Introduction to proteomics data analysis: maxLFQ summarization

Lieven Clement

statOmics, Ghent University (https://statomics.github.io)

Contents

1 Background **1**

2 Data **2**

3 Data **2**

3.1 Data exploration . . . 3

4 Preprocessing **3**

4.1 Log transform the data . . . 4

4.2 Filtering . . . 4

4.3 Normalize the data using median centering . . . 4

4.4 Explore normalized data . . . 4

5 Data Analysis **6**

5.1 Estimation . . . 6

5.2 Inference . . . 6

5.3 Plots . . . 7

6 Session Info **10**

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

1 Background

This case-study is a subset of the data of the 6th study of the Clinical Proteomic Technology Assessment for Cancer (CPTAC). In this experiment, the authors spiked the Sigma Universal Protein Standard mixture 1 (UPS1) containing 48 different human proteins in a protein background of 60 ng/μL Saccharomyces cerevisiae strain BY4741. Two different spike-in concentrations were used: 6A (0.25 fmol UPS1 proteins/μL) and 6B (0.74 fmol UPS1 proteins/μL) [5]. We limited ourselves to the data of LTQ-Orbitrap W at site 56. The data were searched with MaxQuant version 1.5.2.8, and detailed search settings were described in Goeminne et al. (2016) [1]. Three replicates are available for each concentration.

- NOTE THAT maxLFQ SUMMARISATION IS SUBOPTIMAL!
- THIS IS FOR DIDACTICAL PURPOSES ONLY.

2 Data

We first import the data from proteinGroups.txt file. This is the file containing maxLFQ summarized protein-level intensities. For a MaxQuant search [6], this proteinGroups.txt file can be found by default in the "path_to_raw_files/combined/txt/" folder from the MaxQuant output, with "path_to_raw_files" the folder where the raw files were saved. In this vignette, we use a MaxQuant proteinRaws file which is a subset of the cptac study. To import the data we use the `QFeatures` package.

We generate the object proteinRawFile with the path to the proteinGroups.txt file. Using the `grepEcols` function, we find the columns that contain the LFQ expression data of the proteinRaws in the proteinGroups.txt file.

3 Data

We first import the data from proteinGroups.txt file. This is the file containing maxLFQ summarized protein-level intensities. For a MaxQuant search [6], this proteinGroups.txt file can be found by default in the "path_to_raw_files/combined/txt/" folder from the MaxQuant output, with "path_to_raw_files" the folder where the raw files were saved. In this vignette, we use a MaxQuant proteinRaws file which is a subset of the cptac study. To import the data we use the QFeatures package.

We generate the object proteinRawFile with the path to the proteinGroups.txt file. Using the grepEcols function, we find the columns that contain the LFQ expression data of the proteinRaws in the proteinGroups.txt file.

```
library(tidyverse)
library(limma)
library(QFeatures)
library(msqrob2)
library(plotly)
proteinsFile <- "https://raw.githubusercontent.com/statOmics/PDA22GTPB/data/quantification/cptacAvsB_la

ecols <- grep("LFQ\\.intensity\\.", names(read.delim(proteinsFile)))
```

Next, we read the data and store it in QFeatures object

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

The QFeatures object pe currently contains a single assay, named proteinRaw.

We extract the column names from the proteinRaw assay and see that this contains information about the spike-in condition.

```
colnames(pe[["proteinRaw"]])
```

```
## [1] "LFQ.intensity.6A_7" "LFQ.intensity.6A_8" "LFQ.intensity.6A_9"
## [4] "LFQ.intensity.6B_7" "LFQ.intensity.6B_8" "LFQ.intensity.6B_9"
```

We rename the colnames by dropping the "LFQ.intensity." from the name.

```
(newNames <-  sub(
  pattern = "LFQ\\.intensity\\.",
  replacement = "",
  colnames(pe[["proteinRaw"]]))
 )
```

```
## [1] "6A_7" "6A_8" "6A_9" "6B_7" "6B_8" "6B_9"
```

```
pe <- renameColname(pe,
                    i = "proteinRaw",
                    newNames)
pe <- renamePrimary(pe, newNames)
colnames(pe[["proteinRaw"]])
```

```
## [1] "6A_7" "6A_8" "6A_9" "6B_7" "6B_8" "6B_9"
```

In the following code chunk, we add the spikein condition that we can read in the raw file name to the colData.

```
colData(pe)$condition <-
  colnames(pe[["proteinRaw"]]) %>%
  substr(start = 2, stop = 2) %>%
  as.factor
colData(pe)$condition
```

```
## [1] A A A B B B
## Levels: A B
```

We calculate how many non zero intensities we have per protein and this will be useful for filtering.

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

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

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

3.1 Data exploration

45% of all peptide intensities are missing and for some proteins we do not even measure a signal in any sample.

4 Preprocessing

This section preforms preprocessing for the peptide data. This include

- log transformation,
- filtering

4.1 Log transform the data

```
pe <- logTransform(pe, base = 2, i = "proteinRaw", name = "proteinLog")
```

4.2 Filtering

1. Remove reverse sequences (decoys) and contaminants

We now remove the contaminants and proteins that map to decoys.

```
pe <- filterFeatures(pe,~ Reverse != "+")
pe <- filterFeatures(pe,~ Potential.contaminant != "+")
```

We keep 1537 peptides upon filtering.

4.3 Normalize the data using median centering

We normalize the data by substracting the sample median from every intensity for peptide p in a sample i:

$$y_{ip}^{\text{norm}} = y_{ip} - \hat{\mu}_i$$

with $\hat{\mu}_i$ the median intensity over all observed peptides in sample i.

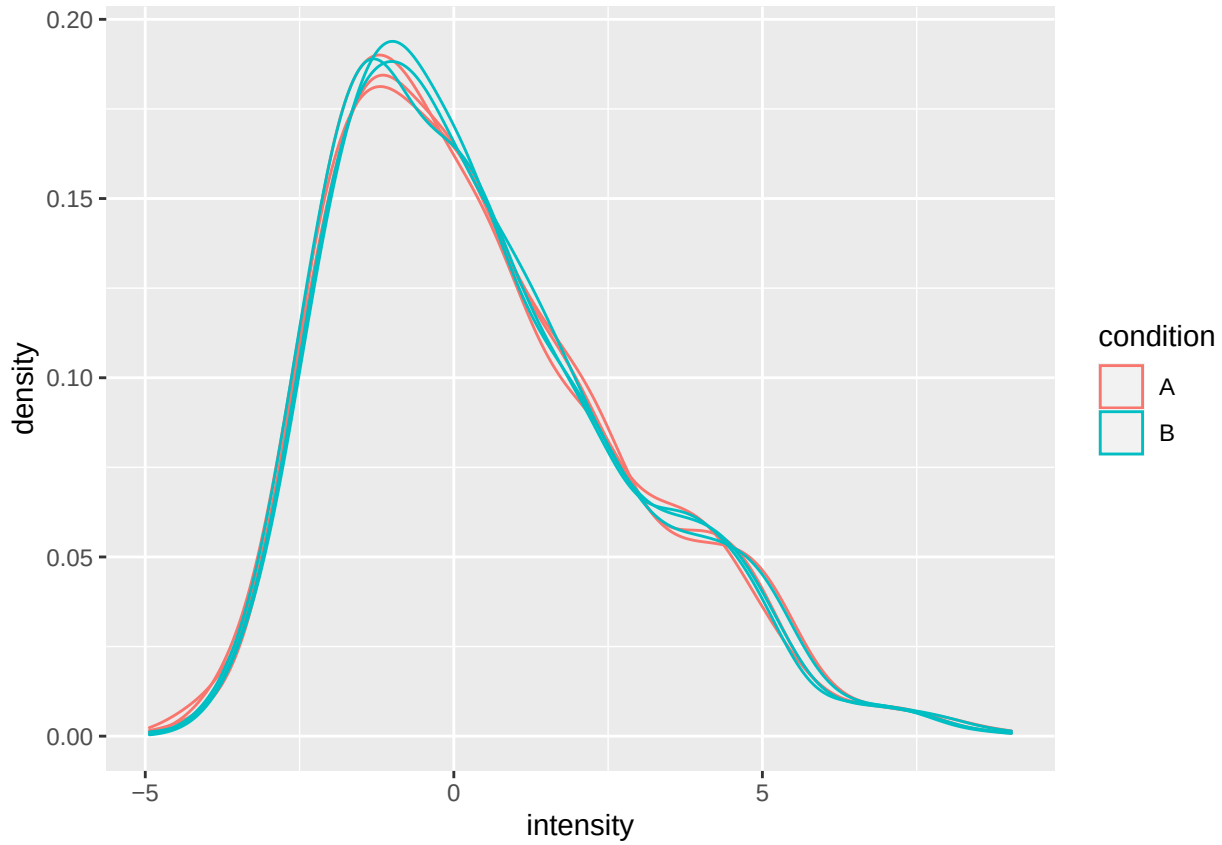
```
pe <- normalize(pe,
                i = "proteinLog",
                name = "protein",
                method = "center.median")
```

4.4 Explore normalized data

Upon the normalisation the density curves are nicely registered

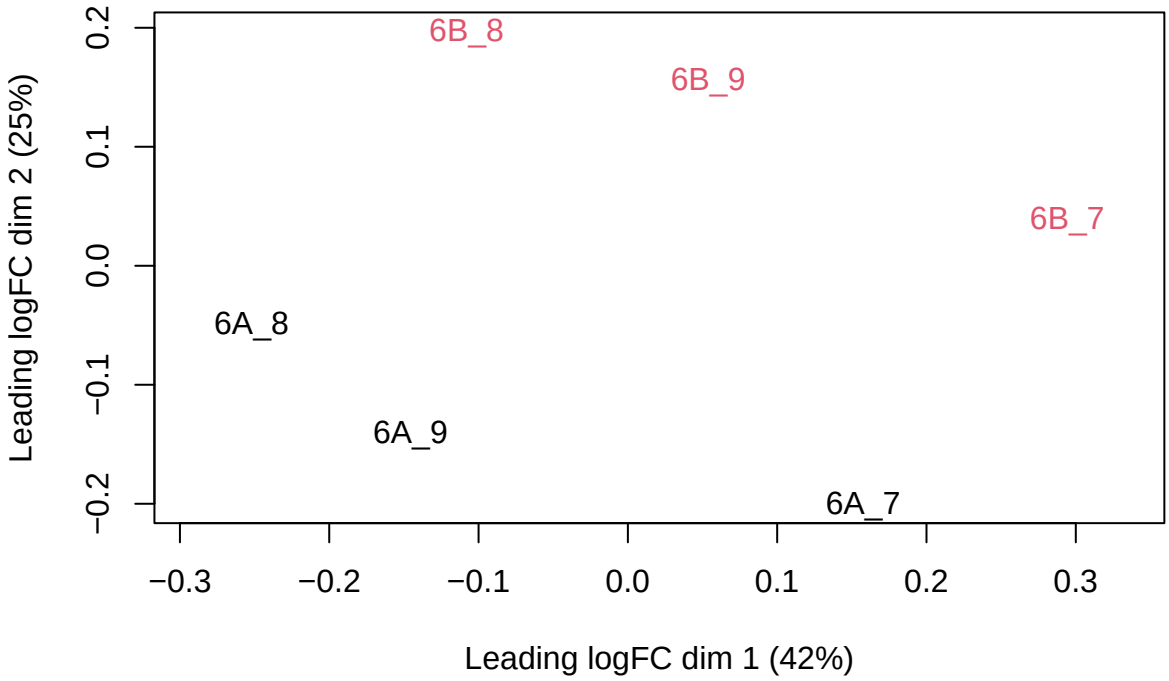
```
pe[["protein"]] %>%
  assay %>%
  as.data.frame() %>%
  gather(sample, intensity) %>%
  mutate(condition = colData(pe)[sample,"condition"]) %>%
  ggplot(aes(x = intensity,group = sample,color = condition)) +
    geom_density()
```

```
## Warning: Removed 4052 rows containing non-finite values (stat_density).
```



We can visualize our data using a Multi Dimensional Scaling plot, eg. as provided by the `limma` package.

```
pe[["protein"]] %>%
  assay %>%
  limma::plotMDS(col = as.numeric(colData(pe)$condition))
```



Note that the samples show a clear separation according to the spike-in condition in the second dimension of the MDS plot.

5 Data Analysis

5.1 Estimation

We model the protein level expression values using `msqrob`. By default `msqrob2` estimates the model parameters using robust regression.

We will model the data with a different group mean. The group is incoded in the variable `condition` of the colData. We can specify this model by using a formula with the factor condition as its predictor: `formula = ~condition`.

Note, that a formula always starts with a symbol '~'.

```
pe <- msqrob(object = pe, i = "protein", formula = ~condition)
```

5.2 Inference

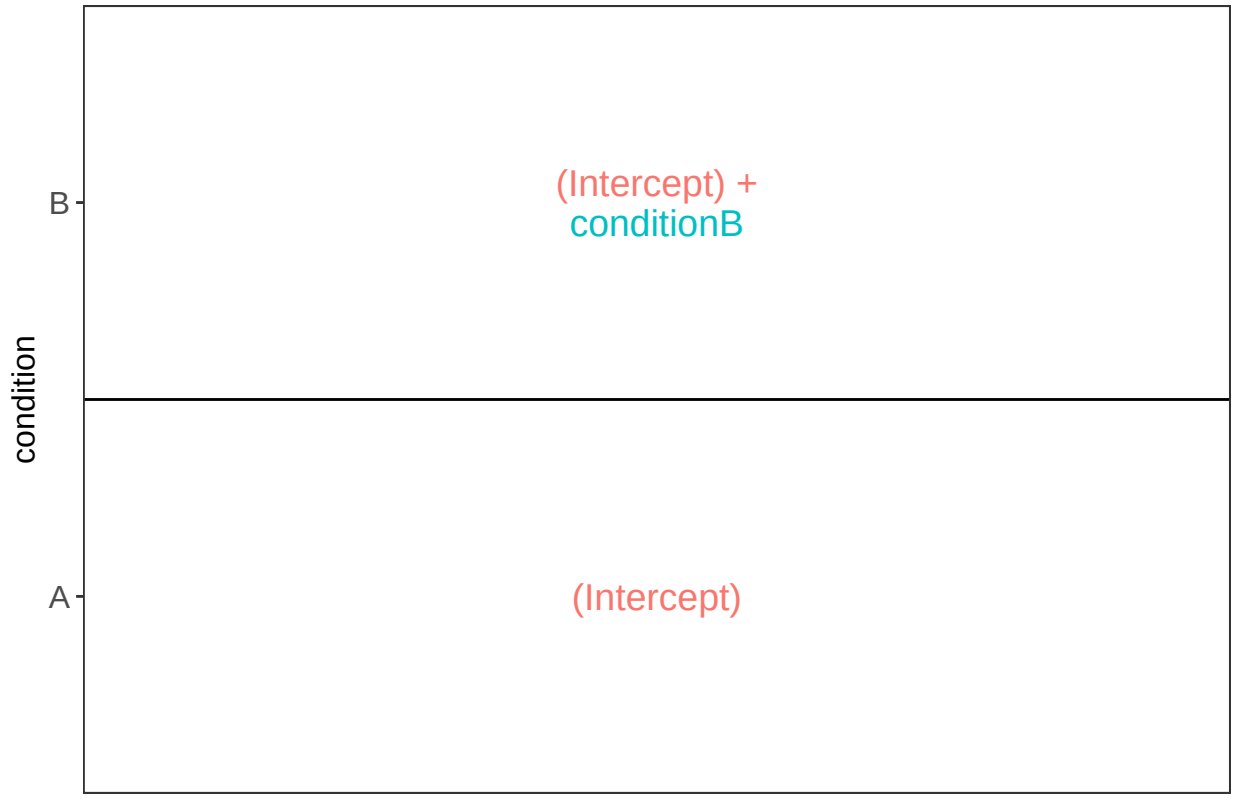
First, we extract the parameter names of the model by looking at the first model. The models are stored in the row data of the assay under the default name msqrobModels.

```
getCoef(rowData(pe[["protein"]])$msqrobModels[[1]])
```

```
## [1] NA
```

We can also explore the design of the model that we specified using the the package `ExploreModelMatrix`

```
library(ExploreModelMatrix)
VisualizeDesign(colData(pe),~condition)$plotlist[[1]]
```



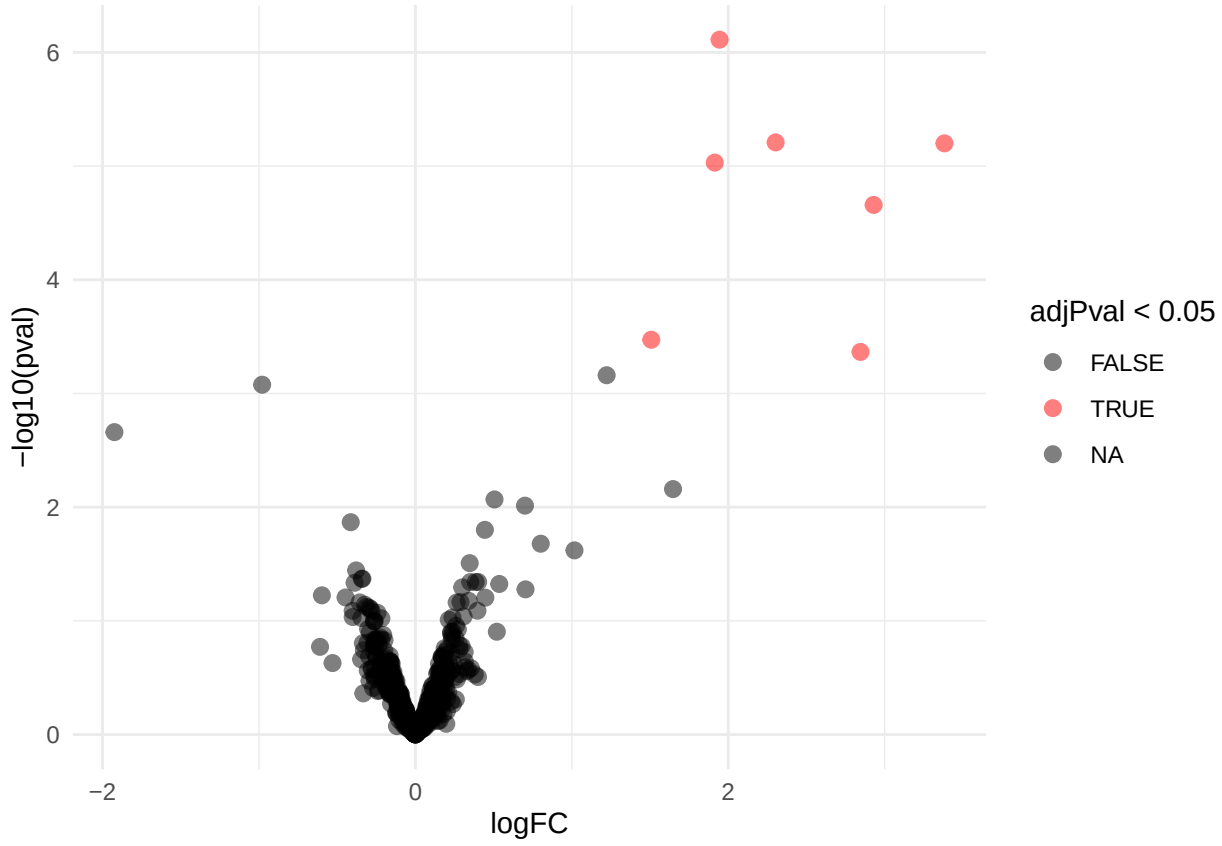
Spike-in condition `A` is the reference class. So the mean log2 expression for samples from condition A is '(Intercept). The mean log2 expression for samples from condition B is'(Intercept)+conditionB'. Hence, the average log2 fold change between condition b and condition a is modelled using the parameter 'conditionB'. Thus, we assess the contrast 'conditionB = 0' with our statistical test.

```
L <- makeContrast("conditionB=0", parameterNames = c("conditionB"))
pe <- hypothesisTest(object = pe, i = "protein", contrast = L)
```

5.3 Plots

5.3.1 Volcano-plot

```
volcano <- ggplot(rowData(pe[["protein"]])$conditionB,
                  aes(x = logFC, y = -log10(pval), color = adjPval < 0.05)) +
  geom_point(cex = 2.5) +
  scale_color_manual(values = alpha(c("black", "red"), 0.5)) + theme_minimal()
volcano
```

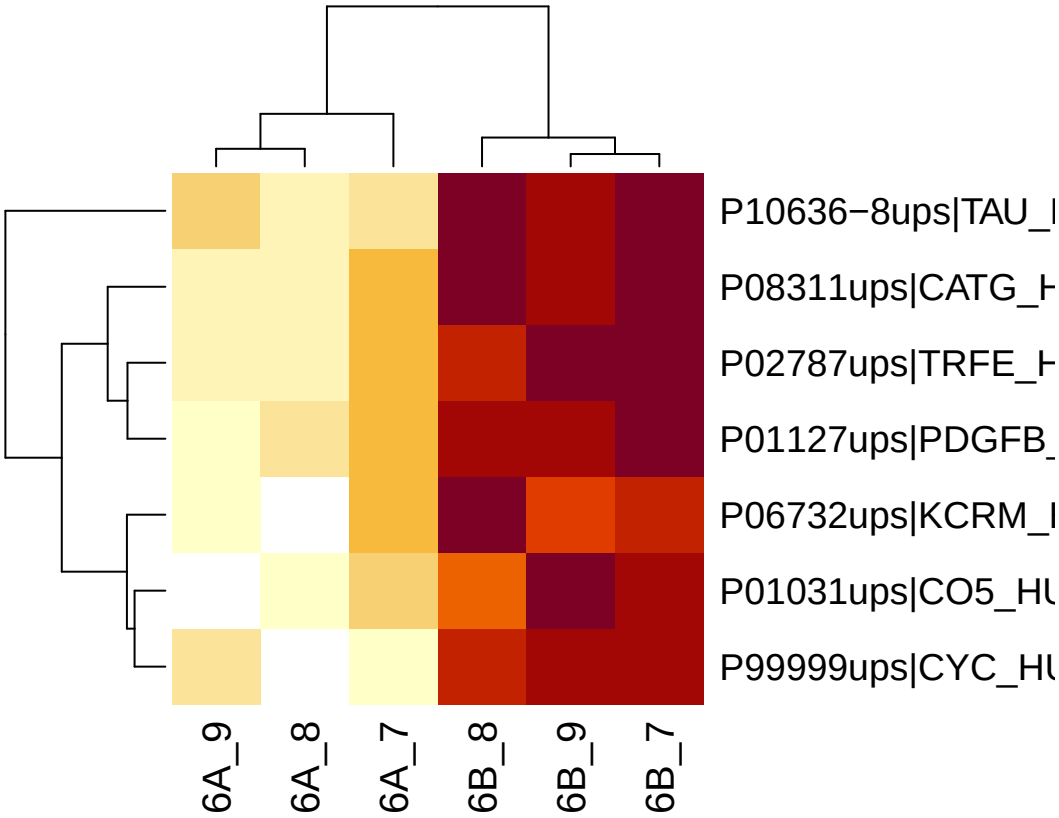


Note, that only 7 proteins are found to be differentially abundant.

5.3.2 Heatmap

We first select the names of the proteins that were declared signficant.

```
sigNames <- rowData(pe[["protein"]])$conditionB %>%
  rownames_to_column("protein") %>%
  filter(adjPval<0.05) %>%
  pull(protein)
heatmap(assay(pe[["protein"]])[sigNames, ])
```



The majority of the proteins are indeed UPS proteins. 1 yeast protein is returned. Note, that the yeast protein indeed shows evidence for differential abundance.

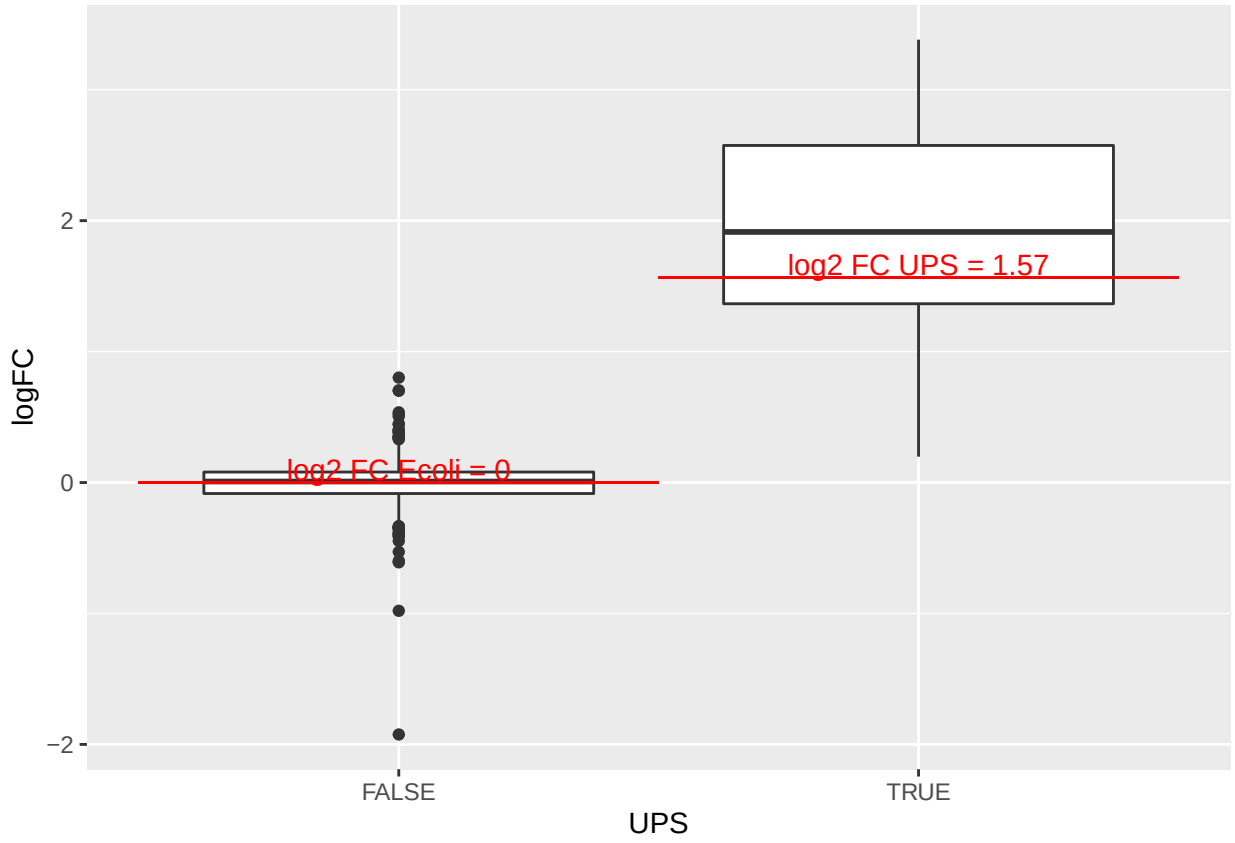
5.3.3 Boxplots

We make boxplot of the log2 FC and stratify according to the whether a protein is spiked or not.

```
rowData(pe[["protein"]])$conditionB %>%
  rownames_to_column(var = "protein") %>%
  ggplot(aes(x=grepl("UPS",protein),y=logFC)) +
  geom_boxplot() +
  xlab("UPS") +
  geom_segment(
    x = 1.5,
    xend = 2.5,
    y = log2(0.74/0.25),
    yend = log2(0.74/0.25),
    colour="red") +
  geom_segment(
    x = 0.5,
    xend = 1.5,
    y = 0,
    yend = 0,
    colour="red") +
  annotate(
```

```
    "text",
    x = c(1,2),
    y = c(0,log2(0.74/0.25))+.1,
    label = c(
      "log2 FC Ecoli = 0",
      paste0("log2 FC UPS = ",round(log2(0.74/0.25),2))
      ),
    colour = "red")
```

```
## Warning: Removed 746 rows containing non-finite values (stat_boxplot).
```



What do you observe?

6 Session Info

With respect to reproducibility, it is highly recommended to include a session info in your script so that readers of your output can see your particular setup of R.

```
sessionInfo()
```

```
## R version 4.2.2 (2022-10-31)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 22.04.1 LTS
```

```
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
## LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.20.so
##
## locale:
##  [1] LC_CTYPE=C.UTF-8       LC_NUMERIC=C           LC_TIME=C.UTF-8
##  [4] LC_COLLATE=C.UTF-8     LC_MONETARY=C.UTF-8    LC_MESSAGES=C.UTF-8
##  [7] LC_PAPER=C.UTF-8       LC_NAME=C              LC_ADDRESS=C
## [10] LC_TELEPHONE=C         LC_MEASUREMENT=C.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats4    stats     graphics  grDevices datasets  utils     methods
## [8] base
##
## other attached packages:
##  [1] ExploreModelMatrix_1.8.0    plotly_4.10.0
##  [3] msqrob2_1.4.0               QFeatures_1.6.0
##  [5] MultiAssayExperiment_1.22.0 SummarizedExperiment_1.26.1
##  [7] Biobase_2.56.0              GenomicRanges_1.48.0
##  [9] GenomeInfoDb_1.32.2         IRanges_2.30.0
## [11] S4Vectors_0.34.0            BiocGenerics_0.42.0
## [13] MatrixGenerics_1.8.0        matrixStats_0.62.0
## [15] limma_3.52.1                forcats_0.5.1
## [17] stringr_1.4.1               dplyr_1.0.9
## [19] purrr_0.3.4                 readr_2.1.2
## [21] tidyr_1.2.0                 tibble_3.1.7
## [23] ggplot2_3.3.6               tidyverse_1.3.2
##
## loaded via a namespace (and not attached):
##   [1] googledrive_2.0.0       minqa_1.2.4             colorspace_2.0-3
##   [4] ellipsis_0.3.2          XVector_0.36.0          fs_1.5.2
##   [7] clue_0.3-61             farver_2.1.0            DT_0.23
##  [10] fansi_1.0.3             lubridate_1.8.0         xml2_1.3.3
##  [13] codetools_0.2-18        splines_4.2.2           knitr_1.40.1
##  [16] jsonlite_1.8.0          nloptr_2.0.3            broom_0.8.0
##  [19] cluster_2.1.3           dbplyr_2.1.1            shinydashboard_0.7.2
##  [22] shiny_1.7.1             BiocManager_1.30.18     compiler_4.2.2
##  [25] httr_1.4.3              backports_1.4.1         assertthat_0.2.1
##  [28] Matrix_1.4-1            fastmap_1.1.0           lazyeval_0.2.2
##  [31] gargle_1.2.0            cli_3.3.0               later_1.3.0
##  [34] htmltools_0.5.2         tools_4.2.2             igraph_1.3.2
##  [37] gtable_0.3.0            glue_1.6.2              GenomeInfoDbData_1.2.8
##  [40] Rcpp_1.0.8.3            cellranger_1.1.0        jquerylib_0.1.4
##  [43] vctrs_0.4.1             nlme_3.1-157            rintrojs_0.3.0
##  [46] xfun_0.33               lme4_1.1-29             rvest_1.0.2
##  [49] mime_0.12               lifecycle_1.0.1         renv_0.15.4
##  [52] googlesheets4_1.0.0     zlibbioc_1.42.0         MASS_7.3-57
##  [55] scales_1.2.0            promises_1.2.0.1        hms_1.1.1
##  [58] ProtGenerics_1.28.0     parallel_4.2.2          AnnotationFilter_1.20.0
##  [61] yaml_2.3.5              sass_0.4.1              stringi_1.7.8
##  [64] highr_0.9               boot_1.3-28             BiocParallel_1.30.2
##  [67] rlang_1.0.2             pkgconfig_2.0.3         bitops_1.0-7
##  [70] evaluate_0.16           lattice_0.20-45         htmlwidgets_1.5.4
```

```
##  [73] labeling_0.4.2          cowplot_1.1.1           tidyselect_1.1.2
##  [76] magrittr_2.0.3          R6_2.5.1                generics_0.1.2
##  [79] DelayedArray_0.22.0     DBI_1.1.2               pillar_1.7.0
##  [82] haven_2.5.0             withr_2.5.0             MsCoreUtils_1.8.0
##  [85] RCurl_1.98-1.6          modelr_0.1.8            crayon_1.5.1
##  [88] utf8_1.2.2              tzdb_0.3.0              rmarkdown_2.14
##  [91] grid_4.2.2              readxl_1.4.0            data.table_1.14.2
##  [94] reprex_2.0.1            digest_0.6.29           xtable_1.8-4
##  [97] httpuv_1.6.5            munsell_0.5.0           viridisLite_0.4.0
## [100] bslib_0.3.1             shinyjs_2.1.0
```