chapitre 9

Imane Ez-Zammoury

13/05/2021

Setup

```
library("SummarizedExperiment")
library(tidyverse)
library(dplyr)
library(rWSBIM1207)
library(rWSBIM1322)
library(ggplot2)
```

Exo 2

4.

Extract the quantitative information for the peptides AIGVLPQLIIDR, NLDAAPTLR and YGLNHVVSLIENKK for samples 6A_7 and 6B_8.

```
a <- as_tibble(assay(cptac_se))
c <- as_tibble(colData(cptac_se))
r <- as_tibble(rowData(cptac_se))

a1 <- a %>%
  mutate(Sequence = r$Sequence) %>%
  filter(!is.na(Sequence))

seq <- a1 %>%
  filter(Sequence %in% c("AIGVLPQLIIDR",
                         "NLDAAPTLR",
                         "YGLNHVVSLIENKK")) %>%
  select("6A_7", "6B_8", "Sequence")
seq
```

```
## # A tibble: 3 x 3
##   `6A_7` `6B_8` Sequence
##    <dbl>  <dbl> <chr>
## 1  44673  37500 AIGVLPQLIIDR
## 2 562630 401190 NLDAAPTLR
## 3 389550 376080 YGLNHVVSLIENKK
```

6.

What is the average expression of LSAAQAELAYAETGAHDK in the groups 6A and 6B?

```
longer <- a1 %>%
  pivot_longer(cols = c(-Sequence),
               names_to = "groups",
               values_to = "expression")

exp <- longer %>%
  filter(Sequence == "LSAAQAELAYAETGAHDK") %>%
  group_by(Sequence, groups) %>%
  summarise(moyenne =
              mean(expression,
                   na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'Sequence'. You can override using the `.groups` argument.
```

```
exp
```

```
## # A tibble: 6 x 3
## # Groups:   Sequence [1]
##   Sequence       groups moyenne
##   <chr>          <chr>    <dbl>
## 1 LSAAQAELAYAET~ 6A_7   1646200
## 2 LSAAQAELAYAET~ 6A_8   1911500
## 3 LSAAQAELAYAET~ 6A_9   2002700
## 4 LSAAQAELAYAET~ 6B_7   1447500
## 5 LSAAQAELAYAET~ 6B_8   1819800
## 6 LSAAQAELAYAET~ 6B_9   2298700
```

7.

Calculate the average expression of all peptides belonging to protein P02753ups|RETBP_HUMAN_UPS for each sample.

```
aP <- a1 %>%
  mutate(Proteins = r$Proteins)

a2 <- aP %>%
  filter(Proteins == "P02753ups|RETBP_HUMAN_UPS") %>%
  pivot_longer(cols = c(-Proteins, -Sequence),
               names_to = "groups",
               values_to = "expression") %>%
  group_by(Sequence) %>%
  summarise(Moyenne = mean(expression, na.rm = TRUE))
a2
```

```
## # A tibble: 2 x 2
##   Sequence   Moyenne
##   <chr>        <dbl>
## 1 QRQEELCLAR  469325
## 2 YWGVASFLQK  111414
```

Exo 3

2.

Import the data from two tab-separated files into R.

```
?
kem.tsv()
kem_counts <-read_tsv("/usr/local/lib/R/site-library/rWSBIM1207/extdata/kem_counts.tsv")

kem_annot <-read_tsv("/usr/local/lib/R/site-library/rWSBIM1207/extdata/kem_annot.tsv")
```

3.

Convert the counts data into a long table format and annotate each sample using the experimental design.

```
kem_counts_longer <-
  kem_counts %>% pivot_longer(
    cols = c(-ref),
    names_to = "sample_id",
    values_to = "expression")
kem_counts_longer
```

```
## # A tibble: 208 x 3
##    ref     sample_id expression
##    <chr>   <chr>          <dbl>
##  1 ENSG00~ KEM182-01       2504
##  2 ENSG00~ KEM182-02       1714
##  3 ENSG00~ KEM182-03       2087
##  4 ENSG00~ KEM182-04       1991
##  5 ENSG00~ KEM182-05       2304
##  6 ENSG00~ KEM182-06       1820
##  7 ENSG00~ KEM182-07       1714
##  8 ENSG00~ KEM182-08       2969
##  9 ENSG00~ KEM182-09       1513
## 10 ENSG00~ KEM182-10       1554
## # ... with 198 more rows
```

4.

Identity the three transcript identifiers that have the highest expression count over all samples.

```
max <- kem_counts_longer %>%
  arrange(desc(expression)) %>%
  select(-ref)
max
```

```
## # A tibble: 208 x 2
##    sample_id expression
##    <chr>          <dbl>
##  1 KEM182-11     148564
##  2 KEM182-08     145105
##  3 KEM182-12     140030
##  4 KEM182-10     137689
```

```
##  5 KEM182-01     135299
##  6 KEM182-13     131926
##  7 KEM182-09     131579
##  8 KEM182-05     127752
##  9 KEM182-14     122057
## 10 KEM182-08     118973
## # ... with 198 more rows
```

5.

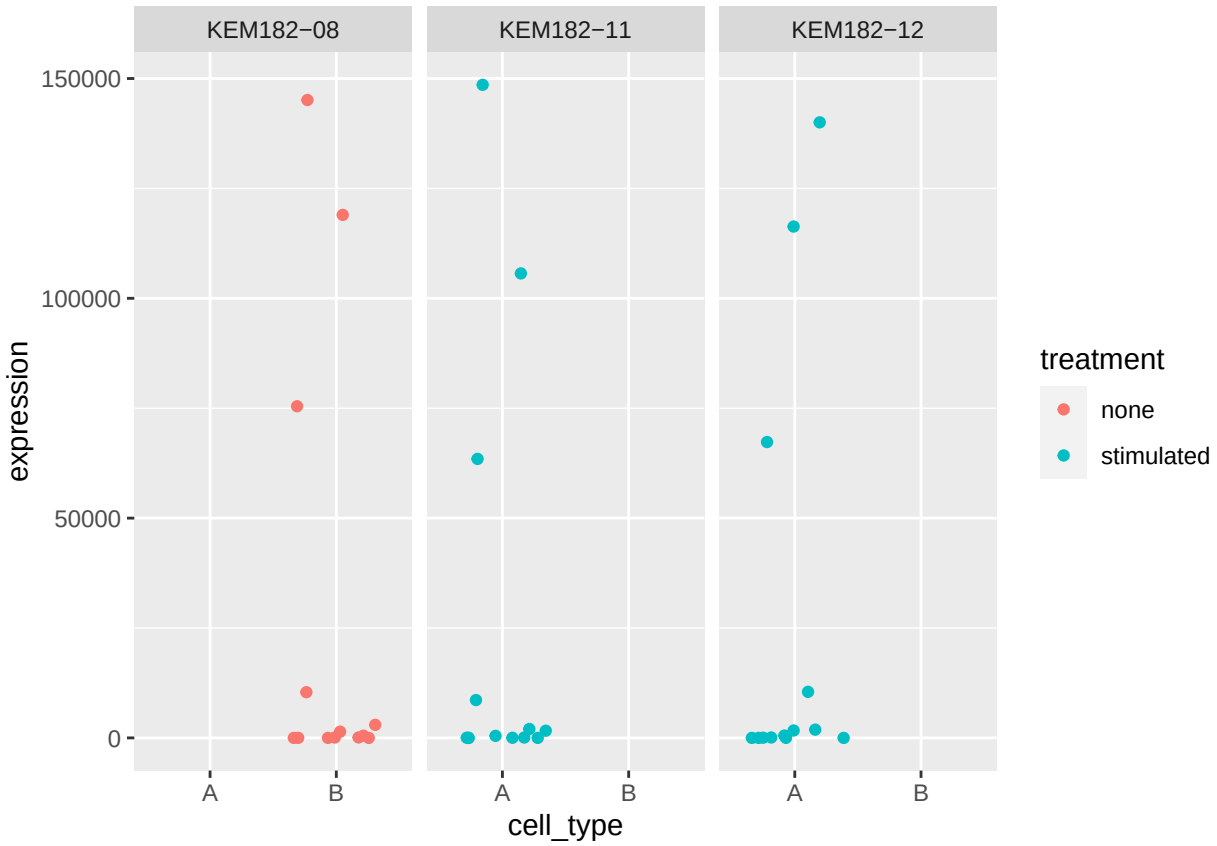
Visualise the distribution of the expression for the three transcripts selected above in cell types A and B under both treatments

```
x <- kem_counts %>% select("KEM182-11",
                    "KEM182-08",
                    "KEM182-12")
x2 <- x %>%
  pivot_longer(cols = c("KEM182-11",
                            "KEM182-08",
                            "KEM182-12"),
             names_to = "sample_id",
             values_to = "expression")

x3 <- full_join(x2, kem_annot) %>% select(-jurkat) %>%
  filter(!is.na(expression))
```

```
## Joining, by = "sample_id"
```

```
x3 %>% ggplot(aes(x = cell_type, y = expression,
       color = treatment)) +
  geom_jitter() +
  facet_grid(~sample_id)
```



6.

For all genes, calculate the mean intensities in each experimental group (as defined by the cell_type and treatment variables).

```
fj <- full_join(kem_counts_longer, kem_annot) %>%
  select(-jurkat, -ref)
```

```
## Joining, by = "sample_id"
```

```
fj %>%
  filter(!is.na(expression)) %>%
  group_by(sample_id) %>%
  summarise(Moyenne = mean(expression))
```

```
## # A tibble: 16 x 2
##    sample_id Moyenne
##    <chr>       <dbl>
##  1 KEM182-01  22438.
##  2 KEM182-02  19258.
##  3 KEM182-03  21922.
##  4 KEM182-04  22194.
##  5 KEM182-05  24587.
##  6 KEM182-06  19377.
##  7 KEM182-07  16736.
##  8 KEM182-08  27311.
##  9 KEM182-09  22970.
```

```
## 10 KEM182-10  26223
## 11 KEM182-11  25431.
## 12 KEM182-12  26030.
## 13 KEM182-13  23108.
## 14 KEM182-14  19888.
## 15 KEM182-15  18529.
## 16 KEM182-16  18872.
```

7.

Focusing only on the three most expressed transcripts and cell type A, calculate the fold-change induced by the treatment. *The fold-change is the ratio between the average expressions in two conditions.*