INTRODUCTION TO MSSTATS

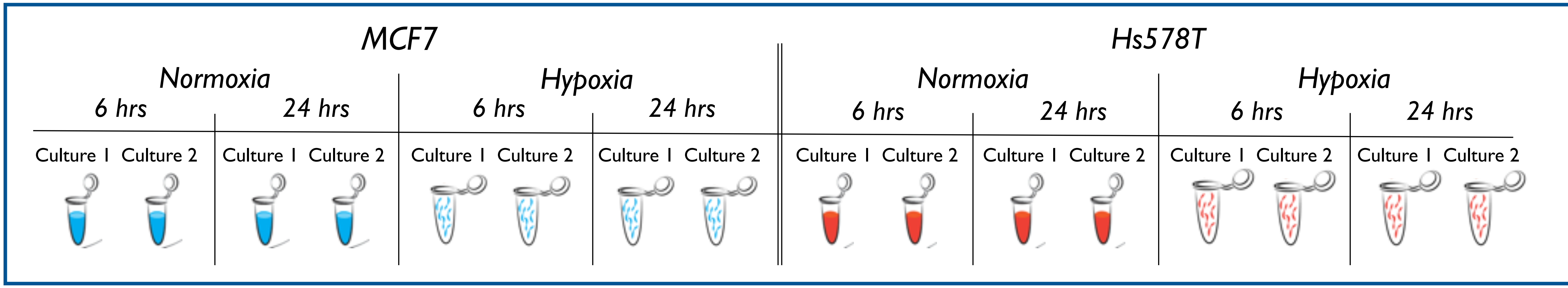
Olga Vitek

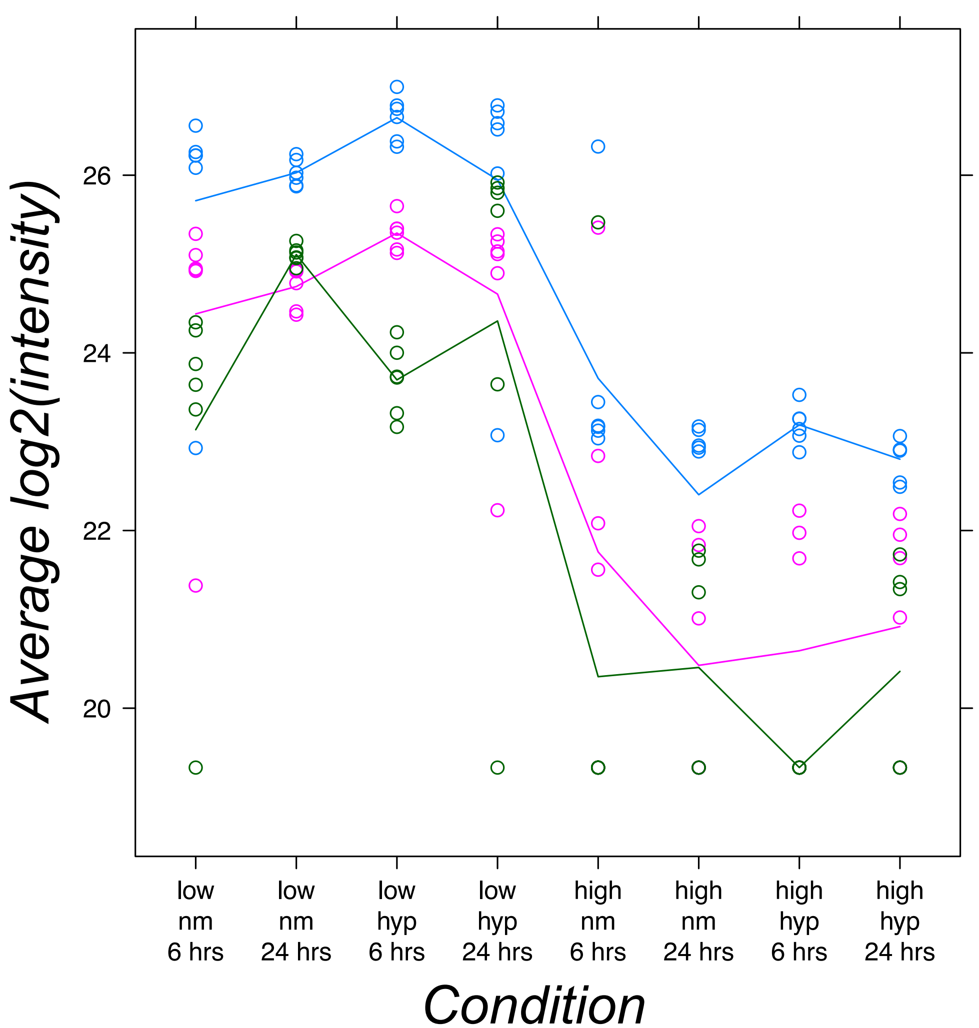
College of Science

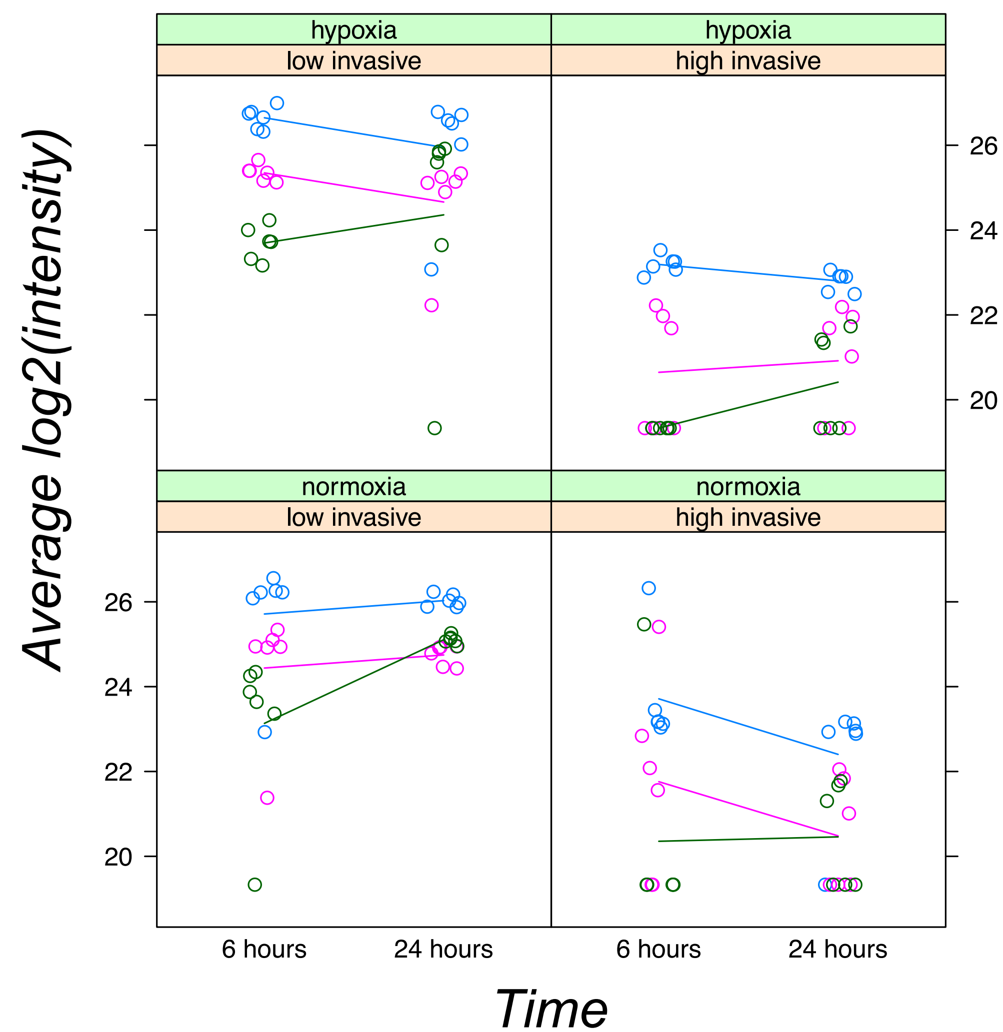
College of Computer and Information Science

Northeastern University

Example: A label-free experiment
Question: which proteins change in abundance?



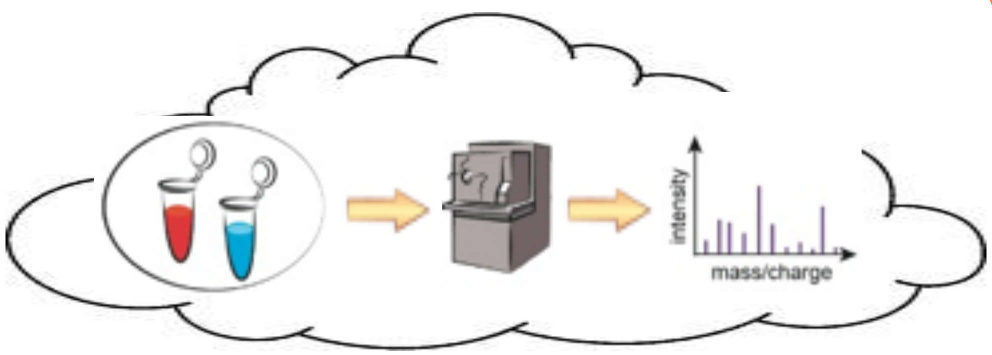




T. Clough et al. *BMC Bioinformatics., 2012*

A typical analysis workflow (also in MSstats)

Experimental design



QC and normalization

Statistical modeling

Model-based conclusions

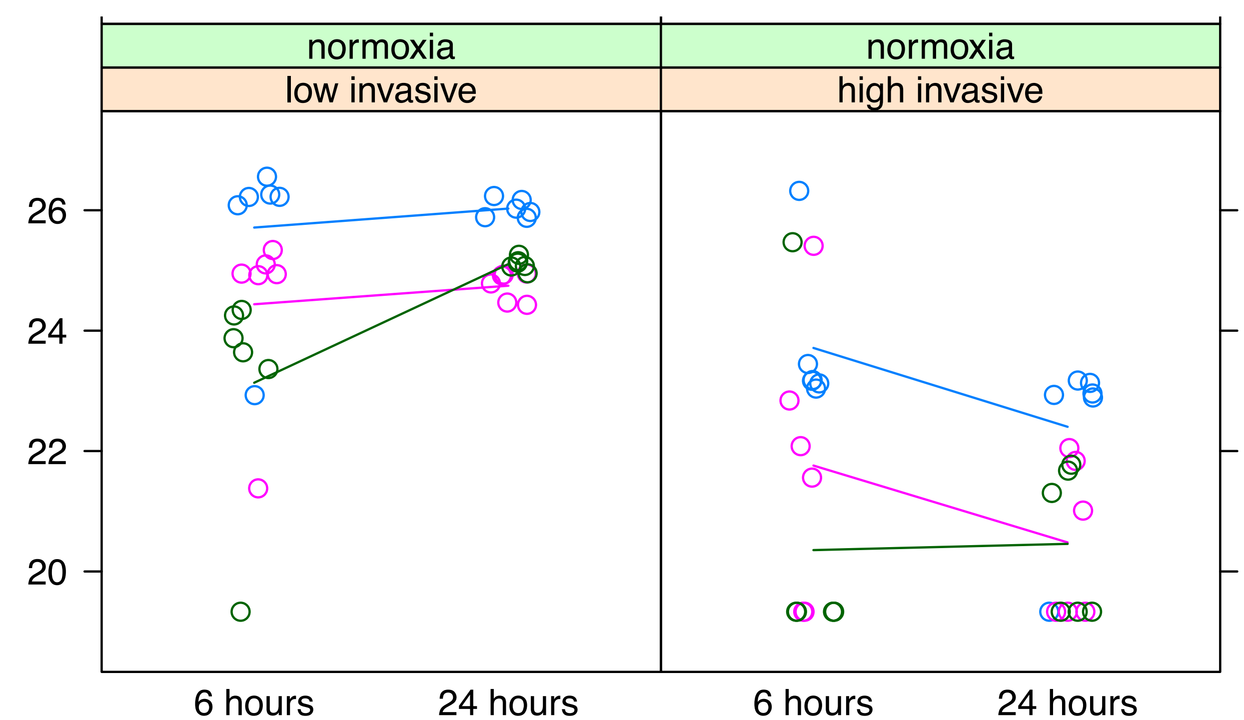
Experimental design

Stresses or conditions

- Time course
- Multiple groups
- Paired design

Problem statement

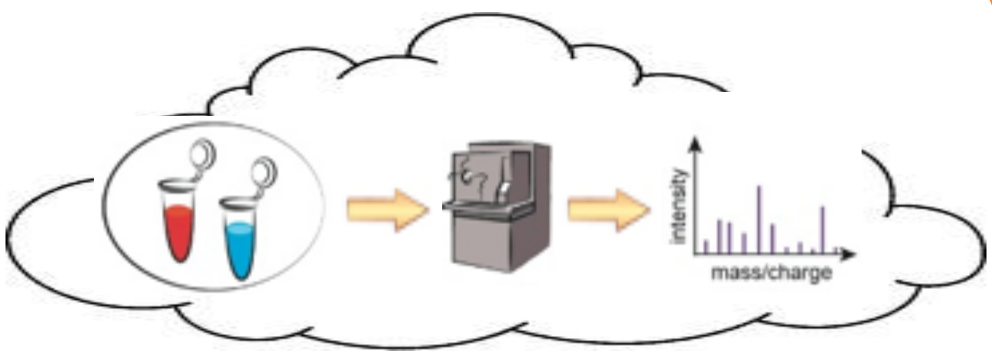
- Screening experiment
- Confirmation experiment



T. Clough et al. *BMC Bioinformatics*, 2012

A typical analysis workflow (also in MSstats)

Experimental design



QC and normalization

Statistical modeling

Model-based conclusions

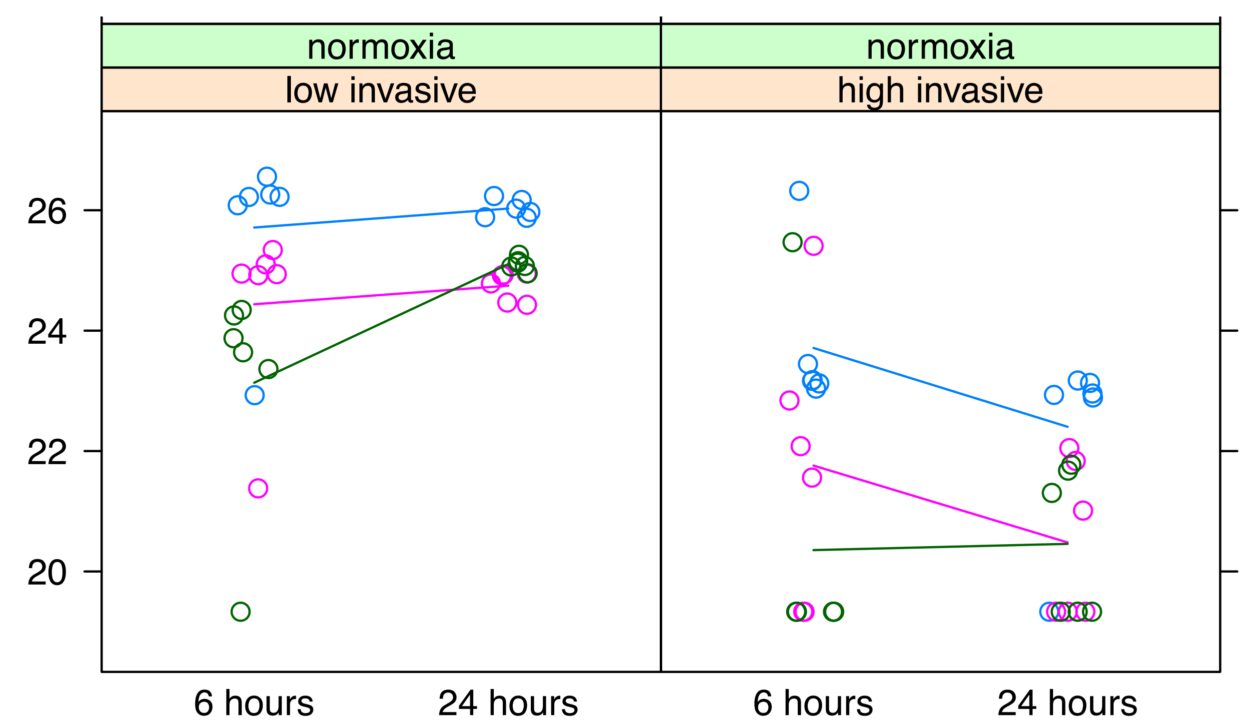
Experimental design

Stresses or conditions

- Time course
- Multiple groups
- Paired design

Problem statement

- Screening experiment
- Confirmation experiment

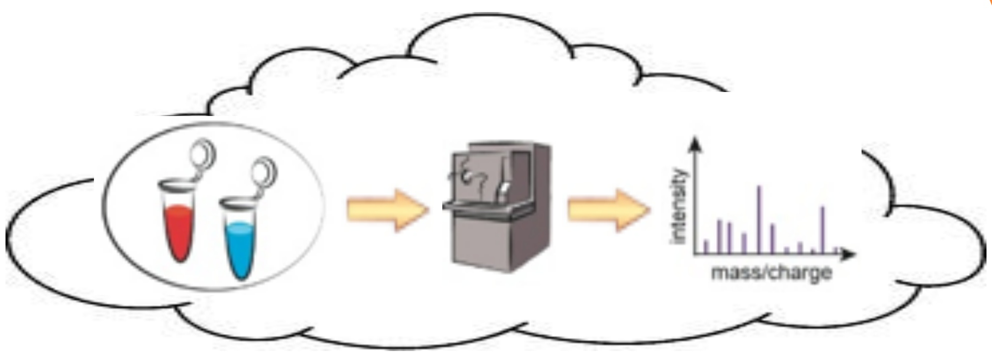


T. Clough et al. *BMC Bioinformatics*, 2012

A typical analysis workflow (also in MSstats)

Experimental design

↓



↓

QC and normalization

↓

Statistical modeling

↓

Model-based conclusions

↓

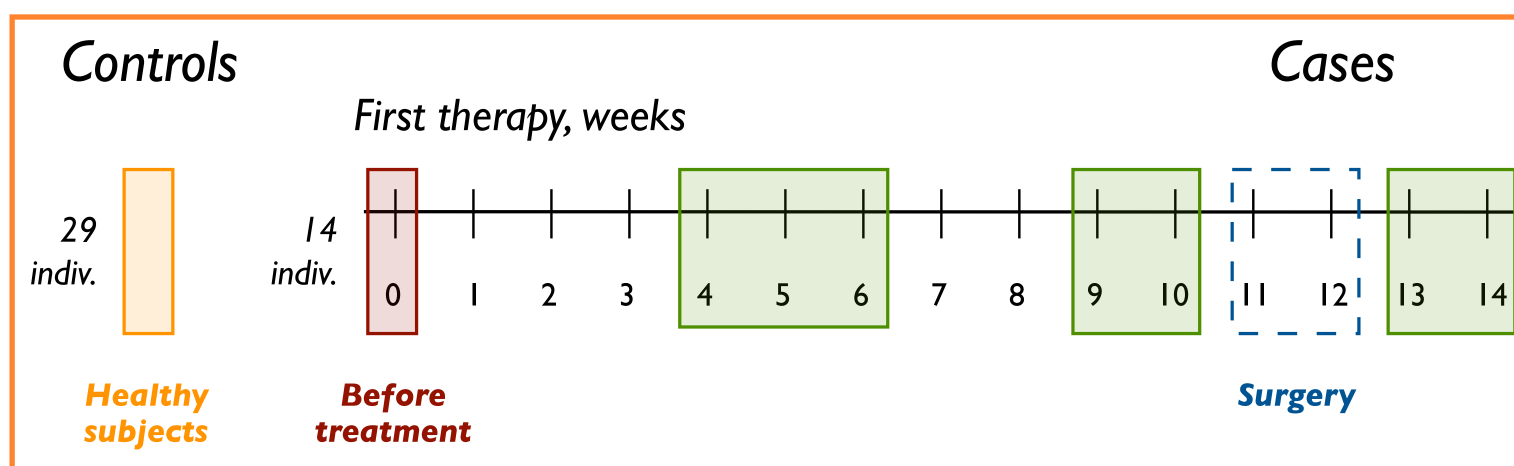
Experimental design

Stresses or conditions

- Time course
- Multiple groups
- Paired design

Problem statement

- Screening experiment
- Confirmation experiment

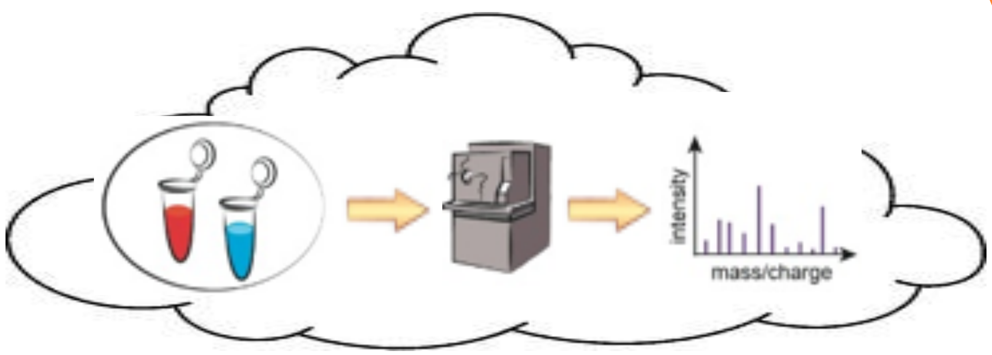


T. Clough et al. *BMC Bioinformatics*, 2012

A typical analysis workflow (also in MSstats)

Experimental design

↓



↓

QC and normalization

↓

Statistical modeling

↓

Model-based conclusions

↓

Experimental design

Stresses or conditions

- Time course
- Multiple groups
- Paired design

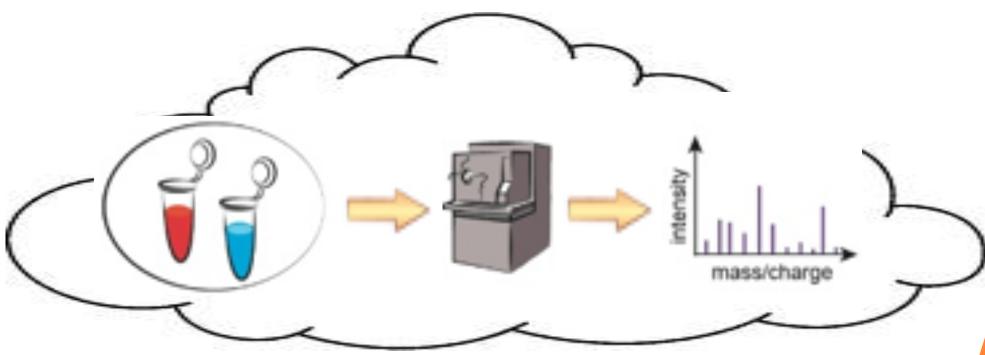
Problem statement

- Screening experiment
- Confirmation experiment

A typical analysis workflow (also in MSstats)

Experimental design

↓



↓

QC and normalization

↓

Statistical modeling

↓

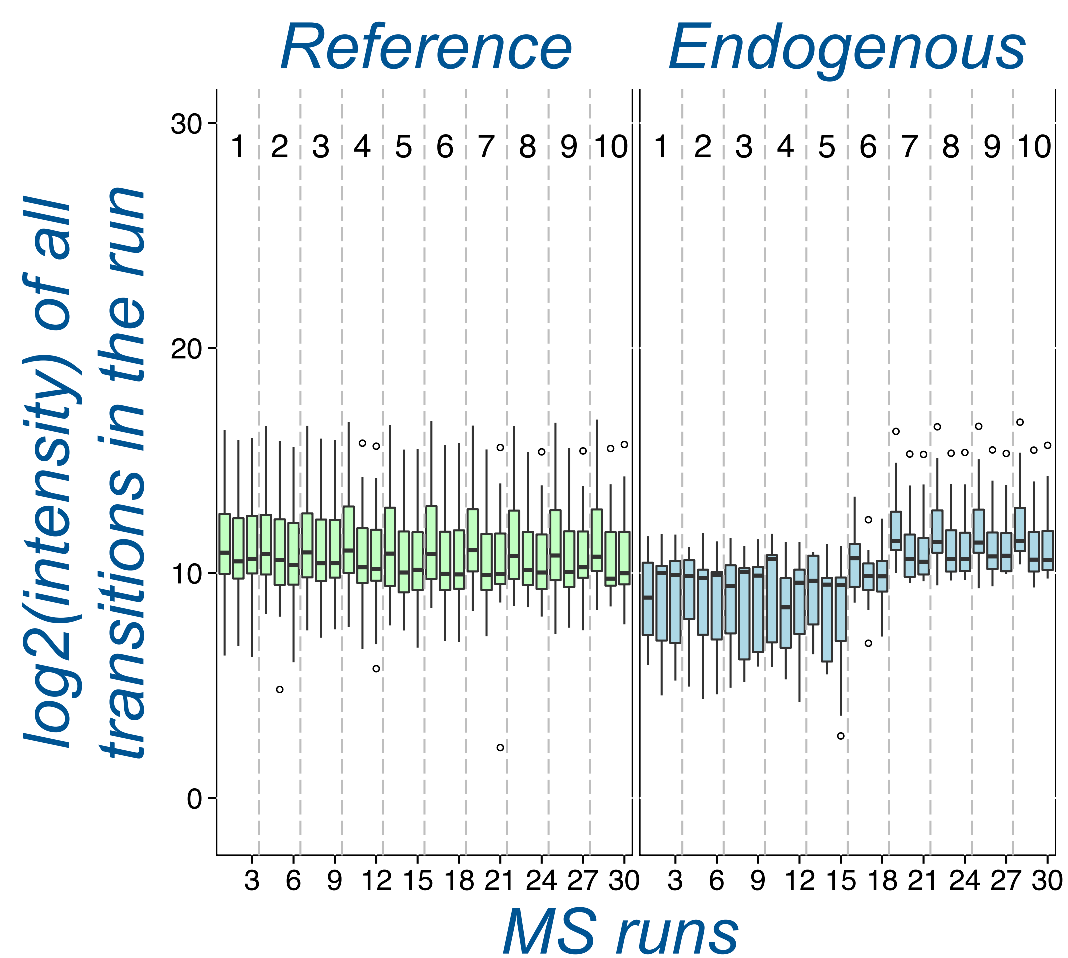
Model-based conclusions

↓

Experimental design

Data preparation

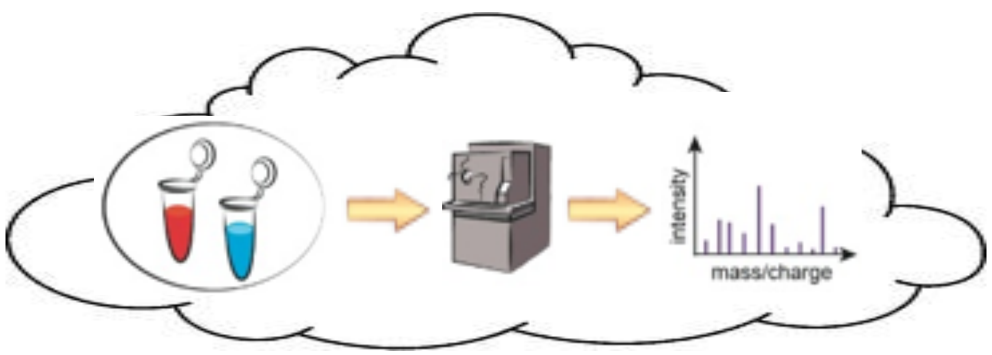
- Formatting
- Visualization
- Remove artifacts
- Normalization



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

A typical analysis workflow (also in MSstats)

Experimental design



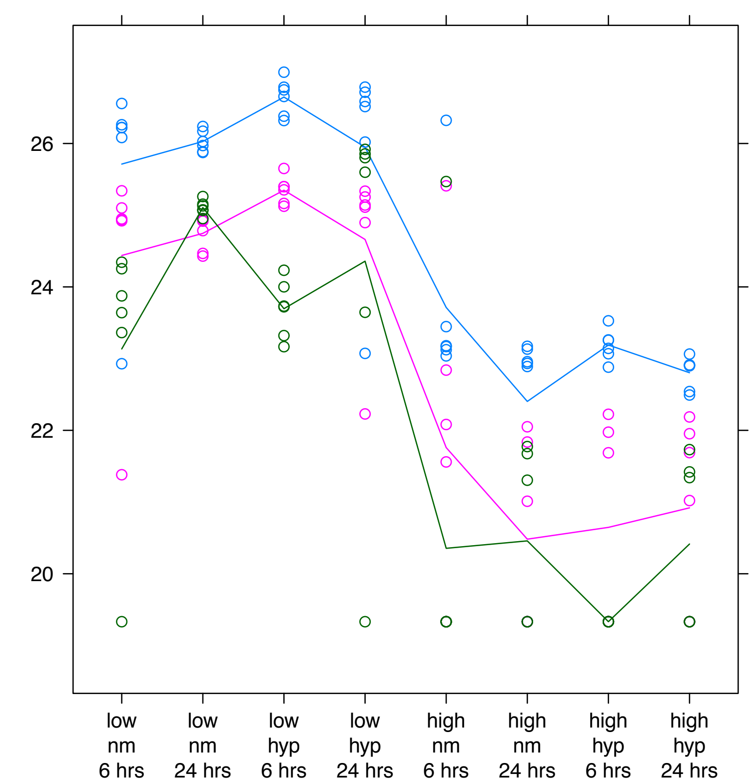
QC and normalization

Statistical modeling

Summarize all protein features in a statistical model

- Systematic variation
- Random variation

Verify the assumptions!

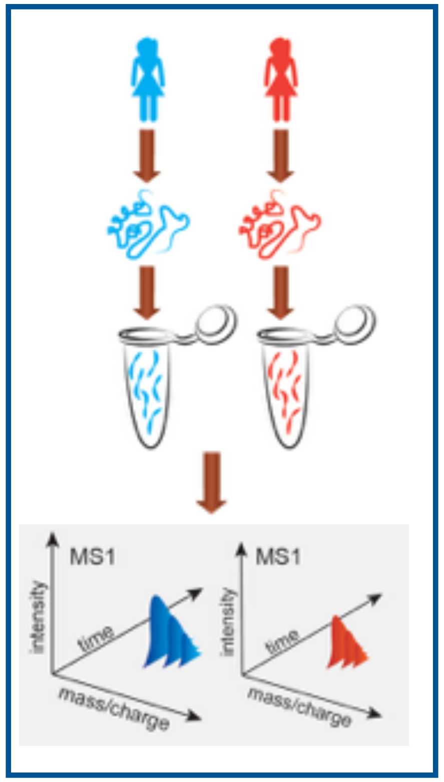


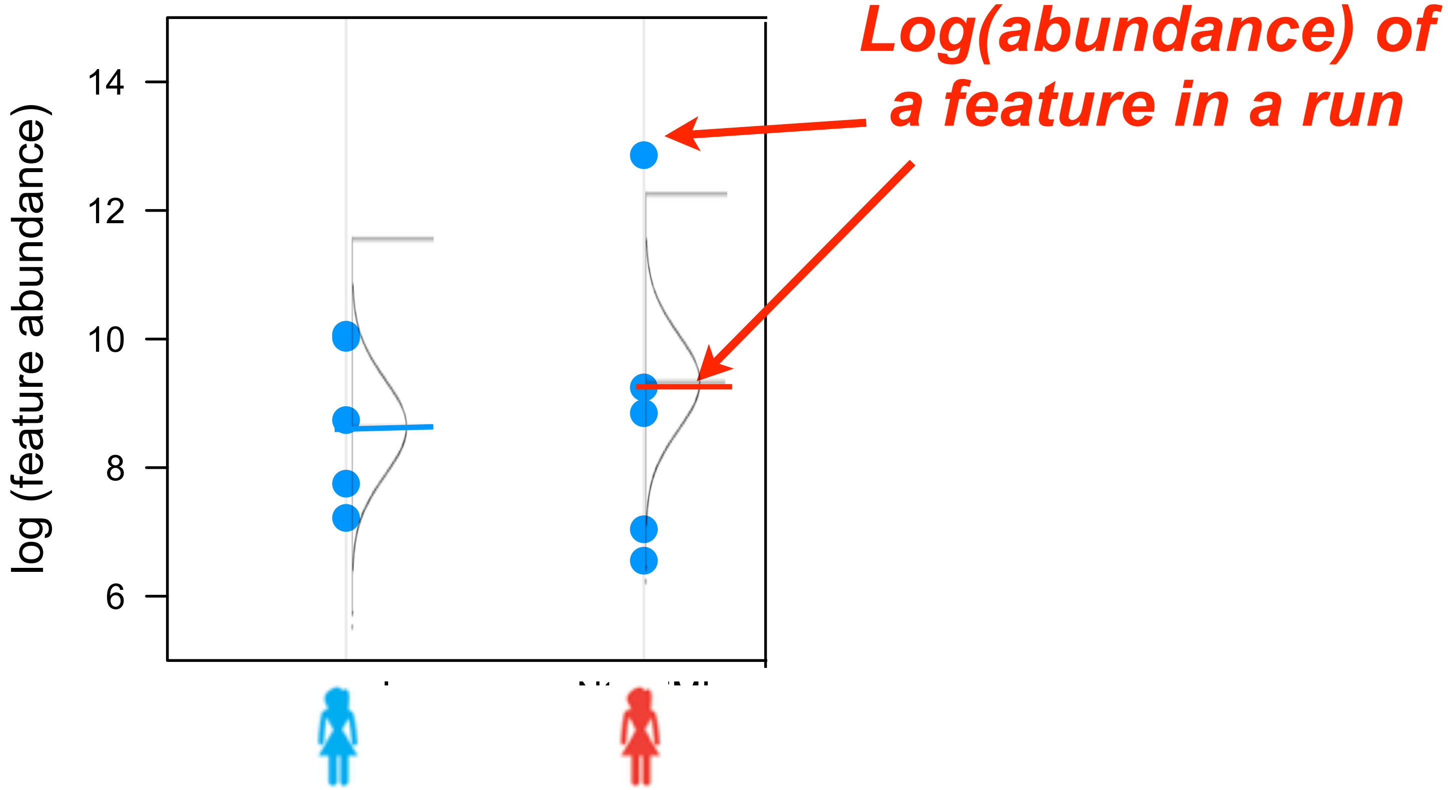
log(peak intensity) = Expected reference abundance + [Deviation from the reference due to: LC-MS feature + condition + feature × condition interaction + biol. replicate] + Random meas. error

$$y_{ijkl} = \mu_{111} + F_i + C_j + (F \times C)_{ij} + S(C)_{k(j)} + \varepsilon_{ijkl}$$

Finding differentially abundant proteins

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



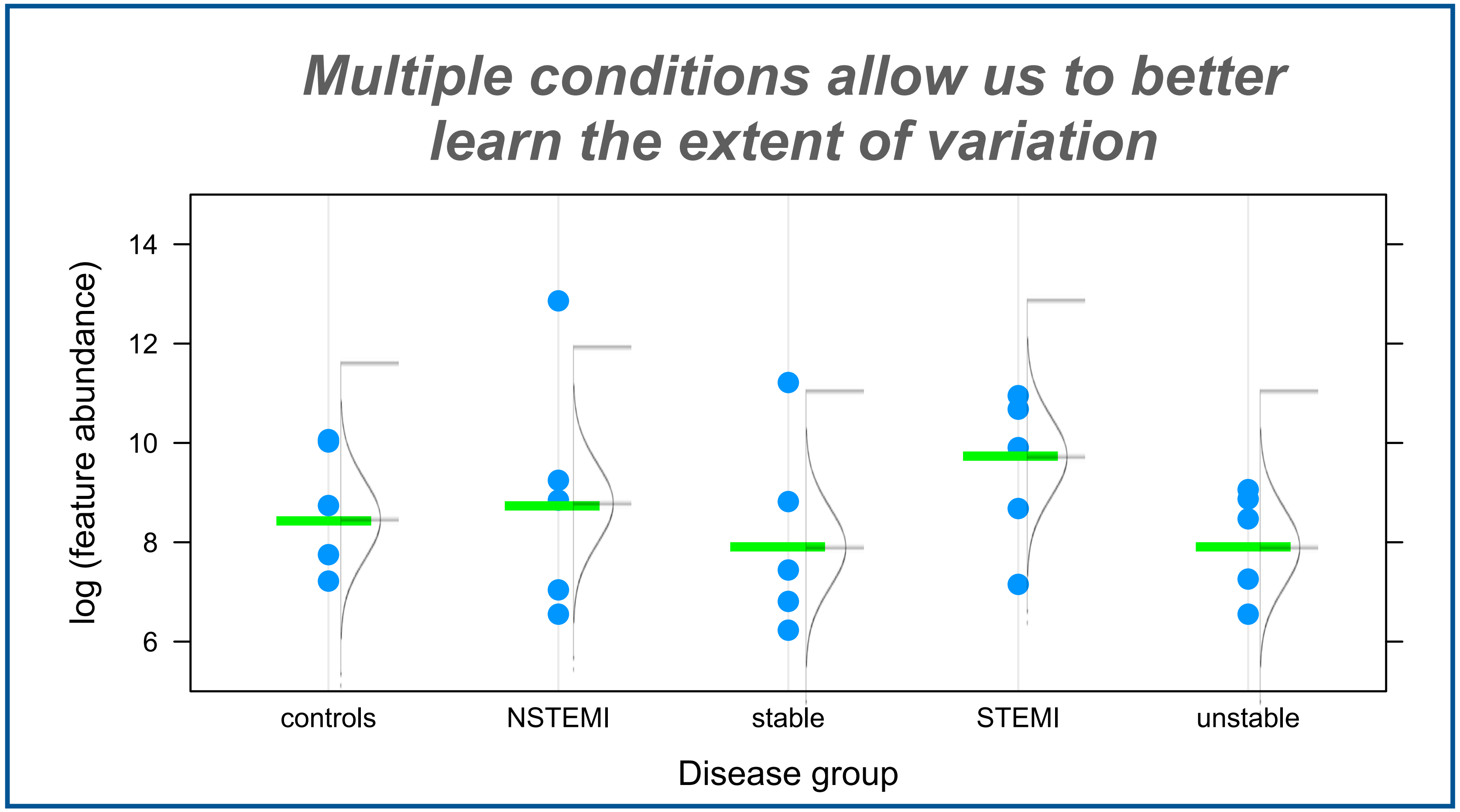


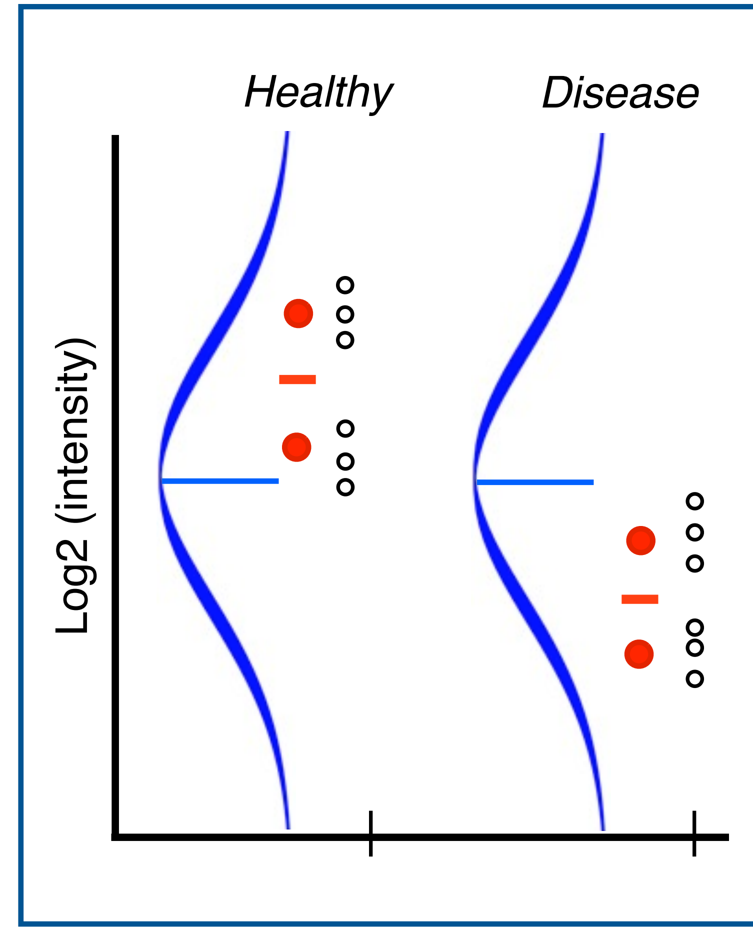
H0: 'status quo', no change in abundance, $\hat{G}_1 - \hat{G}_0 = 0$

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

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

Linear mixed models describe Normal distributions





Deviation from the reference due to

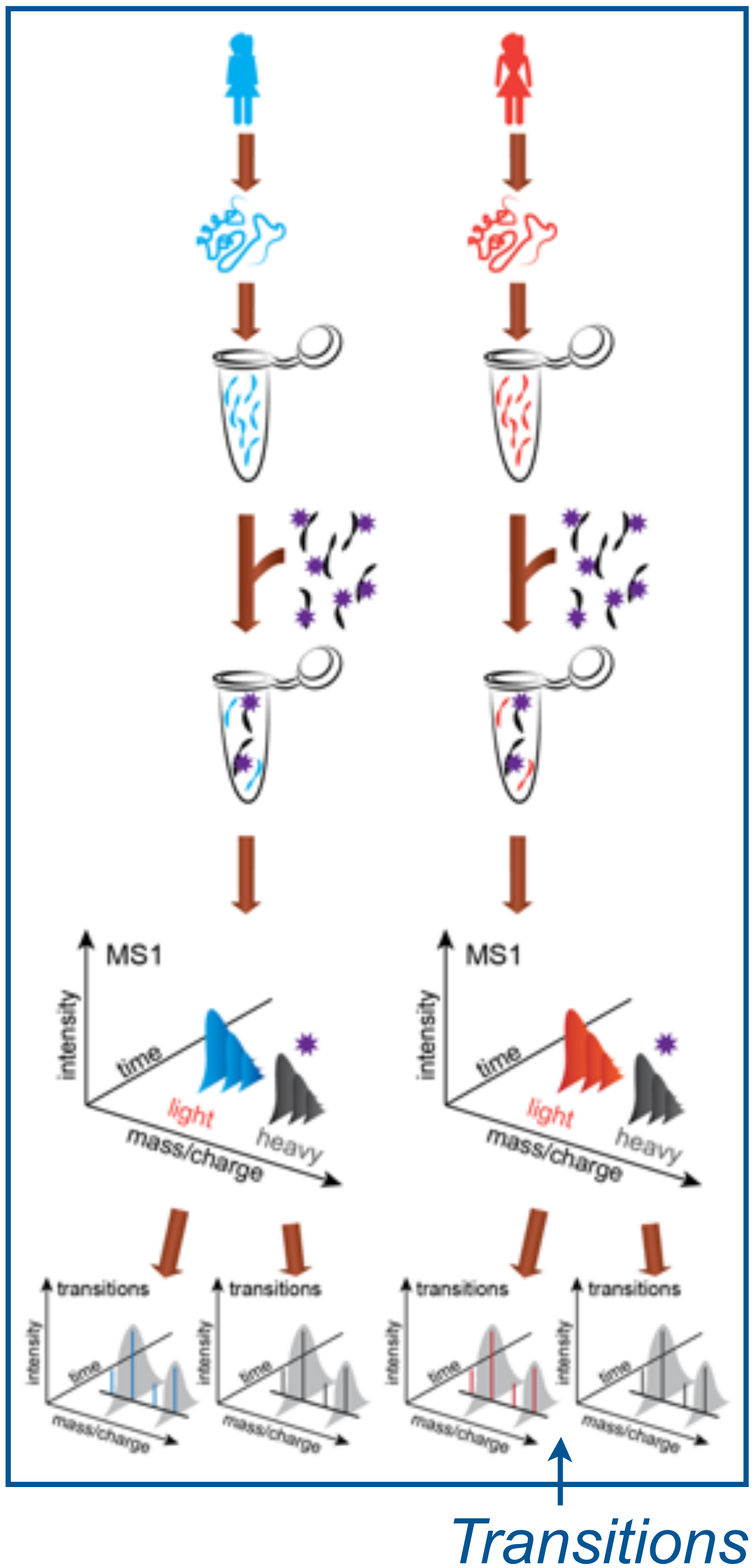
log(peak intensity)	=	Expected reference abundance	+	LC-MS feature	+	condition	+	feature × condition interaction	+	biol. replicate	+	Random meas. error
y_{ijkl}	$=$	μ_{111}	$+$	F_i	$+$	C_j	$+$	$(F \times C)_{ij}$	$+$	$S(C)_{k(j)}$	$+$	ε_{ijkl}

where $F_1 = C_1 = (F \times C)_{i1} = (F \times C)_{1j} = 0$ $\qquad \varepsilon_{ijkl} \overset{iid}{\sim} N\left(0, \sigma^2_{Error,\, ijk}\right)$

expanded scope of biological replication: $S(C)_{k(j)} \overset{iid}{\sim} N\left(0, \sigma^2_S\right)$

Labeled reference peptides help separate the biological and the technological variation

Label-based SRM workflow



Analysis of heavy/light peak pairs

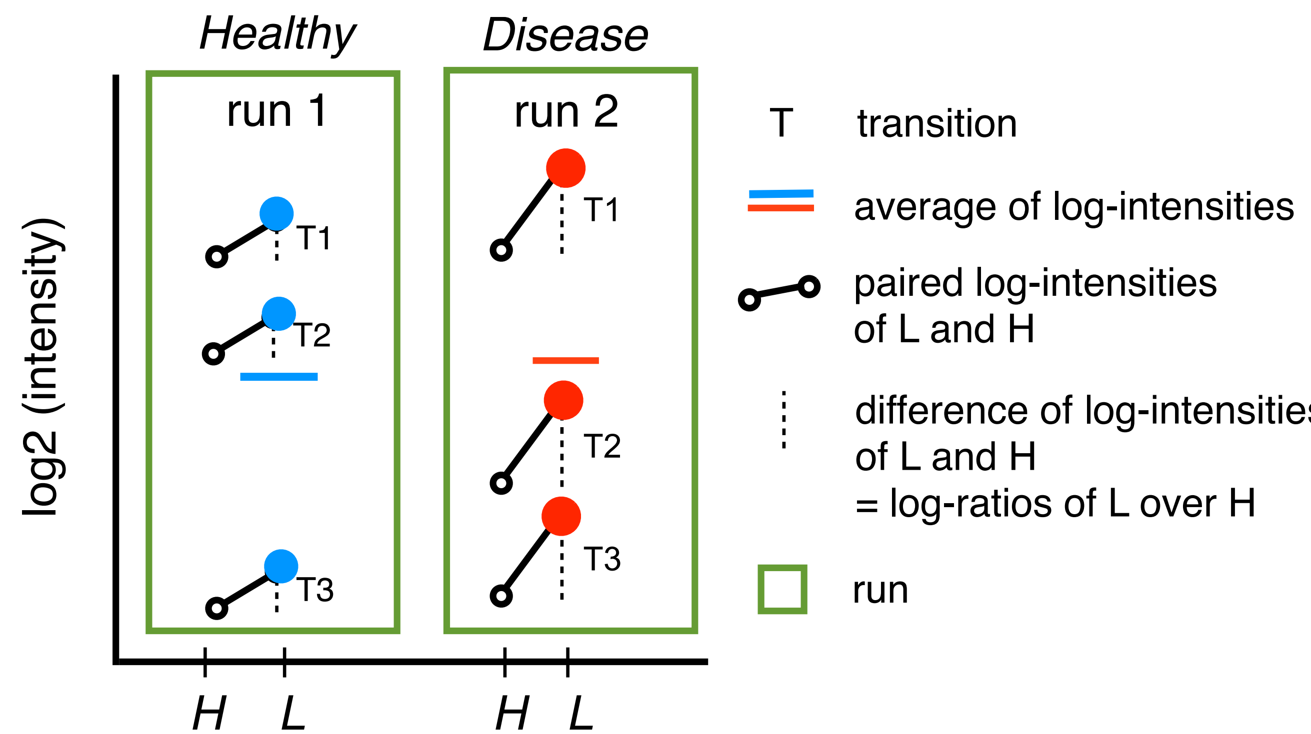


Table of quantified peaks

Label	Peptide	Transition	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

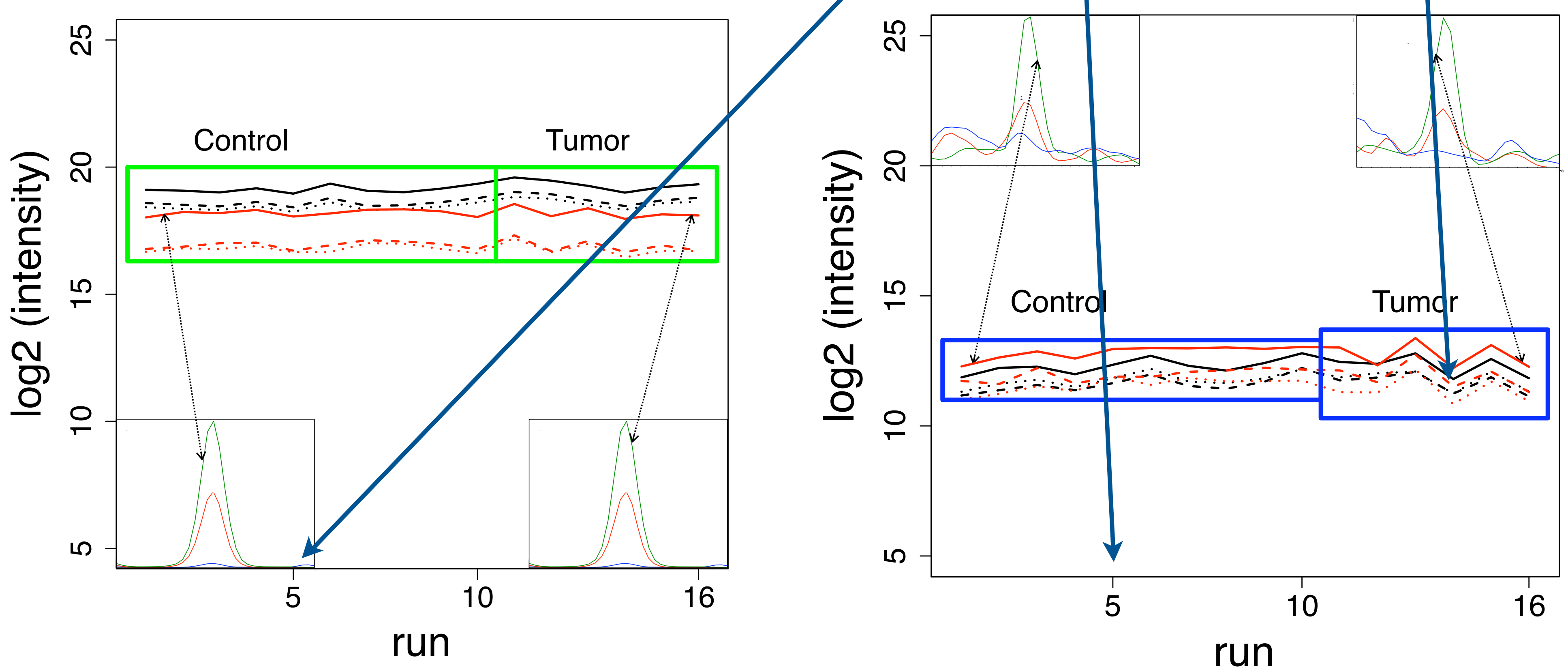
Legend : Label | Feature: Transition/Peptide | Group | Run | Subject

A full linear mixed model for an experiment with labeled reference peptides

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)

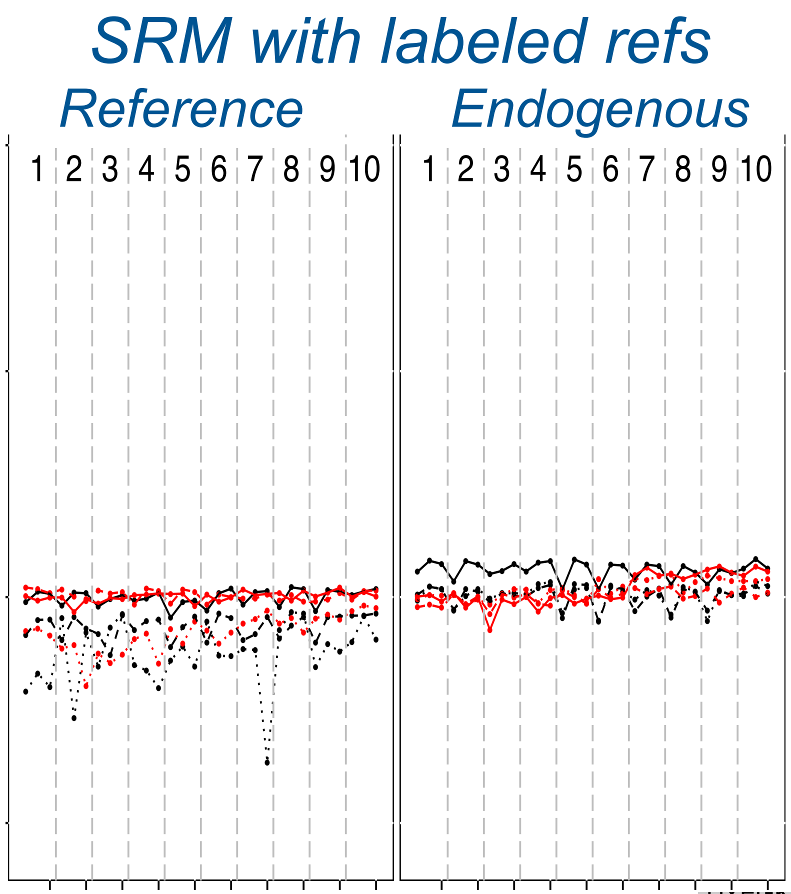
C.-Y. Chang et al., *MCP*, 2012

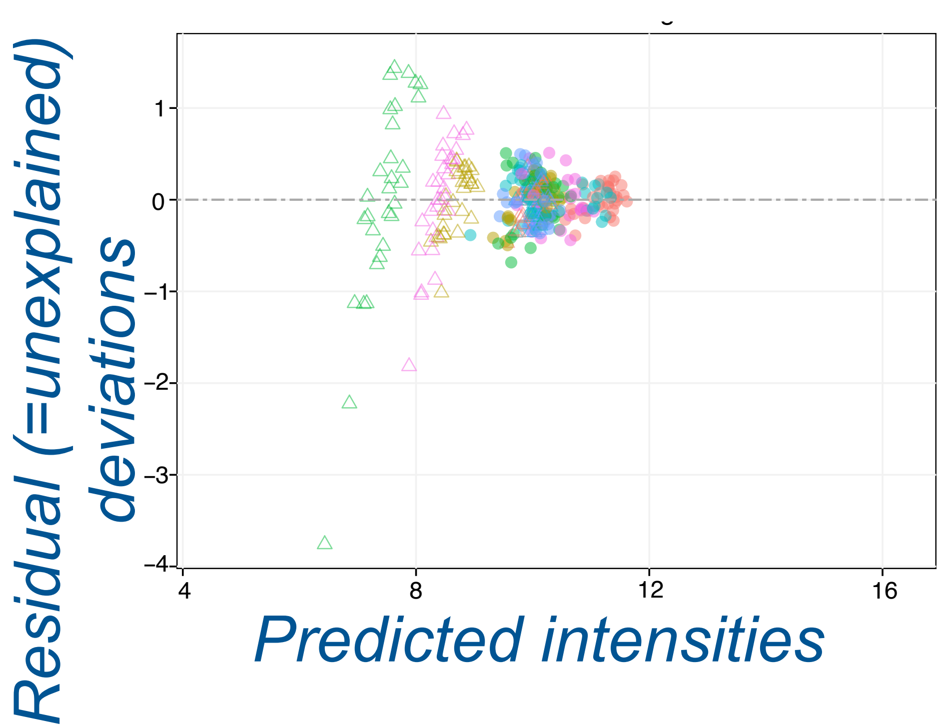


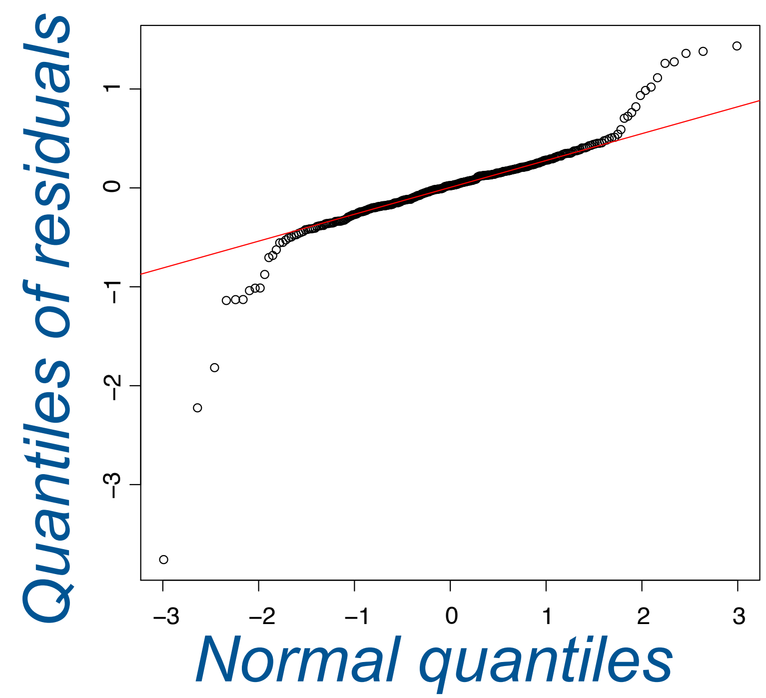
Model log2(int) instead of ratios light/heavy

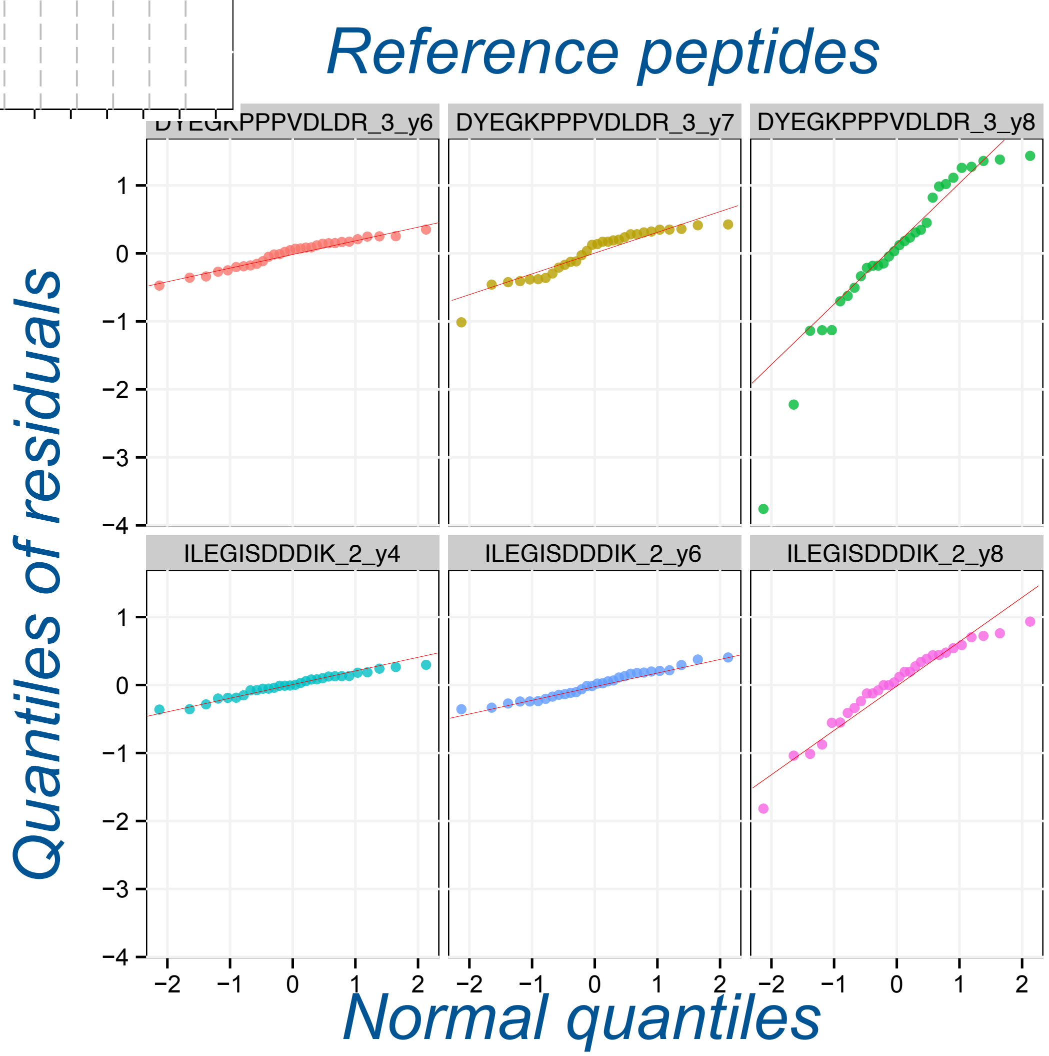
'Run' pairs endogenous and reference transitions from a same run

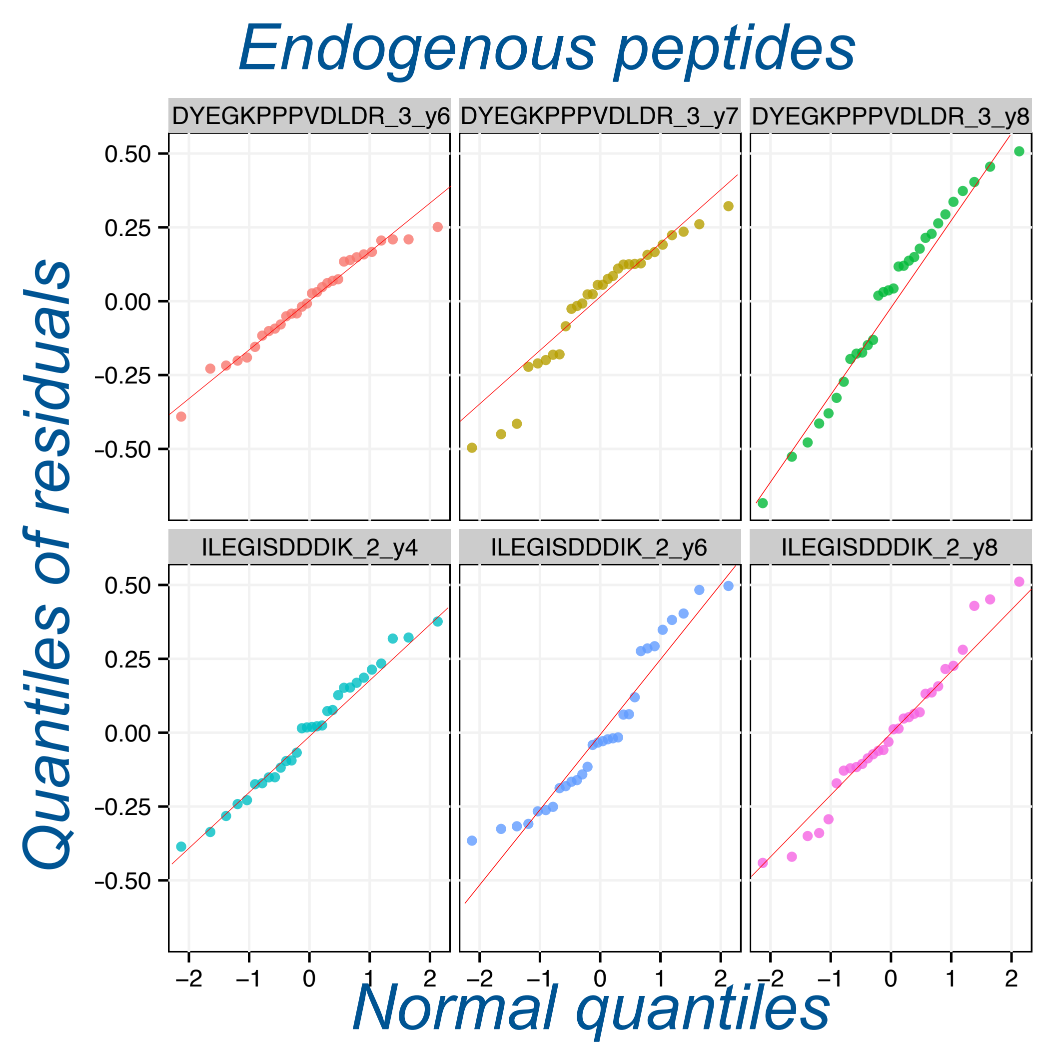
Check model assumptions







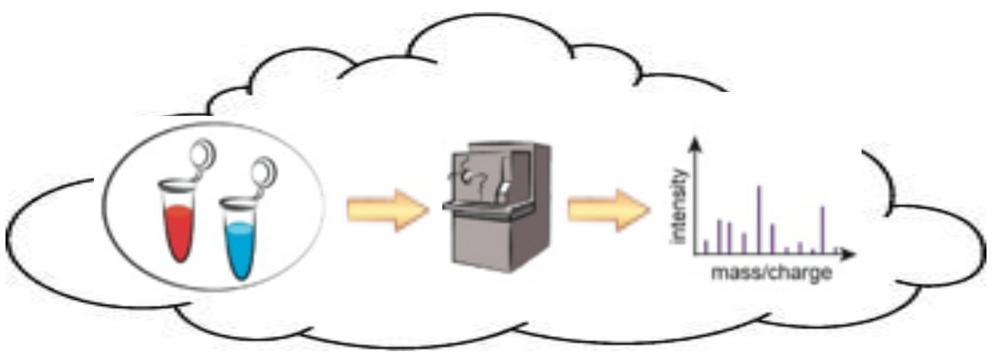




Deviations from independence or from constant variance are often mistaken for deviations from Normality

A typical analysis workflow (also in MSstats)

Experimental design



QC and normalization

Statistical modeling

Model-based conclusions

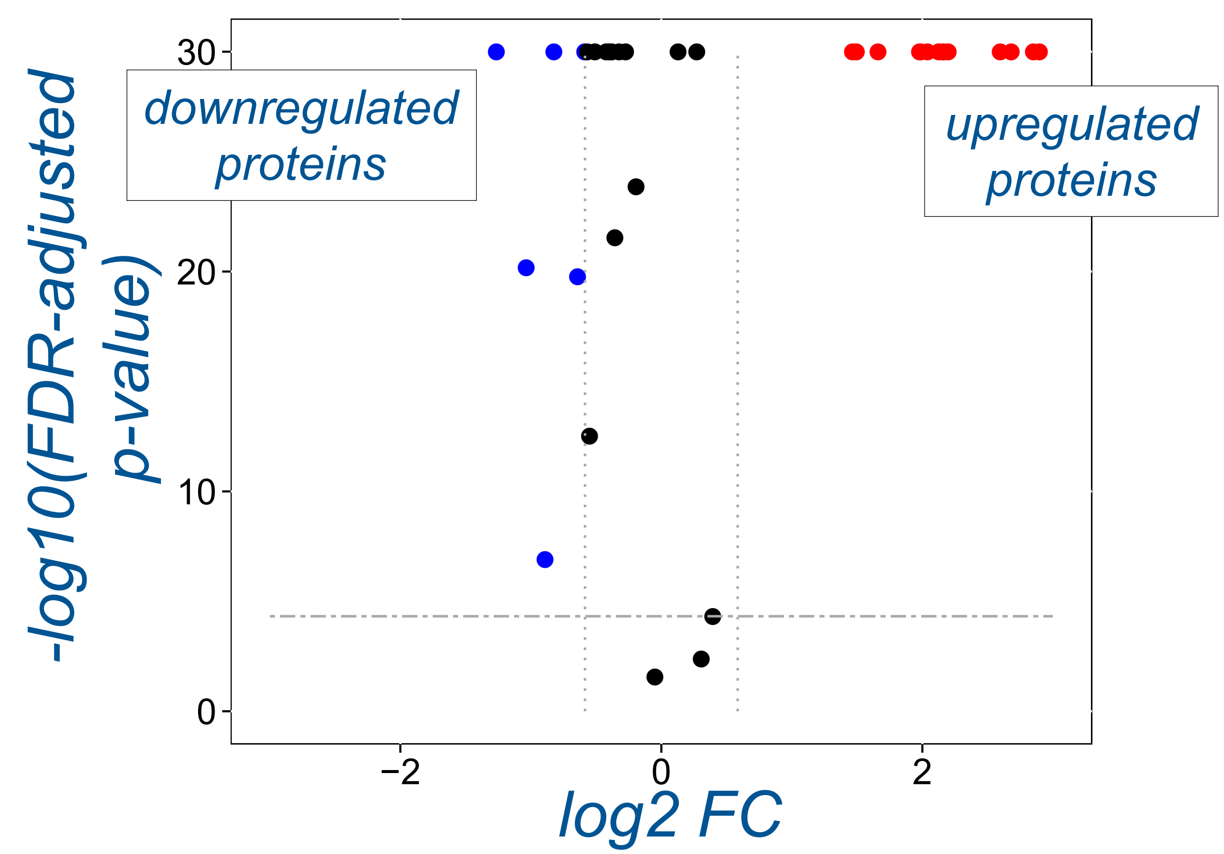
Experimental design

Model-based group comparisons

- Quantify the uncertainty
- Adjust p-values to control FDR

Relative protein quantification

- In one sample
- In one condition

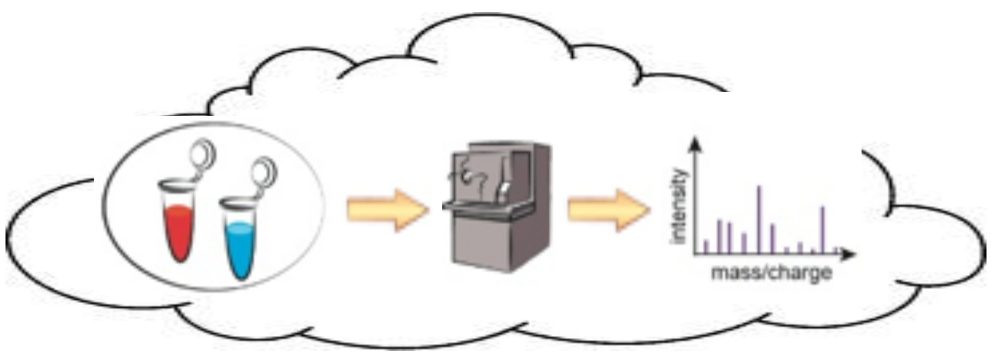


C.-Y. Chang et al. *MCP*, 2012

A typical analysis workflow (also in MSstats)

Experimental design

↓



↓

QC and normalization

↓

Statistical modeling

↓

Model-based conclusions

↓

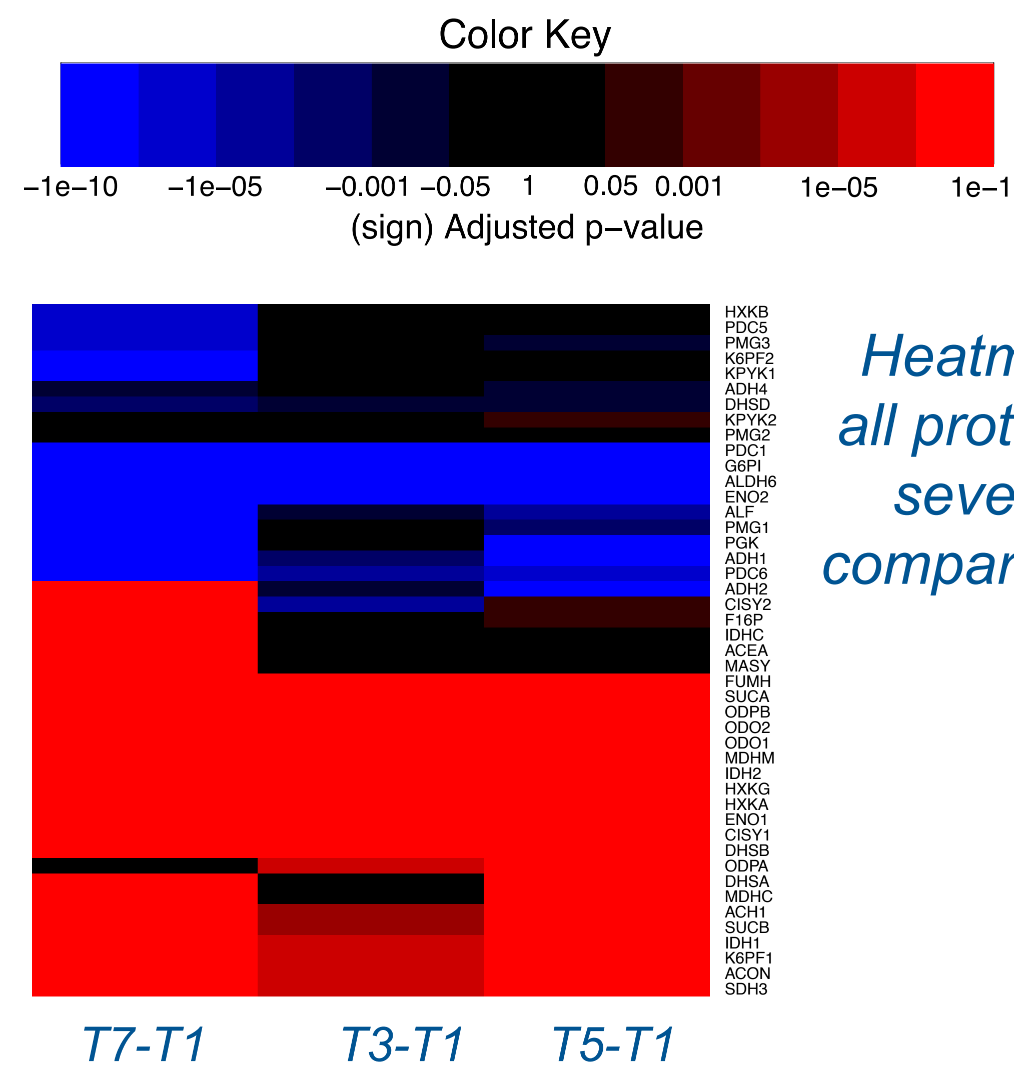
Experimental design

Model-based group comparisons

- Quantify the uncertainty
- Adjust p-values to control FDR

Relative protein quantification

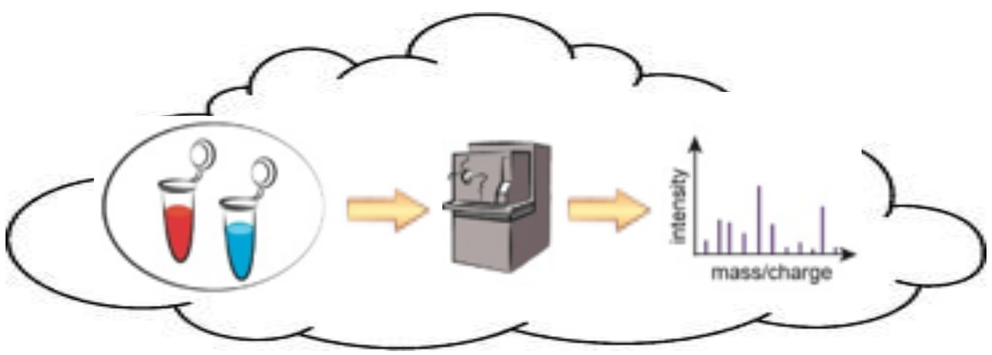
- In one sample
- In one condition



Heatmap: all proteins, several comparisons

A typical analysis workflow (also in MSstats)

Experimental design



QC and normalization

Statistical modeling

Model-based conclusions

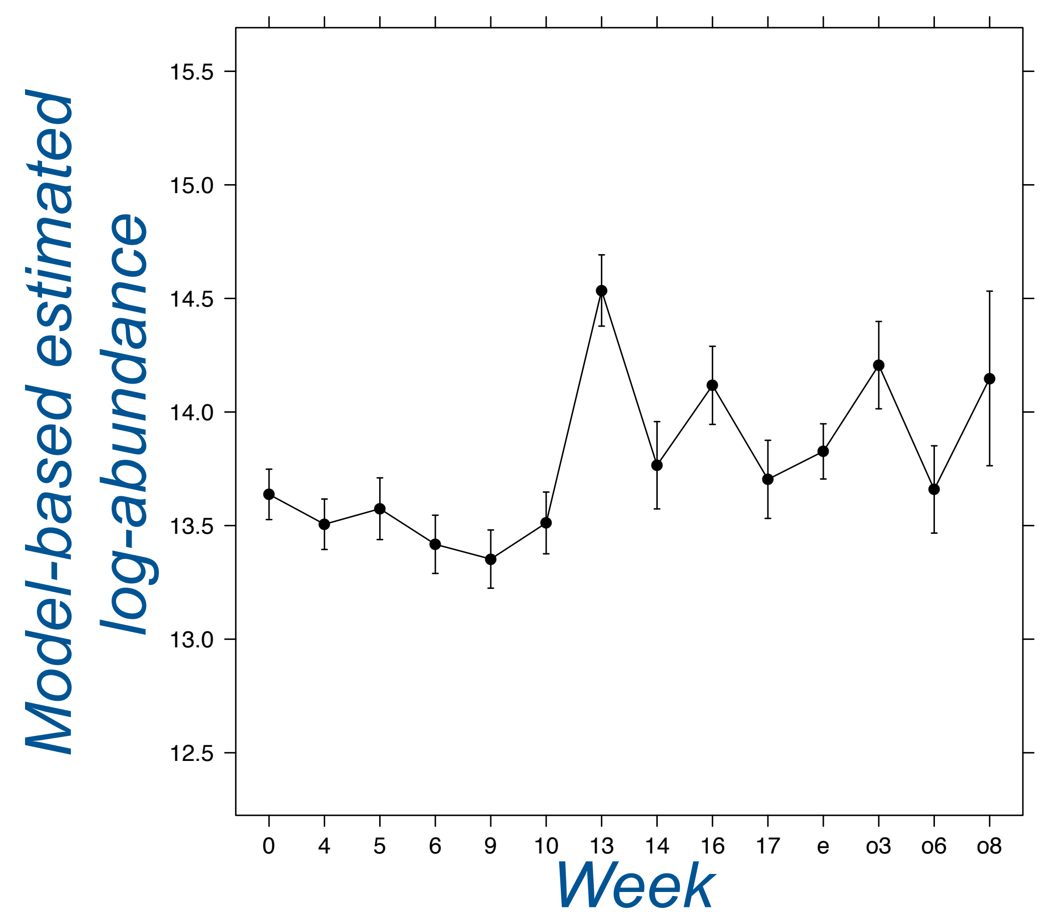
Experimental design

Model-based group comparisons

- Quantify the uncertainty
- Adjust p-values to control FDR

Relative protein quantification

- In one sample
- In one condition

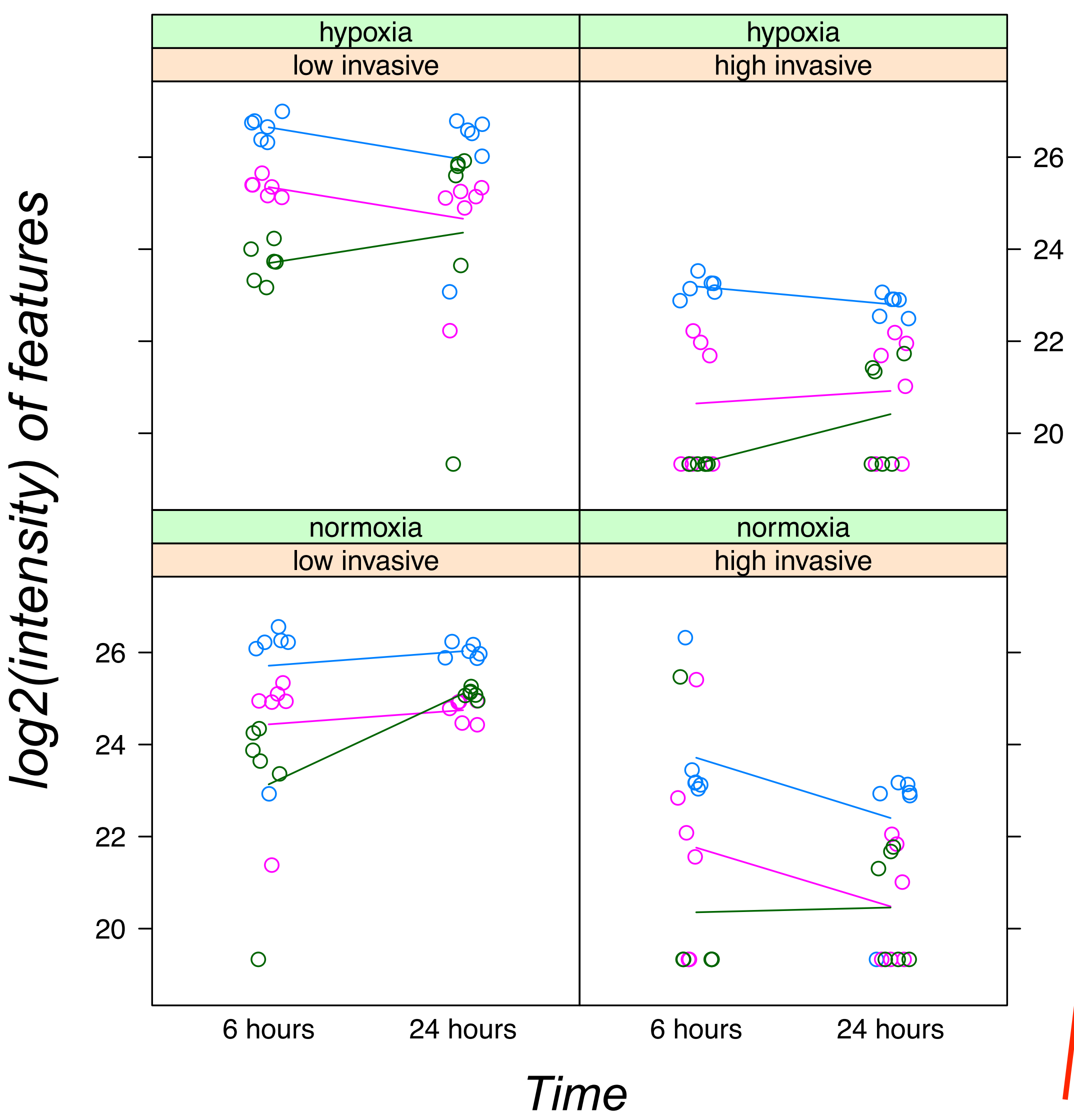


Model-based conclusions

Comparisons between conditions are estimated by linear combinations of model terms

log(peak intensity)	=	Expected reference abundance	+	LC-MS feature	+	condition	+	feature × condition interaction	+	biol. replicate	+	Random meas. error
y_{ijkl}	=	μ_{111}	+	F_i	+	C_j	+	$(F \times C)_{ij}$	+	$S(C)_{k(j)}$	+	ε_{ijkl}

Deviation from the reference due to: LC-MS feature, condition, feature × condition interaction, biol. replicate



Quantity of interest:

$$H_0 : L = \bar{\mu}_{\,[\text{high, nm, 6}]\cdot} - \bar{\mu}_{\,[\text{low, nm, 6}]\cdot} = 0$$

Model-based estimate and test statistic:

$$\hat{L} = \hat{C}_{[\text{high, nm, 6}]} + \frac{1}{I}\sum_{i=1}^{I} \widehat{(F \times C)}_{i,\,[\text{high, nm, 6}]} + \frac{1}{K}\sum_{k=1}^{K} \widehat{S(C)}_{k([\text{high, nm, 6}])}$$

$$- \left(\hat{C}_{[\text{low, nm, 6}]} + \frac{1}{I}\sum_{i=1}^{I} \widehat{(F \times C)}_{i,\,[\text{low, nm, 6}]} + \frac{1}{K}\sum_{k=1}^{K} \widehat{S(C)}_{k([\text{low, nm, 6}])} \right)$$

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

In balanced datasets:

$$\hat{L} = \bar{Y}_{\cdot[\text{high, nm, 6}]\cdot} - \bar{Y}_{\cdot[\text{low, nm, 6}]\cdot}$$

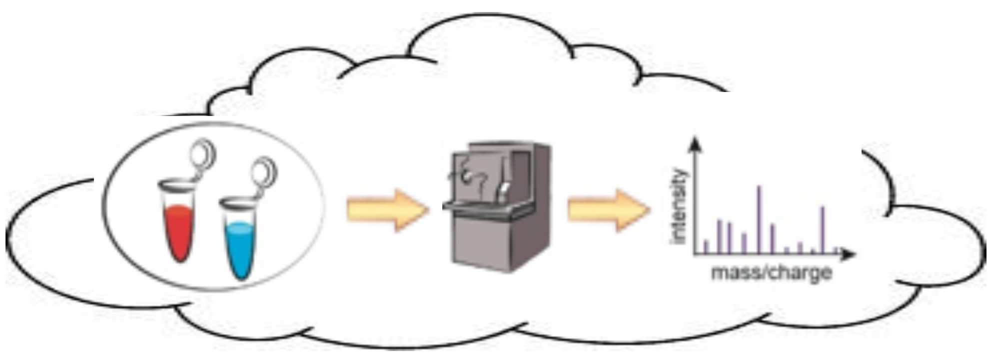
$$t = \frac{\hat{L}}{\sqrt{\frac{2}{IKL}\hat{\sigma}^2_{Error}}} \sim \text{Student}_{IJK(L-1)+(I-1)J(K-1)} \text{ distribution}$$

Model-based S/N

Leads to p-values

A typical analysis workflow (also in MSstats)

Experimental design



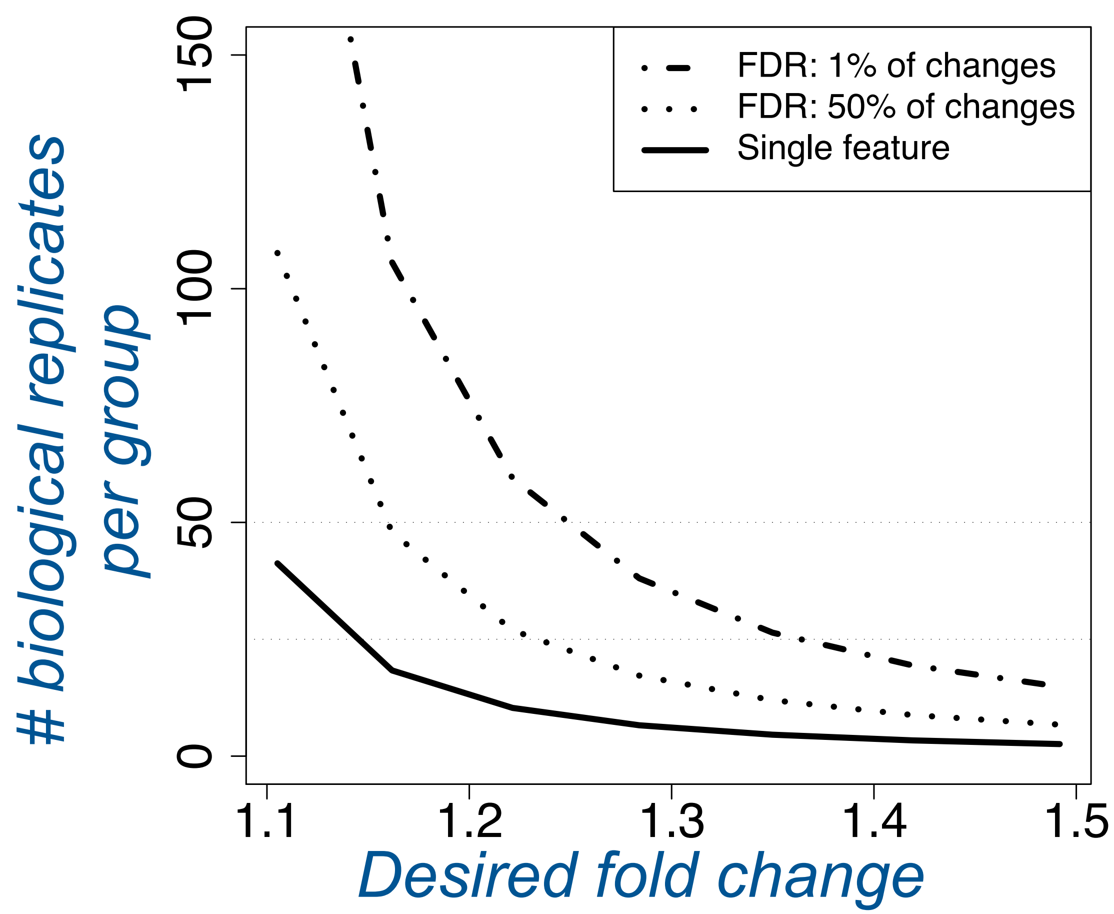
QC and normalization

Statistical modeling

Model-based conclusions

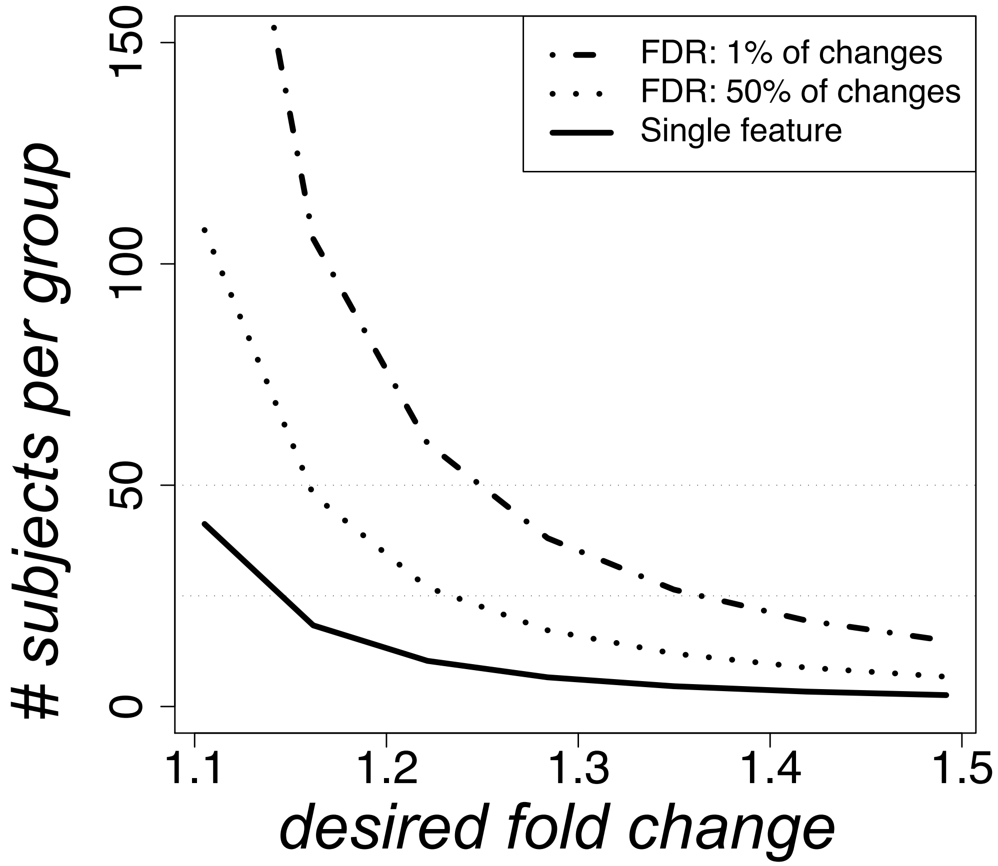
Experimental design

Use the dataset to improve:
- Subject selection: matching
- Resource allocation: blocking
- Calculation of sample size

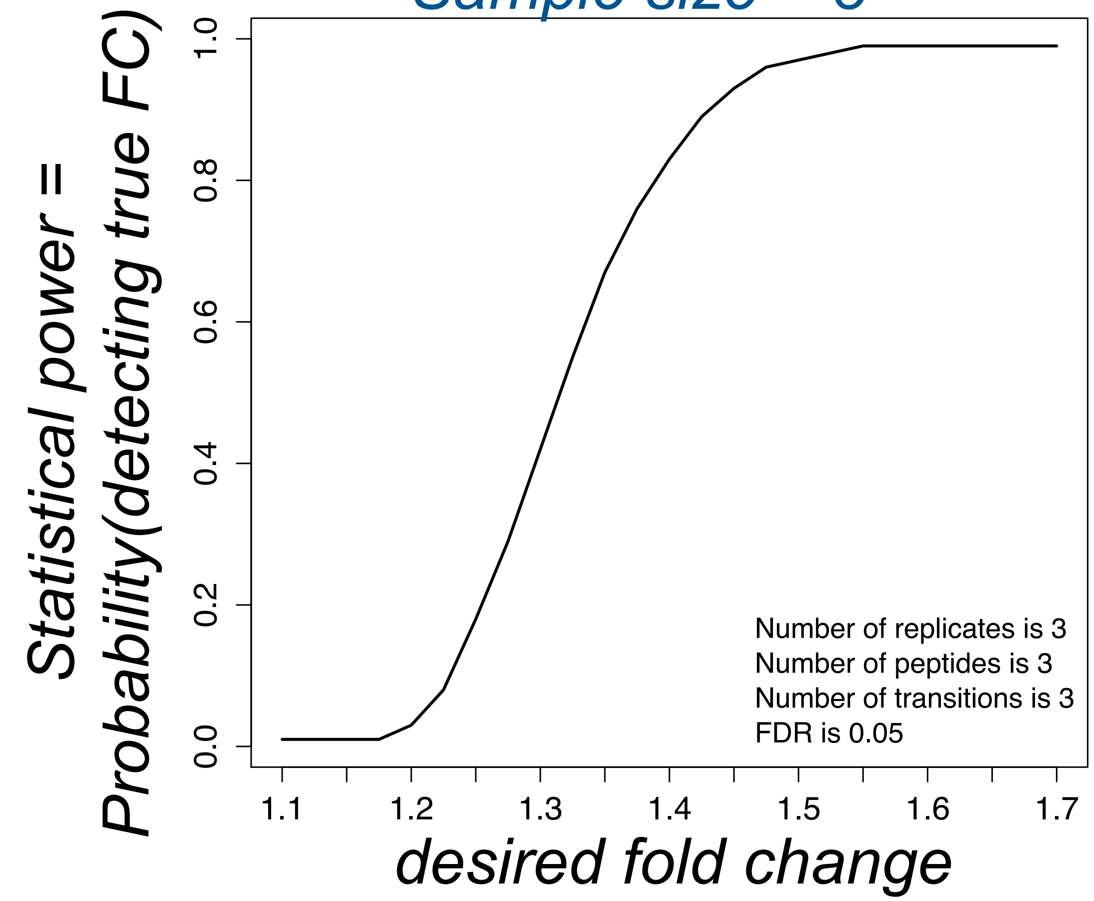


Linear mixed effects models are required to calculate the sample size and the power

Statistical power = P(detect change) = 0.8



Sample size = 3



Need to know in advance:

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

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

β - statistical power (i.e. probability of a true positive discovery)

$\boldsymbol{\Delta}$ - anticipated (log-) fold change

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

Oberg and Vitek, *JPR*, 2009