Part II: Statistical Inference

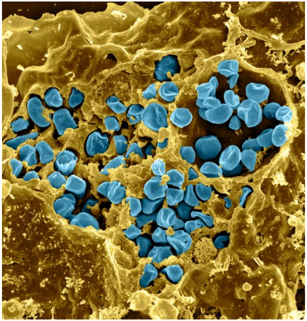
Lieven Clement

Proteomics Data Analysis Shortcourse

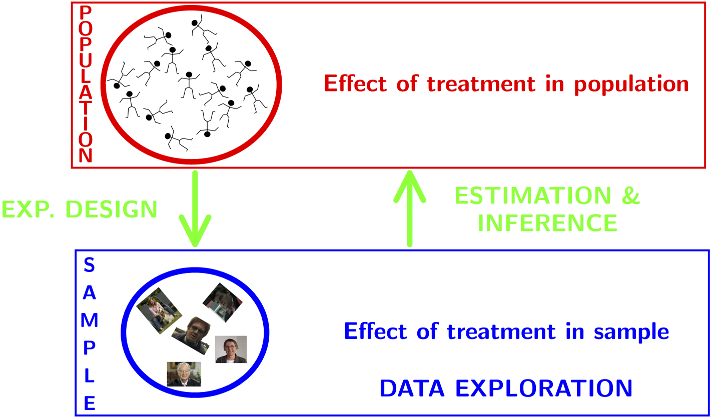
Statistical Inference

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

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

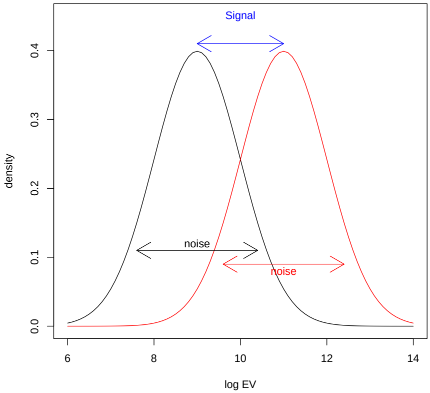
DATA EXPLORATION

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
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Hypothesis testing: a single protein



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

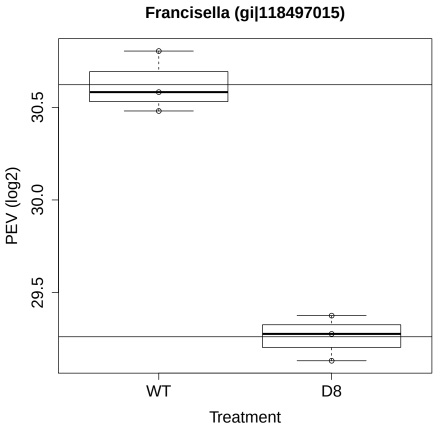
$$T_g = \frac{\Delta}{\text{se}_\Delta}$$

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

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



$$t = \frac{\log_2 \widehat{\text{FC}}}{\text{se}_{\log_2 \widehat{\text{FC}}}} = \frac{-1.4}{0.118} = -11.9$$

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

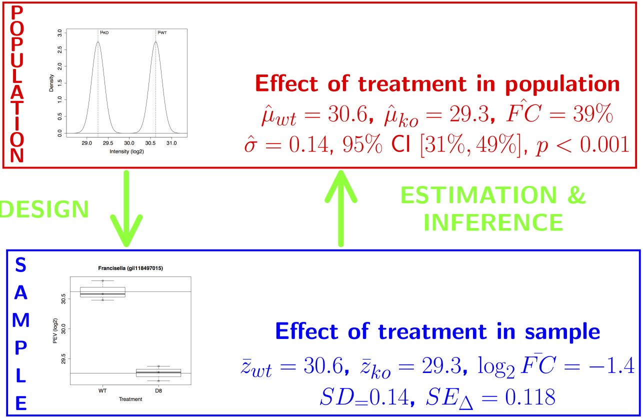
- In general we start from **alternative hypothese** H_A: we want to show an effect of the KO on a protein
    - On average the protein abundance in WT is different from that in KO
- But, we will assess it by falsifying the opposite: **null hypothesis** H_0
    - On average the protein abundance in WT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



Effect of treatment in population

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

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

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?

→ 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.

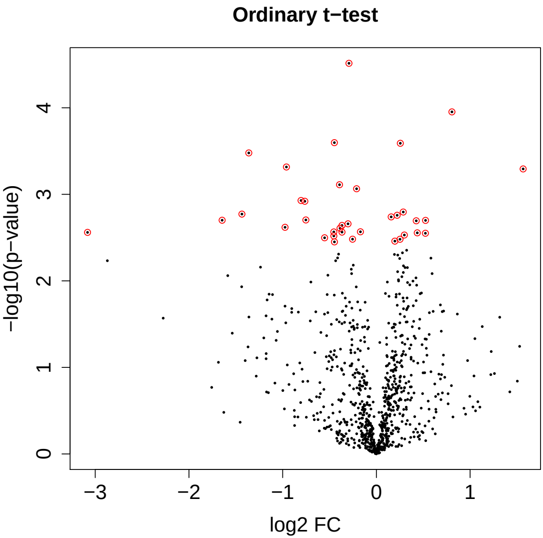
FDR: False discovery rate

- 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

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

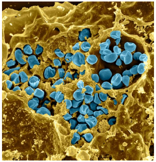
$$\text{FDR}_\text{BH}(|t_\text{thres}|) = \frac{1 \times p_{t_\text{thres}}}{\frac{\#|t_i|\geq t_\text{thres}}{m}}$$

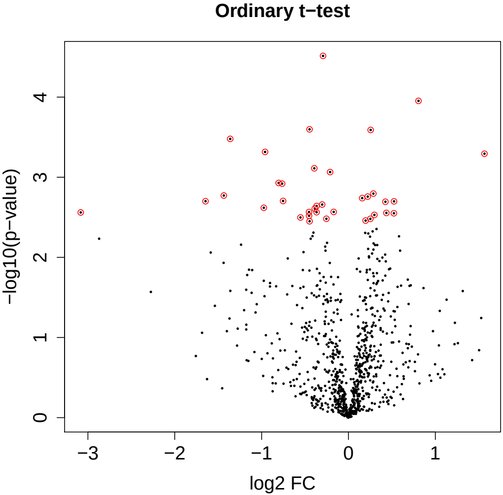
- FDR adjusted p-values can be calculated (e.g. Perseus, R, ...)


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

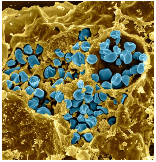
Moderated Statistics

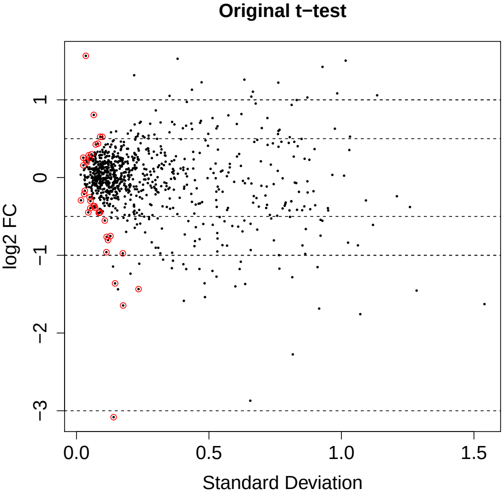
Problems with ordinary t-test





Problems with ordinary t-test





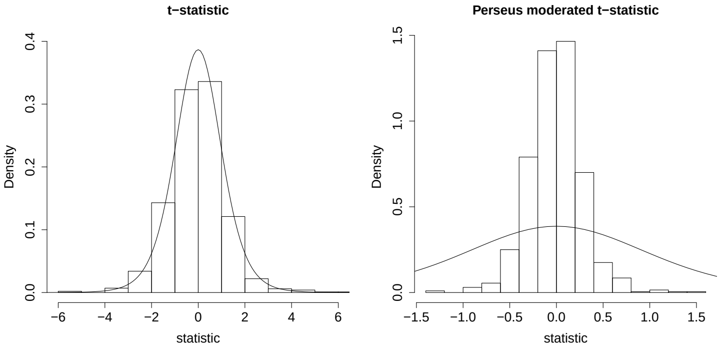
A moderated *t*-test

A general class of moderated test statistics is given by

$$T_g^{mod} = \frac{\bar{Y}_{g1} - \bar{Y}_{g2}}{C \quad \tilde{S}_g},$$

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

- C is a constant depending on the design e.g. $\sqrt{1/n_1 + 1/n_2}$ for a t-test.
- $\tilde{S}_g = S_g + S_0$: add small positive constant to denominator of t-statistic.
- This can be adopted in Perseus.



- The choice of S_0 in Perseus is ad hoc and the t-statistic is no-longer t-distributed.

→ Permutation test, but is difficult for more complex designs.

→ Allows for Data Dredging because user can choose S_0

A moderated *t*-test

A general class of moderated test statistics is given by

$$T_g^{mod} = \frac{\bar{Y}_{g1} - \bar{Y}_{g2}}{C \quad \tilde{S}_g},$$

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

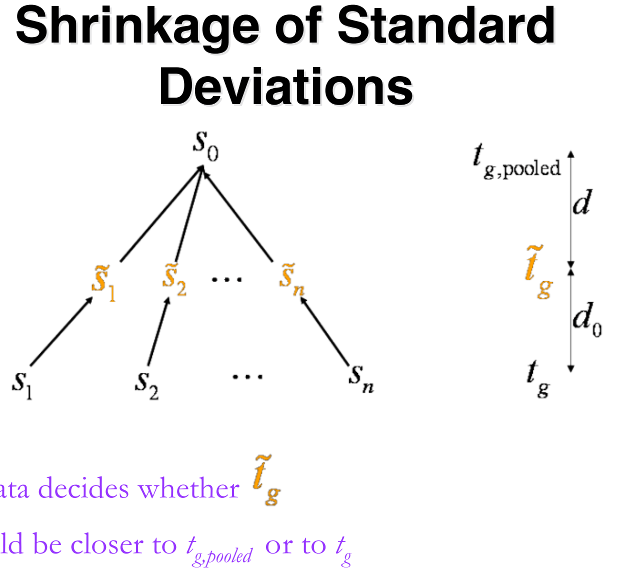
- **empirical Bayes** theory provides formal framework for borrowing strength across proteins,
- Implemented in popular bioconductor package **limma**

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

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

→ Note that the degrees of freedom increase by borrowing strength across proteins!

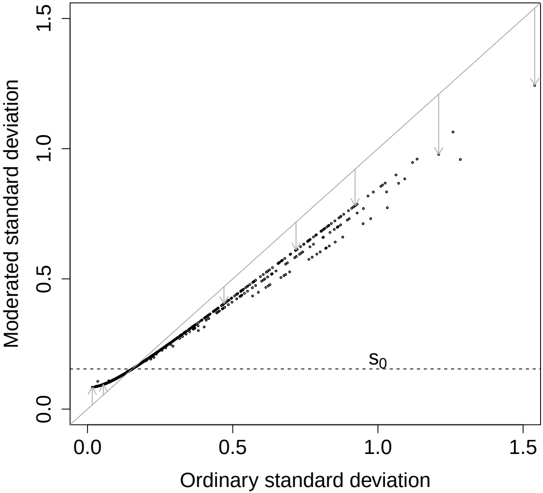
Shrinkage of the variance and moderated t-statistics



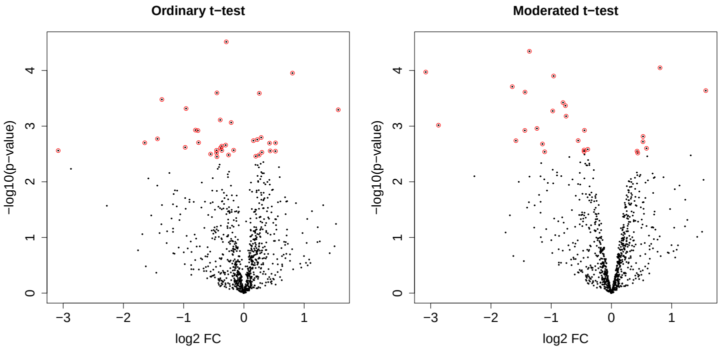
Shrinkage of Standard Deviations

The data decides whether $\tilde{t}_g$ should be closer to $t_{g,pooled}$ or to t_g

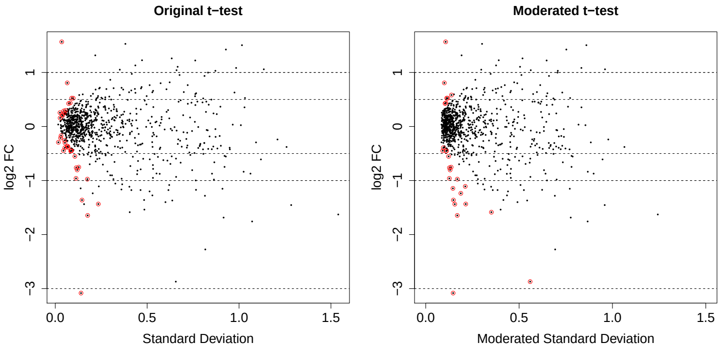
Shrinkage of the variance with limma

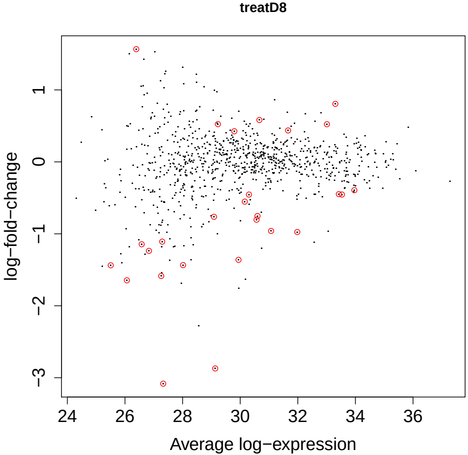


Problems with ordinary t-test solved by moderated EB t-test



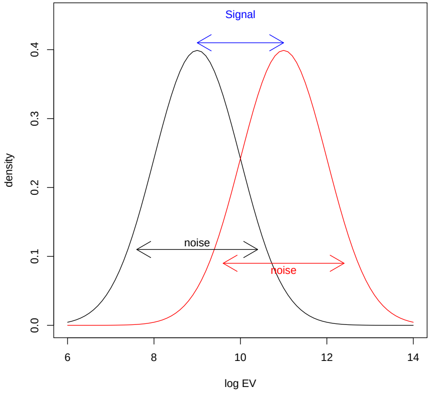
Problems with ordinary t-test solved by moderated EB t-test




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

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:

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

→ 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.

Experimental Design: Blocking

Sources of variability

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

- Biological: fluctuations in protein level between mice, fluctations in protein level between cells, ...
- Technical: cage effect, lab effect, week effect, plasma extraction, MS-run, ...

Blocking Example: mouse T-cells

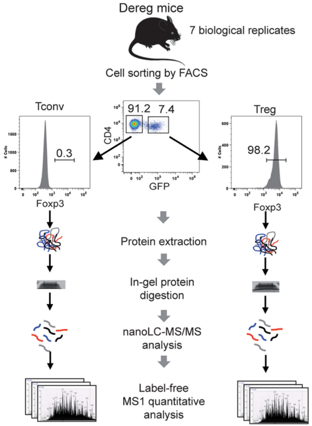
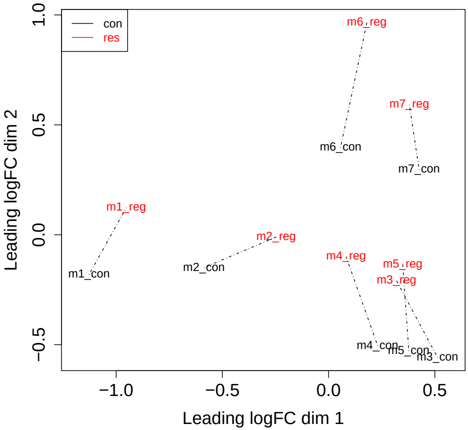


FIG. 1. **Label-free quantitative analysis of conventional and regulatory T cell proteomes.** General analytical workflow based on cell sorting by flow cytometry using the DEREG mouse model and parallel proteomic analysis of Tconv and Treg cell populations by nanoLC-MS/MS and label-free relative quantification.

Blocking Example: mouse T-cells



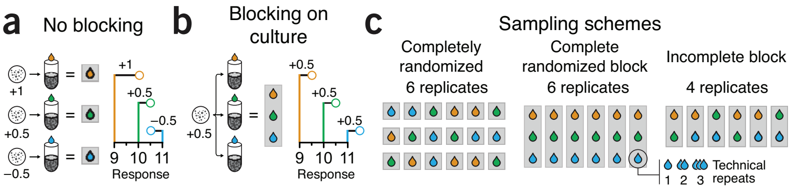
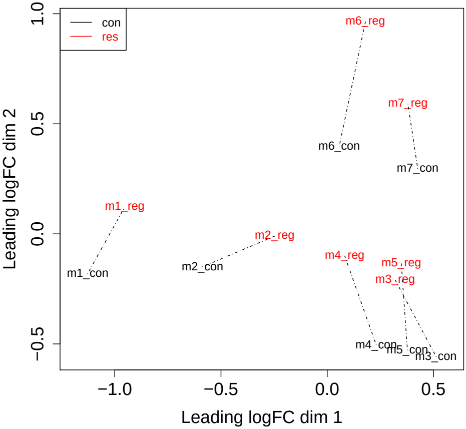


Figure 2 | Blocking improves sensitivity by isolating variation in samples that is independent from treatment effects. (**a**) Measurements from treatment aliquots derived from different cell cultures are differentially offset (e.g., 1, 0.5, –0.5) because of differences in cultures. (**b**) When aliquots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

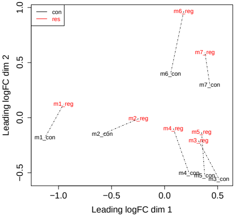
Blocking

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



Blocking

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



→ 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:

$$y \sim \text{type} + \text{mouse}$$

→ Not possible with Perseus!

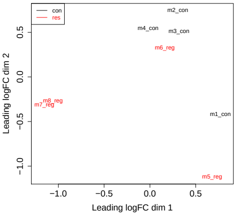
Power gain of blocking

- Completely randomized design (CRD): 8 mice, 4 conventional T-cells, 4 regulatory T-cells.
- Randomized complete block desigh (RBC): 4 mice, for each mouse conventional and regulatory T-cells.

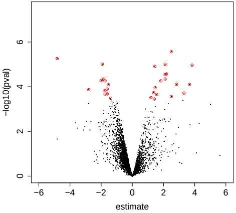
Power gain of blocking

CRD

$y \sim \text{type}$

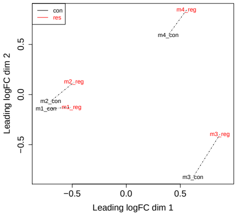


CRD−design:
29 proteins significant

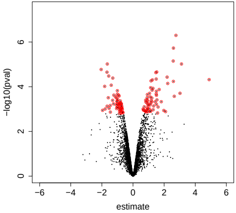


RCB

$y \sim \text{type} + \text{mouse}$

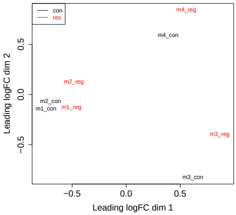


RCB−design:
121 proteins significant

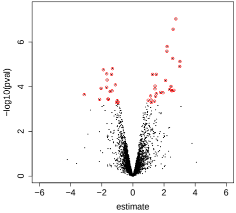


RCB

$y \sim \text{type}$

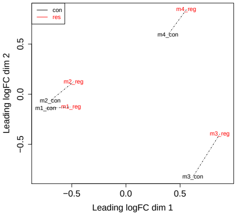


RCB−design, no mouse effect:
43 proteins significant



Anova table: P24452, Capg, Macrophage-capping protein

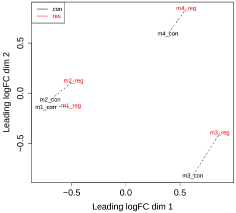
```
### RCB design ###
          Df  Sum Sq Mean Sq  F value    Pr(>F)
type       1 15.2282 15.2282 3720.035 9.71e-06 ***
mouse      3  0.2179  0.0726   17.747  0.02058 *
Residuals  3  0.0123  0.0041
```



```
### RCB design: no mouse effect ###
          Df  Sum Sq Mean Sq F value    Pr(>F)
type       1 15.2282 15.2282  396.87 1.038e-06 ***
Residuals  6  0.2302  0.0384
```

```
### CRD design ###
          Df  Sum Sq Mean Sq F value    Pr(>F)
type       1 11.6350 11.6350  122.86 3.211e-05 ***
Residuals  6  0.5682  0.0947
```

Anova table: P24452, Capg, Macrophage-capping protein

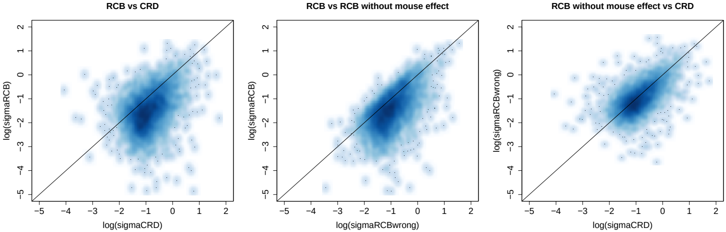


```
### RCB design ###
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.21485    0.05058 439.190 2.60e-08 ***
typereg      2.75937    0.04524  60.992 9.71e-06 ***
mouse2       0.30560    0.06398   4.776   0.0174 *
mouse3      -0.15193    0.06398  -2.375   0.0981 .
mouse4       0.07331    0.06398   1.146   0.3350
---
Residual standard error: 0.06398 on 3 degrees of freedom
```

```
### RCB design: no mouse effect ###
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.27160    0.09794  227.40 4.88e-13 ***
typereg      2.75937    0.13851   19.92 1.04e-06 ***
---
Residual standard error: 0.1959 on 6 degrees of freedom
```

```
### CRD design ###
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.3012     0.1557  149.65 6.00e-12 ***
typereg       2.4956     0.2251   11.08 3.21e-05 ***
---
Residual standard error: 0.3077 on 6 degrees of freedom
```

Comparison residual variance


RCB vs CRD
log(sigmaRCB)
log(sigmaCRD)
RCB vs RCB without mouse effect
log(sigmaRCB)
log(sigmaRCBwrong)
RCB without mouse effect vs CRD
log(sigmaRCBwrong)
log(sigmaCRD)