The pitfalls of summarisation and imputation on label free mass spectrometry based proteomics

Lieven Clement

European Bioconductor Meeting 2019, December 9-11, 2019, Brussels, Belgium

statOmics.github.io

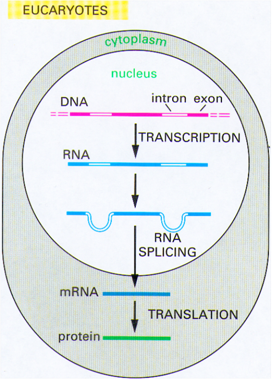
Transcriptomics & single cell omics

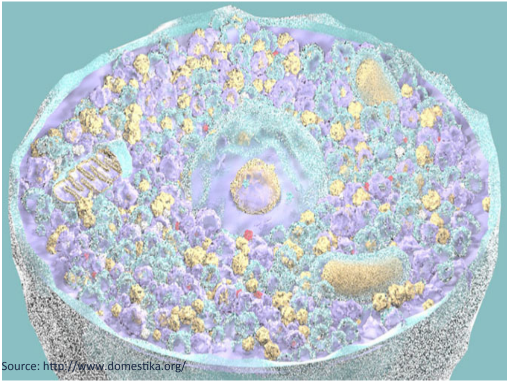
Proteomics

Meta-omics

- Students Adriaan Sticker & Ludger Goeminne
- Lennart Martens Lab (proteomics informatics)
- Kris Gevaert Lab (wetlab)

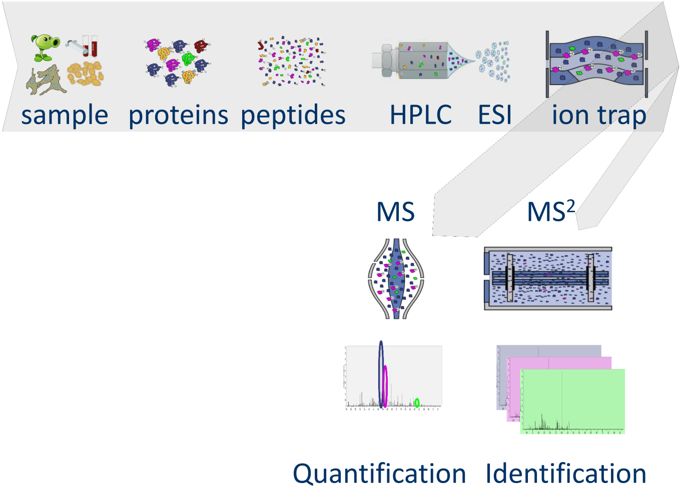
1. Introduction
2. Comparison of popular tools
3. Robust summarisation & Inference
4. Missing Peptides
5. Wrap-up





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

Challenges in Label Free MS-based Quatitative proteomics

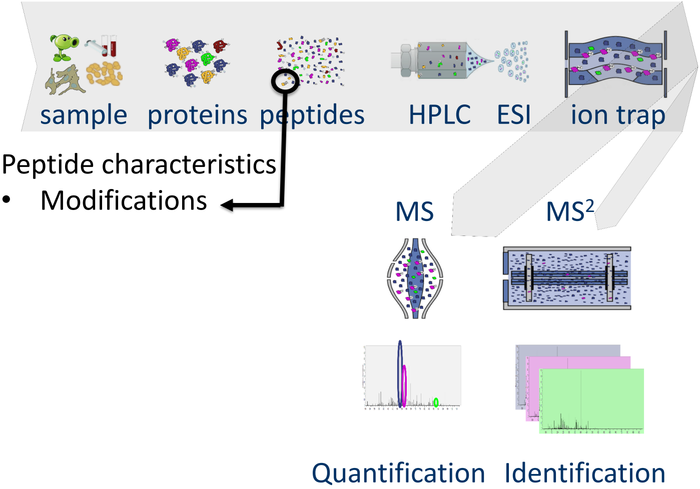


sample proteins peptides HPLC ESI ion trap

MS MS^2

Quantification Identification

Challenges in Label Free MS-based Quatitative proteomics



sample proteins peptides HPLC ESI ion trap

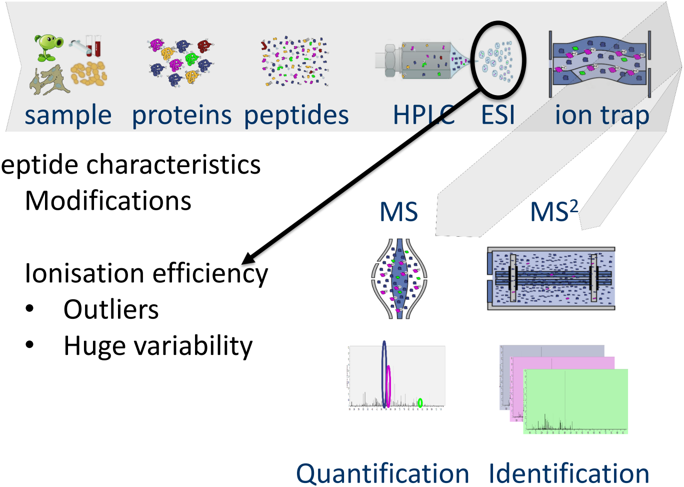
Peptide characteristics

- Modifications

MS MS^2

Quantification Identification

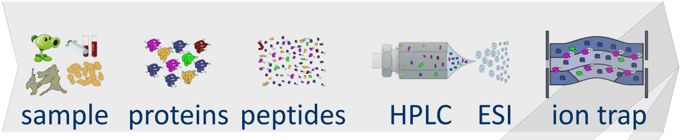
Challenges in Label Free MS-based Quatitative proteomics



Peptide characteristics

- Modifications
- Ionisation efficiency
    - Outliers
    - Huge variability

Challenges in Label Free MS-based Quatitative proteomics

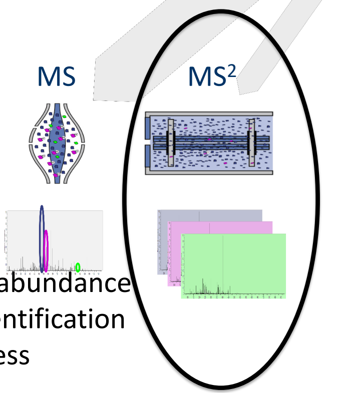


Peptide characteristics

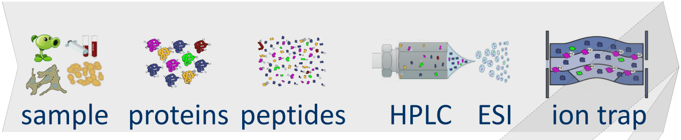
- Modifications

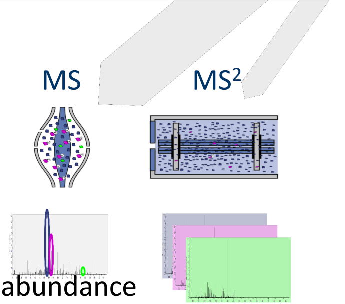
- Ionisation efficiency
  - Outliers
  - Huge variability

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



Challenges in Label Free MS-based Quatitative proteomics





Peptide characteristics

- Modifications

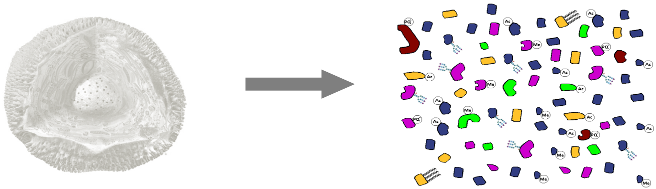
- Ionisation efficiency
  - Outliers
  - Huge variability

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

Unbalanced peptides identifications across samples and messy data

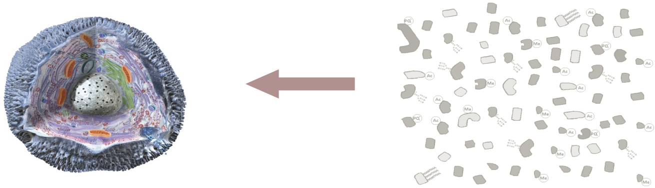
Challenges in Label Free MS-based Quatitative proteomics

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

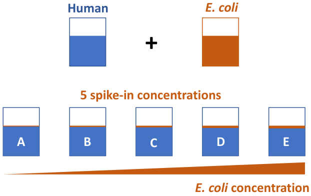


Challenges in Label Free MS-based Quatitative proteomics

We need information on protein level!



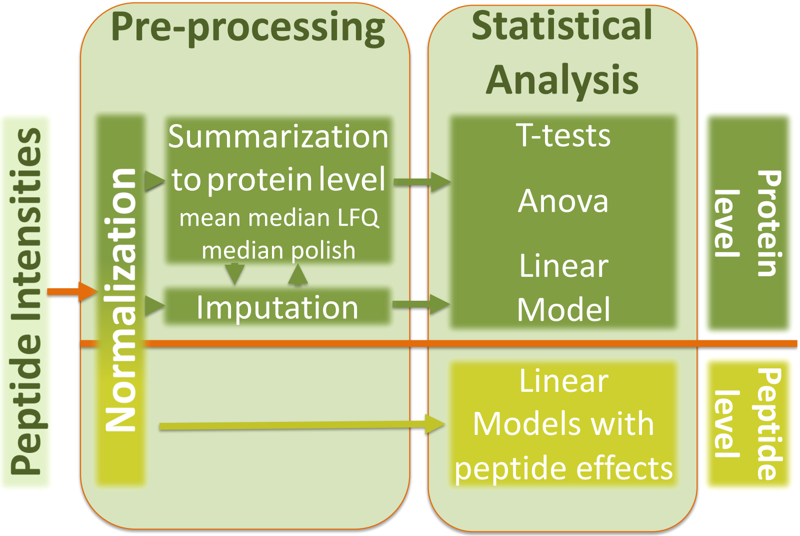
Spike-in study (Shen et al. 2018)



- 4 repeats per spike-in condition
- Trypsin-digested human proteome
- After MaxQuant search with match between runs option
  - Only 50% of all peptides are quantified in all samples

→ **vast amount of missingness**

1. Introduction
2. **Comparison of popular tools**
    1. Overview tools
    2. Difference in performance
    3. Impact of summarisation
3. Robust summarisation & Inference
4. Missing Peptides
5. Wrap-up



Peptide Intensities

Normalization

Pre-processing

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

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

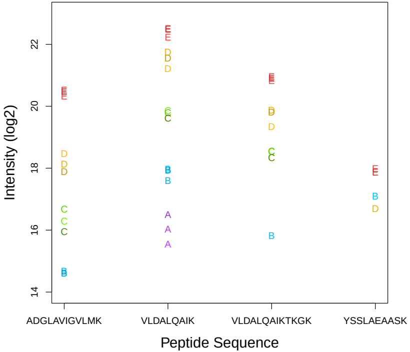
$$y_{grp} = \beta_g^{group} + u_r^{\text{run}} + \beta_p^{\text{pep}} + \epsilon_{rp}$$

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)

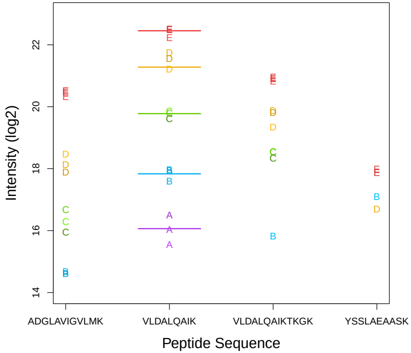
$$y_{grp} = \beta_g^{group} + u_r^{\text{run}} + \beta_p^{\text{pep}} + \epsilon_{rp}$$

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)

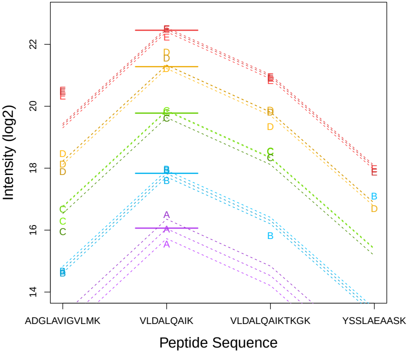
$$y_{grp} = \beta_g^{group} + u_r^{\text{run}} + \beta_p^{\text{pep}} + \epsilon_{rp}$$

protein-level

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

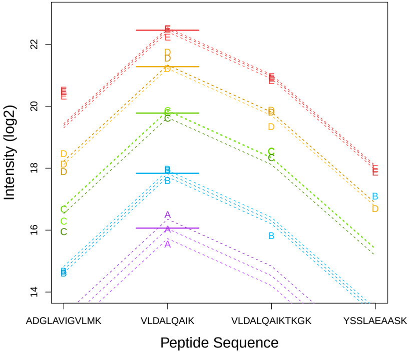
- β_g^{group}: spike-in
- random run effect $u_r^{run} \sim N\left(0, \sigma^2_{run}\right)$
  $\rightarrow$ Addresses pseudo-replication

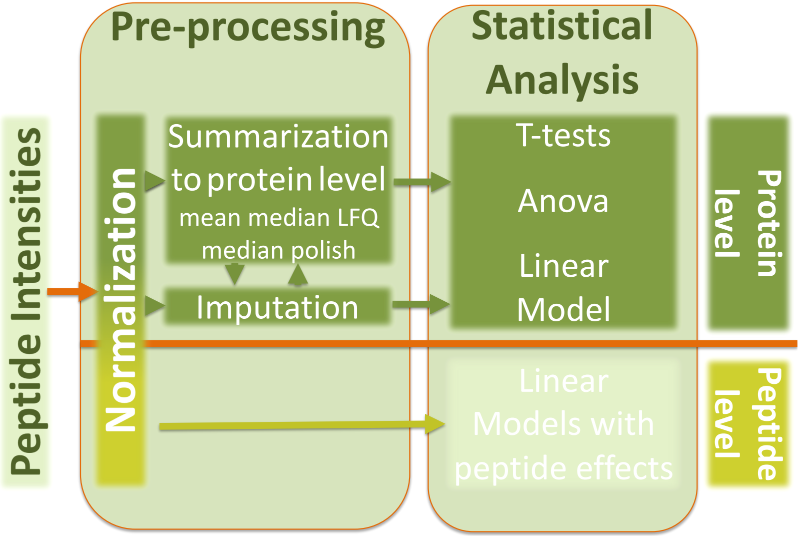
peptide-level

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

Estimation

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





Pre-processing

Statistical Analysis

Peptide Intensities

Normalization

Summarization to protein level

mean median LFQ

median polish

Imputation

T-tests

Anova

Linear Model

Protein level

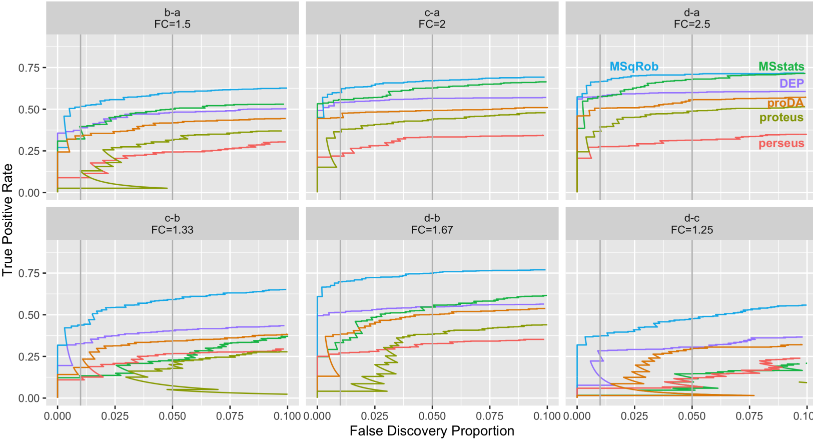
Linear Models with peptide effects

Peptide level

Summarisation based methods

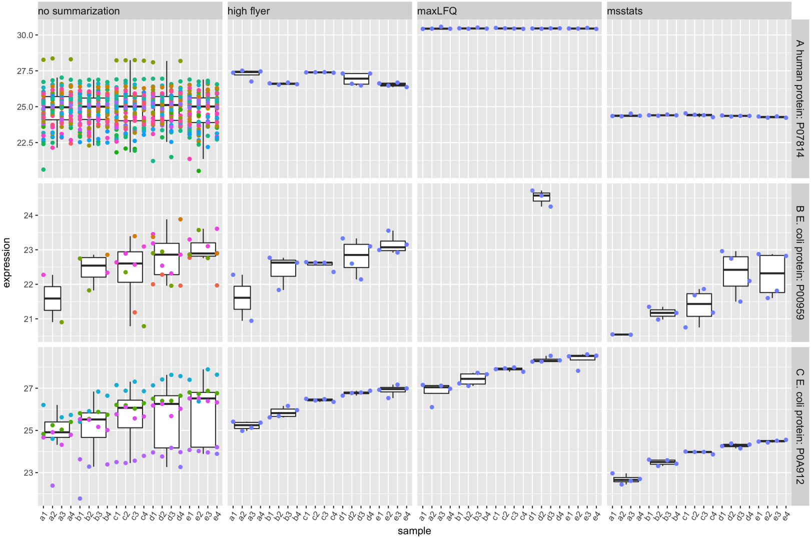
- Perseus: MaxLFQ summarization & Inference with t-test
- Proteus
  - Summarization: average of 3 high-flyers
  - Inference: limma (linear model + EB)
- DEP
  - Summarization: MaxLFQ
  - Imputation at protein level: missingness at random and by low abundance
  - Inference: limma
- proDA
  - Summarization: MaxLFQ
  - probabilistic dropout model
  - Inference: linear model + EB
- MS-stats
  - Summarization with peptide-based model (median polish)
  - Imputation at peptide level: missingness by low abundance
  - Inference: linear model

$$TPR = \frac{TP}{TP + FN} = \frac{\text{E. coli}}{\text{All E. coli}}$$



$$FDP = \frac{FP}{TP + FP} = \frac{\text{Human}}{\text{E. coli+Human}}$$

Summarisation



1. Introduction
2. Comparison of popular tools
3. **Robust summarisation & Inference**
    1. Robust Summarisation
    2. Robust Inference
    3. Results
4. Missing Peptides
5. Wrap-up

Fit MSqRob mixed model in two-stage approach

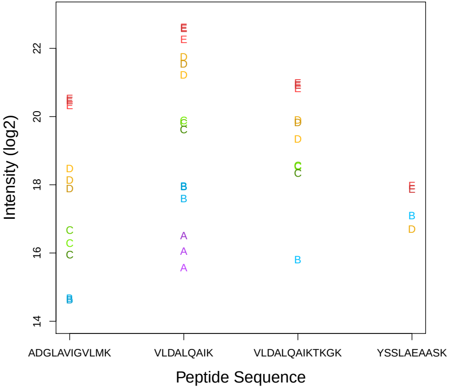
MSqRob

- 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

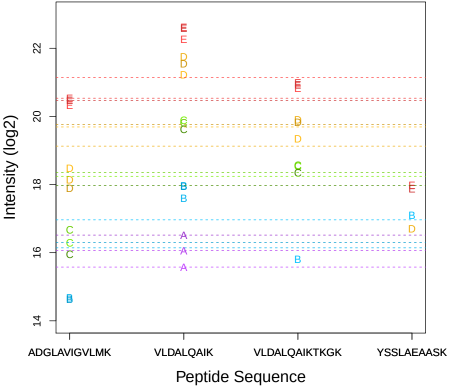
→ Modular approach

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

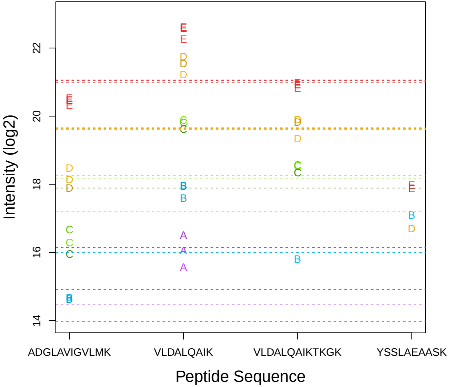


Protein by protein analysis of peptide data with linear model

peptide level protein level

$$y_{rp} = \epsilon_{rp} \quad + \quad \beta_r^{run}$$

Summarisation with peptide based model



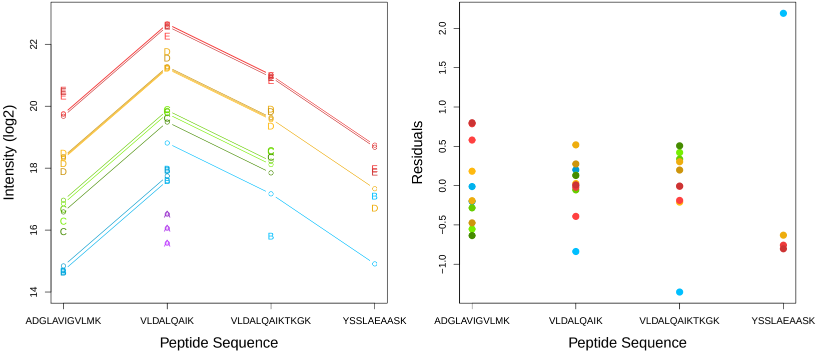
Protein by protein analysis of peptide data with linear model

peptide level protein level

$$y_{rp} = \beta_p^{\text{pep}} + \epsilon_{rp} \quad + \quad \beta_r^{\text{run}}$$

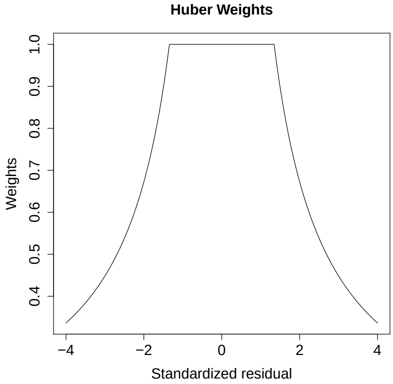
Robust estimation using observation weights

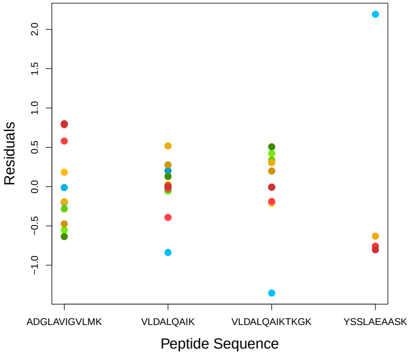
- Outlying peptide intensities: incorrect peptide identification, post-translational modifications, ...



Robust estimation using observation weights

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

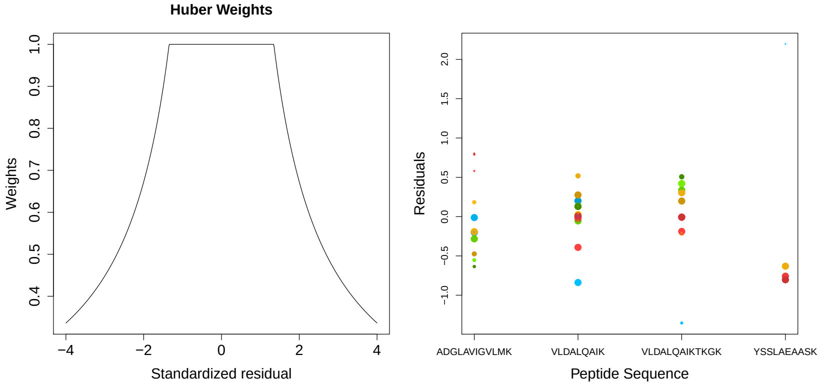




- M-estimation: implemented by iteratively fitting model with observation weights

Robust estimation using observation weights

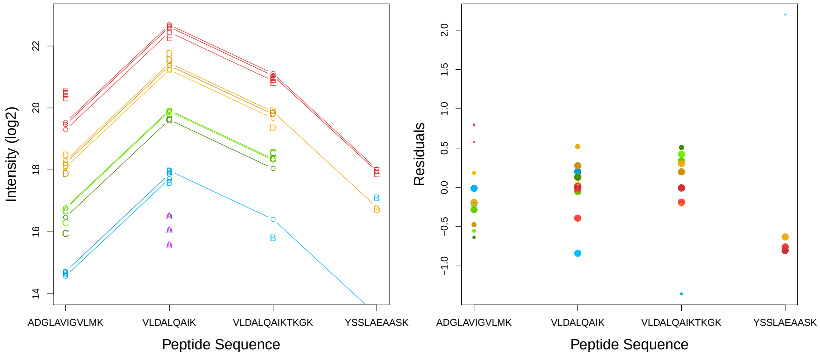
- Outlying peptide intensities: incorrect peptide identification, post-translational modifications, ...



- M-estimation: implemented by iteratively fitting model with observation weights

Robust estimation using observation weights

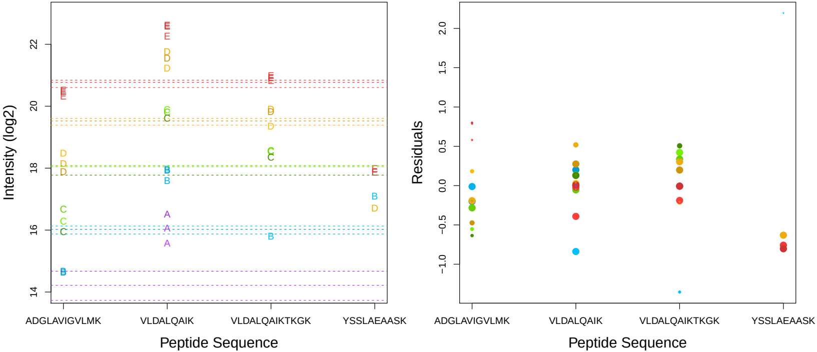
- Outlying peptide intensities: incorrect peptide identification, post-translational modifications, ...



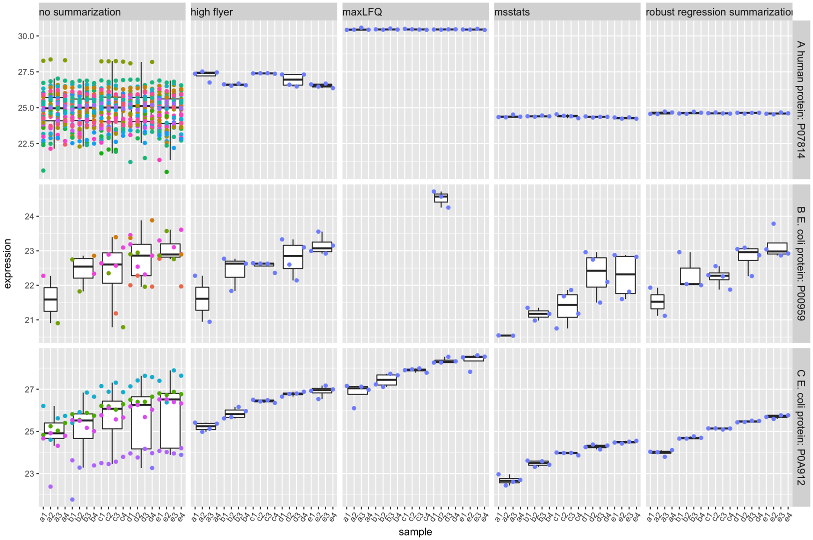
- M-estimation: implemented by iteratively fitting model with observation weights

Robust estimation using observation weights

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



- M-estimation: implemented by iteratively fitting model with observation weights

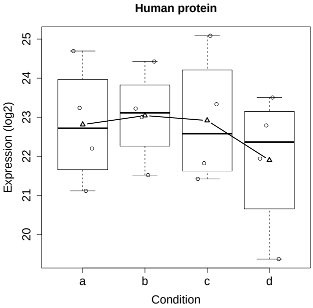

no summarization
high flyer
maxLFQ
msstats
robust regression summarizatio
A human protein: P07814
B E. coli protein: P00959
C E. coli protein: P0A912
expression
30.0
27.5
25.0
22.5
24
23
22
21
27
25
23
sample
a1 a2 a3 a4 b1 b2 b3 b4 c1 c2 c3 c4 d1 d2 d3 d4 e1 e2 e3 e4

1. Introduction
2. Comparison of popular tools
3. Robust summarisation & inference
    1. Robust Summarisation
    2. **Robust Inference**
    3. Results
4. Missing Peptides
5. Wrap-up

Inference upon summarisation: Protein level model

$$y_r = \beta_0 + \beta_{g(r)}^{group} + \epsilon_r$$

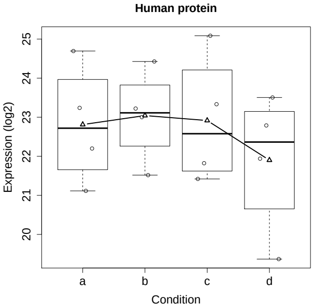
- y_r: protein summary of run r
- $\sum_{g=1}^{G} \beta_g^{group} = 0$



Inference upon summarisation: Protein level model

$$
\begin{aligned}
y_r &= \beta_0 + \beta_{g(r)}^{group} + \epsilon_r \\
&= \mathbf{X}_r^t \boldsymbol{\beta} + \epsilon_r
\end{aligned}
$$

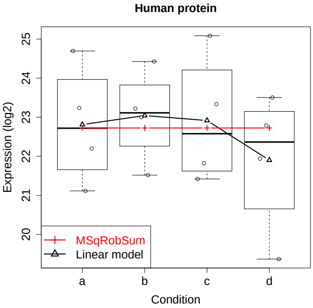
- y_r: protein summary of run r
- $\sum_{g=1}^{G} \beta_g^{group} = 0$
- $\boldsymbol{\beta} = [\beta_0, \beta_1^{group}, \ldots, \beta_G^{group}]^t$
- $\mathbf{X}_r^t = [1 \quad x_{r1}^{group} \ldots x_{rG}^{group}]$
- $x_{rg}^{group} = 1$ if run r in group g
  $x_{rg}^{group} = 0$ otherwise



Inference upon summarisation: Protein level model

$$\begin{aligned} y_r &= \beta_0 + \beta^{group}_{g(r)} + \epsilon_r \\ &= \mathbf{X}_r^t \boldsymbol{\beta} + \epsilon_r \end{aligned}$$

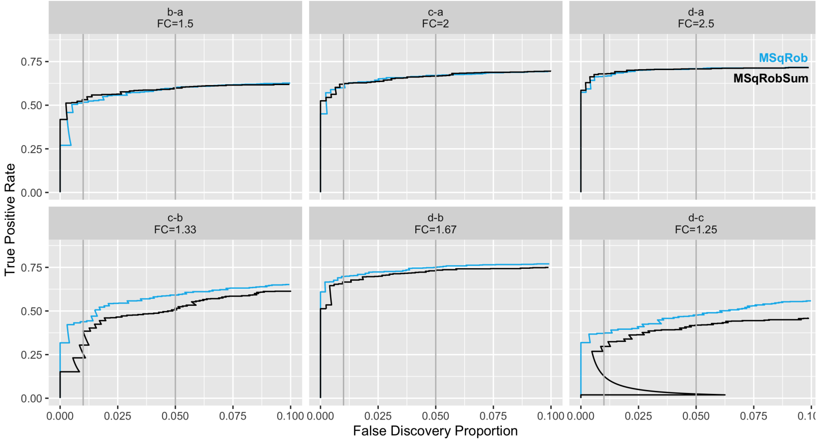
- y_r: protein summary of run r
- $\sum_{g=1}^{G} \beta^{group}_g = 0$
- $\boldsymbol{\beta} = [\beta_0, \beta^{group}_1, \ldots, \beta^{group}_G]^t$
- $\mathbf{X}_r^t = [1 \quad x^{group}_{r1} \ldots x^{group}_{rG}]$
- $x^{group}_{rg} = 1$ if run r in group g
  $x^{group}_{rg} = 0$ otherwise



MSqRobSum: robust M-estimation + ridge regression

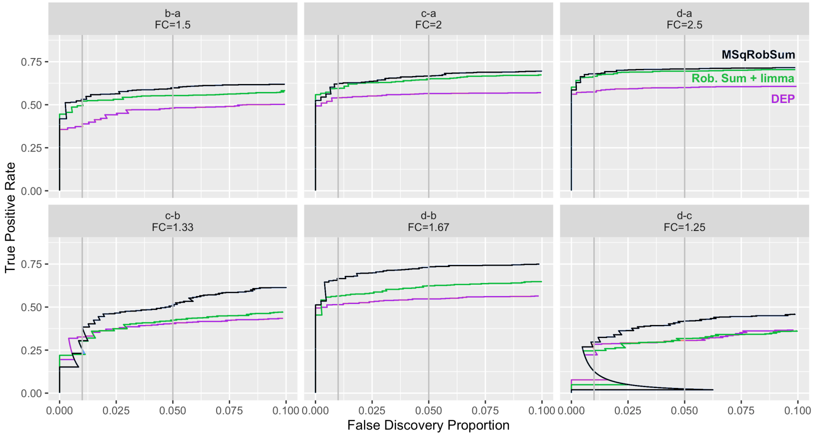
1. Introduction
2. Comparison of popular tools
3. Robust summarisation & Inference
    1. Robust Summarisation
    2. Robust Inference
    3. **Results**
        - MSqRobSum vs MSqRob
        - Modular Approach
        - Fold Change Estimates
4. Missing Peptides
5. Wrap-up

MSqRobSum vs MSqRob

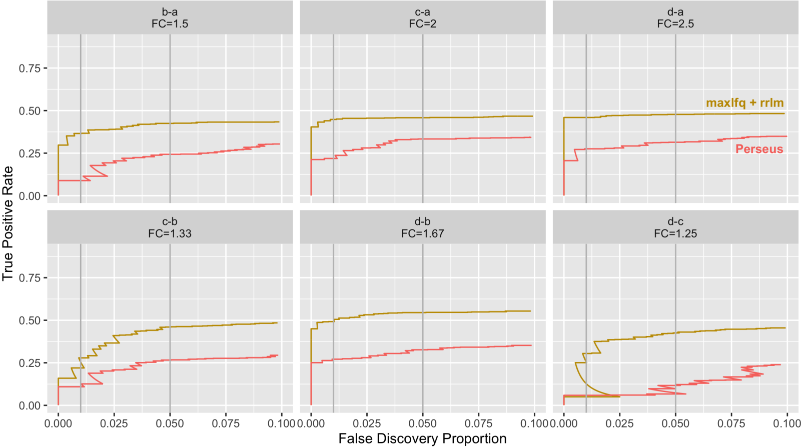


- Still very good performance
- 3 times faster
- df well defined
- Summaries for visualisation

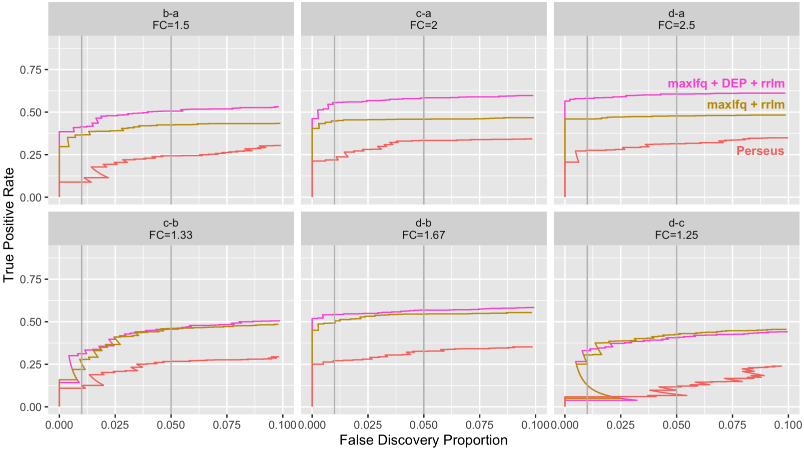
MSqRobSum vs DEP



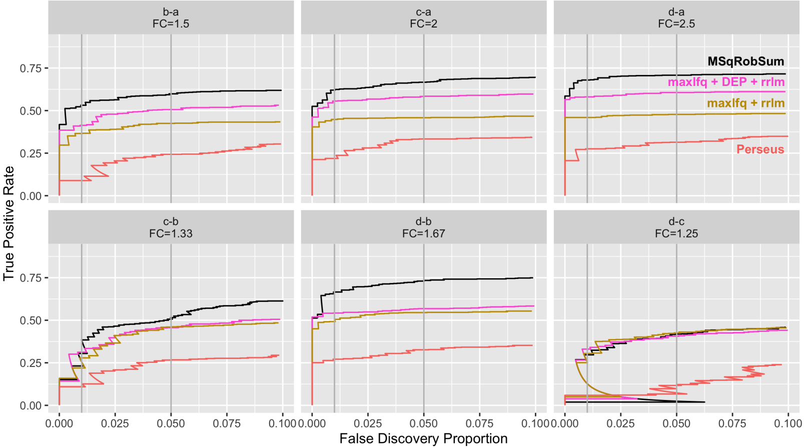
Summarisation & inference are modular



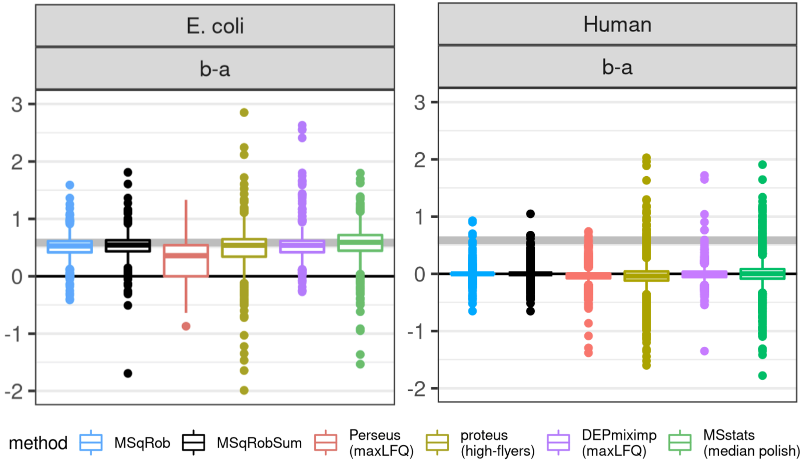
Summarisation & inference are modular



Summarisation & inference are modular

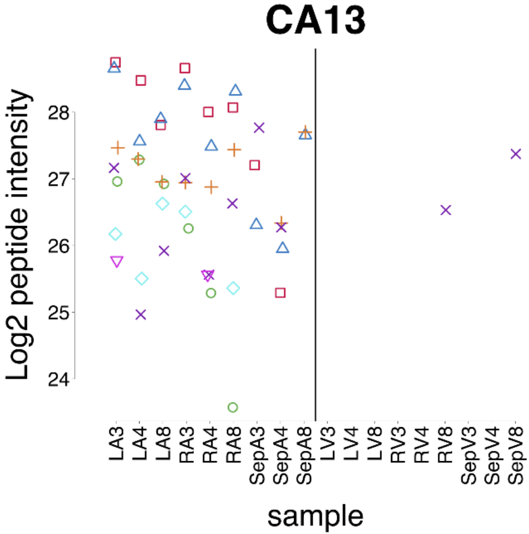


Fold change estimates



1. Introduction
2. Comparison of popular tools
3. Robust summarisation & Inference
4. Missing Peptides
5. Wrap-up

Missing peptides



Hurdle Model

$$\begin{cases} z_{pr} | x_{pr} & \sim & B(\pi_r) \\ y_{pr} | z_{pr} = 1, x_{pr}, u_r^{run} & \sim & N(\mu_{pr}, \sigma^2) \end{cases}$$

- binary component z_{pr} with detection probability π_r

$$\begin{aligned} z_{pr} &= 0\text{: Peptide intensity is missing} \\ z_{pr} &= 1\text{: Peptide intensity is observed} \end{aligned}$$

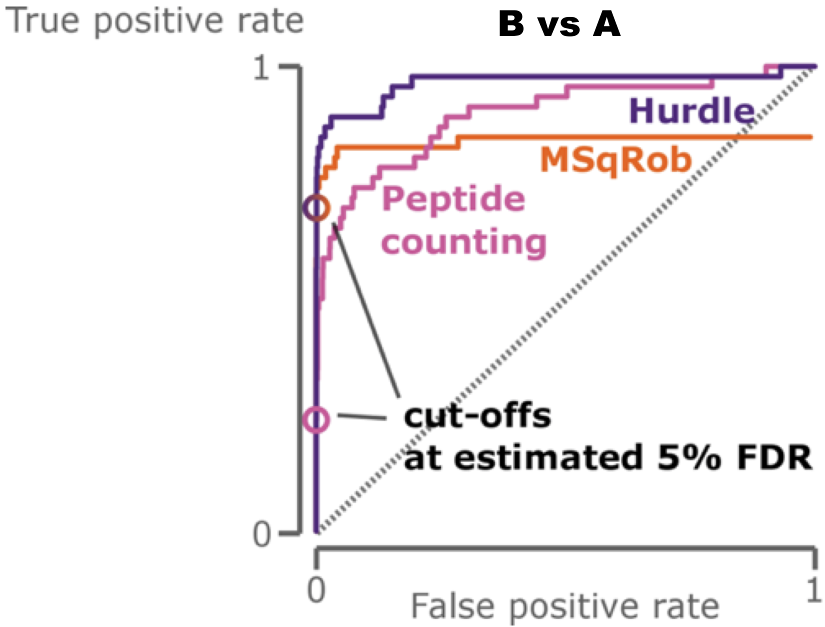
- Normal component for log2-transformed intensities y_{pr} for peptide $p = 1, \ldots, P$ in run $r = 1, \ldots, R$

Hurdle Model

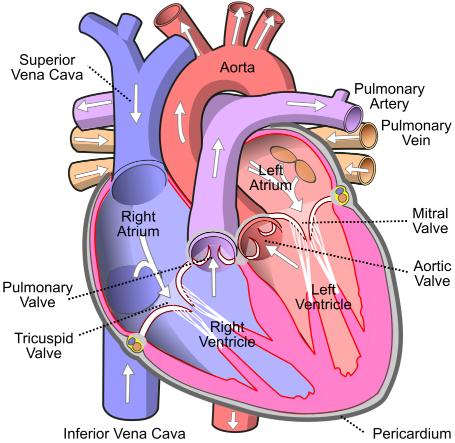
$$\begin{cases} z_{pr}|x_{pr} & \sim & B(\pi_r) \\ y_{pr}|z_{pr} = 1, x_{pr}, u_r^{run} & \sim & N(\mu_{pr}, \sigma^2) \end{cases}$$

- Likelihood of the model implies an estimation ortogonality
- Estimation and inference on π_r via logistic regression of peptide presence absence: differential detection
- Estimation and inference on mu_{pr} via MSqRob model: differential expression given detection
- Combine inference on both components using stageR

B vs A

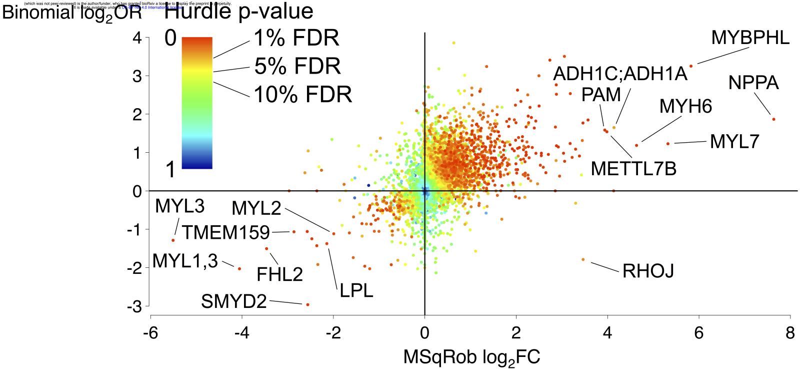

True positive rate
1
0
0
1
False positive rate
Hurdle
MSqRob
Peptide counting
cut-offs at estimated 5% FDR

Case Study: Doll et al. 2017. PRIDE: PXD006675

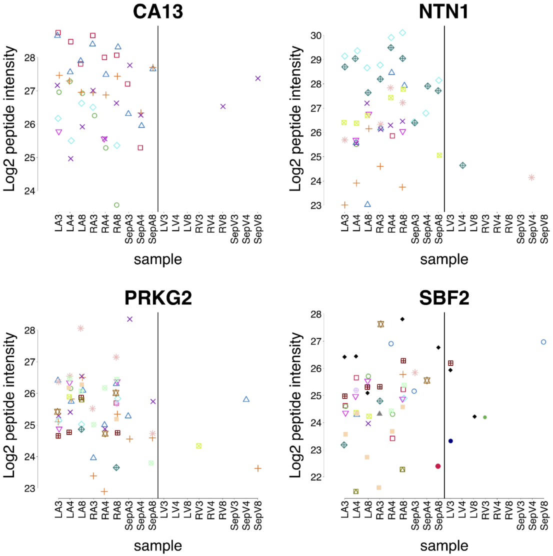


- 3 patients
- biopsies of multiple heart regions
  - Left Atrium
  - Right Atrium
  - Atrial Septum
  - Left Ventriculum
  - Right Ventriculum
  - Ventriculum Septum
  - ...

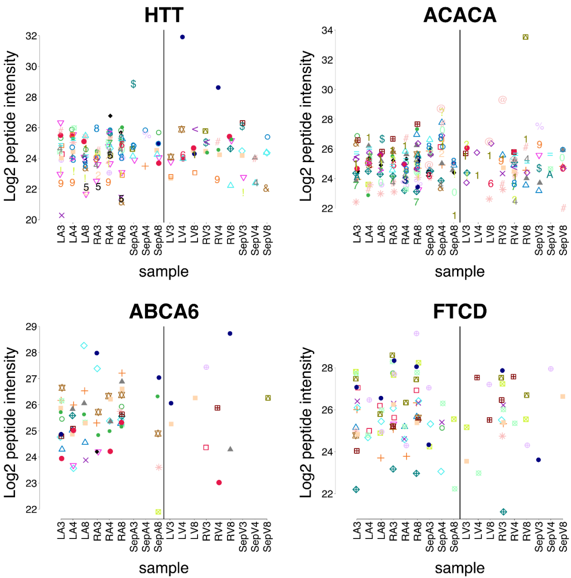
Case Study: Doll et al. 2017. PRIDE: PXD006675



Case Study: Doll et al. 2017. PRIDE: PXD006675



Case Study: Doll et al. 2017. PRIDE: PXD006675



Wrap-up

1. Summarization has to account for peptide effects
2. Imputation can be very detrimental
3. Robust summarisation can avoid imputation to some extend
4. Robust inference with linear models further improves the performance

5. Hurdle model builds upon missing peptides without needing rigid assumptions

6. Preprint on summarization: Sticker et al. 2019 biorxiv http://dx.doi.org/10.1101/668863
7. Preprint on hurdle model: Goeminne et al. 2019 https://doi.org/10.1101/782466
8. Robust summarization is also implemented as a method in the combineFeatures of the MSnBase bioconductor package.

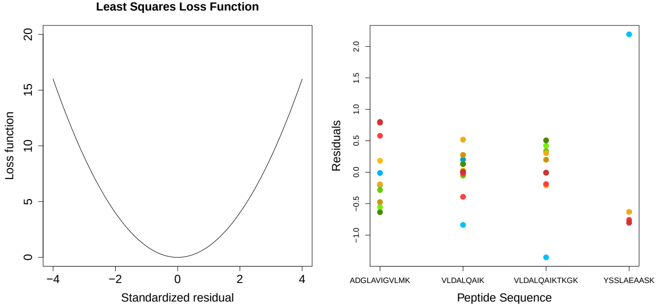
statOmics is hiring predocs and postdocs

statOmics.github.io

Ghent, Belgium

M-estimation

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

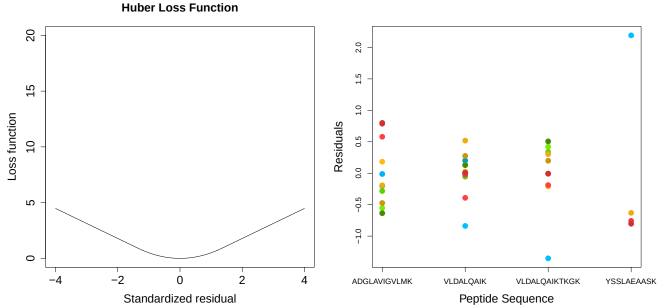


Least Squares Loss Function

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

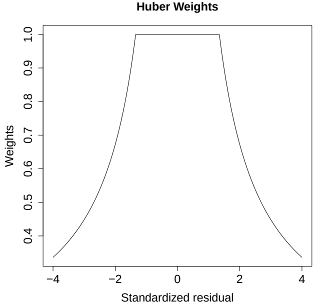
M-estimation

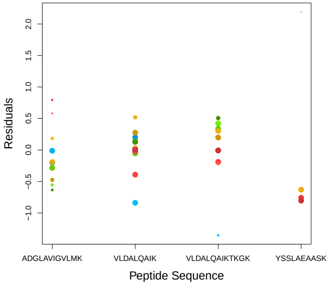
- Outlying peptide intensities: incorrect peptide identification, post-translational modifications, ...



M-estimation

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



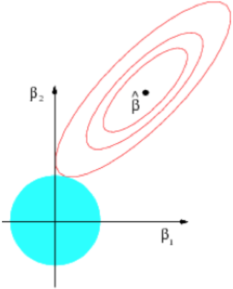


$$\text{argmin}_{\beta^{\text{pep}}_{1\ldots P},\beta^{\text{samp}}_{1\ldots n}}\left[\sum_{r=1}^{n}\sum_{p}^{P} w(\epsilon_{rp})\left(y_{rp}-\beta^{\text{pep}}_{p}-\beta^{\text{run}}_{r}\right)^{2}\right]$$

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

Ridge regression

$$y_r = \mathbf{X}_r^t \boldsymbol{\beta} + \epsilon_r$$



Parameters estimation with loss function:

$$\text{argmin} \sum_{r=1}^{n} w(d_r) \left(y_r - \mathbf{X}_r^t \boldsymbol{\beta}\right)^2 \quad + \quad \lambda \sum_{g \neq 0} (\beta_g)^2$$

with λ: penalty term for regularization of parameters of interest estimated using link between ridge regression and mixed models

Ridge regression

Tune the ridge penalties by exploiting the link between ridge regression and Mixed Models:

$$y_r = \mathbf{X}_r^t \boldsymbol{\beta} + \epsilon_r$$

with

- $\beta_g \sim N\left(0, \frac{\sigma^2}{\lambda}\right)$ with $g = 1, \ldots, J$
- $\epsilon_r \sim N\left(0, \sigma^2\right)$
- Variance components are estimated using lme4 mixed model software
- Predictions of the random effects β_g coincide with solution of ridge estimator.

MaxLFQ summarization

a

>P63208
MPSIK**LQSSDGEIFEVDVEIAK**QSVTIKTMLEDLGMDDEGDD
DPVPLPNVNAAILKKVIQWCTHHKDDPPPPEDDENKEK**RTDD**
IPVWDQEFLKVDQGTLFELILAANYLDIKGLLDVTCK**TVANM**
IKGK**TPEEIRK**TFNIK**NDFTEEEEAQVR**KENQWCEEK

b

Peptide species	Sequence	Charge	Mod.
P_1	LQSSDGEIFEVDVEIAK	2	–
P_2	LQSSDGEIFEVDVEIAK	3	–
P_3	RTDDIPVWDQEFLK	2	–
P_4	TVANMIK	2	–
P_5	TVANMIK	2	Oxid.
P_6	TPEEIRK	3	–
P_7	NDFTEEEEAQVR	2	–

c

Sample	P_1	P_2	P_3	P_4	P_5	P_6	P_7
A		+				+	
B		+	+			+	
C	+	+	+	+		+	+
D	+	+		+		+	+
E		+		+			+
F		+			+		

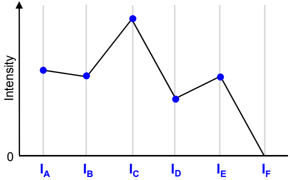
d

	A	B	C	D	E	F
A						
B	r_{BA}					
C	r_{CA}	r_{CB}				
D	r_{DA}	r_{DB}	r_{DC}			
E	r_{EA}	r_{EB}	r_{EC}	r_{ED}		
F	r_{FA}	r_{FB}	r_{FC}	r_{FD}	r_{FE}	

e

$r_{BA} = I_B / I_A$	$r_{CA} = I_C / I_A$	$r_{CB} = I_C / I_B$
$r_{DA} = I_D / I_A$	$r_{DB} = I_D / I_B$	$r_{DC} = I_D / I_C$
$r_{EC} = I_E / I_C$	$r_{ED} = I_E / I_D$	$I_F = 0$

f


Intensity
I_A
I_B
I_C
I_D
I_E
I_F