

# Part I: Normalization & Summarization

Lieven Clement

Proteomics Data Analysis Shortcourse

# Outline

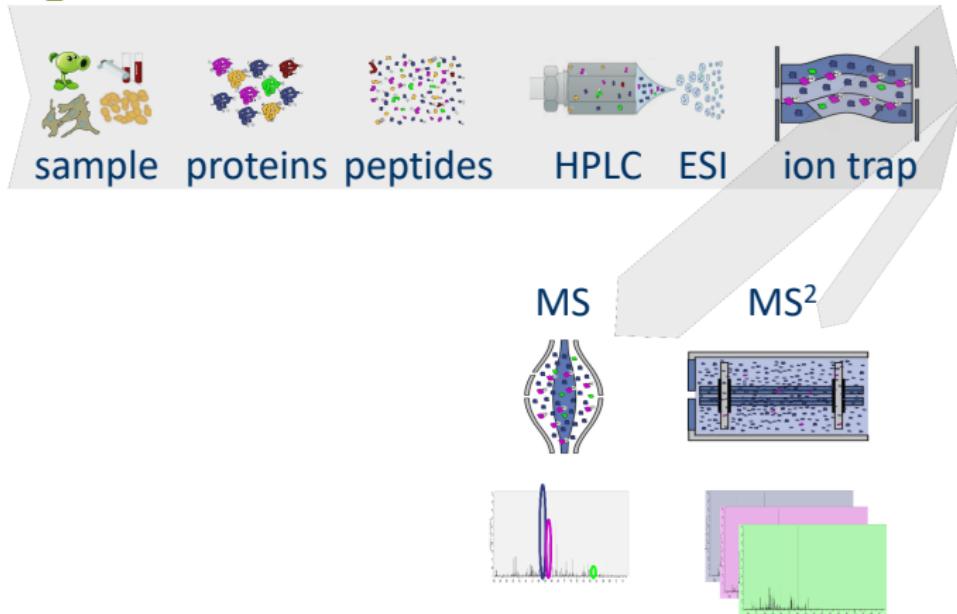
## ① Introduction

- ① Label free MS based Quantitative Proteomics Workflow and Challenges

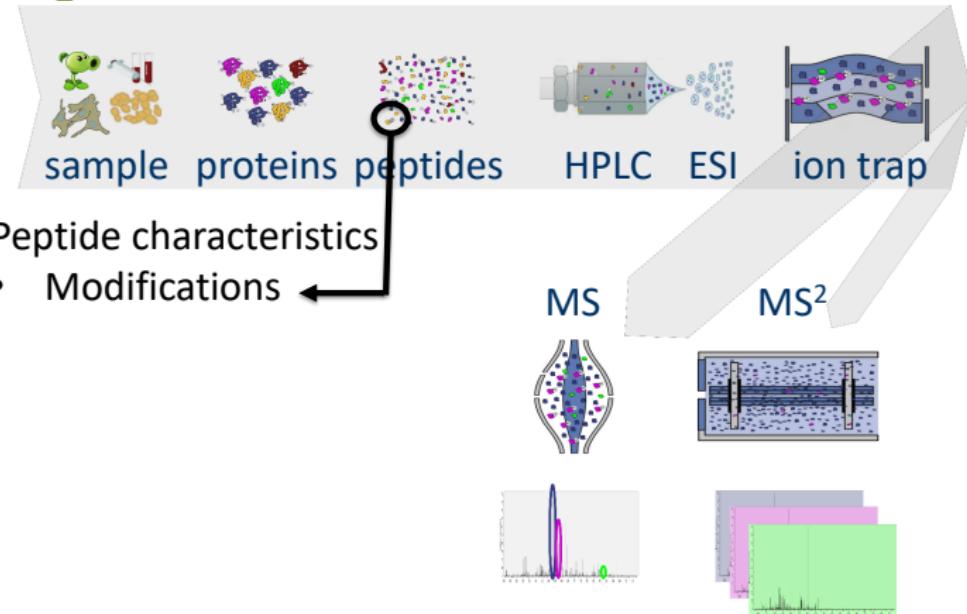
## ② Preprocessing

- ① Filtering
- ② Log transformation
- ③ Normalization
- ④ Summarization

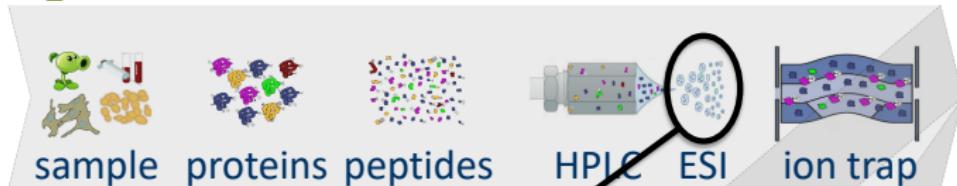
# Challenges in Label Free Quantitative Proteomics



# Challenges in Label Free Quantitative Proteomics

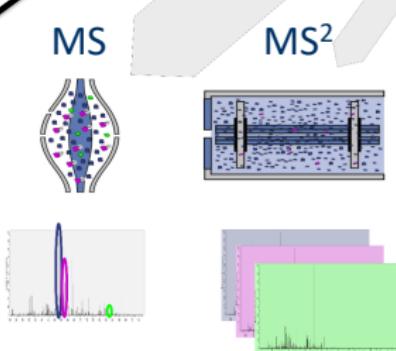


# Challenges in Label Free Quantitative Proteomics



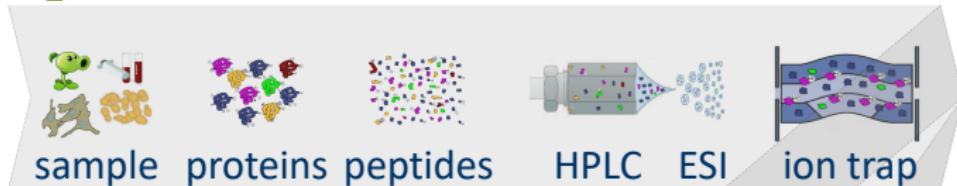
## Peptide characteristics

- Modifications
- Ionisation efficiency
  - Outliers
  - Huge variability



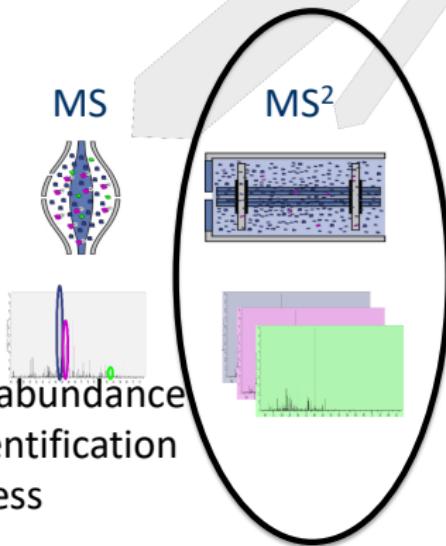
Quantification Identification

# Challenges in Label Free Quantitative Proteomics



## Peptide characteristics

- Modifications
- Ionisation efficiency
  - Outliers
  - Huge variability
- MS<sup>2</sup> selection on peptide abundance
  - Context dependent Identification
  - Non-random missingness

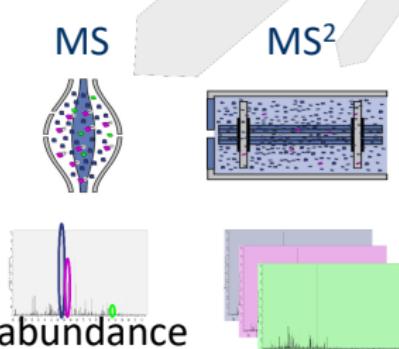


# Challenges in Label Free Quantitative Proteomics



## Peptide characteristics

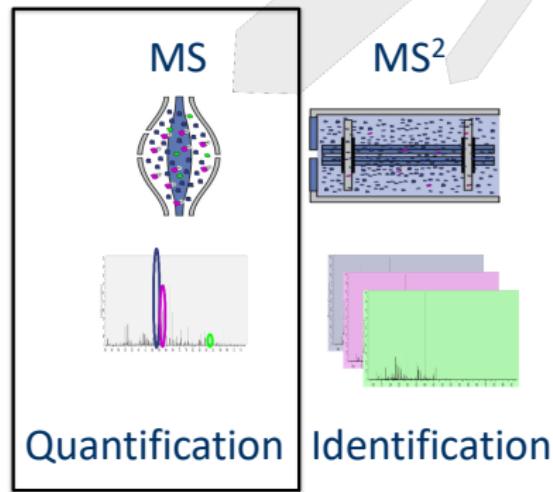
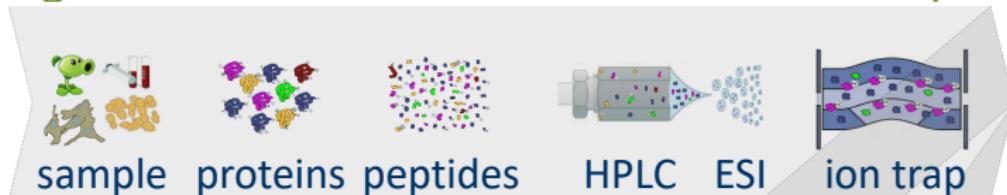
- Modifications
- Ionisation efficiency
  - Outliers
  - Huge variability
- MS<sup>2</sup> selection on peptide abundance
  - Context dependent Identification
  - Non-random missingness



**Unbalanced peptides identifications across samples and messy data**

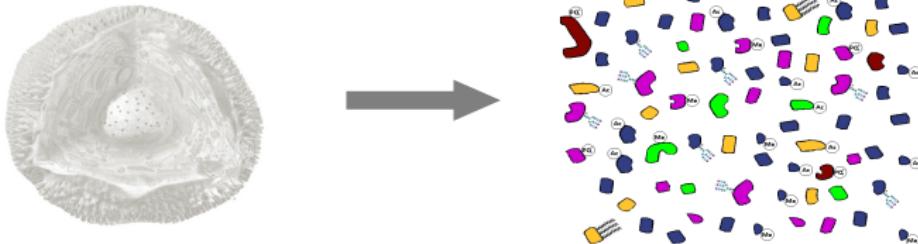


# Challenges in Label Free MS-based Quantitative proteomics



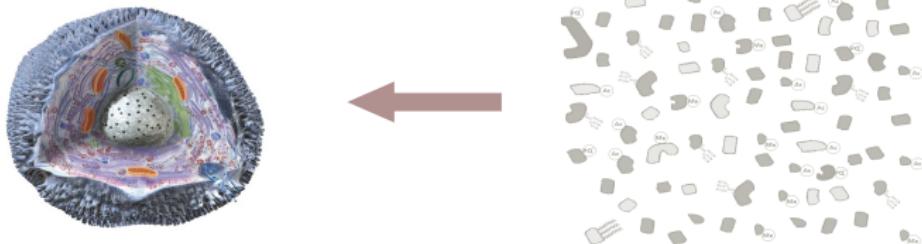
# Challenges in Label Free MS-based Quantitative proteomics

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

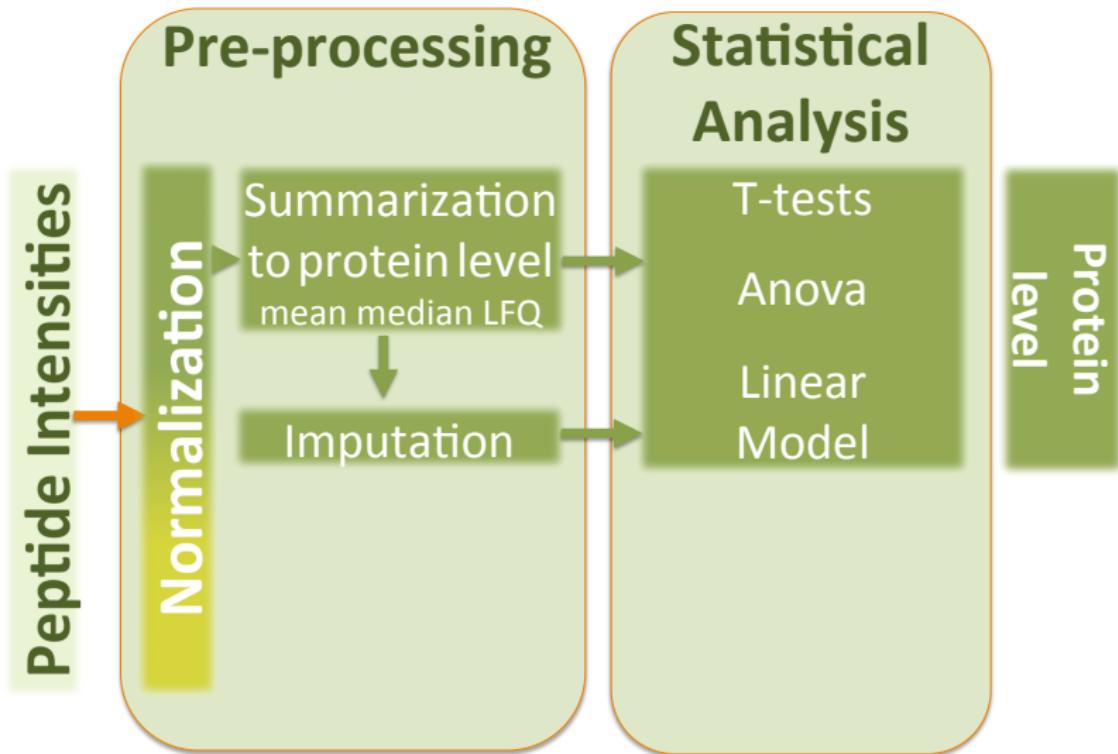


# Challenges in Label Free MS-based Quantitative proteomics

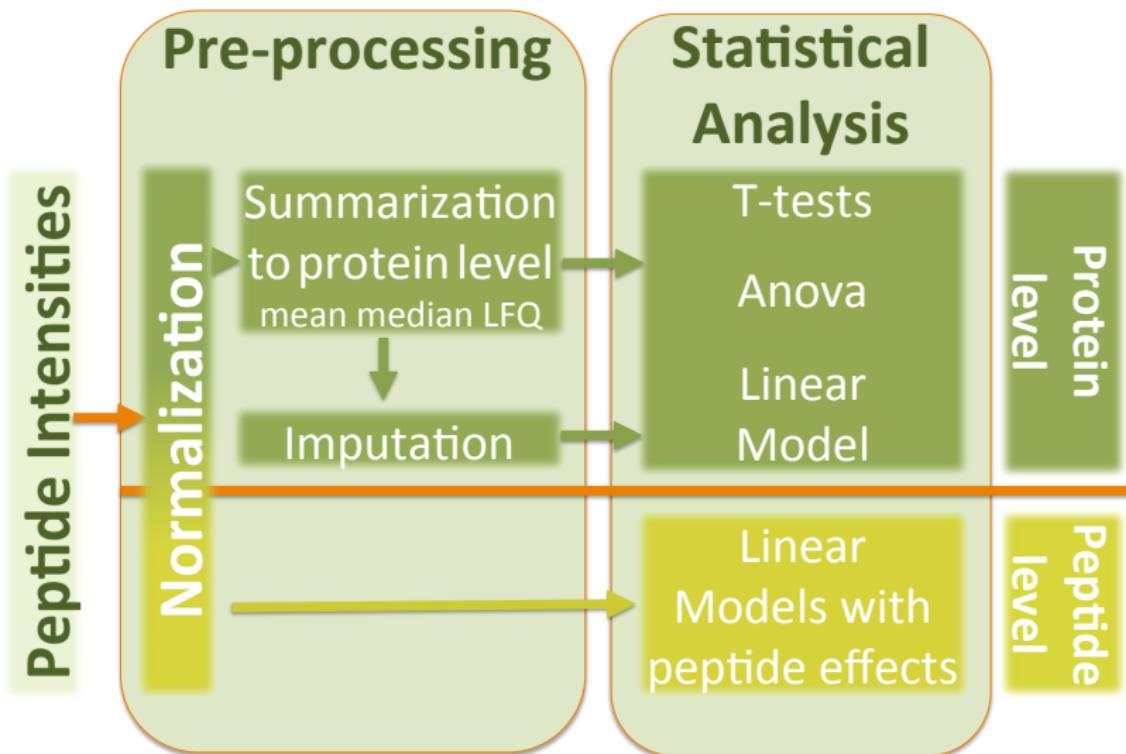
We need information on protein level!



# Label-free Quantitative Proteomics Data Analysis Pipelines



# Label-free Quantitative Proteomics Data Analysis Pipelines



# CPTAC Spike-in Study

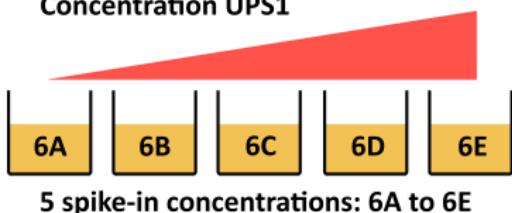
Digested  
UPS1 protein mix



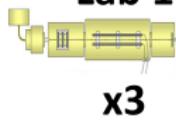
Digested  
yeast proteins



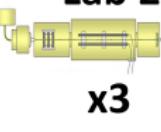
Concentration UPS1



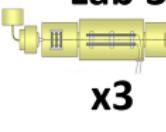
Lab 1



Lab 2



Lab 3



- 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

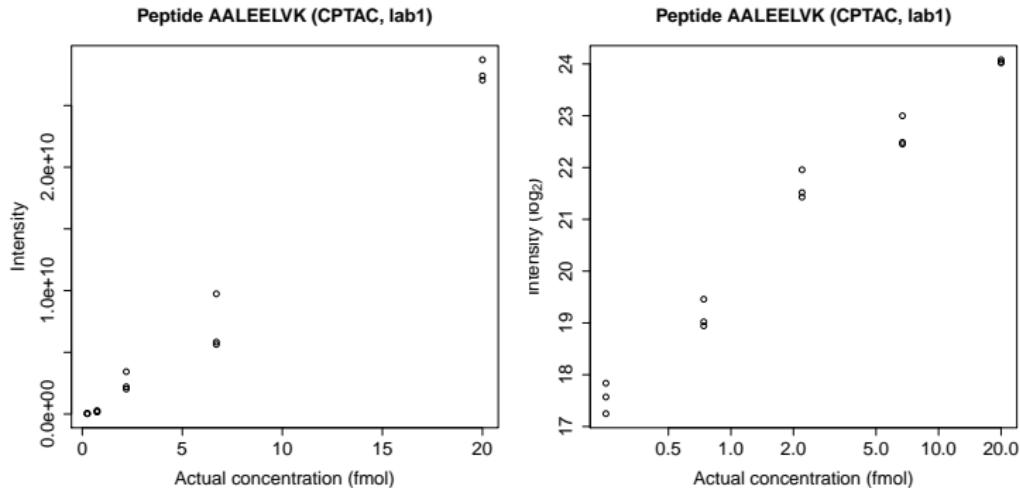
# Preprocessing

- Typical preprocessing steps
  - 1 Filtering
  - 2 Log-transformation
  - 3 Normalization
  - 4 (Summarization)
- Many methods exist

# Filtering

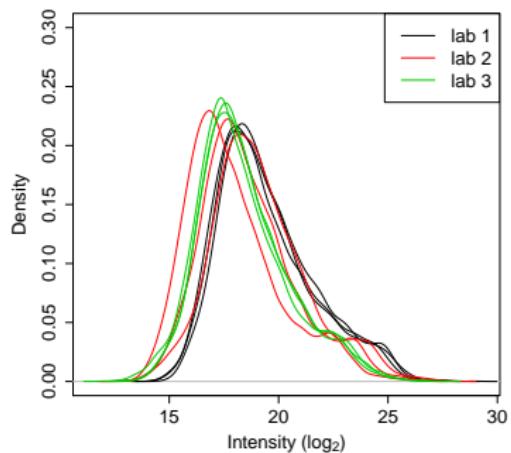
- Reverse sequences
- Only identified by modification site (only modified peptides detected)
- Razor peptides: non-unique peptides assigned to the protein group with the most other peptides
- Contaminants
- Peptides few identifications
- Proteins that are only identified with one or a few peptides
- Filtering does not induce bias if the criterion is independent from the downstream data analysis!

# Log-transformation

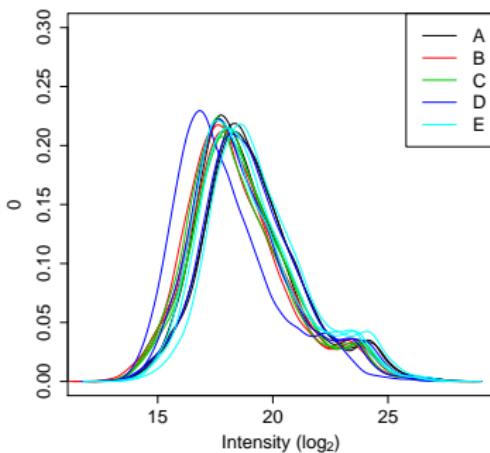


Variability more equal upon log transformation: often multiplicative error structure of intensity-based read-outs

Raw peptide intensity (CPTAC D)



Raw peptide intensity (CPTAC lab2)



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?

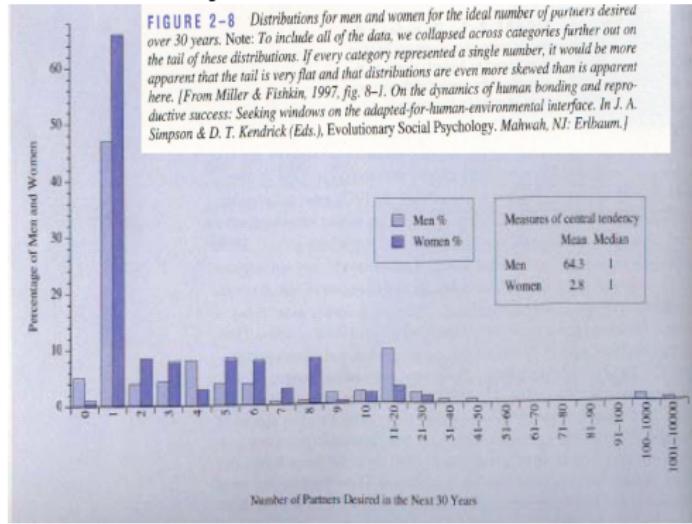
- Over a period of 30 years males desire to have on average 64.3 partners and females 2.8. (Miller and Fishkin, 1997)

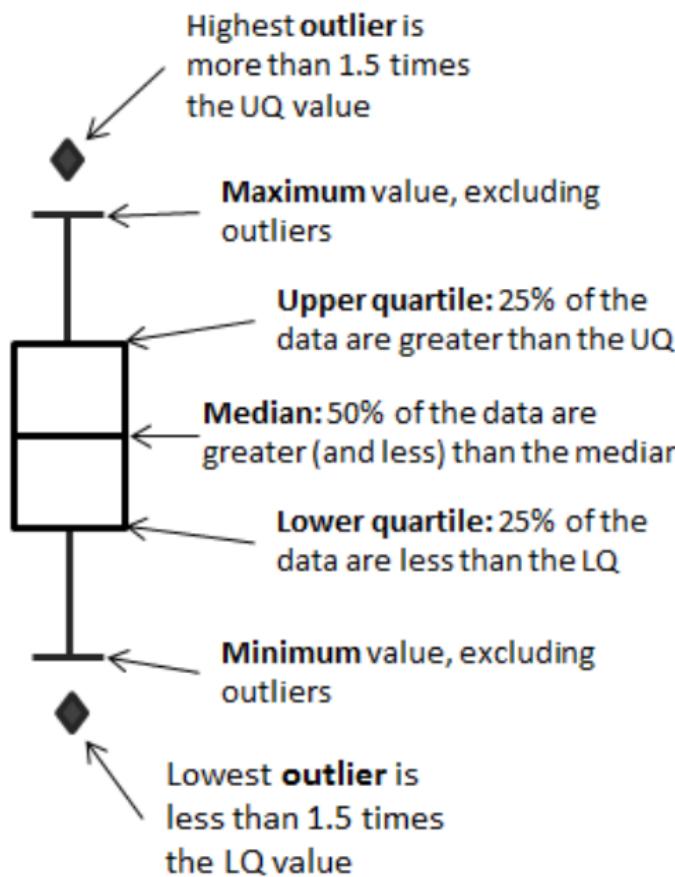
# Mean or median?

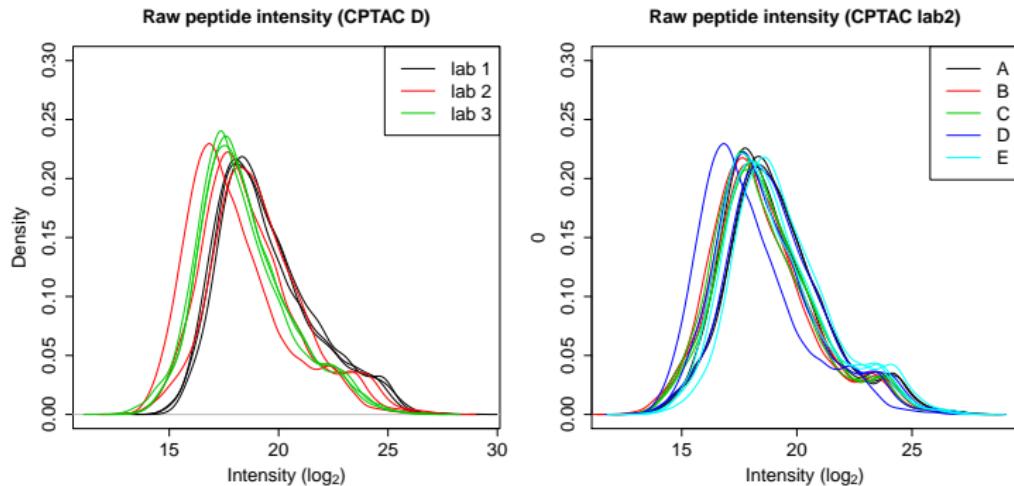
- Over a period of 30 years males desire to have on average 64.3 partners and females 2.8. (Miller and Fishkin, 1997)
- Over a period of 30 years males, is the median of the number of desired partners is 1 for both males and females. (Miller and Fishkin, 1997)

# Mean or median?

Mean is very sensitive to outliers!



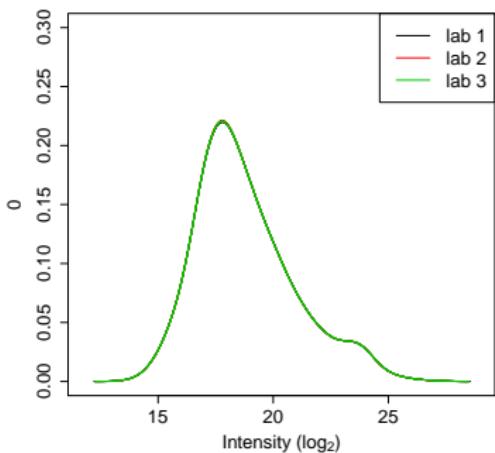




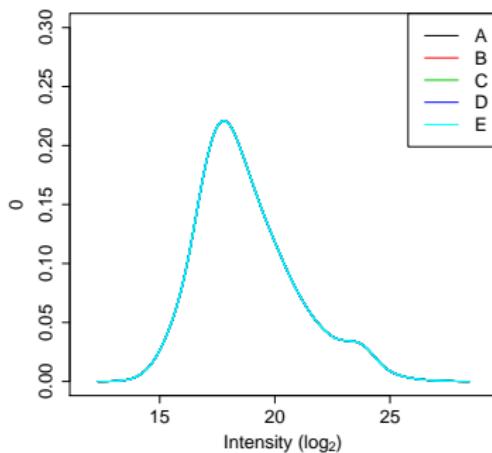
Even in very clean synthetic dataset (same background, only 48 UPS proteins can be different) the marginal peptide intensity distribution across samples can be quite distinct

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

QQ-normalized peptide intensity (CPTAC D)



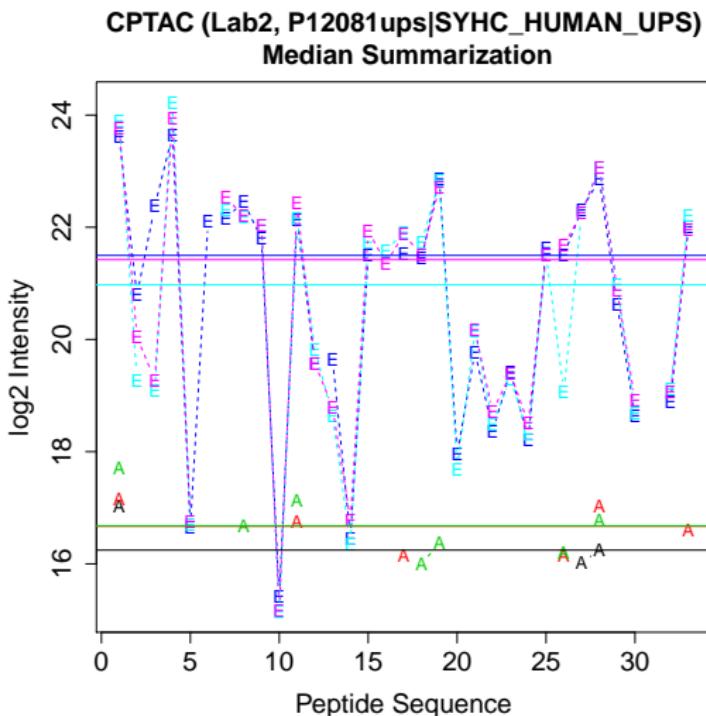
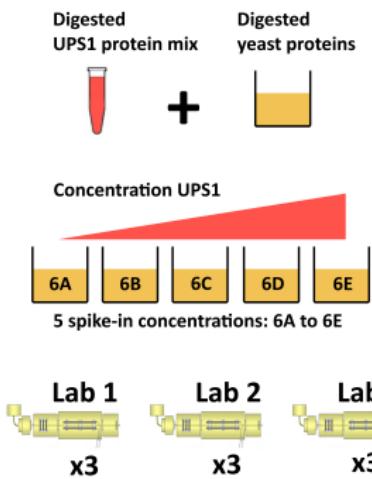
QQ-normalized peptide intensity (CPTAC lab2)



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

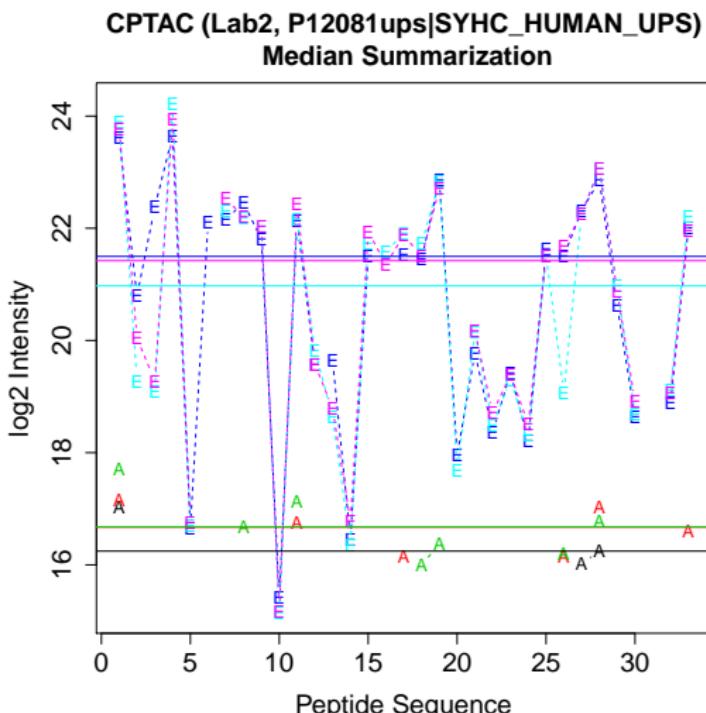
- Considerable effects between and within labs for replicate samples
  - Considerable effects between samples with different spike-in concentration
- Normalization is needed, e.g. **quantile normalization**

# 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:

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

# MaxLFQ summarization

**a**

```
>P63208
MPSIKLQSSDGEIFEVDVEIAKQSTIKTMLEDLGMDDEGDD
DPVPLPNVNNAILKKVIQWCTHKKDDPPPFEDDENKEKRTDD
IPVWDQEFLKVDQGTLFELIAANYLDIKGLLDVTCKTVANM
IKGKTPEEIRKTFNIKNDFTEEEAQVRKENQWCEEK
```

**b**

Peptide species	Sequence	Charge	Mod.
P <sub>1</sub>	<b>LQSSDGEIFEVDV</b> EIAK	2	–
P <sub>2</sub>	<b>LQSSDGEIFEVDV</b> EIAK	3	–
P <sub>3</sub>	<b>RTDDIPVWDQEF</b> FLK	2	–
P <sub>4</sub>	<b>T</b> <b>V</b> <b>A</b> <b>N</b> <b>M</b> <b>I</b> <b>K</b>	2	–
P <sub>5</sub>	<b>T</b> <b>V</b> <b>A</b> <b>N</b> <b>M</b> <b>I</b> <b>K</b>	2	Oxid.
P <sub>6</sub>	<b>T</b> <b>P</b> <b>E</b> <b>E</b> <b>I</b> <b>R</b> <b>K</b>	3	–
P <sub>7</sub>	<b>NDFTEEEE</b> AQVR	2	–

**c**

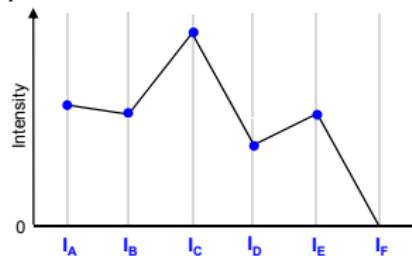
Sample	P <sub>1</sub>	P <sub>2</sub>	P <sub>3</sub>	P <sub>4</sub>	P <sub>5</sub>	P <sub>6</sub>	P <sub>7</sub>
A	+						+
B		+	+				+
C	+	+	+	+		+	+
D	+	+		+		+	+
E		+		+			+
F	+				+		

**d**

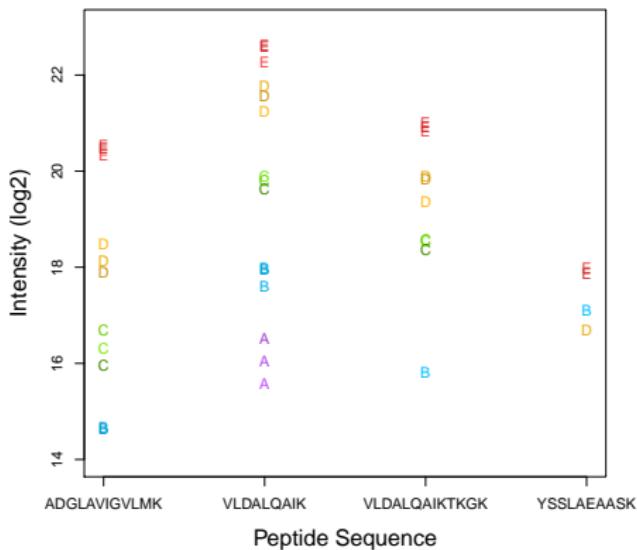
A	$r_{BA}$					
B	$r_{CA}$		$r_{CB}$			
C	$r_{DA}$	$r_{DB}$	$r_{DC}$			
D	$r_{EA}$	$r_{EB}$	$r_{EC}$	$r_{ED}$		
E	$r_{FA}$	$r_{FB}$	$r_{FC}$	$r_{FD}$	$r_{FE}$	
F						
	A	B	C	D	E	F

**e**

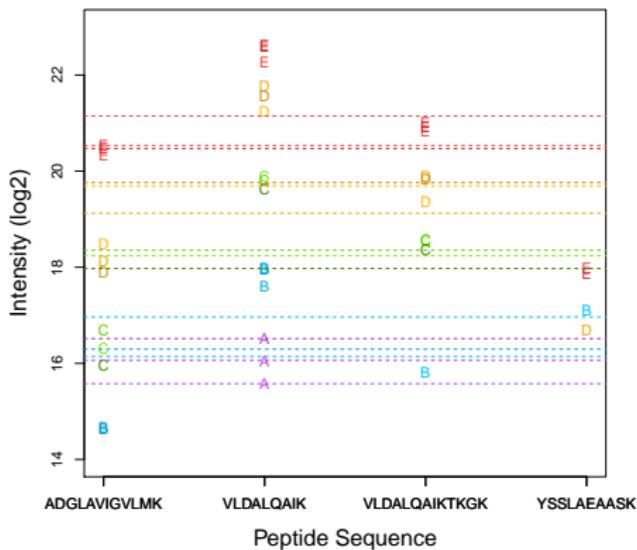
$$\begin{array}{lll} r_{BA} = I_B / I_A & r_{CA} = I_C / I_A & r_{CB} = I_C / I_B \\ r_{DA} = I_D / I_A & r_{DB} = I_D / I_B & r_{DC} = I_D / I_C \\ r_{EC} = I_E / I_C & r_{ED} = I_E / I_D & I_F = 0 \end{array}$$

**f**

# Summarisation with peptide based model



# Summarisation with peptide based model



Protein by protein analysis of peptide data with linear model

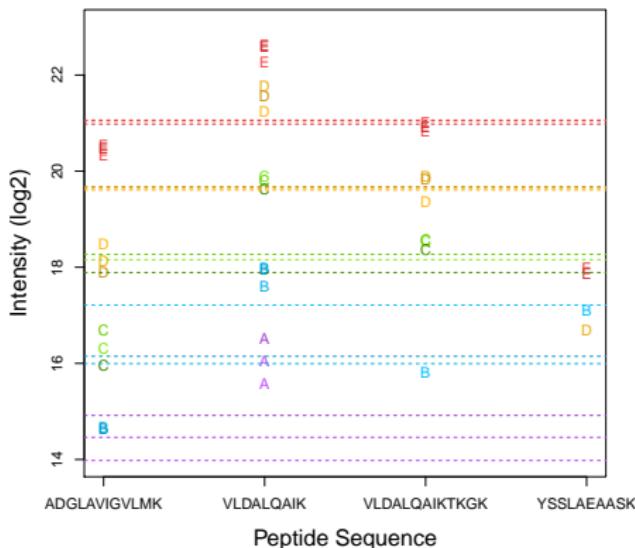
peptide level

$$y_{sp} = \epsilon_{sp}$$

protein level

$$\beta_s^{\text{sample}}$$

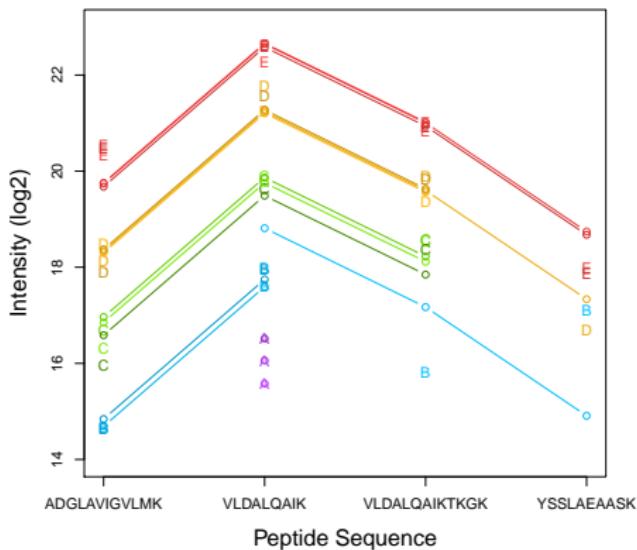
# Summarisation with peptide based model



Protein by protein analysis of peptide data with linear model

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

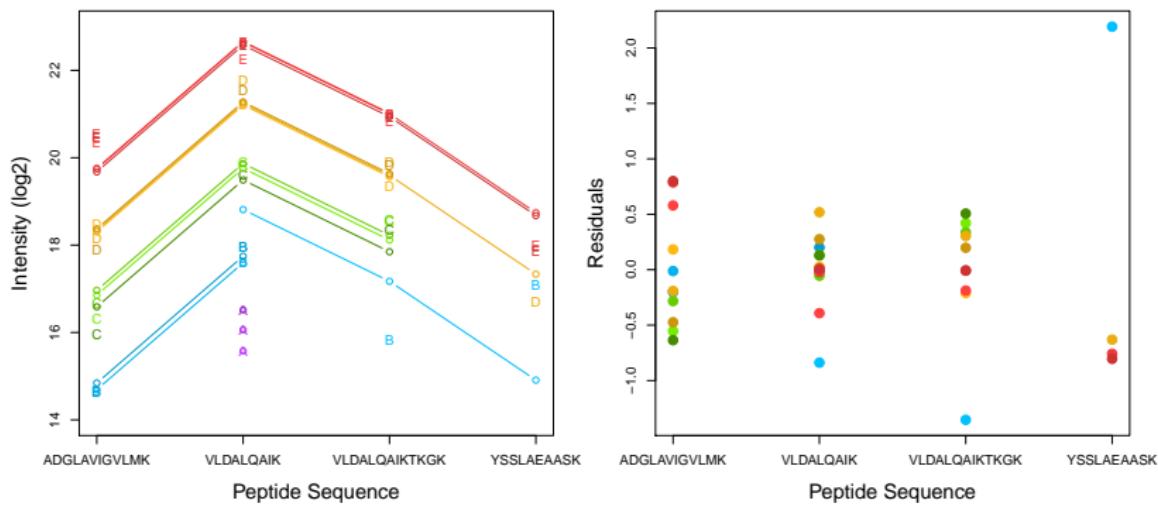
# Summarisation with peptide based model



Protein by protein analysis of peptide data with linear model

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

# Summarisation with peptide based model

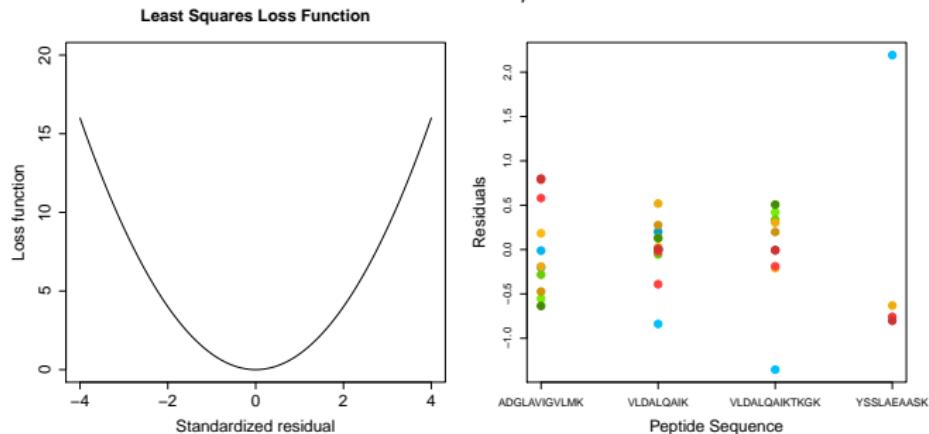


Protein by protein analysis of peptide data with linear model

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

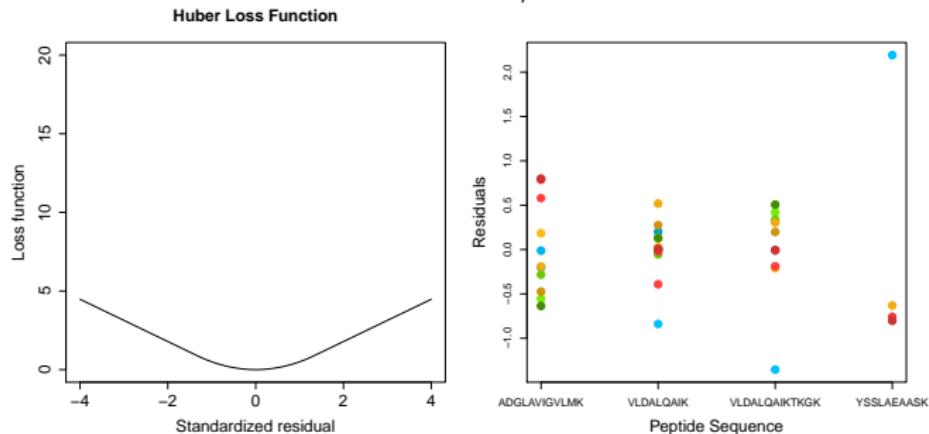
# Robust estimation using observation weights

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



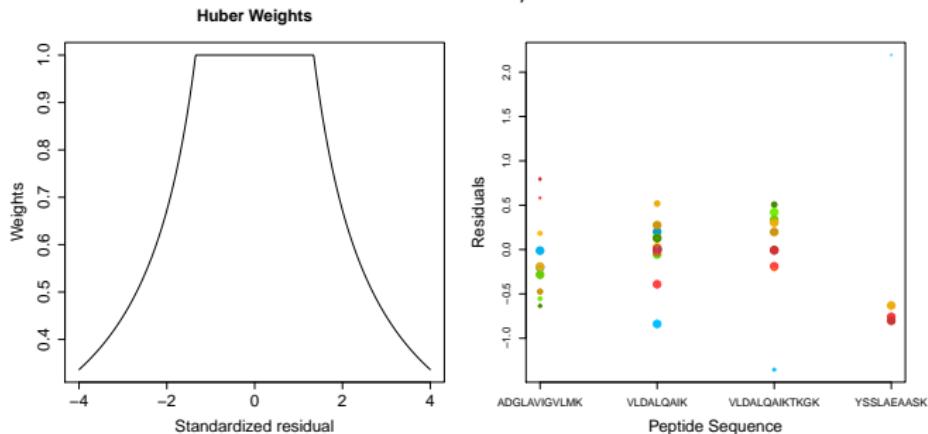
# Robust estimation using observation weights

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



# Robust estimation using observation weights

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

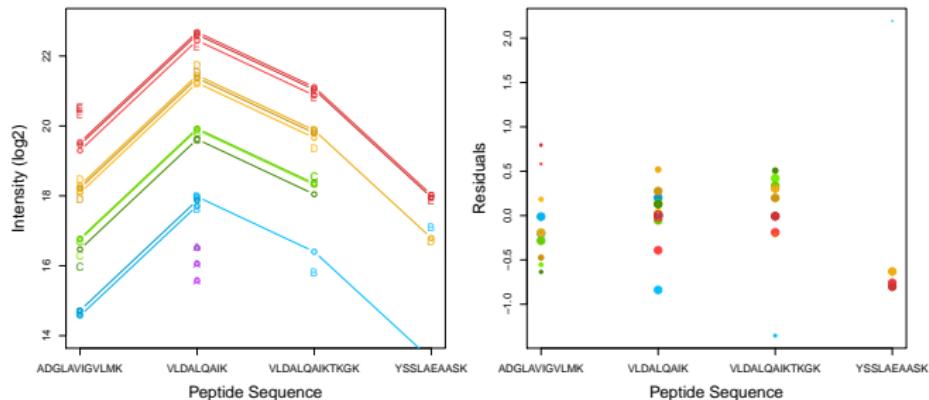


- Iteratively fit model with observation weights  $w(\epsilon_{ip})$

$$\operatorname{argmin}_{\beta_1^{\text{pep}}, \dots, \beta_P^{\text{pep}}, \beta_1^{\text{samp}}, \dots, \beta_n^{\text{samp}}} \left[ \sum_{i=1}^n \sum_{p=1}^P w(\epsilon_{ip}) (y_{ip} - \beta_p^{\text{pep}} - \beta_i^{\text{samp}})^2 \right]$$

# Robust estimation using observation weights

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

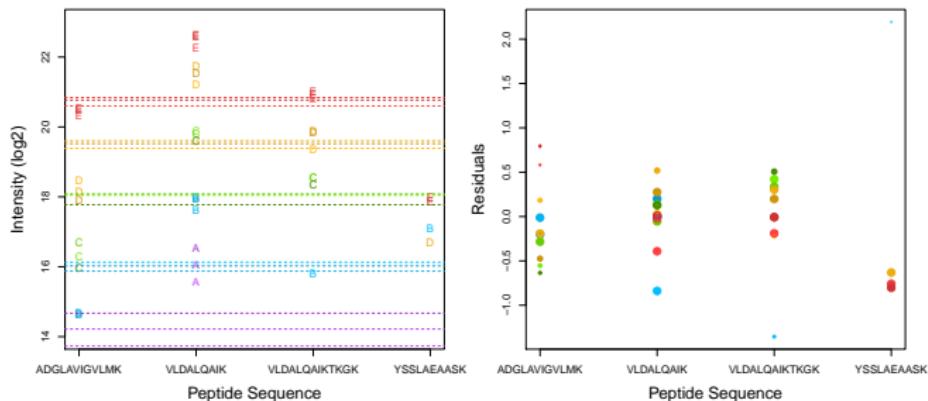


- Iteratively fit model with observation weights  $w(\epsilon_{ip})$

$$\operatorname{argmin}_{\beta_1^{\text{pep}}, \dots, \beta_P^{\text{pep}}, \beta_1^{\text{samp}}, \dots, \beta_n^{\text{samp}}} \left[ \sum_{i=1}^n \sum_{p=1}^P w(\epsilon_{ip}) (y_{ip} - \beta_p^{\text{pep}} - \beta_i^{\text{samp}})^2 \right]$$

# Robust estimation using observation weights

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



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

$$\operatorname{argmin}_{\beta_1^{\text{pep}}, \dots, \beta_P^{\text{pep}}, \beta_1^{\text{samp}}, \dots, \beta_n^{\text{samp}}} \left[ \sum_{i=1}^n \sum_{p=1}^P w(\epsilon_{ip}) (y_{ip} - \beta_p^{\text{pep}} - \beta_i^{\text{samp}})^2 \right]$$