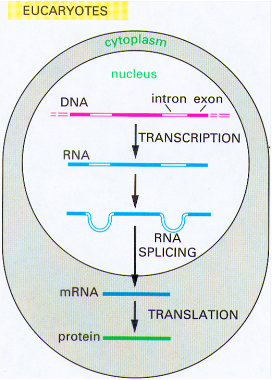
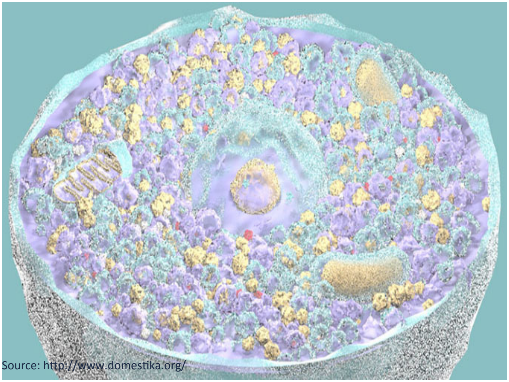
Differential analysis for label free mass spectrometry based proteomics

Lieven Clement

Bioinformatics Summer School 2019, June 1st-5th, UCLouvain, Louvain-la-Neuve, Belgium

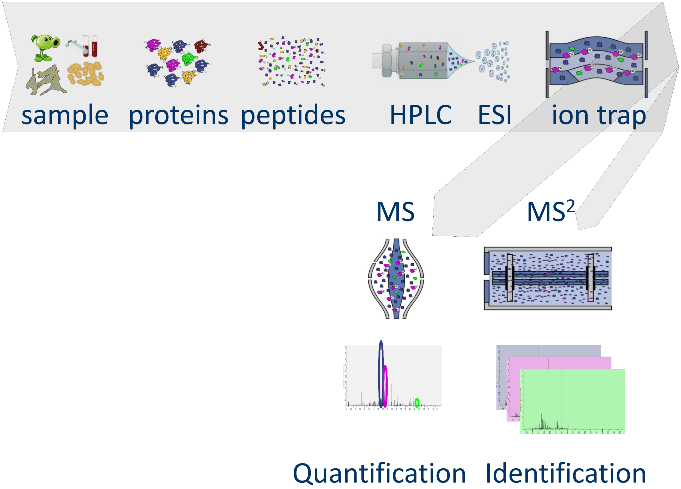
1. Background
2. Peptide based workflow
3. Robust summarisation & Inference
4. Experimental design





Source: http://www.domestika.org/

Challenges in Label Free MS-based Quatitative proteomics

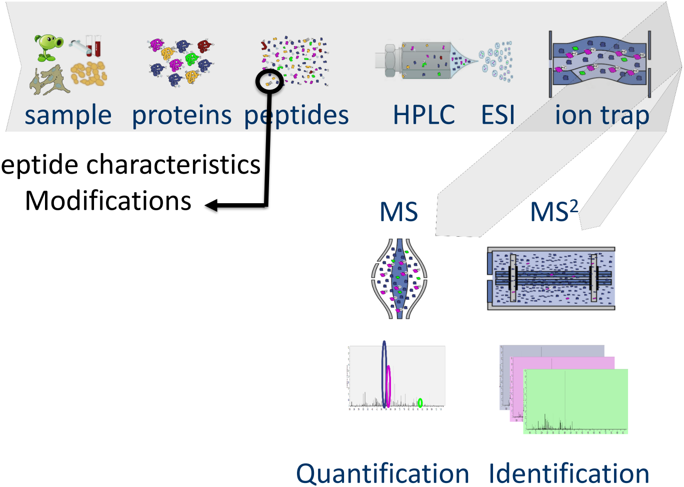


sample proteins peptides HPLC ESI ion trap

MS MS^2

Quantification Identification

Challenges in Label Free MS-based Quatitative proteomics



sample proteins peptides HPLC ESI ion trap

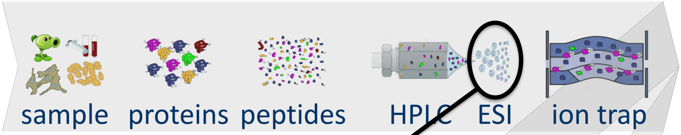
Peptide characteristics

- Modifications

MS MS^2

Quantification Identification

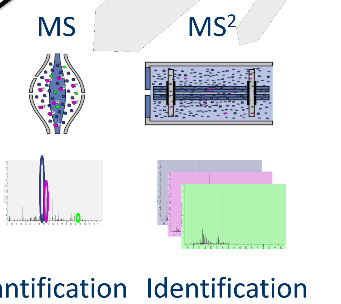
Challenges in Label Free MS-based Quatitative proteomics



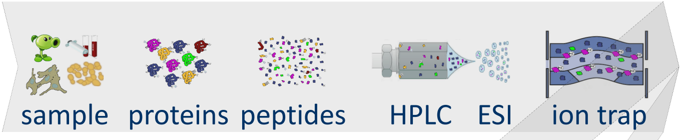
Peptide characteristics

- Modifications

- Ionisation efficiency
  - Outliers
  - Huge variability

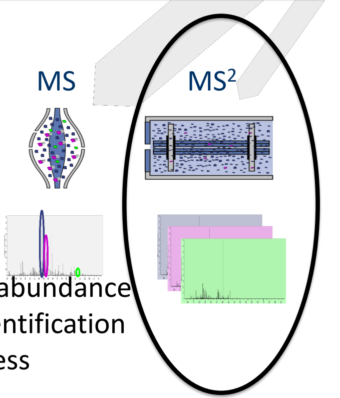


Challenges in Label Free MS-based Quatitative proteomics

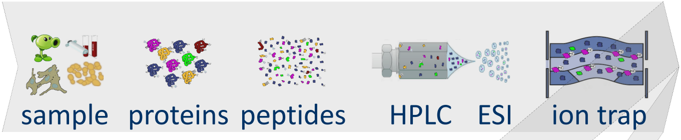


Peptide characteristics

- Modifications
- Ionisation efficiency
  - Outliers
  - Huge variability
- MS^2 selection on peptide abundance
  - Context dependent Identification
  - Non-random missingness

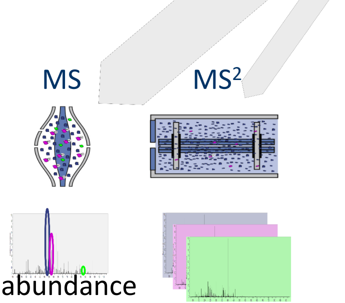


Challenges in Label Free MS-based Quatitative proteomics



Peptide characteristics

- Modifications
- Ionisation efficiency
  - Outliers
  - Huge variability

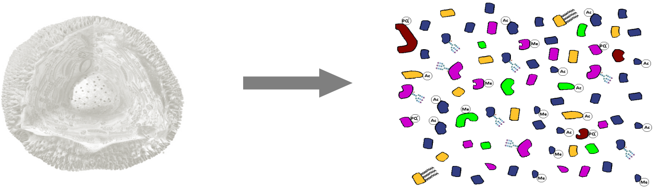


- MS^2 selection on peptide abundance
  - Context dependent Identification
  - Non-random missingness

Unbalanced peptides identifications across samples and messy data

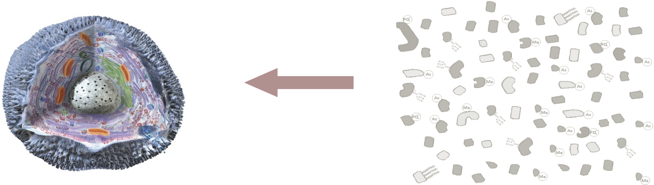
Challenges in Label Free MS-based Quatitative proteomics

MS-based proteomics returns **peptides**: pieces of proteins

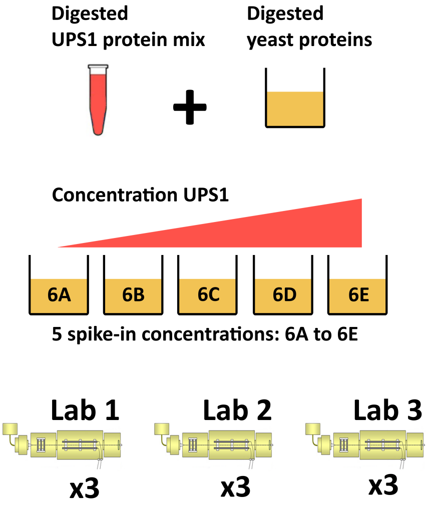


Challenges in Label Free MS-based Quatitative proteomics

We need information on protein level!

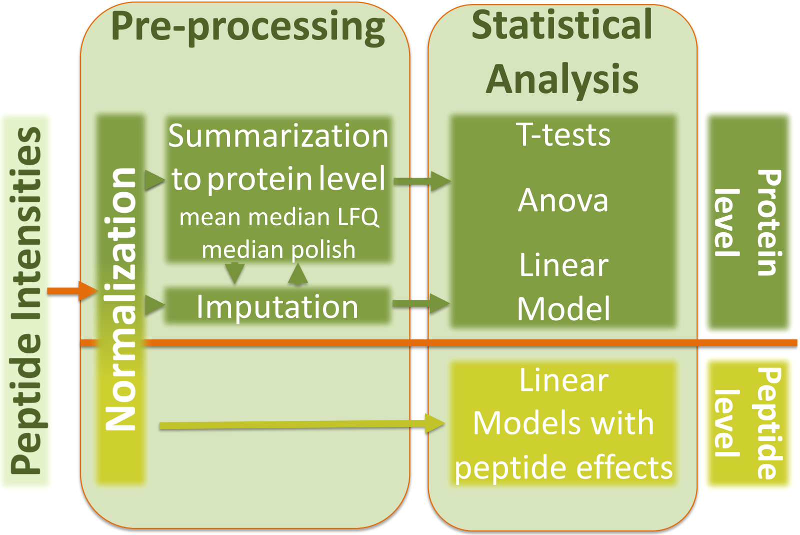


CPTAC Spike-in Study



- Same trypsin-digested yeast proteome background in each sample
- Trypsin-digested Sigma UPS1 standard: 48 different human proteins spiked in at 5 different concentrations (treatment A-E)
- Samples repeatedly run on different instruments in different labs
- After MaxQuant search with match between runs option
  - 41% of all proteins are quantified in all samples
  - 6.6% of all peptides are quantified in all samples

→ **vast amount of missingness**



Peptide Intensities

Pre-processing

Normalization

Summarization to protein level
mean median LFQ
median polish

Imputation

Statistical Analysis

T-tests

Anova

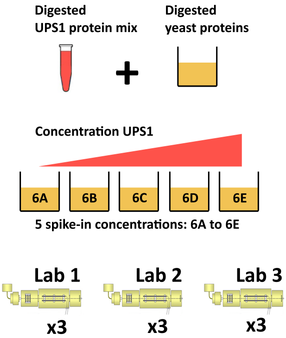
Linear Model

Protein level

Linear Models with peptide effects

Peptide level

Summarization



Digested UPS1 protein mix

Digested yeast proteins

Concentration UPS1

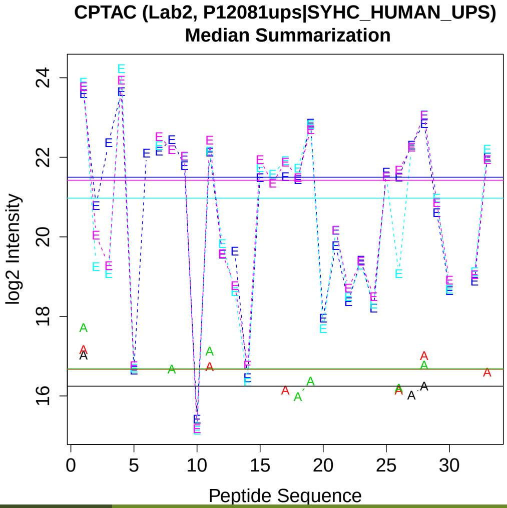
6A 6B 6C 6D 6E

5 spike-in concentrations: 6A to 6E

Lab 1 x3

Lab 2 x3

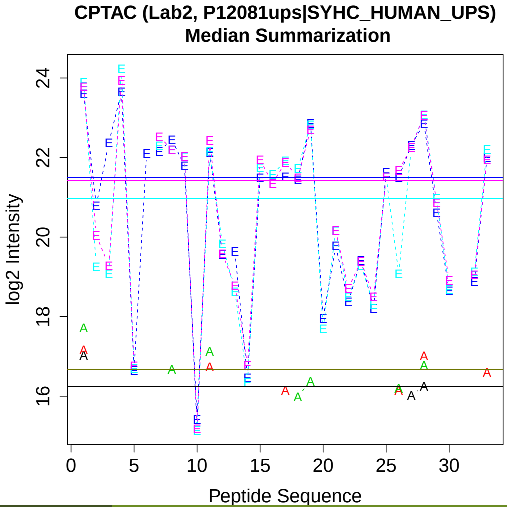
Lab 3 x3

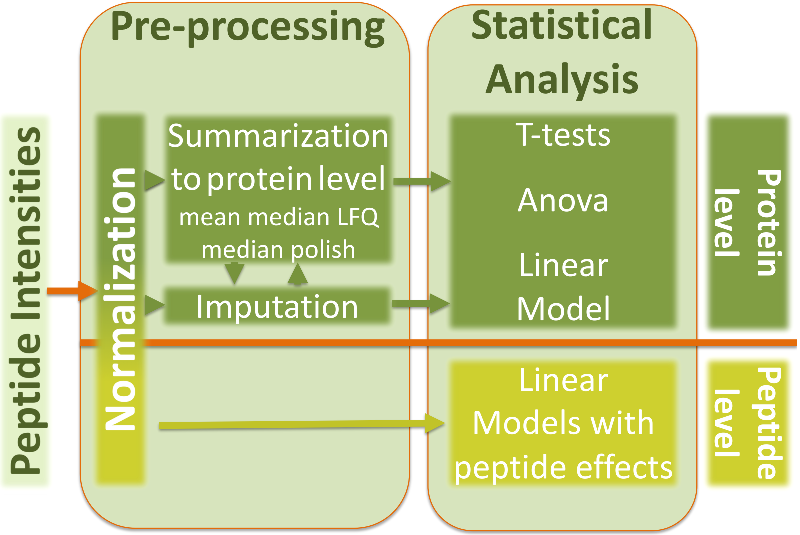


CPTAC (Lab2, P12081ups|SYHC_HUMAN_UPS)
Median Summarization

Summarization

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




Pre-processing
Statistical Analysis
Peptide Intensities
Normalization
Summarization to protein level
mean median LFQ median polish
Imputation
T-tests
Anova
Linear Model
Protein level
Linear Models with peptide effects
Peptide level

MSqRob workflow (Goeminne et al. 2016 MCP, PMID: 26566788)

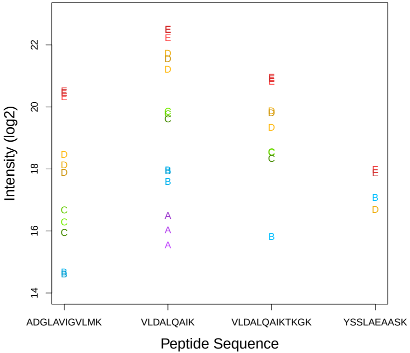
$$y_{grp} = \beta_g^{group} + u_r^{\text{run}} + \beta_p^{\text{pep}} + \epsilon_{rp}$$

protein-level

- β_g^{group}: spike-in
- random run effect $u_r^{\text{run}} \sim N\left(0, \sigma_{\text{run}}^2\right)$
  $\rightarrow$ Addresses pseudo-replication

peptide-level

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



MSqRob workflow (Goeminne et al. 2016 MCP, PMID: 26566788)

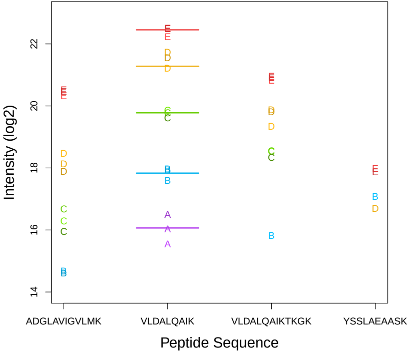
$$y_{grp} = \beta_g^{group} + u_r^{run} + \beta_p^{pep} + \epsilon_{rp}$$

protein-level

- β_g^{group}: spike-in
- random run effect $u_r^{run} \sim N\left(0, \sigma^2_{\text{run}}\right)$
  $\rightarrow$ Addresses pseudo-replication

peptide-level

- peptide specific effect β_p^{pep}
- within run error $\epsilon_{rp} \sim N\left(0, \sigma^2_{\epsilon}\right)$



MSqRob workflow (Goeminne et al. 2016 MCP, PMID: 26566788)

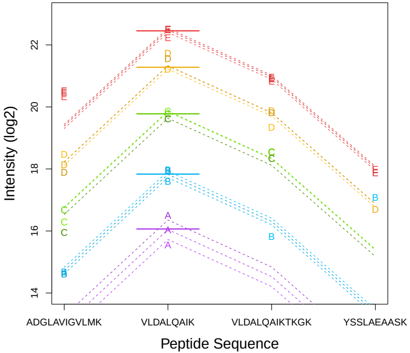
$$y_{grp} = \beta_g^{group} + u_r^{\text{run}} + \beta_p^{\text{pep}} + \epsilon_{rp}$$

protein-level

- β_g^{group}: spike-in
- random run effect $u_r^{\text{run}} \sim N\left(0, \sigma^2_{\text{run}}\right)$
  $\rightarrow$ Addresses pseudo-replication

peptide-level

- peptide specific effect β_p^{pep}
- within run error $\epsilon_{rp} \sim N\left(0, \sigma^2_{\epsilon}\right)$



MSqRob workflow (Goeminne et al. 2016 MCP, PMID: 26566788)

$$y_{grp} = \beta_g^{group} + u_r^{run} + \beta_p^{pep} + \epsilon_{rp}$$

protein-level

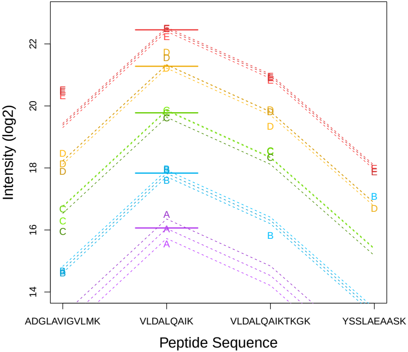
- β_g^{group}: spike-in
- random run effect $u_r^{run} \sim N\left(0, \sigma^2_{run}\right)$
  → Addresses pseudo-replication

peptide-level

- peptide specific effect β_p^{pep}
- within run error $\epsilon_{rp} \sim N\left(0, \sigma^2_{\epsilon}\right)$

Estimation

1. Robust regression for outliers
2. Penalise β^{treat} (Ridge regression)
3. Empirical Bayes variance estimation



Fit MSqRob mixed model in two-stage approach

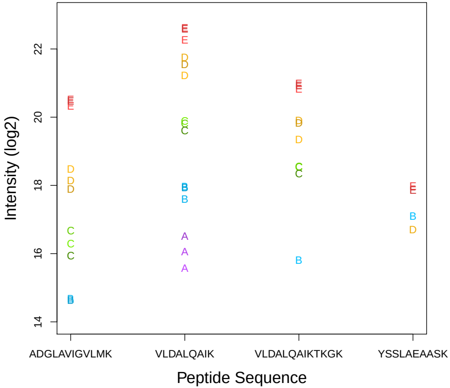
MSqRob

- No protein summaries available
- Difficult to disseminate
- Unclear to calculate degrees of freedom to adopt t-tests for inference in experiments with small sample sizes

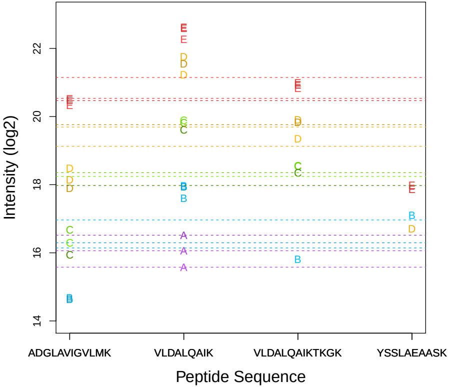
→ Modular approach

1. Summarize peptides to proteins using robust regression
2. Robust penalized regression of protein level summaries

Summarisation with peptide based model



Summarisation with peptide based model

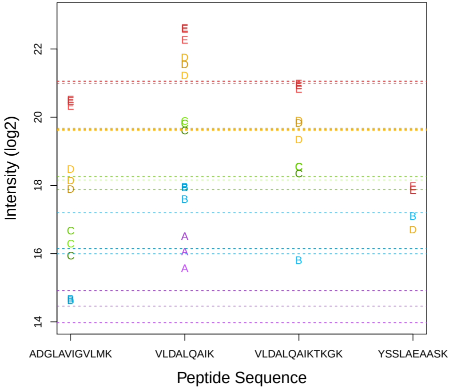


Protein by protein analysis of peptide data with linear model

peptide level protein level

$$y_{sp} = \epsilon_{sp} \quad + \quad \beta_s^{\text{sample}}$$

Summarisation with peptide based model

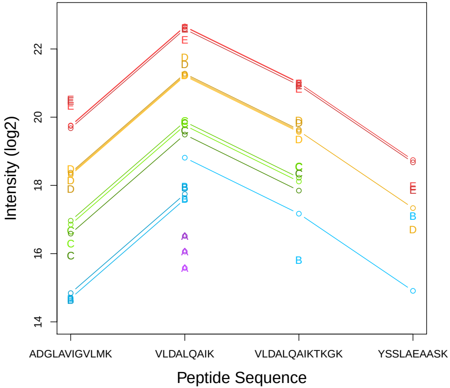


Protein by protein analysis of peptide data with linear model

peptide level protein level

$$y_{sp} = \beta_p^{\text{pep}} + \epsilon_{sp} \quad + \quad \beta_s^{\text{sample}}$$

Summarisation with peptide based model

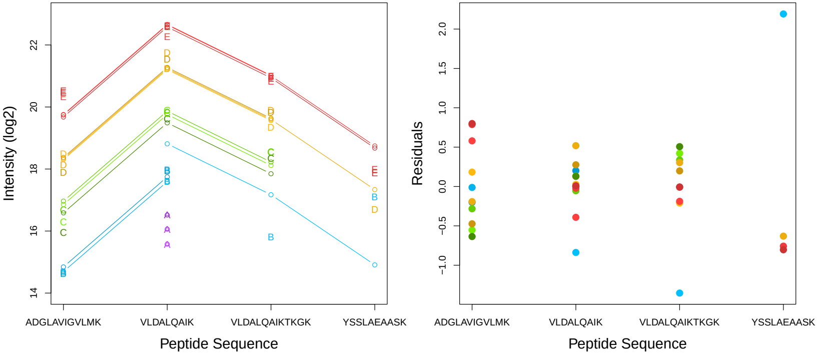


Protein by protein analysis of peptide data with linear model

peptide level protein level

$$y_{sp} = \beta_p^{pep} + \epsilon_{sp} \quad + \quad \beta_s^{sample}$$

Summarisation with peptide based model

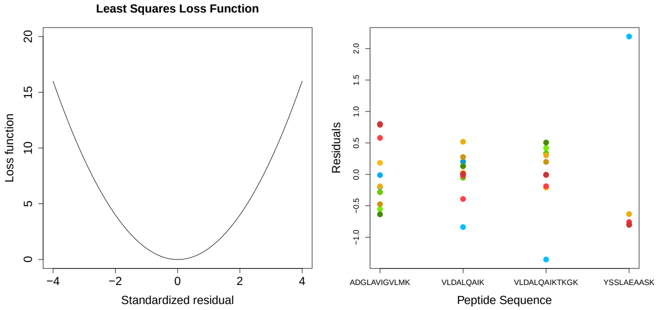


Protein by protein analysis of peptide data with linear model

$$\text{Estimation} \rightarrow \text{argmin}_{\beta^{\text{pep}}_{1\ldots P},\beta^{\text{samp}}_{1\ldots n}} \left[\sum_{i=1}^{n} \sum_{p}^{P} \left(y_{ip} - \beta_p^{\text{pep}} - \beta_i^{\text{samp}} \right)^2 \right]$$

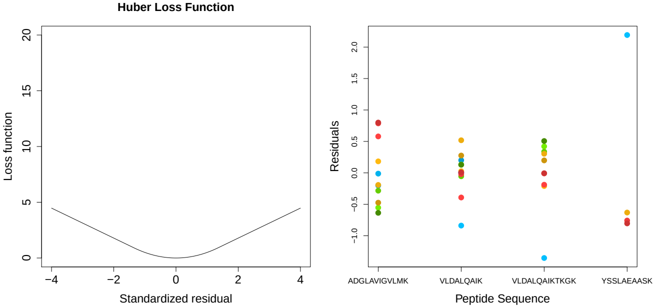
Robust estimation using observation weights

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



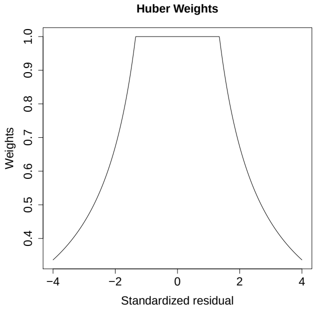
Robust estimation using observation weights

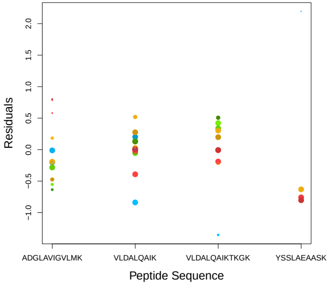
- Outlying peptide intensities: incorrect peptide identification, post-translational modifications, ...



Robust estimation using observation weights

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



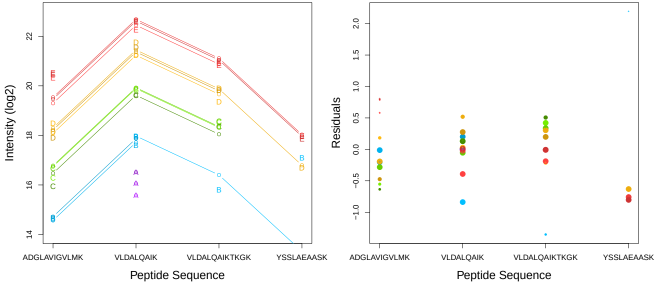


- Iteratively fit model with observation weights $w(\epsilon_{ip})$

$$\text{argmin}_{\beta^{\text{pep}}_{1\ldots P},\beta^{\text{samp}}_{1\ldots n}}\left[\sum_{i=1}^{n}\sum_{p}^{P} w(\epsilon_{ip})\left(y_{ip}-\beta^{\text{pep}}_{p}-\beta^{\text{samp}}_{i}\right)^{2}\right]$$

Robust estimation using observation weights

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

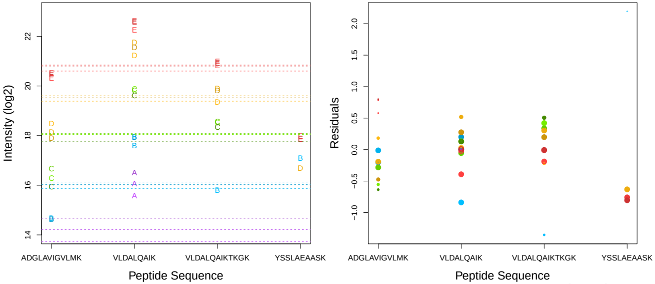


- Iteratively fit model with observation weights $w(\epsilon_{ip})$

$$\text{argmin}_{\beta^{\text{pep}}_{1\ldots P},\beta^{\text{samp}}_{1\ldots n}}\left[\sum_{i=1}^{n}\sum_{p}^{P} w(\epsilon_{ip})\left(y_{ip}-\beta^{\text{pep}}_{p}-\beta^{\text{samp}}_{i}\right)^2\right]$$

Robust estimation using observation weights

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



- Iteratively fit model with observation weights $w(\epsilon_{ip})$

$$\text{argmin}_{\beta^{\text{pep}}_{1\ldots P},\beta^{\text{samp}}_{1\ldots n}} \left[\sum_{i=1}^{n}\sum_{p}^{P} w(\epsilon_{ip})\left(y_{ip}-\beta^{\text{pep}}_{p}-\beta^{\text{samp}}_{i}\right)^2\right]$$

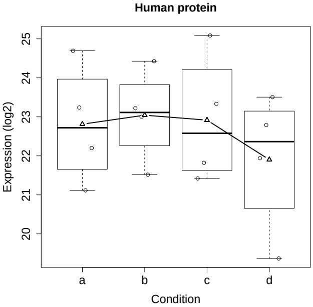
Assess effect of robust summarization

Alter cptacAvsB_lab3_median.Rmd file to use robust summarization:
→ use method="robust" in combineFeatures

Inference upon summarisation: Protein level model

$$y_r \quad = \quad \beta^{group}_{g(r)} + \epsilon_r$$

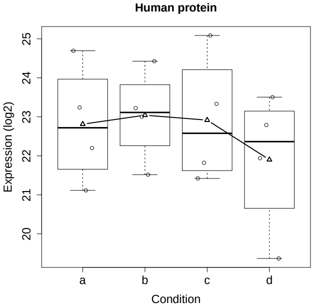
- y_r: protein summary of run r
- $\sum_{g=1}^{G} \beta^{group}_{g} = 0$



Inference upon summarisation: Protein level model

$$\begin{aligned} y_r &= \beta^{group}_{g(r)} + \epsilon_r \\ &= \mathbf{X}^t_r \boldsymbol{\beta} + \epsilon_r \end{aligned}$$

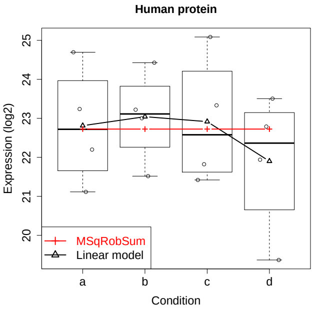
- y_r: protein summary of run r
- $\sum_{g=1}^G \beta^{group}_g = 0$
- $\boldsymbol{\beta} = [\beta^{group}_1, \ldots, \beta^{group}_G]^t$
- $\mathbf{X}^t_r = [\quad x^{group}_{r1} \ldots x^{group}_{rG}]$
- $x^{group}_{rg} = 1$ if run r in group g
  $x^{group}_{rg} = 0$ otherwise



Inference upon summarisation: Protein level model

$$\begin{aligned} y_r &= \beta^{group}_{g(r)} + \epsilon_r \\ &= \mathbf{X}_r^t \boldsymbol{\beta} + \epsilon_r \end{aligned}$$

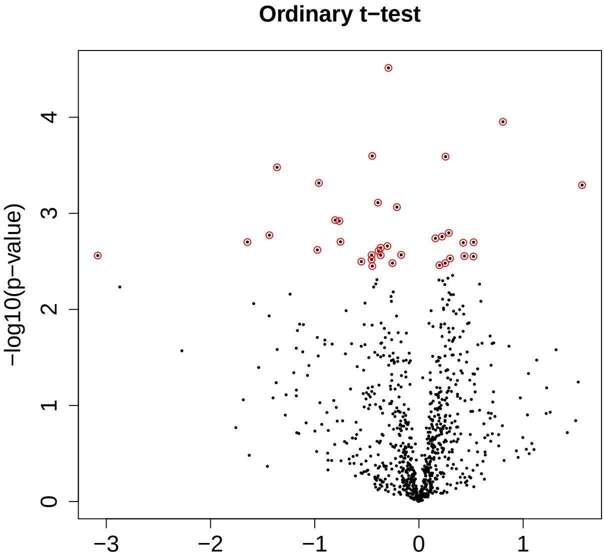
- y_r: protein summary of run r
- $\sum_{g=1}^G \beta_g^{group} = 0$
- $\boldsymbol{\beta} = [\beta_1^{group}, \ldots, \beta_G^{group}]^t$
- $\mathbf{X}_r^t = [\quad x_{r1}^{group} \ldots x_{rG}^{group}]$
- $x_{rg}^{group} = 1$ if run r in group g
  $x_{rg}^{group} = 0$ otherwise



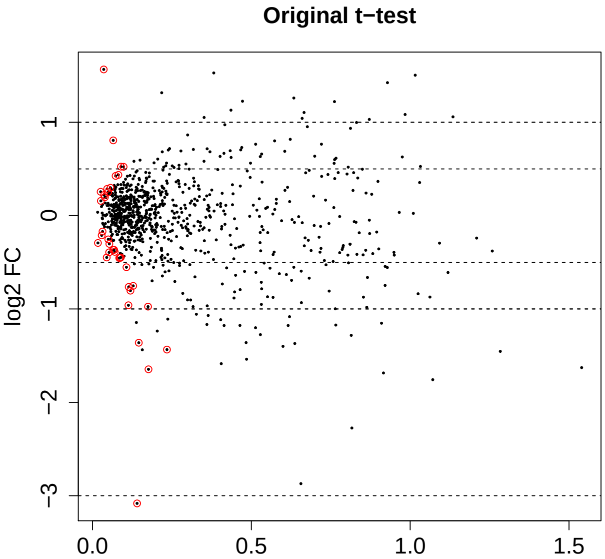
MSqRobSum: robust M-estimation + ridge regression

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.

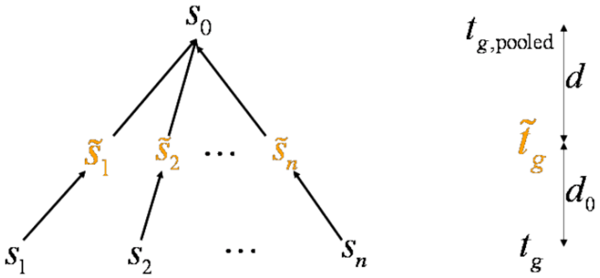
- **empirical Bayes** theory provides formal framework for borrowing strength across genes,
- Implemented in popular bioconductor package **limma**

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

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

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

Shrinkage of Standard Deviations

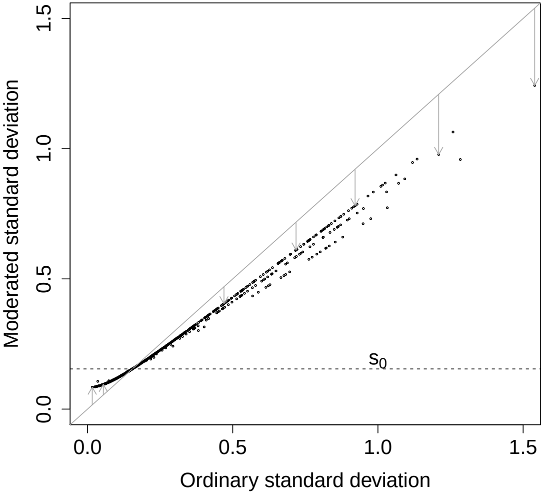


The data decides whether $\tilde{t}_g$

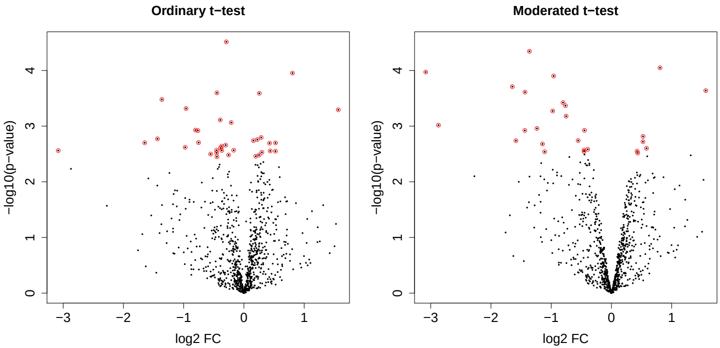
should be closer to $t_{g,pooled}$ or to t_g

Slide courtesy to Rafael Irizarry

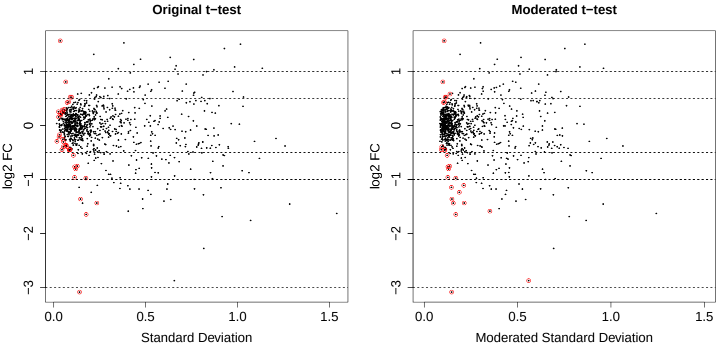
Shrinkage of the variance with limma

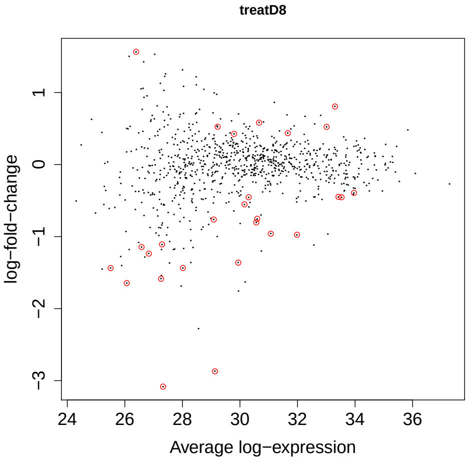


Problems with ordinary t-test solved by moderated EB t-test



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



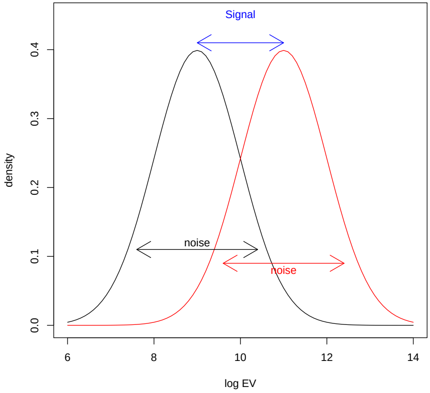

treatD8
log−fold−change
1
0
−1
−2
−3
24
26
28
30
32
34
36
Average log−expression

Breast cancer example

- 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

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!

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

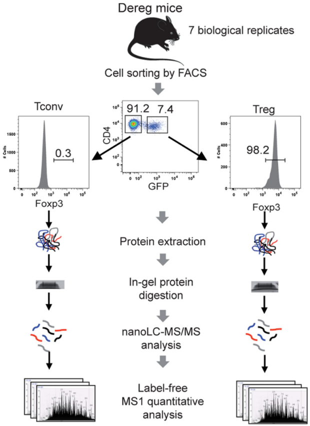
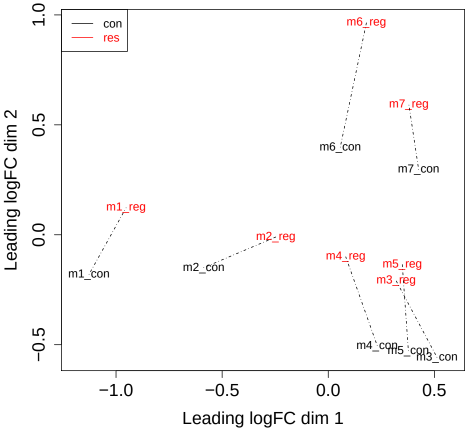


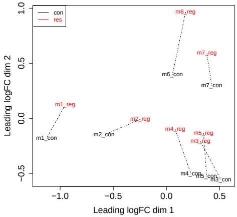
FIG. 1. **Label-free quantitative analysis of conventional and regulatory T cell proteomes.** General analytical workflow based on cell sorting by flow cytometry using the DEREG mouse model and parallel proteomic analysis of Tconv and Treg cell populations by nanoLC-MS/MS and label-free relative quantification.

Blocking Example: mouse T-cells



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:

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

→ use argument fixed=c("type","mouse") in fit.model

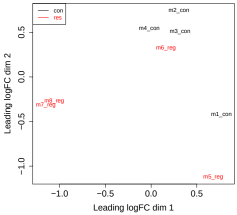
Power gain of blocking

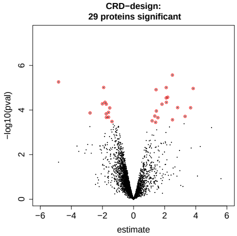
- Completely randomized design (CRD): 8 mice, 4 conventional T-cells, 4 regulatory T-cells.
- Randomized complete block desigh (RBC): 4 mice, for each mouse conventional and regulatory T-cells.

Power gain of blocking

CRD

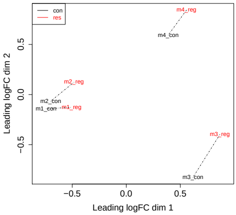
$y \sim \text{type}$

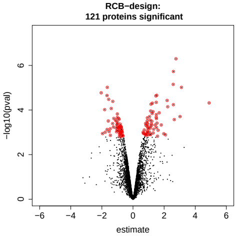




RCB

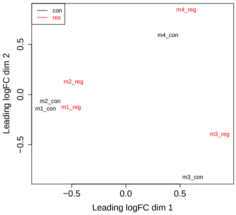
$y \sim \text{type} + \text{mouse}$

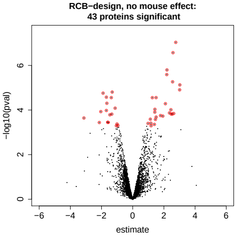




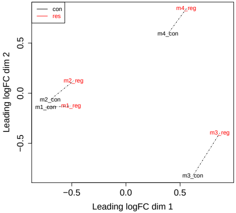
RCB

$y \sim \text{type}$





Anova table: P24452, Capg, Macrophage-capping protein

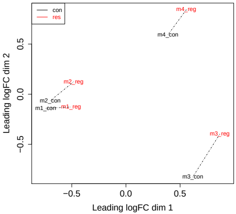


```
### RCB design ###
          Df  Sum Sq Mean Sq  F value   Pr(>F)
type       1 15.2282 15.2282 3720.035 9.71e-06 ***
mouse      3  0.2179  0.0726   17.747  0.02058 *
Residuals  3  0.0123  0.0041
```

```
### 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

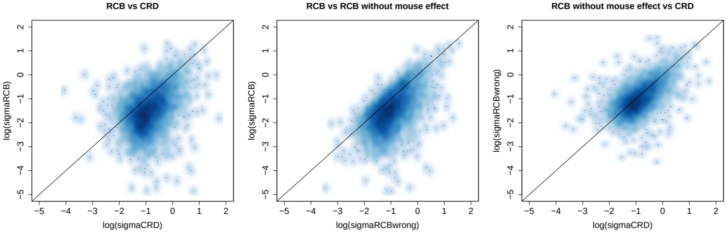


```
### RCB design ###
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.21485    0.05058 439.190 2.60e-08 ***
typereg      2.75937    0.04524  60.992 9.71e-06 ***
mouse2       0.30560    0.06398   4.776   0.0174 *
mouse3      -0.15193    0.06398  -2.375   0.0981 .
mouse4       0.07331    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)