Introduction to transcriptome sequencing analysis methods

Anamaria Necsulea

March 26, 2024

1 Introduction

During this second session, we will continue our analysis of the involvement of Hotair in *HOXD* gene regulation. We will specifically ask which genes are differentially expressed upon *Hotair* knockout. For this, we will start from a read count table, which gives us the number of RNA-seq reads attributed to each gene, for each of the 8 samples that we studied previously.

You can download the data from the following URL:
`http://pbil.univ-lyon1.fr/members/necsulea/MADGT_2024/`

We will use the DESeq2 R package for differential expression analysis. The use of this package is described in detail in the vignette:
`https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html`

You will need to load this library in R with the command `library(DESeq2)`.

2 One factor differential expression analysis

1. Using the `read.table` function in R, load the data in an object named `read.counts`.
2. Create an object named `read.counts.t3`, containing only the columns containing the read counts for the posterior trunk (T3) samples.
3. Create a data frame named `coldata.t3`, including a single column named `genotype`. This column should contain the genotype values for all the T3 samples, given in the same order as in the `read.counts.t3` table.
4. Add row names to the `coldata.t3` object. They should correspond to the `read.counts.t3` column names.
5. Using `DESeqDataSetFromMatrix` function, construct a DESeq object from the read counts and from the column data generated above for the T3 samples. You can find help for this in the *Count matrix input* section of the DESeq2 vignette.
6. Using the `DESeq` and `results` functions in DESeq2, perform a differential expression analysis test, comparing the two genotypes. You can find help for this in the *Differential expression analysis* section of the DESeq2 vignette.
7. Analyze the results. How many statistically significant differentially expressed genes can you detect? What are the statistical criteria on which you base your decision?
8. Is *Hotair* differentially expressed between genotypes? Are the *HoxD* genes differentially expressed? In what direction is the differential expression observed? What does this tell you about the reference used for differential expression analysis in DESeq2?
9. Using the `plotMA` function, analyze the distribution of log2 fold changes as a function of the average read count. What do you observe?

10. Using the lfcShrink function, adjust the log2 fold changes as recommended in the DESeq2 package. You can find help for this in the *Log fold change shrinkage for visualization and ranking* section of the vignette. Redo the MA-plot and compare it to the previous one. How many significantly differentially expressed genes do you obtain after this procedure?

11. Redo the same analysis for FL samples and compare the results with the ones obtained for T3 samples. To do this efficiently, write a function in R that takes the tissue (FL or T3) as argument, and which performs the entire differential expression analysis.

12. Using GOrilla (https://cbl-gorilla.cs.technion.ac.il), perform a gene ontology enrichment analysis, contrasting the genes that were up-regulated in the knockout mice with all annotated genes. Repeat the analysis for down-regulated genes. What do you observe? Is the set of all annotated genes the best control for this GO enrichment analysis? What alternative can you propose?

3 Two-factor differential expression analysis

1. Using the full read count table, implement in DESeq2 an additive model that has two explanatory variables: the tissue (FL or T3) and the genotype (wt or ko).

2. Extract the results of the differential expression analysis for the genotype factor, using the contrast parameter in the results functions in R.

3. Compare the results of the two-factor differential expression analysis with the results of the two one-factor analyses performed above. What do you observe?

4. Implement a model that incorporates an interaction between the two factors. Compare the results of this differential expression analysis with the ones of the additive model.

4 Data transformations and visualisation

For the differential expression analysis, we used the raw read counts as an input for DESeq2. However, for data visualisation we need to transform the counts, mainly to reduce the dependency between the variance and the mean that we previously observed. For that, we will use the variance stabilizing normalization (VST) function in DESeq2. See the corresponding chapter in the DESeq2 vignette (https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html).

1. Extract the VST transformed data from the dds object obtained in the two-factor analysis, using the vst function in DESeq2 with default parameters.

2. Perform a principal component analysis and display the first factorial map using the plotPCA function in R.

3. Extract the normalized counts from the dds object using the counts function in R. Select the lines that correspond to *Hoxd* genes. Display the corresponding results in a heatmap, using the image function in R. Can you visually confirm the differential expression analysis results that you observed previously?