Plan for the day

- **Morning**
  - 9:00am-10:00am Olga: Statistical experimental design
  - 10:00am-10:30am Brendan: Data processing with Skyline
  - 10:30am-11:00am *Coffee*
  - 11:00am-12:00pm Brendan: Data processing with Skyline
- **Afternoon**
  - 1:00pm-2:00pm Olga: Statistical significance analysis
  - 2:00pm-2:30pm Meena: Statistical analysis case studies
  - 2:30pm-3:00pm *Coffee*
  - 3:00pm-4:00pm Meena: Statistical analysis case studies

Steps of statistical significance analysis

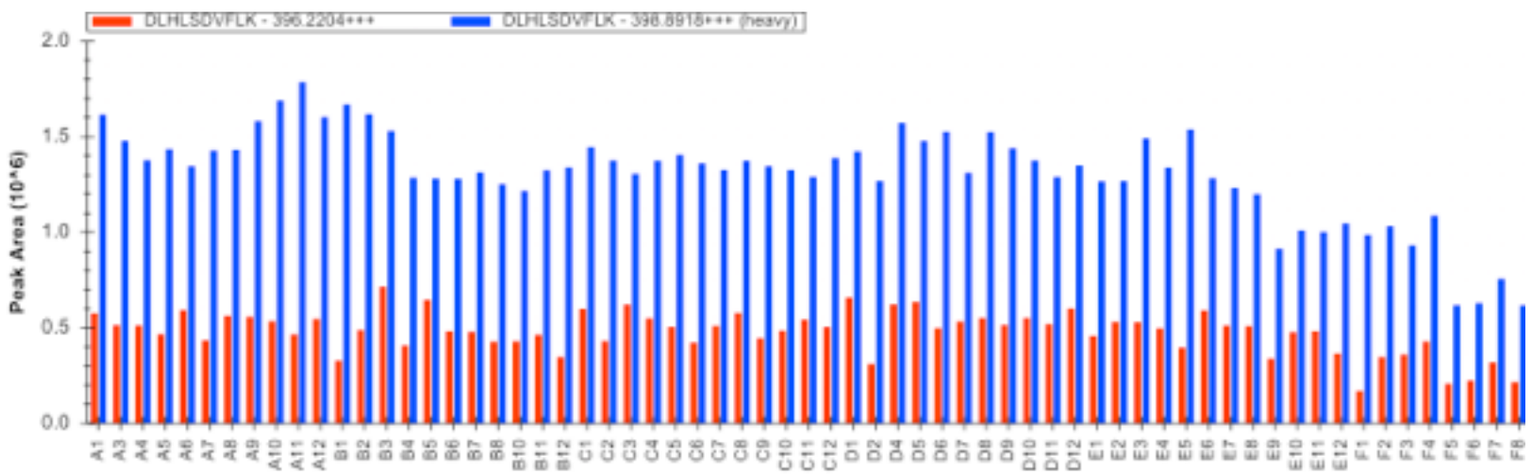
- Define the analysis protocol
  - Type of analysis and comparisons of interest
  - Scope of conclusions
  - Model type
- Normalization and quality control
- Model-based analysis
  - Specify the model
  - Perform-based comparisons
  - Control for multiple testing
- Use the experiment to gain insight into future studies
  - Compare strategies of future resource allocation
  - Calculate sample size of a future similar experiment

Example: ovarian cancer dataset

- 5 cancer patients and 10 controls
- 3 peptides/protein; 3 transitions/peptide

Brendan, with Skyline:

Highlights peak areas



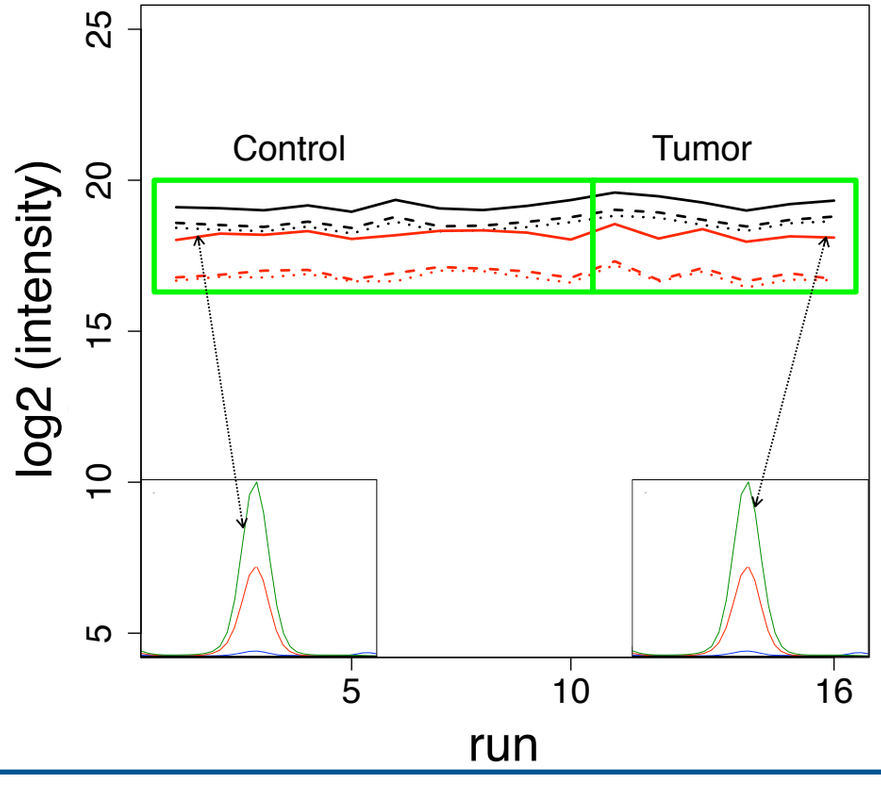
Meena, with MSstats2:

Highlights between-run and between-peptide interferences

Colors = peptides
Line types = transitions

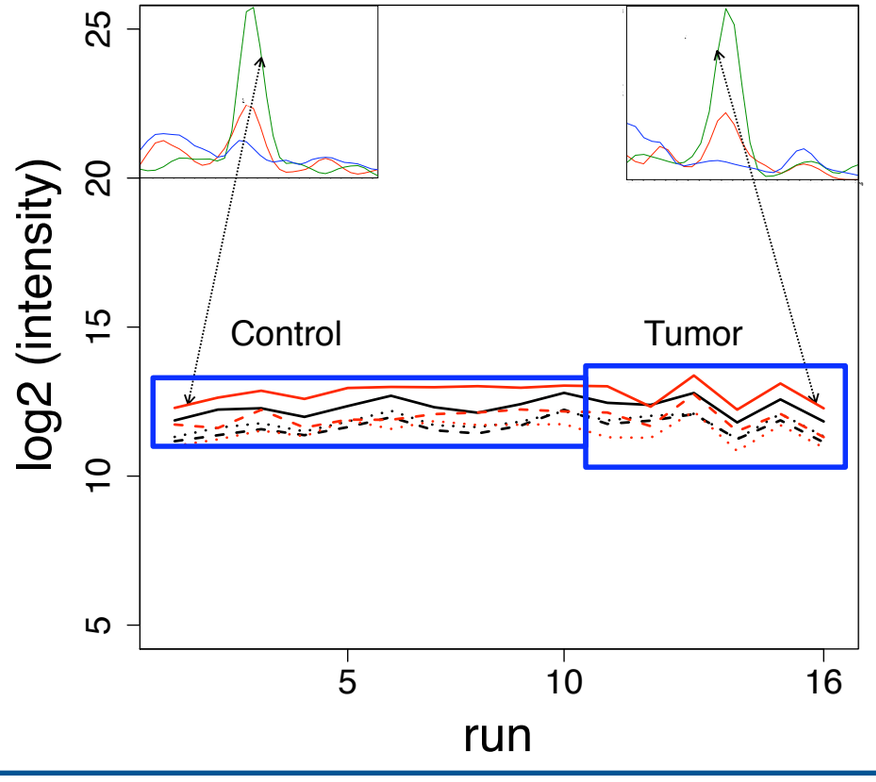
CLU

Stable isotope reference peptides



CLU

Endogenous peptides

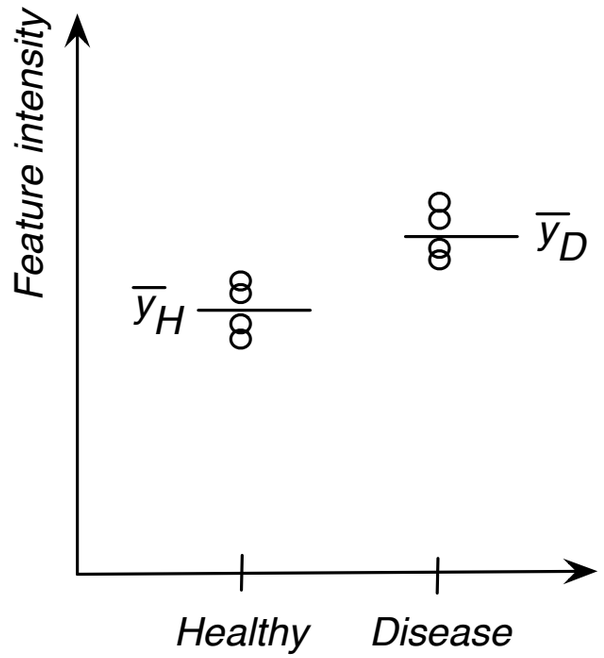


Differentially abundant proteins are not always biomarkers

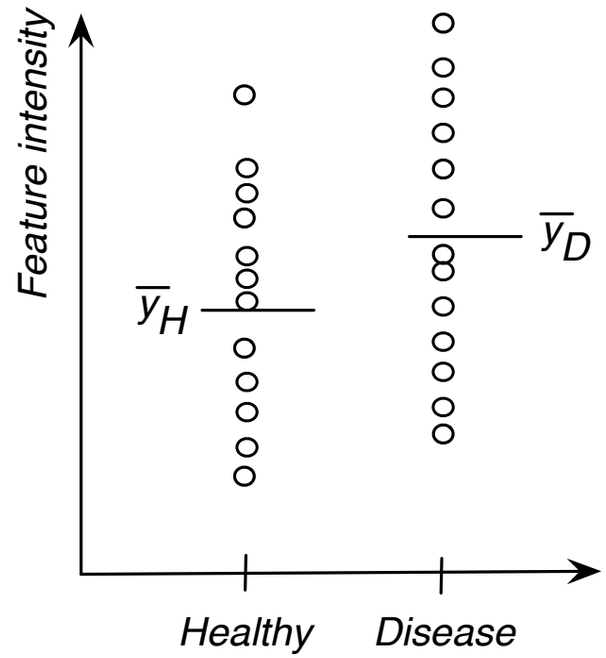
Biomarker: *a (molecular) measurement*

- *predictive of the outcome of the disease or*
- *predictive of therapy response*

$$\frac{\bar{y}_{\mathrm{Disease}} - \bar{y}_{\mathrm{Control}}}{\sqrt{2 \cdot s^2/n}} \sim \mathrm{Student}_d$$



Differentially abundant and predictive



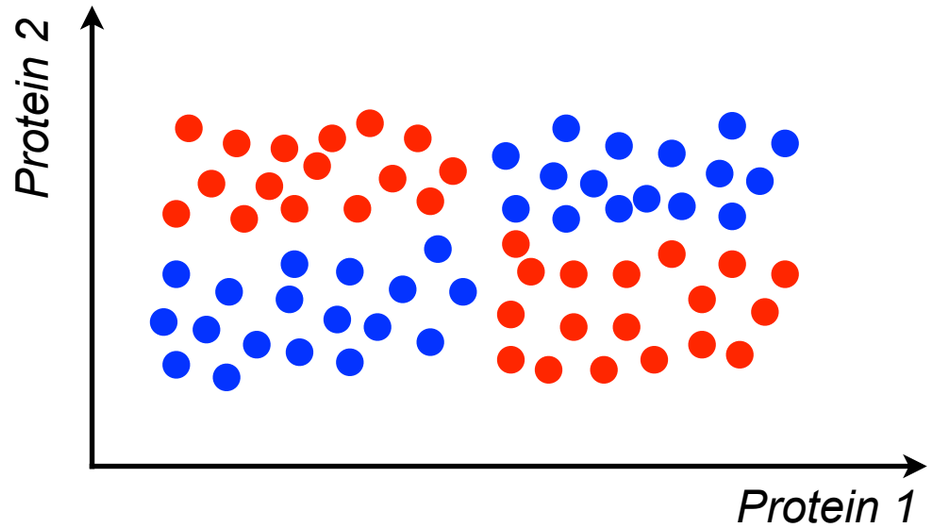
Differentially abundant and not predictive

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

Biomarkers are not always differentially abundant proteins

Biomarker: *a (molecular) measurement*

- *predictive of the outcome of the disease or*
- *predictive of therapy response*

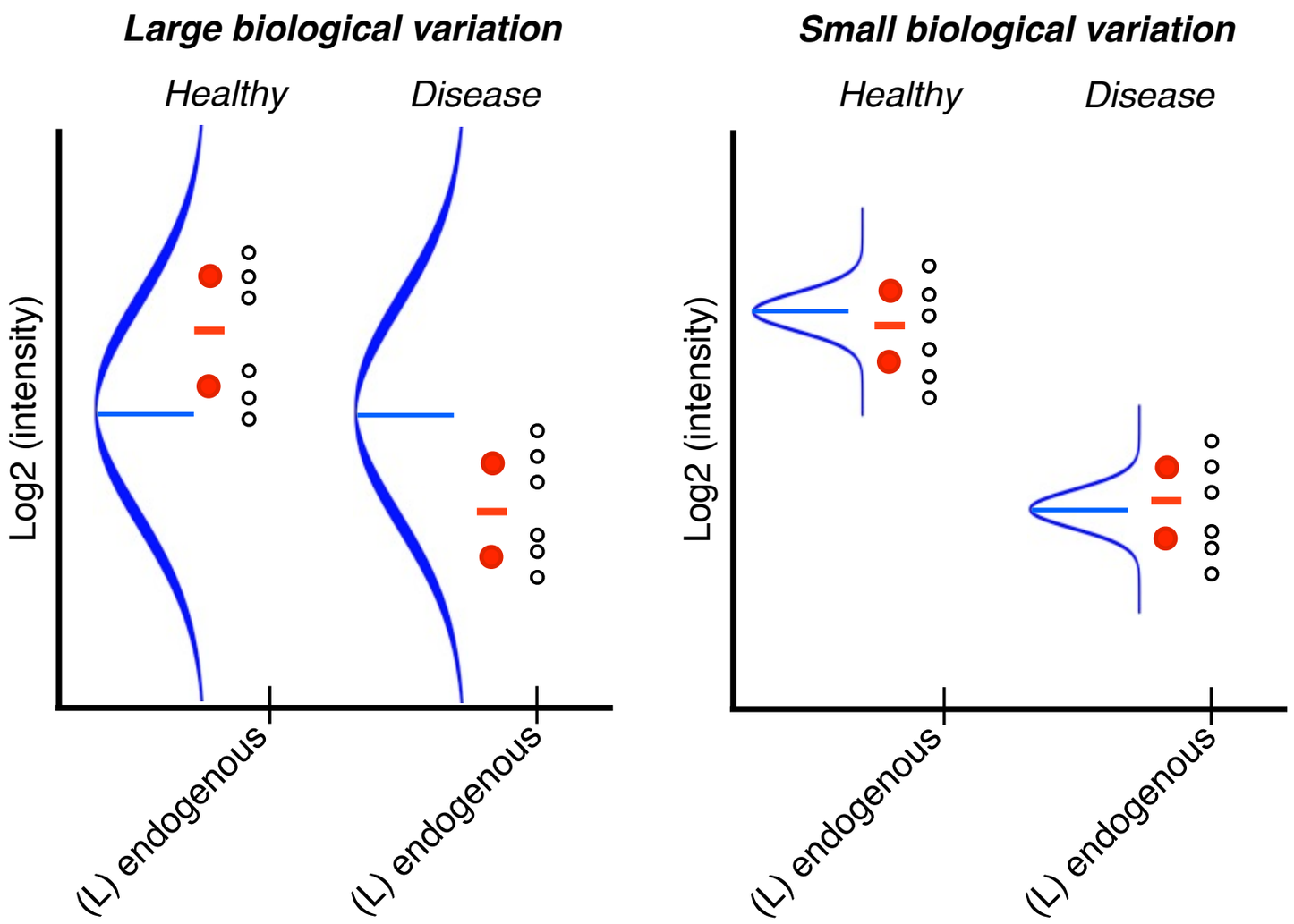


Not differentially abundant but predictive

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

Since the ovarian cancer study is a screening experiment, testing is appropriate

Different scope of conclusions ask different biological questions and leads to different results



Protein abundance in the population

Average abundance in the population

Average abundance in the selected subjects

Abundance in the individual subject

Observed log-intensities of L

Larger biological variation requires more replicates

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

Since the ovarian cancer study is a screening experiment, testing with a restricted scope of conclusions is appropriate

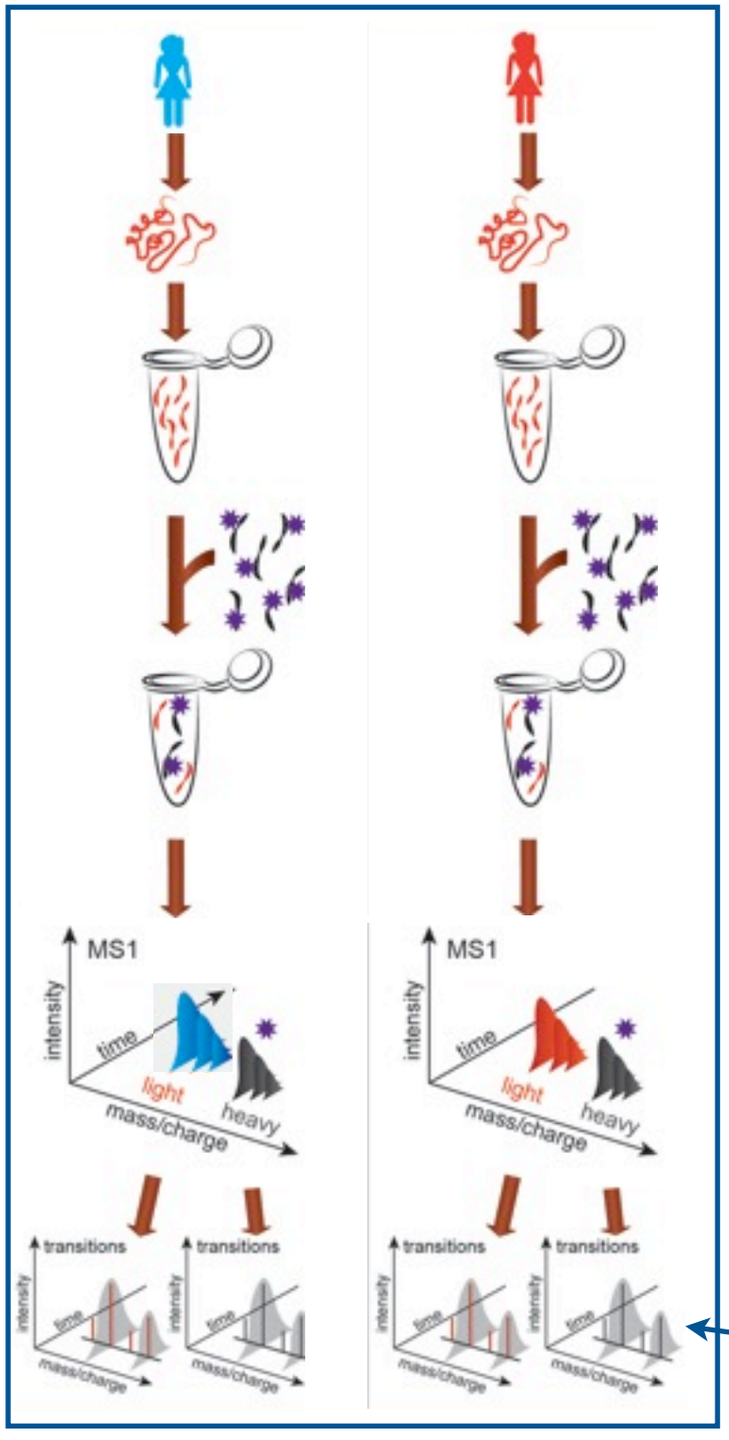
These considerations, and the extent of anticipated interferences in peak intensities, defines the model type

Steps of statistical significance analysis

- Define the analysis protocol
  - Type of analysis and comparisons of interest
  - Scope of conclusions
  - Model type
- Normalization and quality control
- Model-based analysis
  - Specify the model
  - Perform-based comparisons
  - Control for multiple testing
- Use the experiment to gain insight into future studies
  - Compare strategies of future resource allocation
  - Calculate sample size of a future similar experiment

In label-based SRM, reference intensities serve as internal normalization factors

Synthetic standards



			Group 1						Group I	
			Run 1	⋯	⋯	⋯	⋯	⋯	⋯	Run M
			Subject 1	⋯	Subject J	⋯	Subject 1	⋯		Subject J
Endogenous: light labeled peptide	Peptide 1	Transition 1	10.21	⋯	10.57	⋯	15.64	⋯		15.03
		⋮		⋯		⋯		⋯		
		Transition L	10.52	⋯	10.92	⋯	15.29	⋯		15.68
	⋮	⋮		⋯		⋯		⋯		
	Peptide K	Transition 1	11.76	⋯	11.92	⋯	16.22	⋯		16.71
		⋮		⋯		⋯		⋯		
		Transition L	11.65	⋯	11.09	⋯	16.27	⋯		16.51
Reference: heavy labeled peptide	Peptide 1	Transition 1	19.46	⋯	19.77	⋯	19.82	⋯		19.03
		⋮		⋯		⋯		⋯		
		Transition L	19.13	⋯	19.25	⋯	19.67	⋯		19.80
	⋮	⋮		⋯		⋯		⋯		
	Peptide K	Transition 1	19.26	⋯	19.33	⋯	19.58	⋯		19.61
		⋮		⋯		⋯		⋯		
		Transition L	19.73	⋯	19.09	⋯	19.84	⋯		19.55

Legend : Label | Feature: Transition/Peptide | Group | Run | Subject

Normalization is performed as part of the model-based significance analysis

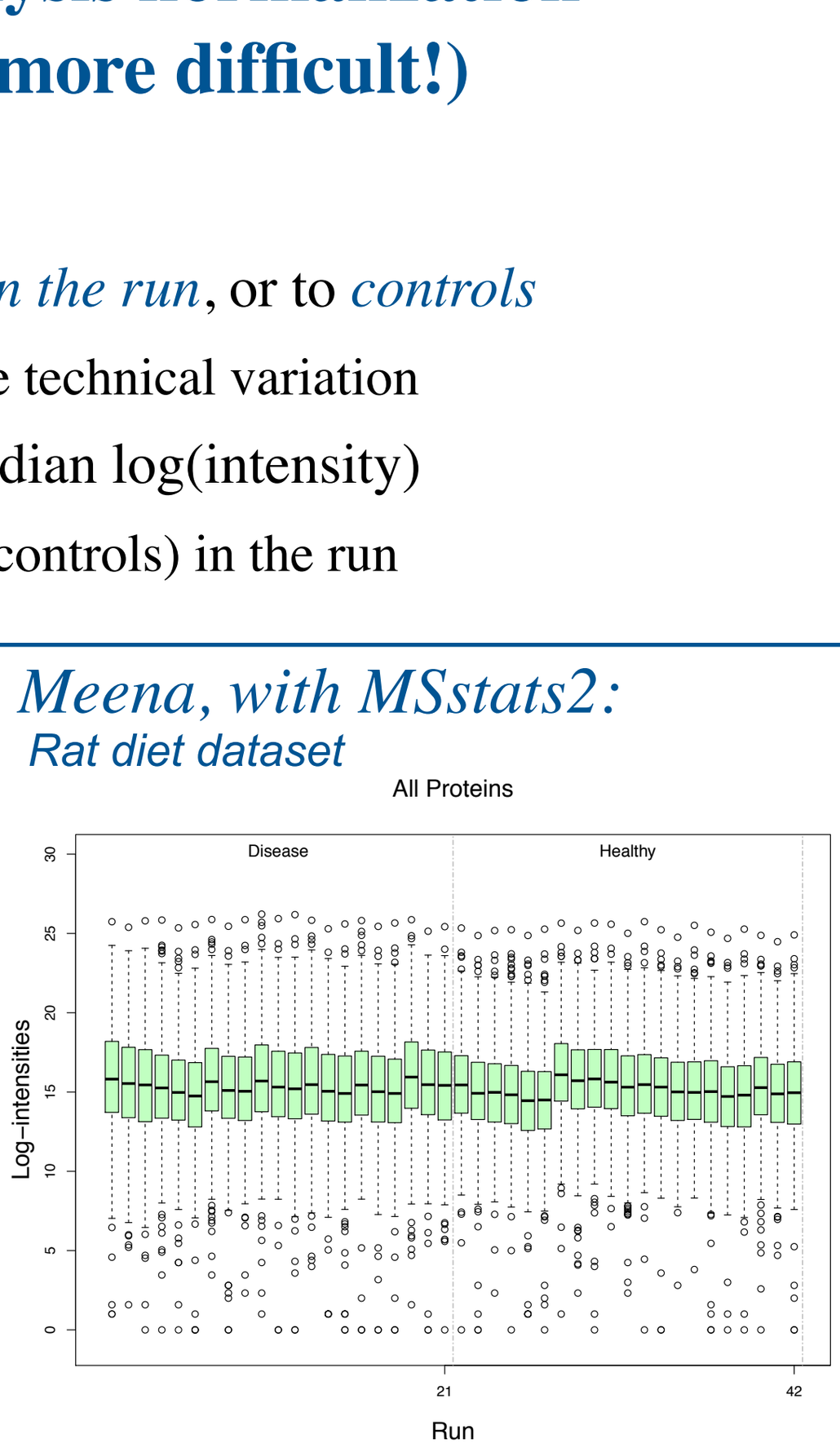
Transitions

In label-free SRM, pre-analysis normalization is more important (and more difficult!)

- Constant normalization
  - Normalize with respect to *all features in the run*, or to *controls*
    - Controls: less biological variation, more technical variation
  - Assumption: all runs have the same median log(intensity)
    - Subtract median[log(intensity)] (of the controls) in the run
    - Add the median of all medians
- Quantile normalization
  - Assumption: all runs have the same distribution of intensities
    - Not just the medians!
    - Too aggressive when the number of features is small (in hundreds)
  - *Global normalization has worked best for us so far*

Meena, with MSstats2:

Rat diet dataset



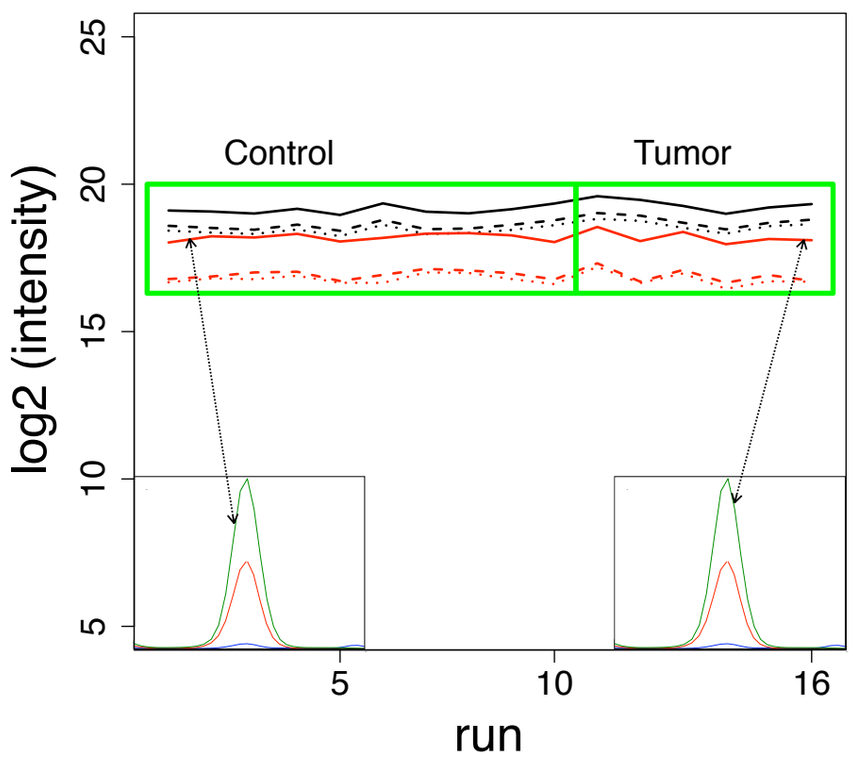
Steps of statistical significance analysis

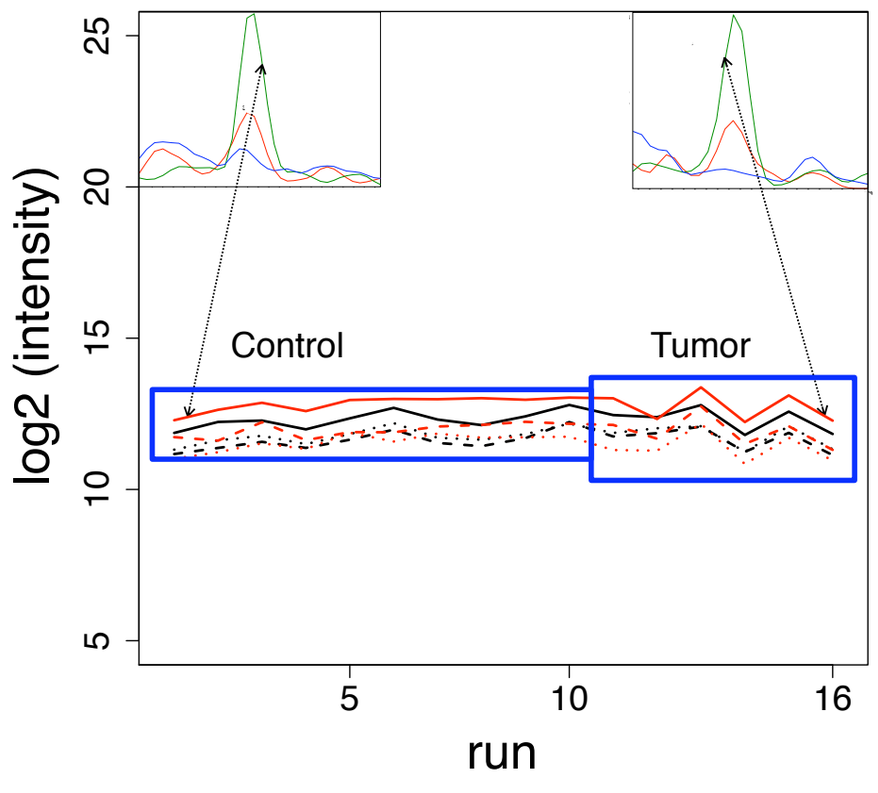
- Define the analysis protocol
  - Type of analysis and comparisons of interest
  - Scope of conclusions
  - Model type
- Normalization and quality control
- Model-based analysis
  - Specify the model
  - Perform-based comparisons
  - Control for multiple testing
- Use the experiment to gain insight into future studies
  - Compare strategies of future resource allocation
  - Calculate sample size of a future similar experiment

Linear mixed effects model describes the systematic and the random sources of variation

Example: ovarian cancer dataset

observed log2(int of peak)	=	overall mean	+	group or time	+ subject +	feature	+	run	+	group by feature	+	run by feature	+	random error
y_{ijklm}	=	μ	+	G_i	$+S(G)_{j(i)}+$	F_{kl}	$+$	R_m	$+$	$(G\times F)^*_{ikl}$	$+$	$(R\times F)^*_{klm}+$		ε_{ijklm}
Fixed/Random		F		F	F/R	F		F/R		F		F/R		R: N(0,σ^2)





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

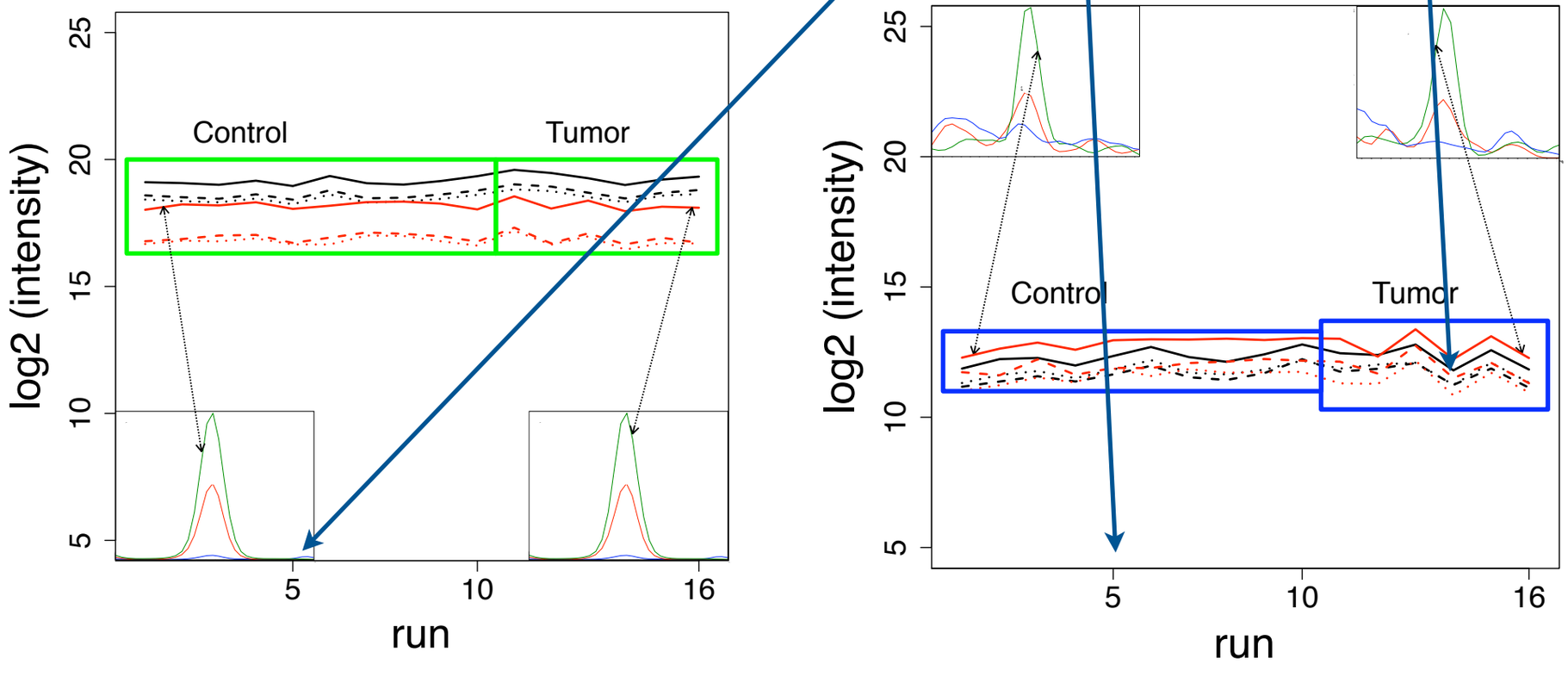
Model log2(int) instead of ratios light/heavy

'Run' pairs the endogenous and reference intensities

Linear mixed effects model describes the systematic and the random sources of variation

Example: ovarian cancer dataset

observed log2(int of peak)	=	overall mean	+	group or time	+ subject +	feature +	run +	group by feature	+	run by feature	+	random error
y_{ijklm}	=	μ	+	G_i	$+S(G)_{j(i)}+$	F_{kl}	$+R_m+$	$(G\times F)^*_{ikl}$	$+(R\times F)^*_{klm}+$			ε_{ijklm}
Fixed/Random		F		F	F/R	F	F/R	F		F/R		R: N(0,σ^2)



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

Model log2(int) instead of ratios light/heavy

'Run' pairs endogenous and reference transitions from a same run

Linear mixed effects model describes the systematic and the random sources of variation

Example: ovarian cancer dataset

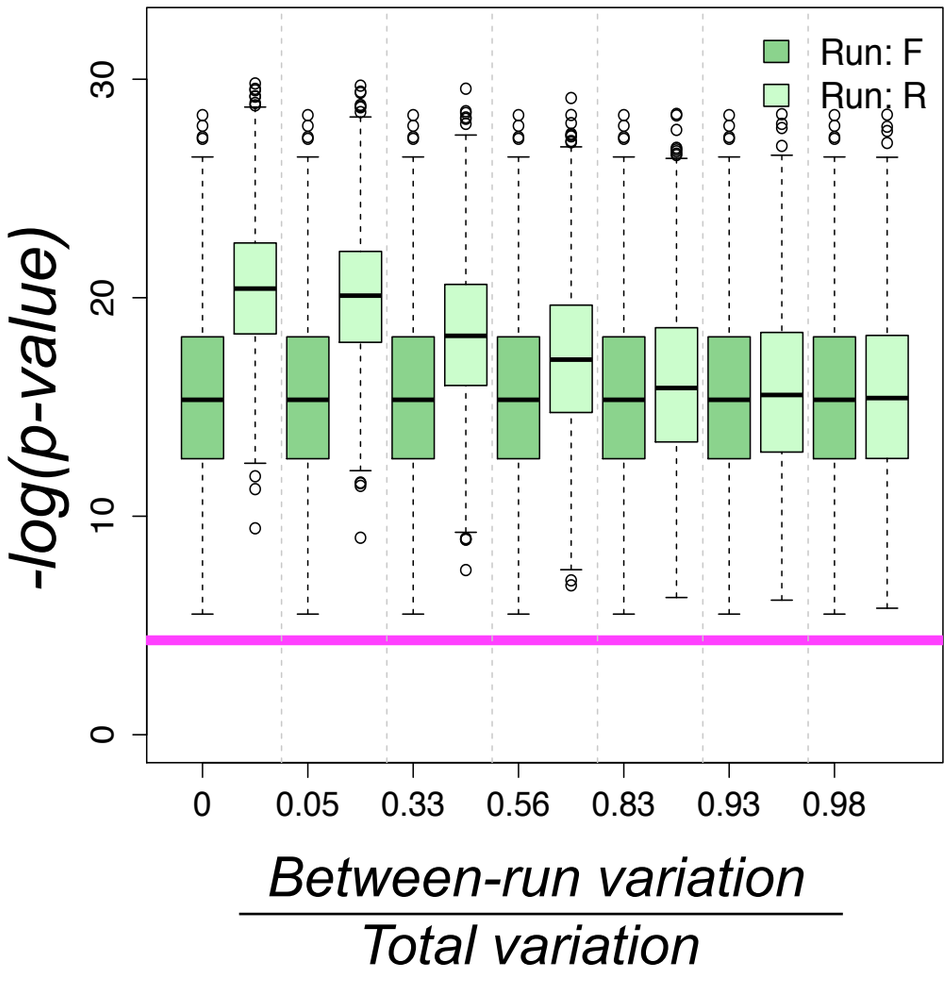
observed log2(int of peak)	=	overall mean	+	group or time	+ subject +	feature	+	run	+	group by feature	+	run by feature	+	random error
y_{ijklm}	=	μ	+	G_i	$+S(G)_{j(i)}+$	F_{kl}	$+$	R_m	$+$	$(G\times F)^*_{ikl}$	$+$	$(R\times F)^*_{klm}+$		ε_{ijklm}
Fixed/Random		F		F	F/R	F		F/R		F		F/R		R: N(0,σ^2)

- Can express the scope of conclusions
  - F: restricted, e.g. $\sum_{j=0}^{J} S(G)_{j(i)} = 0$ R: expanded, e.g. $S(G)_{j(i)} \overset{iid}{\sim} N(0, \sigma^2_S)$
- In some cases, same conclusions as with the ratios
  - When no missing values, restricted scope of run, expanded scope of subject
- Advantage: can be modified to more generality
  - Missing values, flexible scopes, random interferences, unequal variance
- Can express other designs
  - Time course: add *G*x*S* interaction. Label-free: no *R* and *R*x*F* terms

Advantage: appropriately modifying the assumptions improves the accuracy

observed log2(int of peak)	=	overall mean	+	group or time	+ subject	+ feature	+ run	+ group by feature	+	run by feature	+	random error
y_{ijklm}	=	μ	+	G_i	$+S(G)_{j(i)}+$	F_{kl}	$+R_m$	$+(G\times F)^*_{ikl}$		$+(R\times F)^*_{klm}+$		ε_{ijklm}
Fixed/Random		F		F	F/R	F	F/R	F		F/R		R: N(0,σ^2)

Expanded scope of technical replication



Estimated log-fold change:

ratio:
$$\left(\bar{y}_{i\cdots m}-\bar{y}_{0\cdots m}\right)-\left(\bar{y}_{i'\cdots m'}-\bar{y}_{0\cdots m'}\right)$$

intensity:
$$\left(\bar{y}_{i\cdots m}-w\cdot\bar{y}_{0\cdots m}\right)-\left(\bar{y}_{i'\cdots m'}-w\cdot\bar{y}_{0\cdots m'}\right)$$

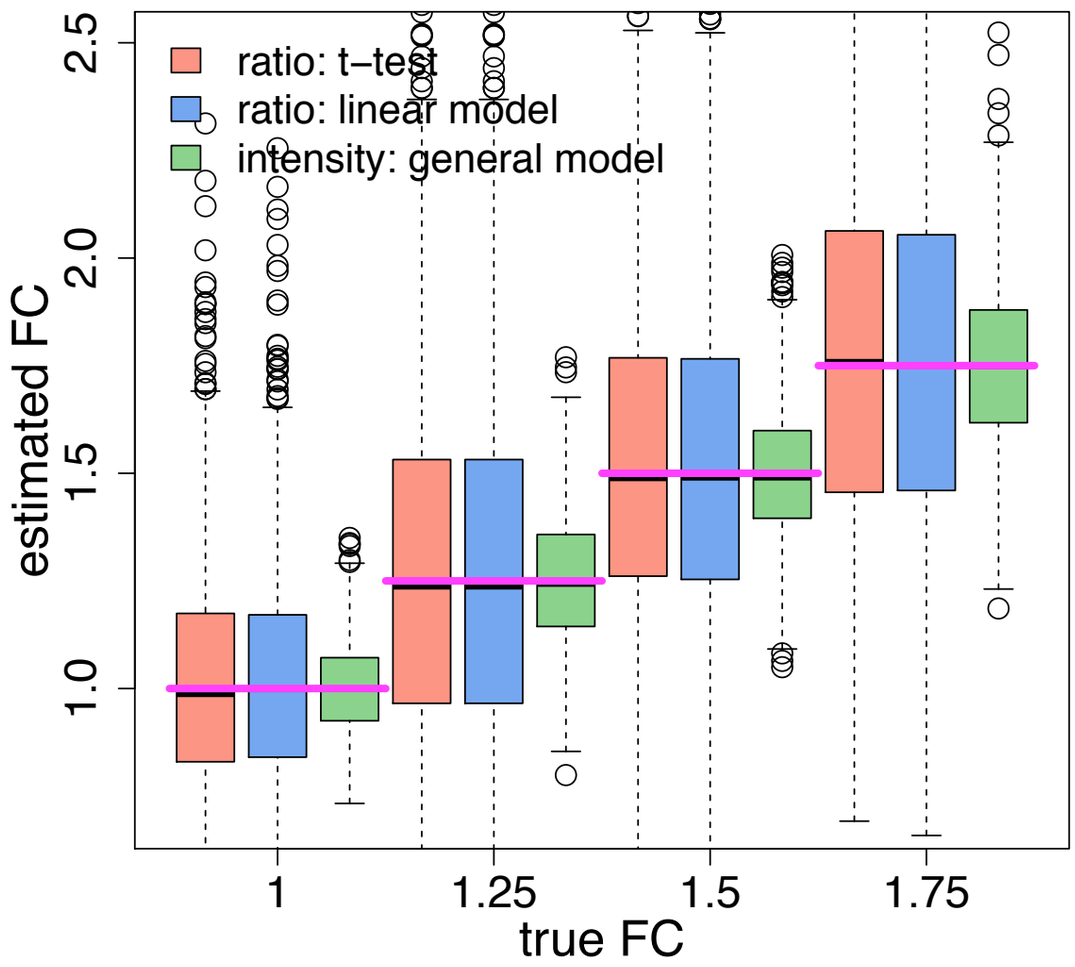
When between-run variation is small, reference intensities are used less

Advantage: better handling of missing data

Do not discard peaks with a missing counterpart

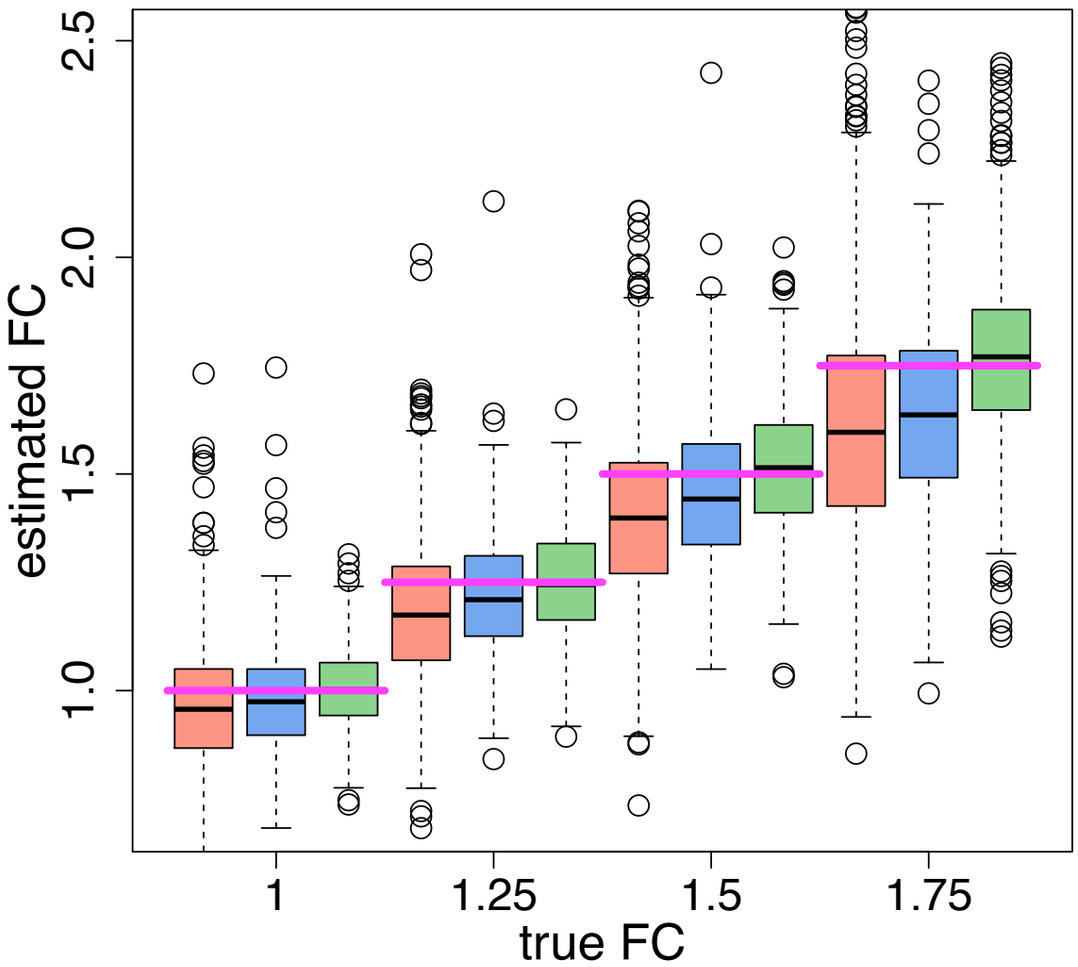
- ratio: t–test
- ratio: linear model
- intensity: general model

Transitions missing at random



Linear models have less variation

Transitions missing at low abundance



Linear models have less bias

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

Advantage: can compare label-free and label-based designs by simulation

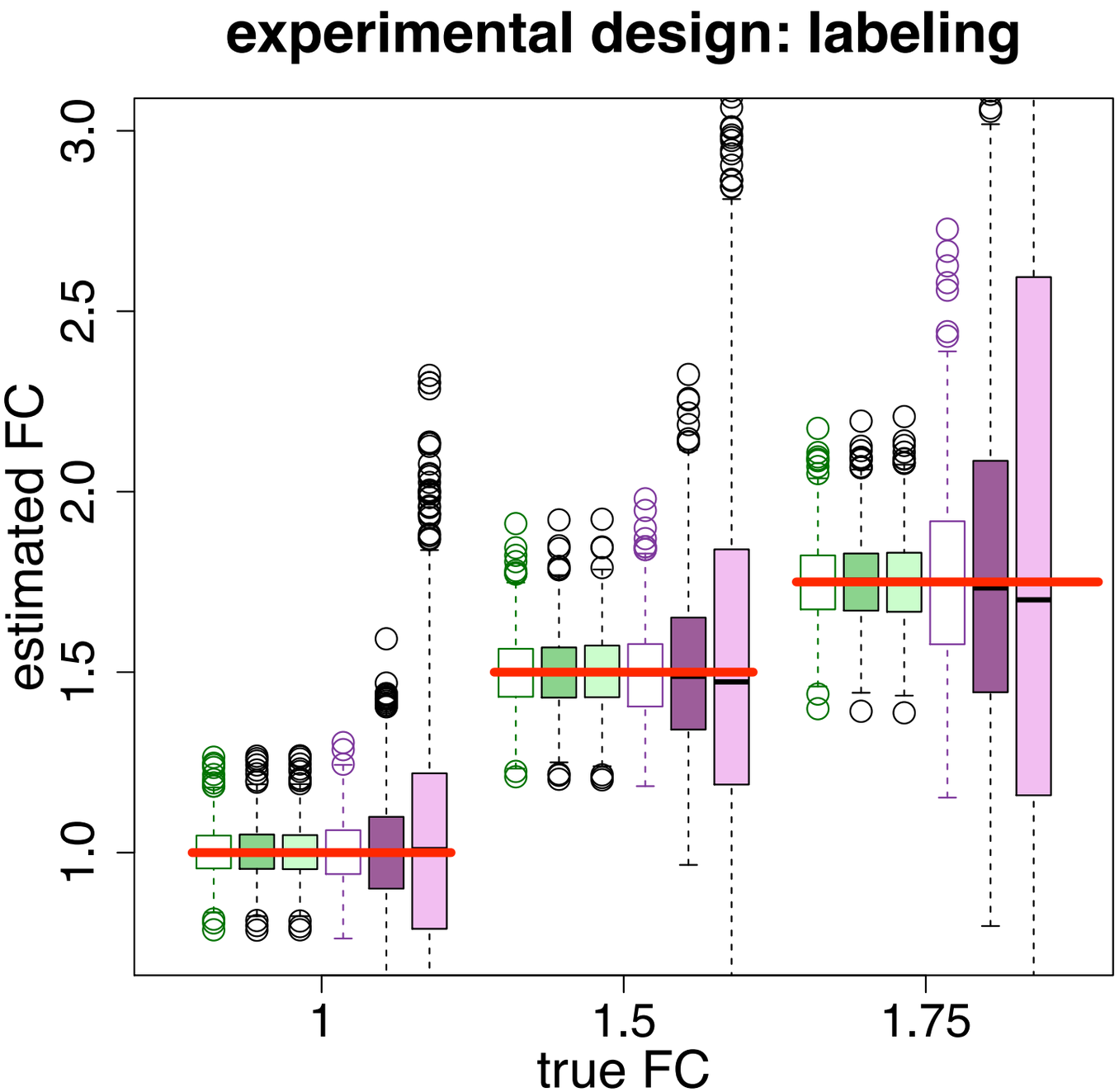
$$\frac{\textit{Between-run variation}}{\textit{Error variation}}$$

Label-based

- 0.5
- 2
- 10

Label-free

- 0.5
- 2
- 10



Caveat: this assumes same technical variation in both workflows.
In practice, label-free experiments can have larger variation.

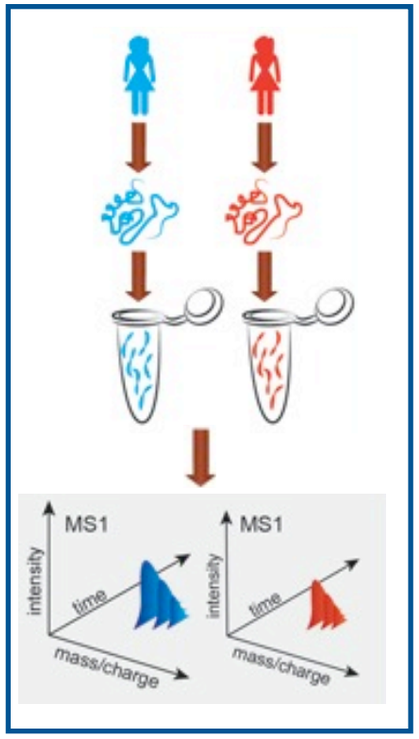
C.-Y. Chang et al., *MCP*, 2012

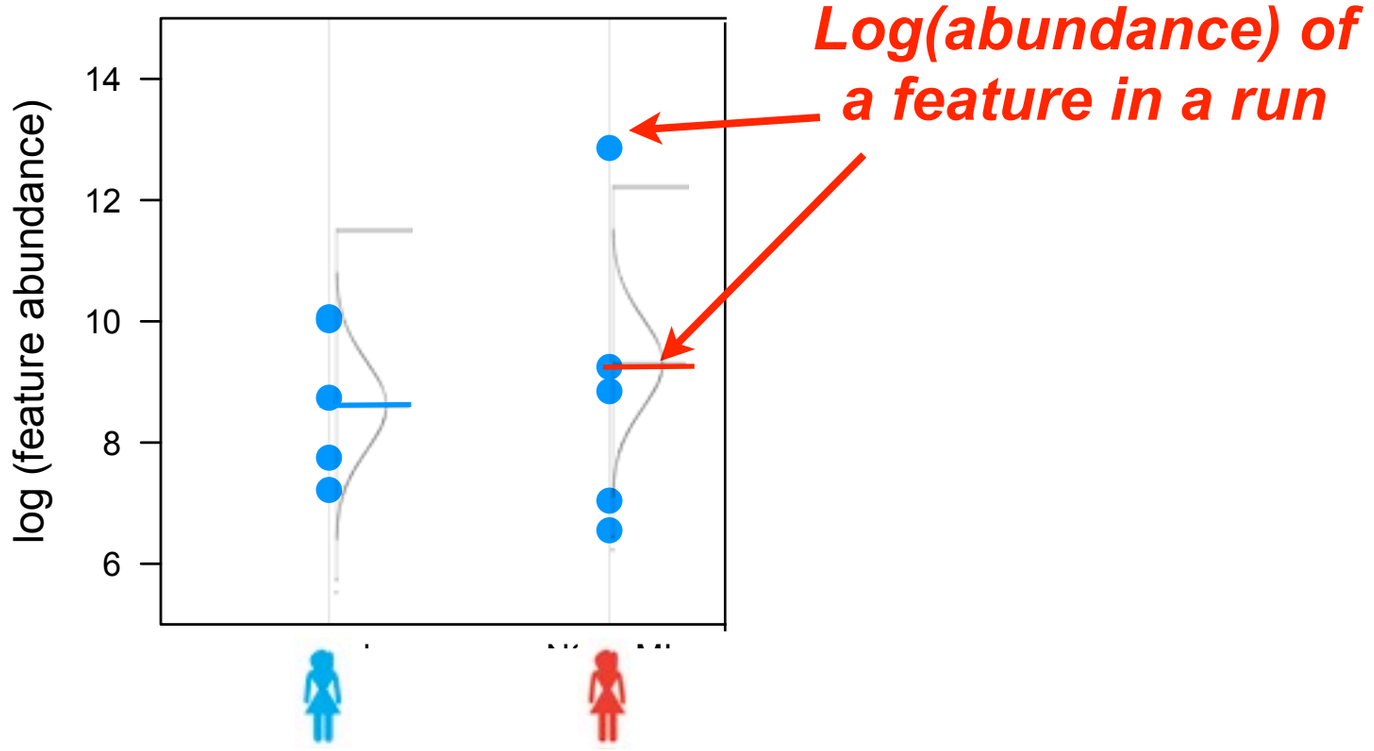
Steps of statistical significance analysis

- Define the analysis protocol
  - Type of analysis and comparisons of interest
  - Scope of conclusions
  - Model type
- Normalization and quality control
- Model-based analysis
  - Specify the model
  - Perform-based comparisons
  - Control for multiple testing
- Use the experiment to gain insight into future studies
  - Compare strategies of future resource allocation
  - Calculate sample size of a future similar experiment

Finding differentially abundant proteins

Simple example: one protein, one feature per protein, label-free





- Two inter-dependent approaches
  - Decision-based
    - For each protein, decide whether it is differentially abundant
  - Ranking-based
    - Rank the proteins for evidence of differential abundance
- Report a measure of confidence; account for # of proteins

False positive rate α

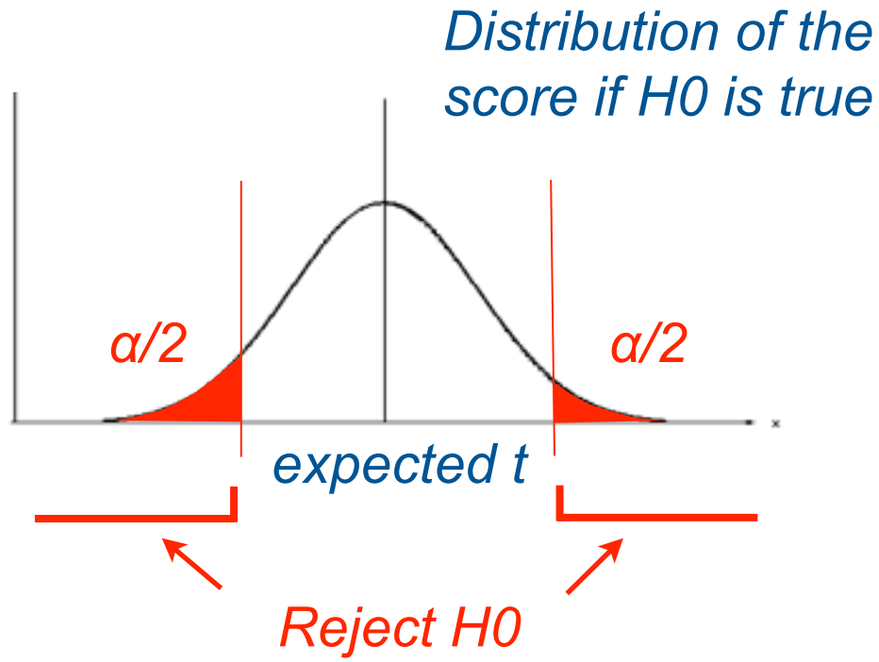
Simple example: one protein, one feature per protein, label-free

H0: 'status quo', no change in abundance, $\hat{G}_1 - \hat{G}_0 = 0$

Ha: change in abundance, $\hat{G}_1 - \hat{G}_0 \neq 0$

$$\text{observed t} = \frac{\hat{G}_1 - \hat{G}_0}{\sqrt{\text{Estimate of variation}}}$$

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



- False positive rate
  - *Property of the decision rule*
  - If
    - H0 is true
    - we infinitely measure the same protein
    - use the same score cutoff
  - FPR is the average proportion of false rejections = α

Output: decision for the protein differentially abundant or not)

P-value

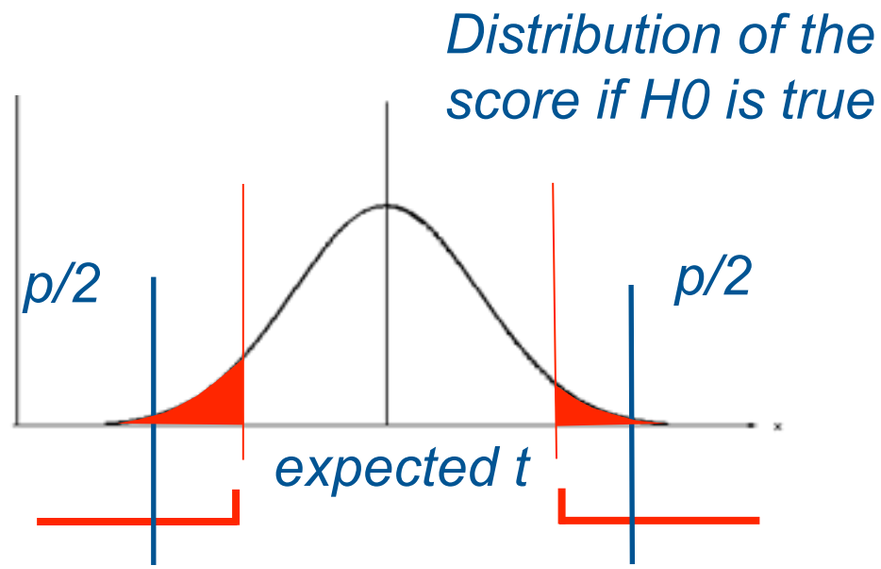
Simple example: one protein, one feature per protein, label-free

H0: 'status quo', no change in abundance, $\hat{G}_1 - \hat{G}_0 = 0$

Ha: change in abundance, $\hat{G}_1 - \hat{G}_0 \neq 0$

$$\text{observed t} = \frac{\hat{G}_1 - \hat{G}_0}{\sqrt{\text{Estimate of variation}}}$$

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



- P-value
  - *Property of the measurement*
  - If
    - H0 is true
    - we infinitely measure the same protein
  - P-value is the average proportion of scores more extreme than ***t***
  - P-value is the lowest α that rejects H0

Output: evidence in favor of differential abundance

More complex models lead to a similar procedure

Example: label-free rat diet dataset

Quantity of interest:

$$H_0 : L = \bar{\mu}_{\text{high}\cdot} - \bar{\mu}_{\text{low}\cdot} = 0$$

Model-based estimate and test statistic:

$$\hat{L} = \hat{C}_{\text{high}} + \frac{1}{I}\sum_{i=1}^{I}\widehat{(F \times C)}_{i,\,\text{high}} \text{ - } \hat{C}_{\text{low}} - \frac{1}{I}\sum_{i=1}^{I}\widehat{(F \times C)}_{i,\,\text{low}}$$

$$t = \frac{\hat{L}}{SE\{\hat{L}\}} \sim \text{Student distribution}$$

In balanced datasets:

$$\hat{L} = \bar{Y}_{\cdot\text{high}\cdot\cdot} - \bar{Y}_{\cdot\text{low}\cdot\cdot}$$

$$t = \frac{\hat{L}}{\sqrt{\frac{2}{IKL}\hat{\sigma}^2_{Error}}} \sim \text{Student}_{IJK(L-1)+(I-1)J(K-1)} \text{ distribution}$$

# of conditions, features, biol and tech reps

A similar signal-to-noise ratio and a similar student distribution

MSstats2 calculates this automatically

Need to account for testing multiple proteins

What happens if we simultaneously test 2 proteins?

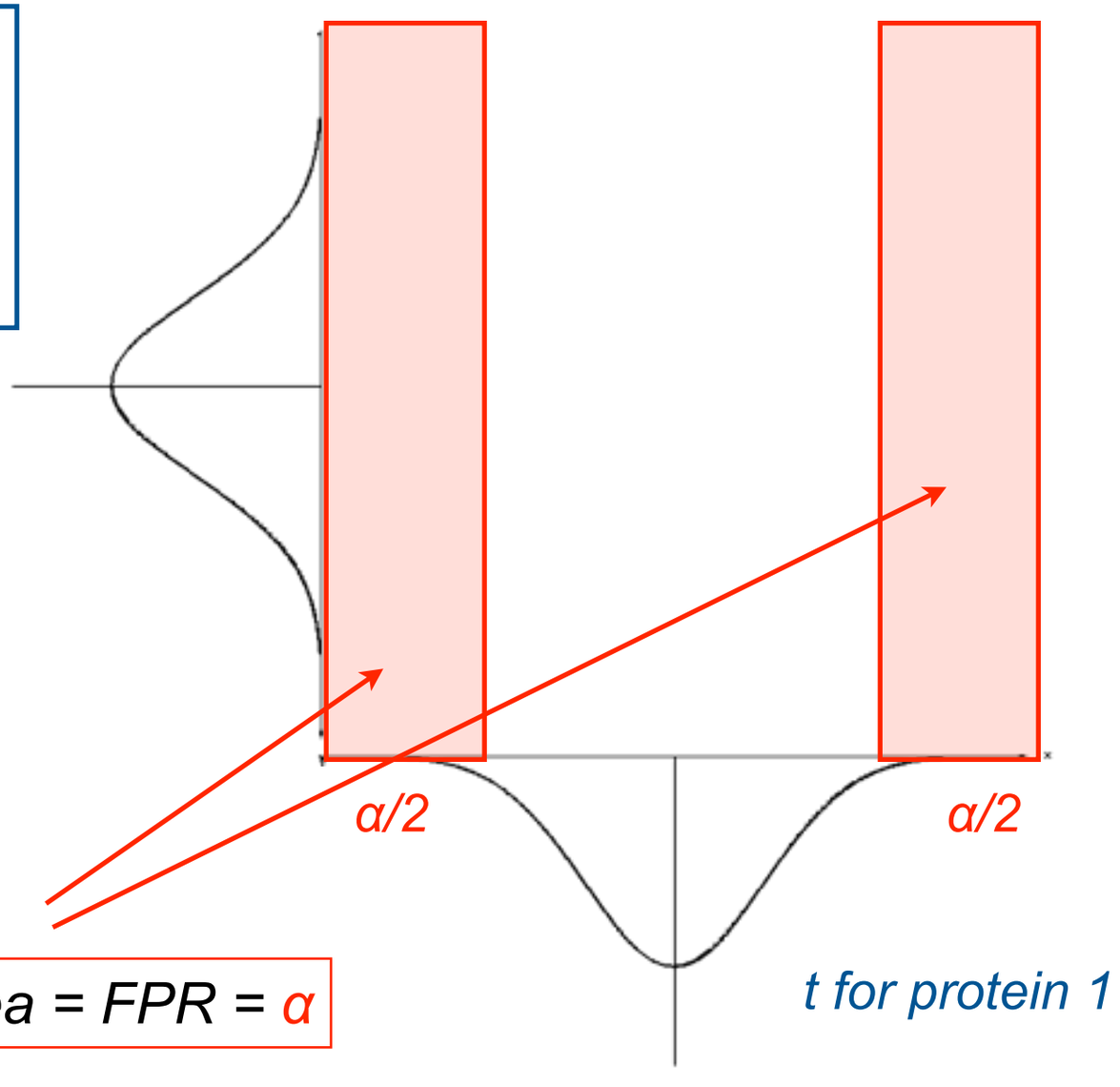
For each protein:

H0: 'status quo', no change in abundance, $\hat{G}_1 - \hat{G}_0 = 0$

Ha: change in abundance, $\hat{G}_1 - \hat{G}_0 \neq 0$

$$\text{observed t} = \frac{\hat{G}_1 - \hat{G}_0}{\sqrt{\text{Estimate of variation}}}$$

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



The area = FPR = α

Need to account for testing multiple proteins

What happens if we simultaneously test 2 proteins?

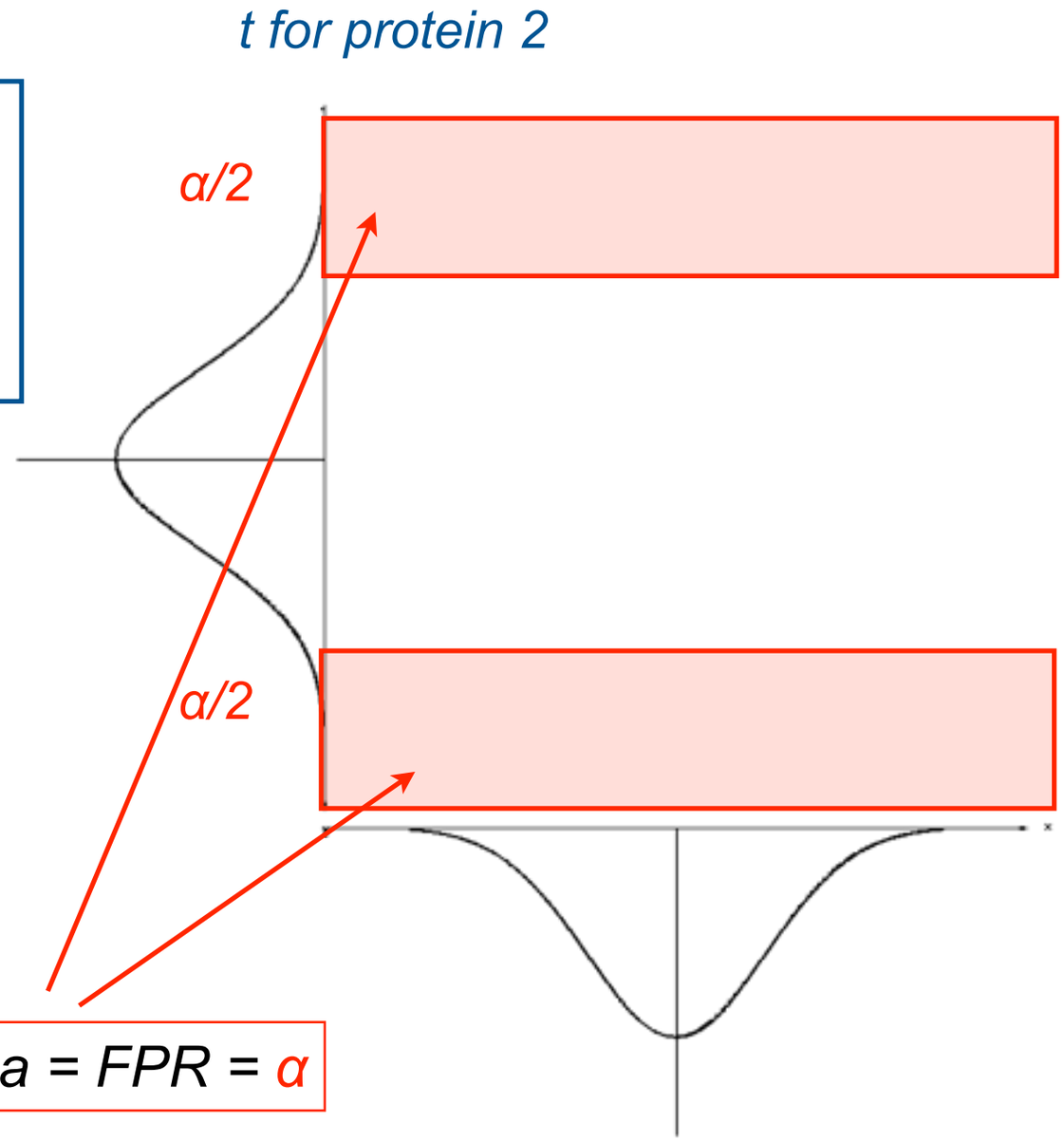
For each protein:

H0: 'status quo', no change in abundance, $\hat{G}_1 - \hat{G}_0 = 0$

Ha: change in abundance, $\hat{G}_1 - \hat{G}_0 \neq 0$

$$\text{observed t} = \frac{\hat{G}_1 - \hat{G}_0}{\sqrt{\text{Estimate of variation}}}$$

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



Need to account for testing multiple proteins

What happens if we simultaneously test 2 proteins?

For each protein:

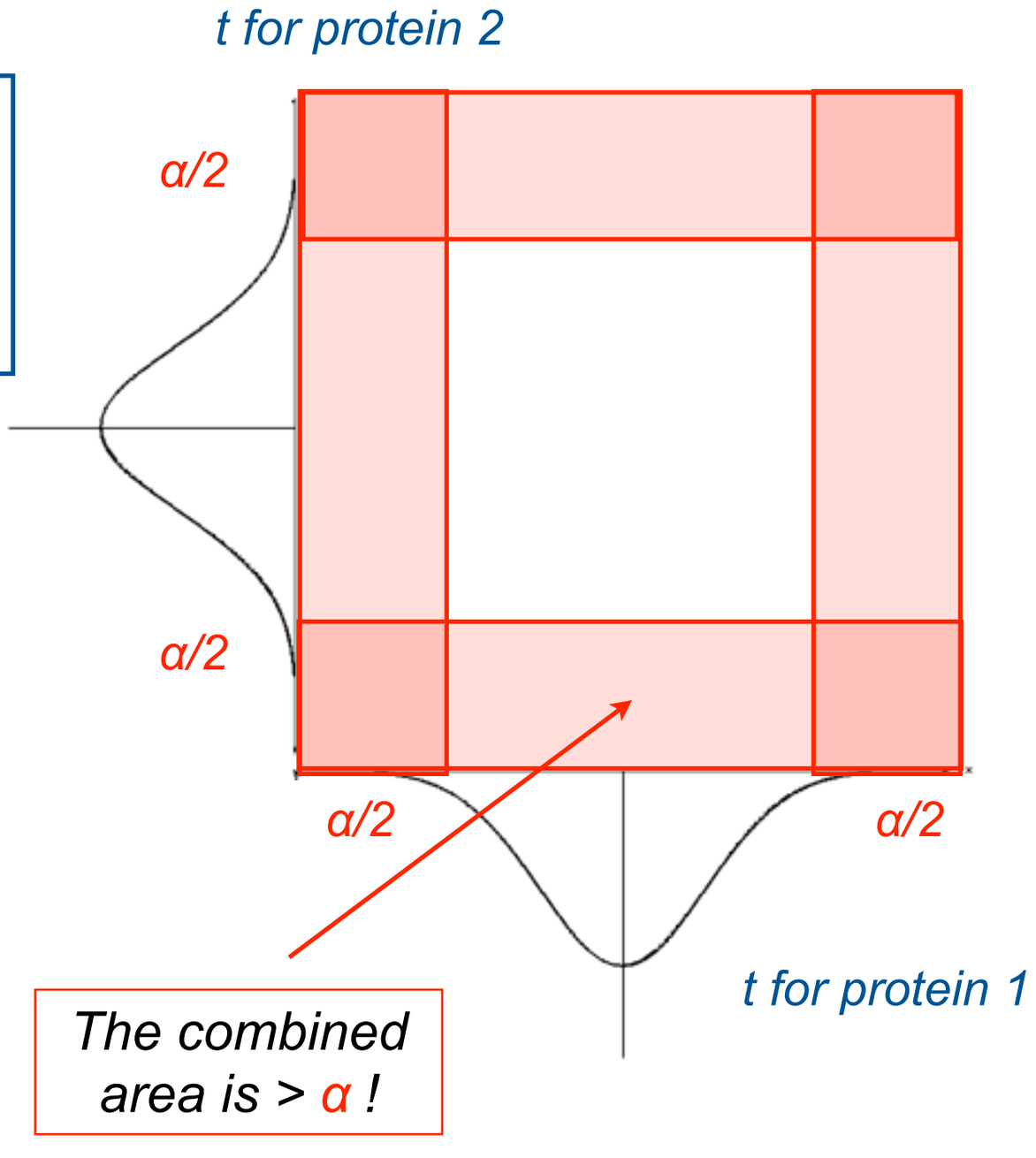
H0: 'status quo', no change in abundance, $\hat{G}_1 - \hat{G}_0 = 0$

Ha: change in abundance, $\hat{G}_1 - \hat{G}_0 \neq 0$

$$\text{observed } t = \frac{\hat{G}_1 - \hat{G}_0}{\sqrt{\text{Estimate of variation}}}$$

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

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



Differentially abundant features: False Discovery Rate (FDR)

The outcome of testing H0 for m features

	# 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)
  - *Property of the testing procedure*
  - If
    - we collect an infinite number of measurements on the same group of proteins
  - FDR is the average proportion of false discoveries in the list of proteins with detected difference

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

Use p-values to control FDR

Vary the threshold while comparing decreasing p-values

- Change decision rule (property of the procedure)

Order	least significant			⟹				most significant
$p-$value	$p_{(m)}$	$p_{(m-1)}$	$\cdots$	$p_{(k+1)}$	$p_{(k)}$	$p_{(k-1)}$	$\cdots$	$p_{(1)}$
Compare to	$\frac{m}{m}q$	$\frac{m-1}{m}q$	$\ldots$	$\frac{k+1}{m}q$	$\frac{k}{m}q$	$\frac{k-1}{m}q$	$\ldots$	$\frac{1}{m}q$
Is $p \leq q$?	No	No	$\ldots$	No	Yes			
Is significant?	No	No	$\ldots$	No	Yes	Yes	Yes	Yes

- Adjust the p-value (property of the test)

$$\tilde{p}_j = \min_{k=j,\ldots,m} \left\{ \min \left(\frac{m}{k} p_{(k)}, 1 \right) \right\}$$

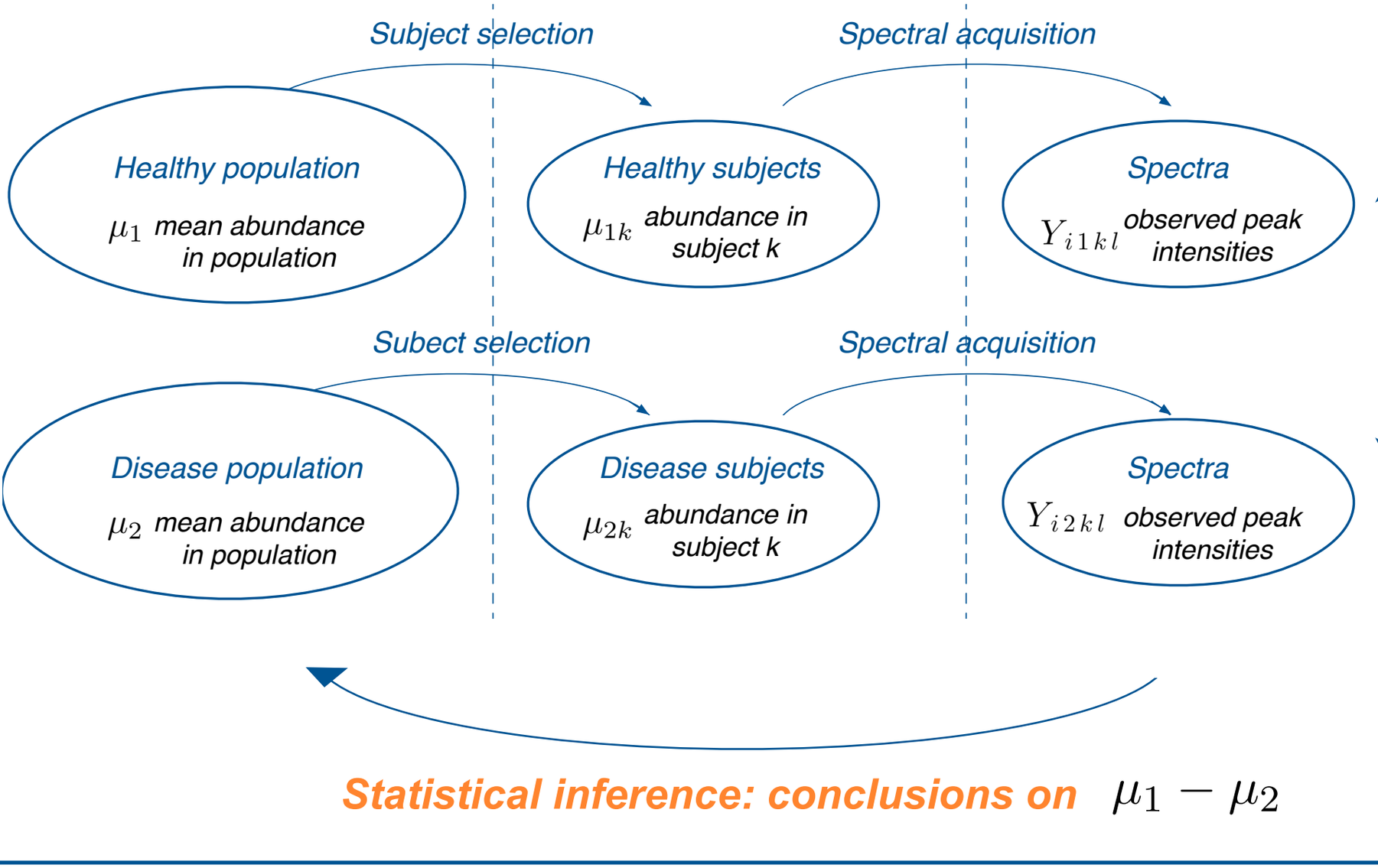
  - adjusted p-value cut-off corresponds to the FDR
  - adjusted p-value (obtained with an alternative procedure) is sometimes referred to as q-value

Benjamini and Hochberg, JRSS B, 57, p. 289, 1995

Steps of statistical significance analysis

- Define the analysis protocol
  - Type of analysis and comparisons of interest
  - Scope of conclusions
  - Model type
- Normalization and quality control
- Model-based analysis
  - Specify the model
  - Perform-based comparisons
  - Control for multiple testing
- Use the experiment to gain insight into future studies
  - Compare strategies of future resource allocation
  - Calculate sample size of a future similar experiment

Recall: H is how a statistician would use the data to perform the comparisons

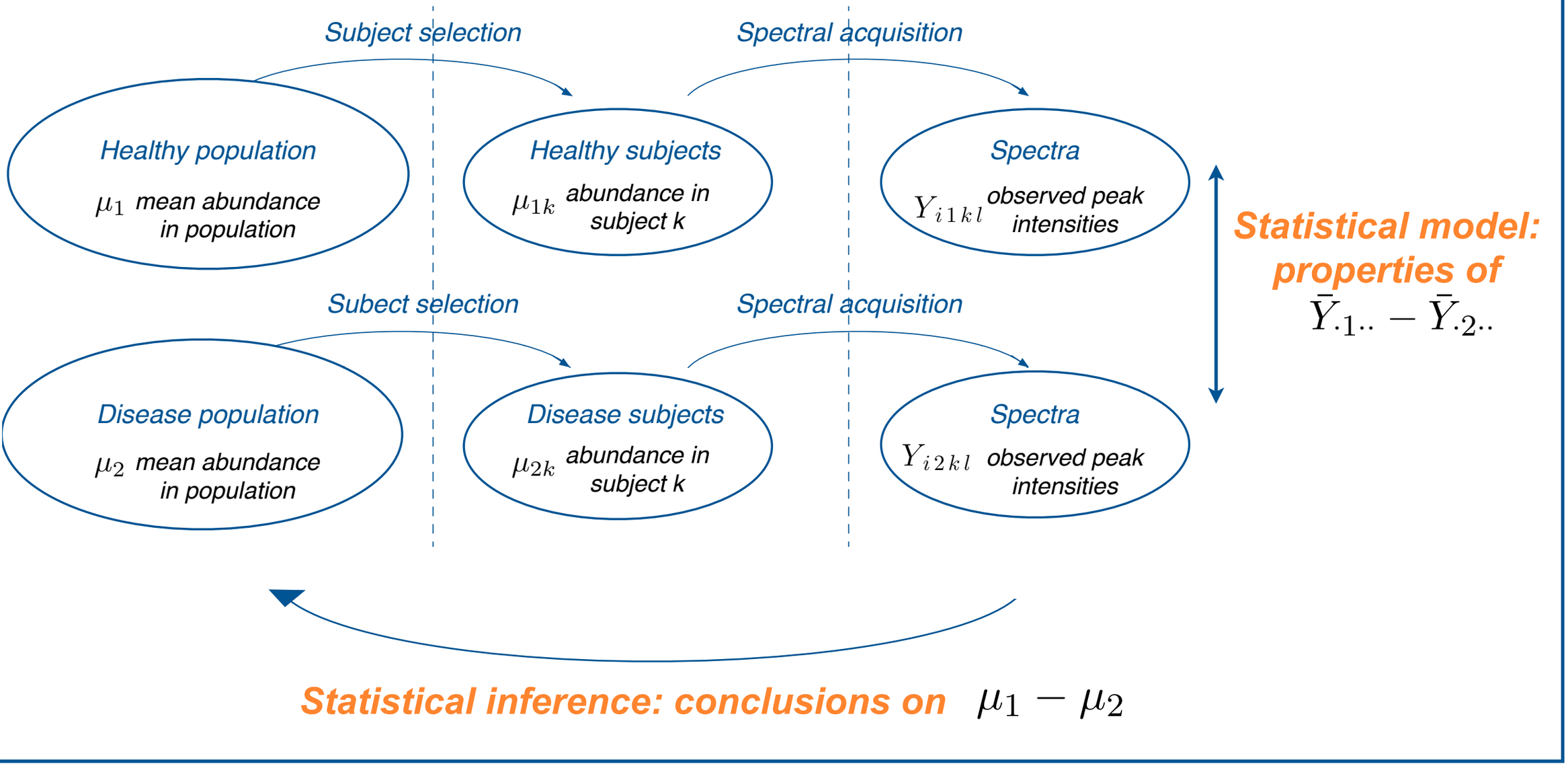


Potential dangers:

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

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

Recall: Here is how a statistician would use the data to perform the comparisons



Potential dangers:

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

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

Focus of resource allocation

Linear mixed effects models are required to evaluate the importance of various replicate types

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	+	$\text{Indiv(Group)}_{j(i)} \sim N\left(0, \sigma^2_{Indiv}\right)$	+	$\text{Prep(Indiv)}_{k(ij)} \sim N\left(0, \sigma^2_{Prep}\right)$	+	$\text{Error}_{l(ijk)} \sim N\left(0, \sigma^2_{Error}\right)$

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

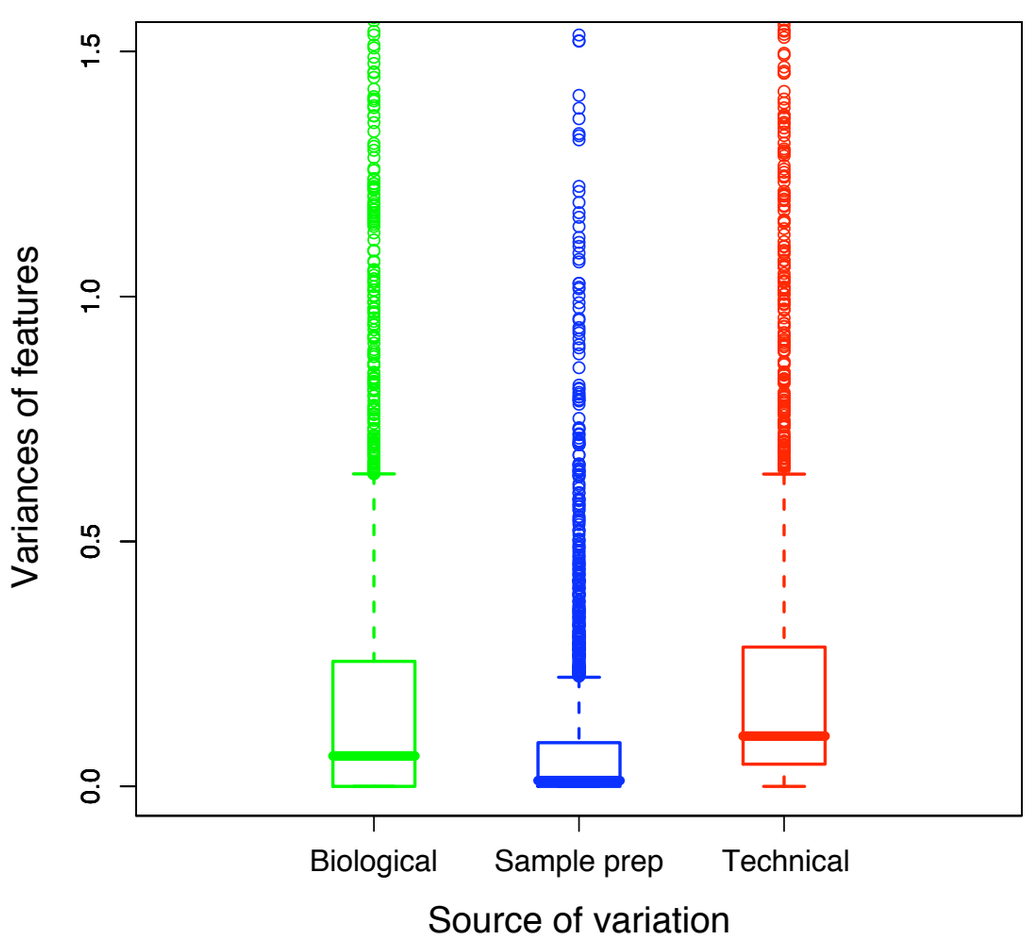
I: # individuals per disease group

J: # sample preps

K: # replicate runs

A pilot experiment

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



Linear mixed effects models are required to evaluate the importance of various replicate types

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	+	$\text{Indiv(Group)}_{j(i)} \sim N\left(0, \sigma^2_{\text{Indiv}}\right)$	+	$\text{Prep(Indiv)}_{k(ij)} \sim N\left(0, \sigma^2_{\text{Prep}}\right)$	+	$\text{Error}_{l(ijk)} \sim N\left(0, \sigma^2_{\text{Error}}\right)$

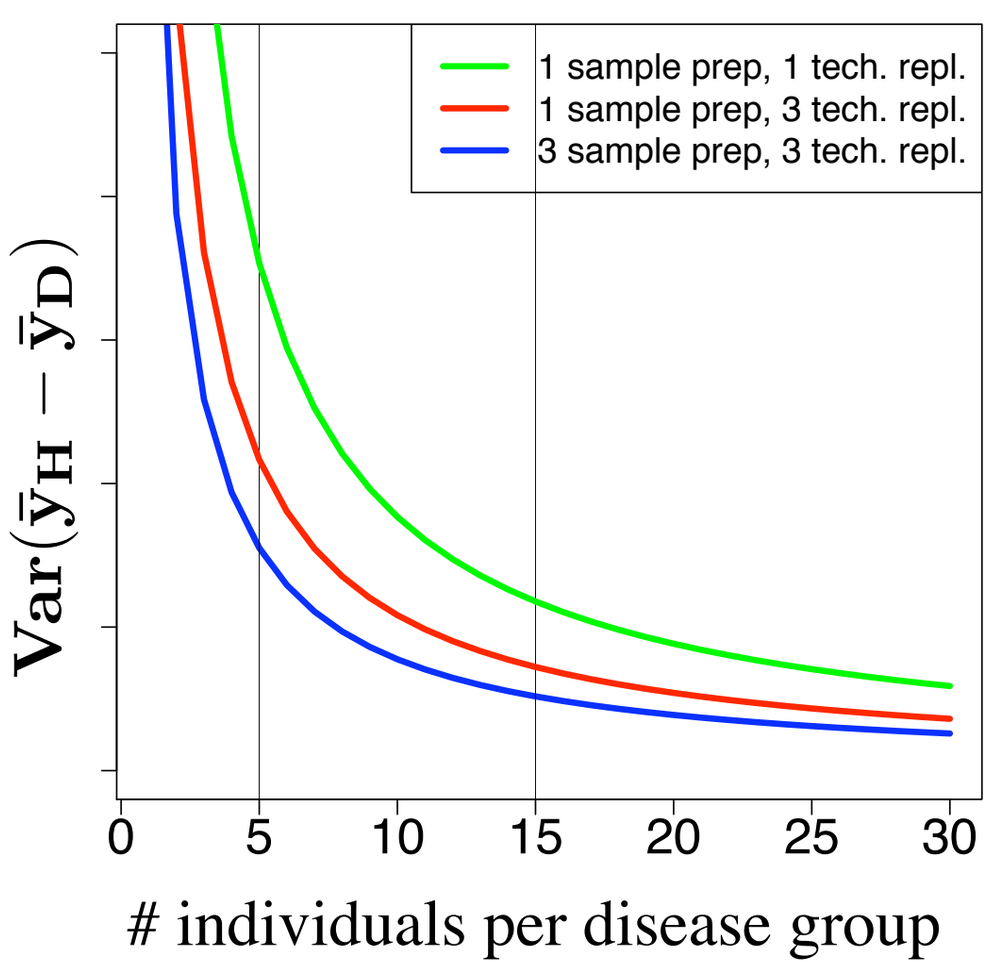
$$\text{Var}(\bar{y}_H - \bar{y}_D) = 2\left(\frac{\sigma^2_{\text{Indiv}}}{I} + \frac{\sigma^2_{\text{Prep}}}{IJ} + \frac{\sigma^2_{\text{Error}}}{IJK}\right)$$

I: # individuals per disease group

J: # sample preps

K: # replicate runs

Conclusion:
Maximize the number of biological replicates



Linear mixed effects models are required to evaluate the value of blocking (e.g. plate or day)

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	+	$\text{Block}_k \sim N\left(0, \sigma^2_{Block}\right)$	+	$\text{Indiv(Group)}_{j(i)} \sim N\left(0, \sigma^2_{Indiv}\right)$	+	$\text{Error}_{l(ijk)} \sim N\left(0, \sigma^2_{Error}\right)$

A completely randomized design

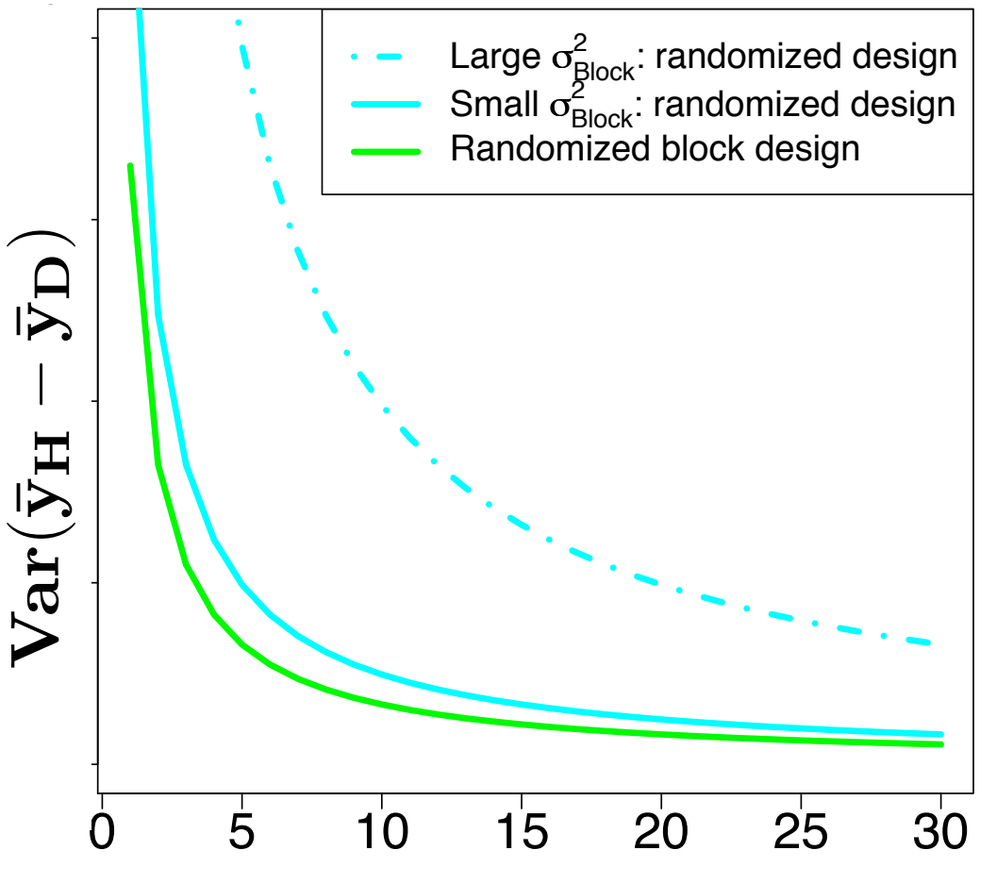
I: # individuals per disease group

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

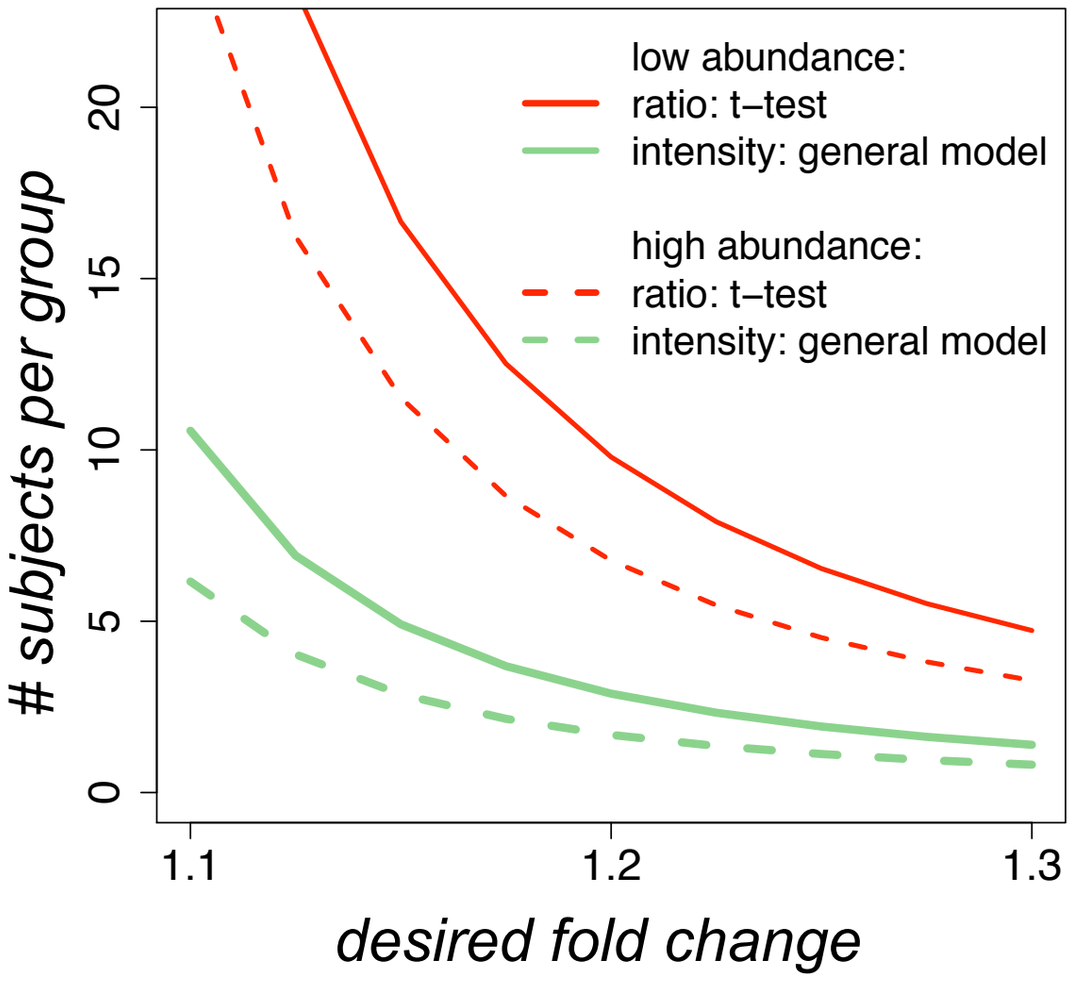
A block-randomized design

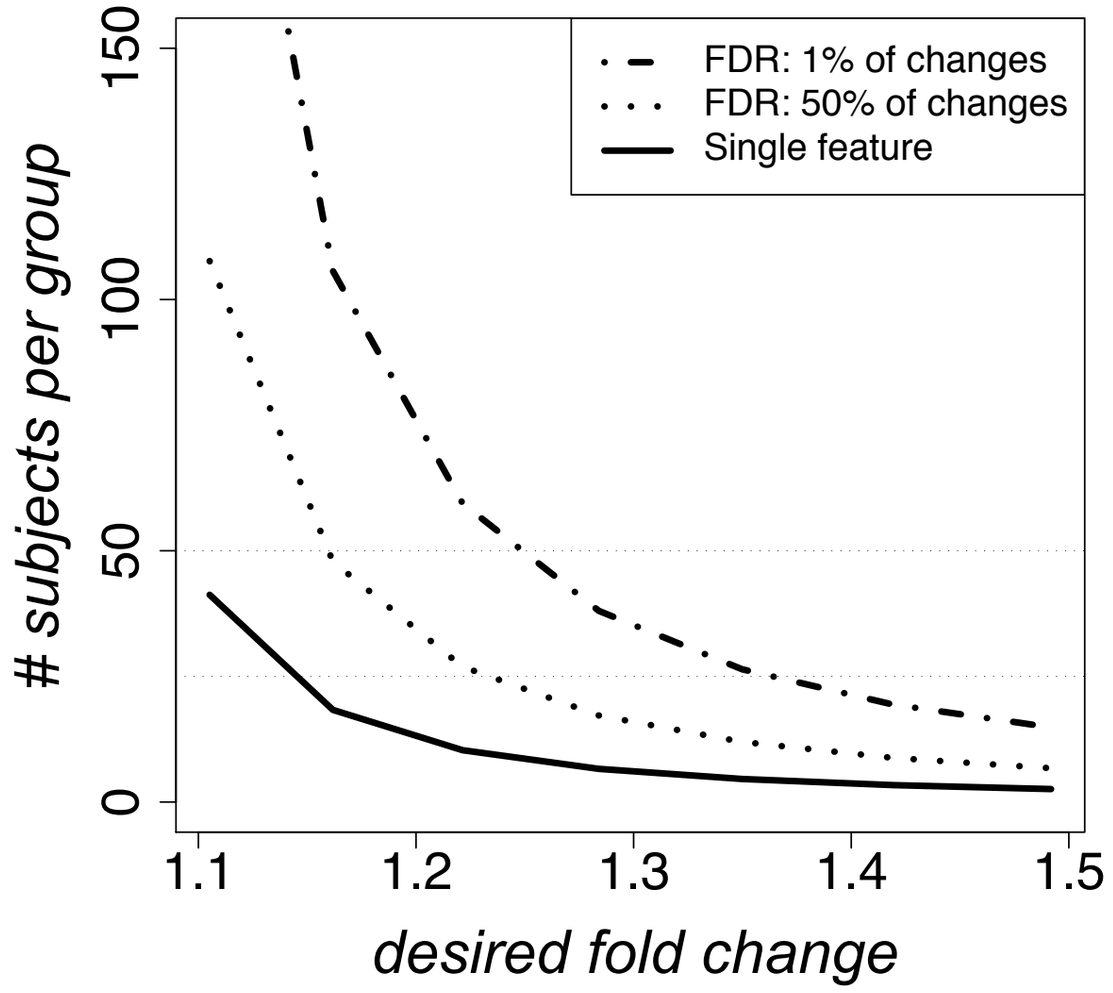
$$\text{Var}(\bar{y}_H - \bar{y}_D) = 2\left(\frac{\sigma^2_{Indiv} + \sigma^2_{Error}}{I}\right)$$

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



Linear mixed effects models are required to calculate the sample size





Oberg and Vitek, *JPR*, 2009

- Need prior information to plan sample size
  - statistical model for data analysis
  - estimates of sources of variation
  - expected proportion of differentially abundant proteins

A lot must be known in advance to calculate the sample size

Need to know in advance:

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

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

β - probability of a true positive discovery

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

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

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

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

Then calculate:

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

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

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

Then solve for the number of replicates

Alternatively, fix sample size and solve for one other number

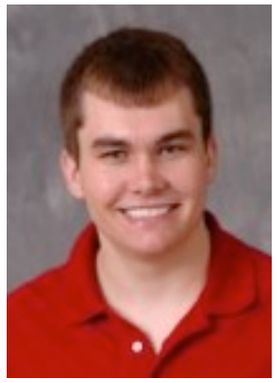
Open-source R-based software for protein quantification

www.stat.purdue.edu/~ovitek

Veavi Chang
Purdue

Meena Choi
Purdue

Tim Clough
Purdue



Statistical protein quantification
Shotgun & SRM
Label-based & label-free

- Recognizes experimental designs
  - time course/group comparison
- Data visualization and quality control
  - data plots, model-checking plots
- Model fitting
  - unequal variance, pooling interactions
- Model-based conclusions
  - group comparison & sample quantification
- Planning future experiments
  - number of replicates, peptides, transitions

Since Dec 2011:

- *285 unique visitors*
- *over 50 unique downloads*
- *over 50 mailing list members*

Now:
integration
with Skyline

Skyline

Concluding thoughts

- More sophisticated models lead to more accurate conclusions
  - It is worthwhile to invest time and effort
  - Software implementation facilitates the task
- More model flexibility means more analysis choices
  - Define the data analysis protocol before seeing the data
  - Do not change the protocol after seeing the data
- Utilize consistent computational tools to facilitate reporting, re-analysis and peer review
  - Skyline is great! Now with the statistical tools.
- Involve a statistician in all steps of planning and analysis!

References

- Skyline
  - B. MacLean et al. *Bioinformatics*, 26, p.966, 2010.
- Statistical analysis tools
  - *SRMstats:*
    C.-Y. Chang et al. *Molecular & Cellular Proteomics*, 2012.
  - *MSstats:*
    T. Clough et al. *BMC Bioinformatics*, 13 (Suppl. 16) 2012.
    T. Clough et al., *Methods in Molecular Biology,* 728, 2011
    T. Clough et al., *Journal of Proteome Research*, 8, p.5275, 2009
- General statistical methodology
  - L. Käll and O. Vitek. *PLoS Computational Biology*, 7, 2011
  - P. Radivojac and O. Vitek (Eds.) *BMC Bioinformatics*, 2012