Statistical mass spectrometry-based proteomics

Olga Vitek

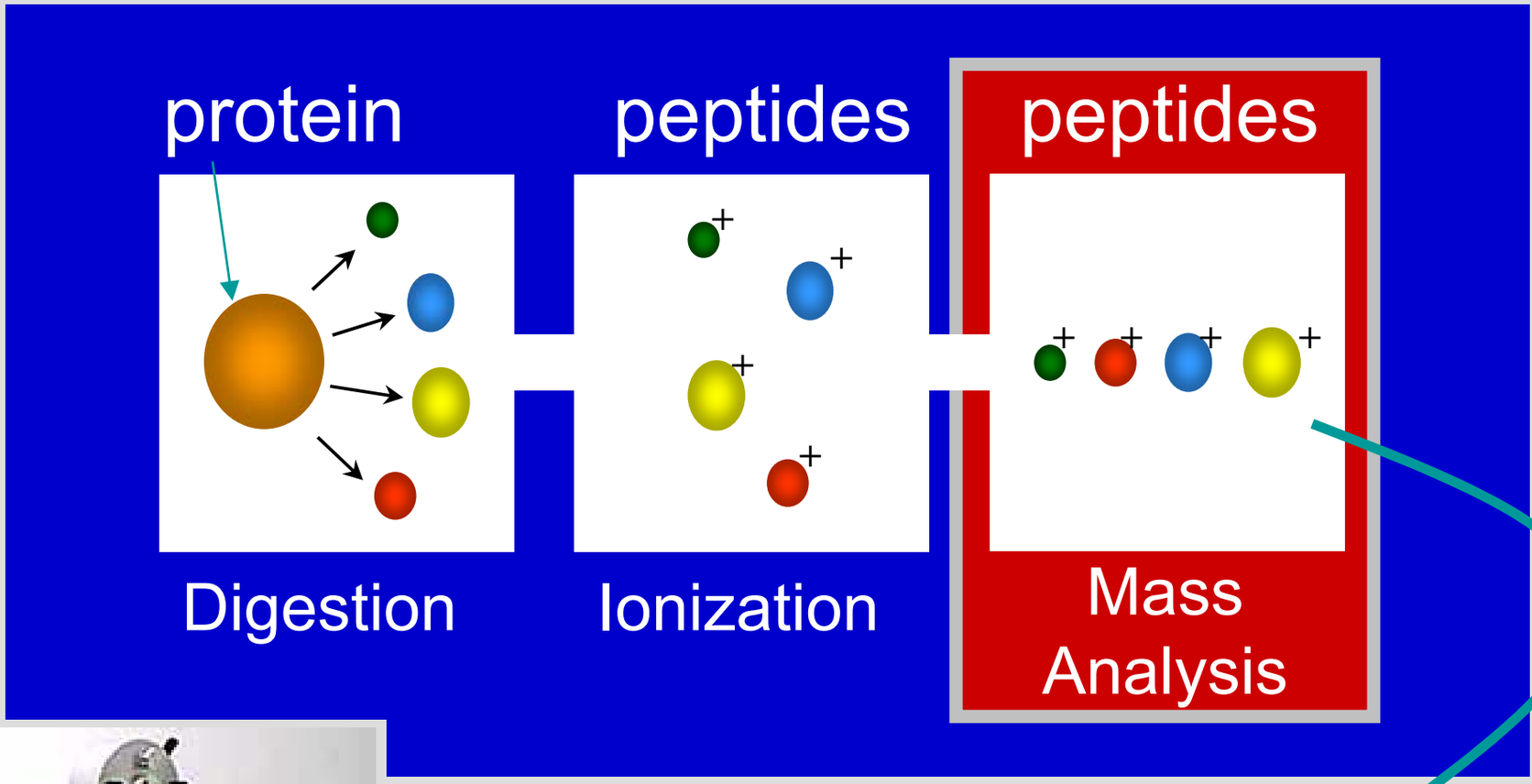
Outline

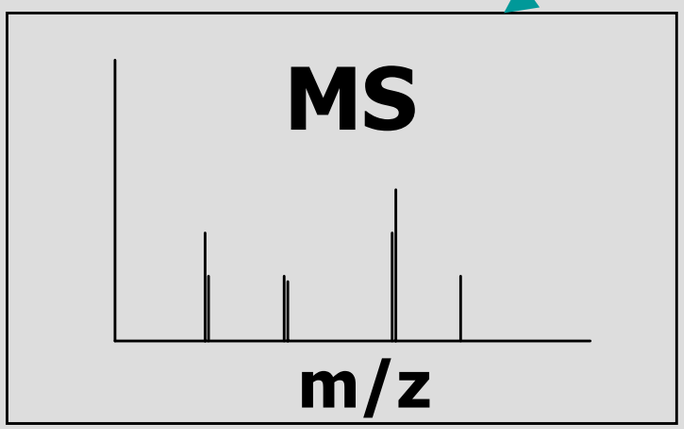
- What is proteomics?
  - Biological questions and technologies
- Protein quantification in label-free workflows
  - Joint analysis of multiple features and conditions
- Protein quantification in label-based workflows
  - Appropriately account for the labeling structure
- Mass spectrometry-based imaging
  - Account for the spacial heterogeneity of spectral data

Goals of proteomics

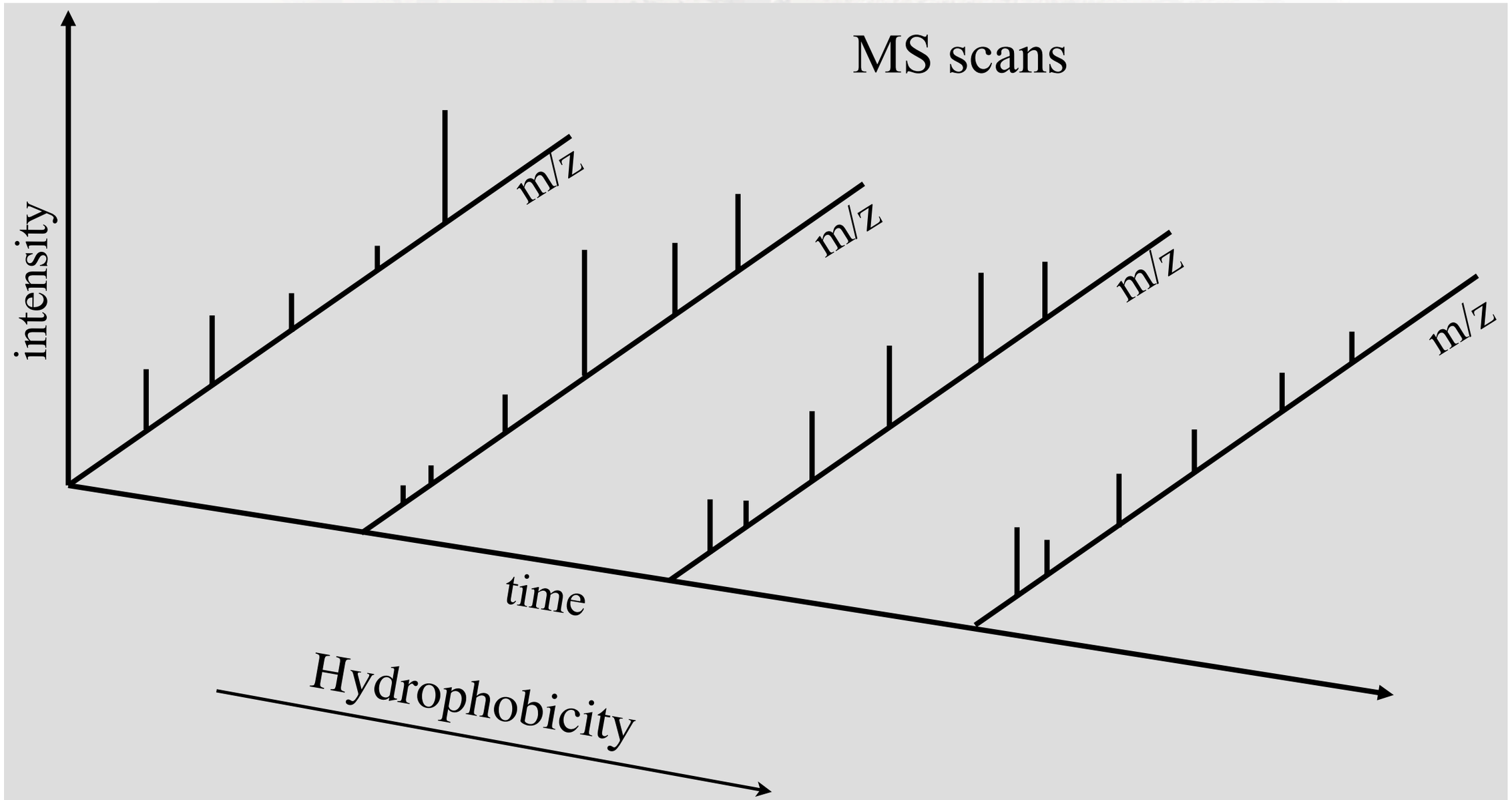
- Proteomics: system-wide characterization of all proteins
  - Sequence, structure, localization, abundance, PTMs, interactions
- More challenging than gene expression
  - *Complexity*
    - Human genome: ~20,000 protein coding genes
    - Their translation+splicing+proteolysis: ~50,000–500,000 proteins
    - Somatic DNA rearrangements and PTM: ~10 million
  - *Dynamic range*
    - > 10 orders of magnitude in plasma
    - Unlike nucleotides, proteins cannot be amplified
- How to make progress?
  - Sample preparation, separation, sensitive instruments
  - Statistical experimental design and accurate analysis

Acquisition of mass spectra

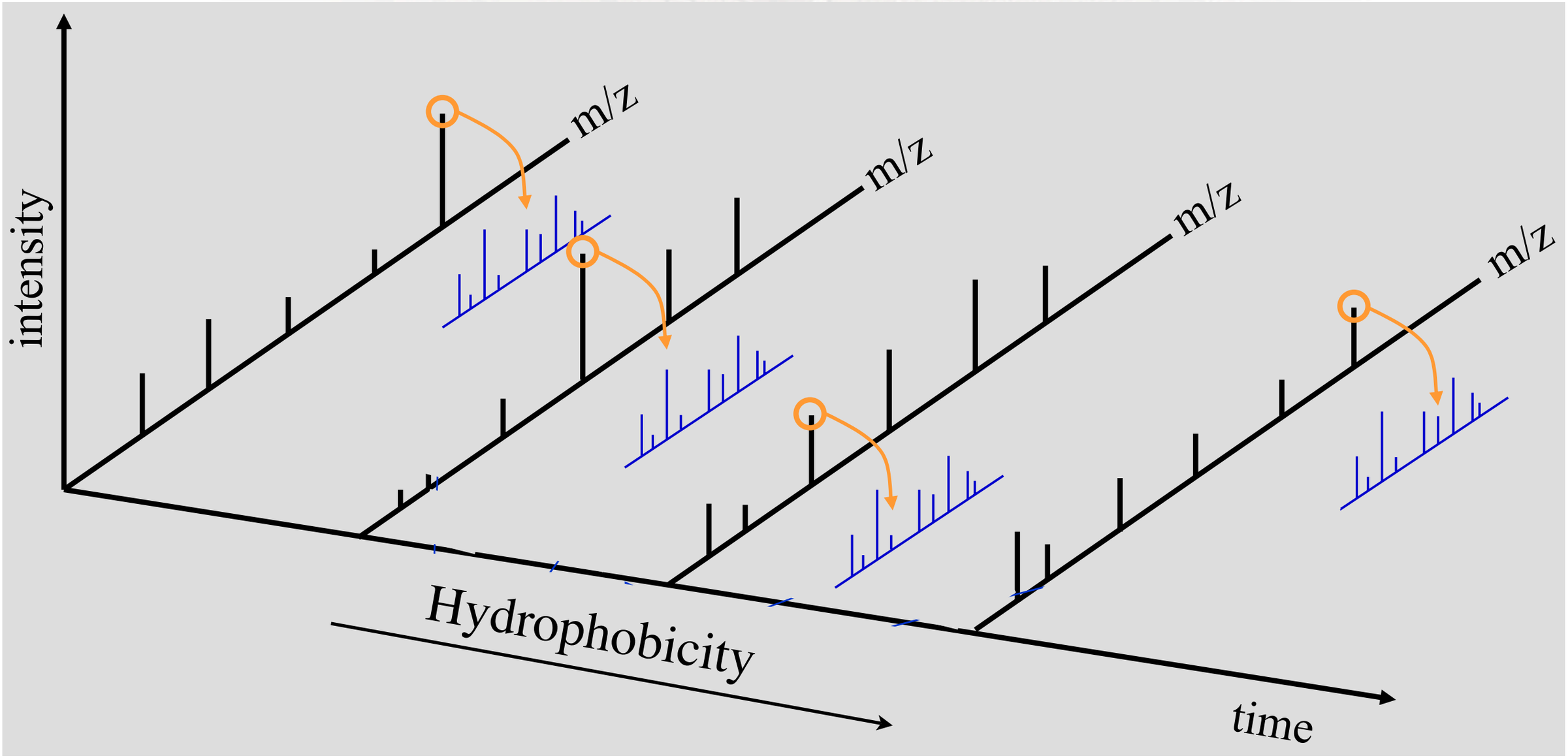




Liquid chromatography coupled with mass spectrometry (LC-MS)

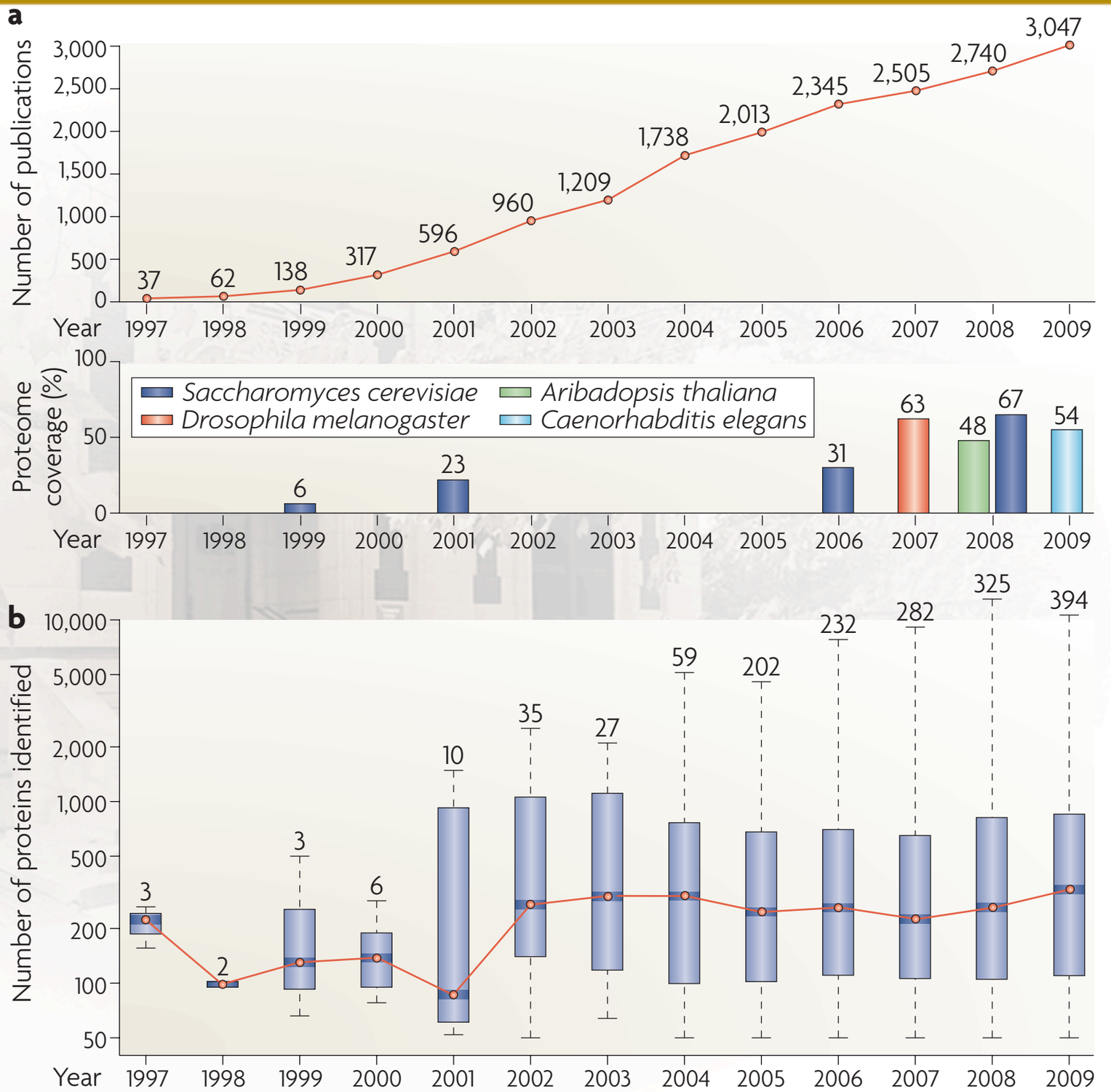


Liquid chromatography coupled with mass spectrometry (LC-MS)



Ahrens *et al.*, *Nature Reviews*, 2010

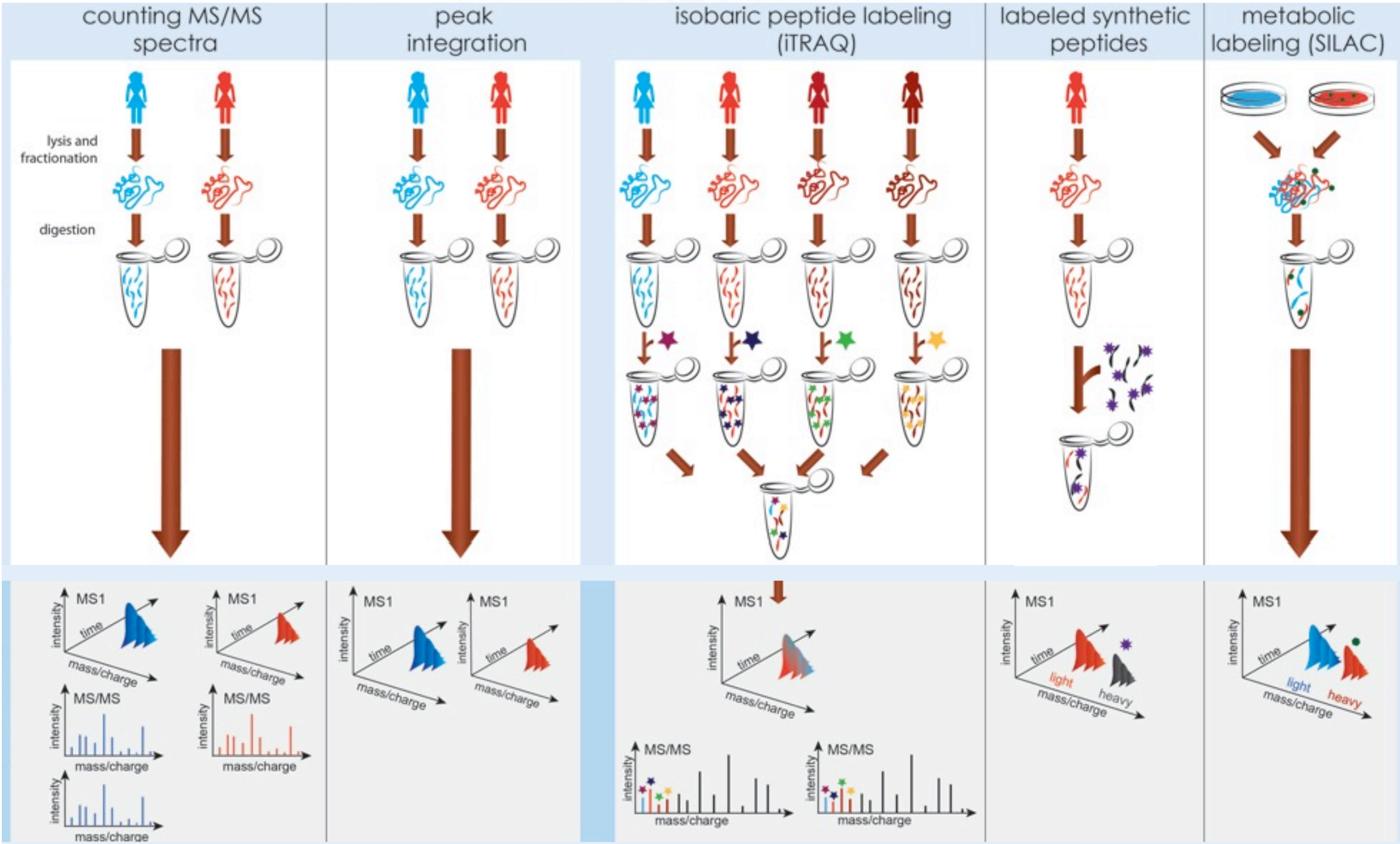
Proteomics is increasingly comprehensive



Global proteomic workflows

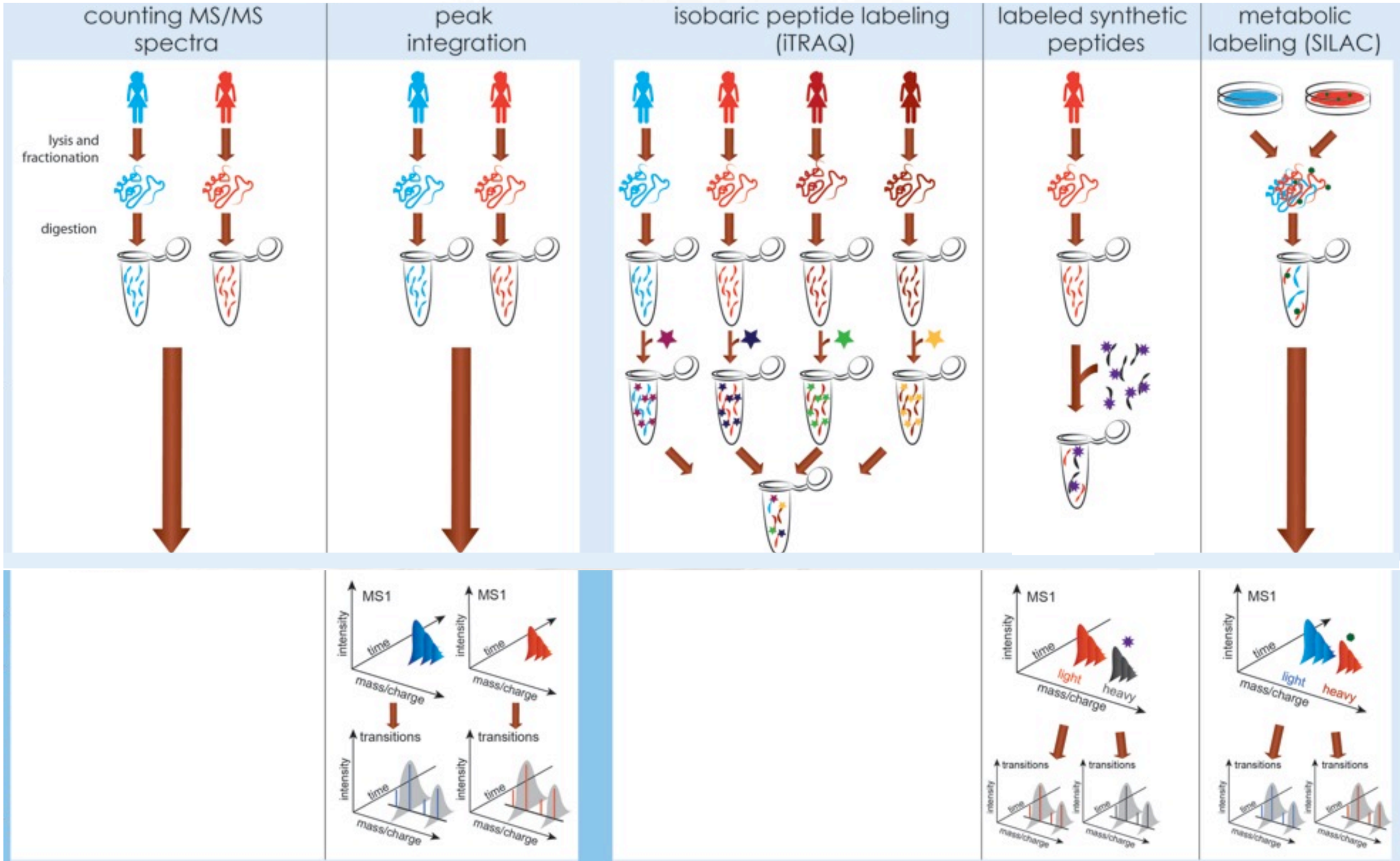
Sample preparation

Global LC-MS/MS



Käll and Vitek, *PLoS Computational Biology*, 7, 2011

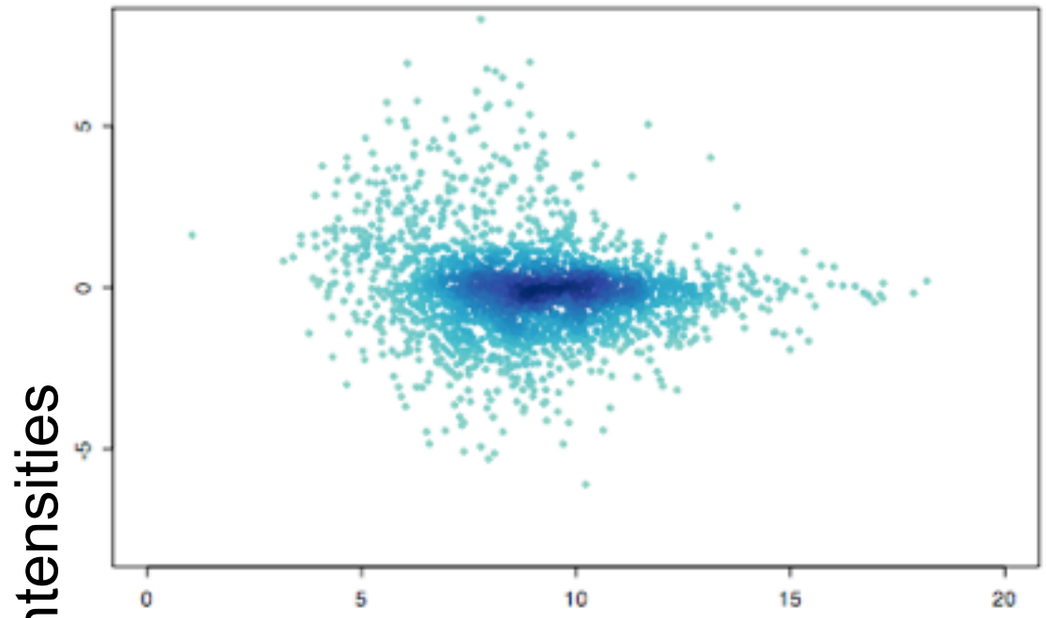
Targeted proteomic workflows



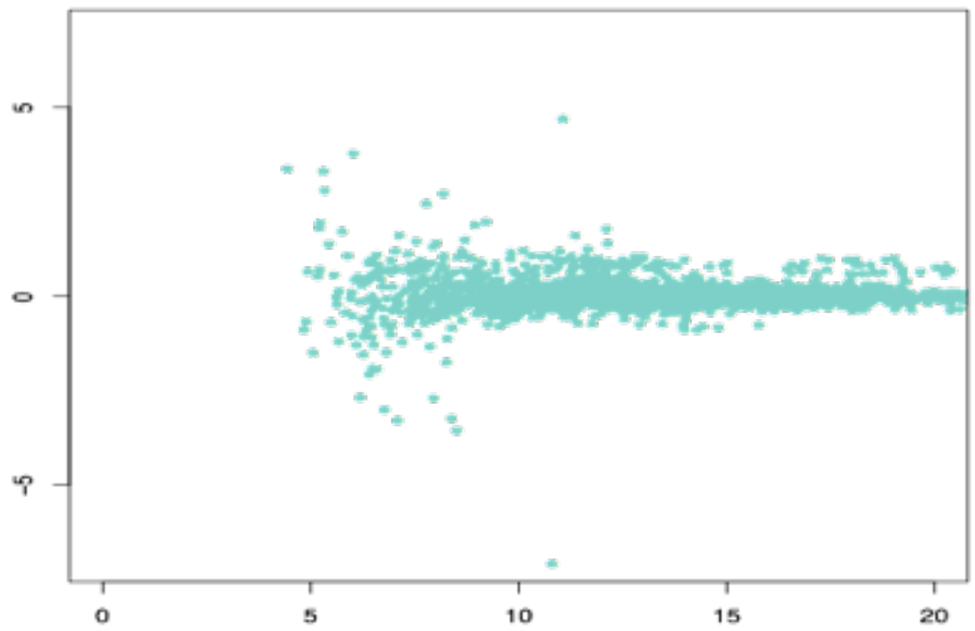
Käll and Vitek, *PLoS Computational Biology*, 7, 2011

Variation is experiment-specific

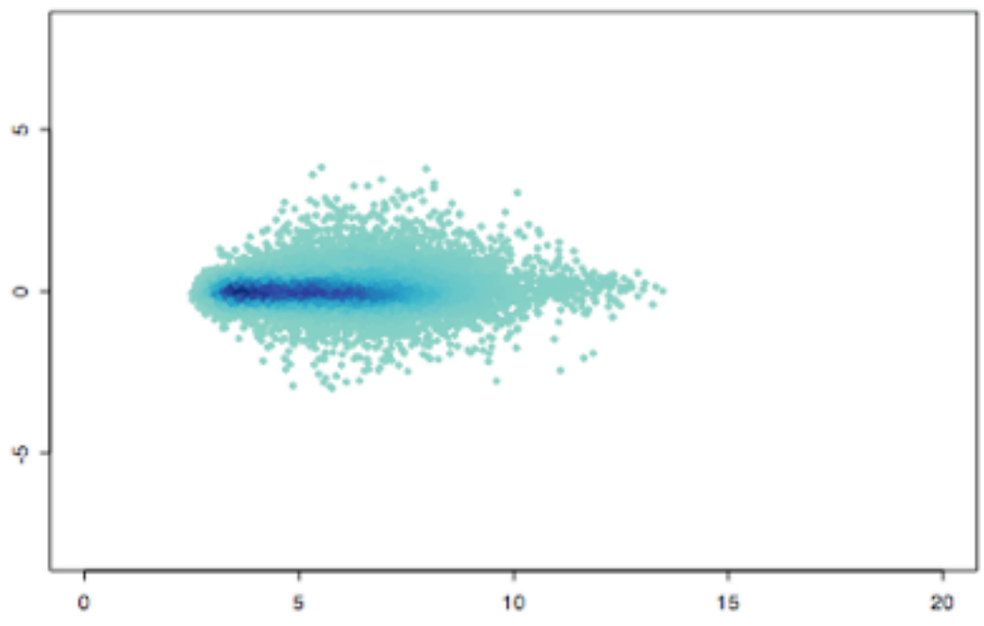
LC-MS: Human diabetes



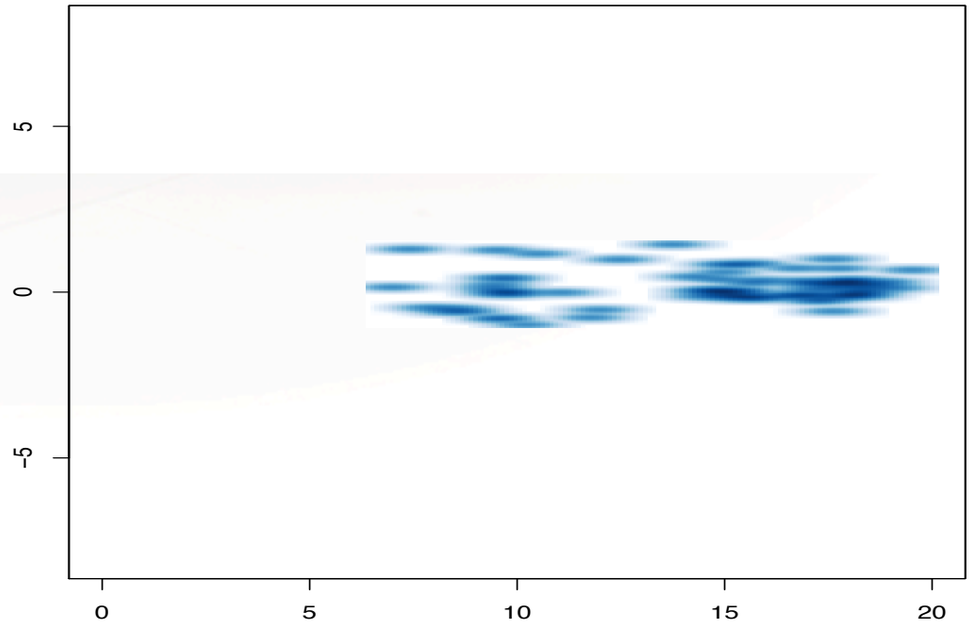
SRM: yeast proteome
(Picotti *et al*, 2009)



Affy: ALL (Acute Leukemia)



SRM: Human ovarian cancer
(Hüttenhain *et al*, 2012)



Outline

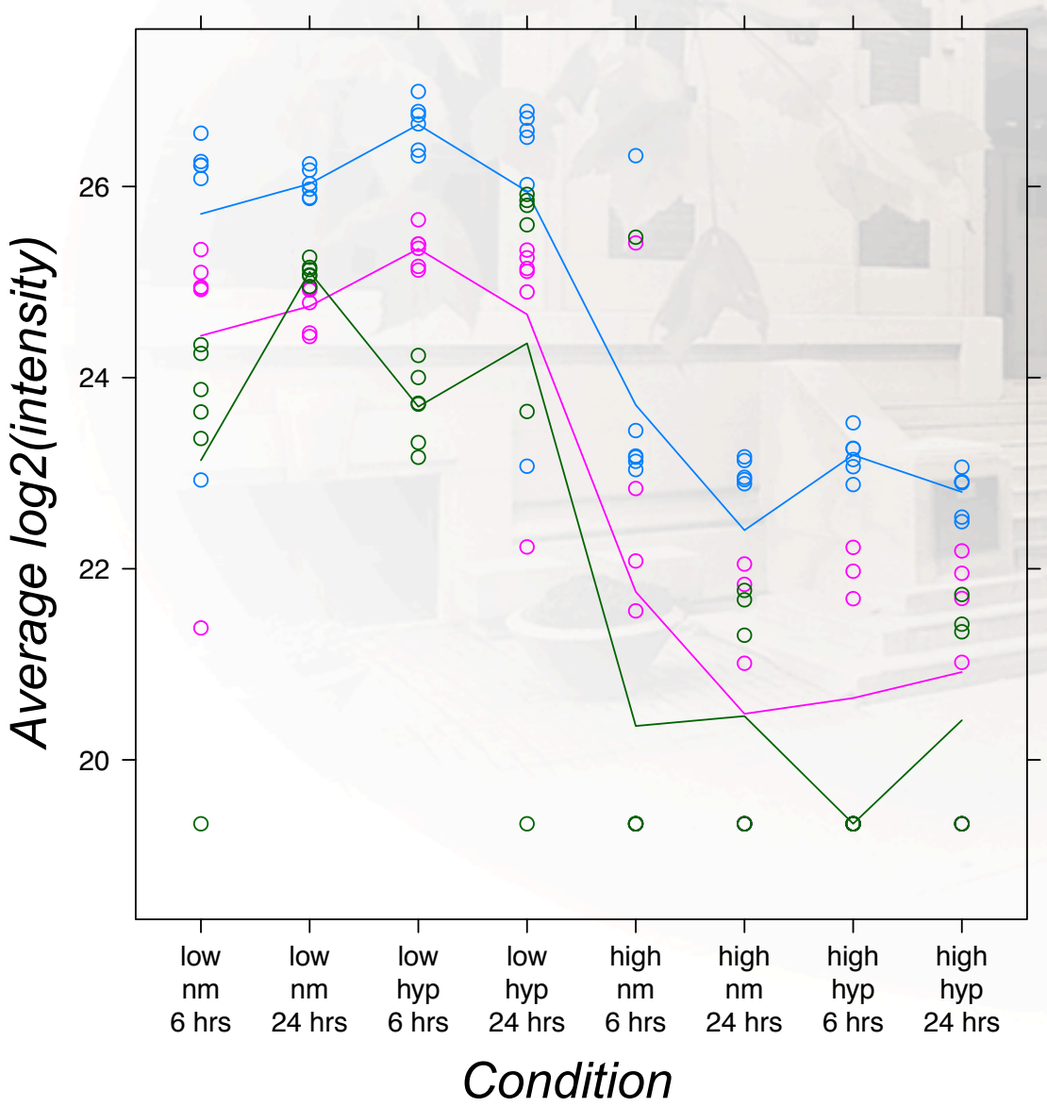
- What is proteomics?
  - Biological questions and technologies
- Protein quantification in label-free workflows
  - Joint analysis of multiple features and conditions
- Protein quantification in label-based workflows
  - Appropriately account for the labeling structure
- Mass spectrometry-based imaging
  - Account for the spacial heterogeneity of spectral data

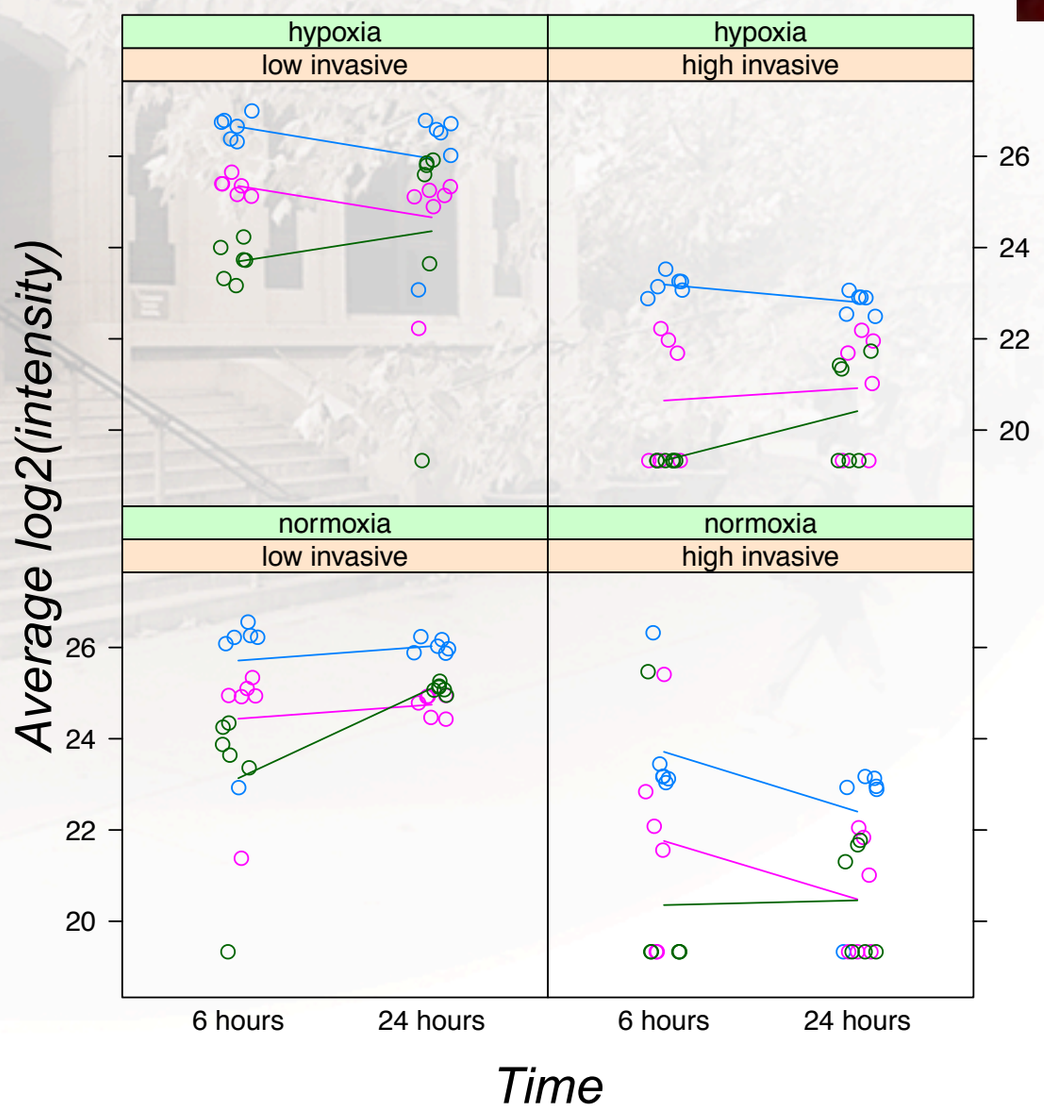
Example: label-free LC-MS

High/low invasive breast cancer cell lines

Goal: protein-level conclusions

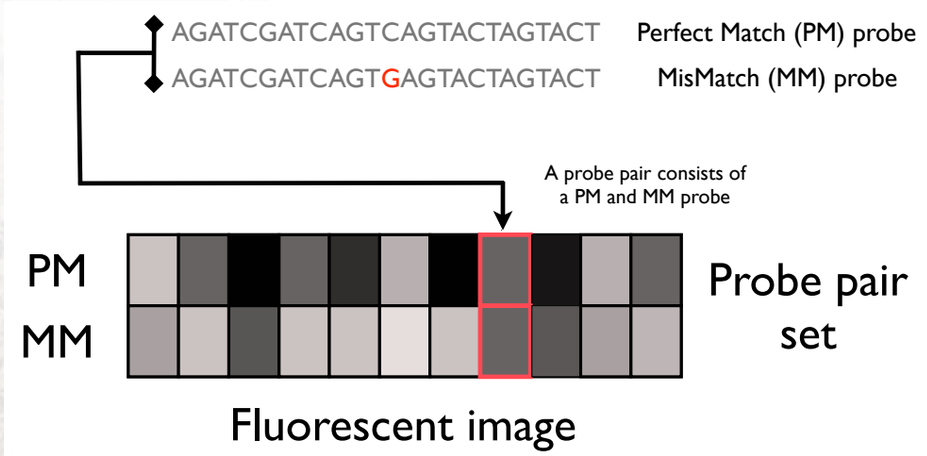
Safia Thaminy ETHZ





Important differences with microarrays

- Oligonucleotide microarrays
  - RMA: Tukey Median Polish
    - Robust averaging of all probes array-specific summary



Sebastiani et al, Statistical Science, 2003.

- Proteomics
  - Number and quality of features per protein vary widely
  - Missing features introduce imbalance
  - Label-based workflows combine multiple samples
    - blocking structure
  - Targeted workflows create a nested structure protein/peptide/transition
- Explicit probabilistic models best represent the data

Linear mixed models for feature intensities

$$\underbrace{\log(\text{peak intensity})}_{y_{ijkl}} = \underbrace{\text{Expected reference abundance}}_{\mu_{111}} + \overbrace{\underbrace{\text{LC-MS feature}}_{F_i} + \underbrace{\text{condition}}_{C_j} + \underbrace{\text{feature} \times \text{condition interaction}}_{(F \times C)_{ij}} + \underbrace{\text{biol. replicate}}_{S(C)_{k(j)}}}^{\text{Deviation from the reference due to}} + \underbrace{\text{Random meas. error}}_{\varepsilon_{ijkl}}$$

where $F_1 = C_1 = (F \times C)_{i1} = (F \times C)_{1j} = 0$ $\qquad \varepsilon_{ijkl} \overset{iid}{\sim} N\left(0, \sigma^2_{Error,\, ijk}\right)$

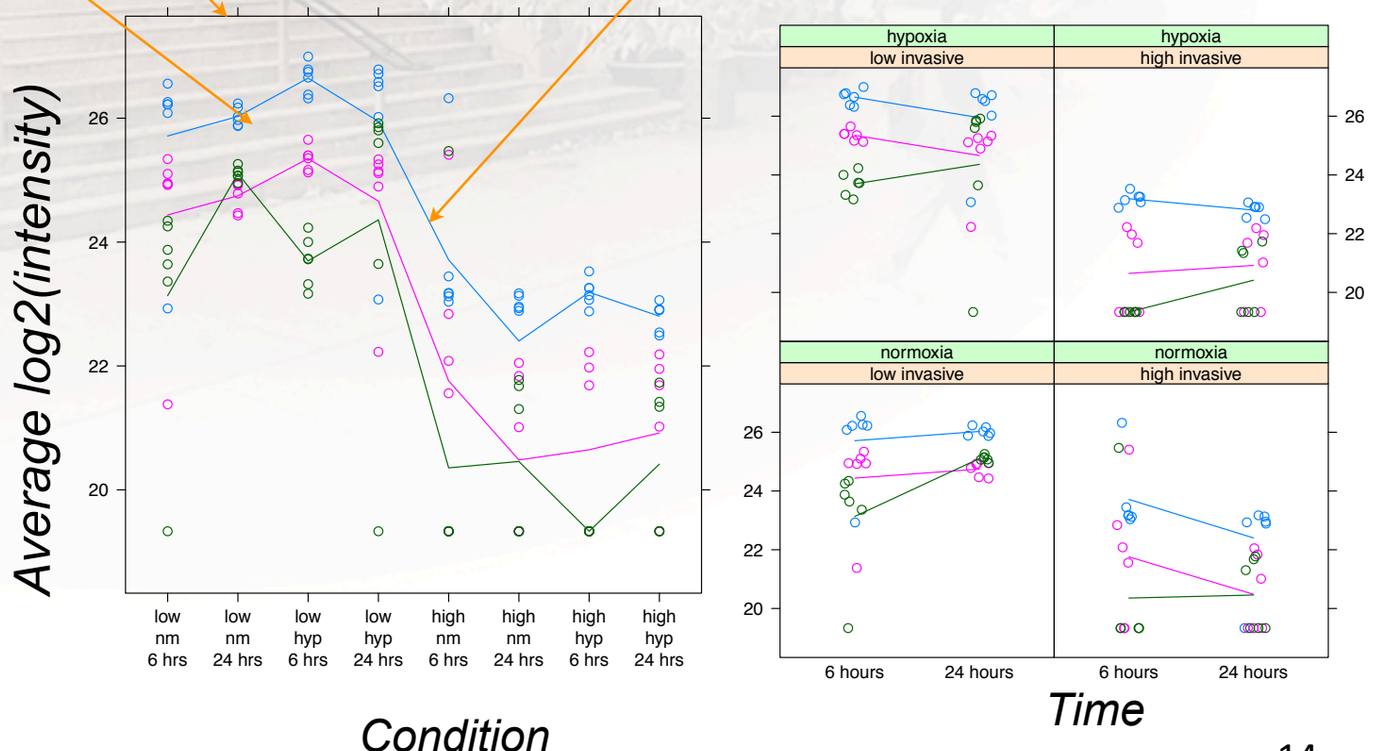
and
(a) reduced scope of biological replication: $S(C)_{1(1)} = 0$
(b) expanded scope of biological replication: $S(C)_{k(j)} \overset{iid}{\sim} N\left(0, \sigma^2_S\right)$

T. Clough et al.
JPR, 2009

T. Clough et al.
Meth. Mol. Biol., 2011

T. Clough et al.
BMC Bioinformatics., 2012

C.-Y. Chang et al.
Mol. Cel. Proteomics, 2012



All inference is based on problem-specific linear combinations of model terms

Quantity of interest:

$$H_0 : L = \bar{\mu}_{[\text{high, nm, 6}]\cdot} - \bar{\mu}_{[\text{low, nm, 6}]\cdot} = 0$$

Model-based estimate and test statistic:

$$\hat{L} = \hat{C}_{[\text{high, nm, 6}]} + \frac{1}{I}\sum_{i=1}^{I}\widehat{(F \times C)}_{i,[\text{high, nm, 6}]} + \frac{1}{K}\sum_{k=1}^{K}\widehat{S(C)}_{k([\text{high, nm, 6}])}$$

$$- \left(\hat{C}_{[\text{low, nm, 6}]} + \frac{1}{I}\sum_{i=1}^{I}\widehat{(F \times C)}_{i,[\text{low, nm, 6}]} + \frac{1}{K}\sum_{k=1}^{K}\widehat{S(C)}_{k([\text{low, nm, 6}])}\right)$$

$$t = \frac{\hat{L}}{SE\{\hat{L}\}} \sim \text{Student distribution}$$

In balanced datasets:

$$\hat{L} = \bar{Y}_{\cdot[\text{high, nm, 6}]\cdot\cdot} - \bar{Y}_{\cdot[\text{low, nm, 6}]\cdot\cdot}$$

$$t = \frac{\hat{L}}{\sqrt{\frac{2}{IKL}\hat{\sigma}^2_{Error}}} \sim \text{Student}_{IJK(L-1)+(I-1)J(K-1)} \text{ distribution}$$

Also methods for model diagnostics, data visualization etc

Tools for quantitative proteomics


Veavi Chang


Meena Choi


Tim Clough

- A variety of quantitative workflow
  - *Global, targeted, data-independent*
  - *Label-free and label-based*
- Accounts for experimental designs
  - *Group comparison, time course*
- Data visualization & QC
- Model-based analysis
  - *Model fitting and inference*
- Planning future experiments
  - *Sample size, resource allocations*

Since 2010:
- *extensive documentation*
- *published case studies*
- *protocols for typical analyses*
- *13 tutorials and workshops*

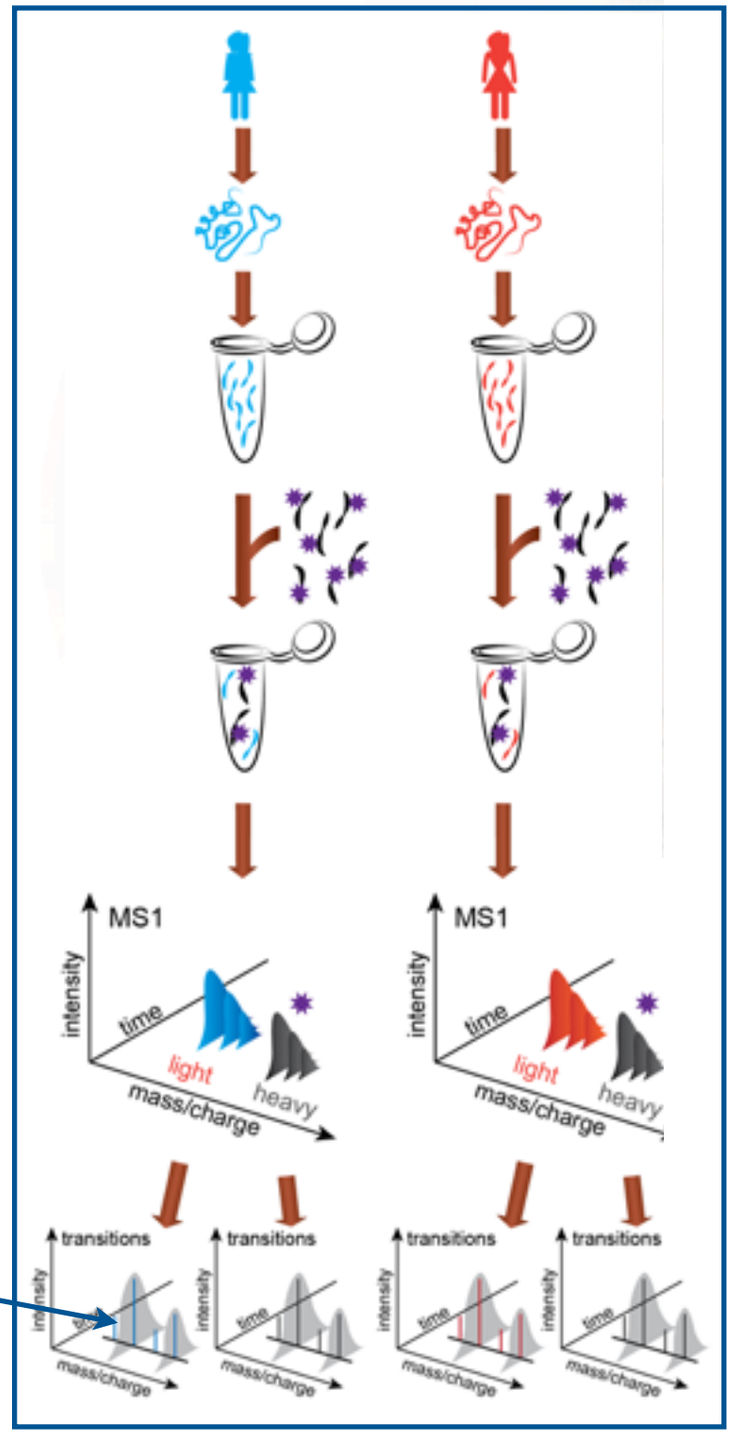
Since 2011:
- *370 unique visitors*
- *over 50 unique downloads*
- *over 50 mailing list members*

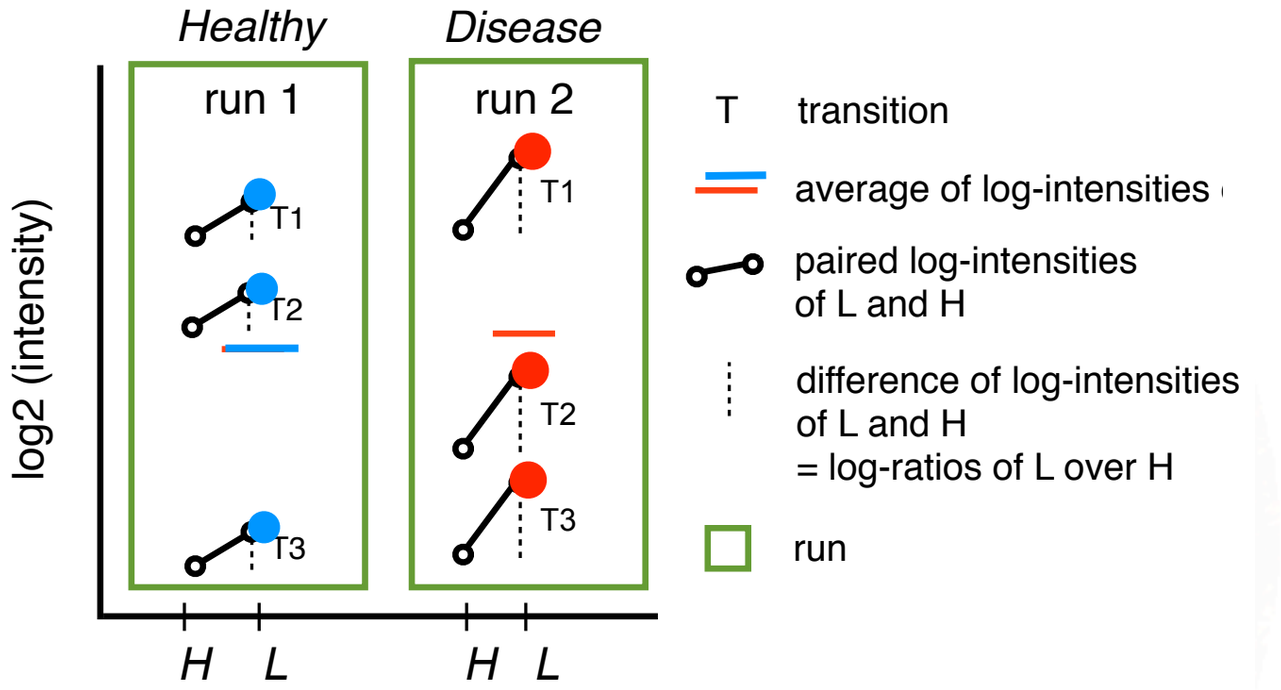
Collaboration: Ruedi Aebersold (ETH Zürich), Michael MacCoss

Outline

- What is proteomics?
  - Biological questions and technologies
- Protein quantification in label-free workflows
  - Joint analysis of multiple features and conditions
- Protein quantification in label-based workflows
  - Appropriately account for the labeling structure
- Mass spectrometry-based imaging
  - Account for the spacial heterogeneity of spectral data

Label-based workflows help separate the biological and the technological variation





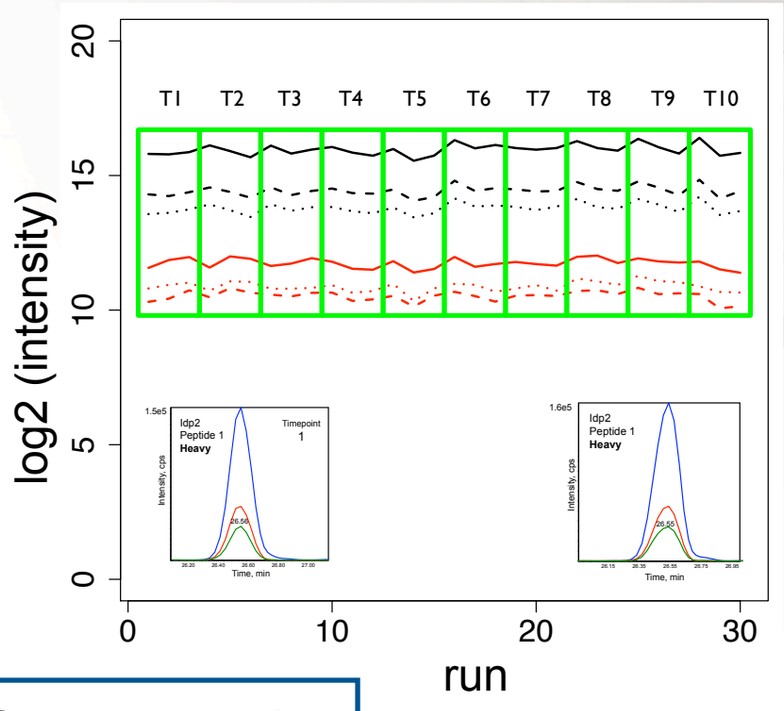
			Group 1				Group I	
			Run 1 Subject 1	Subject J		Subject 1		Run M Subject J
Endogenous: light labeled peptide	Peptide 1	Transition 1	10.21	10.57		15.64		15.03
		Transition L	10.52	10.92		15.29		15.68
	Peptide K	Transition 1	11.76	11.92		16.22		16.71
		Transition L	11.65	11.09		16.27		16.51
Reference: heavy labeled peptide	Peptide 1	Transition 1	19.46	19.77		19.82		19.03
		Transition L	19.13	19.25		19.67		19.80
	Peptide K	Transition 1	19.26	19.33		19.58		19.61
		Transition L	19.73	19.09		19.84		19.55

Legend : Label | Feature: Transition/Peptide | Group | Run | Subject

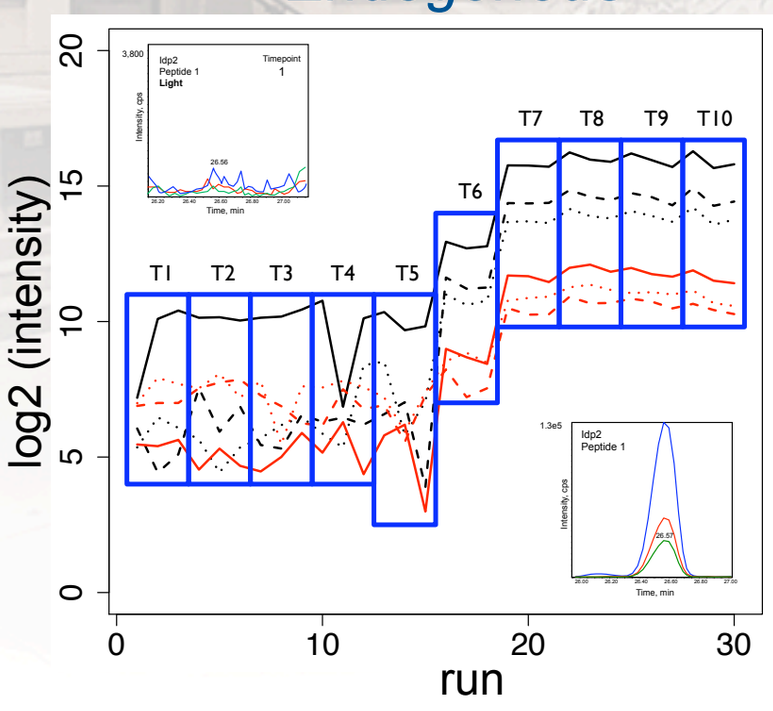
More complex but similar linear mixed effects models

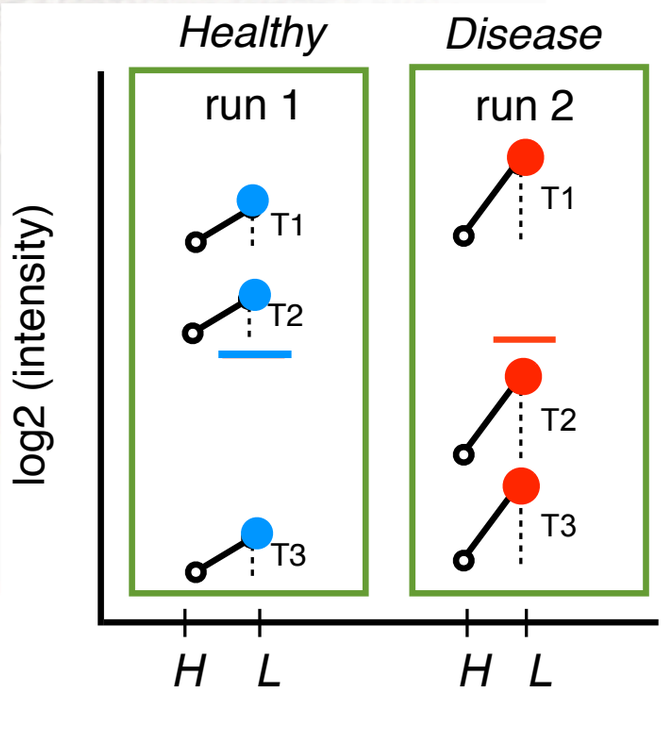
		Systematic deviations from the mean					Statistical interactions			
Observed log2(int of peak)	= **Overall mean**	+ **Group or time**	+ **Subject**	+ **Feature**	+ **Run**	+ **Time by subject**	+ **Time by feature**	+ **Run by feature**	+	**Random error**
y_{ijklm}	= μ	+ T_i	+ $S(G)_{j(i)}$	+ F_{kl}	+ R_m	+ $(T \times S)_{ij}$	+ $(T \times F)_{ikl}$	+ $(R \times F)_{klm}$	+	ε_{ijklm}
Fixed/Random	F	F	F/R	F	F/R	F/R	F	F/R		R: N(0,σ^2)

Reference



Endogenous





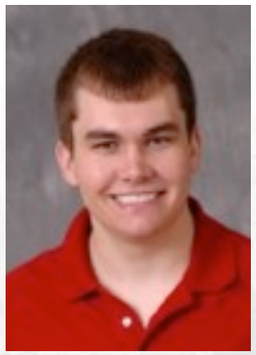
C.-Y. Chang et al. *MCP*, 2012

STATISTICAL PROTEOMICS AND BIOINFORMATICS

Purdue University


Kyle Bemis


Meena Choi


Tim Clough


Veavi Chang


Robert Ness


Danni Yu


Cheng Zheng

ETH Zürich

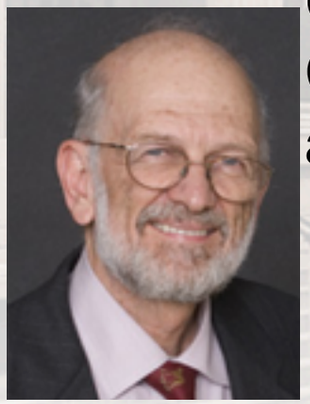

Ruedi Aebersold and lab


Ruth Hüttenhain


Safia Thaminy


Silvia Surinova

Purdue


Graham Cooks and lab

Stanford


Mark Stolowtiz


Parag Mallick

CGR, Spain


Eduard Sabidó

- NSF-SI2-SSE
- NSF-BIO-DBI