History of the Omics Cascade

Åsa Wheelock, Ph.D.
Division of Respiratory Medicine
& Karolinska Biomics Center
asa.wheelock@ki.se

In: *Systems Biology and the Omics Cascade*,
Karolinska Institutet, June 9-13, 2008

In order to understand the presents and the future, we have to understand the past.

The goals of this lecture:

To give a perspective of Omics research in relation to

- Molecular Biology

- Systems Biology

The OMICS Cascade

What CAN happen

GENOMICS

What APPEARS to happen

TRANSCRIPTOMICS

- **1953: DNA structure discovered by James Watson, Francis Crick and Rosalind Franklin**
- **1975: DNA sequenceing by Sanger**
- **1983: PCR developed by Kary Mullis (Nobel prize 1993)**
- **1995: First microarray pub. by Mark Shena et al.**

Poor correlations between mRNA and protein abundance

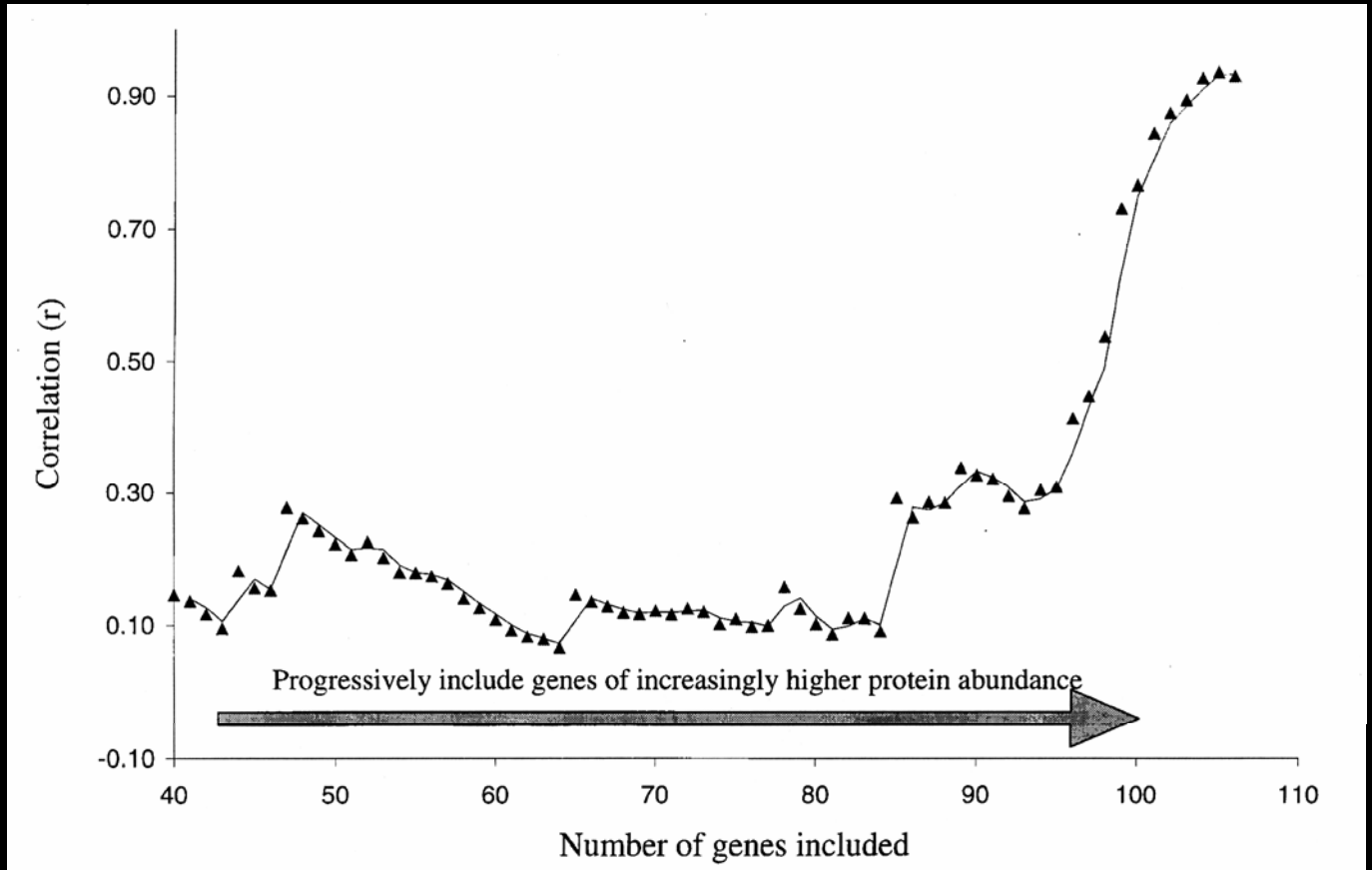
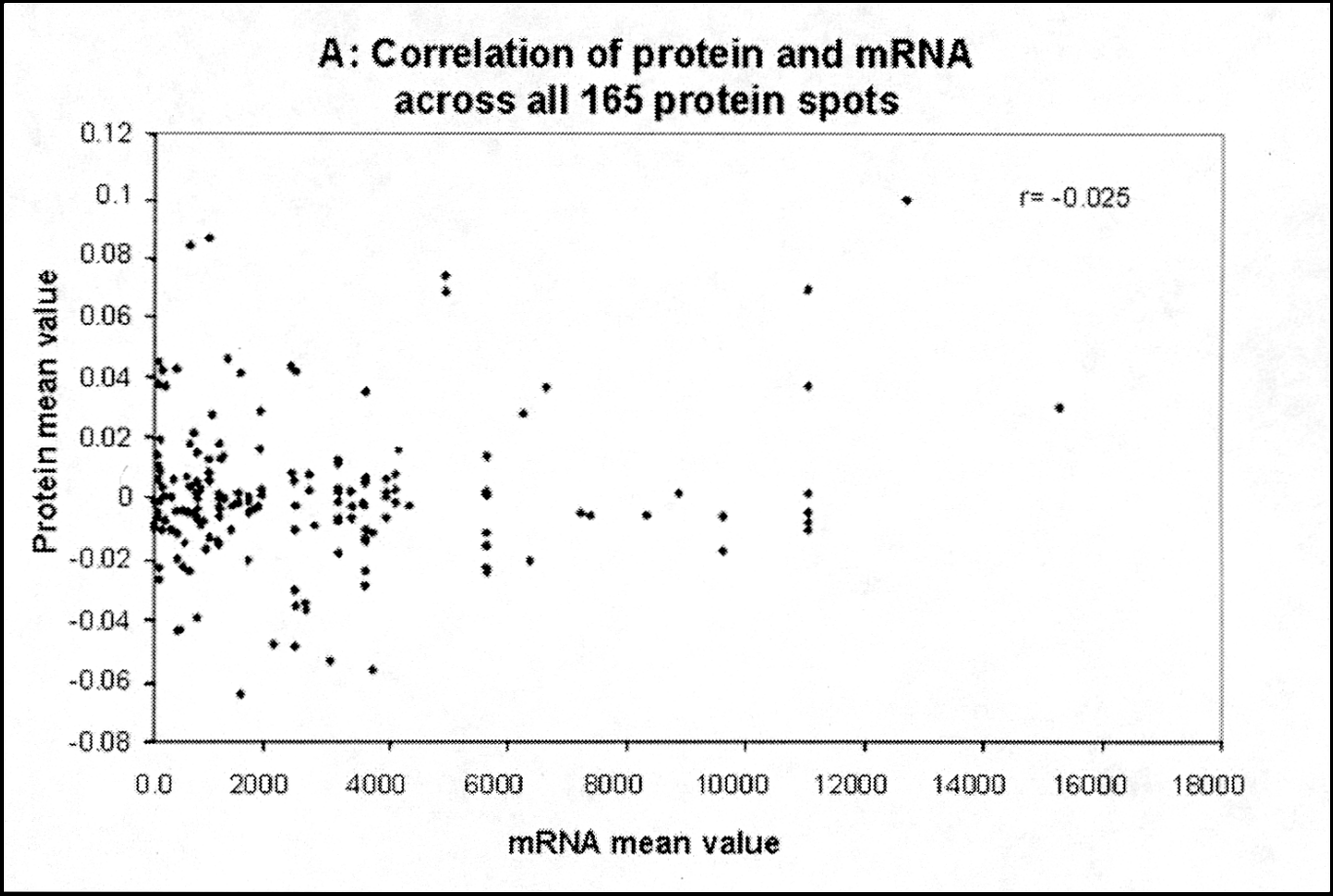


FIG. 6. Effect of highly abundant proteins on Pearson product moment correlation coefficient for mRNA and protein abundance in yeast. The set of 106 genes was ranked according to protein abundance, and the correlation value was calculated by including the 40 lowest-abundance genes and then progressively including the remaining 66 genes in order of abundance. The correlation value climbs as the final 11 highly abundant proteins are included.

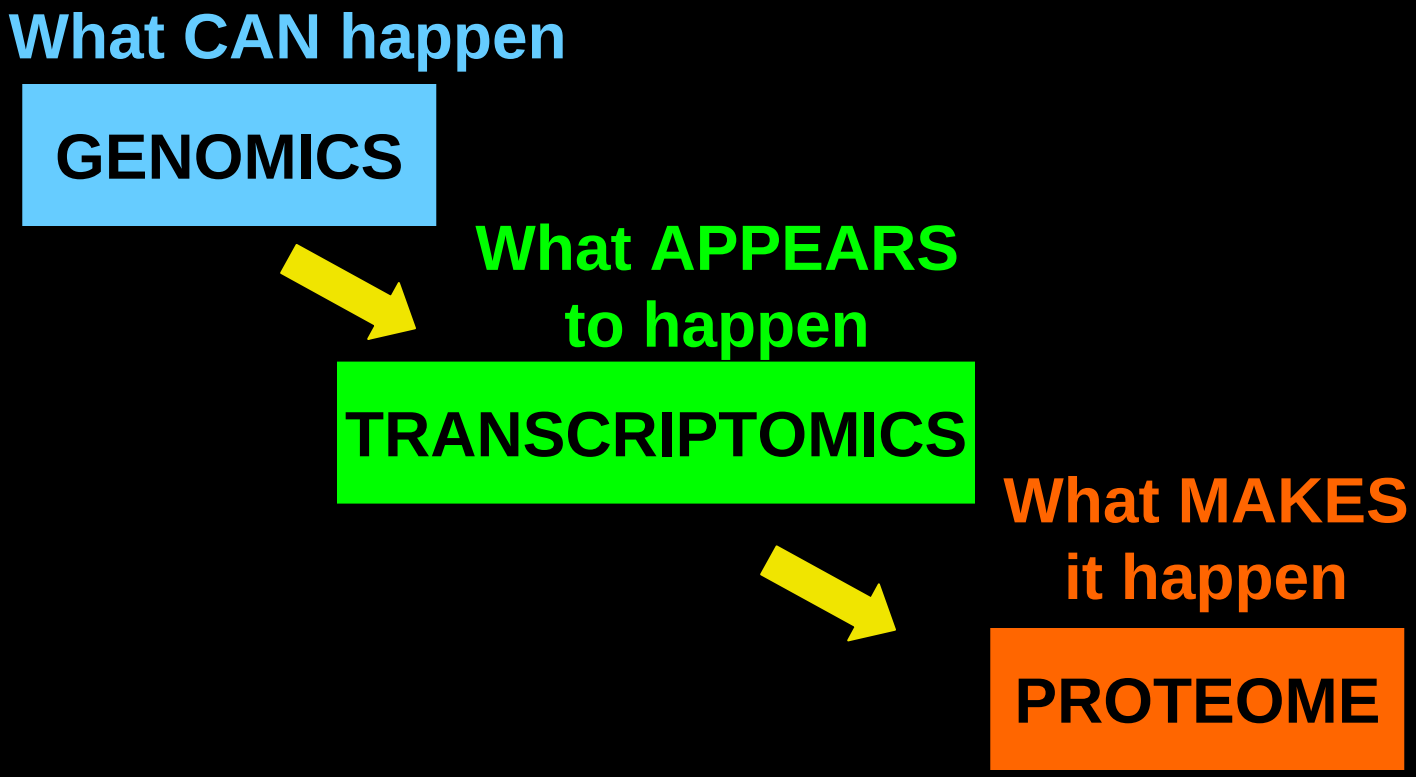
Gygi et al., Mol Cell Biol, 1999

Correlation in 85 samples of human lung adenocarcinoma



Chen et al., MCP, 2002

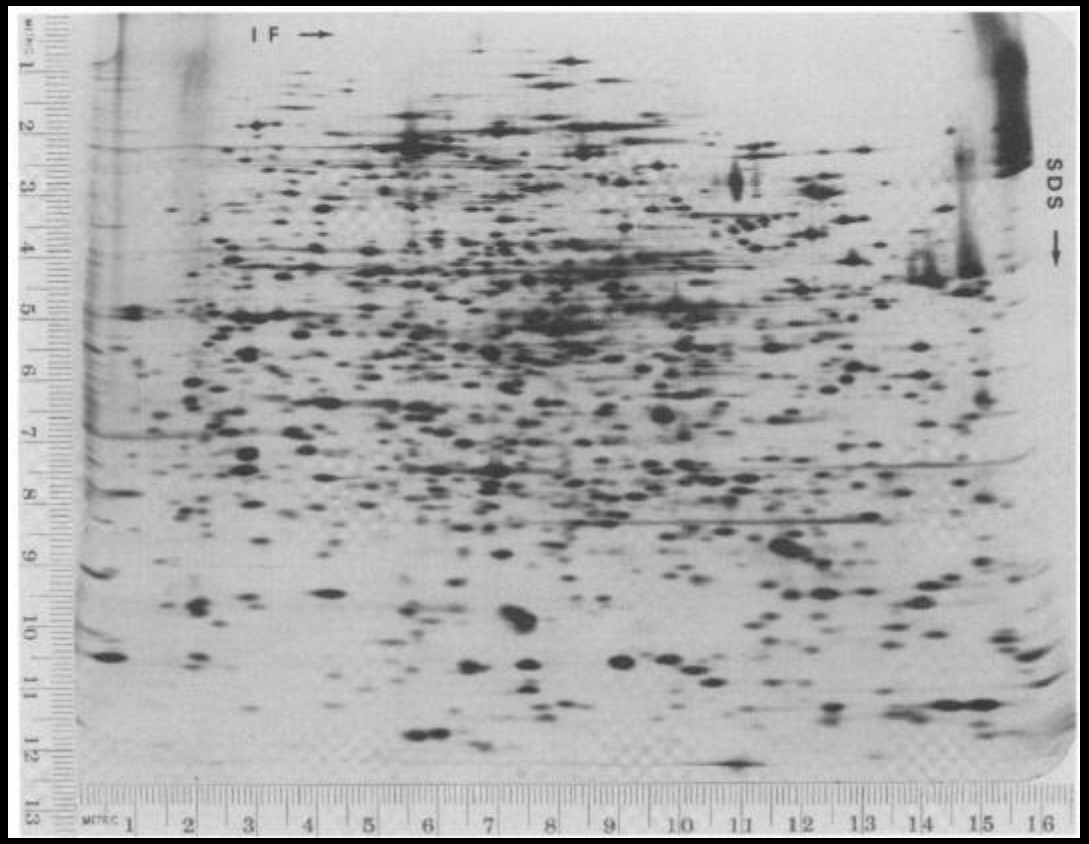
The OMICS Cascade

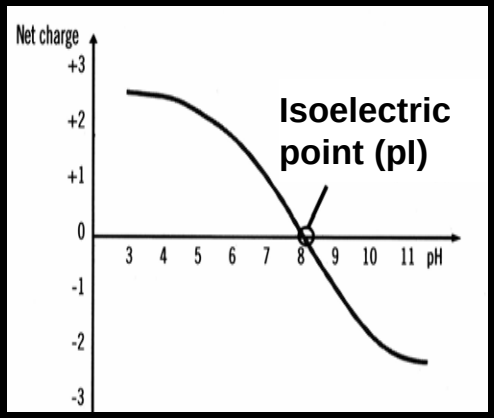


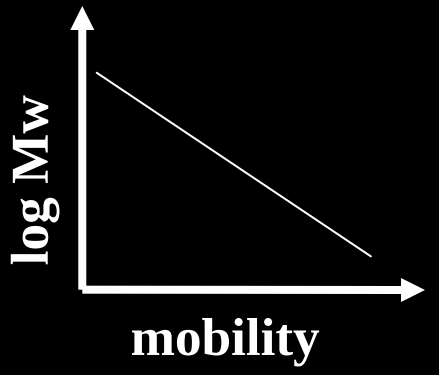
- 1994: Marc Wilkins coins word ”proteome”
  - PROTEin complement of the genOME
- 1997: Yeast genome sequenced

CRAFTSMAN

Patrik O'Farrell (1975) "High Resolution Two-Dimensional Electrophoresis of Proteins







- *Klose, J. 1975. Humangenetic 26, 231-43*

From Protein Chemistry to Proteomics

Protein separation:

- 2-dimensional electrophoresis

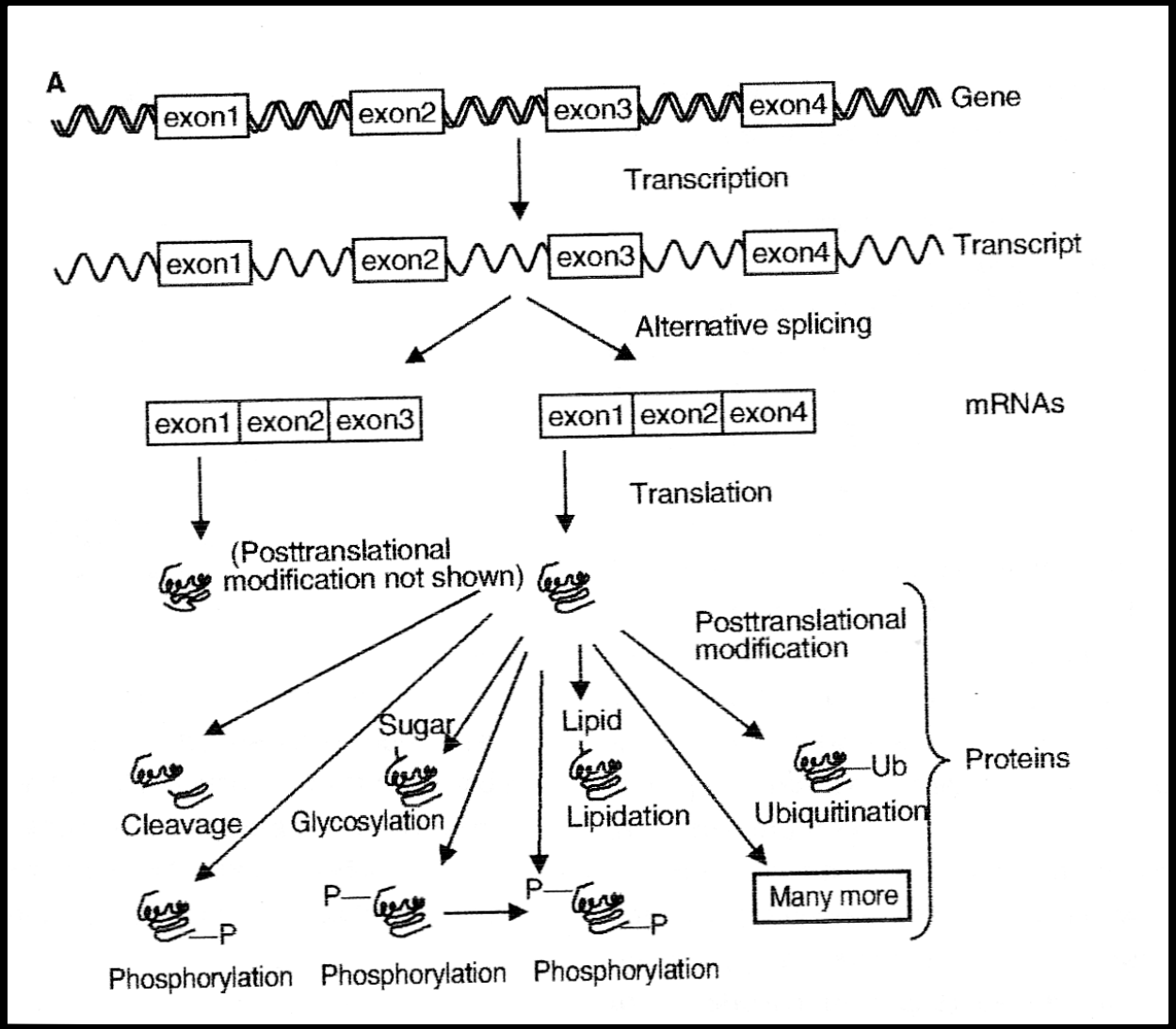
Bioinformatics:

- First generation of 2DE analysis software

Protein identification:

- Edman sequencing
- Mass spectrometry

Complexity of the proteome



30,000 genes per cell coding

Alt.splicing => 2-3 x 30,000 = 90,000 proteins

post-translational Modifications => 10 x 90,000 = 900,000 proteins

Peng and Gygi, JMS: 36:1083, 2001

The OMICS Cascade

What CAN happen

GENOMICS

What APPEARS to happen

TRANSCRIPTOMICS

What MAKES it happen

PROTEOME

What HAS happened

METABOLOME

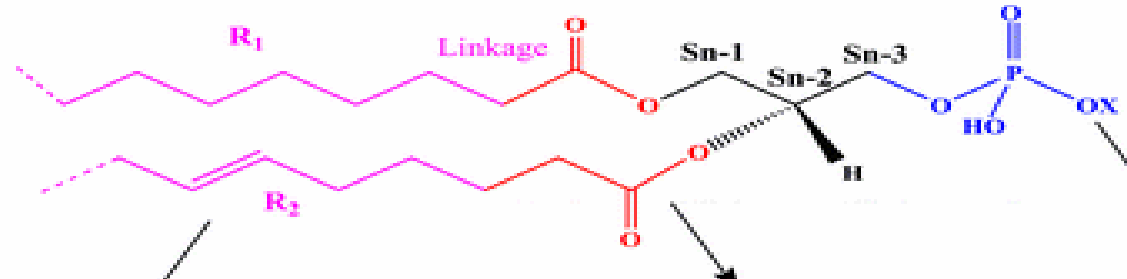
1940s: Chromatography invented by Archer John Porter Martin (Noble Prize 1952)

1970: Robinson & Pauling: chromatographic patterns of urine of vitamin B6-exposure

1946: NMR by Felix Bloch & Edward Mills Purcell (Nobel Prize in 1952)

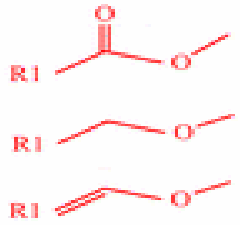
Complexity of Metabolomics

(A) Glycerophospholipid Structure

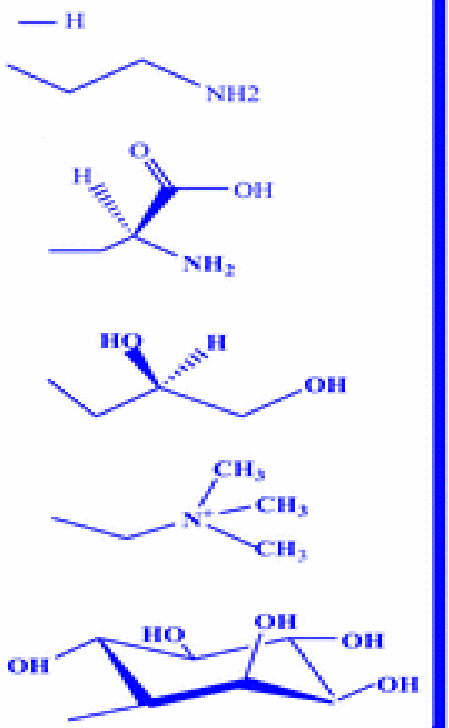


Fatty acid chains (R1,R2)
-Varying length
-Degree of saturation
-Type of double bond (cis/trans)

Linkage
-Ester (shown)
-Ether
-Vinyl Ether



Head group moiety (X)



(B) Theoretical number of molecular lipids

Type of lipid	# of molecular lipids in class
40 common fatty acids	40
40 fatty acid acyl-CoA	40
Monoacylglycerols	120
Diacylglycerols	4800
Triacylglycerols	64000
Phospholipids (6 classes)	9600
Ceramide & Sphingomyelin	>400
Glycosphingolipids	>100000
Cholesterol/Cholesterol ester	41
Total #	**~ 180000**

Powerful tools made Systems Biology possible

What CAN happen

GENOMICS

What APPEARS to happen

TRANSCRIPTOMICS

What MAKES it happen

PROTEOME

What HAS happened

METABOLOME

Omics methods are

not

defined by HIGH THROUGH-PUT...

...but by

HIGH OUT-PUT!

Challenge in Omics Research

Expensive studies =>

- **Small number of replicates (n)**
  - **(microarrays, subjects...)**
- **Large number of variables**
  - **(genes, proteins, etc)**

<u>Results in:</u>

- **Sensitive to type I error**
- **Poor statistical Power**

Statistics revisited

Significant

The null hypothesis can be rejected

The observed difference is unlikely to have occurred by chance. "A statistically significant difference" simply means there is statistical evidence that **there is a difference**.

Significance level

“p-value”: The smaller the p-value, the more certain we are that there is a difference. The probability that the null hypothesis will be rejected in error when it is true (Type I error, or "false positive").

Statistics revisited

Type I error (α)

“false positive”: The error of rejecting a null hypothesis when it is actually true, i.e this is the error of accepting an alternative hypothesis (the real hypothesis of interest) when the results can be attributed to chance.

Type II error (β)

"**false negative**": the error of accepting a null hypothesis when the alternative hypothesis is actually true, i.e the error of failing to observe a difference when in truth there is one.

Statistics revisited

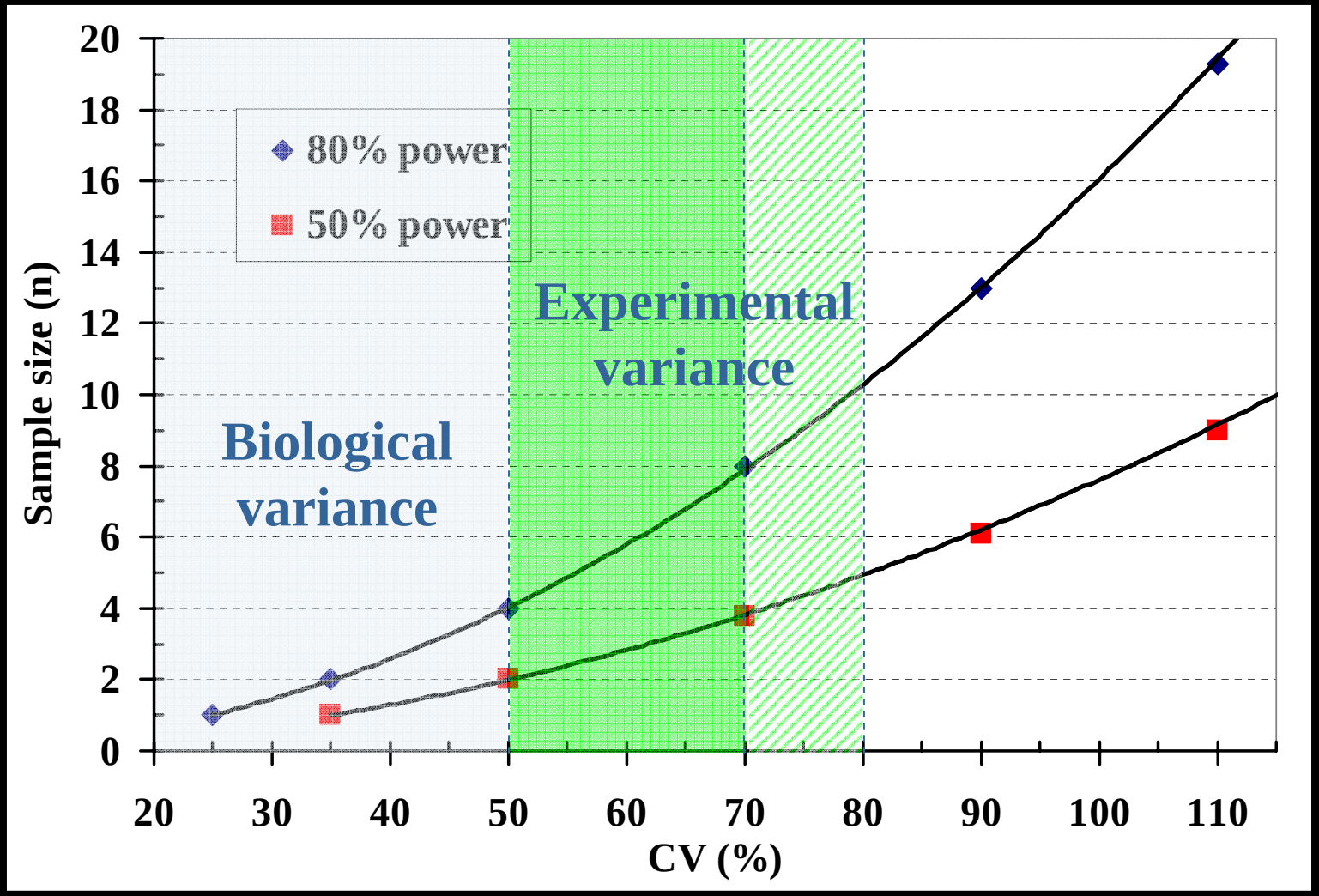
Statistical power

Power = 1-β, i.e. not make a Type II error.

The probability that the test will reject a false null hypothesis, i.e **the power to detect true positives**.

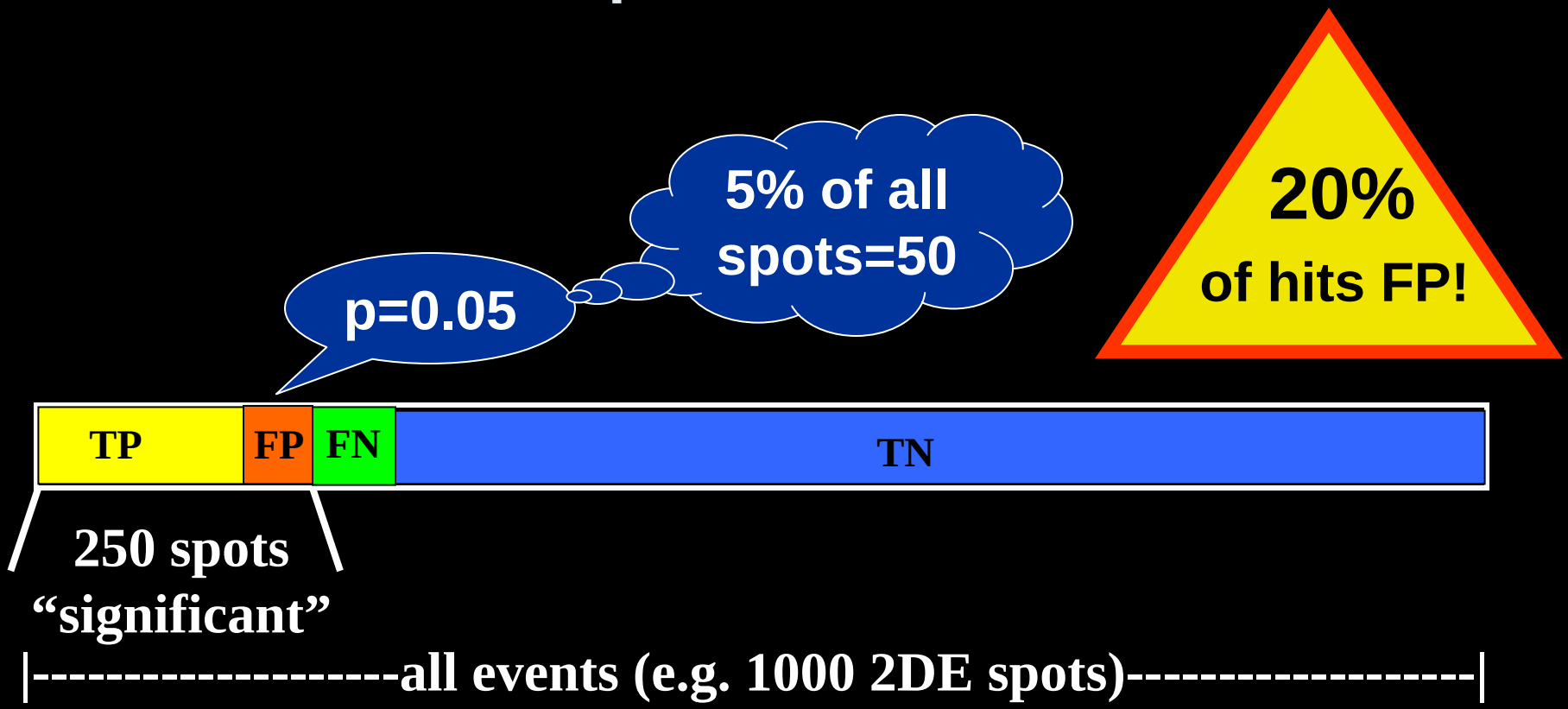
There are no formal standards for power, but a power of 0.50- 0.80 is common.

Statistical Power versus Sample size



Modified from Molloy et al., Proteomics (2003) 3, 1912-19

False positive rate



The false positive rate is the amount of false positives (FP - red) as a proportion of all the spots in an experiment, both altered (positives) and unaltered (negatives).

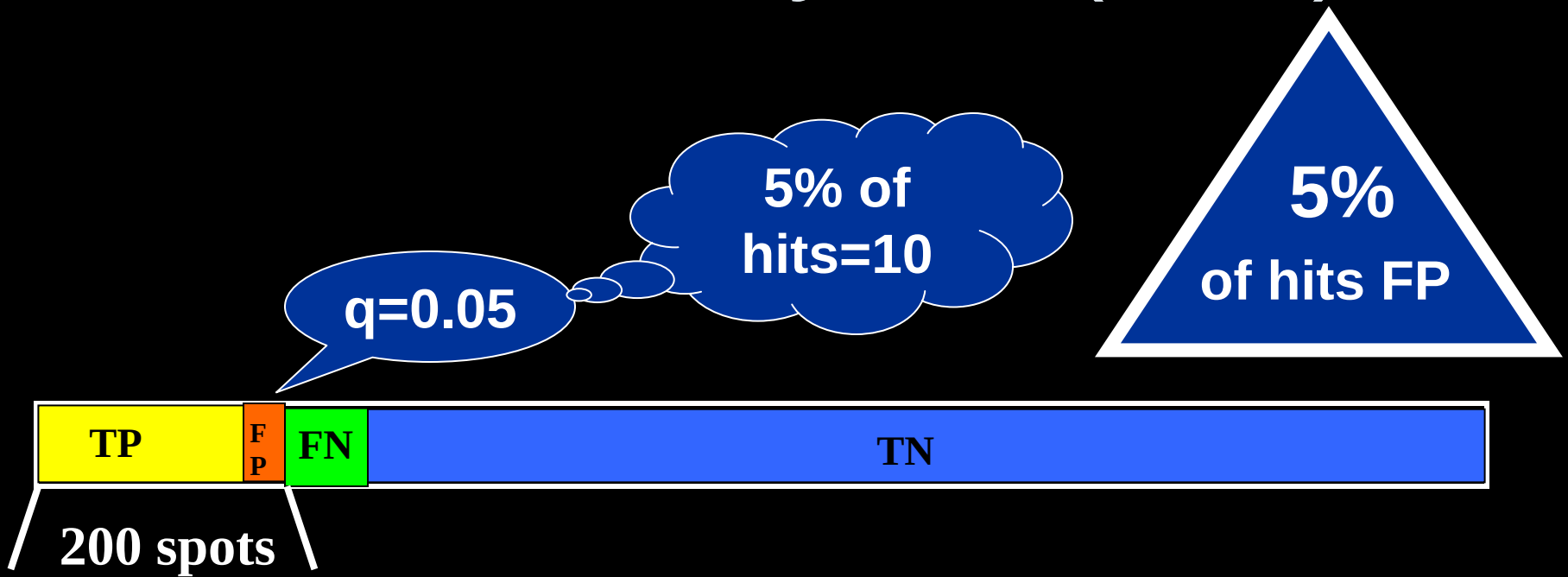
Pros & cons of omics research

+ Can survey all proteins/genes in cell

+ Global analysis => many hits

- Many hits => many false positives

False discovery rate (FDR)



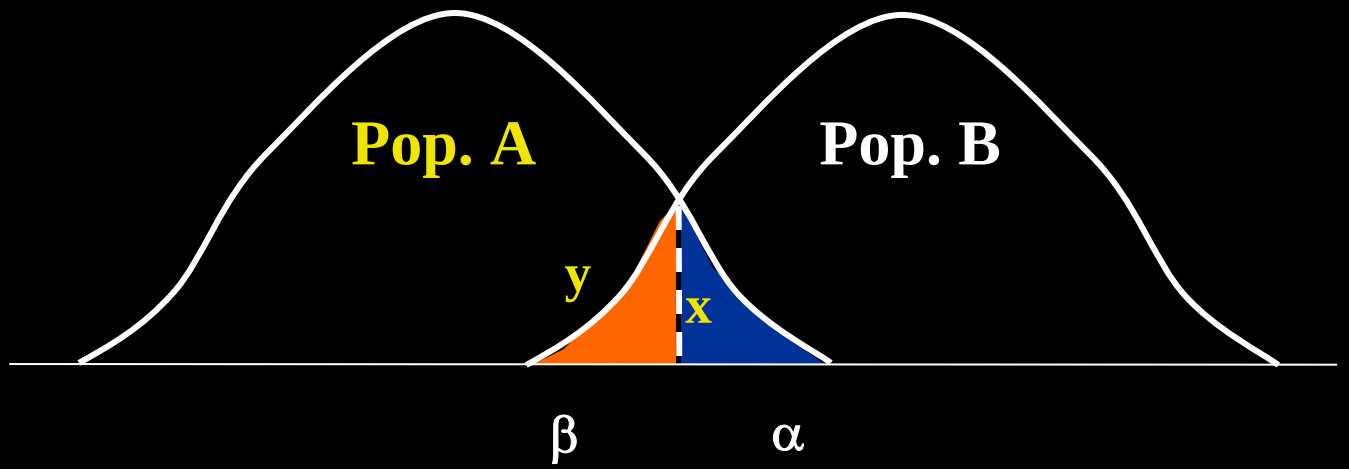
FDR : "Expected proportion of FP among rejected hypotheses". Instead of deciding the number of rejected hypothesis based on all events, an assigned FDR is used to determine cutoff for significance. The resulting cutoff value is calculated from all p-values, and is called the q-value.

FDR decreases the number of FP

What happens to the **POWER** when we switch to FDR?

Type I error (α): false positive
Type II error (β): false negative

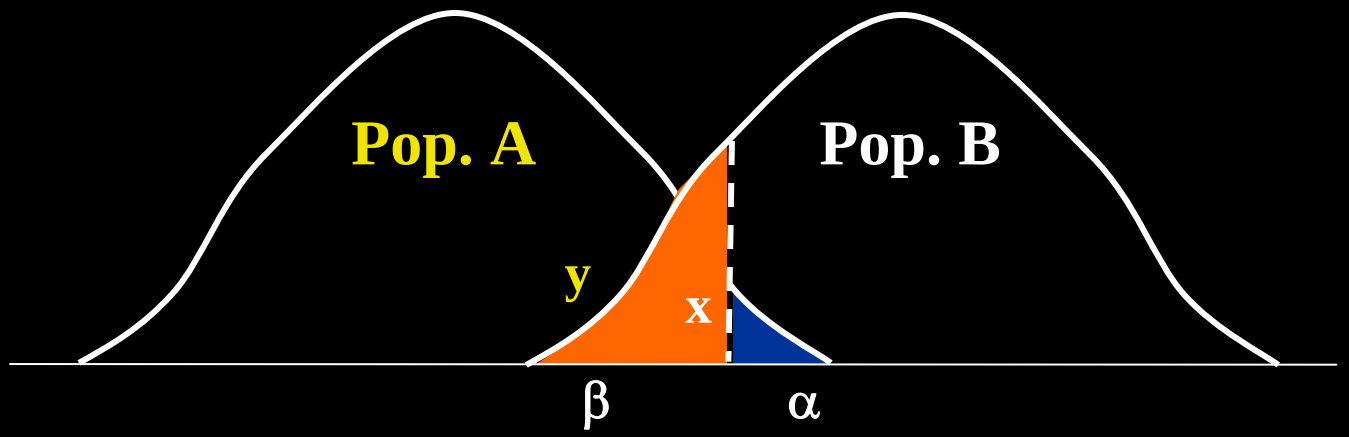
H_0: Sample x and y belong to the same population.
H_i: Sample x and y belong to different populations



Using a p-value as indicated by the dotted line, we reject the null-hypotheses: x belongs to pop. B, and y belongs to pop. A.
If x truly belongs to pop. A, we perform a …
type I error; x is a **false positive.**

Type I error (α): false positive
Type II error (β): false negative

H_0: Sample x and y belong to the same population.
H_i: Sample x and y belong to different populations



If sample x truly belongs to population B, we have performed a… type II error – x is a **false negative**. Increased risk of type II error is a drawback when utilizing p-value corrections.

Increased risk of type II error is a drawback when utilizing p-value corrections (FDR)

↑ type II error (β) => ↓ power (1- β)

PITFALL IN OMICS RESEARCH:

Decrease false positive rates may result in low statistical power to detect true positives

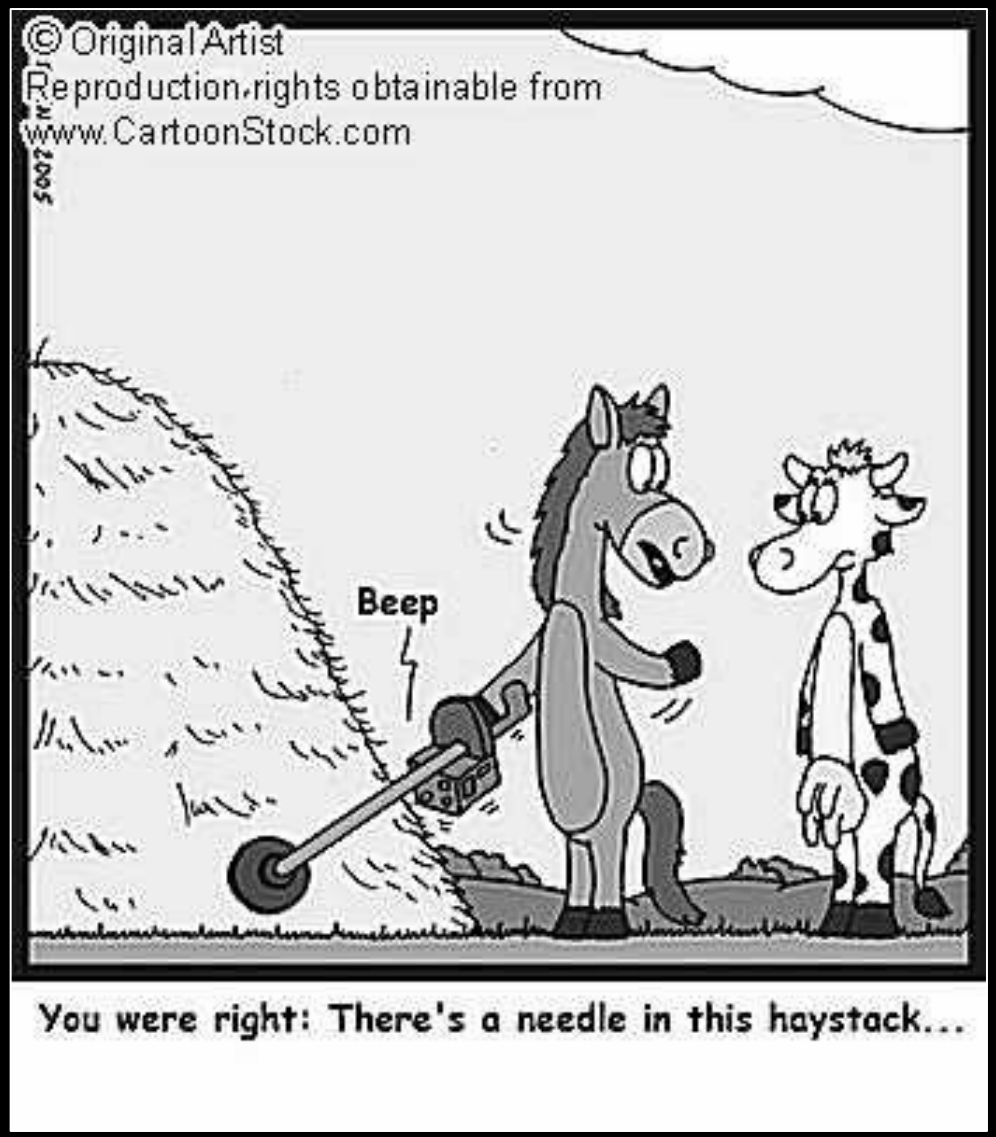
What is Omics good for?

Discovery Science

Hypothesis-generating

”Fishing trip”

Tools to find the needle in the hay stack



Biomarkers of
- Disease
- Exposure
- Response

- Proteins
- Metabolites

ID Drug targets

What is Omics good for?

WHAT? Biomarker / Drug target

Is Specific, Single Biomarker discovery possible?

Protein number:

What seems acceptable today? (evidence based)

- 23’000 genes (23’713)
  - Sequencing, cloning, etc…
- 18,000 transcripts
- Proteins in a cell? 10,000 identified.
  - Conservative: 18,000…

Number of cell types…

- List of distinct cell types in the adult human body from Wikipedia, the free encyclopedia.

–> 320

Protein number:

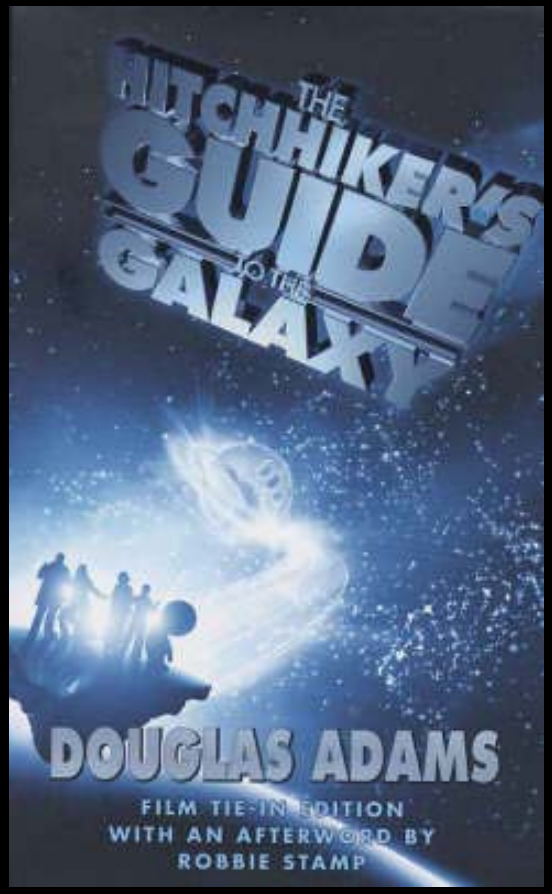
- 2500 proteins common to all cells:
  18,000 - 2500 common in cells = 15,500

- 2000 proteins (max) are secreted:
  15,500 – 2,000 secreted = 13,500

- 320 cells types:
  13,500 / 320 cell types = 42

Denis Hochstrasser,
personal communication

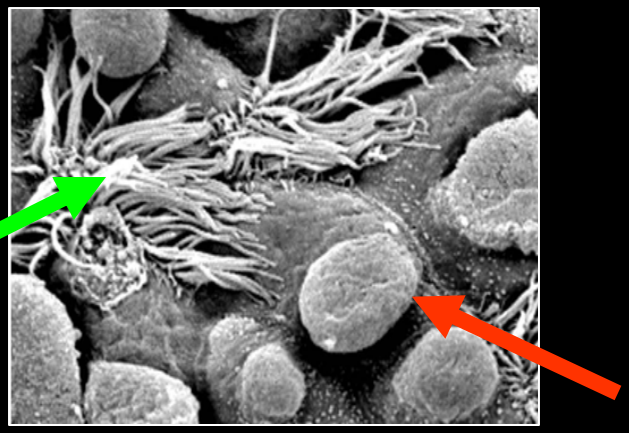
...scooped by Douglas Adams, 1979

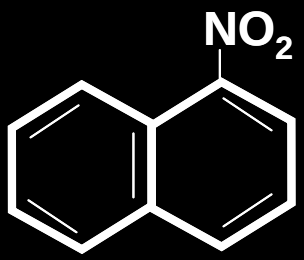


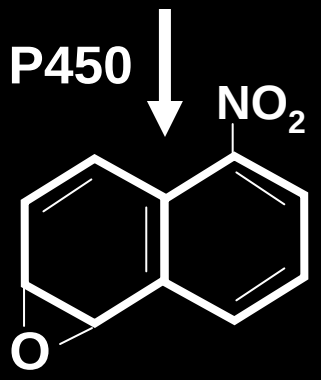
What is Omics good for?

WHAT?	Biomarker / Drug target
WHERE?	Cellular location
WHEN?	Timing of events
HOW?	Complexity

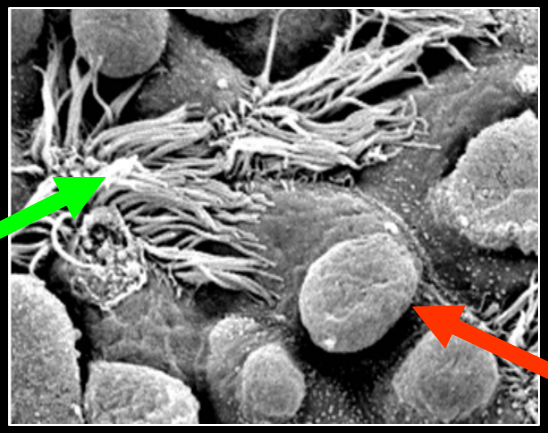
WHERE? Cellular diversity in lung

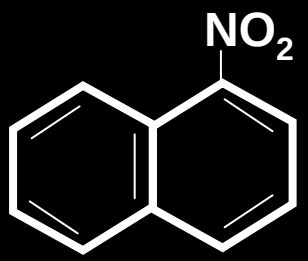


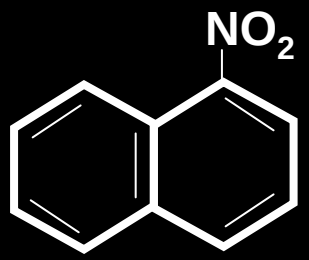


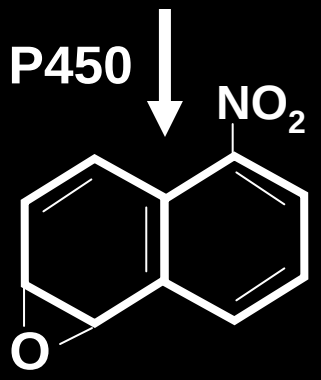


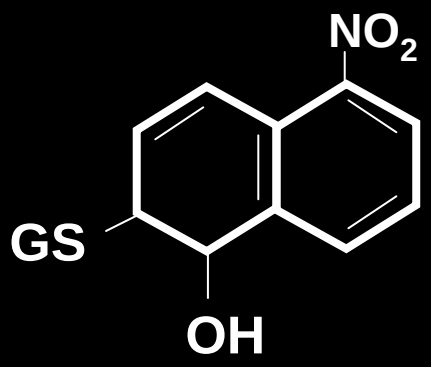
WHEN? Diurnal fluctiations of GSH











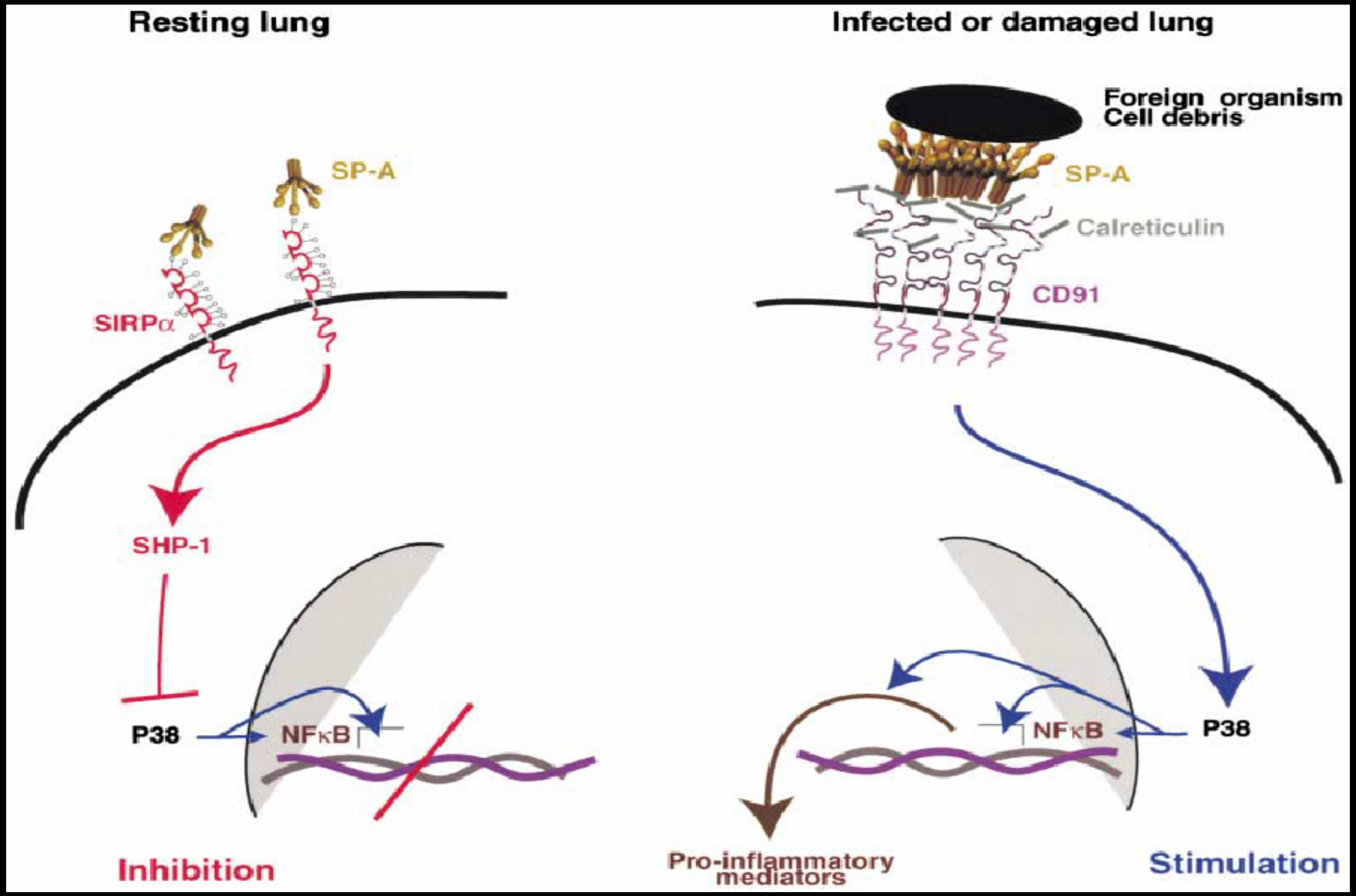
GSH
GST

GSH fluctuations:

AM: ↑ [GSH]

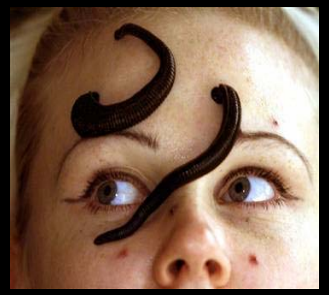
PM: ↓ [GSH]

HOW? SP-A and inflammation



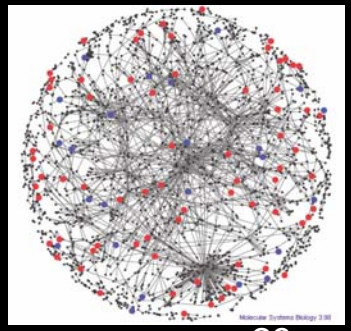
Shift in Philisophy of Medicine...

Tr. Chinese Medicine



1800s: Observe
Effect on whole organism

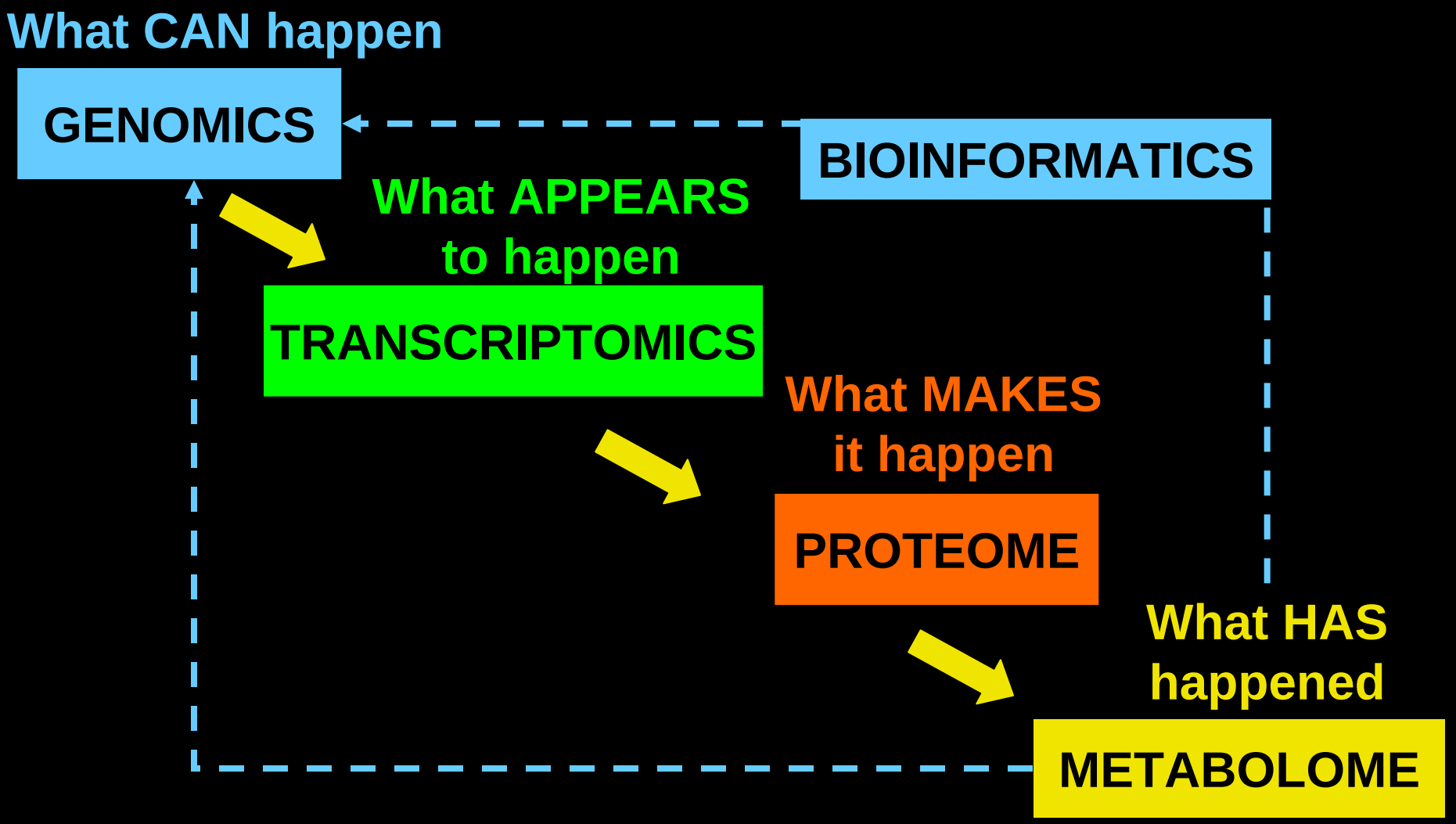
1900s: Manipulate, dissect
Molecular medicine



2000s: Integrate, whole
Systems biology

Personalized Medicine?

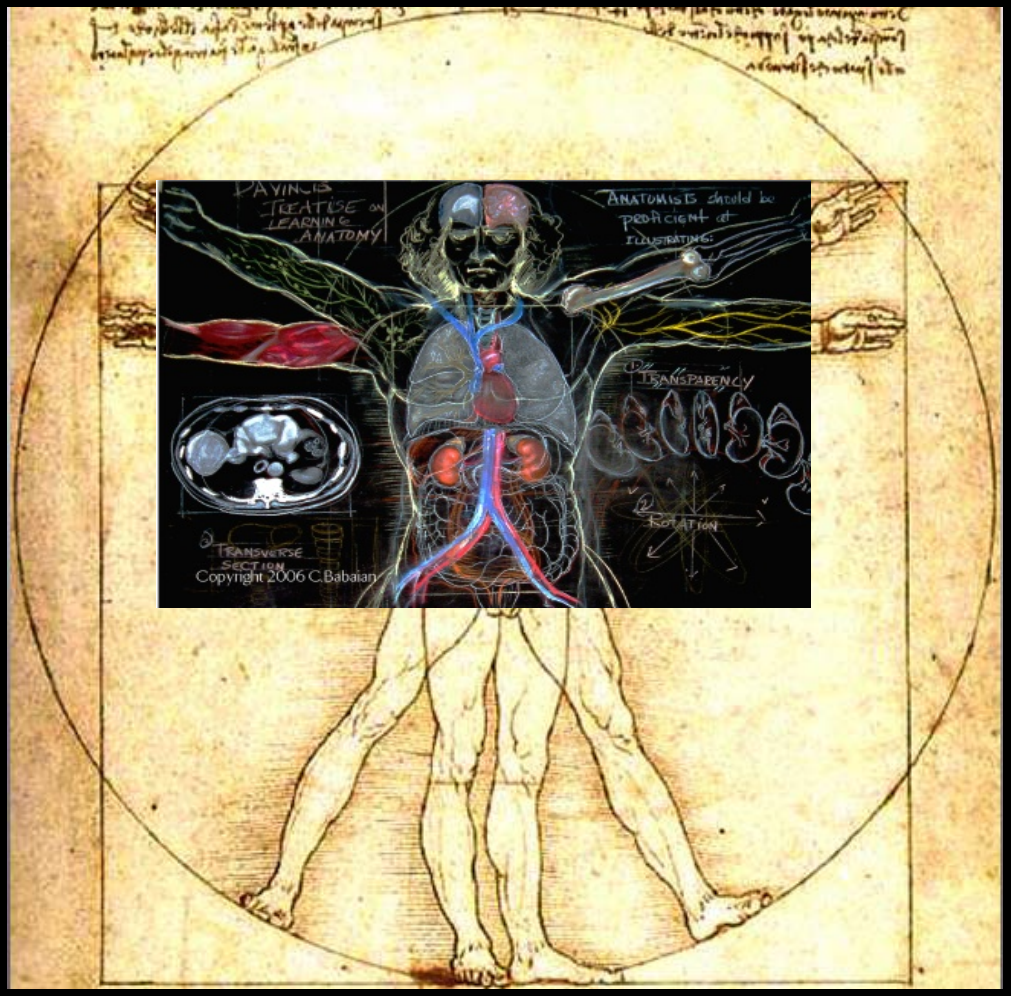
The Omics Cascade ≠ Systems Biology



KEGG – Kyoto Encyclopedia of Genes and Genomes

240 organisms
20,000 organism-specific pathways
782,135 genes

Tools to see the big picture



Systems Biology is about puting together rather than taking apart, integration rather than reduction

- Denis Noble