EXPLORING THE "BUSINESS END" OF THE GENOME THROUGH PROTEOMICS AND TRANSCRIPTOMICS

LARRY HELSETH, PHD

August 23, 2017

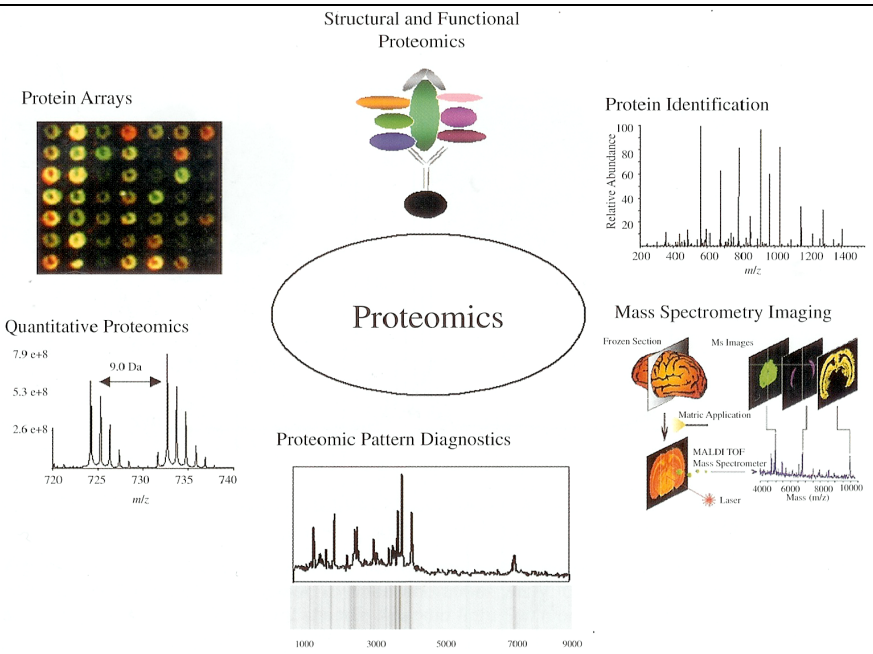
Outline

2

- Describe how proteomics can determine protein structure
- Describe quantitative proteomics techniques
- Describe informatics approaches
- Describe top-down proteomics
- Describe how transcriptomics informs proteomics
- Demo some software for inspection of proteomics results
- Aebersold & Mann Sept 2016 review "Mass-spectrometric exploration of proteome structure and function" http://www.nature.com.proxy.uchicago.edu/nature/journal/v537/n7620/pdf/nature19949.pdf

What is proteomics?

- Proteomics is the study of all proteins in a cell or organism ("proteome")
  - The study of what proteins are expressed at all stages of an organism's or cell's lifecycle.
  - The protein complement to the genome.
- Classical techniques include:
  - 2 dimensional gel electrophoresis
  - Nuclear Magnetic Resonance (NMR)
  - Mass spectroscopy (MS)
  - Database searches to identify proteins



From-TD Veenstra & JR Yates III, "Proteomics for Biological Discovery", Wiley, 2006

MSBI 32400 - Introduction to Bioinformatics 8/23/17

History of Mass Spectrometry

5

- Developed in early 1900s for study of small molecules
- Extended in 1990s to study of peptides and proteins
- Nobel Prize awarded in 2002 for development of two main methods (~5 years before):
  - MALDI
  - Electrospray (ESI)

2002 Nobel Prize in Chemistry

6

Awarded "for the development of methods for identification and structure analyses of biological macromolecules"

- To **John B. Fenn** and **Koichi Tanaka** "for their development of soft desorption ionisation methods for mass spectrometric analyses of biological macromolecules"
- To **Kurt Wüthrich** "for his development of nuclear magnetic resonance spectroscopy for determining the three-dimensional structure of biological macromolecules in solution" (NMR)
- Tanaka's contribution – **hovering through blasting**
- Fenn's contribution – **hovering through spraying**

http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2002/popular.html

Matrix-assisted laser desorption ionization (MALDI) – Koichi Takada

7

- Sample mixed with carrier substance ("matrix") and dried on plate
- Laser heats matrix and causes matrix and sample to sublime off surface
- Ions are accelerated and move down tube in vacuum
- "Time of flight" measures mass and charge of substance

MALDI Sample Plate

8

- Samples are spotted on plate and dried

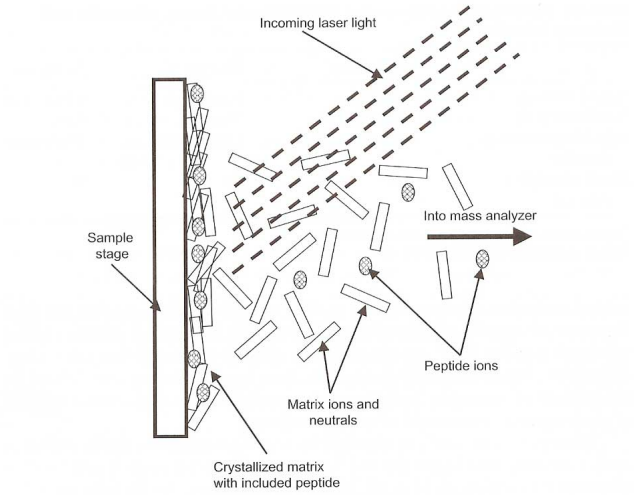
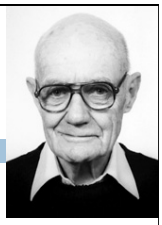


Figure 3.4. A generalized view of the processes associated with matrix-assisted laser desorption/ionization. The protein or peptide analyte are co-crystallized with the matrix compound on the sample stage and are irradiated with UV-laser pulses. The laser pulses vaporize the matrix compound and produce a plume that carries the protonated peptide or protein into the gas phase. The gas-phase ions are directed into the mass analyzer by appropriate electric fields.

From-Kinter & Sherman, "Protein Sequencing and Identification Using Tandem Mass Spectroscopy, 2000

Nobel Prize - John Fenn



10

- Electrospray ionization (ESI) involves dispersing sample as a liquid into vacuum so that sample is ionized and solvent sublimes
- Allows samples to be injected into mass spectrometer as they come off column
- Most commonly used method for proteomics

Electrospray

11

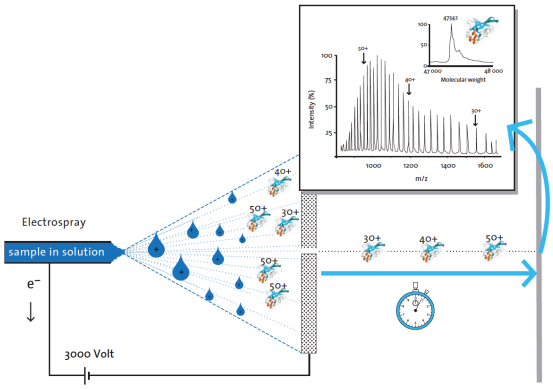


Figure 1. The electrospray process.

http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2002/advanced-chemistryprize2002.pdf

Routine separation protocol (Gel-based)

12

- Samples are separated by SDS polyacrylamide gel electrophoresis
- Proteins are stained to identify bands
- Band is cut out of gel
- Gel slice is cleaved with trypsin to break protein into peptides
- Peptides extracted from gel slice and injected into MS/MS

Routine Separation (Solution)

13

- Separate proteins by Isoelectric Focusing (by pI)
- Purify by liquid chromatography
- Immunoaffinity (using antibodies on beads)

➔ Digest with trypsin before MS/MS

MS/MS Fragmentation

14

- Identifying the sequence in each peptide by **Collision-induced Dissociation** (**CID**) is the key to protein identification.

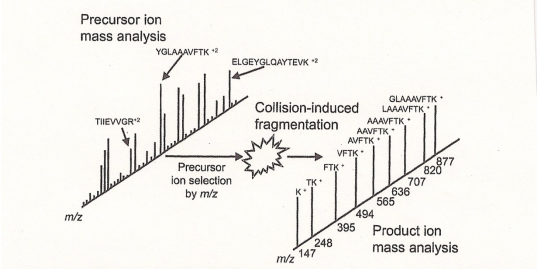
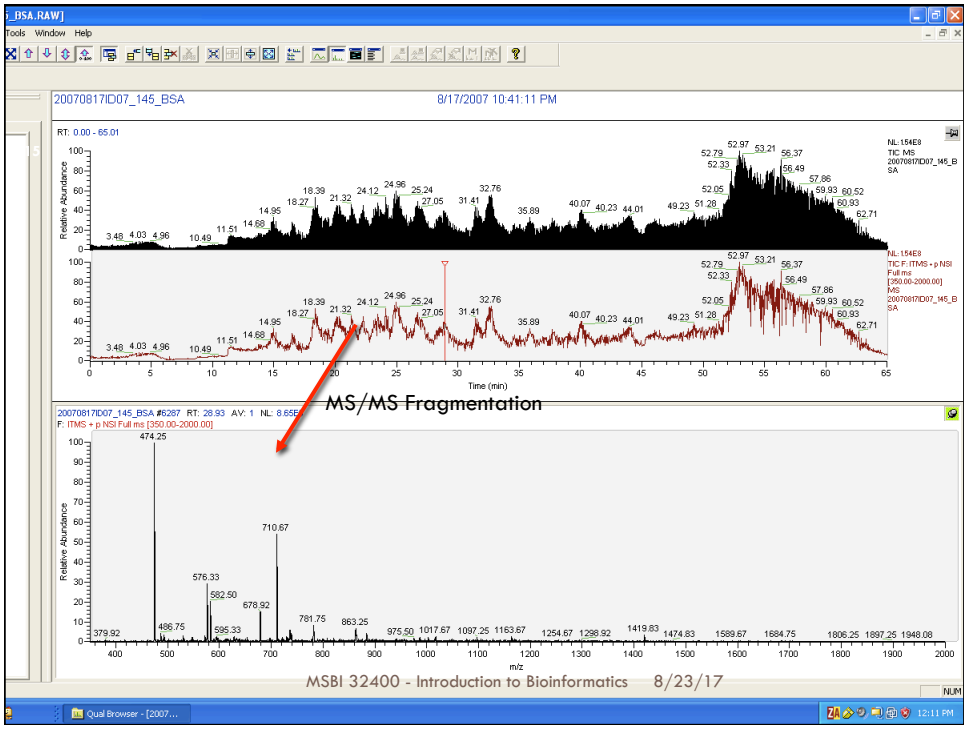


Figure 3.12. Structural characterization of a mass-selected ion by tandem mass spectrometry. Precursor ion mass analysis determines the *m/z* of the peptide ion of interest. That ion is mass-selected by the first stage of mass analysis and is activated by collision with a neutral gas molecule to induce a fragmentation reaction. The ionic products of the fragmentation reaction are mass-analyzed in the second stage of mass analysis to produce a product ion spectrum.

From-Kinter & Sherman, "Protein Sequencing and Identification Using Tandem Mass Spectroscopy, 2000



But…

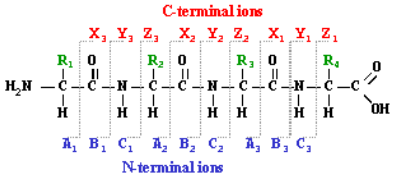
16

- Can you tell which peptides were in the BSA peaks?
  - You need a computer with sophisticated search algorithms to identify your peptides and the proteins they came from.

De Novo Sequencing

17

- Peptide breaks randomly



- Where peptide breaks affects MS spectrum

De Novo Sequencing (cont.)

- The residue mass difference between two peaks tells us which amino acid was next in the sequence.

Table A.1A.1 Compositions and Masses of the Twenty Commonly Occurring Amino Acid Residues[a,b]

Name	Composition	Monoisotopic mass	Average mass	pK_a of ionizing side chain[c]	Occurrence in proteins (%)
Alanine (Ala, A)	C_3H_5NO	71.03711	71.0788	—	8.3
Arginine (Arg, R)	$C_6H_{12}N_4O$	156.10111	156.1876	~11.5–12.5 (12)	5.7
Asparagine (Asn, N)	$C_4H_6N_2O_2$	114.04293	114.1039	—	4.4
Aspartic acid (Asp, D)	$C_4H_5NO_3$	115.02694	115.0886	3.9–4.5 (4)	5.3
Cysteine (Cys, C)	C_3H_5NOS	103.00919	103.1448	8.3–9.5 (9)	1.7
Glutamic acid (Glu, E)	$C_5H_7NO_3$	129.04259	129.1155	4.3–4.5 (4.5)	6.2
Glutamine (Gln, Q)	$C_5H_8N_2O_2$	128.05858	128.1308	—	4.0
Glycine (Gly, G)	C_2H_3NO	57.02146	57.0520	—	7.2
Histidine (His, H)	$C_6H_7N_3O$	137.05891	137.1412	6.0–7.0 (6.3)	2.2
Isoleucine (Ile, I)	$C_6H_{11}NO$	113.08406	113.1595	—	5.2
Leucine (Leu, L)	$C_6H_{11}NO$	113.08406	113.1595	—	9.0
Lysine (Lys, K)	$C_6H_{12}N_2O$	128.09496	128.1742	10.4–11.1 (10.4)	5.7
Methionine (Met, M)	C_5H_9NOS	131.04049	131.1986	—	2.4
Phenylalanine (Phe, F)	C_9H_9NO	147.06841	147.1766	—	3.9
Proline (Pro, P)	C_5H_7NO	97.05276	97.1167	—	5.1
Serine (Ser, S)	$C_3H_5NO_2$	87.03203	87.0782	—	6.9
Threonine (Thr, T)	$C_4H_7NO_2$	101.04768	101.1051	—	5.8
Tryptophan (Trp, W)	$C_{11}H_{10}N_2O$	186.07931	186.2133	—	1.3
Tyrosine (Tyr, Y)	$C_9H_9NO_2$	163.06333	163.1760	9.7–10.1 (10.0)	3.2
Valine (Val, V)	C_5H_9NO	99.06841	99.1326	—	6.6

From-Current Protocols in Protein Science, Appendix I, Useful Data, 2000

De Novo Sequencing (cont.)

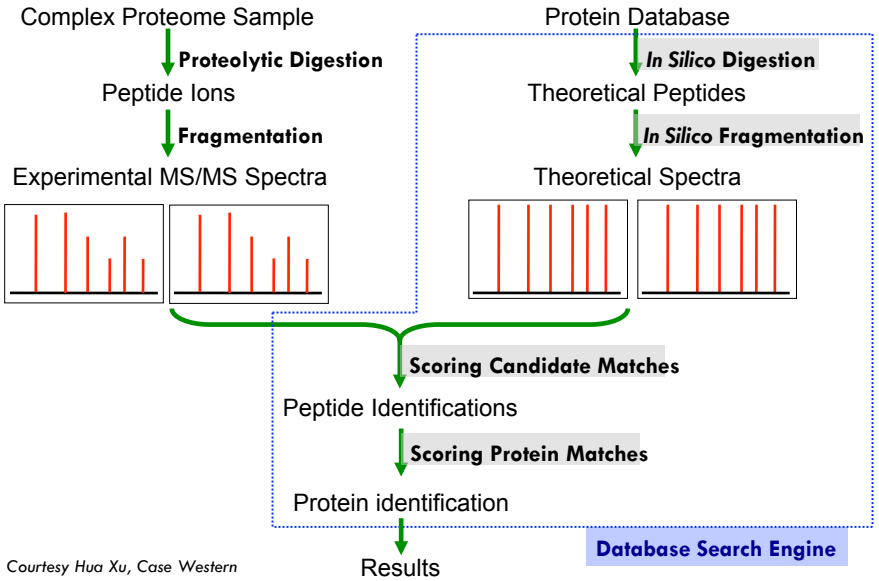
19

- Compare the mass difference between peaks and identify the change (amino acid side chain "R" group)

S-P-A-F-D-S-I-M-A-E-T-L-K
(protonated mass 1410.6)

	mass+	b-ions	y-ions	mass+
	88.1	S	PAFDSIMAETLK	1323.6
88.1 + 97.1 =	185.2	SP	AFDSIMAETLK	1226.4
185.2 + 71.1 =	256.3	SPA	FDSIMAETLK	1155.4
.	403.5	SPAF	DSIMAETLK	1008.2
.	518.5	SPAFD	SIMAETLK	893.1
.	605.6	SPAFDS	IMAETLK	806.0
.	718.8	SPAFDSI	MAETLK	692.3
.	850.0	SPAFDSIM	AETLK	561.7
.	921.1	SPAFDSIMA	ETLK	490.6
.	1050.2	SPAFDSIMAE	TLK	361.5
.	1151.3	SPAFDSIMAET	LK	260.4
	1264.4	SPAFDSIMAETL	K	147.2

Database Search of Tandem MS Data



Courtesy Hua Xu, Case Western

20

Sample Preparation is Key to Success

21

- Guest lecturer at UIC proteomics workshop estimated that successful proteomics was 40% sample prep, 20% instrumentation and 40% bioinformatics.
- There is no "one size fits all" purification scheme.
- Classic approaches work well for some circumstances but most projects require customization and creativity.
- Consult with your proteomics core before spending months preparing your sample so they don't tell you they can't do it that way!

Major Challenge of Proteomics

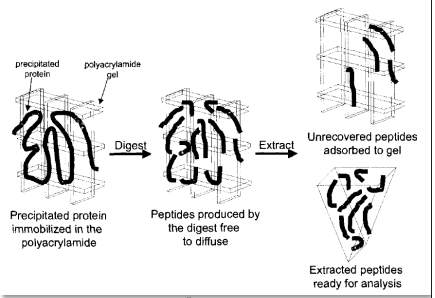
22

- **There is no PCR for proteins!**
- You can't amplify a low abundance signal like you can in the genomics world
- Mass specs only see ions and don't know "good ions" from bad
- Contamination with keratin, detergents or PEG is a major obstacle in classic proteomics approaches
  - Avoid handling gels or the instrument will spend most of it's time sequencing keratin and not see your low abundance protein!
- Just because you "see" something in a Western doesn't mean you've got enough protein to analyze (nor that it's pure)!

Classic Proteomics Approach

23

- SDS PAGE is a tried and true step for proteomics of individual proteins or samples of low complexity
- Concentrates and traps the protein in the gel, then wash away SDS and salts.
- Reduce and alkylate in gel, dehydrate, infuse trypsin then digest overnight and extract peptides



From-Kinter & Sherman, "Protein Sequencing and Identification Using Tandem Mass Spectrometry", Wiley, 2000

Challenges for proteomics

24

Sample preparation

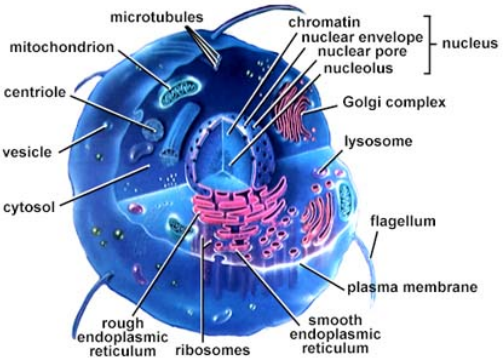
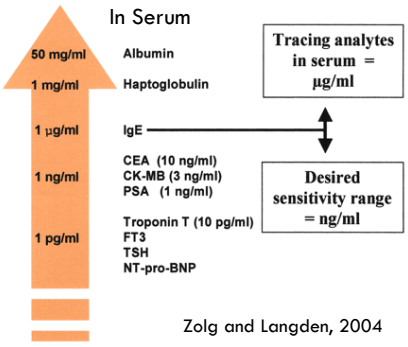


Image from: On-line biology
http://www2.estrellamountain.edu/faculty/farabee/biobk/biobooktoc.html

Protein dynamic range



Zolg and Langden, 2004

Courtesy-Dr. K. Kristjansdottir Midwestern University

Working with Complex Proteomes

25

- Even the best modern mass specs can only see 10~20 ions at a time
  - Anything else eluting at the same time is lost (not sequenced).
- Need to reduce the complexity by spreading out peptides.
- **MudPIT** (multidimensional protein identification technology) is a non-gel approach for the identification of proteins from complex mixtures which involves separation by multiple physical properties (usually charge and hydrophobicity)

MudPIT

26

- Classic approach was to pack a custom column with SCX resin over C18 resin.
- Peptides are eluted with salt "bumps" from SCX then fractionated on C18; repeat
- UIC core prefers to use a preparative isoelectric focusing apparatus (Agilent 3100 OffGel) to separate by charge in solution then run each fraction as a separate mass spec run on C18 HPLC
- UIC core identifies thousands of proteins in a single experiment with this appraoch

Link, AJ and LaBer, J, "Proteomics: A Cold Spring Harbor Laboratory Course Manual", 2009

CFTR - A Protein Misfolding Disease

27

- Yates lab* characterized CFTR and ΔF508 CFTR interactome
- CFTR present ~100 copies/cell
- MudPIT bottom-up analysis identified ~3000 proteins
- Used gene ontology and comparison with WT to identify 638 high confidence CFTR interactors, with 208 unique to ΔF508

➢ Develop RNAi and chemicals to restore interactome

*Pankow, et al. "ΔF508 CFTR interactome remodelling promotes rescue of cystic fibrosis." Nature. 2015 Dec 24;528(7583):510-6.

FASP - Sample Prep in a Tube

28

- Filter-aided sample preparation (FASP)
  - Wiśniewski JR, et al., "Universal sample preparation method for proteome analysis." Nat Methods. 2009 May;6(5): 359-62.
  - Lyse cells w/ SDS, transfer to tube with 10K filter, exchange SDS for 8 M urea, dilute to 2 M urea & add Lys C, dilute to ~1 M & add Trypsin then spin to collect peptides in effluent.
  - Modified to Stage Tips (C18) (Kulak, et al., Nature Methods 11 (March 2014): 319-324
  - ➢ Obtain copy-number estimates for 9667 human proteins

Detecting Post-translational Modifications

29

- The ability to detect PTMs gives proteomics an **advantage** over genomics-related approaches because they reflect the functional state of a protein
- Almost all eukaryotic protein sequences are post-translationally modified and **hundreds** of types of modifications of amino acid residues are known.
- Cf-Aebersold R & Mann M, "Mass-spectrometric exploration of proteome structure and function." Nature. 2016 Sep 15;537(7620):347-55. PMID: 27629641

Post-Translational Modifications modify the residue mass

30

- In principle you just tell the search engine to search for Ser +/- 80 Da, etc.
- Some PTMs (e.g., phosphorylation) require specialized mass spec techniques and bioinformatics
  - See unimod.org

Table A.1A.3 Mass Changes Due to Some Post-Translational Modifications of Peptides and Proteins[a]

Modification[b]	Monoisotopic mass change	Average mass change
Common modifications		
Pyroglutamic acid formation from Gln	−17.0265	−17.0306
Disulfide bond (cystine) formation	−2.0157	−2.0159
C-terminal amide formation from Gly	−0.9840	−0.9847
Deamidation of Asn and Gln	−0.9840	−0.9847
Methylation	14.0157	14.0269
Hydroxylation	15.9949	15.9994
Oxidation of Met	15.9949	15.9994
Proteolysis of a single peptide bond	18.0106	18.0153
Formylation	27.9949	28.0104
Acetylation	42.0106	42.0373
Carboxylation of Asp and Glu	43.9898	44.0098
Phosphorylation	79.9663	79.9799
Sulfation	79.9568	80.0642
Cysteinylation	119.0041	119.1442
Glycosylation with pentoses (Ara, Rib, Xyl)	132.0423	132.1161
Glycosylation with deoxyhexoses (Fuc, Rha)	146.0579	146.1430
Glycosylation with hexosamines (GalN, GlcN)	161.0688	161.1577
Glycosylation with hexoses (Fru, Gal, Glc, Man)	162.0528	162.1424
Modification with lipoic acid (amide bond to lysine)	188.0330	188.3147
Glycosylation with *N*-acetylhexosamines (GalNAc, GlcNAc)	203.0794	203.1950
Farnesylation	204.1878	204.3556
Myristoylation	210.1984	210.3598
Biotinylation (amide bond to lysine)	226.0776	226.2994
Modification with pyridoxal phosphate (Schiff base to lysine)	231.0297	231.1449
Palmitoylation	238.2297	238.4136
Stearoylation	266.2610	266.4674
Geranylgeranylation	272.2504	272.4741
Glycosylation with *N*-acetylneuraminic acid (sialic acid, NeuAc, NANA, SA)	291.0954	291.2579
Glutathionylation	305.0682	305.3117
Glycosylation with *N*-glycolylneuraminic acid (NeuGe)	307.0903	307.2573
5′-Adenosylation	329.0525	329.2091
Modification with 4′-Phosphopantetheine	339.0780	339.3294
ADP-ribosylation (from NAD)	541.0611	541.3052
Adventitious modifications		
Acrylamide	71.0371	71.0788
Glutathione	304.0712	304.3038
2-Mercaptoethanol	75.9983	76.1192

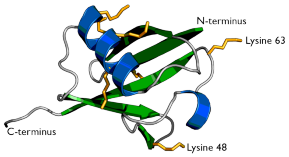
[a]To obtain the molecular mass of a modified peptide or protein, the appropriate mass changes should be algebraically added to the molecular mass calculated for the unmodified molecule.
[b]A more extensive list of modifications is available from the Delta mass site at *http://www.mbcf.dfci.harvard.edu/docs/DeltaMass.html*.

From Current Protocols in Protein Science, Appendix 1, Useful Data, 2000

Example: Ubiquitination

31

- Ubiquitin is a small protein (8564.47 Da) that gets conjugated to the ε -amino group of lysine via the COOH-terminus leading to protein degradation
- COOH terminal sequence is: …Q**K**ESTLHLVL**R**L**R**GG
- Trypsinization of complex leaves GG

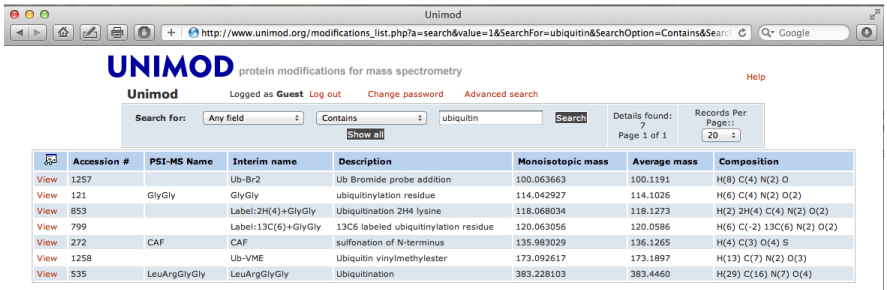


Source- Wikipedia

- But-Beware of possible artifact if use iodoacetamide to alkylate (57Da on K + 57 Da on K looks like 114 Da if there are two K's in partial-cleavage peptide)
  - *Neilsen, et al, "Iodoacetamide-induced artifact mimics ubiquitination in mass spectrometry." Nat Methods. 2008 Jun;5(6):459-60.*

UNIMOD Database

- http://www.unimod.org/ (login as guest)



	Accession #	PSI-MS Name	Interim name	Description	Monoisotopic mass	Average mass	Composition
View	1257		Ub-Br2	Ub Bromide probe addition	100.063663	100.1191	H(8) C(4) N(2) O
View	121	GlyGly	GlyGly	ubiquitinylation residue	114.042927	114.1026	H(6) C(4) N(2) O(2)
View	853		Label:2H(4)+GlyGly	Ubiquitination 2H4 lysine	118.068034	118.1273	H(2) 2H(4) C(4) N(2) O(2)
View	799		Label:13C(6)+GlyGly	13C6 labeled ubiquitinylation residue	120.063056	120.0586	H(6) C(-2) 13C(6) N(2) O(2)
View	272	CAF	CAF	sulfonation of N-terminus	135.983029	136.1265	H(4) C(3) O(4) S
View	1258		Ub-VME	Ubiquitin vinylmethylester	173.092617	173.1897	H(13) C(7) N(2) O(3)
View	535	LeuArgGlyGly	LeuArgGlyGly	Ubiquitination	383.228103	383.4460	H(29) C(16) N(7) O(4)

How Many Proteins in our Proteome?

33

- Estimates of the number of human genes have dropped from 100's of thousands to < 21,000
- The number of proteins is based on the proteome
- Numbers of proteins also vary based on alternative splicing
- Several labs report >10,000 protein IDs

Pennisi, E. "Working the (gene count) numbers: finally, a firm answer?" Science. 2007 May 25;316(5828):1113.

- ENCODE* found 20,687 protein-coding genes with 6.3 alternatively spliced transcripts per locus, whose coding exons encompass 1.22% of the genome (20,687 X 6.3 = 130,328.1 proteins)
- UniProtKB+ lists 131,333 human proteins, of which 68,079 represent the "complete proteome" (as of 11/28/12).

*30 papers published 6 Sept 2012 at http://nature.com/encode
+http://www.uniprot.org/

Polymorphic Proteomics

Just because things look alike doesn't mean they're the same!

34

Polymorphism challenges conventional proteomic analysis

35

- Variations in primary sequences aren't reflected in "consensus sequence" in protein databases used for routine sequence searches.
  - Polymorphic sequences are routinely missed by search engines and clinically relevant information is lost.
  - **RHVFGESDELIGQK** is recognized but **RHVFG*D*SDELIGQK** is not recognized.
- ➢ Need to alter protein databases to include all known polymorphisms and alternate splicings.

SwissProt to the rescue!

36

- The clever bioinformaticians at the European Bioinformatics Institute have a solution
- Their curated SwissProt Knowledgebase database includes all published reports of polymorphism and splice variants in the record for each protein
  - Click the link to see detailed information about each variant
- You can search the variants directly through the SwissVar portal at http://swissvar.expasy.org/

How about some automation?

37

- SwissProt provides **Splicevar,** a Perl program that creates separate records for each reported single amino acid substitution or splice variant
- Splicevar uses **SwissKnife,** a Perl library for manipulating and querying SwissProt fields
- Splicevar took 8+ hours to generate the fully annotated database and reference file, growing SwissProt from 250 MB to 1.4 GB
- Proteomics search engines like Mascot, MaxQuant & ProSight can use the SwissProtVarsplic-modified database to identify polymorphic proteins
  - Server admin needs to update Mascot databases

Use case: TPI variants

38

- Dr. Alfredo Torres-Larios sent UIC core samples for analysis
- After initial ID of TPI he wrote to ask us if we could look for an E→D point mutation. Initially I added the full sequence for VAR_007536 at the end of the database and we successfully ID'd the mutant in the presence of normal peptides (~25%).
- After running Swissvar to build the polymorph database we automatically identified VAR_0749 and found evidence for another, unexpected variant.

P60174-00-03-00 (100%), 26,655.1 Da
1 TPI deficiency-VAR_007536 Displayed of P60174 OS=Homo sapiens GN=TPI1
2 unique peptides, 2 unique spectra, 3 total spectra, 178/249 amino acids (71% coverage)

MAPSRKFFVG	GNWKMNGRKQ	SLGELIGTLN	AAKVPADTEV	VCAPPTAYID
FARQKLDPKI	AVAAQNCYKV	TNGAFTGEIS	PGMIKDCGAT	WVVLGHSERR
HVFCDSDELI	GQKVAHALAE	GLGVIACIGE	KLDEREAGIT	EKVVFEQTKV
IADNVKDWSK	VVLAYEPVWA	IGTGKTATPQ	QAQEVHEKLR	GWLKSNVSDA
VAQSTRIIYG	GSVTGATCKE	LASQPDVDGF	LVGGASLKPE	FVDIINAKQ

Aguirre B, Costas M, Cabrera N, Mendoza-Hernández G, Helseth DL Jr., et al. 2011 A Ribosomal Misincorporation of Lys for Arg in Human Triose-phosphate Isomerase Expressed in Escherichia coli Gives Rise to Two Protein Populations. PLoS ONE 6(6): e21035. doi:10.1371/journal.pone.0021035

Proteomics/Translational Databases

39

- Larry's favorite database is **NextProt** (www.nextprot.org)
  - ***Human only*** but includes views of expression, structural and medical information
- **SwissVar** (swissvar.expassy.org)
  - Polymorphisms
- **UniProt KB** (www.uniprot.org)
  - In silico cleavage tools
- **OMIM** (omim.org)
  - On-line Mendelian Inheritance in Man, Victor McKusick
- **NCBI** (www.ncbi.nlm.nih.gov/protein/)
- **Ensembl** (www.ensembl.org)
- **Peptide Atlas** (www.peptideatlas.org)

Proteomics/Translational Databases (cont)

40

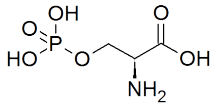
- **Human Protein Atlas** Microarray-based immunohistochemistry catalog of 24,000 antibodies corresponding to ~17,000 protein-encoding genes across 44 major tissues and organs.
  - Science. 2015 Jan 23;347(6220):1260419. doi: 10.1126/science. 1260419.
  - http://www.proteinatlas.org/
- **Human Proteome Maps**:
  - Nature 509 (29 May 2014): 575-581
    - (http://www.humanproteomemap.org/) &
  - Nature 509 (29 May 2014): 582-587
    - ProteomicsDB: http://www.proteomicsdb.org/).
- **Cancer Proteome Atlas**, based on reverse-phase protein array quantitation during The Cancer Genome Atlas project:
  - Nature Methods (Nov 2013): 1046-1047
  - Nature Communications, 5 (29 May 2014): 3887 "A pan-cancer proteomic perspective on The Cancer Genome Atlas"
  - http://app1.bioinformatics.mdanderson.org/tcpa/_design/basic/index.html
- **ProteomeXchange** (http://proteomecentral.proteomexchange.org/)

NextProt (www.nextprot.org)

41

- Gaudet P, et al., "The neXtProt knowledgebase on human proteins: 2017 update" Nucleic Acids Research, 2017, Vol. 45, Database issue D177–D182
- Added Variant Portal with includes over 8000 phenotypic observations for over 4000 variations in a number of genes involved in hereditary cancers and channelopathies.

I Want My PTMs!

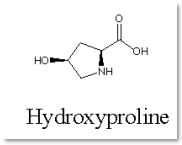


42

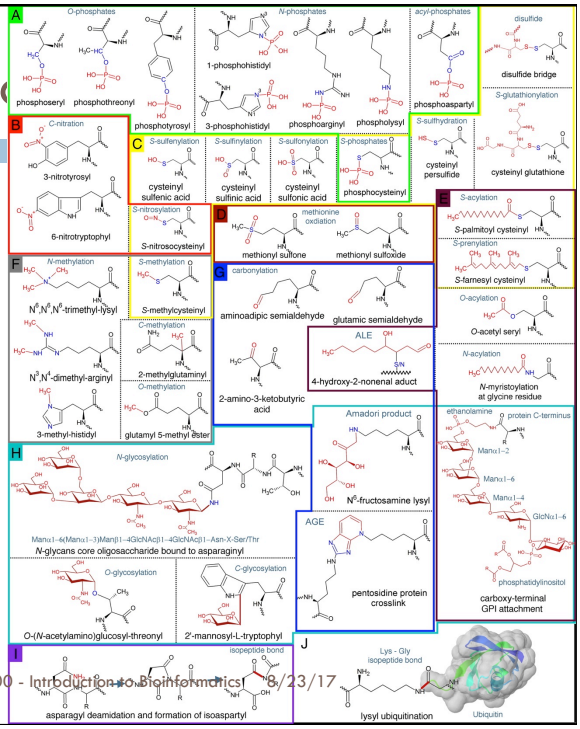
- Phosphopeptides are the most frequently requested PTM at core facilities
- Regulatory phosphopeptides are present at low abundance
- Phosphopeptides are the hardest PTM to identify because the phosphate group just falls off!
- The mass spec sees a "neutral loss" of 80 Da but you don't get any sequence information about the peptide which just lost the phosphate.
- Alternates: Fragment again ("MS3" protocol) or use different fragmentation schemes (ECD, HCD, ETD)
- Ultimately, it's just a mass difference but check with your core about the chemistry, fragmentation, etc.

Chemical stru

43



Cerný M, Skalák J, Cerna H, Brzobohatý B.
Journal of Proteomics, 15 June 2013
PMID: 23777897



Dynamic Range Issues

44

- Low abundance things *usually don't get selected for MS/MS*
- Typical problem with normal instrument methods:
  - A PTM peptide needs to be present at >1% of the total ions coming out at that time in order to be selected
- Avoid FCS (10% FCS ≅ 10 mg/ml albumin, etc.)
- **Have enough sample, and use sample prep to separate it well!**

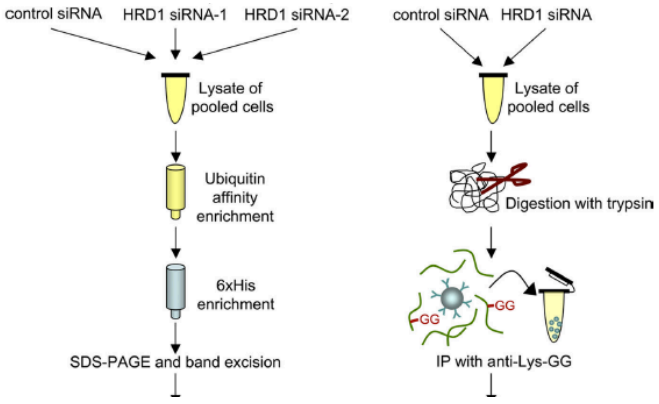
Enrichment Methods

45

- **Best method:**
  - 1) covalent linkage between tag/antibody and bead
    - If not, then antibody/tag can elute with protein, coats the SDS-PAGE gel and is often the major identified protein from gel (instead of the protein of interest)
  - 2) Magnetic beads (low background, fast binding)
- **Types of enrichments:**
  - Immunoprecipitation
    - Antibody against PTM or protein of interest
  - Bio-affinity purifications
    - Clone a tag into your protein of interest
      - His, FLAG, HA, Strep, TAP

Examples using affinity approach

46



Lee, et al., "Ubiquitin Ligase Substrate Identification through Quantitative Proteomics at Both the Protein and Peptide Levels", J Biol Chem, Dec. 2, 2011

Enrichment Methods (cont)

47

- Know your binding capacity
  - If known: For every 1 mg total protein applied, use ~150 pmol binding capacity's worth of beads for your pull-down
  - Sometimes you don't know the molar capacity for IPs, so you have to optimize
- Still do another separation after pull-down
  - e.g. elute pull-down beads using SDS sample loading buffer, run on 1D gel, look for bands or just cut into equally-sized pieces
- Consider isotope labeling to tell the difference between non-specific background and your specific binders

Phosphoproteomics

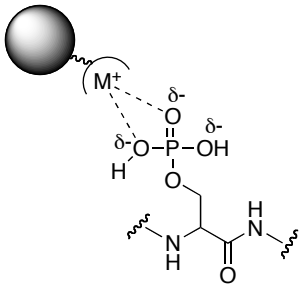
48

- To study the 'Omics of phosphoproteins requires specialized purification techniques
- Most investigators use either IMAC (Immobilized Metal Affinity Chromatography) or TiO_2 or some combination of the two to enrich for phosphopeptides prior to characterization
- This can mean you identify individual phosphopeptide sequences but may not be able to tell which protein they came from

Phosphoproteomics (cont)

49

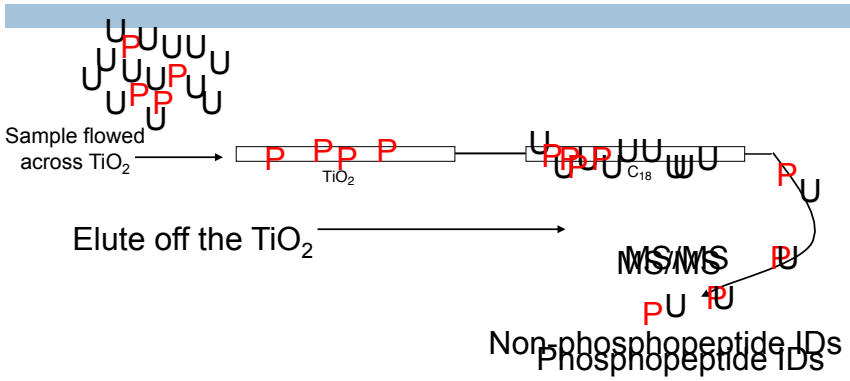
- **Phospho-enrichment**
  - Chemical properties of phosphate



  - IMAC, TiO_2, Phos-Tag, **Pro-Q Diamond**

TiO2 Enrichment of Phosphopeptides

50

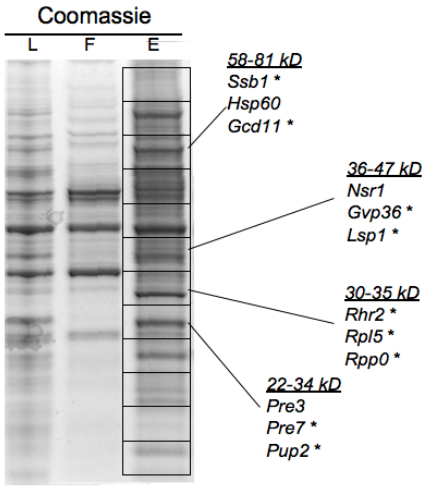


Need at least 1 pmol of phospho material(s) for best chance of ID. For low % phospho, that means >100 pmol total of that protein

1 pmol of 60 kD protein = 60 ng

Pro-Q Diamond resin phosphoprotein affinity pull-down

51



L, Lysate; F, Flow through; E, Eluate

Kristjansdottir, et al. "Phosphoprotein Profiling by PA-GeLC-MS/MS" Journal of Proteome Research 2008, 7, 2812–2824

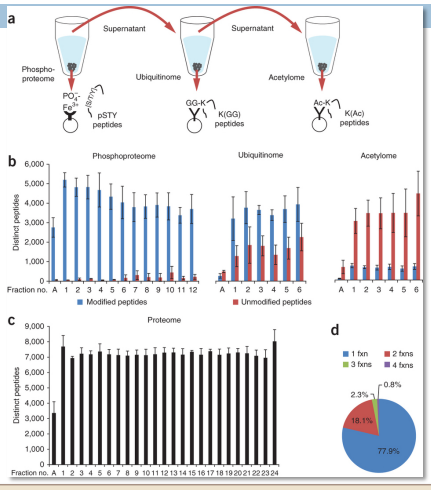
Enrich at the protein or the peptide level?

52

- Protein
  - Pros: allows you to further separate by gel before digestion, you keep all the other peptides from the protein, better overall ID
  - Cons: Sample still very complex, low abundance mods still hard to see
- Peptide
  - Pros: only thing left in sample means more likely to be seen
  - Cons: only thing left in sample means that's the only evidence you have for that protein
- For best experiment, you need to do both

Example of Serial Enrichment

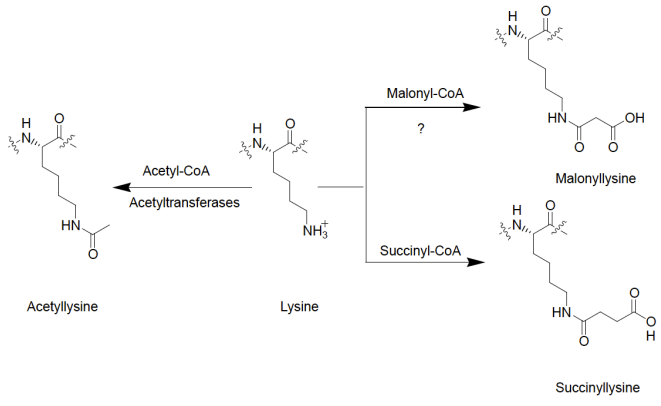
53



Mertins P, Qiao JW, Patel J, Udeshi ND, Clauser KR, Mani DR, Burgess MW, Gillette MA, Jaffe JD, Carr SA. Nat Methods. 2013 Jul;10(7):634-7.

We're still discovering PTMs

54



Peng, et al., "The first identification of lysine malonylation substrates and its regulatory enzyme", Molec Cellular Proteomics, Ahead of print 9/9/11

PTM's role in epigenetics

55

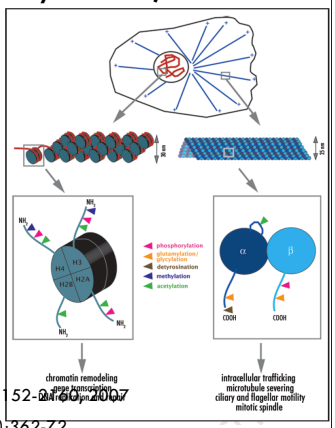
Special issue of **Molecular & Cellular Proteomics** dedicated to "Chromatin Biology and Epigenetics", March 2016; 15 (3)

- From http://www.mcponline.org/content/15/3.cover-expansion

PTMs-Exploring the Tubulin Code

56

- Analogous to the "histone code" of acetylations, phosphorylations, and methylations
- Controls tubulin structural dynamics



Verhey & Gaertig, "The Tubulin Code ", Cell Cycle 6:17, 2152-2160, 2007

Janke & Kneussel, Trends Neurosci. 2010 Aug;33(8):362-72.

Identifying new tubulin PTMs

57

- Yuyu Song, PhD (UIC) with Dr. Scott Brady, UIC Cell Biology & Anatomy
- Transglutaminase modification of tubulins lead to long term stability
- Isolate modified tubulins using IEF then SDS-PAGE
- Found modifications of specific sites near α-β interface

Song Y, Kirkpatrick LL, Schilling AB, Helseth DL, Chabot N, Keillor JW, Johnson GV, Brady ST. Neuron. 2013 Apr 10;78(1):109-23

Informatics for proteomics

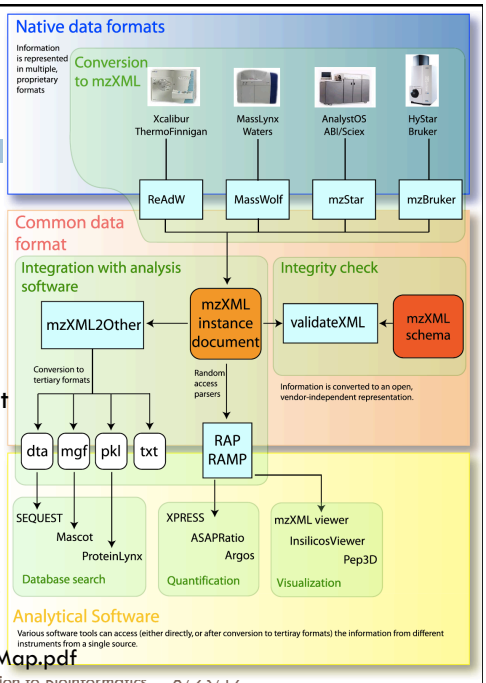
58

- All vendors use Windows-based software to run instruments, so convert data files to text for parsing
- Text files can be either peak list (mass & intensity) or XML + graphs (mzXML)

Tower of Babel

59

- Search engines like Mascot and X! Tandem require that your spectral files be converted from .RAW to another format before submitting. Mascot requires MGF files ("Mascot Generic Format"), while X!Tandem and other search engines require that you convert your files to the mzXML or mzML format.
- Links to conversion tools are at http://tools.proteomecenter.org/wiki/index.php?title=Formats:mzXML



http://sashimi.sourceforge.net/extra/softwareMap.pdf

Search engines

60

- Core labs use commercial search engines like Mascot
- Also use open source search engines: MaxQuant, X! Tandem, TPP
  - http://maxquant.org/
  - http://tools.proteomecenter.org/wiki/index.php?title=Main_Page
- Thermo products: Bioworks and Proteome Discoverer
- Specialty search engines for disulfide bonds, post-translational modifications, etc.
- Use Scaffold (commercial editor required) to prepare user friendly results from Mascot results

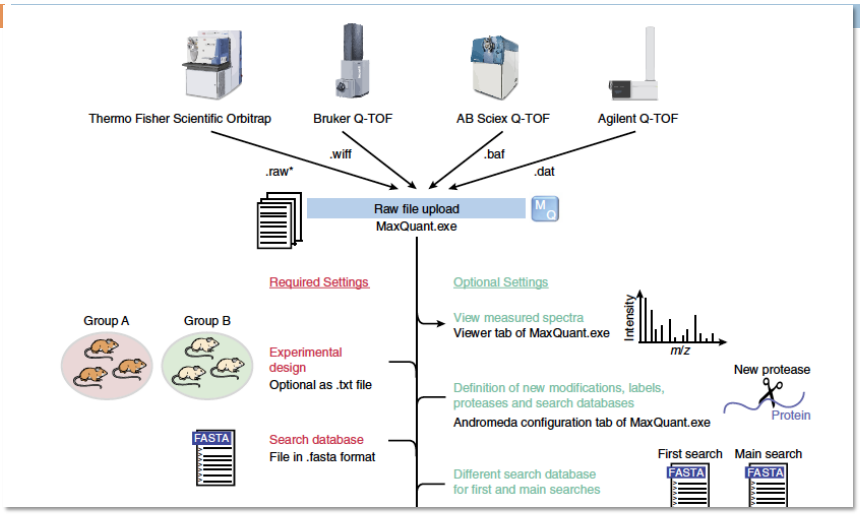
MaxQuant

61

- U of C proteomics core lab recommends using MaxQuant for quantitation work
  - Tyanova S, Temu T, Cox J, "The MaxQuant computational platform for mass spectrometry-based shotgun proteomics." Nat Protoc. 2016 Dec;11(12):2301-2319. PMID: 27809316
  - The MaxQuant algorithms are efficiently parallelized on multiple processors and scale well from desktop computers to servers with many cores. The software is written in C# and is freely available at http://www.maxquant.org.
    - Works on Thermo RAW files; no need for conversion.
- Runs on Windows desktop/server

MaxQuant workflow

62



What do you find?

63

- Core labs typically ID a protein based upon 10% or more of sequence (unless you give them tons of a pure protein)
- The Scaffold ID is statistically validated at 95% confidence in the protein ID, 95% confidence in each peptide assignment, and requires 2 or more unique peptides
- Search against a consensus protein database for your species unless told otherwise.
- Routinely identify >1500 proteins from a single MudPIT experiment so use Scaffold to visualize, show sequence coverage and Gene Ontology info for each protein.

http://www.proteomesoftware.com/products/free-viewer/

Working with Protein Databases

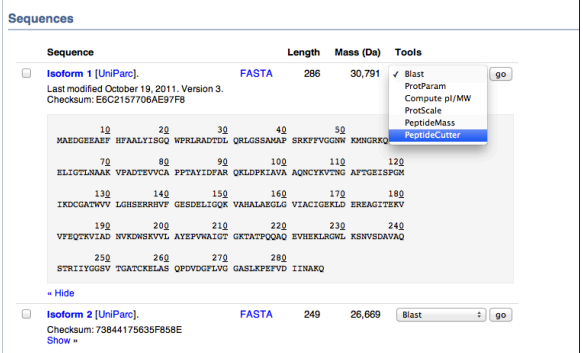
64

- Core facilities maintain common NCBI and UniProt databases on Mascot, MassMatrix and X!Tandem servers
- MassMatrix allows users to upload custom databases
- NCBI website allows you to search the protein database for a species, change to FASTA display then download all protein sequences
- UniProtKB has similar feature

Which Enzyme Should I Use?

65

- UniProt allows you to do in silico cleavage
- Search for your protein of interest then change to PeptideCutter and select trypsin
- Check resulting peptides to see if target area will be visible. If not, change enzyme.



Source: http://www.uniprot.org/uniprot/P60174

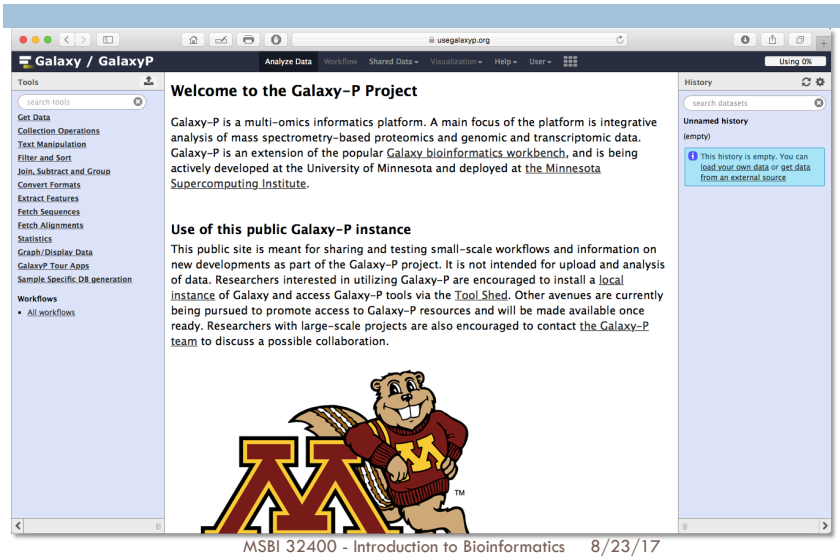
Proteomics Databases & Resources

66

- www.thehpp.org (Human Proteome Project)
- www.c-hpp.org (Chromosome-centric Human Proteome Project)
- www.humanproteomemap.org - Kim MS, et al. Nature. 2014 May 29;509(7502):575-81 PMID: 24870542
- www.proteomicsDB.org (93% of human proteome) - Wilhelm M, et al. Nature. 2014 May 29;509(7502):582-7. PMID: 24870543
- www.proteinatlas.org - Protein expression profiles based on immunohistochemistry for a large number of human tissues, cancers and cell lines
- Cell-surface proteome coming soon
- **Galaxy-P** (usegalaxyp.org) – **Galaxy for Proteomics** (Tim Griffin, U MN)
  - Sheynkman, et al. "Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. BMC Genomics. 2014 Aug 22;15:703. PubMed PMID: 25149441
  - Jagtap PD, et al. "Flexible and accessible workflows for improved proteogenomic analysis using the Galaxy framework." J Proteome Res. 2014 Oct 10. [Epub ahead of print] PubMed PMID: 25301683.

Galaxy-P for Proteomics (usegalaxy-p.org)

67



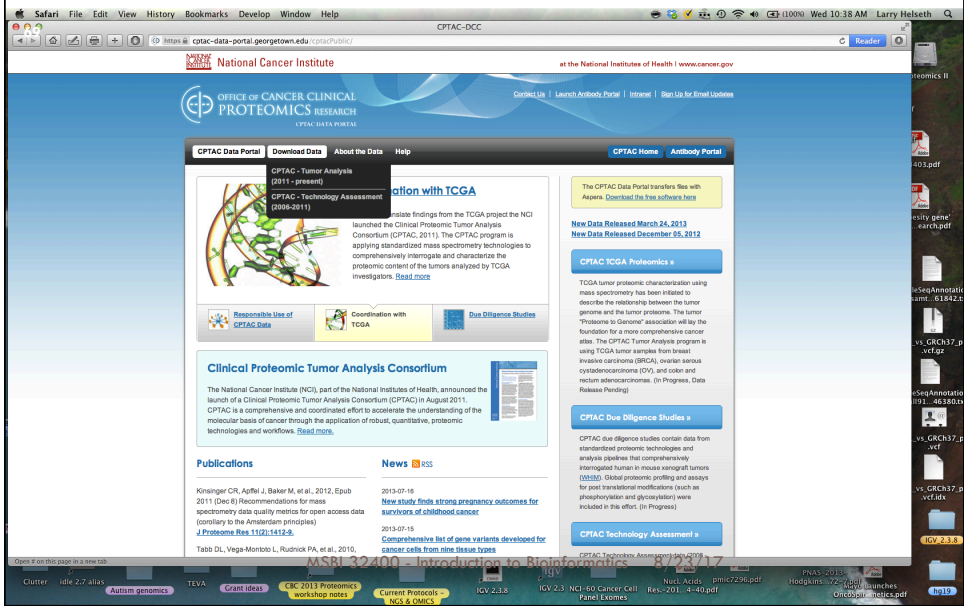
Finding data for software development and testing

68

Proteomics Data Repositories:

- www.proteomexchange.org
- PRIDE (www.ebi.ac.uk/pride) - Search for MS/MS spectra
- www.peptideatlas.org (w/ SRM Atlas, etc) – MS/MS Spectra library
- cptac-data-portal.georgetown.edu/cptacPublic/ - TCGA CPTAC Cancer proteome (cf- proteomics.cancer.gov)
- TCGA has Parallel proteomics projects
  - http://cancergenome.nih.gov/abouttcga/overview/howitworks/proteomecharacterization
  - Clinical Proteomic Tumor Analysis Consortium (CPTAC)
  - http://proteomics.cancer.gov with Data Portal

TCGA Proteomics Data Portal



MSBI 32400 - Introduction to Bioinformatics 8/23/17

Isotopes and Mass Spec

71

- Naturally abundance of C, N & O complicate things

Atom	Mass	Abundance
12C 13C 14C	12.000000 13.003355 14.003242	98.93 1.07 *
14N 15N	14.003074 15.000109	99.632 0.368
16O 17O 18O	15.994915 16.999132 17.999160	99.757 0.038 0.205

Source: http://www.chem.ualberta.ca/~massspec/atomic_mass_abund.pdf

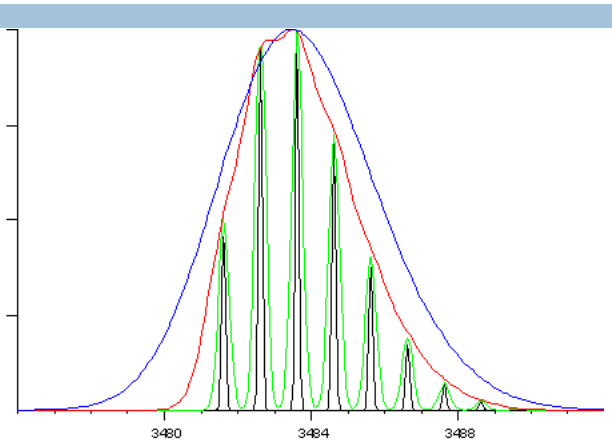
Isotopes complicates MS analysis

72

- A typical peptide will have multiple forms due to natural abundance of each isotope
- A small protein (insulin) gets even more complicated
- A medium sized protein (BSA) is a mess
- Resolving power is critical to working with larger proteins

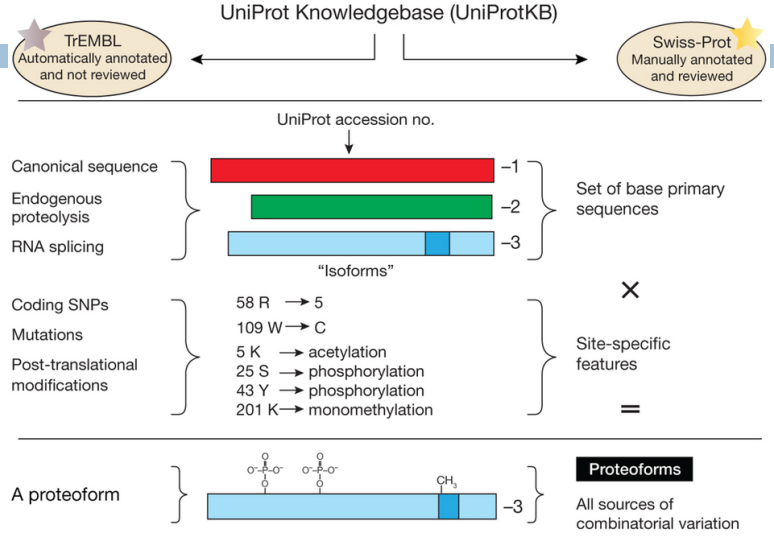
http://www.matrixscience.com/help/mass_accuracy_help.html

73



74

Proteoforms: Protein Variability in Higher Proteomes



Smith LM, Kelleher NL, Nat Methods. 2013 Mar;10(3):186-7

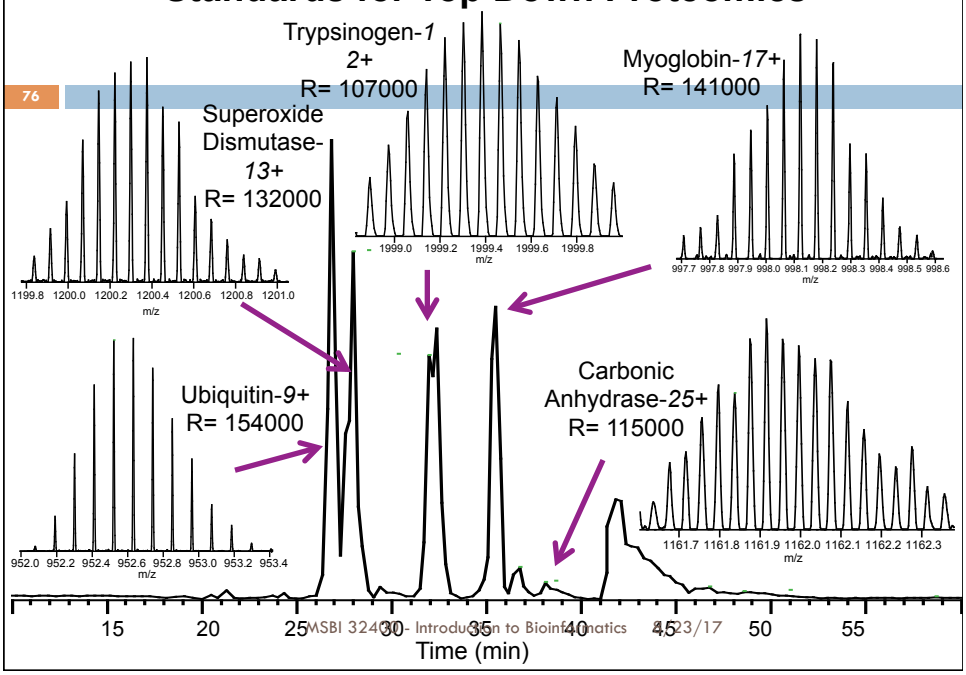
Top-Down Proteomics

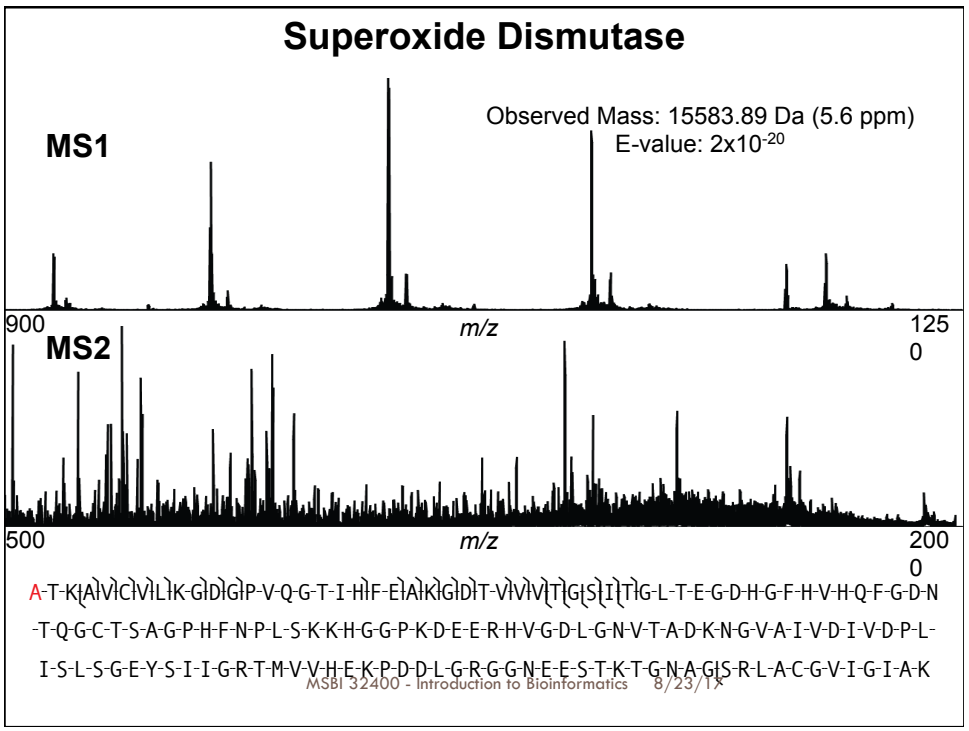
75

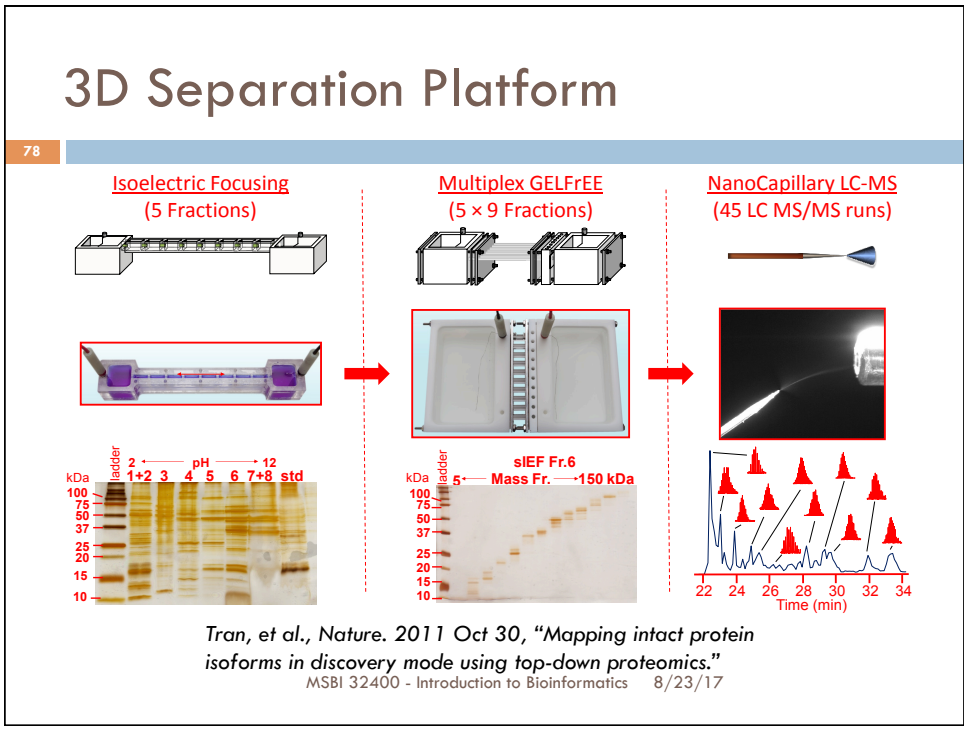
- Fragment intact proteins then use sophisticated informatics to match fragments to known protein sequences
- Requires high resolution mass spec (7T or higher)
- Used to characterize complex signaling isoforms from selected proteins
- "High throughput" protocols and instrumentation still needed for routine screening
- MW range typically limited to <100 KDa without specialized methods or partial cleavage ("middle down")

Standards for Top Down Proteomics

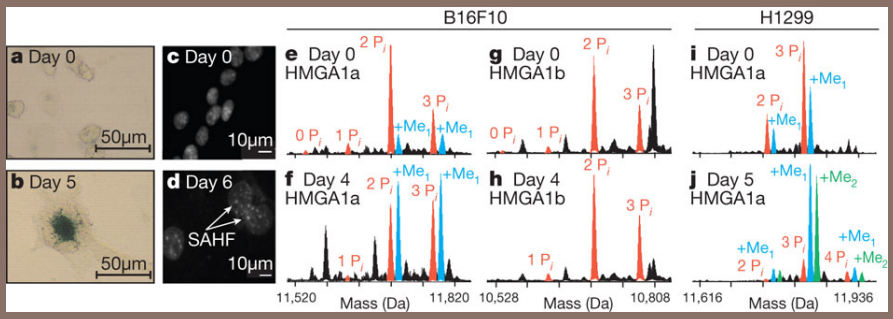
76




Superoxide Dismutase
MS1
Observed Mass: 15583.89 Da (5.6 ppm)
E-value: 2x10-20
900
m/z
1250
MS2
500
m/z
2000
A-T-K-A-V-C-V-L-K-G-D-G-P-V-Q-G-T-I-H-F-E-A-K-G-D-T-V-V-V-T-G-S-I-T-G-L-T-E-G-D-H-G-F-H-V-H-Q-F-G-D-N
-T-Q-G-C-T-S-A-G-P-H-F-N-P-L-S-K-K-H-G-G-P-K-D-E-E-R-H-V-G-D-L-G-N-V-T-A-D-K-N-G-V-A-I-V-D-I-V-D-P-L-
I-S-L-S-G-E-Y-S-I-I-G-R-T-M-V-V-H-E-K-P-D-D-L-G-R-G-G-N-E-E-S-T-K-T-G-N-A-G-S-R-L-A-C-G-V-I-G-I-A-K
MSBI 32400 - Introduction to Bioinformatics 8/23/17


3D Separation Platform
78
Isoelectric Focusing
(5 Fractions)
Multiplex GELFrEE
(5 × 9 Fractions)
NanoCapillary LC-MS
(45 LC MS/MS runs)
kDa
ladder
pH
1+2 3 4 5 6 7+8 std
sIEF Fr.6
Mass Fr.
150 kDa
Time (min)
Tran, et al., Nature. 2011 Oct 30, "Mapping intact protein isoforms in discovery mode using top-down proteomics."
MSBI 32400 - Introduction to Bioinformatics 8/23/17

Monitoring dynamics of HMGA1 isoforms during senescence in B16F10 and H1299 cells.



JC Tran *et al. Nature* **000**, 1-5 (2011) doi:10.1038/nature10575 nature

Top down requires different software

80

- High-throughput search using ProSightPC
- Developed by Kelleher lab and licensed to Thermo
- Works on Thermo RAW files directly
- Available for PC and as web interface
- NU's Proteomics Center of Excellence uses a Varsplic database for ProSight to cover known polymorphisms

ProSightPC™ 2.0
Software for Precision Proteomics

Quantitative Proteomics

81

Different approaches:

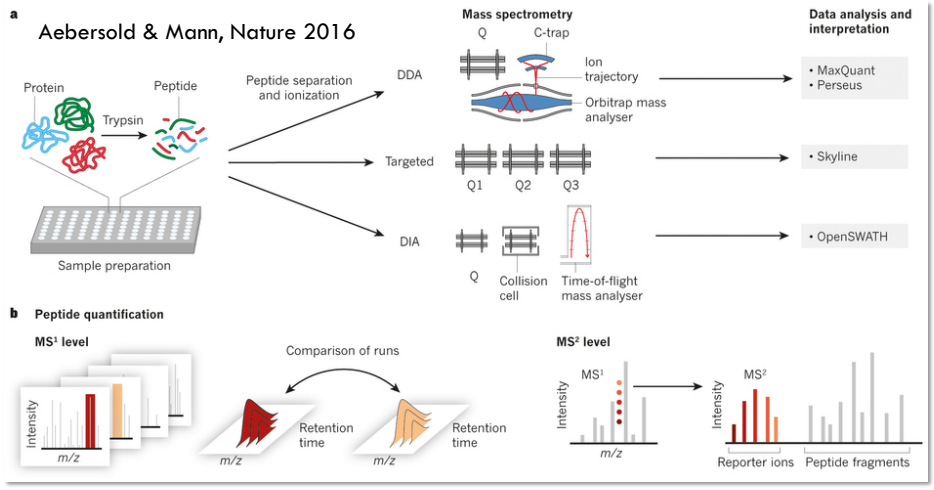
- Spectral counting (label-free)
- SILAC
- Isobaric tags
- SRM (biomarkers)

Spectral counting

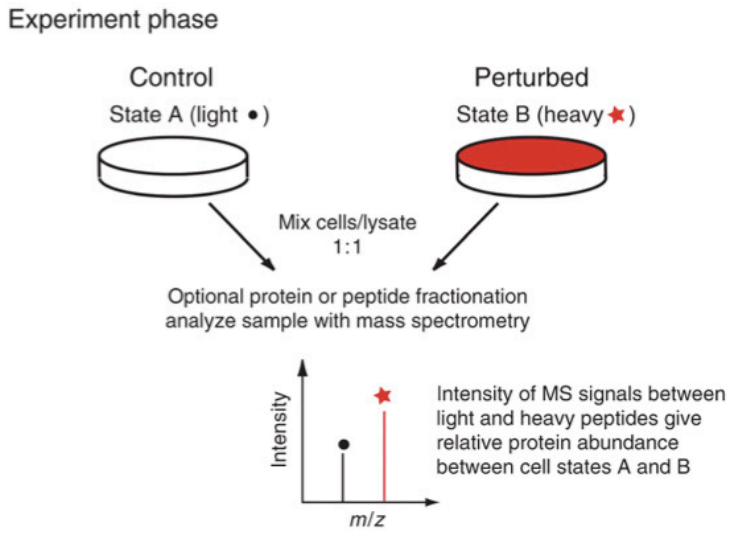
82

- Quantitation based on how many times peptides from each protein are seen in a run
- Requires triplicate (at minimum) runs with blanks in between to minimize carry-over
- Requires dedicated instrument: A recent study of post-synaptic density proteome required two months of dedicated mass spec time and years of informatics
  - *Nanavati D, et al., "The effects of chronic treatment with mood stabilizers on the rat hippocampal post-synaptic density proteome." J Neurochem. 2011 Nov;119(3):617-29.*

83



SILAC = "Stable Isotope Labeling of Cell Culture"



Ong S. E. and Mann M., Nature Protocols, 2007, 1, 2650-2660

84

Pros and Cons of SILAC

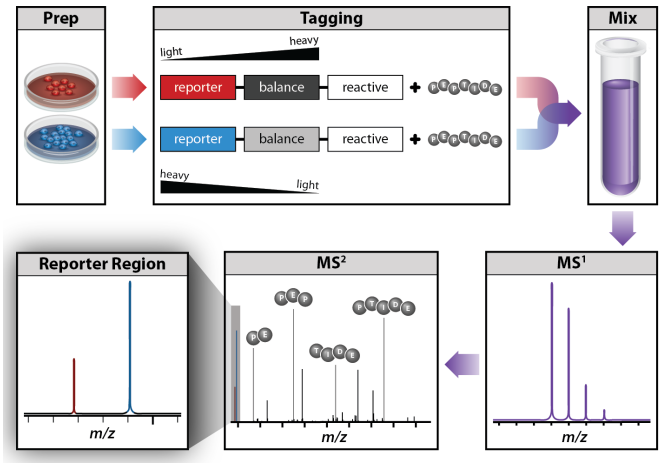
CONS:

- ❑ Costly (need isotopic Lys &/or Arg, dialyzed FCS & custom medium)
- ❑ May require several weeks to reach optimal label incorporation
- ❑ Some conversion of Arg -> Pro complicating analysis
- ❑ Ultimately limited to binary (+/-) or in some cases 3-plex but must avoid overlapping isotopic envelopes (>4 Da apart)
- ❑ Cannot be done on biopsy material *(can do ^{18}O labeling)*

PROS:

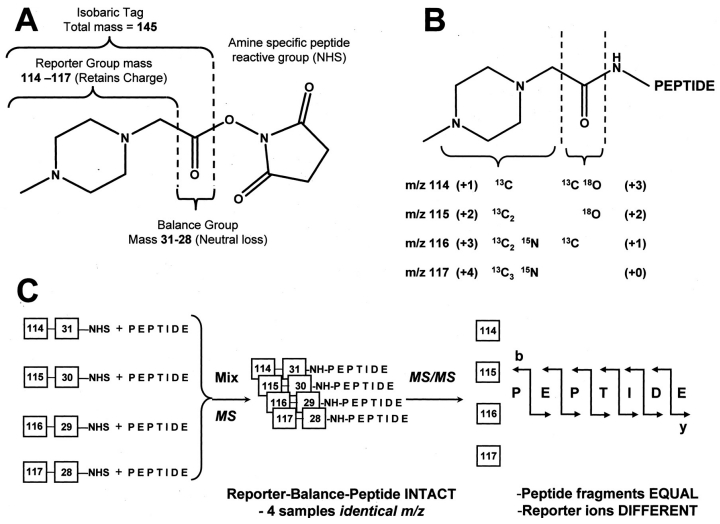
- ❑ Closer to the biosynthetic difference so less risk for non-systematic loss
- ❑ Can be coupled with IMAC or other separation because SILAC doesn't affect charge or hydrophobicity of peptides
- ❑ Quantitation done in MS1 instead of MS2 which is usually more accurate

Quantitation with Isobaric tagging



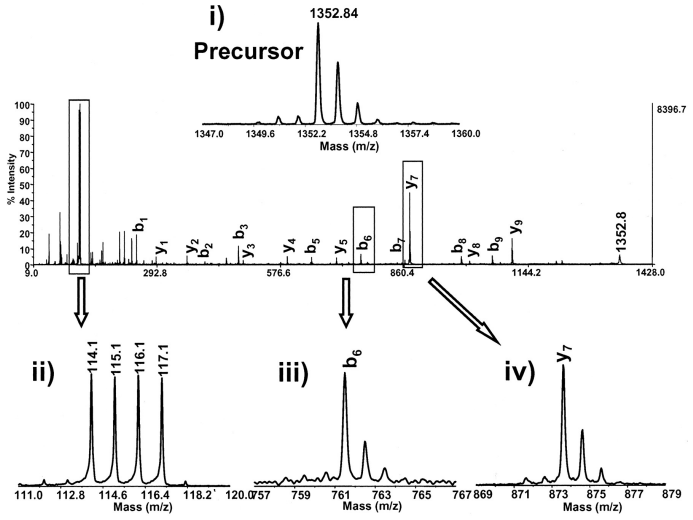
Source-A.J. Bureta, http://en.wikipedia.org/wiki/Isobaric_labeling

iTRAQ: Isobaric Tagging Reagent and Quantitation



Philip L. Ross et al. Mol Cell Proteomics 2004;3:1154-1169

Example MS/MS spectrum of peptide TPHPALTEAK from a protein digest mixture prepared by labeling four separate digests with each of the four isobaric reagents and combining the reaction mixtures in a 1:1:1:1 ratio.



Philip L. Ross et al. Mol Cell Proteomics 2004;3:1154-1169

Basic iTRAQ Workflow

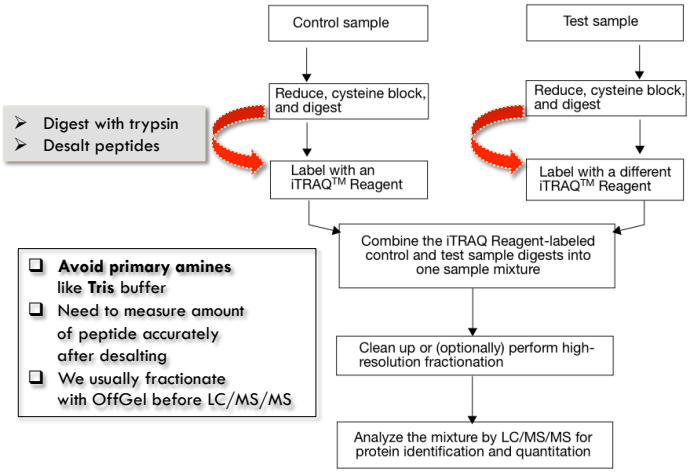
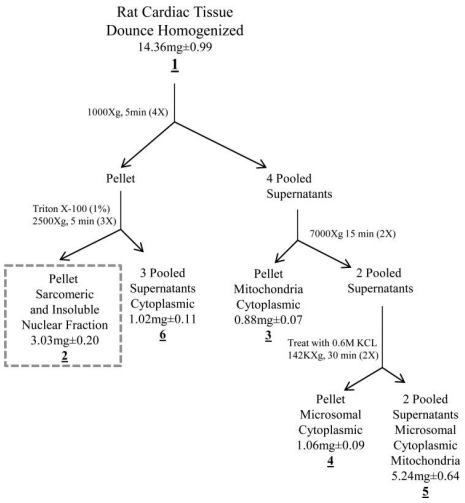


Figure 1-1 Overview of iTRAQ™ Reagents methodology

Modifed from Applied Biosystems/AB Sciex



Sub-cellular fractionation of rat cardiac tissue

The flowchart depicts the sub-cellular separation based mainly on differential centrifugation. The underlined numbers represent fractions taken during the procedure. The dashed box represents the sarcomeric enriched fraction used in the study. Protein concentration was determined at each of the fractions and is represented as mean (mg) ±SEM n=4. There were no significant differences between the non-ischemic and ischemic tissue protein concentrations. The fractionation yielded approximately 78% recovery when comparing the whole homogenate (fraction 1) vs fractions 2-6.

From-Warren, et al., "Sub-proteomic fractionation, iTRAQ, and OFFGEL-LC-MS/MS approaches to cardiac proteomics." J Proteomics. 2010 Jun 16;73(8):1551-61.

Uses of iTRAQ

91

- Comparing samples under different conditions, time courses, etc.
  - Doesn't require cell culture
- Kinobeads for kinase inhibitor binding curves

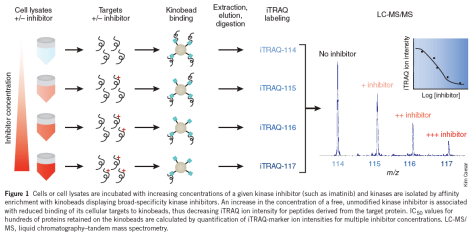


Figure 1 Cells or cell lysates are incubated with increasing concentrations of a given kinase inhibitor (such as imatinib) and kinases are isolated by affinity enrichment with kinobeads displaying broad-specificity kinase inhibitors. An increase in the concentration of a free, unmodified kinase inhibitor is associated with reduced binding of its cellular targets to kinobeads, thus decreasing iTRAQ ion intensity for peptides derived from the target protein. IC_{50} values for hundreds of proteins retained on the kinobeads are calculated by quantification of iTRAQ-marker ion intensities for multiple inhibitor concentrations. LC-MS/MS, liquid chromatography–tandem mass spectrometry.

From News & Views by Forest White (Nature Biotech. 25(2007):994-996) summarizing Bantscheff, et al., "Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors", Nature Biotech, 25 (2007):1035-1044.

Protein Expression Levels Across Populations

92

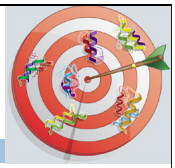
- Mike Snyder's group-"Variation and genetic control of protein abundance in humans", Nature 2013 Jul 4;499(7456):79-82.
- Using iTRAQ/TMT to quantitate protein expression instead of mRNA levels.
- Quantitated relative expression levels of ~6000 gene products in lymphoblastoid cell lines from 95 individuals genotyped in the HapMap Project (CEU, YRI and ASN populations)

Using isobaric tagging to study tumors

93

- Wang X, et al., "Breast tumors educate the proteome of stromal tissue in an individualized but coordinated manner." Sci Signal. 2017 Aug 8;10(491) PMID: 28790197
- New York Univ School of Medicine, Wash U St Louis, Baylor College of Medicine & Thermo research group in Rockford
- Used patient-derived xenograft (PDX) in mice combined with TMT labeling to study how a tumor changes the microenvironment around it.

SRM Quantitation



94

- "Targeted Proteomics" was selected as the Method to Watch for 2010 by Nature Methods*
- SRM (Selective Reaction Monitoring; aka Multiple Reaction Monitoring) uses triple quadropole mass specs. The method involves letting specific ions through the first quad, fragment in the second, and look for a specific fragment in the third quad
- This "transition" is unique to a target peptide and can be used to quantitate
- Usually mix in a known amount of a heavy form of the target peptide and compare unknown intensity with spiked pepide
- Rapid assays
- Ability to multiplex and look for >20 transitions simultaneously

*"Targeted proteomics" *Nature Methods* **8**, 43 (January 2011)

SRM is useful for…

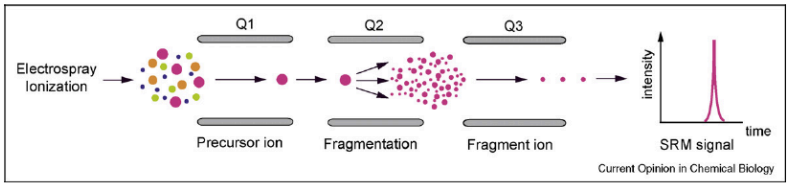
95

- Quantitating something without fractionation
- Small molecule quantitation
- **Peptides**
  - Use free software (Skyline-U Wash) that converts a "discovery" MS/MS run into a targeted run on triple quad mass spec
  - Relies on "proteotypic peptides", i.e., peptides we know we'll see every time from a specific protein.

SRM with Triple Quad

96

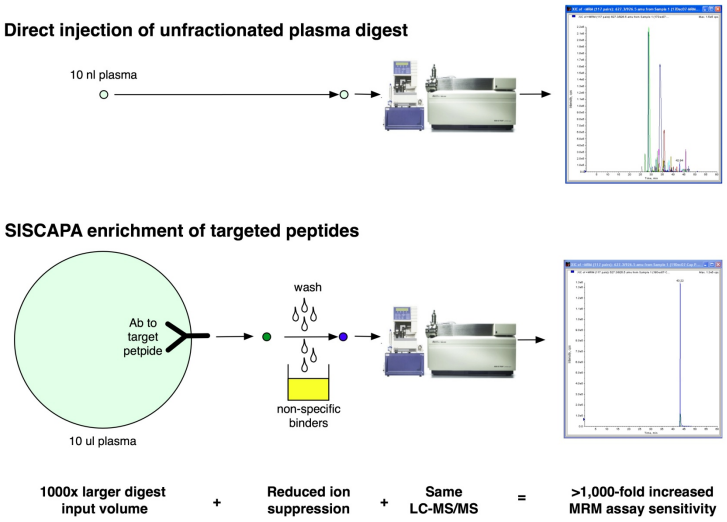
Figure 1



Selected reaction monitoring. The figure shows a schematic illustration of the principle of a triple quadrupole mass spectrometer (QQQ). In the first quadrupole (Q1), a specific precursor ion of a PTP is selected based on its mass-to-charge (*m/z*) ratio. The precursor ion is fragmented in the second quadrupole (Q2) by collision-induced dissociation, which allows for the selection of a specific fragment of the target peptide ion, according to its *m/z* ratio, in the third quadrupole (Q3). The signal intensity of this fragment is reported over time. The pair of *m/z* ratios for the precursor and fragment ions is a so-called SRM transition. A series of the best SRM transitions for the target peptide in combination with its retention time and instrument parameters, serve as a fingerprint for a PTP and constitute a definitive SRM assay.

Hüttenhain, et al. "Perspectives of targeted mass spectrometry for protein biomarker verification" Current Opinion in Chemical Biology 2009, 13:1–8

SISCAPA: Increasing Target Peptide & Decreasing Ion Suppression ("Immuno-SRM")

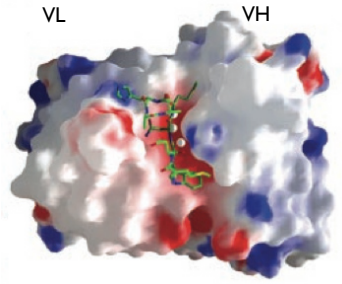


Modified from-Leigh Anderson, Plasma Proteome Institute

Anti-Peptide Antibodies As Analyte-Specific Reagents

Anti-peptide antibody (APA)

- Typically recognize 5-8 amino acid linear epitope in a combining site groove surrounded by complementarity-determining residues of Ig L and H chains
- 6-8 amino acid sequences are typically unique in the human proteome
- An anti-peptide antibody has the potential to select a single tryptic peptide from the human proteome
- The peptide immunogen is easy to synthesize, and thus the antibody is easy to make (e.g., in rabbits)
  - Can generate clones from rabbit spleens
- OriGene Technologies developing panels of APAs under NCI contract



8-mer peptide bound to antibody groove
N.K.Vyas et al (2003) PNAS 100:15023–15028.

Modified from-Leigh Anderson, Plasma Proteome Institute

SRM in Translational Science

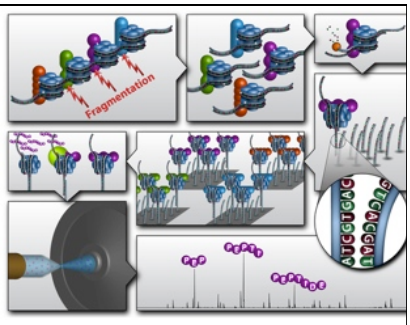
99

Clinical Proteomic Tumor Analysis Consortium (CPTAC):

http://proteomics.cancer.gov

- CPTAC public track hub available in UCSC Genome Browser for human hg19 assembly. Contains peptides from CPTAC studies of breast, colorectal & ovarian cancer from TCGA ("CPTAC Hub v1").
- NCI announced (Dec 2014) initiative to develop SRM assays for ~100 proteins involved in RAS signaling.
  - RAS oncogene linked to ~30% of human cancers
  - Combination of conventional SRM and Immuno SRM for low abundance proteins
  - CPTAC Assay Portal: https://assays.cancer.gov
    - 1219 assays
    - 1146 unique peptide sequences
    - 630 unique proteins

How low can we go?



100

SRM of ChIP

- July 2011 PLoS ONE article from Lloyd Smith (UW Madison) and Michael Olivier (UW Milwaukee):
  http://www.plosone.org/article/info:doi/10.1371/journal.pone.0026217
- Using "GENECAPP (Global ExoNuclease-based Enrichment of Chromatin-Associated Proteins for Proteomics)
- Formaldeyhyde cross-links protein to DNA then digest and SRM to measure targeted proteins bound to DNA

Can we ID single copy proteins?

101

- IFF need picomole of protein = 6 X 10^{11} molecules = 50 ng for 50 KDa protein
- IFF need picomole of cells (single copy protein) that's 6 X 10^{11} cells = 6000 liters (yeast)!
- IFF 6 copies per cell and only need 1 femptomole → 1 liter of media
- Olivier's group able to identify 200 atomoles of a transcription factor (Gal4p)
- By isolating nuclei first from 8 liters of yeast were able to ID 8 out of 10 targeted proteins in ChIP

Remember – There is no PCR for proteins!

Can we study single cell proteomics?

102

- Jonathan Sweedler (UIUC) is studying metabolites and peptides from single cells
  - Using MALDI combined with microfluidics to study peptides released from individual neurons
  - See Ong, et al., "Mass spectrometry-based characterization of endogenous peptides and metabolites in small volume samples", BBA, (2015), http://dx.doi.org/10.1016/j.bbapap.2015.01.008

Imaging Proteomics – Extract then MS/MS

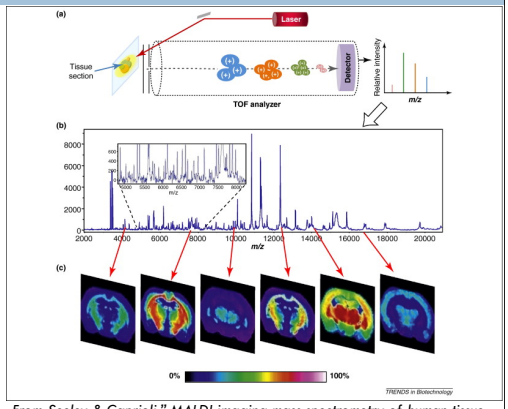
103

- Paraffin-embedded or frozen tissue
- Use laser capture microdissection (LCM) to capture area of interest
- Collect enough material then
  - Deparaffinize (FFPE) and boil in SDS to reverse X-links
  - Allow trypsin to digest O/N
  - Extract peptides and run long LC/MS/MS gradient to characterize

Imaging MS of Tissue Sections

104

- Two serial sections of a tissue sample are collected: one on a MALDI target and one on a standard microscope slide for histological staining
- Coat 5-10 μm sections with matrix then MALDI
- Can incubate sections with trypsin before coating for MALDI
- Can identify lipids, metabolites and proteins
- Caprioli's group (Vanderbilt) starting to apply clinically
- cf-**Luke Hanley's group at UIC**
  - Studying biofilms



From-Seeley & Caprioli," MALDI imaging mass spectrometry of human tissue: method challenges and clinical perspectives", Trends in Biotechnology 29, March 2011, Pages 136-143
See also-Casadonte & Caprioli, "Proteomic analysis of formalin-fixed paraffin-embedded tissue by MALDI imaging mass spectrometry." Nature Protocols (Oct. 2011) 1695-1709

Is this the future of Medical 'Omics?

106

- Snyder group* at Stanford published his genome, his mom's genome, and his proteome, metabolome and RNA-Seq expression results over 15 months
- All data available for download

*Cell 148, 1293–1307, March 16, 2012

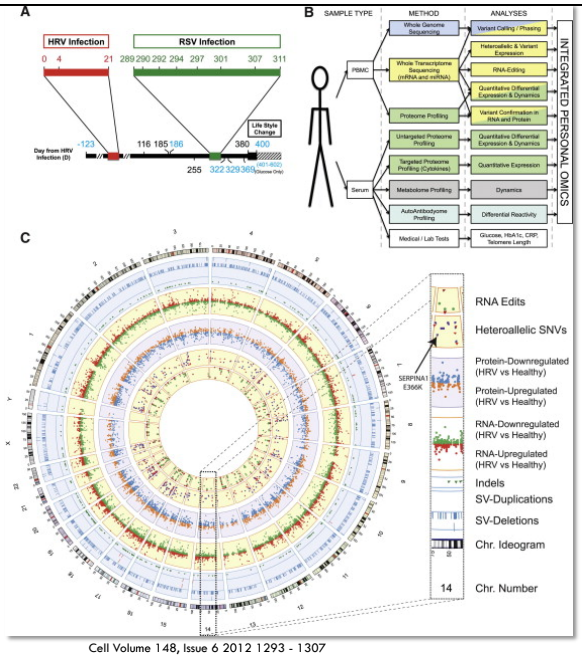


Figure 1 Summary of Study (A) Time course summary. The subject was monitored for a total of 726 days, during which there were two infections (red bar, HRV; green bar, RSV). The black bar indicates the period when the subject: (1) increased exercise, (2) i...

Cell Volume 148, Issue 6 2012 1293 - 1307

107

Analyzing Snyder RNA-Seq Data

- Downloaded from SRA
- Analyzed only two time points
- Galaxy RNA-Seq workflow

Expression View of Snyderome

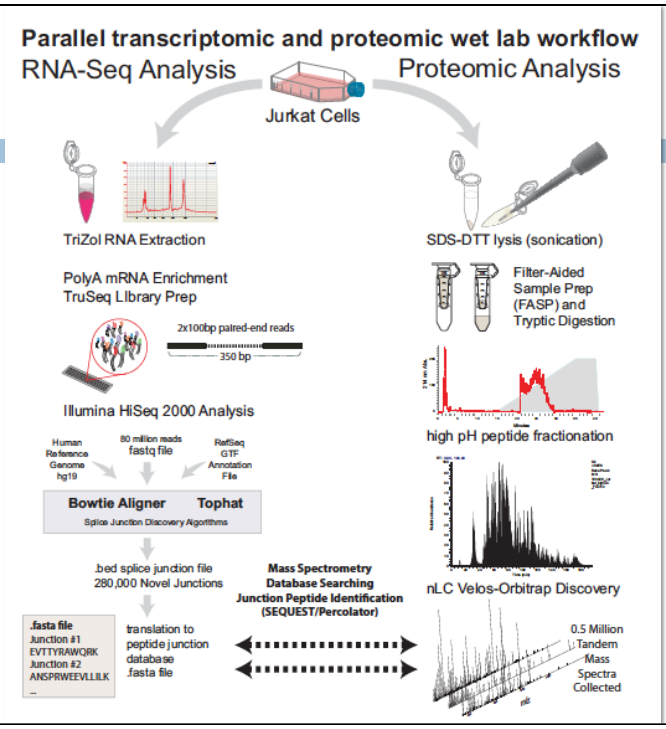
- Analyzed using the Tuxedo suite on Galaxy
- Opened "Accepted Hits" BAMs generated by TopHat in IGV
- Added Splice Junctions called by Cufflinks/ Cuffcompare/CuffDiff as separate tracks in IGV

Pushing the limits of Proteomics

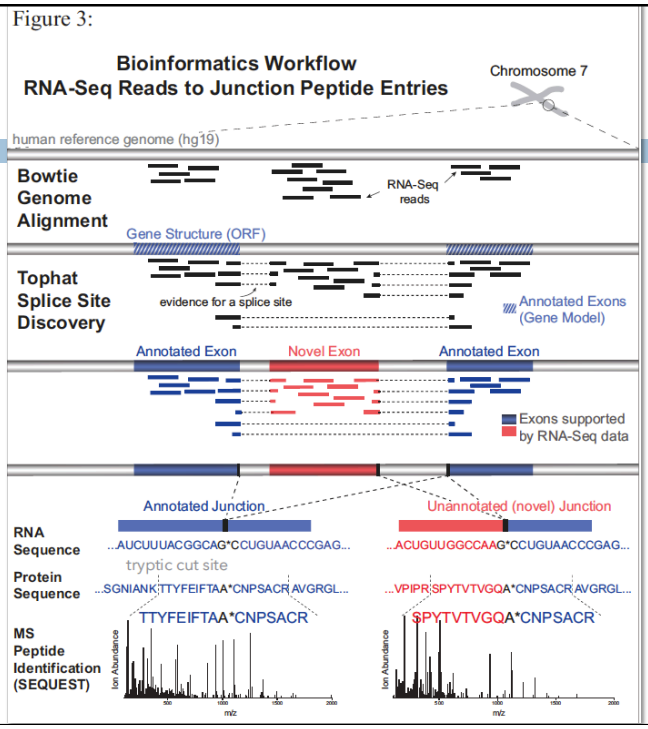
110

- Lloyd Smith's lab (UW Madison) – Sheynkman GM, et al. "Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq" Mol Cell Proteomics. 2013 Aug;12(8):2341-53.
  - *Data deposited to PeptideAtlas*
- Use RNA-Seq data to identify new splice junctions, then add to peptide database to improve identifications of novel peptides.
- Discovered 57 splice junction peptides not present in the Uniprot-Trembl proteomic database comprising an array of different splicing events, including skipped exons, alternative donors and acceptors, and noncanonical transcriptional start sites.

111



112



Specialized Proteomics Examples

113

- Estimate of protein content and localization in synaptic boutons through immunohistochemistry
  - Wilhelm, et al. Science 344 (30 May 2014): 1023-1028
- Mass spec studies using isotopic labeling to study turnover of extremely long-lived nuclear pore proteins in rat brain
  - Savas, et al., Science 335 (24 Feb 2012): 942-943 & Toyama, Savas, et al., Cell, 154 (29 Aug 2013): 971-982.
- SILAC quantitative proteomic analysis of total proteome and tyrosine phosphoproteome from cells isolated from three sites of pancreatic CA metastasis in one patient
  - Kim, et al., Mol Cell Proteomics. 2014 Nov;13(11):2803-11. doi: 10.1074/mcp.M114.038547.
- Genentech and others using Proteogenomics approaches to sequence mAbs (when clone missing) or predict immunogenic tumor mutations

Proteogenomics approach

114

- Use RNA-seq data to add to proteomics database

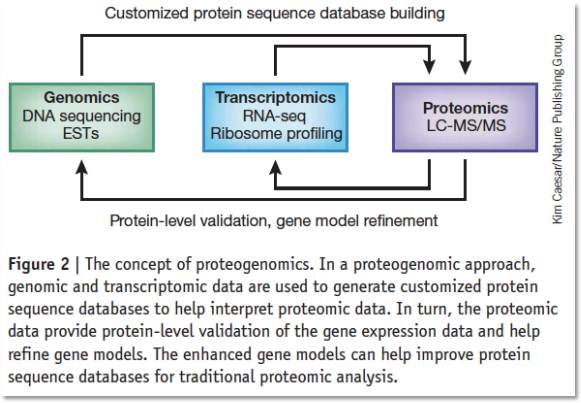


Figure 2 | The concept of proteogenomics. In a proteogenomic approach, genomic and transcriptomic data are used to generate customized protein sequence databases to help interpret proteomic data. In turn, the proteomic data provide protein-level validation of the gene expression data and help refine gene models. The enhanced gene models can help improve protein sequence databases for traditional proteomic analysis.

Nesvizhskii AI, "Proteogenomics: concepts, applications and computational strategies." Nat Methods. 2014 Nov;11(11):1114-25.

Reality check

115

- Ruggles KV, et al., "An Analysis of the Sensitivity of Proteogenomic Mapping of Somatic Mutations and Novel Splicing Events in Cancer" Molecular & Cellular Proteomics 15 (March 2016): 1060–1071
- Applied a proteogenomic data integration tool (QUILTS) to illustrate protein variant discovery using whole genome, whole transcriptome, and global proteome datasets generated from a pair of luminal and basal-like breast-cancer patient-derived xenografts (PDX).
- Despite analysis of over 30 sample process replicates, only about 10% of SNVs (somatic and germline) detected by both DNA and RNA sequencing were observed as peptides. An even smaller proportion of peptides corresponding to Novel Splice Junctions observed by RNA sequencing were detected (<0.1%).

Local Resources

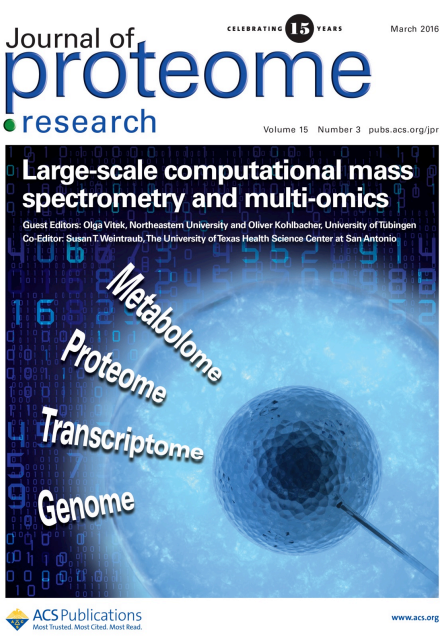
116

- U of C Proteomics core: https://proteomics.uchicago.edu/ or donw@uchicago.edu
- UIC Mass Spectrometry, Metabolomics & Proteomics Facility
  - http://www.rrc.uic.edu/mmpf
  - proteomics@uic.edu
- Chicago Mass Spec Discussion Group
  - http://www.cmsdg.org/ (meet at NU-Evanston & UIC)
- Top Down Proteomics & Center for Excellence in Proteomics – Northwestern University Kelleher lab (http://www.kelleher.northwestern.edu or contact paul-thomas@northwestern.edu)
- Annual Chicago Biomedical Research Proteomics & Informatics Workshop (late summer?)
  - http://chicagobiomedicalconsortium.org/
- Metabolomics labs in Chicago:
  - http://breemen.lab.uic.edu/

New software is continually published

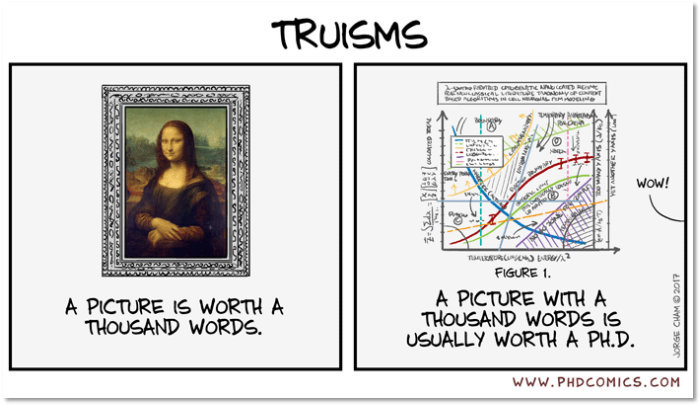
117

- http://pubs.acs.org/toc/jprobs/15/3
  - March 2016 special issue of J Proteome Research



PhD Comics' perspective on my obsession with networks and pathways

118



- http://www.phdcomics.com/comics.php?f=1926

Thanks for your attention!

119

- Let me know if you have any questions about your final projects (**preferred e-mail:** lhelseth@gmail.com)
- Please document what you do, show what you found, and send your write-up to me by 6 pm on August 25th.