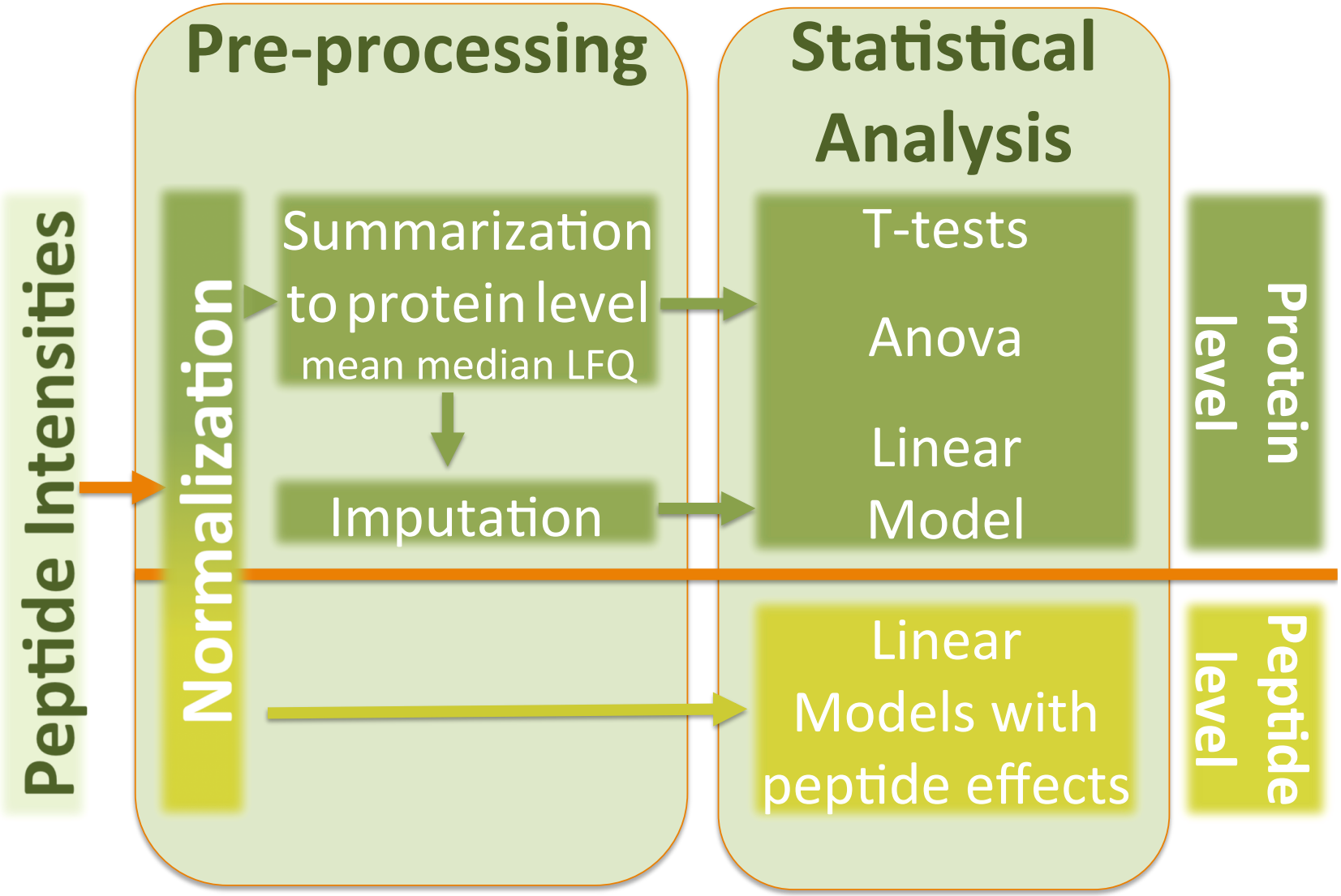
Part II: Statistical Inference

Lieven Clement

Proteomics Data Analysis 2018, Gulbenkian Institute, May 28 -June 1 2018.

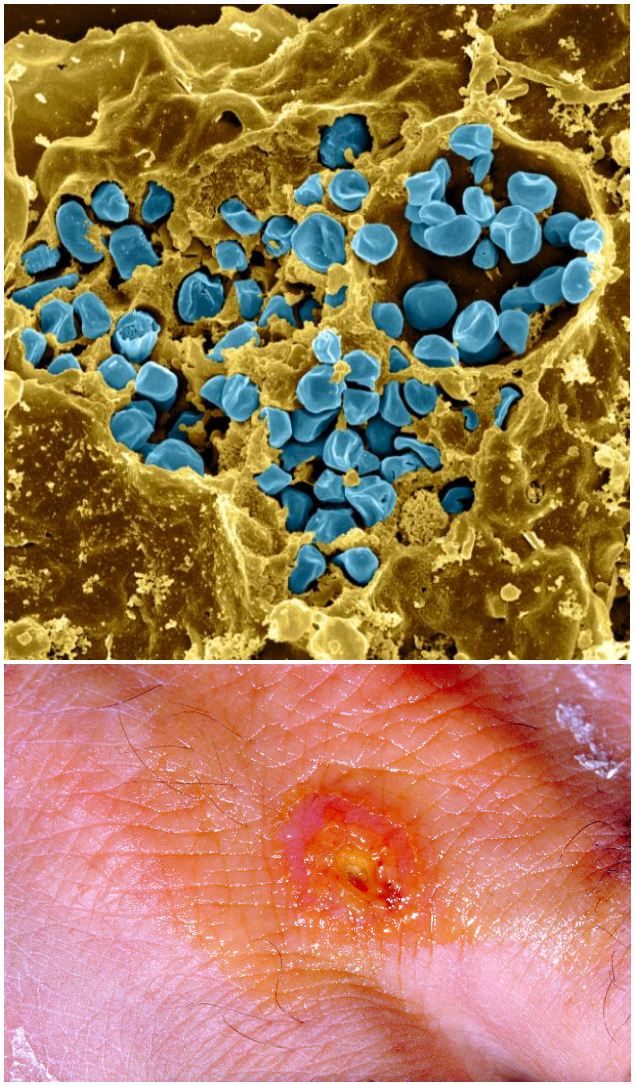
Label-free Quantitative Proteomics Data Analysis Pipelines



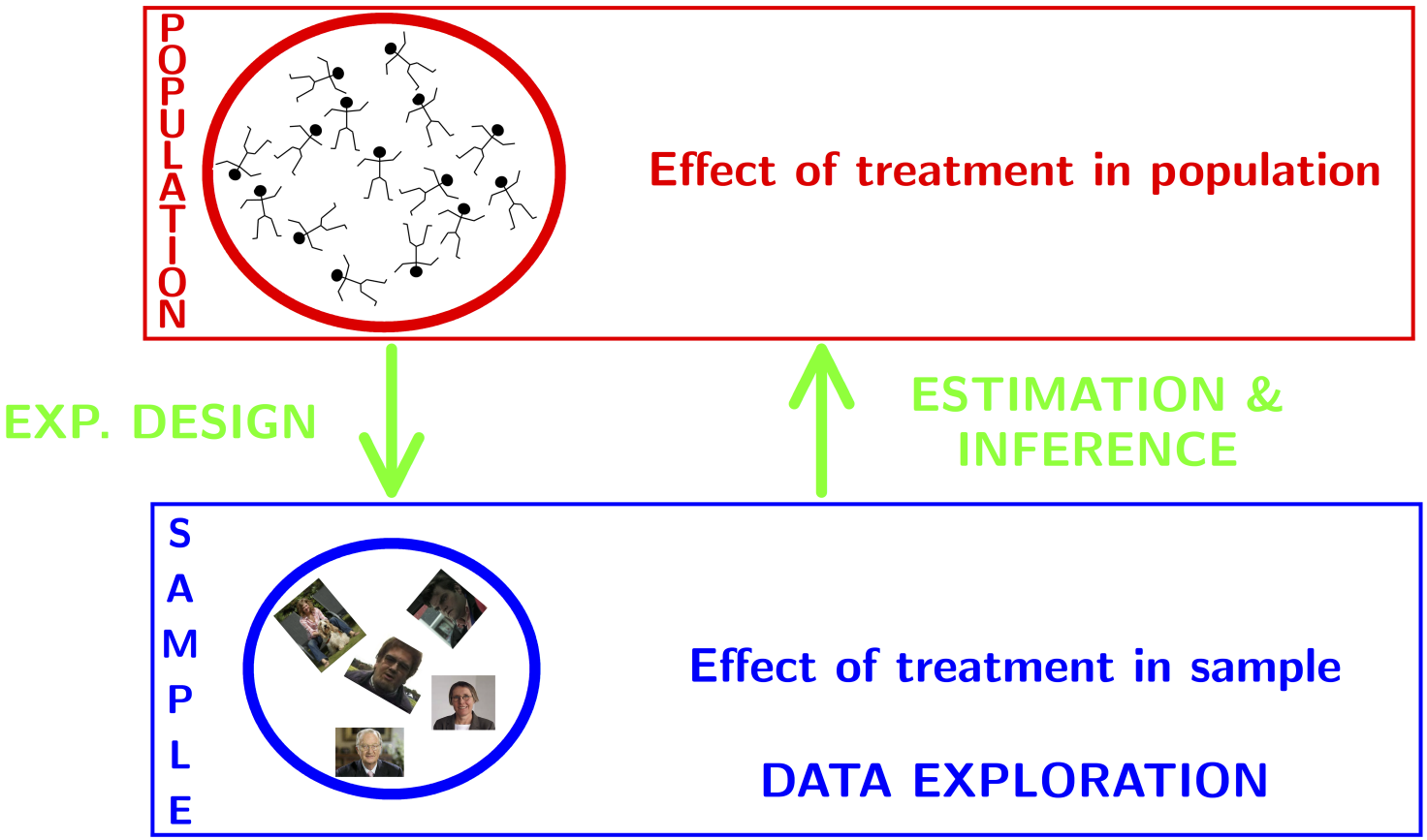
Statistical Inference

1. Francisella tularensis Example
2. Hypothesis testing
3. Multiple testing
4. Moderated statistics
5. Experimental design
6. Peptide based models

Francisella tularensis experiment



- Pathogen: causes tularemia
- Metabolic adaptation key for intracellular life cycle of pathogenic microorganisms.
- Upon entry into host cells quick phasomal escape and active multiplication in cytosolic compartment.
- Francisella is auxotroph for several amino acids, including arginine.
- Inactivation of arginine transporter delayed bacterial phagosomal escape and intracellular multiplication.
- Experiment to assess difference in proteome using 3 WT vs 3 ArgP KO mutants

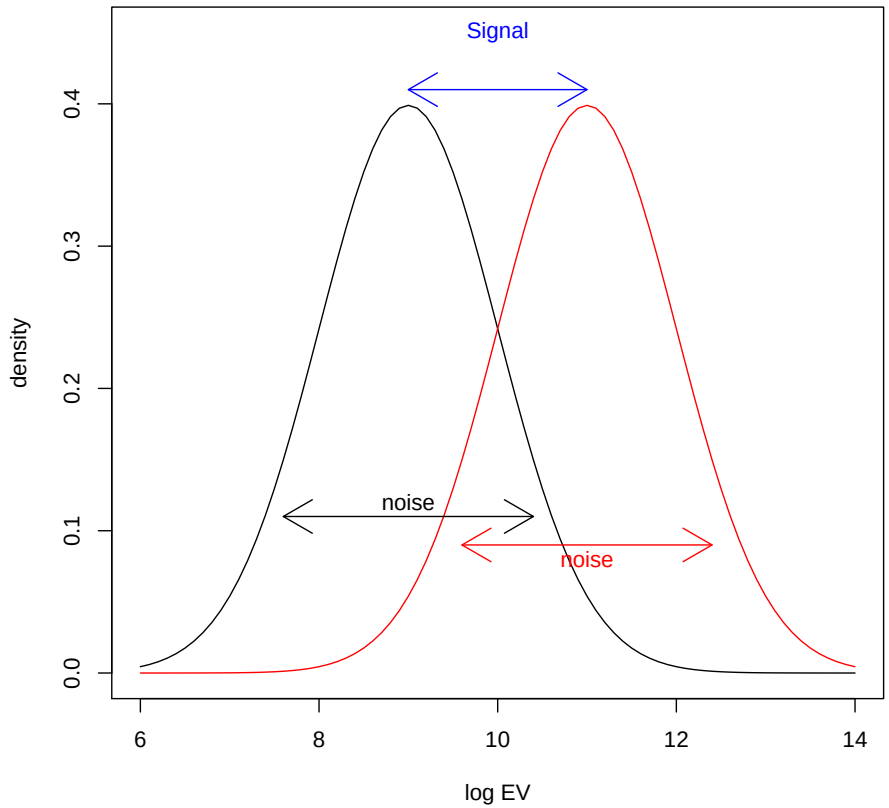

POPULATION
Effect of treatment in population
EXP. DESIGN
ESTIMATION & INFERENCE
SAMPLE
Effect of treatment in sample
DATA EXPLORATION

Summarized data structure

- WT vs KO
- 3 vs 3 repeats
- 882 proteins

Protein	WT_1	WT_2	WT_3	KO_1	KO_2	KO_3
gi|118496616	29.83	29.77	29.91	29.70	29.86	29.80
gi|118496617	31.28	31.23	31.51	31.30	31.51	31.76
gi|118496635	32.39	32.27	32.24	32.25	32.14	32.22
gi|118496636	30.74	30.54	30.64	30.65	30.49	30.60
gi|118496637	29.56	29.35	29.56	29.30	29.24	29.14
gi|118498323	31.38	30.52	30.62	31.04	27.38	NA
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Hypothesis testing: a single protein



$$\Delta = \bar{z}_{p1} - \bar{z}_{p2}$$

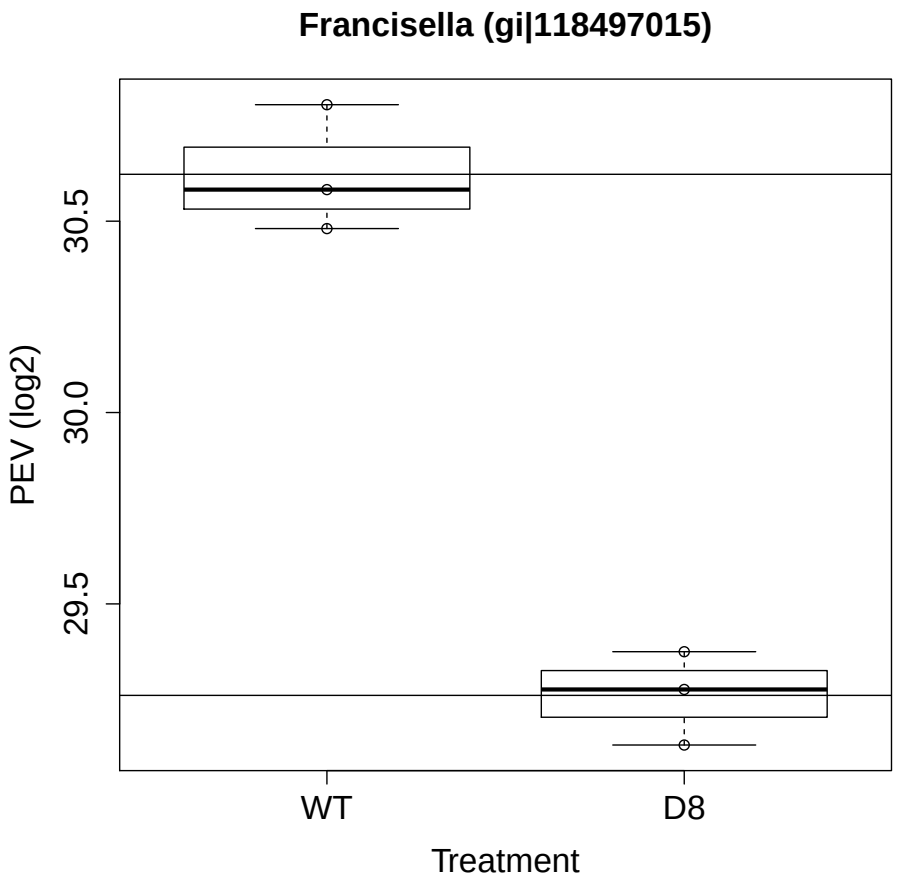
$$T_g = \frac{\Delta}{\text{se}_\Delta}$$

$$T_g = \frac{\widehat{\text{signal}}}{\widehat{\text{Noise}}}$$

If we can assume equal variance in both treatment groups:

$$\text{se}_\Delta = \text{SD}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Hypothesis testing: a single protein



$$t = \frac{\log_2 \widehat{FC}}{se_{\log_2 \widehat{FC}}} = \frac{-1.4}{0.118} = -11.9$$

Is $t = -11.9$ indicating that there is an effect?

How likely is it to observe $t = -11.8$ when there is no effect of the argP KO on the protein expression?

Null hypothesis and alternative hypothesis

- In general we start from **alternative hypothese** H_A: we want to show an effect of the KO on a protein
  - On average the protein abundance in WT is different from that in KO

Null hypothesis and alternative hypothesis

- In general we start from **alternative hypothese** H_A: we want to show an effect of the KO on a protein
  - On average the protein abundance in WT is different from that in KO
- But, we will assess it by falsifying the opposite: **null hypothesis** H_0
  - On average the protein abundance in WT is equal to that in KO

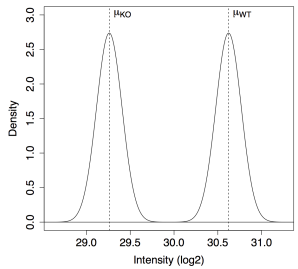
```
Two Sample t-test

data:  z by treat
t = -11.449, df = 4, p-value = 0.0003322
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.031371 -1.691774
sample estimates:
mean in group D8 mean in group WT
        29.26094         30.62251
```

- How likely is it to observe an equal or more extreme effect than the one observed in the sample when the null hypothesis is true?
- When we make assumptions about the distribution of our test statistic we can quantify this probability: **p-value**. The p-value will only be calculated correctly if the underlying assumptions hold!
- When we repeat the experiment, the probability to observe a fold change more extreme than a 2.6 fold ($\log_2 FC = -1.36$) down or up regulation by random change (if H_0 is true) is 3 out of 10.000.
- If the p-value is below a significance threshold α we reject the null hypothesis. **We control the probability on a false positive result at the α-level (type I error)**

Hypothesis testing: a single protein

POPULATION



Effect of treatment in population

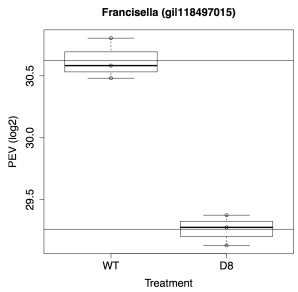
$\hat{\mu}_{wt} = 30.6$, $\hat{\mu}_{ko} = 29.3$, $\hat{FC} = 39\%$

$\hat{\sigma} = 0.14$, 95% CI $[31\%, 49\%]$, $p < 0.001$

EXP. DESIGN

ESTIMATION & INFERENCE

SAMPLE



Effect of treatment in sample

$\bar{z}_{wt} = 30.6$, $\bar{z}_{ko} = 29.3$, $\log_2 \bar{FC} = -1.4$

$SD_{=}0.14$, $SE_{\Delta} = 0.118$

Multiple hypothesis testing

Problem of multiple hypothesis testing

- Consider testing DA for all $m = 882$ proteins simultaneously
- What if we assess each individual test at level α?

$\rightarrow$ Probability to have a false positive among all m simultatenous test $>>>$ $\alpha = 0.05$

Suppose that 600 proteins are non-DA, then we could expect to discover on average $600 \times 0.05 = 30$ false positive proteins. Hence, we are bound to call false positive proteins each time we run the experiment.

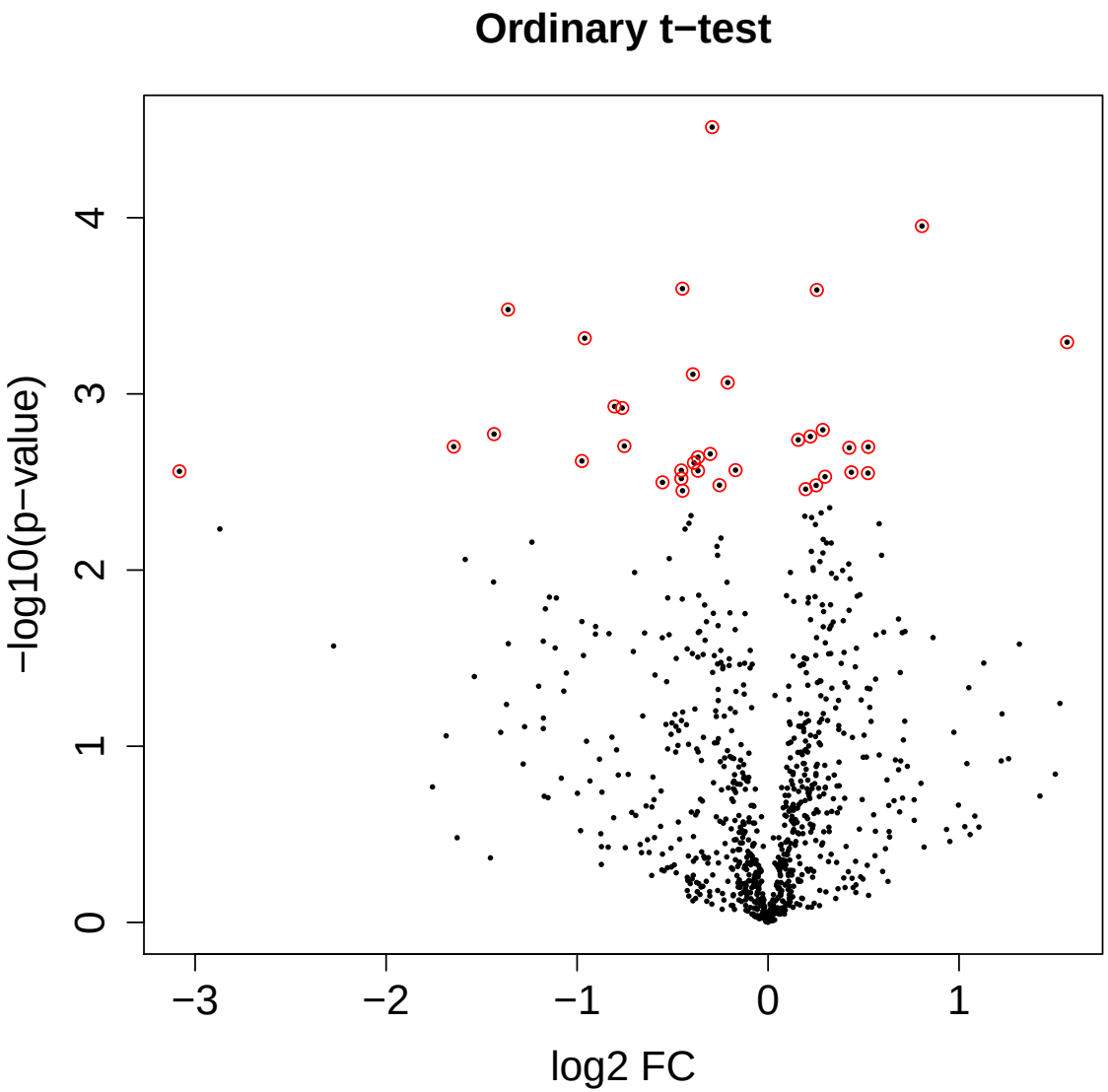
FDR: False discovery rate

- FDR: Expected proportion of false positives on the total number of positives you return.
- An FDR of 1% means that on average we expect 1% false positive proteins in the list of proteins that are called significant.
- Defined by Benjamini and Hochberg in 1995

$$\text{FDR}(|t_{\text{thres}}|) = \text{E}\left[\frac{FP}{FP+TP}\right] = \frac{\pi_0 Pr(|T| \geq t_{\text{thres}}|H_0)}{Pr(|T| \geq t_{\text{thres}})}$$

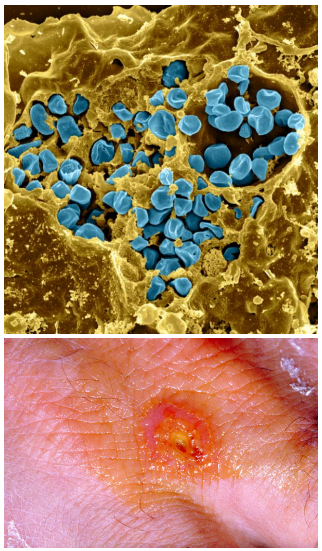
$$\text{FDR}_{\text{BH}}(|t_{\text{thres}}|) = \frac{1 \times p_{t_{\text{thres}}}}{\frac{\#|t_i| \leq t_{\text{thres}}}{m}}$$

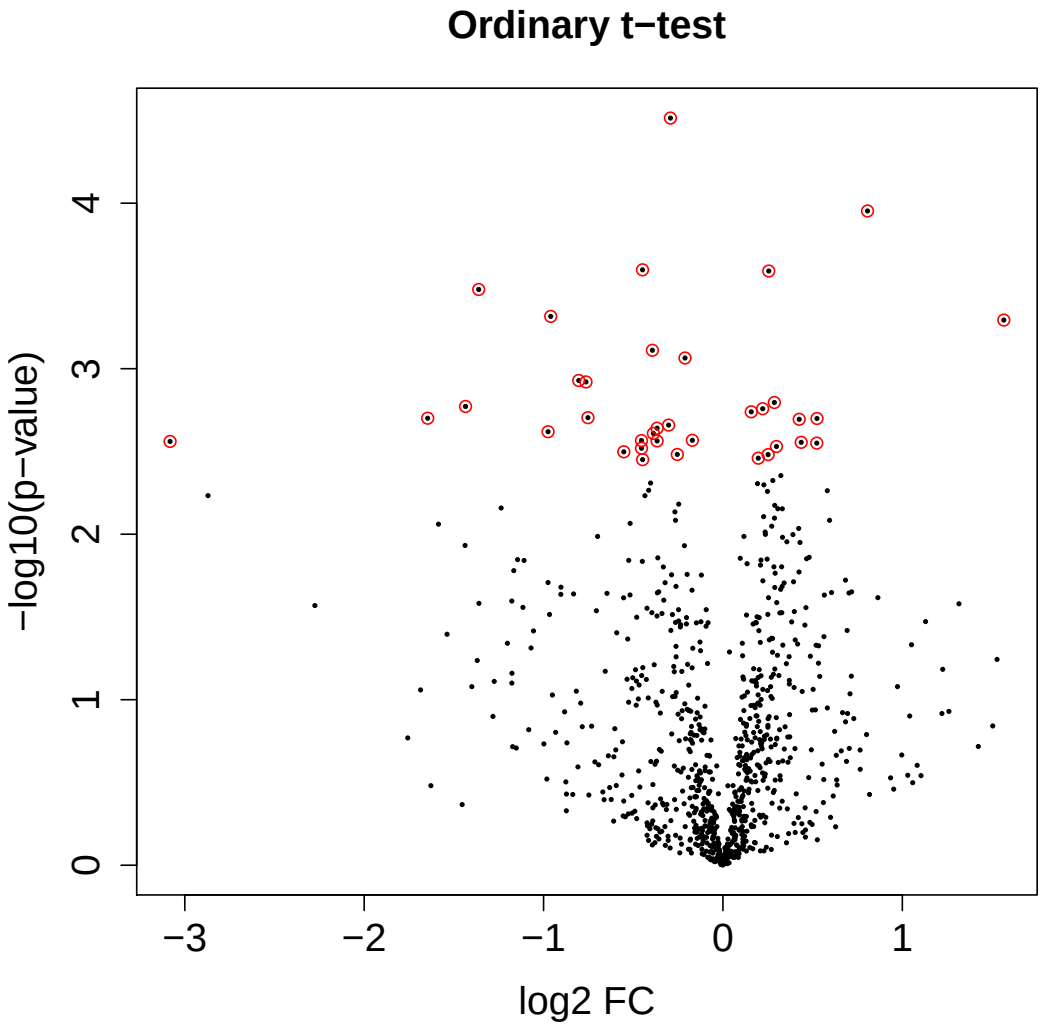
- FDR adjusted p-values can be calculated (e.g. Perseus, R, ...)


Ordinary t−test
−log10(p−value)
0
1
2
3
4
−3
−2
−1
0
1
log2 FC

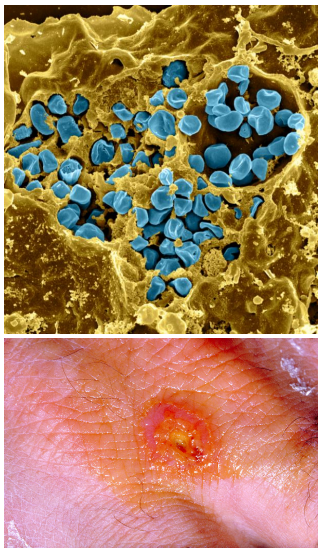
Moderated Statistics

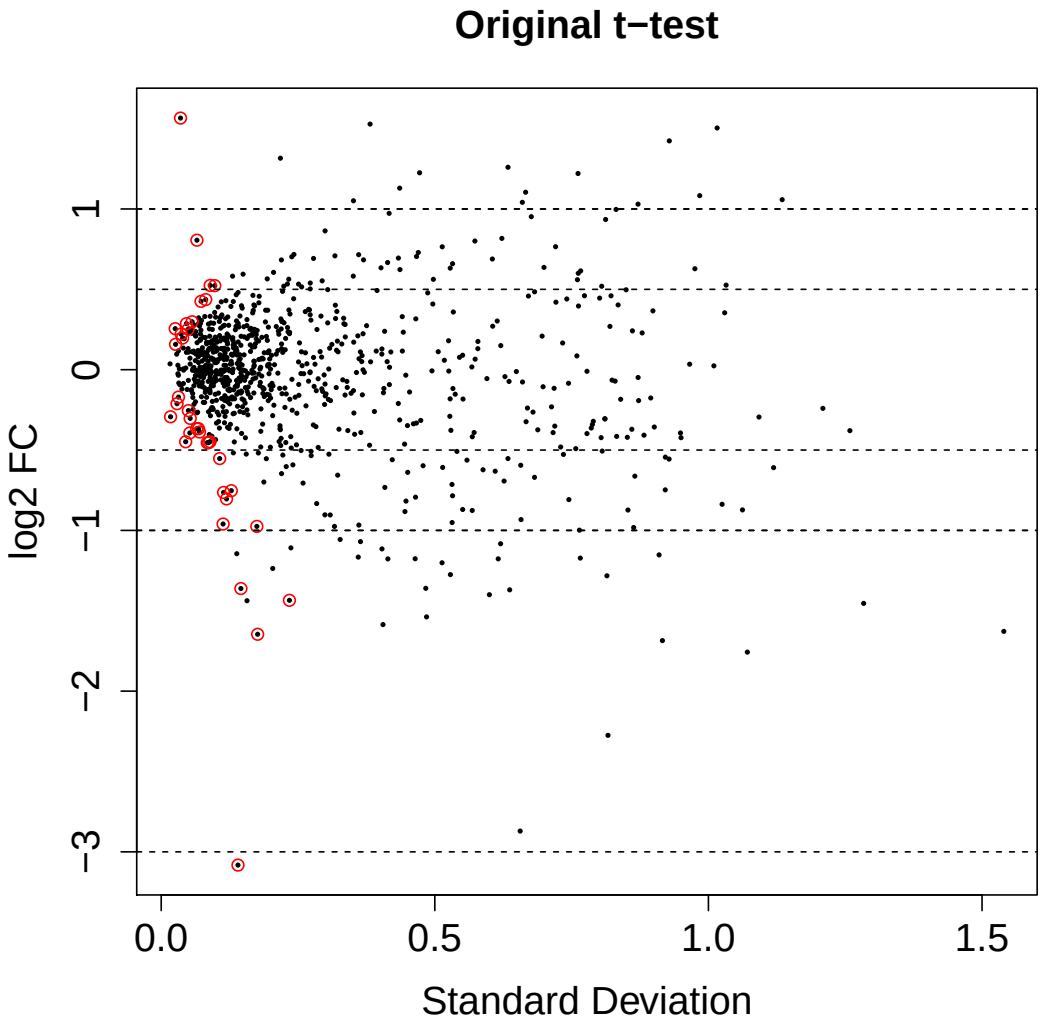
Problems with ordinary t-test





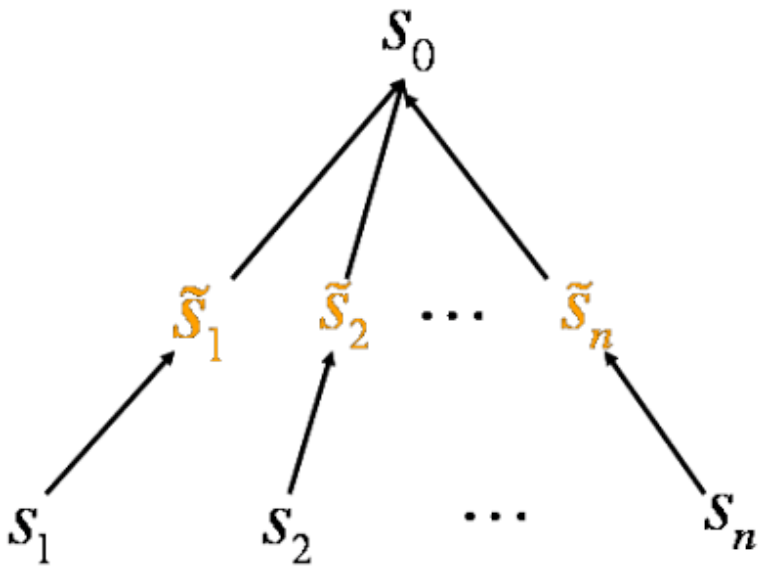
Problems with ordinary t-test

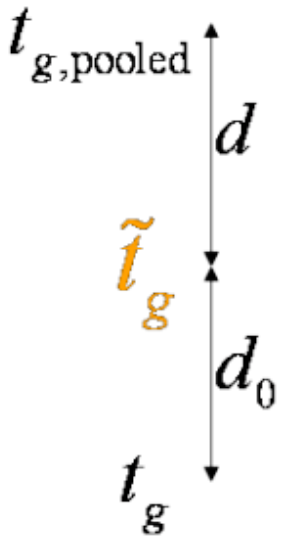




Shrinkage of the variance and moderated t-statistics

Shrinkage of Standard Deviations

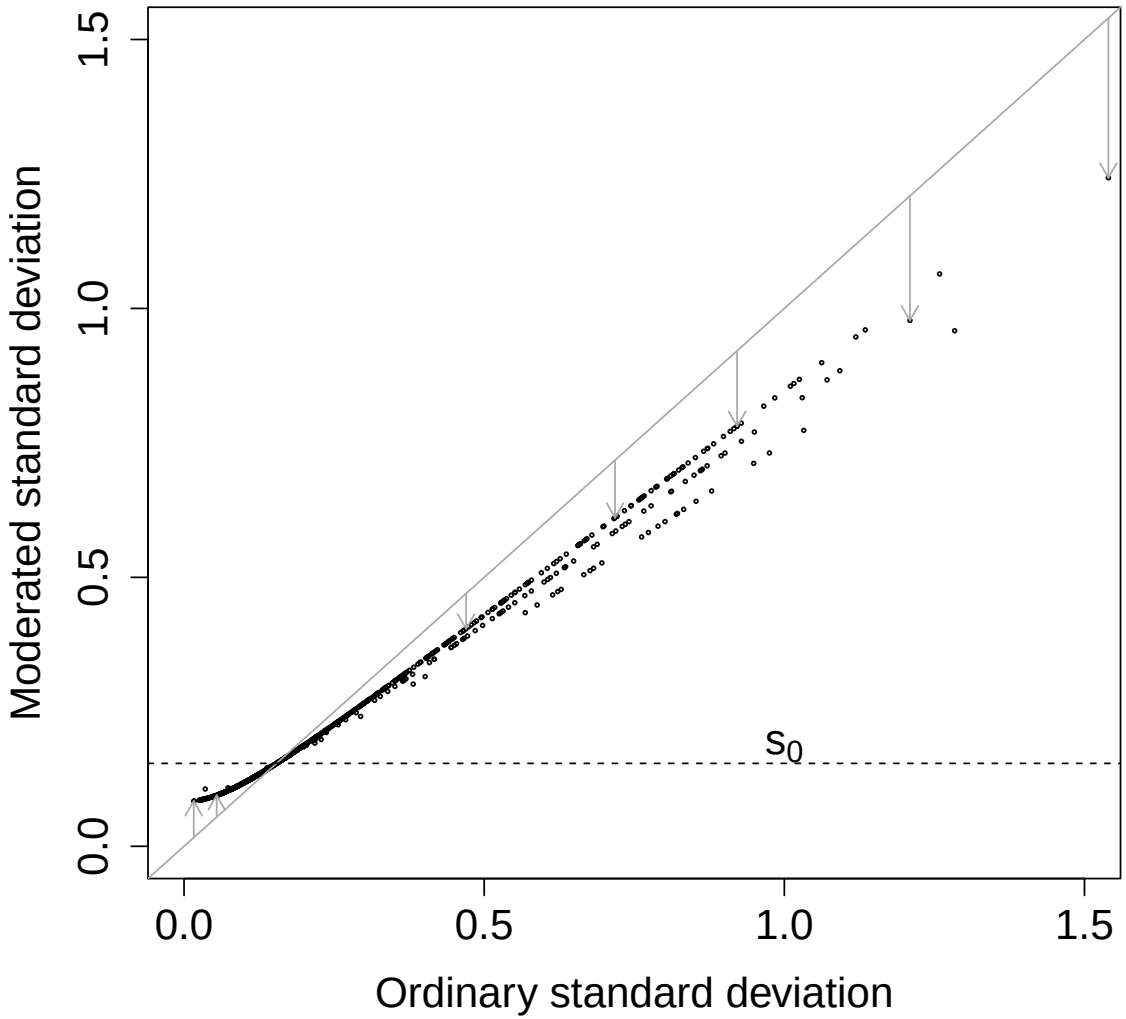




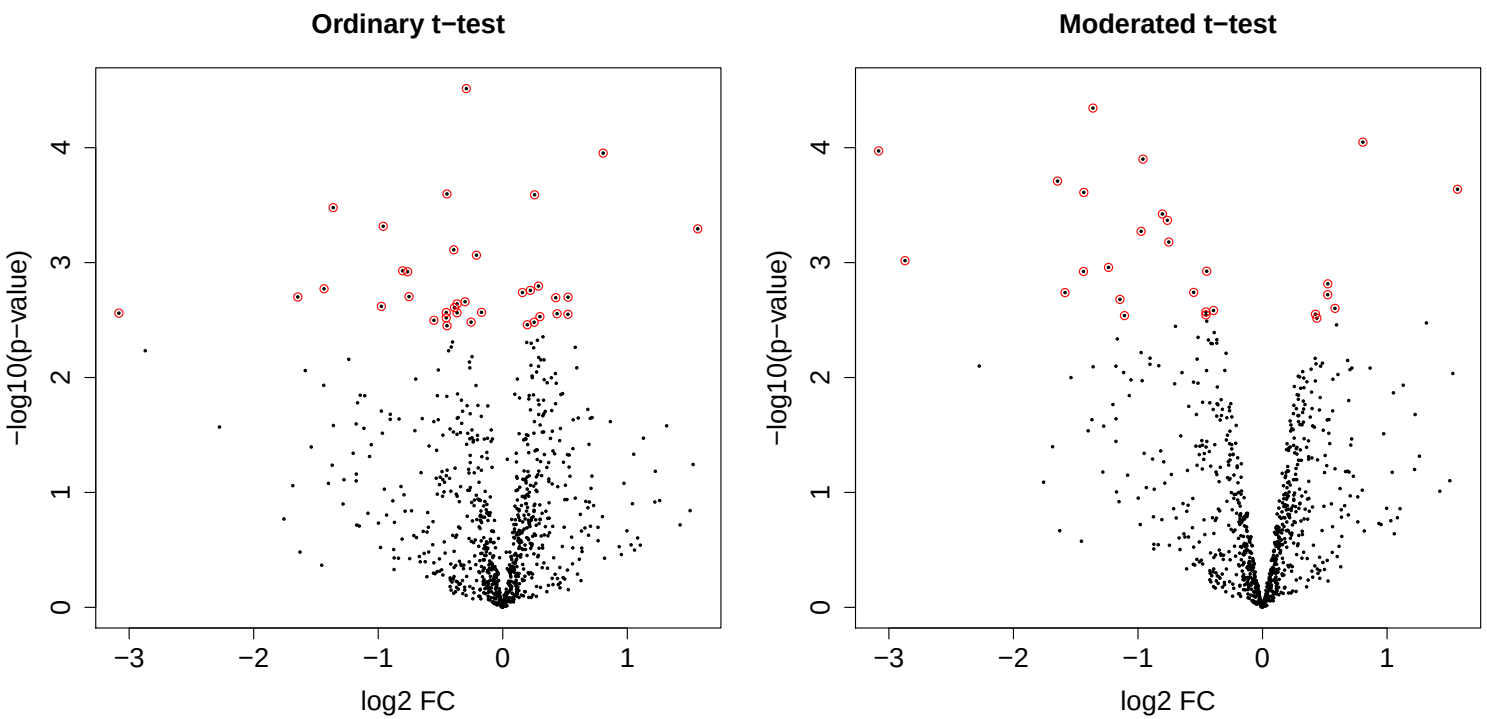
The data decides whether $\tilde{t}_g$

should be closer to $t_{g,pooled}$ or to t_g

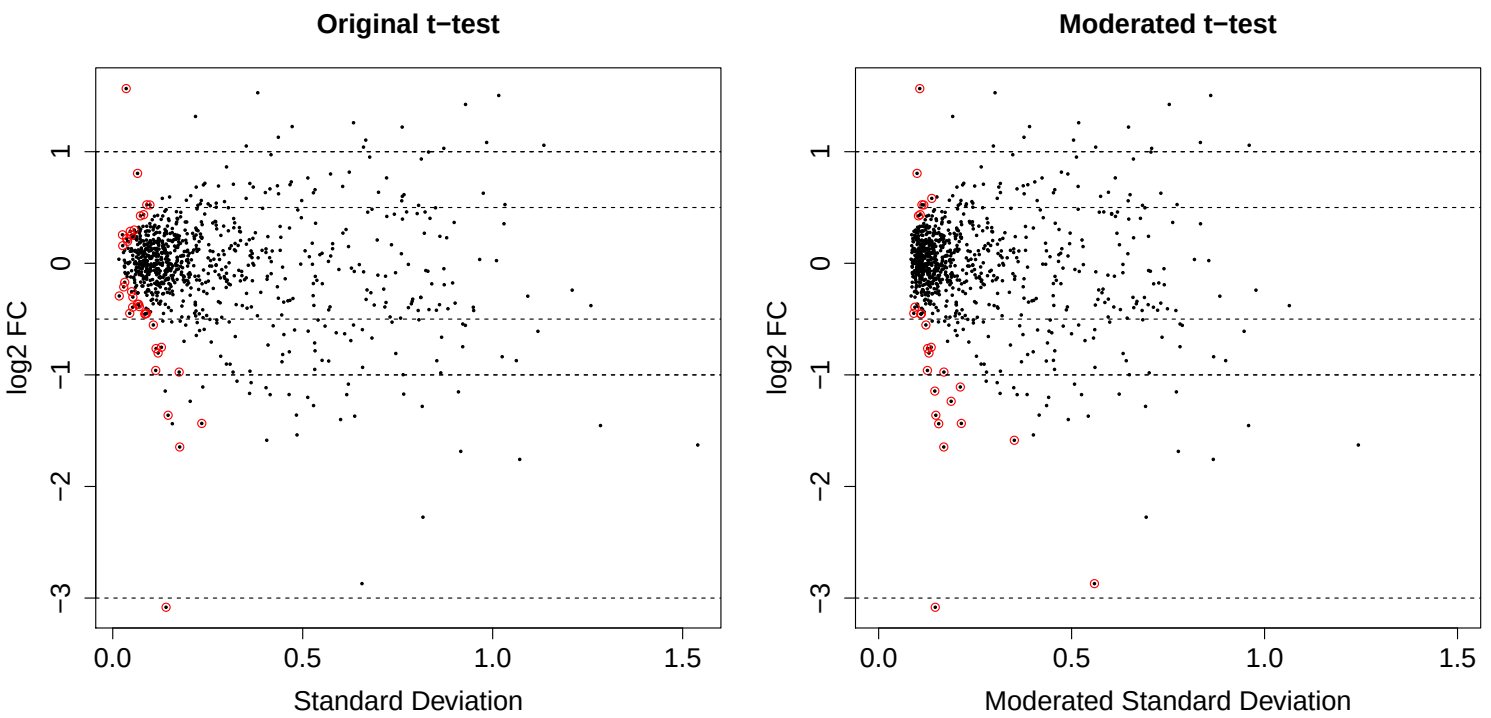
Shrinkage of the variance with limma

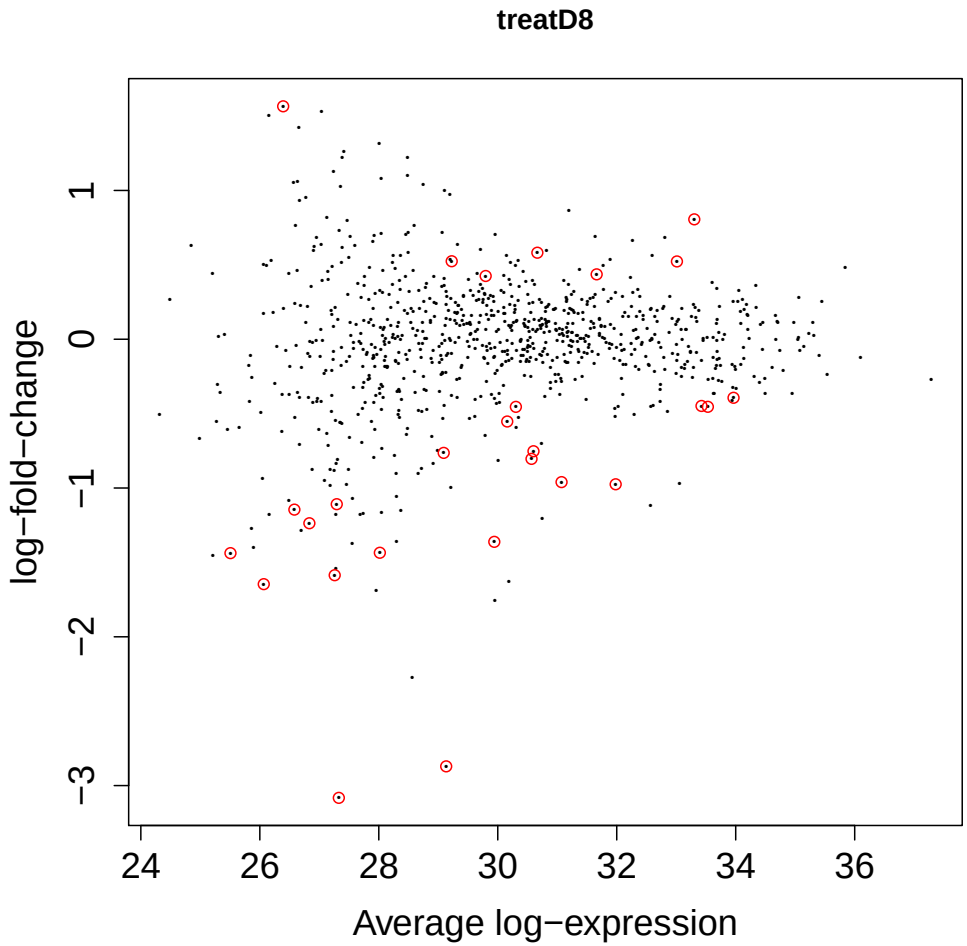


Problems with ordinary t-test solved by moderated EB t-test



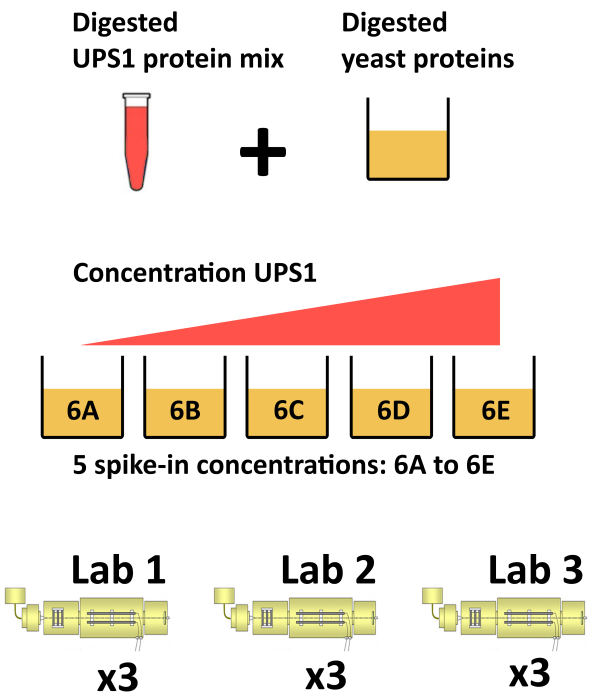
Problems with ordinary t-test solved by moderated EB t-test




treatD8
log−fold−change
1
0
−1
−2
−3
24
26
28
30
32
34
36
Average log−expression

Peptide-based models

Inference with Peptide Based Methods



- Protein by protein analysis of peptide level data with linear model

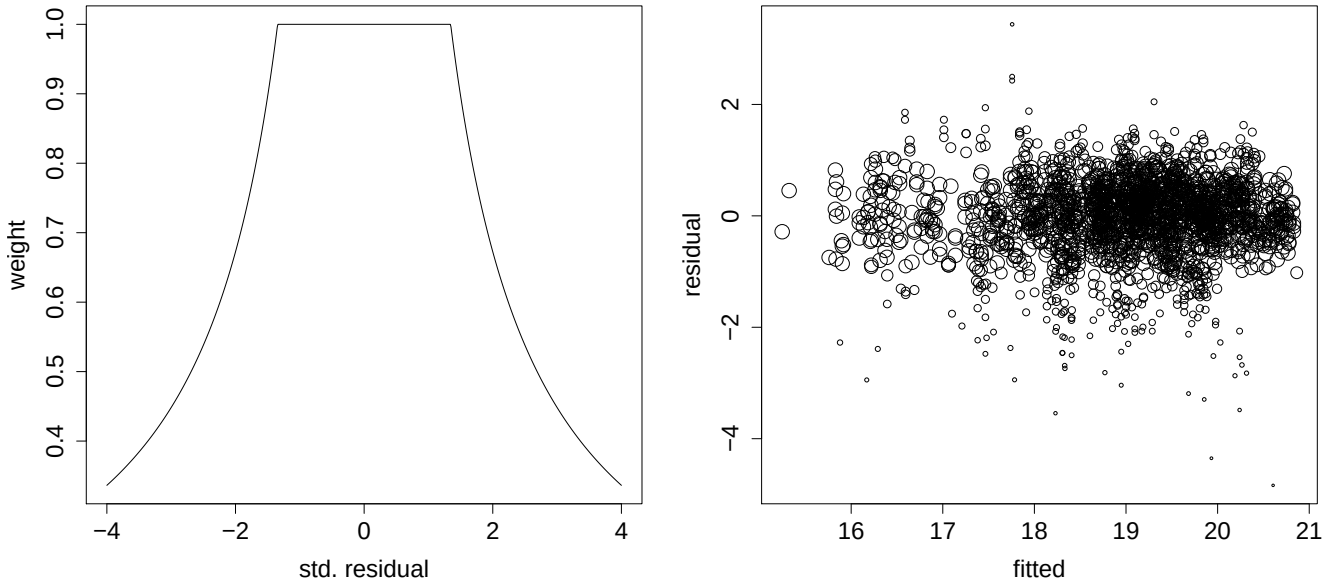
  peptide level — protein level

  $$y_{pept} \sim peptide + treatment + lab$$

- Variance estimation in the literature: protein-wise (LM) or via limma-style EB (LM-Sq).
- t-tests on model parameters

Extension I: Robust estimation using observation weights (Ex I: LM-Sq-Rob)

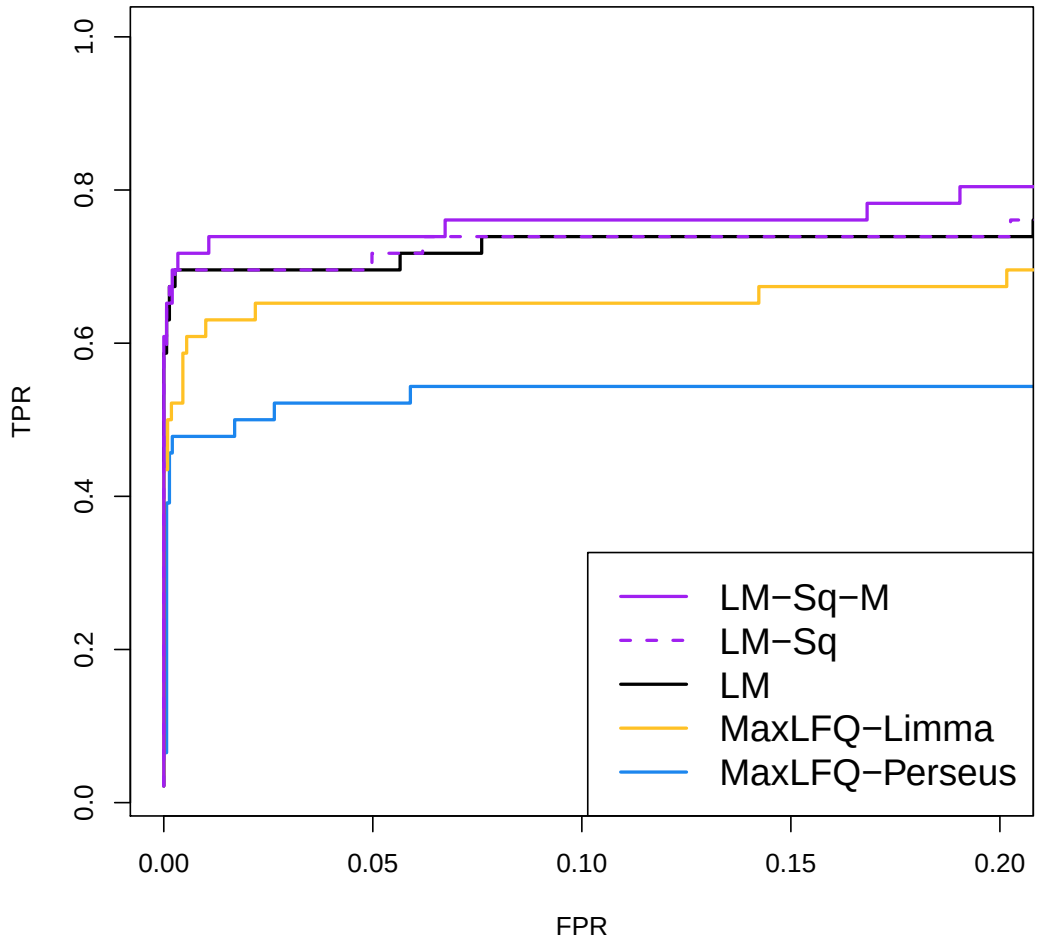
- Outlying peptide intensities: incorrect peptide identification, post-translational modifications, ...



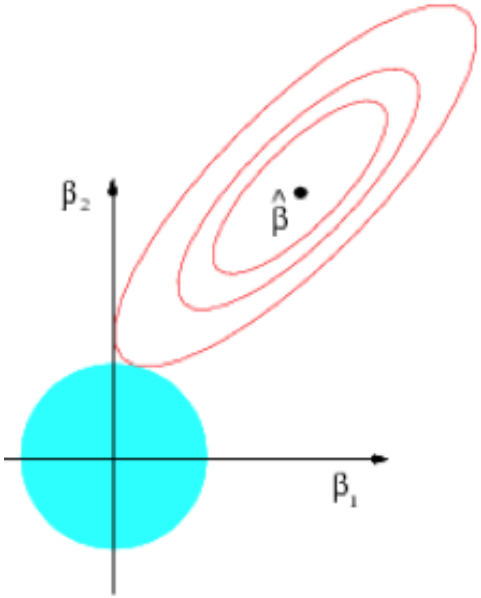
- Iteratively fit model with observation weights $w(d_{ijp})$

$$\text{argmin} \left[\sum_{i=1}^{n} \sum_{p}^{P_j} w(d_{ijp}) \left(y_{ijp} - \mathbf{x}_i^T \boldsymbol{\beta}_j^{\text{treat}} - \beta_{jp}^{\text{pep}} \right)^2 \right]$$

Method performance



Extension II: Ridge regression

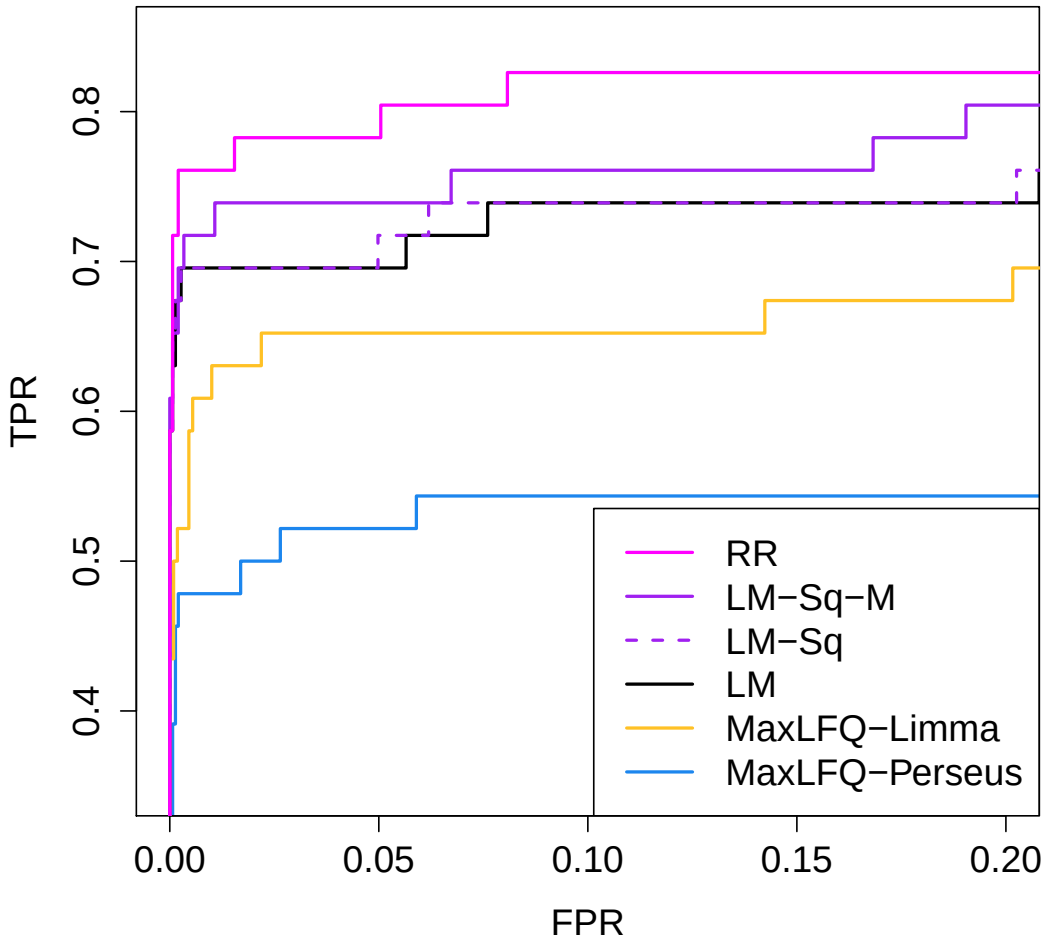


Parameters estimation via ridge regression, loss function:

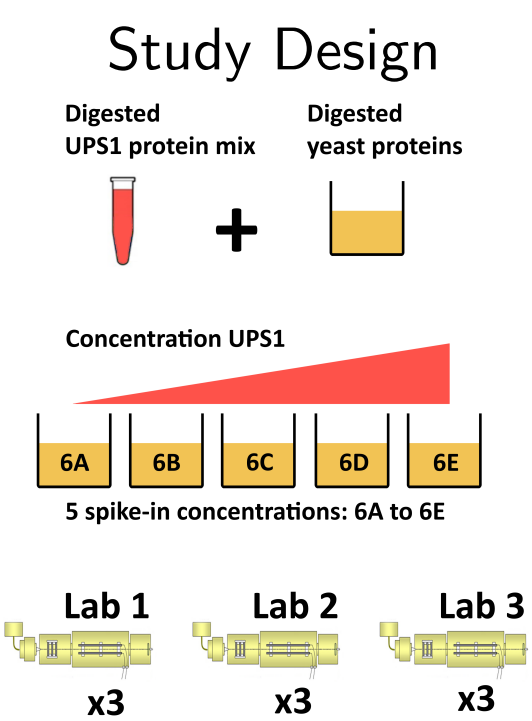
$$\text{argmin} \left[\sum_{i=1}^{n} \sum_{p}^{P_j} w(d_{ijp}) \left(y_{ijp} - \mathbf{X}_i^T \boldsymbol{\beta}_j^{\text{treat}} - \beta_{jp}^{\text{pep}} \right)^2 + \lambda_j^{\text{treat}} \sum \left(\beta_j^{\text{treat}} \right)^2 + \lambda_j^{\text{pep}} \sum \left(\beta_{jp}^{\text{pep}} \right)^2 \right]$$

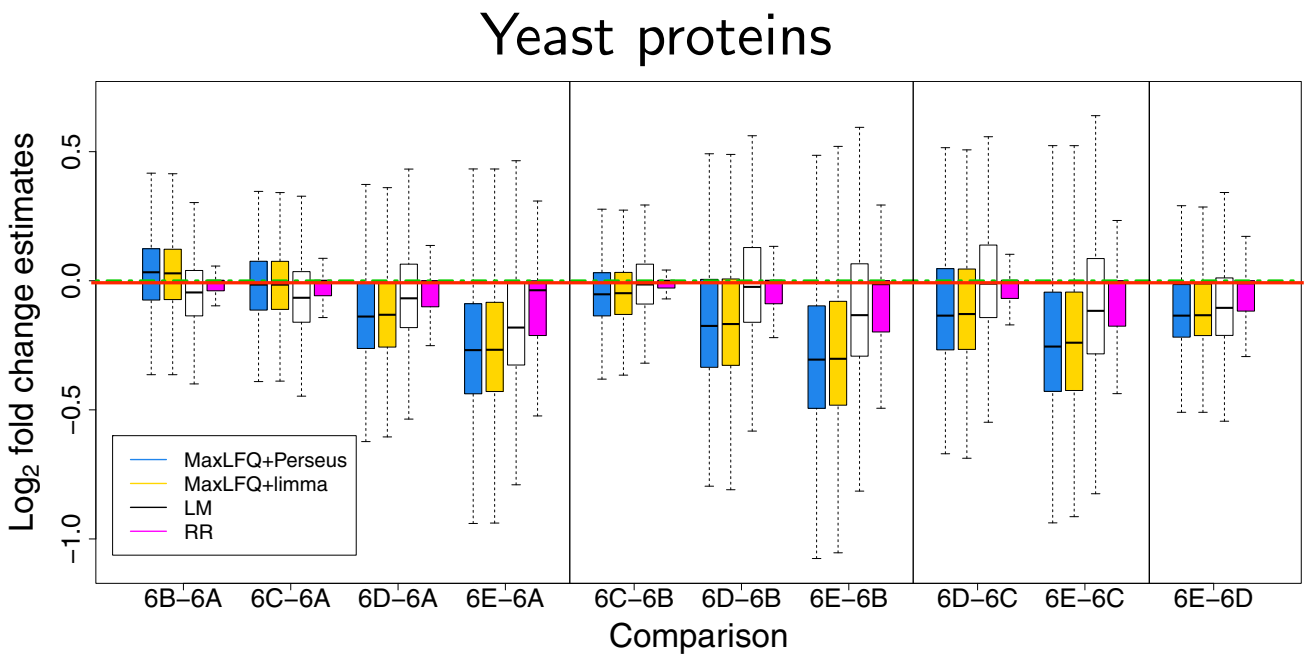
with

- λ_{treat}: penalty term for regularization of parameters of interest
- λ_{pep}: penalty term for regularization of peptide specific parameters


TPR
0.4
0.5
0.6
0.7
0.8
FPR
0.00
0.05
0.10
0.15
0.20
RR
LM−Sq−M
LM−Sq
LM
MaxLFQ−Limma
MaxLFQ−Perseus

Fold Change Estimates: Accuracy & Precision

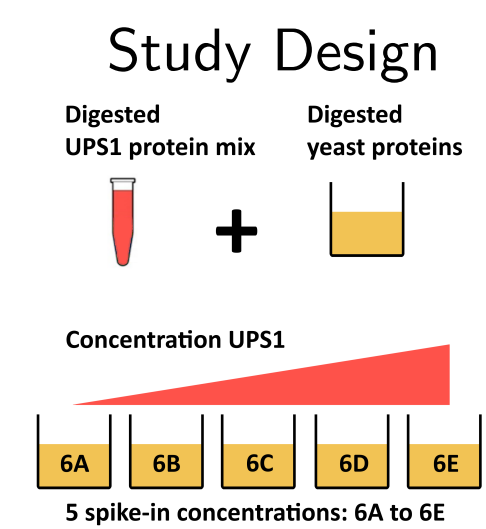


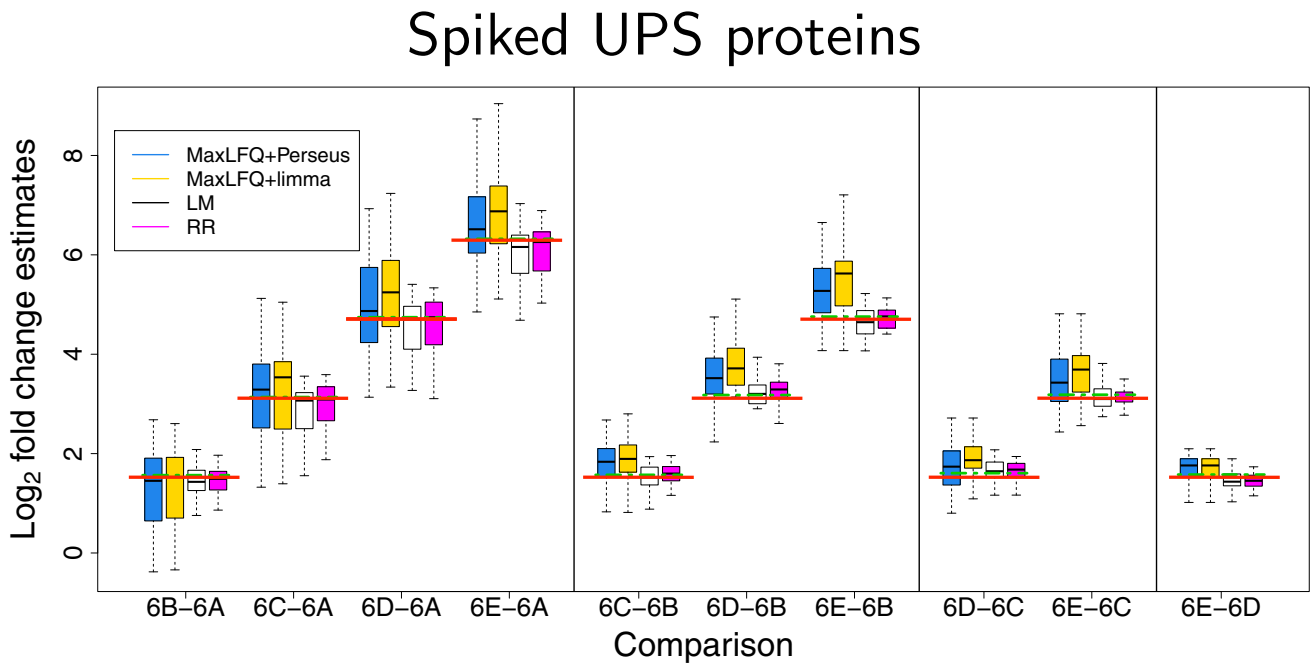


- Shrinkage: more precise and accurate FC estimates
- Note, negative bias of the log2 FC estimates as spike-in concentration increases

  Ionization suppression effects + Violation of normalization assumptions

Fold Change Estimates: Accuracy & Precision





- MaxLFQ- Perseus and MaxLFQ-limma are always more biased and more variable
- Again MSqRob has a higher precision
- Shrinkage does not affect accuracy if there is evidence for DA!

MSqRob

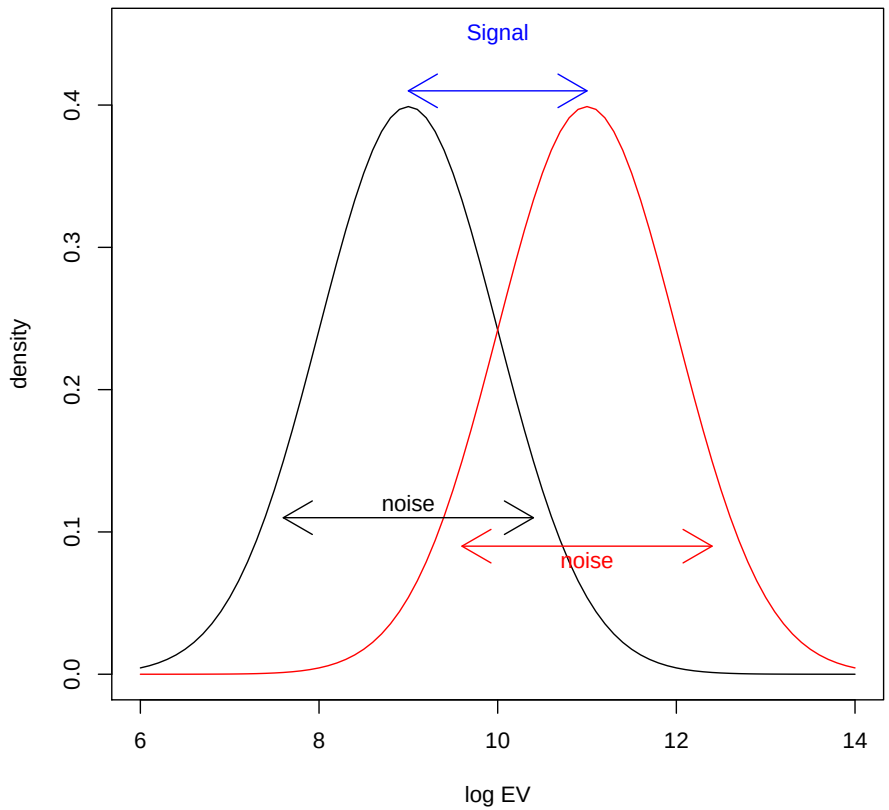
GitHub

https://github.com/statOmics/MSqRob

Goeminne, L., Gevaert, K. and Clement, L. (2016). Molecular and Cellular Proteomics, 15(2), 657-668
Goeminne, L., Gevaert, K. and Clement, L. (2017). Journal of Proteomics, In Press.
http://dx.doi.org/10.1016/j.jprot.2017.04.004

Experimental Design

Power?



$$\Delta = \bar{z}_{p1} - \bar{z}_{p2}$$

$$T_g = \frac{\Delta}{\text{se}_\Delta}$$

$$T_g = \frac{\widehat{\text{signal}}}{\widehat{\text{Noise}}}$$

If we can assume equal variance in both treatment groups:

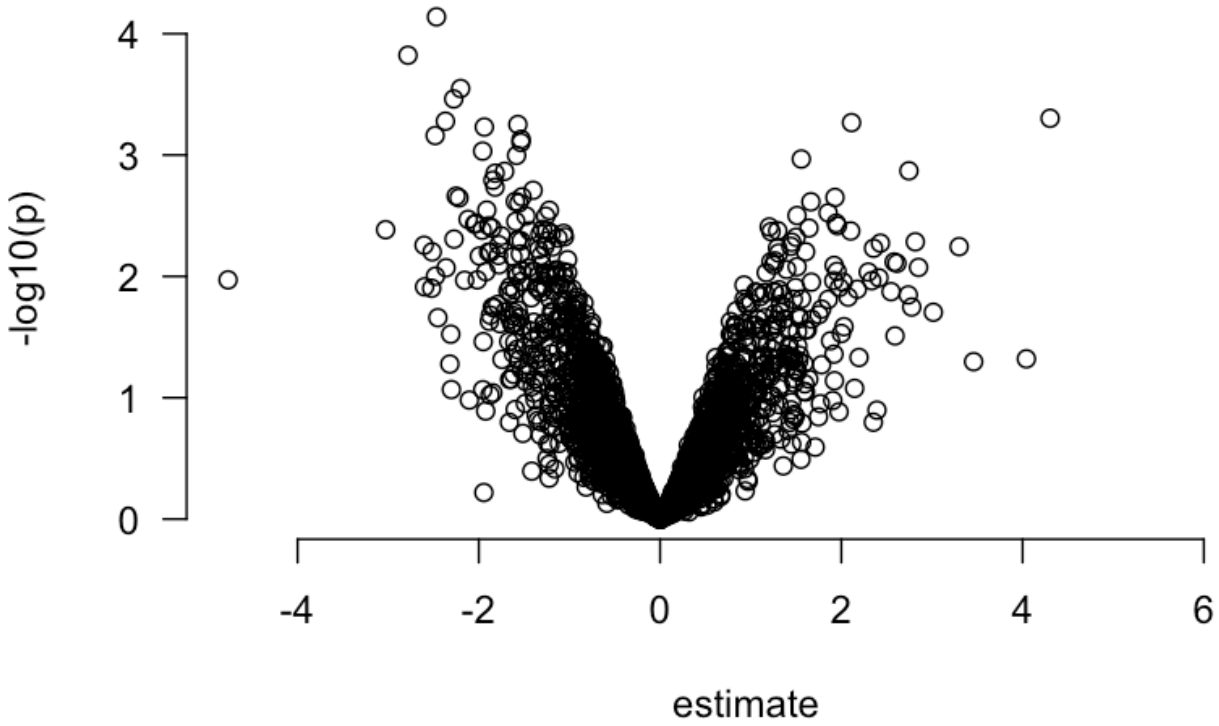
$$\text{se}_\Delta = \text{SD}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$\rightarrow$ Design: if number of bio-repeats increases we have a higher power!

- Study on tamoxifen treated Estrogen Receptor (ER) positive breast cancer patients
- Proteomes for tumors of patients with good and poor outcome upon recurrence.

- Study on tamoxifen treated Estrogen Receptor (ER) positive breast cancer patients
- Proteomes for tumors of patients with good and poor outcome upon recurrence.

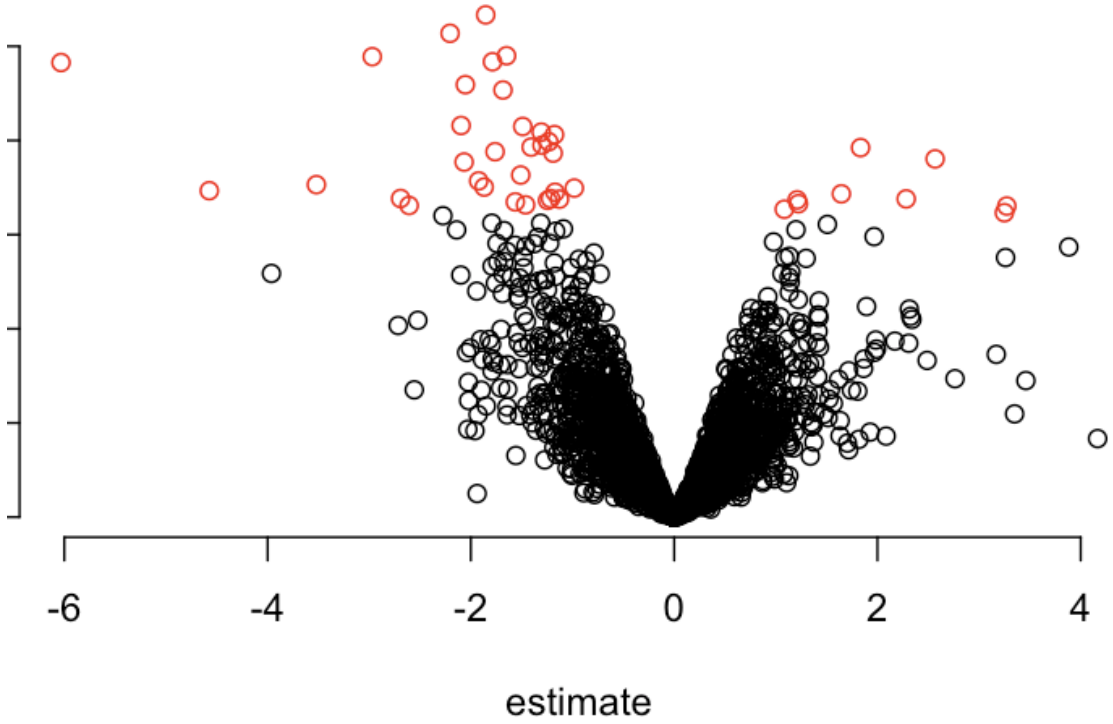
3 vs 3



0 proteins at 5% FDR

- Study on tamoxifen treated Estrogen Receptor (ER) positive breast cancer patients
- Proteomes for tumors of patients with good and poor outcome upon recurrence.

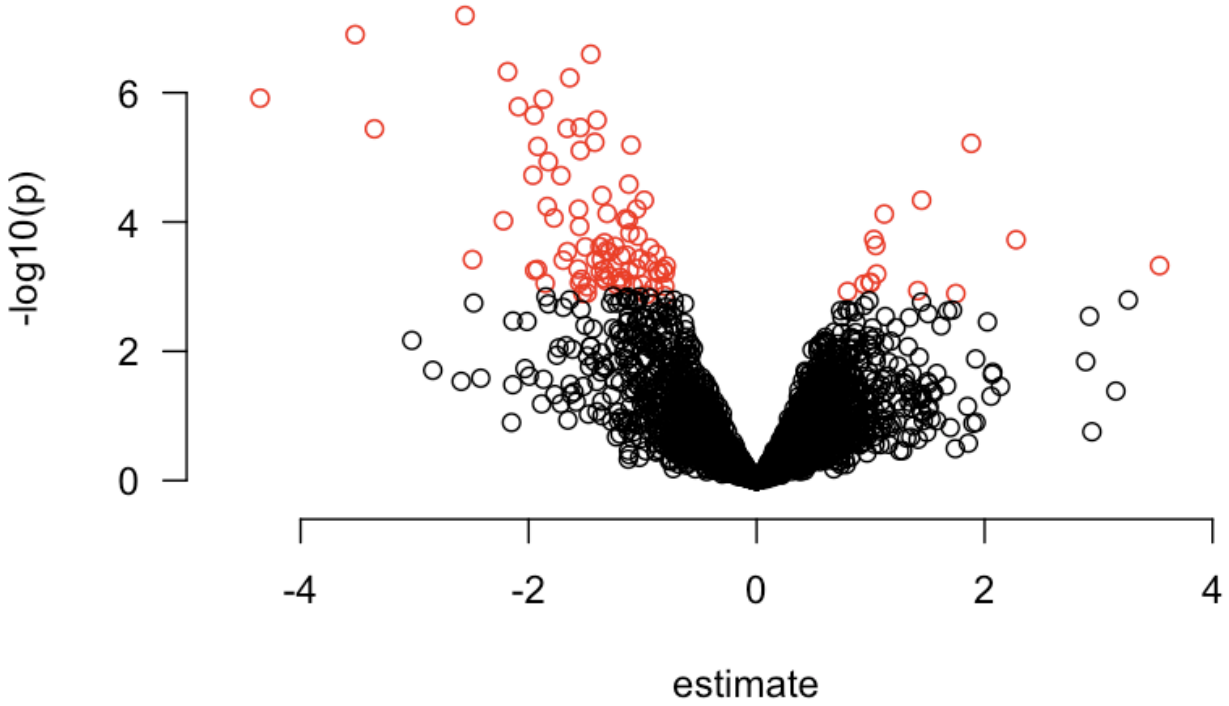
6 vs 6



41 proteins at 5% FDR

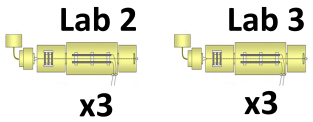
- Study on tamoxifen treated Estrogen Receptor (ER) positive breast cancer patients
- Proteomes for tumors of patients with good and poor outcome upon recurrence.

9 vs 9



96 proteins at 5% FDR

Blocking

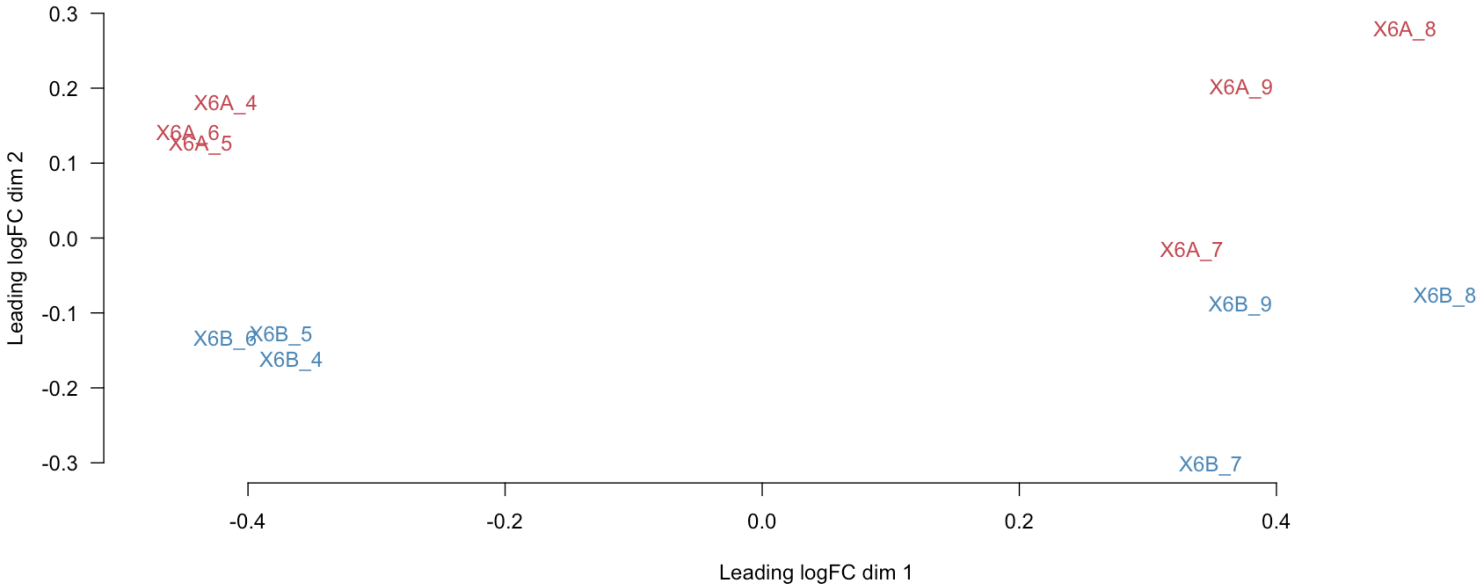


Color variable [?]

condition

MDS plot after full preprocessing [?]

☐ Plot MDS points
☑ Plot MDS labels



Experimental Design: Blocking

Sources of variability

$$\sigma^2 = \sigma^2_{bio} + \sigma^2_{\text{lab}} + \sigma^2_{\text{extraction}} + \sigma^2_{\text{run}} + \dots$$

- Biological: fluctuations in protein level between rats of the same litter, between rats of different litters.
- Technical: cage effect, lab effect, week effect, plasma extraction, MS-run, ...

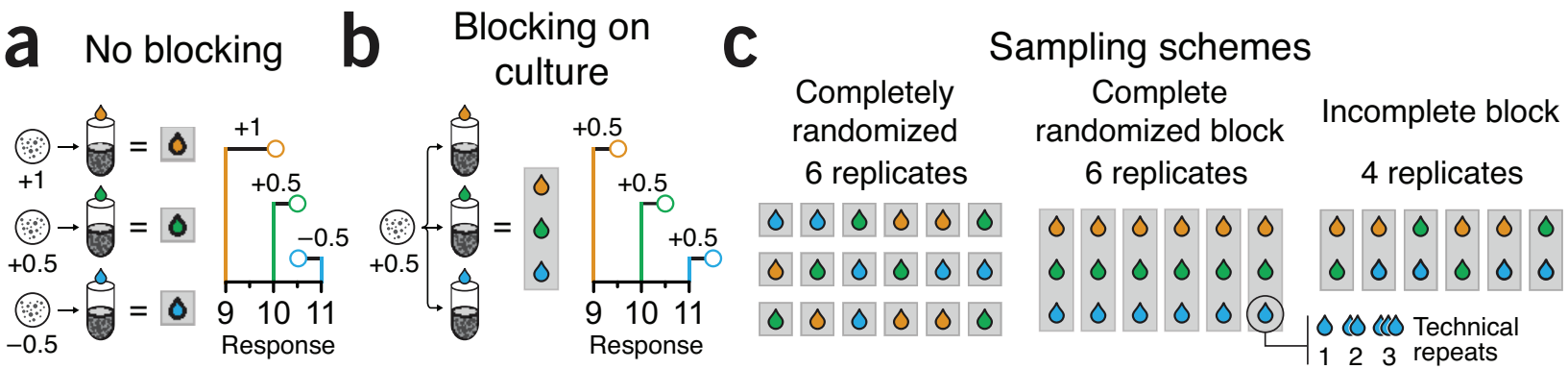


Figure 2 | Blocking improves sensitivity by isolating variation in samples that is independent from treatment effects. (**a**) Measurements from treatment aliquots derived from different cell cultures are differentially offset (e.g., 1, 0.5, –0.5) because of differences in cultures. (**b**) When aliquots are derived from the same culture, measurements are uniformly offset (e.g., 0.5). (**c**) Incorporating blocking in data collection schemes. Repeats within blocks are considered technical replicates. In an incomplete block design, a block cannot accommodate all treatments.

Blocking

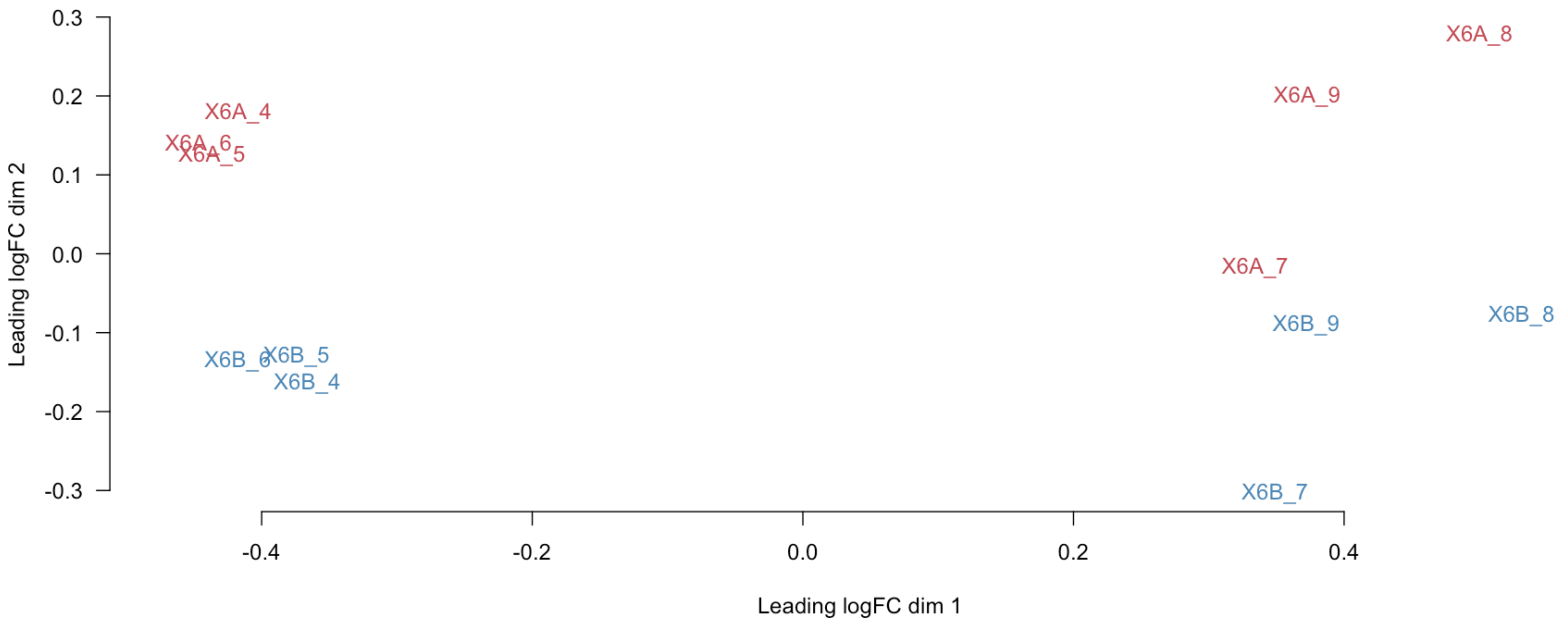
$$\sigma^2 = \sigma^2_{\text{within lab}} + \sigma^2_{\text{between lab}}$$

Color variable [?]

condition

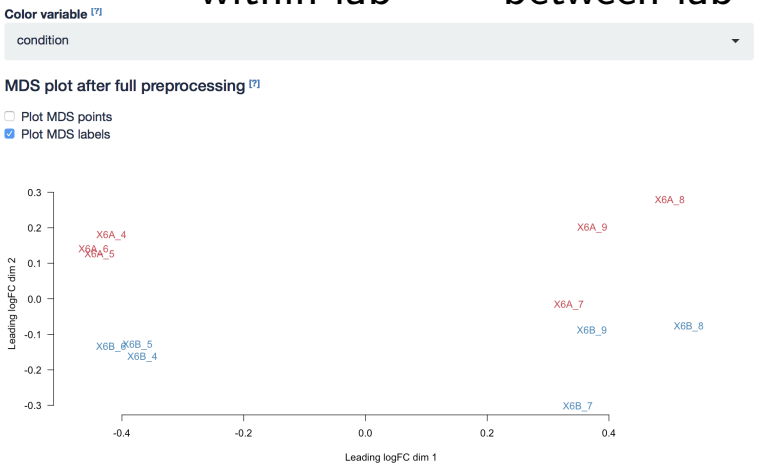
MDS plot after full preprocessing [?]

☐ Plot MDS points
☑ Plot MDS labels



Blocking

$$\sigma^2 = \sigma^2_{\text{within lab}} + \sigma^2_{\text{between lab}}$$



→ All treatments of interest are present within block!

→ We can estimate the effect of the treatment within block!

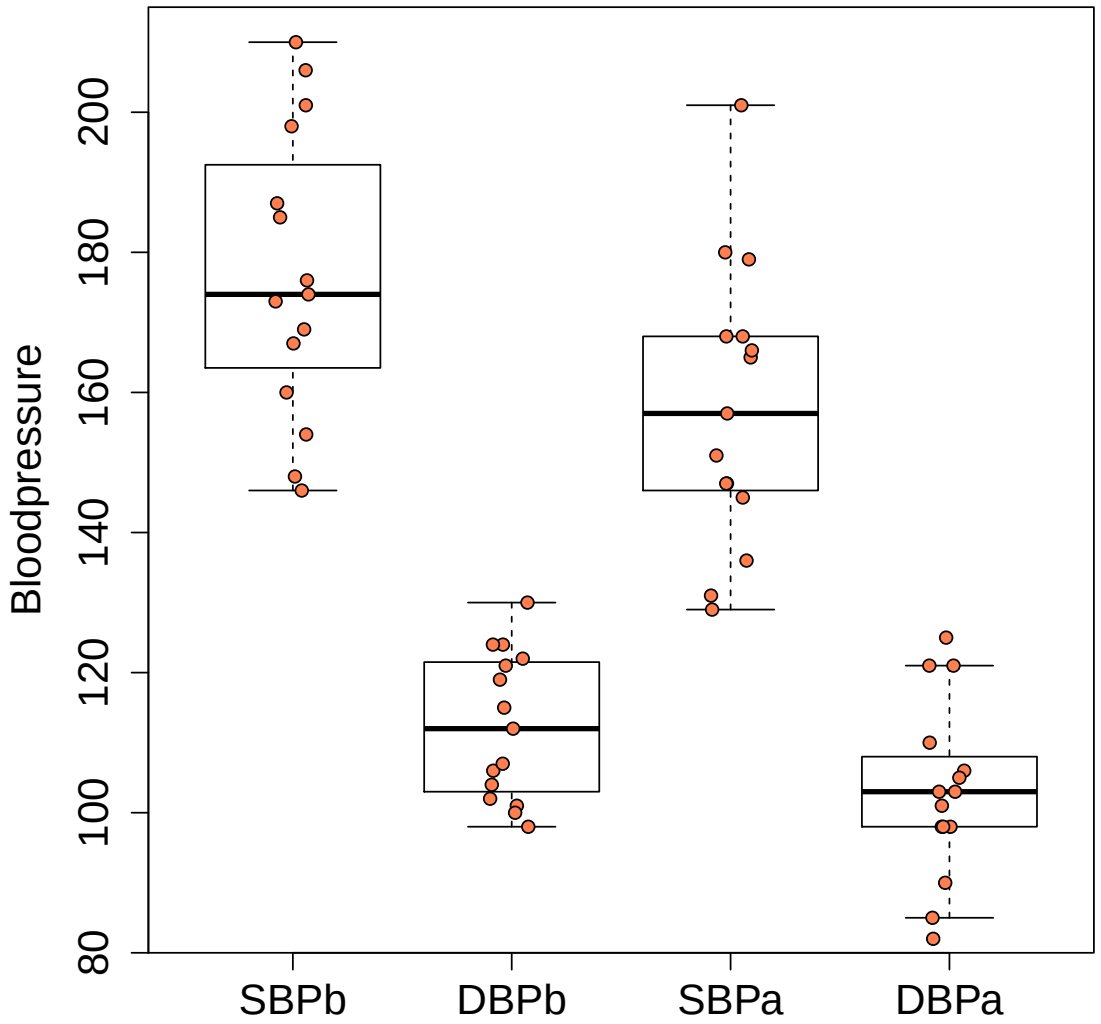
→ We can isolate the between block variability from the analysis

→ linear model:

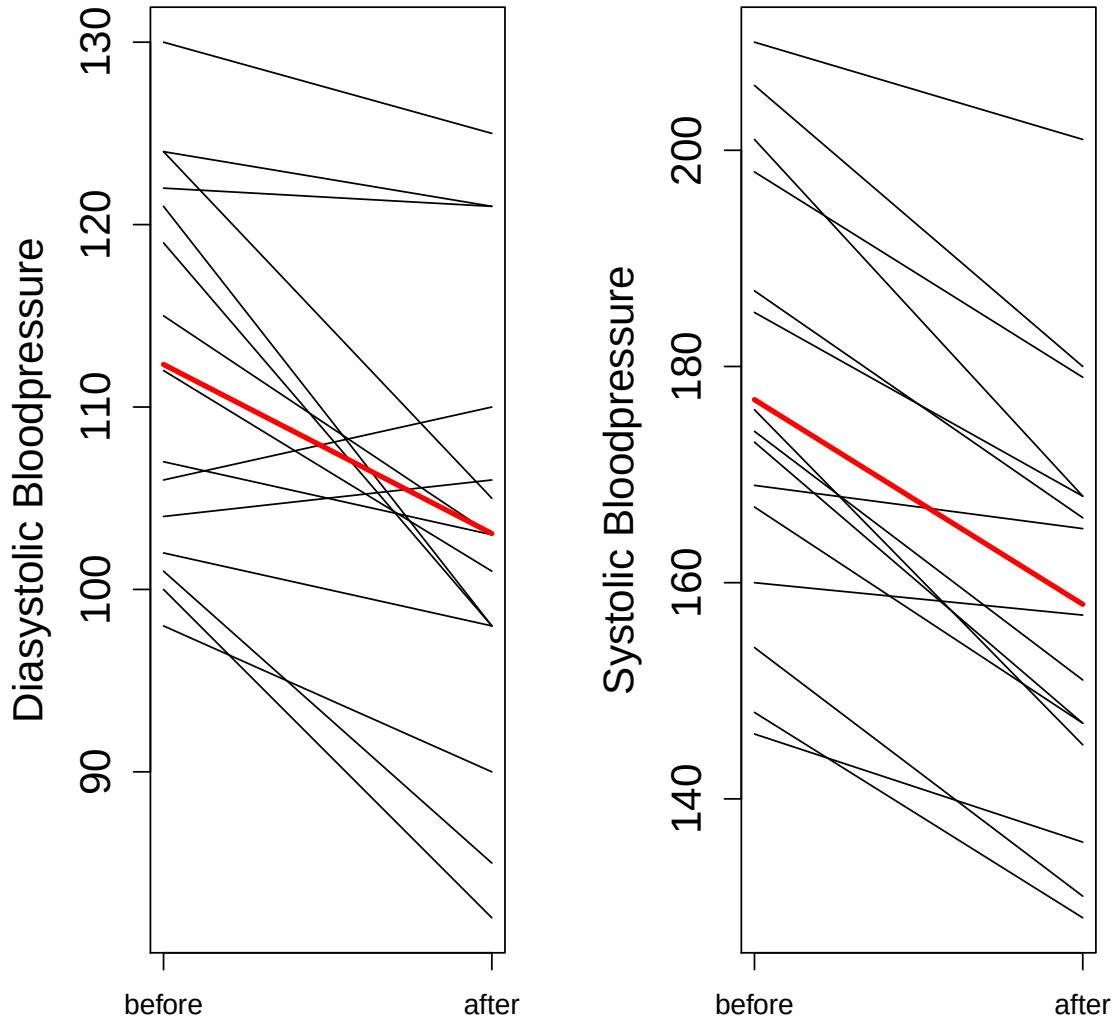
$$y \sim \text{treatment} + \text{lab}$$

→ Not possible with Perseus!

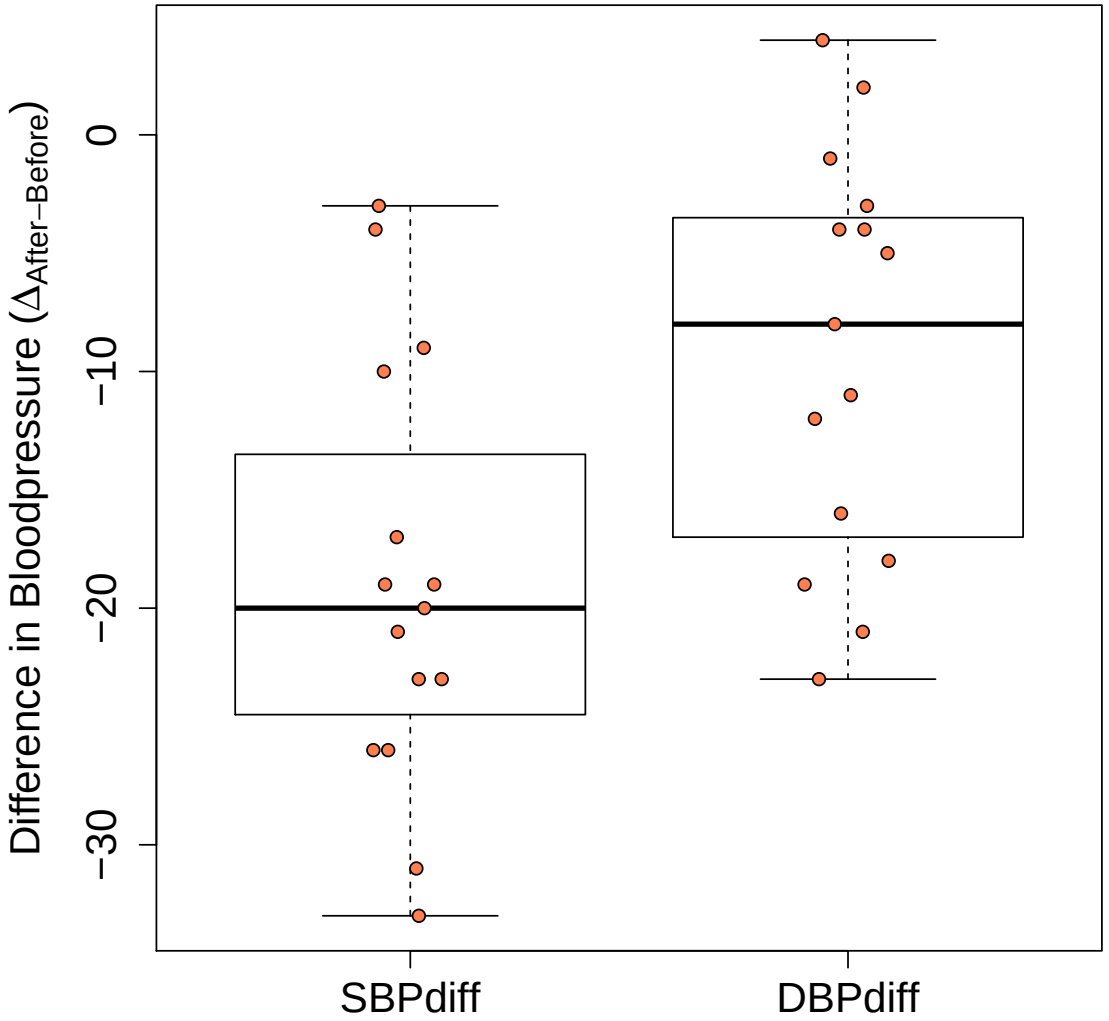
Power gain of blocking



Power gain of blocking



Power gain of blocking



Power gain of blocking

- Completely randomized design: 14 people, 7 baseline BP, 7 BP upon treatment.
- Randomized complete block desigh: 7 people, 7 baseline BP and BP upon treatment.

Power gain of blocking

```
Completely randomized design

Call:
lm(formula = bp ~ treat, data = captoprilCRD)

Residuals:
    Min      1Q  Median      3Q     Max
-26.714 -11.643  -3.929  11.179  30.857

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  179.143      7.036  25.461 8.19e-12
treatT       -23.429      9.950  -2.355   0.0364

(Intercept) ***
treatT      *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.62 on 12 degrees of freedom
Multiple R-squared:  0.316,Adjusted R-squared:  0.259
F-statistic: 5.544 on 1 and 12 DF,  p-value: 0.03641
```

Power gain of blocking

```
Randomized complete block design

Call:
lm(formula = bp ~ treat + patient, data = captoprilRCB)

Residuals:
   Min     1Q Median     3Q    Max
    -8     -3      0      3      8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  213.000      5.442  39.138 1.86e-08
treatT       -15.000      3.848  -3.898 0.008004
patientp2    -38.500      7.200  -5.348 0.001749
patientp3    -29.000      7.200  -4.028 0.006896
patientp4    -47.000      7.200  -6.528 0.000617
patientp5    -48.500      7.200  -6.737 0.000521
patientp6    -45.000      7.200  -6.250 0.000777
patientp7    -29.000      7.200  -4.028 0.006896

(Intercept) ***
treatT      **
patientp2   **
patientp3   **
patientp4   ***
patientp5   ***
patientp6   ***
patientp7   **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Power gain of blocking

```
Randomized complete block bad analysis

Call:
lm(formula = bp ~ treat, data = captoprilRCB)

Residuals:
    Min      1Q  Median      3Q     Max
-19.143 -11.643  -1.143   5.357  36.857

Coefficients:
            Estimate Std. Error t value
(Intercept)  179.143      6.694  26.763
treatT       -15.000      9.466  -1.585
            Pr(>|t|)
(Intercept) 4.55e-12 ***
treatT         0.139
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
  0.1 ' ' 1

Residual standard error: 17.71 on 12 degrees of freedom
Multiple R-squared:  0.173,Adjusted R-squared:  0.1041
F-statistic: 2.511 on 1 and 12 DF,  p-value: 0.1391
```