Tutorial on differential expression analysis of bulk RNA-seq data

Lieven Clement

In this tutorial we apply the edgeR quasi-likelihood workflow of the bulk RNA-seq lecture to a new dataset. In the first exercise we get acquainted with the workflow on a simple two-group comparison. In the second exercise we analyse the full experiment, which has a factorial design.

Users who work with scripts can adapt the code of the macrophage workflow in the bulk RNA-seq lecture.

Influenza infection and the mouse brain

Influenza A virus infects the lungs, but patients often also report neurological symptoms. Blackmore et al. (2017) studied how a peripheral influenza infection affects the central nervous system in mice. They infected 8-week-old C57BL/6 mice intranasally with influenza A and sequenced the RNA of the cerebellum of

- non-infected mice (day 0),
- mice 4 days after infection, i.e. around the peak of the infection in the lungs,
- mice 8 days after infection, i.e. when the virus is being cleared by the adaptive immune response.

Both female and male mice were used. The data are available in GEO under accession number GSE96870. The reads were aligned to the mouse genome with STAR and the reads that map to each gene were counted.

We use the count table that the Carpentries are available through the link https://github.com/carpentries-incubator/bioc-rnaseq/tree/main/episodes/data:

- `GSE96870_counts_cerebellum.csv`: the count table, with one row per gene and one column per sample,
- `GSE96870_coldata_cerebellum.csv`: the sample information, i.e. sex, infection, time and mouse identifier for every sample,

- `GSE96870_rowranges.tsv`: the gene annotation, i.e. gene symbol, chromosome (`seqnames`), Entrez identifier and type of the gene.

The files can be imported with the code below. `fread` of the data.table package detects the format of the file automatically, so the same function reads both the comma-separated (`.csv`) and the tab-separated (`.tsv`) files.

```
library(tidyverse)
library(edgeR)
library(data.table)

dataUrl <- "https://github.com/carpentries-incubator/bioc-rnaseq/raw/main/episodes/data/"
counts <- fread(paste0(dataUrl, "GSE96870_counts_cerebellum.csv"), data.table = FALSE) |>
  column_to_rownames("gene")
coldata <- fread(paste0(dataUrl, "GSE96870_coldata_cerebellum.csv"), data.table = FALSE)
rowdata <- fread(paste0(dataUrl, "GSE96870_rowranges.tsv"), data.table = FALSE) |>
  mutate(ENTREZID = as.character(ENTREZID))
```

Exercise 1: sex differences before the infection

We first analyse a simple two-group comparison: female vs male mice before the infection (day 0). The columns of the count table are in the same order as the rows of the sample information, so we can use the same selection for both.

```
all(colnames(counts) == coldata$sample)
day0 <- coldata$time == "Day0"
counts0 <- counts[, day0]
coldata0 <- coldata[day0, ]
table(coldata0$sex)
```

1. Import the data and select the samples of the day-0 mice with the code above. How many female and male mice are there? What is the experimental unit?
2. Store the counts, the sample information and the gene annotation in a `DGEList` object. Visualise the library sizes. Are there samples with a much lower sequencing depth than the others?
3. Filter the genes with `filterByExpr` and compute the TMM normalisation factors. How many genes do you keep?
4. Make an MDS plot and colour the samples according to sex. What do you observe?
5. Fit the model `~ sex` with `glmQLFit` and test for a sex effect with `glmQLFTest`. How many genes are differentially expressed at an FDR of 5%?

6. Which genes do you expect to differ between female and male mice? Look at the top table and the chromosomes on which the significant genes are located. Can this comparison serve as a positive control for your workflow?
7. Why are the log fold changes of the top genes so large?
8. Some significant genes on the X chromosome have a much smaller fold change. Can you explain why they are higher in females?
9. Make an MD plot or a volcano plot of the results.

Exercise 2: the full experiment

We now analyse all mice. The researchers want to answer the following questions:

- Is there an early effect of the infection (day 4 vs day 0) in female mice? And in male mice?
- Is there a late effect of the infection (day 8 vs day 0) in female mice? And in male mice?
- Does the expression change between the early and the late phase (day 8 vs day 4) in female mice? And in male mice?
- Do these effects differ between the sexes?
- What are the average early effect, late effect and change between the early and late phase over both sexes?
- Is there a difference between male and female mice before the infection?

1. Tabulate the number of mice per sex and time point, and per infection status and time point. What is the experimental unit? Which type of design is this? Can you estimate the effect of the infection separately from the effect of time?
2. Visualise the library sizes of all samples. Are they related to sex or time? Why would that be a problem?
3. Mistakes in the sample annotation happen. Check the recorded sex of each mouse with the expression of *Xist*, which is only expressed in females, and *Ddx3y*, which is located on the Y chromosome.
4. Filter and normalise the data. Make an MDS plot, colour the samples according to time and use a different symbol for each sex. Which sources of variability do the first two dimensions capture? Only a handful of genes differ between the sexes. Why can sex still dominate a dimension of the MDS plot?
5. Fit the appropriate model to the data, write the full QL model, write the mean log expression of each of each group in terms of the model parameters and interpret all model parameters.

6. Translate each research question into a contrast, i.e. a linear combination of the model parameters, and build the contrast matrix **L**.

7. Test all contrasts with `glmQLFTest`. How many genes are differentially expressed for each contrast?

8. Test whether the response to the infection differs between the sexes at any time point with an omnibus F-test. How many degrees of freedom does this test have, and why? No gene has a significant interaction. Would you drop the interaction from the model? Motivate your answer.

9. Use `glmTreat` to select the genes with a fold change that is significantly larger than 2 for the average contrasts. Visualise the results with a volcano plot, an MD plot and a heatmap of the top genes.

10. Perform an over-representation analysis of GO biological processes with `goana` (`species = "Mm"`) for the up- and down-regulated genes of the average early and late effects. Which biological processes are affected at day 4 and at day 8? How does the response of the cerebellum change over the course of the infection?

11. The contrast for the sex difference before the infection uses the same day-0 mice as exercise 1. The full model, however, estimates the dispersions with all mice. Would you expect the full model to detect more genes? Does it? Estimate the common biological coefficient of variation (BCV) separately for each time point with `estimateDisp` and use it to explain your findings.

References

Blackmore, Stephen, Jessica Hernandez, Michal Juda, et al. 2017. "Influenza Infection Triggers Disease in a Genetic Model of Experimental Autoimmune Encephalomyelitis." *Proceedings of the National Academy of Sciences* 114 (30): E6107–16. https://doi.org/10.1073/pnas.1620415114.