MSSTATS

Meena Choi, Ting Huang, Olga Vitek

College of Science

College of Computer and Information Science

Northeastern University

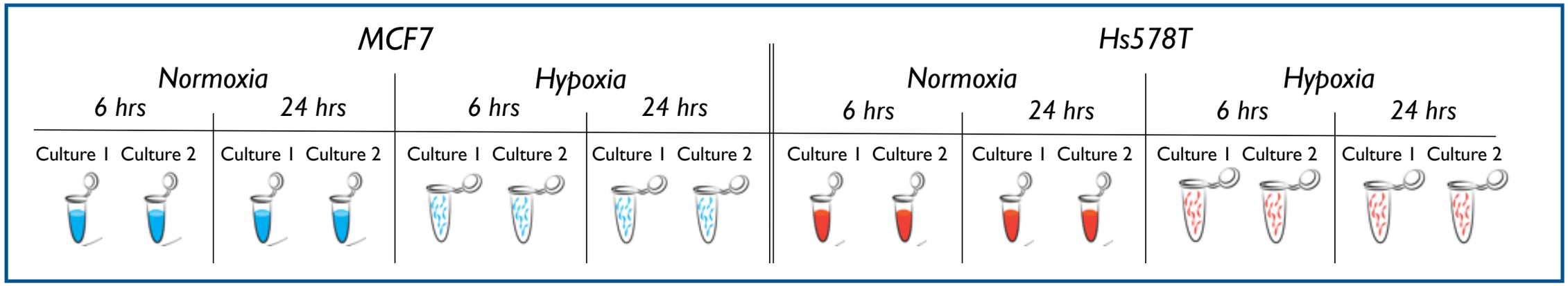
SCHEDULE

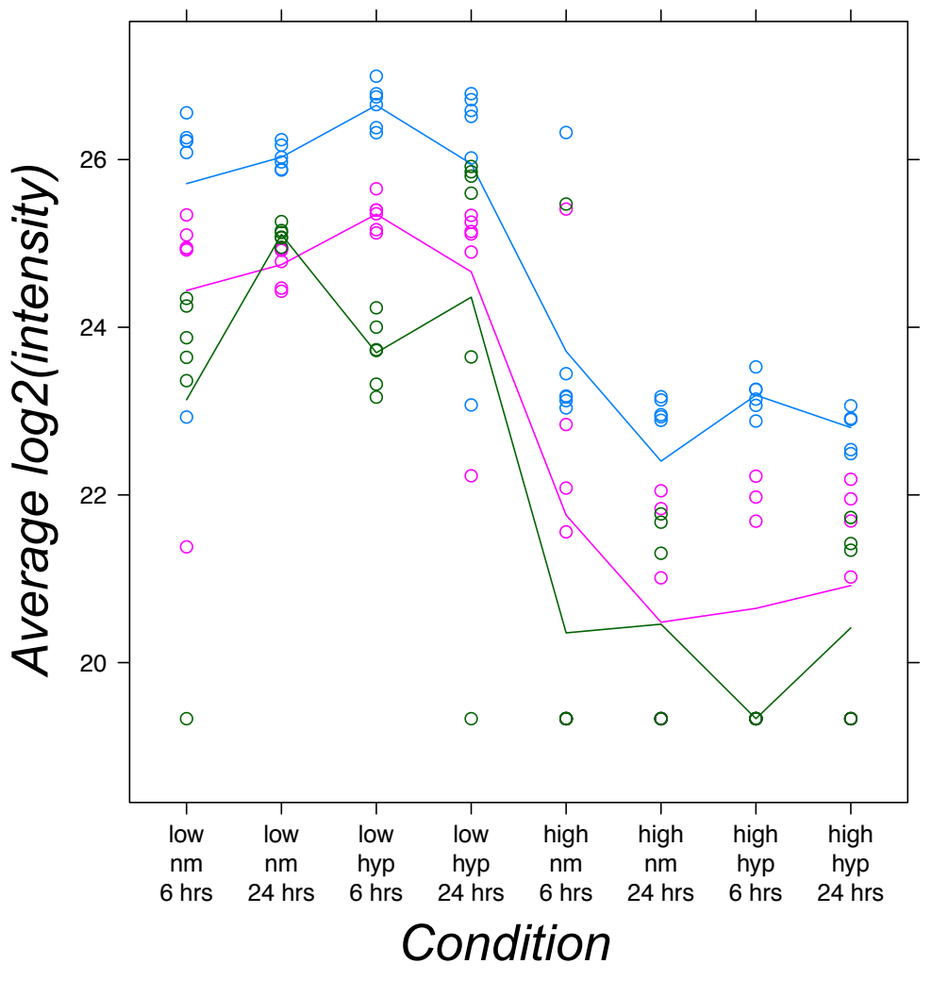
	March 10	March 11
		Skyjam
9:00-9:30	Section1 : Lecture : motivating example, experimental design	Section6 : Differential Analysis of DDA Data with Skyline
9:30-10:00		
10:00-10:30	break	break
10:30-11:00	Section2 : Introduction to Skyline	Section7: Lecture - Msstats (45 mins)
11:00-11:30	Section3 : Processing DDA Data with Skyline	
11:30-12:00		Section8 : Hands-on : different abundance analysis
12:00-12:30	Lunch	Lunch
12:30-13:00		
13:00-13:30	Section4 : Lecture : statistical inference, multiple testing	Section9 : Manual Inspection of Differential Results with Skyline
13:30-14:00		
14:00-14:30	break	break
14:30-15:00	Section5 : Intro R, data exploration, basic statistics	Section10 : From DDA Quantification to SRM, PRM and DIA
15:00-15:30		
15:30-16:00		

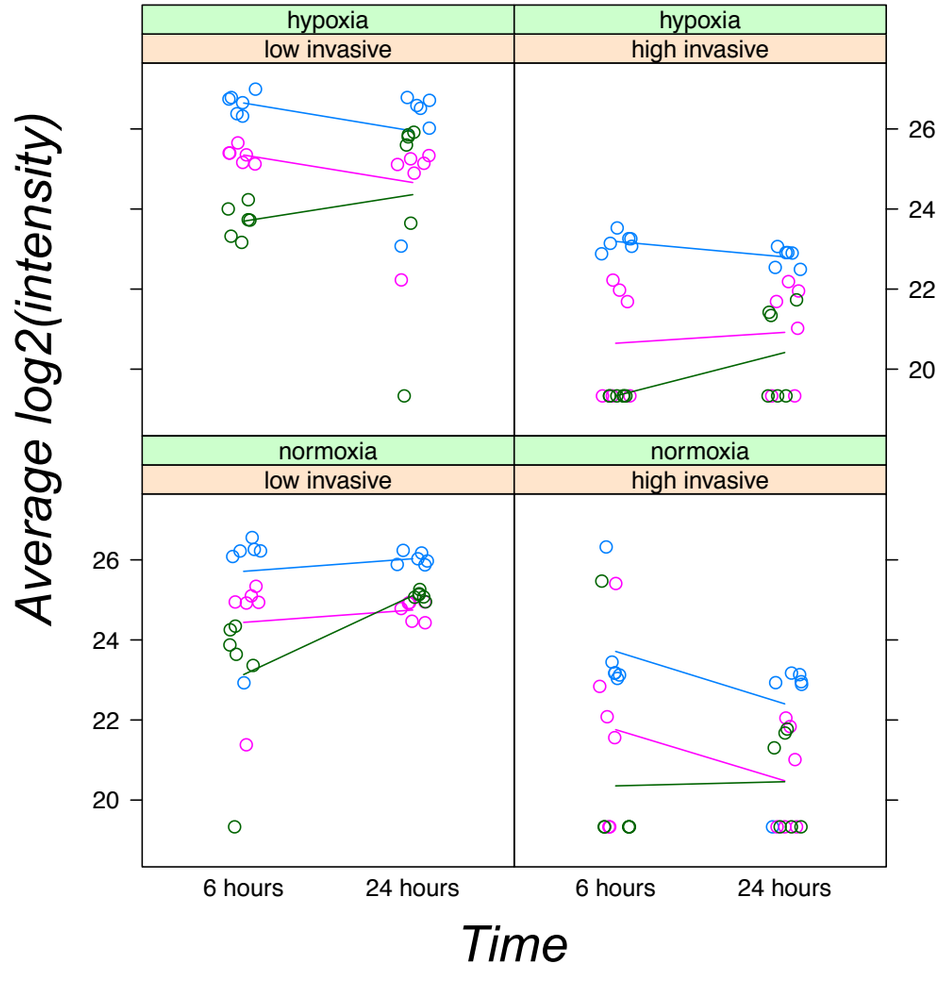
https://goo.gl/EyPprx

EXAMPLE: A LABEL-FREE EXPERIMENT

Question: which proteins change in abundance?





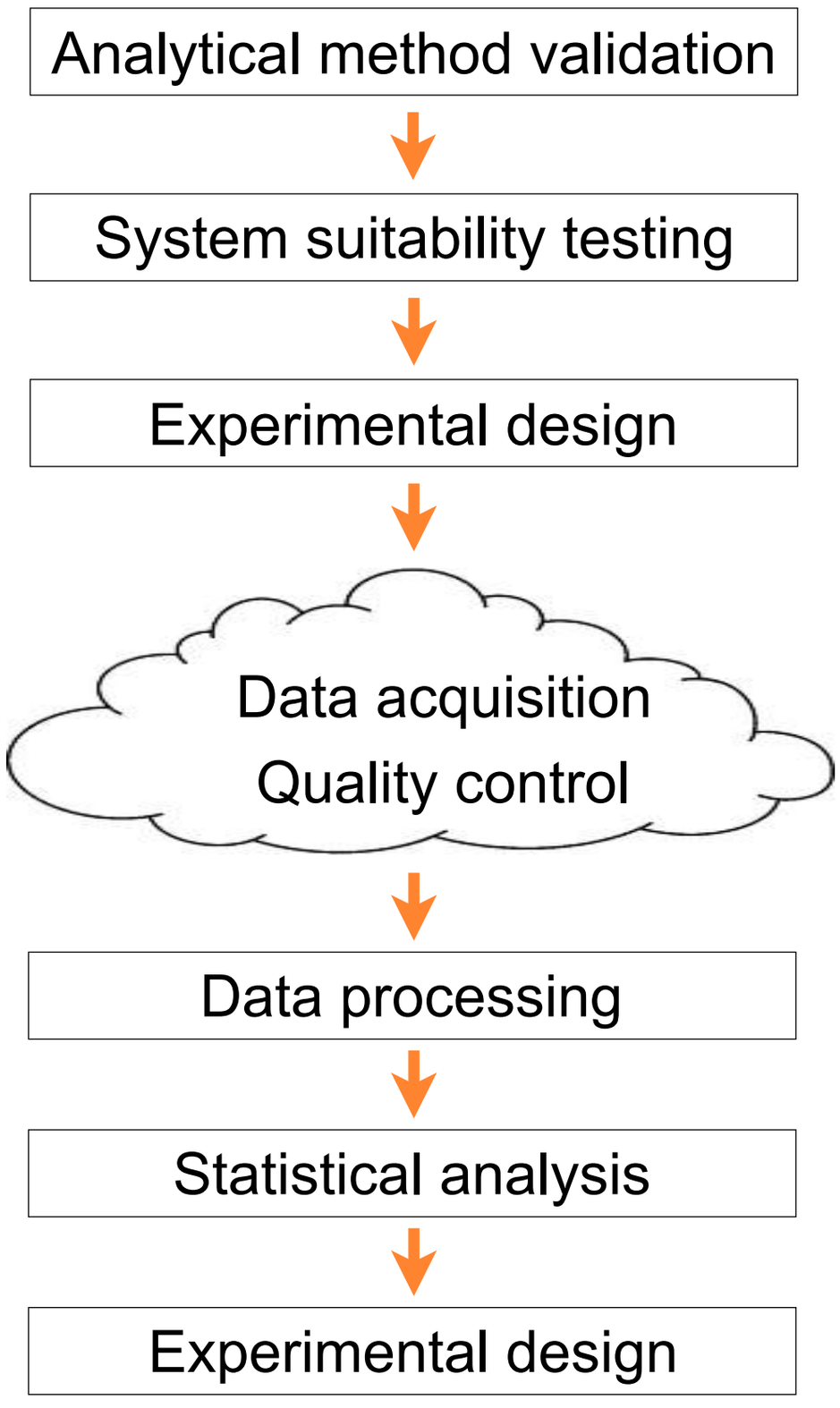


T. Clough et al. *BMC Bioinformatics., 2012*

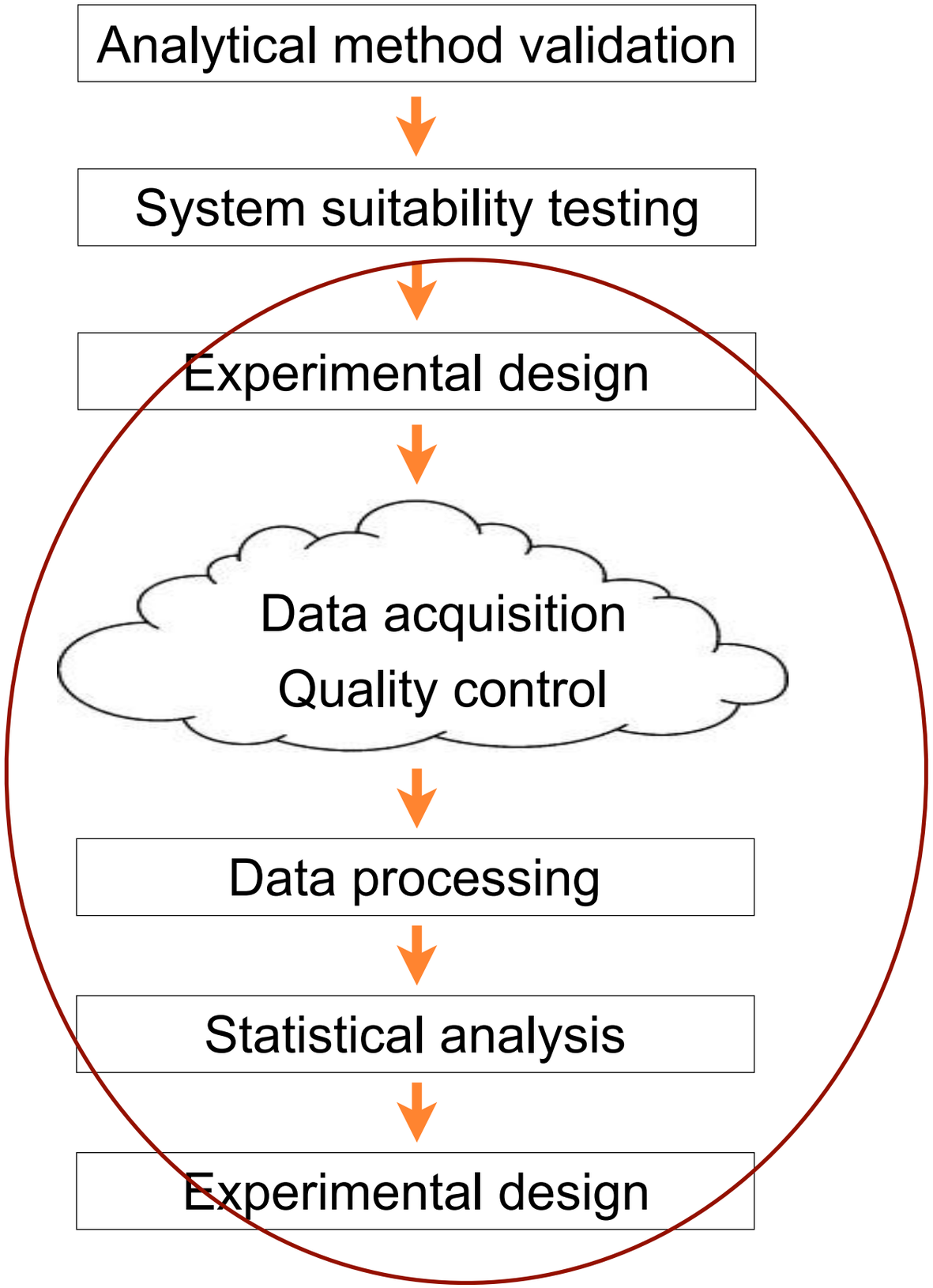
MSSTATS

- Statistical relative quantification of proteins and peptides
  - Which protein changes in abundance?
- Complex experimental designs
  - Multiple conditions, factorial experiments, paired designs, time course
- Chromatography-based quantification
  - Shotgun DDA, targeted SRM, data independent DIA/SWATH, PRM
- Label-free or label-based
  - Simple summaries and models
- Multiple functionalities
  - Data visualization, statistical modeling and inference, sample size
- Free, open-source and inter-operable with other tools
  - Skyline external tool, converters from MaxQuant, Progenesis, OpenMS…

MS EXPERIMENT: STATISTICIAN'S VIEW



MS EXPERIMENT: STATISTICIAN'S VIEW



MS EXPERIMENT: STATISTICIAN'S VIEW

Analytical method validation

↓

System suitability testing

↓

Experimental design

↓

Data acquisition
Quality control

Study goals

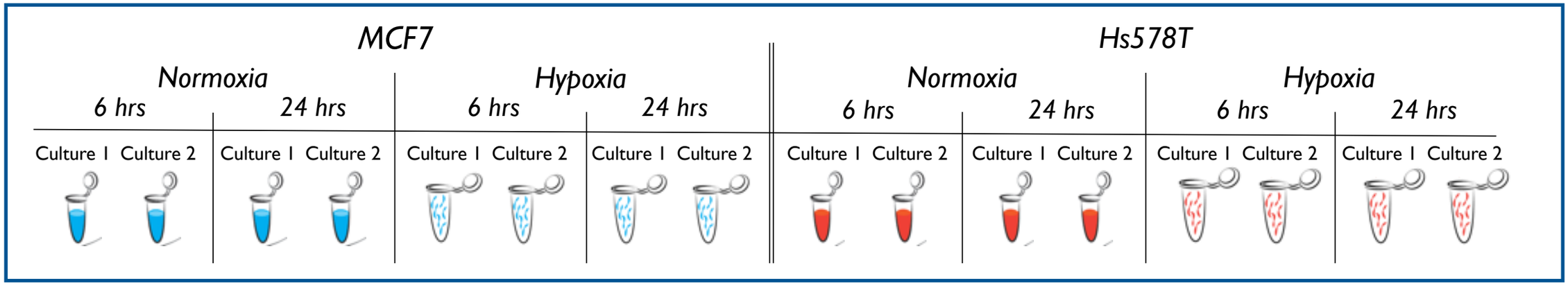
- Exploratory analysis
- Differential analysis
- Biomarker discovery

Biological aspects

- Selection of conditions
- Selection of replicates

Technological aspects

- Sample prep
- Data acquisition
- Randomization



T. Clough et al. *BMC Bioinformatics., 2012*

MS EXPERIMENT: STATISTICIAN'S VIEW

Analytical method validation

↓

System suitability testing

↓

Experimental design

↓

Data acquisition
Quality control

Study goals

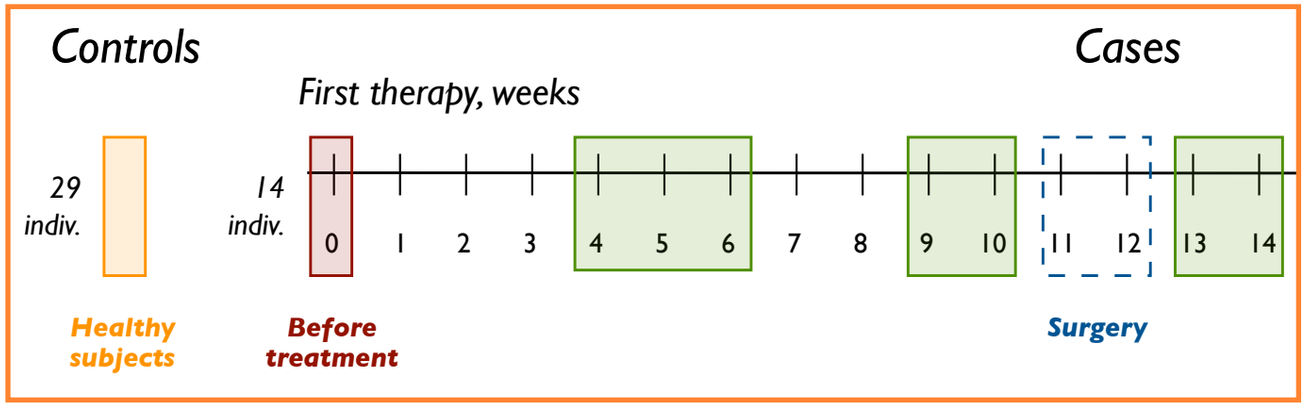
- Exploratory analysis
- Differential analysis
- Biomarker discovery

Biological aspects

- Selection of conditions
- Selection of replicates

Technological aspects

- Sample prep
- Data acquisition
- Randomization



T. Clough et al. *BMC Bioinformatics., 2012*

MS EXPERIMENT: STATISTICIAN'S VIEW

Analytical method validation

↓

System suitability testing

↓

Experimental design

↓

Data acquisition
Quality control

↓

Data processing

↓

Statistical analysis

↓

Experimental design

Input: list of identifies and quantified peaks

- MaxQuant
- OpenMS
- Progenesis
- Skyline
- Spectrnaut
- DIAUmpire
- …

Steps

- Data viz & QC
- Normalization
- Missing & outlying peaks
- Quantify protein in a run
- …

www.msstats.org

INPUT DATA

	A	B	C	D	E	F	G	H	I	J
1	ProteinName	PeptideSequence	PrecursorCharge	FragmentIon	ProductCharge	IsotopeLabelType	Condition	BioReplicate	Run	Intensity
2	ACEA	EILGHEIFFDWELP	3	y3	0	H	1	ReplA	1	66472.3847
3	ACEA	EILGHEIFFDWELP	3	y3	0	L	1	ReplA	1	5764.16228
4	ACEA	EILGHEIFFDWELP	3	y4	0	H	1	ReplA	1	101005.166
5	ACEA	EILGHEIFFDWELP	3	y4	0	L	1	ReplA	1	61.65238
6	ACEA	EILGHEIFFDWELP	3	y5	0	H	1	ReplA	1	90055.4993
7	ACEA	EILGHEIFFDWELP	3	y5	0	L	1	ReplA	1	472.691803
8	ACEA	TDSEAATLISSTID	2	y10	0	H	1	ReplA	1	43506.5425
9	ACEA	TDSEAATLISSTID	2	y10	0	L	1	ReplA	1	217.203553
10	ACEA	TDSEAATLISSTID	2	y7	0	H	1	ReplA	1	68023.0377
11	ACEA	TDSEAATLISSTID	2	y7	0	L	1	ReplA	1	725.284308
12	ACEA	TDSEAATLISSTID	2	y8	0	H	1	ReplA	1	68276.0489
13	ACEA	TDSEAATLISSTID	2	y8	0	L	1	ReplA	1	243.658527

MS EXPERIMENT: STATISTICIAN'S VIEW

Analytical method validation

↓

System suitability testing

↓

Experimental design

↓

Data acquisition
Quality control

↓

Data processing

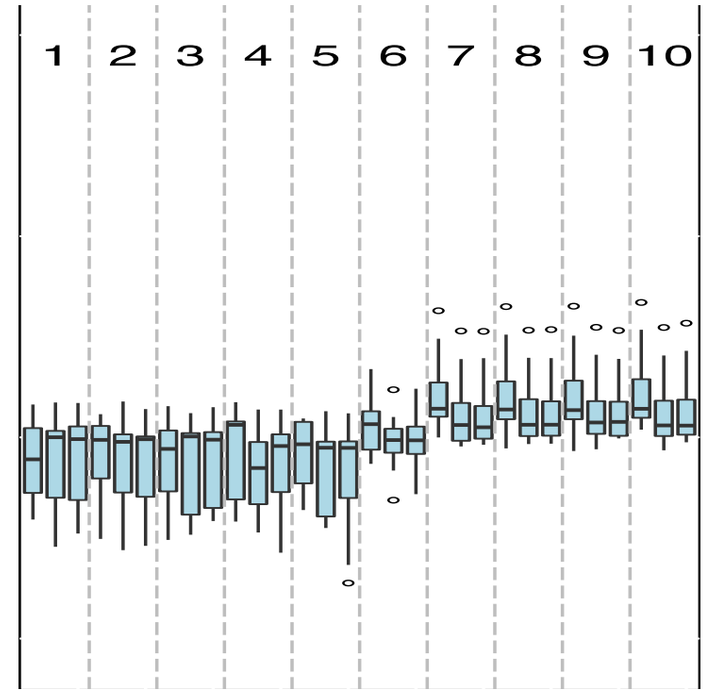
↓

Statistical analysis

↓

Experimental design

NORMALIZATION



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

MS EXPERIMENT: STATISTICIAN'S VIEW

Analytical method validation

↓

System suitability testing

↓

Experimental design

↓

Data acquisition
Quality control

↓

Data processing

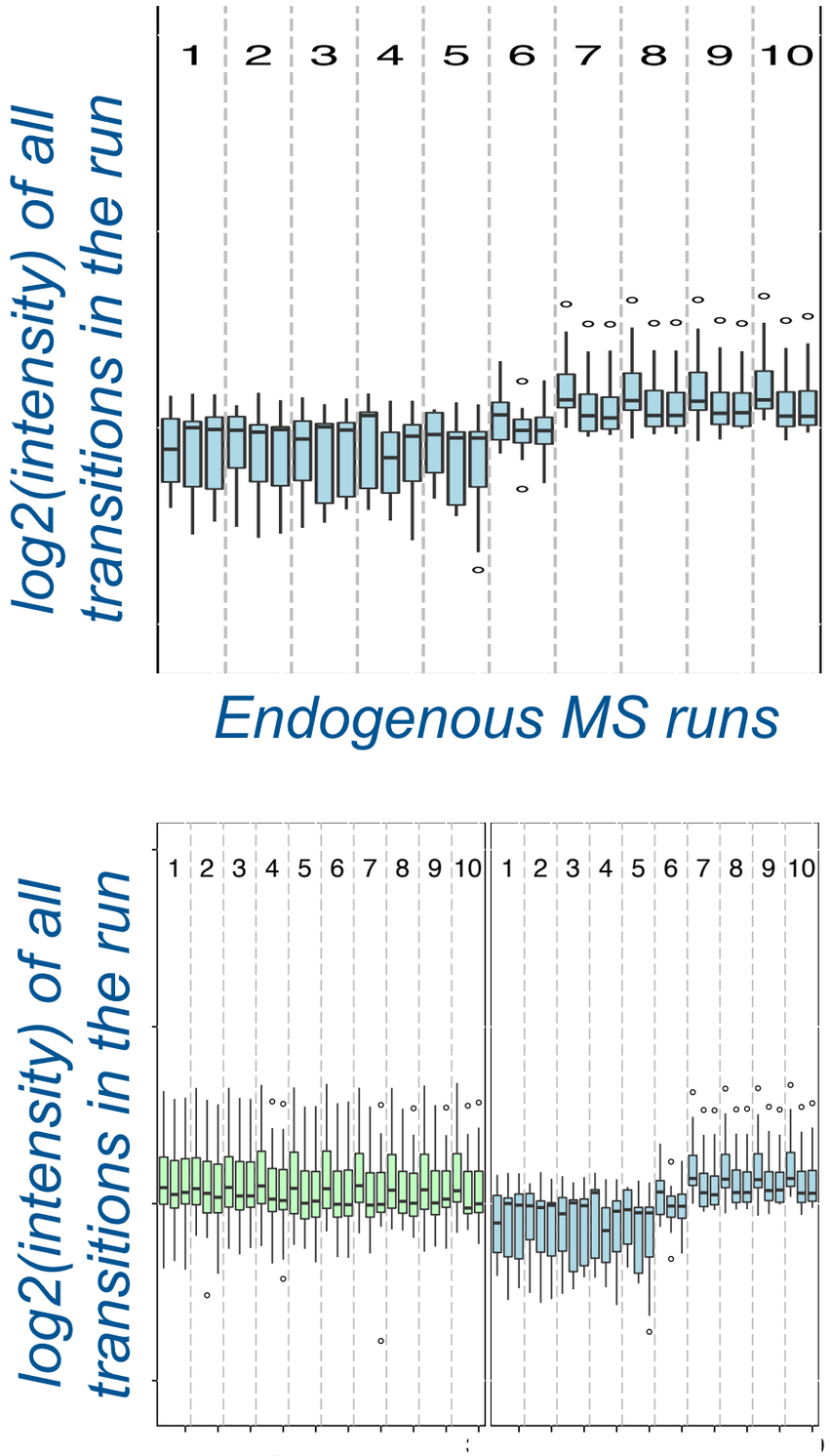
↓

Statistical analysis

↓

Experimental design

NORMALIZATION



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

MS EXPERIMENT: STATISTICIAN'S VIEW

Analytical method validation

↓

System suitability testing

↓

Experimental design

↓

Data acquisition
Quality control

↓

Data processing

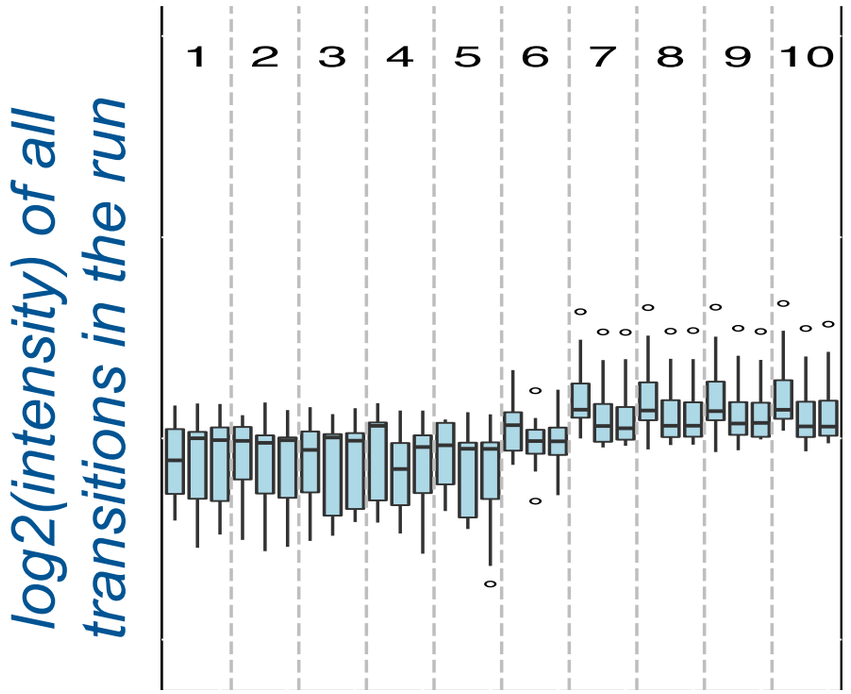
↓

Statistical analysis

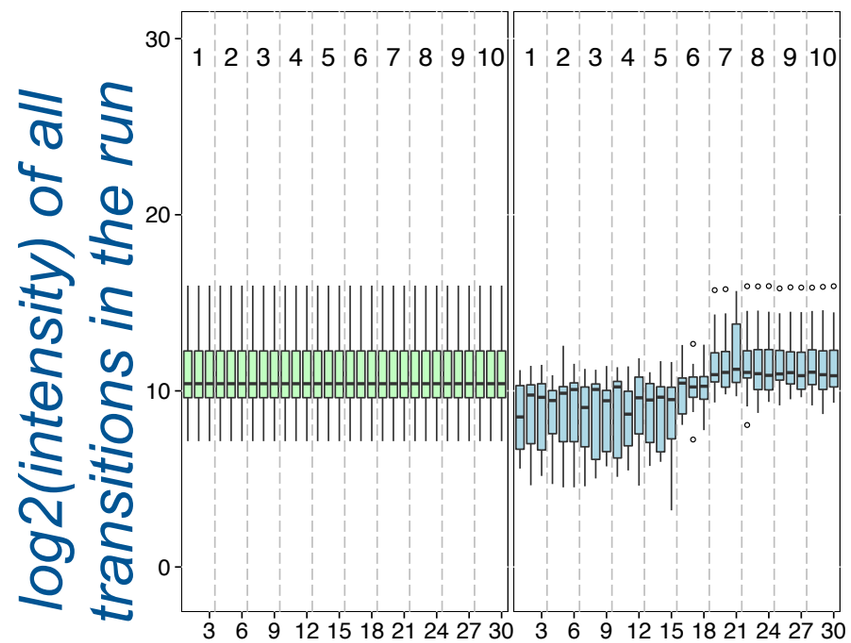
↓

Experimental design

NORMALIZATION



Endogenous MS runs



Reference Endogenous MS runs

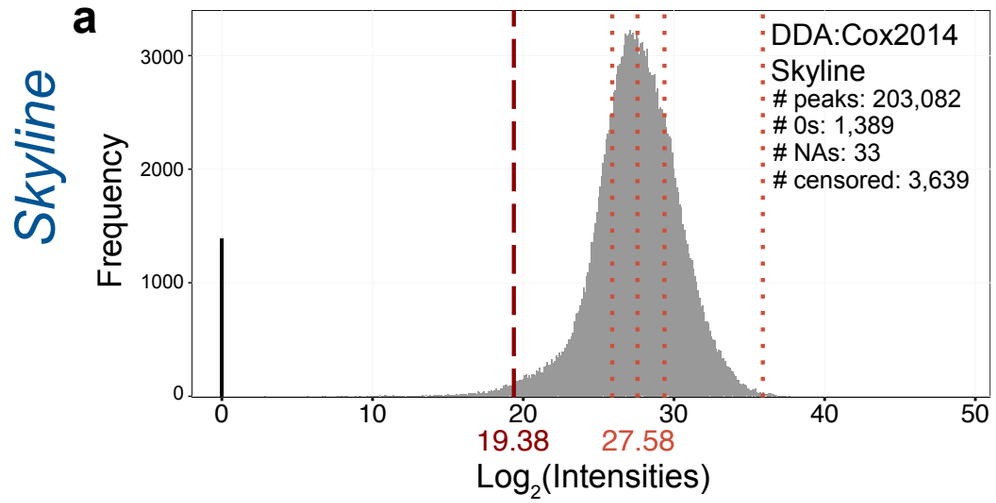
C.-Y. Chang et al. *MCP*, 2012

PROPERTIES OF PEAK INTENSITIES VARY BETWEEN DATA PROCESSING TOOLS

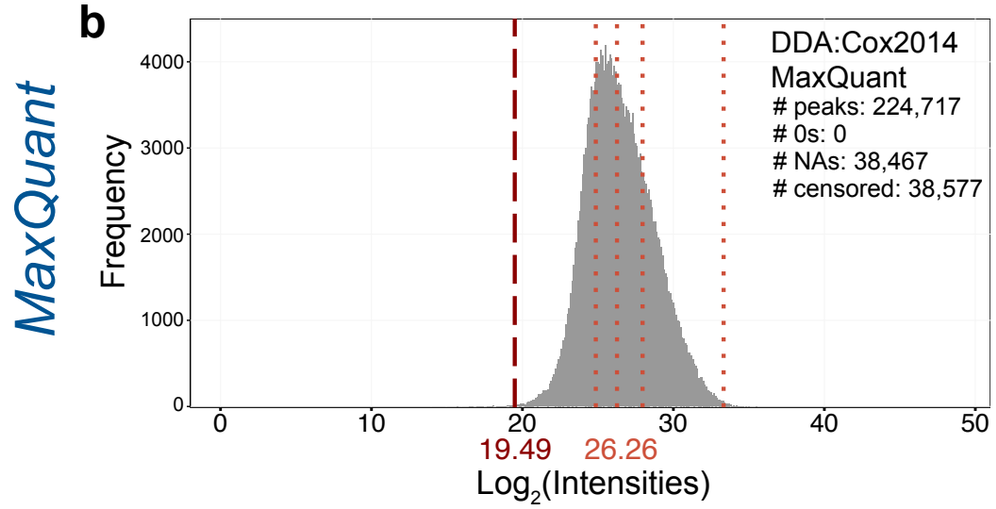
DDA: Cox 2014

DIA: Bruderer 2015

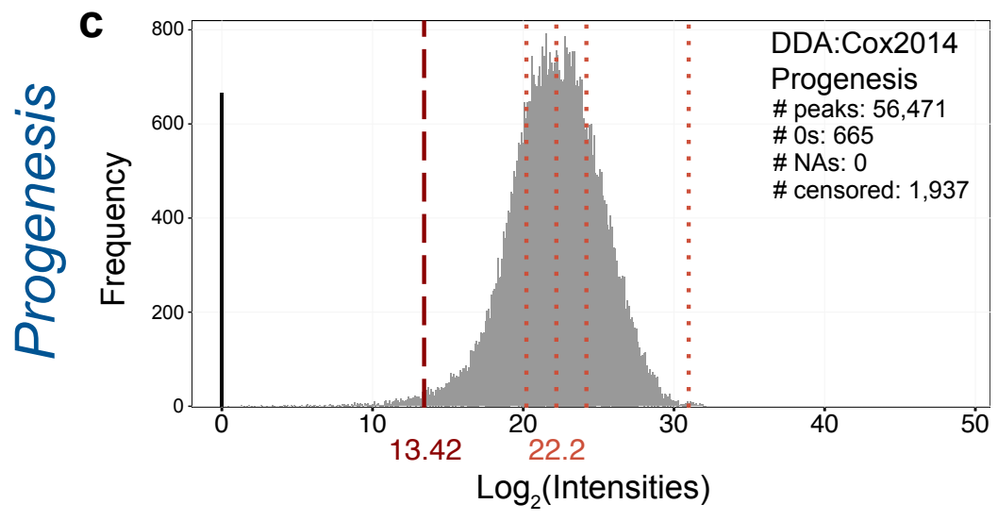
Skyline



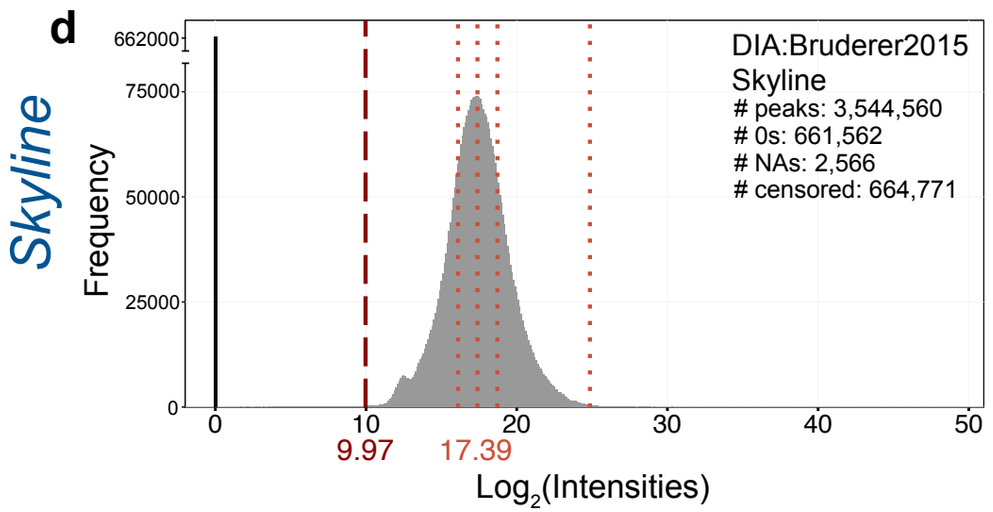
MaxQuant



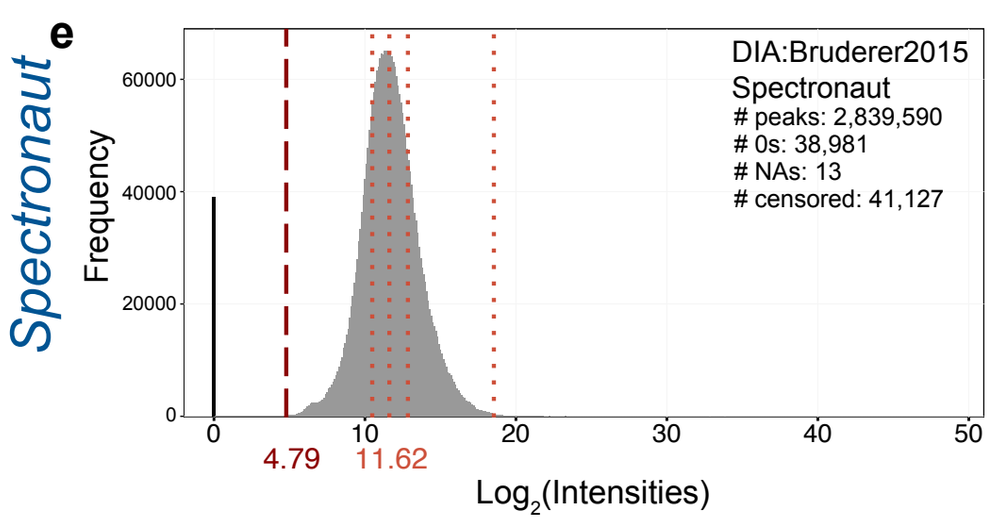
Progenesis

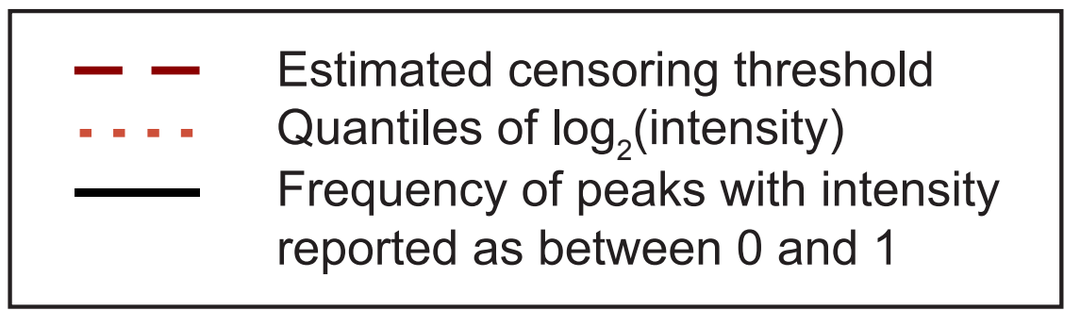


Skyline



Spectronaut





MS EXPERIMENT: STATISTICIAN'S VIEW

Analytical method validation

↓

System suitability testing

↓

Experimental design

↓

Data acquisition
Quality control

↓

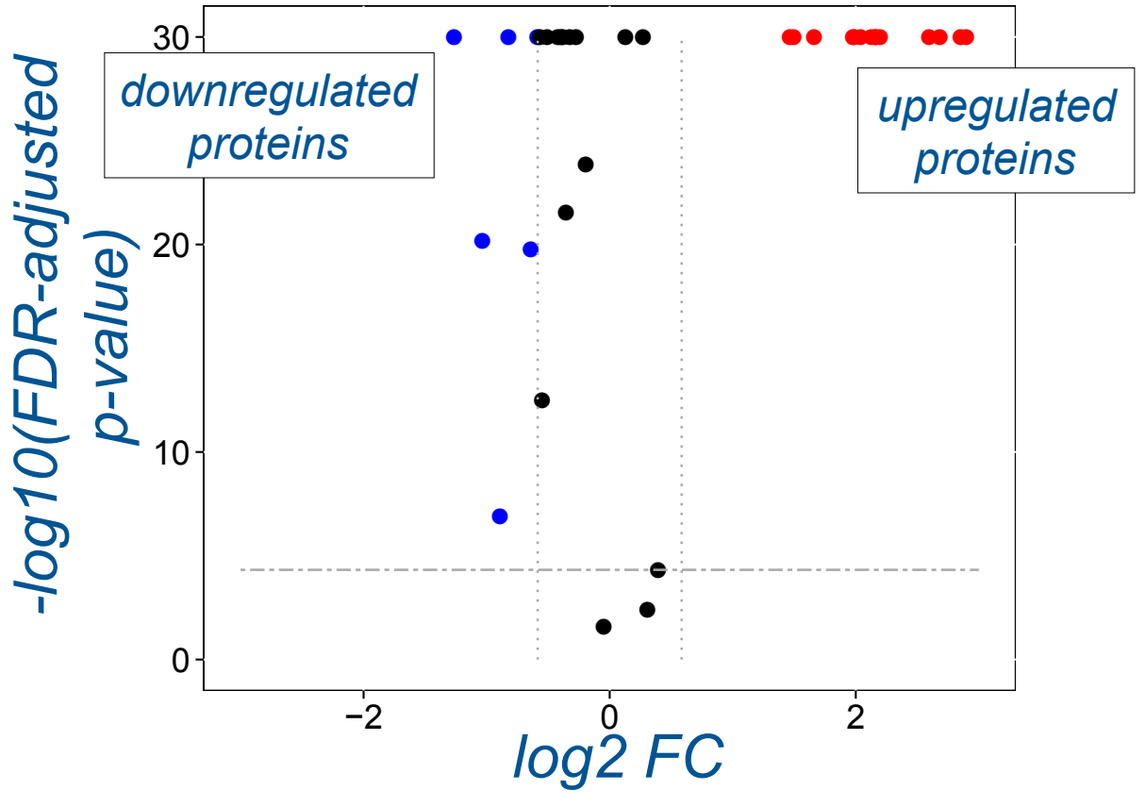
Data processing

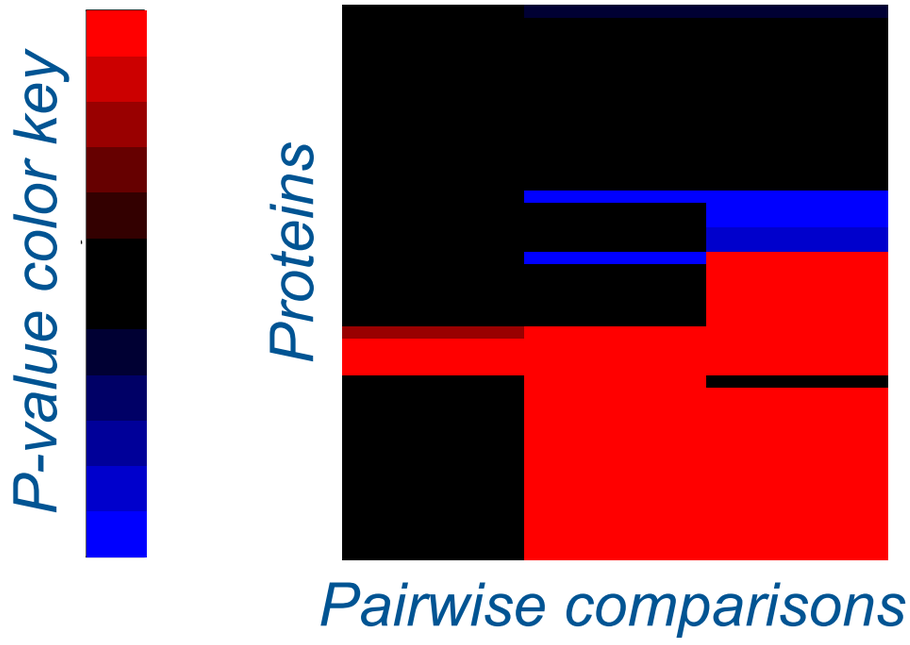
↓

Statistical analysis

↓

Experimental design





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

LINEAR MIXED MODELS

A split plot approach

Whole plot

Subplot	$Condition_1$						...				...	$Condition_I$									
	$Subject_1$			$Subject_2$			...	$Subject_J$			...	$Subject_{(I-1)J+1}$			$Subject_{(I-1)J+2}$			...	$Subject_{IJ}$		
	Run_1	Run_2	Run_3	Run_4	Run_5	Run_6	...	Run_{JK-2}	Run_{JK-1}	Run_{JK}	...	$Run_{(I-1)JK+1}$	$Run_{(I-1)JK+2}$	$Run_{(I-1)JK+3}$	$Run_{(I-1)JK+4}$	$Run_{(I-1)JK+5}$	$Run_{(I-1)JK+6}$	...	Run_{IJK-2}	Run_{IJK-1}	Run_{IJK}
$Feature_1$	y	y	y	y	y	y	...	y	y	y	...	y	y	y	y	y	y	...	y	y	y
$Feature_2$	y	y	y	y	y	**NA**	...	y	y	y	...	y	y	y	y	y	y	...	y	y	y
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
$Feature_L$	y	**NA**	y	**NA**	**NA**	y	...	**NA**	y	y	...	**NA**	y	y	y	y	y	...	y	**NA**	y

Whole plot | Subplot

$$y_{ijkl} = \mu + \underbrace{\mathbf{Condition}_i + \underbrace{\mathbf{Subject(Condition)_{j(i)}}}_{\text{Whole-plot biological variation}}}_{\text{Whole plot}} + \underbrace{\underbrace{\mathbf{Run_{ijk}}}_{\text{Whole-plot technical variation}} + \mathbf{Feature}_l + \underbrace{\mathbf{Run \times Feature_{ijkl}}}_{\text{Subplot error}}}_{\text{Subplot}}$$

where $\sum_{i=1}^{I} \mathbf{Condition}_i = 0$, $\sum_{j=1}^{L} \mathbf{Feature}_l = 0$

$$\mathbf{Subject(Condition)_{j(i)}} \overset{\mathbf{iid}}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2_{\mathbf{Subject}})$$

$$\mathbf{Run_{ijk}} = \psi_{\mathbf{ijk}} \overset{\mathbf{iid}}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2_{\psi})$$

$$\mathbf{Run \times Feature_{ijkl}} = \epsilon_{\mathbf{ijkl}} \overset{\mathbf{iid}}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2_{\epsilon})$$

INTERPRETING CENSORED VALUES

	Run_1	Run_2	Run_3	Run_4	Run_5	Run_6	· · ·	Run_{JK-2}	Run_{JK-1}	Run_{JK}	· · ·	$Run_{(I-1)JK+1}$	$Run_{(I-1)JK+2}$	$Run_{(I-1)JK+3}$	$Run_{(I-1)JK+4}$	$Run_{(I-1)JK+5}$	$Run_{(I-1)JK+6}$	· · ·	Run_{IJK-2}	Run_{IJK-1}	Run_{IJK}
$Feature_1$	y	y	y	y	y	y	· · ·	y	y	y	· · ·	y	y	y	y	y	y	· · ·	y	y	y
$Feature_2$	y	y	y	y	y	$\mathbf{NA}_{rand}$	· · ·	y	y	y	· · ·	y	y	y	y	y	y	· · ·	y	y	y
· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·
$Feature_L$	y	$\mathbf{NA}_{cen}$	y	$\mathbf{NA}_{cen}$	$\mathbf{NA}_{cen}$	y	· · ·	$\mathbf{NA}_{cen}$	y	y	· · ·	$\mathbf{NA}_{cen}$	y	y	y	y	y	· · ·	y	$\mathbf{NA}_{cen}$	y

↓

	Run_1	Run_2	Run_3	Run_4	Run_5	Run_6	· · ·	Run_{JK-2}	Run_{JK-1}	Run_{JK}	· · ·	$Run_{(I-1)JK+1}$	$Run_{(I-1)JK+2}$	$Run_{(I-1)JK+3}$	$Run_{(I-1)JK+4}$	$Run_{(I-1)JK+5}$	$Run_{(I-1)JK+6}$	· · ·	Run_{IJK-2}	Run_{IJK-1}	Run_{IJK}
Summarized	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	· · ·	$\hat{y}$	$\hat{y}$	$\hat{y}$	· · ·	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	· · ·	$\hat{y}$	$\hat{y}$	$\hat{y}$

↓

Model-based inference by whole plot

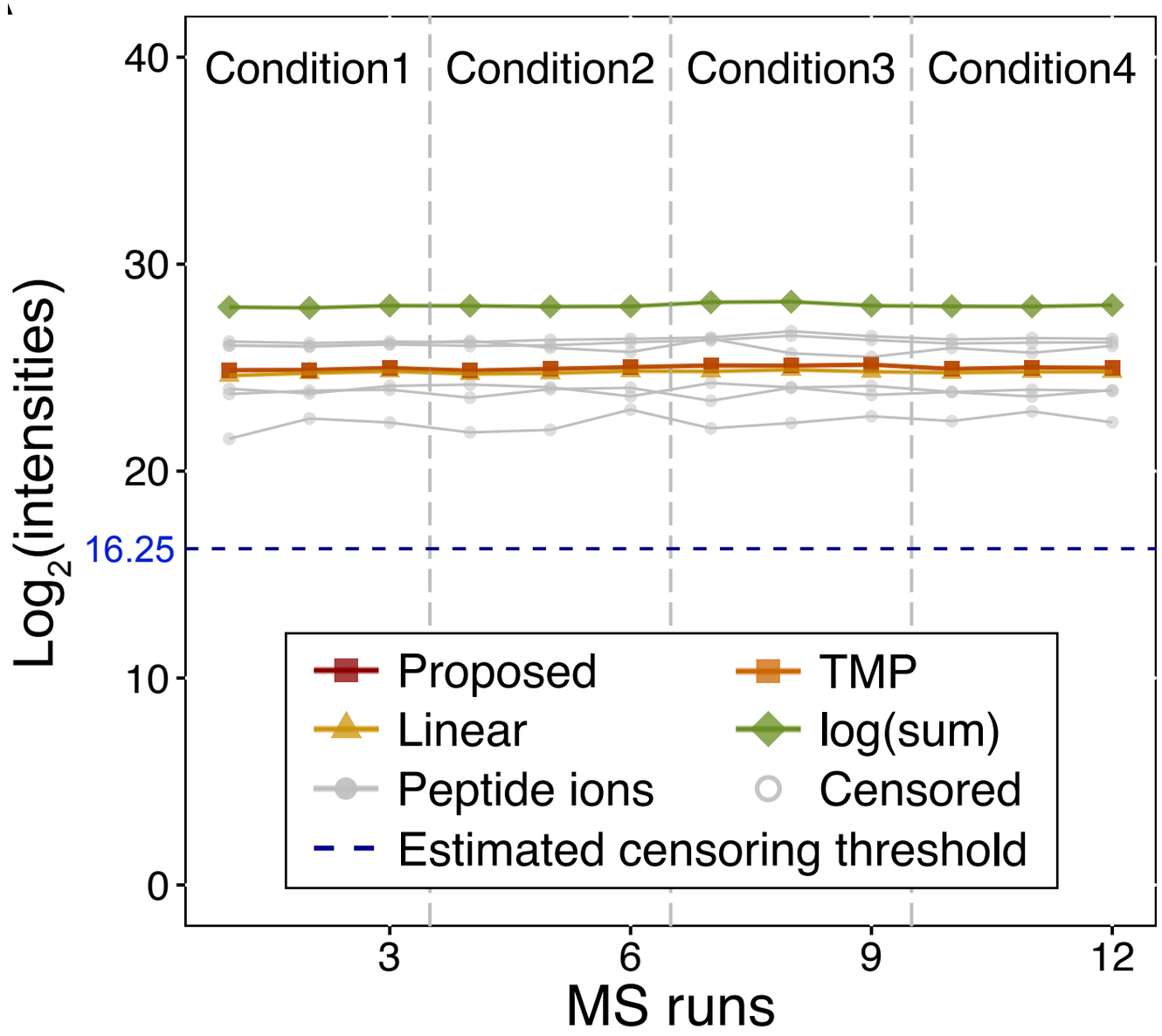
$$\hat{y}_{ijk} = \mu + Condition_i + Subject(Condition)_{j(i)} + \psi_{ijk}, \text{where}$$

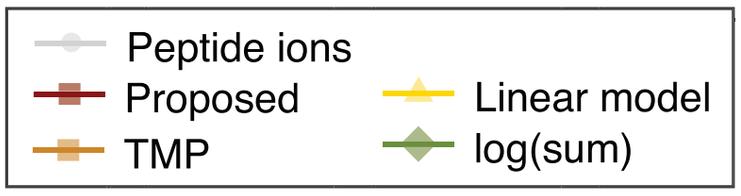
$$\sum_i Condition_i = 0,\ Subject(Condition)_{j(i)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{Subject}),\ \psi_{ijk} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{\psi})$$

	$Condition_1$						· · ·				· · ·	$Condition_I$									
	$Subject_1$			$Subject_2$			· · ·	$Subject_J$			· · ·	$Subject_{(I-1)J+1}$			$Subject_{(I-1)+2}$			· · ·	$Subject_{IJ}$		
	Run_1	Run_2	Run_3	Run_4	Run_5	Run_6	· · ·	Run_{JK-2}	Run_{JK-1}	Run_{JK}	· · ·	$Run_{(I-1)JK+1}$	$Run_{(I-1)JK+2}$	$Run_{(I-1)JK+3}$	$Run_{(I-1)JK+4}$	$Run_{(I-1)JK+5}$	$Run_{(I-1)JK+6}$	· · ·	Run_{IJK-2}	Run_{IJK-1}	Run_{IJK}
Summarized	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	· · ·	$\hat{y}$	$\hat{y}$	$\hat{y}$	· · ·	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	· · ·	$\hat{y}$	$\hat{y}$	$\hat{y}$

ROBUSTNESS TO OUTLIERS

Methods perform similarly with high quality data



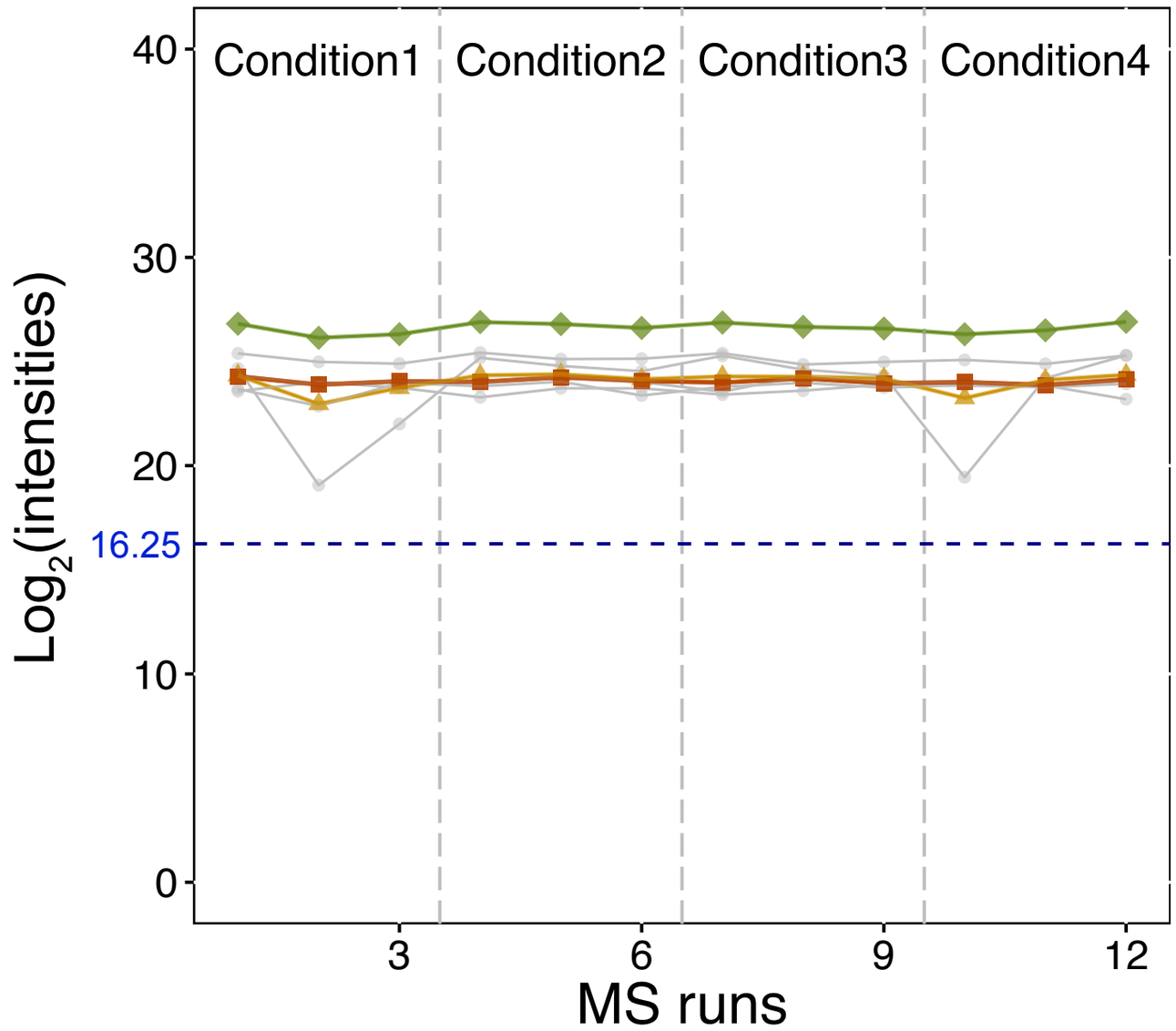


Condition2-Condition1 : True fold change=1

	EstimatedFC	Adj.pvalue
Proposed	1.016	0.999
TMP	1.016	0.999
Linear model	1.020	0.999
log(sum)	1.019	0.999

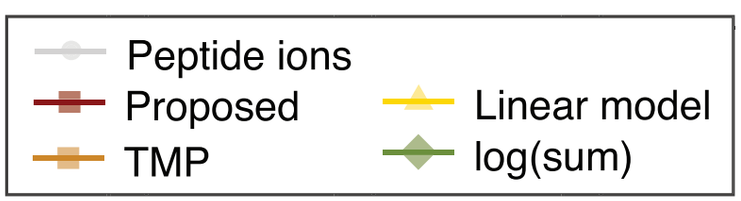
ROBUSTNESS TO OUTLIERS

Outliers in low intensities: robust summarization with TMP improves upon linear model



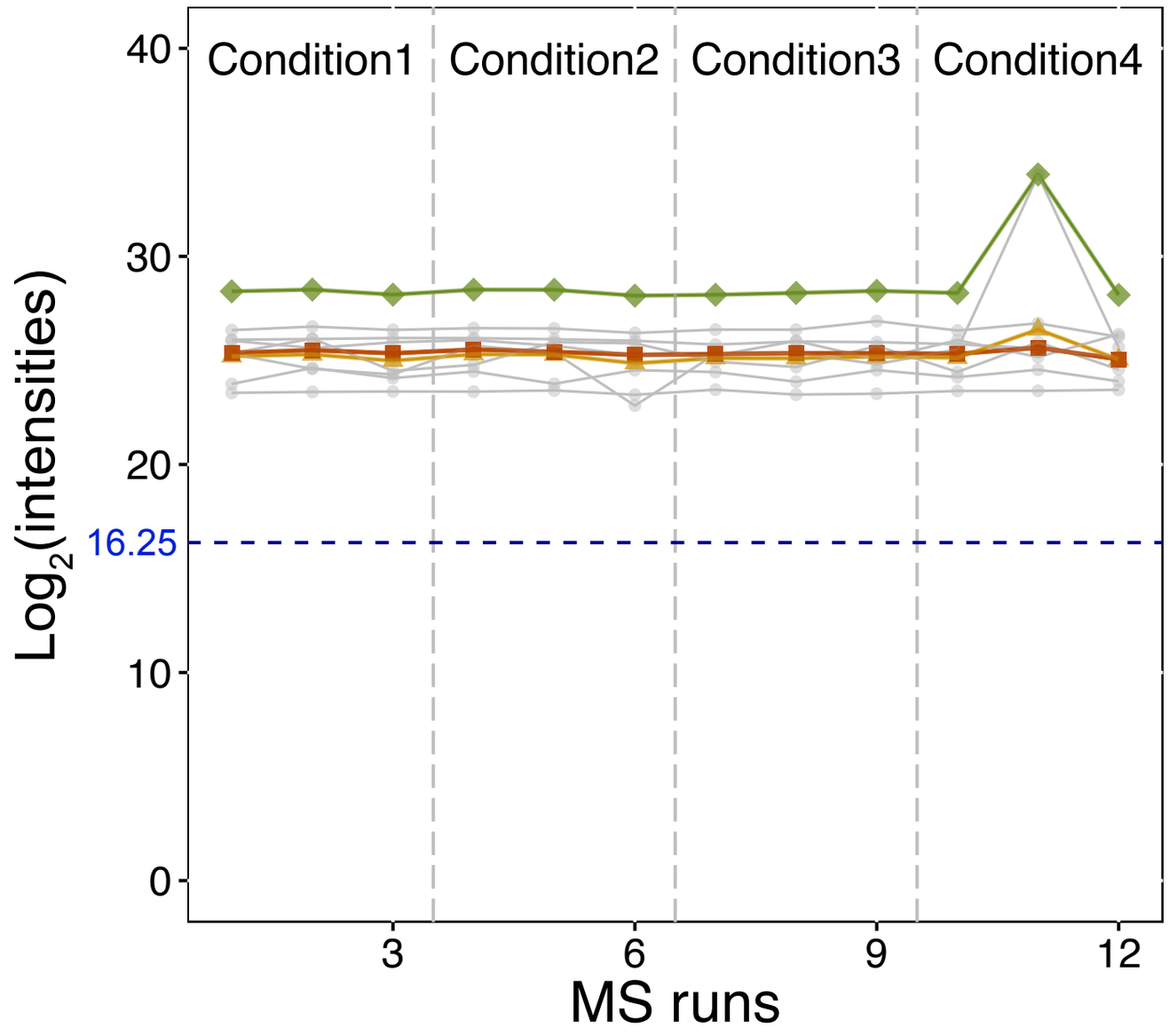
Condition3-Condition1 : True fold change=1

	EstimatedFC	Adj.pvalue
Proposed	0.979	0.952
TMP	0.979	0.956
Linear model	1.488	0.815
log(sum)	1.218	0.734



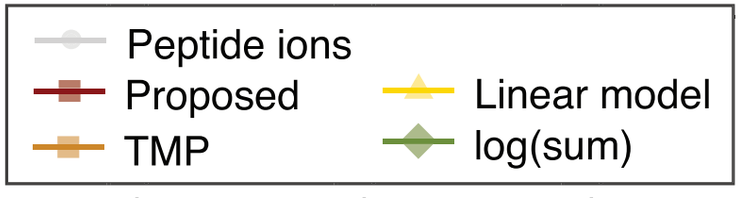
ROBUSTNESS TO OUTLIERS

Outliers in high intensities: robust summarization with TMP improves upon log(sum)



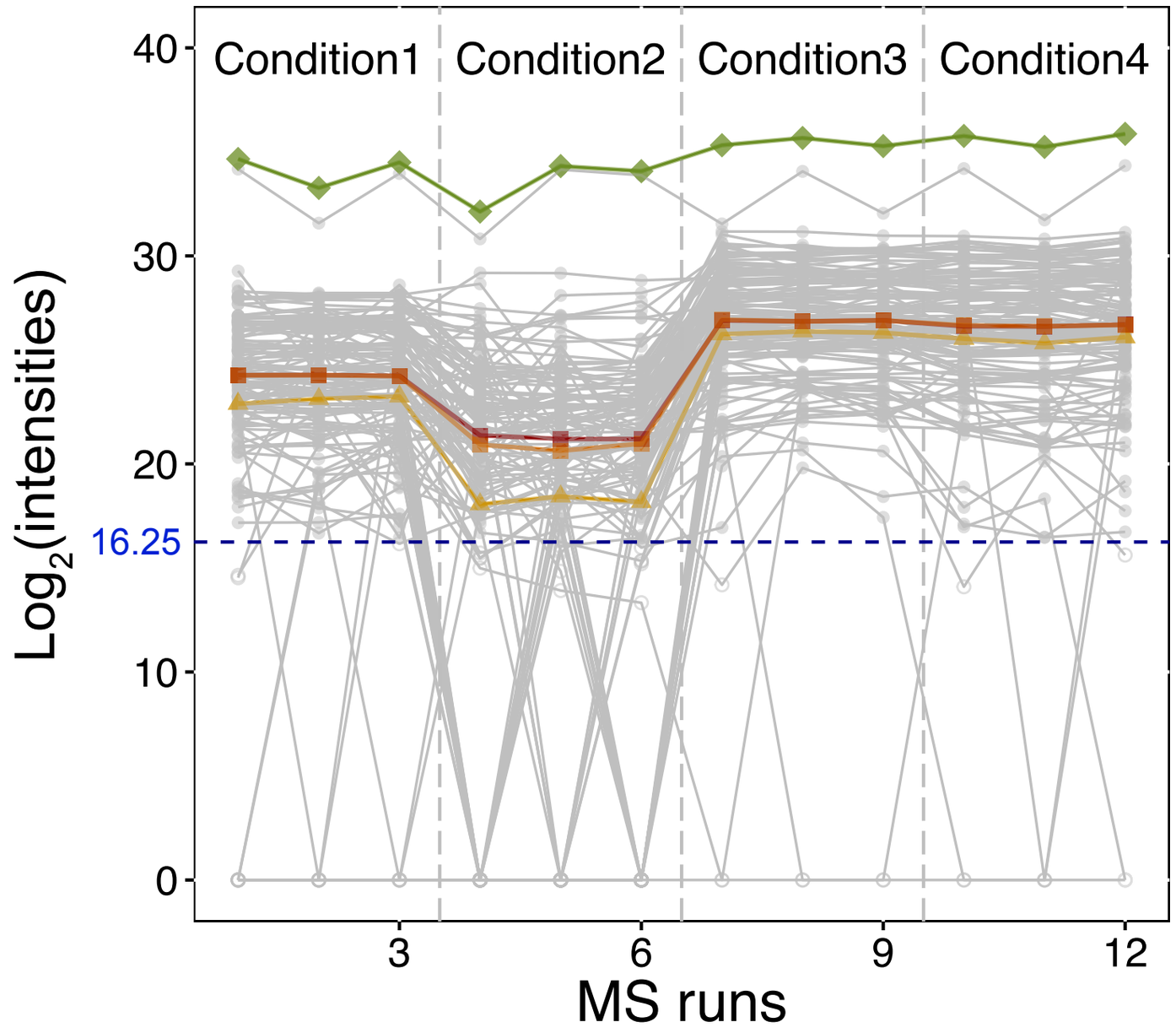
Condition4-Condition1 : True fold change=1

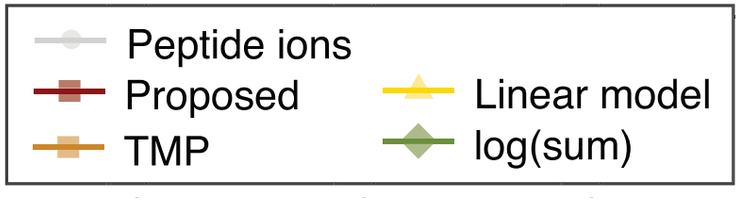
	EstimatedFC	Adj.pvalue
Proposed	0.951	0.948
TMP	0.951	0.948
Linear model	1.317	0.881
log(sum)	3.514	0.741



ROBUSTNESS TO OUTLIERS

Outliers in both high and low intensities: TMP improves upon linear model and log(sum)





Condition1-Condition2 : True fold change=7.5

	EstimatedFC	Adj.pvalue
Proposed	8.015	< 0.001
TMP	10.605	< 0.001
Linear model	29.106	< 0.001
log(sum)	1.552	0.999

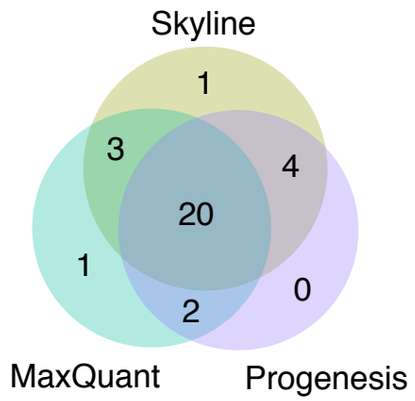
BETTER STATISTICAL METHODS ENHANCE REPRODUCIBLE RESEARCH

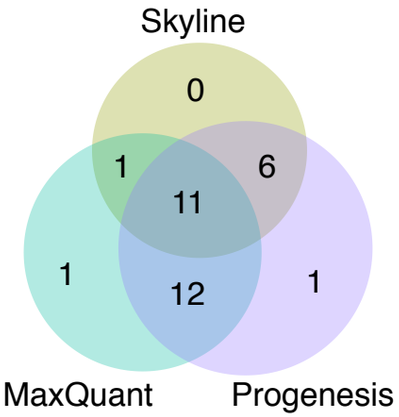
Better agreement in #differentially abundant proteins between tools

Proposed

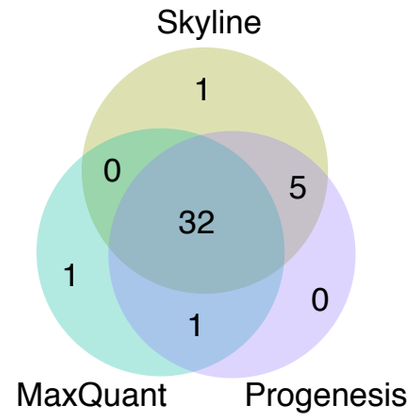
Log(sum)

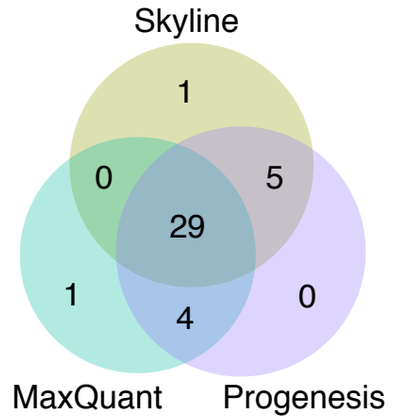
DDA: iPRG2015



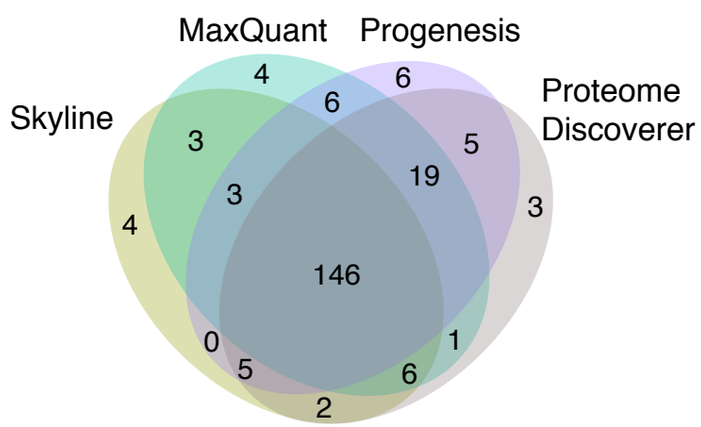


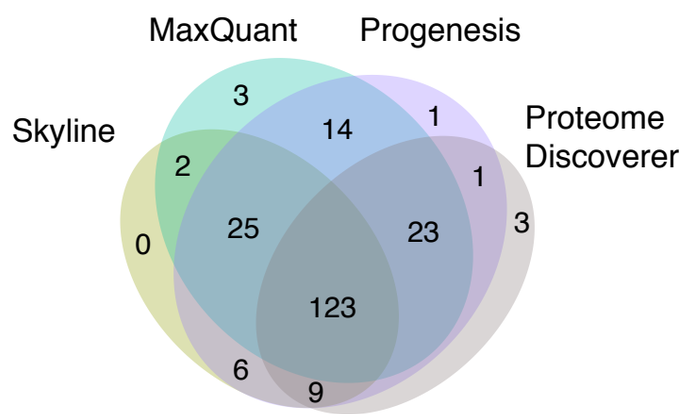
DDA: Cox 2014





DDA: Spike-in

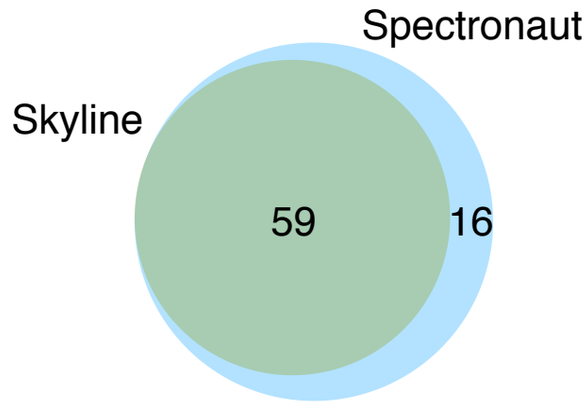




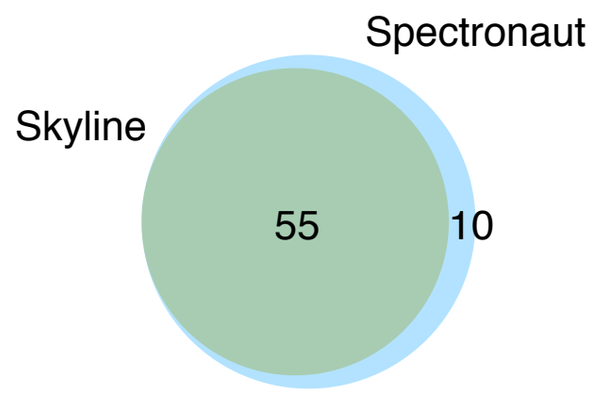
BETTER STATISTICAL METHODS ENHANCE REPRODUCIBLE RESEARCH

DIA: Bruderer2015

Proposed

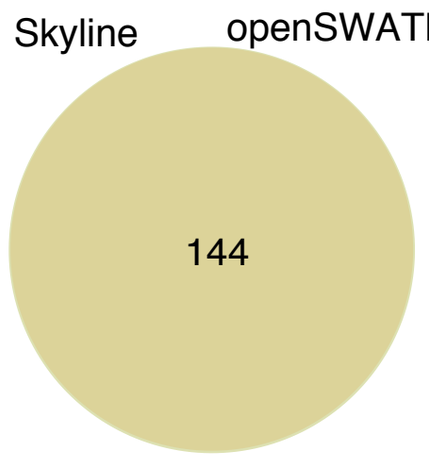


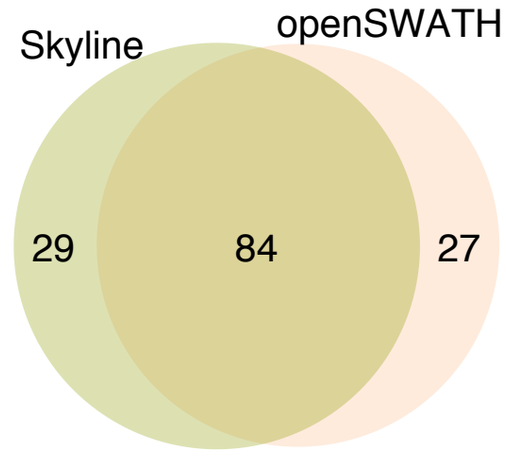
Log(sum)



Better agreement in #differentially abundant proteins between tools

DIA; Roest2014





MS EXPERIMENT: STATISTICIAN'S VIEW

Analytical method validation

↓

System suitability testing

↓

Experimental design

↓

Data acquisition
Quality control

↓

Data processing

↓

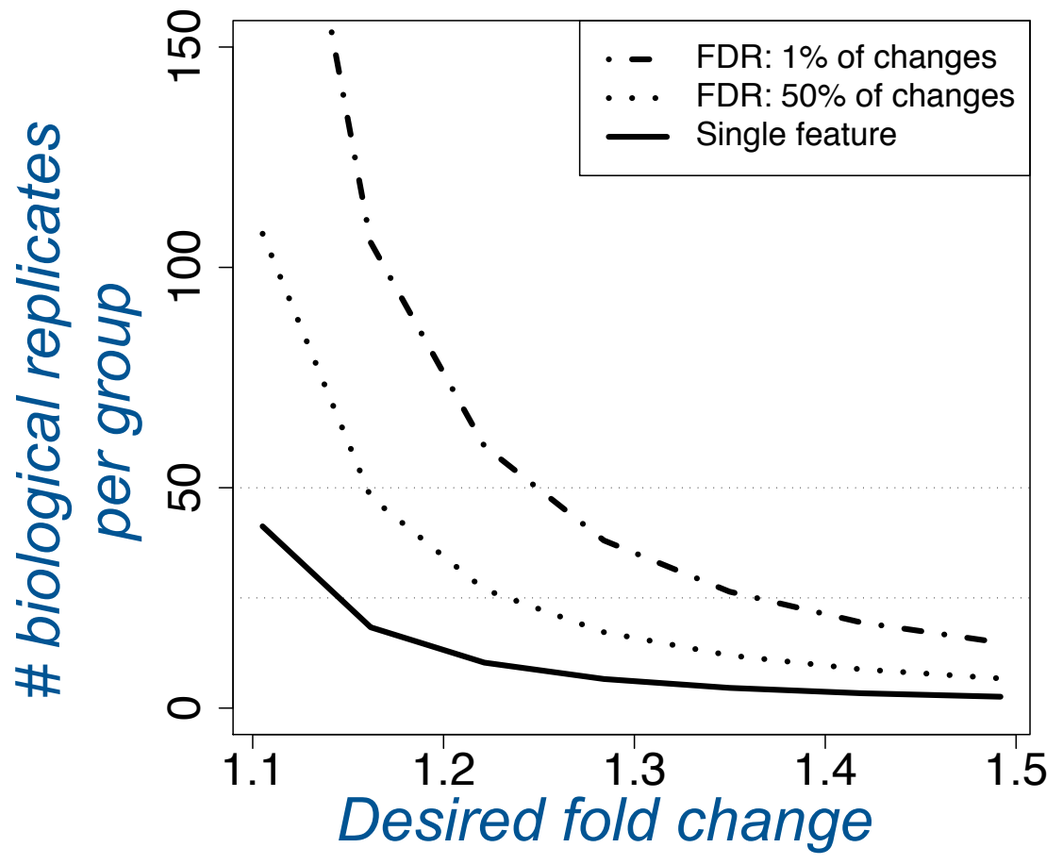
Statistical analysis

↓

Experimental design

Use the dataset to improve:

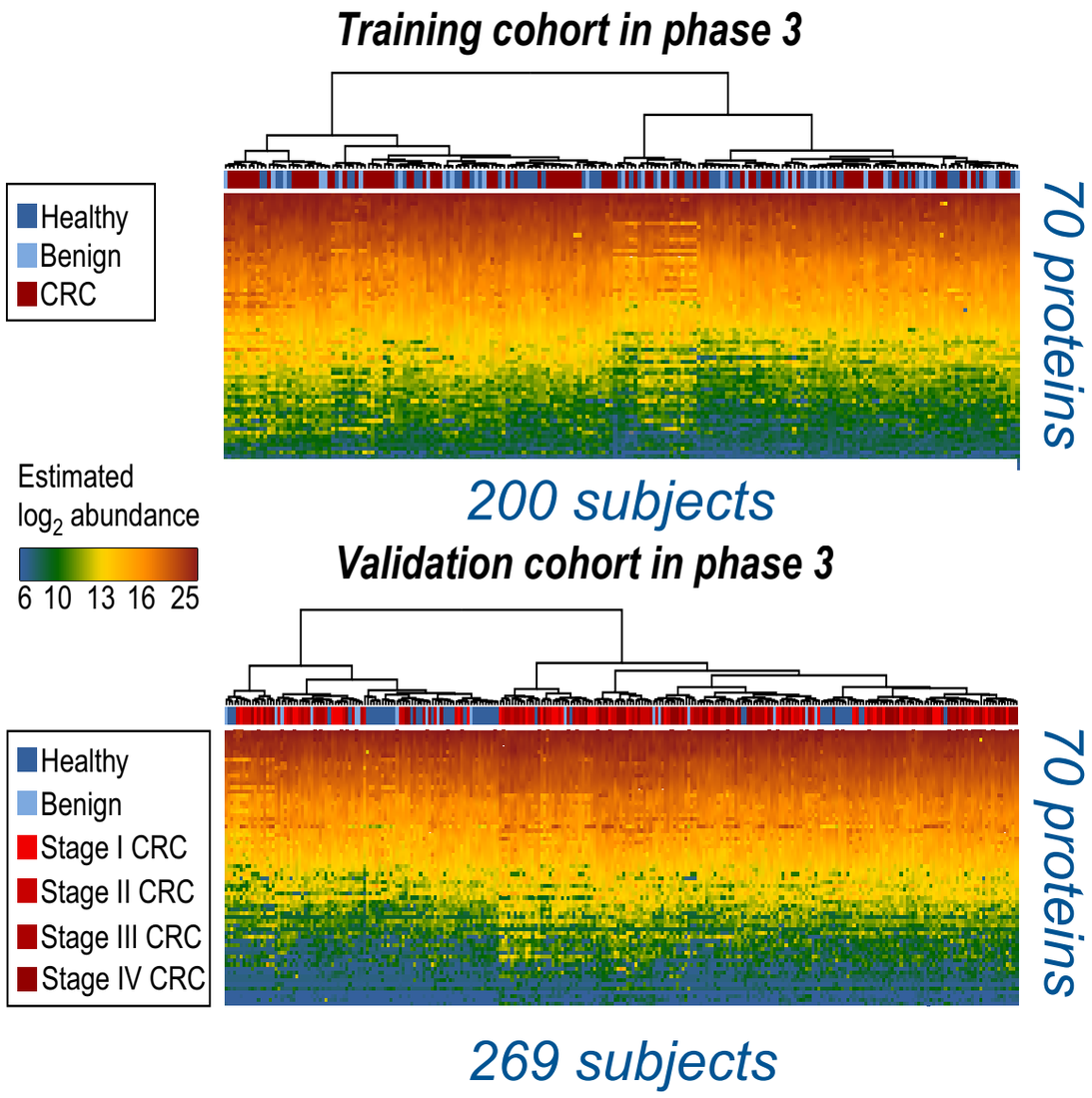
- Subject selection: matching
- Resource allocation: blocking
- Calculation of sample size



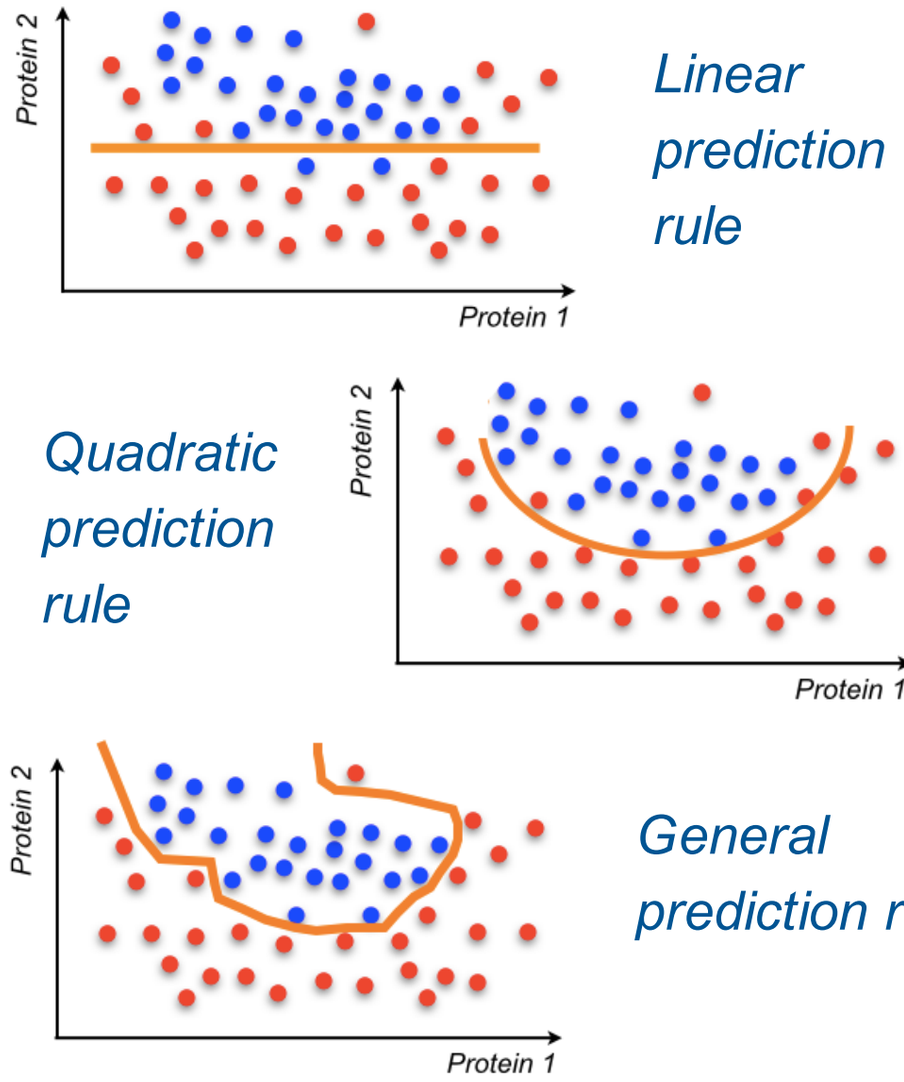
CASE STUDY:BIOMARKERS OF DISEASE

Statistical objective: a model classify each subject into correct class

Input: array of quantified proteins in subjects



Output: subset of predictive features

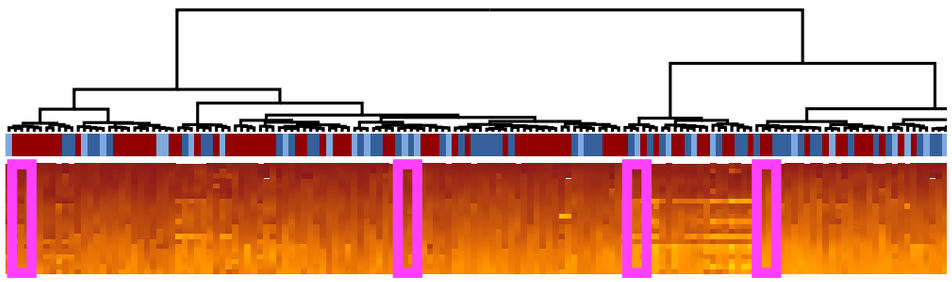


Surinova et al., EMBO Molecular Medicine, 2015

REPRODUCIBILITY EVALUATION

Resample subjects in training set, while fixing proteins number

Training cohort in phase 3



Protein rank: random forest classifier

10 predictive proteins

Protein	Mean accuracy decrease
ACTG1	4.29792551
H2AFX	4.17272432
ACTA2	3.64730166
HIST2H2AC	3.26940544
HIST1H2AB	3.10230189
ACTA1	3.07322088
TUBA1C	2.76765106
ACTG2	2.28619635
HSP90AA1	2.02174233
TUBB4A	1.99468791
HIST1H4A	1.86925866
SPTAN1	1.84453388
TUBB2B	1.75602473
COL6A3	1.72005229
TUBB4B	1.70743012
HIST1H2BK	1.68147503
HBB	1.61645207
HIST2H2AB	1.50013107

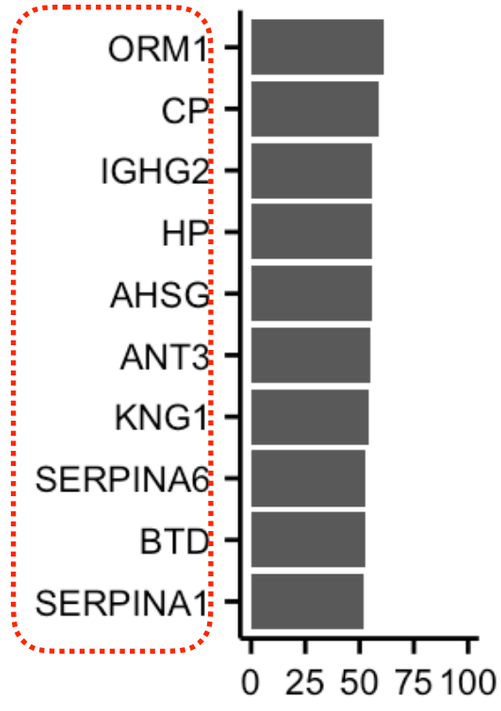
Repeat 100 times

Decision

Truth

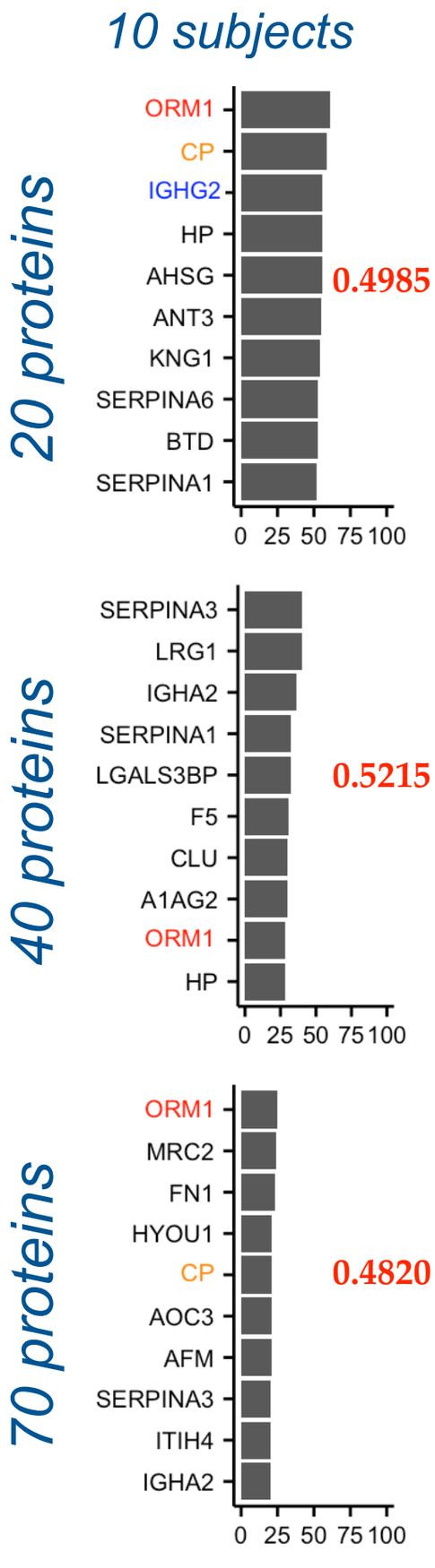
	CRC	Control
CRC	TP	FN
Control	FP	TN

Top 10 most frequently selected proteins

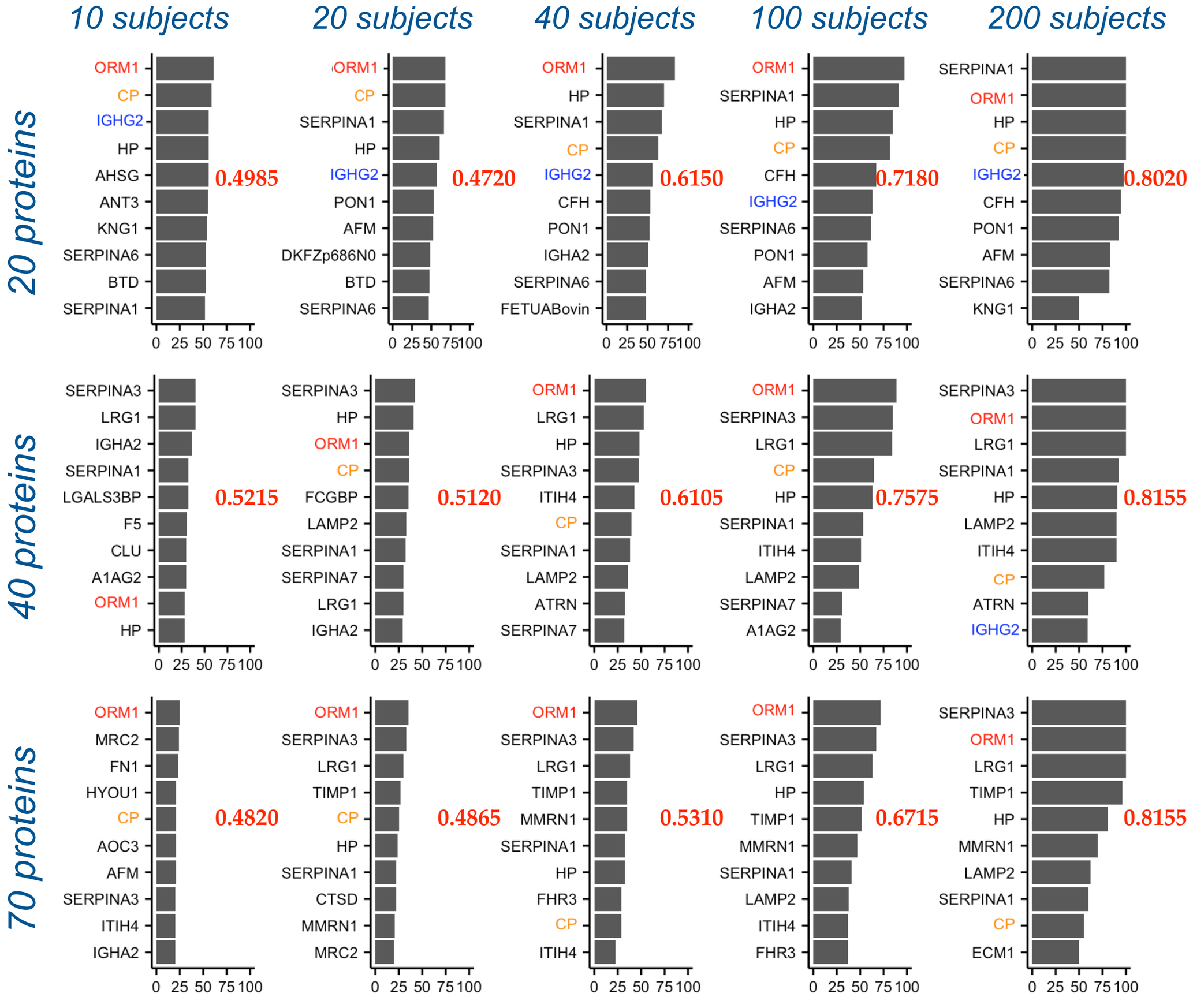


Number of iterations that a protein is identified predictive

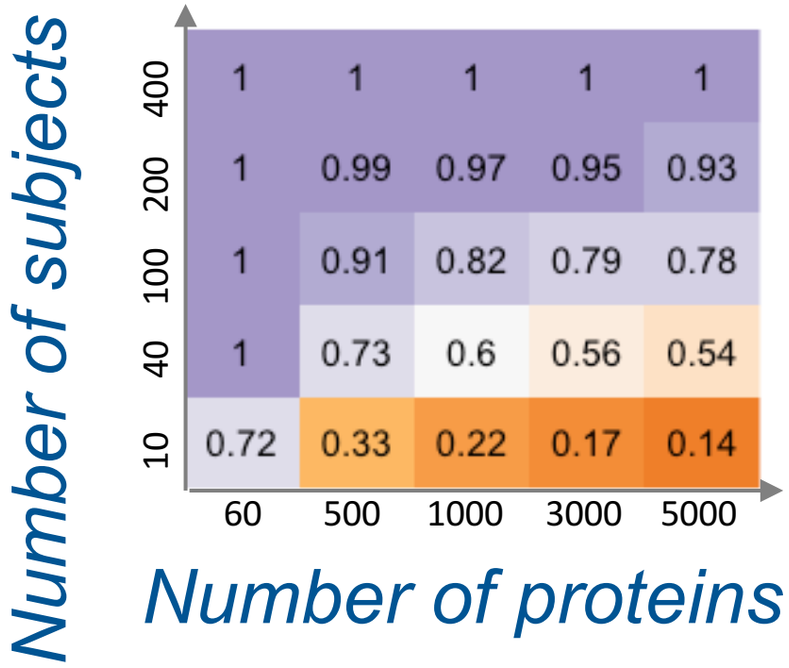
200 COLORECTAL CANCER SUBJECTS: SRM



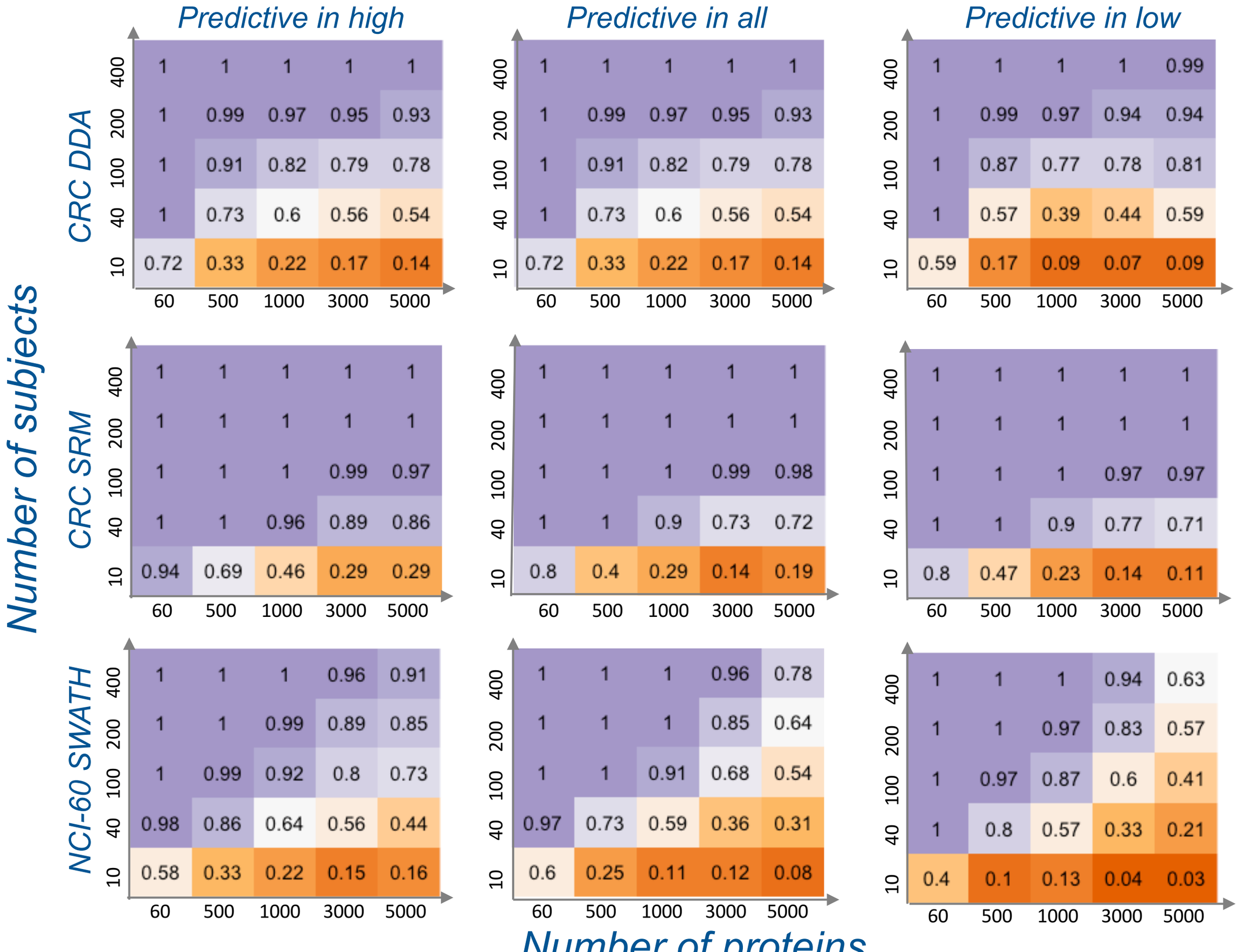
200 COLORECTAL CANCER SUBJECTS: SRM



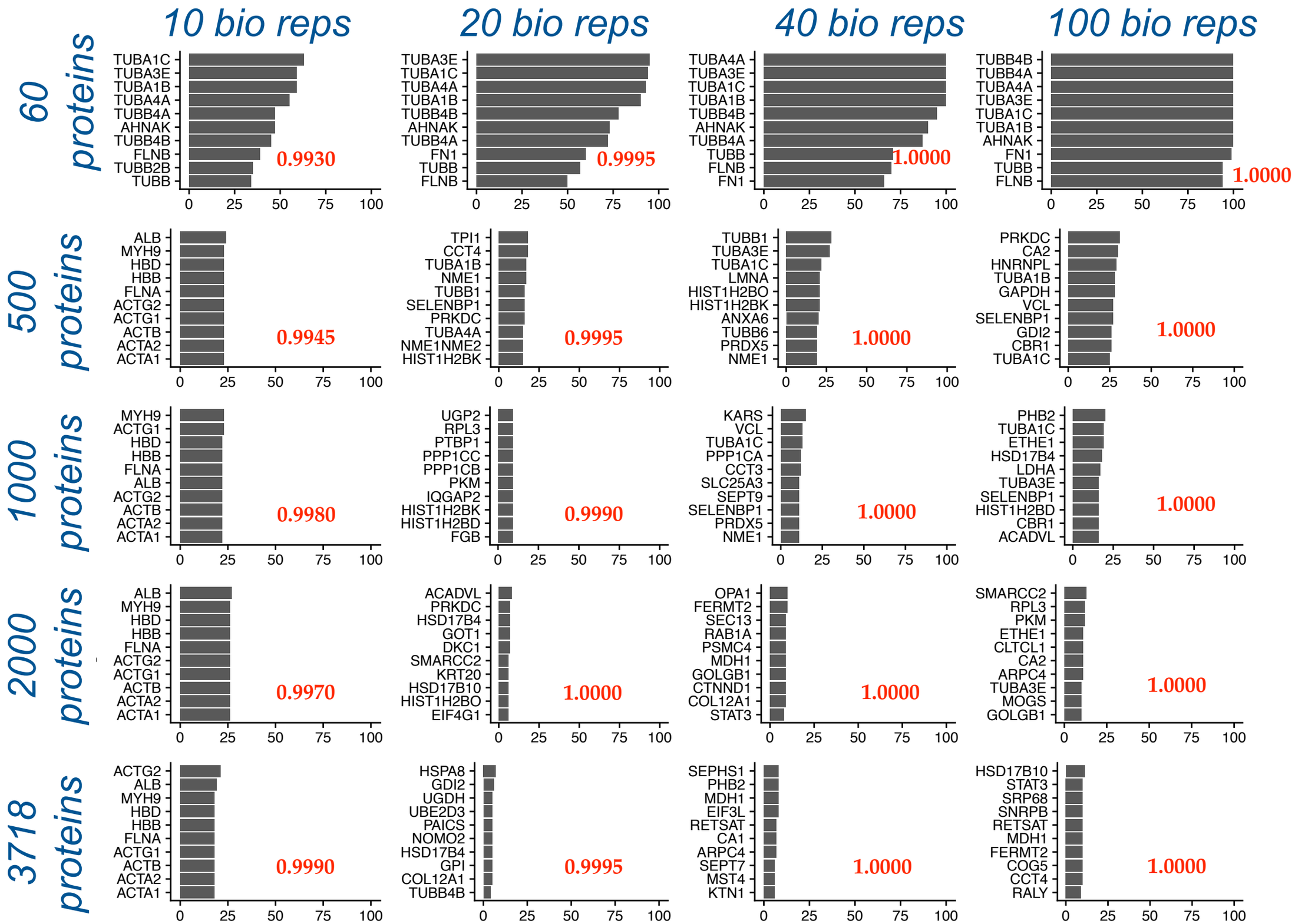
INCREASING PROTEIN NUMBER DECREASES ACCURACY



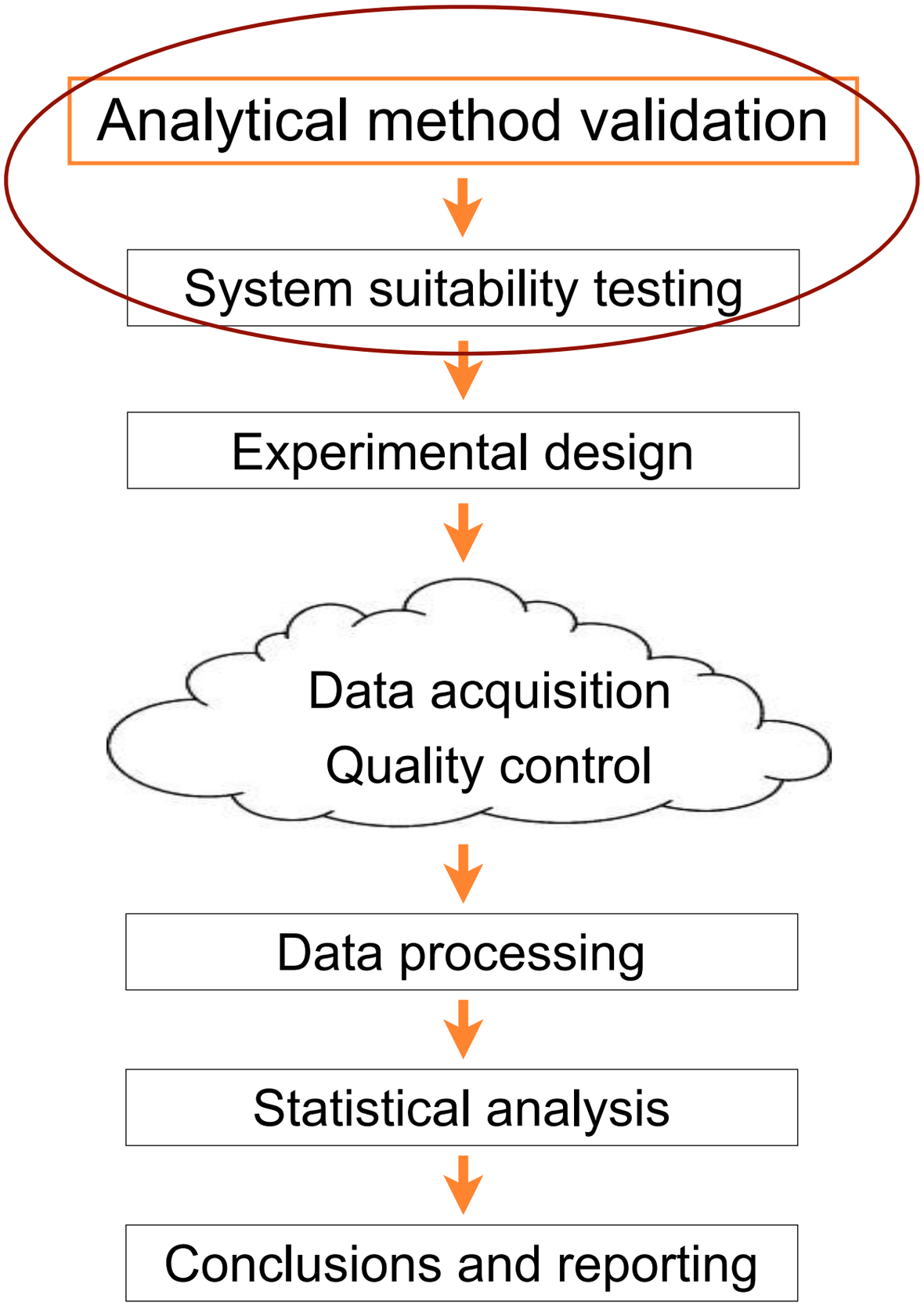
MORE REPLICATES SELECT MORE PREDICTIVE PROTEINS



SUBJECTS WITH COLORECTAL CANCER AND CONTROLS: FLAWED DESIGN

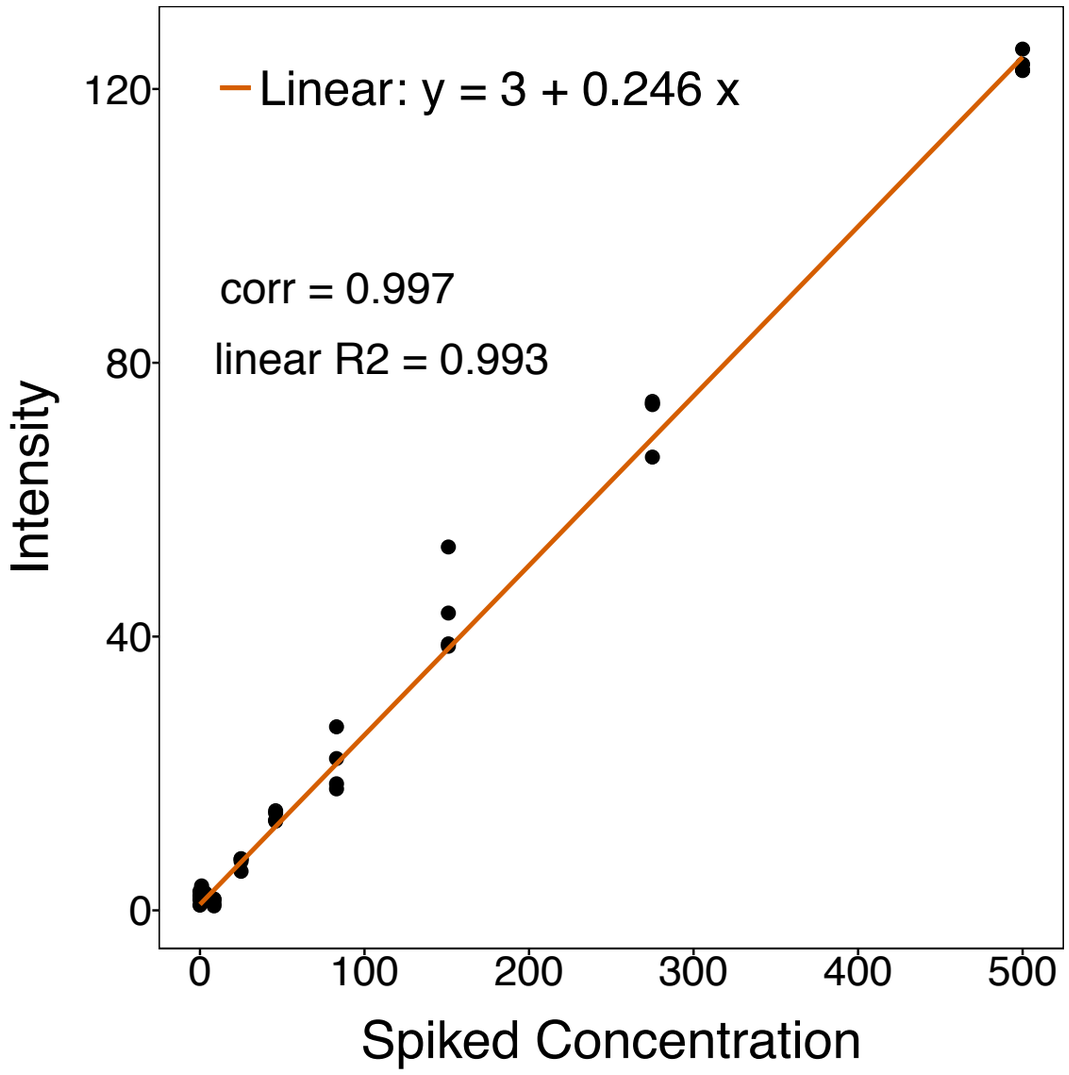


MS EXPERIMENT: STATISTICIAN'S VIEW



ASSAY CHARACTERIZATION

Statistical method: linear regression



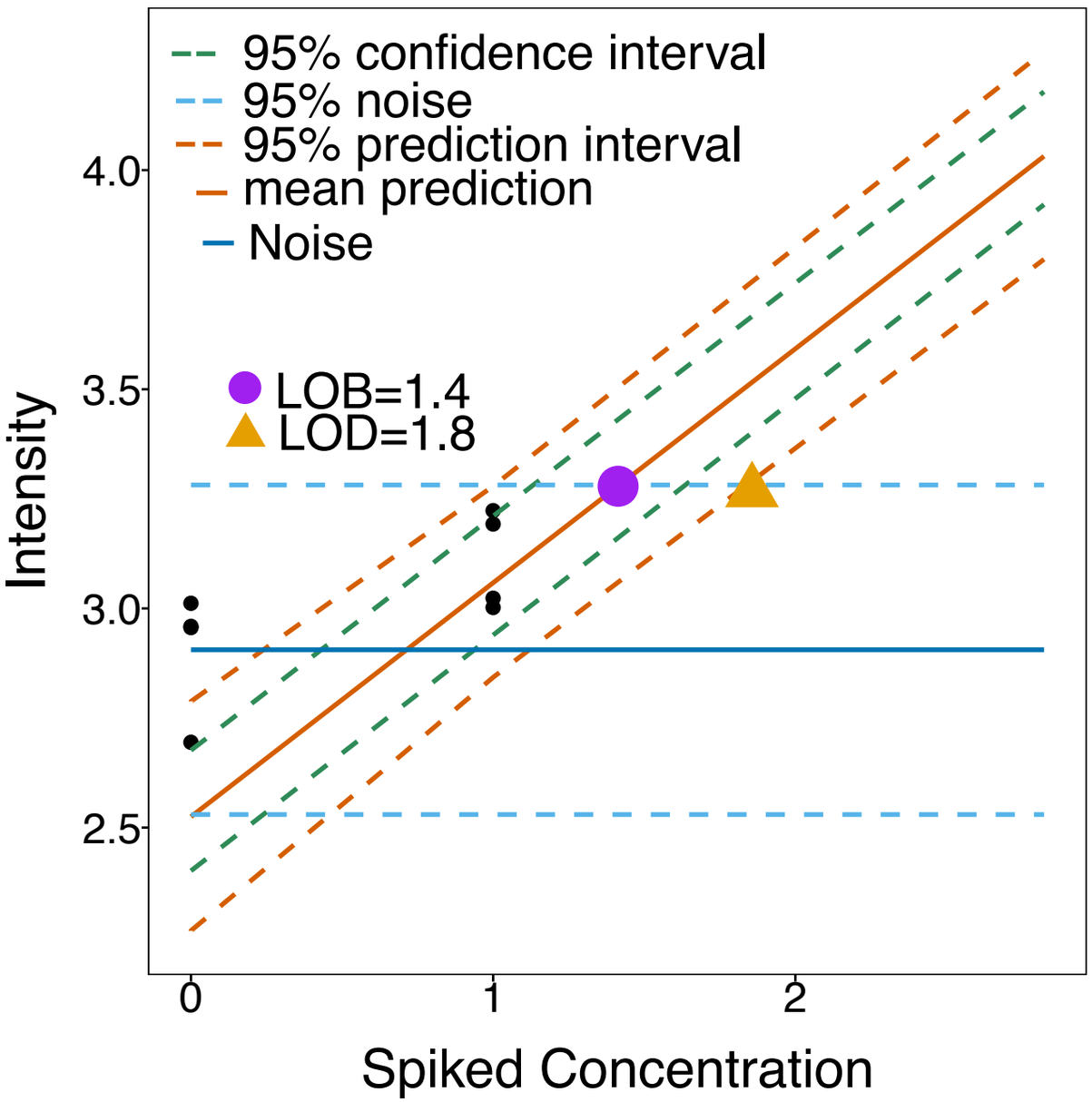
- Motivating example
  - DIA calibration experiment
  - Peptide SSAAPPPPPR
- Goal: quantify
  - Background noise
  - Slope (assay efficiency)
  - Quantify LoD and LoQ

- From the graph
  - Linear relationship is theoretically plausible
  - High correlation, high R^2

→ *False perception of good quality*

FIGURES OF MERIT

Statistical method: linear regression

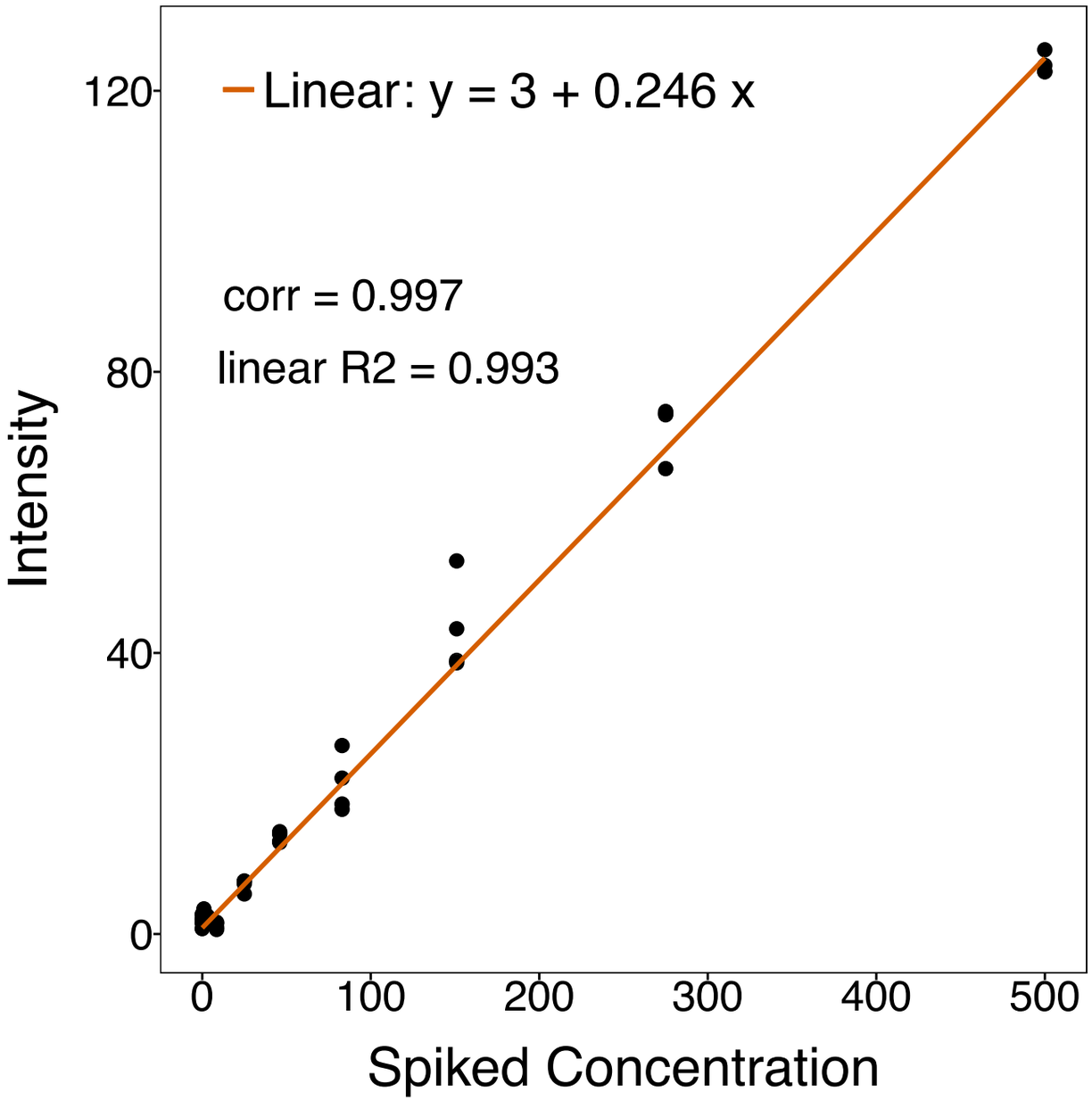


- Limit of blank (LoB)
  - upper limit of prediction interval of blank intersects curve fit
- Limit of detection (LoD)
  - upper limit of prediction interval of blank intersects lower limit of prediction interval of curve fit

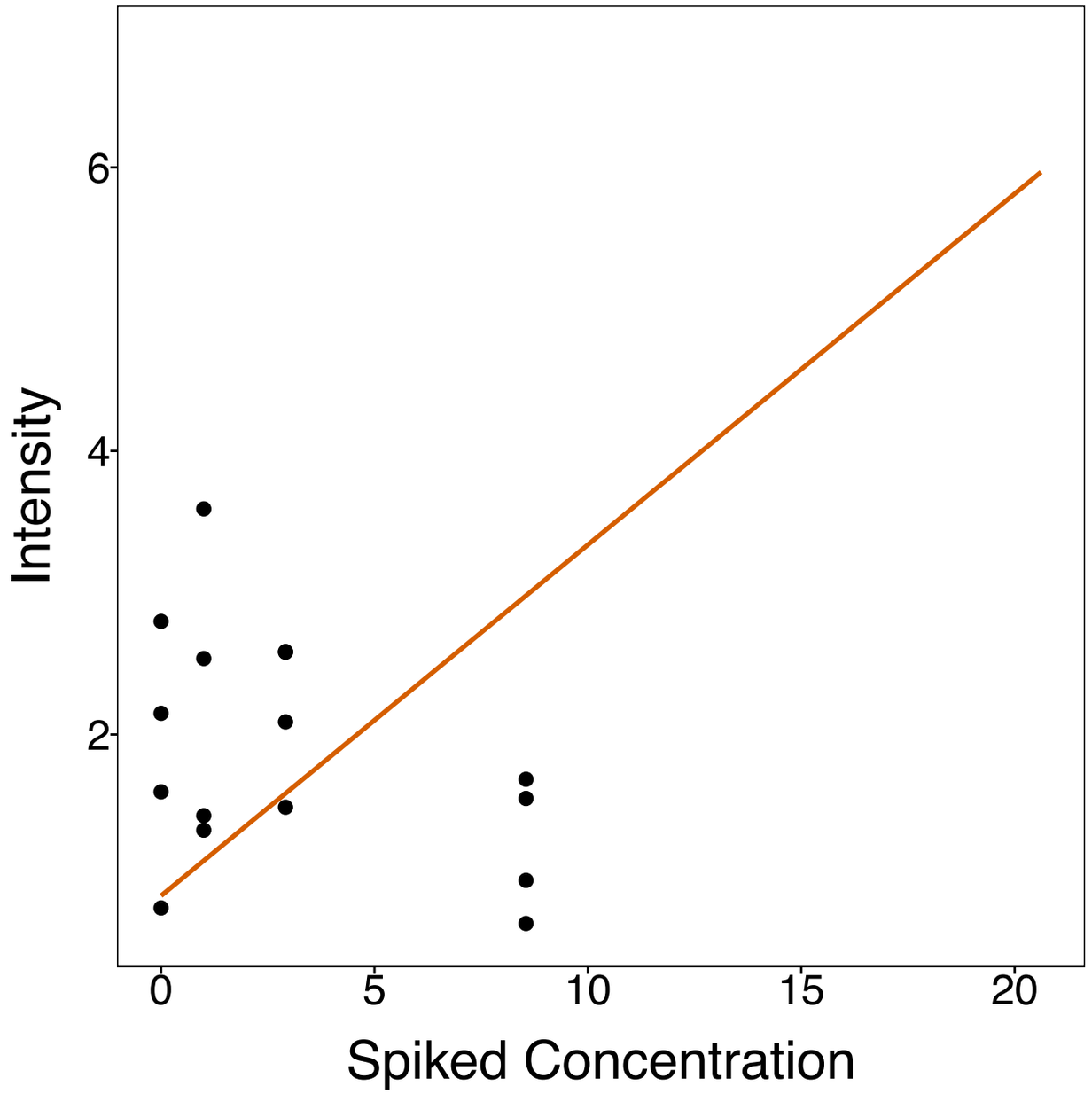
Used to both characterize the assays and compare the technologies

ZOOM INTO LOW CONCENTRATIONS

High R^2 does not always mean good fit



Zoom out

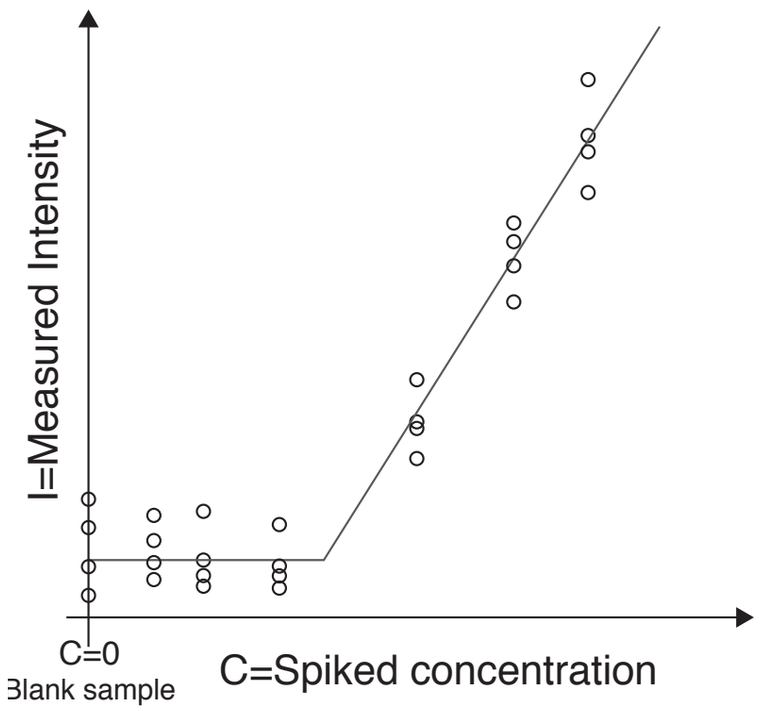


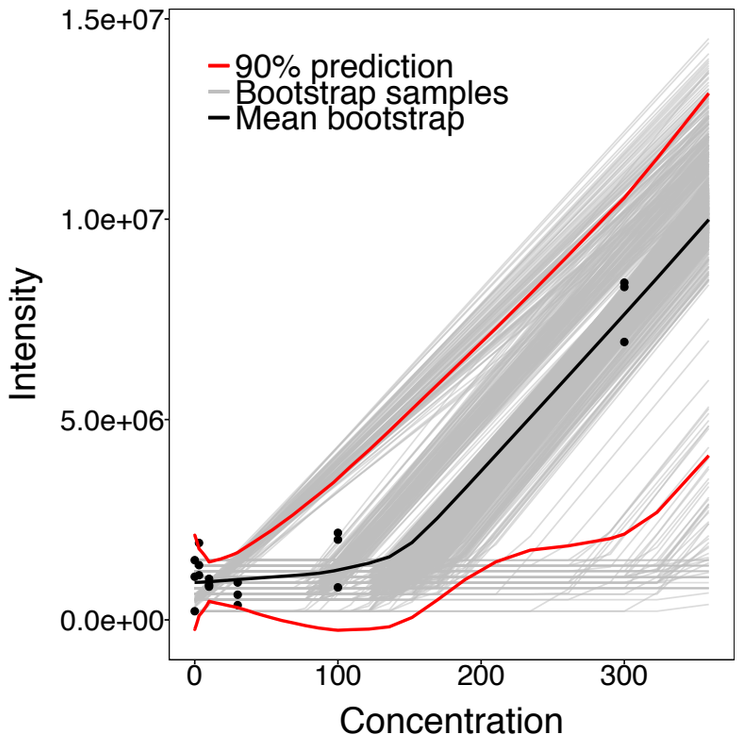
Zoom in

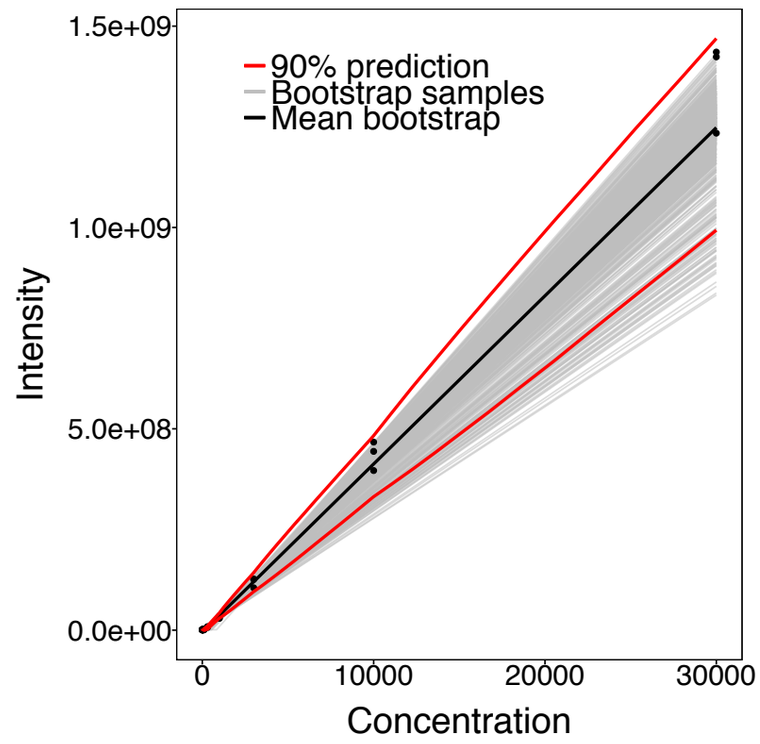
Calibration experiment, SRM , CPTAC

PROPOSED APPROACH

Canonical calibration curve + resampling (bootstrap)







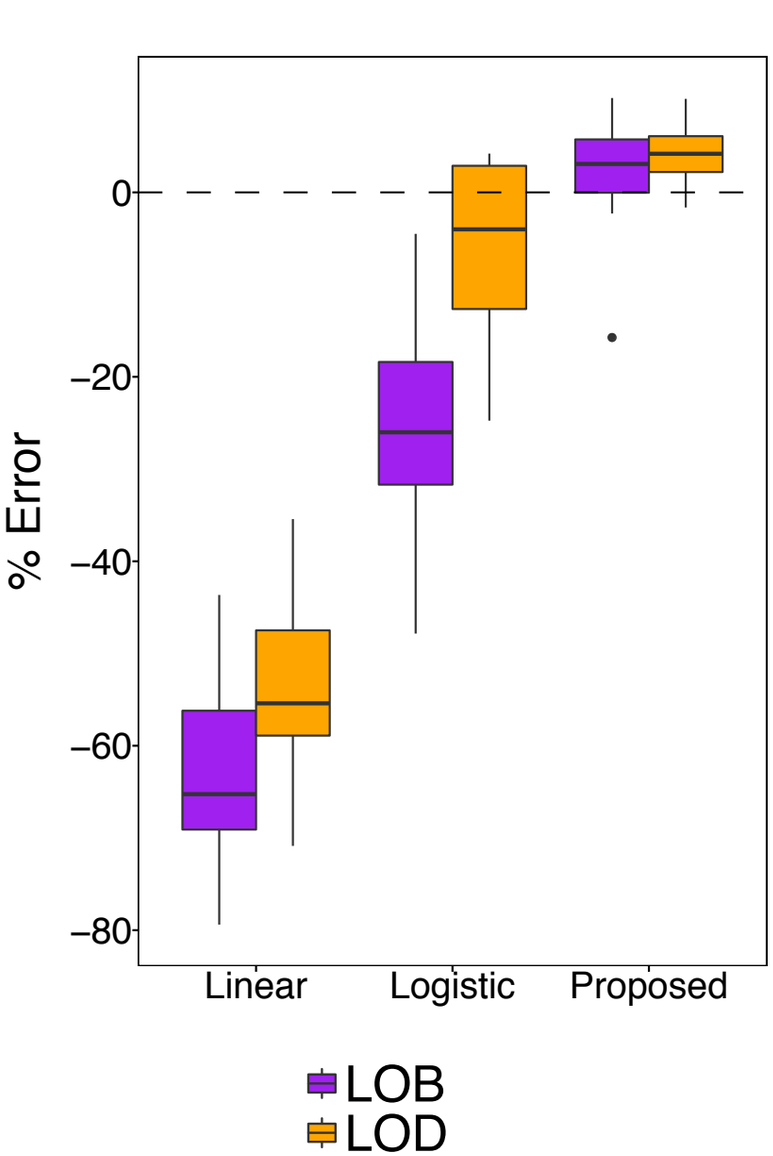
$$Y_{ij} = \begin{cases} \text{Intercept} + \text{Noise}_{ij} \\ \text{Intercept} + \text{Slope} \times (C_i - \text{Change}) + \text{Noise}_{ij}, \quad \text{if } C_i \geq \text{Change} \end{cases}$$

The nonlinear fit expresses uncertainty in change point location

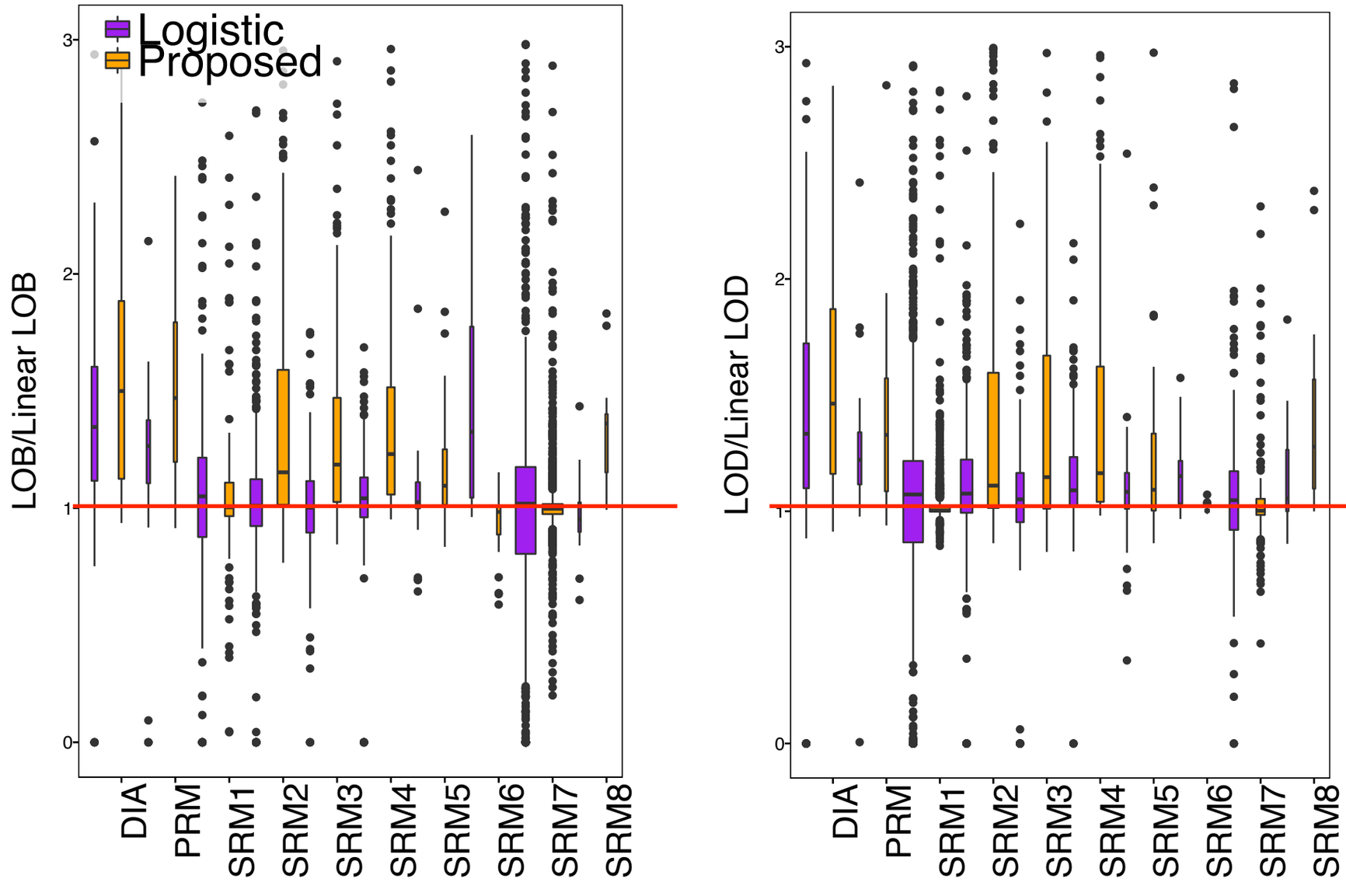
Yields more accurate & conservative figures of merit

ADVANTAGES OF NON-LINEAR REGRESSION

More accurate & conservative figures of merit



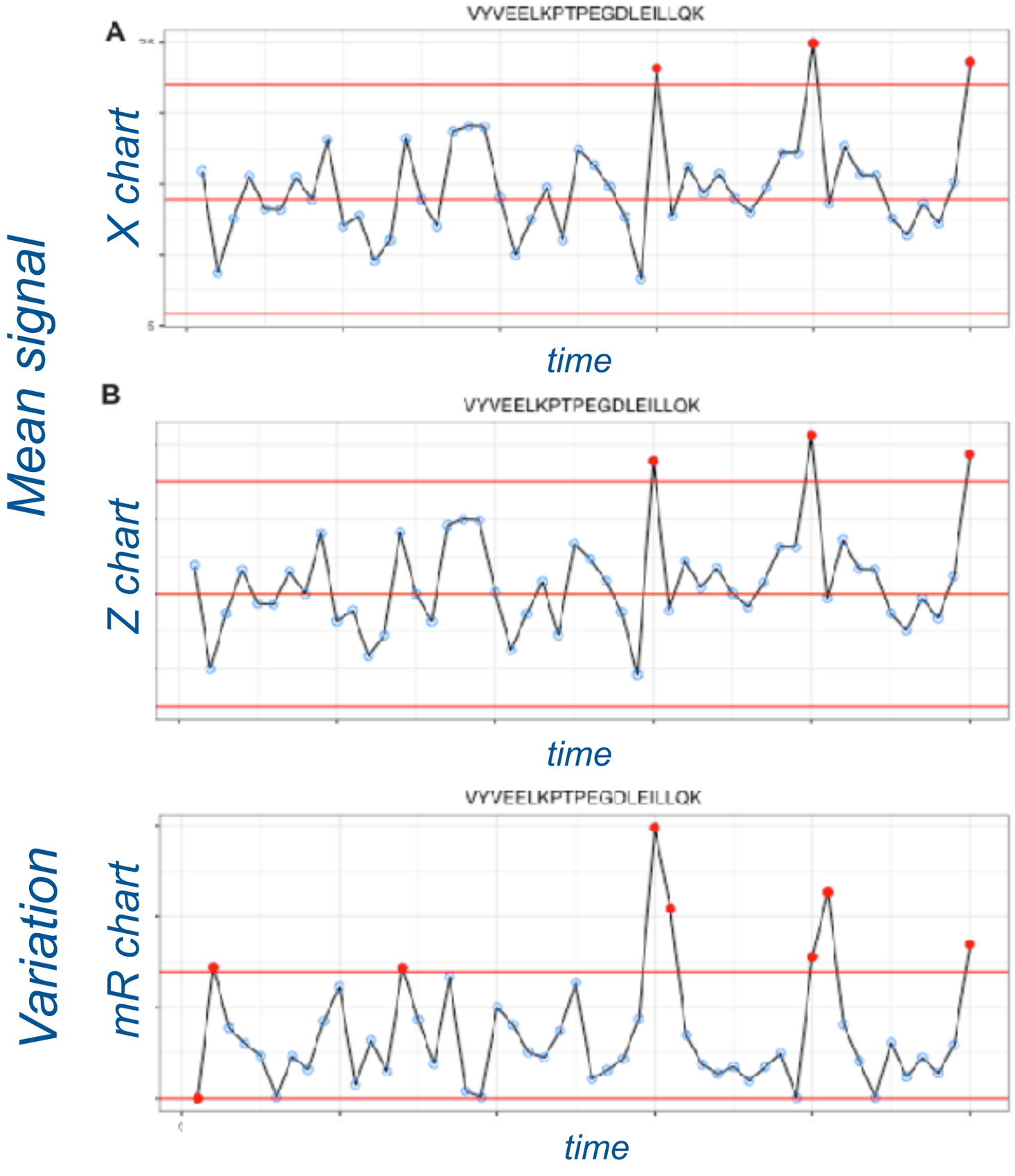
Simulation



Experimental datasets

Statistical process control

Monitor longitudinal profiles of a standard (e.g. peak area)



monitor mean

monitor standardized mean

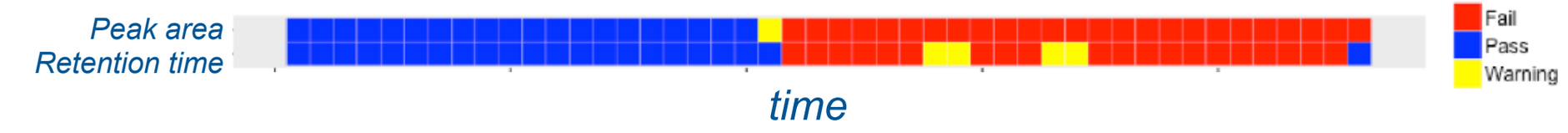
monitor variation moving range

Detects large changes

E. Dogu *et all Molecular and Cellular Proteomics*, 2017.

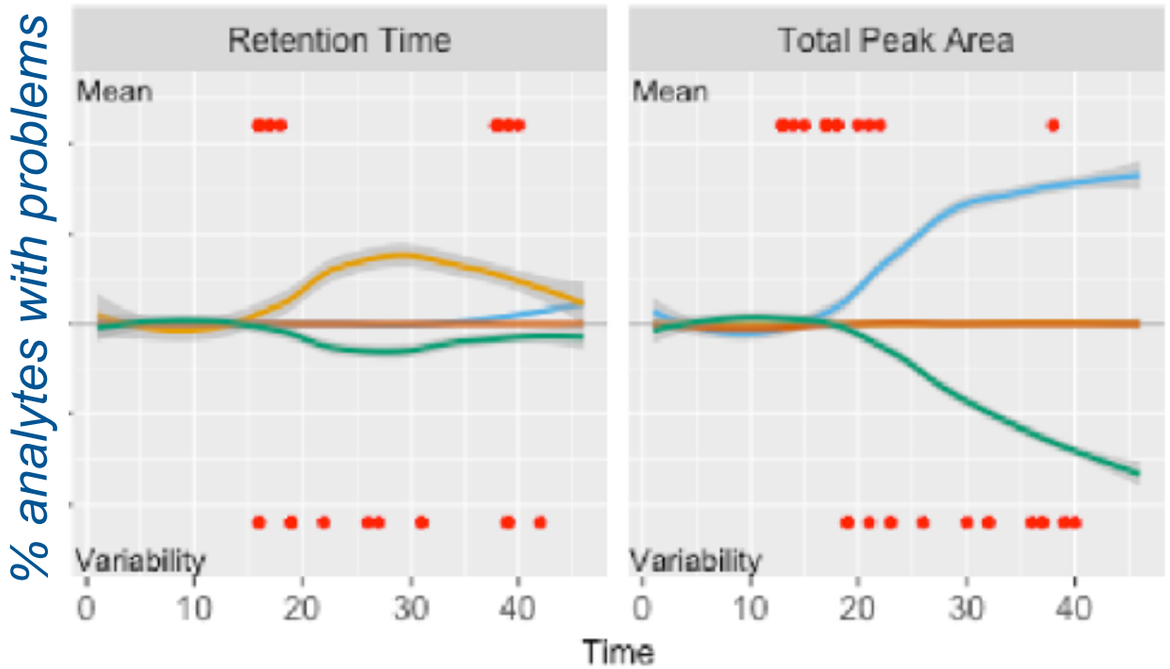
Multi-chart summary

Decision map: sustained drift in mean

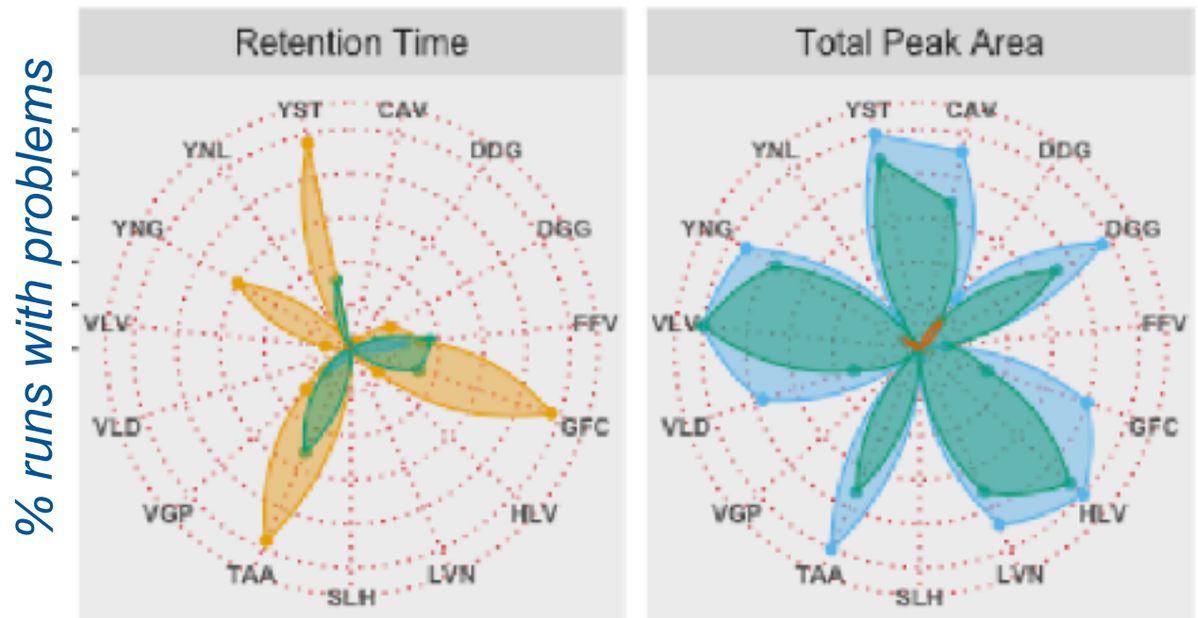


Decision map: sustained drift in variance

River plots: sustained drift in mean & variance



Radar plots: sustained drift in mean & variance



E. Dogu *et all Molecular and Cellular Proteomics*, 2017.

TAKE-AWAY

Statistical mindset is key for reproducible MSI research

Statistical software:

- Documentable, automated re-analysis workflows
- Fully transparent, open algorithms and code

Data analysis:

- Statistical modeling to handle variation
- Assay characterization, system suitability, QC

Experimental design:

- Selection of conditions/subjects/replicates
- SOP for the entire workflow

May Institute
Computation and statistics for mass spectrometry and proteomics

April 30–May 11, 2018, Northeastern University, Boston MA
Organizers : Meena Choi, Brendan MacLean and Olga Vitek

Program Overview

- April 30 – May 2, 2018 : Targeted proteomics with Skyline
- April 30 – May 2, 2018 : Proteomics and metabolomics with OpenMS
- May 2-4, 2018 : Beginner's statistics in R
- May 2-4, 2018 : Advanced R
- May 7-9, 2018 : Statistics for quantitative mass spectrometry
- May 7-9, 2018 : Visualization of biomolecular data
- May 9-11, 2018 : Capstone – case studies in data-independent acquisition (DIA)

NIH

de NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

http://computationalproteomics.ccis.northeastern.edu/

Sue Abbatiello

Ruedi Aebersold

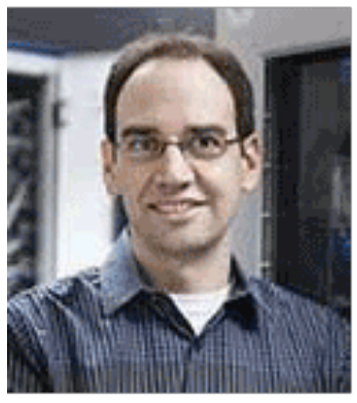
Kylie Bemis

Meena Choi

Ben Collins

Laurent Gatto

Andy Hoofnagle



Oliver Kohlbacher

Mike MacCoss

Brendan MacLean

Olga Vitek

ACKNOWLEDGEMENTS

Northeastern University

Kylie Bemis
Meena Choi
Eralp Dogu
Dan Guo
April Harry
Ting Huang
Cyril Galitzine
Robert Ness
Sara Taheri
Tsung-Heng Tsai

ABRF iPRG

Henry Lam
Eugene Kapp
Brett Phinney
John Cottrell
Michael Hoopman
Sangtae Kim
Thomas Neubert
Magnus Palmblad
Sue Weintraub

University of Washington

Michael MacCoss
Brendan MacLean
Jarrett Egertson

Mugla University

Eralp Dogu

ETH Zurich

Ruedi Aebersold
Tiannan Guo
Ruth Huttenhain
Paola Picotti
Silvia Surinova
Bernd Wollscheid

Support:

NSF
NIH
Sternberg Chair
Canary Center
Roche
Genentech
Eli Lilly