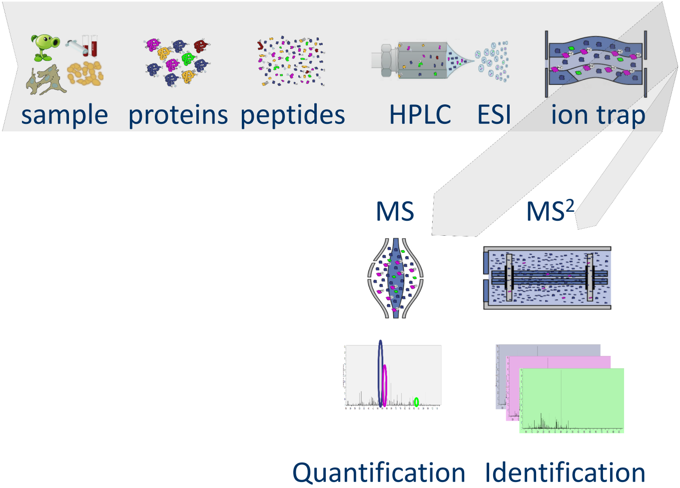
Part I: Normalization & Summarization

Lieven Clement

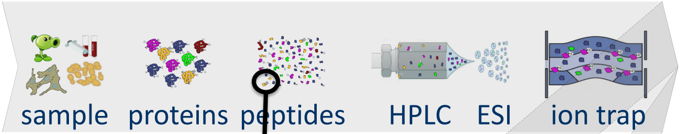
Proteomics Data Analysis Shortcourse

Outline

Challenges in Label Free Quantitative Proteomics

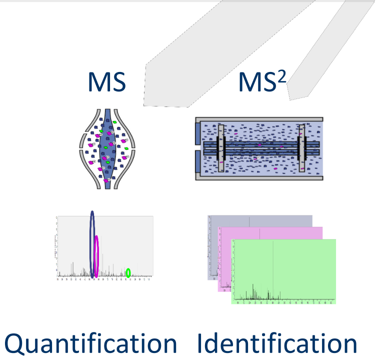


Challenges in Label Free Quantitative Proteomics

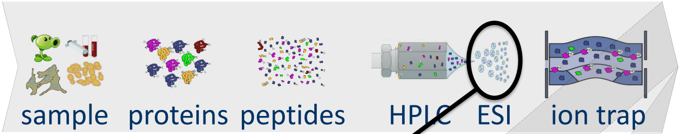


Peptide characteristics

- Modifications



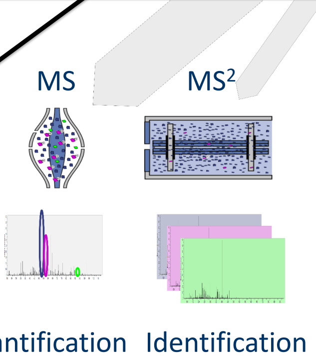
Challenges in Label Free Quantitative Proteomics



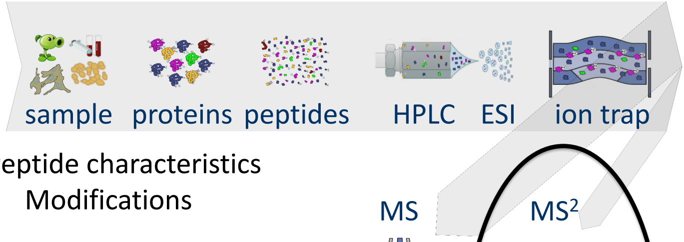
Peptide characteristics

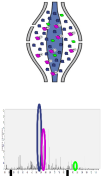
- Modifications

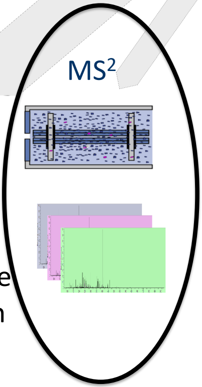
- Ionisation efficiency
  - Outliers
  - Huge variability



Challenges in Label Free Quantitative Proteomics



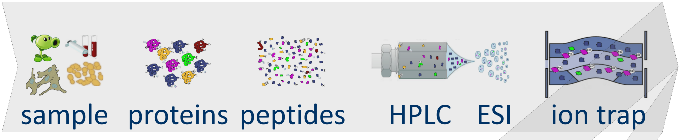




Peptide characteristics

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

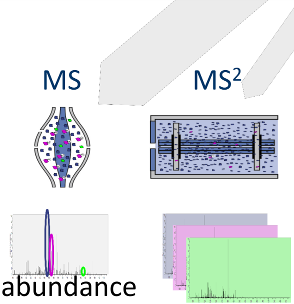
Challenges in Label Free Quantitative Proteomics



Peptide characteristics

- Modifications

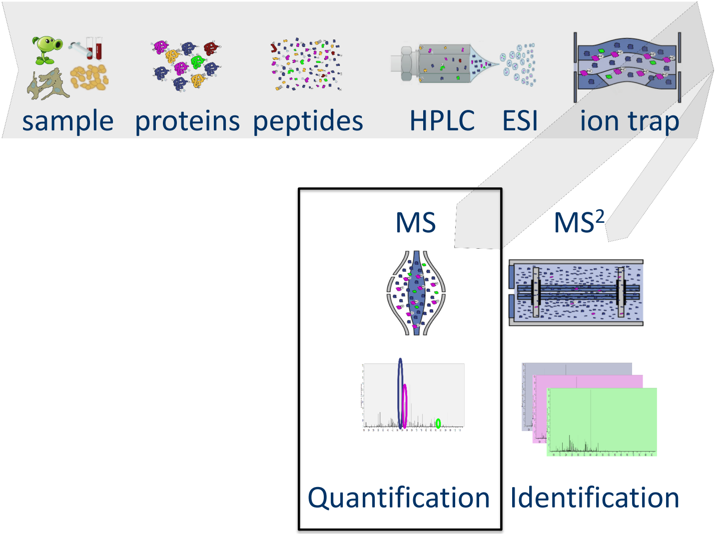
- Ionisation efficiency
    - Outliers
    - Huge variability



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

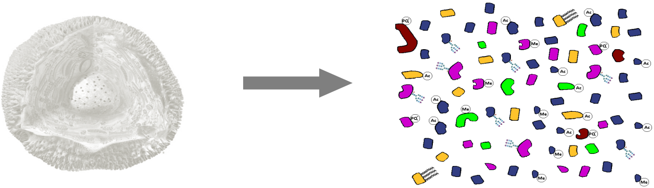
Unbalanced peptides identifications across samples and messy data

Challenges in Label Free MS-based Quatitative proteomics



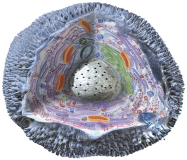
Challenges in Label Free MS-based Quatitative proteomics

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

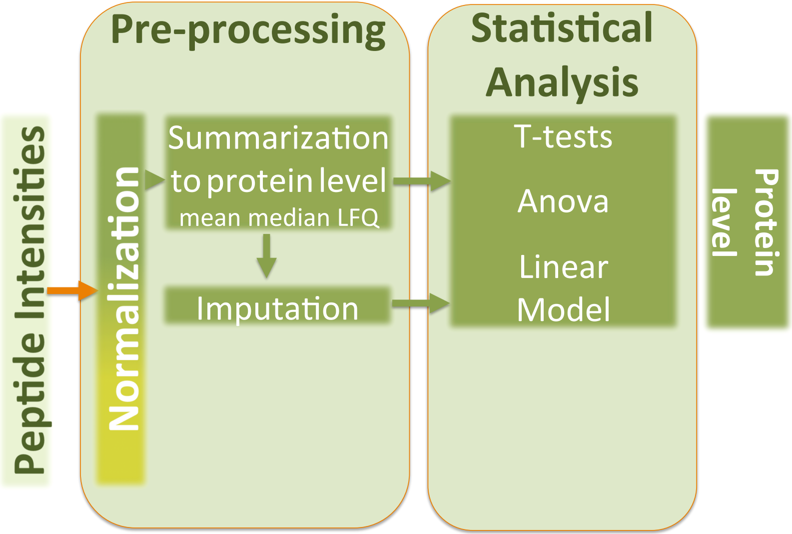


Challenges in Label Free MS-based Quatitative proteomics

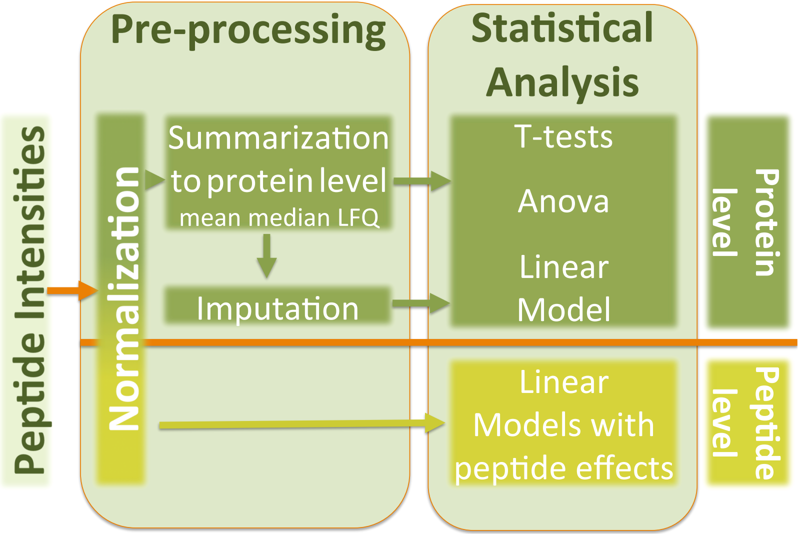
We need information on protein level!



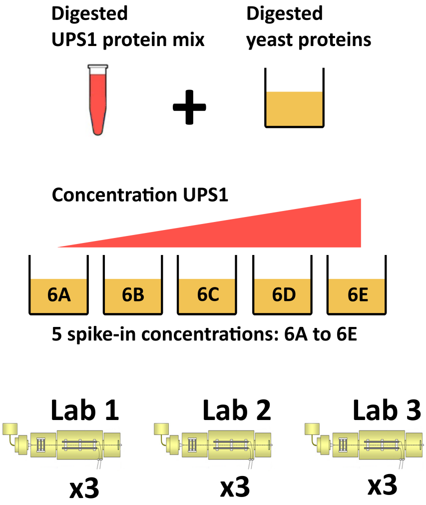
Label-free Quantitative Proteomics Data Analysis Pipelines



Label-free Quantitative Proteomics Data Analysis Pipelines



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**

Preprocessing

- Typical preprocessing steps
  1. Filtering
  2. Log-transformation
  3. Normalization
  4. (Summarization)

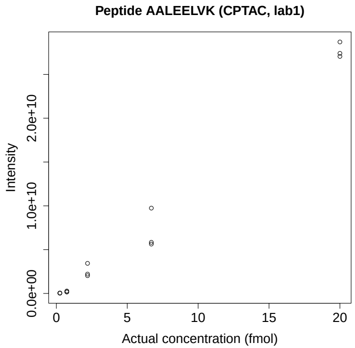
- Many methods exist

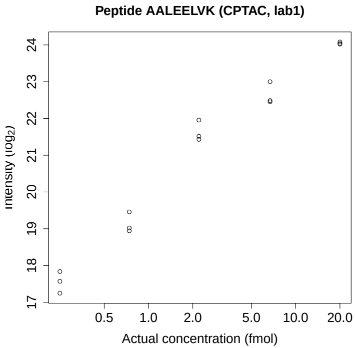
Filtering

- Reverse sequences
- Only identified by modification site (only modified peptides detected)
- Razor peptides: non-unique peptides assigned to the protein group with the most other peptides
- Contaminants
- Peptides few identifications
- Proteins that are only identified with one or a few peptides

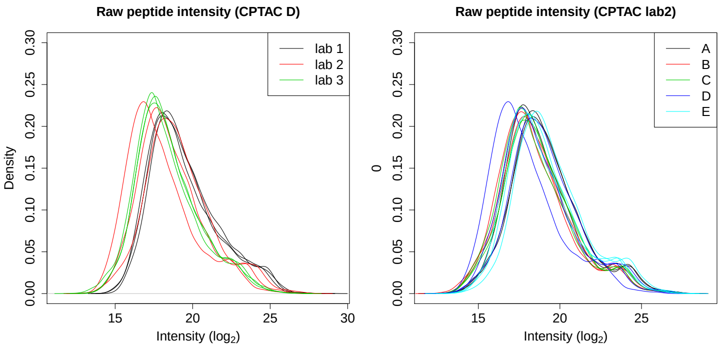
- Filtering does not induce bias if the criterion is independent from the downstream data analysis!

Log-transformation





Variability more equal upon log transformation: often multiplicative error structure of intensity-based read-outs



Even in very clean synthetic dataset (same background, only 48 UPS proteins can be different) the marginal peptide intensity distribution across samples can be quite distinct

- Considerable effects between and within labs for replicate samples
- Considerable effects between samples with different spike-in concentration

→ Normalization is needed

Mean or median?

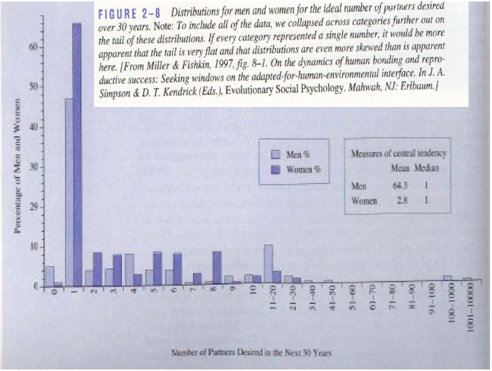
- Over a period of 30 years males desire to have on average 64.3 partners and females 2.8. (Miller and Fishkin, 1997)

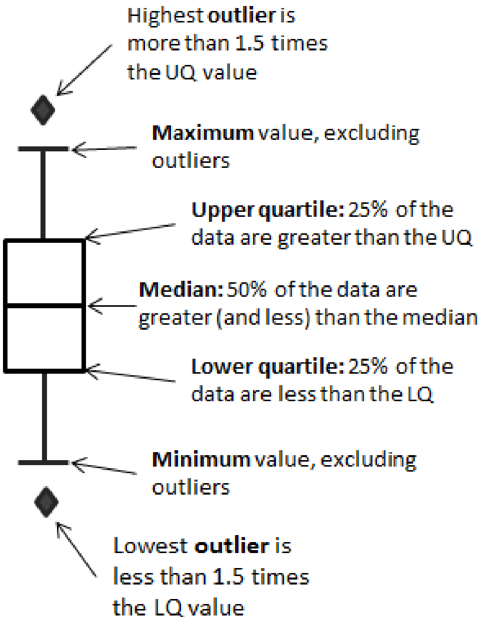
Mean or median?

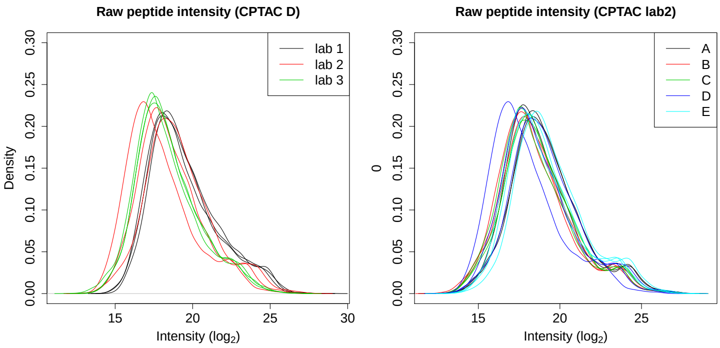
- Over a period of 30 years males desire to have on average 64.3 partners and females 2.8. (Miller and Fishkin, 1997)
- Over a period of 30 years males, is the median of the number of desired partners is 1 for both males and females. (Miller and Fishkin, 1997)

Mean or median?

Mean is very sensitive to outliers!



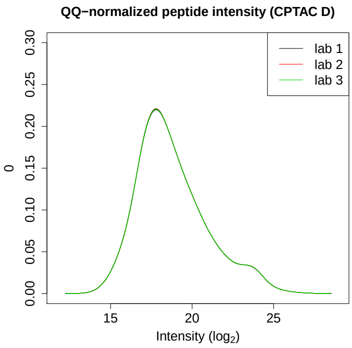

Highest **outlier** is more than 1.5 times the UQ value
Maximum value, excluding outliers
Upper quartile: 25% of the data are greater than the UQ
Median: 50% of the data are greater (and less) than the median
Lower quartile: 25% of the data are less than the LQ
Minimum value, excluding outliers
Lowest **outlier** is less than 1.5 times the LQ value

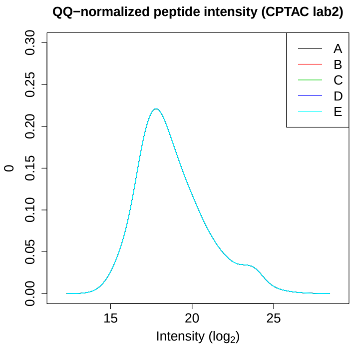


Even in very clean synthetic dataset (same background, only 48 UPS proteins can be different) the marginal peptide intensity distribution across samples can be quite distinct

- Considerable effects between and within labs for replicate samples
- Considerable effects between samples with different spike-in concentration

$\rightarrow$ Normalization is needed



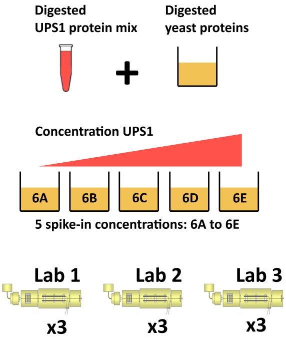


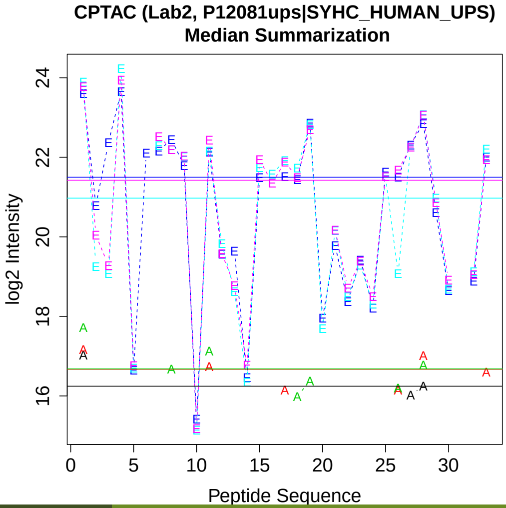
Even in very clean synthetic dataset (same background, only 48 UPS proteins can be different) the marginal peptide intensity distribution across samples can be quite distinct

- Considerable effects between and within labs for replicate samples
- Considerable effects between samples with different spike-in concentration

$\rightarrow$ Normalization is needed, e.g. **quantile normalization**

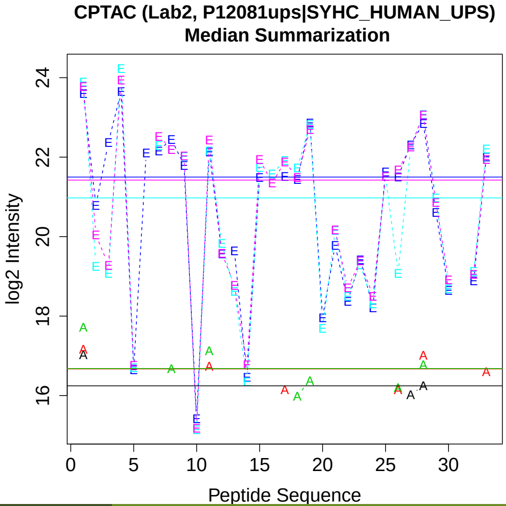
Summarization





Summarization

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



Impact of summarization?

Subsample of cptac study

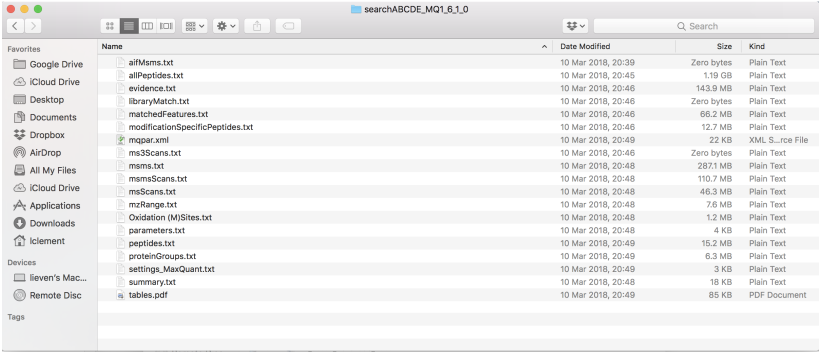
- Lab 3
- Spike in concentration A vs B

→ Two group comparison

1. Median summarization
2. MaxLFQ (MaxQuant)
3. Robust summarization:

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

MaxQuant output



searchABCDE_MQ1_6_1_0

Name	Date Modified	Size	Kind
aifMsms.txt	10 Mar 2018, 20:39	Zero bytes	Plain Text
allPeptides.txt	10 Mar 2018, 20:45	1.19 GB	Plain Text
evidence.txt	10 Mar 2018, 20:46	143.9 MB	Plain Text
libraryMatch.txt	10 Mar 2018, 20:46	Zero bytes	Plain Text
matchedFeatures.txt	10 Mar 2018, 20:46	66.2 MB	Plain Text
modificationSpecificPeptides.txt	10 Mar 2018, 20:46	12.7 MB	Plain Text
mqpar.xml	10 Mar 2018, 20:49	22 KB	XML S...rce File
ms3Scans.txt	10 Mar 2018, 20:46	Zero bytes	Plain Text
msms.txt	10 Mar 2018, 20:48	287.1 MB	Plain Text
msmsScans.txt	10 Mar 2018, 20:48	110.7 MB	Plain Text
msScans.txt	10 Mar 2018, 20:48	46.3 MB	Plain Text
mzRange.txt	10 Mar 2018, 20:48	7.6 MB	Plain Text
Oxidation (M)Sites.txt	10 Mar 2018, 20:48	1.2 MB	Plain Text
parameters.txt	10 Mar 2018, 20:48	4 KB	Plain Text
peptides.txt	10 Mar 2018, 20:49	15.2 MB	Plain Text
proteinGroups.txt	10 Mar 2018, 20:49	6.3 MB	Plain Text
settings_MaxQuant.txt	10 Mar 2018, 20:49	3 KB	Plain Text
summary.txt	10 Mar 2018, 20:48	18 KB	Plain Text
tables.pdf	10 Mar 2018, 20:49	85 KB	PDF Document

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		+			+		

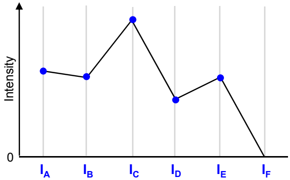
d

	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}	

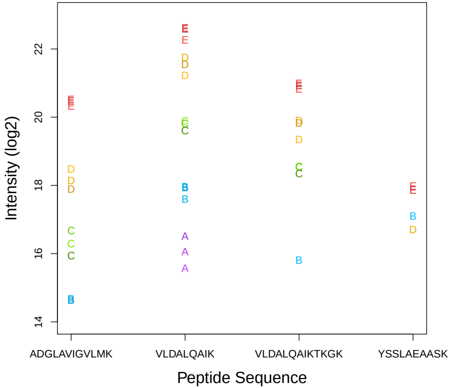
e

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

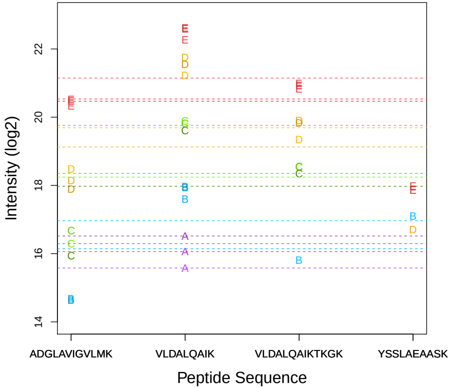
f



Summarisation with peptide based model



Summarisation with peptide based model

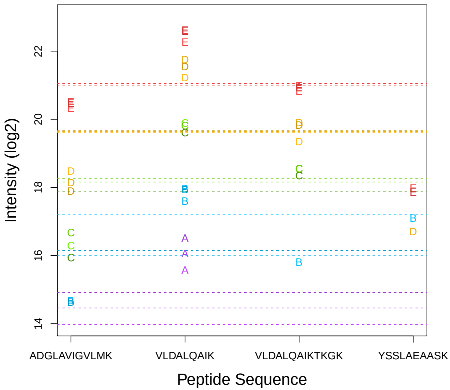


Protein by protein analysis of peptide data with linear model

peptide level protein level

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

Summarisation with peptide based model

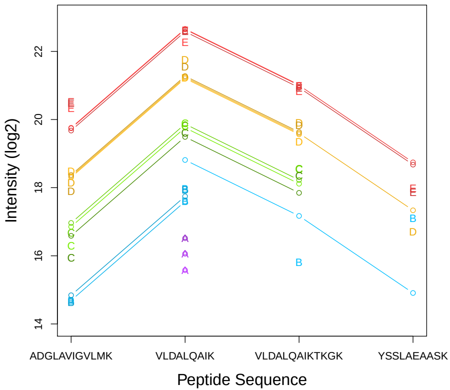


Protein by protein analysis of peptide data with linear model

peptide level protein level

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

Summarisation with peptide based model

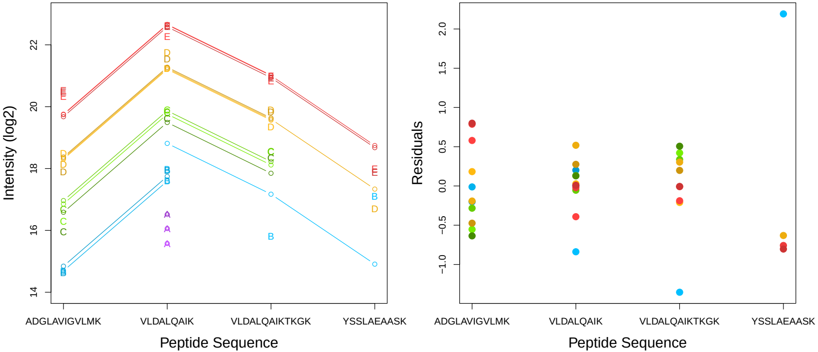


Protein by protein analysis of peptide data with linear model

peptide level protein level

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

Summarisation with peptide based model

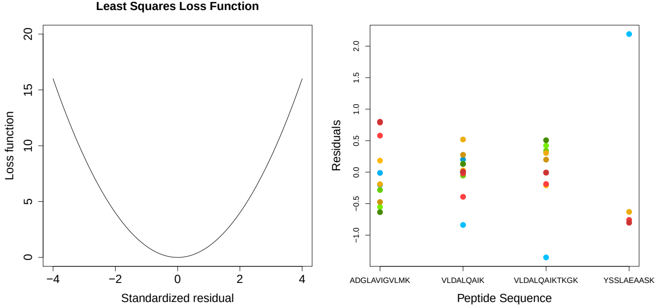


Protein by protein analysis of peptide data with linear model

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

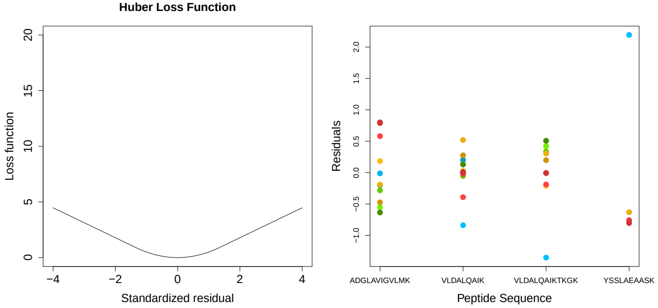
Robust estimation using observation weights

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



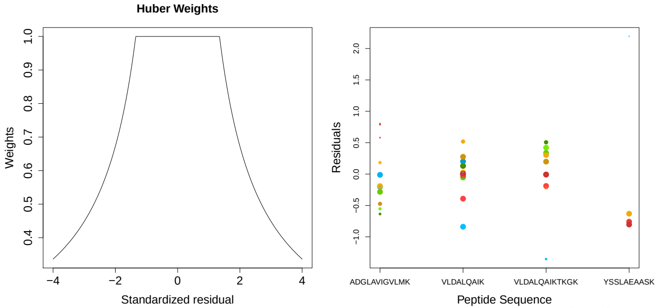
Robust estimation using observation weights

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



Robust estimation using observation weights

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

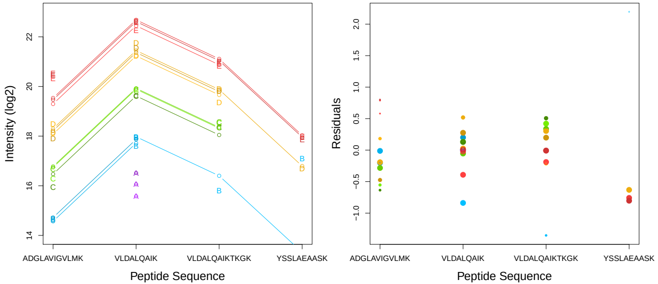


- Iteratively fit model with observation weights $w(\epsilon_{ip})$

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

Robust estimation using observation weights

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

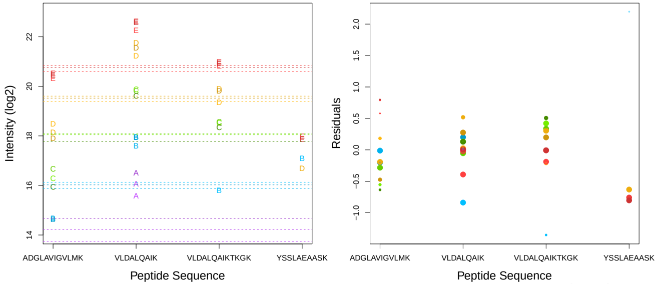


- Iteratively fit model with observation weights $w(\epsilon_{ip})$

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

Robust estimation using observation weights

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



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

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