# STATISTICAL METHODS FOR HIGH-THROUGHPUT BIOLOGY

Olga Vitek

College of Science

College of Computer and Information Science

Northeastern University

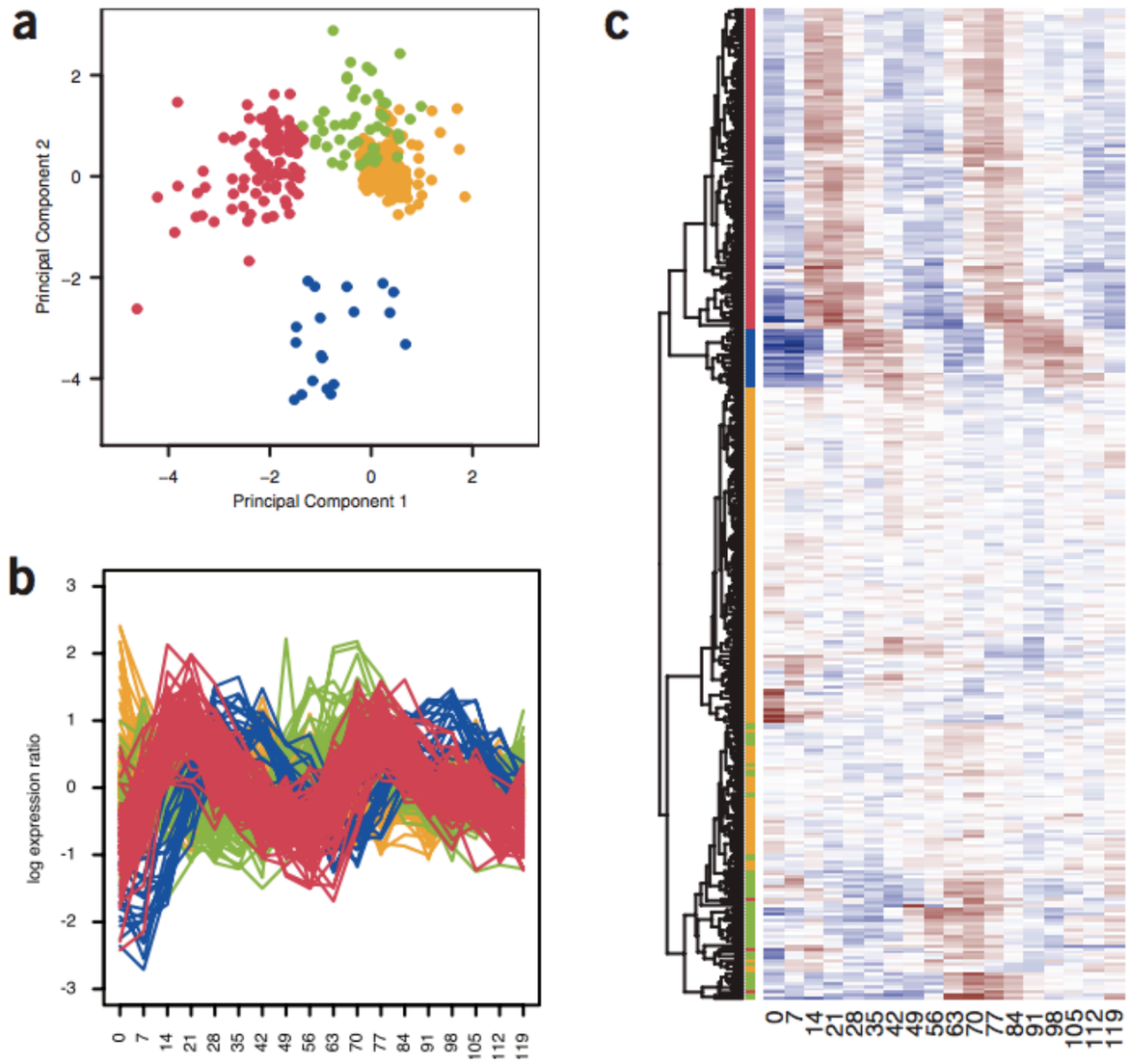
# CHALLENGES

- Statistical goals of the experiments
  - Class discovery, class comparison, class prediction
- Class comparison: methods from genomics
  - Limma: continuous data, small n (microarrays)
  - DEseq2: count data, small n (RNA-seq)
- Class discovery
  - Principle component analysis
  - Hierarchical clustering

# 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

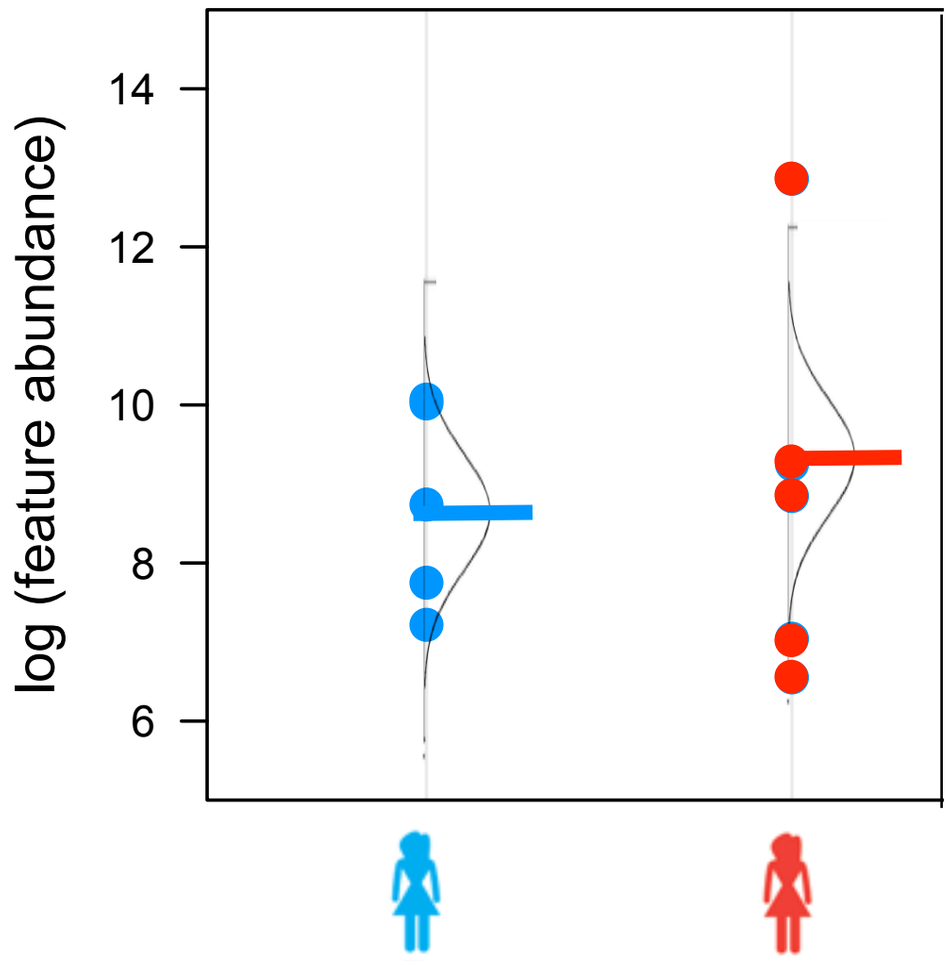



Gehlenborg *et al*, Nature Methods, 2010

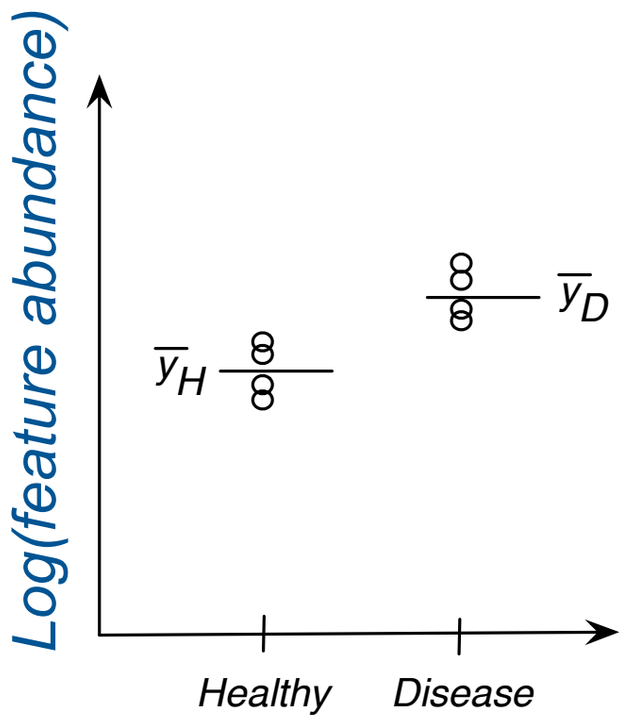
# STATISTICAL GOAL 2: CLASS COMPARISON

## Compare mean abundances in subject groups

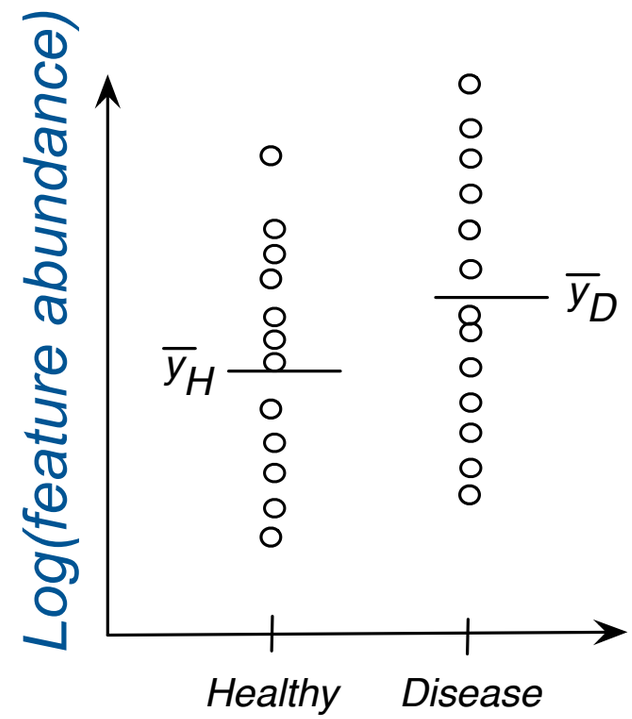
- Known class labels
  - Compare group averages
  - Report p-values, posterior probabilities etc
- Useful when compare groups of subjects
  - Best used for basic biology
  - Initial (Tier III) biomarker discovery screen



# DIFFERENTIALLY ABUNDANT PROTEINS ARE NOT ALWAYS BIOMARKERS



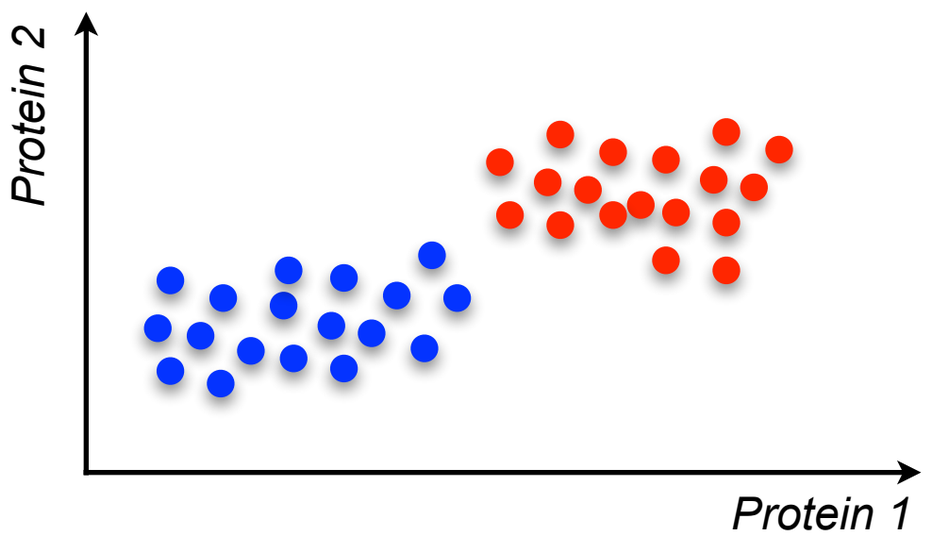

*Differentially abundant and predictive*



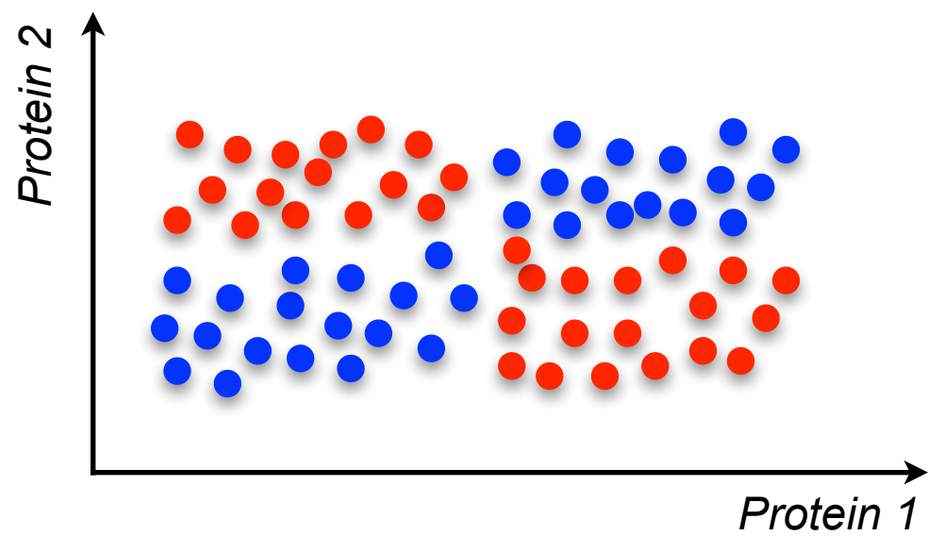

*Differentially abundant and not predictive*

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

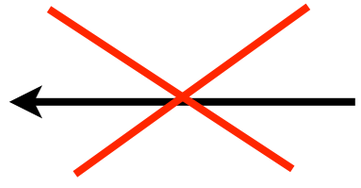
# BIOMARKER PROTEINS ARE NOT ALWAYS DIFFERENTIALLY ABUNDANT




*Differentially abundant and predictive*



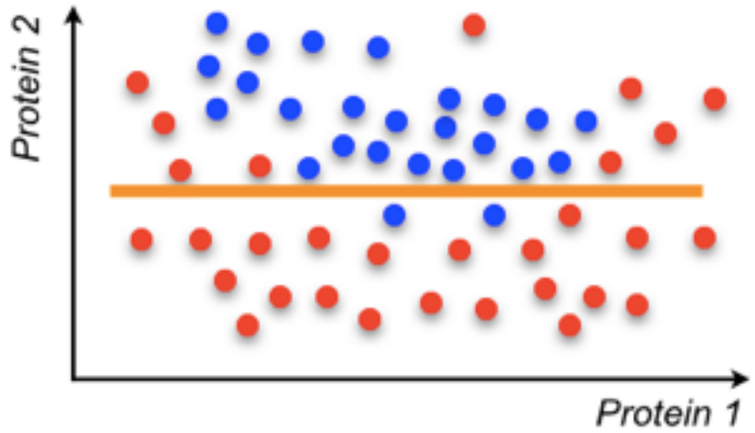

*Not differentially abundant but predictive*

*Single protein:* *Differentially abundant*  *Predictive*

# STATISTICAL GOAL 3: CLASS PREDICTION

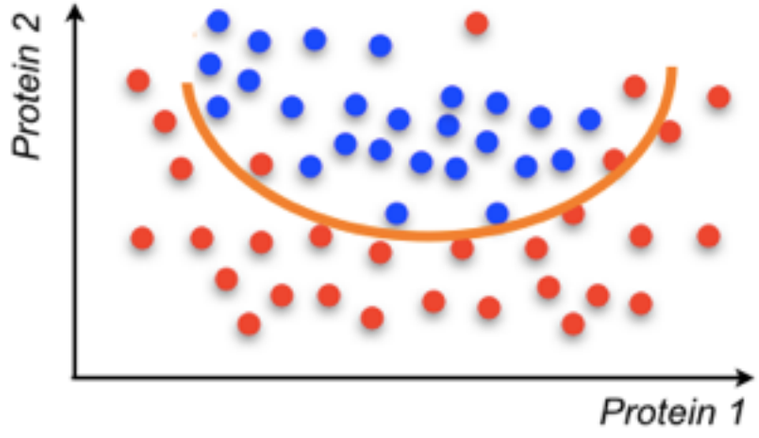
## Classify each subject into a known group

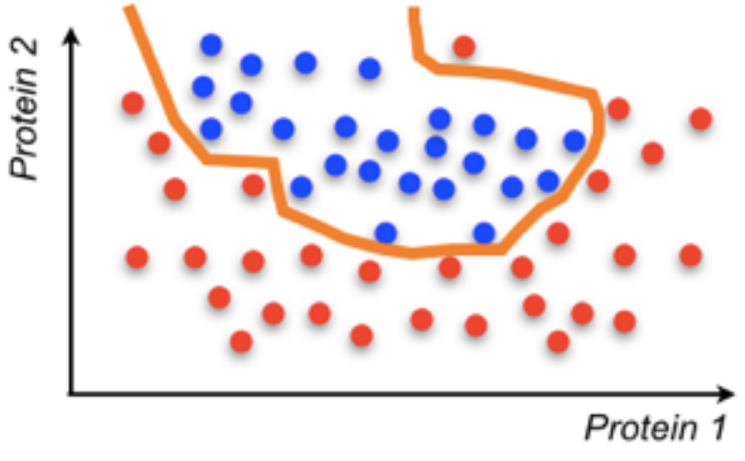
- Known class labels
  - Predict individual subjects
  - Report misclassification error (sensitivity, specificity, predictive value etc)
- Useful when focus on an individual
  - Tier I or Tier II biomarker discovery studies




*Linear prediction rule*

*Quadratic prediction rule*



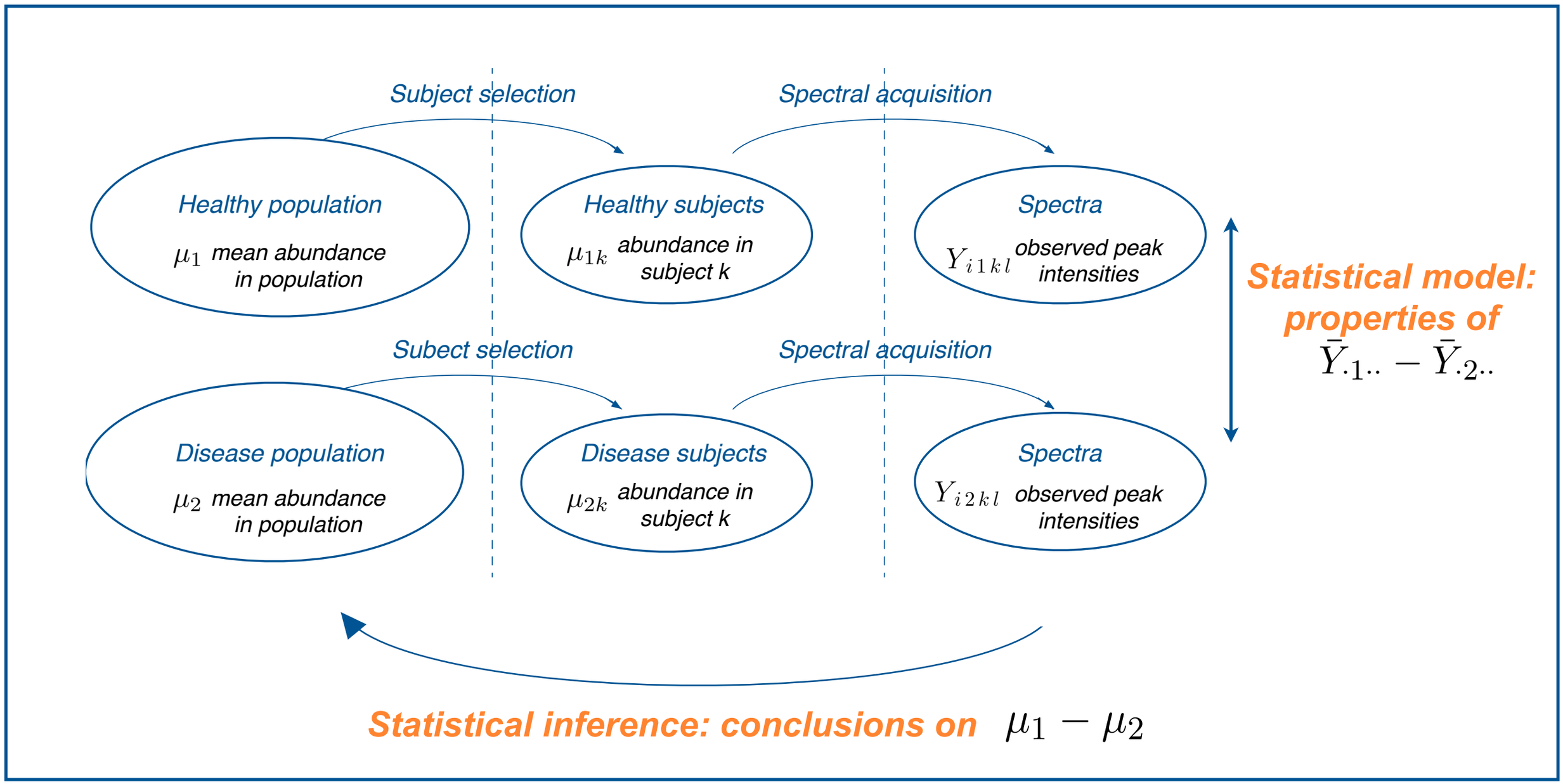




*General prediction rule*

# CHALLENGES

- Statistical goals of the experiments
  - Class discovery, class comparison, class prediction
- Class comparison: methods from genomics
  - Limma: continuous data, small n (microarrays)
  - DEseq2: count data, small n (RNA-seq)
- Class discovery
  - Principle component analysis
  - Hierarchical clustering

# DEFINITION OF BIAS AND INEFFICIENCY

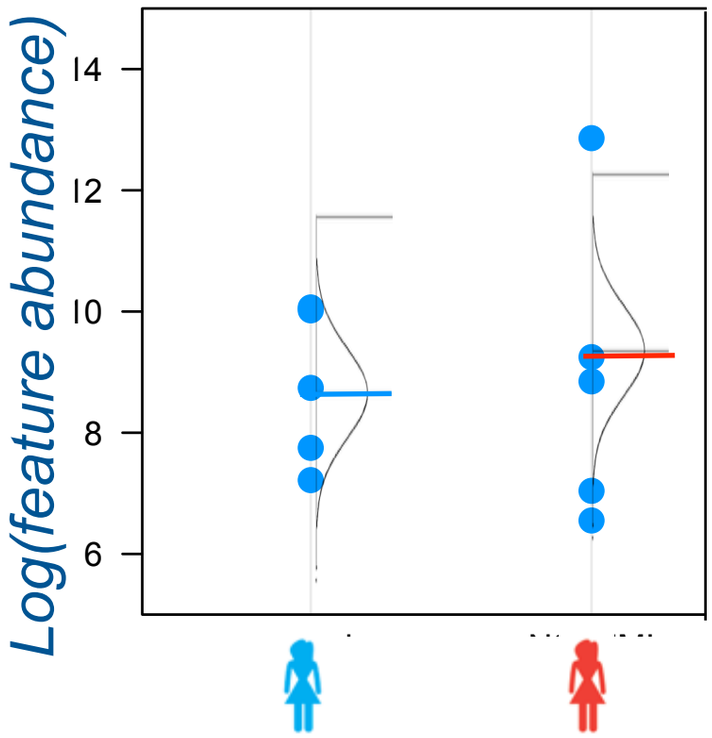



**Bias:** $\bar{Y}_{.1..} - \bar{Y}_{.2..}$ *systematically different from* $\mu_{1k} - \mu_{2k}$

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

# TESTING MANY HYPOTHESES IN PARALLEL

## Recall 2-sample t-test




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

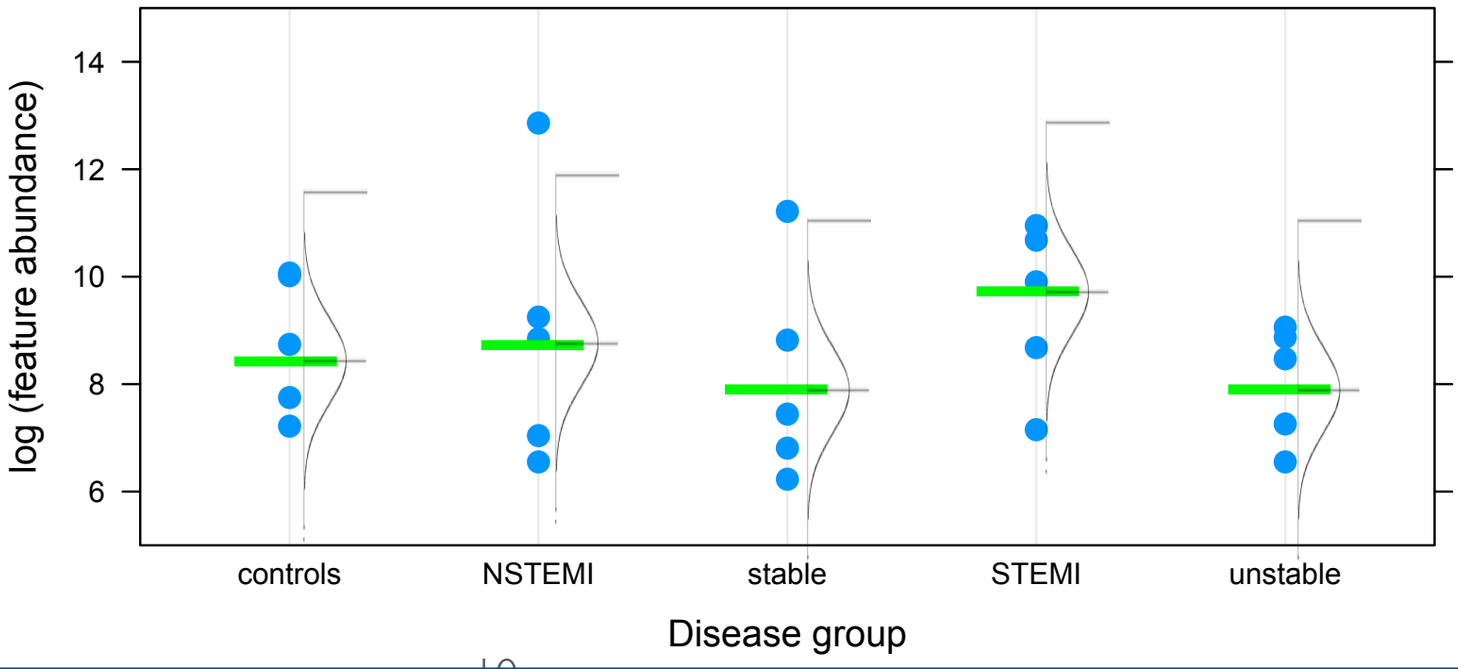
*Ha: change in abundance,* $\mu_1 - \mu_2 \neq 0$

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

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

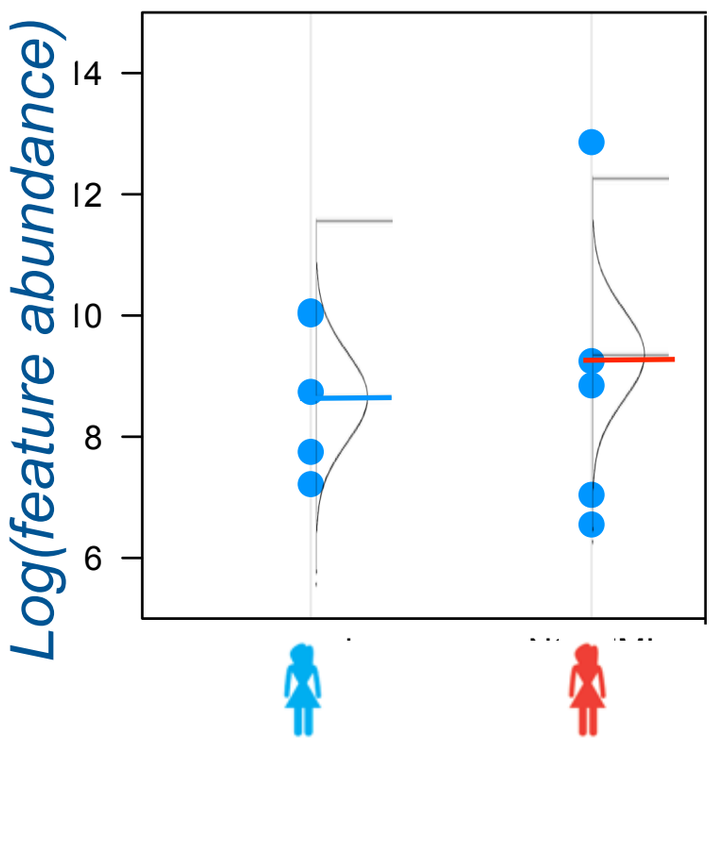
*Modification: assume same variance in both groups*

***Multiple conditions allow us to better learn the extent of variation***



# 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,* $\mu_1 - \mu_2 \neq 0$

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

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

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

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

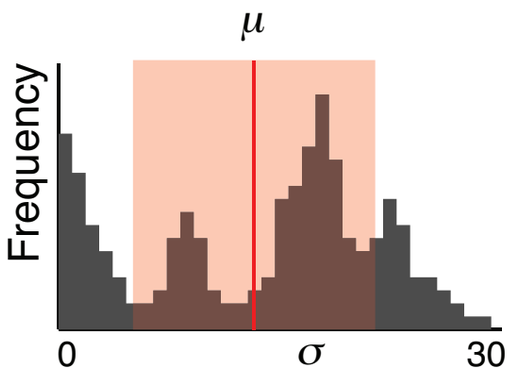
*replace with*

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

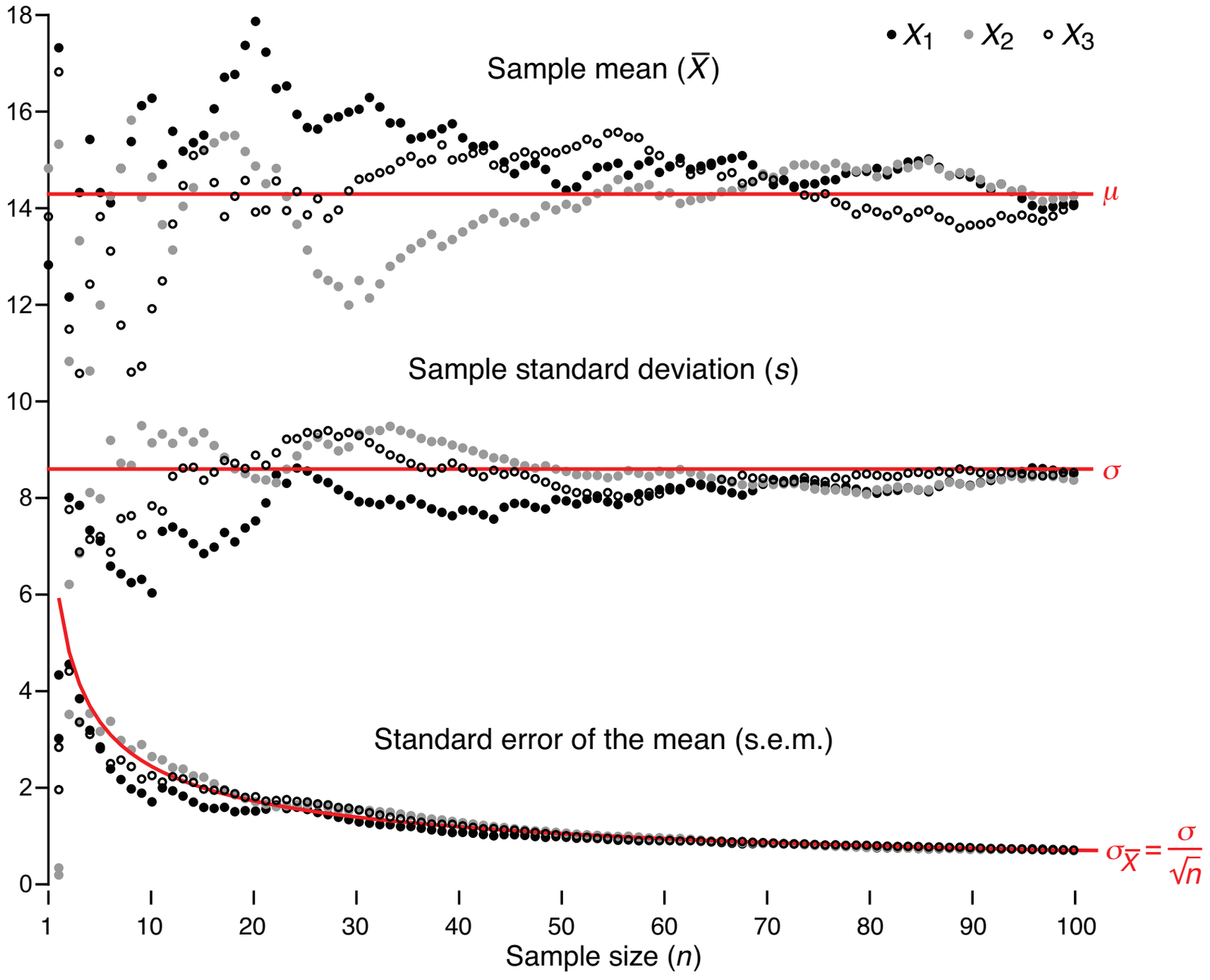
*Solution: we have many proteins. Can we learn from them?*

# EFFECT OF SAMPLE SIZE

## As n increases, the estimates stabilize




*Probability distribution of the data*




Simulated example

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

# LIMMA BORROWS INFORMATION FROM ALL PROTEINS

*Traditional approach:*

$$t = \frac{\text{difference of group means}}{\text{estimate of variation}} = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \text{Student}_d$$

*protein-specific variance*

*degree of freedom, reflects the number of replicates*

*Solution by linear models for microarrays (limma):*

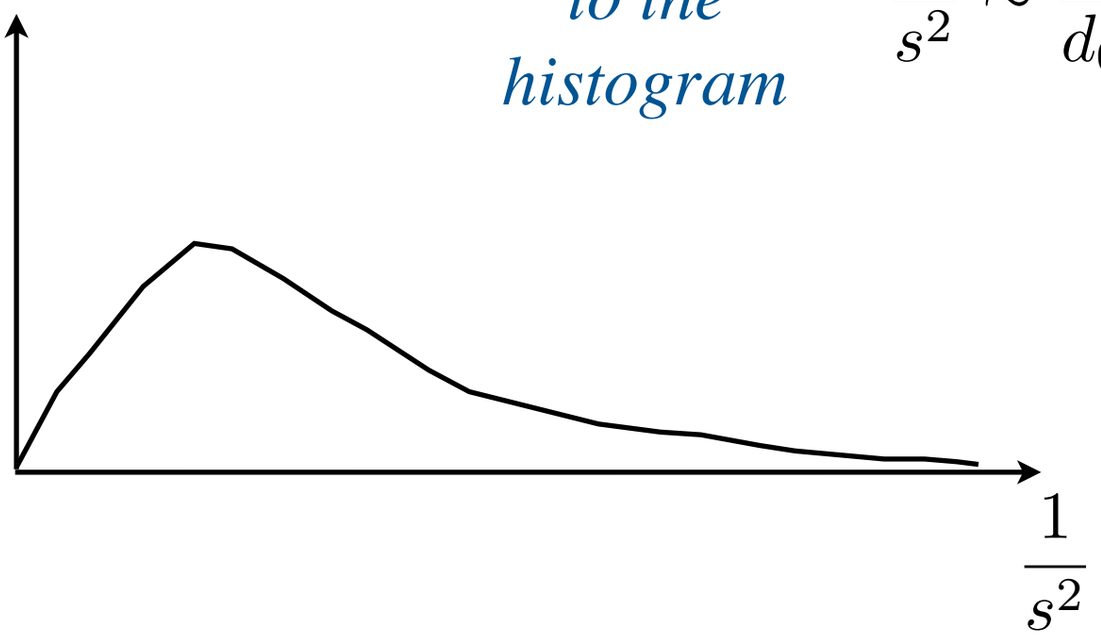
Smyth, 2005

*histogram over all proteins*

*fit the curve to the histogram*

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

*"average degree of freedom" over all proteins*




*"average variance" over all proteins*

# LIMMA BORROWS INFORMATION FROM ALL PROTEINS

*Traditional approach:*

$$t = \frac{\text{difference of group means}}{\text{estimate of variation}} = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \text{Student}_d$$

*protein-specific variance*

*degree of freedom, reflects the number of replicates*

*Solution by linear models for microarrays (limma):*

Smyth, 2005

$$t = \frac{\text{difference of group means}}{\text{estimate of variation}} = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{\tilde{s}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \text{Student}_{\tilde{d}}$$

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

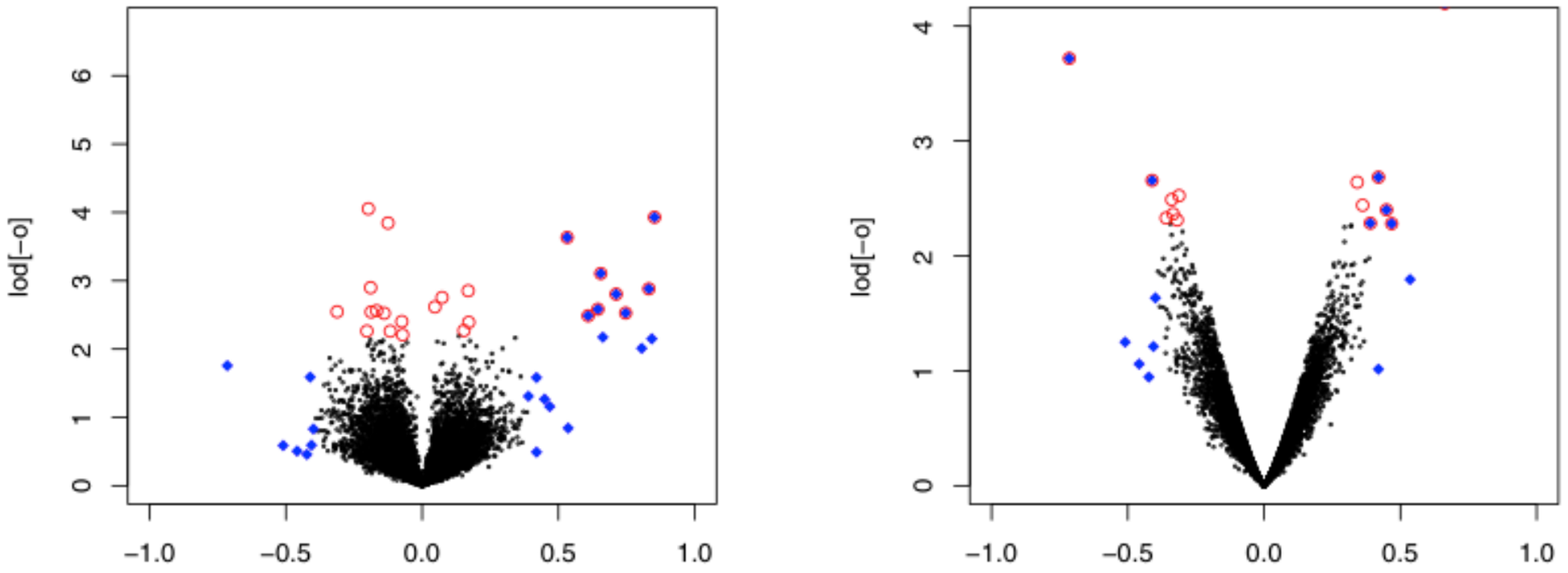
*"average variance" over all proteins*

*"average degree of freedom" over all proteins*

# 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



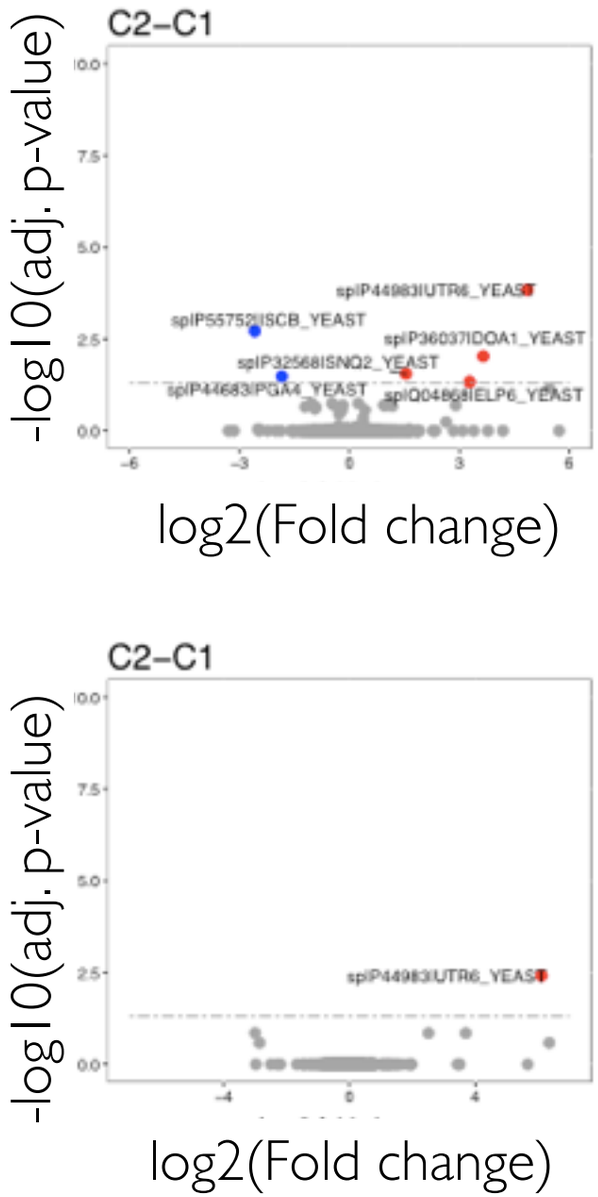

- Variance: sample (left), moderated (right)
- X axis: estimated log-fold change
- Y axis: $-log$(pValue) (here: $-log$(posteriorOdds))
- Moderated variance yields better agreement between log-fold change and significance

# LIMMA HAS LIMITATIONS IN PROTEOMICS

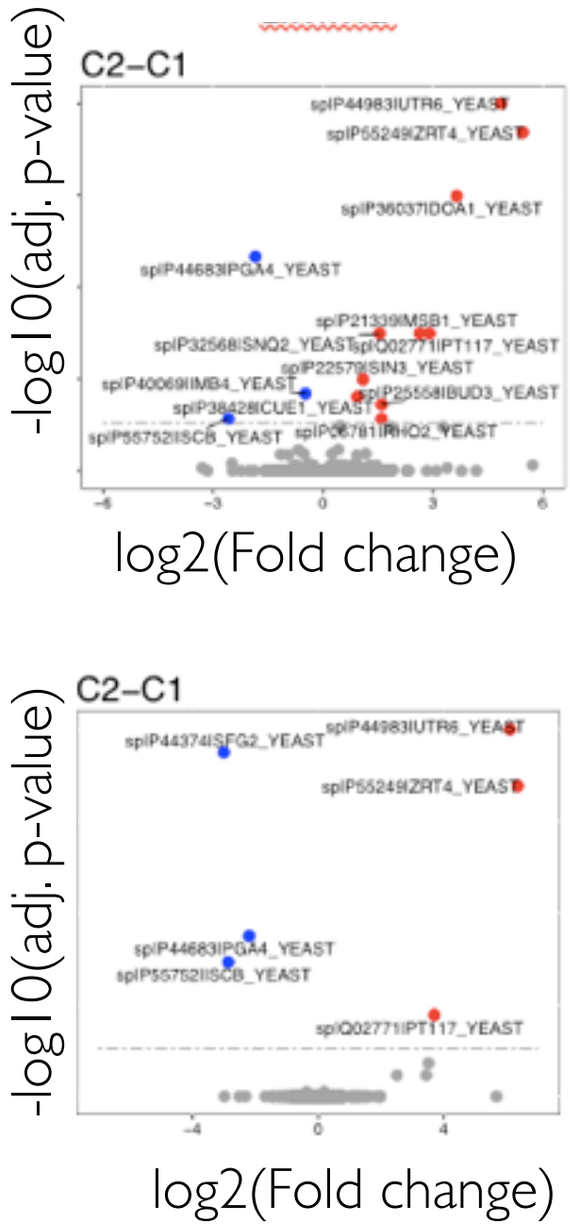
## Does not address summarization; limited experimental designs

**log(sum)**

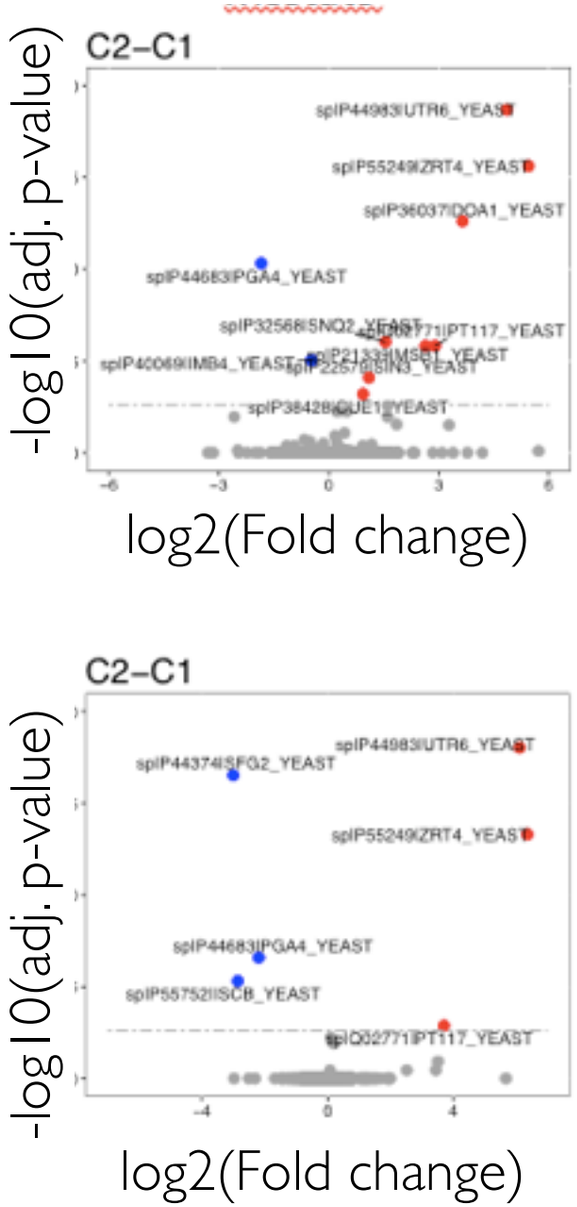
**T-test**



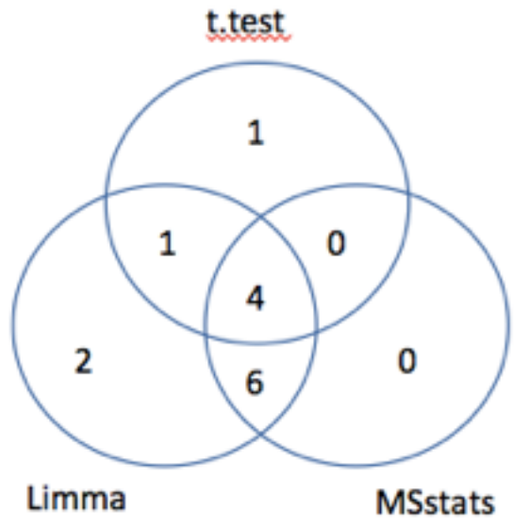

**Limma**




**MSstats**




adj. p-value cutoff




**MSstats summary**

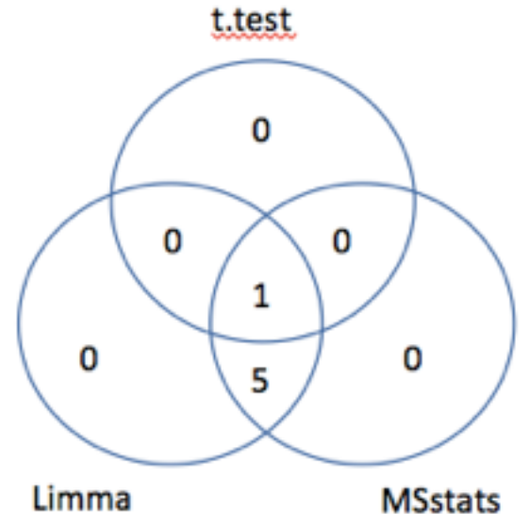








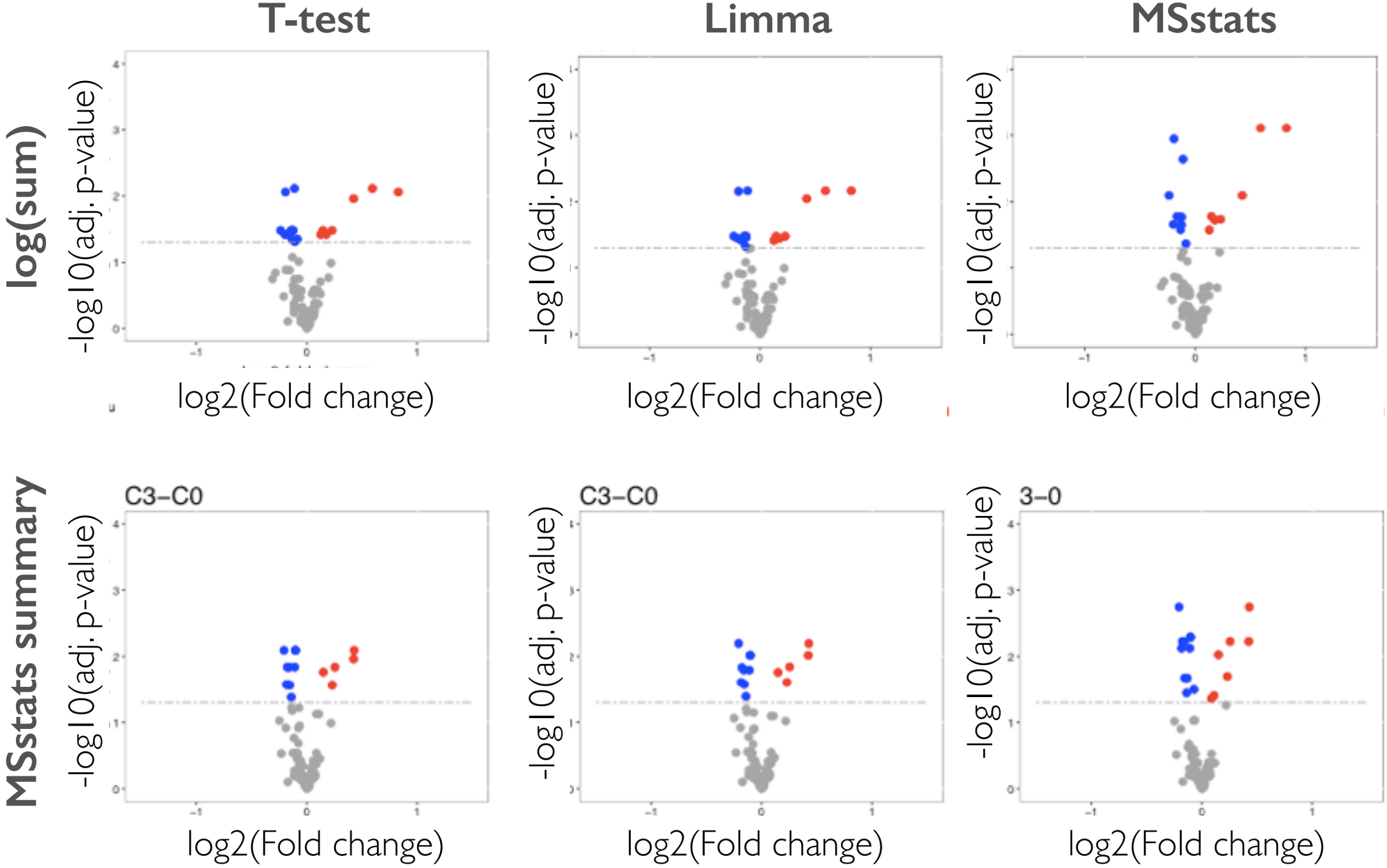

adj. p-value cutoff




iPRG peak intensities, comparing C1 vs C2, 3 replicates/group

# LIMMA HELPS MOSTLY WHEN SAMPLE SIZE IS SMALL

## No difference in results when sample size is large

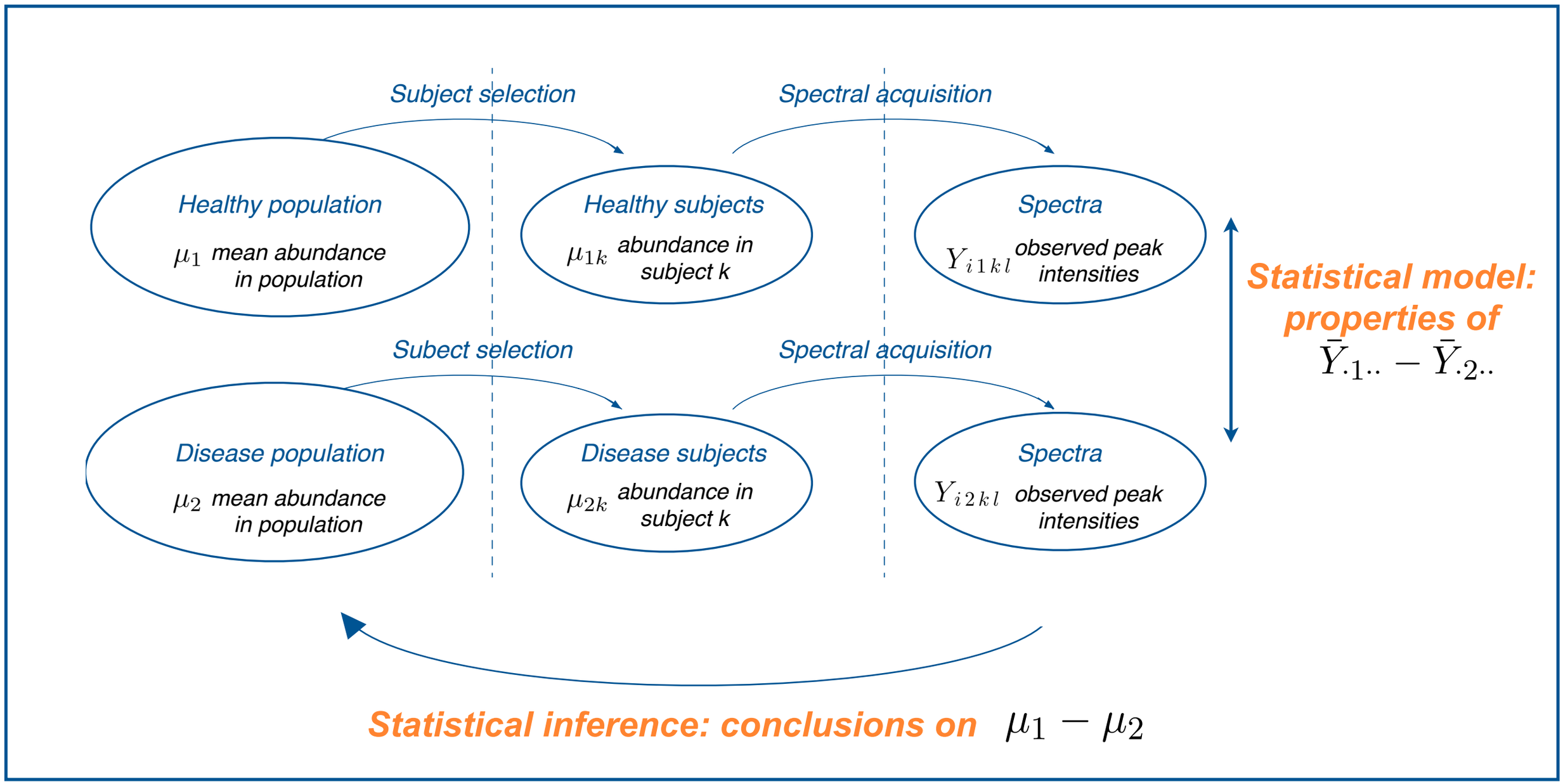



## Clinical experiment, 5- replicates/group

# CHALLENGES

- Statistical goals of the experiments
  - Class discovery, class comparison, class prediction
- Class comparison: methods from genomics
  - Limma: continuous data, small n (microarrays)
  - DEseq2: count data, small n (RNA-seq)
- Class discovery
  - Principle component analysis
  - Hierarchical clustering

# DEFINITION OF BIAS AND INEFFICIENCY




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

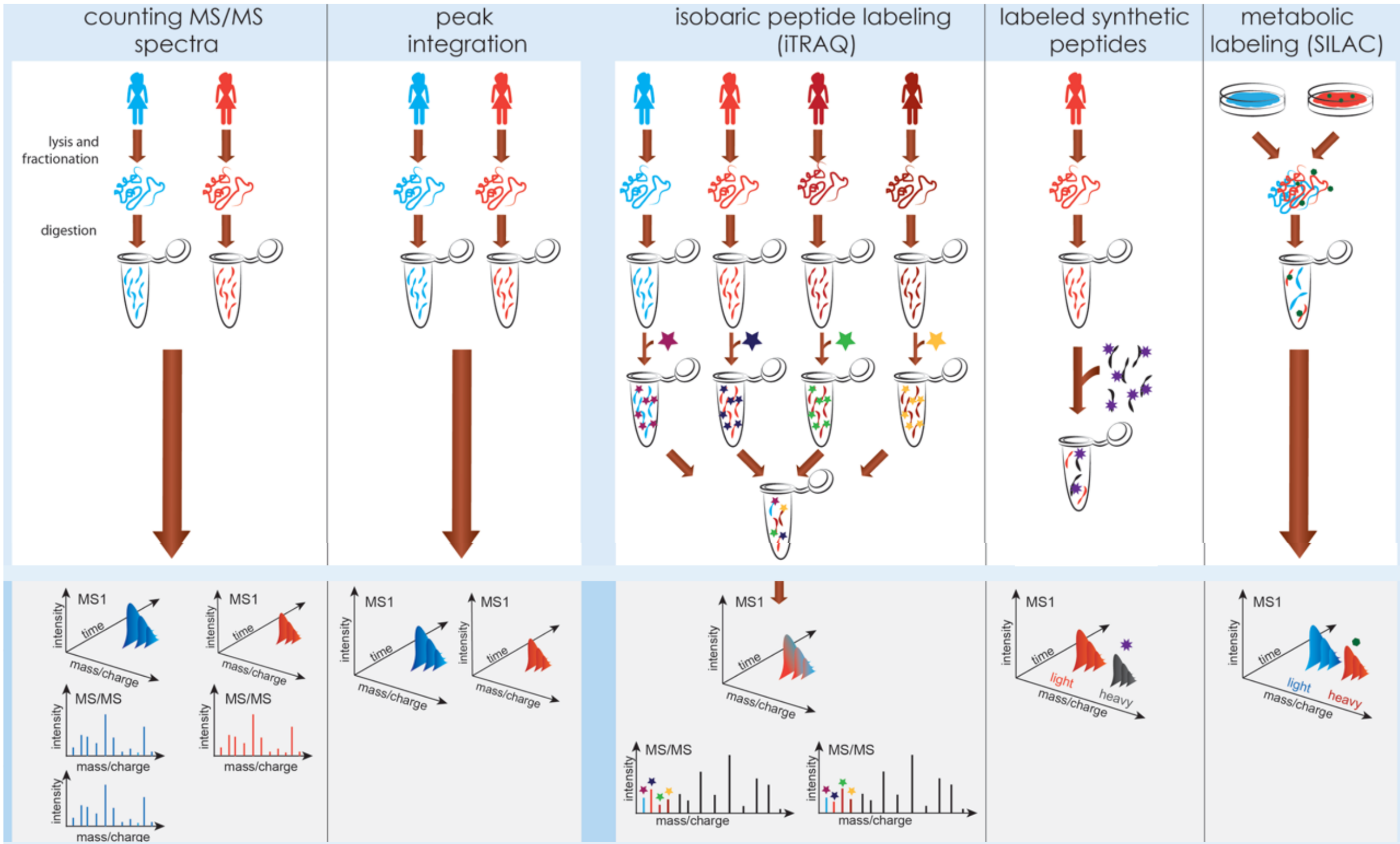
**Inefficiency:** *Large* $Var(\bar{Y}_{.1..} - \bar{Y}_{.2..})$

# COMPLEX DESIGNS: MULTIPLEXING

***Label-free*** ***Label-based***

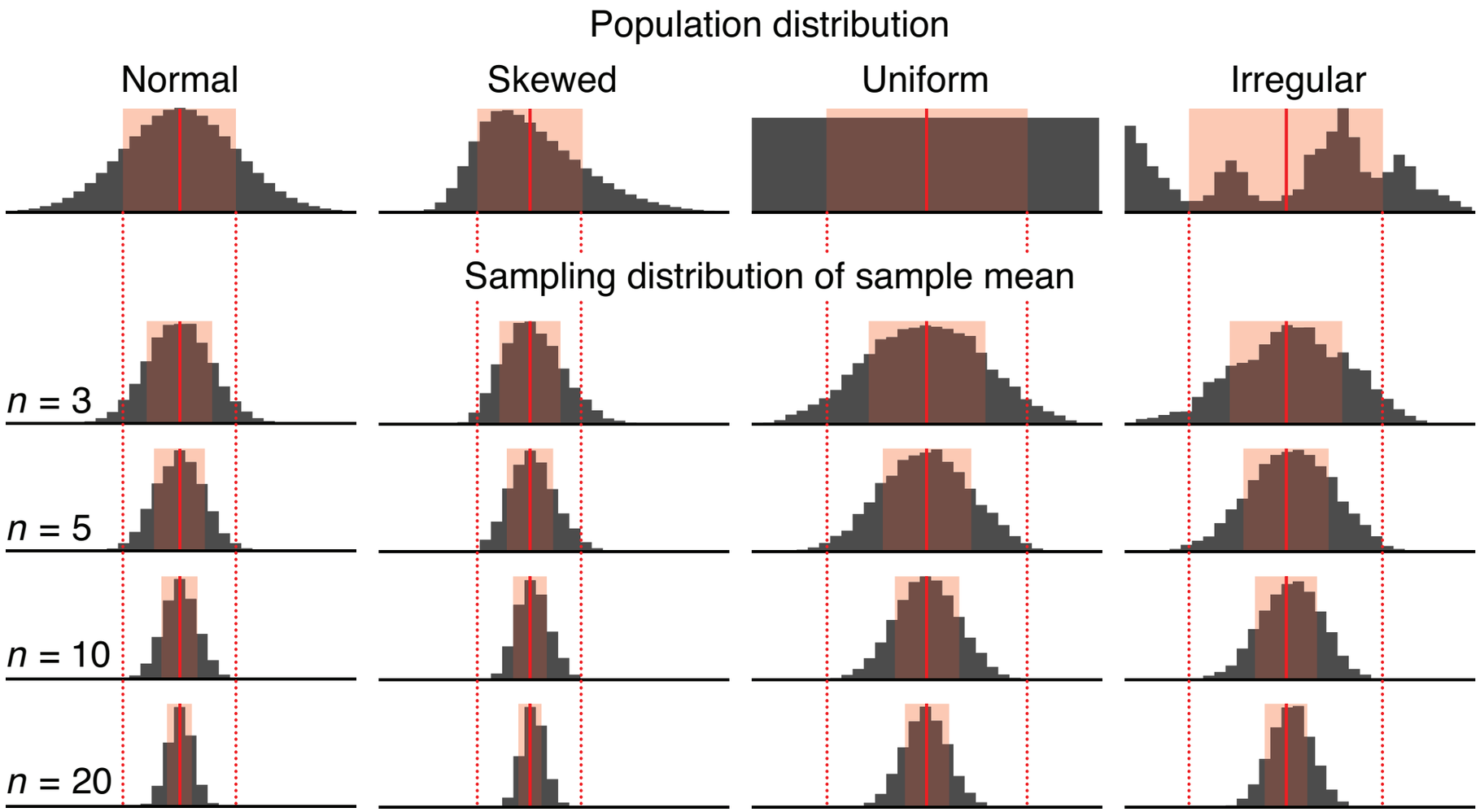
Sample preparation

Global LC-MS/MS




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

# RECALL THE CENTRAL LIMIT THEOREM




*Probability distribution of the data*

*Repeatedly selecting n data points and calculating means*

*Challenges with spectral counts:*

- *Spectral counts have skewed discrete distributions*
- *Summaries (e.g., the mean) do not have Normal distributions when sample size is small*

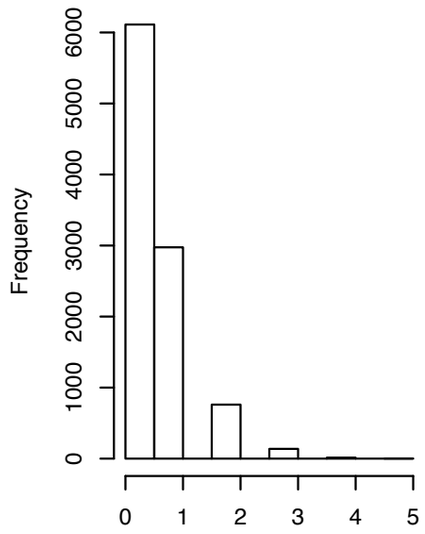
Krzywinski and Altman, Points of Significance Collection, *Nature Methods*

# STATISTICAL MODELS FOR COUNT DATA

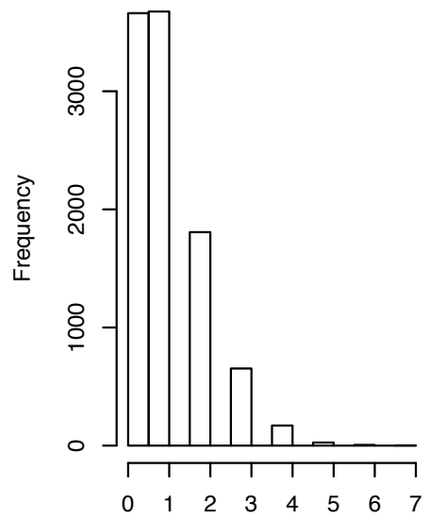


## One-parameter Poisson distribution

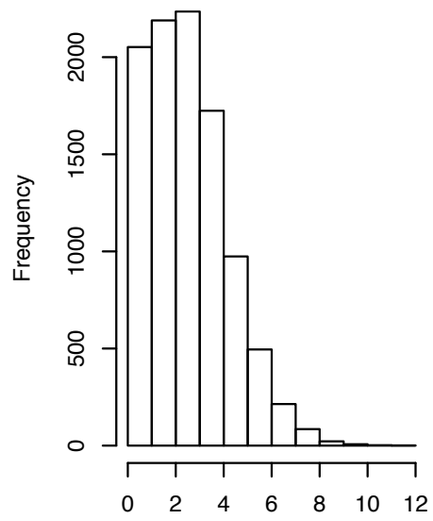
$\lambda = 0.5$



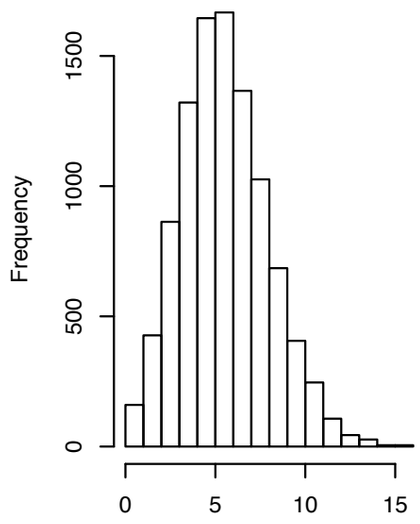

$\lambda = 1$



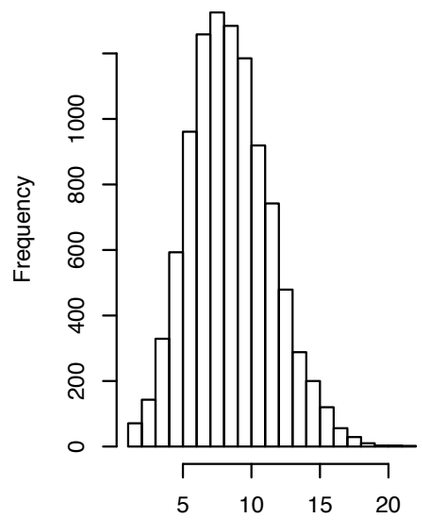

$\lambda = 3$



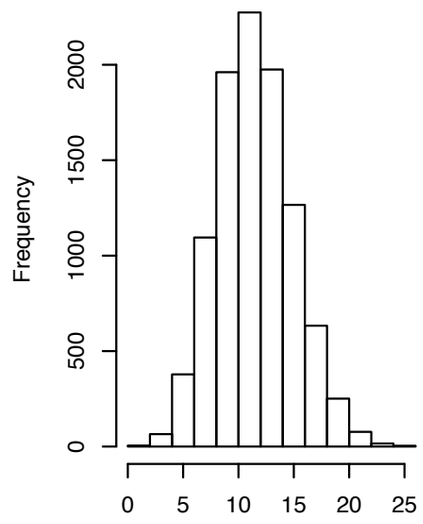

$\lambda = 6$




$\lambda = 9$




$\lambda = 12$




Poisson distribution: counts of MS/MS spectra in a condition

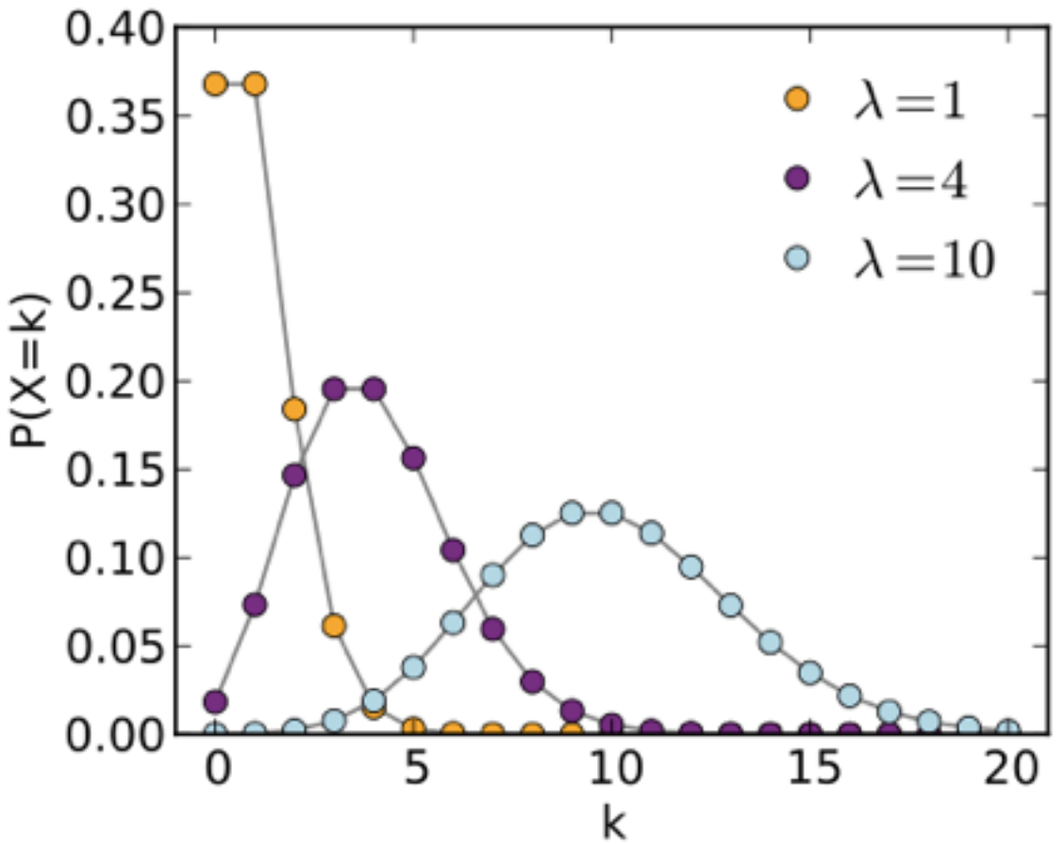
$\lambda$ is the mean spectral count in the population

Goal: compare population means between conditions

- *Single parameter: mean = variance = $\lambda$*
- *Normal distribution is appropriate for large $\lambda$*

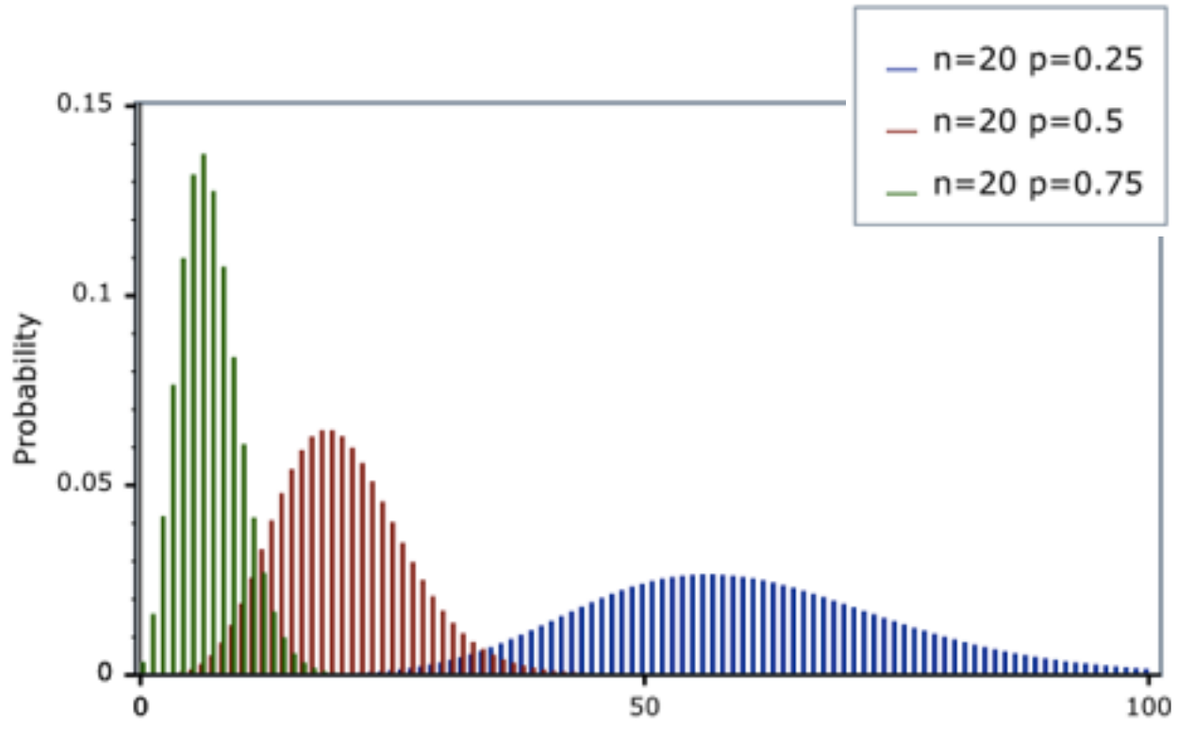
# STATISTICAL MODELS FOR COUNT DATA

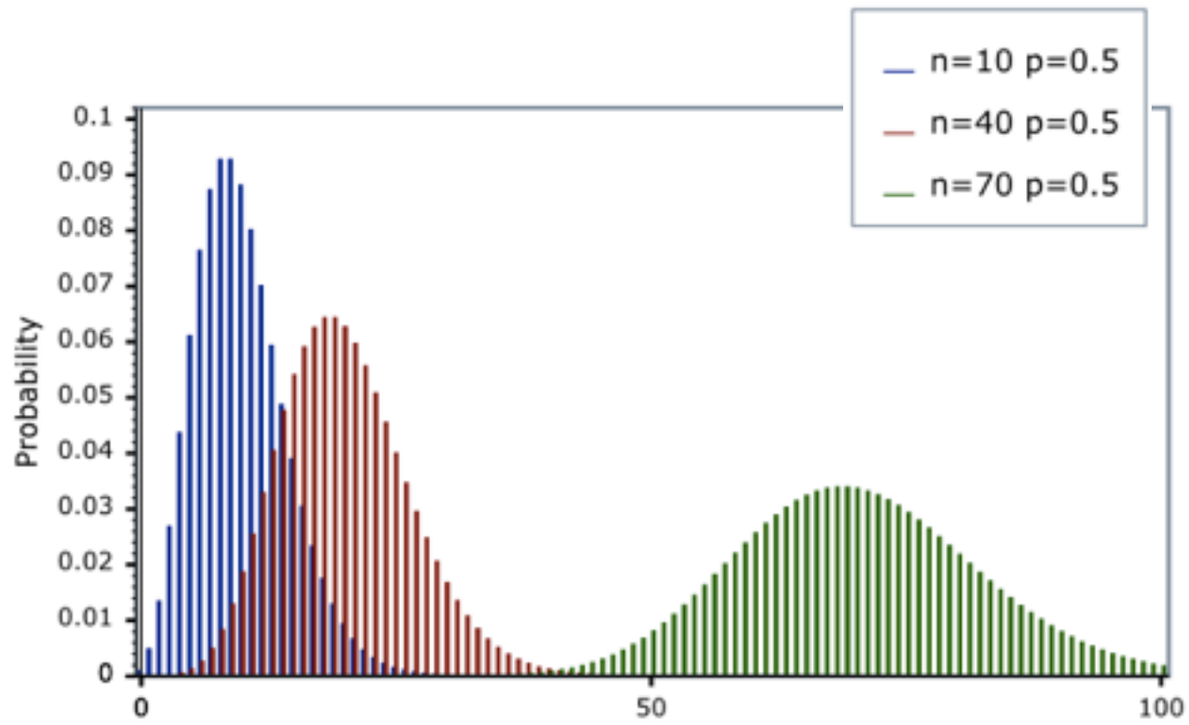
## Two-parameter Negative Binomial distribution




*Poisson distribution*

- *Two parameters help represent biological variation*




*Negative Binomial distribution*

http://www.boost.org/
https://en.wikipedia.org/wiki/Poisson_distribution

# 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 $\phi_{gi}$ 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 $\hat{\phi}_{gi}$

More convenient notation

$\phi_{gi}$

Nuisance variation

*(borrows from all genes)*

$\mu_{gi}$

Population mean of gene g in group i

*(of interest)*

$M_{ij}$

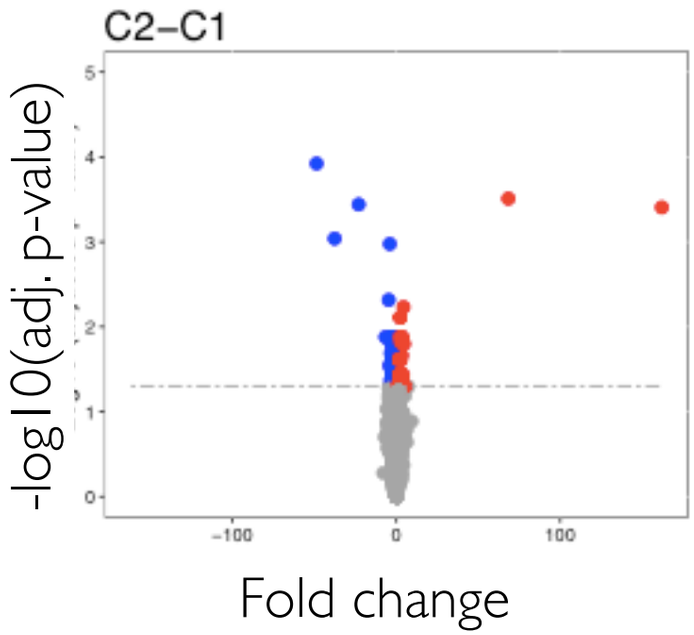
Size of the library = total # of MS/MS spectra in run

*(normalization factor)*

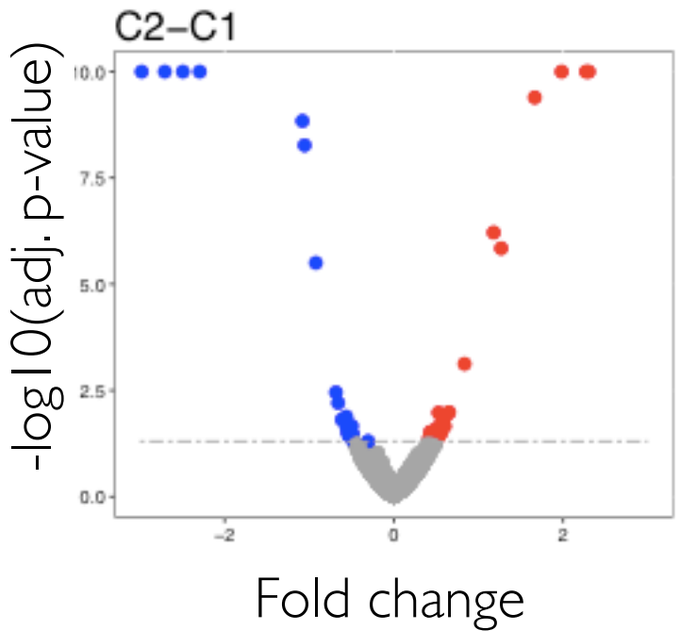
Anders and Huber,. Genome Biology, 2010

# Accurate representation of count data is the main advantage of DESeq2 over t-test

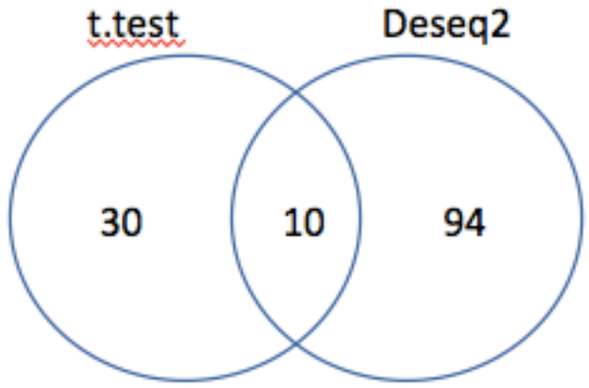
**T-test**

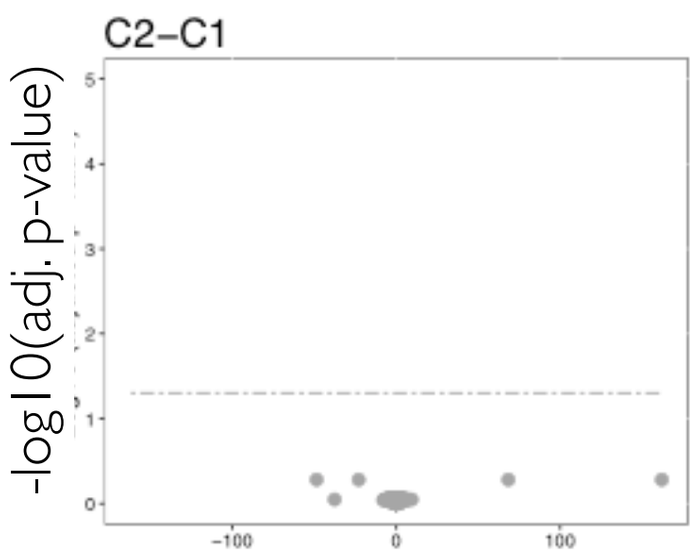


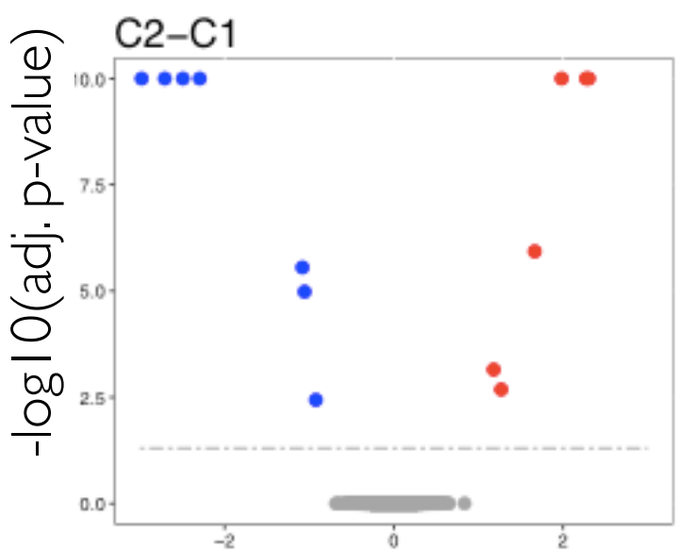

**DESeq2**



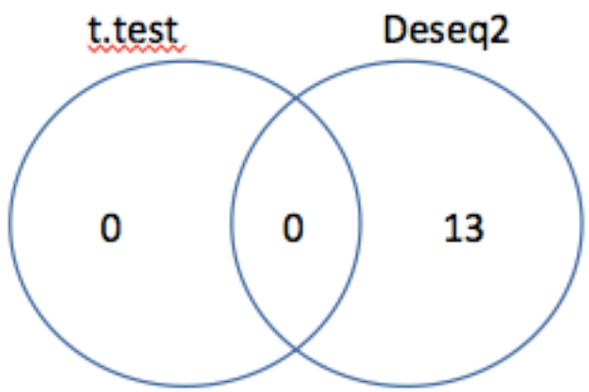

p-value cutoff




adj. p-value cutoff



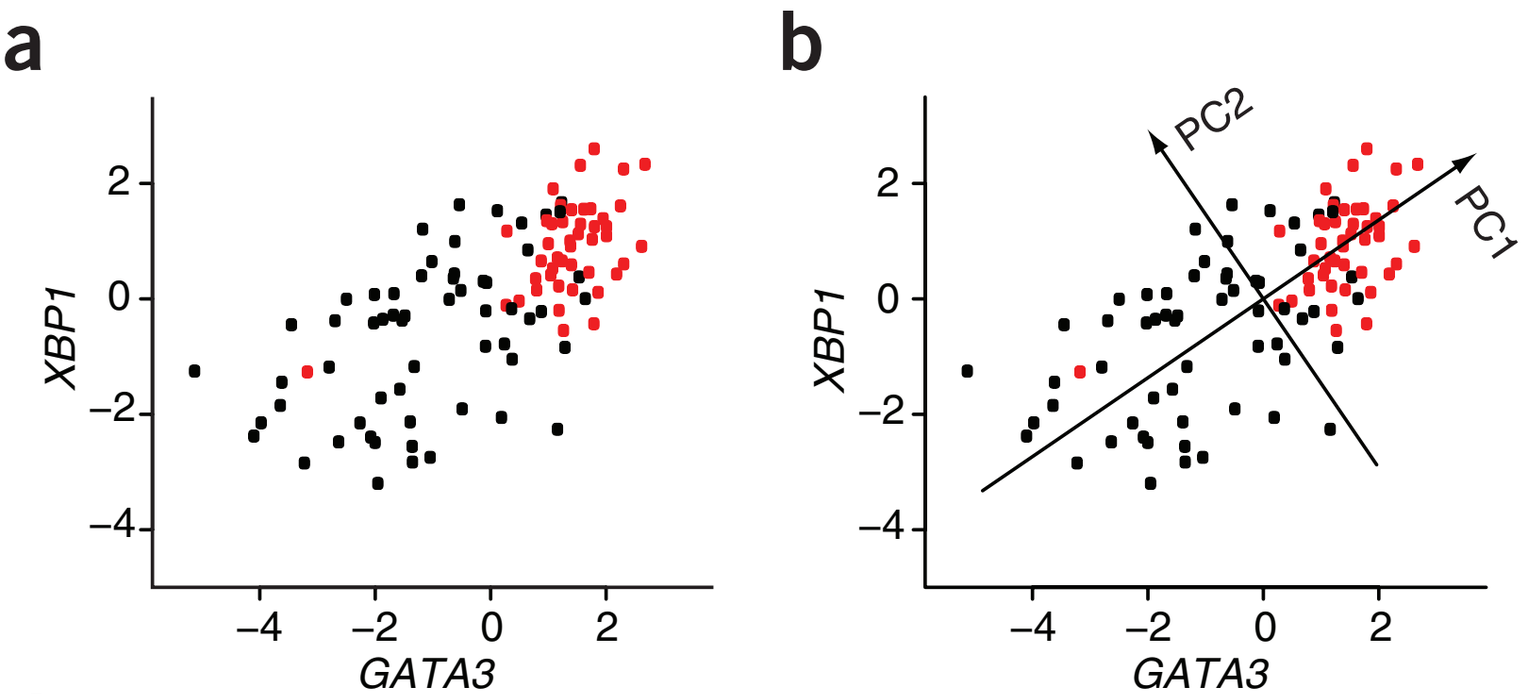

iPRG peak intensities, comparing C1 vs C2, 3 replicates/group

# CHALLENGES

- Statistical goals of the experiments
  - Class discovery, class comparison, class prediction
- Class comparison: methods from genomics
  - Limma: continuous data, small n (microarrays)
  - DEseq2: count data, small n (RNA-seq)
- Class discovery
  - Principle component analysis
  - Hierarchical clustering

# PRINCIPLE COMPONENT ANALYSIS

Ringnér, *Nature Biotechnology*, 2008




*Samples in the coordinates of 2 proteins*

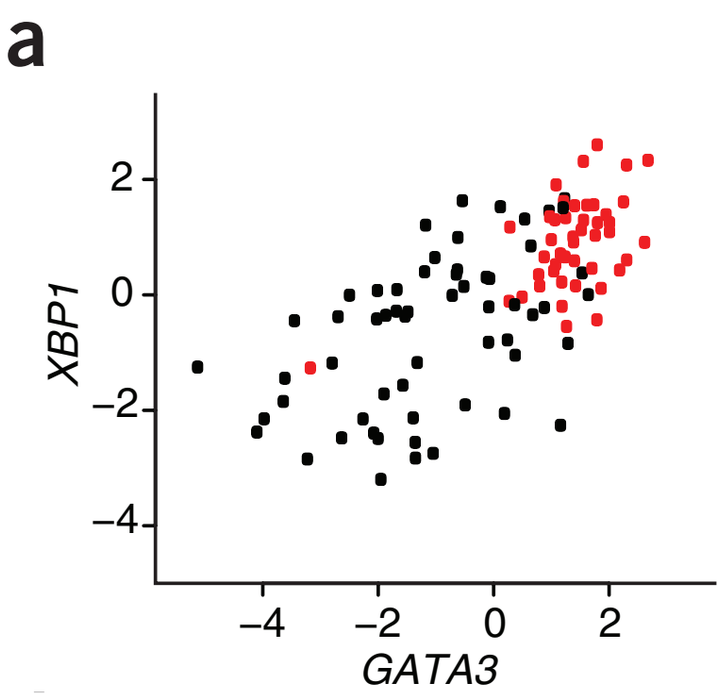
*Samples in the rotated coordinates of 2 proteins*

| | | Genes $g_1$ | ... | $g_G$ |
|---|---|---|---|---|
| Samples | $S_1$ | $x_{11} - \bar{x}_{\cdot 1}$ | ... | $x_{1G} - \bar{x}_{\cdot G}$ |
| | ... | | ... | |
| | $S_I$ | $x_{I1} - \bar{x}_{\cdot 1}$ | ... | $x_{IG} - \bar{x}_{\cdot G}$ |
| | $S_\cdot$ | $\bar{x}_{\cdot 1}$ | ... | $\bar{x}_{\cdot G}$ |

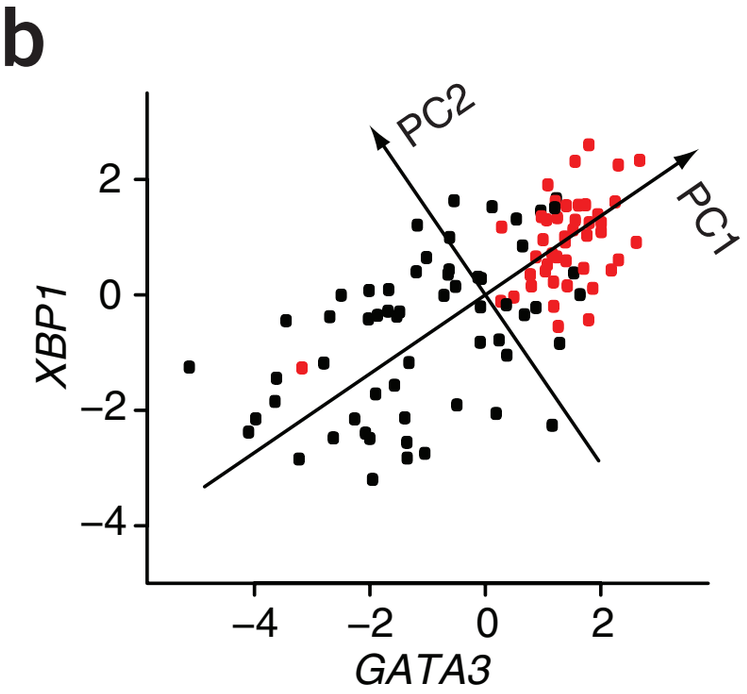
*Shifting the coordinates to the center of the cloud*

# PRINCIPLE COMPONENT ANALYSIS

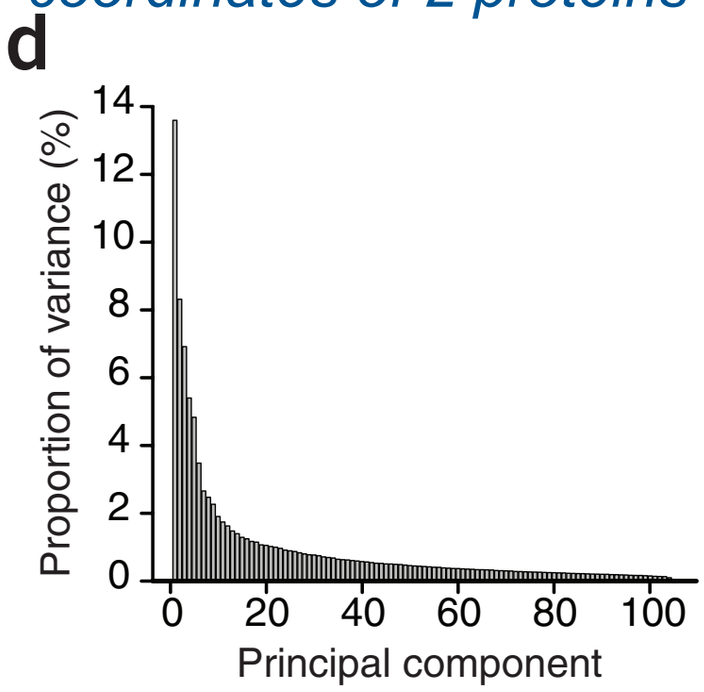
Ringnér, *Nature Biotechnology*, 2008



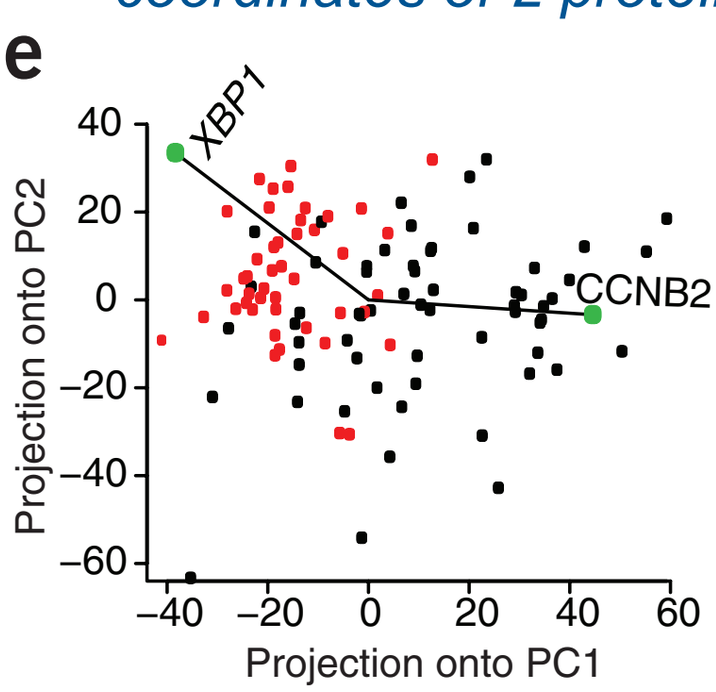

*Samples in the coordinates of 2 proteins*




*Samples in the rotated coordinates of 2 proteins*




*% of total variance explained by principle components*



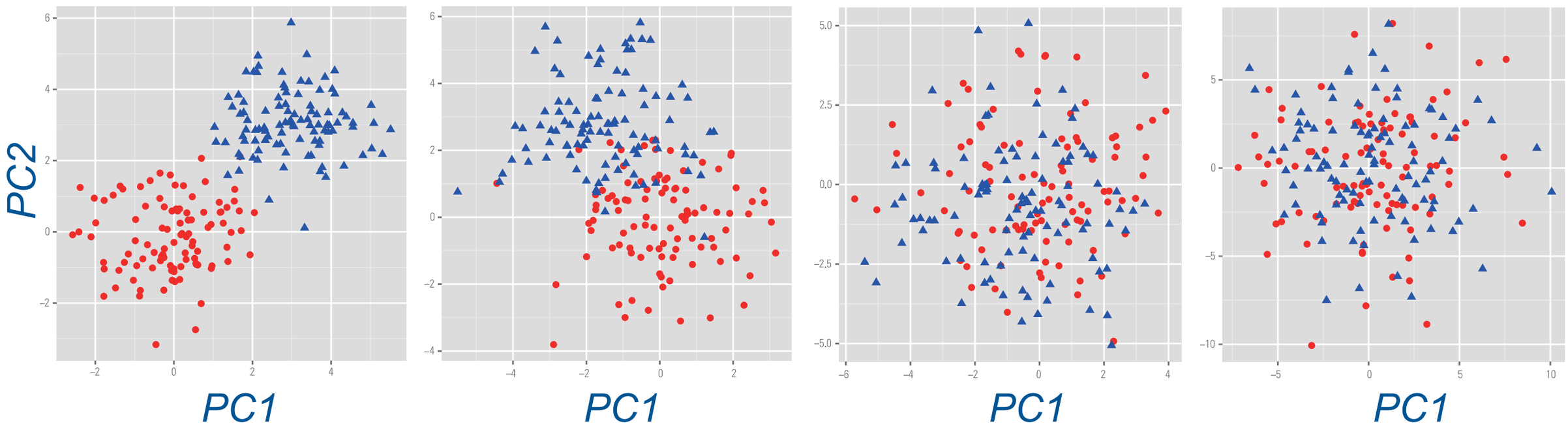

*Samples projected on principle components*

# CHALLENGE IN HIGH-DIMENSIONAL STUDIES

## As the number of (noisy) proteins increases, signal may be lost

A simulation study

- Simulate $n = 100$ observations from 2 classes
- Each observation is a point in $m = 2, 40, 200, 1000$ dimensions
- Only first 10 dimensions are informative
- Plot first 2 principle components (i.e., eigenvectors)
- Informative data should show a good separation between the two classes

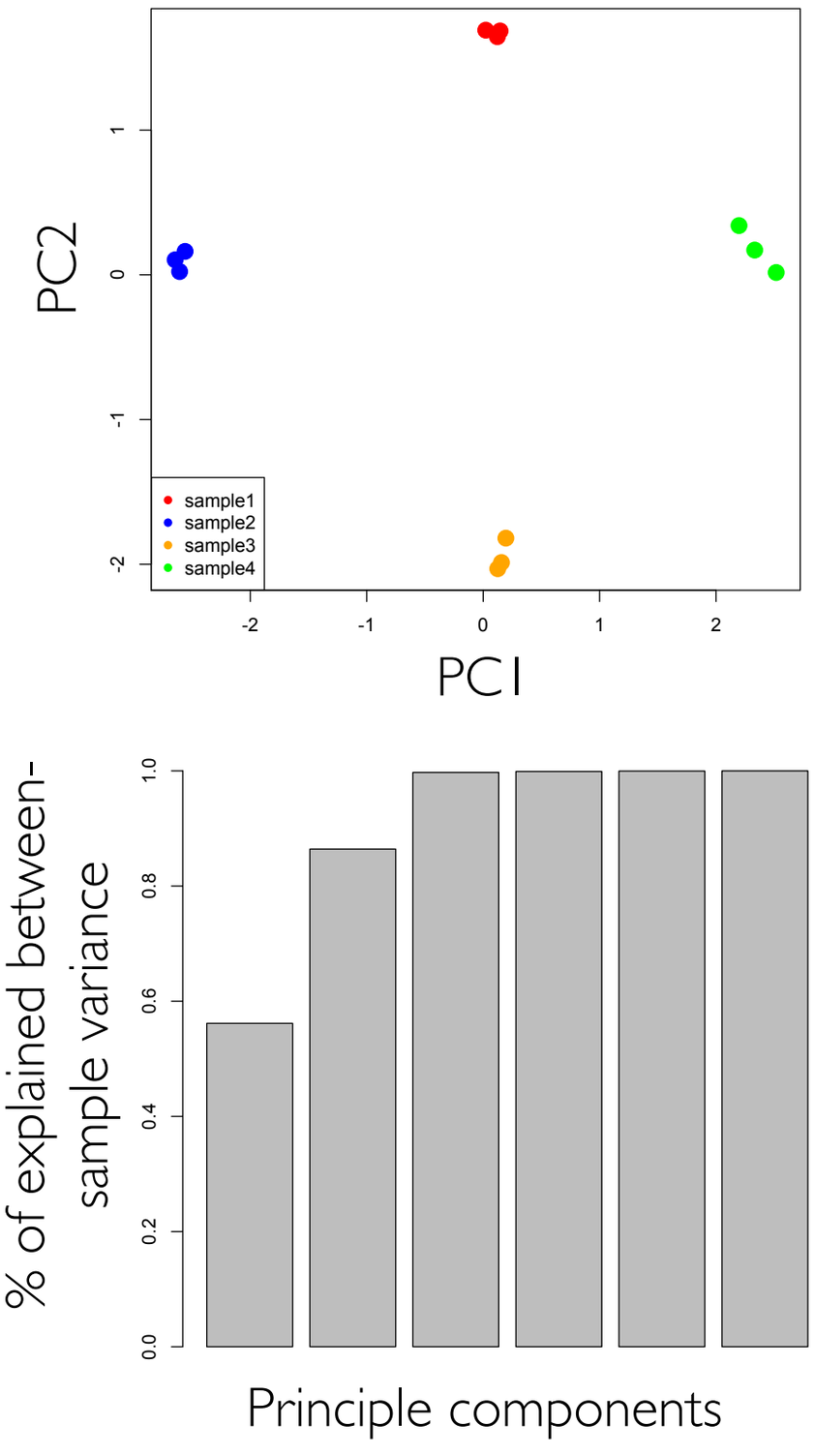



Fan, Han, Liu, *Challenges of Big Data Analysis*, National Science Review, 2014

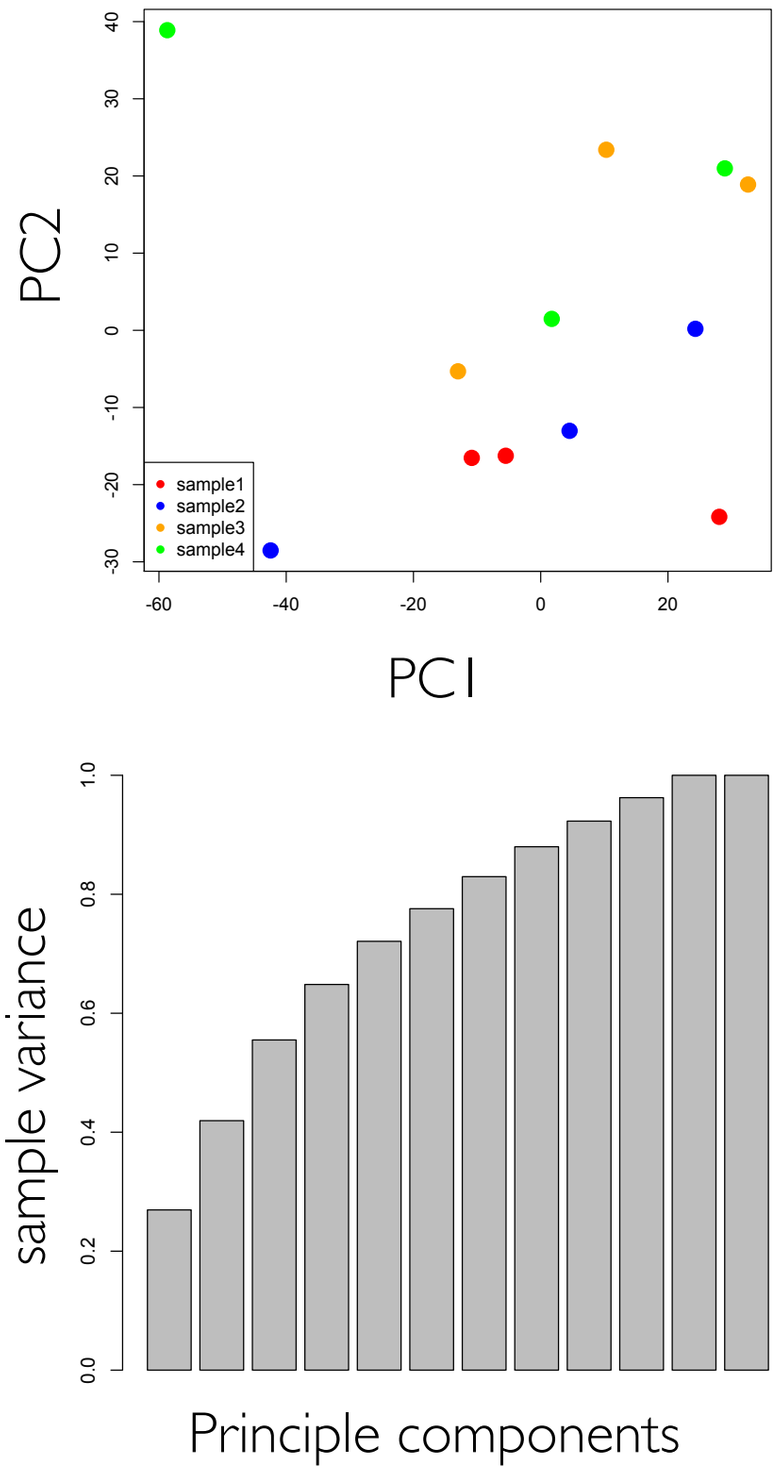
# CHALLENGE IN HIGH-DIMENSIONAL STUDIES

## iPRG peak intensities: signal from spiked proteins lost in high dimensions

### Spiked proteins only



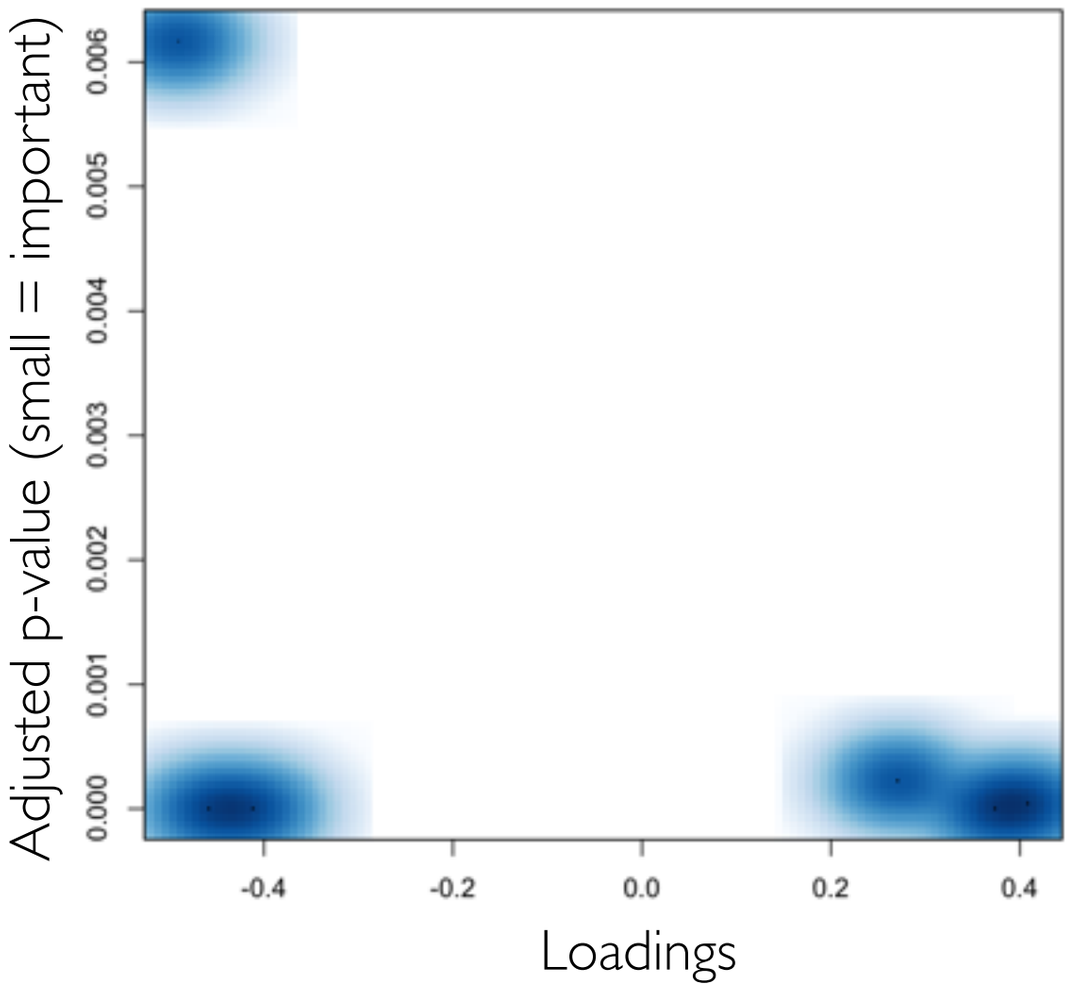

### All proteins



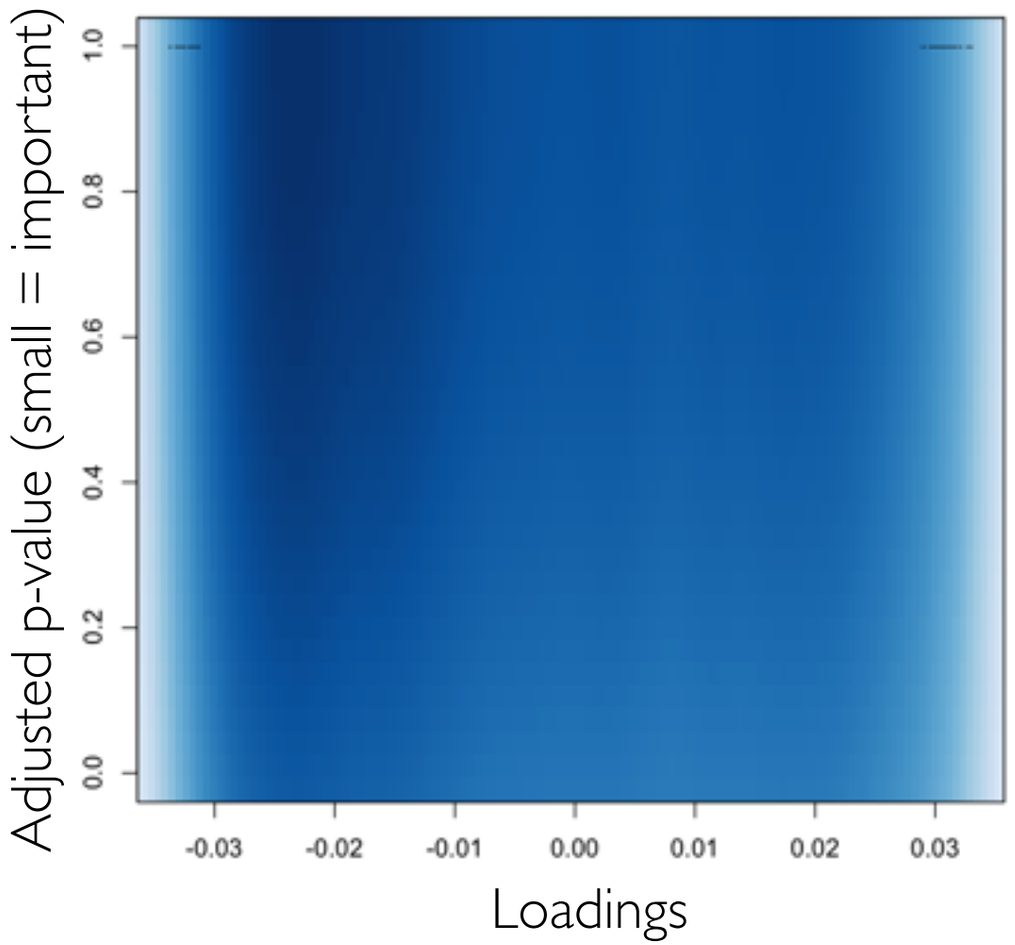
# CHALLENGE IN HIGH-DIMENSIONAL STUDIES

## Proteins with high loadings should not be interpreted as biologically important

**Spiked proteins only**




**All proteins**

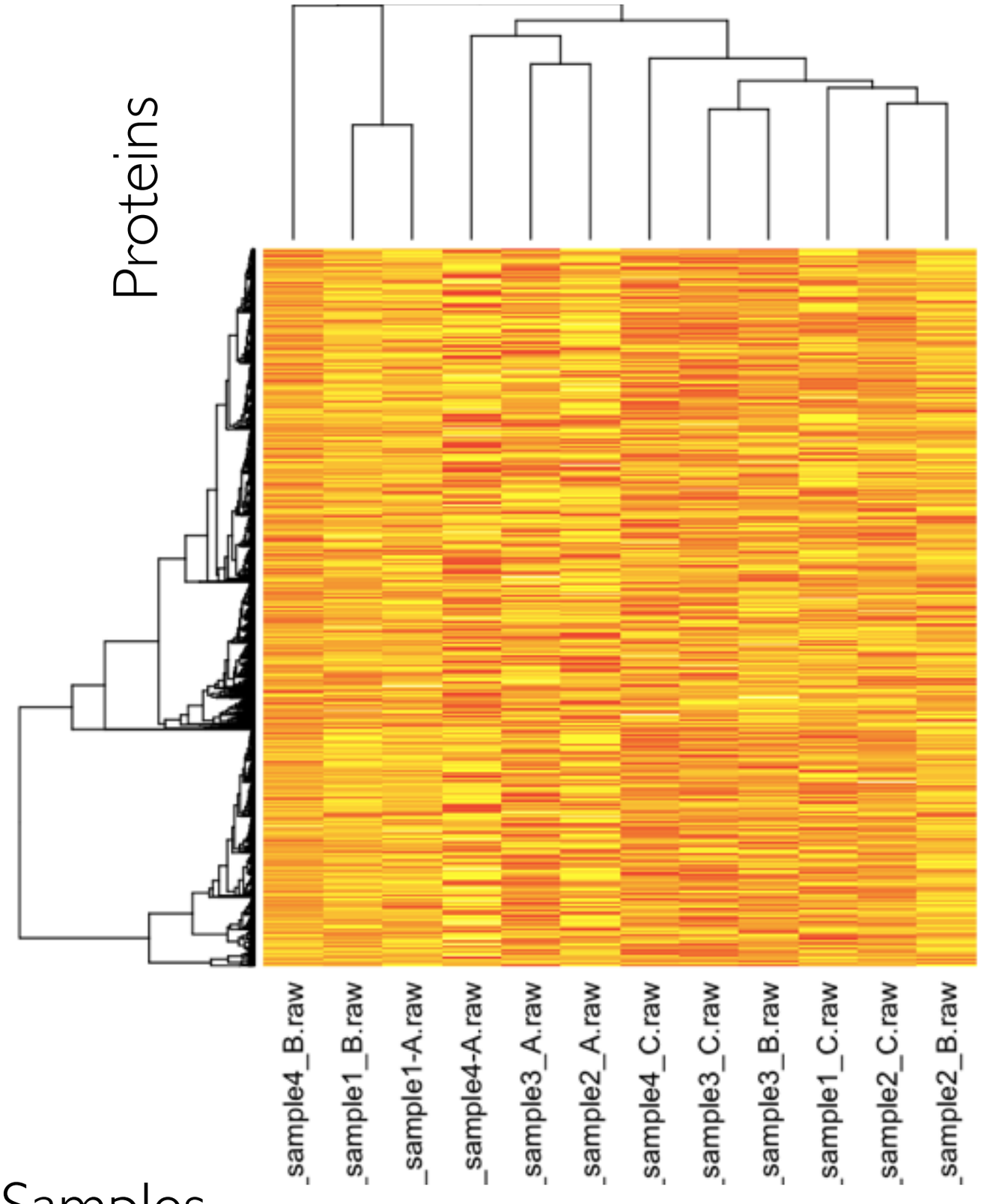


# CHALLENGES

- Statistical goals of the experiments
  - Class discovery, class comparison, class prediction
- Class comparison: methods from genomics
  - Limma: continuous data, small n (microarrays)
  - DEseq2: count data, small n (RNA-seq)
- Class discovery
  - Principle component analysis
  - Hierarchical clustering

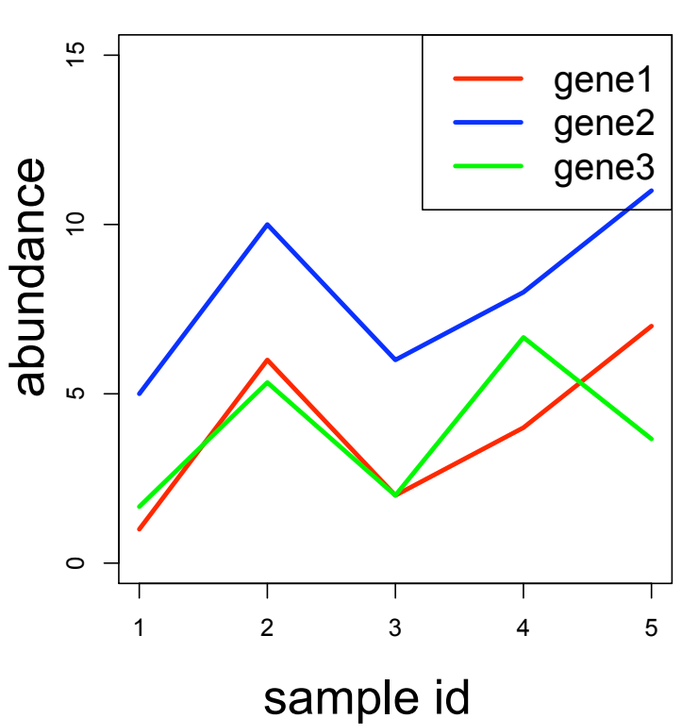
# HEATMAPS DISPLAY QUANTITATIVE HIGH-DIMENSIONAL DATASETS

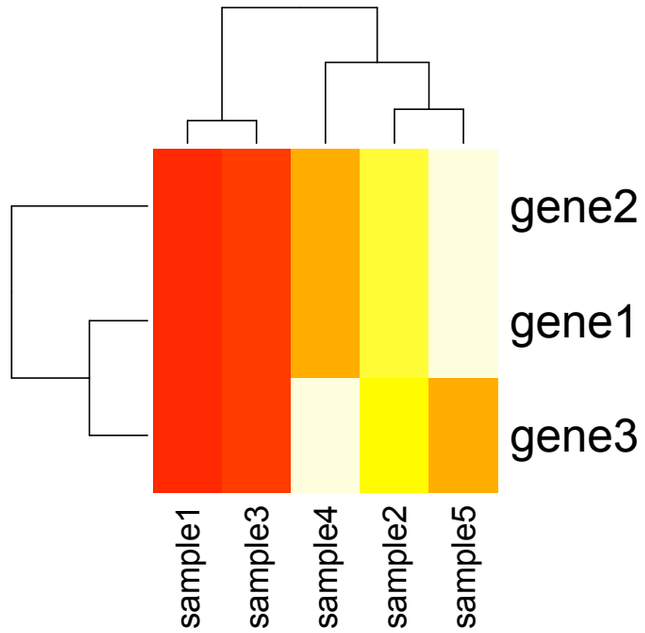
**iPRG dataset, all proteins**

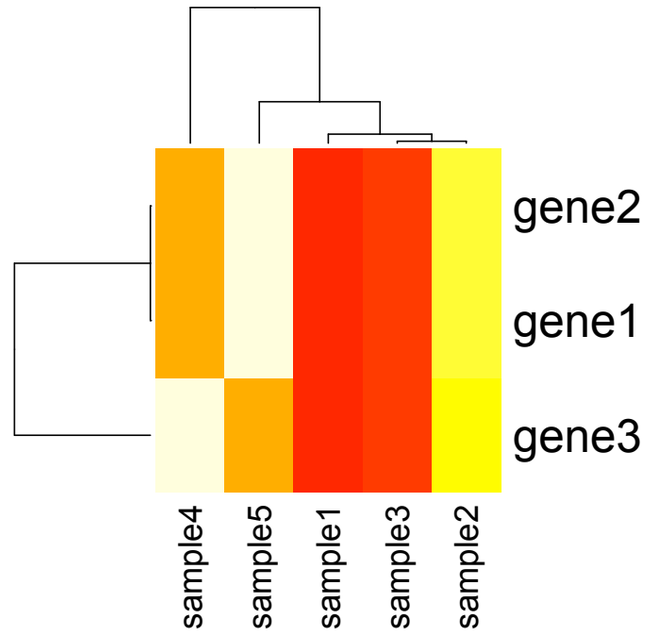


# HIERARCHICAL CLUSTERING

## Clusters depend on the definition of similarity between the profiles



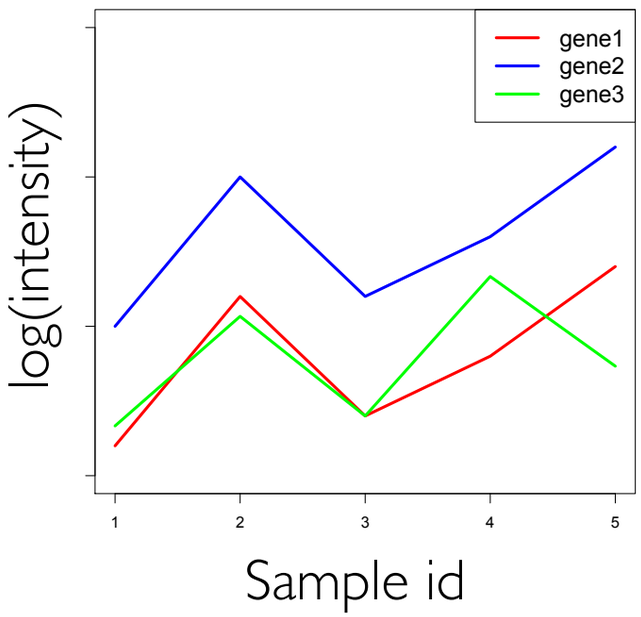




$$Euclidian(\vec{e}_i,\vec{e}_j)=\left\|\vec{e}_i-\vec{e}_j\right\|_2=\sqrt{\sum_{k=1}^{n}(e_{ik}-e_{jk})^2}$$

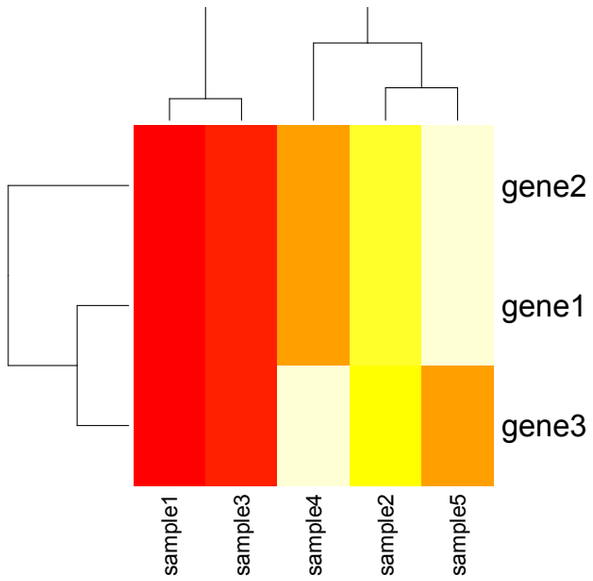
$$Pearson(\vec{e}_i,\vec{e}_j)=\frac{Cov[\vec{e}_i,\vec{e}_j]}{\sqrt{Var[\vec{e}_i]Var[\vec{e}_j]}}=\frac{\sum_{k=1}^{n}(e_{ik}-\mu_i)(e_{jk}-\mu_j)}{\sigma_i\sigma_j}$$

# CENTERING & SCALING AFFECTS SIMILARITIES

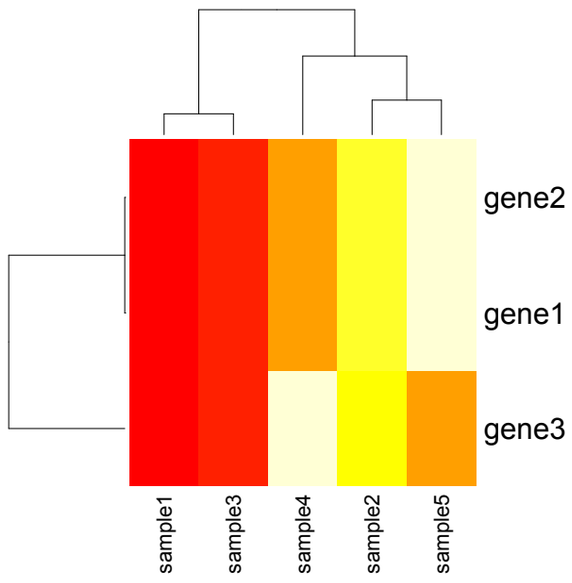
Original



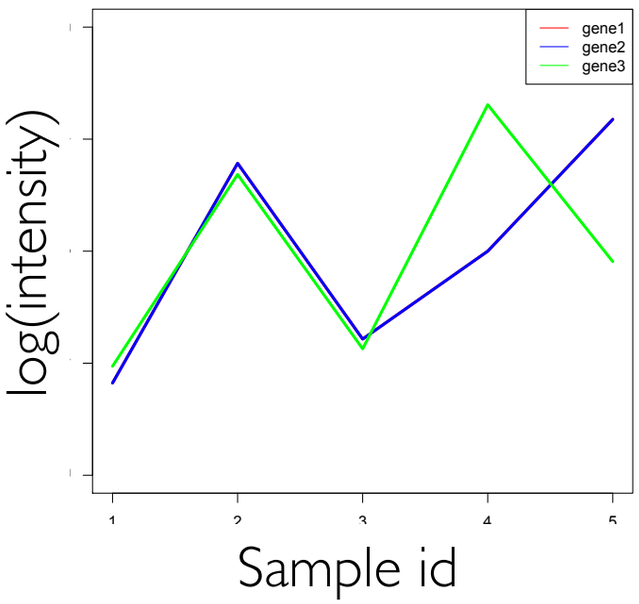

Euclidian distance, original



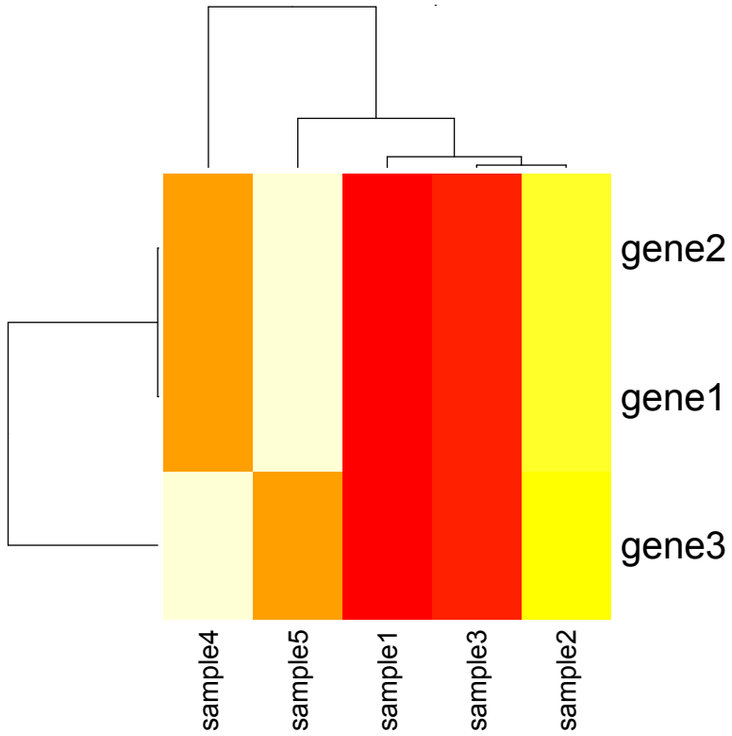

Euclidian distance, centered & scaled



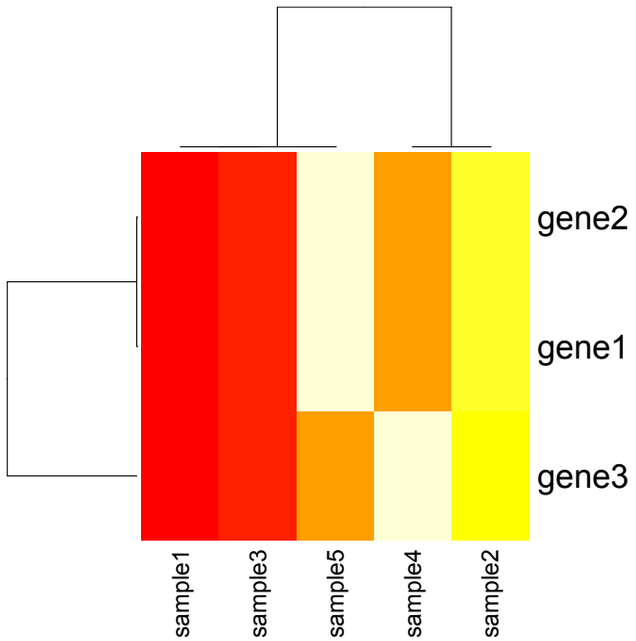

Centered and scaled




Correlation distance, original

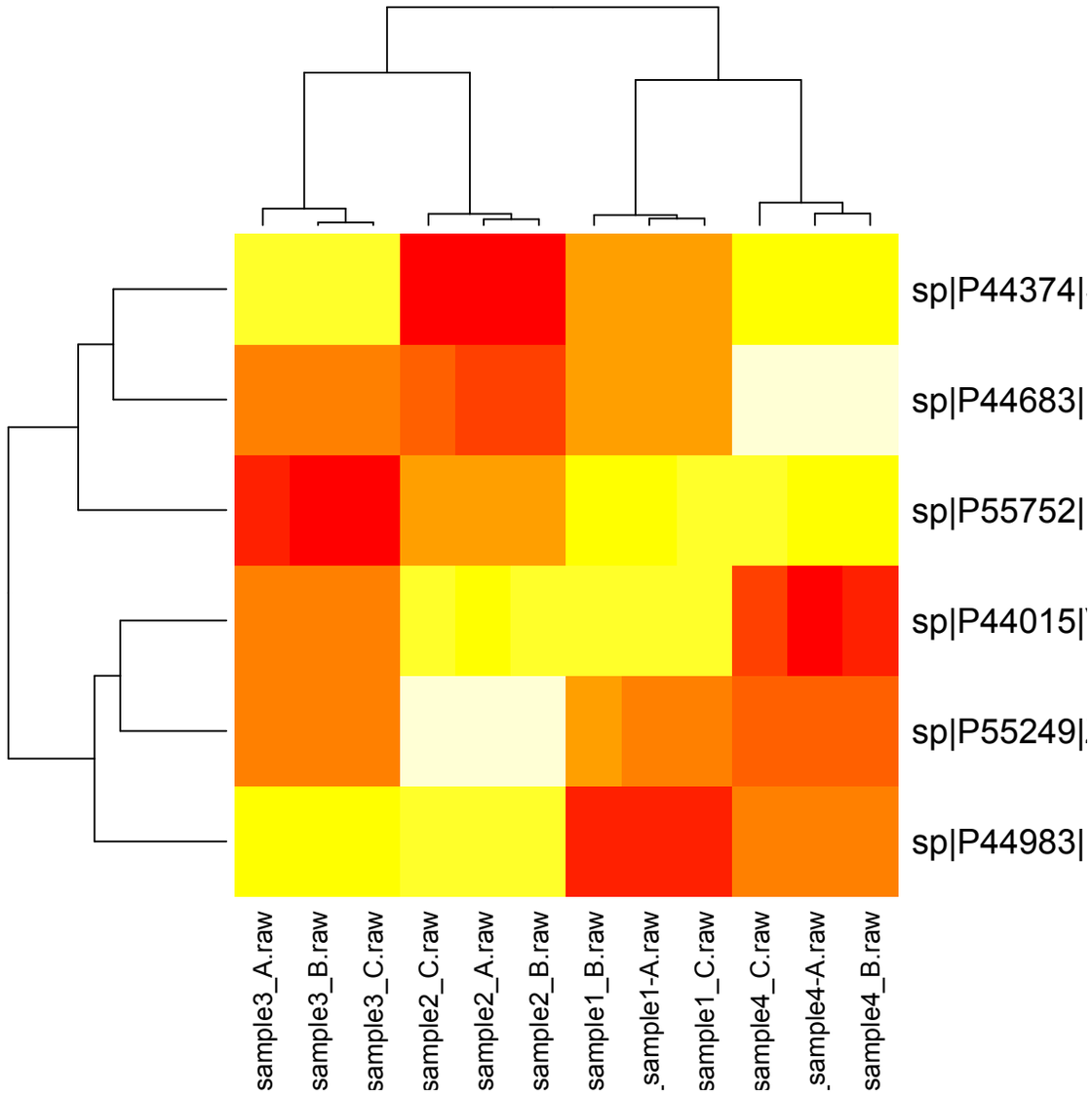



Correlation distance, centered and scaled

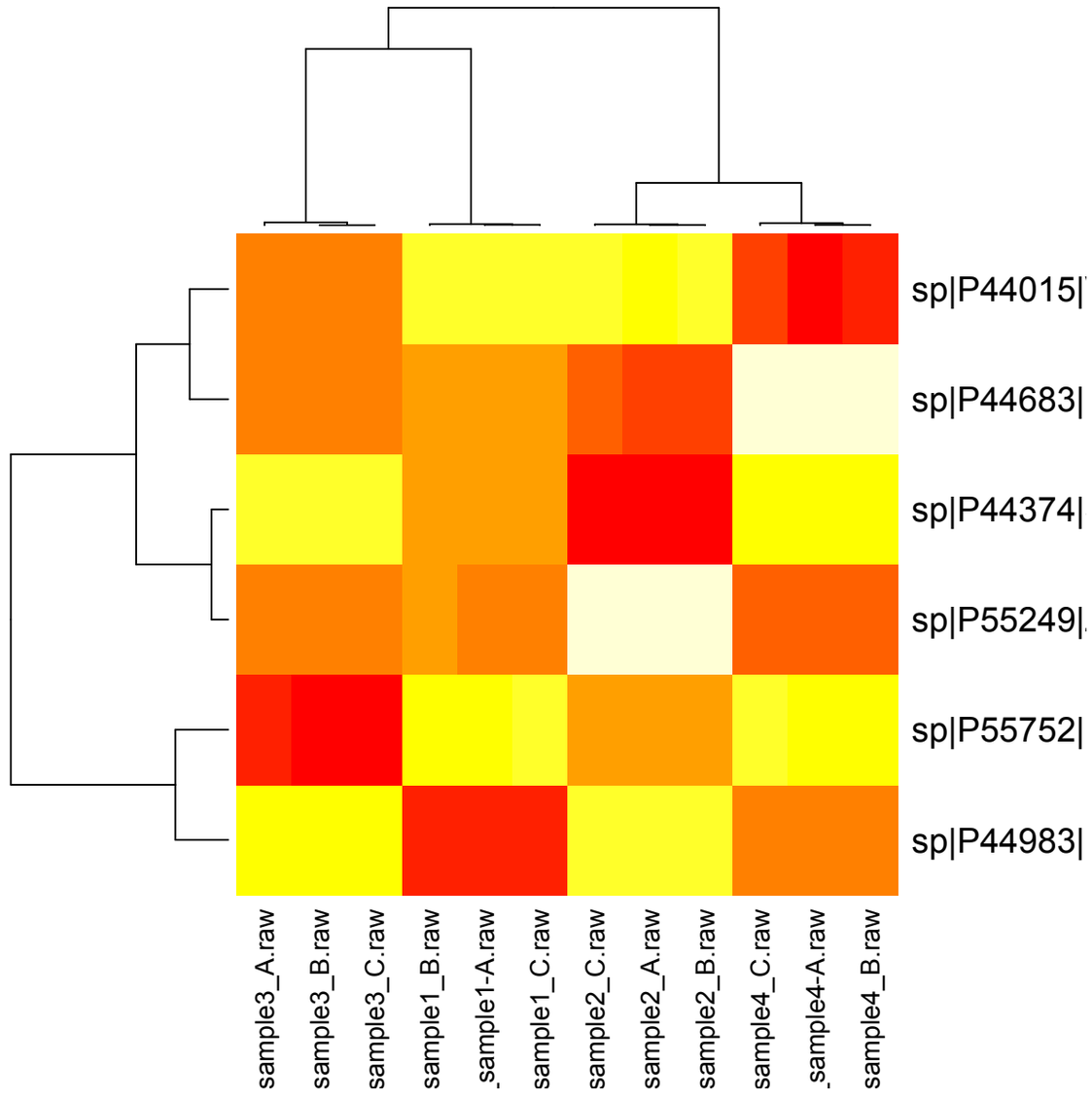


# IPRG PEAK INTENSITIES, SPIKED PROTEINS

## Euclidian distance, original



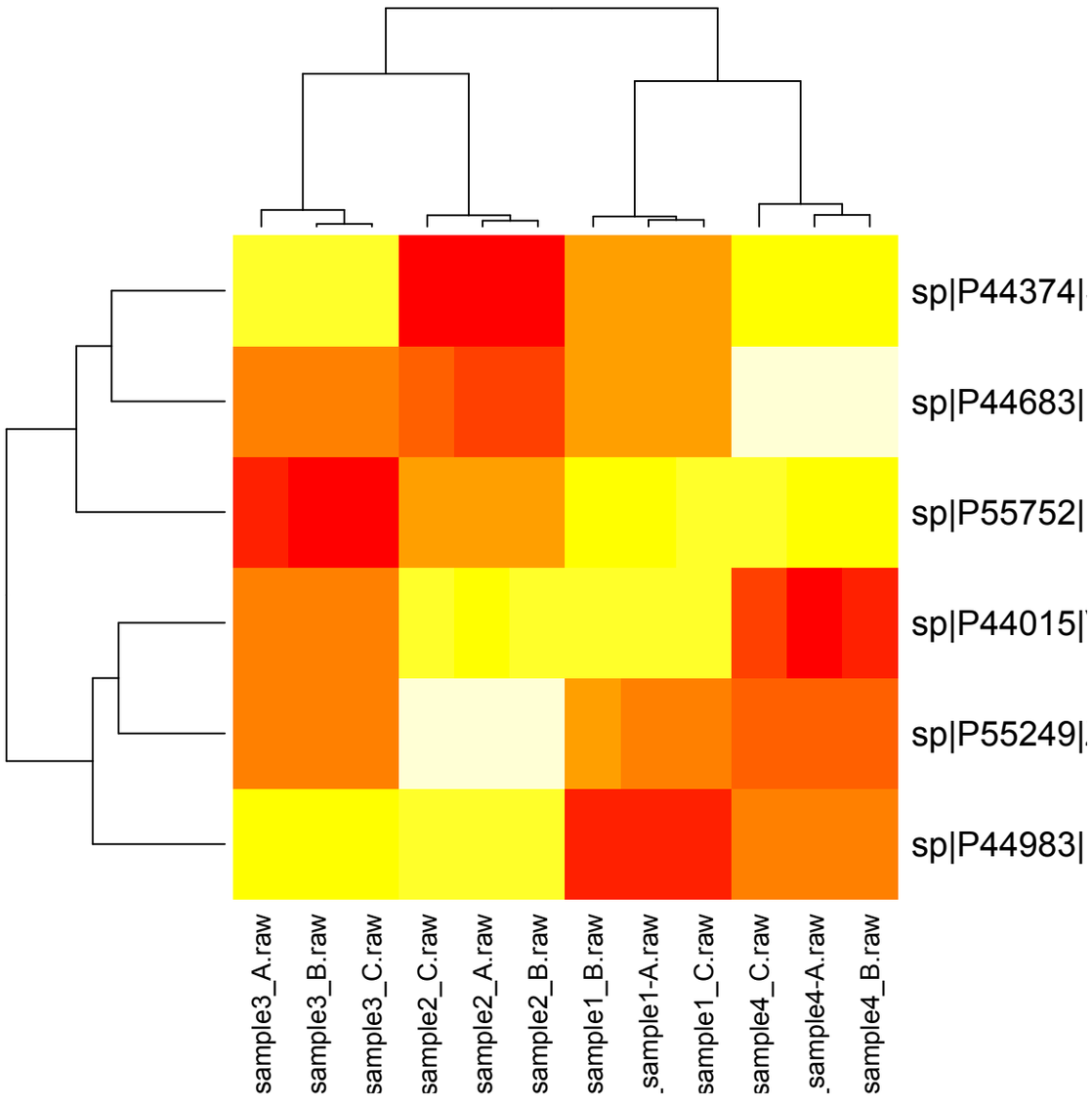

## Correlation distance, original



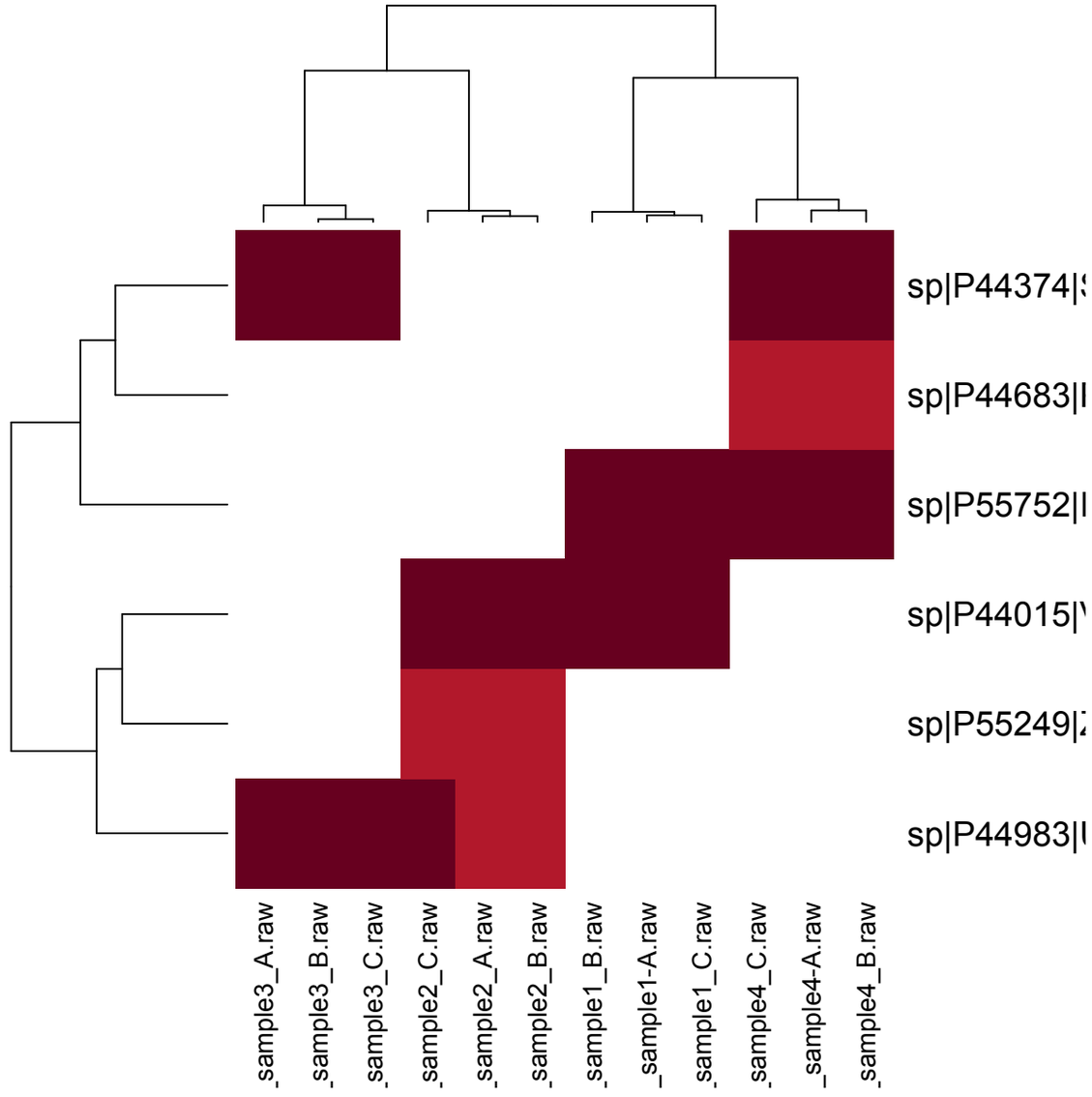

*more similarity between proteins and samples*

# IPRG PEAK INTENSITIES, SPIKED PROTEINS

## Euclidian distance, original

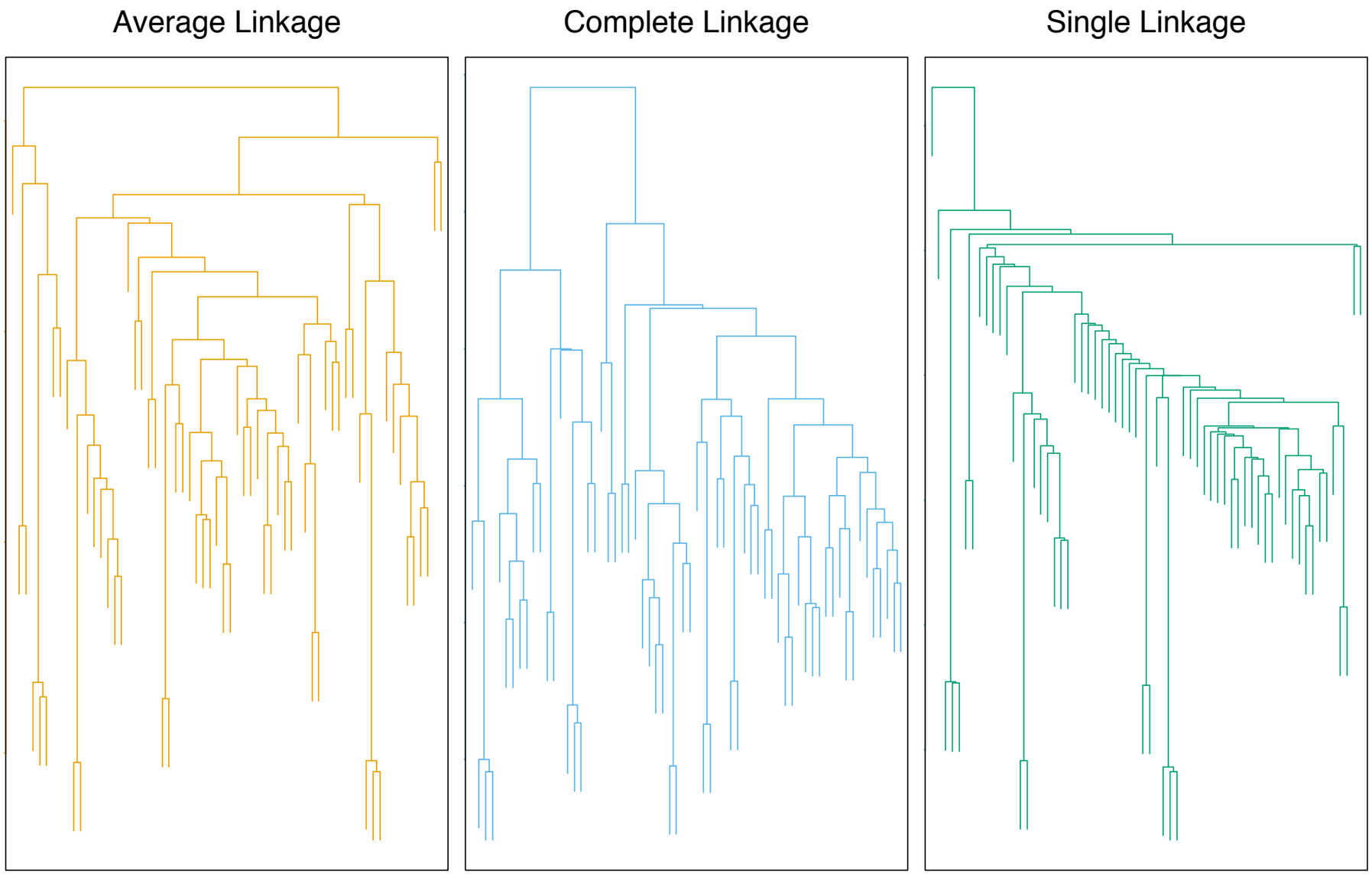



## Euclidian distance, equally spaced color




*high abundances dominate*

# LINKAGE DETERMINES ORDER OF ROWS AND COLUMNS




**Conclusion:** Average linkage leads to most 'balanced' dendrograms

Hastie, Tibshirani, Friedman, *The elements of Statistical Learning* 2008

# CHALLENGE IN HIGH-DIMENSIONAL STUDIES

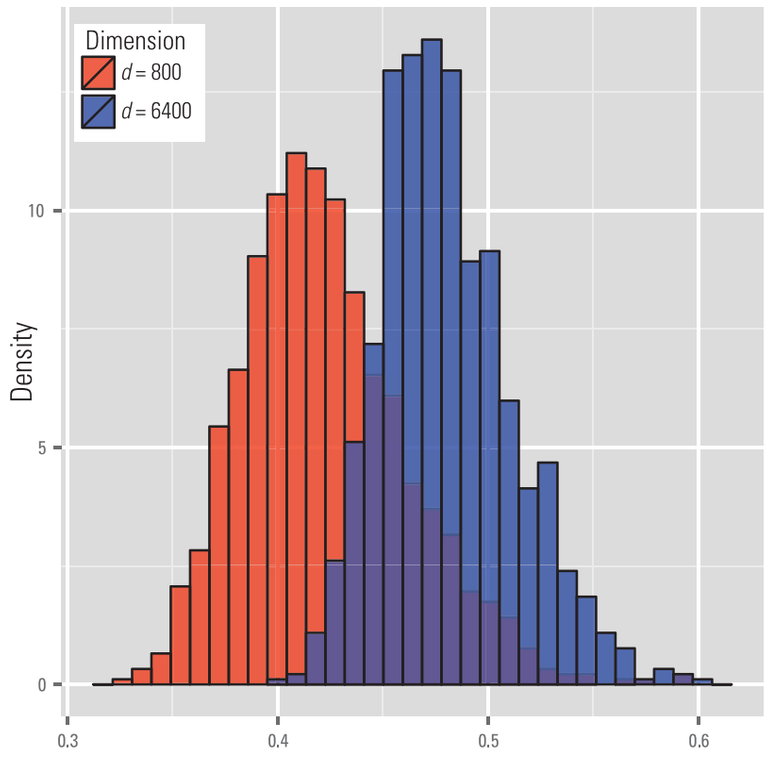
## Spurious correlations may arise

Fan, Han, Liu, *Challenges of Big Data Analysis*, National Science Review, 2014

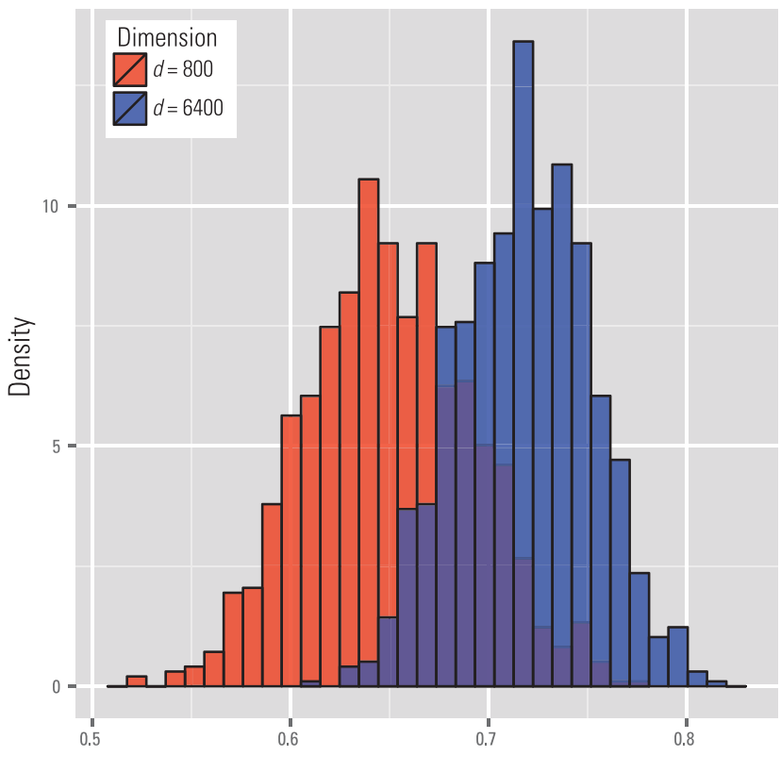
A simulation study

- Simulate $n = 60$ independent observations
- Each observation is in $d = 800, 6400$ dimensions
- Left: max absolute correlation between the first dimension and any other dimensions
- Right: max absolute correlation between the first dimension and a linear combination of any 4 other

*Maximum absolute correlation*




*2 proteins*




*4 proteins*