Proteomics & Protein-Protein Interactions

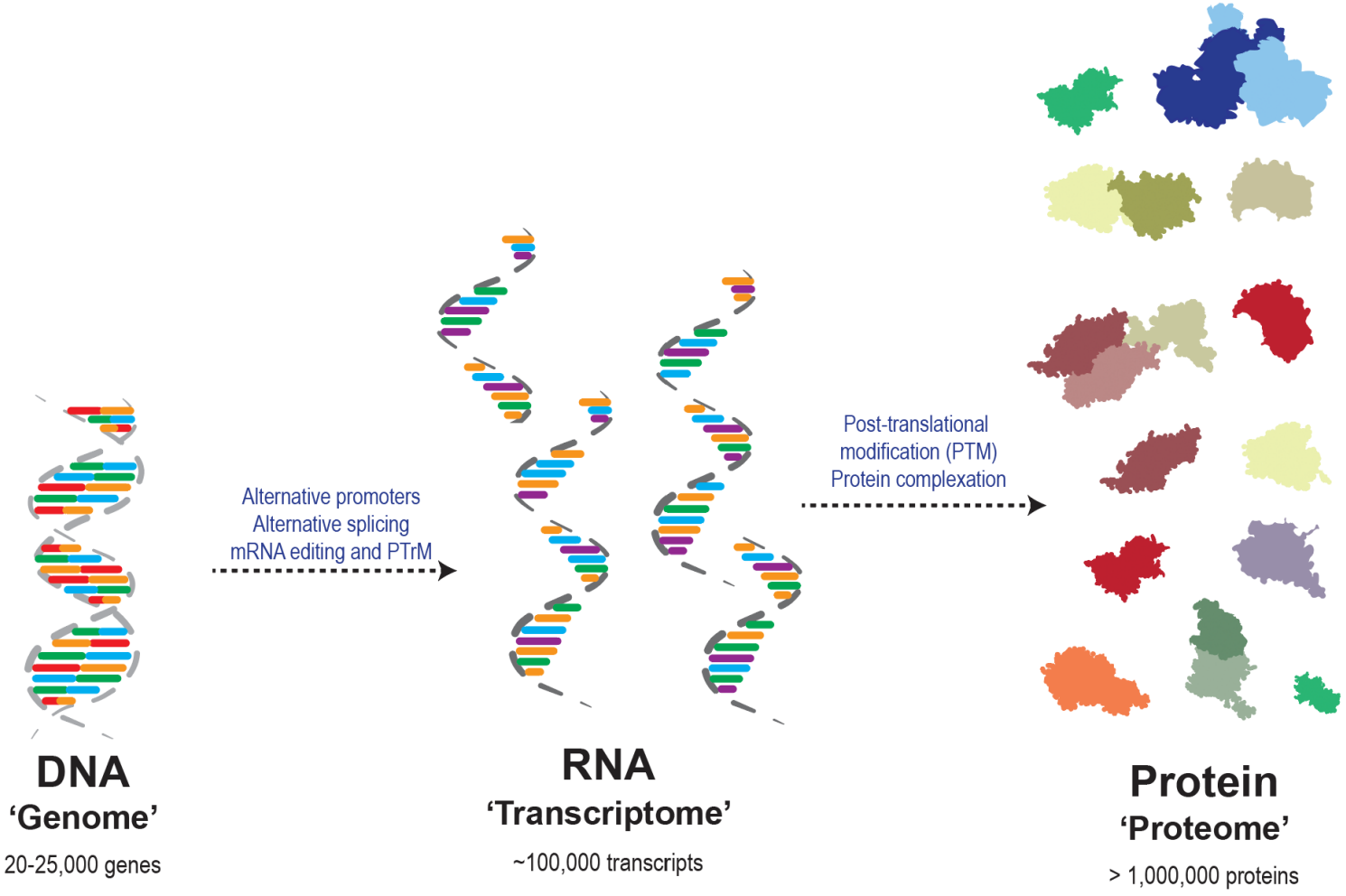
Jesse Rinehart, PhD

Biomedical Data Science: Mining & Modeling

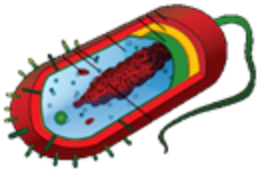
CBB 752, Spring 2025

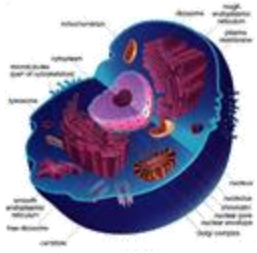
Cellular & Molecular Physiology
Yale University School of Medicine

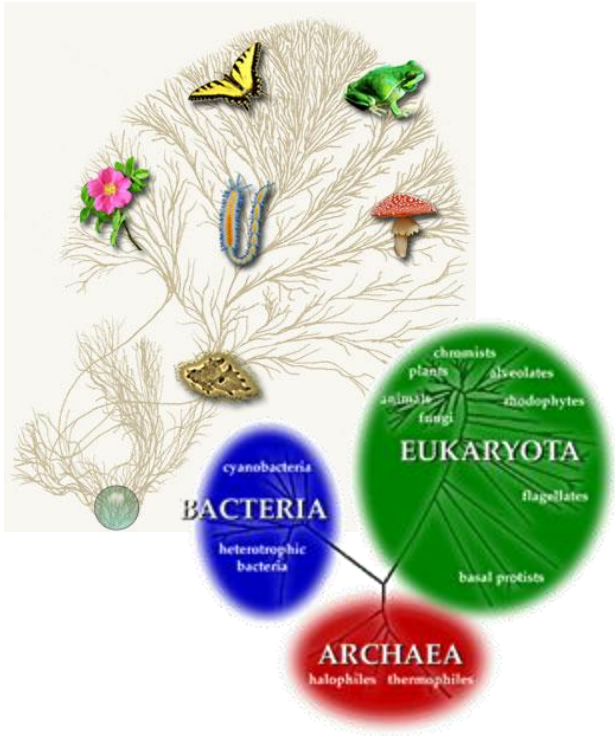
DNA → RNA → PROTEIN

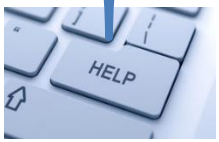


DNA → RNA → PROTEIN

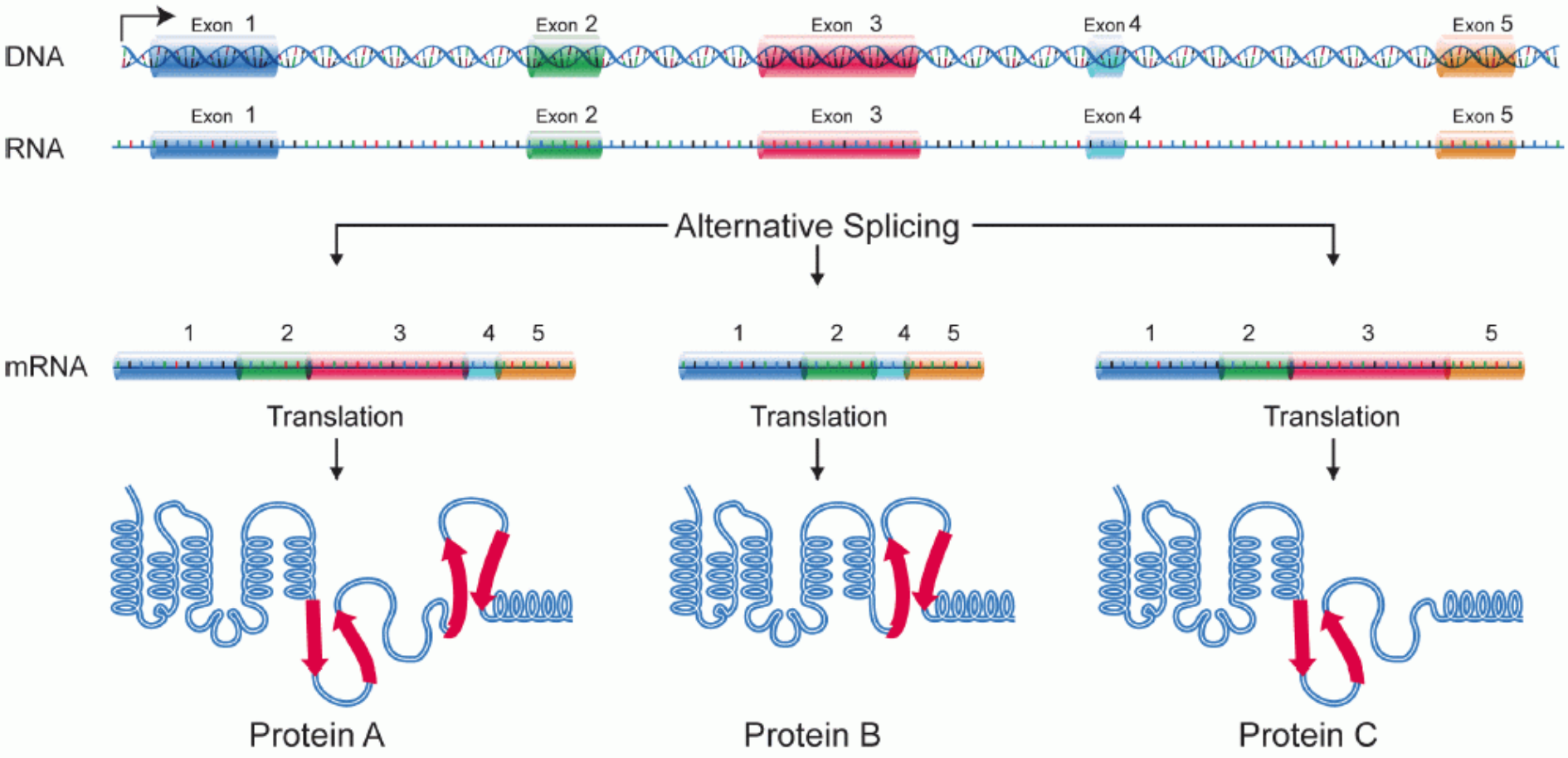






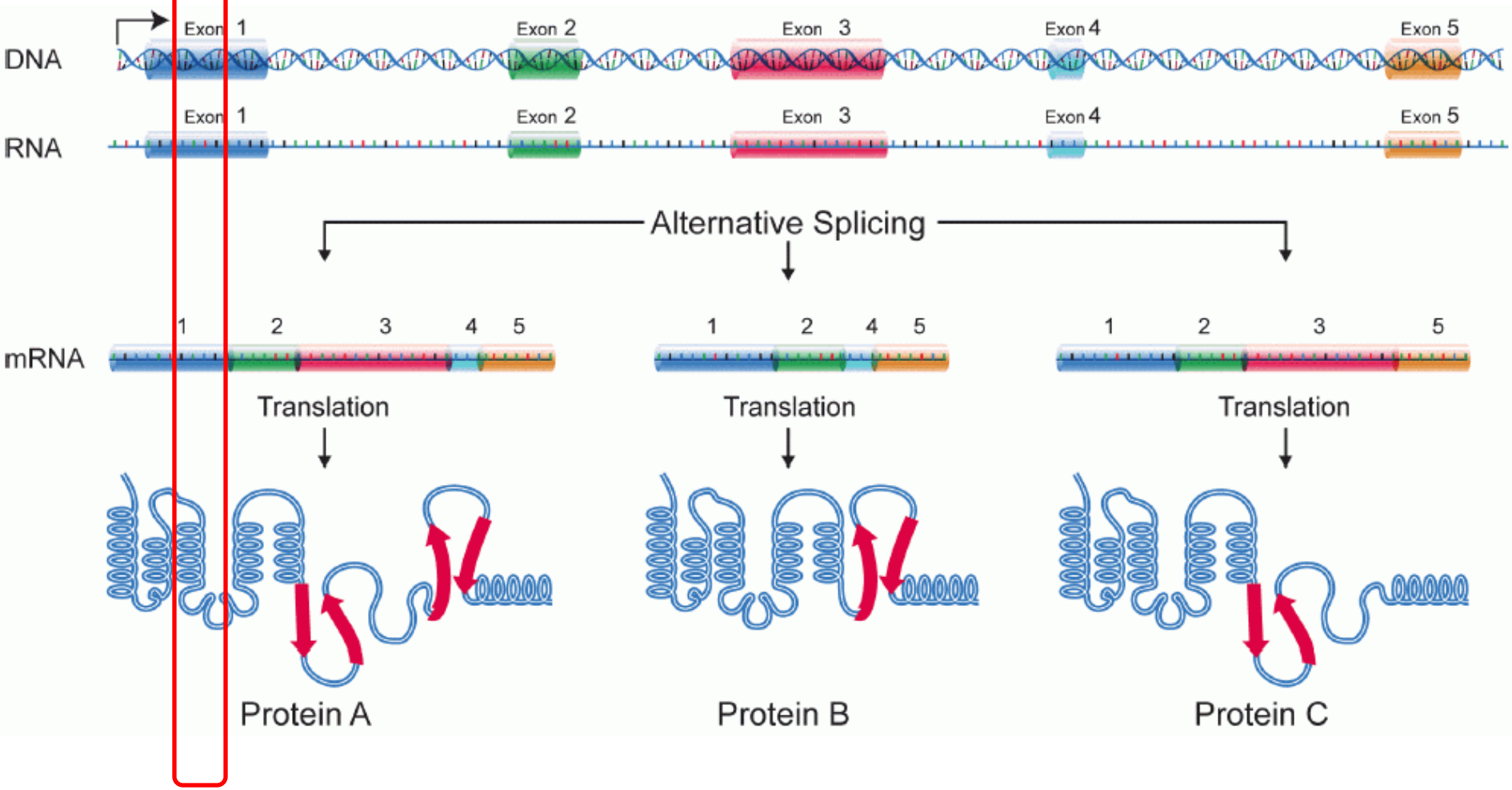


DNA → RNA → PROTEIN



https://en.wikipedia.org/wiki/Alternative_splicing

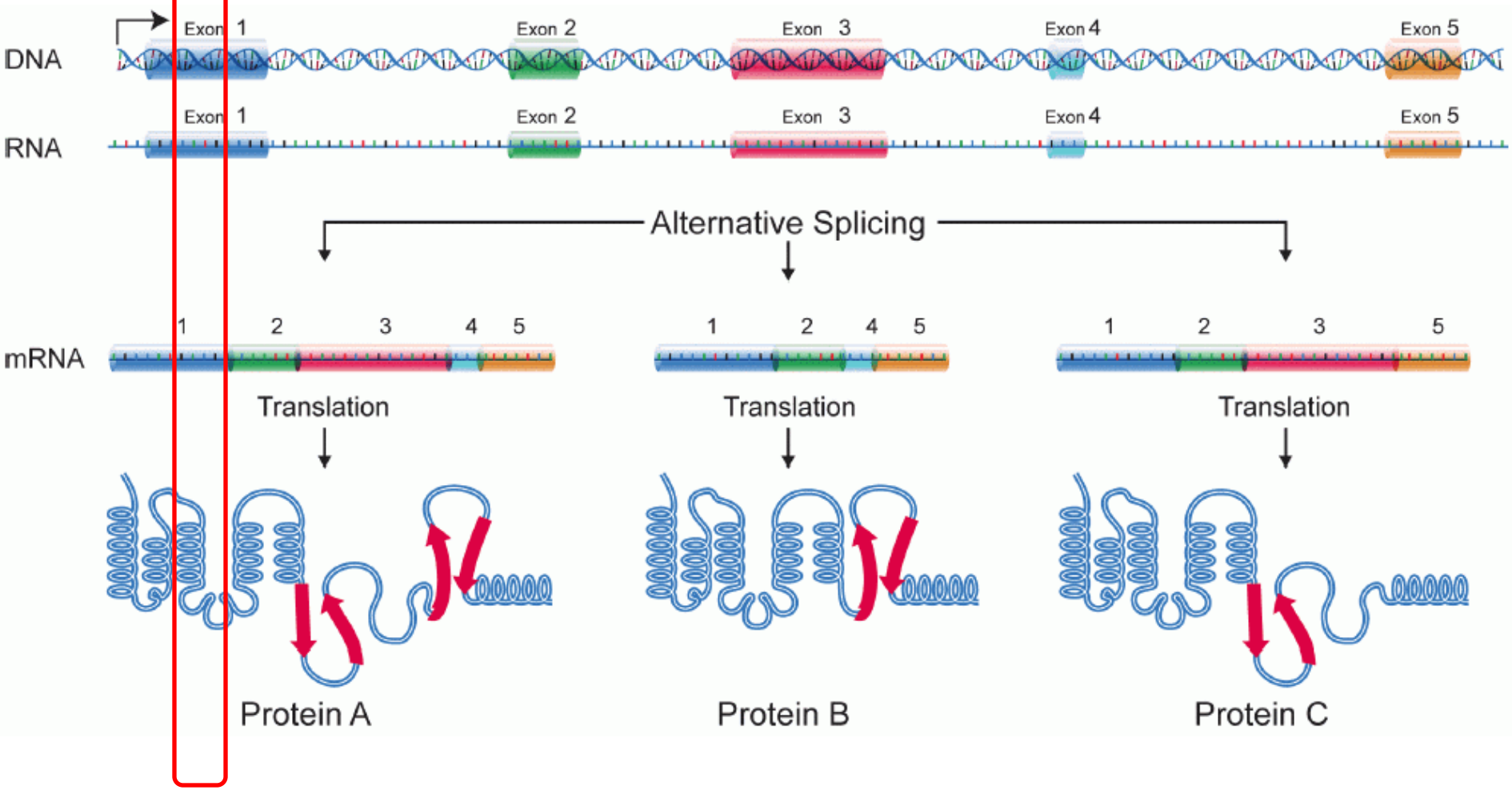
Data capture here



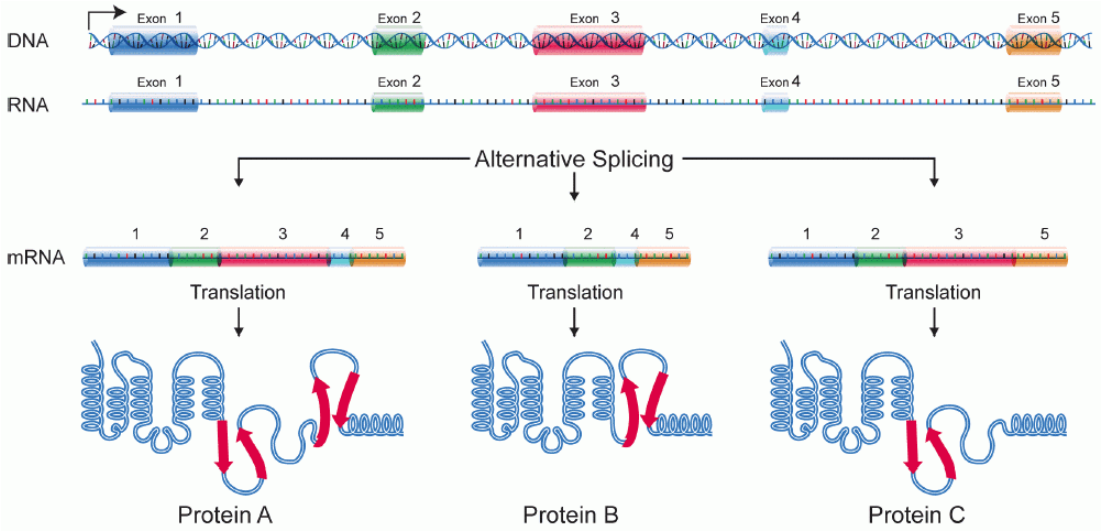
https://en.wikipedia.org/wiki/Alternative_splicing

Data capture here

Missing information



https://en.wikipedia.org/wiki/Alternative_splicing



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

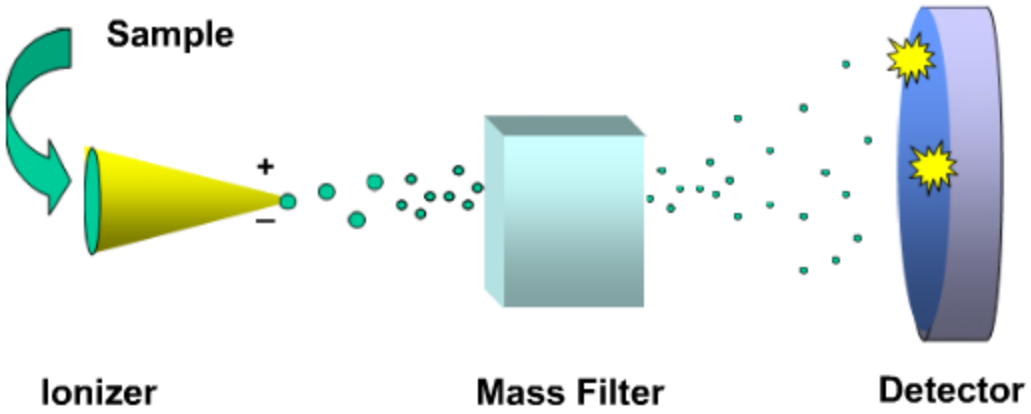
Overview

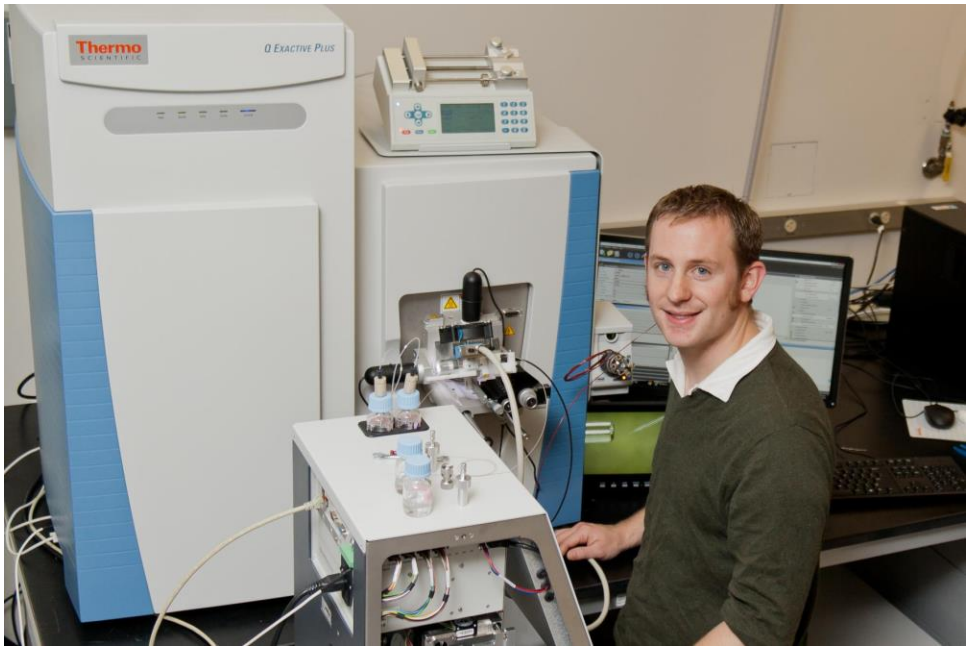
- **Techniques & Technologies**
  - Mass Spectrometry
  - Protein-Protein Interactions
  - Quantitative Proteomics
- **Applications**
  - Representative Studies
- **Putting it all together….**
  - Databases & Pathways

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

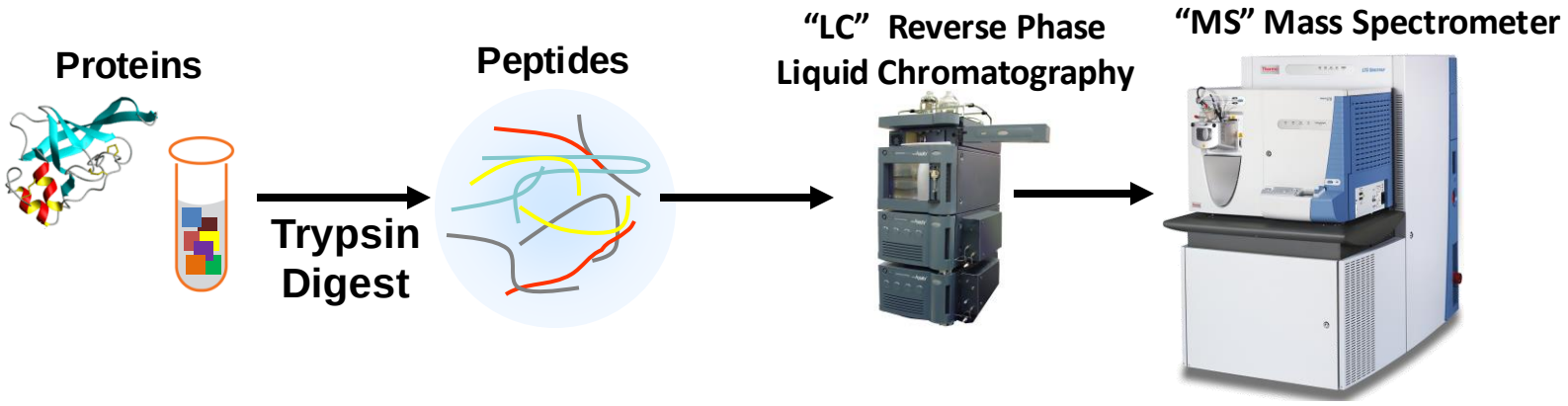
- Matrix Assisted Laser Desorption Ionization (MALDI)
  Tanaka K, Waki H, Ido Y, et al. *Rapid Commun Mass Spectrom* 1988

- 2002 Nobel Prize in Chemistry awarded to
  John B. Fenn & Koichi Tanaka

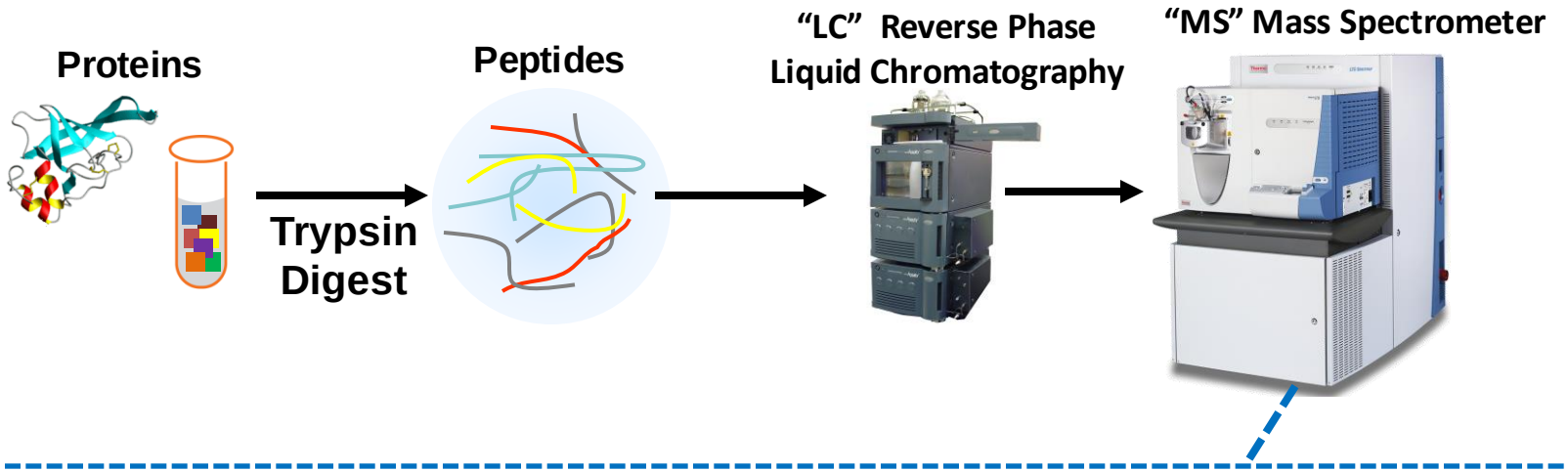
- **Enabled direct measurement and "sequencing" of intact peptides & MS based Proteomics is born**

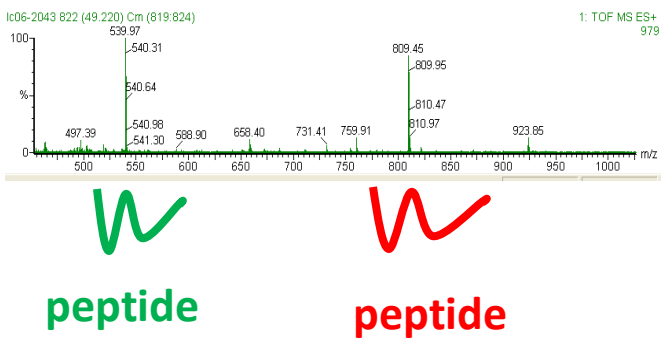
*Matthias Mann (Yale University; Ph.D.; 1988; Chemical Engineering) trained with John Fenn during some of the breakthrough work at Yale

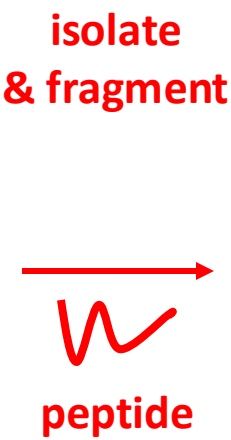
Typical work-flow for LC-MS "shotgun proteomics"

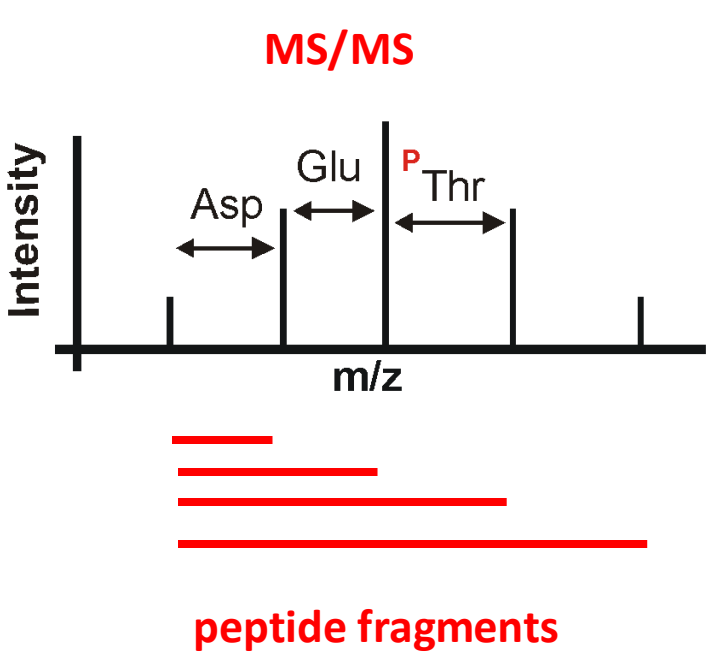


Typical work-flow for LC-MS "shotgun proteomics"

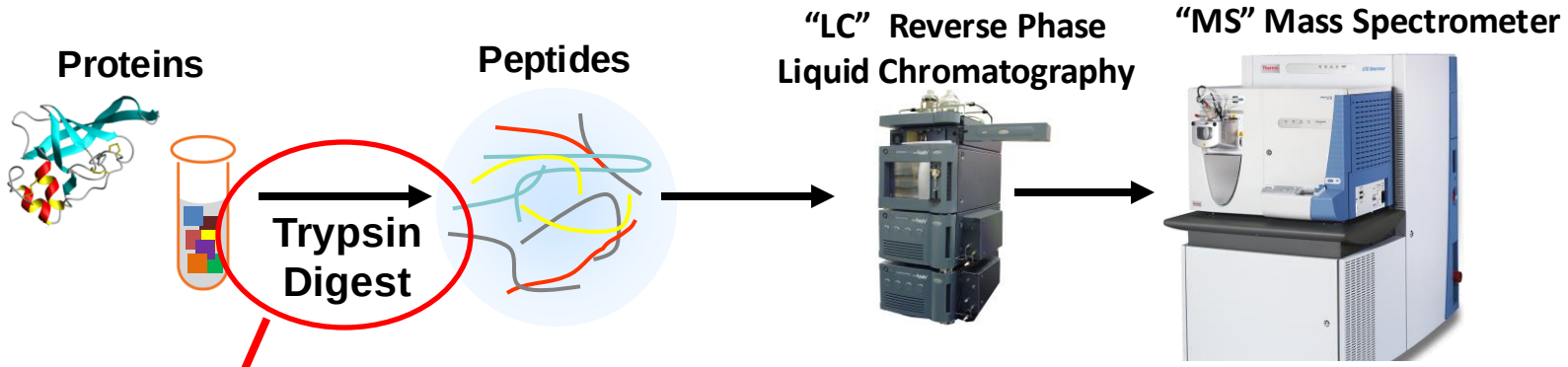




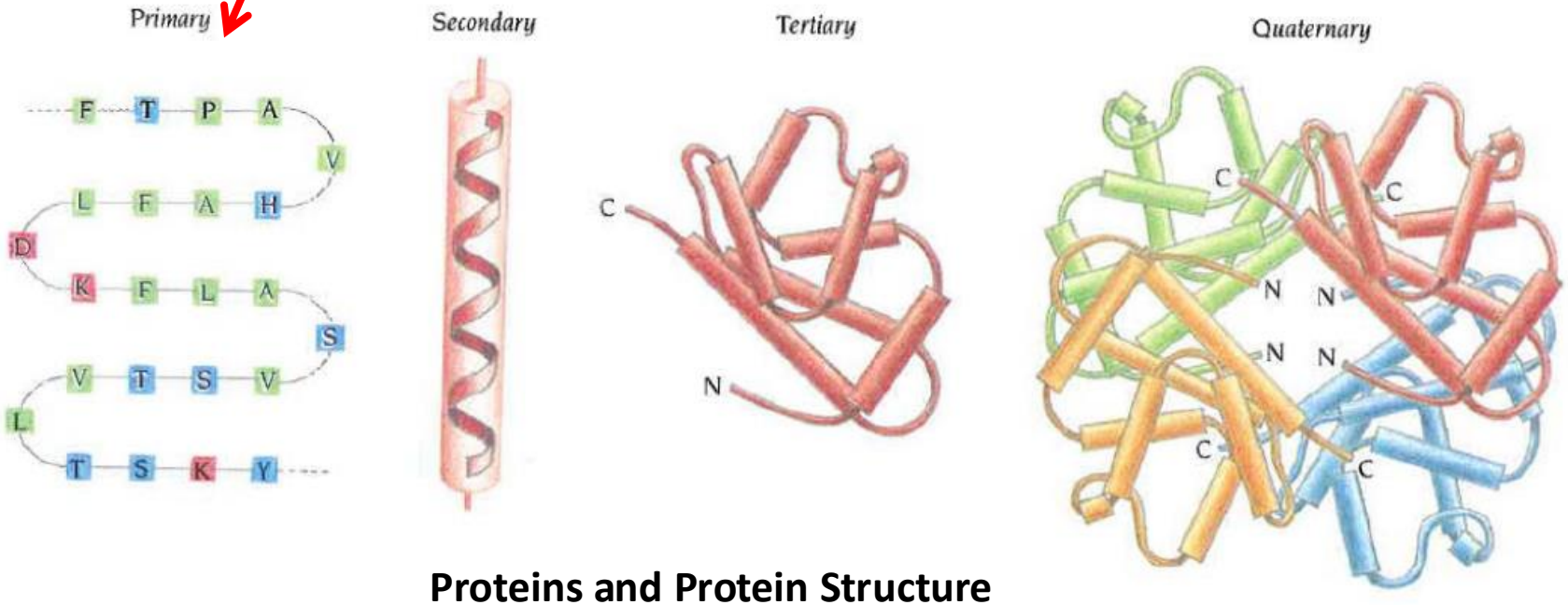




Typical work-flow for LC-MS "shotgun proteomics"



Trypsin cuts after Lys (K) & Arg (R)



Proteins and Protein Structure

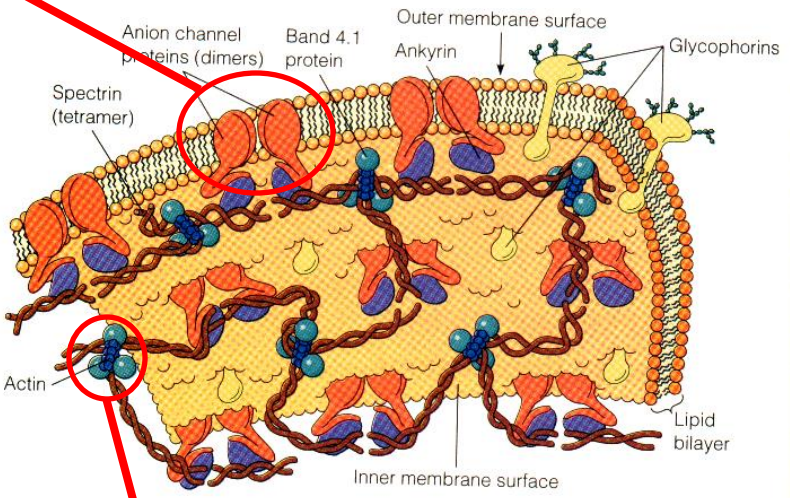
(Branden, C. and Tooze, J. ***Introduction to Protein Structure***)

Trypsin digest followed by LC-MS: Examples of "***Sequence Coverage***"

Band 3 Anion Transporter

Matched peptides shown in Bold Red

```
  1 MEELQDDYED MMEENLEQEE YEDPDIPESQ MEEPAAHDTE ATATDYHTTS
 51 HPGTHKVYVE LQELVMDEKN QELRWMEAAR WVQLEENLGE NGAWGRPHLS
101 HLTFWSLLEL RRVFTKGTVL LDLQETSLAG VANQLLDRFI FEDQIRPQDR
151 EELLRALLLK HSHAGELEAL GGVKPAVLTR SGDPSQPLLP QHSSLETQLF
201 CEQGDGGTEG HSPSGILEKI PPDSEATLVL VGRADFLEQP VLGFVRLQEA
251 AELEAVELPV PIRFLFVLLG PEAPHIDYTQ LGRAAATLMS ERVFRIDAYM
301 AQSRGELLHS LEGFLDCSLV LPPTDAPSEQ ALLSLVPVQR ELLRRRYQSS
351 PAKPDSSFYK GLDLNGGPDD PLQQTGQLFG GLVRDIRRRY PYYLSDITDA
401 FSPQVLAAVI FIYFAALSPA ITFGGLLGEK TRNQMGVSEL LISTAVQGIL
451 FALLGAQPLL VVGFSGPLLV FEEAFFSFCE TNGLEYIVGR VWIGFWLILL
501 VVLVVAFEGS FLVRFISRYT QEIFSFLISL IFIYETFSKL IKIFQDHPLQ
551 KTYNYNVLMV PKPQGPLPNT ALLSLVLMAG TFFFAMMLRK FKNSSYFPGK
601 LRRVIGDFGV PISILIMVLV DFFIQDTYTQ KLSVPDGFKV SNSSARGWVI
651 HPLGLRSEFP IWMMFASALP ALLVFILIFL ESQITTLIVS KPERKMVKGS
701 GFHLDLLLVV GMGGVAALFG MPWLSATTVR SVTHANALTV MGKASTPGAA
751 AQIQEVKEQR ISGLLVAVLV GLSILMEPIL SRIPLAVLFG IFLYMGVTSL
801 SGIQLFDRIL LLFKPPKYHP DVPYVKRVKT WRMHLFTGIQ IICLAVLWVV
851 KSTPASLALP FVLILTVPLR RVLLPLIFRN VELQCLDADD AKATFDEEEG
901 RDEYDEVAMP V
```

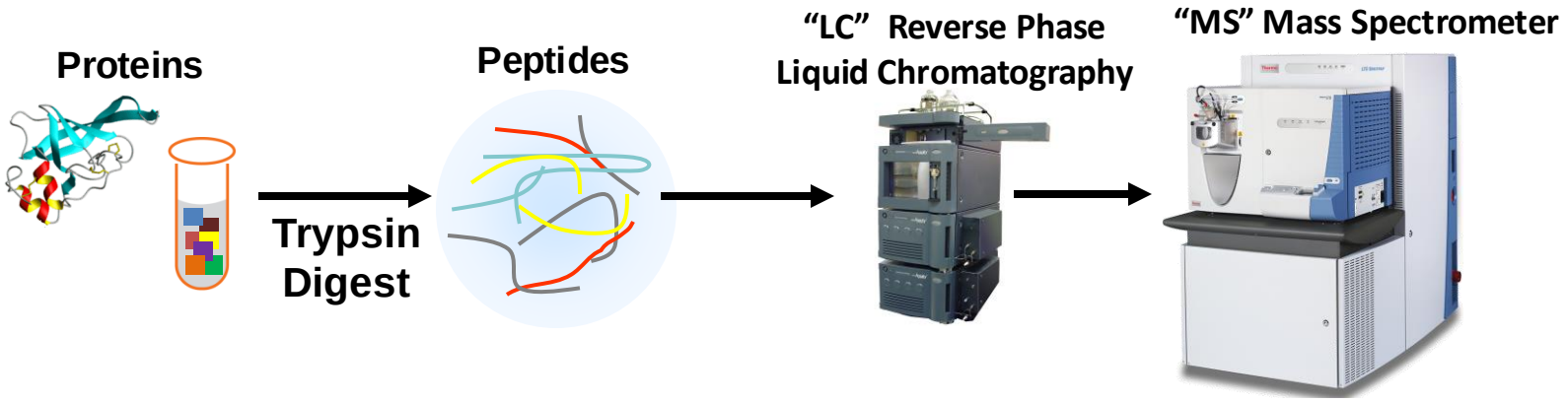


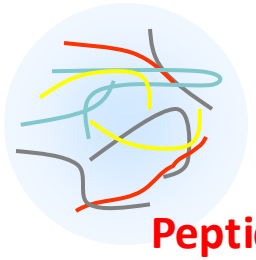
β-actin

Matched peptides shown in Bold Red

```
  1 MDDDIAALVV DNGSGMCKAG FAGDDAPRAV FPSIVGRPRH QGVMVGMGQK
 51 DSYVGDEAQS KRGILTLKYP IEHGIVTNWD DMEKIWHHTF YNELRVAPEE
101 HPVLLTEAPL NPKANREKMT QIMFETFNTP AMYVAIQAVL SLYASGRTTG
151 IVMDSGDGVT HTVPIYEGYA LPHAILRLDL AGRDLTDYLM KILTERGYSF
201 TTTAEREIVR DIKEKLCYVA LDFEQEMATA ASSSSLEKSY ELPDGQVITI
251 GNERFRCPEA LFQPSFLGME SCGIHETTFN SIMKCDVDIR KDLYANTVLS
301 GGTTMYPGIA DRMQKEITAL APSTMKIKII APPERKYSVW IGGSILASLS
351 TFQQMWISKQ EYDESGPSIV HRKCF
```

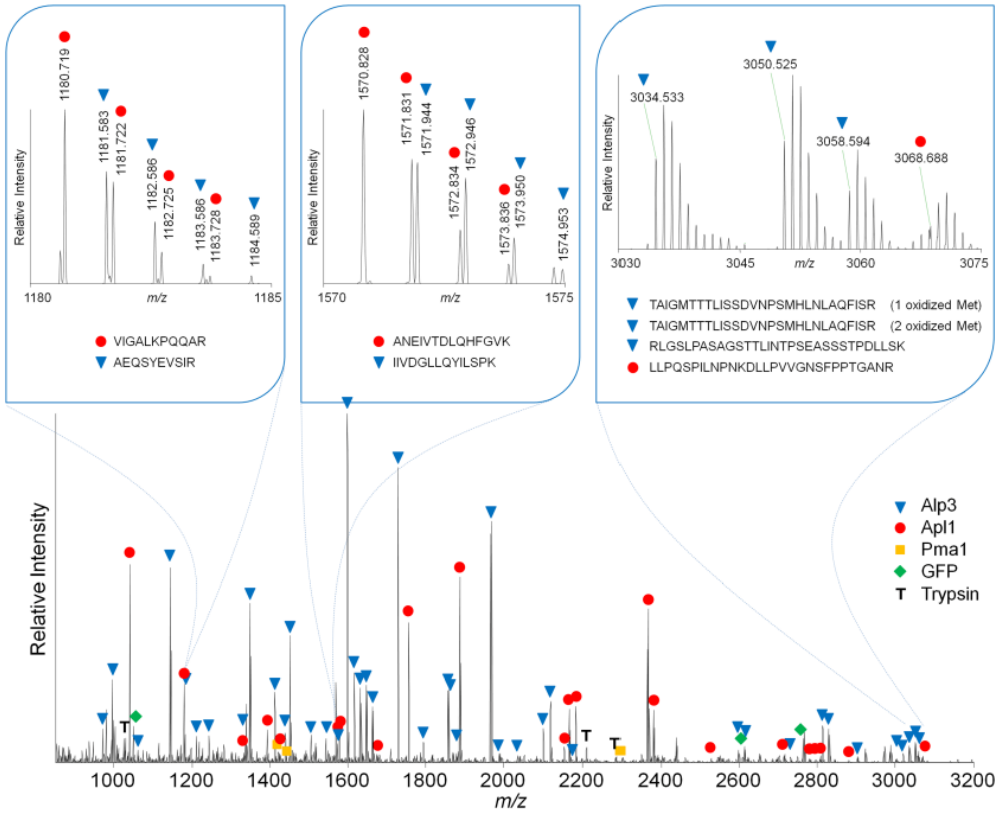
The mass spectra of peptide mixtures are complex



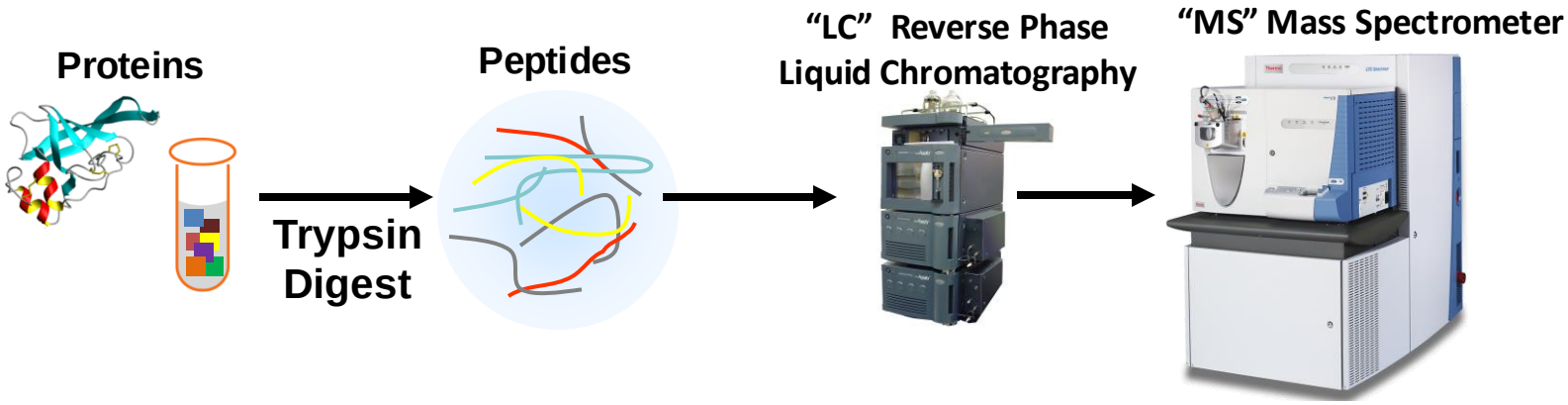


Peptide ions have a mass (m) and a charge (z).

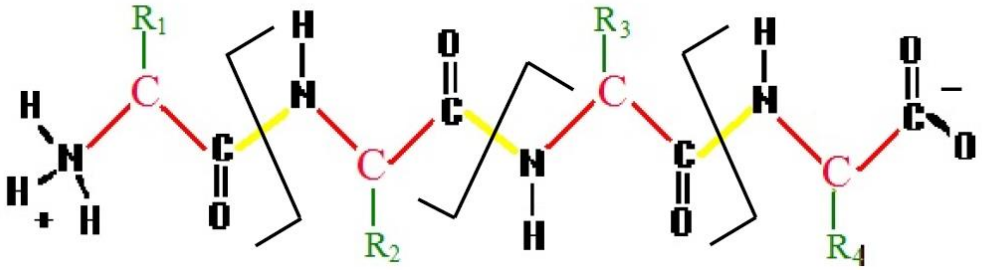
100 Da peptide:
+1 = 100 m/z
+2 = 50 m/z
+3 = 33.3 m/z



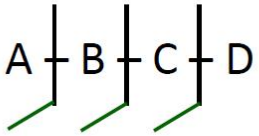
Peptide ions are isolated, fragmented, and “sequenced”



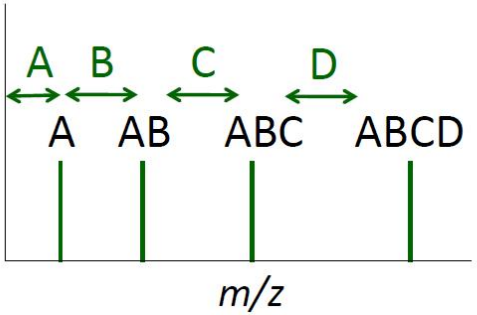
Peptide sequencing



Simplified concept of peptide fragmentation



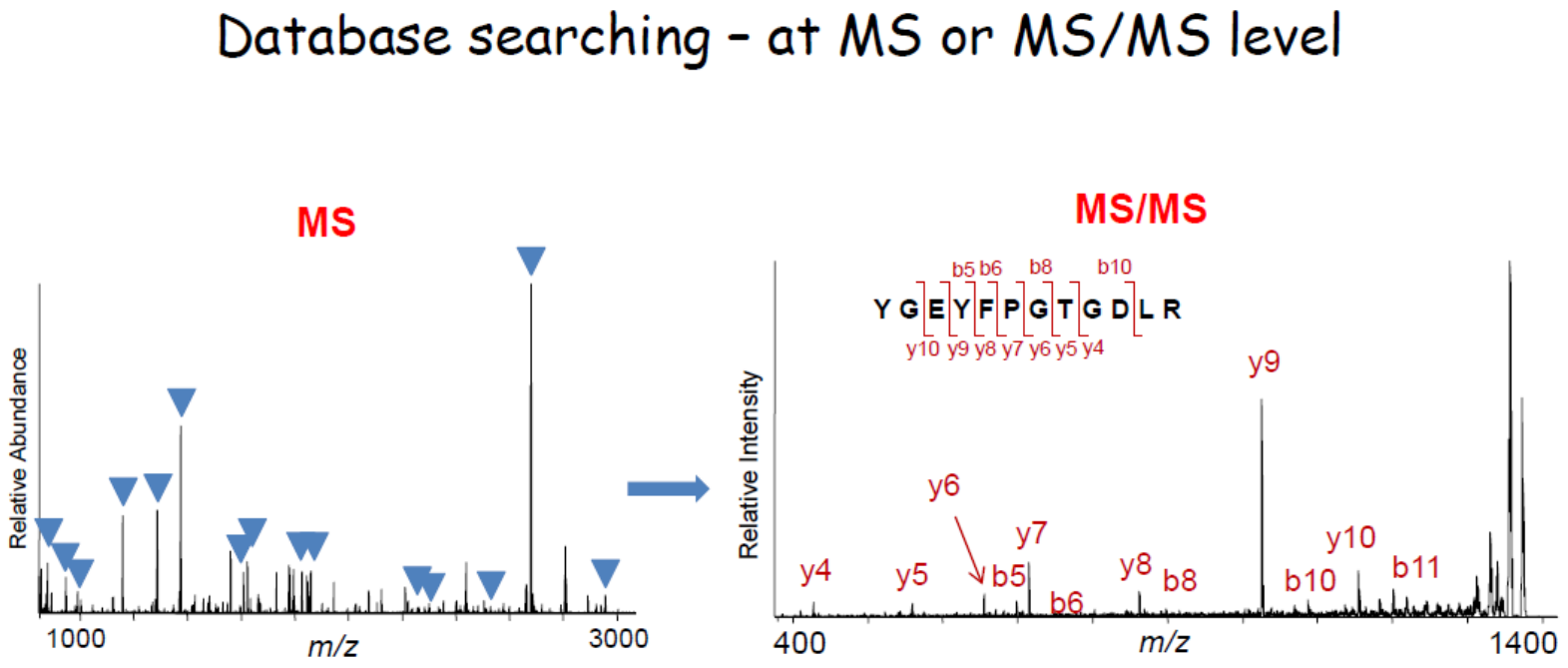
Peptide “ABCD”



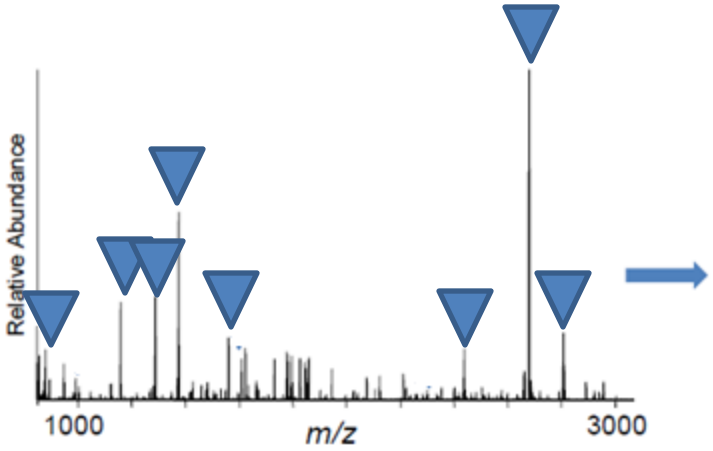
Fragment Spectra of Peptide “ABCD”

Computational Steps:

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

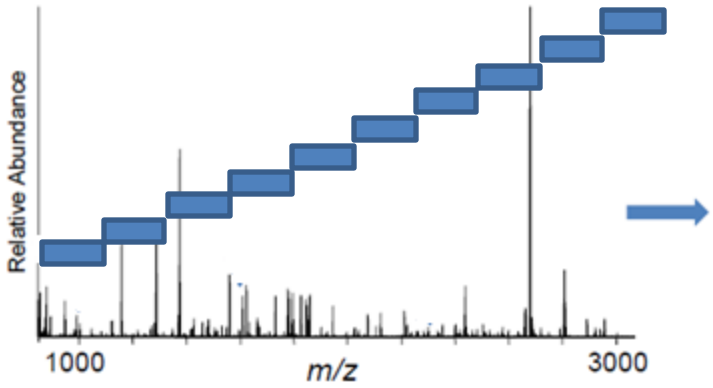


DIA (Data-independent Acquisition) **vs. DDA** (Data-dependent Acquisition)



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

Further Reading: PMID27092249; PMID30104418

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**

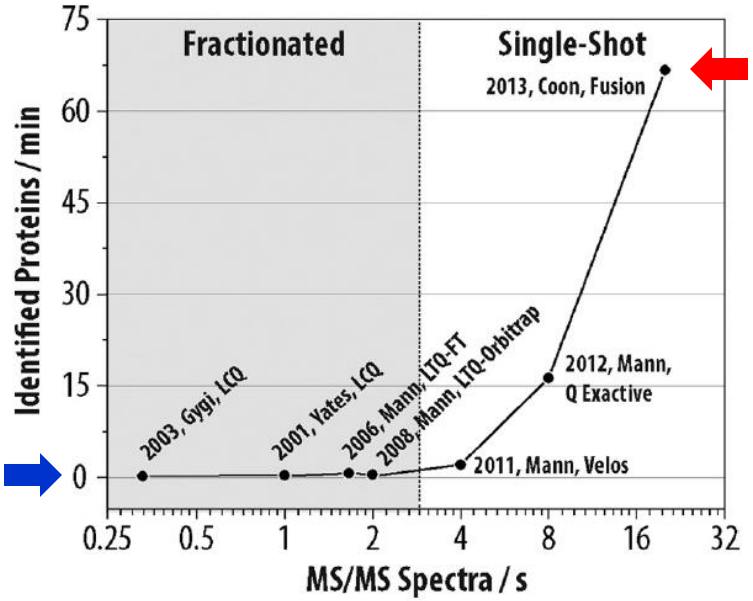


FIG. 5. **Rate of protein identifications as a function of mass spectrometer scan rate for selected large-scale yeast proteome analyses over the past decade.** Each data point is annotated with the year, corresponding author, type of MS system used, and reference number.

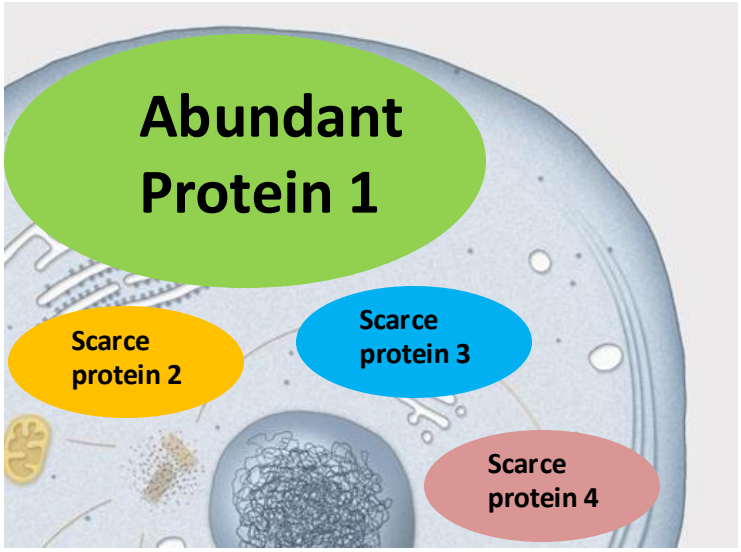
"each **one hour** analysis achieved detection of **3,977 proteins"**

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

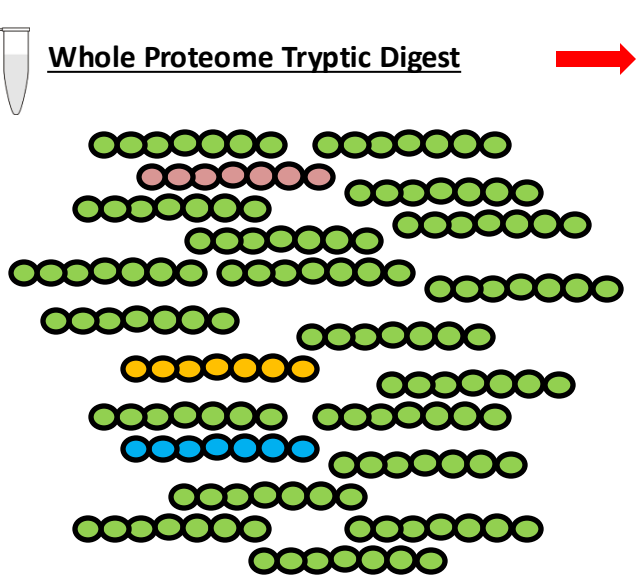
The one hour yeast proteome. Hebert AS, et a, Coon JJ.
Mol Cell Proteomics. 2014 PMID: 24143002 & *Nat Protoc.* 2015. PMID: 25855955

Challenge Question:

Cell with a 4 protein proteome



Whole Proteome Tryptic Digest



One LC-MS run

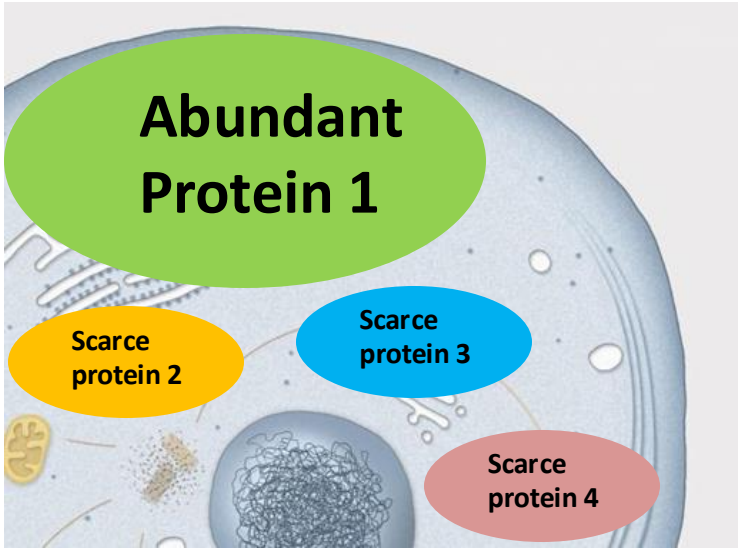
(Hypothetical MS that can only identify one peptide)

Protein 1 Identified

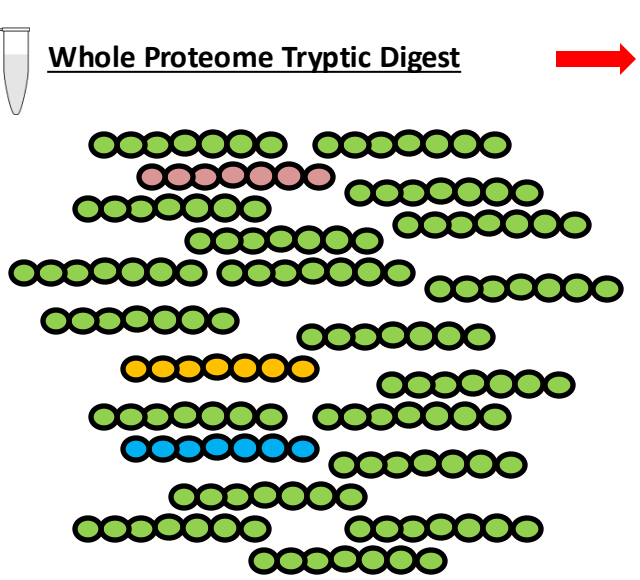
Challenge Question:

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

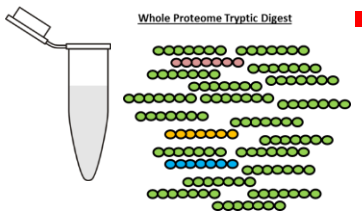


One LC-MS run

(Hypothetical MS that can only identify one peptide)

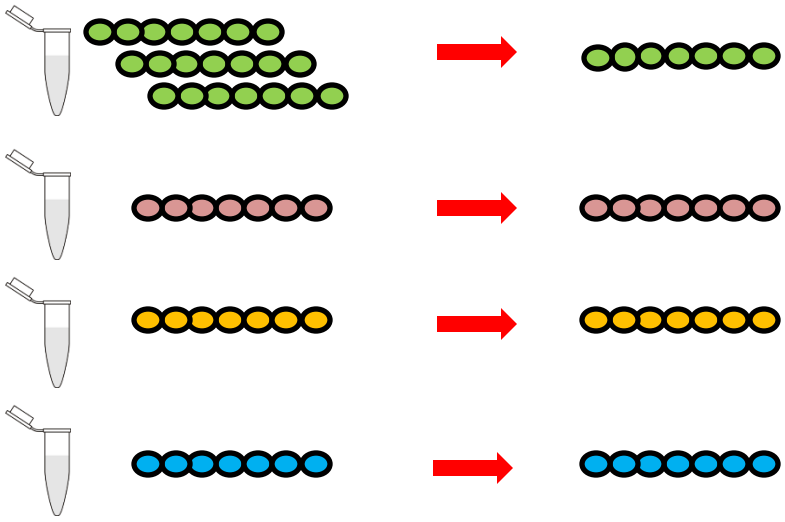
Protein 1 Identified

Option #1: Peptide Fractionation



Chromatography + fractionation

4 separate LC-MS runs



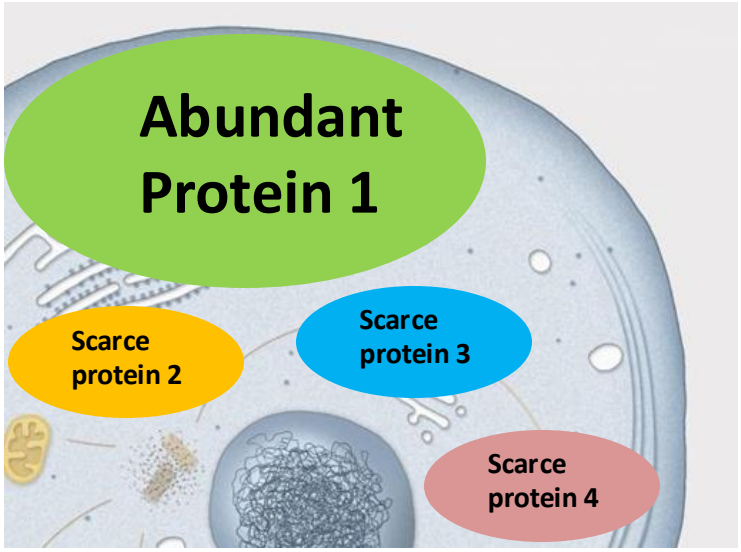
Protein 1 Identified

Protein 2 Identified

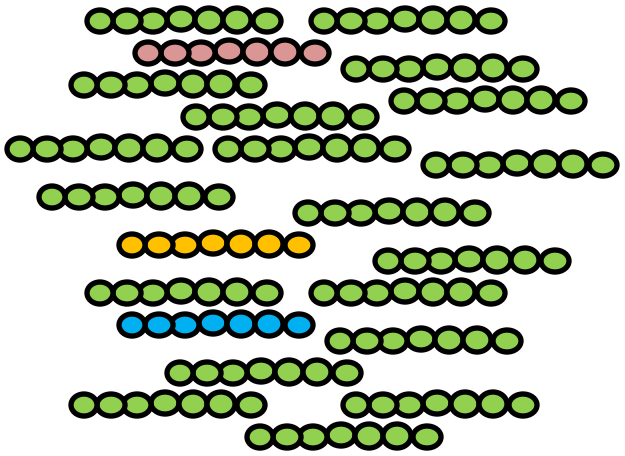
Protein 3 Identified

Protein 4 Identified

Cell with a 4 protein proteome



Whole Proteome Tryptic Digest

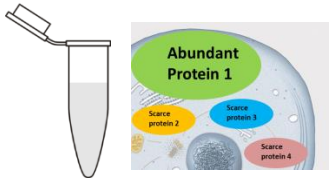


One LC-MS run

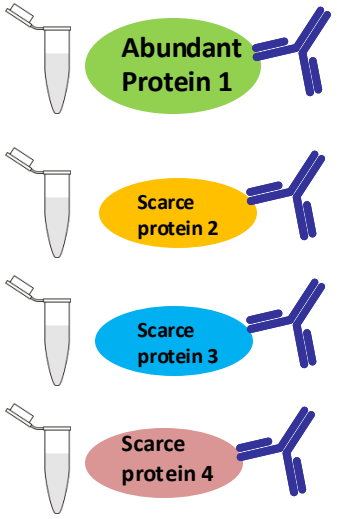
(Hypothetical MS that can only identify one peptide)

Protein 1 Identified

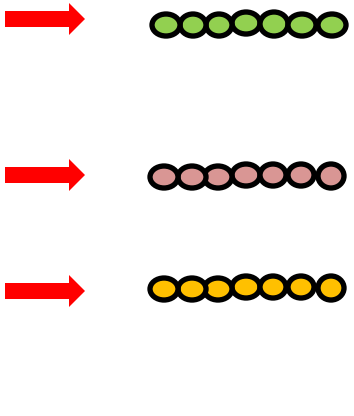
Option #2: Proteome Fractionation (e.g. Immunoprecipitation)



Separate IP Tryptic Digest



4 separate LC-MS runs



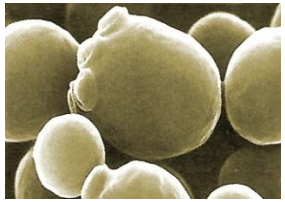
Protein 1 Identified

Protein 2 Identified

Protein 3 Identified

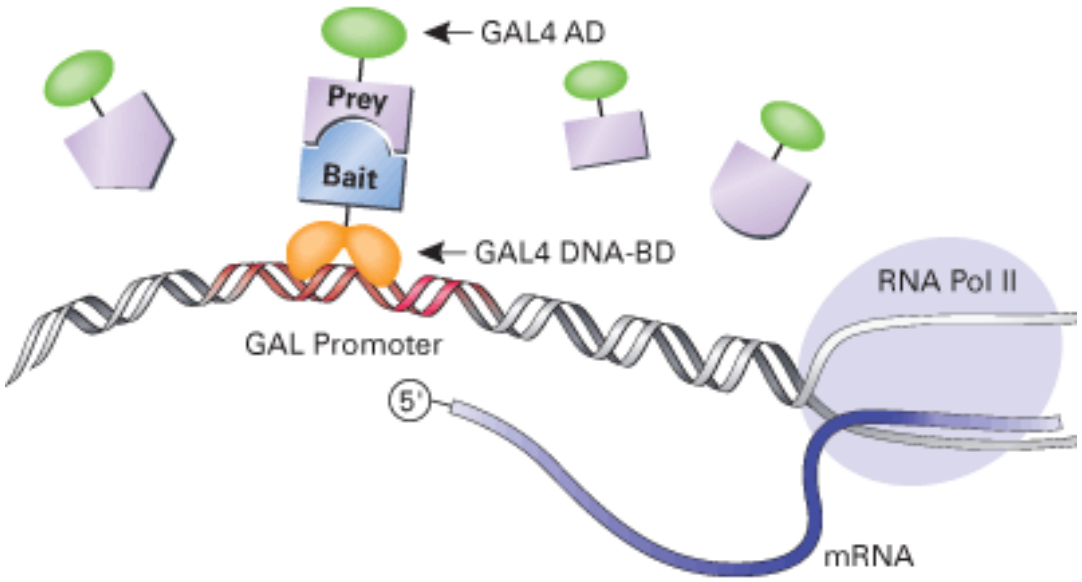
Protein 4 Identified

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



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**

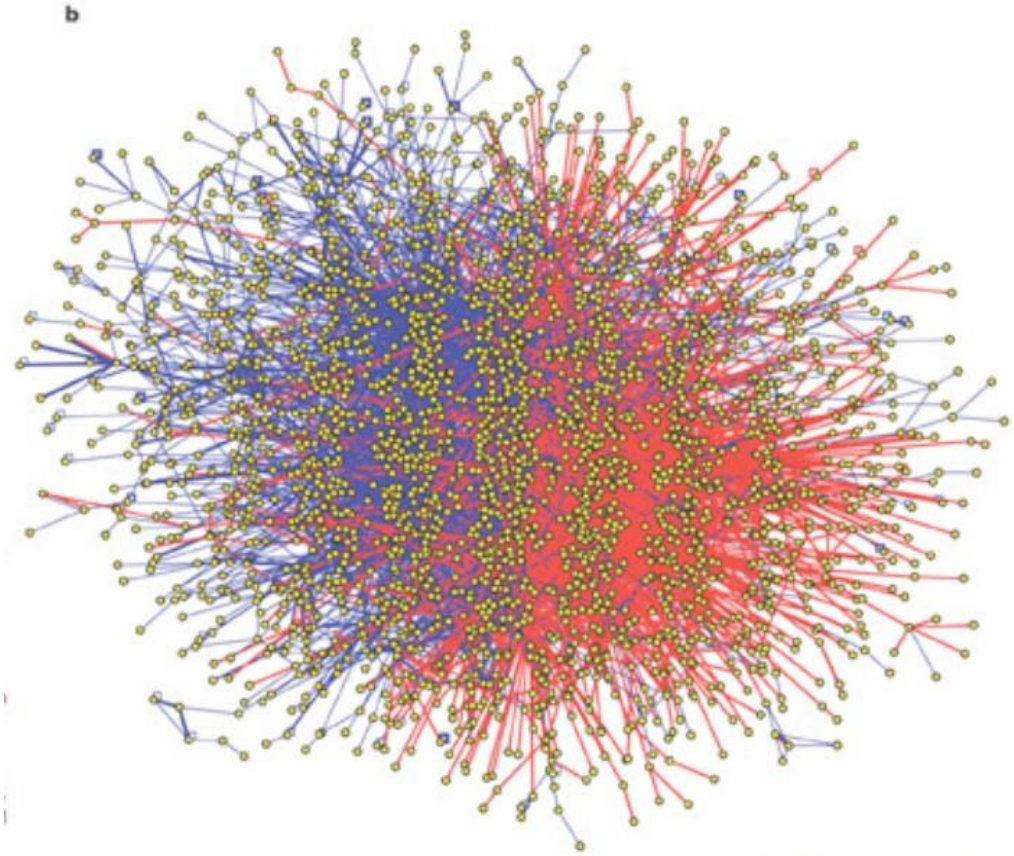
Some Disadvantages

- **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

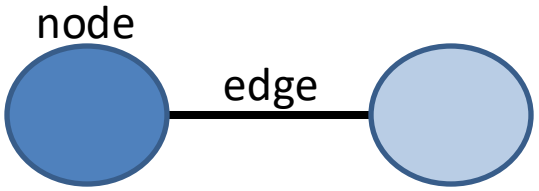
8,100 ORFs (~7,200 genes)

10,597 interactions

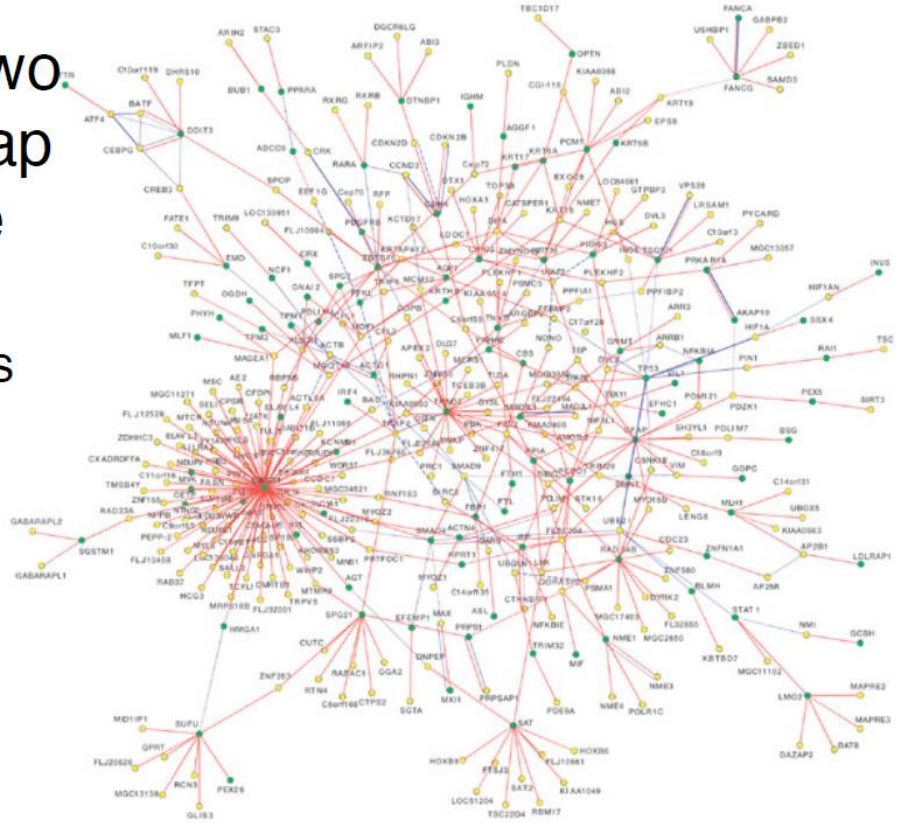


Rual et al. Nature 2005

Protein-Protein interaction maps: Proteins are represented by **nodes** and interactions are represented by **edges** between nodes.



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

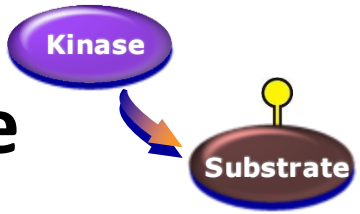


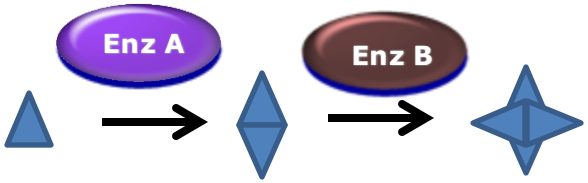
Rual et al. Nature 2005 Vol 437

Protein-Protein interactions:

Some examples:

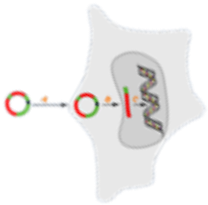
- Physical and direct
- Physical and indirect
  - Multi-protein complexes
  - Scaffolds
- Transient
  - Kinase & substrate
- Metabolic

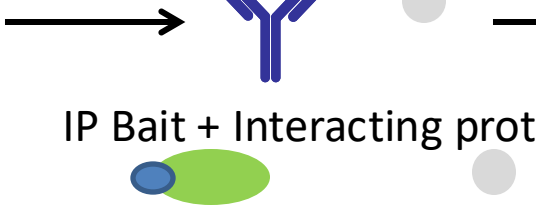




Adding common molecular handles to “tag” every protein

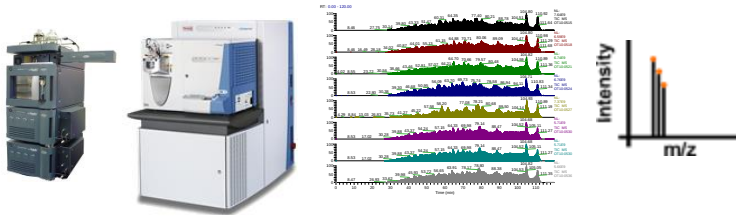
Tagged “bait” protein



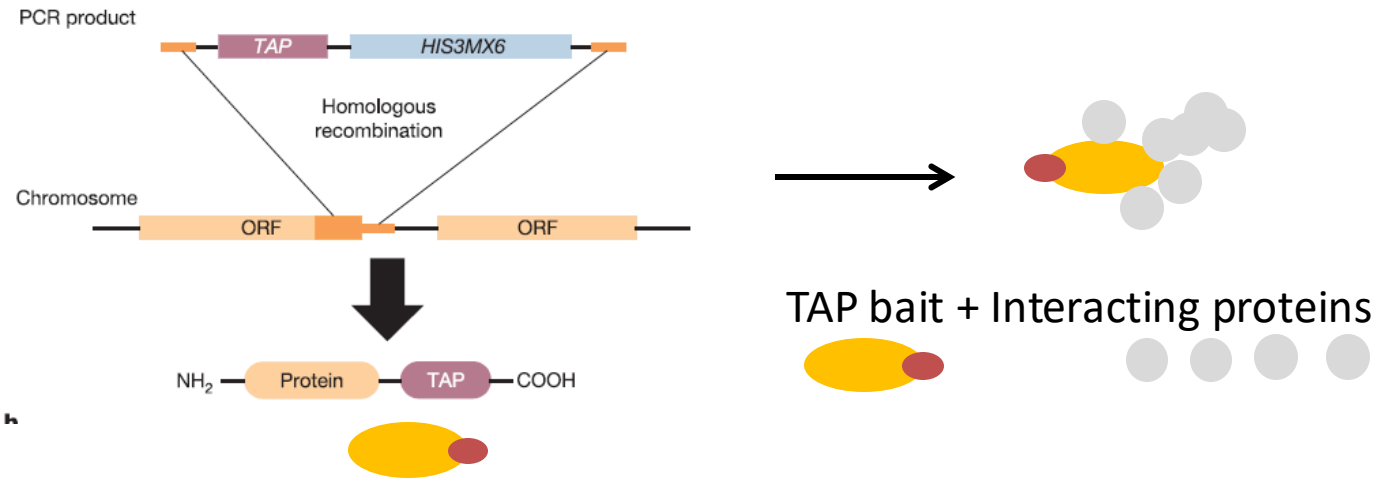


IP Bait + Interacting proteins

Multiple runs of “shot gun” LC-MS/MS



Adding common molecular handles to “tag” every protein



Collection of tagged “bait” expression strains

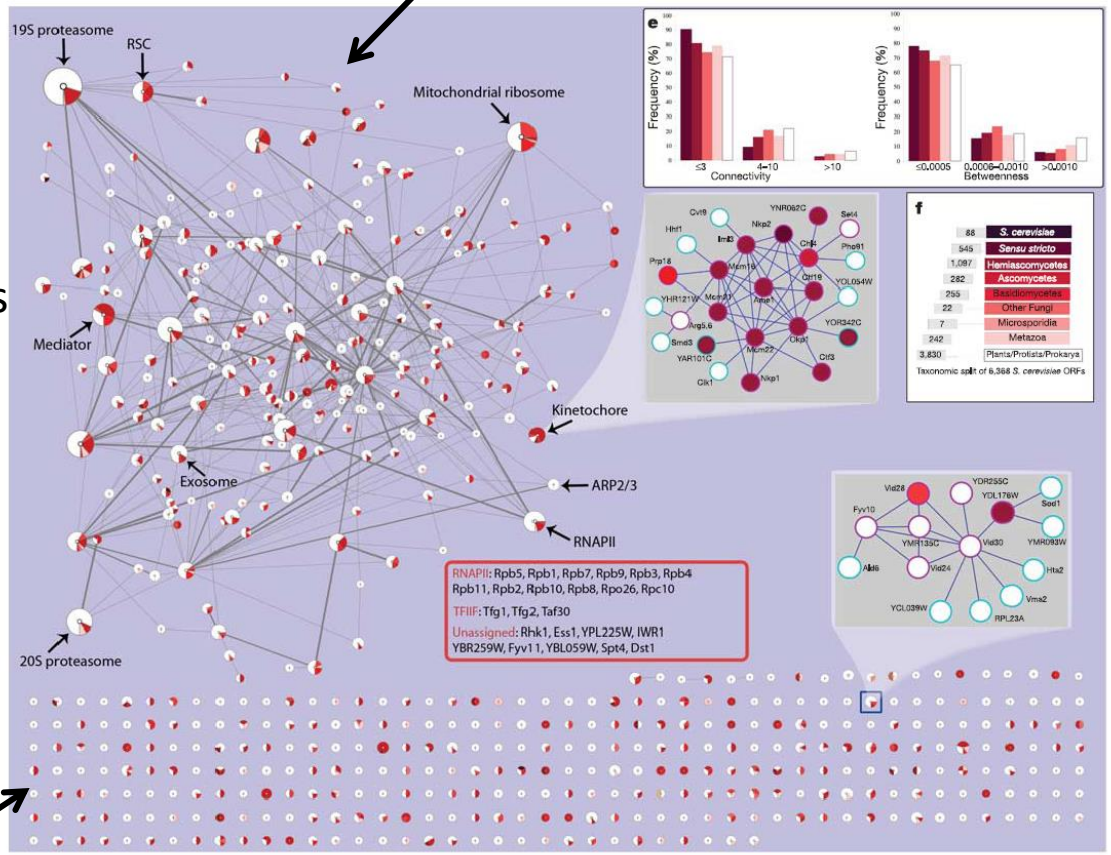
2003

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**

Cellular proteins are organized into complexes

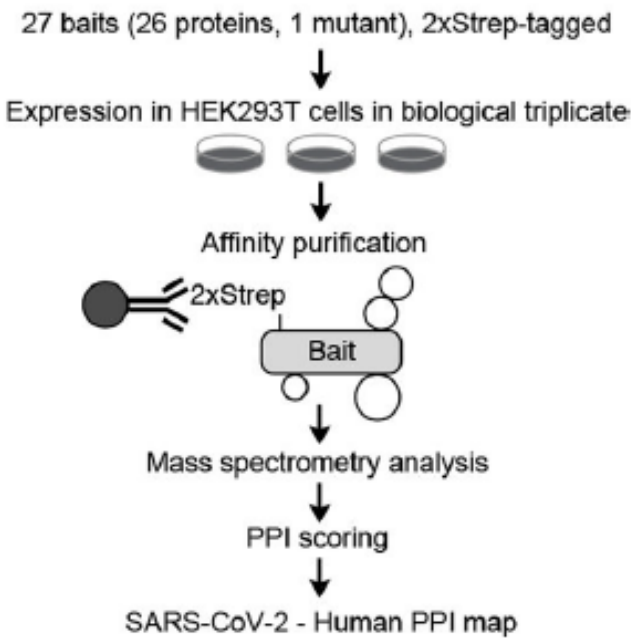
- 4,562 tagged proteins
- 2,357 successful purifications
- Identified 4,087 interacting proteins ~72 % proteome
- Majority of the yeast proteome is organized into complexes
- Many complexes are conserved in other species

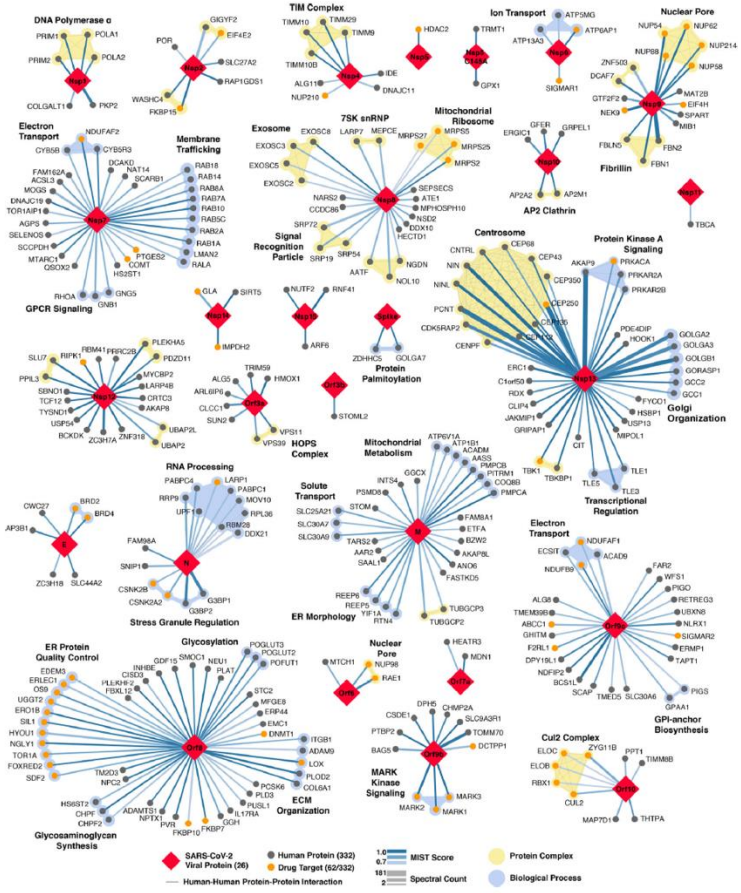


Complexes with little or no interconnectivity

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

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





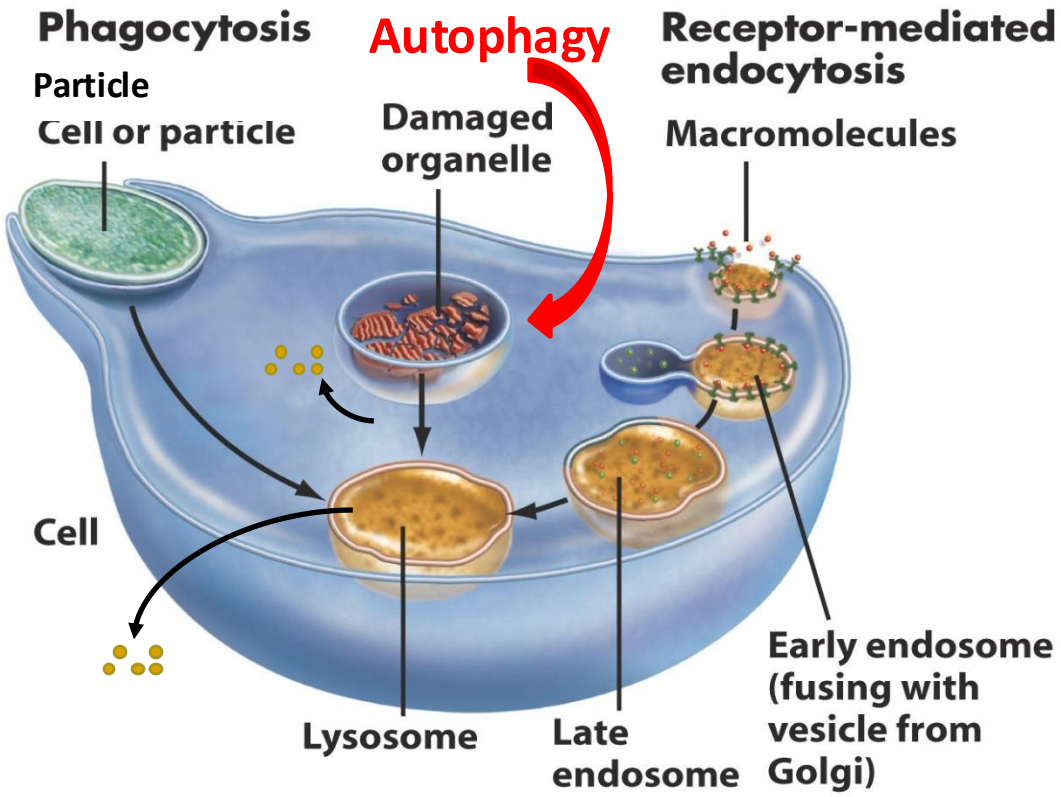
Gordon et al. Nature. 2020
Krogan lab PMID: 32353859

nature Vol 466|1 July 2010|doi:10.1038/nature09204

ARTICLES

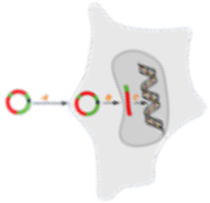
Network organization of the human autophagy system

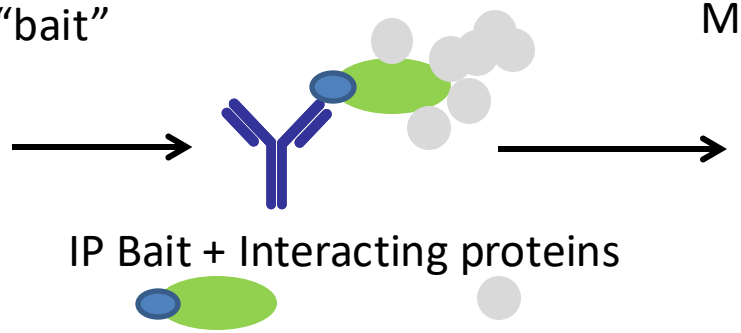
Christian Behrends[1], Mathew E. Sowa[1], Steven P. Gygi[2] & J. Wade Harper[1]



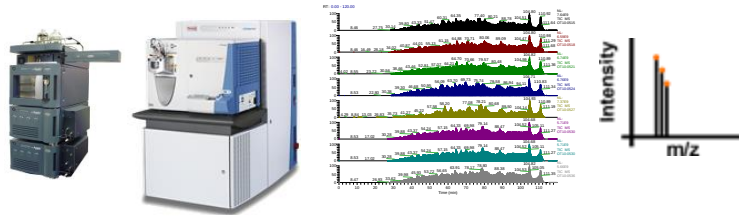
Pearson Prentice Hall, Inc. 2005
www.stolaf.edu/people/giannini/cell/lys.htm

Transfect tagged "bait"





Multiple runs of "shot gun" 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

Behreands et al, Nature 2010

Autophagy Interaction Network

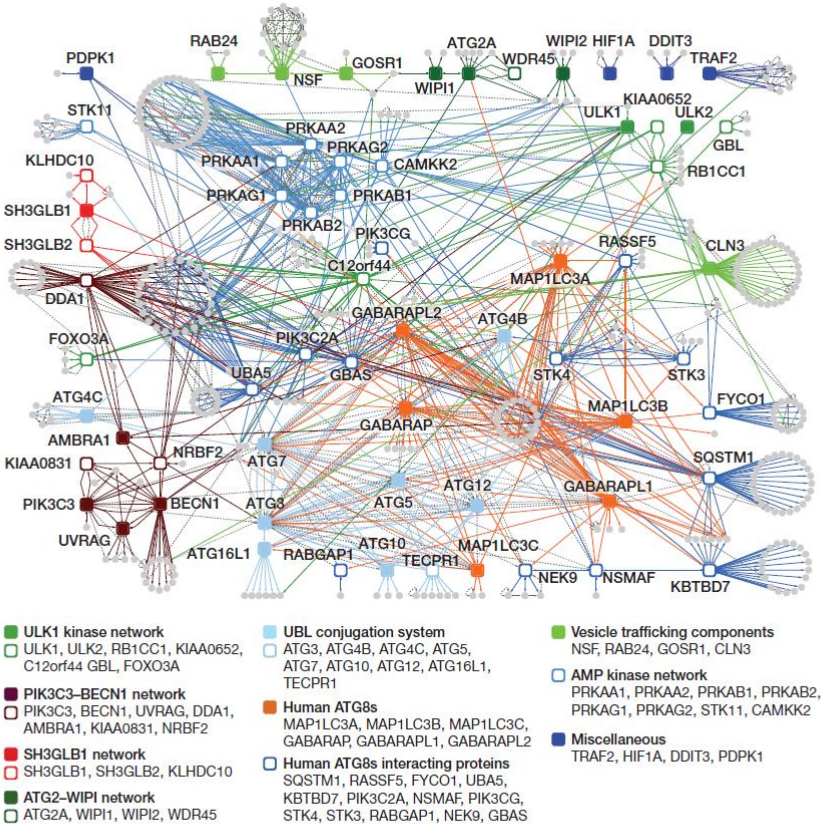


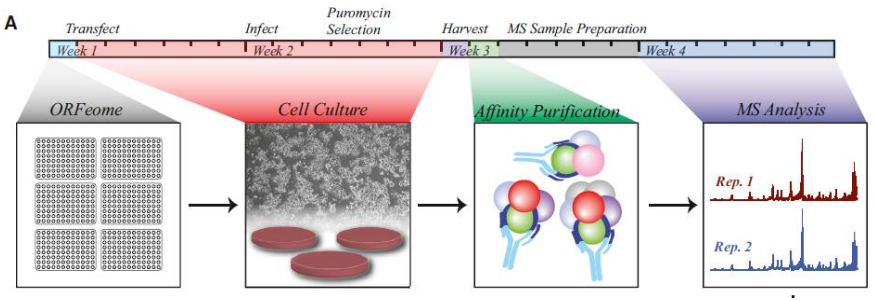
Figure 1 | Overview of the autophagy interaction network (AIN). HCIPs within the autophagy network are shown for 32 primary baits (filled squares) and 33 secondary baits (open squares). Subnetworks are colour-coded. Interacting proteins are indicated by grey circles.

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. Huttlin[1], Raphael J. Bruckner[1], Joao A. Paulo[1], Joe R. Cannon[1], Lily Ting[1], Kurt Baltier[1], Greg Colby[1], Fana Gebreab[1],

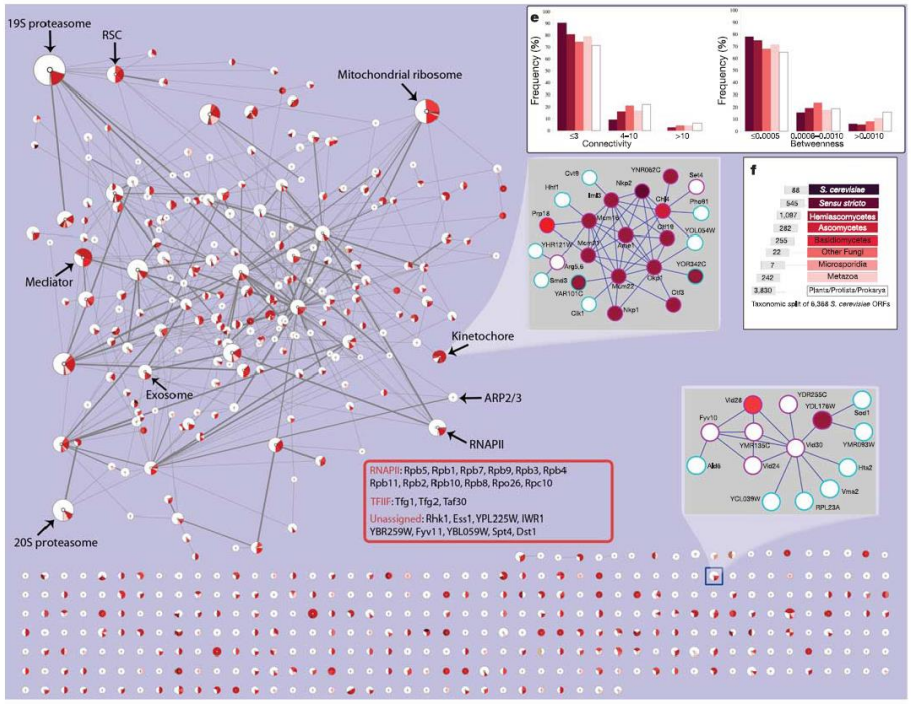
BioPlex 2.0 Huttlin et al, *Nature*. 2017 PMID: 28514442

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.

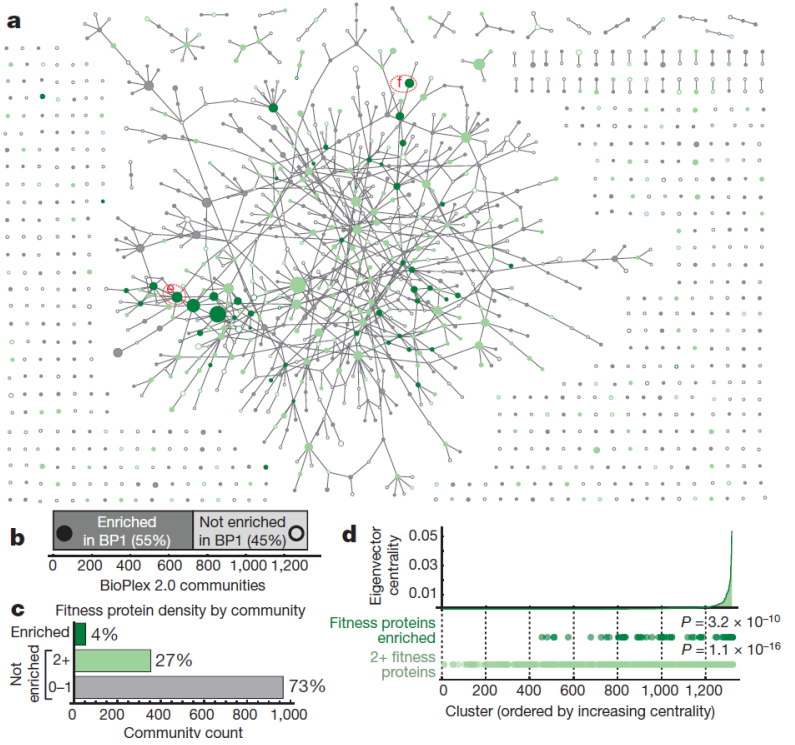
Cellular proteins are organized into complexes and this proteome organization is conserved

Yeast: Interaction Network of Complexes



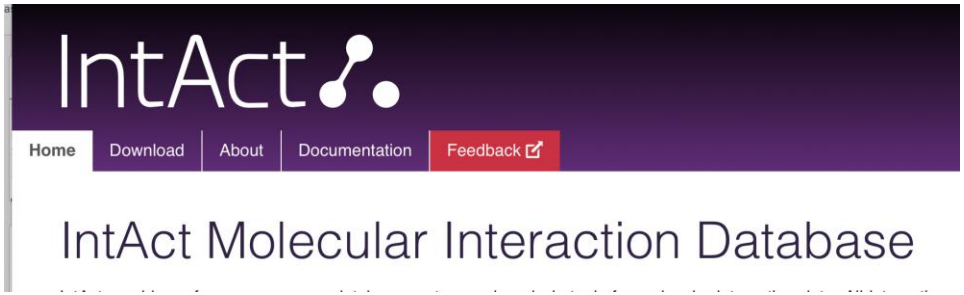
Krogan NJ, et al. *Nature*. 2006 PMID: 16554755

Human: Protein Complex "Communities"

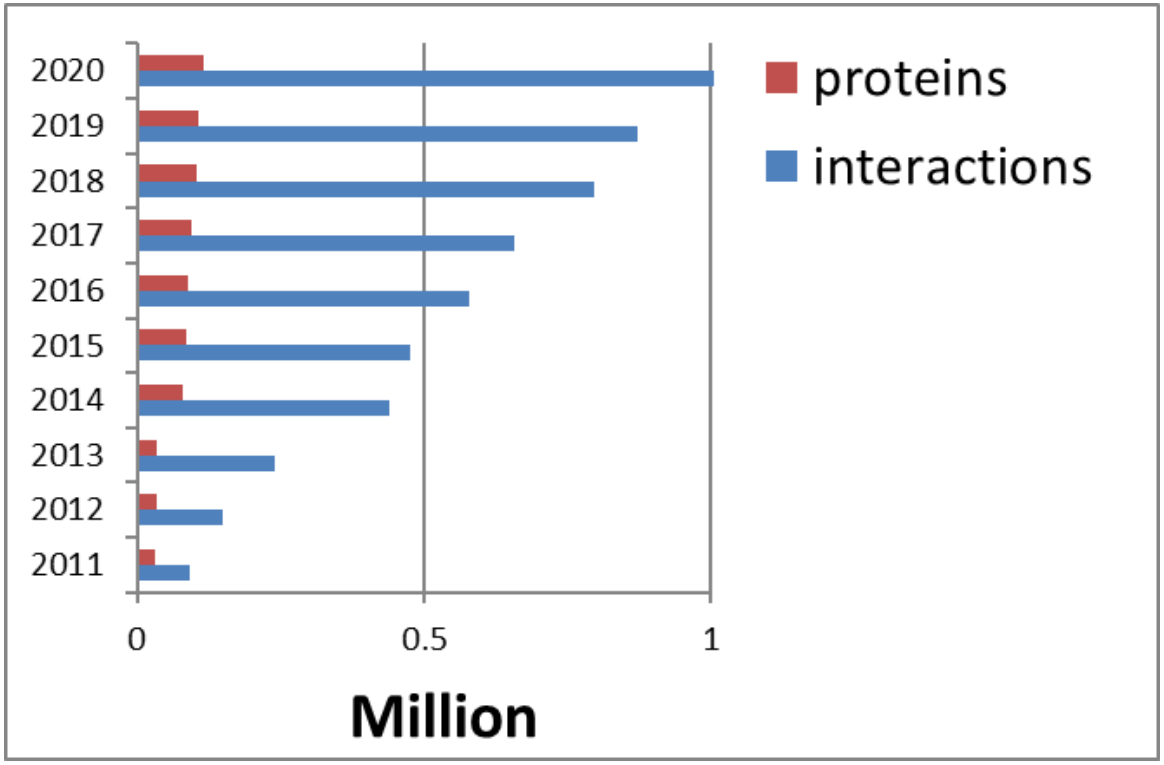


Huttlin et al, *Nature*. 2017 PMID: 28514442

Protein-Protein Interaction Databases



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



2023 Data Content

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

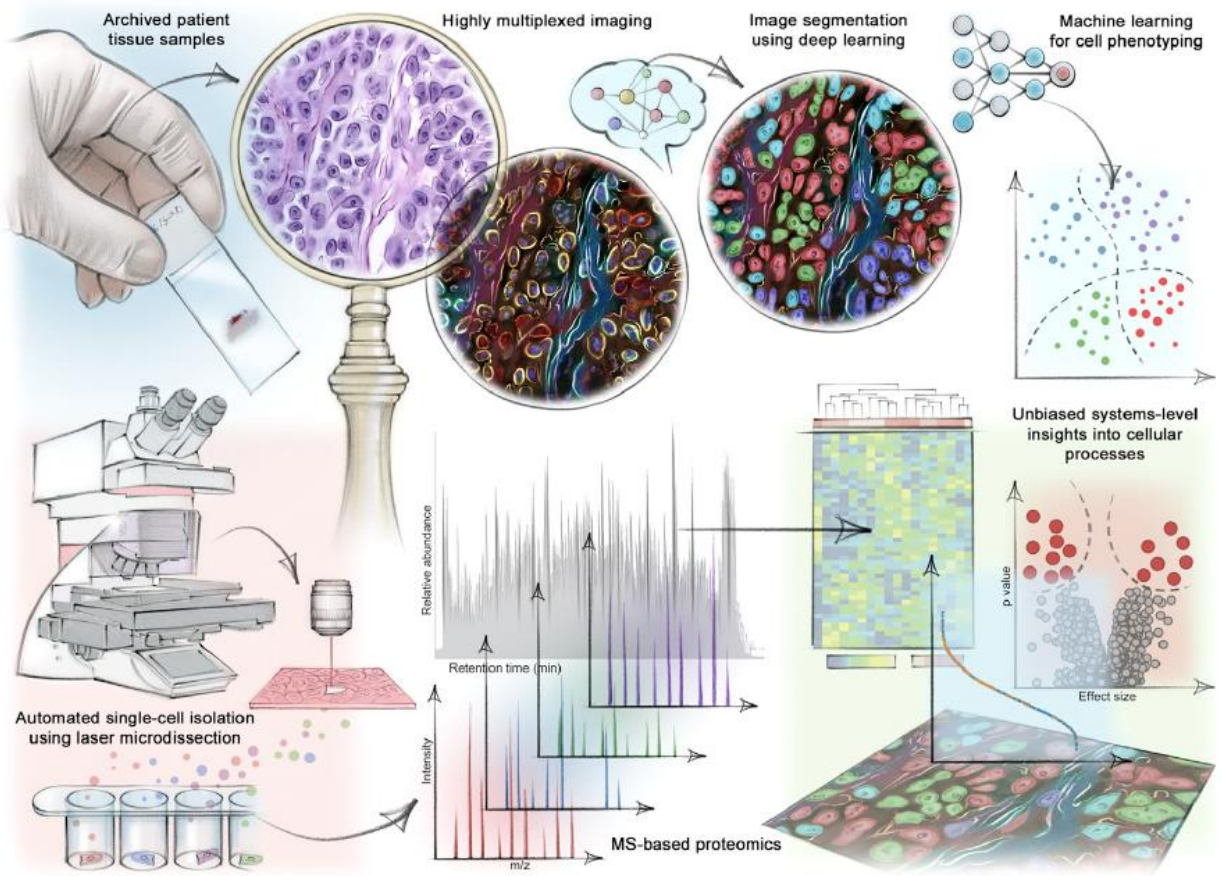
+ 39,393 interactions

+ 903 proteins

2022

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

Single Cell Proteomics & Spatial Proteomics



Molecular Cell

CellPress

Technology review

Unbiased spatial proteomics with single-cell resolution in tissues

PMID: 35714588

Andreas Mund,[1,4] Andreas-David Brunner,[2,3,4] and Matthias Mann[1,2,*]

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

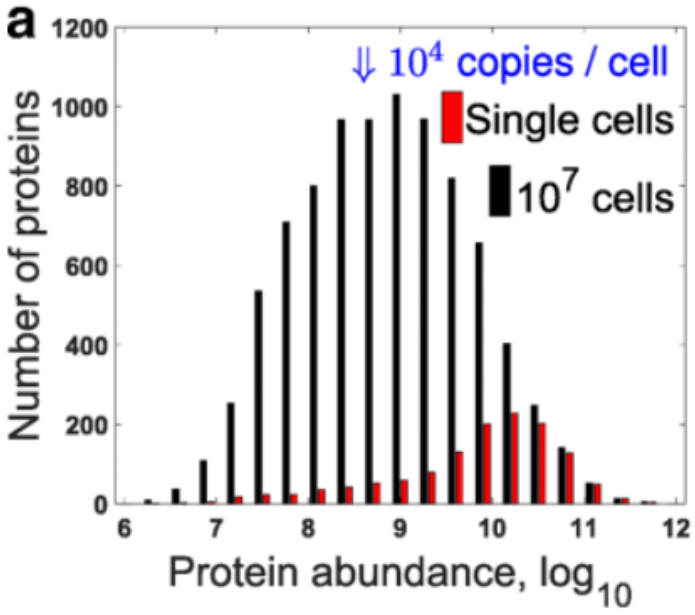
John R. S. Newman[1,2], Sina Ghaemmaghami[1,2]†, Jan Ihmels[1,2], David K. Breslow[1,2], Matthew Noble[1], Joseph L. DeRisi[1,3] & Jonathan S. Weissman[1,2]

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.

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

Increasing sensitivity in MS analysis to reach single cell proteomes



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

PMID: 30343672

Importantly, for our purposes, there has been a dramatic boost in sensitivity in just the last few years. In our own laboratory, for instance, the amount of sample needed to identify thousands of proteins in routine 1-h liquid chromatography-mass spectrometry (LC-MS) measurements has decreased more than 100-fold to the nanogram level (Beck et al., 2015; Meier et al., 2020).

Molecular Cell

Technology review

Unbiased spatial proteomics with single-cell resolution in tissues

Andreas Mund,[1,4] Andreas-David Brunner,[2,3,4] and Matthias Mann[1,2,*]

PMID: 35714588

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

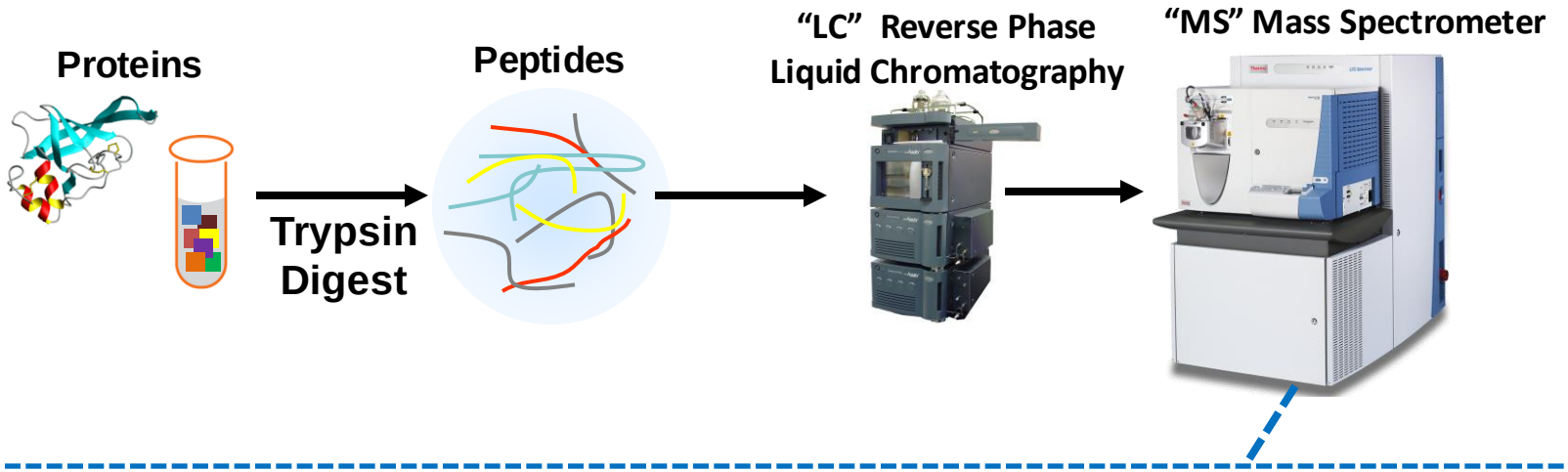
However:

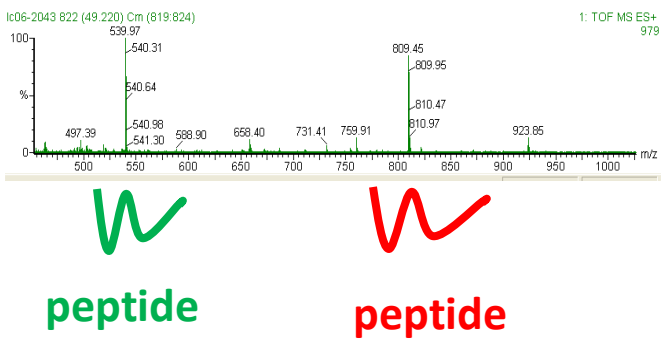
- Interaction data is largely static

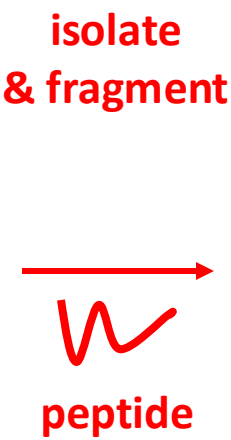
Next Step:

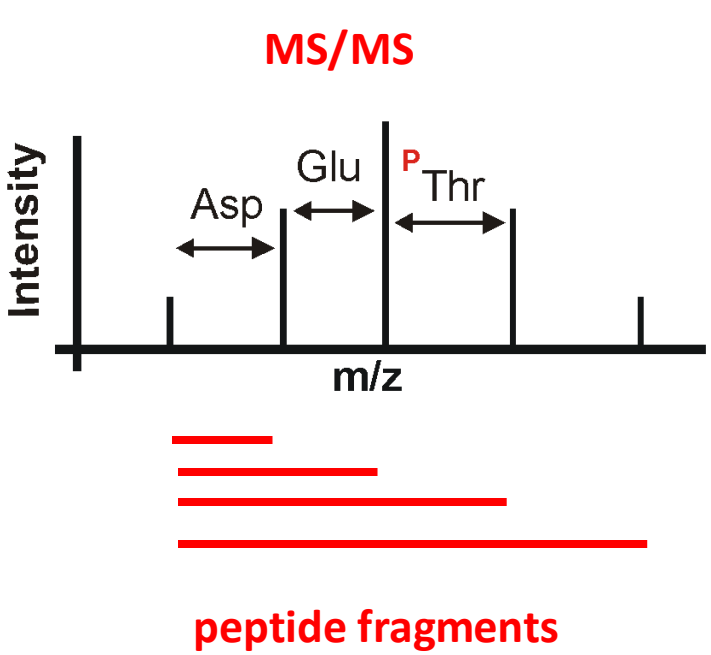
- How do protein interaction networks change over time?

Typical work flow for LC-MS "shotgun proteomics"



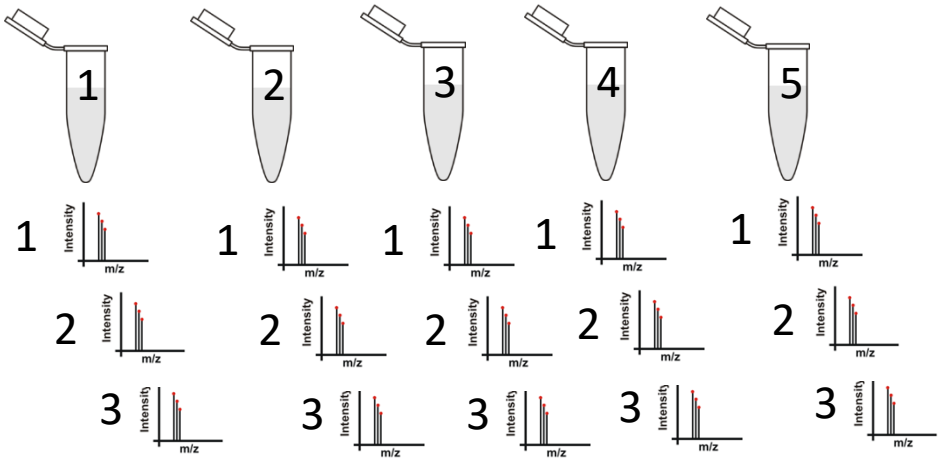




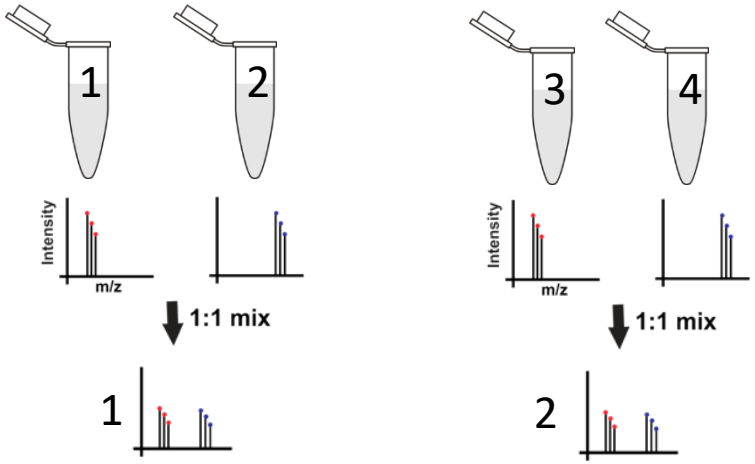


Multiple Techniques Enable Quantitative Proteomics

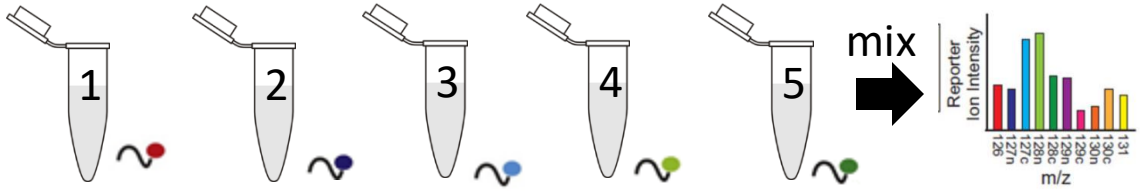
Label Free
-many, many replicates
-indirect quant



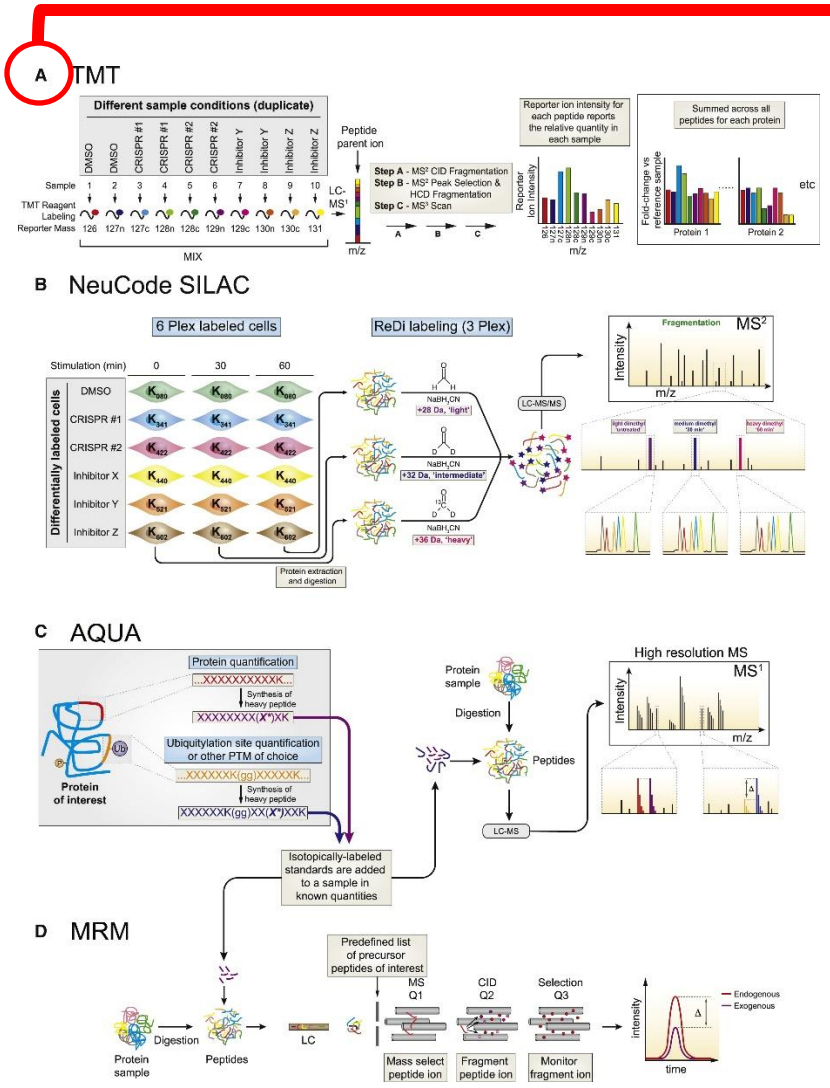
“Metabolic” Labeling
-fewer replicates
-multiplex
-direct quant

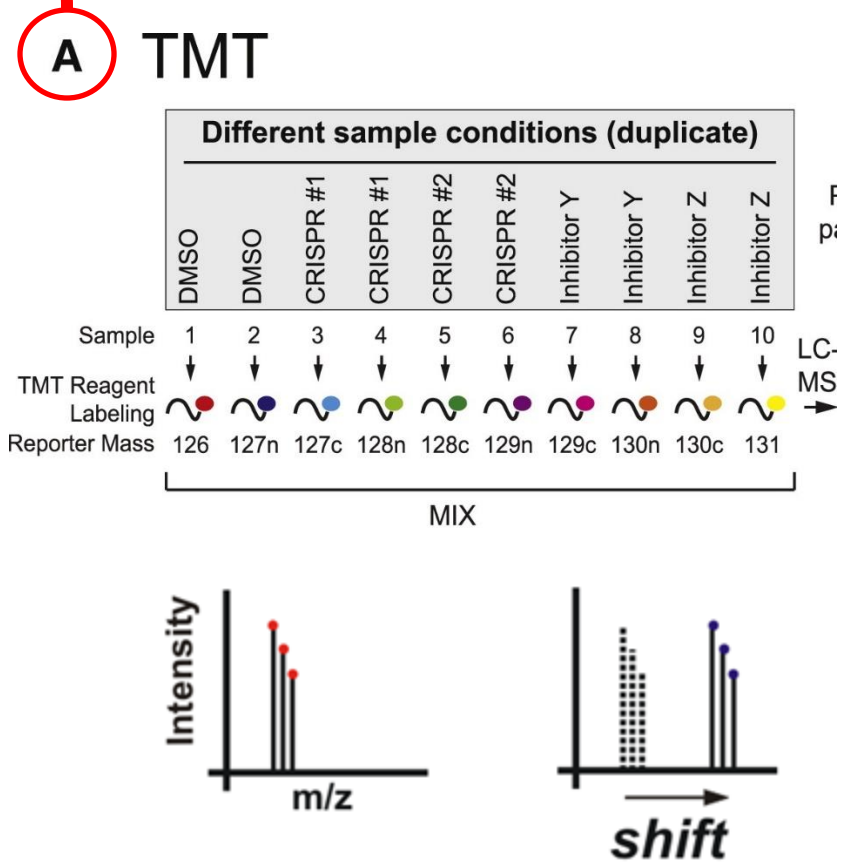


Barcoding
-increased multiplex
-direct quant



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





Quantifying ubiquitin signaling; Ordureau A, Münch C, Harper JW. 2015 **PMID: 26000850**

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

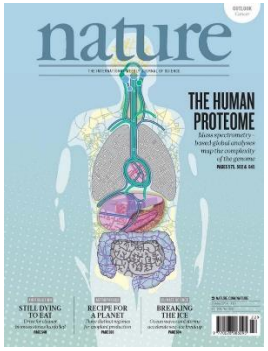
The Sequence of the Human Genome

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 Consortium*



2014

ARTICLE

doi:10.1038/nature13319

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]

ARTICLE

doi:10.1038/nature13302

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[3], Shobhit Jain[3], 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],

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

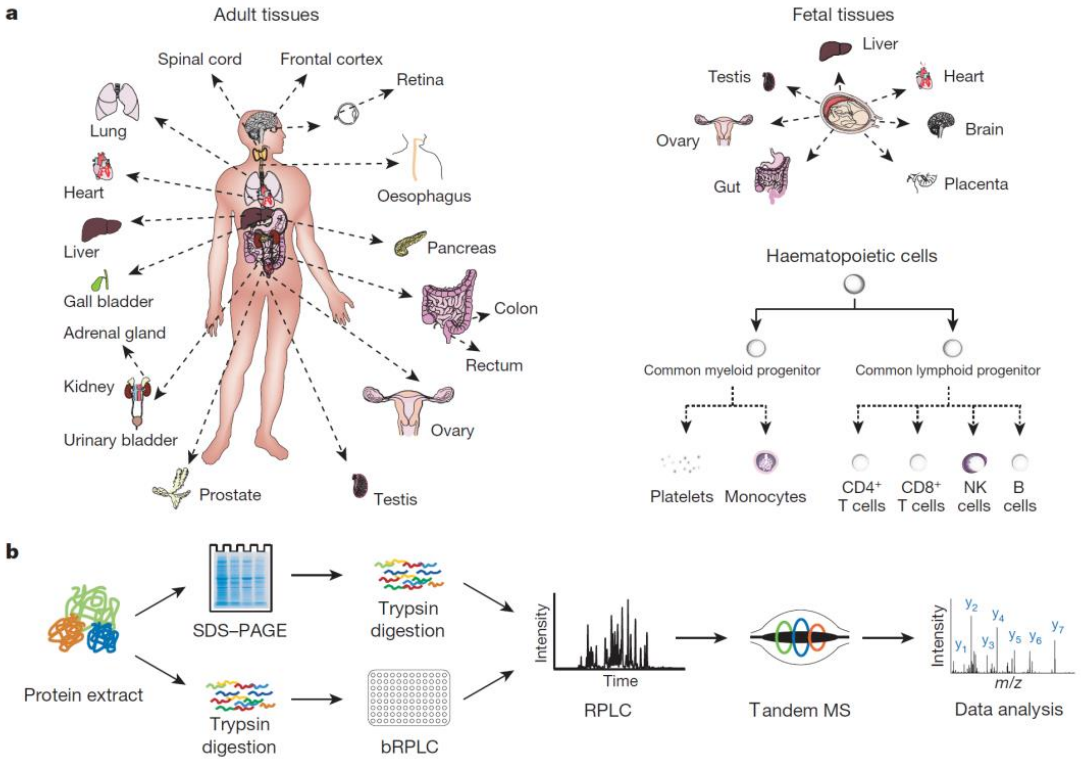
ARTICLE

doi:10.1038/nature13302

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],

- 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.



A draft map of the human proteome; Kim & Akhilesh Pandey et al., PMID: 24870542

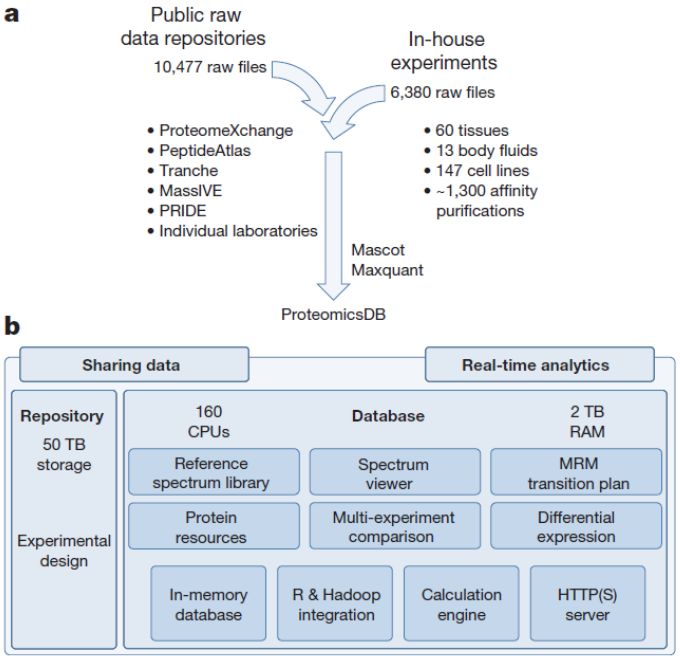
ARTICLE

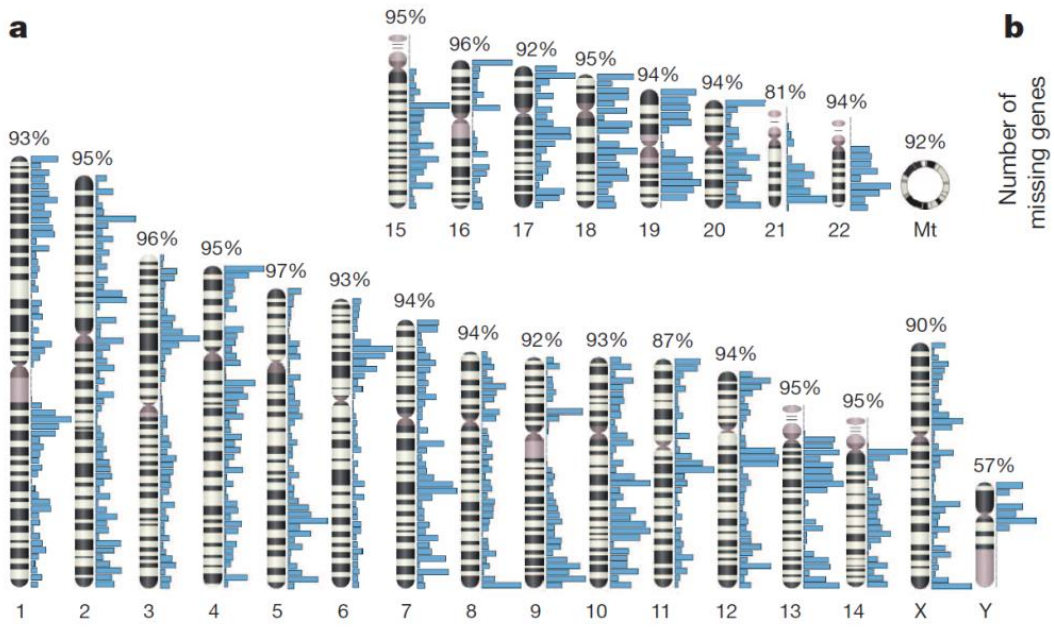
doi:10.1038/nature13319

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]

- 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; Wilhelm & Bernhard Kuster et al., **PMID: 24870543**

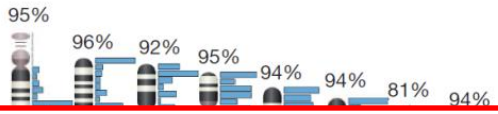
ARTICLE

doi:10.1038/nature13319

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]

- 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):

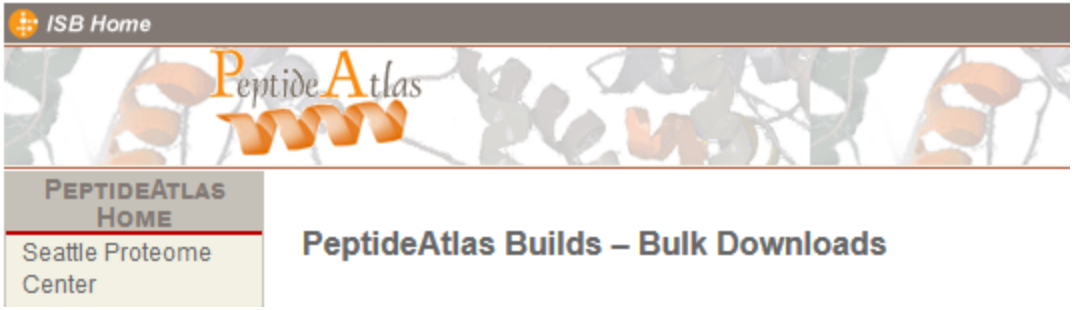
How long would it take to get the same data?

In 2001? ~61 years

In 2014? ~265 Days

Mass-spectrometry-based draft of the human proteome; Wilhelm & Bernhard Kuster et al., **PMID: 24870543**

Proteomics Databases: Peptide depositories



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

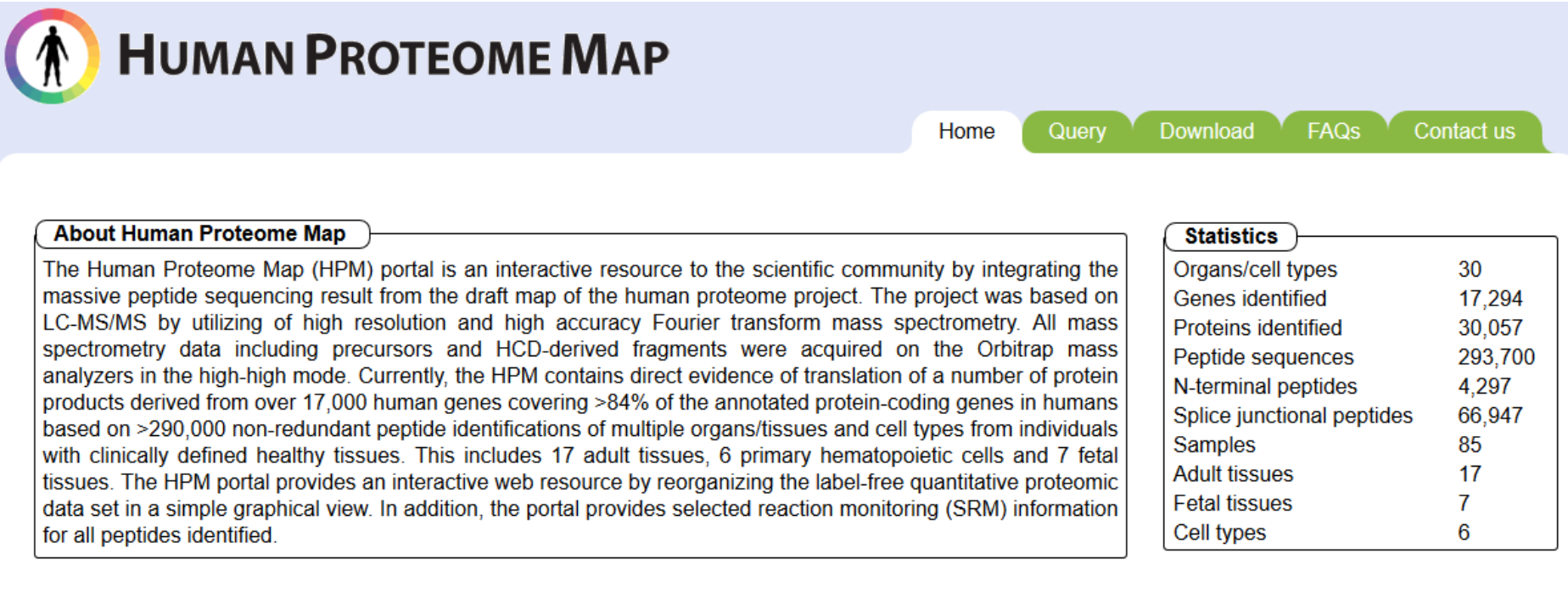
TaxID	Date	Number of Samples	Peptide Inclusion Cutoff	Number of Peptide-Spectrum Matches (PSMs)	Number of Distinct Peptides	Reference Database	Peptide Sequences	Peptide CDS Coordinates	Peptide CDS and Chromosomal Coordinates	Database Tables
9606	Mar 2015	1011	PSM FDR = 0.0002	133,638,335	1,025,698	Ensembl 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



ARTICLE

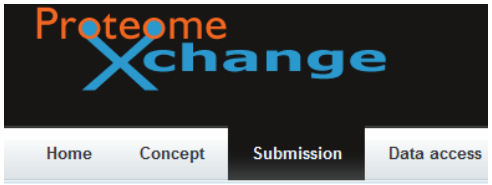
doi:10.1038/nature13302

A draft map of the human proteome

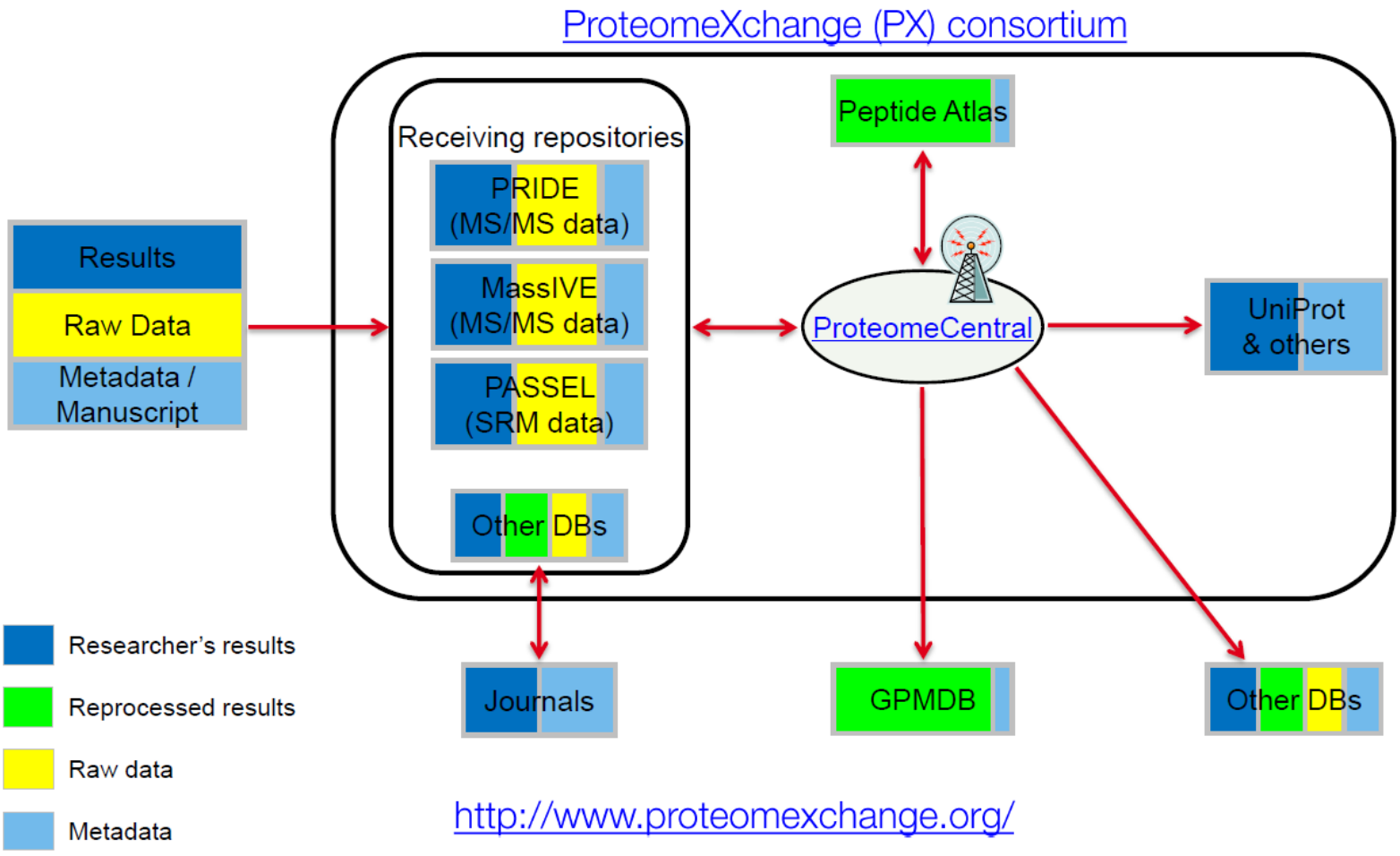
Min-Sik Kim[1,2], Sneha M. Pinto[3], Derese Getnet[1,4], Raja Sekhar Nirujogi[3], Srikanth S. Manda[3], Raghothama Chaerkady[1,2], Anil K. Madugundu[3], Dhanashree S. Kelkar[3], Ruth Isserlin[5], Shobhit Jain[5], Joji K. Thomas[3], Babylakshmi Muthusamy[3], Pamela Leal-Rojas[1,6], Praveen Kumar[3], Nandini 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],

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

Proteomics Databases: Integrated Resources

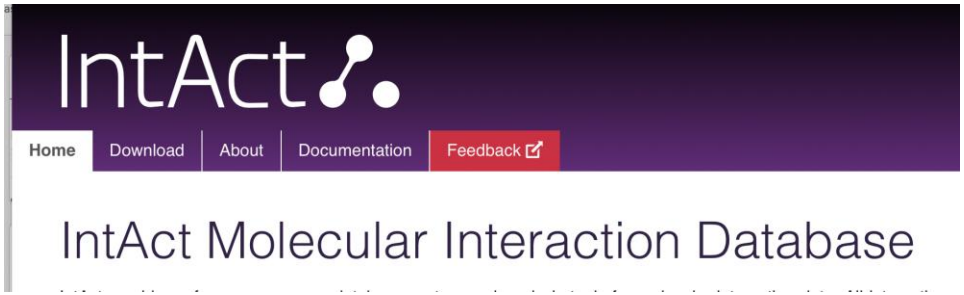


http://www.proteomexchange.org/

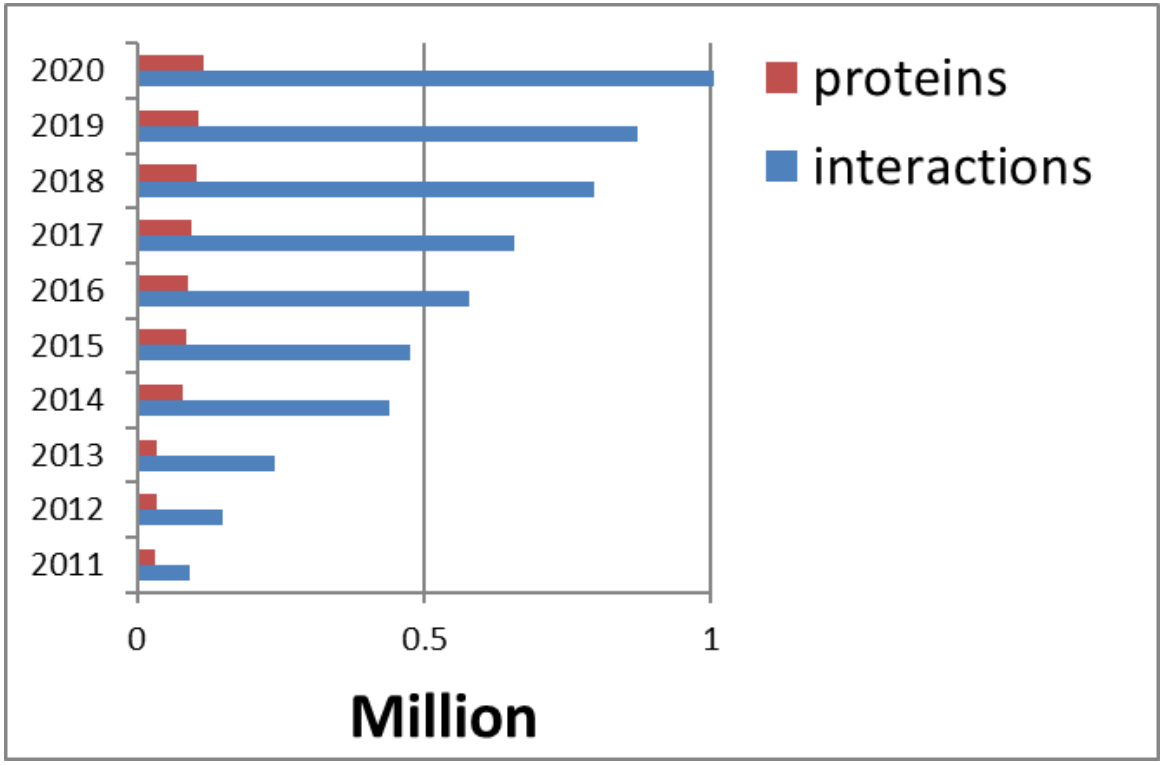


Slide modified from "Computational Mass Spectrometry-Based Proteomics 6th Maxquant Summer School" 21-25 July 2014
Emanuele Alpi, UniProt and PRIDE Development

Protein-Protein Interaction Databases



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



2023 Data Content

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

+ 39,393 interactions

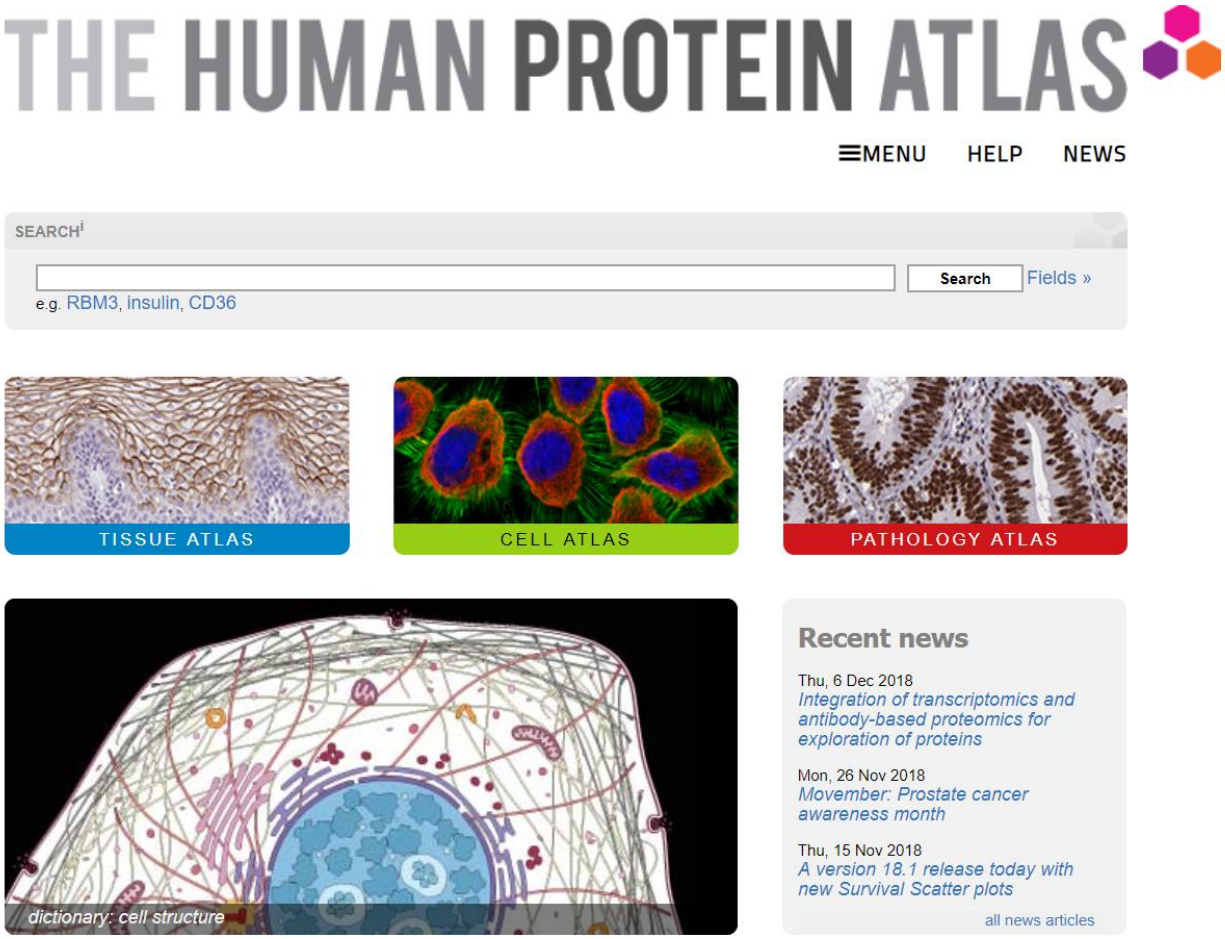
+ 903 proteins

2022

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

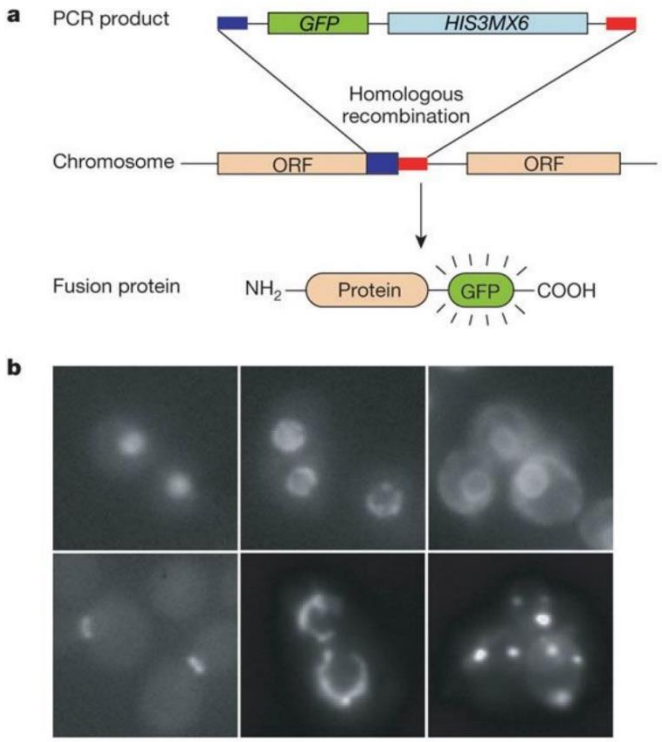
Proteomics Databases: Integrated Resources Beyond Mass Spectrometry

http://www.proteinatlas.org/



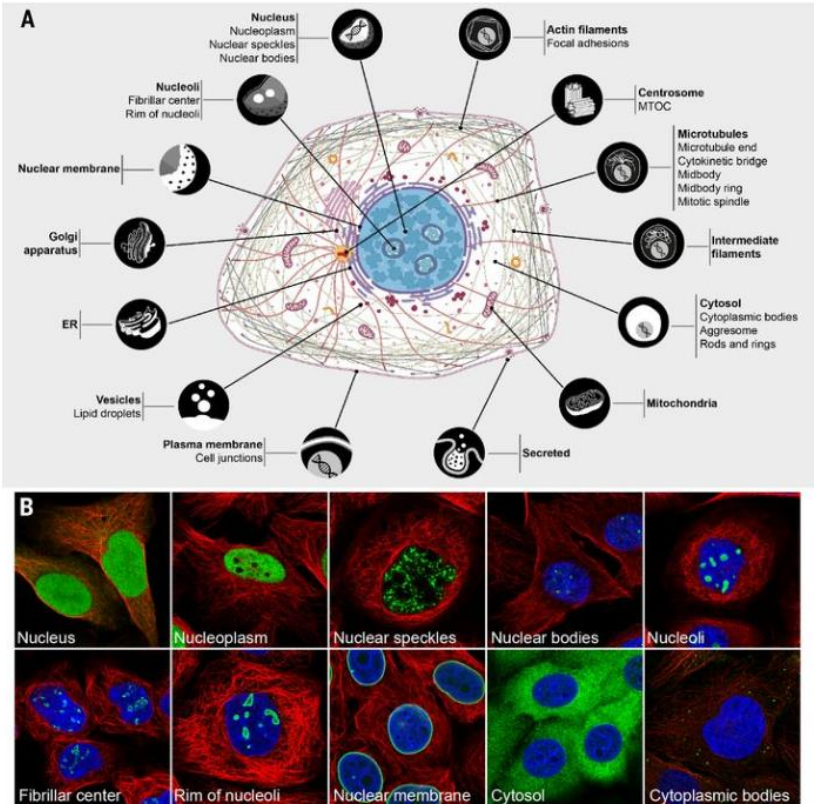
Proteomics Databases: Integrated Resources Beyond Mass Spectrometry

>4,000 GFP-Gene Fusions



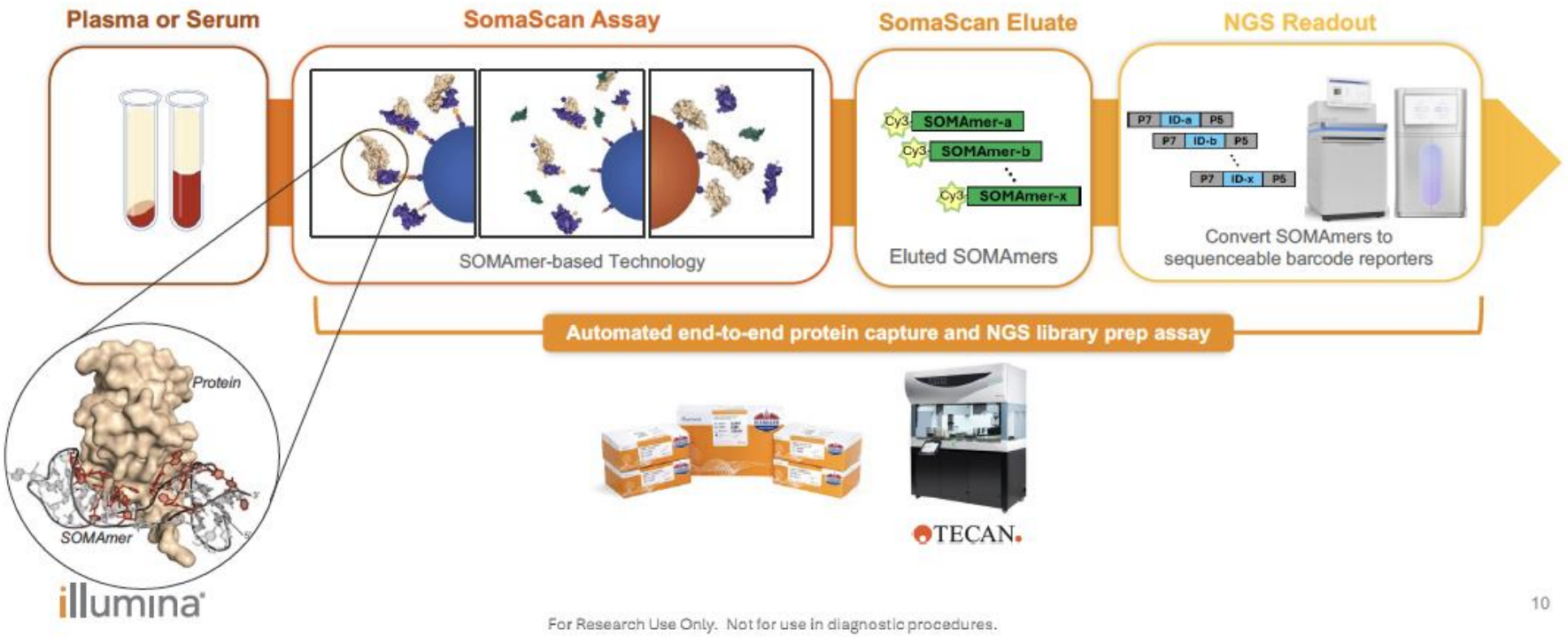
Huh et al., Global analysis of protein localization in budding yeast. Nature. 2003 PubMed:14562095

>13,000 Antibodies



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

Proteomics Technology: Beyond Mass Spectrometry



nature medicine

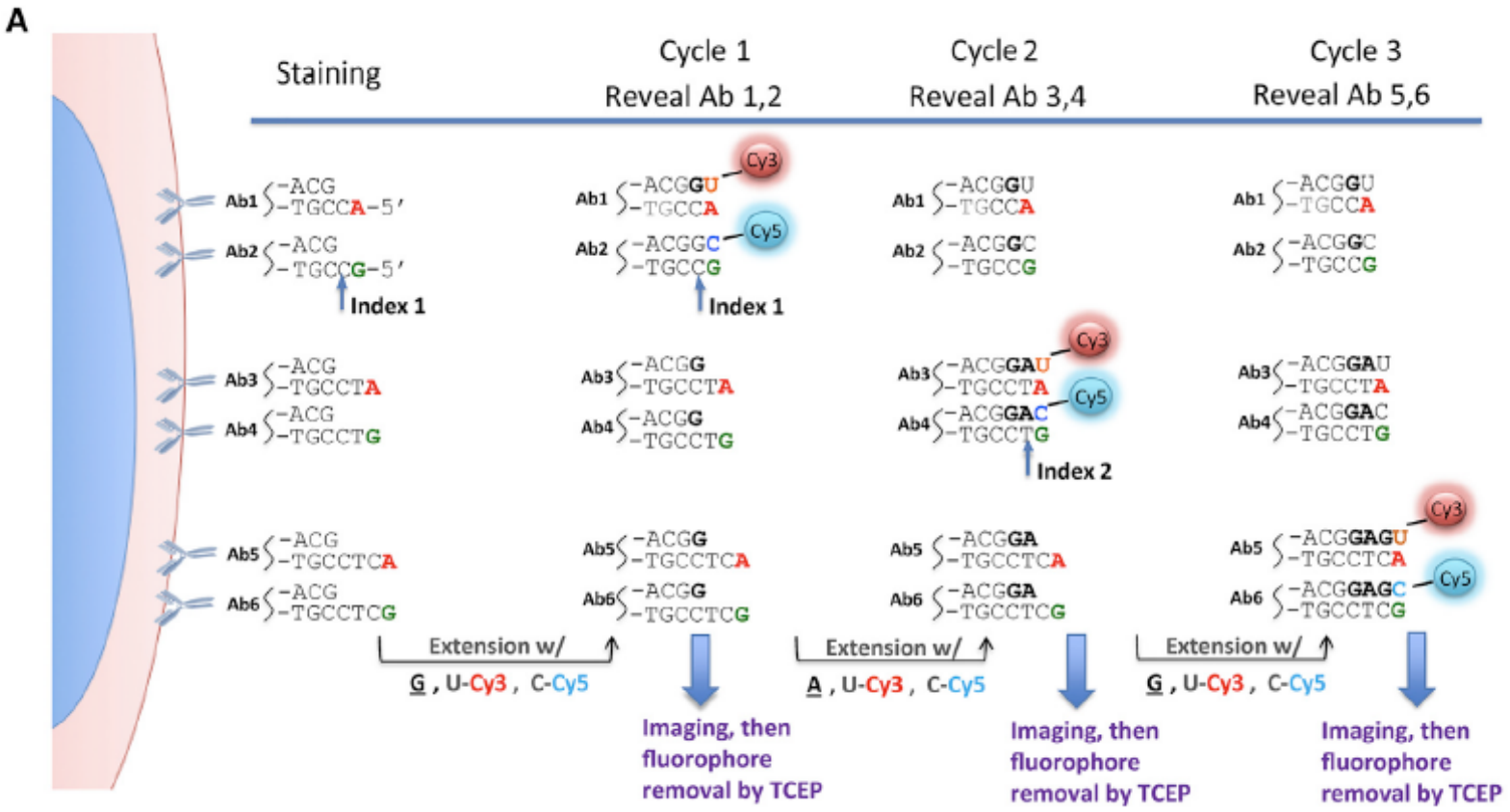
Article

https://doi.org/10.1038/s41591-024-03092-6

A unified metric of human immune health

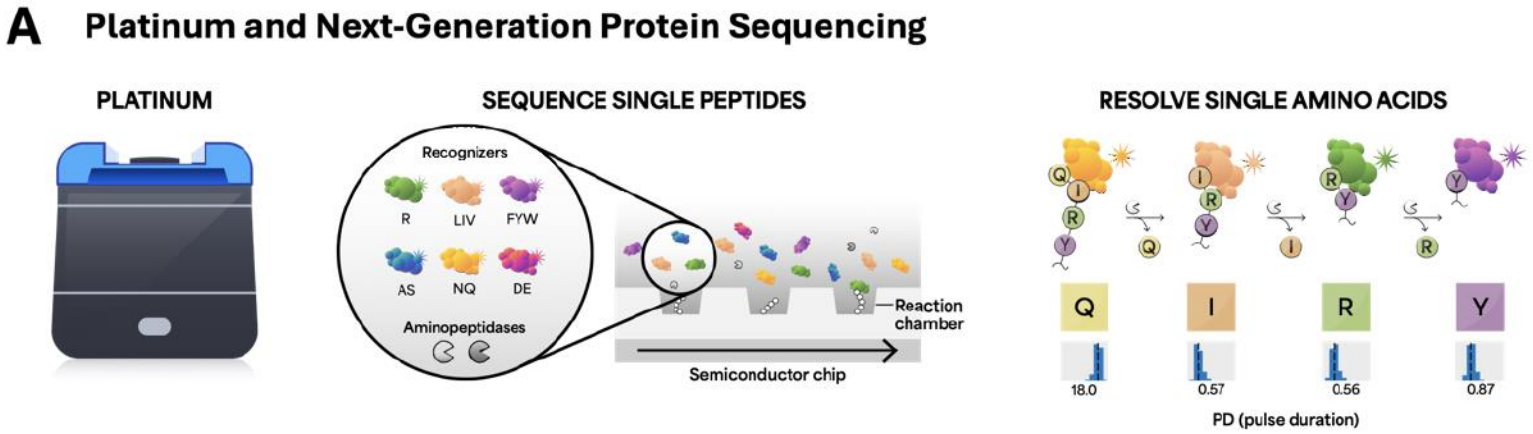
Sparks R, et al. Nat Med. 2024. PMID: 38961223

Proteomics Technology: Beyond Mass Spectrometry



Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging Cell 2018. Yury Goltsev et al. PMID: 30078711

Proteomics Technology: Beyond Mass Spectrometry



Protein Barcoding and Next-Generation Protein Sequencing for Multiplexed Protein Selection, Analysis, and Tracking

Mathivanan Chinnaraj, et al
https://www.biorxiv.org/content/10.1101/2024.12.31.630920v1