Cancer Proteomics

Bing Zhang, Ph.D.

Professor of Molecular and Human Genetics

Lester & Sue Smith Breast Center

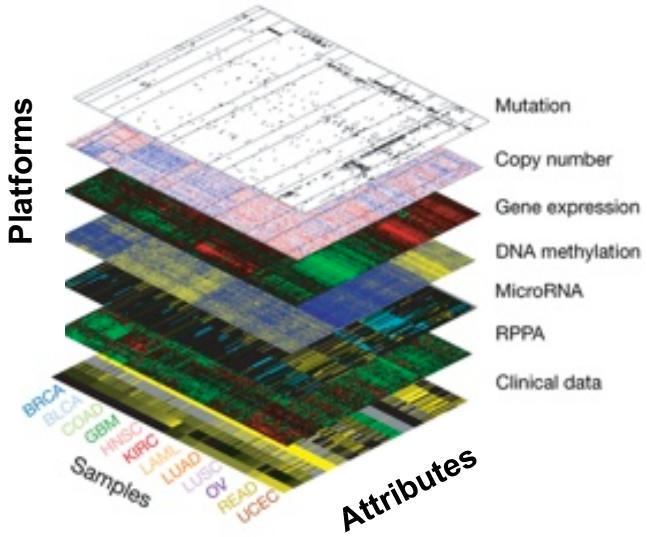
Baylor College of Medicine

bing.zhang@bcm.edu

Overview

- Why proteomics
- Proteomics technology
- Protein identification
- Protein quantification
- Protein-protein interaction

Cancer genomics: success and challenges

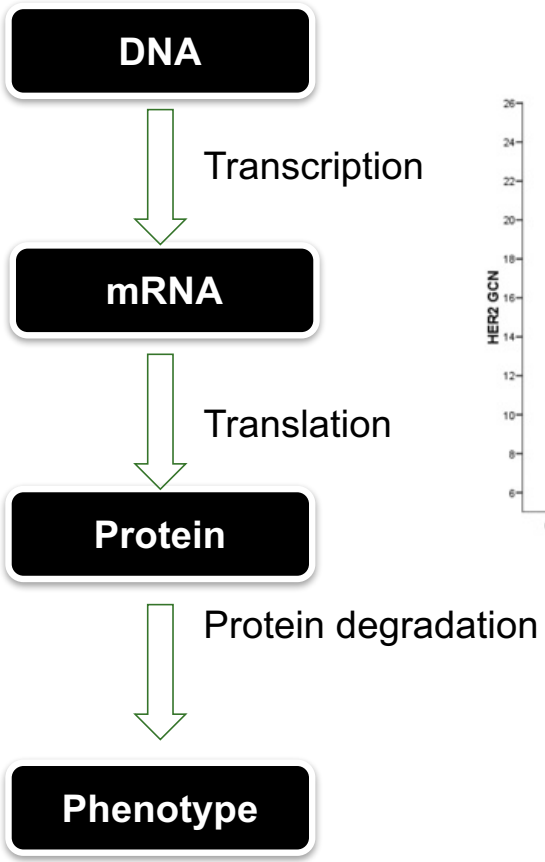


The Cancer Genome Atlas

TCGA, Nat Genet, 2013

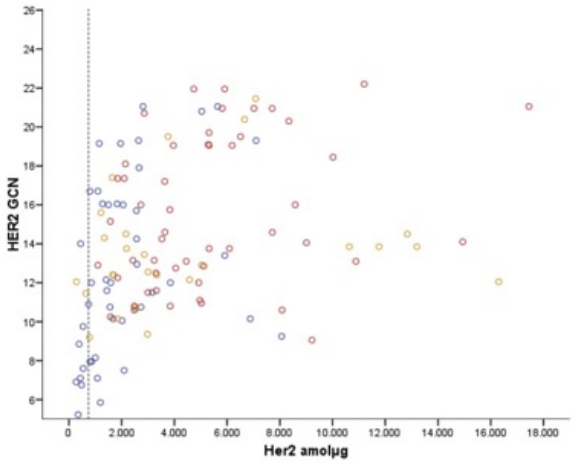
- Comprehensive catalogs of the key genomic alterations
- Actionable genomic abnormalities
  - BCR-ABL translocation in CML
  - HER2 amplification in breast cancer
  - BRAF mutation in melanoma
  - EGFR mutations or ALK rearrangements in lung cancer
- Genome-driven oncology
  - FoundationOne CDx™
- Challenges
  - Only covers a subset of patients
  - Lower than expected response rates
  - Drug resistance

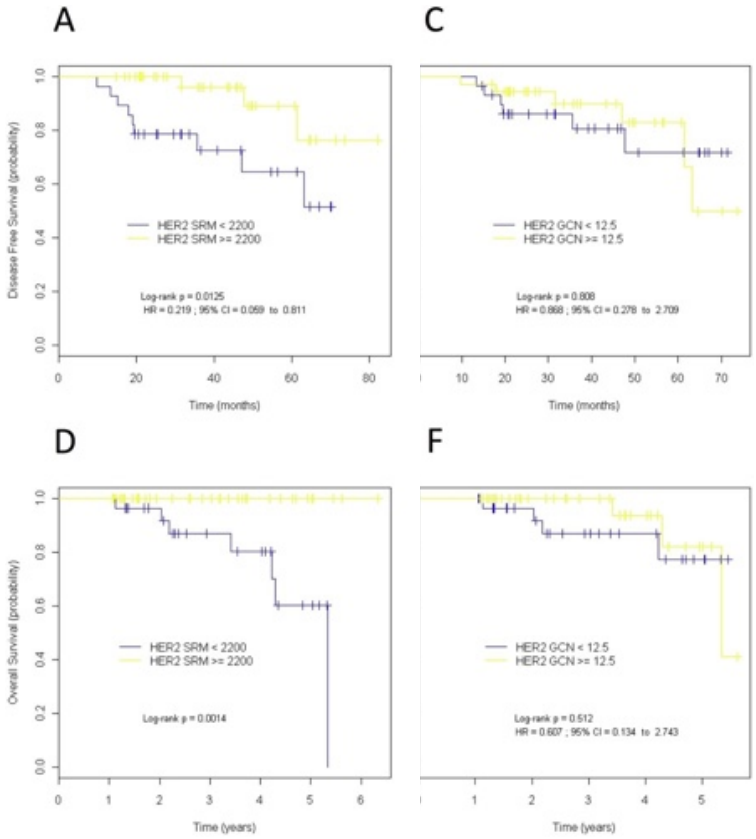
Missing biology



HER2 protein abundance *vs* copy number amplification

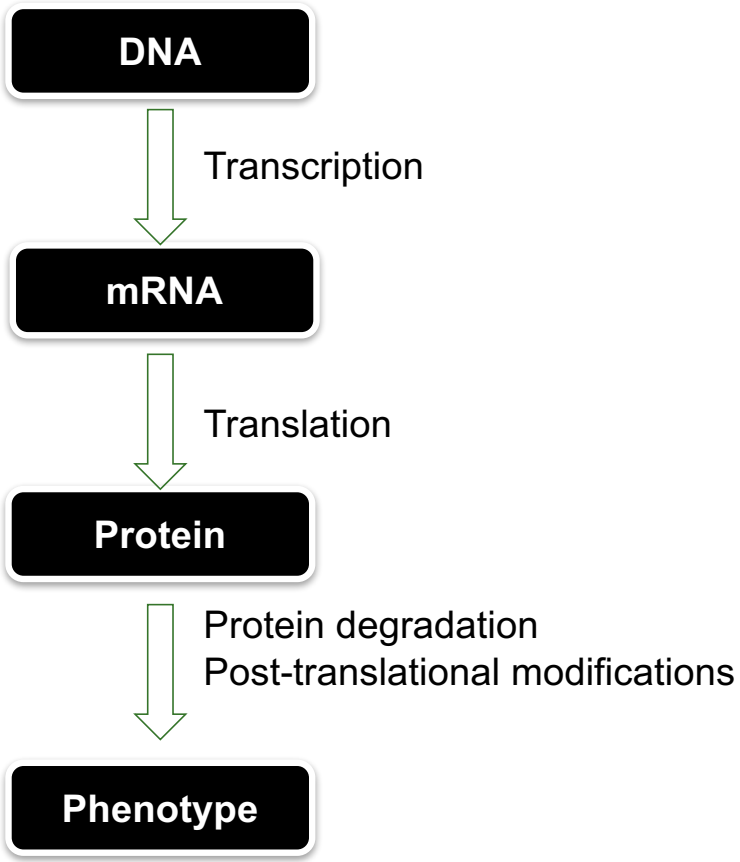
Nuciforo et al., Mol Oncol, 2016





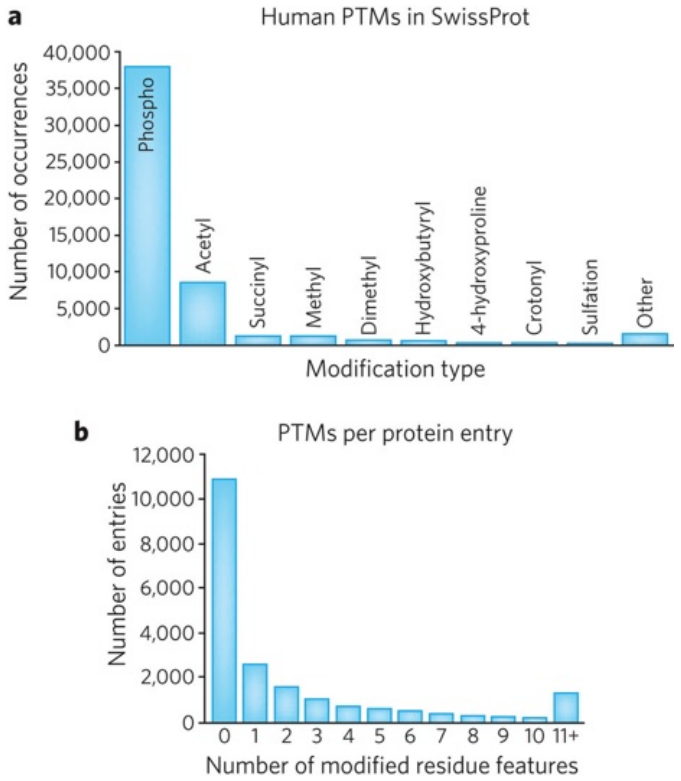
BCM QCB, April 2018

Missing biology

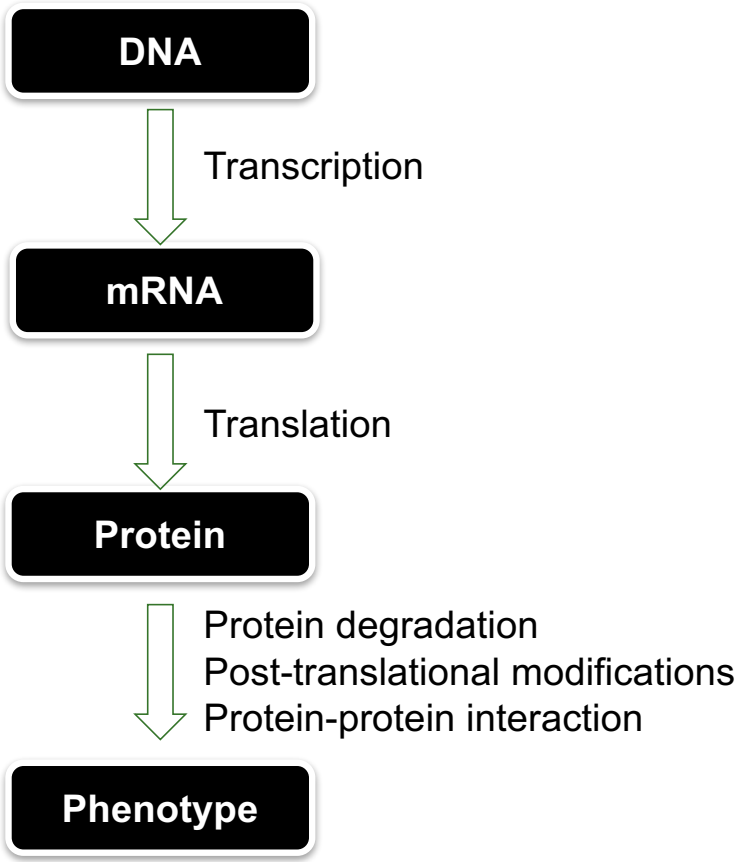


How many human proteoforms are there?

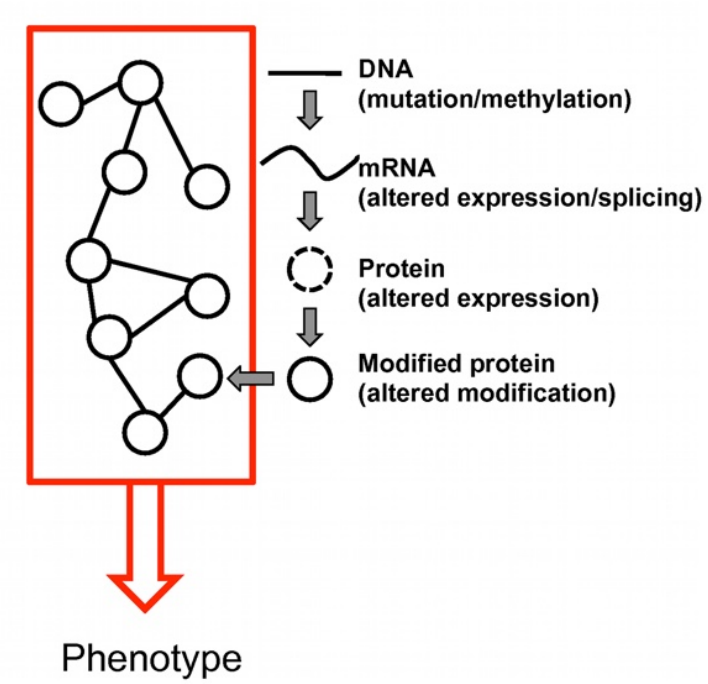
Aebersold et al., Nat Chem Biol, 2018



Missing biology

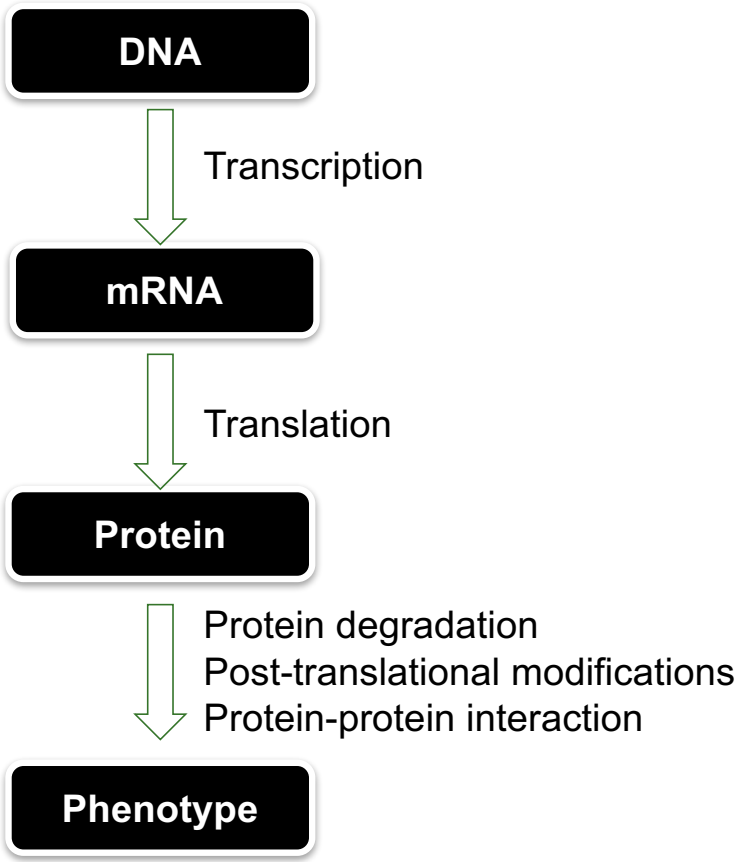


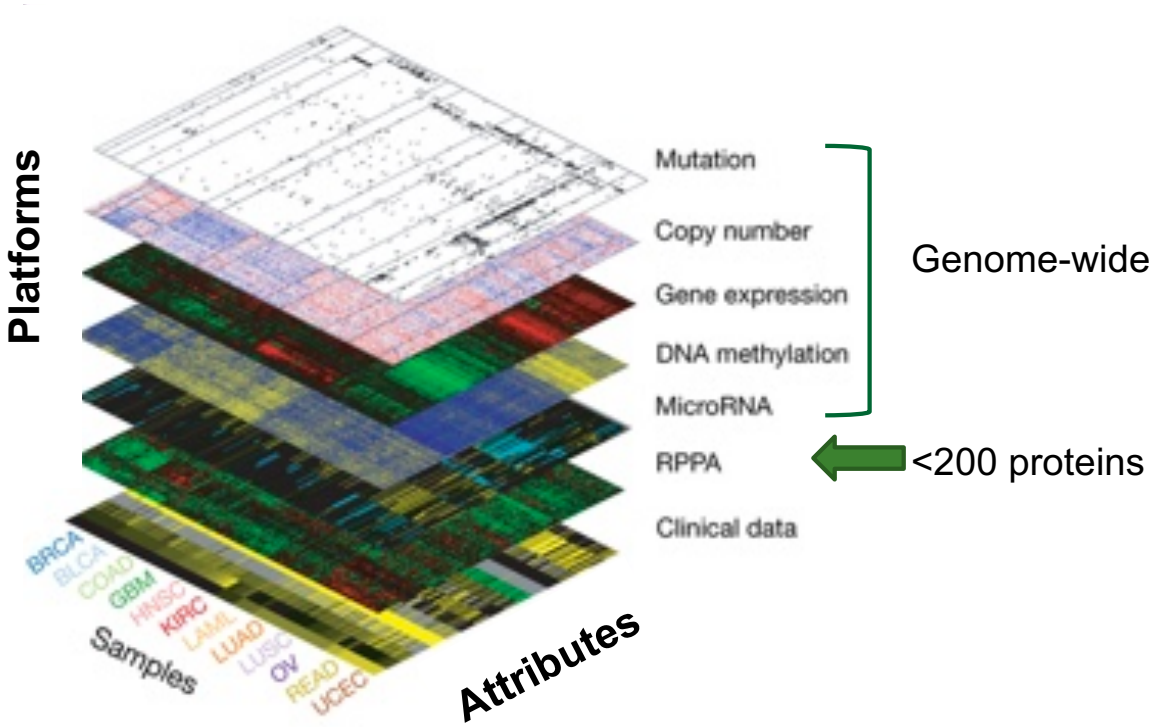
Cancer is a disease of network dysregulation



Shi et al., PLoS One, 2012

Missing biology

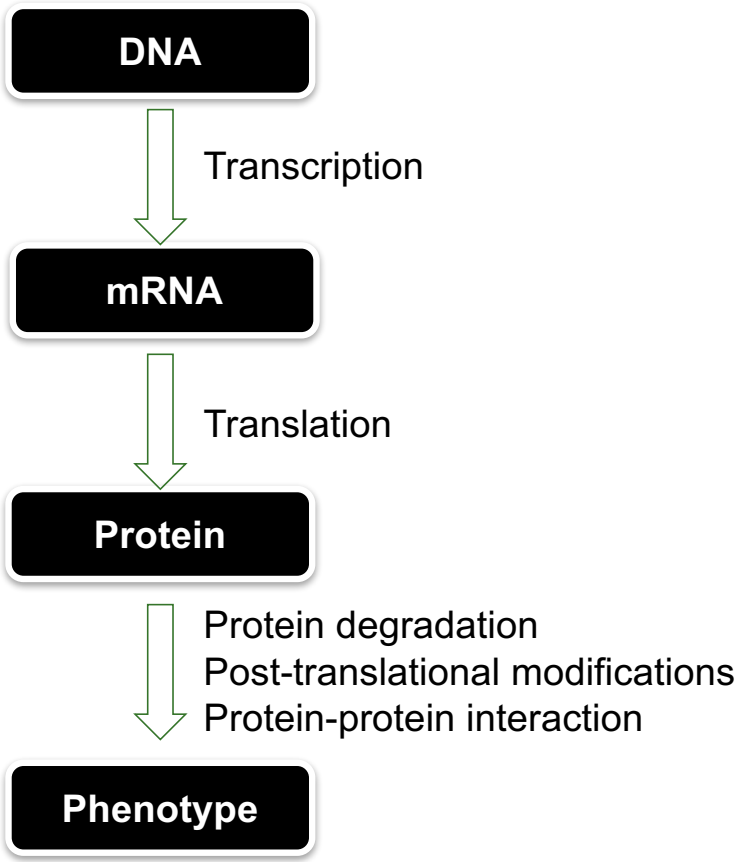




The Cancer Genome Atlas

TCGA, Nat Genet, 2013

Proteome and proteomics



- Proteome
  - The entire set of proteins expressed in a given type of cell, tissue, or organism, at a given time, under defined conditions.
- Proteomics
  - The study of the proteome.

Overview

- Why proteomics
- **Proteomics technology**
- Protein identification
- Protein quantification
- Protein-protein interaction
- Proteogenomic integration and systems biology

Mass spectrometers used in proteomics research

AB SCIEX
TripleTOF 5600

Thermo Scientific
LTQ Orbitrap Velos

AB SCIEX
QTRAP 5500

Thermo Scientific
Q Exactive

Bruker Daltonics
UltrafleXtreme

Thermo Scientific
Q Exactive HF

AB SCIEX
TripleTOF 6600

Thermo Scientific
Orbitrap Fusion

Thermo Scientific
LTQ Orbitrap Elite

Slice courtesy of Bo Wen

Three principle components of the mass spectrometers

Produce ions (charged molecules)

Ion Source

- Matrix-assisted laser desorption/ionization (MALDI)
- Electrospray ionization (ESI)

Sort ions by mass to charge ratio (m/z)

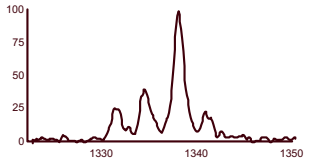
Mass Selection

- Time-of-flight (TOF)
- Quadrupole mass filter (Q)
- Ion traps (IT)
- Fourier transform ion cyclotron resonance (FTICR)

Report intensity of ions in mass spectrum

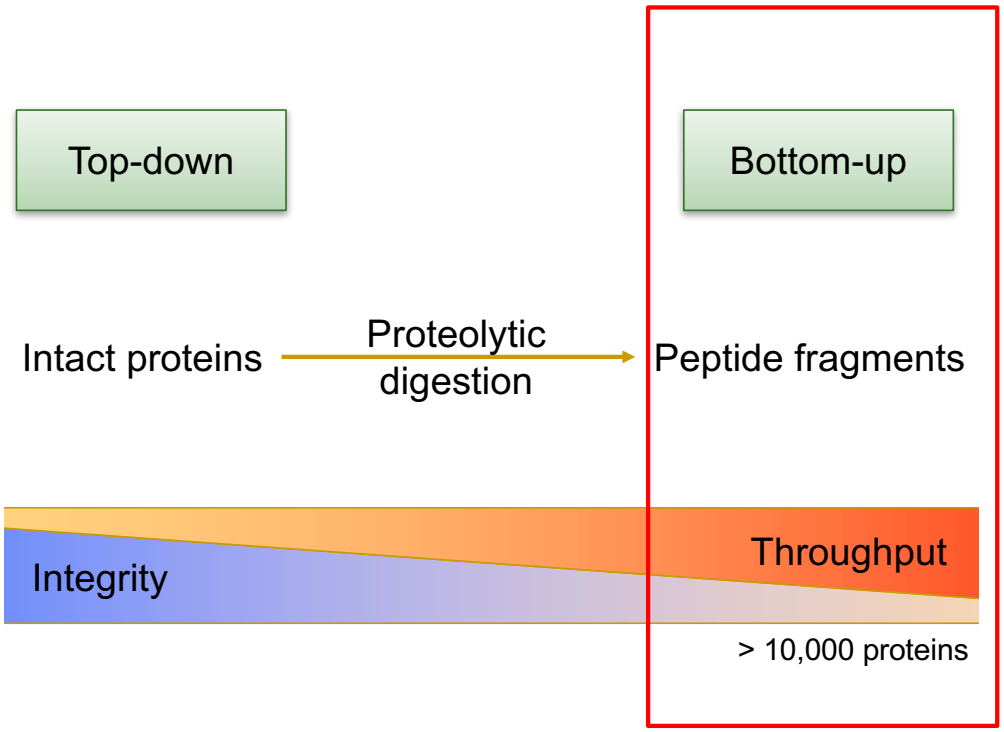
Detector

- Electron multiplier
- Array detector

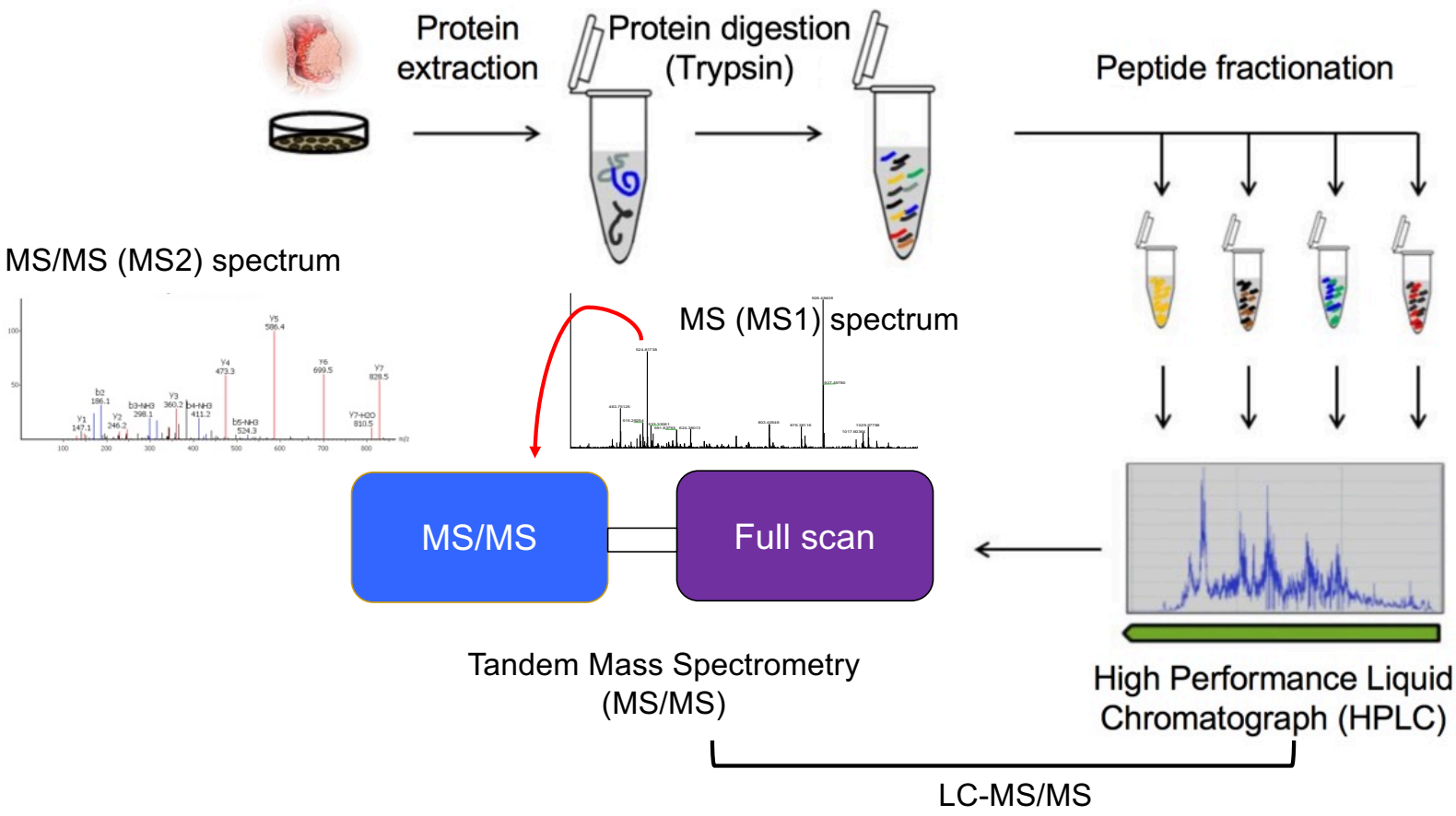


Mass Spectrum

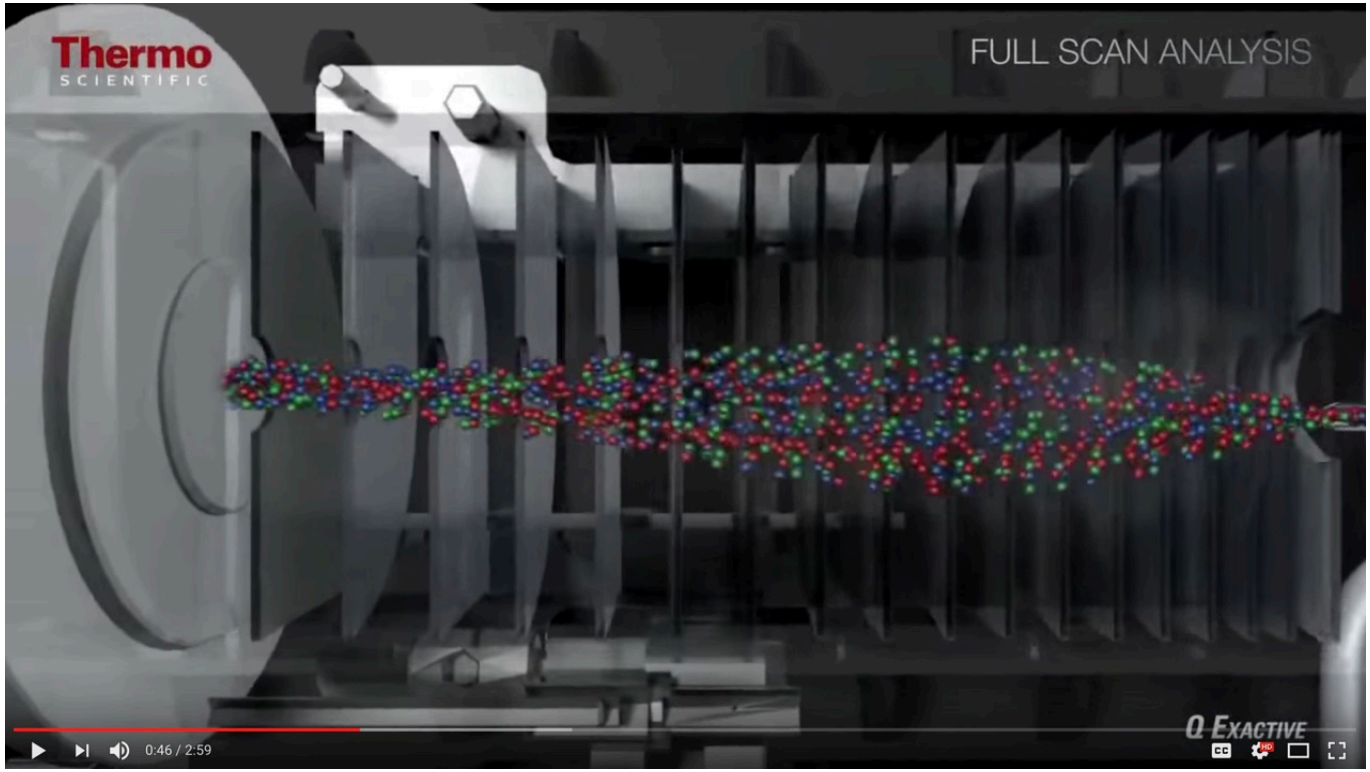
Mass spectrometry-based proteomic strategies



Bottom-up proteomics (shotgun proteomics)



MS/MS analysis using Q Exactive

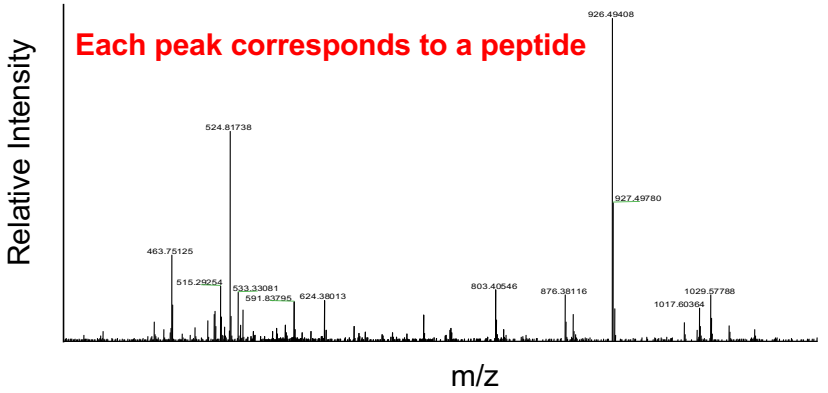


Overview

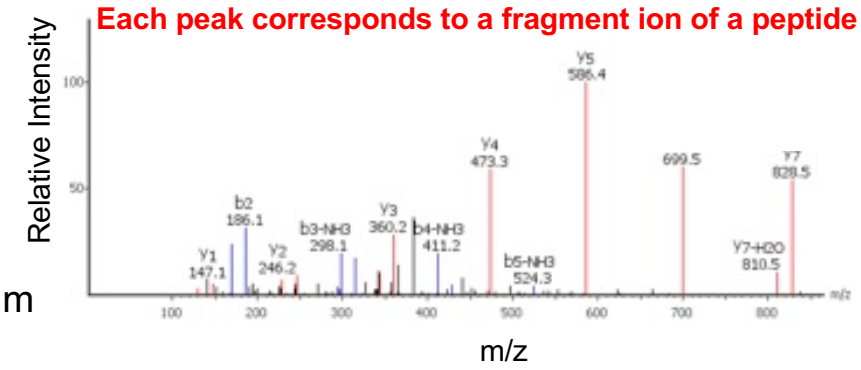
- Why proteomics
- Proteomics technology
- **Protein identification**
- Protein quantification
- Protein-protein interaction
- Proteogenomic integration and systems biology

MS/MS data and peptide identification

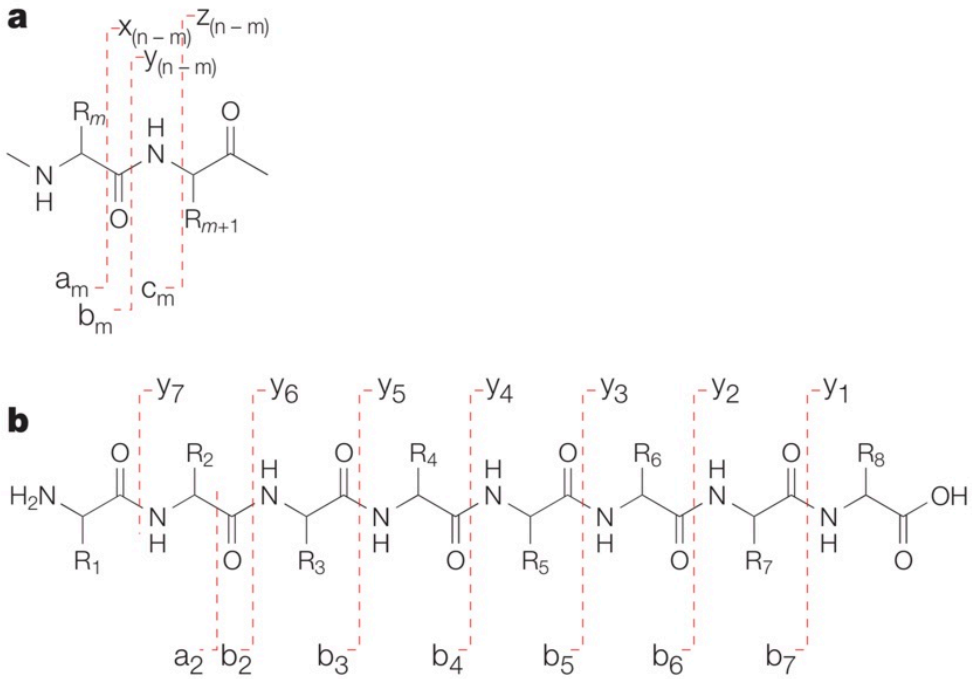
MS (MS1) spectrum



MS/MS (MS2) spectrum



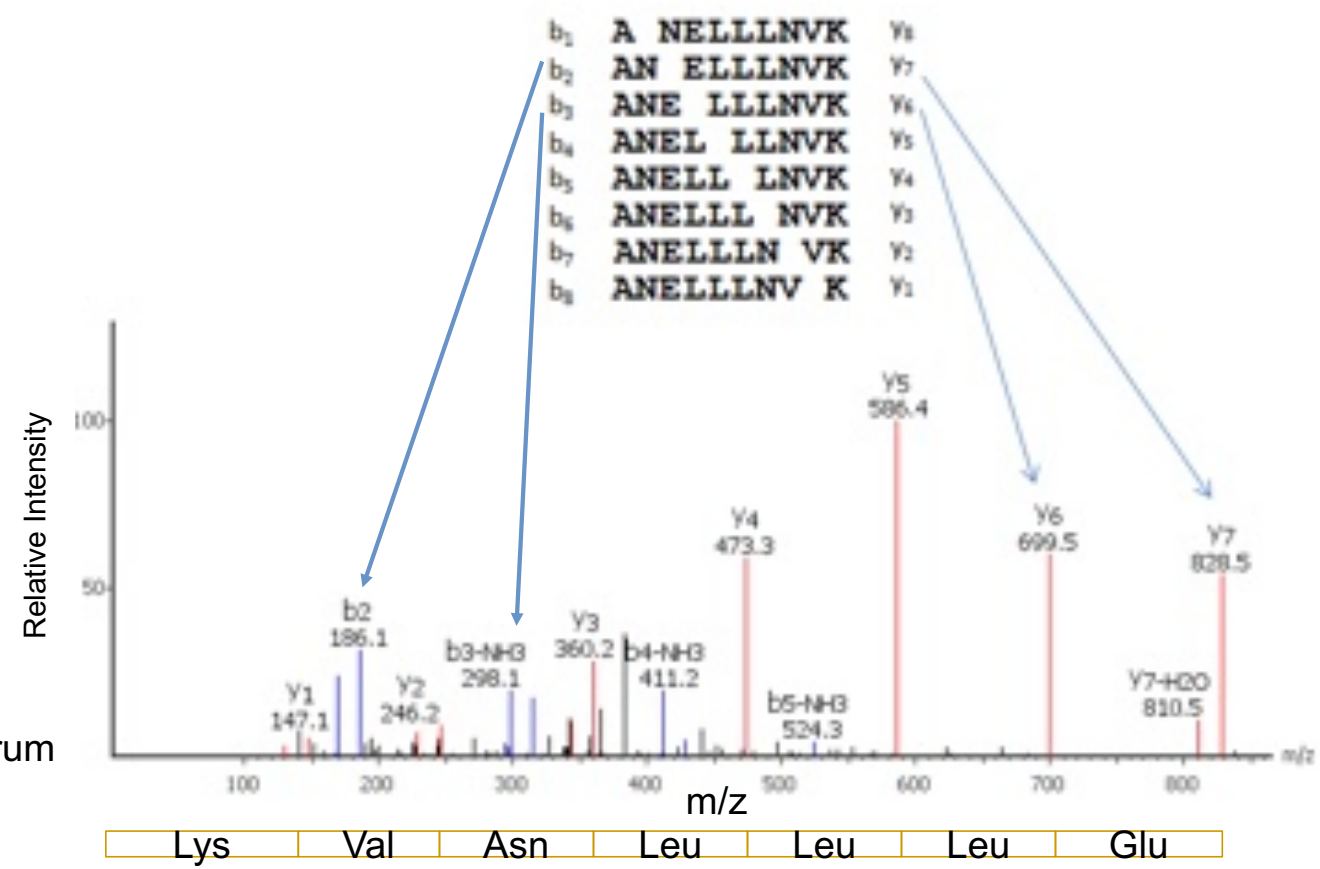
Fragment generation



Roepstorff and Fohlman, Biological Mass Spectrometry, 1984
Steen and Mann. Nat Rev Mol Cell Biol, 2004

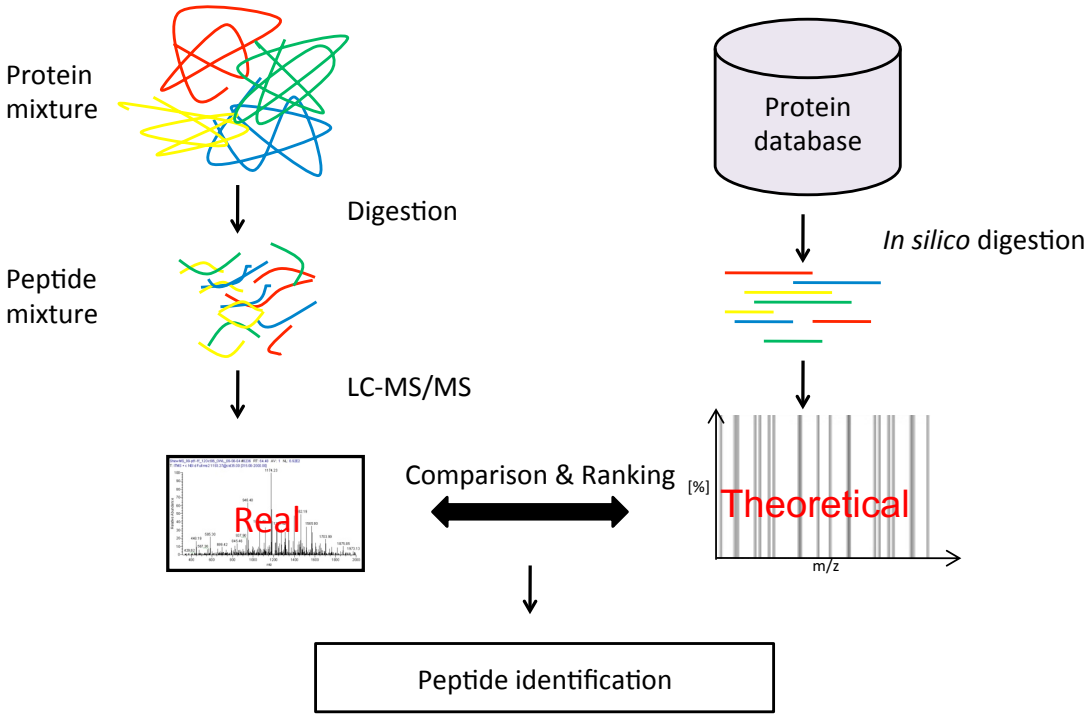
Peptide identification methods

- De novo analysis
  - Missing ions
  - Noise
- Spectrum library search
  - Incomplete
- Database search



MS/MS (MS2) spectrum

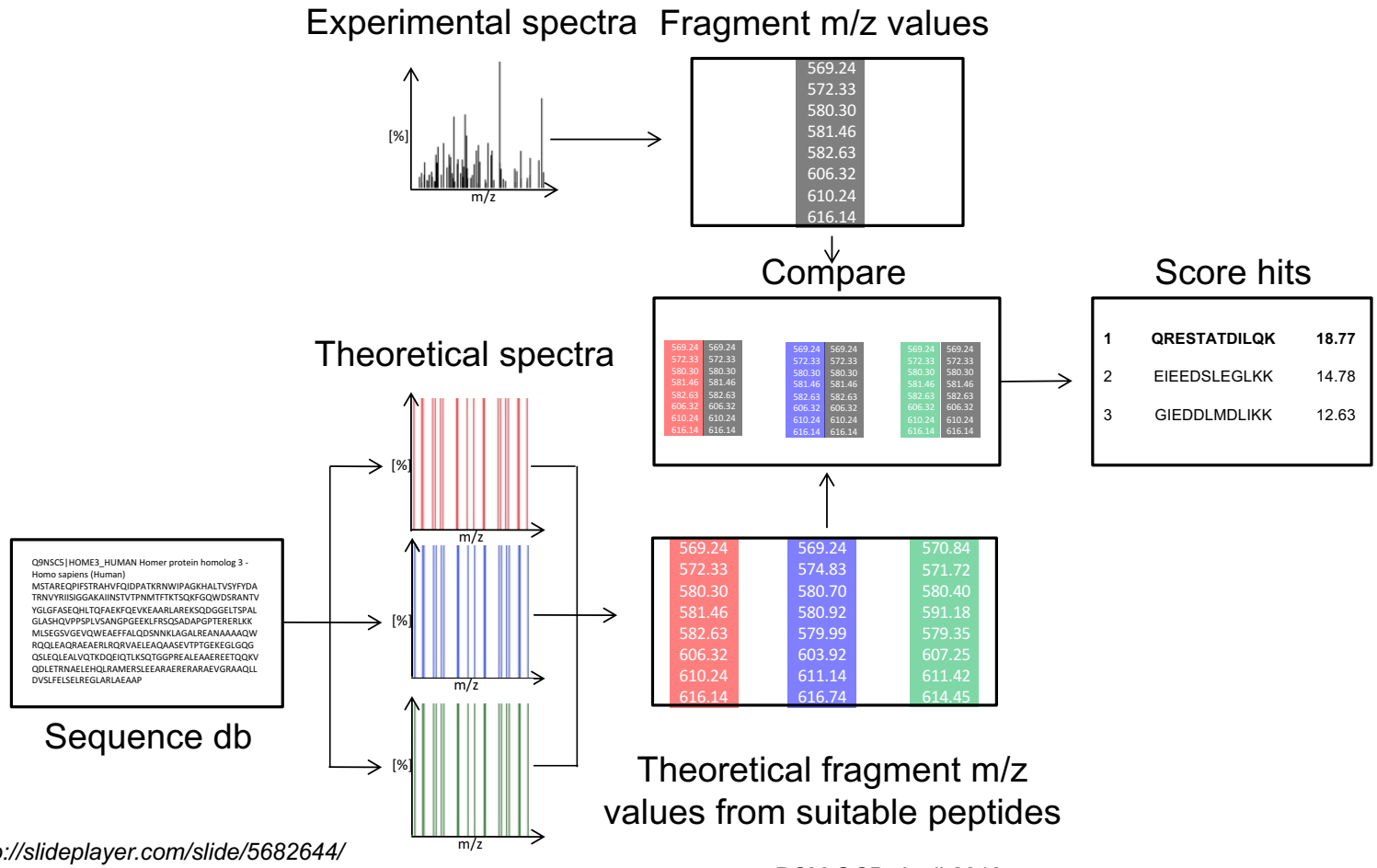
Database search for peptide identification



Wang et al., J Proteome Res, 2014

- Each proteome only includes a subset of the sequences in the database.
- Identify sequences that match the ions present in the MS/MS spectrum.

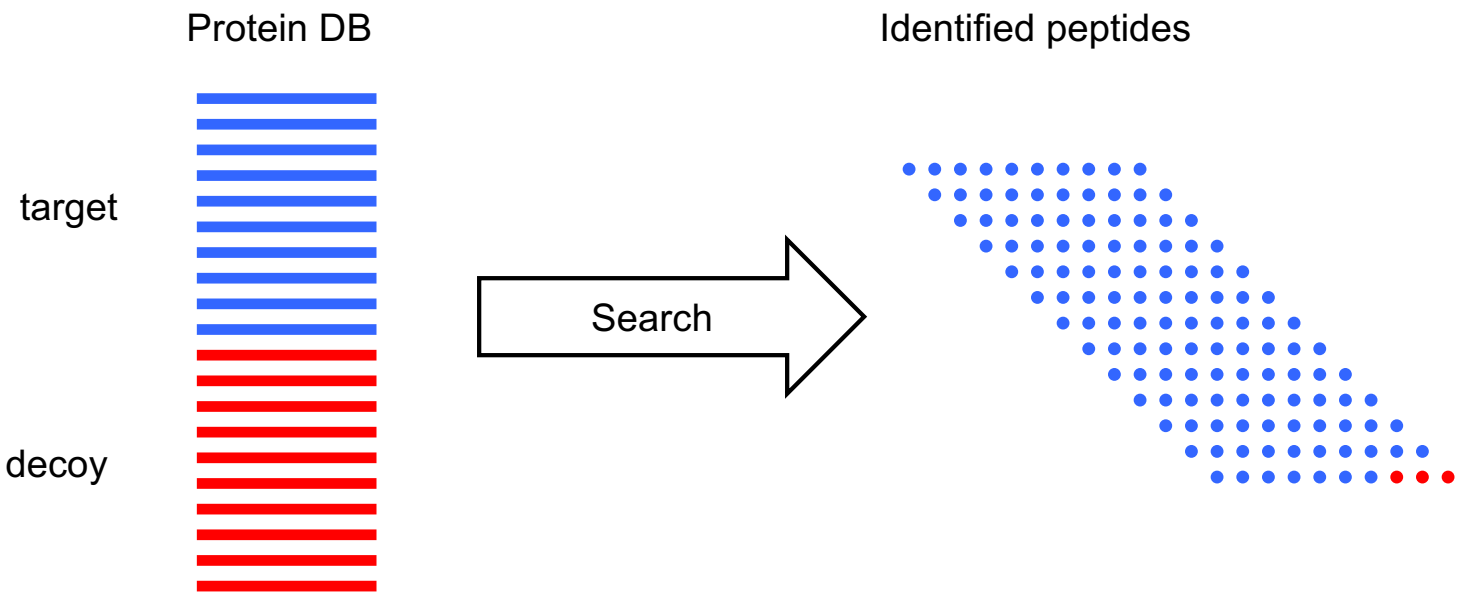
Database search for peptide identification



Algorithms:
Sequest
X!Tandem
Mascot
MyriMatch
MS-GF+
MaxQuant
…

http://slideplayer.com/slide/5682644/

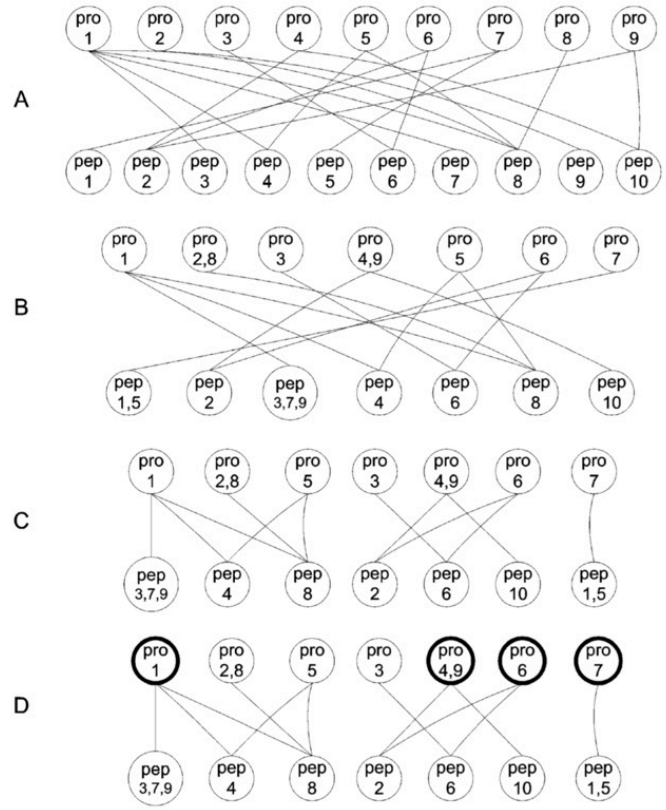
Quality control

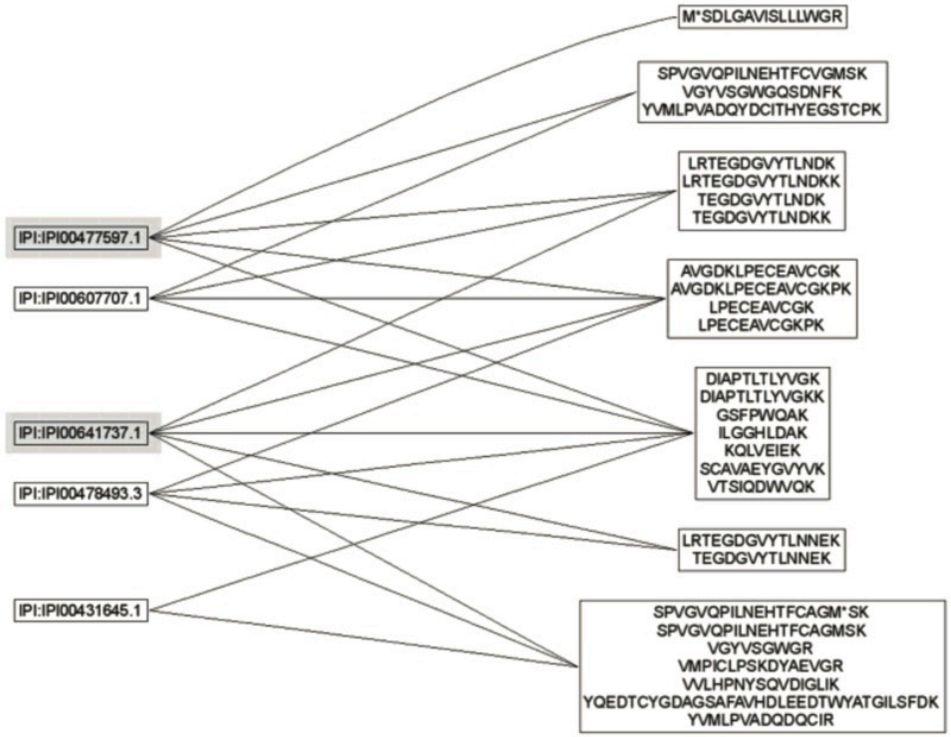


$$\text{False discovery rate (FDR)} = \frac{2 \times \#\text{decoy}}{\#\text{decoy} + \#\text{target}}$$

Elias and Gygi, Nat Methods, 2007

Protein assembly

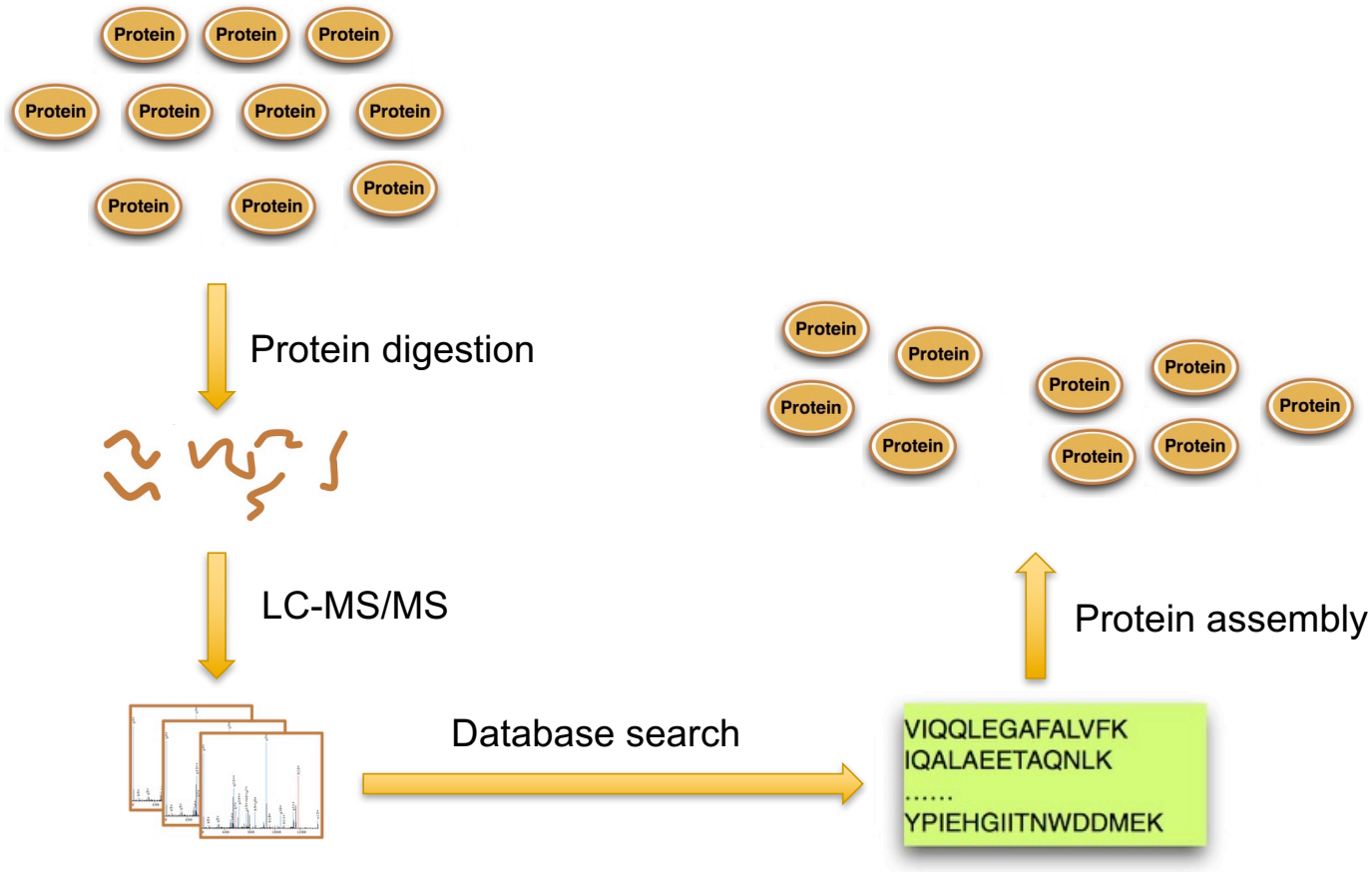




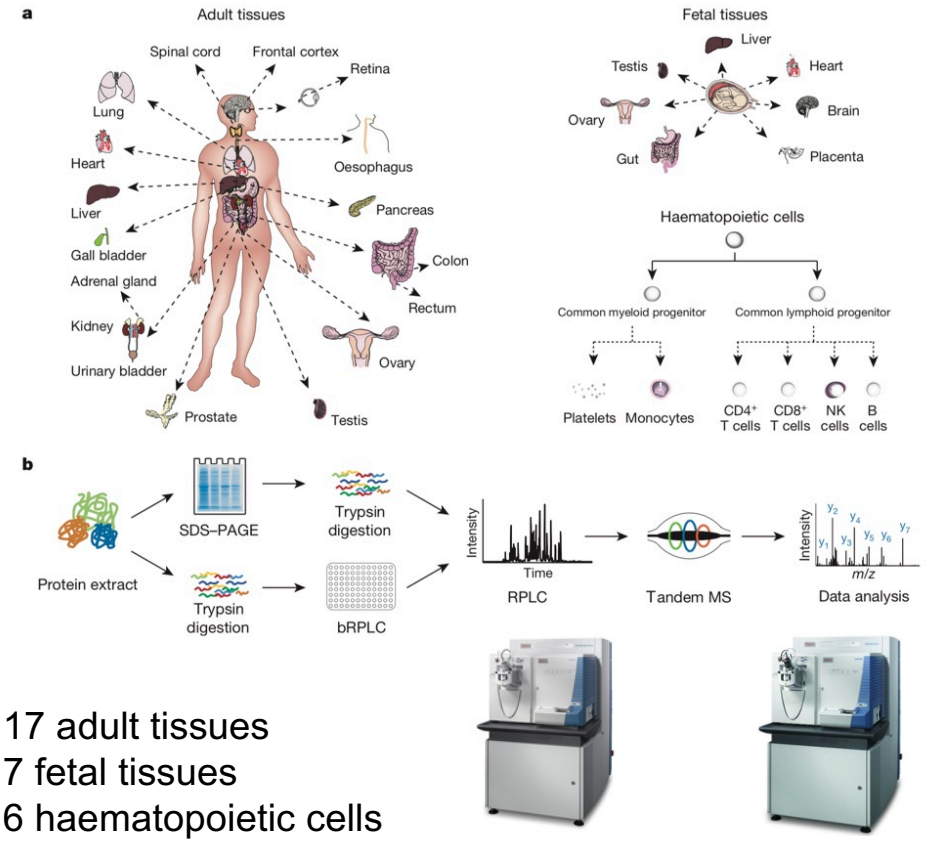
Protein parsimony through bipartite graph analysis

Zhang et al. JPR, 2006

Protein identification summary



Case study 1: draft map of the human proteome (Kim et al., 2014)



17 adult tissues
7 fetal tissues
6 haematopoietic cells

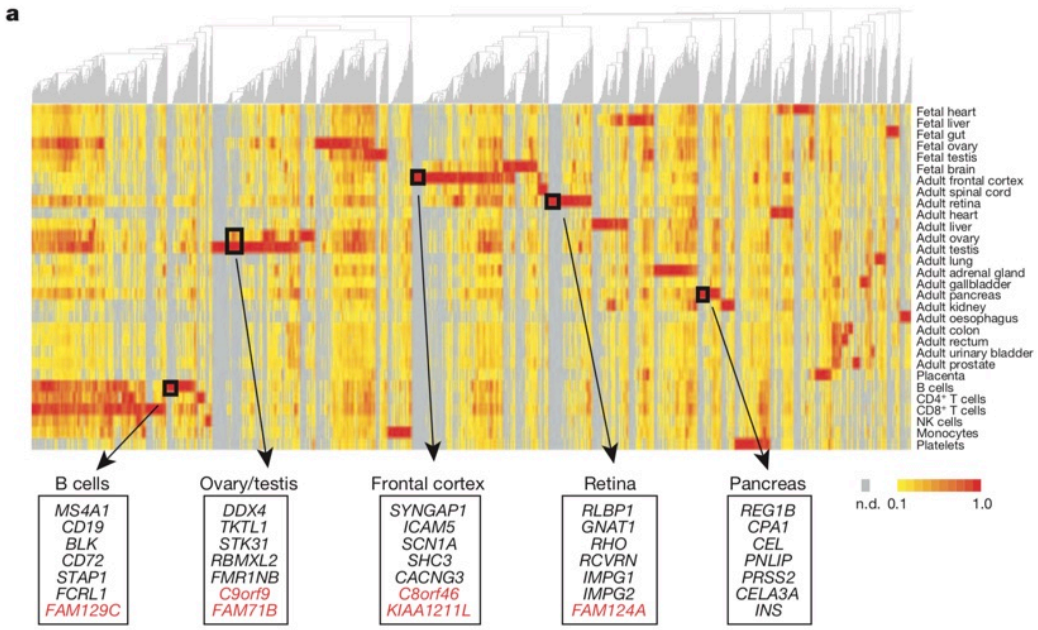
LTQ Orbitrap Velos

LTQ Orbitrap Elite

- 2,000 LC-MS/MS runs
- 25 million MS/MS spectra
- MASCOT and SEQUEST
- 293,000 peptides (FDR<1%)
- Proteins encoded by 17,294 genes
- 2535 protein products from the “missing proteins” (66% of all missing proteins)

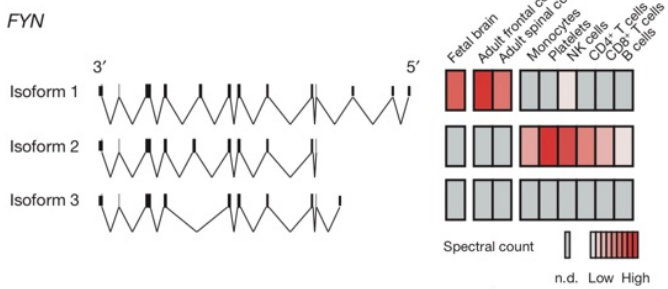
Kim et al., Nature, 2014

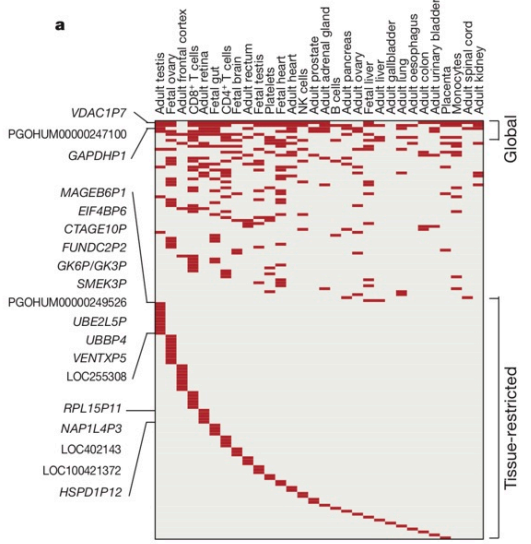
Case study 1: draft map of the human proteome (Kim et al., 2014)



Tissue-specific proteins

Tissue-specific isoforms

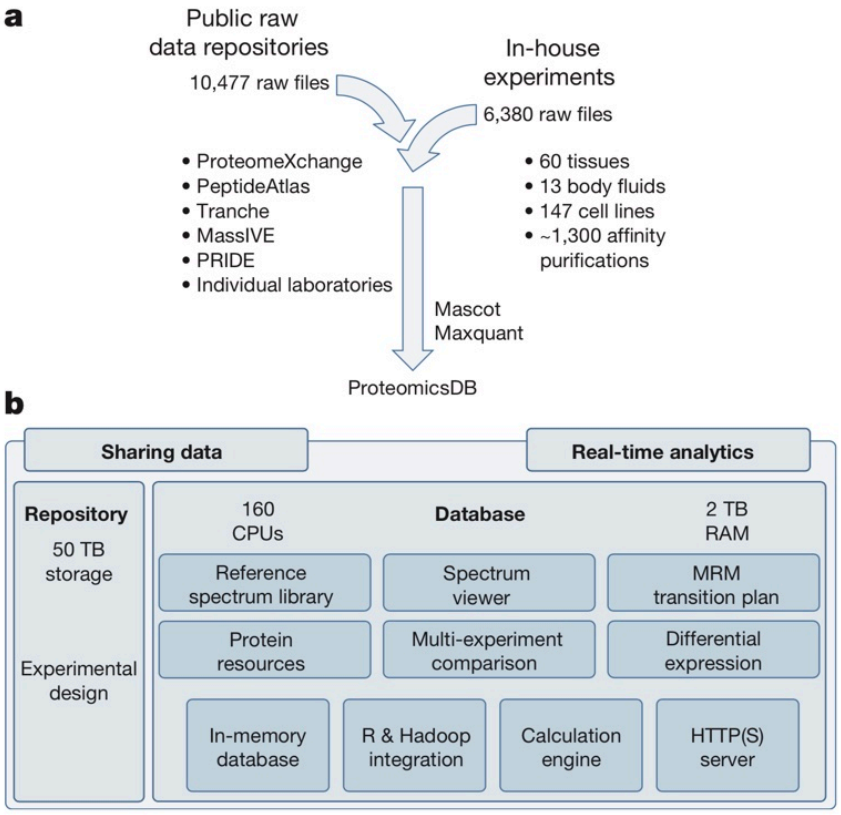




Translated pseudogenes

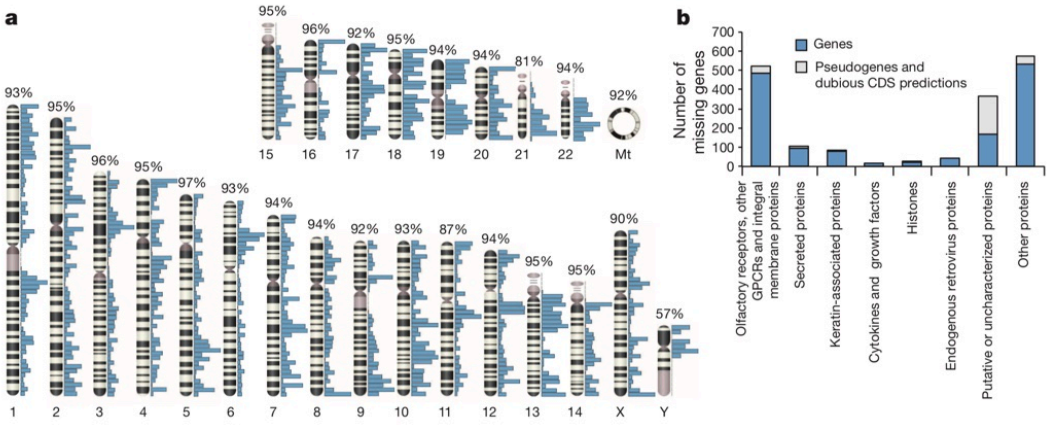
Kim et al., Nature, 2014

Case study 2: draft map of the human proteome (Wilhelm et al., 2014)

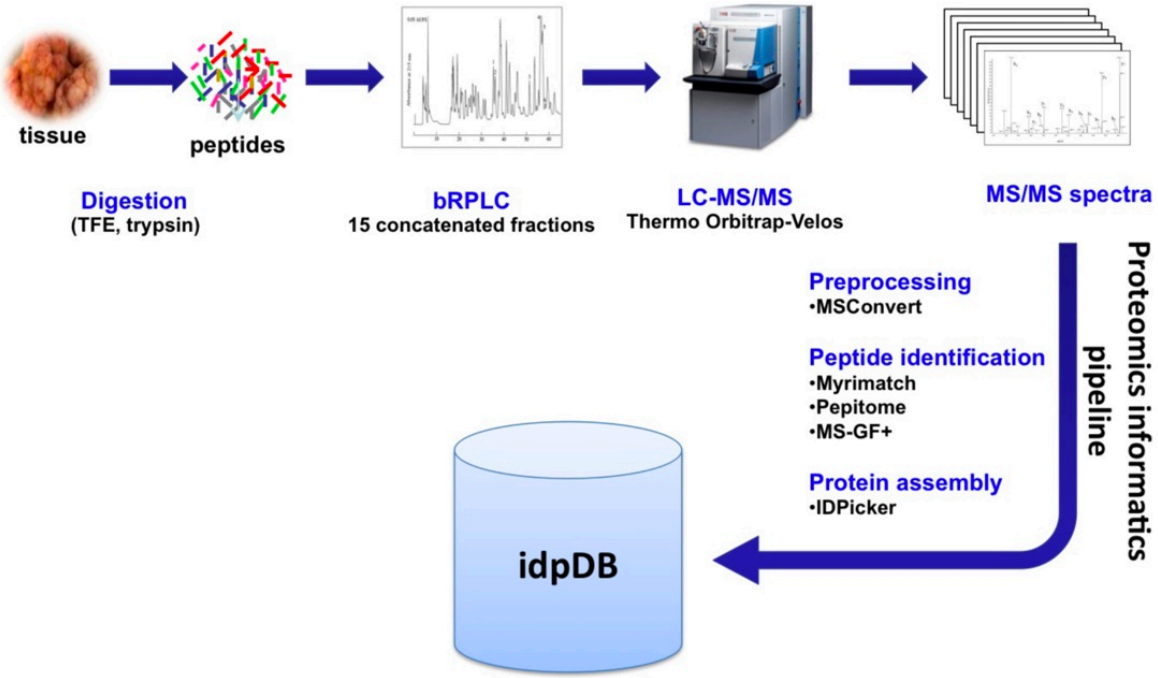


Wilhelm et al., Nature, 2014

- 16,857 LC-MS/MS runs
- MASCOT and Maxquant
- Proteins encoded by 18,097 genes (FDR<1%)



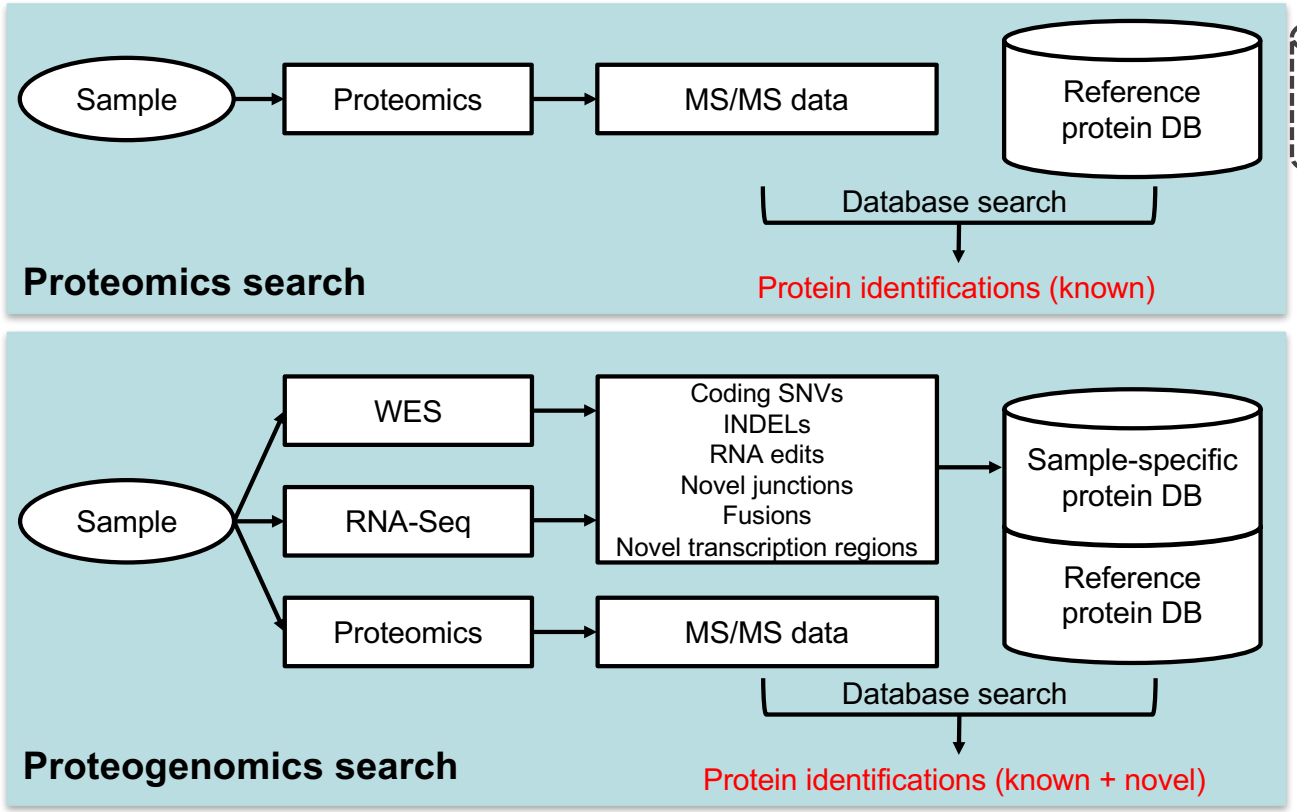
Case study 3: human colorectal cancer proteome (Zhang et al., 2014)

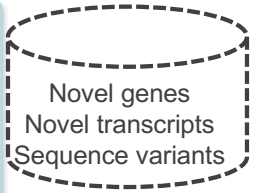


- 95 colorectal tumor samples
- 1,425 LC-MS/MS runs
- 124,823 peptides (FDR<1%)
- 7,526 proteins

Zhang et al., Nature, 2014

Case study 3: human colorectal cancer proteome (Zhang et al., 2014)





Biomarkers
Drug targets
Neoantigens

CustomizedProDB

R/Bioconductor
Wang et al., J Proteome Res, 2012
Wang & Zhang, Bioinformatics, 2013

Case study 3: human colorectal cancer proteome (Zhang et al., 2014)

- 1,065 variant peptides
- 796 unique Single Amino Acid Variants (SAAVs)
  - 64 in TCGA reported somatic variations
  - 101 in the COSMIC database
  - 526 in the dbSNP database
- 647 variant proteins
  - Known cancer genes: KRAS, CTNNB1, SF3B1, ALDH2, FH, etc.
  - Targets of FDA approved drugs: ALDH2, HSD17B4, PARP1, P4HB, TST, GAK, SLC25A24, SUPT16H, etc

Zhang et al., Nature, 2014

Overview

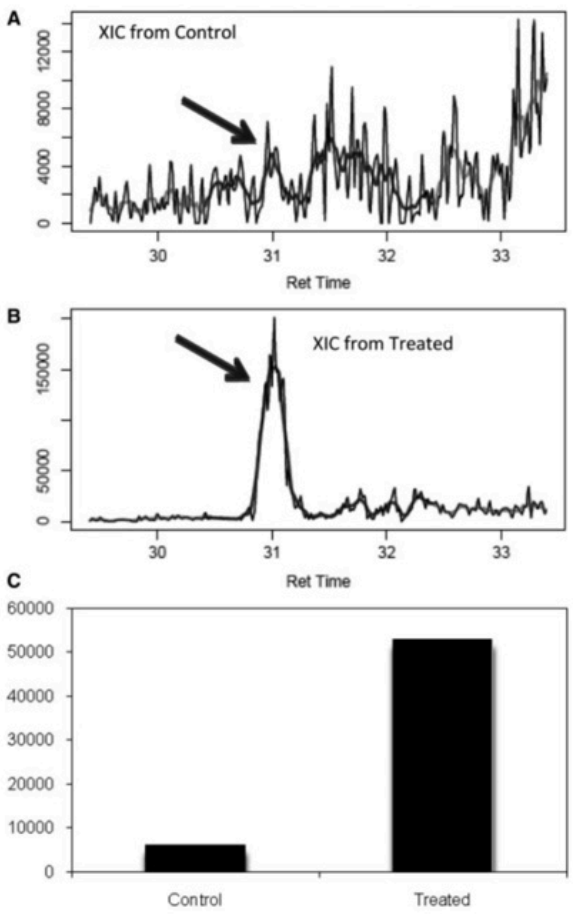
- Why proteomics
- Proteomics technology
- Protein identification
- **Protein quantification**
- Protein-protein interaction

Protein quantification methods

- Label-free
  - Ion intensity (MS1)
  - Spectral counting (MS2)
- Labeling
  - Metabolic labeling
    - SILAC (stable isotope labeling with amino acids in cell culture)
  - Chemical labeling
    - ICAT (isotope-coded affinity tag)
    - ITRAQ (isobaric tags for relative and absolute quantitation)
    - TMT (tandem mass tag)

Ion intensity (MS1)

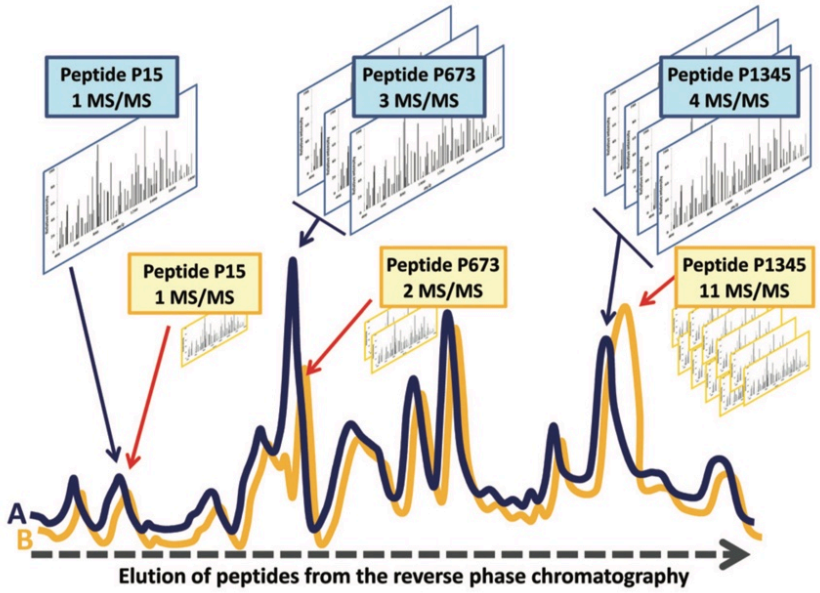
- Sum of the signal intensity of peptide precursor ions belonging to a particular protein.
- Extracted ion chromatograms (XICs) for the mass to charge ratios determined for each peptide.



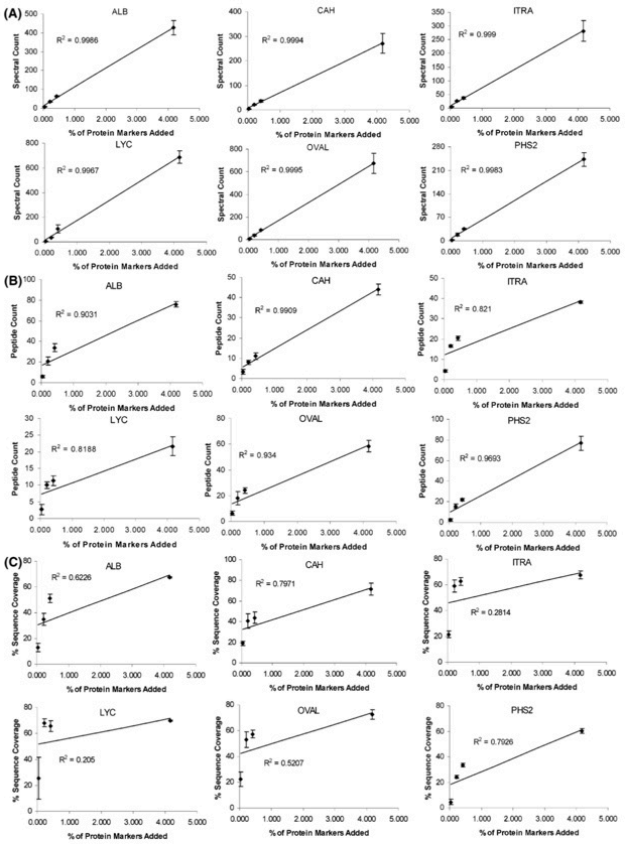
Wang et al., Brief Funct Genomic Proteomic, 2008

Spectral counting (MS2)

- Counting the number of fragment spectra identifying peptides of a given protein.



Armengaud, Environ Microbiol, 2013



Liu et al., Anal Chem, 2004

Spectral counting (MS2)

Spectral counting methods

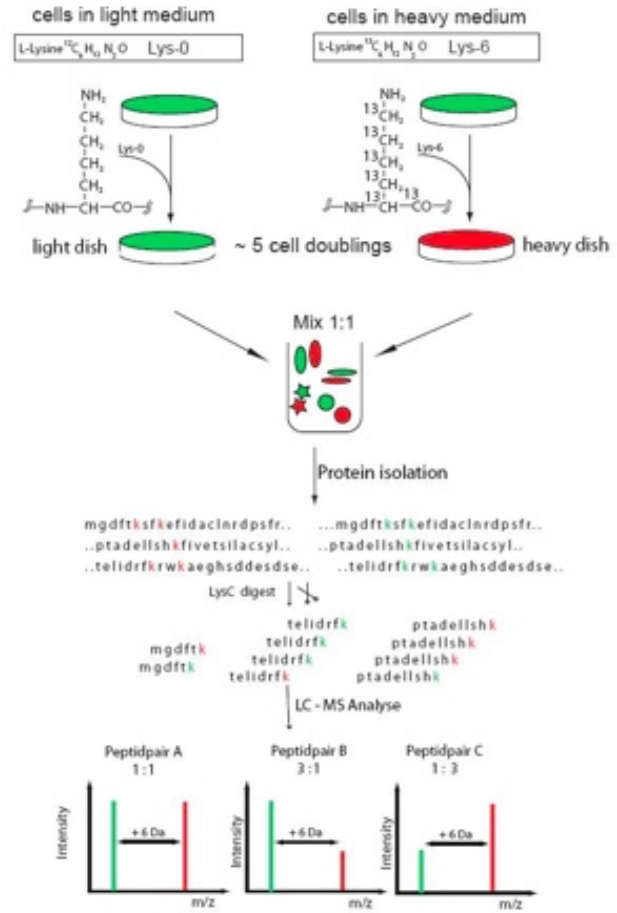
Table 1 Selected methods for label-free quantification based on spectrum counting

Method	Principle	Comments	References
Peptide count	Use of the number of peptides for an identified protein as a measure of abundance	Less useful than spectrum count	[80]
Spectrum count	Use of the number of PSMs for an identified protein as a measure of abundance	Higher dynamic range and better reproducibility than peptide counting	[81]
MS/MS intensity	Average total intensity of all fragment ion spectra matched to a protein	Intensity dimension provides extended dynamic range and better accuracy than spectrum counting	[90]
PAI and exponentially modified PAI	Protein abundance index and exponentially modified PAI (10^{PAI-1}). PAI is the number of identified peptides divided by the number of observable peptides	Implemented in some search engines such as Mascot. Designed for absolute quantification	[165]
Spectrum count/ molecular weight	Spectrum count is divided by the molecular weight of a protein	Similar to SAF	[162]
SAF	The spectrum count is normalized for protein length	Absolute and relative quantification	[163]
NSAF	SAF normalized for the sum of all protein abundances in the sample	Absolute and relative quantification	[89, 220]
RSC	Includes normalization of run-to-run variations and a correction factor	Relative quantification	[83]
APEX	Improves spectral counting by focusing on counting peptides that will likely be detected by MS techniques	Uses machine learning classification to derive peptide detection probabilities. This can be used to predict detectable peptides for any protein. Absolute and relative quantification	[145, 166]
SIn	Combines spectrum count with fragment-ion intensity (sum of all fragment ion intensities of all PSMs for a protein)	Variants include normalization by the sum of all protein spectral indexes in an experiment and normalization for the length of the protein. Absolute and relative quantification	[91]
mSCI	Number of observed peptides divided by protein relative identification probability	Conceptually similar to APEX. Absolute and relative quantification	[231]
RIBAR	Average of log 2 peptide ratios for a protein. Peptide ratios are calculated using the sum of all fragment ion intensities across all PSMs	Pairwise relative protein quantification	[92]

APEX absolute protein expression, *MS* mass spectrometry, *mSCI* modified spectrum count index, *MS/MS* tandem mass spectrometry, *NSAF* normalized spectral abundance factor, *PAI* protein abundance index, *PSM* peptide-to-spectrum match, *RIBAR* robust intensity-based averaged ratio, *RSC* relative spectral count, *SAF* spectral abundance factor, *SIn* normalized signal intensity

Bantscheff, et al., Anal Bioanal Chem. 2012

SILAC (stable isotope labeling with amino acids in cell culture)

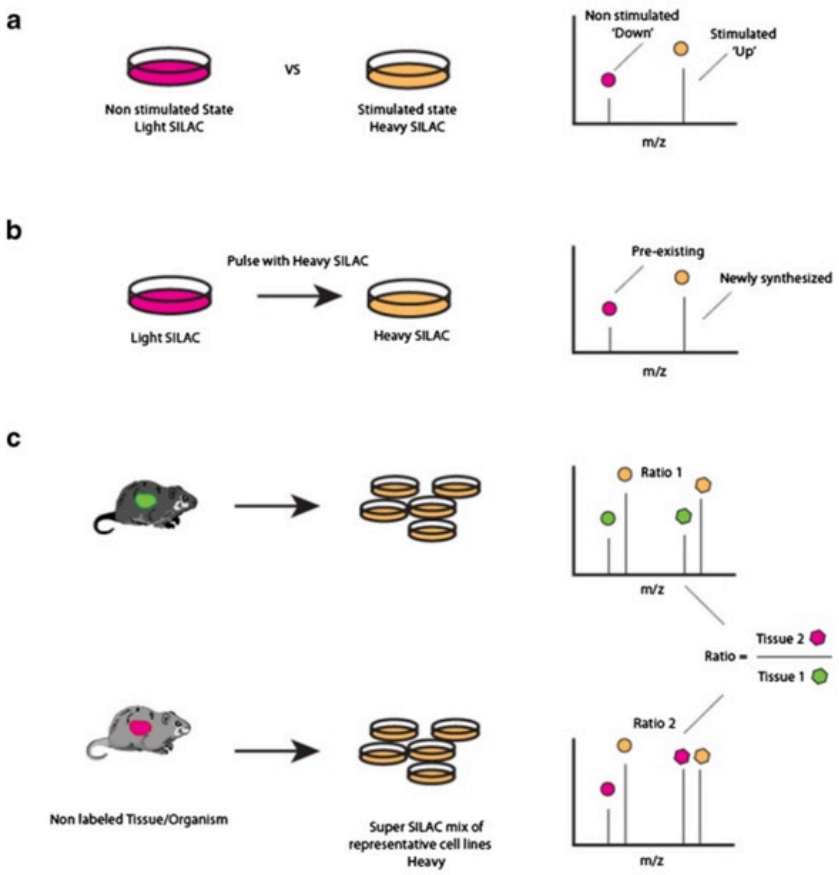


http://www.silantes.com/silac.htm

- Reagent
  - Stable-isotope-labeled amino acids
- Label and mix
  - Cell culture level
- Quantification
  - Ratio of heavy to low signals in peptide mass spectra (MS1)
- Cell culture only

Ong et al., MCP, 2002

SILAC (stable isotope labeling with amino acids in cell culture)



Bantscheff, et al., Anal Bioanal Chem. 2012

SILAC

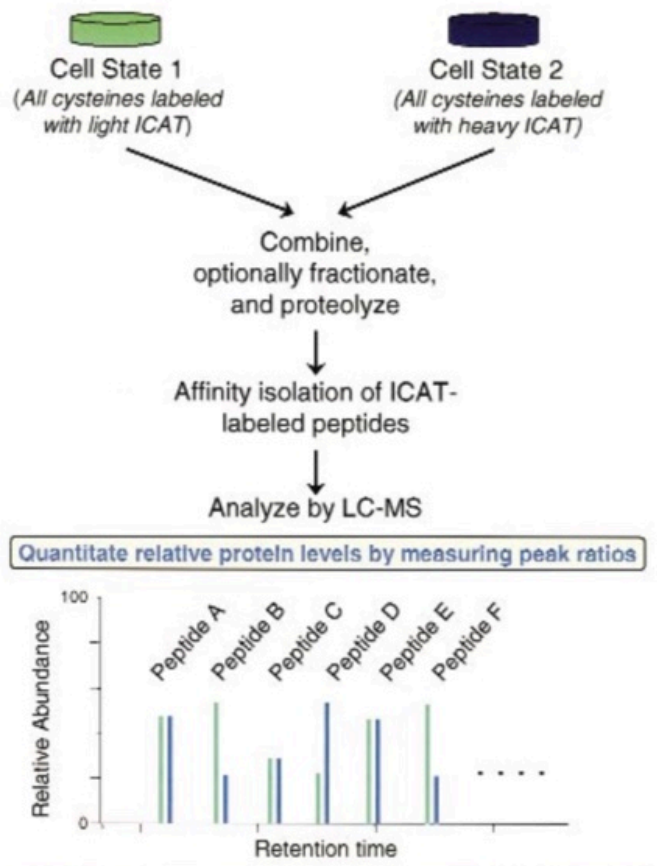
Pulsed SILAC

- Cells are grown in normal medium and pulse-labeled for a certain time with heavy SILAC.
- The ratio of heavy to low signals in such experiments is a measure of protein synthesis and degradation.

Super SILAC

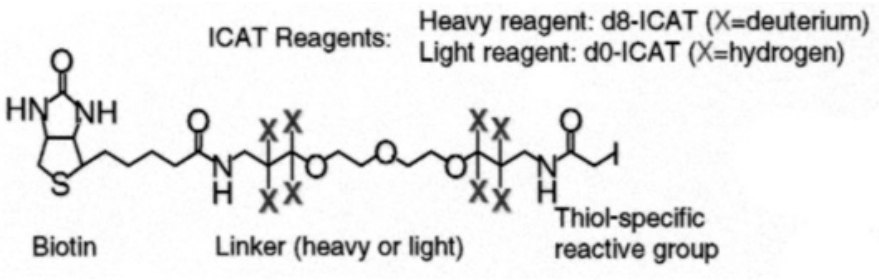
- Use a mixture of heavy-SILAC-labeled cell lines as a common internal standard.
- Unlabeled tissues from, *e.g.*, an animal model system, are compared with the common internal standard.
- The ratio of two (or more) such experiments allows comparison of protein expression between animals.

ICAT (isotope-coded affinity tag)



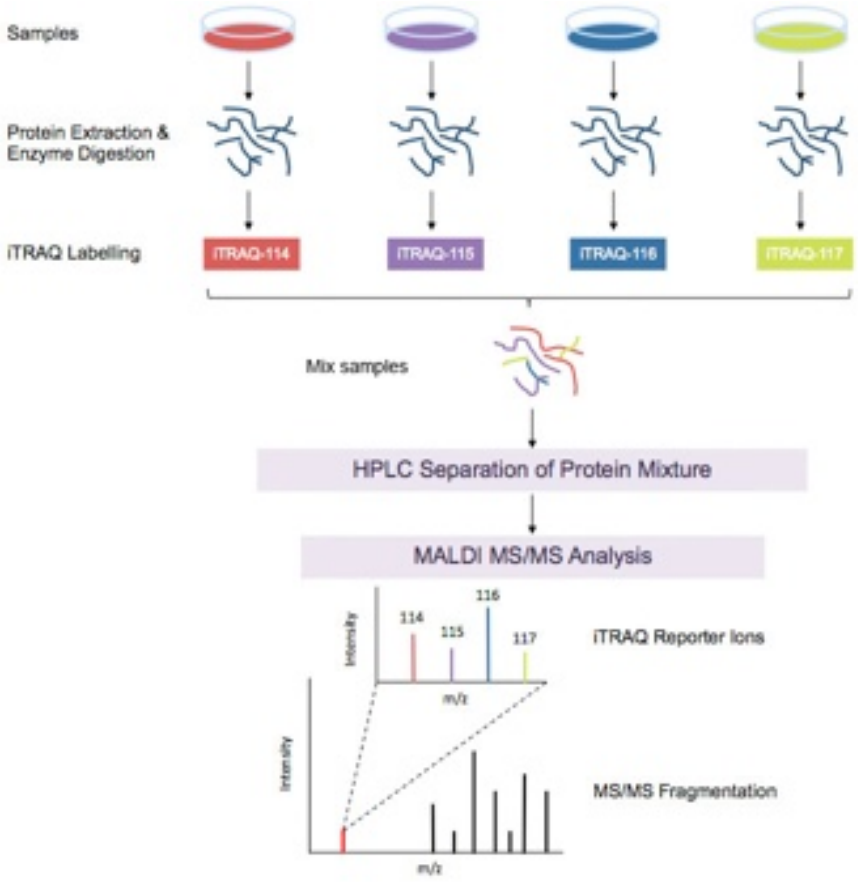
Gygi, et al., Nat Biotech, 1999

- Reagent



- Label and mix
  - Protein level
- Quantification
  - Ratio of heavy to light signals in peptide mass spectra (MS1)

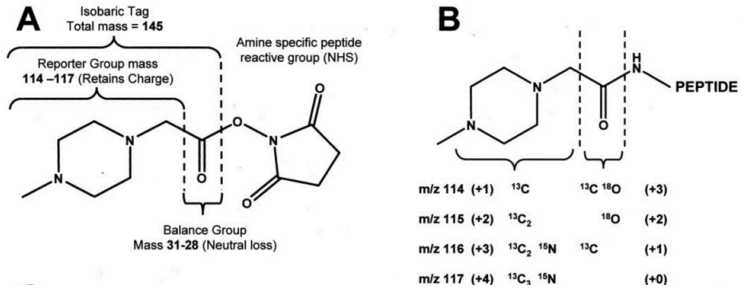
ITRAQ (isobaric tags for relative and absolute quantitation)



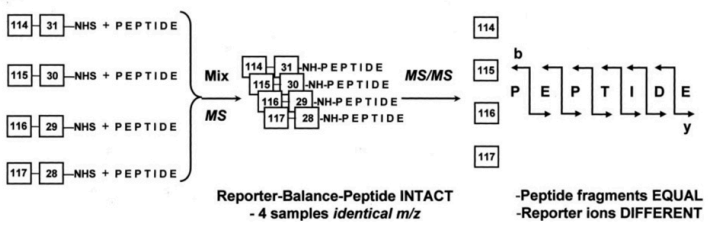
https://www.creative-proteomics.com/services/itraq-based-proteomics-analysis.htm

- Reagent

Ross, et al., MCP, 2004

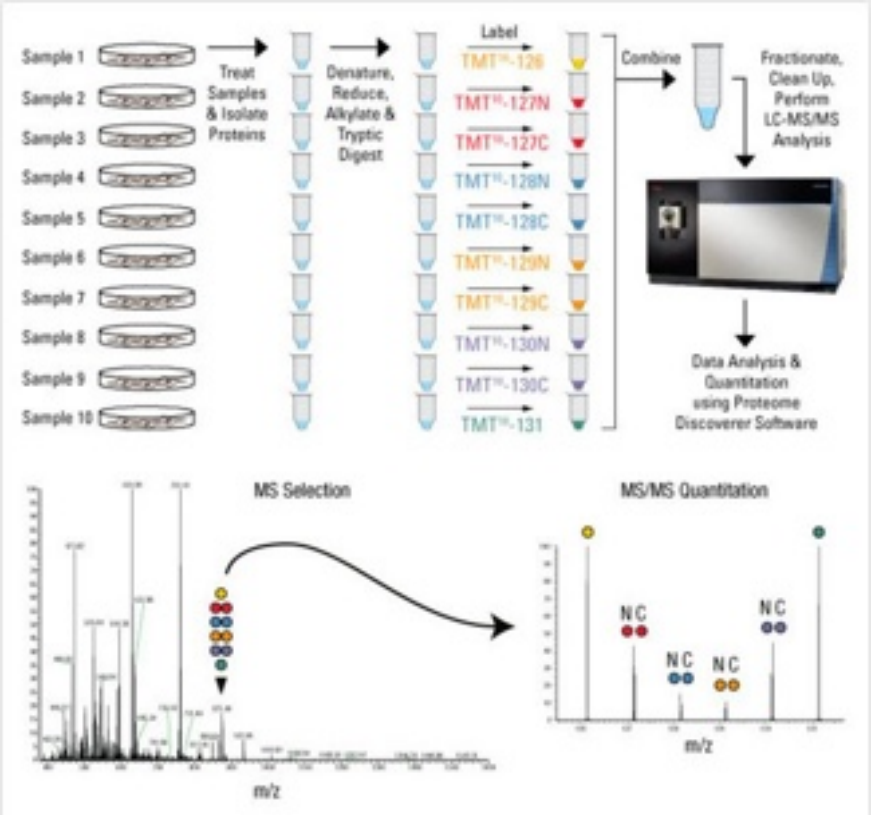


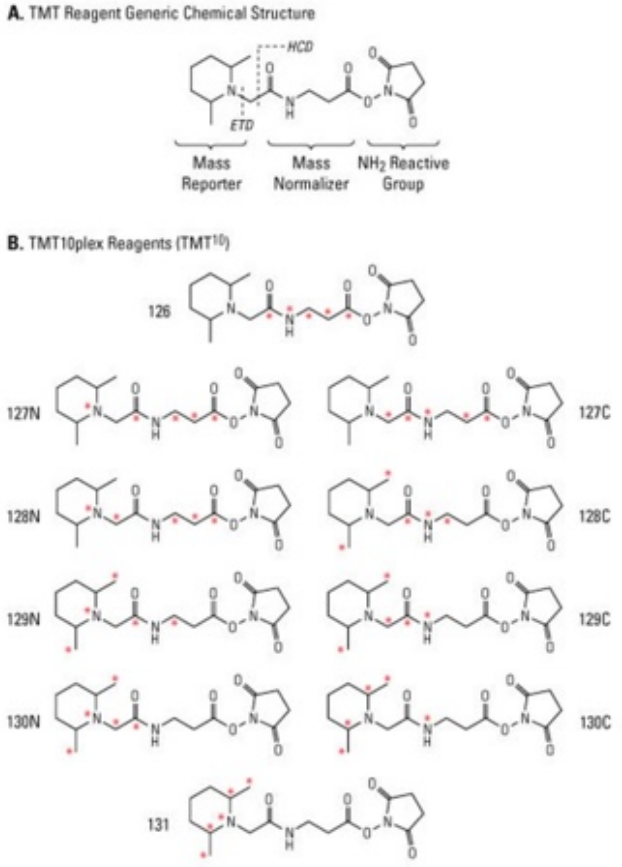
- Label and mix
  - Peptide level



- Quantification
  - Relative intensities of the reporter ions (MS2)

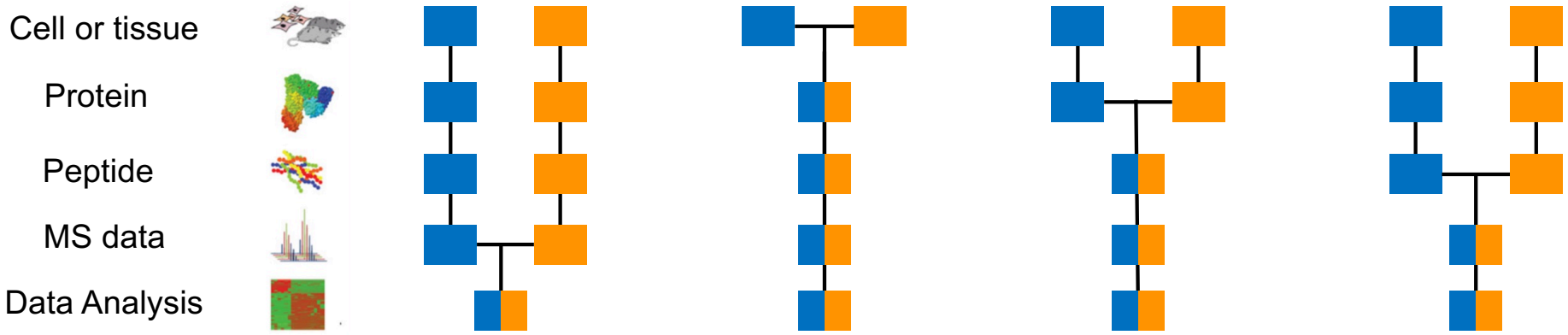
TMT (tandem mass tag)





https://www.thermofisher.com/order/catalog/product/90110

Proteomic quantification methods comparison



	Label-free	SILAC	ICAT	ITRAQ / TMT
Label	None	Metabolic	Chemical	Chemical
Sample type	Cell or tissue	Cell	Cell or tissue	Cell or tissue
Label and mix	Data	Cell culture	Protein	Peptide
Quantitative accuracy	+	++++	+++	++
Comparisons	Many	2-3	2	2-11
Quantitative level	MS1 or MS2	MS1	MS1	MS2
Quantification data	Intensity or count	ratio	ratio	ratio
Dynamic range	3-4	1-2	1-2	2

Schulze and Usadel, Annu Rev Plant Biol, 2010

Bantscheff, et al., Anal Bioanal Chem. 2012

Software for MS-based protein quantification

Label-free quantification		Labeling-based quantification	
Commercial	Open	Commercial	Open
• Mascot • ProteinPilot • Scaffold • SIEVE • Progenesis LC-MS • ProteoIQ • PEAKS Q	• **MaxQuant** • IDEAL-Q • PVIEW • TPP • LFQuant • Skyline • OpenMS • APEX • QSPEC	• Mascot • ProteinPilot • Scaffold • Proteome Discoverer • PEAKS Q	• MaxQuant • TPP • Multi-Q • iQuantitator • IsobariQ • Isobar • Diffprot • MilQuant • Iquant • PyQuant

Slide Courtesy of Bo Wen

Proteomic quantification data

Samples

Proteins

RefSeq ID	C01CO005	C01CO006	C01CO008	C01CO013	C01CO014	C01CO015	C01CO019	C01CO022	C05CO002	C05CO003	.....
NP_000005	-0.472	-0.325	-1.863	-0.334	-0.642	-0.974	-0.921	1.426	-1.34	-0.45	.....
NP_000007	0.407	0.458	0.591	-0.142	-0.267	-0.344	-1.455	-0.38	-0.303	-0.255	.....
NP_000008	-0.203	0.409	0.347	-0.889	-0.532	-0.95	-0.695	-0.837	-0.184	-0.286	.....
NP_000009	0.224	0.304	0.657	0.314	0.297	-0.374	-0.324	-0.883	-0.404	0.299	.....
NP_000010	0.154	-0.075	0.176	-0.019	-0.122	-0.691	-0.167	-0.703	-0.882	-0.473	.....
NP_000012	0.414	0.247	1.004	0.876	-0.236	0.156	0.486	-0.105	0.461	-0.673	.....
NP_000013	-0.366	-0.022	-0.494	-0.138		-0.432	-0.438	0.069	-0.424	0.136	.....
NP_000017	-0.134	0.241	-0.53	0.152	-0.117	-0.127	0.008	0.32	-0.464	-0.411	.....
NP_000019	-0.558	-0.091	-0.337	-0.341	-0.28	-0.411	-0.569	0.159	-0.46	0.142	.....
NP_000020	-0.375	-0.041	-1.226	-0.129	-0.059	-0.741	-0.547	-0.085	-0.404	-0.343	.....
NP_000022	-0.245	-0.706	-0.316	0.27	-0.46	-0.164	-0.1	-0.169	-0.01	0.235	.....
NP_000024	-0.082	-1.295	-0.162	-0.076	-0.549		-0.141		-0.229		.....
NP_000025	-0.497	-0.647	-0.356	-0.358	0.988	-0.516	0.085	0.09	-0.193	0.256	.....
NP_000026	-0.312	-0.454	-1.411	-0.371	1.331	-0.441	0.08	-0.041	-0.208	0.342	.....
NP_000030	-0.403	-0.561	-1.597	-0.203	-0.251	-1.014	-0.788	0.547	-1.294	0.124	.....
NP_000031	0.097	-0.872	-0.598	0.595	-1.314	-0.908	-1.078	0.174	-1.217	1.737	.....
NP_000032	-1.046	0.169	-1.523	-0.09	-0.876	-0.969	-0.89	0.03	-0.772	0.618	.....
NP_000033	-0.605	-1.117	-0.818	-1.262	-0.362	-1.019	-0.417	0.373	-0.919	0.01	.....
NP_000034	-0.258	0.52	0.068	-0.099	0.256	-0.749	-0.025	-0.388	-0.616	-0.003	.....
NP_000036	-0.386	-2.142	-1.689	-0.416	-0.267	-0.652	-0.738	0.847	-0.24	1.742	.....
NP_000037	-0.023	0.553	-0.068	-0.711	-0.56	-0.483	-0.177	-0.338	-0.296	-0.149	.....
NP_000038	-0.524	0.185	0.114	-0.444	-0.19	-0.093	-0.346	-0.739	0.079	-1.757	.....
NP_000039	0.517	0.321	0.282	0.047	0.062	-0.282	0.392	-0.198	0.054	-0.359	.....
NP_000040	-0.168			-2.599							.....
NP_000041	1.048	0.004	-0.205	0.976	1.517	-0.894	0.308	0.302	0.932	0.181	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....

- Data preprocessing
  - QC and filtering
  - Missing value imputation
  - Normalization
  - Transformation
  - Aggregation to genes (optional)

Applications

- **Differential expression** (supervised analysis)
  - Input: protein expression data, class label of the samples
  - Output: differentially expressed proteins
  - *e.g.*, disease biomarker discovery; biological mechanisms
- **Clustering** (unsupervised analysis)
  - Input: protein expression data
  - Output: groups of similar samples or proteins
  - *e.g.*, disease subtype identification; co-expressed protein group identification
- **Prediction** (machine learning)
  - Input: protein expression data, class label of the samples (training data)
  - Output: prediction model
  - *e.g.*, disease diagnosis and prognosis

Case study 1: protein biomarker discovery

- Question
  - To identify novel protein biomarkers in stool for detecting colorectal cancer (CRC) and advanced adenomas.
- Data
  - Series 1: spectral count data from 12 CRC vs 10 control
  - Series 2: spectral count data from 81 CRC, 40 advanced adenomas, 43 low-level adenomas, and 129 control
- Analysis
  - Differential expression analysis: beta-binomial model + multiple test adjustment
  - Validation of selected proteins using antibody-based assays in series 3 with 14 CRC, 16 advanced adenomas, 18 low-level adenomas, and 24 control

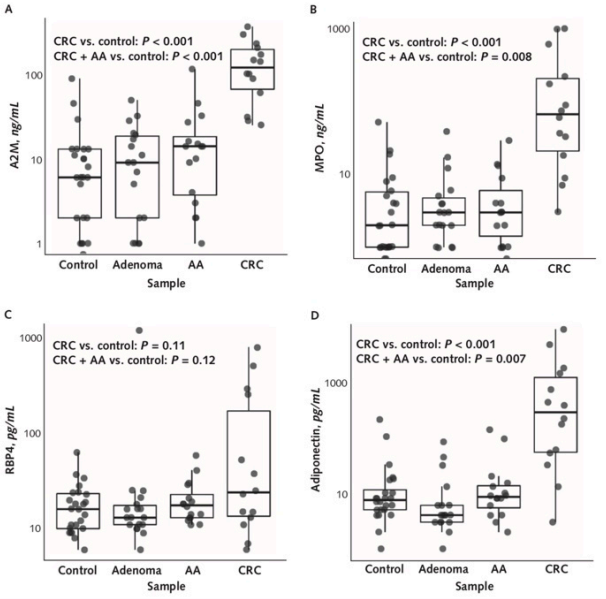
Bosch et al., Ann Intern Med, 2017

Case study 1: protein biomarker discovery

Table. Twenty-Nine Candidate Proteins: Mean Spectral Counts per Sample With Positive Results, Fold Changes, and *P* and *Q* Values

Gene Symbol	HGNC Gene Name	UniProt Identifier	Samples in Series 1							Samples in Series 2						
			Control (n = 10)		CRC (n = 12)		CRC vs. Control*			Control (n = 129)		CRC (n = 79)		CRC vs. Control*		
			Total Samples With Positive Results, n	Mean Spectral Counts per Sample With Positive Results	Total Samples With Positive Results, n	Mean Spectral Counts per Sample With Positive Results	Fold Change	*P* Value	*Q* Value	Total Samples With Positive Results, n	Mean Spectral Counts per Sample With Positive Results	Total Samples With Positive Results, n	Mean Spectral Counts per Sample With Positive Results	Fold Change	*P* Value	*Q* Value
A2M	Alpha-2-macroglobulin	P01023	6	10.7	12	120.4	34.0	<0.001	0.001	118	21.1	79	145.5	5.7	<0.001	<0.001
S100A8	S100 calcium binding protein A8	P05109	8	3.9	12	10.8	3.9	<0.001	0.001	128	6.5	79	15.6	1.8	<0.001	<0.001
S100A9	S100 calcium binding protein A9	P06702	9	2.8	12	10.5	5.0	<0.001	0.001	129	9	79	23.1	1.9	<0.001	<0.001
CP	Ceruloplasmin	P00450	1	1	11	32.9	583.5	<0.001	0.001	38	4.4	64	21.8	11.1	<0.001	<0.001
TF	Transferrin	P02787	1	1	11	68.4	1227.2	<0.001	0.002	21	9.6	53	61.2	21.9	<0.001	<0.001
CAT	Catalase	P04040	0	0	9	22.2	∞	<0.001	0.002	48	5.1	59	20.3	5.3	<0.001	<0.001
C9	Complement C9	P02748	0	0	9	3.4	∞	<0.001	0.002	2	1.5	12	2.4	11.2	<0.001	0.003
LTF	Lactotransferrin	P02788	2	3.5	10	34.3	80.5	<0.001	0.002	109	13.9	79	57	3.78	<0.001	<0.001
HBB	Hemoglobin subunit beta	P68871	6	10.5	12	73.2	9.9	<0.001	0.002	68	12.1	77	49.9	5.8	<0.001	<0.001
HPX	Hemopexin	P02790	0	0	8	14.8	∞	<0.001	0.002	31	5	59	29.7	13.7	<0.001	<0.001
HBA1	Hemoglobin subunit alpha 1	P69905	2	1	10	46.7	79.9	<0.001	0.003	52	8.7	72	36.2	7.9	<0.001	<0.001
HP	Haptoglobin	P00738	5	3.8	11	55.5	17.0	<0.001	0.003	24	9.9	69	39.6	17.4	<0.001	<0.001
GPI	Glucose-6-phosphate isomerase	P06744	0	0	8	4.5	∞	<0.001	0.005	60	5.3	60	12.9	2.8	<0.001	<0.001
MPO	Myeloperoxidase	P05164-2	0	0	7	6	∞	<0.001	0.005	85	8.7	77	29.5	3.8	<0.001	<0.001
HBD	Hemoglobin subunit delta	P02042	0	0	7	13.4	∞	<0.001	0.005	8	6	33	7.7	8.2	<0.001	<0.001
C3	Complement C3	P01024	4	3.5	11	99.8	68.0	<0.001	0.006	103	17.6	79	112.5	6.6	<0.001	<0.001
SERPINF2	Serpin family F member 2	P08697-2	1	5	9	4.3	13.2	<0.001	0.008	93	3	71	10.6	3.2	<0.001	<0.001
CDA	Cytidine deaminase	P32320	0	0	8	1.4	∞	<0.001	0.008	65	1.5	63	2.5	2.0	<0.001	<0.001
FGG	Fibrinogen gamma chain	P02679-2	0	0	6	10.8	∞	0.001	0.015	1	1	20	4.7	218.3	<0.001	<0.001
AZU1	Azurocidin 1	P20160	0	0	6	1.3	∞	0.001	0.016	63	2.7	72	6	3.1	<0.001	<0.001
VTN	Vitronectin	P04004	0	0	6	2.5	∞	0.002	0.021	3	2.7	31	5.1	38.1	<0.001	<0.001
RBP4	Retinol binding protein 4	Q5VY30	2	4	8	5.6	7.4	0.002	0.022	37	3	68	8.1	5.6	<0.001	<0.001
KNG1	Kininogen 1	P01042-2	0	0	5	2	∞	0.002	0.022	2	1	21	2.4	39.2	<0.001	<0.001
PSMA5	Proteasome subunit alpha 5	P28066	0	0	5	1.8	∞	0.004	0.034	36	2.3	52	2.8	1.8	<0.001	<0.001
C5	Complement C5	P01031	0	0	5	8.6	∞	0.004	0.034	11	6.5	57	10.4	13.0	<0.001	<0.001
FN1	Fibronectin 1	P02751-11	1	2	8	8.9	69.9	0.005	0.037	18	5.1	61	18.9	17.7	<0.001	<0.001
LDHA	Lactate dehydrogenase A	P00338	0	0	5	2.6	∞	0.006	0.043	24	1.6	29	3.2	2.3	0.014	0.044
PRTN3	Proteinase 3	P24158	1	1	6	2.8	22.3	0.006	0.046	72	3.3	71	6.6	2.4	<0.001	<0.001
GSR	Glutathione-disulfide reductase	P00390-2	1	1	7	5.6	31.5	0.007	0.049	93	3.6	70	6.8	1.5	0.002	0.007

Figure 3. Biomarker detection in FIT fluids from sample series 3.



Antibody-based assays

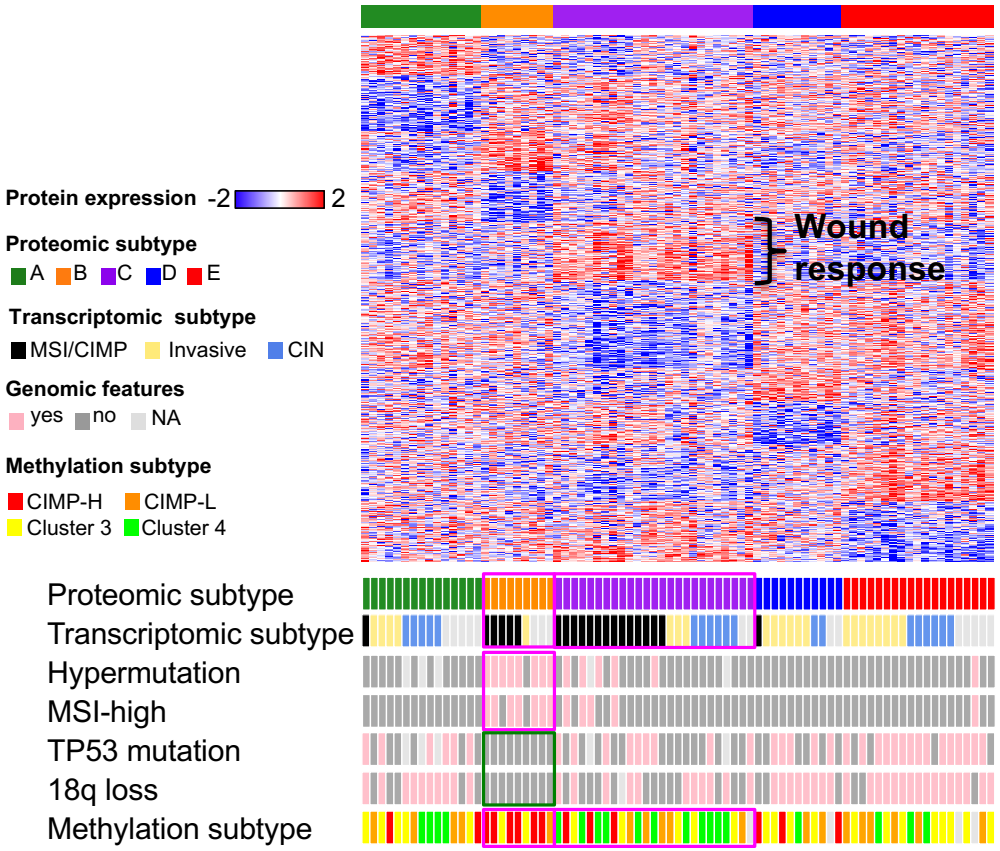
BCM QCB, April 2018

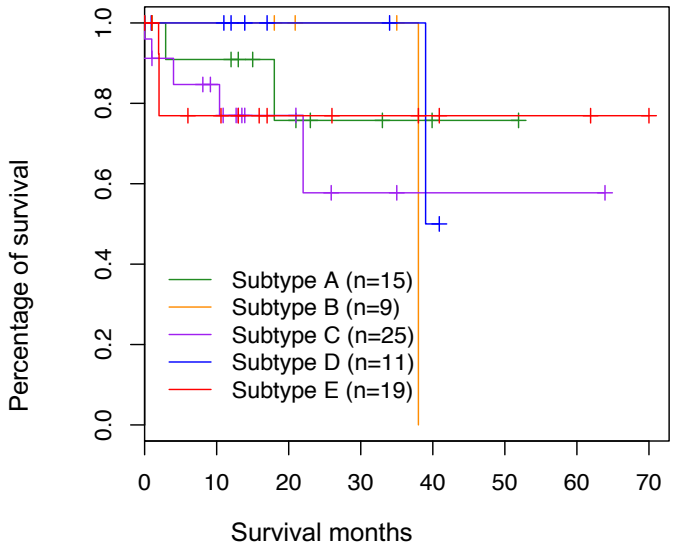
Case study 2: tumor subtyping

- Question
  - Can we stratify CRC tumors based on proteomics data?
  - Is the stratification biologically and clinically meaningful?
- Data
  - Spectral count data from 90 CRC tumors
- Analysis
  - Clustering analysis: consensus clustering of the samples
  - Biological interpretation: enrichment analysis
  - Clinical interpretation: survival analysis

Zhang et al., Nature, 2014

Case study 2: tumor subtyping

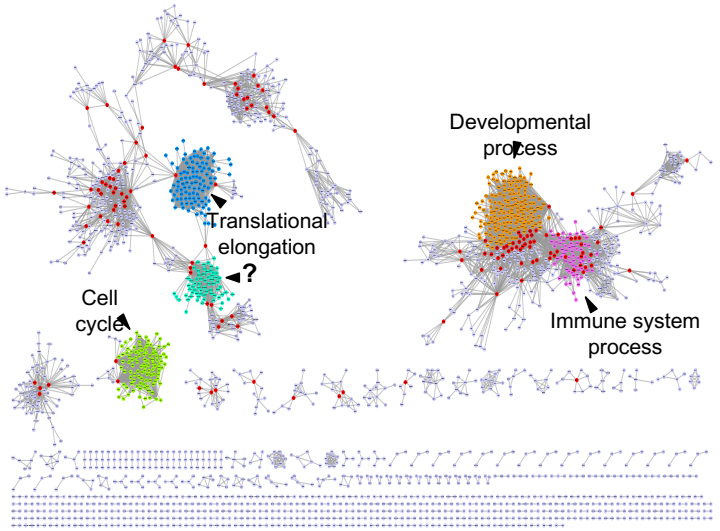




Zhang et al., Nature, 2014

Case study 3: co-expression network and co-expression modules

- Question
  - Are mRNA and protein co-expression networks wired similarly?
- Data
  - 77 breast tumors: RNA-Seq and ITRAQ
  - 87 colorectal tumors: RNA-Seq and Label-free
  - 174 ovarian tumors: microarray and ITRAQ
- Analysis
  - Co-expression network construction: k-nearest neighbor
  - Co-expression module identification: NetSAM
  - Biological interpretation: enrichment analysis

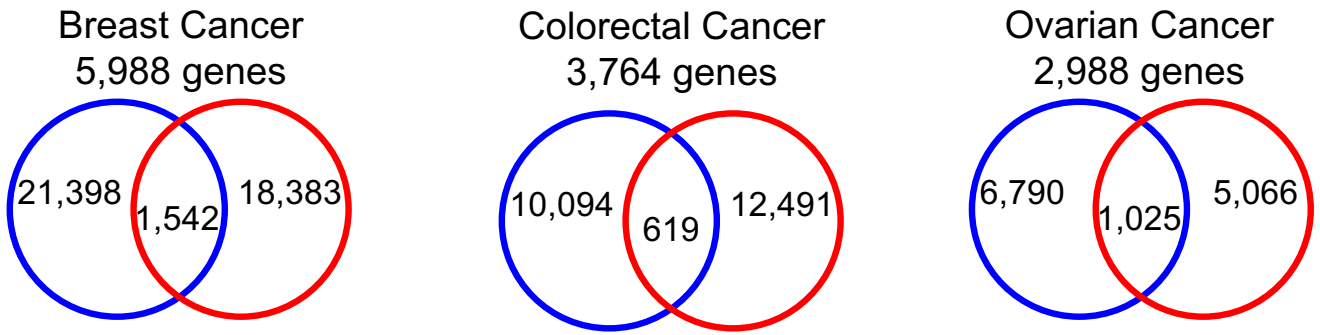


Wang et al. MCP, 2017

Tripathi et al. Cancer Res, 2014

Case study 3: co-expression network and co-expression modules

Protein and mRNA networks are wired differently



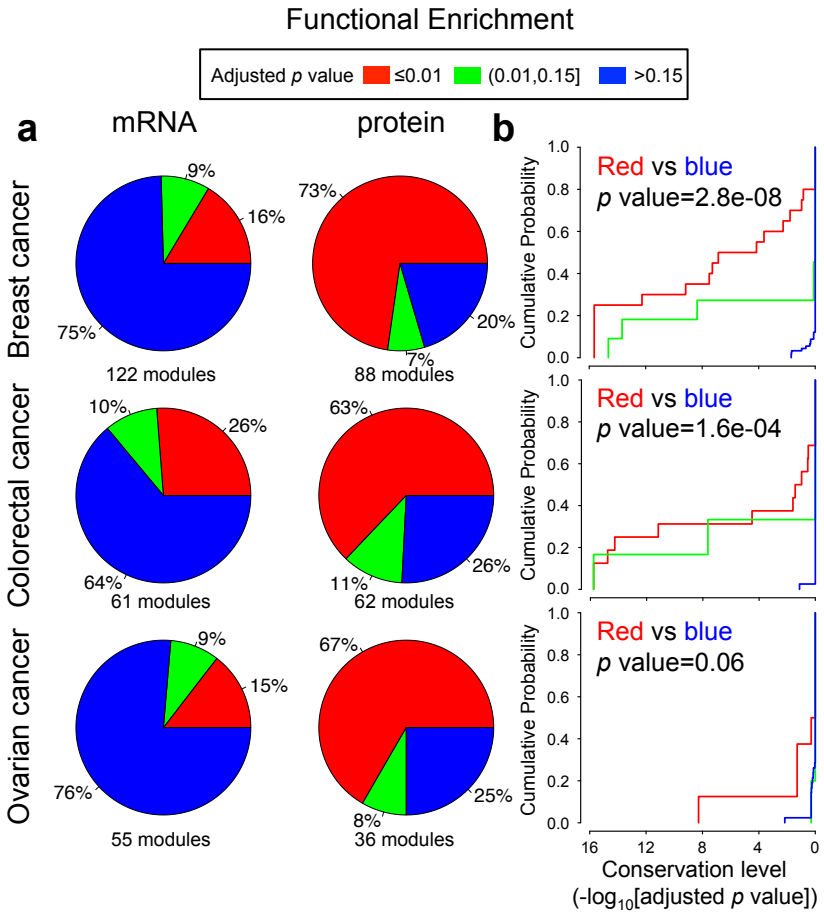
Edges in mRNA co-expression network

Edges in protein co-expression network

Edge overlap between mRNA and protein co-expression networks

Wang et al. MCP, 2017

Case study 3: co-expression network and co-expression modules

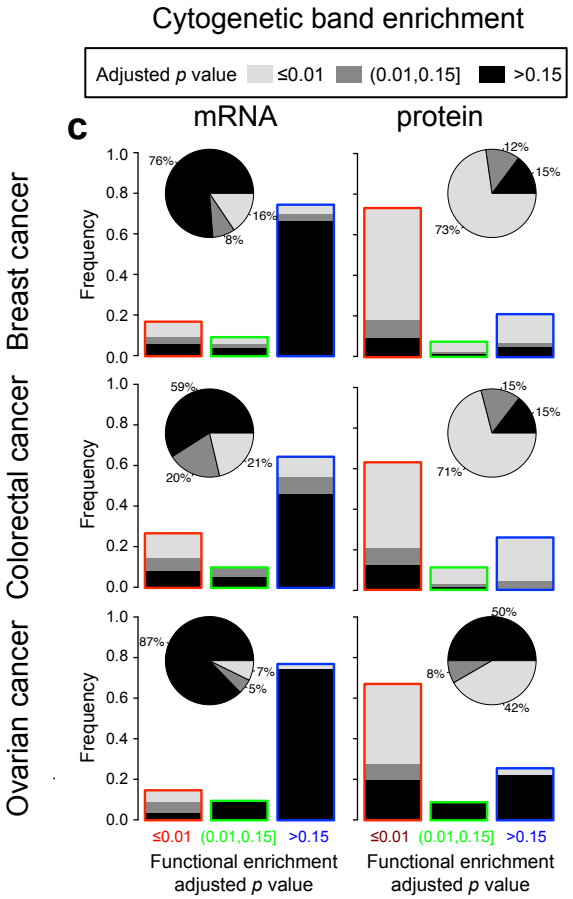


- More protein modules showed significant functional enrichment
- Functionally coherent mRNA modules are more likely to be preserved in protein networks

Protein modules *vs* mRNA modules

Wang et al. MCP, 2017

Case study 3: co-expression network and co-expression modules



- More protein modules showed significant functional enrichment
- Functionally coherent mRNA modules are more likely to be preserved in protein networks
- More mRNA modules showed significant cytogenetic band enrichment

Protein modules *vs* mRNA modules

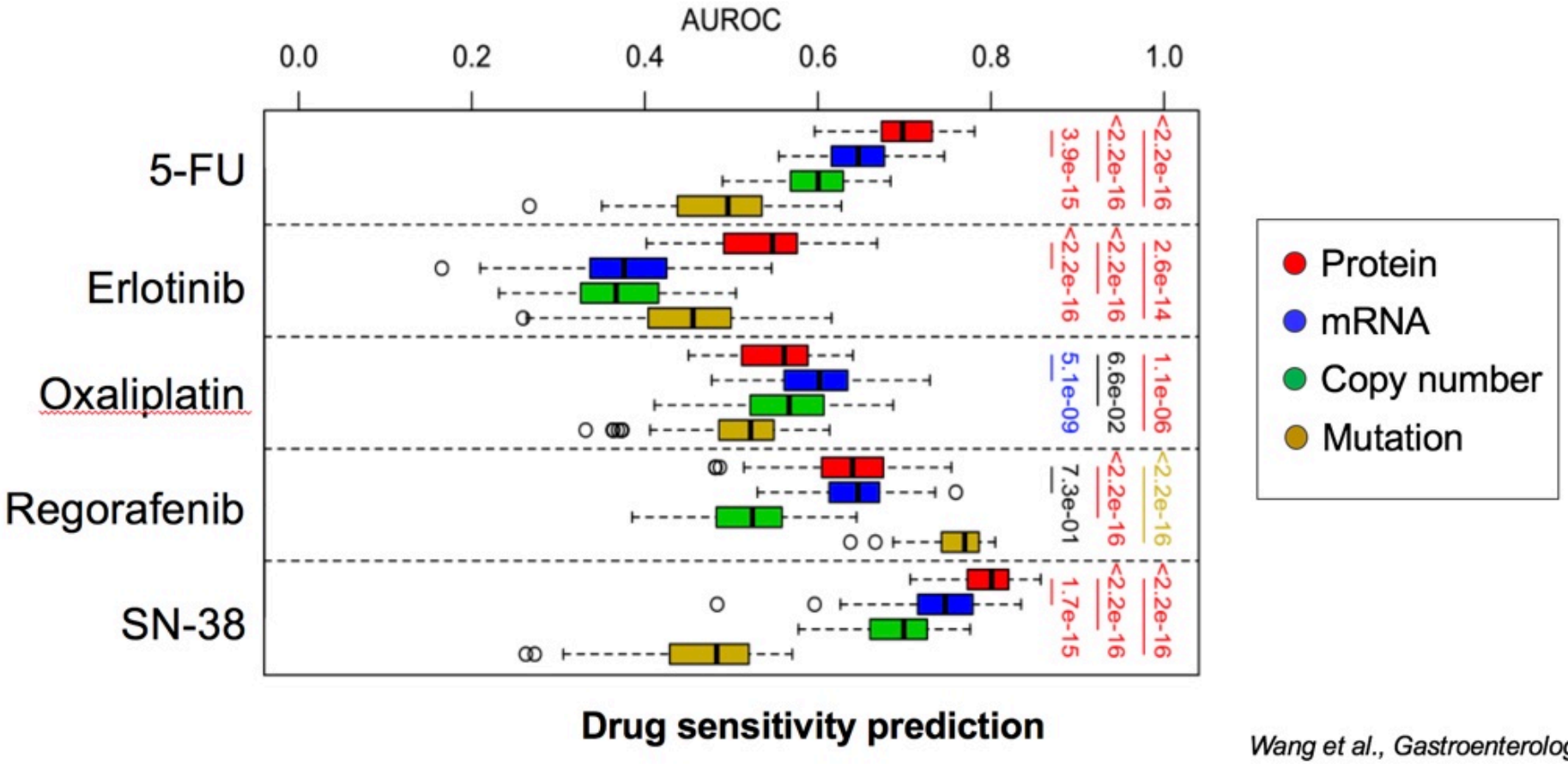
Wang et al. MCP, 2017

Case study 4: drug sensitivity prediction

- Question
  - Can omics data predict drug sensitivity of colorectal cancer (CRC) cell lines?
- Data
  - 44 CRC cell lines: mutation, copy number, RNA-Seq, label-free proteomics, drug sensitivity
- Analysis
  - Random forests + five-fold cross validation

Wang et al., Gastroenterology, 2017

Case study 4: drug sensitivity prediction



Drug sensitivity prediction

BCM QCB, April 2018

Wang et al., Gastroenterology, 2017

Overview

- Why proteomics
- Proteomics technology
- Protein identification
- Protein quantification
- **Protein-protein interaction**

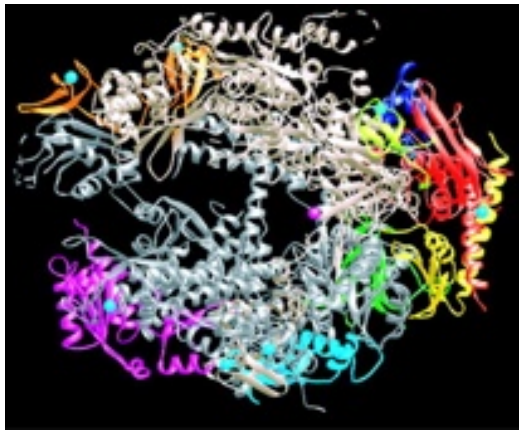
Protein-protein interaction (PPI)

- Definition
  - Physical association of two or more protein molecules
- Examples
  - Receptor-ligand interactions
  - Kinase-substrate interactions
  - Transcription factor-co-activator interactions
  - Multiprotein complex, *e.g.*, multimeric enzymes

Significance of protein interaction

- Most proteins mediate their function through interacting with other proteins
  - To form molecular machines
  - To participate in various regulatory processes
- Distortions of protein interactions can cause diseases

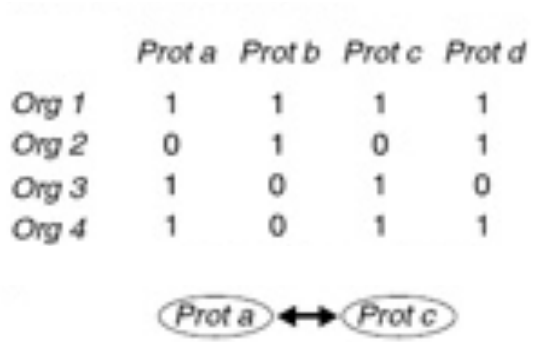
RNA polymerase II, 12 subunits



Cramer et al. Science 292:1863, 2001

Protein-protein interaction identification

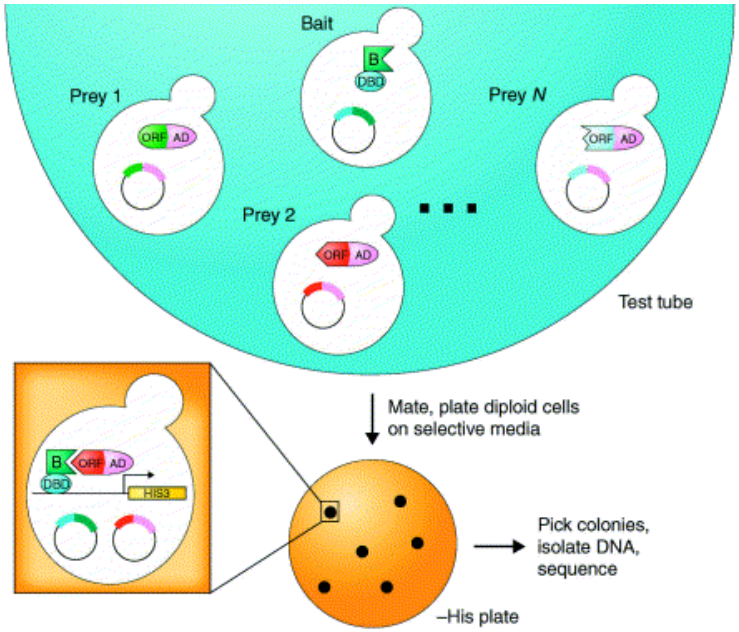
- Computational
  - Gene fusion
  - Conservation of gene neighborhood
  - Phylogenetic profiling
  - Co-evolution
  - Co-expression
  - Ortholog interaction
  - Domain interaction
- Experimental
  - Yeast two-hybrid
  - Tandem affinity purification



	Prot a	*Prot b*	*Prot c*	*Prot d*
Org 1	1	1	1	1
Org 2	0	1	0	1
Org 3	1	0	1	0
Org 4	1	0	1	1

Valencia et al. Curr. Opin. Struct. Biol, 12:368, 2002

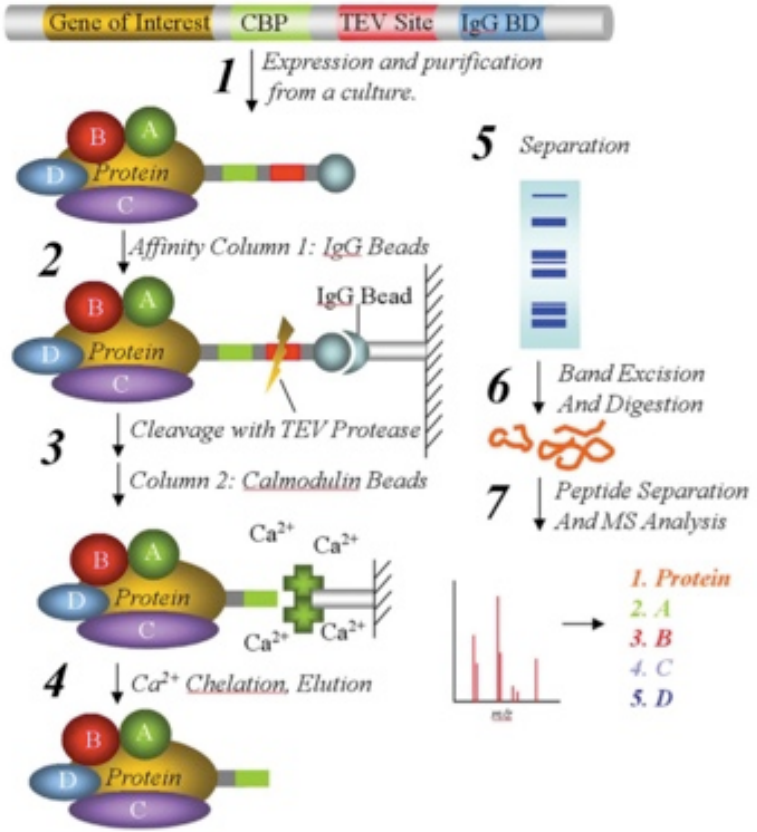
Yeast two-hybrid



Uetz P. Curr Opin Chem Biol. 6:57, 2002

- Method
  - Bait strain: a protein of interest, bait (B), fused to a DNA-binding domain (DBD)
  - Prey strains: ORFs fused to a transcriptional activation domain (AD)
  - Mate the bait strain to prey strains and plate diploid cells on selective media (e.g. without Histidine)
  - If bait and prey interact in the diploid cell, they reconstitute a transcription factor, which activates a reporter gene whose expression allows the diploid cell to grow on selective media
  - Pick colonies, isolate DNA, and sequence to identify the ORF interacting with the bait
- Pros
  - High-throughput
  - Can detect transient interactions
- Cons
  - False positives
  - Non-physiological (done in the yeast nucleus)
  - Can't detect multiprotein complexes

Tandem affinity purification



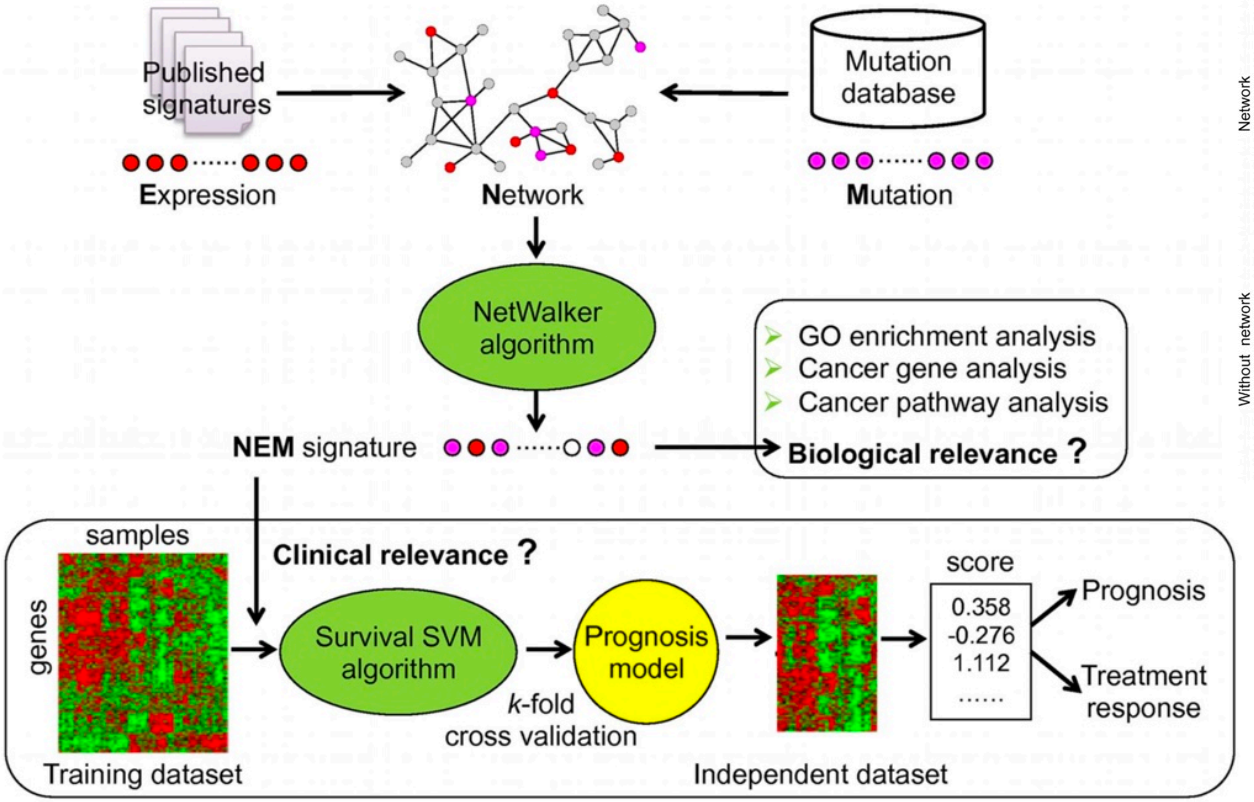
Chepelev et al. Biotechnol & Biotechnol 22:1, 2008

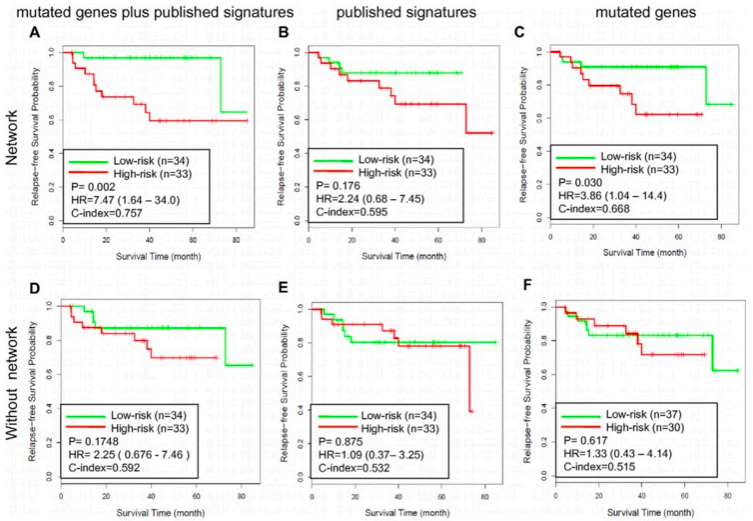
- Method
  - TAP tag: IgG binding domain, TEV protease cleavage site, calmodulin binding domain,
  - Bait protein gene is fused with the DNA sequences encoding TAP tag
  - Tagged bait is expressed in cells and forms native complexes
  - Complexes purified by TAP method
  - Components of each complex are identified through gel separation followed by MS/MS
- Pros
  - High-throughput
  - Physiological setting
  - Can detect large stable protein complexes
- Cons
  - High false positives
  - Can't detect transient interactions
  - Can't detect interactions not present under the given condition
  - Tagging may disturb complex formation
  - Binary interaction relationship is not clear

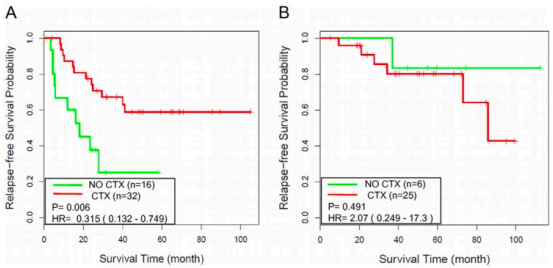
PPI data in the public domain

- Database of Interacting Proteins (DIP)
  http://dip.doe-mbi.ucla.edu/
- The Molecular INTeraction database (MINT)
  http://mint.bio.uniroma2.it/mint/
- The Biomolecular Object Network Databank (BOND)
  http://bond.unleashedinformatics.com/
- The General Repository for Interaction Datasets (BioGRID)
  http://www.thebiogrid.org/
- Human Protein Reference Database (HPRD)
  http://www.hprd.org
- Online Predicted Human Interaction Database (OPHID)
  http://ophid.utoronto.ca
- iRef
  http://wodaklab.org/iRefWeb
- The International Molecular Exchange Consortium (IMEX)
  http://www.imexconsortium.org

Case study 1: network-based gene expression signature



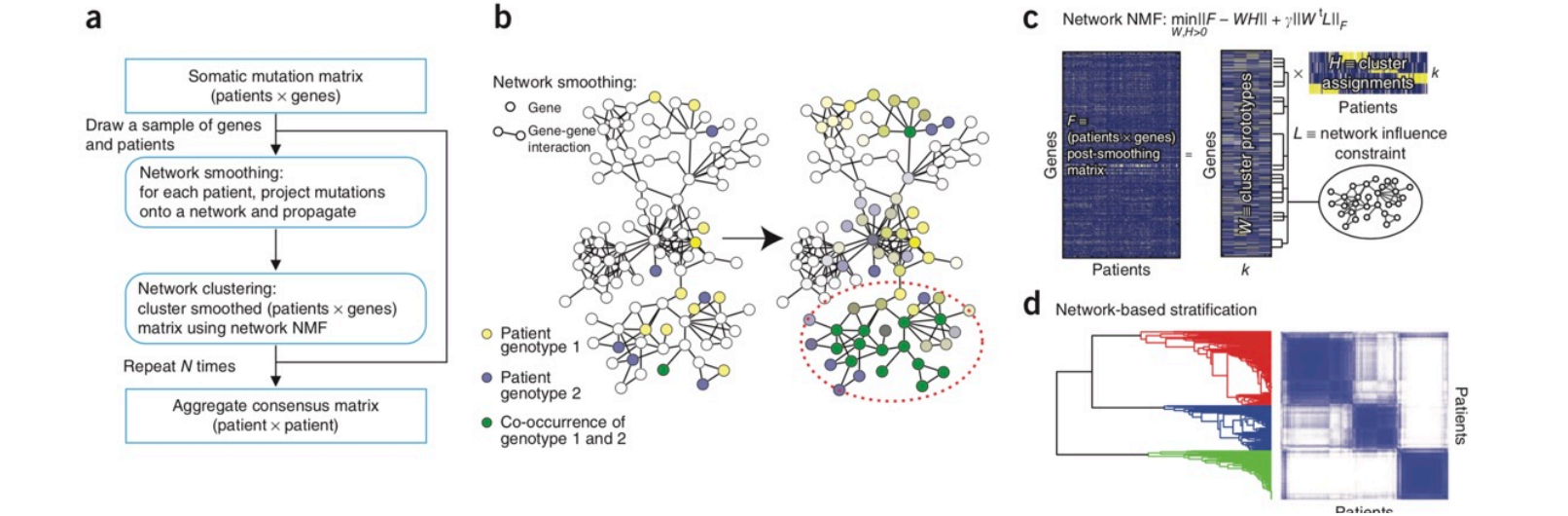


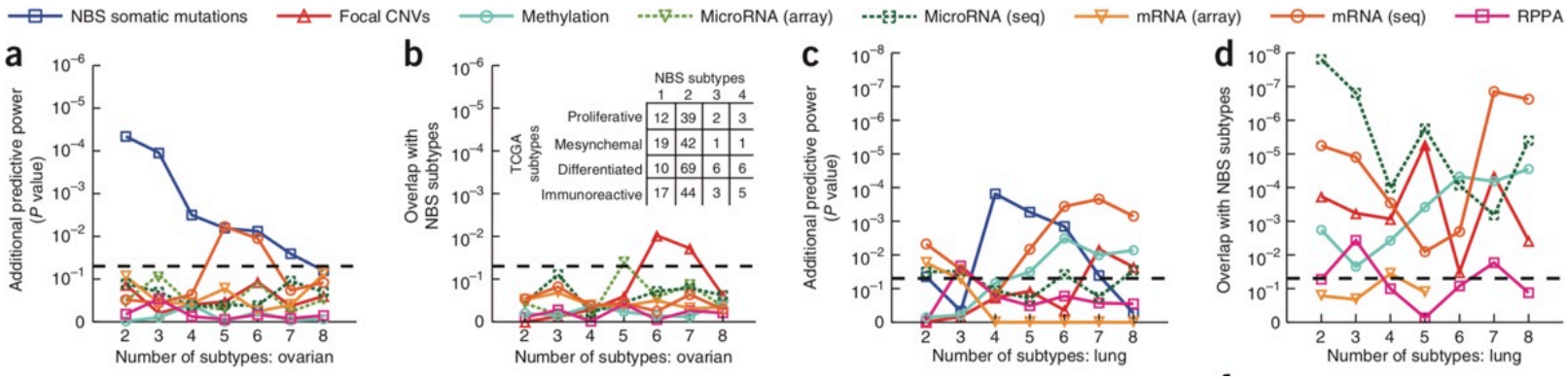


Shi et al., PLoS One, 2012

"high-risk" group "low-risk" group

Case study 2: network-based stratification of tumor mutations





Hofree et al, Nat Methods, 2013

Summary

- Why proteomics
- Proteomics technology
- Protein identification
- Protein quantification
- Protein-protein interaction