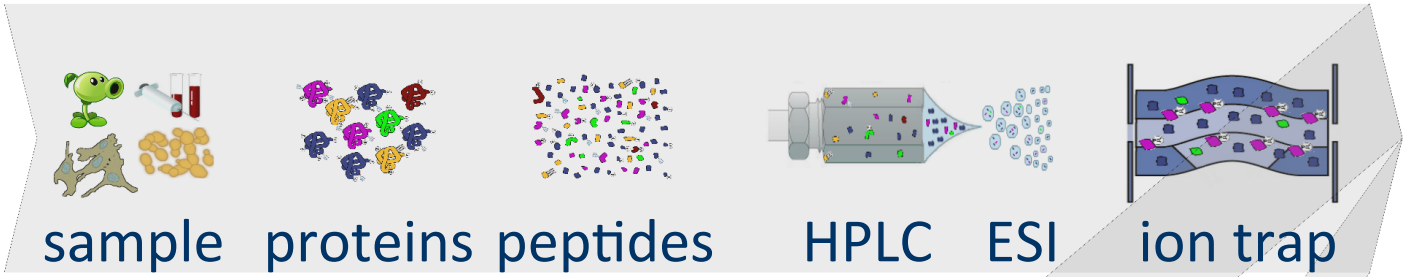
Part I: Normalization & Summarization

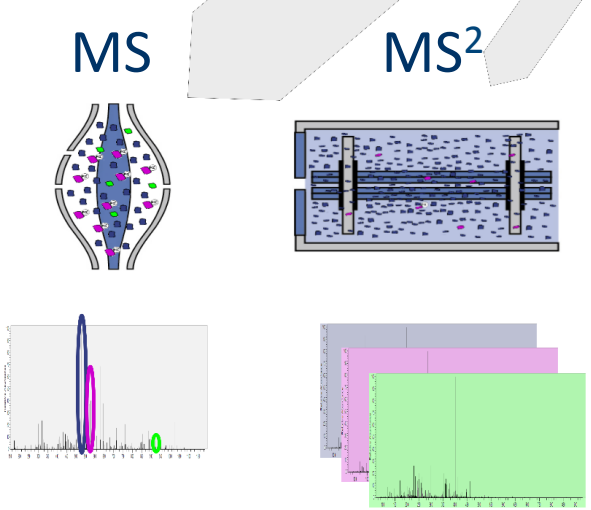
Lieven Clement

Proteomics Data Analysis 2018, Gulbenkian Institute, May 28 -June 1 2018.

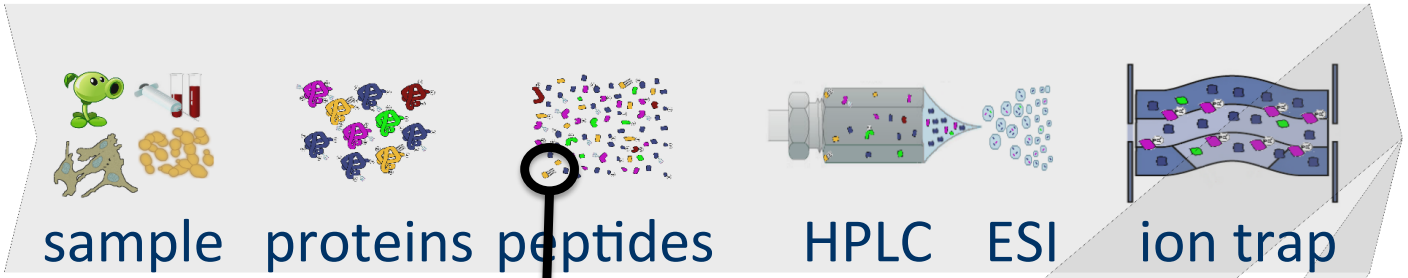
Outline

Challenges in Label Free Quantitative Proteomics



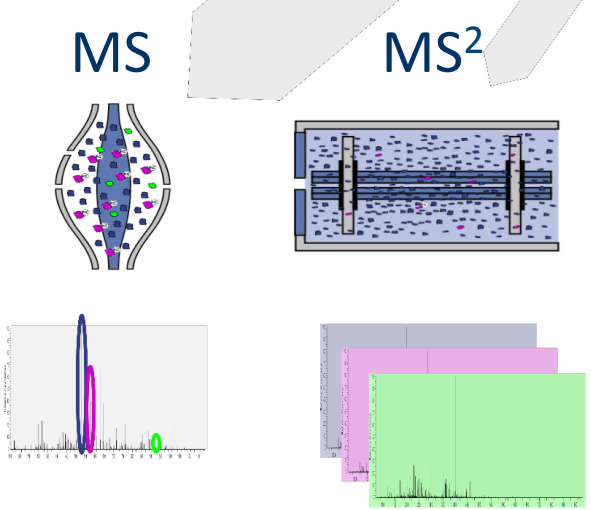


Challenges in Label Free Quantitative Proteomics

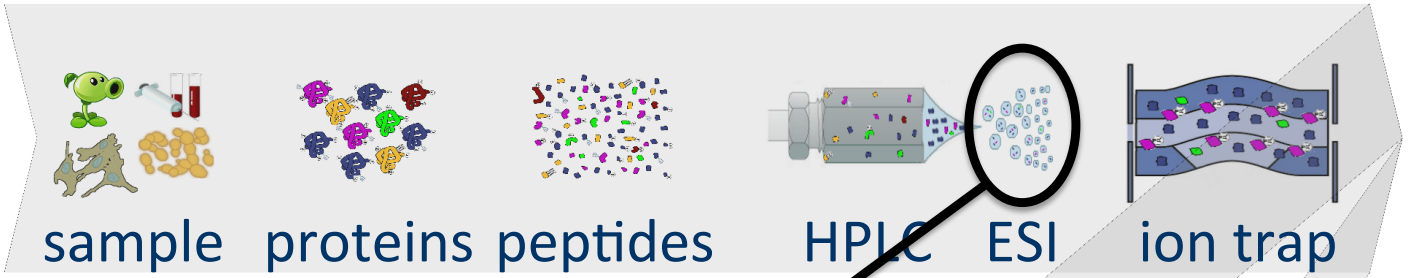


Peptide characteristics

- Modifications

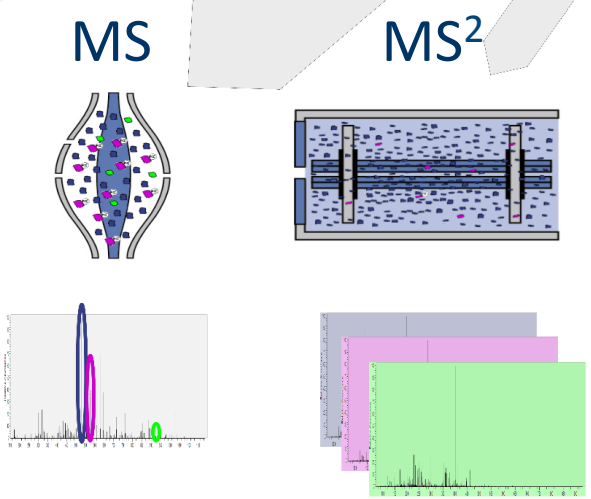


Challenges in Label Free Quantitative Proteomics

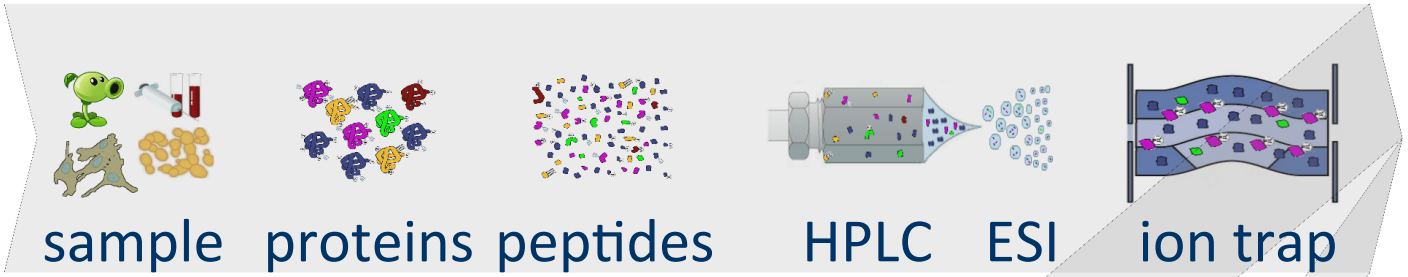


Peptide characteristics

- Modifications
- Ionisation efficiency
  - Outliers
  - Huge variability

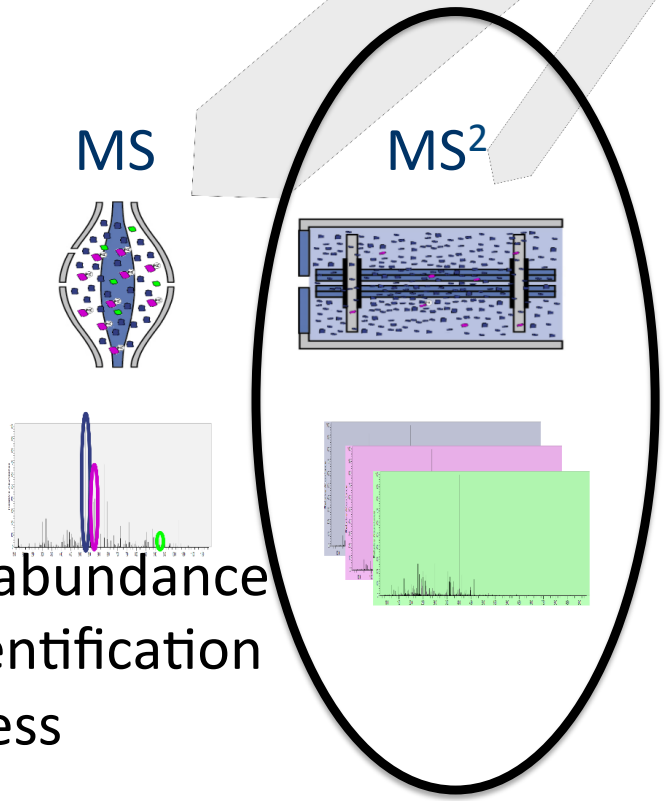


Challenges in Label Free Quantitative Proteomics

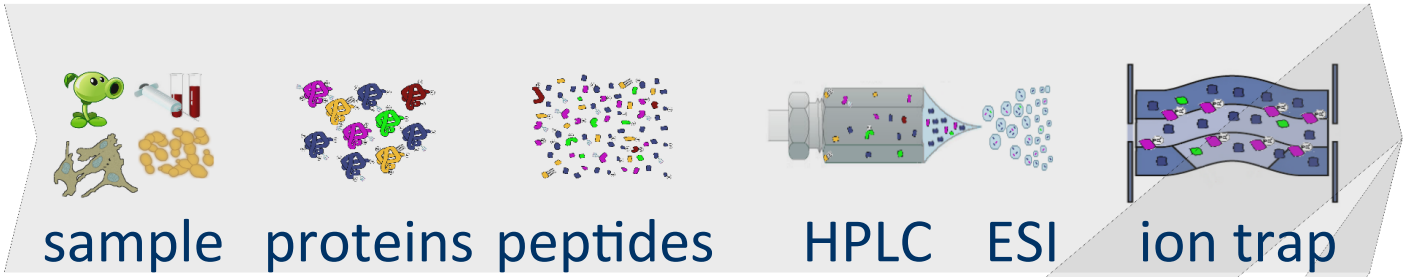


Peptide characteristics

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

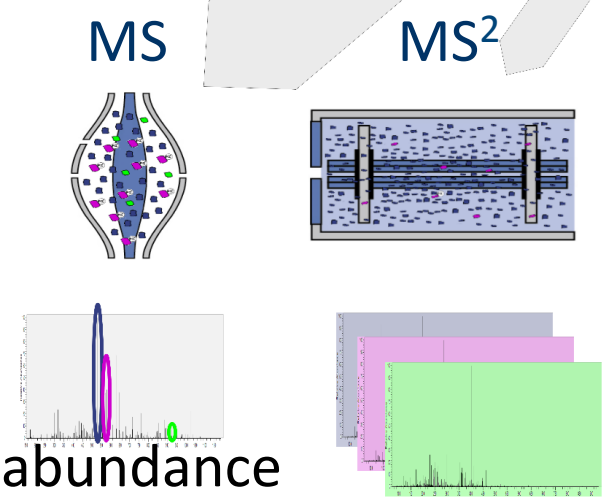


Challenges in Label Free Quantitative Proteomics



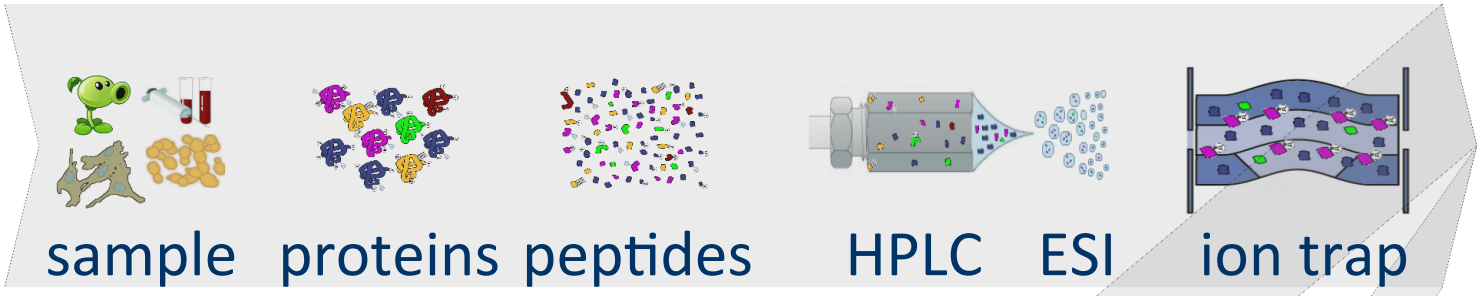
Peptide characteristics

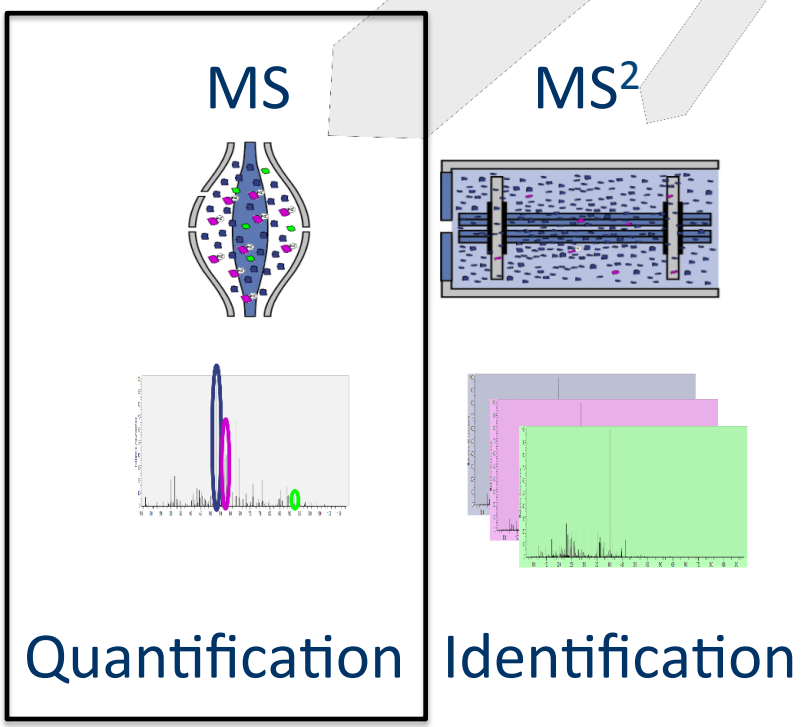
- Modifications
- Ionisation efficiency
  - Outliers
  - Huge variability
- MS^2 selection on peptide abundance
  - Context dependent Identification
  - Non-random missingness



Unbalanced peptides identifications across samples and messy data

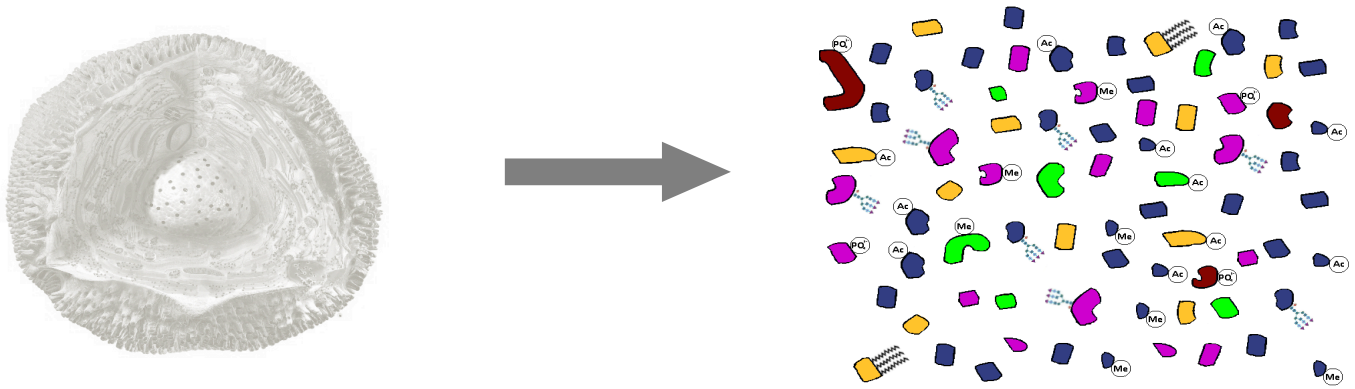
Challenges in Label Free MS-based Quatitative proteomics





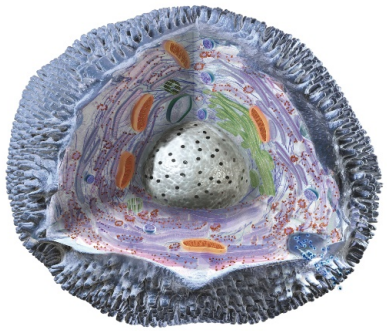
Challenges in Label Free MS-based Quatitative proteomics

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

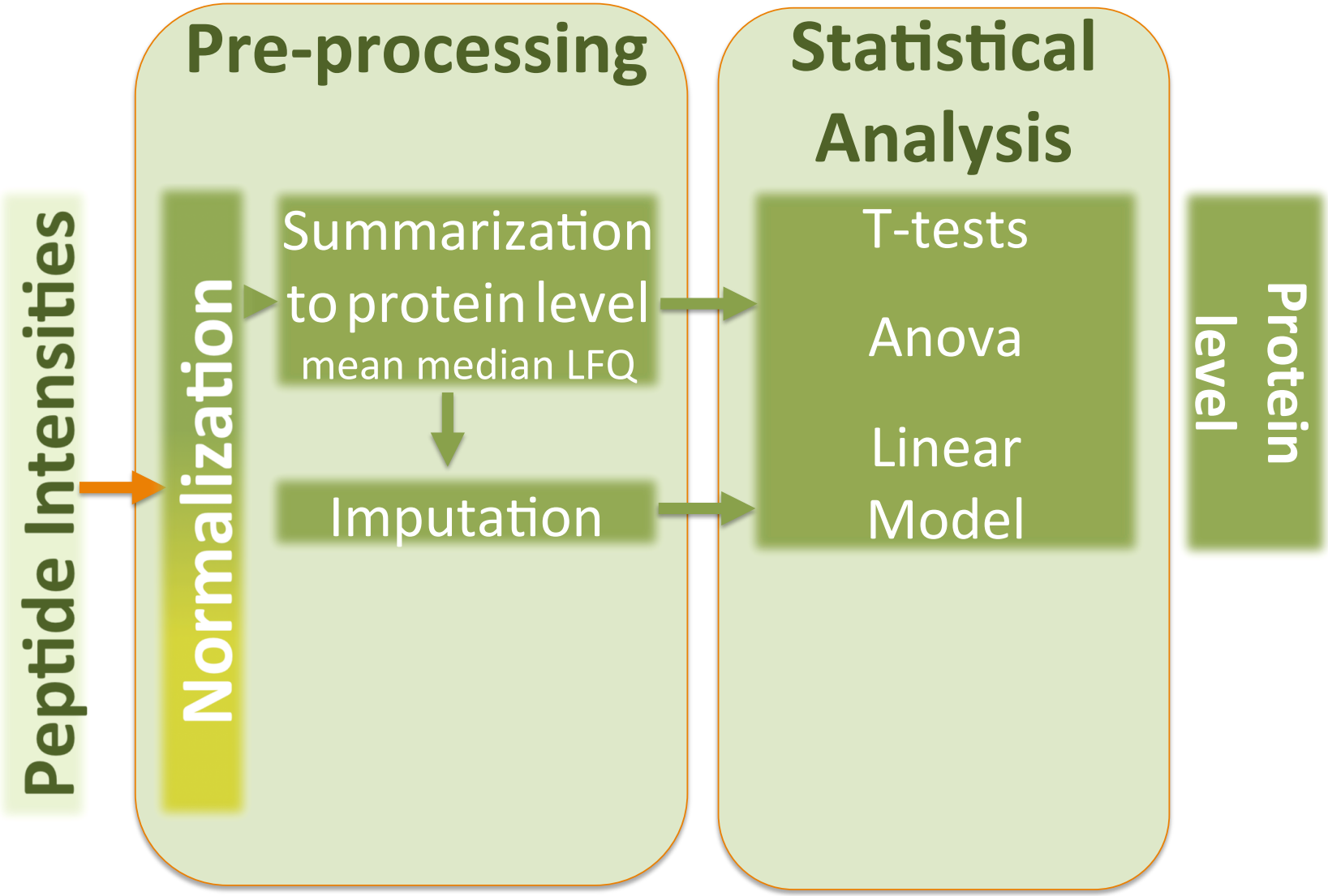


Challenges in Label Free MS-based Quatitative proteomics

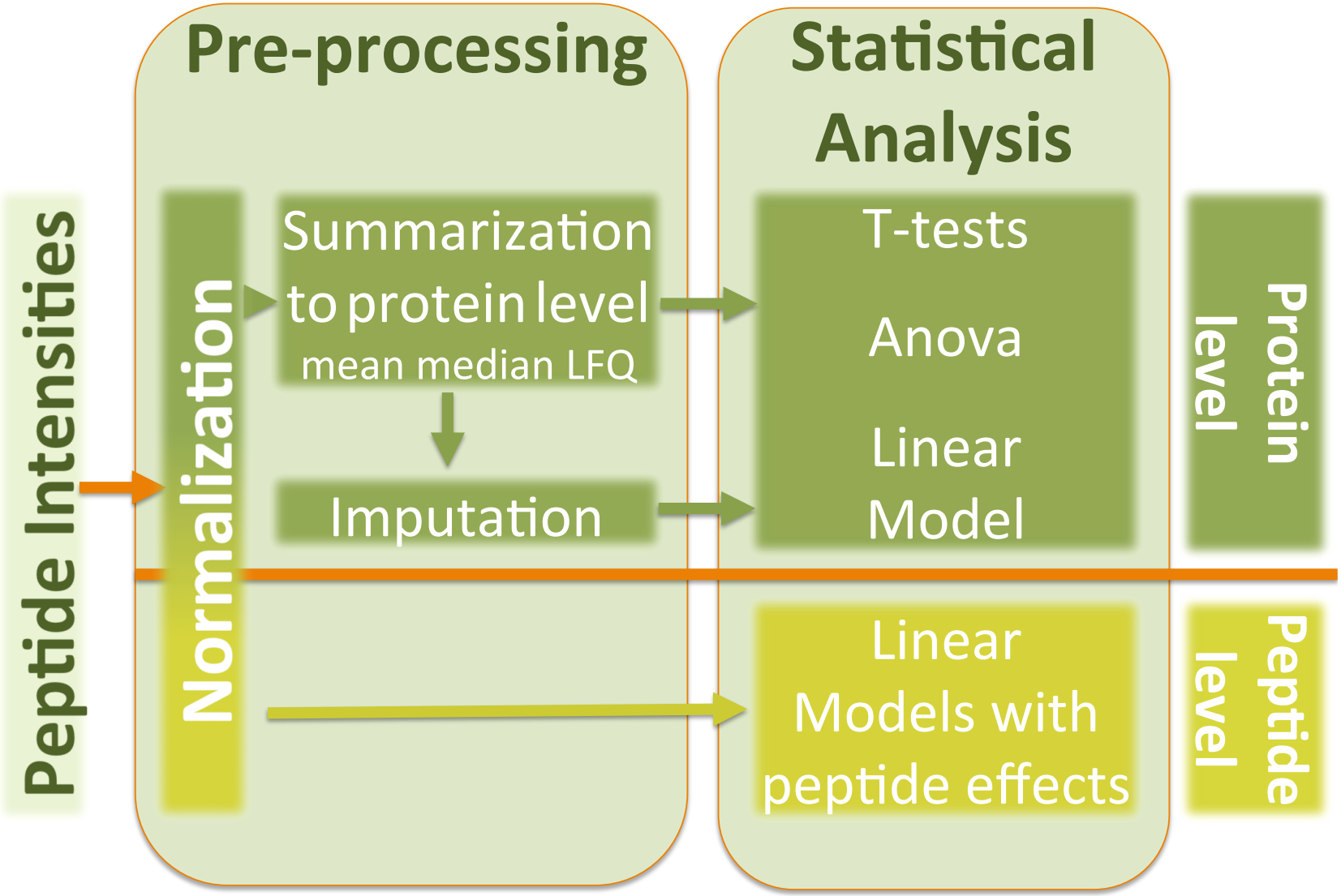
We need information on protein level!



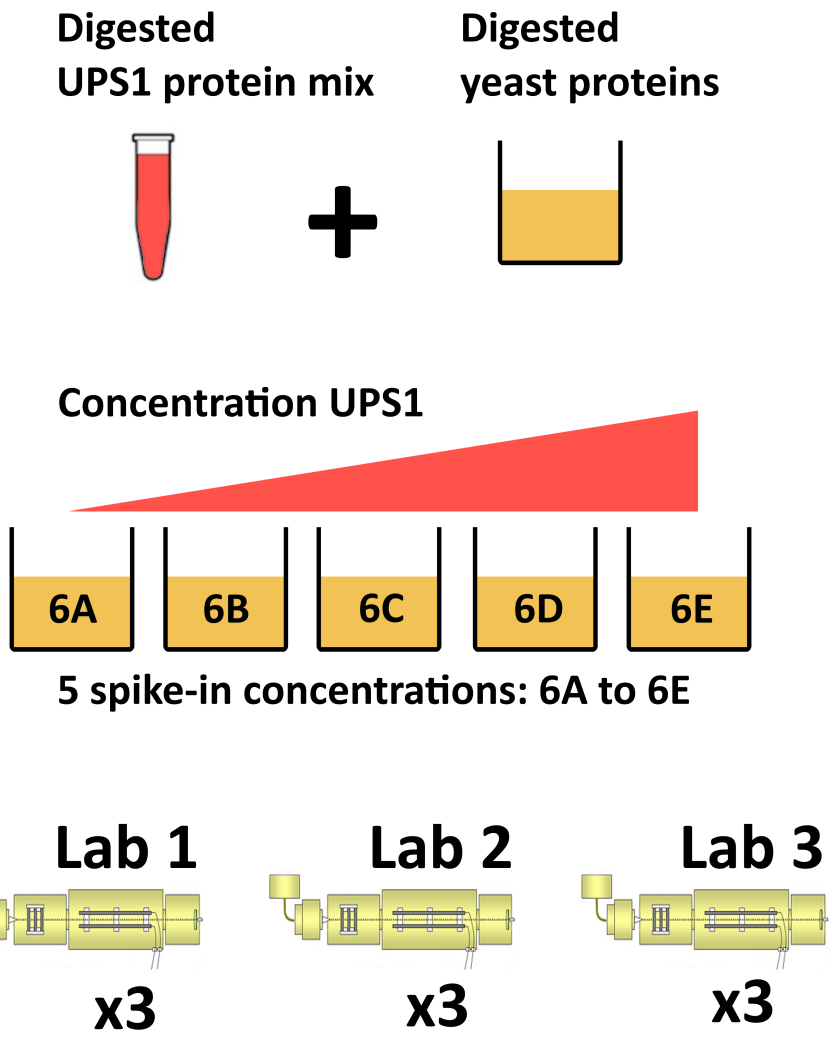
Label-free Quantitative Proteomics Data Analysis Pipelines



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)

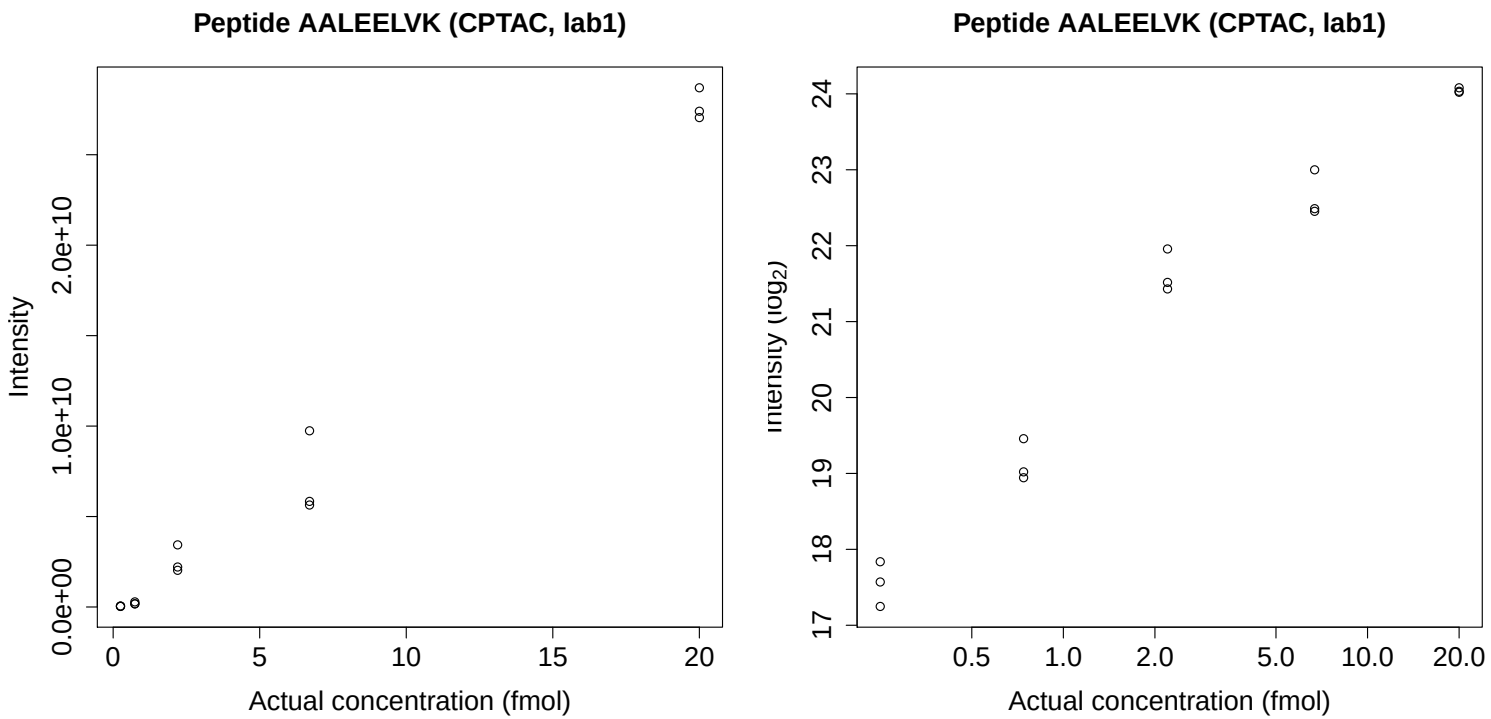
- Many methods exist

Filtering

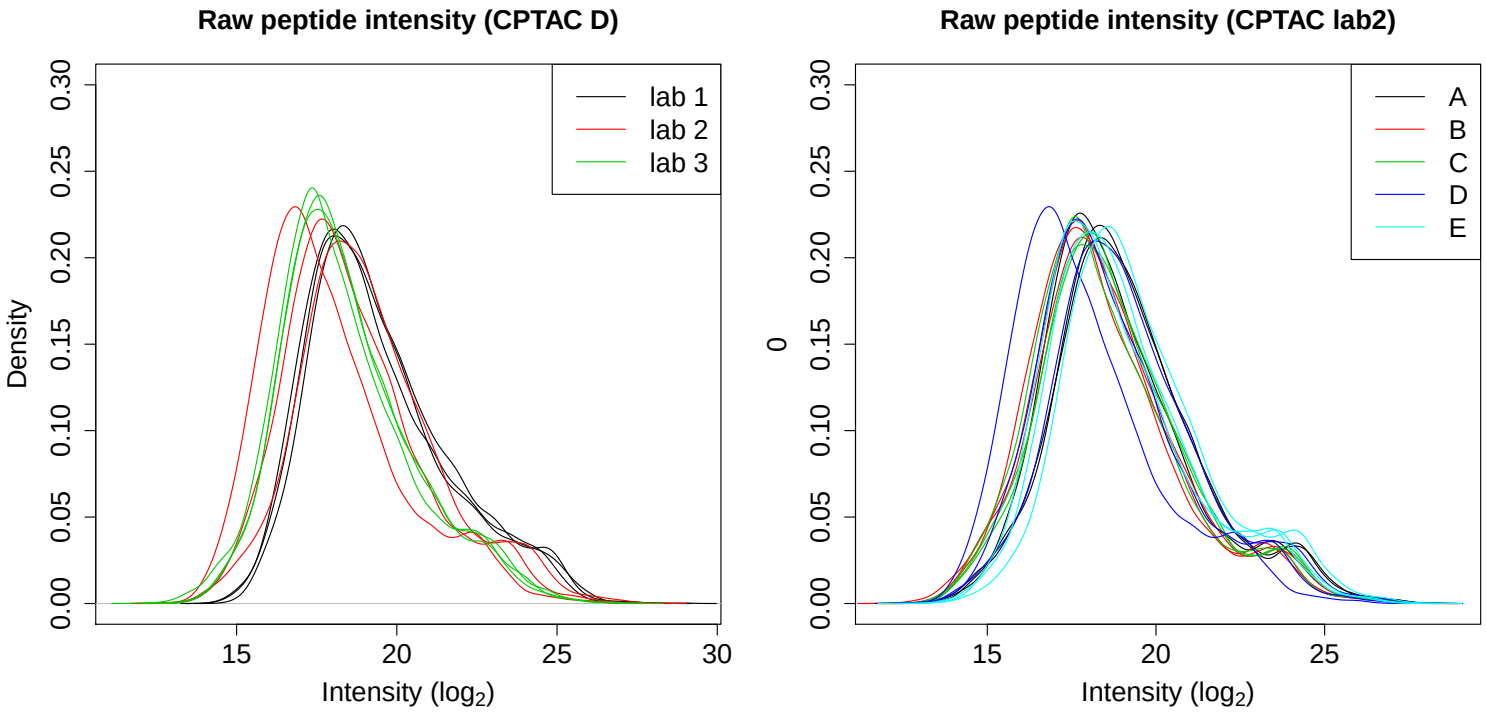
- Reverse sequences
- Only identified by modification site (only modified peptides detected)
- Razor peptides: non-unique peptides assigned to the protein group with the most other peptides
- Contaminants
- Peptides few identifications
- Proteins that are only identified with one or a few peptides

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

Log-transformation



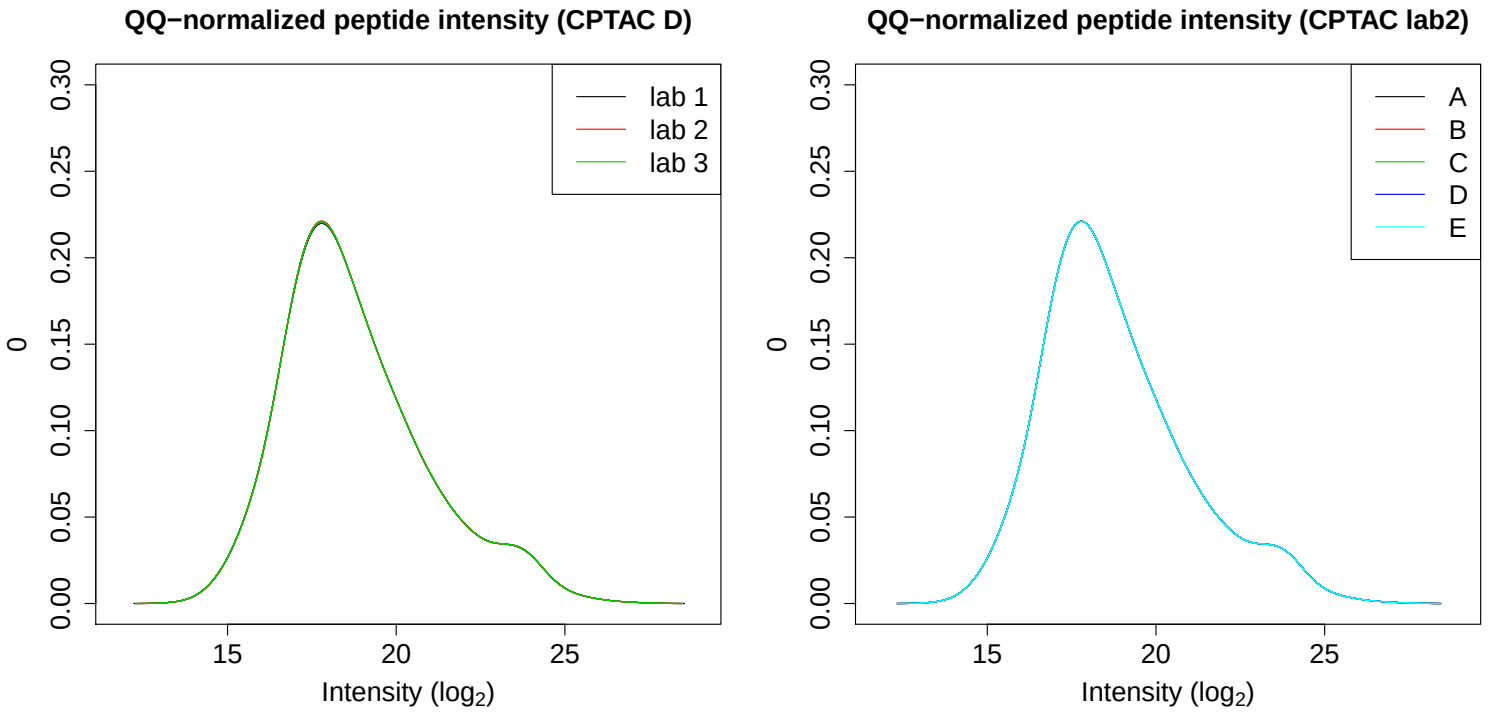
Variability more equal upon log transformation: often multiplicative error structure of intensity-based read-outs



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

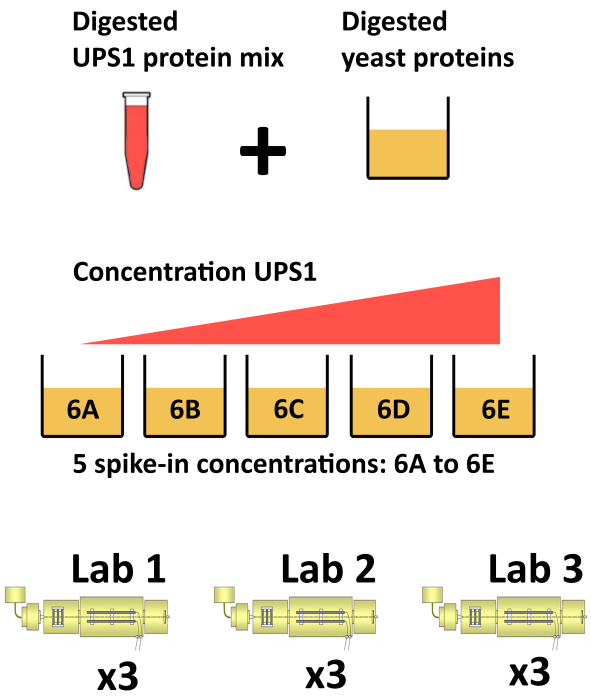


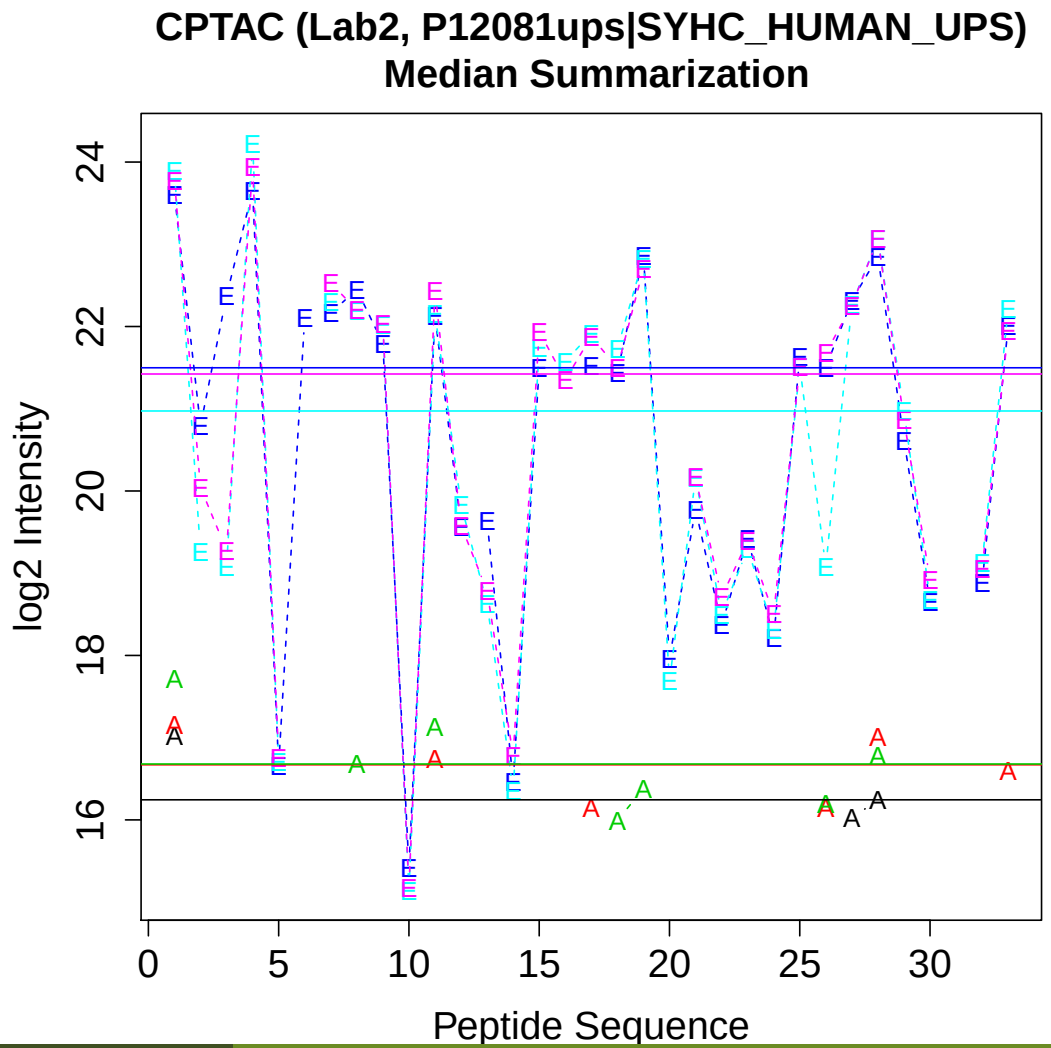
Even in very clean synthetic dataset (same background, only 48 UPS proteins can be different) the marginal peptide intensity distribution across samples can be quite distinct

- 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**

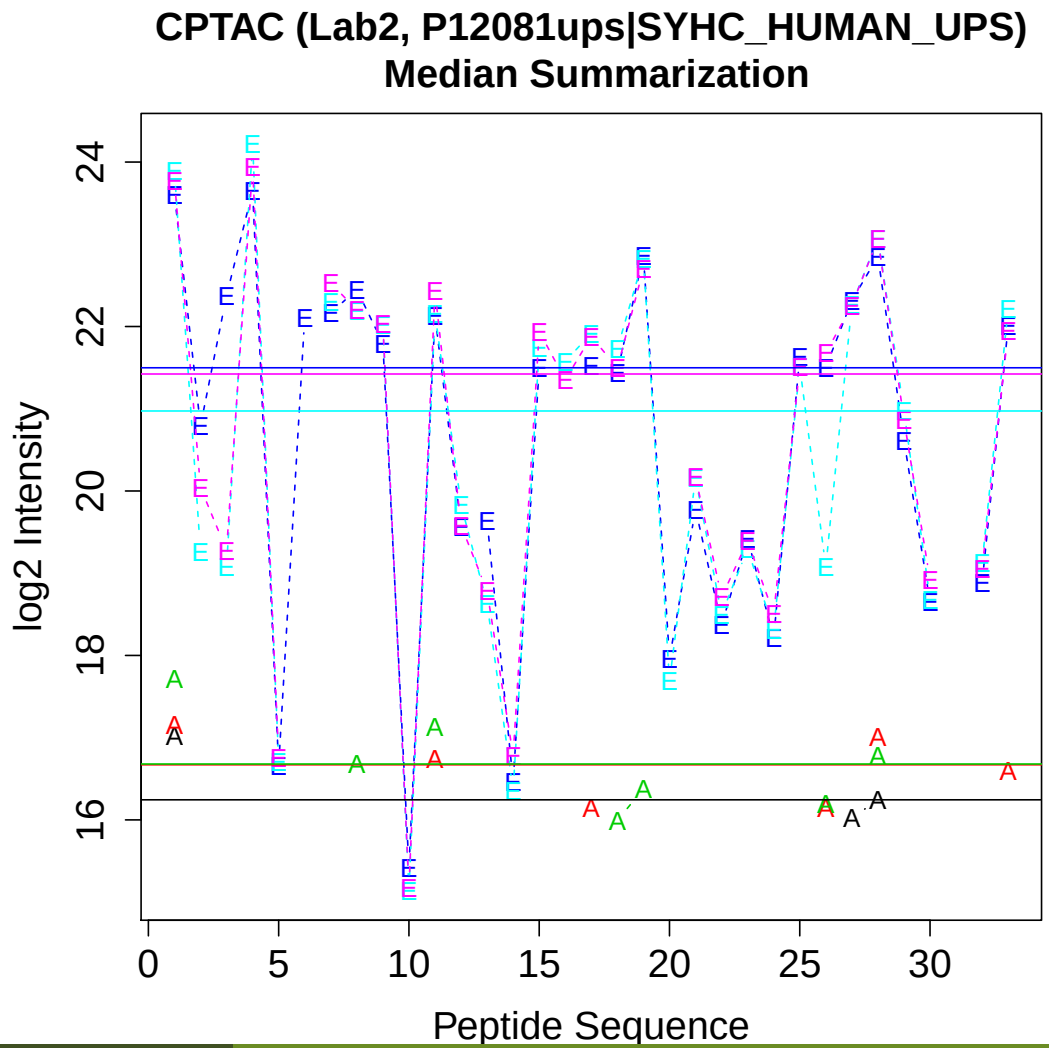
Summarization





Summarization

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



MaxLFQ summarization

a

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

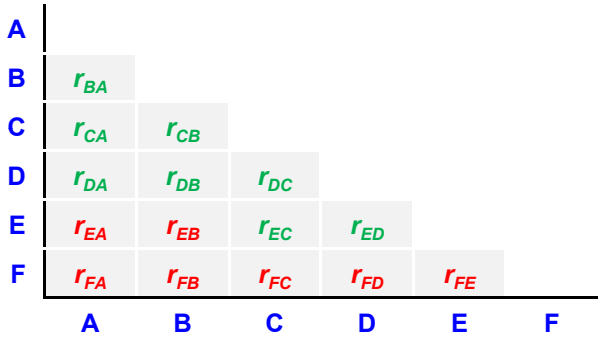
b

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		+			+		

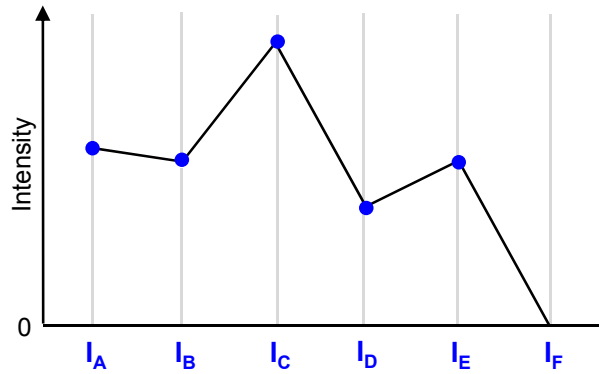
d



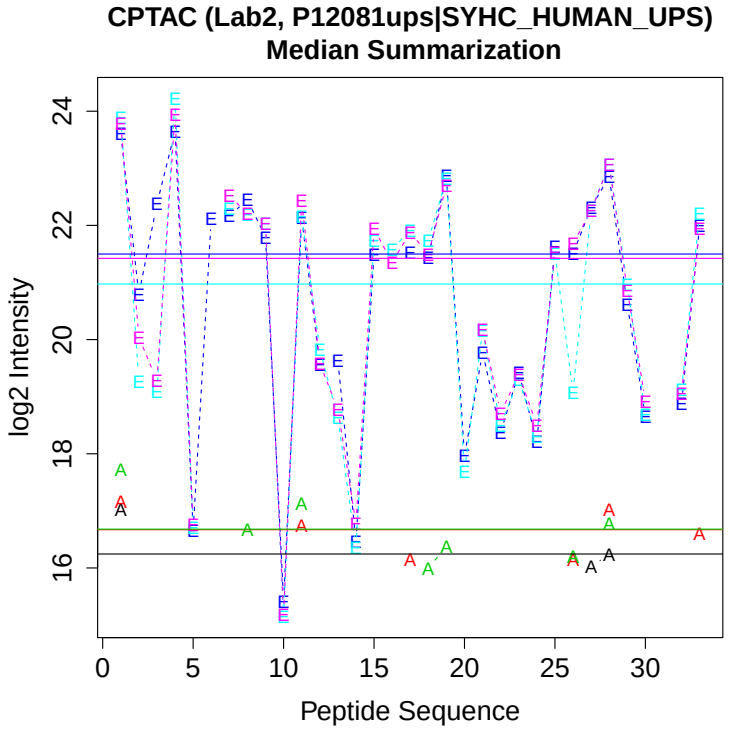
e

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

f

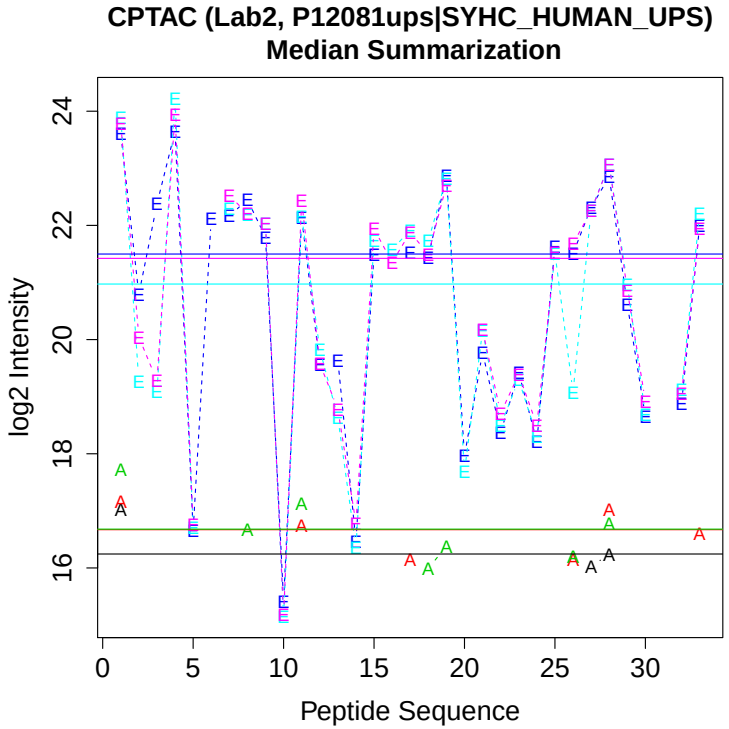


Peptide based model



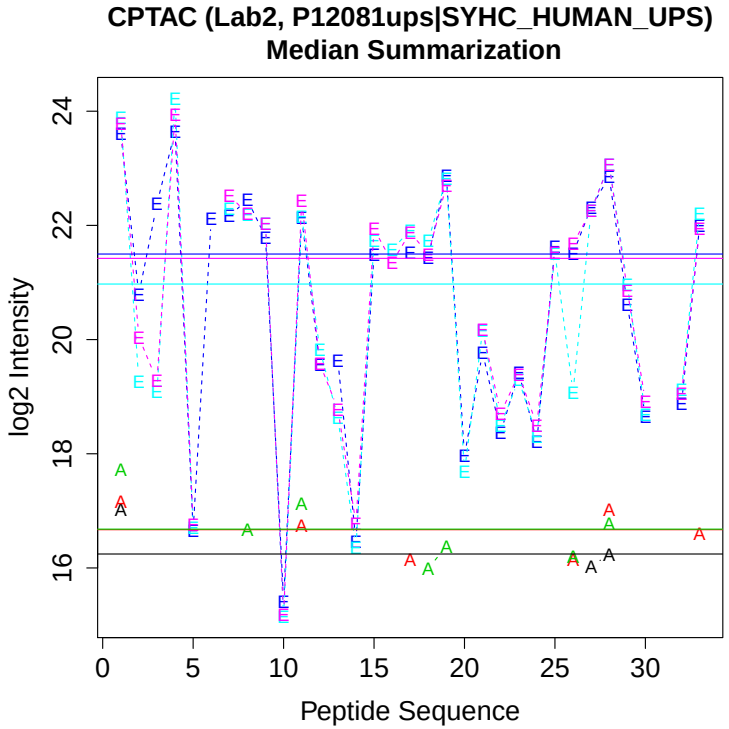
1. y: log2 transformed peptide intensities

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

$$y_{pept} \sim \underbrace{peptide}_{\text{peptide level}} + \underbrace{sample}_{\text{protein level}}$$