UNIVERSITEIT GENT

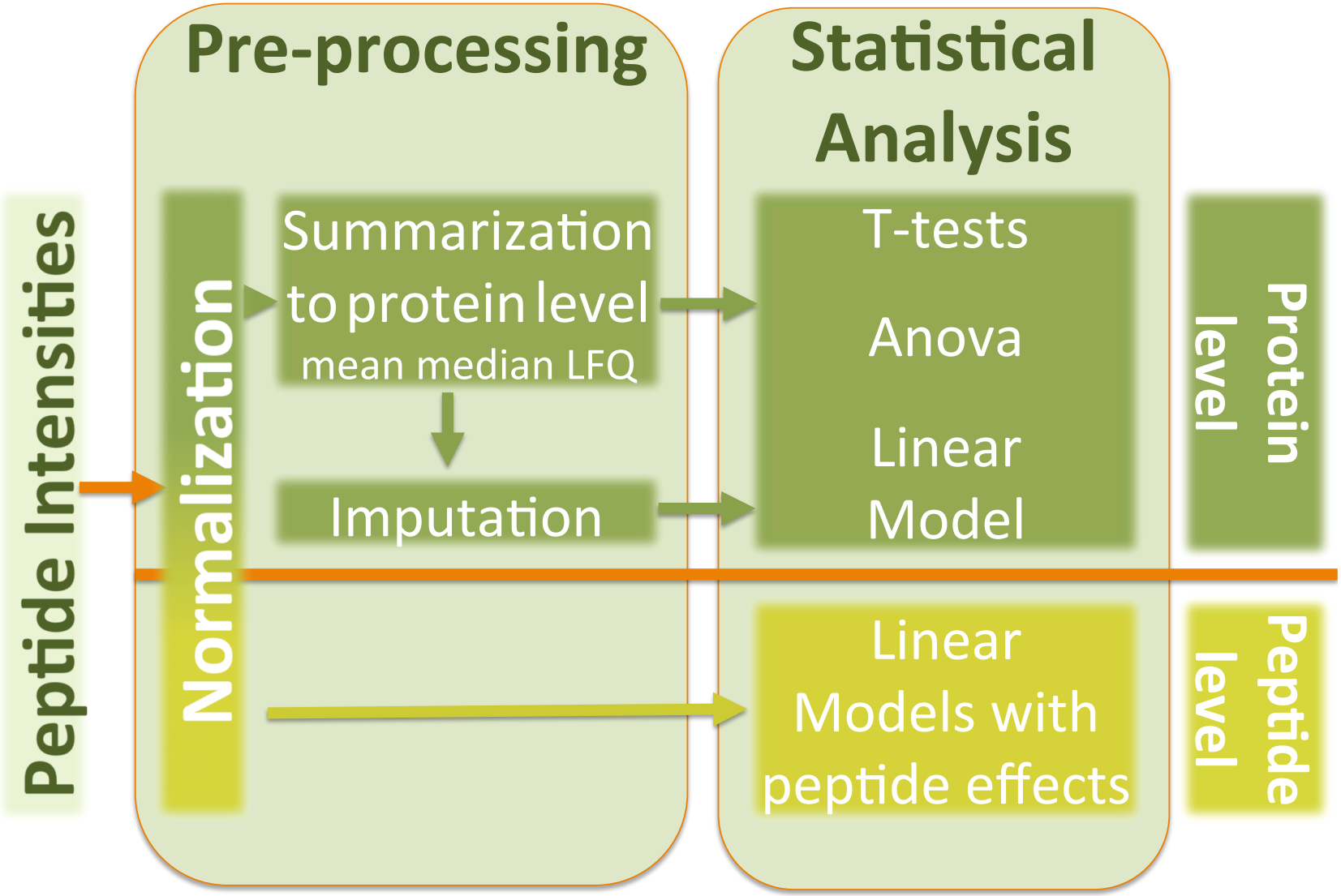
BIG N2N
Bioinformatics Institute Ghent
from nucleotides to networks

Part II: Statistical Inference

Lieven Clement

Proteomics Data Analysis Shortcourse

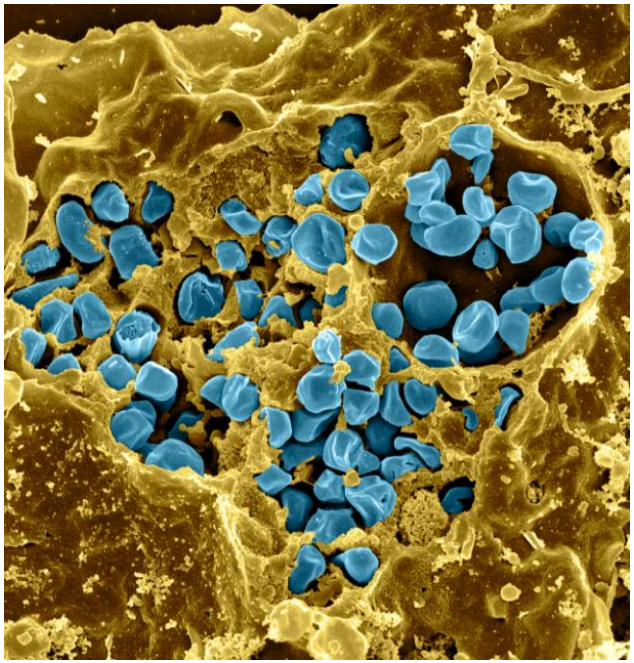
Label-free Quantitative Proteomics Data Analysis Pipelines



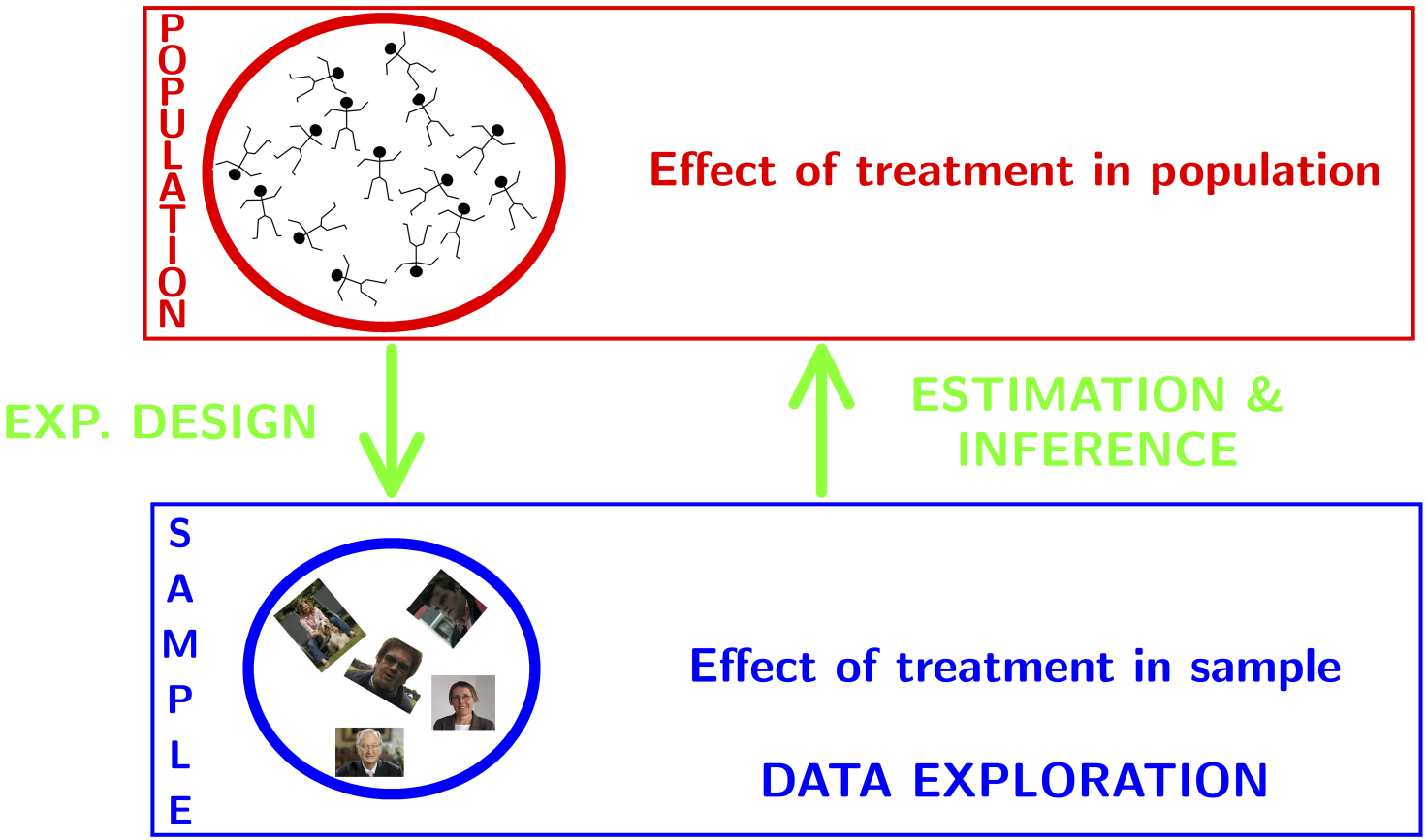
Statistical Inference

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

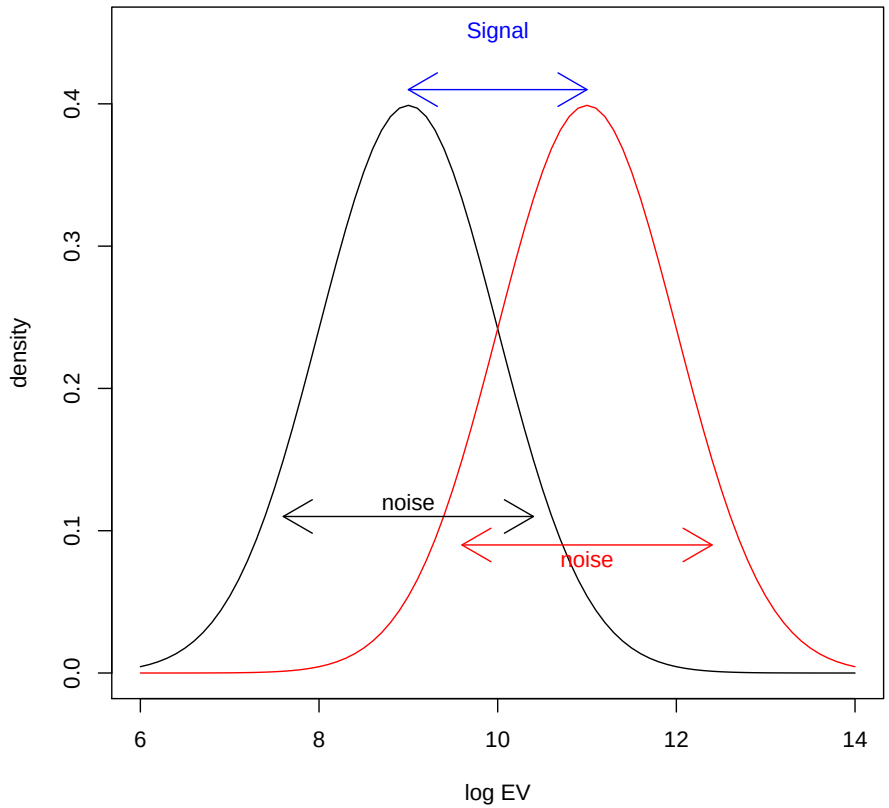

POPULATION
Effect of treatment in population
EXP. DESIGN
ESTIMATION & INFERENCE
SAMPLE
Effect of treatment in sample
DATA EXPLORATION

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



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

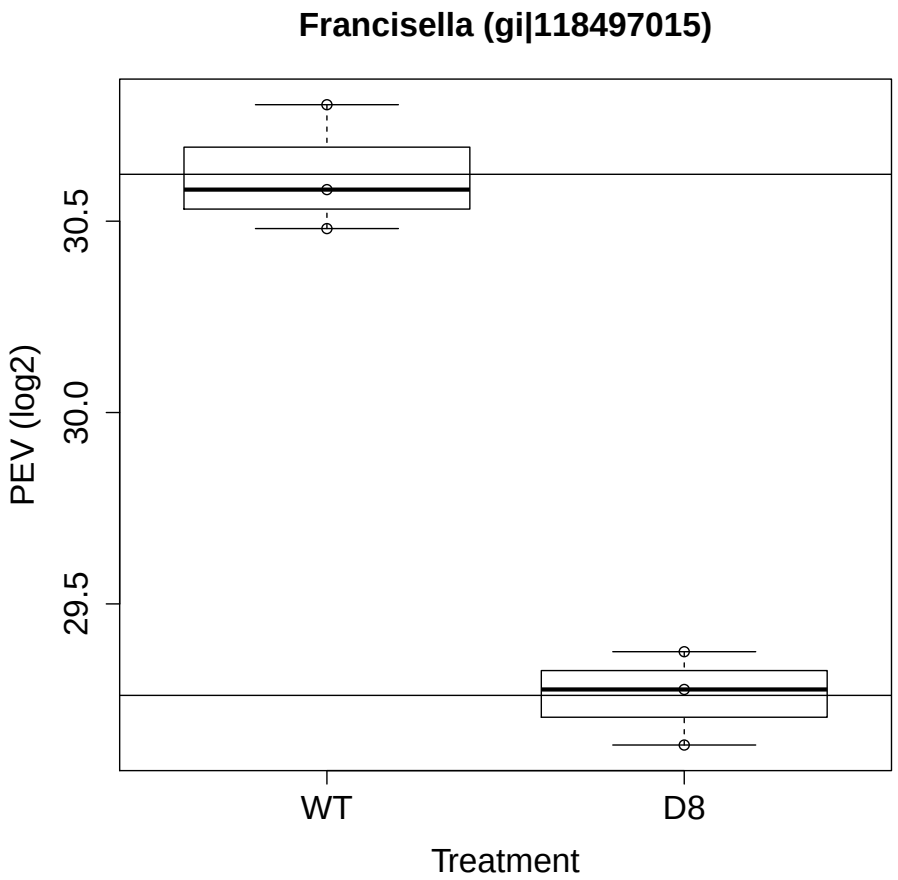
$$T_g = \frac{\Delta}{\text{se}_\Delta}$$

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

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

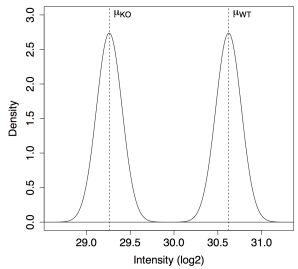
```
Two Sample t-test

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

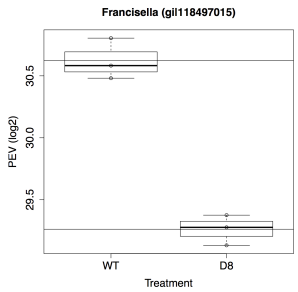
$\hat{\mu}_{wt} = 30.6$, $\hat{\mu}_{ko} = 29.3$, $\hat{FC} = 39\%$

$\hat{\sigma} = 0.14$, 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.

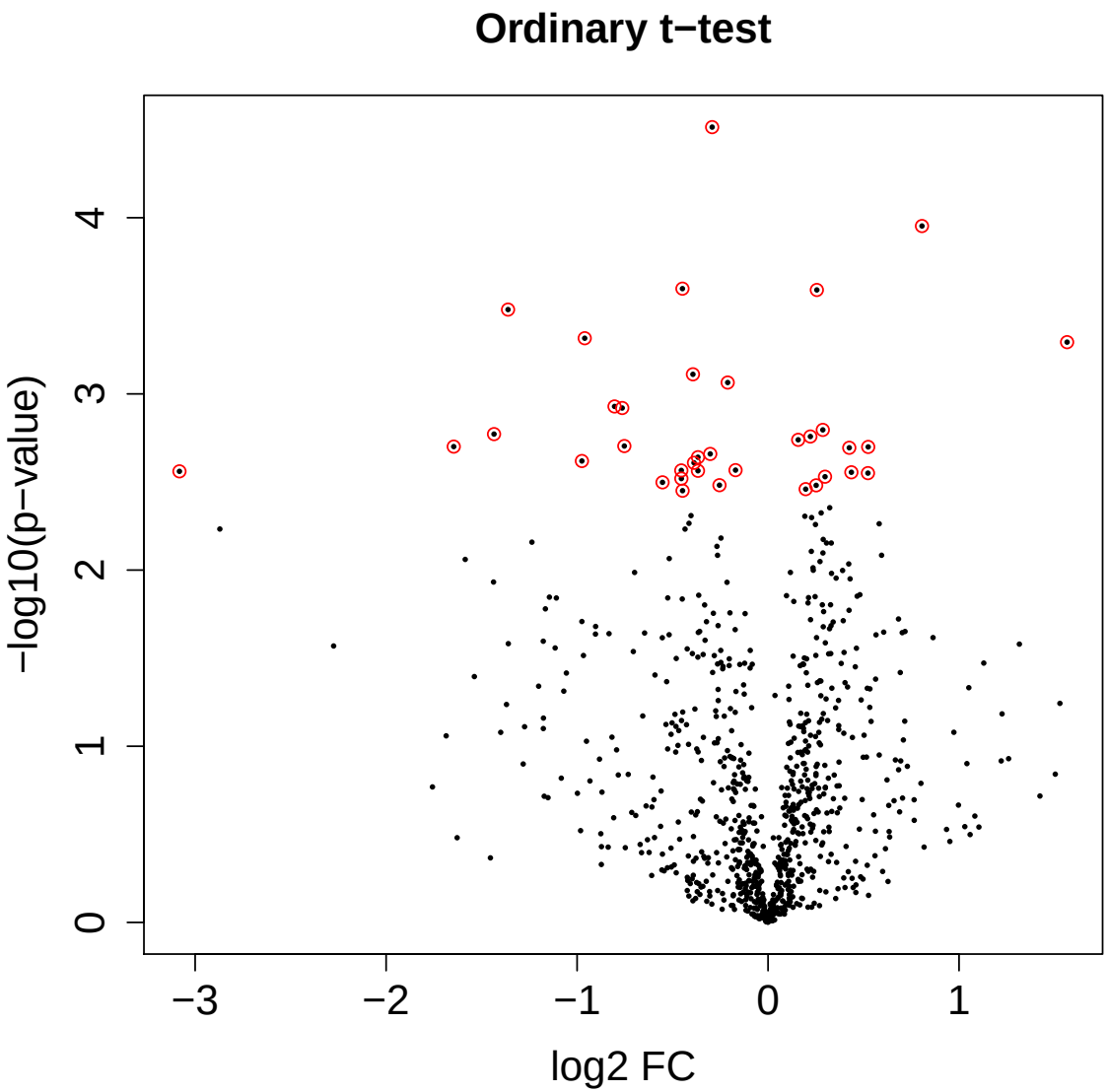
FDR: False discovery rate

- 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

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

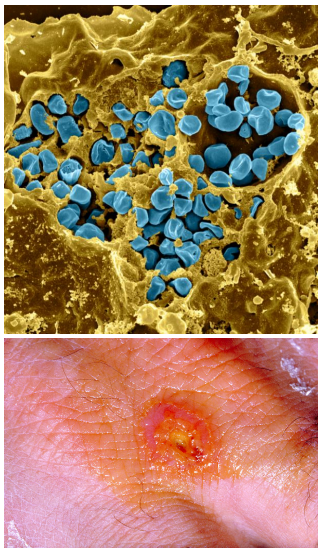
$$\text{FDR}_{\text{BH}}(|t_{\text{thres}}|) = \frac{1 \times p_{t_{\text{thres}}}}{\frac{\#|t_i| \geq t_{\text{thres}}}{m}}$$

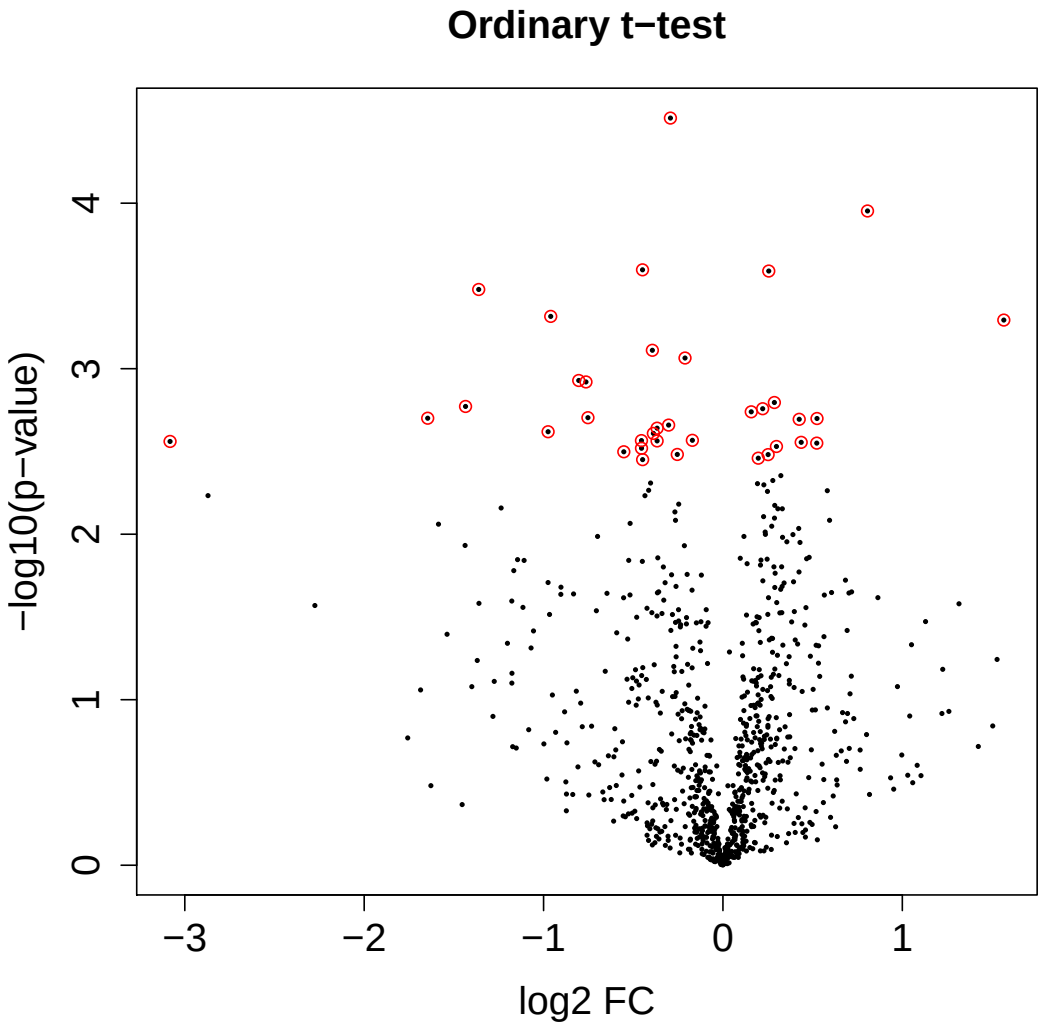
- FDR adjusted p-values can be calculated (e.g. Perseus, R, ...)


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

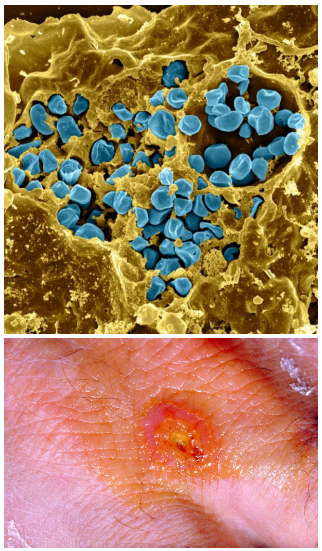
Moderated Statistics

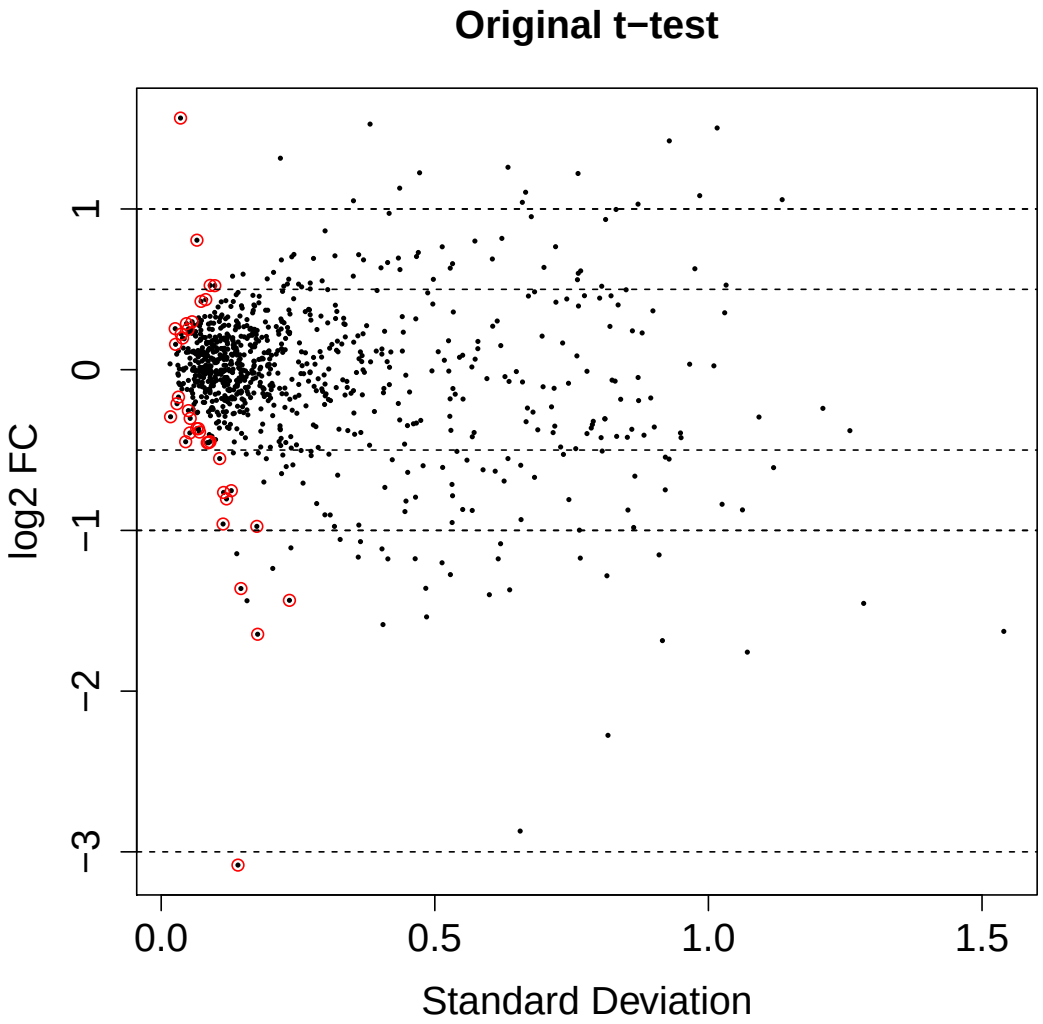
Problems with ordinary t-test





Problems with ordinary t-test





A moderated *t*-test

A general class of moderated test statistics is given by

$$T_g^{mod} = \frac{\bar{Y}_{g1} - \bar{Y}_{g2}}{c\left(\tilde{S}_g\right)},$$

where $\tilde{S}_g$ is a moderated standard deviation estimate.

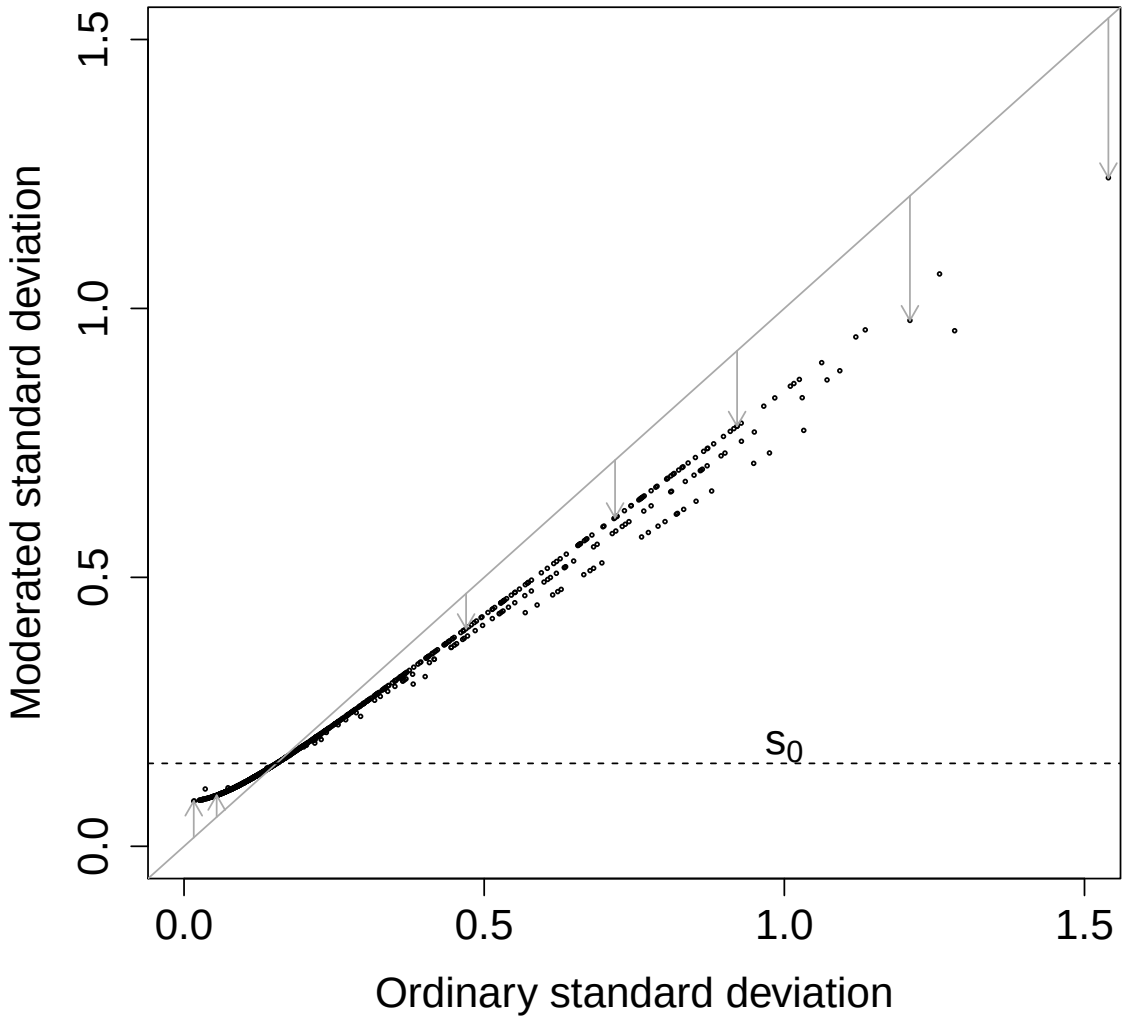
- **empirical Bayes** theory provides formal framework for borrowing strength across genes,
- Implemented in popular bioconductor package **limma**

$$\tilde{S}_g = \sqrt{\frac{d_g S_g^2 + d_0 S_0^2}{d_g + d_0}},$$

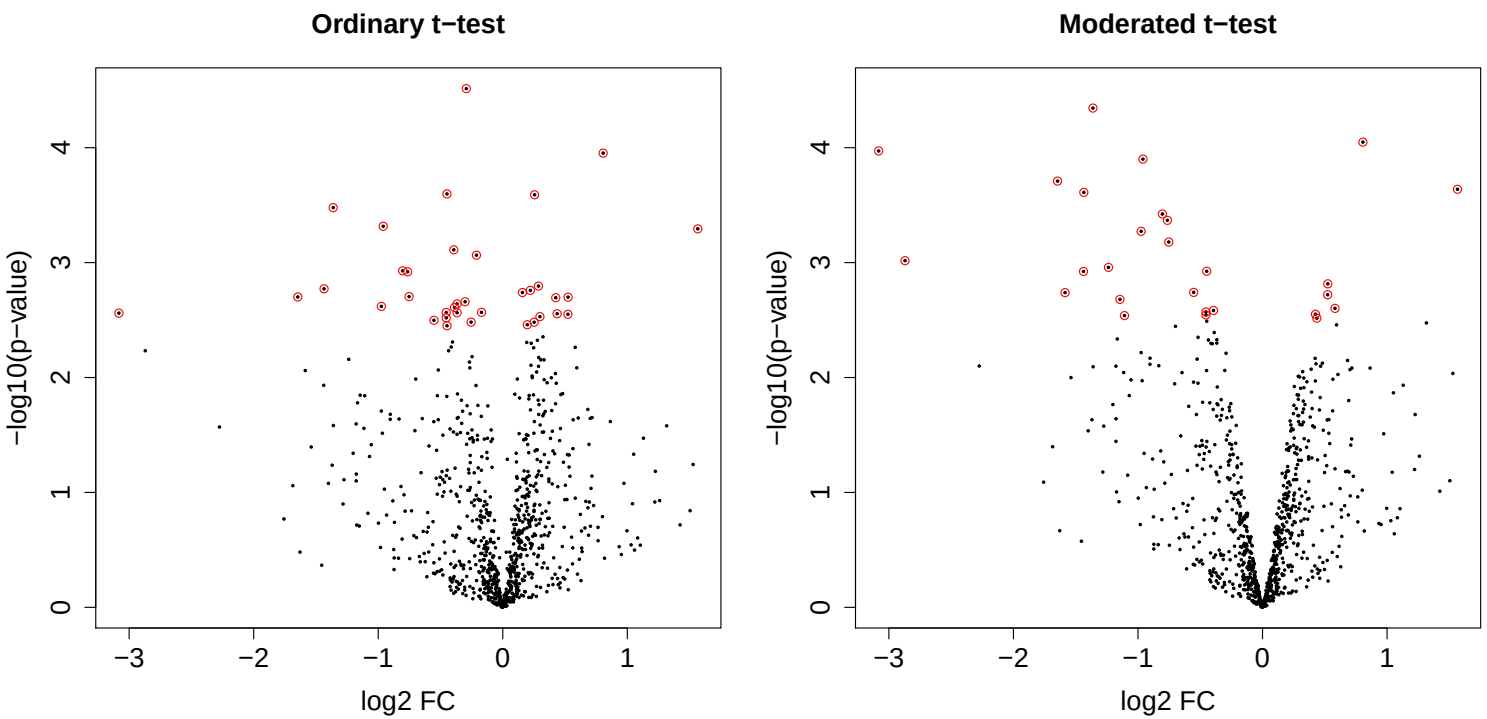
- S_0^2: common variance (over all proteins)
- Moderated t-statistic is t-distributed with $d_0 + d_g$ degrees of freedom.

$\rightarrow$ Note that the degrees of freedom increase by borrowing strength across genes!

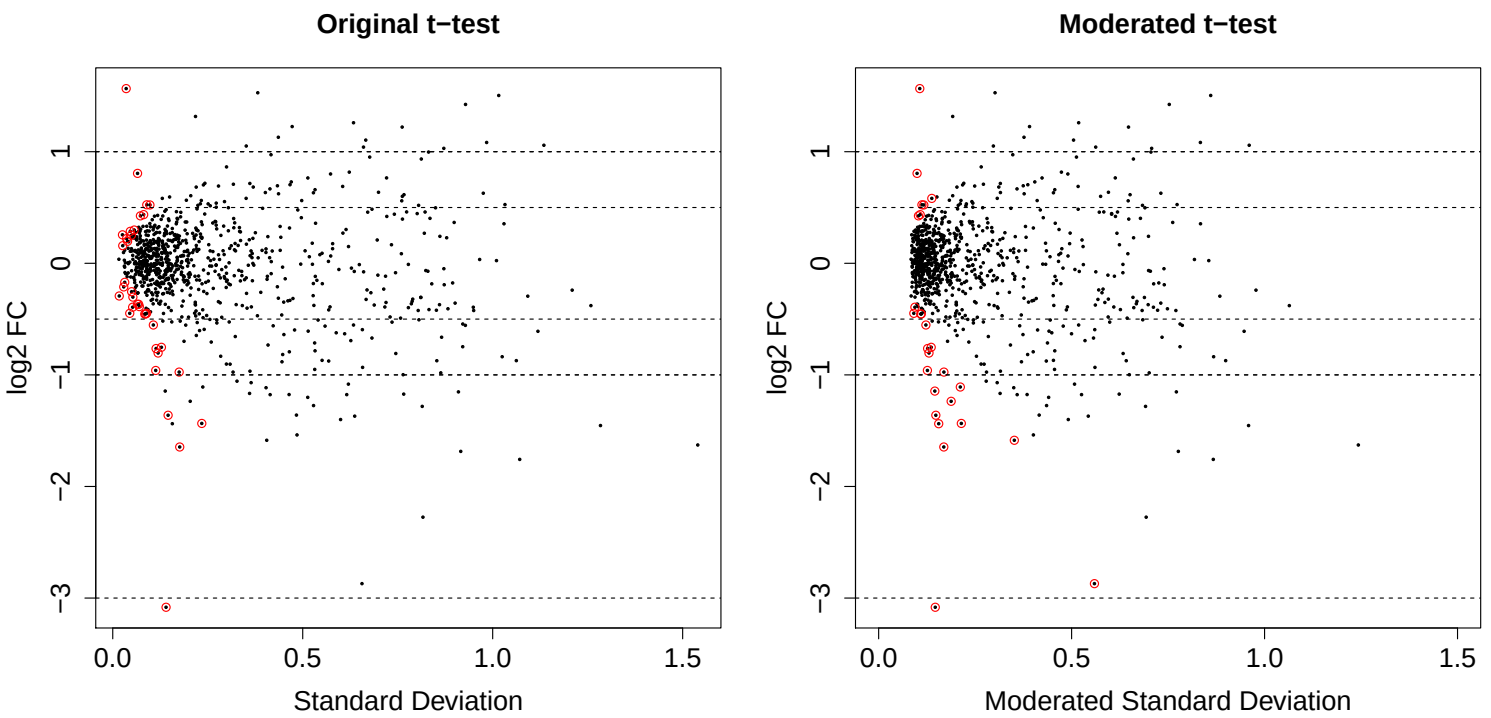
Shrinkage of the variance with limma

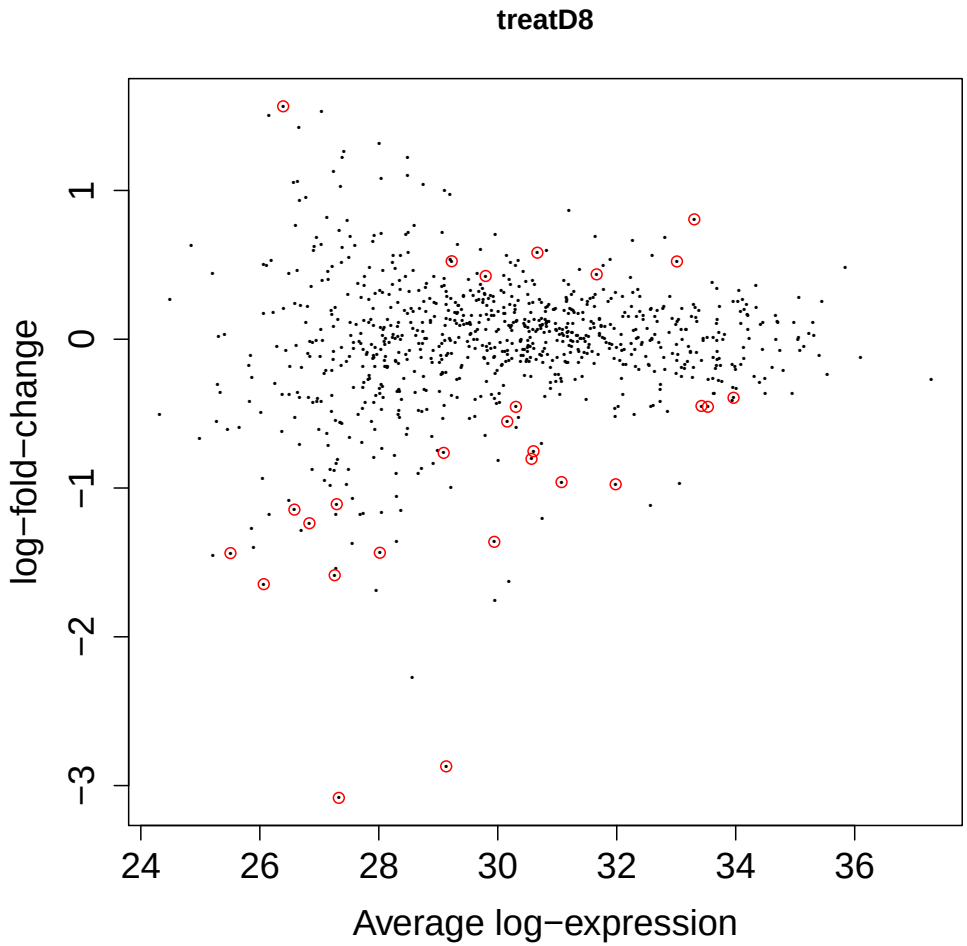


Problems with ordinary t-test solved by moderated EB t-test



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

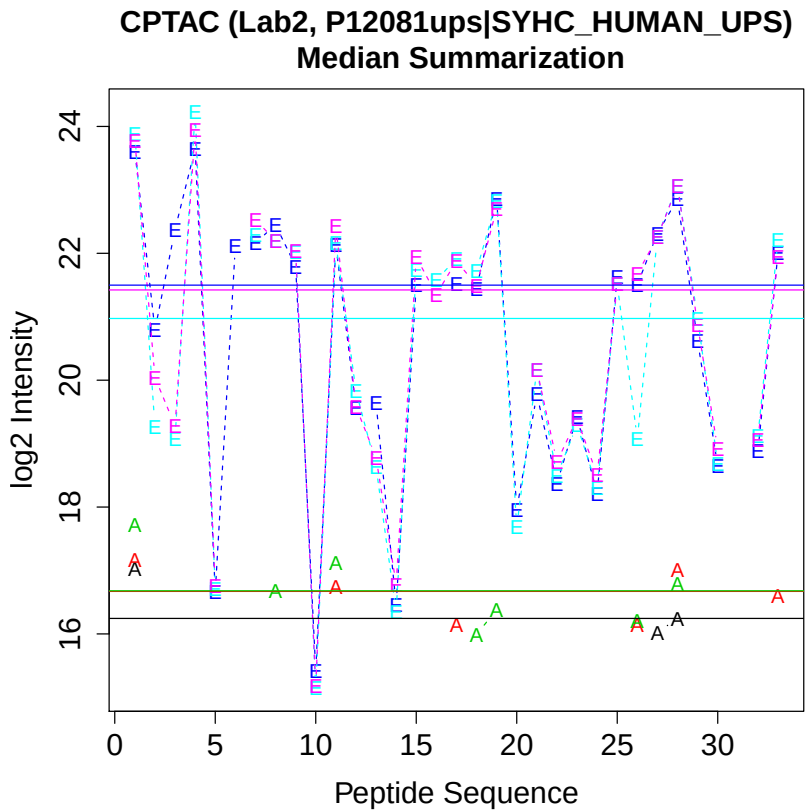



treatD8
log−fold−change
1
0
−1
−2
−3
24
26
28
30
32
34
36
Average log−expression

Peptide-based models

MS measures abundances at peptide level, but we have to assess differences at the protein level

- Pseudoreplication!
- Strong peptide effect
- Unbalanced peptide identification
- Summarization bias
- Different precision of protein level summaries



MSqRob workflow

Linear model on normalized and log2 peptide intensities (y_{glsp})

$$y_{glsp} = \beta_g^{group} + \beta_l^{lab} + u_s^{\text{samp}} + \beta_p^{\text{pep}} + \epsilon_{sp}$$

protein-level

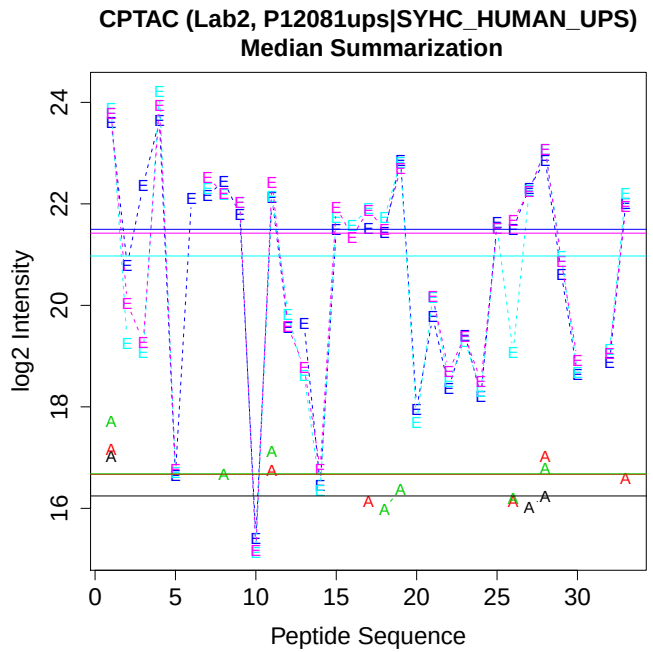
- β_g^{group} and β_l^{lab} related to spike-in and lab
- random sample effect
  $u_s^{\text{samp}} \sim N\left(0, \sigma_s^2\right)$
  $\rightarrow$ Addresses pseudo-replication

peptide-level

- peptide specific effect β_p^{pep}
- within sample error $\epsilon_{ip} \sim N\left(0, \sigma_\epsilon^2\right)$

Estimation

1. Outliers $\rightarrow$ robust regression
2. Penalisation on $\boldsymbol{\beta}^{\text{treat}}$



Linear model on normalized and log2 peptide intensities (y_{glsp})

$$y_{glsp} = \beta_g^{group} + \beta_l^{lab} + u_s^{\text{samp}} + \beta_p^{\text{pep}} + \epsilon_{sp}$$

- Fitting this model is computationally cumbersome
- Inference with linear mixed models is more difficult
- We therefore use a two step procedure to fit the model:
  1. Fit peptide specific model to summarize at peptide intensities into protein level summaries at the sample level

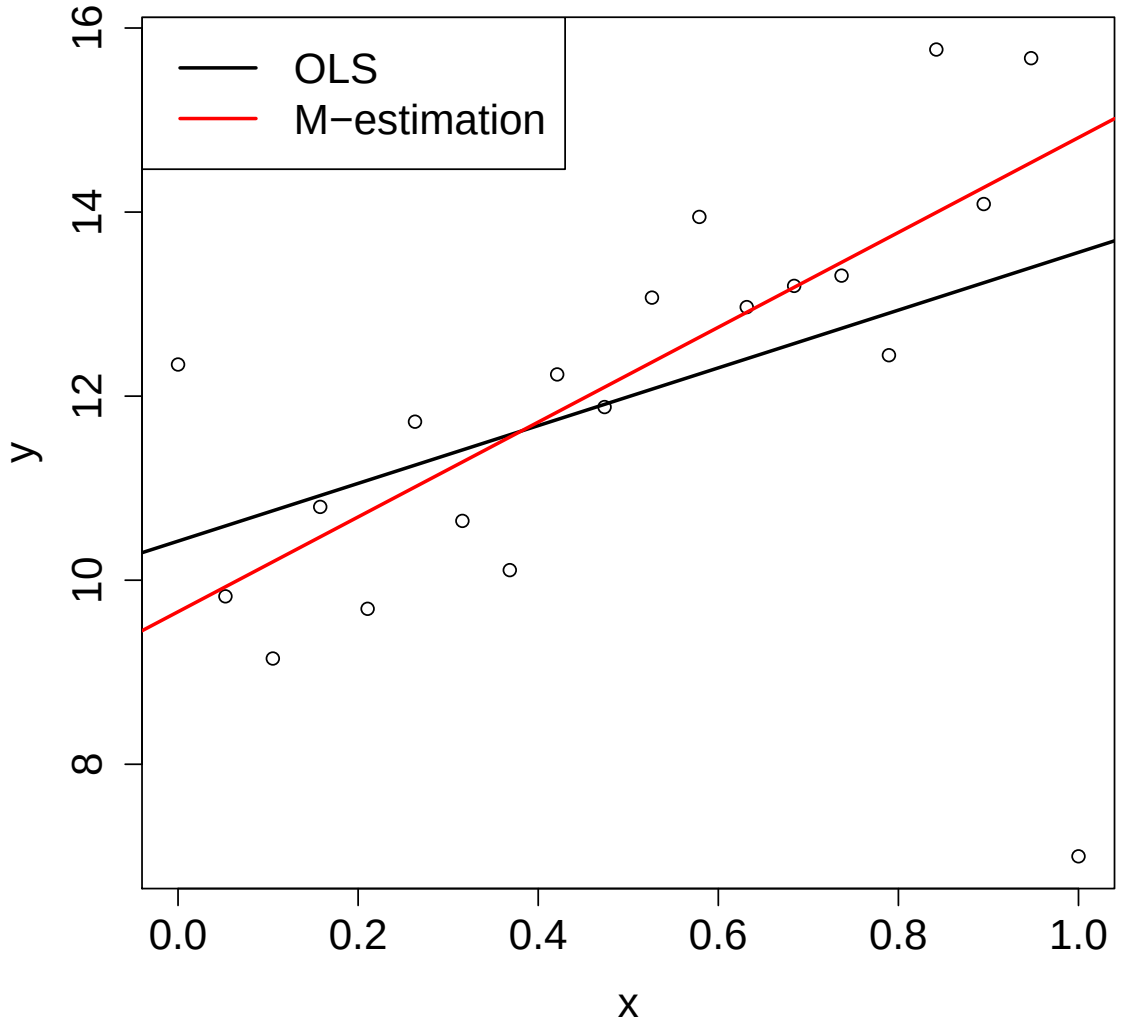
$$y_{sp} = z_s^{\text{samp}} + \beta_p^{\text{pep}} + \epsilon_{sp}$$

  2. Fit linear model to assess effect of treatment

$$z_s = X_s\beta + \epsilon_s$$

- ϵ_s lumps between and within sample variance.
- Fit both models using robust regression
- Advantage: much faster, more straightforward to explain
- Disadvantage: difference in precision of protein summaries is not accounted for
- Performance is very similar to peptide level model

Robust regression



Robust regression

- Robust fit minimises the maximal bias of the estimators
- CI and statistical tests are based on asymptotic theory
- If ϵ is normal, the M-estimators have a high efficiency!
- ordinary least squares (OLS): minimize loss function

$$\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2$$

- M-estimation: minimize loss function

$$\sum_{i=1}^{n}\rho\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right)$$

with

- ρ is symmetric, i.e. $\rho(z) = \rho(-z)$
- ρ has a minimum at $\rho(0) = 0$, is positive for all $z \neq 0$
- $\rho(z)$ increases as $|z|$ increases

The estimator $\hat{\beta}$ is also the solution to the equation

$$\sum_{i=1}^{n} \Psi(y_i - \mathbf{x}_i\boldsymbol{\beta}) = 0,$$

where Ψ is the derivative of ρ. For $\hat{\beta}$ possessing the robustness property, Ψ should be bounded.

Example: least squares
$\rho(z) = z^2$, and thus $\Psi(z) = 2z\frac{\partial z}{\partial \beta}$ (unbounded!). $\hat{\boldsymbol{\beta}}$ is the solution of

$$\sum_{i=1}^{n} 2\mathbf{x}_i(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) = 0 \text{ or } \hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbf{y}$$

with $\mathbf{X} = [\mathbf{x}_1 \ldots \mathbf{x}_G]^T$

When a location and a scale parameter, say σ, have to be estimated simultaneously, we write

$$(\hat{\beta}, \hat{\sigma}) = \text{ArgMin}_{\beta,\sigma} \sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x}_i^T\beta}{\sigma}\right) \text{ and } \sum_{i=1}^{n} \Psi\left(\frac{y_i - \mathbf{x}_i^T\beta}{\sigma}\right) = 0.$$

Define $u_i = \frac{y_i - \mathbf{x}_i^T\beta}{\sigma}$. The last estimation equation is equivalent to

$$\sum_{i=1}^{n} w(u_i)u_i = 0,$$

with weight function $w(u) = \Psi(u)/u$. This is the typical form that appears when solving the *iteratively reweighted least squares problem*,

$$(\hat{\beta}, \hat{\sigma}) = \text{ArgMin}_{\mu,\sigma} \sum_{i=1}^{n} w(u_i^{(k-1)}) \left(u_i^{(k)}\right)^2,$$

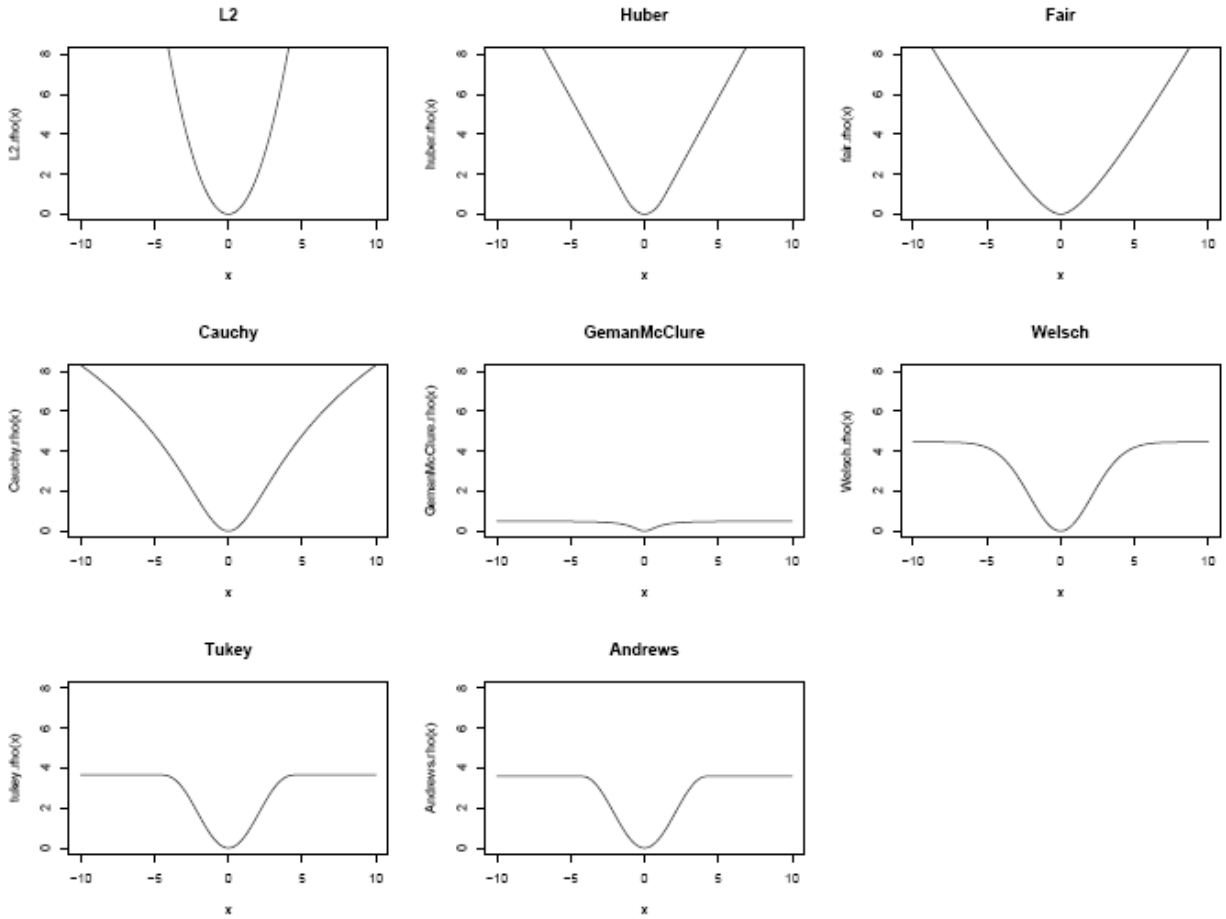
where k represents the iteration number.

Some Examples of Robust Functions

Name	$\rho(x)$	$\psi(x)$	$w(x)$
Huber	$\begin{cases} x^2/2 & \text{if } \lvert x\rvert \le k \\ k(\lvert x\rvert - k/2) & \text{if } \lvert x\rvert > k \end{cases}$	$\begin{cases} x & \text{if } \lvert x\rvert \le k \\ k\,\mathrm{sgn}(x) & \text{if } \lvert x\rvert > k \end{cases}$	$\begin{cases} 1 & \text{if } \lvert x\rvert \le k \\ \frac{k}{\lvert x\rvert} & \text{if } \lvert x\rvert > k \end{cases}$
"Fair"	$c^2\left(\frac{\lvert x\rvert}{c} - \log\left(1+\frac{\lvert x\rvert}{c}\right)\right)$	$\frac{x}{1+\frac{\lvert x\rvert}{c}}$	$\frac{1}{1+\frac{\lvert x\rvert}{c}}$
Cauchy	$\frac{c^2}{2}\log\left(1+(x/c)^2\right)$	$\frac{x}{1+(x/c)^2}$	$\frac{1}{1+(x/c)^2}$
Geman-McClure	$\frac{x^2/2}{1+x^2}$	$\frac{x}{\left(1+x^2\right)^2}$	$\frac{1}{\left(1+x^2\right)^2}$
Welsch	$\frac{c^2}{2}\left(1-\exp\left(-\left(\frac{x}{c}\right)^2\right)\right)$	$x\exp\left(-(x/c)^2\right)$	$\exp\left(-(x/c)^2\right)$
Tukey	$\begin{cases} \frac{c^2}{6}\left(1-\left(1-(x/c)^2\right)^3\right) & \text{if } \lvert x\rvert \le c \\ \frac{c^2}{6} & \text{if } \lvert x\rvert > c \end{cases}$	$\begin{cases} x\left(1-(x/c)^2\right)^2 & \text{if } \lvert x\rvert \le c \\ 0 & \text{if } \lvert x\rvert > c \end{cases}$	$\begin{cases} \left(1-(x/c)^2\right)^2 & \text{if } \lvert x\rvert \le c \\ 0 & \text{if } \lvert x\rvert > c \end{cases}$
Andrews	$\begin{cases} k^2(1-\cos(x/k)) & \text{if } \lvert x\rvert \le k\pi \\ 2k^2 & \text{if } \lvert x\rvert > k\pi \end{cases}$	$\begin{cases} k\sin(x/k) & \text{if } \lvert x\rvert \le k\pi \\ 0 & \text{if } \lvert x\rvert > k\pi \end{cases}$	$\begin{cases} \frac{\sin(x/k)}{x/k} & \text{if } \lvert x\rvert \le k\pi \\ 0 & \text{if } \lvert x\rvert > k\pi \end{cases}$

PhD thesis Bolstad 2004

The ρ functions



PhD thesis Bolstad 2004

Common Ψ-Functions

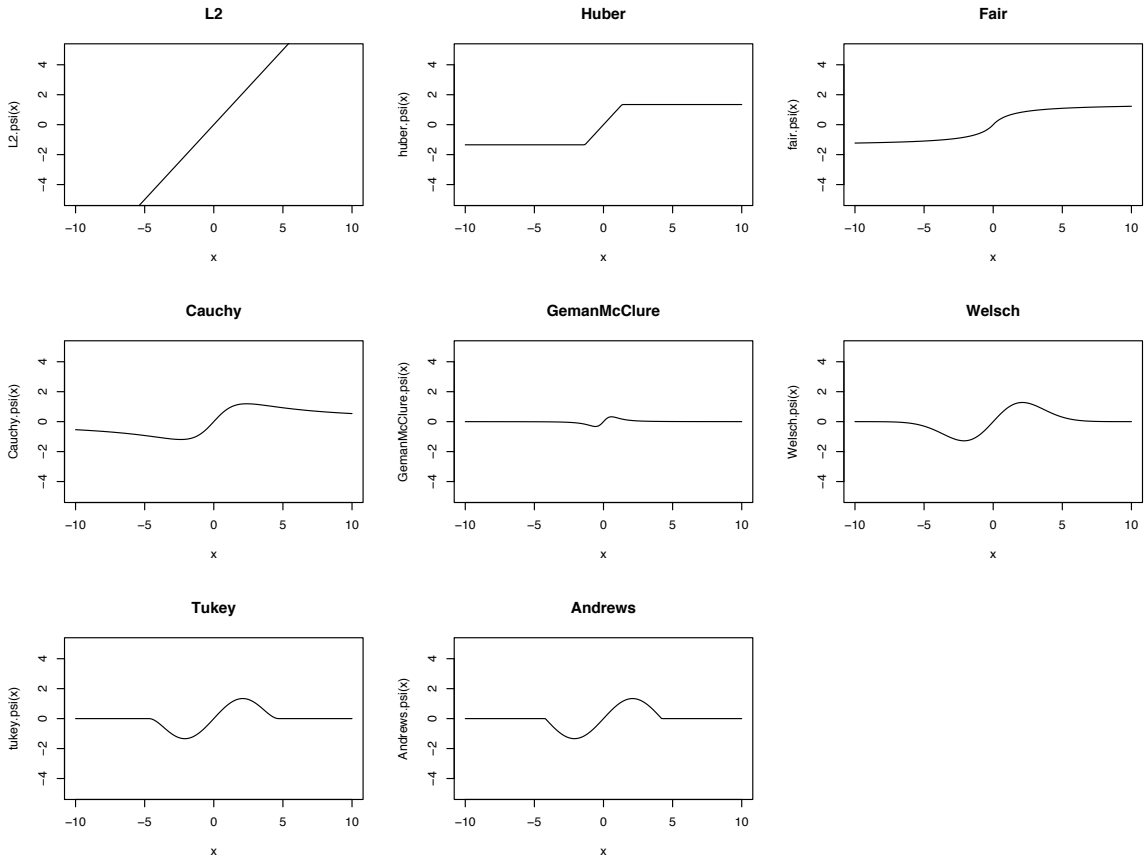


Figure 4.2: The ψ functions for some common M-estimators.

PhD thesis Bolstad 2004

Corresponding Weight Functions

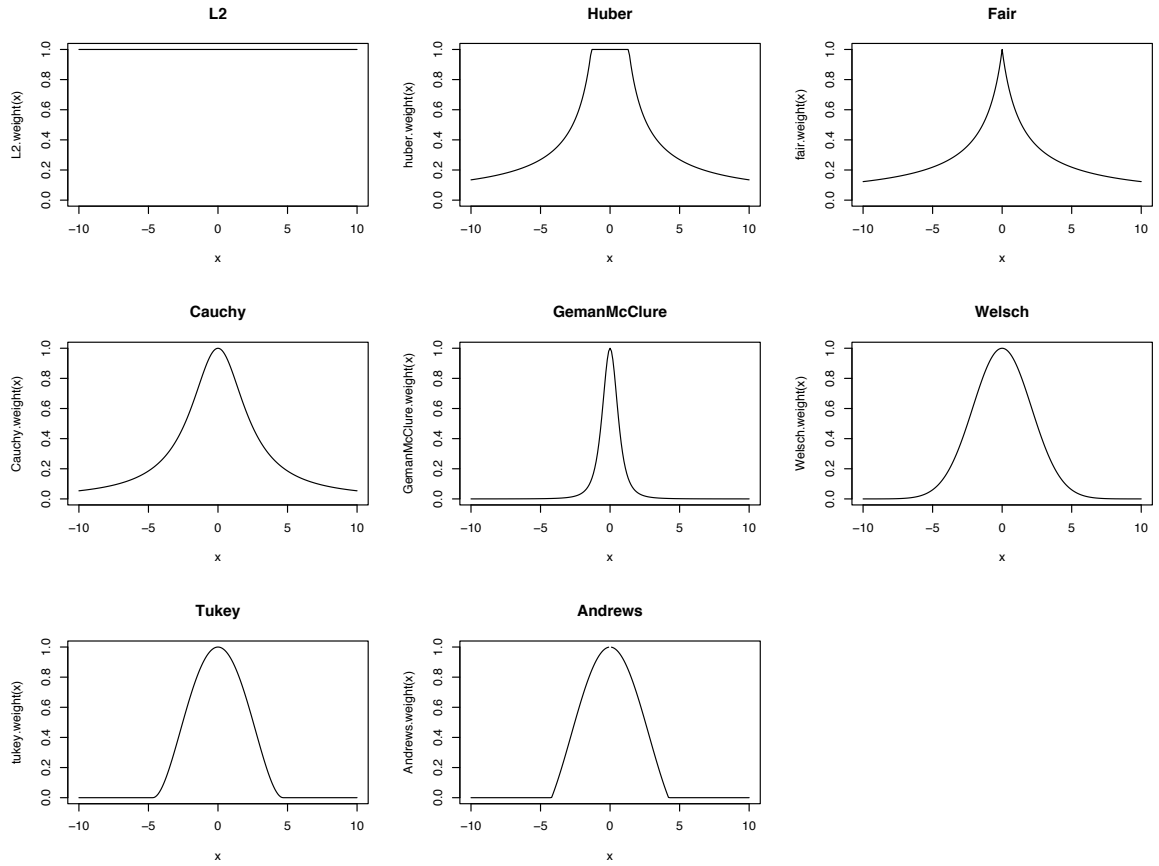
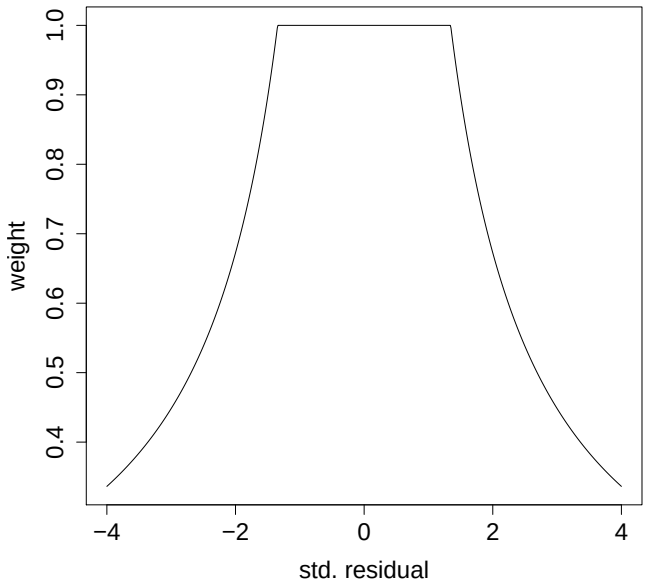


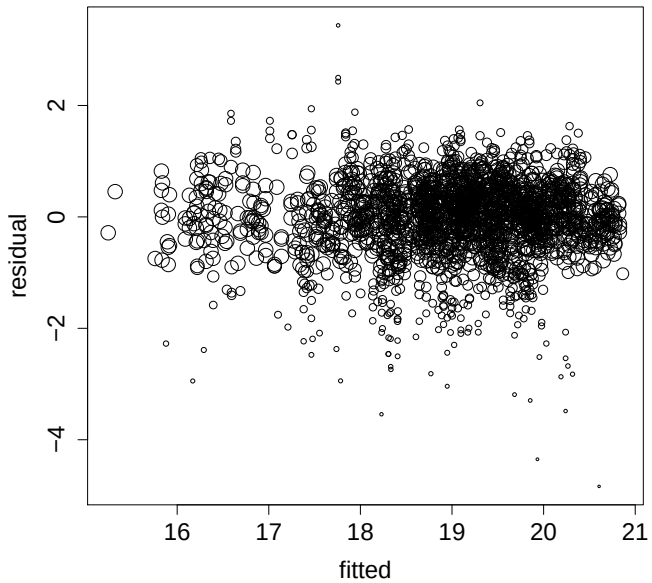
Figure 4.3: The weight functions for some common M-estimators.

PhD thesis Bolstad 2004

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

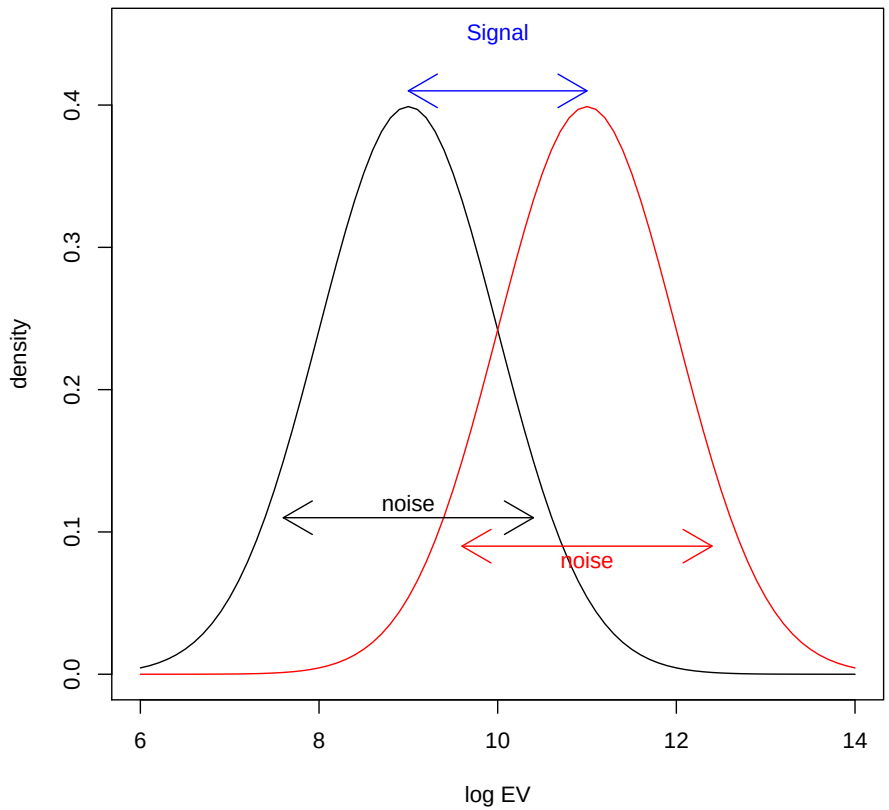
- Outlying peptide intensities: incorrect peptide identification, post-translational modifications, ...





Experimental Design

Power?



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

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

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

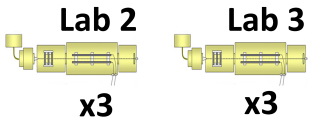
If we can assume equal variance in both treatment groups:

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

$\rightarrow$ 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.
- Assess difference in power between 3vs3, 6vs6 and 9vs9 patients.

Blocking

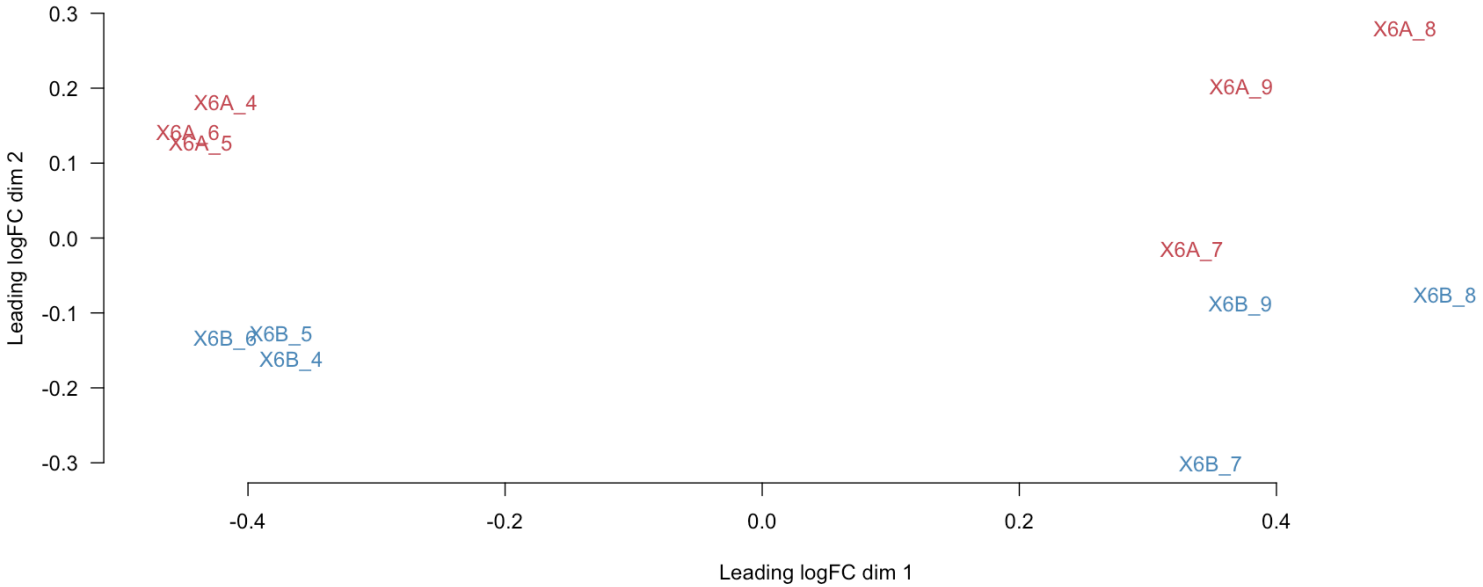


Color variable [?]

condition

MDS plot after full preprocessing [?]

- [ ] Plot MDS points
- [x] 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 experimental units.
- Technical: lab effect, time effect, plasma extraction, MS-run, ...

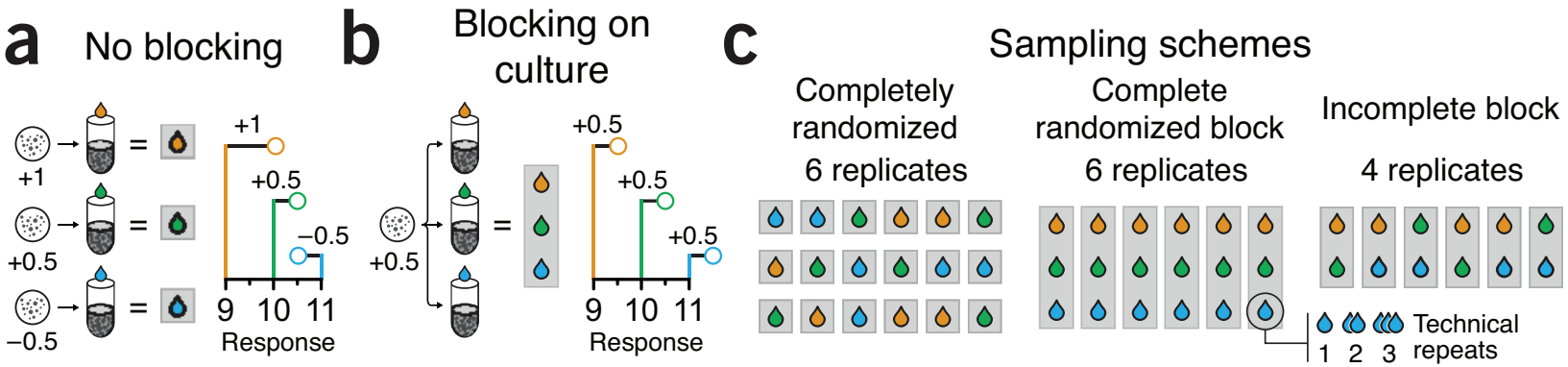
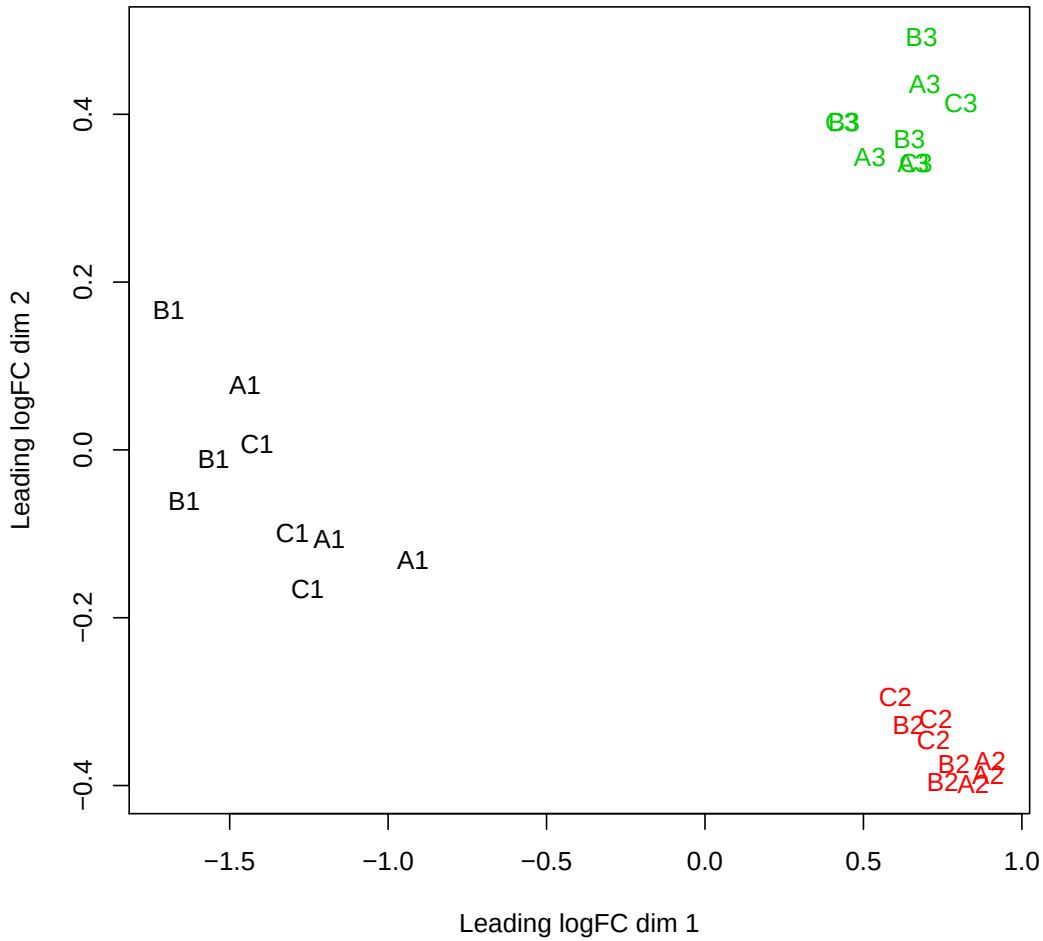


Figure 2 | Blocking improves sensitivity by isolating variation in samples that is independent from treatment effects. (**a**) Measurements from treatment aliquots derived from different cell cultures are differentially offset (e.g., 1, 0.5, –0.5) because of differences in cultures. (**b**) When aliquots 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.

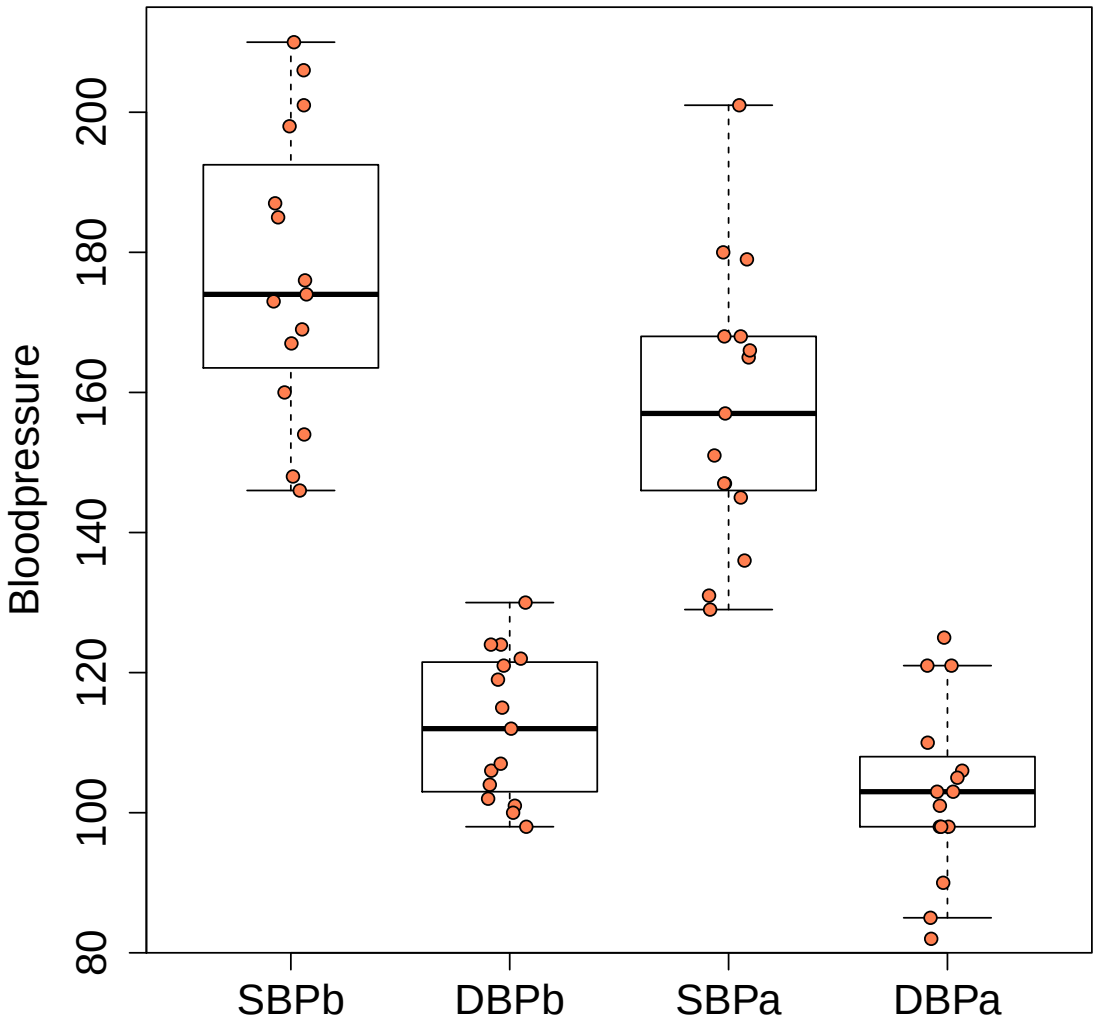
Nature Methods 2014, 11(7) 699–700.

Effect of treatment and lab: strong blocking

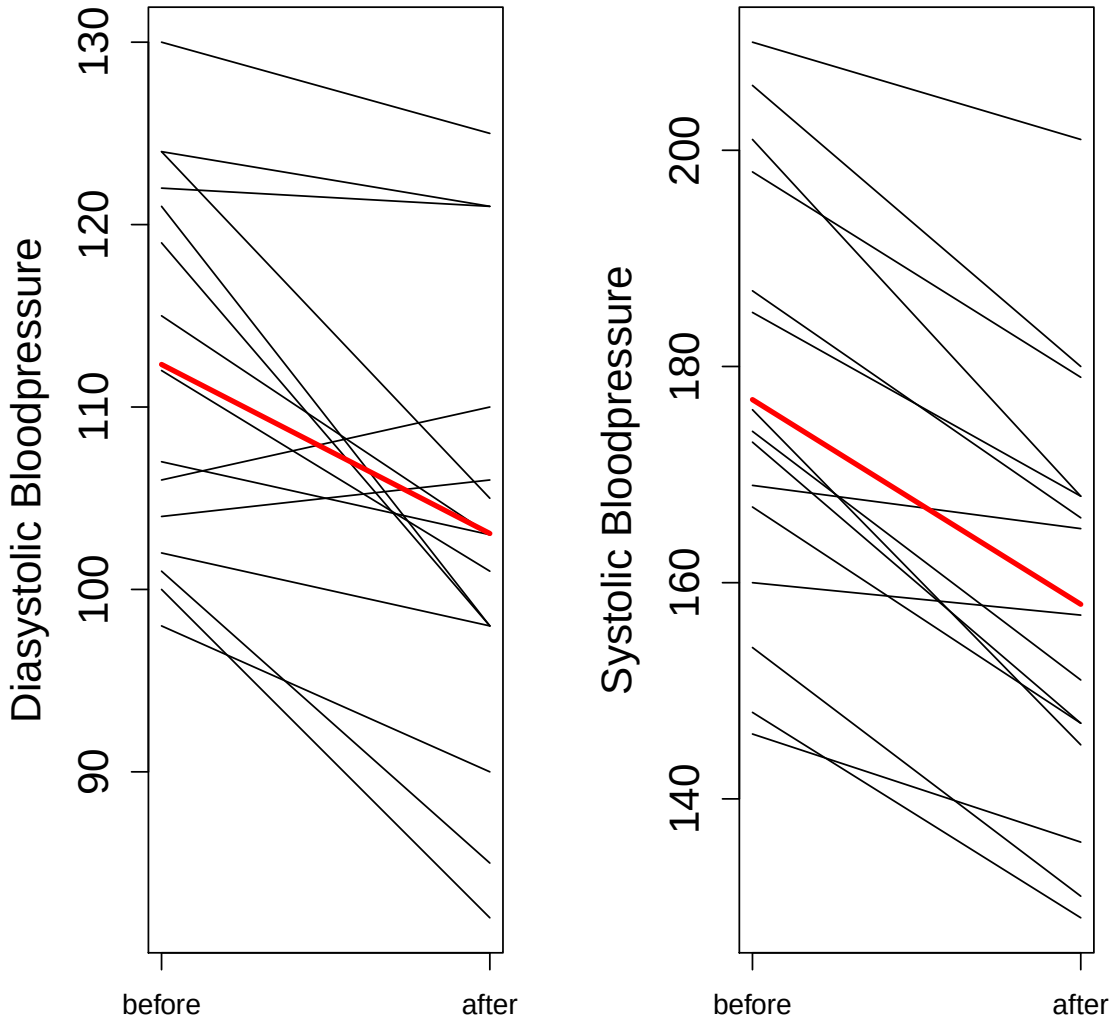
- Unwanted variability is often present (e.g. batch, lab period, cage, ...)
- Exploit it
- Randomize all treatment within each block
- Isolate between-block variability from FC estimates in the data analysis



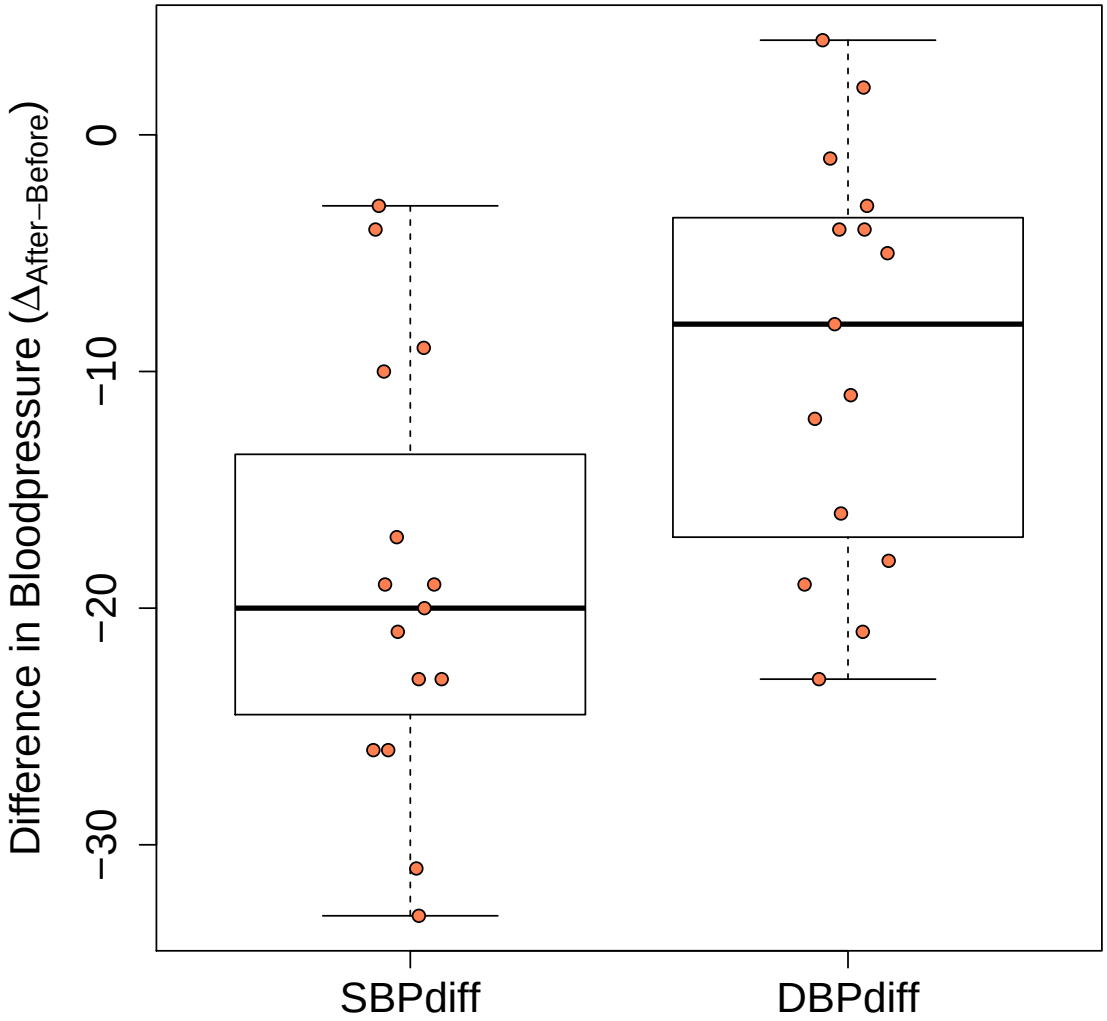
Intermezzo: Power gain of blocking



Intermezzo: Power gain of blocking



Intermezzo: 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 design

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

Residual standard error: 7.2 on 6 degrees of freedom
Multiple R-squared:  0.9317,Adjusted R-squared:  0.8519
F-statistic: 11.69 on 7 and 6 DF,  p-value: 0.00404
```

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
```