Exercise 8

Ewa Szczurek
MIM UW

November 24, 2015

Exercise 1. Differential expression with RNA-seq data. We will follow the steps of DESeq vs edgeR tutorial in R.

For DESeq tutorial, see `http://bioconductor.org/packages/release/bioc/vignettes/DESeq/inst/doc/DESeq.pdf`

1. Construct the count matrix and the meta data for pasilla experiments. Note that we are dealing with a multi-factor design (we have different conditions and different samples).
2. In usual setup, count matrix should be made from bam files (alignment) and an annotation file (gtf format). This can be done
   - using python and the pysam module
   - using R
   - using HTSeq
3. Make a new countDataSet
4. Estimate the normalisation factors (referred to as library sizes)
5. To see the estimated library sizes use

   ```
   sizeFactors(d)
   ```

   - Which of the samples has the largest estimated factor?
   - What does this tell about this sample in comparison to other samples?
   - To do normalisation, should we multiply or divide the counts in the samples by these values?
6. To store the normalised counts, run

   ```
   dn = counts( d, normalized=TRUE )
   ```

7. Plot the box plots for the raw counts in each sample, and box plots of the normalised counts. Set the 0 values to NA. Use

   ```
   boxplot
   ```

   function and

   ```
   par(mfrow=c(1,2))
   ```

8. Estimate and plot dispersion

9. The dispersion can be understood as the square of the coefficient of biological variation. So, if a genes expression typically differs from replicate to replicate sample by 20%, this genes dispersion is 0.22 = .04.

10. Note that in DESeq the variance seen between counts is the sum of two components: Poisson noise, i.e., the uncertainty in measuring a concentration by counting reads (technical noise) and the biological sample-to-sample variation (dispersion). The former is the dominating noise source for lowly expressed genes. The latter dominates for highly expressed genes. The sum of both, Poisson noise and dispersion, is considered in the differential expression inference.

    - What is on the x axis?
    - What is on the y axis?
    - Where is the dispersion the largest compared to mean?
    - What is the red line in the plot?

11. See what the variance stabilising transformation does

    ```
    library("vsn")
    par(mfrow=c(1,2))
    notAllZero = (rowSums(counts(d))>0)
    meanSdPlot(log2(counts(d)[notAllZero, ] + 1), ylim = c(0,2.5))
    vsd = vsd = varianceStabilizingTransformation( d )
    meanSdPlot(vsd[notAllZero, ], ylim = c(0,2.5))
    ```

12. Make the PCA plot for this data. What do the principal components correspond to?

13. For inference of differential expression, we now specify two models by formulas. The full model regresses the genes expression on both the library type and the treatment condition, the reduced model regresses them only on the library type. For each gene, we fit generalized linear models (GLMs) according to the two models, and then compare them in order to infer whether the additional specification of the treatment improves the fit and hence, whether the treatment has significant effect.

    - Inspect fitting results calling head(dfit1) and head(dfit0).

  - The first columns show the fitted coefficients, converted to a logarithm base
  - Which column stands for the $\log_2$ fold change (log ratio of ’untreated’ versus ’treated’?
  - The column deviance is the deviance of the fit. (Comparing the deviances with a χ^2 likelihood ratio test is how nbinomGLMTest calculates the p values.)
  - The last column, converged, indicates whether the calculation of coefficients and deviance has fully converged.

14. Make results table with pvalues and adjusted p-values.
15. Repeat for edgeR
16. Compare the results on the final plot and using

    ```
    addmargins(table(sig.edgeR=etable$FDR<0.05, sig.DESeq=dtable$padj<0.05))
    ```

    - Which method finds more differential genes?

Homework 1. No homework as of today!