DESIGN AND ANALYSIS OF QUANTITATIVE PROTEOMIC EXPERIMENTS

Introduction to statistical methods
and
practical examples
using Skyline, R and MSstats

Meena Choi
Northeastern University

Ting Huang
Northeastern University

Brendan MacLean
University of Washington

Birgit Schilling
Buck Institute for Aging

Olga Vitek
Northeastern University

https://github.com/MeenaChoi/USHUPO2017_shortcourse

SCHEDULE

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

STATISTICAL EXPERIMENTAL DESIGN

Olga Vitek

College of Science

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

- Challenges of reproducibility
  - Motivating example: iPRG 2015-2016
- Translate scientific question into statistics
  - Statistical terms for 'biomarker' (or 'signature')
- Experimental design
  - Replication, randomization, blocking

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

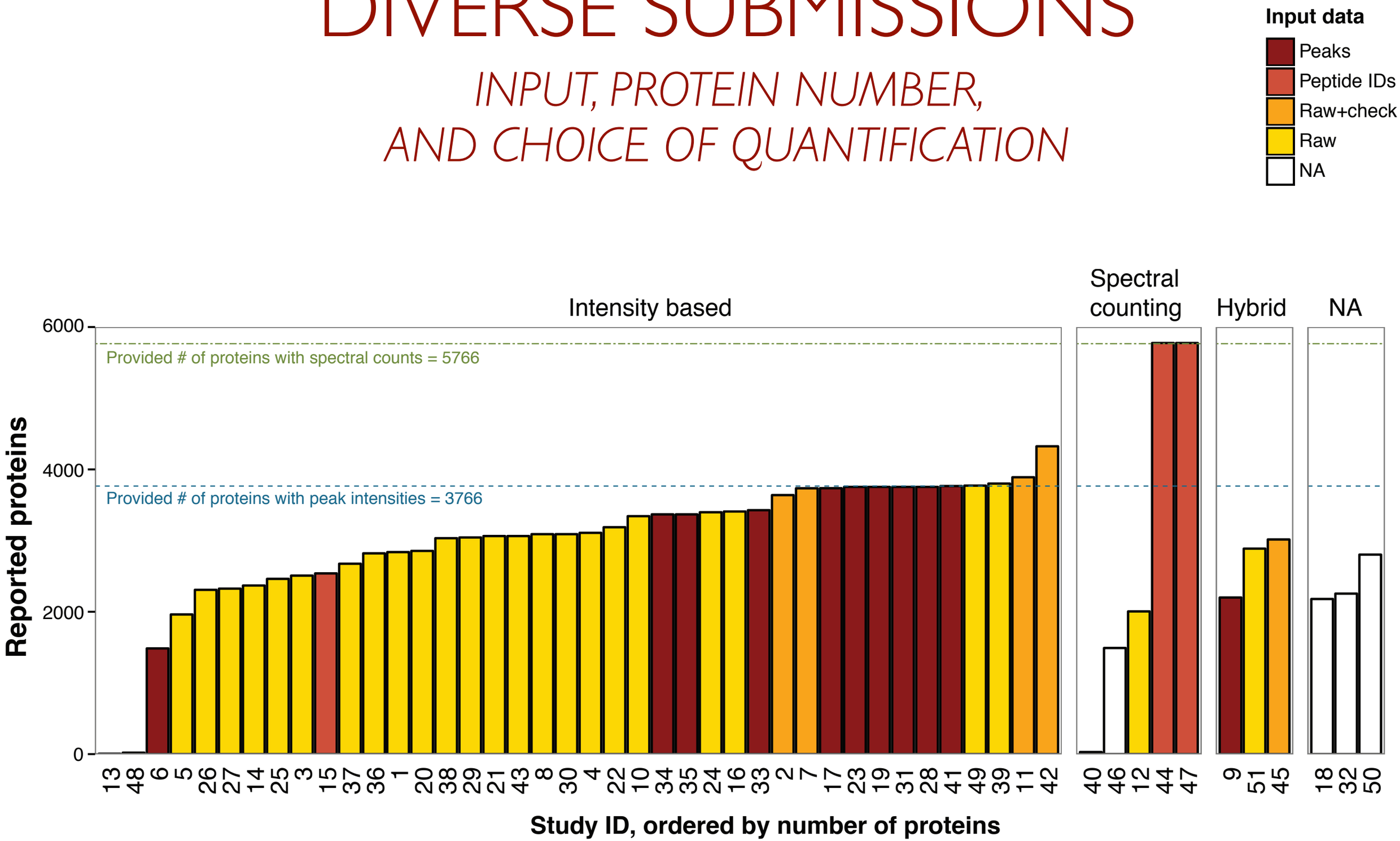
Spiked into a constant background: tryptic digests of S. cerevisiae

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

Choi et al., *Journal of Proteome Research,* 2017.

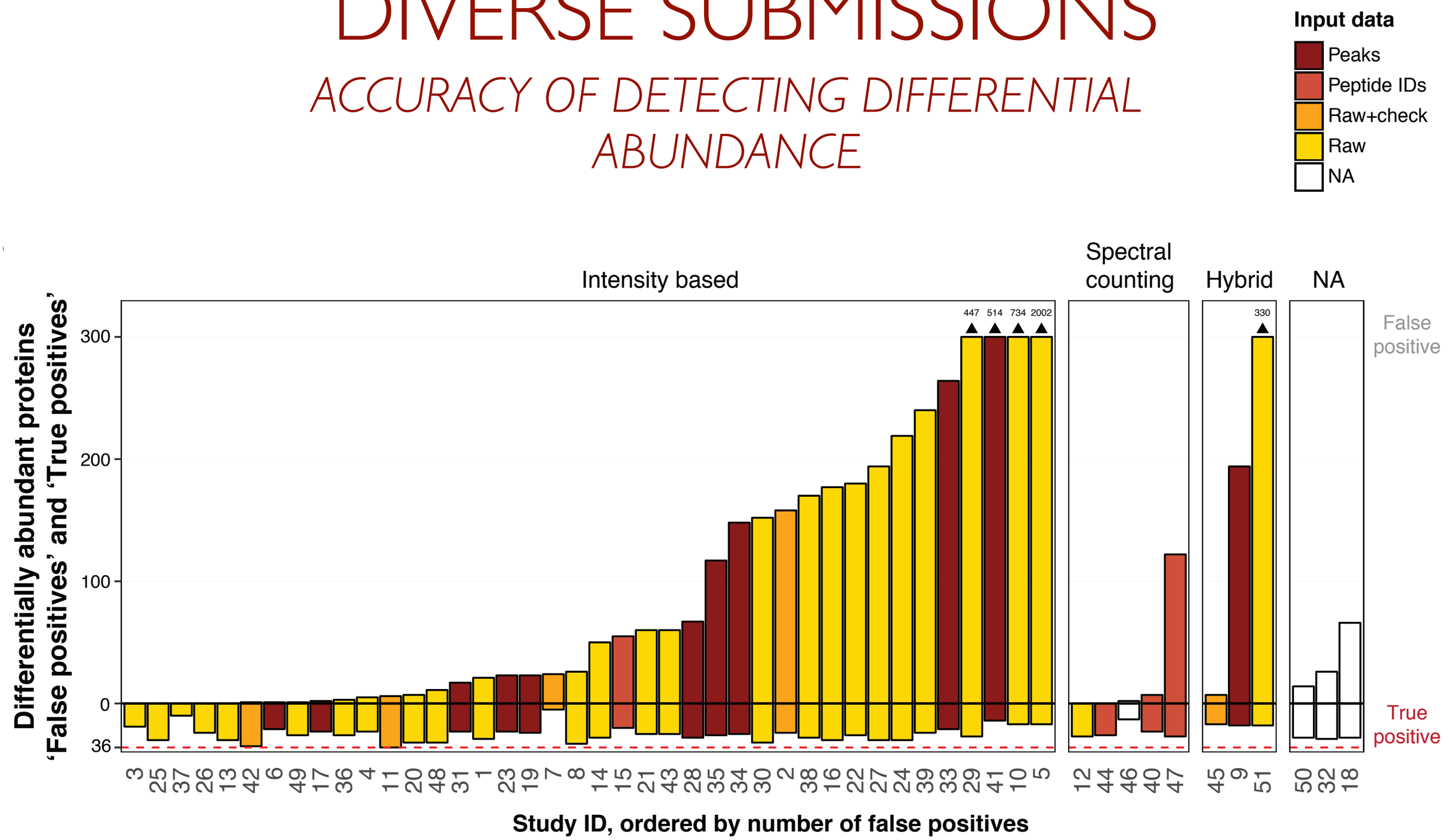
DIVERSE SUBMISSIONS

INPUT, PROTEIN NUMBER, AND CHOICE OF QUANTIFICATION



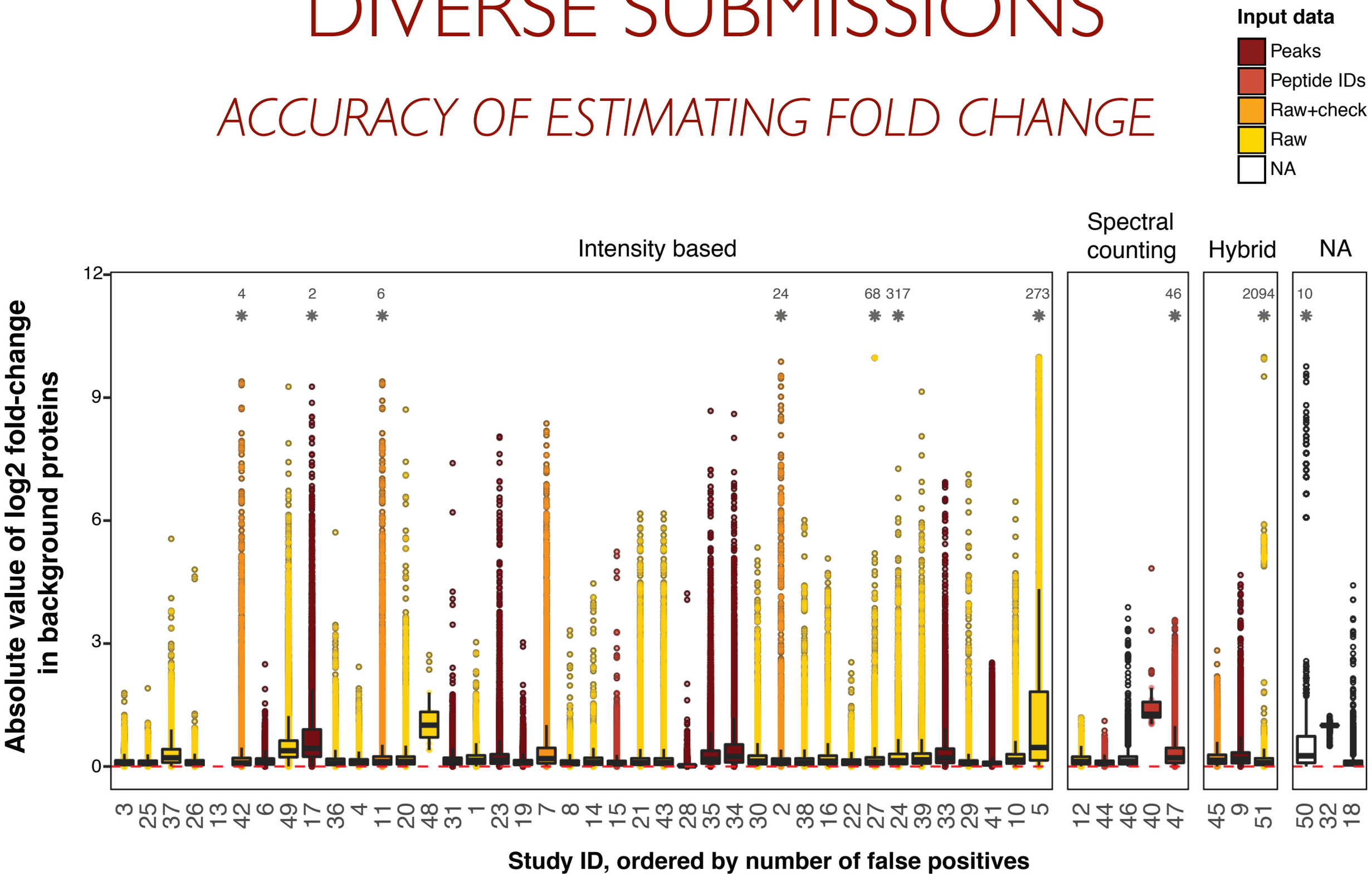
DIVERSE SUBMISSIONS

ACCURACY OF DETECTING DIFFERENTIAL ABUNDANCE



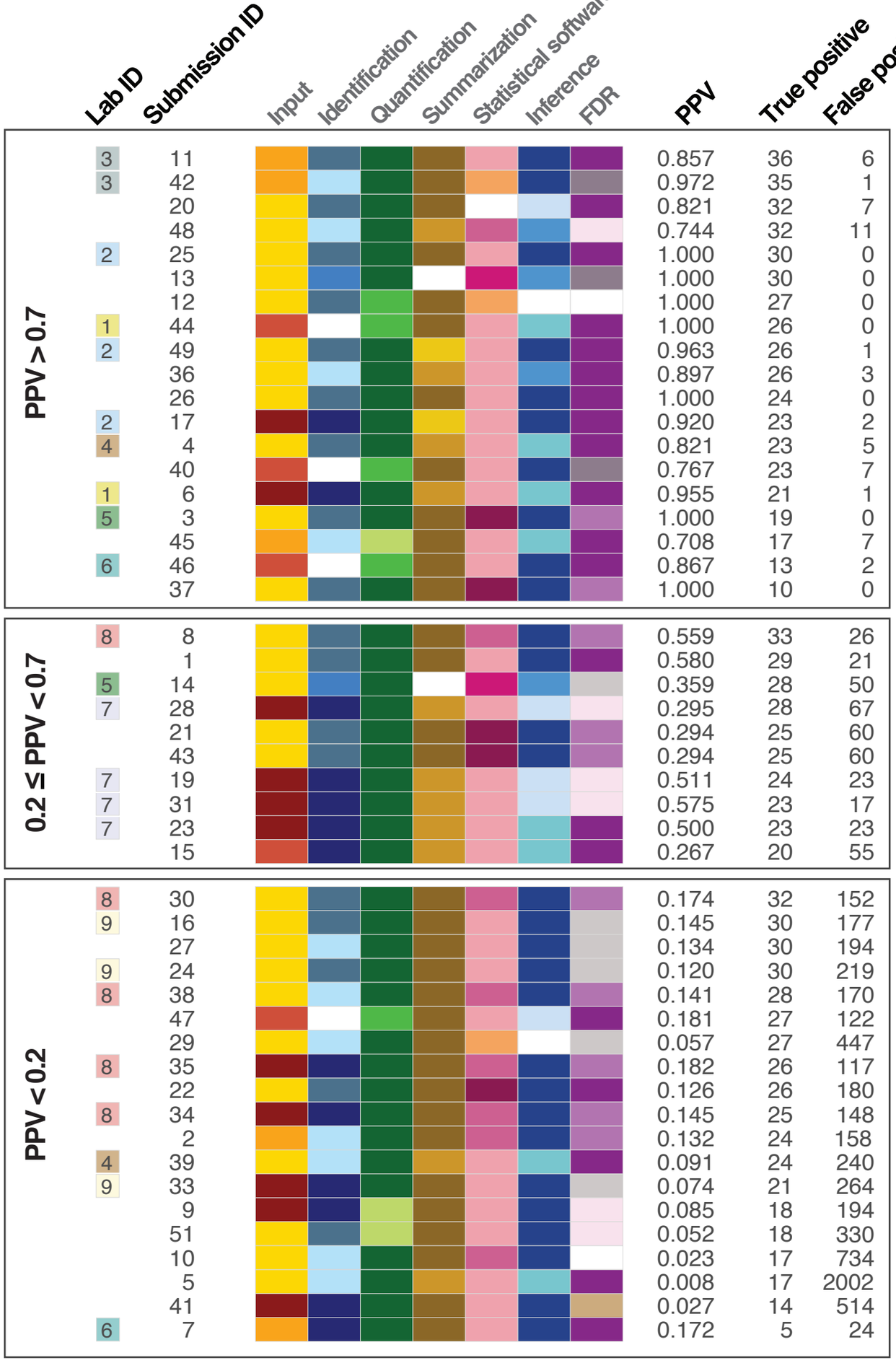
DIVERSE SUBMISSIONS

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 =

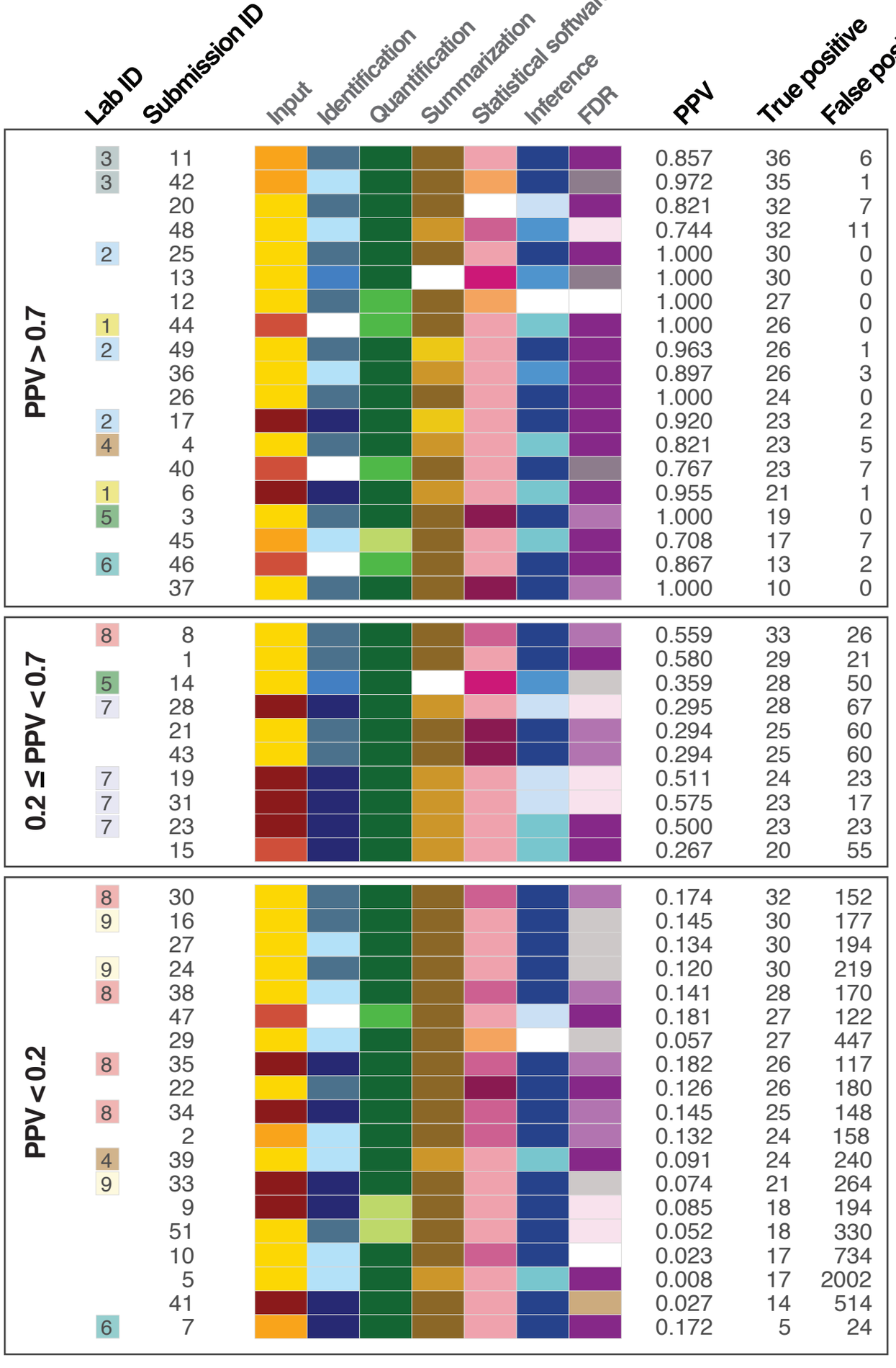
$$\frac{\text{\# true differentially abundant proteins}}{\text{\# claimed differentially abundant proteins}}$$

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

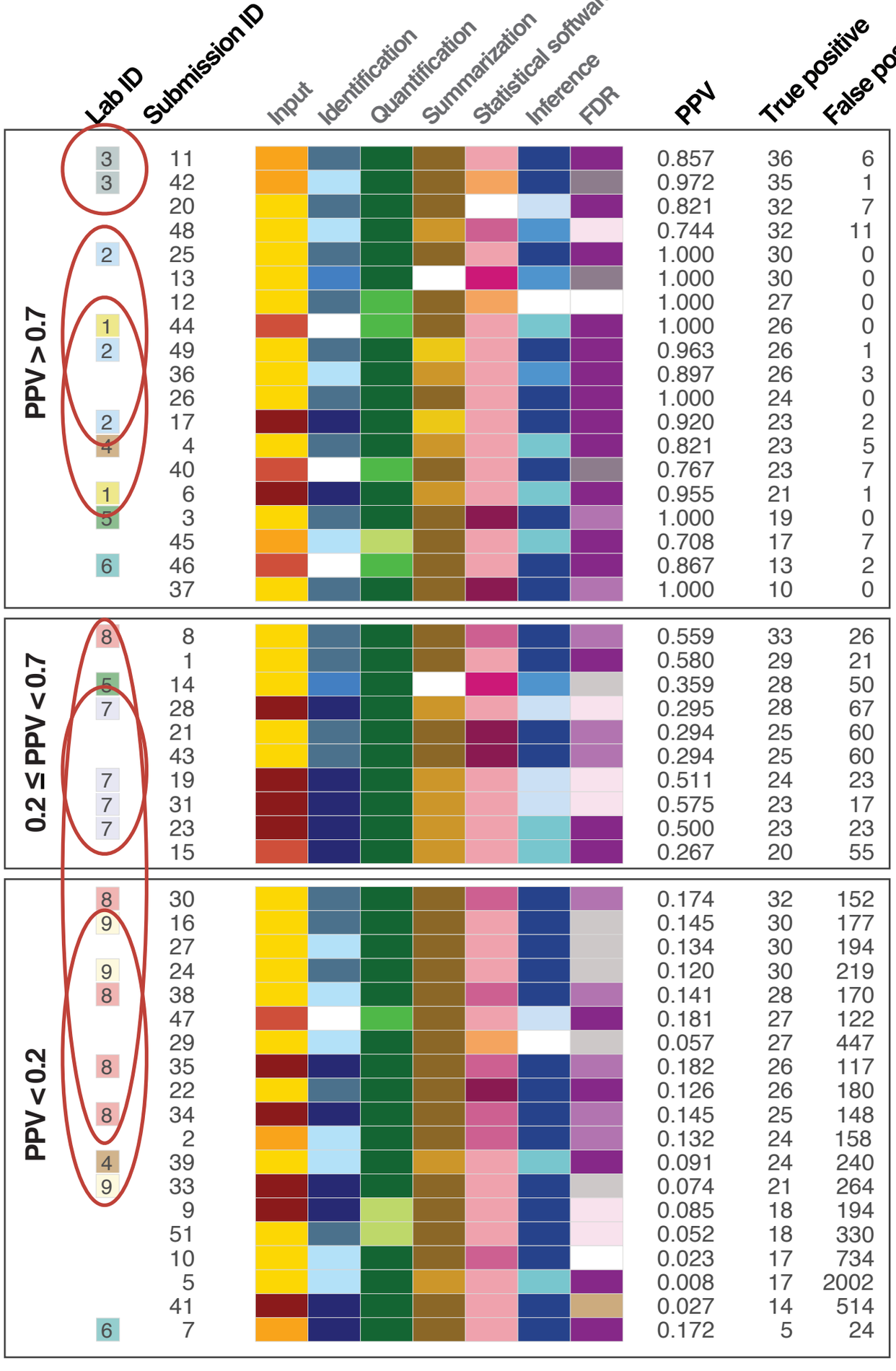
Skyline and linear modeling in R

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

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∥, 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
∥ Matrix Science Ltd., London W1U 7QB, 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.

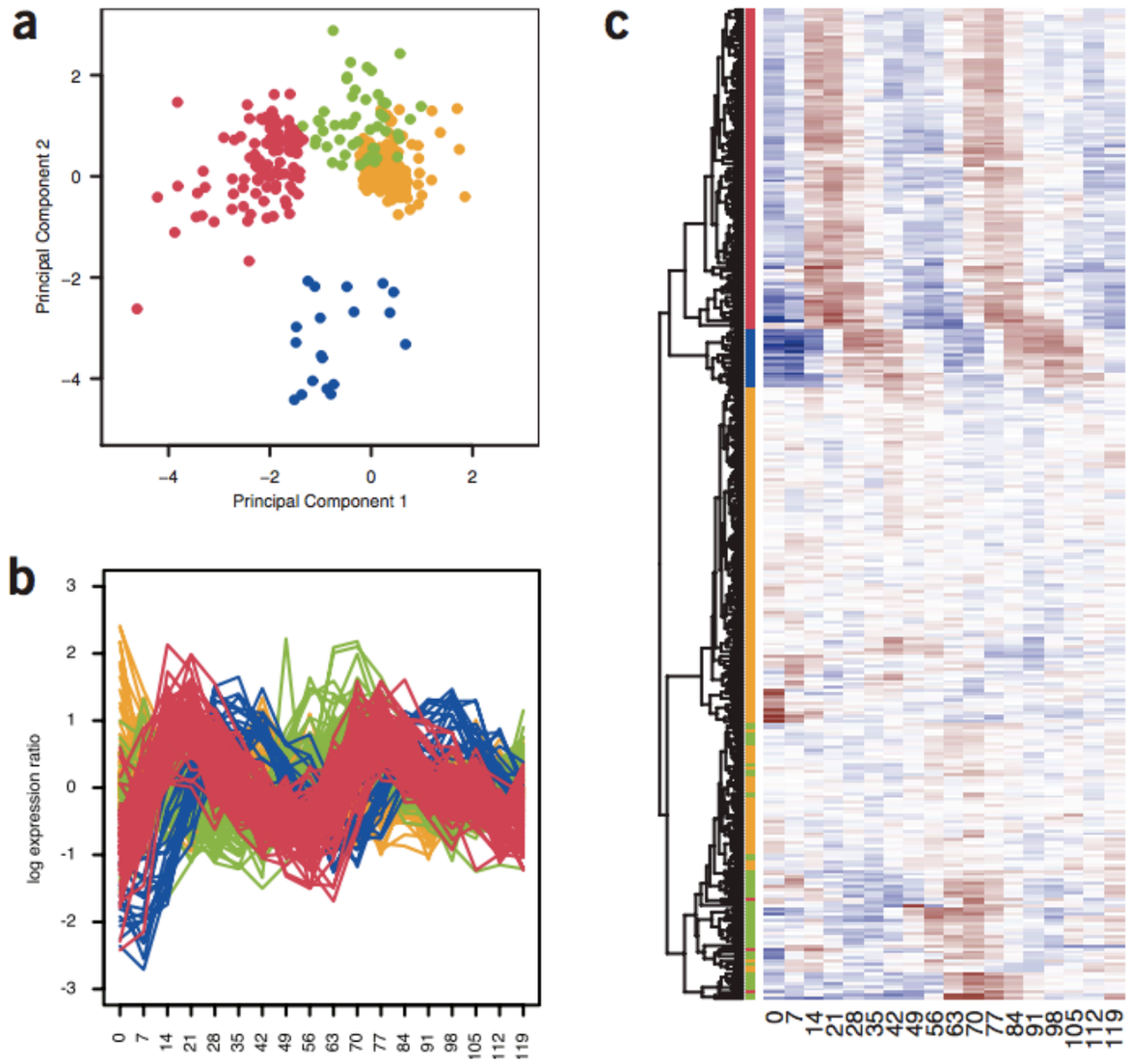
OUTLINE

- Challenges of reproducibility
  - Motivating example: iPRG 2015-2016
- Translate scientific question into statistics
  - Statistical terms for 'biomarker' (or 'signature')
- Experimental design
  - Replication, randomization, blocking

STATISTICAL GOAL 1: CLASS DISCOVERY

Discover proteins or subjects with similar patterns

- No known class labels
  - E.g., no 'healthy' or 'disease'
  - All variation treated equally
  - No error rates
- Can't find something meaningful if unsure what we look for
  - Best used for visualization

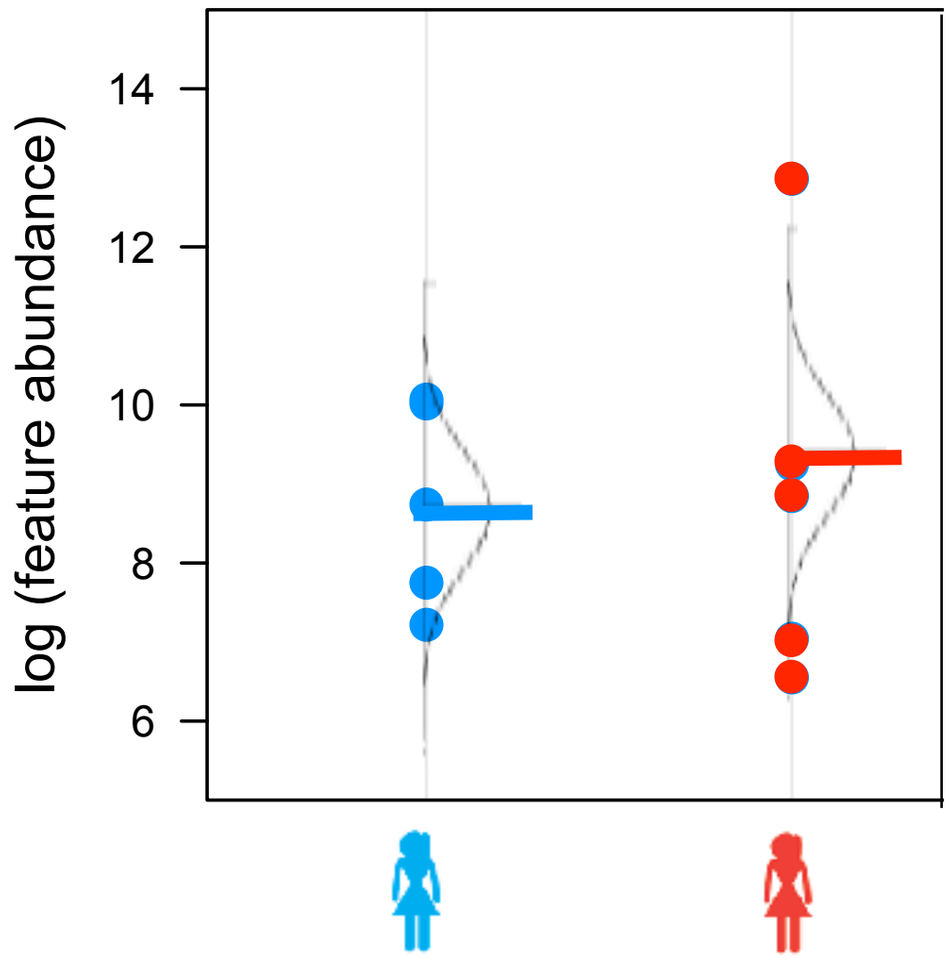


Gehlenborg *et al*, Nature Methods, 2010

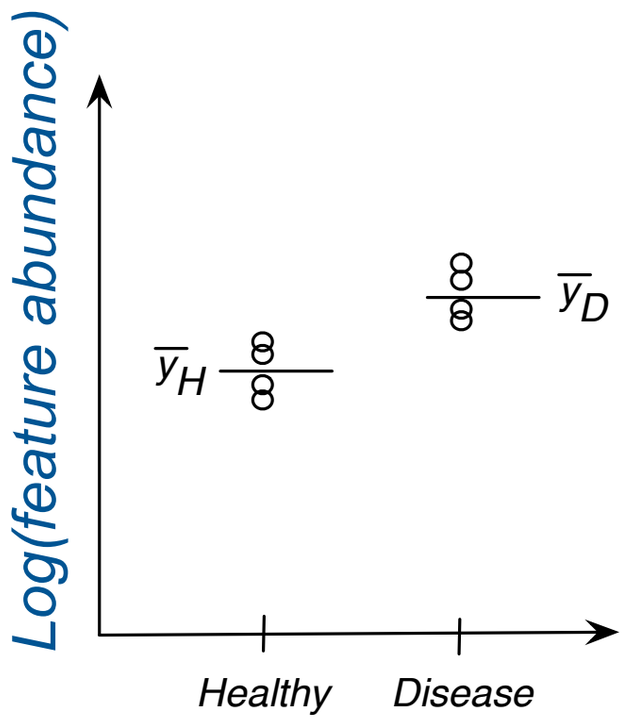
STATISTICAL GOAL 2: CLASS COMPARISON

Compare mean abundances in subject groups

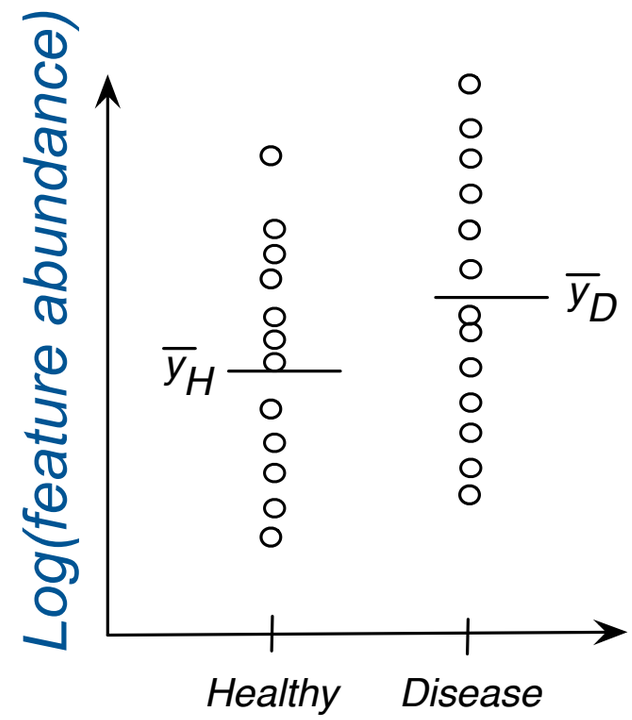
- Known class labels
  - Compare group averages
  - Report p-values, posterior probabilities etc
- Useful when compare groups of subjects
  - Best used for basic biology
  - Initial (Tier III) biomarker discovery screen



DIFFERENTIALLY ABUNDANT PROTEINS ARE NOT ALWAYS BIOMARKERS



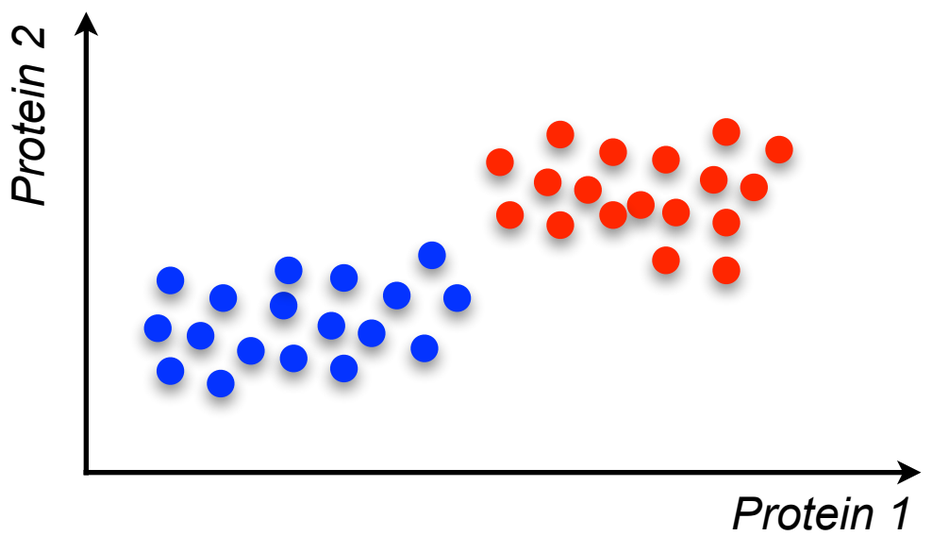
Differentially abundant and predictive



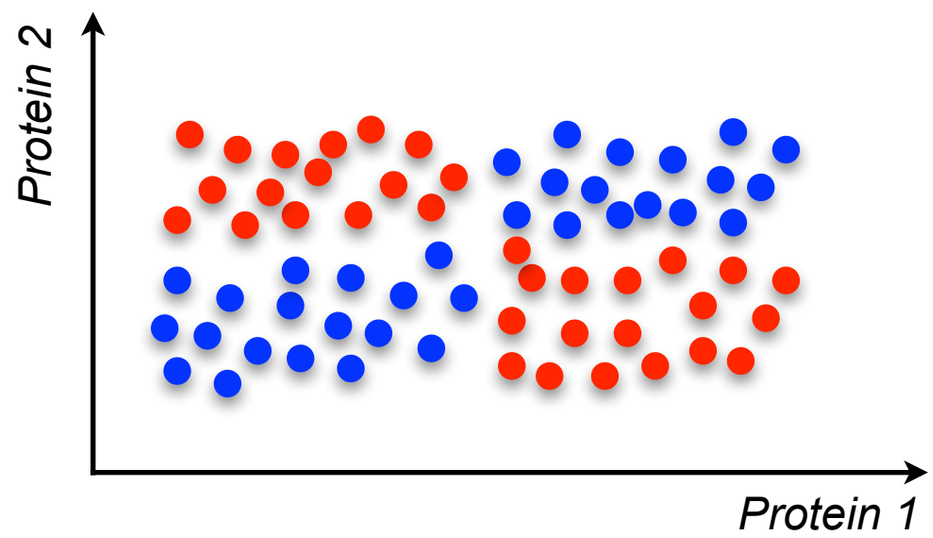
Differentially abundant and not predictive

Single protein: *Differentially abundant* → ✗ *Predictive*

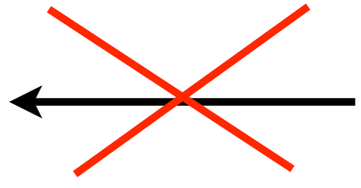
BIOMARKER PROTEINS ARE NOT ALWAYS DIFFERENTIALLY ABUNDANT



Differentially abundant and predictive



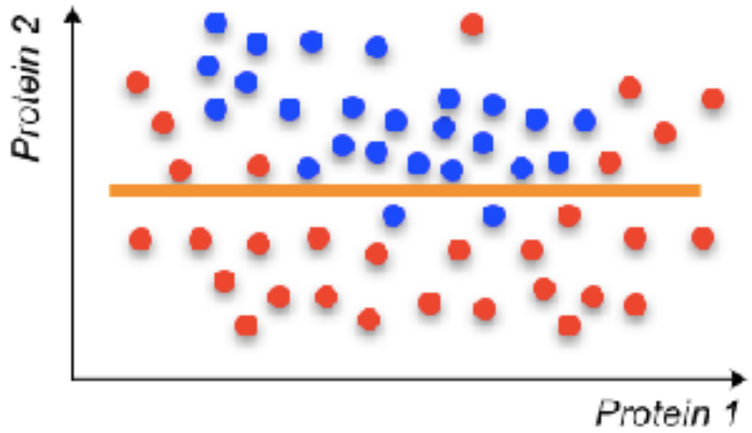
Not differentially abundant but predictive

Single protein: *Differentially abundant*  *Predictive*

STATISTICAL GOAL 3: CLASS PREDICTION

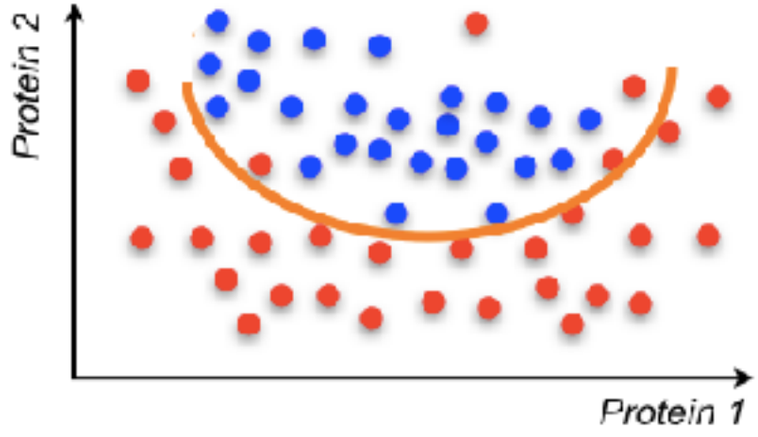
Classify each subject into a known group

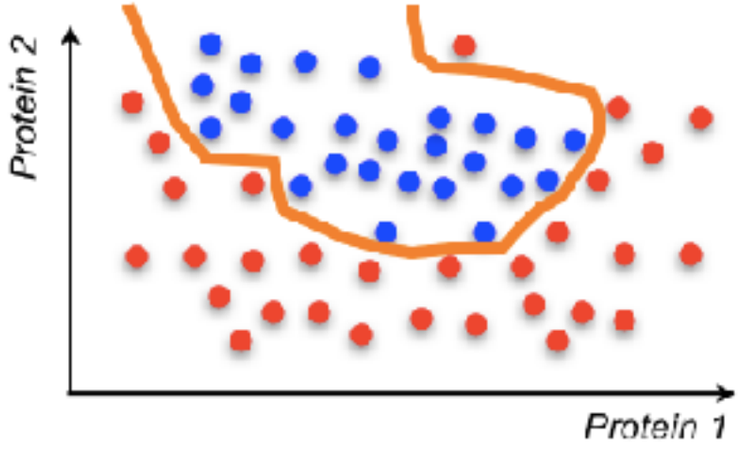
- Known class labels
  - Predict individual subjects
  - Report misclassification error (sensitivity, specificity, predictive value etc)
- Useful when focus on an individual
  - Tier I or Tier II biomarker discovery studies



Linear prediction rule

Quadratic prediction rule



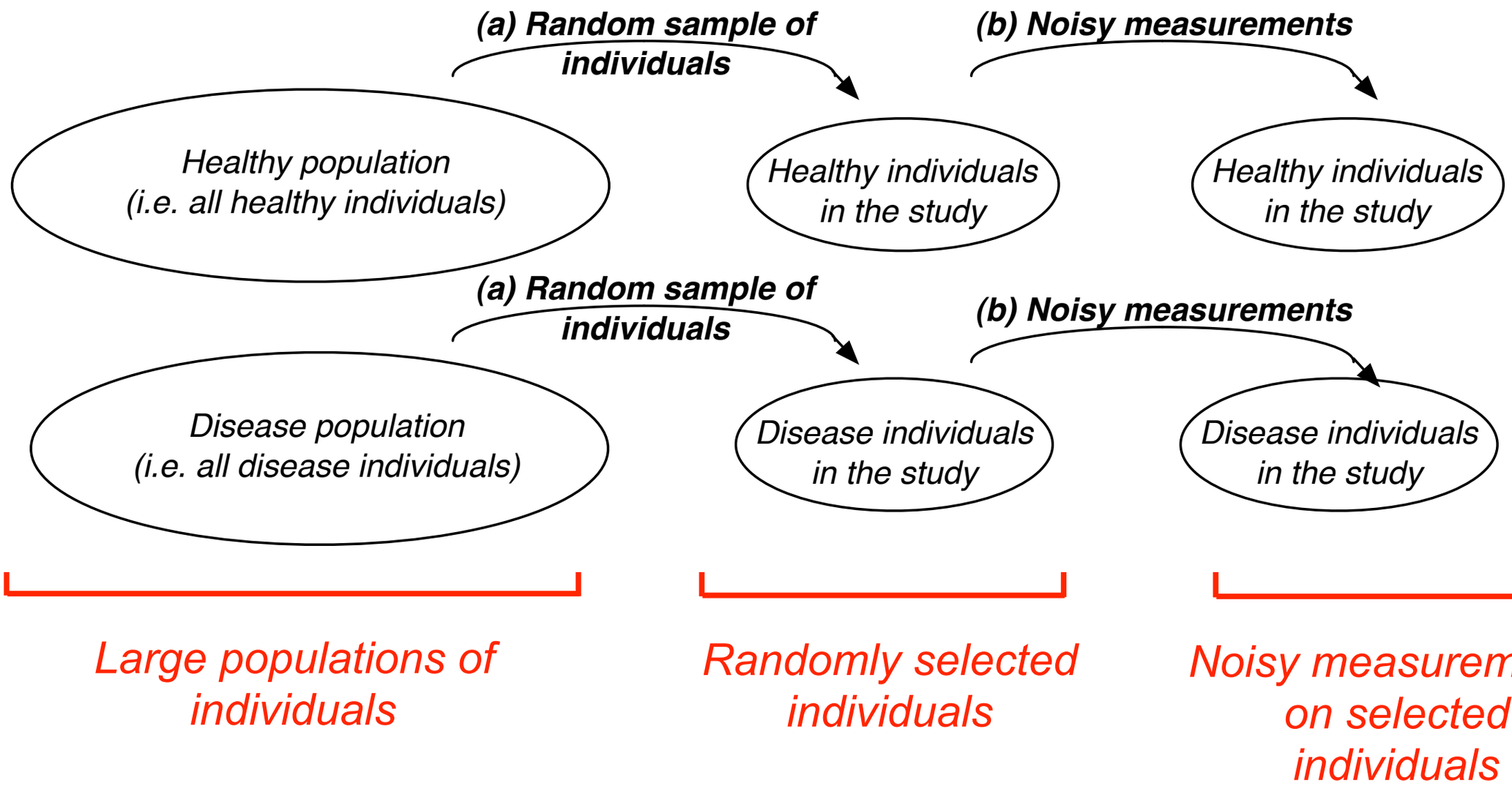


General prediction rule

OUTLINE

- Challenges of reproducibility
  - Motivating example: iPRG 2015-2016
- Translate scientific question into statistics
  - Statistical terms for 'biomarker' (or 'signature')
- Experimental design
  - Replication, randomization, blocking

A STATISTICIAN'S VIEW OF THE EXPERIMENT

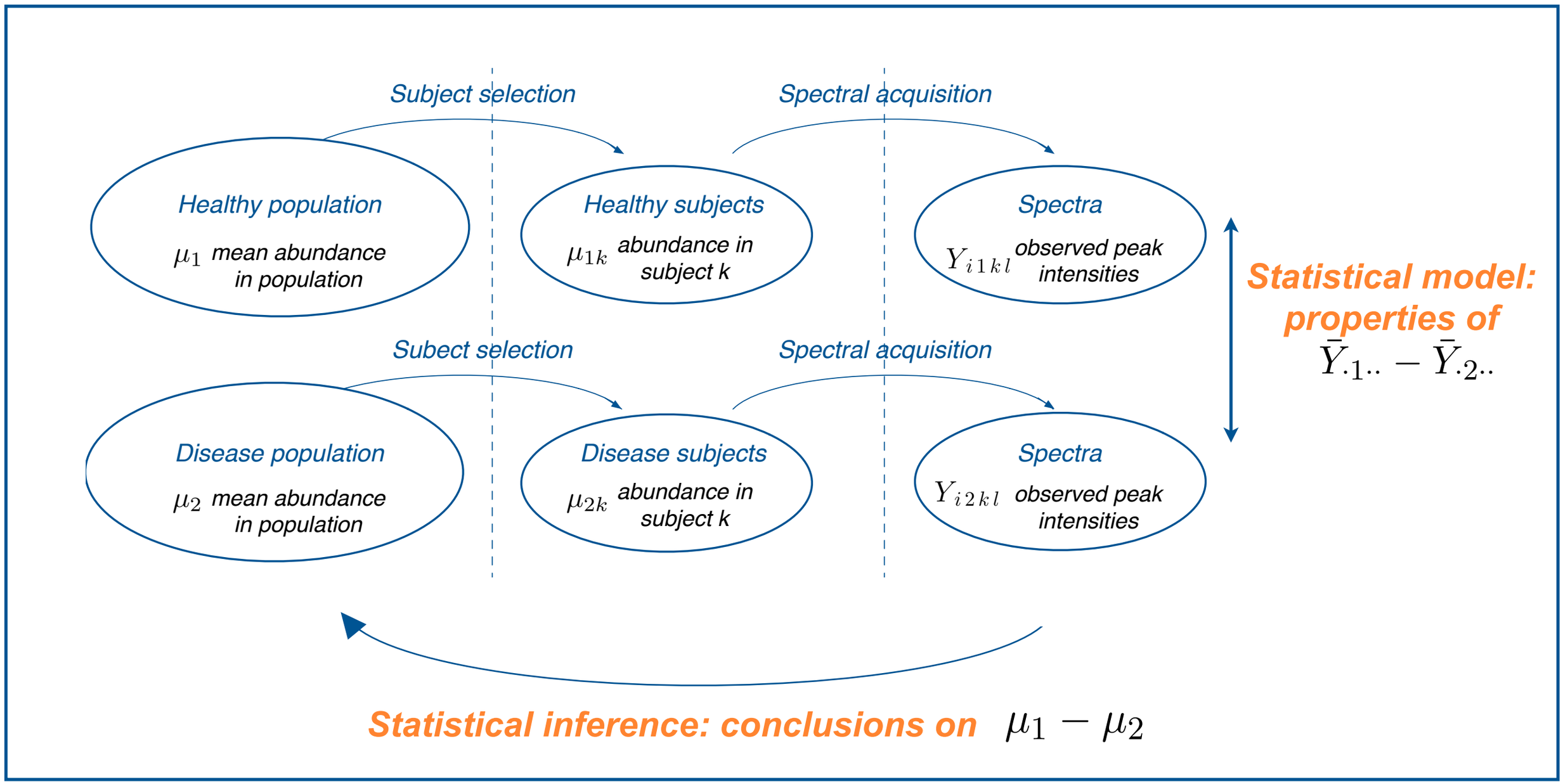


Dangers:

Bias: conclusions systematically differ from truth

Inefficiency: unnecessary variation in the data

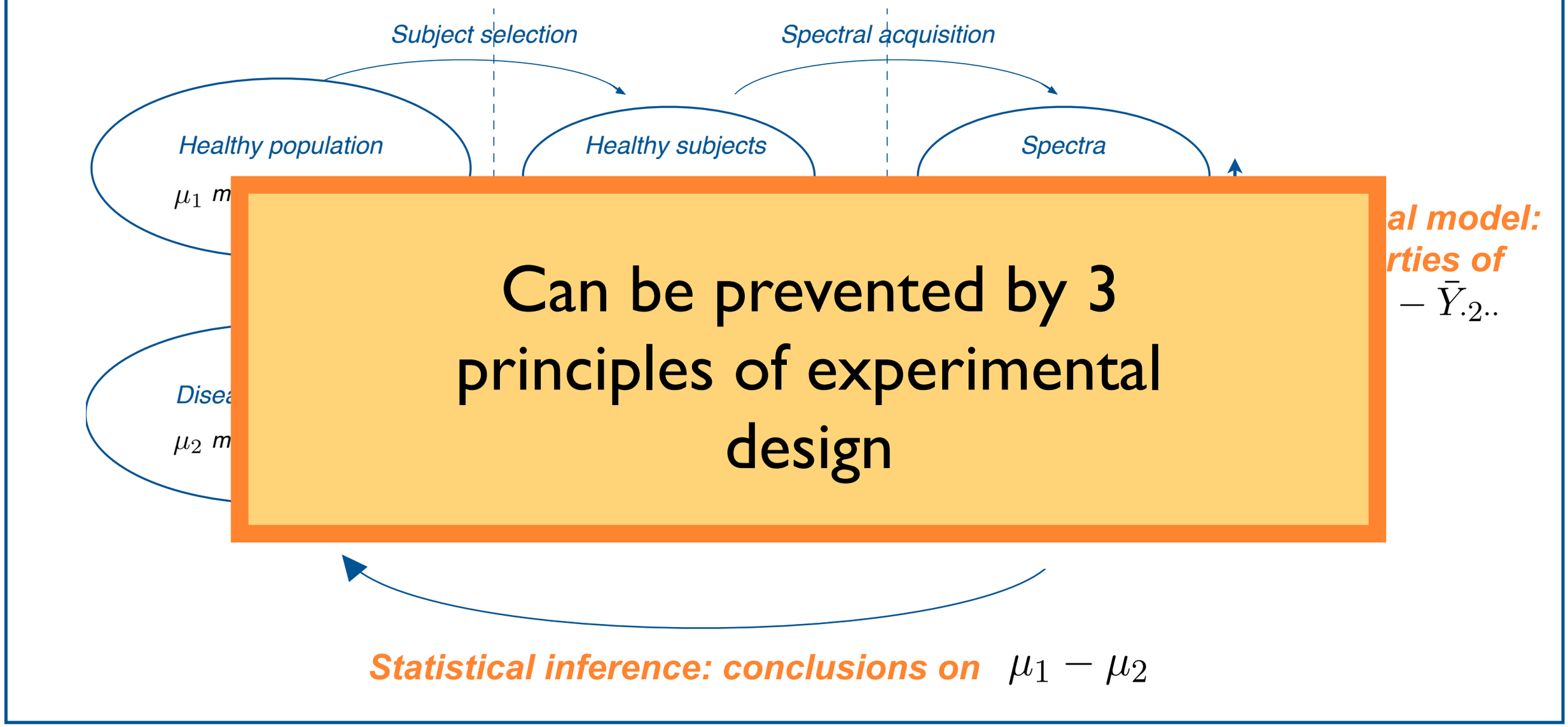
DEFINITION OF BIAS AND INEFFICIENCY



Bias: $\bar{Y}_{.1..} - \bar{Y}_{.2..}$ *systematically different from* $\mu_{1k} - \mu_{2k}$

Inefficiency: *Large* $Var(\bar{Y}_{.1..} - \bar{Y}_{.2..})$

DEFINITION OF BIAS AND INEFFICIENCY

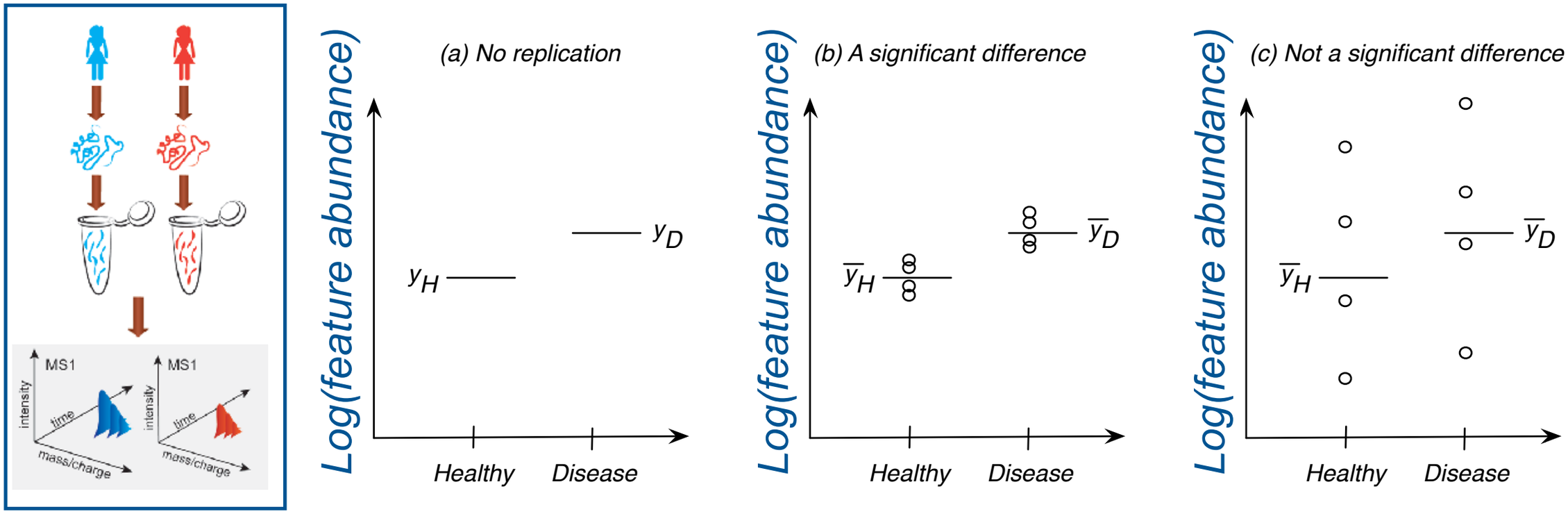


Bias: $\bar{Y}_{.1..} - \bar{Y}_{.2..}$ *systematically different from* $\mu_{1k} - \mu_{2k}$

Inefficiency: *Large* $Var(\bar{Y}_{.1..} - \bar{Y}_{.2..})$

PRINCIPLE 1: REPLICATION

(1) carries out the inference and (2) minimizes inefficiencies



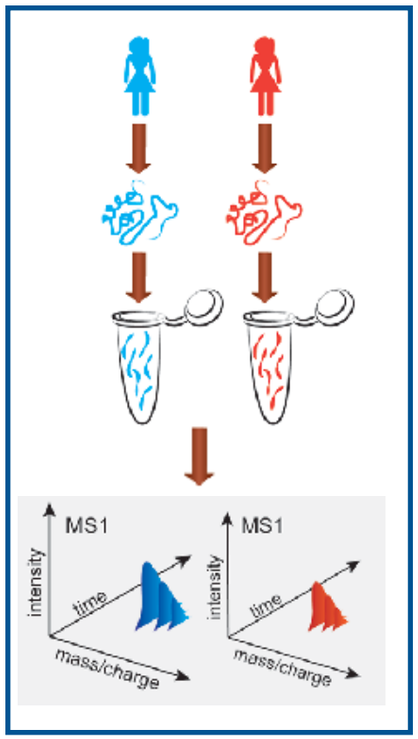
Two levels of randomness imply two types of replication:

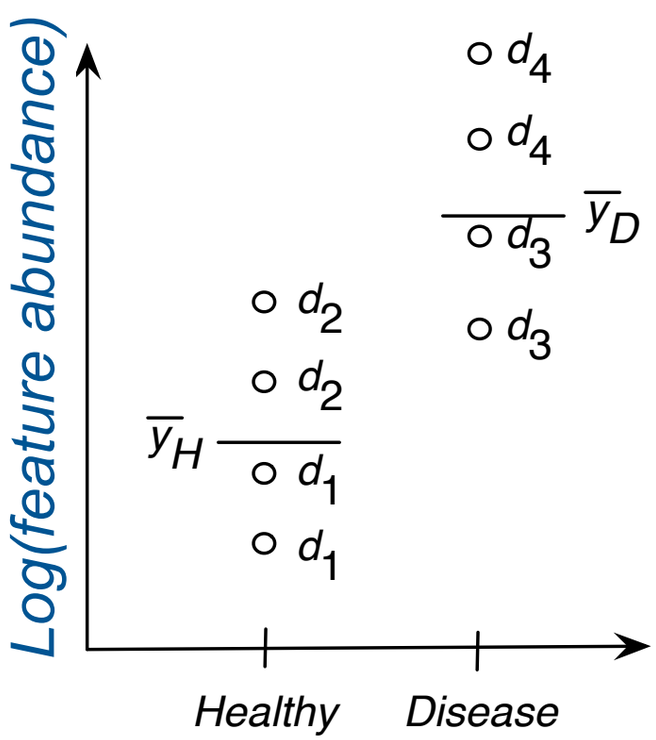
- *Biological replicates:* selecting multiple subjects from the population
- *Technical replicates:* multiple runs per subject

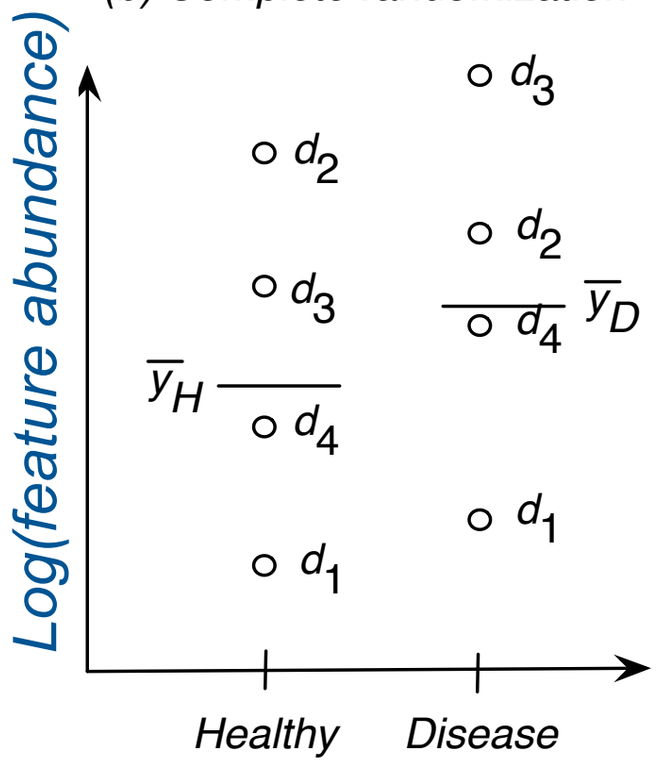
Oberg and Vitek, *J. Proteome Research*, 8, 2009

PRINCIPLE 2: RANDOMIZATION

Prevents bias







No randomization
= confounding
= bias

Complete randomization
= no bias

Two levels of randomness imply two types of randomization:

- *Biological replicates:* random selection of subjects from the population
- *Technical replicates:* random allocation of samples to all processing steps

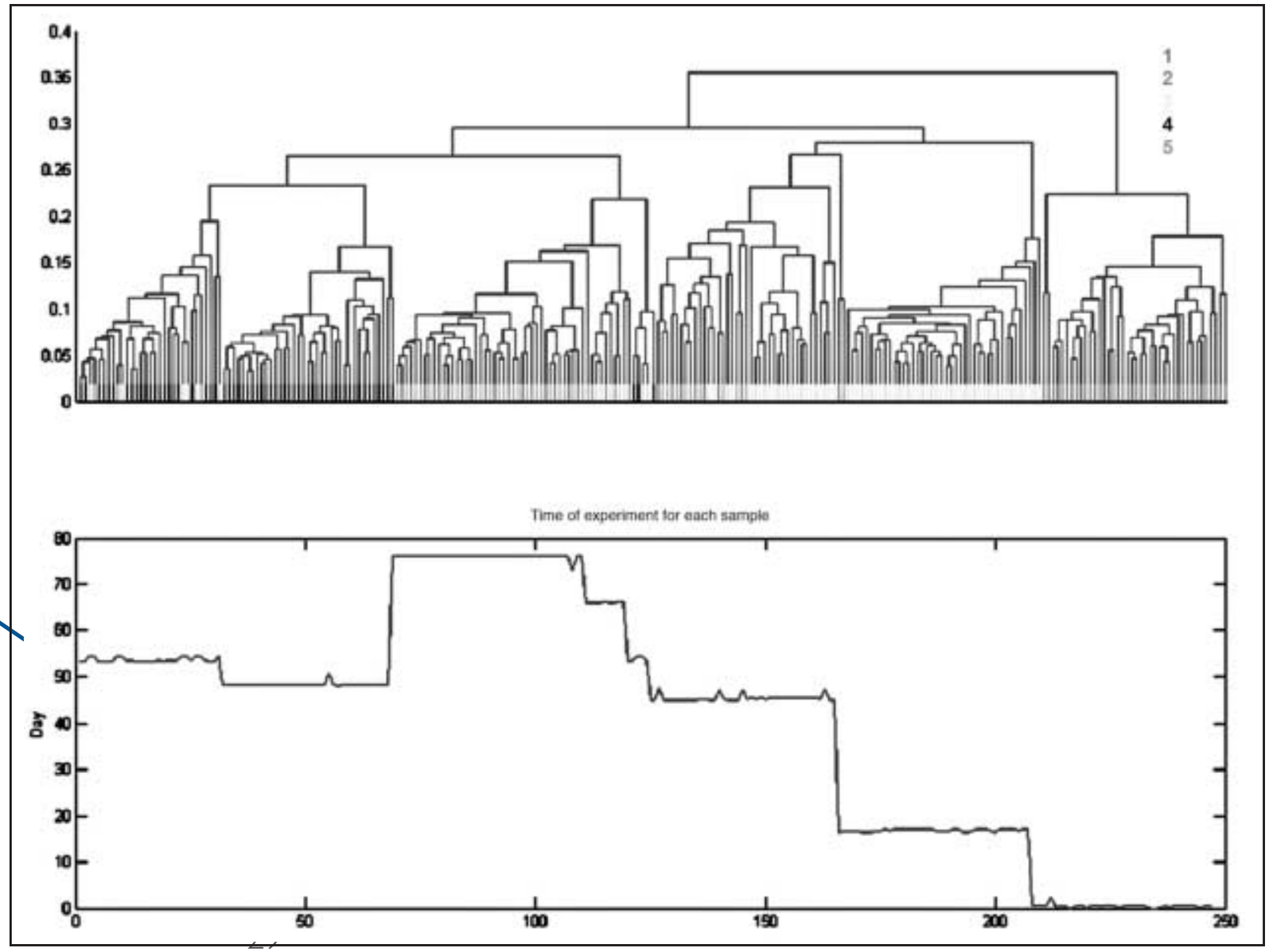
Oberg and Vitek, *J. Proteome Research*, 8, 2009

EXAMPLE: LACK OF RANDOMIZATION

Hu, Coombes, Morris, Baggerly, Briefings in Functional Genomics, 2005

- Serum samples with five types of cancer
- SELDI-TOF MS
  - normalized, peak picked

Hierarchical clustering of samples



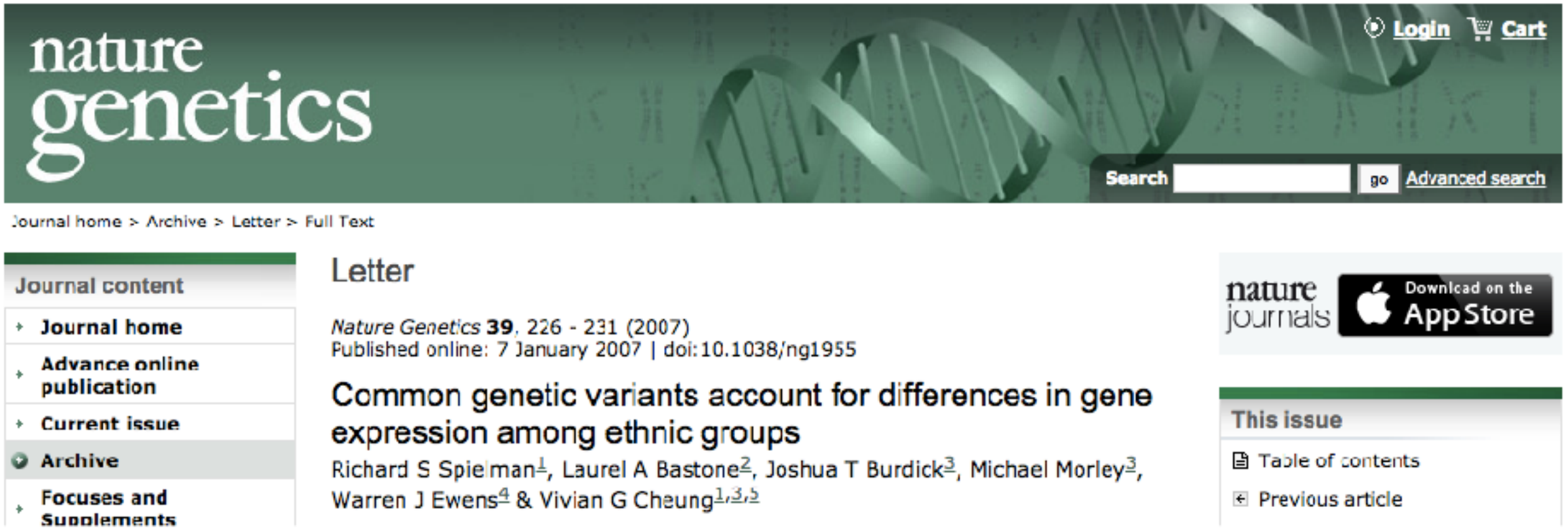
Cancer subtype confounded with time

Same time-based clustering on the QC samples!

BEWARE OF BIG EFFECTS

THEY ARE LIKELY TO REFLECT FLAWS OF THE DESIGN

- Study of gene expression between Asians and Europeans
- Found that 78% of genes are differentially
  - Asians were profiles in one year, and Europeans in another
  - The difference therefore likely reflects a batch effect

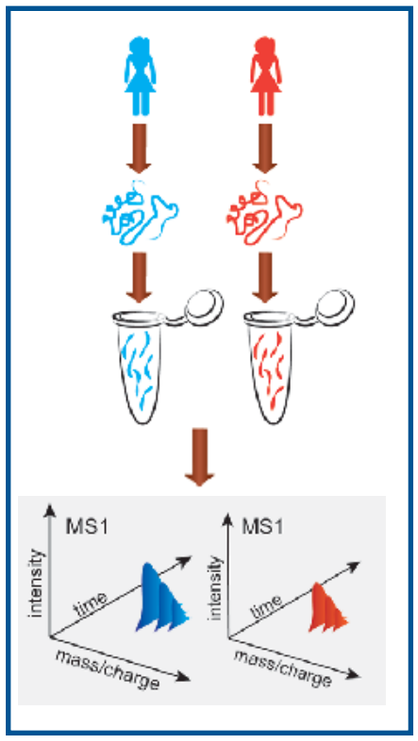


Source: a blog by Jeff Leek, Biostatistics, John Hopkins University

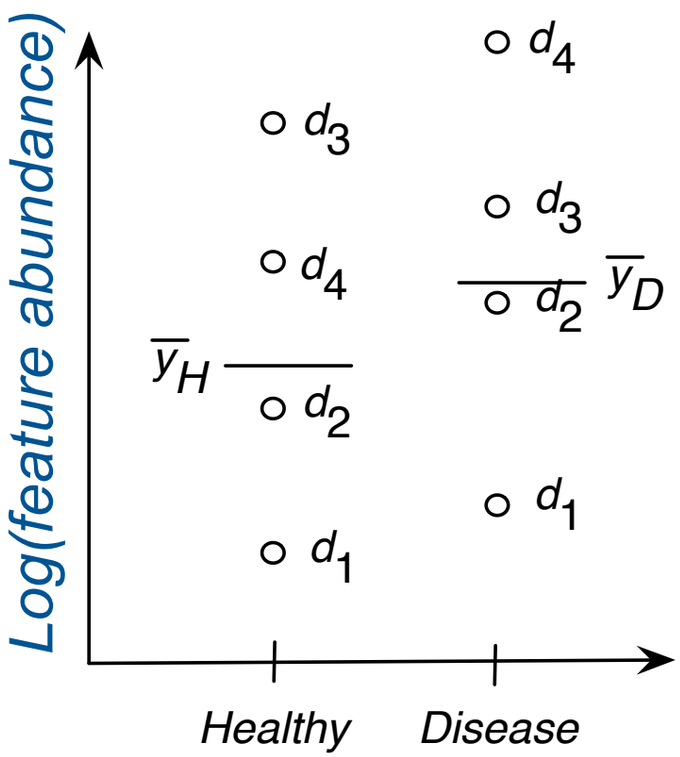
http://simplystatistics.org/2016/02/01/a-menagerie-of-messed-up-data-analyses-and-how-to-avoid-them/

PRINCIPLE 3: BLOCKING

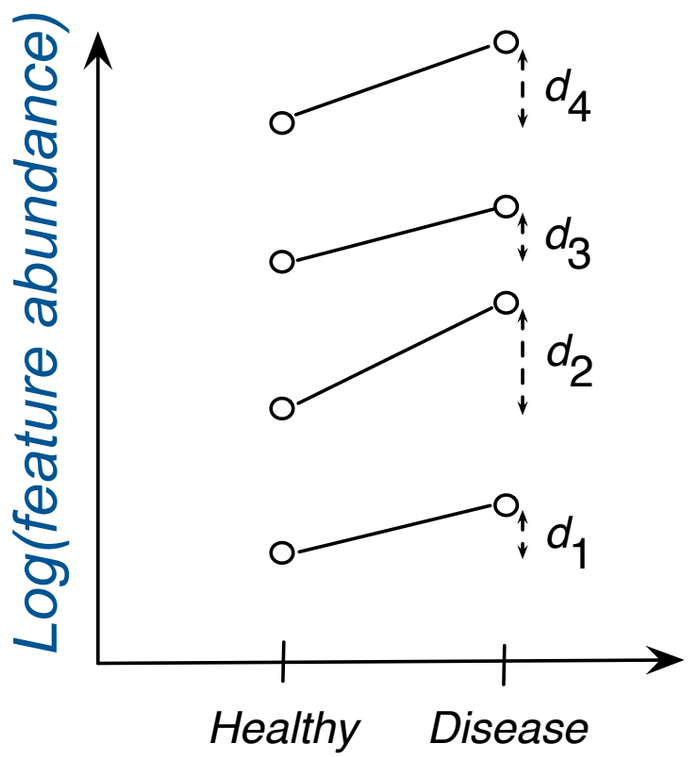
Helps reduce both bias and inefficiency



(b) Complete randomization



(c) Day = block



Complete randomization
= inflated variance

Block-randomization
= restriction on randomization
= systematic allocation

Two levels of randomness imply two types of blocks:

- *Biological replicates:* subjects having similar characteristics (e.g. age)
- *Technical replicates:* samples processed together (e.g. in a same day)

Oberg and Vitek, *J. Proteome Research*, 8, 2009

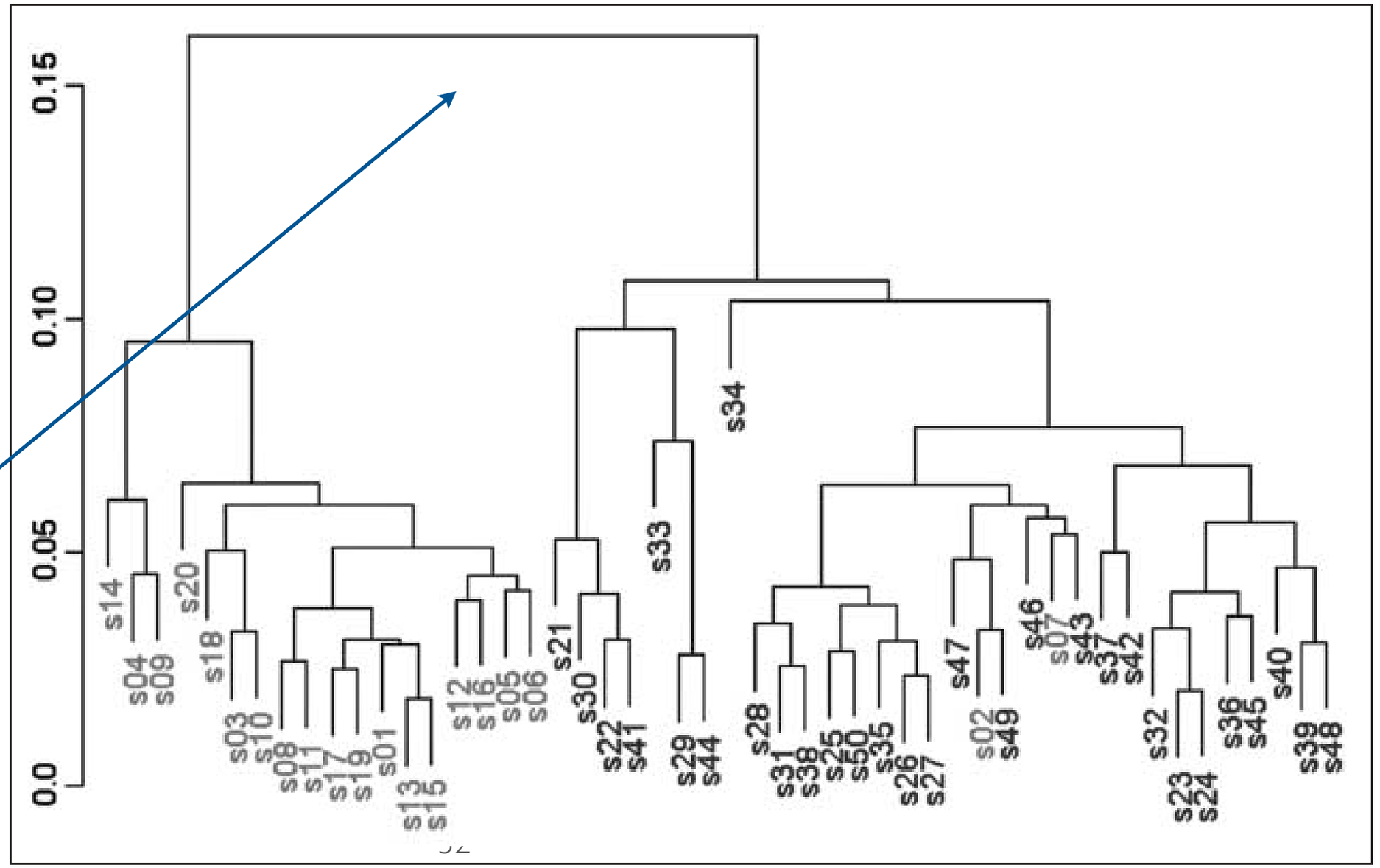
EXAMPLE: LACK OF BLOCKING

Hu, Coombes, Morris, Baggerly, Briefings in Functional Genomics, 2005

- Serum samples with two types of cancer
- SELDI-TOF MS, 3 fractions
  - normalized, peak picked

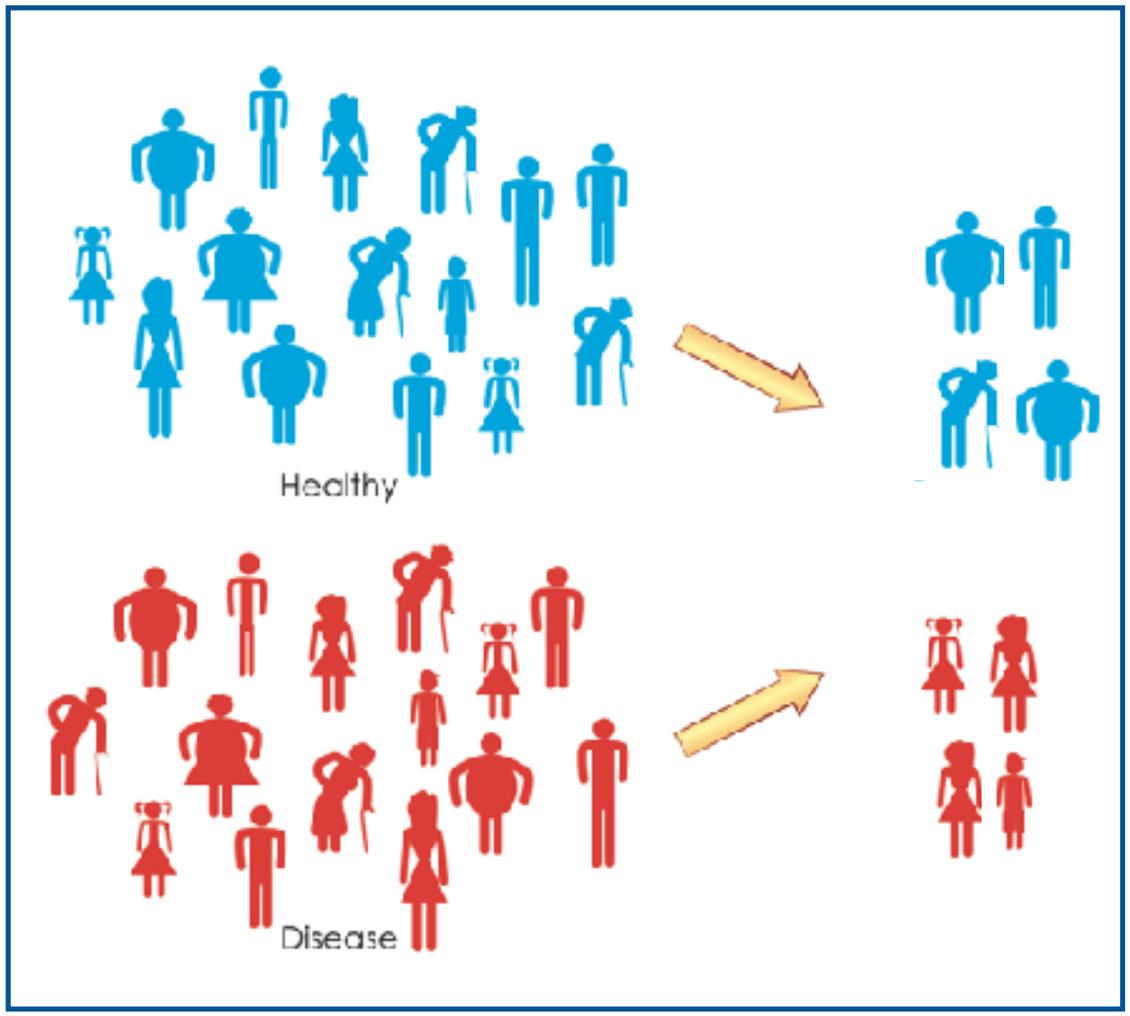
Hierarchical clustering of samples

Protocol change

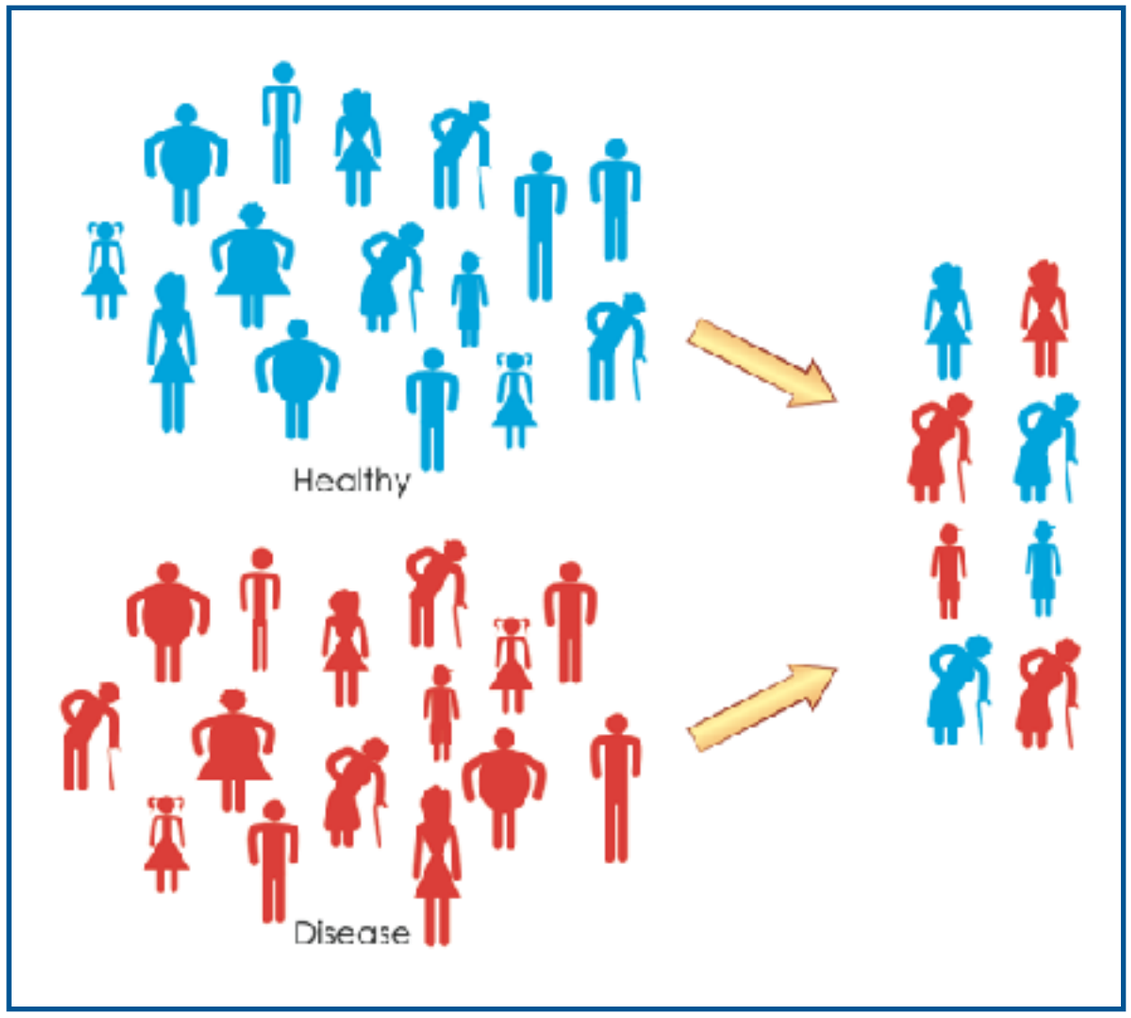


MATCHING

Blocking with respect to biological risk factors



Complete randomization
= inflated variance



Block-randomization
= restriction on randomization
= systematic allocation

Käll and Vitek, *PLoS Computational Biology*, 7, 2011

EXAMPLE

Block-randomized selection of subjects from repository

		Disease group				
		Control	Stable angina	Unstable angina	NSTEMI	STEMI
Stratification	$\geq$ 58 y.o; Female	354	300	49	39	29
	$\geq$ 58 y.o; Male	701	843	143	86	54
	$<$ 58 y.o; Female	80	56	5	5	8
	$<$ 58 y.o; Male	264	190	34	23	27

Counts in the initial repository of samples

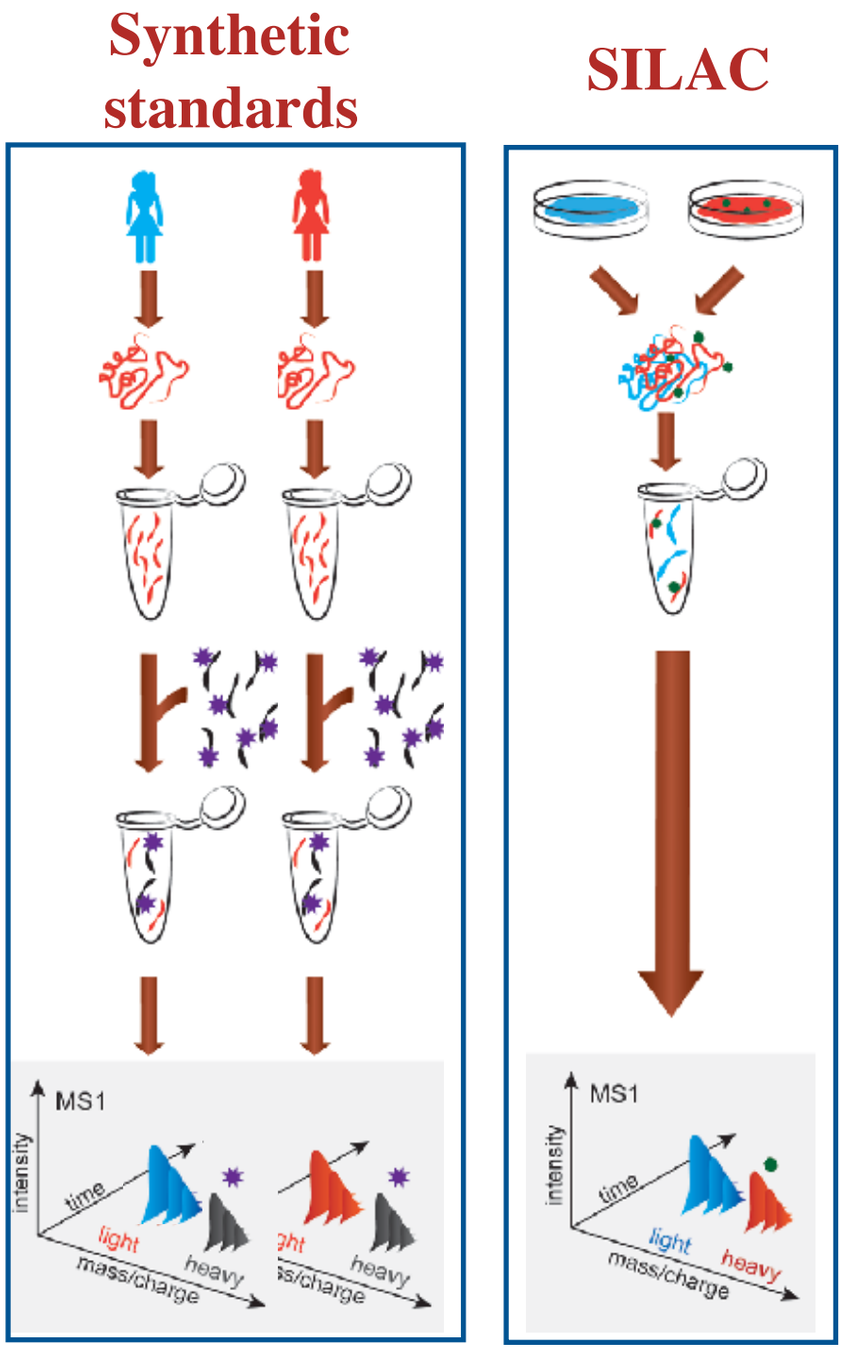
		Disease group				
		Control	Stable angina	Unstable angina	NSTEMI	STEMI
Stratification	$\geq$ 58 y.o; Female	3	3	3	3	3
	$\geq$ 58 y.o; Male	3	3	3	3	3
	$<$ 58 y.o; Female	2	2	2	2	2
	$<$ 58 y.o; Male	2	2	2	2	2

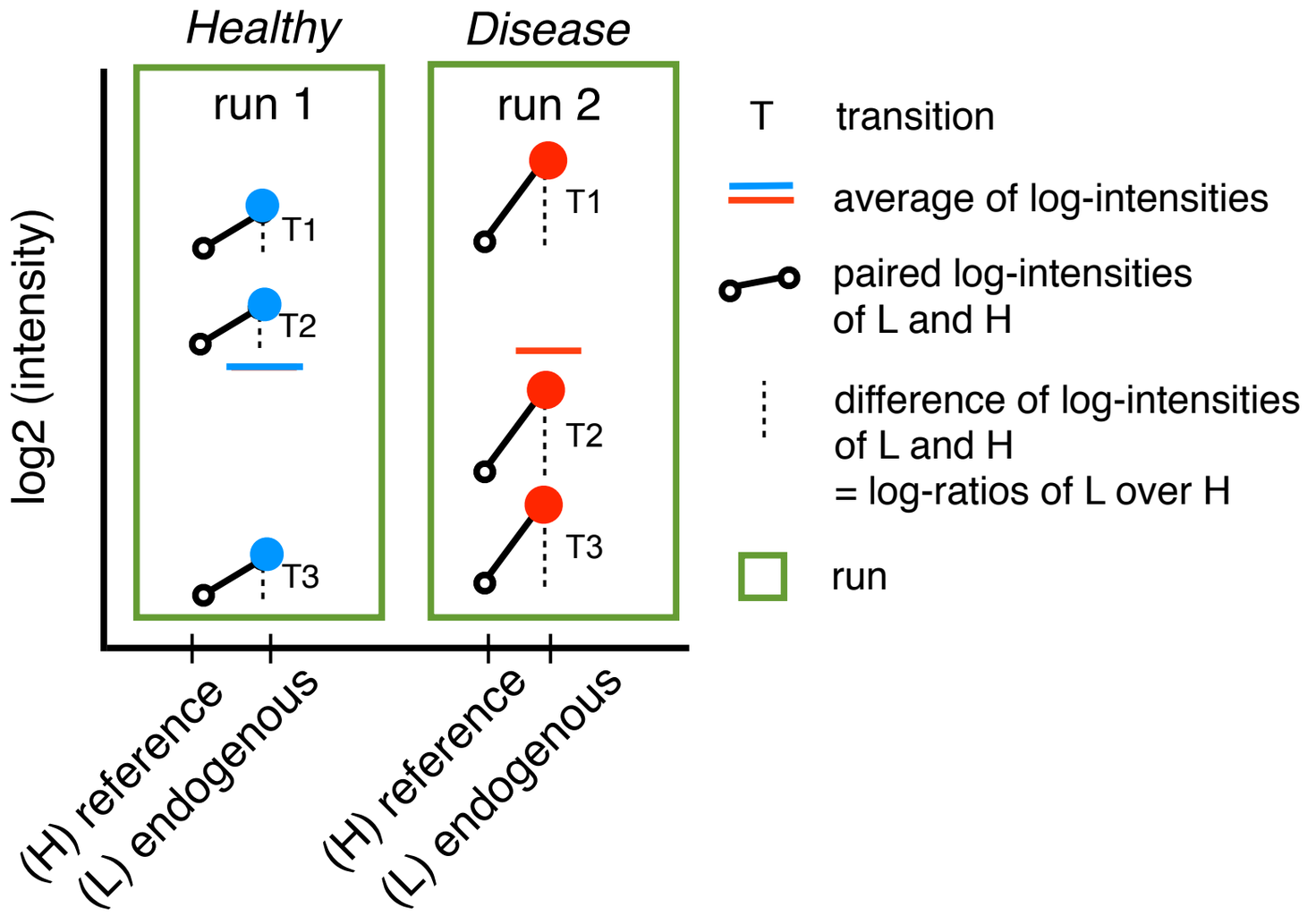
Counts of subjects included in the study

Mass spectra acquired without technical replication

LABELING (=MULTIPLEXING)

Blocking with respect to mass spectrometry run





Multiplexing reduces both bias and variance

(assuming that extra sample handling does not introduce extra variation)

Martin Krzywinski & Naomi Altman

nature methods
Techniques for life scientists and chemists

NATURE METHODS | THIS MONTH

Points of significance: Importance of being uncertain

Martin Krzywinski & Naomi Altman

Affiliations

Nature Methods **10**, 809–810 (2013) | doi:10.1038/nmeth.2613

Points of significance: Comparing samples—part I

Martin Krzywinski & Naomi Altman

Points of Significance: Error bars

Martin Krzywinski & Naomi Altman

Affiliations

Nature Methods **10**, 921–922 (2013) | doi:10.1038/nmeth.2659

Points of significance: Power and sample size

Martin Krzywinski & Naomi Altman

Affiliations

Nature Methods **10**, 1139–1140 (2013) | doi:10.

Points of significance: Significance, *P* values and *t*-tests

Martin Krzywinski & Naomi Altman