STATENS SERUM INSTITUT

1:40 p.m. - 2:05 p.m.

Biofluid Proteomics: Potential, Pitfalls, and Solutions

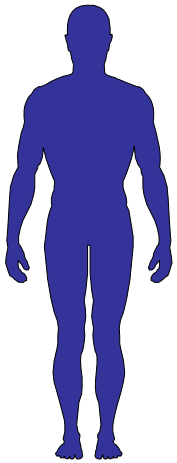
Niels Heegaard, M.D., D.Sc. (Med)
Director, Department of Clinical Biochemistry and Autoimmunology
Statens Serum Institut (Denmark)

STATENS SERUM INSTITUT

Potential:

- Enormeous complexity of the protein and peptide universe present in humans are beginning to be matched by analytical methods
- Promises to give us new biomarkers for diagnosis & individualized medicine

Proteome potential



~ 20 k genes =>
~ 1,000 k proteins

Concentrations vary (log 7-9)

Different over time and location.

Reflects events in the body

Biomarker potential

Stratification

Predisposition to disease

Early indicator of disease

Disease type

Drug selection

Dose selection

Toxicity avoidance

Monitoring

Disease initiation

Progression

Severity

Drug choice/effect

STATENS SERUM INSTITUT

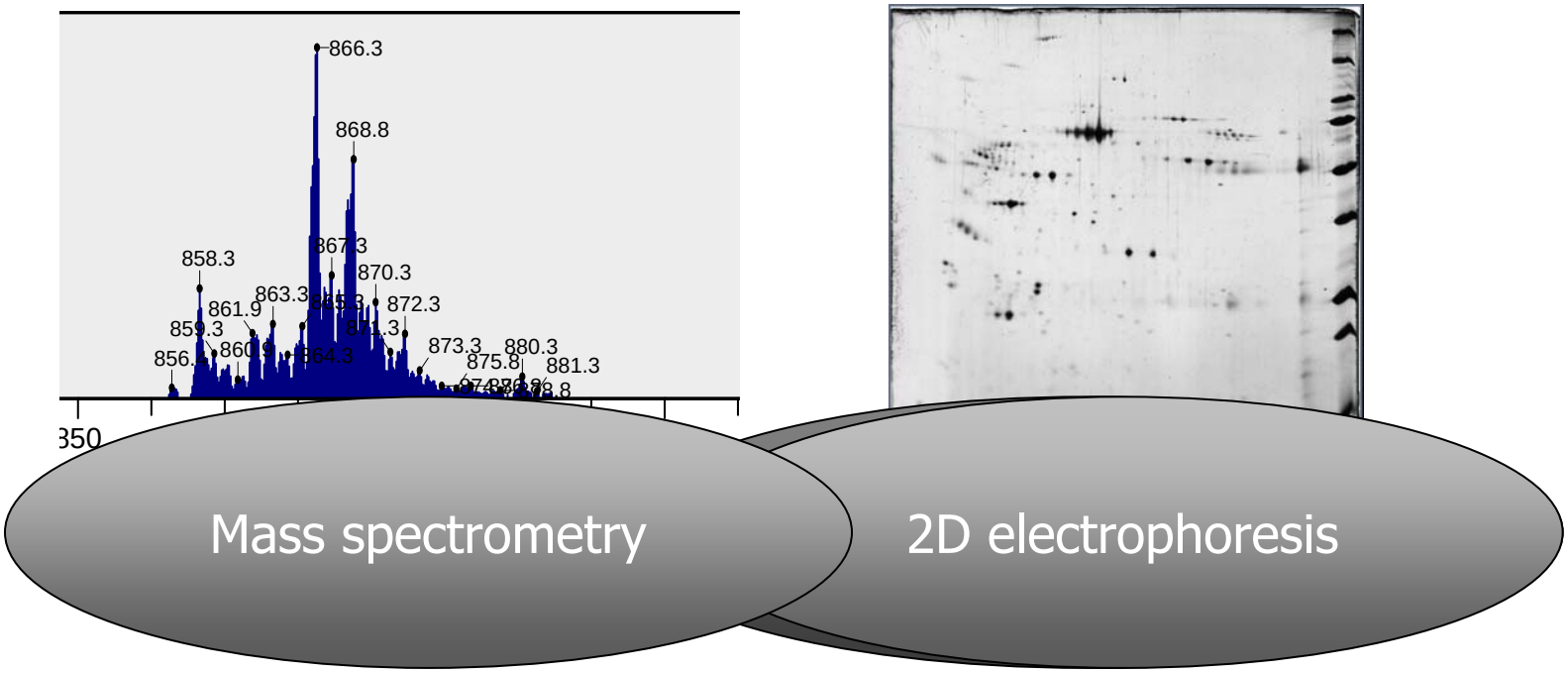
Evolution of protein analytical methods

- Zone gel electrophoresis
  - 5-10 bands
- Immunoelectrophoresis
  - 20-50 proteins
- SDS-polyacrylamide gel electrophoresis
  - 50-100 protein bands
- Two-dimensional gel electrophoresis – *more than 1000 proteins*
- Mass spectrometry - *~30,000 data points/sample; 100 of peptides/proteins*
- Two-dimensional on-line separations (e.g. LC-mass spec)
  - 1,000 of proteins

Biofluid Proteomics: Potential, Pitfalls, and Solutions

STATENS SERUM INSTITUT

- New biomarkers in the peptide and protein universe (proteomes) of biological fluids?



Output:	30.000 m/z intensity values	up to 2.000 spots
Years in use:	~10	34
New markers in clinical use:	None	1-2

Issues in multiparameter analyses of biofluids for biomarker discovery

- Seeing differences when there are no relevant differences
- Not seeing differences when there are relevant differences
- Seeing differences but not the relevant ones

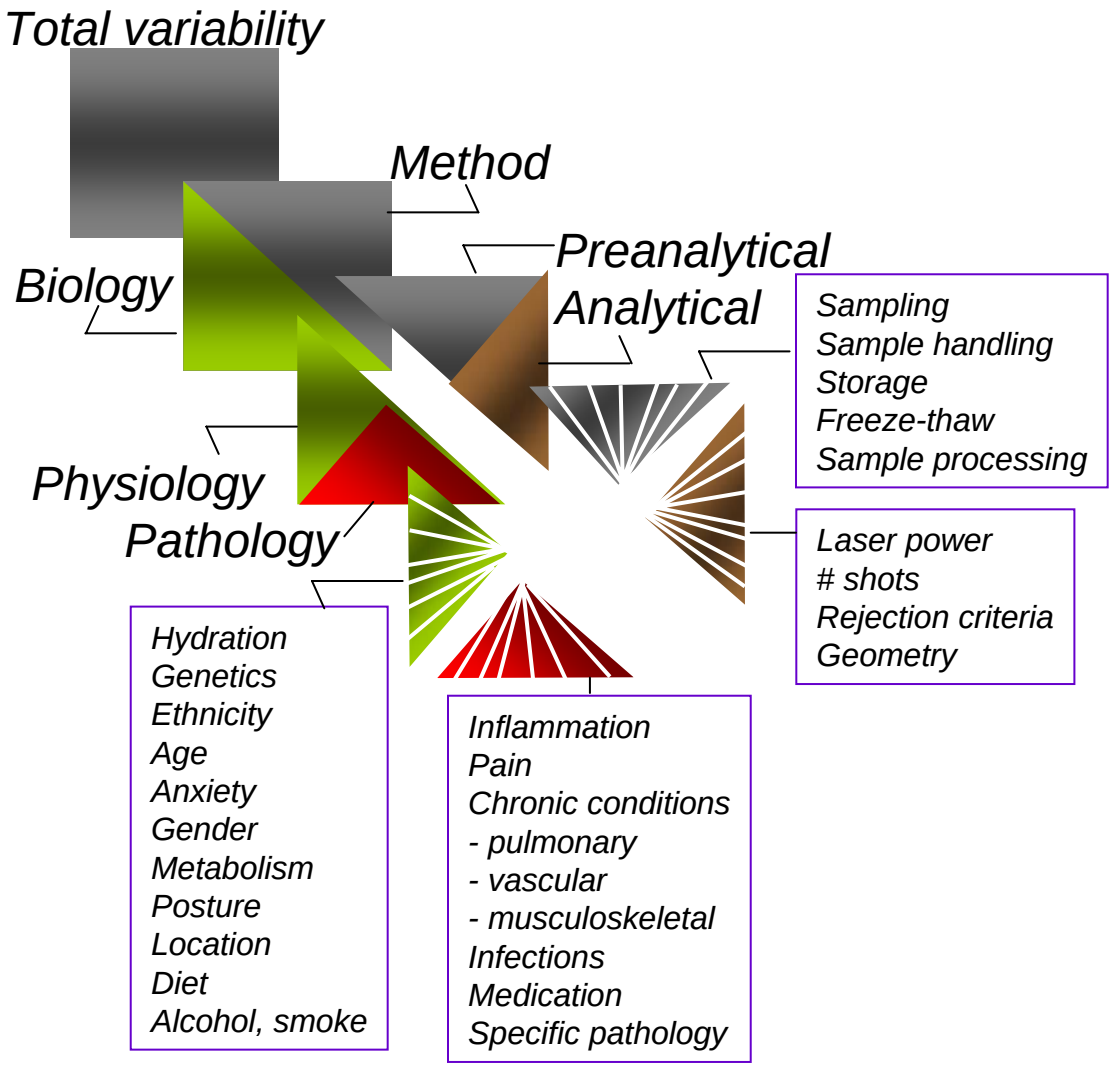
STATENS SERUM INSTITUT

Variability in biofluid analysis

- We are interested in the biological variables associated with a specific pathological condition
- This is obscured by tremendous noise

STATENS SERUM INSTITUT

Variability contributors in high-resolution analysis of biofluids



Seeing differences between groups using analyses yielding multiple variables

- Overfitting is an obvious risk – too many data and too few samples
  - We need at least 10 times more samples than features!
  - Look for univariate effects
  - Bonferroni-correct significance levels (*p*/N)
  - Use PCA first or other unsupervised method
  - Do permutation analysis
  - Independently train, cross-validate, and validate

STATENS SERUM INSTITUT

Seeing differences

- Minimize noise
  - Study design from patient to analysis: focus on reducing variability
  - Reduce complexity of material
  - Optimize analytical procedures for reproducibility
  - Reduce amount of data (=variable selection)

STATENS SERUM INSTITUT

Minimize noise - Study Design

Ovarian Cancer proteomics
Pelvic Mass Study

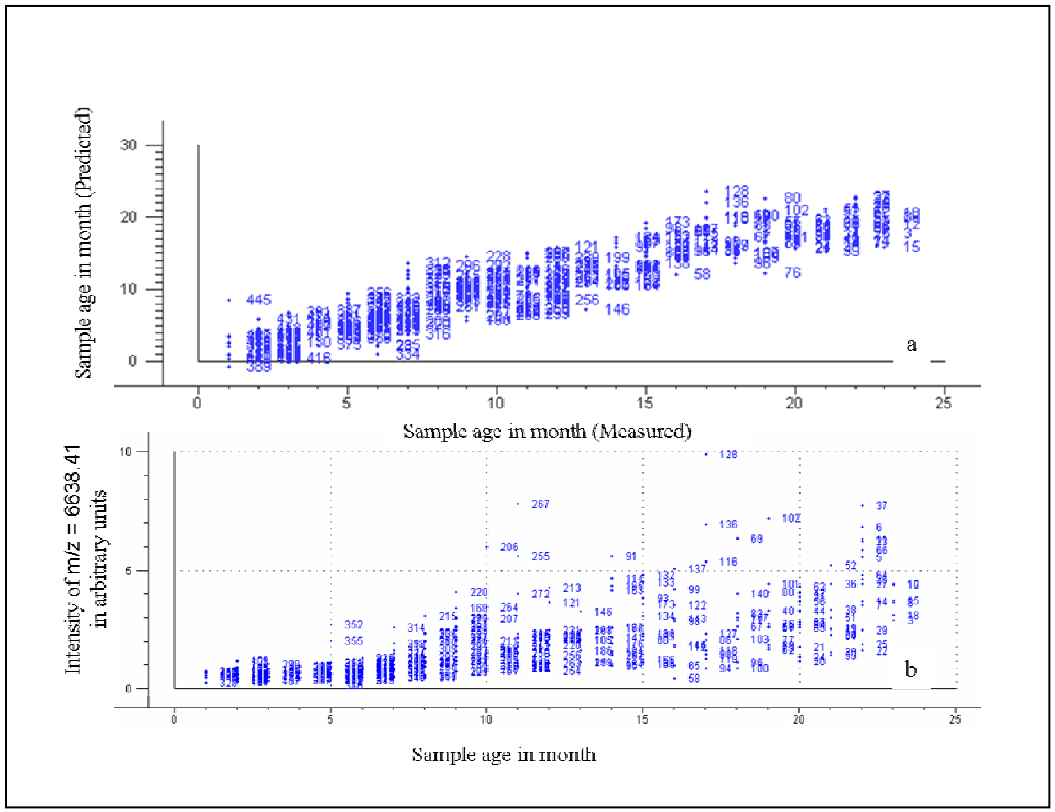
- All samples (disease and disease-control) collected in a specialized clinic following uniform and strict guidelines. Only women with a risk-of-malignancy index RMI > 150, and age > 18 years are included.

- >400 sera were collected. 25 % expected malignant (borderline to stage IV). 75 % are expected to be benign conditions

- Samples are thawed on ice and handled on a fractionation robot using WCX beads and IMAC-Cu beads at 22°C and humidity set to 35%

STATENS SERUM INSTITUT

Minimize noise. The introduction of new noise – systematic changes caused by sample handling



a) Long-term storage effects at -20°C. a) Modelling sample age using PLS. Samples were collected over a period of 23 months.

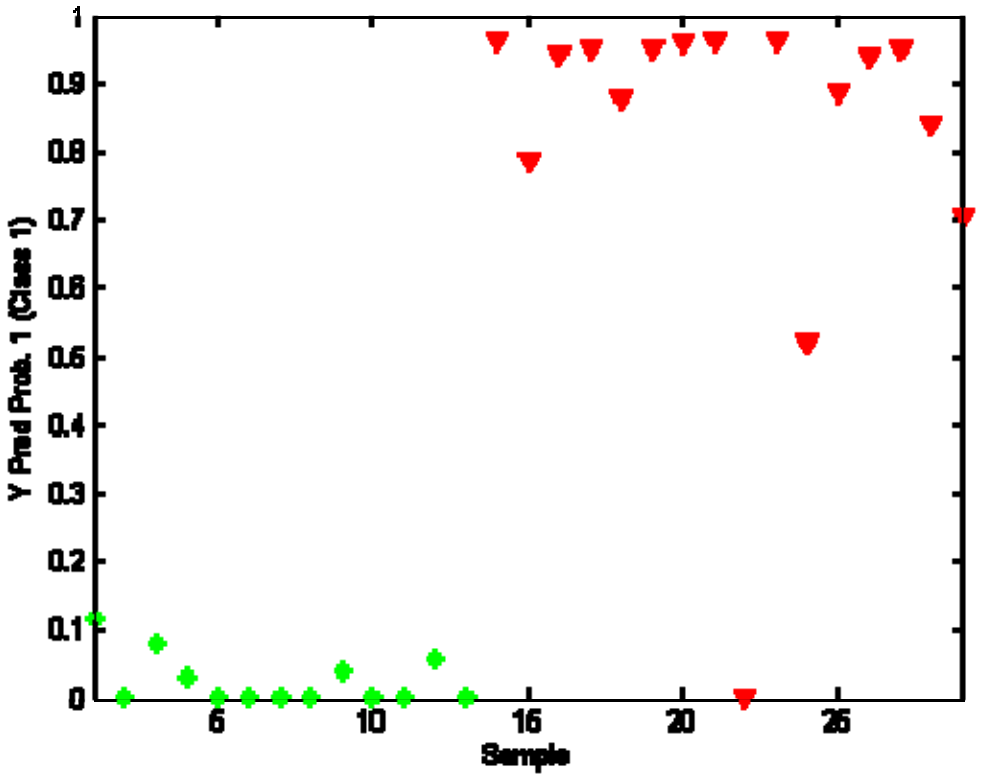
b) The figure shows a variable with m/z 6638.41 with intensity increasing as a function of storage time. The variance between samples becomes larger with increased storage time.

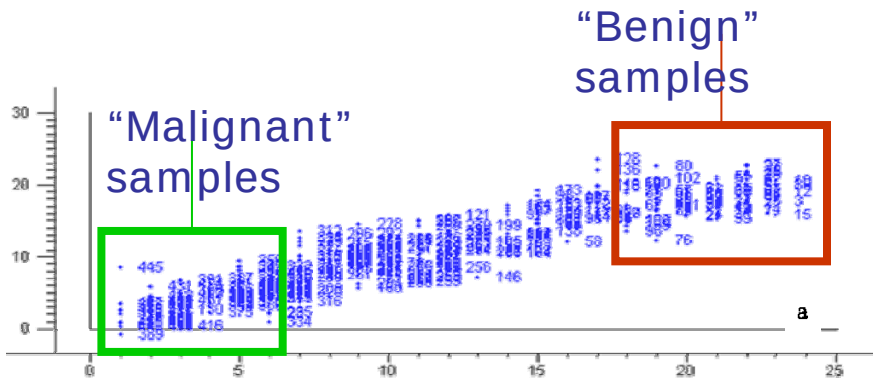
West-Nørager M et al. Anal Chem 2009 (ePub)

STATENS SERUM INSTITUT

Possible consequence of systematic noise

Example of spurious classification model caused by storage artifacts.





STATENS SERUM INSTITUT

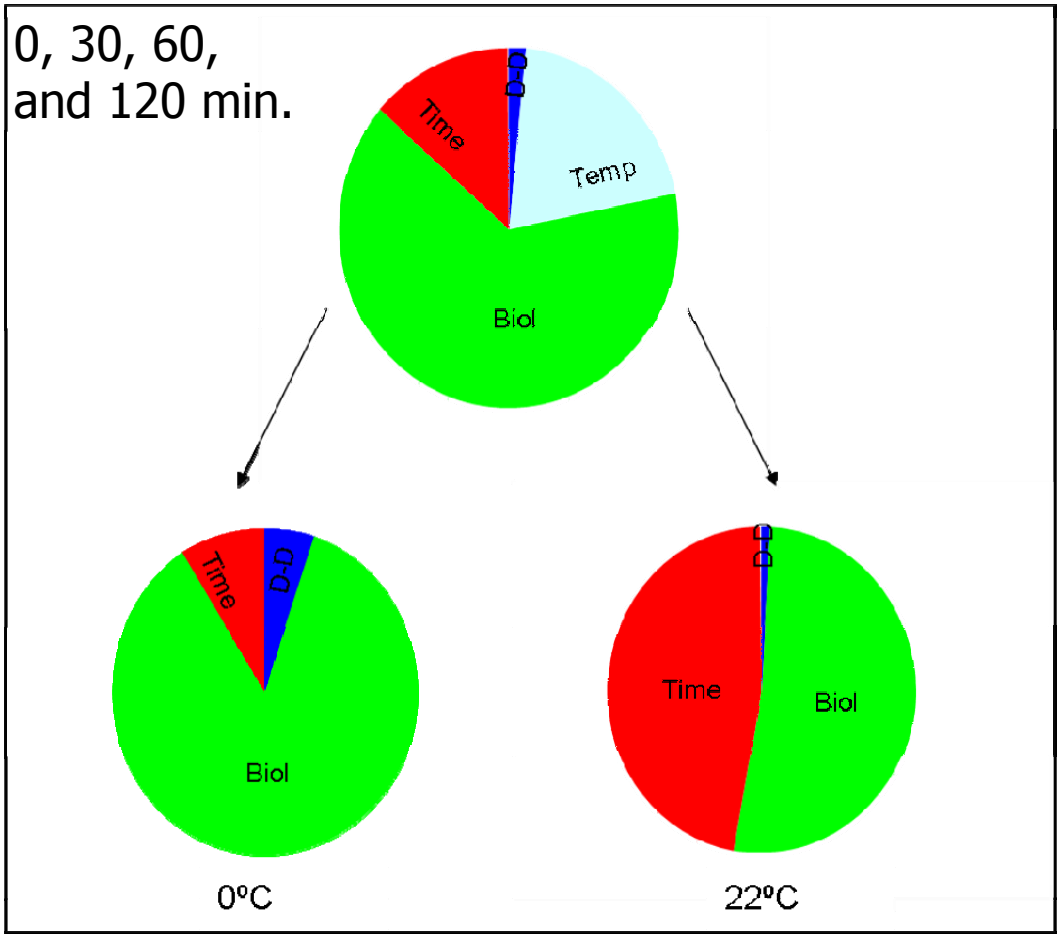
Variability arising *in vitro.* Table of typical peaks involved in post-sampling changes in normal sera

Name/sequence	Fragm. mass
Complement C3f: aa 2-16 (SKITHRIHWESASLL)	**1777.9**
Complement C3f: aa 1-16 (SSKITHRIHWESASLL)	**1864.9**
Complement C3f: aa 1-17 (SSKITHRIHWESASLLR)	**2021.7**
Fibrin alpha C term fragment: aa 81-105 (SSSYSKQFTSSTSYNRGDSTFESKS)	**2767.4**
Fibrin alpha C term fragment: aa 81-106 (SSSYSKQFTSSTSYNRGDSTFESKSY)	**2931.5**
Fibrinopeptide A: aa 1-12 (EGDFLAEGGGVR)	**1206.5**
Fibrinopeptide A: aa 3-16 (SGEGDFLA EGGGVR)	**1350.7**
Fibrinopeptide A: aa 2-16 (DSGEGDFLAEGGGVR)	**1465.5**
Fibrinopeptide A (Modifications: 3 Phosphorylated): aa 1-16 (ADSGEGDFLAEGGGVR)	**1616.9**
Kininogen: aa 439-456 (HNLGHGHKHERDQGHGHQ)	**2080.9**

West-Nørager *et al.* J Chromatogr B 2007

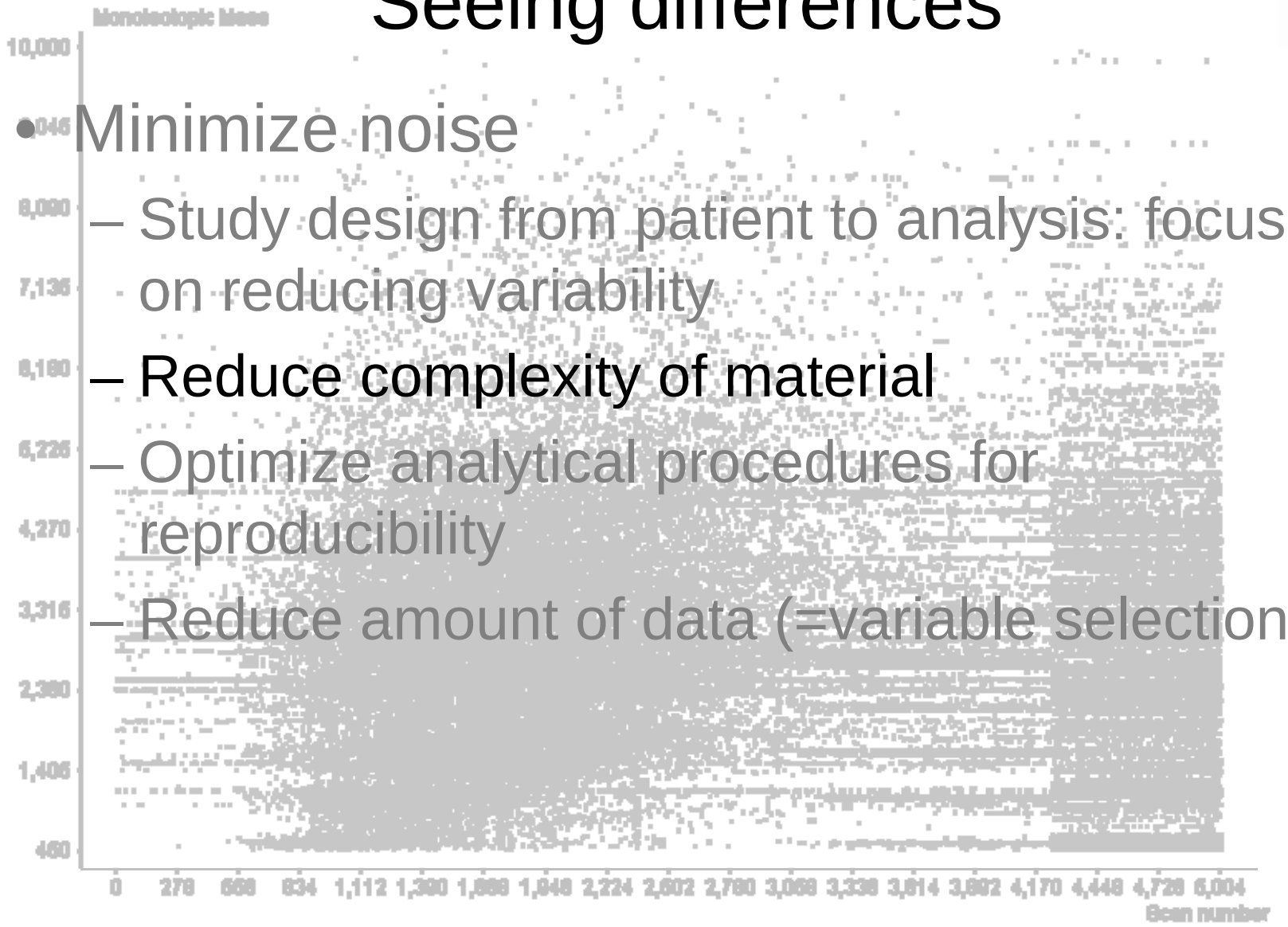
STATENS SERUM INSTITUT

Minimizing noise.
Serum proteomics, WCX-beads
Time, temperature, biology, and day-to-day variation



STATENS SERUM INSTITUT

Seeing differences



- Minimize noise
  - Study design from patient to analysis: focus on reducing variability
  - Reduce complexity of material
  - Optimize analytical procedures for reproducibility
  - Reduce amount of data (=variable selection)

STATENS SERUM INSTITUT

Simplify input to analytical machines

Sample fractionation

- Chromatographic – ”non-biological”
  - Charge, hydrophobicity, size
  - Removal of abundant species

- Biological (*Intelligent Proteomics*)
  - Enrich for post-translational modifications
    - Phosphorylations
    - Glycosylations
    - Others
  - Microparticle isolation
    - Subfractionate on surface markers, size
  - Highly targeted affinity isolation
    - Antibodies
    - Lectins

- COMBINATIONS

STATENS SERUM INSTITUT

Reducing biofluid complexity: CSF-phosphoproteomics.

Cerebrospinal fluid

↓

Concentrate on 3 kD spin filter to
35 mg/ml total protein

↓

Dilute 1:1 in 8 M urea/0.4 M NH_4HCO_3,
reduce and alkylate

↓

Lys-C endoproteinase at 2%, 37 C, 3hrs
followed by trypsin at 2%, 37 C, overnight

↓

N-glycosidase F treatment, 3 hrs
dilute 1:5 in 1 M glycolic acid, 5% trifluoroacetic acid, 80% acetonitrile
and load onto TiO_2-microcolumn & wash with loading buffer

↓

Wash with 1% trifluoroacetic acid, 80% acetonitrile,
elute with NH_4OH, pH 11 and acidify with 1/10 vol. 100% formic acid.
Desalt and analyze

Bahl JMC et al., Anal Chem 2008

Biofluid Proteomics: Potential, Pitfalls, and Solutions

Phosphorylation sites identified in this study are underscored. Previously unknown phosphorylation sites are highlighted.

Bahl JMC et al., Anal Chem 2008

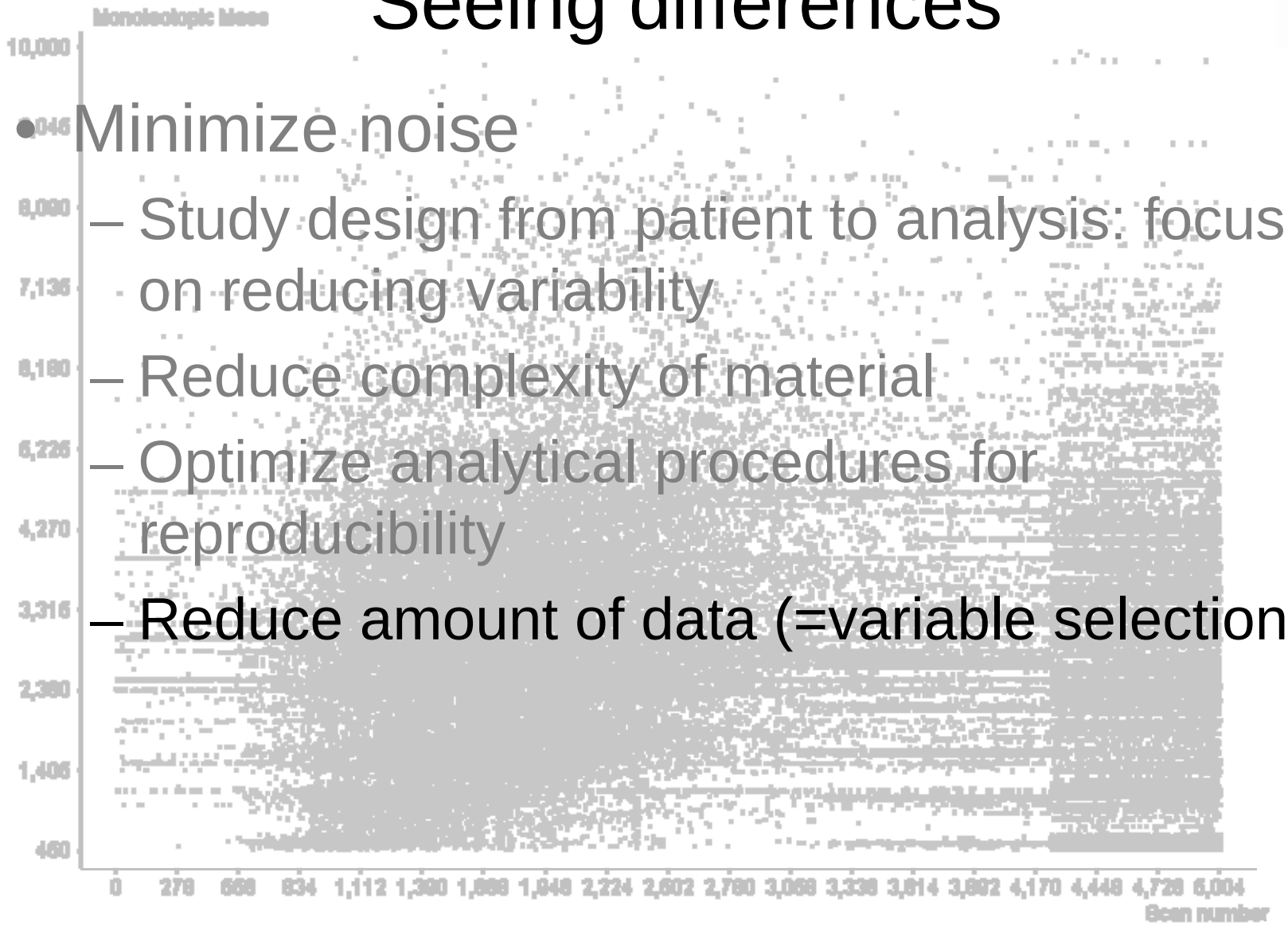
STATENS SERUM INSTITUT

Phosphoproteins in control CSF
- 56 new phosphosites in 38 proteins

Protein	Peptide	Sequence
Osteopontin	52-70	QNLLAPQNAVpSSEETNDFK
	176-203	RPDIQYPDApTDEDITpSHMEpSEELNGAYK
	204-220	AIPVAQDLNAPpSDWDpSR
	223-241	DpSYETpSQLDDQpSAETHSHK
	250-268	ANDEpSNEHSDVIDpSQELpSK
	269-290	VpSREFHpSHEFHpSHEDMLVVDPK
	300-314	FRIpSHELDpSASpSEVN
SPARC like protein 1	76-89	SKEESHEQSAEQGK
	90-98	SSSQELGLK
	187-220	DQGNQEQDPNISNGEEEEEKEPGEVGTHNDNQER
	256-286	MQEDEFDQGNQEQEDNSNAEMEEENASNVNK
	287-299	HIQETEWQSQEGK
	409-426	KAENSSNEEETSSEGNMR
Secretogranin 1	123-131	WAEGGGHSR
	134-153	ADEPQWSLYPSDSQVpSEEVK
	179-202	GEDpSpSEEKHLEEPGETQNAFLNER
	235-254	SSQESGEEAGSQENHPQESK
	259-276	pSQEEpSEEGEEDATSEVDK
	305-324	GHPQEEpSEESNVSMASLGEK
	372-387	APRPQpSEESWDEEDKR
	397-409	MAHGYGEEpSEEER
	617-640	pSAEFPDFYDpSEEPVSTHQEAENEK
Secretogranin 2	257-269	IESQTQEEVRDSK
	549-561	EHLNQGSSQETDK
Receptor-type tyrosine-protein phosphatase N2	355-368	AALGESGEQADGPK
	428-448	SEHPESSLSSEEETAGVENVK
Dickkopf related protein 3	81-103	ASSEVNLANLPPSYHNETNTDTK
Apolipoprotein E	138-152	GEVQAMLGQSTEELR
Fibulin 1	147-163	SQETGDLDVGGLQETDK
Kininogen 1	325-343	ETTCSKESNEELTESCETK
Cystatin C	35-51	LVGGPMDASVEEEGVRR
Secretogranin 3	35-53	ELSAERPLNEQIAEAEEDK
Trans Golgi network integral membrane protein 2	65-83	DSPSKSSAEAQTPEDTPNK
Extracellular matrix protein 2	209-227	EALQSEEDEEVKEEDTEQK
Cadherin 2	121-145	LSLKPTLTEESVKESAEVEEIVFPR
Neuron Navigator 2	1-9	MPAILVASK
Antithrombin III	61-78	KATEDEGSEQKIPEATNR
Serum Albumin	76-88	TCVADESAENCDK
Versican core protein	2108-2124	QEIESETTSEEQIQEEK.
Sushi domain-containing protein 5	290-298	GSGEQQIMR
Proenkephalin A	237-252	FAEALPSDEEGESYSK
Cocaine- and amphetamine-regulated transcript protein	37-51	ALDIYSAVDDASHEK
Selenoprotein P	262-274	DMPASEDLQDLQK
Apolipoprotein L	306-320	VTEPISAESGEQVER
Fructose-bisphosphate aldolase A	29-42	GILAADESTGSIAK
Reticulocalbin-1	71-81	TFDQLTPDESK
Proprotein convertase	683-692	HLAQASQELQ
Matrix remodelling-associated protein 7	125-142	GPSSEGPEEEDGEGFSFK
Amyloid precursor protein	439-450	VESLEQEAANER
Plasminogen	349-386	IPSCDSSPVSTEQLAPTAPPELTPVVQDCYHGDGQSYR
Follistatin-related protein 1	163-170	LDSSEFLK
Kallikrein-6	81-91	ESSQEQSSVVR
Cofilin-1	2-13	ASGVAVSDGVIK
Myrostoyl alanine-rich C-kinase substrate	12-30	GEAAAERPGEAAVASSPSK
Golgi phosphoprotein 2	253-267	EETNEIQVVNEEPQR
Latent-transforming growth factor beta binding protein isoform 1L	1280-1297	GFVPAGESSSEAGGENYK
Testican -3	33-42	SDGGNFLDDK
Ubiquitin	55-63	TLSDYNIQK
Thy-1 membrane protein	88-97	VLYLSAFTSK

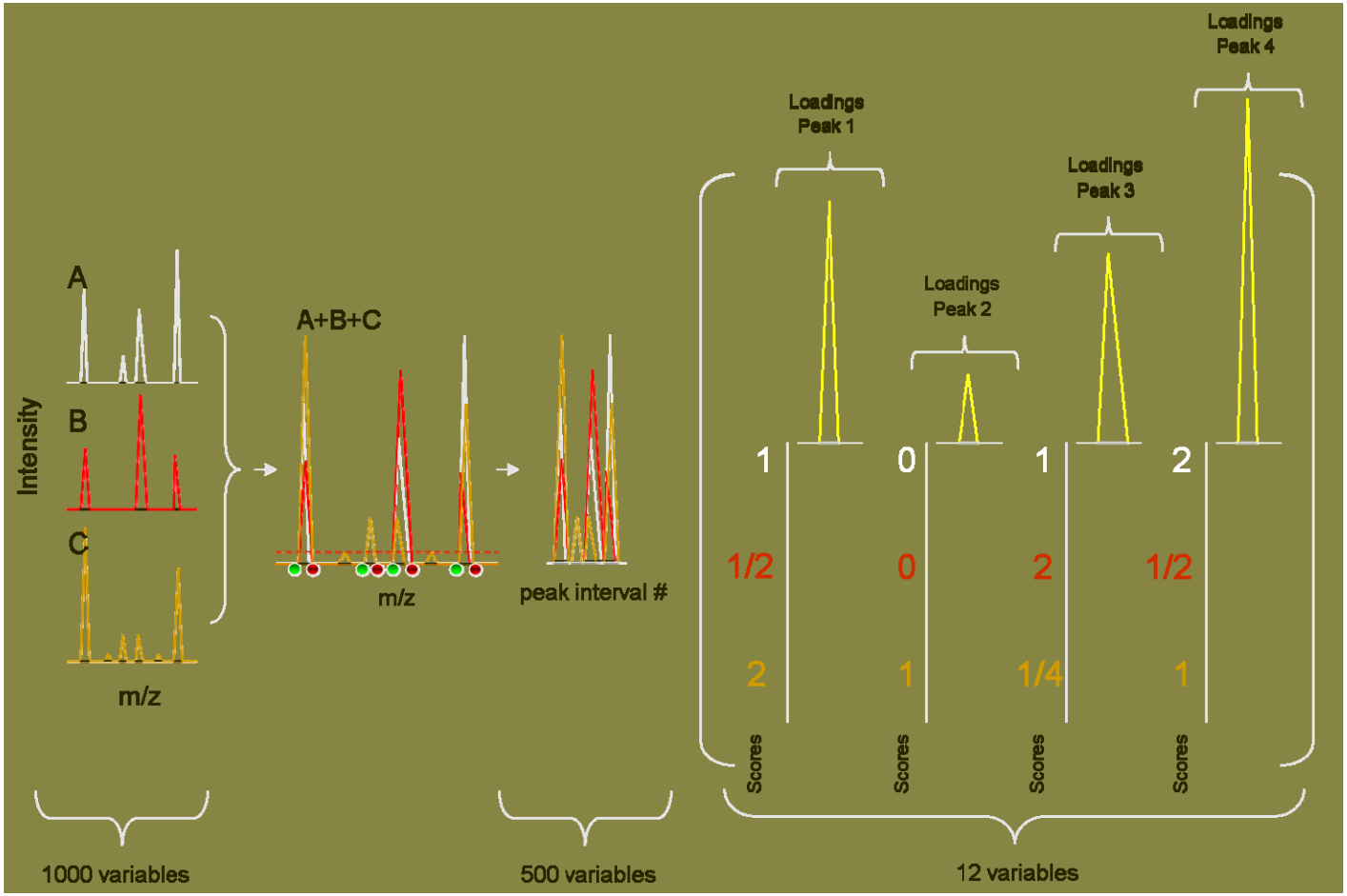
STATENS SERUM INSTITUT

Seeing differences



- Minimize noise
  - Study design from patient to analysis: focus on reducing variability
  - Reduce complexity of material
  - Optimize analytical procedures for reproducibility
  - Reduce amount of data (=variable selection)

A method for data reduction that involves both peak picking and principal component analysis



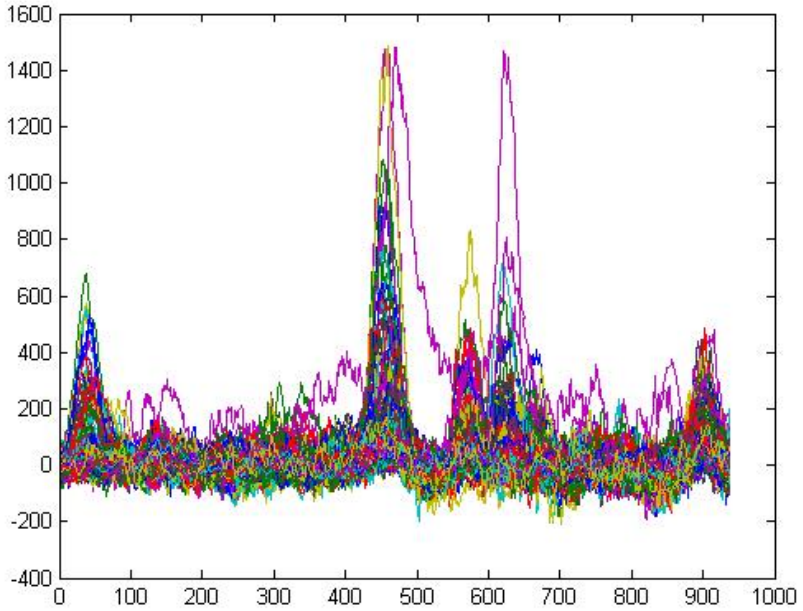
300 samples —› 33.000 points = matrix with 9.900.000 points

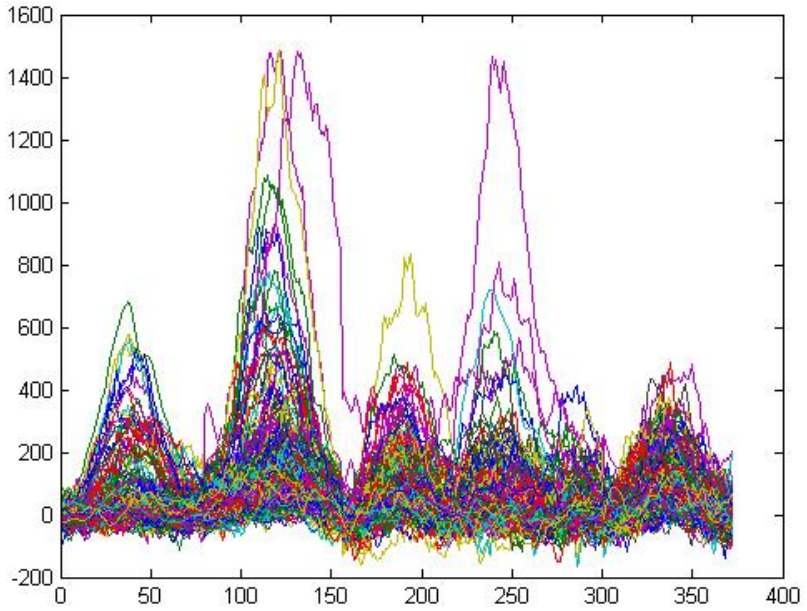
Through data reduction, matrix reduced to 165.000 points

98% in reduction with less than 1% loss of data

STATENS SERUM INSTITUT

Data Reduction I.
Peak picking (aka feature selection).

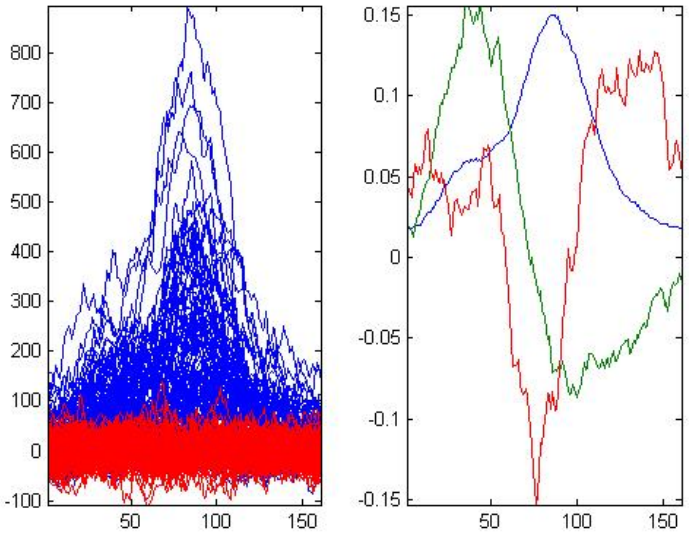




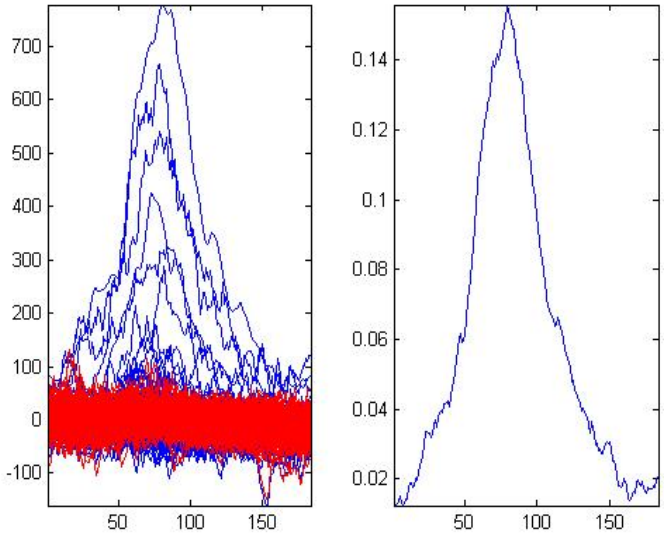
Removal of data between peaks. Yields 50% variable reduction

STATENS SERUM INSTITUT

Data Reduction II.
PCA (principal component analysis) performed in each peak interval (MatLab-script available)



Peak #107



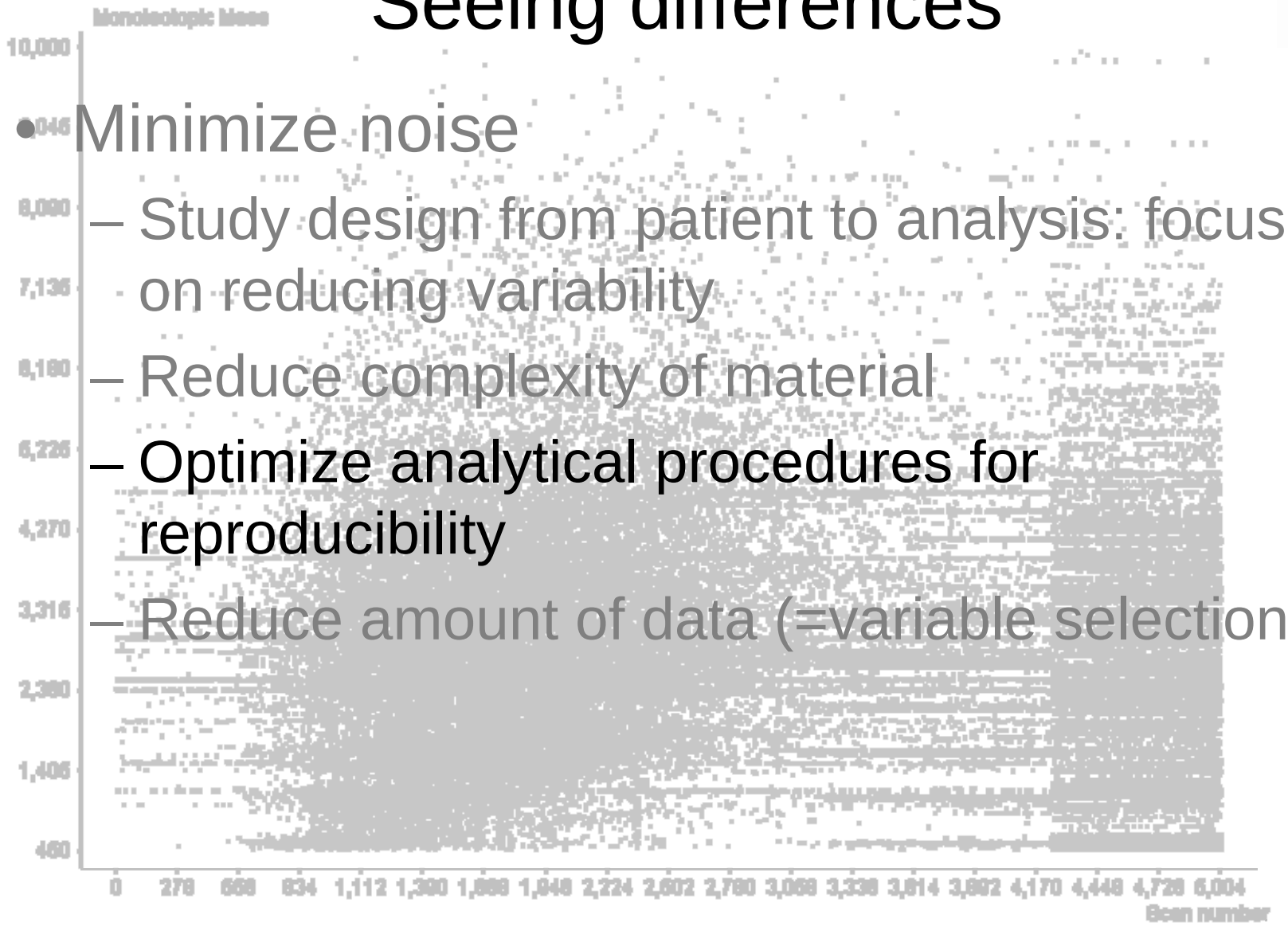
Peak #114

Procedure reduces the cumulated data matrix from 8,745,000 to 137,000 data sets. Thus, every sample is represented by 520 variables, - a reduction of 98% without loosing information

West-Nørager *et al.* Anal Chem ePub 2009

STATENS SERUM INSTITUT

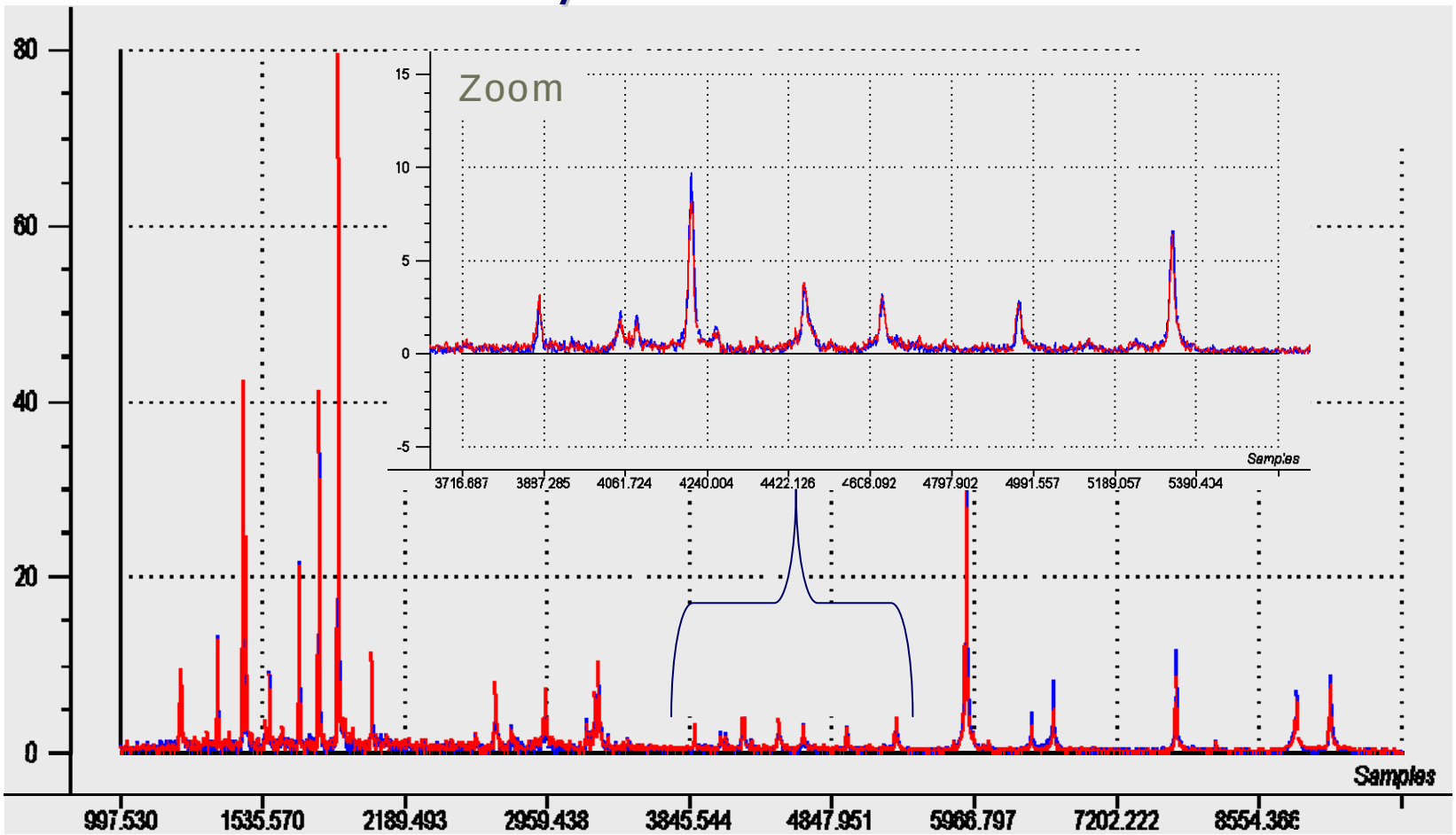
Seeing differences



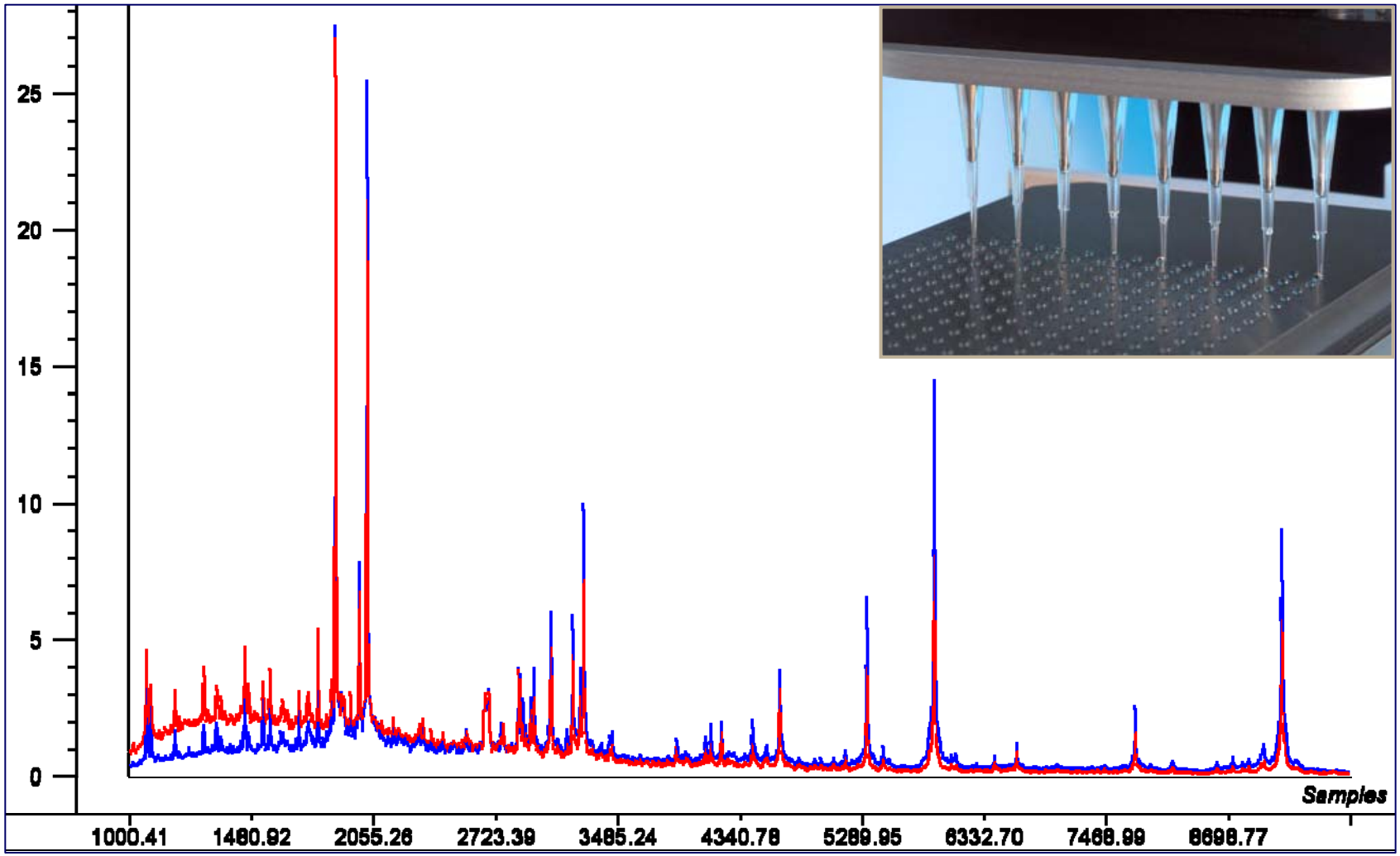
- Minimize noise
  - Study design from patient to analysis: focus on reducing variability
  - Reduce complexity of material
  - Optimize analytical procedures for reproducibility
  - Reduce amount of data (=variable selection)

MALDI-TOF MS reproducibility

- Spot to spot variation in same run
  - R^2 usually ≥ 0.99

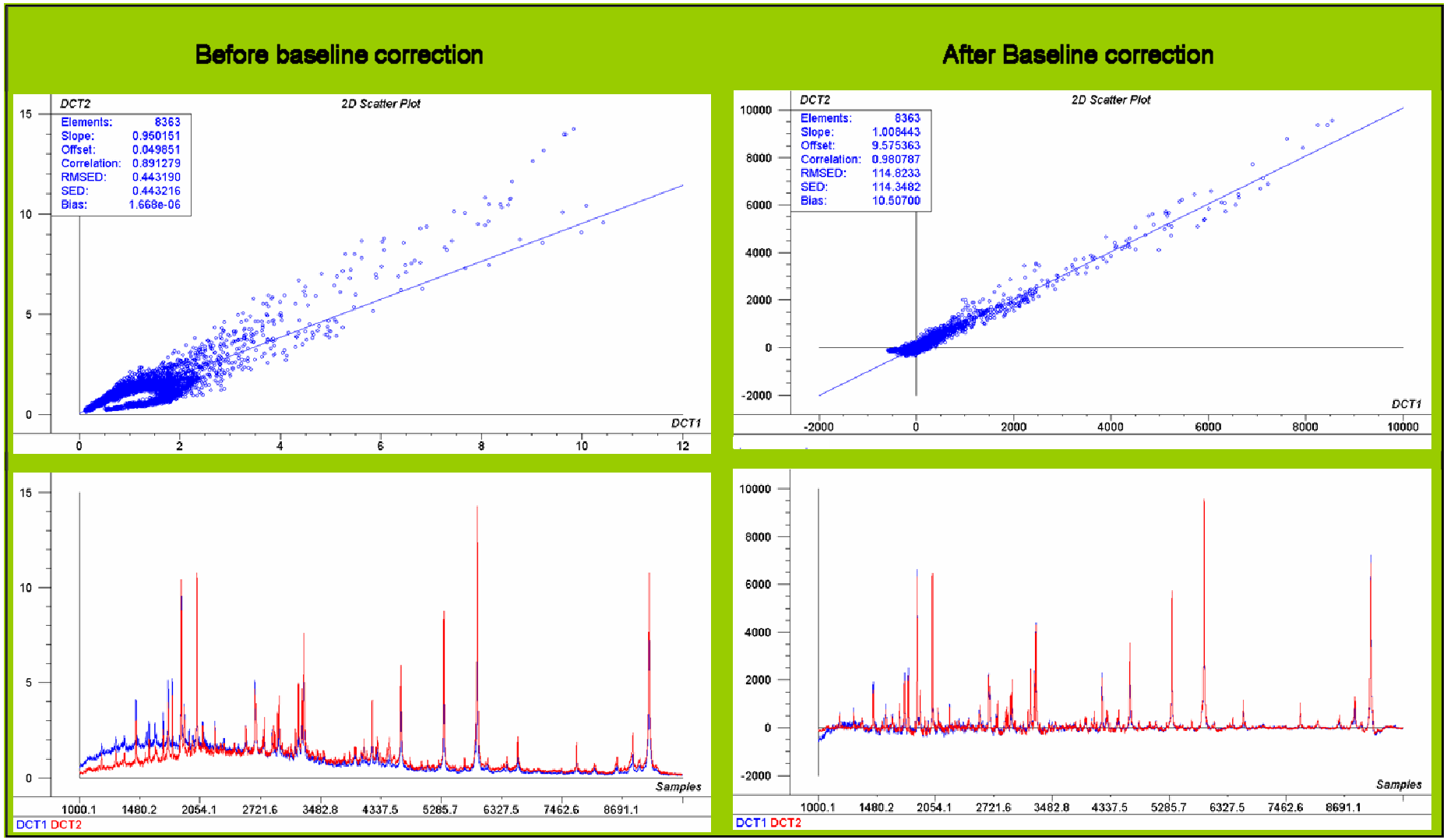


Same sample, different targets

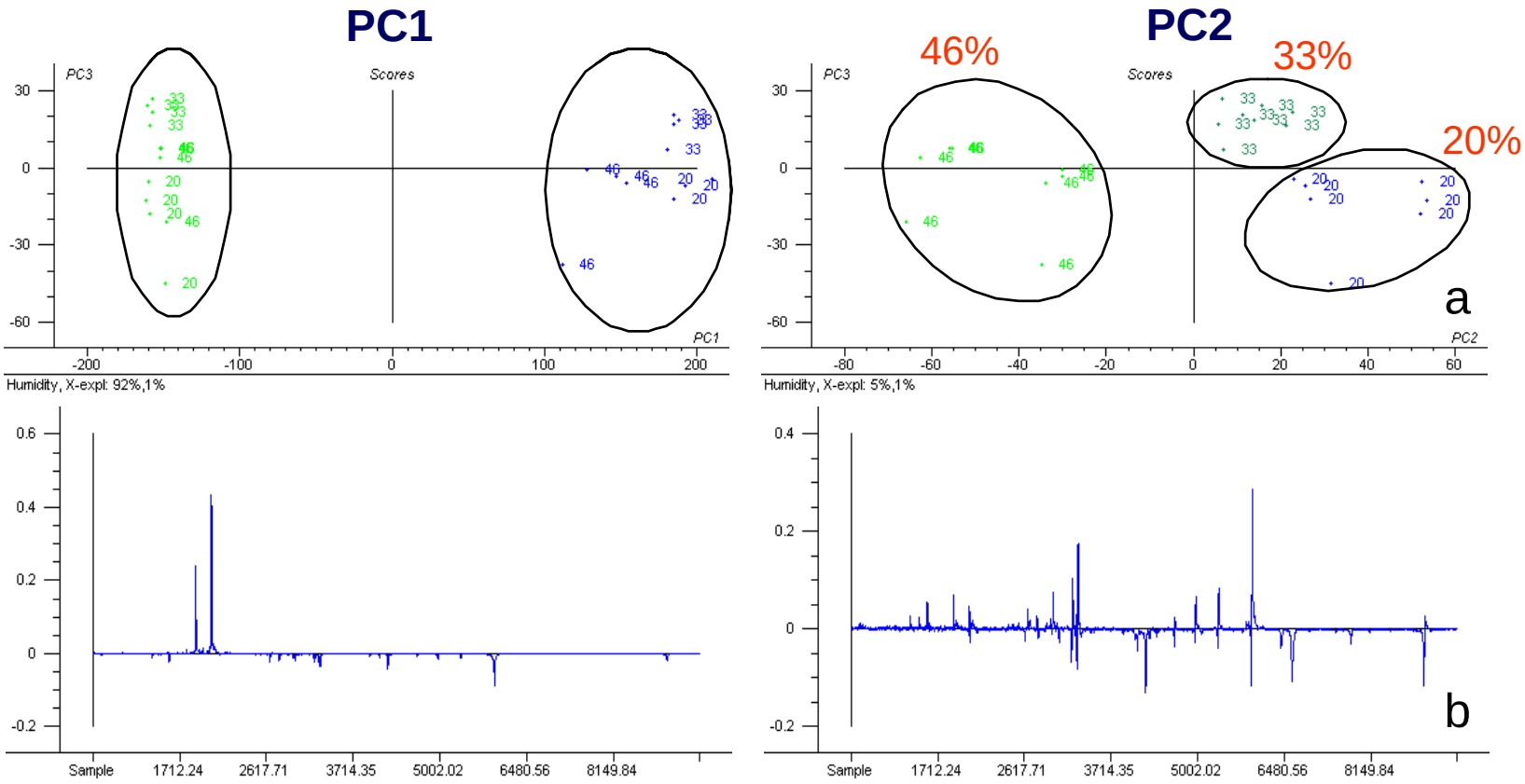


Same sample, different targets

Baseline correction removes most of the target-induced variation



Ambient humidity during sample crystallization affects results

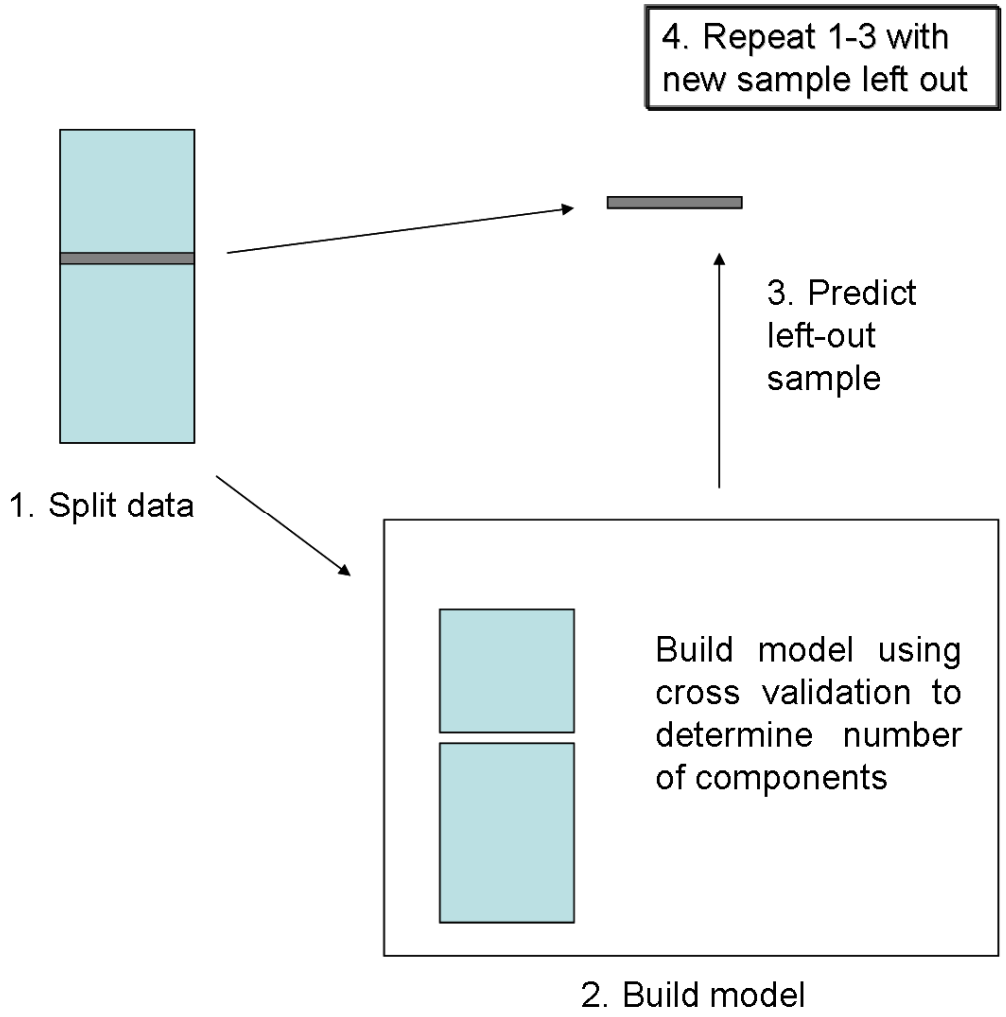


PCA analysis of the crystallization process at three different humidity levels (20%, 33%, and 46%) using replicates of two different samples. Score plots show both the biological difference (PC1) and the different crystallization dependent on the humidity level (PC2). The related loading plots outline the variables accountable for the object pattern in the score plot.

Seeing differences between groups using analyses yielding multiple variables

- Overfitting is an obvious risk – too many data and too few samples
  - We need at least 10 times more samples than features!
  - Look for univariate effects
  - Bonferroni-correct significance levels (p/N)
  - Use PCA first or other unsupervised method
  - Do permutation analysis
  - Independently train, cross-validate, and validate

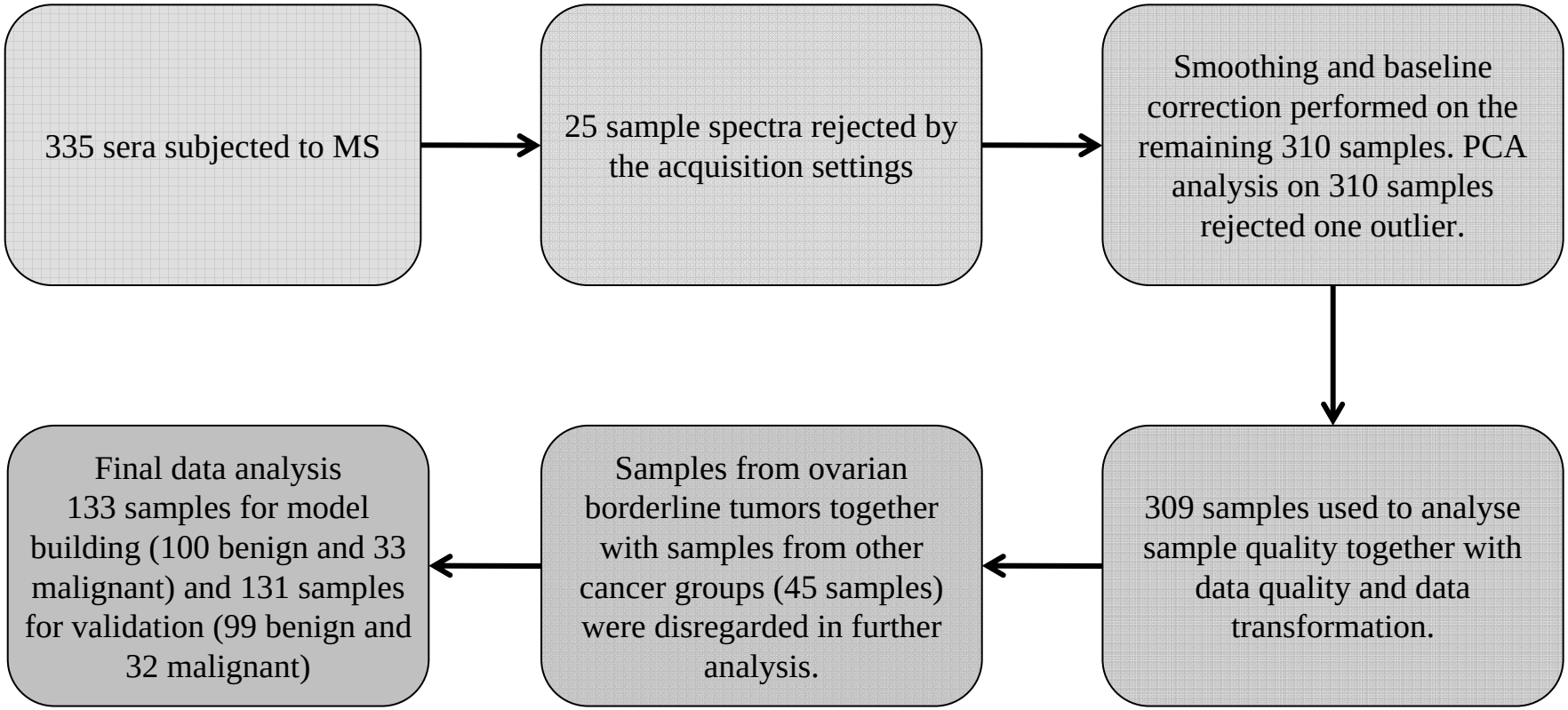
Cross-Model-Validation is more stringent than usual *leave-one-out* cross validation for testing model



West-Nørager *et al., Anal Chem. 2009 ePub*

STATENS SERUM INSTITUT

Pelvic Mass Study
Serological biomarkers for ovarian cancer

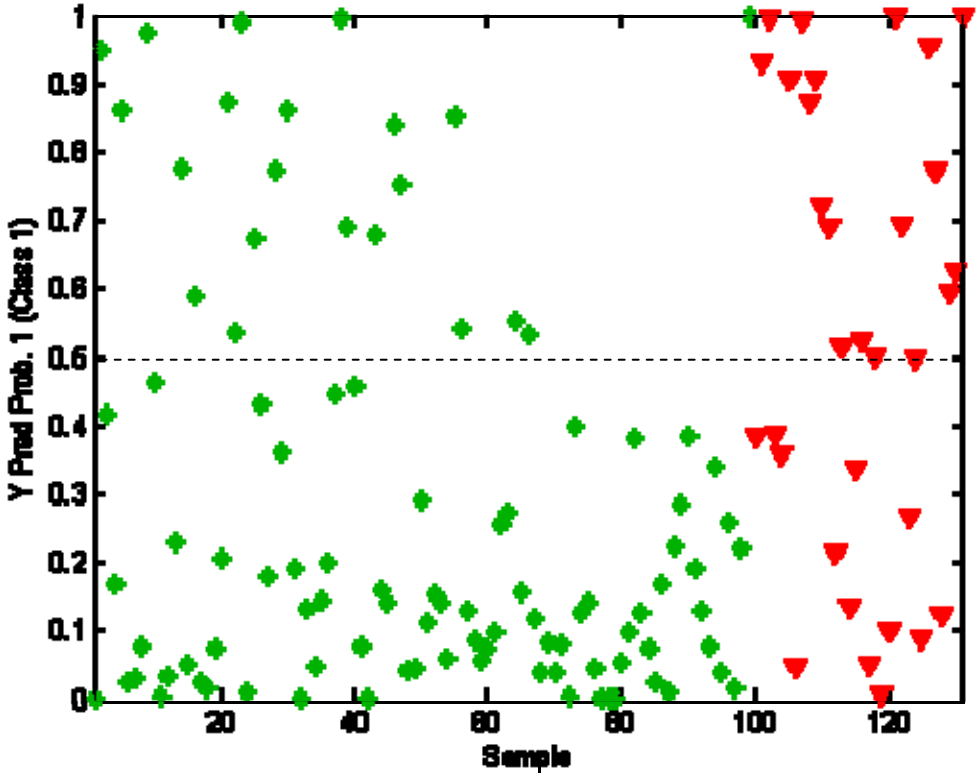


STATENS SERUM INSTITUT

Pelvic Mass Study
Serological biomarkers for ovarian cancer

Overview of the performance of the final model

Independent validation sample set of 99 benign (green asterisks) and 32 malignant (red triangles) from the PLS-DA model – 29 variables (26 peaks)



The final model discriminates malignant from benign conditions in 68% of the cases with 56% sensitivity and 79% specificity.

Model / Correctly predicted	Benign	Malignant
Cross validation (CV)	86 %	74 %
Model prediction (Final validation)	79 %	56 %

West-Nørager et al. Anal. Chem. 2009

Clinical proteomics

Goal: Measurements that discriminate and characterize disease vs. health

What the Analyst wants: Comprehensible, quantitative, and reproducible measurements of the contents of complex biosamples

What the Statistician wants: The best descriptors for maximizing the distinction between two or more groups of samples

What the Physician wants: Laboratory tests that increase life expectancy of patients

Acknowledgements

Mikkel West-Nørager
Estrid Høgdall
Claus Høgdall

Rasmus Bro
Martin R.Larsen
Søren Skov Jensen

Justyna Czarna Bahl
Keld Poulsen
Mark Lim, NIH