SciLifeLab

An insight into computational and statistical mass spectrometry-based proteomics

Lukas Käll
Royal Institute of Technology - KTH
School of Biotechnology
Stockholm, Sweden

KTH
VETENSKAP
OCH KONST

http://per-colator.com
http://kaell.org

Outline

1. What is proteomics?
2. Background on Mass spectrometry
3. Peptide identification in shotgun proteomics
4. Multiple hypothesis corrections
5. The statistics of shotgun proteomics
6. Some open problems

Outline

1. What is proteomics?
2. Background on Mass spectrometry
3. Peptide identification in shotgun proteomics
4. Multiple hypothesis corrections
5. The statistics of shotgun proteomics
6. Some open problems

Central Dogma

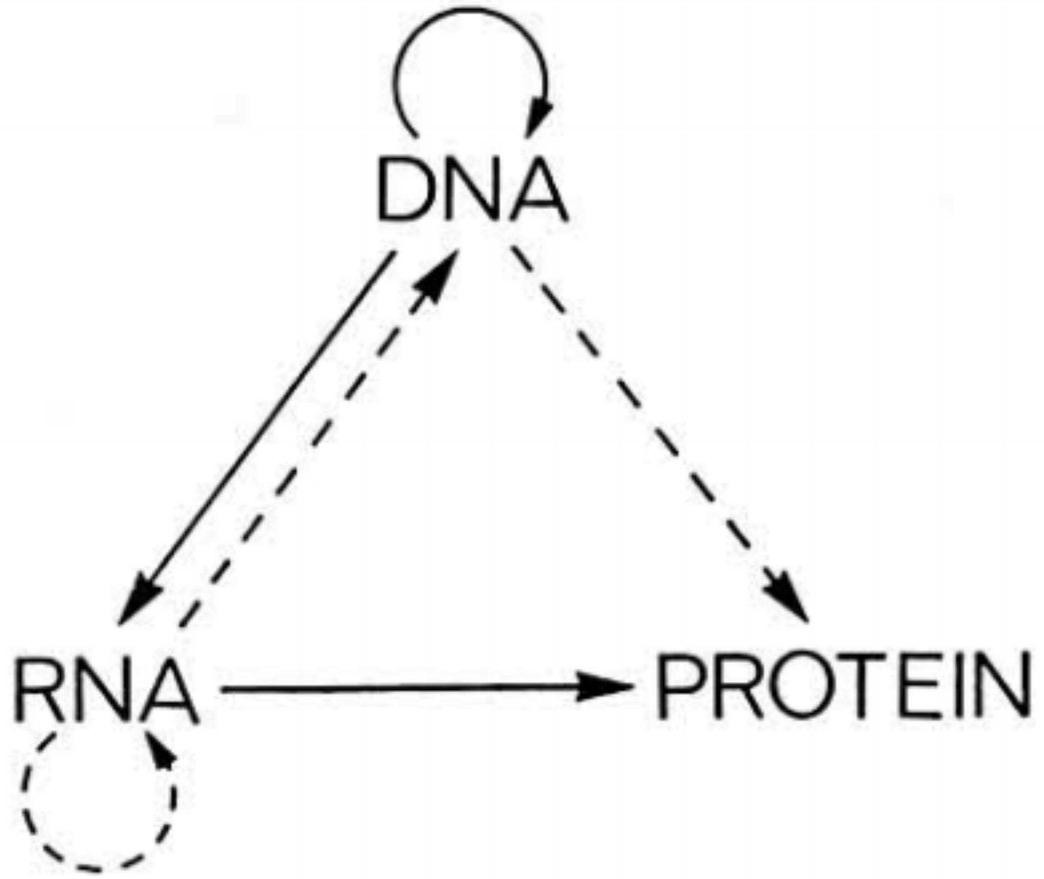
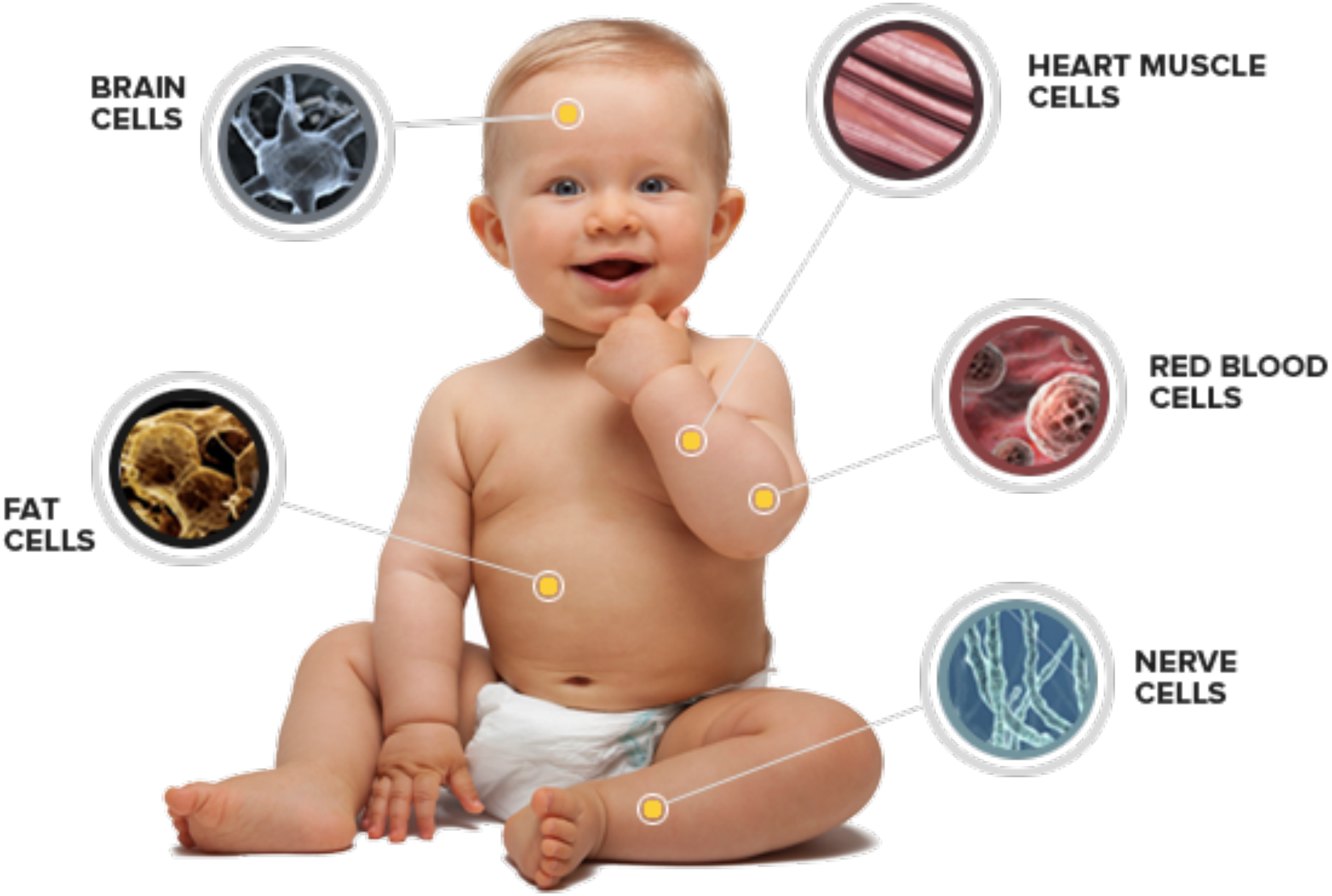


Fig. 3. A tentative classification for the present day. Solid arrows show general transfers; dotted arrows show special transfers. Again, the absent arrows are the undetected transfers specified by the central dogma.

[Crick, Nature 1970]

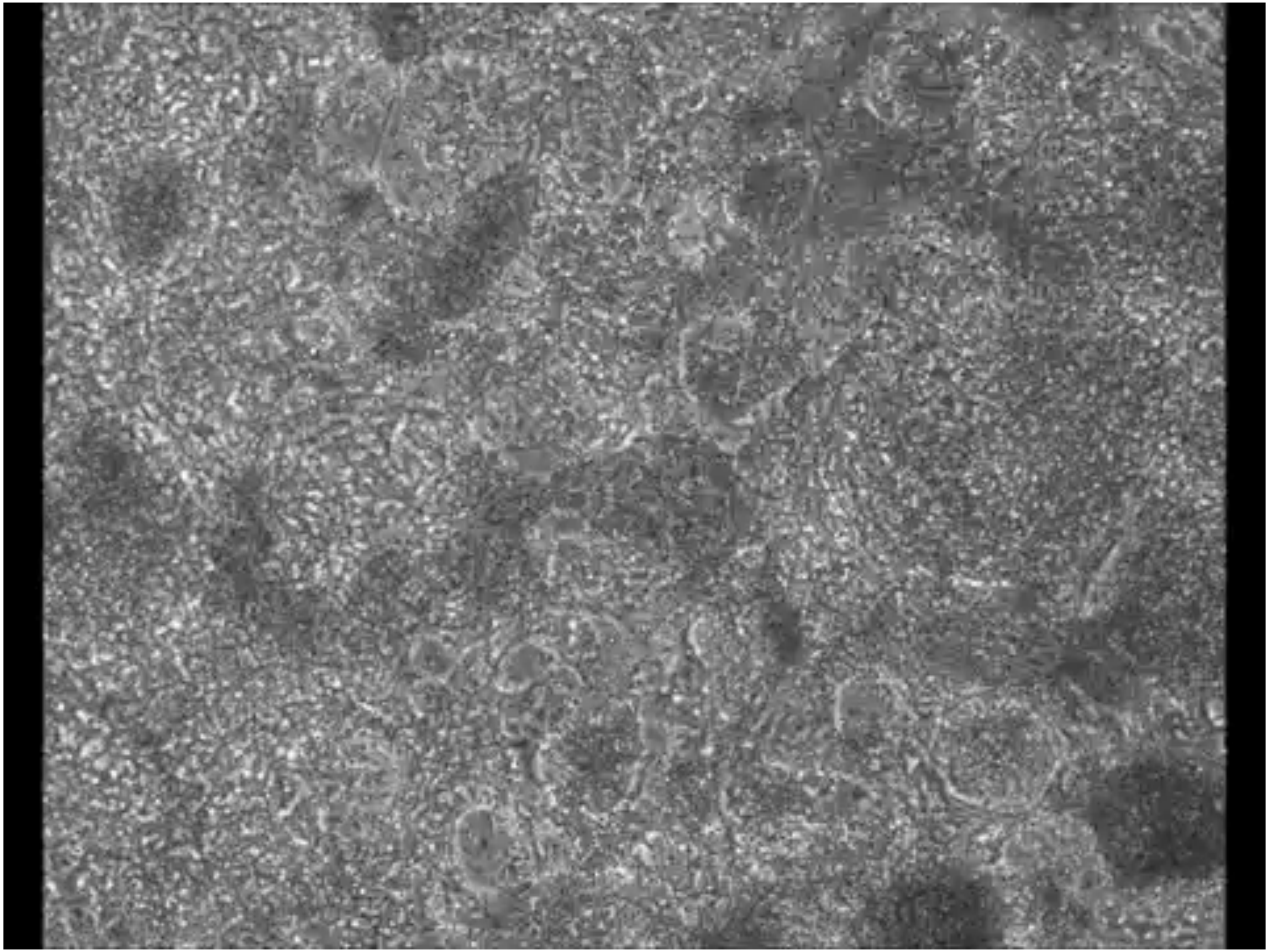
DNA -> RNA -> Proteins

Same DNA, different configuration of proteins



https://youtu.be/jEtaqmW3ZK4

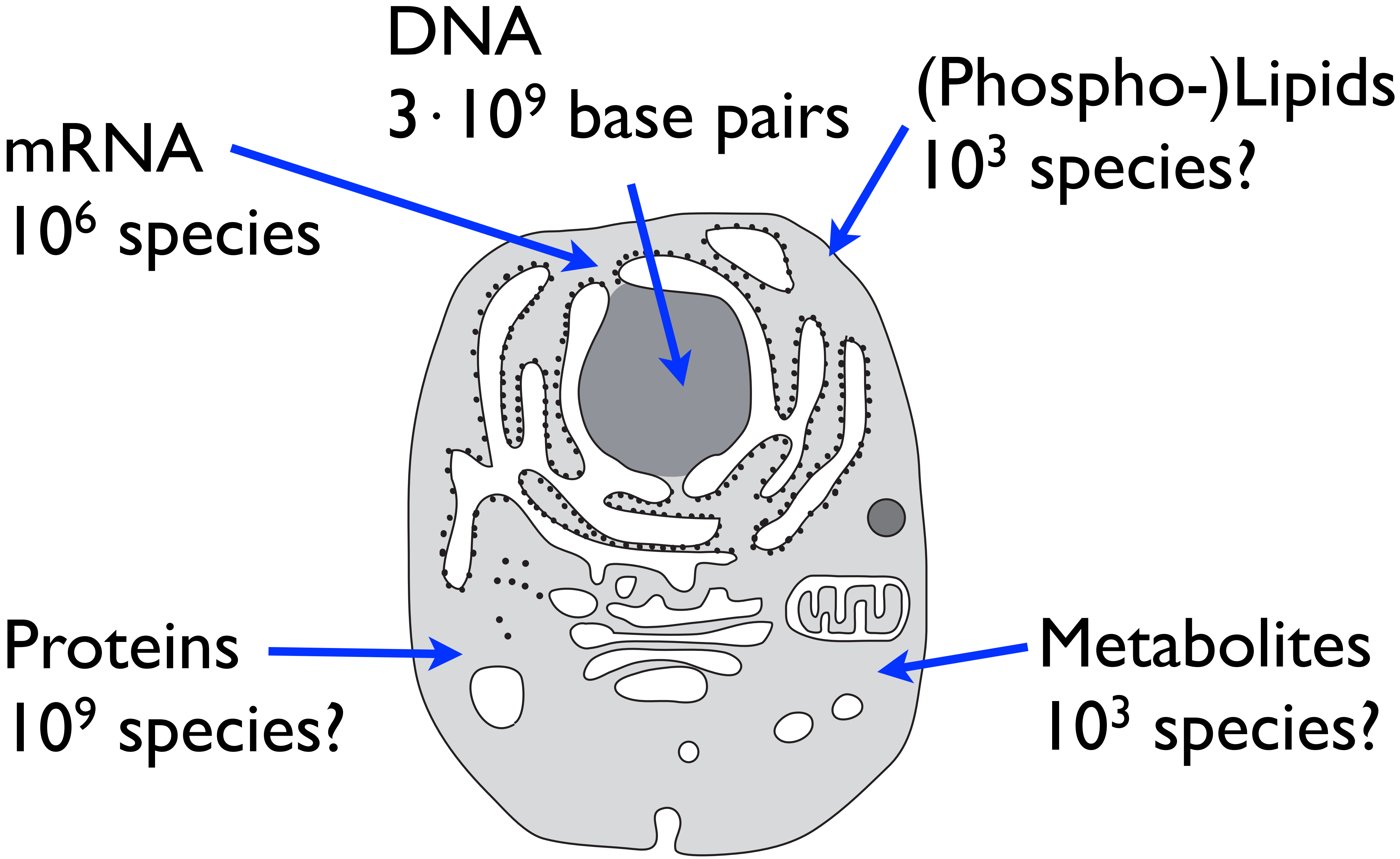
Pluripotent stem cells reprogrammed as cardiomyocytes



Same DNA, different configuration of proteins

An organism’s proteins are closer its DNA to its phenotype, *i.e.* its observable traits

A human cell - a system



DNA
$3 \cdot 10^9$ base pairs

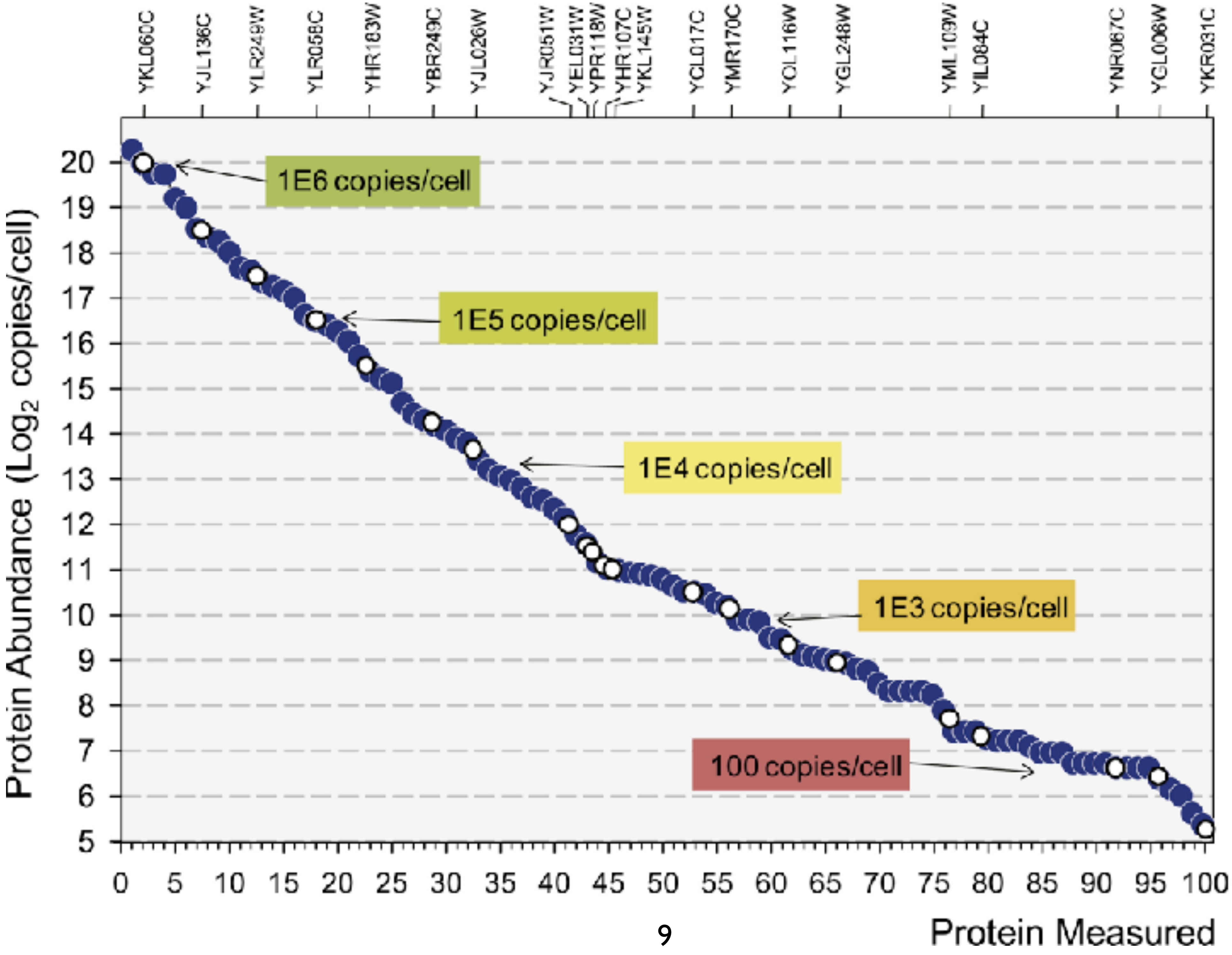
mRNA
10^6 species

(Phospho-)Lipids
10^3 species?

Proteins
10^9 species?

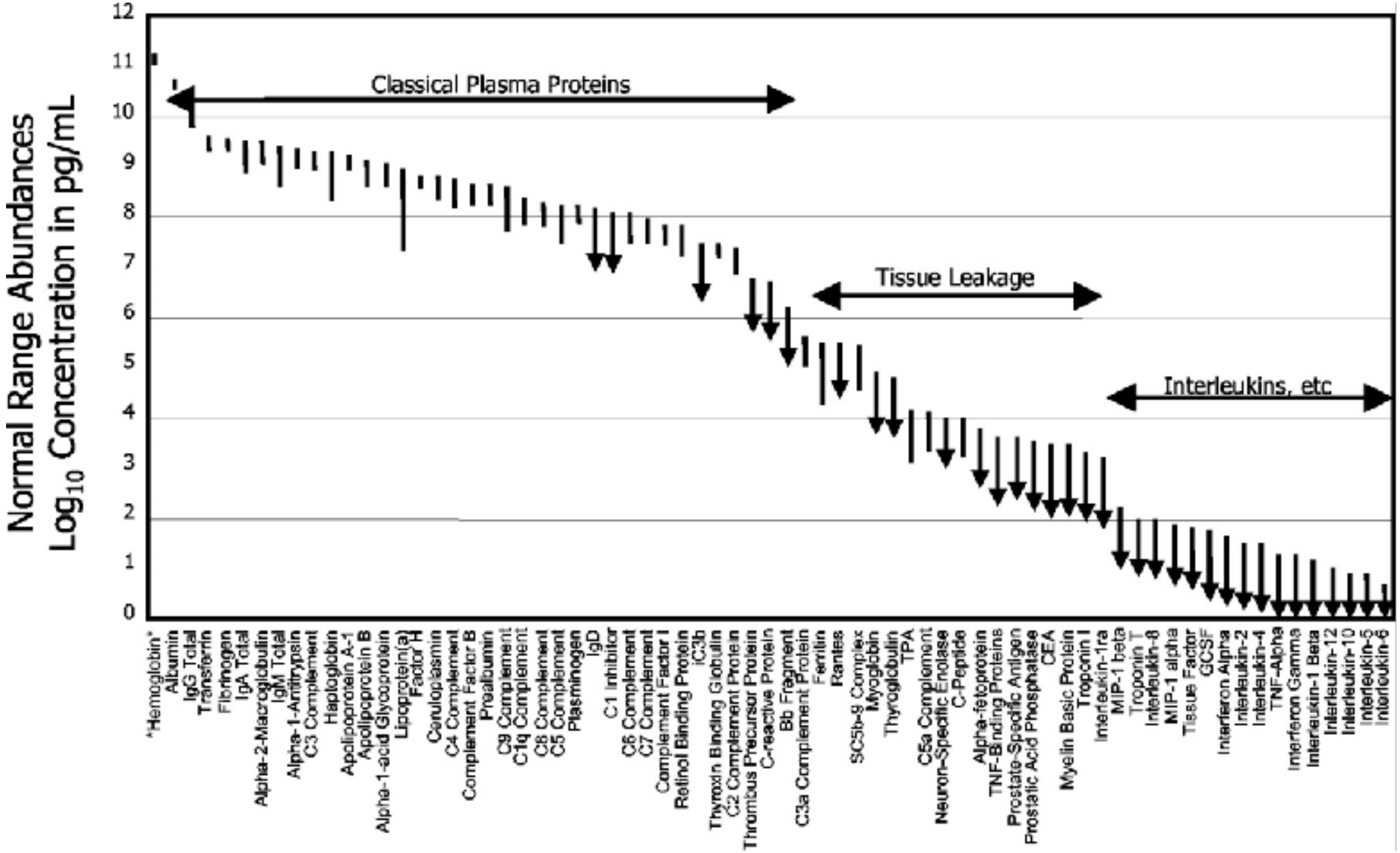
Metabolites
10^3 species?

Proteins concentration in yeast range >4 orders of magnitude



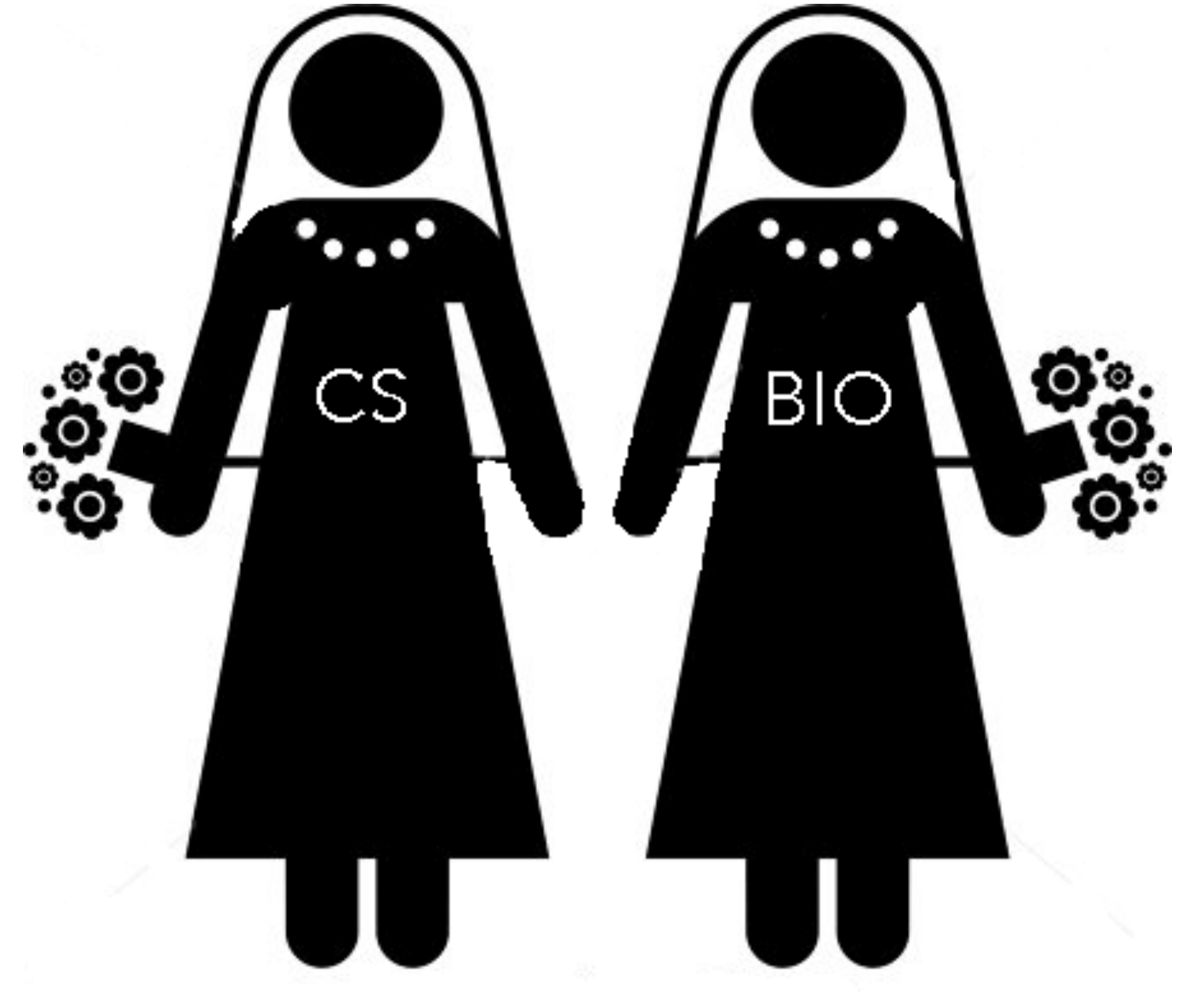
[Picotti et al Cell 2009]

Protein concentration in blood plasma range >10 orders of magnitude



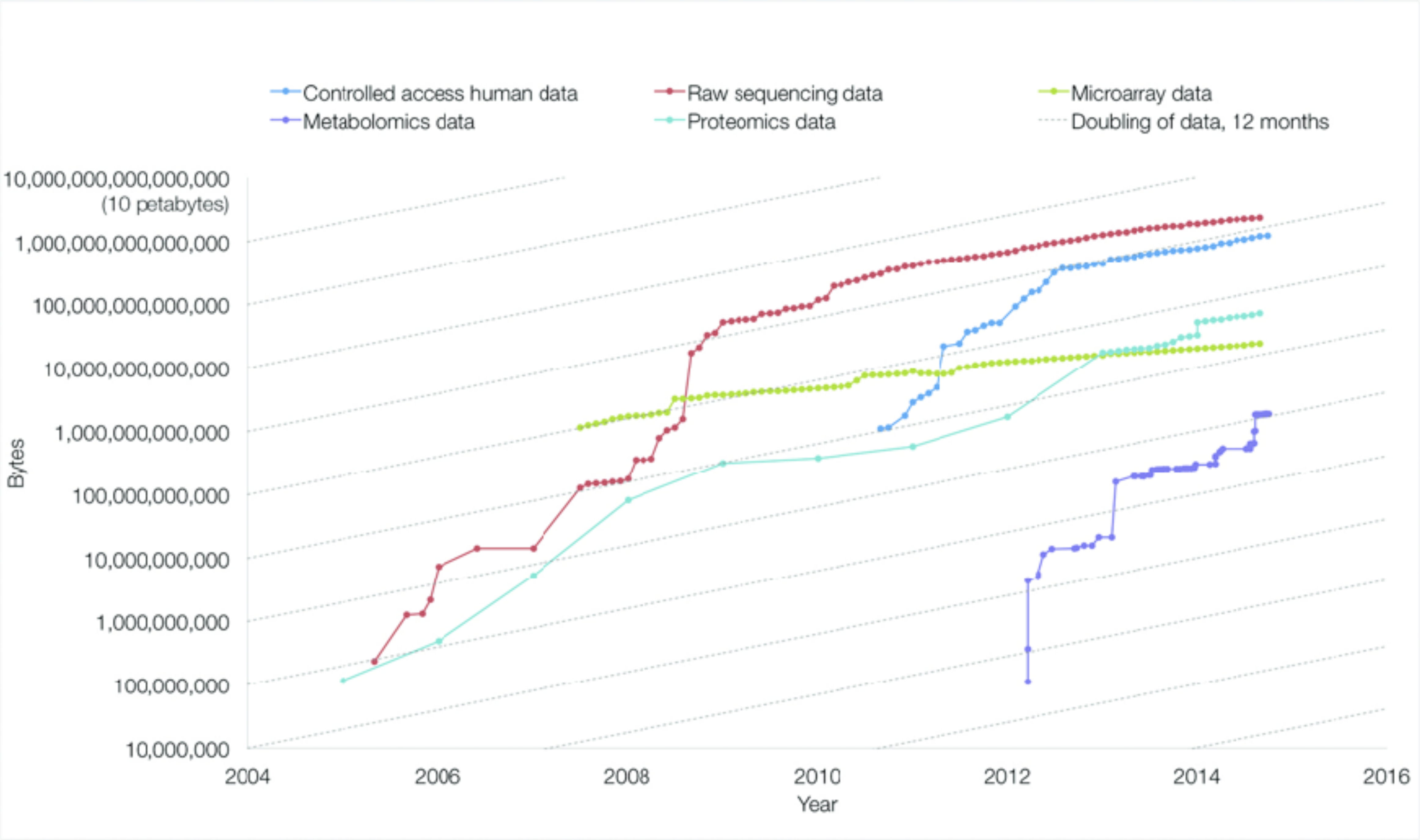
[Andersson & Andersson, MCP 2002]

What is Bioinformatics?



Bioinformatics is an interdisciplinary field that develops and applies computational methods to analyze biological data, to make new predictions or discover new biology.

The amount of biological data is expanding exponentially

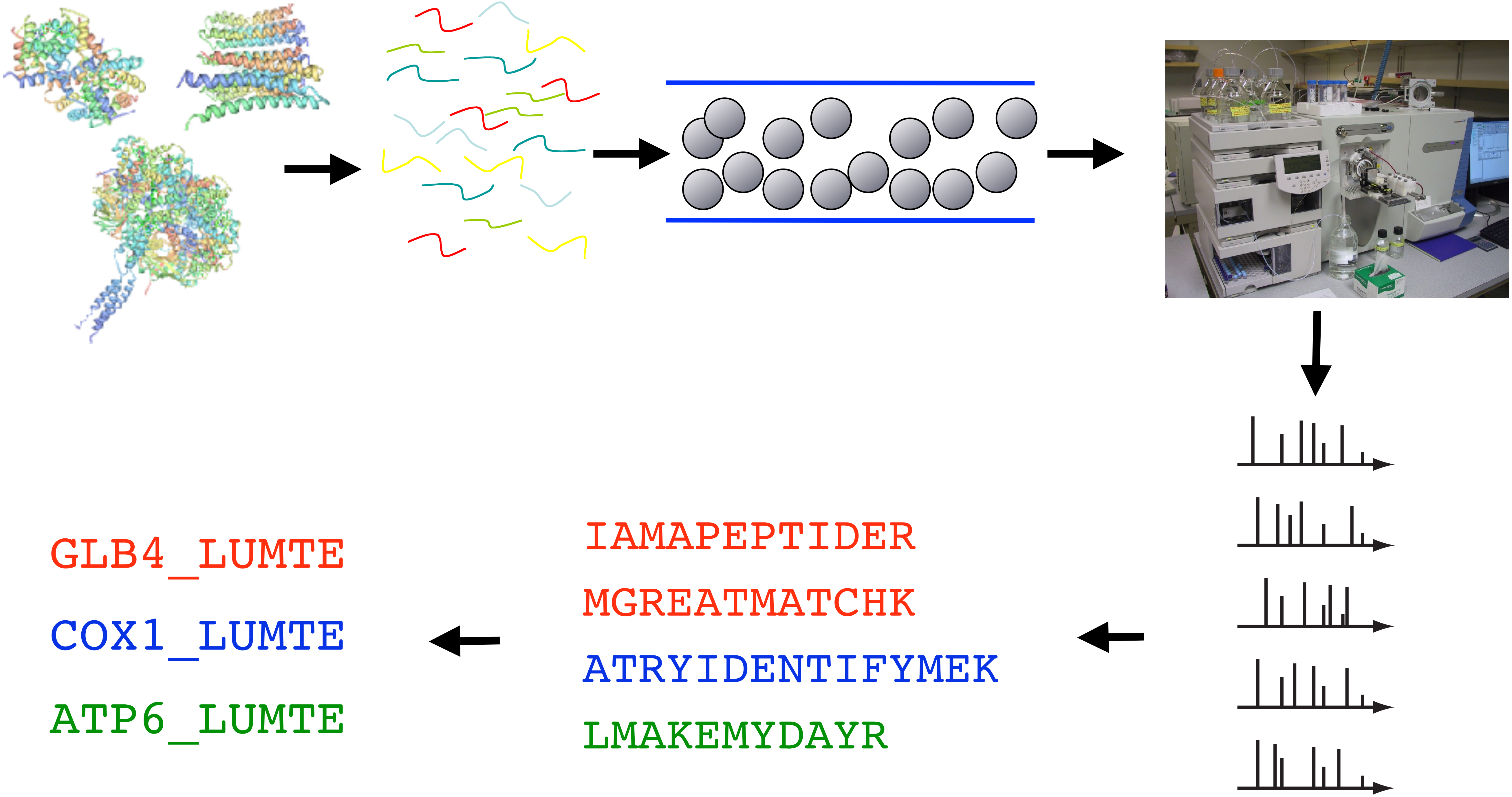


Data growth curves of 5 major EMBL-EBI resources (European Genome-phenome Archive (EGA); European Nucleotide Archive (ENA); Proteomics data repository (PRIDE); Metabolomics resource (MetaboLights); and Functional genomics database (ArrayExpress) over the years 2005-2013. Source: EMBL-EBI.

Outline

1. What is proteomics?
2. Background on Mass spectrometry
3. Peptide identification in shotgun proteomics
4. Multiple hypothesis corrections
5. The statistics of shotgun proteomics
6. Some open problems

Shotgun proteomics



IAMAPEPTIDER

MGREATMATCHK

ATRYIDENTIFYMEK

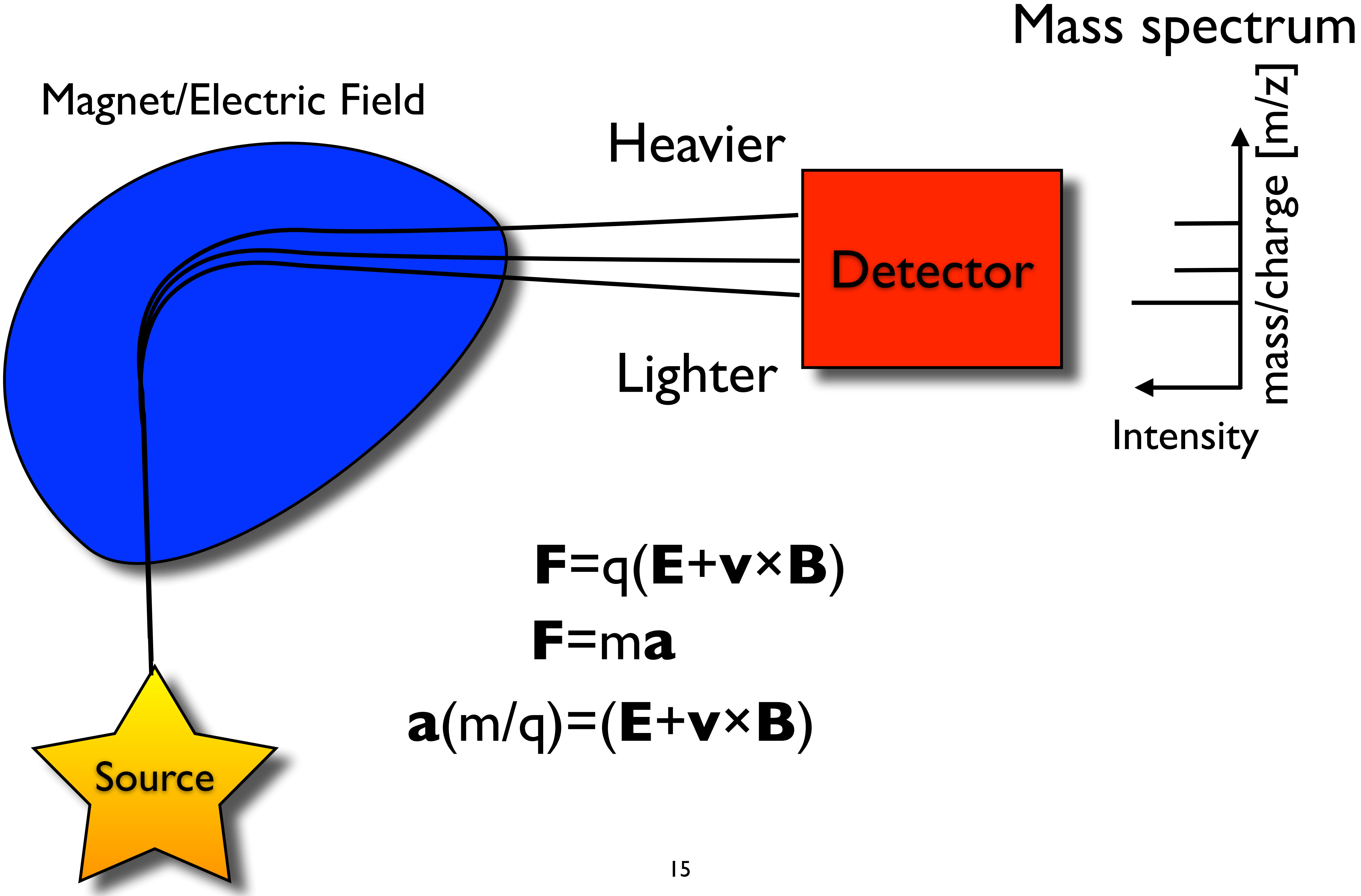
LMAKEMYDAYR

GLB4_LUMTE

COX1_LUMTE

ATP6_LUMTE

Mass spectrometry



Magnet/Electric Field

Heavier

Lighter

Detector

Mass spectrum

mass/charge [m/z]

Intensity

Source

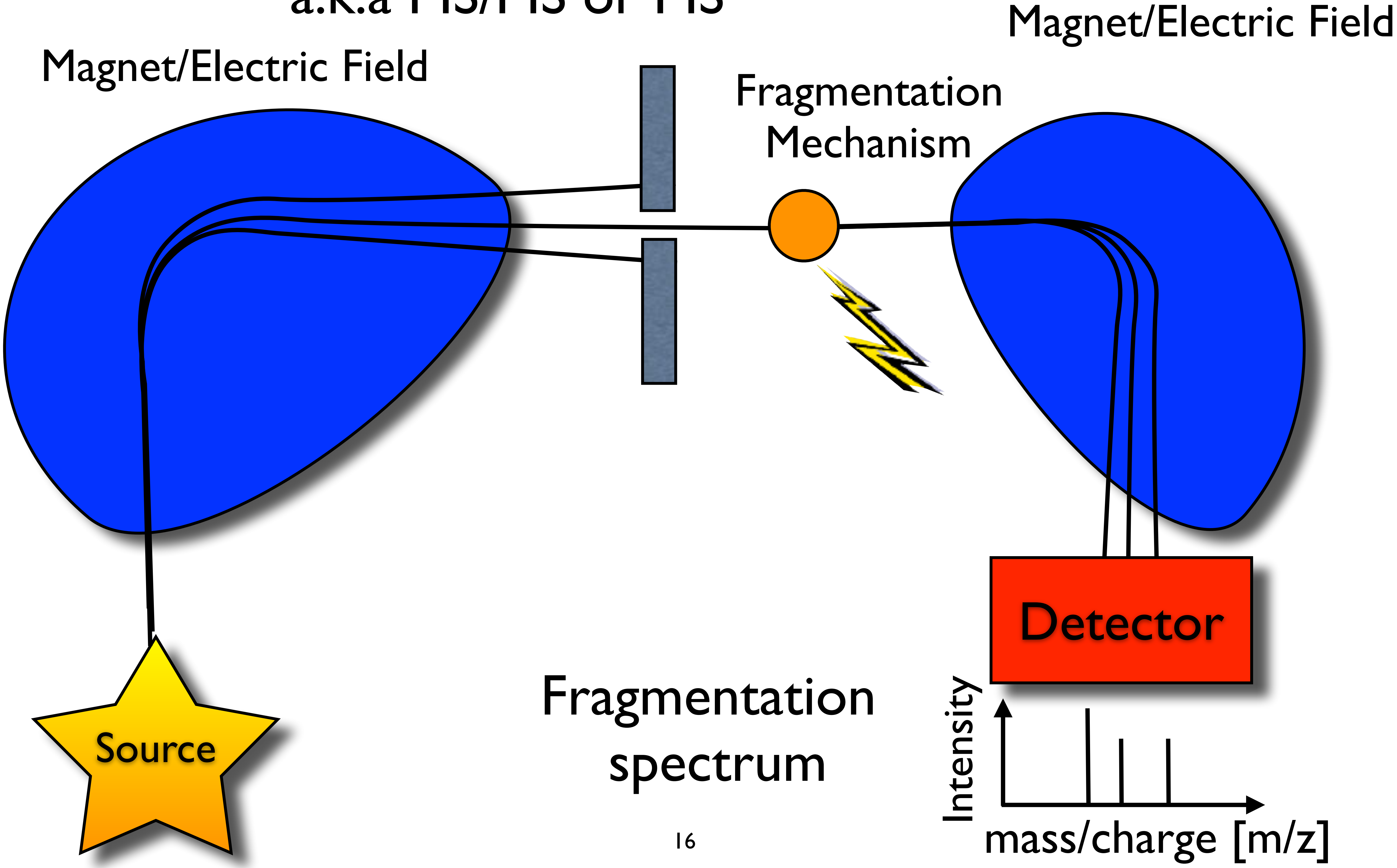
$\mathbf{F}=q(\mathbf{E}+\mathbf{v}\times\mathbf{B})$

$\mathbf{F}=m\mathbf{a}$

$\mathbf{a}(m/q)=(\mathbf{E}+\mathbf{v}\times\mathbf{B})$

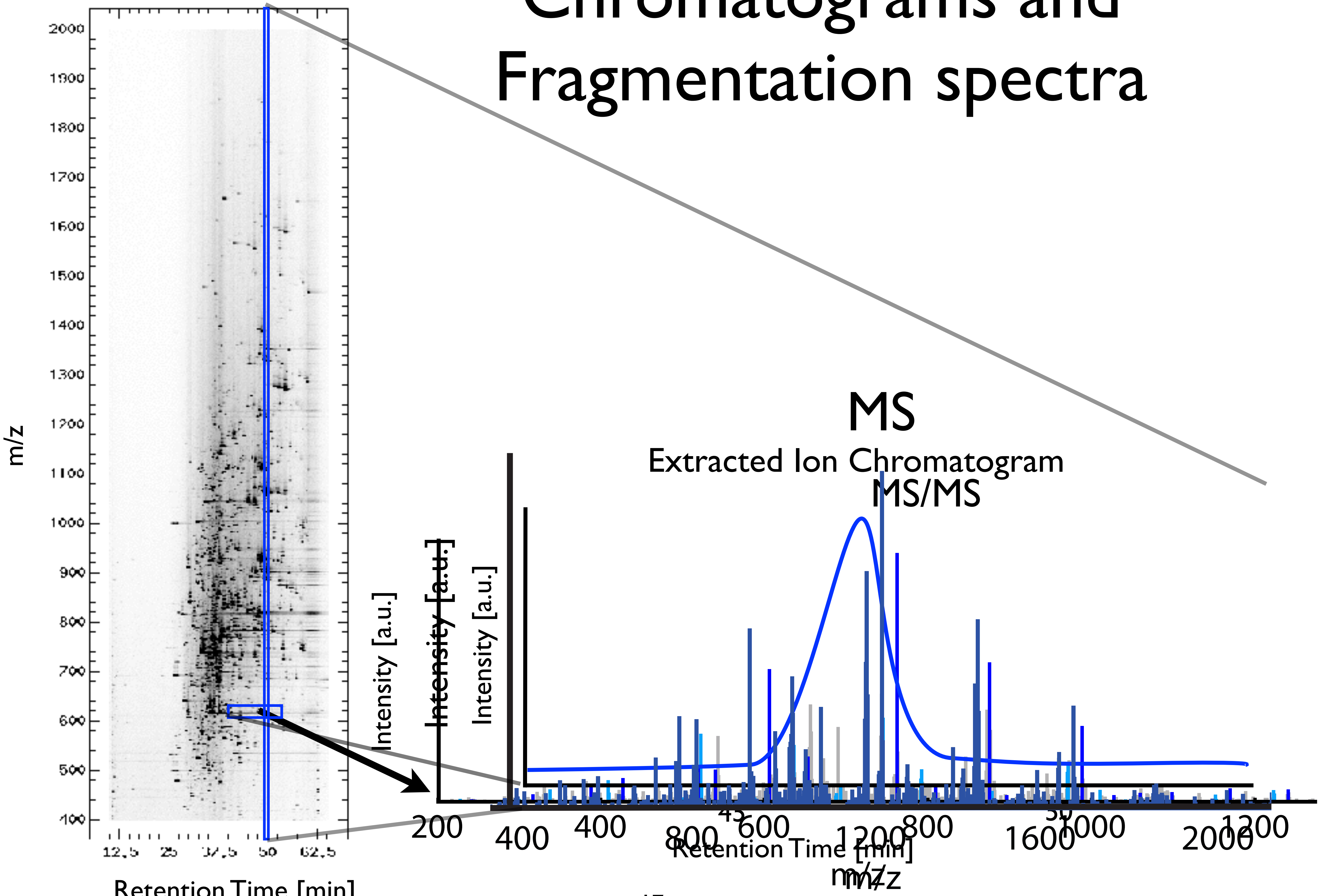
Tandem mass spectrometry

a.k.a MS/MS or MS^2



Fragmentation spectrum

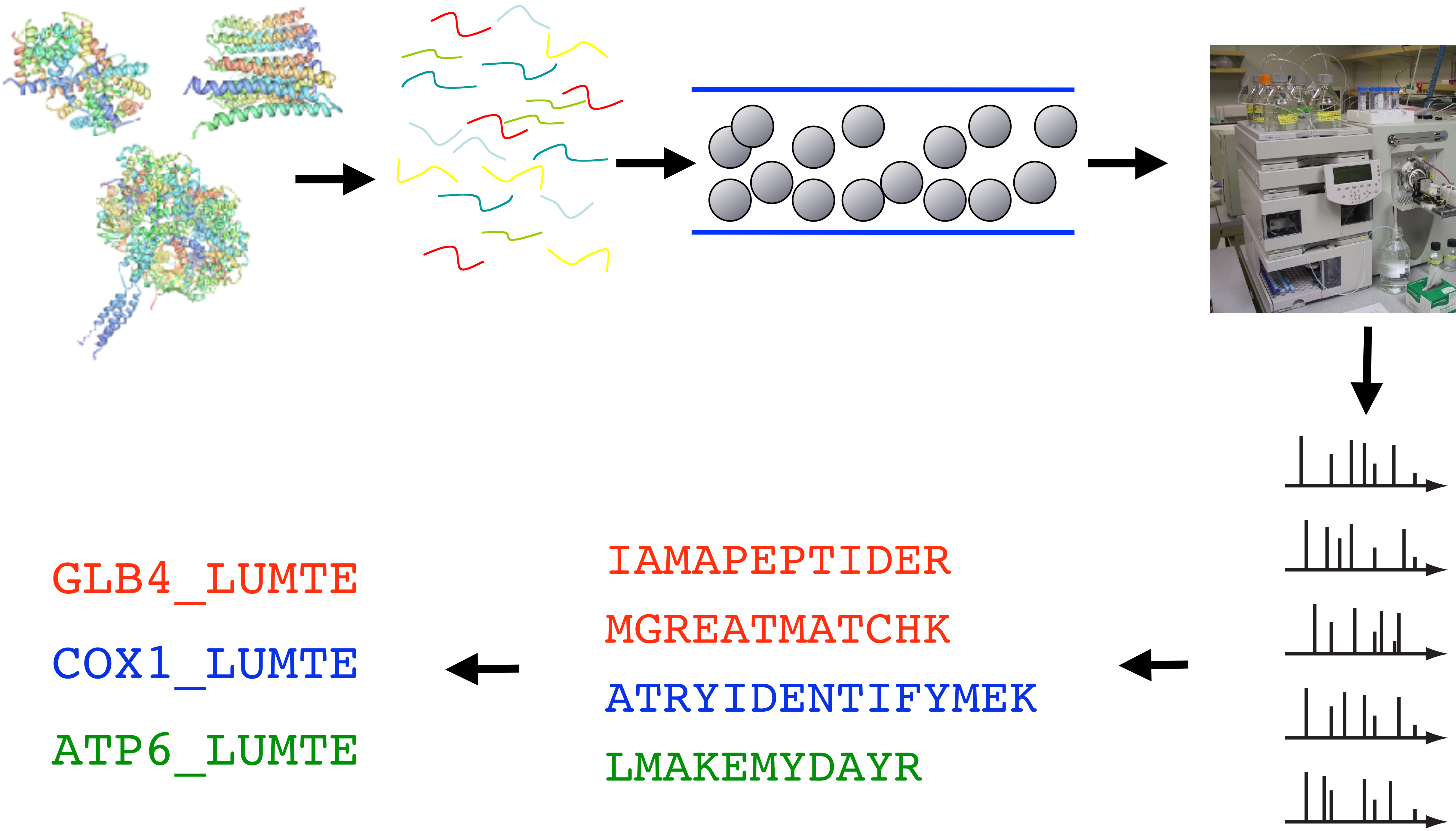
Chromatograms and Fragmentation spectra



Outline

1. What is proteomics?
2. Background on Mass spectrometry
3. Peptide identification in shotgun proteomics
4. Multiple hypothesis corrections
5. The statistics of shotgun proteomics
6. Some open problems

Shotgun proteomics



IAMAPEPTIDER

MGREATMATCHK

ATRYIDENTIFYMEK

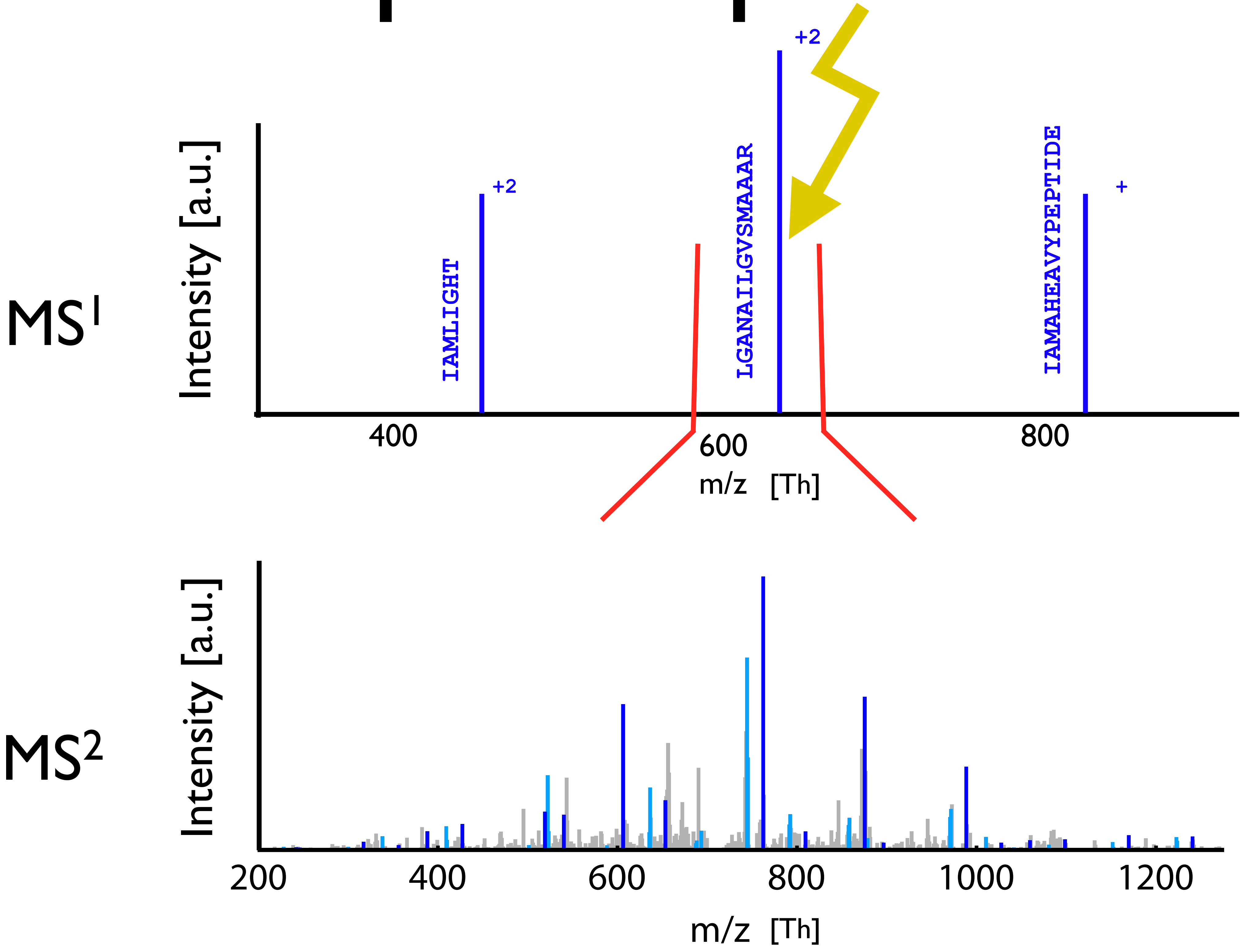
LMAKEMYDAYR

GLB4_LUMTE

COX1_LUMTE

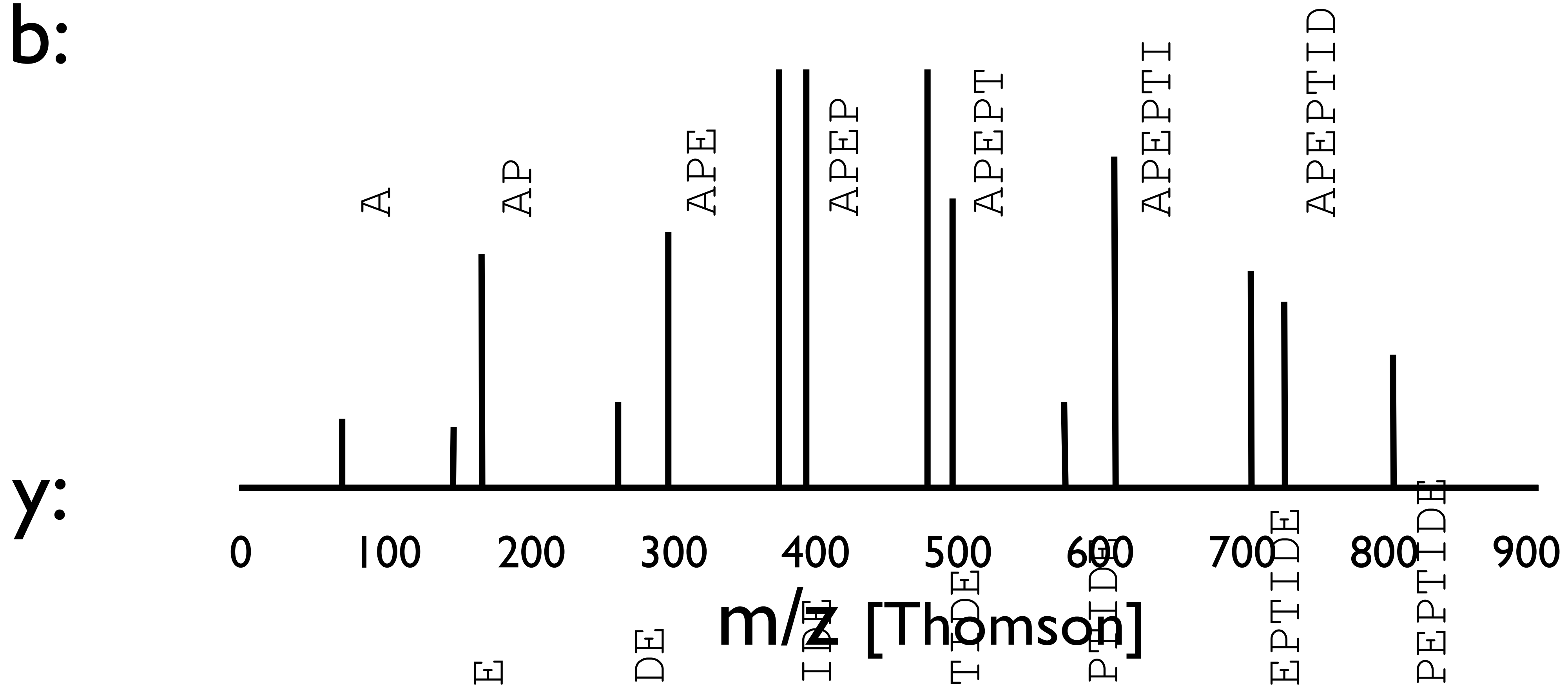
ATP6_LUMTE

Peptide spectra

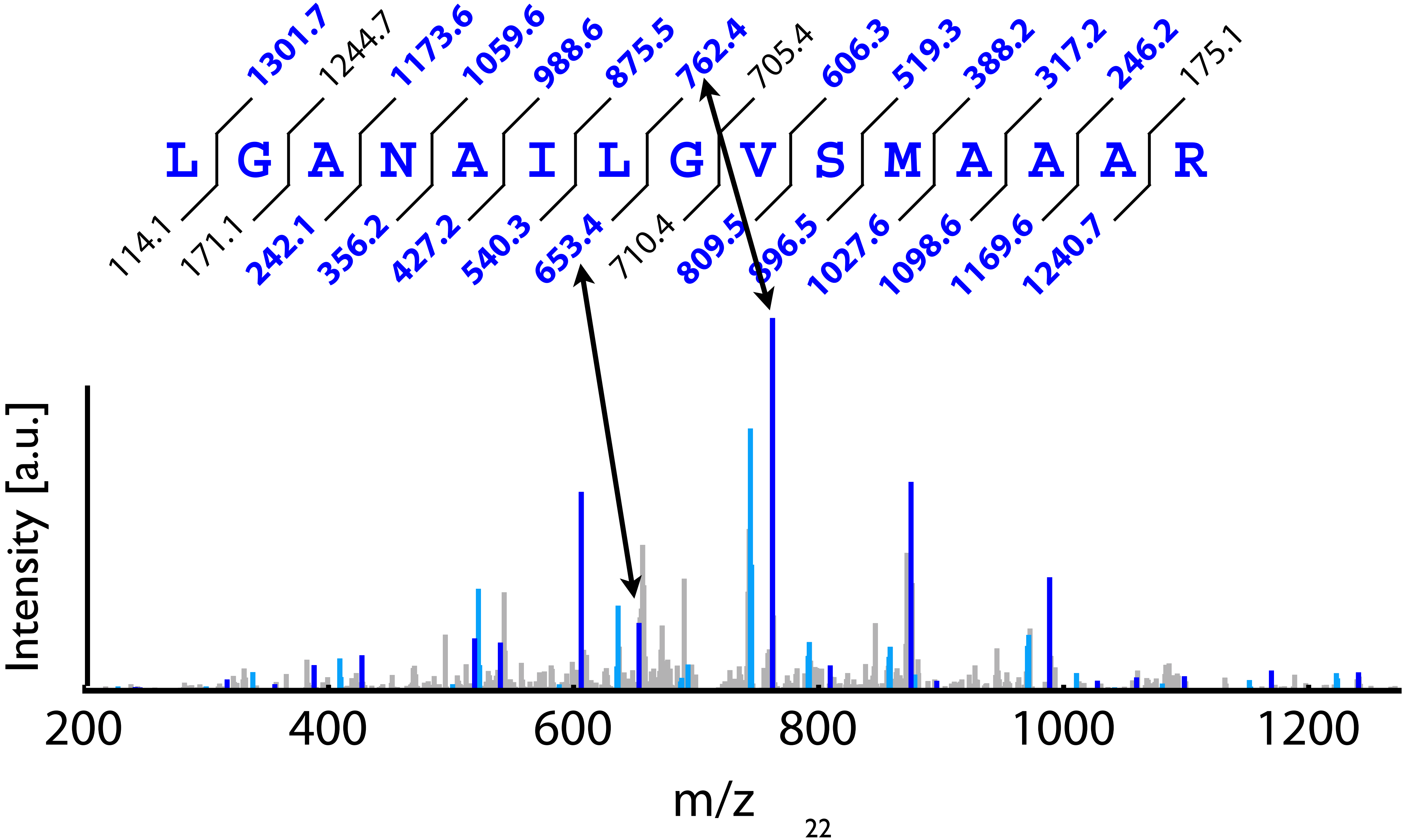


Fragmentation Spectrum

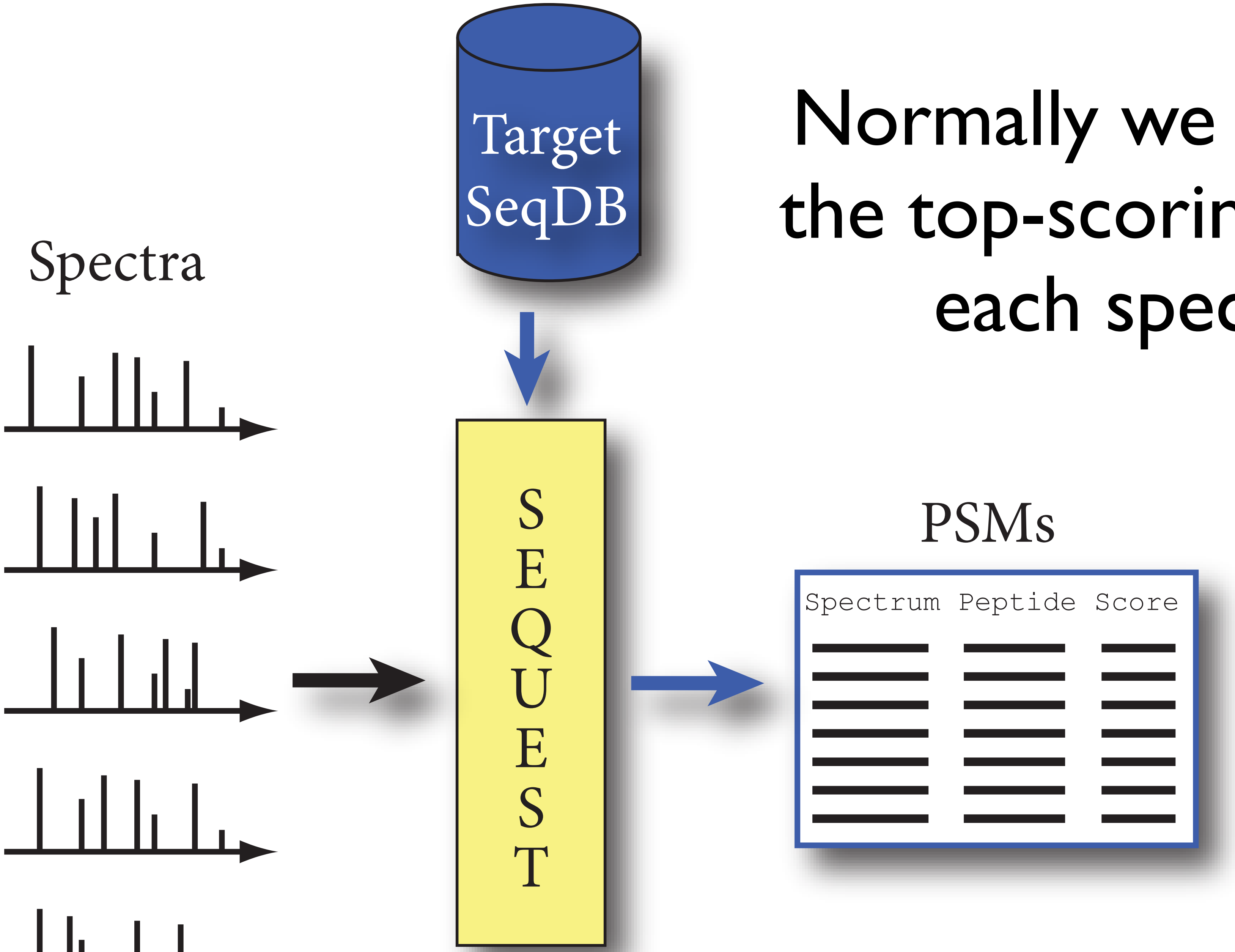
A|P|E|P|T|I|D|E



Peptide fragmentation spectrum



Peptide identification

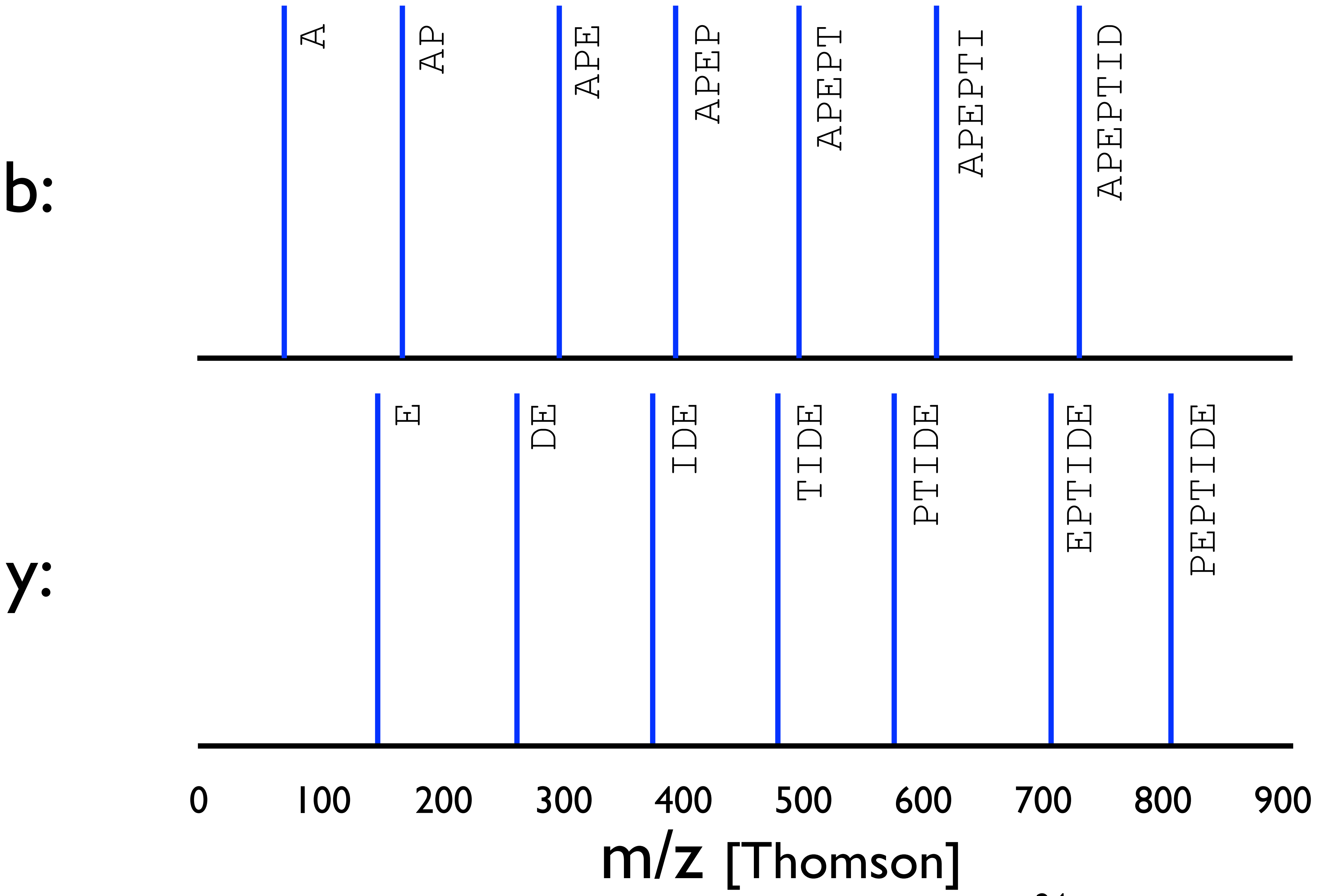


Normally we keep only the top-scoring PSM for each spectrum

PSM - Peptide Spectrum Match

Theoretical Spectrum of a peptide

A|P|E|P|T|I|D|E



Search engine

SEQUEST:

Observed Spectrum

Scoring Function

$$\text{matched\_peptide}(s,D) = \underset{p\in D}{\operatorname{argmax}}\, f(s,T(p))$$

Database of peptides

peptide

Theoretical spectrum of *p*

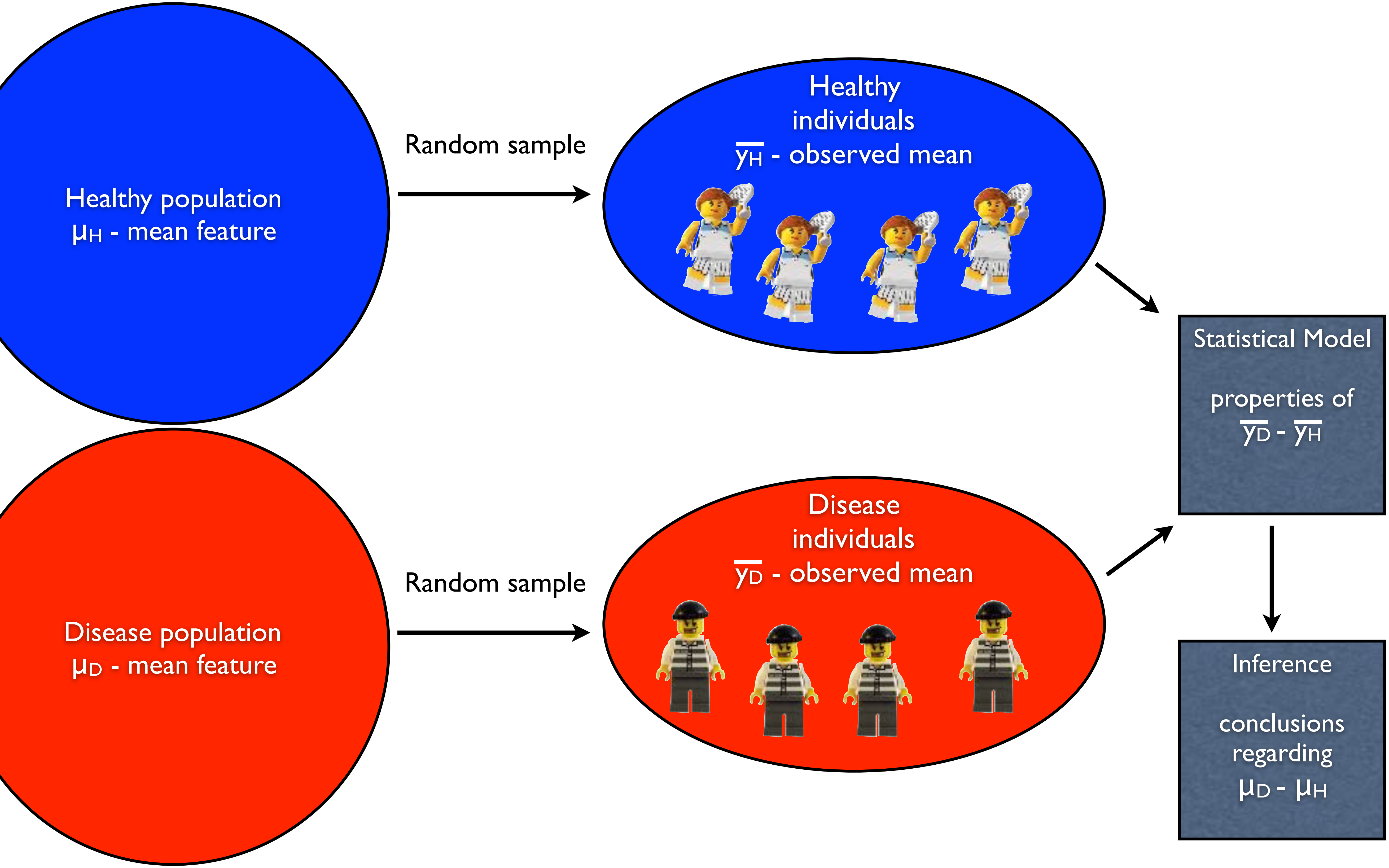
other:

$$\text{matched\_peptide}(s,D) = \underset{p\in D}{\operatorname{argmax}}\, f(s,p)$$

Outline

1. What is proteomics?
2. Background on Mass spectrometry
3. Peptide identification in shotgun proteomics
4. Multiple hypothesis corrections
5. The statistics of shotgun proteomics
6. Some open problems

Statistical inference procedure



Healthy population
μ_H - mean feature

Random sample

Healthy individuals
$\overline{y}_H$ - observed mean

Disease population
μ_D - mean feature

Random sample

Disease individuals
$\overline{y}_D$ - observed mean

Statistical Model

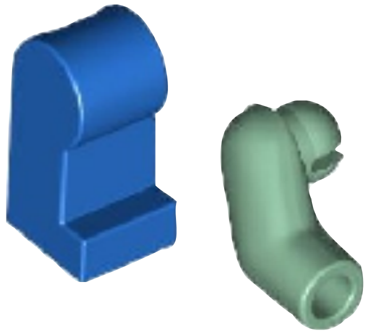
properties of
$\overline{y}_D - \overline{y}_H$

Inference

conclusions regarding
$\mu_D - \mu_H$

Hypothesis testing

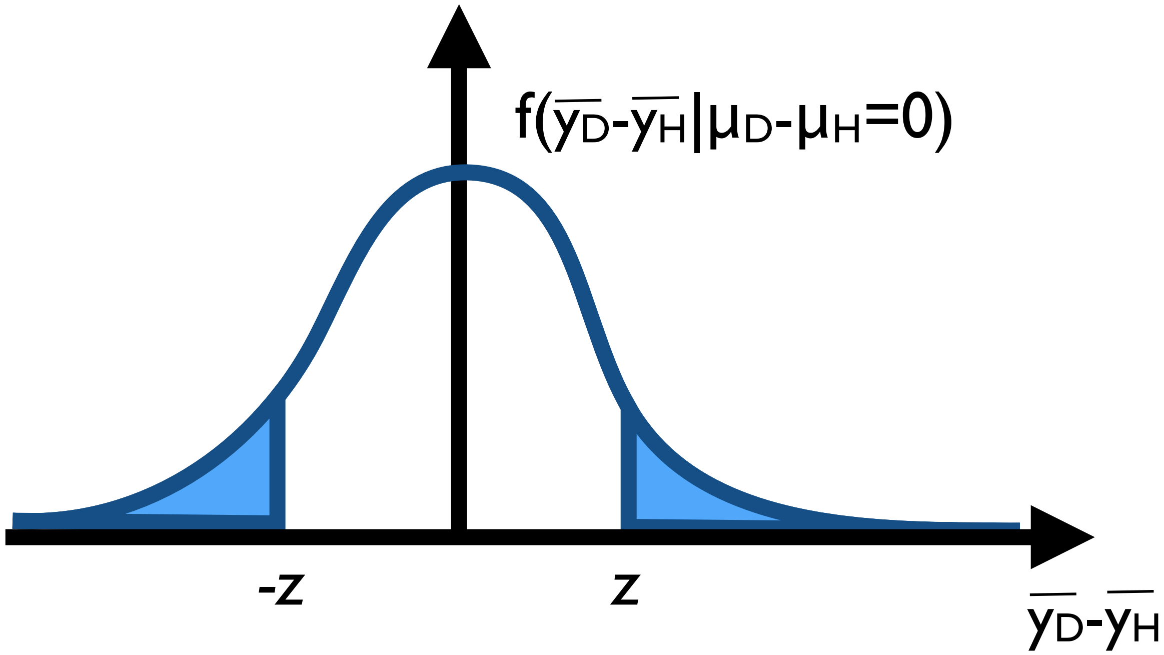
- H_0: The *null* hypothesis. The situation we are not interested in (typically $\mu_D - \mu_H = 0$)



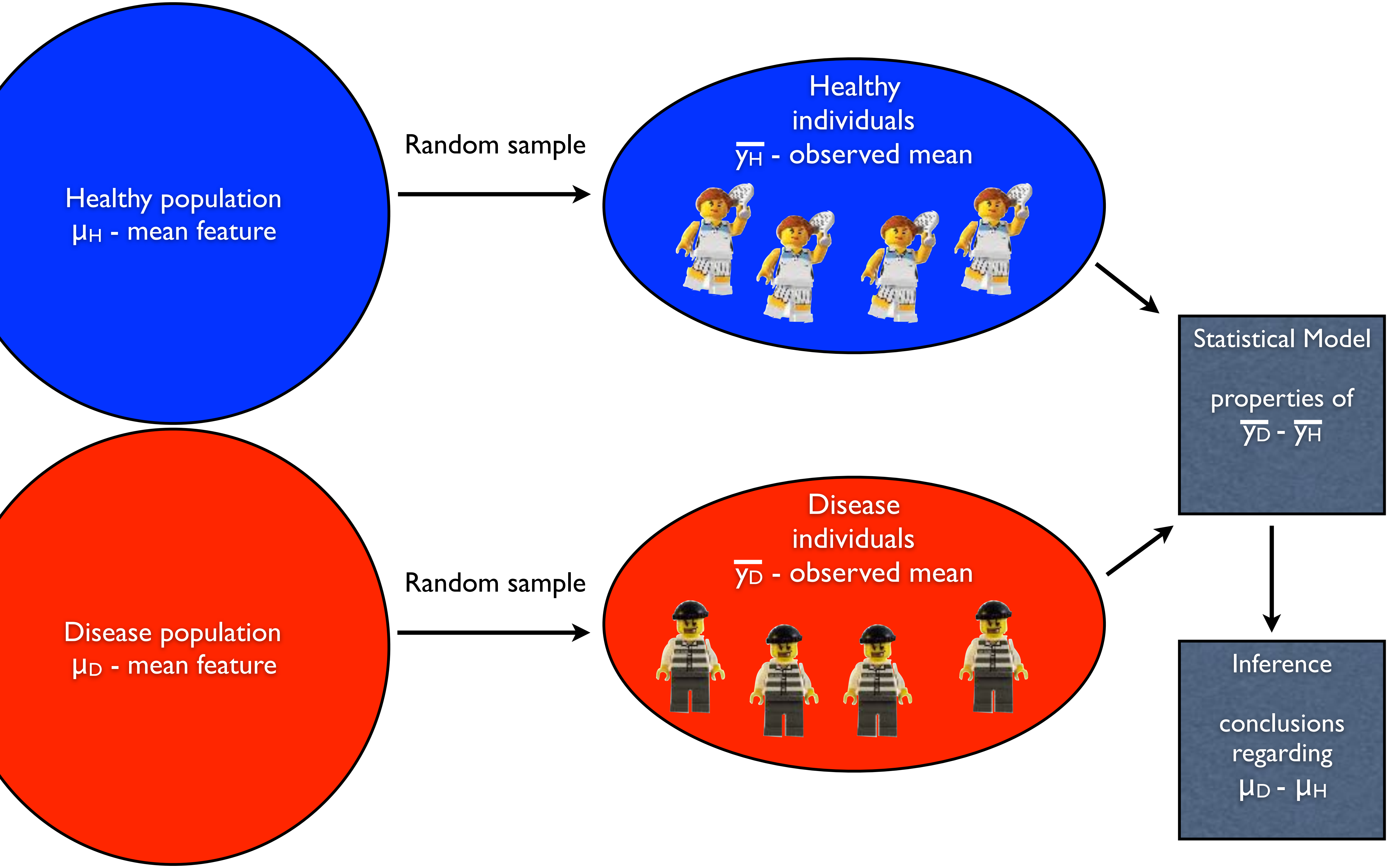
- H_1: The *alternative* hypothesis. The situation we want to detect (typically $\mu_D - \mu_H \neq 0$)

p value

- $\Pr(|\bar{y}_D - \bar{y}_H| \geq z | \mu_D - \mu_H = 0)$, *i.e.* the probability to a result at least as extreme as the one that was observed given H_0.
- *p* values are uniformly distributed under H_0.



Statistical inference procedure



Healthy population
μ_H - mean feature

Random sample

Healthy individuals
$\bar{y}_H$ - observed mean

Disease population
μ_D - mean feature

Random sample

Disease individuals
$\bar{y}_D$ - observed mean

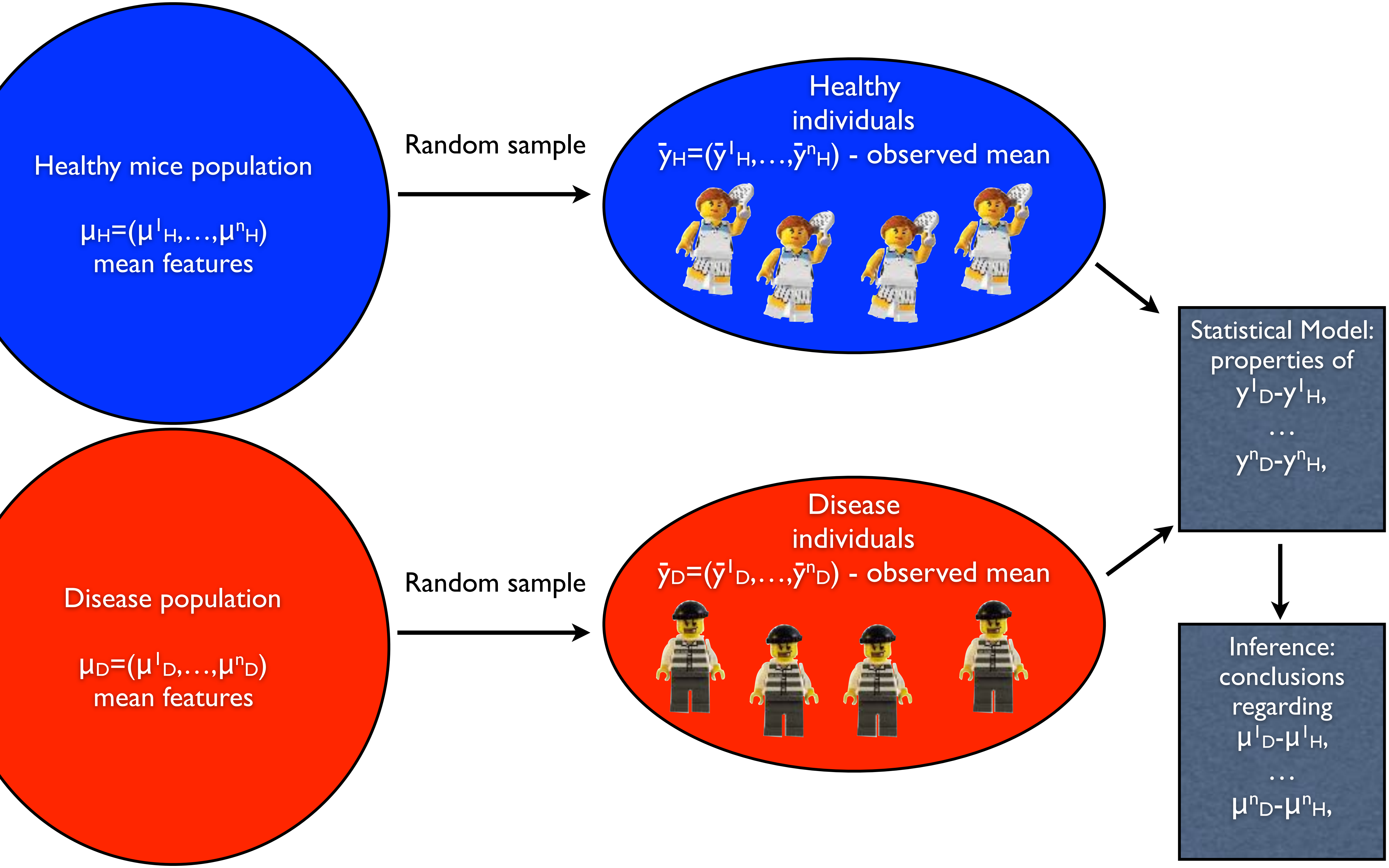
Statistical Model

properties of
$\bar{y}_D - \bar{y}_H$

Inference

conclusions regarding
$\mu_D - \mu_H$

Multiple measurements per sampled individual



Healthy mice population

$\mu_H=(\mu^1_H,\ldots,\mu^n_H)$
mean features

Random sample

Healthy
individuals
$\bar{y}_H=(\bar{y}^1_H,\ldots,\bar{y}^n_H)$ - observed mean

Statistical Model:
properties of
$y^1_D-y^1_H,$
…
$y^n_D-y^n_H,$

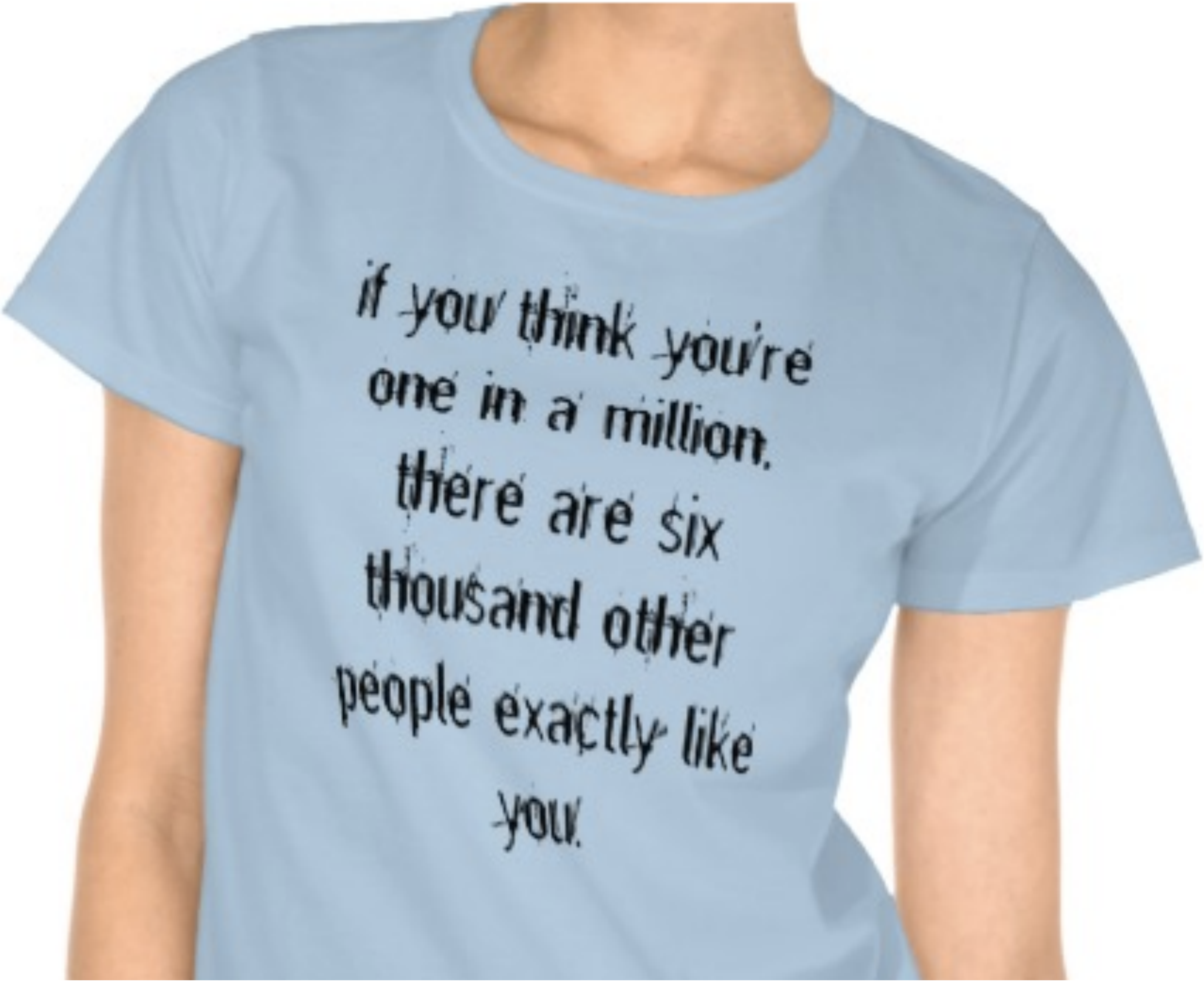
Disease population

$\mu_D=(\mu^1_D,\ldots,\mu^n_D)$
mean features

Random sample

Disease
individuals
$\bar{y}_D=(\bar{y}^1_D,\ldots,\bar{y}^n_D)$ - observed mean

Inference:
conclusions
regarding
$\mu^1_D-\mu^1_H,$
…
$\mu^n_D-\mu^n_H,$


if you think you're
one in a million.
there are six
thousand other
people exactly like
you.

False Discovery Rate

score	type
0.0001	alternative (H_1)
0.00015	alternative (H_1)
0.00017	alternative (H_1)
0.0002	alternative (H_1)
0.00022	null (H_0)
0.00023	alternative (H_1)
0.00034	alternative (H_1)
0.00042	alternative (H_1)
0.00046	null (H_0)
0.00055	alternative (H_1)
0.00065	null (H_0)
0.00073	alternative (H_1)
0.00084	null (H_0)
...	...

threshold

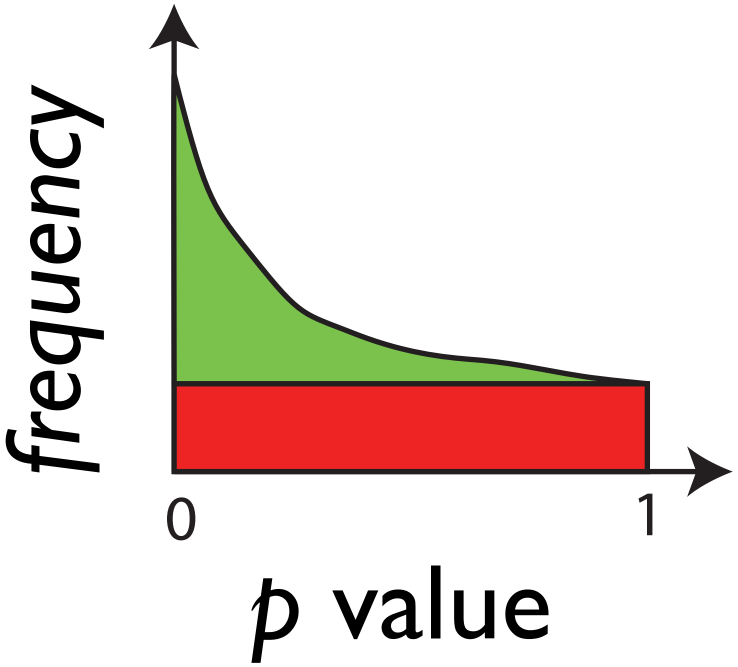
$$\frac{2}{10}$$

FDR(x) is the expectation value of the fraction of tests below threshold *x* that are generated under the null hypothesis

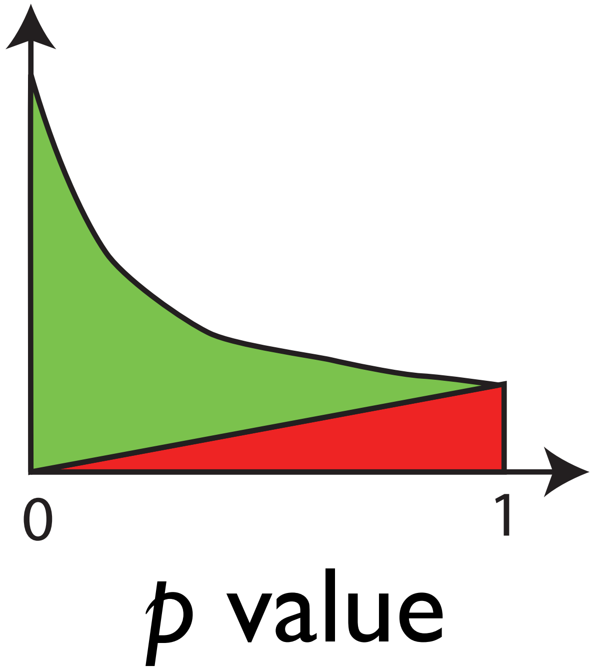
Concept test: distribution of *p* values

Which of the following histograms would be a likely outcome from a well calibrated high throughput experiment?

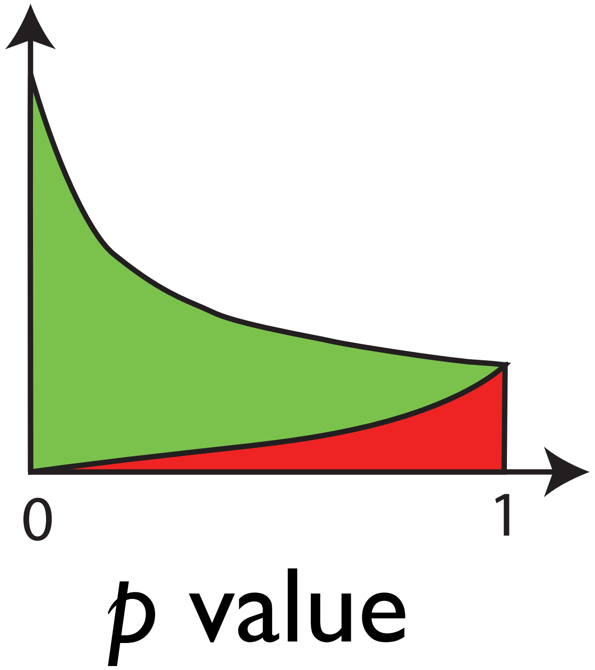
[✓ A]



B



C



🟢 H_1 🔴 H_0

	Called significant	Called not significant	Total
Null true	F	$m_0 - F$	m_0
Alternative true	T	$m_1 - T$	m_1
Total	S	$m - S$	m

idéa [Benjamini and Hochberg 1995] - control for:

$$\frac{\text{no. false positive features}}{\text{no. significant features}} = \frac{F}{F+T} = \frac{F}{S},$$

$$\mathrm{FDR} = \mathrm{E}\left[\frac{F}{F+T}\right] = \mathrm{E}\left[\frac{F}{S}\right].$$

Statistical significance for genomewide studies

John D. Storey*[†] and Robert Tibshirani[‡]

*Department of Biostatistics, University of Washington, Seattle, WA 98195; and [‡]Departments of Health Research and Policy and Statistics, Stanford University, Stanford, CA 94305

Edited by Philip P. Green, University of Washington School of Medicine, Seattle, WA, and approved May 30, 2003 (received for review January 28, 2003)

With the increase in genomewide experiments and the sequencing of multiple genomes, the analysis of large data sets has become commonplace in biology. It is often the case that thousands of features in

to the method in ref. 5 under certain assumptions. Also, ideas similar to FDRs have appeared in the genetics literature (1, 13).

Similarly to the p value, the q value gives each feature its own

We got *m p* values, $p_1, p_2, \ldots, p_m$

for a threshold t we may say that:

$$F(t) = \#\,\{\text{null}\, p_i \leq t; i = 1, \ldots, m\} \text{ and}$$

$$S(t) = \#\,\{p_i \leq t; i = 1, \ldots, m\}.$$

$$\text{FDR}(t) = \text{E}\left[\frac{F(t)}{S(t)}\right].$$

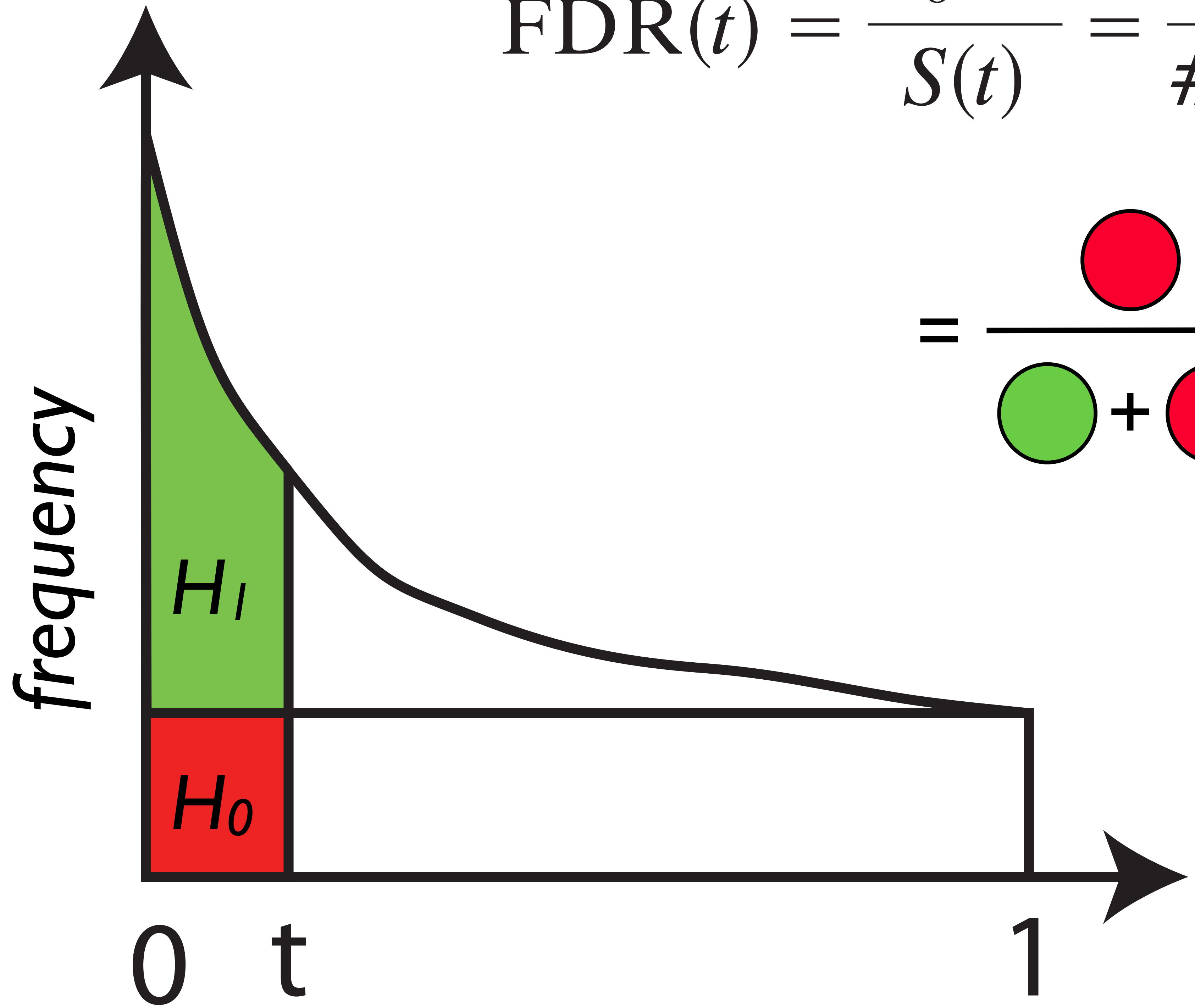
Evenly distributed *p* values: *F(t)*=*m*$_0$*t*=π$_0$*mt*

$$\widehat{\text{FDR}}(t) = \frac{\hat{\pi}_0 m \cdot t}{S(t)} = \frac{\hat{\pi}_0 m \cdot t}{\#\{p_i \leq t\}}.$$

Illustration of $\widehat{\text{FDR}}$

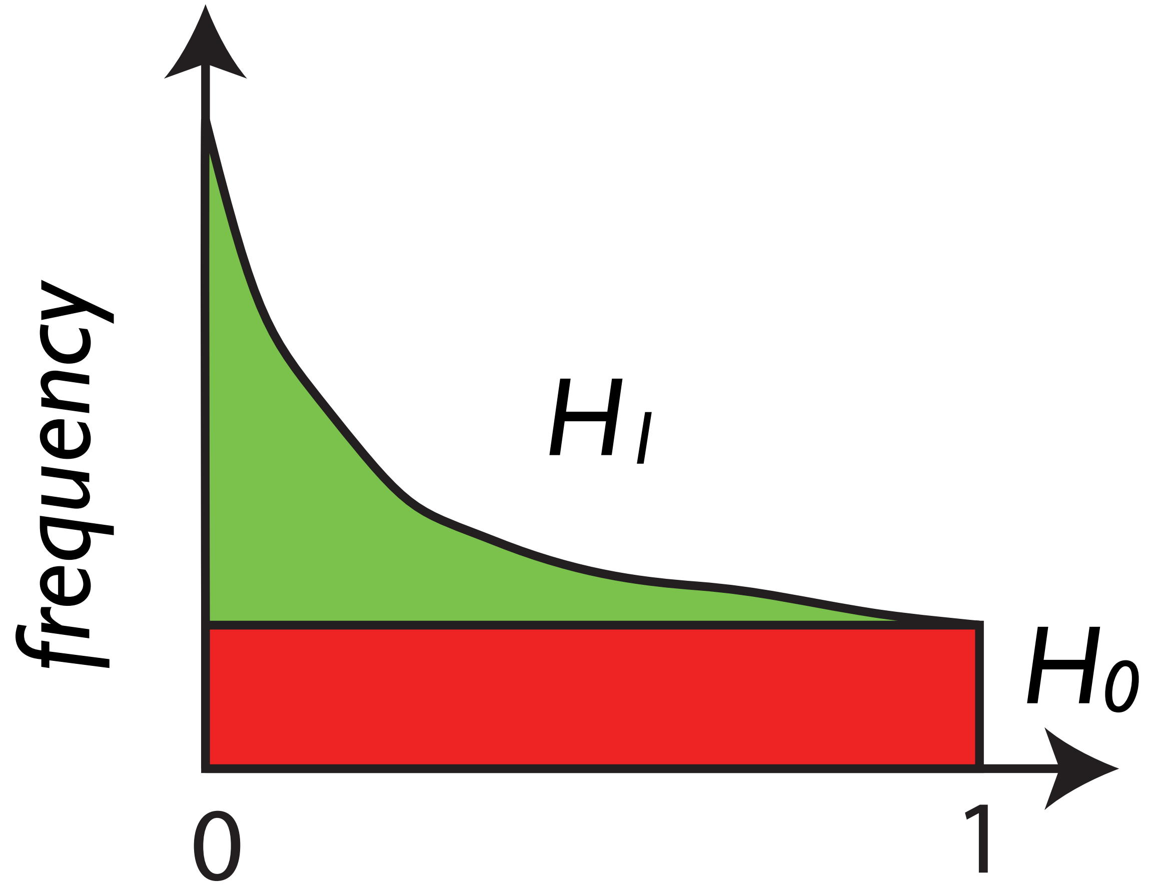
$$\widehat{\text{FDR}}(t) = \frac{\hat{\pi}_0 m \cdot t}{S(t)} = \frac{\hat{\pi}_0 m \cdot t}{\#\{p_i \leq t\}}.$$

= 🔴 / (🟢 + 🔴)



π_0

π_0 is the prior probability that a statistic is derived under *H_0 i.e.* Pr(H=*H_0*)

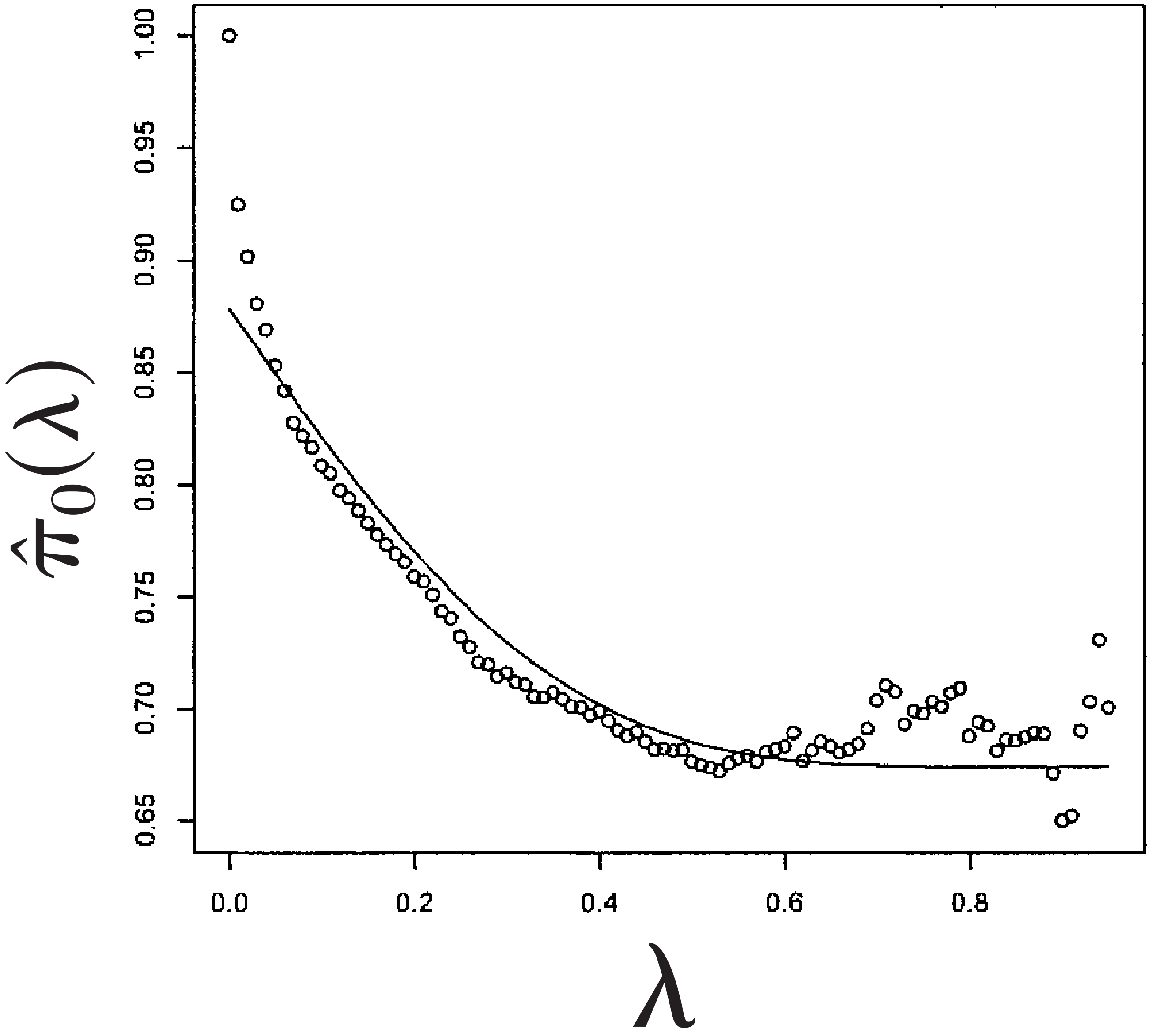


π_0 = 🔴 / (🟢 + 🔴)

π_0 estimation

Investigate the higher (close to 1) *p* values

$$\hat{\pi}_0(\lambda) = \frac{\#\,\{p_i > \lambda; i = 1, \ldots, m\}}{m(1 - \lambda)},$$



q value

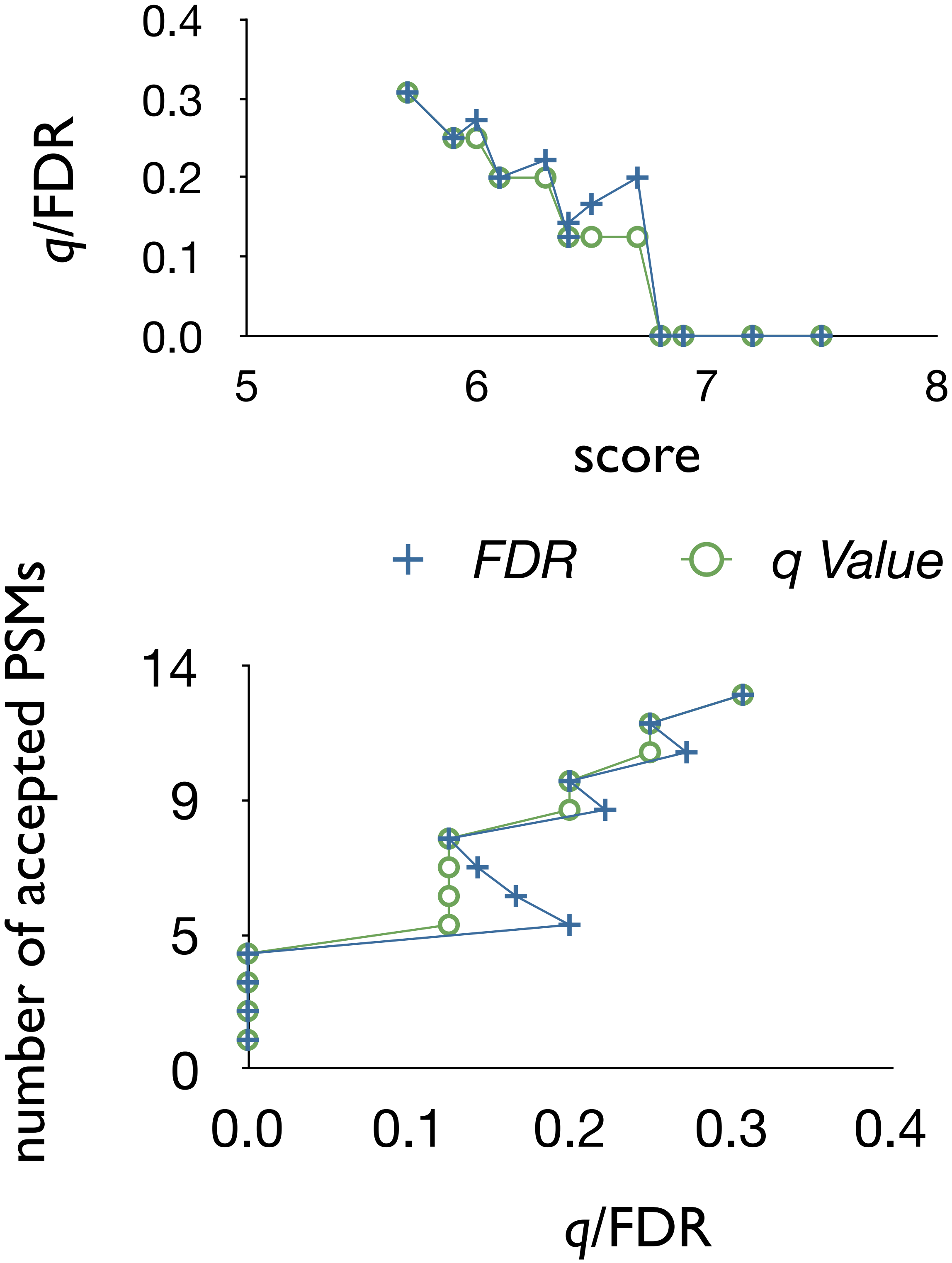
To assign relevant measures to individual identifications and to ensure a monotonically increasing function with the threshold, the *q* value is defined as

$$\hat{q}(p_i) = \min_{t \geq p_i} \widehat{\mathrm{FDR}}(t).$$

q value

$$q(x)=\min_{x\geq x'}\{FDR(x')\}$$

score	type
7.5	correct
7.2	correct
6.9	correct
6.8	correct
6.7	incorrect
6.5	correct
6.4	correct
6.4	correct
6.3	incorrect
6.1	correct
6	incorrect
5.9	correct
5.7	incorrect
...	...



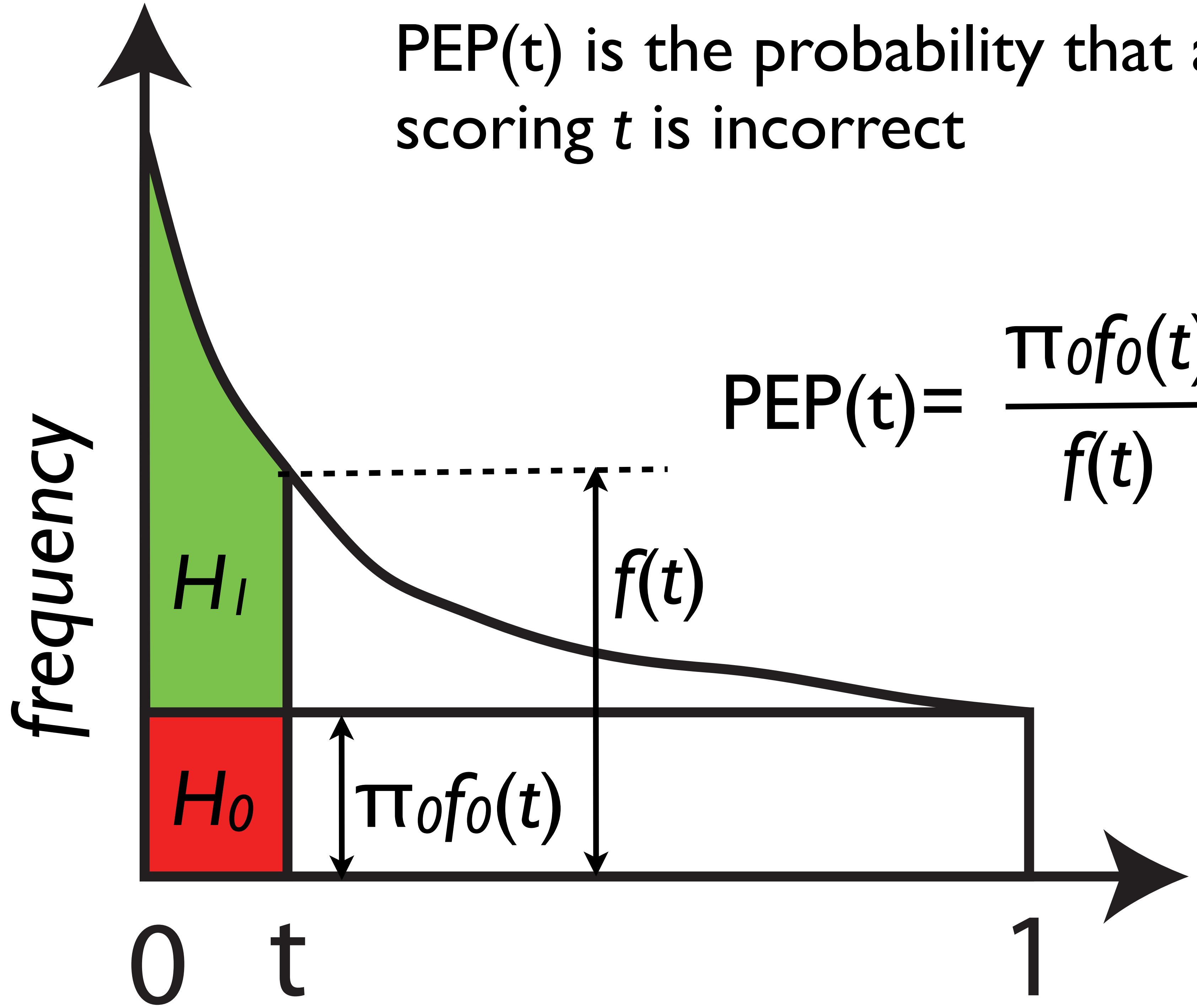
FDRs from empirical null models

- If we have an empirical null model, i.e. a mechanism z(y) that models readouts under the null model a *p* value can be estimated as $p(t)=\#\{z(y^i)\geq t\}/(m+1)$

Posterior Error Probability a.k.a. local FDR

PEP(t) is the probability that an identification scoring *t* is incorrect

$$\mathrm{PEP(t)} = \frac{\pi_0 f_0(t)}{f(t)} = \frac{\pi_0 f_0(t)}{\pi_0 f_0(t) + \pi_1 f_1(t)}$$



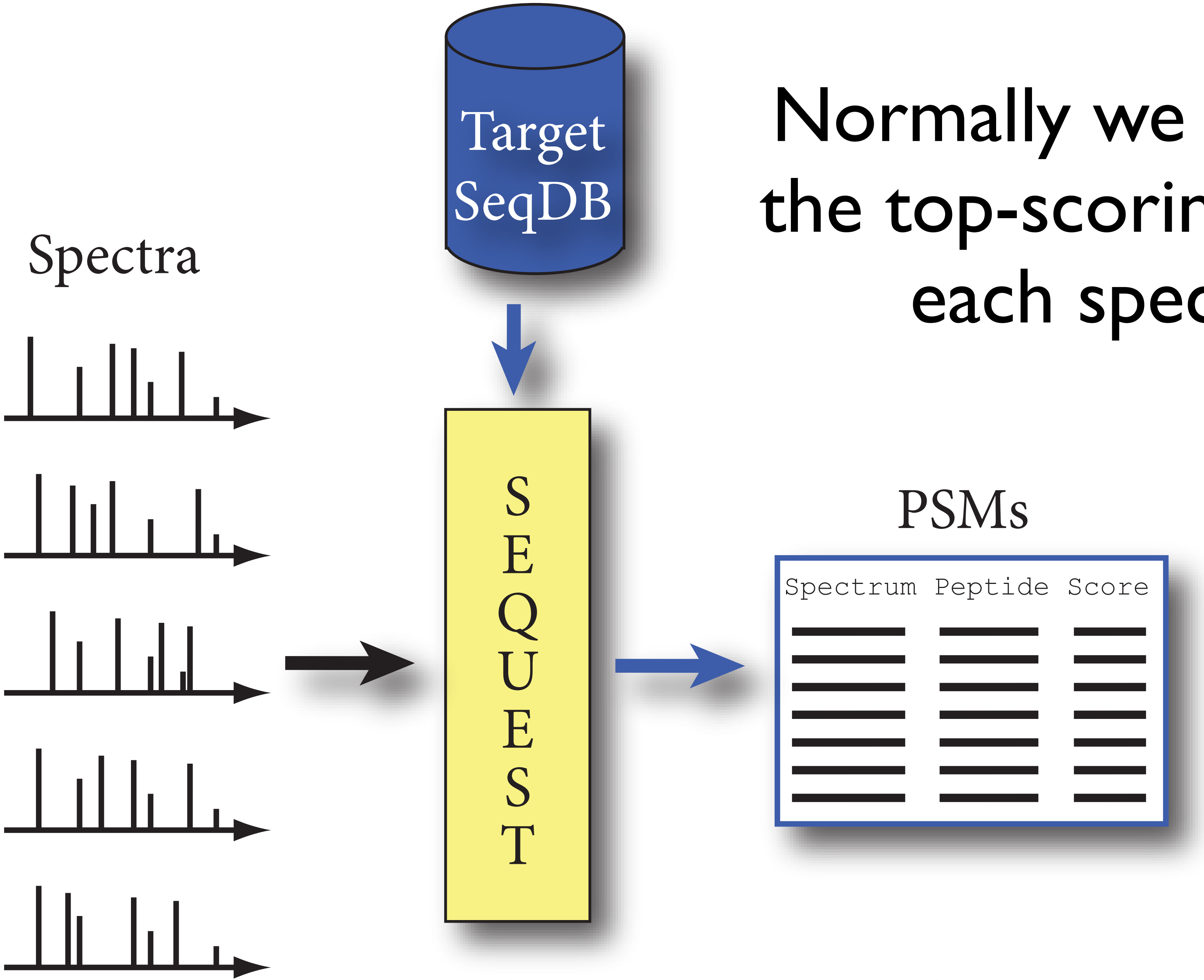
Some popular confidence metrics

- False Discovery Rate - *FDR*(*x*) is the expectation value of the fraction of identifications with score above threshold *x* that are incorrect
- q value - *q*(*x*) is the minimal *FDR*(*x*') out of all thresholds *x*' that includes *x*
- Posterior Error Probability - *PEP*(*x*) is the probability that an identification with score *x* is incorrect
- p value - *p*(*x*) is the probability that an incorrect identification gets a score higher than or as high as *x*

Outline

1. What is proteomics?
2. Background on Mass spectrometry
3. Peptide identification in shotgun proteomics
4. Multiple hypothesis corrections
5. The statistics of shotgun proteomics
6. Some open problems

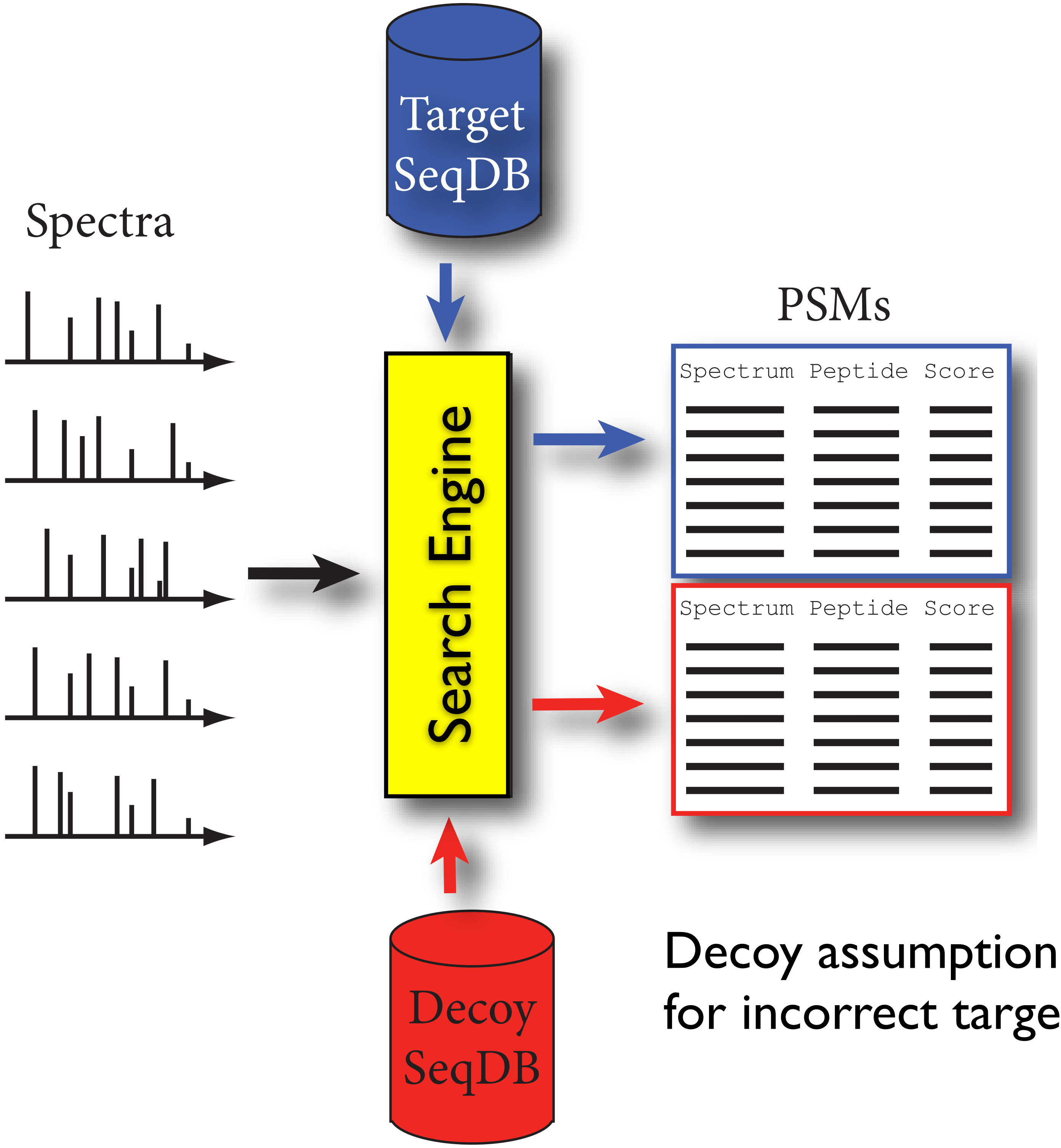
Peptide identification



Normally we keep only the top-scoring PSM for each spectrum

PSM - Peptide Spectrum Match

We can use target-decoy analysis to calculate *q* values



Spectra

Target SeqDB

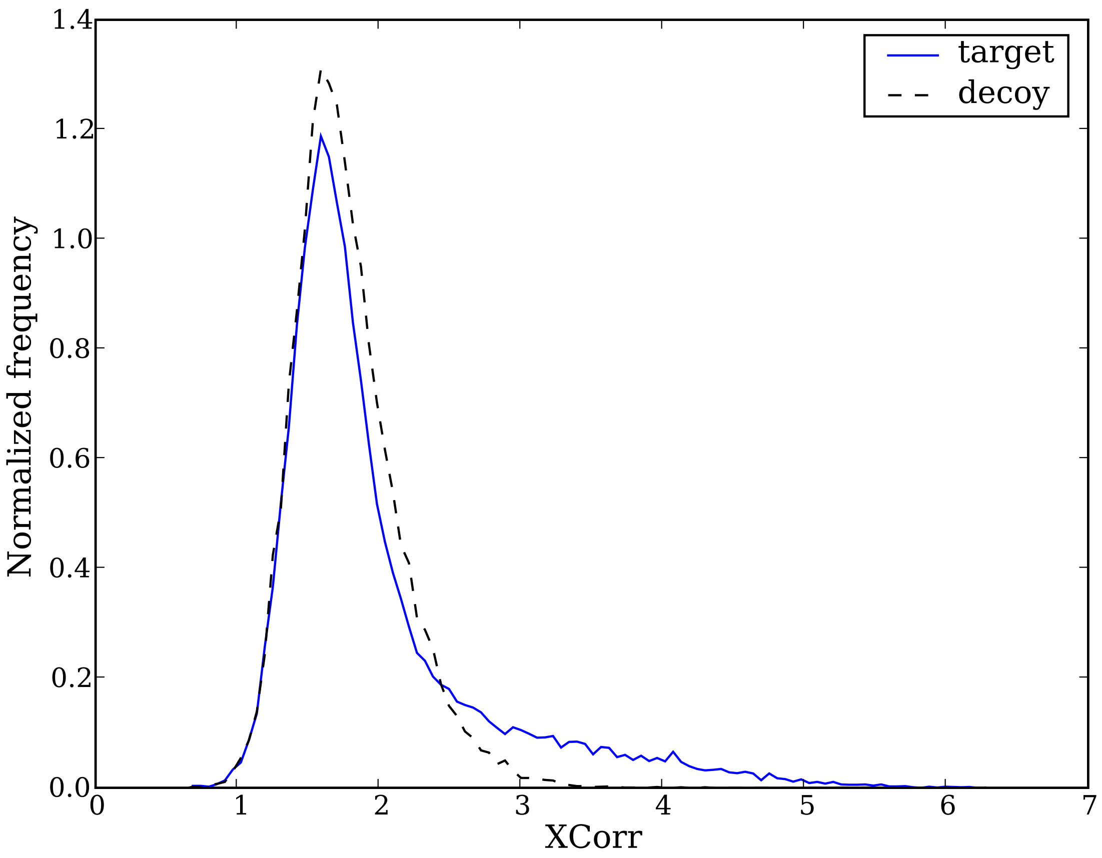
Search Engine

Decoy SeqDB

PSMs

Spectrum	Peptide	Score

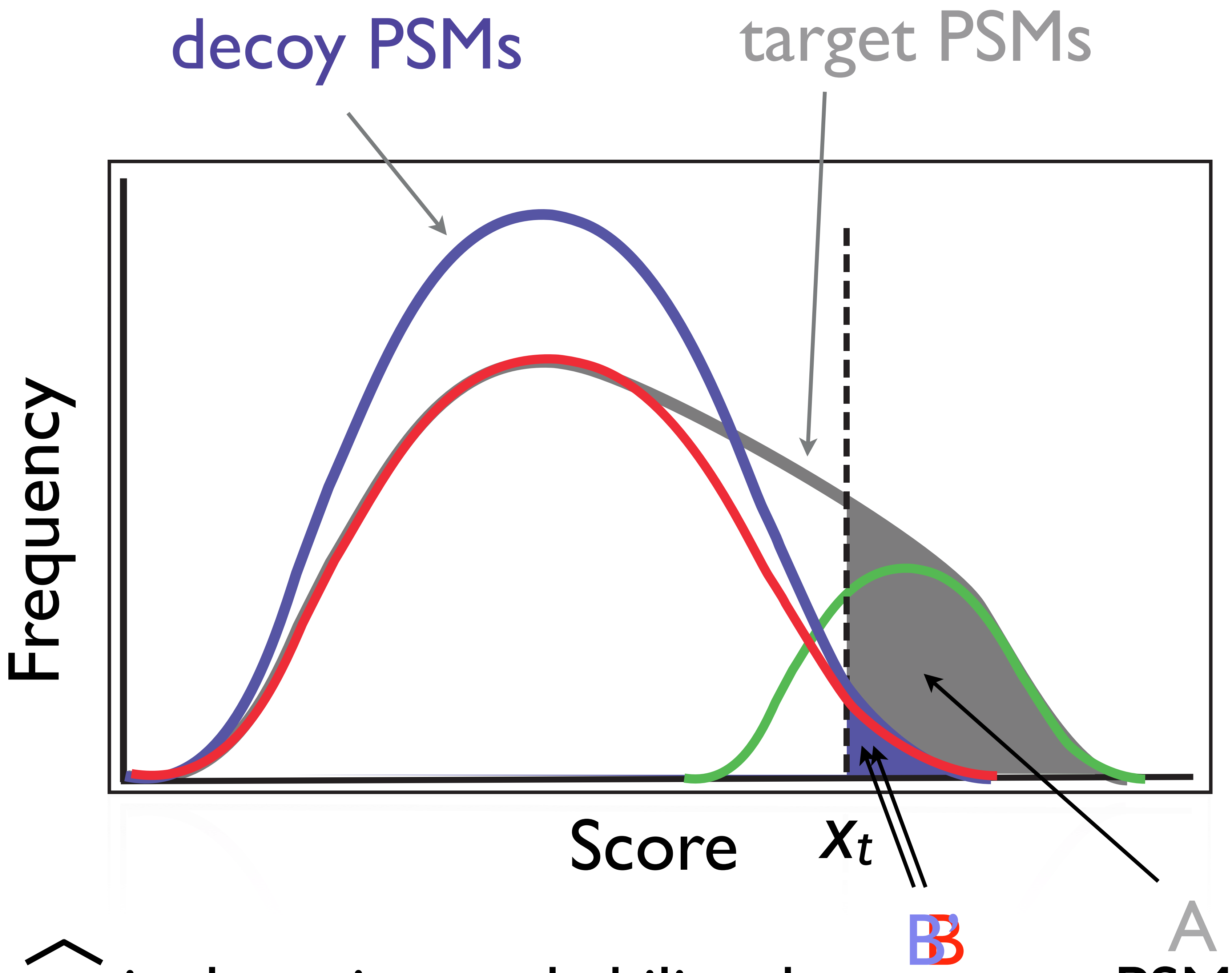
Spectrum	Peptide	Score



Decoy assumption: decoy identifications are good proxies for incorrect target identifications

[Moore et al. JASMS 2002]

Using decoy PSMs to estimate false discovery rate



$$\mathrm{FDR}(x_t) = \frac{\Pr(x \geq x_t, H=0)}{\Pr(x \geq x_t)}$$

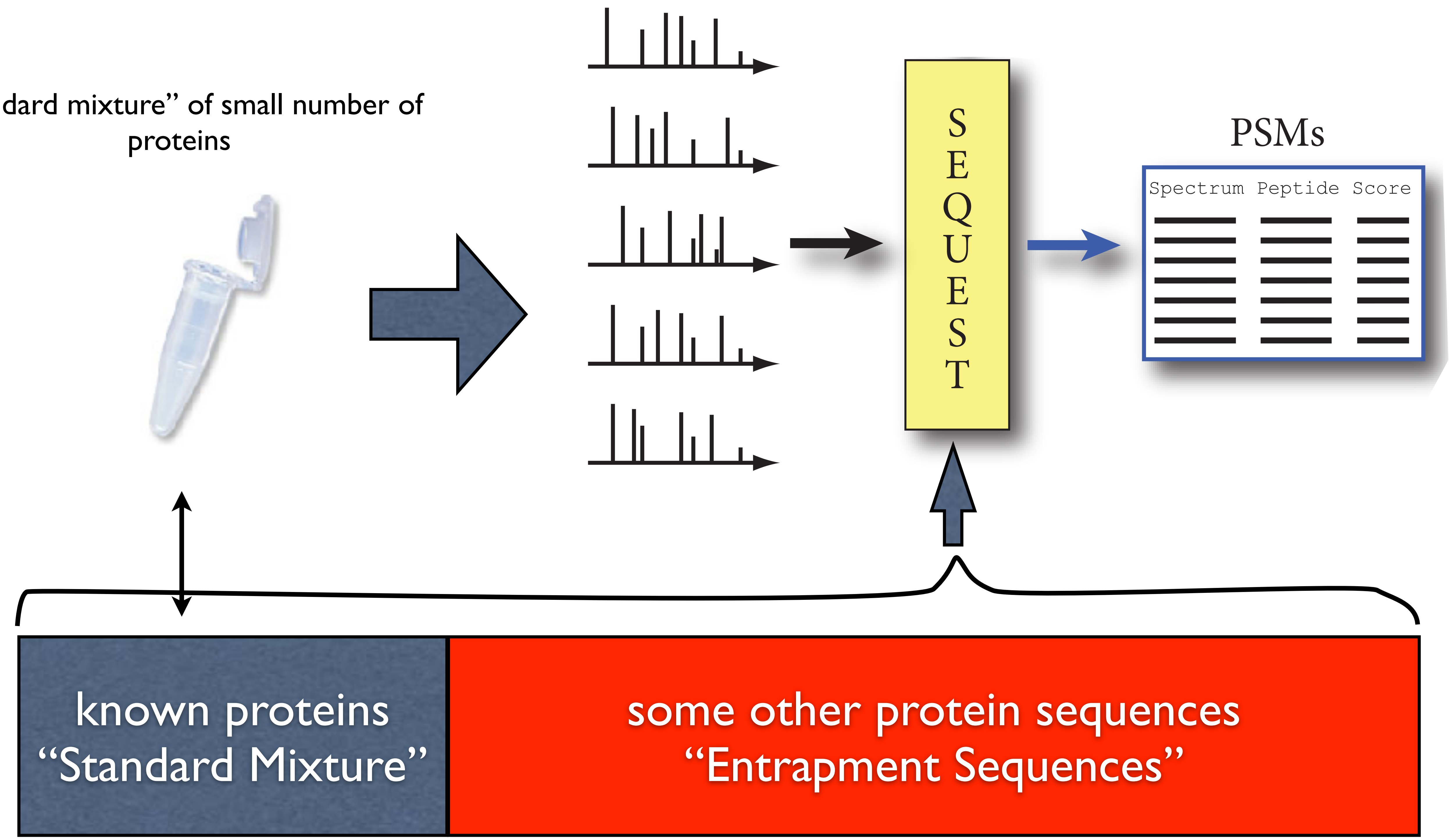
$$\widehat{\mathrm{FDR}} = \frac{B}{A} = \frac{\hat{\pi}_0 \cdot B'}{A}$$

$$\widehat{q}(x_t) = \inf_{x \leq x_t}\{\widehat{\mathrm{FDR}}(x)\}$$

[Käll *et al.* JPR 2008]

$\hat{\pi}_0$ is the prior probability that a target PSM is incorrectly matched

Known Sample



“Standard mixture” of small number of proteins

Spectra

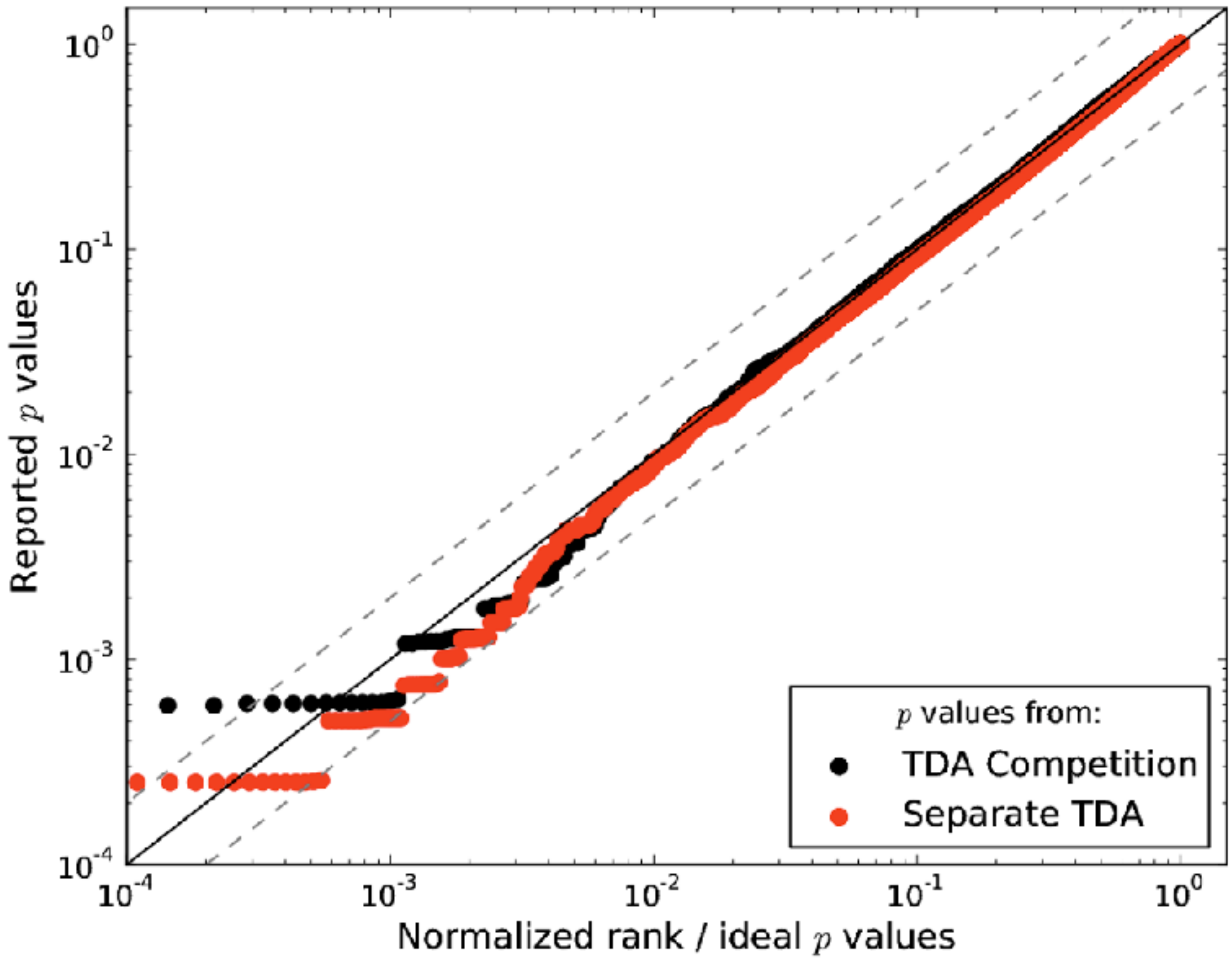
SEQUEST

PSMs

known proteins
“Standard Mixture”

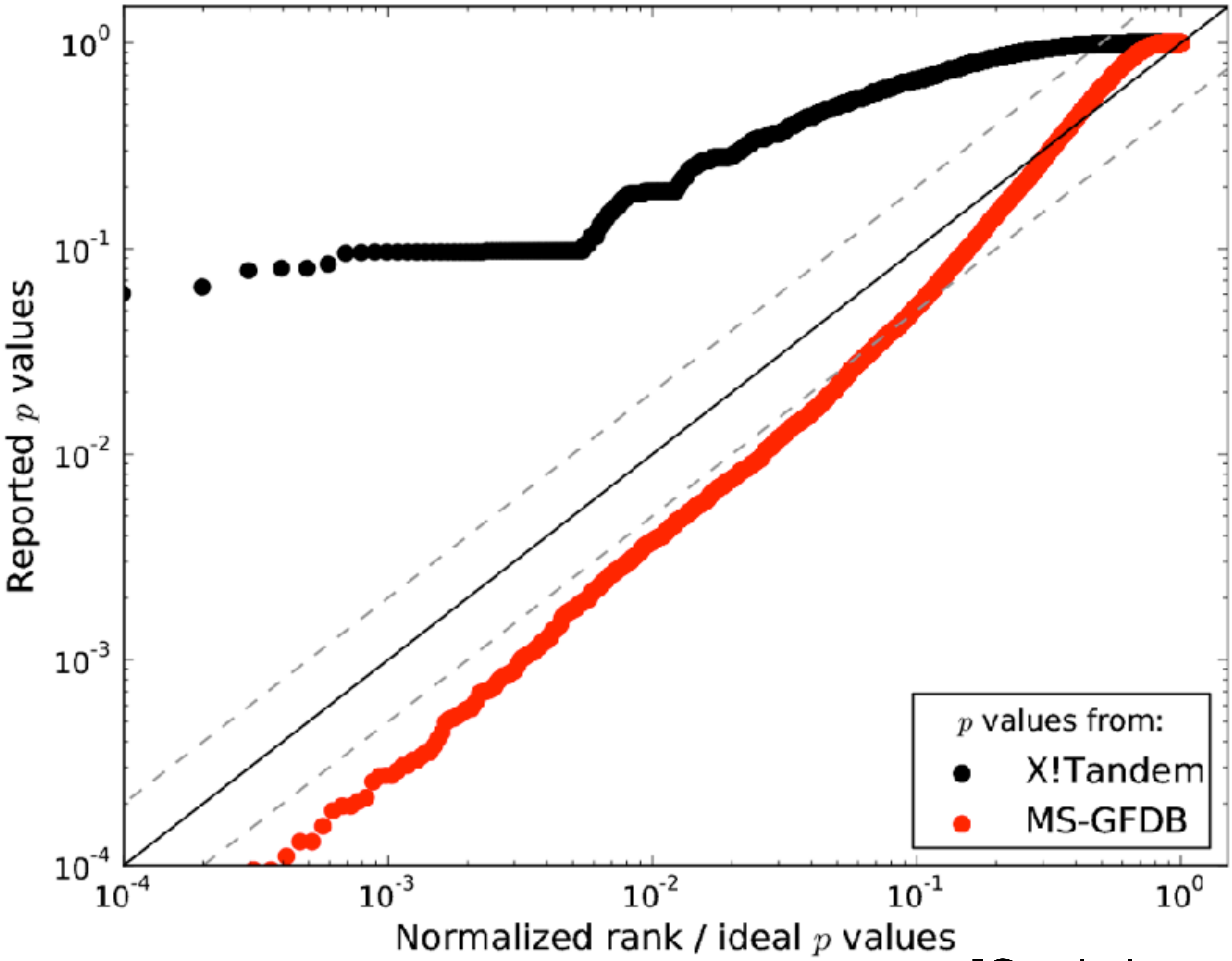
some other protein sequences
“Entrapment Sequences”

Calibration: Quantile-quantile plots



[Granholm *et al. JPR* 2011]

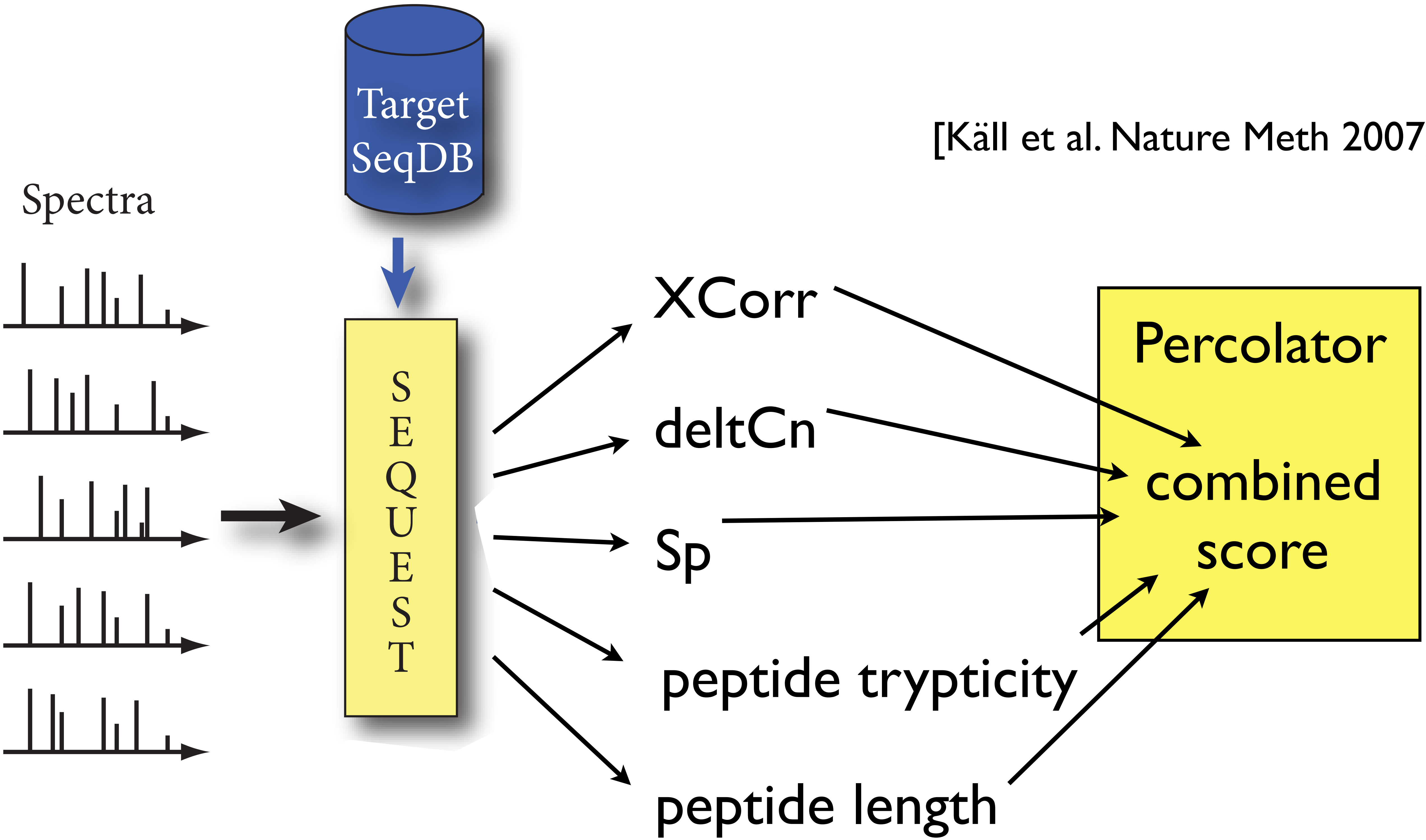
Conservative (black) and anti-conservative (red) scores



[Granholm *et al. JPR* 2011]

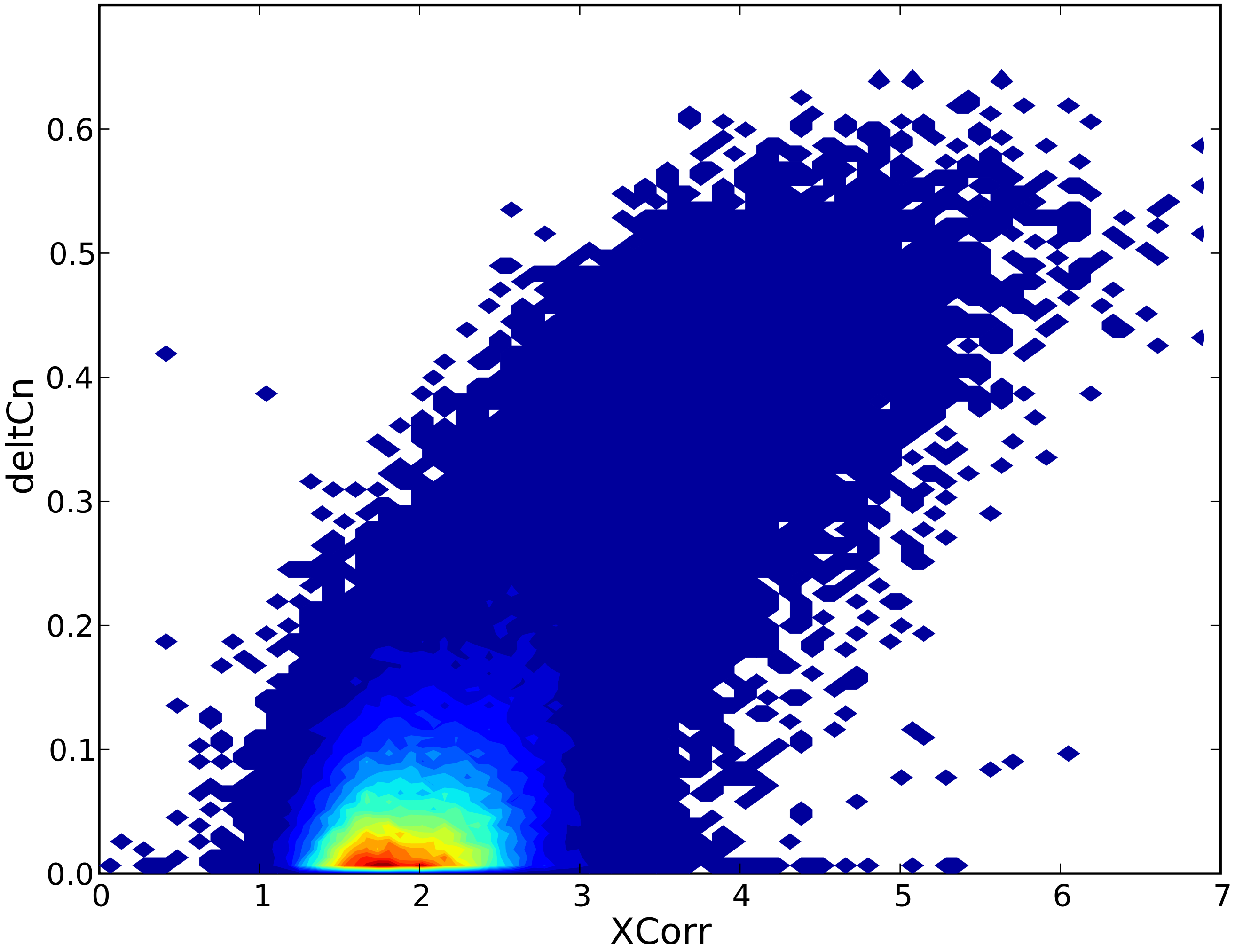
Percolator combines different PSM features in an optimal manner

[Käll et al. Nature Meth 2007]



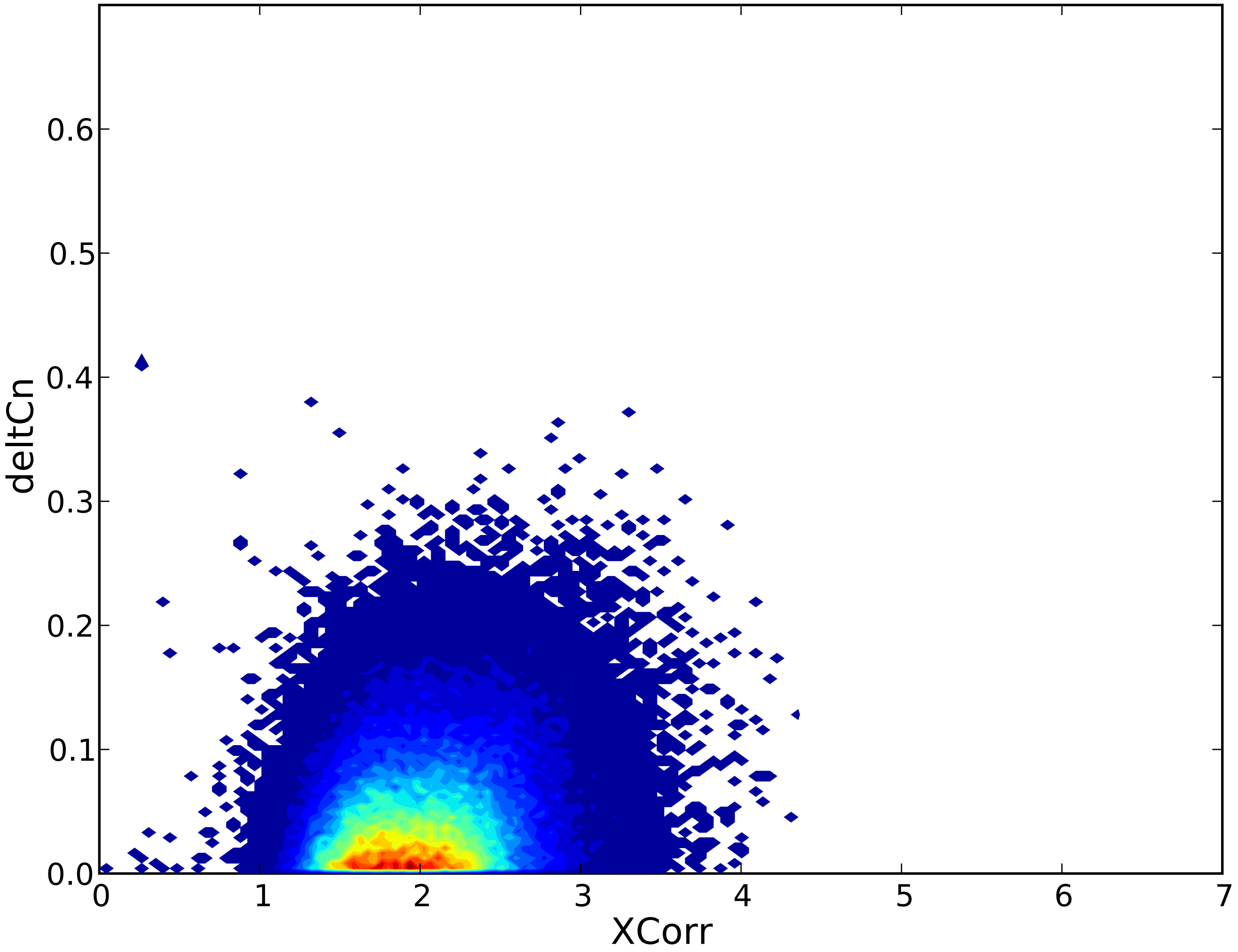
Target PSMs consist of a mixture of correct and incorrect PSMs and are hence not always good examples of correct PSMs

Target



Label +1

Decoy



Label -1

Machine learning strategies

Set of Target PSMs contain mostly null PSMs.

Possible workarounds:

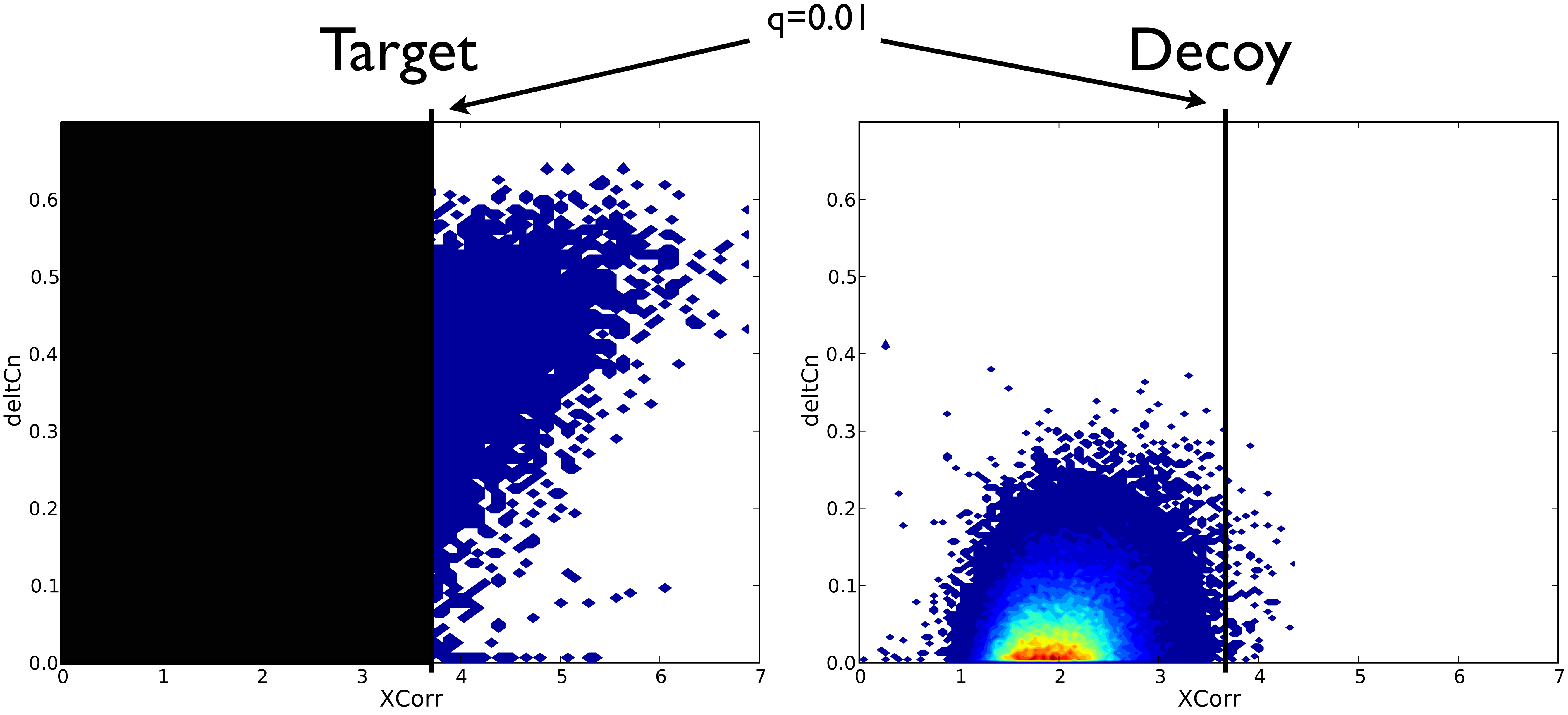
1. Curate a set of known correct PSMs

   Anderson *et al.* (2003), Keller *et al.* (2002) [PeptideProphet]

2. Better algorithms:

   -Semi-supervised learning

Self-training



q=0.01

Target

Decoy

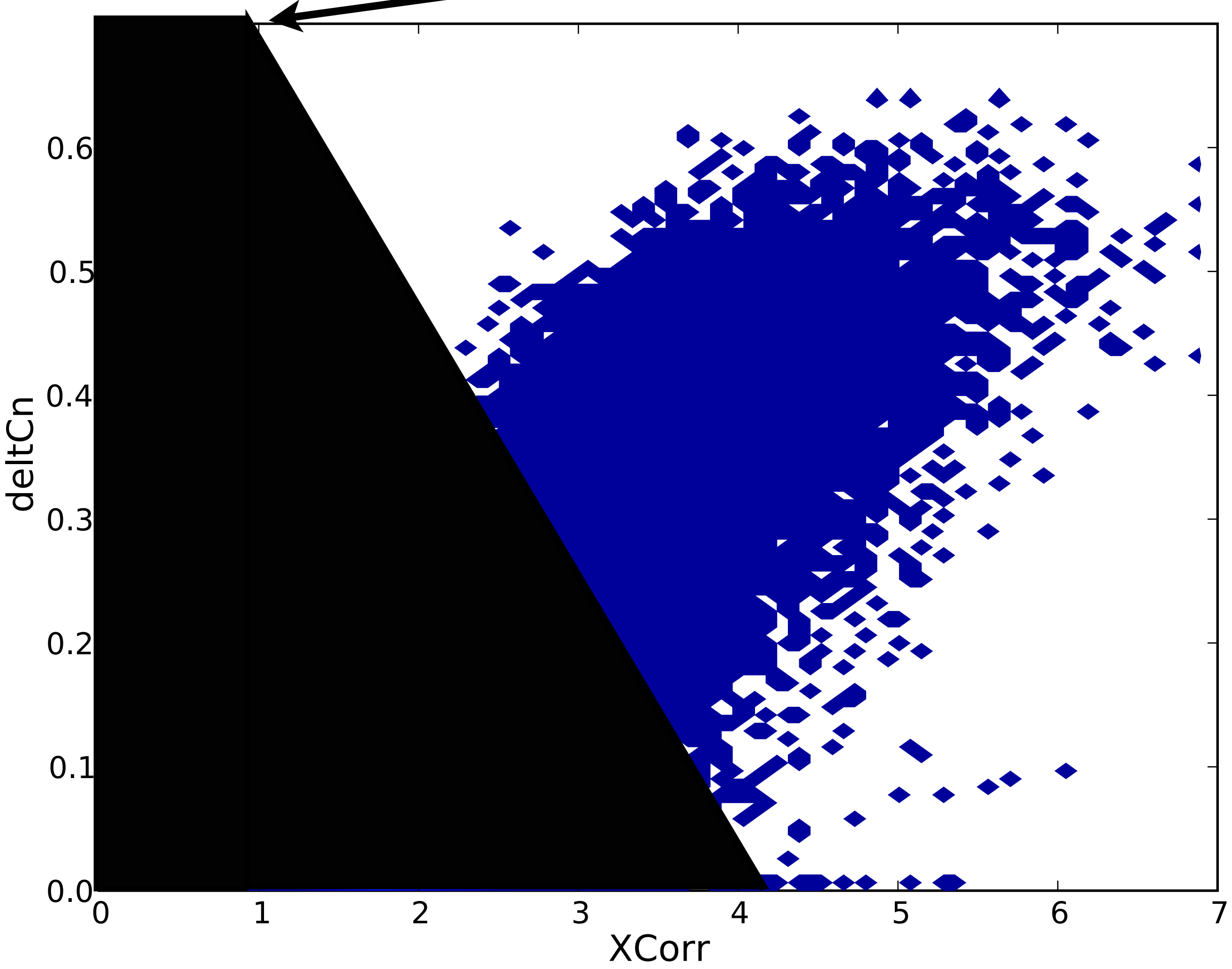
Label 1

Label -1

Self-training

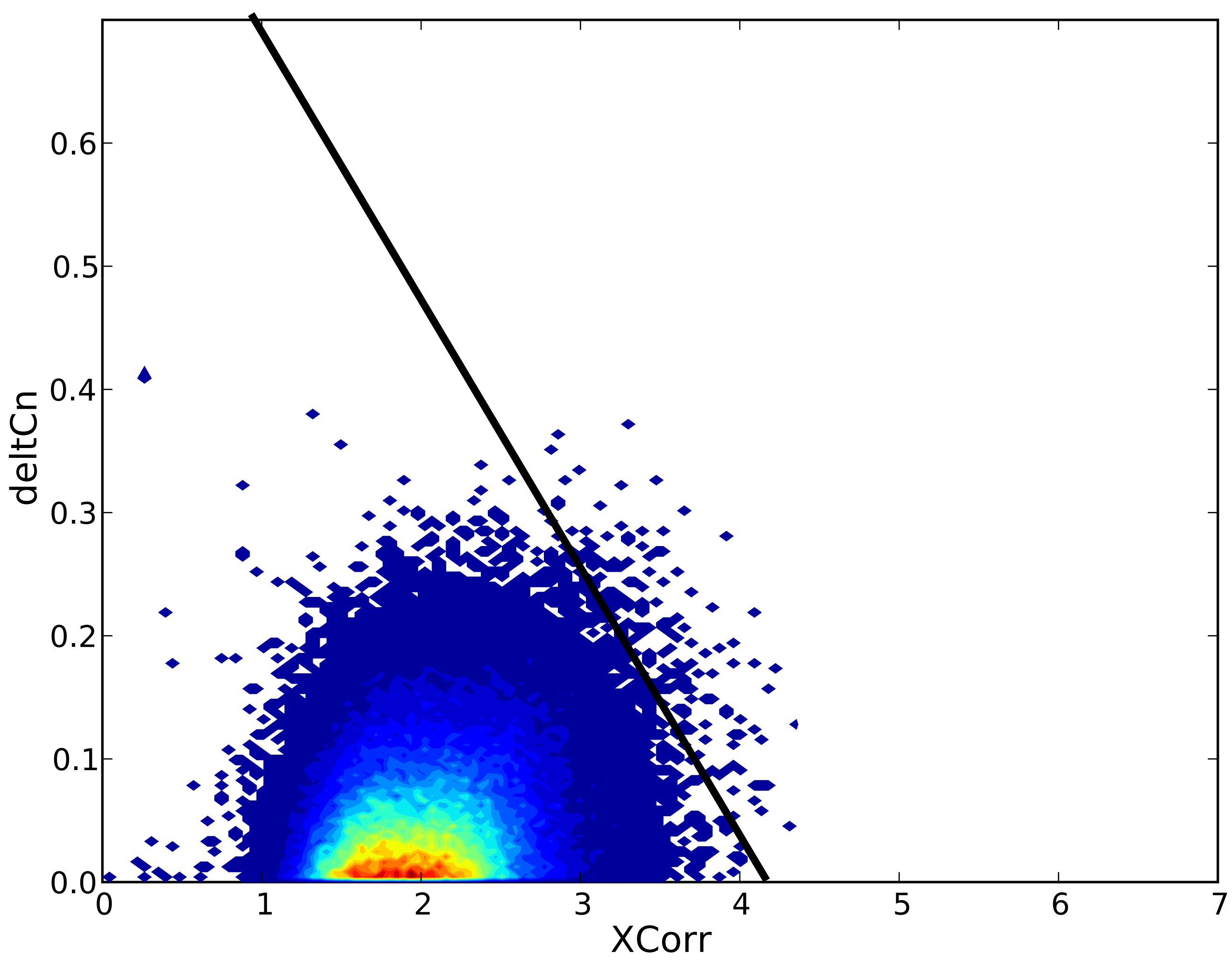
q=0.01

Target



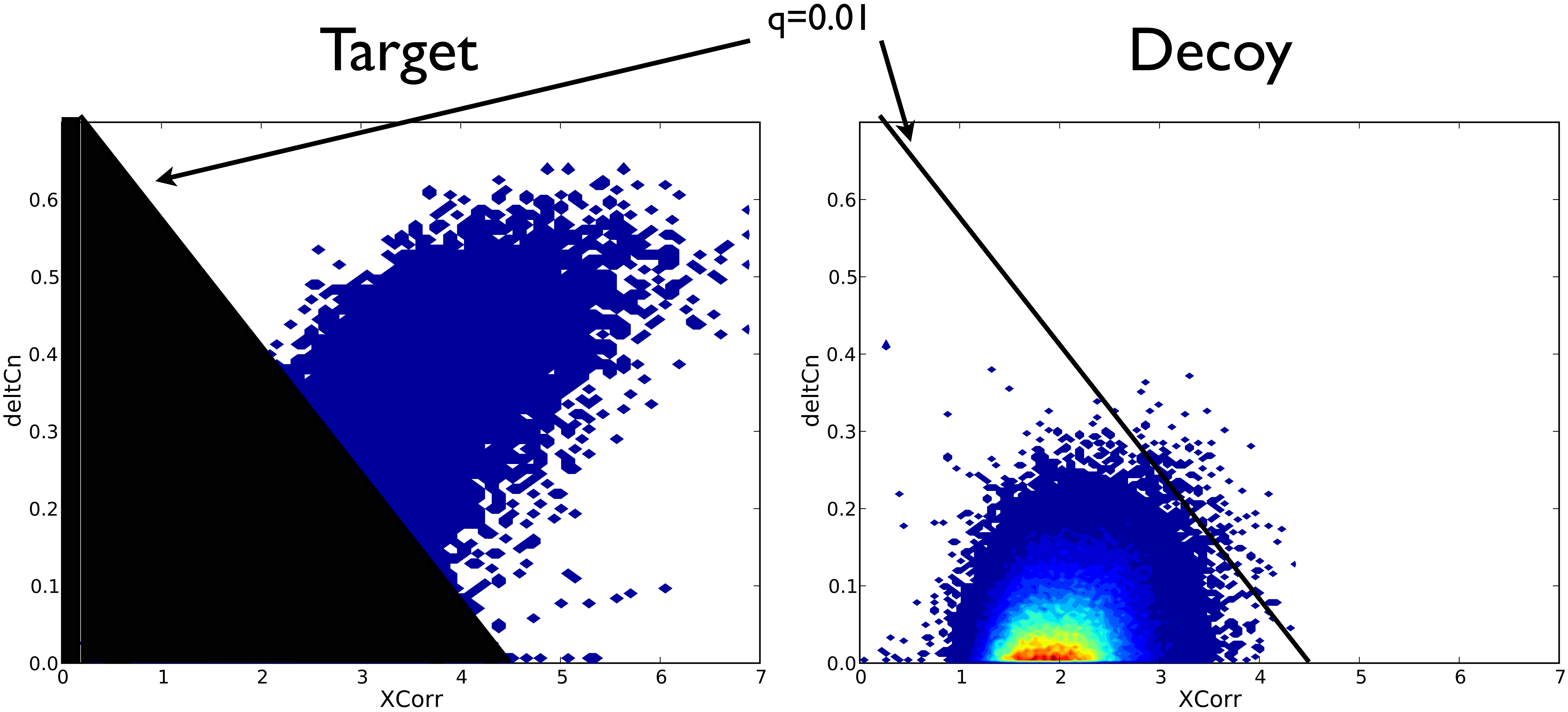
Label 1

Decoy



Label -1

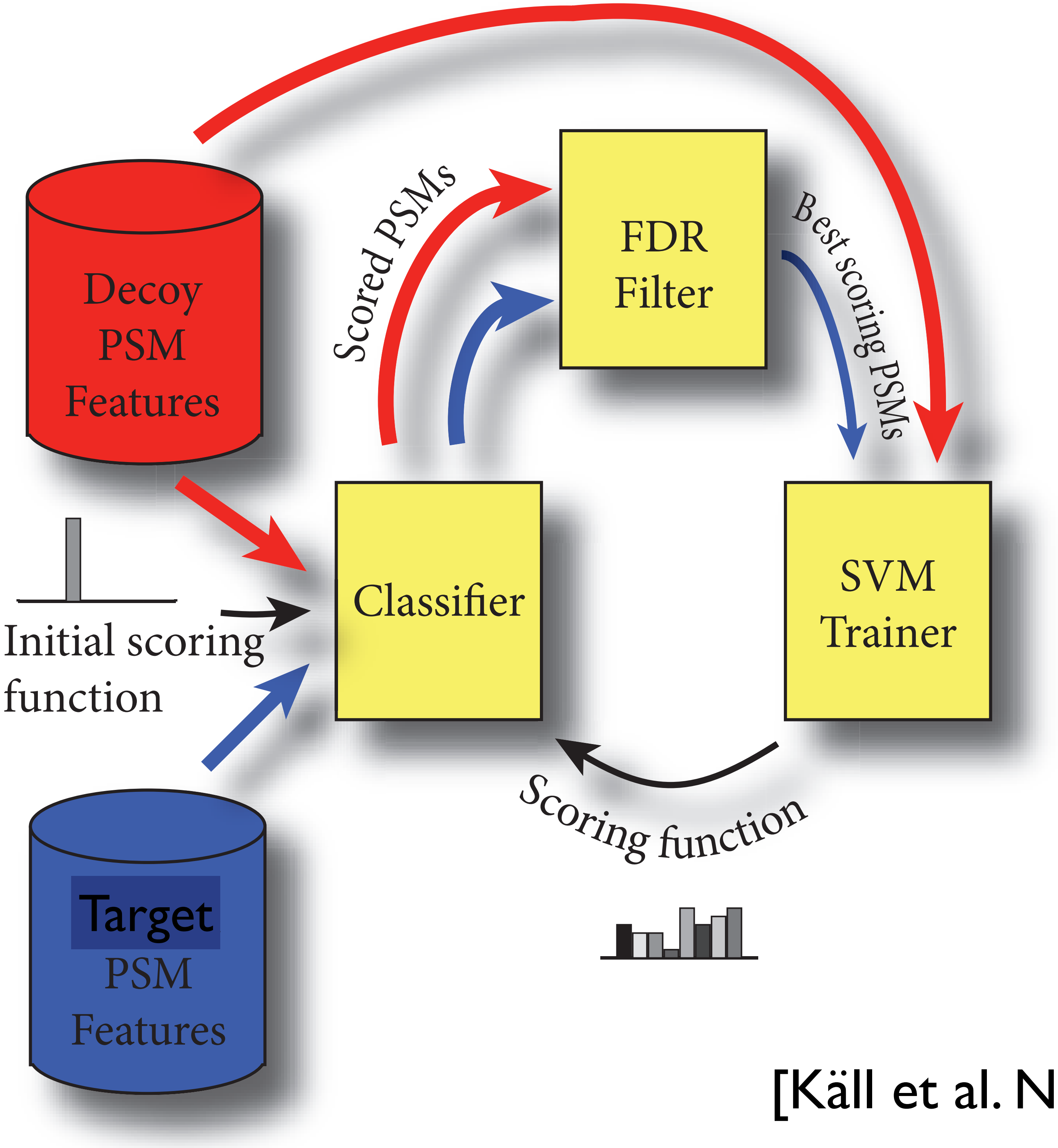
Self-training



Label 1

Label -1

Percolator algorithm

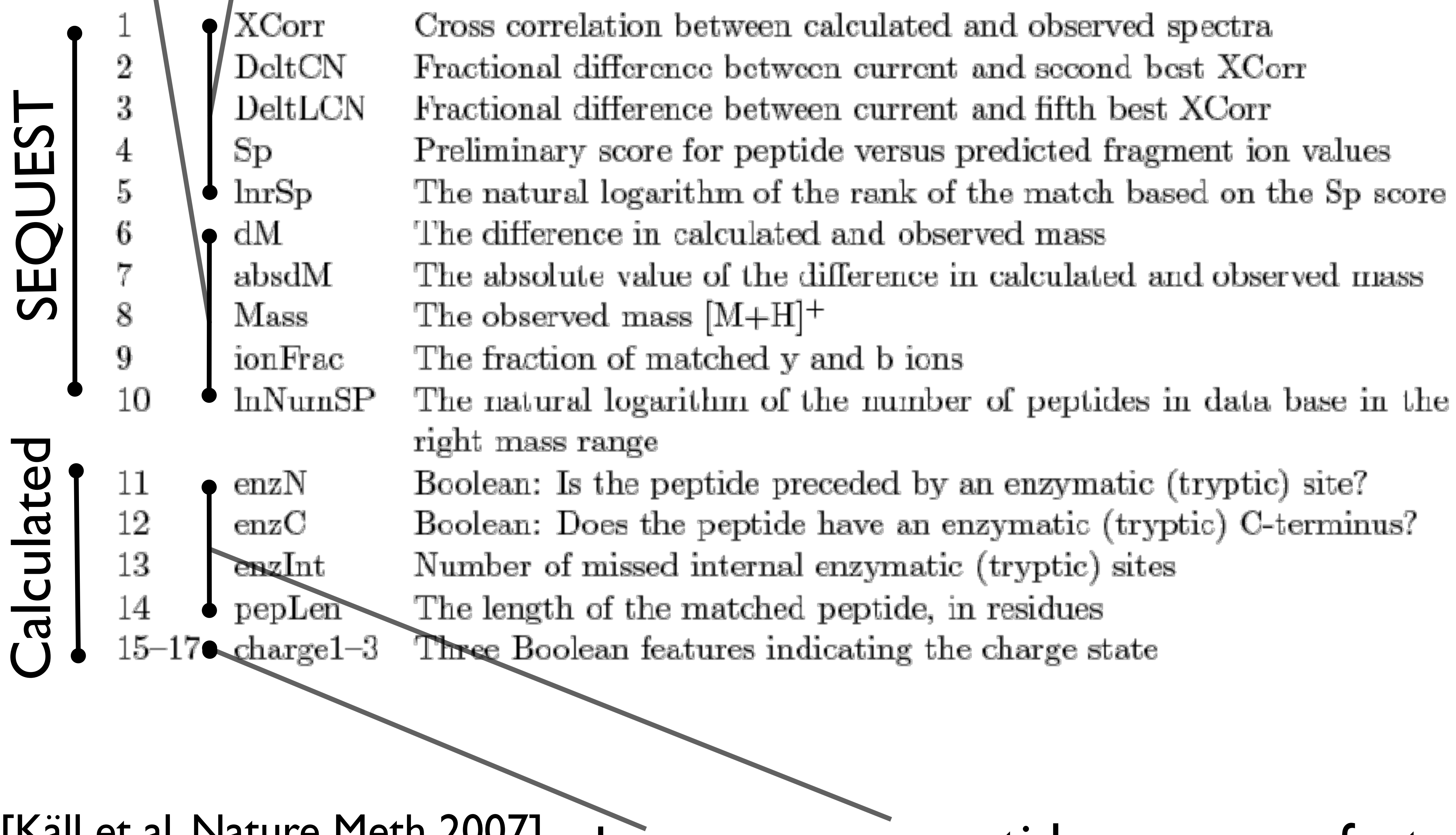


[Käll et al. Nature Meth 2007]

precursor mass features

PSM features

scores



1	XCorr	Cross correlation between calculated and observed spectra
2	DeltCN	Fractional difference between current and second best XCorr
3	DeltLCN	Fractional difference between current and fifth best XCorr
4	Sp	Preliminary score for peptide versus predicted fragment ion values
5	lnrSp	The natural logarithm of the rank of the match based on the Sp score
6	dM	The difference in calculated and observed mass
7	absdM	The absolute value of the difference in calculated and observed mass
8	Mass	The observed mass $[M+H]^+$
9	ionFrac	The fraction of matched y and b ions
10	lnNumSP	The natural logarithm of the number of peptides in data base in the right mass range
11	enzN	Boolean: Is the peptide preceded by an enzymatic (tryptic) site?
12	enzC	Boolean: Does the peptide have an enzymatic (tryptic) C-terminus?
13	enzInt	Number of missed internal enzymatic (tryptic) sites
14	pepLen	The length of the matched peptide, in residues
15–17	charge1–3	Three Boolean features indicating the charge state

SEQUEST

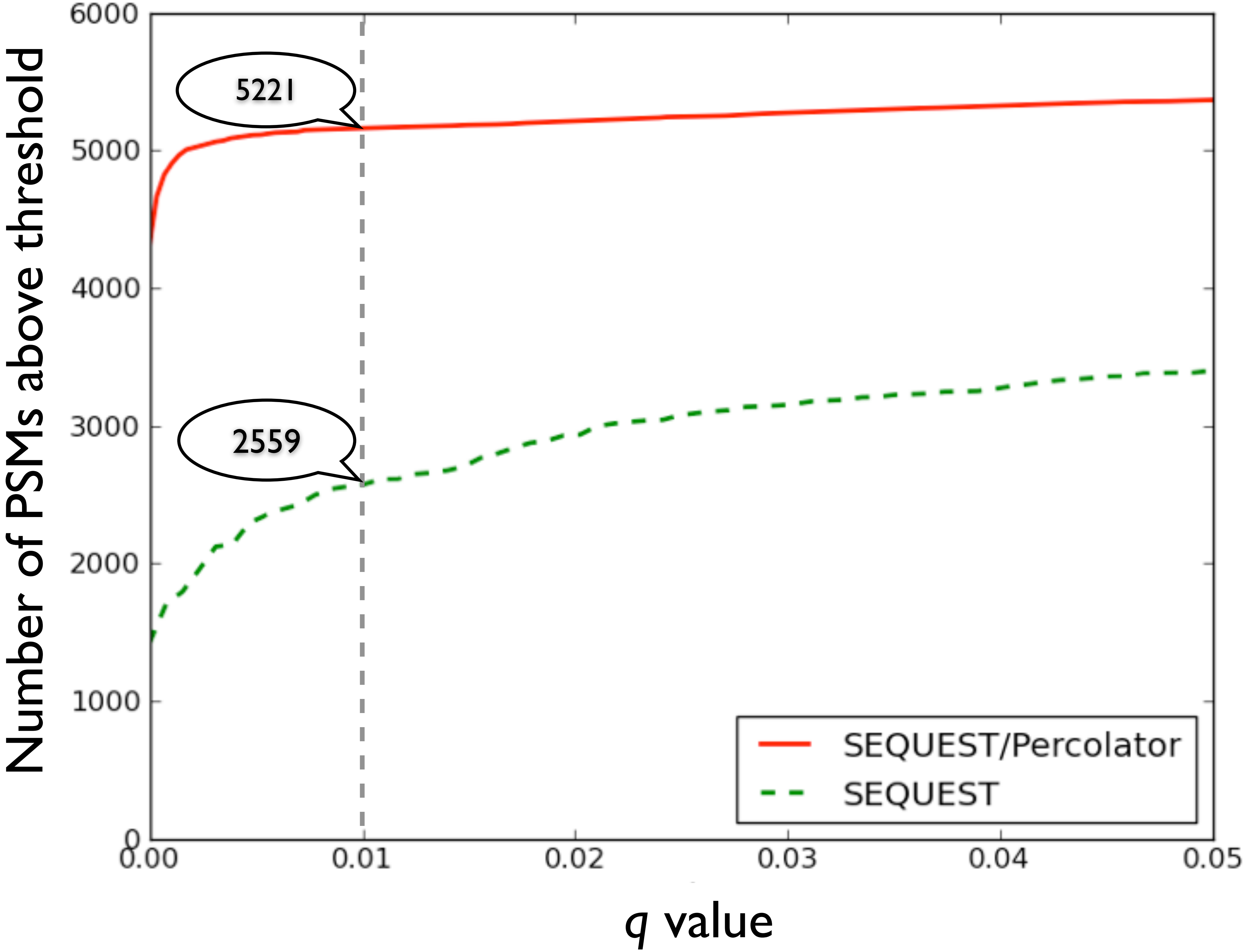
Calculated

[Käll et al. Nature Meth 2007]

charge

peptide sequence features

Percolator greatly increase the yield from Sequest matching results

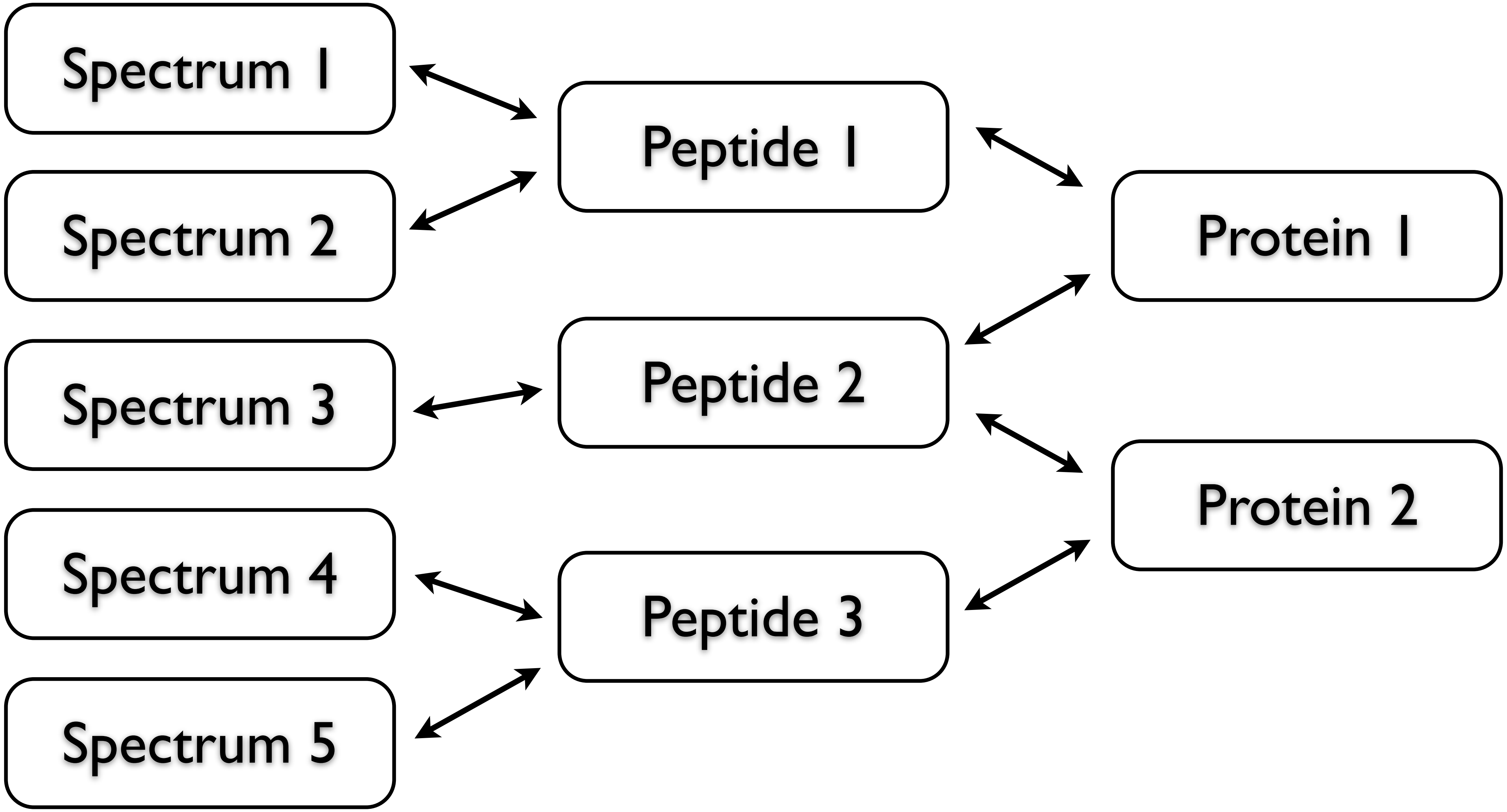


7829 CID-PSMs, from trypsinized HEK293 cells [Kim *et al.* MCP 2010]

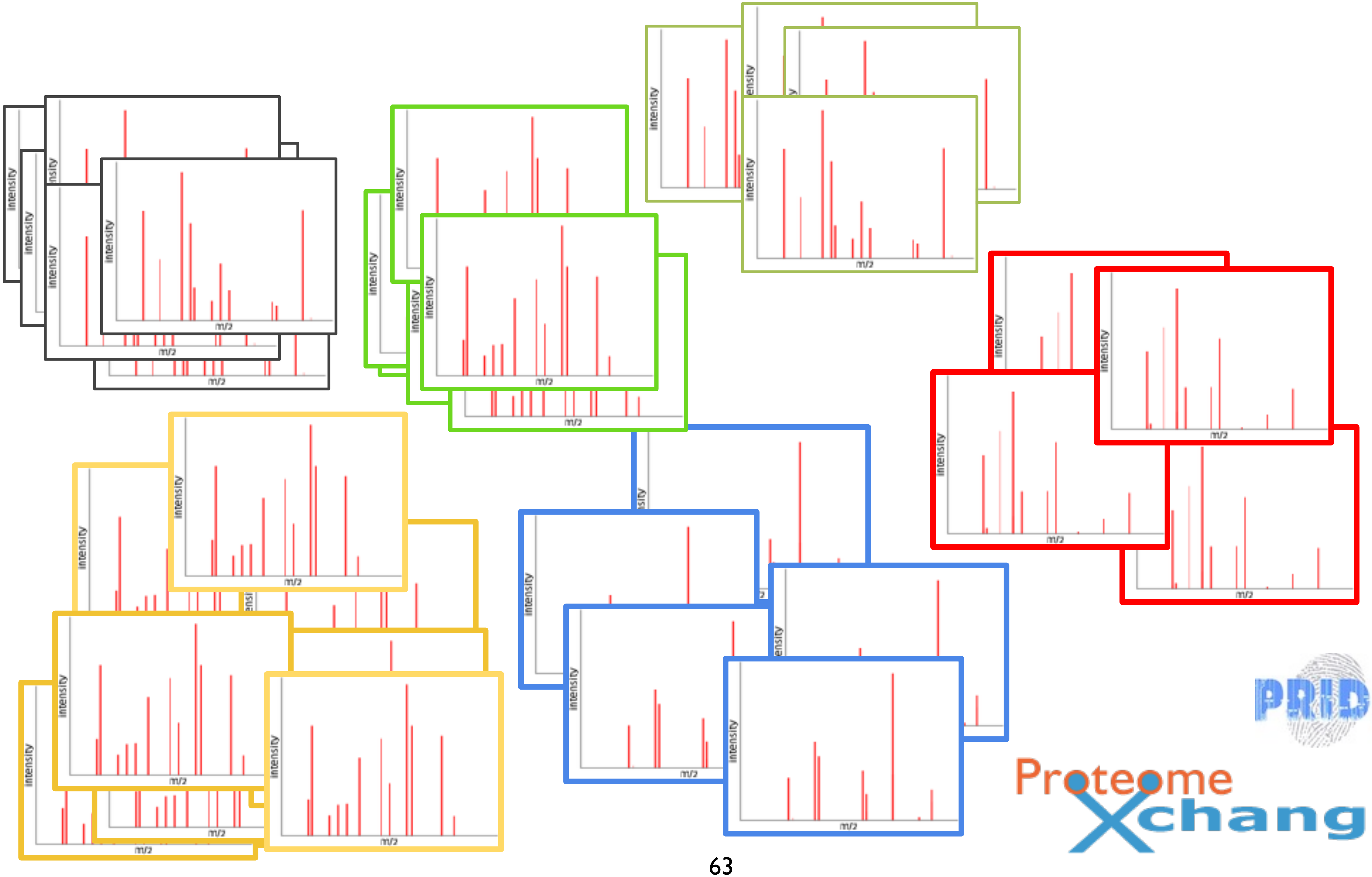
Outline

1. What is proteomics?
2. Background on Mass spectrometry
3. Peptide identification in shotgun proteomics
4. Multiple hypothesis corrections
5. The statistics of shotgun proteomics
6. Some open problems

PSM/Peptide/Protein level statistics



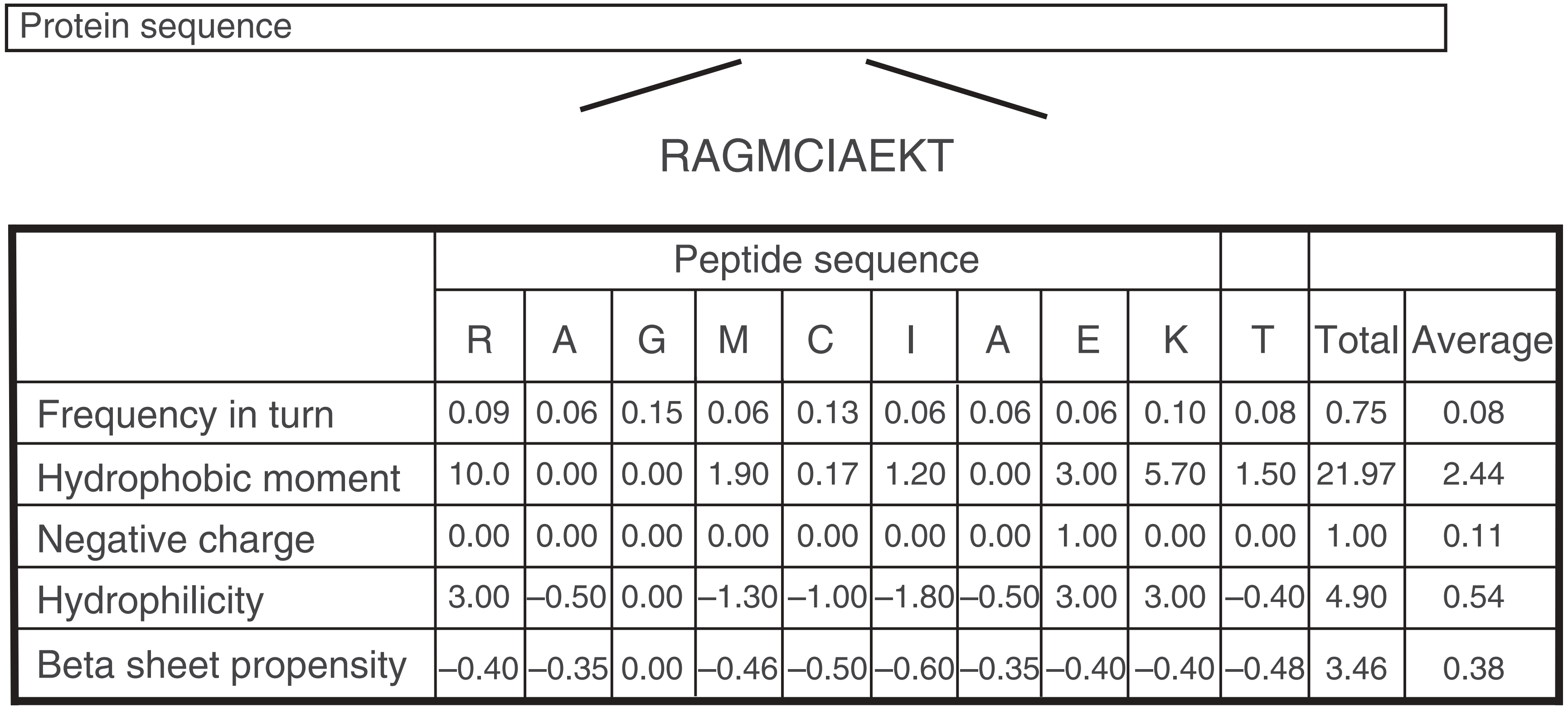
Clustering of Fragment Spectra



[The & Käll, JPR in press]

Proteotypic peptide prediction

- Some peptides are more prone to be detected than other peptides. We may predict such "proteotypic" peptides using classical machine learning.



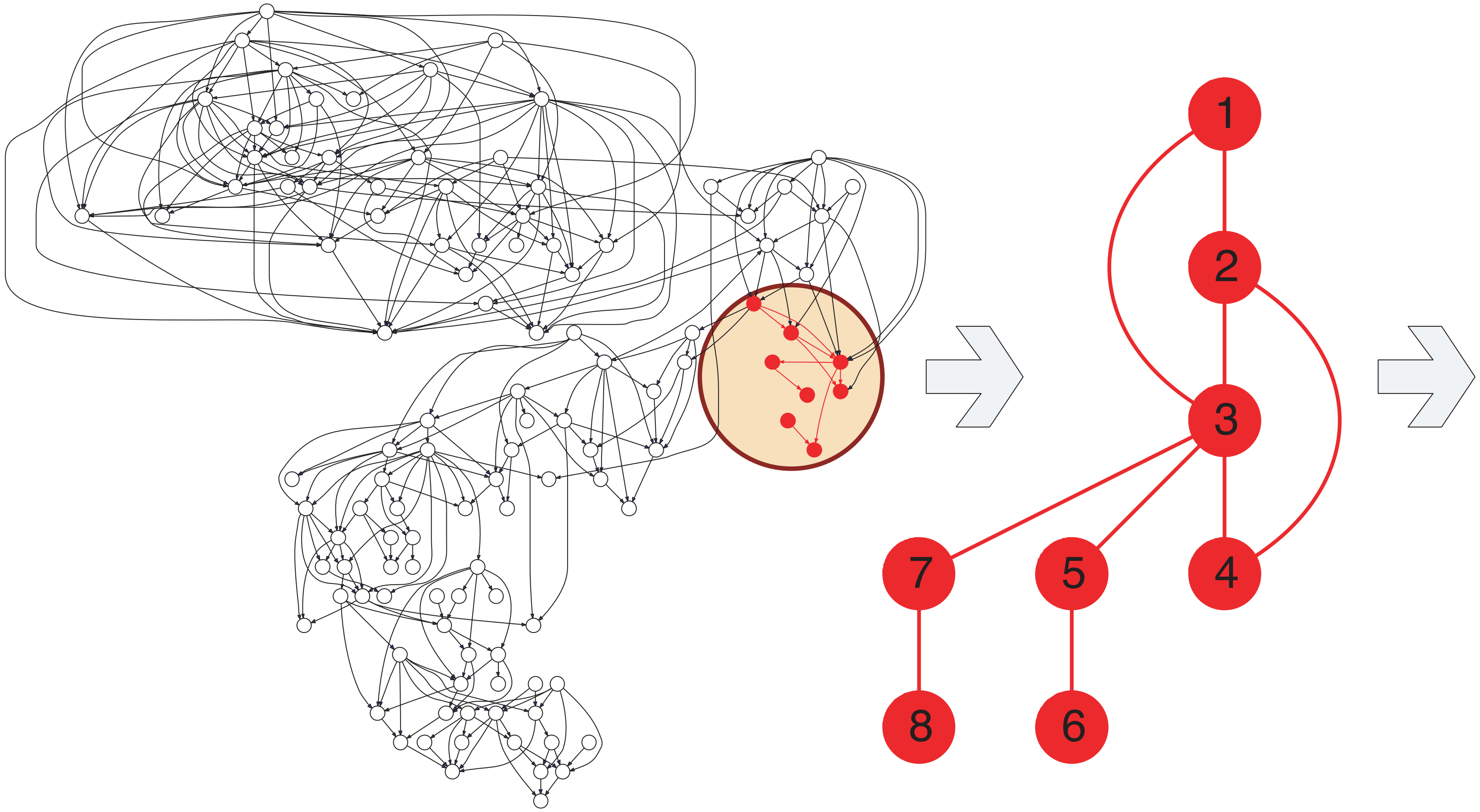
Protein sequence

RAGMCIAEKT

	Peptide sequence											
	R	A	G	M	C	I	A	E	K	T	Total	Average
Frequency in turn	0.09	0.06	0.15	0.06	0.13	0.06	0.06	0.06	0.10	0.08	0.75	0.08
Hydrophobic moment	10.0	0.00	0.00	1.90	0.17	1.20	0.00	3.00	5.70	1.50	21.97	2.44
Negative charge	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.11
Hydrophilicity	3.00	–0.50	0.00	–1.30	–1.00	–1.80	–0.50	3.00	3.00	–0.40	4.90	0.54
Beta sheet propensity	–0.40	–0.35	0.00	–0.46	–0.50	–0.60	–0.35	–0.40	–0.40	–0.48	3.46	0.38

[Mallick *et al.* Nat Biotech 2007]

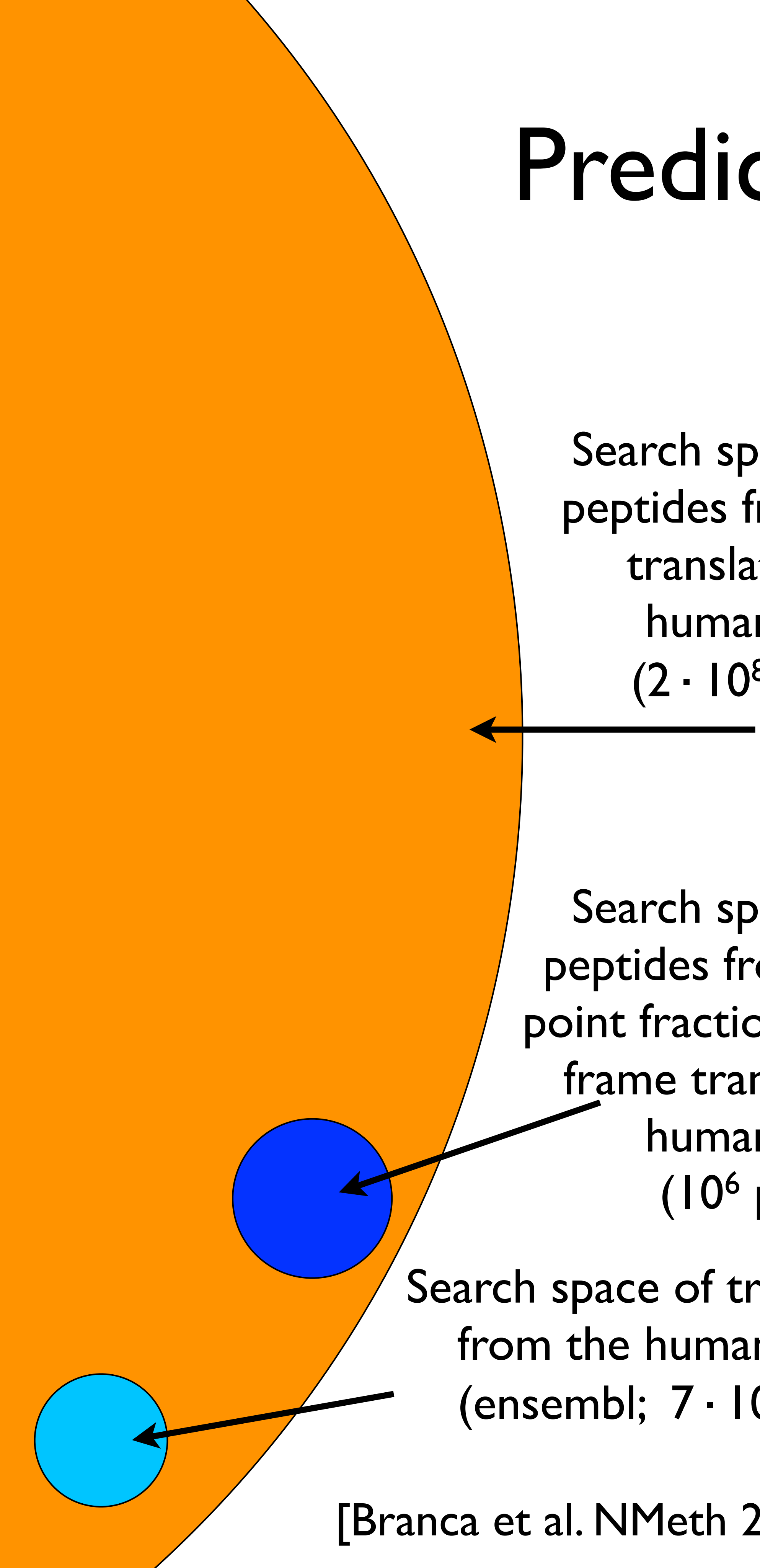
Spectral Alignment



1	KQGGTLDD	LEE	QAREL
2	KQGGTLDD	LEE	QARE
3	KQGGTLDD	LEE	QAR
4	KQGGTLDD	LEE	QA
5	KQGGTLDD	LEE^{-18}	QAR
6	KQGGTLDD	LEE^{-18}	Q
7	QGGTLDD	LEE	QAR
8	QGGTLDD^{-53}	LEE	QAR

[Bandeira *et al.* PNAS 2007]

Predicting properties of peptides

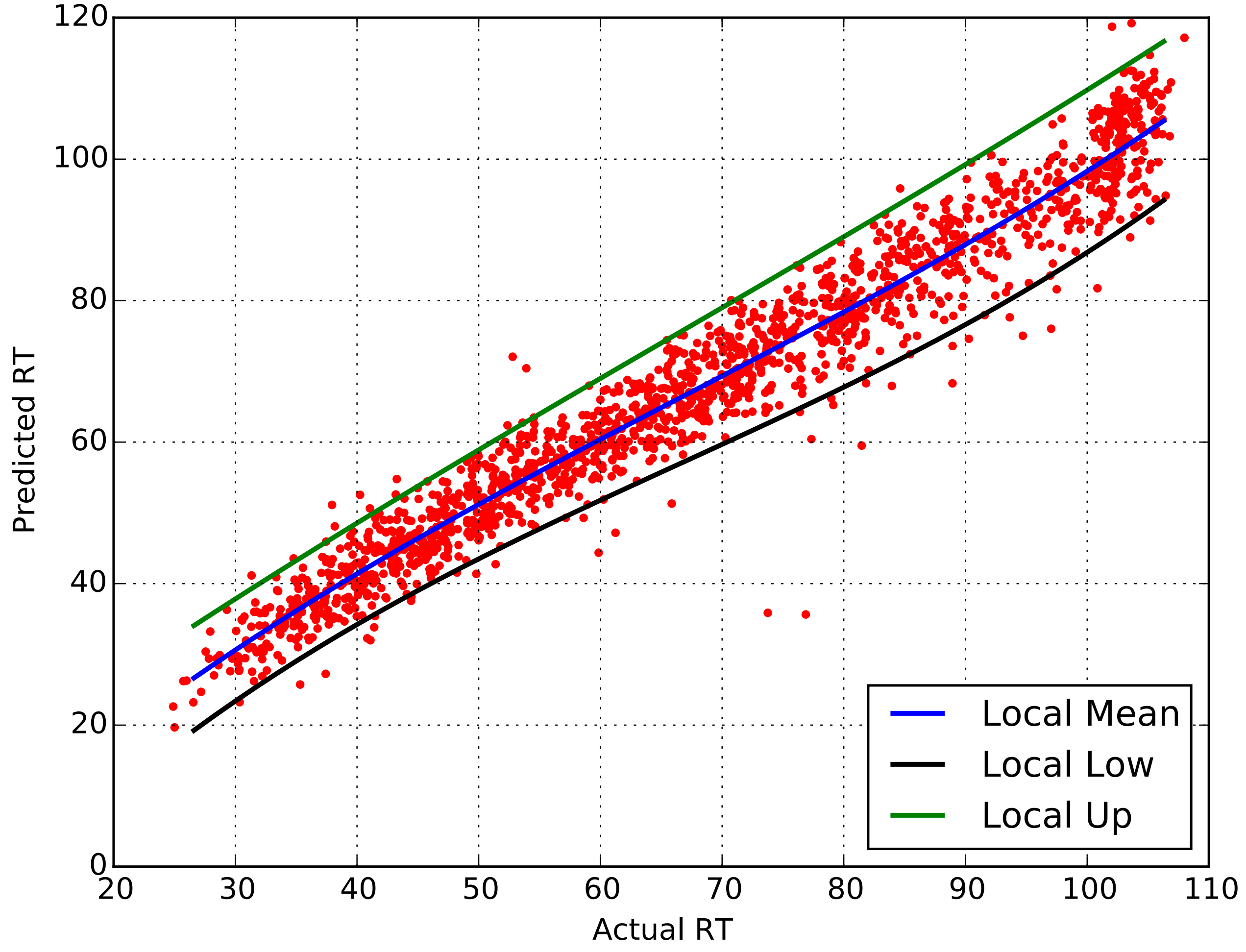


Search space of tryptic peptides from six frame translation of the human genome ($2 \cdot 10^8$ peptides)

Search space of tryptic peptides from iso-electric point fractionation of an six frame translation of the human genome (10^6 peptides)

Search space of tryptic peptides from the human proteome (ensembl; $7 \cdot 10^5$ peptides)

[Branca et al. NMeth 2014]



[Afkham et al. manuscript]

Conclusions

- Shotgun proteomics is currently the most accurate technique to analyze protein content of biological mixtures; detect protein complexes; and to detect and localize post translational modifications
- There is a large need of statistical and bioinformatical method development and education
- There are ample amount of data available waiting for your even more advanced analysis