R Fundamentals and Best Practices for Mass Spectrometry Data Analysis

Saturday, November 14 (12:00-3:15pm Eastern)

Meena Choi, Genentech

Olga Vitek, Northeastern University

Module #5: Basic statistics in R

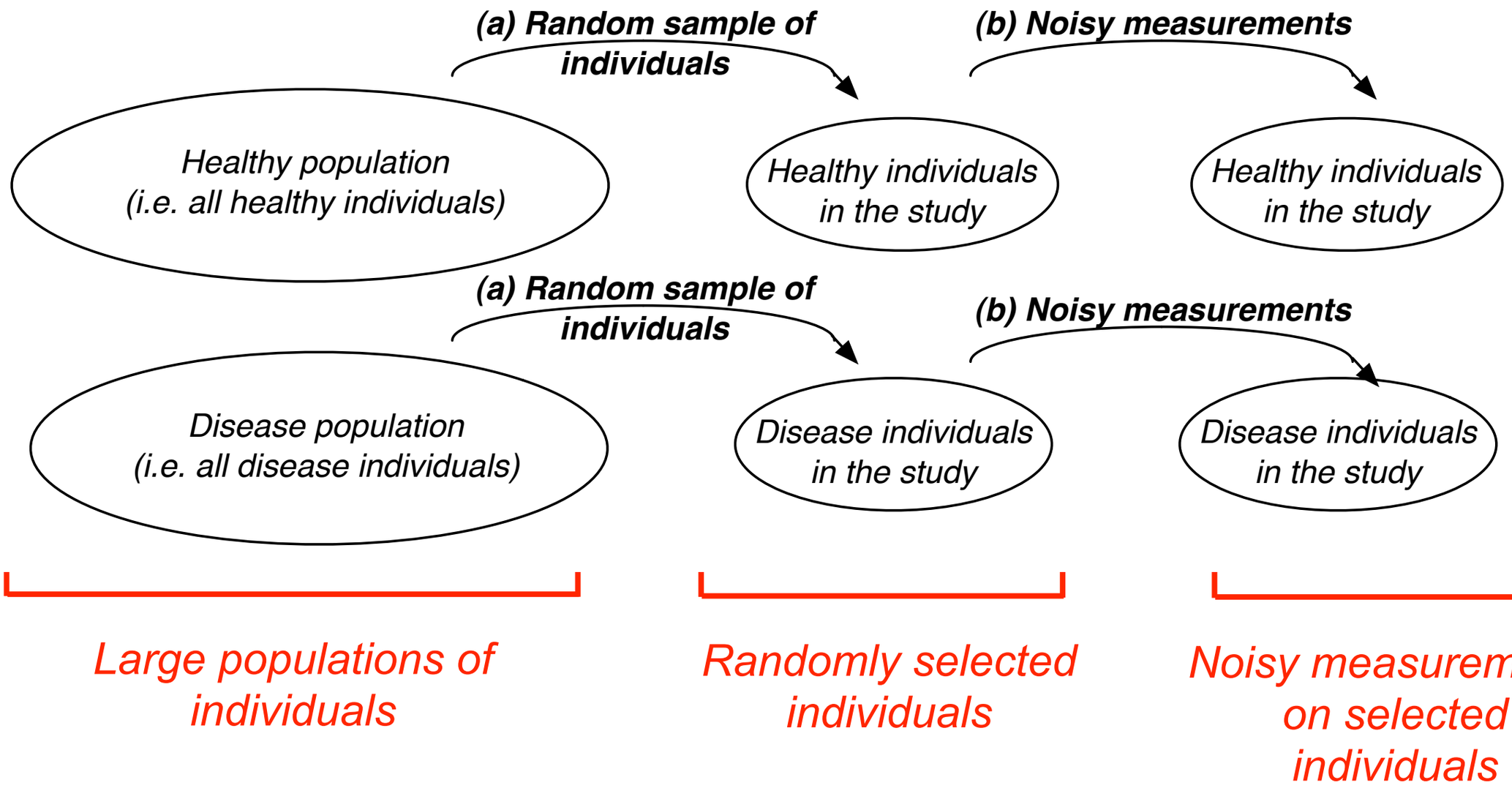
Module #6: Reproducible Data Analysis with RMarkdown

STATISTICAL DESIGN AND ANALYSIS OF EXPERIMENTS: A CRASH COURSE

Beware: this is not enough info!

- Fundamental principles of experimental design
  - Replication, randomization, blocking
- Basics of statistical inference
  - T-test, p-values and error bars
- Adjustments for multiple testing
  - False discovery rate

A STATISTICIAN'S VIEW OF THE EXPERIMENT



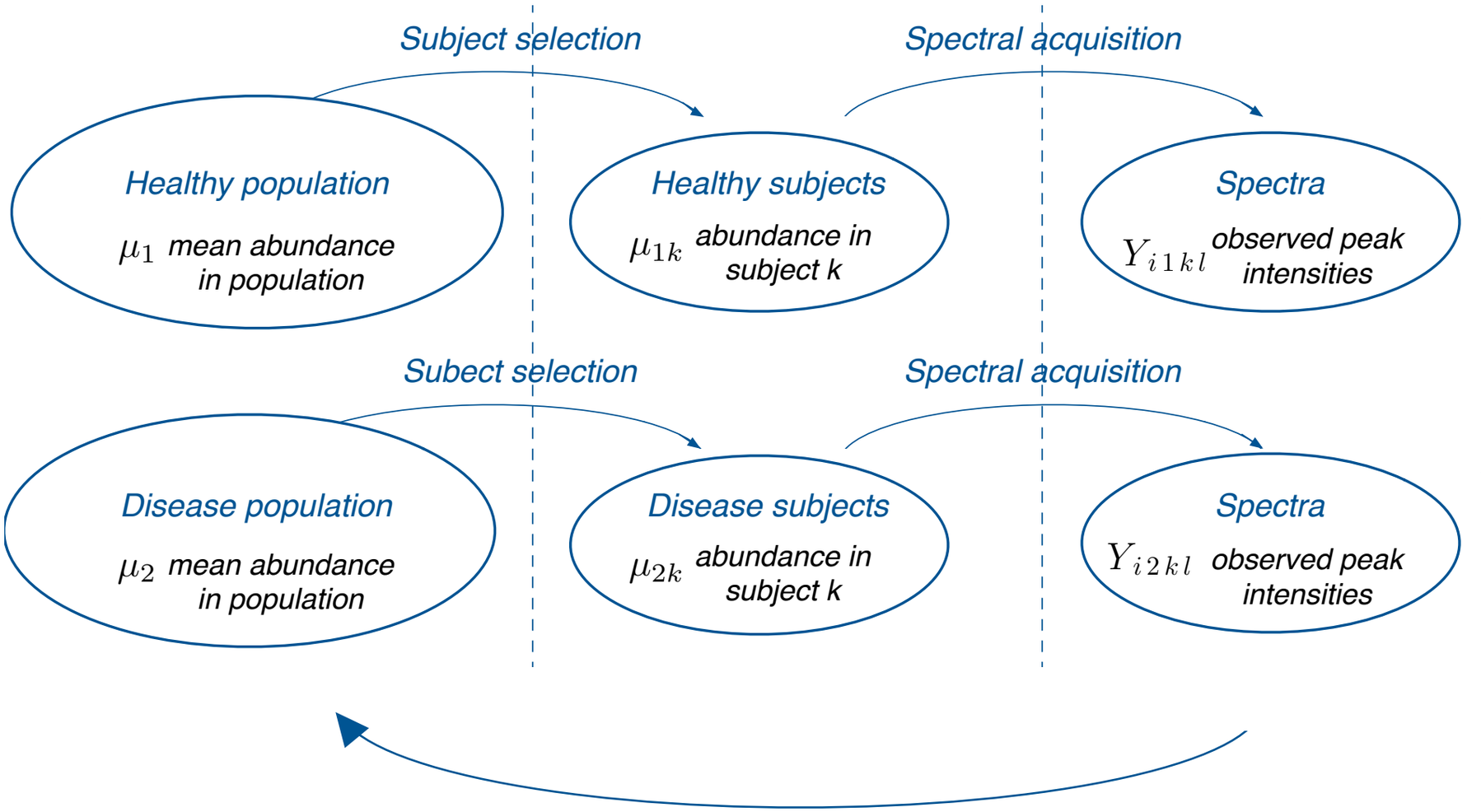
Dangers:

Bias: conclusions systematically differ from truth

Inefficiency: unnecessary variation in the data

COMPARE DESIGNS

In terms of bias and (in)-efficiency



Statistical model: properties of $\bar{Y}_{.1..} - \bar{Y}_{.2..}$

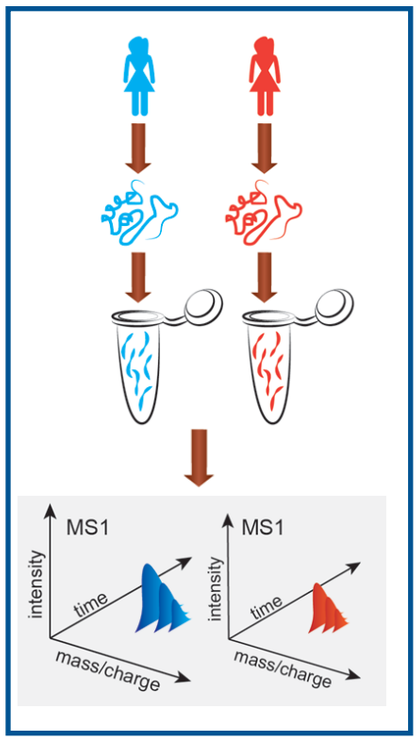
Statistical inference: conclusions on $\mu_1 - \mu_2$

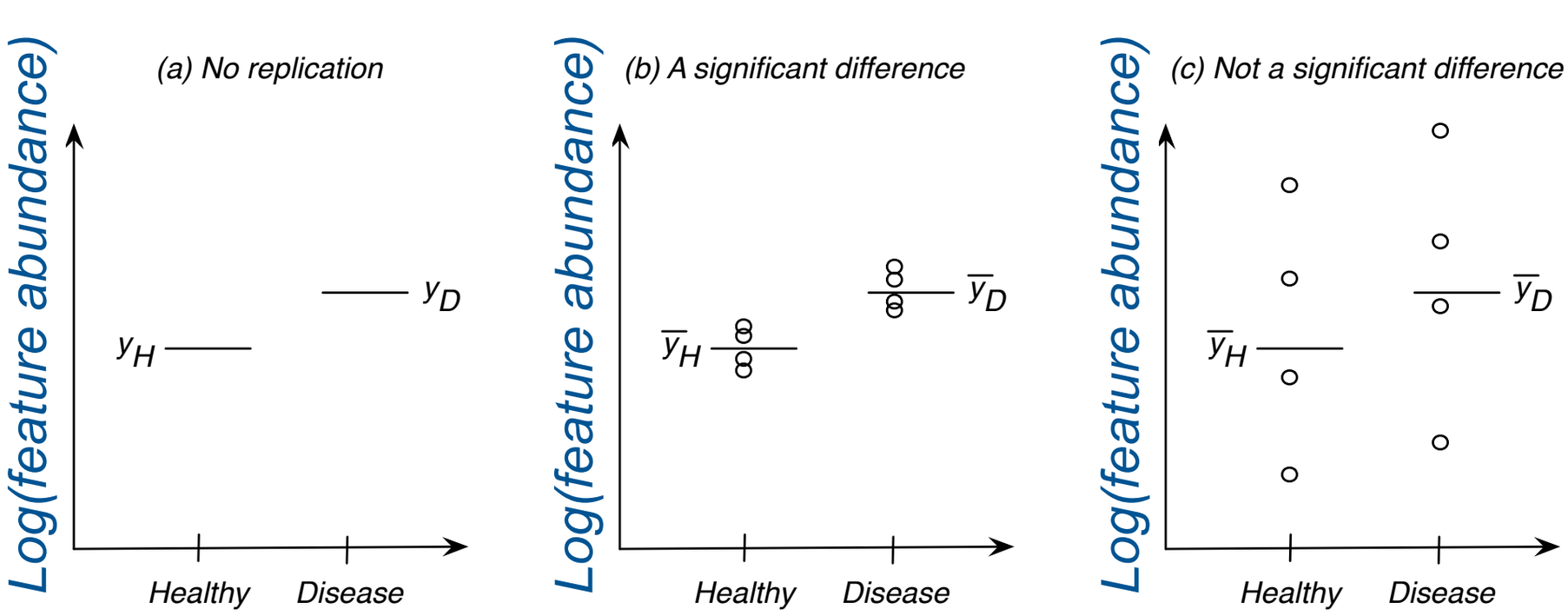
Bias: $\bar{Y}_{.1..} - \bar{Y}_{.2..}$ *systematically different from* $\mu_{1k} - \mu_{2k}$

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

PRINCIPLE 1: REPLICATION

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





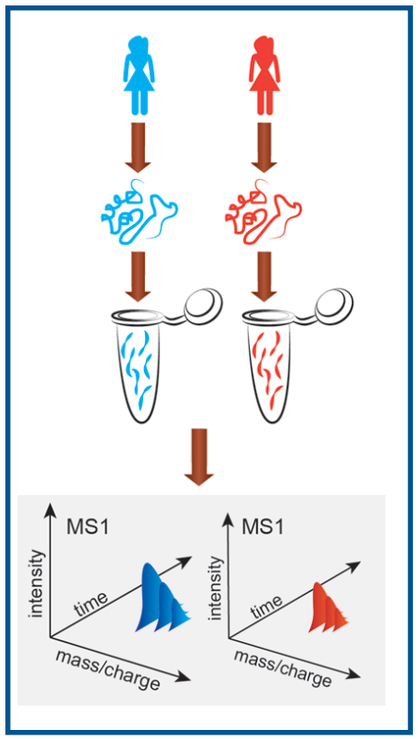
Two levels of randomness imply two types of replication:

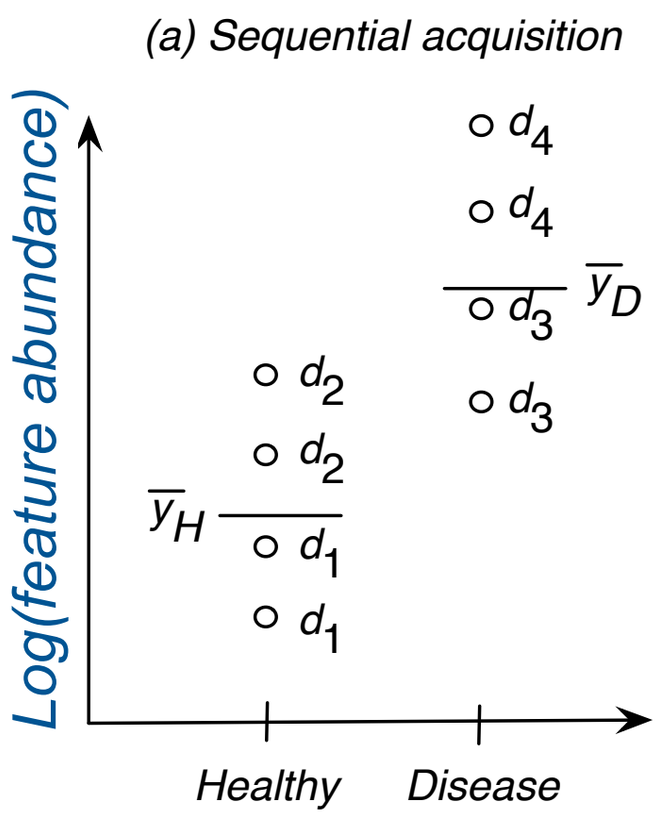
- *Biological replicates:* selecting multiple subjects from the population
- *Technical replicates:* multiple runs per subject

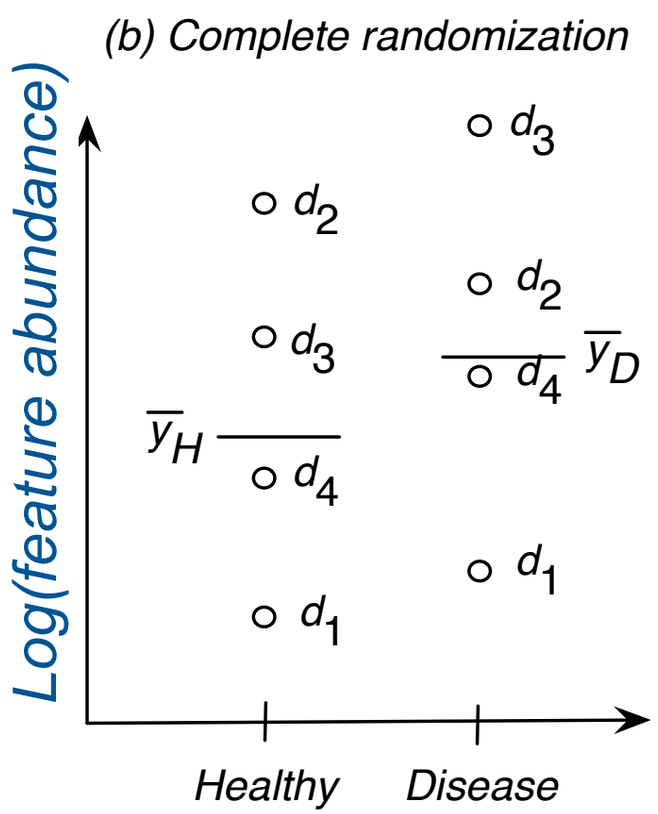
Oberg and Vitek, *J. Proteome Research*, 8, 2009

PRINCIPLE 2: RANDOMIZATION

Prevents bias







No randomization
= confounding
= bias

Complete randomization
= no bias

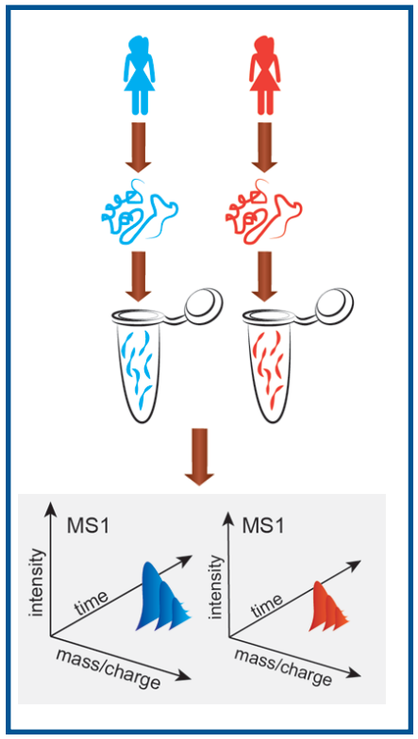
Two levels of randomness imply two types of randomization:

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

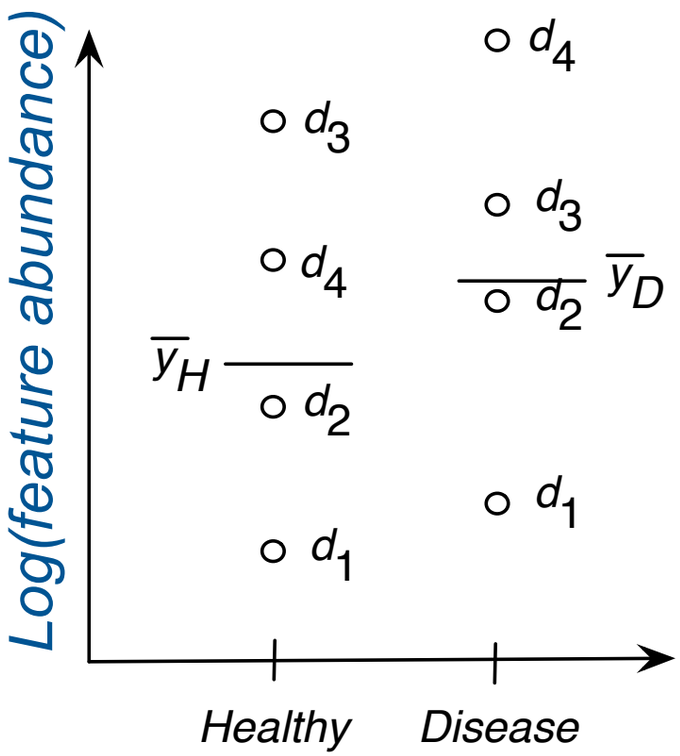
Oberg and Vitek, *J. Proteome Research*, 8, 2009

PRINCIPLE 3: BLOCKING

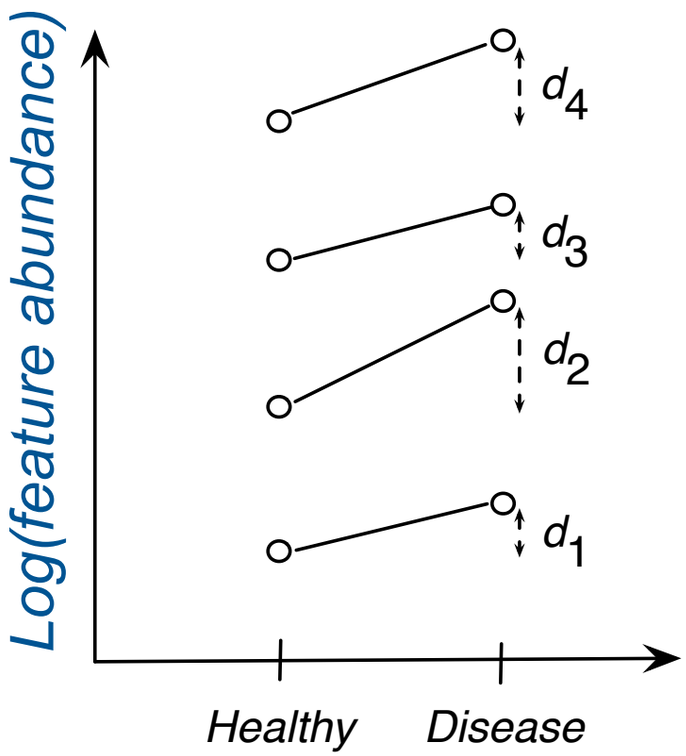
Helps reduce both bias and inefficiency



(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

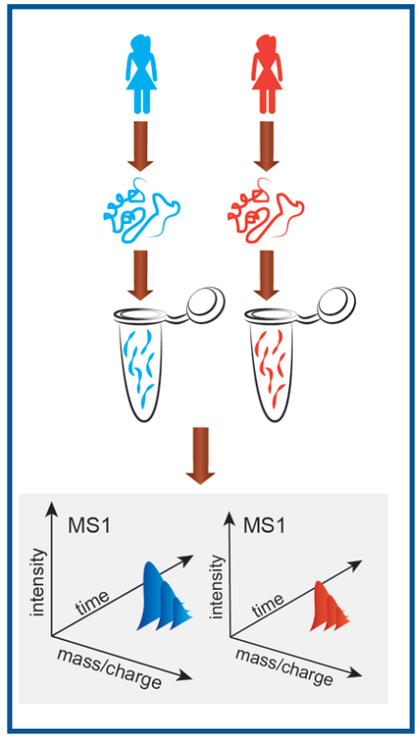
STATISTICAL DESIGN AND ANALYSIS OF EXPERIMENTS: A CRASH COURSE

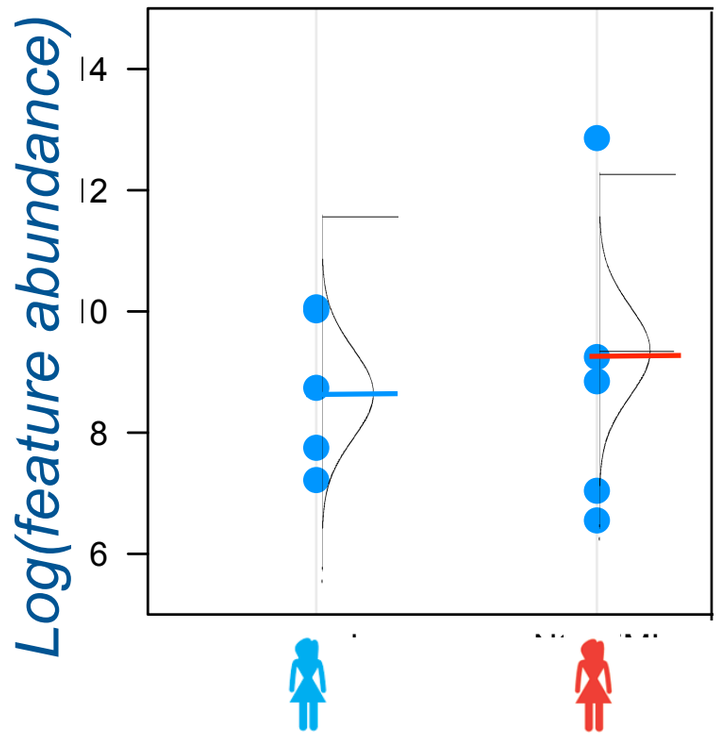
Beware: this is not enough info!

- Fundamental principles of experimental design
  - Replication, randomization, blocking
- Basics of statistical inference
  - T-test, p-values and error bars
- Adjustments for multiple testing
  - False discovery rate

TWO-SAMPLE T-TEST

Simple example: label-free experiment, one feature/protein





Sample means in each group

H0: 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$

Ha: change in abundance, $\mu_1 - \mu_2 \neq 0$

$$\textbf{observed t} = \frac{\textbf{difference of group means}}{\textbf{estimate of variation}} = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

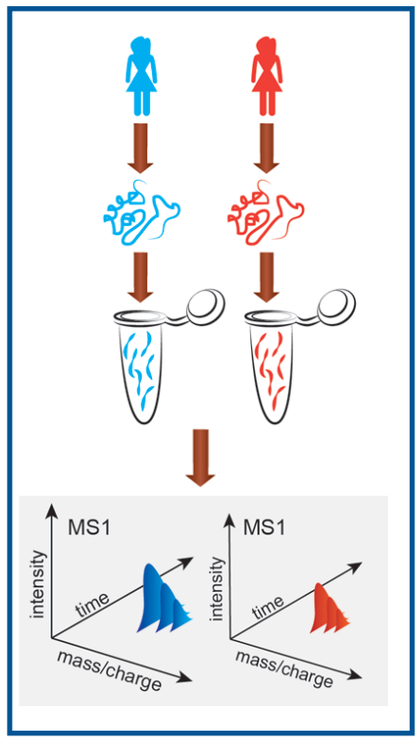
Number of replicates

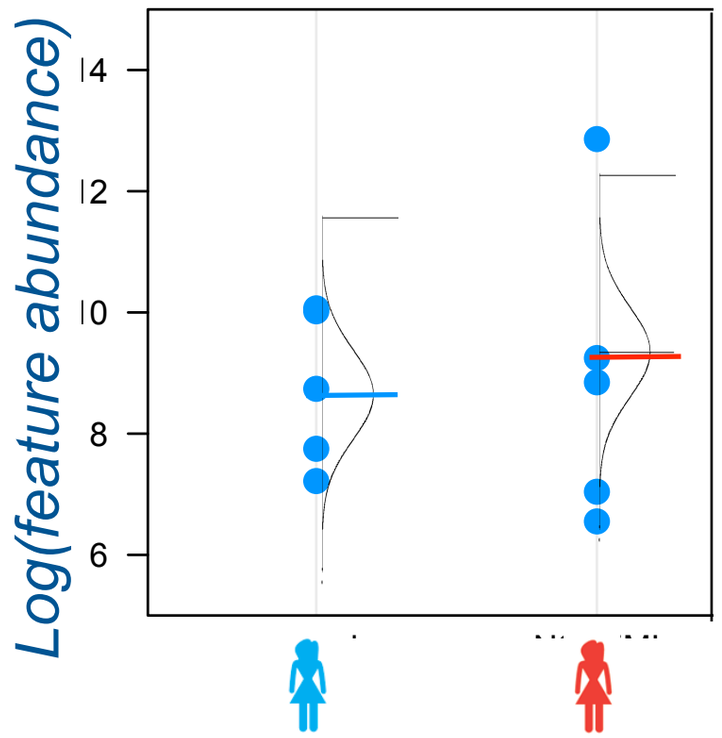
$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_{1\cdot})^2$$

Sample variance

TWO-SAMPLE T-TEST

Simple example: label-free experiment, one feature/protein





H0: 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$

Ha: change in abundance, $\mu_1 - \mu_2 \neq 0$

$$\textbf{observed t} = \frac{\textbf{difference of group means}}{\textbf{estimate of variation}} = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_{1\cdot})^2$$

Properties of the means

$\frac{s_1^2}{n_1}$

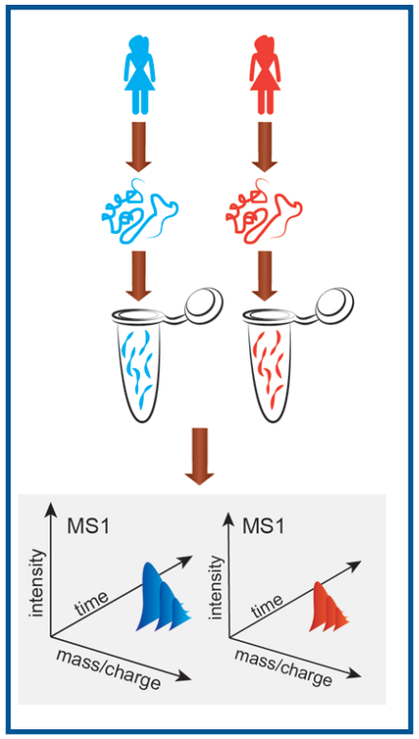
Variance of the sampling distribution of first mean

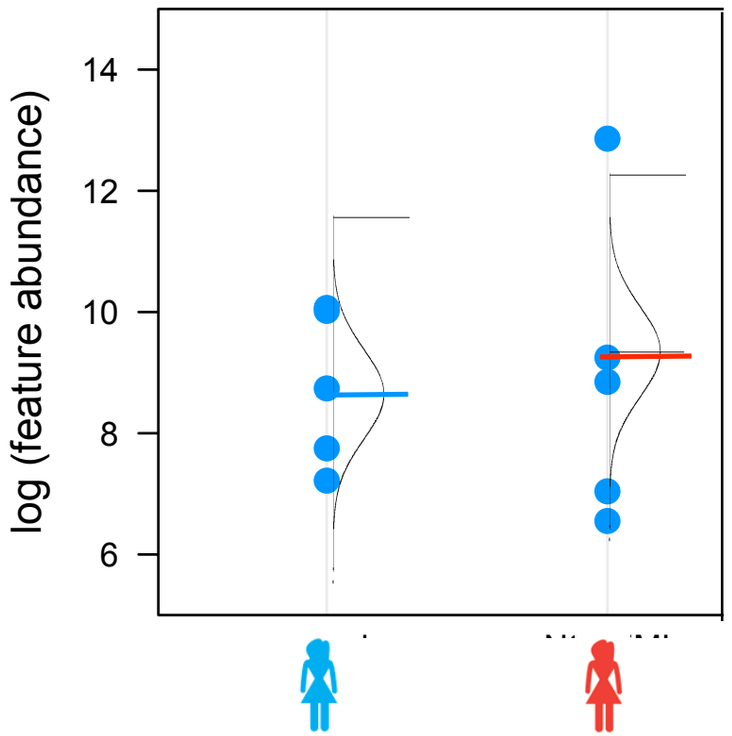
$\sqrt{\frac{s_1^2}{n_1}}$

Standard error of the first mean

FINDING DIFFERENTIALLY ABUNDANT PROTEINS

P-value





H0: 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$

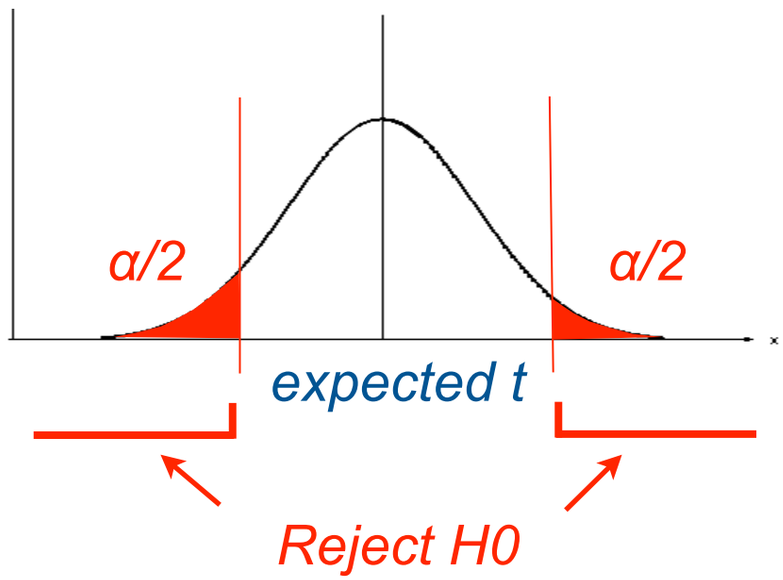
Ha: change in abundance, $\mu_1 - \mu_2 \neq 0$

$$\mathrm{t} = \frac{\textbf{difference of group means}}{\textbf{estimate of variation}} = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

$$\overset{\text{no difference}}{\sim} \textbf{Student distribution}$$

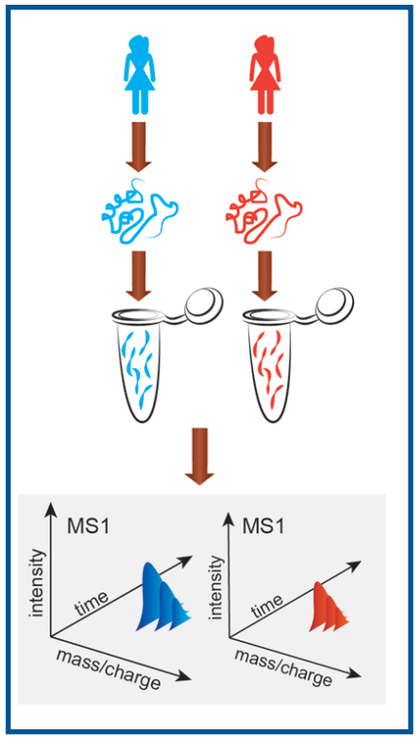
Distribution of the score if H0 is true

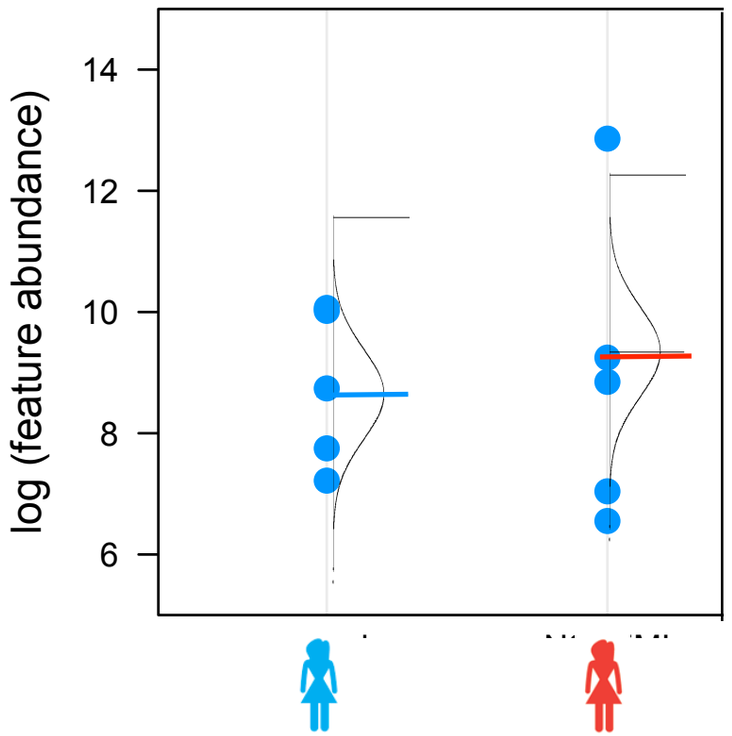
α = False Positive Rate



FINDING DIFFERENTIALLY ABUNDANT PROTEINS

P-value





H0: 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$

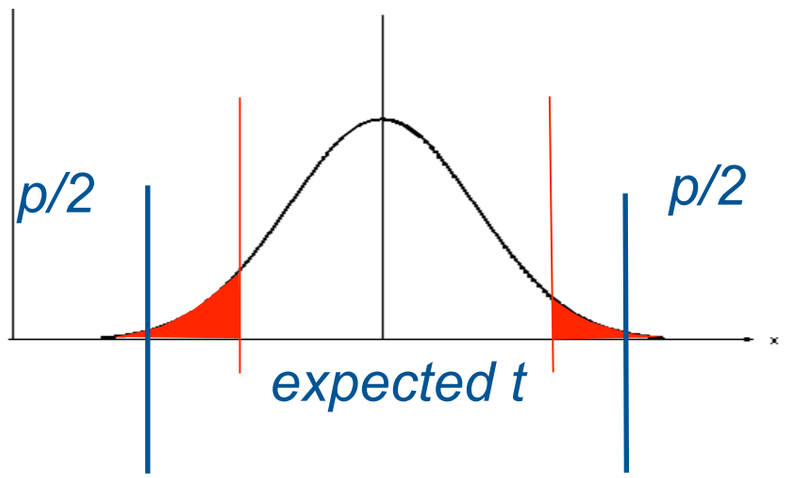
Ha: change in abundance, $\mu_1 - \mu_2 \neq 0$

$$\mathbf{t} = \frac{\textbf{difference of group means}}{\textbf{estimate of variation}} = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

$$\overset{\text{no difference}}{\sim} \textbf{Student distribution}$$

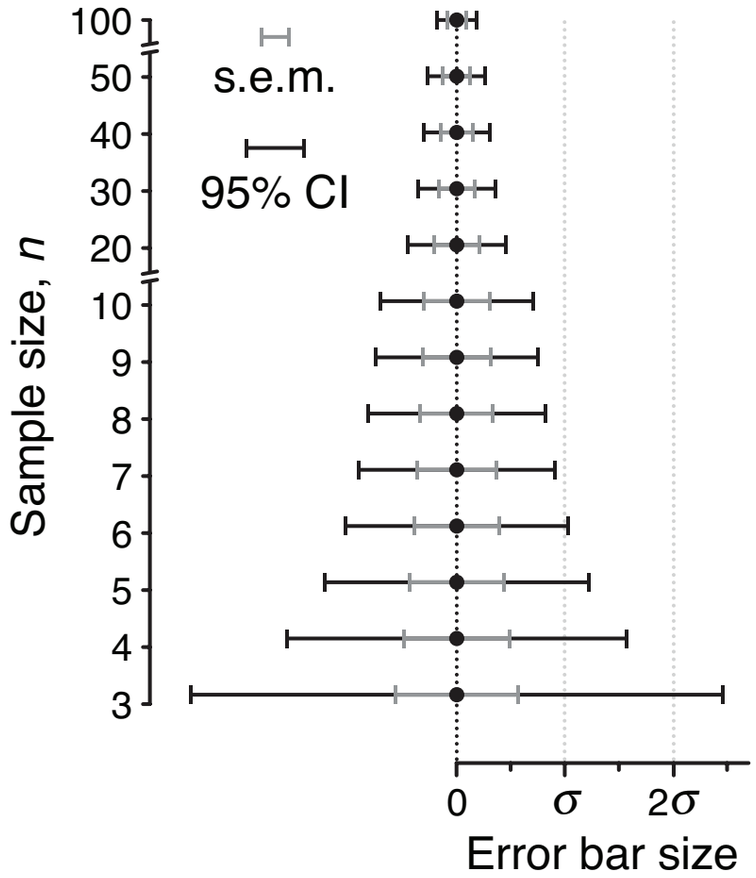
Distribution of the score if H0 is true

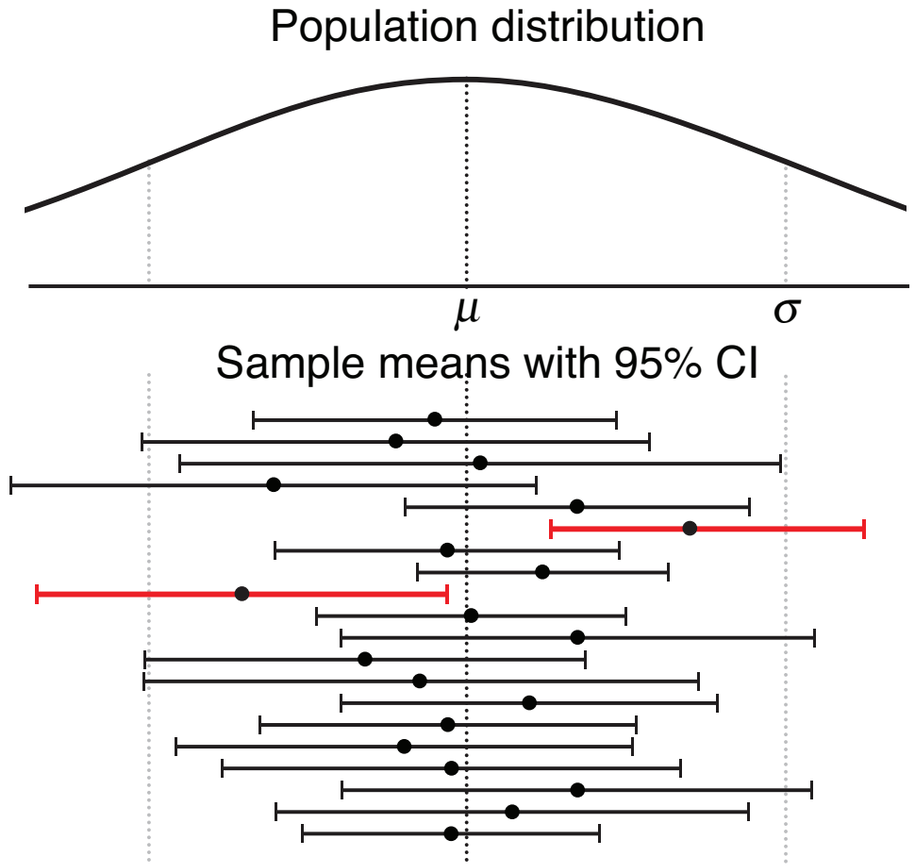
p = p-value



ALTERNATIVE TO TESTING: CONFIDENCE INTERVALS

Not all error bars are made equal





Note on sample size:

- ↑ biological replicates
- ↓ interval width

A 95% CI: if we repeatedly collect data and draw intervals, then 95% of them will contain the true mean

$$\left[(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}) - t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \;,\; (\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}) + t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

Krzywinski and Altman, Points of Significance Collection, *Nature Methods*

STATISTICAL DESIGN AND ANALYSIS OF EXPERIMENTS: A CRASH COURSE

Beware: this is not enough info!

- Fundamental principles of experimental design
  - Replication, randomization, blocking
- Basics of statistical inference
  - T-test, p-values and error bars
- Adjustments for multiple testing
  - False discovery rate

MULTIPLE TESTING

Control False Positive Rate for two proteins

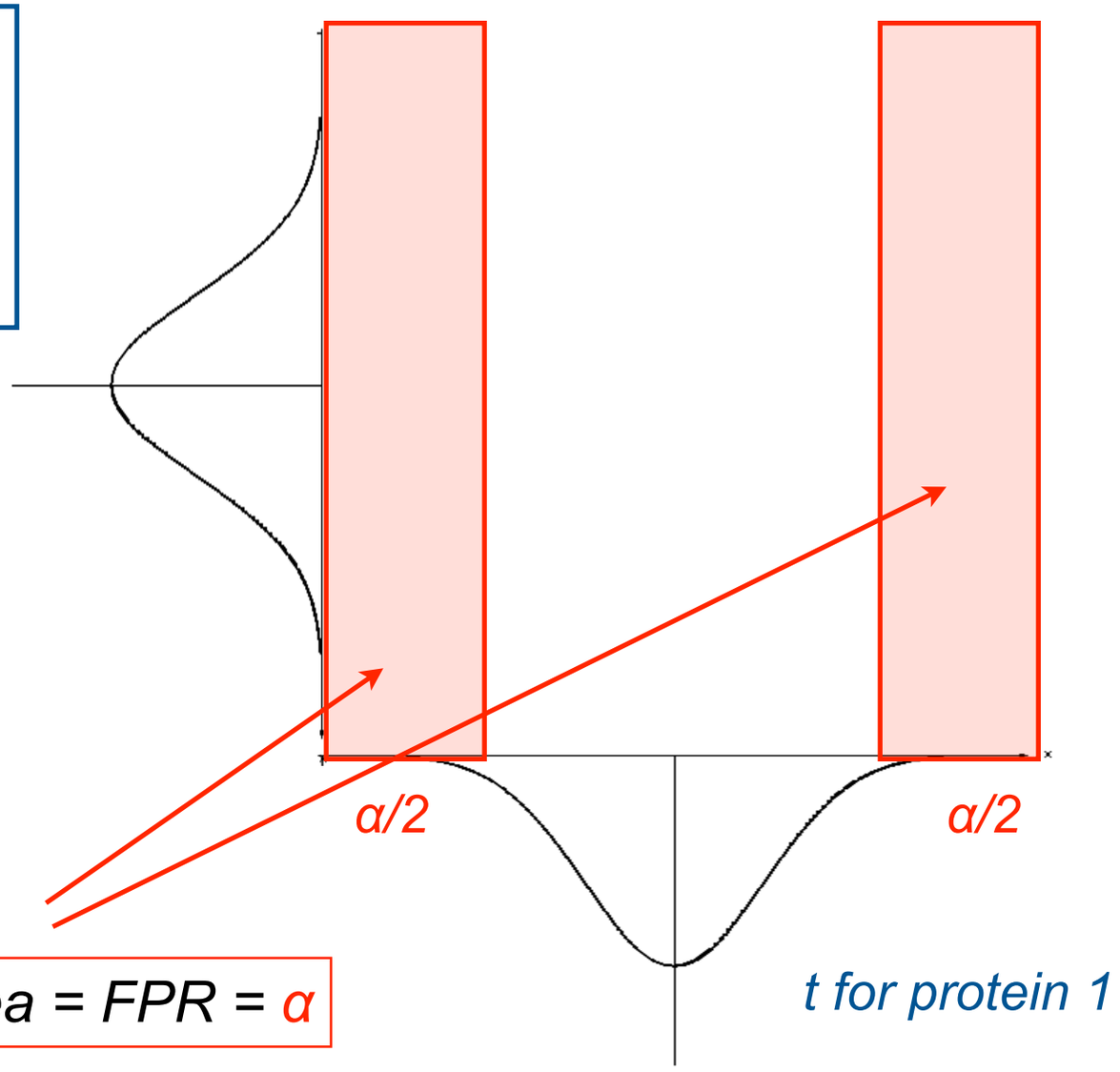
For each protein:

H0: 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$

Ha: change in abundance, $\mu_1 - \mu_2 \neq 0$

$$\textbf{observed t} = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

$$\overset{\textbf{no difference}}{\sim} \textbf{Student distribution}$$



MULTIPLE TESTING

Control False Positive Rate for two proteins

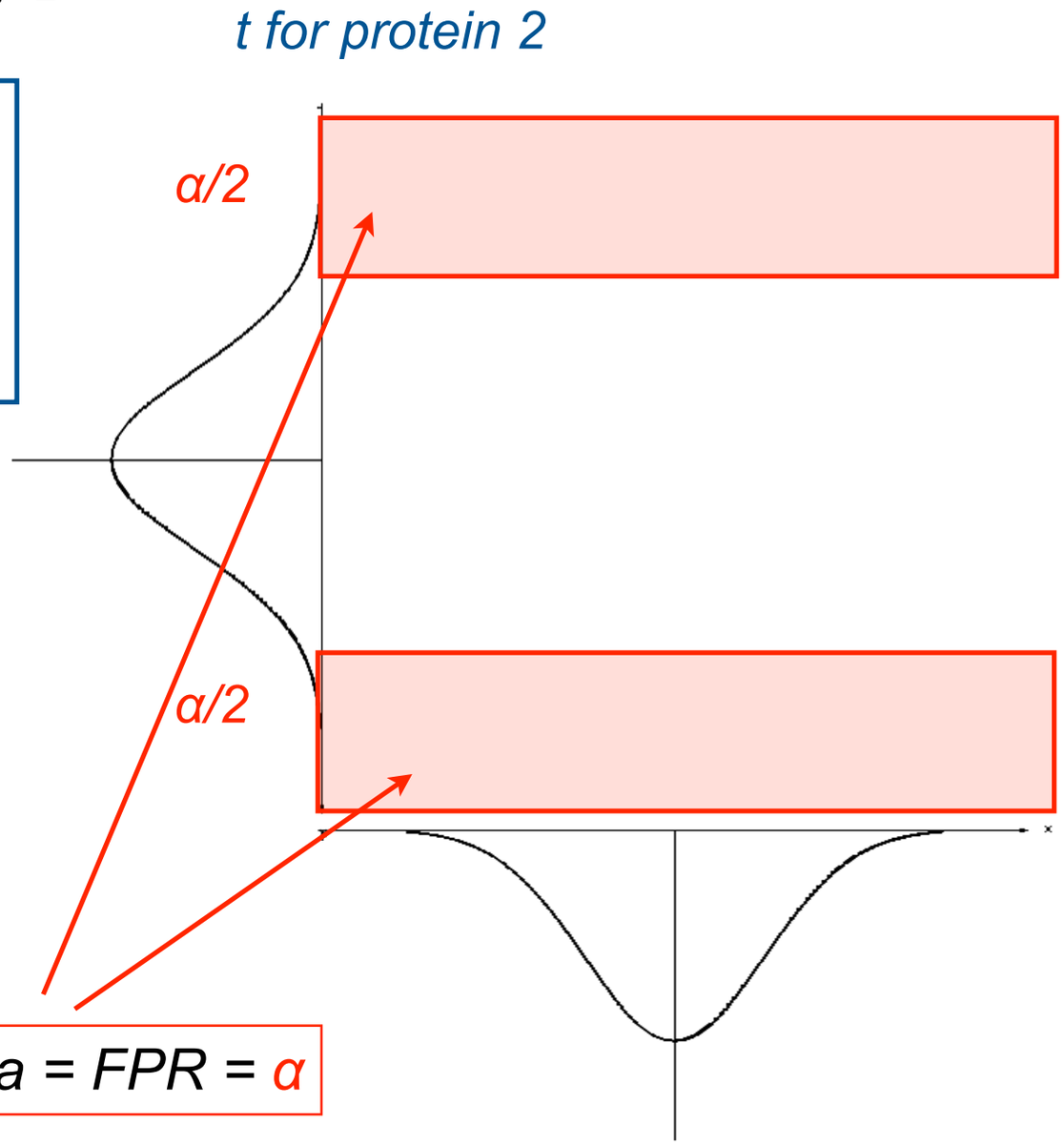
For each protein:

H0: 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$

Ha: change in abundance, $\mu_1 - \mu_2 \neq 0$

$$\textbf{observed t} = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

$$\overset{\textbf{no difference}}{\sim} \textbf{Student distribution}$$



MULTIPLE TESTING

Control False Positive Rate for two proteins

For each protein:

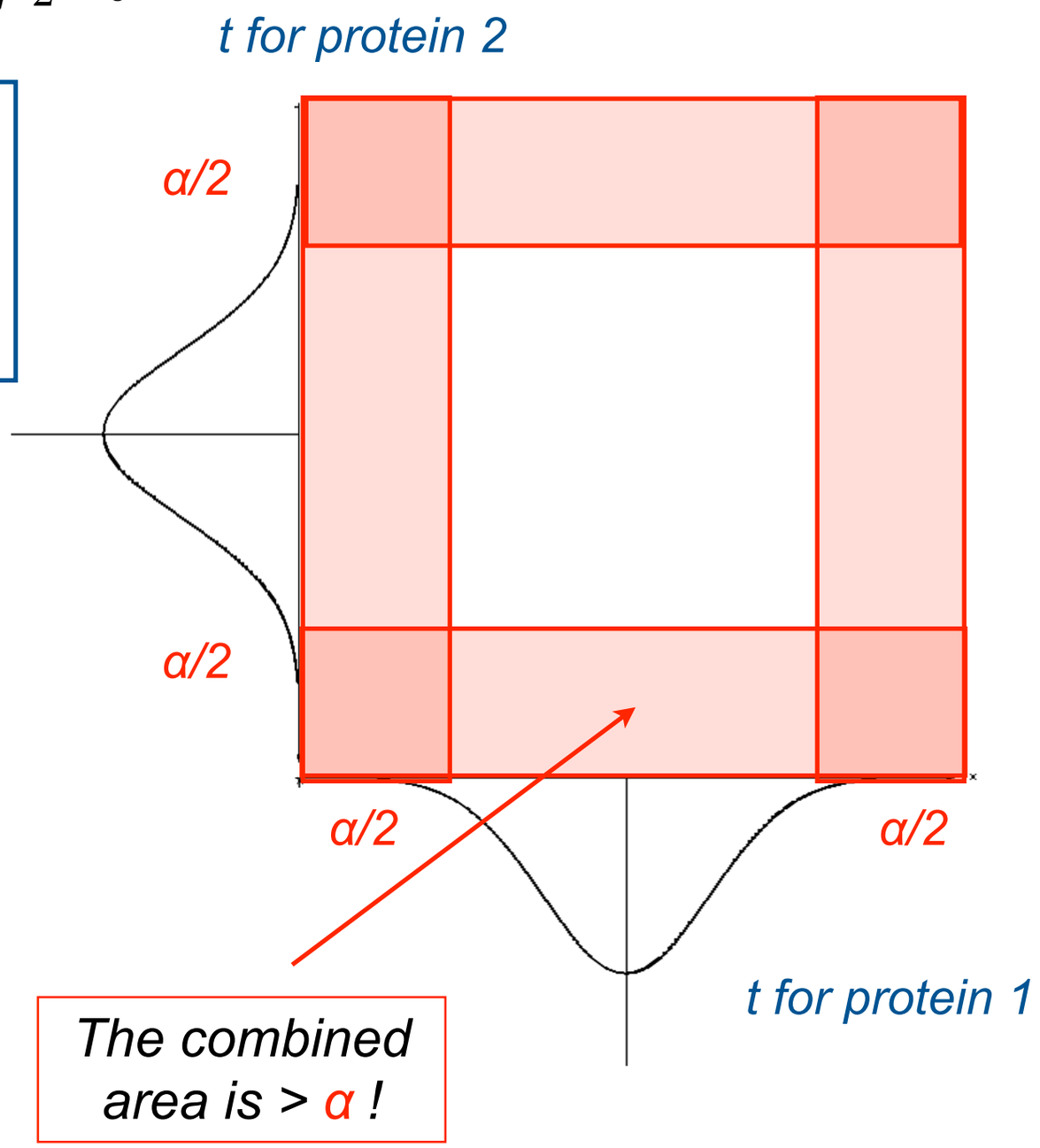
H0: 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$

Ha: change in abundance, $\mu_1 - \mu_2 \neq 0$

$$\textbf{observed t} = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

$$\overset{\textbf{no difference}}{\sim} \textbf{Student distribution}$$

- P(at least one incorrect decision) > α
- The univariate FPR does not hold for the list
- Need to define a *multivariate* error rate



MULTIPLE TESTING

Control False Positive Rate for two proteins

For each protein:

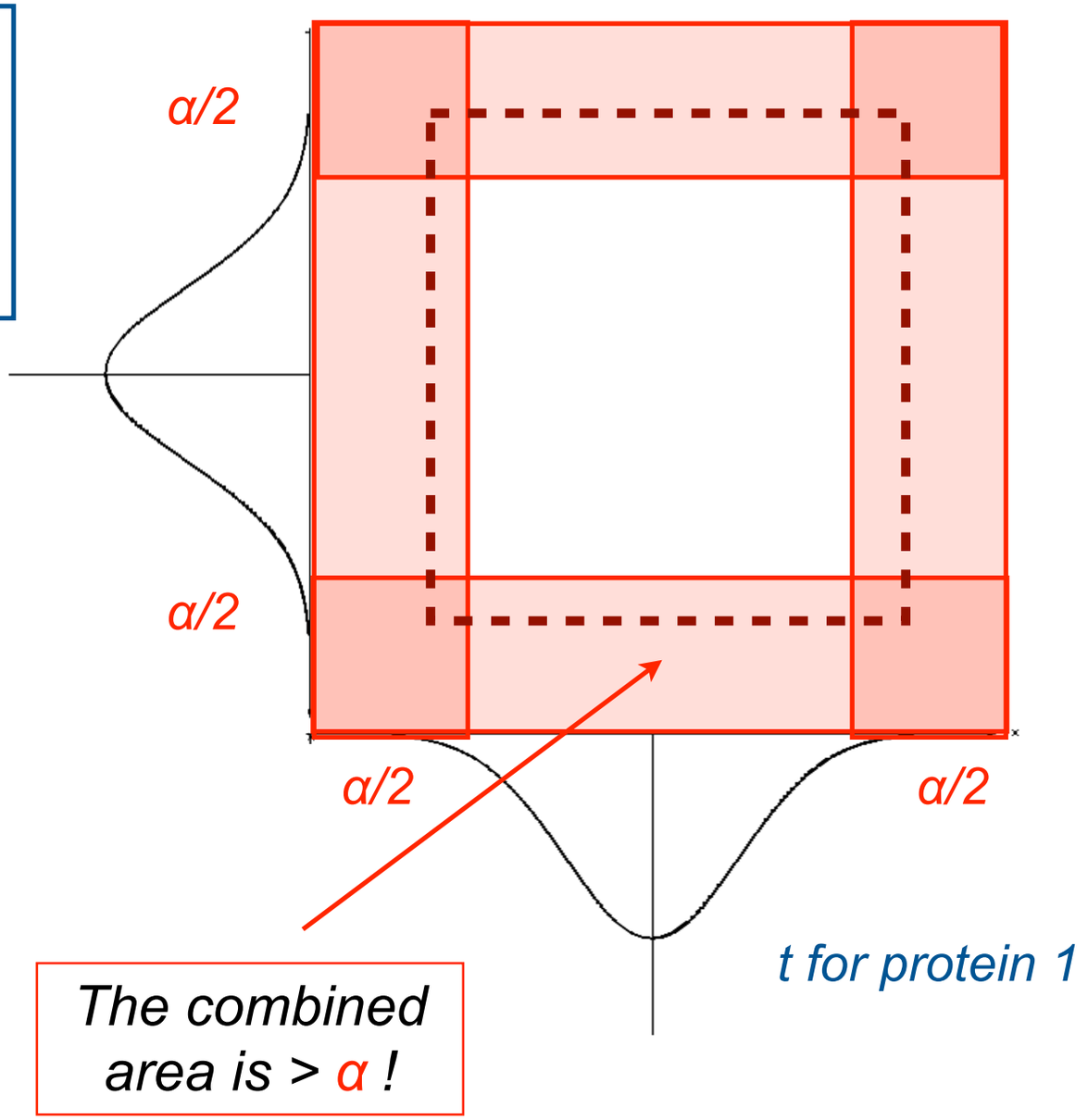
H0: 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$

Ha: change in abundance, $\mu_1 - \mu_2 \neq 0$

$$\text{observed t} = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

$$\overset{\text{no difference}}{\sim} \textbf{Student distribution}$$

- P(at least one incorrect decision) > α
- The univariate FPR does not hold for the list
- Need to define a *multivariate* error rate



TESTING M PROTEINS

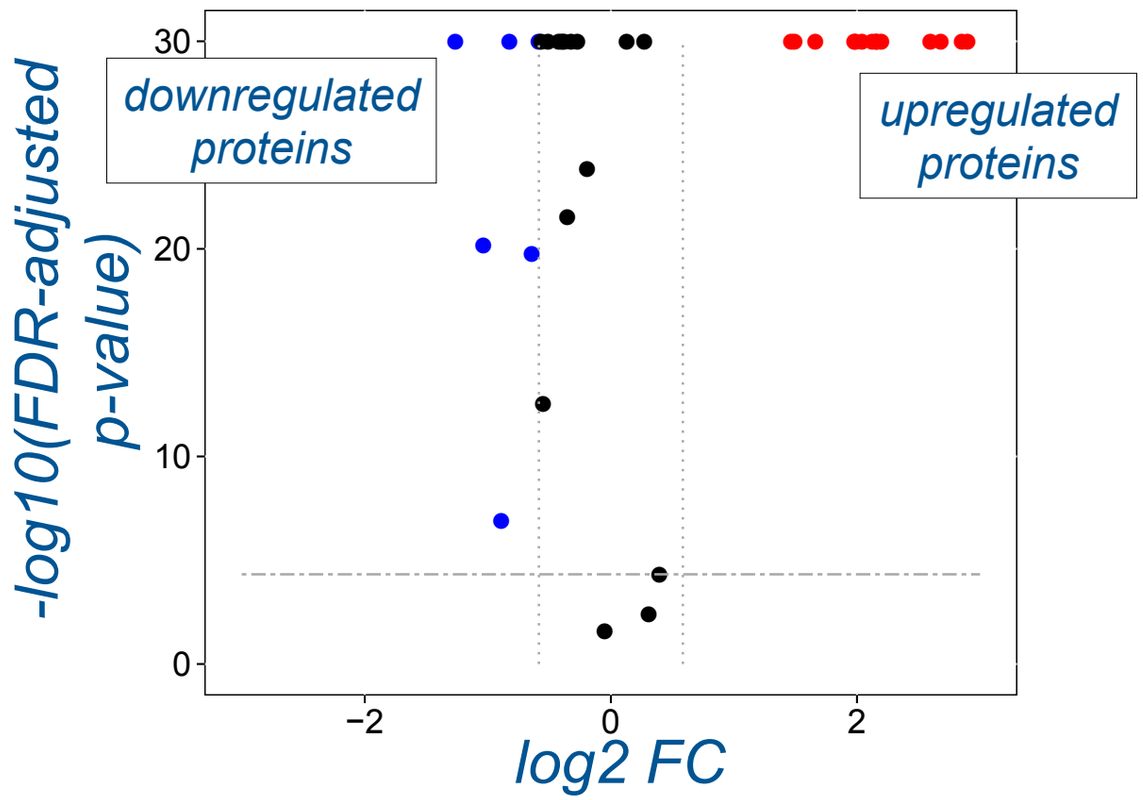
Change criteria from False Positive Rate to False Discovery Rate

	# of proteins with no detected difference	# of proteins with detected difference	Total
# true non-diff. proteins	U	V	m_0
# true diff. proteins	T	S	$m_1 = m - m_0$
Total	$m - R$	R	m

- False discovery rate (FDR)
  - An infinite number of measurements on same proteins
  - FDR: the *average* proportion of false discoveries

$$\mathbf{FDR} = \mathbf{E}\left[\frac{\mathbf{V}}{\max(\mathbf{R}, 1)}\right]$$

Bonferroni approach controls family-wise error rate = $P(V > 0)$



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/