Project report

Mario Limón

2025-02-02

Contents

Data process

```
library("recount3")
```

Libraries

```
## Cargando paquete requerido: SummarizedExperiment

## Cargando paquete requerido: MatrixGenerics

## Cargando paquete requerido: matrixStats

##
## Adjuntando el paquete: 'MatrixGenerics'

## The following objects are masked from 'package:matrixStats':
##
##     colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars
```

```
## Cargando paquete requerido: GenomicRanges

## Cargando paquete requerido: stats4

## Cargando paquete requerido: BiocGenerics

##
## Adjuntando el paquete: 'BiocGenerics'

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, saveRDS, setdiff,
##     table, tapply, union, unique, unsplit, which.max, which.min

## Cargando paquete requerido: S4Vectors

##
## Adjuntando el paquete: 'S4Vectors'

## The following object is masked from 'package:utils':
##
##     findMatches

## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname

## Cargando paquete requerido: IRanges

##
## Adjuntando el paquete: 'IRanges'

## The following object is masked from 'package:grDevices':
##
##     windows

## Cargando paquete requerido: GenomeInfoDb

## Cargando paquete requerido: Biobase
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.

##
## Adjuntando el paquete: 'Biobase'

## The following object is masked from 'package:MatrixGenerics':
##
##     rowMedians

## The following objects are masked from 'package:matrixStats':
##
##     anyMissing, rowMedians
```

```
library("ggplot2")
library("edgeR")
```

```
## Cargando paquete requerido: limma

##
## Adjuntando el paquete: 'limma'

## The following object is masked from 'package:BiocGenerics':
##
##     plotMA
```

```
library("limma")
library("pheatmap")
library("RColorBrewer")
```

Download data Downloading data from recount 3 database of project 'Transcriptome profiling in knock-in mouse models of Huntington' 's disease [Striatum_mRNA] (SRP053398). The data of this study is to analyze the expression profiles of RNA in mouse models with Hungtington's diseaase at different ages with different lengths of CAG repeats, to understand better the molecular changes related with this neurodegenarative disease.

```
rse_gene_SRP053398 <- recount3::create_rse_manual(
  project = "SRP053398",
  project_home = "data_sources/sra",
  organism = "mouse",
  annotation = "gencode_v23",
  type = "gene")
```

```
## 2025-02-03 22:46:05.151958 downloading and reading the metadata.
```

```
## 2025-02-03 22:46:05.896468 caching file sra.sra.SRP053398.MD.gz.
```

```
## 2025-02-03 22:46:06.999361 caching file sra.recount_project.SRP053398.MD.gz.

## 2025-02-03 22:46:07.712445 caching file sra.recount_qc.SRP053398.MD.gz.

## 2025-02-03 22:46:08.525597 caching file sra.recount_seq_qc.SRP053398.MD.gz.

## 2025-02-03 22:46:09.172936 caching file sra.recount_pred.SRP053398.MD.gz.

## 2025-02-03 22:46:09.346805 downloading and reading the feature information.

## 2025-02-03 22:46:10.178305 caching file mouse.gene_sums.M023.gtf.gz.

## 2025-02-03 22:46:11.638169 downloading and reading the counts: 208 samples across 55421 features.

## 2025-02-03 22:46:12.245009 caching file sra.gene_sums.SRP053398.M023.gz.

## Warning in grep(pattern, bfr, value = TRUE): unable to translate ' El n<a3>mero
## de serie del volumen es: B435-11DF' to a wide string

## Warning in grep(pattern, bfr, value = TRUE): input string 2 is invalid

## Warning in grep(pattern, bfr, value = TRUE): unable to translate ' El n<a3>mero
## de serie del volumen es: B435-11DF' to a wide string

## Warning in grep(pattern, bfr, value = TRUE): input string 2 is invalid

## Warning in grep(pattern, bfr, value = TRUE): unable to translate ' El n<a3>mero
## de serie del volumen es: B435-11DF' to a wide string

## Warning in grep(pattern, bfr, value = TRUE): input string 2 is invalid

## 2025-02-03 22:46:16.317789 constructing the RangedSummarizedExperiment (rse) object.
```

```r
# Transform the counts
assay(rse_gene_SRP053398, "counts") <- compute_read_counts(rse_gene_SRP053398)

# Before expanding the data we should explore the attributes
rse_gene_SRP053398$sra.sample_attributes[1:3]
```

Processing data

```
## [1] "age;;2 month|genotype;;Het (Q140) Knock-In|housing condition;;housed with Het (Q140) mice|Sex;;
## [2] "age;;2 month|genotype;;Het (Q140) Knock-In|housing condition;;housed with Het (Q140) mice|Sex;;
## [3] "age;;2 month|genotype;;Het (Q140) Knock-In|housing condition;;housed with Het (Q140) mice|Sex;;
```

```
# Now we expand the attributes to process the data to make easier the analysis
rse_gene_SRP053398 <- expand_sra_attributes(rse_gene_SRP053398)
colData(rse_gene_SRP053398)[
    ,
    grepl("^sra_attribute", colnames(colData(rse_gene_SRP053398)))]
```

```
## DataFrame with 208 rows and 7 columns
##             sra_attribute.age sra_attribute.genotype
##                   <character>            <character>
## SRR1795046           2 month    Het (Q140) Knock-In
## SRR1795047           2 month    Het (Q140) Knock-In
## SRR1795048           2 month    Het (Q140) Knock-In
## SRR1795049           2 month    Het (Q140) Knock-In
## SRR1795050           2 month    Het (Q140) Knock-In
## ...                      ...                    ...
## SRR1795224          10 month              Wild Type
## SRR1795225          10 month              Wild Type
## SRR1795226          10 month              Wild Type
## SRR1795227          10 month              Wild Type
## SRR1795228          10 month              Wild Type
##             sra_attribute.housing_condition sra_attribute.Sex
##                                 <character>       <character>
## SRR1795046          housed with Het (Q14..            female
## SRR1795047          housed with Het (Q14..            female
## SRR1795048          housed with Het (Q14..            female
## SRR1795049          housed with Het (Q14..              male
## SRR1795050          housed with Het (Q14..              male
## ...                                    ...               ...
## SRR1795224          housed with Het (Q20..            female
## SRR1795225          housed with Het (Q20..              male
## SRR1795226          housed with Het (Q20..              male
## SRR1795227          housed with Het (Q20..              male
## SRR1795228          housed with Het (Q20..              male
##             sra_attribute.source_name sra_attribute.strain sra_attribute.tissue
##                           <character>          <character>          <character>
## SRR1795046                  striatum              C57BL/6             striatum
## SRR1795047                  striatum              C57BL/6             striatum
## SRR1795048                  striatum              C57BL/6             striatum
## SRR1795049                  striatum              C57BL/6             striatum
## SRR1795050                  striatum              C57BL/6             striatum
## ...                              ...                  ...                  ...
## SRR1795224                  striatum              C57BL/6             striatum
## SRR1795225                  striatum              C57BL/6             striatum
## SRR1795226                  striatum              C57BL/6             striatum
## SRR1795227                  striatum              C57BL/6             striatum
## SRR1795228                  striatum              C57BL/6             striatum
```

```
# To perform the statistical analysis we need to use a correct R format
rse_gene_SRP053398$sra_attribute.age <- as.numeric(gsub(" month", "", rse_gene_SRP053398$sra_attribute.a
rse_gene_SRP053398$sra_attribute.genotype <- factor(tolower(rse_gene_SRP053398$sra_attribute.genotype))
rse_gene_SRP053398$sra_attribute.Sex <- factor(rse_gene_SRP053398$sra_attribute.Sex)

# We can check the summary of the interest variables
```

```
summary(as.data.frame(colData(rse_gene_SRP053398)[
    ,
    grepl("^sra_attribute.[age|disease|RIN|sex]", colnames(colData(rse_gene_SRP053398)))
]))
```

```
##  sra_attribute.age         sra_attribute.genotype sra_attribute.source_name
##  Min.   : 2        het (q111) knock-in:24         Length:208
##  1st Qu.: 2        het (q140) knock-in:24         Class :character
##  Median : 6        het (q175) knock-in:24         Mode  :character
##  Mean   : 6        het (q20) knock-in :24
##  3rd Qu.:10        het (q80) knock-in :24
##  Max.   :10        het (q92) knock-in :24
##                    wild type          :64
##  sra_attribute.strain
##  Length:208
##  Class :character
##  Mode  :character
##
##
##
##
```

We will check the quality of the data calculating the proportion of assigned genes.

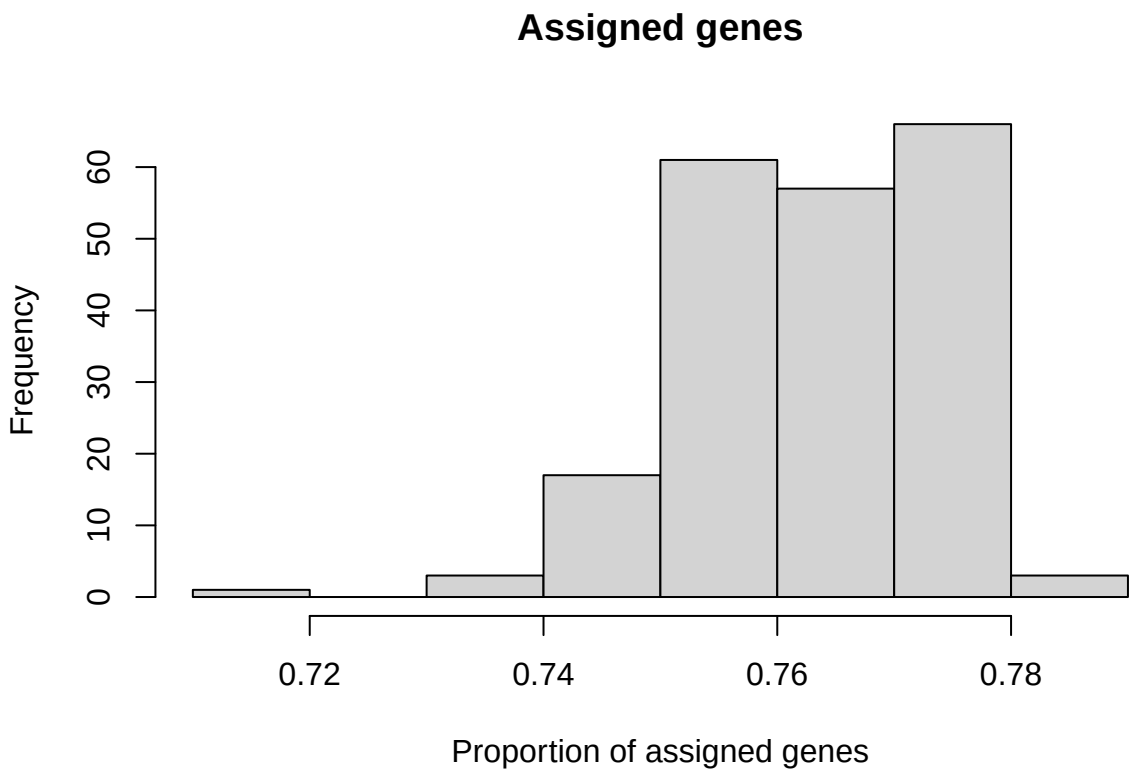
```
rse_gene_SRP053398$assigned_gene_prop <- rse_gene_SRP053398$recount_qc.gene_fc_count_all.assigned /
  rse_gene_SRP053398$recount_qc.gene_fc_count_all.total
summary(rse_gene_SRP053398$assigned_gene_prop)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.7192  0.7569  0.7650  0.7634  0.7718  0.7841
```

```
# We can create another variable to make more general the analysis
rse_gene_SRP053398$state <- factor(ifelse(rse_gene_SRP053398$sra_attribute.genotype != 'wild type' , "kn
table(rse_gene_SRP053398$state)
```

```
##
## knock-in wildtype
##      144       64
```

```r
hist(rse_gene_SRP053398$assigned_gene_prop, main = "Assigned genes", xlab = "Proportion of assigned gene
```




```r
# We can also check if there is difference in the quality of the data between the different states
with(colData(rse_gene_SRP053398), tapply(assigned_gene_prop, state, summary))
```

```
## $`knock-in`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.7192  0.7555  0.7605  0.7614  0.7700  0.7841
##
## $wildtype
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.7406  0.7665  0.7703  0.7679  0.7730  0.7793
```

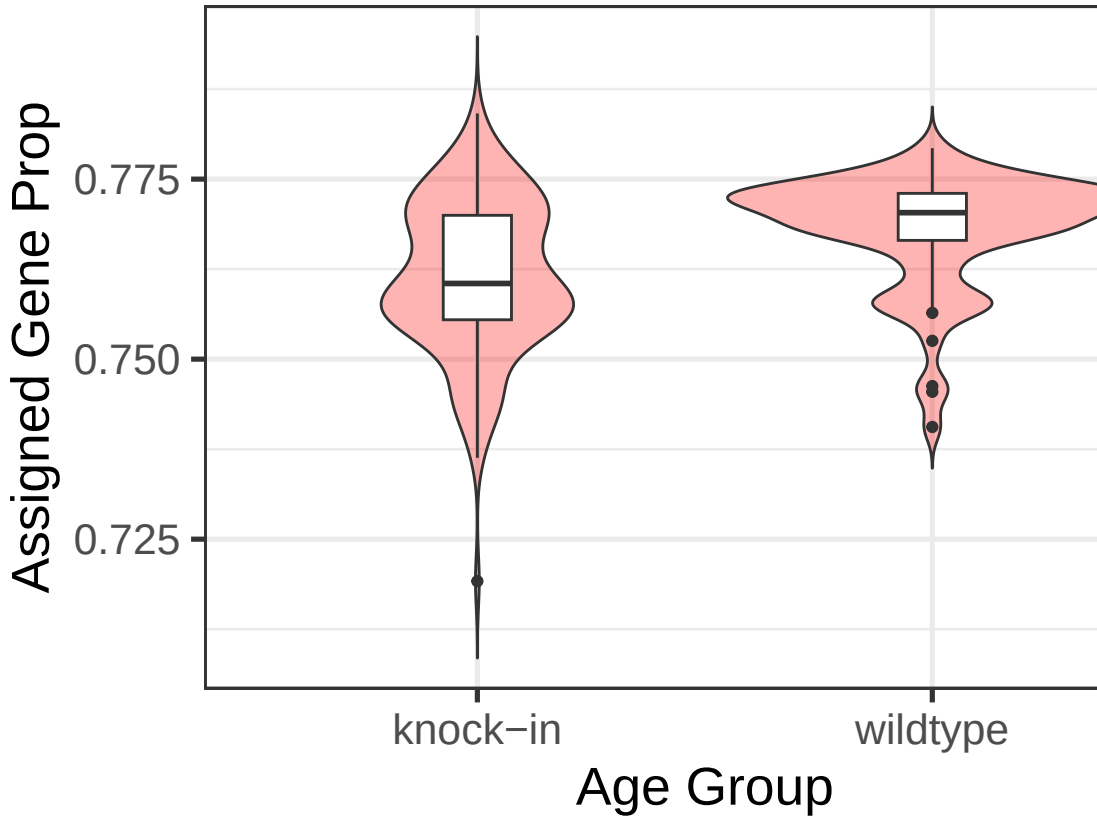
Normalization of the data Normalizing data is essential because it ensures that all features contribute equally to the analysis, preventing bias towards variables with larger scales.

```r
# With the library of edgeR we will normalize the data
dge <- DGEList(
    counts = assay(rse_gene_SRP053398, "counts"),
    genes = rowData(rse_gene_SRP053398)
)
dge <- calcNormFactors(dge)
```

Expression data analysis

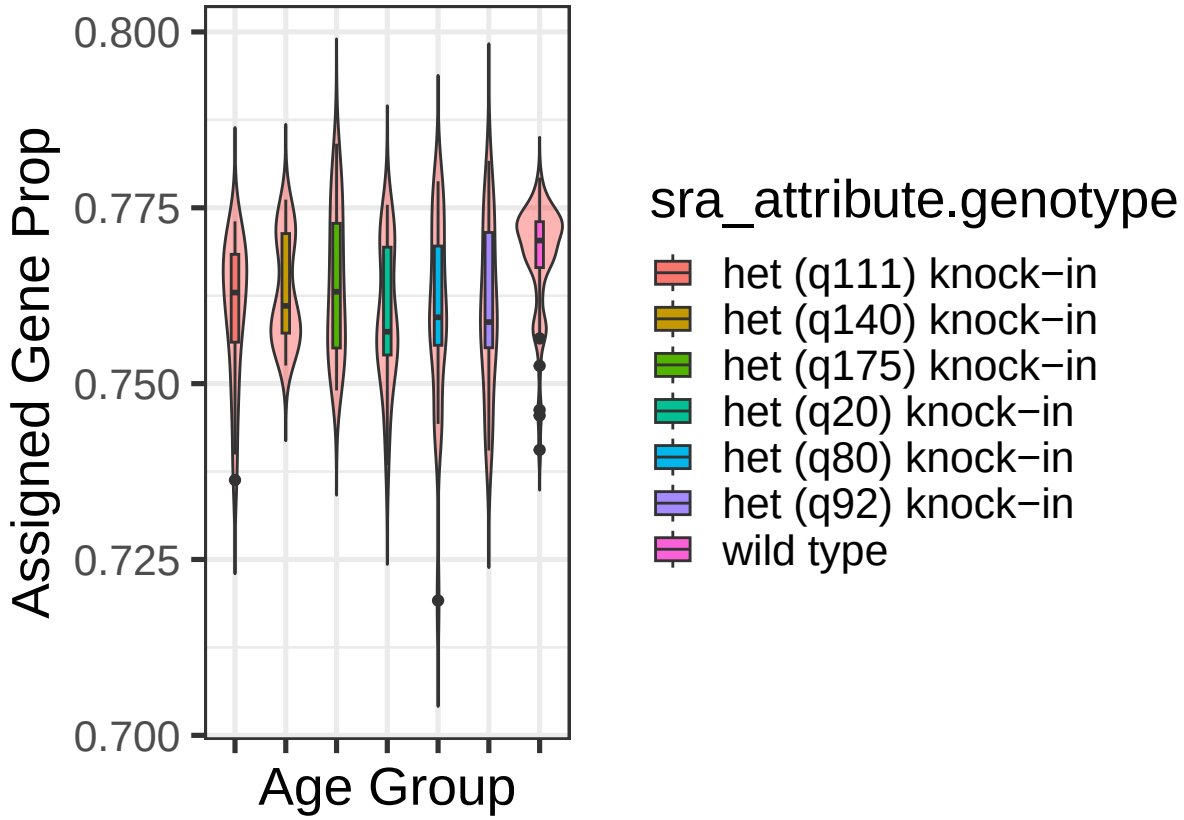
Here we will compare the differences between the different states of the mice (knock-in and wildtype), and the differences of the knock-in mice CAG repeats lengths.

```
ggplot(as.data.frame(colData(rse_gene_SRP053398)), aes(y = assigned_gene_prop, x = state,)) +
    geom_violin(trim = FALSE, alpha = 0.3, fill = 'red') +
    geom_boxplot(width = 0.15) +
    theme_bw(base_size = 20) +
    ylab("Assigned Gene Prop") +
    xlab("Age Group")
```



Data distribution plots

```
ggplot(as.data.frame(colData(rse_gene_SRP053398)), aes(y = assigned_gene_prop, x = sra_attribute.genotyp
    fill = sra_attribute.genotype)) +
    geom_violin(trim = FALSE, alpha = 0.3, fill = 'red') +
    geom_boxplot(width = 0.15) +
    theme_bw(base_size = 20) +
    ylab("Assigned Gene Prop") +
    xlab("Age Group") +
    theme(axis.text.x = element_blank())
```




```
mod <- model.matrix(~ state + sra_attribute.Sex + sra_attribute.age + assigned_gene_prop,
    data = colData(rse_gene_SRP053398)
)
colnames(mod)
```

Statistical model

```
## [1] "(Intercept)"           "statewildtype"         "sra_attribute.Sexmale"
## [4] "sra_attribute.age"     "assigned_gene_prop"
```

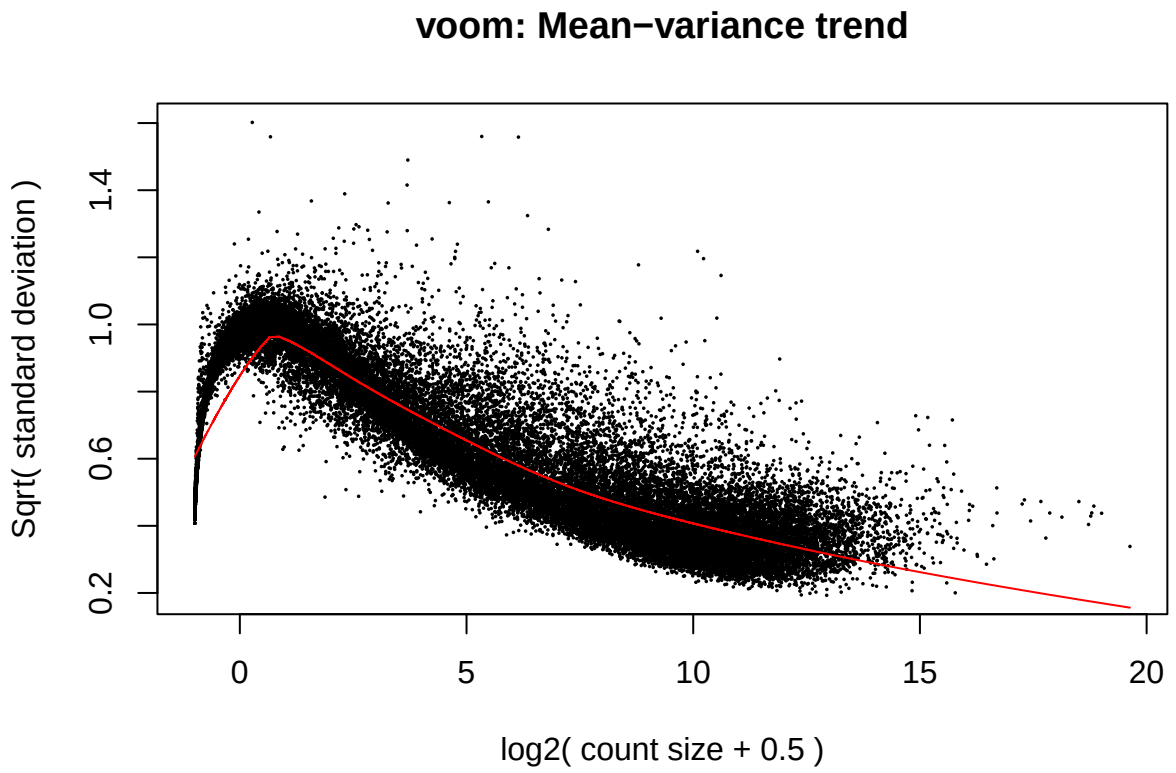
```
mod2 <- model.matrix(~ sra_attribute.genotype + sra_attribute.Sex + sra_attribute.age + assigned_gene_pr
    data = colData(rse_gene_SRP053398)
)
colnames(mod2)
```

```
##  [1] "(Intercept)"
##  [2] "sra_attribute.genotypehet (q140) knock-in"
##  [3] "sra_attribute.genotypehet (q175) knock-in"
##  [4] "sra_attribute.genotypehet (q20) knock-in"
##  [5] "sra_attribute.genotypehet (q80) knock-in"
##  [6] "sra_attribute.genotypehet (q92) knock-in"
##  [7] "sra_attribute.genotypewild type"
```

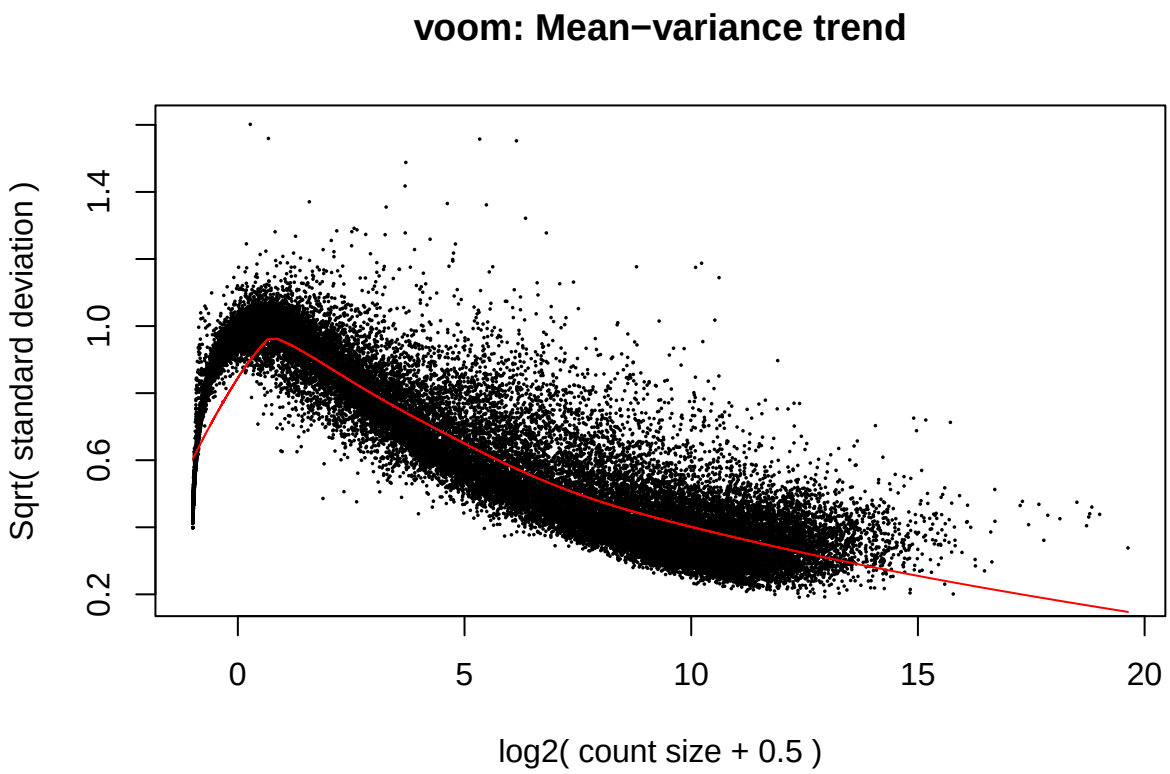
```
##  [8] "sra_attribute.Sexmale"
##  [9] "sra_attribute.age"
## [10] "assigned_gene_prop"
```

Now with the statistical model, we can visualize the expression data between the two models.

```
vGene <- voom(dge, mod, plot = TRUE)
```




```
vGene2 <- voom(dge, mod2, plot = TRUE)
```



Lets see the data in an easier way.

```
eb_results <- eBayes(lmFit(vGene))
eb_results2 <- eBayes(lmFit(vGene2))
```

```
de_results <- topTable(
    eb_results,
    coef = 2,
    number = nrow(rse_gene_SRP053398),
    sort.by = "none"
)
dim(de_results)
```

```
## [1] 55421    17
```

```
de_results2 <- topTable(
    eb_results2,
    coef = 2,
    number = nrow(rse_gene_SRP053398),
    sort.by = "none"
)
dim(de_results2)
```

```
## [1] 55421    17
```

```
head(de_results)
```

```
##                         source type bp_length phase              gene_id
## ENSMUSG00000079800.2 ENSEMBL gene      1271    NA ENSMUSG00000079800.2
## ENSMUSG00000095092.1 ENSEMBL gene       366    NA ENSMUSG00000095092.1
## ENSMUSG00000079192.2 ENSEMBL gene       255    NA ENSMUSG00000079192.2
## ENSMUSG00000079794.2 ENSEMBL gene       255    NA ENSMUSG00000079794.2
## ENSMUSG00000094799.1 ENSEMBL gene       366    NA ENSMUSG00000094799.1
## ENSMUSG00000095250.1 ENSEMBL gene       366    NA ENSMUSG00000095250.1
##                            gene_type  gene_name level mgi_id havana_gene  tag
## ENSMUSG00000079800.2 protein_coding AC125149.3     3   <NA>        <NA> <NA>
## ENSMUSG00000095092.1 protein_coding AC125149.5     3   <NA>        <NA> <NA>
## ENSMUSG00000079192.2 protein_coding AC125149.1     3   <NA>        <NA> <NA>
## ENSMUSG00000079794.2 protein_coding AC125149.2     3   <NA>        <NA> <NA>
## ENSMUSG00000094799.1 protein_coding AC125149.4     3   <NA>        <NA> <NA>
## ENSMUSG00000095250.1 protein_coding AC133103.4     3   <NA>        <NA> <NA>
##                              logFC   AveExpr           t   P.Value adj.P.Val
## ENSMUSG00000079800.2 -0.006062492 -6.178024 -0.07442493 0.9407445 0.9982871
## ENSMUSG00000095092.1  0.091059089 -6.268930  1.63293946 0.1040093 0.3557332
## ENSMUSG00000079192.2 -0.034630316 -6.215590 -0.56118870 0.5752789 0.9298723
## ENSMUSG00000079794.2  0.072265592 -6.185644  0.94882576 0.3438210 0.7342276
## ENSMUSG00000094799.1  0.209568778 -5.732321  1.47977583 0.1404608 0.4331247
## ENSMUSG00000095250.1  0.161777870 -5.743961  1.26925806 0.2057812 0.5500965
##                               B
## ENSMUSG00000079800.2 -6.030338
## ENSMUSG00000095092.1 -4.802441
## ENSMUSG00000079192.2 -5.922852
## ENSMUSG00000079794.2 -5.602003
## ENSMUSG00000094799.1 -4.721735
## ENSMUSG00000095250.1 -5.003934
```

```
head(de_results2)
```

```
##                         source type bp_length phase              gene_id
## ENSMUSG00000079800.2 ENSEMBL gene      1271    NA ENSMUSG00000079800.2
## ENSMUSG00000095092.1 ENSEMBL gene       366    NA ENSMUSG00000095092.1
## ENSMUSG00000079192.2 ENSEMBL gene       255    NA ENSMUSG00000079192.2
## ENSMUSG00000079794.2 ENSEMBL gene       255    NA ENSMUSG00000079794.2
## ENSMUSG00000094799.1 ENSEMBL gene       366    NA ENSMUSG00000094799.1
## ENSMUSG00000095250.1 ENSEMBL gene       366    NA ENSMUSG00000095250.1
##                            gene_type  gene_name level mgi_id havana_gene  tag
## ENSMUSG00000079800.2 protein_coding AC125149.3     3   <NA>        <NA> <NA>
## ENSMUSG00000095092.1 protein_coding AC125149.5     3   <NA>        <NA> <NA>
## ENSMUSG00000079192.2 protein_coding AC125149.1     3   <NA>        <NA> <NA>
## ENSMUSG00000079794.2 protein_coding AC125149.2     3   <NA>        <NA> <NA>
## ENSMUSG00000094799.1 protein_coding AC125149.4     3   <NA>        <NA> <NA>
## ENSMUSG00000095250.1 protein_coding AC133103.4     3   <NA>        <NA> <NA>
##                              logFC   AveExpr          t   P.Value adj.P.Val
## ENSMUSG00000079800.2  0.31110646 -6.178024  2.1146628 0.0356913 0.2830508
## ENSMUSG00000095092.1  0.07487521 -6.268930  0.7457858 0.4566684 0.7345529
## ENSMUSG00000079192.2 -0.10512004 -6.215590 -0.9378334 0.3494559 0.6508992
## ENSMUSG00000079794.2  0.11525556 -6.185644  0.8140742 0.4165661 0.7038785
## ENSMUSG00000094799.1 -0.21946569 -5.732321 -0.9205291 0.3584001 0.6581476
```

```
## ENSMUSG00000095250.1 -0.26389307 -5.743961 -1.3038389 0.1937803 0.4971736
##                                B
## ENSMUSG00000079800.2 -3.643142
## ENSMUSG00000095092.1 -5.325141
## ENSMUSG00000079192.2 -5.168688
## ENSMUSG00000079794.2 -5.233149
## ENSMUSG00000094799.1 -4.972487
## ENSMUSG00000095250.1 -4.692043
```

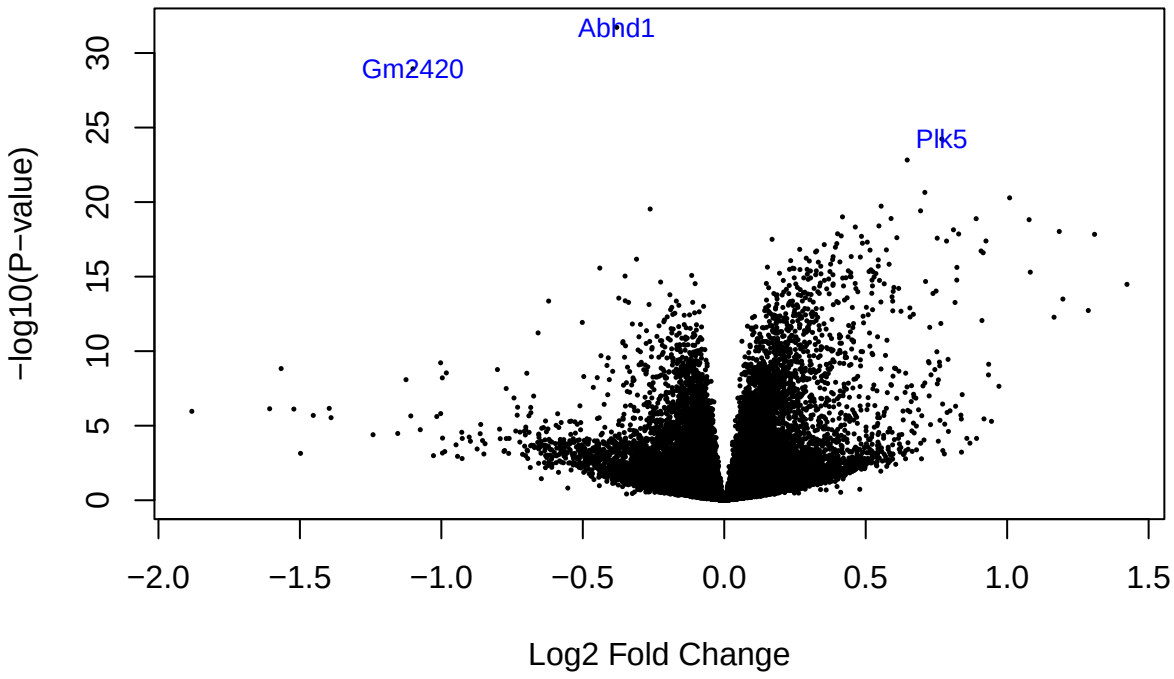
```r
# Differentially expressed genes
table(de_results$adj.P.Val < 0.05)
```

```
##
## FALSE  TRUE
## 47196  8225
```

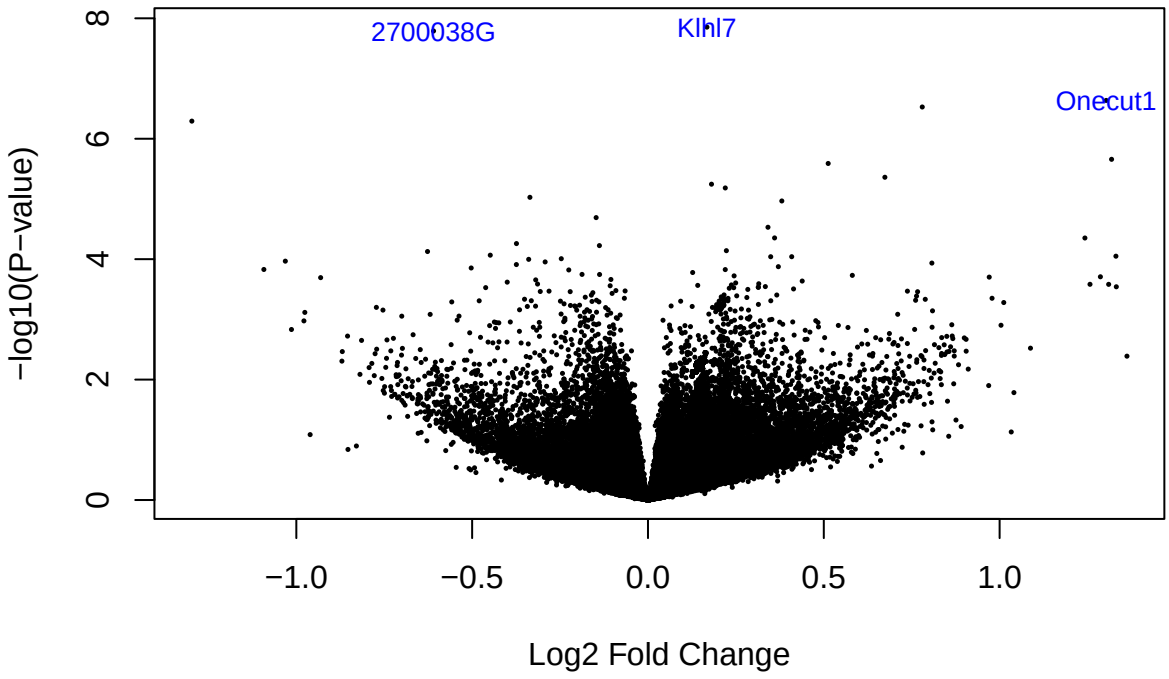
```r
# Differentially expressed genes
table(de_results2$adj.P.Val < 0.05)
```

```
##
## FALSE  TRUE
## 55409    12
```

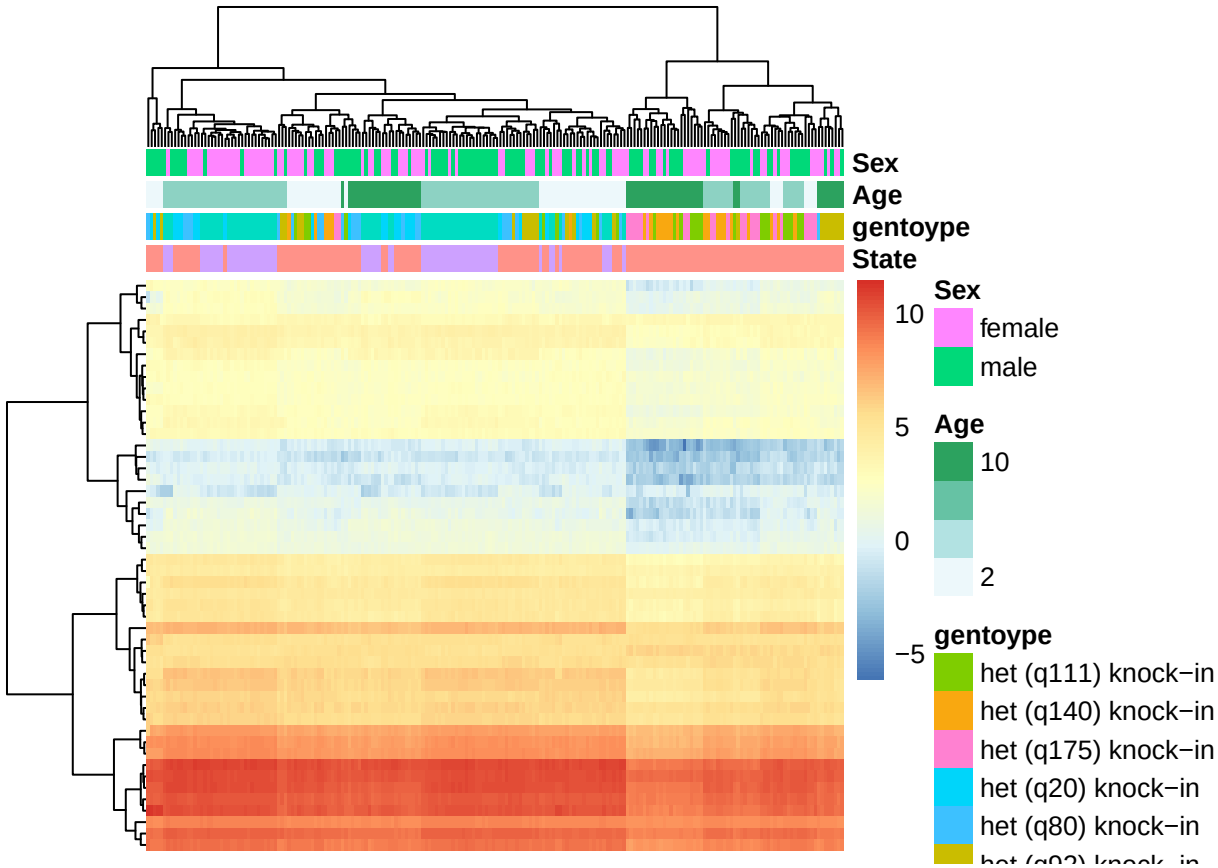
```r
# Performing volcano plot
volcanoplot(eb_results, coef = 2, highlight = 3, names = de_results$gene_name)
```



```
volcanoplot(eb_results2, coef = 2, highlight = 3, names = de_results$gene_name)
```




```
# Performing cluster heatmap

# Extracting over/sub expressed genes
exprs_heatmap <- vGene$E[rank(de_results$adj.P.Val) <= 50, ]

df <- as.data.frame(colData(rse_gene_SRP053398)[, c("state", "sra_attribute.genotype", "sra_attribute.ag
colnames(df) <- c("State", "gentoype", "Age", "Sex")

library("pheatmap")
pheatmap(
    exprs_heatmap,
    cluster_rows = TRUE,
    cluster_cols = TRUE,
    show_rownames = FALSE,
    show_colnames = FALSE,
    annotation_col = df
)
```

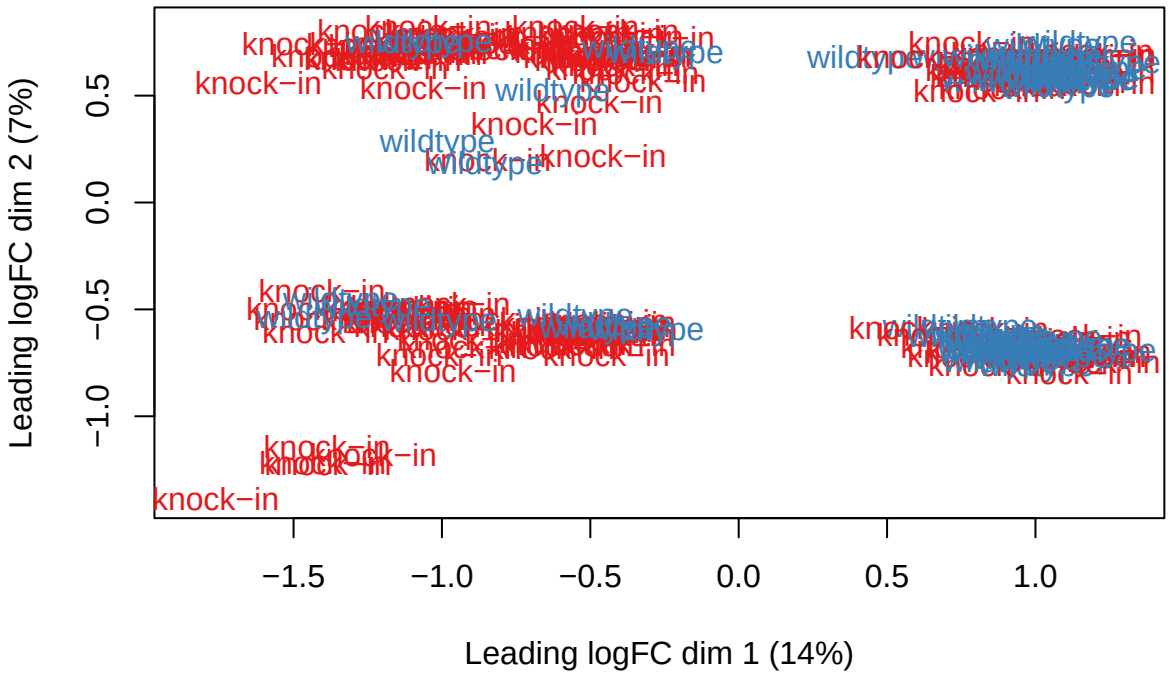


Multidimensional scaling

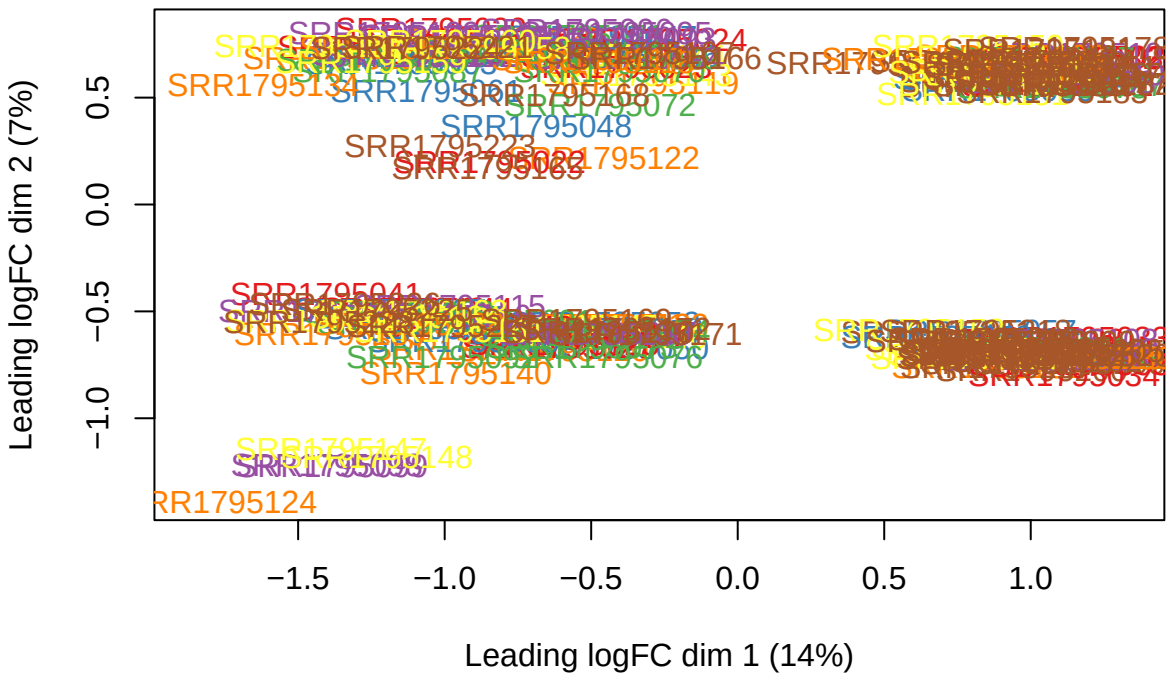
```
col.group <- df$State
levels(col.group) <- brewer.pal(nlevels(col.group), "Set1")
```

```
## Warning in brewer.pal(nlevels(col.group), "Set1"): minimal value for n is 3, returning requested pal
```

```
col.group <- as.character(col.group)

plotMDS(vGene$E, labels = df$State, col = col.group)
```



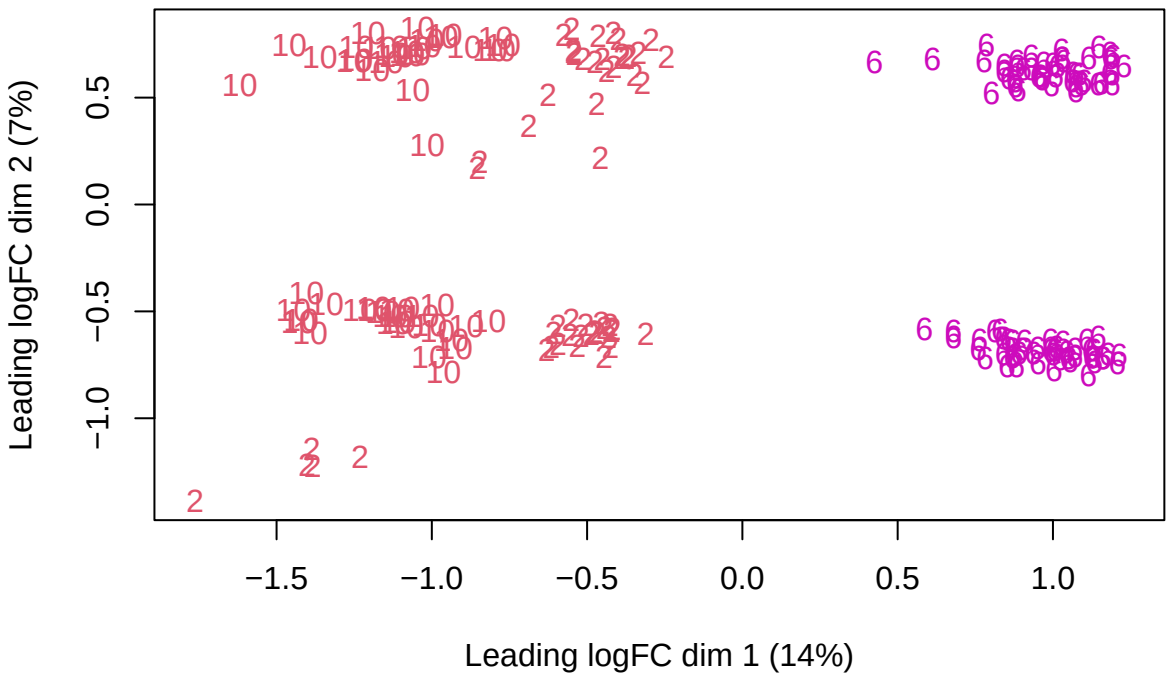

```
col.group2 <- df$gentoype
levels(col.group2) <- brewer.pal(nlevels(col.group2), "Set1")

col.group2 <- as.character(col.group2)

plotMDS(vGene2$E, labels = df$genotype, col = col.group2)
```




```
col.group <- df$Age
levels(col.group) <- brewer.pal(nlevels(col.group), "Set1")
```

```
## Warning in brewer.pal(nlevels(col.group), "Set1"): minimal value for n is 3, returning requested pal
```

```
col.group <- as.character(col.group)

plotMDS(vGene$E, labels = df$Age, col = col.group)
```



Biological discussion

The analysis of the data shows that there are differences in the expression of both models. In the first model only comparing the state of the mice, there were more differential expressed genes than the model that considers also the knock-in mice CAG repeats lengths. In the first model, there was a clear separation in the MDS Plot, that indicates that shows the effect of the knock in. In the second model the effect of the CAG repeats lengths has an impact more specific in some genes like the Onecut1 or Klf7, that suggests that some genes are regulated in a dependant way of the CAG repeats lengths. In conclusion the knock-in the mutation of the Htt has a strong effect in gene expression, that could be reflected in the modifications of multiple biological pathways, like inflammation, oxidative stress or neurodegeneration. The CAG repeats lengths has a less impact in the gene expression, leading to a more gradual or restricted effect in specific processes. These few overexpressed genes and subexpressed genes, could be related to the severity of the disease. In the study already performed by Langelder et al. (2016), (Langfelder et al. 2016) they mentioned that the CAG length and age are highly correlated with 13 striatal and 5 cortical modules, implicated in the dysregulation of cyclic AMP signaling, cell death and procadherin genes.

Langfelder, Peter, Jeffrey P Cantle, Doxa Chatzopoulou, Nan Wang, Fuying Gao, Ismael Al-Ramahi, Xiao-Hong Lu, et al. 2016. “Integrated Genomics and Proteomics Define Huntingtin CAG Length-Dependent Networks in Mice.” *Nature Neuroscience* 19 (4): 623–33. https://doi.org/10.1038/nn.4256.