STATISTICAL METHODS FOR HIGH-THROUGHPUT BIOLOGY

Meena Choi, Ting Huang, Olga Vitek

College of Science

College of Computer and Information Science

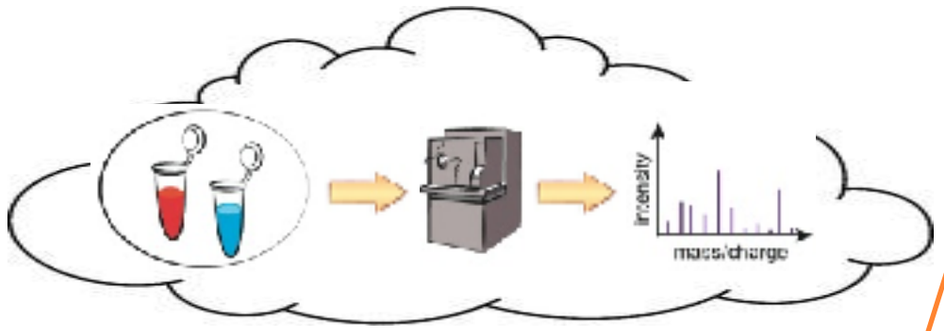
Northeastern University

CHALLENGES

- Normalization
  - Remove systematic shifts in signals between runs
- Statistical modeling
  - Limma: continuous features, small n
  - MSstats: many continuous features per analyte
  - DEseq2: counts per analyte
- Hypothesis testing
  - Multiple comparisons

A TYPICAL ANALYSIS WORKFLOW

Experimental design



QC and normalization

Statistical modeling

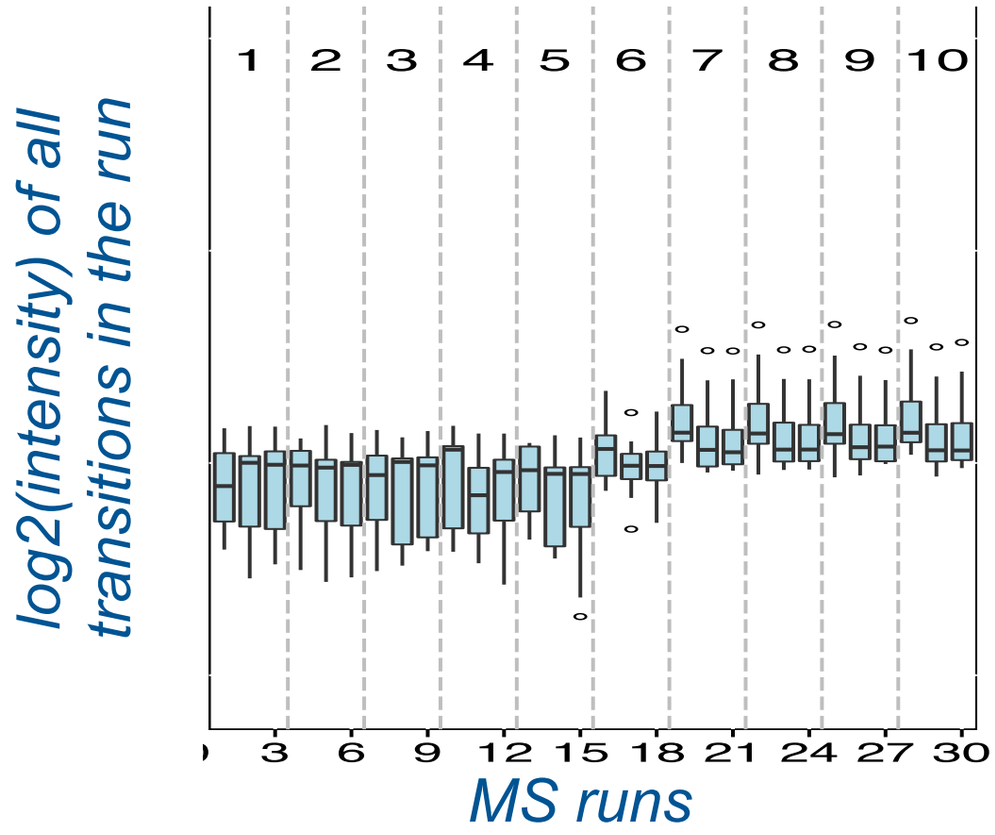
Model-based conclusions

Experimental design

Data preparation

- Formatting
- Visualization
- Remove artifacts
- Normalization

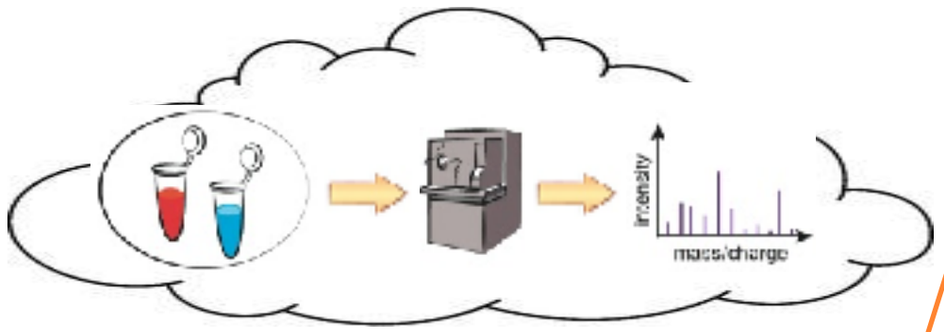
Endogenous



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

A TYPICAL ANALYSIS WORKFLOW

Experimental design

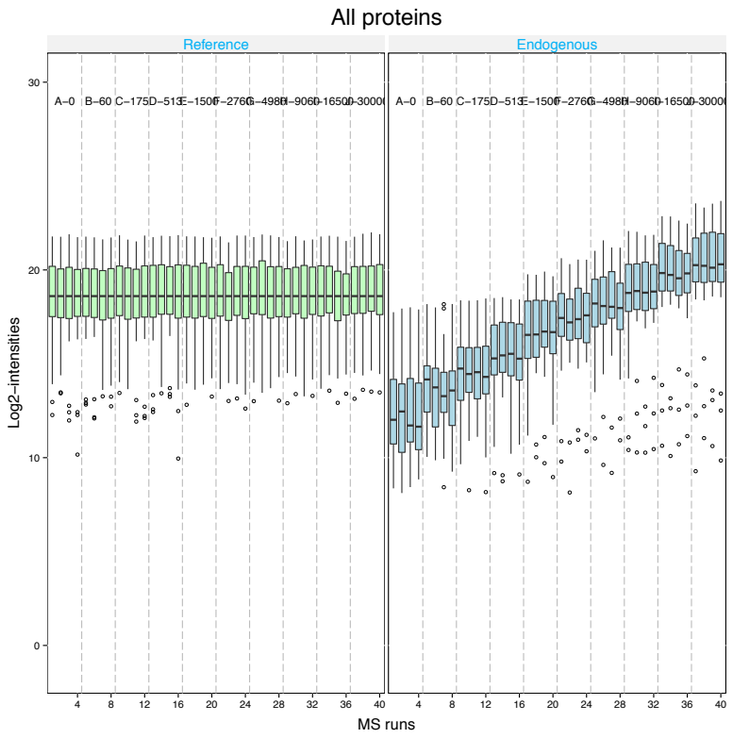


QC and normalization

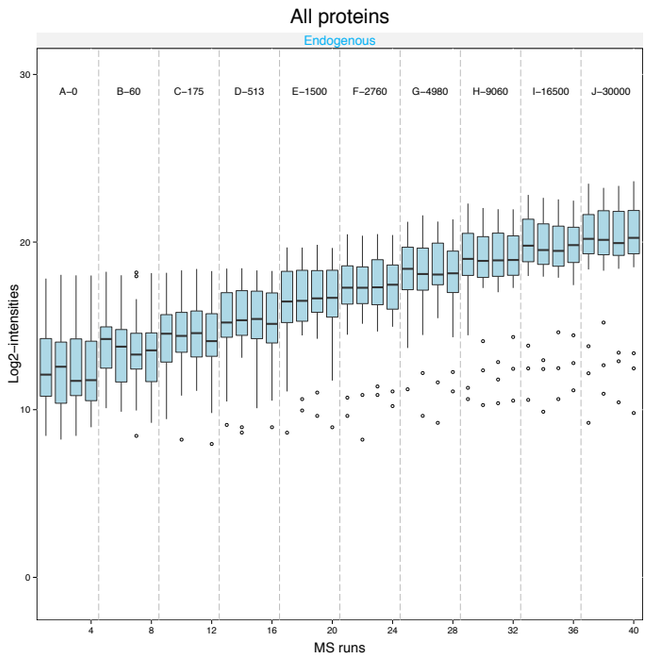
Data preparation

- Formatting
- Visualization
- Remove artifacts
- Normalization

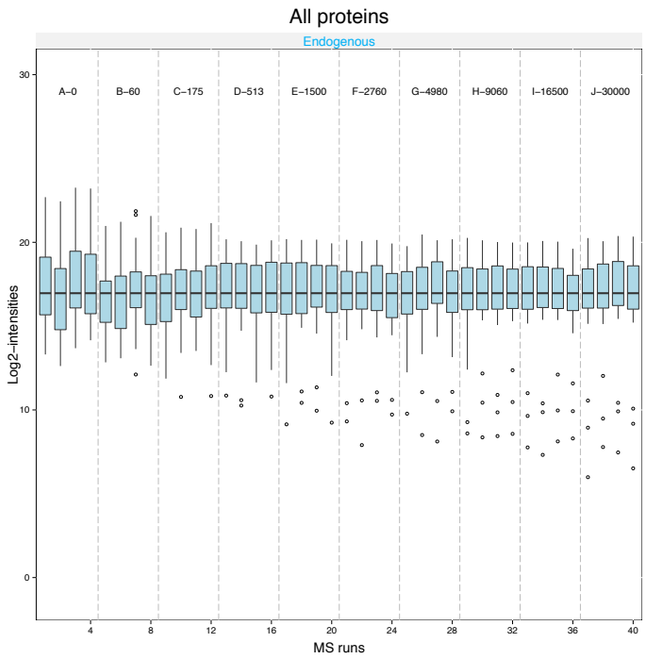
Equalize medians normalization



No normalization

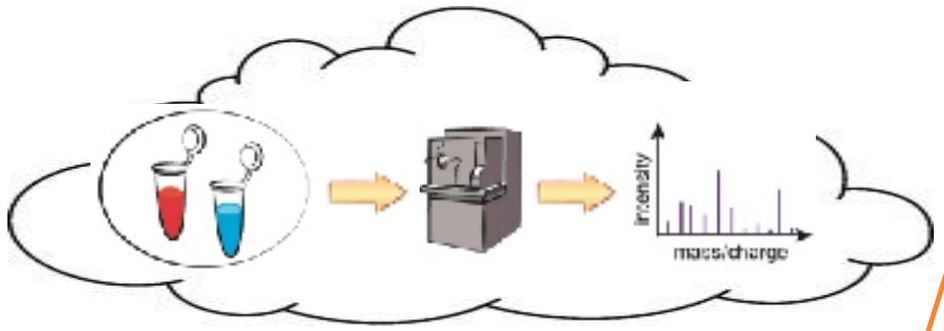


Equalize medians normalization



A TYPICAL ANALYSIS WORKFLOW

Experimental design



QC and normalization

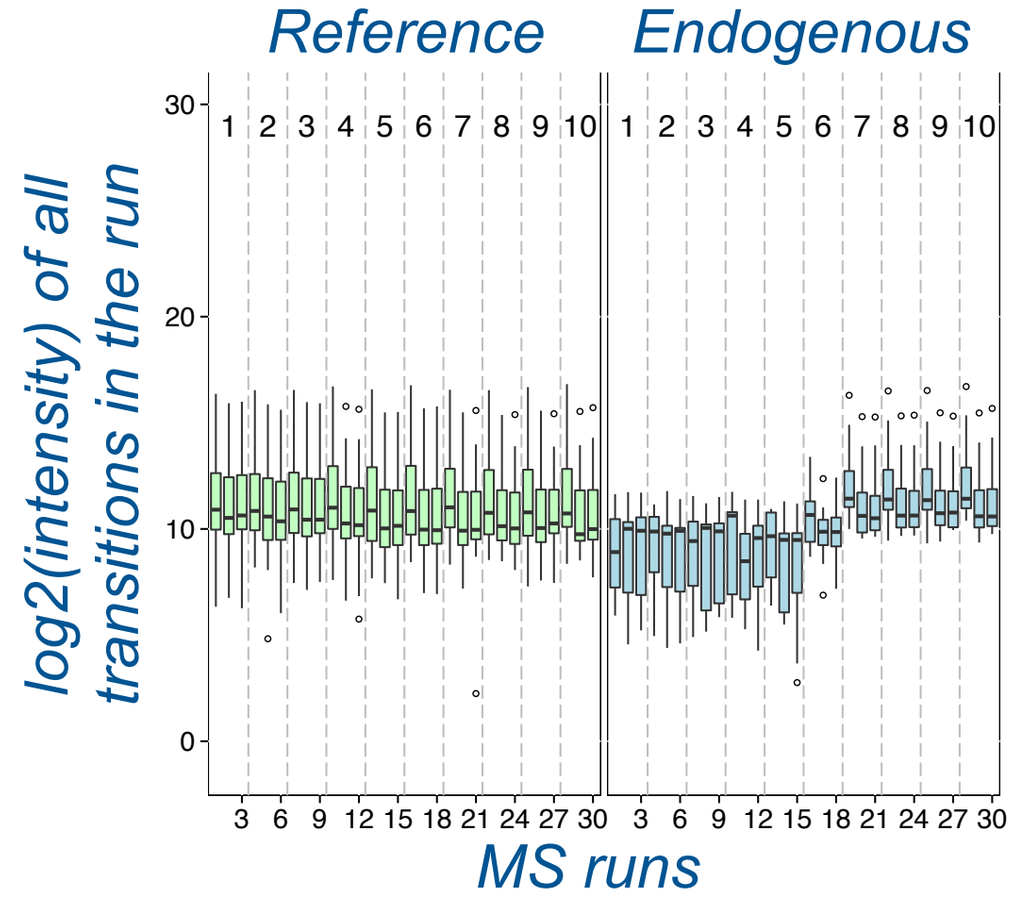
Statistical modeling

Model-based conclusions

Experimental design

Data preparation

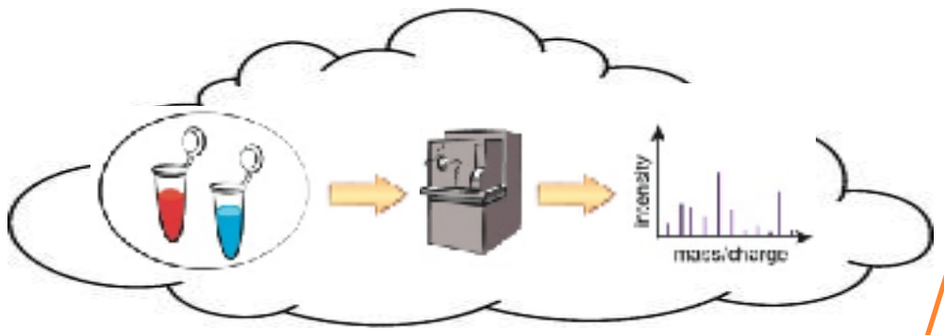
- Formatting
- Visualization
- Remove artifacts
- Normalization



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

A TYPICAL ANALYSIS WORKFLOW

Experimental design



QC and normalization

Statistical modeling

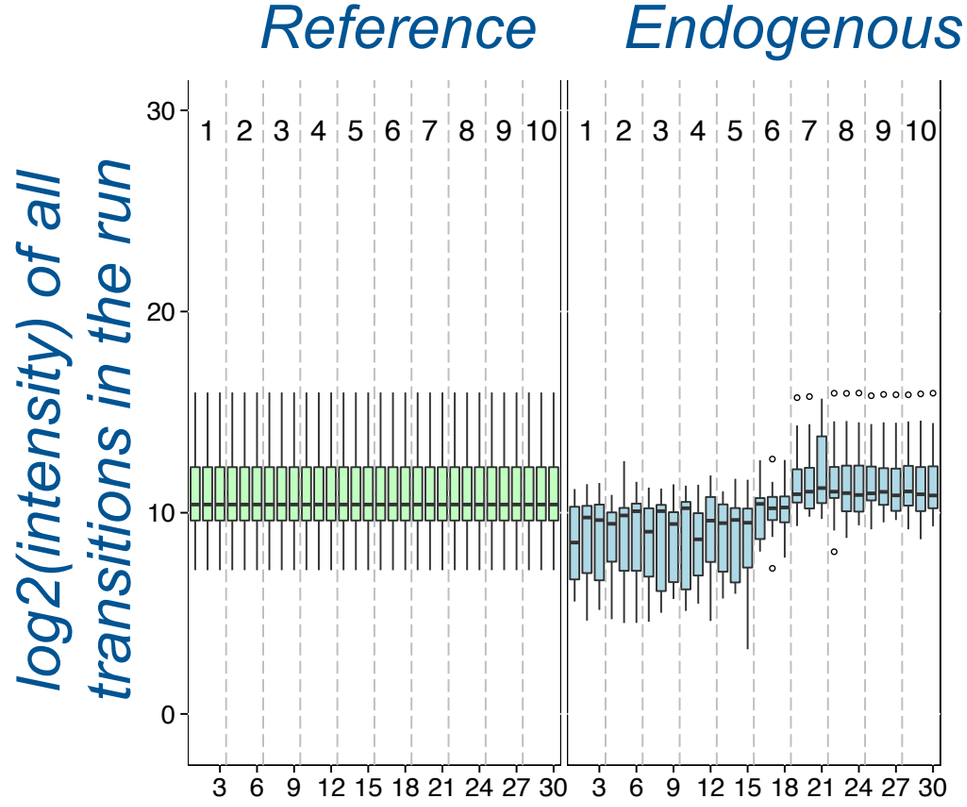
Model-based conclusions

Experimental design

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

Data preparation

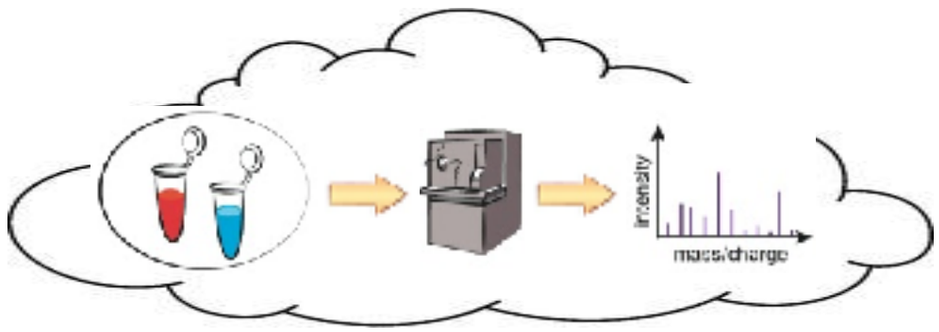
- Formatting
- Visualization
- Remove artifacts
- Normalization



Quantile normalization equalizes all the quantiles of the reference signals in a run

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!

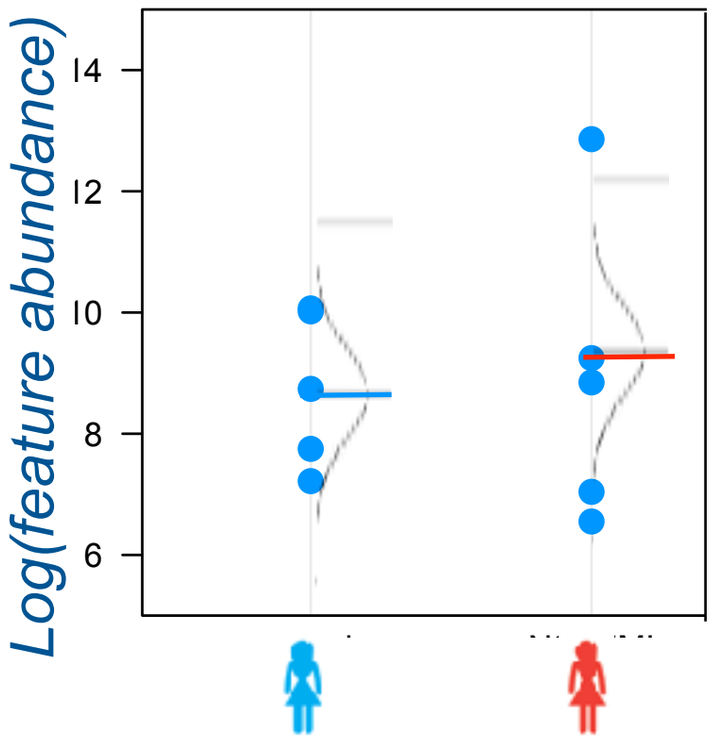
- Describe statistical properties of
  - experimental design
  - biological variation
  - measurement technology

CHALLENGES

- Normalization
  - Remove systematic shifts in signals between runs
- Statistical modeling
  - Limma: continuous features, small n
  - MSstats: many continuous features per analyte
  - DEseq2: counts per analyte
- Hypothesis testing
  - Multiple comparisons

TESTING MANY HYPOTHESES IN PARALLEL

Recall 2-sample t-test



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

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

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

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

Modification: assume same variance in both groups

$$\textbf{t} = \frac{\textbf{difference of group means}}{\textbf{estimate of variation}} = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{1\cdot}}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Problem: with small sample size, s^2 *is unreliable*

LIMMA BORROWS INFO ON VARIATION ACROSS GENES

ANOVA-based comparison:

$$\frac{\bar{y}_{Disease} - \bar{y}_{Control}}{s \cdot constant} \sim Student_d$$

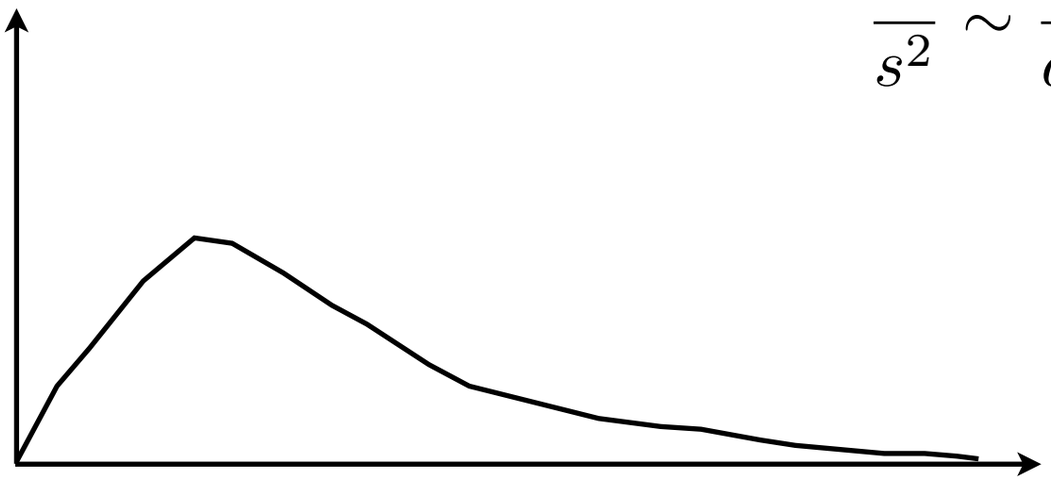
Gene-specific variance

Reflects the number of data points in the feature

Gene-specific variance estimation can be unstable for small sample size

Smyth, 2005

Frequency over all features



$$\frac{1}{s^2} \sim \frac{1}{d_0 \cdot s_0^2} \cdot \chi^2_{d_0}$$

"Average degree of freedom" estimated over all features

"Average variance" estimated over all features

ESTIMATION OF THE HYPERPARAMETERS

Anova-based comparison:

$$\frac{\bar{y}_{Disease} - \bar{y}_{Control}}{s \cdot constant} \sim Student_d$$

Feature-specific variance

Reflects the number of data points in the feature

Moderated Anova-based comparison (Limma, implemented in Corra):

Smyth, 2005

$$\frac{\bar{y}_{Disease} - \bar{y}_{Control}}{\tilde{s} \cdot constant} \sim Student(\tilde{d})$$

$$\tilde{s}^2 = \frac{d_0 \cdot s_0^2 + d \cdot s^2}{d_0 + d}$$

$$\tilde{d} = d_0 + d$$

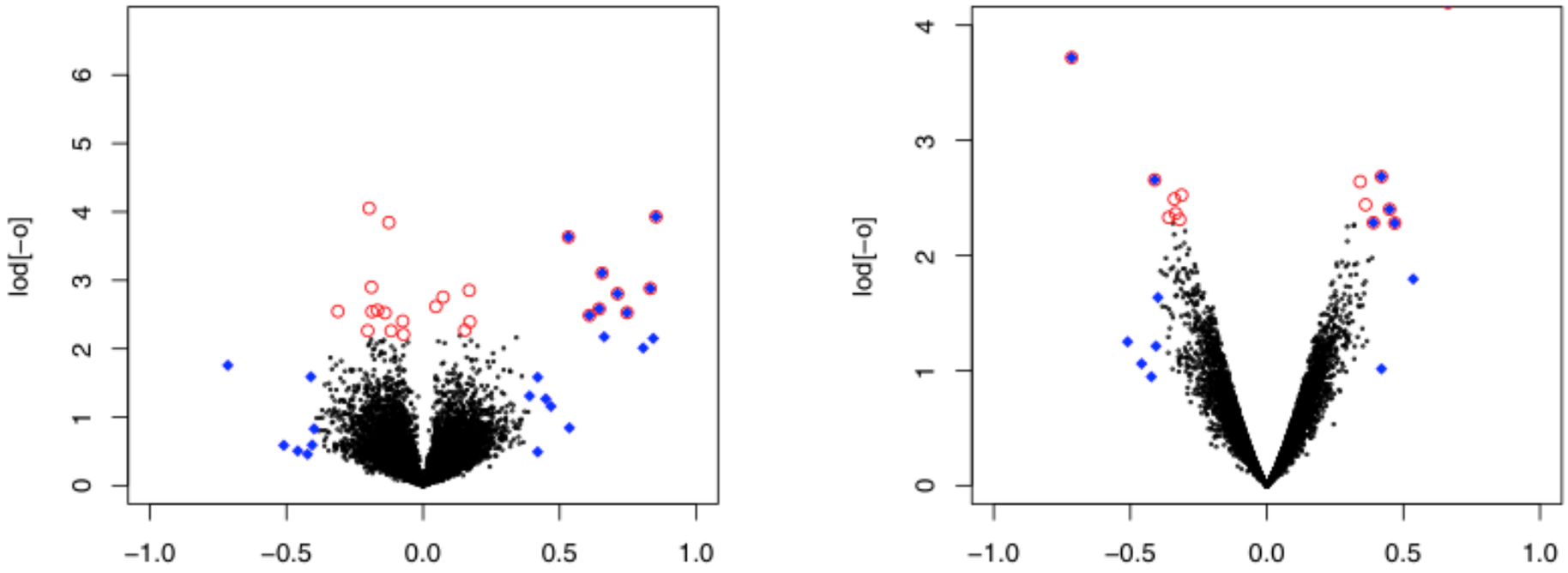
“Average variance” estimated over all features

“Average degree of freedom” estimated over all features

IMPROVED ESTIMATION OF VARIATION IMPROVES HYPOTHESIS TESTING

Primarily in experiments with a small number of replicates

Smyth, Stat. Appl. in Genetics and Mol. Biol., 2005



- Variance: sample (left), moderated (right)
- X axis: estimated log-fold change
- Y axis: $-log$(pValue) (here: $-log$(posteriorOdds))
- Moderated variance yields better agreement between log-fold change and significance

CHALLENGES

- Normalization
  - Remove systematic shifts in signals between runs
- Statistical modeling
  - Limma: continuous features, small n
  - MSstats: many continuous features per analyte
  - DEseq2: counts per analyte
- Hypothesis testing
  - Multiple comparisons

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

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

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

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

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

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

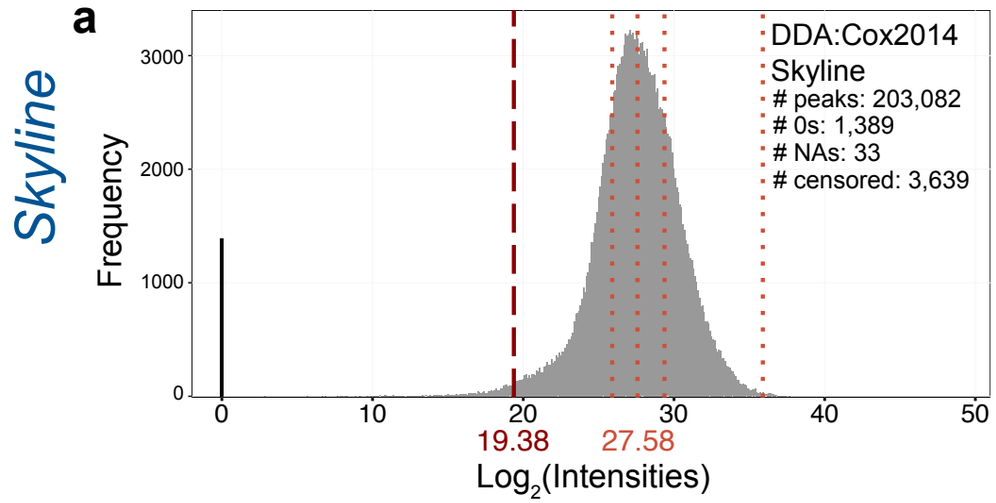
T. Clough et al. *BMC Bioinformatics.*, 2012
C.-Y. Chang et al. *MCP*, 2012

PROPERTIES OF PEAK INTENSITIES VARY BETWEEN DATA PROCESSING TOOLS

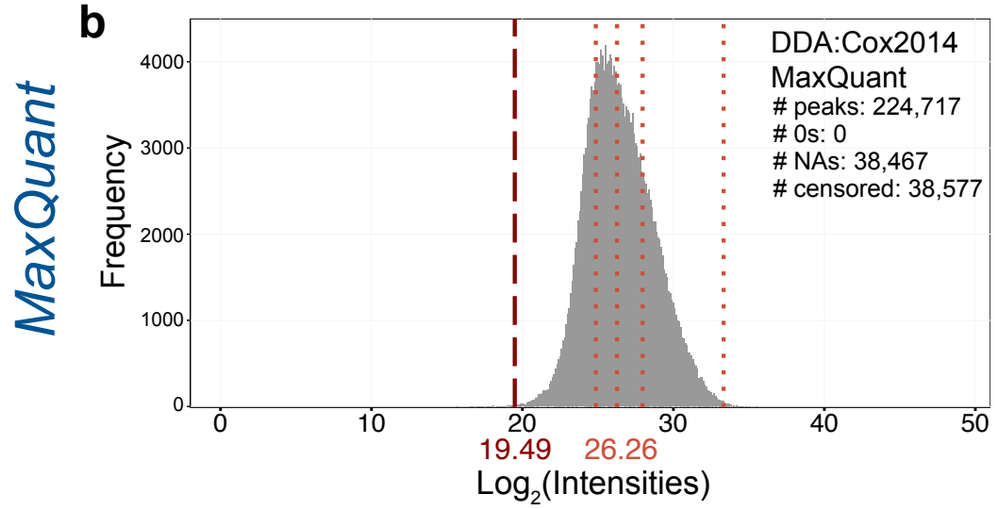
DDA: Cox 2014

DIA: Bruderer 2015

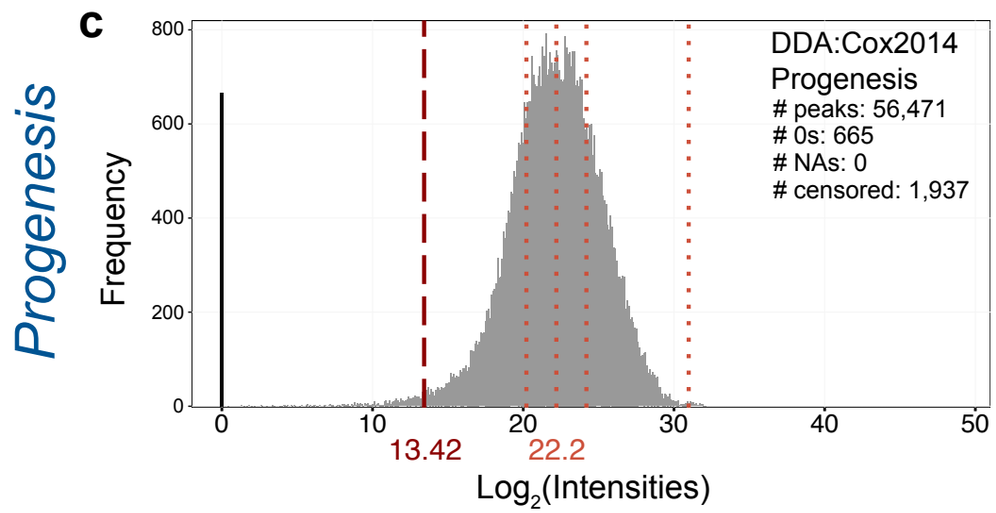
Skyline



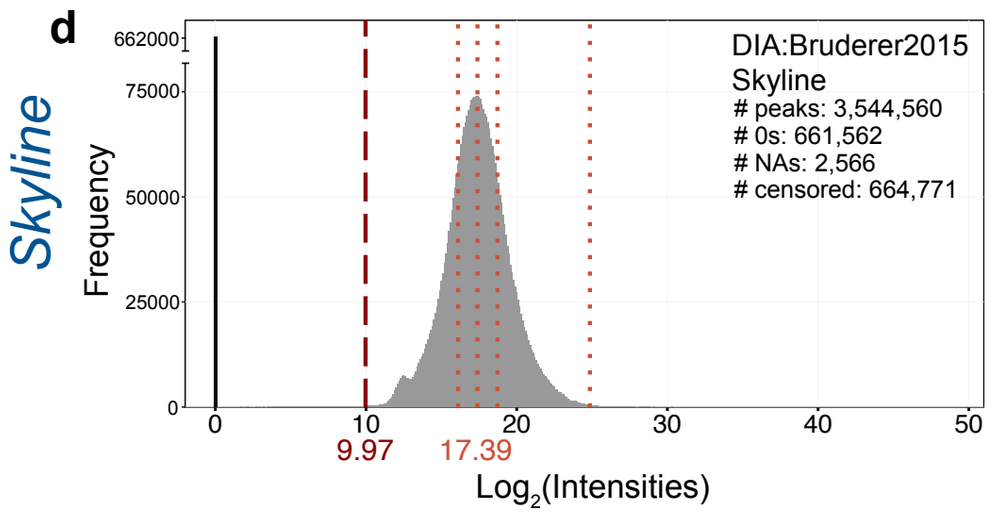
MaxQuant



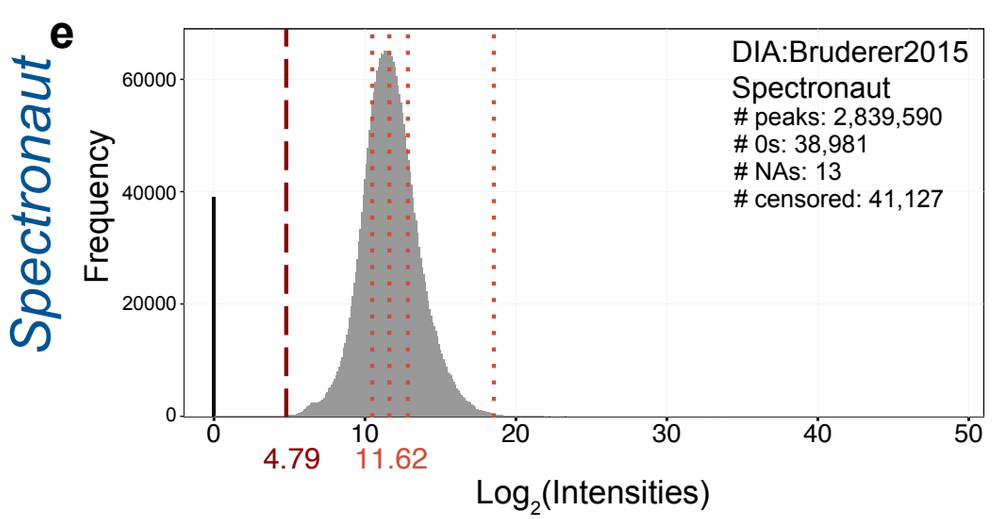
Progenesis

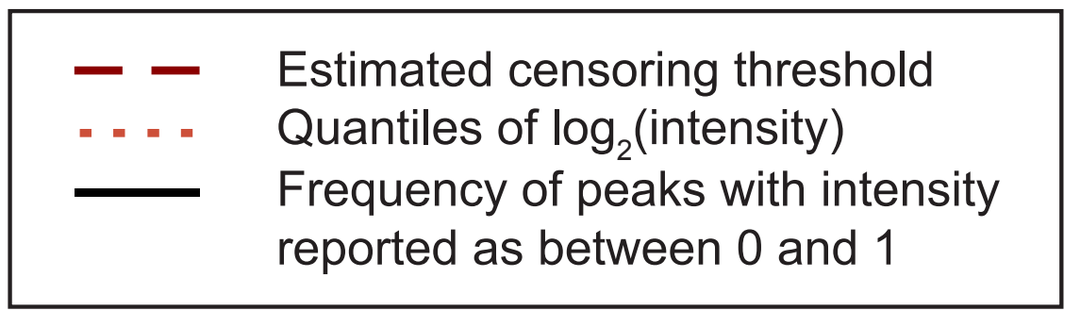


Skyline



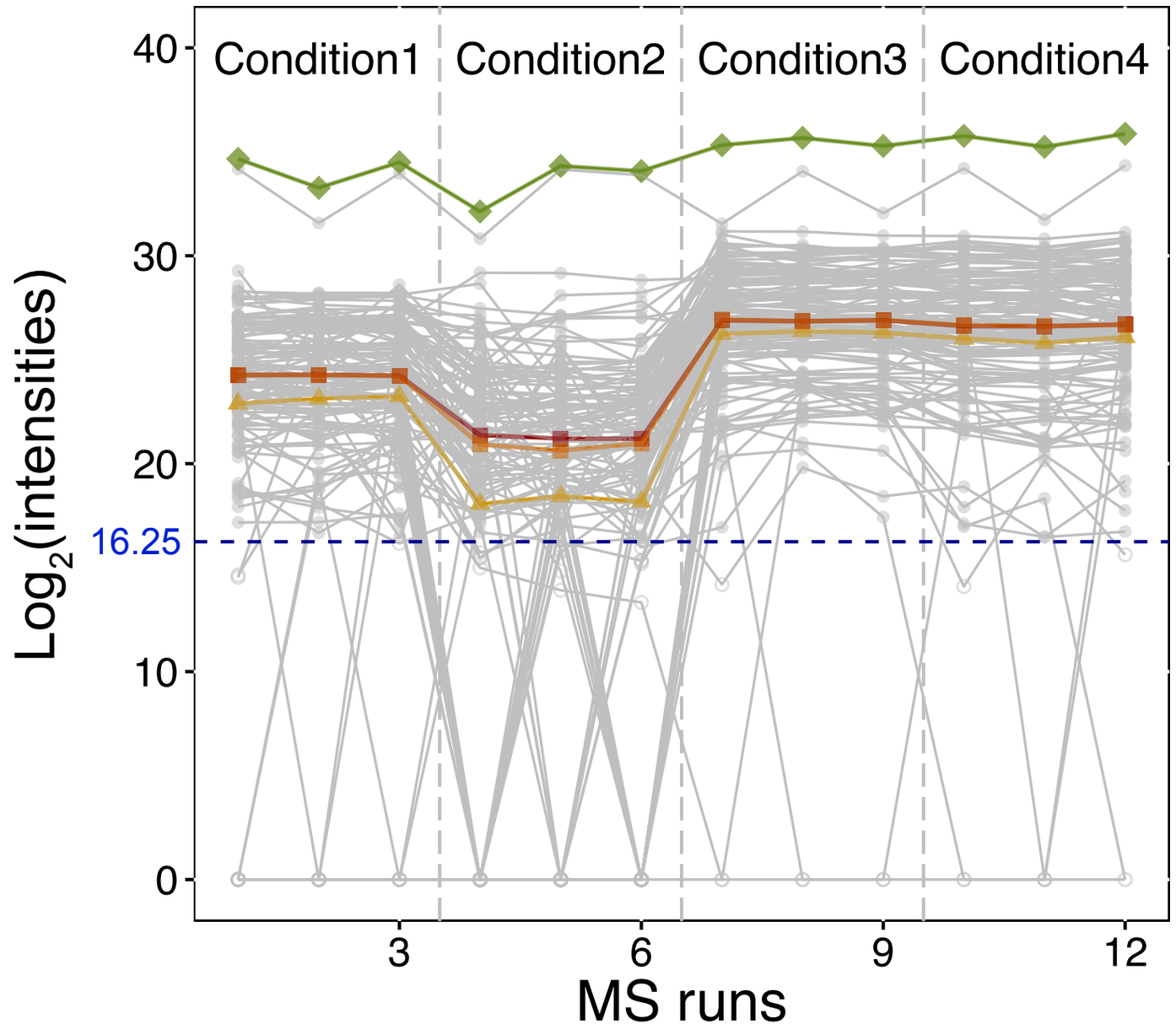
Spectronaut

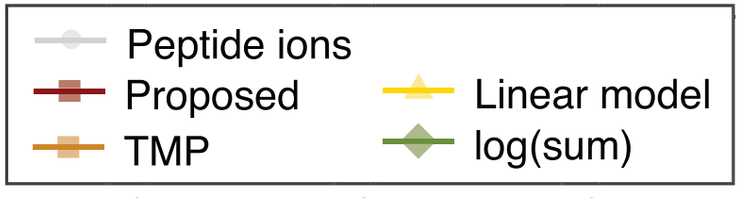




ROBUSTNESS TO OUTLIERS

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

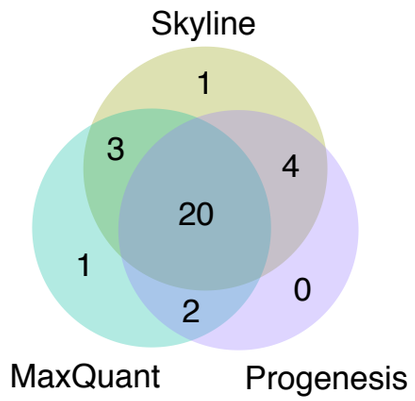
BETTER STATISTICAL METHODS ENHANCE REPRODUCIBLE RESEARCH

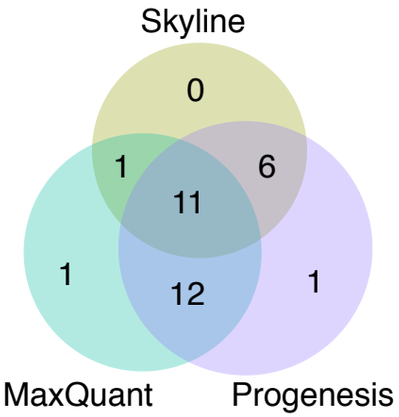
Better agreement in #differentially abundant proteins between tools

Proposed

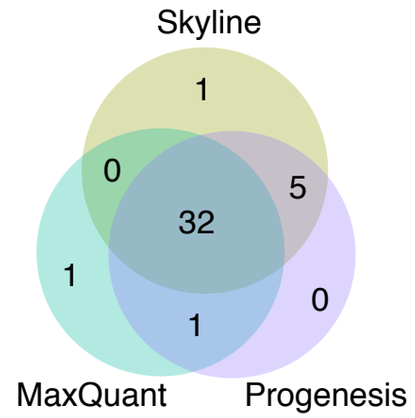
Log(sum)

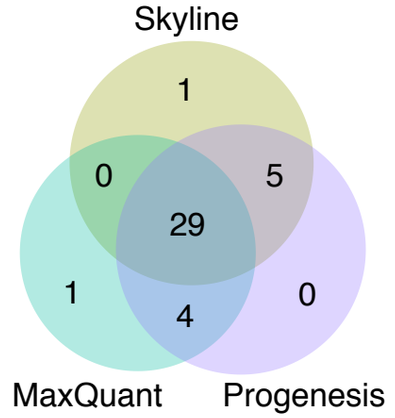
DDA: iPRG2015



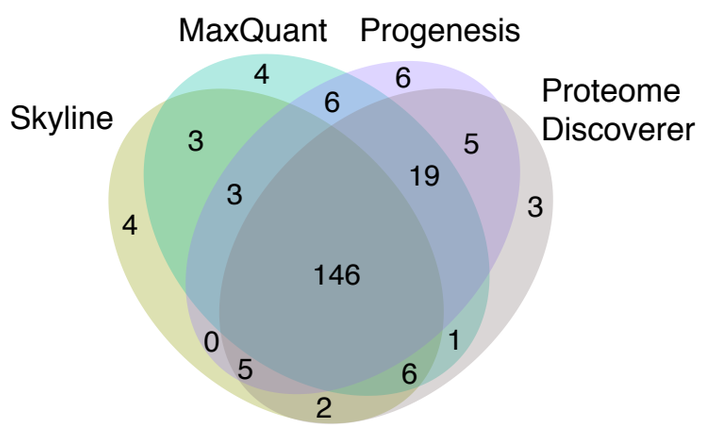


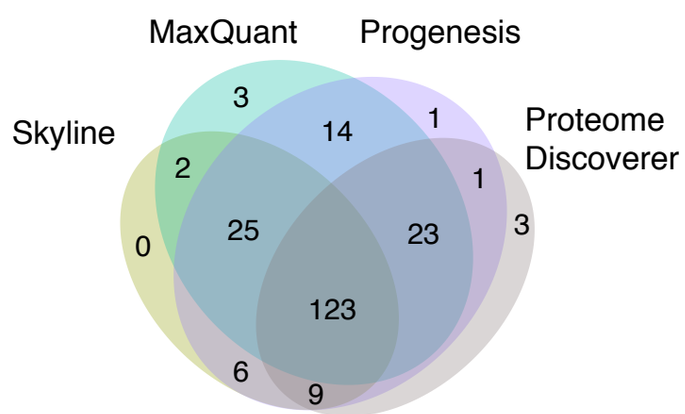
DDA: Cox 2014





DDA: Spike-in





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 th previous version

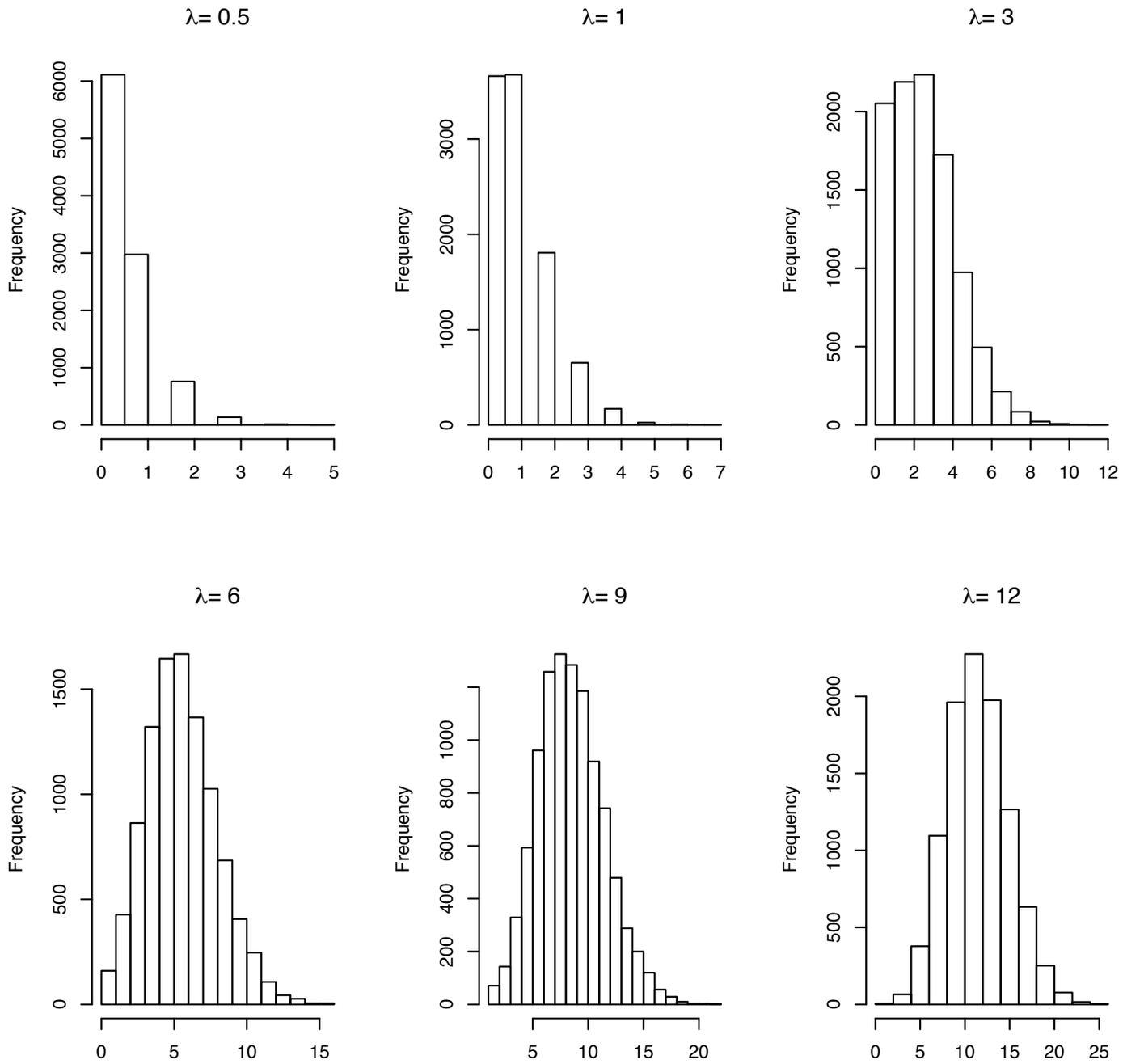
www.msstats.org

CHALLENGES

- Normalization
  - Remove systematic shifts in signals between runs
- Statistical modeling
  - Limma: continuous features, small n
  - MSstats: many continuous features per analyte
  - DEseq2: counts per analyte
- Hypothesis testing
  - Multiple comparisons

GENERALIZED LINEAR MIXED MODELS

Extension of linear models for counts data



Poisson distribution: counts of gene reads in a condition

λ population mean

Goal: compare population means between conditions

DESEQ2

Models variation: Negative Binomial distribution

- Model: Negative Binomial distribution

$$Y_{gij} \sim \mathcal{NB}\left(M_{ij} \cdot \mu_{gi},\ \phi_{gi}\right)$$

  - Allow a different ϕ_{gi} per condition

- Calculate size factors for normalization
  - Median ratio of observed counts to reference

$$\widehat{M}_{ij} = median_i\ M_{ij}\ /\ (\prod_{j=1}^{J} M_{ij})^{1/J}$$

- Estimate variance as function of the mean
  - Estimate $\widehat{Var}\{Y_{gij}\}$ by the method of moments (i.e. by the per-gene sample variance)
  - Model variance as function of the mean
  - Use a non-parametric model, or

$$\hat{V}(\hat{\mu}_{gi}) = \hat{M}_{ij} \cdot \hat{\mu}_{gi} + \hat{M}_{ij}^2 \cdot (a_0 + a_1/\hat{\mu}_{gi})$$

- From the fit, obtain $\hat{\phi}_{gi}$

M_{ij}

Size of library *j* of group *i* (used for normalization)

μ_{gi}

Population mean of gene *g* in group *i* (of interest for the comparison)

ϕ_{gi}

Nuisance variation (estimated from all genes)

Anders and Huber,. Genome Biology, 2010

CHALLENGES

- Normalization
  - Remove systematic shifts in signals between runs
- Statistical modeling
  - Limma: continuous features, small n
  - MSstats: many continuous features per analyte
  - DEseq2: counts per analyte
- Hypothesis testing
  - Multiple comparisons

PITFALL: MULTIPLE TESTING

- An fMRI on dead fish
- Found many active brain regions
  - All background noise and random variation

Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

[1] Psychology Department, University of California Santa Barbara, Santa Barbara, CA; [2] Department of Psychology, Vassar College, Poughkeepsie, NY;
[3] Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

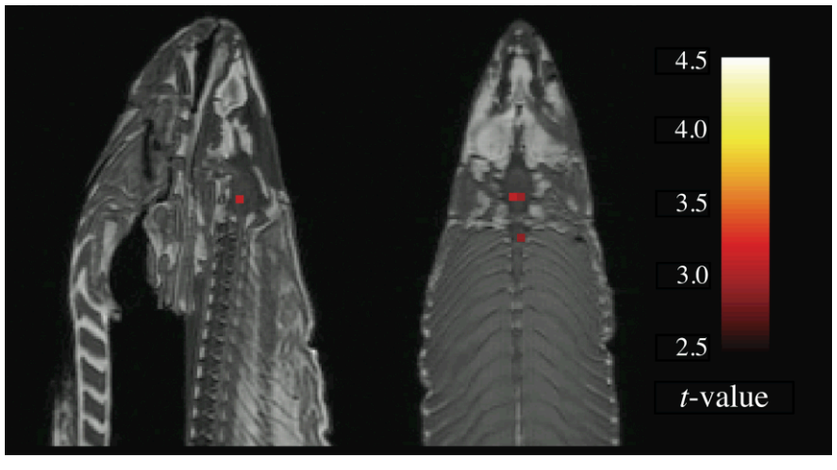
INTRODUCTION

With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the 130,000 voxels in a typical fMRI volume the probability of a false positive is almost certain. Correction for multiple comparisons should be completed with these datasets, but is often ignored by investigators. To illustrate the magnitude of the problem we carried out a real experiment that demonstrates the danger of not correcting for chance properly.

METHODS

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at

GLM RESULTS



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/

MULTIPLE TESTING

Control False Positive Rate for two proteins

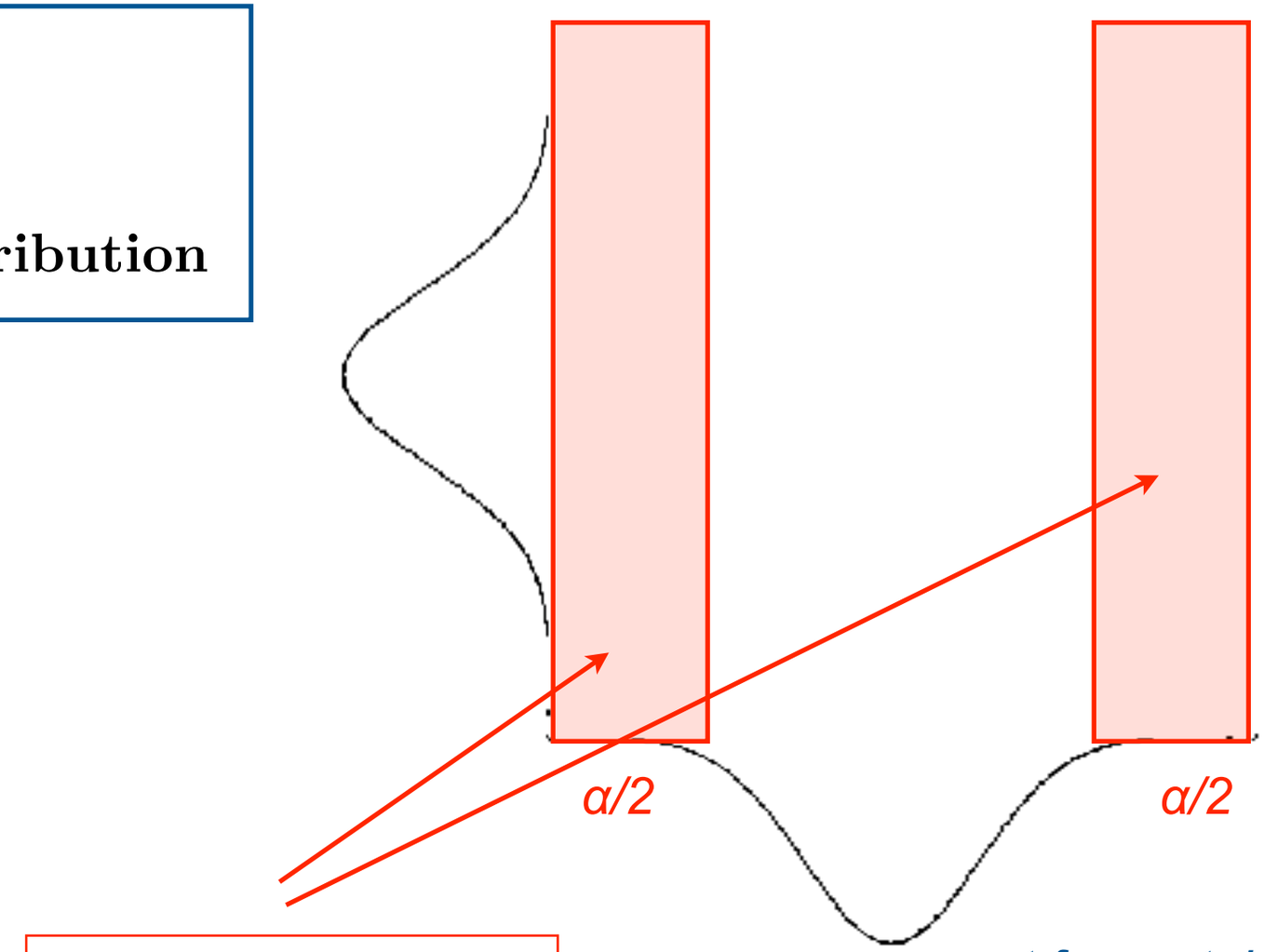
For each protein:

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

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

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

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



The area = FPR = α

MULTIPLE TESTING

Control False Positive Rate for two proteins

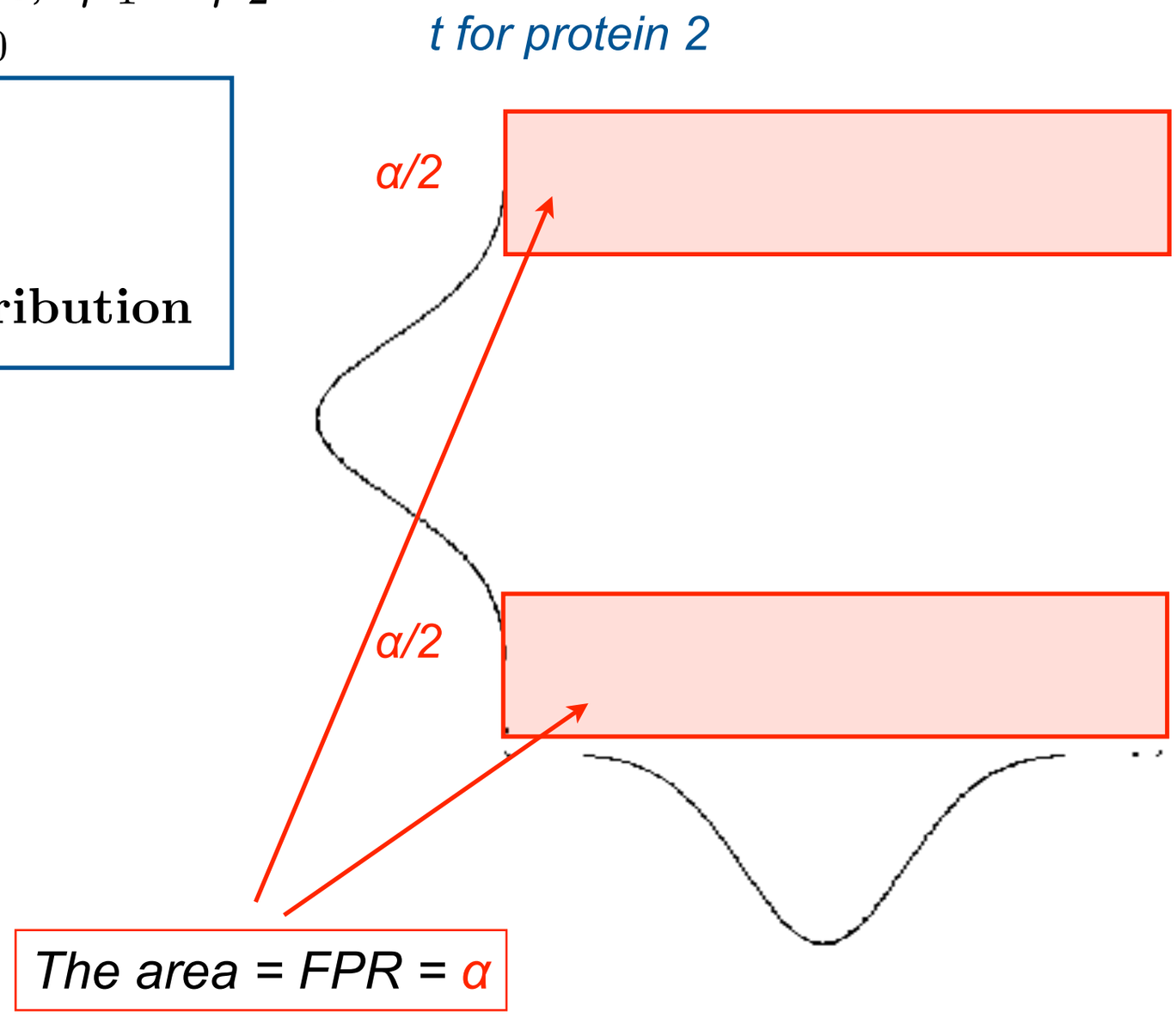
For each protein:

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:

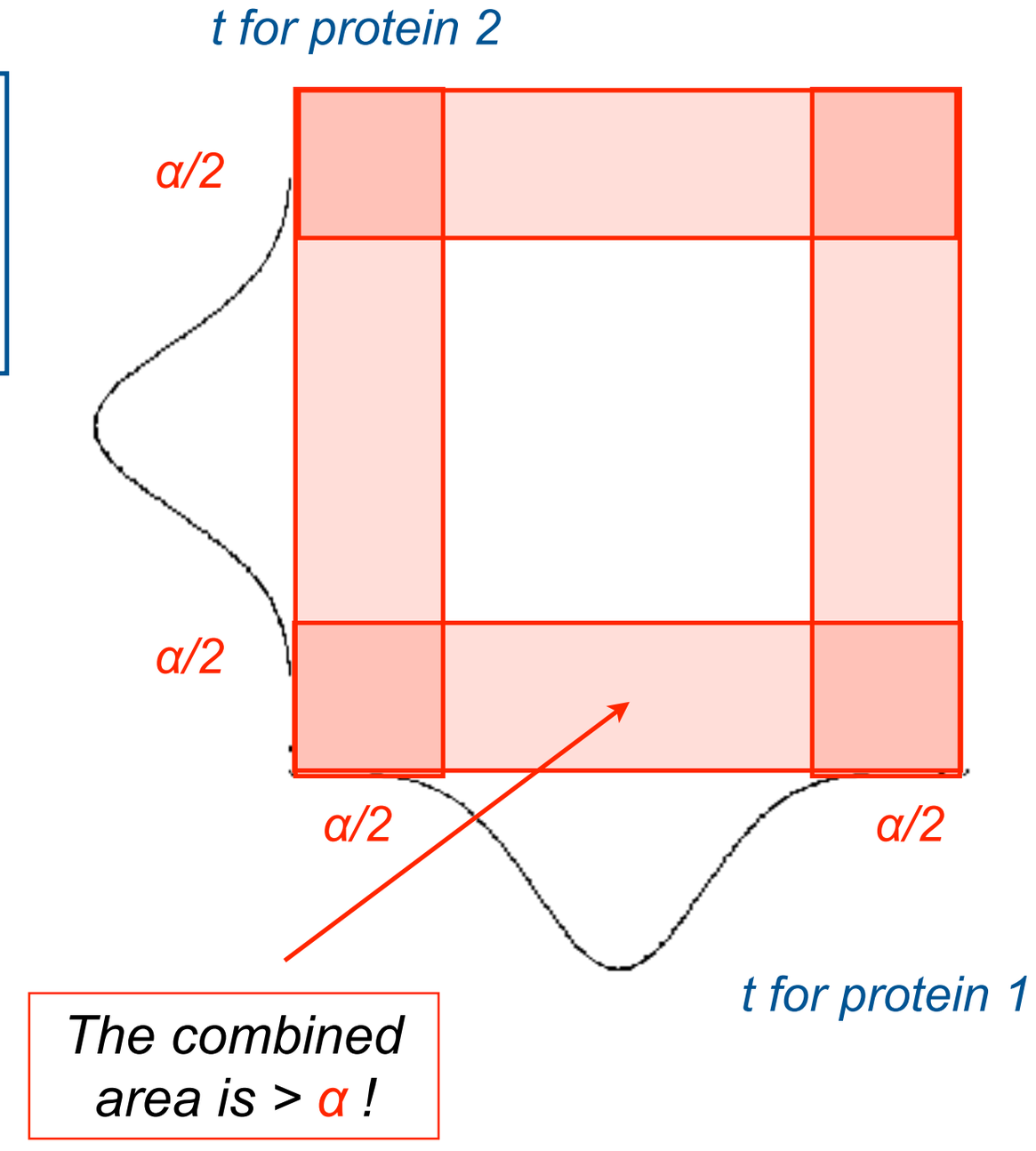
H0: 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$

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:

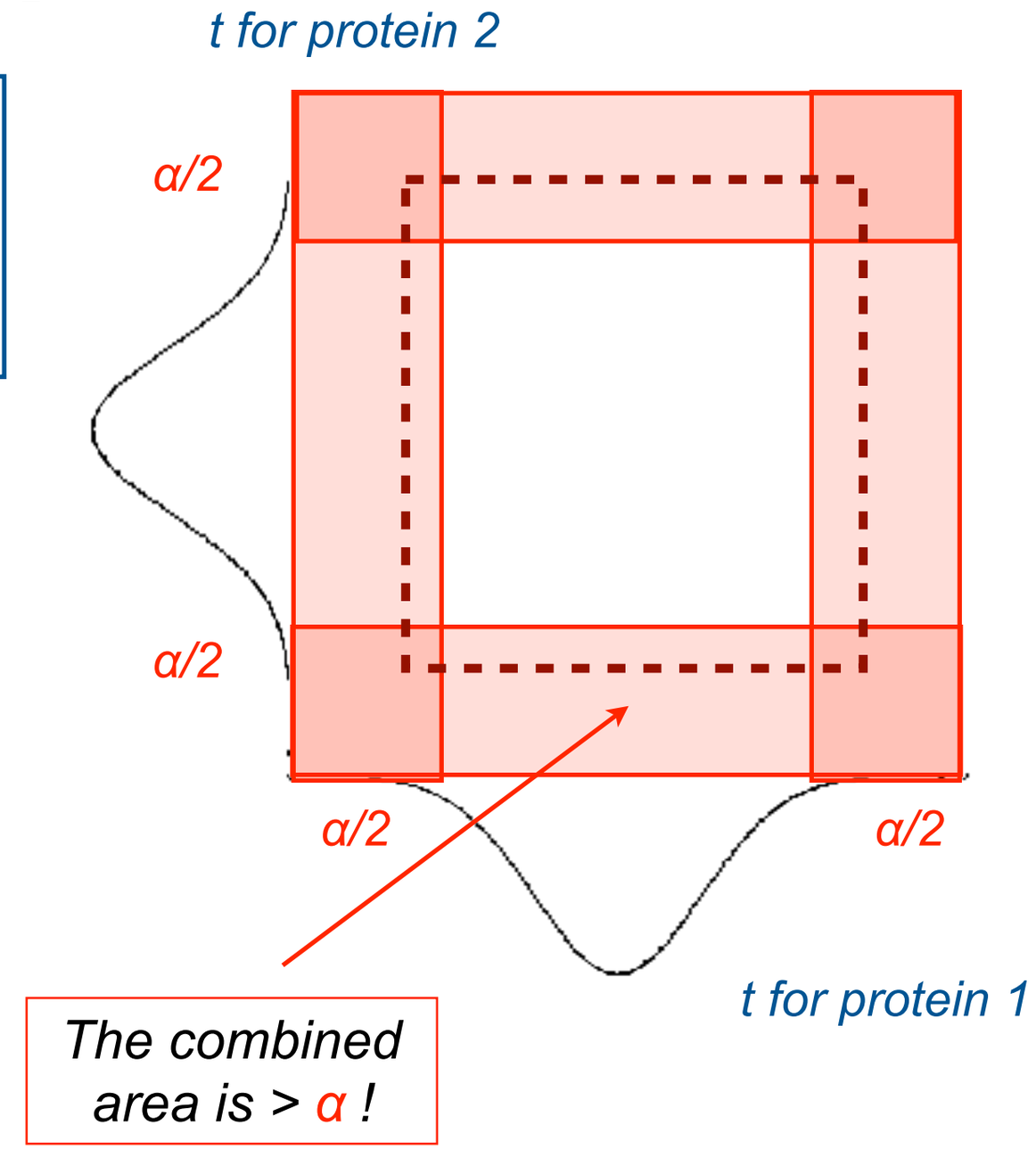
H0: 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$

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

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

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

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



TESTING M PROTEINS

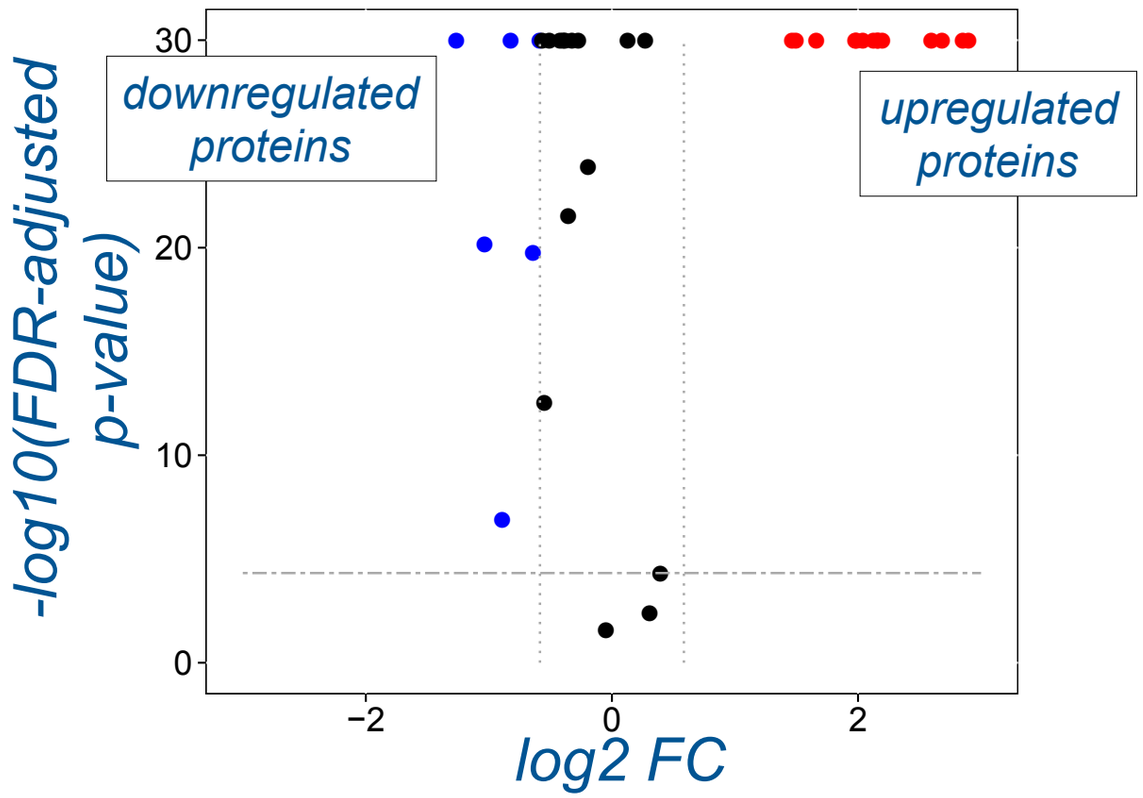
Change criteria from False Positive Rate to False Discovery Rate

	# 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

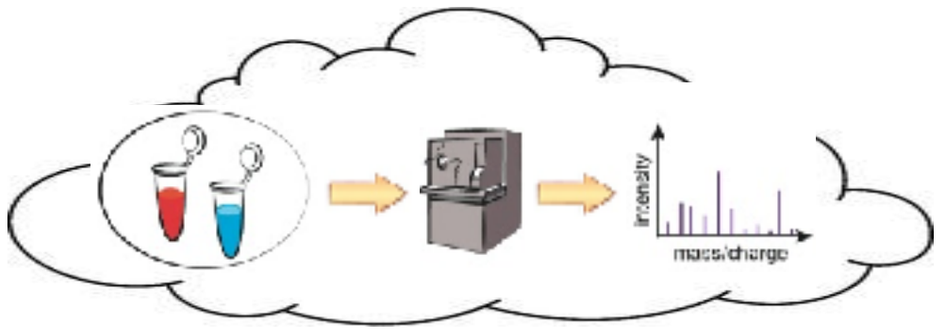
$$\mathbf{FDR} = \mathbf{E}\left[\frac{\mathbf{V}}{\max(\mathbf{R}, 1)}\right]$$

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



A TYPICAL ANALYSIS WORKFLOW

Experimental design



QC and normalization

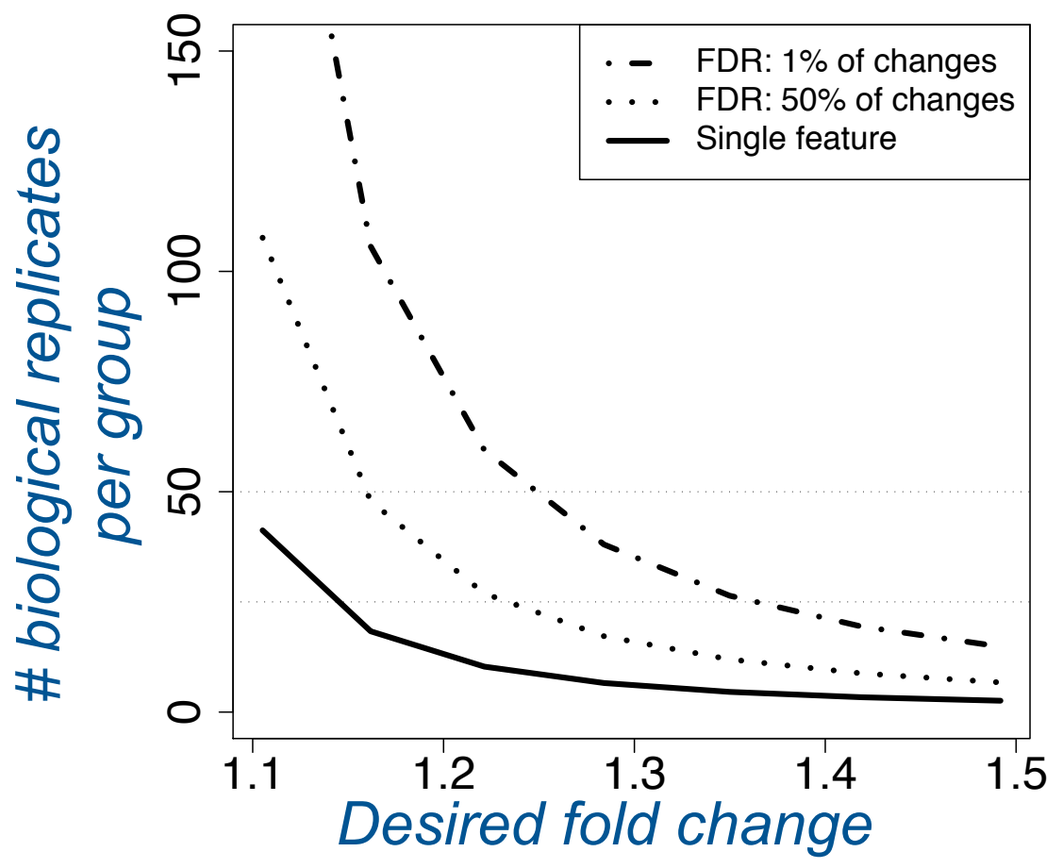
Statistical modeling

Model-based conclusions

Experimental design

Use the dataset to improve:

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



DIFFICULTY: MANY FEATURES ARE OF INTEREST

Would like to control the False Discovery Rate:

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

$$\mathbf{q} = \mathbf{E}\left[\frac{\mathbf{V}}{\mathbf{max(R, 1)}}\right]$$ = the "average" proportion of false positives

This changes the sample size calculation:

Fix: $\mathbf{q}$ - the False Discovery Rate

$\mathbf{m_0/m_1}$ - anticipated ratio of unchanging features

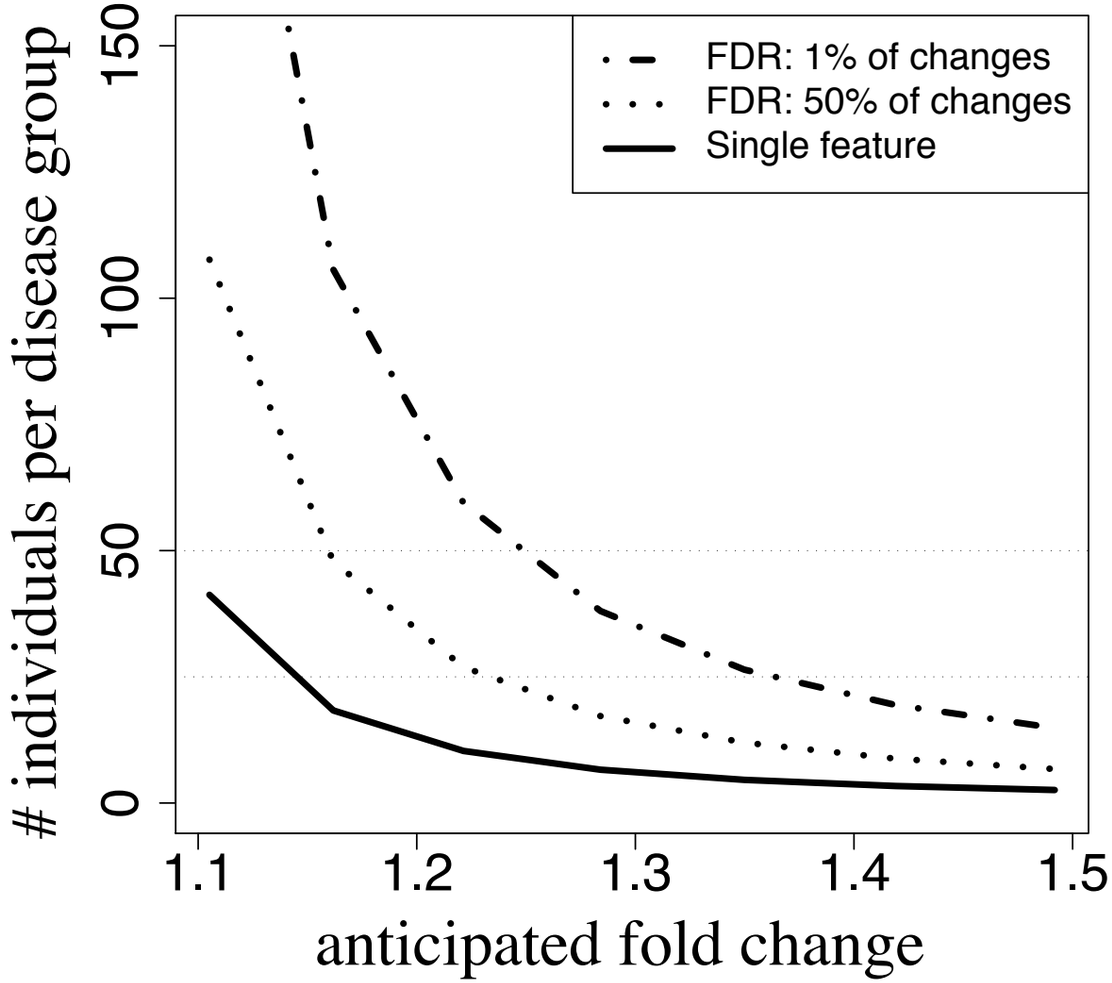
This defines

$$\alpha_{\text{ave}} \leq (\mathbf{1} - \beta)_{\text{ave}} \cdot \mathbf{q}\frac{\mathbf{1}}{\mathbf{1} + (\mathbf{1} - \mathbf{q}) \cdot \mathbf{m_0/m_1}},$$

- average probability of a false positive discovery

SO HOW MANY REPLICATES DO I NEED?

Example: pilot study with diabetes patients.
A block-randomized design



Conclusion:
The fewer changes we expect, the larger the sample size

Oberg and Vitek, *JPR*, 2009