QFeatures Structure

Lieven Clement

statOmics, Ghent University

Contents

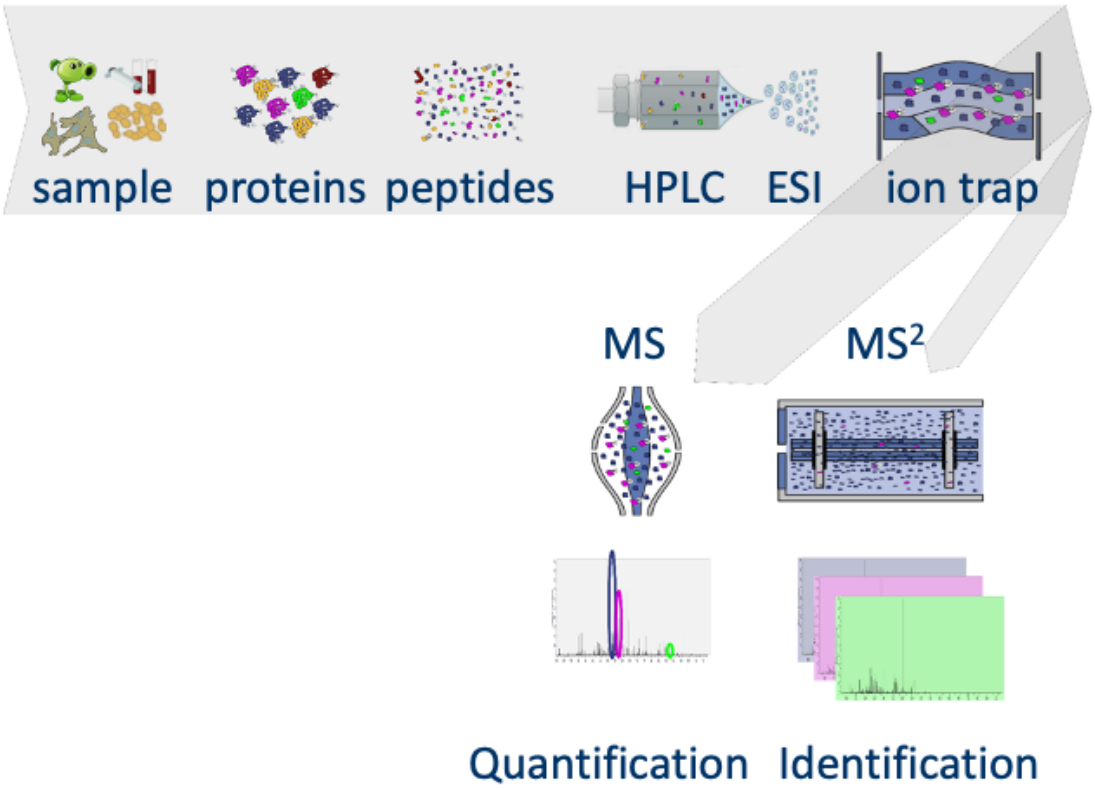
Outline

1. Introduction
2. Preprocessing
    - Log-transformation
    - Filtering
    - Normalization
    - Summarization

---

1 Intro: Challenges in Label-Free Quantitative Proteomics

1.1 MS-based workflow



- Peptide Characteristics
  - Modifications
  - Ionisation Efficiency: huge variability
  - Identification
    - Misidentification → outliers
    - MS^2 selection on peptide abundance
    - Context depending missingness
    - Non-random missingness

→ Unbalanced pepide identifications across samples and messy data

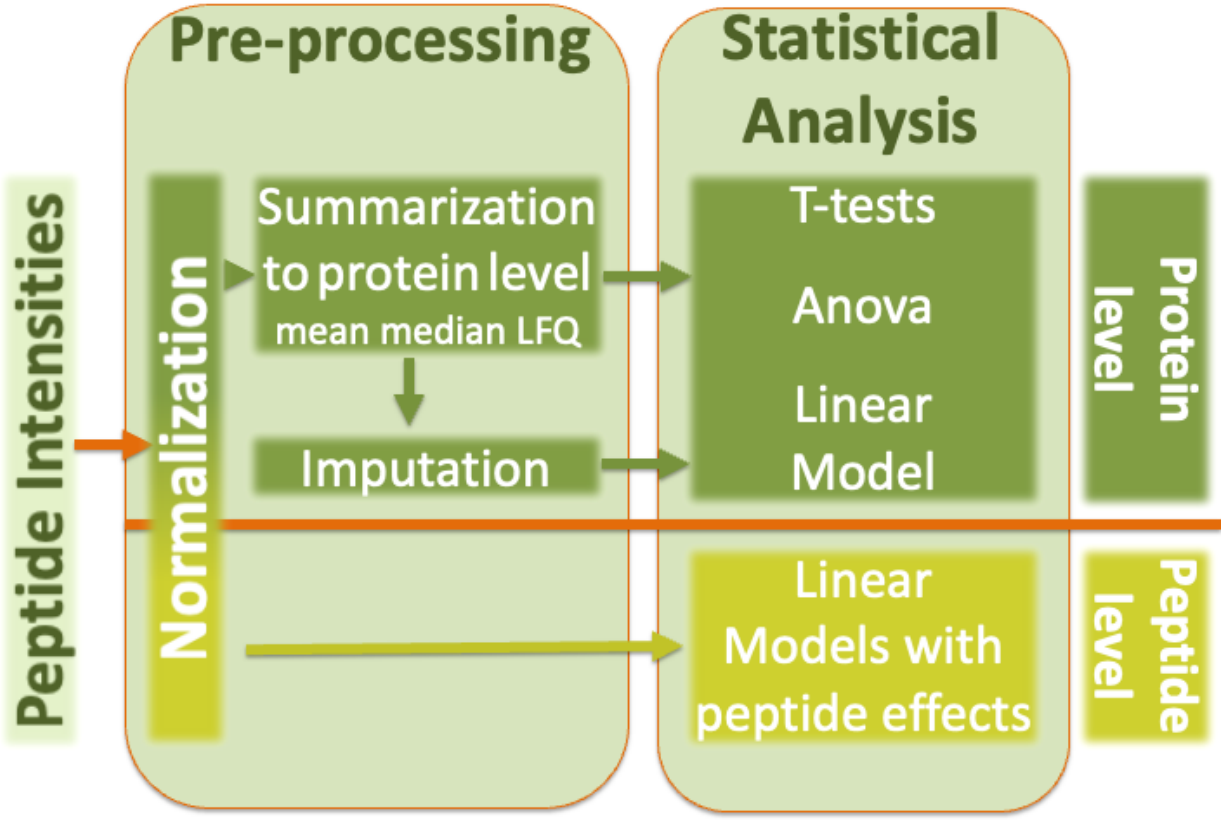
---

1.2 Level of quantification

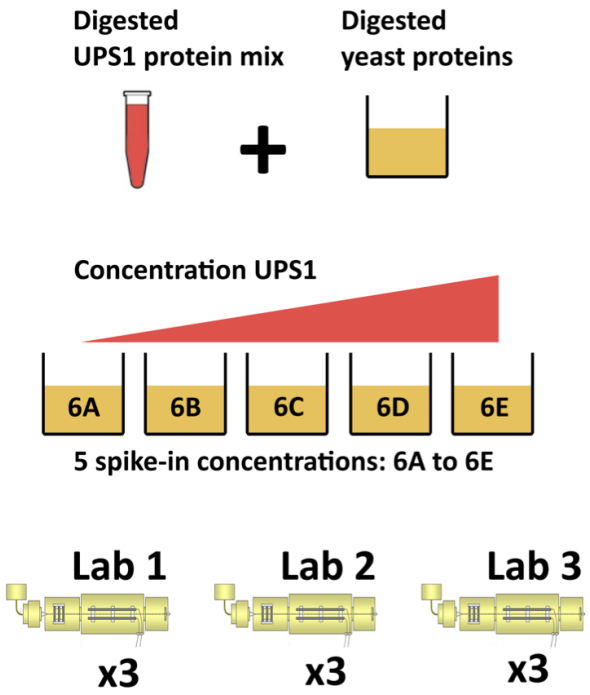
- MS-based proteomics returns peptides: pieces of proteins

- Quantification commonly required on the protein level

1.3 Label-free Quantitative Proteomics Data Analysis Workflows



1.4 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

2 Import the data in R

2.1 Data infrastructure

- We use the `QFeatures` package that provides the infrastructure to
  - store,
  - process,
  - manipulate and
  - analyse quantitative data/features from mass spectrometry experiments.
- It is based on the `SummarizedExperiment` and `MultiAssayExperiment` classes.

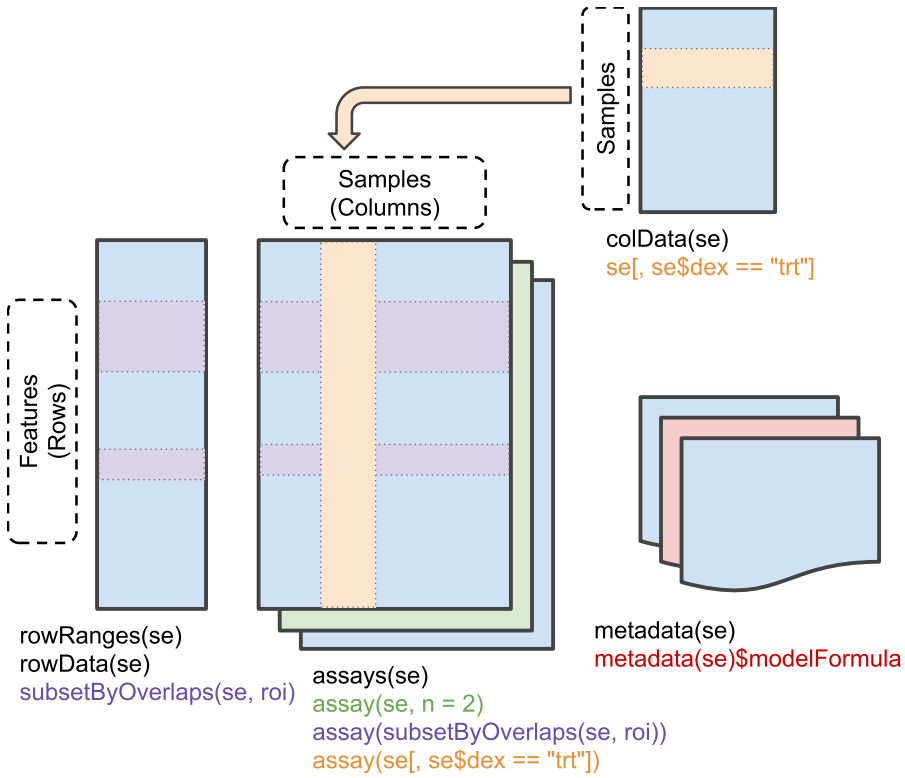


Figure 1: Conceptual representation of a 'SummarizedExperiment' object. Assays contain information on the measured omics features (rows) for different samples (columns). The 'rowData' contains information on the omics features, the 'colData' contains information on the samples, i.e. experimental design etc.

- Assays in a QFeatures object have a hierarchical relation:
  - proteins are composed of peptides,
  - themselves produced by spectra
  - relations between assays are tracked and recorded throughout data processing

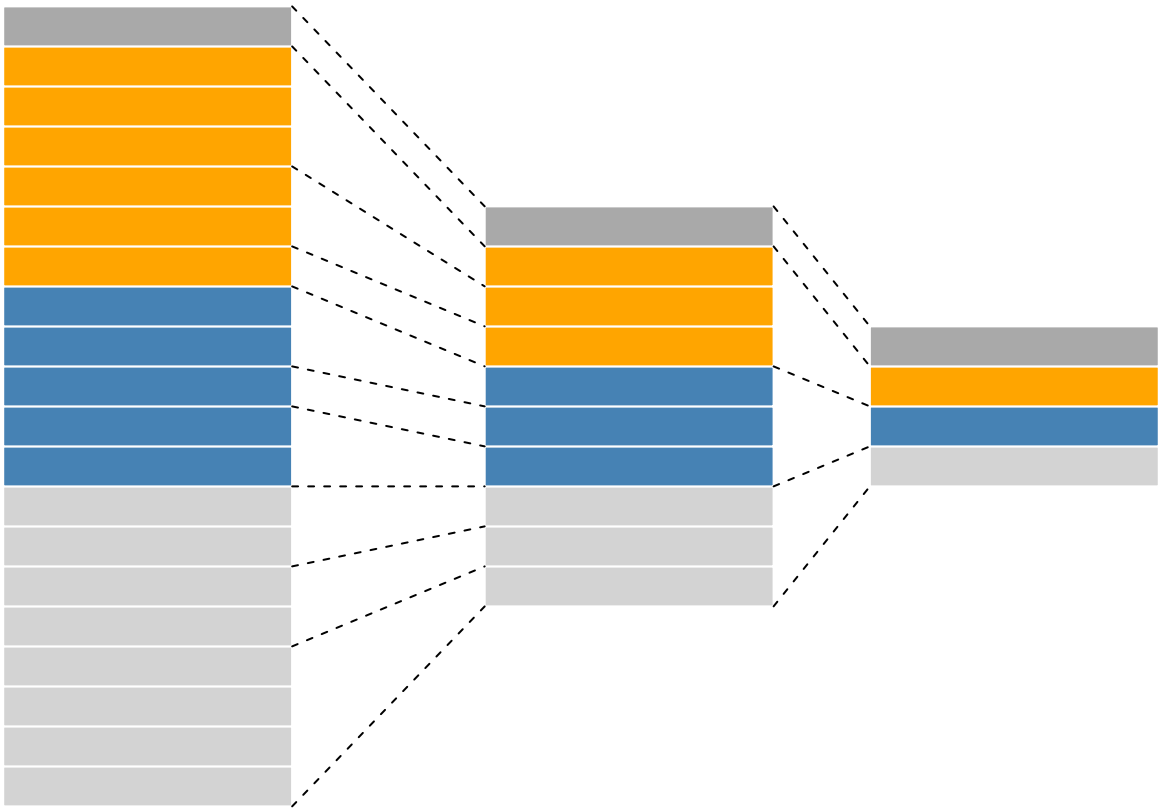


Figure 2: Conceptual representation of a `QFeatures` object and the aggregative relation between different assays.

2.2 Import data in R

2.2.1 Load libraries

Click to see code

```
library(tidyverse)
library(limma)
library(QFeatures)
library(msqrob2)
library(plotly)
library(ggplot2)
```

2.2.2 Read data

Click to see background and code

1. We use a peptides.txt file from MS-data quantified with maxquant that contains MS1 intensities summarized at the peptide level.

```
peptidesFile <- "https://raw.githubusercontent.com/statOmics/PDA/data/quantification/fullCptacDatasSetN
```

2. Maxquant stores the intensity data for the different samples in columnns that start with Intensity. We can retreive the column names with the intensity data with the code below:

```
ecols <- grep("Intensity\\.", names(read.delim(peptidesFile)))
```

3. Read the data and store it in QFeatures object

```
pe <- readQFeatures(
  table = peptidesFile,
  fnames = 1,
  ecol = ecols,
  name = "peptideRaw", sep="\t")
```

2.2.3 Explore object

Click to see background and code

- The rowData contains information on the features (peptides) in the assay. E.g. Sequence, protein, ...

```
head(rowData(pe[["peptideRaw"]])[,c("Proteins","Sequence","Charges","Intensity","Experiment.6A_1","Expe
```

```
## DataFrame with 6 rows and 6 columns
##                          Proteins      Sequence     Charges Intensity
##                       <character>   <character> <character> <numeric>
## AAAAGAGGAGDSGDAVTK sp|P38915|... AAAAGAGGAG...           2   1190800
## AAAALAGGK          sp|Q3E792|...     AAAALAGGK           2 280990000
## AAAALAGGKK         sp|Q3E792|...    AAAALAGGKK           2  33360000
## AAADALSDLEIK       sp|P09938|... AAADALSDLE...           2  54622000
## AAADALSDLEIKDSK    sp|P09938|... AAADALSDLE...           3  18910000
## AAAEEFQR           sp|P53075|...      AAAEEFQR           2   1158600
##                    Experiment.6A_1 Experiment.6A_2
##                          <integer>       <integer>
## AAAAGAGGAGDSGDAVTK              NA               1
## AAAALAGGK                       NA               1
## AAAALAGGKK                      NA               1
## AAADALSDLEIK                     1               1
## AAADALSDLEIKDSK                  1               1
## AAAEEFQR                        NA              NA
```

- The colData contains information on the samples

```
colData(pe)
```

```
## DataFrame with 45 rows and 0 columns
```

- No information is stored yet on the design.

```
pe %>% colnames
```

```
## CharacterList of length 1
## [["peptideRaw"]] Intensity.6A_1 Intensity.6A_2 ... Intensity.6E_9
```

- Note, that the sample names include the spike-in condition.
- They also end on a number.
  - 1-3 is from lab 1,
  - 4-6 from lab 2 and
  - 7-9 from lab 3.
- We update the colData with information on the design

```
colData(pe)$lab <- rep(rep(paste0("lab",1:3),each=3),5) %>% as.factor
colData(pe)$condition <- pe[["peptideRaw"]] %>% colnames %>% substr(12,12) %>% as.factor
colData(pe)$spikeConcentration <- rep(c(A = 0.25, B = 0.74, C = 2.22, D = 6.67, E = 20),each = 9)
```

- We explore the colData again

```
colData(pe)
```

```
## DataFrame with 45 rows and 3 columns
##                      lab condition spikeConcentration
##                 <factor>  <factor>          <numeric>
## Intensity.6A_1     lab1         A               0.25
## Intensity.6A_2     lab1         A               0.25
## Intensity.6A_3     lab1         A               0.25
## Intensity.6A_4     lab2         A               0.25
## Intensity.6A_5     lab2         A               0.25
## ...                 ...       ...                ...
## Intensity.6E_5     lab2         E                 20
## Intensity.6E_6     lab2         E                 20
## Intensity.6E_7     lab3         E                 20
## Intensity.6E_8     lab3         E                 20
## Intensity.6E_9     lab3         E                 20
```