SP1 Analysis

Orfeas Gkourlias

3/10/2022

3. Exploratory Data Analysis

The data will be need to be extracted from the count file first. Since it's in .tsv format, the seperator is going to be tab based. The inclusion of headers adds an X to the sequence ID's, because R is unable to make headers out of just integers.

3.1 Loading the data

```
file <- c("..\\data\\GSE152262_RNAseq_Raw_Counts.tsv")
raw_data <- read.table(file, sep = '\t', header = TRUE)
raw_data[1:5,]
```

```
##                 X X4275 X4277 X4279 X4280 X4280a X4281
## 1 ENSG00000000003    23    30    11    43     31     8
## 2 ENSG00000000005     0     0     0     0      2     0
## 3 ENSG00000000419   778   910   838   911   1113  1051
## 4 ENSG00000000457   378   438   441   772    738   389
## 5 ENSG00000000460    44    51    58    61     65    28
```

```
dim(raw_data)
```

```
## [1] 58307     7
```

```
str(raw_data)
```

```
## 'data.frame':    58307 obs. of  7 variables:
##  $ X     : chr  "ENSG00000000003" "ENSG00000000005" "ENSG00000000419" "ENSG00000000457" ...
##  $ X4275 : int  23 0 778 378 44 14575 30 54 213 546 ...
##  $ X4277 : int  30 0 910 438 51 21109 23 89 206 589 ...
##  $ X4279 : int  11 0 838 441 58 7164 94 105 333 452 ...
##  $ X4280 : int  43 0 911 772 61 11710 151 77 419 407 ...
##  $ X4280a: int  31 2 1113 738 65 11846 148 69 384 373 ...
##  $ X4281 : int  8 0 1051 389 28 27759 68 50 180 561 ...
```

The data is now loaded in as a data frame. Every row shows the raw counts of a specific gene being expressed. 4275, 4277 and 4281 are the variant types. The datatypes are correct in this case. There should only be integers included, except for the gene names.

Now that the data has been properly loaded, objects can be made to differentiate the control and case counts.

```
case <- raw_data[,c(1:3,7)]
control <- raw_data[,c(1,4:6)]

case[1,]
```

```
##                 X X4275 X4277 X4281
## 1 ENSG00000000003    23    30     8
```

```
control[1,]
```

```
##                 X X4279 X4280 X4280a
## 1 ENSG00000000003    11    43     31
```