

Part I: Normalization & Summarization

Lieven Clement

Proteomics Data Analysis Shortcourse

Outline

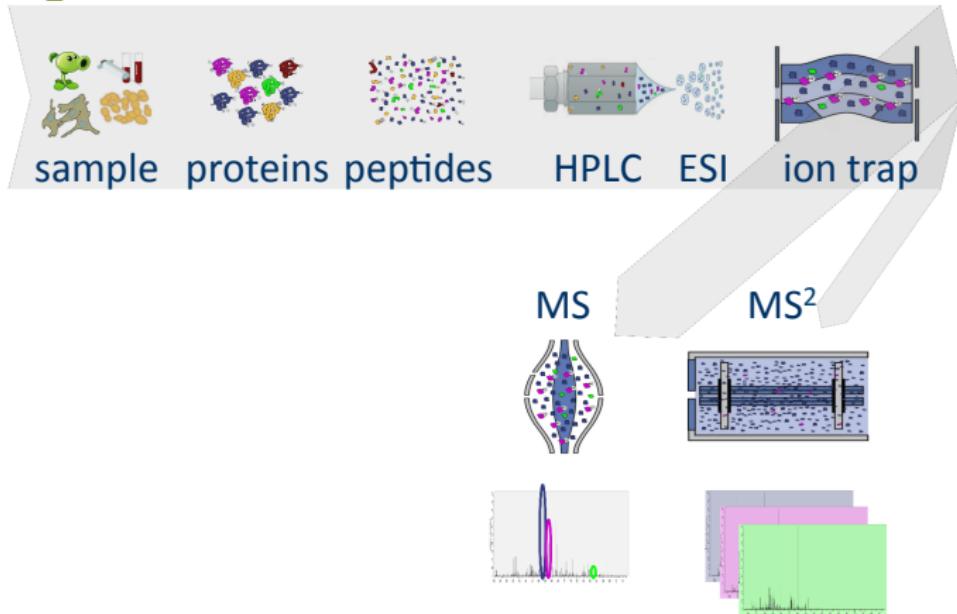
① Introduction

- ① Label free MS based Quantitative Proteomics Workflow and Challenges

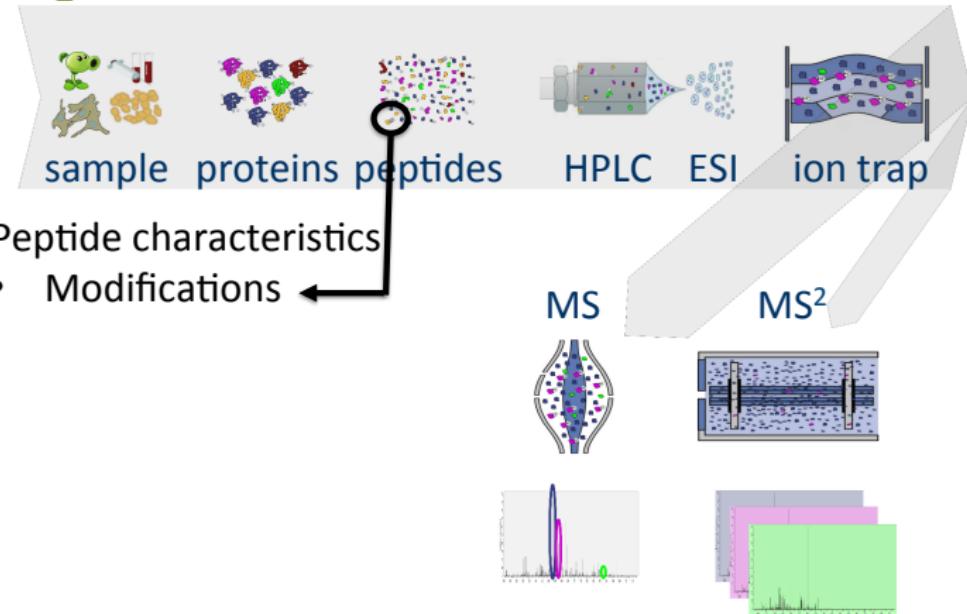
② Preprocessing

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

Challenges in Label Free Quantitative Proteomics

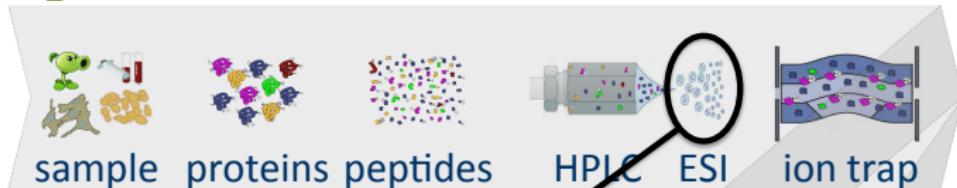


Challenges in Label Free Quantitative Proteomics



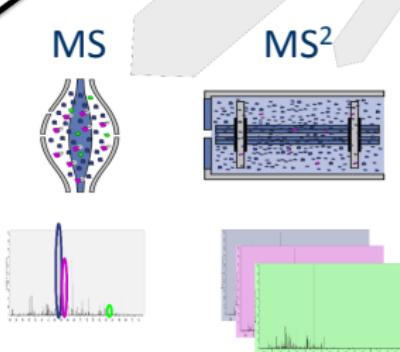
Quantification Identification

Challenges in Label Free Quantitative Proteomics



Peptide characteristics

- Modifications
- Ionisation efficiency
 - Outliers
 - Huge variability



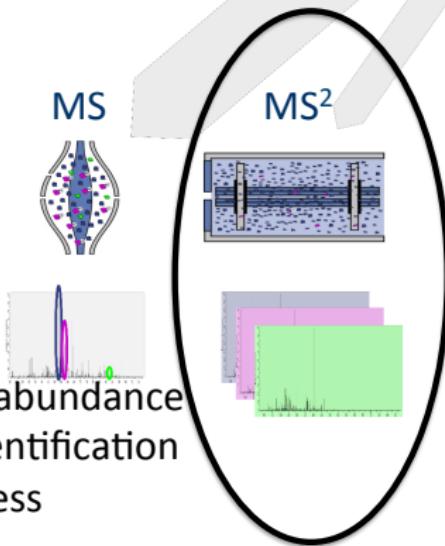
Quantification Identification

Challenges in Label Free Quantitative Proteomics



Peptide characteristics

- Modifications
- Ionisation efficiency
 - Outliers
 - Huge variability
- MS² selection on peptide abundance
 - Context dependent Identification
 - Non-random missingness

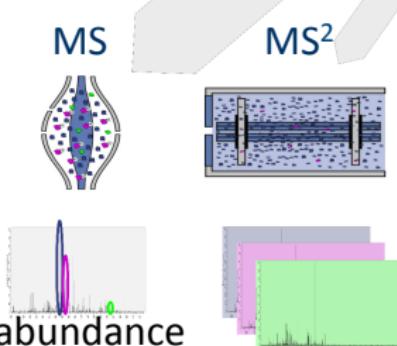


Challenges in Label Free Quantitative Proteomics



Peptide characteristics

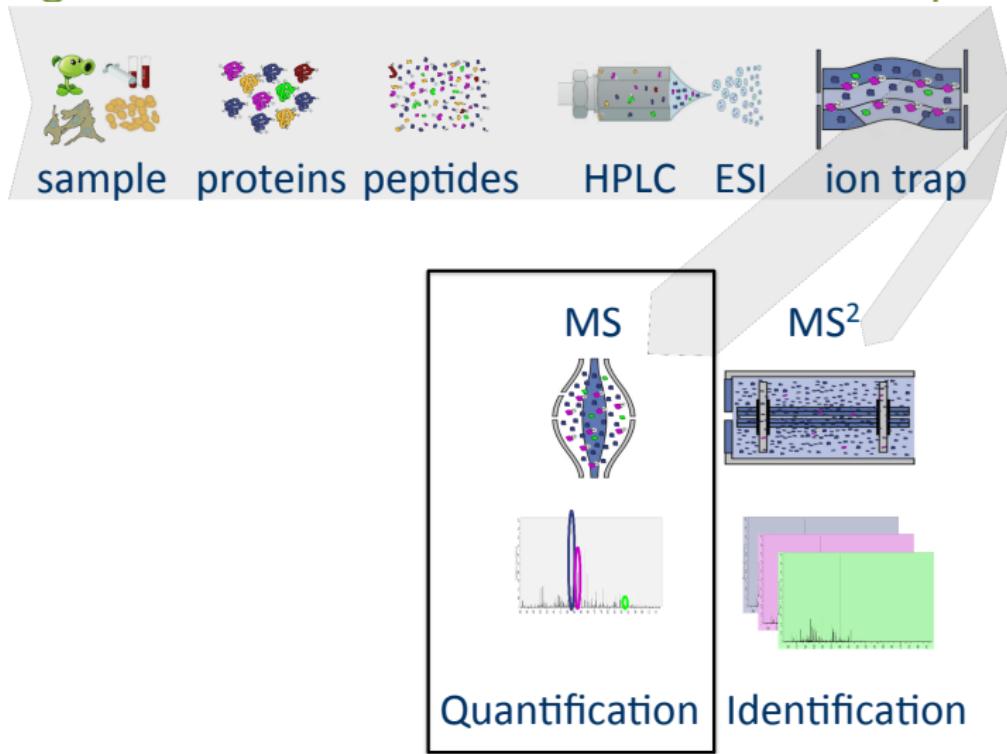
- Modifications
- Ionisation efficiency
 - Outliers
 - Huge variability
- MS² selection on peptide abundance
 - Context dependent Identification
 - Non-random missingness



Unbalanced peptides identifications across samples and messy data

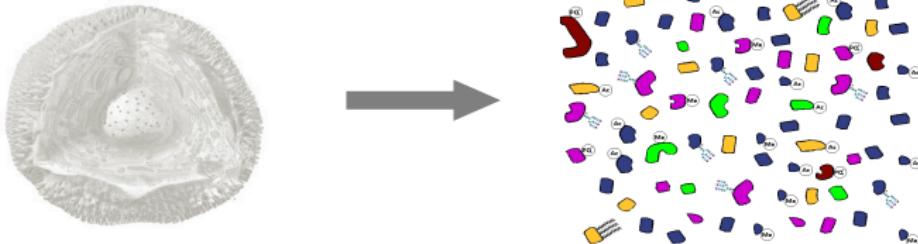


Challenges in Label Free MS-based Quantitative proteomics



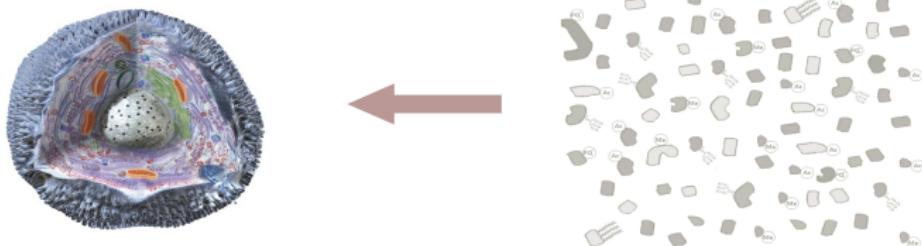
Challenges in Label Free MS-based Quantitative proteomics

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

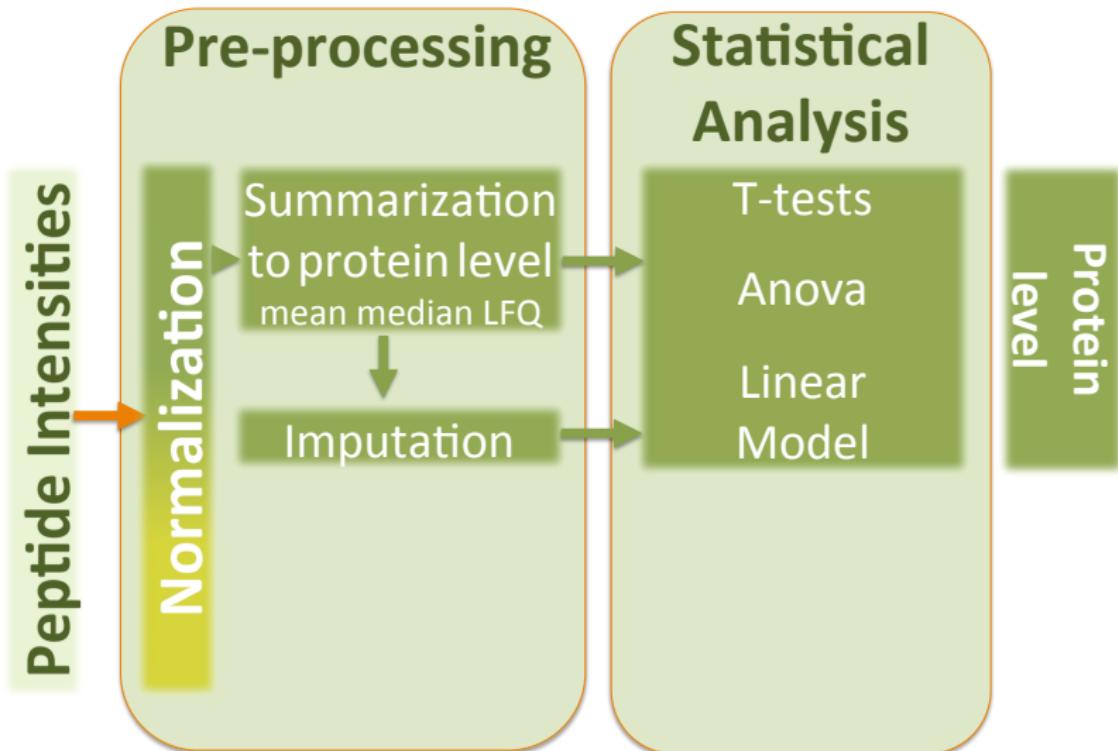


Challenges in Label Free MS-based Quantitative proteomics

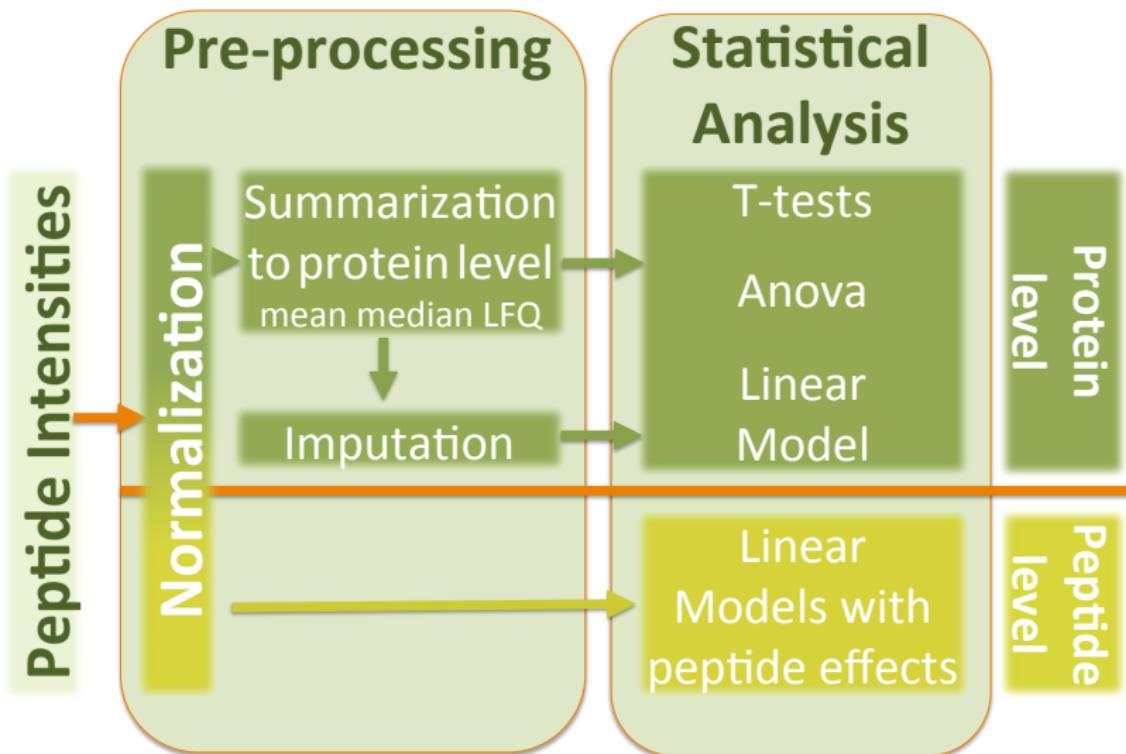
We need information on protein level!



Label-free Quantitative Proteomics Data Analysis Pipelines



Label-free Quantitative Proteomics Data Analysis Pipelines



CPTAC Spike-in Study

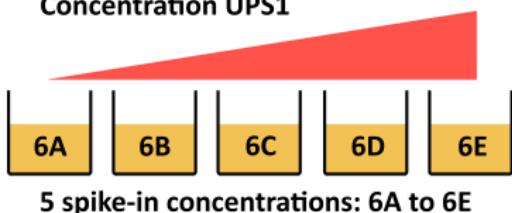
Digested
UPS1 protein mix



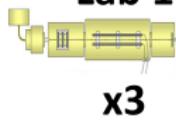
Digested
yeast proteins



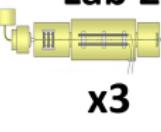
Concentration UPS1



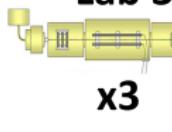
Lab 1



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

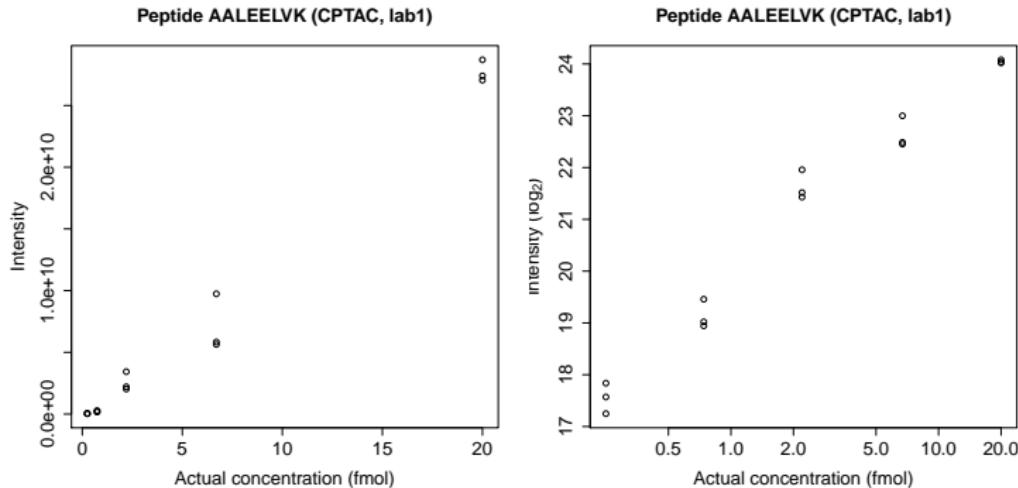
Preprocessing

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

Filtering

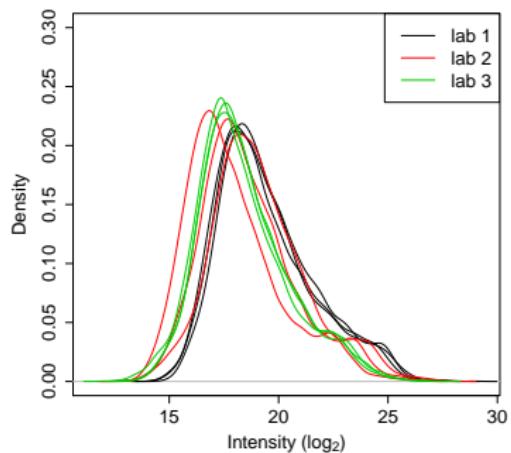
- Reverse sequences
- Only identified by modification site (only modified peptides detected)
- Razor peptides: non-unique peptides assigned to the protein group with the most other peptides
- Contaminants
- Peptides few identifications
- Proteins that are only identified with one or a few peptides
-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

Log-transformation

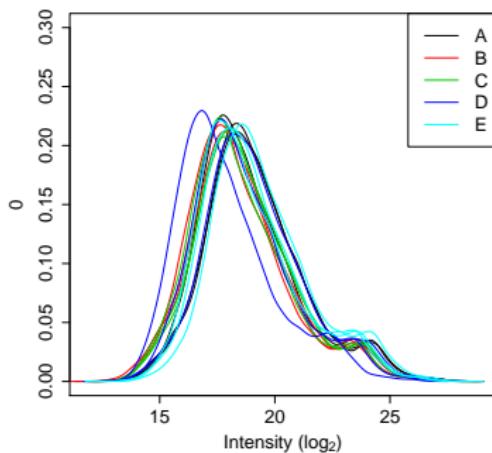


Variability more equal upon log transformation: often multiplicative error structure of intensity-based read-outs

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?

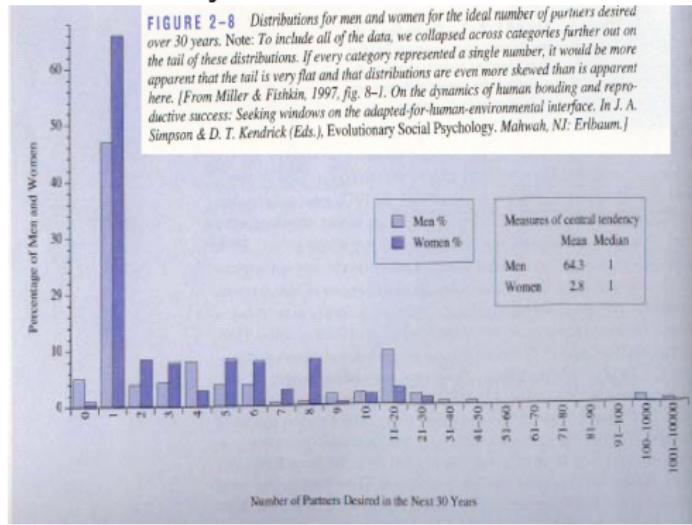
- Over a period of 30 years males desire to have on average 64.3 partners and females 2.8. (Miller and Fishkin, 1997)

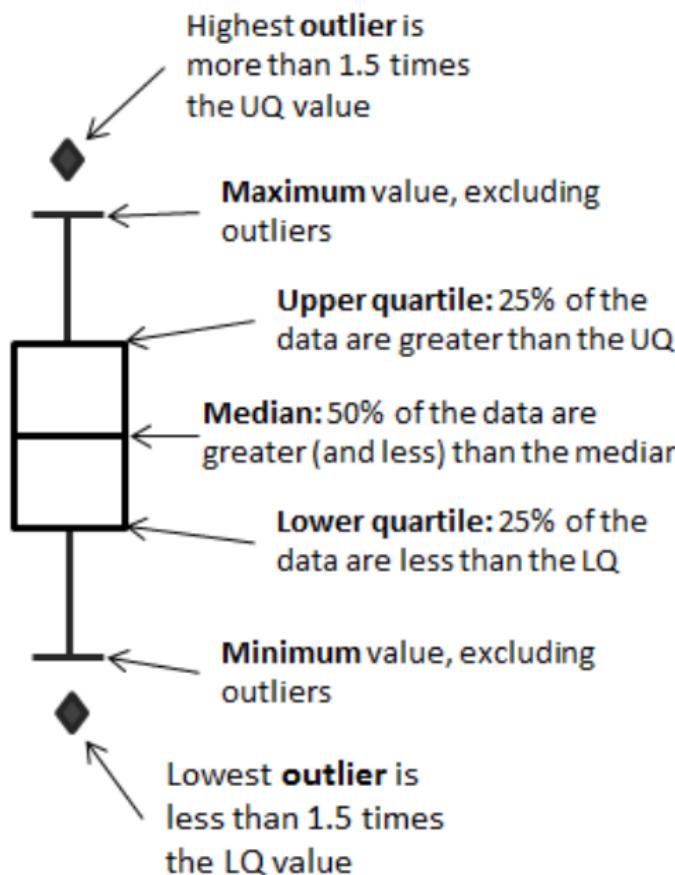
Mean or median?

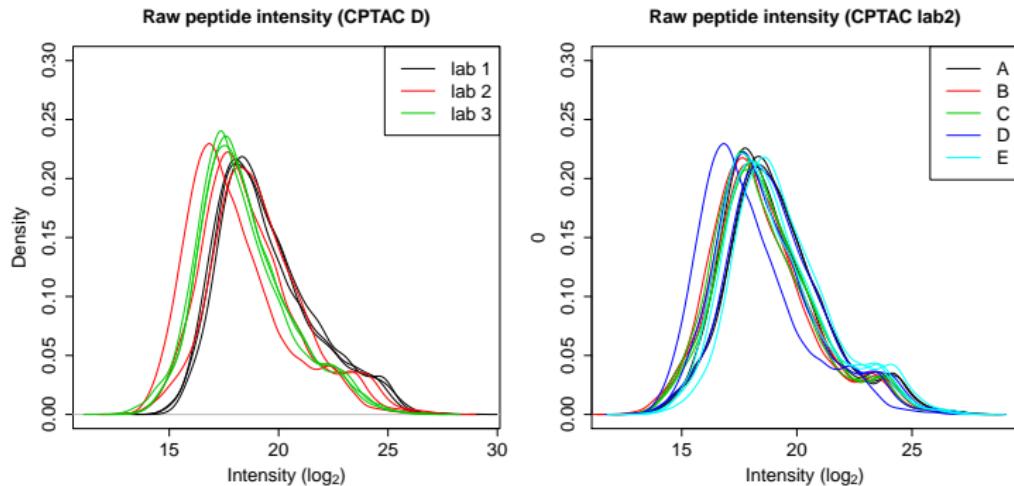
- Over a period of 30 years males desire to have on average 64.3 partners and females 2.8. (Miller and Fishkin, 1997)
- Over a period of 30 years males, is the median of the number of desired partners is 1 for both males and females. (Miller and Fishkin, 1997)

Mean or median?

Mean is very sensitive to outliers!





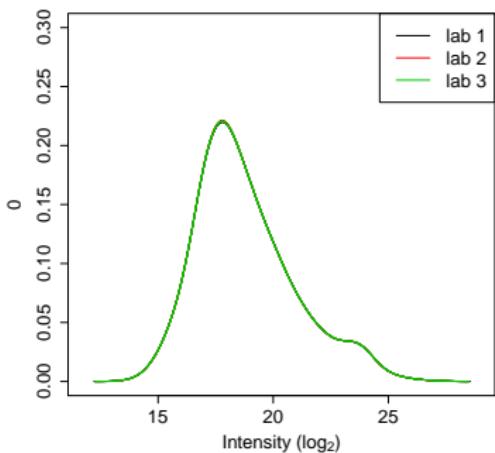


Even in very clean synthetic dataset (same background, only 48 UPS proteins can be different) the marginal peptide intensity distribution across samples can be quite distinct

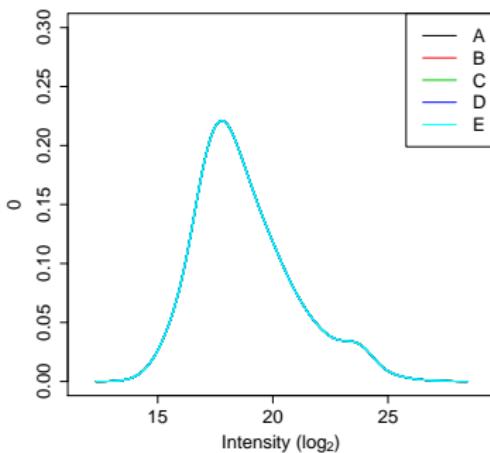
- Considerable effects between and within labs for replicate samples
- Considerable effects between samples with different spike-in concentration
- Normalization is needed



QQ-normalized peptide intensity (CPTAC D)



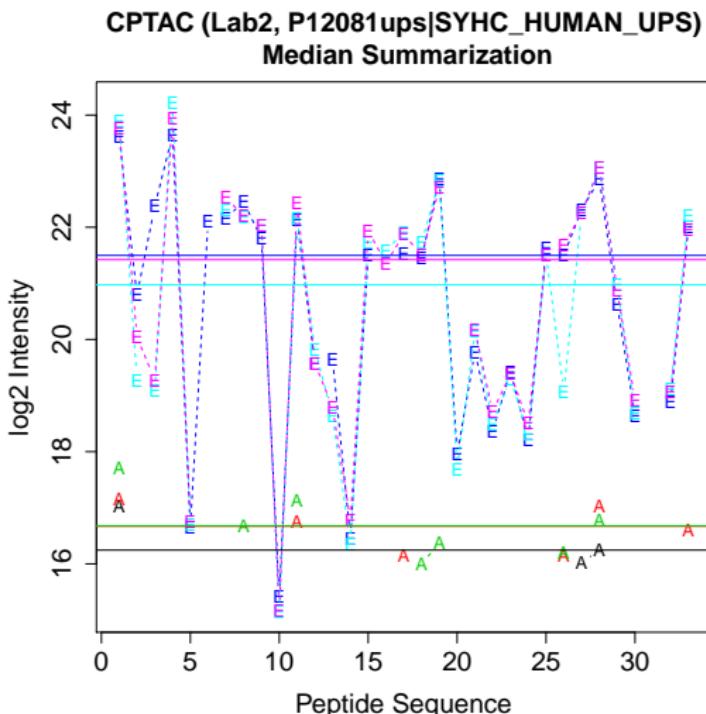
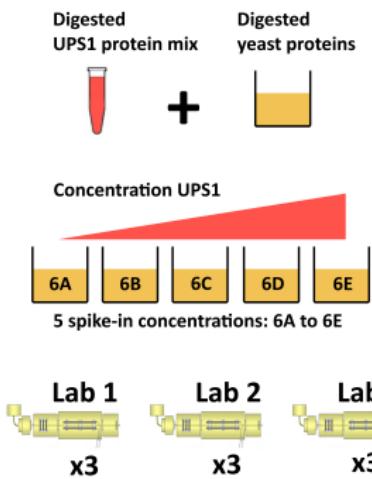
QQ-normalized peptide intensity (CPTAC lab2)



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

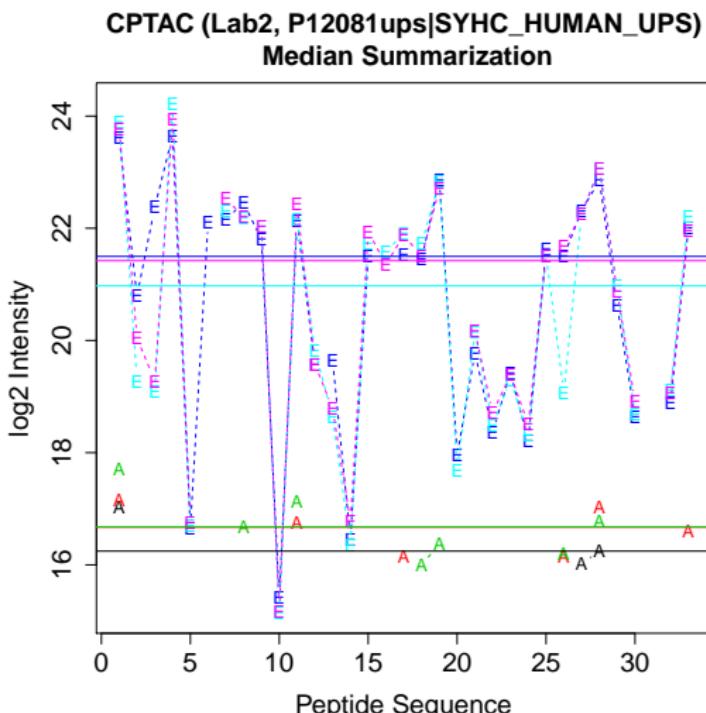
- Considerable effects between and within labs for replicate samples
 - Considerable effects between samples with different spike-in concentration
- Normalization is needed, e.g. **quantile normalization**

Summarization



Summarization

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



MaxLFQ summarization

a

```
>P63208
MPSIKLQSSDGEIFEVDVEIAKQSTIKTMLEDLGMDDEGDD
DPVPLPNVNNAILKKVIQWCTHKKDDPPPFEDDENKEKRTDD
IPVWDQEFFLKVDQGTLFELILAANYLDIKGLLDVTCKTVANM
IKGKTFEEEIRKTFNIKNDFTEEEAQVRKENQWCEEK
```

b

Peptide species	Sequence	Charge	Mod.
P ₁	LQSSDGEI FEVDVEIAK	2	–
P ₂	LQSSDGEI FEVDVEIAK	3	–
P ₃	RTDDI PVWDQEF <u>FLK</u>	2	–
P ₄	T <u>V</u> A N M I K	2	–
P ₅	T <u>V</u> A N M I K	2	Oxid.
P ₆	T <u>P</u> E E I R K	3	–
P ₇	N <u>D</u> F <u>T</u> E E E <u>A</u> Q V R	2	–

c

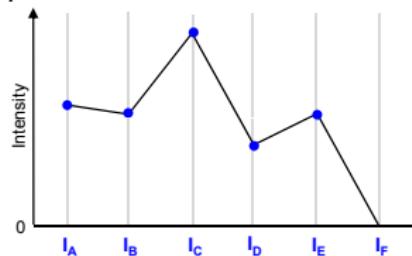
Sample	P ₁	P ₂	P ₃	P ₄	P ₅	P ₆	P ₇
A	+						+
B		+	+				+
C	+	+	+	+		+	+
D	+	+		+		+	+
E		+		+			+
F	+				+		

d

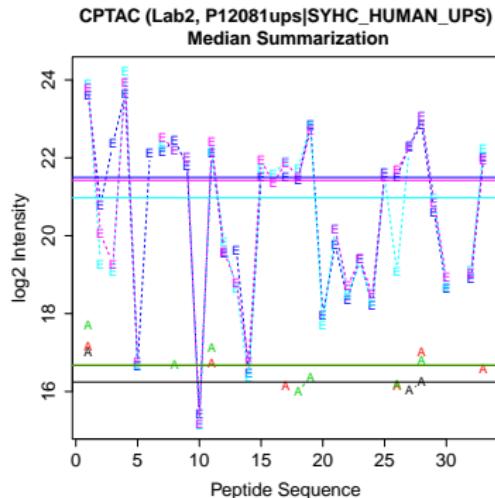
A	r_{BA}					
B	r_{CA}		r_{CB}			
C	r_{DA}	r_{DB}	r_{DC}			
E	r_{EA}	r_{EB}	r_{EC}	r_{ED}		
F	r_{FA}	r_{FB}	r_{FC}	r_{FD}	r_{FE}	
	A	B	C	D	E	F

e

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

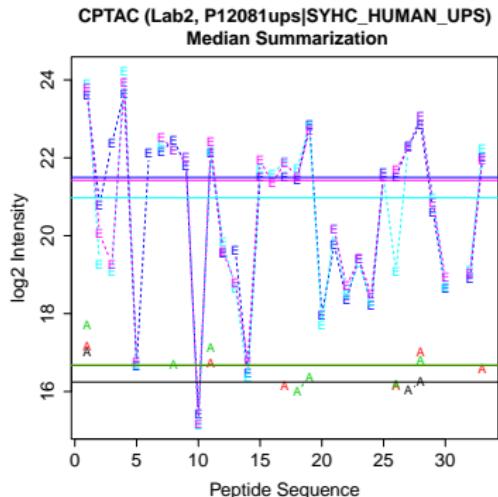
f

Peptide based model



- ① y: log₂ transformed peptide intensities

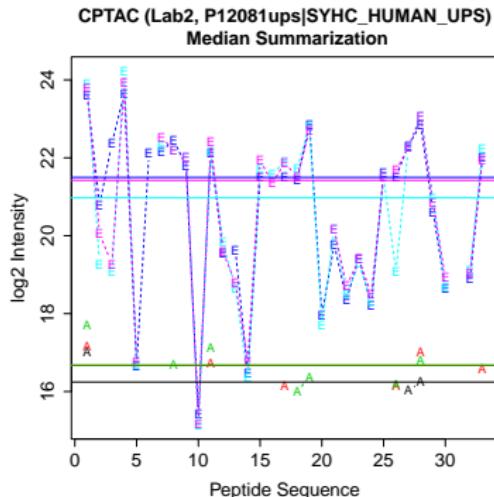
Peptide based model



- ① y: log₂ transformed peptide intensities
- ② Protein by protein analysis of peptide level data with linear model



Peptide based model



- ① y : log₂ transformed peptide intensities
- ② Protein by protein analysis of peptide level data with linear
model peptide level protein level
 $y_{pept} \sim peptide + sample$

