CATEGORICAL DATA AND SAMPLE SIZE

Meena Choi, Ting Huang, Olga Vitek

College of Science

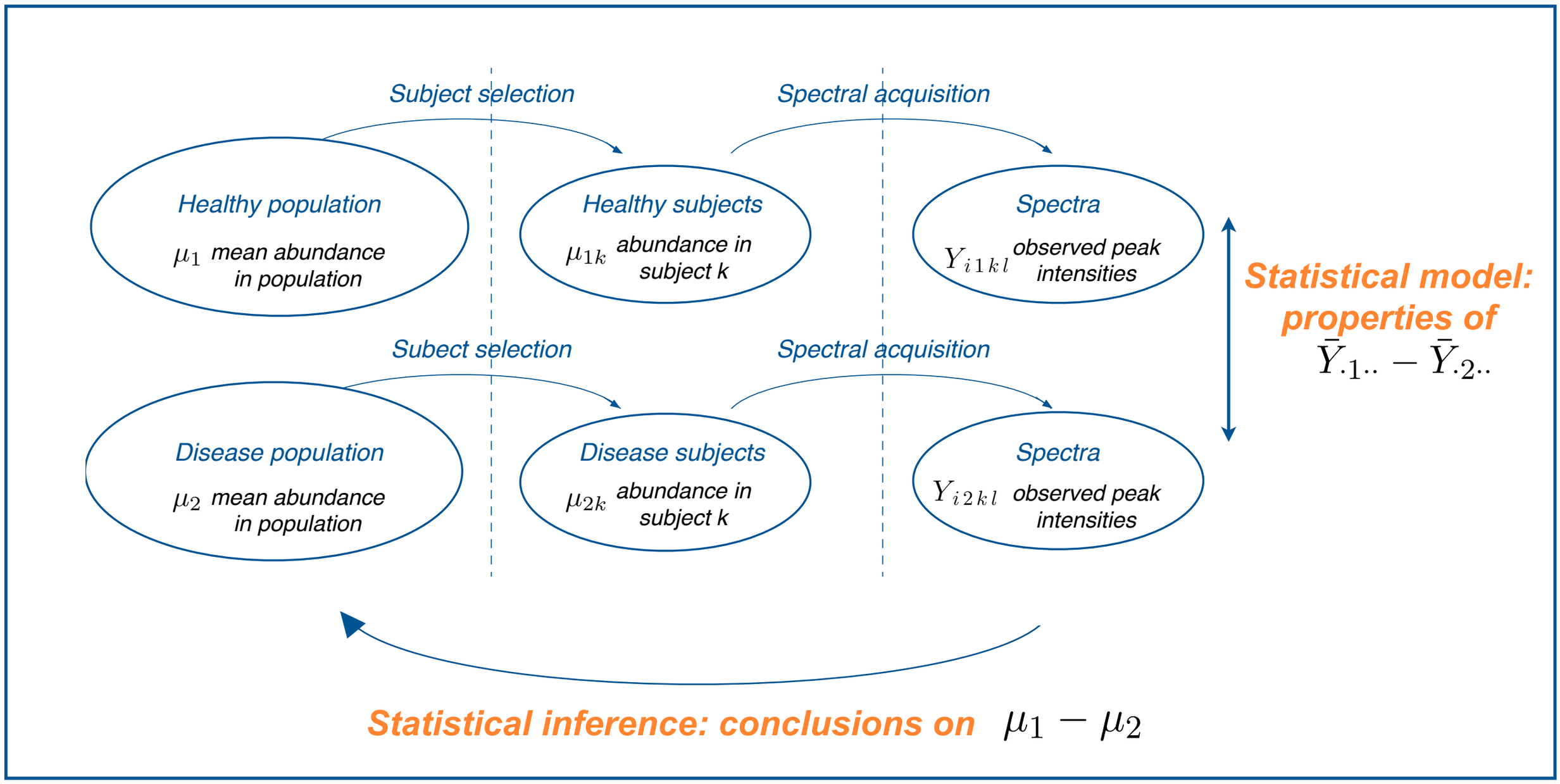
College of Computer and Information Science

Northeastern University

OUTLINE

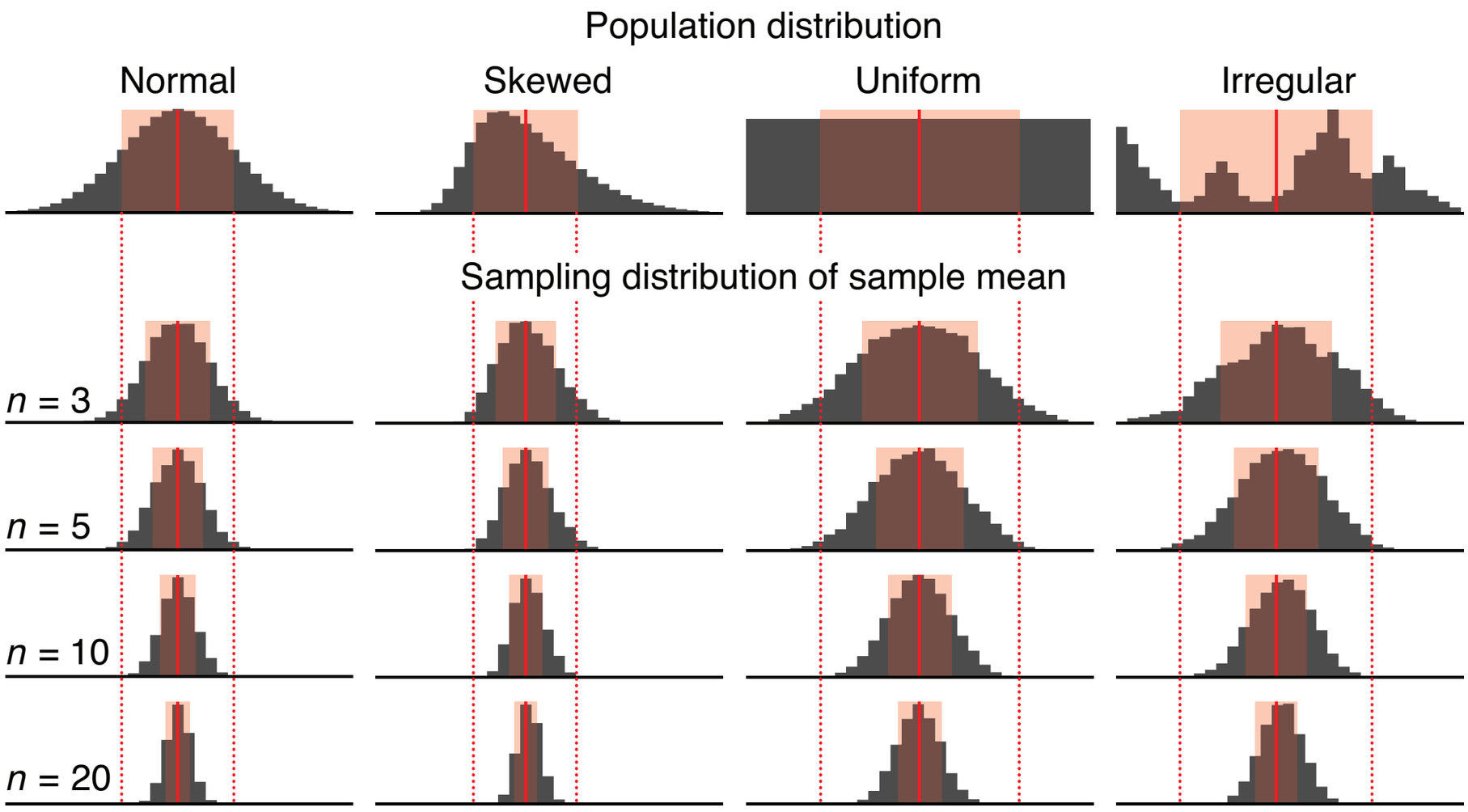
- Statistical analysis of count data
  - Comparison of proportions
  - 2-way tables: chi2 test, Fisher's exact test
- So how many replicates do I need?
  - Design of complex experiments

RECALL THE GOAL OF GROUP COMPARISON



What if measurements are binary?
(E.g., up or down?)

RECALL THE CENTRAL LIMIT THEOREM



Probability distribution of the data

Repeatedly selecting n data points and calculating means

Averages of binary variables will approach Normal distribution as sample size increases

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

ANALYZING PROPORTIONS

Hypothesis testing and confidence interval

Population proportions

H0: 'status quo', no change in abundance, $\pi_1 - \pi_2 = 0$

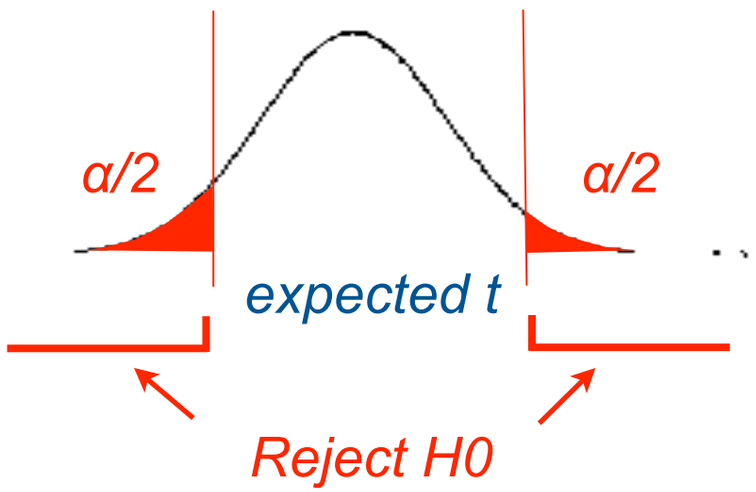
Ha: change in abundance, $\pi_1 - \pi_2 \neq 0$

Proportions in each group

$$\textbf{observed t} = \frac{\textbf{difference of group means}}{\textbf{estimate of variation}} = \frac{p_1 - p2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

Distribution of the score if H0 is true ~ Normal (0,1)

α = False Positive Rate



Confidence interval for difference of proportions

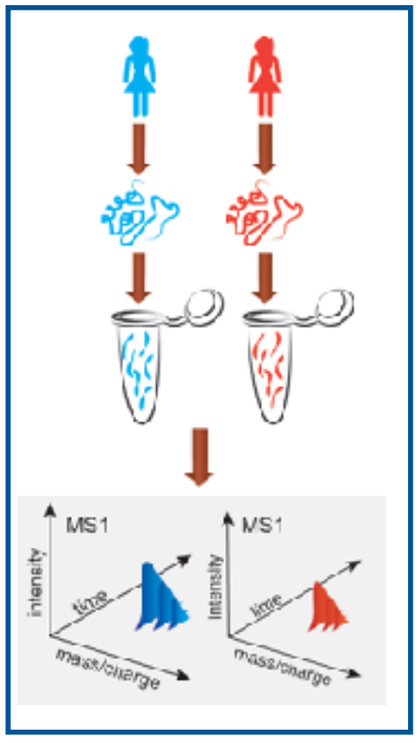
$$p_1 - p_2 \pm z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

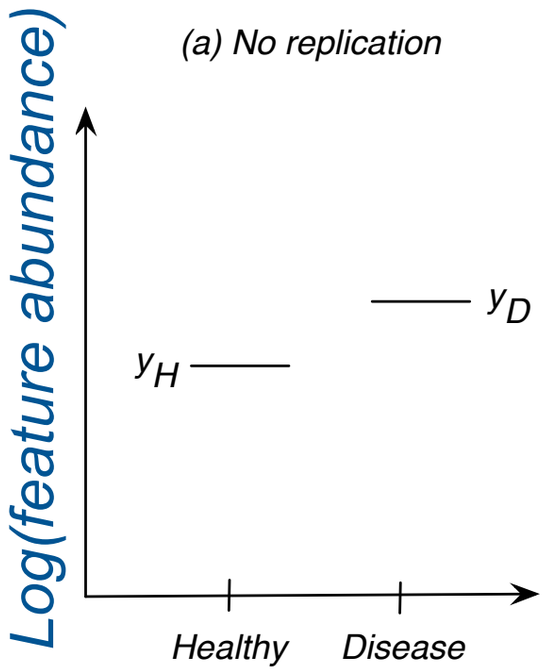
OUTLINE

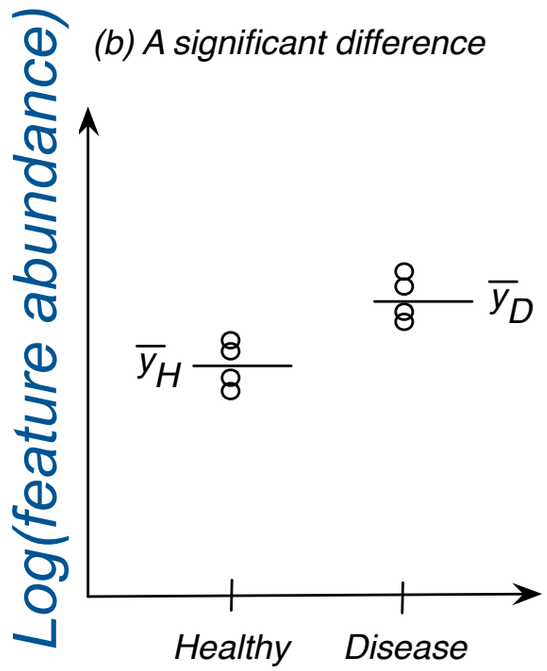
- Statistical analysis of count data
  - Comparison of proportions
  - 2-way tables: chi2 test, Fisher's exact test
- So how many replicates do I need?
  - Design of complex experiments

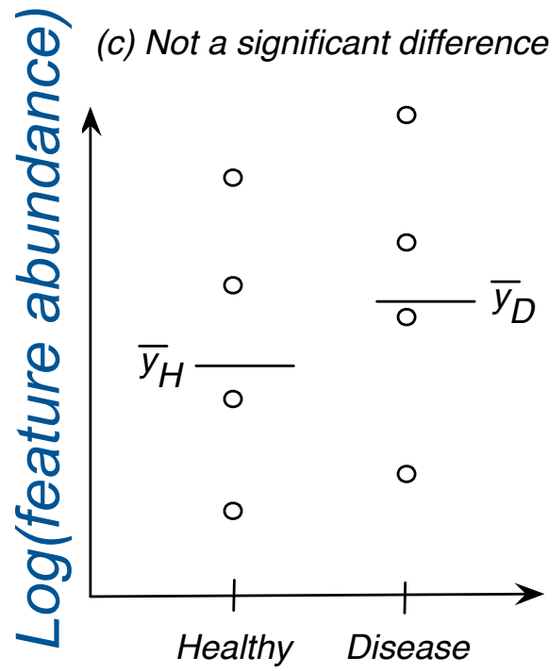
PRINCIPLE 1: REPLICATION

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









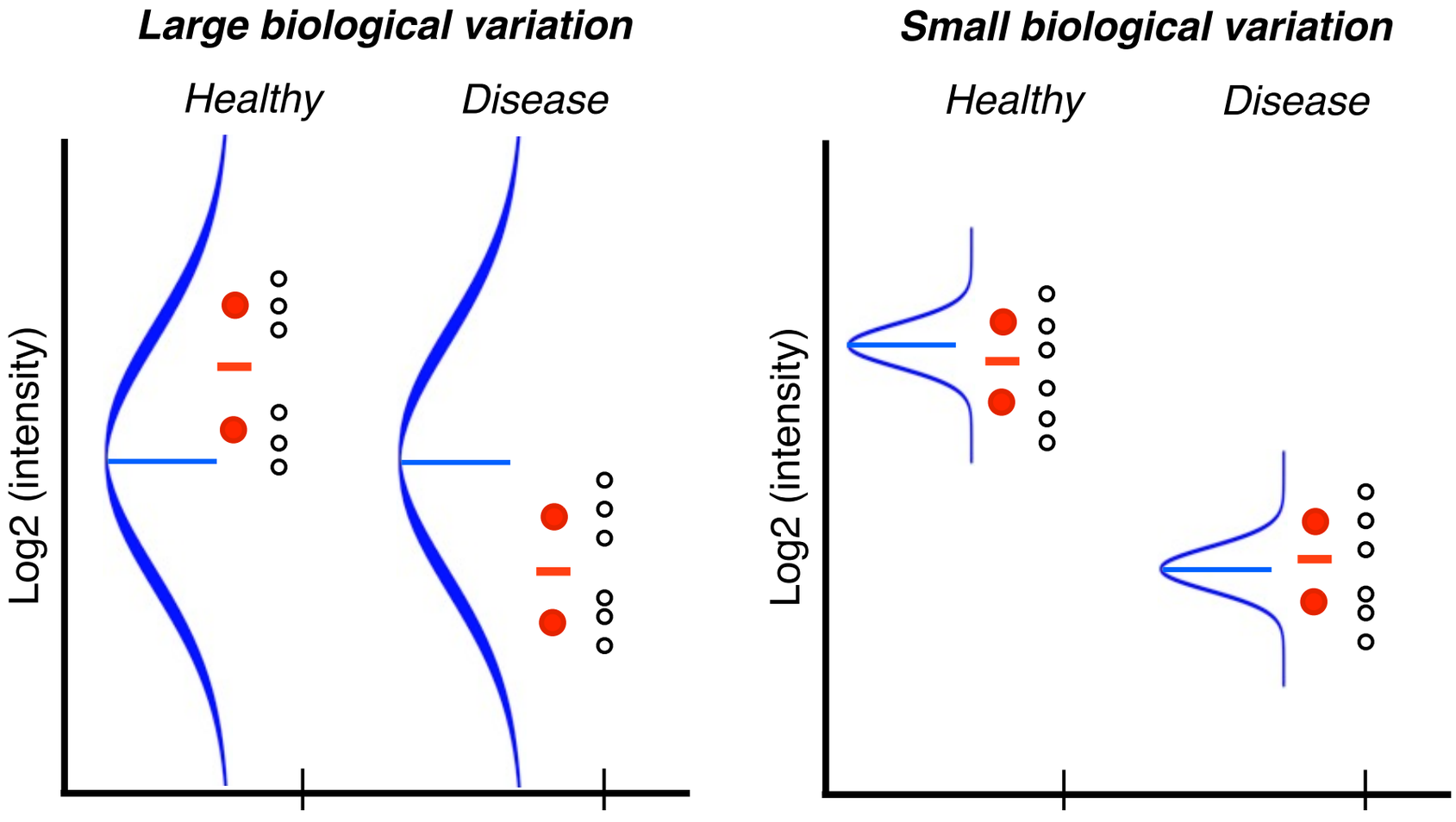
Two levels of randomness imply two types of replication:

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

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

MULTI-LAYER DESIGN AND ANALYSIS

Multiple sources are responsible for variation in measurements



When biological variance is large, more biological replicates are needed to accurately estimate the variance

Chang et al, *Molecular and Cellular Proteomics,* 11:M111.014662, 2012

BIOLOGICAL REPLICATION IS MOST IMPORTANT

Observed feature intensity	=	Systematic mean signal of disease group	+	Random deviation due to individual	+	Random deviation due to sample preparation	+	Random deviation due to measurement error
y_{ijkl}	=	Group mean_i	+	$\text{Indiv(Group)}_{j(i)} \sim N\left(0, \sigma^2_{\text{Indiv}}\right)$	+	$\text{Prep(Indiv)}_{k(ij)} \sim N\left(0, \sigma^2_{\text{Prep}}\right)$	+	$\text{Error}_{l(ijk)} \sim N\left(0, \sigma^2_{\text{Error}}\right)$

$$\text{Var}(\bar{y}_H - \bar{y}_D) = 2\left(\frac{\sigma^2_{\text{Indiv}}}{I} + \frac{\sigma^2_{\text{Prep}}}{IJ} + \frac{\sigma^2_{\text{Error}}}{IJK}\right)$$

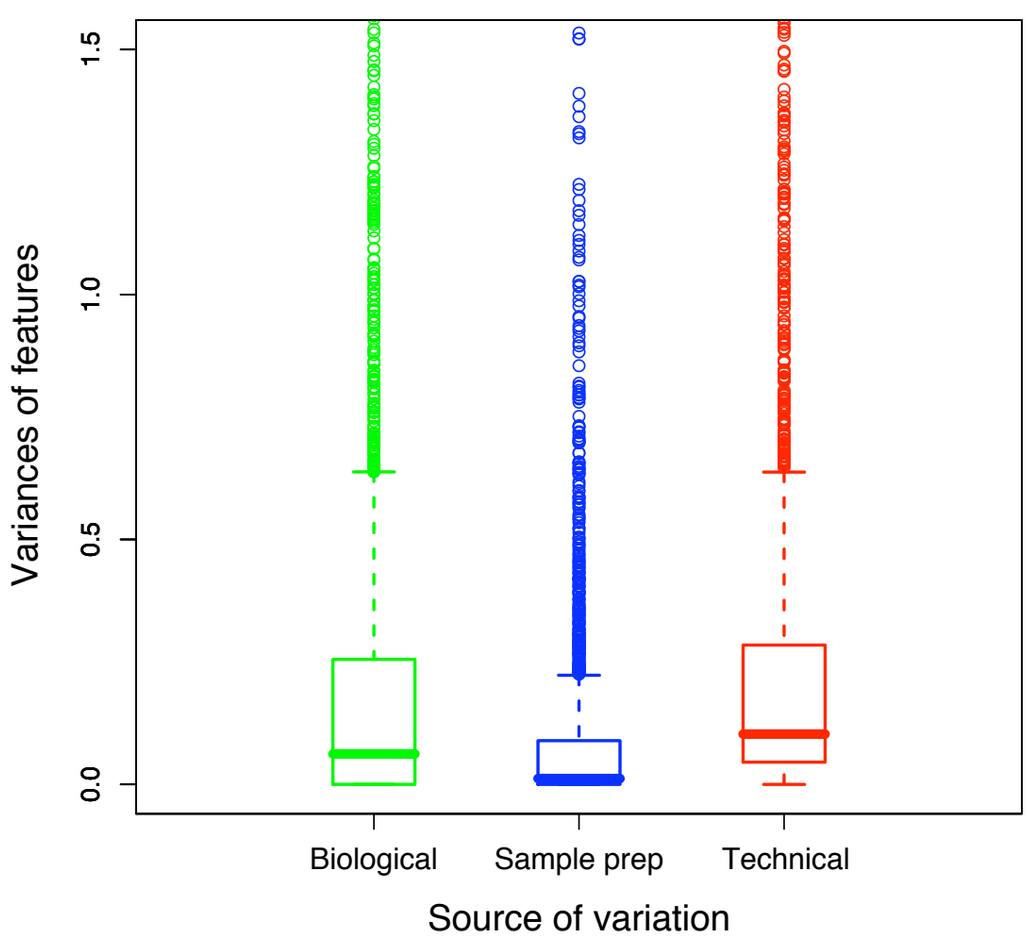
I: # individuals per disease group

J: # sample preps

K: # replicate runs

A pilot experiment

- 2 healthy individuals, 2 with diabetes
- multiple sample preparations
- multiple LC-MS replicates



BIOLOGICAL REPLICATION IS MOST IMPORTANT

Observed feature intensity	=	Systematic mean signal of disease group	+	Random deviation due to individual	+	Random deviation due to sample preparation	+	Random deviation due to measurement error
y_{ijkl}	=	Group mean_i	+	$\text{Indiv(Group)}_{j(i)} \sim N\left(0, \sigma^2_{\text{Indiv}}\right)$	+	$\text{Prep(Indiv)}_{k(ij)} \sim N\left(0, \sigma^2_{\text{Prep}}\right)$	+	$\text{Error}_{l(ijk)} \sim N\left(0, \sigma^2_{\text{Error}}\right)$

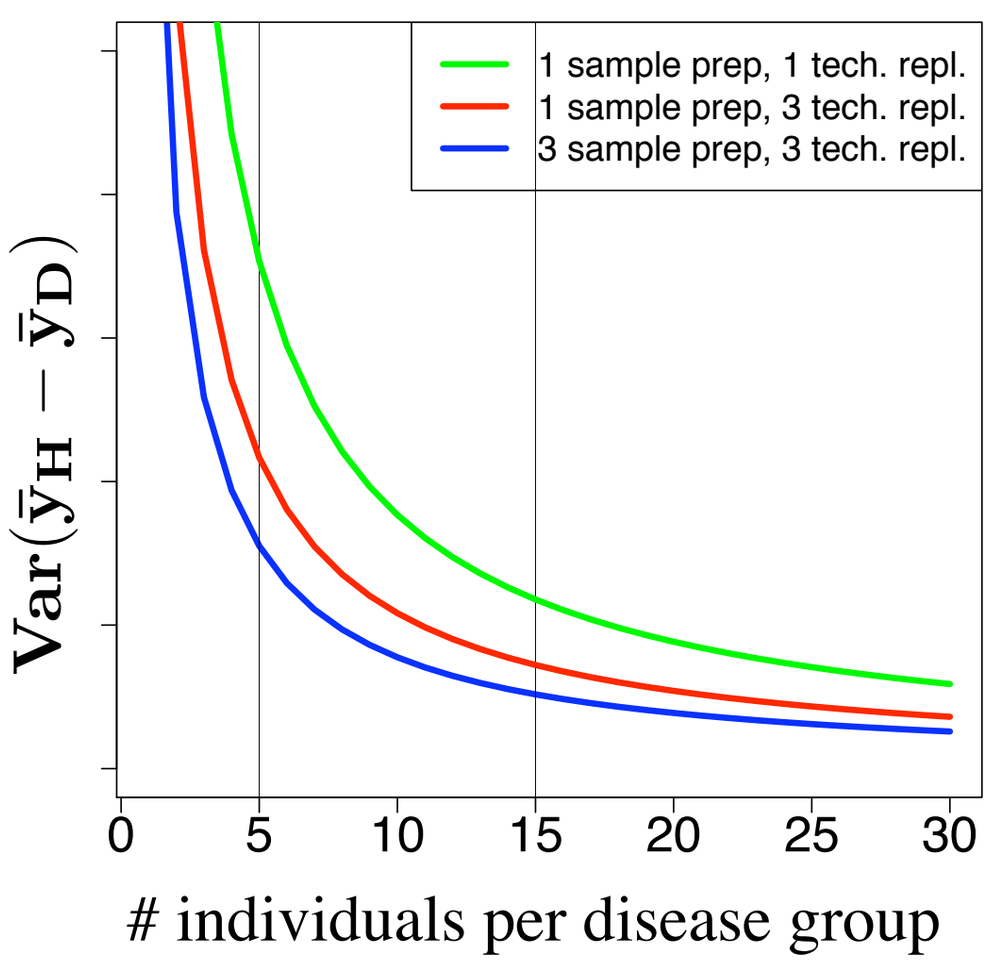
$$\text{Var}(\bar{y}_H - \bar{y}_D) = 2\left(\frac{\sigma^2_{\text{Indiv}}}{I} + \frac{\sigma^2_{\text{Prep}}}{IJ} + \frac{\sigma^2_{\text{Error}}}{IJK}\right)$$

I: # individuals per disease group

J: # sample preps

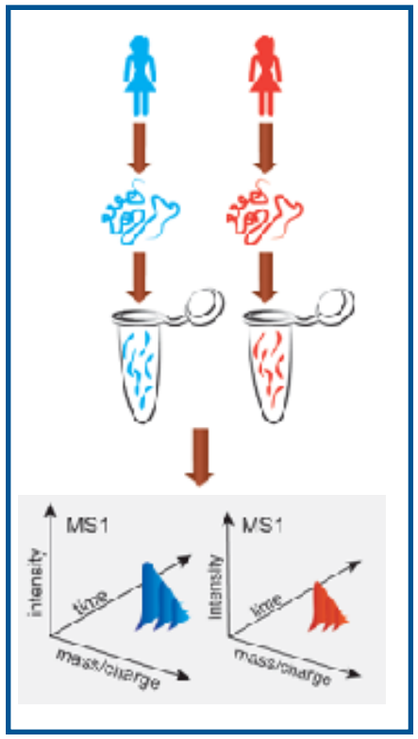
K: # replicate runs

Conclusion:
Maximize the number of biological replicates

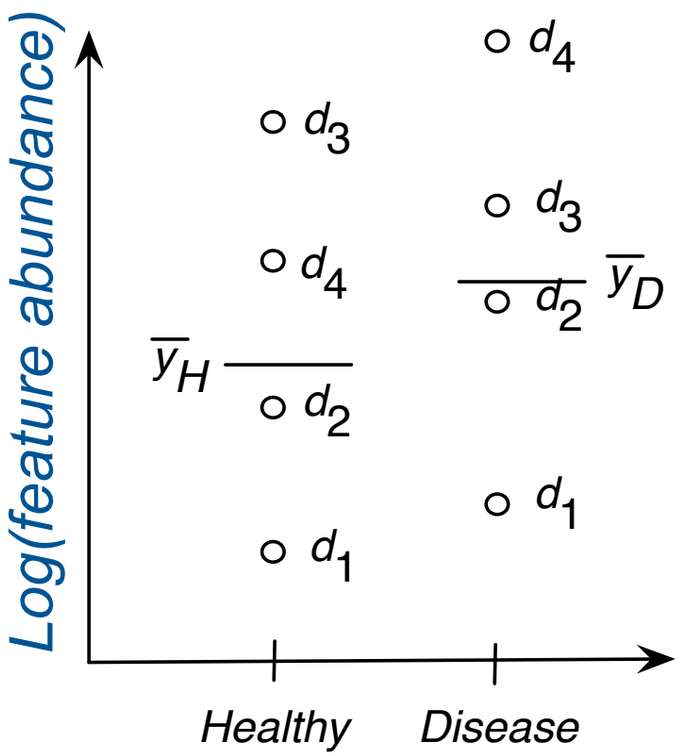


PRINCIPLE 3: BLOCKING

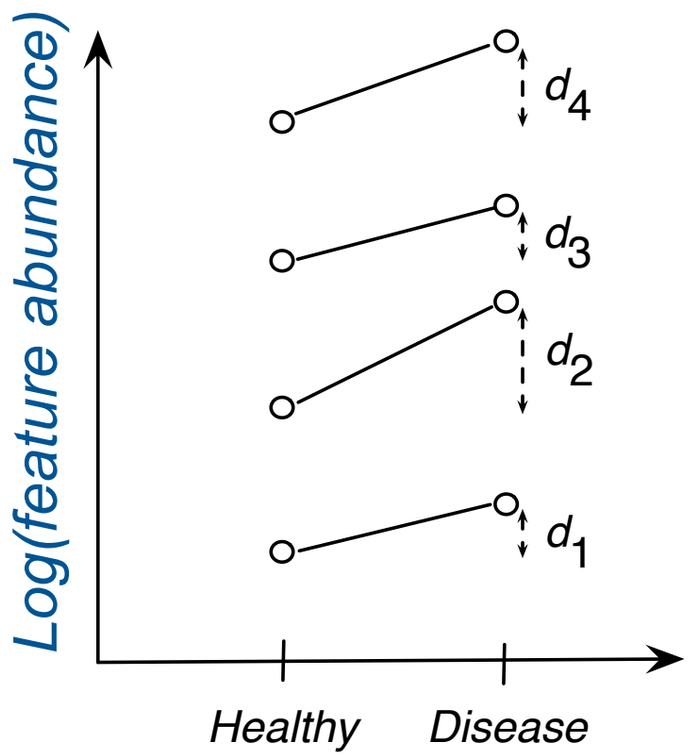
Helps reduce both bias and inefficiency



(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)

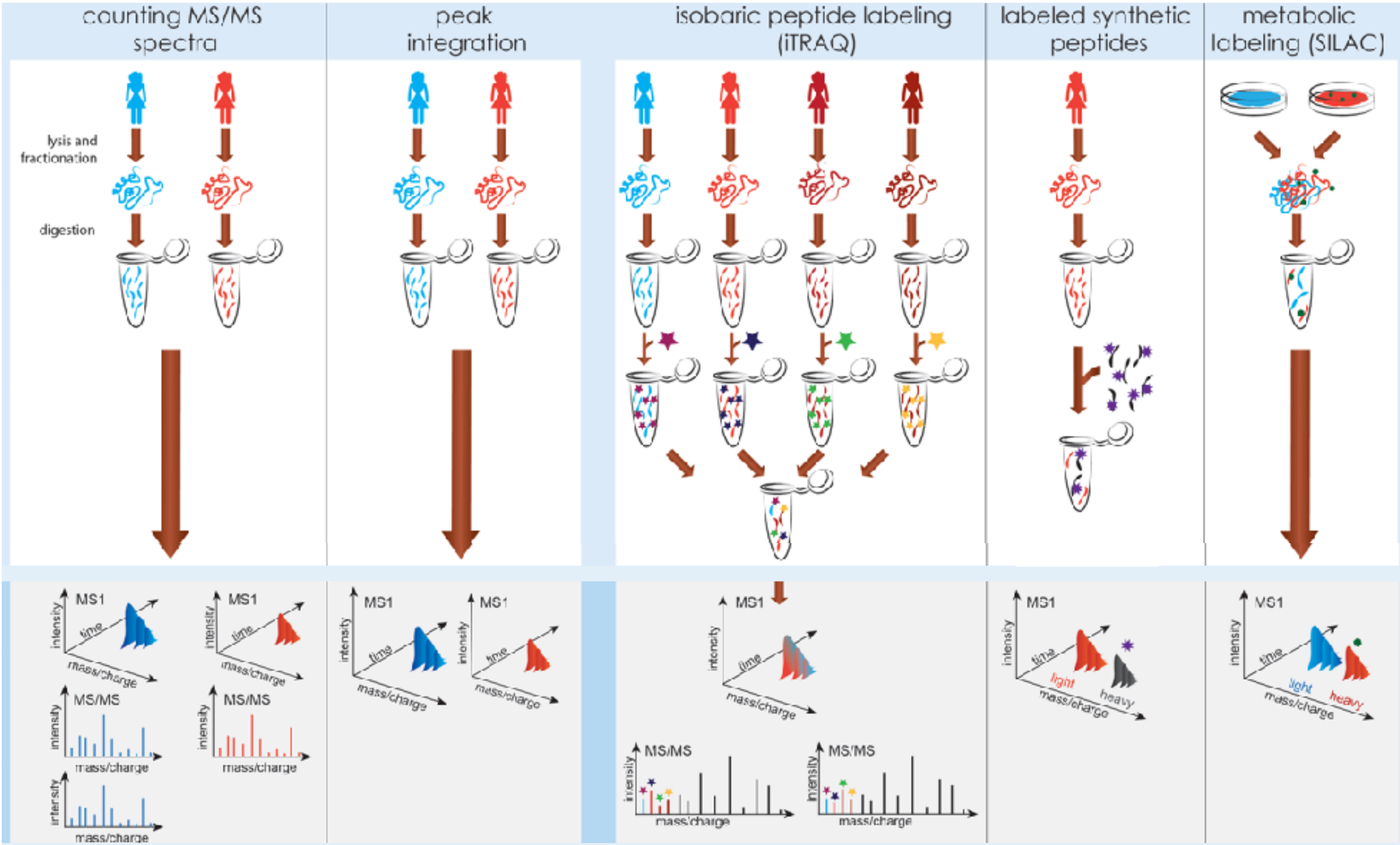
Oberg and Vitek, *J. Proteome Research*, 8, 2009

LABELED QUANTITATIVE WORKFLOWS

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

Sample preparation

Global LC-MS/MS



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

BLOCKING IS HELPFUL

Especially when between-block variance is large

Observed feature intensity	=	Systematic mean signal of disease group	+	Random deviation due to block (e.g. plate or day)	+	Random deviation due to individual	+	Random deviation due to measurement error
y_{ijkl}	=	Group mean_i	+	$\text{Block}_k \sim N(0, \sigma^2_{Block})$	+	$\text{Indiv(Group)}_{j(i)} \sim N(0, \sigma^2_{Indiv})$	+	$\text{Error}_{l(ijk)} \sim N(0, \sigma^2_{Error})$

A completely randomized design

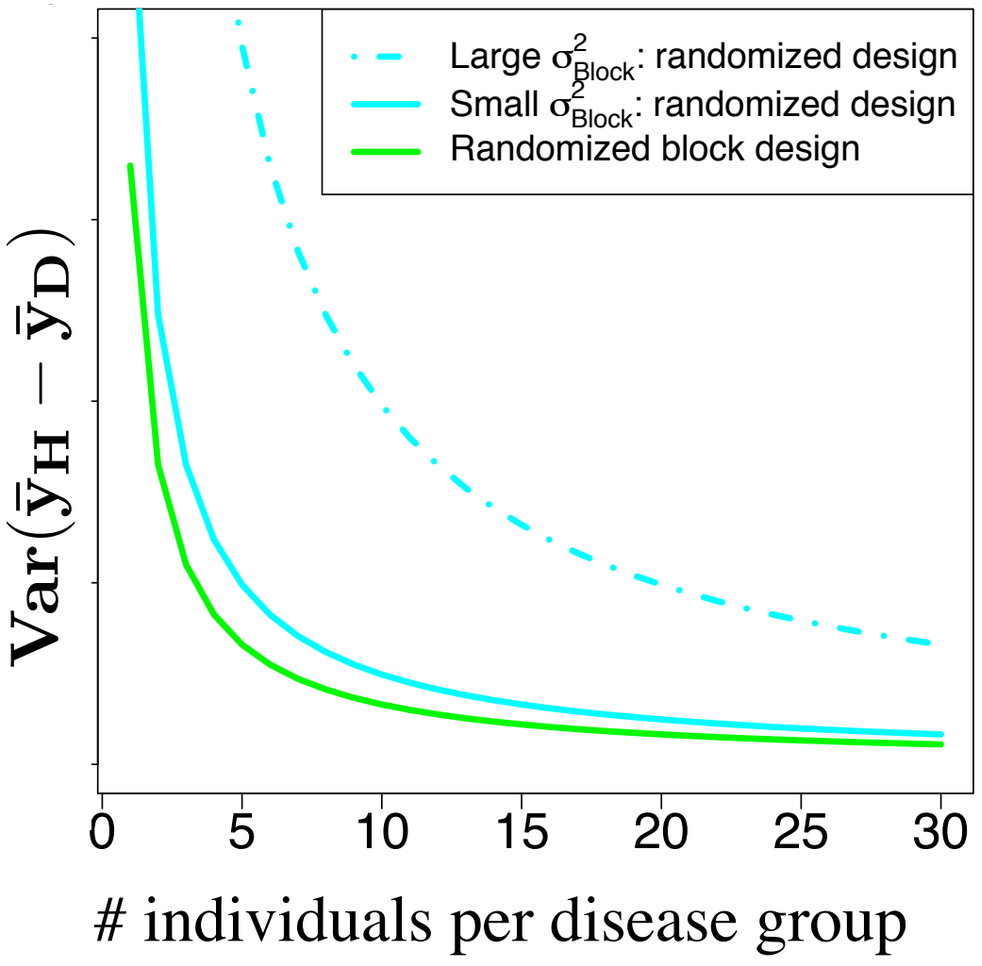
I: # individuals per disease group

$$\text{Var}(\bar{y}_H - \bar{y}_D) = 2\left(\frac{\sigma^2_{Block} + \sigma^2_{Indiv} + \sigma^2_{Error}}{I}\right)$$

A block-randomized design

$$\text{Var}(\bar{y}_H - \bar{y}_D) = 2\left(\frac{\sigma^2_{Indiv} + \sigma^2_{Error}}{I}\right)$$

Conclusion: Block-randomize
- if can not control a large source of variation
- if moderate sample size



SO HOW MANY REPLICATES DO I NEED?

If we only have one feature:

Fix: α - probability of a false positive discovery

β - probability of a true positive discovery

$\boldsymbol{\Delta}$ - anticipated fold change

$\sigma^2_{\mathbf{Indiv}}$ and $\sigma^2_{\mathbf{Error}}$ - anticipated variability

Write:

$$\mathbf{Var}(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}) \leq \left(\frac{\boldsymbol{\Delta}}{\mathbf{z_{1-\beta}} + \mathbf{z_{1-\alpha/2}}}\right)^{\mathbf{2}}$$

where $\mathbf{z_{1-\beta}}$ and $\mathbf{z_{1-\alpha/2}}$ are quantiles of Normal distribution

$$\mathbf{Var}(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}) = s_1^2/n_1 + s_2^2/n_2 \leq \left(\frac{\boldsymbol{\Delta}}{\mathbf{z_{1-\beta}} + \mathbf{z_{1-\alpha/2}}}\right)^{\mathbf{2}}$$

→ solve for the number of individuals n_1 and n_2

SO HOW MANY REPLICATES DO I NEED?

If we only have one feature:

Fix: α - probability of a false positive discovery

β - probability of a true positive discovery

$\boldsymbol{\Delta}$ - anticipated fold change

$\sigma^2_{\mathbf{Indiv}}$ and $\sigma^2_{\mathbf{Error}}$ - anticipated variability

Write:

$$\mathbf{Var}(\bar{\mathbf{y}}_{\mathbf{H}} - \bar{\mathbf{y}}_{\mathbf{D}}) \leq \left(\frac{\boldsymbol{\Delta}}{\mathbf{z}_{\mathbf{1}-\beta} + \mathbf{z}_{\mathbf{1}-\alpha/\mathbf{2}}} \right)^{\mathbf{2}}$$

where $\mathbf{z}_{\mathbf{1}-\beta}$ and $\mathbf{z}_{\mathbf{1}-\alpha/\mathbf{2}}$ are quantiles of the Normal distribution

A completely randomized design

I: # individuals per disease group

$$\mathbf{Var}(\bar{\mathbf{y}}_{\mathbf{H}} - \bar{\mathbf{y}}_{\mathbf{D}}) = \mathbf{2}\left(\frac{\sigma^2_{\mathbf{Block}} + \sigma^2_{\mathbf{Indiv}} + \sigma^2_{\mathbf{Error}}}{\mathbf{I}} \right)$$

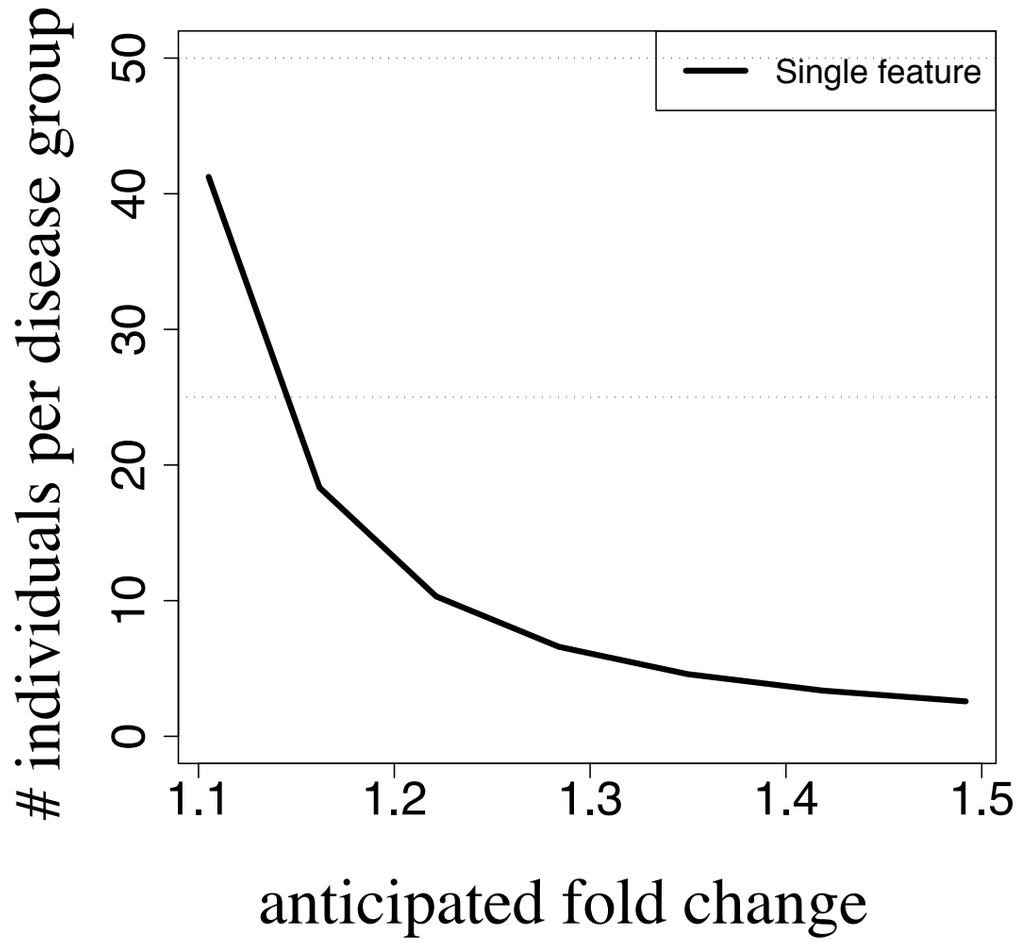
A block-randomized design

$$\mathbf{Var}(\bar{\mathbf{y}}_{\mathbf{H}} - \bar{\mathbf{y}}_{\mathbf{D}}) = \mathbf{2}\left(\frac{\sigma^2_{\mathbf{Indiv}} + \sigma^2_{\mathbf{Error}}}{\mathbf{I}} \right)$$

→ solve for the number of individuals I

SO HOW MANY REPLICATES DO I NEED?

Example: pilot study with diabetes patients.
A block-randomized design
If we only had one feature:



Conclusion:
The smaller the anticipated difference, the larger the sample size