

Proteomics & Protein-Protein Interactions

Jesse Rinehart, PhD

Biomedical Data Science: Mining & Modeling

CBB 752, Spring 2023

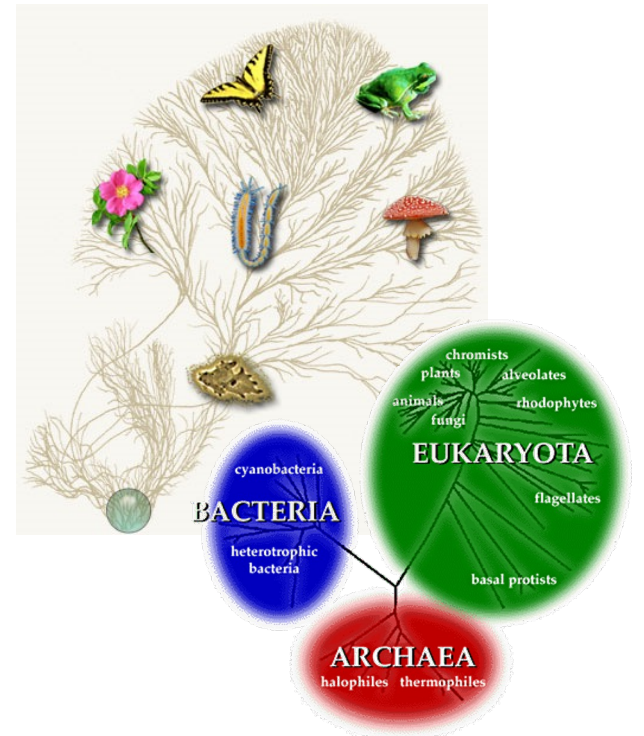
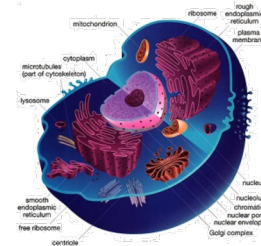
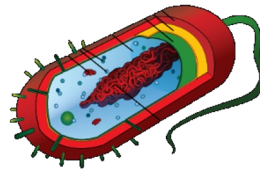
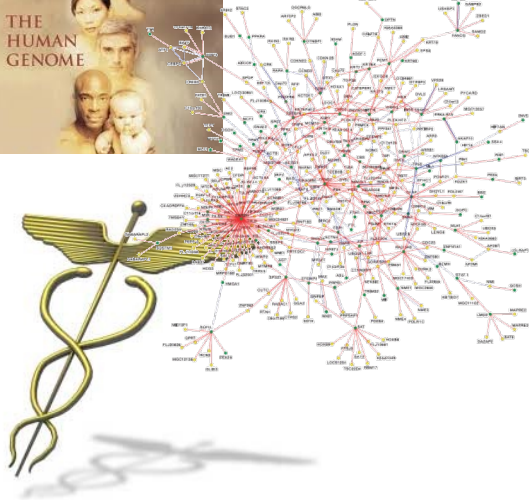
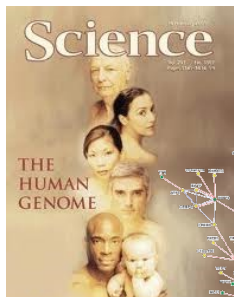


**Cellular & Molecular Physiology
Yale University School of Medicine**

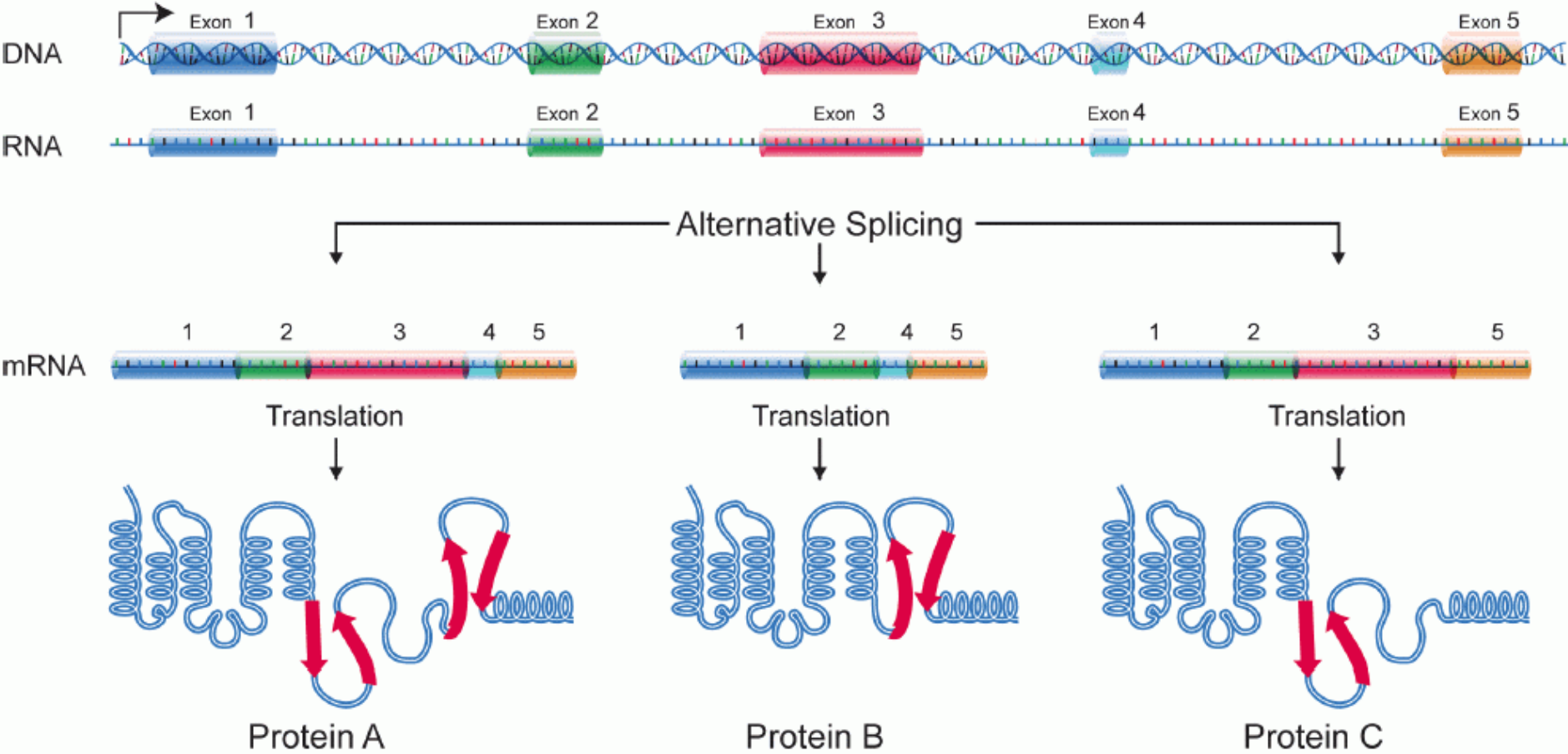


DNA → RNA → PROTEIN

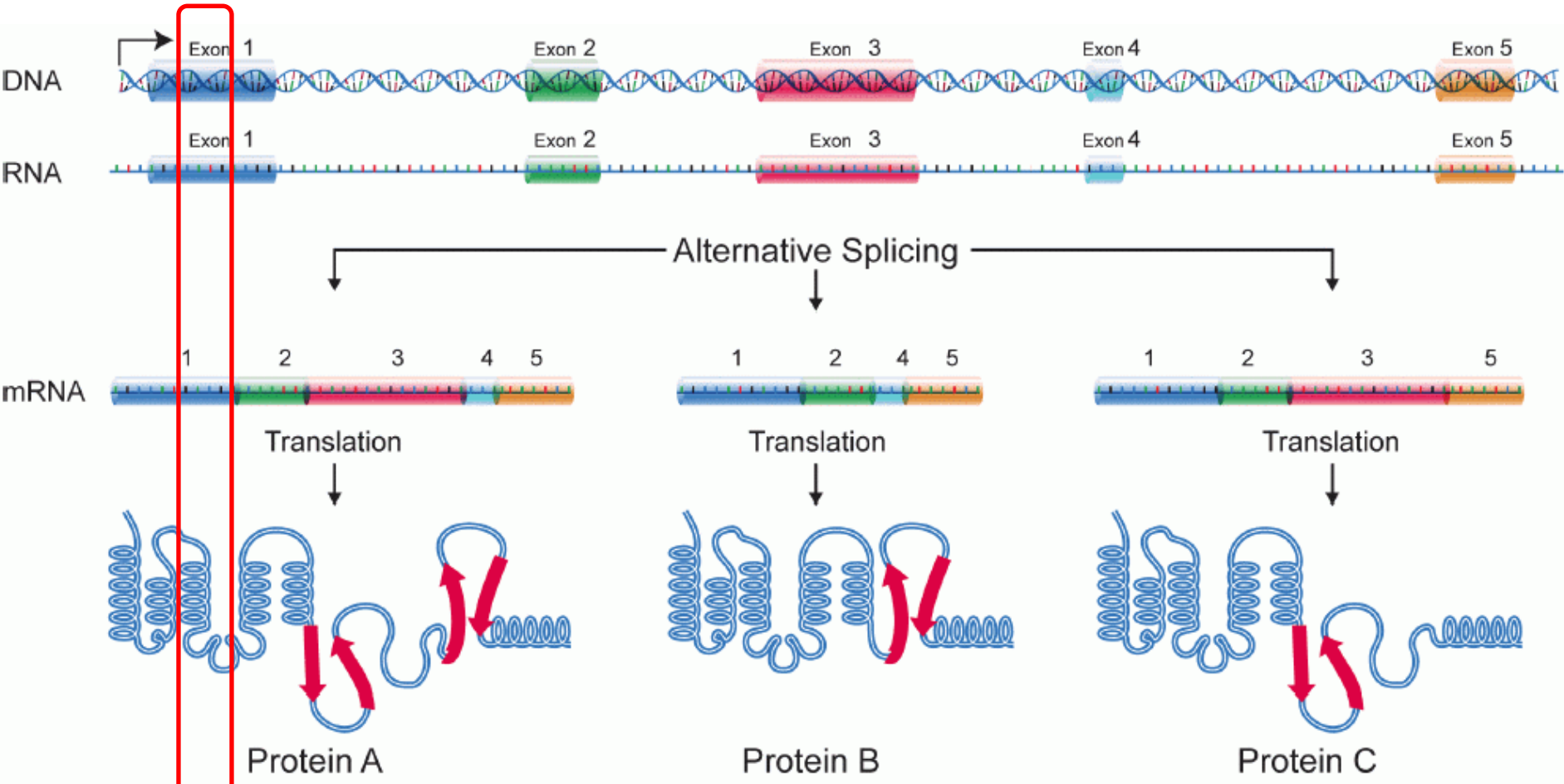
DNA → RNA → PROTEIN



DNA → RNA → PROTEIN

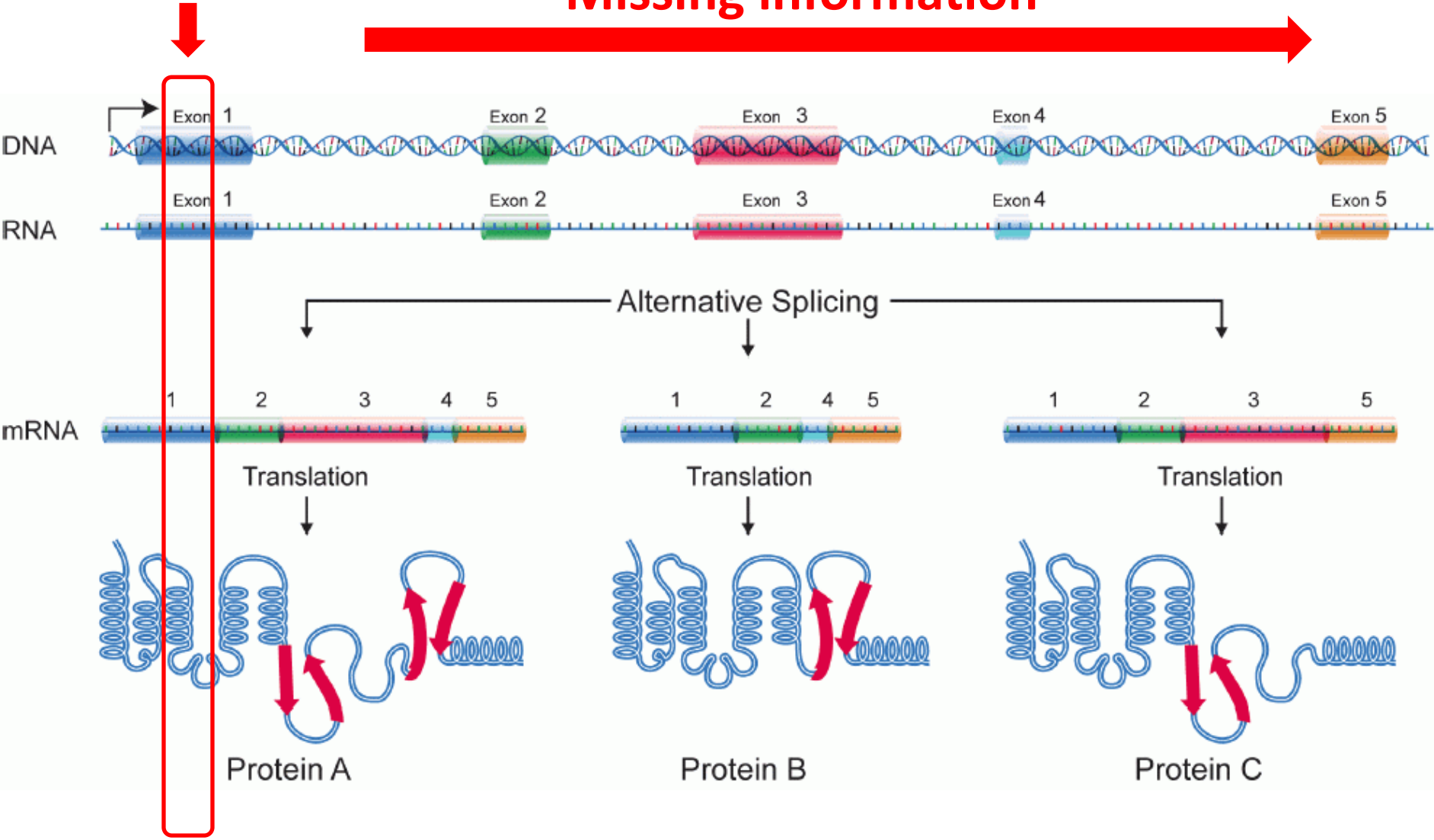


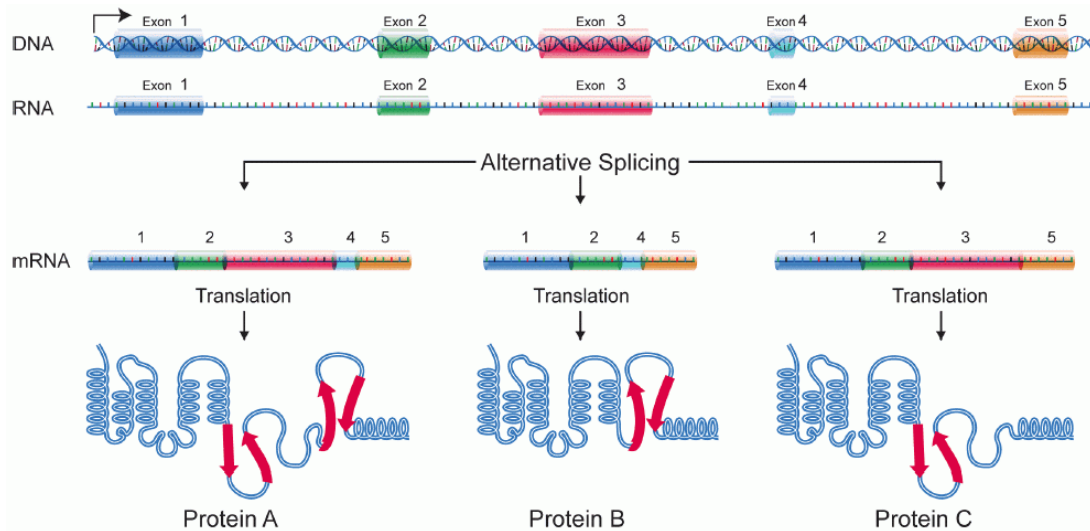
Data capture here



Data capture here

Missing information





Solutions are coming that will completely change databases

- Multiple fields developing methods and technology
- Extend read length
- Measure intact molecules
- Push towards single molecule measurements

Major challenges prevent complete proteome analysis

- **Proteomics is sample limited**
 - Recombinant DNA polymerases revolutionized genome sequencing by allowing for amplification of DNA samples
 - Proteomics has no “polymerase” or amplification method and must contend with natural abundancies
- **Mass spectrometry has limitations**
 - No mass spectrometer, or method, can yet provide full amino-acid resolution of a proteome

Proteomics

The study of the expression, location, modification, interaction, function, and structure of all the proteins in a given cell, organelle, tissue, organ, or whole organism.

Proteomics & Protein-Protein Interactions

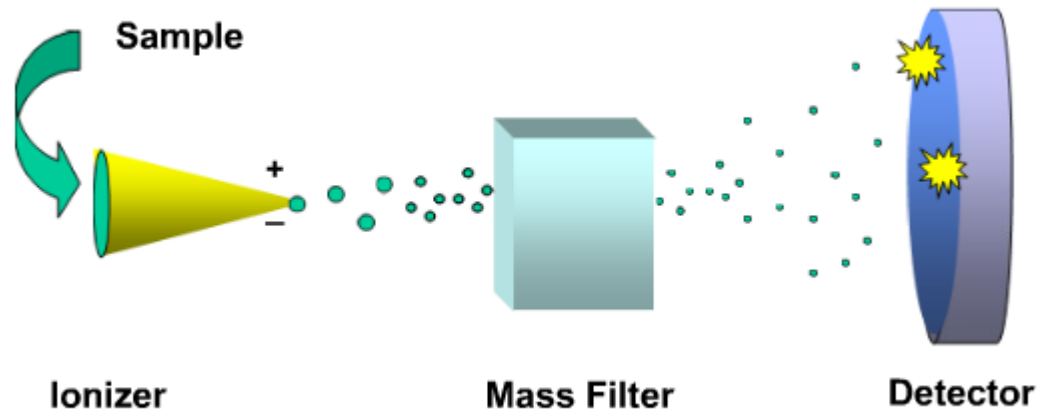
Overview

- **Techniques & Technologies**
 - Mass Spectrometry
 - Protein-Protein Interactions
 - Quantitative Proteomics
- **Applications**
 - Representative Studies
- **Putting it all together....**
 - Databases & Pathways

Principles of Mass Spectrometry (MS)

- In a mass spectrum we measure m/z (mass-to-charge)
- For proteins we measure peptide m/z
- A sample must be ionizable in order to be analyzed

Basic Components of a Mass Spectrometer



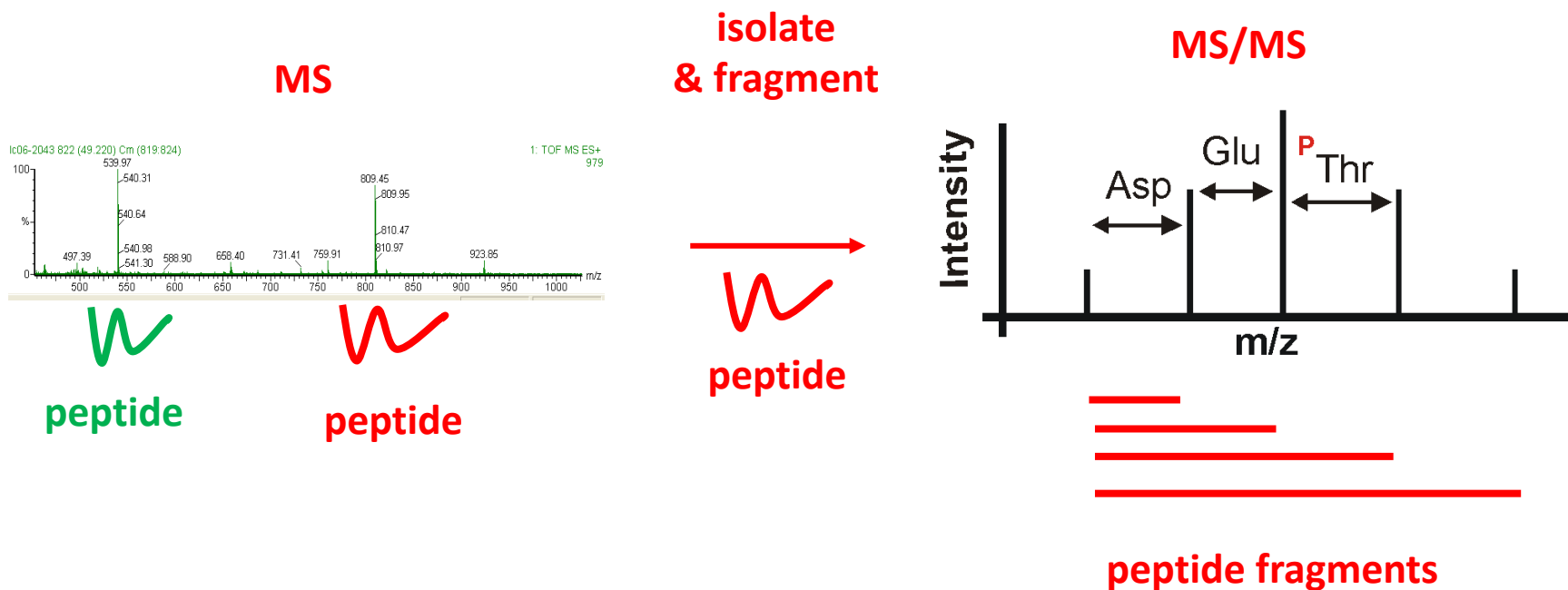
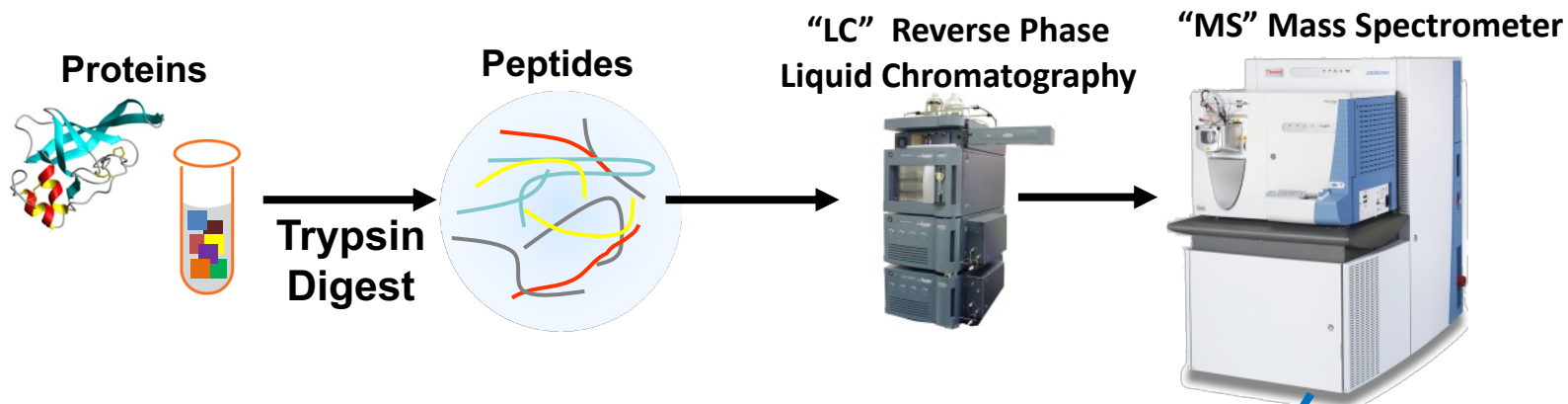
Two major ionization techniques enabled the success of mass spectrometry in the life sciences.

- Electrospray Ionization (ESI)
Fenn JB, *Mann M, Meng CK, Wong SF, Whitehouse CM. *Science*. 1989
- Matrix Assisted Laser Desorption Ionization (MALDI)
Tanaka K, Waki H, Ido Y, et al. *Rapid Commun Mass Spectrom* 1988
- 2002 Nobel Prize in Chemistry awarded to
John B. Fenn & Koichi Tanaka
- Enabled direct measurement and “sequencing” of intact peptides & MS based Proteomics is born

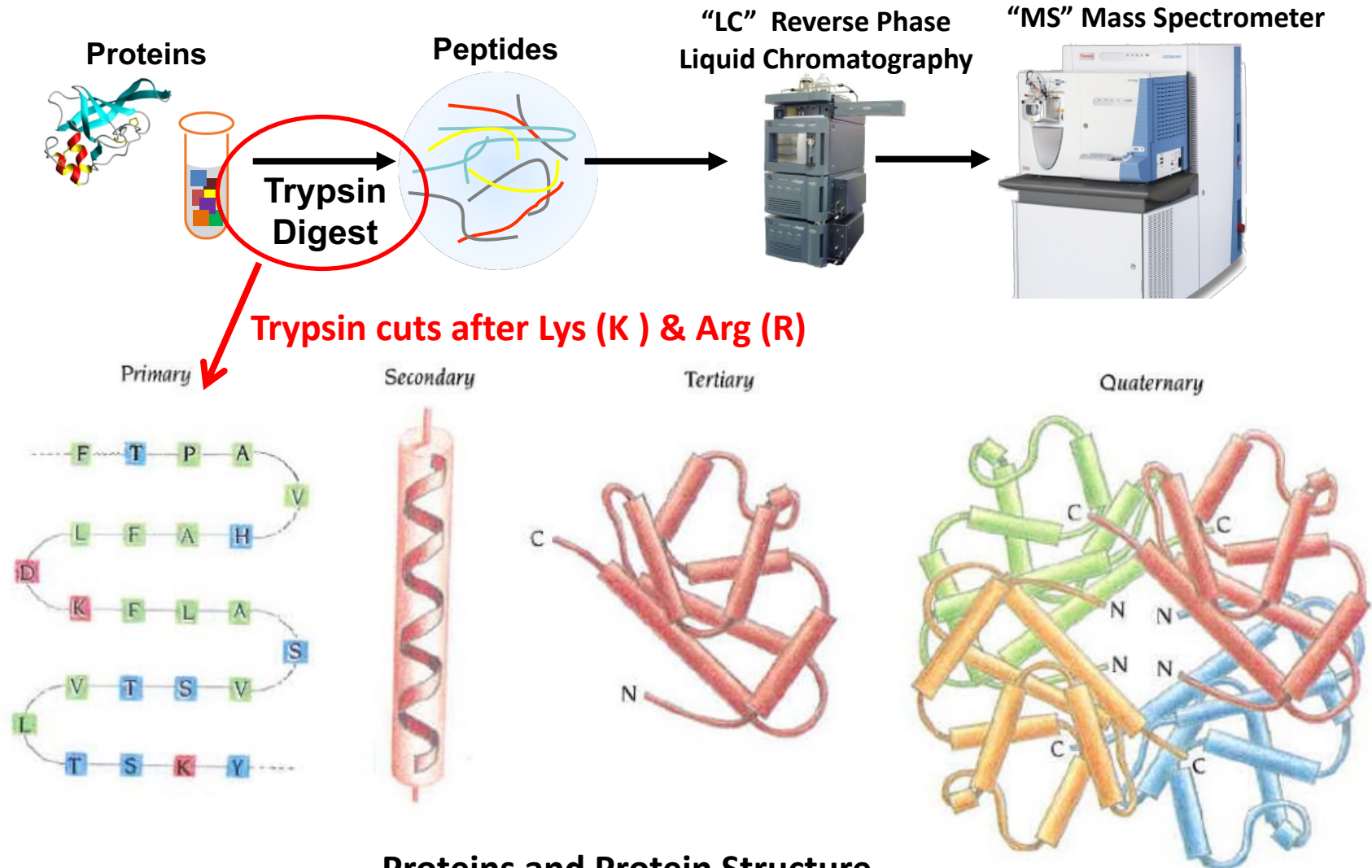
*

Matthias Mann (Yale University; Ph.D.; 1988; Chemical Engineering) trained with John Fenn during some of the breakthrough work at Yale

Typical work-flow for LC-MS "shotgun proteomics"



Typical work-flow for LC-MS “shotgun proteomics”



Proteins and Protein Structure

(Branden, C. and Tooze, J. *Introduction to Protein Structure*)

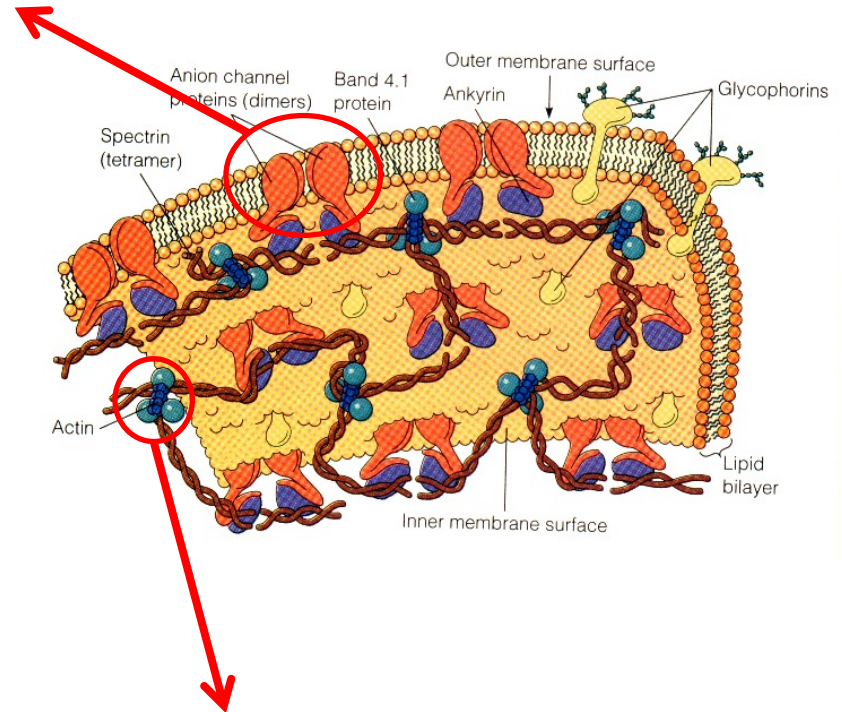
Trypsin digest followed by LC-MS: Examples of “Sequence Coverage”

Band 3 Anion Transporter

Matched peptides shown in Bold Red

```

1 MEELQDDYED MMEENLEQEE YEDPDIPESQ MEEPAAHDE ATATDYHTS
51 HPGTHKVVVE LQELVMDEKN QELRWMEEAR WVQLEENLGE NGAWGRPHLS
101 HLTFWSLLEL RRVFTKGTVL LDLQETSLAG VANQLLDRFI FEDQIRPQDR
151 EELLRALLLK HSHAGELEAL GGVKPAVLTR SGDPSQPLL QHSSLETQLF
201 CEQGDGGTEG HSPSGILEKI PPDSEATLVL VGRADFLEQP VLGFVRLQEA
251 AELEAVELPV PIRFLFVLLG PEAPHIDYTQ LGRAAATLMS ERVFRIDAYM
301 AQSRGELLHS LEGFLDCSLV LPPTDAPSEQ ALLSLVPVQR ELLRRRYQSS
351 PAKPDSSFYK GLDLNGGPDD PLQQTGQLFG GLVRDIRRRY PYLSDITDA
401 FSPQVLAAVI FIYFAALSPA ITFGGLLGEK TRNQMGVSEL LISTAVQGIL
451 FALLGAQPLL VVGFSGPLLV FEEAFFSFCE TNGLEYIVGR VWIGFWLILL
501 VVLVVAFEFS FLVRFISRYT QEIFSFLISL IFIYETFSKL IKIFQDHPLQ
551 KTYNYNVLV PKPQGPLPNT ALLSLVLMAG TFFFAMMLRK FKNSSYFPGK
601 LRRVIGDFGV PISILIMVLV DFFIQDITYTQ KLSVPDGFVK SNSSARGWVI
651 HPLGLRSEFP IWMMFASALP ALLVFILIFL ESQITTLIVS KPERKMKVGS
701 GFHLDLLLIV GMGGVAALFG MPWLSATTVR SVTHANALTV MGKASTPGAA
751 AQIQEVKEQR ISGLLVAVLV GLSILMEPIL SRIPLAVLFG IFLYMGVTSL
801 SGIQLFDRIL LLFKPPKYHP DVPYVKRVKT WRMHLFTGIQ IICLAVLWVV
851 KSTPASLALP FVLILTIVPLR RVLLPLIFRN VELQCLDADD AKATFDEEEG
901 RDEYDEVAMP V
    
```



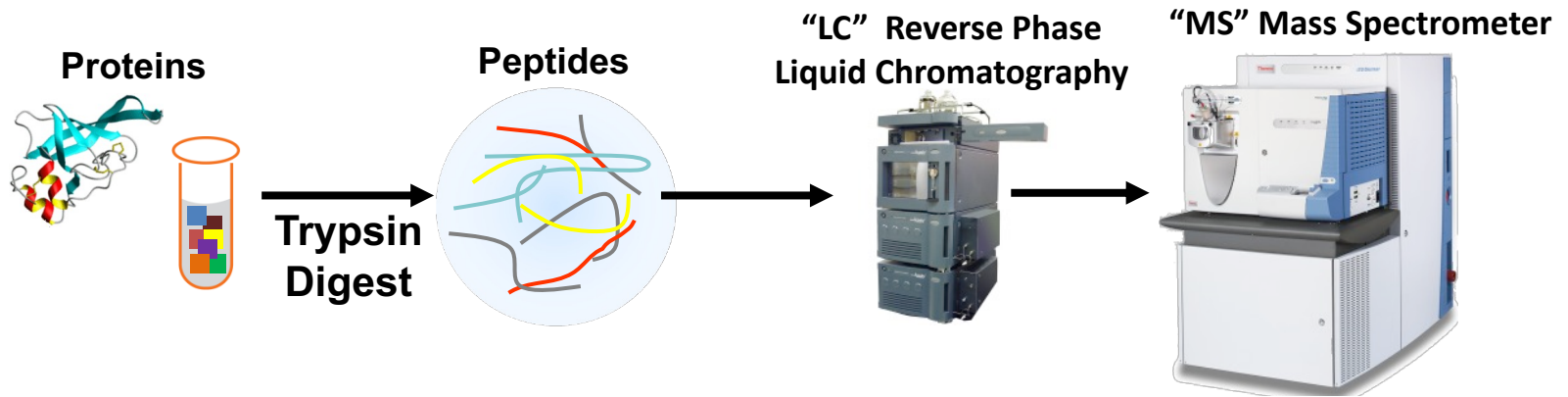
β -actin

Matched peptides shown in Bold Red

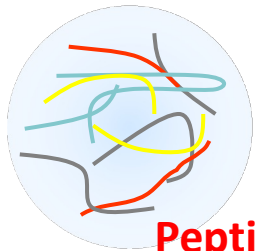
```

1 MDDDIAALVV DNGSGMCKAG FAGDDAPRAV FPSIVGRPRH QGVVMGMGQK
51 DSYVGDEAQS KRGILTLYYP IEHGIVTNWD DMEKIWHHTF YNELRVAPEE
101 HPVLLTEAPL NPKANREKMT QIMPETFNTF AMYVAIQAVL SLYASGRTTG
151 IVMDSGDGVT HTVPIYEGYA LPHAILRLDL AGRDLTDYLM KILTERGYSF
201 TTTAEREIVR DIKEKLCYVA LDPEQEMATA ASSSSLEKSY ELPDQVITI
251 GNERFRCPEA LFQPSFLGME SCGIHETTFN SIMKCDVDIR KDLYANTVLS
301 GGTMYPGIA DRMQKEITAL APSTMKIKII APPERKYSVW IGGASILASLS
351 TFQQMWISKQ EYDESGPSIV HRKCF
    
```

The mass spectra of peptide mixtures are complex

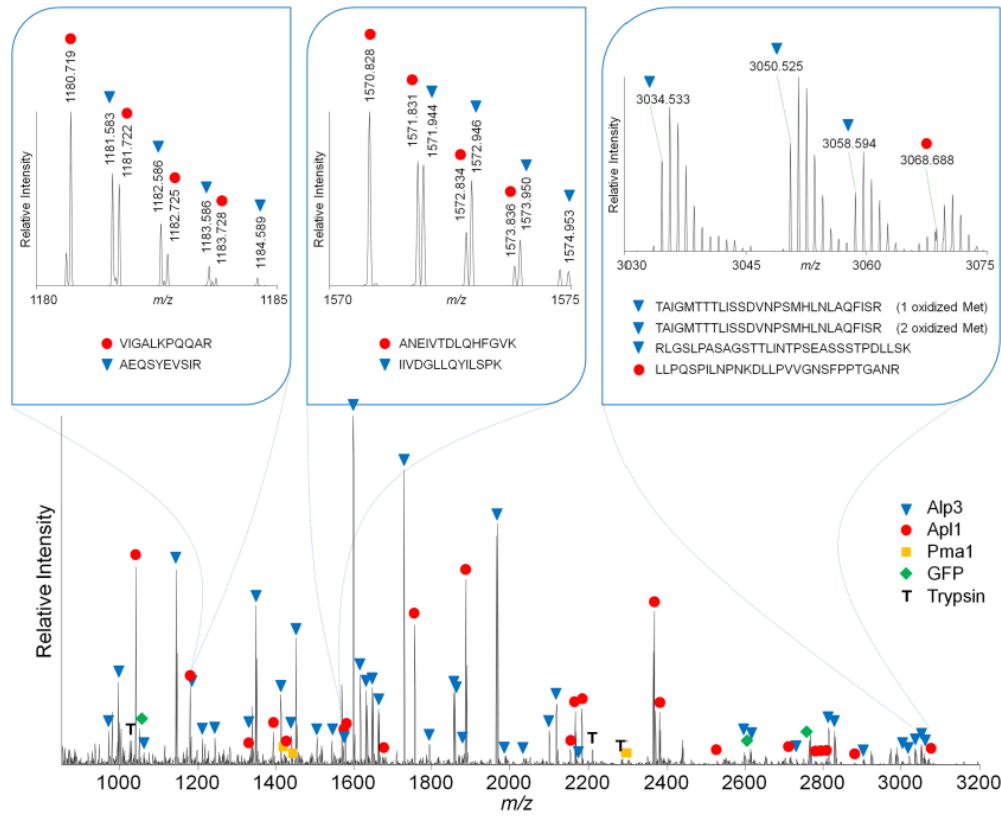


Mass Spectrum

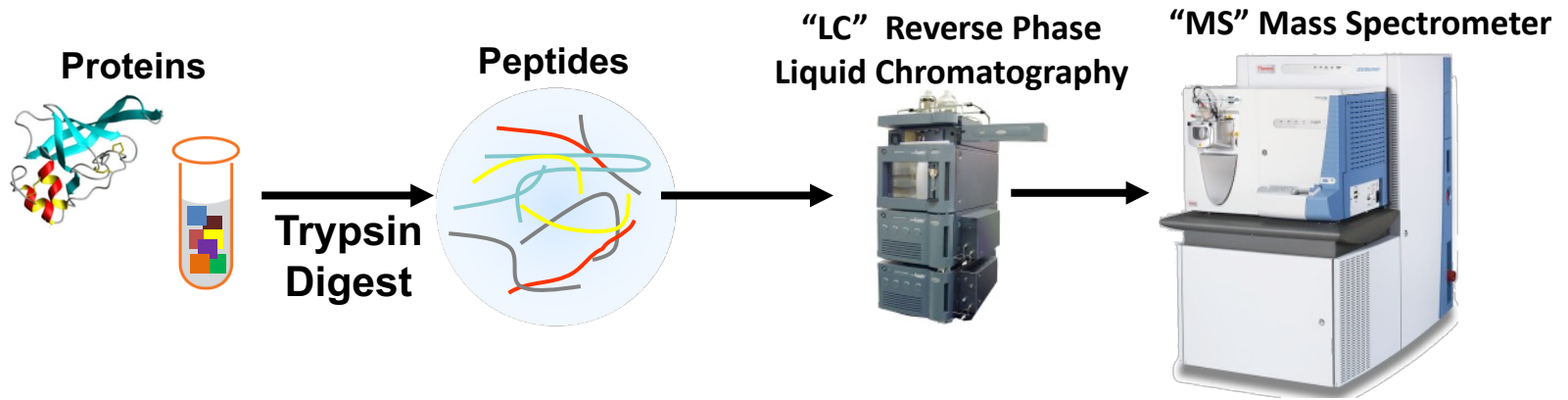


Peptide ions have a mass (m) and a charge (z).

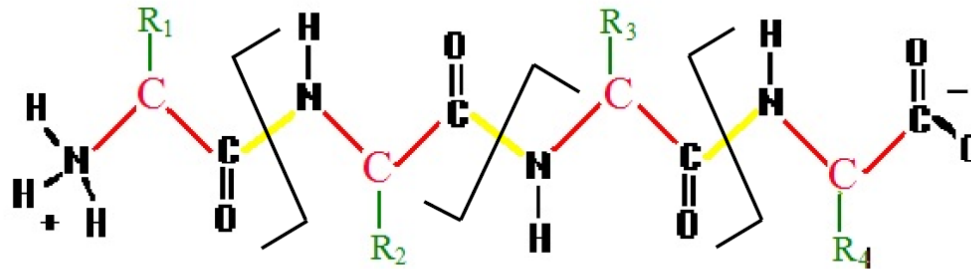
100 Da peptide:
 +1 = 100 m/z
 +2 = 50 m/z
 +3 = 33.3 m/z



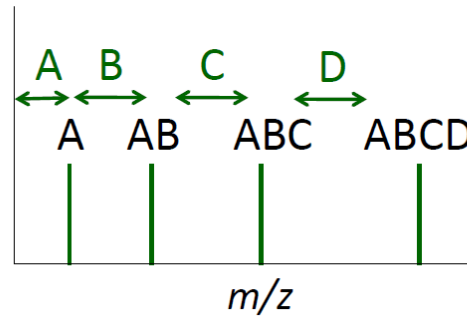
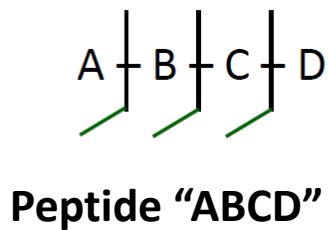
Peptide ions are isolated, fragmented, and “sequenced”



Peptide sequencing



Simplified concept of peptide fragmentation

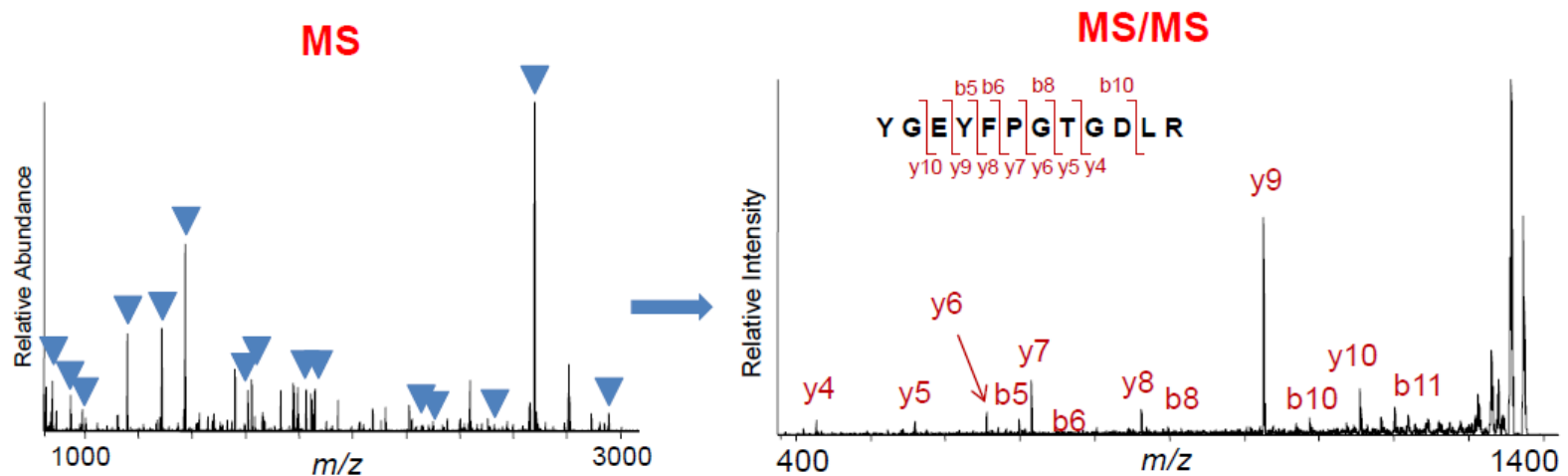


Fragment Spectra of Peptide “ABCD”

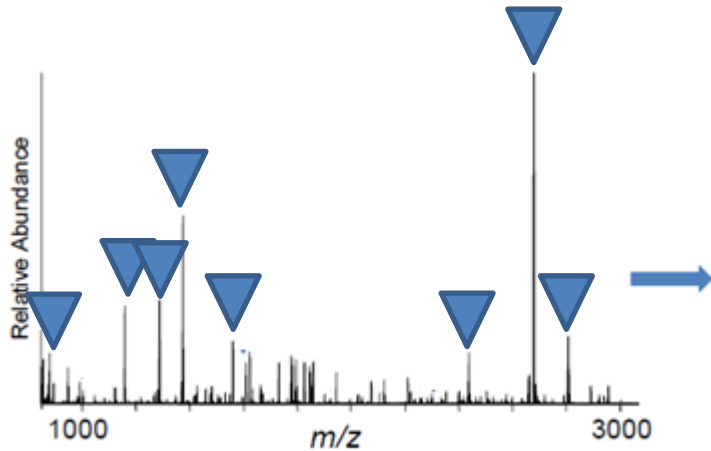
Computational Steps:

- Massive amounts of MS and MS/MS data need interpretation
- Genome databases define proteome
- Proteome database used to “match” peptide sequence data

Database searching - at MS or MS/MS level

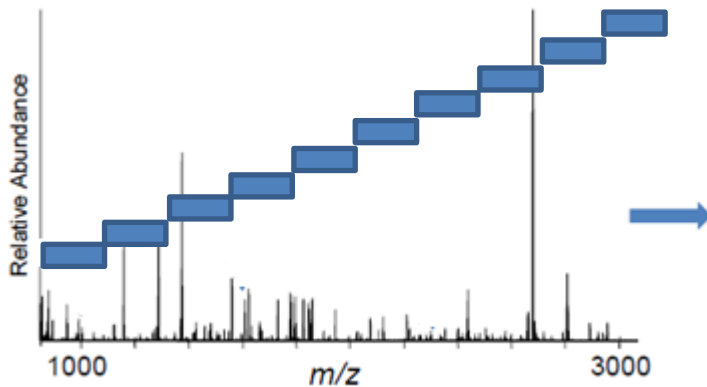


DIA (Data-independent Acquisition) vs. DDA (Data-dependent Acquisition)



DDA (Data-dependent Acquisition)

▼ The ***most intense/“abundant”*** ions are selected for MS/MS sequencing



DIA (Data-independent Acquisition)

■ ***All ions*** in small M/Z windows are selected for MS/MS sequencing

The ***pace of proteomics is set by a combination of techniques and technological advances.**

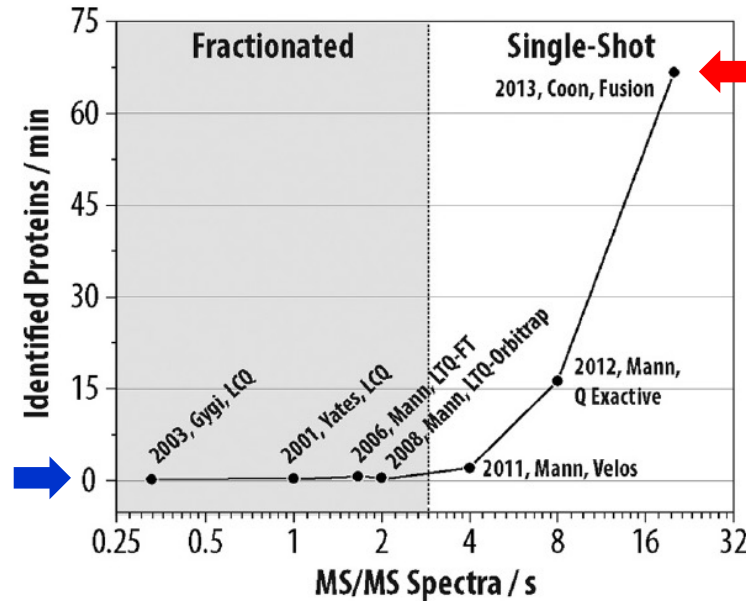
***orders of magnitude behind genome technologies (*sequencing*)**

**Yeast proteome reported
in Washburn et al.**

***Nature Biotech* 2001:**

**~82 hours* = 1,484
proteins ~0.3
proteins/ min**

***estimates from paper: 3
fractions @ 15 X 110 minute
“runs” for each fraction**



**“each one hour
analysis achieved
detection of 3,977
proteins”**

**The one hour yeast
proteome. Hebert et al
Mol Cell Proteomics. 2014**

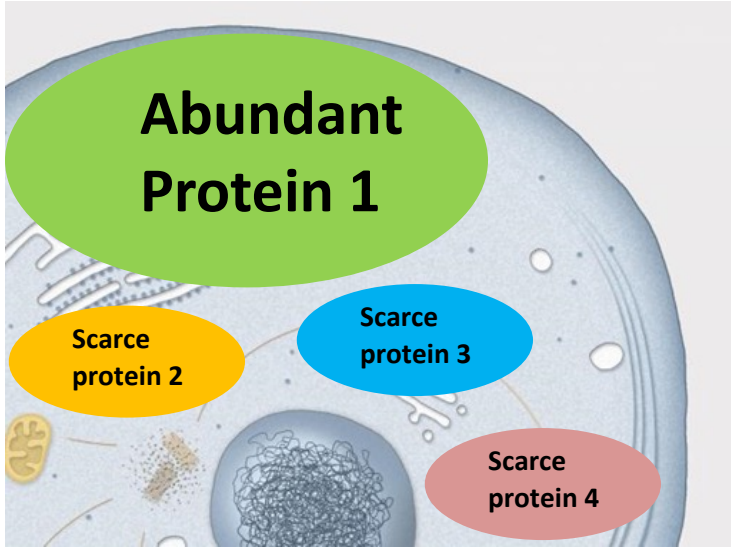
FIG. 5. Rate of protein identifications as a function of mass spectrometer scan rate for selected large-scale yeast proteome analyses over the past decade. Each data point is annotated with the year, corresponding author, type of MS system used, and reference number.

The one hour yeast proteome. Hebert AS, et al, Coon JJ.

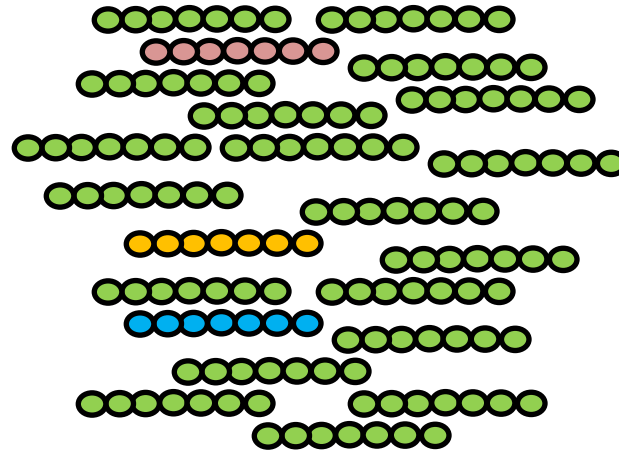
Mol Cell Proteomics. 2014 PMID: 24143002 & *Nat Protoc*. 2015. PMID: 25855955

Challenge Question:

Cell with a 4 protein proteome



Whole Proteome Tryptic Digest



One LC-MS run

(Hypothetical MS that can only identify one peptide)

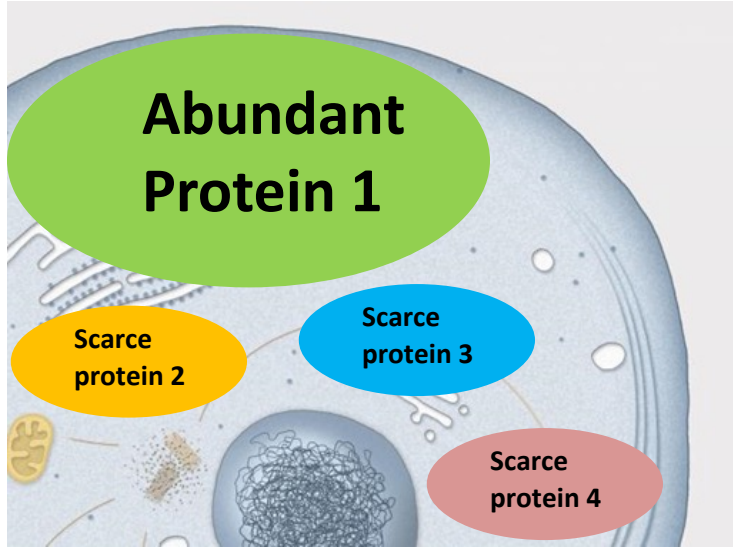


Protein 1
Identified

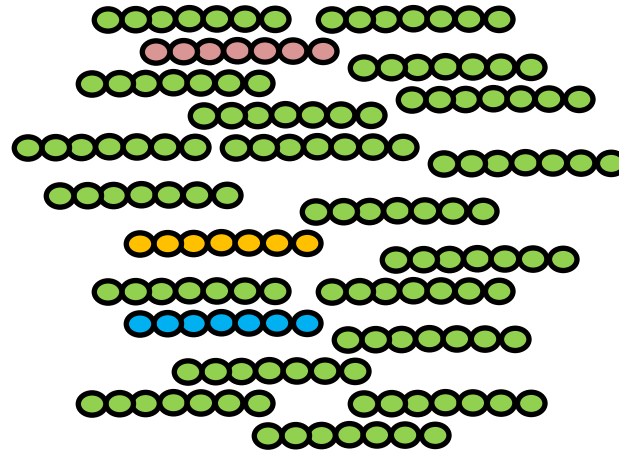
Challenge Question:

How would you detect all four proteins in this cell using a mass spectrometer that can only identify one peptide?

Cell with a 4 protein proteome



Whole Proteome Tryptic Digest



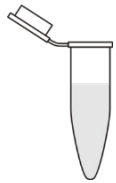
One LC-MS run

(Hypothetical MS that can only identify one peptide)

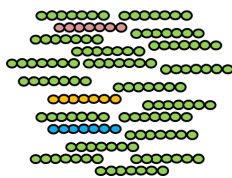


Protein 1
Identified

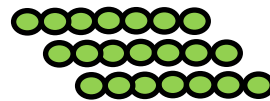
Option #1: Peptide Fractionation



Whole Proteome Tryptic Digest



Chromatography + fractionation



Protein 1
Identified



Protein 2
Identified

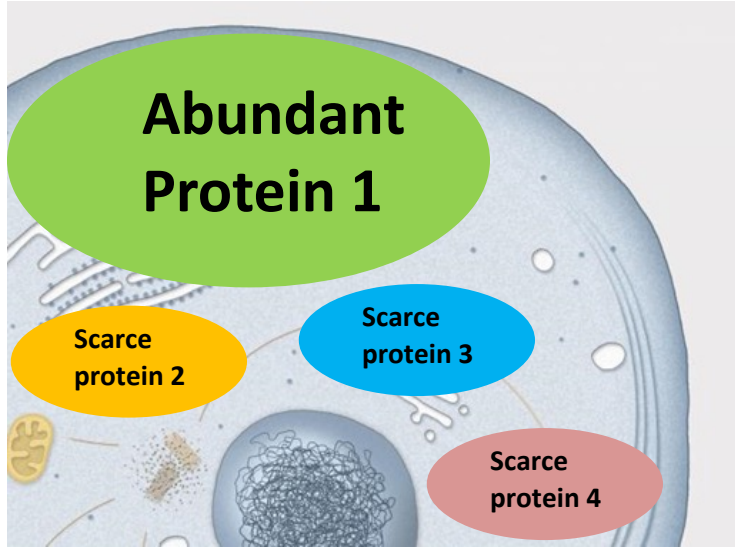


Protein 3
Identified



Protein 4
Identified

Cell with a 4 protein proteome

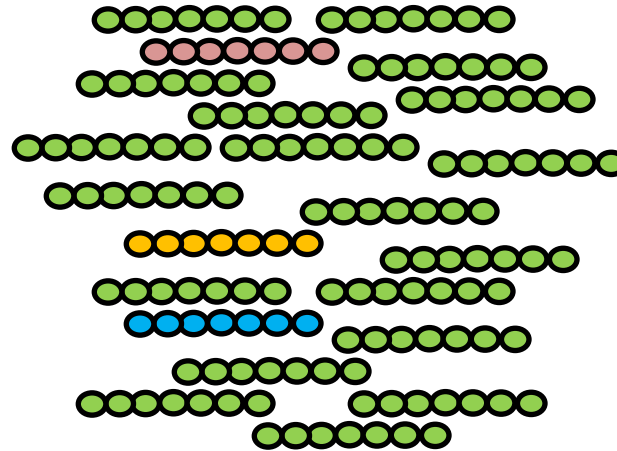


Whole Proteome Tryptic Digest



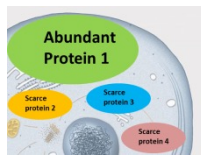
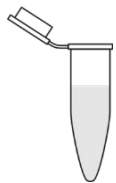
One LC-MS run

(Hypothetical MS that can only identify one peptide)



Protein 1
Identified

Option #2: Proteome Fractionation (e.g. Immunoprecipitation)



Separate IP Tryptic Digest



Abundant Protein 1



Protein 1
Identified



Scarce protein 2



Protein 2
Identified



Scarce protein 3



Protein 3
Identified

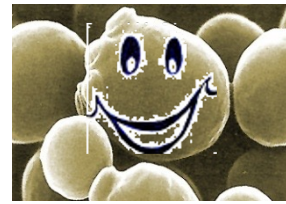


Scarce protein 4



Protein 4
Identified

A tour of proteomics: Studies with the budding yeast *Saccharomyces cerevisiae*



2000 & 2001

Uetz et al, A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* .
& Ito et al, A comprehensive two-hybrid analysis to explore the yeast protein interactome . *PNAS*.

➔ **Large scale yeast two hybrid screens to map proteome wide interactions.**

2001

Washburn, et al. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnol.*

➔ **Established the 'shotgun' technology by showing that many proteins in a yeast-cell lysate could be identified in a single experiment.**

2002

Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*.

& Gavin, A. C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* .

➔ **Protein-protein interaction maps can be obtained by MS; the yeast cell is organized into protein complexes.**

2003

Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature*. & Huh, W. K. *et al.* Global analysis of protein localization in budding yeast. *Nature*.

➔ **TAP-Tag and expression studies & GFP-Tag and localization studies**

2006

Krogan NJ, *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*.

➔ **TAP-Tag and Protein-Protein Interaction**

2008

de Godoy LM, *et al.* Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*.

➔ **SILAC based quantitation of an entire proteome.**

2009

Picotti P, *et al.* Full dynamic range proteome analysis of *S. cerevisiae* by targeted proteomics. *Cell*.

➔ **Towards proteome wide targeted proteomics.**

A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*.

Uetz et al, Nature 2000

Ito et al, PNAS 2001

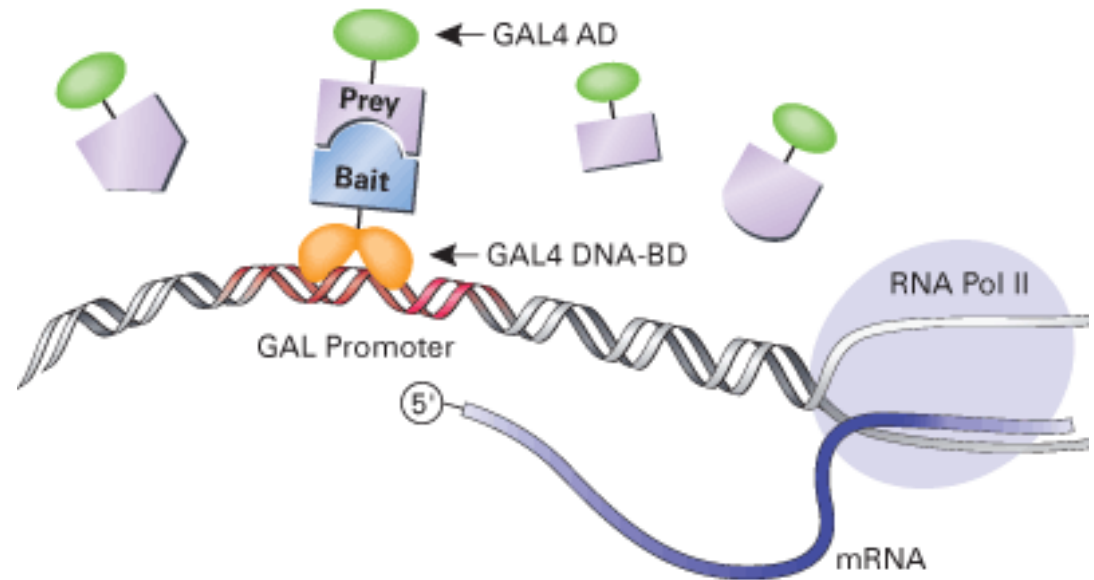
Yeast Two Hybrid Assay

Advantages:

- *In vivo* assay
- Simple

Some Disadvantages

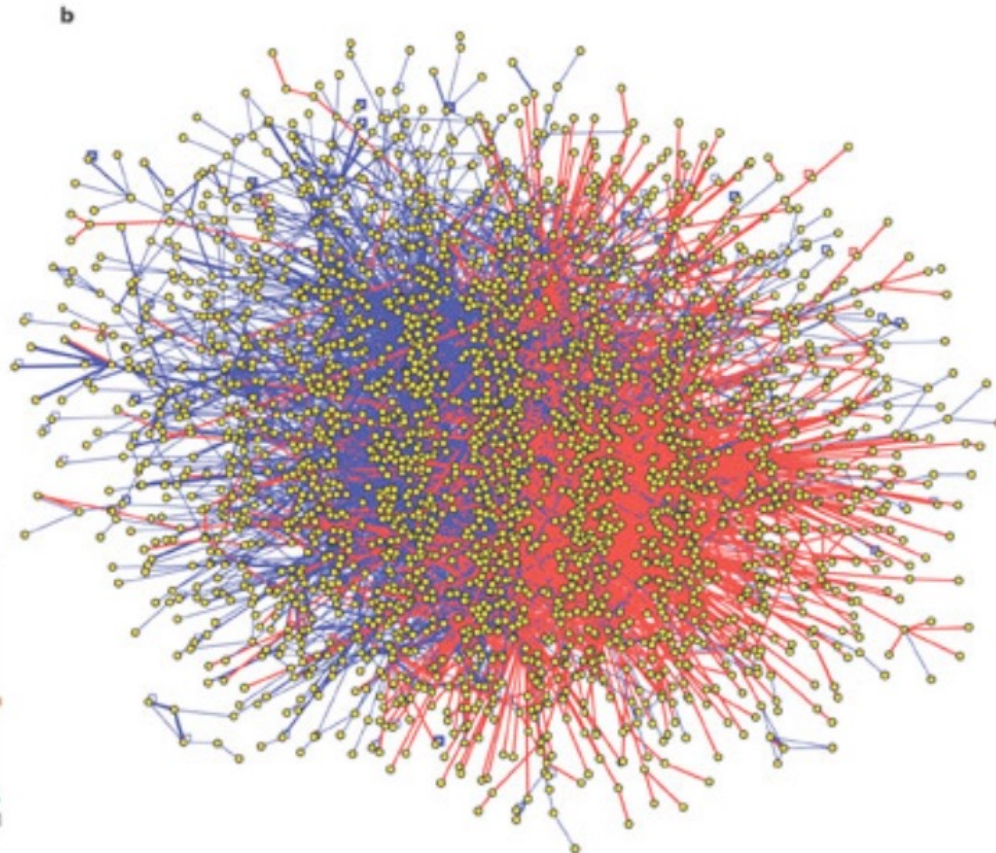
- Hard to execute on large scale
- False positives: a real interaction or “possible” interaction
- Interaction in nucleus (required for GAL system)
- Clones are fusion proteins and sometimes “partial” proteins
- Multiple protein complexes not “captured”



Human Two Hybrid Map

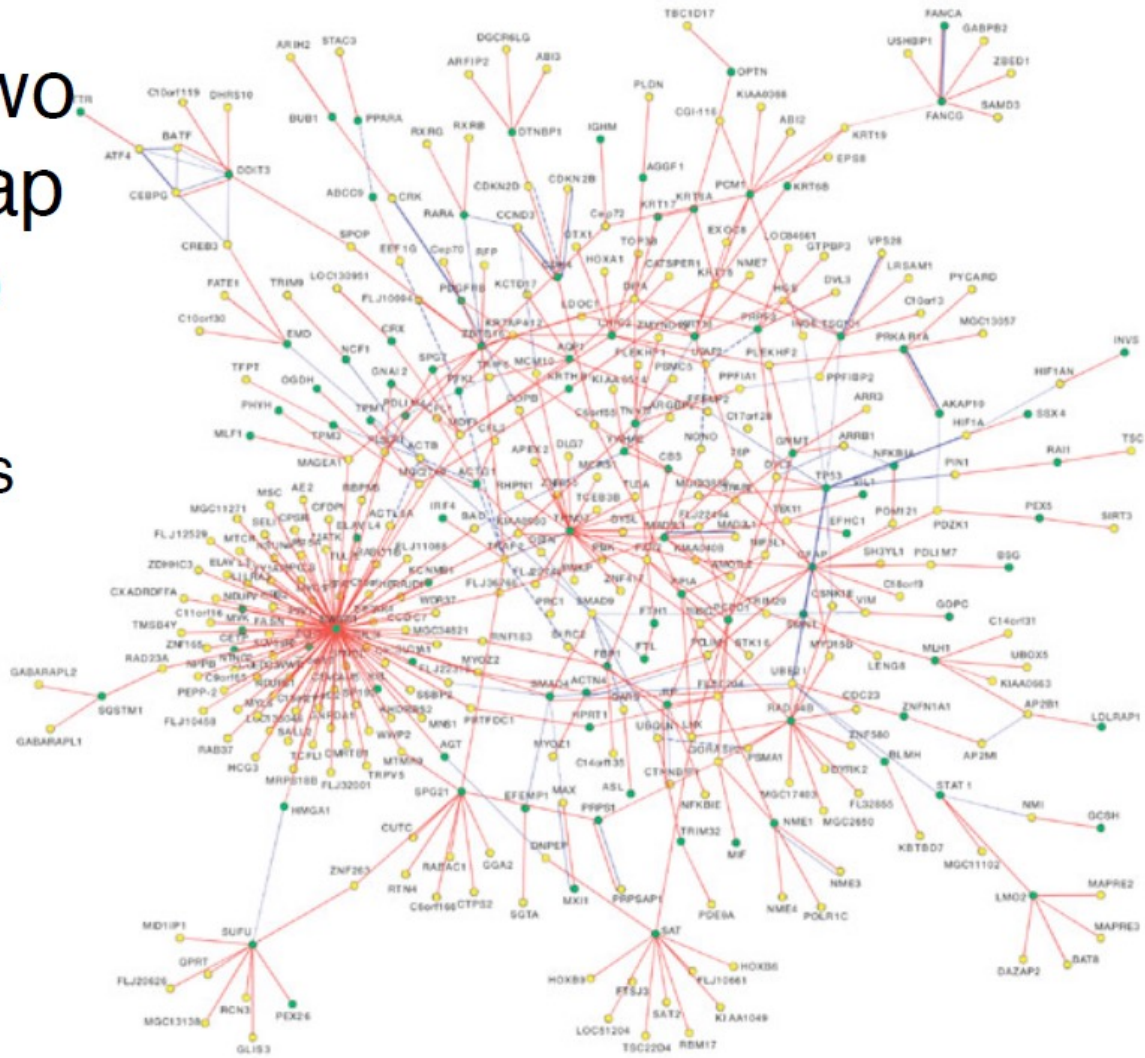
8,100 ORFs (~7,200 genes)

10,597 interactions



Rual et al. Nature 2005

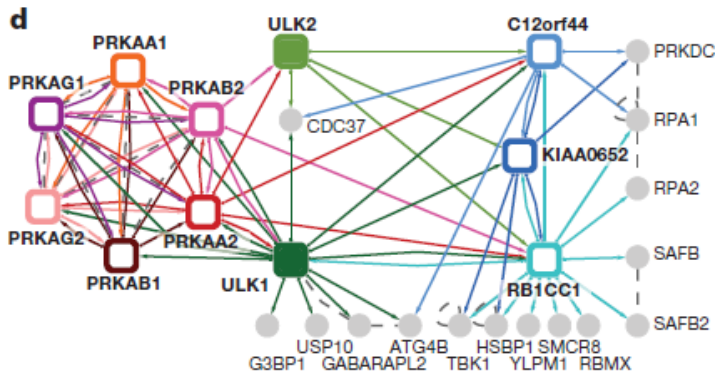
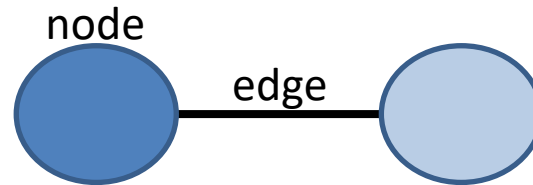
Human Two Hybrid Map Disease Genes (121 genes (green))



Rual et al. Nature 2005 Vol 437

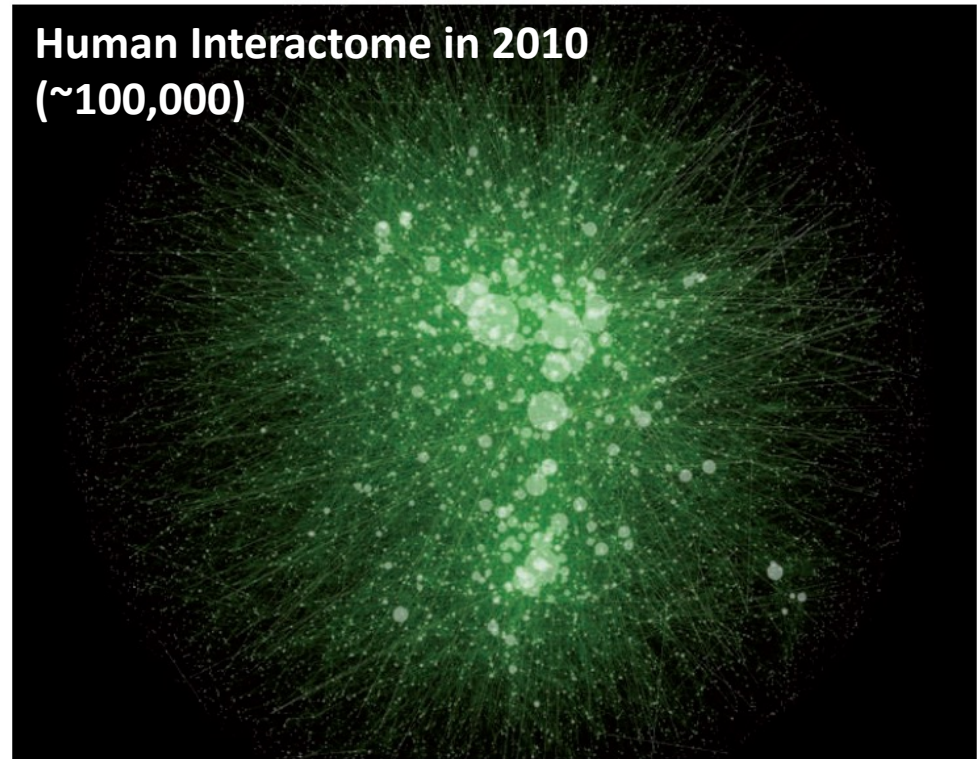
Protein-Protein interaction maps:

Proteins are represented by **nodes** and interactions are represented by **edges** between nodes.



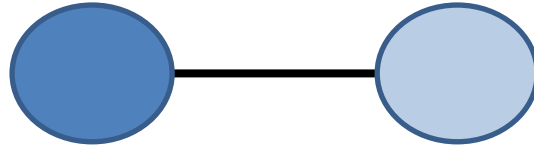
K. ONO/UC SAN DIEGO/CYTOSCAPE

**Human Interactome in 2010
(~100,000)**



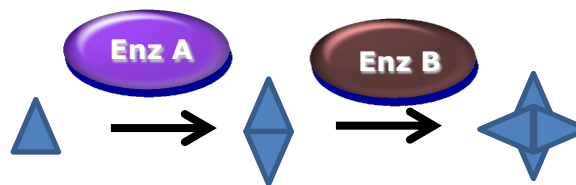
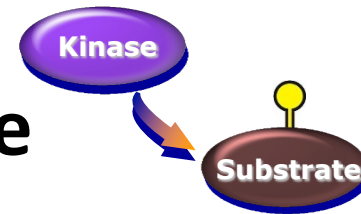
Bonetta, *Nature* 2010

Protein-Protein interactions:

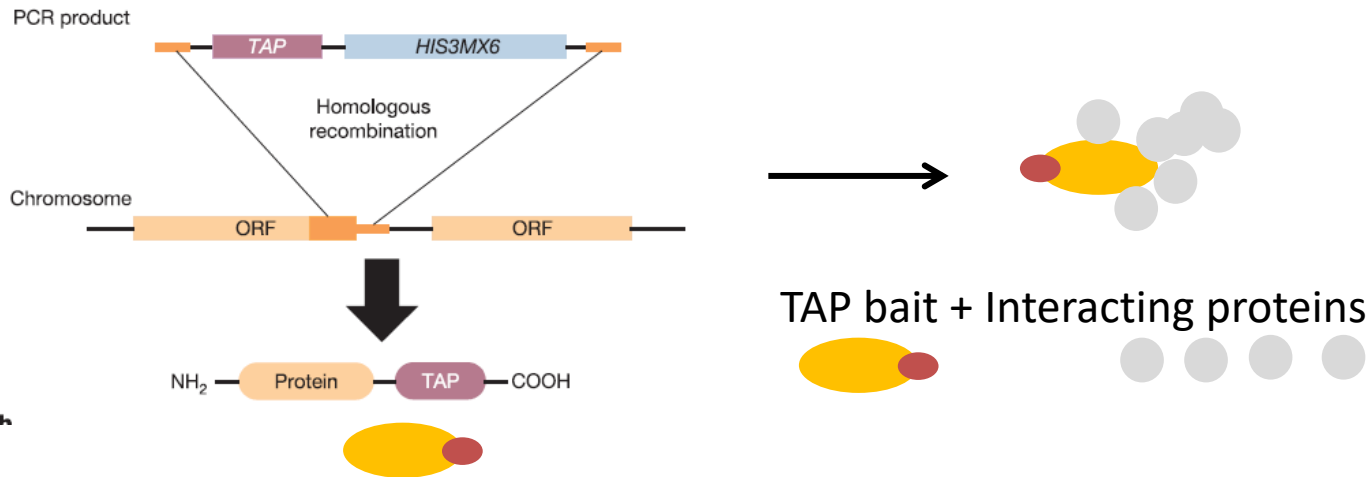


Some examples:

- Physical and direct
- Physical and indirect
- Multi-protein complexes
- Scaffolds
- Transient
- Kinase & substrate
- Metabolic



Adding common molecular handles to every yeast protein: “TAP” tag



Collection of **tagged** “bait” expression strains

2003

Ghaemmaghami, S. et al. Global analysis of protein expression in yeast. *Nature*. & Huh, W. K. et al. Global analysis of protein localization in budding yeast. *Nature*.

➔ TAP-Tag and expression studies & GFP-Tag and localization studies

2002

Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*.
& Gavin, A. C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* .

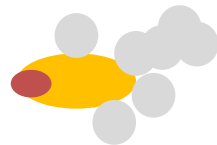
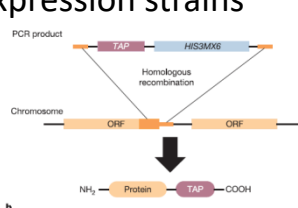
➔ Protein–protein interaction maps can be obtained by MS; the yeast cell is organized into protein complexes.

2006

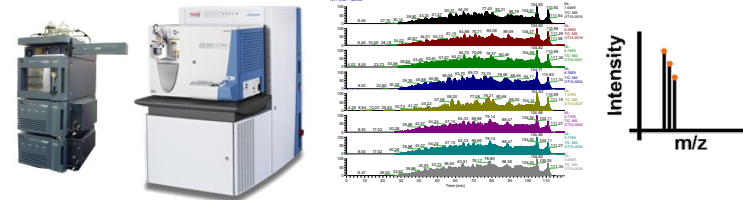
Krogan NJ, *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*.

➔ TAP-Tag and Protein-Protein Interaction

Collection of tagged “bait”
expression strains



Multiple runs of “shotgun” MS
& SDS-PAGE with MS on individual proteins



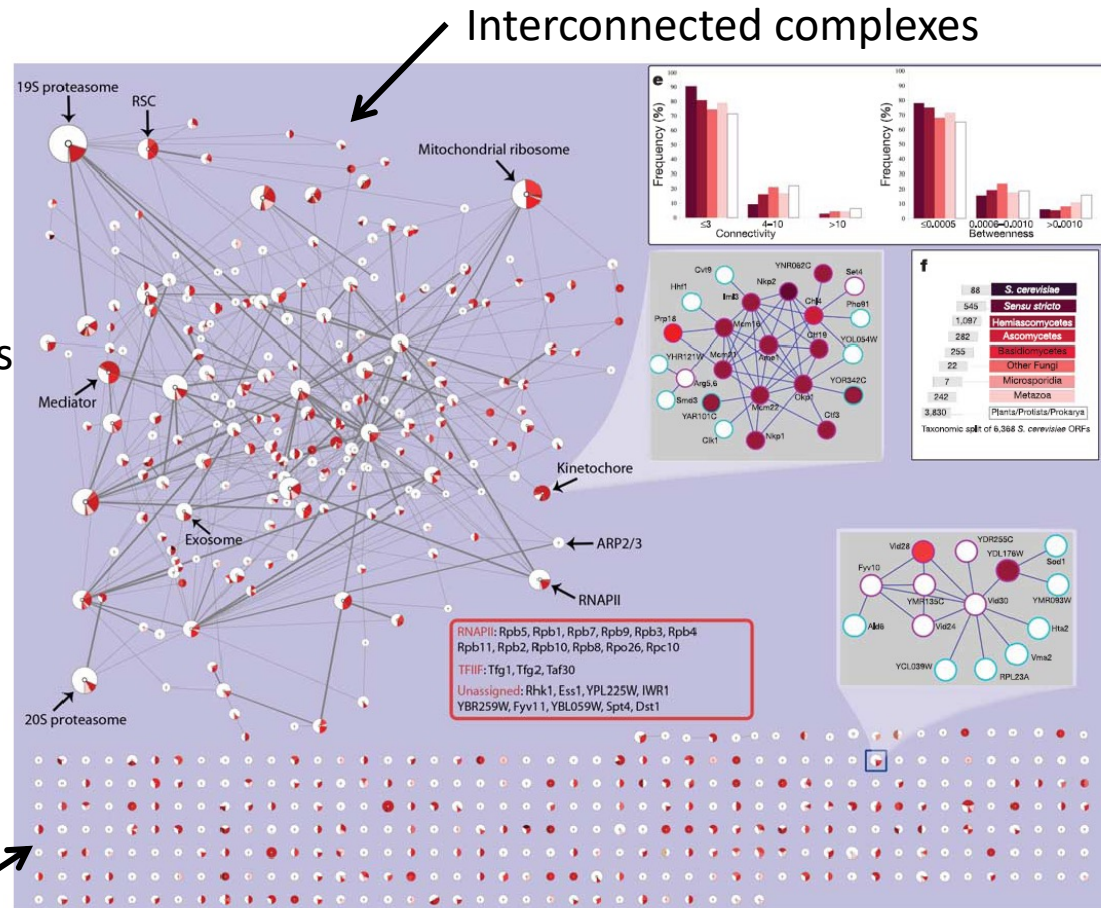
Krogan et al. observed 7,123 protein–protein interactions:

Important aspects:

- Tagged the native genes and did not overexpress the fusion proteins
- Could immediately validate partners (reciprocal purification in data set)
- Complementary MS techniques, deeper coverage of complexes
- Authors state, “...rigorous computational procedures to assign confidence values to our predictions...”

Cellular proteins are organized into complexes

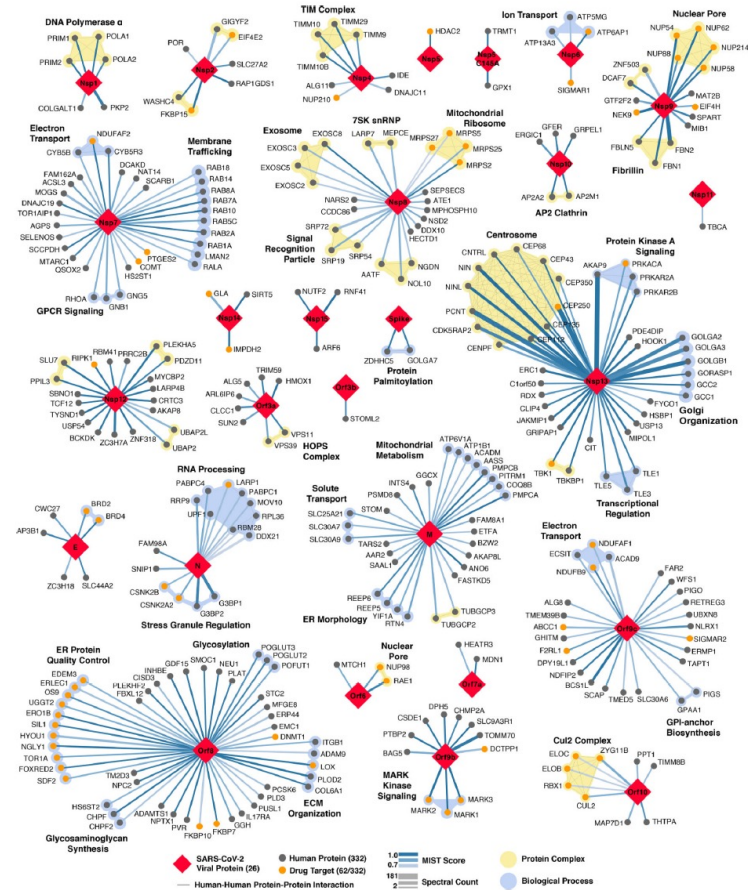
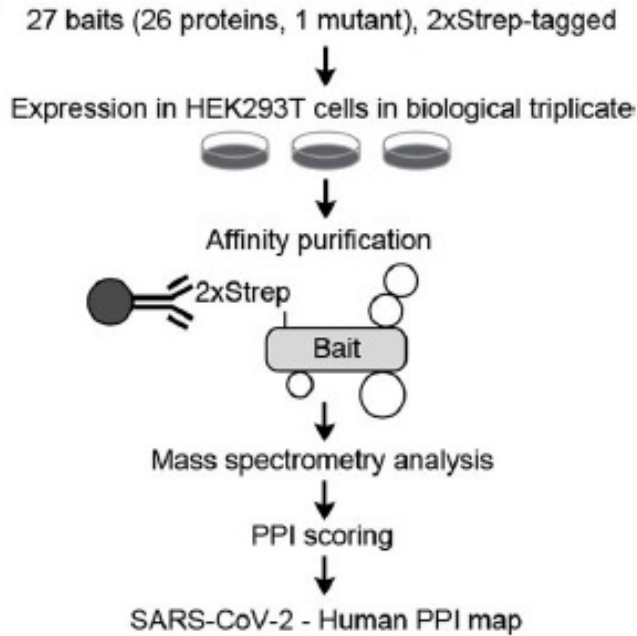
- 4,562 tagged proteins
- 2,357 successful purifications
- Identified 4,087 interacting proteins
~72 % proteome
- Majority of the yeast proteome is organized into complexes
- Many complexes are conserved in other species



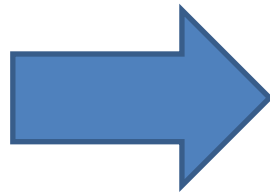
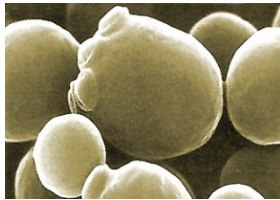
Complexes with little or no interconnectivity

Krogan NJ, et al. *Nature*. 2006

A SARS-CoV-2 Protein Interaction Map Reveals Targets for Drug-Repurposing



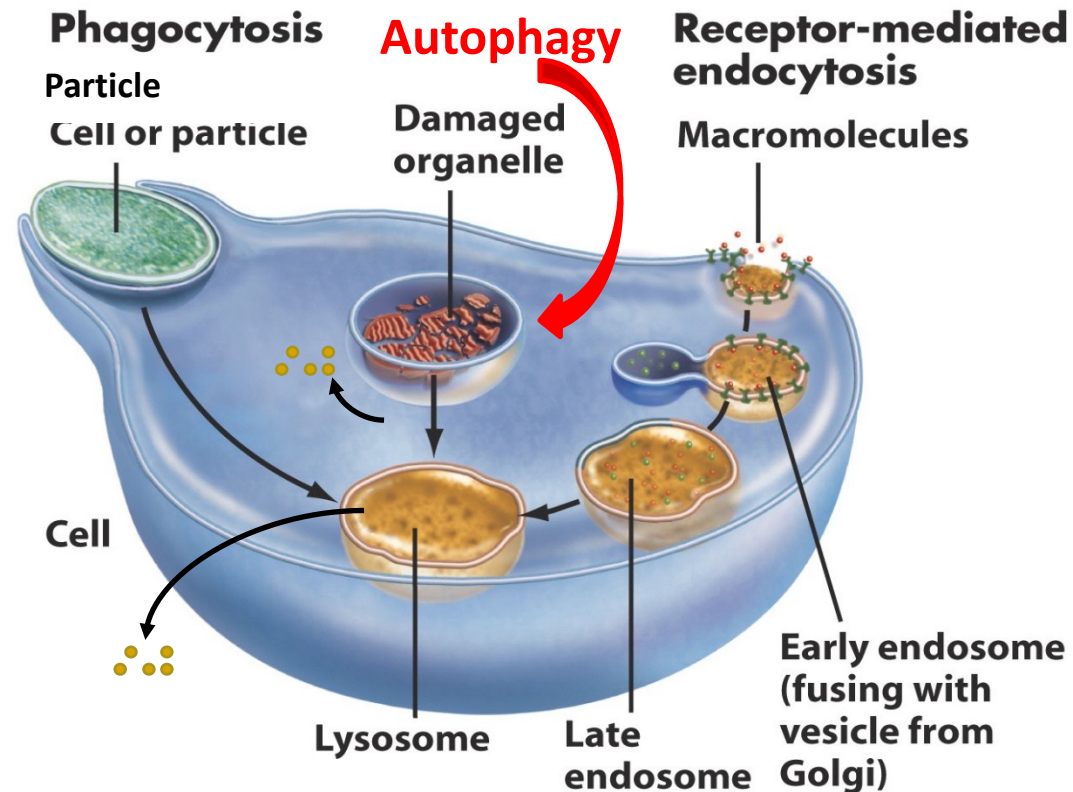
How do we learn more about the organization of the human proteome?



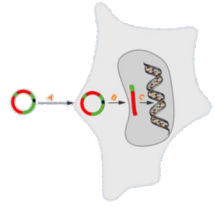
ARTICLES

Network organization of the human autophagy system

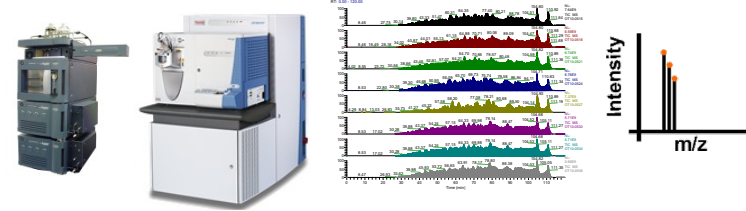
Christian Behrends¹, Mathew E. Sowa¹, Steven P. Gygi² & J. Wade Harper¹



Transfect tagged "bait"



Multiple runs of "shotgun" LC-MS/MS



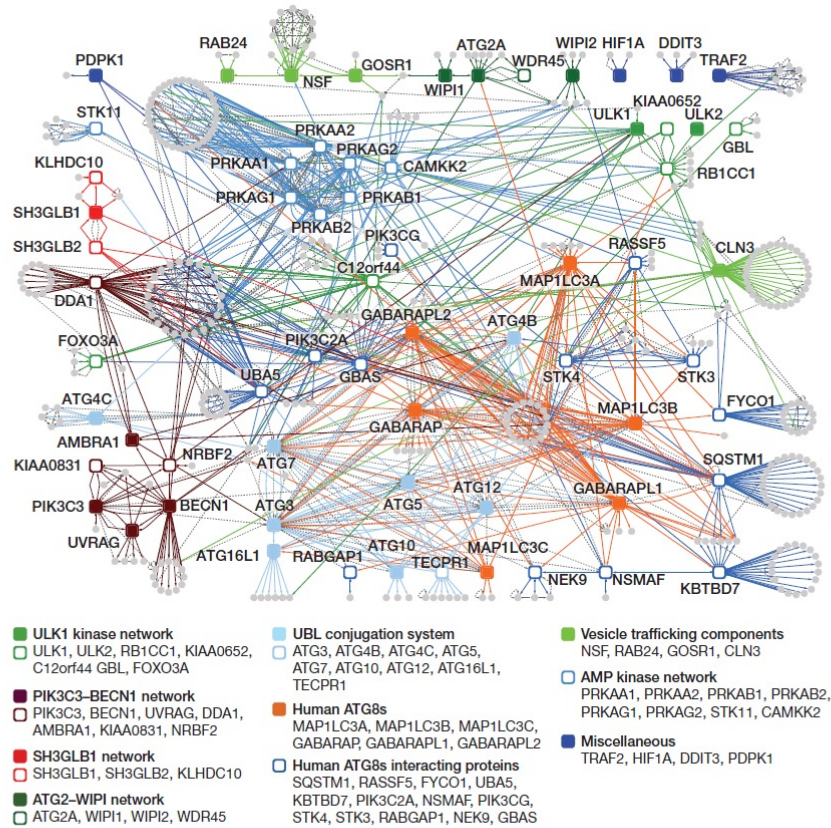
~65 bait proteins
LC-MS/MS identifies
2,553 proteins

Data analysis to sort out real
interaction from background

Authors use CompPASS
to identify High-Confidence
Interacting Proteins (HCIP)

763 HCIPs identified that compose
The Autophagy Interaction Network

Autophagy Interaction Network



Behreands et al, Nature 2010

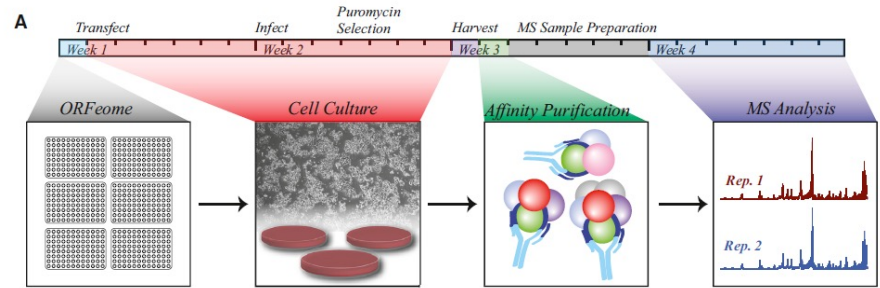
Figure 1 | Overview of the autophagy interaction network (AIN). HCIPs within the autophagy network are shown for 32 primary baits (filled squares) and 33 secondary baits (open squares). Subnetworks are colour-coded. Interacting proteins are indicated by grey circles.

BioPlex (Biophysical Interactions of ORFeome-derived complexes)

~25% of human genes used as baits

5,891 IP-MS experiments

56,553 interactions from 10,961 proteins



<http://wren.hms.harvard.edu/bioplex/>

The BioPlex Network: A Systematic Exploration of the Human Interactome

Edward L. Huttlin,¹ Lily Ting,¹ Raphael J. Bruckner,¹ Fana Gebreab,¹ Melanie P. Gygi,¹ John Szpyt,¹ Stanley Tam,¹

BioPlex 1.0 Huttlin et al, *Cell*. 2015, PMID: 26186194

Architecture of the human interactome defines protein communities and disease networks

Edward L. Huttlin¹, Raphael J. Bruckner¹, Joao A. Paulo¹, Joe R. Cannon¹, Lily Ting¹, Kurt Baltier¹, Greg Colby¹, Fana Gebreab¹,

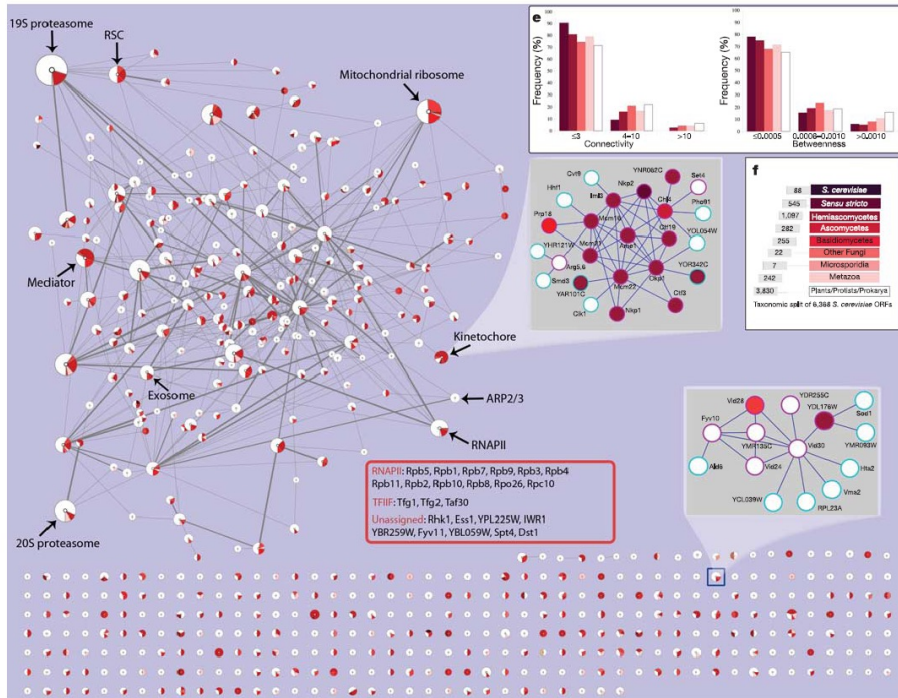
BioPlex 2.0 Huttlin et al, *Nature*. 2017 PMID: 28514442

BioPlex 3.0 Huttlin et al, *Cell* 2021 PMID: 33961781

This dataset contains ~120,000 interactions detected in HEK293T cells using 10128 baits.
<https://bioplex.hms.harvard.edu/interactions.php>.

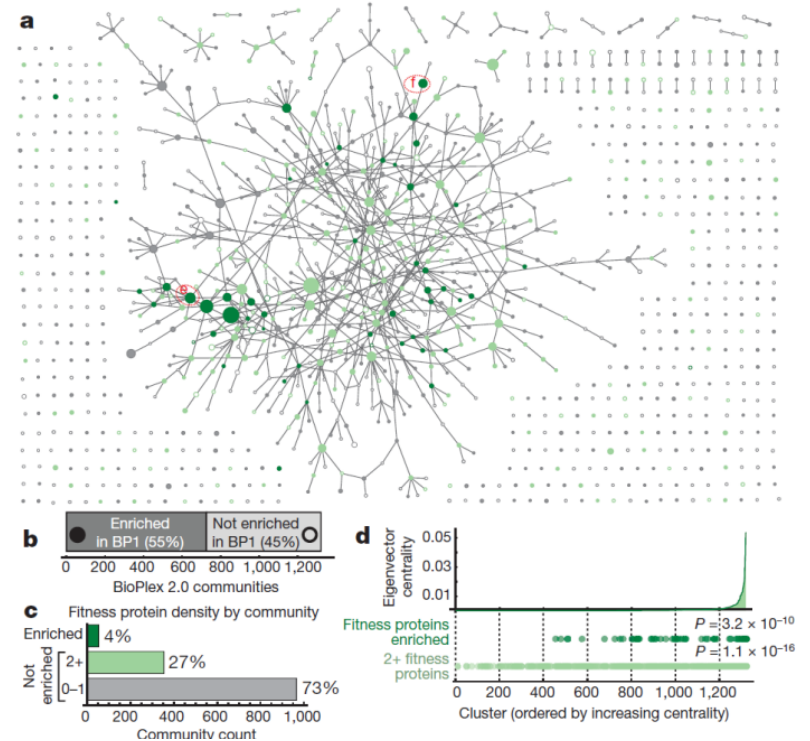
Cellular proteins are organized into complexes and this proteome organization is conserved

Yeast: Interaction Network of Complexes



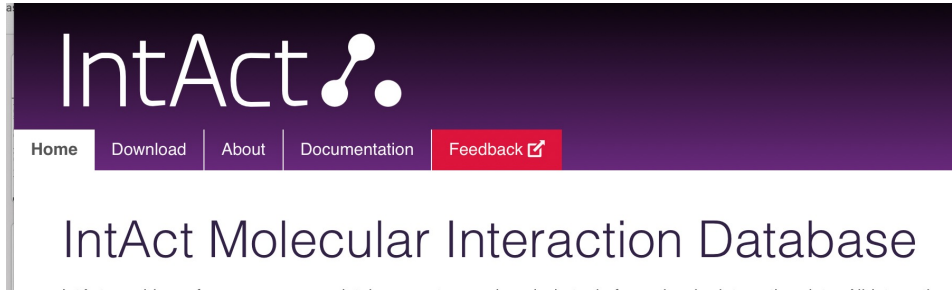
Krogan NJ, et al. *Nature*. 2006 PMID: 16554755

Human: Protein Complex "Communities"

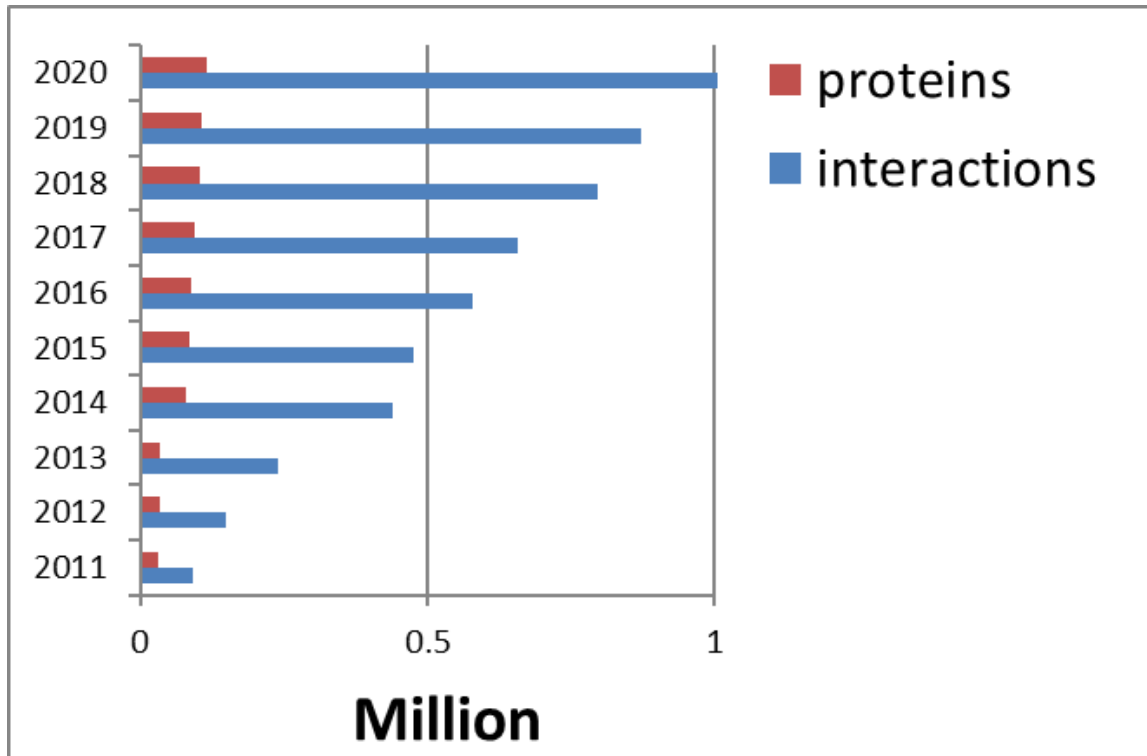


Huttlin et al, *Nature*. 2017 PMID: 28514442

Protein-Protein Interaction Databases



<https://www.ebi.ac.uk/intact/home>



2023 📊 Data Content

- Interactors: 118,924
- Interactions: 1,194,594

↑
+ **39,393 interactions**
+ **903 proteins**

2022

- Interactors: 118,213
- Interactions: 1,155,201

Proteomics & Protein-Protein Interactions

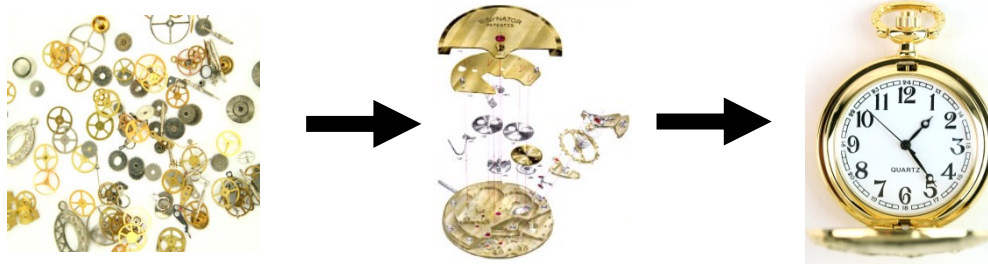
Overview

- **Techniques & Technologies**
 - Mass Spectrometry
 - Protein-Protein Interactions
 - Quantitative Proteomics
- **Applications**
 - Representative Studies
- **Putting it all together....**
 - Databases & Pathways

Protein interaction networks:

Some of the many important aspects:

- Parts List
- Organization and assembly
- Biological function can be inferred

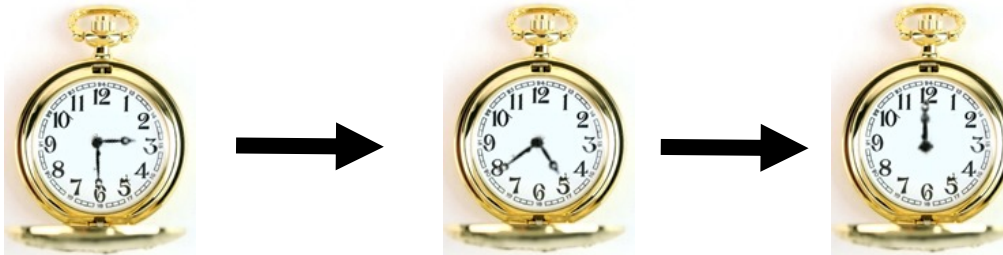


However:

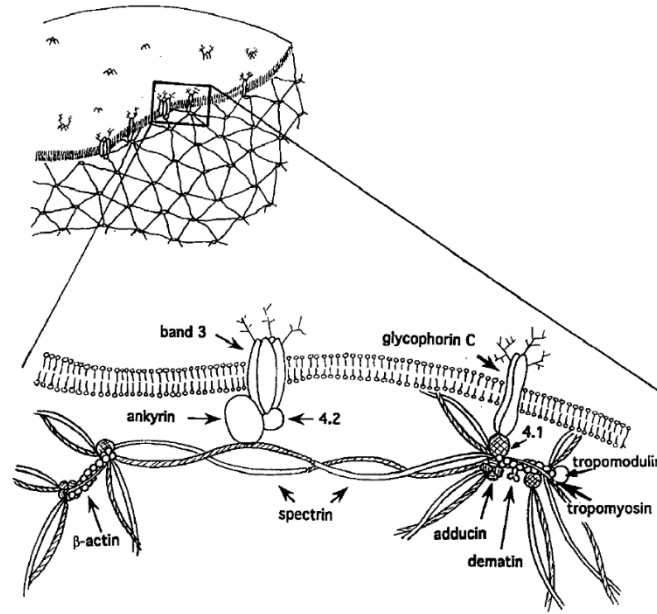
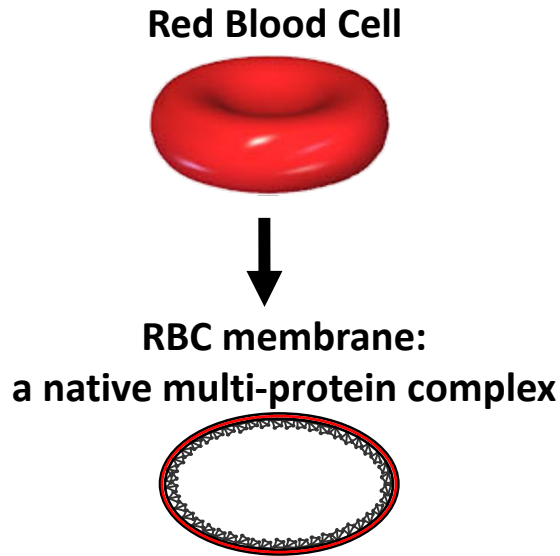
- Interaction data is largely static

Next Step:

- **How do protein interaction networks change over time?**

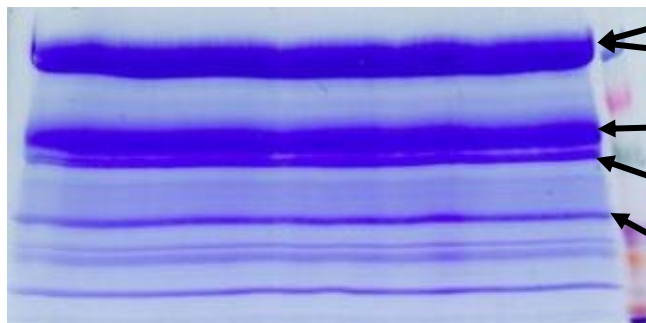


MS Data is not inherently quantitative



RBC membrane proteome
Coomassie Stained
SDS-PAGE (250 ug Protein)
~16 bands

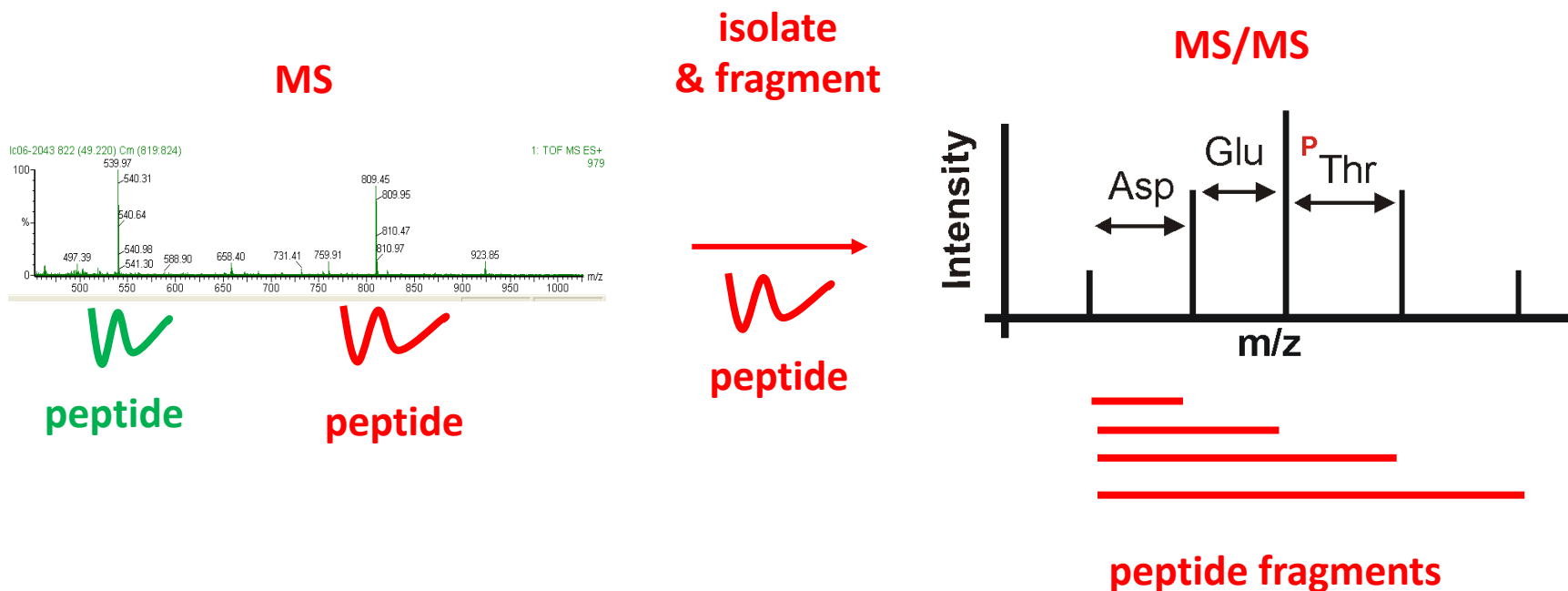
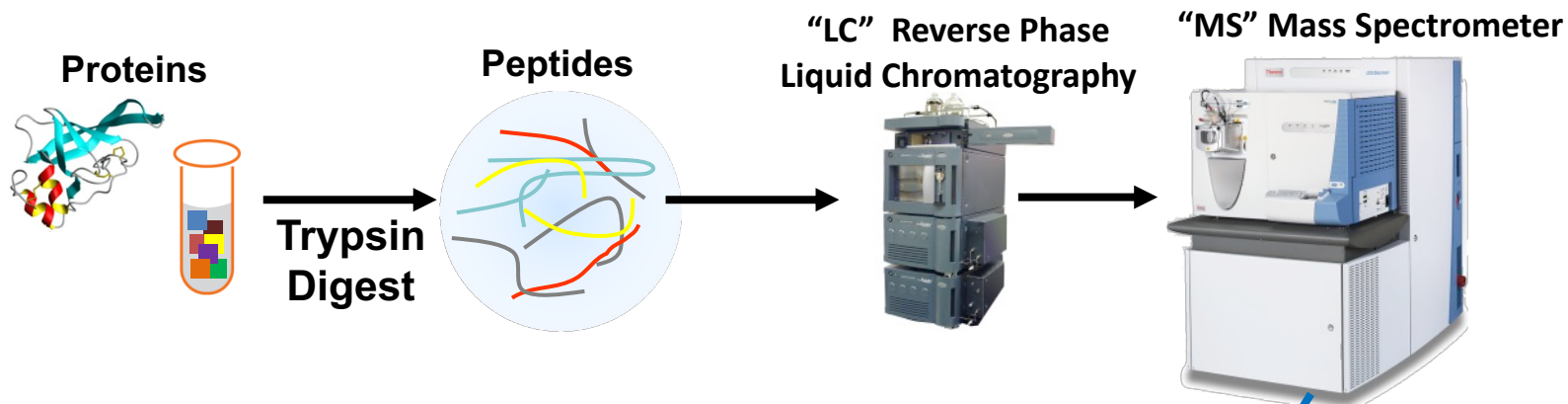
RBC membrane proteome
Shotgun Proteomics
1ug Peptides (242 Proteins)



peptides (unique)

Spectrin α	352 (291)	Spectrin alpha chain, erythrocyte OS=Homo sapiens GN=SPTA1 PE=1 SV=5
Spectrin β	291 (233)	Spectrin beta chain, erythrocyte OS=Homo sapiens GN=SPTB PE=1 SV=5
	172 (134)	Ankyrin-1 OS=Homo sapiens GN=ANK1 PE=1 SV=3
Band 3	57 (46)	Band 3 anion transport protein OS=Homo sapiens GN=SLC4A1 PE=1 SV=3
	52 (39)	Erythrocyte membrane protein band 4.2 OS=Homo sapiens GN=EPB42 PE=1 SV=5
Band 4.1	43 (34)	Actin, cytoplasmic 1 OS=Homo sapiens GN=ACTB PE=1 SV=1
	30 (20)	Actin, alpha cardiac muscle 1 OS=Homo sapiens GN=ACTC1 PE=1 SV=1
	22 (9)	Beta-actin-like protein 2 OS=Homo sapiens GN=ACTBL2 PE=1 SV=2
β -actin	28 (6)	POTE ankyrin domain family member J OS=Homo sapiens GN=POTEJ PE=3 SV=1
	68 (49)	Protein 4.1 OS=Homo sapiens GN=EPB41 PE=1 SV=4

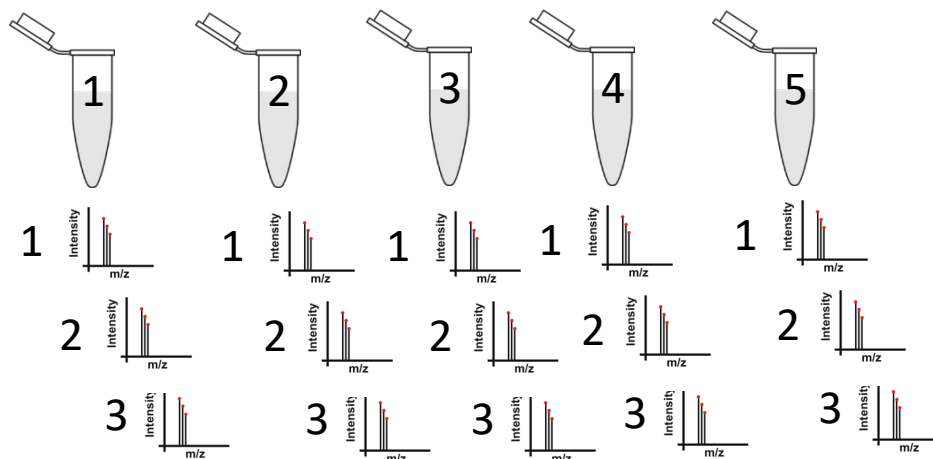
Typical work flow for LC-MS "shotgun proteomics"



Multiple Techniques Enable Quantitative Proteomics

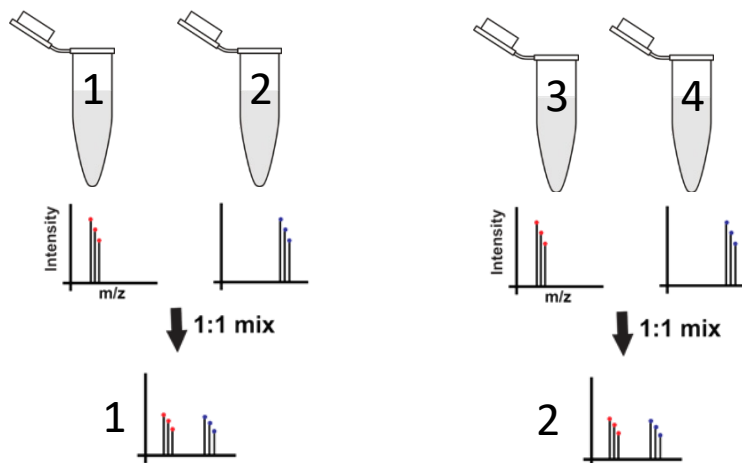
Label Free

- many, many replicates
- indirect quant



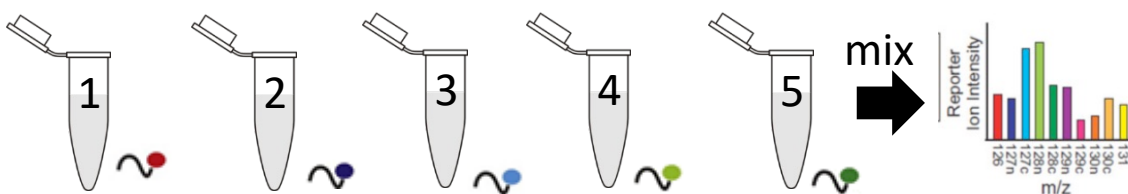
“Metabolic” Labeling

- fewer replicates
- multiplex
- direct quant



Barcoding

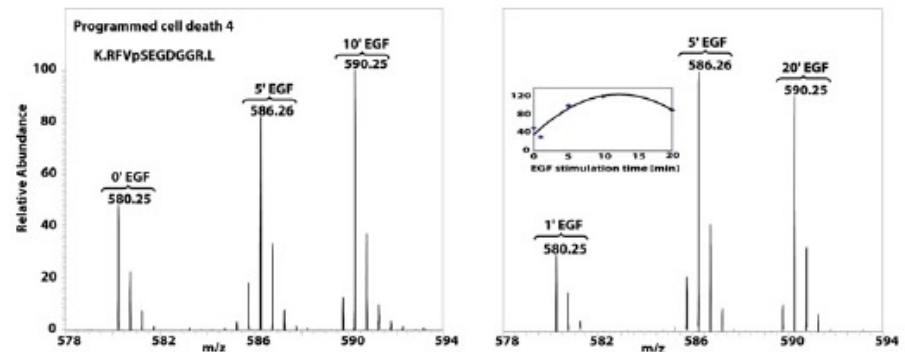
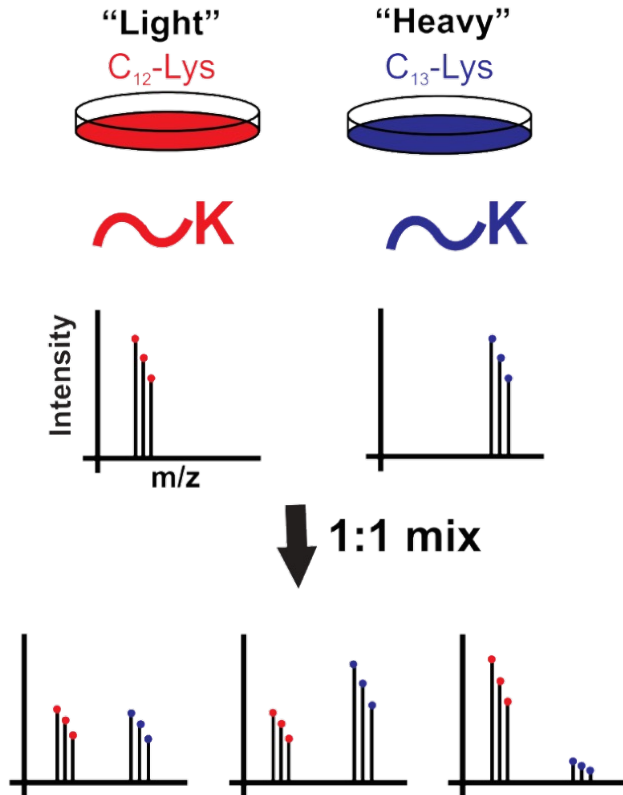
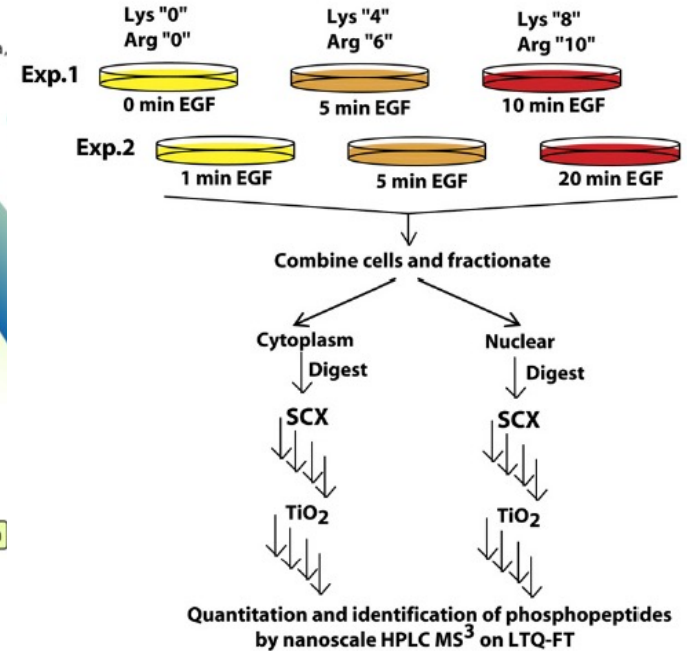
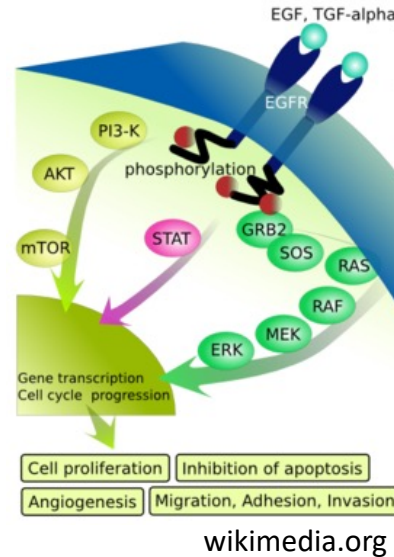
- increased multiplex
- direct quant



"Metabolic" Labeling: proteins and peptides heavy labeled in cell culture

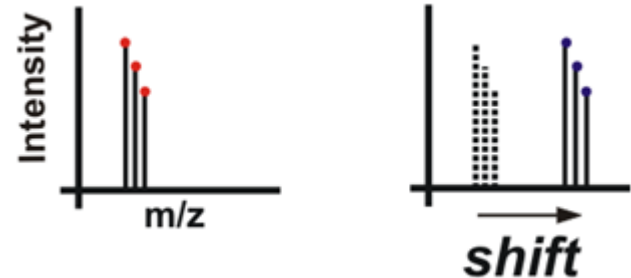
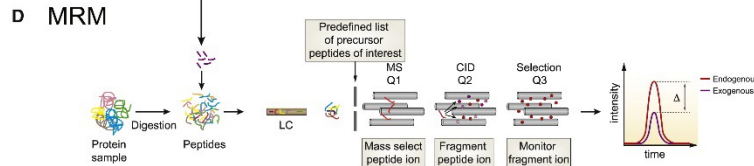
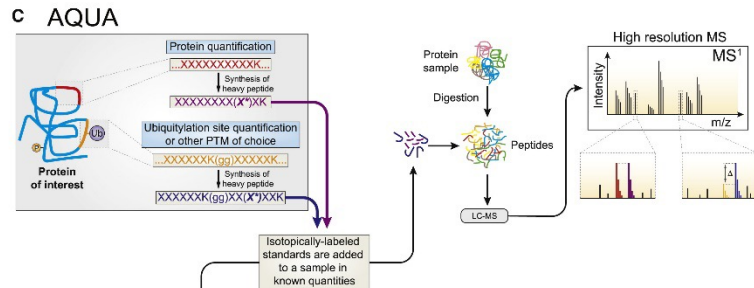
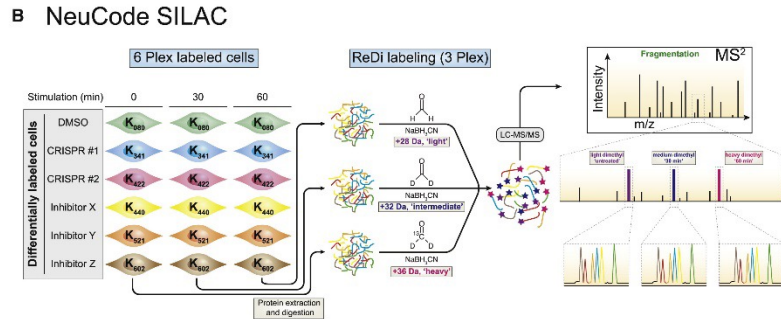
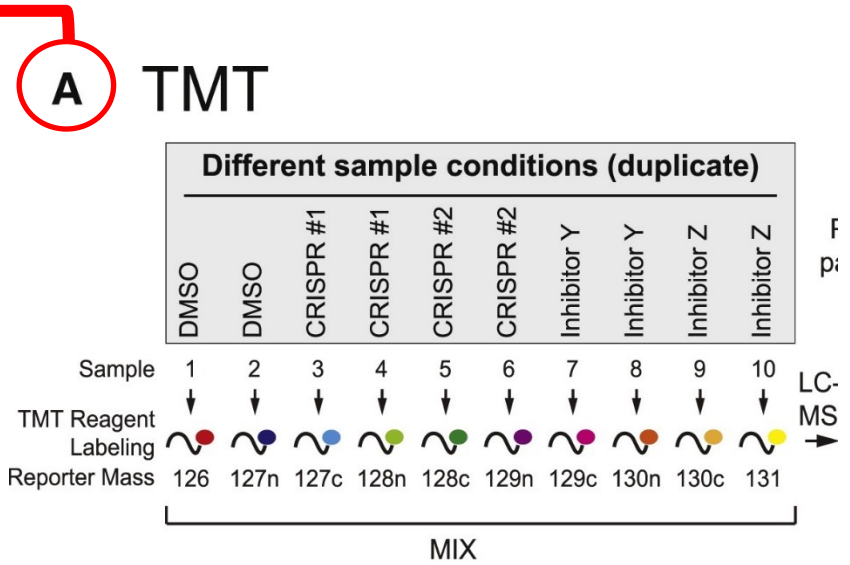
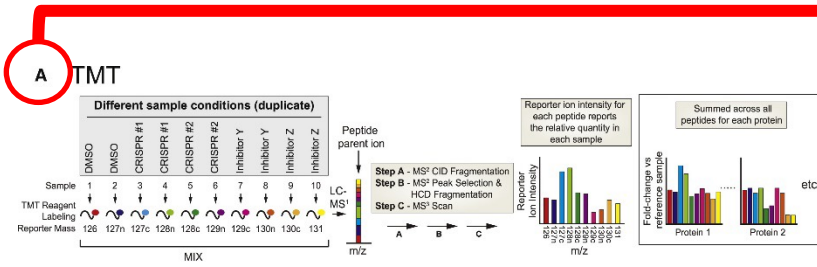
- SILAC approach enables dynamic analysis

- Olsen, et al. *Cell*, 2006: Phosphorylation dynamics after EGF stimulation



MS spectra triplets

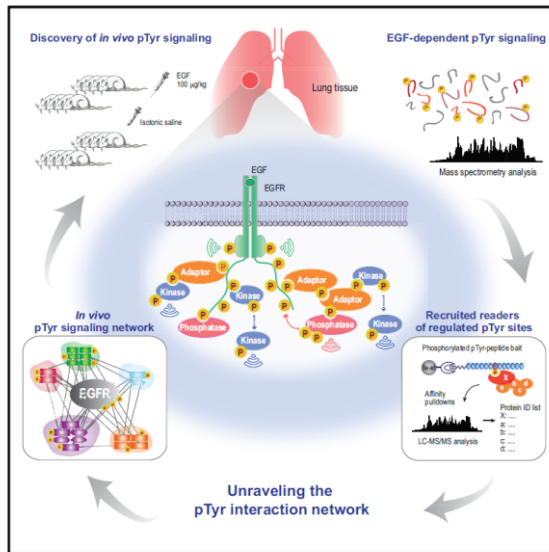
Barcoding: Heavy labels can be used for “barcoding” proteomes



Combinations of technologies are enabling quantitative proteomics at increasing complexity and with greater biological insight

Oncogenic Mutations Rewire Signaling Pathways by Switching Protein Recruitment to Phosphotyrosine Sites

Graphical Abstract



Authors

Alicia Lundby, Giulia Franciosa, Kristina B. Emdal, ..., Guillermo Montoya, Lars J. Jensen, Jesper V. Olsen

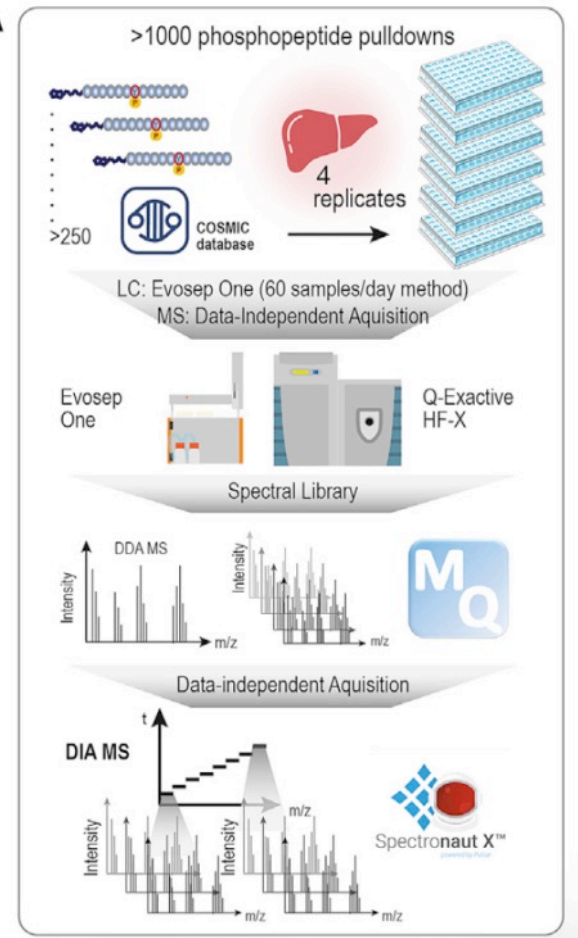
Correspondence

alicia.lundby@sund.ku.dk (A.L.),
jesper.olsen@cpr.ku.dk (J.V.O.)

In Brief

Lundby et al. used mass-spectrometry-based proteomics to analyze *in vivo* EGF signaling in lung tissue. They analyzed changes in the EGF-regulated phosphotyrosine interactome and found that cancer mutations in the vicinity of phosphotyrosine sites induce molecular switches that alter protein signaling networks.

A

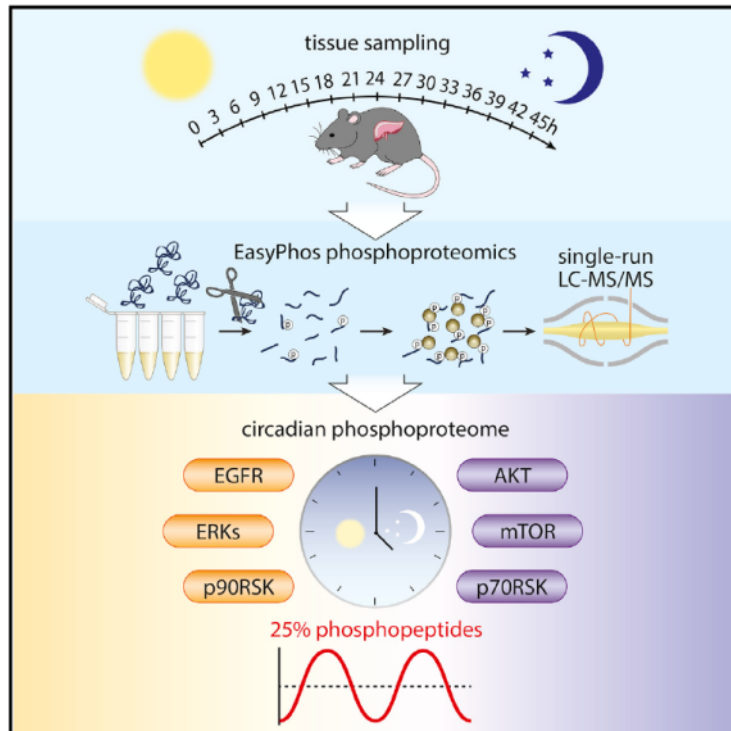


Combinations of technologies are enabling quantitative proteomics at increasing complexity and with greater biological insight

Cell Metabolism

Phosphorylation Is a Central Mechanism for Circadian Control of Metabolism and Physiology

Graphical Abstract



Authors

Maria S. Robles, Sean J. Humphrey,
Matthias Mann

Correspondence

mmann@biochem.mpg.de

In Brief

Robles et al. profile the global in vivo circadian phosphoproteome of the mouse liver and reveal that 25% of the quantified phosphopeptides oscillate with very high amplitudes compared to the transcriptome and proteome. Phosphorylation-dependent tuning of signaling pathways is a key circadian mechanism for metabolic regulation.

Proteomics & Protein-Protein Interactions

Overview

- **Techniques & Technologies**
 - Mass Spectrometry
 - Protein-Protein Interactions
 - Quantitative Proteomics
- **Applications**
 - Representative Studies
- **Putting it all together....**
 - Databases & Pathways

DNA → RNA → PROTEIN



2001

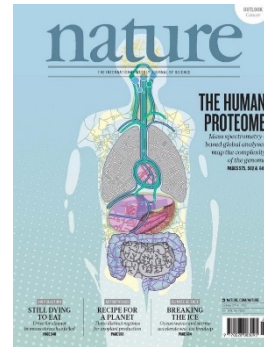
The Sequence of the Human Genome

J. Craig Venter,^{1*} Mark D. Adams,¹ Eugene W. Myers,¹ Peter W. Li,¹ Richard J. Mural,¹ Granger G. Sutton,¹ Hamilton O. Smith,¹ Mark Yandell,¹ Cheryl A. Evans,¹ Robert A. Holt,¹

articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium¹



2014

ARTICLE

doi:10.1038/nature13319

Mass-spectrometry-based draft of the human proteome

Mathias Wilhelm^{1,2*}, Judith Schlegel^{3*}, Hannes Hahne^{4*}, Amin Moghaddas Gholami^{4*}, Marcus Lieberenz², Mikhail M. Savitski⁵, Emanuel Ziegler⁵, Lars Butzmann⁵, Siegfried Gesualdo⁵, Harald Marx¹, Toby Mathison¹, Simone Lemeer⁵, Karsten Schmittaumann⁵, Ulf Reimer⁵, Holger Wenschuh⁵, Martin Mollenhauer⁵, Julia Stotta-Husperina⁵, Joos Hendrik Boese⁵, Marcus Bantscheff⁶, Anja Gerstmaier², Franz Faerber² & Bernhard Kuster^{1,6}

ARTICLE

doi:10.1038/nature13302

A draft map of the human proteome

Min-Sik Kim^{1,2}, Sneha M. Pinto³, Derese Getnet^{1,4}, Raja Sekhar Nirujogi³, Srikanth S. Manda³, Raghobama Chaerkady^{1,2}, Anil K. Madugundu³, Dhanashree S. Kelkar³, Ruth Isserlin⁵, Shobhit Jain³, Joji K. Thomas³, Babylakshmi Muthusamy³, Pamela Leal-Rojas^{1,6}, Praveen Kumar³, Nandini A. Sahasrabudhe³, Lavanya Balakrishnan³, Jayshree Advani³, Bijesh George³, Santosh Renuse³, Lakshmi Dhevi N. Selvan³, Arun H. Patil³, Vishalakshi Nanjappa³, Aneesh Radhakrishnan³, Sumarjet Prasad³,

The Sequence of the Human Genome. PMID: 11181995

Initial sequencing and analysis of the human genome. PMID: 11237011

A draft map of the human proteome. PMID: 24870542

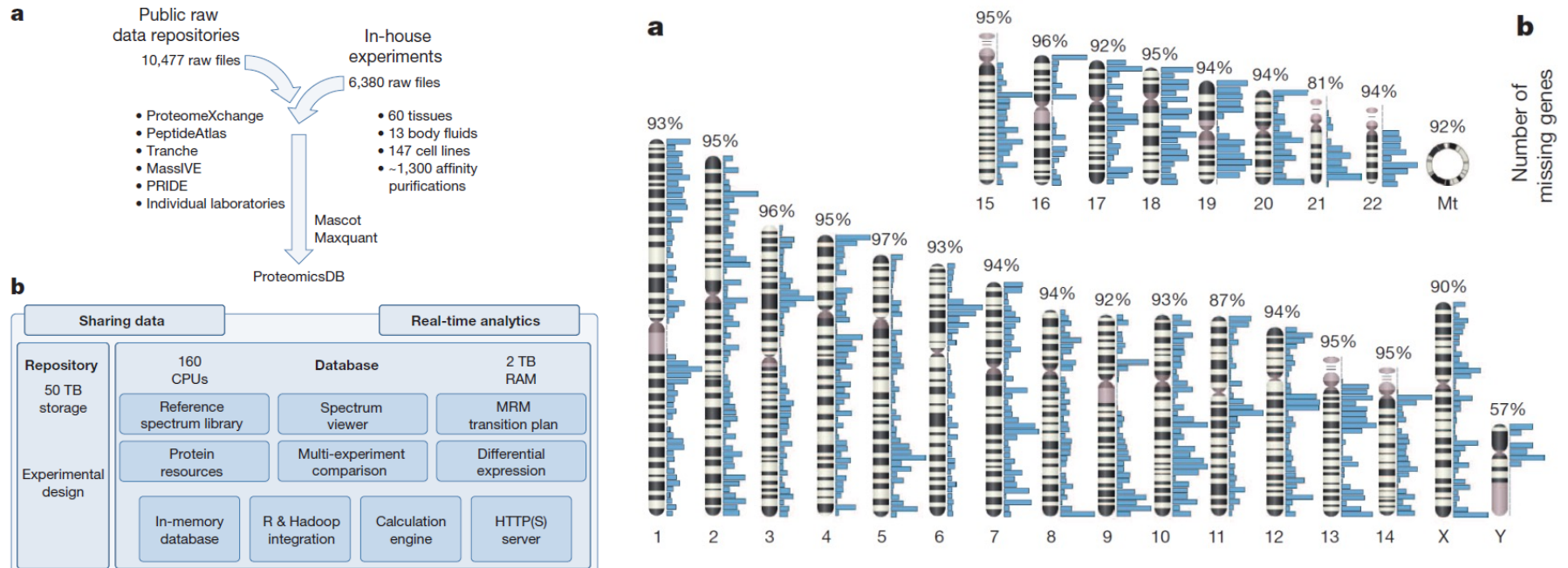
Mass-spectrometry-based draft of the human proteome. PMID: 24870543

Mass-spectrometry-based draft of the human proteome

Mathias Wilhelm^{1,2*}, Judith Schlegl^{2*}, Hannes Hahne^{1*}, Amin Moghaddas Gholami^{1*}, Marcus Lieberenz², Mikhail M. Savitski³, Emanuel Ziegler², Lars Butzmann², Siegfried Gessulat², Harald Marx¹, Toby Mathieson³, Simone Lemeer¹, Karsten Schnatbaum⁴, Ulf Reimer², Holger Wenschuh⁴, Martin Mollenhauer⁵, Julia Slotta-Huspenina⁵, Joos-Hendrik Boese², Marcus Bantscheff³, Anja Gerstmair², Franz Paerber² & Bernhard Kuster^{1,6}

- Large Assembly of new and existing data:
- ProteomicsDB, database designed for the real-time analysis of big data

<https://www.proteomicsdb.org>



Mass-spectrometry-based draft of the human proteome

Mathias Wilhelm^{1,2*}, Judith Schlegl^{2*}, Hannes Hahne^{1*}, Amin Moghaddas Gholami^{1*}, Marcus Lieberenz², Mikhail M. Savitski³, Emanuel Ziegler², Lars Butzmann², Siegfried Gessulat², Harald Marx¹, Toby Mathieson³, Simone Lemeer¹, Karsten Schnatbaum⁴, Ulf Reimer⁴, Holger Wenschuh⁴, Martin Mollenhauer⁵, Julia Slotta-Huspenina⁵, Joos-Hendrik Boese², Marcus Bantscheff³, Anja Gerstmair², Franz Faerber² & Bernhard Kuster^{1,6}

- Large Assembly of new and existing data:
- ProteomicsDB, database designed for the real-time analysis of big data

<https://www.proteomicsdb.org>



Wilhelm *et al.* carried out 6,380 LC-MS experiments (or runs):

How long would it take to get the same data?

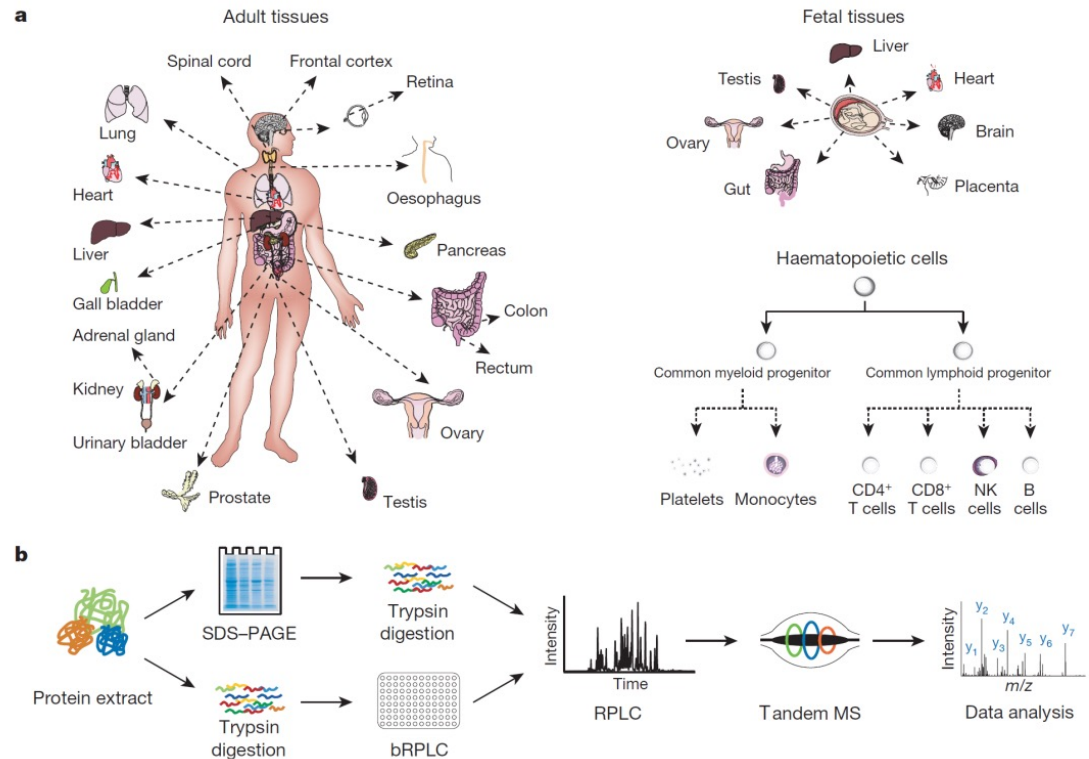
In 2001? ~61 years

In 2014? ~265 Days

A draft map of the human proteome

Min-Sik Kim^{1,2}, Sneha M. Pinto³, Derese Getnet^{1,4}, Raja Sekhar Nirujogi³, Srikanth S. Manda³, Raghothama Chaerkady^{1,2}, Anil K. Madugundu³, Dhanashree S. Kelkar³, Ruth Isserlin⁵, Shobhit Jain⁵, Joji K. Thomas³, Babylakshmi Muthusamy³, Pamela Leal-Rojas^{1,6}, Praveen Kumar³, Nandini A. Sahasrabudhe³, Lavanya Balakrishnan³, Jayshree Advani³, Bijesh George³, Santosh Renuse³, Lakshmi Dhevi N. Selvan³, Arun H. Patil³, Vishalakshi Nanjappa³, Aneesh Radhakrishnan³, Samarjeet Prasad¹,

- New, large collection of proteomics data
 - 30 histologically normal human samples
 - 17 adult tissues,
 - 7 fetal tissues
 - 6 purified primary haematopoietic cells
- 17,294 genes accounting for approximately 84% of the total annotated protein-coding genes in humans.



Proteomics Databases: Peptide depositories



ISB Home

PeptideAtlas

PEPTIDEATLAS HOME

Seattle Proteome Center

PeptideAtlas Builds – Bulk Downloads

<http://www.peptideatlas.org/builds/>

TaxID	Date	Number of Samples	Peptide Inclusion Cutoff	Number of Peptide-Spectrum Matches (PSMs)	Number of Distinct Peptides	Reference Database	Peptide Sequences	Peptide CDS Coordinates	Peptide CDS and Chromosomal Coordinates	Database Tables
9606	Mar 2015	1011	PSM FDR = 0.0002	133,638,335	1,025,698	Ensembl v78+UPSP+Trembl201412+14IPI 3.87+cRAP+nextprotSNP	APD_Hs_all.fasta	prot_map	chrom_map	MYSQL.XML

Protein Identification Terminology used in PeptideAtlas

http://www.peptideatlas.org/docs/protein_ident_terms.php

- Each PeptideAtlas build is associated with a reference database usually a combination of several protein sequence databases (Swiss-Prot, IPI, Ensembl ...)
- From the reference database, any protein that contains any observed peptide is considered to be a member of the Atlas.
- It is easy to see that the entire list of proteins in an Atlas is going to be highly redundant. Thus, we label each Atlas protein using the terminology below.
 - The term "observed peptides" in this context refers to the set of peptides in the PeptideAtlas build.
 - These peptides are selected using a PSM (peptide spectrum match)

Proteomics Databases: Peptide depositories



HUMAN PROTEOME MAP

[Home](#)[Query](#)[Download](#)[FAQs](#)[Contact us](#)

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

ARTICLE

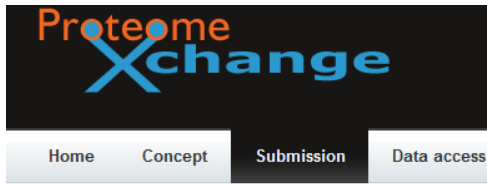
doi:10.1038/nature13302

A draft map of the human proteome

Min-Sik Kim^{1,2}, Sneha M. Pinto³, Derese Getnet^{1,4}, Raja Sekhar Nirujogi³, Srikanth S. Manda³, Raghothama Chaerkady^{1,2}, Anil K. Madugundu³, Dhanashree S. Kelkar³, Ruth Isserlin⁵, Shobhit Jain⁵, Joji K. Thomas³, Babyakshmi Muthusamy³, Pamela Leal-Rojas^{1,6}, Praveen Kumar³, Nandini A. Sahasrabudhe³, Lavanya Balakrishnan³, Jayshree Advani³, Bijesh George³, Santosh Renuse³, Lakshmi Dhevi N. Selvan³, Arun H. Patil³, Vishalakshi Nanjappa³, Aneesh Radhakrishnan³, Samarjeet Prasad¹,

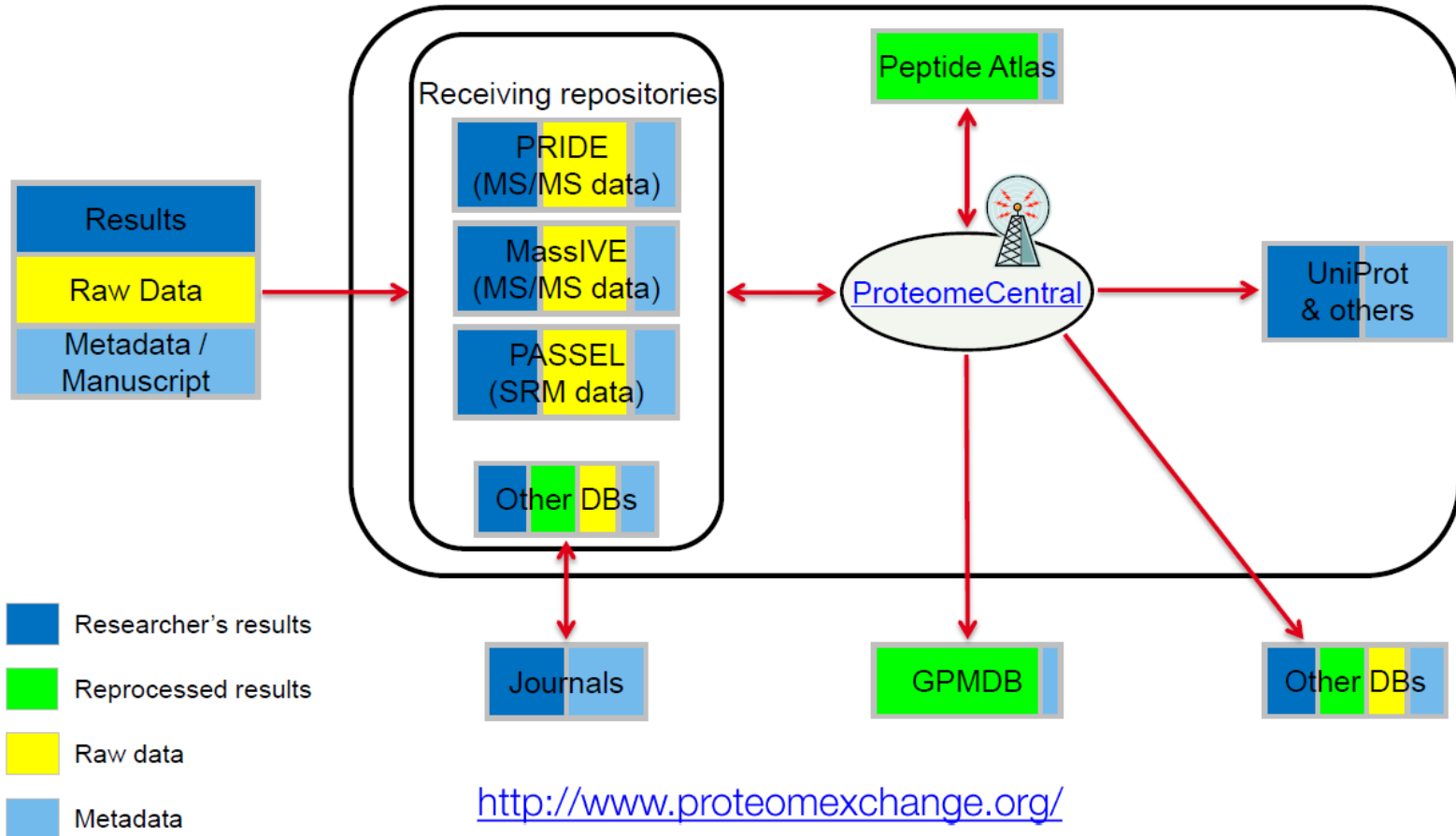
Kim & Akhilesh Pandey et al., *Nature*, 2014. PMID: 24870542

Proteomics Databases: Integrated Resources

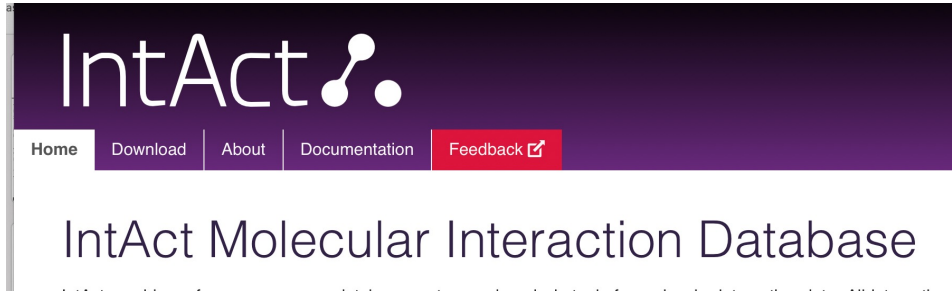


<http://www.proteomexchange.org/>

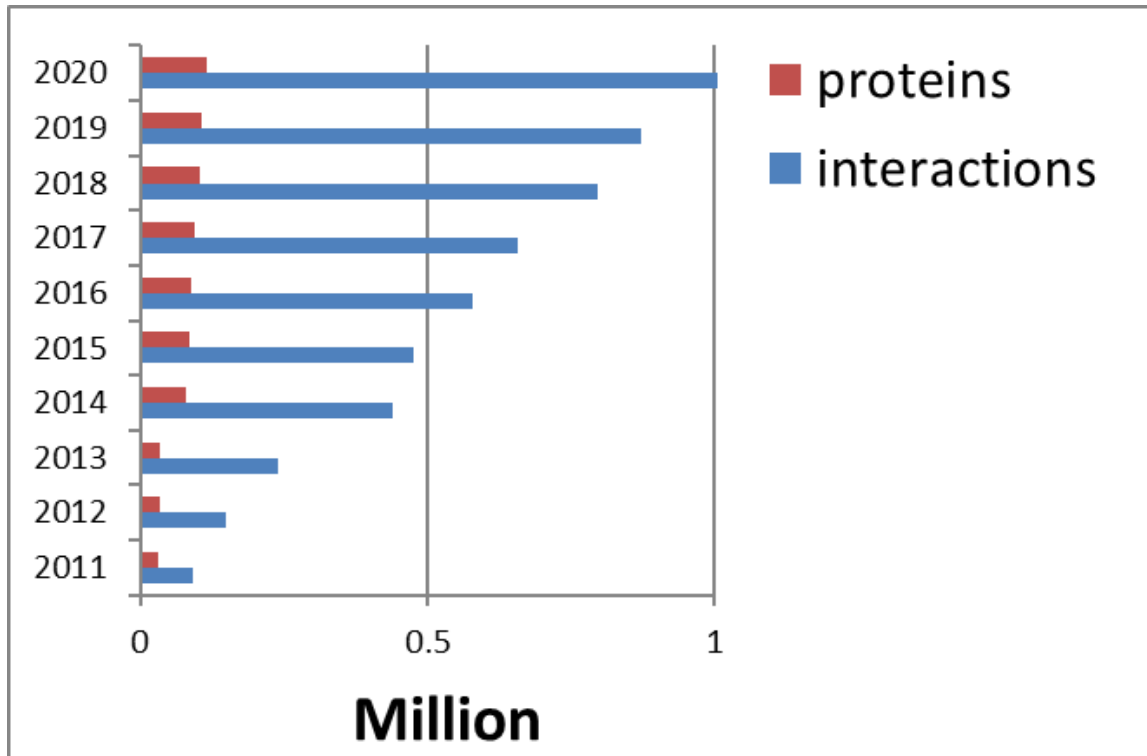
[ProteomeXchange \(PX\) consortium](#)



Protein-Protein Interaction Databases



<https://www.ebi.ac.uk/intact/home>



2023 📊 Data Content

- Interactors: 118,924
- Interactions: 1,194,594

↑
+ **39,393 interactions**
+ **903 proteins**

2022

- Interactors: 118,213
- Interactions: 1,155,201

Proteomics Databases: Integrated Resources Beyond Mass Spectrometry

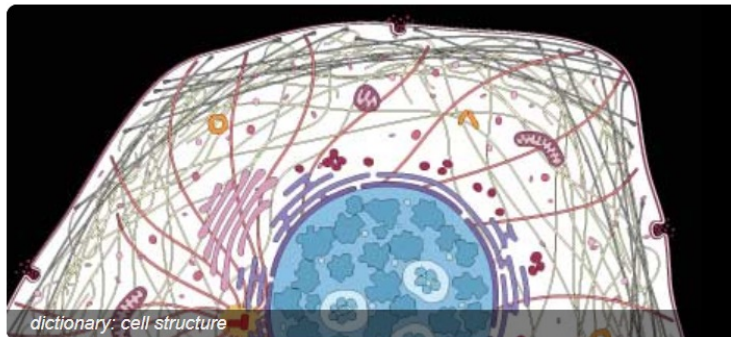
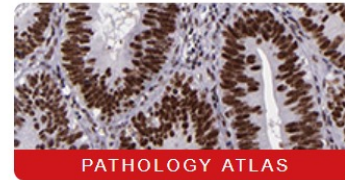
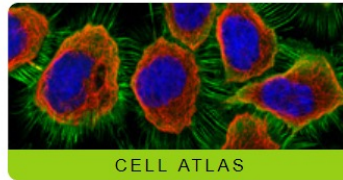
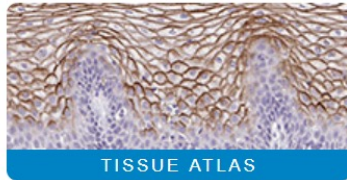
<http://www.proteinatlas.org/>

THE HUMAN PROTEIN ATLAS

[MENU](#) [HELP](#) [NEWS](#)

SEARCH[†]

[Fields »](#)
e.g. RBM3, insulin, CD36



Recent news

Thu, 6 Dec 2018
Integration of transcriptomics and antibody-based proteomics for exploration of proteins

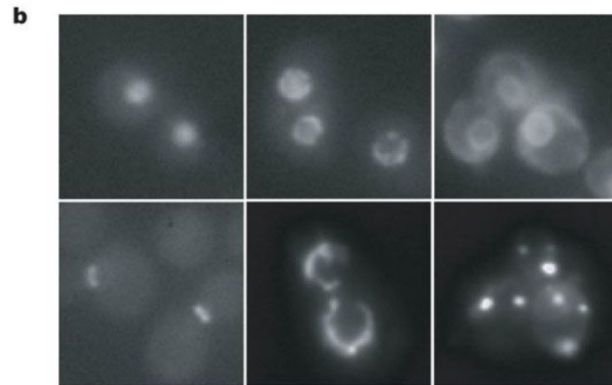
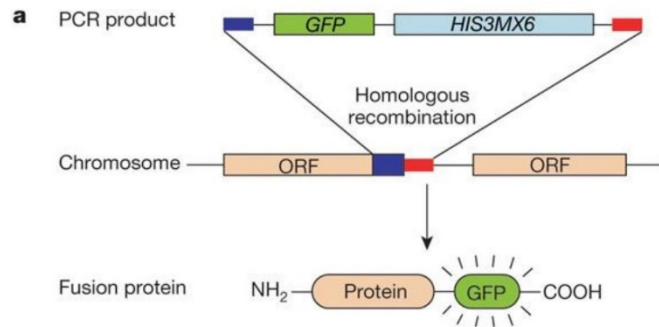
Mon, 26 Nov 2018
November: Prostate cancer awareness month

Thu, 15 Nov 2018
A version 18.1 release today with new Survival Scatter plots

[all news articles](#)

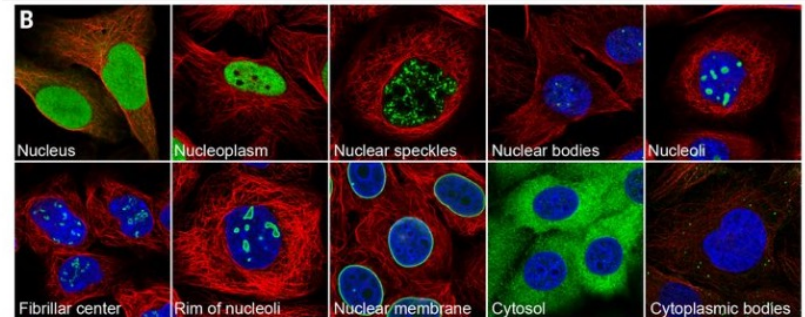
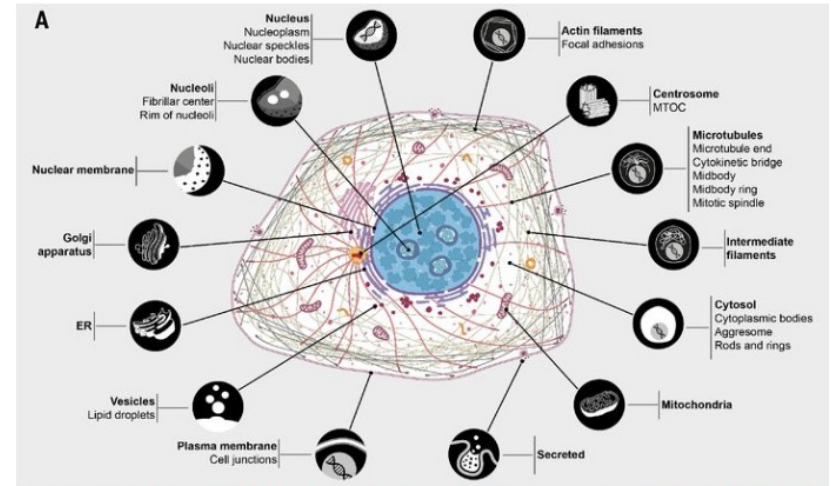
Proteomics Databases: Integrated Resources Beyond Mass Spectrometry

>4,000 GFP-Genes Fusions



Huh et al., Global analysis of protein localization in budding yeast. *Nature*. 2003
PubMed:14562095

>13,000 Antibodies



Thul PJ, et al. A subcellular map of the human proteome. *Science*. 2017. PubMed:28495876

Proteomics at single cell resolution in 2006 ?

nature

Vol 441|15 June 2006|doi:10.1038/nature04785

ARTICLES

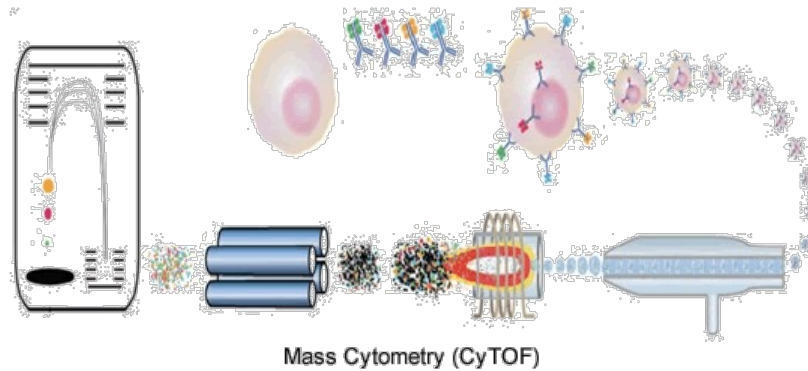
Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise

John R. S. Newman^{1,2}, Sina Ghaemmaghani^{1,2,†}, Jan Ihmels^{1,2}, David K. Breslow^{1,2}, Matthew Noble¹, Joseph L. DeRisi^{1,3} & Jonathan S. Weissman^{1,2}

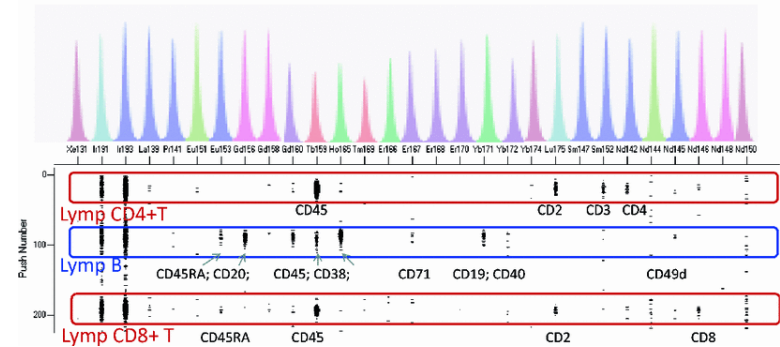
A major goal of biology is to provide a quantitative description of cellular behaviour. This task, however, has been hampered by the difficulty in measuring protein abundances and their variation. Here we present a strategy that pairs high-throughput flow cytometry and a library of GFP-tagged yeast strains to monitor rapidly and precisely protein levels at single-cell resolution. Bulk protein abundance measurements of >2,500 proteins in rich and minimal media provide a detailed view of the cellular response to these conditions, and capture many changes not observed by DNA microarray analyses. Our single-cell data argue that noise in protein expression is dominated by the stochastic production/ destruction of messenger RNAs. Beyond this global trend, there are dramatic protein-specific differences in noise that are strongly correlated with a protein's mode of transcription and its function. For example, proteins that respond to environmental changes are noisy whereas those involved in protein synthesis are quiet. Thus, these studies reveal a remarkable structure to biological noise and suggest that protein noise levels have been selected to reflect the costs and potential benefits of this variation.

Proteomics at single cell resolution

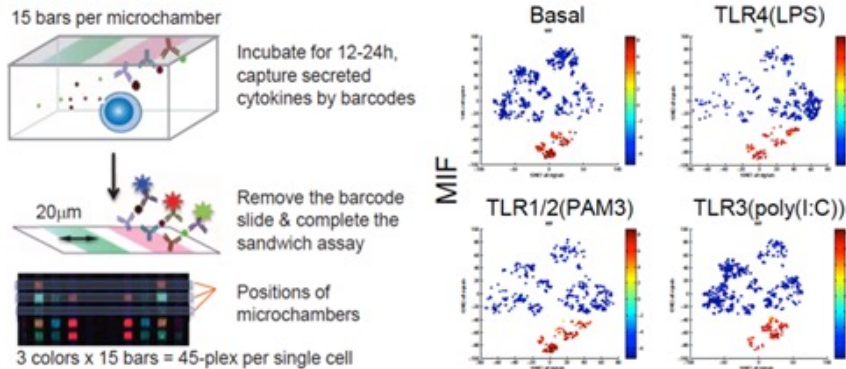
CytoTOF



CD2	¹⁷⁵ Lu	CD10	¹⁶⁸ Er	CD20	¹⁵⁶ Gd	CD38	¹⁶⁵ Ho	CD49d	¹⁴⁵ Nd	CD117	¹⁴⁷ Sm
CD3	¹⁵² Sm	CD11b	¹⁵⁴ Gd	CD31	¹⁴⁴ Nd	CD40	¹⁷² Yb	CD56	¹⁷⁸ Yb	HLA-DR	¹⁶⁰ Gd
CD4	¹⁴² Nd	CD13	¹⁶⁴ Er	CD33	¹⁴¹ Pr	CD44	¹⁵¹ Eu	CD64	¹⁴⁸ Nd		
CD7	¹³⁹ La	CD15	¹⁷⁰ Er	CD34	¹⁶⁹ Tm	CD45	¹⁵⁹ Tb	CD71	¹⁶⁷ Er		
CD8	¹⁴⁶ Nd	CD19	¹⁷¹ Yb	CD36	¹⁵⁰ Nd	CD45RA	¹⁵³ Eu	CD90	¹⁷⁴ Yb		



Single cell protein "capture" technology



Lu Y #, Xue Q #, Eisele MR, Sulistijo E, Brower K, Han L, Amir ED, Pe'er D, Miller-Jensen K *, and Fan R *, Highly multiplexed profiling of single-cell effector functions reveals deep functional heterogeneity in response to pathogenic ligands, *Proc. Natl. Acad. Sci. U.S.A.*, 112(7), 607-615 (2015).

Major challenges prevent complete proteome analysis

- **Proteomics is sample limited**
 - Recombinant DNA polymerases revolutionized genome sequencing by allowing for amplification of DNA samples
 - Proteomics has no “polymerase” or amplification method and must contend with natural abundancies
- **Mass spectrometry has limitations**
 - No mass spectrometer, or method, can yet provide full amino-acid resolution of a proteome

Transformative Opportunities for Single-Cell Proteomics

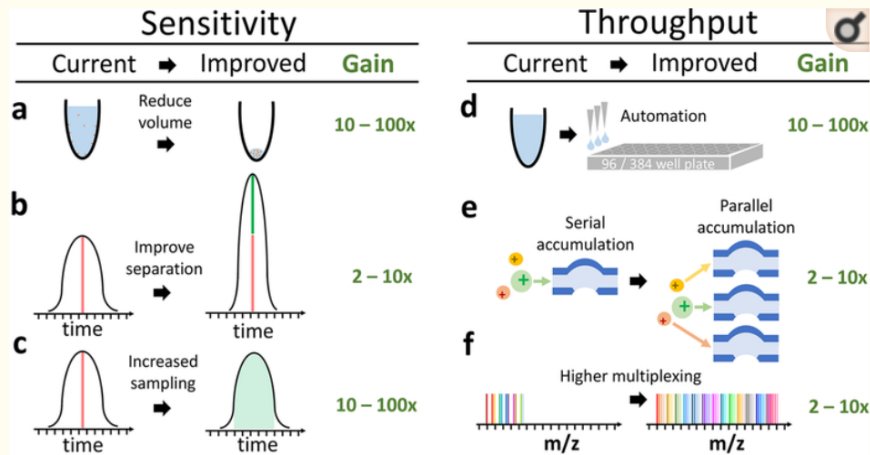
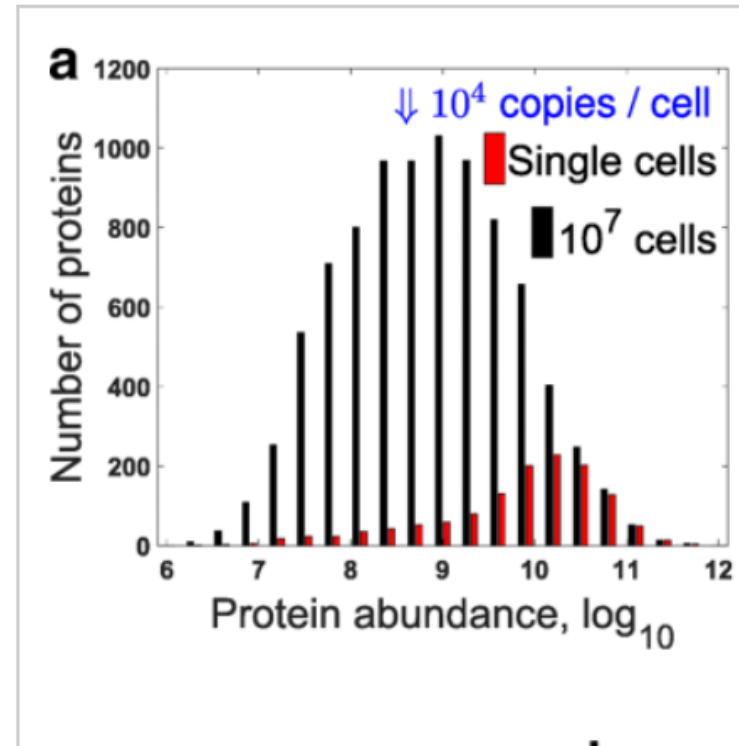
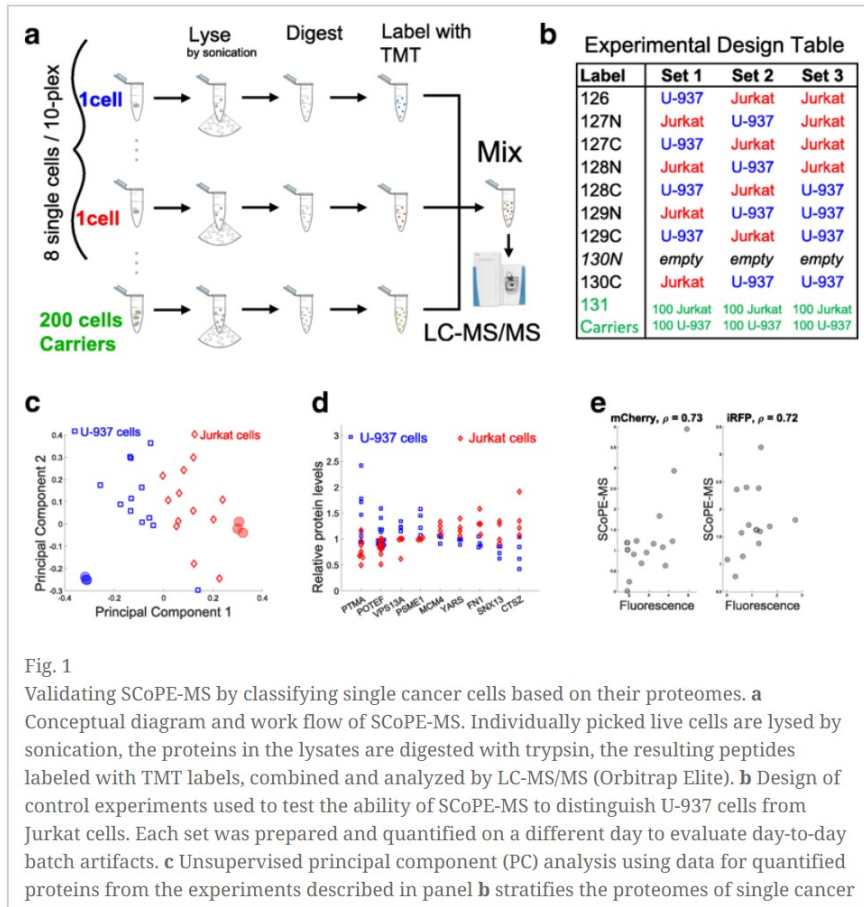


Figure 3.

Transformative opportunities for improving the quantification of single-cell proteomes. (a) Most bulk samples prepared for MS have volume of 10–100 μL .^{11,12,17} Reducing the volume for sample preparation to 1 to 2 nL ¹³ can significantly reduce protein losses from surface adsorption. (b) The sharper the separation peaks, the larger the fraction of the ions can be analyzed for a fixed sampling (injection) time. Sharper peaks can be achieved by reducing the bore of LC columns, using monolithic columns, PLOT columns,²⁷ or capillary electrophoresis.²⁵ (c) Typically elution peaks have a full width at the base of ~ 60 s and about 10–15 s at midheight, whereas ions for MS2 are sampled for mere milliseconds. These settings are typical for bulk proteomics and result in sampling $<1\%$ of the ions delivered to the instruments. Thus increasing the sampling time $100\times$ can substantially increase the ions analyzed by MS, the sensitivity, and the accuracy of quantification. While, the panel displays sampling during the apex of the peak, this cannot always be achieved for all ions. (d) Automated liquid handling and 96/384-well plates can increase the consistency of sample preparation, decrease volumes to the nanoliter range, and increase throughput. (e) Parallel accumulation and serial injection of ions can afford increased ion sampling without reducing throughput. (f) A larger number of barcodes will increase the number cellular proteomes quantified per run without reducing proteome coverage or ion sampling.

Achieving high chromatographic resolution and quantifying thousands of proteins requires an hour of LC–MS/MS time or more. Thus to quantify the proteomes of thousands of single cells within hours, we need to quantify many cells per LC–MS/MS run. Such multiplexing can be achieved by isobaric chemical barcoding.^{37,38} These barcodes are chemically identical but distinguishable by MS due to their different isotopic compositions.



Method | Open Access

SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation

Bogdan Budnik , Ezra Levy, Guillaume Harmange and Nikolai Slavov 

Genome Biology 2018 19:161

<https://doi.org/10.1186/s13059-018-1547-5> | © The Author(s). 2018

Received: 20 February 2018 | Accepted: 19 September 2018 | Published: 22 October 2018