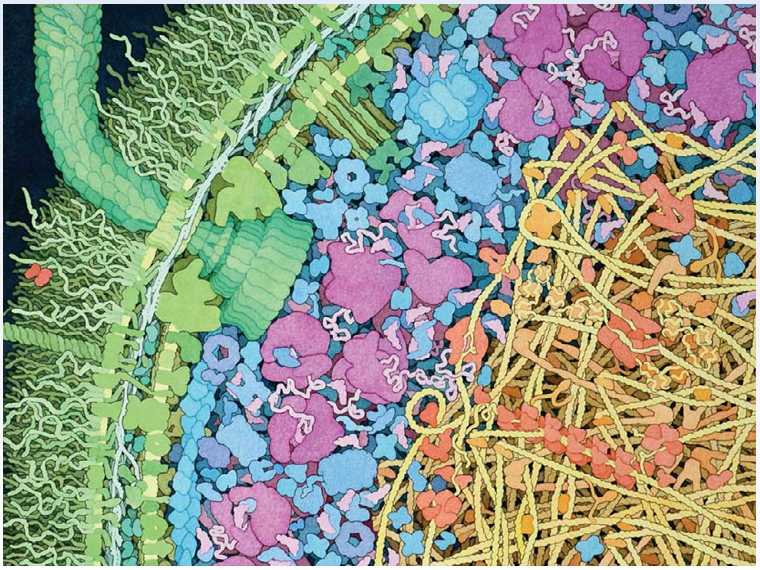
Introduction to Proteomics

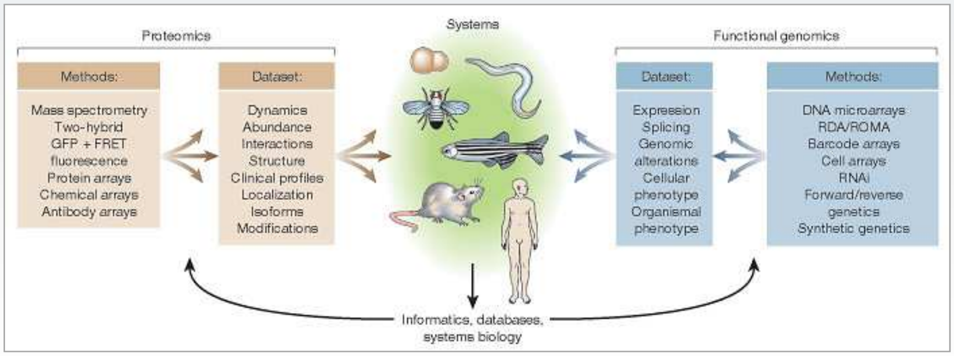


"The Machinery of Life"
Watercolor by David S. Goodsell, 2009

What is the Proteome?

Proteome – All proteins expressed by a genome, cell, tissue, organism, etc.

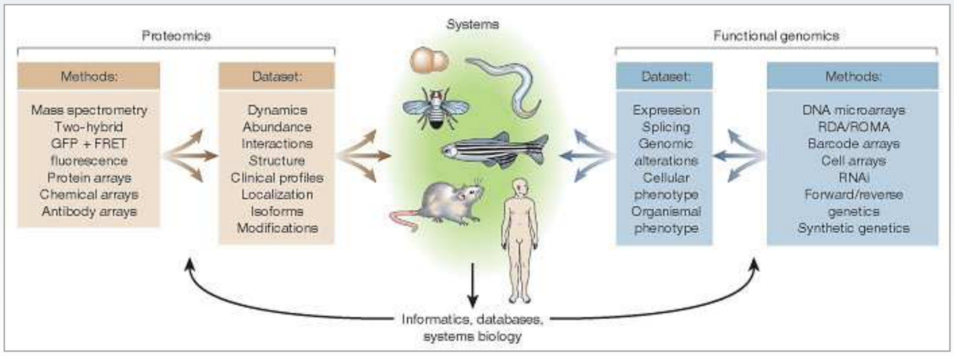
- Complete proteome
- Organ/tissue proteome
- Cellular proteome
- Serum proteome, urine proteome



Tyers and Mann (2003) *Nature*

What is Proteomics?

"Proteomics includes not only the identification and quantification of proteins, but also the determination of their localization, modifications, interactions, activities, and, ultimately, their function." -Stan Fields in *Science, 2001.*



Tyers and Mann (2003) *Nature*

Genome vs. Proteome

The Proteome is complex!

Organism	Genes	Protein-coding	Proteins
Human	43,334	20,848	79,481
Mouse	36,817	22,670	54,567
C. elegans	45,461	20,389	29,872
Drosophila	14,892	13,804	21,999
Yeast	7,130	6,696	6,696
E. coli	4,377	4, 290	
L. Monocytogenes	2,926	2,853	
M. Genitalium	485	485	381 are essential
Human mitochondria	37	13	
φX174 (*E. coli* virus)	11	11	
mimivirus (amoeba virus)	1,018	979	

Genome vs. Proteome

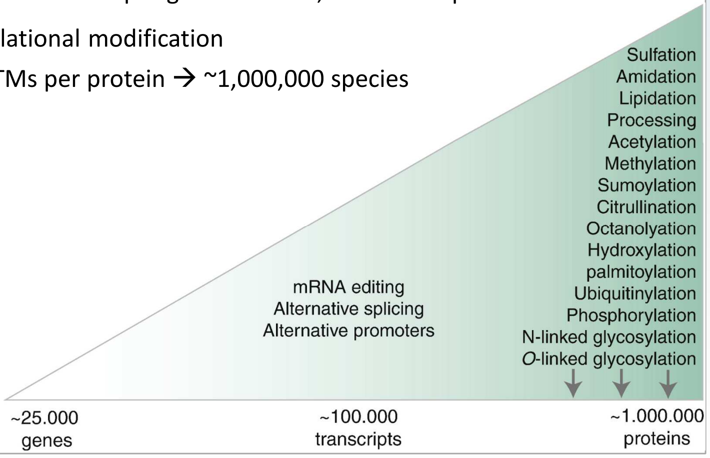
Proteome Complexity

Alternative splicing

5-6 splice variants per gene → ~100,000 transcripts

Post-translational modification

8-10 PTMs per protein → ~1,000,000 species


Sulfation
Amidation
Lipidation
Processing
Acetylation
Methylation
Sumoylation
Citrullination
Octanolyation
Hydroxylation
palmitoylation
Ubiquitinylation
Phosphorylation
N-linked glycosylation
O-linked glycosylation

mRNA editing
Alternative splicing
Alternative promoters

~25.000 genes

~100.000 transcripts

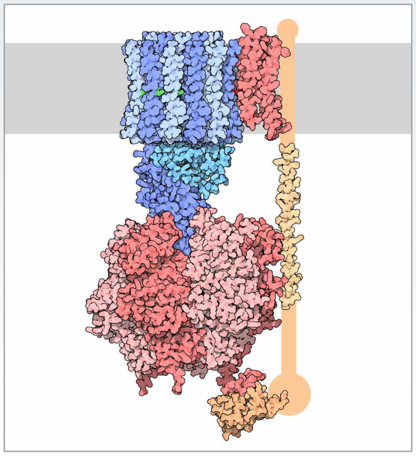
~1.000.000 proteins

Schjoldager, *et al*. (2012) *Biochim Biophys*

Proteins do not act alone!

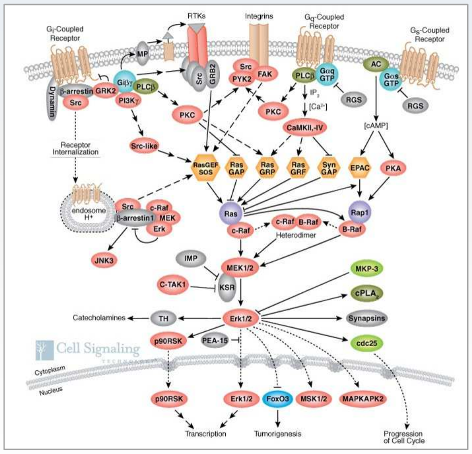
Proteins interact as complexes and function in pathways

ATP Synthase Complex



http://www.rcsb.org/pdb/101/motm.do?momID=72

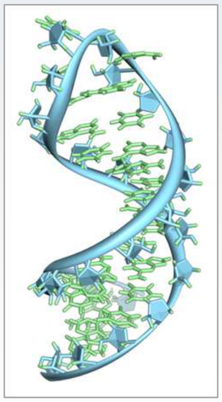
MAPK G-Coupled Protein Receptor Pathway

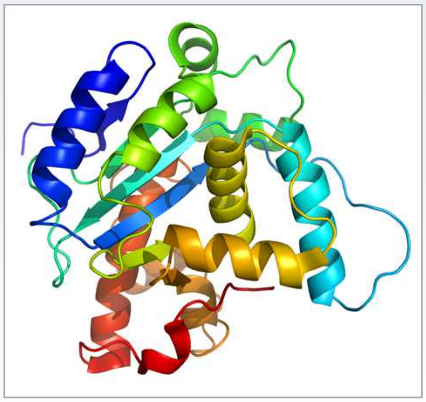


http://dlab.cl/complex-systems/mapk_g_protein

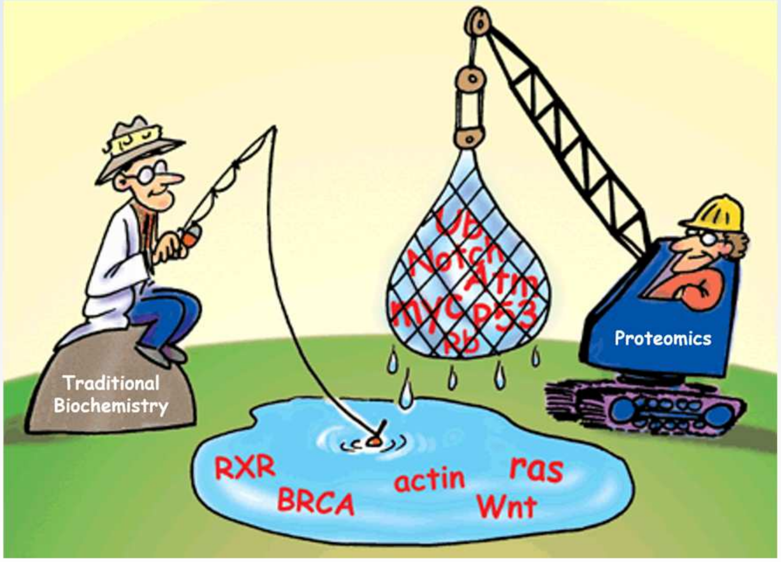
Proteins can be difficult to work with as well

mRNA	**Protein**
Able to amplify (PCR)	No amplification
Readily soluble	High dynamic range (>10^9)
Simple extraction	Variable solubility/extractibility





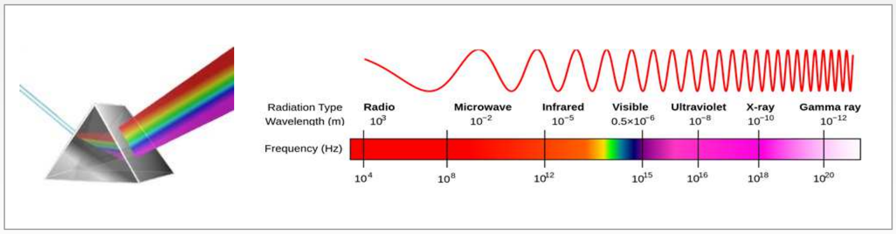
How do we overcome the complexity of the proteome?



JOE SUTLIFF

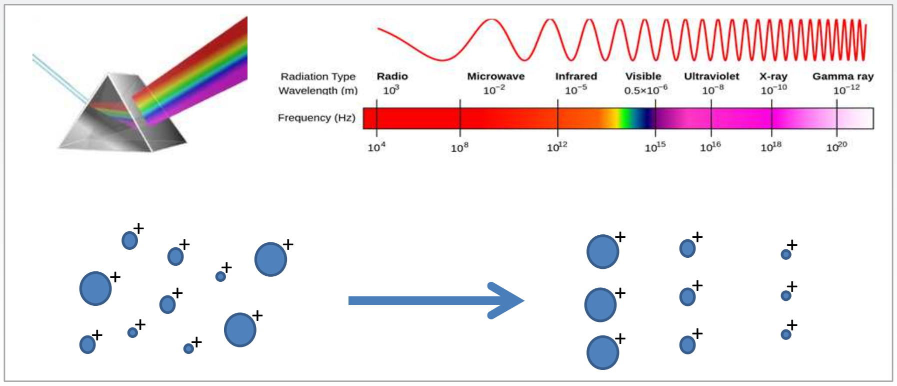
Mass Spectrometry

Ion Source → **Mass Analyzer** → **Detector**



Mass Spectrometry

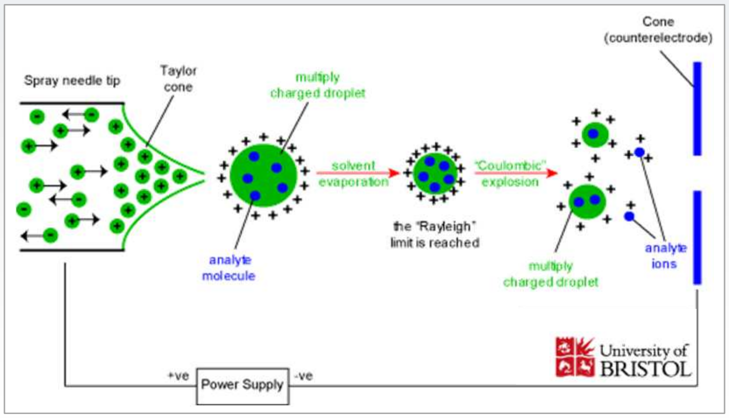
Ion Source → **Mass Analyzer** → **Detector**



Ion Source

Ion Source → Mass Analyzer → Detector

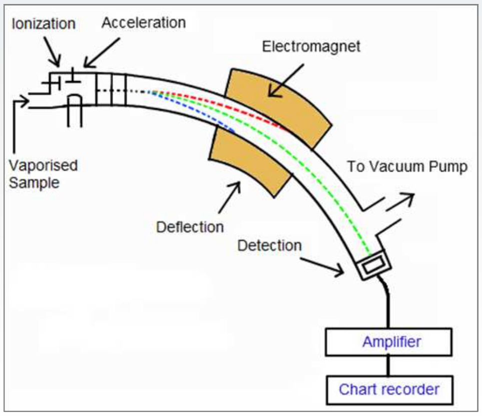
Electrospray Ionization (ESI)



Mass Analyzer

Ion Source → Mass Analyzer → Detector

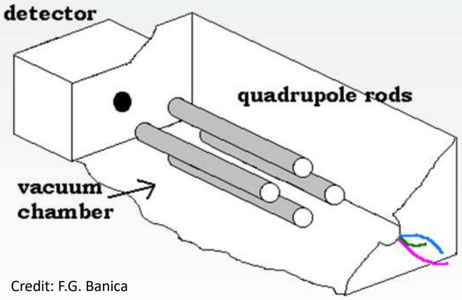
Magnetic Sector MS



Mass Analyzer

Ion Source → **Mass Analyzer** → Detector

Quadrupole



Credit: F.G. Banica

AC and DC electric fields are applied

If AC > DC, low m/z destabilized

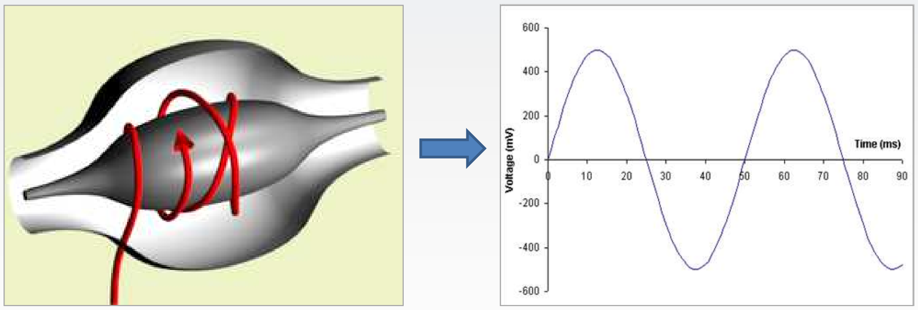
If AC < DC, high m/z destabilized

Only ions of particular m/z will pass

Mass Analyzer

Ion Source → **Mass Analyzer** → Detector

Orbitrap: Fourier Transform MS



A single ion “orbits” the central electrode at a specific frequency based on its mass and charge, and this frequency can be visualized as a waveform.

Mass Analyzer

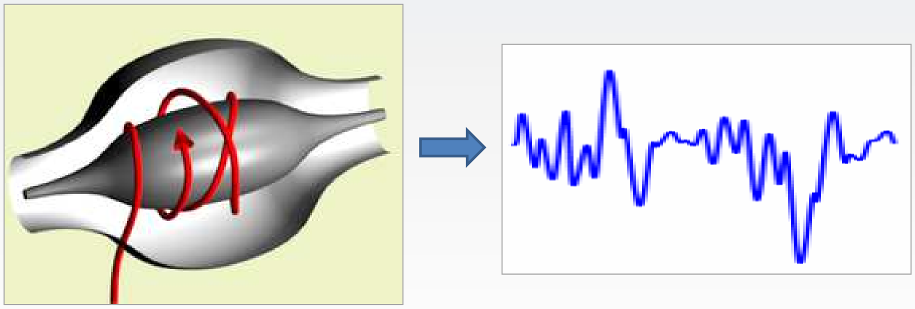
Ion Source → **Mass Analyzer** → Detector

Orbitrap: Fourier Transform MS

Mass Analyzer

Ion Source → **Mass Analyzer** → Detector

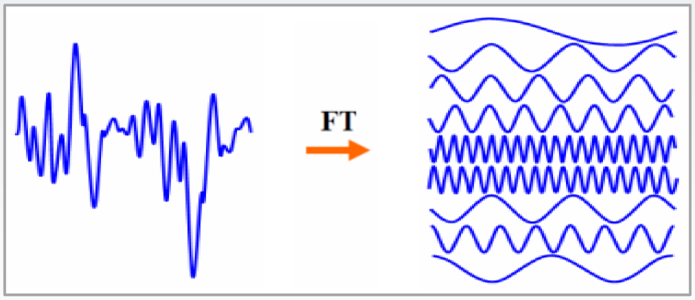
Orbitrap: Fourier Transform MS



When multiple ions of distinct m/z are present, the various waveforms are combined in the readout, leading to a complex waveform.

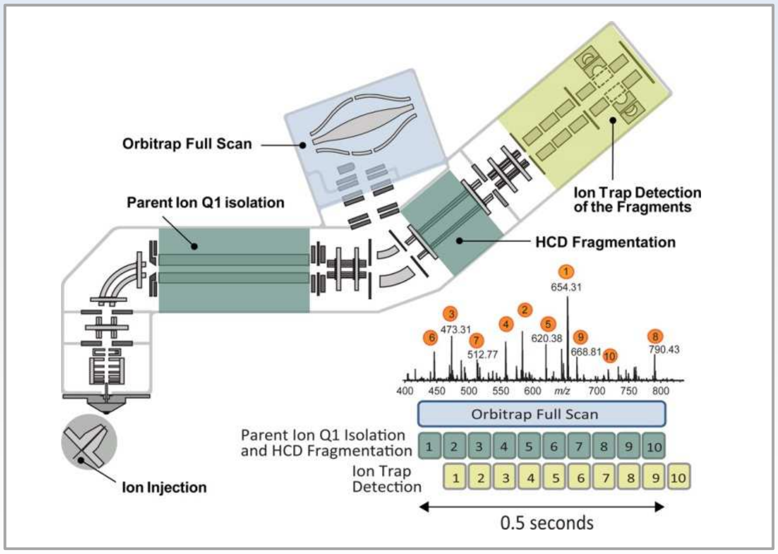
Mass Analyzer

Orbitrap: Fourier Transform MS



Fortunately, the complex waveform can be mathematically deconvoluted into its individual components by Fourier transform, with each individual component corresponding to a specific ion in the scan.

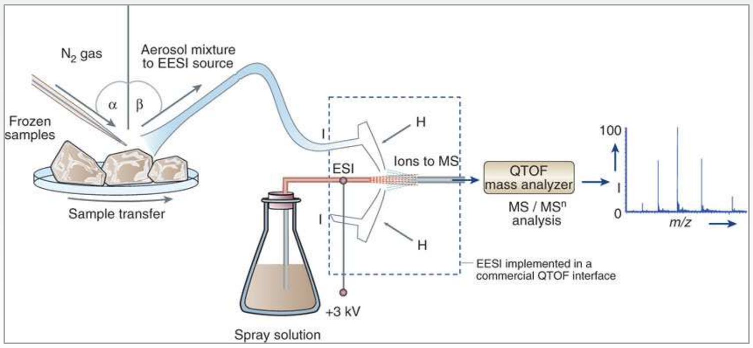
Orbitrap Fusion Tribrid Mass Spectrometer



Ion Source

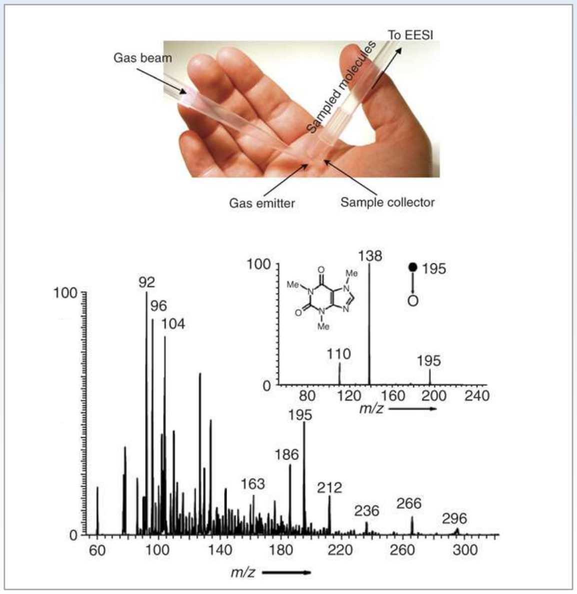
Ion Source → **Mass Analyzer** → **Detector**

Neutral Desorption Extractive Electrospray Ionization (ND-EESI)



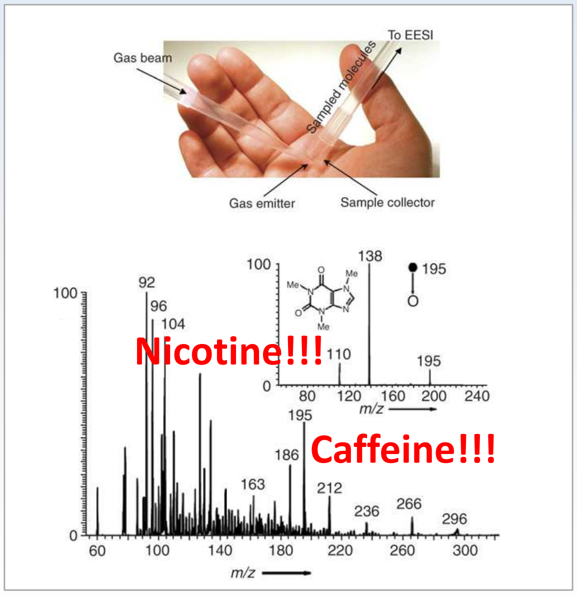
Chen, *et al.* (2008) *Nature Protocols*

ND-EESI



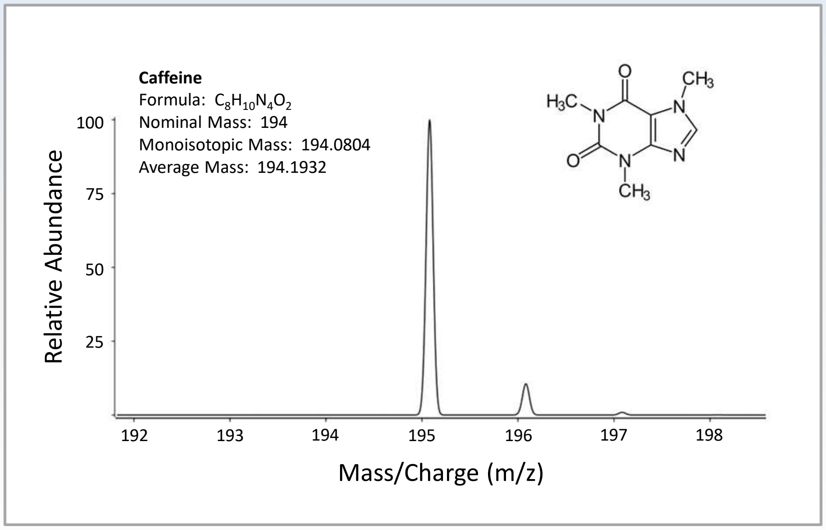
Chen, *et al.* (2008) *Nature Protocols*

ND-EESI



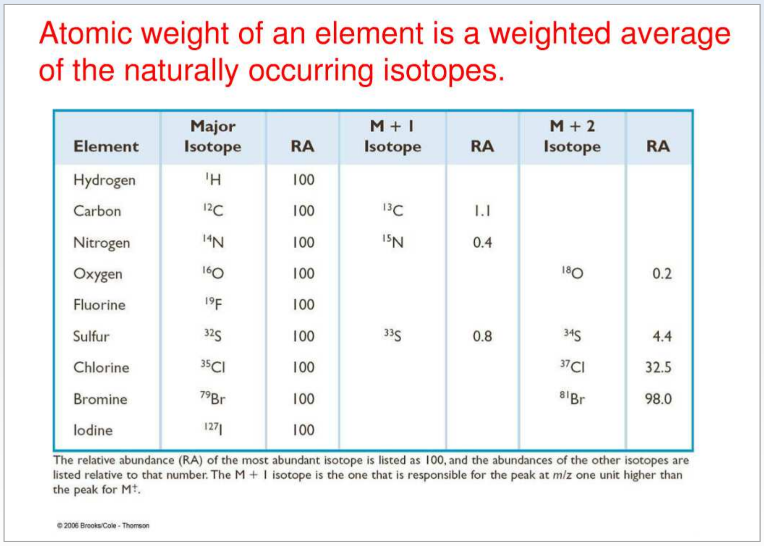
Chen, *et al.* (2008) *Nature Protocols*

What does a mass spectrum tell us?



Caffeine
Formula: $C_8H_{10}N_4O_2$
Nominal Mass: 194
Monoisotopic Mass: 194.0804
Average Mass: 194.1932

Relative Abundance of Isotopes

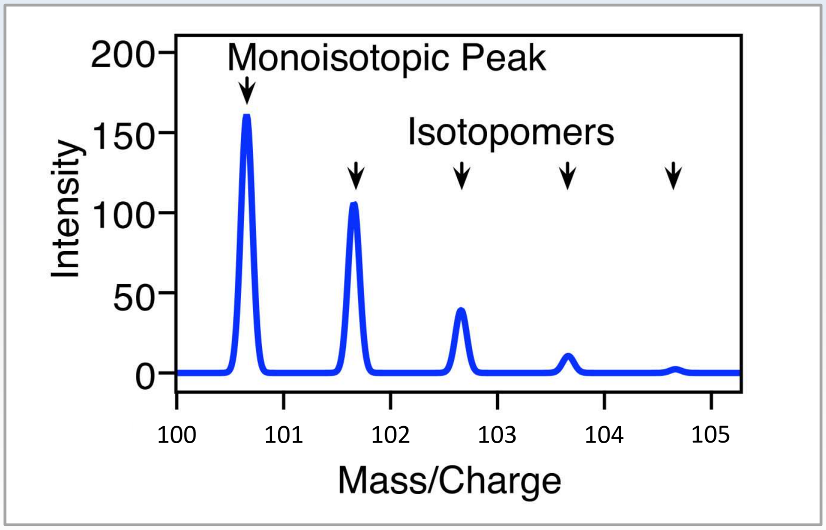


Atomic weight of an element is a weighted average of the naturally occurring isotopes.

Element	Major Isotope	RA	M + 1 Isotope	RA	M + 2 Isotope	RA
Hydrogen	^{1}H	100				
Carbon	^{12}C	100	^{13}C	1.1		
Nitrogen	^{14}N	100	^{15}N	0.4		
Oxygen	^{16}O	100			^{18}O	0.2
Fluorine	^{19}F	100				
Sulfur	^{32}S	100	^{33}S	0.8	^{34}S	4.4
Chlorine	^{35}Cl	100			^{37}Cl	32.5
Bromine	^{79}Br	100			^{81}Br	98.0
Iodine	^{127}I	100				

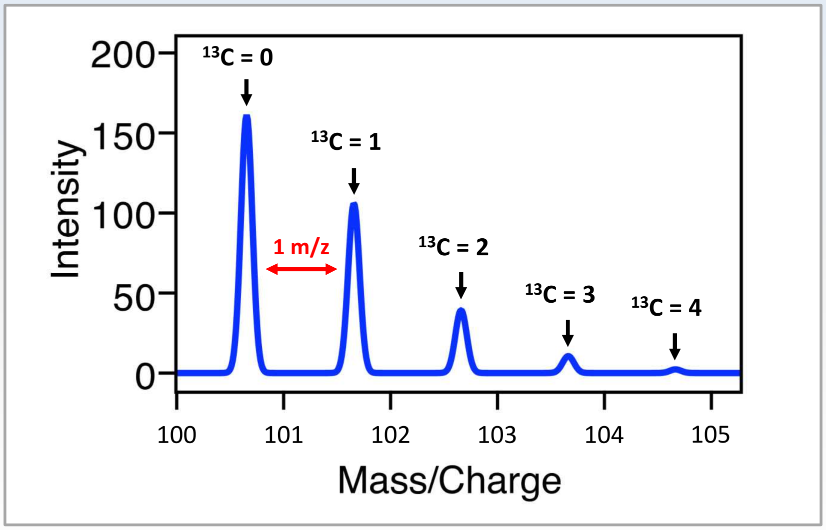
The relative abundance (RA) of the most abundant isotope is listed as 100, and the abundances of the other isotopes are listed relative to that number. The M + 1 isotope is the one that is responsible for the peak at *m/z* one unit higher than the peak for $M^{+\cdot}$.

What does a mass spectrum tell us?



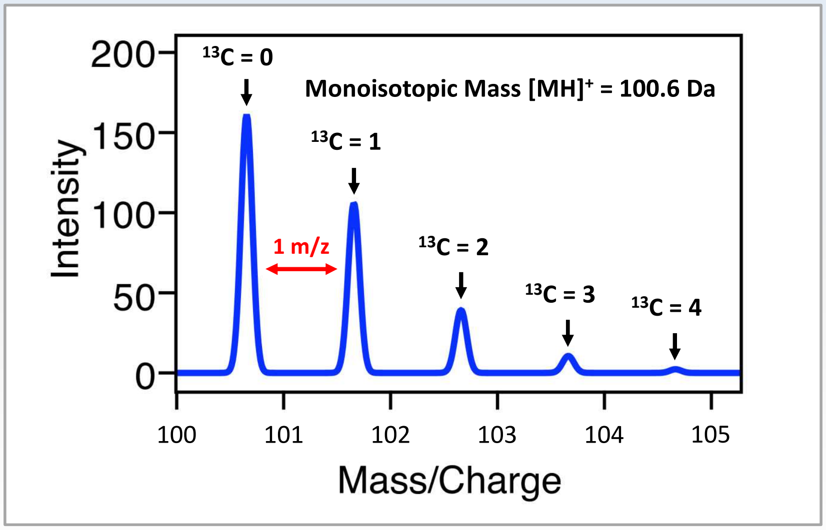
Sykes and Williamson, BMC Bioinformatics, 2008

What does a mass spectrum tell us?



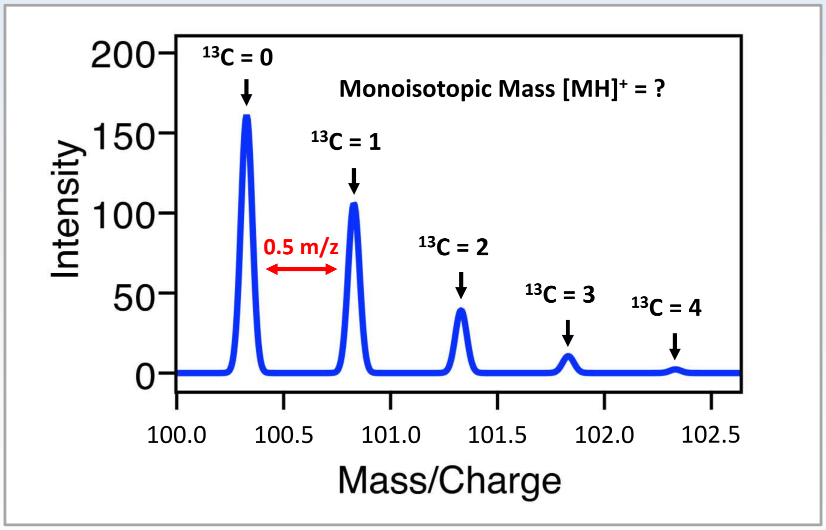
Sykes and Williamson, BMC Bioinformatics, 2008

What does a mass spectrum tell us?



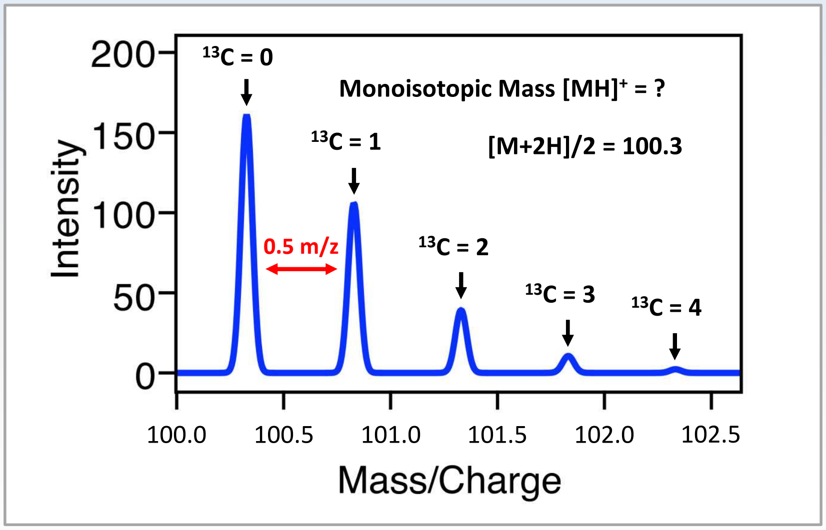
Sykes and Williamson, BMC Bioinformatics, 2008

What does a mass spectrum tell us?



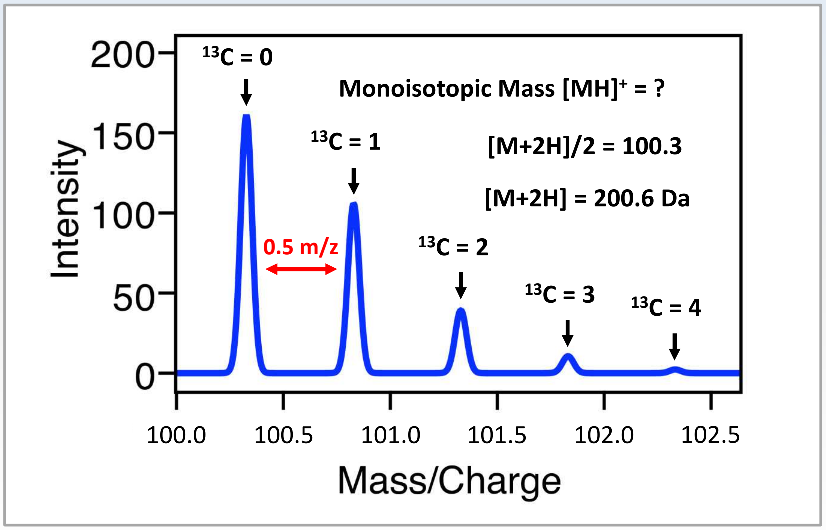
Sykes and Williamson, BMC Bioinformatics, 2008

What does a mass spectrum tell us?



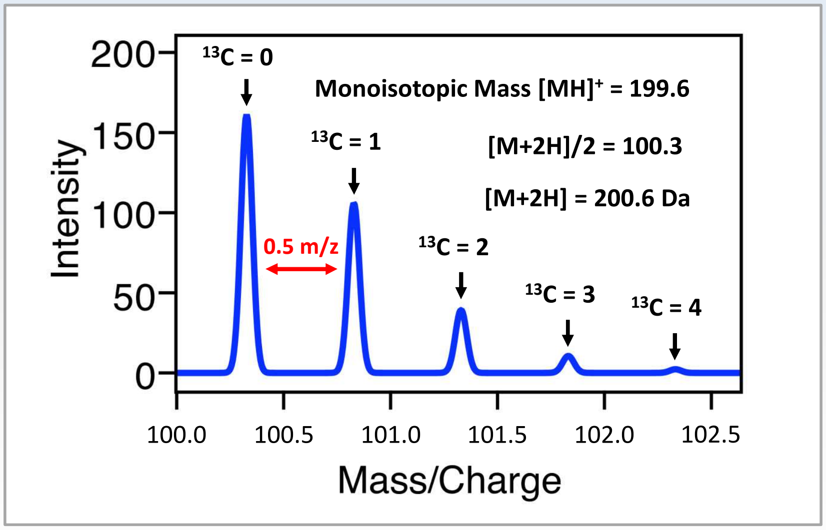
Sykes and Williamson, BMC Bioinformatics, 2008

What does a mass spectrum tell us?



Sykes and Williamson, BMC Bioinformatics, 2008

What does a mass spectrum tell us?

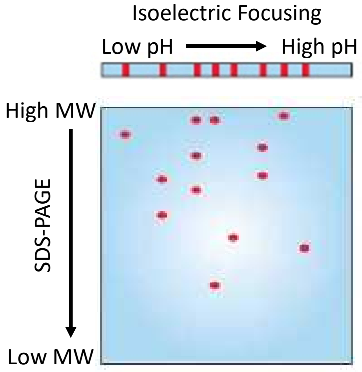


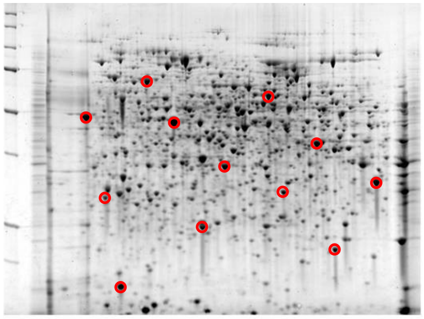
Sykes and Williamson, BMC Bioinformatics, 2008

Proteomic analysis requires novel methods

Classic approach: 2D gel electrophoresis and mass spectrometry

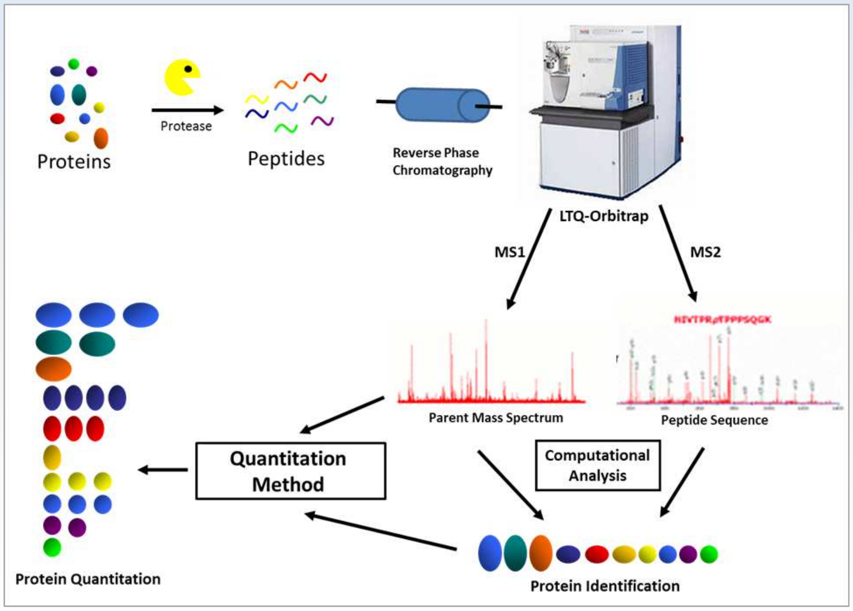
Protein separations reduce sample complexity, increase sensitivity





Individual bands excised for MS analysis

A Typical Shotgun Proteomics Experiment



Proteins

Protease

Peptides

Reverse Phase Chromatography

LTQ-Orbitrap

MS1

MS2

HIVTPR$_p$TPPPSQGK

Parent Mass Spectrum

Peptide Sequence

Quantitation Method

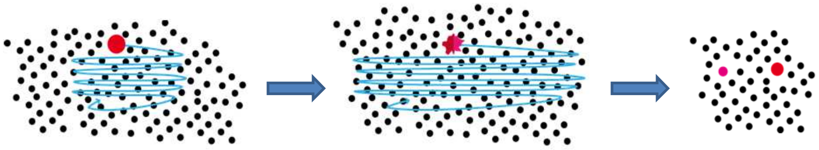
Computational Analysis

Protein Identification

Protein Quantitation

Collision Activated Dissociation (CAD)

Collision Induced Dissociation (CID)

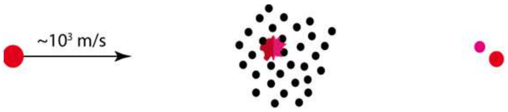


Trapping

Activation
10-1000s of collisions

Fragments

Higher Energy Collision Dissociation (HCD)



acceleration

collide with target gas
~10 - 100 collisions

fragments

33

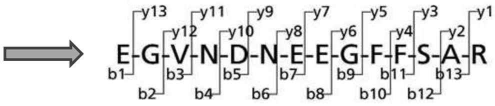
Precursor Ion Fragmentation

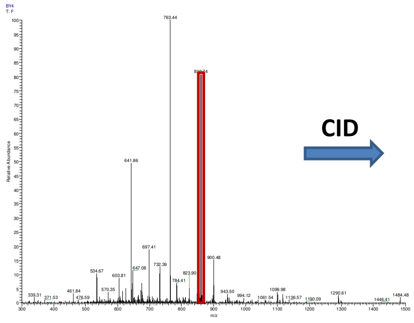
A precursor ion is selected from the full spectrum (MS1) and isolated

The ion is energetically fragmented (CID, HCD, ETD, etc.)

The fragmentation spectrum (MS2) reveals a ladder of sequential amino acids

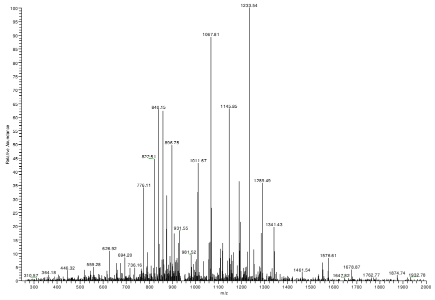
E-G-V-N-D-N-E-E-G-F-F-S-A-R





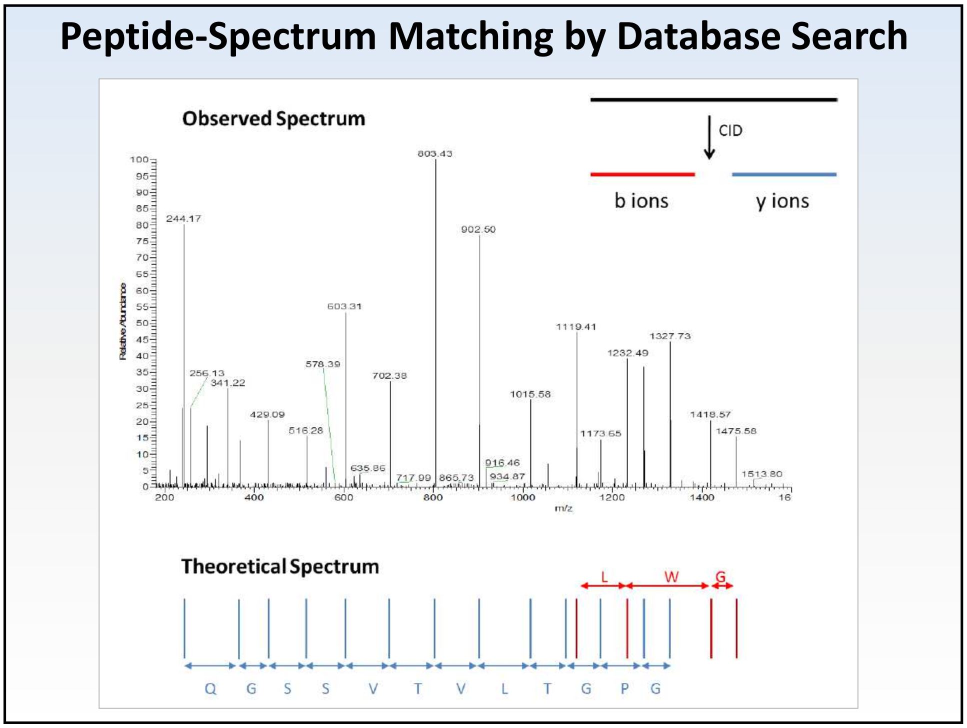
MS1

CID

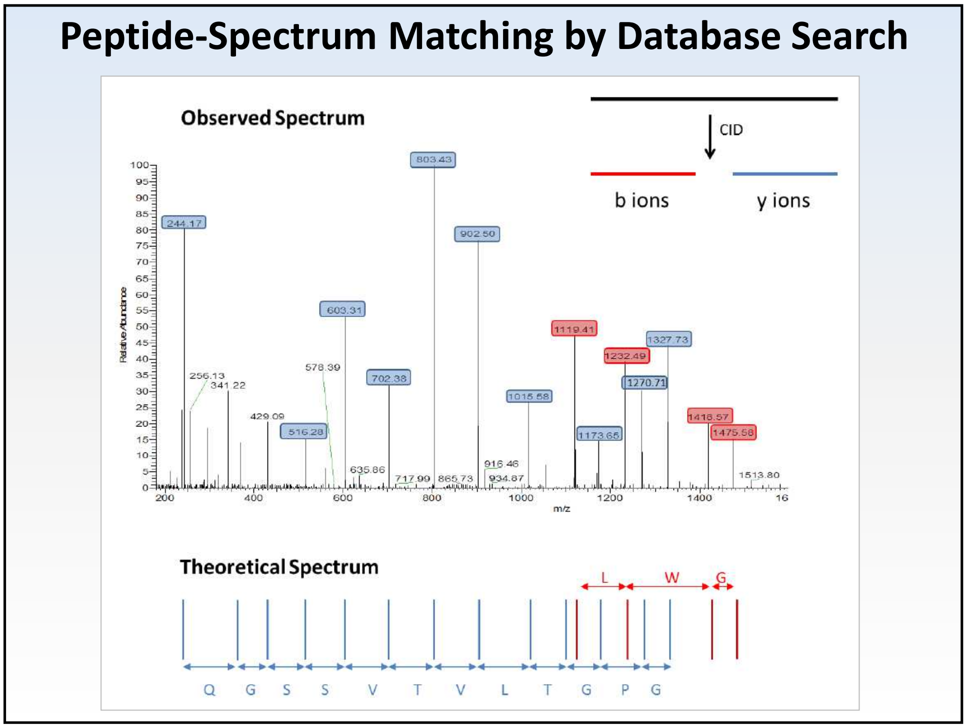


MS2

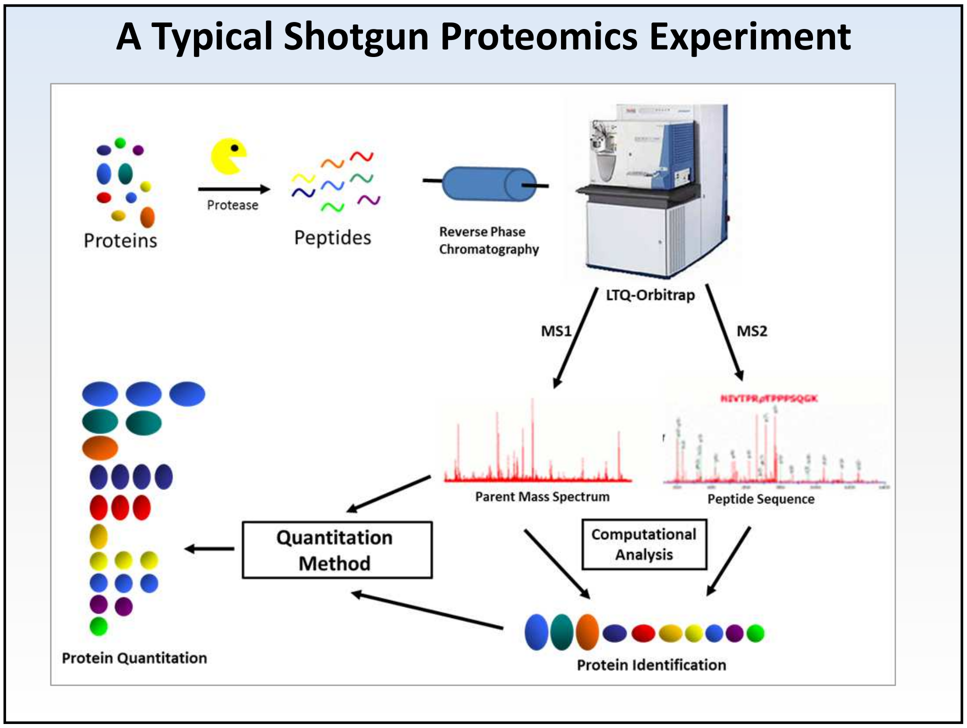
34



Peptide-Spectrum Matching by Database Search

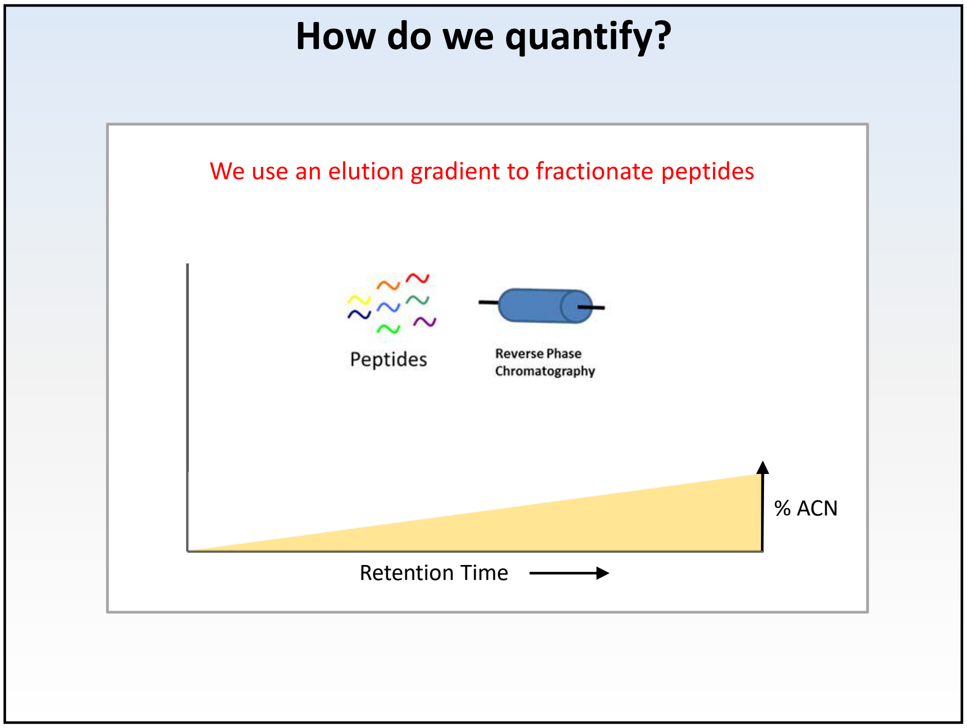


Peptide-Spectrum Matching by Database Search



A Typical Shotgun Proteomics Experiment

37



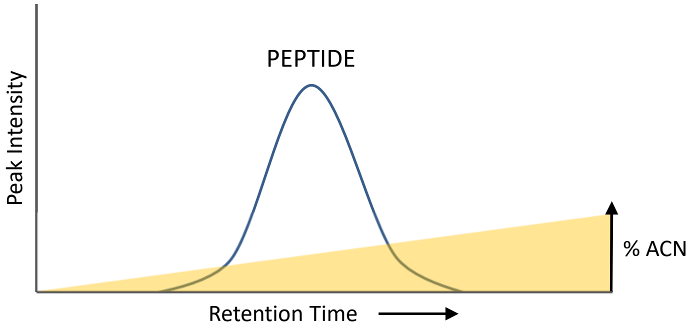
How do we quantify?

We use an elution gradient to fractionate peptides

38

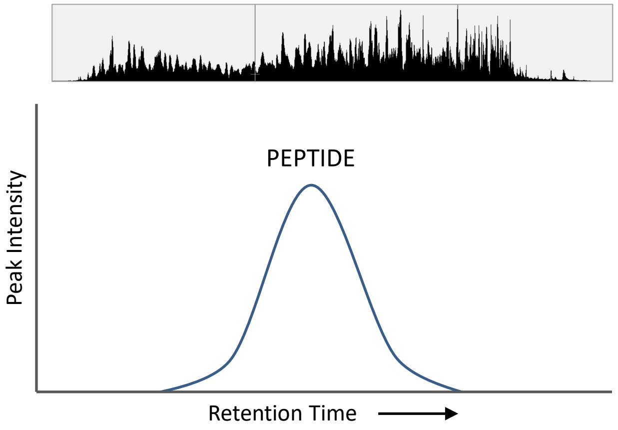
How do we quantify?

Peptides elute as individual peaks at specific retention times

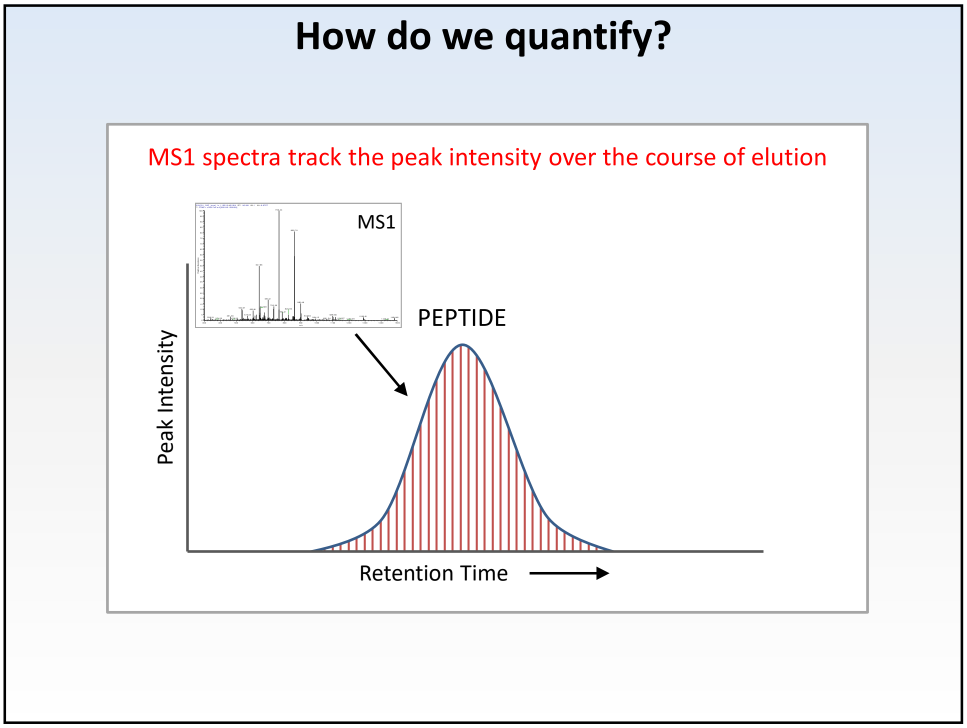


39

How do we quantify?

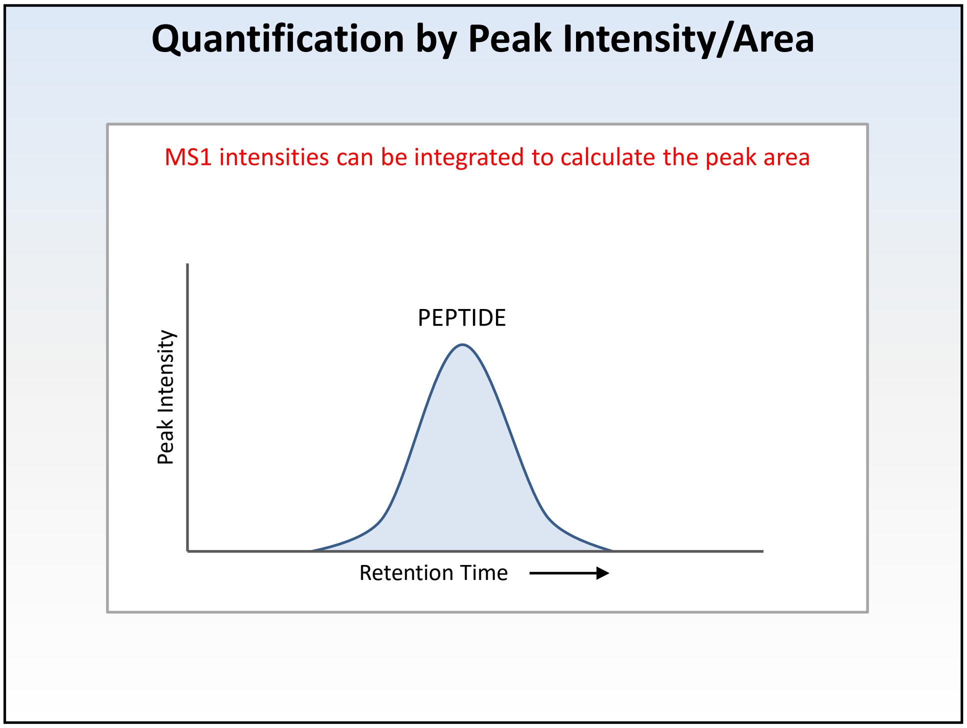


40



How do we quantify?

MS1 spectra track the peak intensity over the course of elution

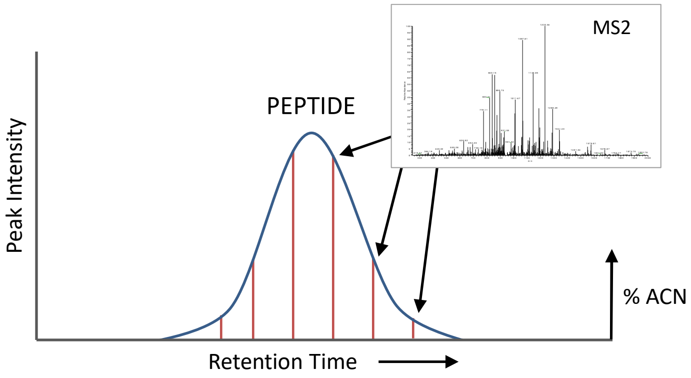


Quantification by Peak Intensity/Area

MS1 intensities can be integrated to calculate the peak area

Quantification by Spectral Counting

MS2 spectra identify the peptide-spectral matches (PSMs)
The sum of PSMs for a protein is the spectral count

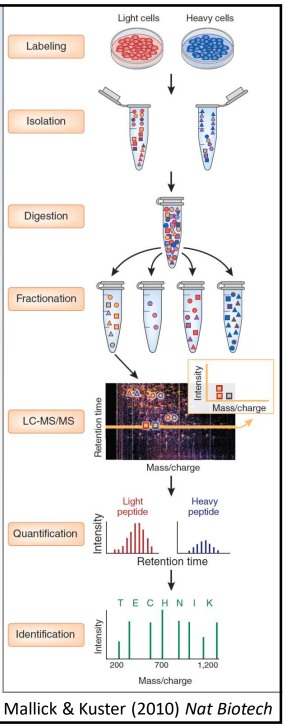


One common strategy for <u>relative</u> quantification using isotopically labeled samples (e.g. ^{15}N vs. ^{14}N, ^{13}C vs. ^{12}C, etc.)

SILAC = stable isotope labeling with amino acids in cell culture

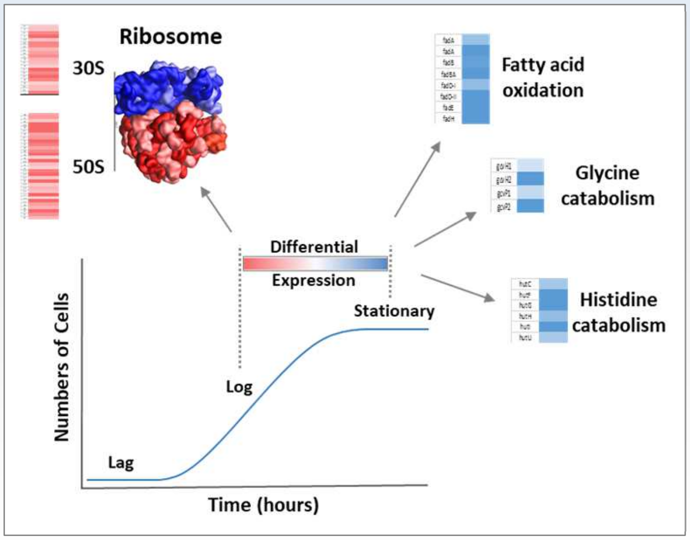
What's the difference between <u>relative</u> and <u>absolute</u> quantification?

How might we measure absolute quantities of proteins?



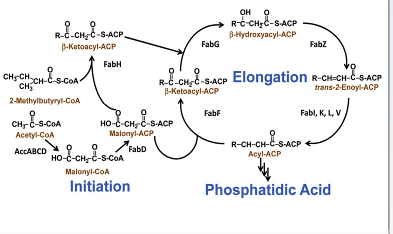
Mallick & Kuster (2010) *Nat Biotech*

Differential Protein Expression Analysis



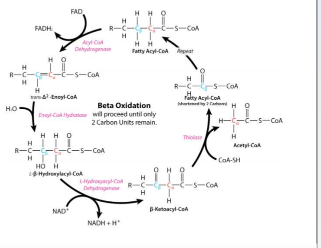
Differential Protein Expression Analysis

Fatty Acid Biosynthesis

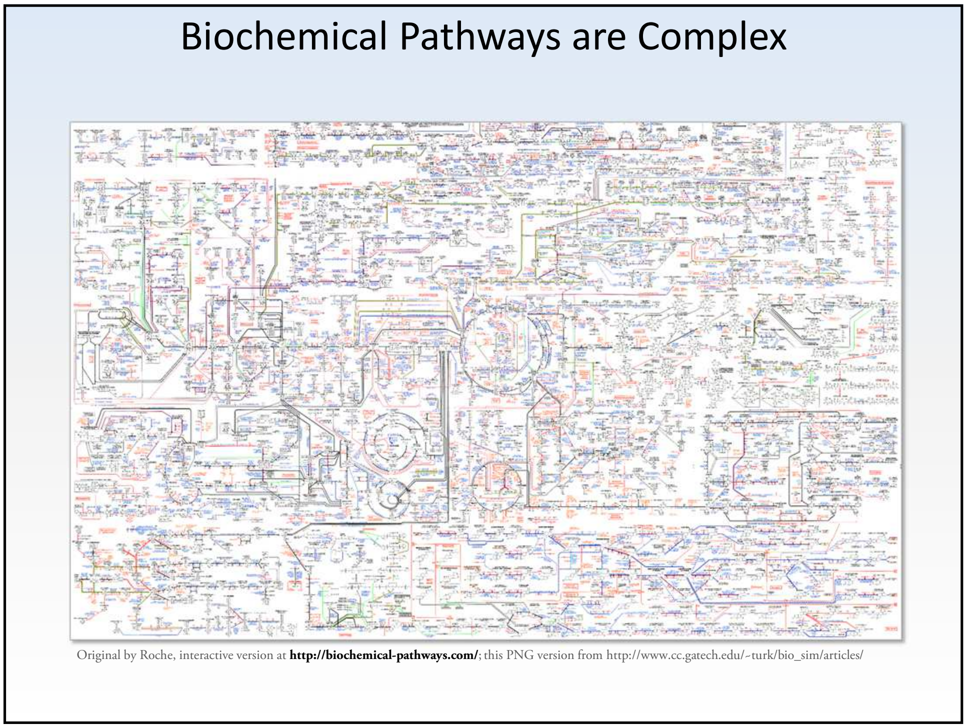


Protein	MidLog	Stat	Z-score	Ratio (PSMs)	Ratio (XIC)
accA	319	265	5.26	1.54	1.92
accB	291	200	6.77	1.85	3.25
accC	672	551	7.90	1.56	1.28
accD	156	141	3.06	1.42	1.93
PP_4379	218	117	7.89	2.39	2.58
fabA	175	55	9.90	3.98	2.29
fabB	601	428	9.51	1.80	1.37
fabD	267	209	5.45	1.64	1.17
fabF	198	97	8.06	2.58	3.48
fabG	209	167	4.65	1.62	1.48
fabZ	169	84	7.44	2.67	5.22

Fatty Acid Beta Oxidation

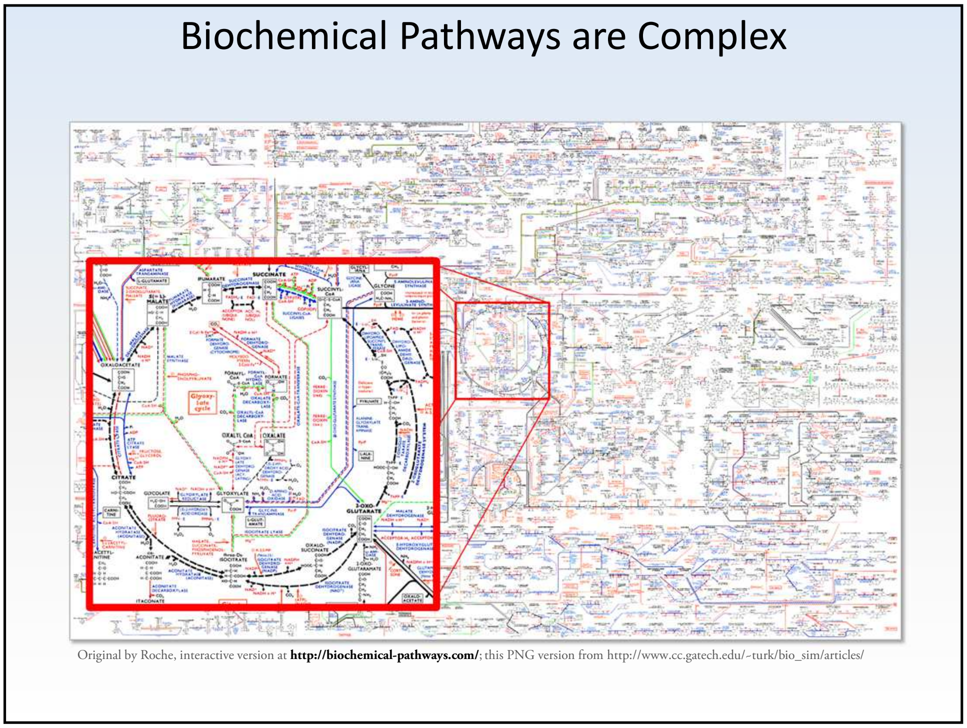


Protein	MidLog	Stat	Z-score	Ratio (PSMs)	Ratio (XIC)
fadA	108	325	-7.81	-2.31	-5.11
paaJ	0	288	-14.95	-113.80	-64852.07
bktB	117	760	-18.20	-6.11	-32.96
yqeF	45	518	-17.09	-8.86	-42.71
fadB	202	515	-8.32	-1.97	-8.79
PP_1845	77	170	-3.94	-1.69	-3.67
PP_2217	66	287	-9.41	-3.75	-13.33
paaF	0	69	-7.21	-27.40	-3970.44
PP_4030	0	24	-4.31	-10.12	-2978.20
fadE	8	708	-22.99	-91.68	-2672.30
acd	22	185	-9.60	-9.84	-114.34
PP_2437	0	27	-4.53	-11.21	-1054.34
PP_2793	0	10	-2.66	-4.78	-737.20



Biochemical Pathways are Complex

Original by Roche, interactive version at **http://biochemical-pathways.com/**; this PNG version from http://www.cc.gatech.edu/~turk/bio_sim/articles/



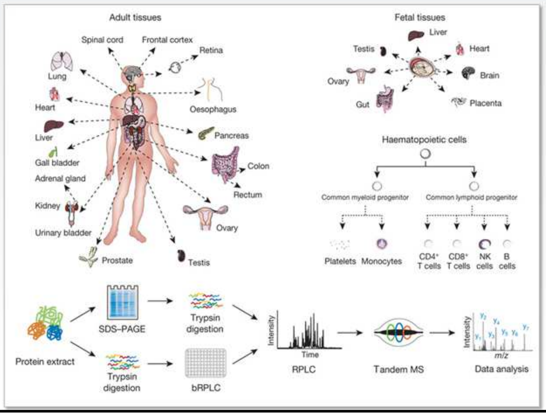
Biochemical Pathways are Complex

Original by Roche, interactive version at **http://biochemical-pathways.com/**; this PNG version from http://www.cc.gatech.edu/~turk/bio_sim/articles/

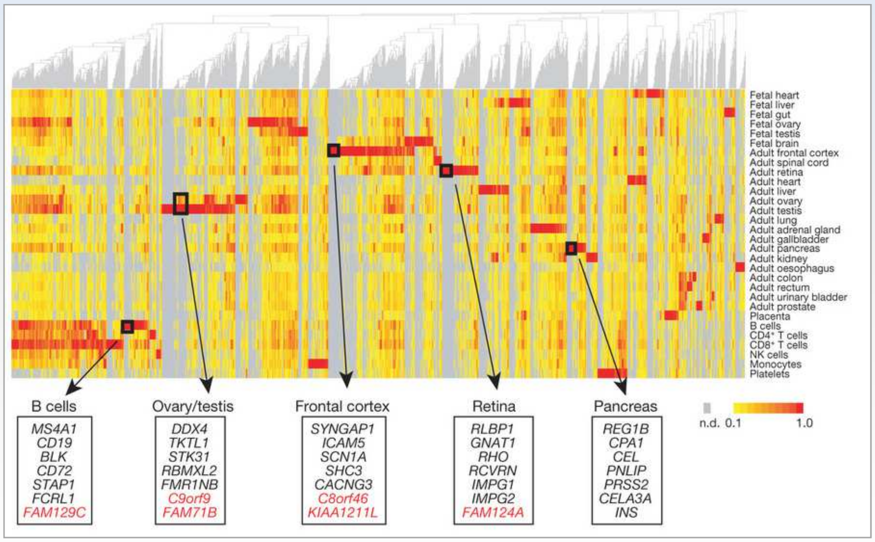
How Complete is the Annotated Proteome?

A draft map of the human proteome

Min-Sik Kim[1,2], Sneha M. Pinto[3], Derese Getnet[1,4], Raja Sekhar Nirujogi[3], Srikanth S. Manda[3], Raghothama Chaerkady[1,2], Anil K. Madugundu[3], Dhanashree S. Kelkar[3], Ruth Isserlin[5], Shobhit Jain[5], Joji K. Thomas[3], Babylakshmi Muthusamy[3], Pamela Leal-Rojas[1,6], Praveen Kumar[3], Nandini A. Sahasrabuddhe[3], Lavanya Balakrishnan[3], Jayshree Advani[3], Bijesh George[3], Santosh Renuse[3], Lakshmi Dhevi N. Selvan[3], Arun H. Patil[3], Vishalakshi Nanjappa[3], Aneesha Radhakrishnan[3], Samarjeet Prasad[1], Tejaswini Subbannayya[3], Rajesh Raju[3], Manish Kumar[3], Sreelakshmi K. Sreenivasamurthy[3], Arivusudar Marimuthu[3], Gajanan J. Sathe[3], Sandip Chavan[3], Keshava K. Datta[3], Yashwanth Subbannayya[3], Apeksha Sahu[3], Soujanya D. Yelamanchi[3], Savita Jayaram[3], Pavithra Rajagopalan[3], Jyoti Sharma[3], Krishna R. Murthy[3], Nazia Syed[3], Renu Goel[3], Aafaque A. Khan[3], Sartaj Ahmad[3], Gourav Dey[3], Keshav Mudgal[7], Aditi Chatterjee[3], Tai-Chung Huang[1], Jun Zhong[1], Xinyan Wu[1,2], Patrick G. Shaw[1], Donald Freed[1], Muhammad S. Zahari[2], Kanchan K. Mukherjee[8], Subramanian Shankar[9], Anita Mahadevan[10,11], Henry Lam[12], Christopher J. Mitchell[1], Susarla Krishna Shankar[10,11], Parthasarathy Satishchandra[13], John T. Schroeder[14], Ravi Sirdeshmukh[3], Anirban Maitra[15,16], Steven D. Leach[1,17], Charles G. Drake[16,18], Marc K. Halushka[15], T. S. Keshava Prasad[3], Ralph H. Hruban[15,16], Candace L. Kerr[19]†, Gary D. Bader[5], Christine A. Iacobuzio-Donahue[15,16,17], Harsha Gowda[3] & Akhilesh Pandey[1,2,3,4,15,16,20]



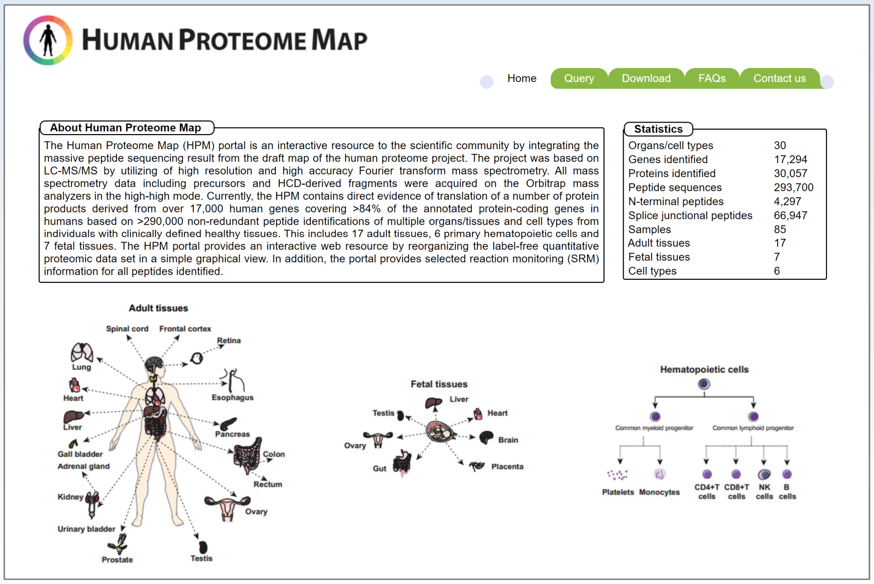
The (Annotated) Human Proteome



B cells	Ovary/testis	Frontal cortex	Retina	Pancreas
MS4A1	DDX4	SYNGAP1	RLBP1	REG1B
CD19	TKTL1	ICAM5	GNAT1	CPA1
BLK	STK31	SCN1A	RHO	CEL
CD72	RBMXL2	SHC3	RCVRN	PNLIP
STAP1	FMR1NB	CACNG3	IMPG1	PRSS2
FCRL1	C9orf9	C8orf46	IMPG2	CELA3A
FAM129C	FAM71B	KIAA1211L	FAM124A	INS

Kim, *et al.* (2014) *Nature*

The (Annotated) Human Proteome



About Human Proteome Map

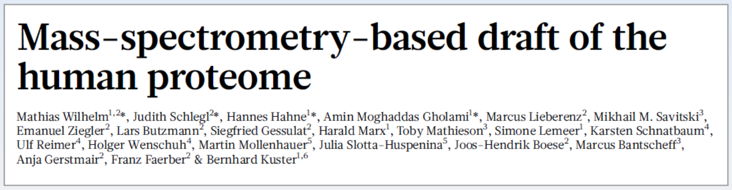
The Human Proteome Map (HPM) portal is an interactive resource to the scientific community by integrating the massive peptide sequencing result from the draft map of the human proteome project. The project was based on LC-MS/MS by utilizing of high resolution and high accuracy Fourier transform mass spectrometry. All mass spectrometry data including precursors and HCD-derived fragments were acquired on the Orbitrap mass analyzers in the high-high mode. Currently, the HPM contains direct evidence of translation of a number of protein products derived from over 17,000 human genes covering >84% of the annotated protein-coding genes in humans based on >290,000 non-redundant peptide identifications of multiple organs/tissues and cell types from individuals with clinically defined healthy tissues. This includes 17 adult tissues, 6 primary hematopoietic cells and 7 fetal tissues. The HPM portal provides an interactive web resource by reorganizing the label-free quantitative proteomic data set in a simple graphical view. In addition, the portal provides selected reaction monitoring (SRM) information for all peptides identified.

Statistics

Organs/cell types	30
Genes identified	17,294
Proteins identified	30,057
Peptide sequences	293,700
N-terminal peptides	4,297
Splice junctional peptides	66,947
Samples	85
Adult tissues	17
Fetal tissues	7
Cell types	6

www.humanproteomemap.org/

The (Annotated) Human Proteome

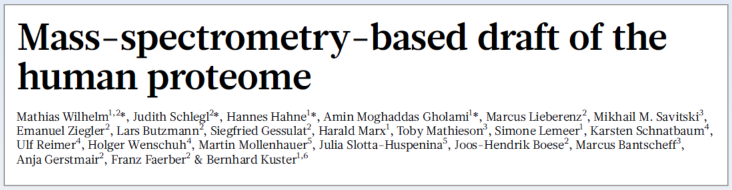


Mass-spectrometry-based draft of the human proteome

Mathias Wilhelm[1,2]*, Judith Schlegl[2]*, Hannes Hahne[1]*, Amin Moghaddas Gholami[1]*, Marcus Lieberenz[2], Mikhail M. Savitski[3], Emanuel Ziegler[2], Lars Butzmann[2], Siegfried Gessulat[2], Harald Marx[1], Toby Mathieson[3], Simone Lemeer[1], Karsten Schnatbaum[4], Ulf Reimer[4], Holger Wenschuh[4], Martin Mollenhauer[5], Julia Slotta-Huspenina[5], Joos-Hendrik Boese[2], Marcus Bantscheff[3], Anja Gerstmair[2], Franz Faerber[2] & Bernhard Kuster[1,6]

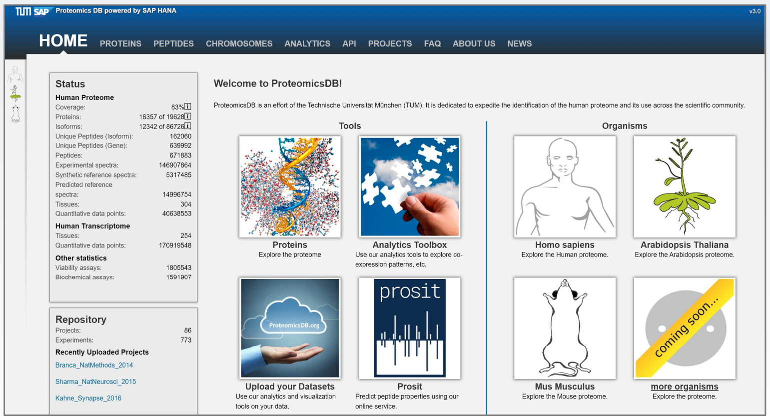
	Pandey	Kuster
Genes identified (19,629)	17,294	18,290
Proteins identified (86,771)	30,057	20,855
Unique peptide sequences	293,700	781,976
Isoform-specific peptides	66,947	220,410
N-terminal peptides	4,297	7,977
Number of samples	85	387
Adult tissues	17	17
Fetal tissues	7	7
Cell types	6	8

The (Annotated) Human Proteome



Mass-spectrometry-based draft of the human proteome

Mathias Wilhelm[1,2]*, Judith Schlegl[2]*, Hannes Hahne[1]*, Amin Moghaddas Gholami[1]*, Marcus Lieberenz[2], Mikhail M. Savitski[3], Emanuel Ziegler[2], Lars Butzmann[2], Siegfried Gessulat[2], Harald Marx[1], Toby Mathieson[3], Simone Lemeer[1], Karsten Schnatbaum[4], Ulf Reimer[4], Holger Wenschuh[4], Martin Mollenhauer[5], Julia Slotta-Huspenina[5], Joos-Hendrik Boese[2], Marcus Bantscheff[3], Anja Gerstmair[2], Franz Faerber[2] & Bernhard Kuster[1,6]



Status

Human Proteome

Coverage:	83%
Proteins:	16357 of 19628
Isoforms:	12342 of 86726
Unique Peptides (Isoform):	162060
Unique Peptides (Gene):	639992
Peptides:	671883
Experimental spectra:	146907864
Synthetic reference spectra:	5317485
Predicted reference spectra:	14996754
Tissues:	304
Quantitative data points:	40638553

Human Transcriptome

Tissues:	254
Quantitative data points:	170919548

Other statistics

Viability assays:	1805543
Biochemical assays:	1591907

Repository

Projects:	86
Experiments:	773

Recently Uploaded Projects

Branca_NatMethods_2014

Sharma_NatNeurosci_2015

Kahne_Synapse_2016

Welcome to ProteomicsDB!

ProteomicsDB is an effort of the Technische Universität München (TUM). It is dedicated to expedite the identification of the human proteome and its use across the scientific community.

Tools

Proteins
Explore the proteome

Analytics Toolbox
Use our analytics tools to explore co-expression patterns, etc.

Upload your Datasets
Use our analytics and visualization tools on your data.

Prosit
Predict peptide properties using our online service.

Organisms

Homo sapiens
Explore the Human proteome.

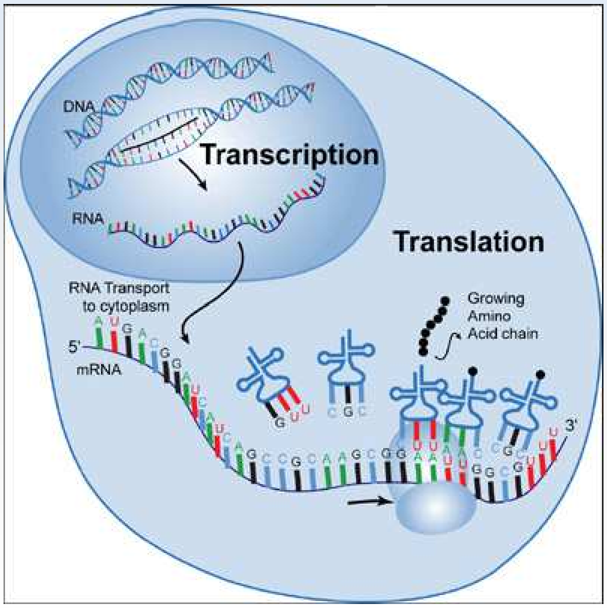
Arabidopsis Thaliana
Explore the Arabidopsis proteome.

Mus Musculus
Explore the Mouse proteome.

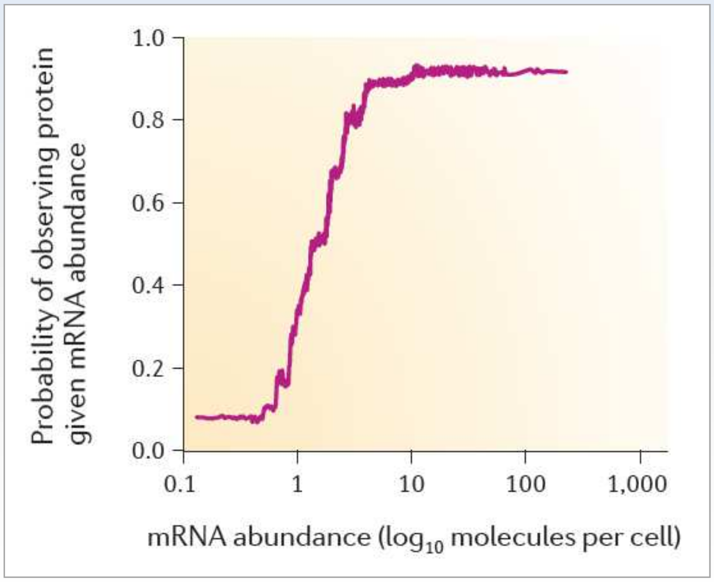
more organisms
Explore the proteome.

www.proteomicsdb.org/

Transcriptome vs. Proteome



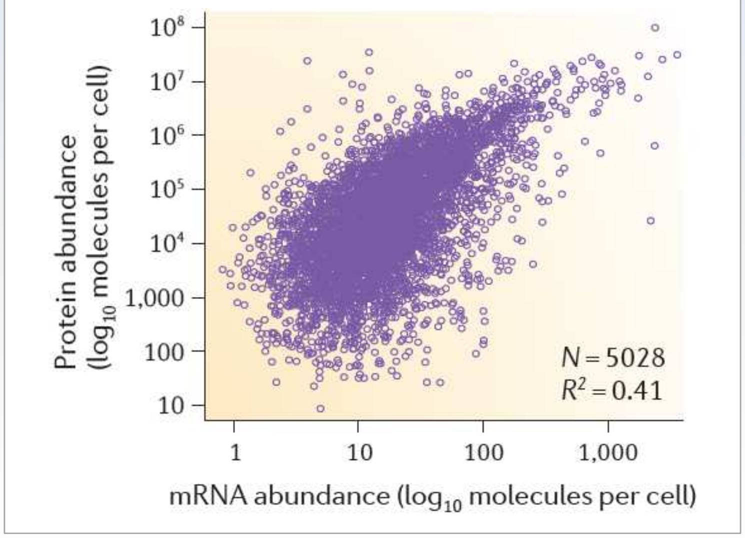
RNA levels are an excellent proxy (in general) for protein presence



Vogel, *et al*. (2010) *Mol Sys Biol*

Yet transcript abundances only partially predict protein abundances

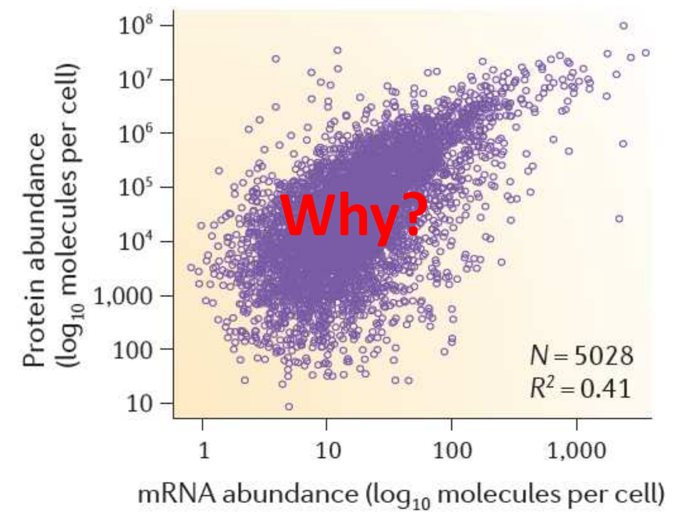
e.g. in a mouse cell line:



Vogel, *et al*. (2010) *Mol Sys Biol*

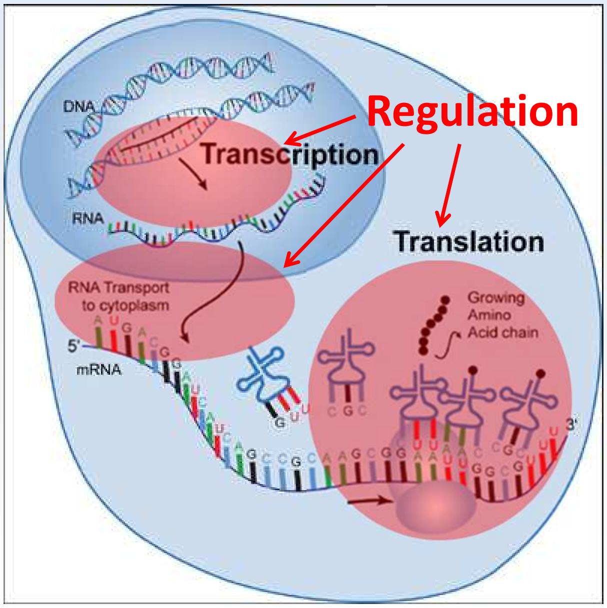
Yet transcript abundances only partially predict protein abundances

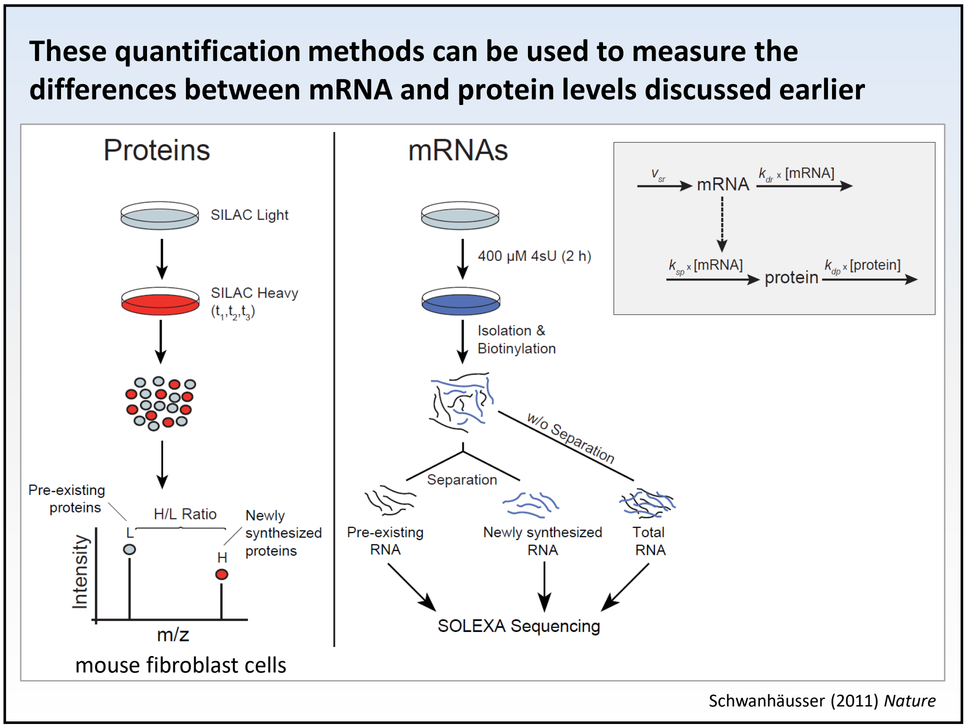
e.g. in a mouse cell line:



Vogel, *et al*. (2010) *Mol Sys Biol*

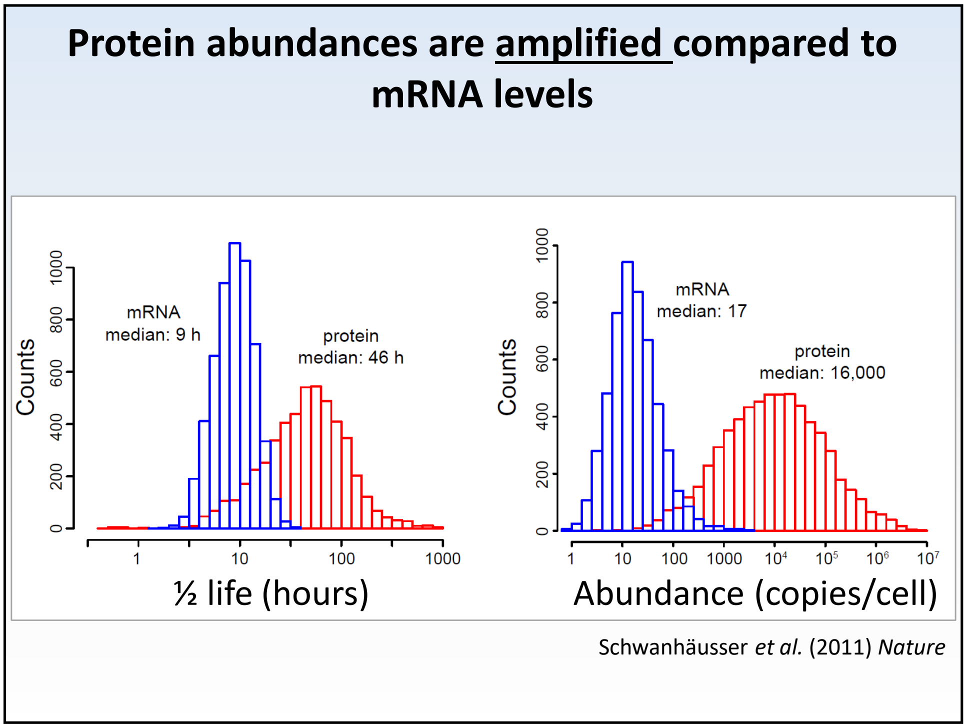
Transcriptome vs. Proteome





These quantification methods can be used to measure the differences between mRNA and protein levels discussed earlier

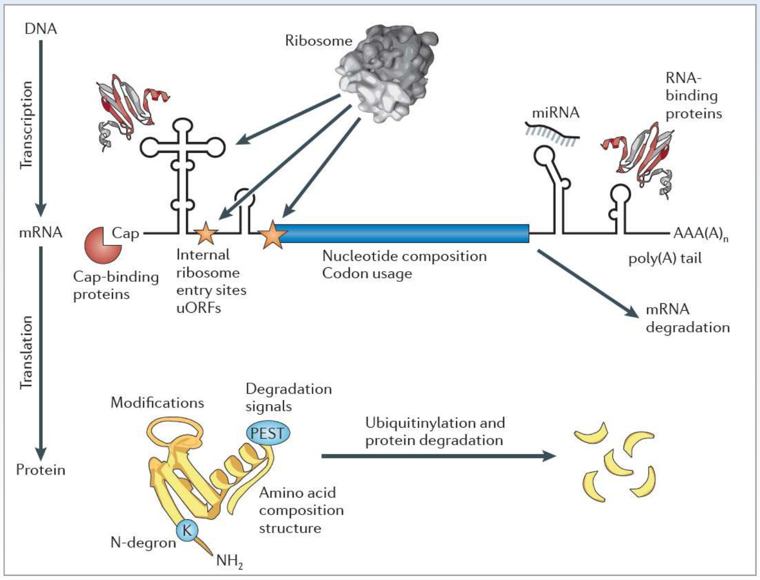
Schwanhäusser (2011) *Nature*



Protein abundances are amplified compared to mRNA levels

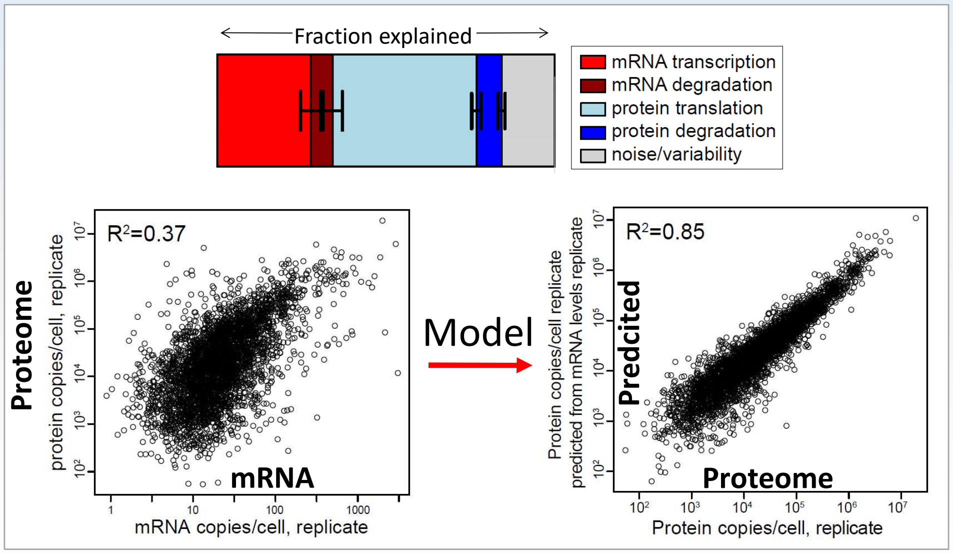
Schwanhäusser *et al.* (2011) *Nature*

Steady-state protein abundances are determined by a balance of regulation occurring at the levels of both RNA and protein



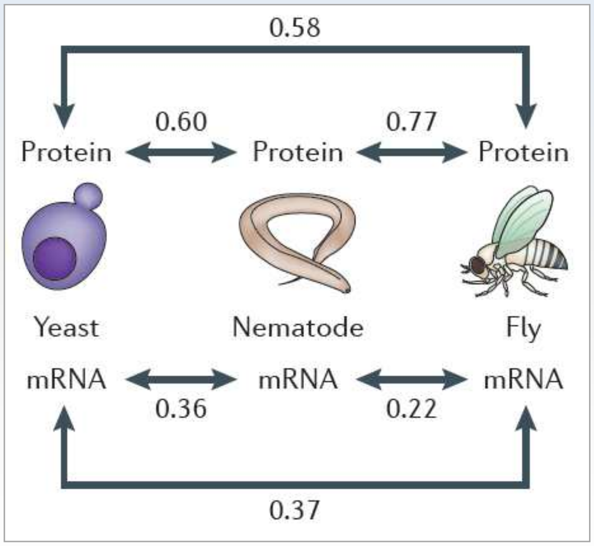
Vogel, *et al.* (2012) *Nat Rev Genetics*

~85% of the variation in protein abundance can be explained



Schwanhäusser *et al.* (2011) *Nature*

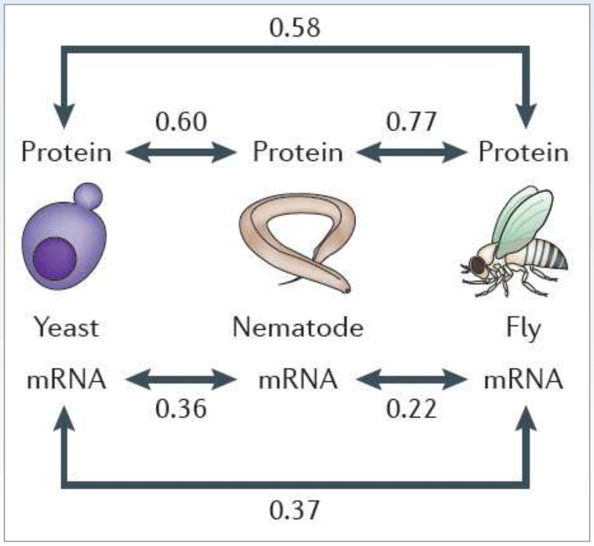
Protein levels show far greater conservation across species than mRNA levels



The #'s are correlation coefficients.

Laurent, *et al.* (2010) *Proteomics*

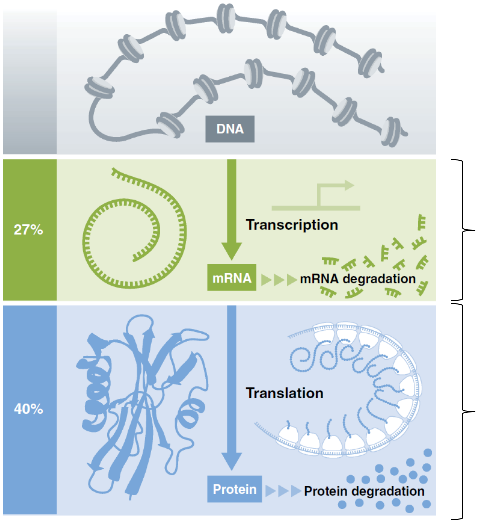
Protein levels show far greater conservation across species than mRNA levels



The #'s are correlation coefficients.

Laurent, *et al.* (2010) *Proteomics*

Transcription is the switch, downstream regulation the rheostat

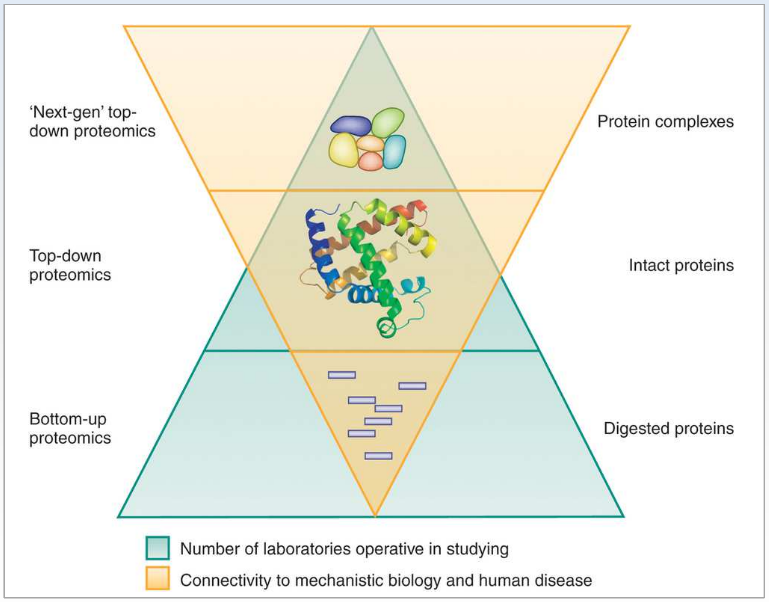


Controls on/off state, roughly sets protein abundance

Fine-tune protein abundances

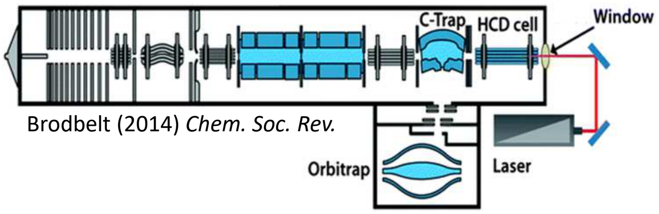
Figure adapted from Plotkin, *Mol Sys Bio* 2010

Top Down Proteomics



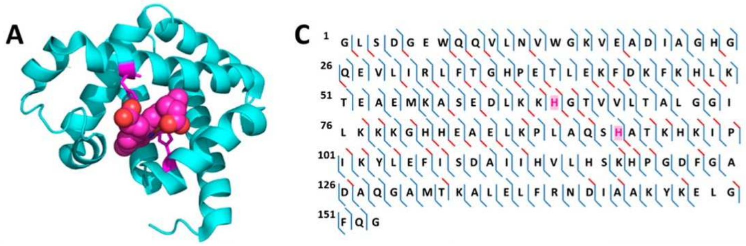
Compton and Kelleher (2012) *Nat. Methods*

Top Down Proteomics



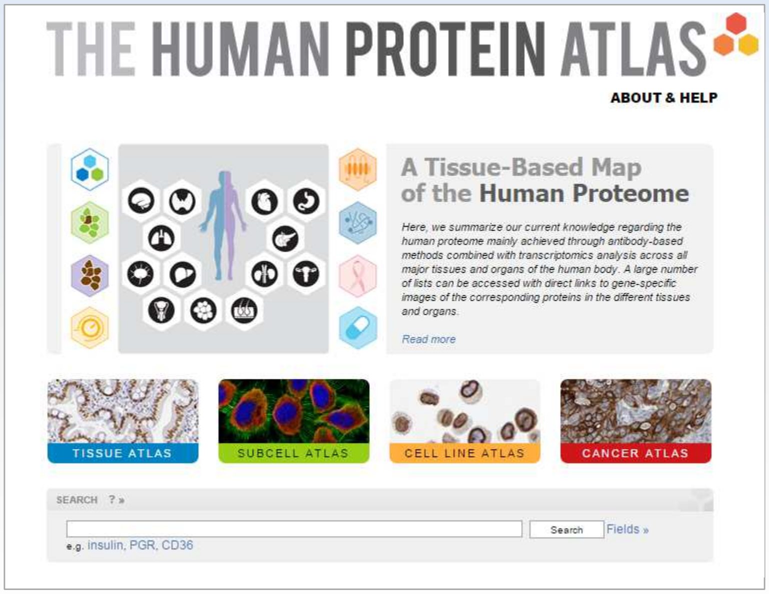
Brodbelt (2014) *Chem. Soc. Rev.*

Mass Specs with friggin' lasers on them!



O'Brien *et al.* (2014) *J. Am. Chem. Soc.*

Beyond Mass Spec: High-throughput Protein Mapping



A Tissue-Based Map of the Human Proteome

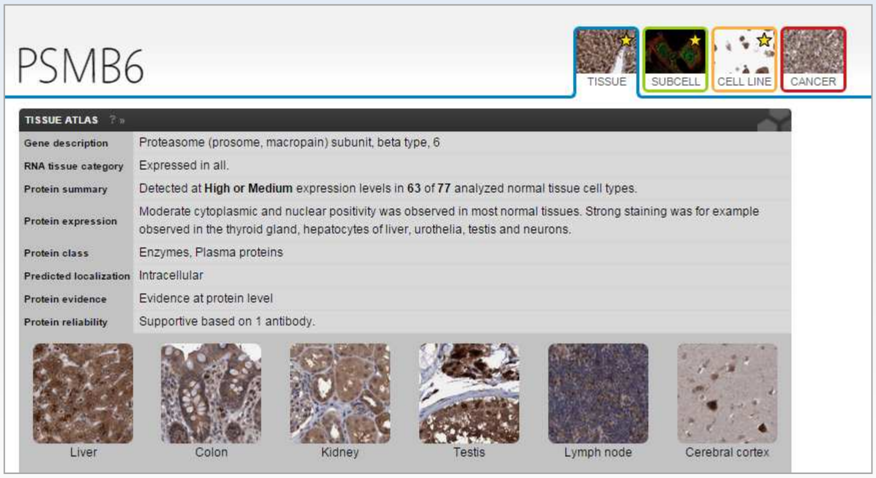
Here, we summarize our current knowledge regarding the human proteome mainly achieved through antibody-based methods combined with transcriptomics analysis across all major tissues and organs of the human body. A large number of lists can be accessed with direct links to gene-specific images of the corresponding proteins in the different tissues and organs.

Read more

e.g. insulin, PGR, CD36

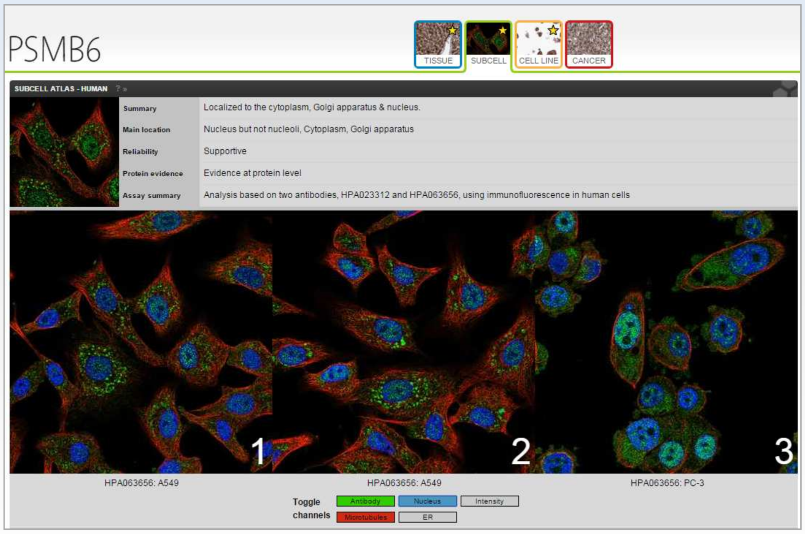
www.proteinatlas.org

Human Protein Atlas



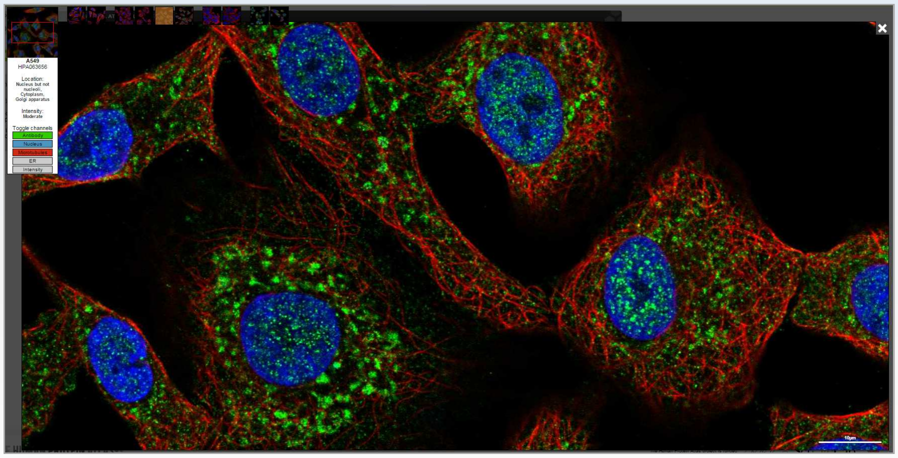
Gene description	Proteasome (prosome, macropain) subunit, beta type, 6
RNA tissue category	Expressed in all.
Protein summary	Detected at **High or Medium** expression levels in **63** of **77** analyzed normal tissue cell types.
Protein expression	Moderate cytoplasmic and nuclear positivity was observed in most normal tissues. Strong staining was for example observed in the thyroid gland, hepatocytes of liver, urothelia, testis and neurons.
Protein class	Enzymes, Plasma proteins
Predicted localization	Intracellular
Protein evidence	Evidence at protein level
Protein reliability	Supportive based on 1 antibody.

Human Protein Atlas



Summary	Localized to the cytoplasm, Golgi apparatus & nucleus.
Main location	Nucleus but not nucleoli, Cytoplasm, Golgi apparatus
Reliability	Supportive
Protein evidence	Evidence at protein level
Assay summary	Analysis based on two antibodies, HPA023312 and HPA063656, using immunofluorescence in human cells

Human Protein Atlas

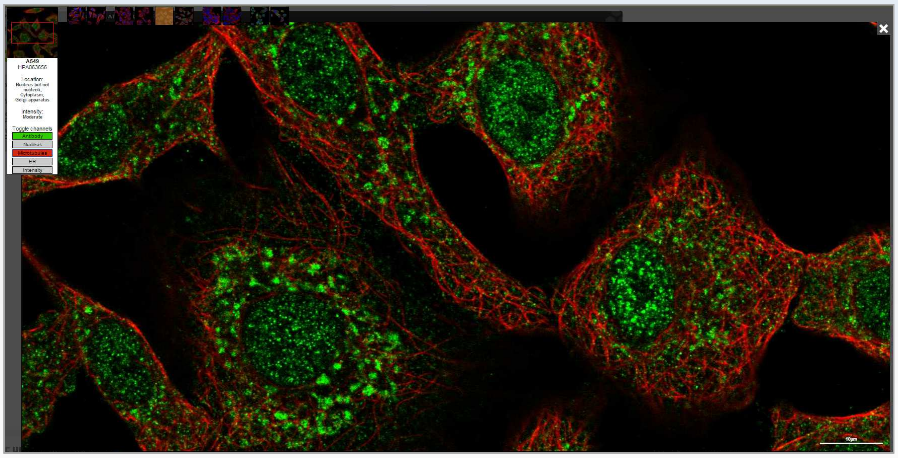


A549
HPA063656

Location:
Nucleus but not nucleoli, Cytoplasm, Golgi apparatus

Intensity:
Moderate

Human Protein Atlas



A549
HPA063656

Location:
Nucleus but not nucleoli, Cytoplasm, Golgi apparatus

Intensity:
Moderate