Chapitre 9 exercices supplémentaires

Armelle de le Court

23/05/2021

```
library("SummarizedExperiment")
library("rWSBIM1322")
data(cptac_se)
class(cptac_se)
```

```
## [1] "SummarizedExperiment"
## attr(,"package")
## [1] "SummarizedExperiment"
```

Extract the quantitative information for the peptides AIGVLPQLIIDR, NLDAAPTLR and YGLNHVVSLIENKK for samples 6A_7 and 6B_8. Subsetting works as we have seen for data.frames in chapter 3.

```
colnames(cptac_se)
```

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

```
cptac_se3 <- cptac_se[c("AIGVLPQLIIDR", "NLDAAPTLR","YGLNHVVSLIENKK"),c(1,5)]
cptac_se3
```

```
## class: SummarizedExperiment
## dim: 3 2
## metadata(3): MSnbaseFiles MSnbaseProcessing MSnbaseVersion
## assays(1): ''
## rownames(3): AIGVLPQLIIDR NLDAAPTLR YGLNHVVSLIENKK
## rowData names(3): Proteins Sequence nNA
## colnames(2): 6A_7 6B_8
## colData names(3): group sample nNA
```

```
assay(cptac_se3)
```

```
##                  6A_7   6B_8
## AIGVLPQLIIDR    44673  37500
## NLDAAPTLR      562630 401190
## YGLNHVVSLIENKK 389550 376080
```

Look and interpret the experimental design stored in the sample metadata of this experiment. To help you out, you can also read its documentation.

```
metadatac <- colData(cptac_se)
metadatac
```

```
## DataFrame with 6 rows and 3 columns
##            group    sample       nNA
##      <character> <integer> <integer>
## 6A_7          6A         7      4669
## 6A_8          6A         8      5388
## 6A_9          6A         9      5224
## 6B_7          6B         7      4651
## 6B_8          6B         8      5470
## 6B_9          6B         9      5207
```

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

```
mean(assay(cptac_se)["LSAAQAELAYAETGAHDK",1:3])
```

```
## [1] 1853467
```

```
mean(assay(cptac_se)["LSAAQAELAYAETGAHDK",4:6])
```

```
## [1] 1855333
```

Calculate the average expression of all peptides belonging to protein P02753ups|RETBP_HUMAN_UPS for each sample. You can indentify which peptides to use by looking for that protein in the object's rowData slot.

```
rowData(cptac_se)
```

```
## DataFrame with 4051 rows and 3 columns
##                              Proteins        Sequence       nNA
##                           <character>     <character> <integer>
## AAAALAGGK       sp|Q3E792|RS25A_YEAS..       AAAALAGGK         0
## AAAALAGGKK      sp|Q3E792|RS25A_YEAS..      AAAALAGGKK         0
## AAADALSDLEIK      sp|P09938|RIR2_YEAST    AAADALSDLEIK         0
## AAADALSDLEIKDSK   sp|P09938|RIR2_YEAST AAADALSDLEIKDSK         0
## AAALVNK           sp|P05030|PMA1_YEAST         AAALVNK         0
## ...                                ...             ...       ...
## YVVLASHLGRPNGER    sp|P00560|PGK_YEAST YVVLASHLGRPNGER         0
## YWGVASFLQK      P02753ups|RETBP_HUMA..      YWGVASFLQK         0
## YYGGNEIIDK        sp|P37292|GLYM_YEAST      YYGGNEIIDK         0
## YYPSYIVSK         sp|P22147|XRN1_YEAST       YYPSYIVSK         0
## YYSISSSSLSEK      sp|P16603|NCPR_YEAST    YYSISSSSLSEK         0
```

```
avg <- rowData(cptac_se)$Proteins== "P02753ups|RETBP_HUMAN_UPS"
min(which(avg == TRUE))
```

```
## [1] 2742
```

```
which(avg == TRUE)
```

```
## [1] 2742 4048
```

```
pp <- cptac_se[c(2742,4048)]
pp
```

```
## class: SummarizedExperiment
## dim: 2 6
```

```
## metadata(3): MSnbaseFiles MSnbaseProcessing MSnbaseVersion
## assays(1): ''
## rownames(2): QRQEELCLAR YWGVASFLQK
## rowData names(3): Proteins Sequence nNA
## colnames(6): 6A_7 6A_8 ... 6B_8 6B_9
## colData names(3): group sample nNA
```

```
app <- assay(pp)
mean(app)
```

```
## [1] 290369.5
```

Exercices with KEM

Import the data from two tab-separated files into R. The full paths to the two files can be accessed with kem.tsv(). Read ?kem for details on the content of the two files. In brief, the kem_counts.tsv file contains RNA-Seq expression counts for 13 genes and 18 samples and kem_annot.tsv contains annotation about each sample. Read the data into two tibbles names kem and annot respectively and familiarise yourself with the content of the two new tables.

```
library(tidyverse)
library(dplyr)

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

```
## # A tibble: 16 x 4
##    sample_id jurkat cell_type treatment
##    <chr>     <chr>  <chr>     <chr>
##  1 KEM182-01 yes    A         none
##  2 KEM182-02 yes    A         none
##  3 KEM182-03 yes    A         none
##  4 KEM182-04 yes    A         none
##  5 KEM182-05 yes    B         none
##  6 KEM182-06 yes    B         none
##  7 KEM182-07 yes    B         none
##  8 KEM182-08 yes    B         none
##  9 KEM182-09 yes    A         stimulated
## 10 KEM182-10 yes    A         stimulated
## 11 KEM182-11 yes    A         stimulated
## 12 KEM182-12 yes    A         stimulated
## 13 KEM182-13 yes    B         stimulated
## 14 KEM182-14 yes    B         stimulated
## 15 KEM182-15 yes    B         stimulated
## 16 KEM182-16 yes    B         stimulated
```

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

```
kem_long <- kem%>%
  pivot_longer(names_to = "sample_id", values_to="gene_expression",-ref)
```

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

```
# kem_long%>%
# top_n(3, counts)
```

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

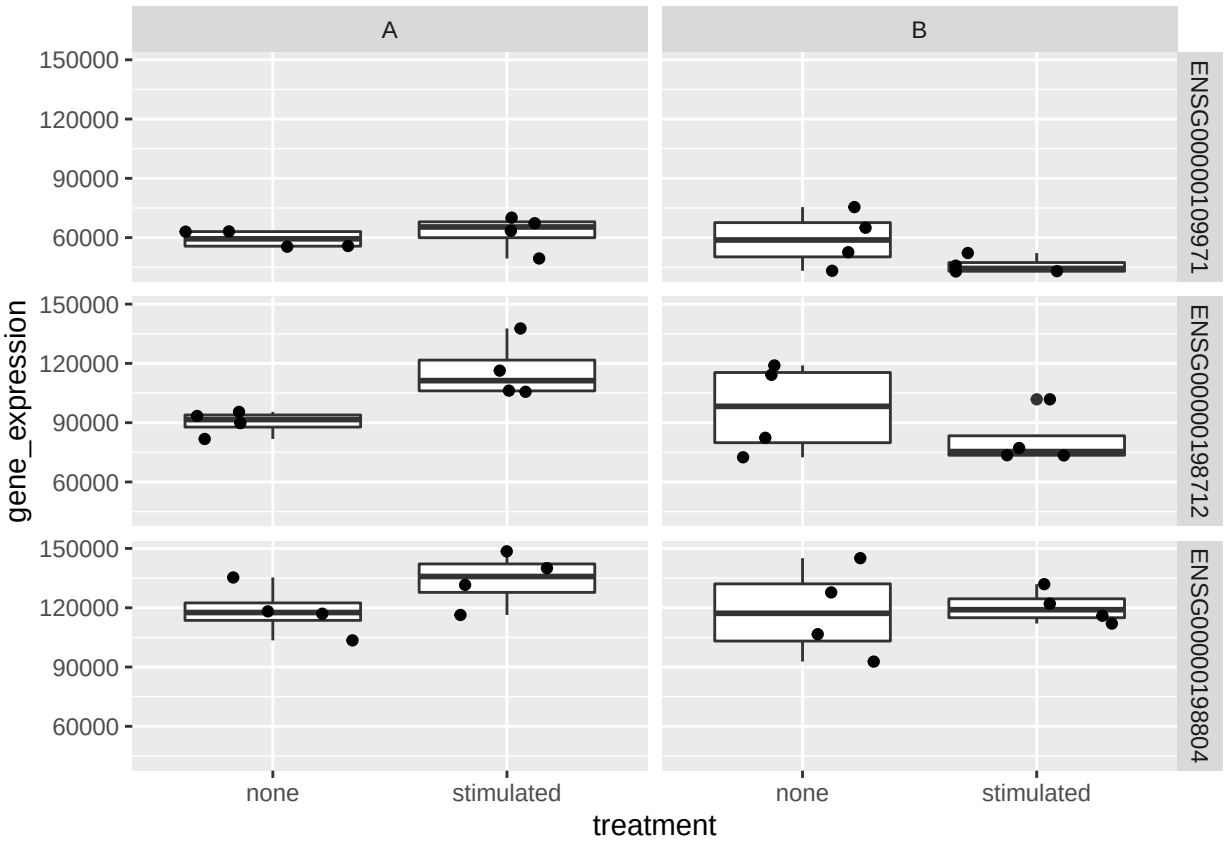
```
full_join(kem_long, annot) %>%
  group_by(ref) %>% summarise(counts=sum(gene_expression)) %>%
  arrange(desc(counts))
```

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

```
## # A tibble: 13 x 2
##    ref              counts
##    <chr>             <dbl>
##  1 ENSG00000198804 1964649
##  2 ENSG00000198712 1540373
##  3 ENSG00000109971  907840
##  4 ENSG00000188229  135118
##  5 ENSG00000055917   32478
##  6 ENSG00000063245   23426
##  7 ENSG00000090263    6675
##  8 ENSG00000137193    1682
##  9 ENSG00000198182     782
## 10 ENSG00000226259     197
## 11 ENSG00000267890      77
## 12 ENSG00000268262      33
## 13 ENSG00000224936       2
```

```
full_join(kem_long, annot) %>%
  filter(ref %in% c("ENSG00000198804","ENSG00000198712","ENSG00000109971")) %>%
  ggplot(aes(x=treatment, y= gene_expression)) +
  geom_boxplot() +
  geom_jitter()+
  facet_grid(ref~cell_type)
```

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



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

```
full_join(kem_long, annot) %>%
  group_by(cell_type,treatment,ref) %>%
  summarise(m=mean(gene_expression))
```

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

```
## `summarise()` has grouped output by 'cell_type', 'treatment'. You can override using the `.groups` a
```

```
## # A tibble: 52 x 4
## # Groups:   cell_type, treatment [4]
##    cell_type treatment ref                 m
##    <chr>     <chr>     <chr>             <dbl>
##  1 A         none      ENSG00000055917   2074
##  2 A         none      ENSG00000063245   1290
##  3 A         none      ENSG00000090263    425.
##  4 A         none      ENSG00000109971  59334.
##  5 A         none      ENSG00000137193     87.5
##  6 A         none      ENSG00000188229   7018.
##  7 A         none      ENSG00000198182     48.5
##  8 A         none      ENSG00000198712  90102.
##  9 A         none      ENSG00000198804 118486.
## 10 A         none      ENSG00000224936      0.5
## # ... with 42 more rows
```

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.

```
kem_long_annot <- full_join(kem_long, annot)
```

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

```
kem_long_annot2 <- kem_long_annot%>%
  select(ref, cell_type, sample_id, gene_expression, treatment) %>%
  filter(ref %in% c("ENSG00000198804", "ENSG00000198712", "ENSG00000109971"),
         cell_type == "A")

kem_long_annot2
```

```
## # A tibble: 24 x 5
##    ref             cell_type sample_id gene_expression treatment
##    <chr>           <chr>     <chr>               <dbl> <chr>
##  1 ENSG00000109971 A         KEM182-01           55400 none
##  2 ENSG00000109971 A         KEM182-02           55768 none
##  3 ENSG00000109971 A         KEM182-03           63149 none
##  4 ENSG00000109971 A         KEM182-04           63017 none
##  5 ENSG00000109971 A         KEM182-09           49444 stimulated
##  6 ENSG00000109971 A         KEM182-10           70096 stimulated
##  7 ENSG00000109971 A         KEM182-11           63470 stimulated
##  8 ENSG00000109971 A         KEM182-12           67300 stimulated
##  9 ENSG00000198712 A         KEM182-01           89785 none
## 10 ENSG00000198712 A         KEM182-02           81772 none
## # ... with 14 more rows
```

```
kem_long_annot2 %>%
  na.omit() %>%
  group_by(sample_id) %>%
  pivot_wider(names_from = "treatment",
              values_from = "gene_expression",
              c(ref, cell_type, sample_id))
```

```
## # A tibble: 24 x 5
## # Groups:   sample_id [8]
##    ref             cell_type sample_id  none stimulated
##    <chr>           <chr>     <chr>     <dbl>      <dbl>
##  1 ENSG00000109971 A         KEM182-01 55400         NA
##  2 ENSG00000109971 A         KEM182-02 55768         NA
##  3 ENSG00000109971 A         KEM182-03 63149         NA
##  4 ENSG00000109971 A         KEM182-04 63017         NA
##  5 ENSG00000109971 A         KEM182-09    NA      49444
##  6 ENSG00000109971 A         KEM182-10    NA      70096
##  7 ENSG00000109971 A         KEM182-11    NA      63470
##  8 ENSG00000109971 A         KEM182-12    NA      67300
##  9 ENSG00000198712 A         KEM182-01 89785         NA
## 10 ENSG00000198712 A         KEM182-02 81772         NA
## # ... with 14 more rows
```