R Fundamentals and Best Practices for Mass Spectrometry Data Analysis

Sunday, November 15 (12:00-3:15pm Eastern)

Meena Choi, Genentech
Olga Vitek, Northeastern University

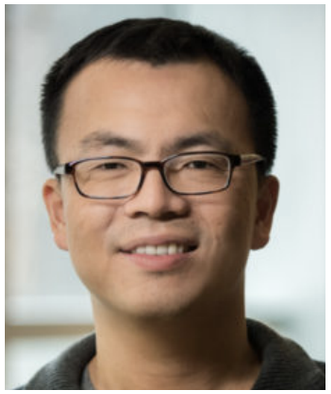
Module #7: Analysis of proteomic data with MSstats

Introduction to MSSTATS

Meena Choi
Genentech

Ting Huang
Northeastern

Mateusz Staniak
University of Wrocław



Tsung-Heng Tsai
Kent State

and

Veavi Chang
Tim Clough
Eralp Dogu
Cyril Galitzine
Devon Kohler
Sara Taheri
Mateusz Staniak
Akshay Kulkarni
Sumedh Sankhe

Olga Vitek
Khoury College of
Computer Sciences
Northeastern University

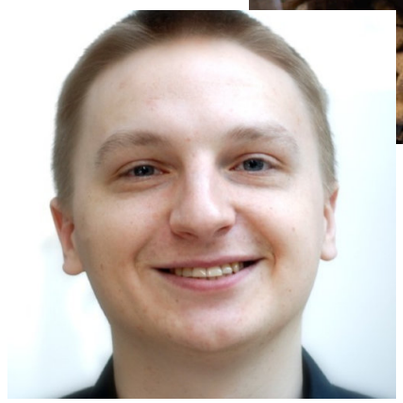
MSSTATS

www.msstats.org

R-based open-source tool for quantitative proteomics

- Which proteins change in abundance?
  - Complex designs: factorial experiments, paired designs, time course
  - Abundance of each subject on a relative scale
- Label-free or label-based
  - Shotgun DDA, data independent DIA/SWATH, PRM, targeted SRM
  - TMT; soon: differential PTMs
- Data processing
  - Free, open-source and inter-operable with other tools
  - Converteres to Skyline, MaxQuant, OpenMS, Spectronaut, ProteomeDiscoverer…
- Statistical analysis
  - Flexible models, account for complex designs, missing values and outliers
  - Stable parameter estimation
- Method validation, system suitability, experimental design
  - System suitability/QC, assay characterization
  - Sample size for testing and classification

Relative protein quantification in MS-based proteomics



Meena Choi
Ting Huang
Mateusz Staniak
Tsung-Heng Tsai

- Analytical method validation
- System suitability testing
- Experimental design
- Data acquisition
  Quality control
- Data processing
- Statistical analysis
- Data for future designs

Assay characterization
LOB/LOD

MSstatsQC
Software package
System suitability monitoring

Experimental design
Sample size for testing and classification

R software package	*Open source software*	*Commercial software*
MSnbase	**Skyline, MaxQuant**	**Proteome Discoverer**
RforProteomics	**OpenMS, OpenSWATH**	**Spectronaut, SpectroMine**
	DIA-Umpire	**Progenesis**

MSstats
Software package
Significance analysis for DDA, SRM, DIA

MSstatsTMT
Software package
Significance analysis for TMT

MSstatsBioData
Experiment package
Published studies with DDA or SRM

42 datasets, > 178 reanalyses

MassIVE
Mass Spectrometry
Interactive **V**irtual **E**nvironment

INTRODUCTION TO MSSTATS

- Overview
- Statistical modeling
- Community resources

MS EXPERIMENT: DESIGN

Analytical method validation

System suitability testing

Experimental design

Data acquisition
Quality control

Study goals

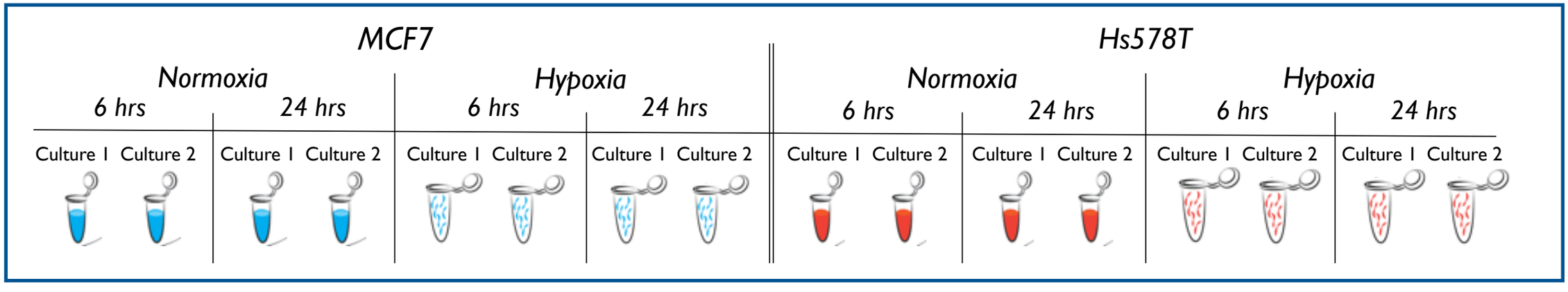
- Exploratory analysis
- Differential analysis
- Biomarker discovery

Biological aspects

- Selection of conditions
- Selection of replicates

Technological aspects

- Sample prep
- Data acquisition
- Randomization



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

MS EXPERIMENT: DESIGN

Analytical method validation

↓

System suitability testing

↓

Experimental design

↓

Data acquisition
Quality control

Study goals

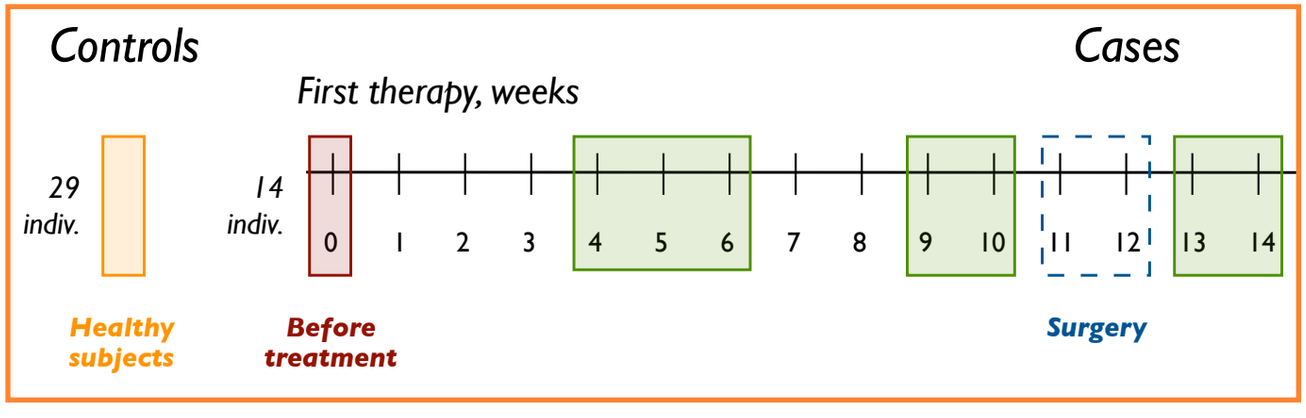
- Exploratory analysis
- Differential analysis
- Biomarker discovery

Biological aspects

- Selection of conditions
- Selection of replicates

Technological aspects

- Sample prep
- Data acquisition
- Randomization



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

MS EXPERIMENT: DATA PROCESSING

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 FORMAT

	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

Reports produced by (or converted from) data processing tools

MS EXPERIMENT: NORMALIZATION

Analytical method validation

↓

System suitability testing

↓

Experimental design

↓

Data acquisition
Quality control

↓

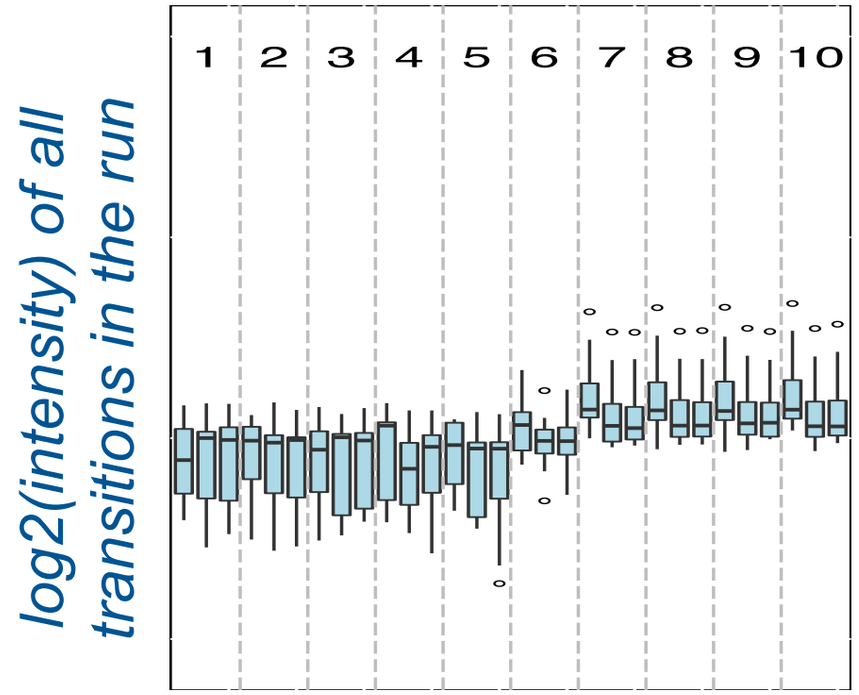
Data processing

↓

Statistical analysis

↓

Experimental design



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

MS EXPERIMENT: NORMALIZATION

Analytical method validation

↓

System suitability testing

↓

Experimental design

↓

Data acquisition
Quality control

↓

Data processing

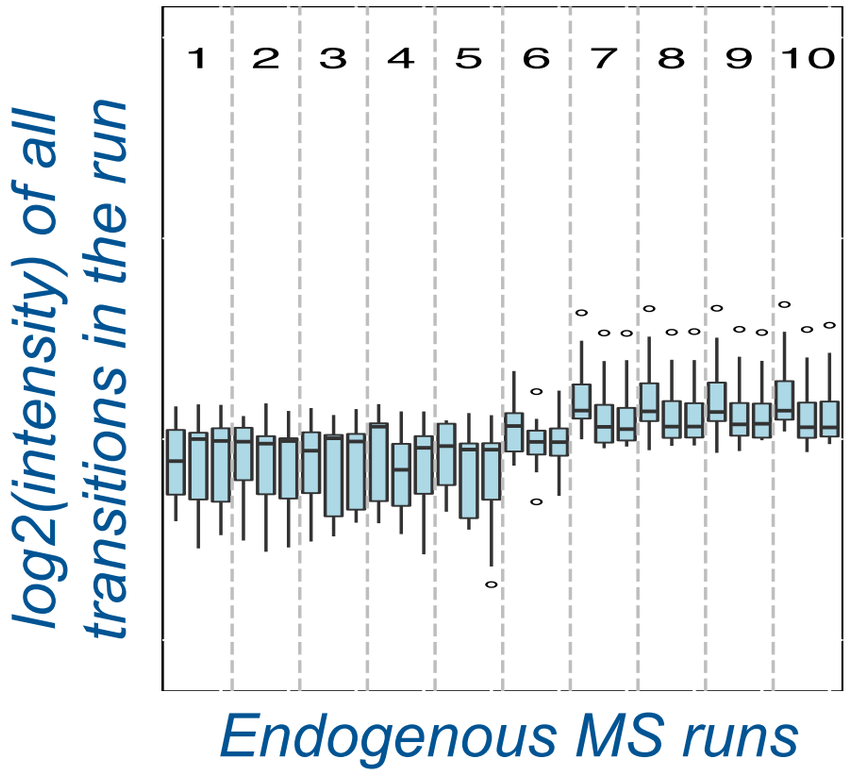
↓

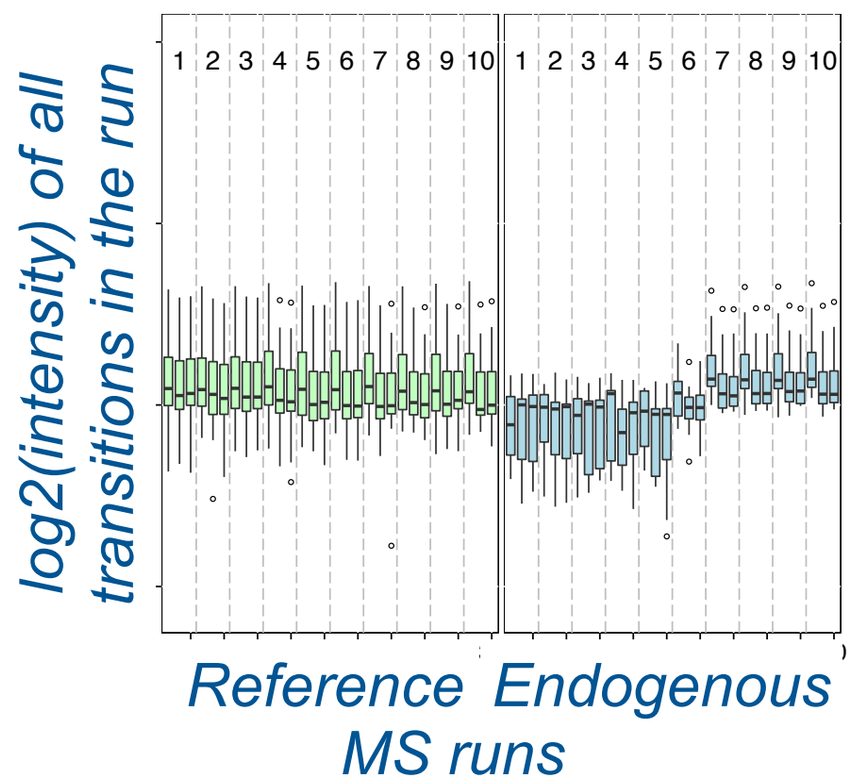
Statistical analysis

↓

Experimental design

NORMALIZATION





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

MS EXPERIMENT: NORMALIZATION

Analytical method validation

↓

System suitability testing

↓

Experimental design

↓

Data acquisition
Quality control

↓

Data processing

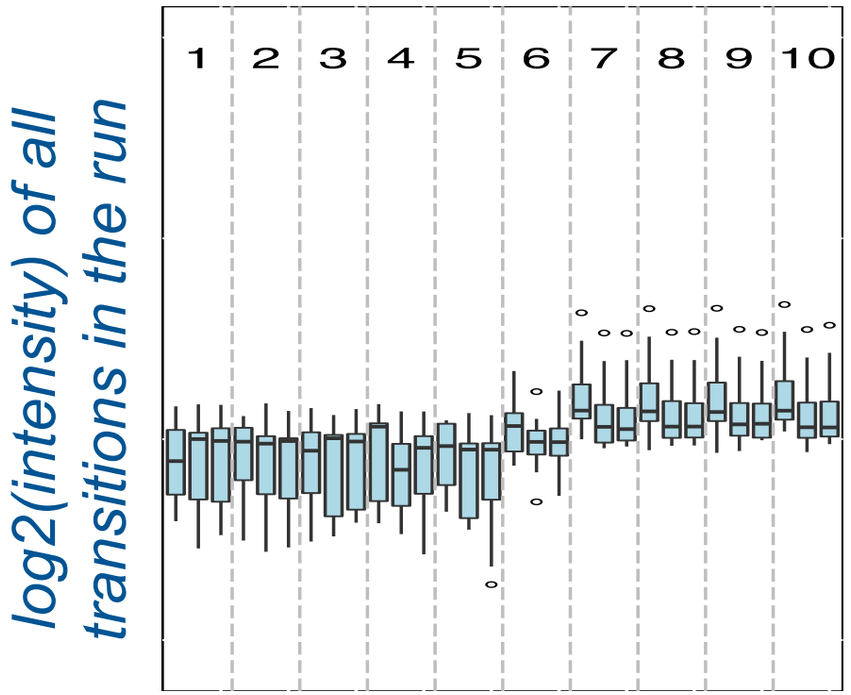
↓

Statistical analysis

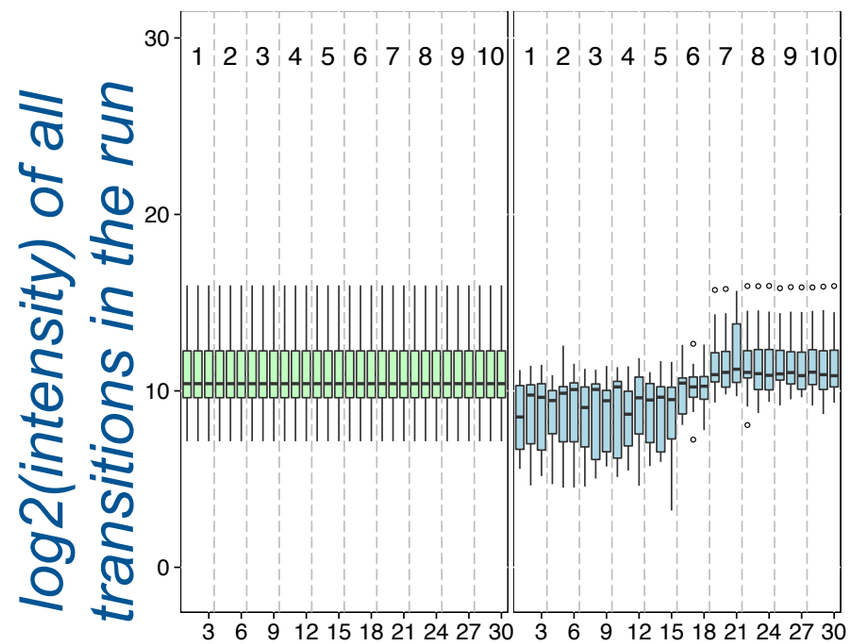
↓

Experimental design

NORMALIZATION



Endogenous MS runs



Reference Endogenous MS runs

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

MS EXPERIMENT: INFERENCE

Analytical method validation

↓

System suitability testing

↓

Experimental design

↓

Data acquisition
Quality control

↓

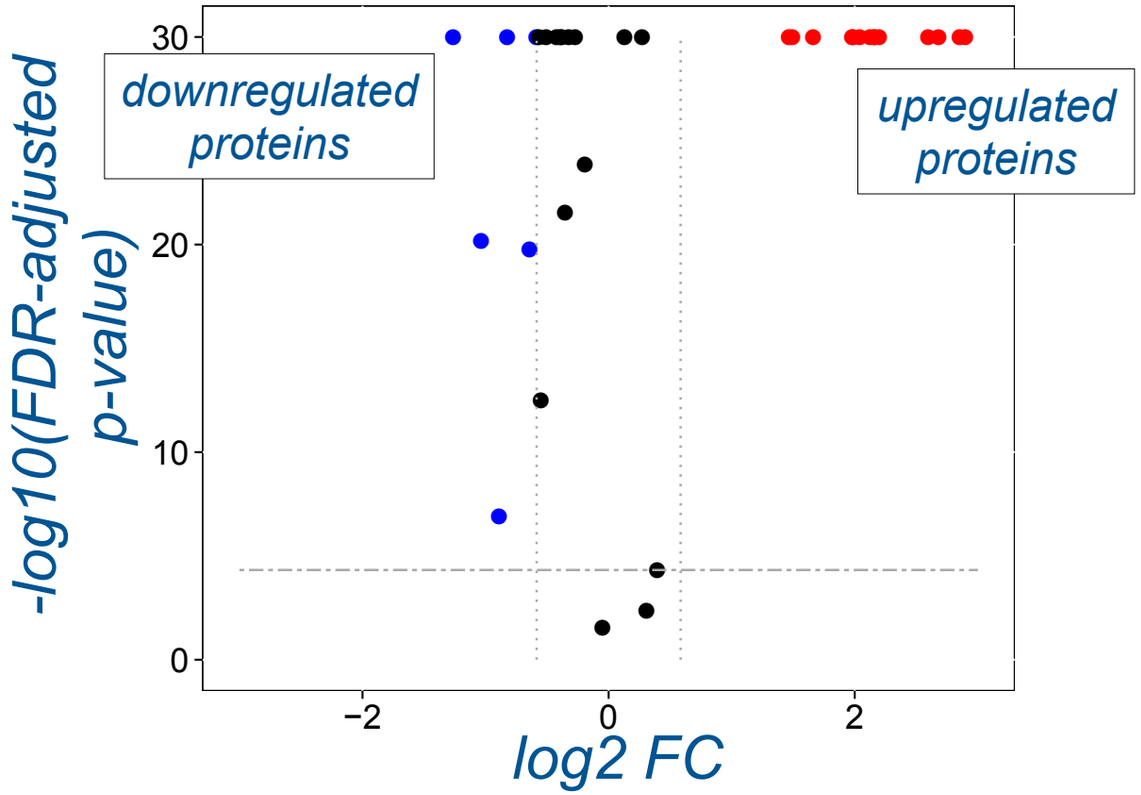
Data processing

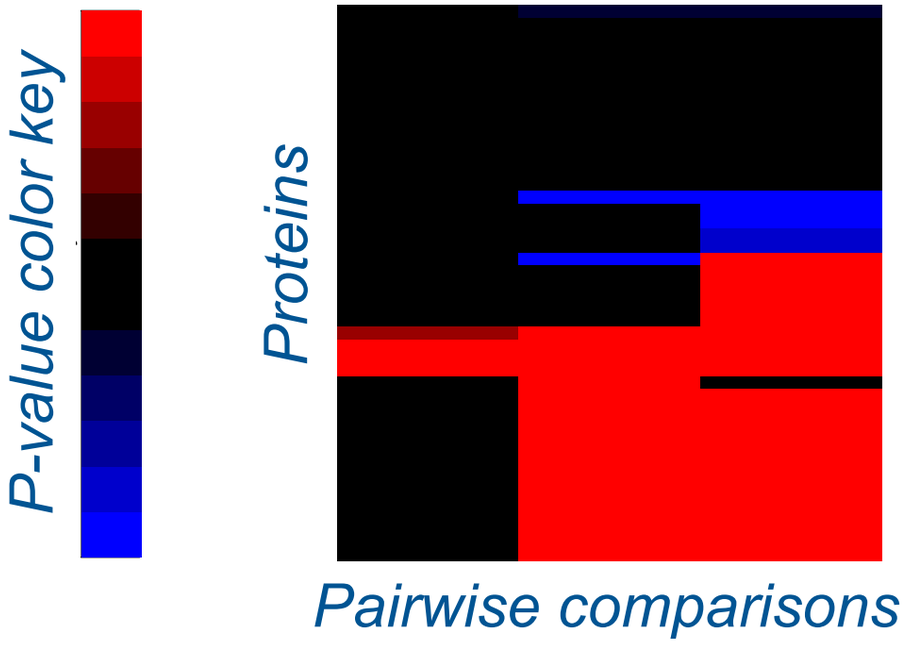
↓

Statistical analysis

↓

Experimental design





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

MS EXPERIMENT: DESIGN

Analytical method validation

↓

System suitability testing

↓

Experimental design

↓

Data acquisition
Quality control

↓

Data processing

↓

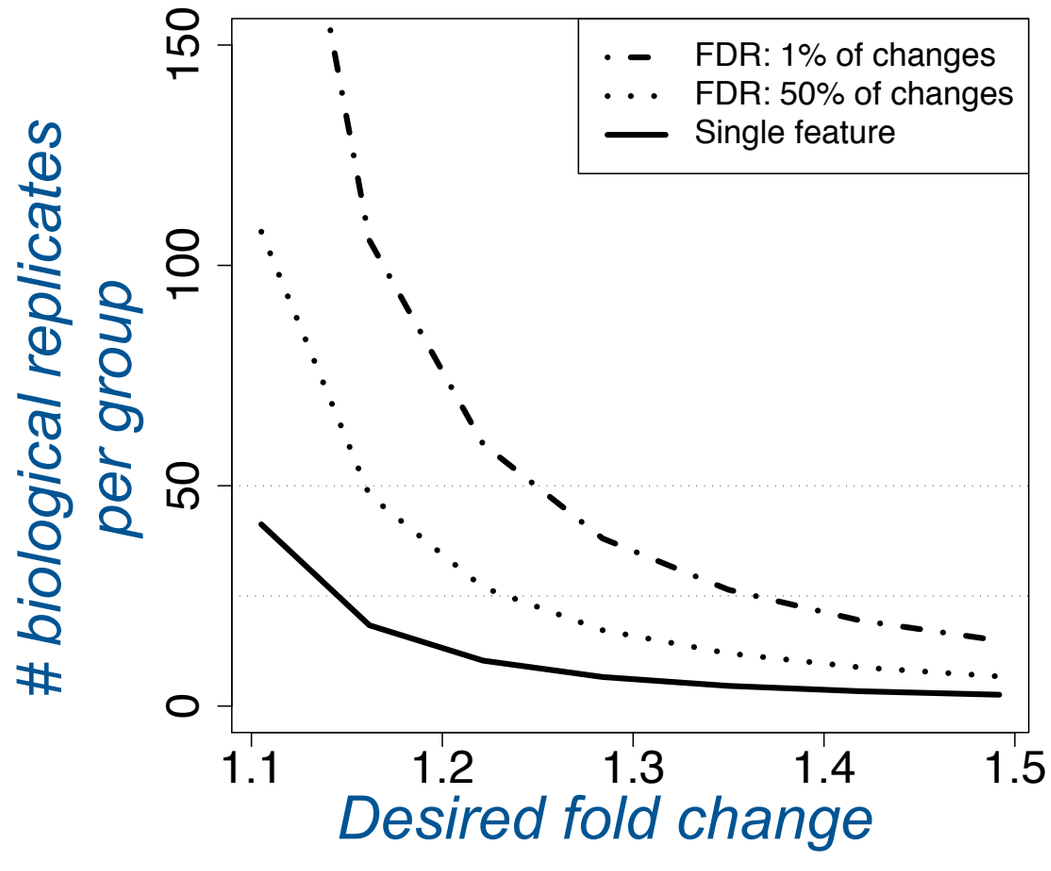
Statistical analysis

↓

Experimental design

Use the dataset to improve:

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



INTRODUCTION TO MSSTATS

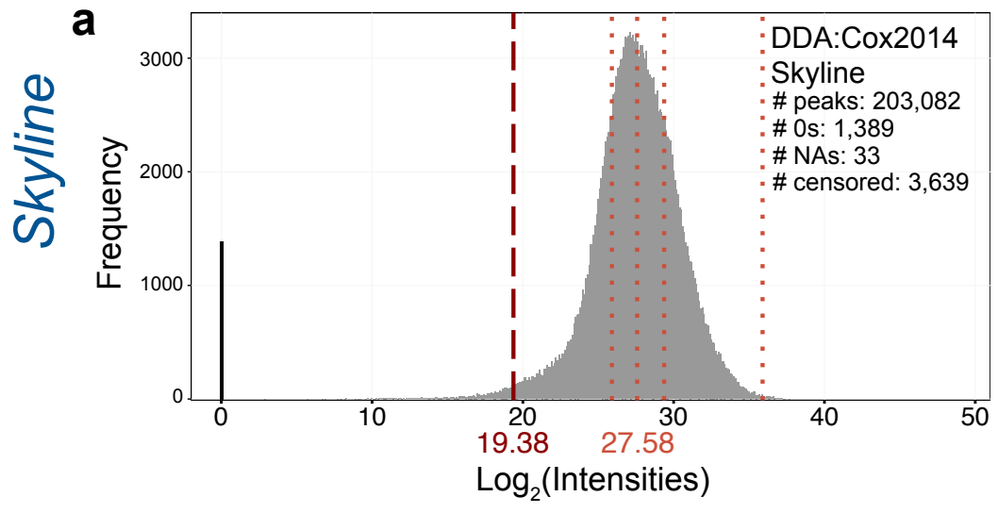
- Overview
- Statistical modeling
- Community resources

PROPERTIES OF PEAK INTENSITIES VARY BETWEEN DATA PROCESSING TOOLS

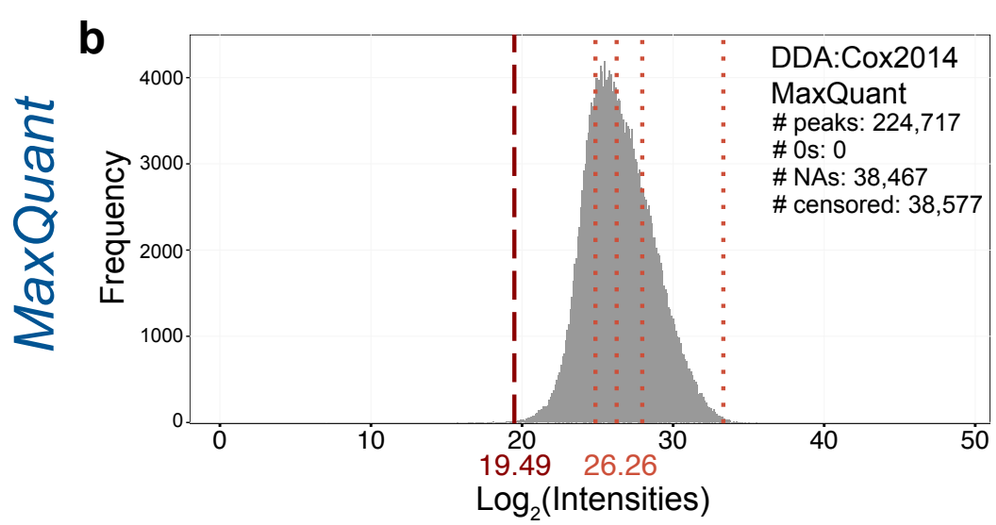
DDA: Cox 2014

DIA: Bruderer 2015

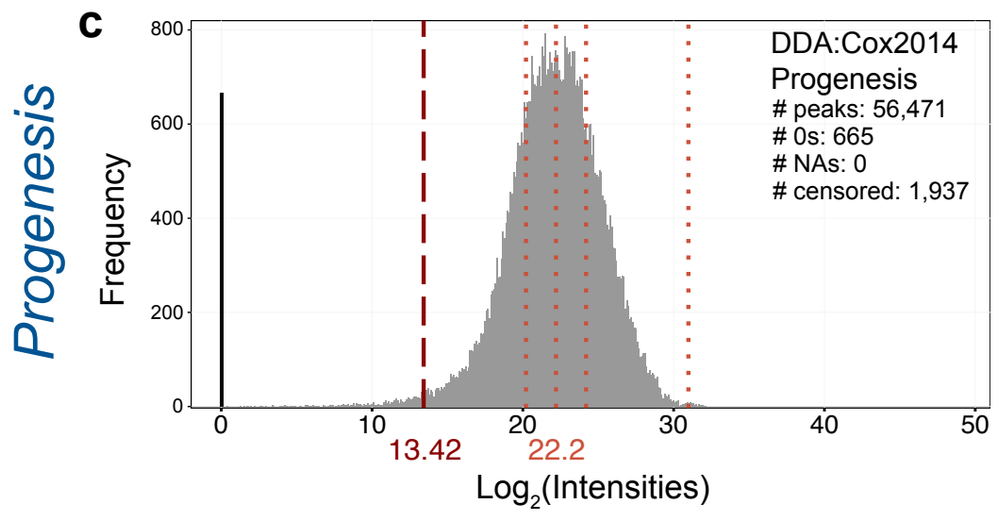
Skyline



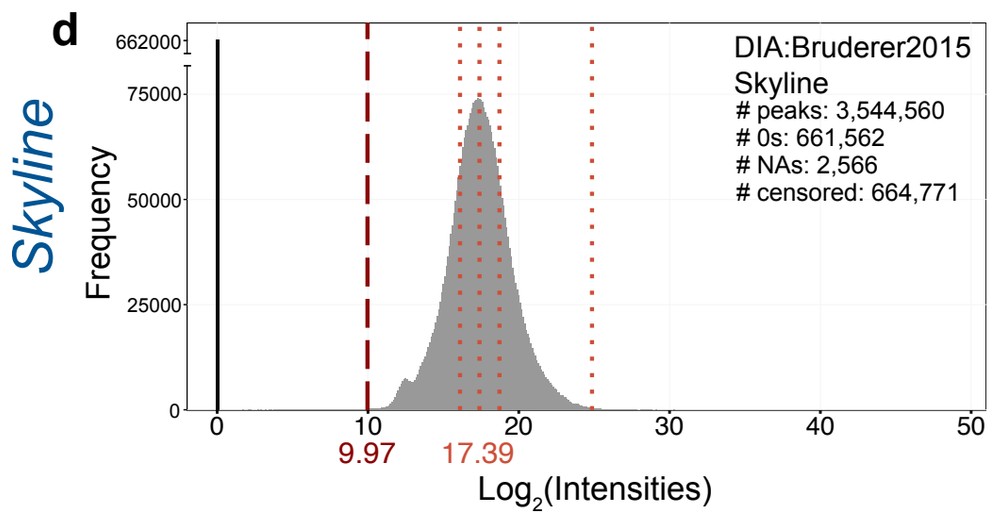
MaxQuant



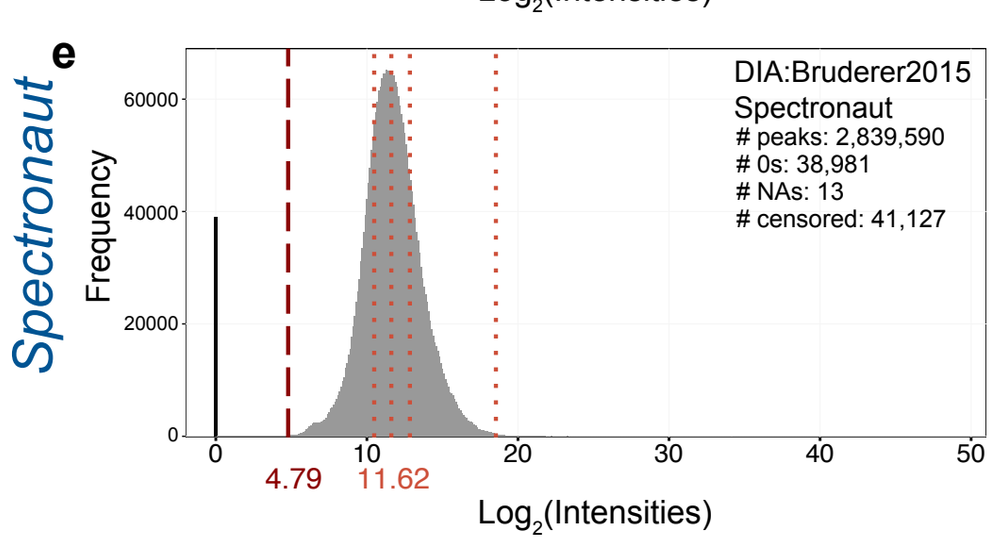
Progenesis

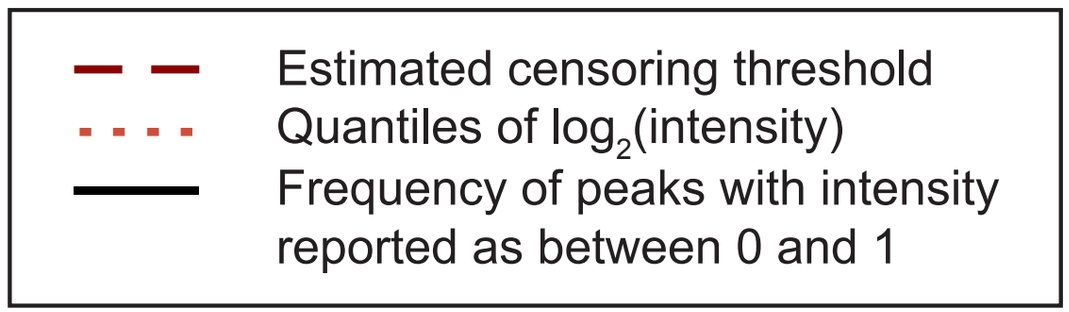


Skyline



Spectronaut





SCHEMATIC DATA REPRESENTATION

Repeat for every protein

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	**Cens**	y	**Cens**	**Cens**	y	. . .	**Cens**	y	y	. . .	**NA**	y	y	y	y	y	. . .	y	**Cens**	y

Spectral features: technological aspects of the experiment

Missing values

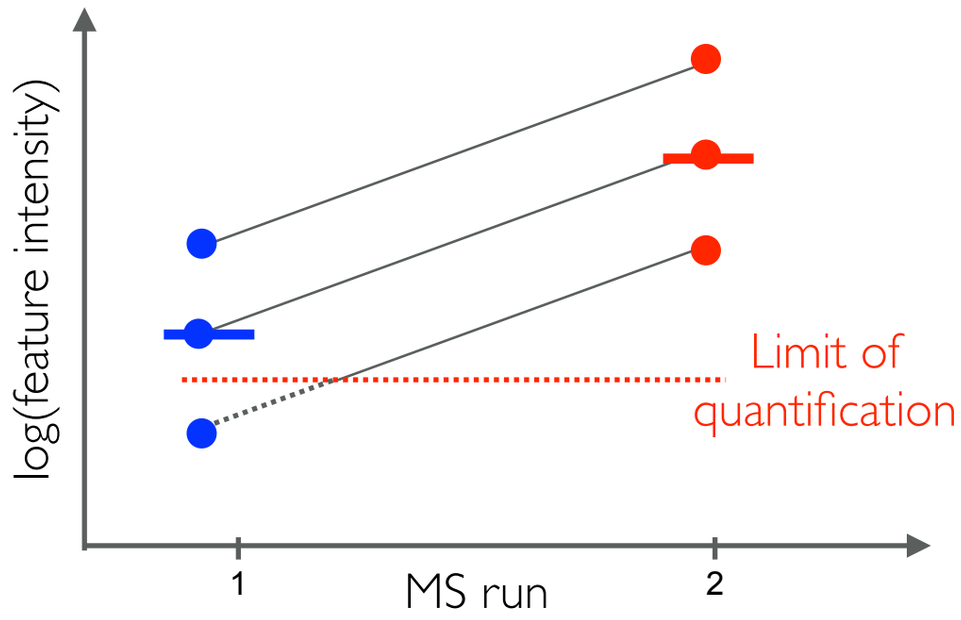
View feature intensities below the threshold as censored

Log(feature intensities)

Challenges

- Represent experimental design
- Represent peak intensities
- Estimate variance components & inference

IMPUTATION OF MISSING VALUES

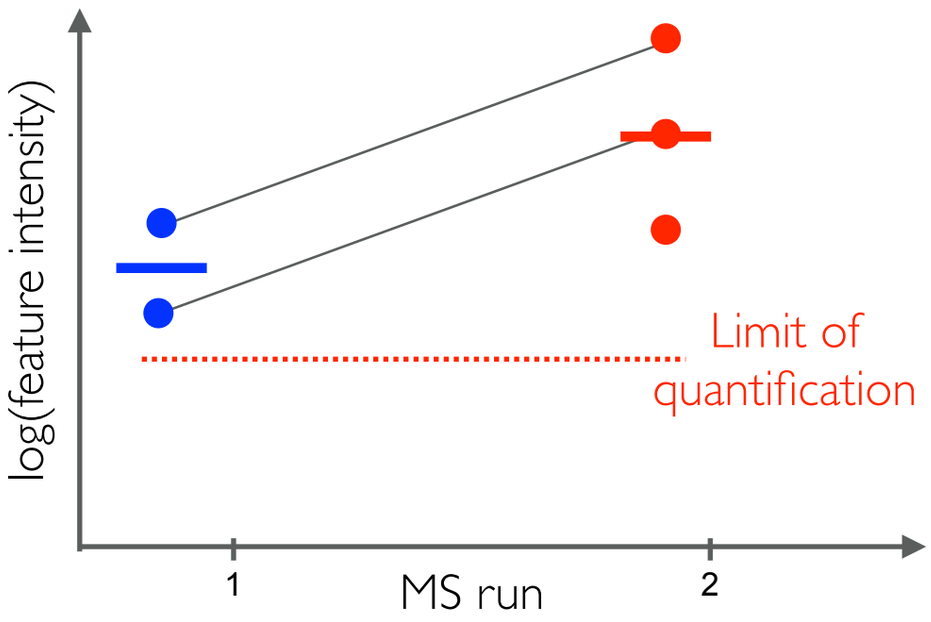


MSstats

Relies on information from other features and runs

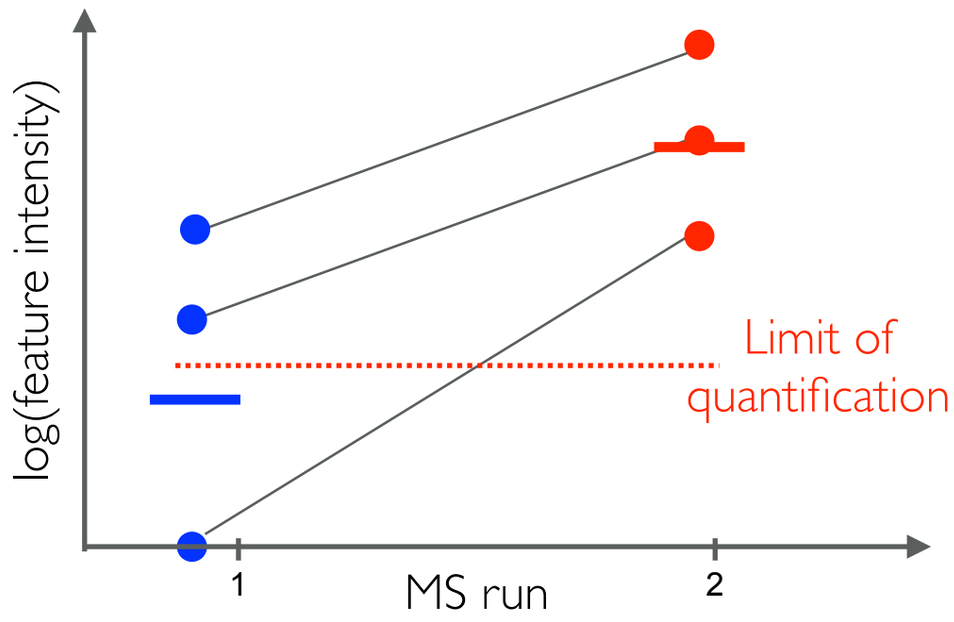
No imputation

Underestimates the regulation



Imputation with 0

Overestimates the regulation



MSSTATS: MODEL AND INFERENCE

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
$Feature_2$	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 accelated failure model

$$y_{ijkl} = \mu + Run_{ijk} + Feature_l + \epsilon_{ijkl} \text{ , where } \sum_{ijk} Run_{ijk} = 0,\ \sum_{l} Feature_l = 0, \epsilon_{ijkl} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_\epsilon)$$

	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

↓

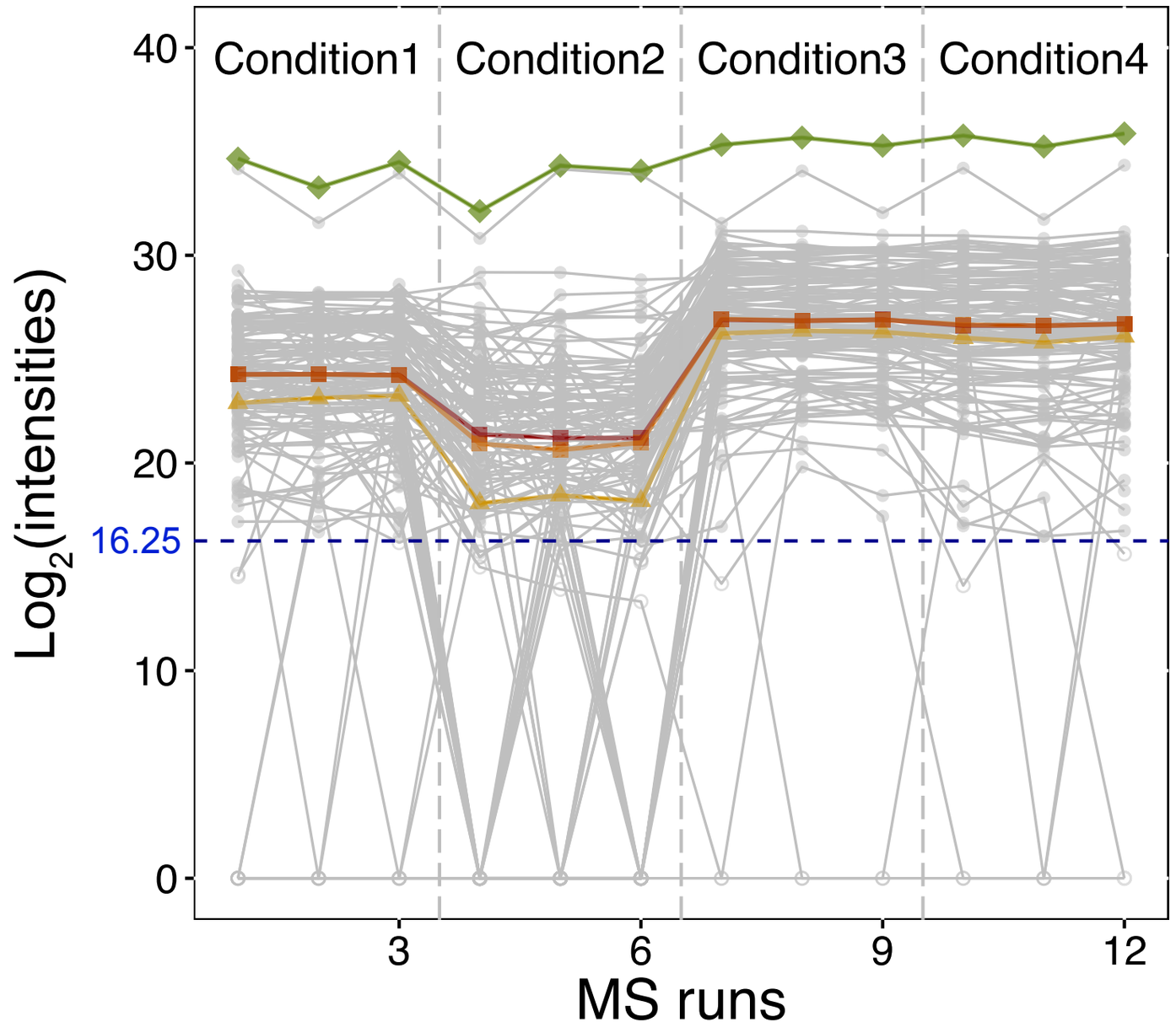
TMP : Parameter estimation by robust method

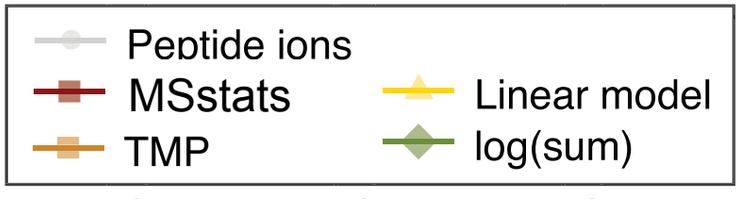
Whole plot model

	$Condition_1$										· · ·	$Condition_I$									
	$Subject_1$			$Subject_2$			· · ·	$Subject_J$			· · ·	$Subject_{(I-1)J+1}$			$Subject_{(I-1)+2}$			· · ·	$Subject_{IJ}$		
	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

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

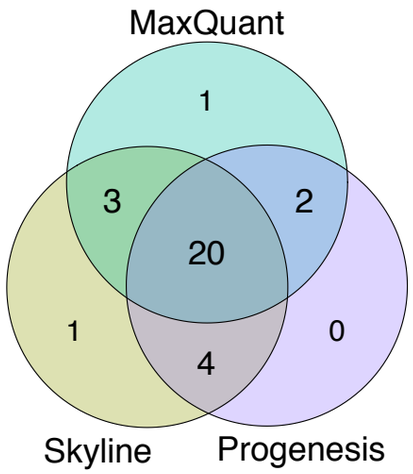
STATISTICAL METHODS ENHANCE REPRODUCIBLE RESEARCH

Better agreement in correct detection of differentially abundant proteins between tools

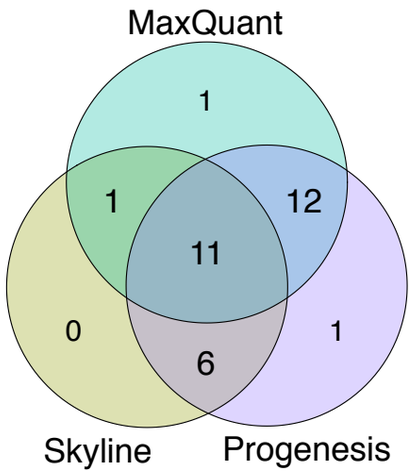
Meena Choi

DDA: iPRG2015

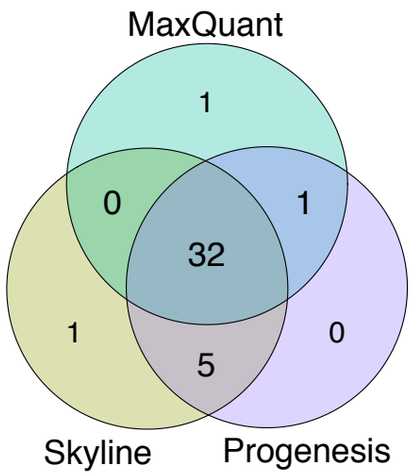
MSstats

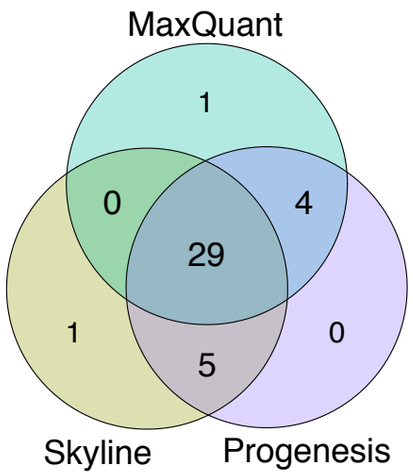


Log(sum)

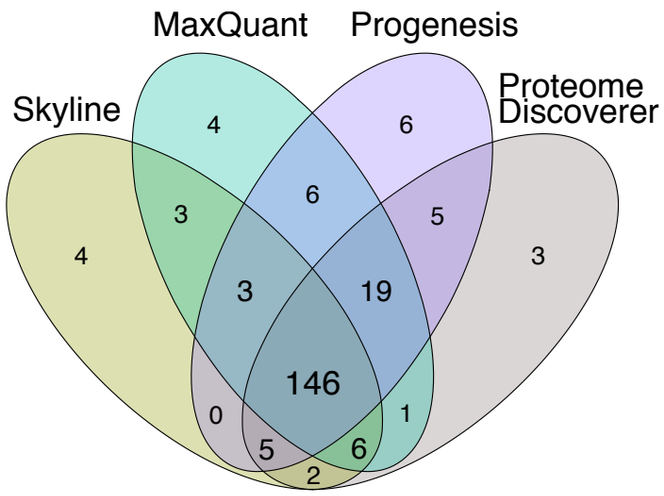


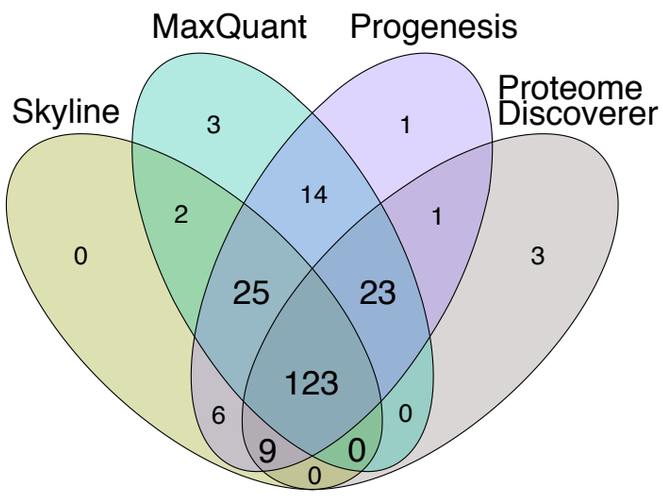
DDA: Cox 2014





DDA: Spike-in





INTRODUCTION TO MSSTATS

- Overview
- Statistical modeling
- Community resources

DATA PROCESSING IMPACTS STATISTICAL ANALYSIS

Publicly available data are key for method development and evaluation

https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp

MassIVE

Mass Spectrometry
Interactive **V**irtual **E**nvironment

MassIVE.quant Statistics

Public Datasets:	46
Quant Reanalyses:	188
Number of Datasets by Species:	
Homo sapiens	23
Mus musculus	12
Saccharomyces cerevisiae	5
Human immunodeficiency virus 1	4
Others	14

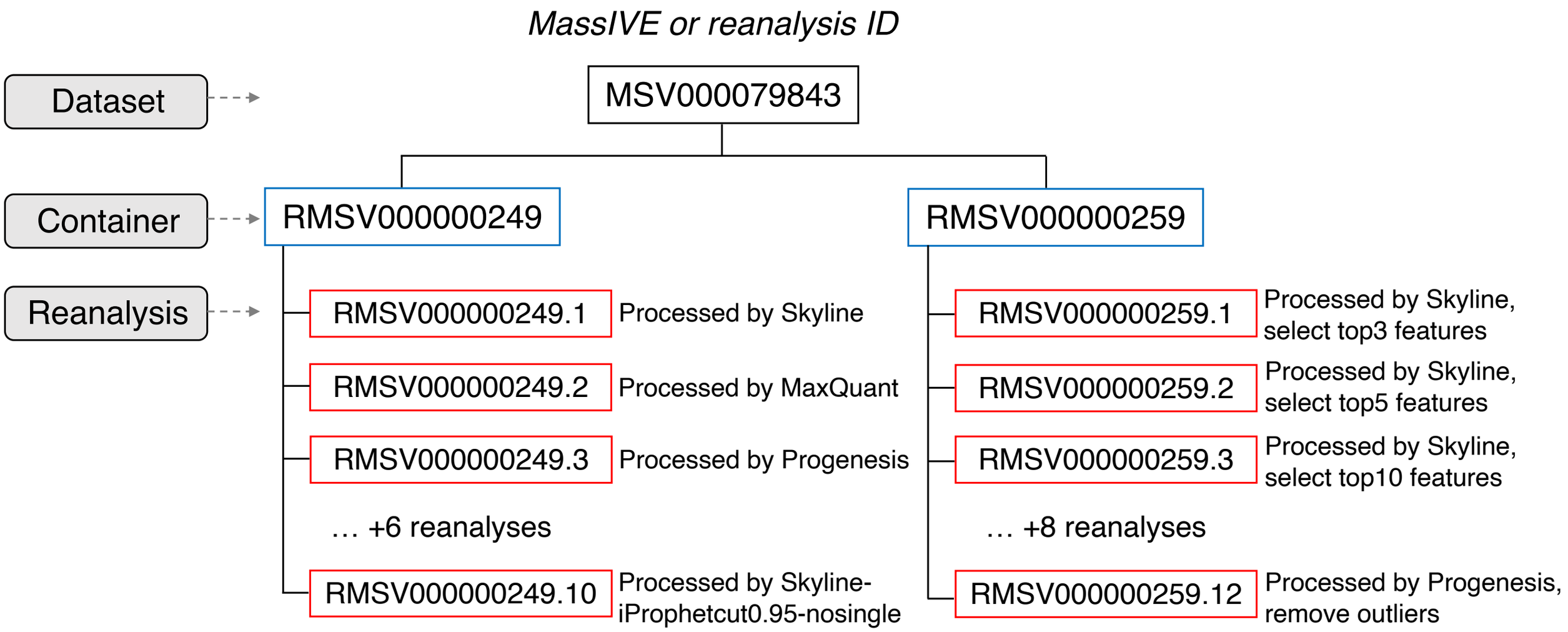
nature methods
BRIEF COMMUNICATION
https://doi.org/10.1038/s41592-020-0955-0

MassIVE.quant: a community resource of quantitative mass spectrometry–based proteomics datasets

Meena Choi[1], Jeremy Carver[2], Cristina Chiva[3,4], Manuel Tzouros[5], Ting Huang[1], Tsung-Heng Tsai[1], Benjamin Pullman[2], Oliver M. Bernhardt[6], Ruth Hüttenhain[7], Guo Ci Teo[8], Yasset Perez-Riverol[9], Jan Muntel[6], Maik Müller[10], Sandra Goetze[10,11], Maria Pavlou[10], Erik Verschueren[7], Bernd Wollscheid[10,11], Alexey I. Nesvizhskii[8], Lukas Reiter[6], Tom Dunkley[5], Eduard Sabidó[3,4], Nuno Bandeira[2]✉ and Olga Vitek[1]✉

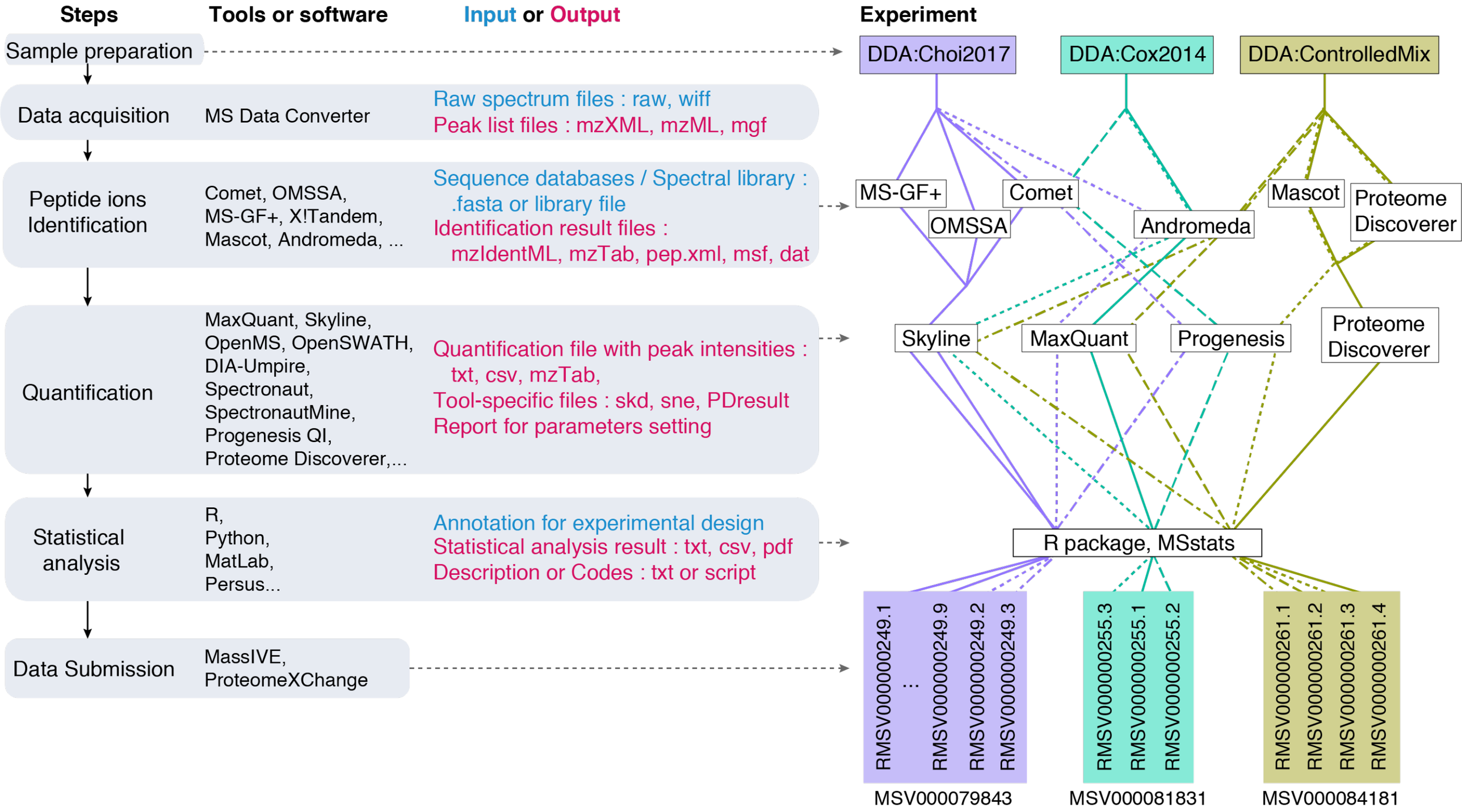
MASSIVE.QUANT

Repository for quantitative data and reanalysis



MASSIVE.QUANT

Repository for quantitative data and reanalysis



https://computationalproteomics.khoury.northeastern.edu/

May Institute goes ONLINE!
Computation and statistics for mass spectrometry and proteomics

April 27 – May 8, 2020, Northeastern University, Boston MA
Organizers : Meena Choi and Olga Vitek

Week 1:

April 27, 11:00am-1:00pm	**Alicia Williams:** Scientific writing workshop (live)-registration closed	Docs Video
April 28, 11:00am-12:30pm	**Lindsay Pino:** Targeted analysis with Skyline, a PRM perspective (live)-registration closed	Docs Video
April 28, 12:45pm-2:15pm	**Brian Searle:** Introduction to DIA (live)-registration closed	Docs Video
April 29, 11:00am-12:30pm	**Bernhard Kuster:** Large-scale proteomics with TMT (live)-registration closed	Docs Video
April 29, 12:45pm-2:15pm	**Ruedi Aebersold:** Modern technology for modern biology (live)-registration closed	Docs Video
April 30, 11:00am-1:00pm	**Brendan MacLean:** Skyline for DIA: Intro with a controlled mixture and exploring Prosit spectrum and retention time prediction (video+live)-registration closed	Docs Video
May 1, 11:00am-1:00pm	**Brendan MacLean:** Comparing DIA with a deep fractionation library versus a refined subset in Skyline (video+live)-registration closed	Docs Video


Sue Abbatiello

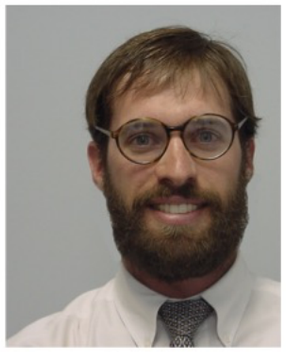

Ruedi Aebersold


Nuno Bandeira


Kylie Bemis


Meena Choi


Laurent Gatto


Andy Hoofnagle


Oliver Kohlbacher


Bernhard Kuster


Brendan MacLean


Brian Searle


Olga Vitek

ACKNOWLEDGEMENTS

Northeastern University

Kylie Bemis
Meena Choi
Eralp Dogu
Dan Guo
April Harry
Ting Huang
Cyril Galitzine
Devon Kohler
Akshay Kulkarni
Sai Lakkimsetty
Robert Ness
Sumedh Sankhe
Sara Taheri
Tsung-Heng Tsai

University of Wroclaw

Mateusz Staniak
Malgorzata Bogdan

ETH Zurich

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

University of Washington

Michael MacCoss
Brendan MacLean
Jarrett Egertson

Mugla University

Eralp Dogu

Hoffman LaRoche

Tom Dunkley
Balazs Banfai

Genentech

Don Kirkpatrick
Erik Verschueren

UCSD

Nuno Bandeira
Jeremy Carver

Support:

NSF
NIH
Sternberg Chair
Canary Center
Roche
Genentech
Eli Lilly
Chan Zuckerberg foundation

Thank you to the instructors and to the teaching assistants!

Ryan Benz
Meena Choi
Niyati Chopra
Miguel Cosenza
Matthias Fahrner
Amanda Figueroa-Navedo
Melanie Foell
Omkar Reddy Gojala
Dan Guo
Shubhanshu Gupta
Ting Huang
Maanasa Kaza
Smit Anish Kiri
Devon Kohler
Sai Srikanth Lakkimsetty
Danielle LaMay
Ajeya Makanahalli Kempegowda
Yogesh Nizzer
Harish Ramani
Ruthvik Ravindra
Abdul Rehman
Sai Divya Sangeetha Bhagavatula
Siddarth Sathyanarayanan
Gopalika Shama
Rishabh Rajesh Shanbhag
Sagar Singh
Mateusz Staniak
Sara Taheri
Anuska Tak
Derrie Susan Varghese
Amrutha Vempati

Video of the presentation: https://www.youtube.com/channel/UCnbUMFlIRLaY7fwfSintWuQ/