Statistical Methods for Quantitative MS-based Proteomics: Peptide-level Models for Summarization and Inference

Lieven Clement

statOmics, Ghent University

Contents

This is part of the online course Proteomics Data Analysis 2021 (PDA21)

1 Import the data in R

1.1 Load libraries

Click to see code

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

1.2 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/PDA21/data/quantification/fullCptacDatasSe
```

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")
```

1.3 Explore object

Click to see background and code

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

```
rowData(pe[["peptideRaw"]])
```

```
## DataFrame with 11466 rows and 143 columns
##                          Sequence N.term.cleavage.window C.term.cleavage.window
##                       <character>            <character>            <character>
## AAAAGAGGAGDSGDAVTK  AAAAGAGGAG...          EHQHDEQKAA...          DSGDAVTKIG...
## AAAALAGGK              AAAALAGGK          QQLSKAAKAA...          AAALAGGKKS...
## AAAALAGGKK            AAAALAGGKK          QQLSKAAKAA...          AALAGGKKSK...
## AAADALSDLEIK        AAADALSDLE...          MPKETPSKAA...          ALSDLEIKDS...
## AAADALSDLEIKDSK     AAADALSDLE...          MPKETPSKAA...          DLEIKDSKSN...
## ...                           ...                    ...                    ...
## YYSIYDLGNNAVGLAK    YYSIYDLGNN...          VGDAFLRKYY...          NNAVGLAKAI...
## YYTFNGPNYNENETIR    YYTFNGPNYN...          FKDGSYPKYY...          YNENETIRHI...
## YYTITEVATR             YYTITEVATR          QEWDINERYY...          TITEVATRAK...
## YYTVFDRDNNR         YYTVFDRDNN...          LGDVFIGRYY...          VFDRDNNRVG...
## YYTVFDRDNNRVGFAEAAR YYTVFDRDNN...          LGDVFIGRYY...          VGFAEAARL_...
##                     Amino.acid.before First.amino.acid Second.amino.acid
##                           <character>      <character>       <character>
## AAAAGAGGAGDSGDAVTK                  K                A                 A
## AAAALAGGK                           K                A                 A
## AAAALAGGKK                          K                A                 A
## AAADALSDLEIK                        K                A                 A
```

```
## AAADALSDLEIKDSK                    K               A               A
## ...                              ...             ...             ...
## YYSIYDLGNNAVGLAK                   K               Y               Y
## YYTFNGPNYNENETIR                   K               Y               Y
## YYTITEVATR                         R               Y               Y
## YYTVFDRDNNR                        R               Y               Y
## YYTVFDRDNNRVGFAEAAR                R               Y               Y
##                      Second.last.amino.acid Last.amino.acid Amino.acid.after
##                                   <character>     <character>      <character>
## AAAAGAGGAGDSGDAVTK                         T               K                I
## AAAALAGGK                                  G               K                K
## AAAALAGGKK                                 K               K                S
## AAADALSDLEIK                               I               K                D
## AAADALSDLEIKDSK                            S               K                S
## ...                                      ...             ...              ...
## YYSIYDLGNNAVGLAK                           A               K                A
## YYTFNGPNYNENETIR                           I               R                H
## YYTITEVATR                                 T               R                A
## YYTVFDRDNNR                                N               R                V
## YYTVFDRDNNRVGFAEAAR                        A               R                L
##                       A.Count   R.Count   N.Count   D.Count   C.Count   Q.Count
##                     <integer> <integer> <integer> <integer> <integer> <integer>
## AAAAGAGGAGDSGDAVTK          7         0         0         2         0         0
## AAAALAGGK                   5         0         0         0         0         0
## AAAALAGGKK                  5         0         0         0         0         0
## AAADALSDLEIK                4         0         0         2         0         0
## AAADALSDLEIKDSK             4         0         0         3         0         0
## ...                       ...       ...       ...       ...       ...       ...
## YYSIYDLGNNAVGLAK            2         0         2         1         0         0
## YYTFNGPNYNENETIR            0         1         4         0         0         0
## YYTITEVATR                  1         1         0         0         0         0
## YYTVFDRDNNR                 0         2         2         2         0         0
## YYTVFDRDNNRVGFAEAAR         3         3         2         2         0         0
##                       E.Count   G.Count   H.Count   I.Count   L.Count   K.Count
##                     <integer> <integer> <integer> <integer> <integer> <integer>
## AAAAGAGGAGDSGDAVTK          0         5         0         0         0         1
## AAAALAGGK                   0         2         0         0         1         1
## AAAALAGGKK                  0         2         0         0         1         2
## AAADALSDLEIK                1         0         0         1         2         1
## AAADALSDLEIKDSK             1         0         0         1         2         2
## ...                       ...       ...       ...       ...       ...       ...
## YYSIYDLGNNAVGLAK            0         2         0         1         2         1
## YYTFNGPNYNENETIR            2         1         0         1         0         0
## YYTITEVATR                  1         0         0         1         0         0
## YYTVFDRDNNR                 0         0         0         0         0         0
## YYTVFDRDNNRVGFAEAAR         1         1         0         0         0         0
##                       M.Count   F.Count   P.Count   S.Count   T.Count   W.Count
##                     <integer> <integer> <integer> <integer> <integer> <integer>
## AAAAGAGGAGDSGDAVTK          0         0         0         1         1         0
## AAAALAGGK                   0         0         0         0         0         0
## AAAALAGGKK                  0         0         0         0         0         0
## AAADALSDLEIK                0         0         0         1         0         0
## AAADALSDLEIKDSK             0         0         0         2         0         0
## ...                       ...       ...       ...       ...       ...       ...
```

```
## YYSIYDLGNNAVGLAK            0         0         0         1         0         0
## YYTFNGPNYNENETIR            0         1         1         0         2         0
## YYTITEVATR                  0         0         0         0         3         0
## YYTVFDRDNNR                 0         1         0         0         1         0
## YYTVFDRDNNRVGFAEAAR         0         2         0         0         1         0
##                       Y.Count   V.Count   U.Count    Length Missed.cleavages
##                     <integer> <integer> <integer> <integer>        <integer>
## AAAAGAGGAGDSGDAVTK          0         1         0        18                0
## AAAALAGGK                   0         0         0         9                0
## AAAALAGGKK                  0         0         0        10                1
## AAADALSDLEIK                0         0         0        12                0
## AAADALSDLEIKDSK             0         0         0        15                1
## ...                       ...       ...       ...       ...              ...
## YYSIYDLGNNAVGLAK            3         1         0        16                0
## YYTFNGPNYNENETIR            3         0         0        16                0
## YYTITEVATR                  2         1         0        10                0
## YYTVFDRDNNR                 2         1         0        11                1
## YYTVFDRDNNRVGFAEAAR         2         2         0        19                2
##                          Mass      Proteins Leading.razor.protein
##                     <numeric>   <character>           <character>
## AAAAGAGGAGDSGDAVTK   1445.675 sp|P38915|...         sp|P38915|...
## AAAALAGGK             728.418 sp|Q3E792|...         sp|Q3E792|...
## AAAALAGGKK            856.513 sp|Q3E792|...         sp|Q3E792|...
## AAADALSDLEIK         1215.635 sp|P09938|...         sp|P09938|...
## AAADALSDLEIKDSK      1545.789 sp|P09938|...         sp|P09938|...
## ...                       ...          ...                 ...
## YYSIYDLGNNAVGLAK      1759.88 sp|P07267|...         sp|P07267|...
## YYTFNGPNYNENETIR      1993.88 sp|Q00955|...         sp|Q00955|...
## YYTITEVATR            1215.61 sp|P38891|...         sp|P38891|...
## YYTVFDRDNNR           1461.66 P07339ups|...         P07339ups|...
## YYTVFDRDNNRVGFAEAAR   2263.08 P07339ups|...         P07339ups|...
##                     Start.position End.position Unique..Groups.
##                          <integer>    <integer>     <character>
## AAAAGAGGAGDSGDAVTK              97          114             yes
## AAAALAGGK                       13           21             yes
## AAAALAGGKK                      13           22             yes
## AAADALSDLEIK                     9           20             yes
## AAADALSDLEIKDSK                  9           23             yes
## ...                            ...          ...             ...
## YYSIYDLGNNAVGLAK               388          403             yes
## YYTFNGPNYNENETIR              1275         1290             yes
## YYTITEVATR                     311          320             yes
## YYTVFDRDNNR                    225          235             yes
## YYTVFDRDNNRVGFAEAAR            225          243             yes
##                     Unique..Proteins.     Charges        PEP     Score
##                           <character> <character>  <numeric> <numeric>
## AAAAGAGGAGDSGDAVTK                yes           2 1.1843e-05    82.942
## AAAALAGGK                          no           2 7.4562e-06   134.810
## AAAALAGGKK                         no           2 3.3094e-09   143.730
## AAADALSDLEIK                      yes           2 9.1593e-23   182.230
## AAADALSDLEIKDSK                   yes           3 1.5319e-04    73.927
## ...                               ...         ...         ...       ...
## YYSIYDLGNNAVGLAK                  yes           2 7.7415e-37   174.240
## YYTFNGPNYNENETIR                  yes           2 4.2208e-21   147.750
```

```
## YYTITEVATR                        yes           2 1.3566e-04   109.160
## YYTVFDRDNNR                       yes           2 6.1425e-04   110.930
## YYTVFDRDNNRVGFAEAAR               yes           3 8.9859e-04    59.728
##                     Identification.type.6A_1 Identification.type.6A_2
##                                    <character>              <character>
## AAAAGAGGAGDSGDAVTK            By matchin...                 By MS/MS
## AAAALAGGK                     By matchin...            By matchin...
## AAAALAGGKK                    By matchin...            By matchin...
## AAADALSDLEIK                       By MS/MS                 By MS/MS
## AAADALSDLEIKDSK               By matchin...            By matchin...
## ...                                    ...                      ...
## YYSIYDLGNNAVGLAK              By matchin...            By matchin...
## YYTFNGPNYNENETIR              By matchin...            By matchin...
## YYTITEVATR                         By MS/MS            By matchin...
## YYTVFDRDNNR                   By matchin...            By matchin...
## YYTVFDRDNNRVGFAEAAR           By matchin...            By matchin...
##                     Identification.type.6A_3 Identification.type.6A_4
##                                    <character>              <character>
## AAAAGAGGAGDSGDAVTK            By matchin...                 By MS/MS
## AAAALAGGK                     By matchin...                 By MS/MS
## AAAALAGGKK                    By matchin...                 By MS/MS
## AAADALSDLEIK                  By matchin...                 By MS/MS
## AAADALSDLEIKDSK               By matchin...                 By MS/MS
## ...                                    ...                      ...
## YYSIYDLGNNAVGLAK              By matchin...                 By MS/MS
## YYTFNGPNYNENETIR              By matchin...                 By MS/MS
## YYTITEVATR                    By matchin...            By matchin...
## YYTVFDRDNNR                   By matchin...            By matchin...
## YYTVFDRDNNRVGFAEAAR           By matchin...            By matchin...
##                     Identification.type.6A_5 Identification.type.6A_6
##                                    <character>              <character>
## AAAAGAGGAGDSGDAVTK            By matchin...            By matchin...
## AAAALAGGK                     By matchin...            By matchin...
## AAAALAGGKK                    By matchin...            By matchin...
## AAADALSDLEIK                       By MS/MS                 By MS/MS
## AAADALSDLEIKDSK                    By MS/MS                 By MS/MS
## ...                                    ...                      ...
## YYSIYDLGNNAVGLAK                   By MS/MS                 By MS/MS
## YYTFNGPNYNENETIR                   By MS/MS                 By MS/MS
## YYTITEVATR                    By matchin...            By matchin...
## YYTVFDRDNNR                   By matchin...            By matchin...
## YYTVFDRDNNRVGFAEAAR           By matchin...            By matchin...
##                     Identification.type.6A_7 Identification.type.6A_8
##                                    <character>              <character>
## AAAAGAGGAGDSGDAVTK                 By MS/MS                 By MS/MS
## AAAALAGGK                          By MS/MS                 By MS/MS
## AAAALAGGKK                         By MS/MS                 By MS/MS
## AAADALSDLEIK                       By MS/MS            By matchin...
## AAADALSDLEIKDSK                    By MS/MS                 By MS/MS
## ...                                    ...                      ...
## YYSIYDLGNNAVGLAK              By matchin...            By matchin...
## YYTFNGPNYNENETIR              By matchin...            By matchin...
## YYTITEVATR                         By MS/MS            By matchin...
## YYTVFDRDNNR                   By matchin...            By matchin...
```

```
## YYTVFDRDNNRVGFAEAAR            By matchin...            By matchin...
##                       Identification.type.6A_9 Identification.type.6B_1
##                                     <character>              <character>
## AAAAGAGGAGDSGDAVTK                     By MS/MS            By matchin...
## AAAALAGGK                              By MS/MS                 By MS/MS
## AAAALAGGKK                             By MS/MS            By matchin...
## AAADALSDLEIK                           By MS/MS                 By MS/MS
## AAADALSDLEIKDSK                        By MS/MS            By matchin...
## ...                                         ...                      ...
## YYSIYDLGNNAVGLAK               By matchin...            By matchin...
## YYTFNGPNYNENETIR               By matchin...            By matchin...
## YYTITEVATR                     By matchin...                 By MS/MS
## YYTVFDRDNNR                    By matchin...            By matchin...
## YYTVFDRDNNRVGFAEAAR            By matchin...            By matchin...
##                       Identification.type.6B_2 Identification.type.6B_3
##                                     <character>              <character>
## AAAAGAGGAGDSGDAVTK             By matchin...            By matchin...
## AAAALAGGK                      By matchin...            By matchin...
## AAAALAGGKK                           By MS/MS                 By MS/MS
## AAADALSDLEIK                         By MS/MS            By matchin...
## AAADALSDLEIKDSK                By matchin...            By matchin...
## ...                                       ...                      ...
## YYSIYDLGNNAVGLAK               By matchin...            By matchin...
## YYTFNGPNYNENETIR               By matchin...            By matchin...
## YYTITEVATR                     By matchin...            By matchin...
## YYTVFDRDNNR                    By matchin...            By matchin...
## YYTVFDRDNNRVGFAEAAR            By matchin...            By matchin...
##                       Identification.type.6B_4 Identification.type.6B_5
##                                     <character>              <character>
## AAAAGAGGAGDSGDAVTK             By matchin...            By matchin...
## AAAALAGGK                      By matchin...            By matchin...
## AAAALAGGKK                     By matchin...            By matchin...
## AAADALSDLEIK                         By MS/MS                 By MS/MS
## AAADALSDLEIKDSK                      By MS/MS                 By MS/MS
## ...                                       ...                      ...
## YYSIYDLGNNAVGLAK                     By MS/MS            By matchin...
## YYTFNGPNYNENETIR                     By MS/MS                 By MS/MS
## YYTITEVATR                           By MS/MS                 By MS/MS
## YYTVFDRDNNR                    By matchin...            By matchin...
## YYTVFDRDNNRVGFAEAAR            By matchin...            By matchin...
##                       Identification.type.6B_6 Identification.type.6B_7
##                                     <character>              <character>
## AAAAGAGGAGDSGDAVTK             By matchin...            By matchin...
## AAAALAGGK                      By matchin...                 By MS/MS
## AAAALAGGKK                     By matchin...                 By MS/MS
## AAADALSDLEIK                         By MS/MS                 By MS/MS
## AAADALSDLEIKDSK                      By MS/MS                 By MS/MS
## ...                                       ...                      ...
## YYSIYDLGNNAVGLAK               By matchin...            By matchin...
## YYTFNGPNYNENETIR                     By MS/MS            By matchin...
## YYTITEVATR                     By matchin...            By matchin...
## YYTVFDRDNNR                    By matchin...            By matchin...
## YYTVFDRDNNRVGFAEAAR            By matchin...            By matchin...
##                       Identification.type.6B_8 Identification.type.6B_9
```

```
##                                 <character>              <character>
## AAAAGAGGAGDSGDAVTK                 By MS/MS                 By MS/MS
## AAAALAGGK                          By MS/MS                 By MS/MS
## AAAALAGGKK                         By MS/MS                 By MS/MS
## AAADALSDLEIK                    By matchin...            By matchin...
## AAADALSDLEIKDSK                    By MS/MS                 By MS/MS
## ...                                     ...                      ...
## YYSIYDLGNNAVGLAK                By matchin...            By matchin...
## YYTFNGPNYNENETIR                By matchin...            By matchin...
## YYTITEVATR                         By MS/MS             By matchin...
## YYTVFDRDNNR                     By matchin...            By matchin...
## YYTVFDRDNNRVGFAEAAR             By matchin...            By matchin...
##                       Identification.type.6C_1 Identification.type.6C_2
##                                 <character>              <character>
## AAAAGAGGAGDSGDAVTK              By matchin...            By matchin...
## AAAALAGGK                       By matchin...                 By MS/MS
## AAAALAGGKK                      By matchin...                 By MS/MS
## AAADALSDLEIK                       By MS/MS              By matchin...
## AAADALSDLEIKDSK                 By matchin...            By matchin...
## ...                                     ...                      ...
## YYSIYDLGNNAVGLAK                By matchin...            By matchin...
## YYTFNGPNYNENETIR                By matchin...            By matchin...
## YYTITEVATR                      By matchin...            By matchin...
## YYTVFDRDNNR                     By matchin...            By matchin...
## YYTVFDRDNNRVGFAEAAR             By matchin...            By matchin...
##                       Identification.type.6C_3 Identification.type.6C_4
##                                 <character>              <character>
## AAAAGAGGAGDSGDAVTK              By matchin...            By matchin...
## AAAALAGGK                       By matchin...                 By MS/MS
## AAAALAGGKK                      By matchin...                 By MS/MS
## AAADALSDLEIK                       By MS/MS                 By MS/MS
## AAADALSDLEIKDSK                 By matchin...                 By MS/MS
## ...                                     ...                      ...
## YYSIYDLGNNAVGLAK                By matchin...                 By MS/MS
## YYTFNGPNYNENETIR                By matchin...                 By MS/MS
## YYTITEVATR                         By MS/MS              By matchin...
## YYTVFDRDNNR                     By matchin...            By matchin...
## YYTVFDRDNNRVGFAEAAR             By matchin...            By matchin...
##                       Identification.type.6C_5 Identification.type.6C_6
##                                 <character>              <character>
## AAAAGAGGAGDSGDAVTK                 By MS/MS              By matchin...
## AAAALAGGK                       By matchin...            By matchin...
## AAAALAGGKK                      By matchin...            By matchin...
## AAADALSDLEIK                       By MS/MS                 By MS/MS
## AAADALSDLEIKDSK                    By MS/MS                 By MS/MS
## ...                                     ...                      ...
## YYSIYDLGNNAVGLAK                   By MS/MS                 By MS/MS
## YYTFNGPNYNENETIR                By matchin...            By matchin...
## YYTITEVATR                      By matchin...            By matchin...
## YYTVFDRDNNR                     By matchin...            By matchin...
## YYTVFDRDNNRVGFAEAAR             By matchin...            By matchin...
##                       Identification.type.6C_7 Identification.type.6C_8
##                                 <character>              <character>
## AAAAGAGGAGDSGDAVTK                 By MS/MS              By matchin...
```

```
## AAAALAGGK                          By MS/MS                 By MS/MS
## AAAALAGGKK                         By MS/MS                 By MS/MS
## AAADALSDLEIK                    By matchin...                 By MS/MS
## AAADALSDLEIKDSK                    By MS/MS                 By MS/MS
## ...                                   ...                       ...
## YYSIYDLGNNAVGLAK                By matchin...            By matchin...
## YYTFNGPNYNENETIR                By matchin...            By matchin...
## YYTITEVATR                      By matchin...                 By MS/MS
## YYTVFDRDNNR                     By matchin...            By matchin...
## YYTVFDRDNNRVGFAEAAR             By matchin...            By matchin...
##                       Identification.type.6C_9 Identification.type.6D_1
##                                     <character>              <character>
## AAAAGAGGAGDSGDAVTK              By matchin...            By matchin...
## AAAALAGGK                          By MS/MS             By matchin...
## AAAALAGGKK                         By MS/MS             By matchin...
## AAADALSDLEIK                       By MS/MS                 By MS/MS
## AAADALSDLEIKDSK                    By MS/MS                 By MS/MS
## ...                                   ...                       ...
## YYSIYDLGNNAVGLAK                By matchin...            By matchin...
## YYTFNGPNYNENETIR                By matchin...            By matchin...
## YYTITEVATR                         By MS/MS             By matchin...
## YYTVFDRDNNR                     By matchin...            By matchin...
## YYTVFDRDNNRVGFAEAAR             By matchin...            By matchin...
##                       Identification.type.6D_2 Identification.type.6D_3
##                                     <character>              <character>
## AAAAGAGGAGDSGDAVTK              By matchin...            By matchin...
## AAAALAGGK                       By matchin...            By matchin...
## AAAALAGGKK                      By matchin...            By matchin...
## AAADALSDLEIK                    By matchin...            By matchin...
## AAADALSDLEIKDSK                    By MS/MS             By matchin...
## ...                                   ...                       ...
## YYSIYDLGNNAVGLAK                By matchin...            By matchin...
## YYTFNGPNYNENETIR                By matchin...            By matchin...
## YYTITEVATR                         By MS/MS                 By MS/MS
## YYTVFDRDNNR                     By matchin...            By matchin...
## YYTVFDRDNNRVGFAEAAR             By matchin...            By matchin...
##                       Identification.type.6D_4 Identification.type.6D_5
##                                     <character>              <character>
## AAAAGAGGAGDSGDAVTK              By matchin...            By matchin...
## AAAALAGGK                       By matchin...            By matchin...
## AAAALAGGKK                         By MS/MS             By matchin...
## AAADALSDLEIK                       By MS/MS                 By MS/MS
## AAADALSDLEIKDSK                    By MS/MS                 By MS/MS
## ...                                   ...                       ...
## YYSIYDLGNNAVGLAK                   By MS/MS                 By MS/MS
## YYTFNGPNYNENETIR                   By MS/MS                 By MS/MS
## YYTITEVATR                      By matchin...            By matchin...
## YYTVFDRDNNR                     By matchin...            By matchin...
## YYTVFDRDNNRVGFAEAAR             By matchin...            By matchin...
##                       Identification.type.6D_6 Identification.type.6D_7
##                                     <character>              <character>
## AAAAGAGGAGDSGDAVTK                 By MS/MS             By matchin...
## AAAALAGGK                       By matchin...                 By MS/MS
## AAAALAGGKK                      By matchin...                 By MS/MS
```

```
## AAADALSDLEIK                      By MS/MS            By matchin...
## AAADALSDLEIKDSK                By matchin...                 By MS/MS
## ...                                     ...                      ...
## YYSIYDLGNNAVGLAK                  By MS/MS            By matchin...
## YYTFNGPNYNENETIR                  By MS/MS            By matchin...
## YYTITEVATR                     By matchin...            By matchin...
## YYTVFDRDNNR                    By matchin...                 By MS/MS
## YYTVFDRDNNRVGFAEAAR            By matchin...            By matchin...
##                      Identification.type.6D_8 Identification.type.6D_9
##                                   <character>              <character>
## AAAAGAGGAGDSGDAVTK             By matchin...            By matchin...
## AAAALAGGK                         By MS/MS                 By MS/MS
## AAAALAGGKK                        By MS/MS                 By MS/MS
## AAADALSDLEIK                      By MS/MS                 By MS/MS
## AAADALSDLEIKDSK                   By MS/MS                 By MS/MS
## ...                                     ...                      ...
## YYSIYDLGNNAVGLAK               By matchin...            By matchin...
## YYTFNGPNYNENETIR               By matchin...            By matchin...
## YYTITEVATR                        By MS/MS            By matchin...
## YYTVFDRDNNR                       By MS/MS            By matchin...
## YYTVFDRDNNRVGFAEAAR            By matchin...            By matchin...
##                      Identification.type.6E_1 Identification.type.6E_2
##                                   <character>              <character>
## AAAAGAGGAGDSGDAVTK             By matchin...            By matchin...
## AAAALAGGK                      By matchin...            By matchin...
## AAAALAGGKK                     By matchin...            By matchin...
## AAADALSDLEIK                      By MS/MS                 By MS/MS
## AAADALSDLEIKDSK                   By MS/MS                 By MS/MS
## ...                                     ...                      ...
## YYSIYDLGNNAVGLAK               By matchin...            By matchin...
## YYTFNGPNYNENETIR               By matchin...            By matchin...
## YYTITEVATR                     By matchin...            By matchin...
## YYTVFDRDNNR                    By matchin...            By matchin...
## YYTVFDRDNNRVGFAEAAR            By matchin...            By matchin...
##                      Identification.type.6E_3 Identification.type.6E_4
##                                   <character>              <character>
## AAAAGAGGAGDSGDAVTK             By matchin...            By matchin...
## AAAALAGGK                      By matchin...                 By MS/MS
## AAAALAGGKK                     By matchin...            By matchin...
## AAADALSDLEIK                   By matchin...                 By MS/MS
## AAADALSDLEIKDSK                   By MS/MS            By matchin...
## ...                                     ...                      ...
## YYSIYDLGNNAVGLAK               By matchin...                 By MS/MS
## YYTFNGPNYNENETIR               By matchin...                 By MS/MS
## YYTITEVATR                     By matchin...            By matchin...
## YYTVFDRDNNR                    By matchin...                 By MS/MS
## YYTVFDRDNNRVGFAEAAR            By matchin...            By matchin...
##                      Identification.type.6E_5 Identification.type.6E_6
##                                   <character>              <character>
## AAAAGAGGAGDSGDAVTK             By matchin...            By matchin...
## AAAALAGGK                      By matchin...            By matchin...
## AAAALAGGKK                     By matchin...            By matchin...
## AAADALSDLEIK                      By MS/MS            By matchin...
## AAADALSDLEIKDSK                   By MS/MS                 By MS/MS
```

```
## ...                                    ...                    ...
## YYSIYDLGNNAVGLAK                    By MS/MS                By MS/MS
## YYTFNGPNYNENETIR                    By MS/MS                By MS/MS
## YYTITEVATR                       By matchin...           By matchin...
## YYTVFDRDNNR                         By MS/MS             By matchin...
## YYTVFDRDNNRVGFAEAAR              By matchin...                By MS/MS
##                      Identification.type.6E_7 Identification.type.6E_8
##                                     <character>              <character>
## AAAAGAGGAGDSGDAVTK               By matchin...           By matchin...
## AAAALAGGK                           By MS/MS                By MS/MS
## AAAALAGGKK                          By MS/MS                By MS/MS
## AAADALSDLEIK                        By MS/MS                By MS/MS
## AAADALSDLEIKDSK                  By matchin...                By MS/MS
## ...                                    ...                    ...
## YYSIYDLGNNAVGLAK                 By matchin...           By matchin...
## YYTFNGPNYNENETIR                 By matchin...           By matchin...
## YYTITEVATR                       By matchin...           By matchin...
## YYTVFDRDNNR                         By MS/MS                By MS/MS
## YYTVFDRDNNRVGFAEAAR              By matchin...           By matchin...
##                      Identification.type.6E_9 Experiment.6A_1 Experiment.6A_2
##                                     <character>       <integer>       <integer>
## AAAAGAGGAGDSGDAVTK               By matchin...              NA               1
## AAAALAGGK                           By MS/MS                NA               1
## AAAALAGGKK                          By MS/MS                NA               1
## AAADALSDLEIK                        By MS/MS                 1               1
## AAADALSDLEIKDSK                     By MS/MS                 1               1
## ...                                    ...                ...             ...
## YYSIYDLGNNAVGLAK                 By matchin...              NA              NA
## YYTFNGPNYNENETIR                    By MS/MS                NA              NA
## YYTITEVATR                       By matchin...                1              NA
## YYTVFDRDNNR                         By MS/MS                NA              NA
## YYTVFDRDNNRVGFAEAAR              By matchin...              NA              NA
##                      Experiment.6A_3 Experiment.6A_4 Experiment.6A_5
##                            <integer>       <integer>       <integer>
## AAAAGAGGAGDSGDAVTK                NA               1               1
## AAAALAGGK                          2               1               1
## AAAALAGGKK                        NA               1              NA
## AAADALSDLEIK                       1               1               1
## AAADALSDLEIKDSK                   NA               1               1
## ...                               ...             ...             ...
## YYSIYDLGNNAVGLAK                  NA               1               1
## YYTFNGPNYNENETIR                  NA               1               1
## YYTITEVATR                         1              NA              NA
## YYTVFDRDNNR                       NA              NA              NA
## YYTVFDRDNNRVGFAEAAR               NA              NA              NA
##                      Experiment.6A_6 Experiment.6A_7 Experiment.6A_8
##                            <integer>       <integer>       <integer>
## AAAAGAGGAGDSGDAVTK                 1               1               1
## AAAALAGGK                          1               2               1
## AAAALAGGKK                         1               1               1
## AAADALSDLEIK                       1               1               1
## AAADALSDLEIKDSK                    1               1               1
## ...                               ...             ...             ...
## YYSIYDLGNNAVGLAK                   1              NA              NA
```

```
## YYTFNGPNYNENETIR                 1               1              NA
## YYTITEVATR                       1               1              NA
## YYTVFDRDNNR                     NA              NA              NA
## YYTVFDRDNNRVGFAEAAR             NA              NA              NA
##                     Experiment.6A_9 Experiment.6B_1 Experiment.6B_2
##                           <integer>       <integer>       <integer>
## AAAAGAGGAGDSGDAVTK                1              NA              NA
## AAAALAGGK                         1               1               1
## AAAALAGGKK                        1              NA               1
## AAADALSDLEIK                      1               1               1
## AAADALSDLEIKDSK                   1              NA               1
## ...                             ...             ...             ...
## YYSIYDLGNNAVGLAK                 NA              NA              NA
## YYTFNGPNYNENETIR                  1              NA              NA
## YYTITEVATR                       NA               1               1
## YYTVFDRDNNR                      NA              NA              NA
## YYTVFDRDNNRVGFAEAAR              NA              NA              NA
##                     Experiment.6B_3 Experiment.6B_4 Experiment.6B_5
##                           <integer>       <integer>       <integer>
## AAAAGAGGAGDSGDAVTK               NA              NA               1
## AAAALAGGK                         1               2               1
## AAAALAGGKK                        1               1              NA
## AAADALSDLEIK                      1               1               1
## AAADALSDLEIKDSK                  NA               1               1
## ...                             ...             ...             ...
## YYSIYDLGNNAVGLAK                 NA               1               1
## YYTFNGPNYNENETIR                 NA               1               1
## YYTITEVATR                        1               1               1
## YYTVFDRDNNR                      NA              NA              NA
## YYTVFDRDNNRVGFAEAAR              NA              NA              NA
##                     Experiment.6B_6 Experiment.6B_7 Experiment.6B_8
##                           <integer>       <integer>       <integer>
## AAAAGAGGAGDSGDAVTK                1              NA               1
## AAAALAGGK                        NA               2               1
## AAAALAGGKK                       NA               1               1
## AAADALSDLEIK                      1               1               1
## AAADALSDLEIKDSK                   1               1               1
## ...                             ...             ...             ...
## YYSIYDLGNNAVGLAK                  1              NA              NA
## YYTFNGPNYNENETIR                  1               1              NA
## YYTITEVATR                        1              NA               1
## YYTVFDRDNNR                      NA              NA              NA
## YYTVFDRDNNRVGFAEAAR              NA              NA              NA
##                     Experiment.6B_9 Experiment.6C_1 Experiment.6C_2
##                           <integer>       <integer>       <integer>
## AAAAGAGGAGDSGDAVTK                1              NA              NA
## AAAALAGGK                         2              NA               1
## AAAALAGGKK                        1              NA               1
## AAADALSDLEIK                      1               1               1
## AAADALSDLEIKDSK                   1               1               1
## ...                             ...             ...             ...
## YYSIYDLGNNAVGLAK                 NA              NA              NA
## YYTFNGPNYNENETIR                 NA              NA              NA
## YYTITEVATR                       NA               1               1
```

```
## YYTVFDRDNNR                      NA              NA              NA
## YYTVFDRDNNRVGFAEAAR              NA              NA              NA
##                      Experiment.6C_3 Experiment.6C_4 Experiment.6C_5
##                            <integer>       <integer>       <integer>
## AAAAGAGGAGDSGDAVTK               NA               1               1
## AAAALAGGK                         2               2              NA
## AAAALAGGKK                       NA               1              NA
## AAADALSDLEIK                      1               1               1
## AAADALSDLEIKDSK                   1               1               1
## ...                             ...             ...             ...
## YYSIYDLGNNAVGLAK                 NA               1               1
## YYTFNGPNYNENETIR                 NA               1               1
## YYTITEVATR                        1               1              NA
## YYTVFDRDNNR                      NA              NA              NA
## YYTVFDRDNNRVGFAEAAR              NA              NA              NA
##                      Experiment.6C_6 Experiment.6C_7 Experiment.6C_8
##                            <integer>       <integer>       <integer>
## AAAAGAGGAGDSGDAVTK                1               1               1
## AAAALAGGK                        NA               2               1
## AAAALAGGKK                       NA               1               1
## AAADALSDLEIK                      1               1               1
## AAADALSDLEIKDSK                   1               1               1
## ...                             ...             ...             ...
## YYSIYDLGNNAVGLAK                  1              NA              NA
## YYTFNGPNYNENETIR                  1               1               1
## YYTITEVATR                        1              NA               1
## YYTVFDRDNNR                       1              NA               1
## YYTVFDRDNNRVGFAEAAR              NA              NA              NA
##                      Experiment.6C_9 Experiment.6D_1 Experiment.6D_2
##                            <integer>       <integer>       <integer>
## AAAAGAGGAGDSGDAVTK                1              NA              NA
## AAAALAGGK                         1              NA               1
## AAAALAGGKK                        1              NA              NA
## AAADALSDLEIK                      1               1               1
## AAADALSDLEIKDSK                   1               1               1
## ...                             ...             ...             ...
## YYSIYDLGNNAVGLAK                 NA              NA              NA
## YYTFNGPNYNENETIR                  1              NA              NA
## YYTITEVATR                        1              NA               1
## YYTVFDRDNNR                      NA              NA              NA
## YYTVFDRDNNRVGFAEAAR              NA              NA              NA
##                      Experiment.6D_3 Experiment.6D_4 Experiment.6D_5
##                            <integer>       <integer>       <integer>
## AAAAGAGGAGDSGDAVTK               NA               1               1
## AAAALAGGK                         1               1               1
## AAAALAGGKK                       NA               1              NA
## AAADALSDLEIK                      1               1               1
## AAADALSDLEIKDSK                   1               1               1
## ...                             ...             ...             ...
## YYSIYDLGNNAVGLAK                 NA               1               1
## YYTFNGPNYNENETIR                 NA               1               1
## YYTITEVATR                        1               1               1
## YYTVFDRDNNR                      NA               1               1
## YYTVFDRDNNRVGFAEAAR              NA               1              NA
```

```
##                     Experiment.6D_6 Experiment.6D_7 Experiment.6D_8
##                           <integer>       <integer>       <integer>
## AAAAGAGGAGDSGDAVTK                1               1              NA
## AAAALAGGK                        NA               2               1
## AAAALAGGKK                       NA               1               1
## AAADALSDLEIK                      1               1               1
## AAADALSDLEIKDSK                   1               1               1
## ...                             ...             ...             ...
## YYSIYDLGNNAVGLAK                  1               1              NA
## YYTFNGPNYNENETIR                  1               1               1
## YYTITEVATR                        1              NA               1
## YYTVFDRDNNR                       1               1               1
## YYTVFDRDNNRVGFAEAAR              NA              NA              NA
##                     Experiment.6D_9 Experiment.6E_1 Experiment.6E_2
##                           <integer>       <integer>       <integer>
## AAAAGAGGAGDSGDAVTK               NA              NA               1
## AAAALAGGK                         2              NA               1
## AAAALAGGKK                        1              NA              NA
## AAADALSDLEIK                      1               1               1
## AAADALSDLEIKDSK                   1               1               1
## ...                             ...             ...             ...
## YYSIYDLGNNAVGLAK                 NA              NA              NA
## YYTFNGPNYNENETIR                  1              NA              NA
## YYTITEVATR                       NA              NA               1
## YYTVFDRDNNR                       1               1              NA
## YYTVFDRDNNRVGFAEAAR              NA              NA              NA
##                     Experiment.6E_3 Experiment.6E_4 Experiment.6E_5
##                           <integer>       <integer>       <integer>
## AAAAGAGGAGDSGDAVTK               NA              NA               1
## AAAALAGGK                         2               2               1
## AAAALAGGKK                       NA               1              NA
## AAADALSDLEIK                      1               1               1
## AAADALSDLEIKDSK                   1               1               1
## ...                             ...             ...             ...
## YYSIYDLGNNAVGLAK                  1               1               1
## YYTFNGPNYNENETIR                 NA               1               1
## YYTITEVATR                        1               1               1
## YYTVFDRDNNR                       1               1               1
## YYTVFDRDNNRVGFAEAAR              NA               1               1
##                     Experiment.6E_6 Experiment.6E_7 Experiment.6E_8
##                           <integer>       <integer>       <integer>
## AAAAGAGGAGDSGDAVTK                1              NA              NA
## AAAALAGGK                        NA               2               2
## AAAALAGGKK                       NA               1               1
## AAADALSDLEIK                      1               1               1
## AAADALSDLEIKDSK                   1              NA               1
## ...                             ...             ...             ...
## YYSIYDLGNNAVGLAK                  1              NA              NA
## YYTFNGPNYNENETIR                  1               1               1
## YYTITEVATR                       NA              NA              NA
## YYTVFDRDNNR                       1               1               1
## YYTVFDRDNNRVGFAEAAR               1               1               1
##                     Experiment.6E_9 Intensity     Reverse Potential.contaminant
##                           <integer> <numeric> <character>           <character>
```

```
## AAAAGAGGAGDSGDAVTK               NA   1190800
## AAAALAGGK                         1 280990000
## AAAALAGGKK                        1  33360000
## AAADALSDLEIK                      1  54622000
## AAADALSDLEIKDSK                   1  18910000
## ...                             ...       ...         ...                   ...
## YYSIYDLGNNAVGLAK                 NA   2145900
## YYTFNGPNYNENETIR                  1   5608800
## YYTITEVATR                       NA  13034000
## YYTVFDRDNNR                       1   8702500
## YYTVFDRDNNRVGFAEAAR               1   2391100
##                           id Protein.group.IDs Mod..peptide.IDs  Evidence.IDs
##                    <integer>       <character>      <character>   <character>
## AAAAGAGGAGDSGDAVTK         0               859                0 0;1;2;3;4;...
## AAAALAGGK                  1               230                1 24;25;26;2...
## AAAALAGGKK                 2               230                2 74;75;76;7...
## AAADALSDLEIK               3               229                3 99;100;101...
## AAADALSDLEIKDSK            4               229                4 144;145;14...
## ...                      ...               ...              ...           ...
## YYSIYDLGNNAVGLAK       11461               196            12240 331367;331...
## YYTFNGPNYNENETIR       11462              1254            12241 331384;331...
## YYTITEVATR             11463               854            12242 331411;331...
## YYTVFDRDNNR            11464                34            12243 331439;331...
## YYTVFDRDNNRVGFAEAAR    11465                34            12244 331455;331...
##                       MS.MS.IDs Best.MS.MS Oxidation..M..site.IDs MS.MS.Count
##                     <character>  <integer>            <character>   <integer>
## AAAAGAGGAGDSGDAVTK  0;1;2;3;4;...          0                                10
## AAAALAGGK           10;11;12;1...         21                                18
## AAAALAGGKK          30;31;32;3...         31                                21
## AAADALSDLEIK        51;52;53;5...         72                                29
## AAADALSDLEIKDSK     85;86;87;8...         94                                32
## ...                           ...        ...                   ...         ...
## YYSIYDLGNNAVGLAK    169138;169...     169147                                13
## YYTFNGPNYNENETIR    169151;169...     169159                                14
## YYTITEVATR          169165;169...     169173                                12
## YYTVFDRDNNR         169177;169...     169180                                 7
## YYTVFDRDNNRVGFAEAAR        169184     169184                                 1
```

- 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
```

1.4 Preprocessing

1.4.1 Log-transform

Click to see code to log-transfrom the data

- We calculate how many non zero intensities we have for each peptide and this can be useful for filtering.

```
rowData(pe[["peptideRaw"]])$nNonZero <- rowSums(assay(pe[["peptideRaw"]]) > 0)
```

- Peptides with zero intensities are missing peptides and should be represent with a `NA` value rather than `0`.

```
pe <- zeroIsNA(pe, "peptideRaw") # convert 0 to NA
```

- Logtransform data with base 2

```
pe <- logTransform(pe, base = 2, i = "peptideRaw", name = "peptideLog")
```

1.4.2 Filtering

Click to see code to filter the data

1. Handling overlapping protein groups

In our approach a peptide can map to multiple proteins, as long as there is none of these proteins present in a smaller subgroup.

```
pe[["peptideLog"]] <-
 pe[["peptideLog"]][rowData(pe[["peptideLog"]])$Proteins
 %in% smallestUniqueGroups(rowData(pe[["peptideLog"]])$Proteins),]
```

2. Remove reverse sequences (decoys) and contaminants

We now remove the contaminants, peptides that map to decoy sequences, and proteins which were only identified by peptides with modifications.

```
pe[["peptideLog"]] <- pe[["peptideLog"]][rowData(pe[["peptideLog"]])$Reverse != "+", ]
pe[["peptideLog"]] <- pe[["peptideLog"]][rowData(pe[["peptideLog"]])$
    Potential.contaminant != "+", ]
```

3. Drop peptides that were only identified in one sample

We keep peptides that were observed at last twice.

```
pe[["peptideLog"]] <- pe[["peptideLog"]][rowData(pe[["peptideLog"]])$nNonZero >= 2, ]
nrow(pe[["peptideLog"]])
```

```
## [1] 10478
```

We keep 10478 peptides upon filtering.

1.5 Normalization

Click to see R-code to normalize the data

```
pe <- normalize(pe,
                i = "peptideLog",
                name = "peptideNorm",
                method = "center.median")
```

---

2 Peptide-level models

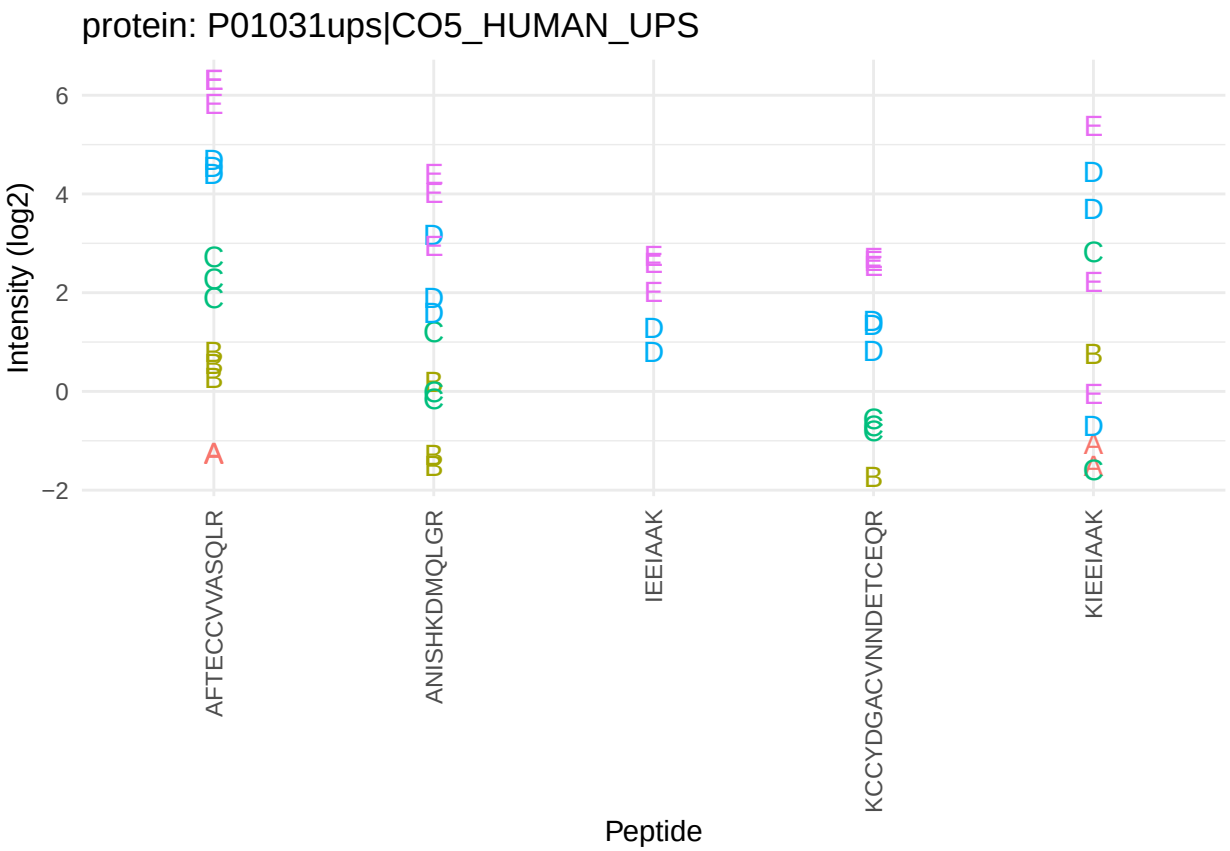
2.1 Summarization

Click to see code to make plot

```
prot <- "P01031ups|CO5_HUMAN_UPS"
data <- pe[["peptideNorm"]][
  rowData(pe[["peptideNorm"]])$Proteins == prot,
  colData(pe)$lab=="lab3"] %>%
  assay %>%
  as.data.frame %>%
  rownames_to_column(var = "peptide") %>%
  gather(sample, intensity, -peptide) %>%
  mutate(condition = colData(pe)[sample,"condition"]) %>%
  na.exclude
sumPlot <- data %>%
  ggplot(aes(x = peptide, y = intensity, color = condition, group = sample, label = condition), show.leg
  geom_text(show.legend = FALSE) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  xlab("Peptide") +
  ylab("Intensity (log2)") +
  ggtitle(paste0("protein: ",prot))
```

Here, we will focus on the summarization of the intensities for protein P01031ups|CO5_HUMAN_UPS.

```
sumPlot
```



2.1.1 Median summarization

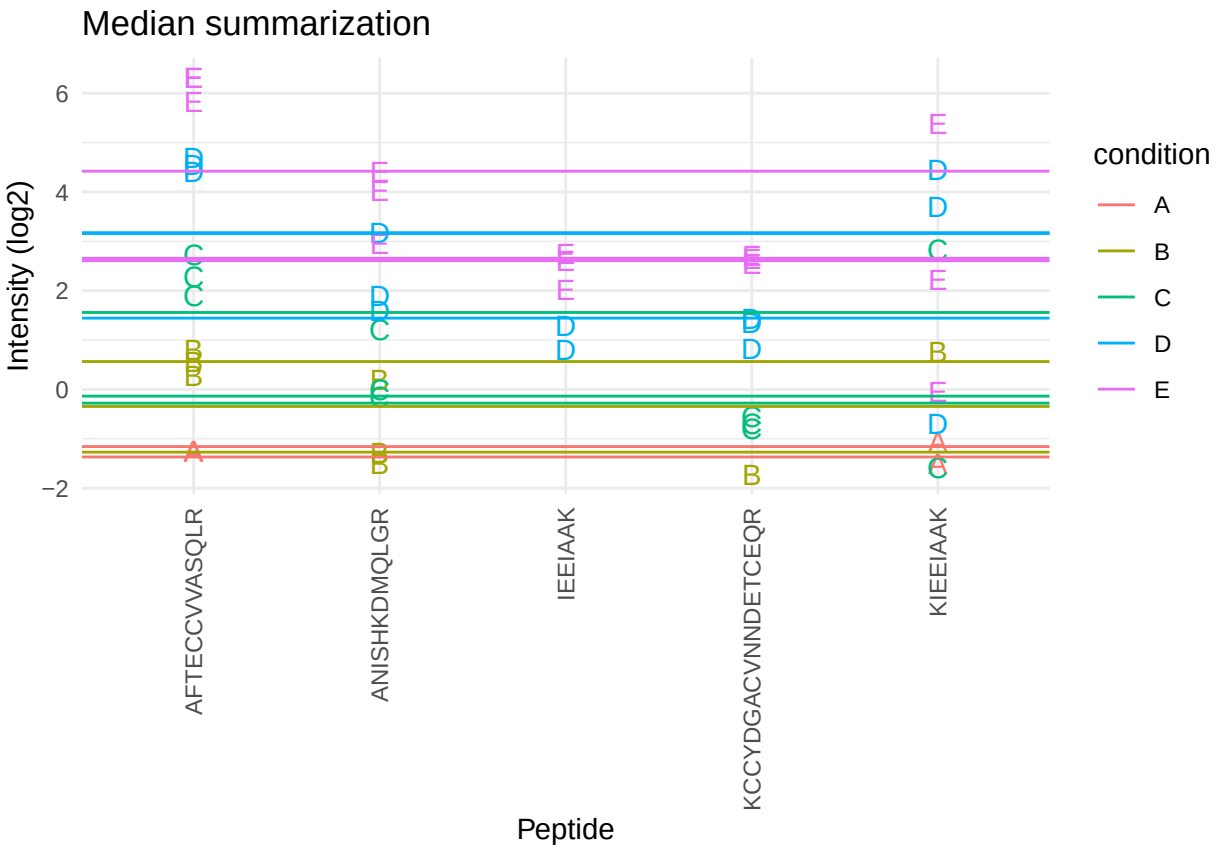
We first evaluate median summarization for protein P01031ups|CO5_HUMAN_UPS.

Click to see code to make plot

```
dataHlp <- pe[["peptideNorm"]][
    rowData(pe[["peptideNorm"]])$Proteins == prot,
    colData(pe)$lab=="lab3"] %>% assay

sumMedian <- data.frame(
  intensity= dataHlp
    %>% colMedians(na.rm=TRUE)
  ,
  condition= colnames(dataHlp) %>% substr(12,12) %>% as.factor )

sumMedianPlot <- sumPlot +
  geom_hline(
    data = sumMedian,
    mapping = aes(yintercept=intensity,color=condition)) +
  ggtitle("Median summarization")
```

```
sumMedianPlot
```

```
## Warning: Removed 1 rows containing missing values (geom_hline).
```



- The sample medians are not a good estimate for the protein expression value.
- Indeed, they do not account for differences in peptide effects
- Peptides that ionize poorly are also picked up in samples with high spike-in concencentration and not in samples with low spike-in concentration
- This introduces a bias.

2.1.2 Model based summarization

We can use a linear peptide-level model to estimate the protein expression value while correcting for the peptide effect, i.e.

$$y_{ip} = \beta_i^{\text{sample}} + \beta_p^{peptide} + \epsilon_{ip}$$

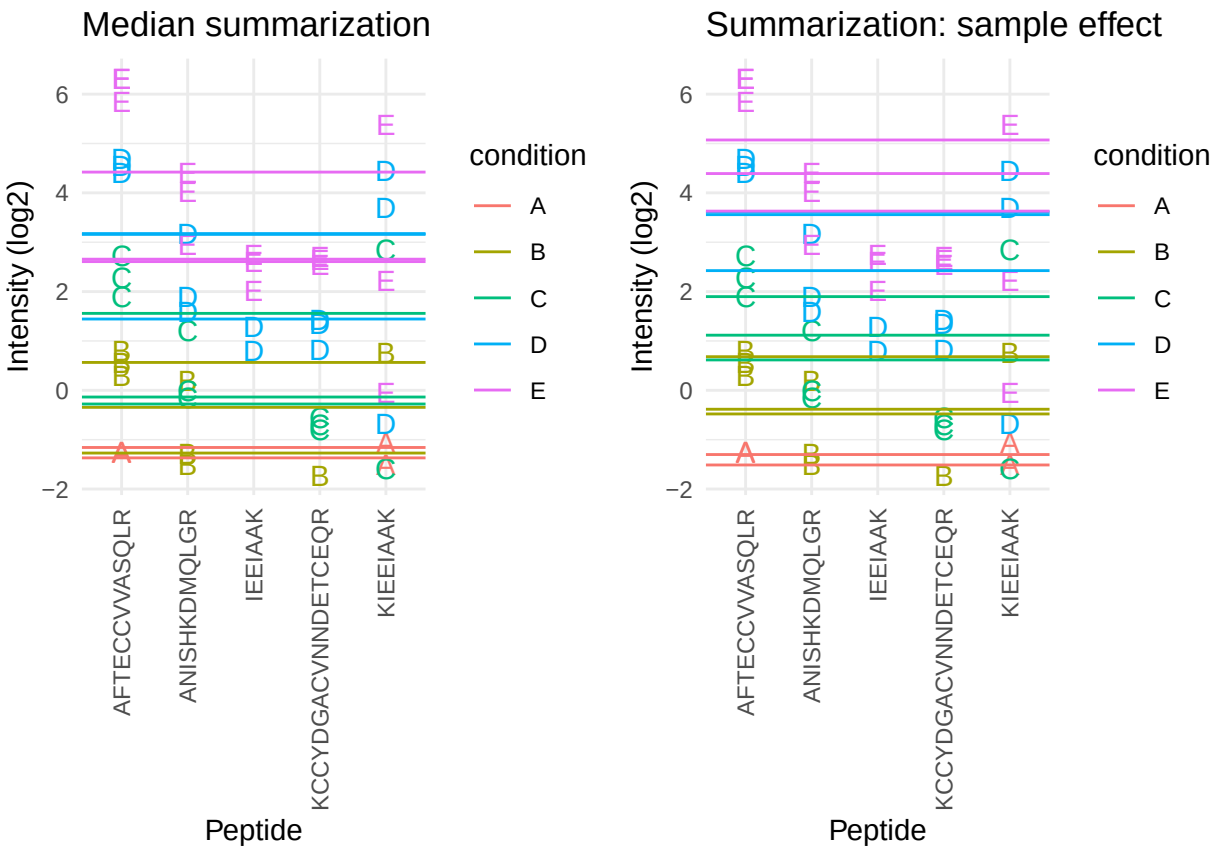
Click to see code to make plot

```
sumMeanPepMod <- lm(intensity ~ -1 + sample + peptide,data)

sumMeanPep <- data.frame(
  intensity=sumMeanPepMod$coef[grep("sample",names(sumMeanPepMod$coef))] + mean(data$intensity) - mean(s
  condition= names(sumMeanPepMod$coef)[grep("sample",names(sumMeanPepMod$coef))] %>% substr(18,18) %>% a


fitLmPlot <-  sumPlot + geom_line(
    data = data %>% mutate(fit=sumMeanPepMod$fitted.values),
```

```
    mapping = aes(x=peptide, y=fit,color=condition, group=sample)) +
    ggtitle("fit: ~ sample + peptide")
sumLmPlot <- sumPlot + geom_hline(
    data = sumMeanPep,
    mapping = aes(yintercept=intensity,color=condition)) +
    ggtitle("Summarization: sample effect")
```

```
grid.arrange(sumMedianPlot, sumLmPlot, ncol=2)
```

```
## Warning: Removed 1 rows containing missing values (geom_hline).
```



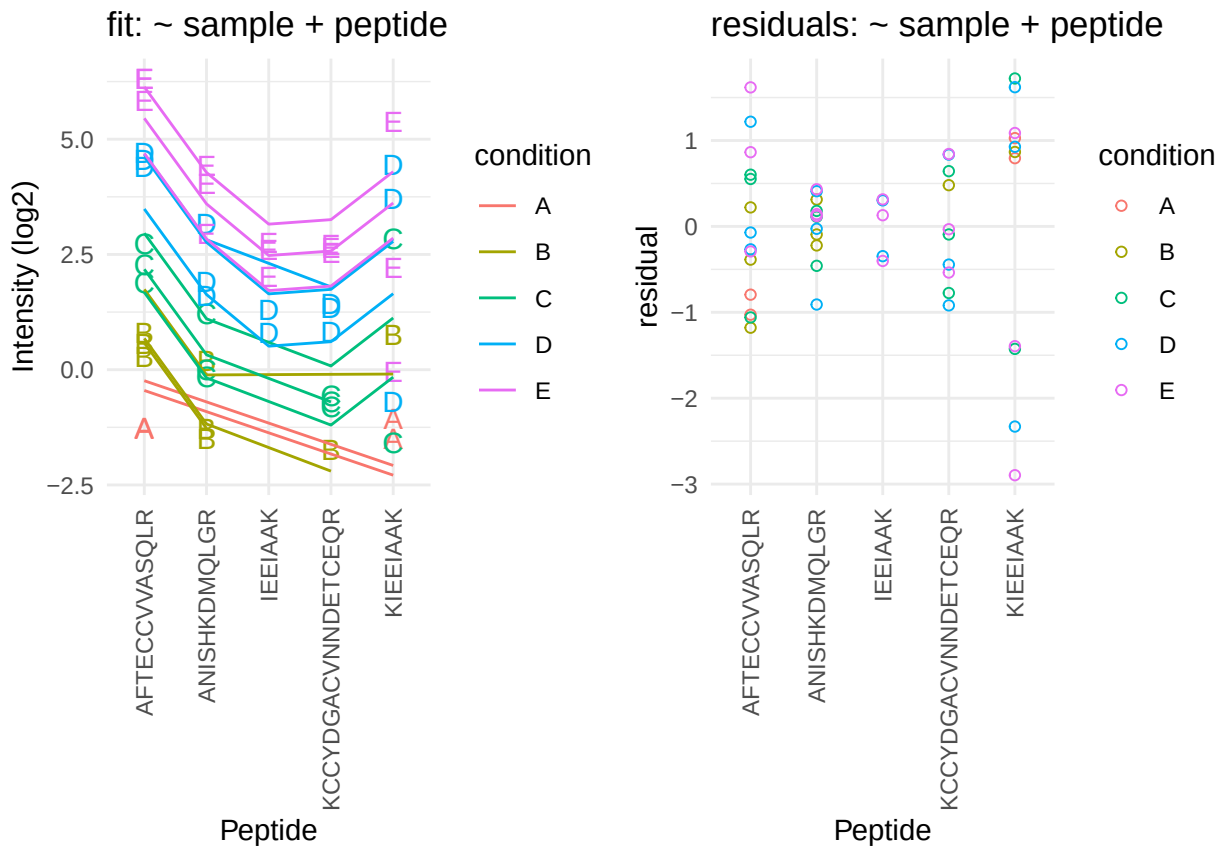
- By correcting for the peptide species the protein expression values are much better separated an better reflect differences in abundance induced by the spike-in condition.
- Indeed, it shows that median summarisation that does not account for the peptide effect indeed over-estimated the protein expression value in the small spike-in conditions and underestimated that in the large spike-in conditions.
- Still there seem to be some issues with samples that for which the expression values are not well separated according to the spike-in condition.

A residual analysis clearly indicates potential issues:

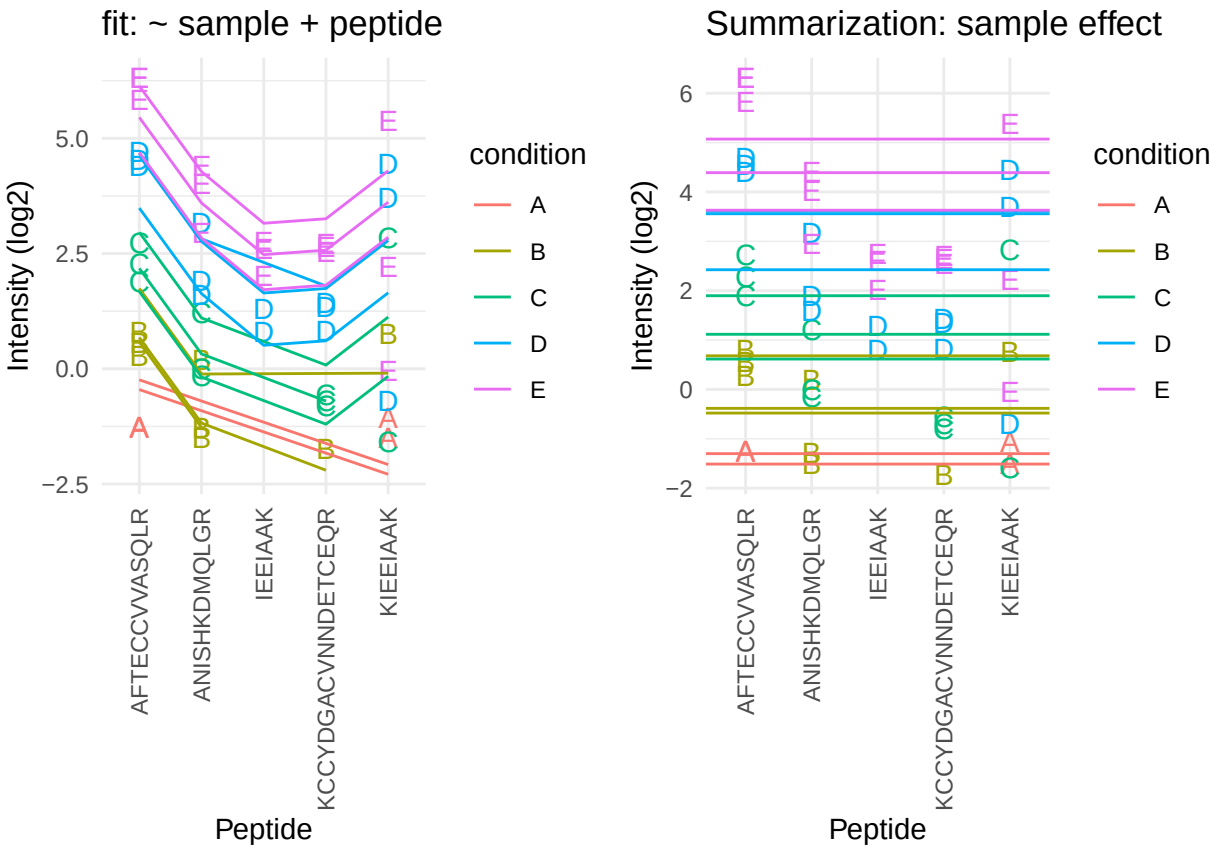
Click to see code to make plot

```
resPlot <- data %>%
  mutate(res=sumMeanPepMod$residuals) %>%
  ggplot(aes(x = peptide, y = res, color = condition, label = condition), show.legend = FALSE) +
  geom_point(shape=21) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  xlab("Peptide") +
  ylab("residual") +
  ggtitle("residuals: ~ sample + peptide")
```

```
grid.arrange(fitLmPlot, resPlot, nrow = 1)
```




```
grid.arrange(fitLmPlot, sumLmPlot, nrow = 1)
```



- The residual plot shows some large outliers for peptide KIEEIAAK.
- Indeed, in the original plot the intensities for this peptide do not seem to line up very well with the concentration.
- This induces a bias in the summarization for some of the samples (e.g. for D and E)

2.1.3 Robust summarization using a peptide-level linear model

$$y_{ip} = \beta_i^{\text{sample}} + \beta_p^{peptide} + \epsilon_{ip}$$

- Ordinary least squares: estimate β that minimizes

$$\text{OLS} : \sum_{i,p} \epsilon_{ip}^2 = \sum_{i,p} (y_{ip} - \beta_i^{\text{sample}} - \beta_p^{\text{peptide}})^2$$

We replace OLS by M-estimation with loss function

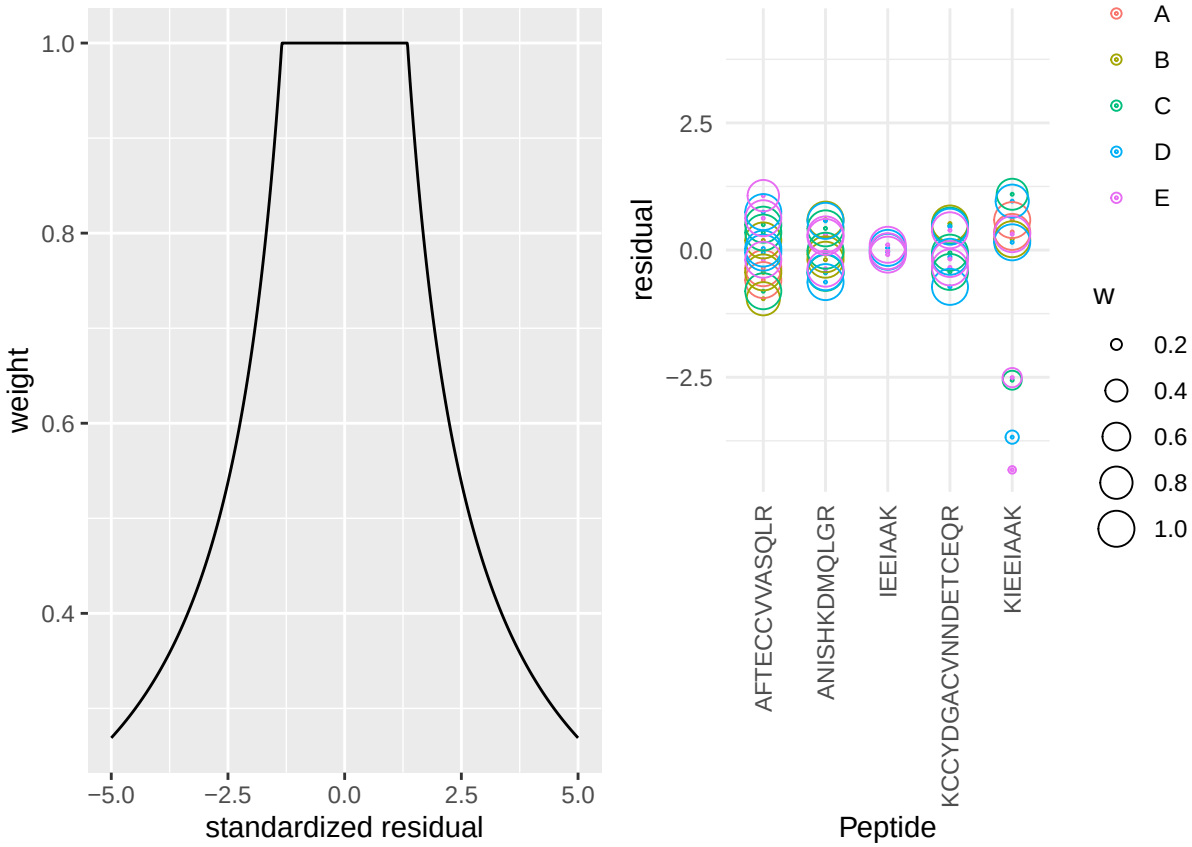
$$\text{OLS} : \sum_{i,p} w_{ip}\epsilon_{ip}^2 = \sum_{i,p} w_{ip}(y_{ip} - \beta_i^{\text{sample}} - \beta_p^{\text{peptide}})^2$$

- Iteratively fit model with observation weights w_{ip} until convergence
- The weights are calculated based on standardized residuals

Click to see code to make plot

```
sumMeanPepRobMod <- MASS::rlm(intensity ~ -1 + sample + peptide,data)
resRobPlot <- data %>%
  mutate(res = sumMeanPepRobMod$residuals,
         w = sumMeanPepRobMod$w) %>%
  ggplot(aes(x = peptide, y = res, color = condition, label = condition,size=w), show.legend = FALSE) +
  geom_point(shape=21,size=.2) +
  geom_point(shape=21) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  xlab("Peptide") +
  ylab("residual") +
  ylim(c(-1,1)*max(abs(sumMeanPepRobMod$residuals)))
weightPlot <- qplot(
  seq(-5,5,.01),
  MASS::psi.huber(seq(-5,5,.01)),
  geom="path") +
  xlab("standardized residual") +
  ylab("weight")
```

```
grid.arrange(weightPlot,resRobPlot,nrow=1)
```

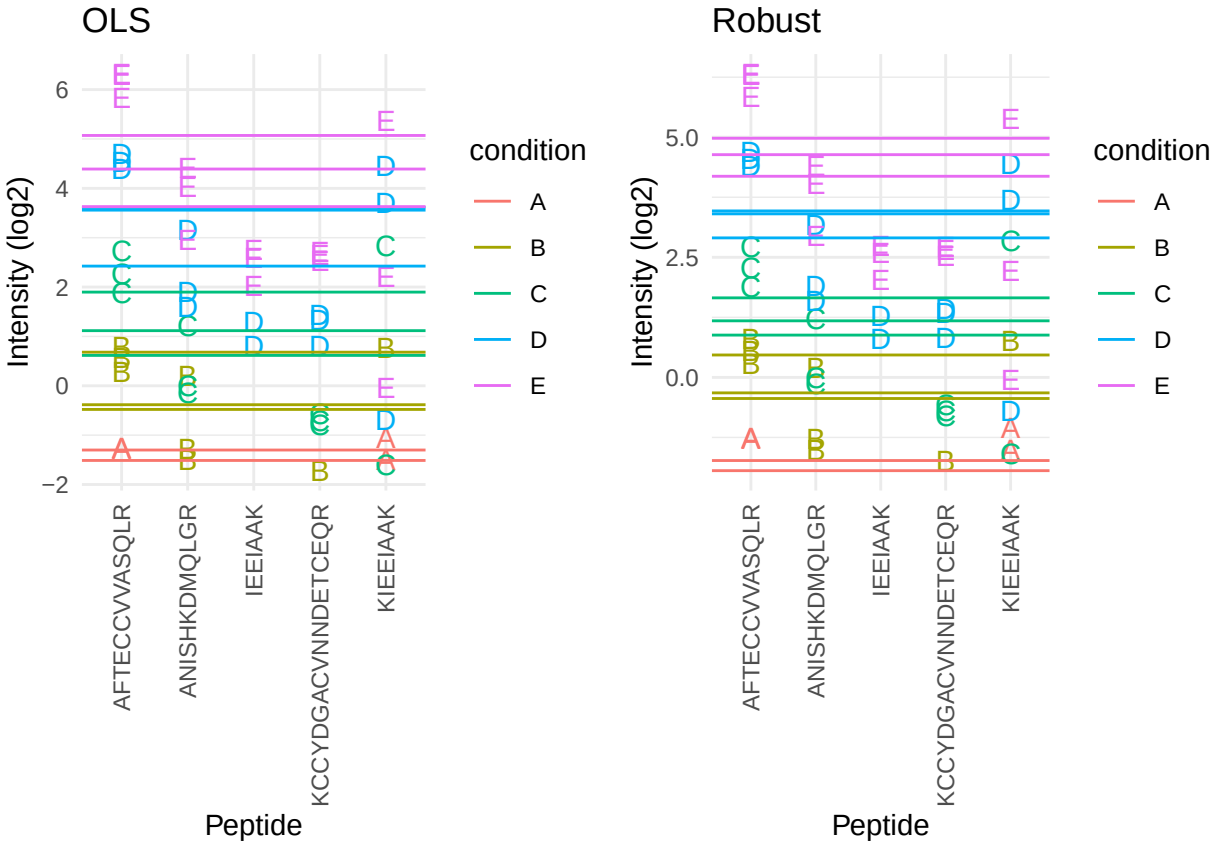


- We clearly see that the weights in the M-estimation procedure will down-weight errors associated with outliers for peptide KIEEIAAK.

Click to see code to make plot

```
sumMeanPepRob <- data.frame(
  intensity=sumMeanPepRobMod$coef[grep("sample",names(sumMeanPepRobMod$coef))] + mean(data$intensity) -
  condition= names(sumMeanPepRobMod$coef)[grep("sample",names(sumMeanPepRobMod$coef))] %>% substr(18,18)

sumRlmPlot <- sumPlot + geom_hline(
    data=sumMeanPepRob,
    mapping=aes(yintercept=intensity,color=condition)) +
    ggtitle("Robust")
```

```
grid.arrange(sumLmPlot + ggtitle("OLS"), sumRlmPlot, nrow = 1)
```



- Robust regresion results in a better separation between the protein expression values for the different samples according to their spike-in concentration.

2.2 Estimation of differential abundance using peptide level model

- Instead of summarising the data we can also directly model the data at the peptide-level.
- But, we will have to address the pseudo-replication.

$$y_{iclp} = \beta_0 + \beta_c^{\text{condition}} + \beta_l^{\text{lab}} + \beta_p^{\text{peptide}} + u_s^{\text{sample}} + \epsilon_{iclp}$$

- protein-level
  - $\beta_c^{\text{condition}}$: spike-in condition

  - $\beta_c^{\text{condition}}$: lab effect
  - $u_r^{\text{run}} \sim N\left(0, \sigma^2_{\text{run}}\right) \rightarrow$ random effect addresses pseudo-replication

- peptide-level

  - β_p^{peptide}: peptide effect
  - $\epsilon_{rp} \sim N\left(0, \sigma^2_{\epsilon}\right)$ within sample (run) error

- DA estimates:

$$\log_2 FC_{B-A} = \beta_B^{\text{condition}}$$

$$\log_2 FC_{C-B} = \beta_C^{\text{condition}} - \beta_B^{\text{condition}}$$

- Mixed peptide-level models are implemented in msqrob2

- It has the advantages that

  1. it correctly addresses the difference levels of variability in the data
  2. it avoids summarisation and therefore also accounts for the difference in the number of peptides that are observed in each sample
  3. more powerful analysis

- It has the disadvantage that

  1. protein summaries are no longer available for plotting
  2. it is difficult to correctly specify the degrees of freedom for the test-statistic leading to inference that is too liberal in experiments with small sample size
  3. sometimes sample level random effect variance are estimated to be zero, then the pseudo-replication is not addressed leading to inference that is too liberal for these specific proteins
  4. they are much more difficult to disseminate to users with limited background in statistics

Hence, for this course we opted to use peptide-level models for summarization, but not for directly inferring on the differential expression at the protein-level.