Proteomics Technology and its Application for Cancer Analysis

Robert Chalkley

Mass Spectrometry Facility

UCSF

chalkley@cgl.ucsf.edu

Outline

- Mass Spectrometry and Peptide/Protein ID
- Mass Spectrometry and Quantification
- Cancer Proteomic Data

What is Mass Spectrometry?

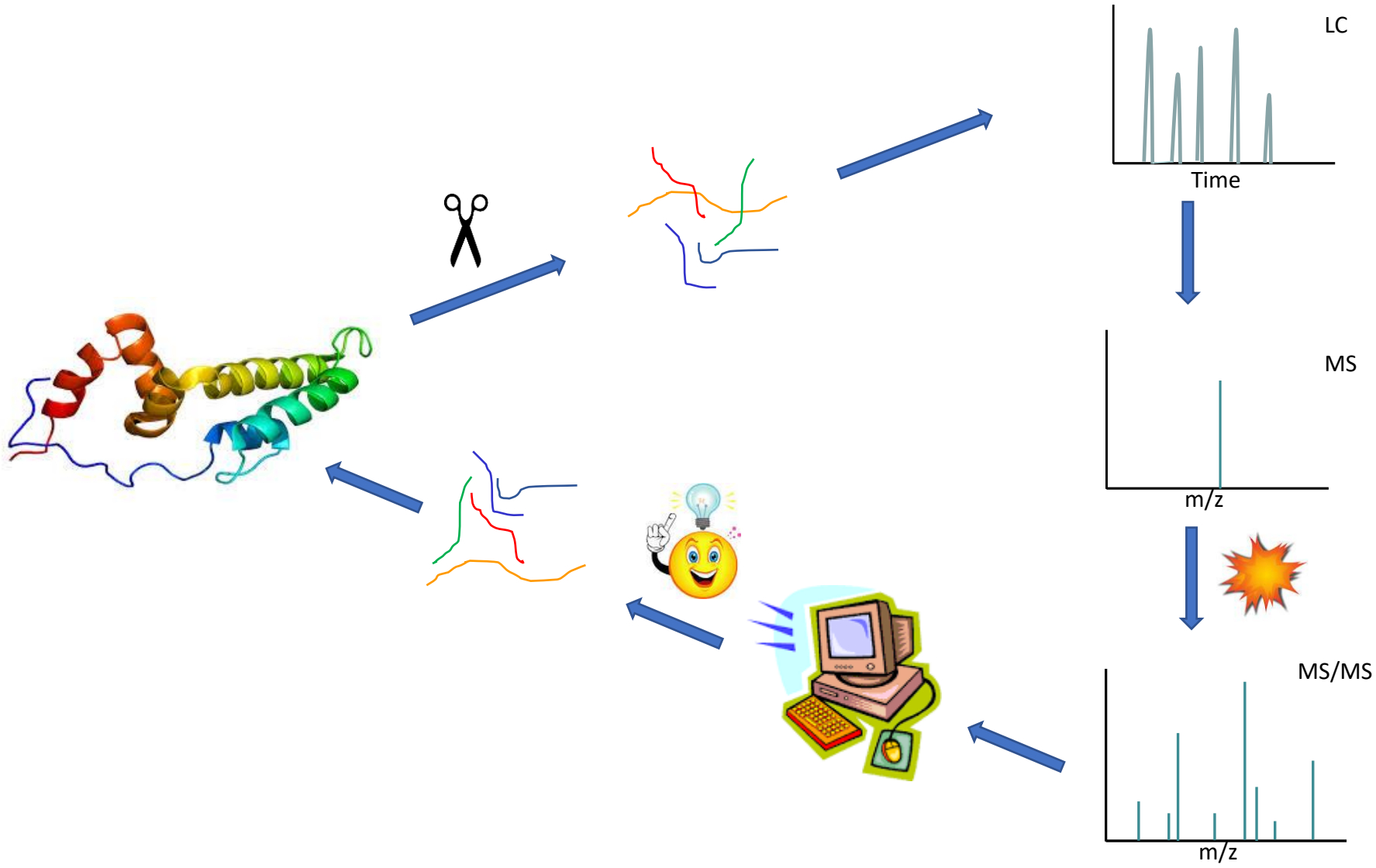
Analytical technique used to measure the mass-to-charge ratio (m/z) of gaseous ions. It is most generally used to find the composition of a physical sample by generating a mass spectrum representing the masses of sample components. The technique has several applications, including:

- Compound ID
- Determination of isotopic composition.
- Quantifying the amount of a compound (relative or absolute amount)
- Characterization of molecular structure
- Studying the fundamentals of gas phase ion chemistry (the chemistry of ions and neutrals in vacuum).

Mass Spectrometry and Proteomics

1. Protein identification, either by direct protein analysis, or by digesting the protein into smaller pieces (peptides), then identifying the peptides.
2. Identification of post-translational modifications: e.g. phosphorylation, acetylation.
3. Quantifying relative differences in protein/peptide levels between related samples.
4. Quantifying changes in post-translational modifications.
5. Identifying interactions, surfaces / 3D structure.

Standard Protein Identification Workflow



Why do we normally analyze peptides rather than proteins?

- MW of protein is not informative enough:
  - e.g. in UniProtKB database 8 human proteins mass 20000 +/- 2 Da
- Sequence in database may be wrong: 1 amino acid difference will change the mass, such that you will not identify the protein
- Sequence in the database may be of the pre-protein with signal peptide still attached. May be alternatively spliced product.
- Protein may be post-translationally modified.
- Protein may be a degradation product.
- Analysis of intact proteins is less sensitive than peptides.
- Fragmentation of intact proteins is inefficient and less sensitive in comparison to peptides

- Solution: Digest protein into peptides, then analyze peptides.

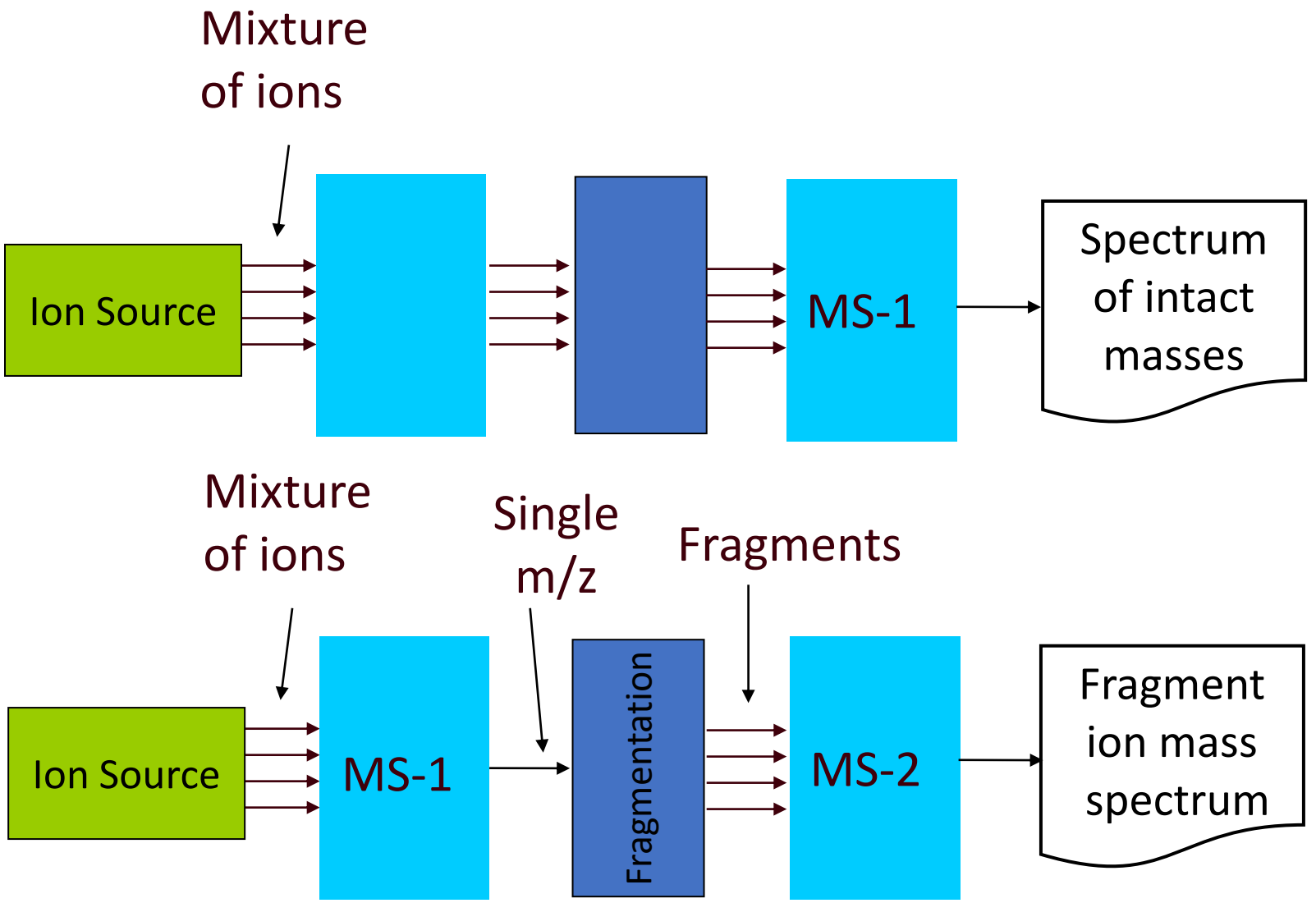
We Generally Use Trypsin to Digest Proteins

- Majority of peptides 7 - 20 amino acids in length
  - Long enough to be protein-specific while maintaining good sensitivity

- High Enzyme Specificity – cuts all Lys and Arg (slower when followed by Pro).
  - Can bioinformatically predict what peptides should be formed from any protein

- Produces peptides with basic C-terminus – give good fragmentation series

MS and MS/MS (Tandem Mass Spectrometry)



Amino Acid Residue Masses

Amino acid residue		Monoisotopic mass
Ala	A	71.03711
Cys	C	103.00919
Asp	D	115.02694
Glu	E	129.04259
Phe	F	147.06841
Gly	G	57.02146
His	H	137.05891
Ile	I	113.08406
Lys	K	128.09496
Leu	L	113.08406
Met	M	131.04049
Asn	N	114.04293
Pro	P	97.05276
Gln	Q	128.05858
Arg	R	156.10111
Ser	S	87.03203
Thr	T	101.04768
Val	V	99.06841
Trp	W	186.07931
Tyr	Y	163.06333

Modified Amino Acid Residue	Monoisotopic Mass
Homoserine Lactone	83.03712
Pyroglutamic acid	111.03203
Hydroxyproline	113.04768
Oxidised Methionine	147.03541
Carbamidomethylcysteine	160.03065

Peptide Identification Strategies

1. Compare observed fragmentation spectrum to theoretical fragmentation of peptides
    - Most common approach
    - Most flexible

2. Compare to previously acquired spectra

    Spectral Library

    +Can identify peptides based on lower quality spectra

        Know which peaks should be observed

        Can make use of peak intensities

    -Cannot identify anything new

        e.g. may not be good for PTM analysis

Is my Best Match Correct?

•Software gives you have a score for all peptides in the database that have the same precursor mass as your spectrum.

•You have a best match.

How do you decide whether this top scoring match is correct?

- Difficult to calculate a probability that it is correct due to potential homology.
  - Easier to calculate a probability that it is a random match.

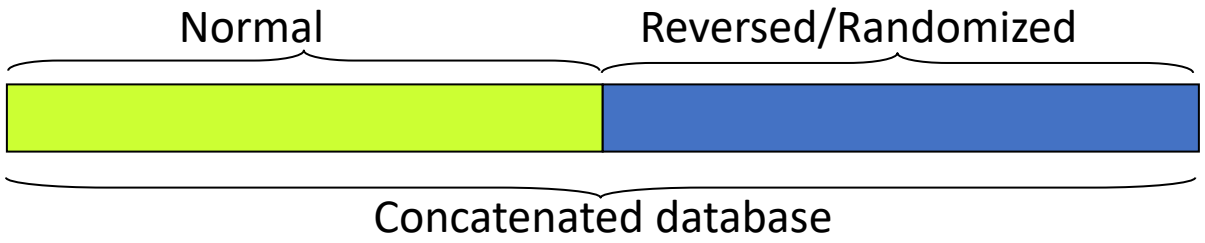
Most search engines report an Expectation value.

•The number of times a given score (or greater) is predicted to be achieved by random (incorrect) matches.

•Low Expectation value => probably correct

How Do You Assess the Reliability of a Dataset?

- Create a randomized/reversed version of the database and concatenate it onto the end of the normal database. (target-decoy)



- Search data against a concatenated database.
- At a given threshold, for every match to the random part of the database you predict one incorrect match to the normal part of the database.

- E.g. if in your database search you match 10 spectra above your score / expectation value threshold to the random part of the database, then there are probably 10 incorrect matches among those assigned to peptides in the normal part of the database.

- This is calculating a False Discovery Rate (reliability) for the dataset as a whole; not individual matches.

Nat Methods (2007) 4 3 207-14

Mass Spectometry and Quantification

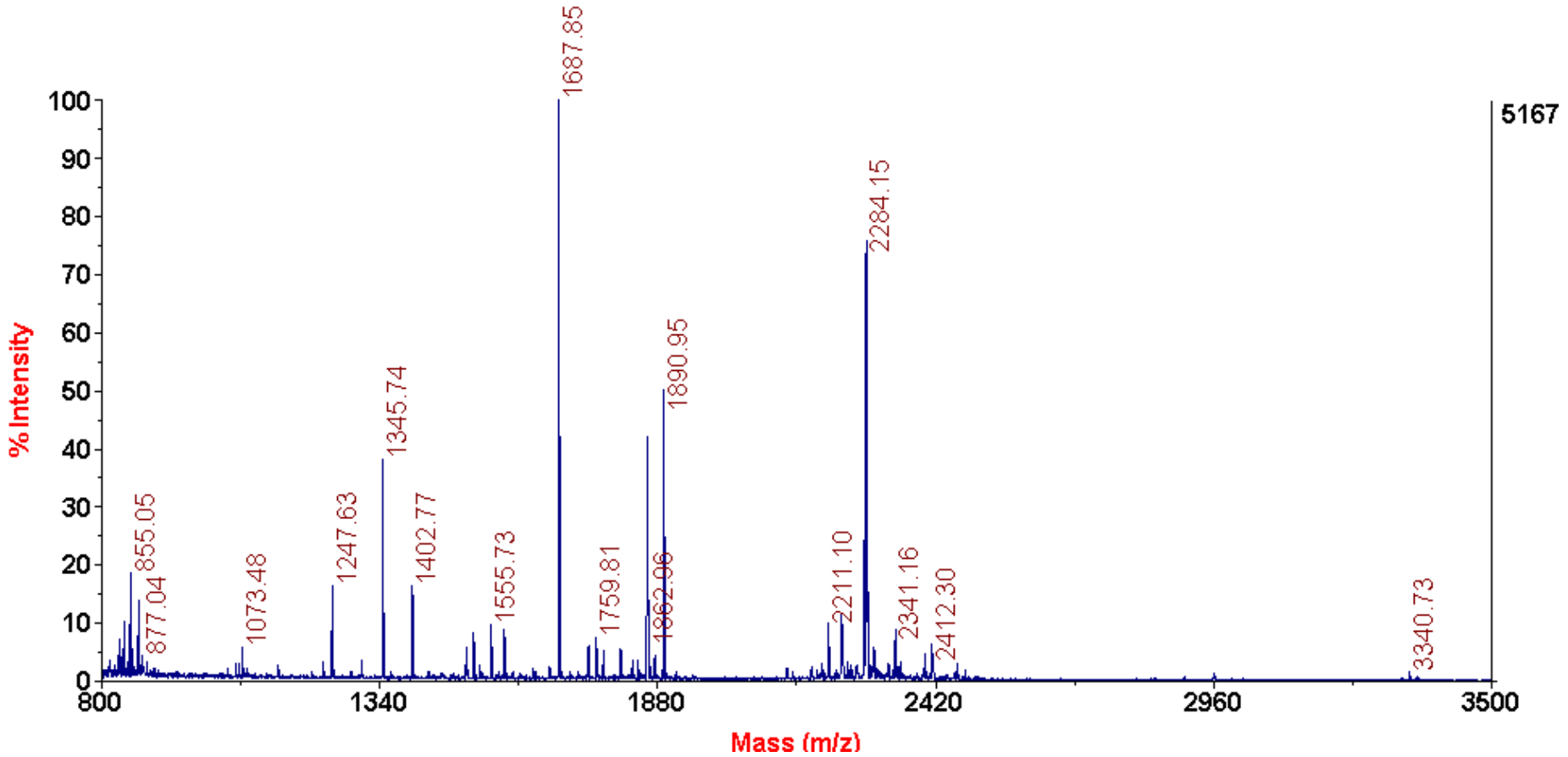
- MS signal is not inherently quantitative

- Factors that contribute to peak intensity
  - Absolute amount of component
  - Ionizability
  - Number of charge states component is observed in
  - Relative amount to other components (complexity of sample)

MS signal is not inherently quantitative

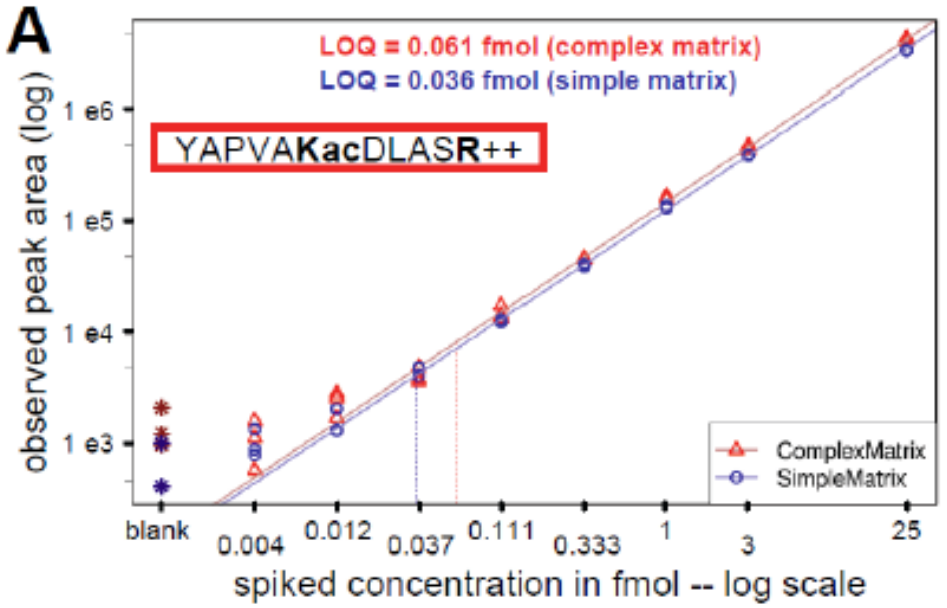
Components of the same concentration may give drastically different peak intensities

- e.g. different peptides from the same protein



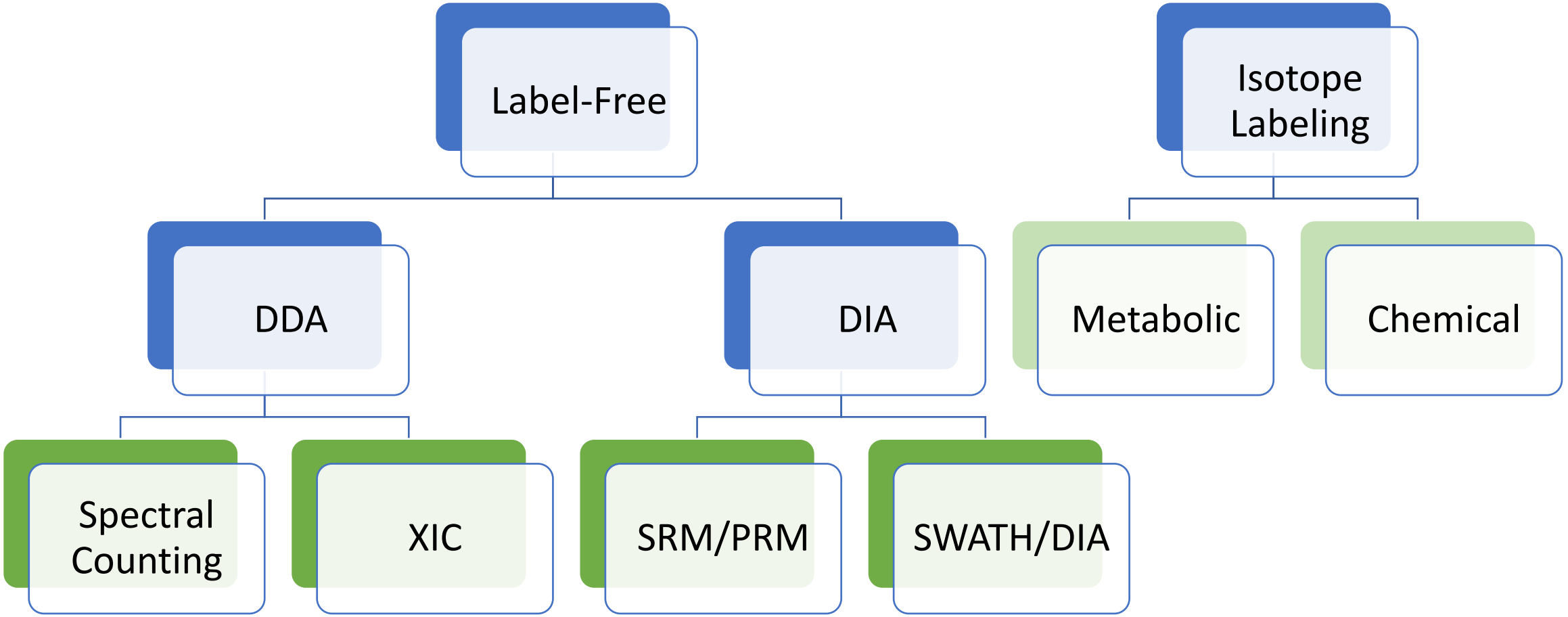
MS signal is not inherently quantitative

- Practically all MS-based quantification is comparing relative peak intensities of the same component
  - For label-free approaches this is comparing between samples
  - For isotope-labeled samples the comparison is within the same sample
- For absolute quantification one measures the peak intensity of a known amount (preferably a calibration curve of different known amounts) and then compares observed signal
  - Due to sample complexity affecting ionization/signal, the known amount must be spiked into same background



Mol Cell Proteomics (2012) 11 5 202-14

Mass Spectrometry Based Quantification Methods



Match Your Quantification Method To Your Experimental Design

How many samples do you want to compare?

- One vs. one (e.g. WT vs. mutant)
- One vs. many (e.g. time course)
- Many vs. many (e.g. drugs vs. genotypes)

How accurate do you need the quantification?

- Is knowing something has changed enough, or do you need to know the magnitude?
- Are you going to confirm the result by a different method?

Do you have enough statistical power in your experimental design?

- You may need multiple controls for biological variability

Consider amount of material, instrument time, cost of reagents

Mass Spectrometry Acquisition Methods

- If peptide identification is based on fragmentation (MS/MS) data, how do you decide what is fragmented?

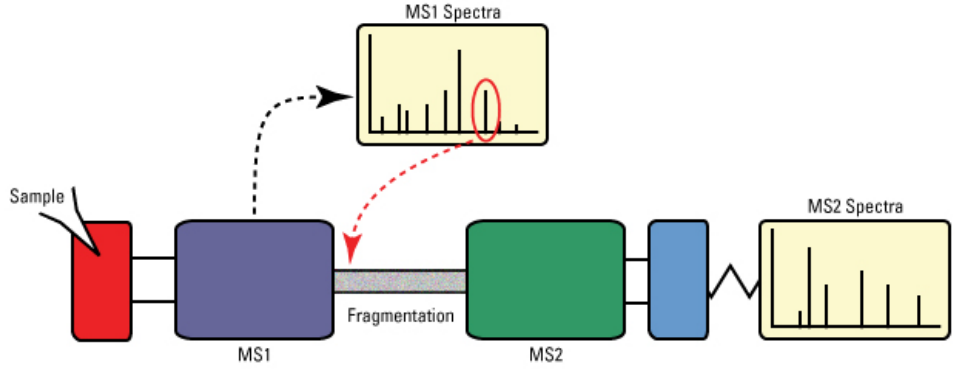
Data-Dependent Acquisition (DDA)

- Mass spectrometer automatically selects the most abundant peak in the MS spectrum for fragmentation analysis, then the 2nd most intense… Once a peak has been selected it is put on an exclusion list.

Data-Independent Acquisition (DIA)

- Mass spectrometer is given a pre-determined list/range of masses to fragment, independent of what is eluting.
  - This includes targeted analysis.

DDA for Quantitative Proteomics



Advantages

- Easiest to implement
  - Can use the same acquisition method for all projects/samples
  - Do not need any prior knowledge about the protein composition of the sample
- Can provide quantitative information about most proteins detected

Disadvantages

- Somewhat biased toward more abundant components
- If you are comparing between samples, the same components may not be selected for MS/MS
  - May have missing data points when comparing samples
    - Lack of MSMS does not mean it was not present
  - Can try to match MS1 peaks between runs independent of MS/MS analysis
    - Need to control for errors

DDA Label-free Quantification

- Spectral counting
- Extracted ion chromatogram (XIC) (MS1 filtering)

Spectral Counting

- Use count of spectra identified to a protein as a measure of protein abundance

Rank	Acc #	Exploris_20220418/SProt_mouse			Protein MW	Species	Protein Name
		Num Unique	Peptide Count	% Cov			
1	Q8VDD5	148	574	59.2	226374.0	MOUSE	Myosin-9
2	Q05920	92	1118	74.2	129685.7	MOUSE	Pyruvate carboxylase, mitochondrial
3	E9Q557	106	176	30.2	332915.0	MOUSE	Desmoplakin
4	Q91ZA3	70	681	72.1	79922.5	MOUSE	Propionyl-CoA carboxylase alpha chain, mitochondrial
5	Q5SWU9	90	124	46.1	265259.0	MOUSE	Acetyl-CoA carboxylase 1
6	Q99MR8	67	353	75.7	79344.5	MOUSE	Methylcrotonoyl-CoA carboxylase subunit alpha, mitochondrial
7	Q05793	74	102	27.4	398297.7	MOUSE	Basement membrane-specific heparan sulfate proteoglycan core protein
8	Q9JHU4	91	95	23.9	532050.0	MOUSE	Cytoplasmic dynein 1 heavy chain 1
9	Q9ERU9	72	86	29.6	341123.8	MOUSE	E3 SUMO-protein ligase RanBP2
10	O70133	66	132	48.6	149476.2	MOUSE	ATP-dependent RNA helicase A
11	Q9QXS1	87	94	22.2	534192.0	MOUSE	Plectin
12	Q99104	73	111	41.2	215540.1	MOUSE	Unconventional myosin-Va
13	Q99MN9	48	477	77.8	58409.5	MOUSE	Propionyl-CoA carboxylase beta chain, mitochondrial
14	E9Q7G0	67	75	41.3	235632.3	MOUSE	Nuclear mitotic apparatus protein 1
15	Q68FD5	61	95	46.5	191558.3	MOUSE	Clathrin heavy chain 1

Limitations of Spectral Counting

- Granular and stochastic
  - Need at least 4 peptides (in one sample) to be able to make any quantitative estimate

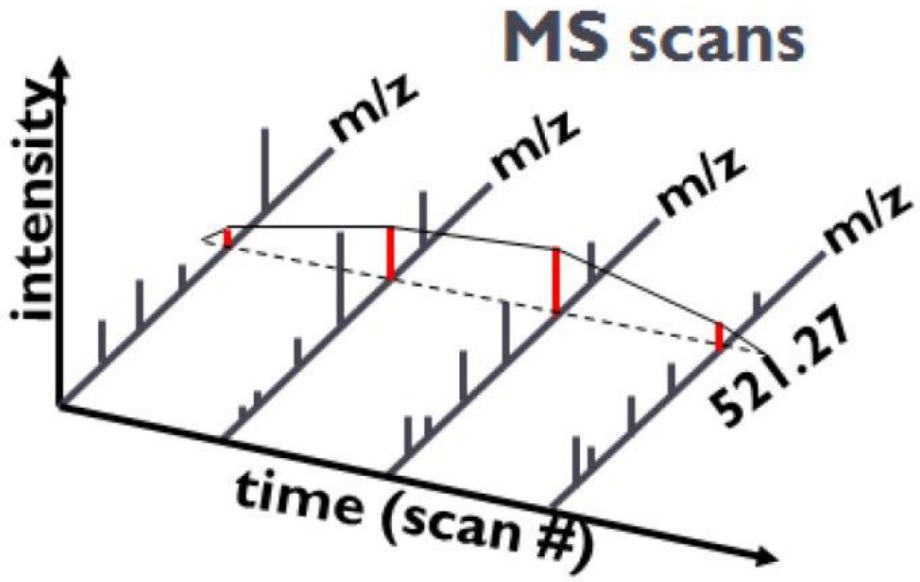
1150	Q9R1P0	3	3	10.0	29471.0	MOUSE	Proteasome subunit alpha type-4
1151	Q9CY50	1	1	5.2	32065.3	MOUSE	Translocon-associated protein subunit alpha
1152	P63073	1	1	6.5	25053.4	MOUSE	Eukaryotic translation initiation factor 4E
1153	P06801	2	2	5.1	63954.3	MOUSE	NADP-dependent malic enzyme
1154	Q64676	2	2	3.3	61249.8	MOUSE	2-hydroxyacylsphingosine 1-beta-galactosyltransferase
1155	Q05CL8	1	1	2.5	64802.8	MOUSE	La-related protein 7
1156	Q8BTS4	1	1	3.3	55732.3	MOUSE	Nuclear pore complex protein Nup54
1157	Q9CXZ1	1	2	8.6	19784.9	MOUSE	NADH dehydrogenase [ubiquinone] iron-sulfur protein 4, mitochondrial
1158	Q8C0E2	2	2	6.8	39124.9	MOUSE	Vacuolar protein sorting-associated protein 26B
1159	Q8BGA7	1	1	4.0	49908.8	MOUSE	Integrator complex subunit 15
1160	Q9Z2I8	3	3	9.5	46840.4	MOUSE	Succinate--CoA ligase [GDP-forming] subunit beta, mitochondrial

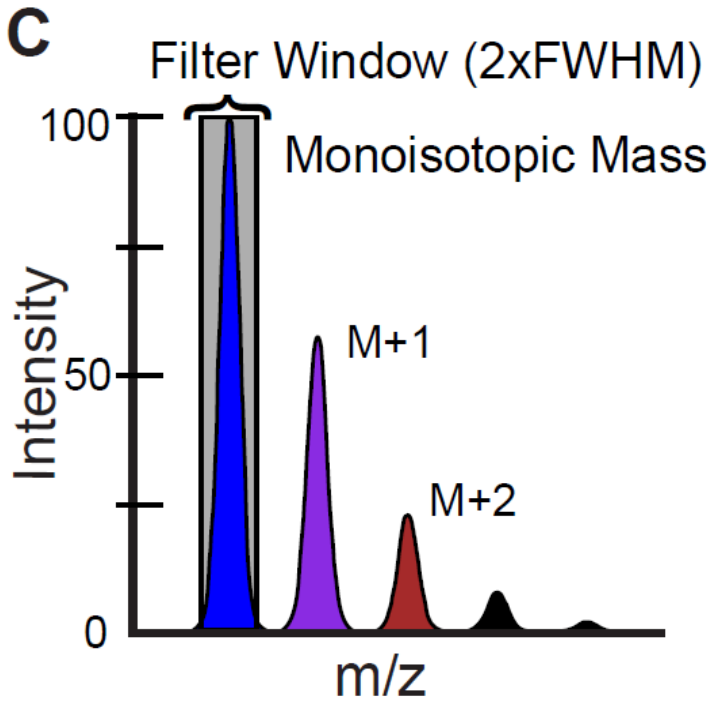
In general, spectral counting can identify proteins that are changing in level, but not the magnitude of change.

XIC / MS1 Filtering

Plot intensity of peak across all scans as it elutes

Area under plot (XIC) is used as quantitative measure of amount of peptide





Need MS/MS ID for confident peak assignment, but can use accurate mass and retention time to try to match peaks across related runs

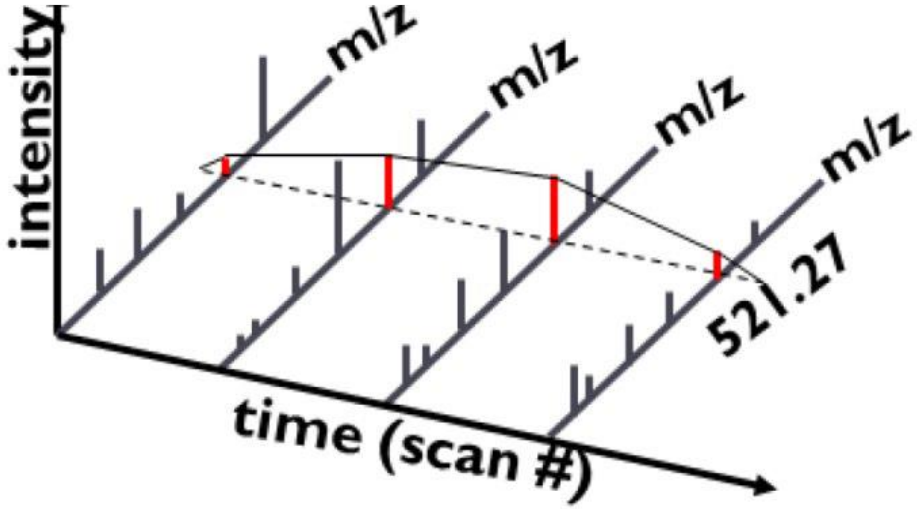
At protein-level can consider sum or median of all or top# peptides

Mol Cell Proteomics (2012) 11 5 202-14

Targeted MS/MS

- Need to compile a list of peptides of interest to target
  - This is normally data from DDA analysis
- Can only target a certain number of components at any time; a few tens
  - If from DDA data you know elution time, can schedule when to target peptide
    - With long, reproducible chromatography it is possible to target several hundred peptides.

- Multiple MS/MS spectra are required per precursor
- Quantification is normally by measuring area of fragment ions across elution of peptide



DIA / SWATH

Instead of targeting specific components, fragment the whole mass range in a series of windows

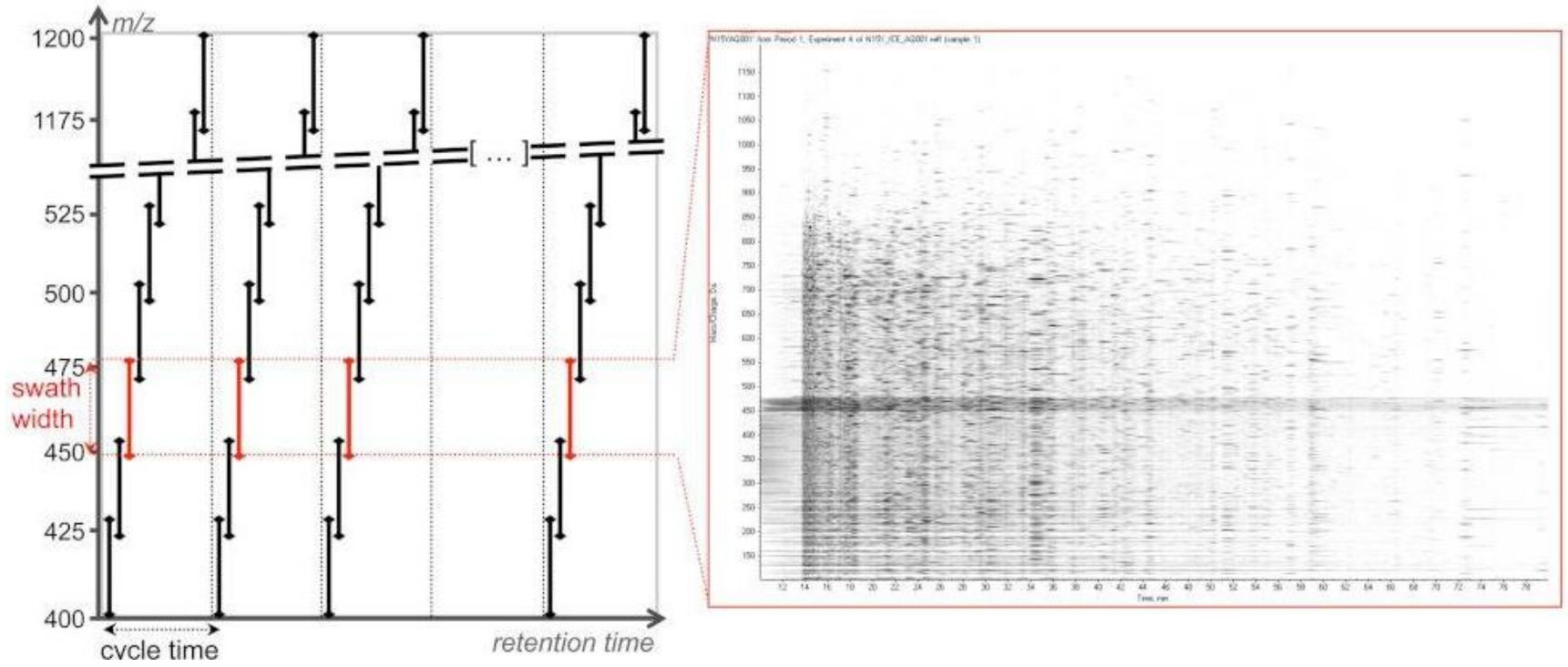
e.g. 25 m/z windows are sequentially fragmented.

Advantage: You have fragmentation data of all components

You can mine data at a later point for new components

No missing datapoints

Disadvantage: Fragmentation spectra contain a mixture of many components



[*Mol Cell Proteomics* (2012) O111.016717]

Normalization in Label-free Quantification

When comparing samples, how do you adjust for differences in protein input amounts?

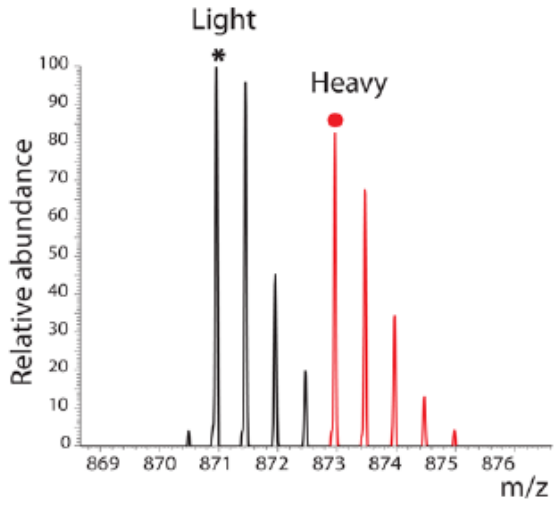
- Assume certain protein/s are not changing between conditions
  - e.g. abundant structural proteins
- Assume the total amount of protein has not changed between conditions
  - Normalize to the sum intensity of all peptides
  - Normalize to the median intensity of all peptides
- Spiked in standard (peptide or digest) can be added for normalization

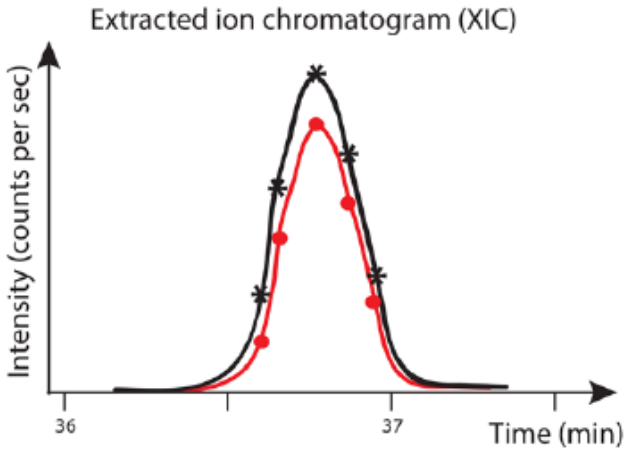
Stable Isotope Labeling for Quantification Within a Sample

^{1}H versus ^{2}H
^{12}C versus ^{13}C
^{14}N versus ^{15}N
^{16}O versus ^{18}O

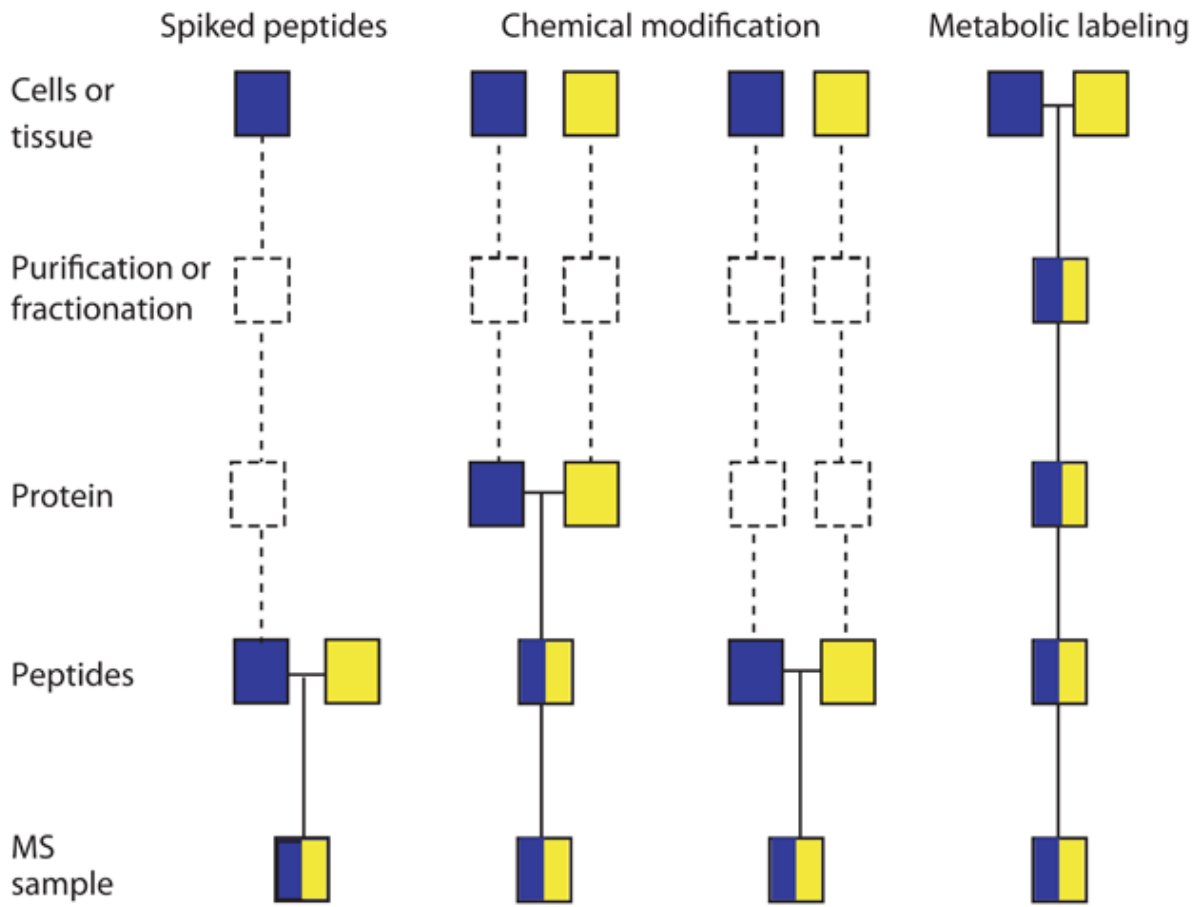
Desired properties:

- Uniform, complete, stable labeling
- Light and heavy co-elute
- Sufficient mass separation
- Label does not interfere with digestion, ionization, fragmentation, etc.

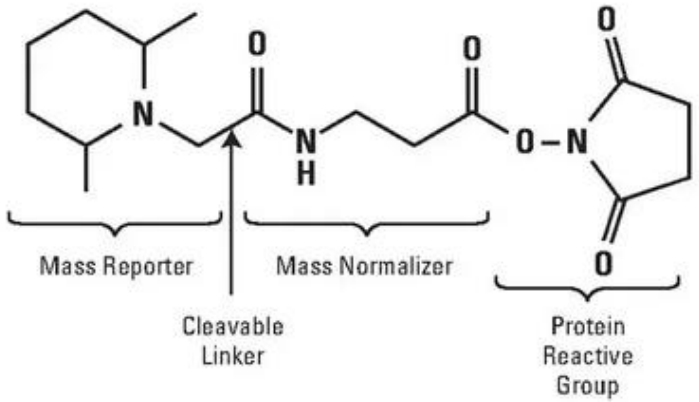


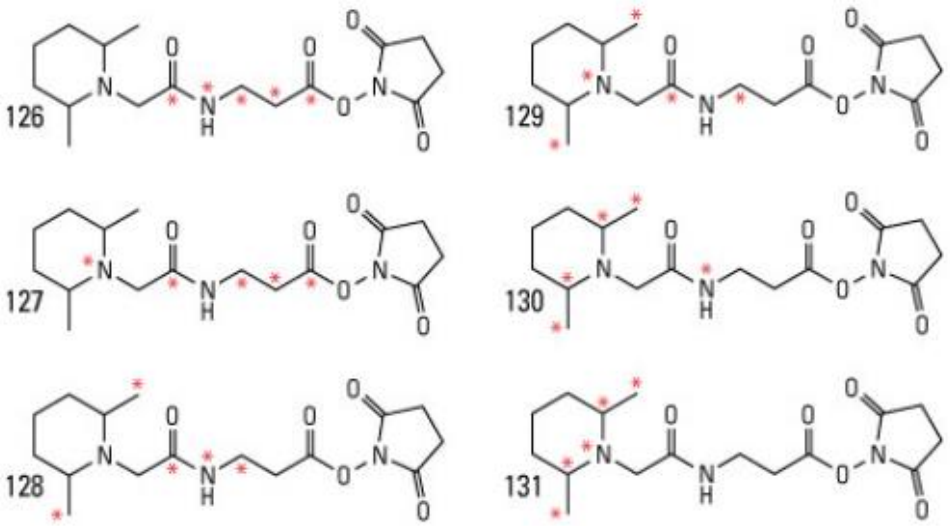


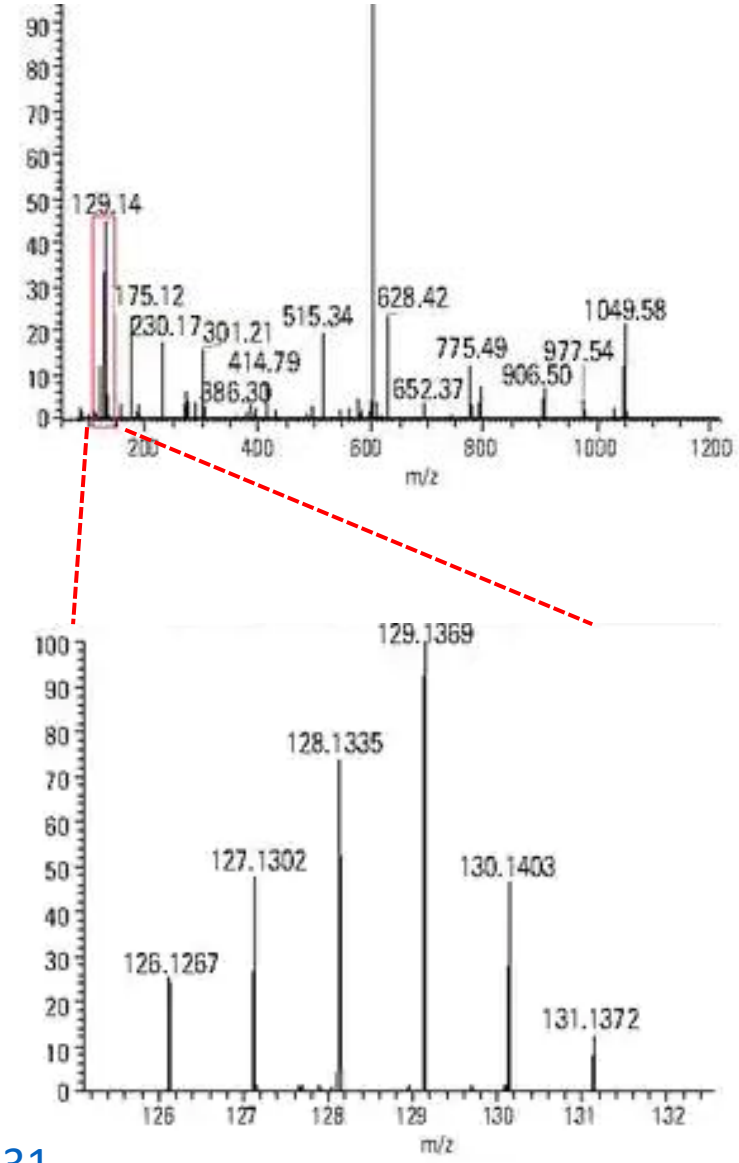
Different experimental designs lead to combination of samples at different points in the analysis



Isobaric Tags: TMT Reagents

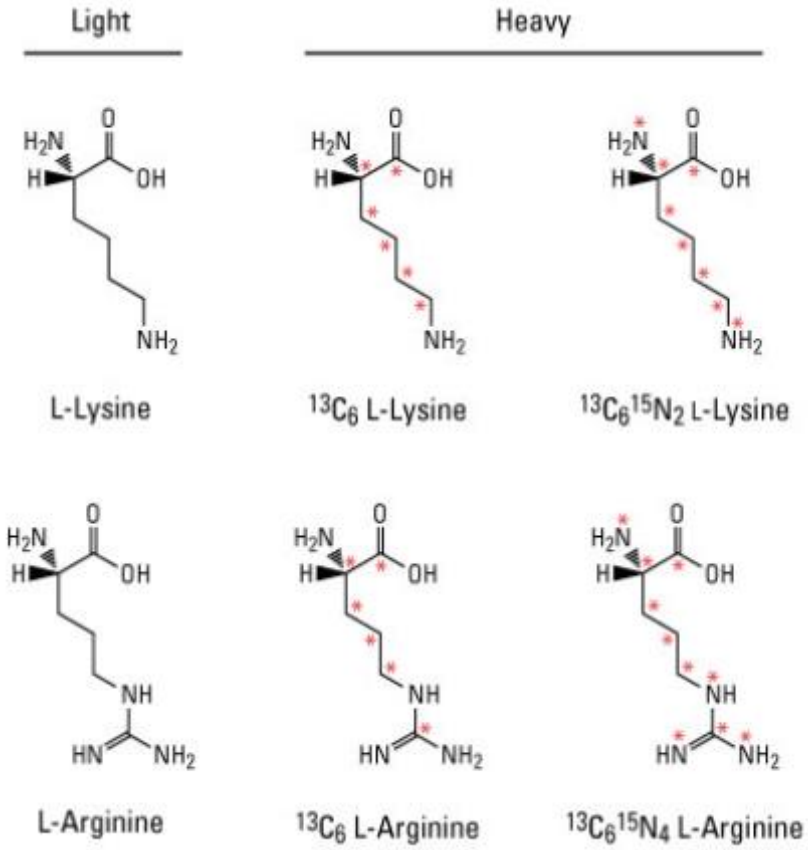


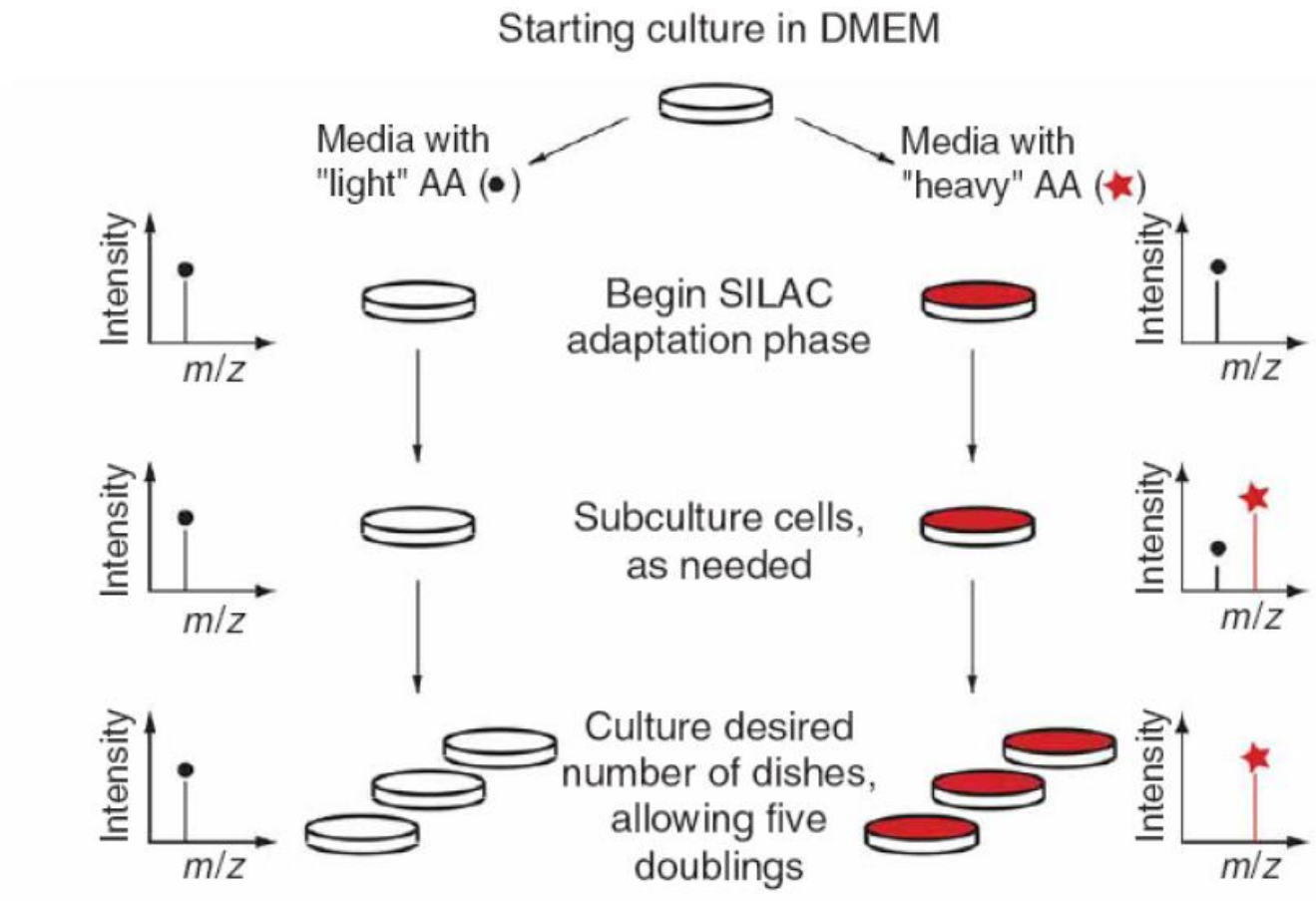




[Anal Chem (2008) 80 8 2921-31]

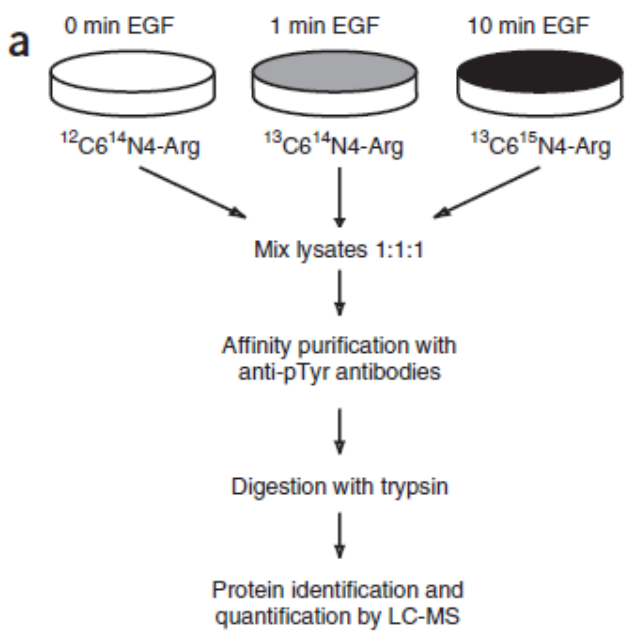
Metabolic Labeling:
Stable Isotope Labeling of Amino Acids in Culture (SILAC)

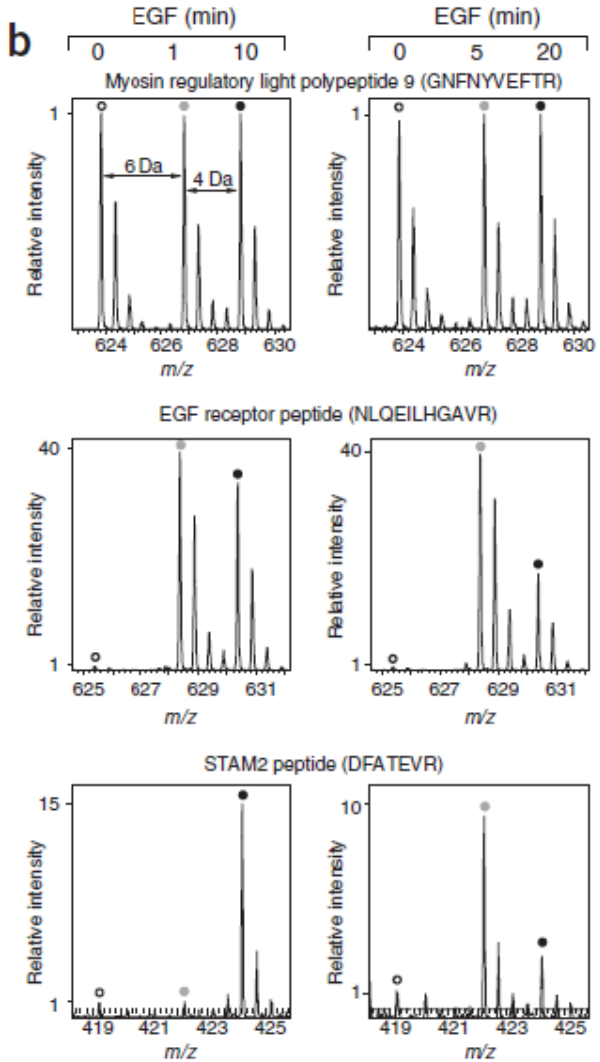




Nat Protoc (2006) 1 6 2650-60

Example: EGF stimulation in HeLa cells





[Nat Biotechnol (2004) 22 9 1139-45]

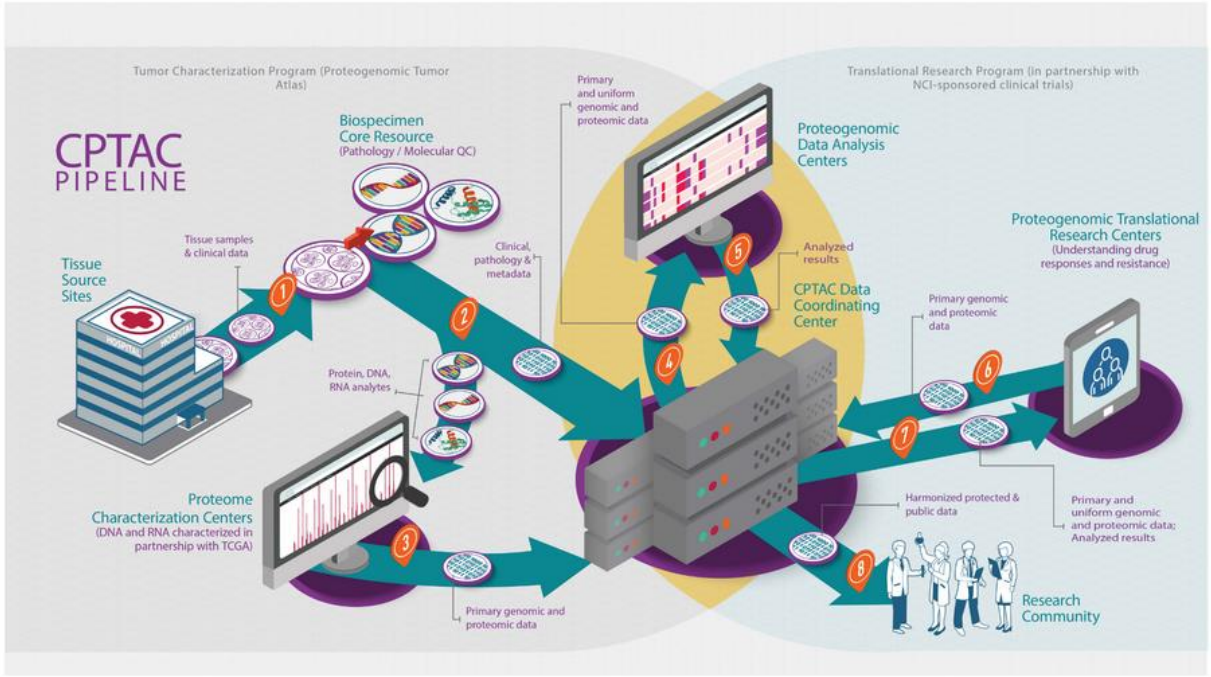
Proteomics Data Availability

- Raw data associated with proteomics publications is freely available
  - A pre-requisite for publishing in most journals
- Most data is submitted to a repository in the proteomeXchange consortium[1].
- Metadata is less consistent.
  - May be in repository submission
  - May be in journal publication
  - Mapping from samples to files is sometimes not clear.

[1] http://www.proteomexchange.org

Cancer Data Creation

CPTAC



Data Portal

Antibody Portal

Assay Portal

CONTACT US

SIGN UP FOR EMAIL UPDATES

The National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC) is a national effort to accelerate the understanding of the molecular basis of cancer through the application of large-scale proteome and genome analysis, or proteogenomics.

- Funded half a dozen centers for proteomic data analysis of samples supplied by NCI.
  - There is matched genomic data for these samples.
- Big efforts in standardizing sample preparation and analysis to make results acquired in different labs more comparable.
- Metadata for these samples is relatively complete.
- Much of emphasis now is on proteogenomics.

https://proteomics.cancer.gov/programs/cptac

Cancer Data Sharing

https://proteomic.datacommons.cancer.gov/pdc/

Cancer Data Application

NATIONAL CANCER INSTITUTE
DCTD Division of Cancer Treatment & Diagnosis

CONTACT US | SIGN UP FOR UP

Search... SEARC

OCCPR Office of Cancer Clinical Proteomics Research

APOLLO HOME

APOLLO Network

The ***A***pplied ***P***roteogenomics ***O***rganizationa***L*** ***L***earning and ***O***utcomes (APOLLO) network is a collaboration between NCI, the Department of Defense (DoD), and the Department of Veterans Affairs (VA) to incorporate proteogenomics into patient care as a way of looking beyond the genome, to the activity and expression of the proteins that the genome encodes. The emerging field of proteogenomics aims to better predict how patients will respond to therapy by screening their tumors for both genetic abnormalities and protein information, an approach that has been made possible in recent years due to advances in proteomic technology.

Screening a patient's proteogenome may enable researchers to more precisely match their tumor types to targeted therapies than screening for genomic mutations alone. Proteogenomics may also help researchers more completely characterize the biologic pathways of cancer development, metastasis, and treatment resistance. Multiple pilot studies have shown that, across tumor types, proteogenomics identifies additional biology, beyond standard genomics alone.

Since most cancer drugs target proteins, researchers hope that combining protein analysis with gene analysis will improve the ability to predict tumor response to treatment and, eventually, to match the tumor with the right drug.

Data Portal

Antibody Portal

Assay Portal

CONTACT US
SIGN UP FOR EMAIL UPDATES

https://proteomics.cancer.gov/programs/apollo-network

NATIONAL CANCER INSTITUTE
DCTD Division of Cancer Treatment & Diagnosis

CONTACT US | SIGN UP FOR UPDATES

Search... SEARCH

OCCPR Office of Cancer Clinical Proteomics Research

ICPC HOME ICPC PORTAL ABOUT US PROTEOGENOMIC PROJECTS

International Cancer Proteogenome Consortium

The International Cancer Proteogenome Consortium (ICPC), is a voluntary scientific organization that provides a forum for collaboration among some of the world's leading cancer and proteogenomic research centers. Catalyzed by the effort of the Cancer Moonshot to encourage international cooperation and investments among nations in cancer research and care, as well as new efforts in precision medicine, the International Cancer Proteogenome Consortium (ICPC) was launched in late 2016. The ICPC brings together more than a dozen countries to study the application of proteogenomic analysis in predicting cancer treatment success and to share data and results with researchers worldwide, hastening progress for patients.

Search

Search...

Sort By

Project

Breast Cancer
Breast
United States

Clear Cell Renal Cell Carcinoma
Kidney
United States

Colon Adenocarcinoma
Intestine
United States

Cutaneous Melanoma
Skin
Sweden

Early-onset Gastric Cancer
Stomach
South Korea

Glioblastoma Multiforme
Brain
United States

HBV+ Hepatocellular Carcinoma
Liver
China

HPV-negative HNSCC
Head and Neck
United States

Intrahepatic Cholangiocarcinoma
Liver
China

ICPC INTERNATIONAL CANCER PROTEOGENOME CONSORTIUM

CONTACT US

SIGN UP FOR EMAIL UPDATES

NEWS AND MEDIA

PUBLICATIONS

https://icpc.cancer.gov/portal/

Proteomics and Cancer: Where Are We?

Coverage

- Proteomics is able to identify close to 10K proteins per sample.
  - It can provide relative quantitative information for the majority of these.

Missing information

- It does not observe the whole sequence of a protein.
  - If you do not see a region, you cannot map a variant/mutation.
- It may not identify the same set of proteins when comparing samples.
  - Not identifying does not mean it is not there.

Comparing Data

- Sample handling/preparation is a major source of variability
  - Often leads to many of the differences when comparing results acquired in different labs.
  - NCI/CPTAC have managed to reduce this problem significantly for their data.
    - Data from a lot of cancer types now available.