Some considerations from the data generation side

May 10th, 2017
Boston

Ruedi Aebersold
ETH Zurich

A draft of the human genome (2000)

...announced that the international Human Genome Project and Celera Genomics Corporation have both completed an initial sequencing of the human genome -- the genetic blueprint for human beings. He congratulated the scientists working in both the public and private sectors on this landmark achievement, which promises to lead to a new era of molecular medicine, an era that will bring new ways to prevent, diagnose, treat and cure disease.

Now, scientists will be able to use the working draft of the human genome to:

* Alert patients that they are at risk for certain diseases.

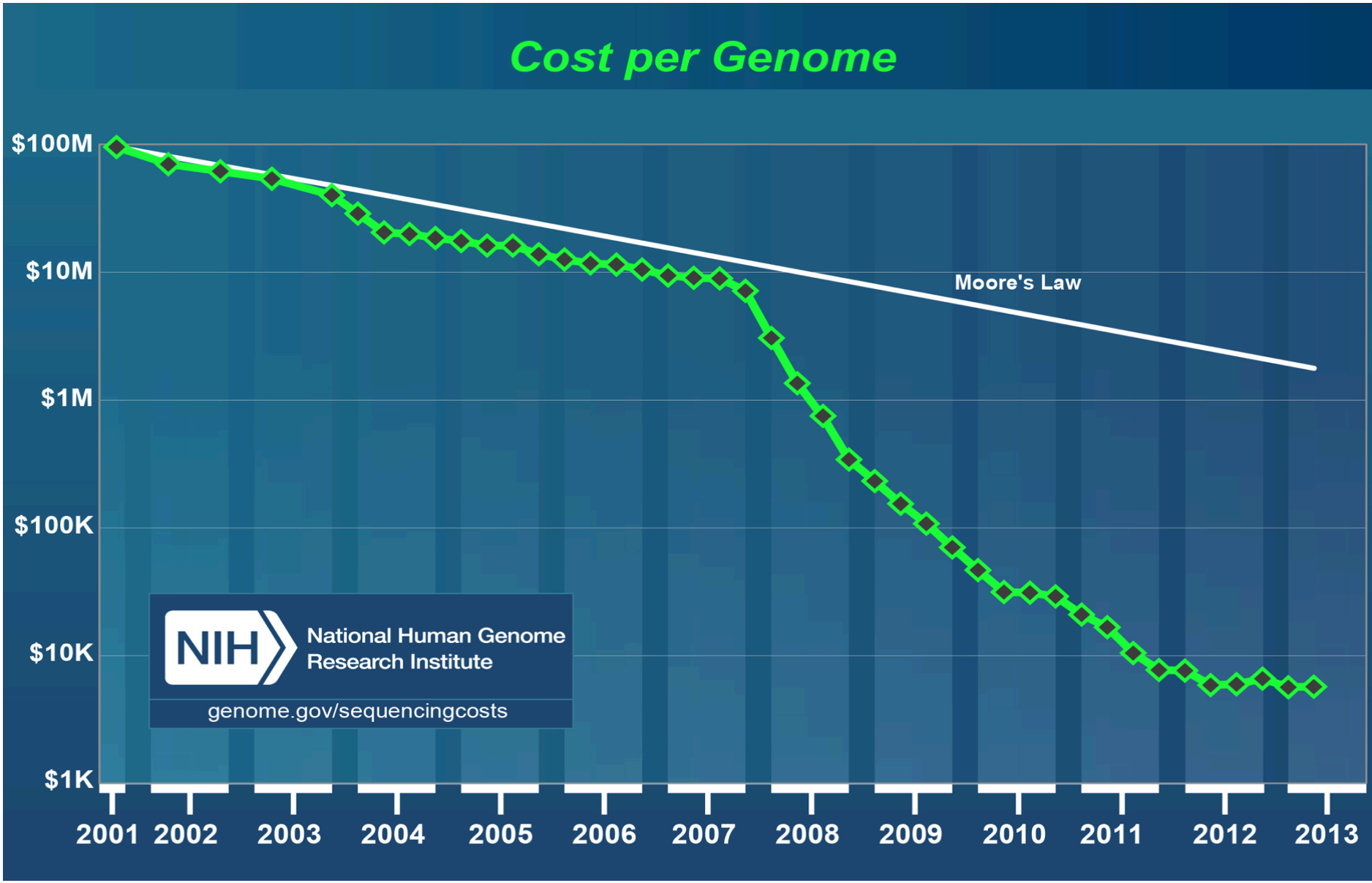
* Reliably predict the course of disease.

•Precisely diagnose disease and ensure the most effective treatment is used.

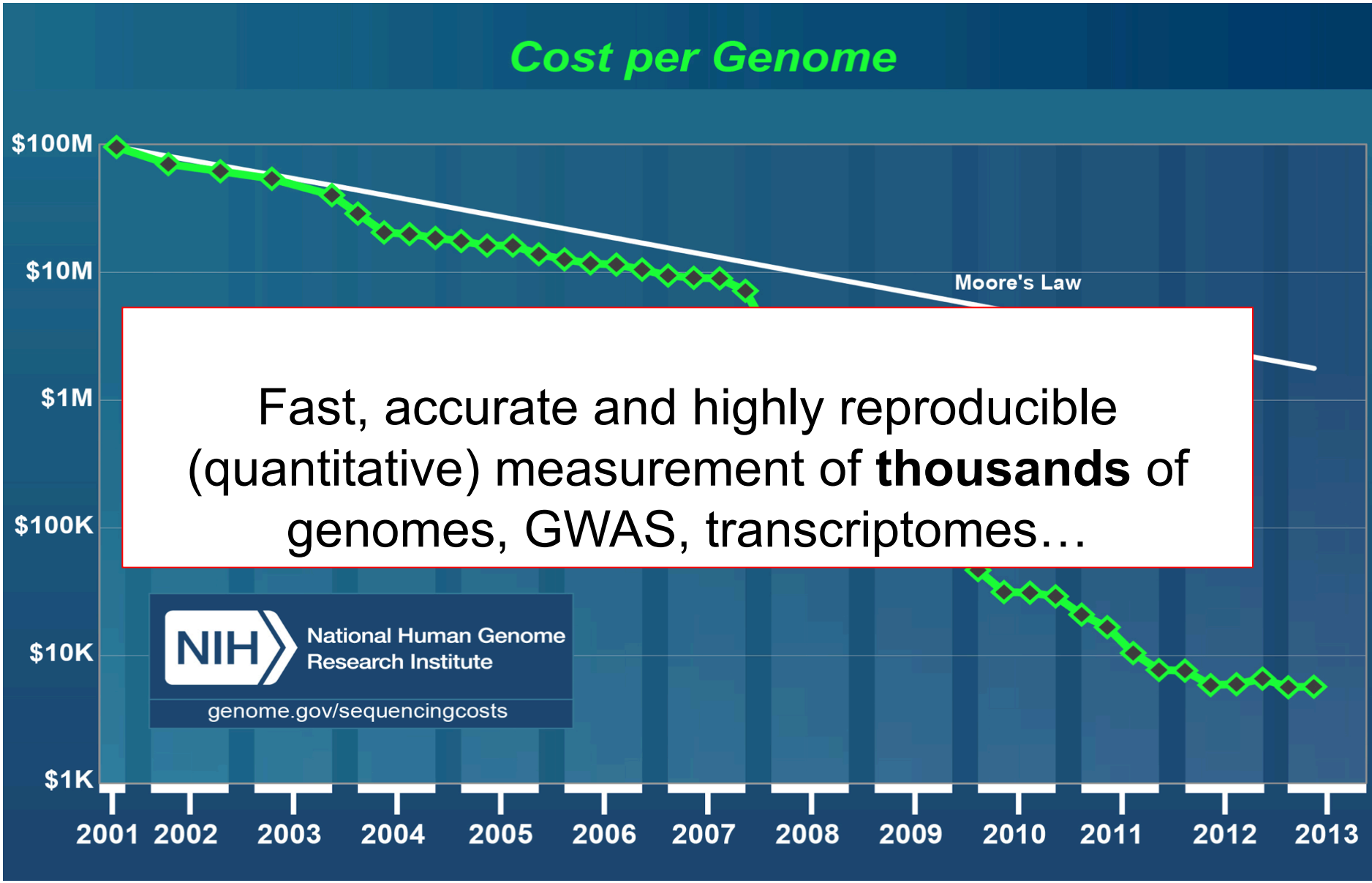
•Developing new treatments at the molecular level

Press release US. Govt.

Then came the genomics revolution



Then came the genomics revolution….



Nature Reviews Molecular Cell Biology | AOP, published online 10 November 2010; doi:10.1038/nrm2973

REVIEWS

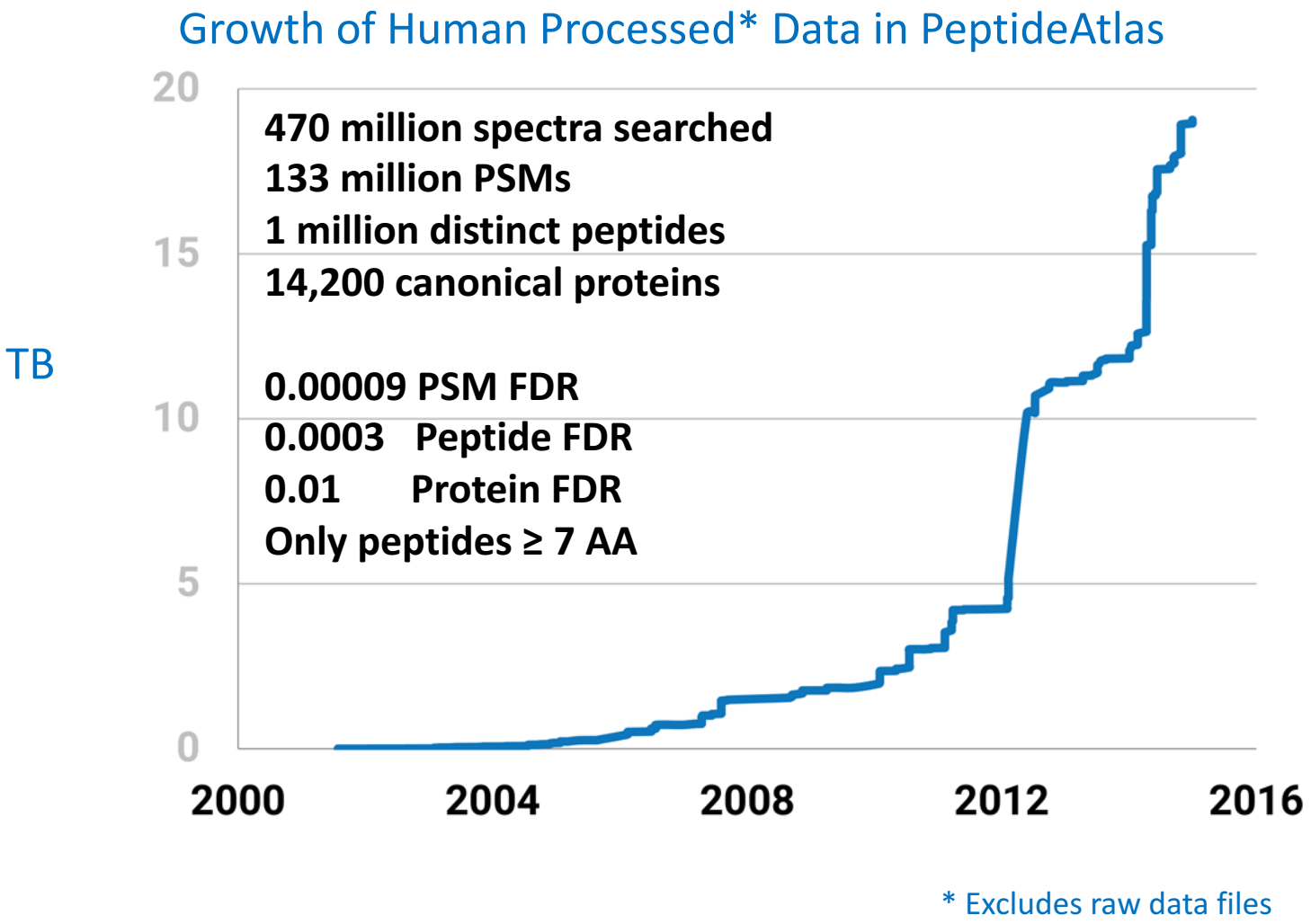
Generating and navigating proteome maps using mass spectrometry

Christian H. Ahrens, Erich Brunner*, Ermir Qeli*, Konrad Basler*‡ and Ruedi Aebersold§*

Abstract | Proteomes, the ensembles of all proteins expressed by cells or tissues, are typically analysed by mass spectrometry. Recent technical and computational advances have greatly increased the fraction of a proteome that can be identified and quantified in a single study.

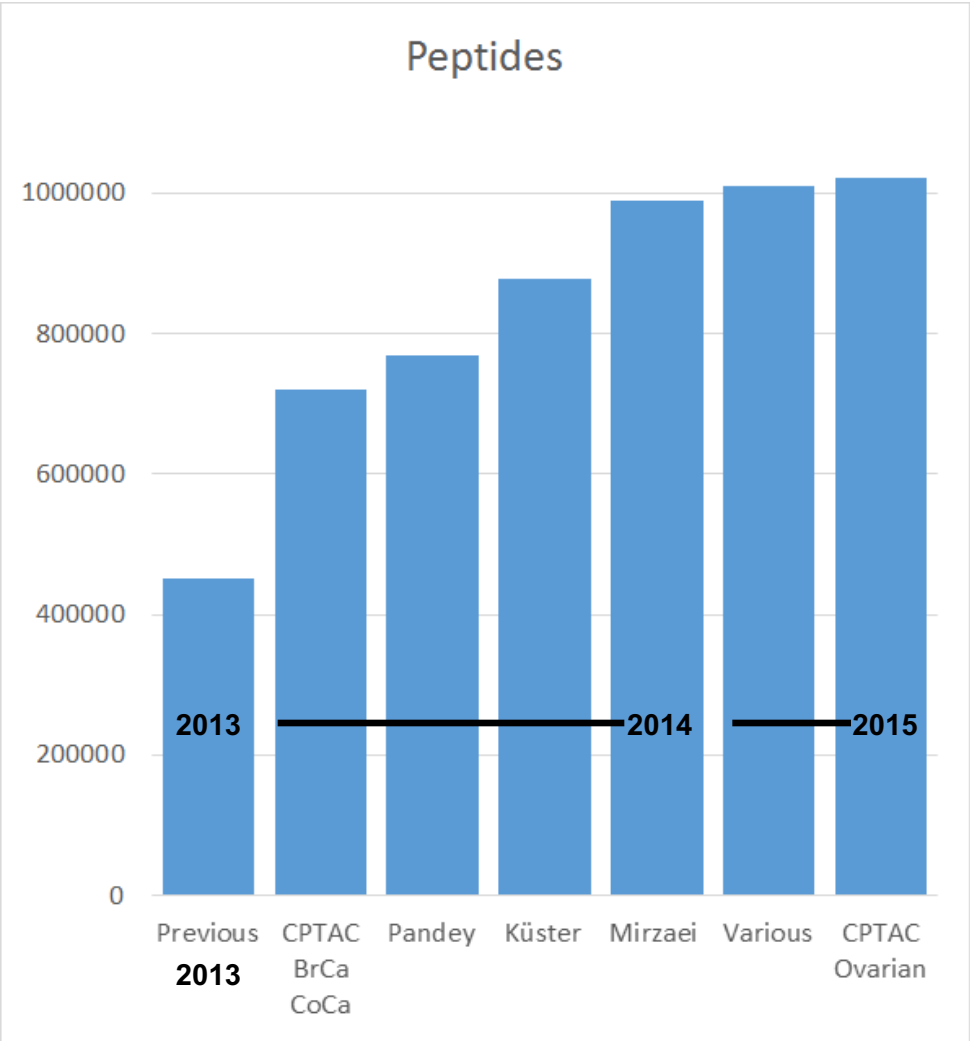
- **Generate (complete) proteome maps by MS.**
  - Discovery proteomics
- **Then use (complete) proteome maps to learn new biology by the reproducible, (quantitative) measurement of large numbers of samples.**

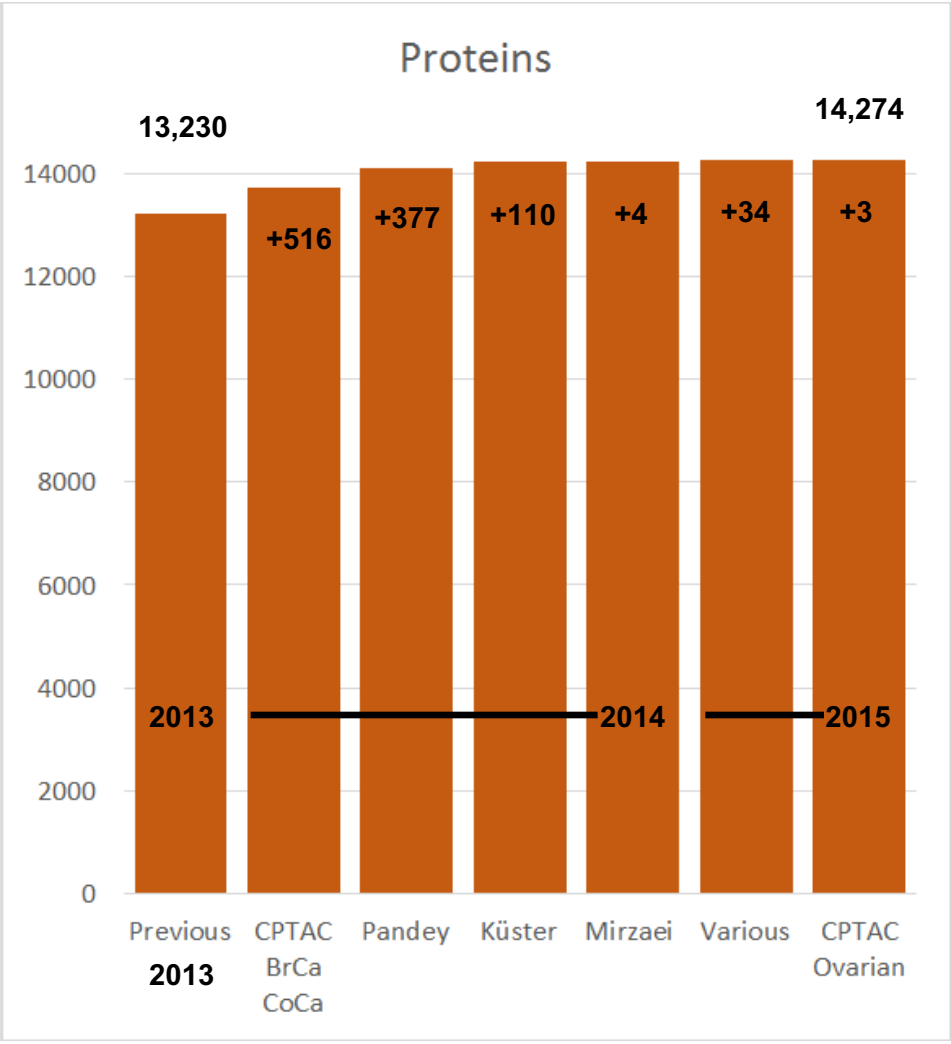
Human PeptideAtlas: Saturation **community** mapping of the human proteome



Courtesy: Eric Deutsch, ISB

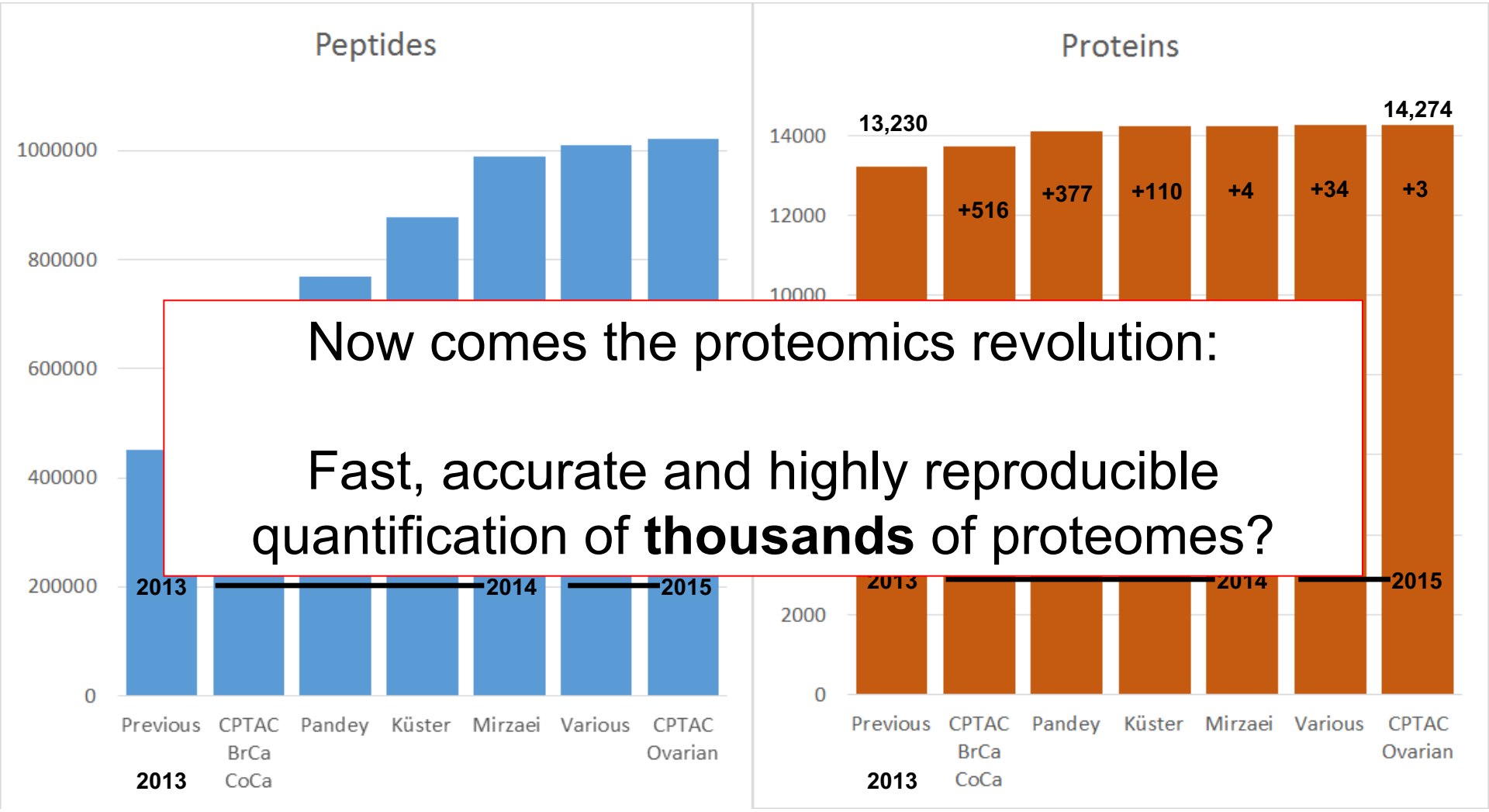
Saturation of the human proteome





Courtesy: Eric Deutsch, ISB

Saturation of the human proteome



Now comes the proteomics revolution:

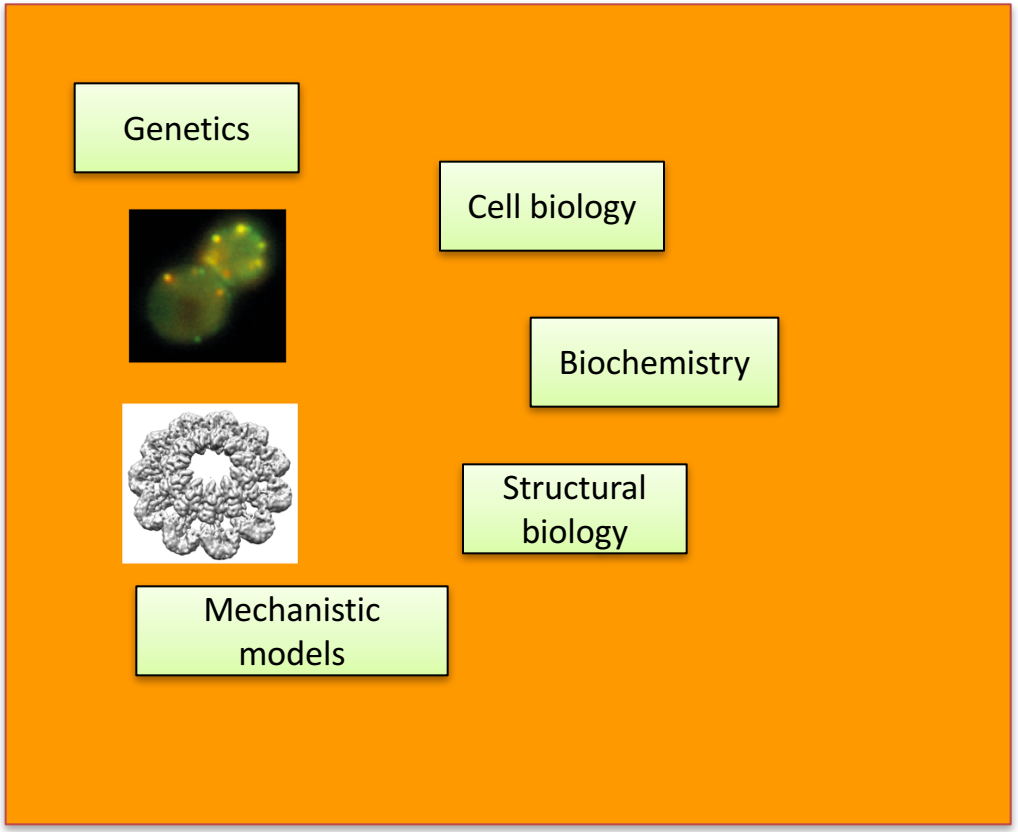
Fast, accurate and highly reproducible quantification of **thousands** of proteomes?

Outline

- The data matrix (proteins vs. samples) as currency of proteomic experimentation
- How do we generate data matrices by mass spectrometry?
- What dimensions should a data matrix have to be maximally useful?
- Can we quantify sources of variation and what are they?
- Can we avoid, recognize and correct artifactual variation?

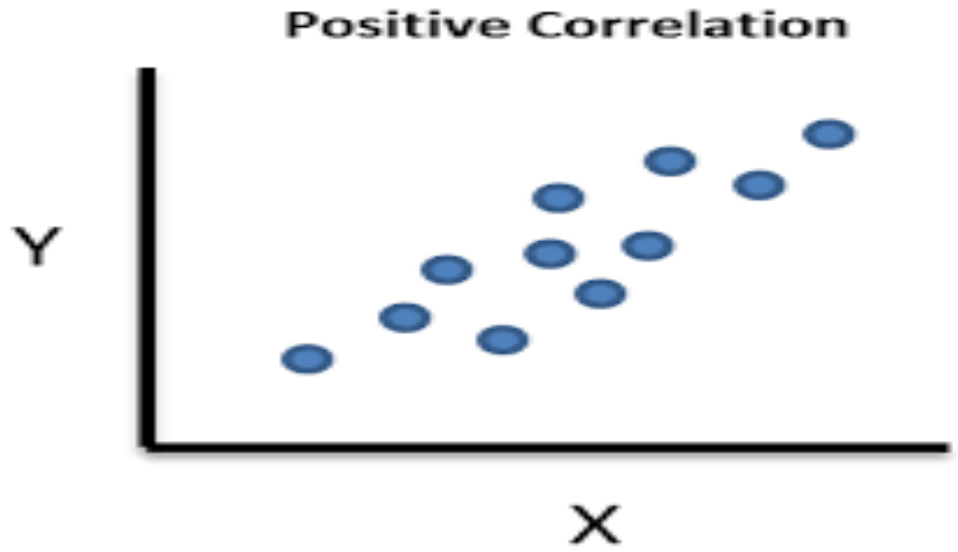
Traditional research approach

Reductionist world: Focus on specific molecules



Discover protein that has a specific function, e.g. growth factor

BigData world: Associations



- Two dependent variables observed under many conditions
- If one is increased the other is increased as well
- Does not imply a causal relationship

- These days we can **collect large proteomic datasets**
- **of dependent variables**

Convergence of research approaches

Big data world: Focus on large datasets

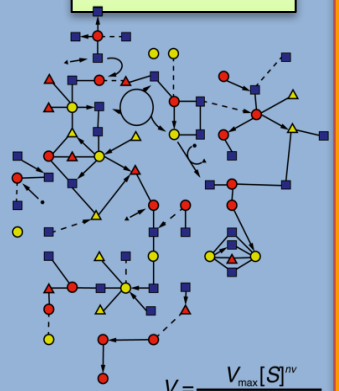
Reductionist world: Focus on specific molecules

Genomics

Transcriptomics

Proteomics

Metabolomics

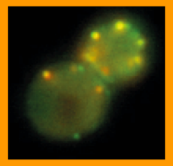


$$V = \frac{V_{max}[S]^{nv}}{V_{max}(K_{0.5}) + [S]^{n}}$$

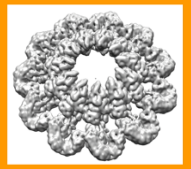
Statistical models

Genetics

Cell biology



Biochemistry



Structural biology

Mechanistic models

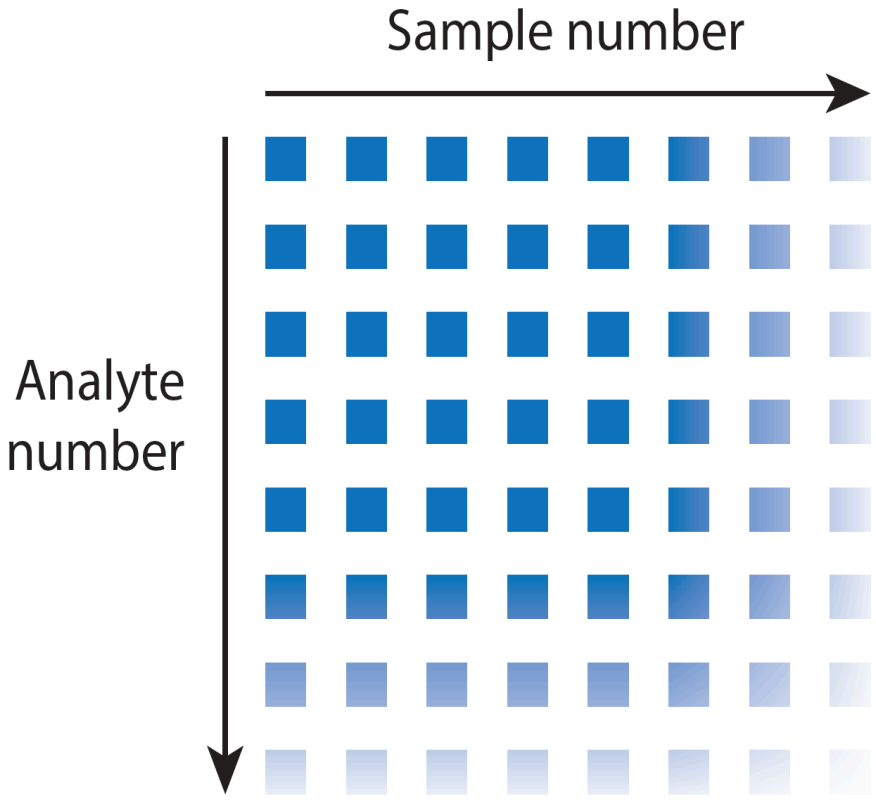
Integrated quantitative models

Learn the behavior of a system from large datasets

Discover protein that has a specific function, e.g. growth factor

Data matrix: The new currency of proteomics data

Data matrix



- Samples are typically derived from perturbed states of a biological specimen
- The data matrix supports correlative analyses, machine learning….

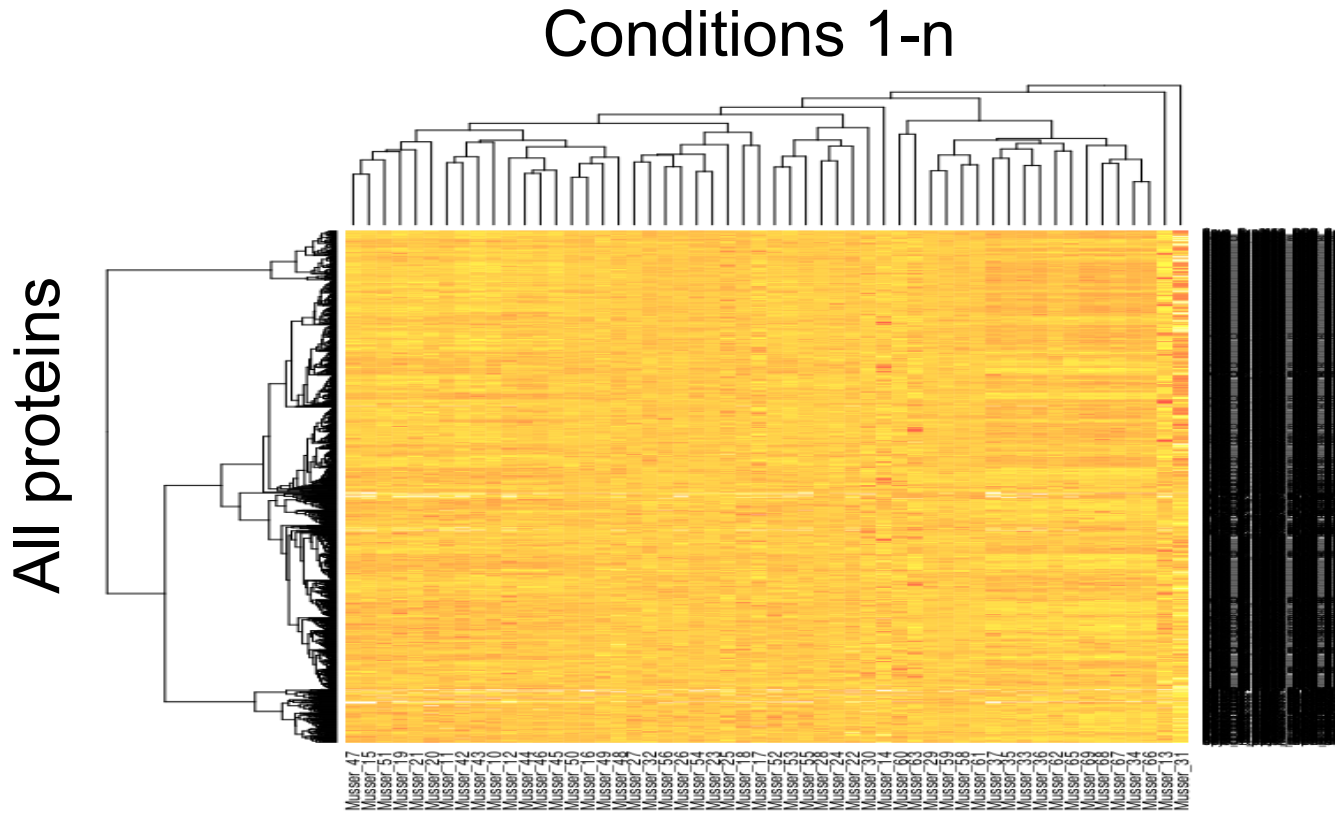
Examples of studies relying on data matrices

- Biomarker studies (control vs. disease)
- (Genome-wide/proteome –wide) association studies
- Identifications of clusters of proteins behaving similarly across conditions
- Network inference
- Population based molecular biology
- Support of mechanistic models

Outline

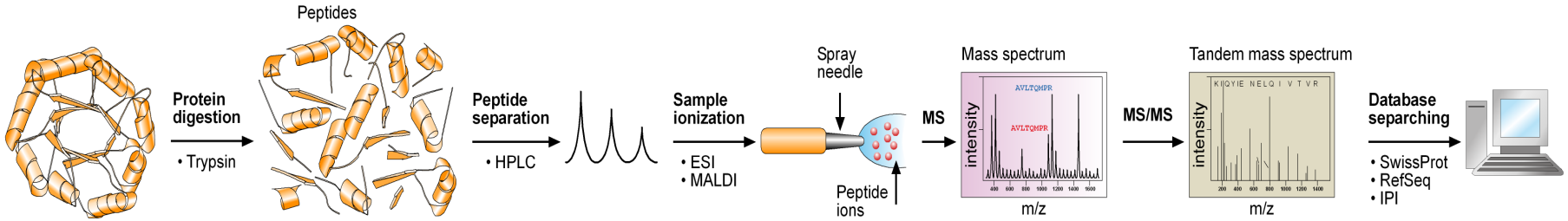
- The data matrix (proteins vs. samples) as currency of proteomic experimentation
- How do we generate data matrices by mass spectrometry?
- What dimensions should a data matrix have to be maximally useful?
- Can we quantify sources of variation and what are they?
- Can we avoid, recognize and correct artifactual variation?

The ideal proteomic data matrix



We would like to quantify all proteins across all conditions but we can't

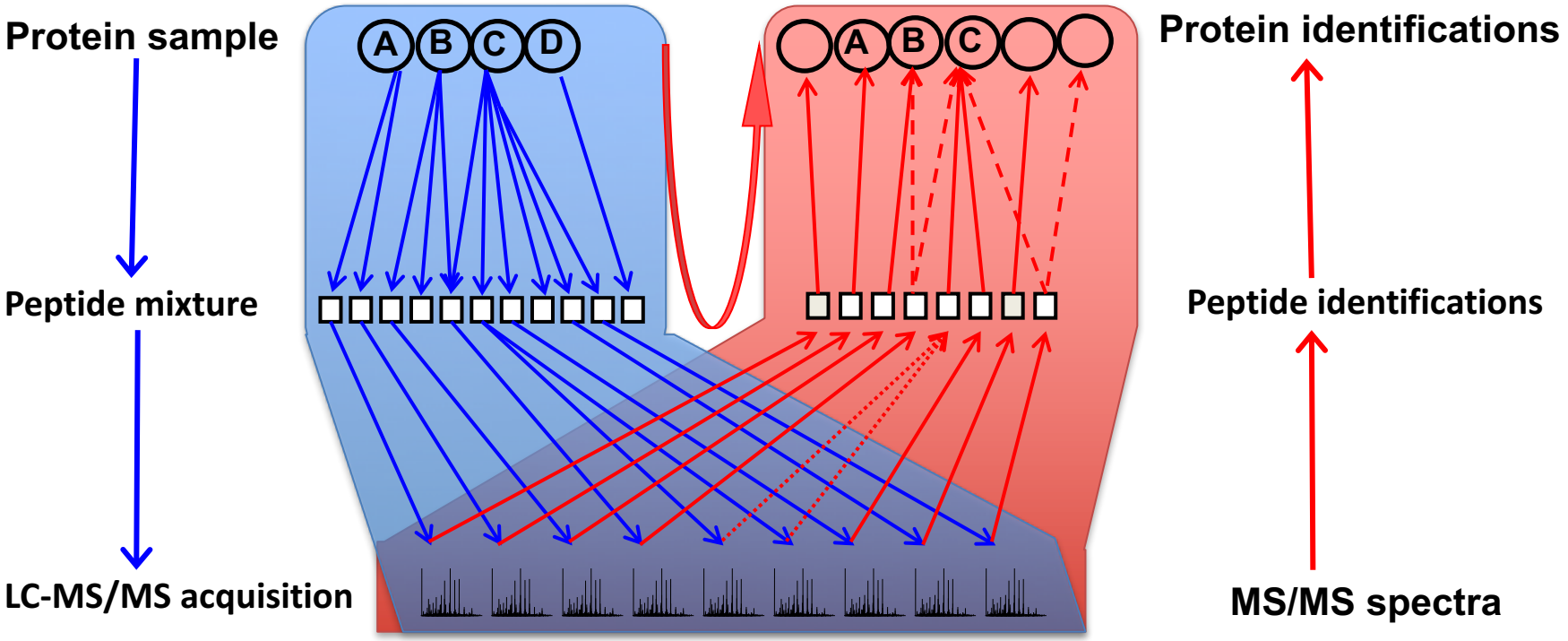
Bottom-up Proteomics



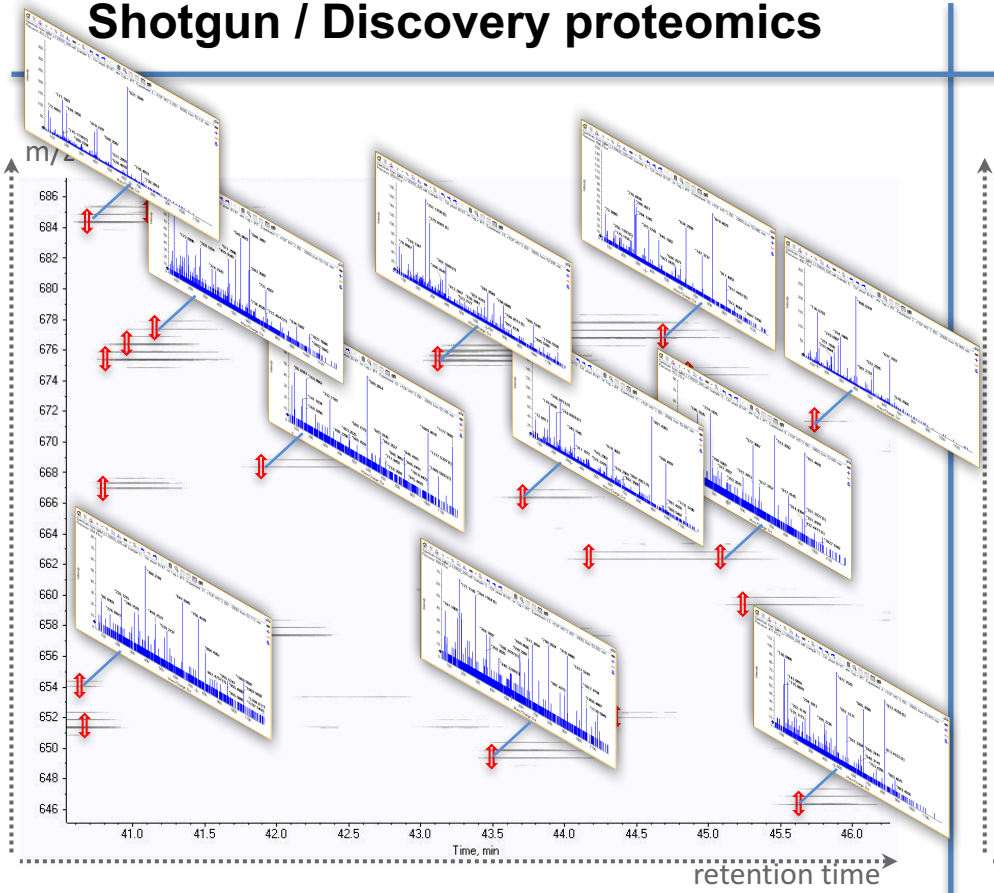
- Proteins are digested to peptides
- Per protein > 100 peptides are generated

Mann and Aebersold Nature, 2003

MS-based proteomics – bottom-up LC-MS/MS in a nutshell



Shotgun / Discovery proteomics



In shotgun:

Each MS2 spectra is a snapshot of a ***different*** peptide

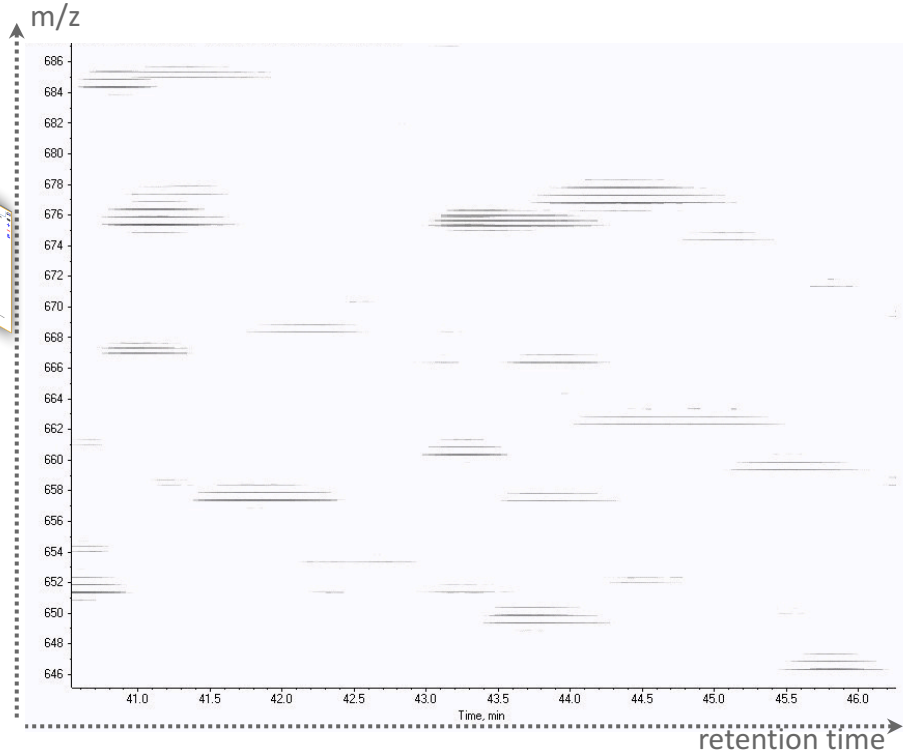
=> Fast (high number of identifications)

=> Not consistent across samples

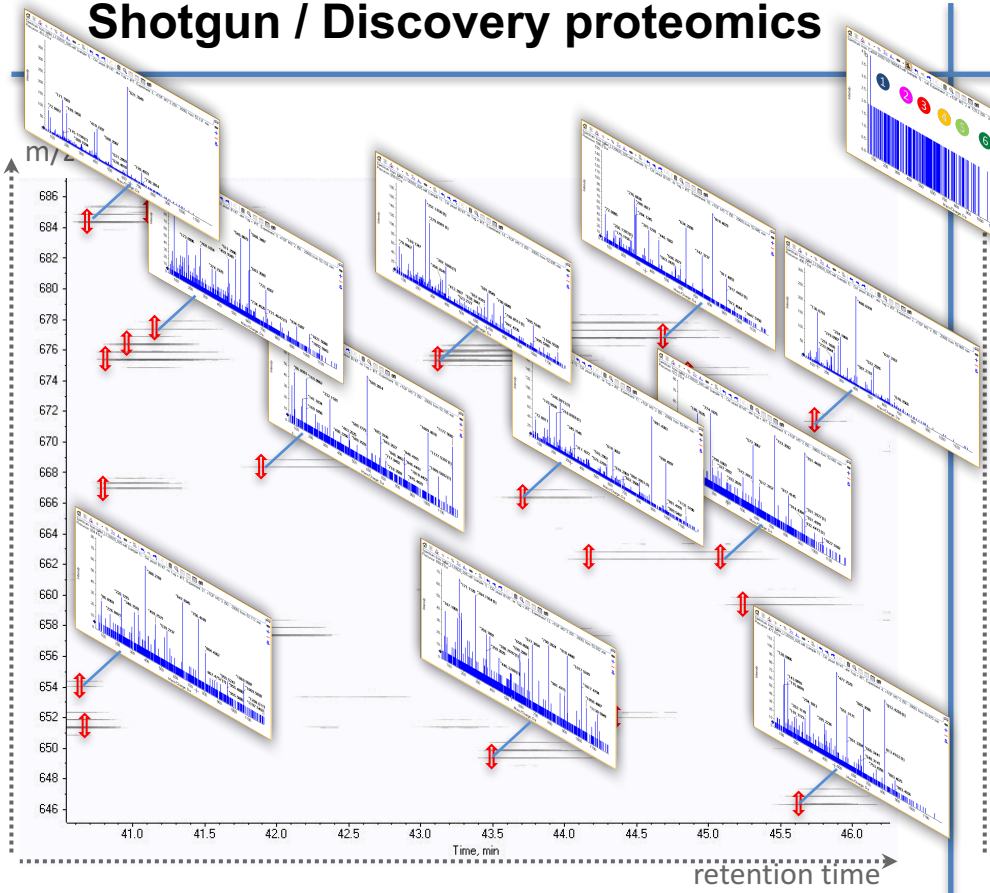
=> Instrument-driven (biased towards high abundant species)

=> Low sensitivity

Targeted (acquisition) proteomics



Shotgun / Discovery proteomics



In shotgun:

Each MS2 spectra is a snapshot of a ***different*** peptide

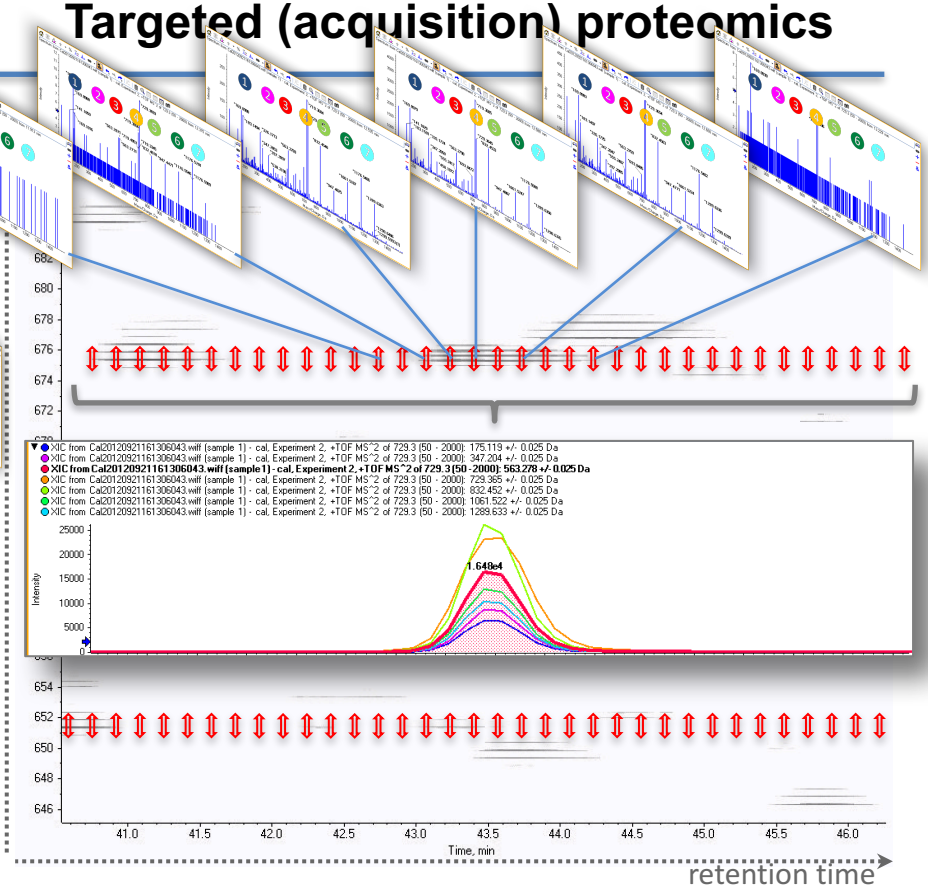
=> Fast (high number of identifications)

=> Not consistent across samples

=> Instrument-driven (biased towards high abundant species)

=> Low sensitivity

Targeted (acquisition) proteomics



In SRM/PRM:

Each MS2 series is a recording of **1** peptide across LC

The fragment ions are acquired over time and are used to reconstitute the elution profile of peptides

=> Consistent + accurate quantification

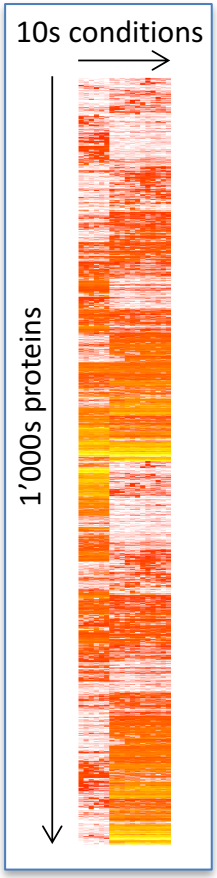
=> User-driven (the machine is forced to select the lower intense signals)

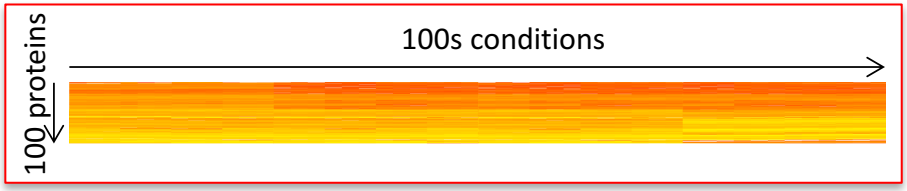
=> High sensitivity

=> Low number of precursors monitored per run

Issue 1 with "standard" LC-MS/MS proteomics strategies

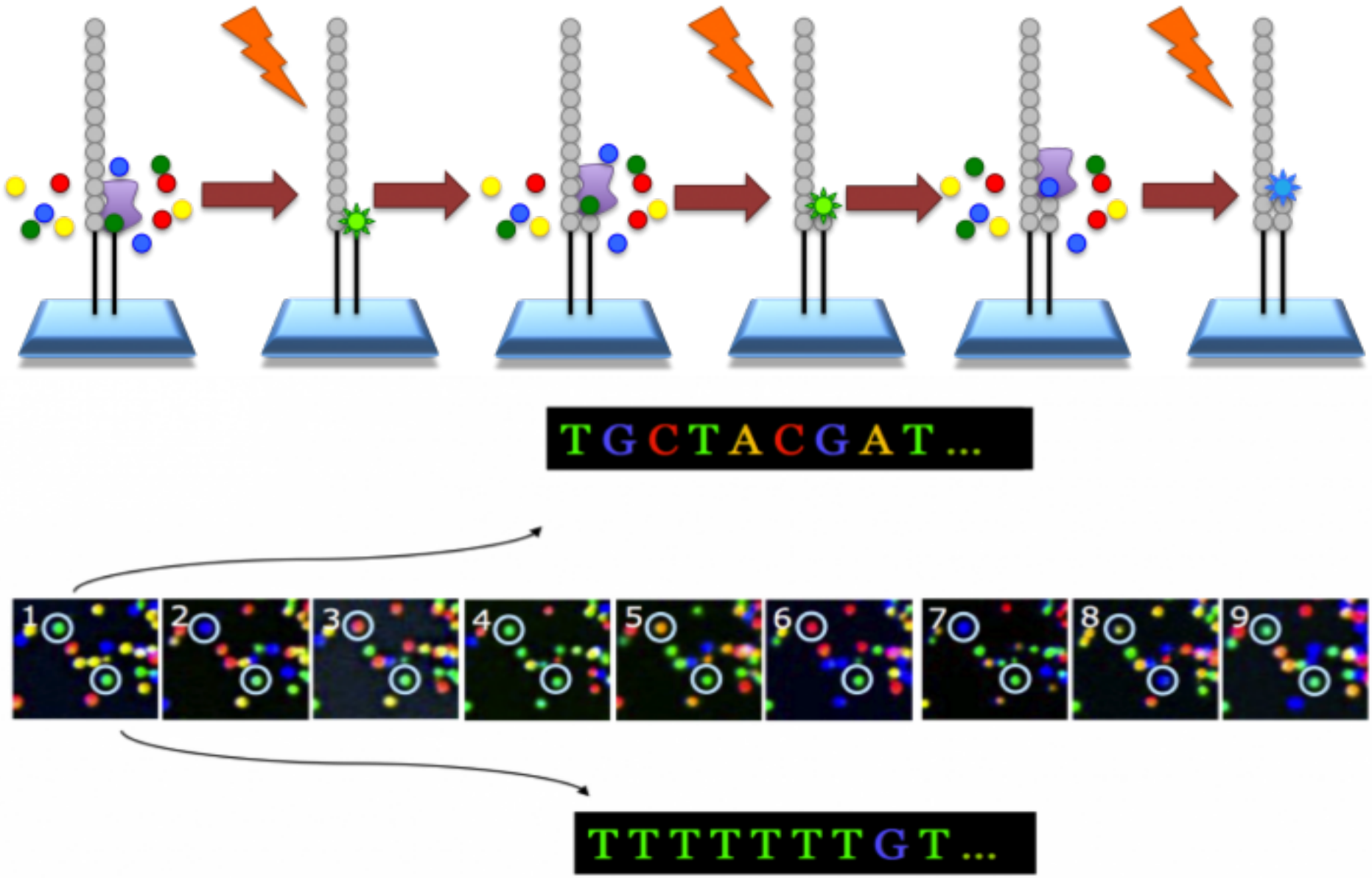
Shotgun / Discovery proteomics		Targeted (acquisition) proteomics
The ideal proteomic method should **quantify large sets of proteins across multiple samples** with:		
- -	consistency	+ +
+ +	rapidity	- -
+	sensitivity	+ +



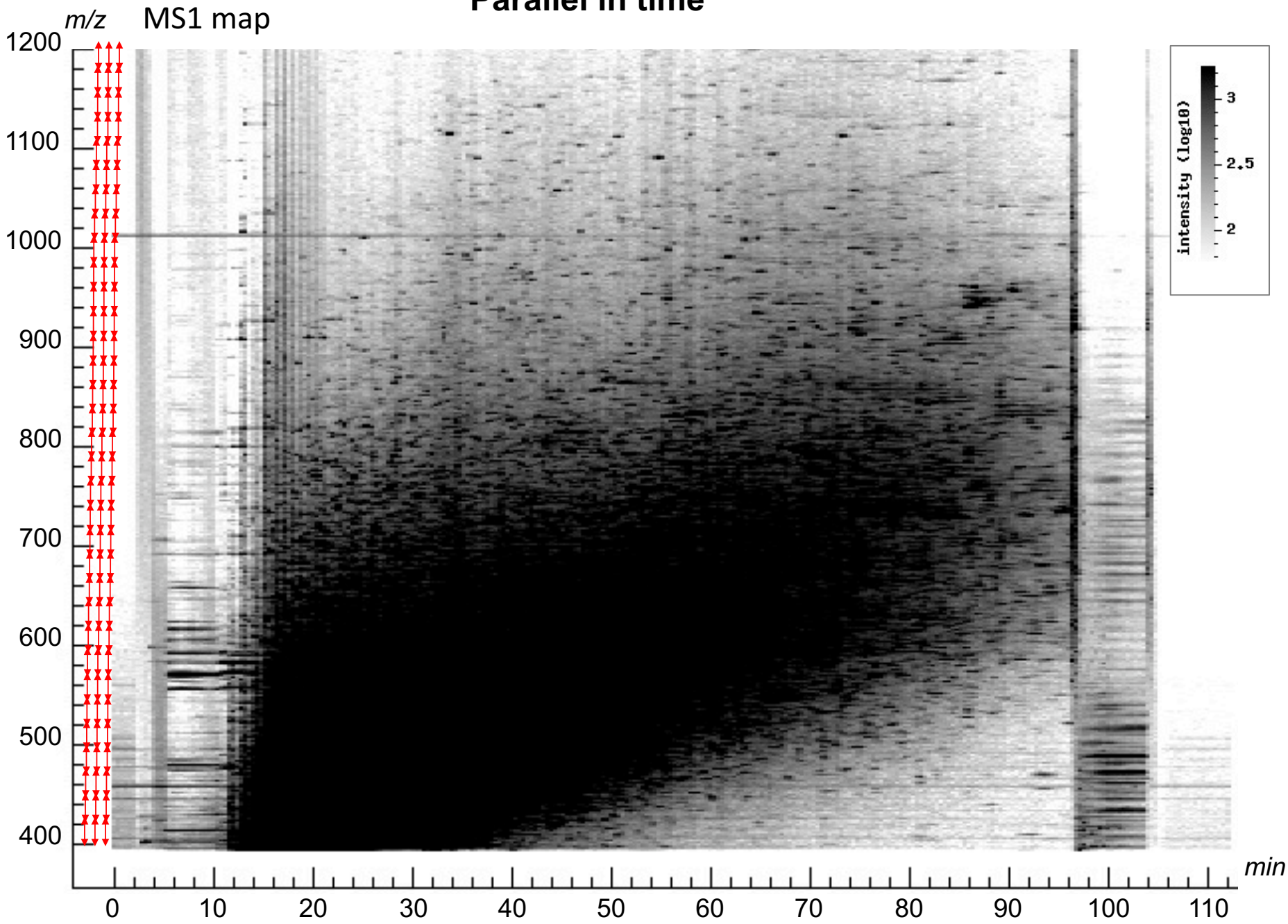


=> need a method to perform consistent acquisition of large number of proteins in large number of samples

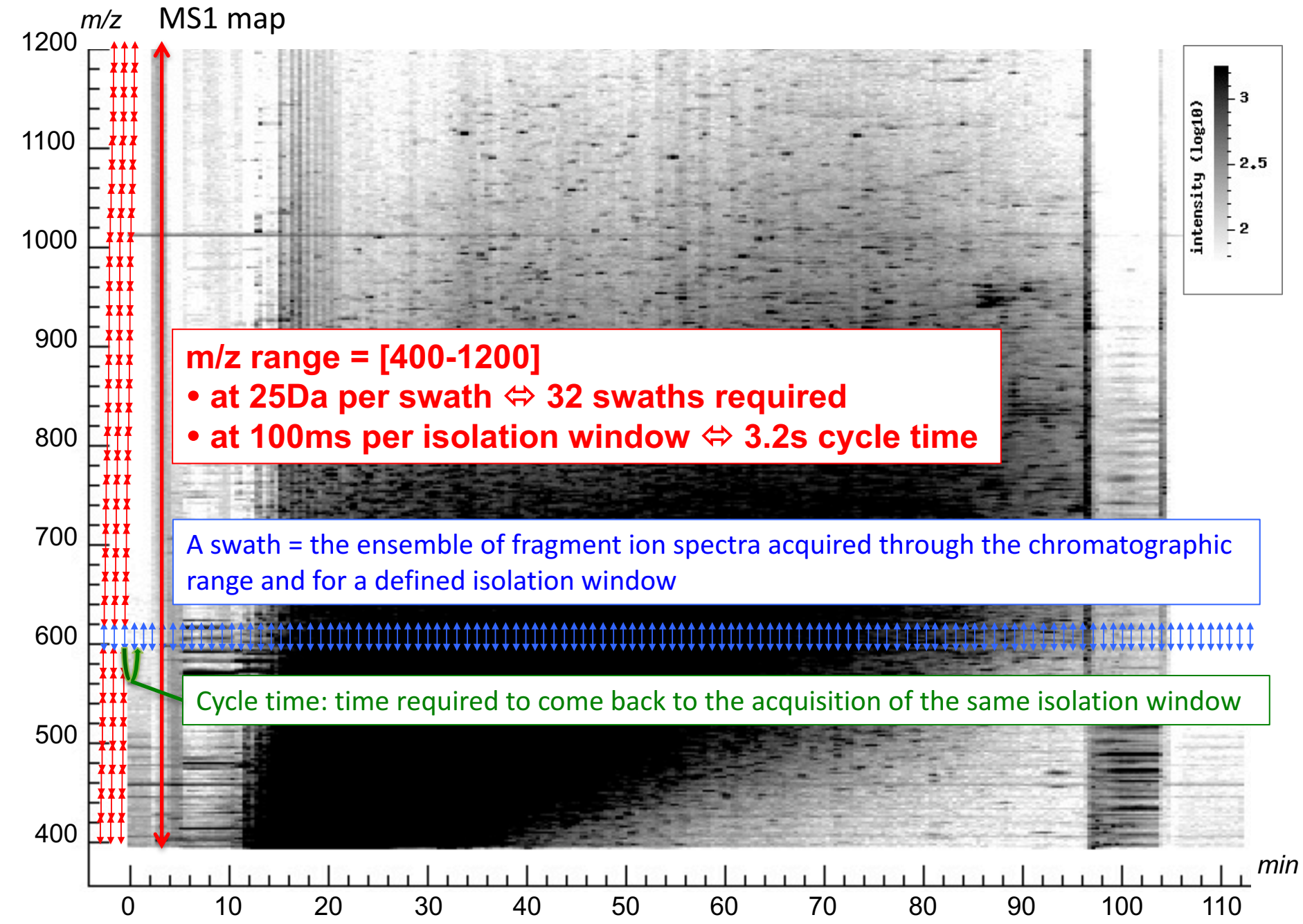
NGS: Highly parallel in space



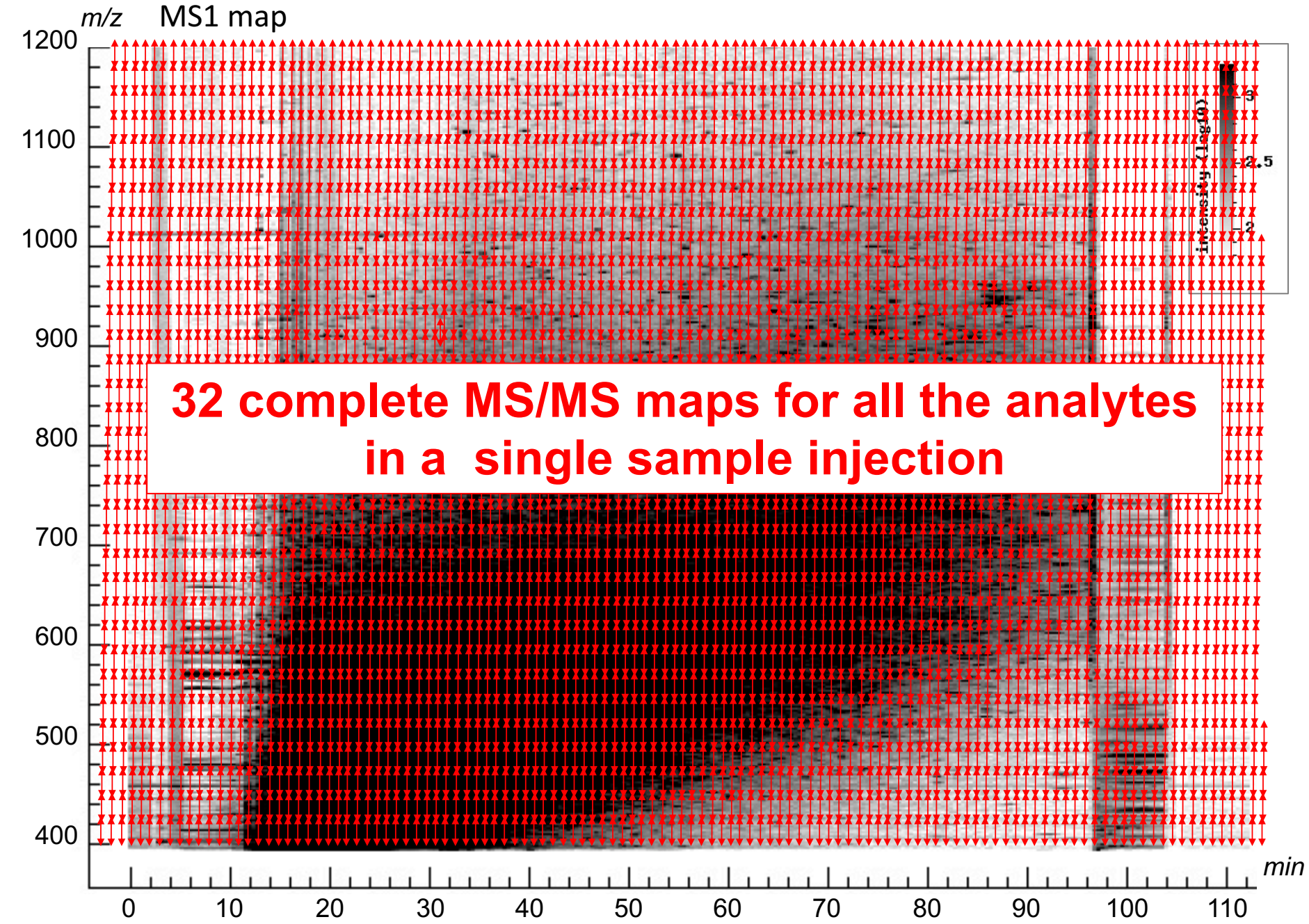
windowed DIA principle: MS/MS acquisition for "all" detectable precursors
Parallel in time



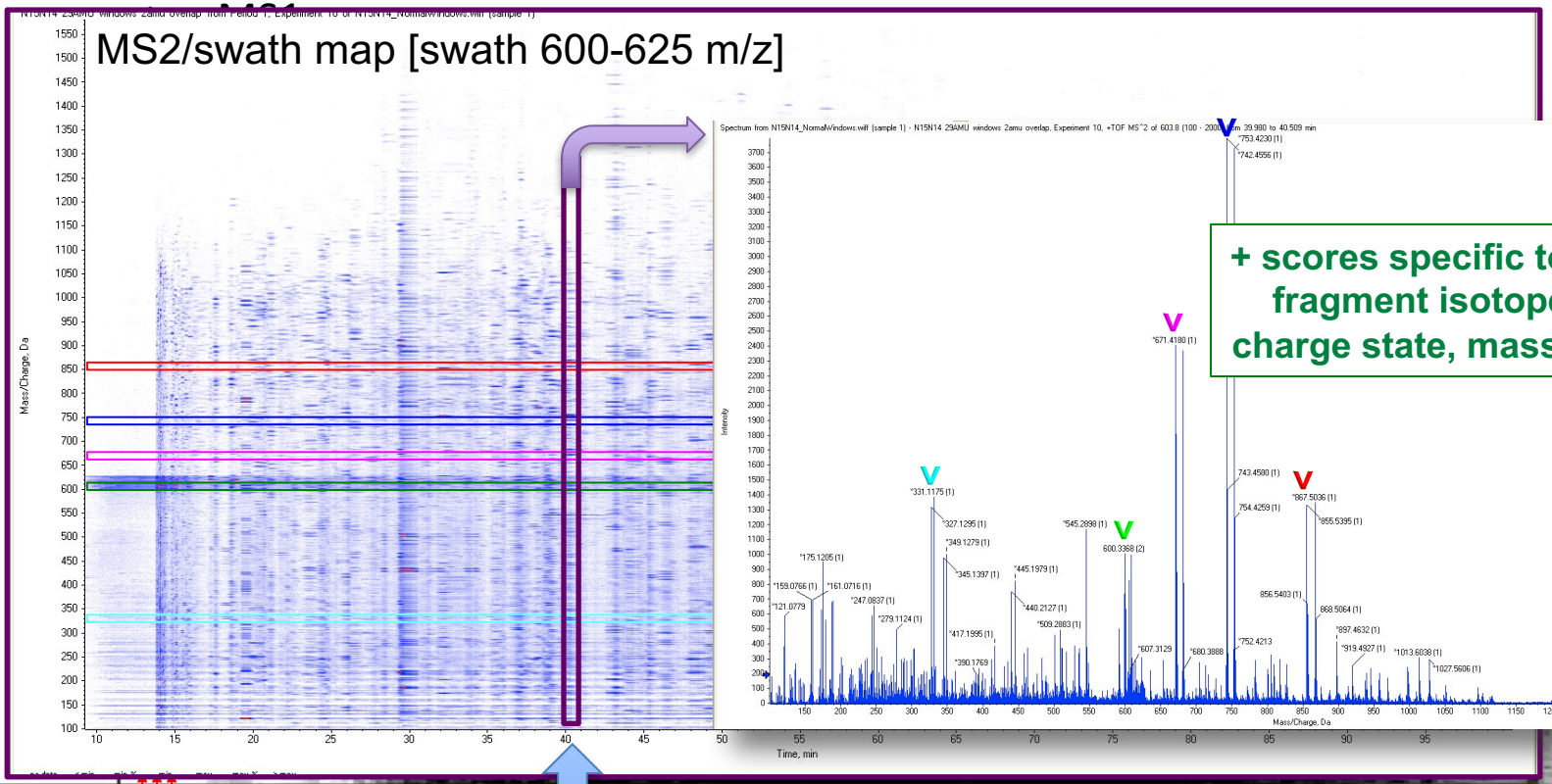
windowed DIA principle: MS/MS acquisition for "all" detectable precursors



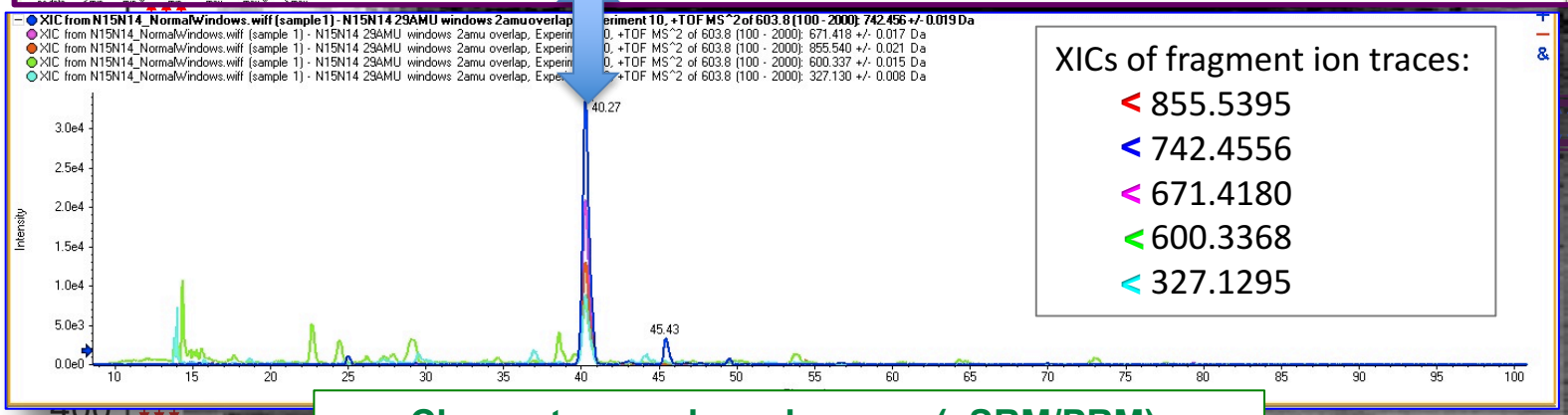
windowed DIA principle: MS/MS acquisition for "all" detectable precursors



SWATH-MS data analysis principle: "peptide-centric" targeted extraction



+ scores specific to high-res. MS2: fragment isotope distribution, charge state, mass accuracy, etc...



XICs of fragment ion traces:

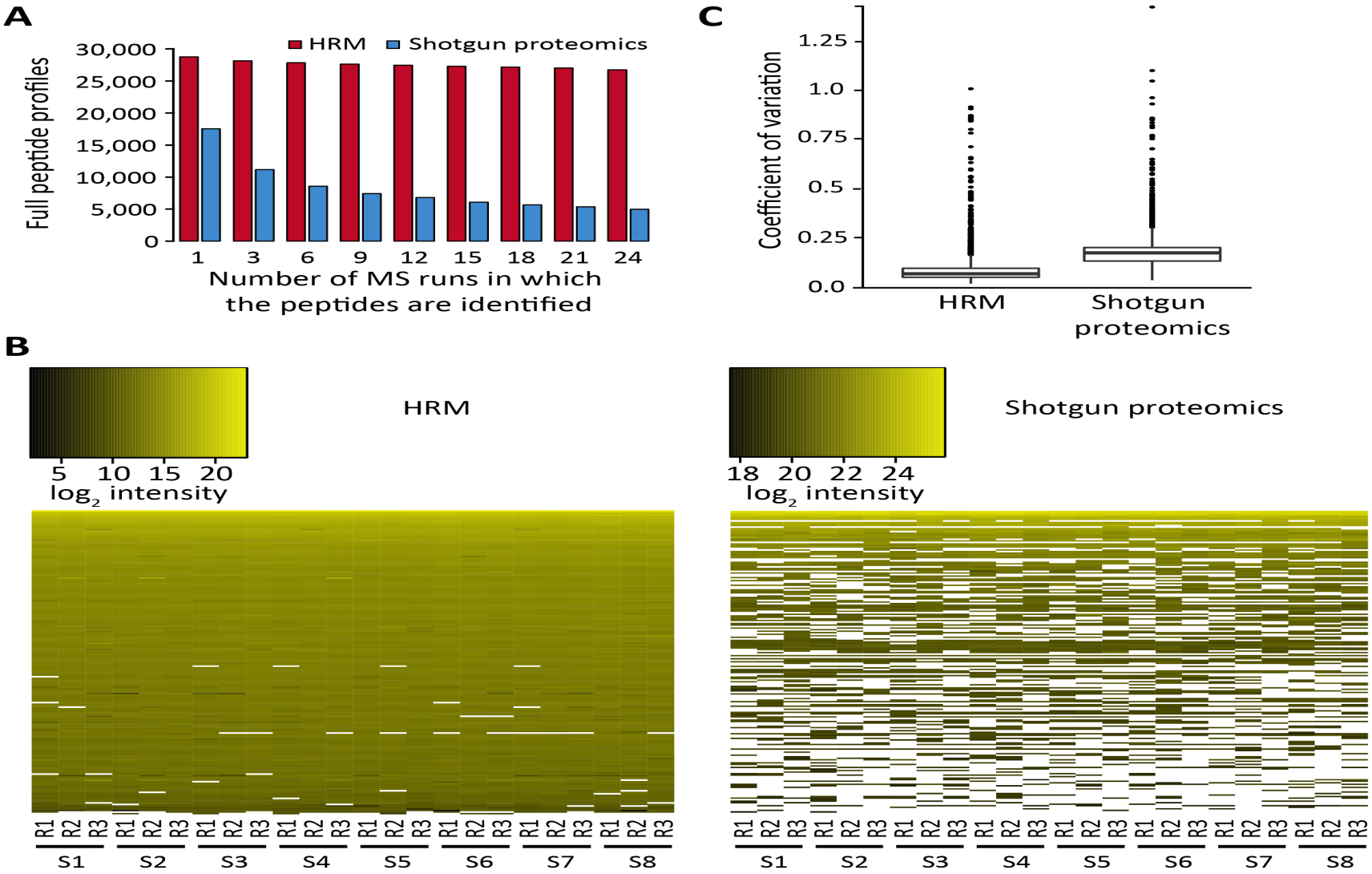
- < 855.5395
- < 742.4556
- < 671.4180
- < 600.3368
- < 327.1295

Chromatogram-based scores (~SRM/PRM): co-elution, peak shape, relative intensities, etc...

mProphet: Reiter et al, Nature Methods (2011)

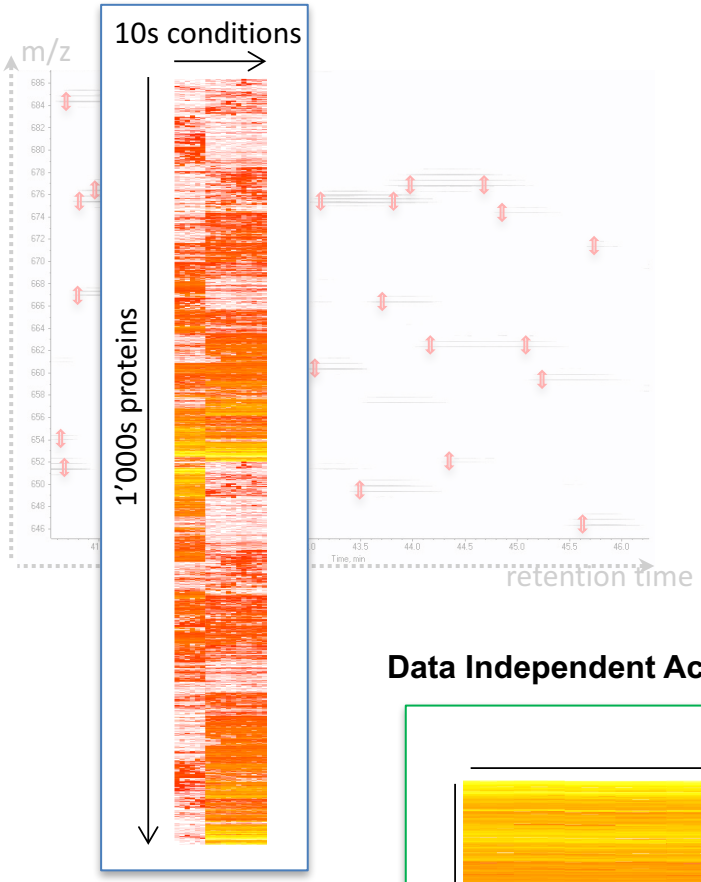
min

….and now with real data

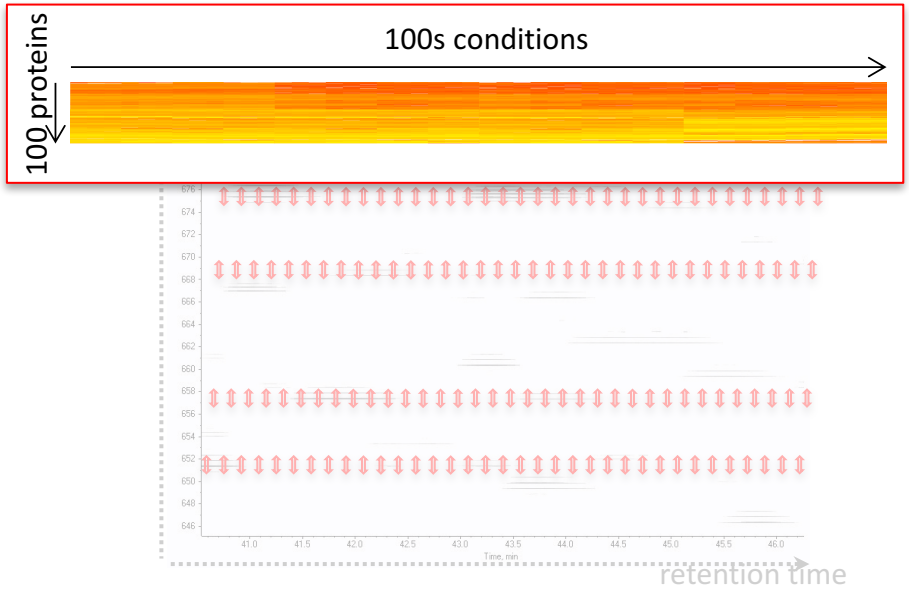


Bruderer et al , MCP 2015

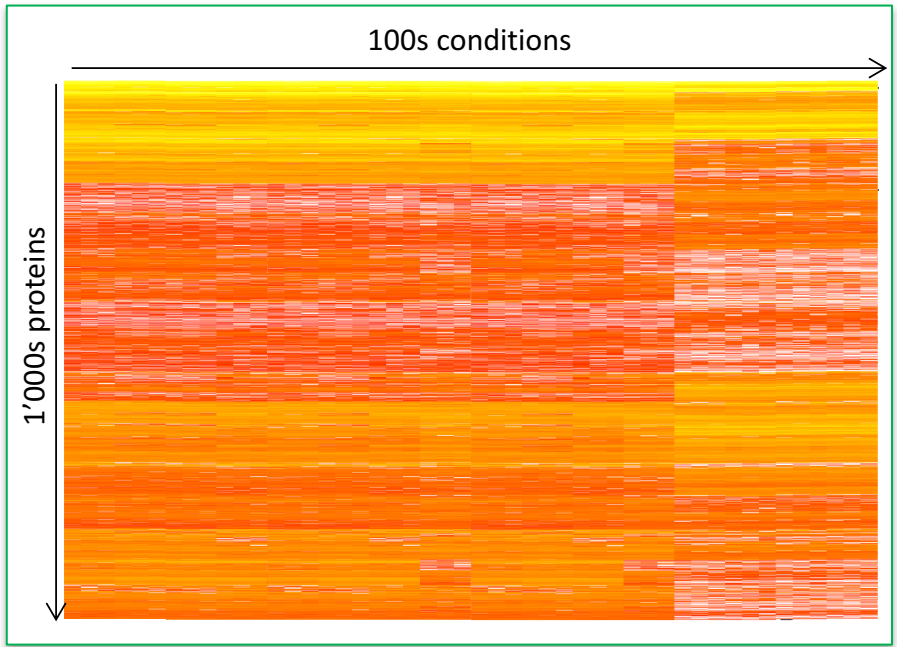
Shotgun / Discovery proteomics



Targeted (acquisition) proteomics



Data Independent Acquisition + Targeted Data Extraction

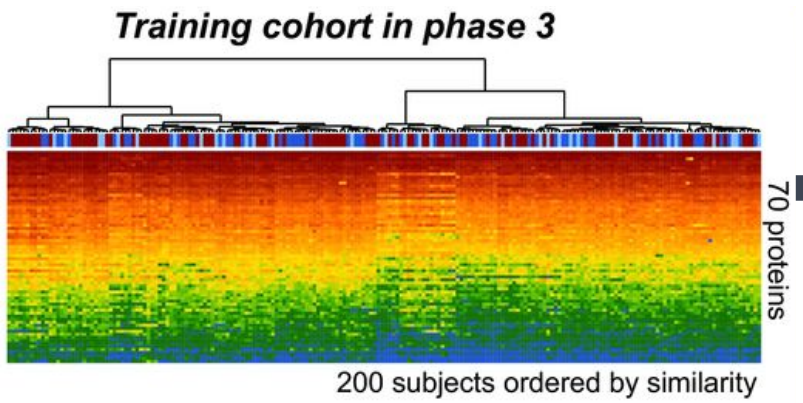


Outline

- The data matrix (proteins vs. samples) as currency of proteomic experimentation
- How do we generate data matrices by mass spectrometry?
- What dimensions should a data matrix have to be maximally useful?
- Can we quantify sources of variation and what are they?
- Can we avoid, recognize and correct artifactual variation?

OPTIMIZATION IN LARGE-SCALE PROTEOMIC EXPERIMENTS

The goal of this work is to determine factors that maximize prediction power, reproducibility and sensitivity of biomarkers



More subjects?

More proteins?

	DDA	SRM/PRM	DIA
Proteomic coverage	High (1000s~10000)	Low (10s~100s)	Medium (1000s)
Sample throughput	Low (10~100s, even one sample)	High (100s~1000s)	Medium (100s)
Sensitivity	Low	High	Medium

Ting Huang , in prep

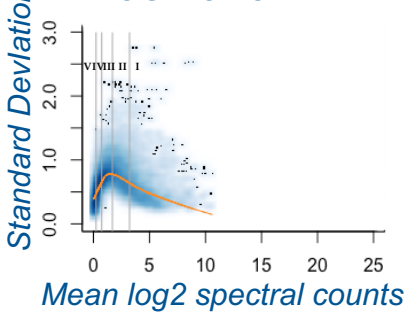
Aebersold and Mann, *Nature*, 537, 347–355, 2016.

DATASETS FOR EXPERIMENTS

Quantified by DDA, SRM and SWATH

- TCGA-CRC samples quantified by DDA (spectral counts)
  - ✓ Zhang, B. et al. (2014). Nature 513, 382-387;
- CRC plasma samples quantized by SRM
  - ✓ Surinova, S. et al. (2015). EMBO mol. Med., 7, 1166-1178;
- NCI-60 cell lines quantified by SWATH
  - ✓ Manuscript in preparation;

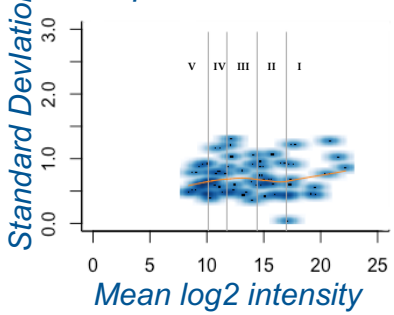
TCGA CRC DDA



7211 proteins

95 colorectal samples

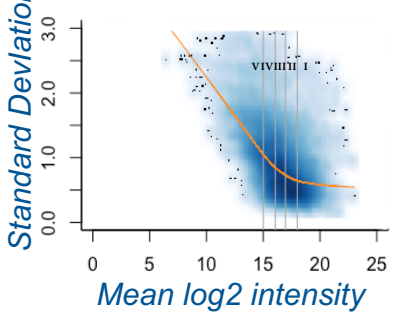
CRC plasma SRM



70 proteins

200 CRC plasma samples

NCI-60 SWATH



3717 proteins

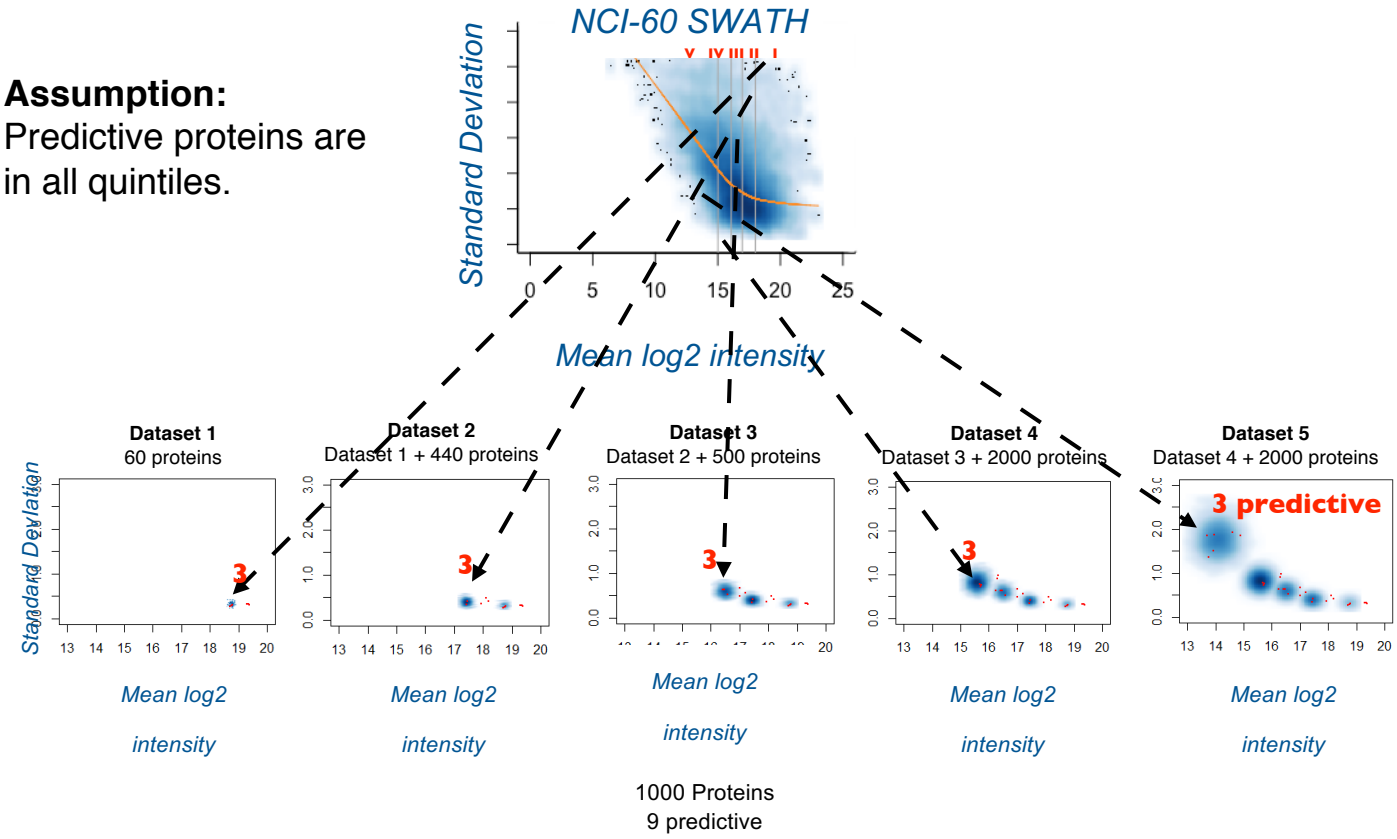
60 Cell lines

Variance depends on protein intensity

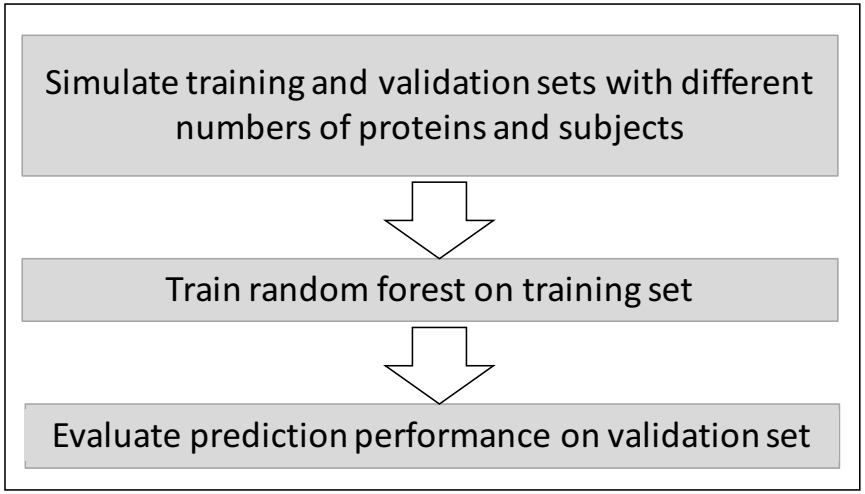
AN EXAMPLE OF SIMULATING DATASETS

Simulate datasets with different protein number and 60 samples

Assumption:
Predictive proteins are in all quintiles.



DESIGN OF SIMULATION EXPERIMENTS

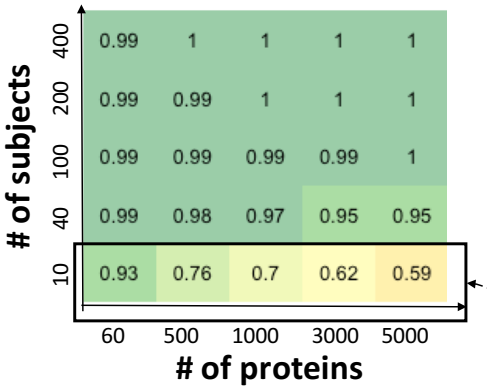


# of simulations: 100

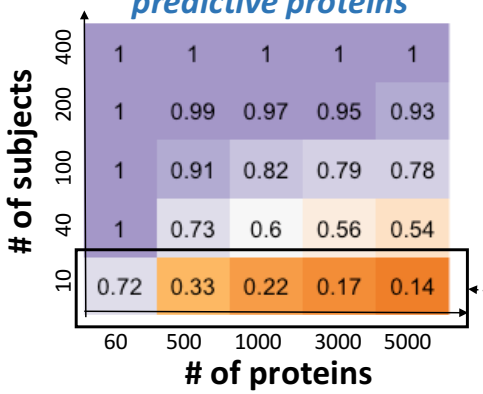
# of proteins: 60, 500, 1000, 3000, 5000
# of biological replicates: 10, 40, 100, 200, 400
Distribution of predictive proteins: high, low, all

EVALUATE PREDICTION PERFORMANCE ON VALIDATION SET

Mean predictive accuracy



Mean sensitivity of identifying predictive proteins



Accuracy = $\frac{\text{\# of correctly classified subjects (TP+TN)}}{\text{\# of subjects}}$

TP+TN	9.3	7.6	7.0	6.2	5.9
# of subjects	10	10	10	10	10
Mean accuracy	0.93	0.76	0.7	0.62	0.59

Sensitivity = $\frac{\text{\# of correctly identified predictive proteins } (TP_p)}{\text{\# of predictive proteins}}$

TP_p	3.6	2.97	2.64	2.38	2.1
# of predictive proteins	5	9	12	14	15
Mean sensitivity	0.72	0.33	0.22	0.17	0.14

CONCLUSION FROM SIMULATION EXPERIMENTS

- More proteins decreased predictive accuracy and sensitivity of identifying predictive proteins.
- More subjects increased accuracy and sensitivity.
- Less technical variability increases performance
- Conclusion: Experimental technologies should optimize for throughput of biological samples, low technical variability and high reproducibility

Outline

- The data matrix (proteins vs. samples) as currency of proteomic experimentation
- How do we generate data matrices by mass spectrometry?
- What dimensions should a data matrix have to be maximally useful?
- Can we quantify sources of variation and what are they?
- Can we **avoid, recognize and correct** artifactual variation?

PubMed says a lot about the plasma proteome:

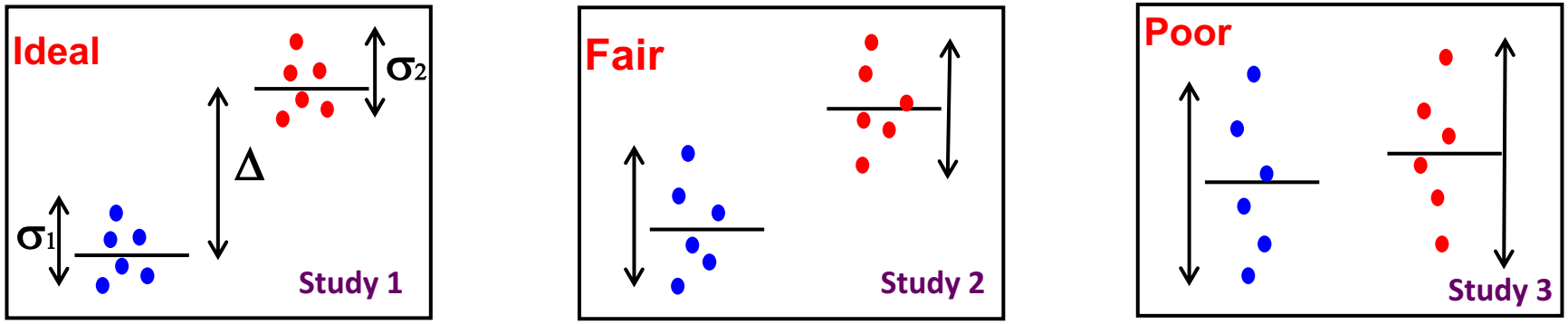
- Protein Biomarker: 476.000 papers
- Plasma protein: > 1.5 Mio papers
- Plasma proteome: > 1500 papers
- Plasma biomarker: > 46.000 papers

However, we do not know….

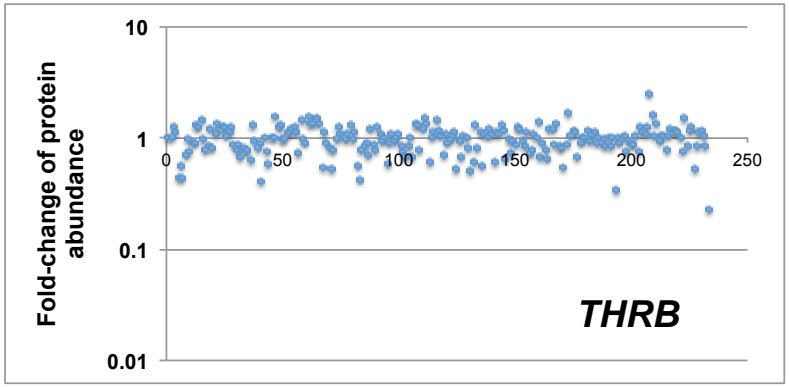
- Variability of protein abundance in a population
- Source of variability (environmental or inherited)
- Variability as a function of time

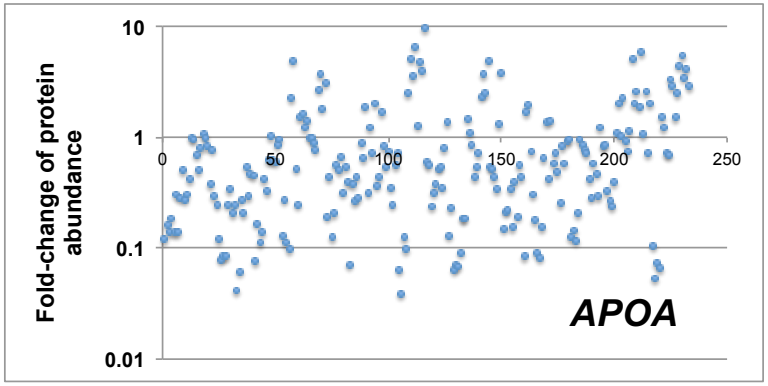
Variability of plasma proteins is important for biomarker studies

Conc. of one plasma protein



Question: Which plasma proteins are more variable in the human population?

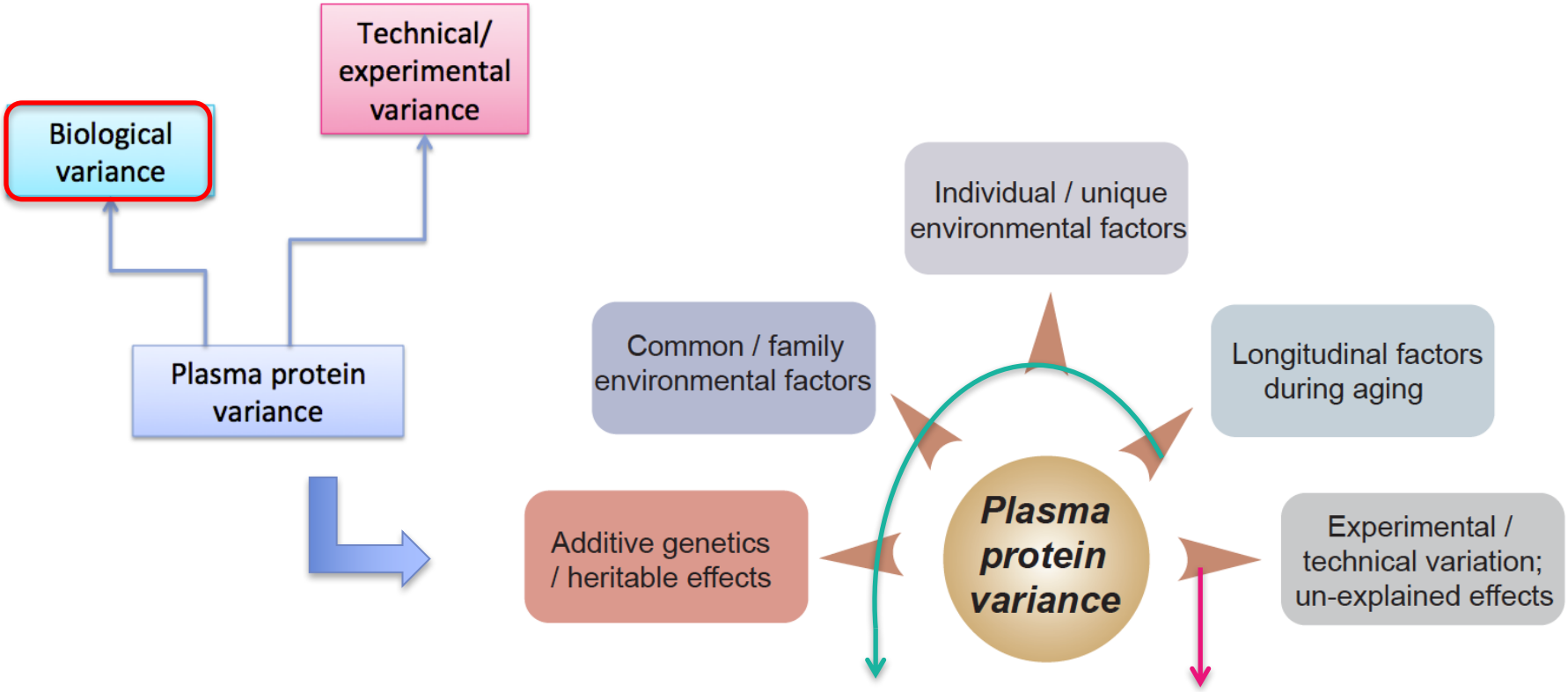




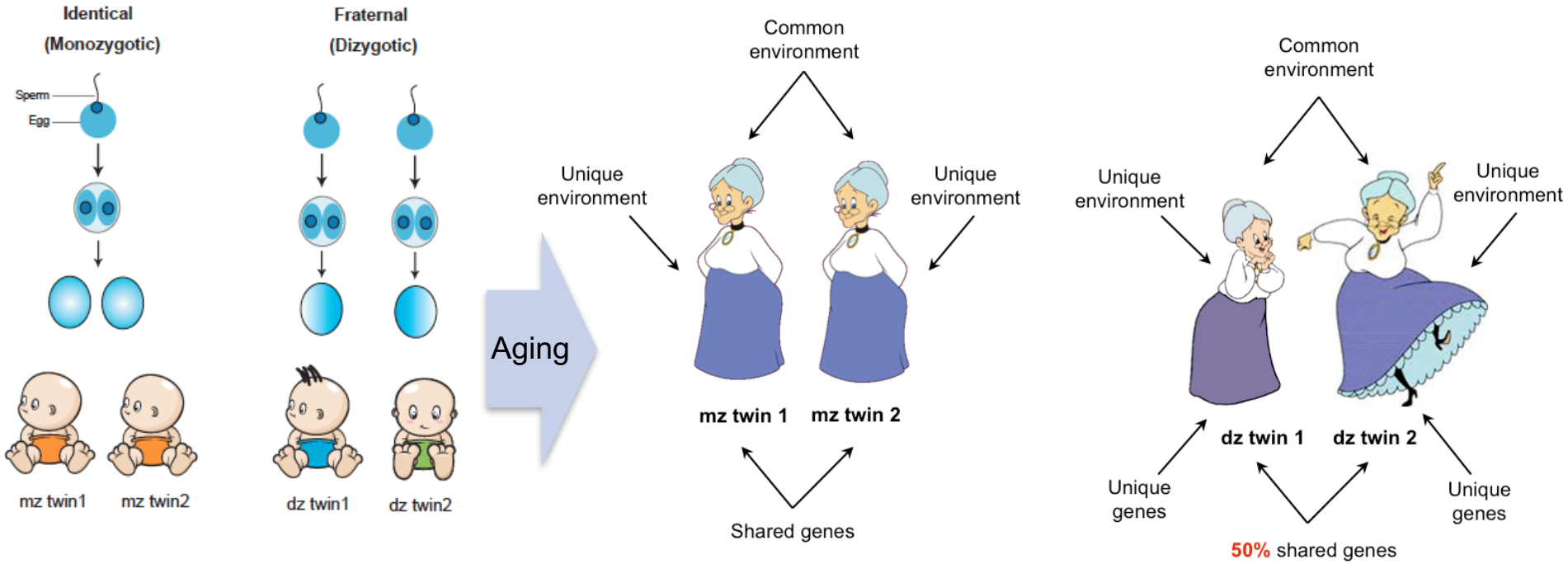
Liu Y et al, *MSB*

Origins of the variability

Question: Which plasma proteins are more variable in the human population?

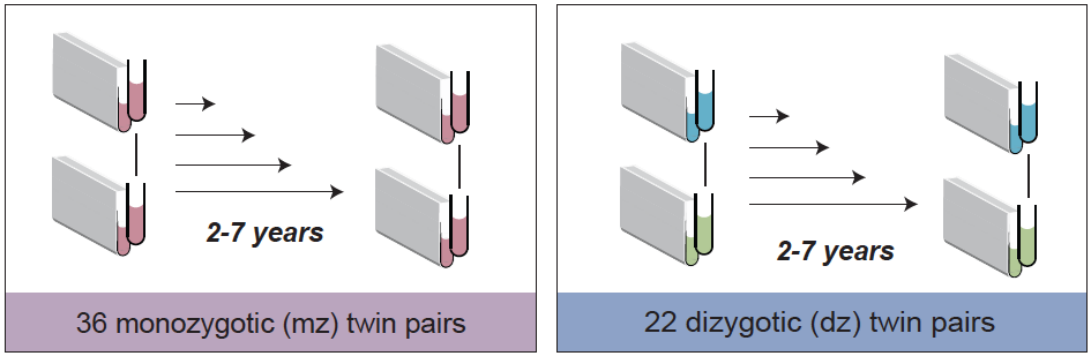


Twins: the natural experiment for human genetic analysis



Experimental design

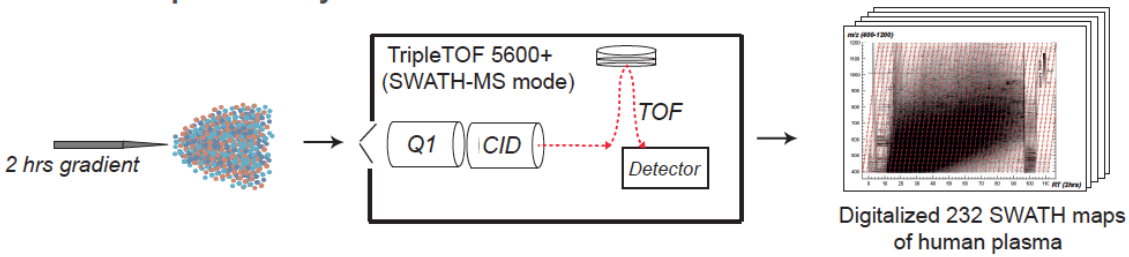
Plasma Sample Collection



Total sample no.
= (36+22) x 2 x 2=232;

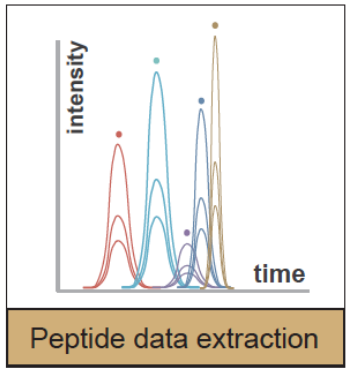
38-78 years old females

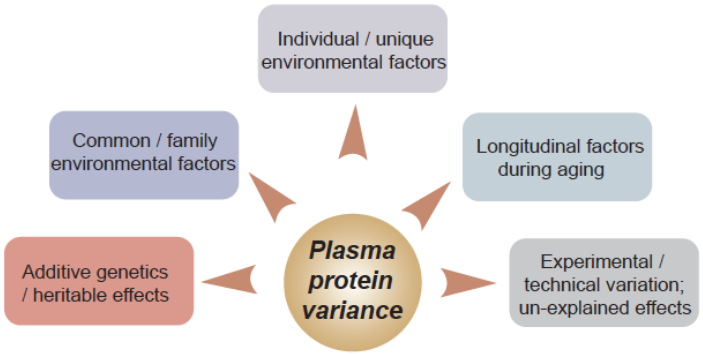
DIA Mass Spectrometry



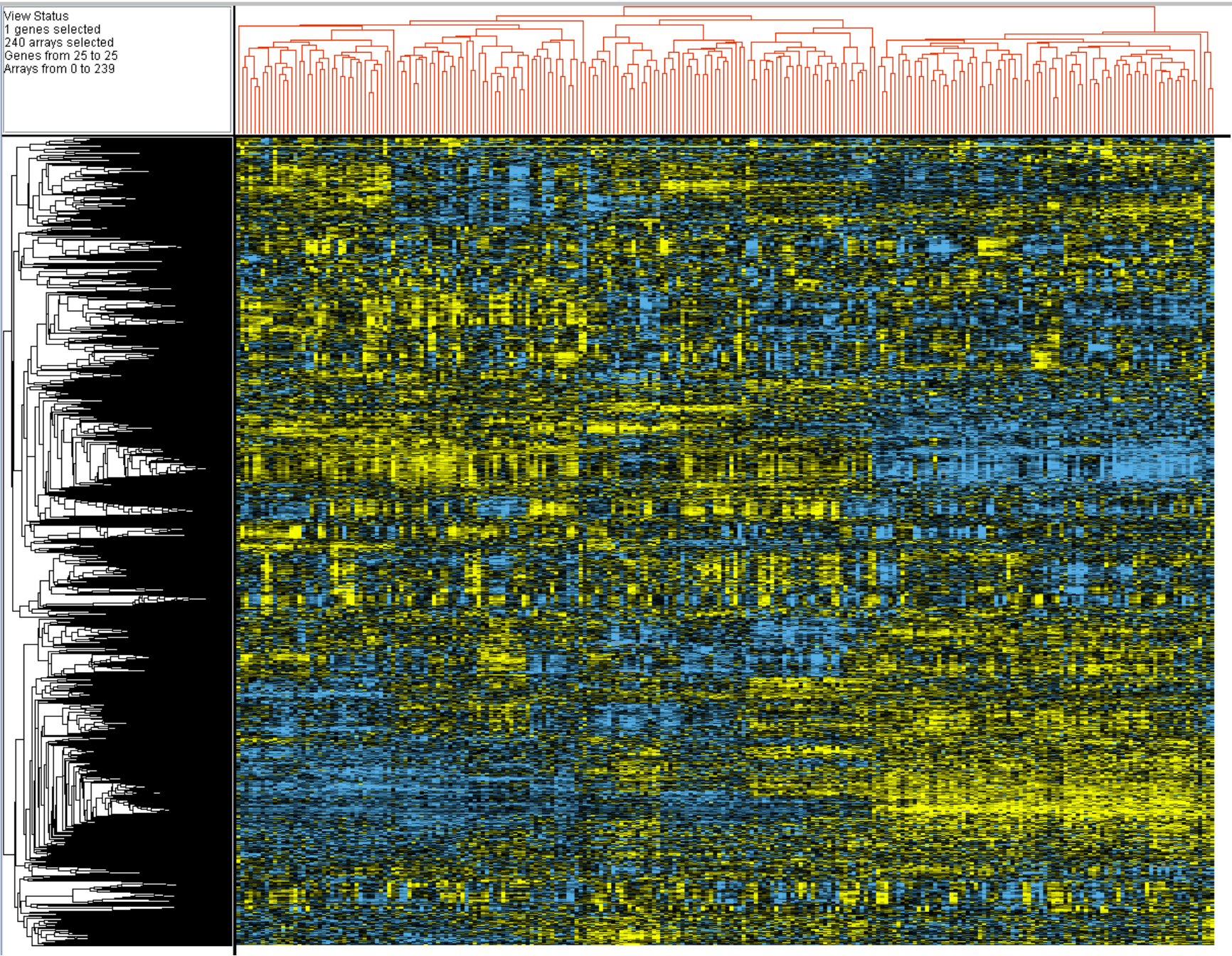
Digitalized 232 SWATH maps of human plasma

Protein I.D. & variance decomposition

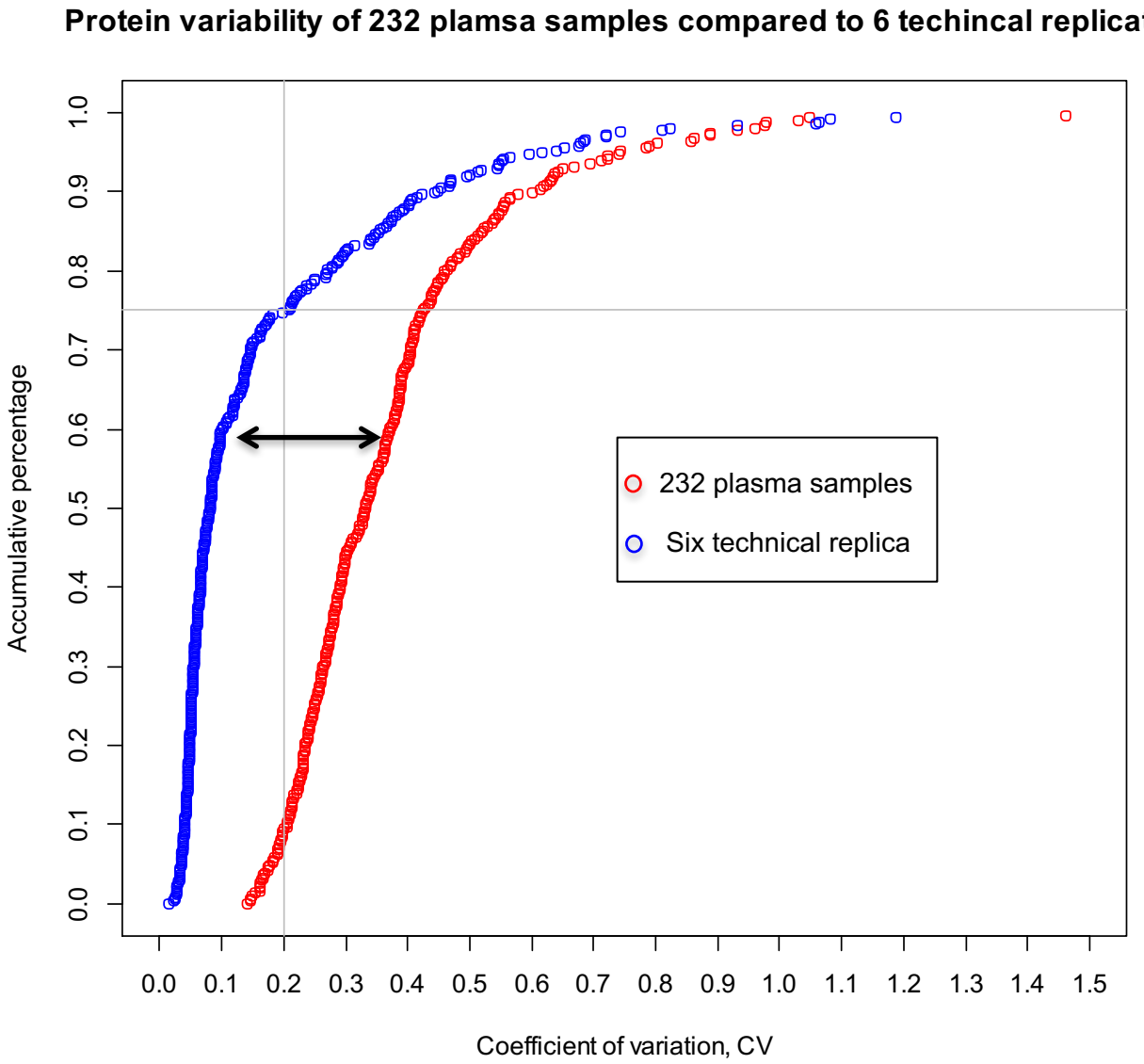




Data Matrix

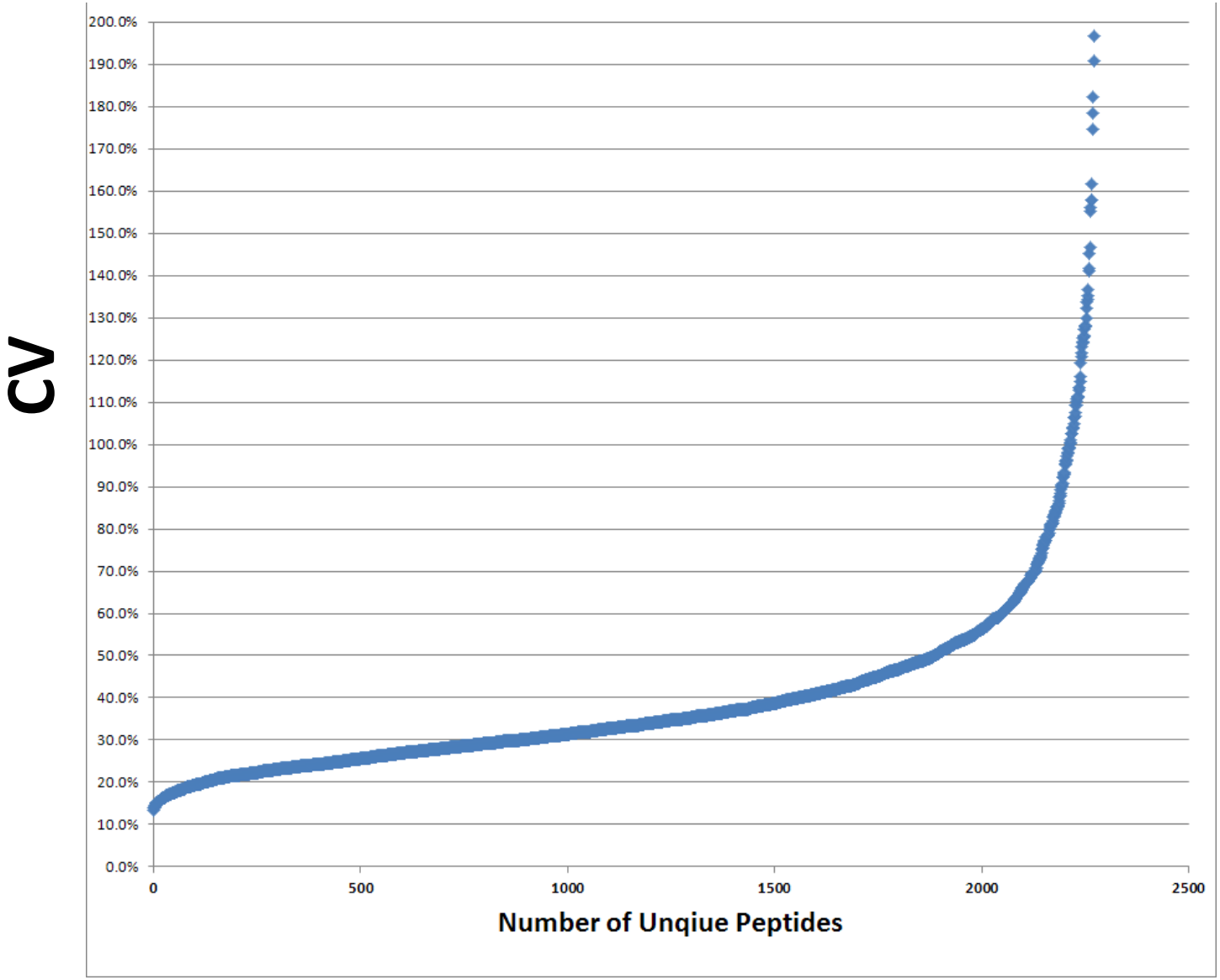

View Status
1 genes selected
240 arrays selected
Genes from 25 to 25
Arrays from 0 to 239
3197 peptides

Validation: Biological vs. technical variability



Based on Top3 peptides per protein

Peptide level variabilities vary greatly



HERITABILITY DETERMINATION

We model the phenotype using **Linear Mixed Models**

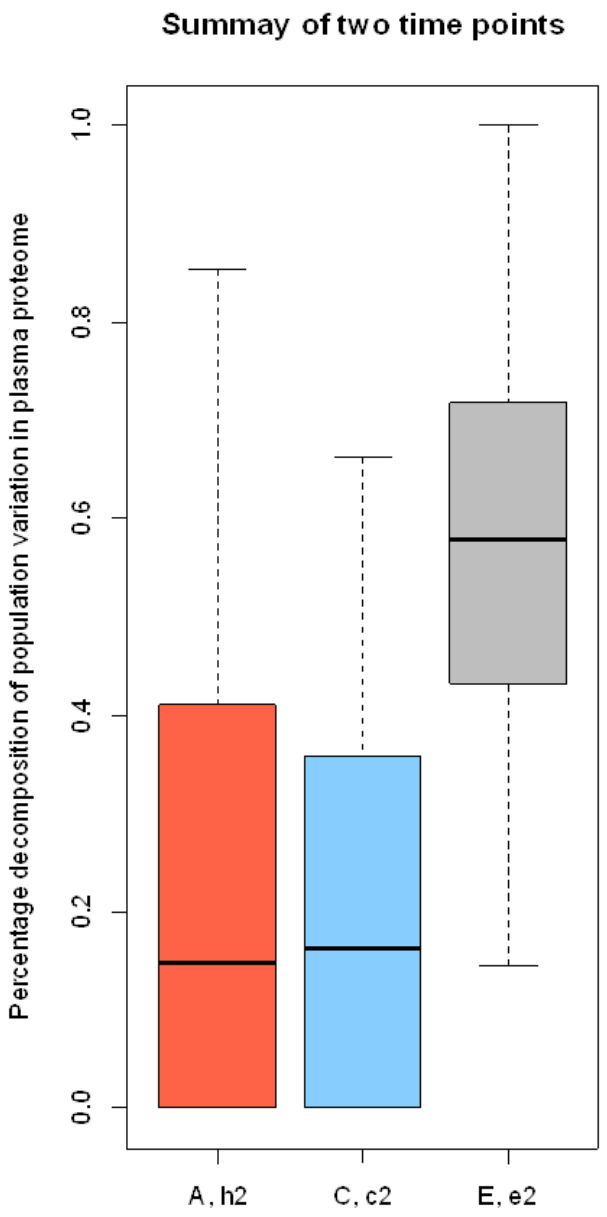
$$y_i = g_i + c_i + e_i$$

- y→ phenotype of interest (protein amount)
- g→ effect of genes (A)
- c→ effect of common environment between the twins (C)
- e→ effect of individual environment (E)

Then, the variance of phenotype can be split into three components; **heritability** is the proportion of variance of the phenotype that is due to the effect of genes:

$$\sigma_y^2 = \sigma_g^2 + \sigma_c^2 + \sigma_e^2$$

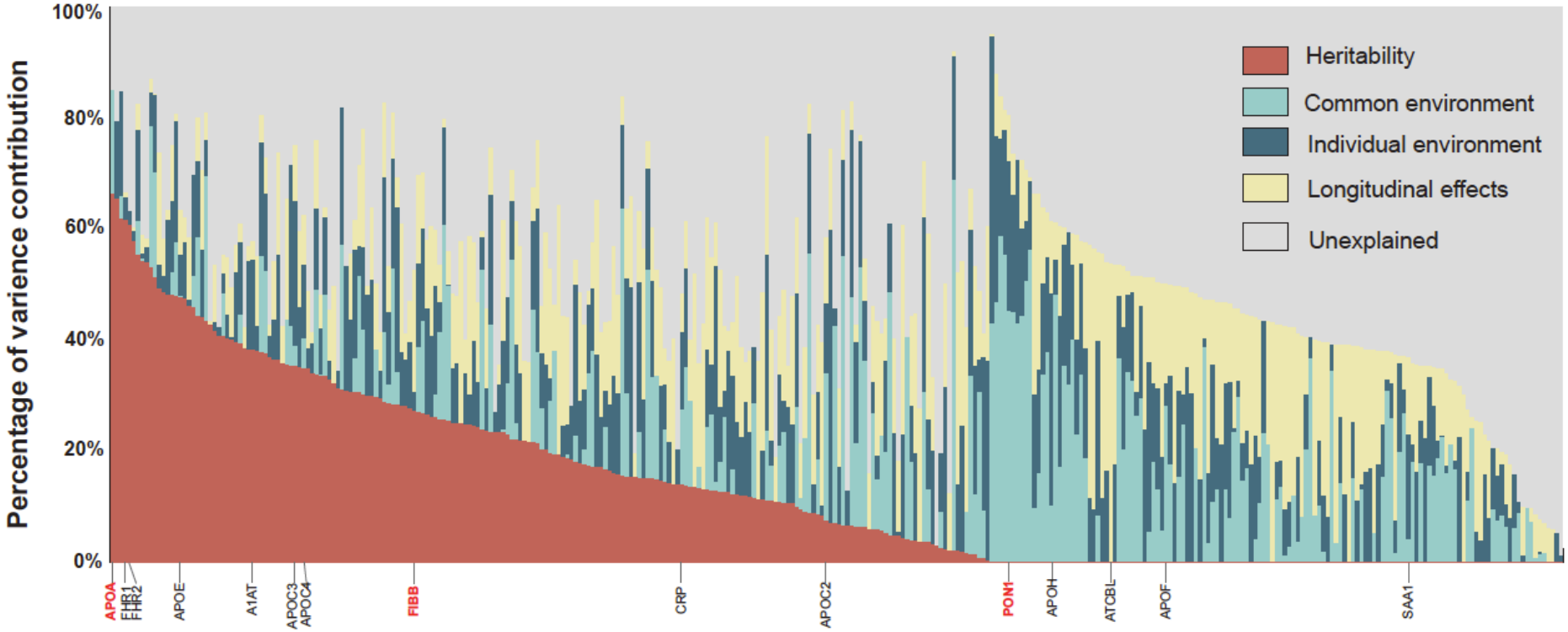
$$h^2 = \sigma_g^2 / \sigma_y^2$$



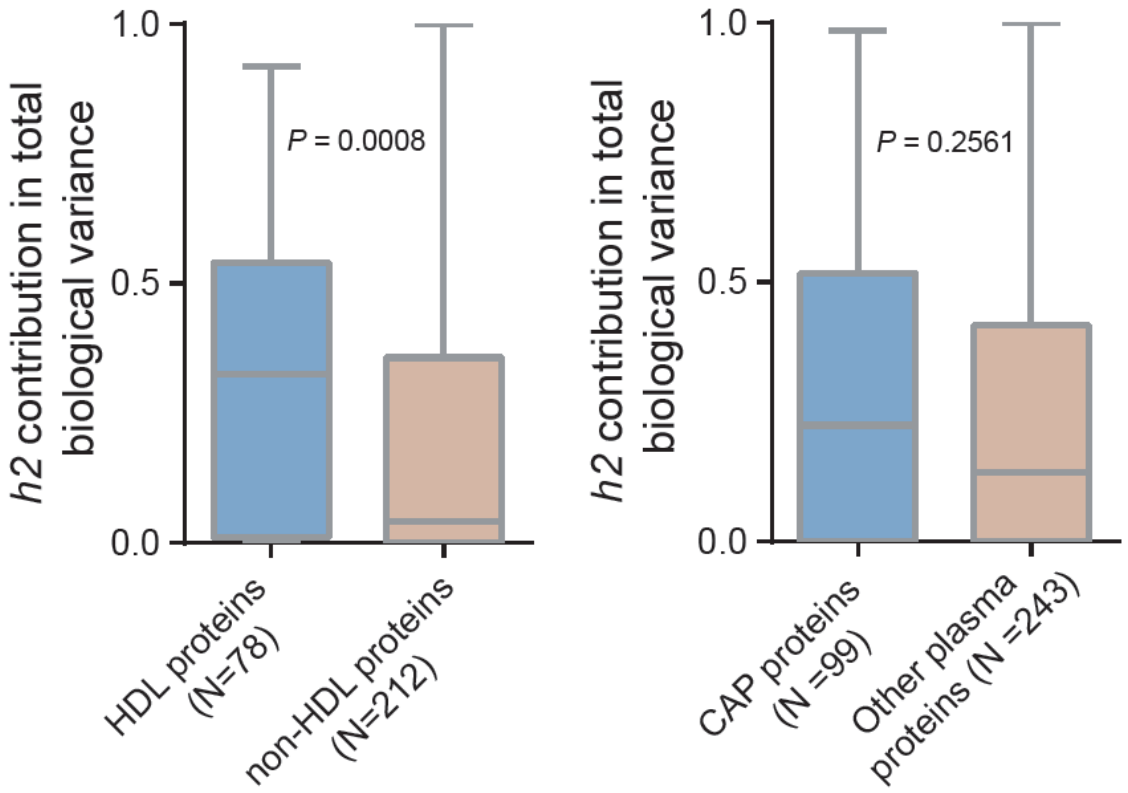
SOLAR version 6.5.8 (Almasy and Blangero 1998 AJHG 62 (5) 1198-1211

Dissection of variance components in plasma proteome

$$\sigma_y^2 = \sigma_g^2 + \sigma_c^2 + \sigma_e^2 + \sigma_w^2 + \sigma_\varepsilon^2$$



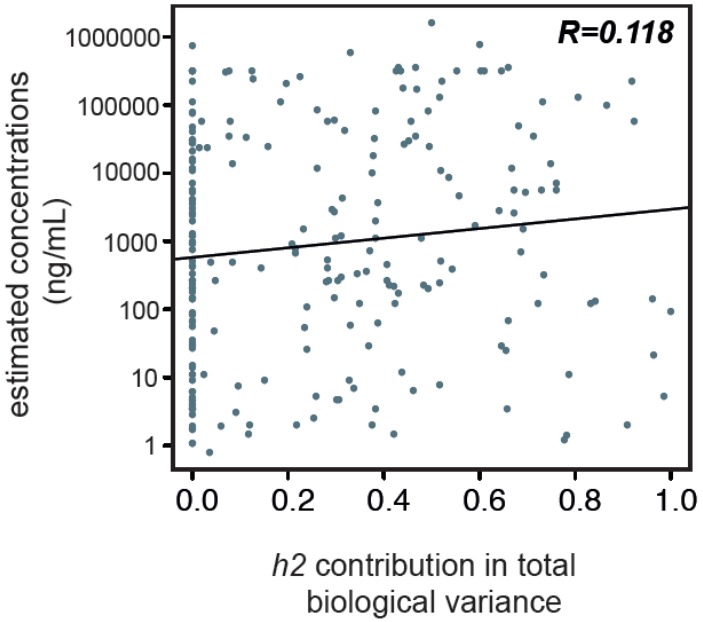
Biological insights and indicators from the results

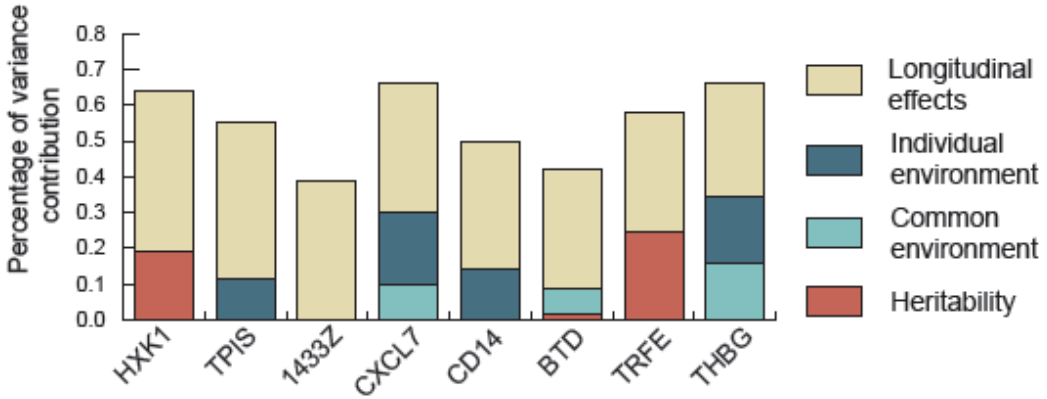


Shah AS *et al,* J Lipid Res. (2013)

Hüttenhain R et al, Sci Transl Med. (2012)

Insights and indicators from the results





Patel, K. et al. Int J Cancer (2011).

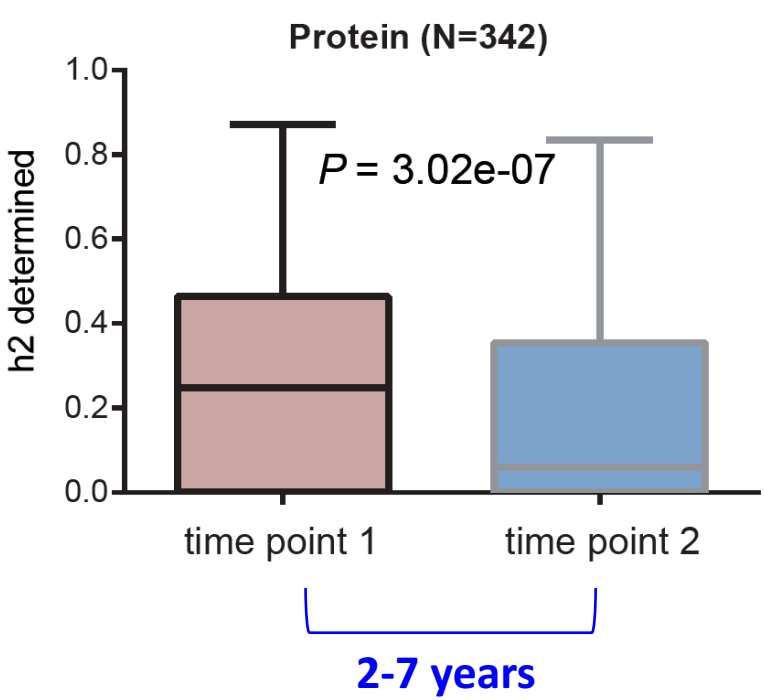
Zhang, X.Z. et al. Cancer Sci (2009).

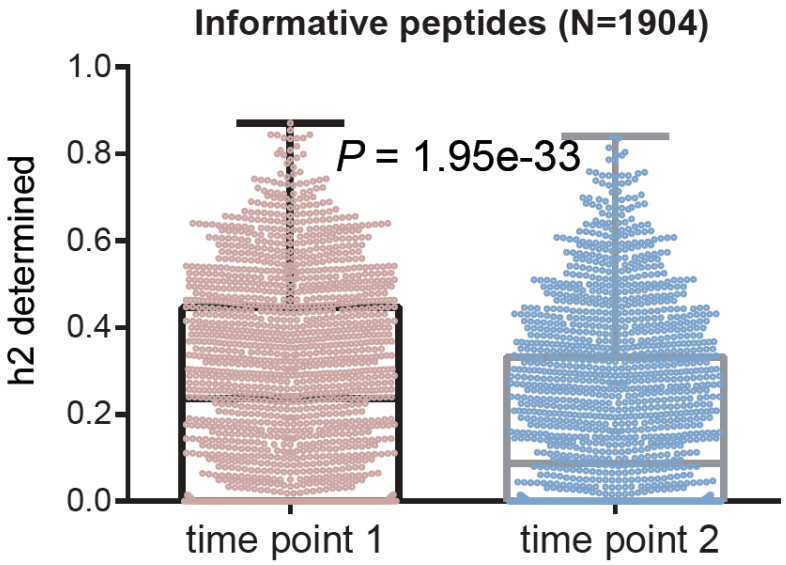
Masui, O. et al. Mol Cell Proteomics (2013).

Hodgkinson, V.C. et al. J Proteomics (2012).

Kang, U.B. et al. BMC Cancer (2010).

Decreased genetic regulation during aging



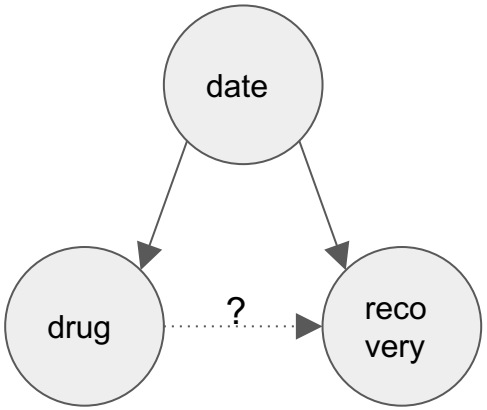


Outline

- The data matrix (proteins vs. samples) as currency of proteomic experimentation
- How do we generate data matrices by mass spectrometry?
- What dimensions should a data matrix have to be maximally useful?
- Can we quantify sources of variation and what are they?
- Can we **avoid, recognize and correct** artifactual variation?

Batch effects can confound biological conclusions

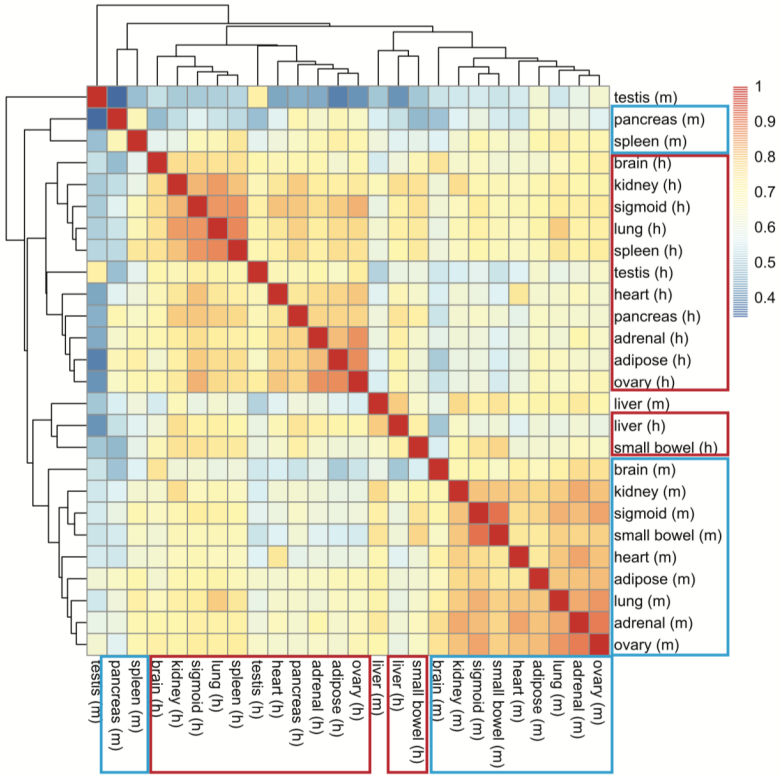
When not accounted for, batch effects can lead to incorrect biological conclusions.



A practical example

The Mouse ENCODE Consortium reported that comparative gene expression data from human and mouse tend to cluster more by species rather than by tissue.

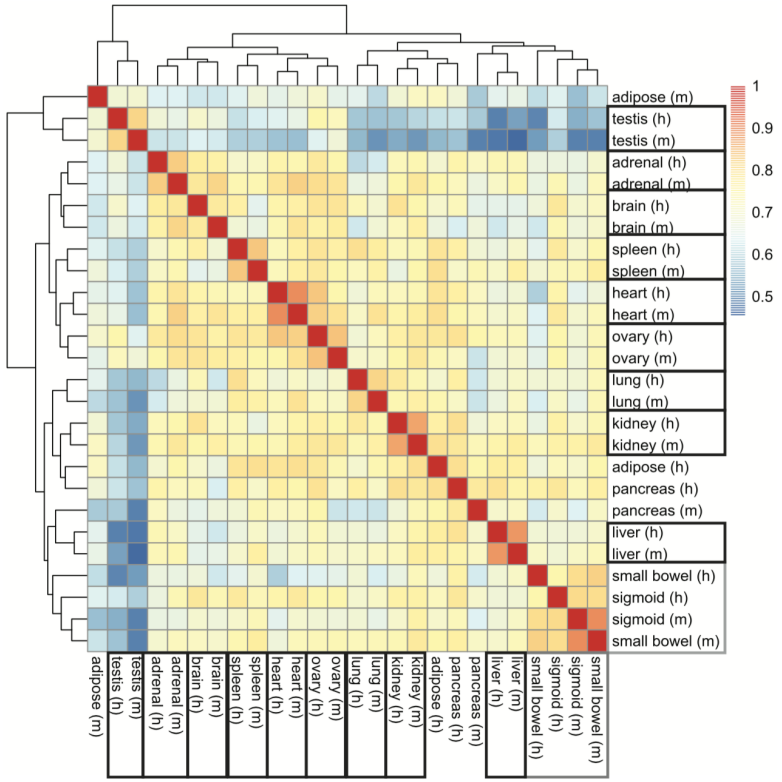
Yue F, Cheng Y, Breschi A, et al.: A comparative encyclopedia of DNA elements in the mouse genome. Nature. 2014; 515(7527): 355–364.



D87PMJN1 (run 253, flow cell D2GUAACXX, lane 7)	D87PMJN1 (run 253, flow cell D2GUAACXX , lane 8)	D4LHBFN1 (run 276, flow cell C2HKJACXX , lane 4)	MONK (run 312, flow cell C2GR3ACXX , lane 6)	HWI-ST373 (run 375, flow cell C3172ACXX , lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	Human
testis		pancreas		Mouse

Gilad et al. repeated the analysis accounting for the fact that the assignment of samples to sequencing flowcells and lanes was nearly completely confounded with the species annotations of the samples.

Gilad, Y. & Mizrahi-Man, O. A reanalysis of mouse ENCODE comparative gene expression data. F1000Research (2015).



"... their conclusions are unwarranted, not wrong, because the study design was simply not suitable for addressing the question of 'tissue' vs. 'species' clustering of the gene expression data."

Possible sources of batch effects

Batch effects which you can reduce before by proper design:

- New Reagents (e.g. fresh prepared vs. different LOTs)
- Machines used (e.g barocycler)
- Students involved, especially when several students work on the same project
- Patient sample handling by the collaborators (dates, labs/technicians need to be tracked as much as feasible)

Batch effects you can hardly influence -> **document properly**

- Track the date
- Special weather conditions (hot days/ air con not working)
- Students involved, especially when several students work on the same project
- Mass spec downtime / column change

Designing to minimize batch effects

Randomizing is useful to guard against unknown/unaccounted effects

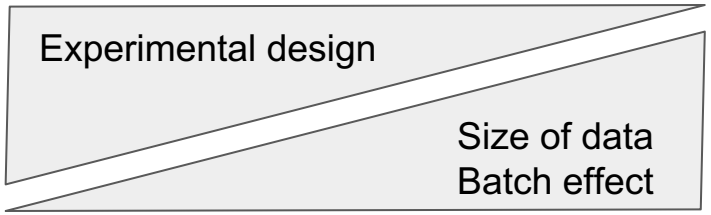
Blocking helps mitigate the effect of known batch effects

Blocks have to be arranged in a way that they do not confound the variable(s) of interest

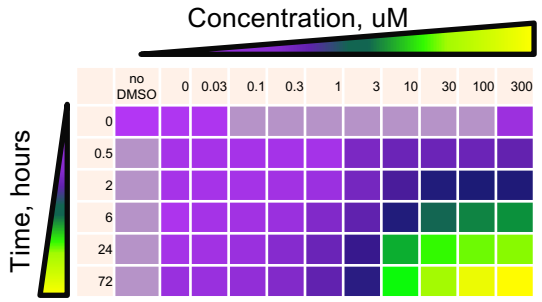
Preferably, the blocks are arranged in a balanced design

For more than one known batch effect, they should be nested if possible

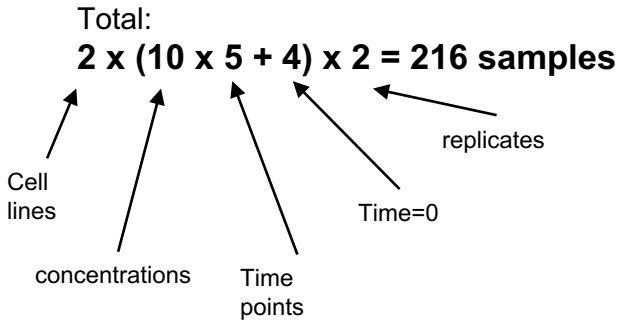
Unfortunately, batch effects increase with increasing data size, whereas our ability to control experimental design decreases.



Case study:
Prostate cancer cell line perturbation



LNCaP	LNCaP-abl
not castration-resistant	castration-resistant
sensitive to Enzalutamide*	not sensitive to Enzalutamide*



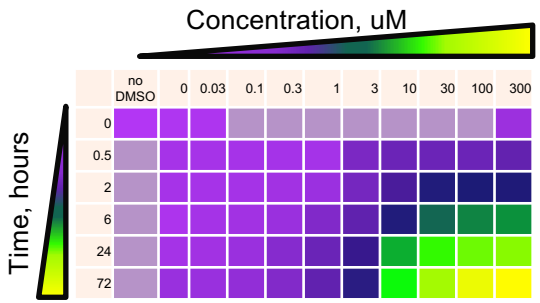
*Enzalutamide - androgen receptor antagonist and FDA-approved drug for CRPC patients

PCT batches

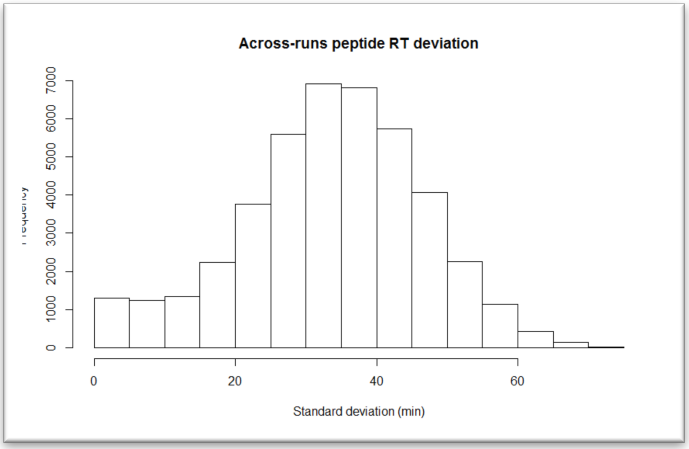
1	2	3	4	5	6	7	8
101a	33a	85a	8a	101b	33b	85b	8b
70a	57a	16a	72a	70b	57b	16b	72b
41a	24a	94a	49a	41b	24b	94b	49b
79a	66a	7a	81a	79b	66b	7b	81b
32a	15a	53a	40a	32b	15b	53b	40b
88a	75a	48a	107a	88b	75b	48b	107b
23a	6a	62a	60a	23b	6b	62b	60b
97a	84a	39a	31a	97b	84b	39b	31b
14a	47a	71a	69a	14b	47b	71b	69b
56a	93a	30a	22a	56b	93b	30b	22b
5a	38a	106a	78a	5b	38b	106b	78b
65a	52a	104a	13a	65b	52b	104b	13b
46a	29a	100a	87a	46b	29b	100b	87b
74a	61a	21a	4a	74b	61b	21b	4b
37a	20a	59a	96a	37b	20b	59b	96b
83a	105a	12a	45a	83b	105b	12b	45b
28a	103a	68a	55a	28b	103b	68b	55b
92a	90a	3a	36a	92b	90b	3b	36b
19a	11a	77a	64a	19b	11b	77b	64b
51a	99a	44a	27a	51b	99b	44b	27b
10a	2a	86a	73a	10b	2b	86b	73b
102a	58a	35a	18a	102b	58b	35b	18b
80a	43a	95a	82a	80b	43b	95b	82b
1a	67a	26a	9a	1b	67b	26b	9b
89a	34a	54a	91a	89b	34b	54b	91b
42a	76a	17a	50a	42b	76b	17b	50b
98a	25a	63a	108a	98b	25b	63b	108b

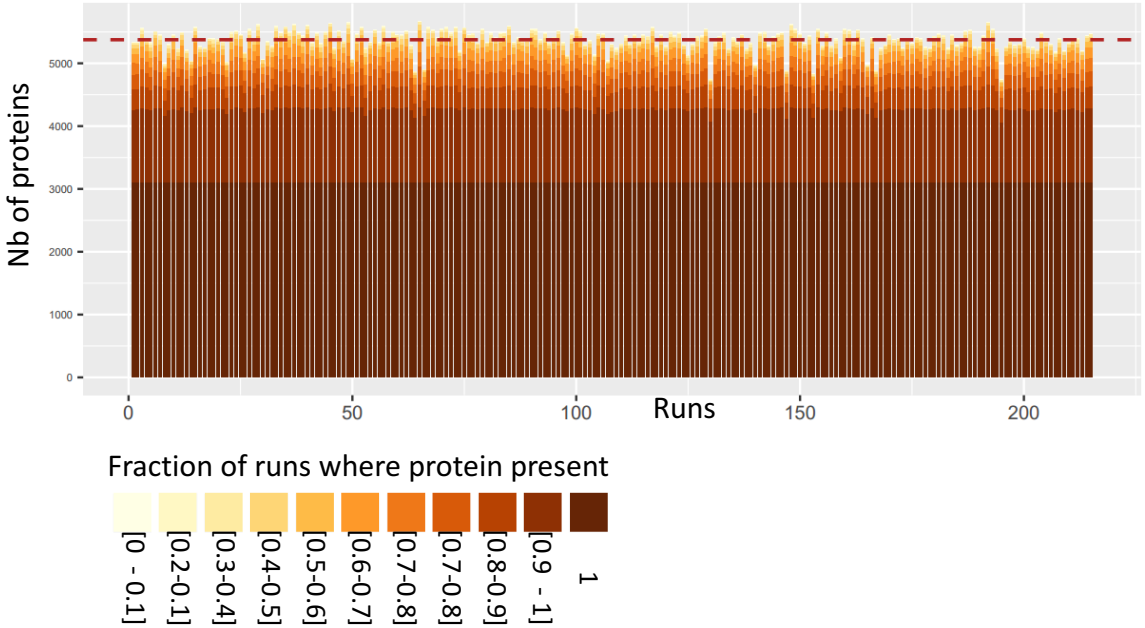
MS run

		Time (hours)					
		0	0.5	2	6	24	72
without DMSO		101a					
Concentration of Enzalutamide (uM) with 1% DMSO	0	102a	1a	11a	21a	31a	41a
	0.03	103a	2a	12a	22a	32a	42a
	0.1		3a	13a	23a	33a	43a
	0.3		4a	14a	24a	34a	44a
	1		5a	15a	25a	35a	45a
	3		6a	16a	26a	36a	46a
	10		7a	17a	27a	37a	47a
	30		8a	18a	28a	38a	48a
	100		9a	19a	29a	39a	49a
	300	104a	10a	20a	30a	40a	50a



Microflow LC SWATH reproducibility and protein identifications





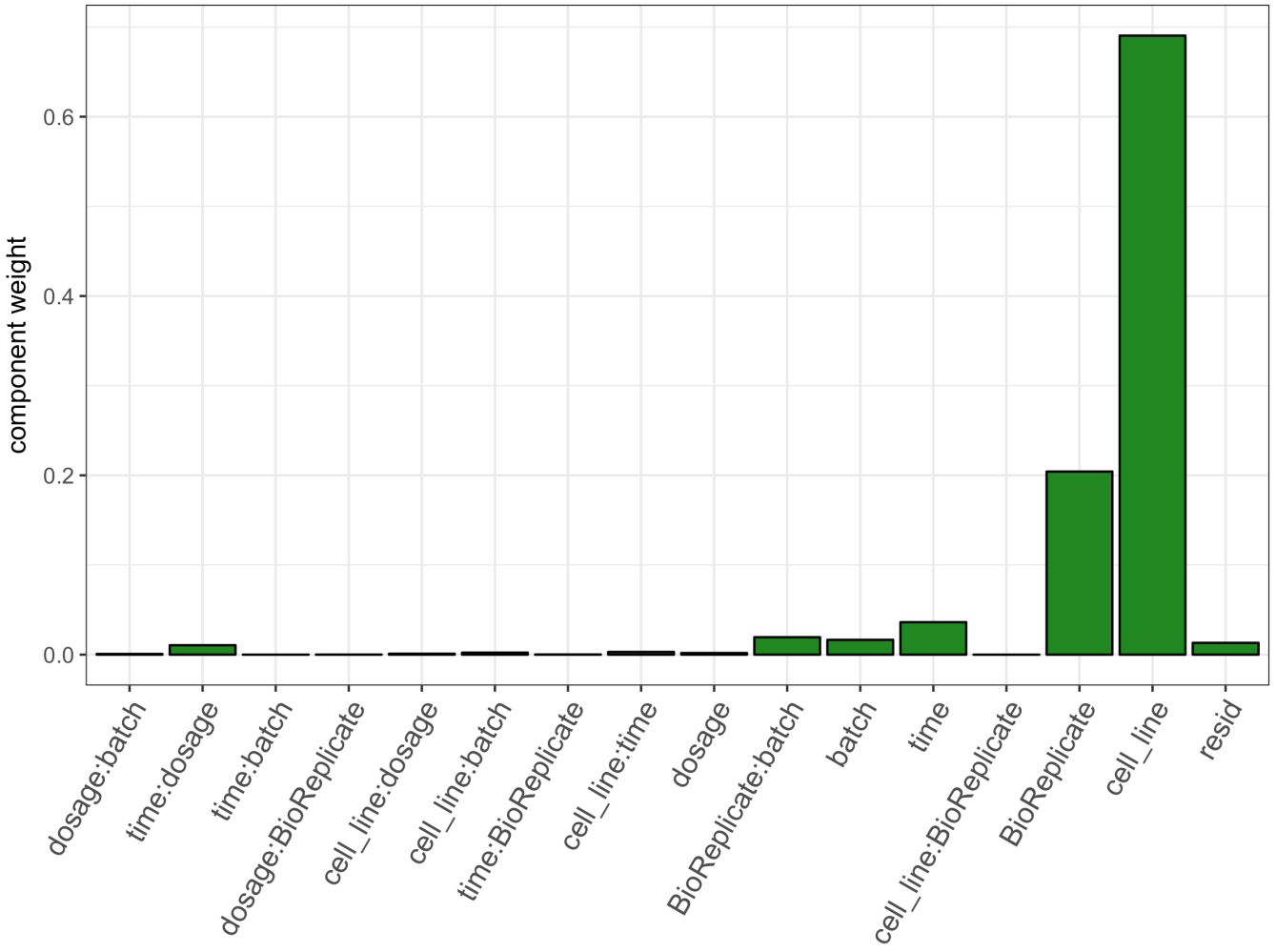
=> 30-second retention time deviation allows confident across-runs chromatographic peak realignment and consistent peptide quantification (removes the need for TRIC?)

=> >4000 proteins are identified throughout the 216 injections

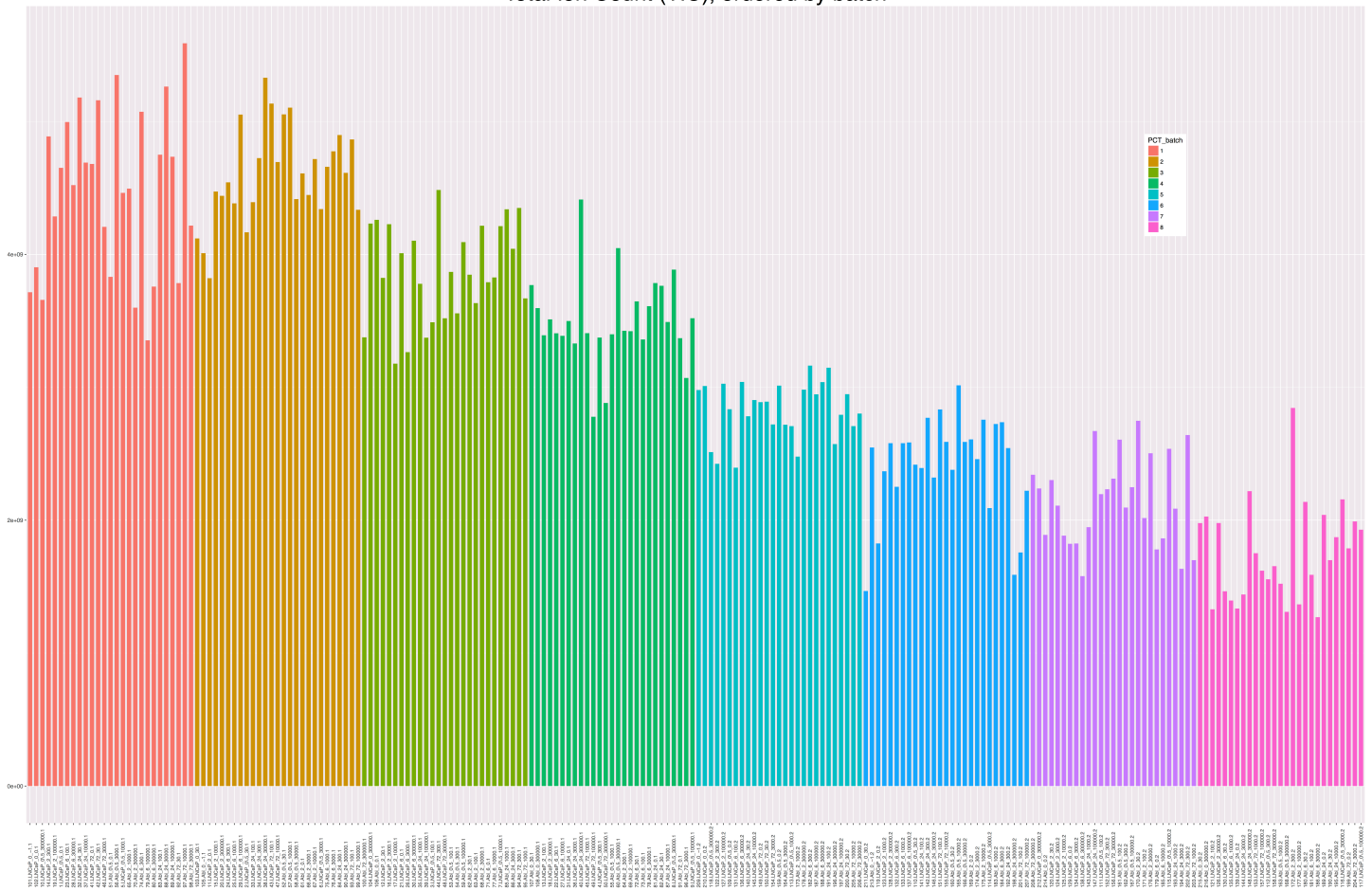
R plots: Jelena Čuklina

PVCA: Principal Variance Component Analysis

shows which

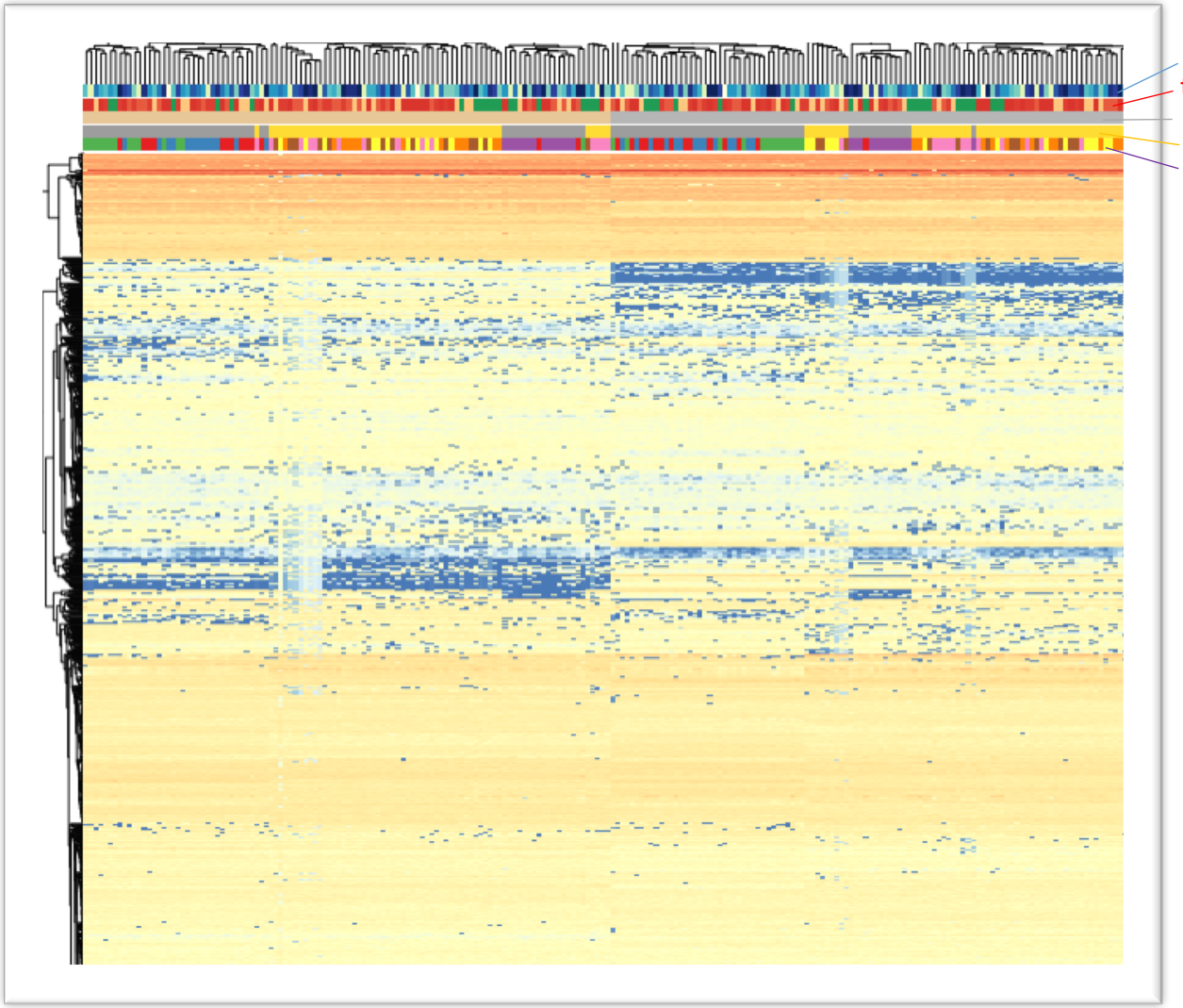


Total Ion Count (TIC), ordered by batch



Data analysis - non-hierarchical clustering

before



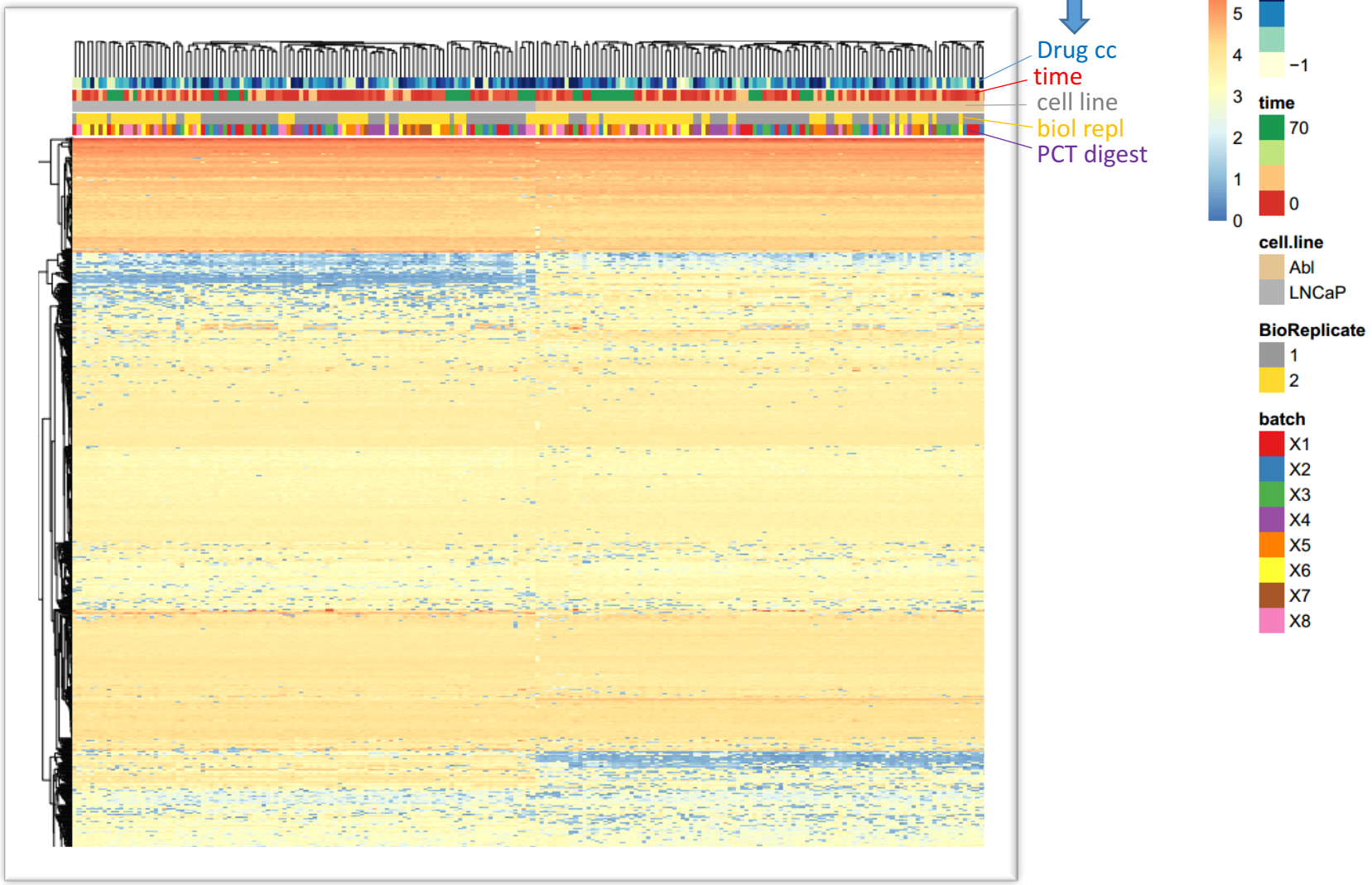
Both cell lines do cluster perfectly from each other

But no clear biological replicate clustering even after PCT batch correction

R plots: Jelena Čuklina

Data analysis - non-hierarchical clustering

after "simple" batch correction (mean centering)



Both cell lines do cluster perfectly from each other

But no clear biological replicate clustering even after PCT batch correction

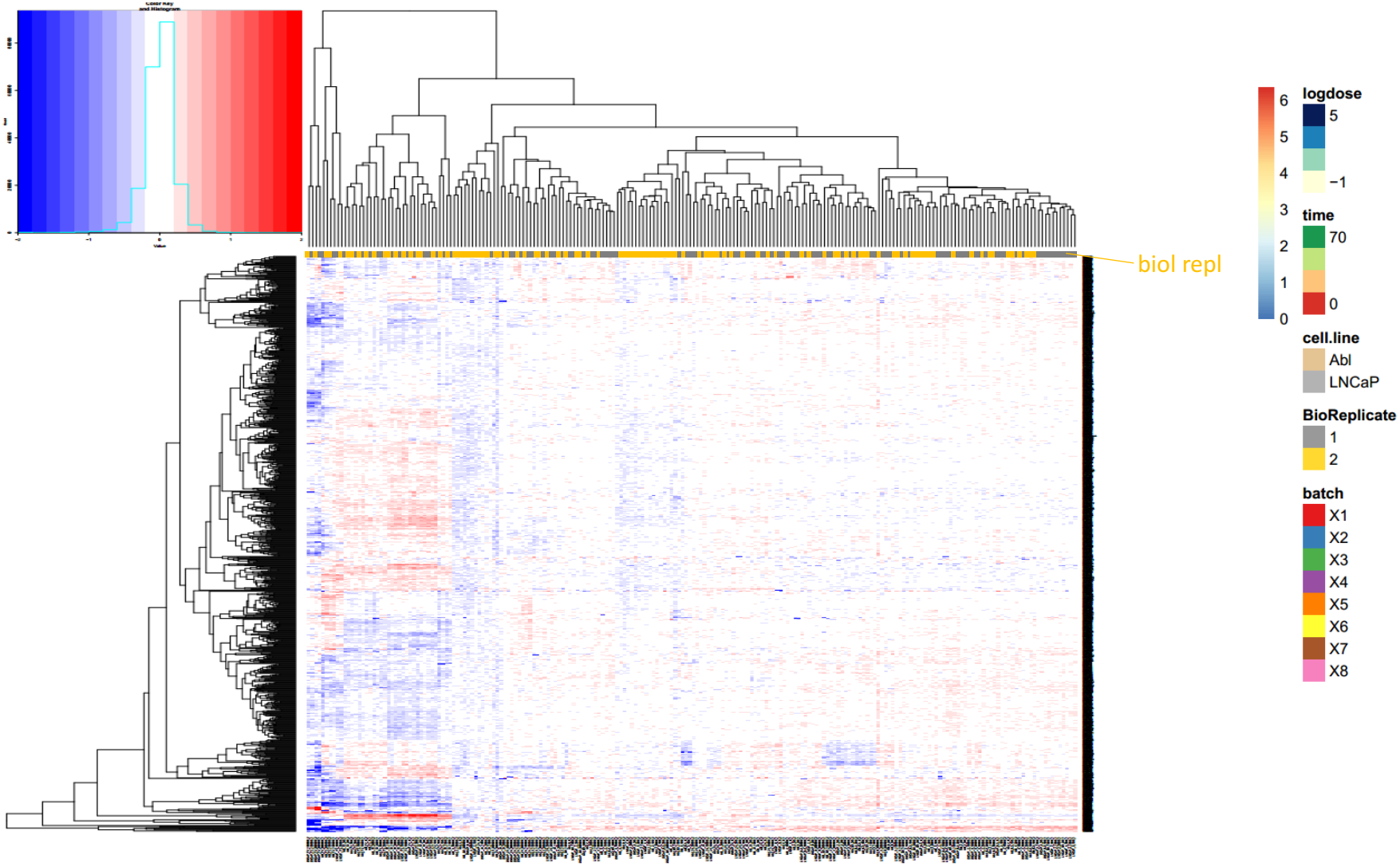
Simple correction does not help better clustering of the biological replicates

R plots: Jelena Čuklina

Data analysis - non-hierarchical clustering

after "sophisticated" batch correction

(lowess fit for each day to correct whatever gradual time-dependent effects/declines over the runs)



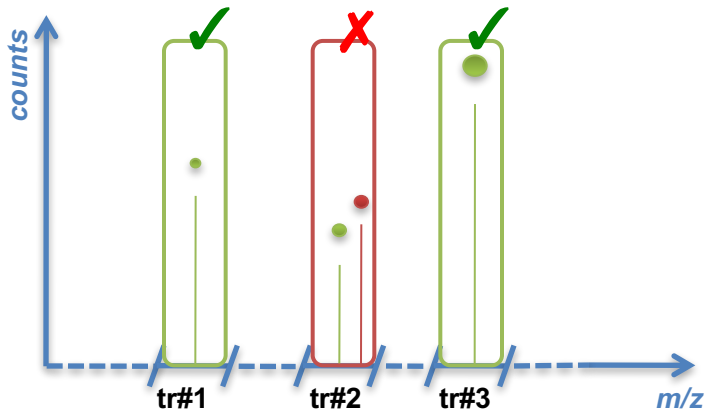
Batch correction and plots: Theodora Katsila

Thank you for your attention!!

SWATH-MS performance: MS/MS data specificity

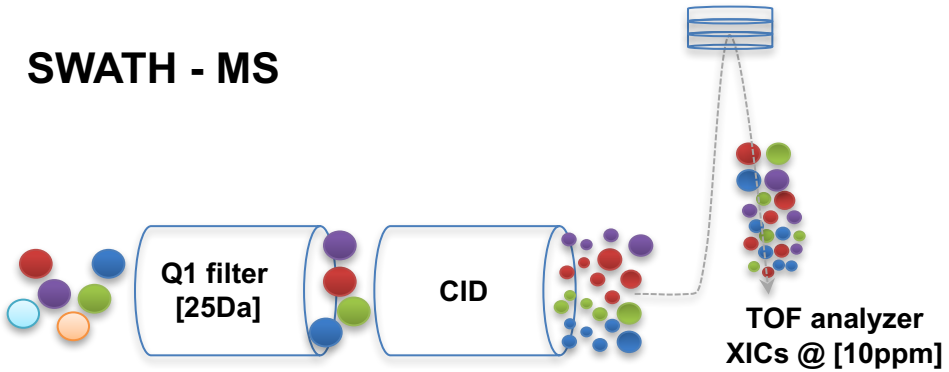
Standard SRM

SRM traces / pseudo-MS2 representation

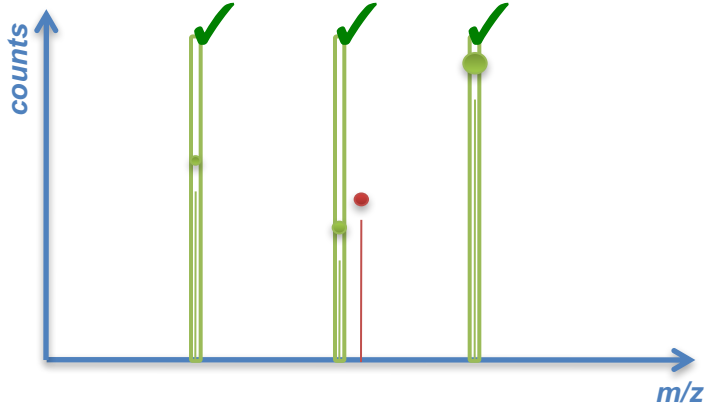


=> some transitions can be contaminated...

SWATH - MS



High res. / high mass acc. full MS2 spectra

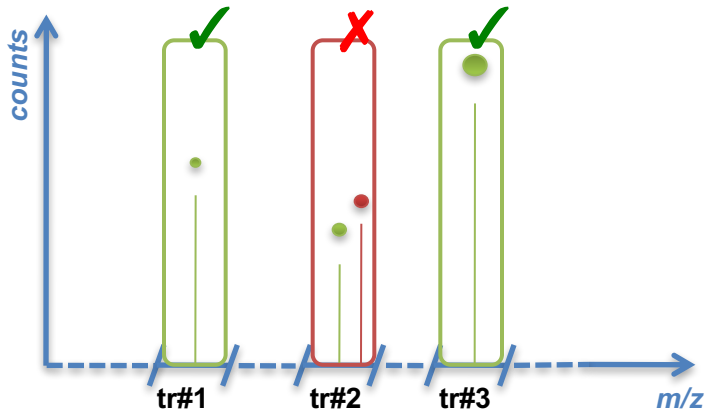


=> some transitions get free of contamination

SWATH-MS performance: MS/MS data specificity

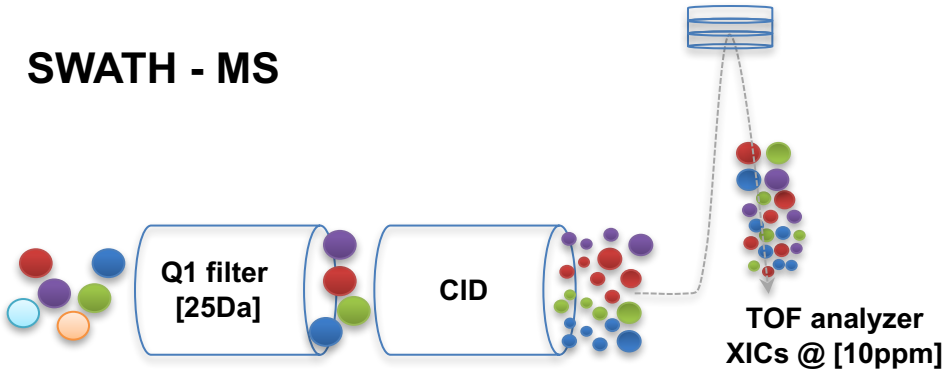
Standard SRM

SRM traces / pseudo-MS2 representation

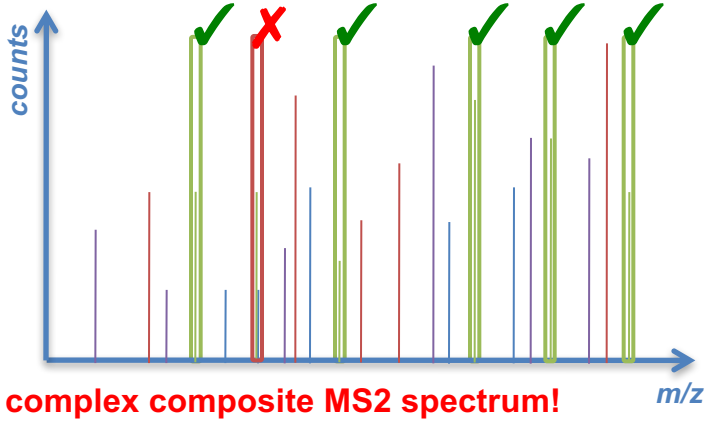


=> some transitions can be contaminated...

SWATH - MS



High res. / high mass acc. full MS2 spectra



=> complex composite MS2 spectrum!

=> but more choice to select contaminant-free transitions

SWATH-MS performance: MS/MS data specificity

- SRM collider : computes the occurrence of fragment ion interferences for various combinations of precursor isolation window width and fragment ion mass accuracy

- computed the theoretical fragment ion spectra for 93000 2+,3+ precursors corresponding to yeast proteins of peptide atlas.