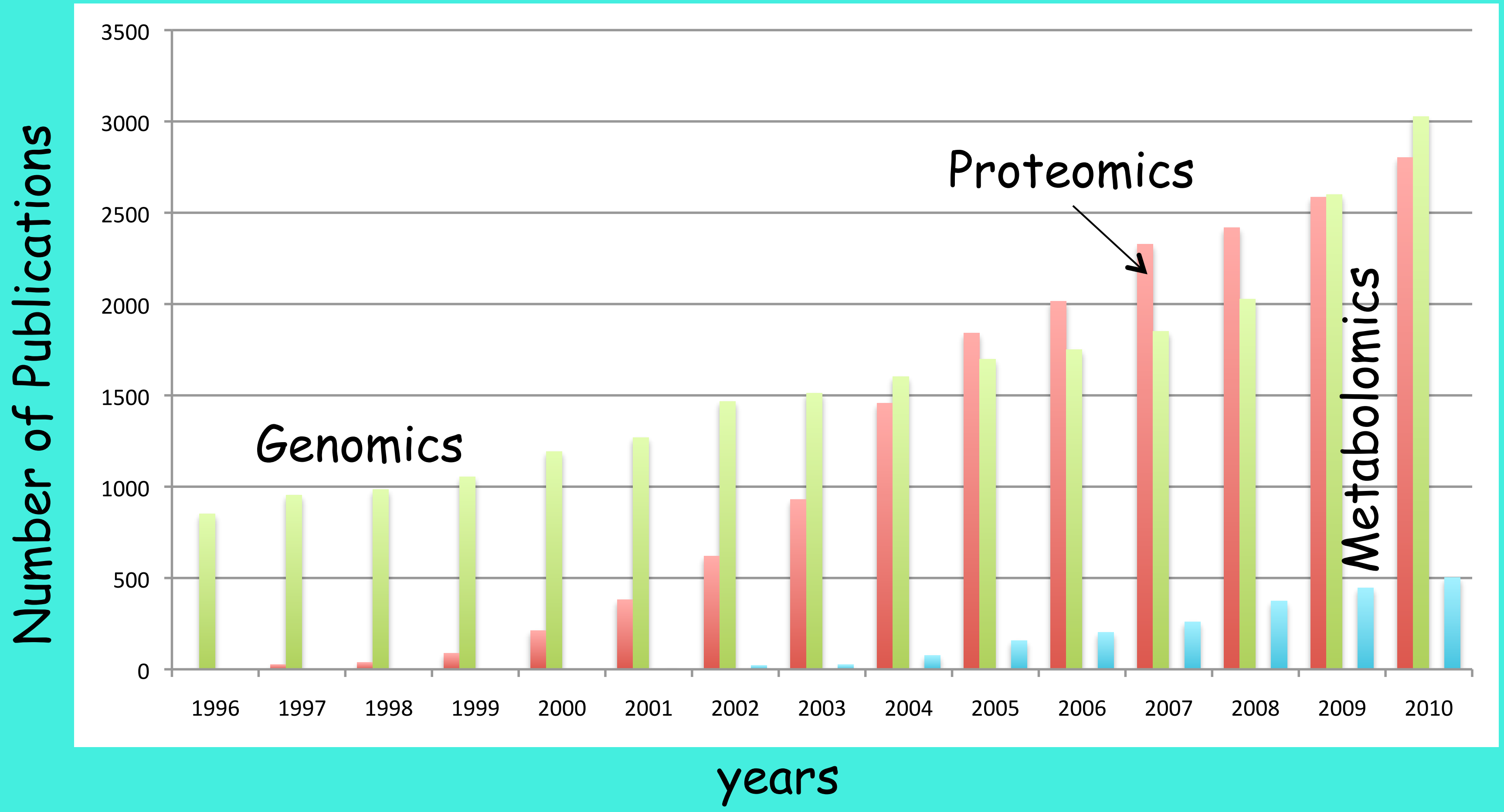
Pathway Analysis
In Expression Proteomics

Roman Zubarev

Roman.Zubarev@ki.se

Molecular Biometry,
Department for Medical Biochemistry & Biophysics,
Karolinska Institutet, Stockholm

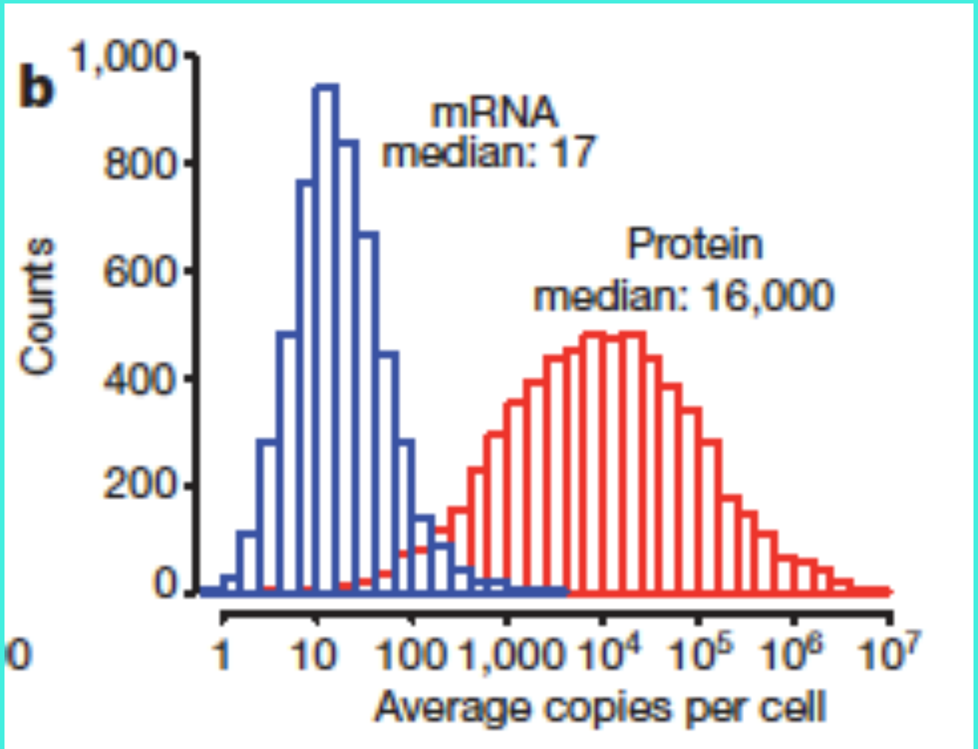
Proteomics vs Transcriptomics and Metabolomics

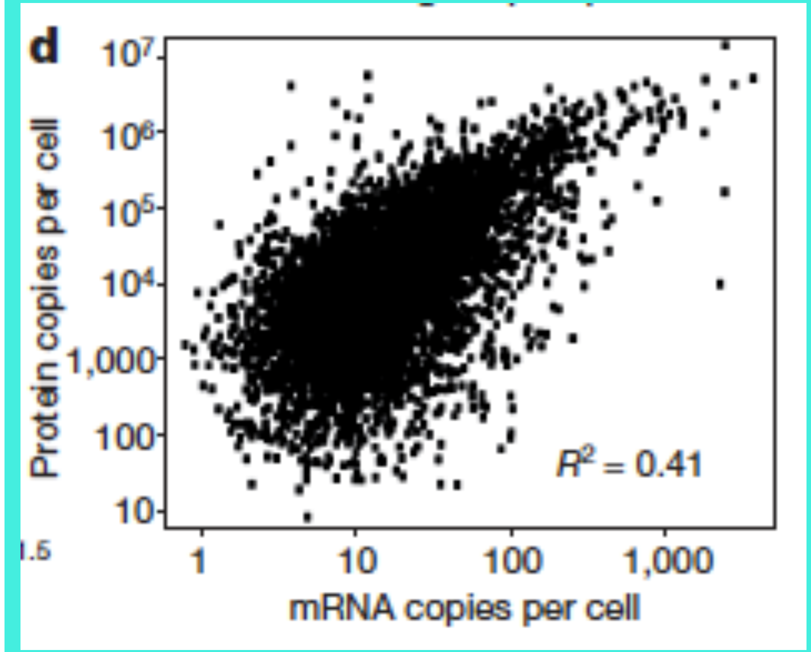


Genomics – what the cell *may* do
Transcriptomics – *wants* to do
Proteomics – *does*
Metabolomics – *has* done

Differences between transcriptomics and proteomics

- The dynamic range – 10^3-10^4 vs 10^7.



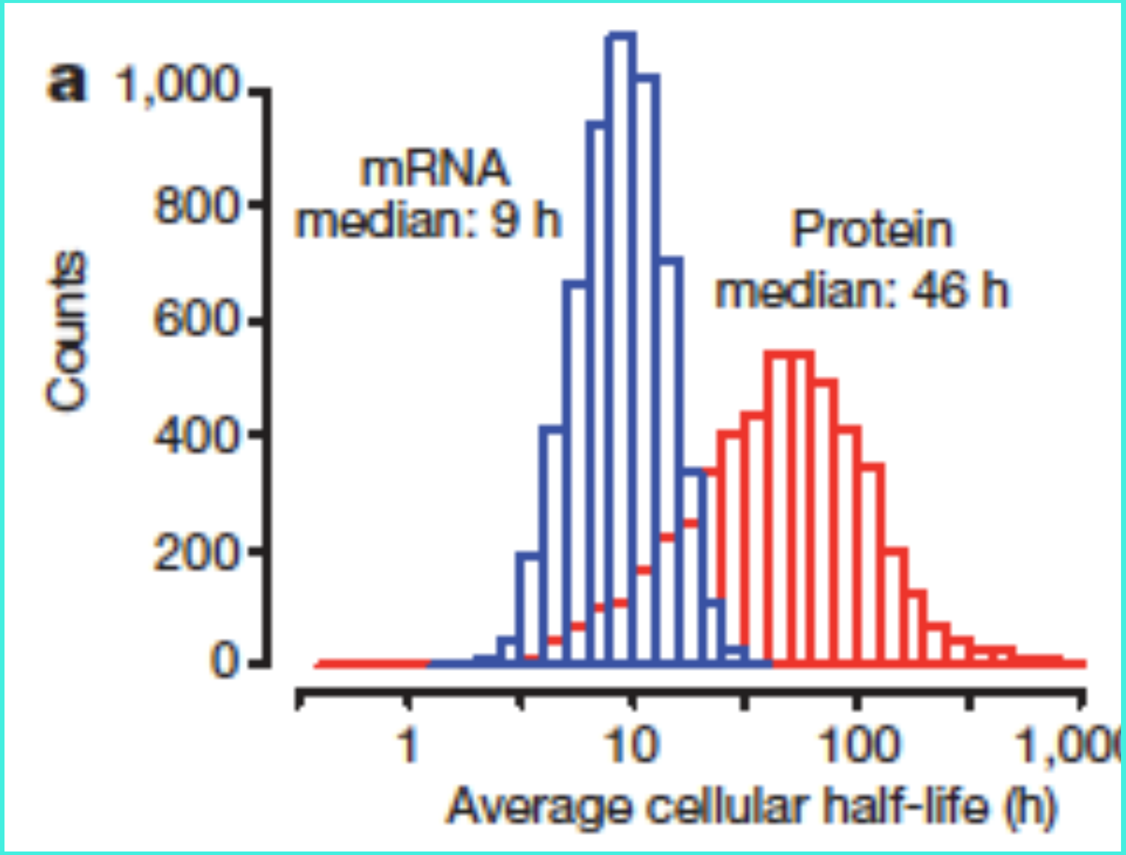


19 MAY 2011 | VOL 473 | NATURE | 337

Since the dynamic range of instrumentation is – 10^3-10^4 , transcriptomics easily covers all 10,000 expressed genes, while proteomics – ca. 5,000 proteins. But false discovery rate for mRNA 5%, for proteins – 1%

Differences between transcriptomics and proteomics

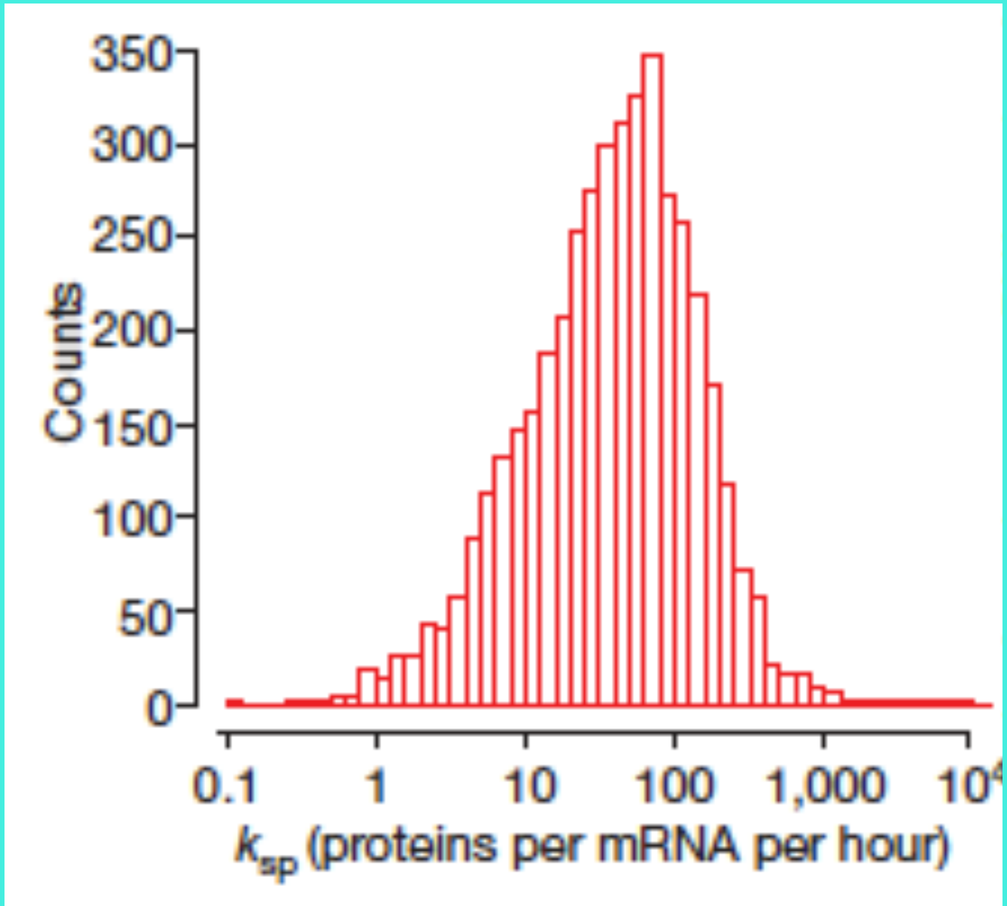
- The cellular half-life:
  - mRNA – 9h
  - proteins – 46 h



19 MAY 2011 | VOL 473 | NATURE | 337

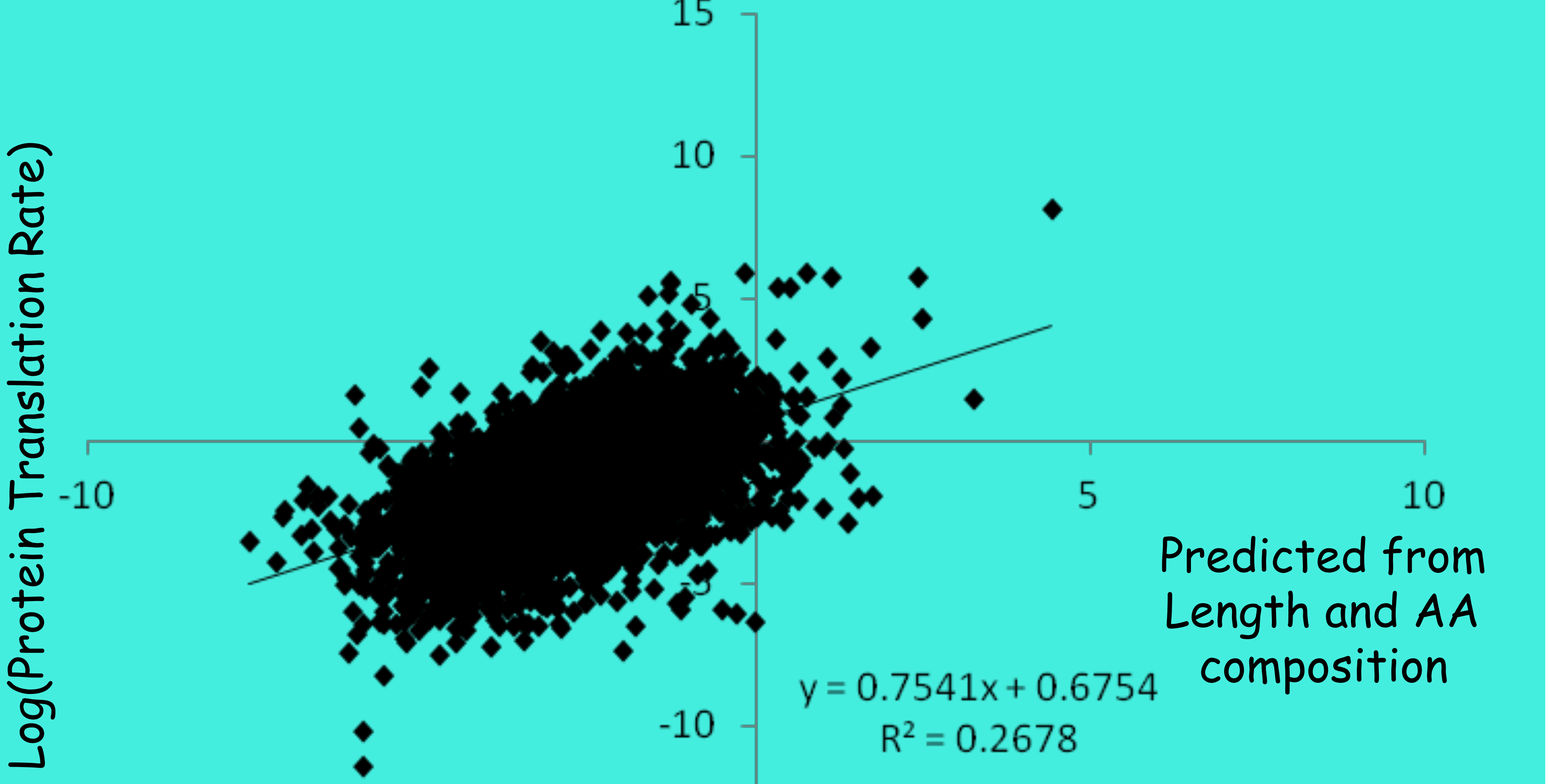
Differences between transcriptomics and proteomics

- The number of protein molecules per mRNA: 1:1 to 1000:1



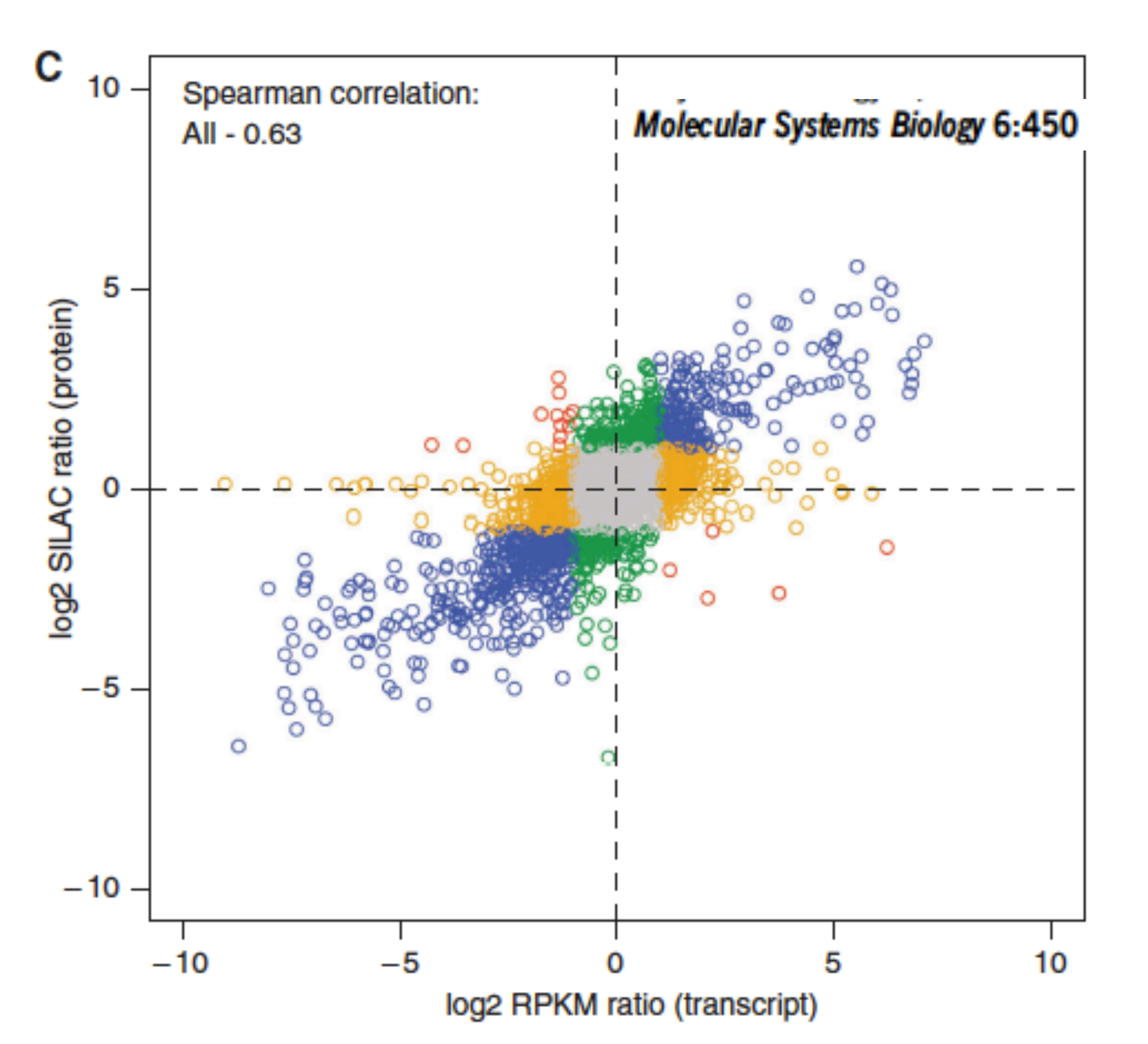
19 MAY 2011 | VOL 473 | NATURE | 337

Combined Predictions – Length and AA Score



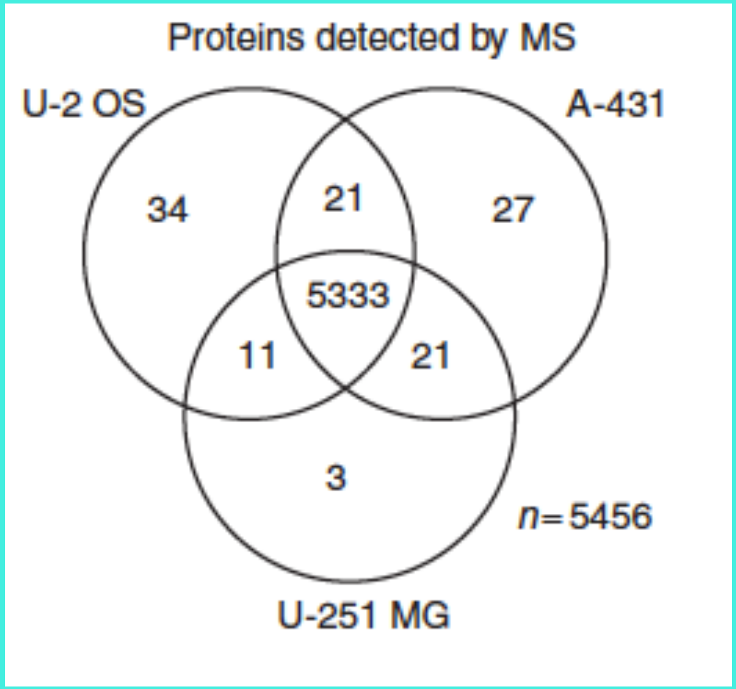
Other factors contribute to translation rate!

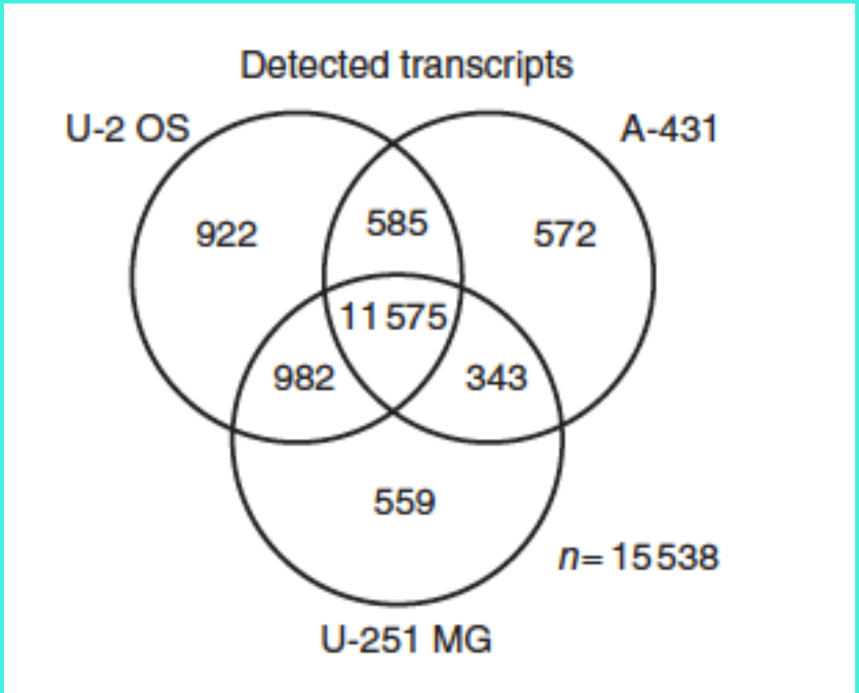
- mRNA abundances predict ca. 40% of the protein abundance, but log(Ratio) for mRNA predict >60% of log(Ratio) for proteins



mRNA data need to be complemented by Proteomics data

In three different cell lines, practically all expressed genes (and proteins) are shared

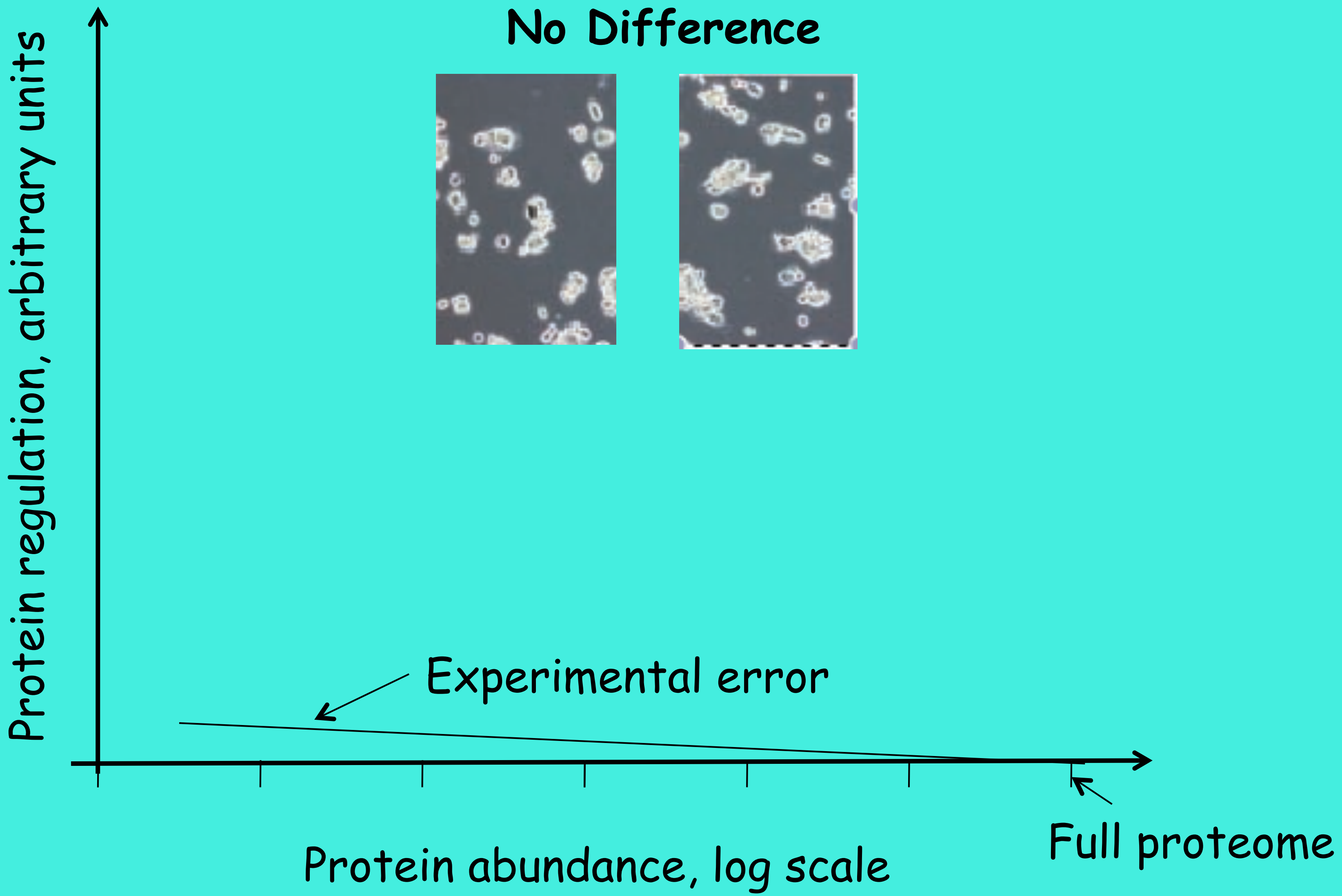




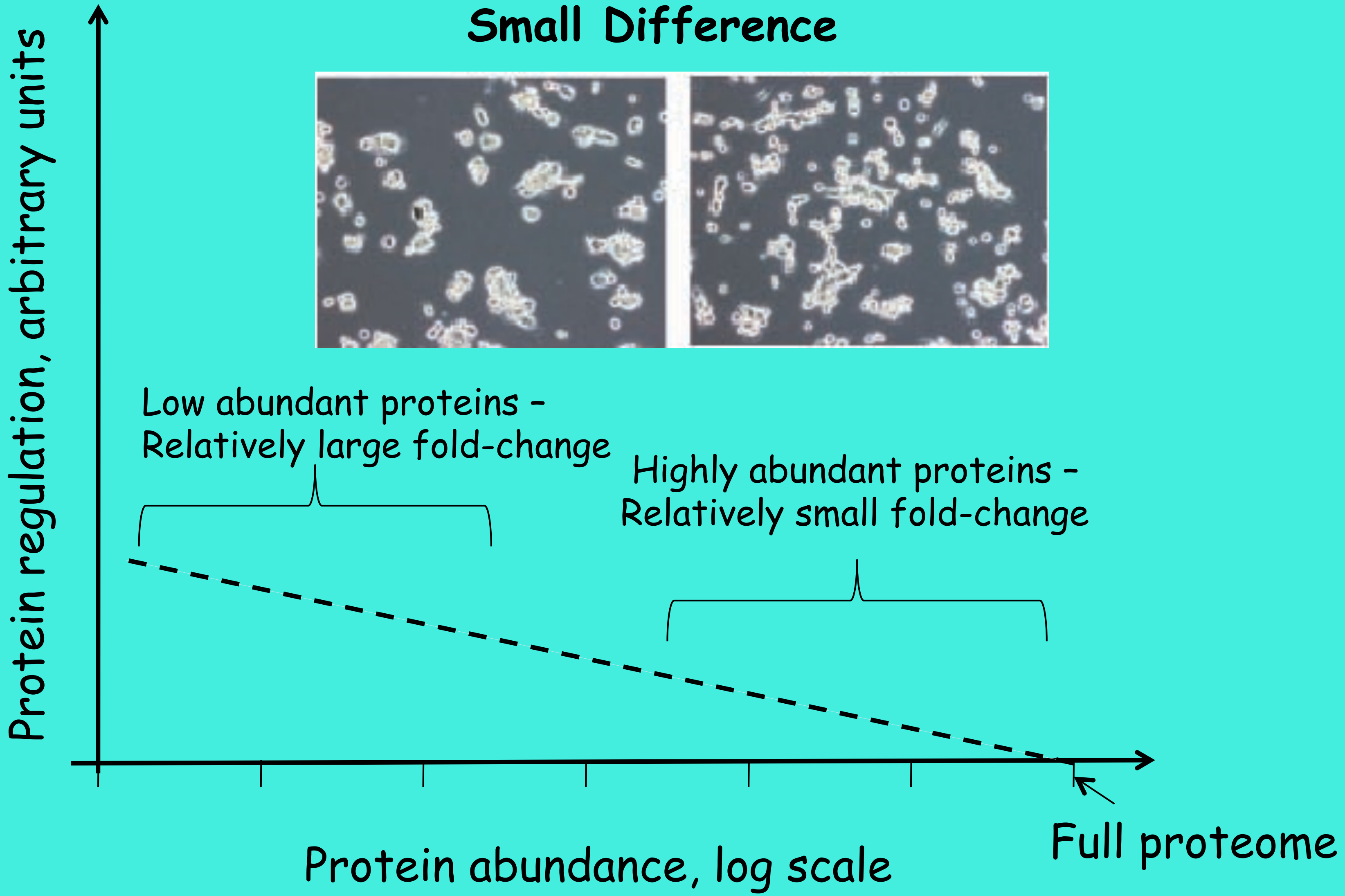
Same proteins are expressed in every cell type, but with different abundances

Molecular Systems Biology 6:450

How does protein regulation depend upon protein abundance?


No Difference
Protein regulation, arbitrary units
Experimental error
Protein abundance, log scale
Full proteome

How does protein regulation depend upon protein abundance?



Small Difference

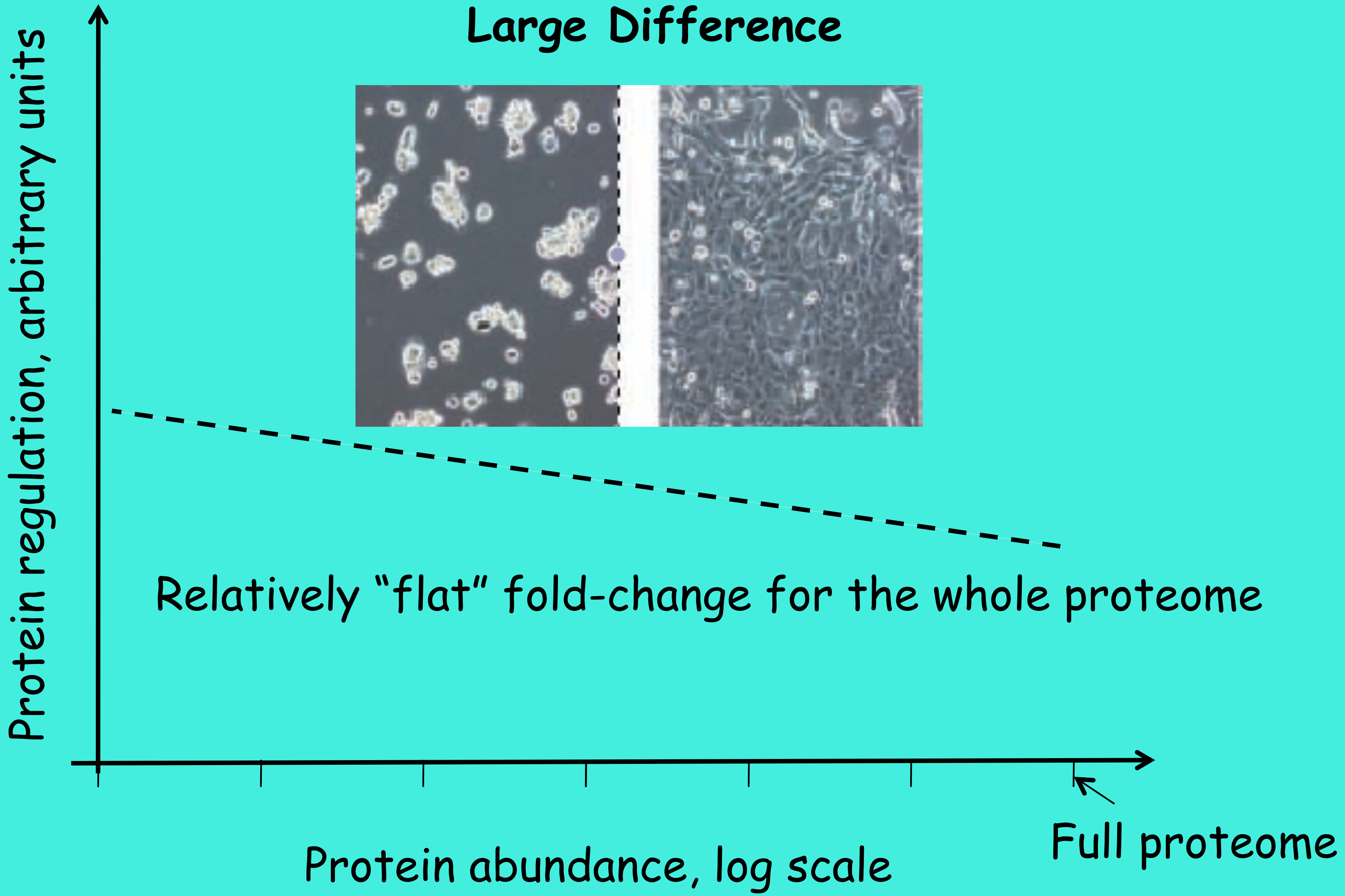
Low abundant proteins –
Relatively large fold-change

Highly abundant proteins –
Relatively small fold-change

Full proteome

Protein regulation, arbitrary units
Protein abundance, log scale

How does protein regulation depend upon protein abundance?


Protein regulation, arbitrary units
Protein abundance, log scale
Full proteome

Large Difference

Relatively “flat” fold-change for the whole proteome

SUMMARY

- Transcriptomics provides large (95%) coverage of expressed genes, but it explains, at best, only 40% of the log(Ratio) of protein abundances.

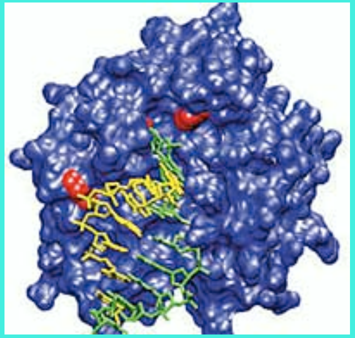
- Proteomics gives lower coverage (50% or less) by expressed proteins, but false discovery rate is only 1%

- For small changes in the proteome (e.g. early stages in time course) , **deep** proteomics is advantageous, as proteins with significant fold-change are those of low-abundance

- For large changes in the proteome (e.g. cell type differentiation), even limited depth proteomics can provide specific fingerprint of cellular state, as protein regulation is largely independent upon abundance

Data Processing in Proteomics

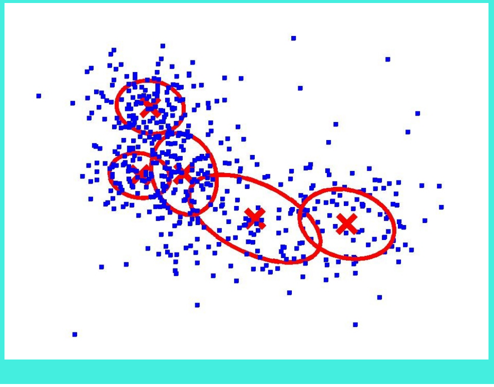
Reductionist Molecular Biology:



"golden bullet"

- detailed interactions, modifications, mechanisms
- lack of total picture

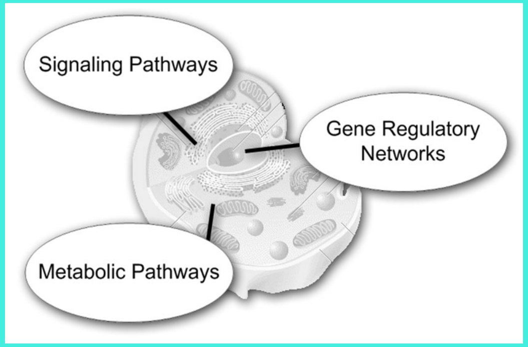
Statistical Approach:



Ad hoc, empirical model

- You get what you see
- Prediction, accuracy
- No explanation

Pathway Biology:



Global model

- prediction based on known pathways
- unknown accuracy
- do pathways exist?...

Protein Identification by Tandem Mass Spectrometry

Protein sequence

ILNKPEDETHLEAQPTDASAQFIRNLQISNE
DLSKEPSISREDLISKEQIVIRSSRQPQSQNPK
LPLSILKEKHLRNATLGSEETTEHTPSDASTT
EGKLMELGHKIMRNLENTVKETIKYLKSLF
SHAFEVVKT

Enzymatic → digest

Tryptic peptides

EDLISK
EQIVIR
LPLSILK
NLENTVK
LMELGHK
QPQSQNPK
NLQISNEDLSK
SLFSHAFEVVK
NATLGSEETTEHTPSDASTTEGK
ILNKPEDETHLEAQPTDASAQFIR

Tandem Mass Spectrometry (MS/MS)

Tryptic peptide

NLENTVK

MS/MS →

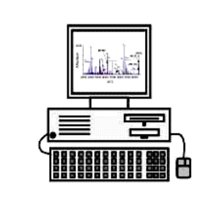
Fragmentation

N L E N T V K

Molecular mass: 817.44

Fragment masses

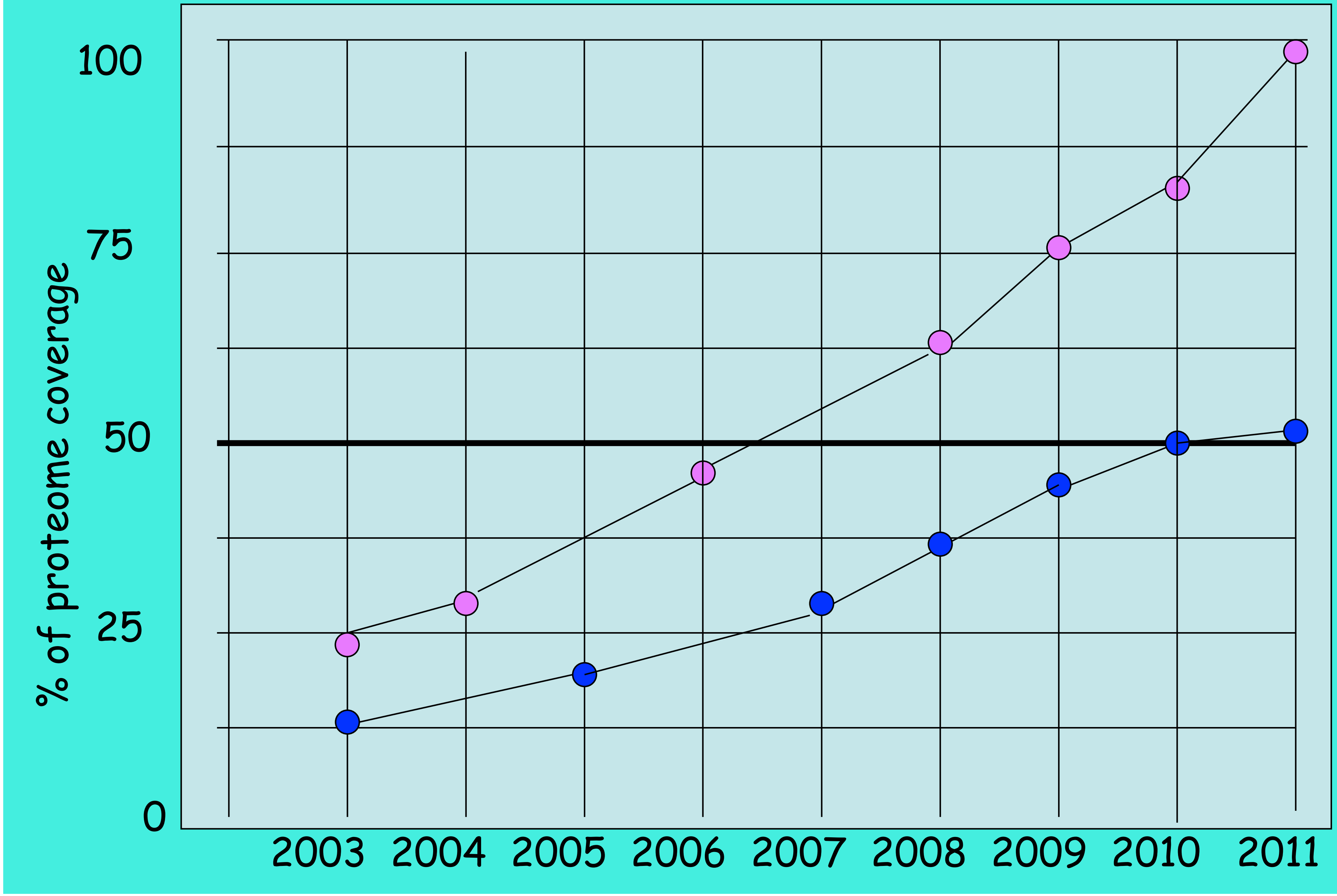
232.17
346.22
388.20
444.28
484.33
511.37
555.40
623.45
666.44
712.52



Your Peptide/ protein is this:

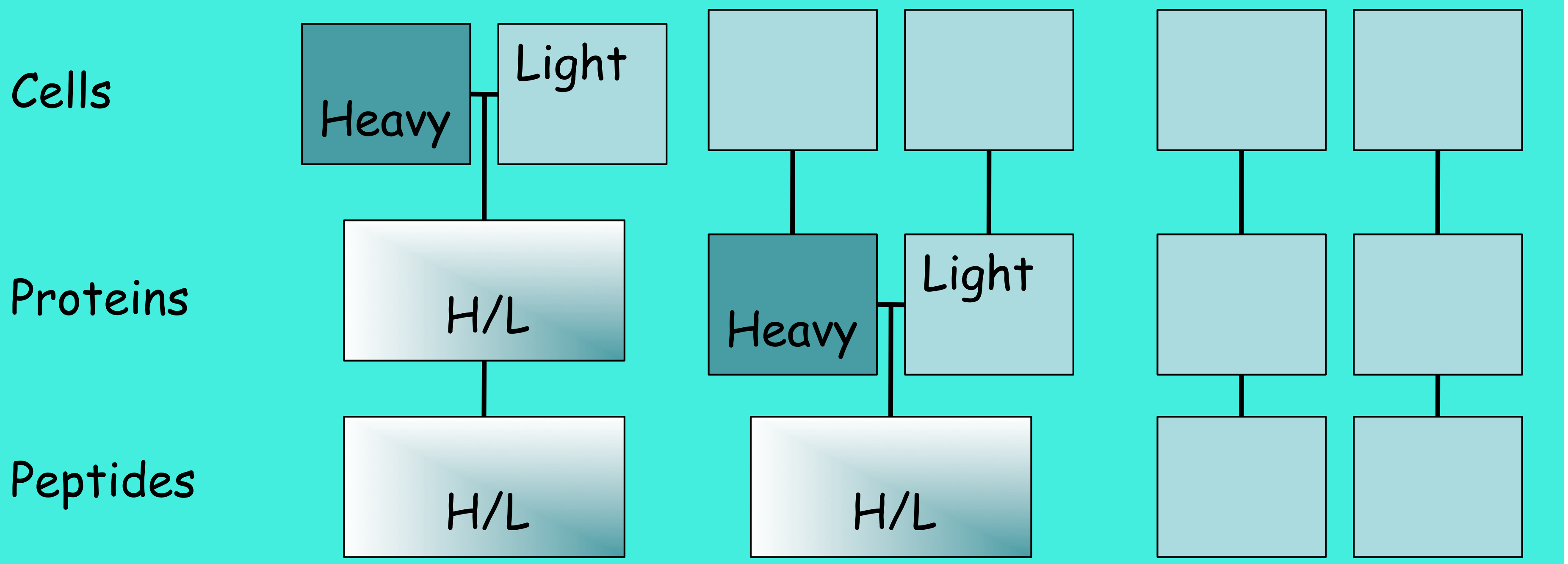
Score = 77

"Deep" vs "Top" Proteomics



MS-based quantitative discovery approaches

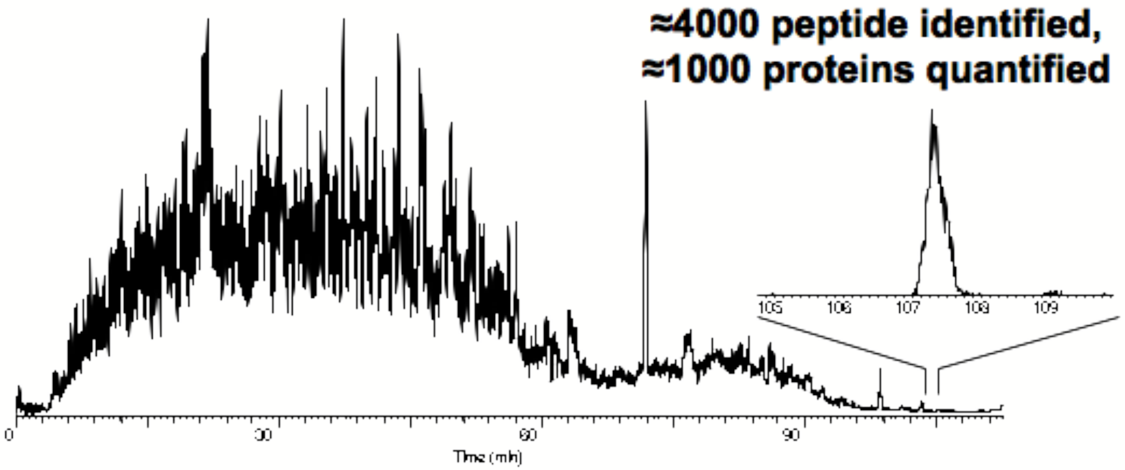
SILAC TMT, iTRAQ Label-free



	SILAC	TMT, iTRAQ	Label-free
Biology	?	0	0
Extraction	0	x	x
Digestion	0	x	x
Tissues, bodily fluids	N/A	N≤8	N-any

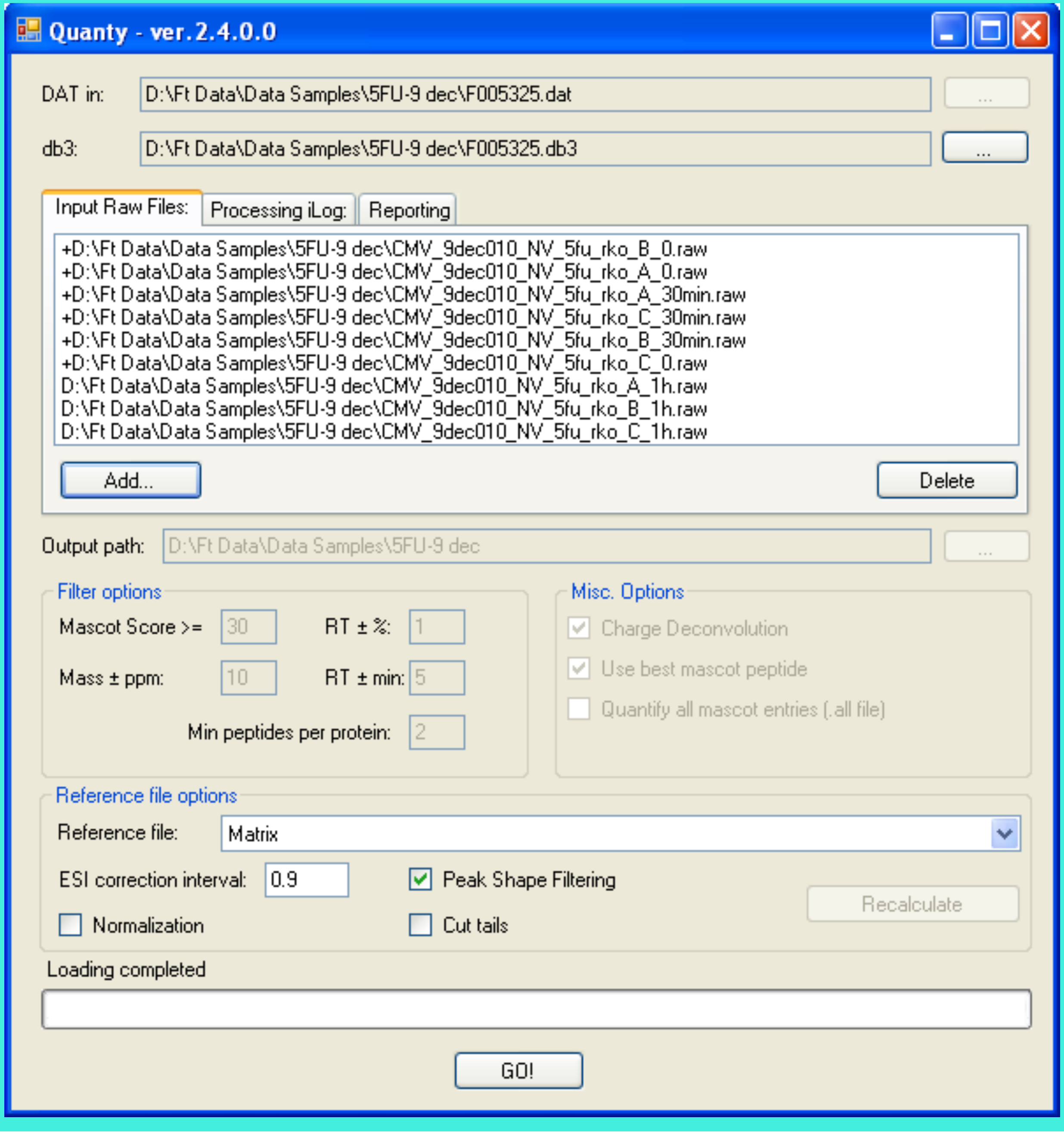
Deep proteomics **Top proteomics**

Top Proteomics


≈4000 peptide identified,
≈1000 proteins quantified

- 'Top proteome' : 1500-3000 proteins, 5000-9000 peptides
- No protein separation
- No peptide separation (on-line reverse-phase LC only)
- Single LC/MS experiment, 0.5-2.0 h long

Quanti 2.4 – February 2011 (2.5 – Feb 2012)



DAT in: D:\Ft Data\Data Samples\5FU-9 dec\F005325.dat

db3: D:\Ft Data\Data Samples\5FU-9 dec\F005325.db3

```
+D:\Ft Data\Data Samples\5FU-9 dec\CMV_9dec010_NV_5fu_rko_B_0.raw
+D:\Ft Data\Data Samples\5FU-9 dec\CMV_9dec010_NV_5fu_rko_A_0.raw
+D:\Ft Data\Data Samples\5FU-9 dec\CMV_9dec010_NV_5fu_rko_A_30min.raw
+D:\Ft Data\Data Samples\5FU-9 dec\CMV_9dec010_NV_5fu_rko_C_30min.raw
+D:\Ft Data\Data Samples\5FU-9 dec\CMV_9dec010_NV_5fu_rko_B_30min.raw
+D:\Ft Data\Data Samples\5FU-9 dec\CMV_9dec010_NV_5fu_rko_C_0.raw
D:\Ft Data\Data Samples\5FU-9 dec\CMV_9dec010_NV_5fu_rko_A_1h.raw
D:\Ft Data\Data Samples\5FU-9 dec\CMV_9dec010_NV_5fu_rko_B_1h.raw
D:\Ft Data\Data Samples\5FU-9 dec\CMV_9dec010_NV_5fu_rko_C_1h.raw
```

Output path: D:\Ft Data\Data Samples\5FU-9 dec

Mascot Score >= 30
RT ± %: 1
Mass ± ppm: 10
RT ± min: 5
Min peptides per protein: 2

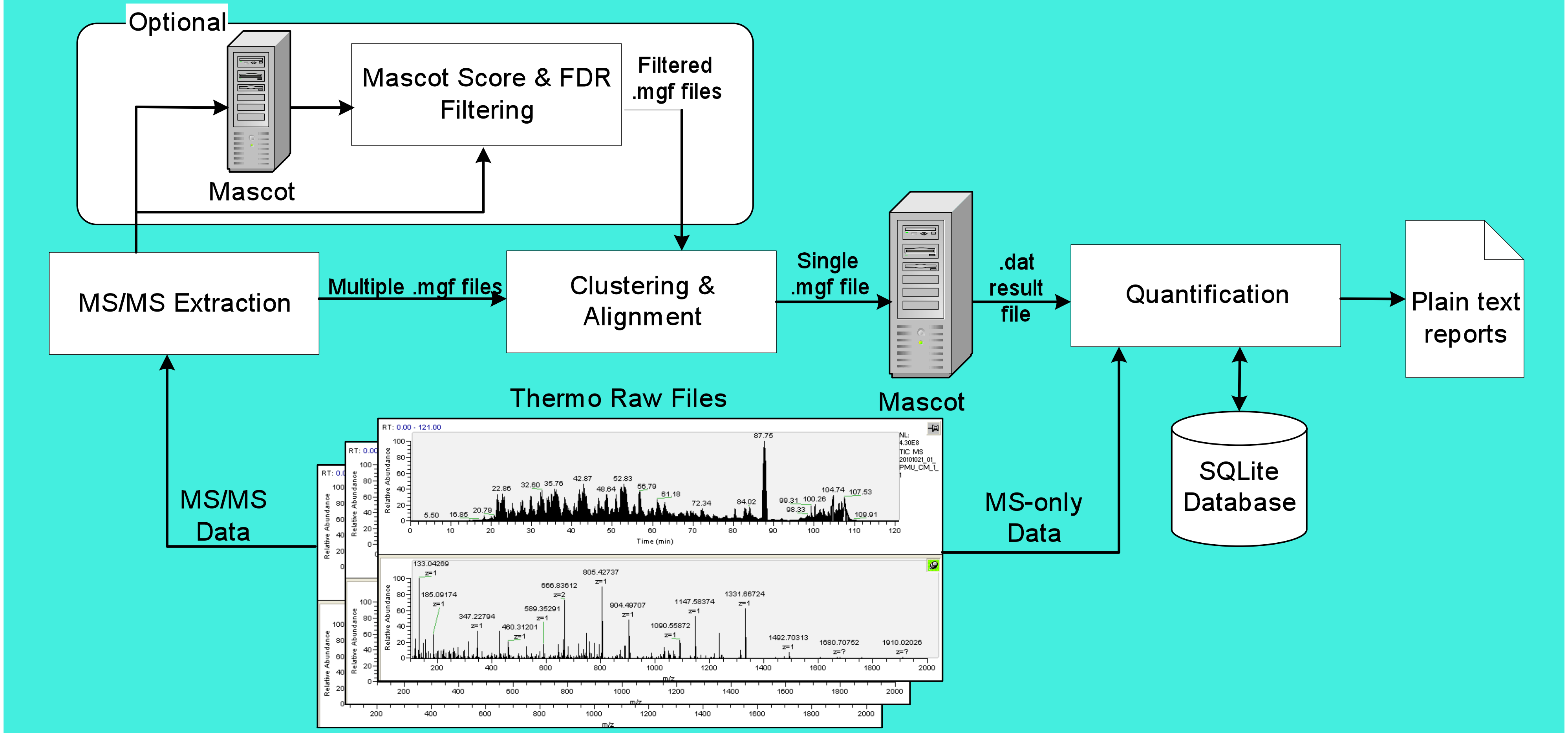
☑ Charge Deconvolution
☑ Use best mascot peptide
☐ Quantify all mascot entries (.all file)

Reference file: Matrix

ESI correction interval: 0.9
☑ Peak Shape Filtering
☐ Normalization
☐ Cut tails

Loading completed

Quanti workflow



Optional

Mascot

Mascot Score & FDR Filtering

Filtered .mgf files

MS/MS Extraction

Multiple .mgf files

Clustering & Alignment

Single .mgf file

Mascot

.dat result file

Quantification

Plain text reports

SQLite Database

MS-only Data

MS/MS Data

Thermo Raw Files

Pathway Analysis

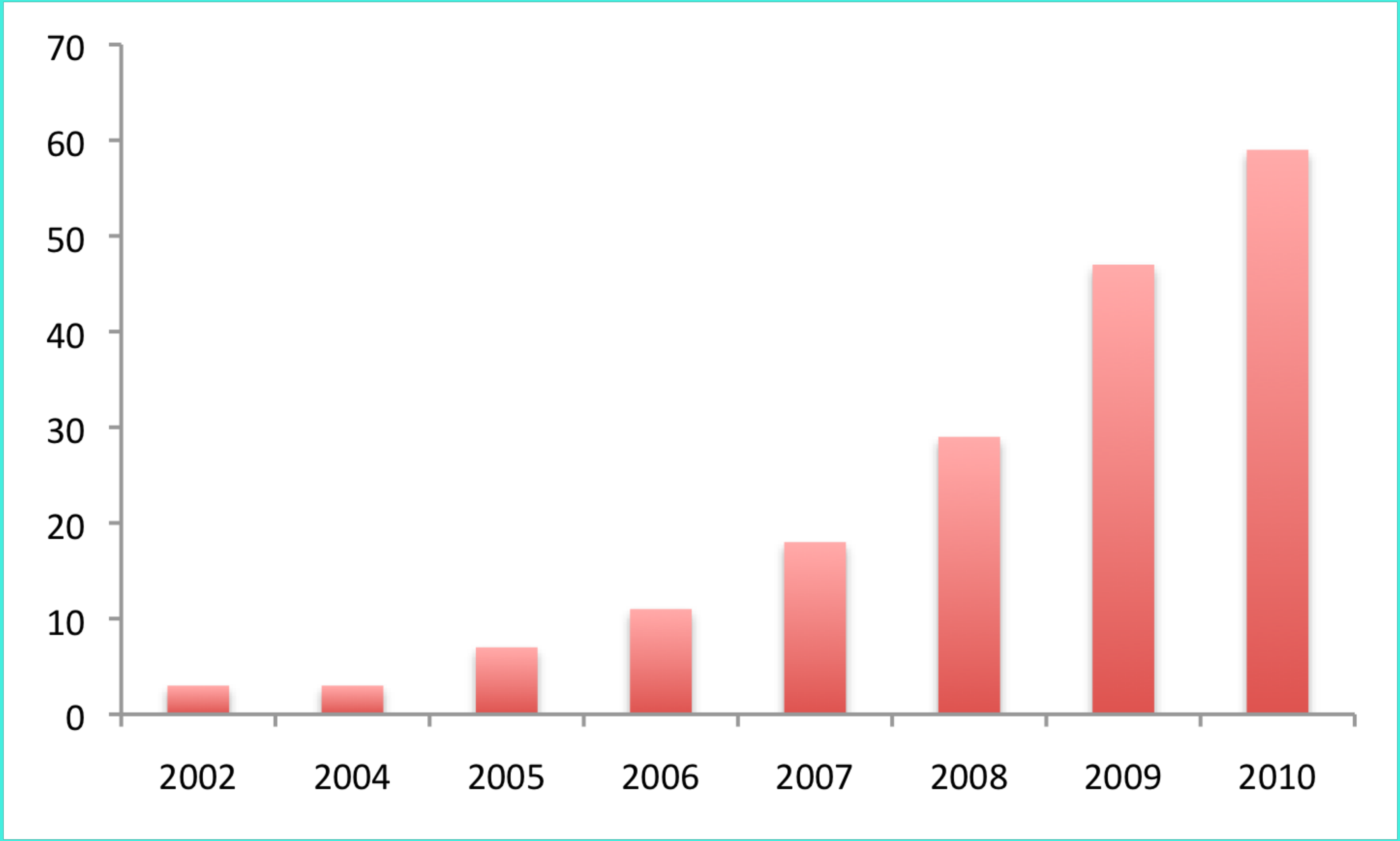
Drug Target Discovery

Disease Modeling

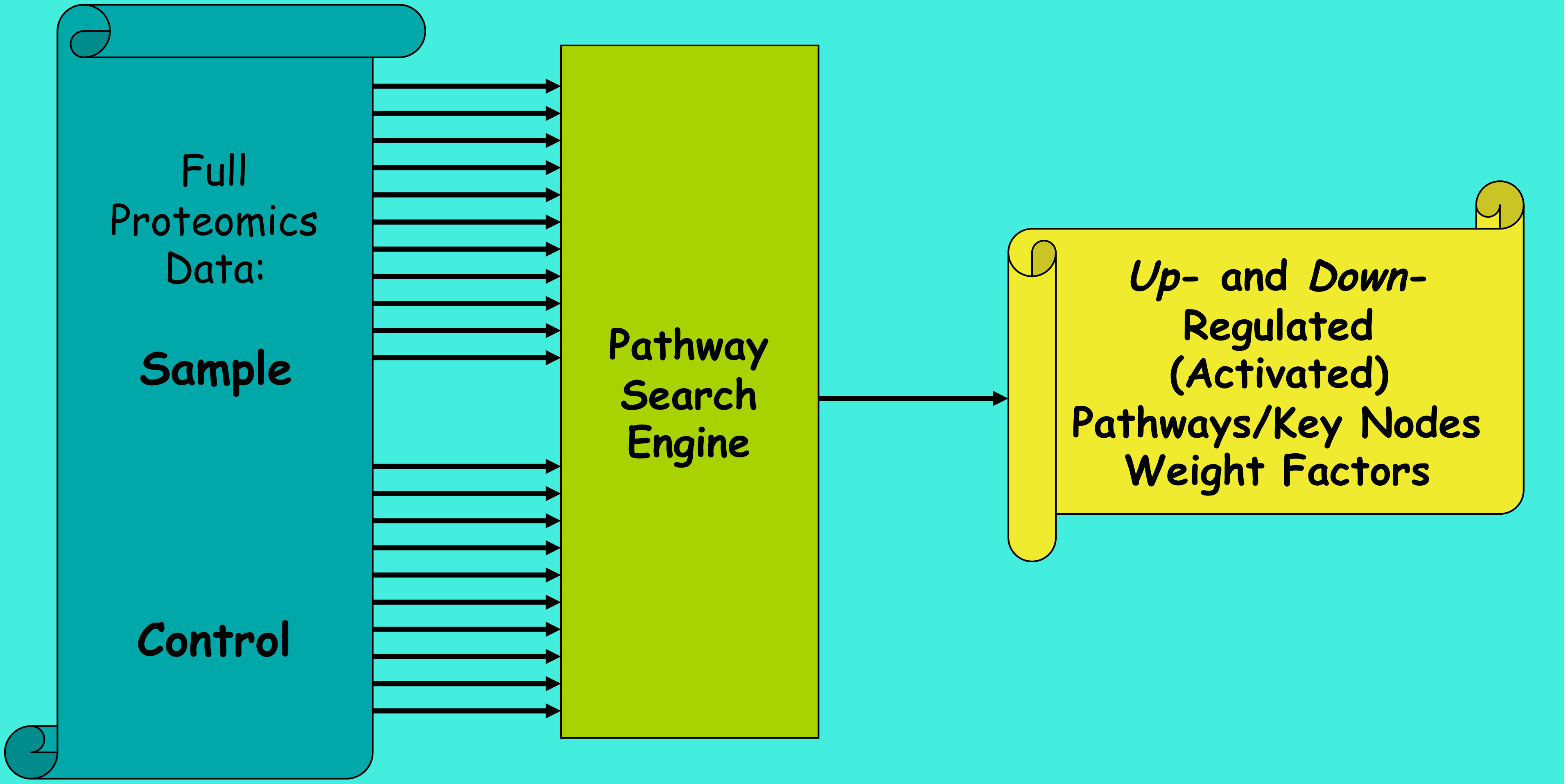
Patient Stratification

Establishing Drug Mechanism

Pathway Analysis & Proteomics

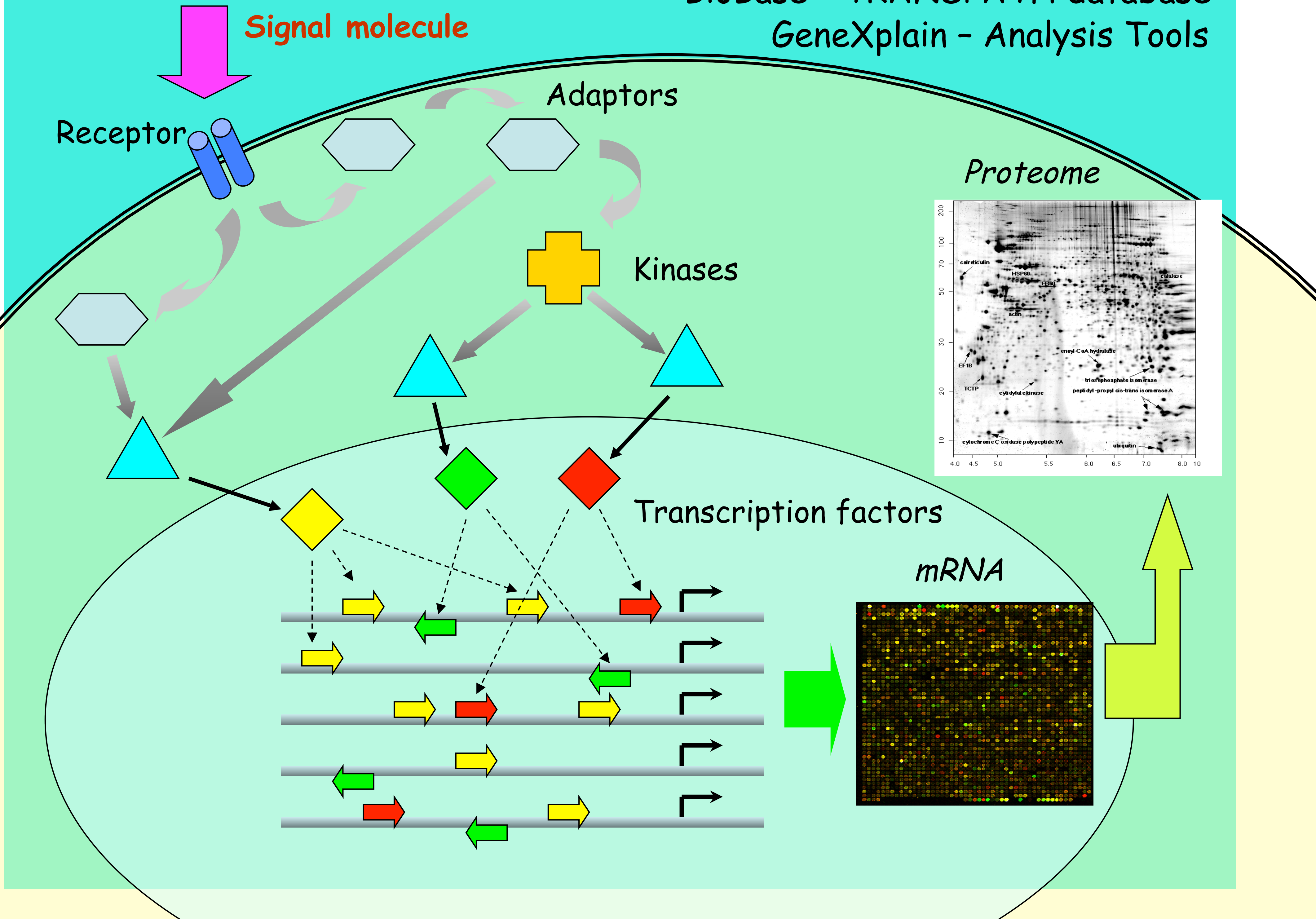


Analytical Pathway Biology

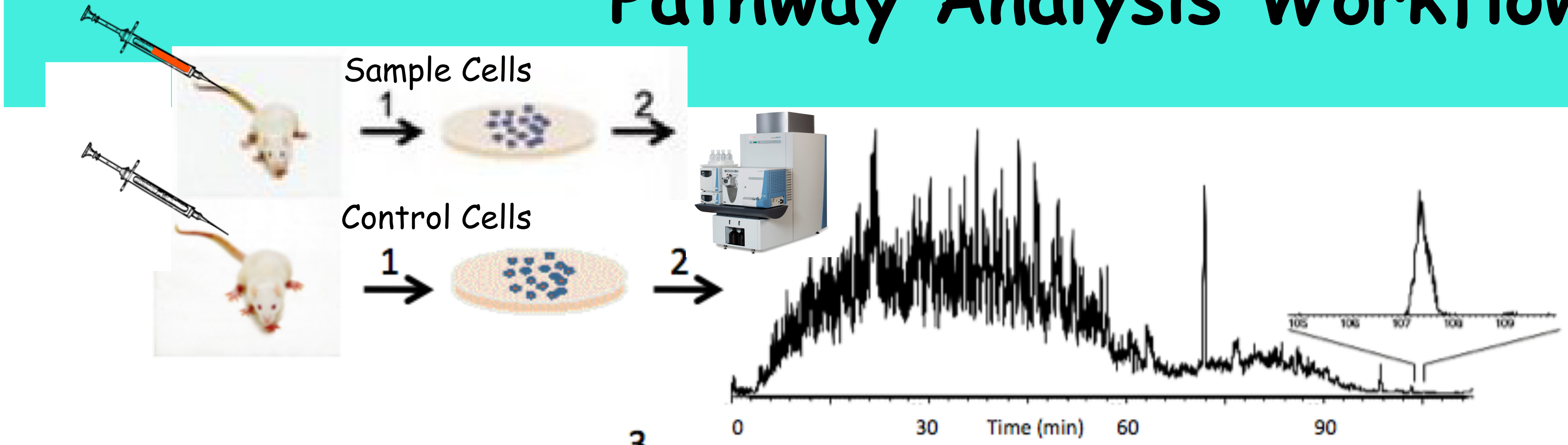


Zubarev, R. A.; Nielsen, M. L.; Savitski, M. M.; Kel-Margoulis, O.; Wingender, E.; Kel, A. Identification of dominant signaling pathways from proteomics expression data, J. Proteomics, 2008, 1, 89-96.

BioBase – TRANSPATH database
GeneXplain – Analysis Tools

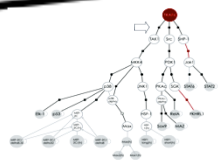


Pathway Analysis Workflow



3

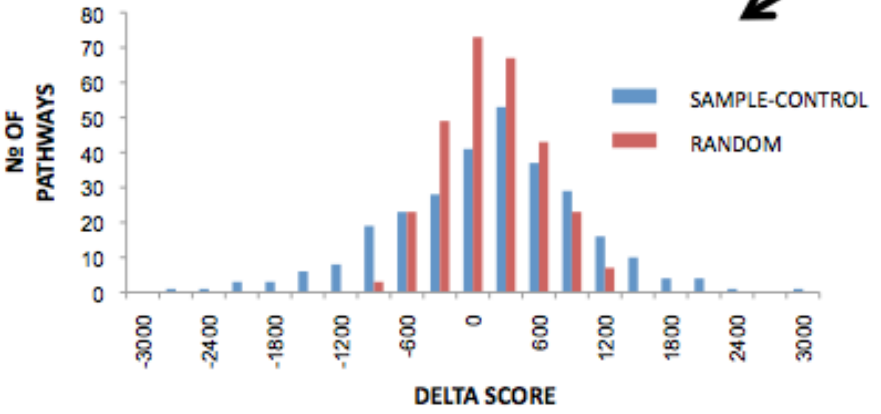
ACCESION №	CONTROL	SAMPLE
IPI00131695	6.91E+08	3.71E+09
IPI00319992	4.06E+09	8.84E+08
IPI00653158	3.23E+08	2.84E+09
IPI00830313	3.86E+09	6.35E+08
IPI00468481	1.28E+09	2.21E+09
IPI00312058	4.21E+09	2.61E+09
IPI00116753	3.62E+09	3.87E+09
IPI00652371	8.01E+07	3.26E+09
IPI00880839	2.64E+09	2.96E+09
IPI00117914	3.66E+09	3.06E+09
IPI00122815	1.09E+09	1.54E+09



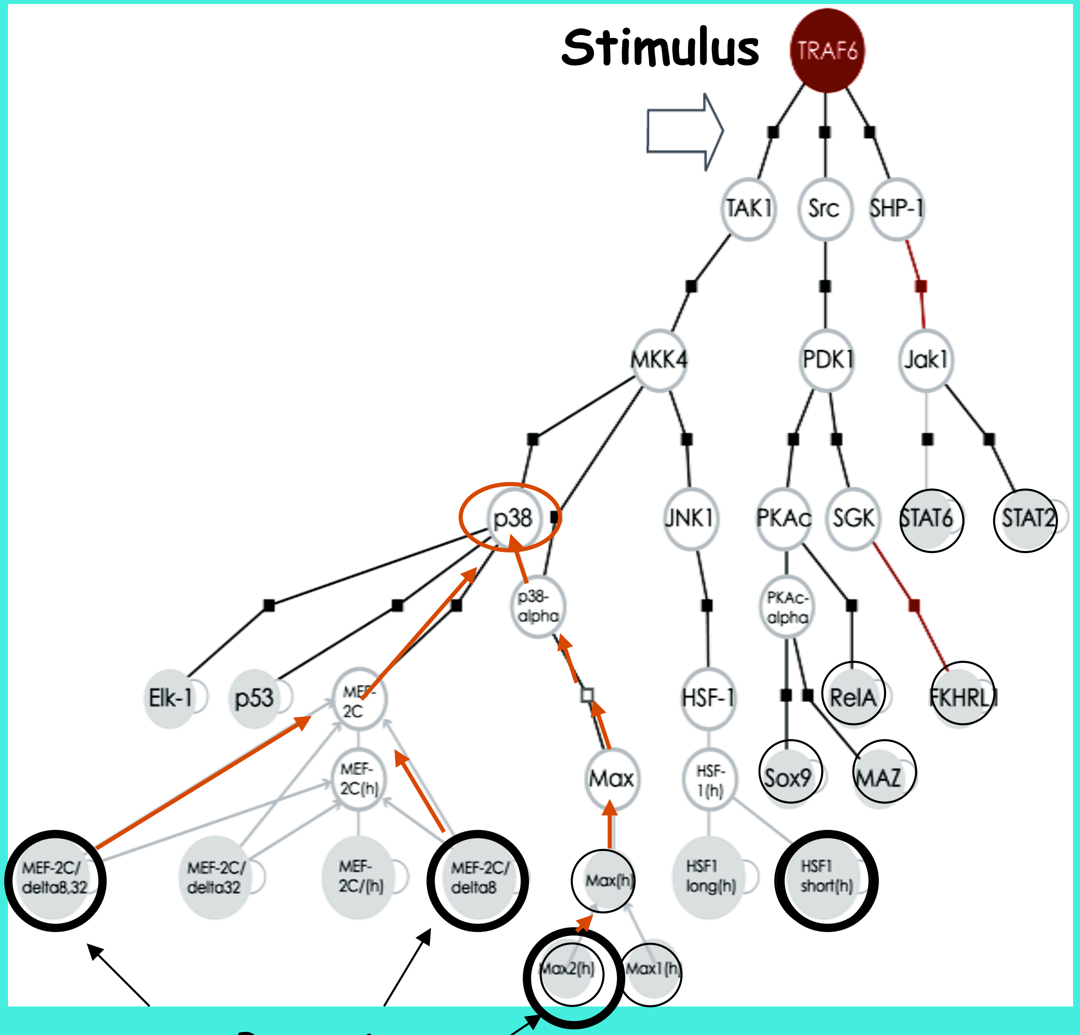
4

PATHWAY NAME	CONTROL	SAMPLE	DELTA
EGFpathway	5.97E+03	3.93E+03	-2.04E+03
JNKpathway	4.54E+03	2.76E+03	-1.78E+03
Faspathway	4.28E+03	2.03E+03	-2.25E+03
Caspasenetwork	3.97E+03	2.03E+03	-1.94E+03
E2Fnetwork	2.65E+03	2.02E+03	-6.26E+02
p53pathway	1.54E+03	2.02E+03	4.78E+02
stress-associatedpathways	1.53E+03	1.51E+03	-2.03E+01
insulinpathway	1.22E+03	1.33E+03	1.09E+02
T-cellantigenreceptorpathway	9.95E+01	1.15E+01	-8.80E+01

5



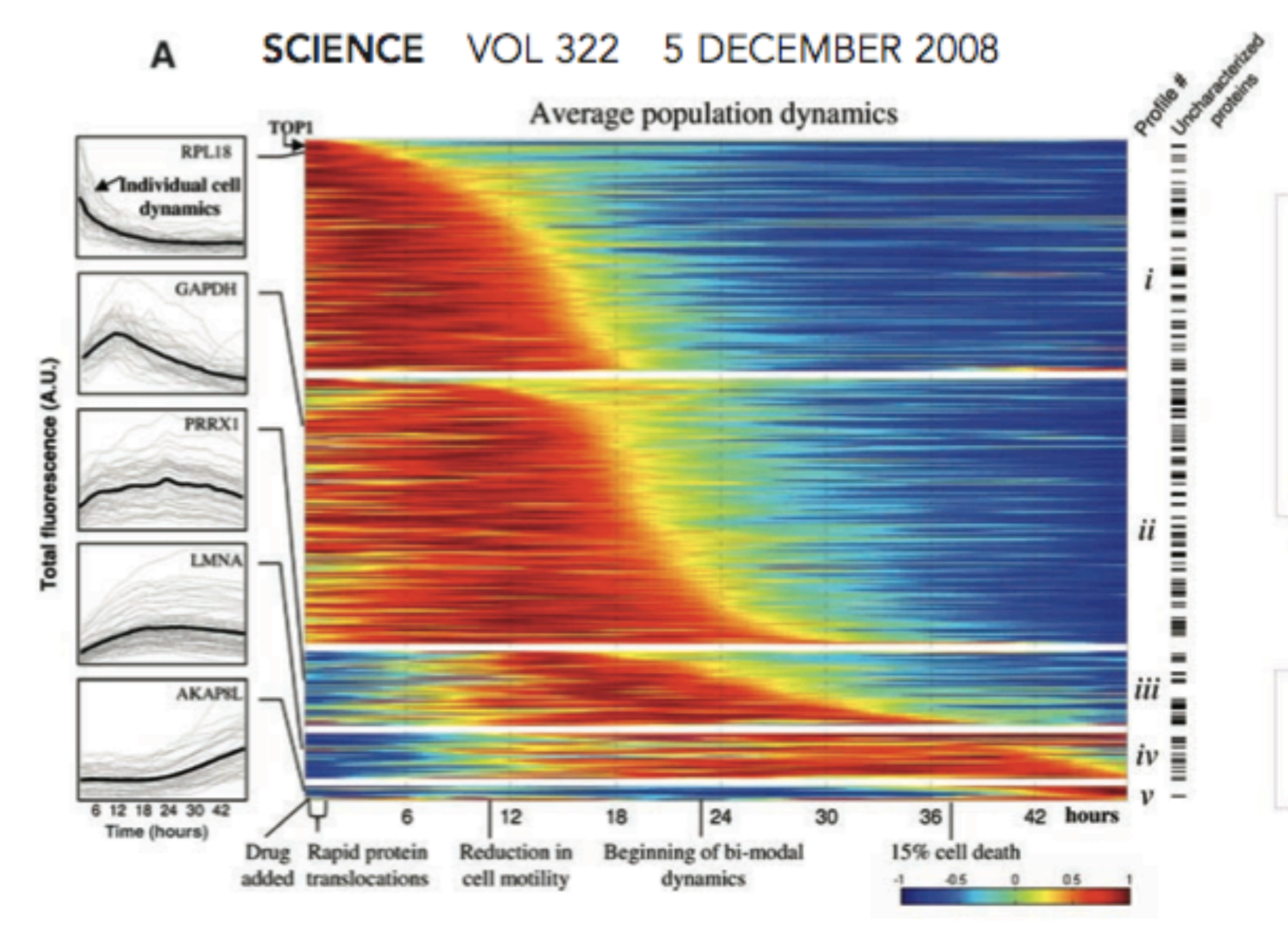
KeyNode-Mediated Analysis: Upstream



Proteins Observed

	Score
KeyNode$_1$	3050
KeyNode$_2$	2987
KeyNode$_3$	2073
...	
KeyNode$_N$	25

Pathway score:
$\sum$(keynode score)



DYNAMIC PROTEOMICS APPROACH

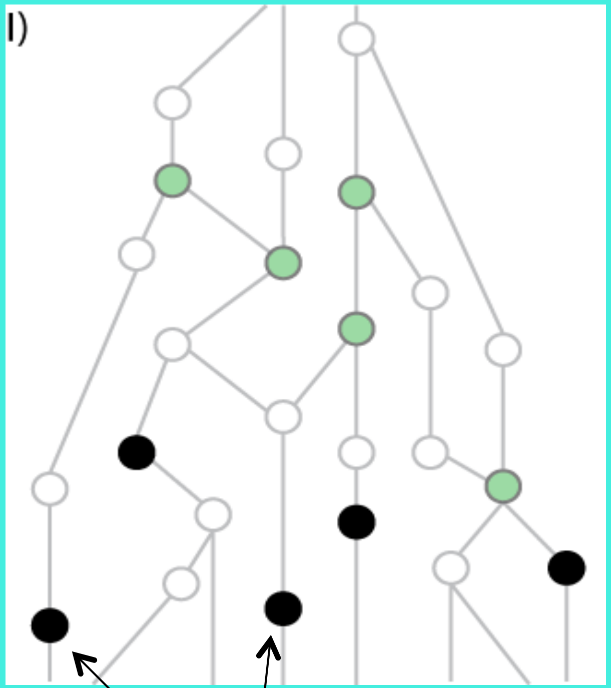
for drug target identification:

- by the **speed** of change (1 h), 10% selection
- by the total change in 48 h, 10% selection

Overall: top 3% (35 proteins)

Pathway Analysis of Dynamic Proteomics Data

I) Protein mapping on Pathways

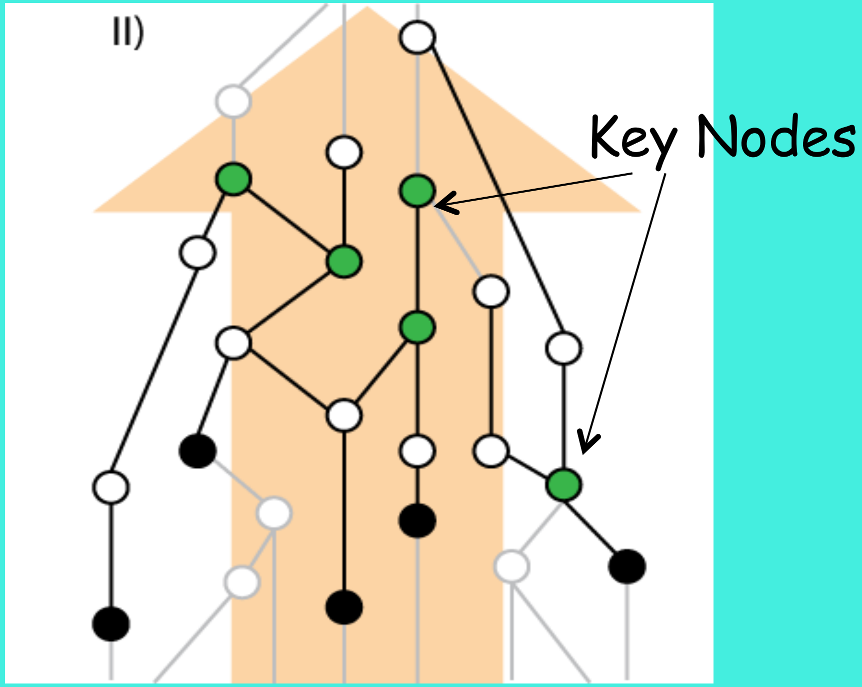


Proteins from input list

Pathway Analysis of Dynamic Proteomics Data

Upstream Search:

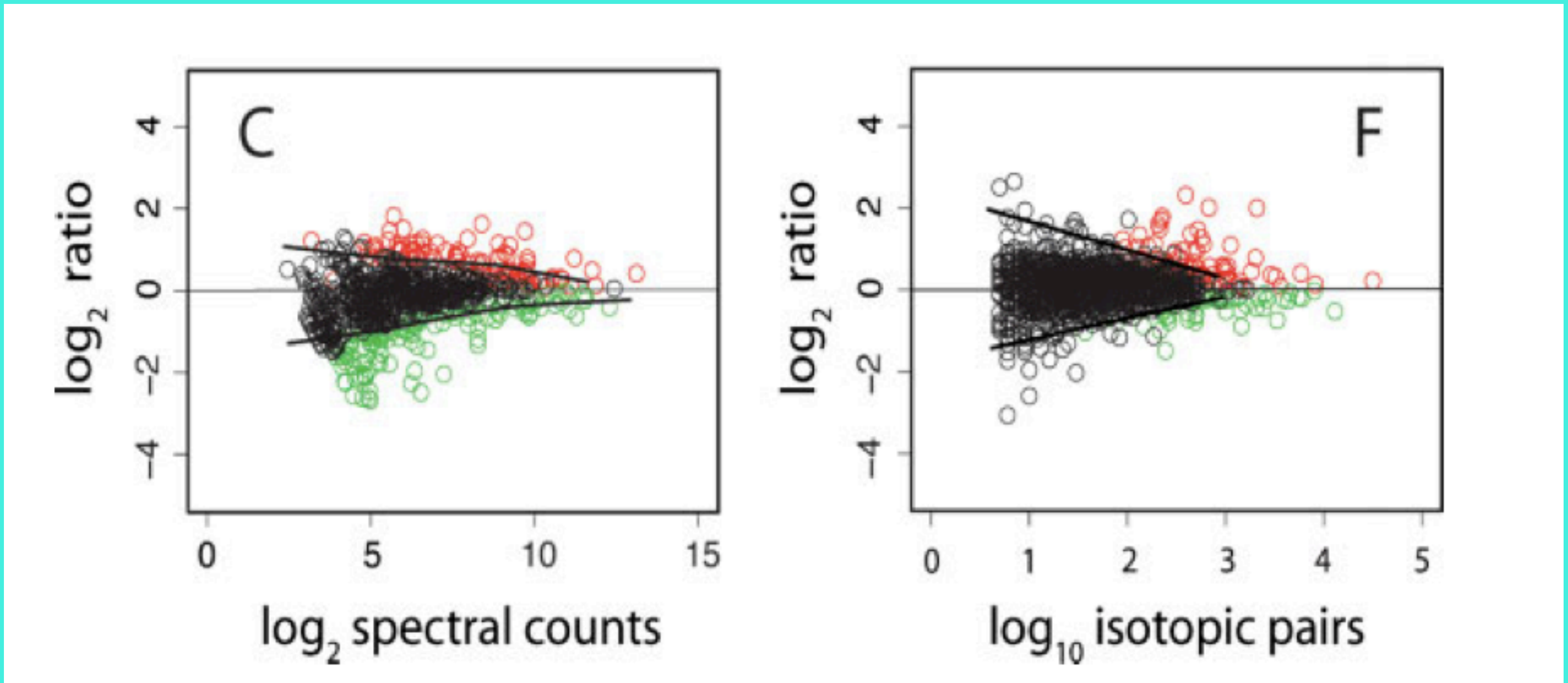
- for Speed, 0-60 min
- for Magnitude, 0-2800 min



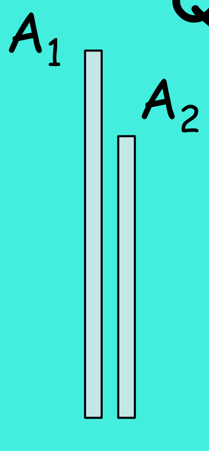
KN Scoring: $\Delta S = (S_A - S_B)*\log_2(S_A/S_B)$

Top KN is selected: one for Speed, one for Magnitude

The threshold problem in proteomics



Hacket M. **Science, Marketing and Wishful Thinking in Quantitative Proteomics**, *Proteomics, 8* (2008).

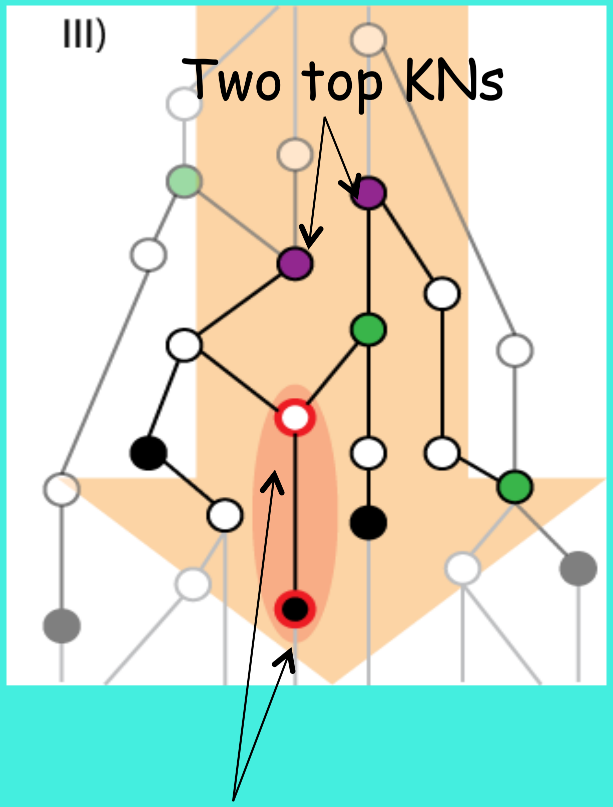


$$G = \mathrm{Abs}(A_1\text{-}A_2)\times\log_2(A_1/A_2)\ \ [\mathrm{ppm}]$$

IF statistical fluctuations of protein abundances follow Poisson distribution, G-threshold is constant

Pathway Analysis of Dynamic Proteomics Data

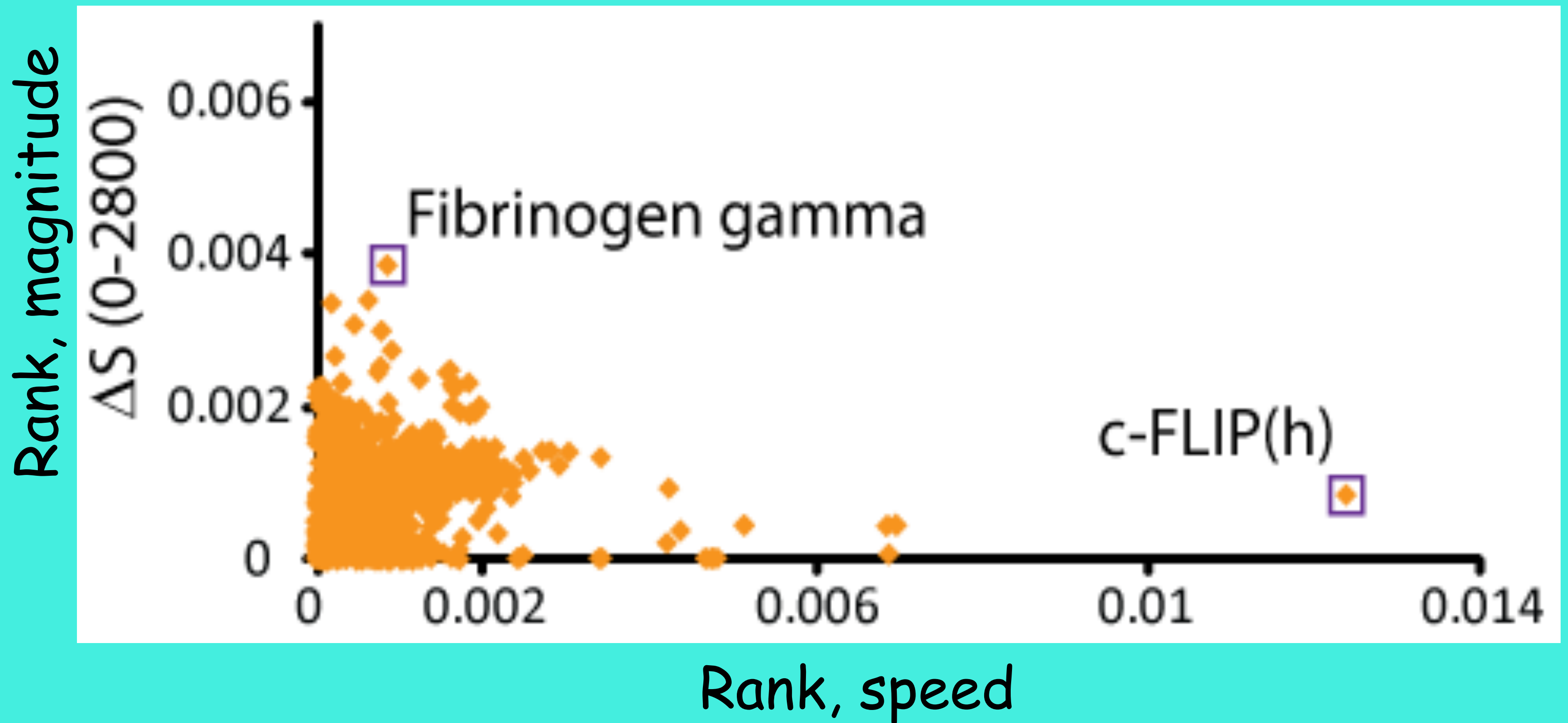
Downstream KN search



Two top KNs

Overlapping Molecules
= Drug Target Candidates

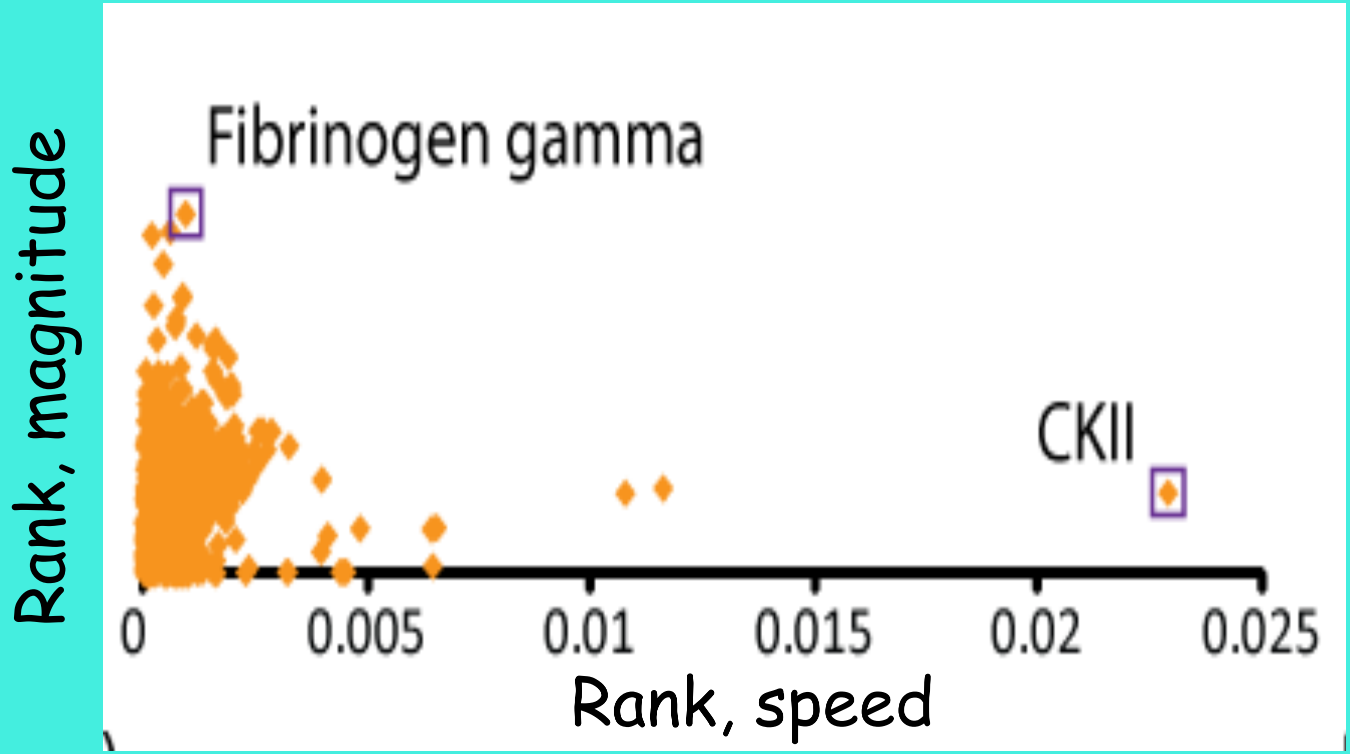
Identification of TOPI as *the* drug target from 812 proteins in the input list



Overlap of downstream lists from F_{gamma}, c-FLIP(h):
9 proteins, of which 2 from input list (known dynamics):

- **TOPI, (speed + magnitude)-rank 228**
- 26S proteasome, (speed+ magnitude)-rank 787

What if TOPI is *removed* from Input list?..



Overlap of downstream lists from F_{gamma}, c-FLIP(h):
4 proteins, none from the input list:

- **TOPI**
- **CKII**
- **Two NR-related proteins**

Take-home messages:

- Transciptomics and proteomics overlap, but proteomics is “closer to action”, and thus produces more relevant data

- Proteomics is currently limited in “depth” due to the large dynamic range of protein abundances, but technology moves forward fast, and the proteomics depth is increasing

- Correlation analysis provides first insight into the biological process, but pathway analysis is necessary to put the results in biological context

-Simple mapping of regulated proteins onto pathways (“direct mapping”) often is insufficient;

- Upstream keynode analysis is superior over direct mapping

- Combining transcriptomics, proteomics and metabolomics data is the future goal of pathway analysis