Sequencing: RNA-seq data intro

Koen Van den Berge

Last compiled on 29 November, 2021

Contents

In this lecture we will start working with a real bulk RNA-seq dataset from Haglund *et al.* (2012). After importing the data, we will be working our way through four major challenges which, together, will form a full RNA-seq differential expression (DE) analysis pipeline where the result of our analysis will be a(n ordered) list of genes that we find to be differently expressed between our conditions of interest. The four main challenges we will look into are

- Choice of modeling assumptions (distribution).
- Normalization.
- Parameter estimation under a limited information setting.
- Statistical inference under high dimensionality (many genes).

1 Experimental design, data import and data exploration

1.1 Experimental design

Let's try to work out the experimental design using the following paragraph from the Methods section of the paper.

> Tissue for cell culturing was obtained from four chief cell parathyroid adenomas collected directly at surgery from female postmenopausal patients. Isolation of cells and culturing were performed essentially as previously described (22). Cells were plated and treated with 100 nM DPN (Tocris Bioscience, Minneapolis, MN) or 100 nM OHT (Sigma-Aldrich, St. Louis, MO) for 24 or 48 h, respectively. Untreated cells cultured in parallel were used as controls. Cells were harvested in RNeasy (QIAGEN AB, Hilden, Germany), and quality control was performed using Bioanalyzer (Agilent Technologies, Santa Clara, CA) and Nanodrop (Nanodrop Technology, Wilmington, DE) for all specimens. RNA samples were isolated from four different adenomas, and one sample (case 4, control 24 h) was omitted before transcriptome sequencing based on low RIN value. The entire sample set used for sequencing consisted of the treatment groups DPN 24 h (n = 4), DPN 48 h (n = 4), OHT 24 h (n = 4), OHT 48 h (n = 4), control 24 h (n = 3), and control 48 h (n = 4).

Figure 1: Figure: A paragraph from the Methods section.

1.2 Data import and exploration

We will be importing the dataset using the parathyroidSE data package from Bioconductor.

```
if (!requireNamespace("BiocManager", quietly = TRUE)){
  install.packages("BiocManager")
}
if(!"SummarizedExperiment" %in% installed.packages()[,1]){
  BiocManager::install("SummarizedExperiment")
}
# install package if not installed.
if(!"parathyroidSE" %in% installed.packages()[,1]) BiocManager::install("parathyroidSE")
```

```
suppressPackageStartupMessages({
  library(parathyroidSE)
  library(SummarizedExperiment)
})

# import data
data("parathyroidGenesSE", package="parathyroidSE")
# rename for convenience
se1 <- parathyroidGenesSE
rm(parathyroidGenesSE)

# three treatments
treatment1 <- colData(se1)$treatment
table(treatment1)
```

```
## treatment1
## Control     DPN     OHT
##       7      10      10
```

```
# two timepoints
time1 <- colData(se1)$time
table(time1)
```

```
## time1
## 24h 48h
##  13  14
```

```
# four donor patients
patient1 <- colData(se1)$patient
table(patient1)
```

```
## patient1
## 1 2 3 4
## 6 8 6 7
```

```
table(patient1, treatment1, time1)
```

```
## , , time1 = 24h
##
##         treatment1
## patient1 Control DPN OHT
##        1       1   1   1
##        2       1   2   2
##        3       1   1   1
##        4       0   1   1
##
## , , time1 = 48h
##
##         treatment1
## patient1 Control DPN OHT
##        1       1   1   1
```

```
##        2       1   1   1
##        3       1   1   1
##        4       1   2   2
```

- We observe that the number of samples that we are observing here is larger than what is described in the paper. As also described in the parathyroidSE vignette, some samples were spread over multiple sequencing runs (i.e., the same sample being sequenced repeatedly) and therefore constitute **technical replication**, rather than biological replication.
- We have previously seen that technical replicates can be considered to be distributed according to a Poisson distribution. One important property of Poisson random variables is that a sum of Poisson random variables still follow a Poisson distribution. Indeed, if $X \sim Poi(\mu_X)$ and $Y \sim Poi(\mu_Y)$, then $X + Y = Z \sim Poi(\mu_X + \mu_Y)$.
- For this reason, it is often suggested to sum technical replicates rather than, for example, averaging, which does not retain the Poisson property (try for yourself!). We'll therefore first sum the technical replicates.

```
dupExps <- as.character(colData(se1)$experiment[duplicated(colData(se1)$experiment)])
dupExps
```

```
## [1] "SRX140511" "SRX140513" "SRX140523" "SRX140525"
```

```
counts <- assays(se1)$counts
newCounts <- counts
cd <- colData(se1)
for(ss in 1:length(dupExps)){
  # check which samples are duplicates
  relevantId <- which(colData(se1)$experiment == dupExps[ss])
  # sum counts
  newCounts[,relevantId[1]] <- rowSums(counts[,relevantId])
  # keep which columns / rows to remove.
  if(ss == 1){
    toRemove <- relevantId[2]
  } else {
    toRemove <- c(toRemove, relevantId[2])
  }
}

# remove after summing counts (otherwise IDs get mixed up)
newCounts <- newCounts[,-toRemove]
newCD <- cd[-toRemove,]

# Create new SummarizedExperiment
se <- SummarizedExperiment(assays = list("counts" = newCounts),
                           colData = newCD,
                           metadata = metadata(se1))

treatment <- colData(se)$treatment
table(treatment)
```

```
## treatment
## Control     DPN     OHT
##       7       8       8
```

```
time <- colData(se)$time
table(time)
```

```
## time
## 24h 48h
##  11  12
```

```
patient <- colData(se)$patient
table(patient)
```

```
## patient
## 1 2 3 4
## 6 6 6 5
```

```
table(patient, treatment, time) # agrees with paper.
```

```
## , , time = 24h
##
##         treatment
## patient Control DPN OHT
##       1       1   1   1
##       2       1   1   1
##       3       1   1   1
##       4       0   1   1
##
## , , time = 48h
##
##         treatment
## patient Control DPN OHT
##       1       1   1   1
##       2       1   1   1
##       3       1   1   1
##       4       1   1   1
```

- After summing the technical replicates and appropriately updating the sample information, we again create a `SummarizedExperiment` object, which is essentially a data container that contains all relevant information about your experiment. Please see the vignette for more information on how to use this class.
- By directly matching columns (samples) and rows (genes) to their relevant metadata, the `SummarizedExperiment` class avoids mistakes by mis-matching columns and rows with each other (provided you haven't mismatched them when you create the object).
- The `SummarizedExperiment` class is modular and extendable, and extensions exist for example for the analysis of single-cell RNA-seq data, i.e., the `SingleCellExperiment` class.
- Due to their convenient organization and widely supported usage within Bioconductor, we will typically work with such classes in the analysis of RNA-seq data.

1.3 Independent filtering

Independent filtering is a strategy to remove features (in this case, genes) prior to the analysis. Removal of these features may lower the multiple testing correction for other genes that pass the filter. We try to

remove genes that have a low power to be found statistically significant, and/or that are biologically less relevant. A common filtering strategy is to remove genes with a generally low expression, as low counts have lower relative uncertainty (hence lower statistical power), and may be considered biologically less relevant.

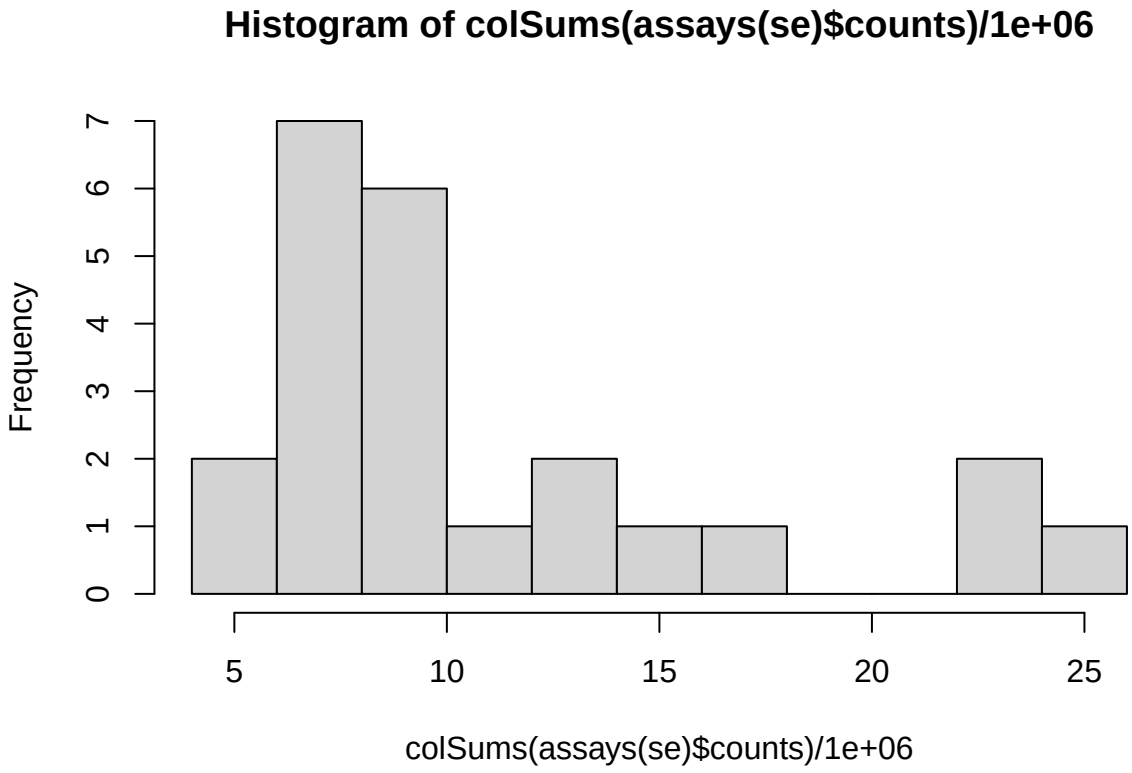
```
suppressPackageStartupMessages({
  library(limma)
  library(edgeR)
})

keep <- rowSums(cpm(se) > 2) >= 3
table(keep)
```

```
## keep
## FALSE  TRUE
## 47837 15356
```

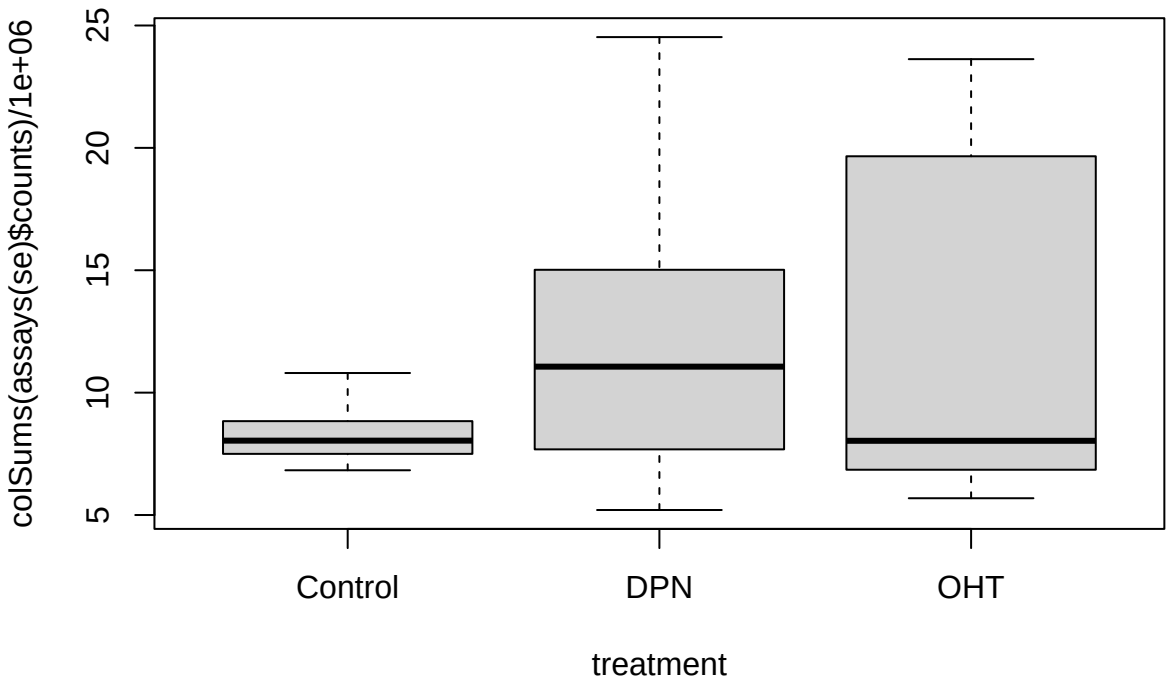
```
se <- se[keep,]
```

1.4 Data exploration

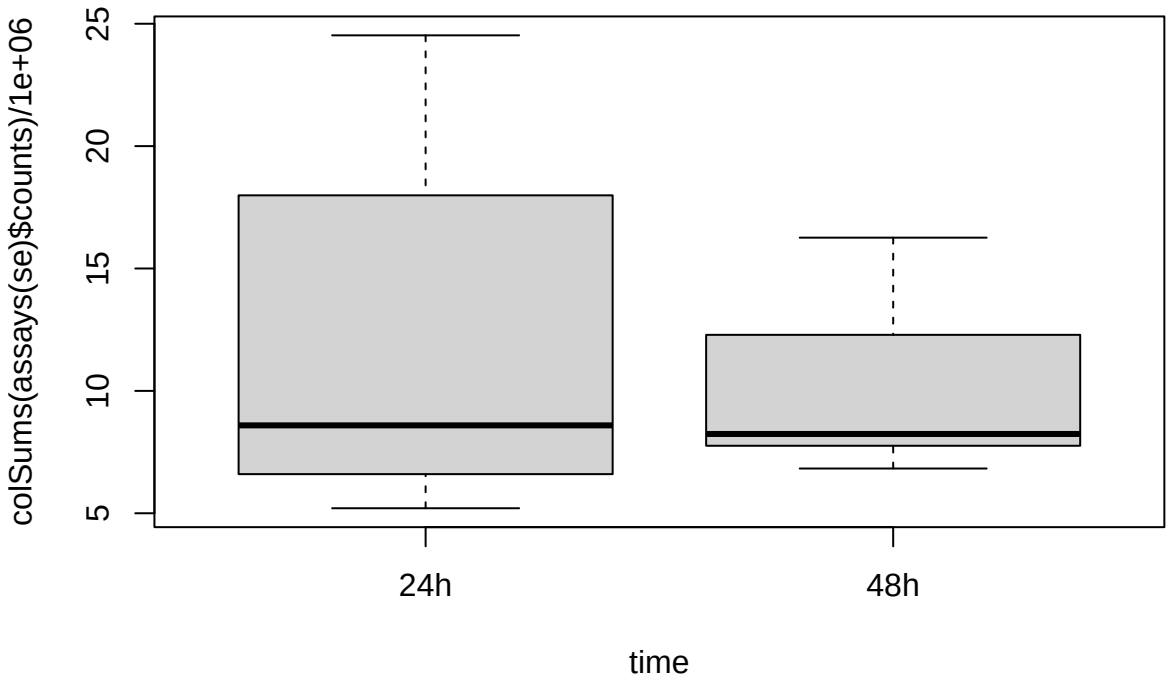
```
# library size distribution
hist(colSums(assays(se)$counts)/1e6, breaks=10)
```



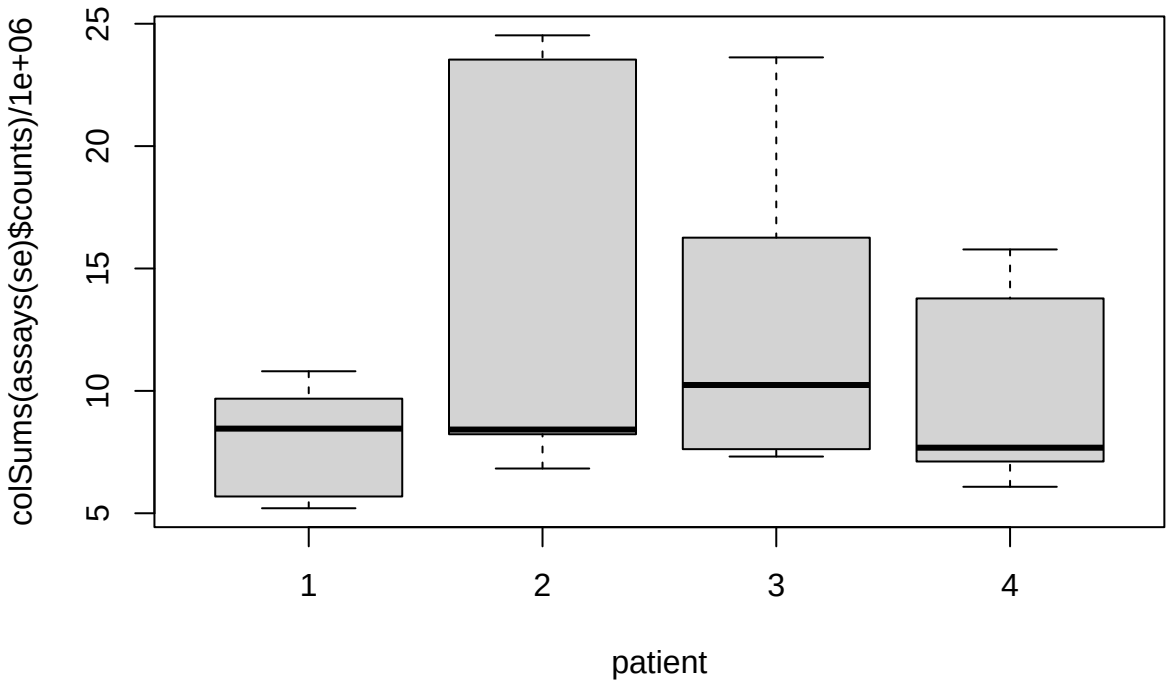
```
boxplot(colSums(assays(se)$counts)/1e6 ~ treatment)
```



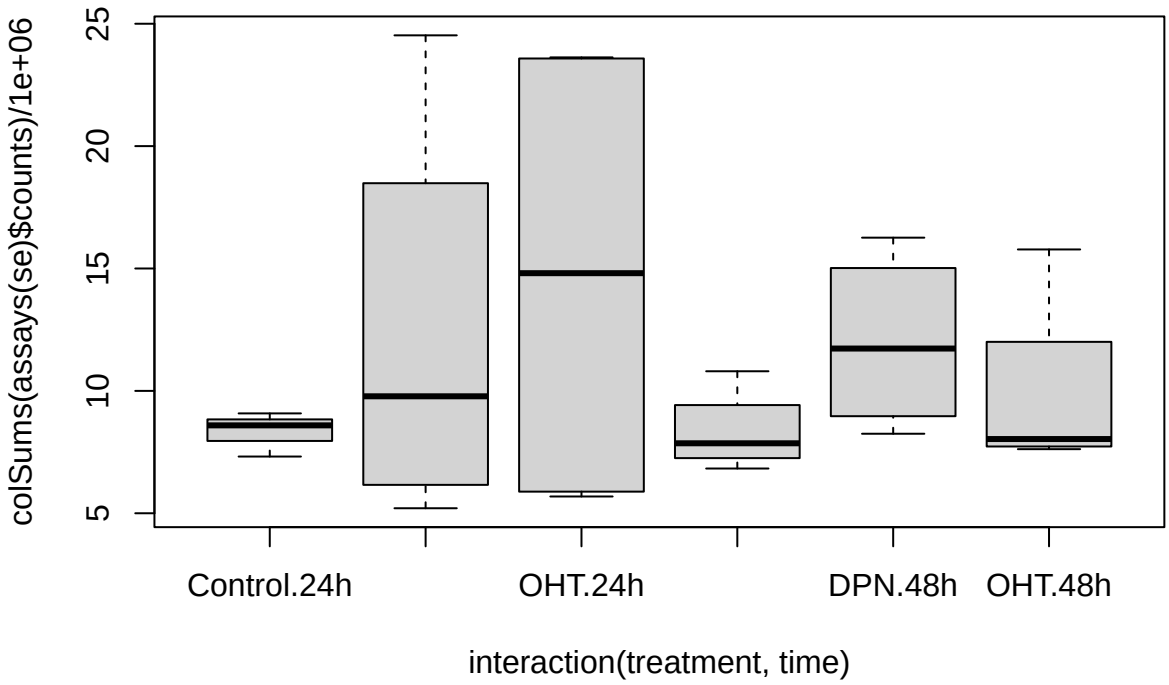

```
boxplot(colSums(assays(se)$counts)/1e6 ~ time)
```



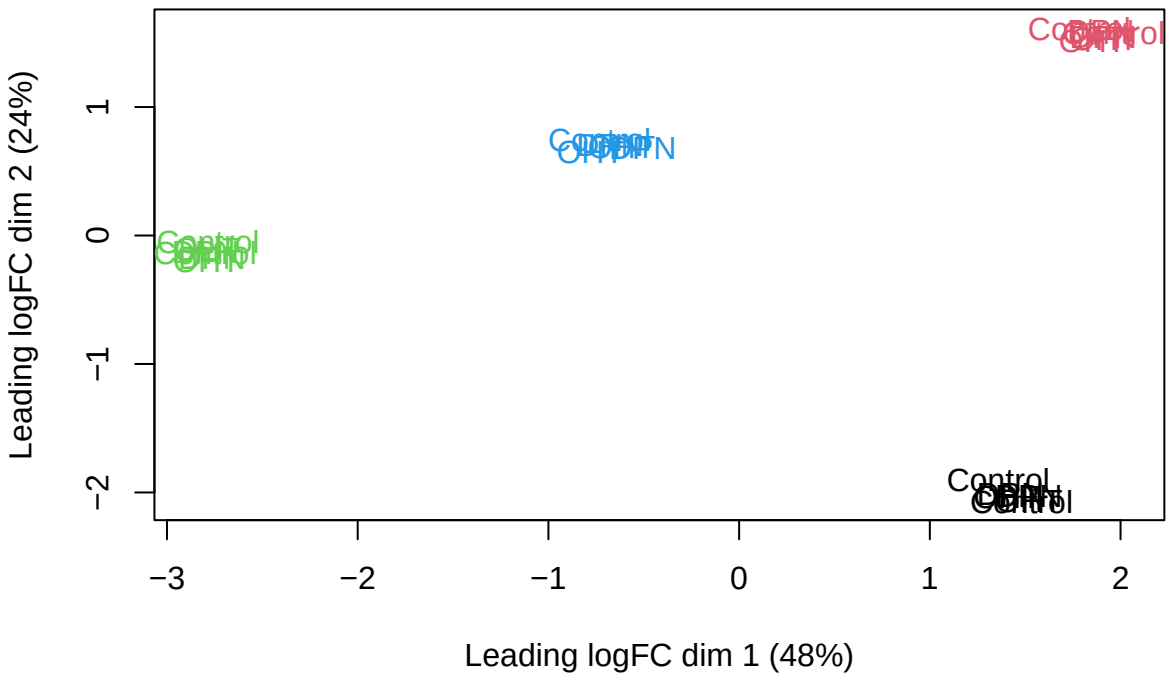

```
boxplot(colSums(assays(se)$counts)/1e6 ~ patient)
```



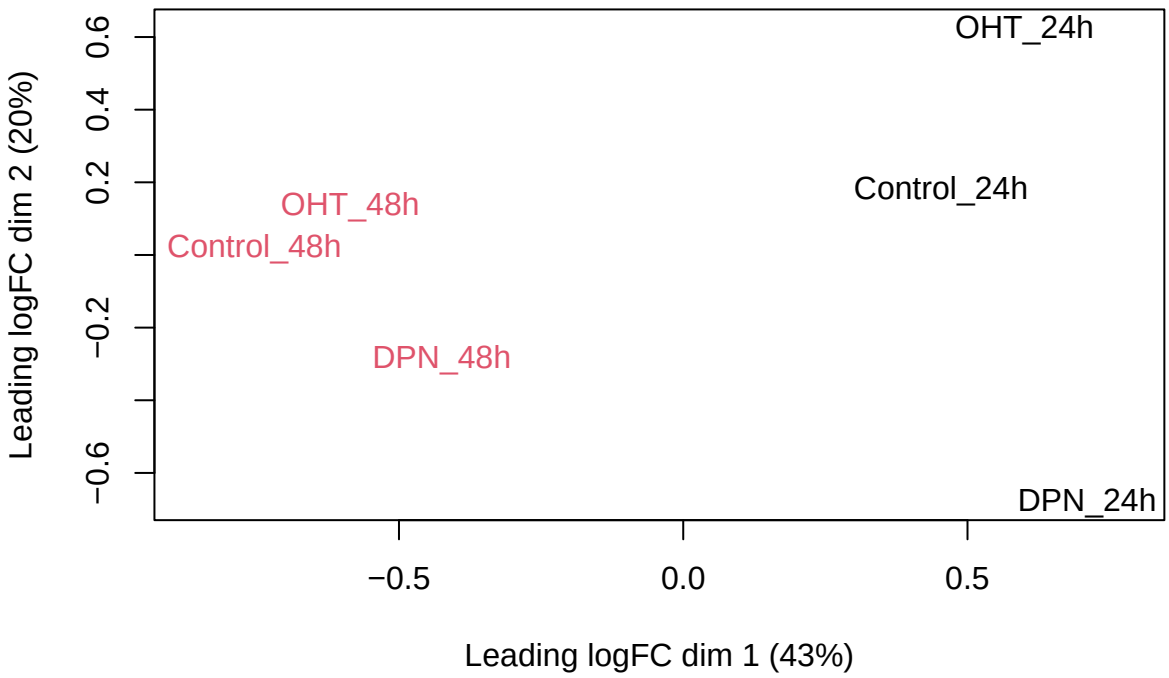

```
boxplot(colSums(assays(se)$counts)/1e6 ~ interaction(treatment, time))
```

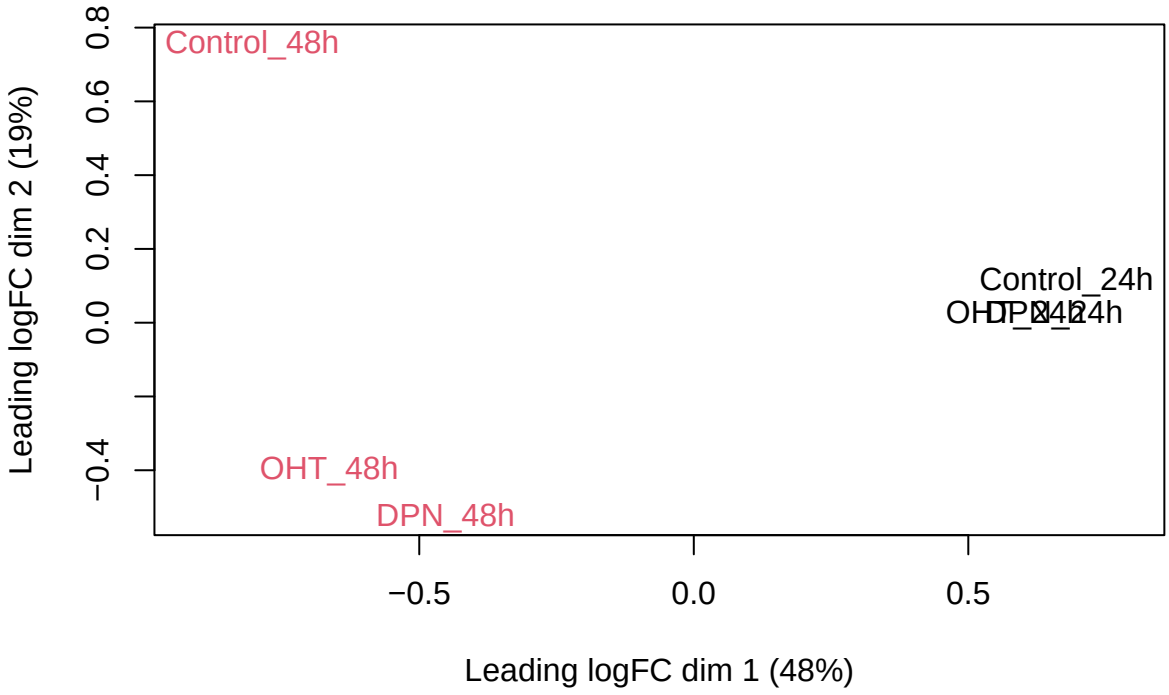


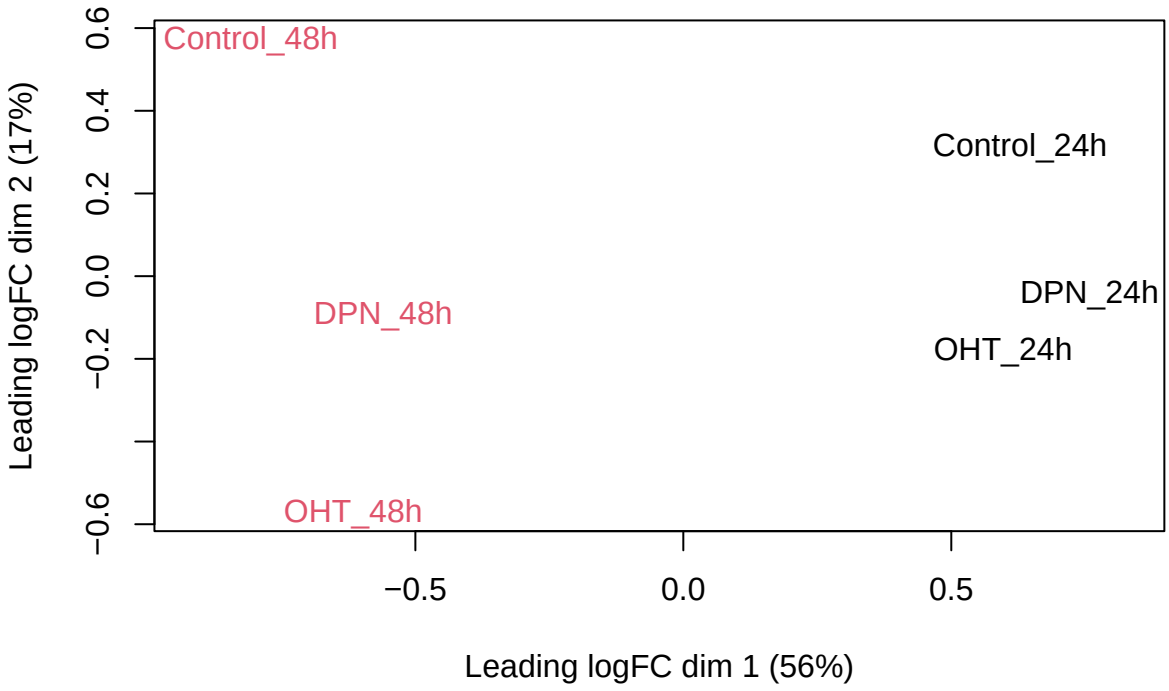

```
# MDS plot
plotMDS(se,
        labels = treatment,
        col=as.numeric(patient))
```

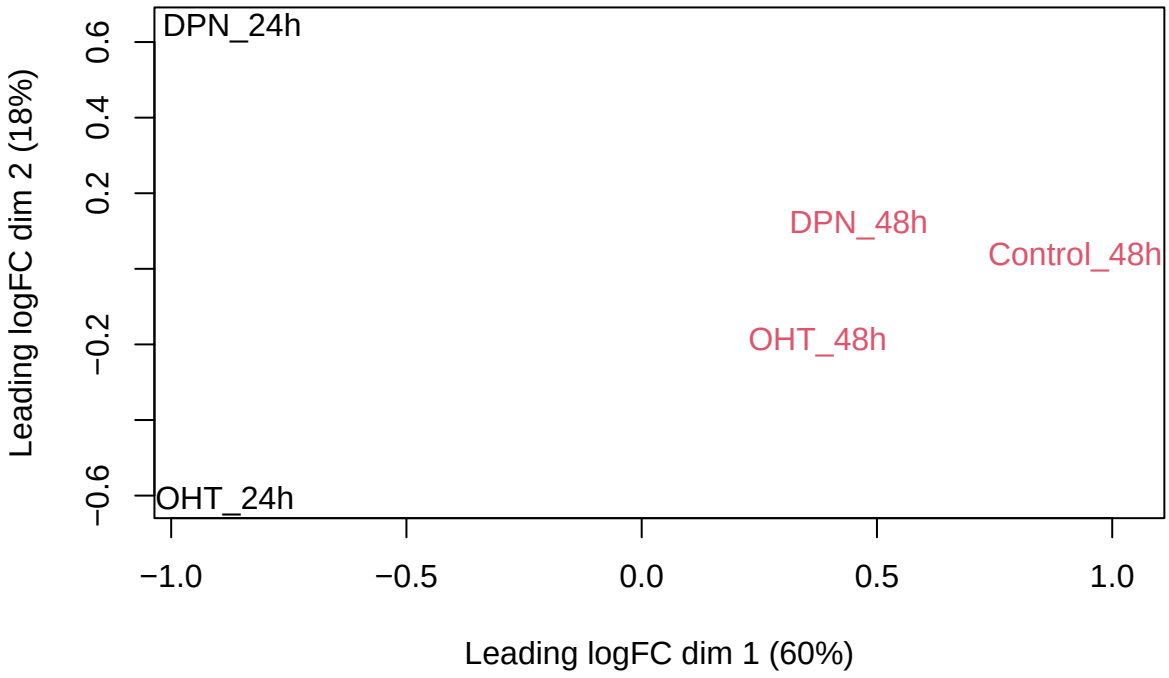



```
## hard to see influence of experimental factors due to large between-patient variation
## we could also make an MDS plot per patient to take a look.
for(kk in 1:4){
  id <- which(patient == kk)
  plotMDS(se[,id],
        labels = paste0(treatment[id],"_",time[id]),
        col=as.numeric(time[id]))
}
```


OHT_24h
Control_24h
OHT_48h
Control_48h
DPN_48h
DPN_24h
0.6
0.4
0.2
−0.2
−0.6
−0.5
0.0
0.5
Leading logFC dim 2 (20%)
Leading logFC dim 1 (43%)


Control_48h
Control_24h
OHT_48h
DPN_48h
Leading logFC dim 2 (19%)
0.8
0.6
0.4
0.2
0.0
−0.4
−0.5
0.0
0.5
Leading logFC dim 1 (48%)


Control_48h
Control_24h
DPN_48h
DPN_24h
OHT_24h
OHT_48h
Leading logFC dim 2 (17%)
0.6
0.4
0.2
0.0
−0.2
−0.6
−0.5
0.0
0.5
Leading logFC dim 1 (56%)



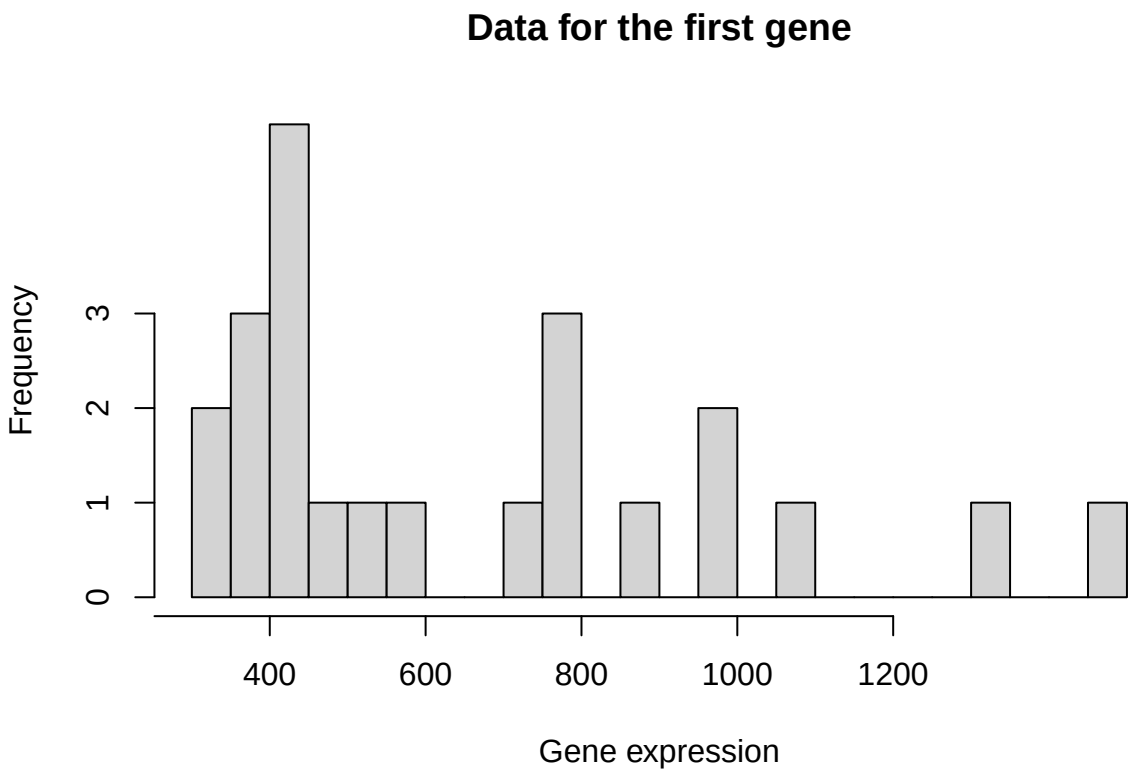
Observations based on MDS plot:

- There is a very large between-patient variability, which is the major source of variation in this dataset. The samples from each patient cluster together tightly.
- Within patient, time consistently explains more variation than the treatments.
- Relative to patients and time, the treatment seems to have a fairly small effect.

2 Challenge I: Choice of modeling assumptions

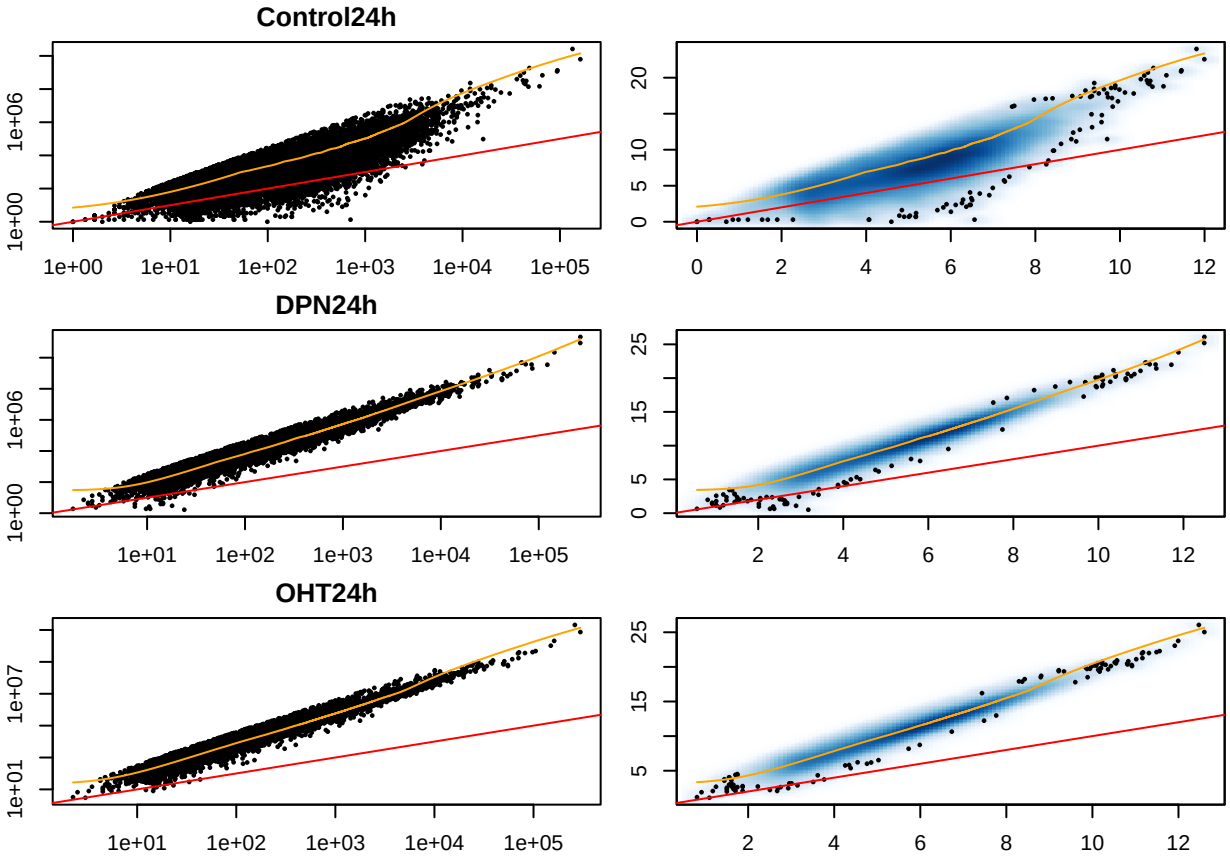
When working with a GLM, as part of the choices of modeling assumptions, we need to pick an appropriate distribution for the expression counts. Below we perform some exploratory analyses to investigate.

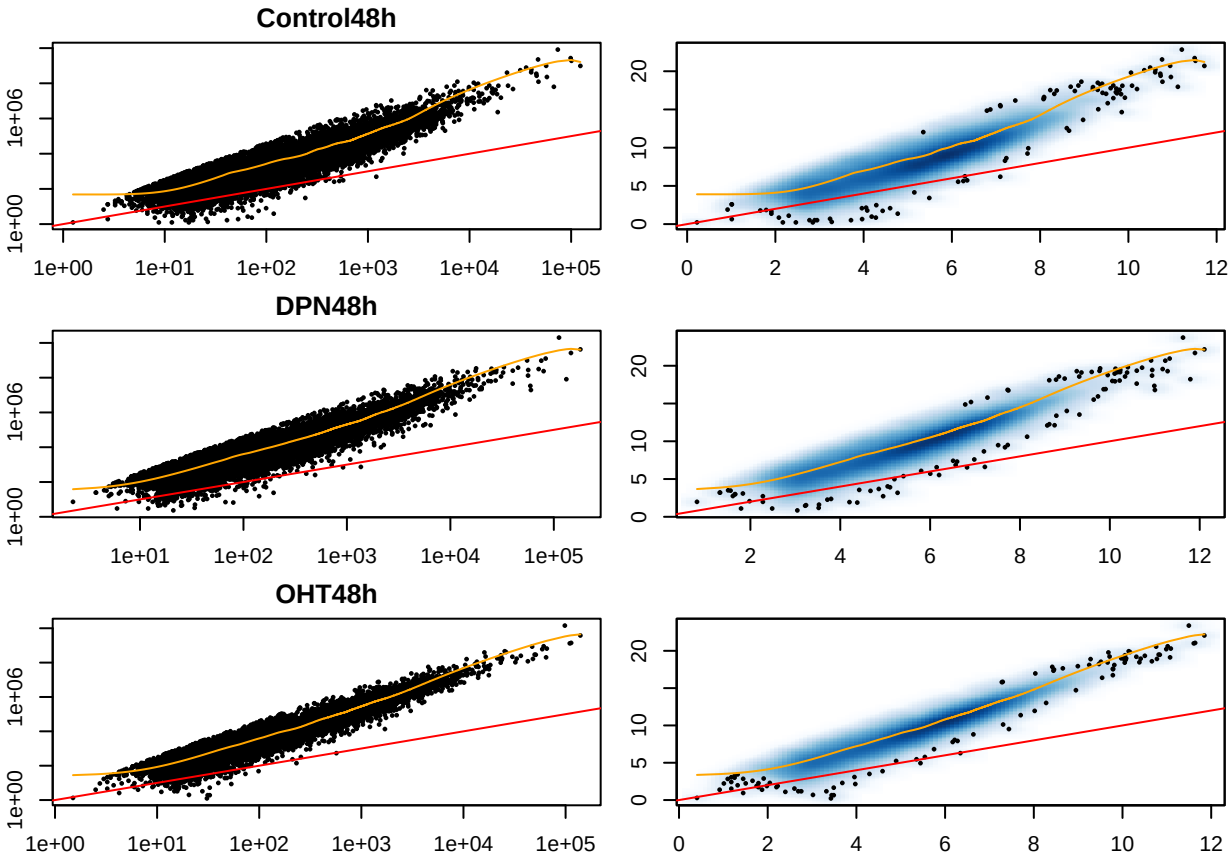
```
y <- assays(se)$counts[1,]
hist(y, breaks = 40,
     xlab = "Gene expression",
     xaxt = "n", yaxt = "n",
     main = "Data for the first gene")
axis(1, at = seq(200, 1200, by=200))
axis(2, at = 0:3)
```




```
# Mean-variance trend within each experimental condition
cont24ID <- which(treatment == "Control" & time == "24h")
DPN24ID <- which(treatment == "DPN" & time == "24h")
OHT24ID <- which(treatment == "OHT" & time == "24h")
cont48ID <- which(treatment == "Control" & time == "48h")
DPN48ID <- which(treatment == "DPN" & time == "48h")
OHT48ID <- which(treatment == "OHT" & time == "48h")
idList <- list(cont24ID, DPN24ID, OHT24ID,
               cont48ID, DPN48ID, OHT48ID)
names(idList) <- paste0(rep(levels(treatment),2), rep(levels(time), each=3))

par(mfrow=c(3,2), mar=c(2,2,2,1))
for(kk in 1:length(idList)){
  # extract counts for each condition
  curCounts <- assays(se)$counts[,idList[[kk]]]
  plot(x = rowMeans(curCounts)+1,
       y = rowVars(curCounts)+1,
       pch = 16, cex=1/2,
       xlab = "Mean", ylab="Variance",
       main = names(idList)[kk],
       log="xy")
  abline(0,1, col="red")
  lw1 <- loess((rowVars(curCounts)+1) ~ (rowMeans(curCounts)+1), span=1/4, lwd=3)
  oo <- order(rowMeans(curCounts)+1)
  lines(rowMeans(curCounts)[oo]+1, lw1$fitted[oo], col="orange")
```

```
  smoothScatter(x = log1p(rowMeans(curCounts)),
       y = log1p(rowVars(curCounts)),
       pch = 16, cex=1/2,
       xlab = "Mean", ylab="Variance")
  abline(0,1, col="red")
  lines(log(rowMeans(curCounts)[oo]+1), log(lw1$fitted[oo]+1), col="orange")
}
```





- Having data on thousands of genes provides the opportunity to empirically assess the mean-variance relationship.
- It is clear that the data is overdispersed with respect to the Poisson distribution (red $y = x$ line). There also seems to be a quadratic trend of the variance as a function of the mean. This has motivated the **negative binomial distribution as the most popular choice to model (bulk) RNA-seq gene expression data**.

- The negative binomial distribution is also referred to as the Gamma-Poisson distribution as it can be formulated as such. Indeed, if

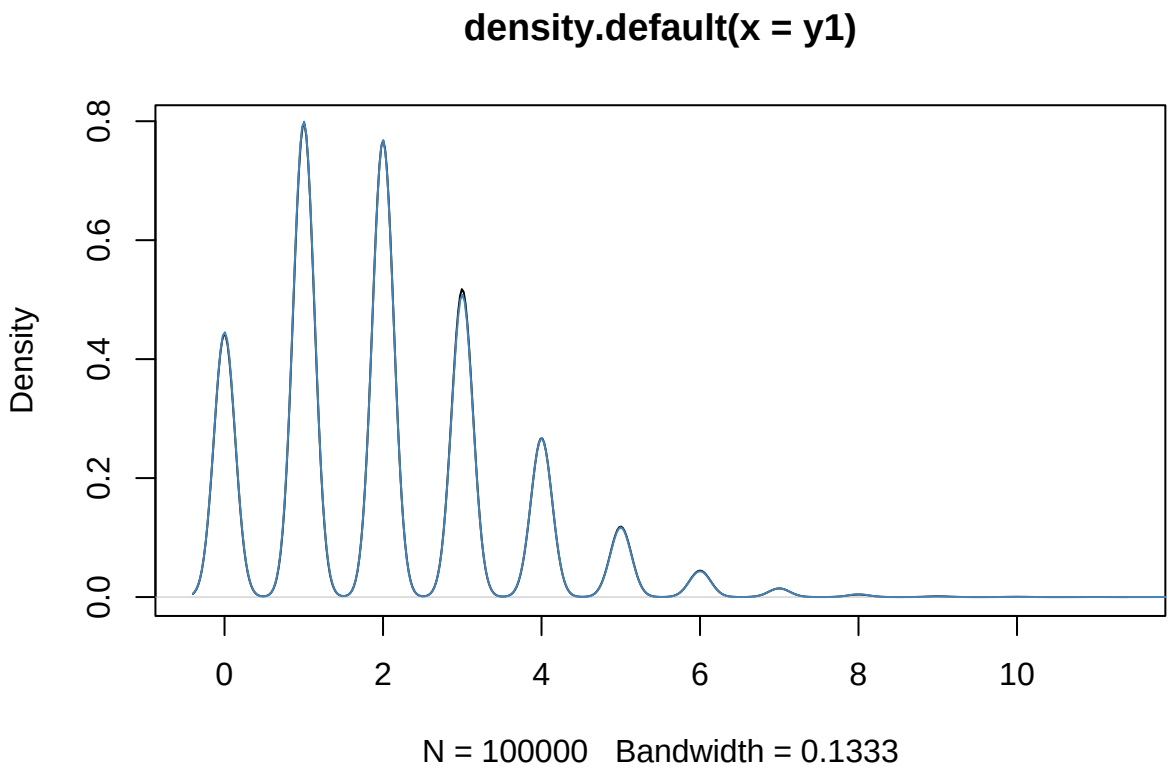
$$\lambda \sim \Gamma(\alpha, \beta) Y|\lambda \sim Poi(\lambda),$$

  then this is equivalent to

$$Y \sim NB(\mu = \alpha/\beta, \phi = 1/\alpha).$$

- This can be shown analytically, but is considered out of the scope of this course. Below, we show it empirically using simulation.
- This theoretical result has got some practical consequences. The Gamma-Poisson formulation makes it clear why we can sum technical replicates as the sum of Poisson random variables is again a Poisson random variable.
- The Poisson statement can thus be considered as capturing technical variation, while the Gamma statement can be considered to capture biological variation, i.e., variation in the mean expression across biological replicates.

```r
alpha <- 20
beta <- 10
lambda <- rgamma(n = 1e5, shape = alpha, rate = beta)
y1 <- rpois(n = 1e5, lambda = lambda)

# note phi = 1 / size
y2 <- rnbinom(n=1e5, mu=alpha / (beta), size=alpha)


plot(density(y1))
lines(density(y2), col="steelblue")
```



3 Challenge II: Normalization

Normalization is necessary to correct for several sources of technical variation:

- **Differences in sequencing depth** between samples. Some samples get sequenced deeper in the sense that they consist of more (mapped) reads and therefore can be considered to contain a higher amount of information, which we should be taking into account. In addition, if a sample is sequenced deeper, it is natural that the counts for each gene will be higher, jeopardizing a direct comparison of the expression counts.
- **Differences in RNA population composition** between samples. As an extreme example, suppose that two samples have been sequenced to the exact same depth. One sample is contaminated and has

a very high concentration of the contaminant cDNA being sequenced, but otherwise the two samples are identical. Since the contaminant will be taking up a significant proportion of the reads being sequenced, the counts will not be directly comparable between the samples. Hence, we may also want to correct for differences in the composition of the RNA population of the samples.

- **Other technical variation** such as sample-specific GC-content or transcript length effects may also be accounted for.

Let's take a look at how comparable different replicates are in the Control condition at 48h in our dataset. We will investigate this using MD-plots (mean-difference plots as introduced by Dudoit *et al.* (2002)), also sometimes referred to as MA-plots.

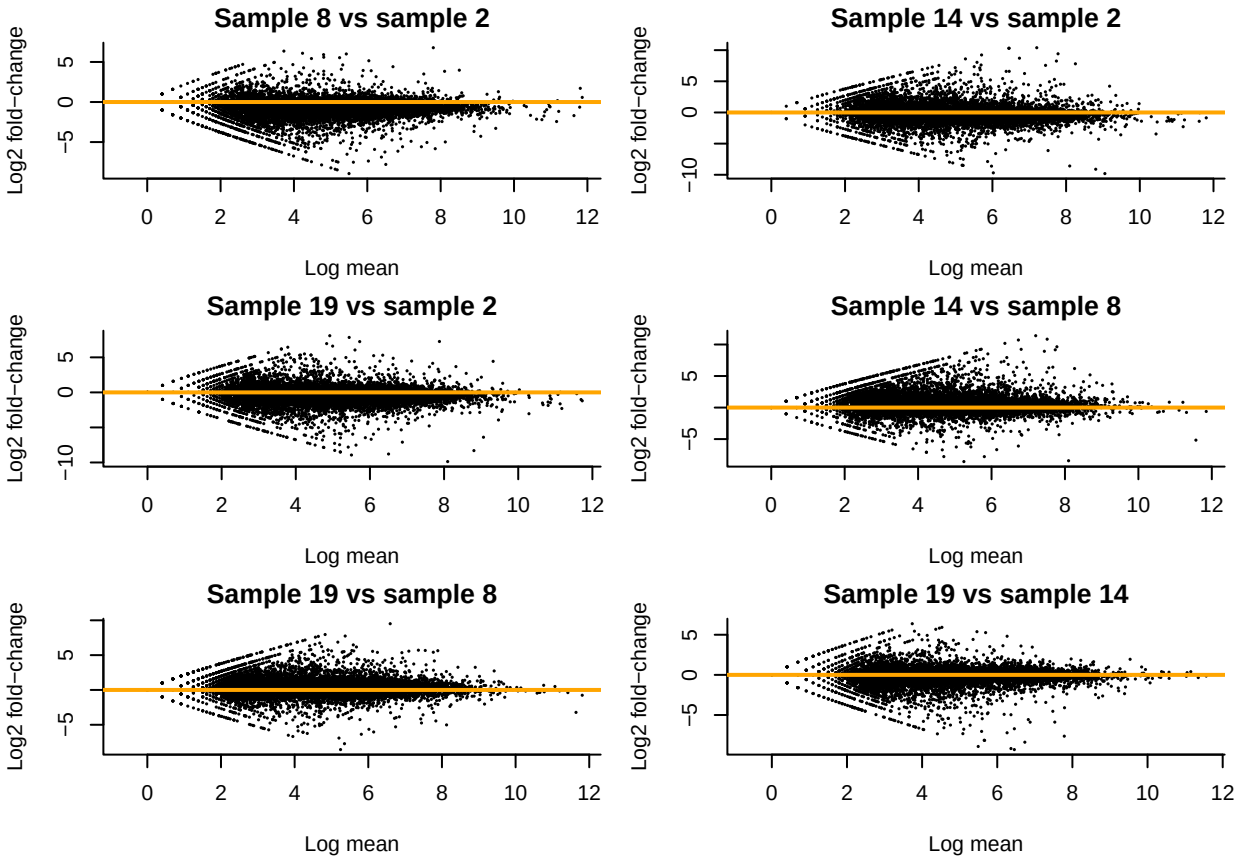
```
cont48ID # relevant samples
```

```
## [1]  2  8 14 19
```

```
colSums(assays(se)$counts[,cont48ID]) / 1e6
```

```
## [1] 10.801193  6.828259  8.038079  7.678421
```

```
combs <- combn(cont48ID, m=2) #pairwise combinations between samples

par(mfrow=c(3,2), mar=c(4,4,2,1))
for(cc in 1:ncol(combs)){
  curSamples <- combs[,cc]
  M <- rowMeans(assays(se)$counts[,curSamples])
  D <- assays(se)$counts[,curSamples[2]] / assays(se)$counts[,curSamples[1]]
  plot(x = log(M), y = log2(D),
       pch = 16, cex=1/3,
       main = paste0("Sample ", curSamples[2], " vs sample ", curSamples[1]),
       xlab = "Log mean", ylab = "Log2 fold-change",
       bty = 'l')
  abline(h = 0, col="orange", lwd=2)
}
```



- We see clear bias for some pairwise comparisons. For example, in the first plot comparing sample 8 versus sample 2, the log fold-changes are biased downwards. This means that, on average, a gene is lower expressed in sample 8 versus sample 2. Looking at the library sizes, we can indeed see that the library size for sample 2 is about 11×10^6 while it is only about 7×10^6 for sample 8! This is a clear library size effect that we should take into account.

3.1 Count scaling versus GLM offsets

- We have previously discussed count scaling transformations such as CPM and TPM.
- A more appropriate and natural way when working with GLMs is through the use of **offsets**. The general use of an offset is to account for the 'effort' performed in order to gather that observation of the response variable. Two examples:

  1. A biologist studying whale migration has one fixed spot where, in the migration season, she counts migrating whales day after day, over several years. For each day she records the number of spotted whales. Of course, the time spent whale-watching may differ from day to day and it is natural that you are more likely to spot more whales if you spend more time looking for them. The time spent spotting whales can then be used as an offset.
  2. In our case, a sample being sequenced deeper contains more information, i.e., more 'effort' has been performed, as compared to a sample being sequenced relatively shallow. We have more confidence of a count from a deeply sequenced sample than from a shallowly sequenced sample. We can therefore use the sequencing depth $N_i = \sum_g Y_{gi}$ as offset in the model.

- Adding an offset to the model is different from adding a new variable to the model. For each new variable we add, we will estimate its average effect β on the response variable. When adding an offset, however, we are implicitly assuming that $\beta = 1$.

- Offsets are typically added on the scale of the linear predictor. Suppose we have a gene g and sample i specific offset O_{gi}, then we can define a negative binomial GLM including the offset as

$$\left\{ \begin{array}{ccc} Y_{gi} & \sim & NB(\mu_{gi}, \phi_g) \\ \log \mu_{gi} & = & \eta_{gi} \\ \eta_{gi} & = & \mathbf{X}_i^T \beta_g + \log(O_{gi}). \end{array} \right.$$

- Please read this page for an intuitive reasoning as to why offsets are preferred over count scaling.

3.2 How to normalize?

Many approaches are available for normalizing RNA-seq data, and we will review a couple of these. Most methods calculate an offset that is used in the GLM used to model gene expression. One notable method, full-quantile normalization, does not calculate an offset, and rather normalizes counts directly, immediately enforcing the same sequencing depth for all samples.

3.2.1 TMM method (default of `edgeR`)

The trimmed mean of M-values (TMM) method introduced by Robinson & Oshlack (2010) is a normalization procedure that calculates a single normalization factor for each sample. As the name suggests, it is based on a trimmed mean of fold-changes (M-values) as the scaling factor. A trimmed mean is an average after removing a set of 'extreme' values. Specifically, TMM calculates a normalization factor $F_i^{(r)}$ across genes g for each sample i as compared to a reference sample r,

$$\log_2(F_i^{(r)}) = \frac{\sum_{g \in G^*} w_{gi}^r M_{gi}^r}{\sum_{g \in G^*} w_{gi}^r},$$

where M_{gi}^r represents the $\log_2$-fold-change of the gene expression fraction as compared to a reference sample r, i.e.,

$$M_{gi}^r = \log_2 \left(\frac{Y_{gi}/N_i}{Y_{gr}/N_r} \right),$$

and w_{gi}^r represents a precision weight calculated as

$$w_{gi}^r = \frac{N_i - Y_{gi}}{N_i Y_{gi}} + \frac{N_r - Y_{gr}}{N_r Y_{gr}},$$

and G^* represents the set of genes after trimming those with the most extreme average expression and fold-change. The weights serve to account for the fact that fold-changes for genes with lower read counts are more variable.

The procedure only takes genes into account where both $Y_{gi} > 0$ and $Y_{gr} > 0$. By default, TMM trims genes with the 30% most extreme M-values and 5% most extreme average gene expression, and chooses as reference r the sample whose upper-quartile is closest to the across-sample average upper-quartile. The normalization factor is then used in conjunction with the library size to calculate an **effective library size**

$$N_i^{eff} = N_i F_i^{(r)}$$

which is used as offset in the GLM. The normalized counts may be given by $\tilde{Y}_{gi} = Y_{gi}/N_i^s$, with size factor

$$N_i^s = \frac{N_i F_i^{(r)}}{\frac{1}{n} \sum_{i=1}^n N_i F_i^{(r)}}.$$

TMM normalization may be performed from the `calcNormFactors` function implemented in `edgeR`:

```
dge <- edgeR::calcNormFactors(se)
dge$samples #normalization factors added to colData
```

```
##          group lib.size norm.factors       run experiment patient treatment
## Sample1      1  9079958    0.9824006 SRR479052  SRX140503       1   Control
## Sample2      1 10801193    0.9730905 SRR479053  SRX140504       1   Control
## Sample3      1  5205034    0.9771680 SRR479054  SRX140505       1       DPN
## Sample4      1  9681399    0.9935090 SRR479055  SRX140506       1       DPN
## Sample5      1  5685671    0.9730827 SRR479056  SRX140507       1       OHT
## Sample6      1  7835627    0.9837862 SRR479057  SRX140508       1       OHT
## Sample7      1  8590615    0.9350218 SRR479058  SRX140509       2   Control
## Sample8      1  6828259    0.9436147 SRR479059  SRX140510       2   Control
## Sample9      1 24525588    0.9445726 SRR479060  SRX140511       2       DPN
## Sample10     1  8248985    0.9320020 SRR479062  SRX140512       2       DPN
## Sample11     1 23535205    0.9384823 SRR479063  SRX140513       2       OHT
## Sample12     1  8228473    0.9311185 SRR479065  SRX140514       2       OHT
## Sample13     1  7317690    1.0567909 SRR479066  SRX140515       3   Control
## Sample14     1  8038079    1.0431594 SRR479067  SRX140516       3   Control
## Sample15     1 12443101    1.0607802 SRR479068  SRX140517       3       DPN
## Sample16     1 16260235    1.0407639 SRR479069  SRX140518       3       DPN
## Sample17     1 23624835    1.0437936 SRR479070  SRX140519       3       OHT
## Sample18     1  7619275    1.0354159 SRR479071  SRX140520       3       OHT
## Sample19     1  7678421    1.0440457 SRR479072  SRX140521       4   Control
## Sample20     1  7114567    1.0541970 SRR479073  SRX140522       4       DPN
## Sample21     1 13777391    1.0365701 SRR479074  SRX140523       4       DPN
## Sample22     1  6081314    1.0611051 SRR479076  SRX140524       4       OHT
## Sample23     1 15778115    1.0414101 SRR479077  SRX140525       4       OHT
##          time submission     study    sample
## Sample1   24h  SRA051611 SRP012167 SRS308865
## Sample2   48h  SRA051611 SRP012167 SRS308866
## Sample3   24h  SRA051611 SRP012167 SRS308867
## Sample4   48h  SRA051611 SRP012167 SRS308868
## Sample5   24h  SRA051611 SRP012167 SRS308869
## Sample6   48h  SRA051611 SRP012167 SRS308870
## Sample7   24h  SRA051611 SRP012167 SRS308871
## Sample8   48h  SRA051611 SRP012167 SRS308872
## Sample9   24h  SRA051611 SRP012167 SRS308873
## Sample10  48h  SRA051611 SRP012167 SRS308874
## Sample11  24h  SRA051611 SRP012167 SRS308875
## Sample12  48h  SRA051611 SRP012167 SRS308876
## Sample13  24h  SRA051611 SRP012167 SRS308877
## Sample14  48h  SRA051611 SRP012167 SRS308878
## Sample15  24h  SRA051611 SRP012167 SRS308879
## Sample16  48h  SRA051611 SRP012167 SRS308880
## Sample17  24h  SRA051611 SRP012167 SRS308881
## Sample18  48h  SRA051611 SRP012167 SRS308882
## Sample19  48h  SRA051611 SRP012167 SRS308883
## Sample20  24h  SRA051611 SRP012167 SRS308884
## Sample21  48h  SRA051611 SRP012167 SRS308885
## Sample22  24h  SRA051611 SRP012167 SRS308886
## Sample23  48h  SRA051611 SRP012167 SRS308887
```

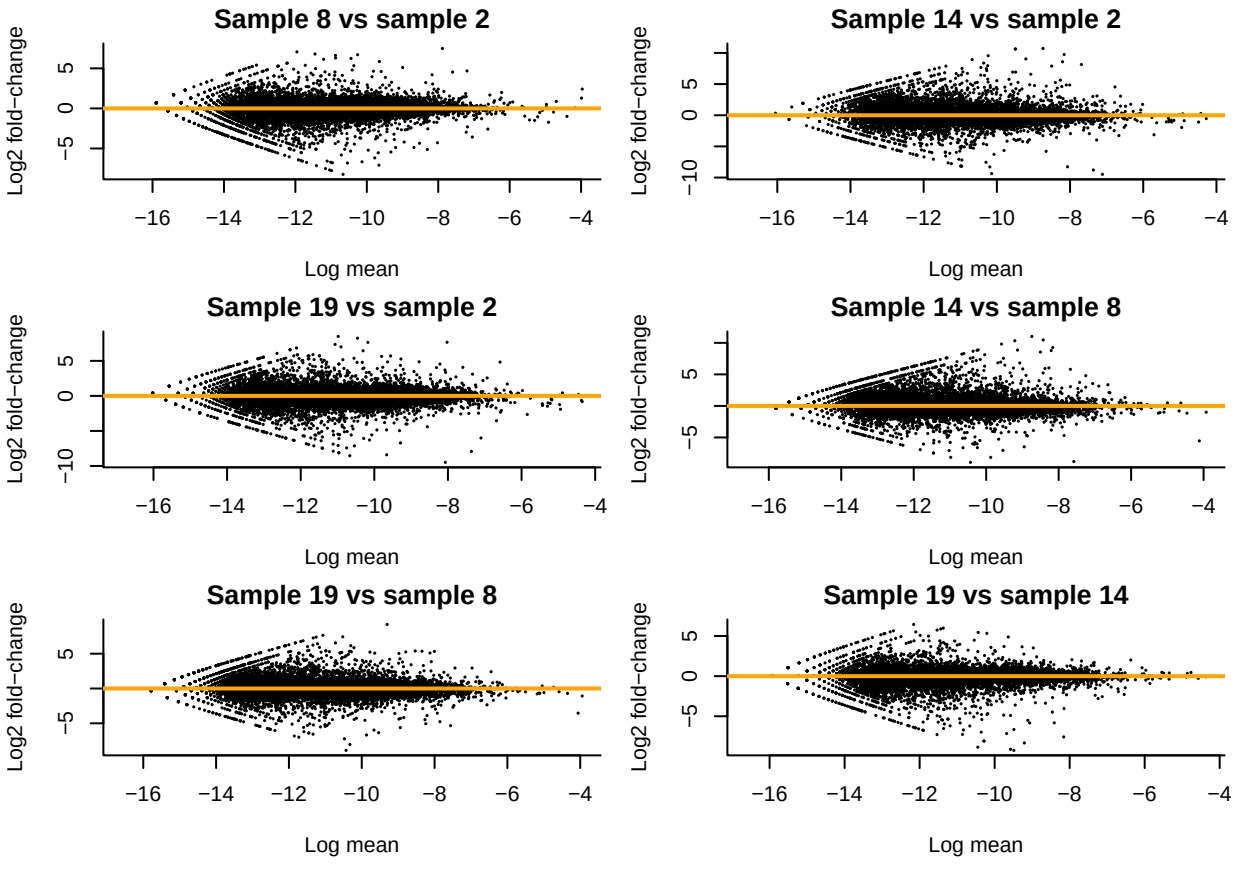
Let's check how our MD-plots look like after normalization. Note that, we can rewrite the GLM as

$$\log\left(\frac{\mu_{gi}}{N_i^s}\right) = \mathbf{X}_i^T \beta_g$$

and so $\frac{\mu_{gi}}{N_i^s}$ can be considered as an 'offset-corrected average count'.

We see that all MD-plots are now nicely centered around a log-fold-change of zero!

```
## normalize
effLibSize <- dge$samples$lib.size * dge$samples$norm.factors
normCountTMM <- sweep(assays(se)$counts, 2, FUN="/", effLibSize)


par(mfrow=c(3,2), mar=c(4,4,2,1))
for(cc in 1:ncol(combs)){
  curSamples <- combs[,cc]
  M <- rowMeans(normCountTMM[,curSamples])
  D <- normCountTMM[,curSamples[2]] / normCountTMM[,curSamples[1]]
  plot(x = log(M), y = log2(D),
       pch = 16, cex=1/3,
       main = paste0("Sample ", curSamples[2], " vs sample ", curSamples[1]),
       xlab = "Log mean", ylab = "Log2 fold-change",
       bty = 'l')
  abline(h = 0, col="orange", lwd=2)
}
```



3.3 Median-of-ratios method (default of DESeq2)

The median-of-ratios method is used in DESeq2 as described in Love *et al.* (2014). It assumes that the expected value $\mu_{gi} = E(Y_{gi})$ is proportional to the true expression of the gene, q_{gi}, scaled by a size factor s_i for each sample,

$$\mu_{gi} = s_i q_{gi}.$$

The size factor s_i is then estimated using the median-of-ratios method compared to a synthetic reference sample r defined based on geometric means of counts across samples

$$s_i = \text{median}_{\{g: Y^*_{gr} \neq 0\}} \frac{Y_{gi}}{Y^*_{gr}},$$

with

$$Y^*_{gr} = \left(\prod_{i=1}^{n} Y_{gi} \right)^{1/n}.$$

We can then use the size factors s_i as offsets to the GLM.

Question. Do you see any issues with the procedure described as such?

Answer.

The procedure relies on genes being expressed in all samples. As sample size increases, the number of genes with this property steadily decreases. This will become crucial in single-cell RNA-seq data analysis.

Median-of-ratios normalization is implemented in the DESeq2 package:

```
dds <- DESeq2::DESeqDataSetFromMatrix(countData = assays(se)$counts,
                                      colData = colData(se),
                                      design = ~ 1) #just add intercept to showcase normalization
```

```
## converting counts to integer mode
```

```
dds <- DESeq2::estimateSizeFactors(dds)
sizeFactors(dds)
```

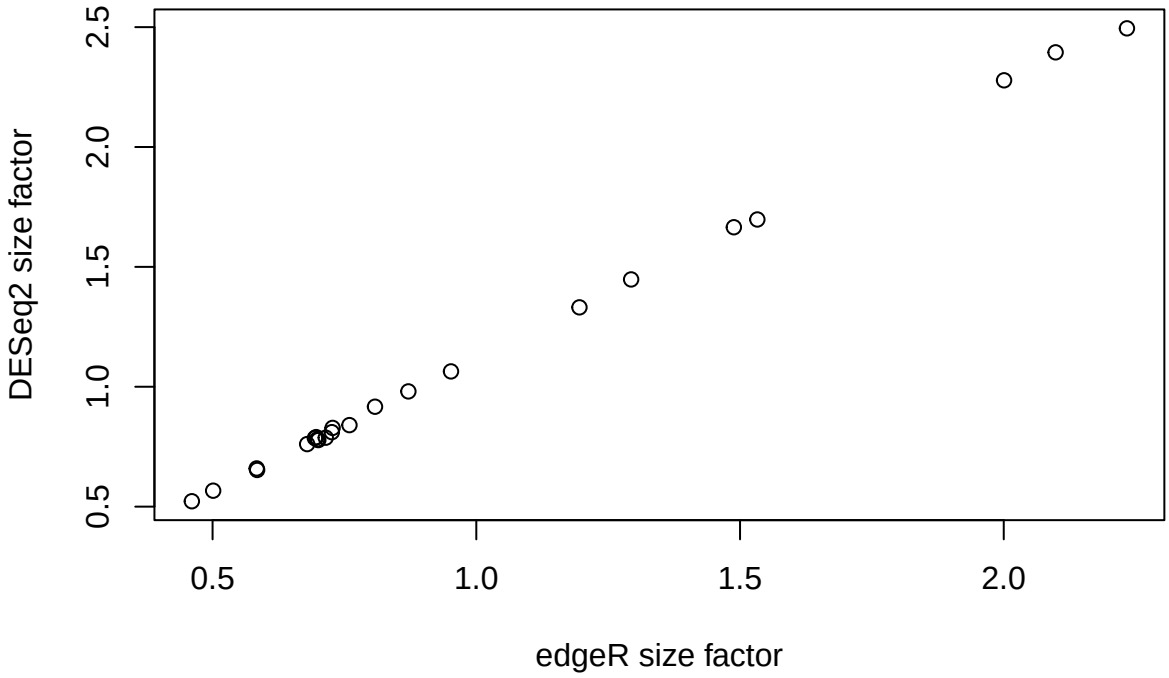
```
##  [1] 0.9166999 1.0640199 0.5224518 0.9806075 0.5664996 0.7802413 0.8282554
##  [8] 0.6592136 2.3949029 0.7898898 2.2781598 0.7858861 0.7775688 0.8399826
## [15] 1.3311498 1.6977165 2.4949218 0.7875376 0.8106158 0.7607974 1.4476678
## [22] 0.6529423 1.6653456
```

You may also want to check out the StatQuest video on DESeq2 normalization.

3.3.1 Comparing TMM with DESeq2 normalization

We can compare the size factors for both normalizations to verify if they agree on the normalization factors. Note we need to scale the effective library sizes from edgeR to enforce a similar scale as the size factors from DESeq2. While below we are using an arithmetic mean, a geometric mean may be used as well, which will be more robust to outlying effective library sizes.

```
plot(effLibSize / mean(effLibSize), sizeFactors(dds),
     xlab = "edgeR size factor",
     ylab = "DESeq2 size factor")
```



3.4 Full-quantile (FQ) normalization

In full-quantile normalization, originally introduced in the context of microarrays by Bolstad *et al.* (2003), the samples are forced to each have a distribution identical to the distribution of the median/average of the quantiles across samples. In practice, we can implement full-quantile normalization using the following procedure

1. Given a data matrix $\mathbf{Y}_{G\times n}$ for G genes (rows) and n samples (columns),
2. sort each column to get $\mathbf{Y}^S$,
3. replace all elements of each row by the median (or average) for that row,
4. obtain the normalized counts $\tilde{\mathbf{Y}}$ by re-arranging (i.e., unsorting) each column.

3.5 Uncertainty in normalization

- The offsets that are being cacluated for normalization purposes are estimates.
- The downstream analyses incorporates these estimates as if they are known, i.e., conditions on the estimates.
- The analysis therefore **ignores the uncertainty** in their estimation.

BOX 1 GLOBAL-SCALING NORMALIZATION FOR scRNA-seq DATA SETS

RNA-seq experiments are inherently stochastic, with reads being randomly sampled from a pool of amplified cDNA molecules. Accordingly, let X_{ij} denote a random variable representing the read count of gene i in cell j. Typically, the parameter of interest is the expression level of each gene (see left panel), i.e., the relative abundance of mRNA molecules for a gene within the population of mRNA molecules in each cell. For the sake of simplicity, we consider here the case of a homogeneous cell population.

Intuitively, a first effect captured through the scaling factor s_j is the endogenous mRNA content n_j, the total number of mRNA molecules per cell (middle panel). Indeed, even within a homogeneous population, n_j can vary across cells. Furthermore, after cell lysis, only a fraction (F_j) of these n_j molecules, are captured and reverse transcribed into cDNA. Consequently, only $n_j \times F_j$ cDNA molecules can potentially be amplified and subsequently sequenced. Critically, the capture and reverse transcription efficiency F_j varies between cells; this introduces cell-to-cell variability that should also be handled by s_j.

Subsequently, because of the minute amount of genetic material contained in a cell, this pool of $n_j \times F_j$ cDNA molecules must be amplified before sequencing library preparation. Variability in amplification efficiency can introduce cell- and gene-specific biases in the measurement of expression levels. We denote the cell-specific amplification factor as A_j, such that amplification leads to a pool of $n_j \times F_j \times A_j$ molecules.

Unlike microarray experiments, RNA-seq is inherently competitive, meaning that a fixed number of reads are distributed between genes. Given this, the amplified pools are subsequently diluted by a cell-specific factor D_j, so that there are $n_j \times F_j \times A_j \times D_j$ amplified cDNA molecules to be sequenced. In principle, the dilution factor D_j can be set so that a library contains the same number of molecules from each cell by carrying out a library quantification step and setting $D_j = m/(n_j \times F_j \times A_j)$, where m is the desired number of molecules per cell. Alternatively, each cell can contribute the same volume of amplified cDNA solution to the library, such that each library will contain a different number of amplified cDNA molecules if the concentration of the solution varies between cells. In this case, $D_j = d$, where d is the proportion of amplified molecules used to prepare the sequencing library. This decision is critical for interpreting the scaling factor s_j, since it affects the number of molecules that are available for sequencing and, consequently, the scale of cell-specific read counts.

Finally, the number of sequenced reads per molecule from each cell (sequencing depth), R_j, also varies stochastically. Consequently, by considering all the above factors, we expect to observe $n_j \times F_j \times A_j \times D_j \times R_j$ reads from cell j. Hence, even within the same sequencing lane, differences in sequencing depth introduce cell-specific artifacts that will be incorporated into the global scaling factor s_j.

While the above discussion assumes a homogeneous population of cells, this interpretation of scaling factors is still valid for more realistic scenarios—with heterogeneous populations—under specific assumptions, such as that the majority of genes is not differentially expressed or that there are roughly equal numbers of upregulated and downregulated genes.

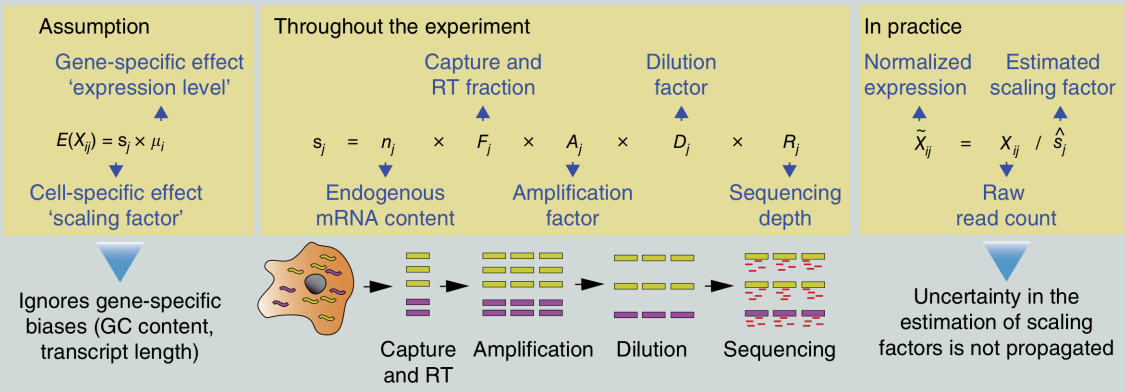


Figure 2: Figure: Box 1 from Vallejos et al. (2017).

4 Challenge III: Parameter estimation (under limited information setting)

There are two challenges to be overcome here:

- First, we need to get the structure of our mean model right, this is, which covariates to include, and how to include them, such that we are capable of capturing important sources of variation in our experiment, in order to derive a correct interpretation of the data in terms of our research question.
- Second, we need to be able to estimate the parameters of our model in an efficient way, while having limited information (i.e., we are often confronted with only a small number of replicates).

4.1 Defining the model for the mean

Let's first check how the authors of the original study parameterized the mean model.

transcripts using Cufflinks (version 1.0.3). Read counts per gene were calculated using HTSeq (version 0.5.1), and differential expression analysis was performed using the edgeR package, (26) employing treatment type, time point, and sample ID as factors in the model. Four different comparisons between sample groups were done: DPN 24 h *vs.* control 24 h, DPN 48 h *vs.* control 48 h, OHT 24 h *vs.* control 24 h, and OHT 48 h *vs.* control 48 h. All raw data are accessible through NCBI GEO Series accession no. GSE37211 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE37211).

Figure 3: Figure: Another paragraph from the Methods section.

The authors write that they are "employing treatment type, time point and sample ID as factors in the model". Concerning experimental variables, this suggests they have added covariates defining the treatment and time point for each sample. The sample ID in the text refers to the original tissue sample and therefore corresponds to the donor patient. While there is also a variable called `sample` in the `colData`, this is not what the authors refer to. Note the ambiguity here and since the authors didn't share their code, this is hard to check! But, more on reproducibility later…

Question. What are the authors assuming when using this structure for the mean model? Do you think that there are extensions or simplifications of the mean model that would be relevant?

Answer.

The authors are acknowledging the relatedness of samples derived from the same donor patient by adding it as a fixed effect to the model, which is great. This blocking strategy has been extensively discussed in the proteomics part of this course. However, by only adding a main effect for treatment and time, they are assuming that the effect of time is identical for all treatments, i.e., the average gene expression in-/decrease at 48h versus 24h is identical for the DPN, OHT or the control samples, which seems like a quite stringent assumption. We can make the model more flexible by allowing for a `treatment * time` interaction.

4.2 Parameter estimation and empirical Bayes

- Even in limited sample size settings, the parameters β of the mean model may be estimated reasonably efficiently, and we have previously discussed the IRLS algorithm to do so.
- However, estimating parameters for the variance (this is, the dispersion parameter ϕ from the negative binomial or the variance parameter σ from the Gaussian) is typically quite a bit harder.

- In genomics, we often take advantage of the parallel structure of the thousands of regression models (one for each gene) to **borrow information across genes** in a procedure called **empirical Bayes**, as also seen in the proteomics part of this course.

- In the Bayesian setting, we use not only the data, but also a prior distribution, to derive our parameter estimates. In a traditional Bayesian analysis, one assumes an *a priori* known prior distribution, which reflects our prior belief into all possible values of the parameter. This prior distribution is completely independent of the observed data and the idea is that one specifies the prior distribution before observing the data. One can then use the data and prior distribution to derive a **posterior distribution** for the parameter(s) θ of interest through Bayes rule

$$p(\theta|\mathbf{Y}) = \frac{p(\mathbf{Y}|\theta)p(\theta)}{\int_{\theta\in\Theta} p(\mathbf{Y}|\theta)p(\theta)d\theta}.$$

  Here, the posterior distribution $p(\theta|\mathbf{Y})$ is calculated using the data likelihood $p(\mathbf{Y}|\theta)$, prior distribution $p(\theta)$ and the 'marginal likelihood' $\int_{\theta\in\Theta} p(\mathbf{Y}|\theta)p(\theta)d\theta$, where Θ denotes the parameter space of θ. We can see that the posterior probability for a specific value of the parameter θ will be high if both the data likelihood as well as prior probability are high.

- In empirical Bayes, we basically take a semi-Bayesian approach to parameter estimation. Indeed, we do not assume a known prior distribution, but estimate it empirically using the data. This empirically estimated prior $\hat{p}(\theta)$ is then used to calculate the posterior distribution.

- While in some settings one can easily calculate the posterior distribution, sometimes it can be a hard problem. In such cases, it may be useful to restrict ourselves to calculating the **maximum a posteriori (MAP)** estimate, which corresponds to the mode of the posterior distribution. This can be considered to be analogous to point estimation in the frequentist setting.

- In our setting, we use genes with a similar average expression to moderate the dispersion estimate for a particular gene. The basic assumption for this to make sense is that genes with similar means might have similar dispersion parameters (or variances), owing to the mean-variance trend.

- Once initial estimates for the gene-wise dispersions have been derived ($\hat{\Phi}_g^{ML}$ in the figure), we use a parametric model to estimate its distribution (typically as a function of the mean) across genes. This distribution is the **prior distribution**.
- Then, each initial estimate is shrunken towards that empirically estimated prior distribution. The amount of shrinkage being performed is data-driven, and depends on the data, taking into account the precision of our initial estimate (i.e., shape of the likelihood) and the variability of the prior distribution.
- These strategies result in impressive performance gains in terms of differential expression analysis and are implemented in all popular differential expression analysis software packages (though in slightly differing ways) like `limma`, `edgeR` and `DESeq2`.

The blog post on understanding empirical Bayes estimation using baseball statistics is a great primer for further reading, as well as the accompanying book by David Robinson.

4.3 In practice

Let's fit the model using `edgeR`.

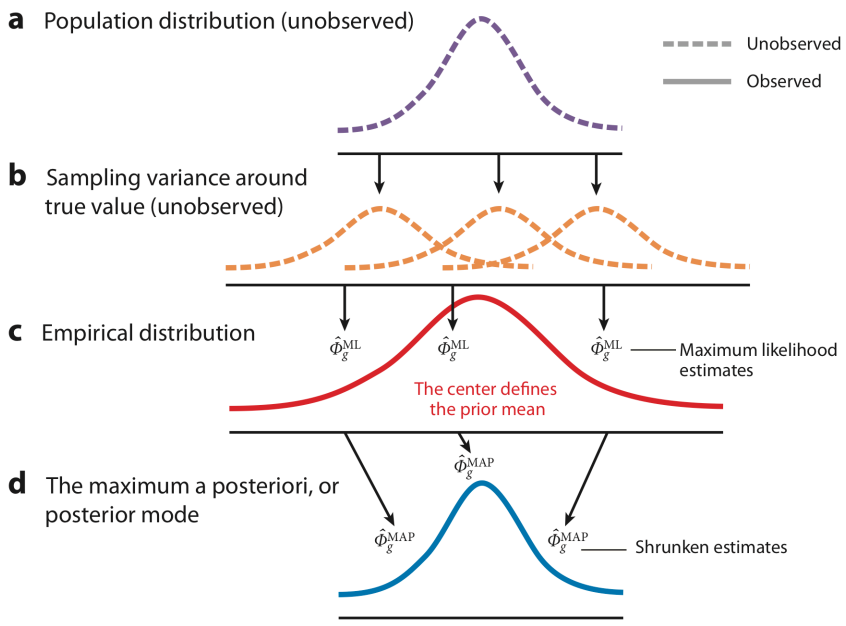
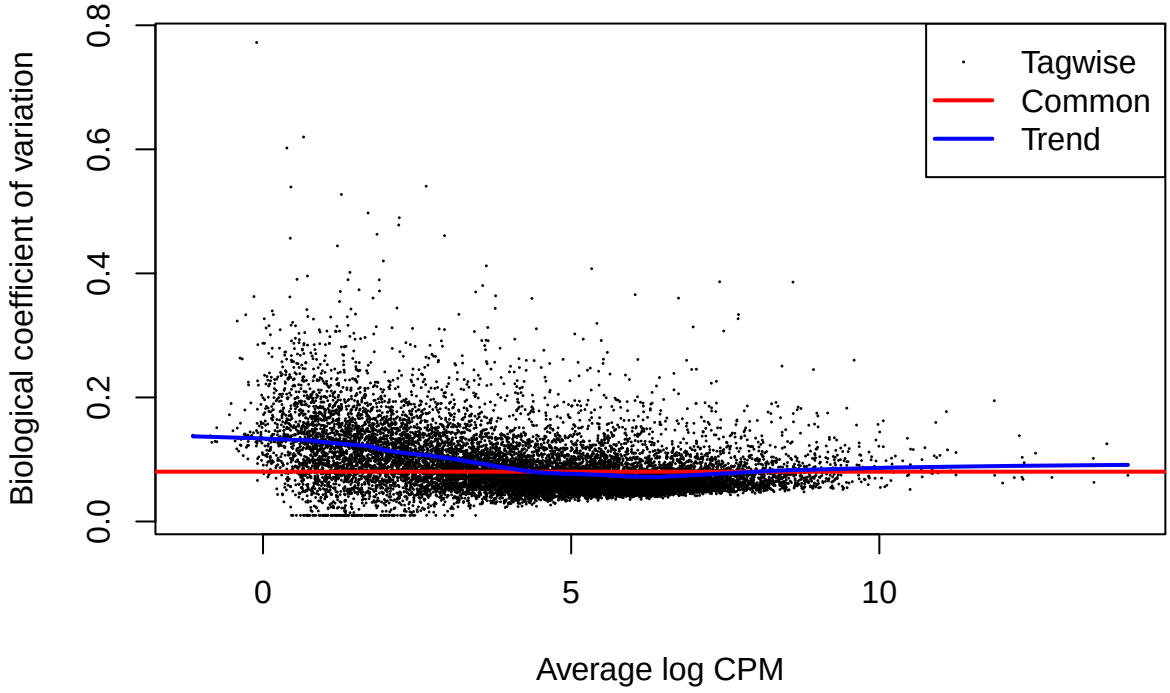


Figure 8

Steps in an empirical Bayes model. In an RNA sequencing experiment, one assesses the observed differences in gene expression across groups of samples with respect to within-group variance. (*a*) The unobserved population distribution for the true within-group variance of each gene. (*b*) Variances are estimated from limited sample size experiments, and so there is sampling variance in our estimate of the variance. A maximum likelihood estimate (MLE) or a bias-corrected estimator for expression variance can be used. (*c*) Thousands of genes are typically observed and estimates are made for each, providing an empirical distribution of MLEs across all genes. This empirical distribution of MLEs can be used to determine a prior distribution for empirical Bayes analysis; the posterior distribution for the variance of each gene is calculated using Bayes' formula. (*d*) Distribution of the maximum a posteriori (MAP), or posterior mode, estimates of variance over all genes. The posterior modes represent shrunken estimates, where the amount of shrinkage is determined by the shape of the likelihood and the width of the prior distribution.

Figure 4: Figure 8 from Van den Berge *et al.* (2019).

```
design <- model.matrix(~ treatment*time + patient, data=colData(se))


dge <- calcNormFactors(se)
dge <- estimateDisp(dge, design) # estimate dispersion estimates
plotBCV(dge)
```




```
fit <- glmFit(dge, design)
head(fit$coefficients)
```

```
##                  (Intercept) treatmentDPN treatmentOHT     time48h   patient2
## ENSG00000000003   -9.332660   0.11215980   0.08617711  0.13277338 -0.5185631
## ENSG00000000419  -10.374585  -0.05233370  -0.03544105 -0.09301681  0.1376257
## ENSG00000000457  -10.935441  -0.12815863  -0.13380545 -0.07347351  0.4574027
## ENSG00000000460  -10.098383   0.08875322   0.13754925 -0.89357390  0.5407554
## ENSG00000000971  -13.689209   0.29526136   0.26736723  0.49224506 -0.1468142
## ENSG00000001036   -8.280283   0.02211192  -0.02416563 -0.08638397 -0.4676508
##                    patient3   patient4 treatmentDPN:time48h
## ENSG00000000003 -0.82994270 -0.6245673          -0.12278804
## ENSG00000000419  0.10515134  0.1023421           0.04810092
## ENSG00000000457 -0.05731193  0.2089687           0.06405330
## ENSG00000000460 -0.33459640 -0.1321134           0.15904868
## ENSG00000000971  0.93857349 -0.2676066          -0.60487781
## ENSG00000001036 -1.52141542 -1.0046446           0.04238597
##                 treatmentOHT:time48h
```

```
## ENSG00000000003          -0.11955914
## ENSG00000000419           0.11561301
## ENSG00000000457           0.13101511
## ENSG00000000460           0.18156925
## ENSG00000000971          -0.71520710
## ENSG00000001036           0.08331487
```

5 Challenge IV: Statistical inference across many genes

5.1 Contrasts on the treatment effects

We will first derive all contrasts that are also investigated in the original manuscript, using our extended model where we are allowing for an interaction effect between treatment and time.

The mean model is

$$\log(\mu_{gi}) = \beta_{g0} + \beta_{g1}x_{DPN} + \beta_{g2}x_{OHT} + \beta_{g3}x_{48h} + \beta_{g4}x_{pat2} + \beta_{g5}x_{pat3} + \beta_{g6}x_{pat4} + \beta_{g7}x_{DPN:48h} + \beta_{g8}x_{OHT:48h}.$$

The intercept corresponds to the log average gene expression in the control group at 24h for patient 1.

DPN 24h vs control 24h. The respective means are

$$\log \mu_{g,DPN,24h} = \beta_{g0} + \beta_{g1},$$

$$\log \mu_{g,con,24h} = \beta_{g0}.$$

And their difference is

$$\delta_g = \beta_{g1}.$$

DPN 48h vs control 48h. The respective means are

$$\log \mu_{g,DPN,48h} = \beta_{g0} + \beta_{g1} + \beta_{g3} + \beta_{g7},$$

$$\log \mu_{g,con,48h} = \beta_{g0} + \beta_{g3}.$$

And their difference is

$$\delta_g = \beta_{g1} + \beta_{g7}.$$

OHT 24h vs control 24h. The respective means are

$$\log \mu_{g,OHT,24h} = \beta_{g0} + \beta_{g2},$$

$$\log \mu_{g,con,24h} = \beta_{g0}.$$

And their difference is

$$\delta_g = \beta_{g2}.$$

OHT 48h vs control 48h. The respective means are

$$\log \mu_{g,OHT,48h} = \beta_{g0} + \beta_{g2} + \beta_{g3} + \beta_{g8},$$

$$\log \mu_{g,con,48h} = \beta_{g0} + \beta_{g3}.$$

And their difference is

$$\delta_g = \beta_{g2} + \beta_{g8}.$$

However, we can also assess the interaction effects: is the time effect different between DPN and OHT treatment versus the control? And how about the DPN vs OHT treatments?

DPN vs control interaction. The time effect for each condition is

$$\delta_{DPN} = \log \mu_{g,DPN,48h} - \log \mu_{g,DPN,24h} = \beta_{g3} + \beta_{g7},$$

$$\delta_{con} = \log \mu_{g,con,48h} - \log \mu_{g,con,24h} = \beta_{g3}.$$

So the interaction effect is

$$\delta_{DPN-con} = \beta_{g7}.$$

OHT vs control interaction. The time effect for each condition is

$$\delta_{OHT} = \log \mu_{g,OHT,48h} - \log \mu_{g,OHT,24h} = \beta_{g3} + \beta_{g8},$$

$$\delta_{con} = \log \mu_{g,con,48h} - \log \mu_{g,con,24h} = \beta_{g3}.$$

So the interaction effect is

$$\delta_{DPN-con} = \beta_{g8}.$$

OHT vs DPN interaction. The time effect for each condition is

$$\delta_{OHT} = \log \mu_{g,OHT,48h} - \log \mu_{g,OHT,24h} = \beta_{g3} + \beta_{g8},$$

$$\delta_{DPN} = \log \mu_{g,DPN,48h} - \log \mu_{g,DPN,24h} = \beta_{g3} + \beta_{g7},$$

So the interaction effect is

$$\delta_{OHT-DPN} = \beta_{g8} - \beta_{g7}.$$

Let's implement all of these in a contrast matrix.

```
L <- matrix(0, nrow = ncol(fit$coefficients), ncol = 7)
rownames(L) <- colnames(fit$coefficients)
colnames(L) <- c("DPNvsCON24", "DPNvsCON48",
                 "OHTvsCON24", "OHTvsCON48",
                 "DPNvsCONInt", "OHTvsCONInt",
                 "OHTvsDPNInt")
# DPN vs control at 24h
L[2,"DPNvsCON24"] <- 1
# DPN vs control at 48h
L[c(2,8),"DPNvsCON48"] <- 1
# OHT vs control at 24h
L[3,"OHTvsCON24"] <- 1
# OHT vs control at 48h
L[c(3,9),"OHTvsCON48"] <- 1
# DPN control interaction
L[8,"DPNvsCONInt"] <- 1
# OHT control interaction
L[9,"OHTvsCONInt"] <- 1
# OHT DPN interaction
L[c(9,8),"OHTvsDPNInt"] <- c(1, -1)

L
```

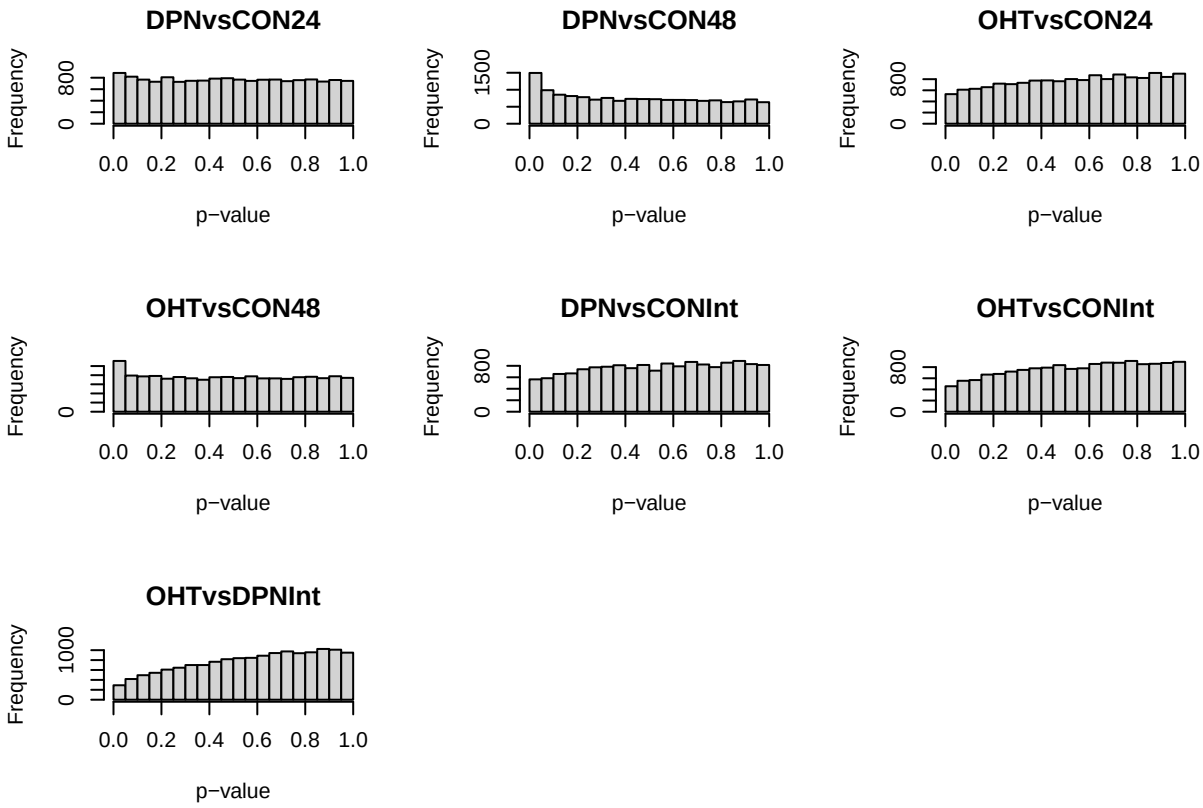
```
##                     DPNvsCON24 DPNvsCON48 OHTvsCON24 OHTvsCON48 DPNvsCONInt
## (Intercept)                  0          0          0          0           0
## treatmentDPN                 1          1          0          0           0
## treatmentOHT                 0          0          1          1           0
## time48h                      0          0          0          0           0
```

```
## patient2                     0          0          0          0          0
## patient3                     0          0          0          0          0
## patient4                     0          0          0          0          0
## treatmentDPN:time48h         0          1          0          0          1
## treatmentOHT:time48h         0          0          0          1          0
##                      OHTvsCONInt OHTvsDPNInt
## (Intercept)                    0           0
## treatmentDPN                   0           0
## treatmentOHT                   0           0
## time48h                        0           0
## patient2                       0           0
## patient3                       0           0
## patient4                       0           0
## treatmentDPN:time48h           0          -1
## treatmentOHT:time48h           1           1
```

And, finally, we can assess each hypothesis using the `glmLRT` function implemented in `edgeR`. We can assess each hypothesis separately by looping over the contrasts.

```
lrtList <- list() #list of results
for(cc in 1:ncol(L)) lrtList[[cc]] <- glmLRT(fit, contrast = L[,cc])

# p-value histograms
pvalList <- lapply(lrtList, function(x) x$table$PValue)
pvalMat <- do.call(cbind,  pvalList)
colnames(pvalMat) <- colnames(L)
par(mfrow=c(3,3))
sapply(1:ncol(pvalMat), function(ii) hist(pvalMat[,ii],
                                          main = colnames(pvalMat)[ii],
                                          xlab = "p-value"))
```

```
##          [,1]            [,2]            [,3]            [,4]
## breaks   numeric,21      numeric,21      numeric,21      numeric,21
## counts   integer,20      integer,20      integer,20      integer,20
## density  numeric,20      numeric,20      numeric,20      numeric,20
## mids     numeric,20      numeric,20      numeric,20      numeric,20
## xname    "pvalMat[, ii]" "pvalMat[, ii]" "pvalMat[, ii]" "pvalMat[, ii]"
## equidist TRUE            TRUE            TRUE            TRUE
##          [,5]            [,6]            [,7]
## breaks   numeric,21      numeric,21      numeric,21
## counts   integer,20      integer,20      integer,20
## density  numeric,20      numeric,20      numeric,20
## mids     numeric,20      numeric,20      numeric,20
## xname    "pvalMat[, ii]" "pvalMat[, ii]" "pvalMat[, ii]"
## equidist TRUE            TRUE            TRUE
```



5.1.1 Multiple testing

```
# number of DE genes
padjMat <- apply(pvalMat, 2, p.adjust, method="fdr")
colSums(padjMat <= 0.05 )
```

```
##  DPNvsCON24  DPNvsCON48  OHTvsCON24  OHTvsCON48 DPNvsCONInt OHTvsCONInt
##           2          62           0          22           0           0
## OHTvsDPNInt
##           0
```

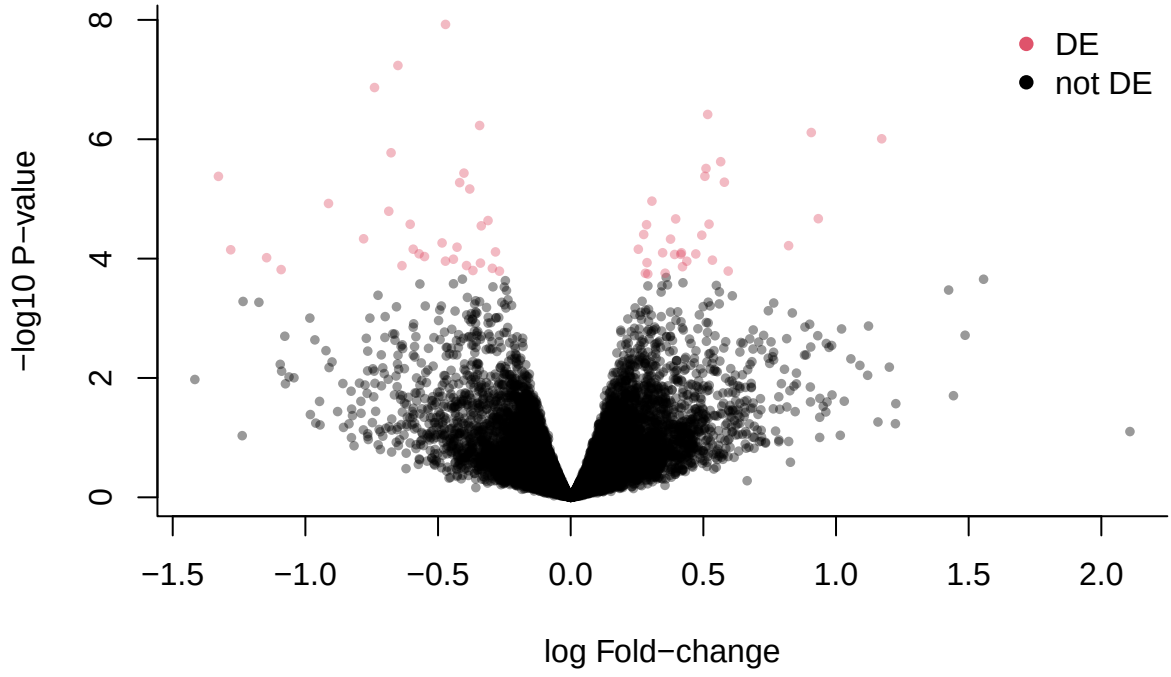
We are finding low numbers of DE genes between treatments at a 5% FDR level. This was already reflected in the the MDS plots.

5.1.2 Visualization

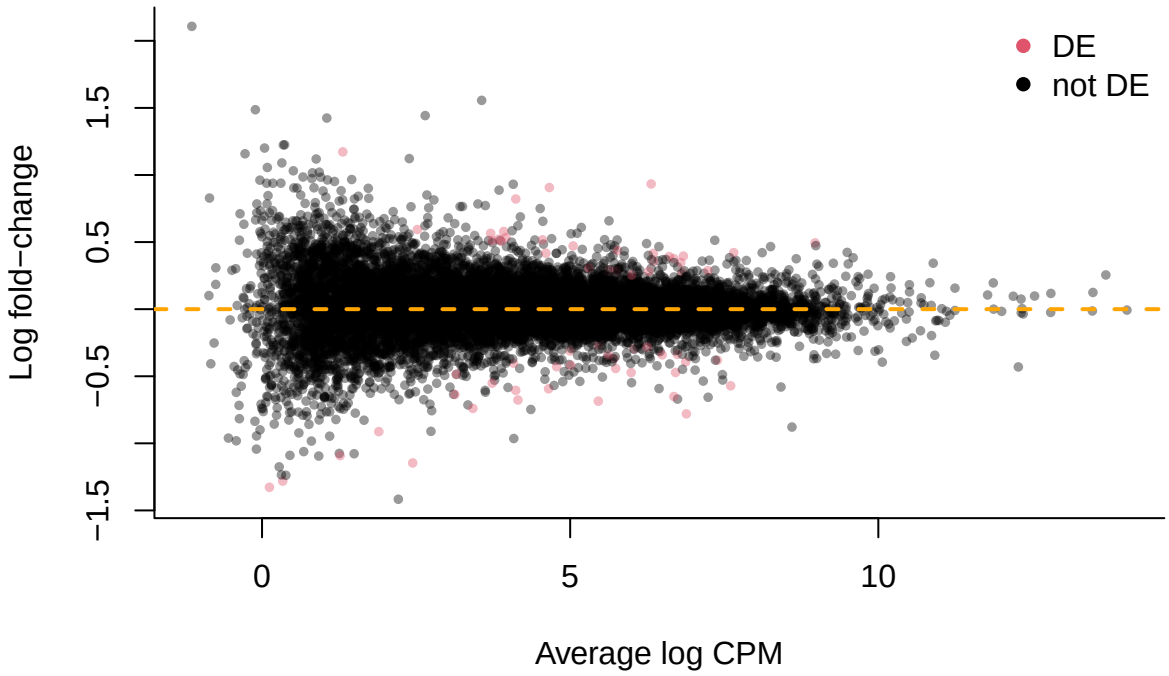
Let's visualize some results for the DPN vs control at 48h contrast.

```
library(scales) # for scales::alpha()
deGenes <- p.adjust(lrtList[[2]]$table$PValue, "fdr") <= 0.05

## volcano plot
```

```r
plot(x = lrtList[[2]]$table$logFC,
     y = -log10(lrtList[[2]]$table$PValue),
     xlab = "log Fold-change",
     ylab = "-log10 P-value",
     pch = 16, col = alpha(deGenes+1, .4),
     cex=2/3, bty='l')
legend("topright", c("DE", "not DE"),
       col = 2:1, pch=16, bty='n')
```



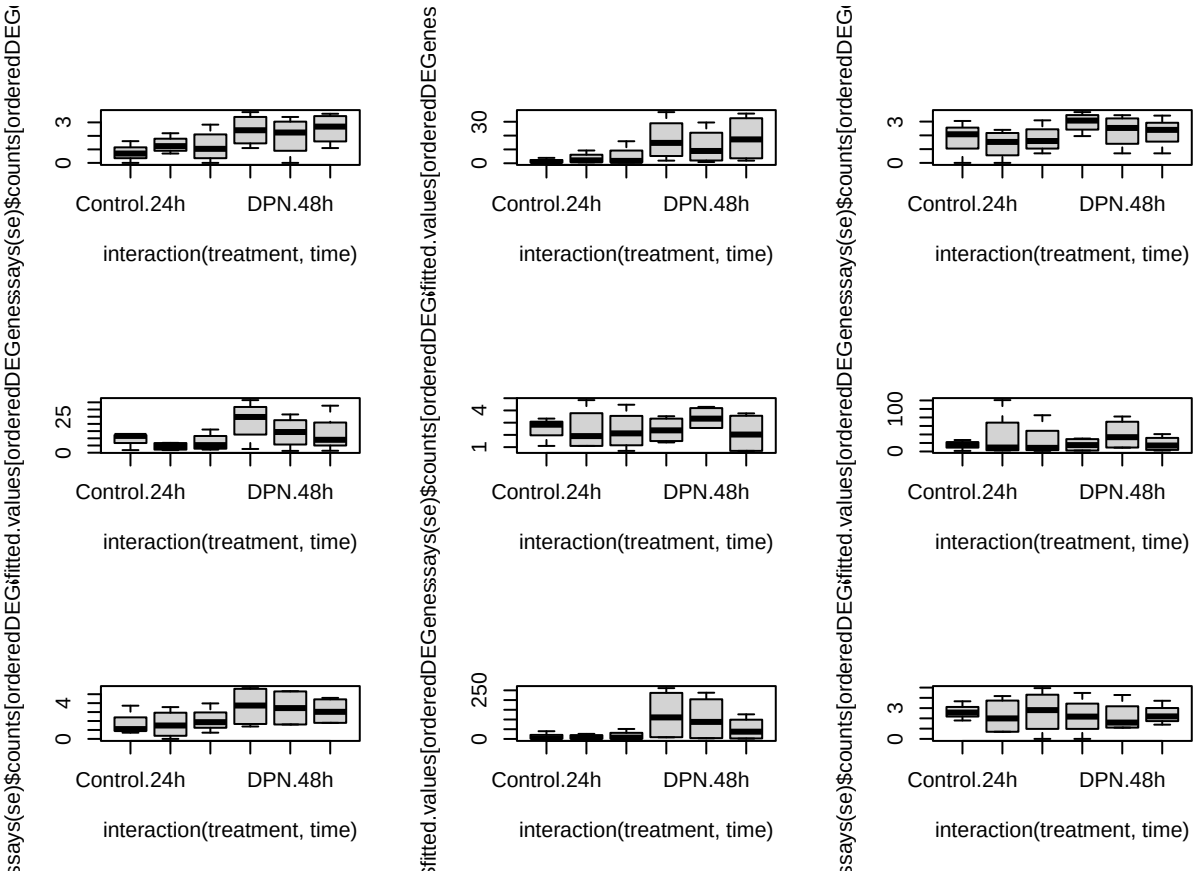

```r
## MD-plot
plot(x = lrtList[[2]]$table$logCPM,
     y = lrtList[[2]]$table$logFC,
     xlab = "Average log CPM",
     ylab = "Log fold-change",
     pch = 16, col = alpha(deGenes+1, .4),
     cex=2/3, bty='l')
legend("topright", c("DE", "not DE"),
       col = 2:1, pch=16, bty='n')
abline(h=0, col="orange", lwd=2, lty=2)
```

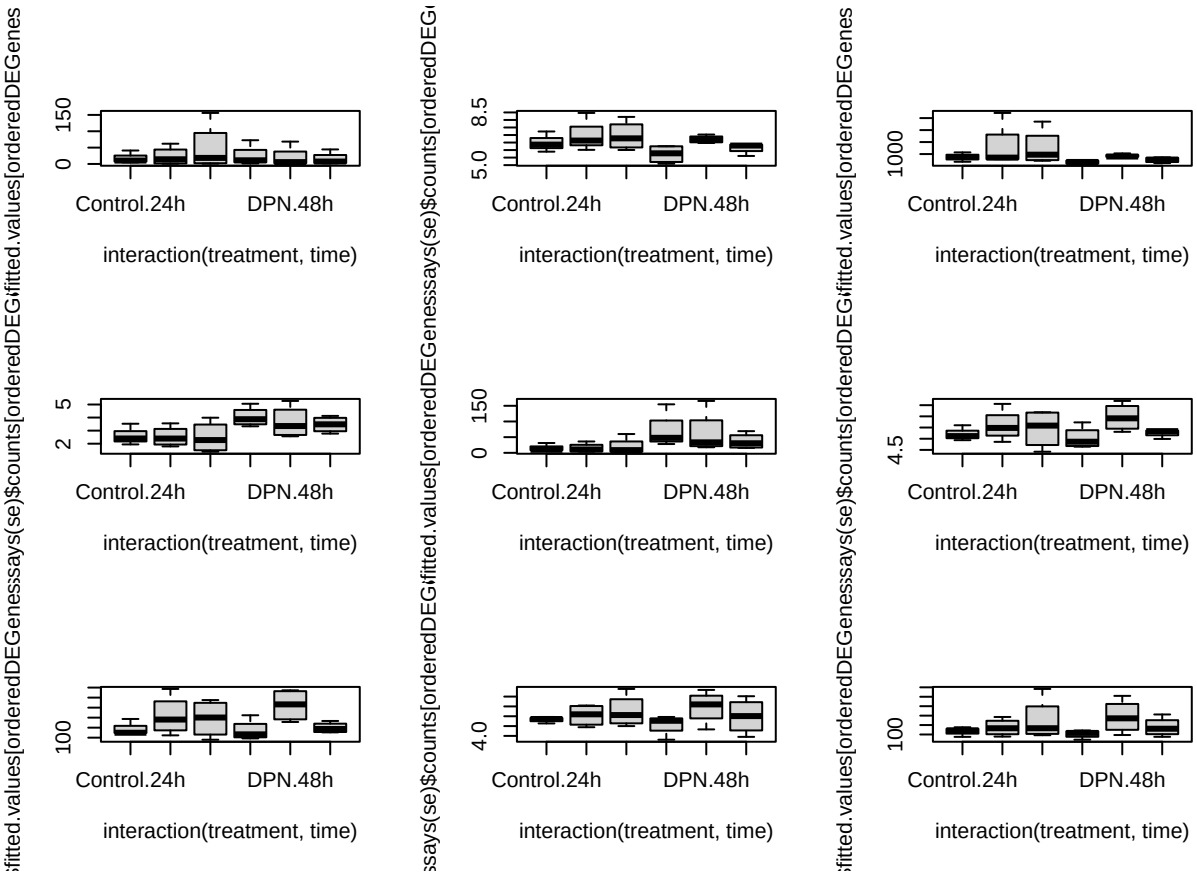


```
# extract all DE genes
deGenes <- rownames(lrtList[[2]]$table)[p.adjust(lrtList[[2]]$table$PValue, "fdr") <= 0.05]
deGenes
```

```
##  [1] "ENSG00000035928" "ENSG00000044574" "ENSG00000070882" "ENSG00000075790"
##  [5] "ENSG00000091137" "ENSG00000092621" "ENSG00000099864" "ENSG00000099875"
##  [9] "ENSG00000100219" "ENSG00000100867" "ENSG00000101255" "ENSG00000101974"
## [13] "ENSG00000102547" "ENSG00000103064" "ENSG00000103257" "ENSG00000103449"
## [17] "ENSG00000107796" "ENSG00000111181" "ENSG00000111790" "ENSG00000119242"
## [21] "ENSG00000120129" "ENSG00000120217" "ENSG00000133935" "ENSG00000135473"
## [25] "ENSG00000135541" "ENSG00000138685" "ENSG00000143127" "ENSG00000145050"
## [29] "ENSG00000145244" "ENSG00000149428" "ENSG00000149485" "ENSG00000152952"
## [33] "ENSG00000155111" "ENSG00000155330" "ENSG00000155660" "ENSG00000164597"
## [37] "ENSG00000166813" "ENSG00000167608" "ENSG00000167703" "ENSG00000168014"
## [41] "ENSG00000169174" "ENSG00000169239" "ENSG00000169762" "ENSG00000170122"
## [45] "ENSG00000170837" "ENSG00000171798" "ENSG00000173210" "ENSG00000175198"
## [49] "ENSG00000181790" "ENSG00000182704" "ENSG00000182836" "ENSG00000183401"
## [53] "ENSG00000185818" "ENSG00000187908" "ENSG00000188783" "ENSG00000197976"
## [57] "ENSG00000219481" "ENSG00000226887" "ENSG00000233705" "ENSG00000234431"
## [61] "ENSG00000236044" "ENSG00000243927"
```

```r
# order according to absolute fold-change
orderedDEGenes <- deGenes[order(abs(lrtList[[2]]$table[deGenes, "logFC"]), decreasing = TRUE)]

par(mfrow=c(3,3))
for(kk in 1:9){
  boxplot(log1p(assays(se)$counts[orderedDEGenes[kk],]) ~ interaction(treatment, time))
  boxplot(fit$fitted.values[orderedDEGenes[kk],] ~ interaction(treatment, time))
}
```





5.2 Contrast on the time effect

Based on the MDS plot, we can expect comparatively more DE genes for the time effect. For didactic purposes, here we assess an **average time effect across the three treatments**. The analysis shows how flexible one can be when using contrasts.

Mean of time 24h:

$$\log \mu_{g,24h} = \frac{1}{3}\left\{ \underbrace{(\beta_{g0})}_{\text{Control, 24h}} + \underbrace{(\beta_{g0}+\beta_{g1})}_{\text{DPN, 24h}} + \underbrace{(\beta_{g0}+\beta_{g2})}_{\text{OHT, 24h}} \right\}$$

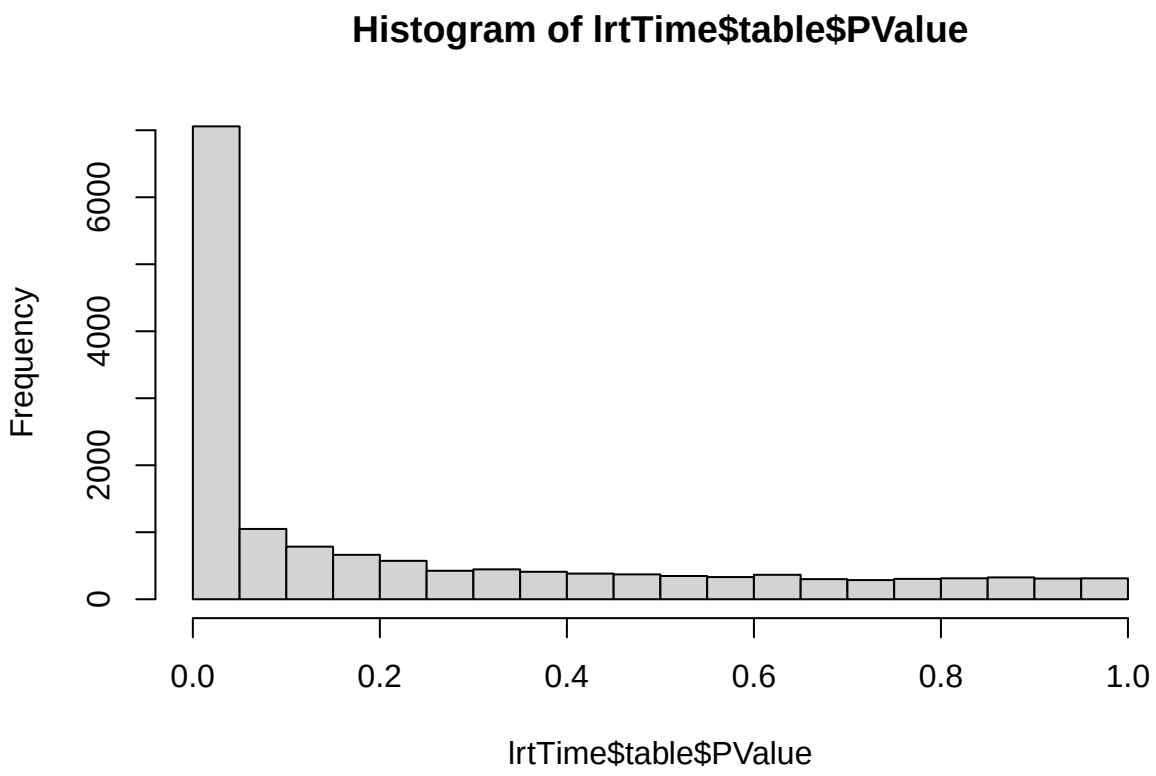
Mean of time 48h:

$$\log \mu_{g,48h} = \frac{1}{3}\left\{ \underbrace{(\beta_{g0}+\beta_{g3})}_{\text{Control, 48h}} + \underbrace{(\beta_{g0}+\beta_{g1}+\beta_{g3}+\beta_{g7})}_{\text{DPN, 48h}} + \underbrace{(\beta_{g0}+\beta_{g2}+\beta_{g3}+\beta_{g8})}_{\text{OHT, 48h}} \right\}$$

Difference:

$$\log\left(\frac{\mu_{g,48h}}{\mu_{g,24h}}\right) = \beta_{g3} + \frac{1}{3}(\beta_{g7}+\beta_{g8})$$

```
Ltime <- matrix(0, nrow = ncol(fit$coefficients), ncol = 1)
rownames(Ltime) <- colnames(fit$coefficients)
Ltime[c("time48h", "treatmentDPN:time48h", "treatmentOHT:time48h"),1] <- c(1, 1/3, 1/3)

lrtTime <- glmLRT(fit, contrast=Ltime)
hist(lrtTime$table$PValue) # very different p-value distribution
```




```
sum(p.adjust(lrtTime$table$PValue, "fdr") <= 0.05) # many DE genes
```

```
## [1] 5938
```

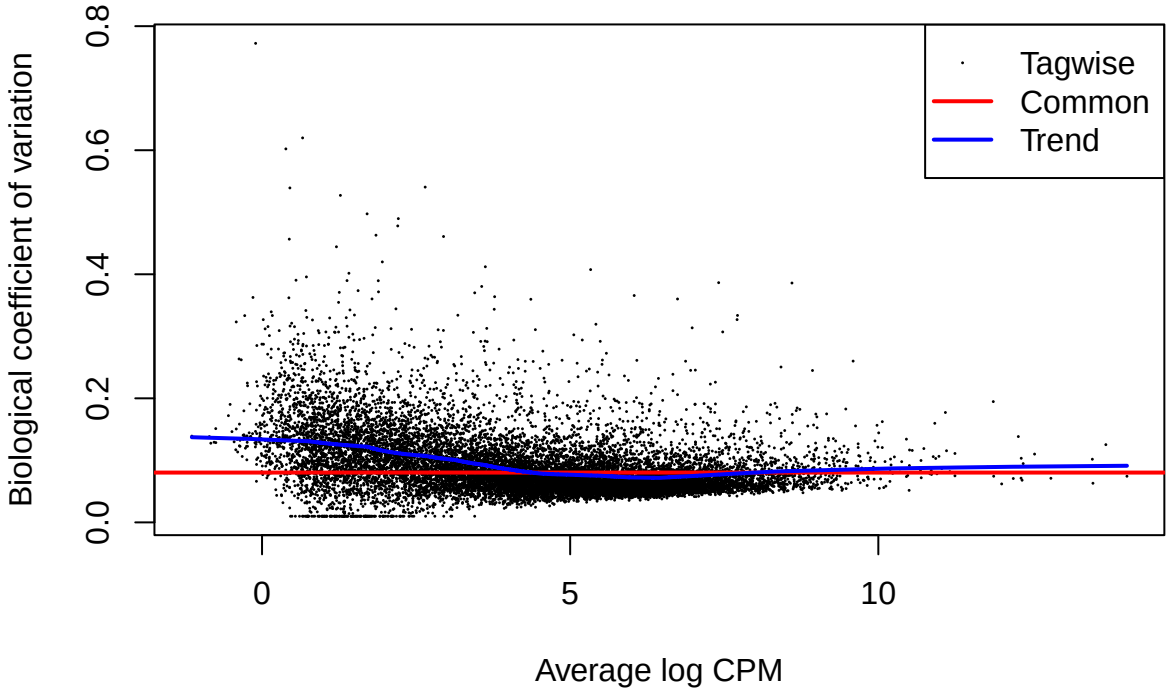
6 Alternative parameterizations

While our design matrix here was parameterized as `~ treatment*time + patient` alternative, equivalent parameterizations are also possible. Below, we demonstrate another parameterization that could work, too, and can be more intuitive. In this parameterization, we estimate a mean for each experimental condition, without an intercept, which can be convenient to think about how to set up contrasts.

```
treatTime <- as.factor(paste0(treatment, time))
table(treatTime)
```

```
## treatTime
## Control24h Control48h     DPN24h     DPN48h     OHT24h     OHT48h
##          3          4          4          4          4          4
```

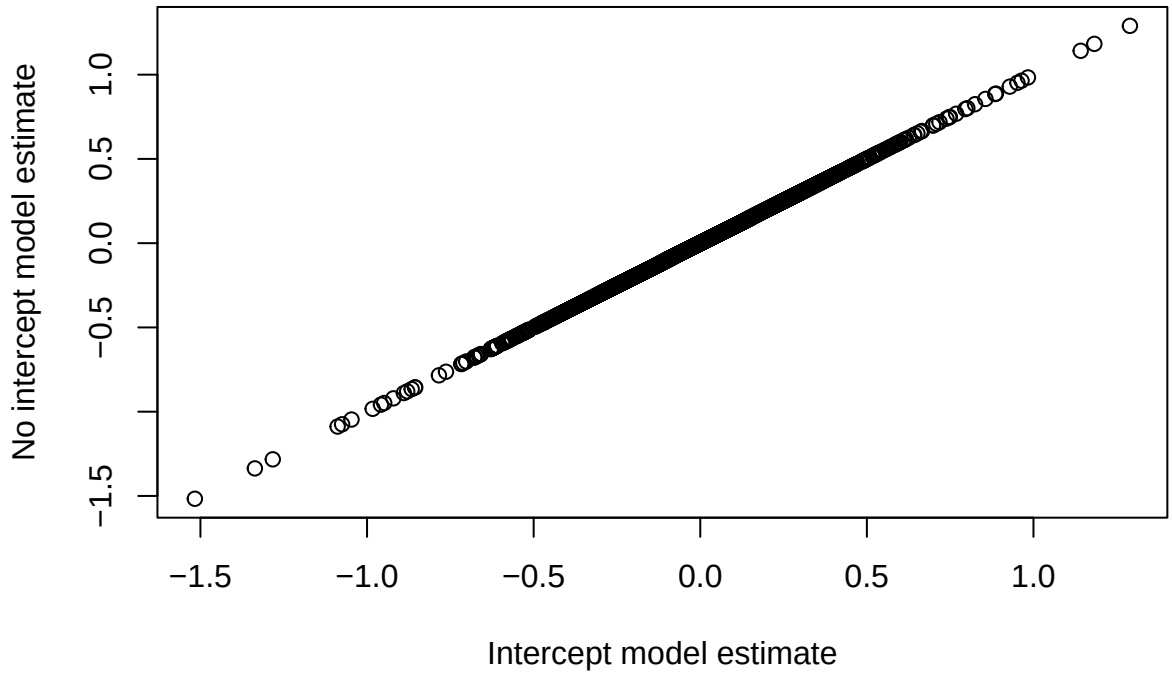
```
design2 <- model.matrix(~ 0 + treatTime + patient)

dge2 <- calcNormFactors(se)
dge2 <- estimateDisp(dge2, design2)
plotBCV(dge2)
```



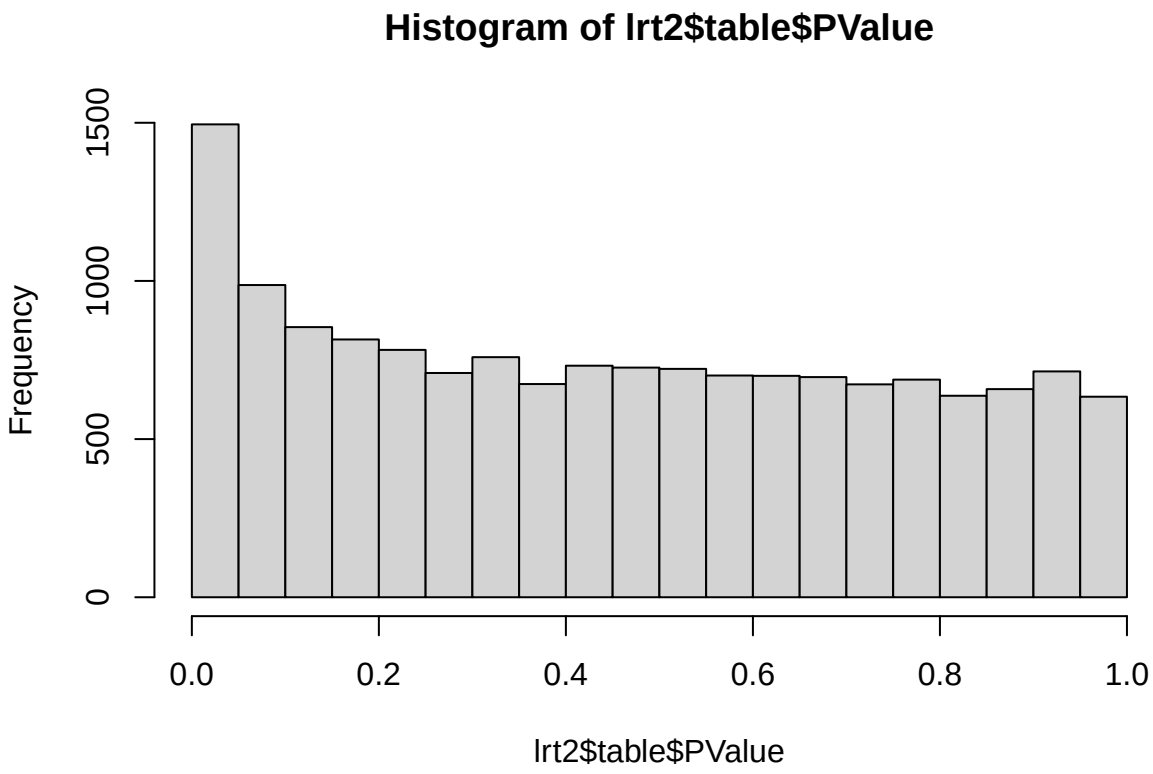

```
fit2 <- glmFit(dge2, design2)
head(fit2$coefficients)
```

```
##                  treatTimeControl24h treatTimeControl48h treatTimeDPN24h
## ENSG00000000003           -9.332660           -9.199887       -9.220500
## ENSG00000000419          -10.374585          -10.467601      -10.426918
## ENSG00000000457          -10.935441          -11.008915      -11.063600
## ENSG00000000460          -10.098383          -10.991957      -10.009630
## ENSG00000000971          -13.689209          -13.196964      -13.393947
## ENSG00000001036           -8.280283           -8.366667       -8.258171
##                  treatTimeDPN48h treatTimeOHT24h treatTimeOHT48h   patient2
## ENSG00000000003       -9.210515       -9.246483       -9.233269 -0.5185631
## ENSG00000000419      -10.471834      -10.410026      -10.387429  0.1376257
## ENSG00000000457      -11.073020      -11.069247      -11.011705  0.4574027
## ENSG00000000460      -10.744155       -9.960834      -10.672839  0.5407554
## ENSG00000000971      -13.506580      -13.421841      -13.644803 -0.1468142
## ENSG00000001036       -8.302169       -8.304449       -8.307518 -0.4676508
##                     patient3   patient4
## ENSG00000000003 -0.82994270 -0.6245673
## ENSG00000000419  0.10515134  0.1023421
## ENSG00000000457 -0.05731193  0.2089687
## ENSG00000000460 -0.33459640 -0.1321134
## ENSG00000000971  0.93857349 -0.2676066
## ENSG00000001036 -1.52141542 -1.0046446
```

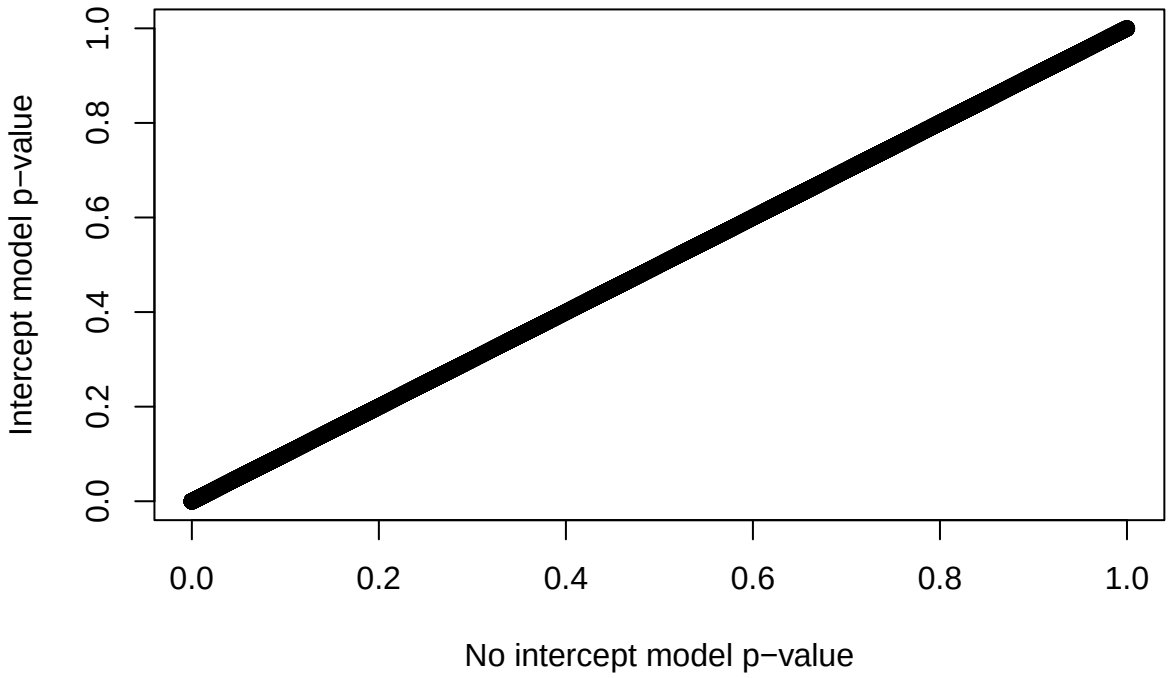
```
## for example: the estimate for the DPN24h vs control 24h is still the same,
## but requires a different combination of parameters
plot(fit$coefficients[,"treatmentDPN"],
     fit2$coefficients[,"treatTimeDPN24h"] - fit2$coefficients[,"treatTimeControl24h"],
     xlab="Intercept model estimate", ylab="No intercept model estimate")
```




```
# Let's implement the DPNvsCON48 contrast
L2 <- matrix(0, nrow = ncol(fit2$coefficients), ncol = 1)
rownames(L2) <- colnames(fit2$coefficients)
L2[c("treatTimeDPN48h", "treatTimeControl48h"),1] <- c(1, -1)
lrt2 <- glmLRT(fit2, contrast=L2[,1])
hist(lrt2$table$PValue)
```




```
plot(x=lrt2$table$PValue, y=lrtList[[2]]$table$PValue,
     xlab="No intercept model p-value",
     ylab="Intercept model p-value")
```



7 Additional Challenge (Opportunity?): The importance of reproducible analysis

Finally, it is crucial to make your analysis reproducible using tools such as RMarkdown and GitHub. Please sit back and watch this amazing lecture from Professor Keith Baggerly on "The Importance of Reproducible Research in High-Throughput Biology.