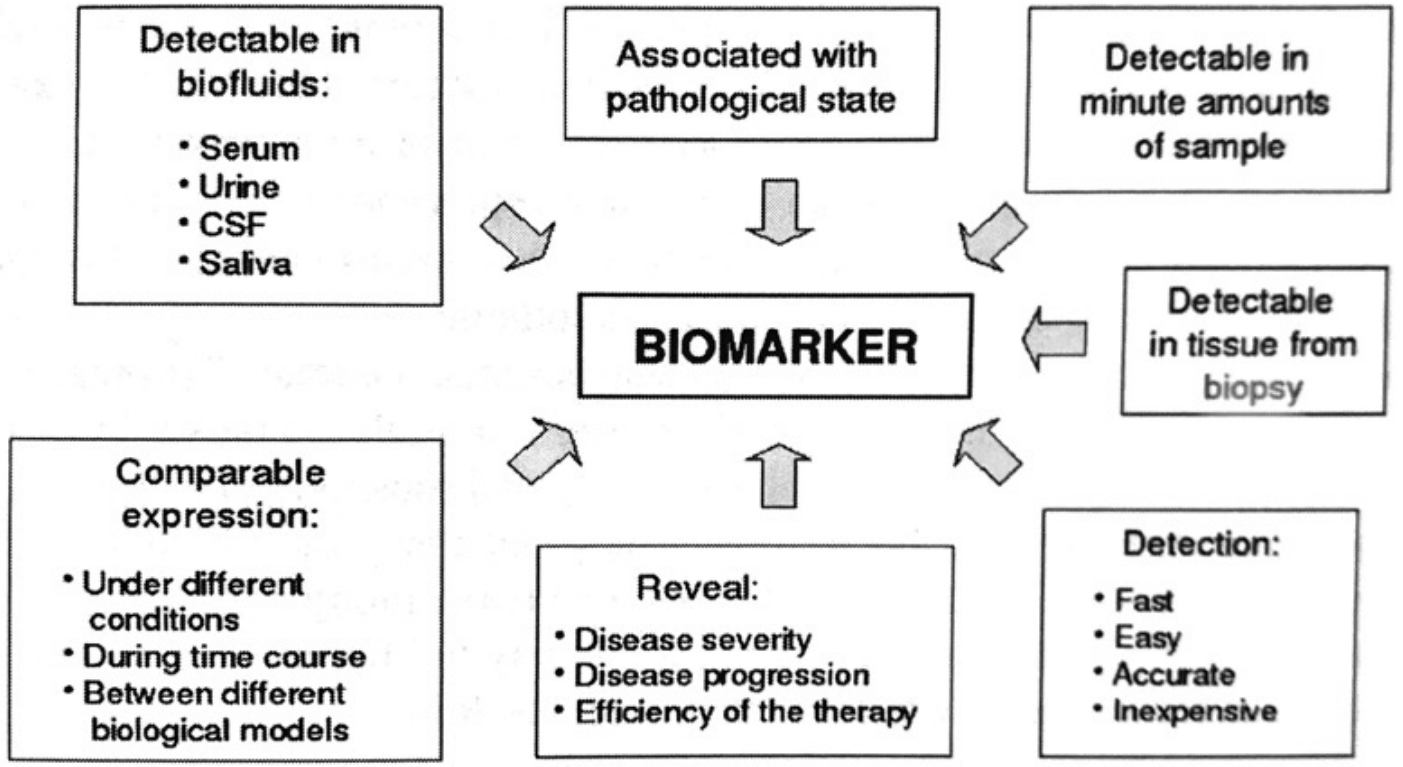
- Biomarker: molecular signature representing a state of a living organ/cell

-- Diagnostics

-- Prognostics

-- Target identification

-- Monitoring drug efficacy



nature protocols

REVIEW ARTICLE

https://doi.org/10.1038/s41596-021-00566-6

Tutorial: best practices and considerations for mass-spectrometry-based protein biomarker discovery and validation

Ernesto S. Nakayasu[1]✉, Marina Gritsenko[1], Paul D. Piehowski[1], Yuqian Gao[1], Daniel J. Orton[1], Athena A. Schepmoes[1], Thomas L. Fillmore[1], Brigitte I. Frohnert[2], Marian Rewers[2], Jeffrey P. Krischer[3], Charles Ansong[1], Astrid M. Suchy-Dicey[4], Carmella Evans-Molina[5], Wei-Jun Qian[1], Bobbie-Jo M. Webb-Robertson[1,6] and Thomas O. Metz[1]✉