Batch effects in large-scale proteomic studies: diagnostics and correction

Jelena Čuklina[1,2,3*], Chloe Lee[1], Evan G. Williams[1], Tatjana Sajic[1], Ben Collins[1], María Rodríguez-Martínez[3], Patrick Pedrioli[1], Ruedi Aebersold[1,4]

[1]ETH Zurich, Institute of Molecular Systems Biology, Zurich, Switzerland, [2]Ph.D. Program in Systems Biology, University of Zurich and ETH Zurich, Zurich, Switzerland,
[3]IBM Zurich Research Laboratory, Rüschlikon, Switzerland, [4]Faculty of Science, University of Zurich, Zurich, Switzerland
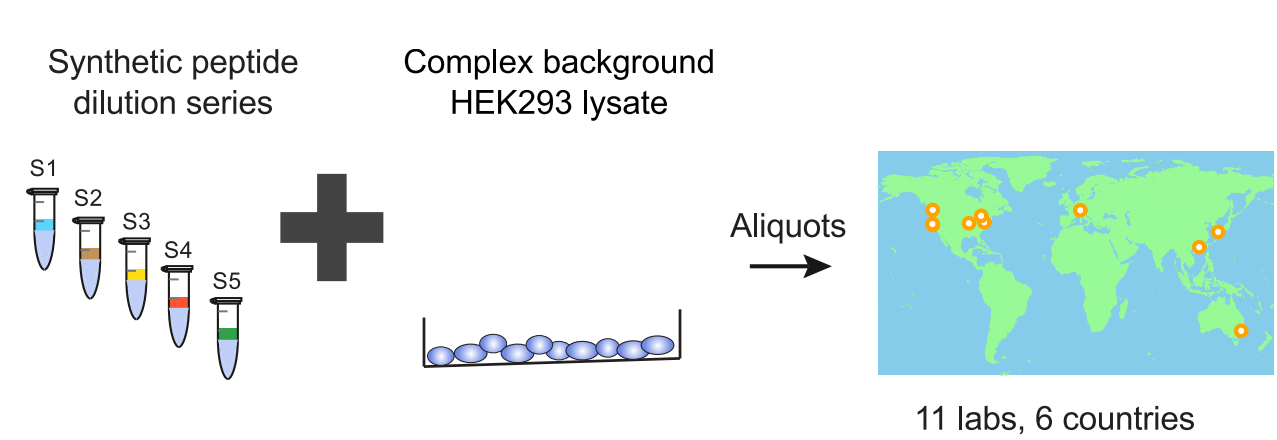
Motivation

In large-scale proteomic studies, logistics restrict the sample number that can be processed in one batch. This inevitably leads to systematic technical bias, known as batch effects. In this study, we analyse batch effects in proteomics. We test existing batch correction tools for their suitability in proteomic studies and introduce new methods for correction and quality control.
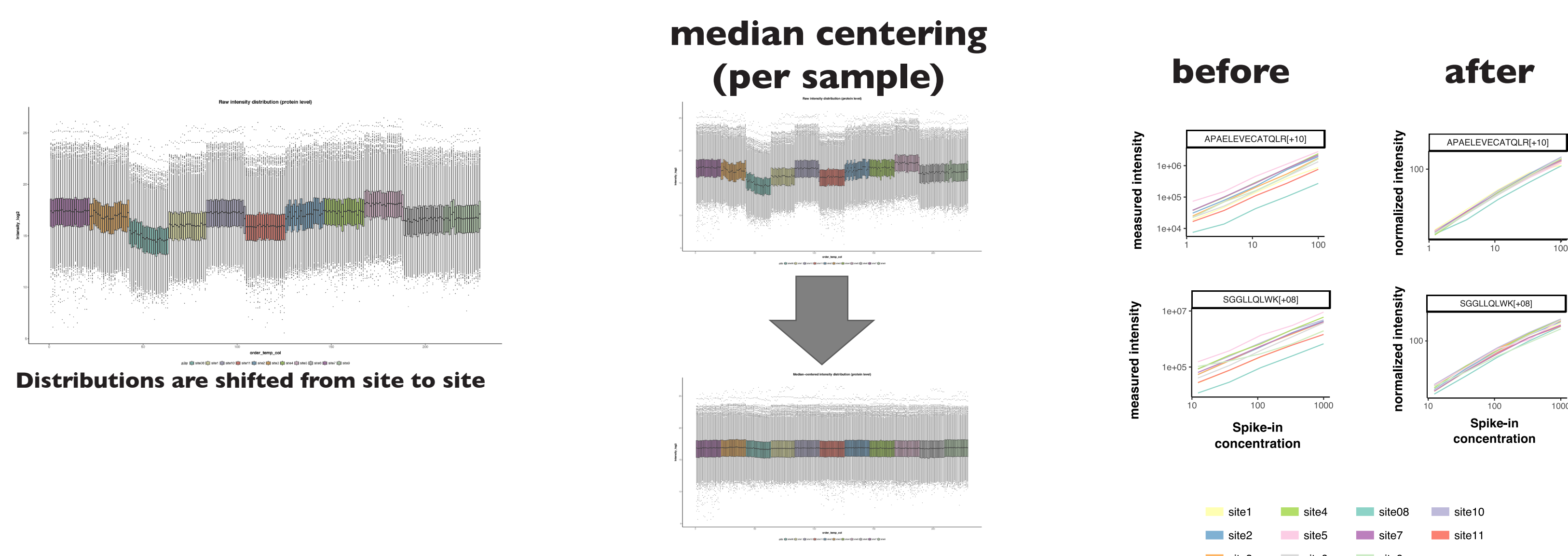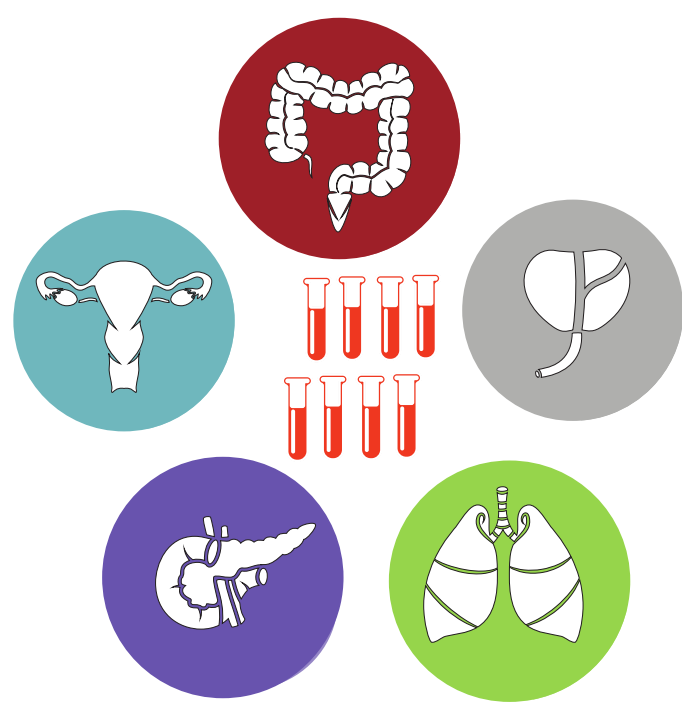
InterLab [1]: 229 samples

Batch effects:
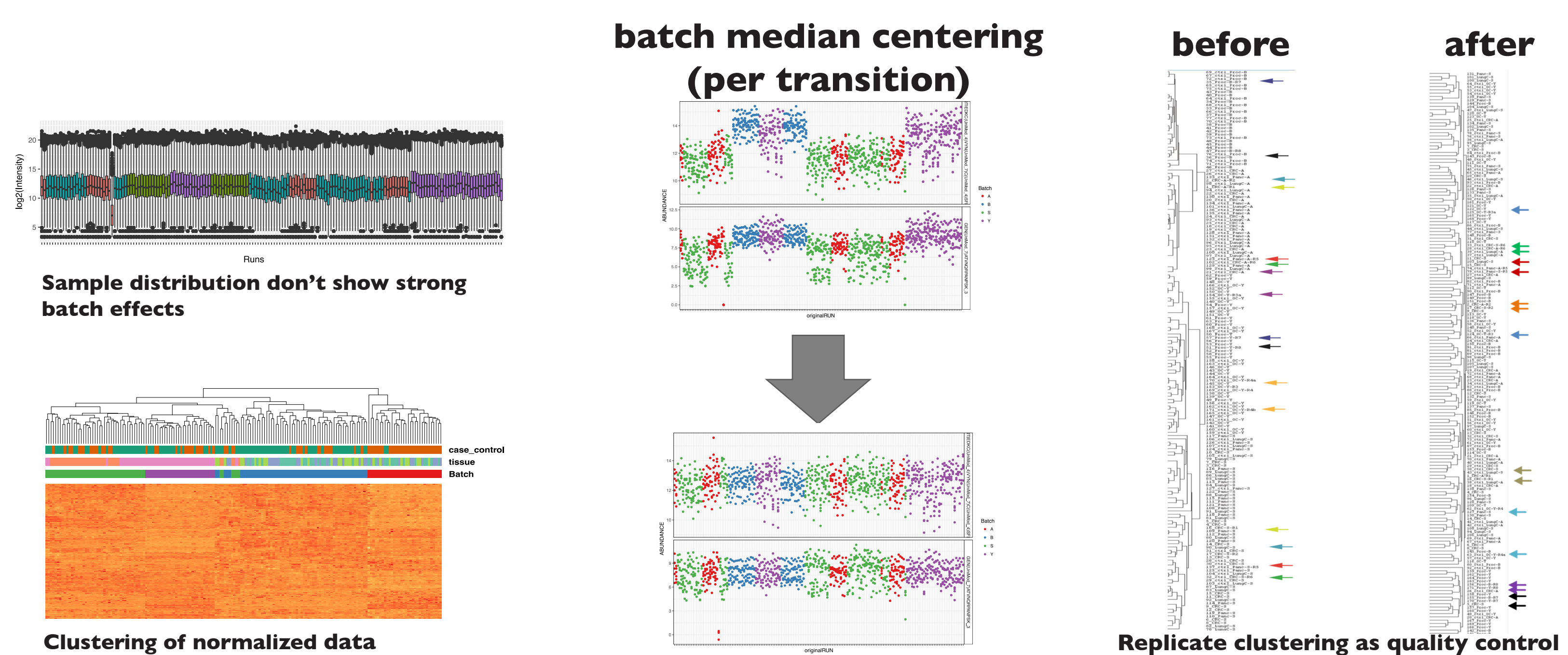Site of measurement
Signal of interest:
Spike-in concentration
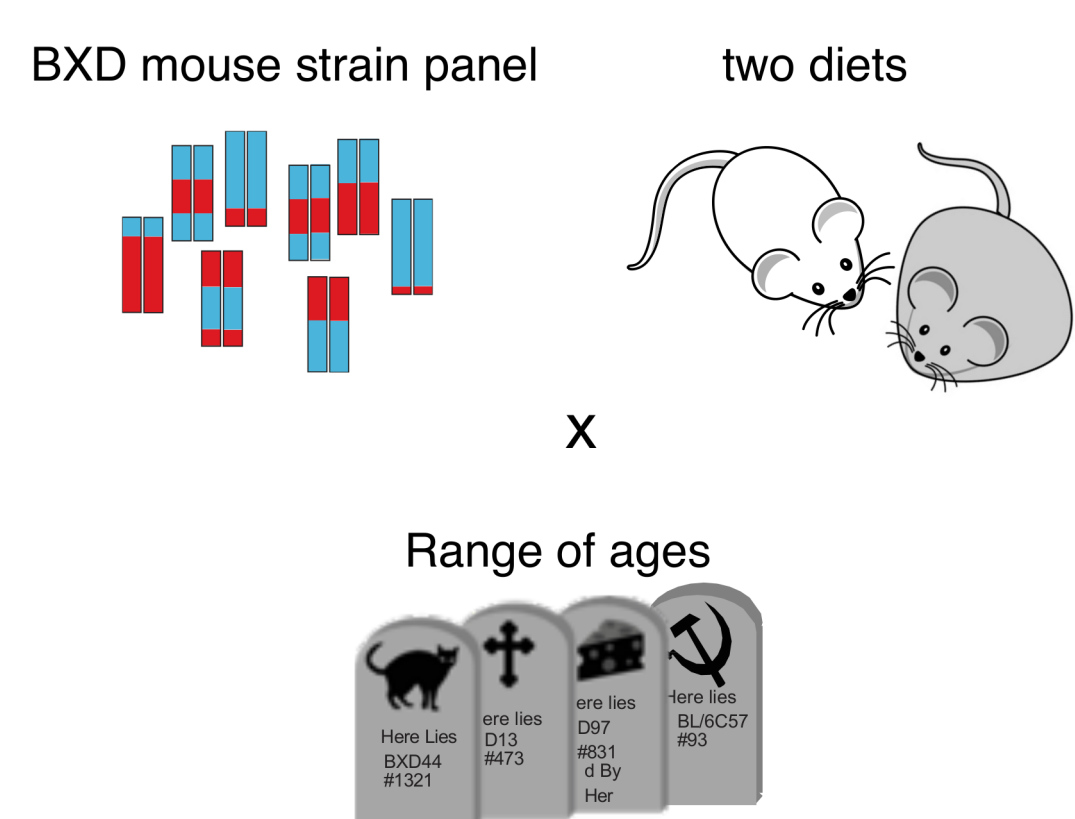
PanCancer [2]: 171 sample

Batch factors: digestion
Biological factors:
1. Tissue type
2. Case/control

Aging mice: 371 sample

Batch factors:
1. Digestion
2. MS batch
3. MS drift
Biological factors:
1. Diet   3. EarTag
2. Strain   4. Age

Batch effects analysis pipeline

Diagnose → **Correct** → **Control**

InterLab

median centering (per sample)

Distributions are shifted from site to site

before / **after**

site1 site4 site08 site10
site2 site5 site7 site11
site3 site6 site9

PanCancer

Sample distribution don't show strong batch effects

Clustering of normalized data

batch median centering (per transition)

before / **after**

Replicate clustering as quality control

Aging mice

LOESS + median centering (per peptide)

Raw data
1146_ADVTPADFSEWSK_3   1997_DGLDAASYYAPVR_2

Normalized data
1146_ADVTPADFSEWSK_3   1997_DGLDAASYYAPVR_2

NEW Fitting LOESS curve
1146_ADVTPADFSEWSK_3   1997_DGLDAASYYAPVR_2

Aligning the medians
1146_ADVTPADFSEWSK_3   1997_DGLDAASYYAPVR_2

Corrected data
1146_ADVTPADFSEWSK_3   1997_DGLDAASYYAPVR_2

Principal Component Analysis   Principal Variance Component Analysis

before / **after**

Distribution of sample correlation

Haao protein: Quantitative trait locus (QTL)

	Raw data	after normalization	after normalization and batch effect correction
Number of QTLs	33	72	87
increase		+118%	+164%

Functions used in analysis implemented as

R package "proBatch"

References

1. Collins, B. C., et al. Multi-Laboratory Assessment of Reproducibility, Qualitative and Quantitative Performance of SWATH-Mass Spectrometry." Nature Communications, 2017
2. Sajic, T., et al. Similarities and Differences of Blood N-Glycoproteins in Five Solid Carcinomas at Localized Clinical Stage Analyzed by SWATH-MS. Cell Reports, 2018

Summary

1. Analysis of batch effects in three large-scale proteomic datasets
2. Implementation of the workflow as 'proBatch' R package.
3. Development of a new method for MS signal drift based on LOESS curve fitting.
4. Introduction of metrics for quality control of batch effects correction.

ETH zürich   FNSNF   NCCR PrECISE   life science zurich   IBM Research