MSSTATS

Meena Choi, Olga Vitek

College of Science

College of Computer and Information Science

Northeastern University

OUTLINE

- iPRG2015: detection of diff. abundance
  - Community effort with label-free shotgun proteomics
- MSstats
  - Normalization, statistical modeling, inference
  - Evaluation
- Extensions to MSstats
  - Assay characterization, longitudinal monitoring

ABRF IPRG STUDY 2015

Detection of differentially abundant proteins in controlled mixture

				Samples			
	Name	Origin	Molecular Weight	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

Choi et al., “ABRF Proteome Informatics Research Group (iPRG) 2015 Study: Detection of differentially abundant proteins in label-free quantitative LC-MS/MS experiments”, *Journal of Proteome Research,* 2017.

EXPERIMENTAL PROCEDURES

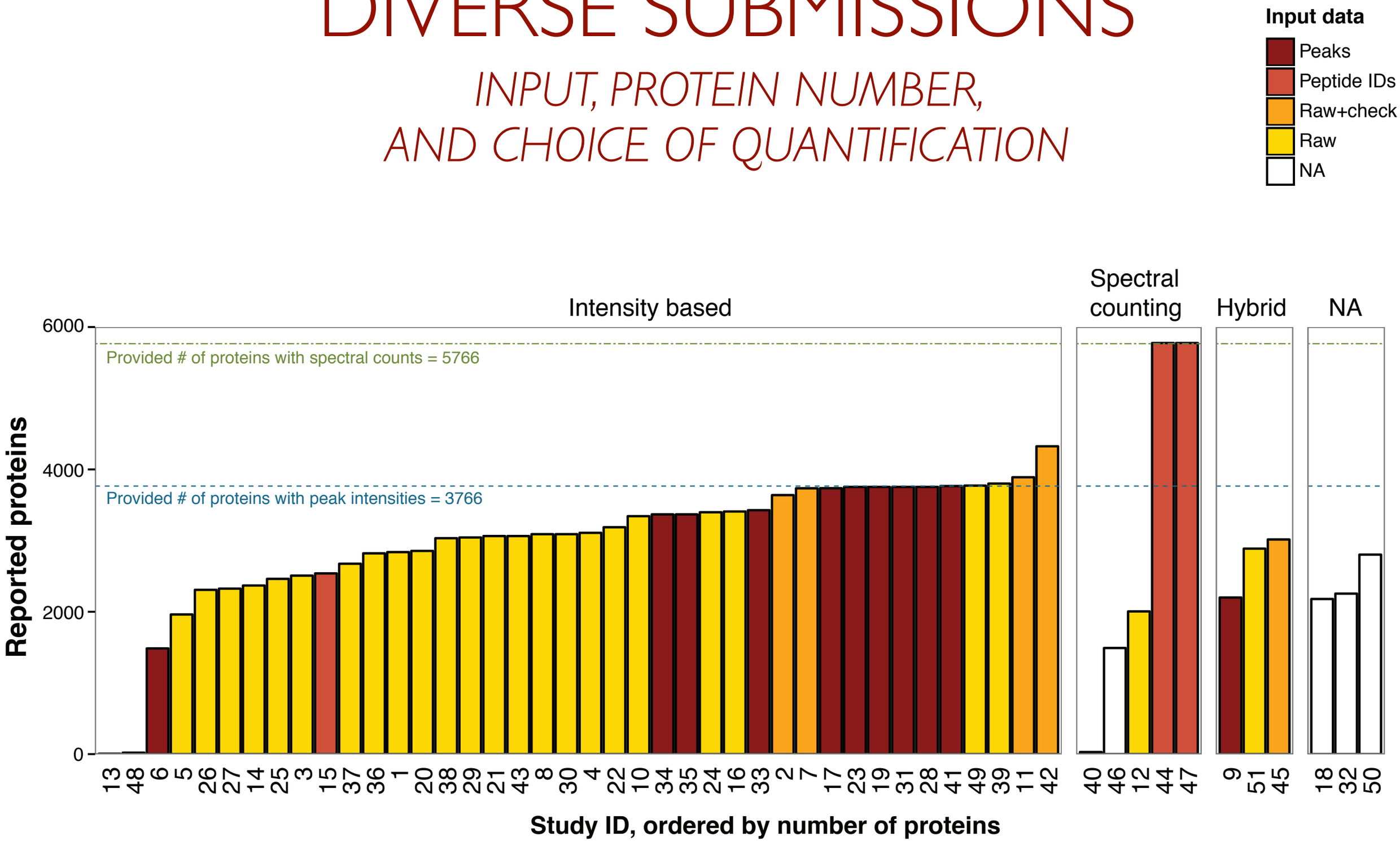
- **Background**
  - 200ng of tryptic digests of *S. cerevisiae*
- **Spectral acquisition**
  - *Three technical replicates per sample*
  - *Randomized order*
  - Separation:
    - Thermo nLC 1000 system
    - 110-min linear gradient
  - Spectral acquisition:
    - *DDA profile mode in Orbitrap*
    - Resolution 70,000 for MS and 17,500 for MS/MS
    - MS1 scan range 300-1650 m/z
    - Normalized collision energy 27%
    - Singly charged ions excluded

SPECTRAL PROCESSING

- **MS/MS identification**
  - Database search: OMSSA7, MS-GF+, Comet
  - Q-value: target-decoy.
    - No filtering!
  - 5,766 proteins, ~26,242 PSMs/run
  - 48% of proteins had 1 or 2 PSMs
  - 29% of proteins had 15 or more PSMs
- **MS1 quantification with Skyline**
  - Original processing
    - 3,766 proteins, 29,575 features
    - median 5 features/protein
  - Post-study processing
    - 3,027 proteins, 34,783 features
    - median 7 features/protein

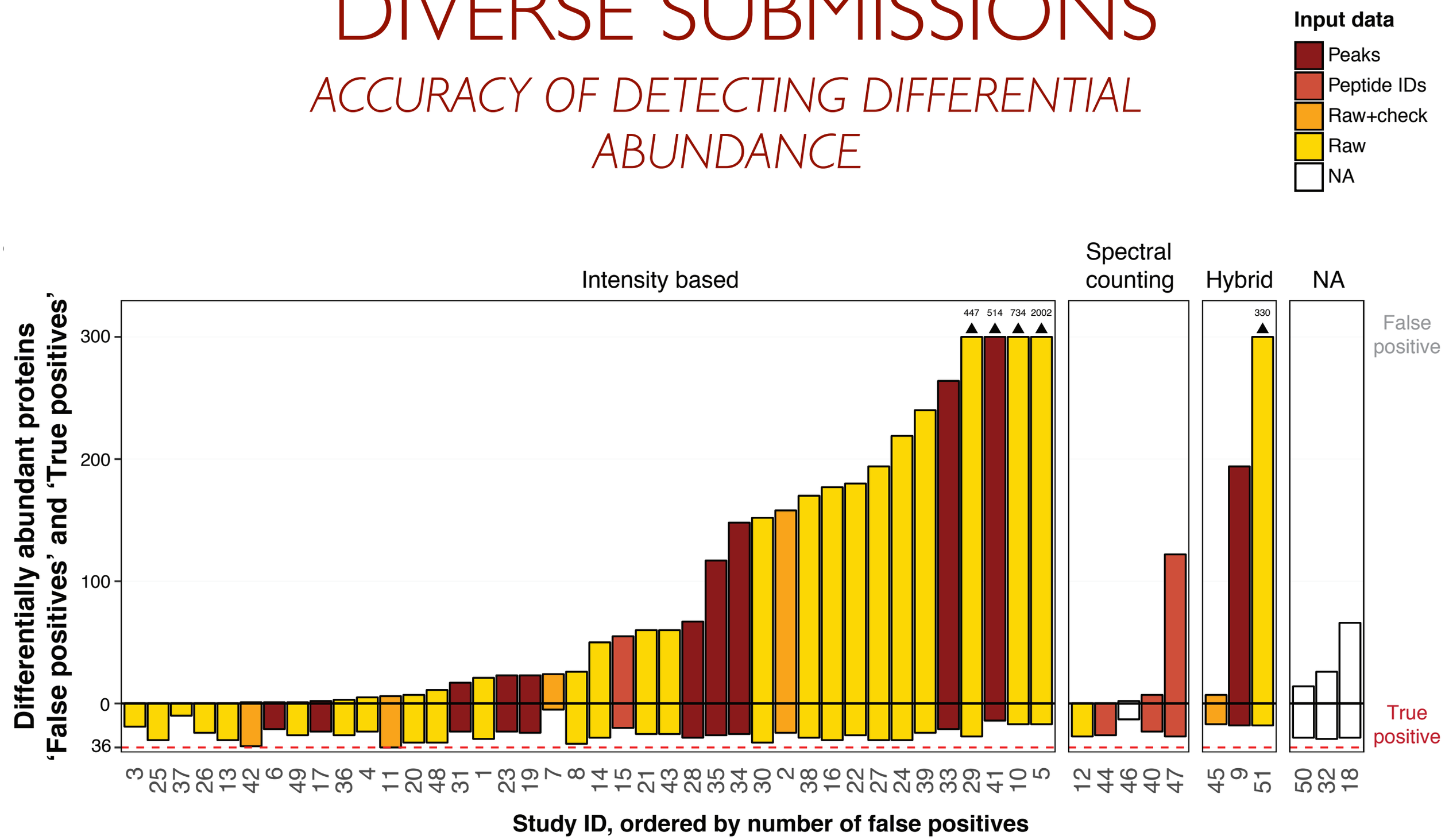
DIVERSE SUBMISSIONS

INPUT, PROTEIN NUMBER, AND CHOICE OF QUANTIFICATION



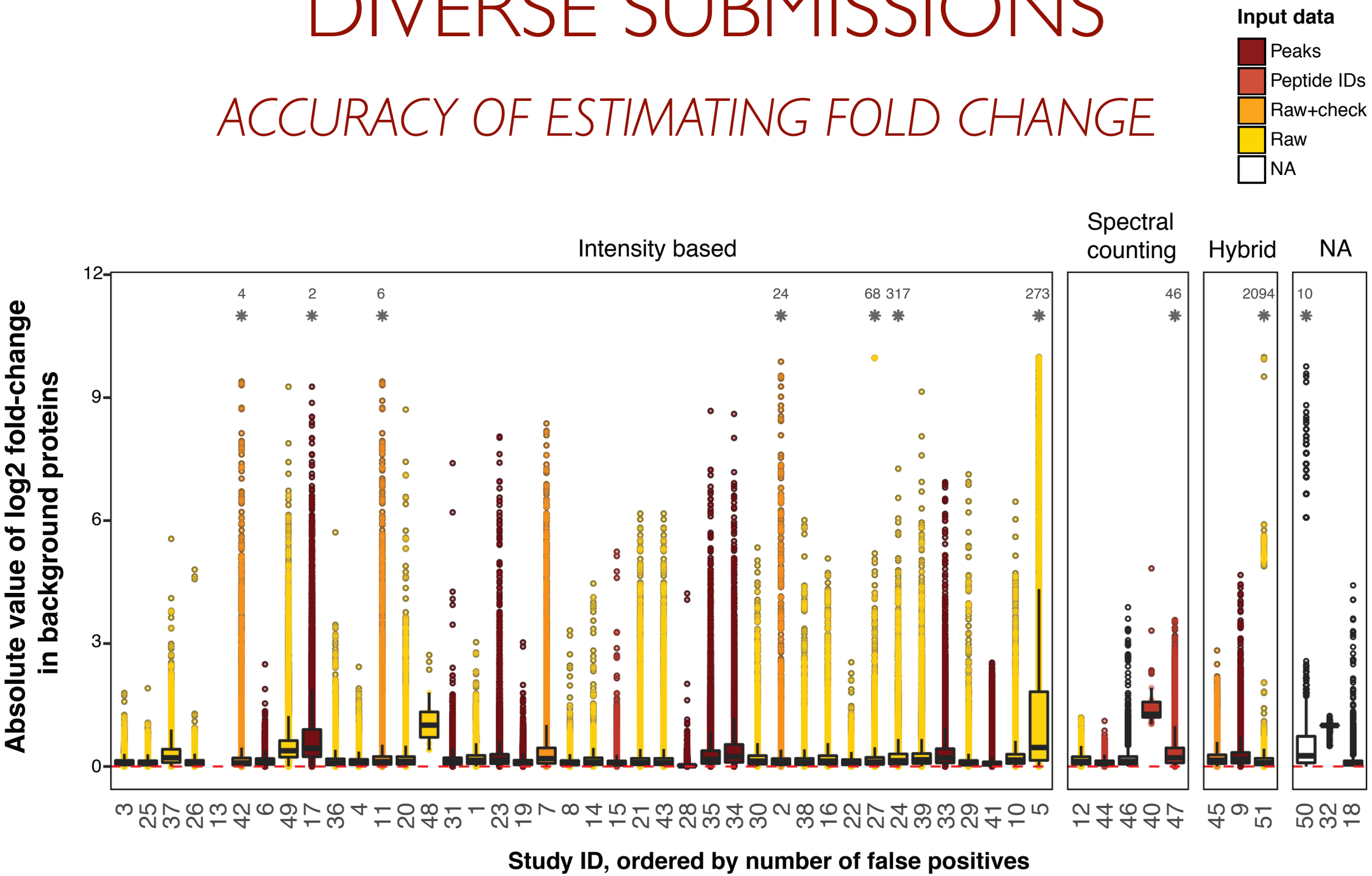
DIVERSE SUBMISSIONS

ACCURACY OF DETECTING DIFFERENTIAL ABUNDANCE



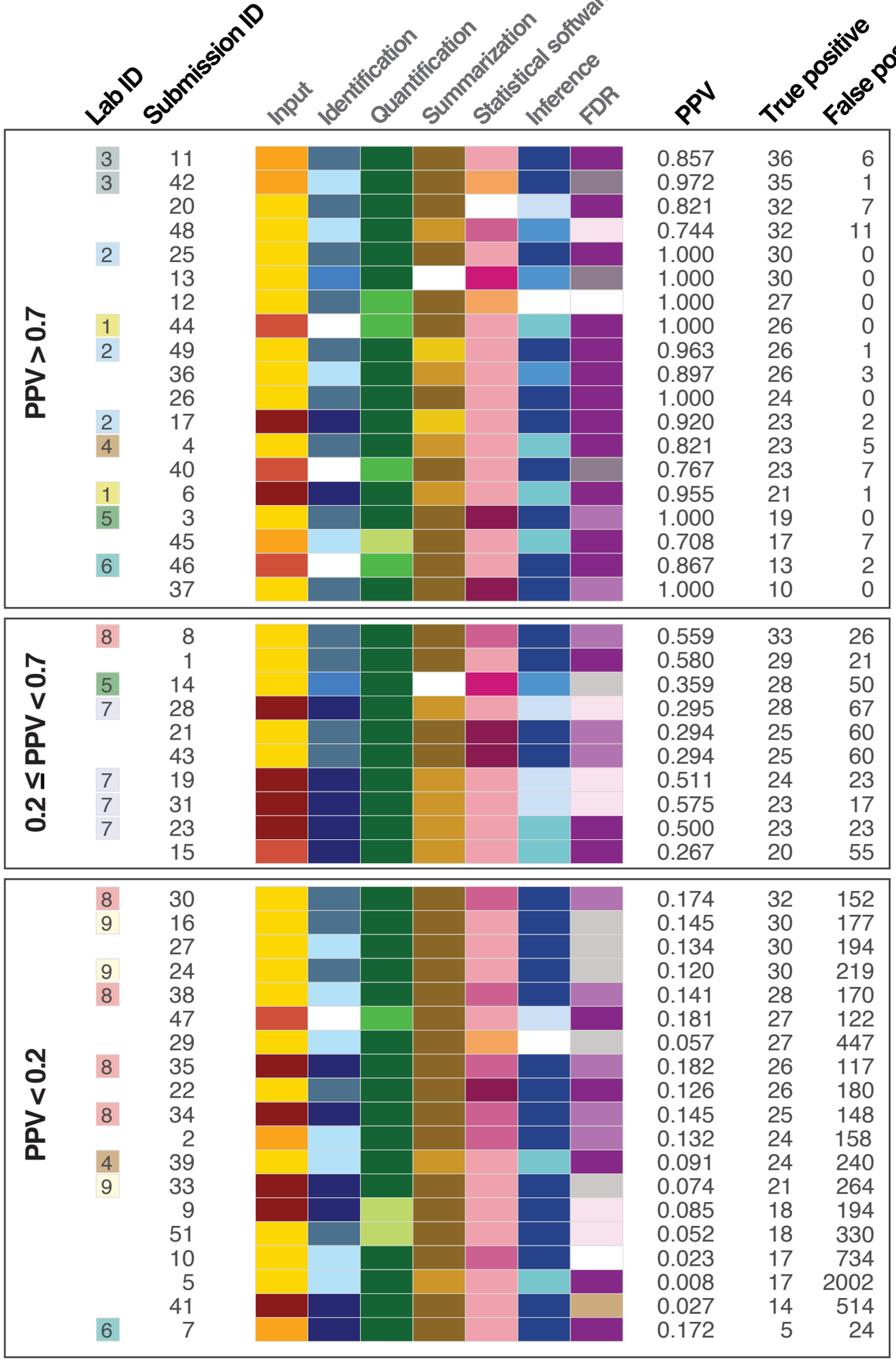
DIVERSE SUBMISSIONS

ACCURACY OF ESTIMATING FOLD CHANGE



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

	Lab ID	Submission ID	Input	Identification	Quantification	Summarization	Statistical softwar	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
		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
		41								0.027	14	514
	6	7								0.172	5	24

Positive predictive value =

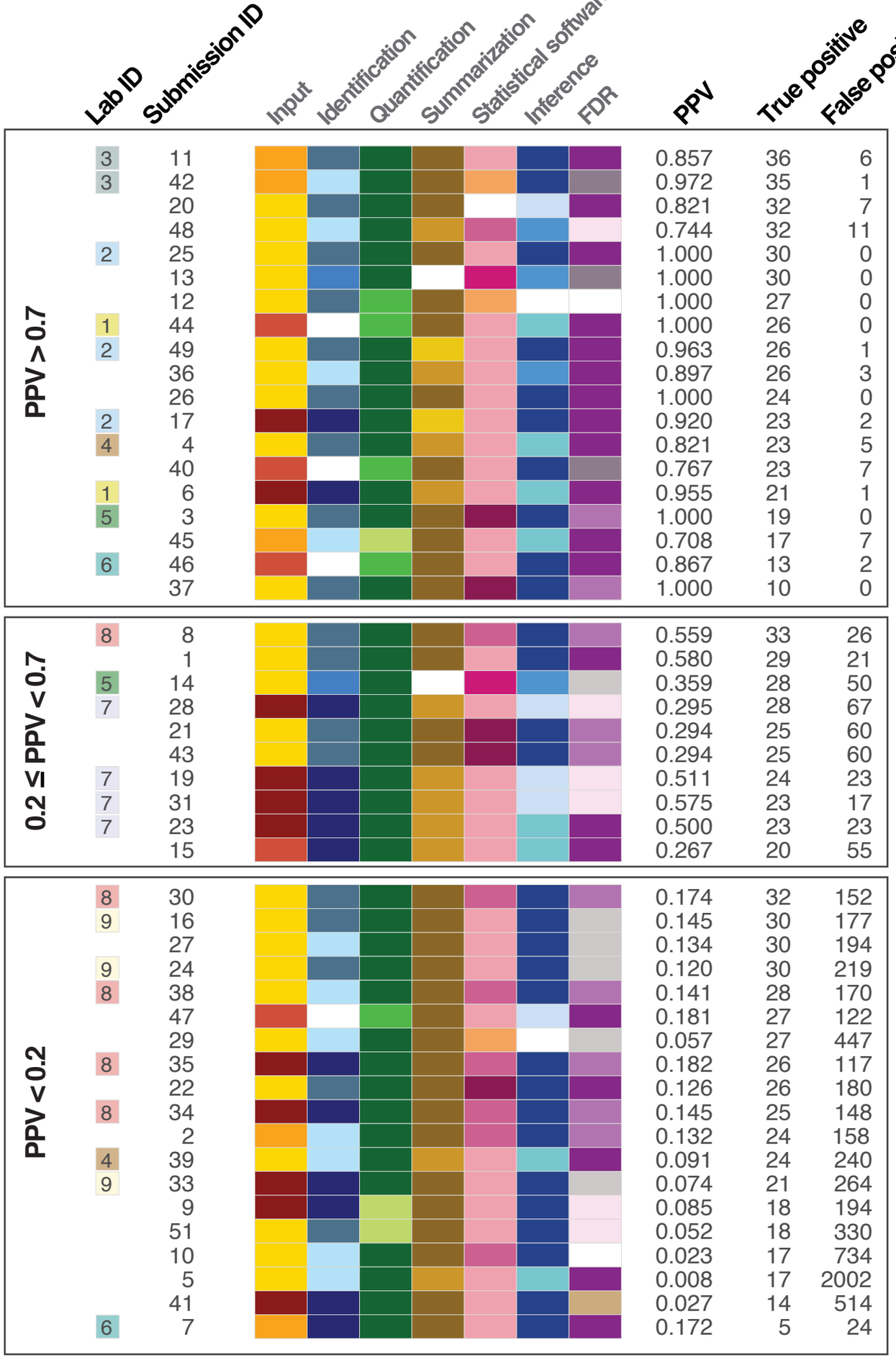
$$\frac{\text{\# true differentially abundant proteins}}{\text{\# claimed differentially abundant proteins}}$$

Good

Bad

Very bad

USER EXPERTISE IS KEY



MaxQuant and Perseus

USER EXPERTISE IS KEY

	Lab ID	Submission ID	Input	Identification	Quantification	Summarization	Statistical softwar	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
		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
		41								0.027	14	514
	6	7								0.172	5	24

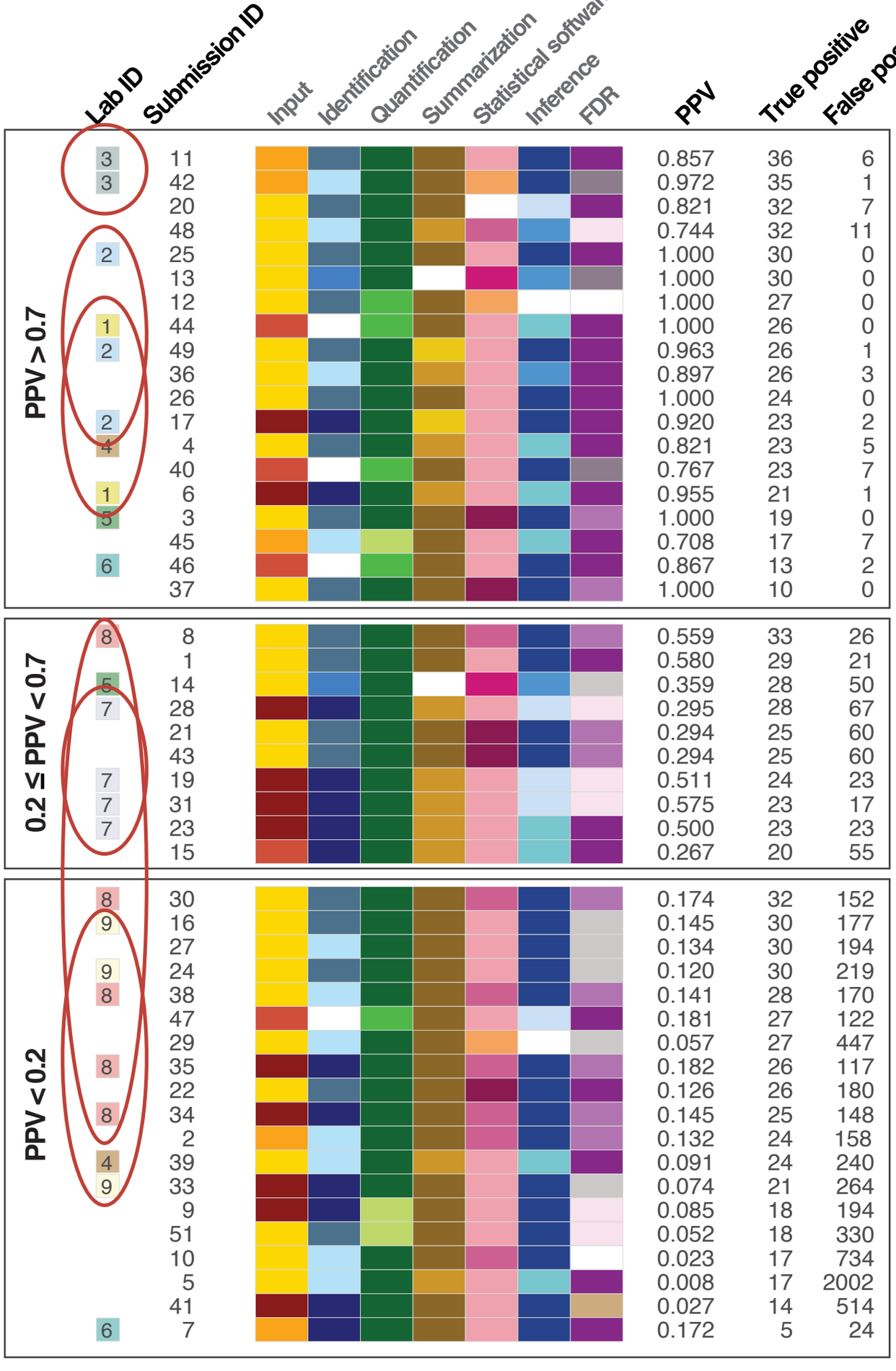
Skyline and linear modeling in R

USER EXPERTISE IS KEY

	Lab ID	Submission ID	Input	Identification	Quantification	Summarization	Statistical softwar	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
		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
		41								0.027	14	514
	6	7								0.172	5	24

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

Journal of proteome research

Search | Citation | Subject | Advanced Search
Enter search text / DOI | Anywhere | Search
J. Proteome Res. | All Publications/Website

Browse the Journal | Articles ASAP | Current Issue | Submission & Review | Open Access | About the Journal

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-Dogu[#‡], Christopher Colangelo[§], John Cottrell[ǁ], 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 Sitki Kocman University, 48000 Mugla, Turkey
[§] Primary Ion, LLC, Old Lyme, Connecticut 06371, United States
[ǁ] 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

*E-mail: o.vitek@neu.edu. Tel: 617-373-2194.

Article Options

ACS ActiveView PDF
Hi-Res Print, Annotate, Reference QuickView
PDF *(3135 KB)*
PDF w/ Links *(886 KB)*
Full Text HTML

Abstract
Supporting Info
Figures
References

Add to ACS ChemWorx

Add to Favorites
Download Citation
Email a Colleague
Order Reprints
Rights & Permissions

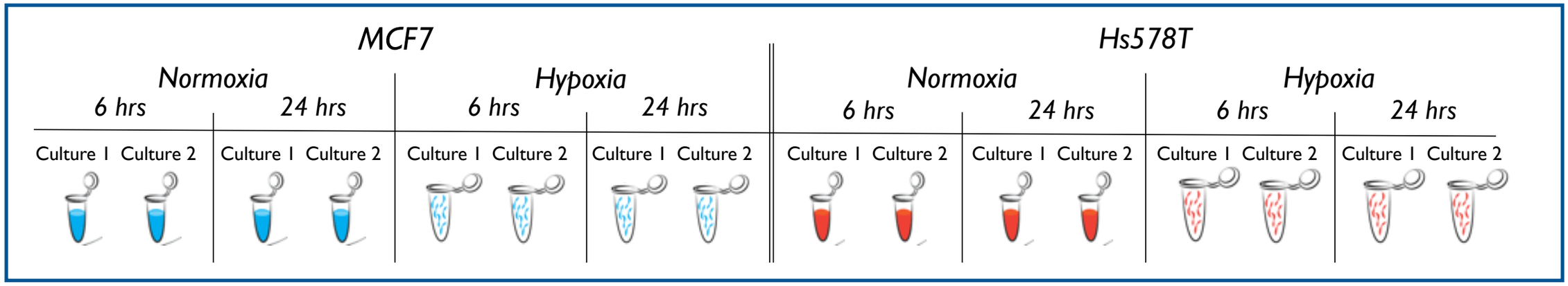
https://github.com/MeenaChoi/ABRF2017_satellite_workshop

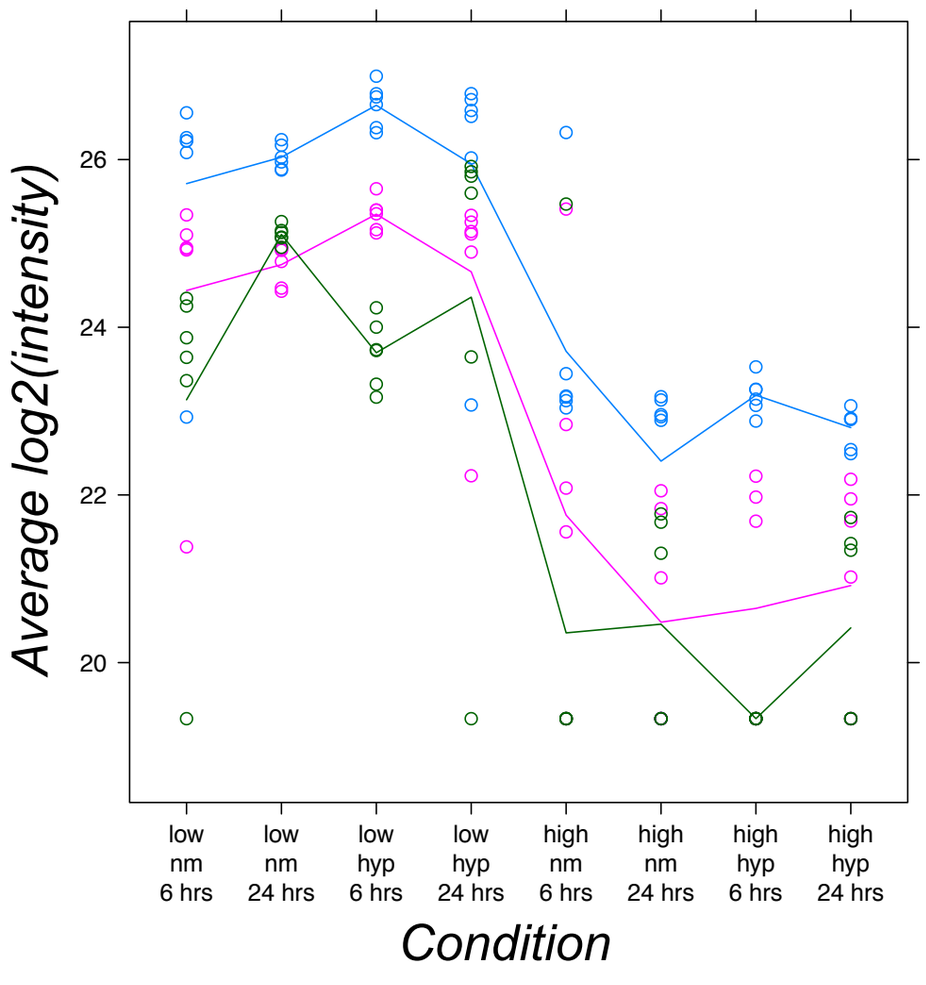
OUTLINE

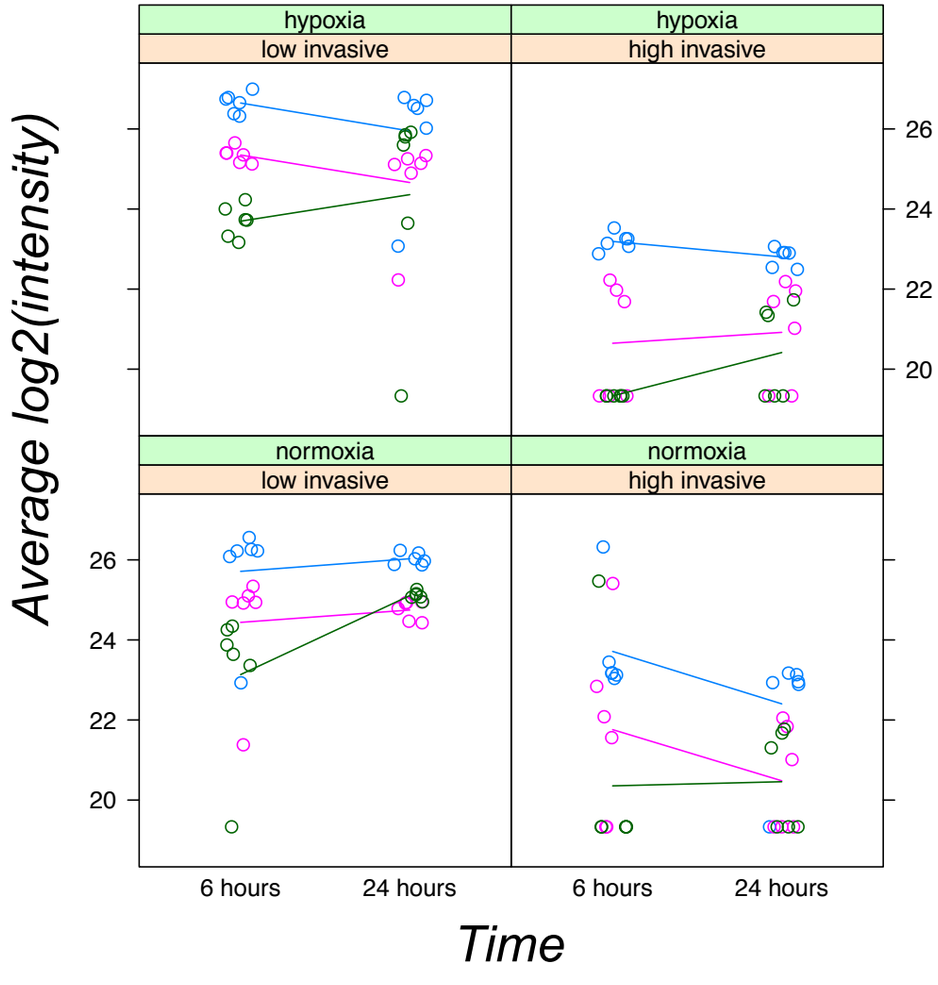
- iPRG2015: detection of diff. abundance
  - Community effort with label-free shotgun proteomics
- MSstats
  - Normalization, statistical modeling, inference
  - Evaluation
- Extensions to MSstats
  - Assay characterization, longitudinal monitoring

EXAMPLE: A LABEL-FREE EXPERIMENT

Question: which proteins change in abundance?

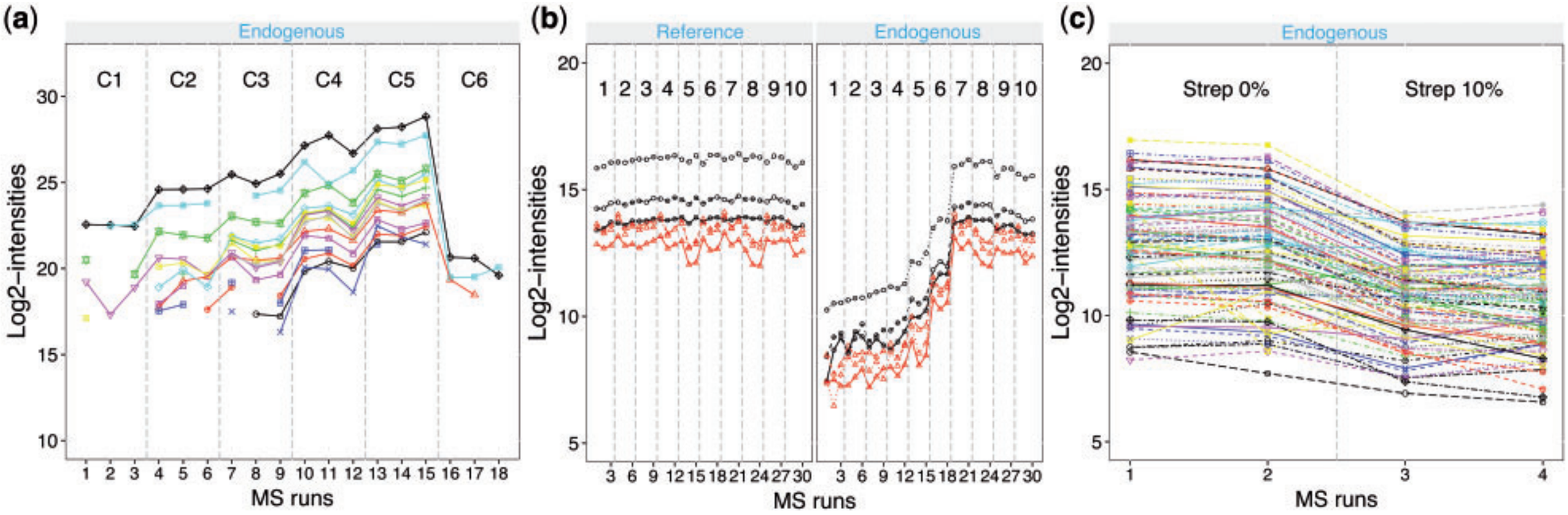






T. Clough et al. *BMC Bioinformatics., 2012*

CHROMATOGRAPHY-BASED QUANTIFICATION YIELDS DATA WITH VARYING PROPERTIES



Data-dependent acquisition (DDA)

Selected reaction monitoring with labeled reference peptides (SRM)

Data-independent acquisition (DIA)

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
- 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
  - External tool for Skyline, converter from MaxQuant

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$_1$	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

Spectral features: technological aspects of the experiment

Missing values

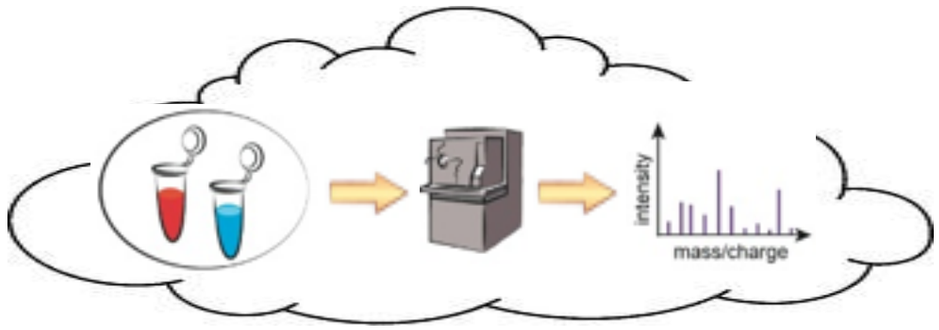
Log(feature intensities)

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

A TYPICAL ANALYSIS WORKFLOW

Experimental design



QC and normalization

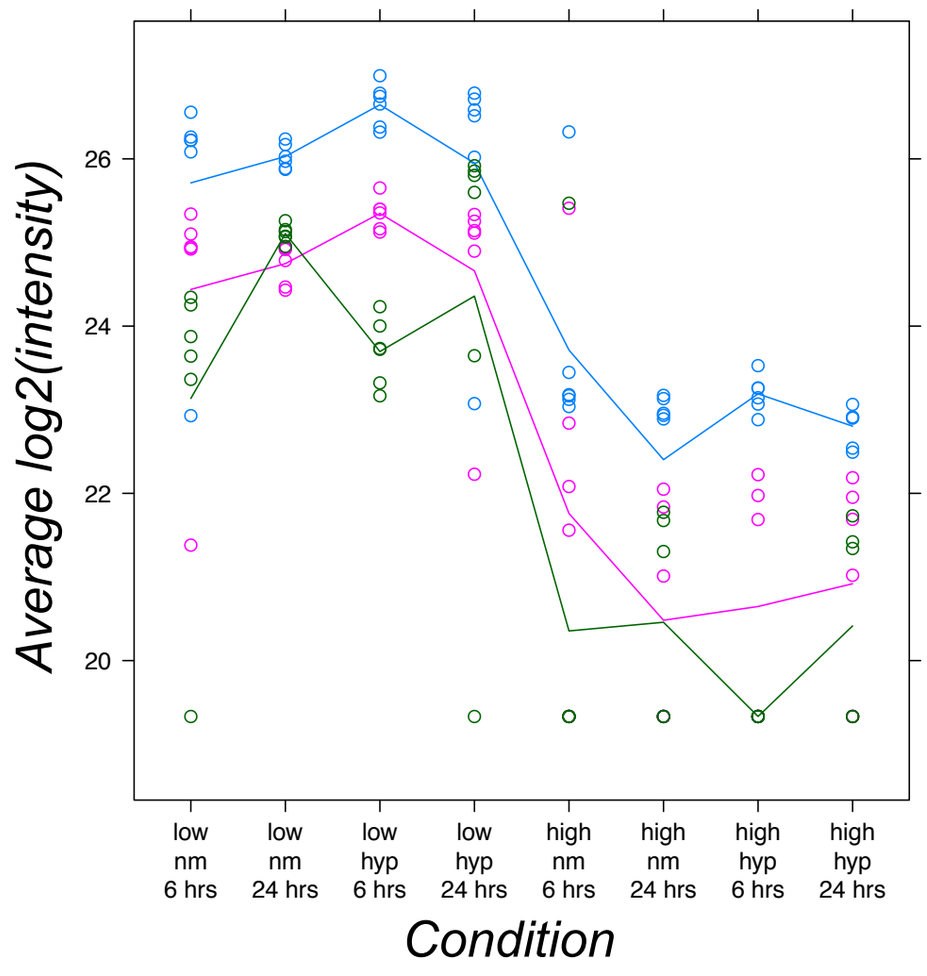
Statistical modeling

Model-based conclusions

Experimental design

Stresses or conditions

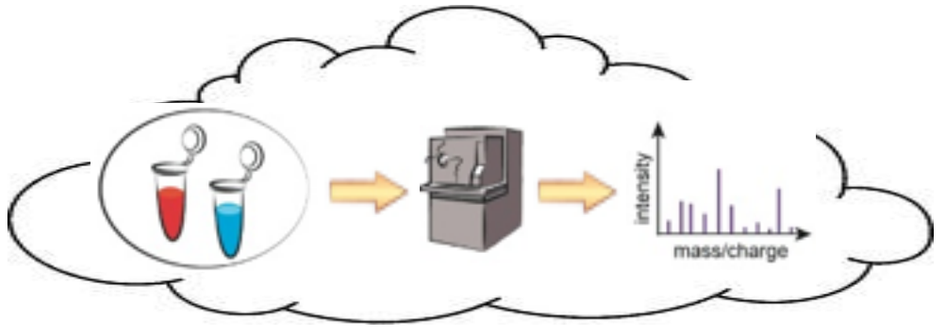
- Time course
- Multiple groups
- Paired design



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

A TYPICAL ANALYSIS WORKFLOW

Experimental design



QC and normalization

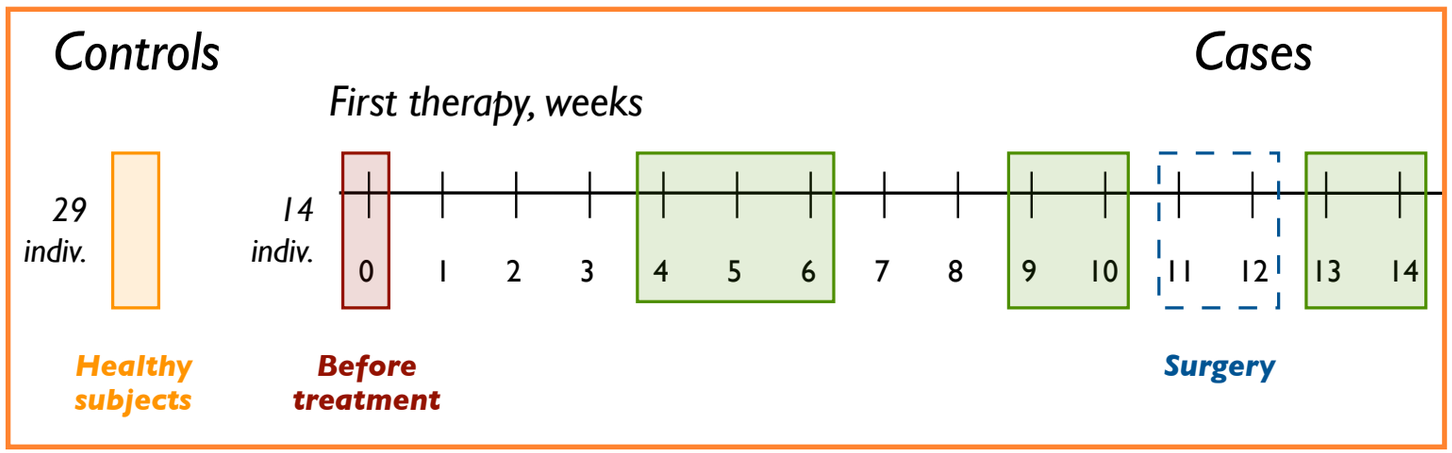
Statistical modeling

Model-based conclusions

Experimental design

Stresses or conditions

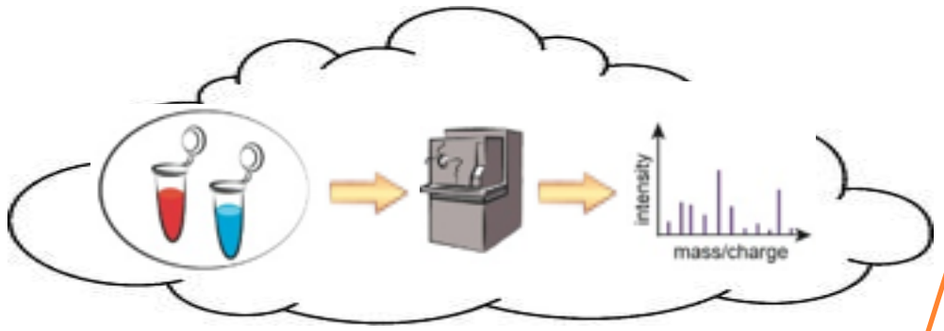
- Time course
- Multiple groups
- Paired design



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

A TYPICAL ANALYSIS WORKFLOW

Experimental design



QC and normalization

Statistical modeling

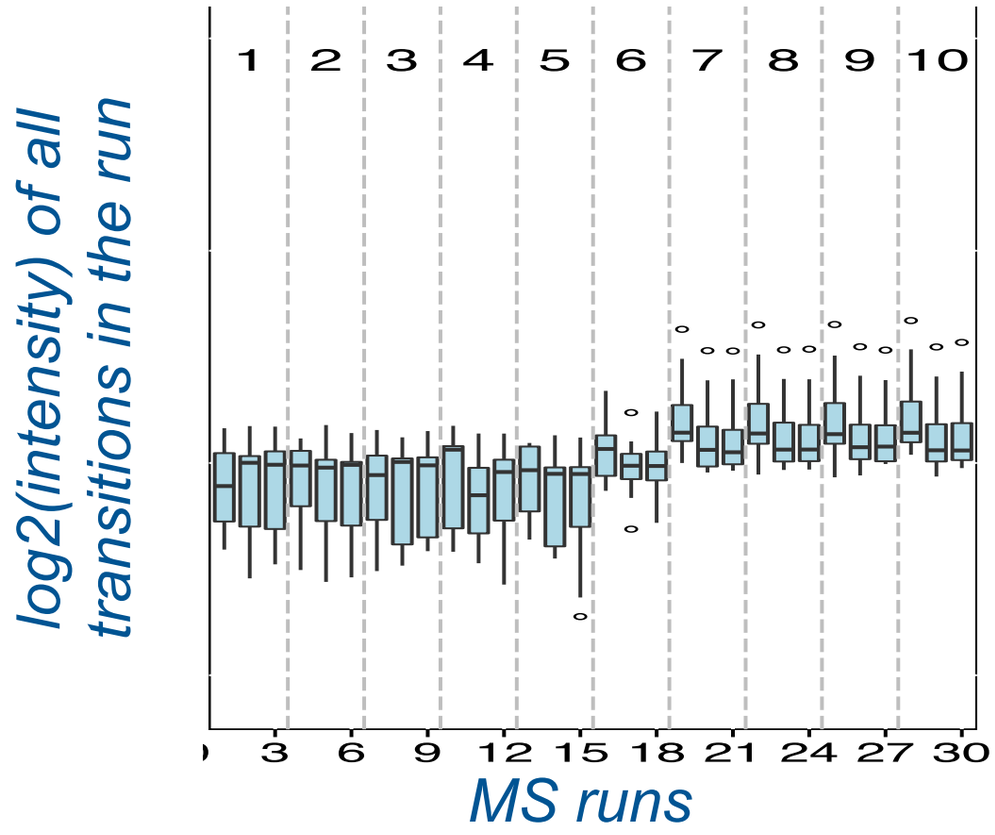
Model-based conclusions

Experimental design

Data preparation

- Formatting
- Visualization
- Remove artifacts
- Normalization

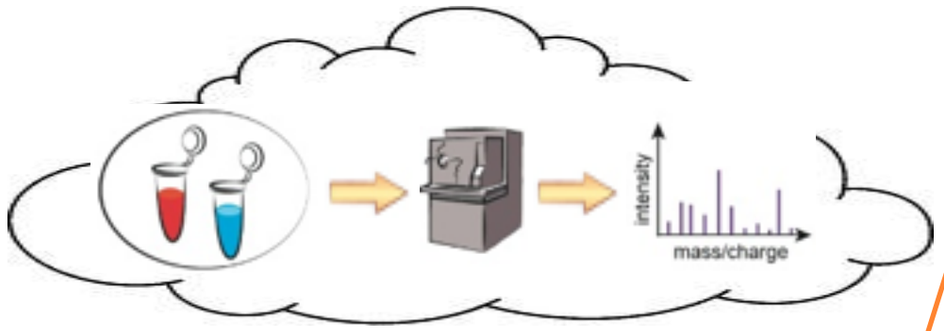
Endogenous



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

A TYPICAL ANALYSIS WORKFLOW

Experimental design



QC and normalization

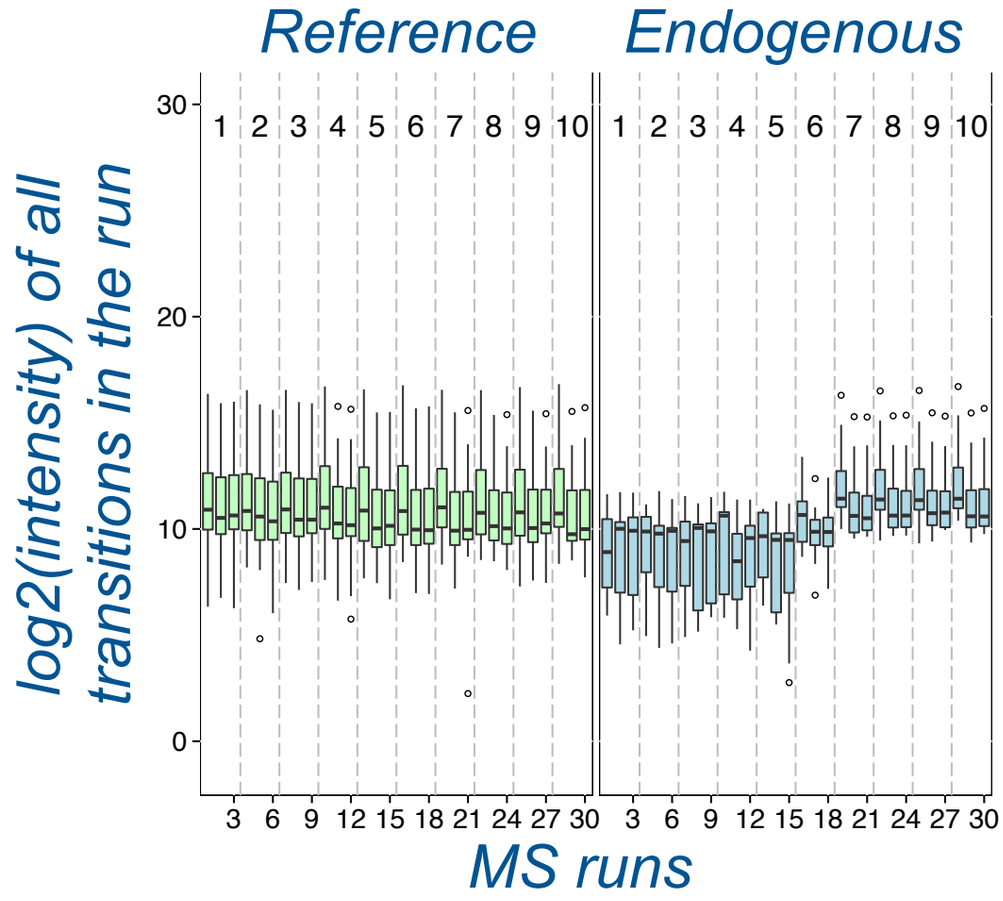
Statistical modeling

Model-based conclusions

Experimental design

Data preparation

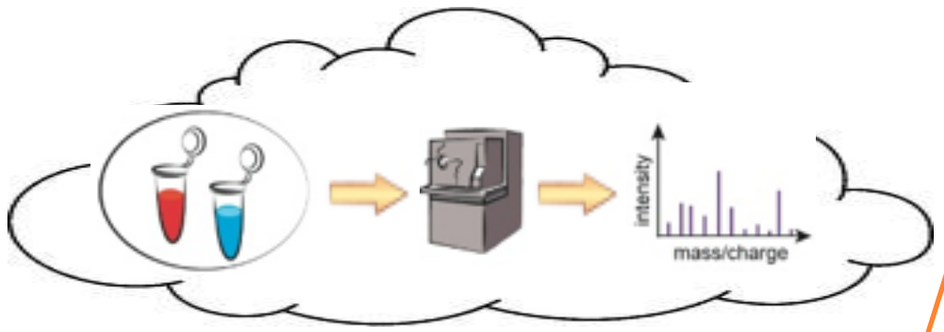
- Formatting
- Visualization
- Remove artifacts
- Normalization



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

A TYPICAL ANALYSIS WORKFLOW

Experimental design



QC and normalization

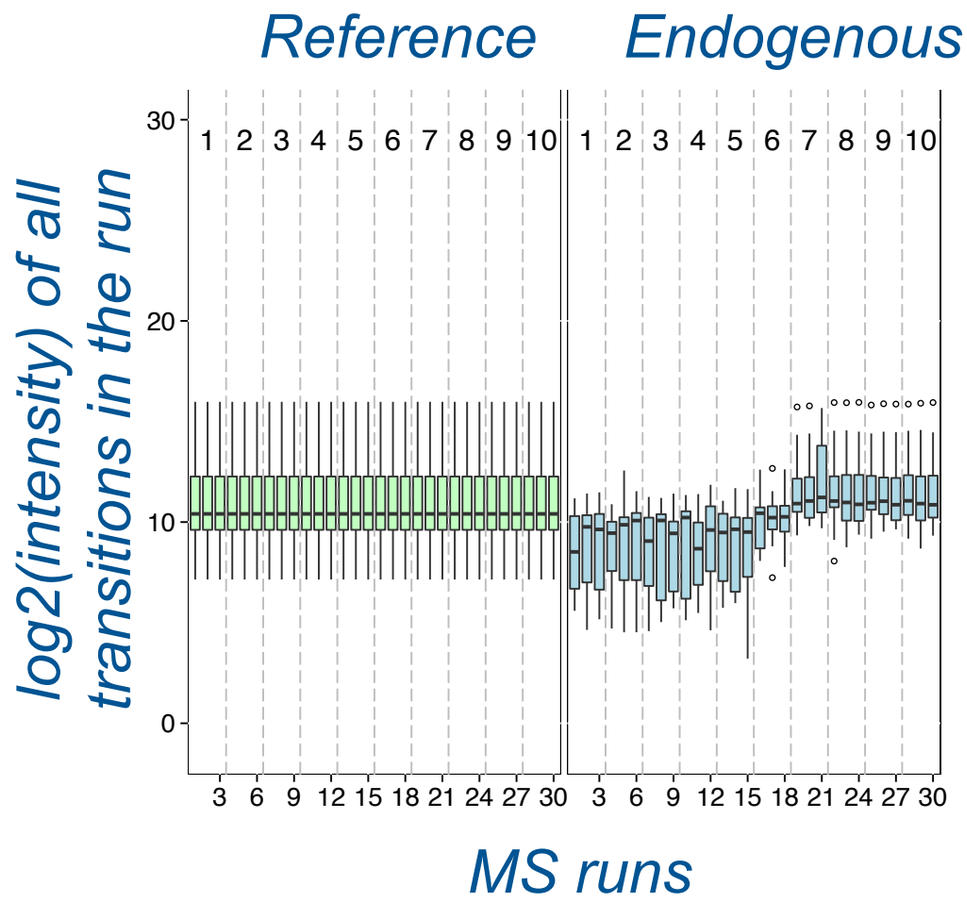
Statistical modeling

Model-based conclusions

Experimental design

Data preparation

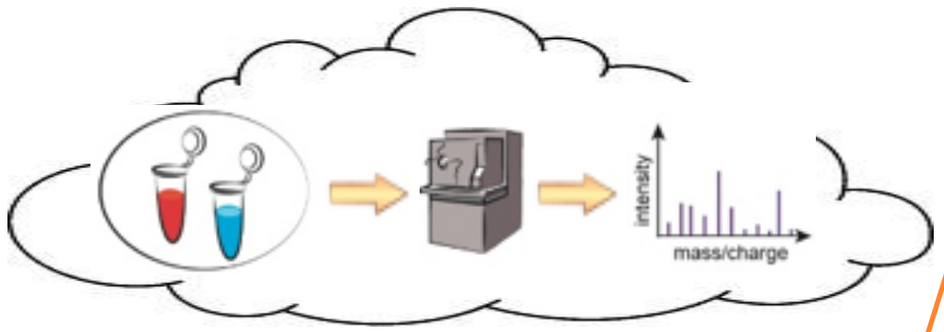
- Formatting
- Visualization
- Remove artifacts
- Normalization



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

A TYPICAL ANALYSIS WORKFLOW

Experimental design

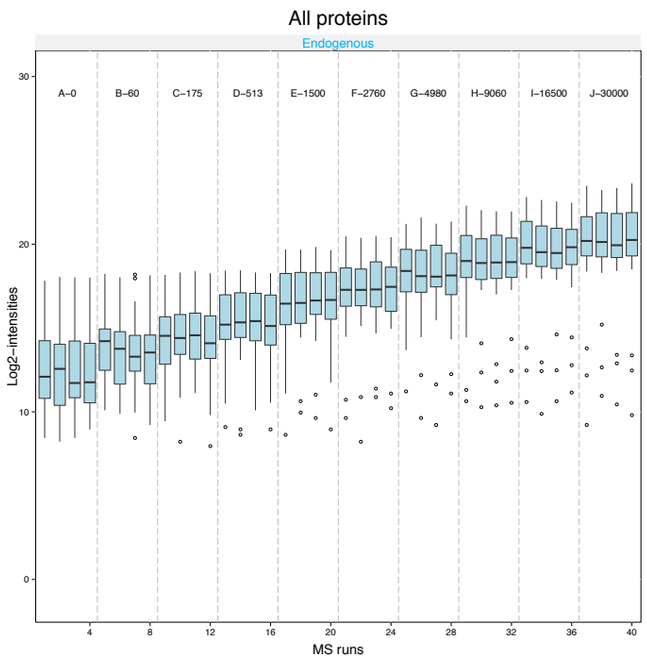


QC and normalization

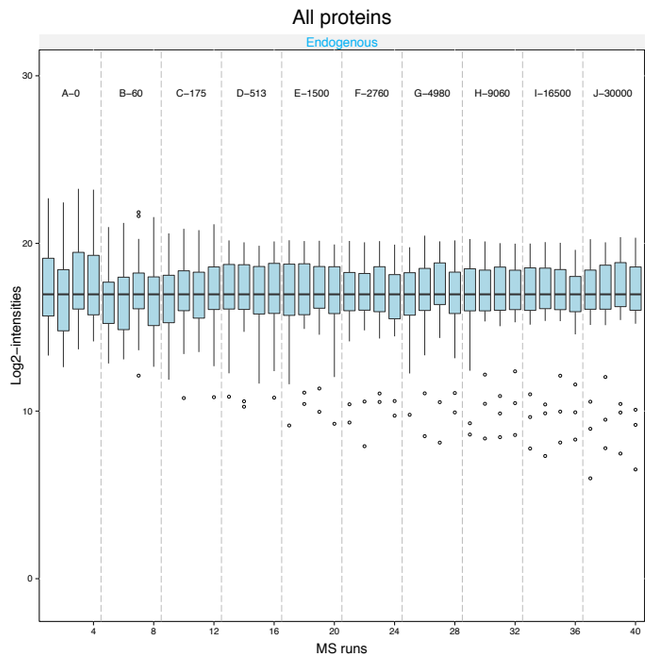
Data preparation

- Formatting
- Visualization
- Remove artifacts
- Normalization

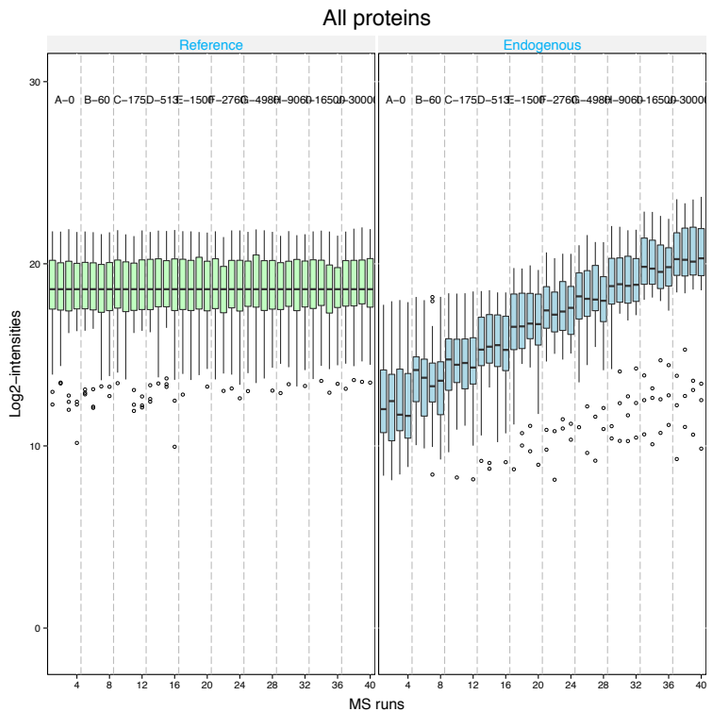
No normalization



Equalize medians normalization



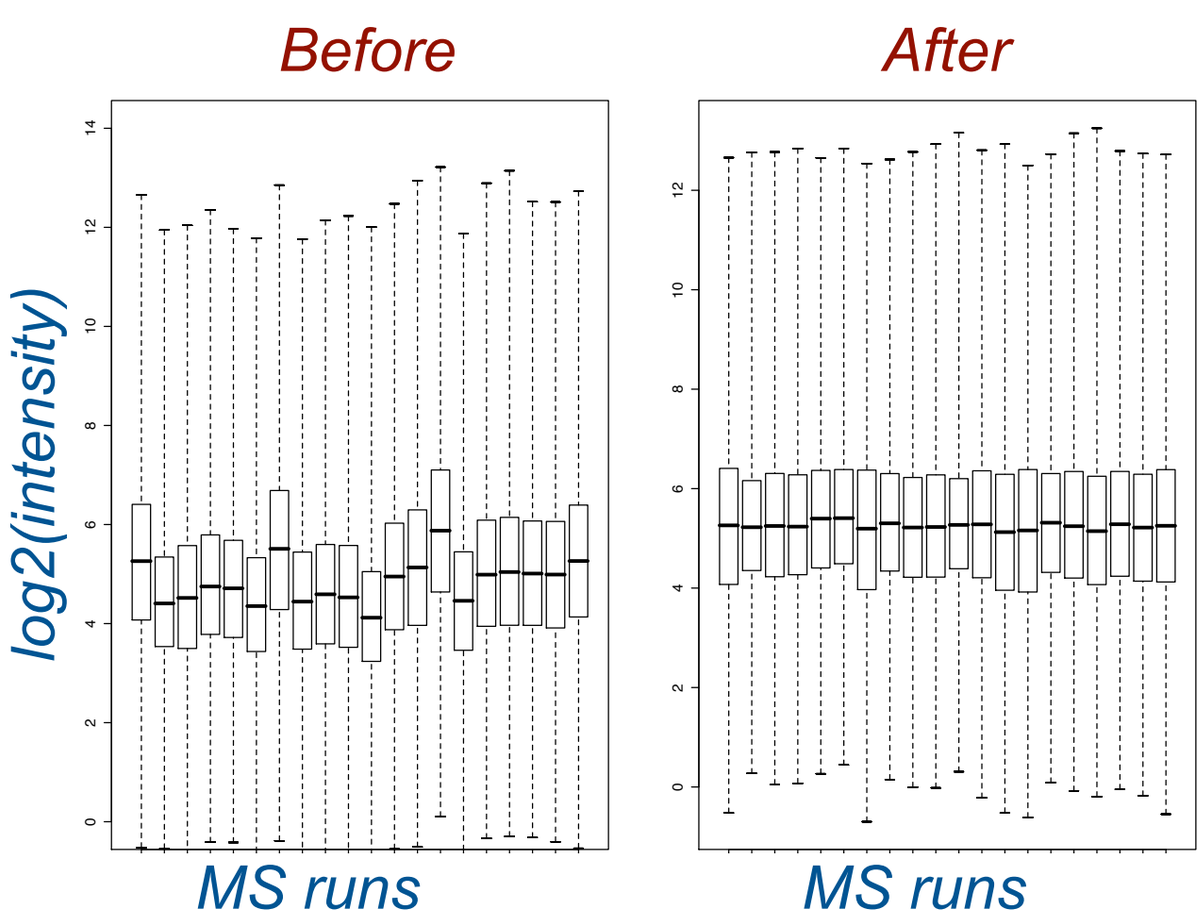
Equalize medians normalization



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

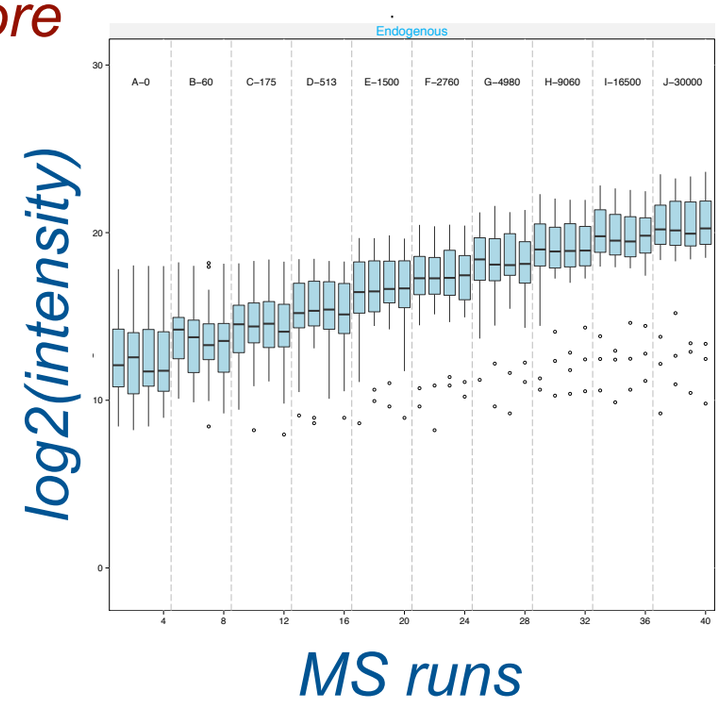


MEDIAN NORMALIZATION

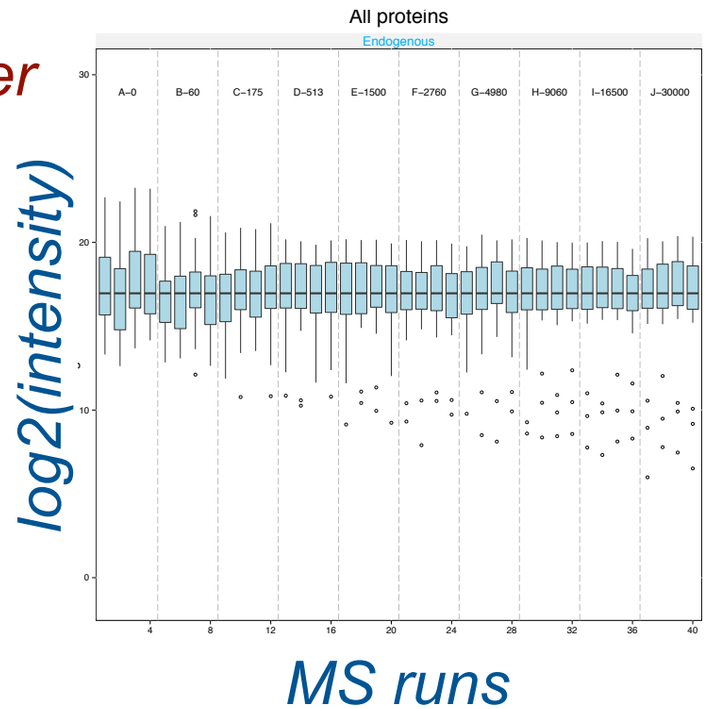
Assumes constant abundance of median of log(int) of endogenous features

- Algorithm
  - Subtract median[log(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



After



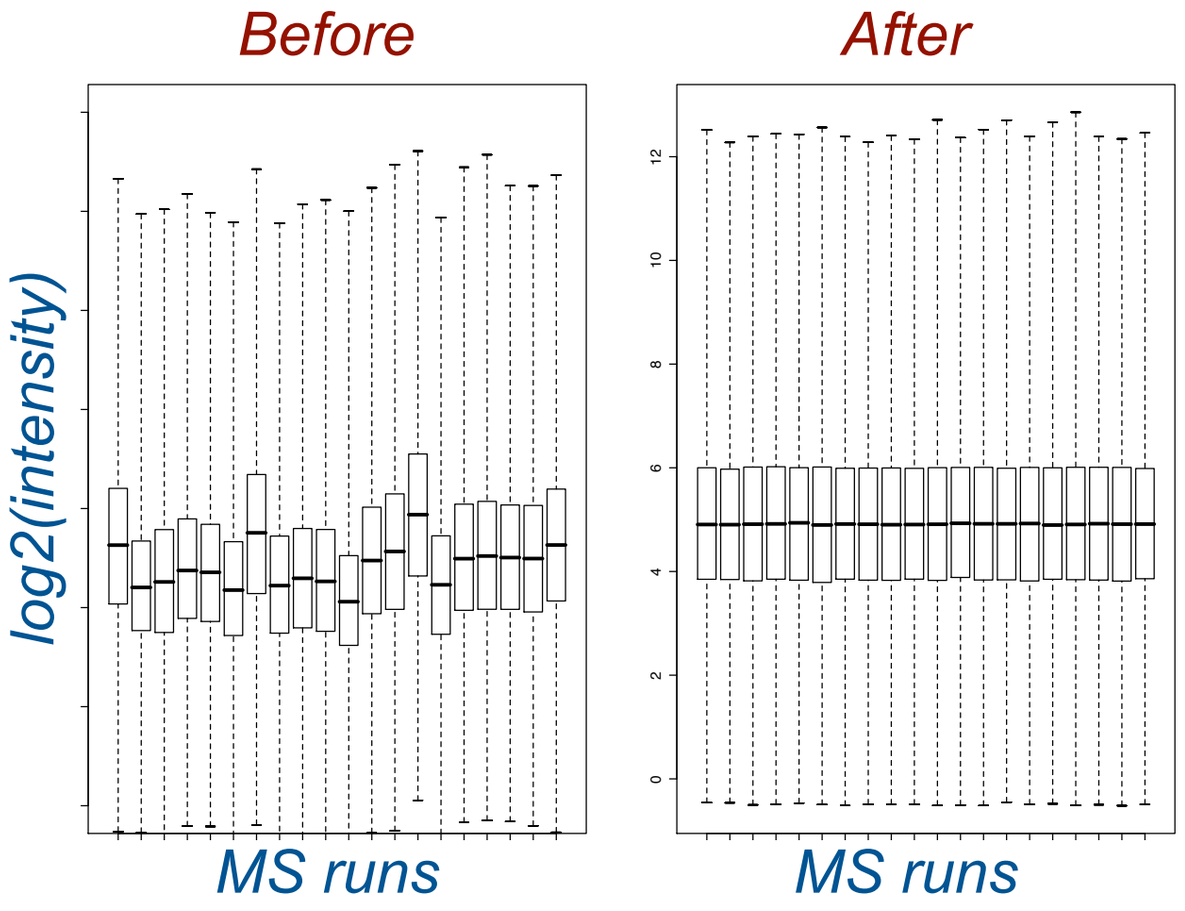
QUANTILE NORMALIZATION

Assumes constant abundance of all quantiles of log(int) of endogenous features

- Algorithm
  - Order log(int) in each run
  - Calculate average of each quantile across runs
  - Substitute each log(int) with the average
  - Re-arrange log(int) in original order

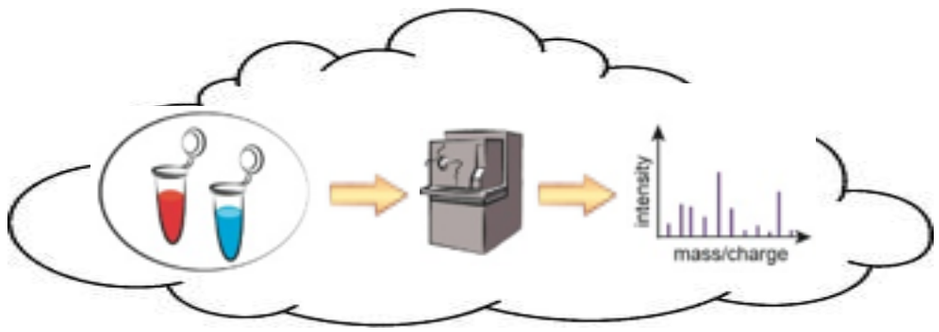
- Comments
  - \+ More stable than a single standard
  - \+ Accounts for all data processing steps
  - -- Assumes that the majority of endogenous proteins are not affected by the conditions
  - -- Often very aggressive

- Best practice
  - Use to normalize multiple standards



A TYPICAL ANALYSIS WORKFLOW

Experimental design



QC and normalization

Statistical modeling

Model-based conclusions

Experimental design

Summarize all protein features in a statistical model

- Systematic variation
- Random variation

Verify the assumptions!

- Describe statistical properties of
  - experimental design
  - biological variation
  - measurement technology

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

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

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

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

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

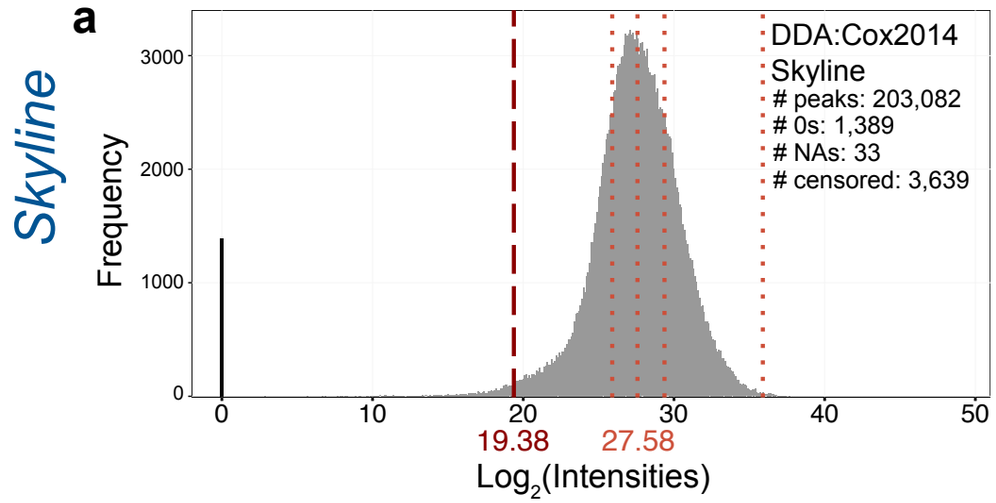
$\mathbf{Run \times Feature_{ijkl}} = \epsilon_{\mathbf{ijkl}} \overset{\mathbf{iid}}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2_{\epsilon})$

PROPERTIES OF PEAK INTENSITIES VARY BETWEEN DATA PROCESSING TOOLS

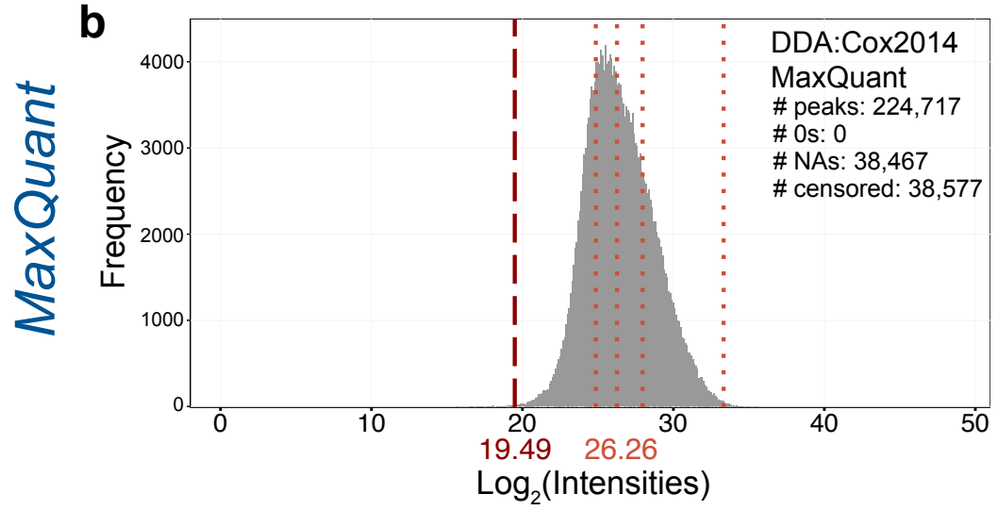
DDA: Cox 2014

DIA: Bruderer 2015

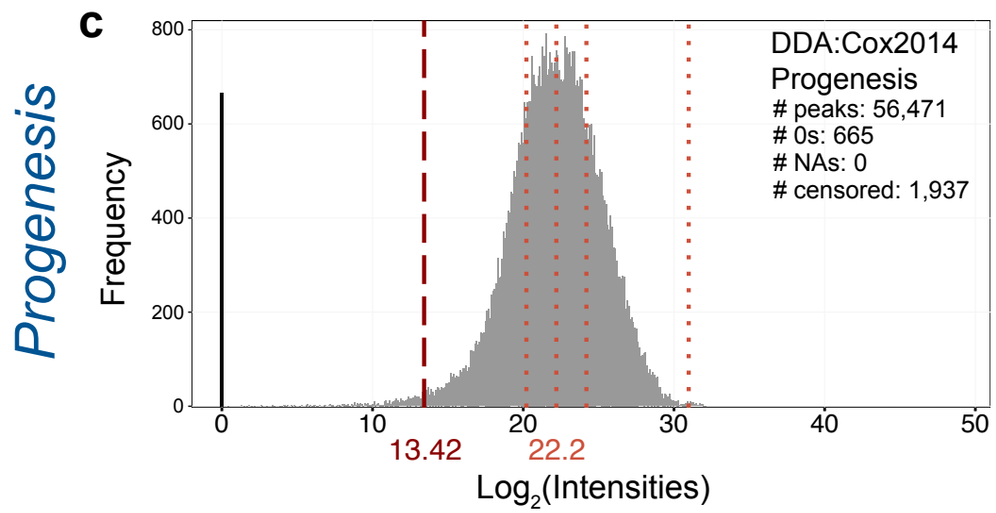
Skyline



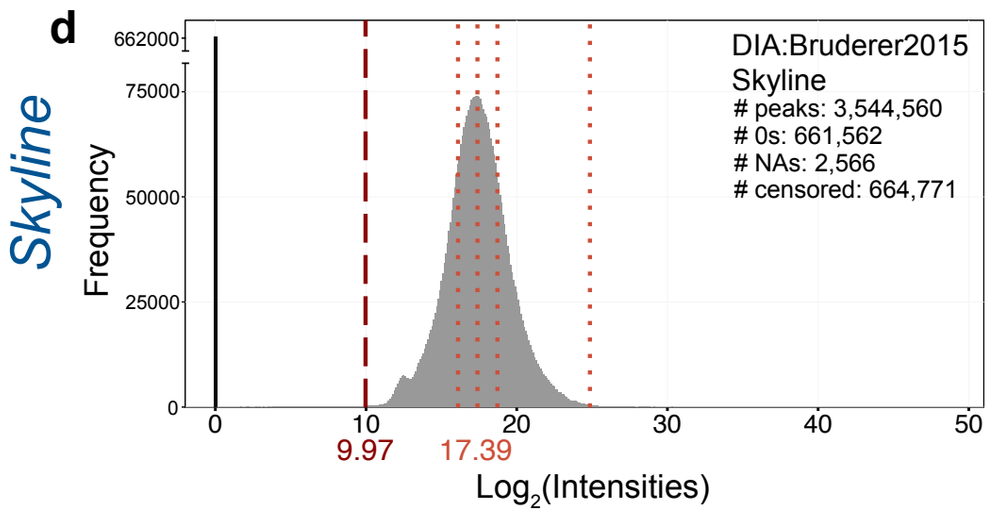
MaxQuant



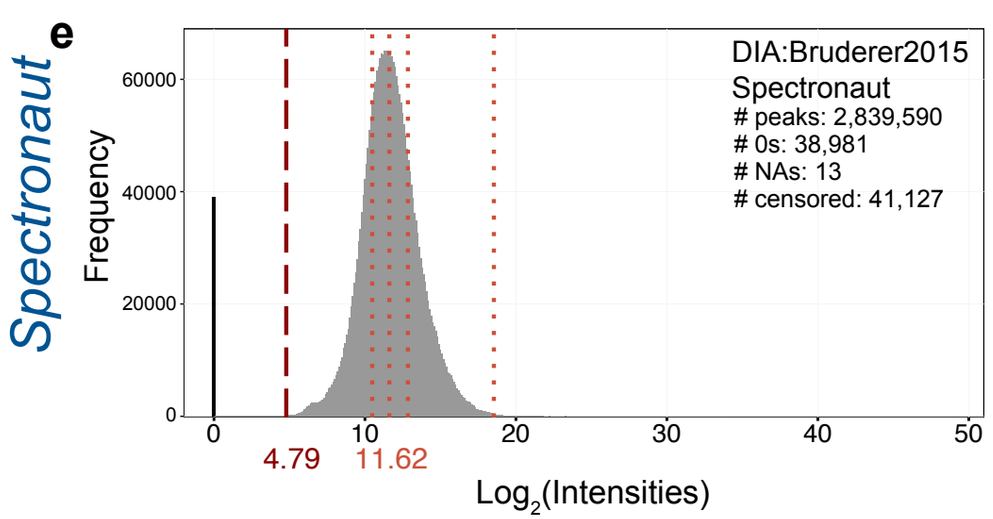
Progenesis

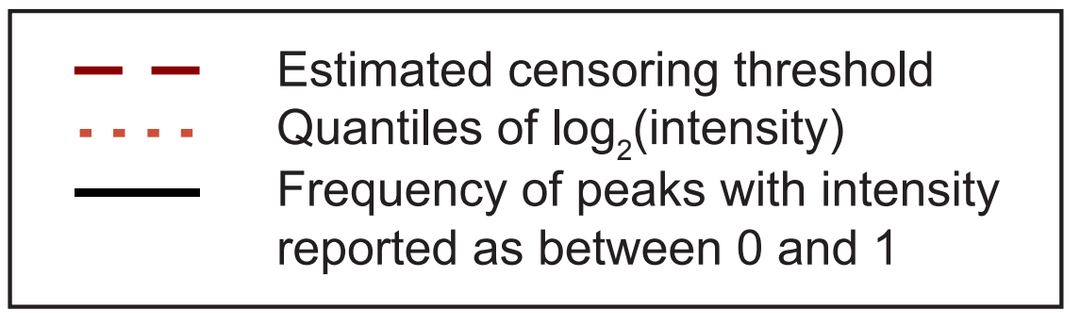


Skyline



Spectronaut





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
$Feature_2$	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

IMPUTING 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
$Feature_2$	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

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	$\mathbf{y}_{imp}$	y	$\mathbf{y}_{imp}$	$\mathbf{y}_{imp}$	y	· · ·	$\mathbf{y}_{imp}$	y	y	· · ·	$\mathbf{y}_{imp}$	y	y	y	y	y	· · ·	y	$\mathbf{y}_{imp}$	y

ROBUST RUN SUMMARIZATION

A

	Run_1	Run_2	Run_3	Run_4	Run_5	Run_6	· · ·	Run_{JK-2}	Run_{JK-1}	Run_{JK}	· · ·	$Run_{(l-1)JK+1}$	$Run_{(l-1)JK+2}$	$Run_{(l-1)JK+3}$	$Run_{(l-1)JK+4}$	$Run_{(l-1)JK+5}$	$Run_{(l-1)JK+6}$	· · ·	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}	· · ·	$Run_{(l-1)JK+1}$	$Run_{(l-1)JK+2}$	$Run_{(l-1)JK+3}$	$Run_{(l-1)JK+4}$	$Run_{(l-1)JK+5}$	$Run_{(l-1)JK+6}$	· · ·	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

$$y_{ijkl} = \mu + Run_{ijk} + Feature_l + \epsilon_{ijkl} \text{ , where}$$
$$median_{ijk}(Run_{ijk}) = 0,\ median_l(Feature_l) = 0,\ and\ median_{ijk}(\epsilon_{ijkl}) = median_l(\epsilon_{ijkl}) = 0$$

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

SUB-PLOT

Summarization over all features in a run

log(Feature int) × *Run*

	1	2	...	12
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

$$x_{ij} = feature_i + run_j + error_{ij}$$

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

LINEAR MODELS FOR THE DESIGN

	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	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	· · ·	$\hat{y}$	$\hat{y}$	$\hat{y}$	· · ·	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	· · ·	$\hat{y}$	$\hat{y}$	$\hat{y}$

C

Step 2 : Model-based inference by whole plot

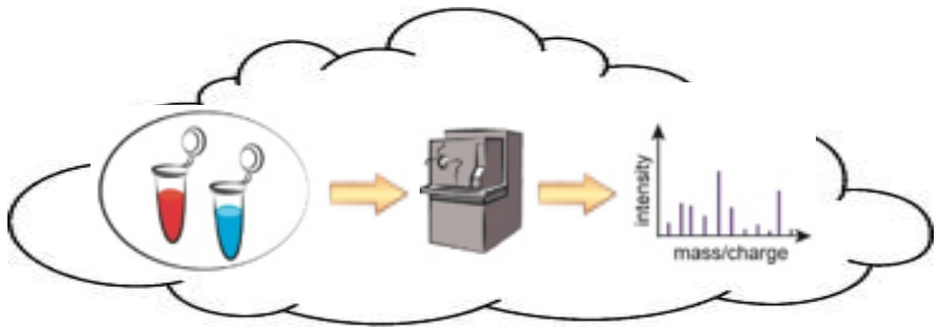
$$\hat{y}_{ijk} = \mu + Condition_i + Subject(Condition)_{j(i)} + \psi_{ijk}, \text{where}$$

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

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

A TYPICAL ANALYSIS WORKFLOW

Experimental design



QC and normalization

Statistical modeling

Model-based conclusions

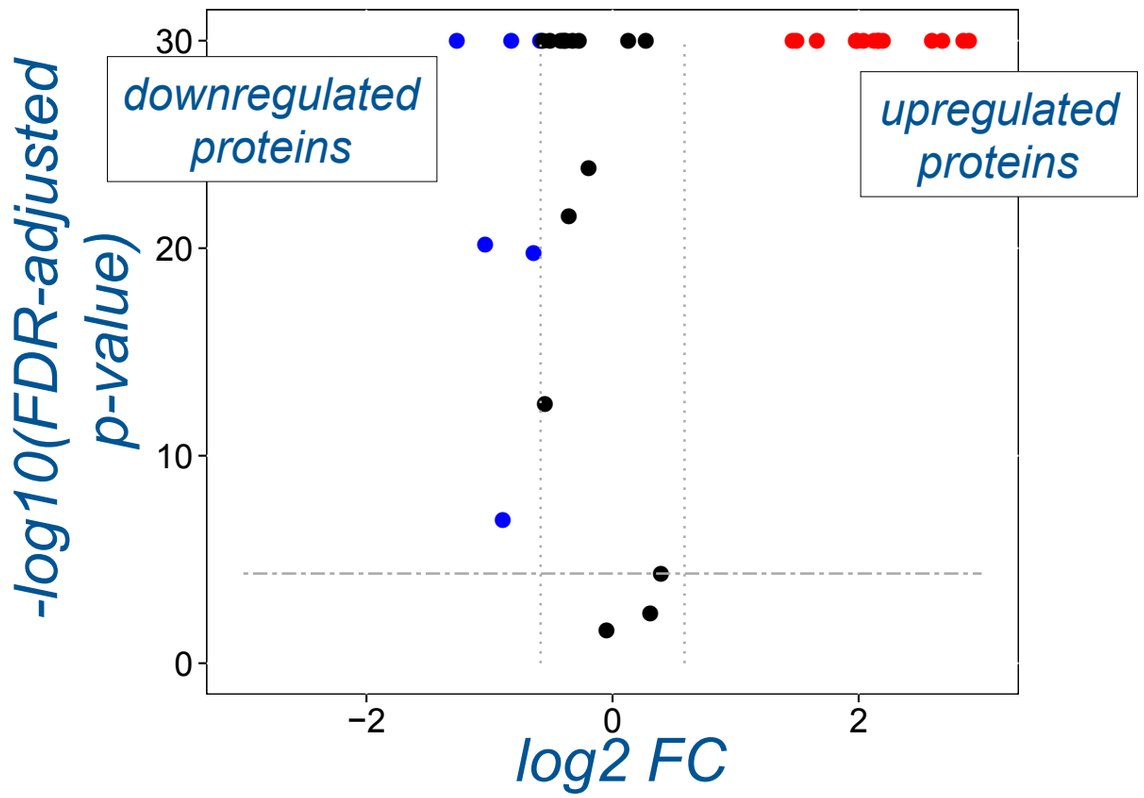
Experimental design

Model-based group comparisons

- Quantify the uncertainty
- Adjust p-values to control FDR

Relative protein quantification

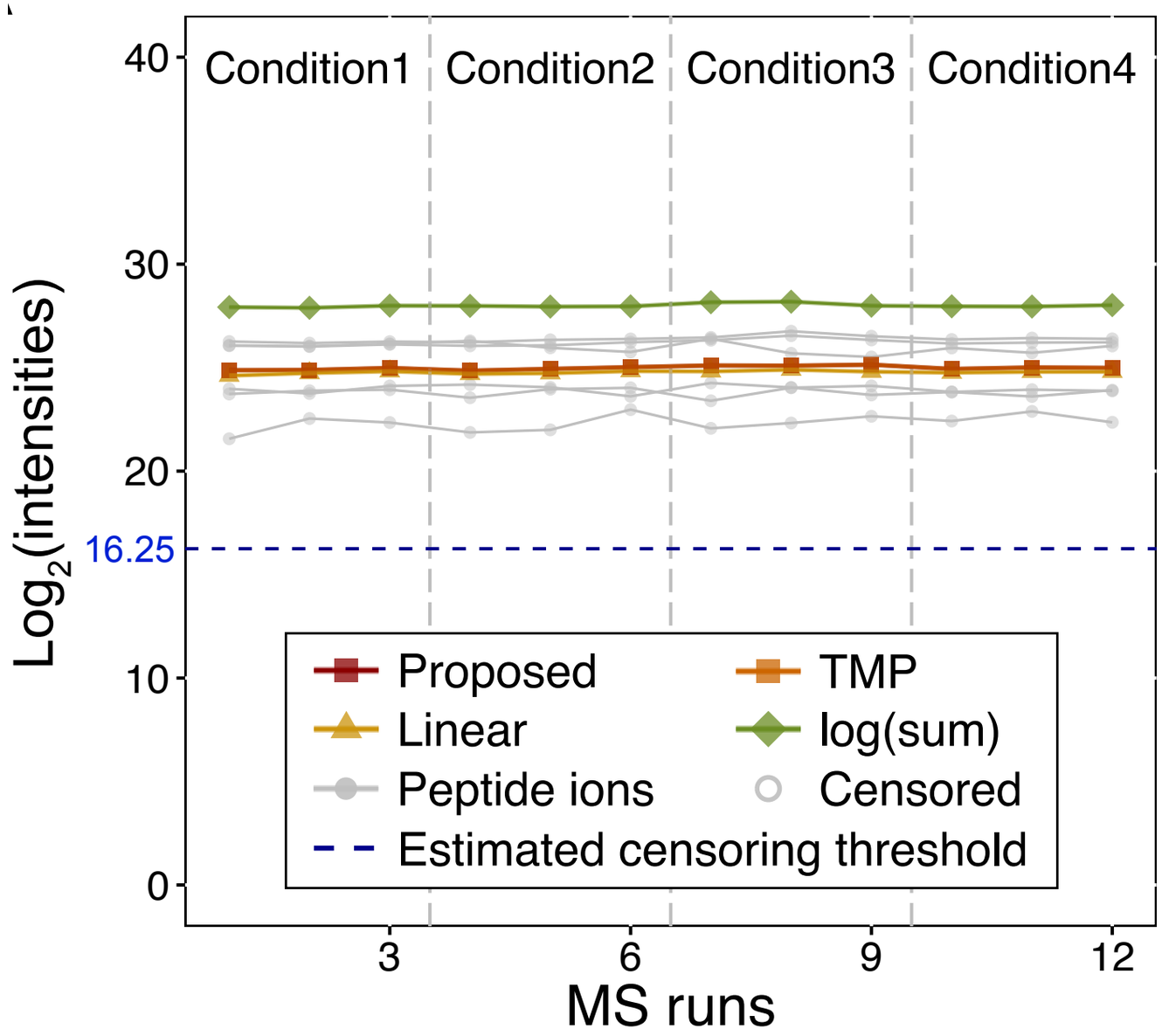
- In one sample
- In one condition

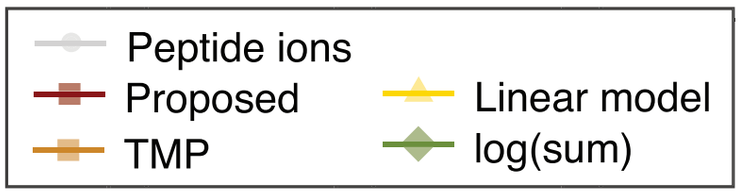


C.-Y. Chang et al. *MCP*, 2012

ROBUSTNESS TO OUTLIERS

Methods perform similarly with high quality data



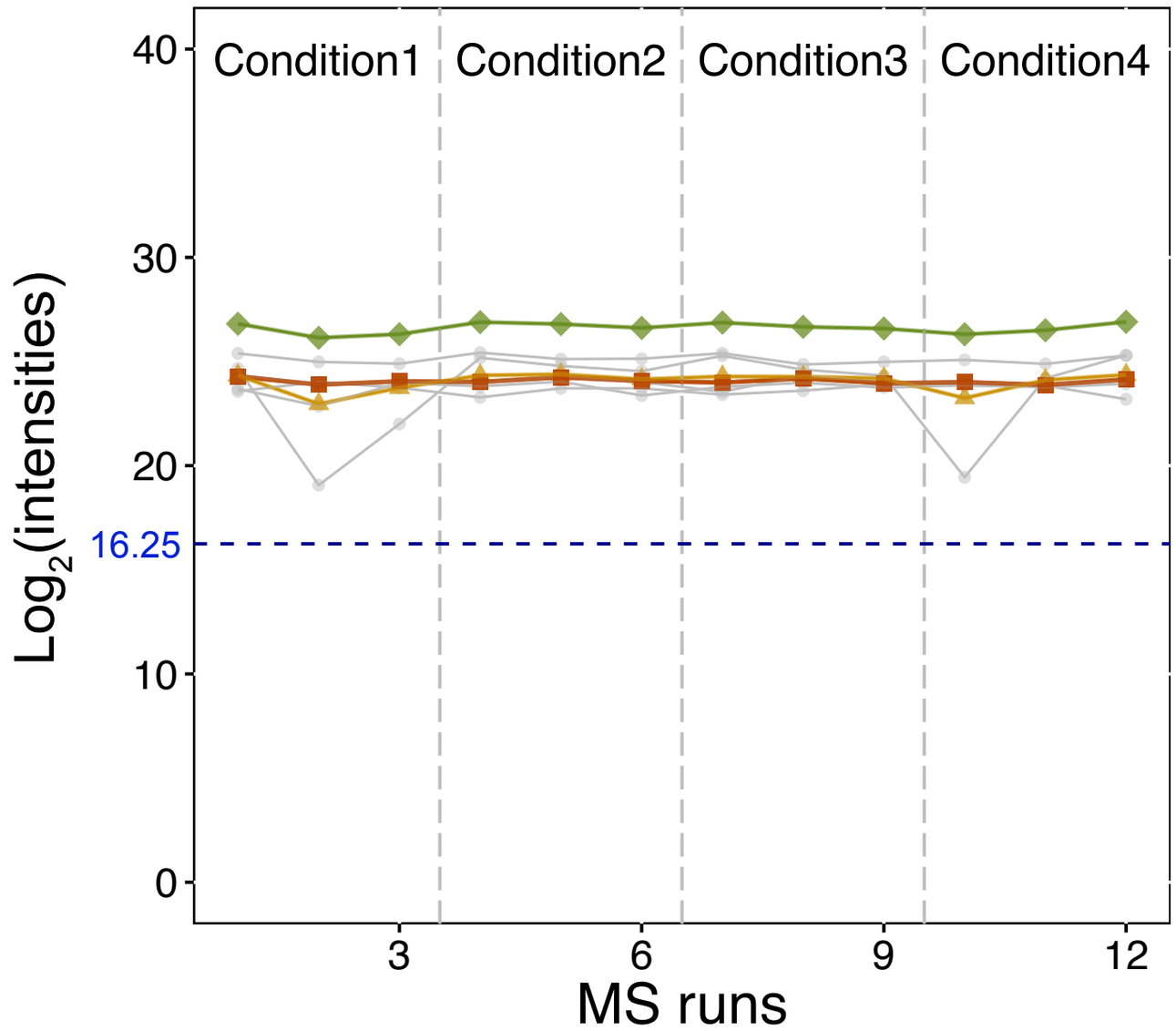


Condition2-Condition1 : True fold change=1

	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

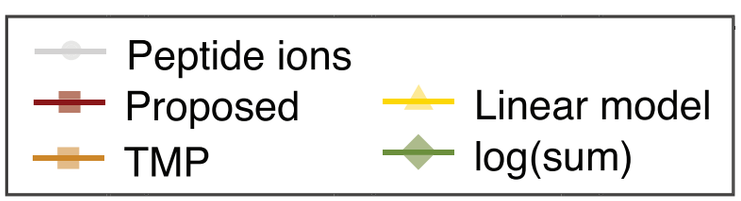
ROBUSTNESS TO OUTLIERS

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



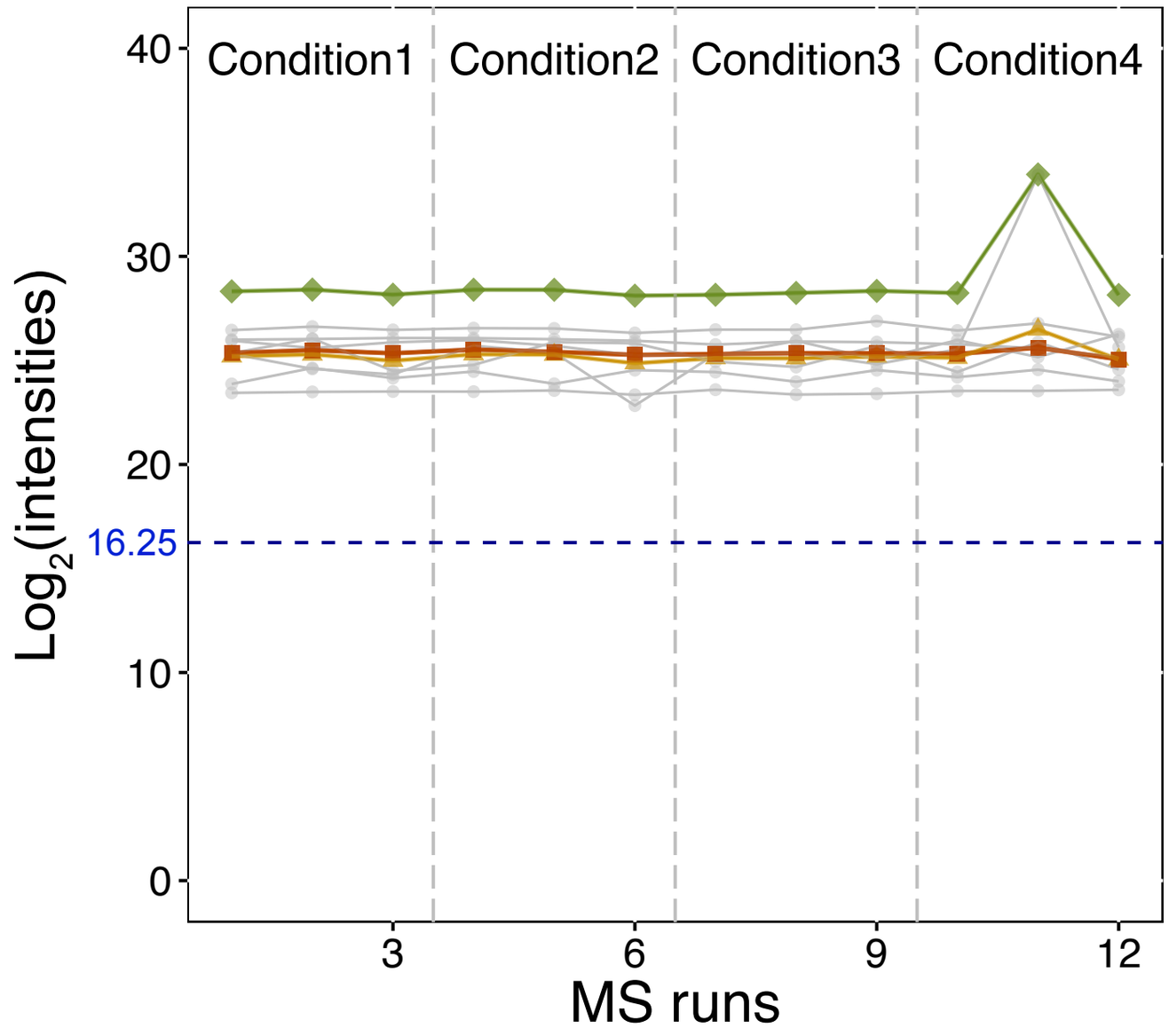
Condition3-Condition1 : True fold change=1

	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



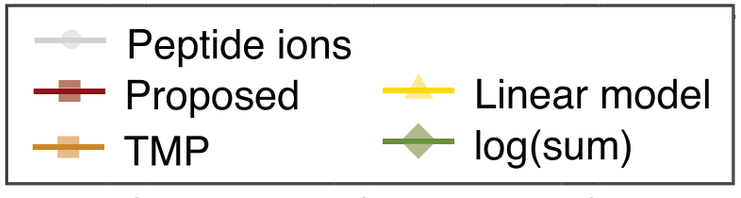
ROBUSTNESS TO OUTLIERS

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



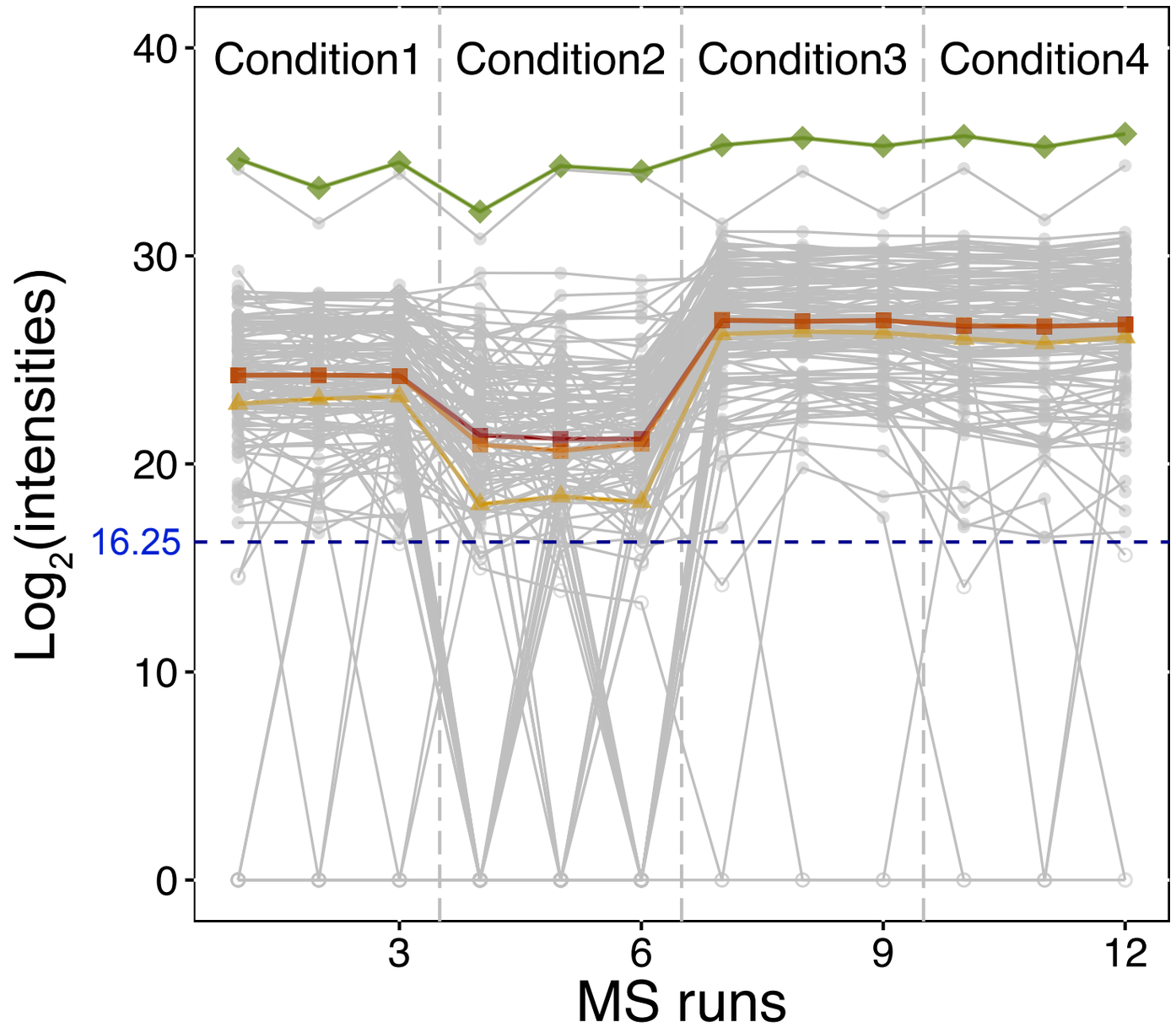
Condition4-Condition1 : True fold change=1

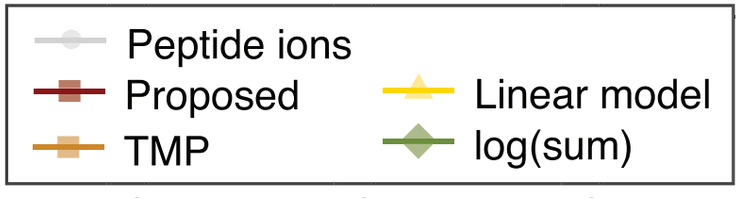
	EstimatedFC	Adj.pvalue
Proposed	0.951	0.948
TMP	0.951	0.948
Linear model	1.317	0.881
log(sum)	3.514	0.741



ROBUSTNESS TO OUTLIERS

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





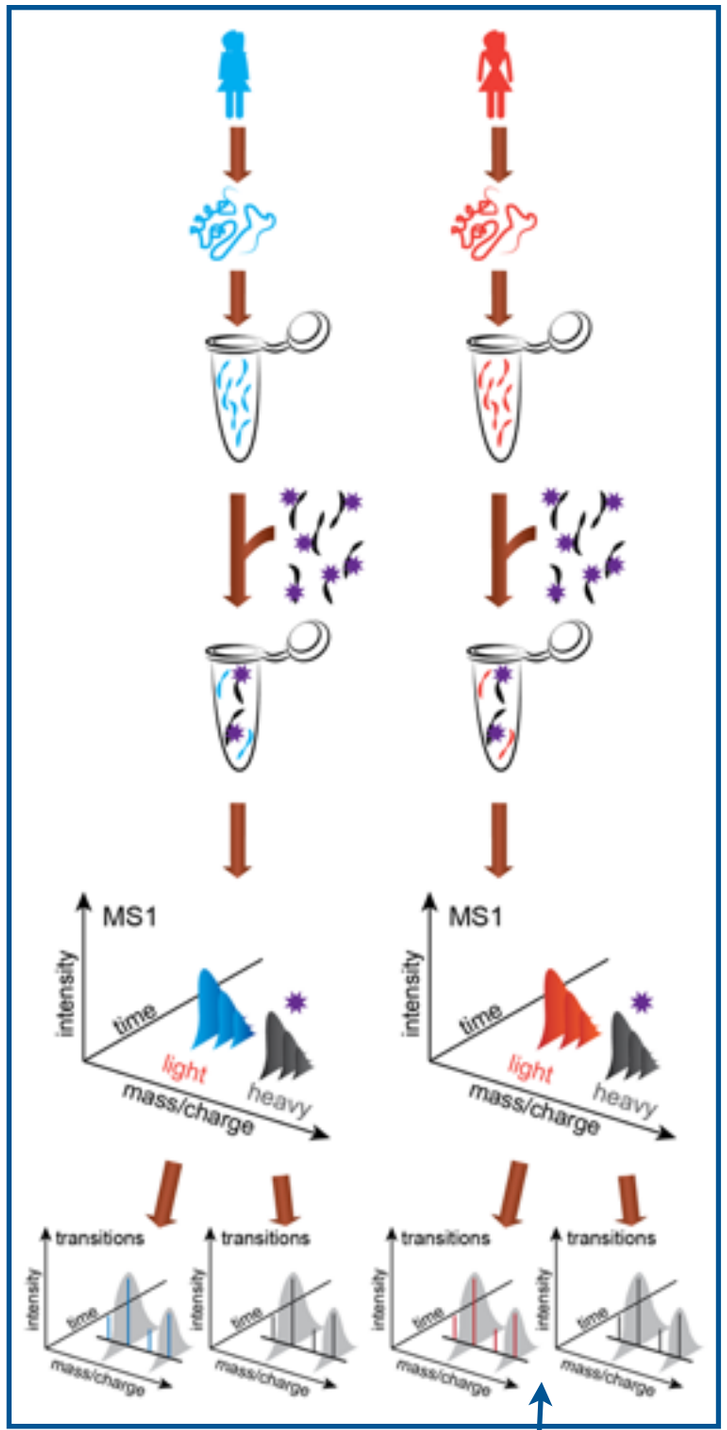
Condition1-Condition2 : True fold change=7.5

	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

EXAMPLE: LABELED REFERENCE PEPTIDES

Labeling + statistical modeling reduces the variation

Label-based SRM workflow



Transitions

Analysis of heavy/light peak pairs

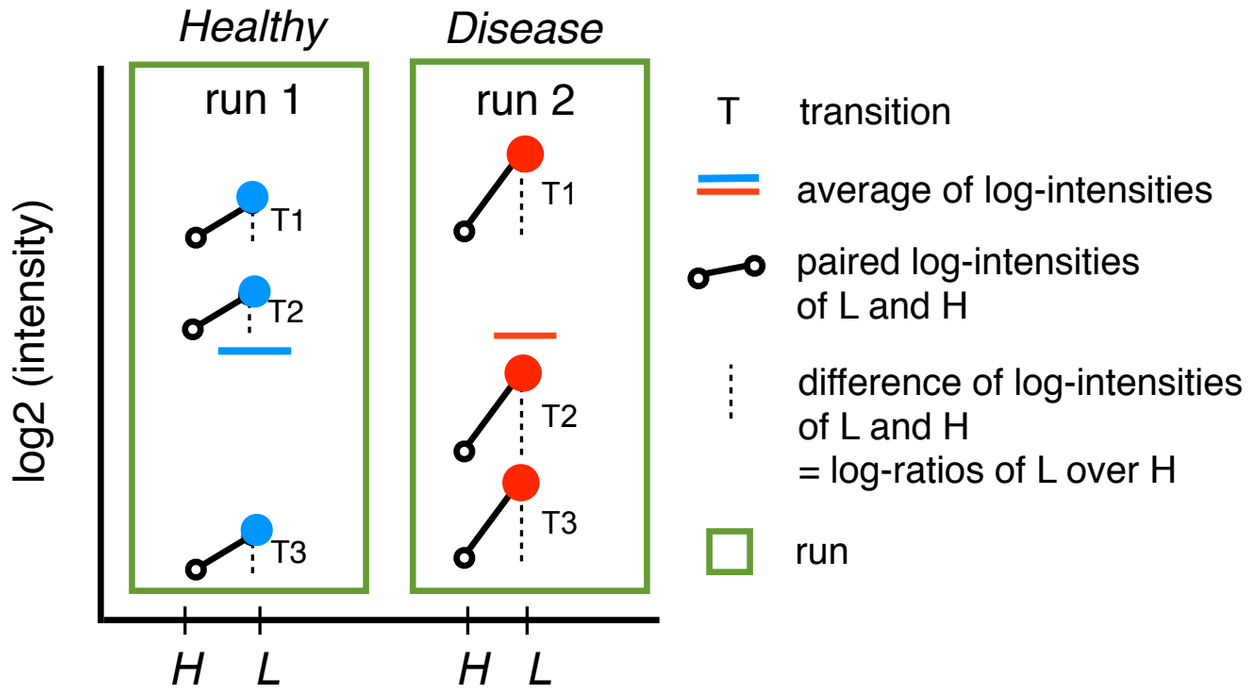


Table of quantified peaks

			Group 1				Group I		
			Run 1	⋯	⋯	⋯	⋯	⋯	Run M
			Subject 1	⋯	Subject J	⋯	Subject 1	⋯	Subject J
Endogenous: light labeled peptide	Peptide 1	Transition 1	10.21	⋯	10.57	⋯	15.64	⋯	15.03
		⋮		⋯		⋯		⋯	
		Transition L	10.52	⋯	10.92	⋯	15.29	⋯	15.68
	⋮	⋮		⋯		⋯		⋯	
	Peptide K	Transition 1	11.76	⋯	11.92	⋯	16.22	⋯	16.71
		⋮		⋯		⋯		⋯	
		Transition L	11.65	⋯	11.09	⋯	16.27	⋯	16.51
Reference: heavy labeled peptide	Peptide 1	Transition 1	19.46	⋯	19.77	⋯	19.82	⋯	19.03
		⋮		⋯		⋯		⋯	
		Transition L	19.13	⋯	19.25	⋯	19.67	⋯	19.80
	⋮	⋮		⋯		⋯		⋯	
	Peptide K	Transition 1	19.26	⋯	19.33	⋯	19.58	⋯	19.61
		⋮		⋯		⋯		⋯	
		Transition L	19.73	⋯	19.09	⋯	19.84	⋯	19.55

Legend : Label | Feature: Transition/Peptide | Group | Run | Subject

EXTENSION: LABELED REFERENCE PEPTIDES

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	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	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$_I$	y		y			y	⋯		y	y	⋯		y	y	y	y	y	⋯	y		y
Reference	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	y
	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
	Feature$_L$	y	y	y	y	y	y	⋯	y	y	y	⋯	y	y	y	y	y	y	⋯	y	y	y

Whole plot | Subplot

$$y_{ijklm} = \mu + \underbrace{\mathbf{Condition}_i + \mathbf{Subject(Condition)_{j(i)}}}_{\text{Whole-plot biological variation}} + \underbrace{\mathbf{Run_{ijk}}}_{\text{Whole-plot technical variation}} + \mathbf{Label_m} + \mathbf{Run \times Label}_{ijkm} + \mathbf{Feature_l} + \underbrace{\epsilon_{ijklm}}_{\text{Subplot error}}$$

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

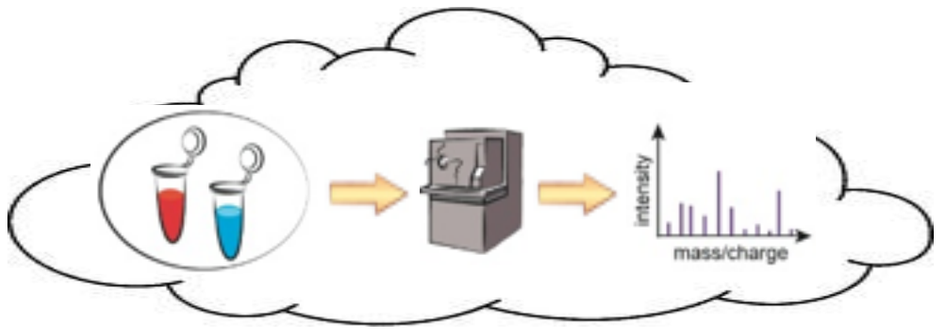
$\mathbf{Subject(Condition)_{j(i)}} \overset{\mathbf{iid}}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2_{\mathbf{Subject}})$

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

$\epsilon_{ijklm} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_\epsilon)$

A TYPICAL ANALYSIS WORKFLOW

Experimental design



QC and normalization

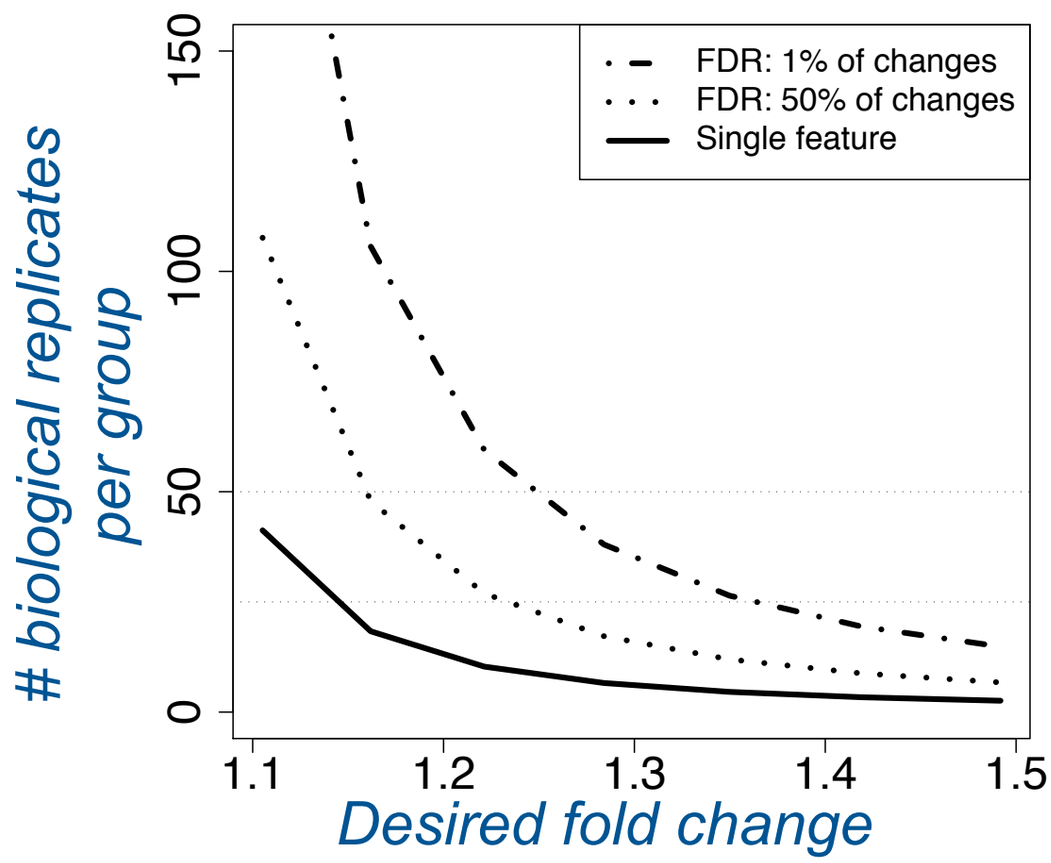
Statistical modeling

Model-based conclusions

Experimental design

Use the dataset to improve:

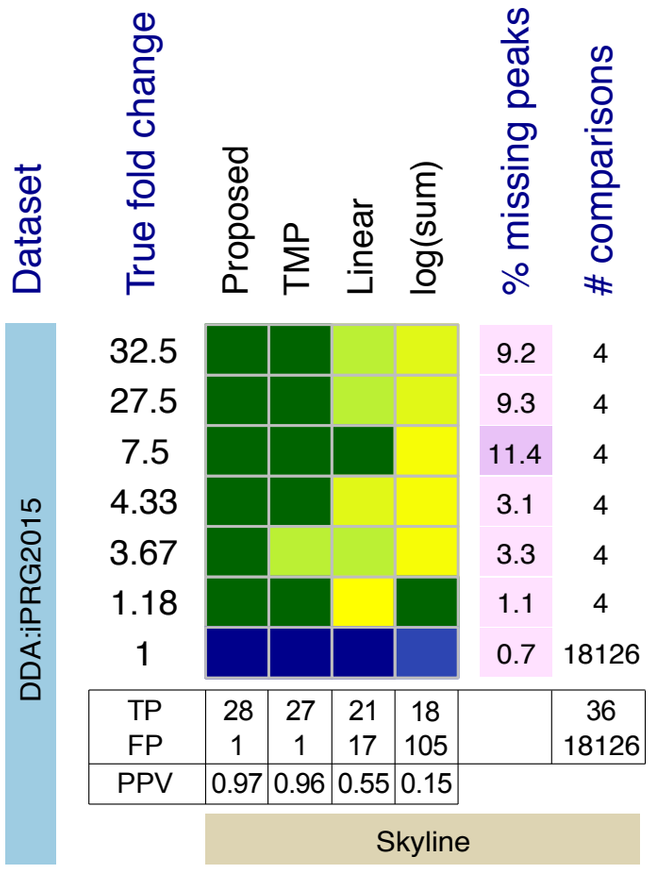
- Subject selection: matching
- Resource allocation: blocking
- Calculation of sample size

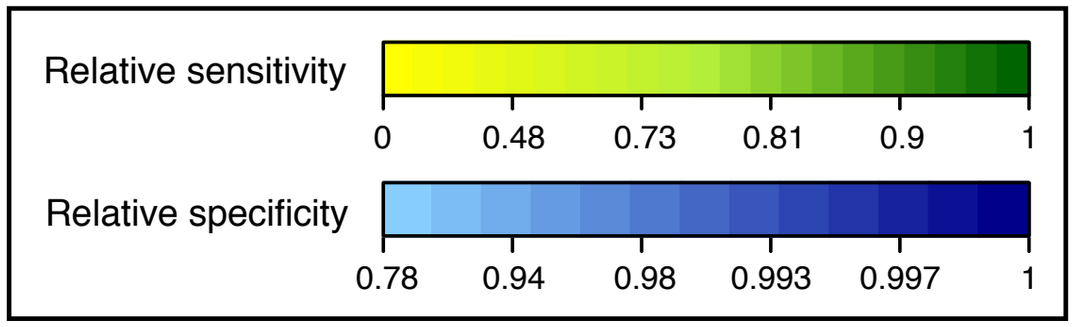


OUTLINE

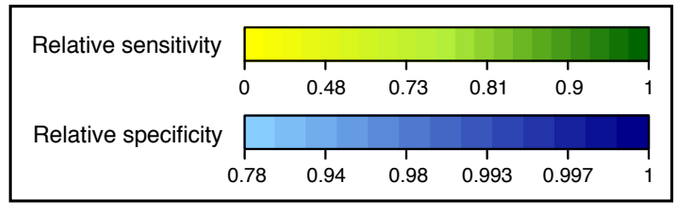
- iPRG2015: detection of diff. abundance
  - Community effort with label-free shotgun proteomics
- MSstats
  - Normalization, statistical modeling, inference
  - Evaluation
- Extensions to MSstats
  - Assay characterization, longitudinal monitoring

TESTING: DDA


Dataset
True fold change
Proposed
TMP
Linear
log(sum)
% missing peaks
# comparisons
DDA:iPRG2015
32.5
27.5
7.5
4.33
3.67
1.18
1
9.2
9.3
11.4
3.1
3.3
1.1
0.7
4
4
4
4
4
4
18126
TP
FP
PPV
28
1
0.97
27
1
0.96
21
17
0.55
18
105
0.15
36
18126
Skyline


Relative sensitivity
0
0.48
0.73
0.81
0.9
1
Relative specificity
0.78
0.94
0.98
0.993
0.997
1

TESTING: DDA



a

Dataset	True fold change	Proposed	TMP	Linear	log(sum)	% missing peaks	# comparisons
DDA:iPRG2015	32.5					9.2	4
	27.5					9.3	4
	7.5					11.4	4
	4.33					3.1	4
	3.67					3.3	4
	1.18					1.1	4
	1					0.7	18126
	TP	28	27	21	18		36
	FP	1	1	17	105		18126
	PPV	0.97	0.96	0.55	0.15		
		Skyline					

		Proposed	TMP	Linear	log(sum)	% missing peaks	# comparisons
DDA:iPRG2015	32.5					40.5	4
	27.5					41.0	4
	7.5					55.8	4
	4.33					20.4	4
	3.67					20.9	4
	1.18					5.5	4
	1					10.7	17205
	TP	26	23	19	25		30
	FP	7	1	5	20		17205
	PPV	0.79	0.96	0.79	0.56		
		MaxQuant					

		Proposed	TMP	Linear	log(sum)	% missing peaks	# comparisons
DDA:iPRG2015	32.5					4.2	4
	27.5					4.2	4
	7.5					6.2	4
	4.33					2.1	4
	3.67					2.1	4
	1.18					0	4
	1					0.4	18054
	TP	26	23	24	30		36
	FP	106	96	183	212		18054
	PPV	0.20	0.19	0.12	0.11		
		Progenesis					

		Proposed	TMP	Linear	log(sum)	% missing peaks	# comparisons
DDA:Cox2014	10					5.5	8
	0.1					1.9	8
	1					1.6	1931
	TP	38	38	34	35		39
	FP	13	13	2	14		1931
	PPV	0.75	0.75	0.94	0.71		
		Skyline					

		Proposed	TMP	Linear	log(sum)	% missing peaks	# comparisons
DDA:Cox2014	10					29.3	8
	0.1					10.2	7
	1					14.0	2046
	TP	34	29	26	34		35
	FP	5	6	5	10		2046
	PPV	0.87	0.83	0.84	0.77		
		MaxQuant					

		Proposed	TMP	Linear	log(sum)	% missing peaks	# comparisons
DDA:Cox2014	10					2.6	8
	0.1					0	7
	1					2.2	1877
	TP	38	37	36	38		38
	FP	17	17	25	31		1877
	PPV	0.69	0.69	0.59	0.55		
		Progenesis					

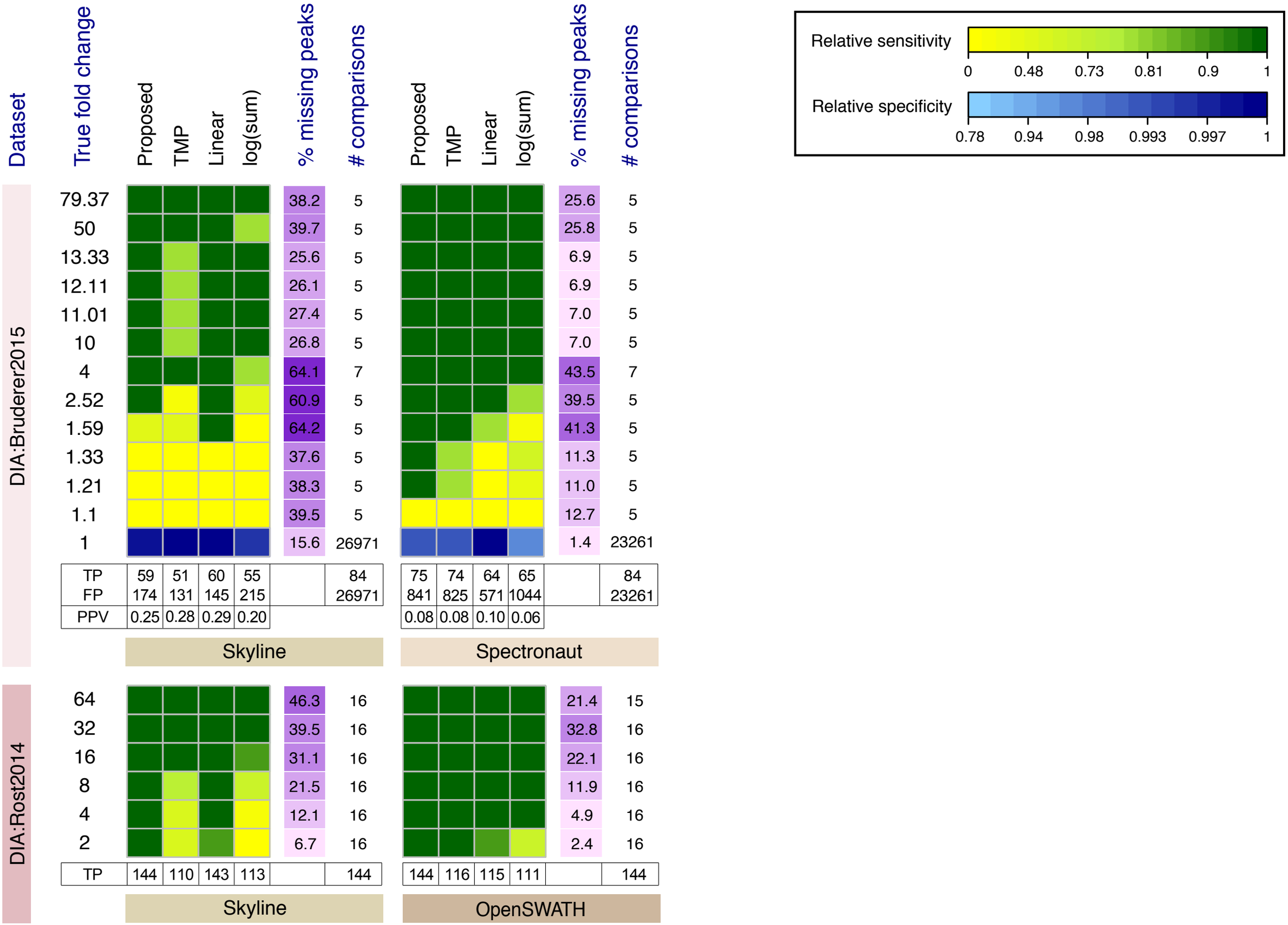
		Proposed	TMP	Linear	log(sum)	% missing peaks	# comparisons
DDA:SpikeIn	4					0.9	48
	2					1.1	144
	1					1.4	12258
	TP	169	168	129	165		192
	FP	101	94	69	178		12258
	PPV	0.63	0.64	0.65	0.48		
		Skyline					

		Proposed	TMP	Linear	log(sum)	% missing peaks	# comparisons
DDA:SpikeIn	4					12.4	56
	2					12.9	168
	1					13.6	15262
	TP	188	192	110	190		224
	FP	76	84	62	109		15262
	PPV	0.71	0.70	0.64	0.64		
		MaxQuant					

		Proposed	TMP	Linear	log(sum)	% missing peaks	# comparisons
DDA:SpikeIn	4					3.1	54
	2					3.0	162
	1					1.0	13084
	TP	190	186	97	202		216
	FP	53	49	40	143		13084
	PPV	0.78	0.79	0.71	0.59		
		Progenesis					

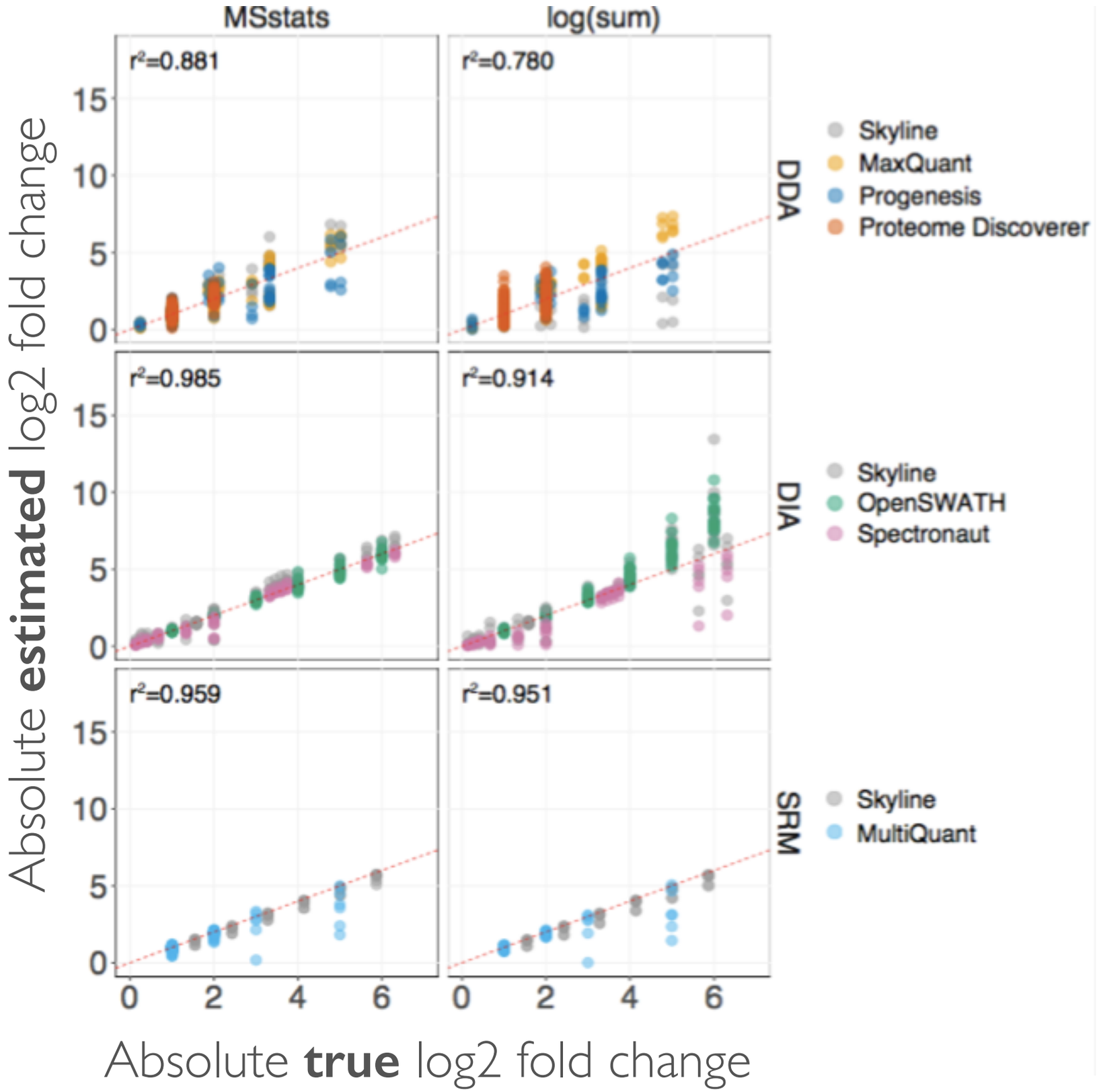
		Proposed	TMP	Linear	log(sum)	% missing peaks	# comparisons
DDA:SpikeIn	4					28.2	54
	2					29.3	162
	1					30.1	12314
	TP	187	203	46	159		216
	FP	98	127	8	101		12314
	PPV	0.66	0.62	0.85	0.61		
		Proteome Discover					

TESTING: DIA


Dataset
True fold change
Proposed
TMP
Linear
log(sum)
% missing peaks
# comparisons
DIA:Bruderer2015
79.37 50 13.33 12.11 11.01 10 4 2.52 1.59 1.33 1.21 1.1 1
38.2 39.7 25.6 26.1 27.4 26.8 64.1 60.9 64.2 37.6 38.3 39.5 15.6
5 5 5 5 5 5 7 5 5 5 5 5 26971
25.6 25.8 6.9 6.9 7.0 7.0 43.5 39.5 41.3 11.3 11.0 12.7 1.4
5 5 5 5 5 5 7 5 5 5 5 5 23261
TP 59 51 60 55 84
FP 174 131 145 215 26971
PPV 0.25 0.28 0.29 0.20
TP 75 74 64 65 84
FP 841 825 571 1044 23261
PPV 0.08 0.08 0.10 0.06
Skyline
Spectronaut
DIA:Rost2014
64 32 16 8 4 2
46.3 39.5 31.1 21.5 12.1 6.7
16 16 16 16 16 16
21.4 32.8 22.1 11.9 4.9 2.4
15 16 16 16 16 16
TP 144 110 143 113 144
TP 144 116 115 111 144
Skyline
OpenSWATH

Relative sensitivity
0 0.48 0.73 0.81 0.9 1
Relative specificity
0.78 0.94 0.98 0.993 0.997 1

ESTIMATION OF LOG-FOLD CHANGE



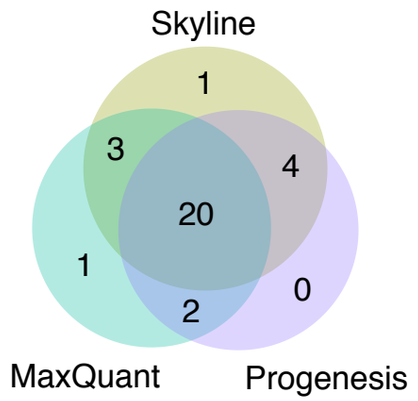
BETTER STATISTICAL METHODS ENHANCE REPRODUCIBLE RESEARCH

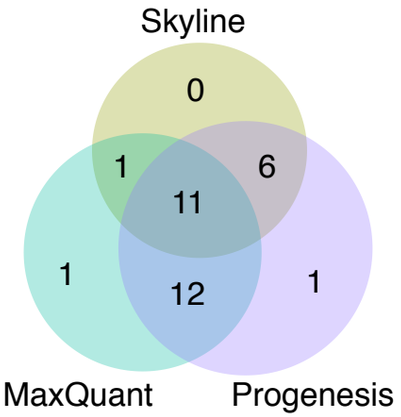
Better agreement in #differentially abundant proteins between tools

Proposed

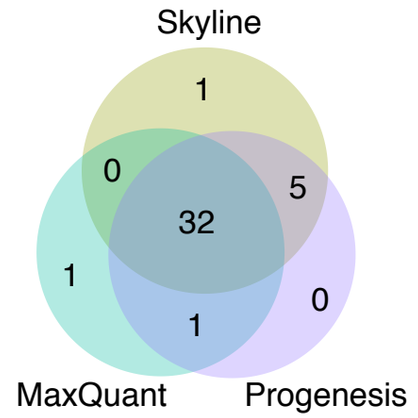
Log(sum)

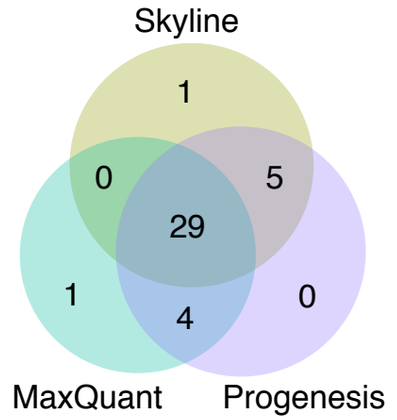
DDA: iPRG2015



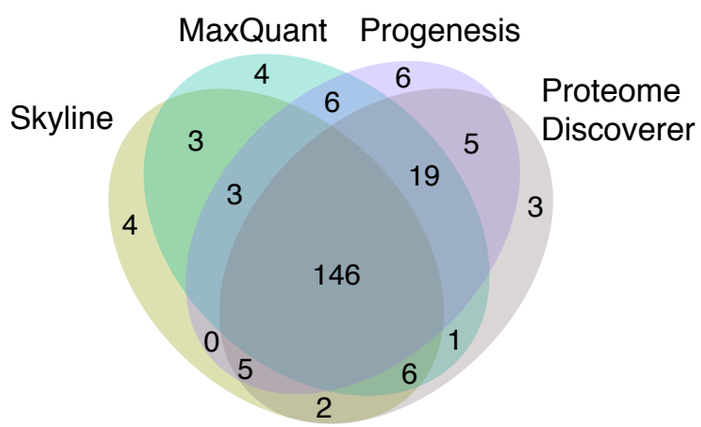


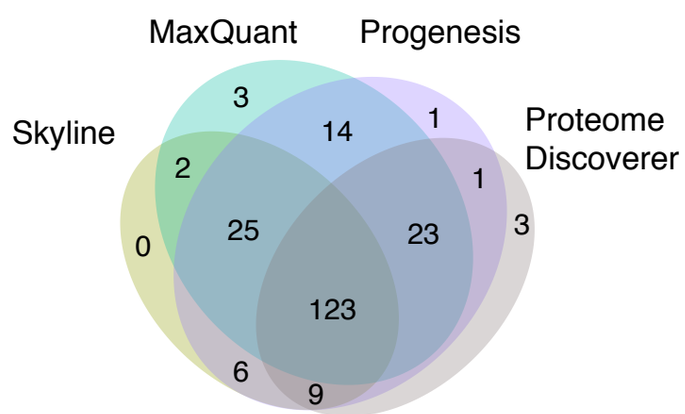
DDA: Cox 2014





DDA: Spike-in

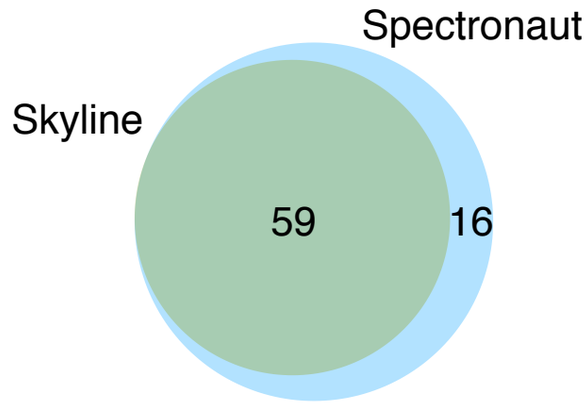




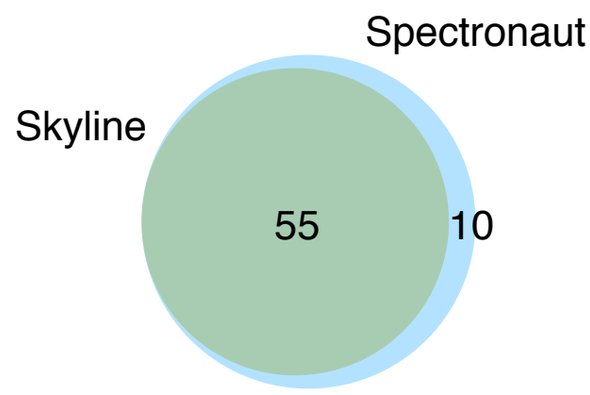
BETTER STATISTICAL METHODS ENHANCE REPRODUCIBLE RESEARCH

DIA: Bruderer2015

Proposed

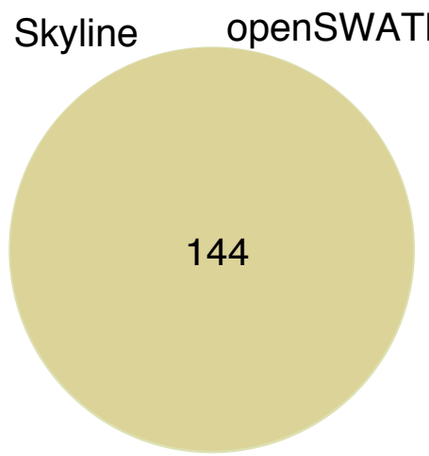


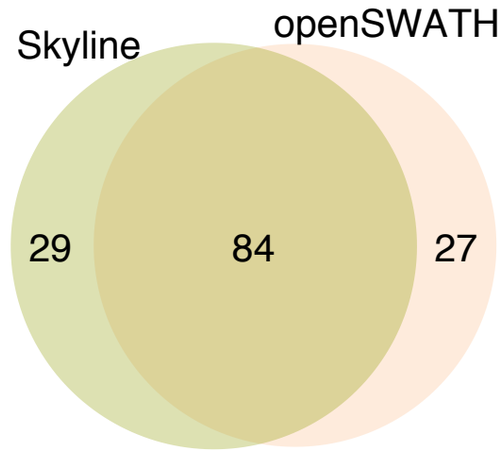
Log(sum)



Better agreement in #differentially abundant proteins between tools

DIA; Roest2014





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

msstats.org

MSstats

Statistical Tool For Quantitative Mass Spectrometry-Based Proteomics

HOME MSSTATS LOB/LOD MSSTATSQC DATASETS TRAINING 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).

News

- The upcoming NEU course (Computation and Statistics for Mass Spectrometry) on May 1-12, 2017. The detailed information will follow shortly.
- MSstats v3.6 is available on Bioconductor 3.4.
- 11/20/2016 – **MSstats v3.7.1 is updated! This is the most recent development version.** Please see 'Installation' page for more details. W will continue to update the versioning of the main package several time year, to synchronize with the Bioconductor release. – Changes since the previous version

MSstats

www.msstats.org

Statistical Tool For Quantitative Mass Spectrometry-Based Proteomics

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

Skyline external tool

Download MSstats

Downloaded: 10036

Tool Information

Organization: Vitek Lab, Purdue University

Authors: Meena Choi, Cyril Galitzine, Tsung-Heng Tsai, Olga Vitek

Languages: R(3.3.1), C#

More Information: http://www.msstats.org/

Bioconductor

Month	Nb of distinct IPs	Nb of downloads
Jan/2016	160	259
Feb/2016	160	295
Mar/2016	222	390
Apr/2016	255	366
May/2016	193	328
Jun/2016	220	378
Jul/2016	170	290
Aug/2016	153	241
Sep/2016	133	215
Oct/2016	172	368
Nov/2016	234	398
Dec/2016	192	363
2016	**1808**	**3891**

MSstats_2016_stats.tab

OUTLINE

- iPRG2015: detection of diff. abundance
  - Community effort with label-free shotgun proteomics
- MSstats
  - Normalization, statistical modeling, inference
  - Evaluation
- Extensions to MSstats
  - Assay characterization, longitudinal monitoring

STATISTICAL METHODS FOR ASSAY CHARACTERIZATION

MSstats

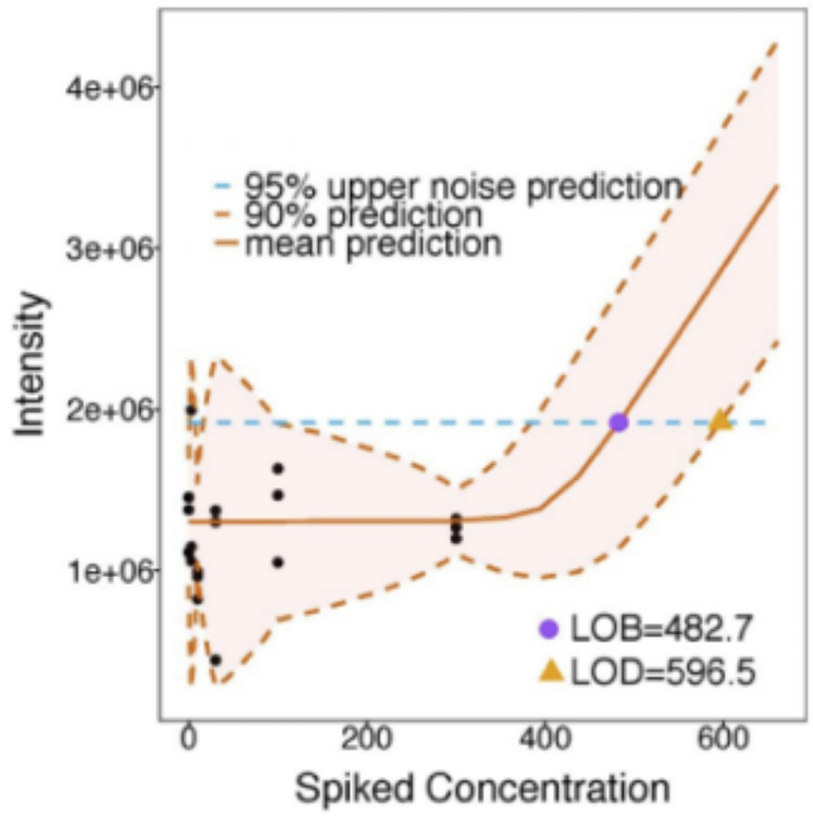
Statistical Tool For Quantitative Mass Spectrometry-Based Proteomics

HOME MSSTATS LOB/LOD MSSTATSQC DATASETS TRAINING CONTACT

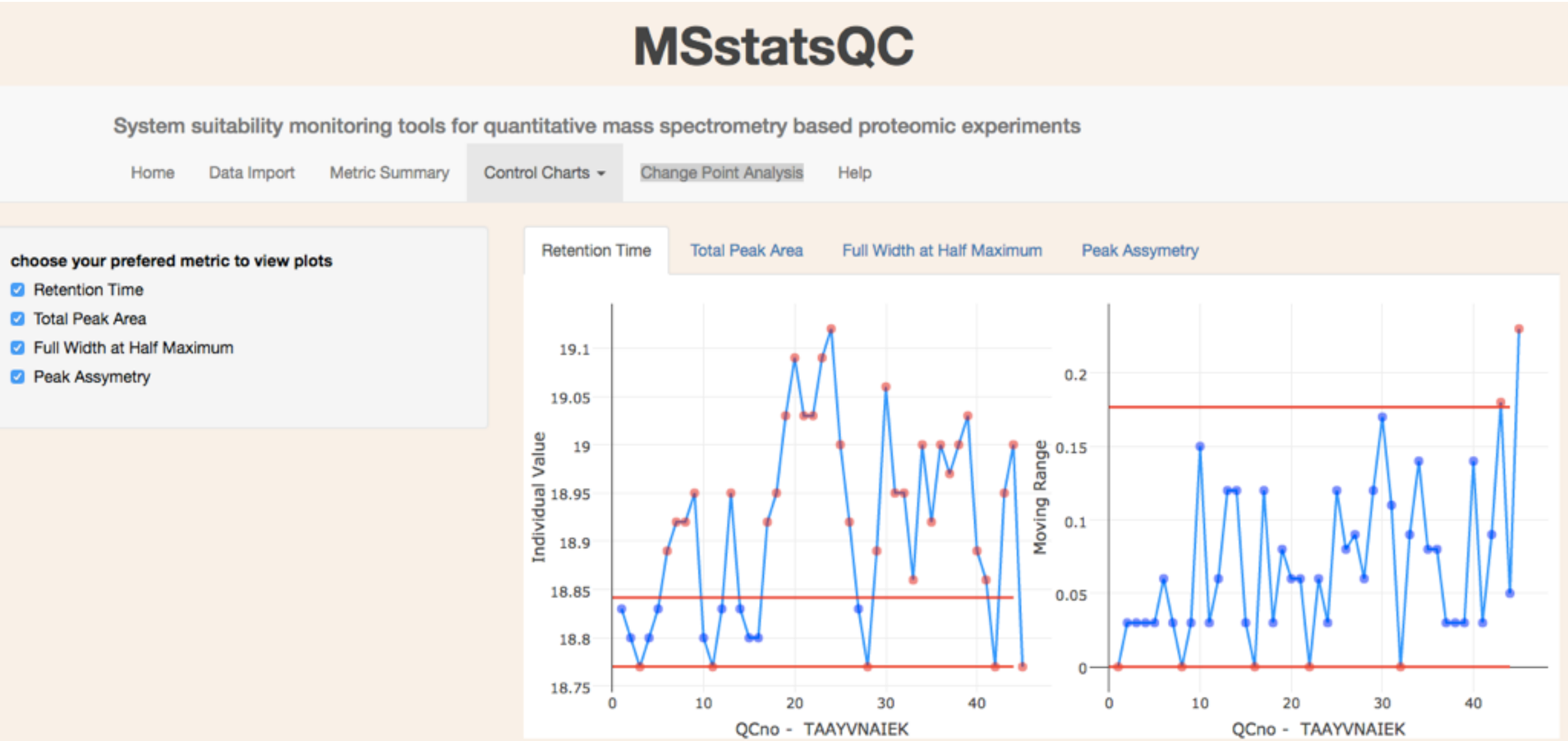
LOB/LOD ANALYSIS

Assay characterization : estimation of limit of blanc (LoB) and limit of detection (LoD)

The need for assay characterization is ubiquitous in quantitative mass spectrometry-based proteomics. Although many assay characteristics exist, the limit of blank (LOB) and limit of detection (LOD) are particularly useful figures of merit. LOB and LOD are determined by repeatedly measuring the peak intensities of peptides in samples with known peptide concentrations, and deriving an intensity versus concentration response curve. Most commonly, a weighted linear regression is fit to the intensity-concentration response, and LOB and LOD are estimated from the fit. Linear methods, however, inaccurately characterize assays containing a noise threshold at low concentrations, which is a very common situation. We propose a new approach based on non-linear regression that correctly captures the noise threshold. In absence of a noise threshold, the estimates of LOB/LOD obtained with non-linear statistical modeling are identical to those of weighted linear regression. However, in presence of a noise threshold the non-linear model changed the estimates of LOB/LOD by up to 20-40%. It improved the accuracy of the results, and avoided the unduly optimistic estimation of these figures of merit. We implemented the non-linear regression approach in the open-source R-based software MSstats, and advocate its general use for mass spectrometric protein assay characterization.



LONGITUDINAL PROFILING FOR SYSTEM SUITABILITY AND QUALITY CONTROL



ACKNOWLEDGEMENTS

Northeastern University

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

ETH Zurich

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

University of Washington

Michael MacCoss
Brendan MacLean
Jarrett Egertson

Biognosis

Lukas Reiter

iPRG 2015

Zeynep Eren-Dogu
Chris Colangelo
John Cottrell
Michael Hopman
Eugene Kapp
Santa Kim
Henry Lam
Tom Neubert
Magnus Palmblad
Brett Phinney
Sue Weintraub,