

INTRODUCTION TO MSSTATS

Meena Choi, Olga Vitek

College of Computer and Information Science



Northeastern University

WHY STATISTICS?

- Variation and uncertainty are unavoidable
 - *Technical variation*: sampling handling, storage, processing
 - *Instrumental variation*: matrix effects, ion suppression
 - *Signal processing*: peak boundaries, identity, intensity
 - *Biological variation*: variation in protein abundance
- Overall goal: effective, reproducible research



OUTLINE

- Motivating example
 - ABRF iPRG study
- MSstats
 - Statistical relative quantification of proteins and peptides
 - Methods evaluation
- Extensions to MSstats
 - Assay characterization
 - System suitability and quality control

ABRF IPRG STUDY 2015

Detection of differentially abundant proteins in controlled mixture

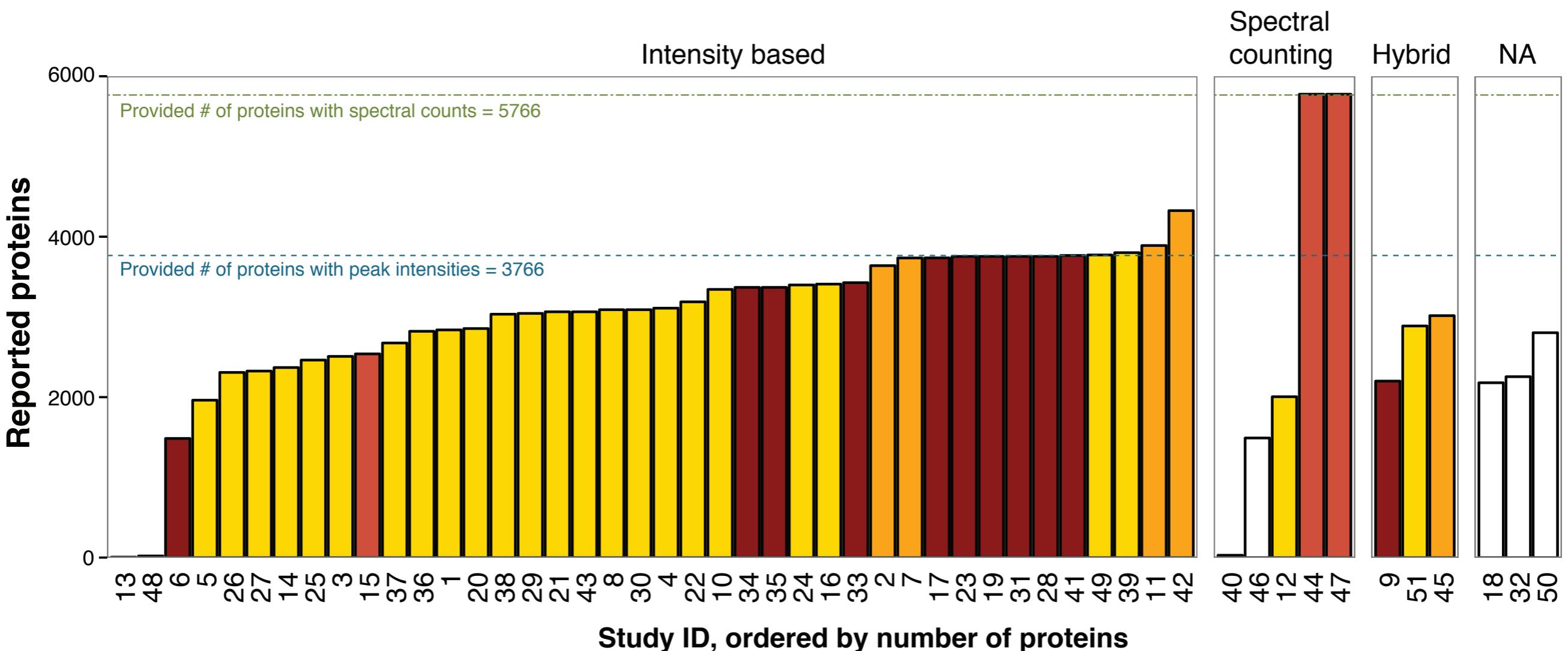
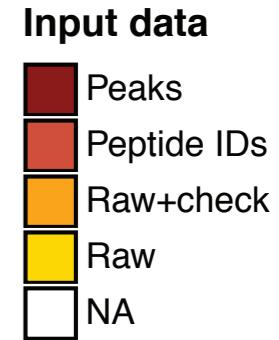
Name	Origin	Molecular Weight	Samples			
			1	2	3	4
A Ovalbumin	Chicken Egg White	45KD	65	55	15	2
B Myoglobin	Equine Heart	17KD	55	15	2	65
C Phosphorylase b	Rabbit Muscle	97KD	15	2	65	55
D Beta-Galactosidase	Escherichia Coli	116KD	2	65	55	15
E Bovine Serum Albumin	Bovine Serum	66KD	11	0.6	10	500
F Carbonic Anhydrase	Bovine Erythrocytes	29KD	10	500	11	0.6

Spiked into a constant background: tryptic digests of S. cerevisiae

- ◆ Three technical replicates per sample
 - ◆ Thermo nLC 1000 system
 - ◆ 110-min linear gradient
 - ◆ DDA profile mode in Orbitrap
 - ◆ Data processing with Skyline

DIVERSE SUBMISSIONS

*INPUT, PROTEIN NUMBER,
AND CHOICE OF QUANTIFICATION*

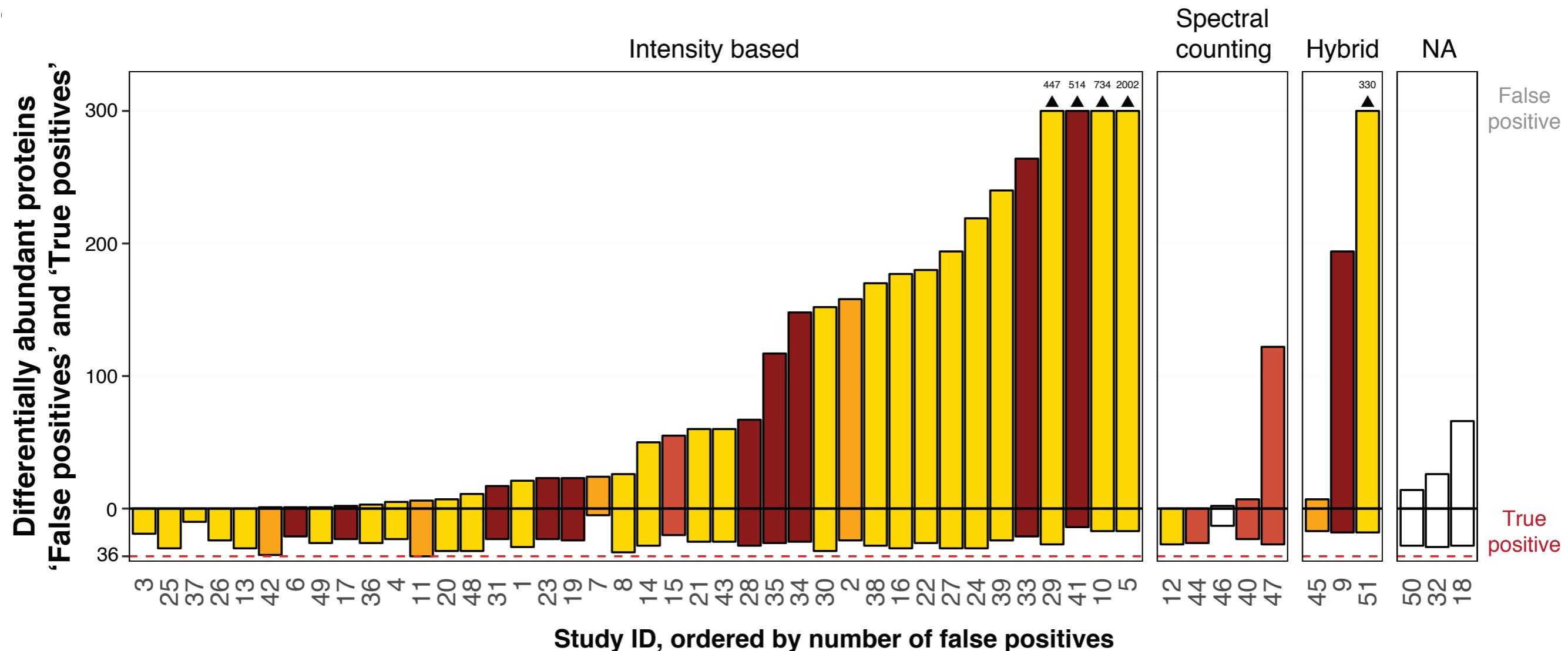


DIVERSE SUBMISSIONS

ACCURACY OF DETECTING DIFFERENTIAL ABUNDANCE

Input data

- Peaks
- Peptide IDs
- Raw+check
- Raw
- NA

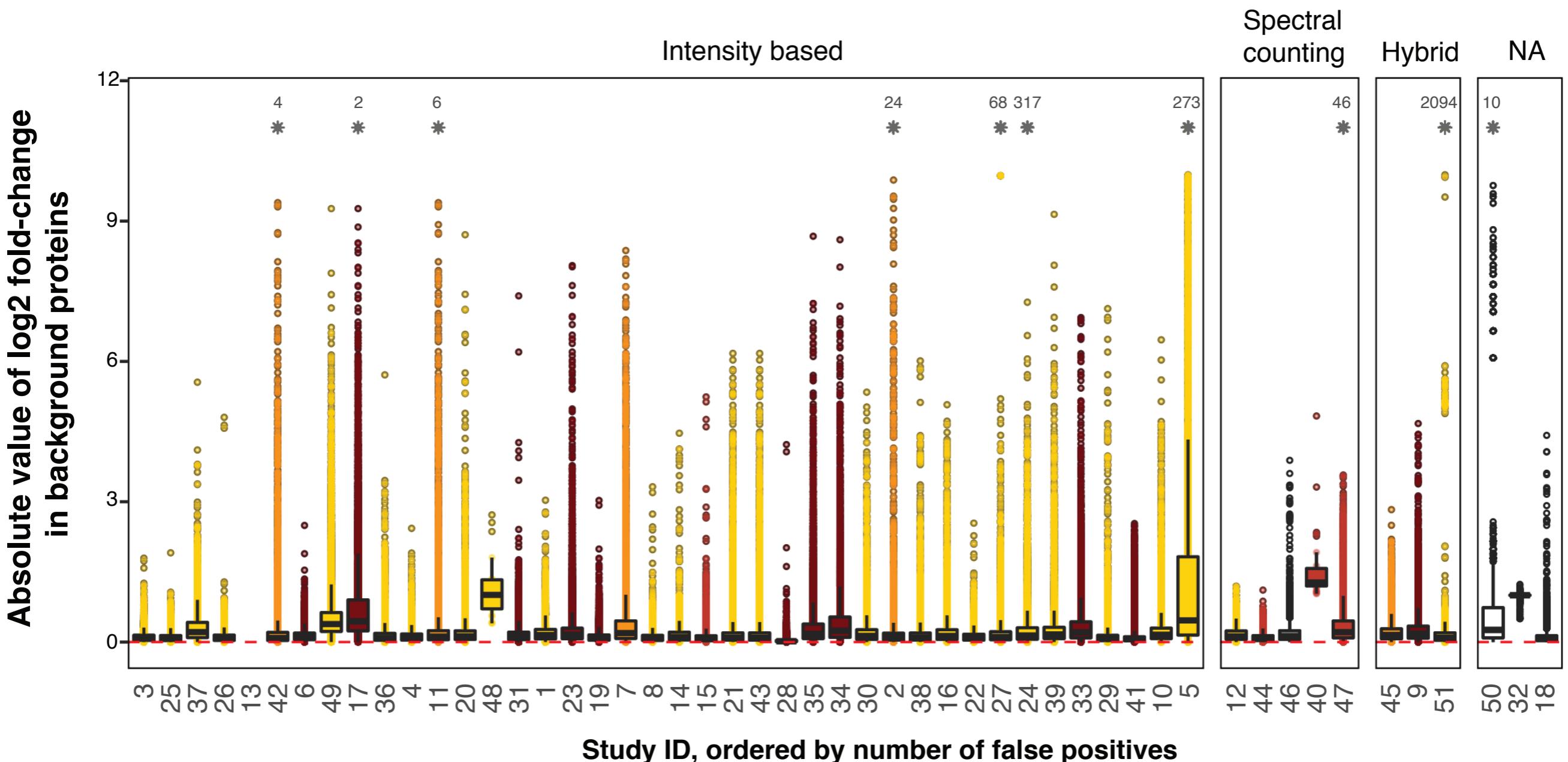


DIVERSE SUBMISSIONS

ACCURACY OF ESTIMATING FOLD CHANGE

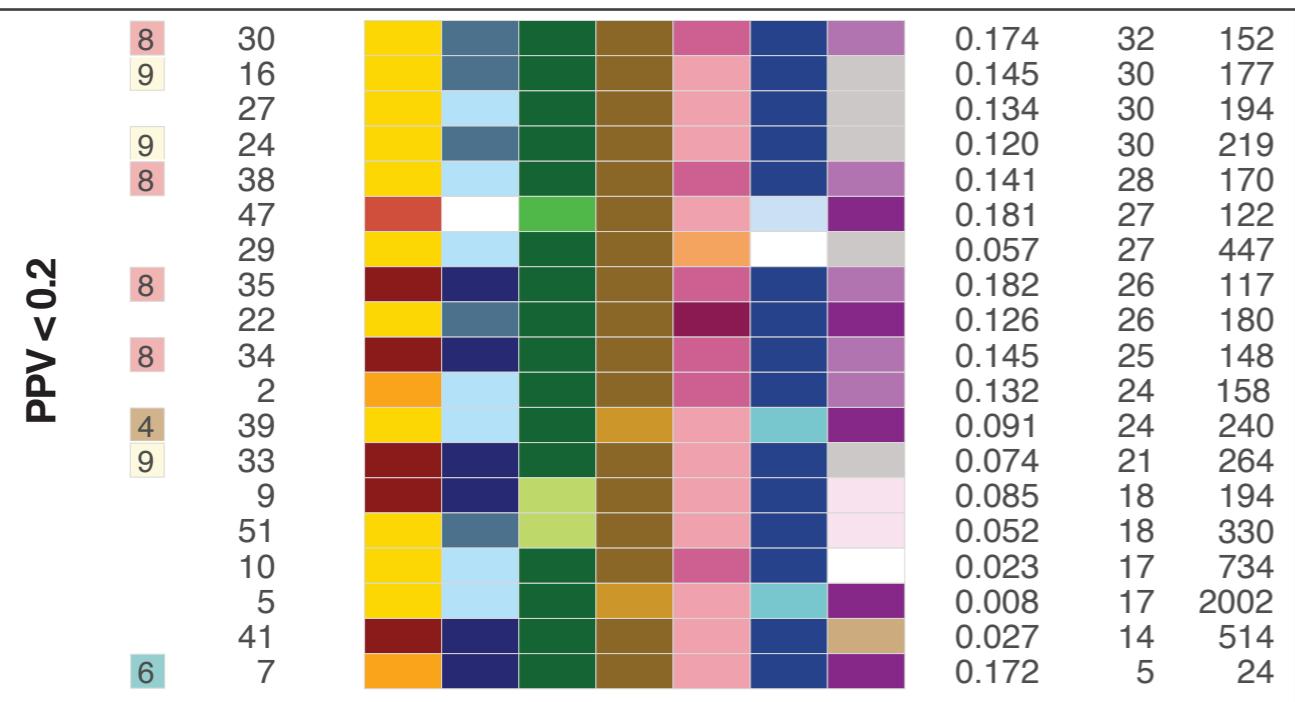
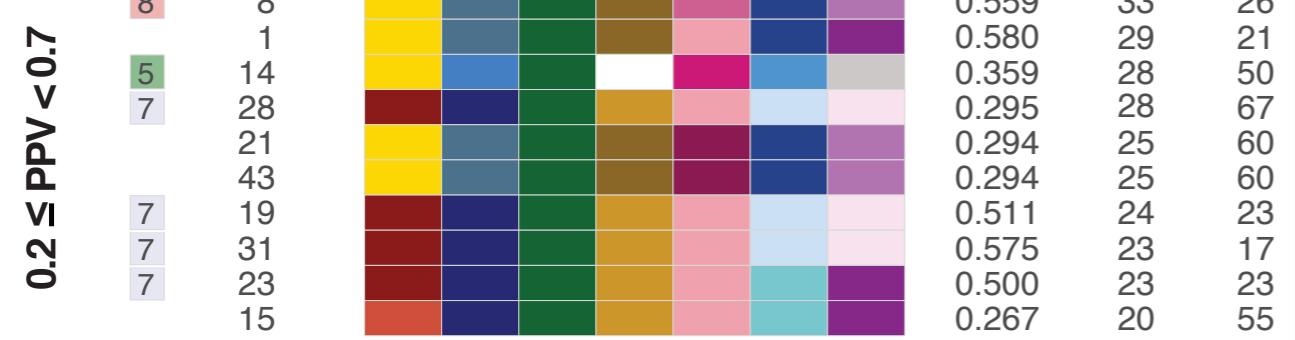
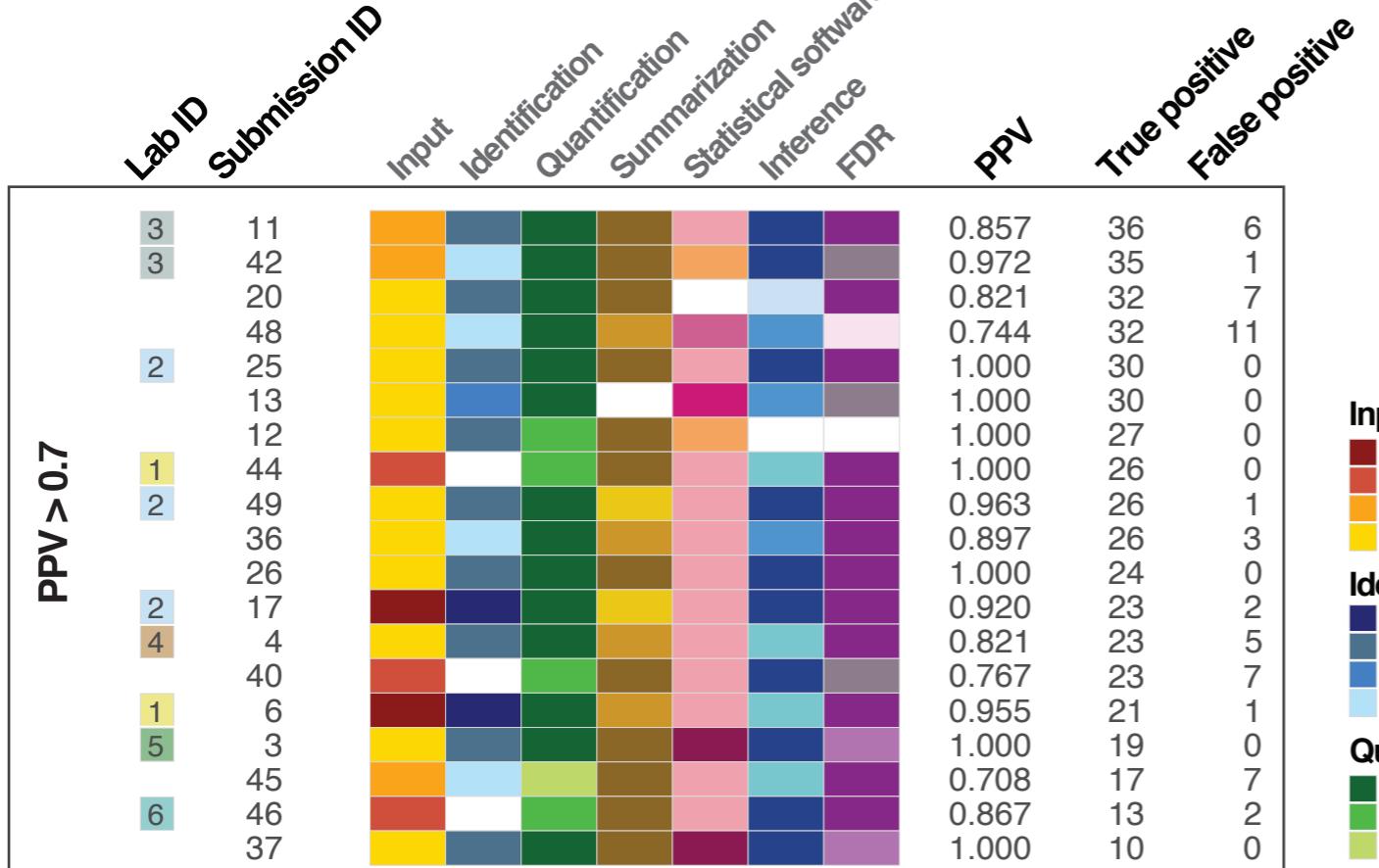
Input data

- Peaks
- Peptide IDs
- Raw+check
- Raw
- NA



SUMMARY OF SUBMISSIONS

USER EXPERTISE IS KEY



Input

- Peaks
- Peptide ids
- Raw+check
- Raw

Identification

- Skyline
- MaxQuant
- Progenesis
- Others

Quantification

- Feature intensity
- Spectral counting
- Hybrid

Summarization

- Protein summarization / Protein-level inference
- Peptide summarization / Protein-level inference
- Peptide summarization / Peptide-level inference

Statistical software

- Persus
- Progenesis QI
- Others
- R, Excel, MatLab, Python
- In-house scripts

Inference

- t-test / SAM's t test
- ANOVA
- Linear (mixed-effects) model
- Others

FDR

- Benjamini Hochberg
- Permutation FDR
- Others
- Manual validation
- FC cutoff
- No adjustment

No information

USER EXPERTISE IS KEY

PPV > 0.7	Lab ID	Submission ID	Process Flow						PPV	True positive	False positive
			Input	Identification	Quantification	Summarization	Statistical software	Inference			
3	11								0.857	36	6
3	42								0.972	35	1
20	20								0.821	32	7
48	48								0.744	32	11
25	25								1.000	30	0
13	13								1.000	30	0
12	12								1.000	27	0
44	44								1.000	26	0
2	49								0.963	26	1
36	36								0.897	26	3
26	26								1.000	24	0
17	17								0.920	23	2
4	4								0.821	23	5
40	40								0.767	23	7
6	6								0.955	21	1
5	3								1.000	19	0
45	45								0.708	17	7
6	46								0.867	13	2
37	37								1.000	10	0

Positive predictive value =

true differentially abundant proteins

claimed differentially abundant proteins

Good

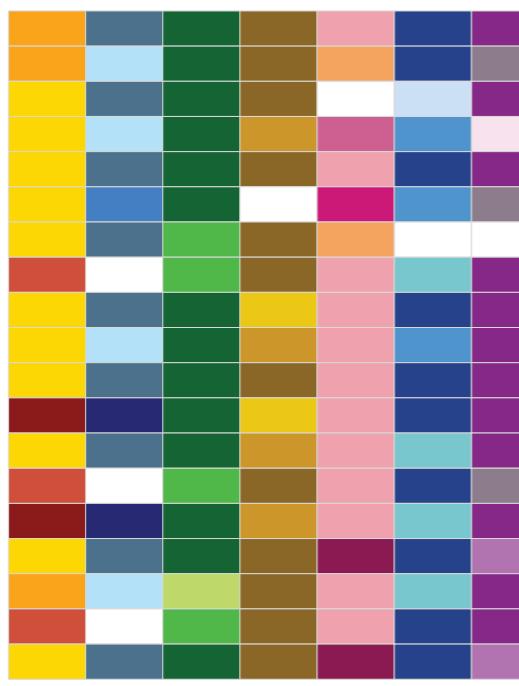
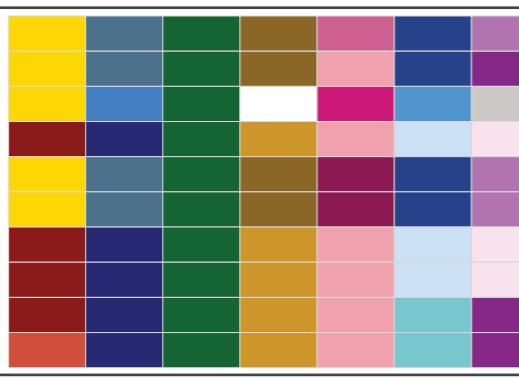
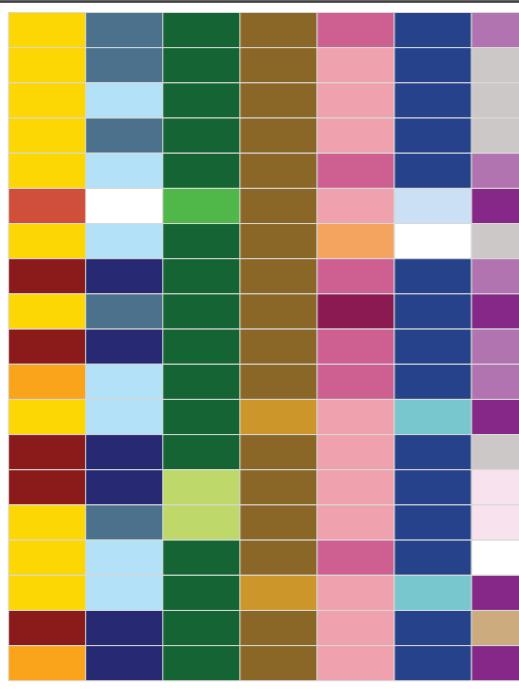
0.2 ≤ PPV < 0.7	Lab ID	Submission ID	Process Flow						PPV	True positive	False positive
			Input	Identification	Quantification	Summarization	Statistical software	Inference			
8	8								0.559	33	26
1	1								0.580	29	21
5	14								0.359	28	50
7	28								0.295	28	67
	21								0.294	25	60
	43								0.294	25	60
7	19								0.511	24	23
7	31								0.575	23	17
7	23								0.500	23	23
	15								0.267	20	55

Bad

PPV < 0.2	Lab ID	Submission ID	Process Flow						PPV	True positive	False positive
			Input	Identification	Quantification	Summarization	Statistical software	Inference			
8	30								0.174	32	152
9	16								0.145	30	177
	27								0.134	30	194
9	24								0.120	30	219
8	38								0.141	28	170
	47								0.181	27	122
	29								0.057	27	447
8	35								0.182	26	117
	35								0.126	26	180
8	22								0.145	25	148
	34								0.132	24	158
	2								0.091	24	240
4	39								0.074	21	264
9	33								0.085	18	194
	51								0.052	18	330
10	10								0.023	17	734
	5								0.008	17	2002
6	41								0.027	14	514
	7								0.172	5	24

Very bad

USER EXPERTISE IS KEY

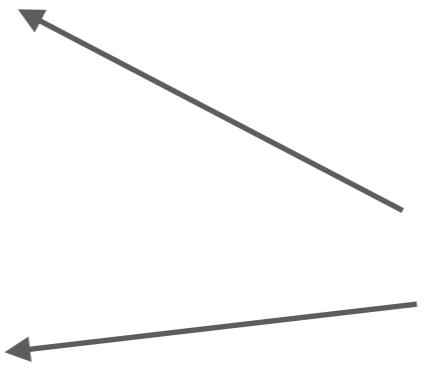
	Lab ID	Submission ID	Input	Identification	Quantification	Summarization	Statistical software	Inference	FDR	PPV	True positive	False positive
PPV > 0.7 	3	11	0.857	36	6							
	3	42	0.972	35	1							
	20	0.821	32	7								
	48	0.744	32	11								
2	25	1.000	30	0								
	13	1.000	30	0								
	12	1.000	27	0								
1	44	1.000	26	0								
2	49	0.963	26	1								
	36	0.897	26	3								
	26	1.000	24	0								
2	17	0.920	23	2								
4	4	0.821	23	5								
	40	0.767	23	7								
1	6	0.955	21	1								
5	3	1.000	19	0								
	45	0.708	17	7								
6	46	0.867	13	2								
	37	1.000	10	0								
0.2 ≤ PPV < 0.7 	8	8	0.559	33	26							
	1	0.580	29	21								
5	14	0.359	28	50								
7	28	0.295	28	67								
	21	0.294	25	60								
	43	0.294	25	60								
7	19	0.511	24	23								
7	31	0.575	23	17								
7	23	0.500	23	23								
	15	0.267	20	55								
PPV < 0.2 	8	30	0.174	32	152							
9	16	0.145	30	177								
	27	0.134	30	194								
9	24	0.120	30	219								
8	38	0.141	28	170								
	47	0.181	27	122								
	29	0.057	27	447								
8	35	0.182	26	117								
	35	0.126	26	180								
8	22	0.145	25	148								
	34	0.132	24	158								
8	2	0.091	24	240								
4	39	0.074	21	264								
9	33	0.085	18	194								
	9	0.052	18	330								
	51	0.023	17	734								
	10	0.008	17	2002								
	5	0.027	14	514								
6	41	0.172	5	24								
	7											

MaxQuant and Perseus

USER EXPERTISE IS KEY

	Lab ID	Submission ID	Input	Identification	Quantification	Summarization	Statistical software	Inference	FDR	PPV	True positive	False positive
PPV > 0.7												
3	11									0.857	36	6
3	42									0.972	35	1
20										0.821	32	7
48										0.744	32	11
2	25									1.000	30	0
13										1.000	30	0
12										1.000	27	0
1	44									1.000	26	0
2	49									0.963	26	1
36										0.897	26	3
2	26									1.000	24	0
4	17									0.920	23	2
4	4									0.821	23	5
40										0.767	23	7
1	6									0.955	21	1
5	3									1.000	19	0
6	45									0.708	17	7
6	46									0.867	13	2
6	37									1.000	10	0
0.2 ≤ PPV < 0.7												
8	8									0.559	33	26
	1									0.580	29	21
5	14									0.359	28	50
7	28									0.295	28	67
	21									0.294	25	60
7	43									0.294	25	60
7	19									0.511	24	23
7	31									0.575	23	17
7	23									0.500	23	23
	15									0.267	20	55
PPV < 0.2												
8	30									0.174	32	152
9	16									0.145	30	177
	27									0.134	30	194
9	24									0.120	30	219
8	38									0.141	28	170
	47									0.181	27	122
8	29									0.057	27	447
8	35									0.182	26	117
8	22									0.126	26	180
8	34									0.145	25	148
	2									0.132	24	158
4	39									0.091	24	240
9	33									0.074	21	264
	9									0.085	18	194
51										0.052	18	330
10										0.023	17	734
5										0.008	17	2002
6	41									0.027	14	514
6	7									0.172	5	24

Skyline and linear modeling in R



USER EXPERTISE IS KEY

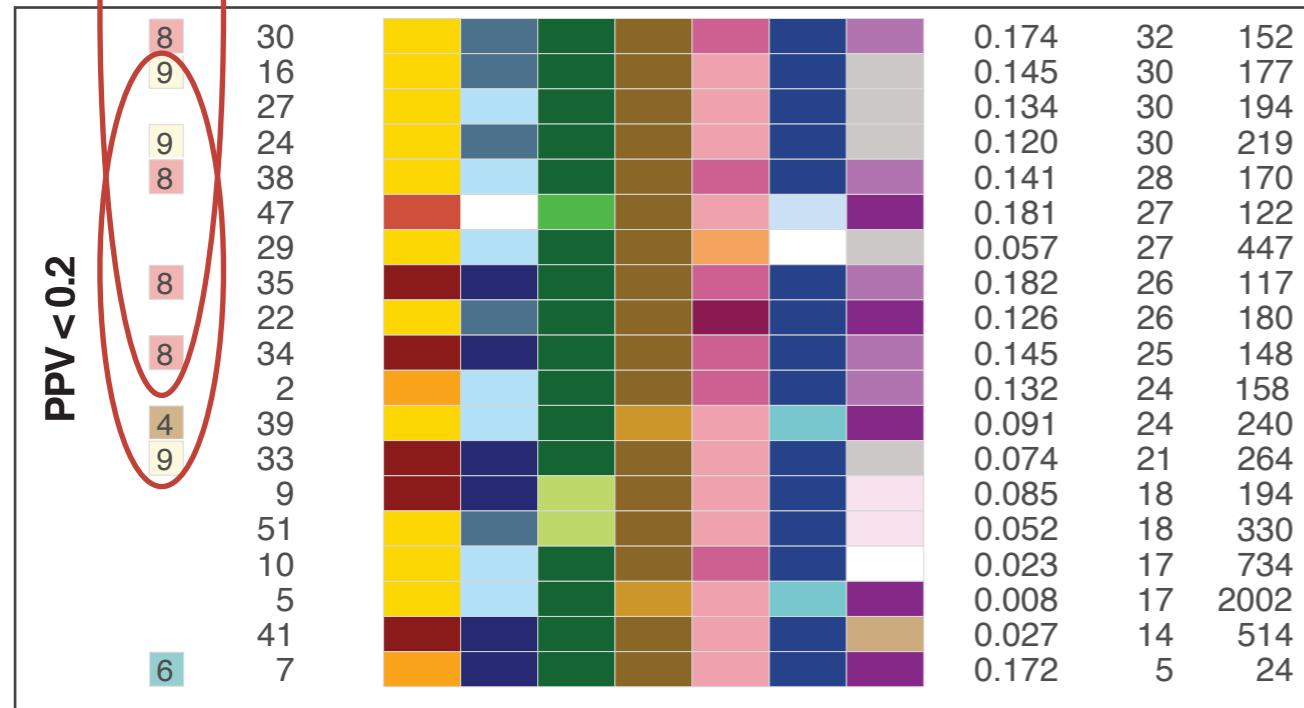
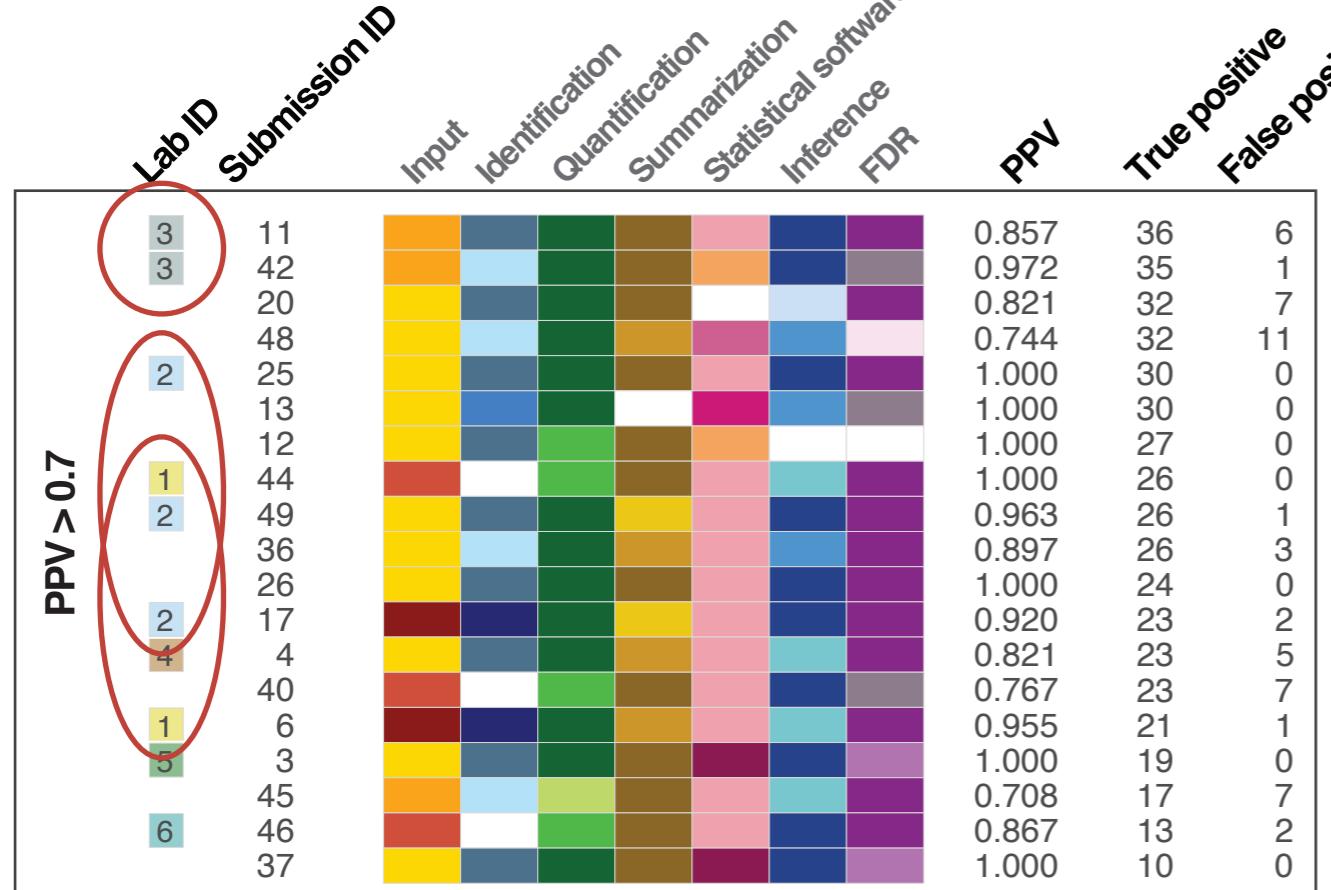
Lab ID	Submission ID	Input	Identification	Quantification	Summarization	Statistical software	Inference	FDR	PPV	True positive	False positive
3	11								0.857	36	6
3	42								0.972	35	1
20	20								0.821	32	7
48	48								0.744	32	11
2	25								1.000	30	0
13	13								1.000	30	0
12	12								1.000	27	0
1	44								1.000	26	0
2	49								0.963	26	1
36	36								0.897	26	3
2	26								1.000	24	0
4	17								0.920	23	2
4	4								0.821	23	5
40	40								0.767	23	7
1	6								0.955	21	1
5	3								1.000	19	0
45	45								0.708	17	7
6	46								0.867	13	2
	37								1.000	10	0

PPV > 0.7	PPV ≤ 0.7	PPV < 0.2
8	8	30
1	1	16
5	14	27
7	28	24
	21	38
	43	47
7	19	29
7	31	35
7	23	35
	15	22

PPV < 0.2	PPV > 0.7	PPV ≤ 0.7
8	30	8
9	16	16
	27	27
9	24	24
8	38	38
	47	47
8	29	29
8	35	35
8	35	35
8	22	22
8	34	34
2	2	2
4	39	39
9	33	33
	9	9
51	51	51
10	10	10
5	5	5
41	41	41
6	7	7

Compared peak intensity vs spectral counts

USER EXPERTISE IS KEY



Input

- Peaks
- Peptide ids
- Raw+check
- Raw

Identification

- Skyline
- MaxQuant
- Progenesis
- Others

Quantification

- Feature intensity
- Spectral counting
- Hybrid

Summarization

- Protein summarization / Protein-level inference
- Peptide summarization / Protein-level inference
- Peptide summarization / Peptide-level inference

Statistical software

- Persus
- Progenesis QI
- Others
- R, Excel, MatLab, Python
- In-house scripts

Inference

- t-test / SAM's t test
- ANOVA
- Linear (mixed-effects) model
- Others

FDR

- Benjamini Hochberg
- Permutation FDR
- Others
- Manual validation
- FC cutoff
- No adjustment

No information

Article

[<> Previous Article](#) [<> Next Article](#) [Table of Contents](#)

ABRF Proteome Informatics Research Group (iPRG) 2015 Study: Detection of Differentially Abundant Proteins in Label-Free Quantitative LC–MS/MS Experiments



Meena Choi^{#†} Zeynep F. Eren-Dogru^{#†}, Christopher Colangelo[§], John Cottrell[¶], Michael R. Hoopmann[¶], Eugene A. Kapp[¶], Sangtae Kim[®], Henry Lam[□], Thomas A. Neubert[¶], Magnus Palmblad[○], Brett S. Phinney^{*}, Susan T. Weintraub[△], Brendan MacLean[▲], and Olga Vitek^{*†}

[#] Northeastern University, Boston, Massachusetts 02115, United States

[†] Mugla Silki Kocman University, 48000 Mugla, Turkey

[§] Primary Ion, LLC, Old Lyme, Connecticut 06371, United States

[¶] Matrix Science Ltd., London W1U 7GB, U.K.

[○] Institute for Systems Biology, Seattle, Washington 98109, United States

[□] Walter and Eliza Hall Institute of Medical Research, Melbourne 3052, Australia

[®] Pacific Northwest National Laboratory, Richland, Washington 99354, United States

[▲] Department of Chemical and Biomolecular Engineering and Division of Biomedical Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

[●] Skirball Institute and Department of Biochemistry and Molecular Pharmacology, New York University School of Medicine, New York, New York 10016, United States

[○] Center for Proteomics and Metabolomics, Leiden University Medical Center, 2333 ZA Leiden, The Netherlands

[▪] University of California at Davis, Davis, California 95616, United States

[△] University of Texas Health Science Center at San Antonio, San Antonio, Texas 78229, United States

[▲] University of Washington, Seattle, Washington 98105, United States

J. Proteome Res., 2017, 16 (2), pp 945–957

DOI: 10.1021/acs.jproteome.6b00881

Publication Date (Web): December 19, 2016

Copyright © 2016 American Chemical Society

*E-mail: o.vitek@neu.edu. Tel: 617-370-2194.

Article Options

ACS ActiveView PDF

Hi Res Print, Annotate, Rotate/Zoom QuickView

[Abstract](#)

[Supporting Info](#)

PDF (3135 KB)

[Figures](#)

PDF w/ Links (885 KB)

[References](#)

Full Text HTML

Add to ACS ChemWorx

Add to Favorites

Download Citation

Email a Colleague

Order Reprints

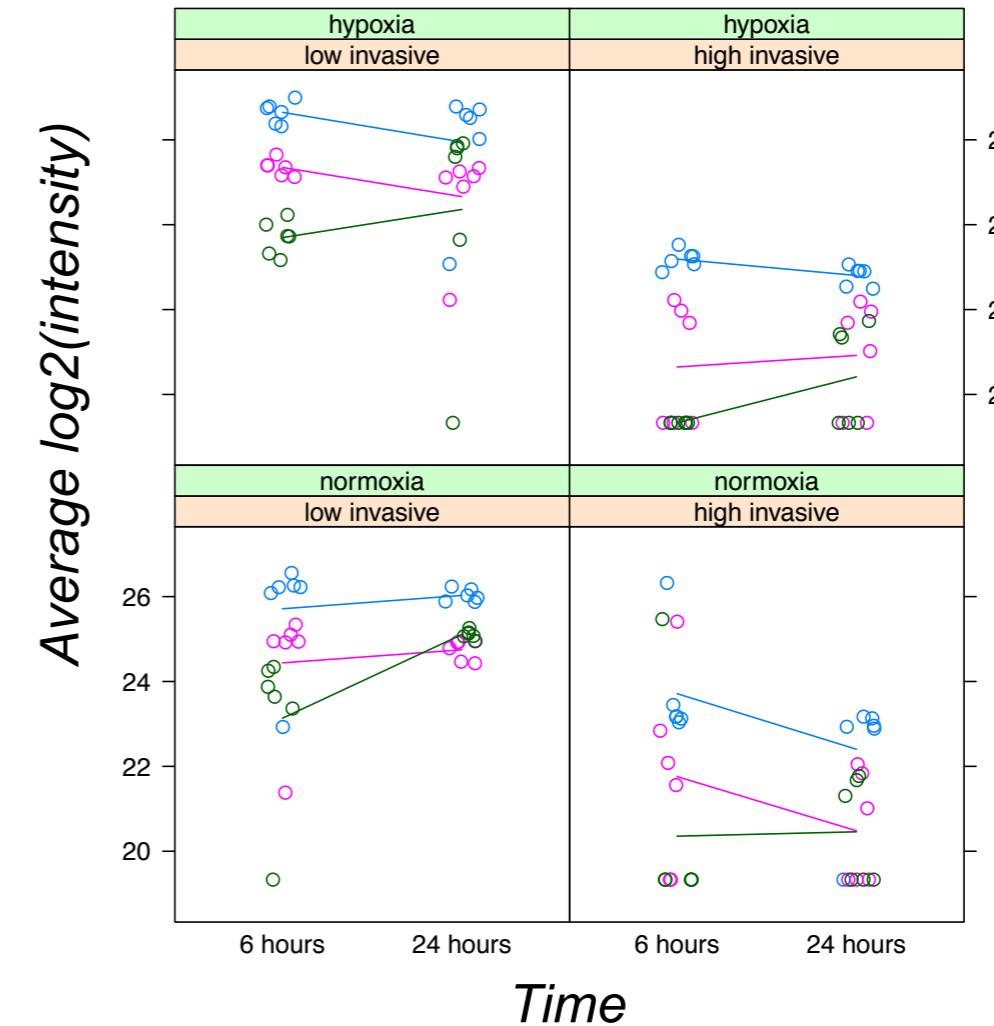
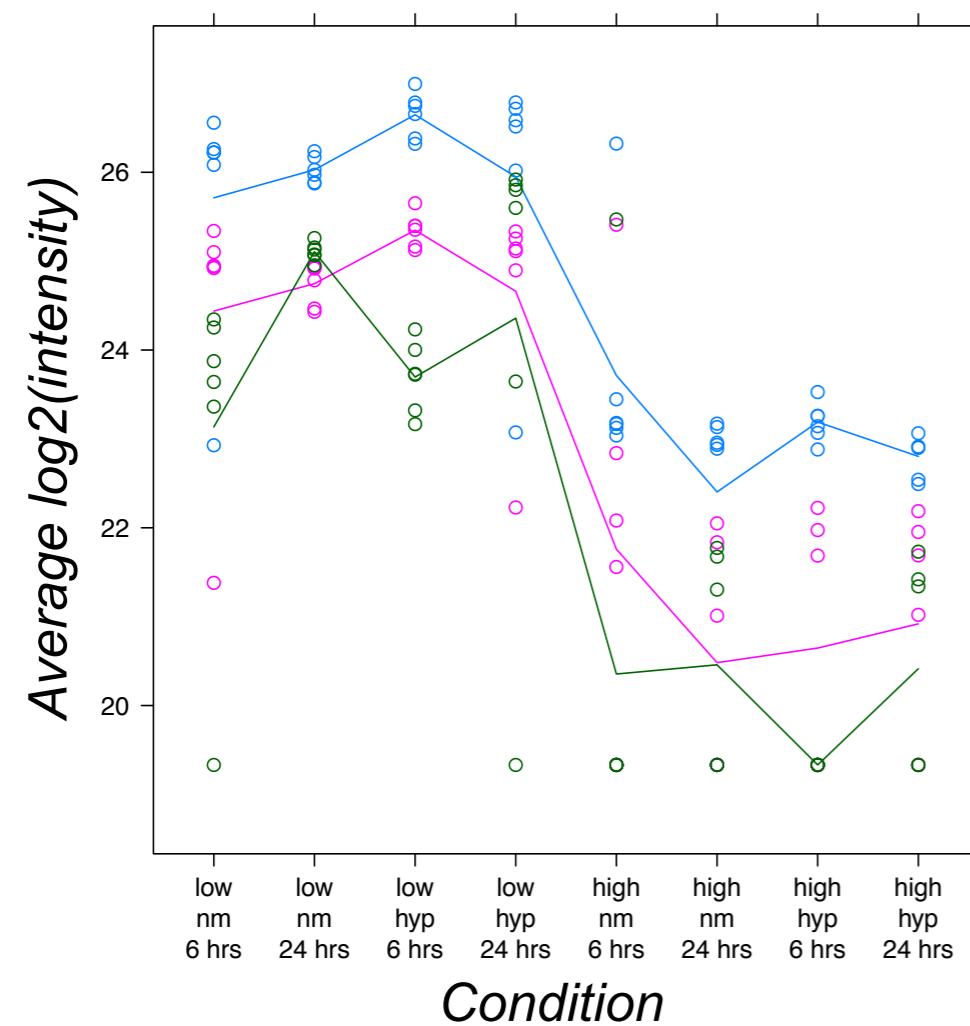
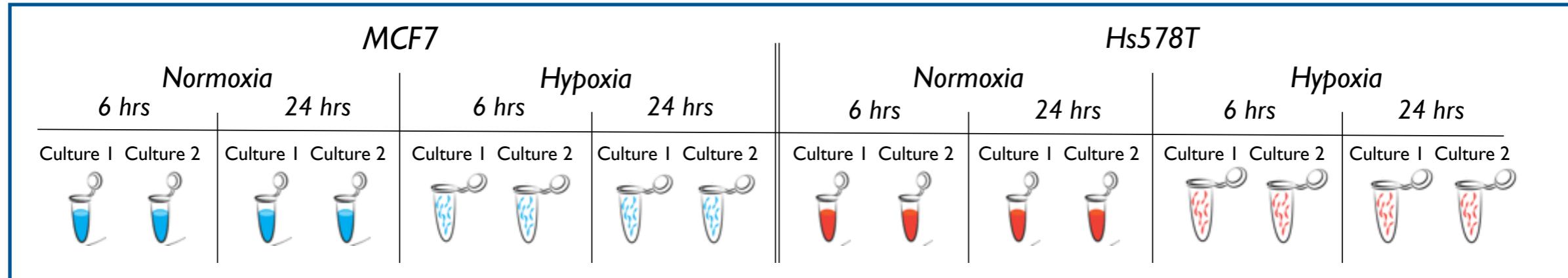
Rights & Permissions

OUTLINE

- Motivating example
 - ABRF iPRG study
- MSstats
 - Statistical relative quantification of proteins and peptides
 - Methods evaluation
- Extensions to MSstats
 - Assay characterization
 - System suitability and quality control

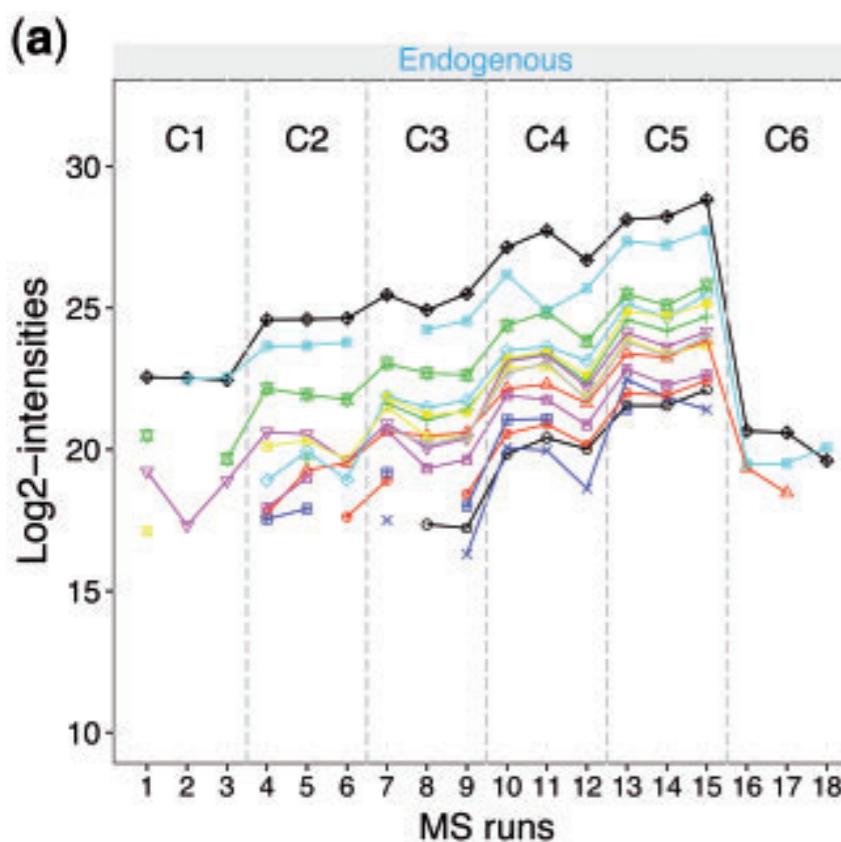
MOTIVATION: A LABEL-FREE EXPERIMENT¹⁶

Which proteins change in abundance?

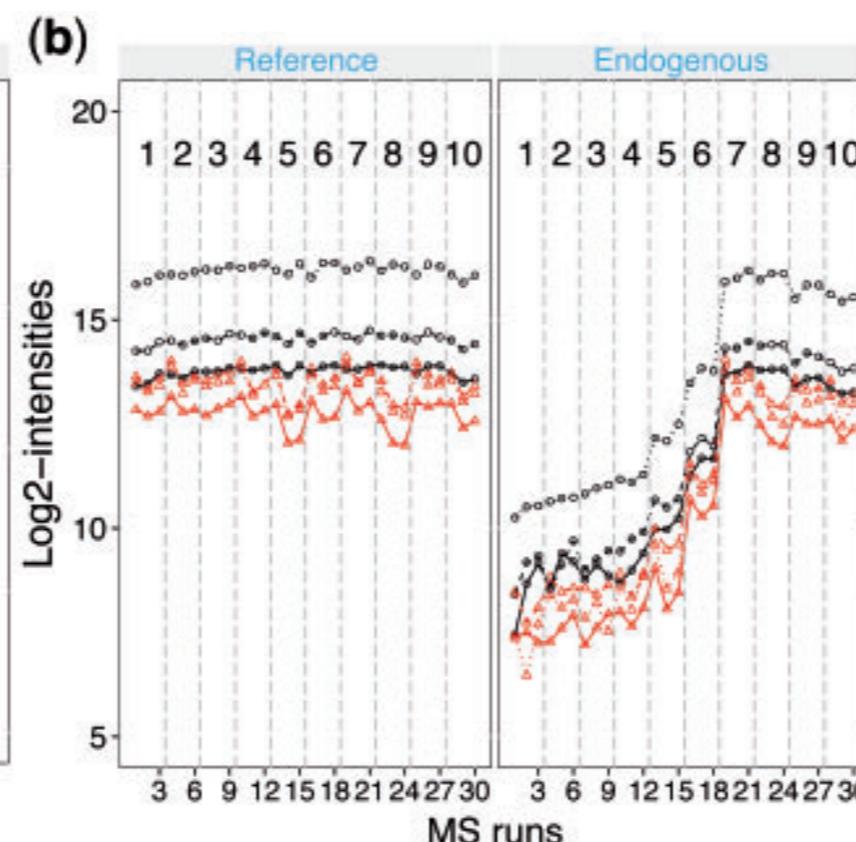


CHROMATOGRAPHY-BASED QUANTIFICATION

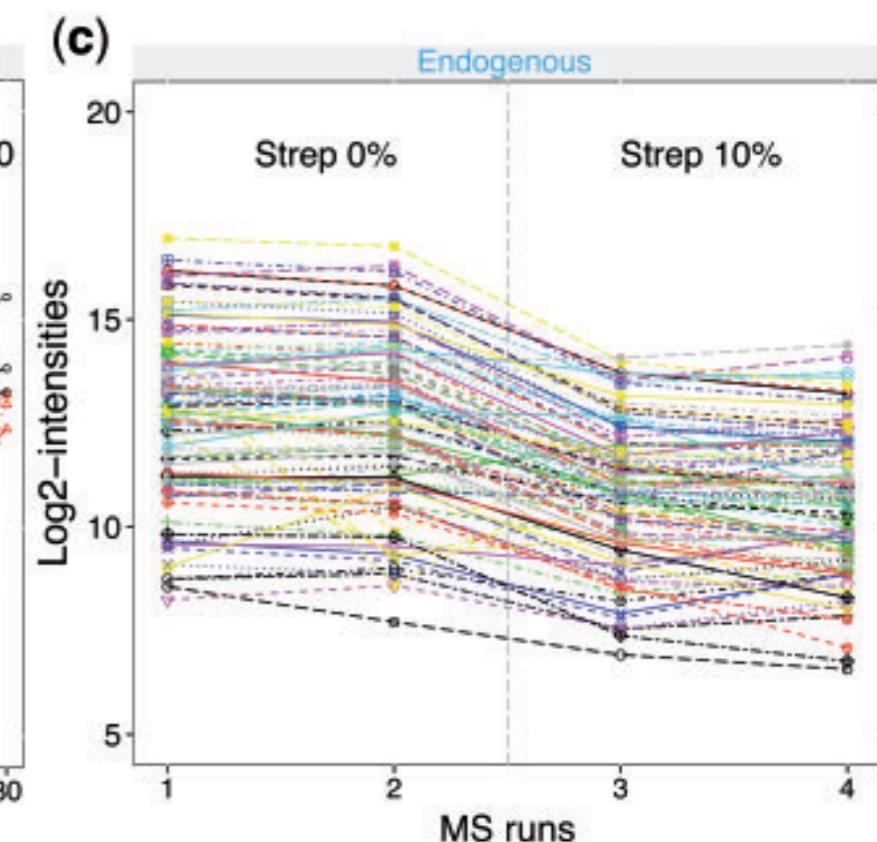
Data structures have similarities and differences



Data-dependent acquisition (DDA)



Selected reaction monitoring with labeled reference peptides (SRM)



Data-independent acquisition (DIA)

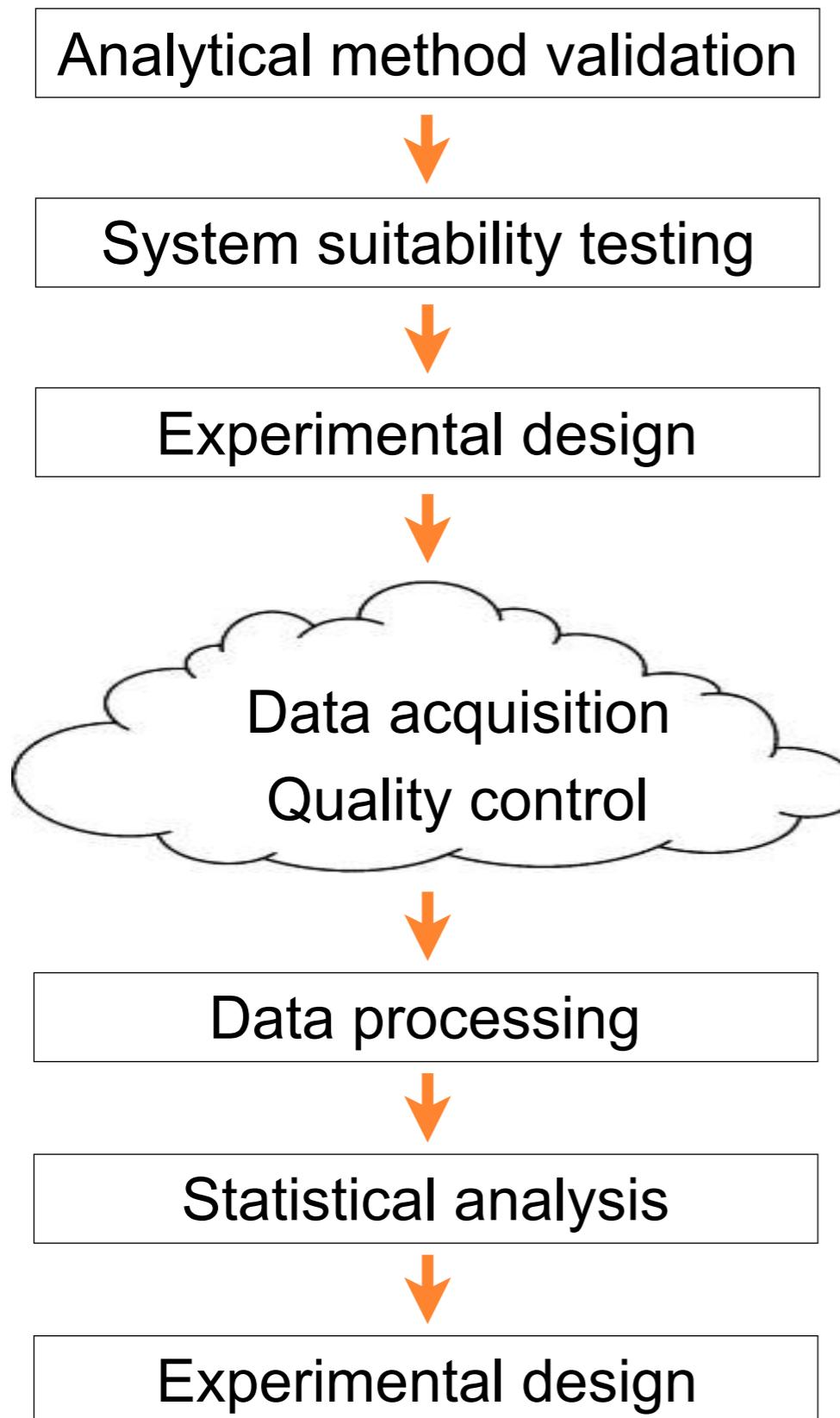
MSSTATS: CURRENT STATUS

- 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...

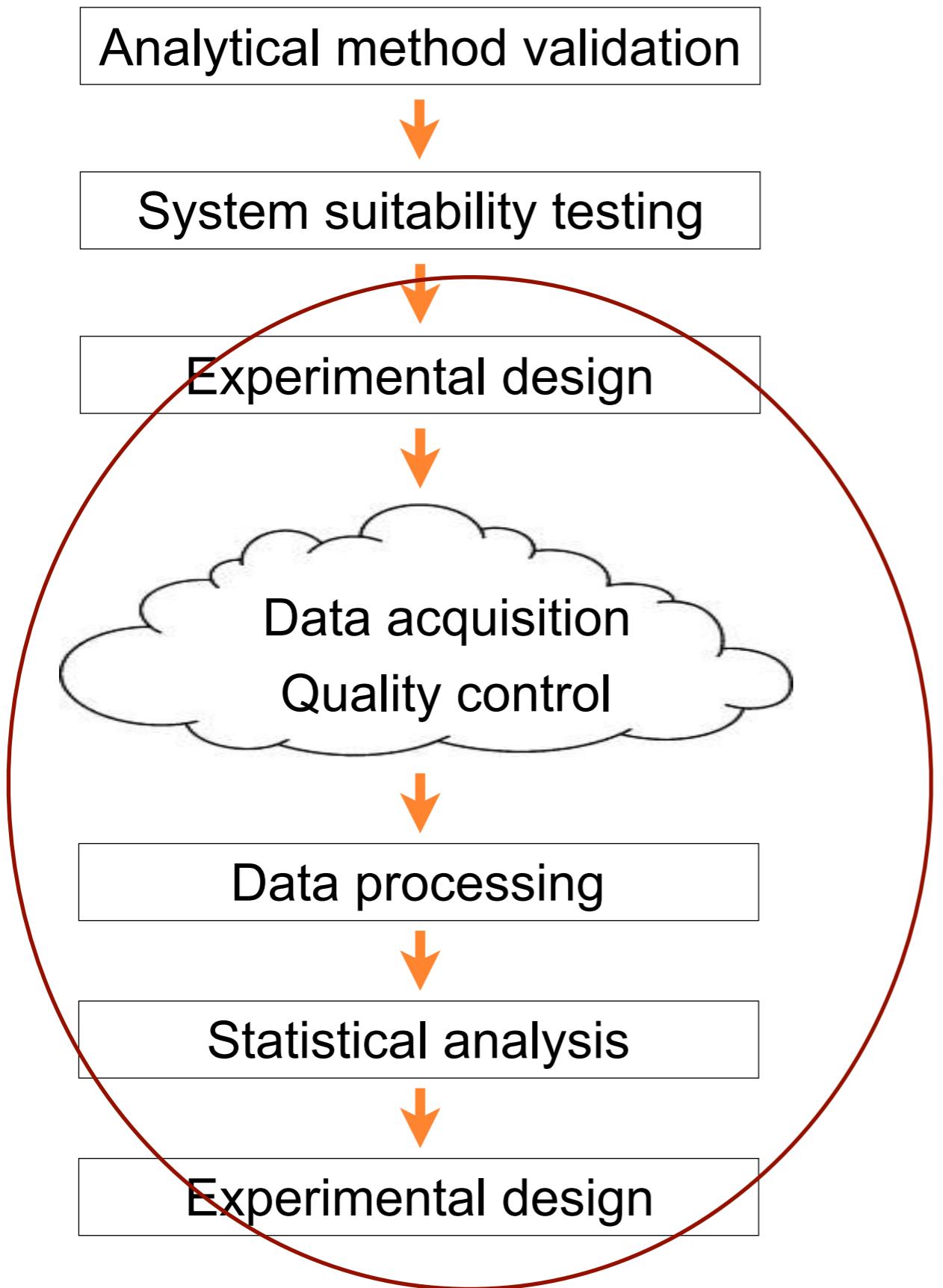
WHY R?

- *Researcher:*
 - Lightweight: minimal software/hardware requirements
 - Interactive: easy data exploration on a laptop
- *Developer:*
 - Large community: leverage state-of-the-art
 - Easy to customize/extend: e.g. include in existing pipelines
- *Science:*
 - Full transparency: open algorithms and code
 - Infrastructure for fully documentable workflows

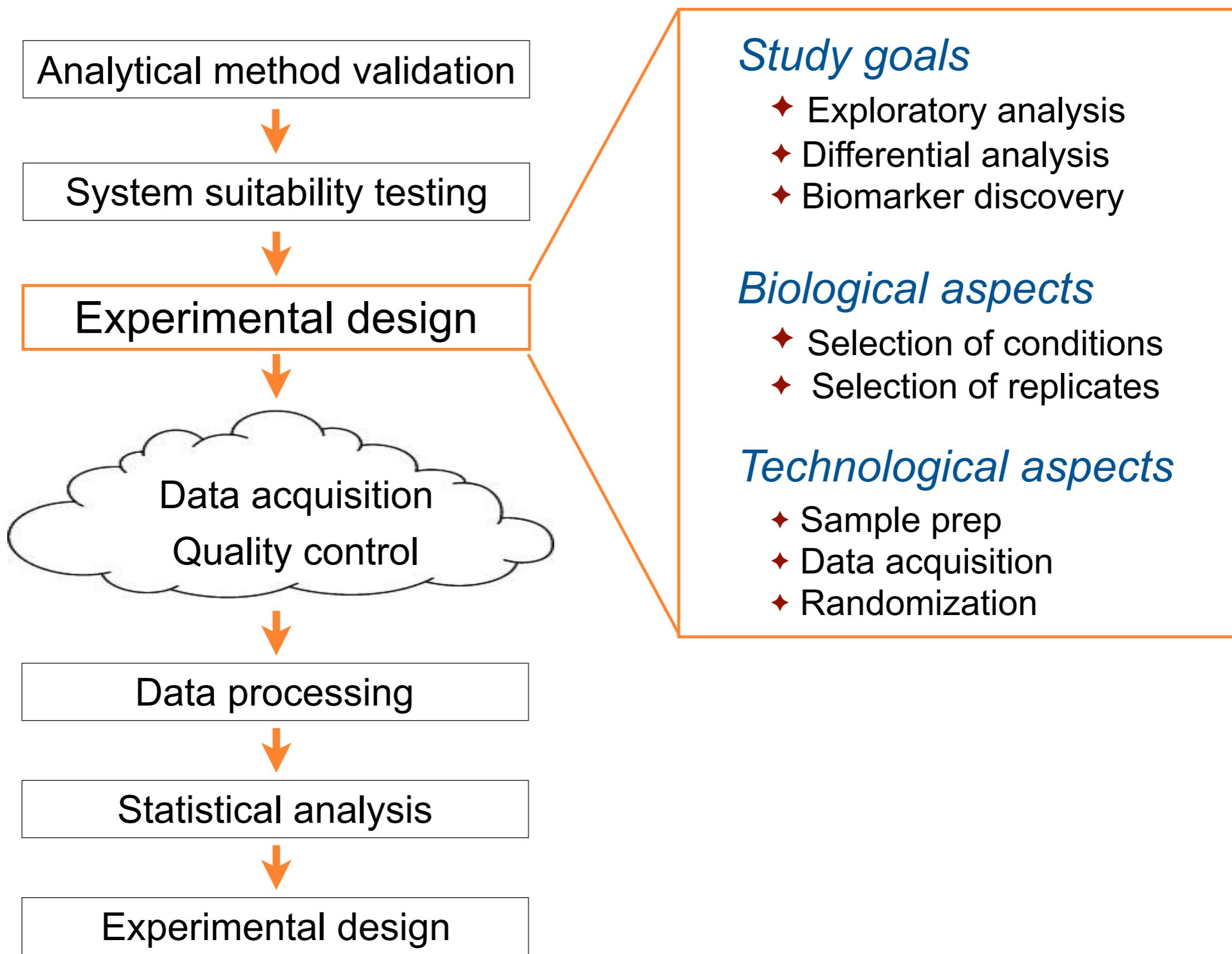
MS EXPERIMENT: STATISTICIAN'S VIEW



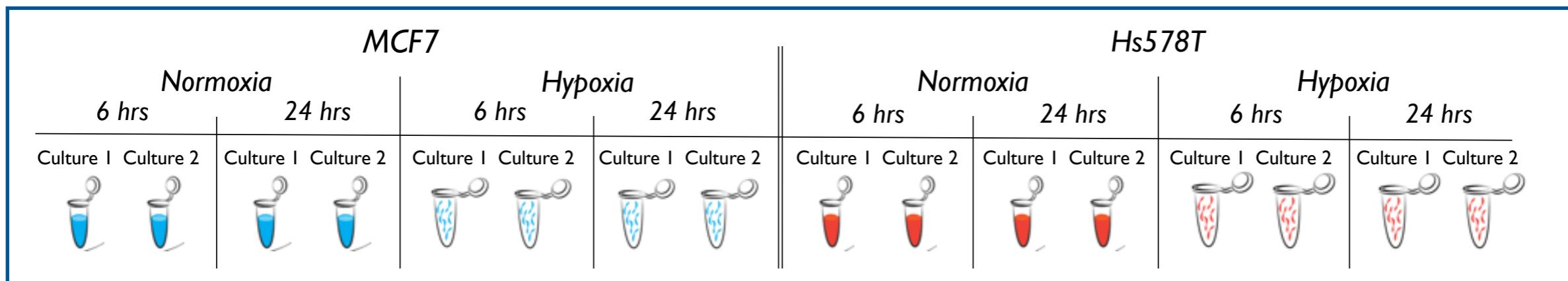
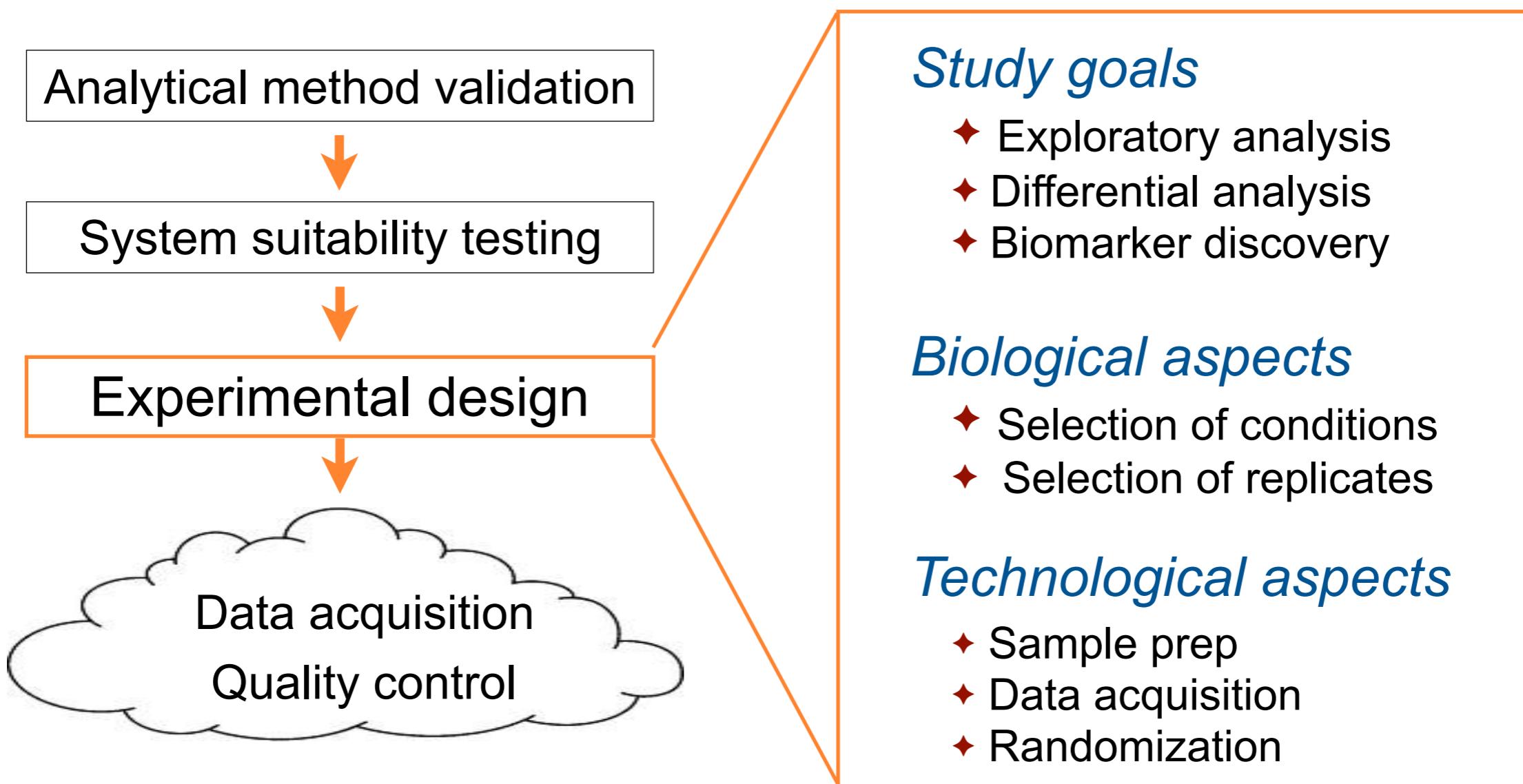
MS EXPERIMENT: STATISTICIAN'S VIEW



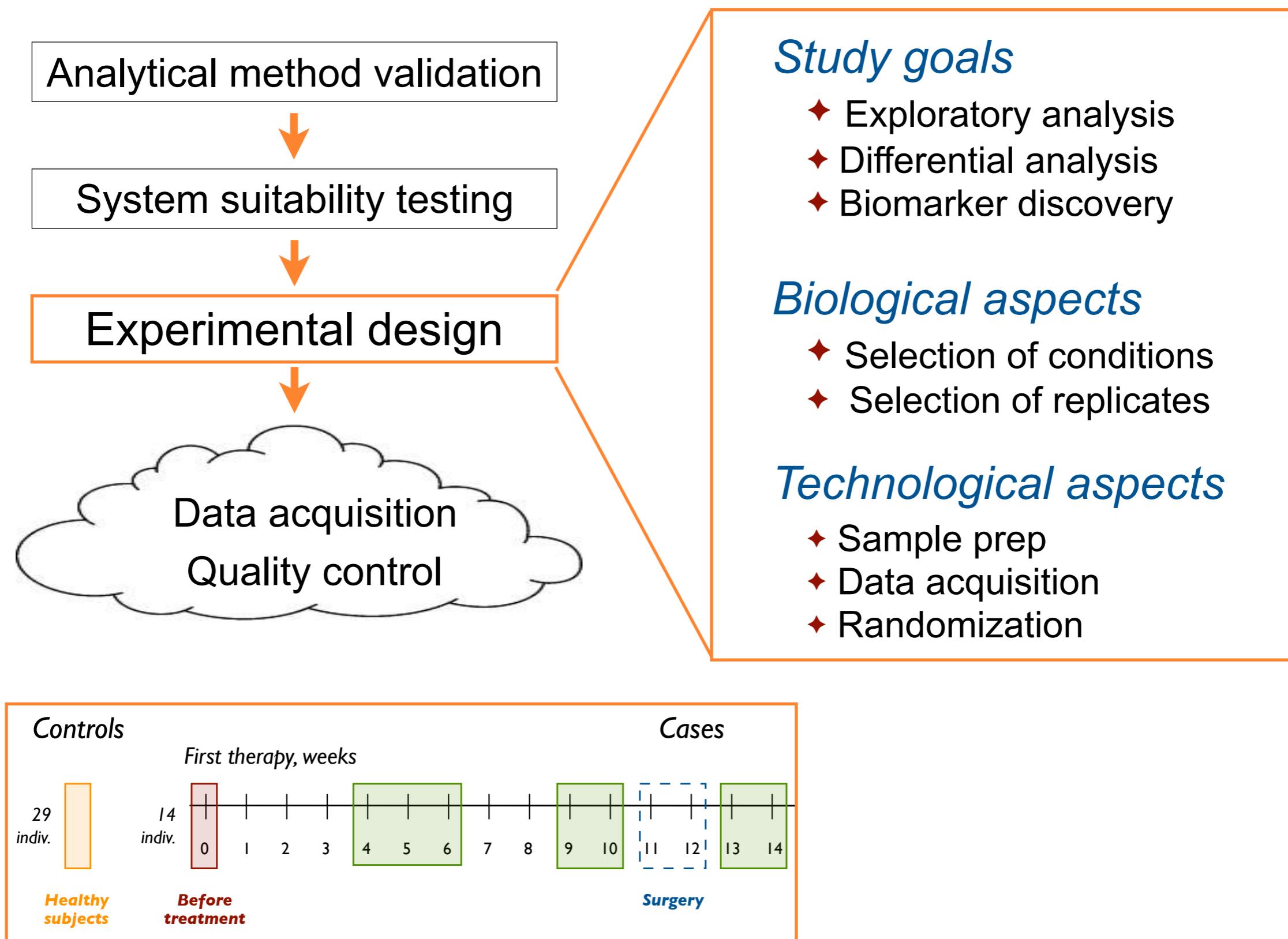
MS EXPERIMENT: STATISTICIAN'S VIEW



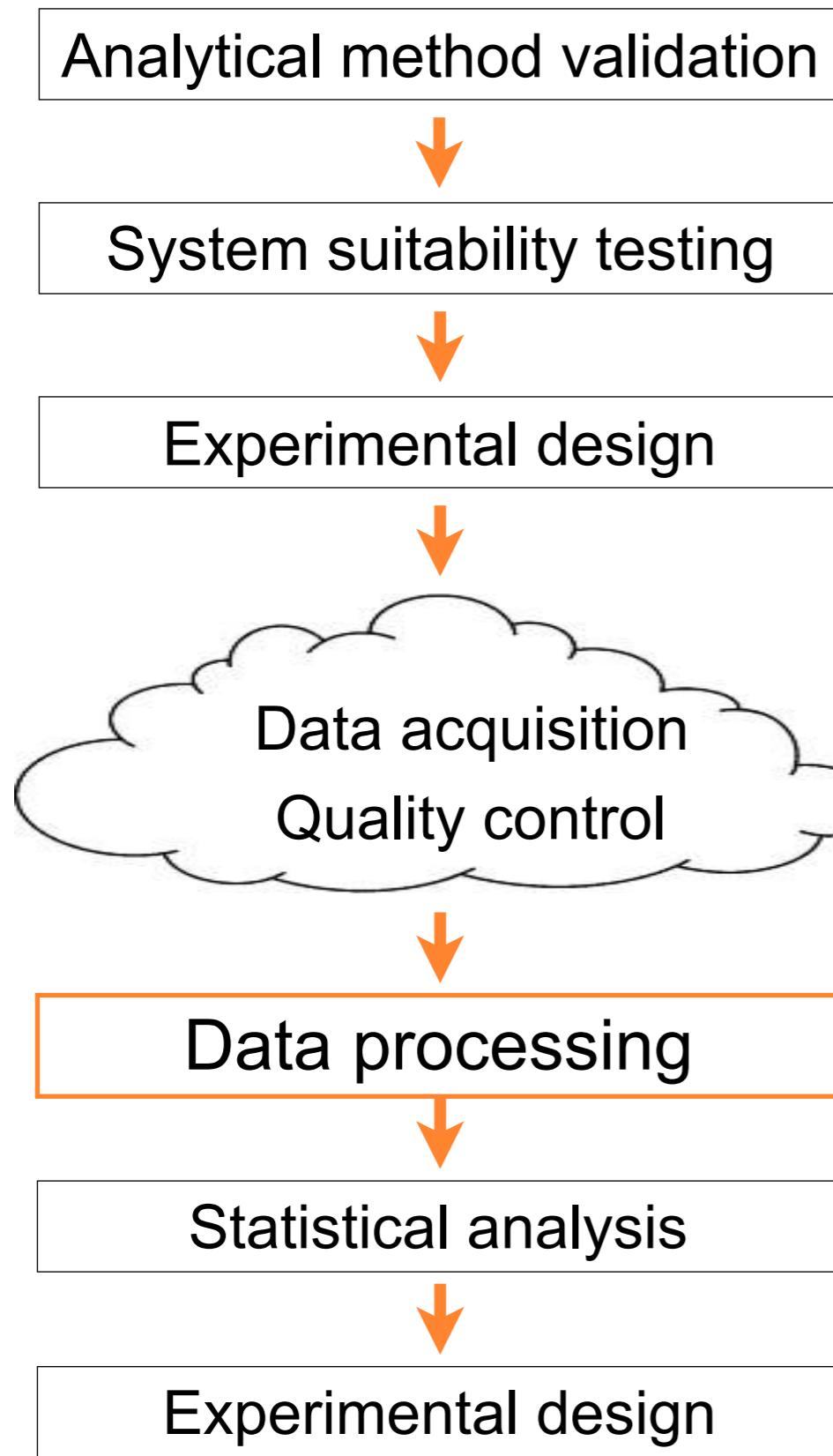
MS EXPERIMENT: STATISTICIAN'S VIEW



MS EXPERIMENT: STATISTICIAN'S VIEW



MS EXPERIMENT: STATISTICIAN'S VIEW



Input: list of identifies and quantified peaks

- ◆ MaxQuant
- ◆ OpenMS
- ◆ Progenesis
- ◆ Skyline
- ◆ SpectraNaut
- ◆ DIAUmpire
- ◆ ...

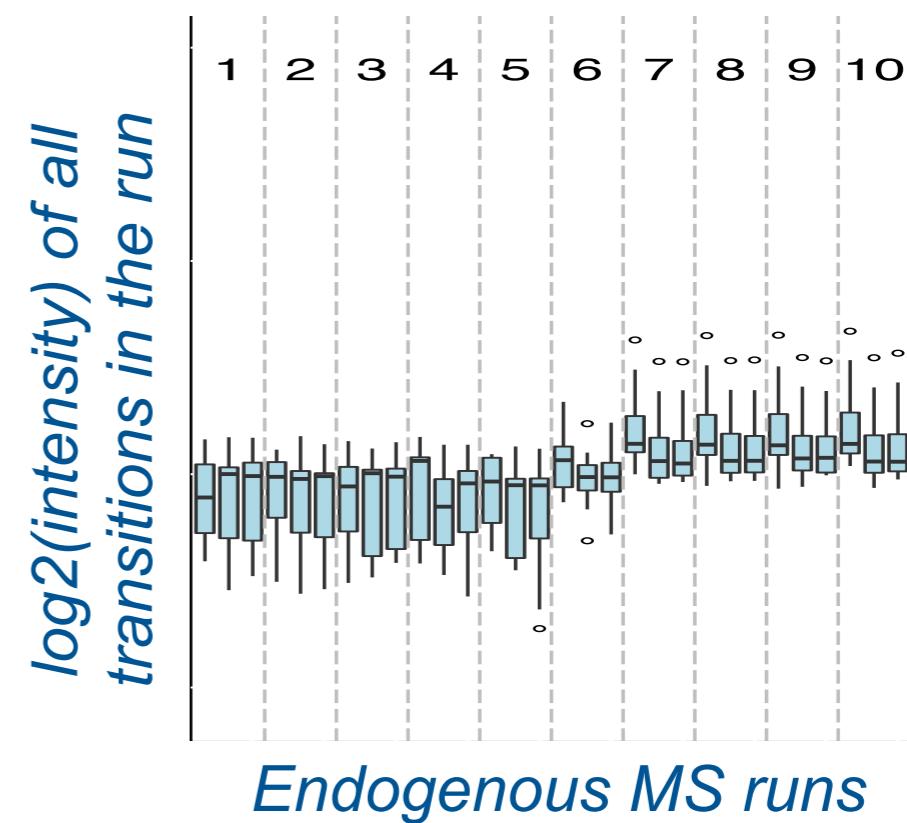
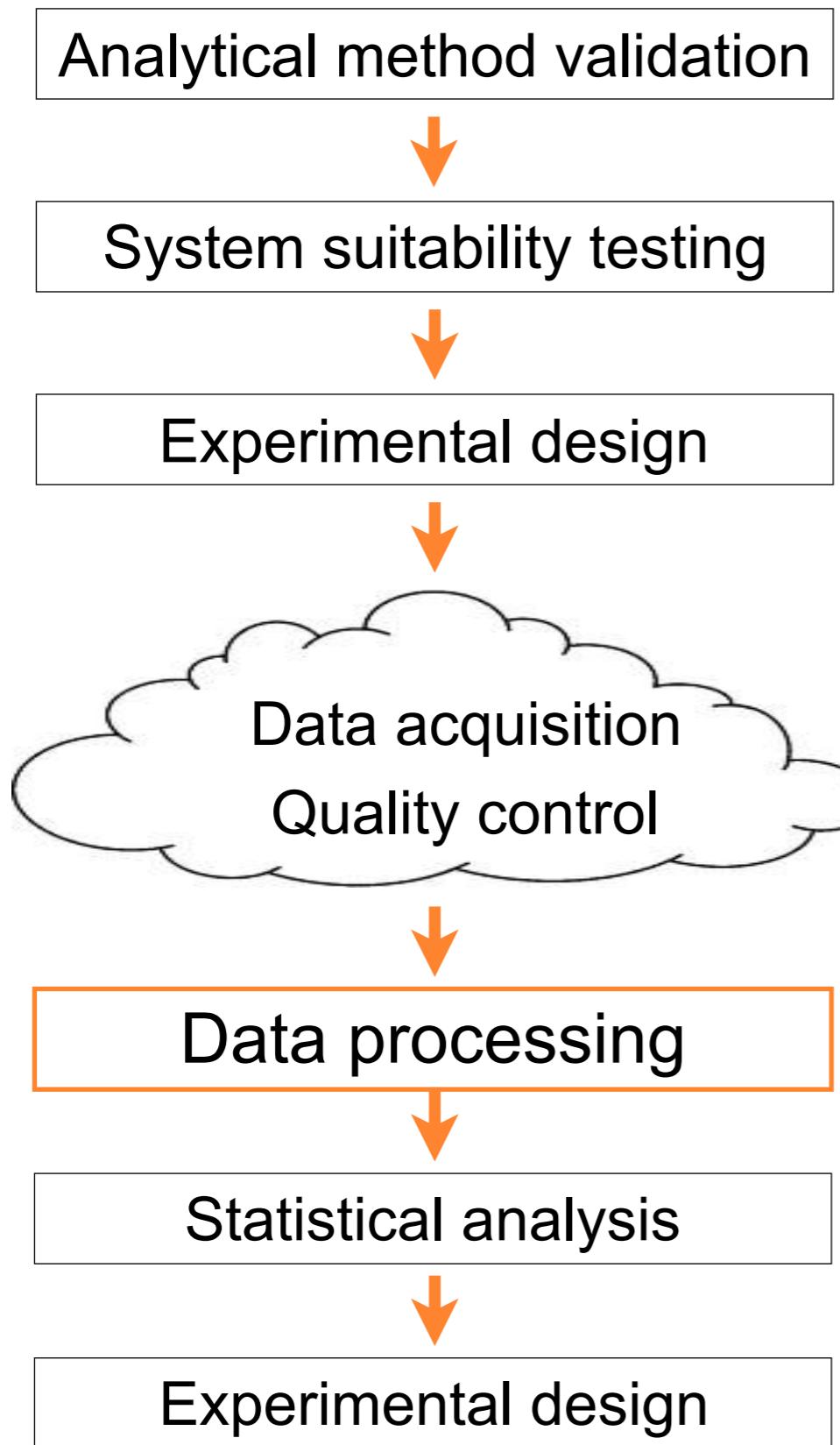
Steps

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

INPUT DATA REPRESENTATION

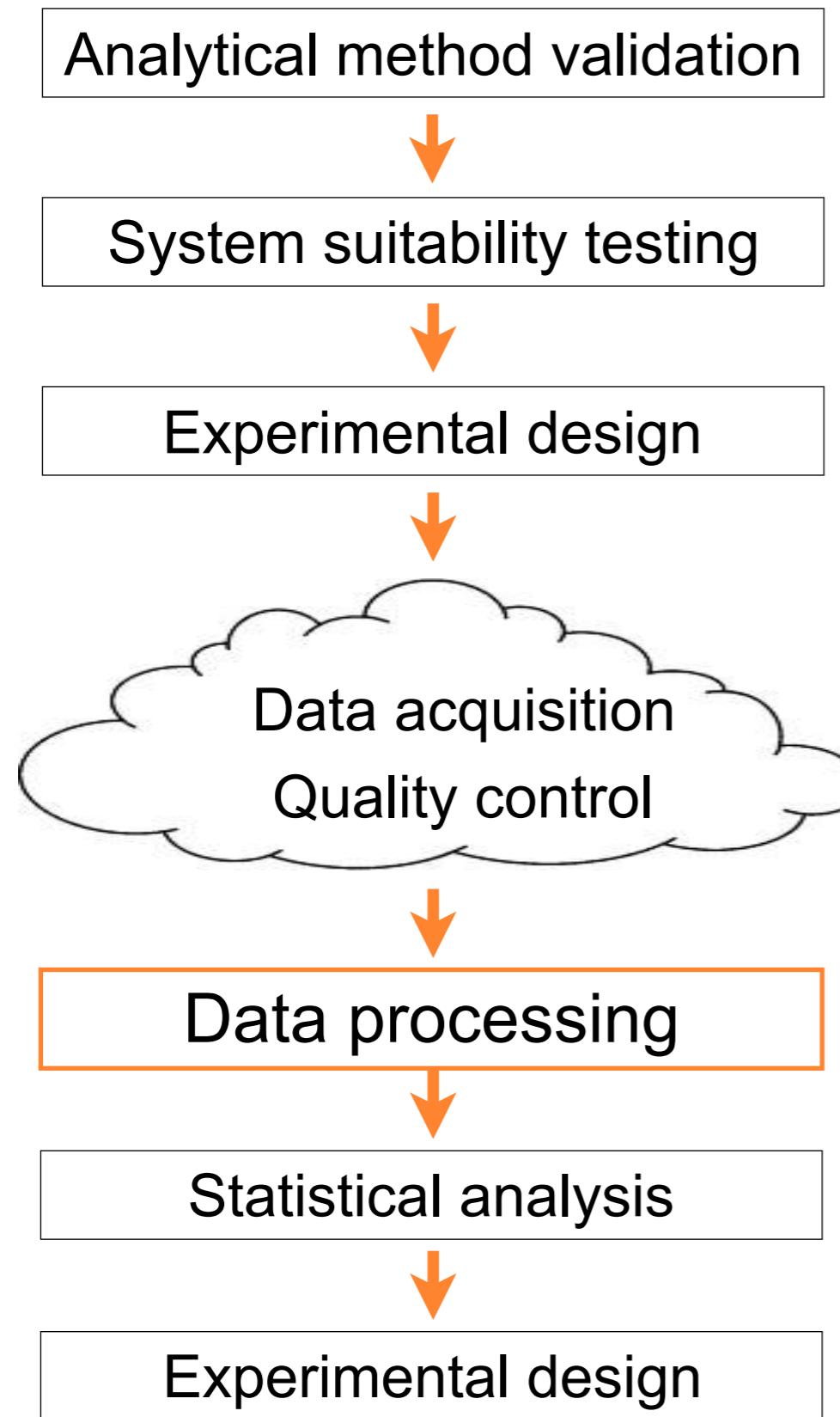
	A	B	C	D	E	F	G	H	I	J
1	ProteinName	PeptideSequence	PrecursorCharge	FragmentIon	ProductCharge	IsotopeLabelType	Condition	BioReplicate	Run	Intensity
2	ACEA	EILGHEIFFDWELP	3	y3	0	H	1	ReplA	1	66472.3847
3	ACEA	EILGHEIFFDWELP	3	y3	0	L	1	ReplA	1	5764.16228
4	ACEA	EILGHEIFFDWELP	3	y4	0	H	1	ReplA	1	101005.166
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

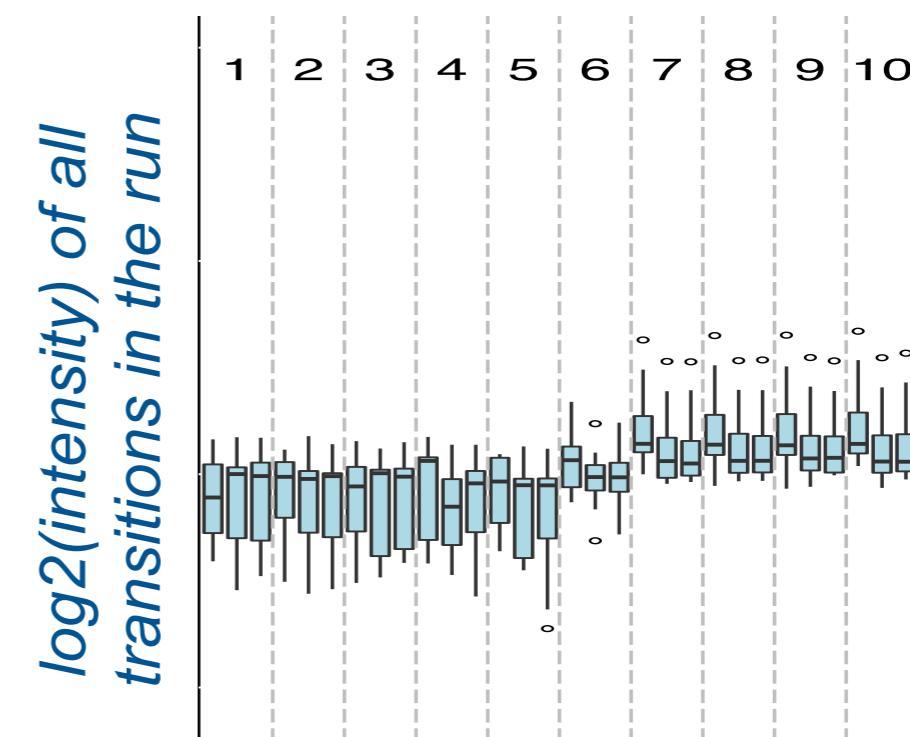


Endogenous MS runs

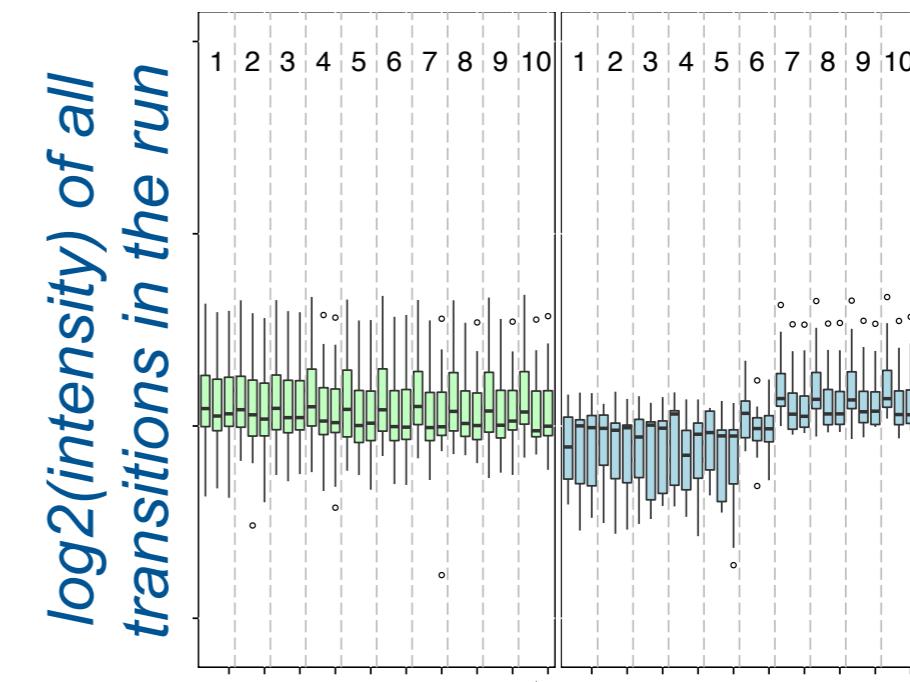
MS EXPERIMENT: STATISTICIAN'S VIEW



NORMALIZATION

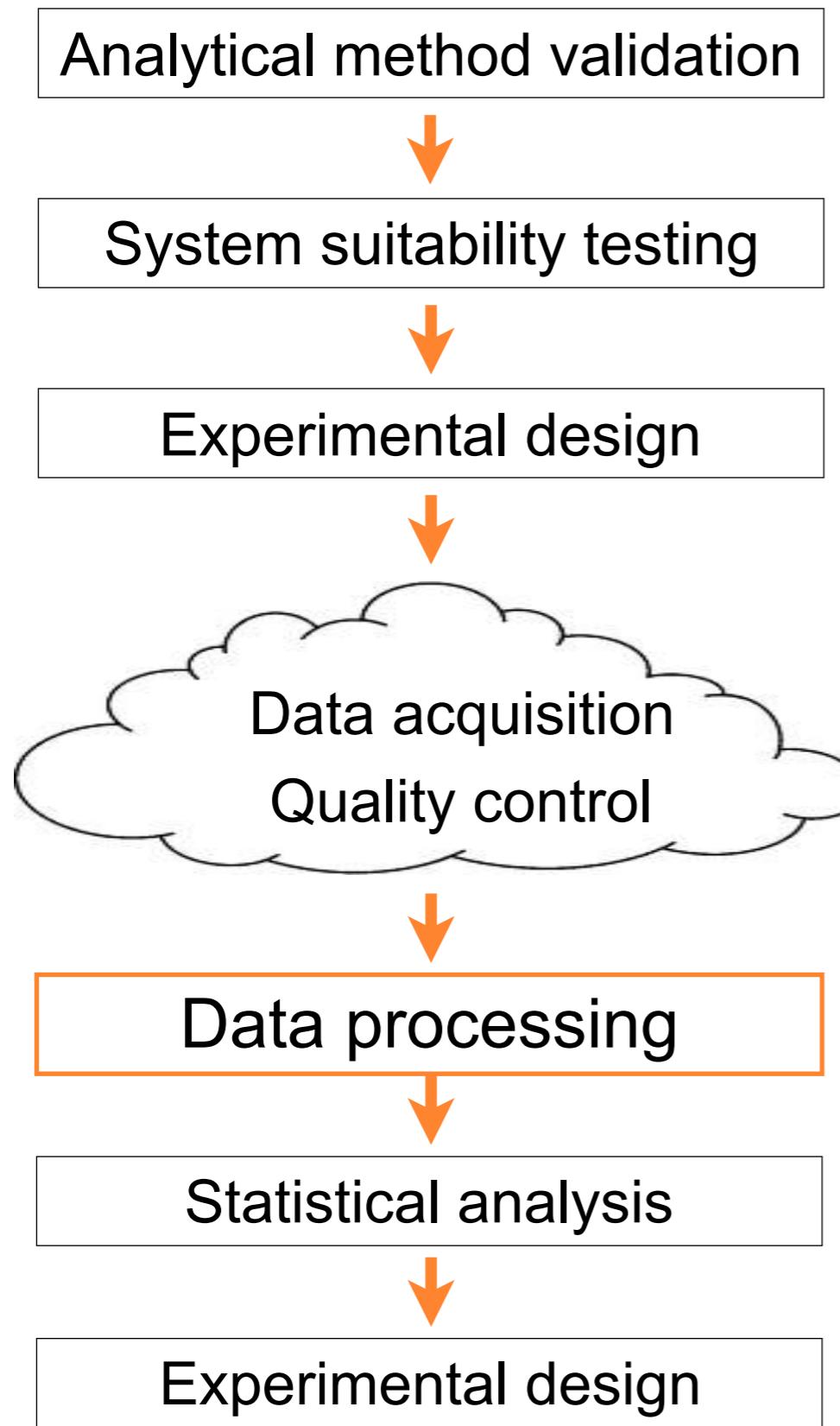


Endogenous MS runs

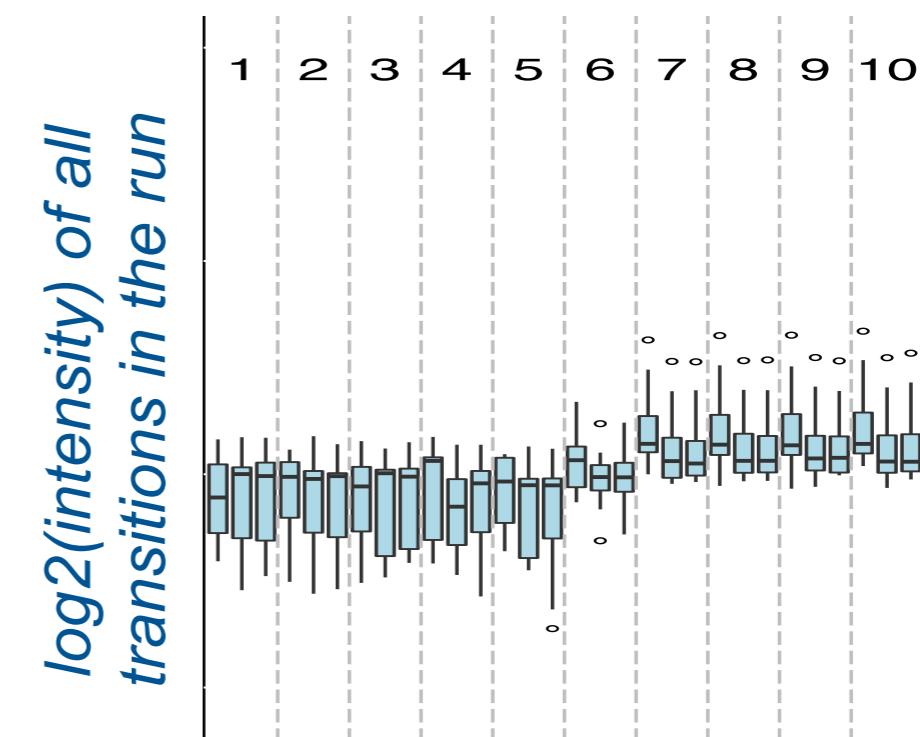


Reference Endogenous MS runs

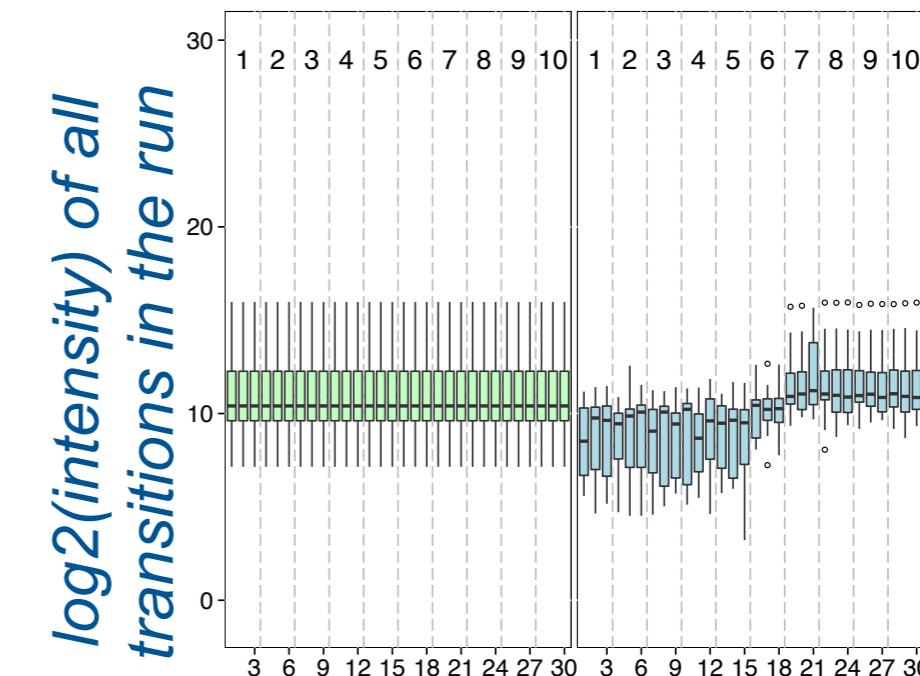
MS EXPERIMENT: STATISTICIAN'S VIEW



NORMALIZATION



Endogenous MS runs

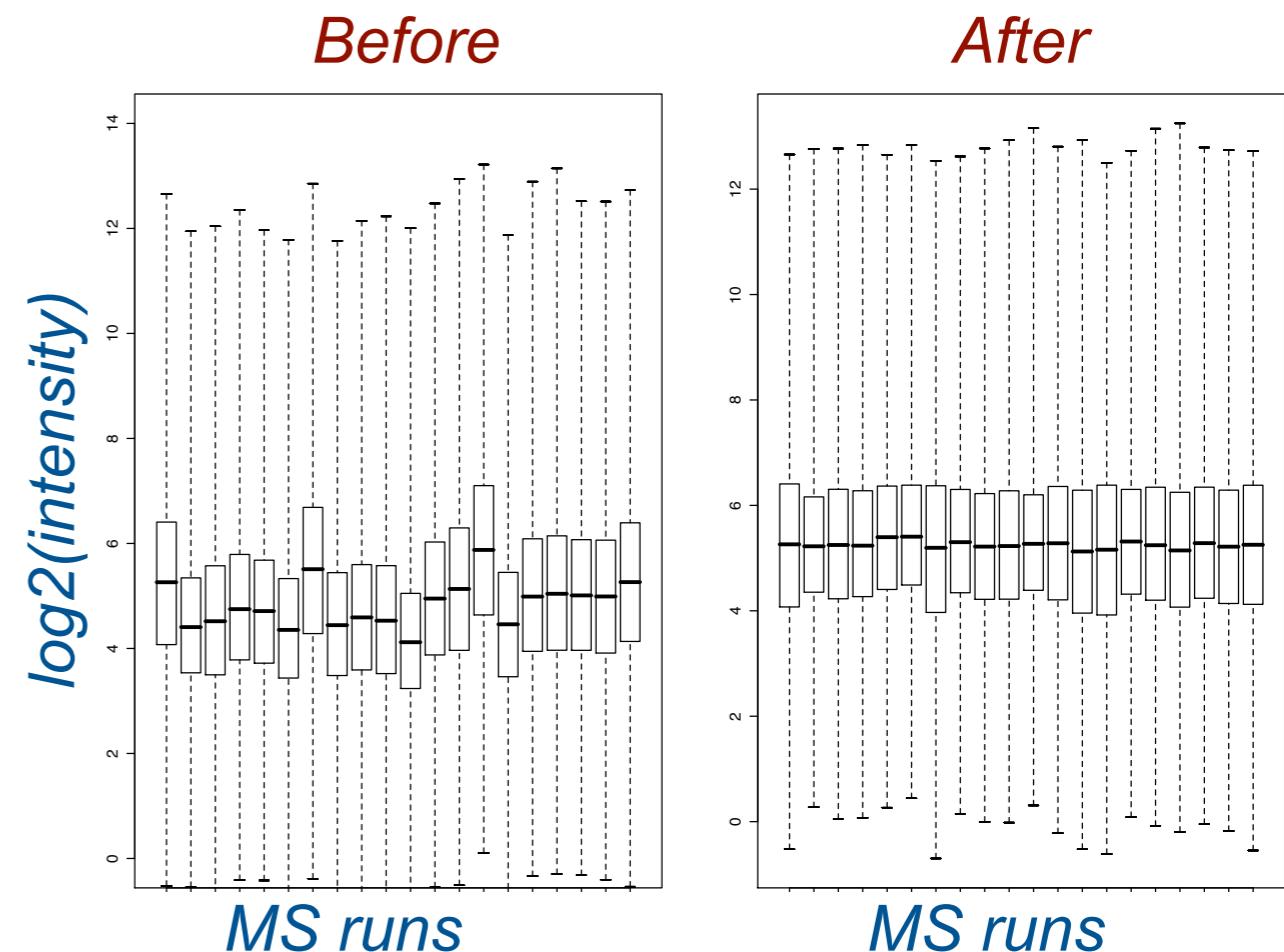


Reference Endogenous MS runs

STANDARD-BASED NORMALIZATION

Assumes constant abundance of features from a standard

- Algorithm
 - Subtract median[log(feature int)] of the standard
 - Add back the median of the medians
- Comments
 - + Standard does not have biological variation
 - + Independent of type/number of features
 - Only accounts for deviations that occur after adding the standard
 - Standards can be noisy (unequal spiked abundance; overlapped peaks)
- Best practice
 - Use one standard for normalization, another for verification

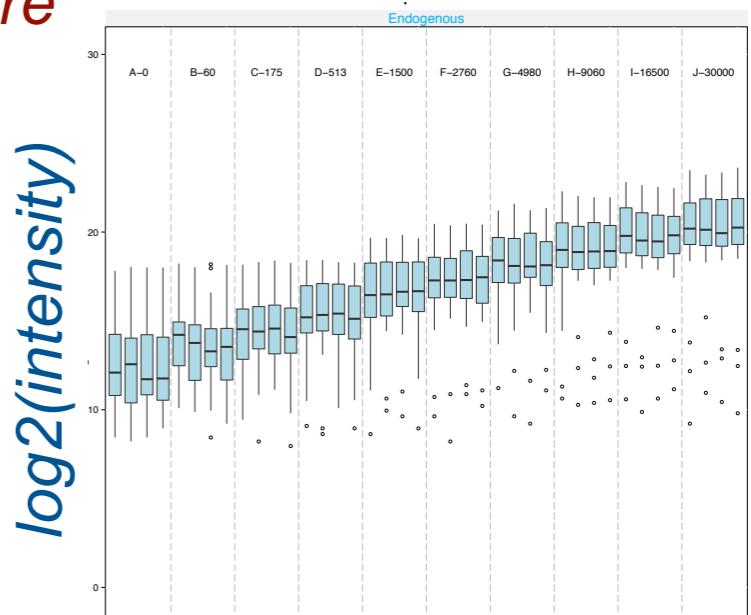


MEDIAN NORMALIZATION

Assumes constant abundance of median of $\log(\text{int})$ of endogenous features

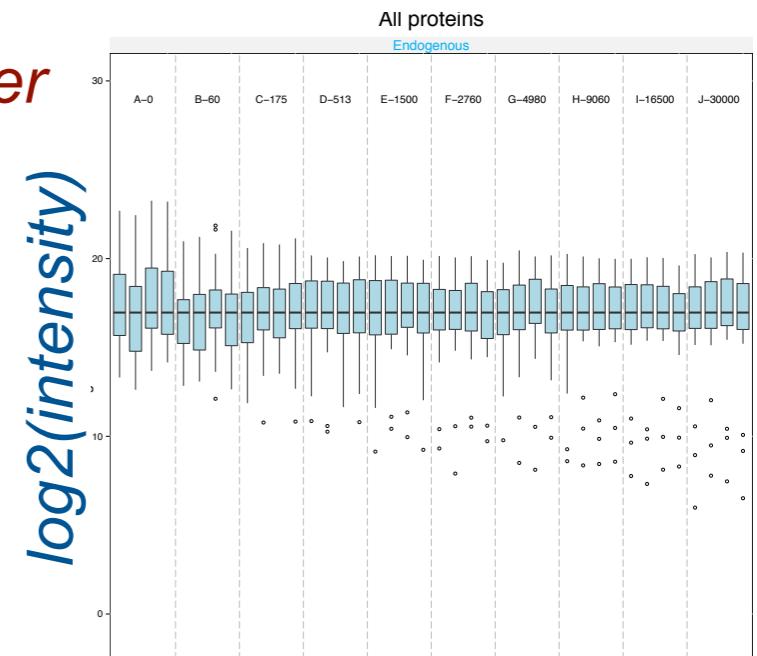
- Algorithm
 - Subtract median[$\log(\text{feature int})$] of all endogenous features
 - Add back the median of the medians
- Comments
 - + More stable than a single standard
 - + Accounts for all data processing steps
 - Assumes that the majority of endogenous proteins are not affected by the conditions
- Best practice
 - Use in discovery experiments with many features

Before



MS runs

After

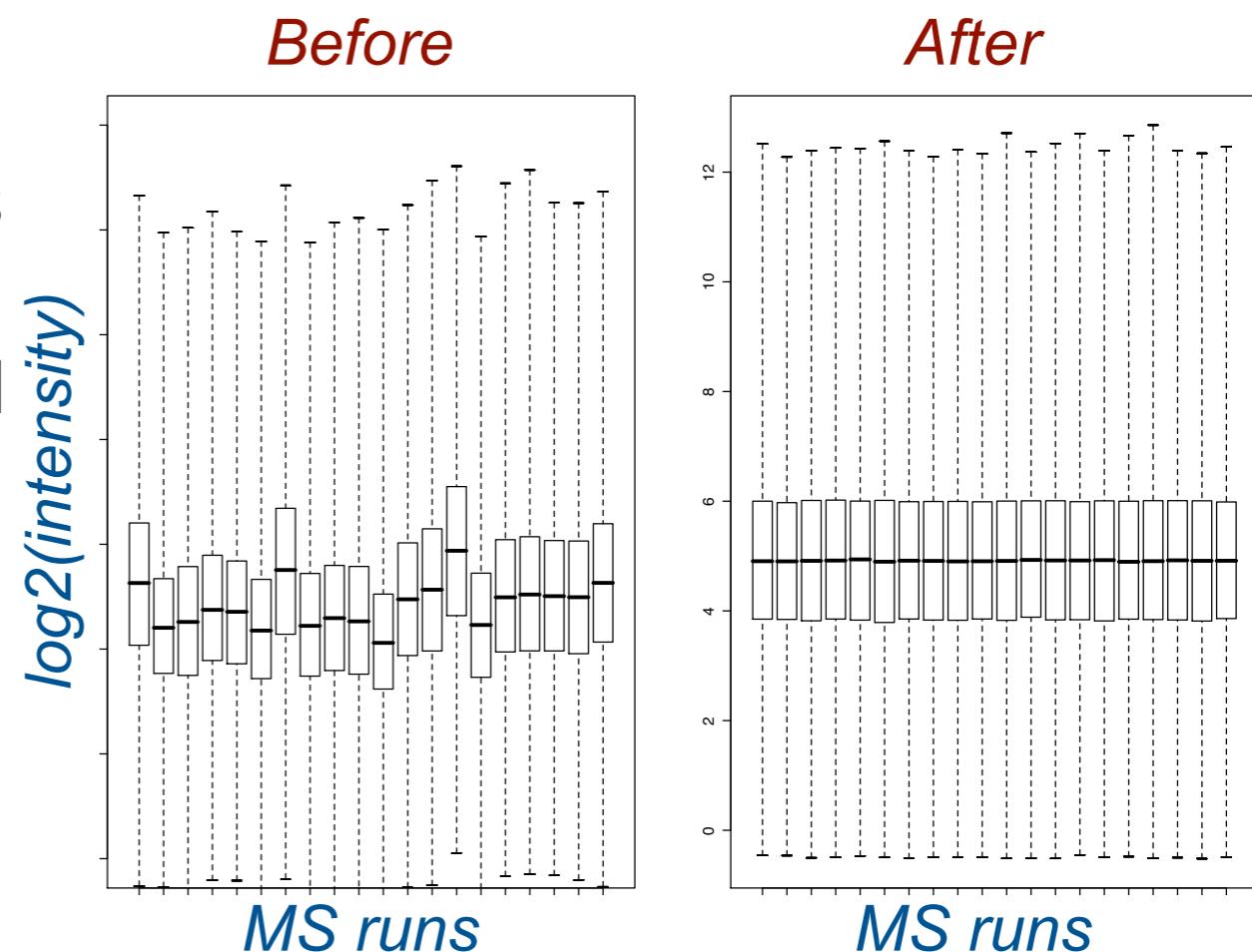


MS runs

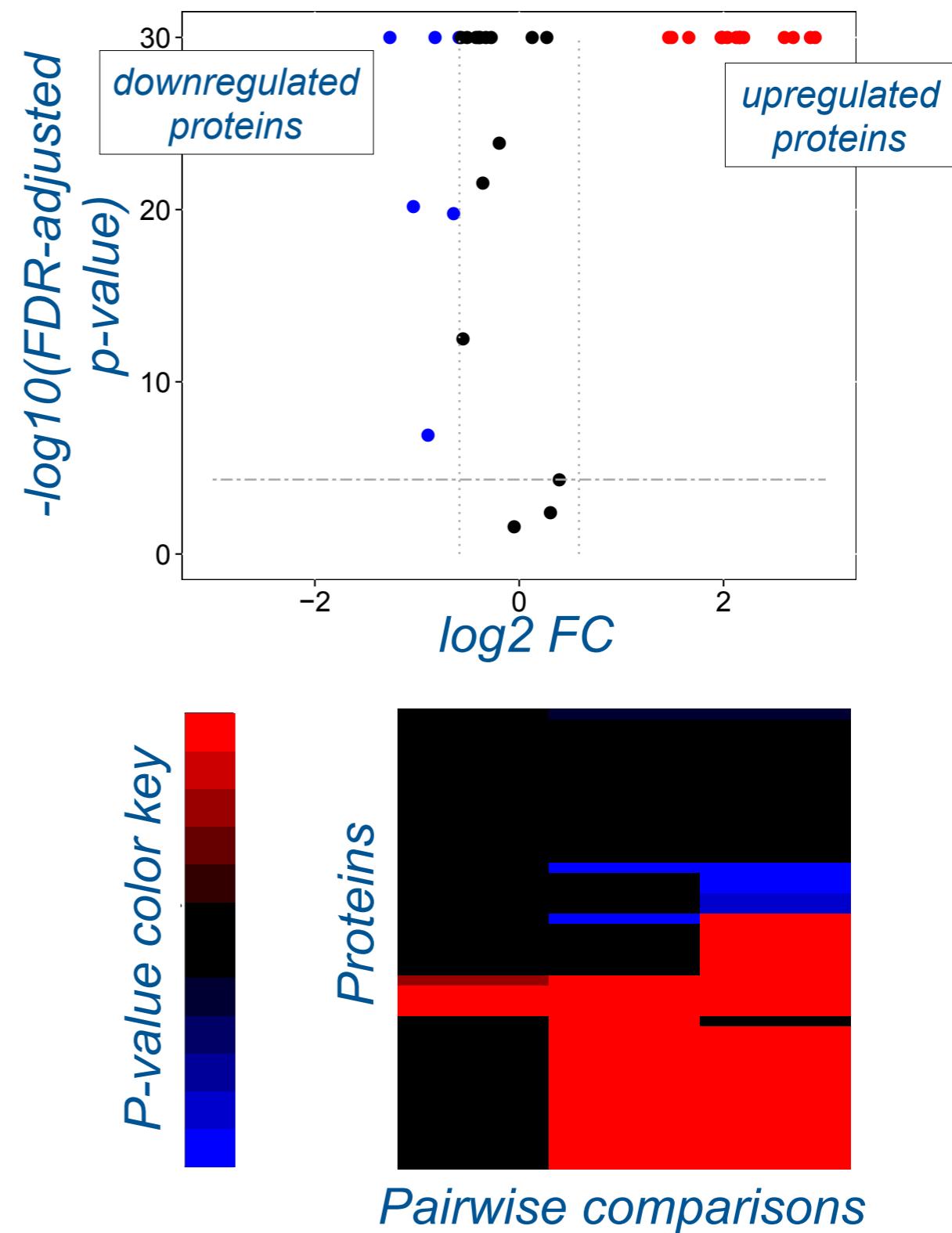
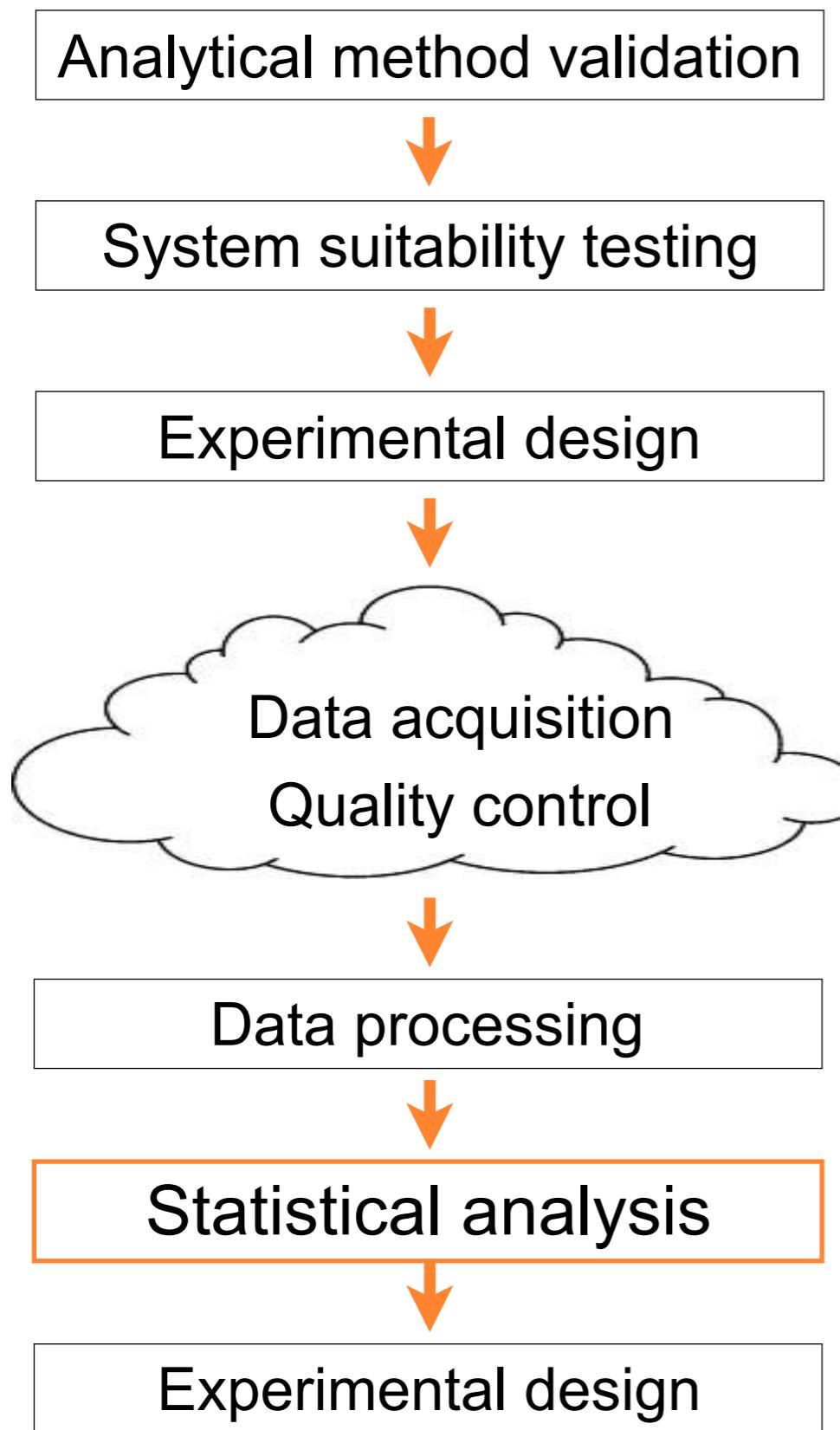
QUANTILE NORMALIZATION

Assumes constant abundance of all quantiles of $\log(\text{int})$ of endogenous features

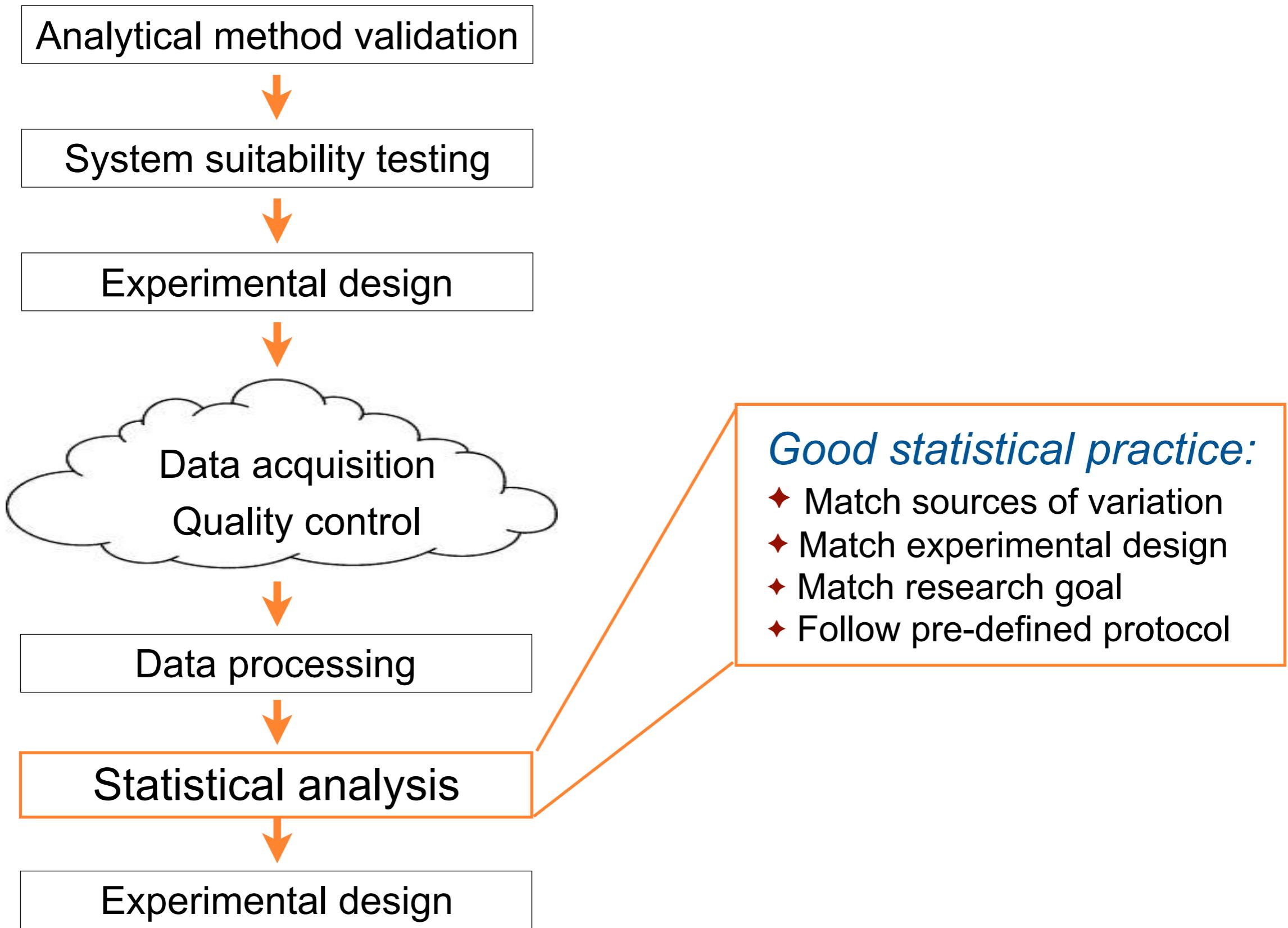
- Algorithm
 - Order $\log(\text{int})$ in each run
 - Calculate average of each quantile across runs
 - Substitute each $\log(\text{int})$ with the average
 - Re-arrange $\log(\text{int})$ in original order
- Comments
 - + More stable than a single standard
 - + Accounts for all data processing steps
 - Assumes that the majority of endogenous proteins are not affected by the conditions
 - Often very aggressive
- Best practice
 - Use to normalize multiple standards



MS EXPERIMENT: STATISTICIAN'S VIEW



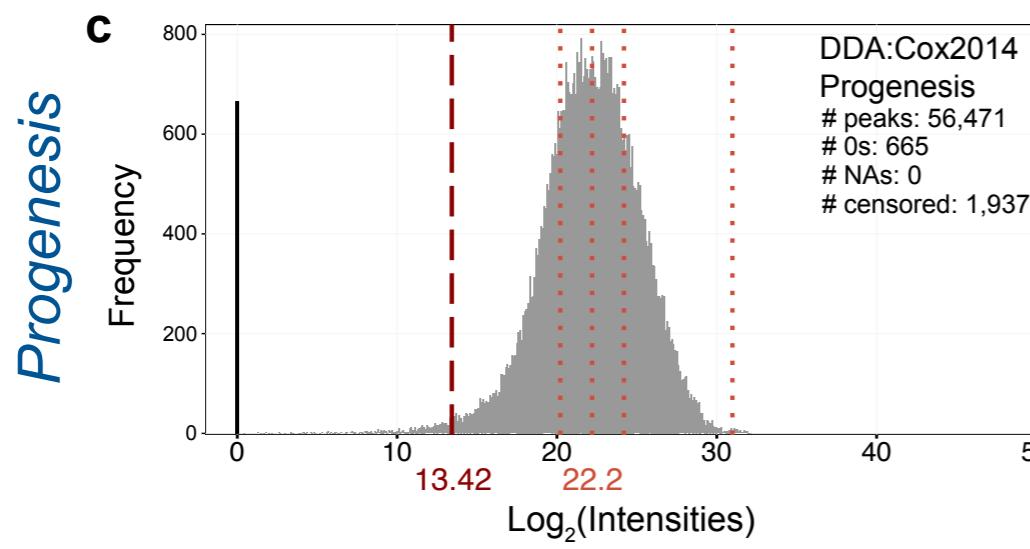
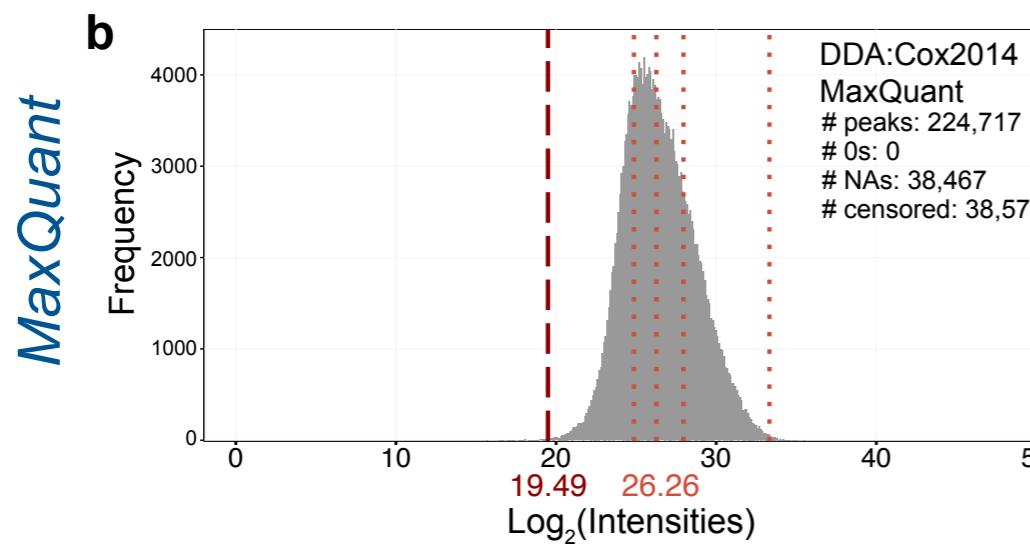
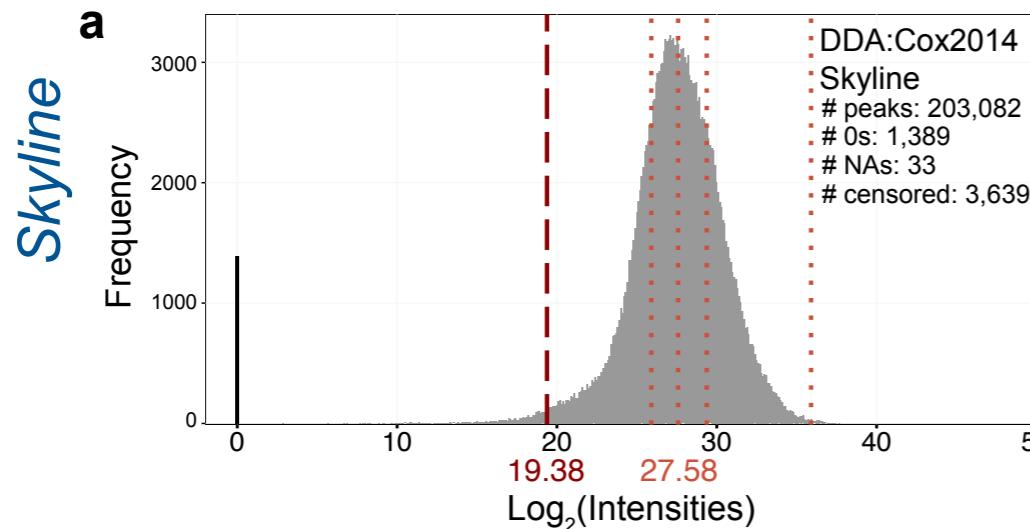
MS EXPERIMENT: STATISTICIAN'S VIEW



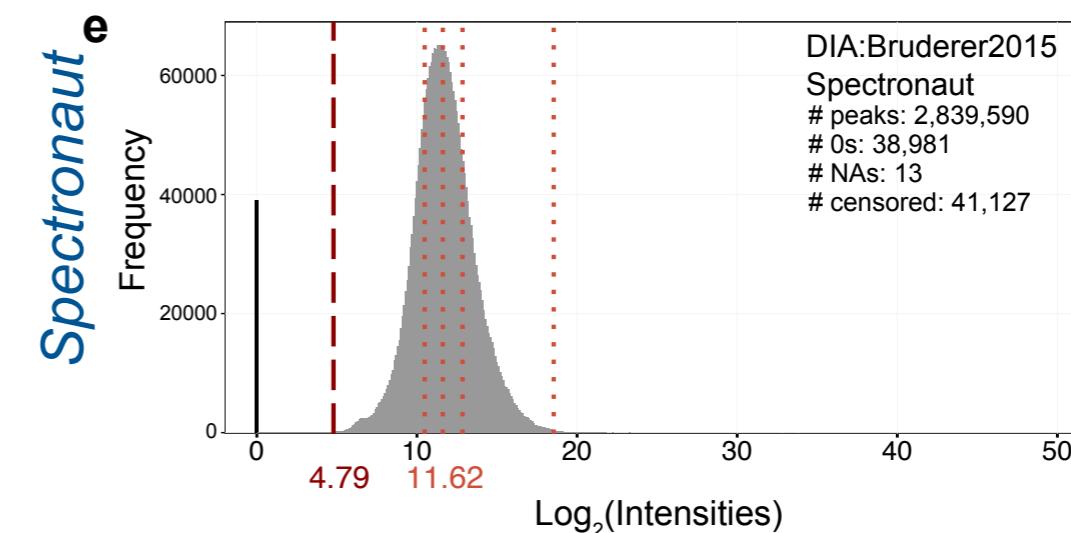
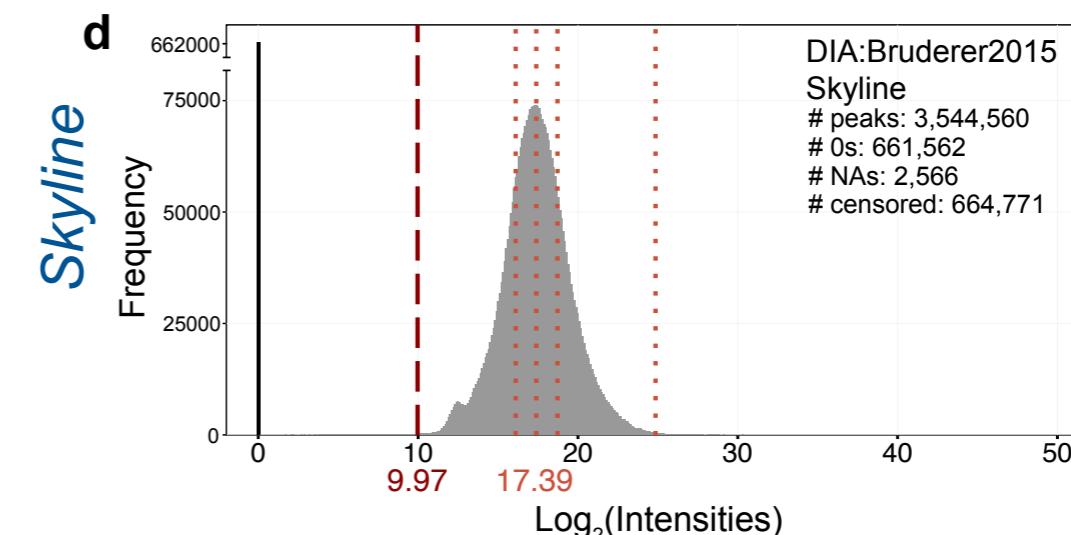
PROPERTIES OF PEAK INTENSITIES VARY BETWEEN DATA PROCESSING TOOLS

35

DDA: Cox 2014



DIA: Bruderer 2015



— — —	Estimated censoring threshold
... . .	Quantiles of log ₂ (intensity)
— — —	Frequency of peaks with intensity reported as between 0 and 1

SCHEMATIC DATA REPRESENTATION

Repeat for every protein

Conditions, subjects and runs: biological aspects of the experiment

Condition ₁												...	Condition _I												
Subject ₁			Subject ₂			...	Subject _J			...	Subject _{(I-1)J+1}			Subject _{(I-1)J+2}			...	Subject _{IJ}			...	Run _{IJK-2}	Run _{IJK-1}	Run _{IJK}	
	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}	...	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	...	y	y	y
Feature ₂	y	y	y	y	y	NA	...	y	y	y	...	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

Spectral features:

technological aspects of the experiment

Missing values

Log(feature intensities)

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

$$y_{ijkl} = \mu + \text{Condition}_i + \text{Subject(Condition)}_{j(i)} + \text{Run}_{ijk} + \text{Feature}_l + \text{Run} \times \text{Feature}_{ijkl}$$

Whole-plot
biological variation Whole-plot
technical variation Subplot
error

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

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

$$\text{Run}_{ijk} = \psi_{ijk} \stackrel{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \sigma_{\psi}^2)$$

$$\text{Run} \times \text{Feature}_{ijkl} = \epsilon_{ijkl} \stackrel{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \sigma_{\epsilon}^2)$$

INTERPRETING CENSORED VALUES

A

	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	y
Feature ₂	y	y	y	y	y	NA_{rand}	...	y	y	y	...	y	y	y	y	y	y	y	...	y	y	y
...
Feature _L	y	NA_{cen}	y	NA_{cen}	NA_{cen}	y	...	NA_{cen}	y	y	...	NA_{cen}	y	y	y	y	y	y	...	y	NA_{cen}	y

INTERPRETING CENSORED VALUES

A

	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	y
Feature ₂	y	y	y	y	y	NA_{rand}	...	y	y	y	...	y	y	y	y	y	y	...	y	y	y	
...
Feature _L	y	NA_{cen}	y	NA_{cen}	NA_{cen}	y	...	NA_{cen}	y	y	...	NA_{cen}	y	y	y	y	y	...	y	NA_{cen}	y	



Step 1 : Run-level subplot summarization

AFT model : Impute censored missing values by accelerated failure model

$$y_{ijkl} = \mu + Run_{ijk} + Feature_l + \epsilon_{ijkl}, \text{ where } \sum_{ijk} Run_{ijk} = 0, \sum_l Feature_l = 0, \epsilon_{ijkl} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$$

	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		...	y	y	y	...	y	y	y	y	y	y	...	y	y	y
...
Feature _L	y	y_{imp}	y	y_{imp}	y_{imp}	y	...	y_{imp}	y	y	...	y_{imp}	y	y	y	y	y	...	y	y_{imp}	y

INTERPRETING CENSORED VALUES

A

	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	y
Feature ₂	y	y	y	y	y	NA_{rand}	...	y	y	y	...	y	y	y	y	y	y	...	y	y	y	
...
Feature _L	y	NA_{cen}	y	NA_{cen}	NA_{cen}	y	...	NA_{cen}	y	y	...	NA_{cen}	y	y	y	y	y	...	y	NA_{cen}	y	



B Step 1 : Run-level subplot summarization

AFT model : Impute censored missing values by accelerated failure model

$$y_{ijkl} = \mu + Run_{ijk} + Feature_l + \epsilon_{ijkl}, \text{ where } \sum_{ijk} Run_{ijk} = 0, \sum_l Feature_l = 0, \epsilon_{ijkl} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$$

	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		...	y	y	y	...	y	y	y	y	y	y	...	y	y	y
...
Feature _L	y	y_{imp}	y	y_{imp}	y_{imp}	y	...	y_{imp}	y	y	...	y_{imp}	y	y	y	y	y	...	y	y_{imp}	y



TMP : Parameter estimation by robust method

$$y_{ijkl} = \mu + Run_{ijk} + Feature_l + \epsilon_{ijkl}, \text{ where}$$

$$\text{median}_{ijk}(Run_{ijk}) = 0, \text{ median}_l(Feature_l) = 0, \text{ and } \text{median}_{ijk}(\epsilon_{ijkl}) = \text{median}_l(\epsilon_{ijkl}) = 0$$

	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}

TUKEY MEDIAN POLISH

Summarization over all features in a run

		<i>Run</i>			
		1	2	...	12
<i>log(Feature int)</i>	1	x_{11}	x_{12}	...	$x_{1,12}$
	2	x_{21}	x_{22}	...	$x_{2,12}$
		
n	x_{n1}	x_{n2}	...	$x_{n,12}$	

Tukey median polish
Represent features and runs in a sub-plot as 2-way Analysis of Variance

$$x_{ij} = \text{feature}_i + \text{run}_j + \text{error}_{ij}$$

Addition: censored data
Impute missing values by assuming that they have intensities below detection threshold

- Robust parameter estimation

- ◆ subtract column median from each value
- ◆ subtract row median from each value
- ◆ continue until no change
- ◆ obtain fitted values
 - subtract the resulting residuals from the original values
- ◆ obtain array-based summary
 - average fitted values over the column

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	NA_{rand}	...	y	y	y	...	y	y	y	y	y	y	...	y	y	y
...
Feature _L	y	NA_{cen}	y	NA_{cen}	NA_{cen}	y	...	NA_{cen}	y	y	...	NA_{cen}	y	y	y	y	y	...	y	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

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

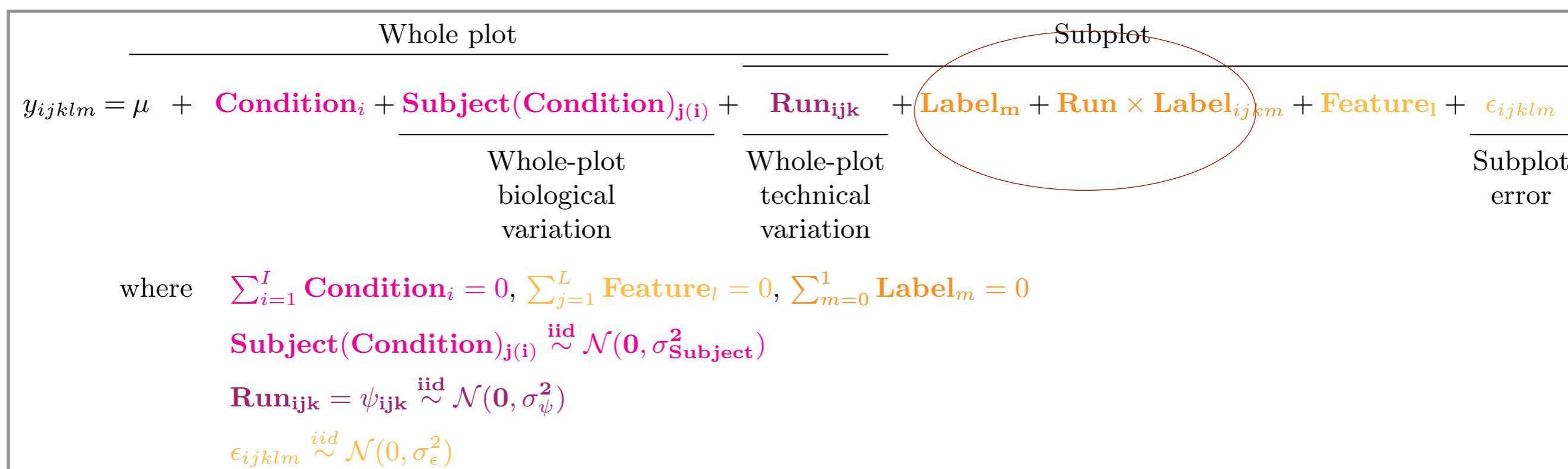
$$\sum_i Condition_i = 0, \quad Subject(Condition)_{j(i)} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_{Subject}^2), \quad \psi_{ijk} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_\psi^2)$$

	Condition ₁						...	Condition _I						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}	

EXTENSION: LABELED REFERENCE PEPTIDES

43

Subplot		Whole plot																
		Condition _i						Condition _j										
		Subject ₁			Subject ₂			Subject _J			Subject _{(I-1)J+1}			Subject _{(I-1)J+2}			Subject _{IJ}	
Run	Run	Run	Run	Run	Run	Run	...	Run	Run	Run	Run	Run	Run	Run	Run	Run	Run	Run
Endogenous	Feature ₁	y	y	y	y	y	y	...	y	y	y	...	y	y	y			
	Feature ₂	y	y	y	y	y	y	y	y	...	y	y	y			
			
	Feature _L	y		y			y	...		y	y			y		y		
Reference	Feature ₁	y	y	y	y	y	y	...	y	y	y	...	y	y	y			
	Feature ₂	y	y	y	y	y	y	...	y	y	y	...	y	y	y			
			
	Feature _L	y	y	y	y	y	y	...	y	y	y	...	y	y	y			



EXTENSION: TMT LABELING

Designs combine biological and technical replication
Run=block

*“Biol”.
replicates
in a run*

*Runs=MS
replicates*

*Mixtures=“biol”.
replicates*

	0.125		0.5		0.667		1		Norm						
	Run 1	127C	129N	Run 2	128N	129C	Run 3	127N	130C	Run 4	128C	130N	Run 5	126	131
Mixture 1	Run 2	127C	129N	Run 3	128N	129C	Run 4	127N	130C	Run 5	128C	130N	Run 6	126	131
	Run 3	127C	129N	Run 4	128N	129C	Run 5	127N	130C	Run 6	128C	130N	Run 7	126	131

	0.125		0.5		0.667		1		Norm						
	Run 1	127C	129N	Run 2	128N	129C	Run 3	127N	130C	Run 4	128C	130N	Run 5	126	131
Mixture 1	Run 2	127C	129N	Run 3	128N	129C	Run 4	127N	130C	Run 5	128C	130N	Run 6	126	131
	Run 3	127C	129N	Run 4	128N	129C	Run 5	127N	130C	Run 6	128C	130N	Run 7	126	131
	Run 1	128C	130N	Run 2	127N	130C	Run 3	128N	129C	Run 4	127C	129N	Run 5	126	131
Mixture 2	Run 2	128C	130N	Run 3	127N	130C	Run 4	128N	129C	Run 5	127C	129N	Run 6	126	131
	Run 3	128C	130N	Run 4	127N	130C	Run 5	128N	129C	Run 6	127C	129N	Run 7	126	131
	Run 1	127N	129C	Run 2	128C	129N	Run 3	127C	130N	Run 4	128N	130C	Run 5	126	131
Mixture 3	Run 2	127N	129C	Run 3	128C	129N	Run 4	127C	130N	Run 5	128N	130C	Run 6	126	131
	Run 3	127N	129C	Run 4	128C	129N	Run 5	127C	130N	Run 6	128N	130C	Run 7	126	131
	Run 1	128N	130C	Run 2	127C	130N	Run 3	128C	129N	Run 4	127N	129C	Run 5	126	131
Mixture 4	Run 2	128N	130C	Run 3	127C	130N	Run 4	128C	129N	Run 5	127N	129C	Run 6	126	131
	Run 3	128N	130C	Run 4	127C	130N	Run 5	128C	129N	Run 6	127N	129C	Run 7	126	131
	Run 1	127C	129N	Run 2	128N	129C	Run 3	127N	130C	Run 4	128C	130N	Run 5	126	131
Mixture 5	Run 2	127C	129N	Run 3	128N	129C	Run 4	127N	130C	Run 5	128C	130N	Run 6	126	131
	Run 3	127C	129N	Run 4	128N	129C	Run 5	127N	130C	Run 6	128C	130N	Run 7	126	131

EXTENSION:TMT LABELING

Label
-free

	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			

Run = block (restriction on randomization)

TMT

TMT	Run ₁										...	Run _k												
	Condition ₁			Condition ₁			...		Condition ₁				Condition _{(I-1)J+1}			Condition _{(I-1)J+2}			...		Condition _I			
	Sub	Sub	Sub	Sub	Sub	Sub	...	Sub	Sub	Sub		Sub	Sub	Sub	Sub	Sub	Sub	...	Sub	Sub	Sub			
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			

Potentially different LC-MS features across runs



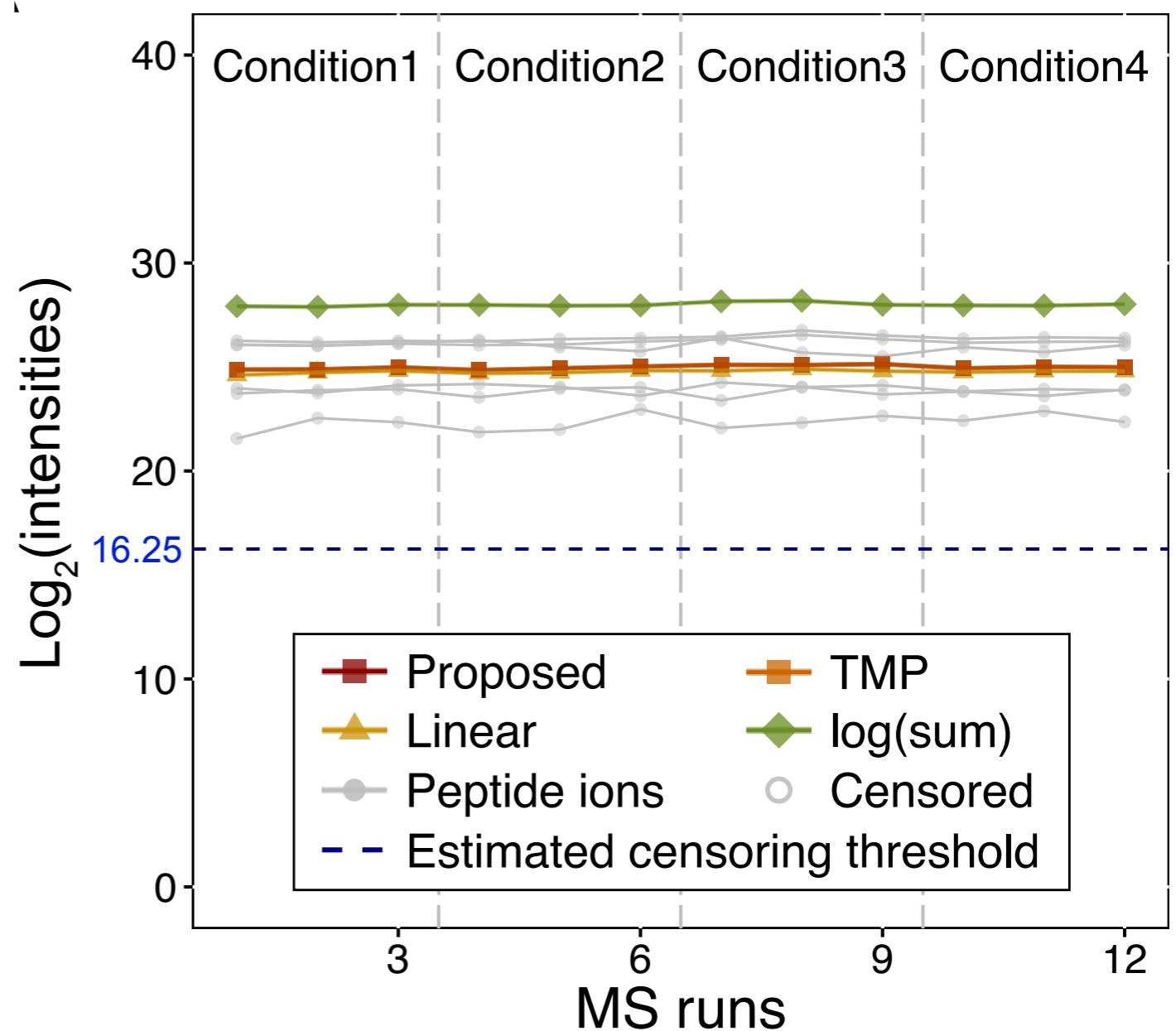
Separate summarization and normalization

OUTLINE

- Motivating example
 - ABRF iPRG study
- MSstats
 - Statistical relative quantification of proteins and peptides
 - Methods evaluation
- Extensions to MSstats
 - Assay characterization
 - System suitability and quality control

ROBUSTNESS TO OUTLIERS

Methods perform similarly with high quality data



Condition2-Condition1 : True fold change=1

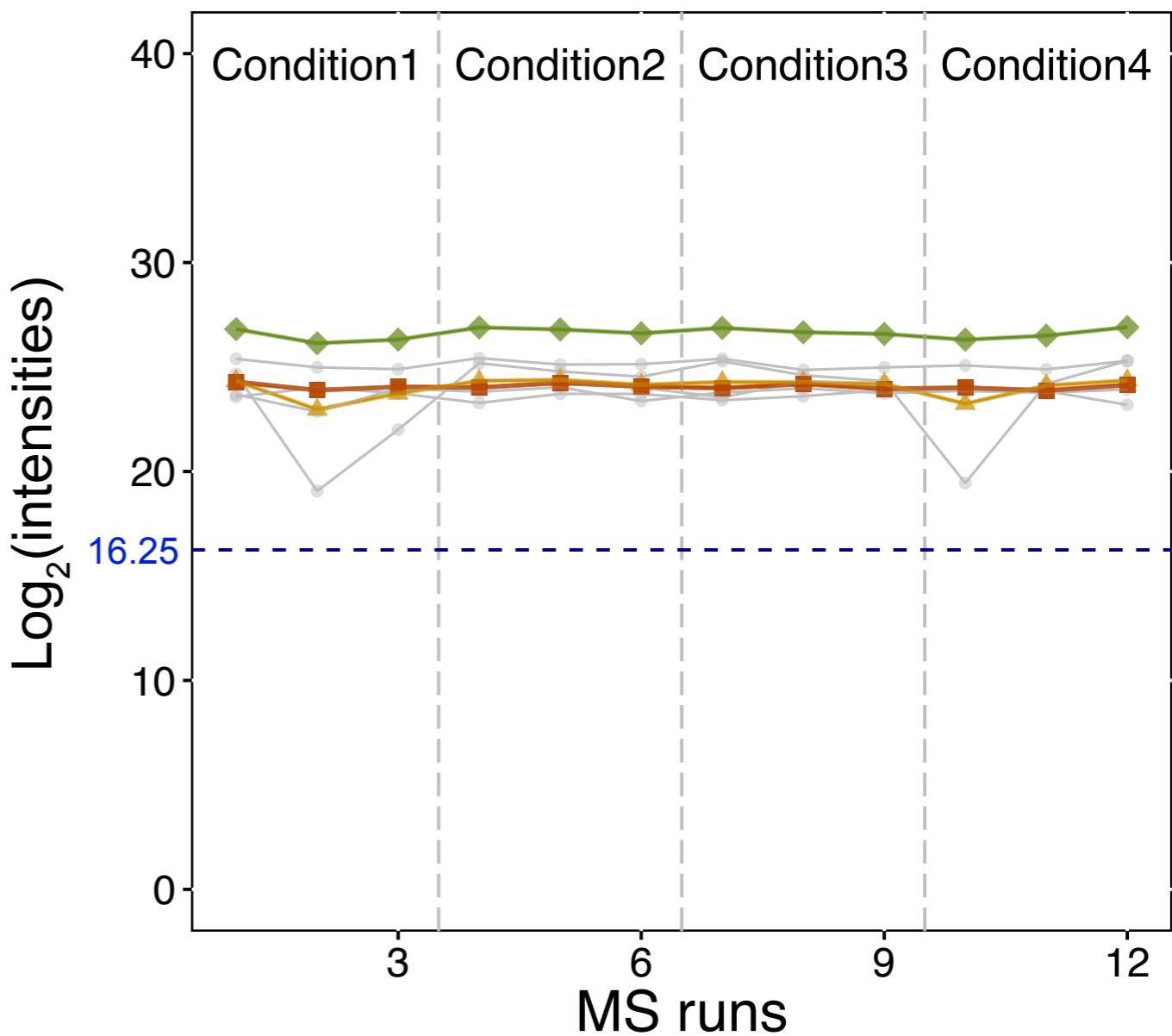
EstimatedFC Adj.pvalue

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

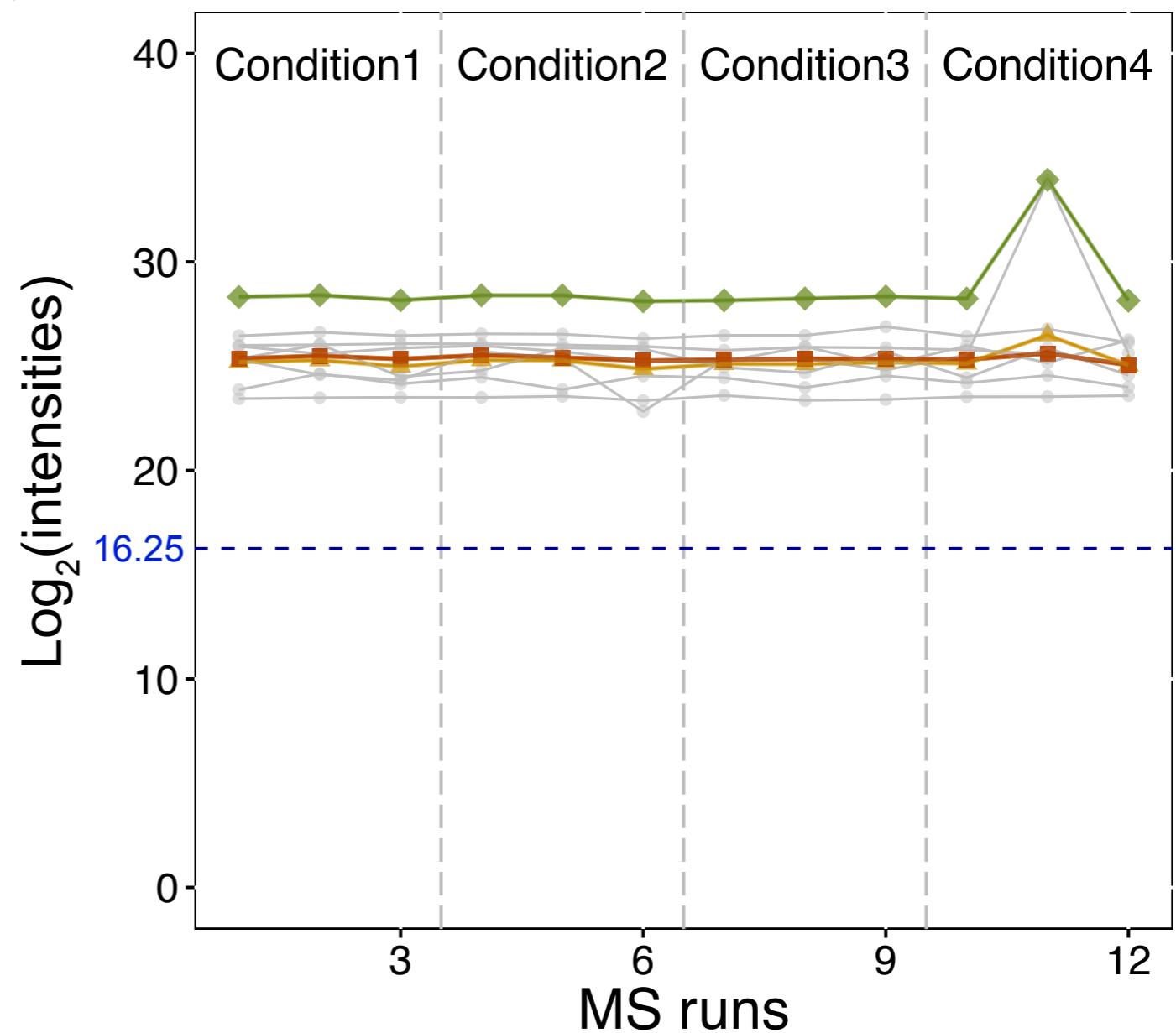
EstimatedFC Adj.pvalue

Peptide ions	
Proposed	
TMP	

	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)*



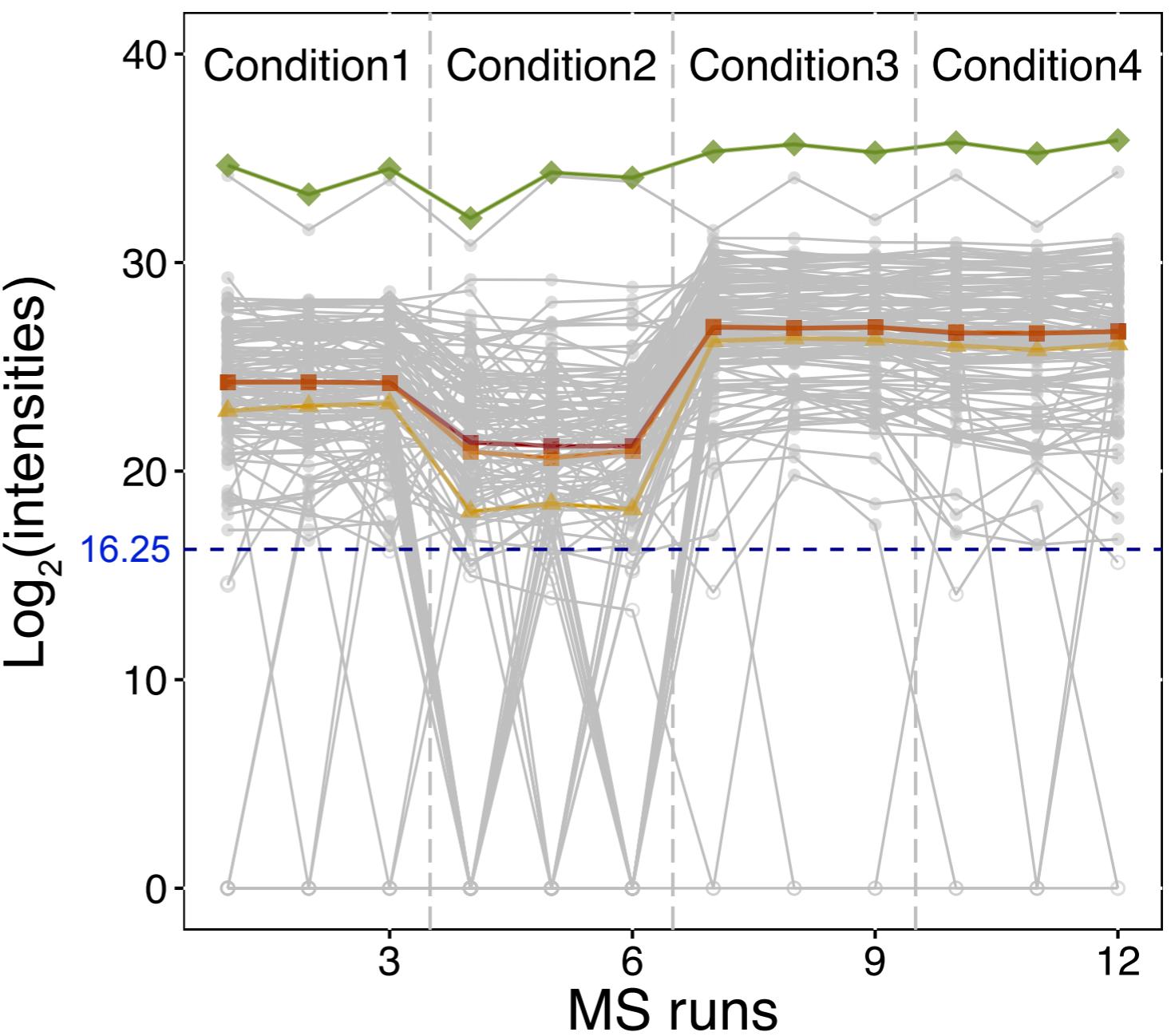
Condition4-Condition1 : True fold change=1
EstimatedFC Adj.pvalue

Peptide ions	
Proposed	
TMP	

	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

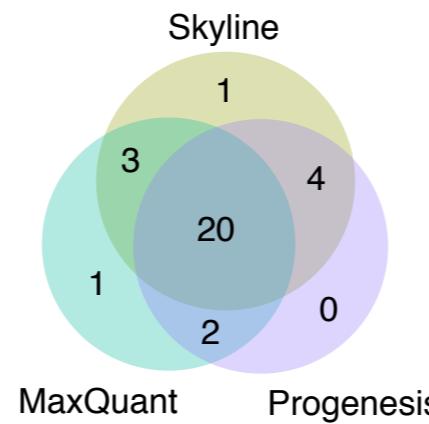
Peptide ions	
Proposed	
TMP	

Proposed	8.015	< 0.001
TMP	10.605	< 0.001
Linear model	29.106	< 0.001
log(sum)	1.552	0.999

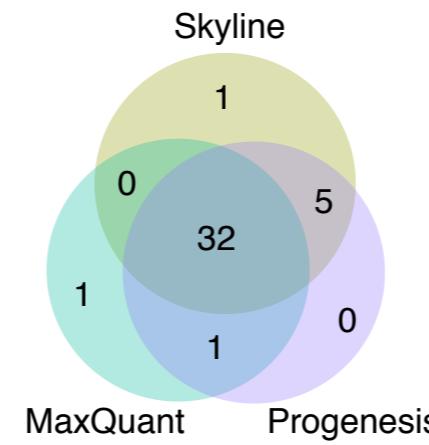
BETTER STATISTICAL METHODS ENHANCE REPRODUCIBLE RESEARCH

Better agreement
in #differentially
abundant proteins
between tools

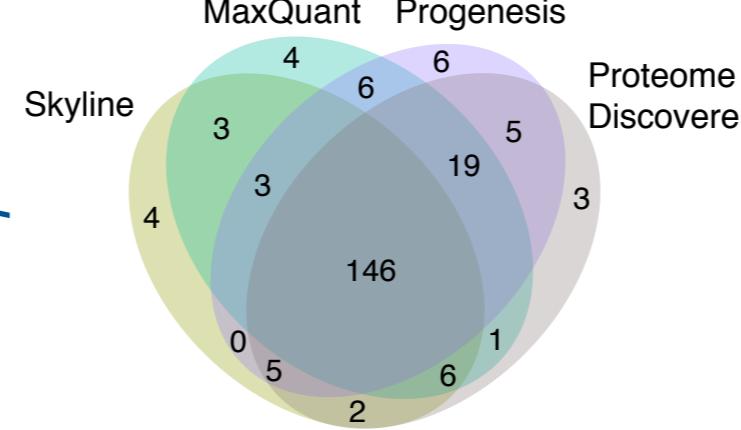
DDA: iPRG2015



DDA: Cox 2014

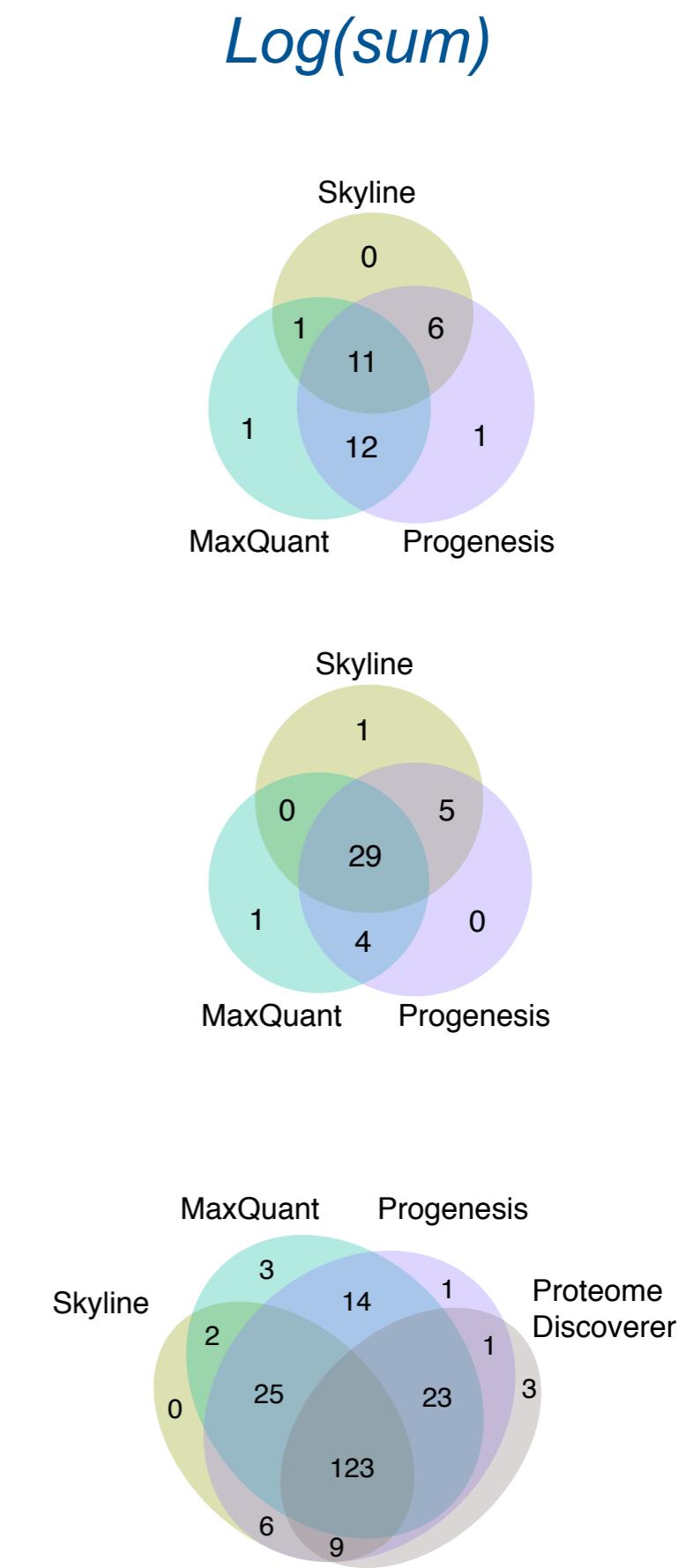
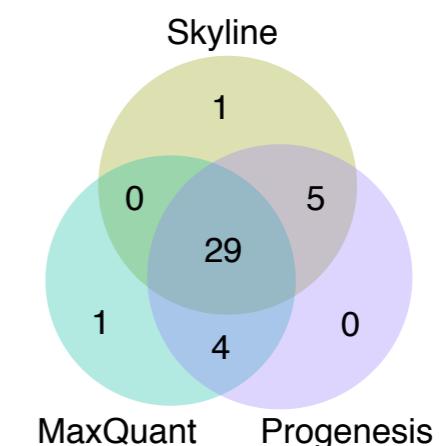
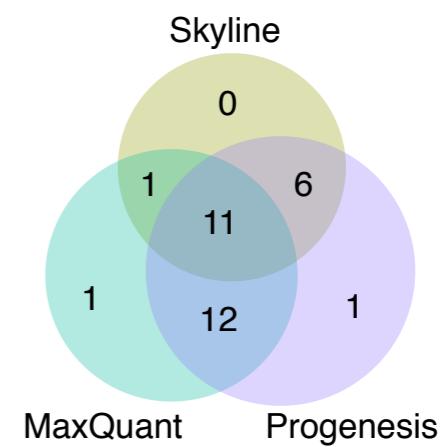


DDA: Spike-in



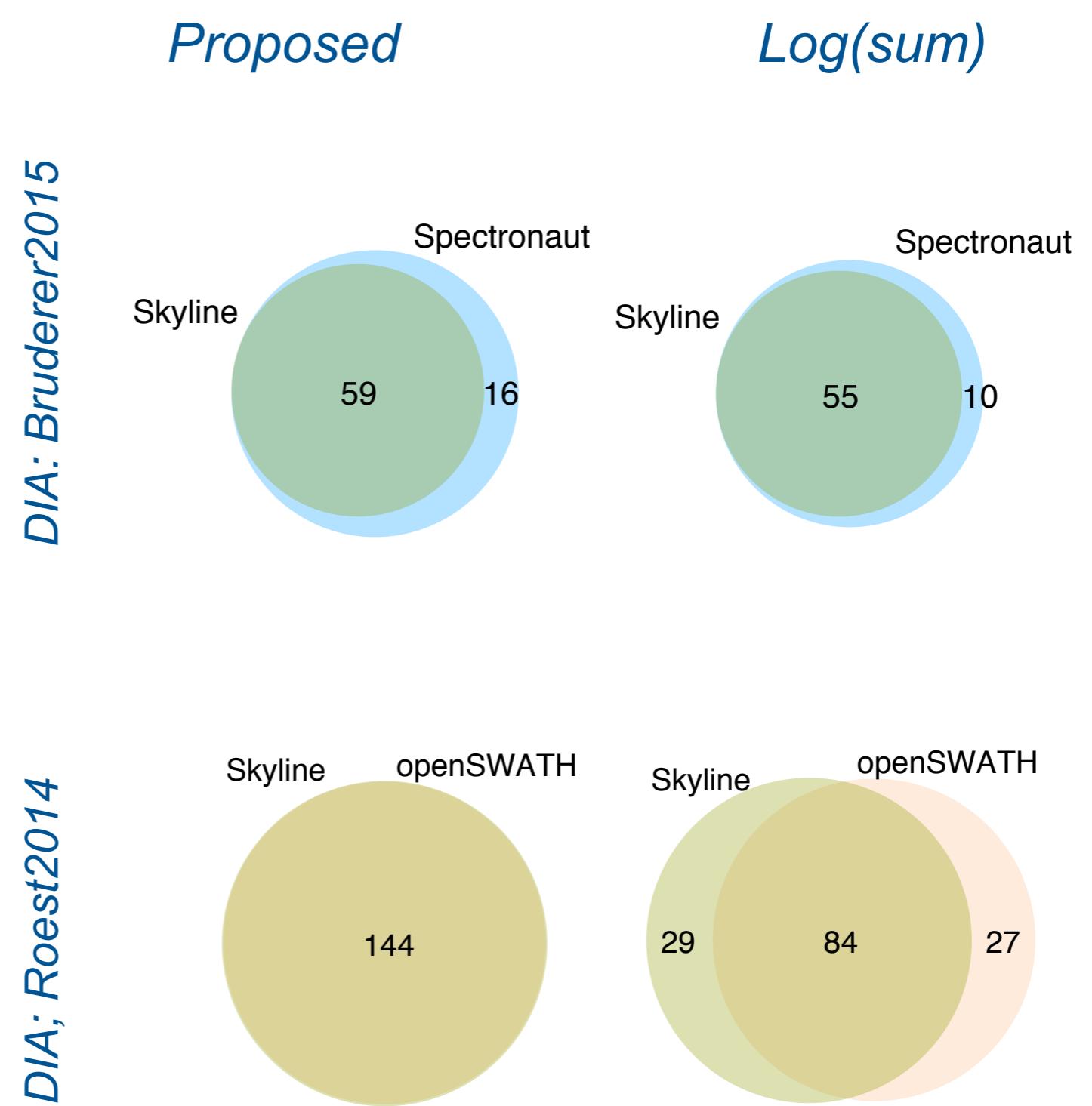
Proposed

Log(sum)

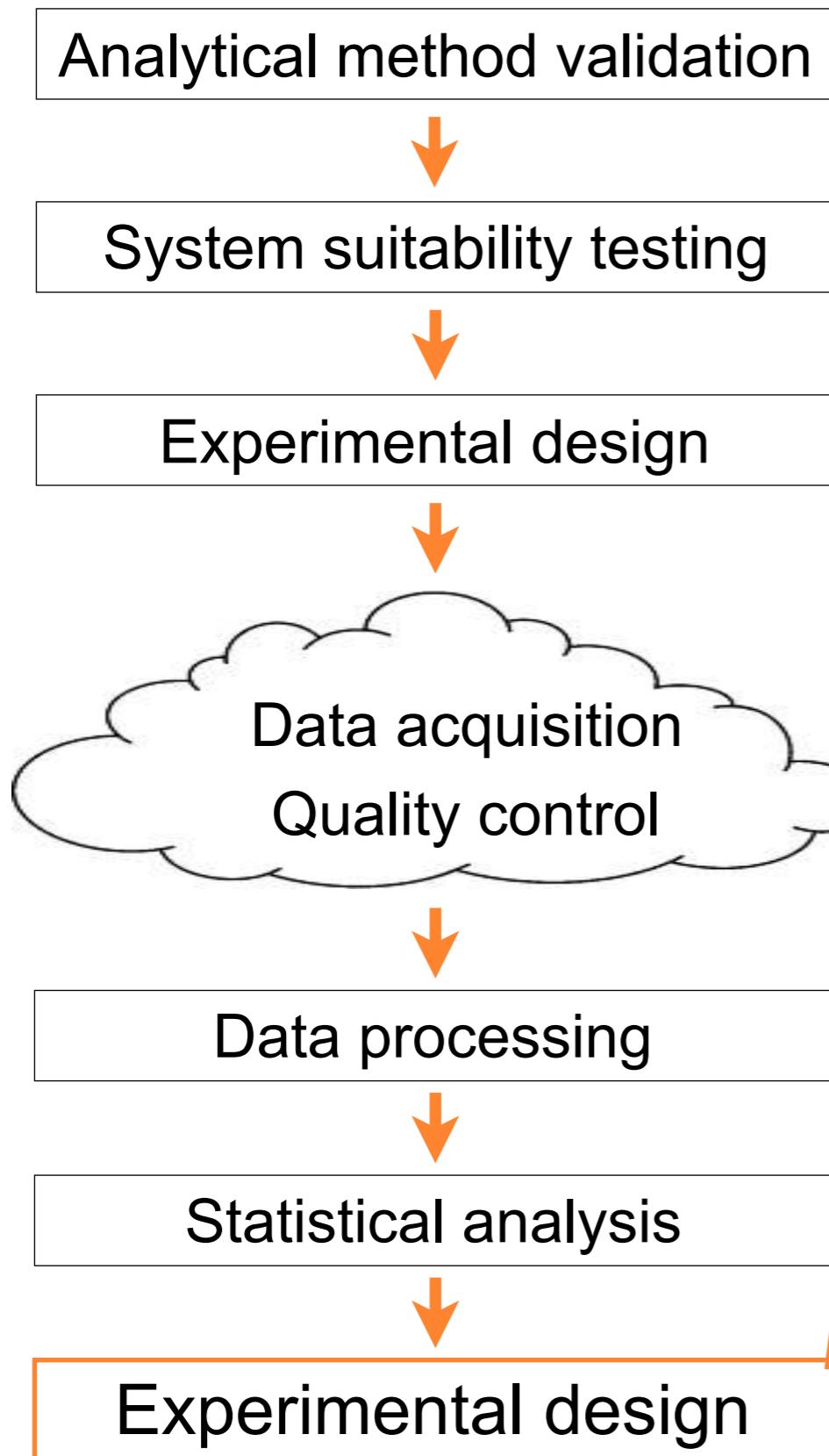


BETTER STATISTICAL METHODS ENHANCE REPRODUCIBLE RESEARCH

Better agreement
in #differentially
abundant proteins
between tools

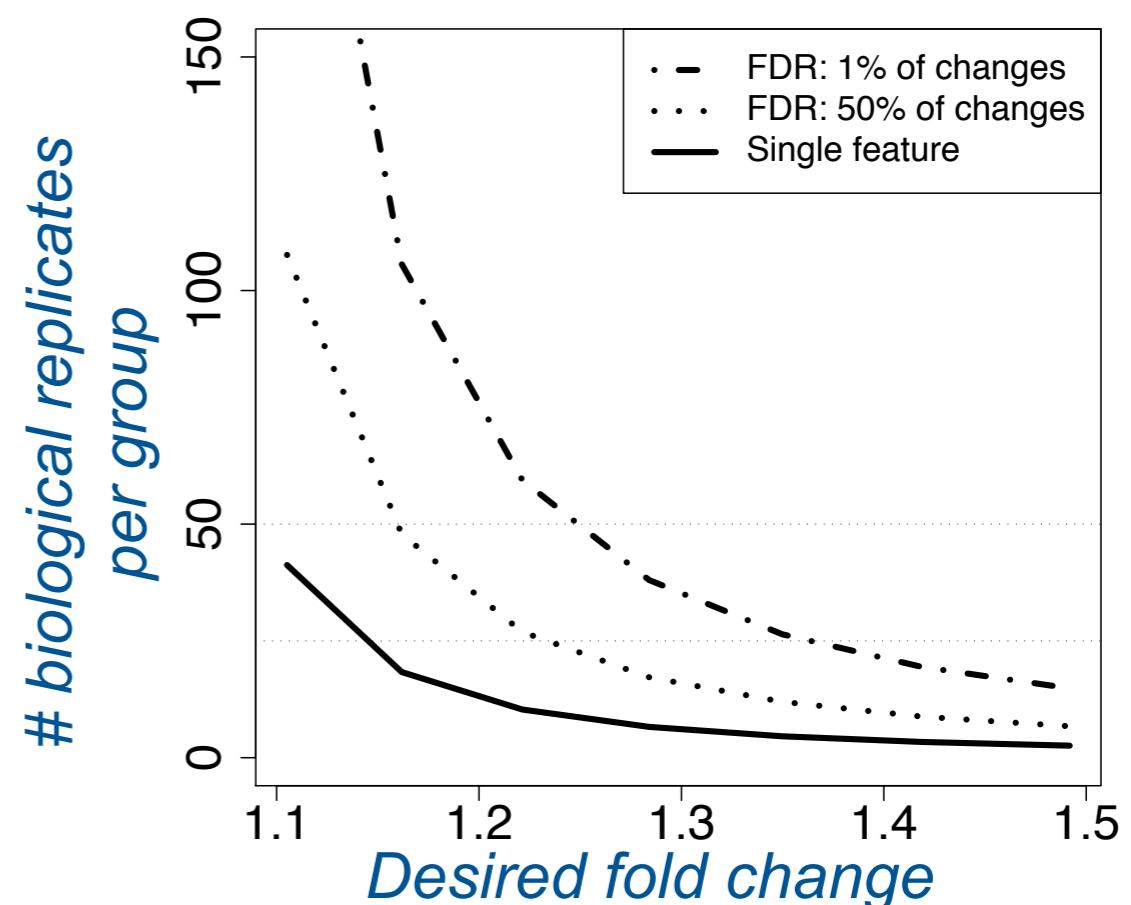


MS EXPERIMENT: STATISTICIAN'S VIEW



Use the dataset to improve:

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



MSSTATS IS OPEN-SOURCE, R-BASED AND PUBLICLY AVAILABLE

<http://msstats.org/>



MSstats

Statistical Tool For Quantitative Mass Spectrometry-Based Proteomics

[HOME](#) [MSSTATS](#) [TMT](#) [CLASSIFICATION](#) [LOB/LOD](#) [QC](#) [DATASETS](#) [COURSE](#) [CONTACT](#)

Overview

MSstats is an open-source R package for statistical relative quantification of proteins and peptides in global, targeted and data-independent proteomics. It provides workflows for

- detecting differentially abundant proteins for MS experiments with chromatography-based quantification, with complex designs.
- characterizing MS assays in terms of limit of blank and limit of detection (LOB/LOD),
- longitudinal monitoring of quality control and system suitability testing (SST).

For comments, please use this [Google group](#).

More information about research in the Olga Vitek lab.

News

- The upcoming NEU course ([Computation and Statistics for Mass Spectrometry](#)) on April 30 – May 11, 2018. The detailed information will follow shortly.
- MSstats v3.10.0 is available on Bioconductor 3.6.
- The most recent development version of the package MSstats is available in [github](#).

 **MSstats**
Statistical Tool For Quantitative Mass Spectrometry-Based Proteomics

www.msstats.org

Skyline
external tool

M. Choi et al. *BMC Bioinformatics.*, 2014

Bioconductor

 **MSstats**

Version 3.8.4 [View All]

Uploaded Dec 10, 2017

[Support Board](#)

[Download MSstats](#)
Downloaded: 12701

Documentation
 KnownIssues-Skyline-MSstatsV2.1.6.pdf
 MSstats-SkylineExternalTool-InstallationAndUserGuide-v2.1.6.pdf
 MSstatsTutorial.zip

Tool Information
Organization: Vitek Lab, Northeastern University
Authors: Meena Choi, Tsung-Heng Tsai, Ching-Yun Chang, Dr. Timothy Clough, Dr. Olga Vitek
Languages: R(3.4.0), C#
More Information: http://www.msstats.org/

downloads top 20%

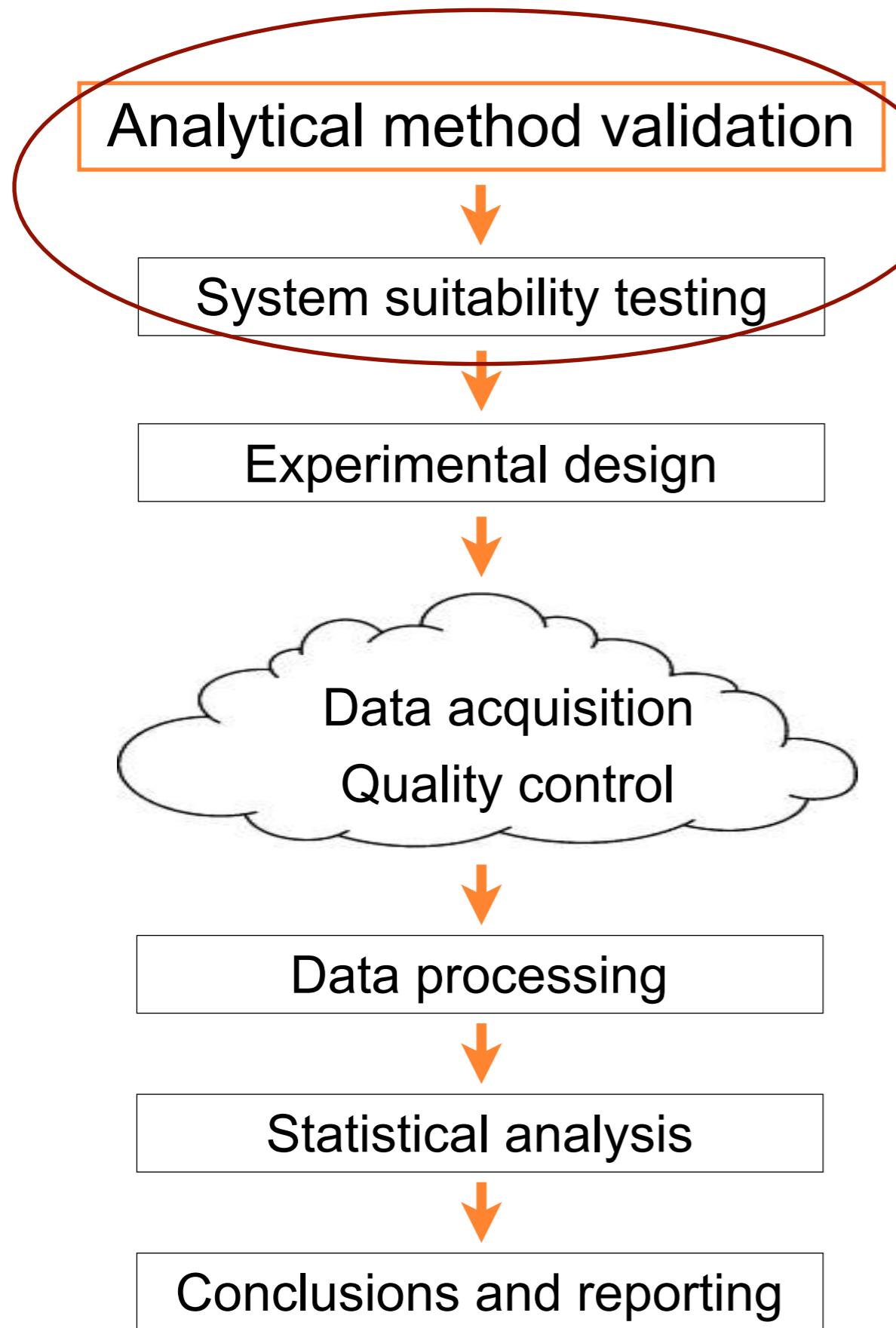
Month	Nb of distinct IPs	Nb of downloads
Jan/2017	172	277
Feb/2017	203	351
Mar/2017	283	614
Apr/2017	232	409
May/2017	317	544
Jun/2017	322	517
Jul/2017	259	379
Aug/2017	222	451
Sep/2017	186	300
Oct/2017	236	559
Nov/2017	300	596
Dec/2017	190	313
2017	2311	5310

[MSstats 2017 stats.tab](#)

OUTLINE

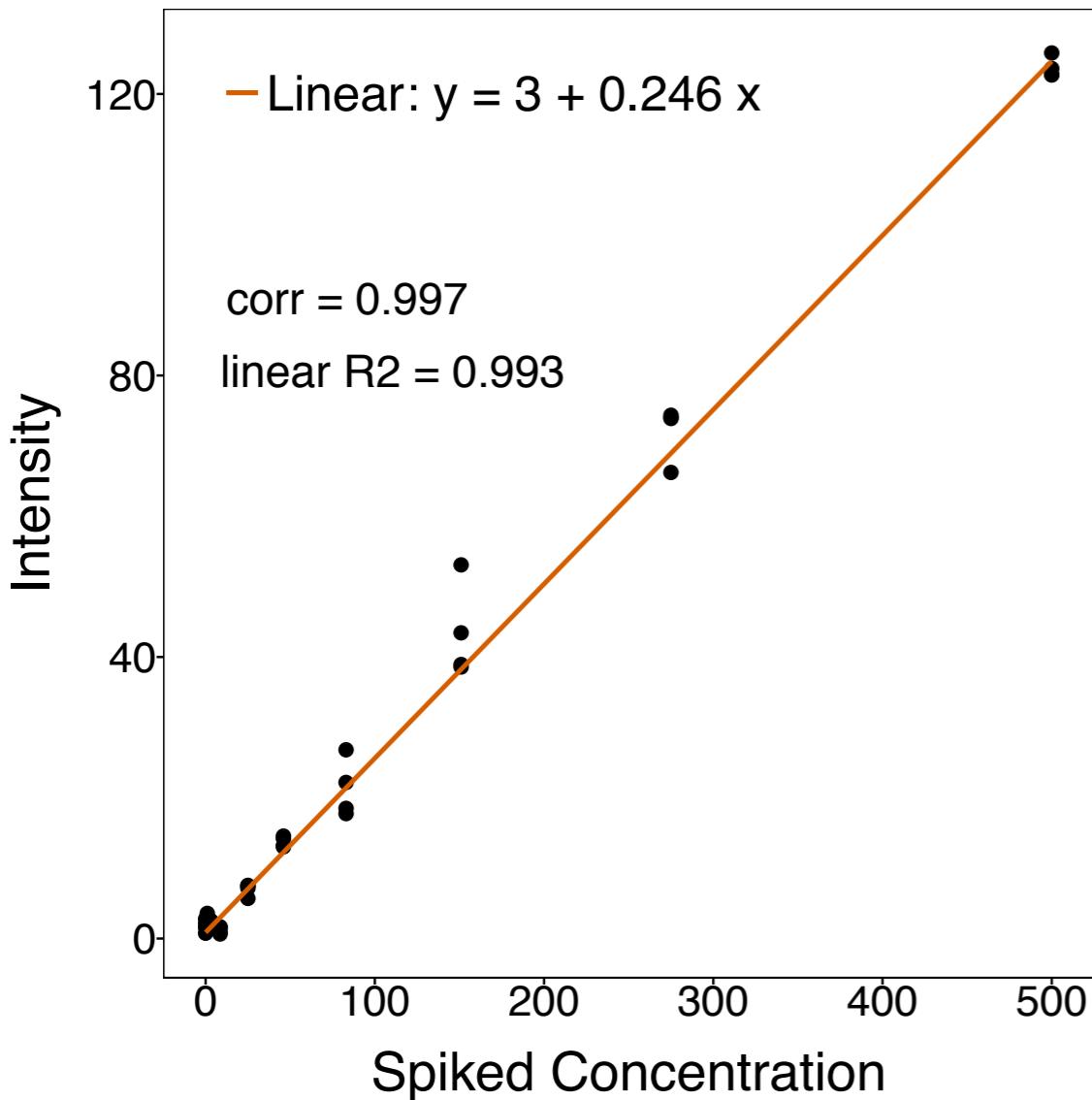
- Motivating example
 - ABRF iPRG study
- MSstats
 - Statistical relative quantification of proteins and peptides
 - Methods evaluation
- Extensions to MSstats
 - Assay characterization
 - System suitability and quality control

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

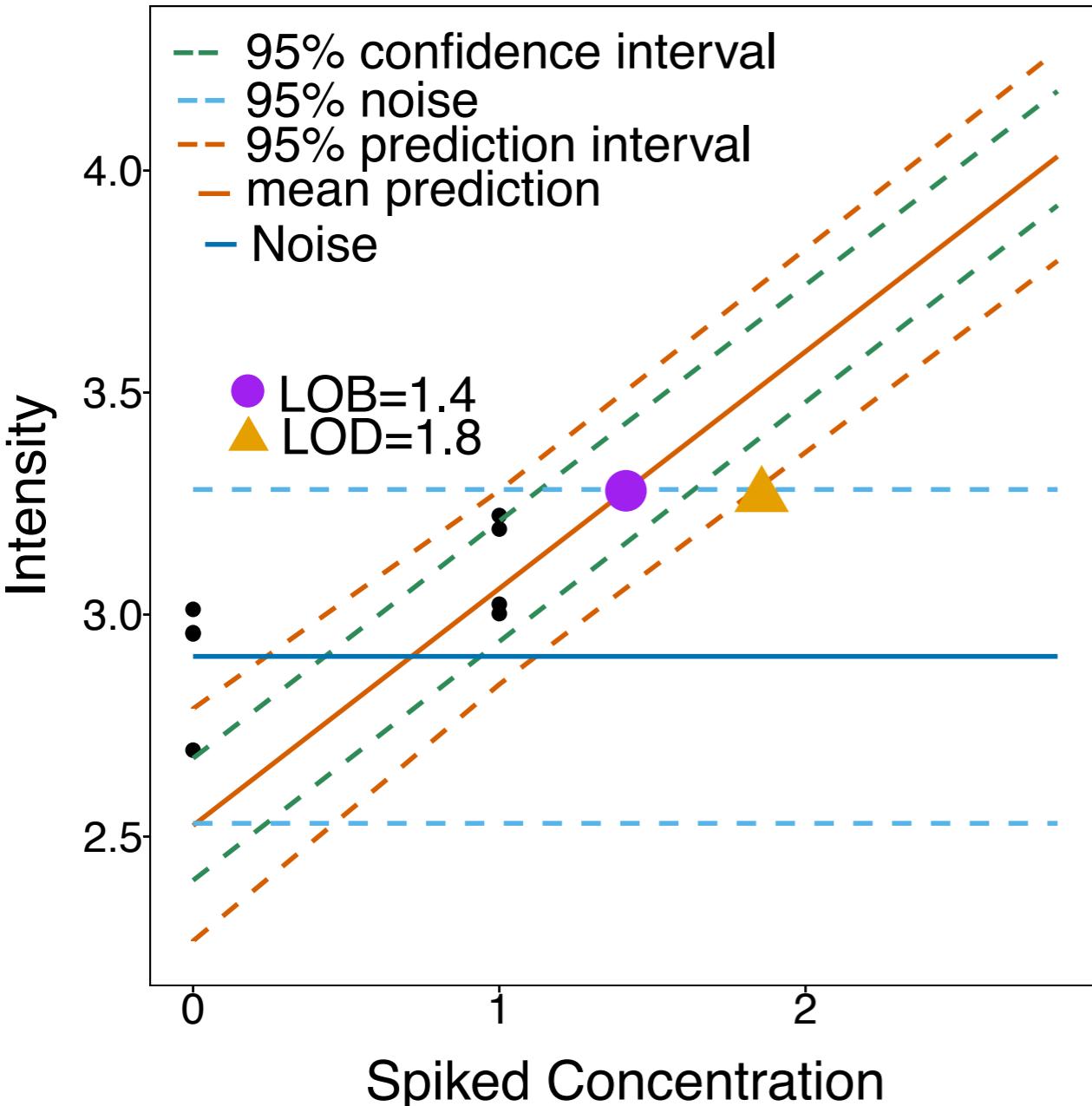
- From the graph
 - ◆ Linear relationship is theoretically plausible
 - ◆ High correlation, high R²



*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

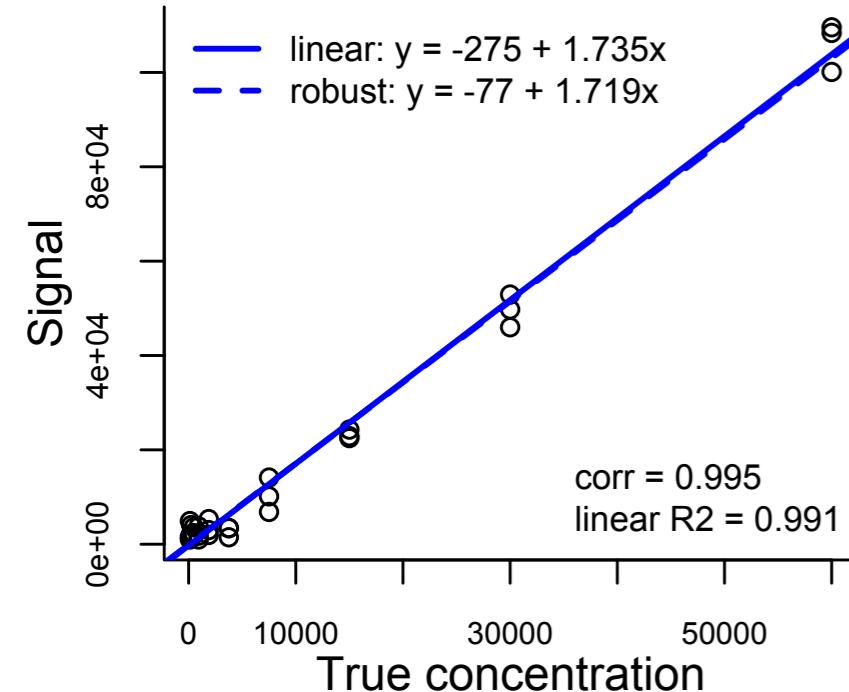
Confidence interval: $\left[\hat{y} - t_{\alpha/1}^{n-2} \cdot \sqrt{\text{var}\{\hat{y}\}}, \hat{y} + t_{\alpha/1}^{n-2} \cdot \sqrt{\text{var}\{\hat{y}\}} \right]$

Prediction interval: $\left[\hat{y} - t_{\alpha/1}^{n-2} \cdot \sqrt{\text{var}\{\hat{y}\} + s^2}, \hat{y} + t_{\alpha/1}^{n-2} \cdot \sqrt{\text{var}\{\hat{y}\} + s^2} \right]$

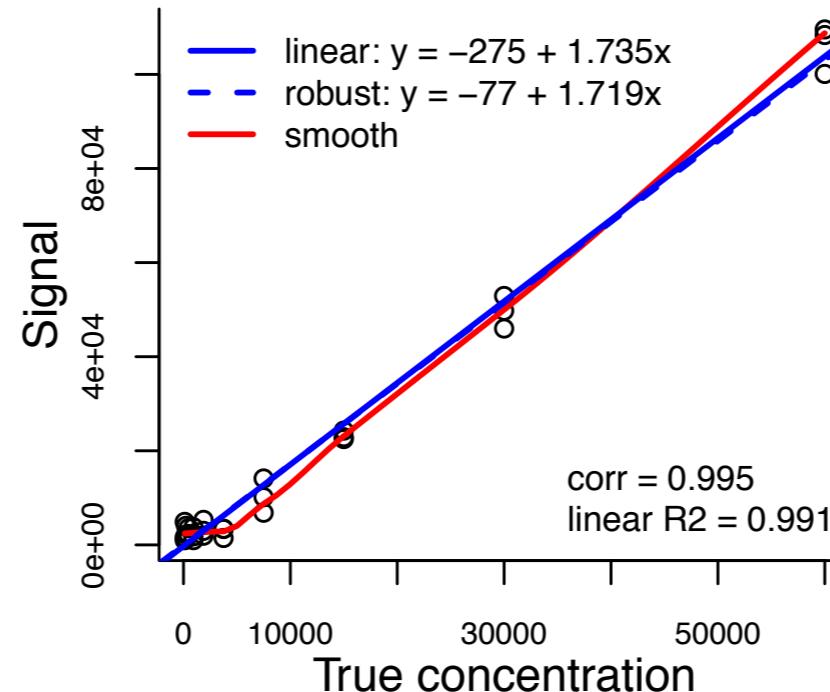
PROBLEM I

Linear fit may be uninterpretable, or wrong

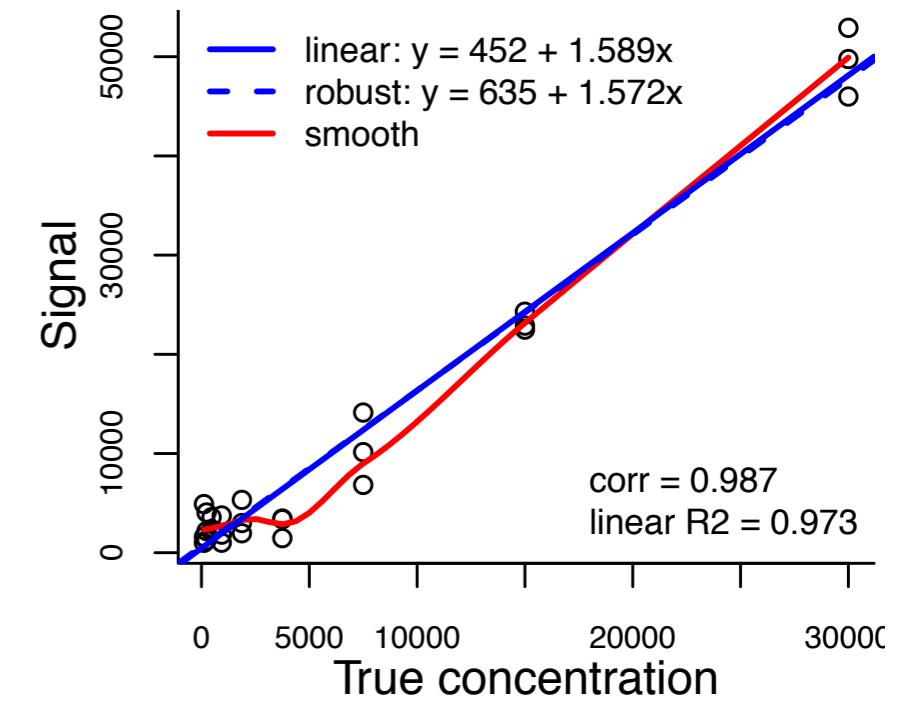
All concentrations



All concentrations,
smoothed



No top concentration,
smoothed



Negative intercept
(no interpretation)

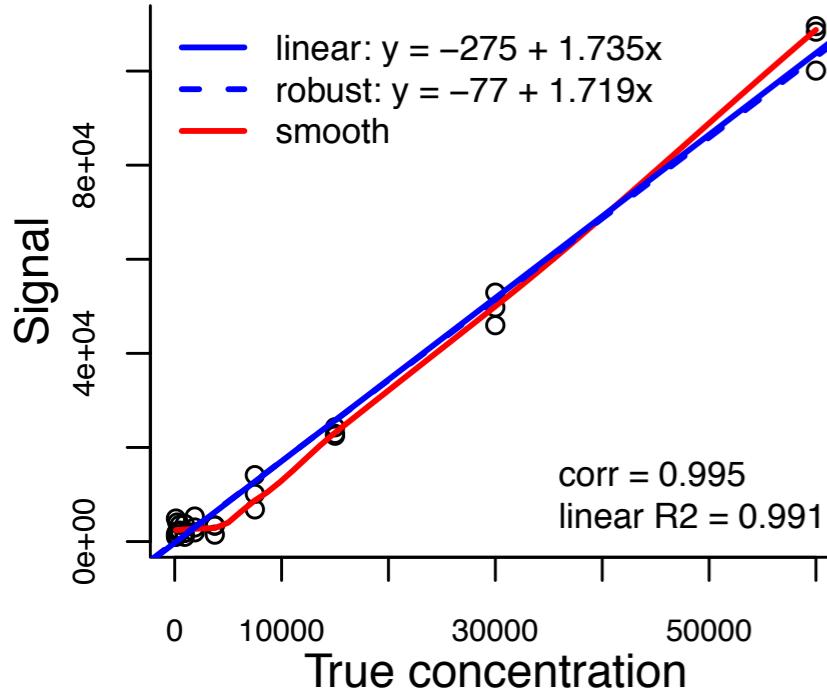
Perception of linearity depends
on the # of concentrations

Correlation, slope and R²
do not quantify linearity

PROBLEM 2

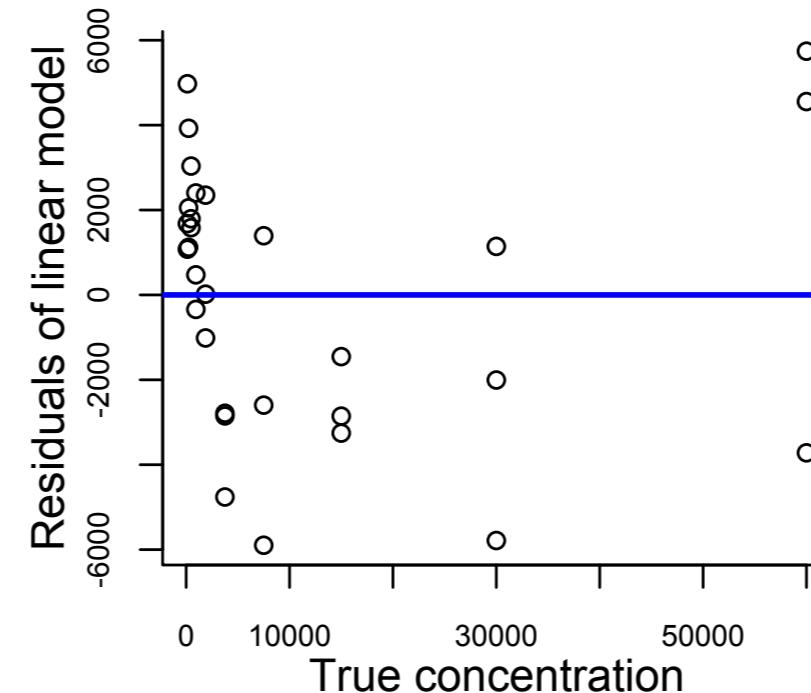
Unequal leverage, unequal variance

All concentrations



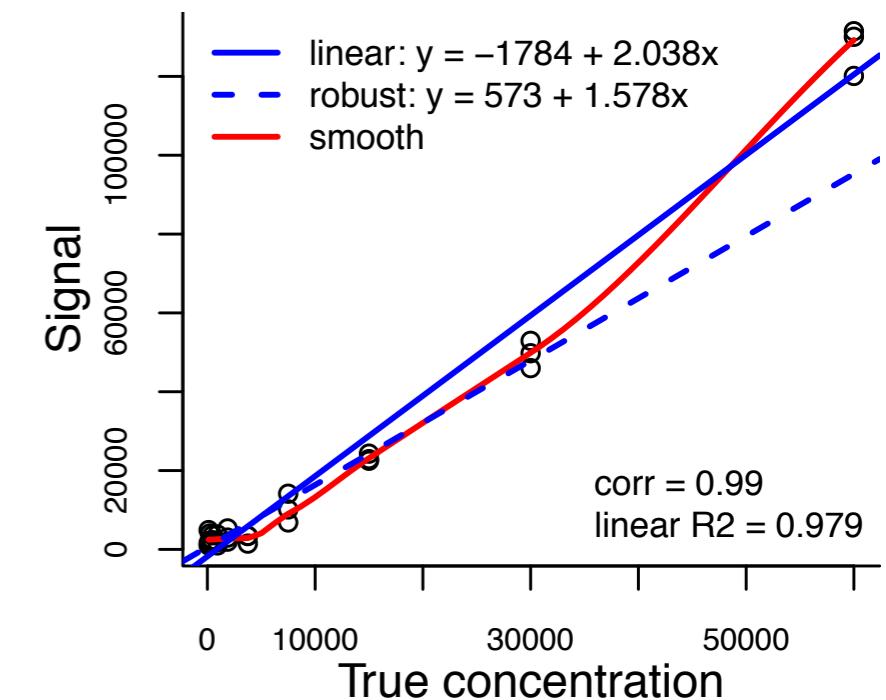
High concentrations are influential

Residuals of the linear model fit



High concentrations have more variance
Low concentrations have worse fit

Intensities at max concentration up by 20%



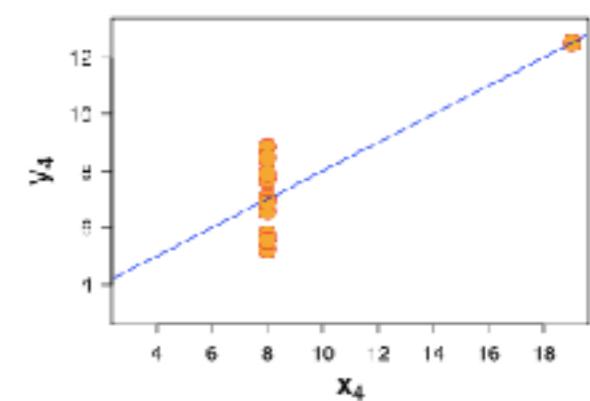
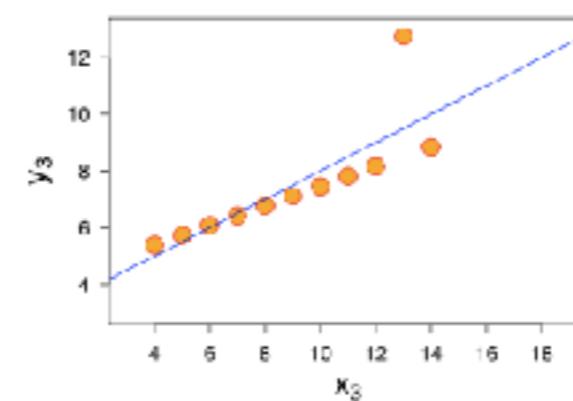
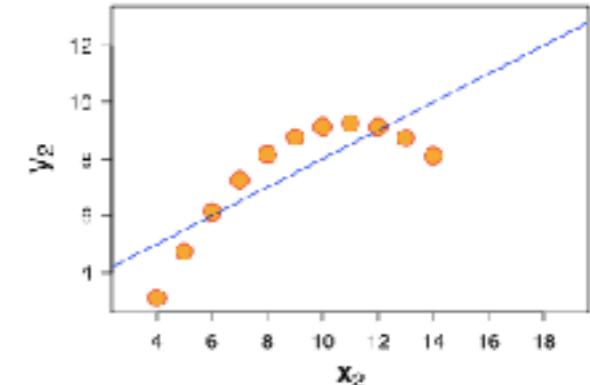
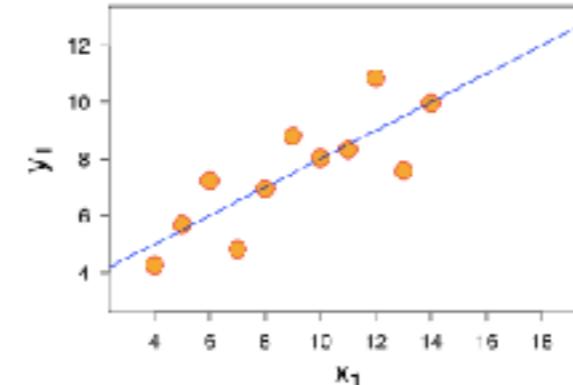
Moving the intensities produces a similarly good fit

The fit is unduly affected by highly variable values
Poor fit at low concentrations is unnoticed

PROBLEM I

Linear fit may be uninterpretable, or wrong

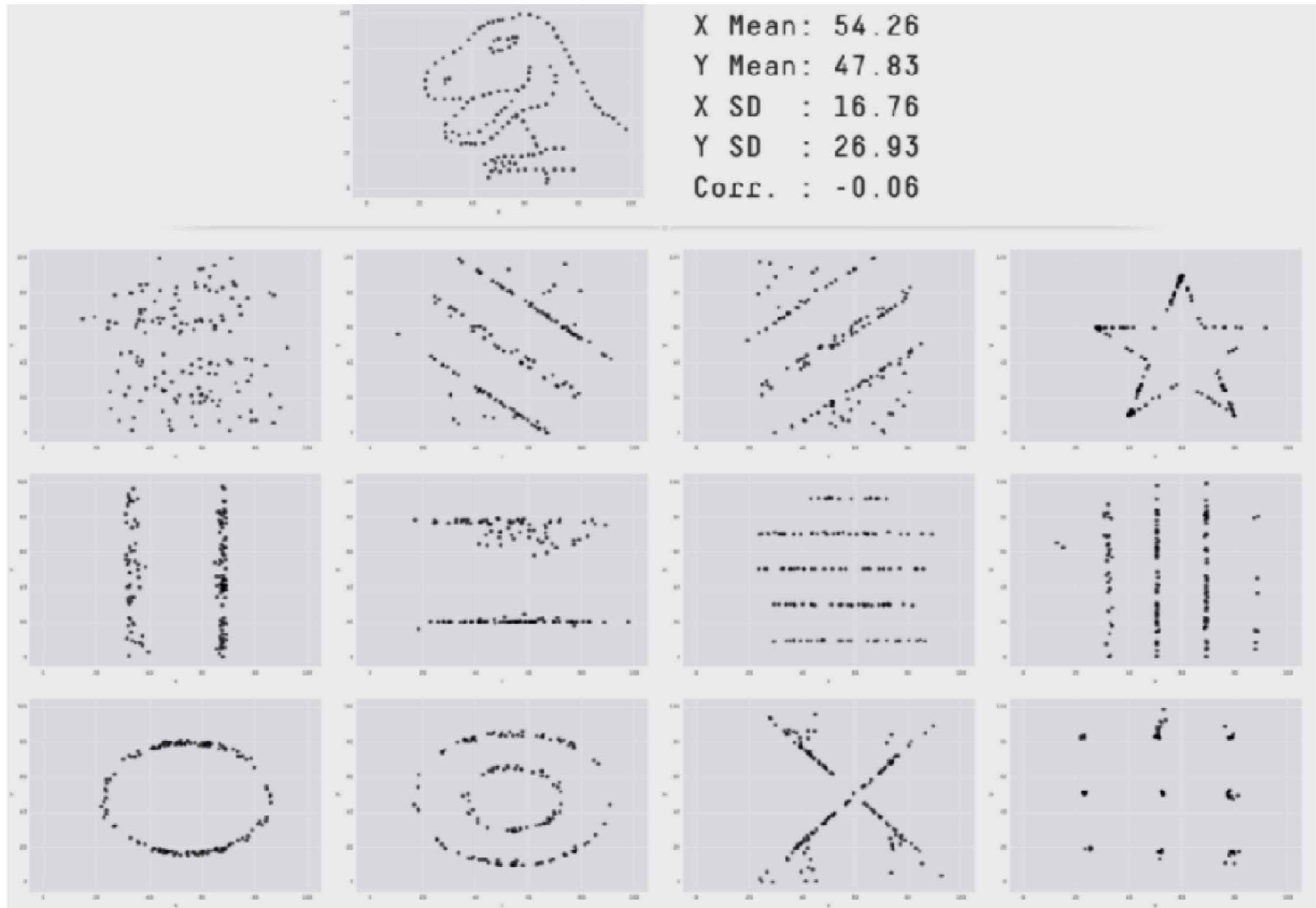
Property	Value
Mean of x	9
Sample variance of x	11
Mean of y	7.50
Sample variance of y	4.125
Correlation between x and y	0.816
Linear regression line	$y = 3.00 + 0.500x$



Anscombe quartet: all datasets have same means, variances, correlations and R^2

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

DATASAURUS

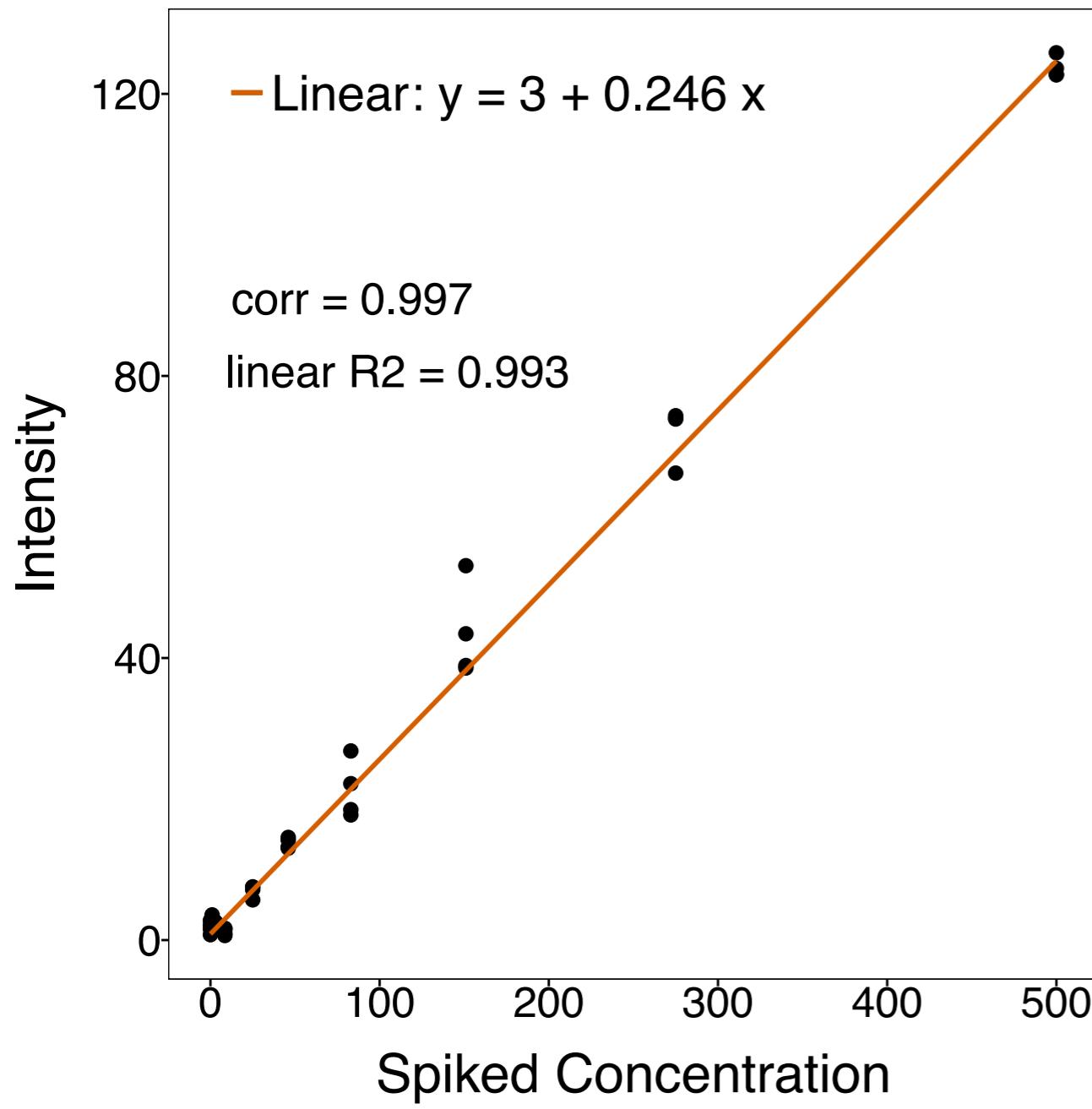


All datasets have same means, variances, correlations and R^2

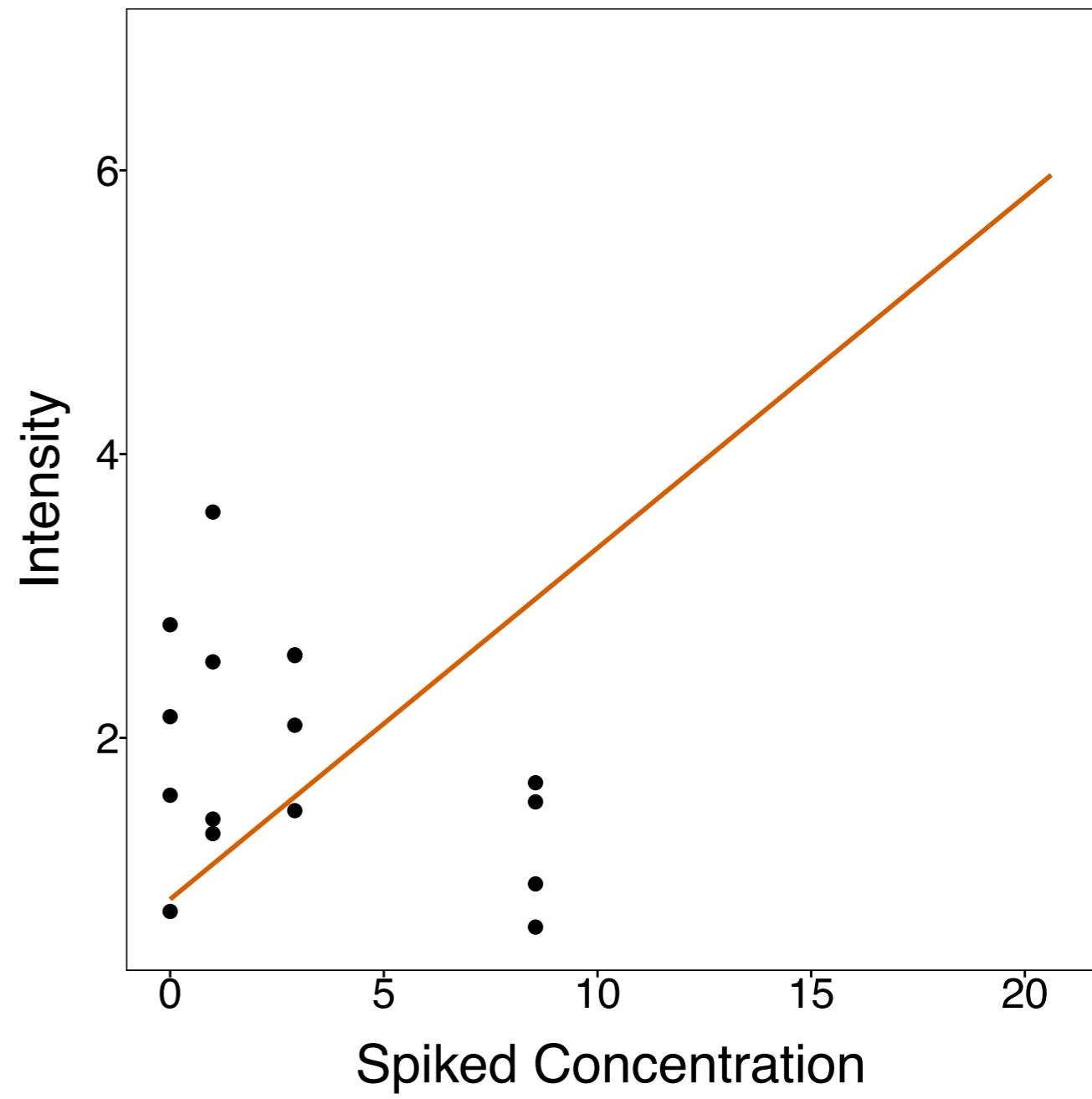
<https://www.autodeskresearch.com/publications/samestats>

ZOOM INTO LOW CONCENTRATIONS

High R² does not always mean good fit



Zoom out

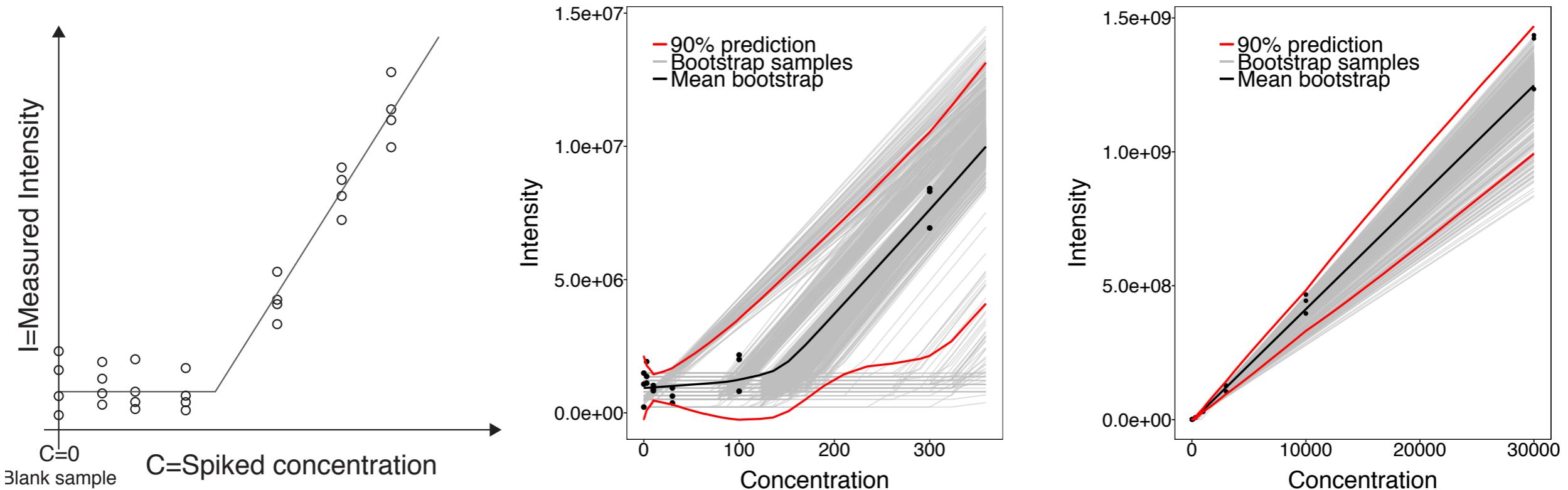


Zoom in

Calibration experiment, SRM , CPTAC

PROPOSED APPROACH

Canonical calibration curve + resampling (bootstrap)

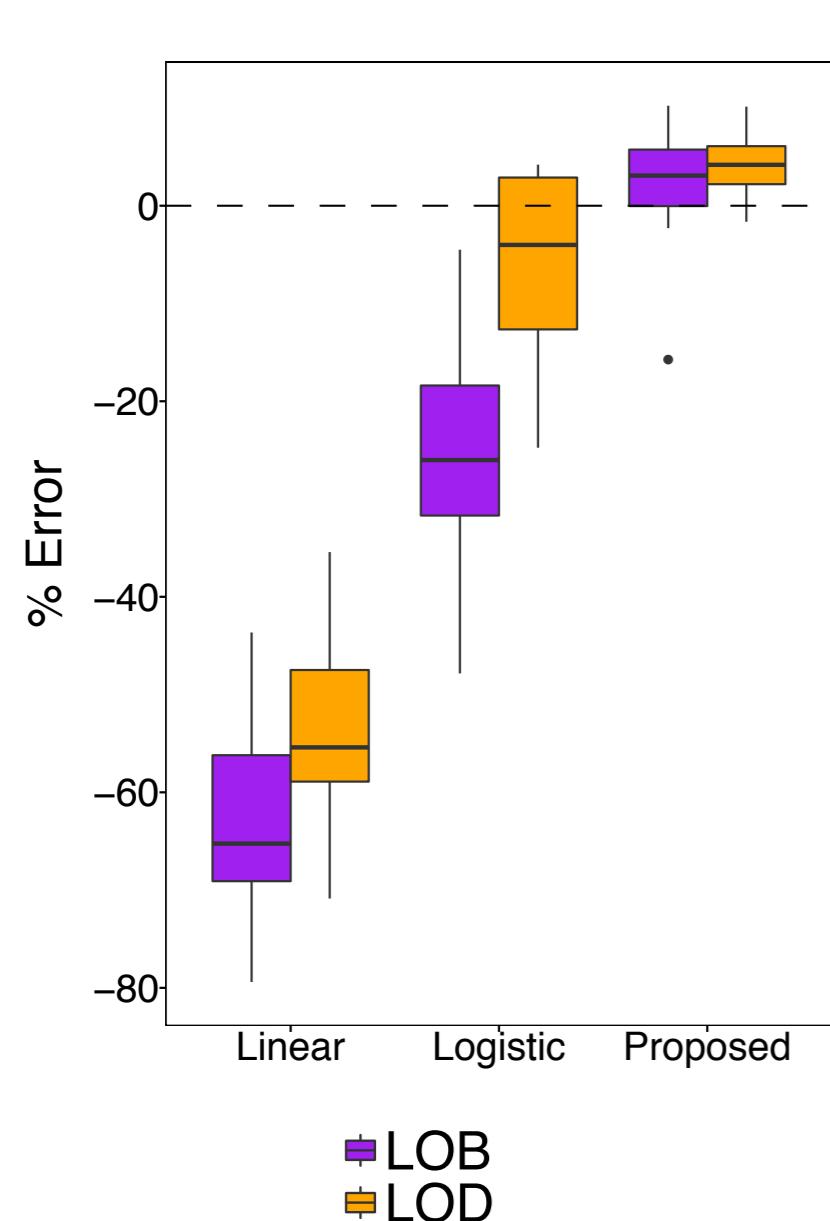


$$Y_{ij} = \begin{cases} \text{Intercept} + \text{Noise}_{ij} \\ \text{Intercept} + \text{Slope} \times (C_i - \text{Change}) + \text{Noise}_{ij}, & \text{if } C_i \geq \text{Change} \end{cases}$$

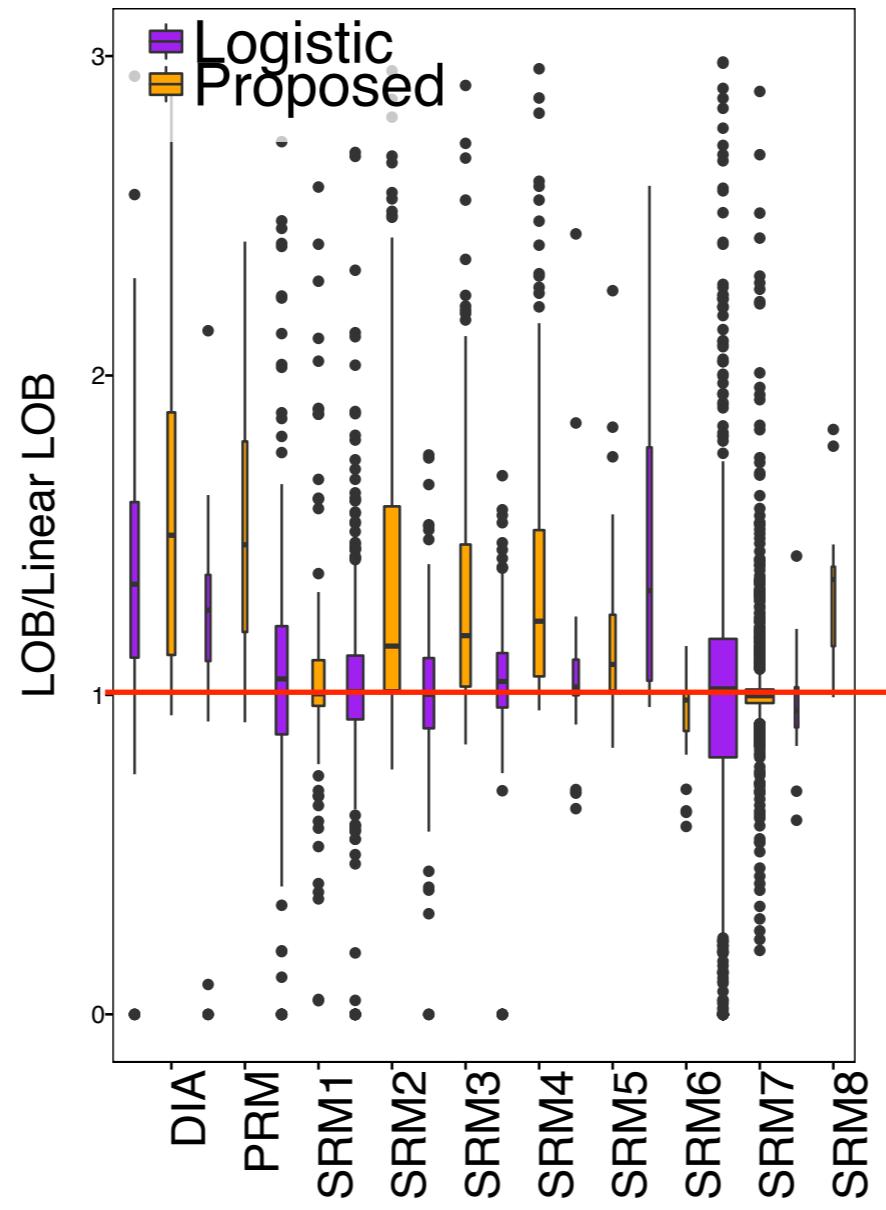
The nonlinear fit expresses uncertainty in change point location
Yields more accurate & conservative figures of merit

ADVANTAGES OF NON-LINEAR REGRESSION

More accurate & conservative figures of merit



Simulation



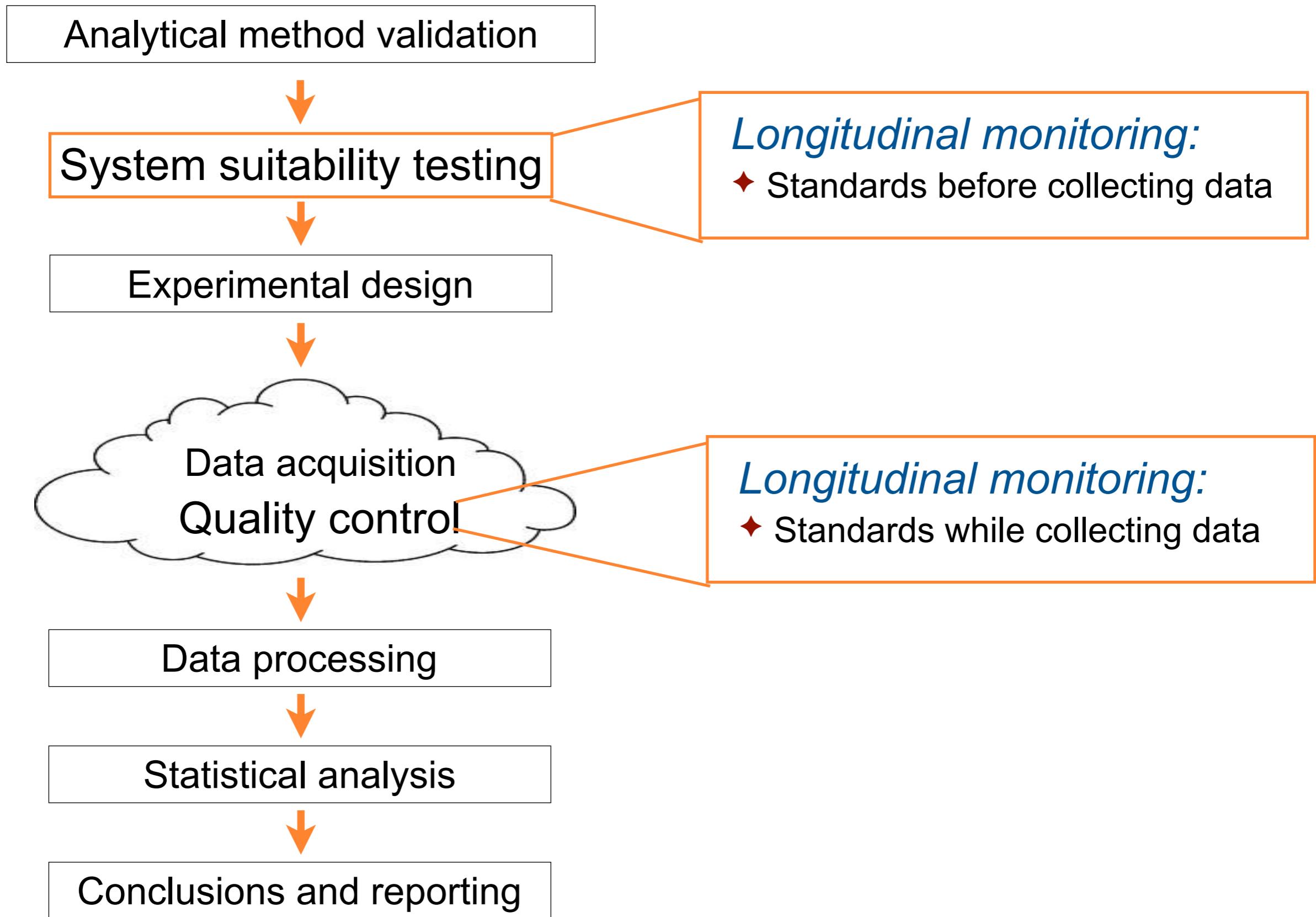
Experimental datasets

Galitzine *et al.*, Molecular and Cellular Proteomics, 2018

OUTLINE

- Motivating example
 - ABRF iPRG study
- MSstats
 - Statistical relative quantification of proteins and peptides
 - Methods evaluation
- Extensions to MSstats
 - Assay characterization
 - System suitability and quality control

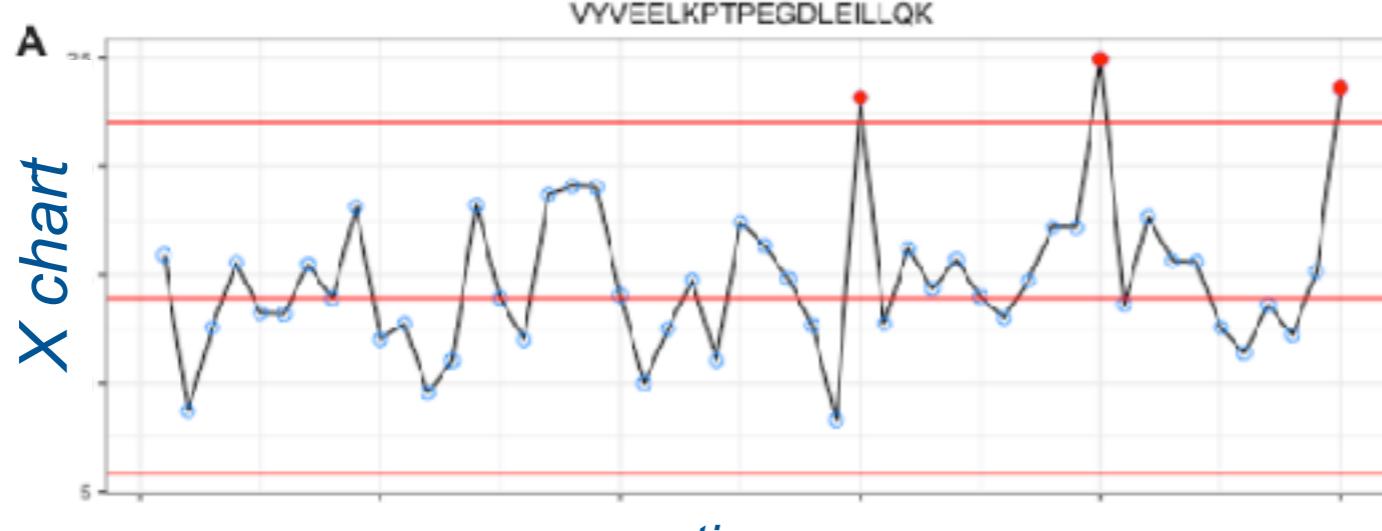
MSI EXPERIMENT: STATISTICIAN'S VIEW



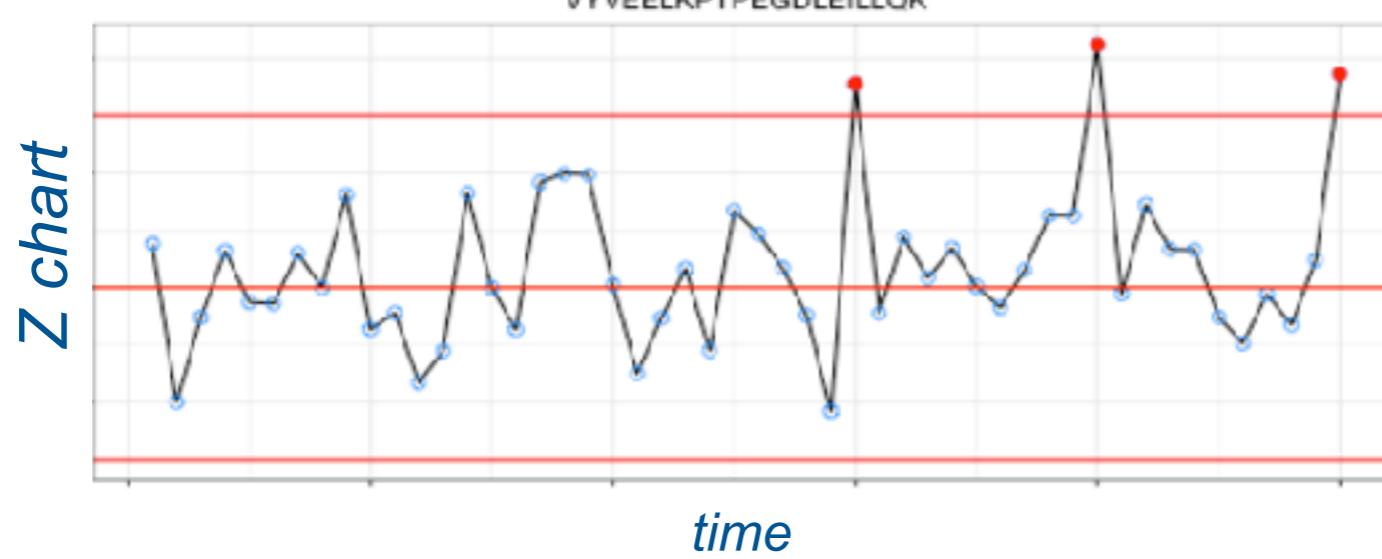
Statistical process control

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

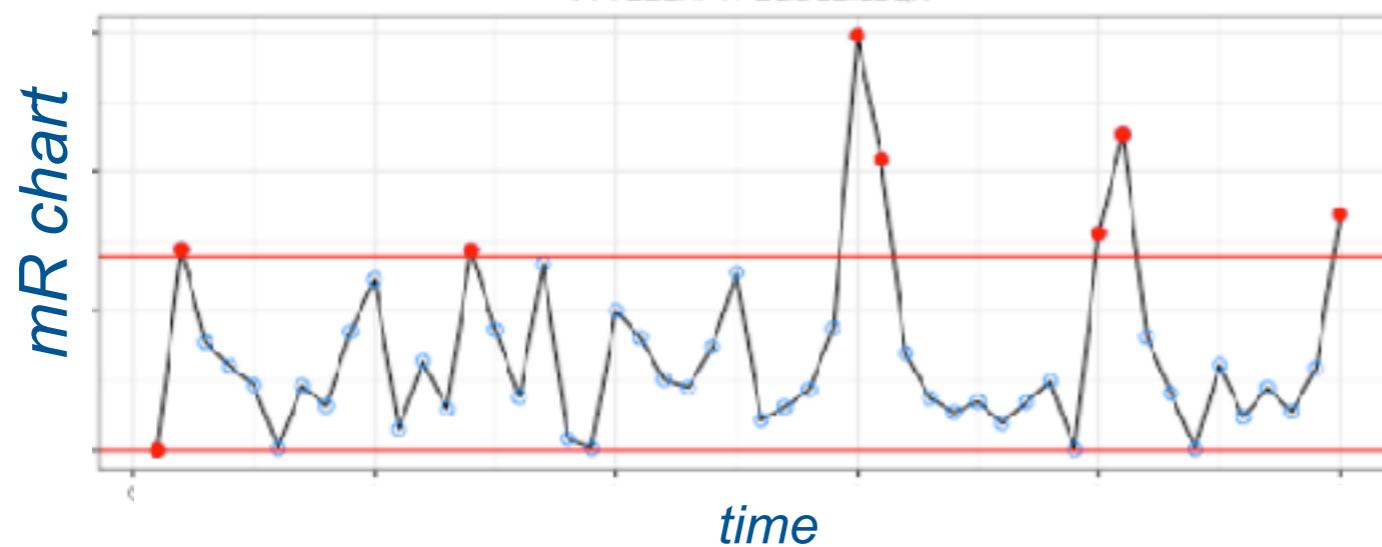
Mean signal



B



Variation



monitor mean

monitor standardized mean

monitor variation
moving range

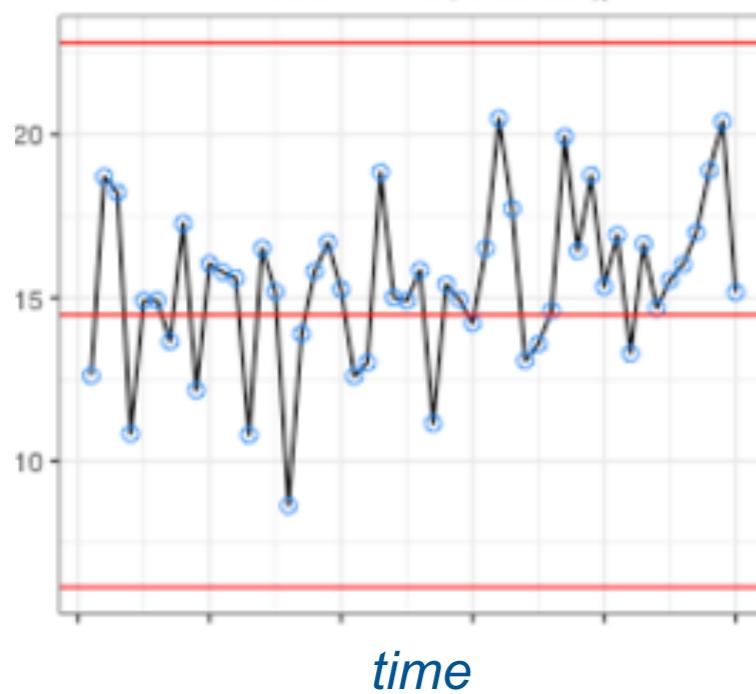
Detects large changes

Longitudinal profiles of a standard (e.g. peak area)

Mean signal

X chart

VYVEELKPTPEGDLEILLQK

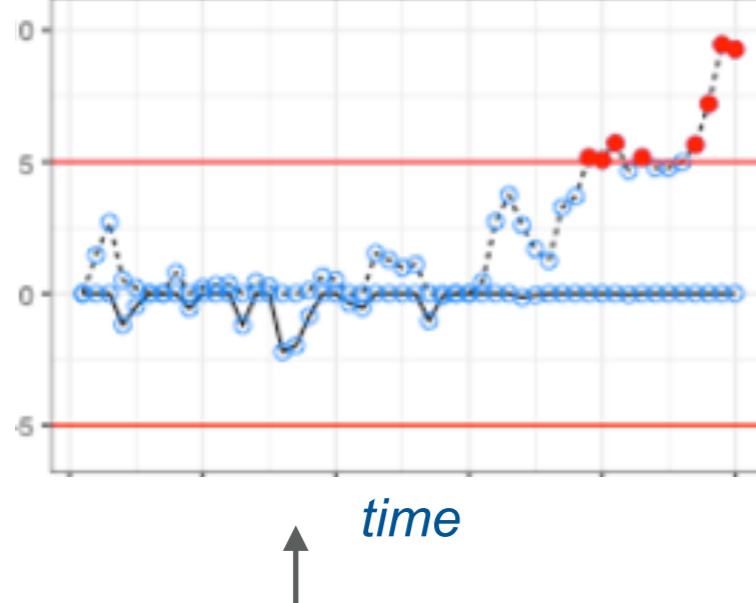


time

Mean signal

CUMSUM \bar{m} chart

VYVEELKPTPEGDLEILLQK



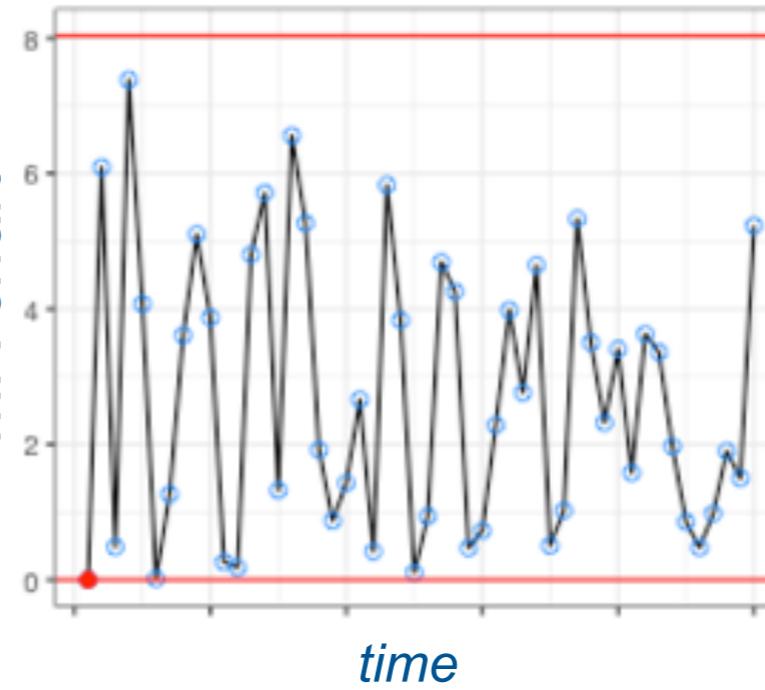
time

cumulative change
in mean

Variation

mR chart

VYVEELKPTPEGDLEILLQK

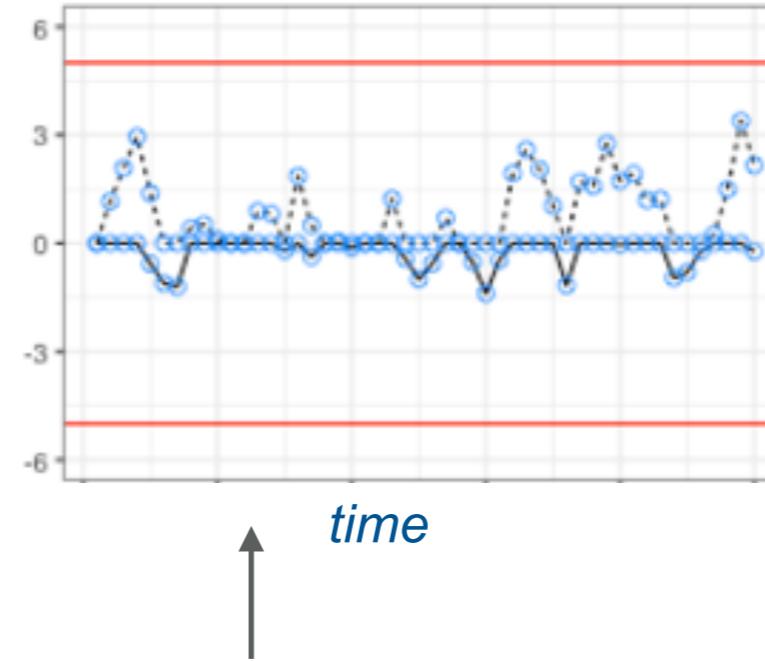


time

Variation

CUMSUMv chart

VYVEELKPTPEGDLEILLQK



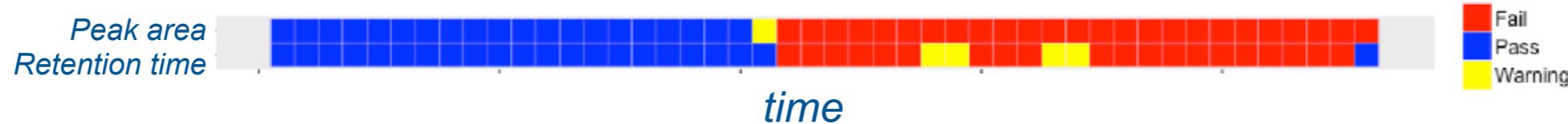
time

cumulative change
in variation

*Detects small
sustained drifts*

Multi-chart summary

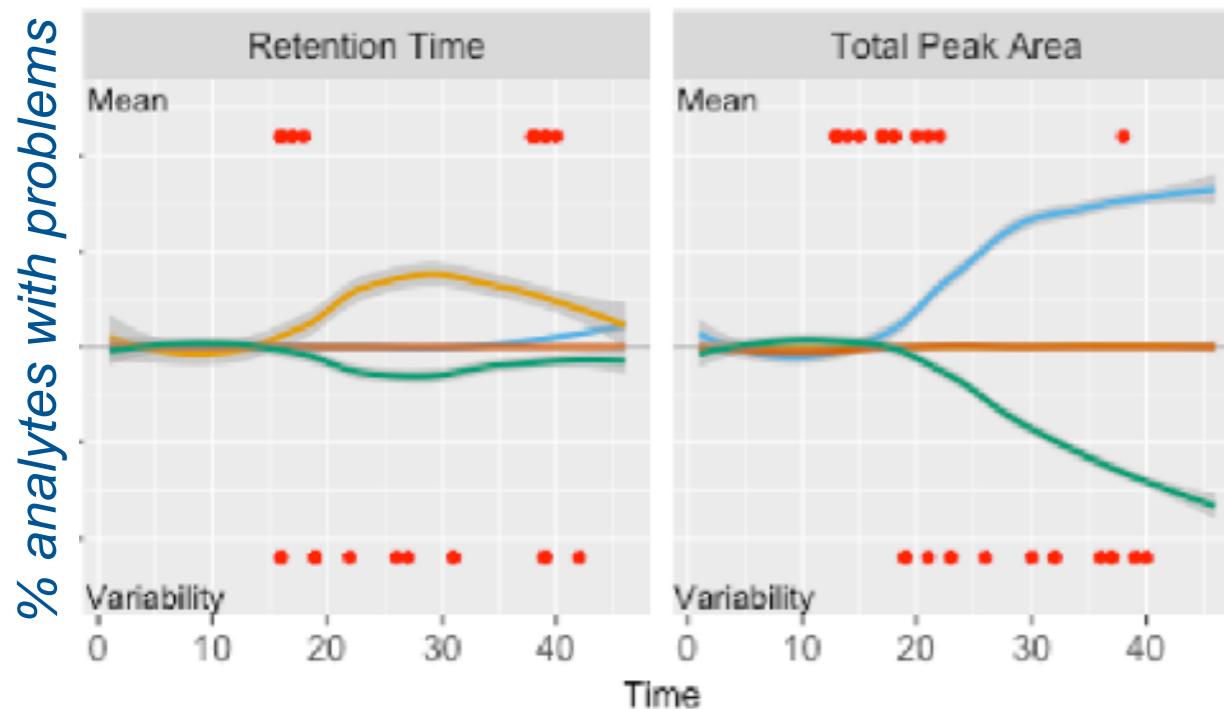
Decision map: sustained drift in mean



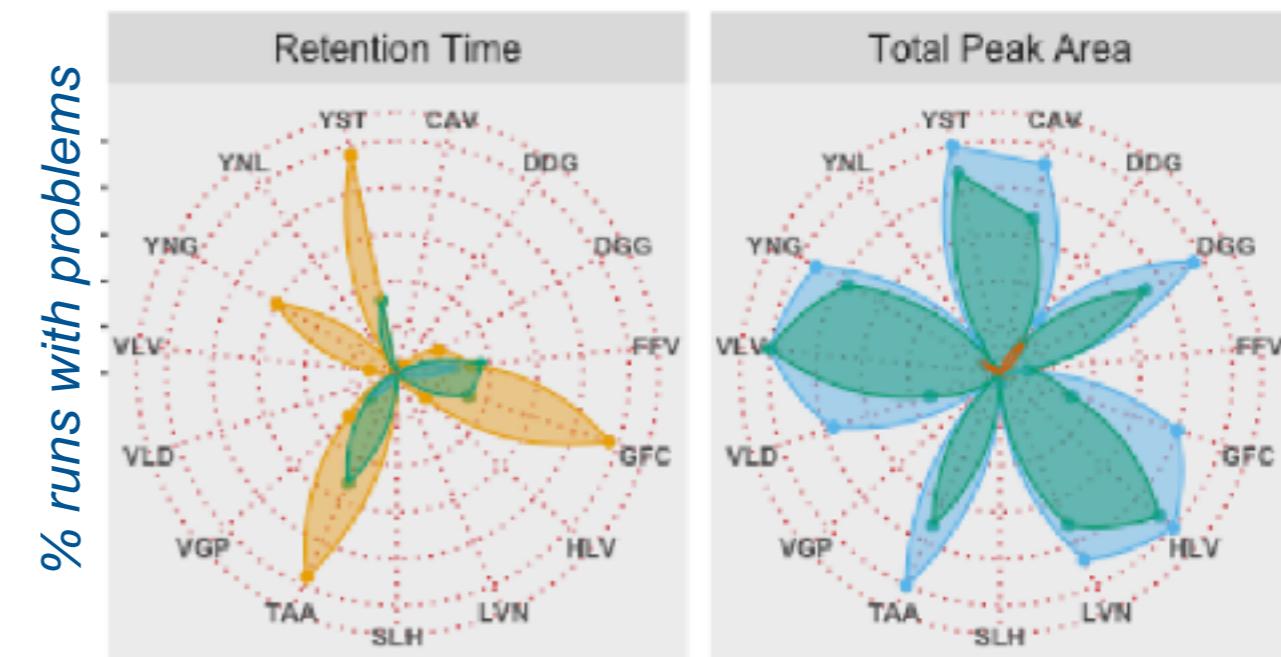
Decision map: sustained drift in variance



River plots: sustained drift in mean & variance



Radar plots: sustained drift in mean & variance



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

ACKNOWLEDGEMENTS

Northeastern University

Kylie Bemis
 Meena Choi
 Dan Guo
 Sicheng Hao
 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

ETH Zurich

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

Mugla University

Eralp Dogu



Support:

NSF
 NIH
 Sternberg Chair
 Canary Center
 Roche
 Genentech
 Eli Lilly