Sequencing: RNA-seq data intro

Koen Van den Berge

8/6/2021

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 Data import

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

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

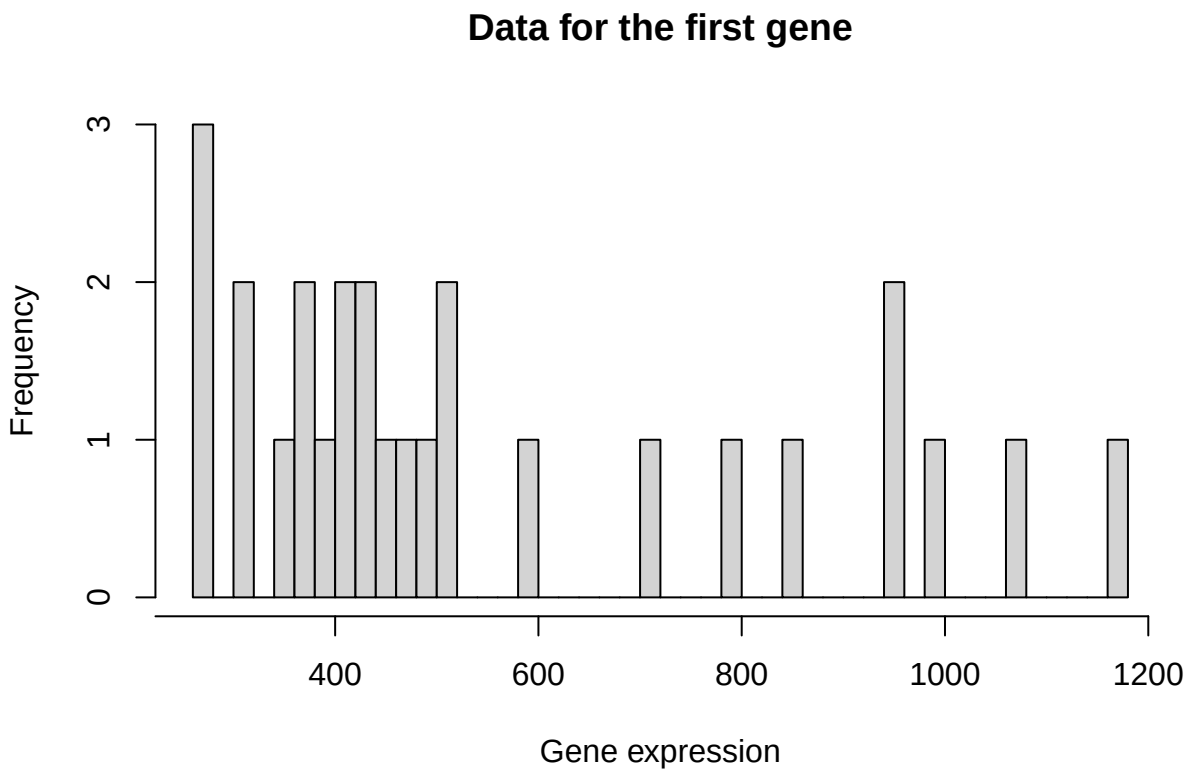
```
library(parathyroidSE)
library(SummarizedExperiment)

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

TODO: look into paper for exp design. Possibly take screenshot and try to decipher together with students.

2 Choice of modeling assumptions

```
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)
```



3 junk

Start with challenges below, and make a section for each. This will simultaneously follow a full RNA-seq analysis pipeline.

Following code should be part of a next file where we start working with RNA-seq data, show the mean-variance trend, discuss normalization, and introduce a DE analysis.

4 Properties of (RNA-seq) count data

Goals:

- Work with real RNA-seq count data, gene by sample matrix
- Explore the data for a single gene: univariate (large range, zeroes, discrete, skewed) as well as bivariate wrt covariate.
-
- Demonstrate the mean-variance relationship by plotting mean and variance across genes. Show that the Poisson doesn't hold across biological replicates and use this as introduction for the negative binomial distribution.

In this lecture, we will introduce working with count data, using a real bulk RNA-seq dataset

- Count data are inherently discrete.

4.1 Mean-variance relationship

4.2 Relative uncertainty, offsets and count scaling

Defer to normalization.

5 Variance-stabilizing transformations

Defer to when talking about dim red.