RNA-seq hands-on tutorial using Chipster: Differential expression analysis and batch effect correction using a Drosophila dataset

Eija Korpelainen and Maria Lehtivaara, CSC – IT Center for Science, chipster@csc.fi

In this tutorial you start with a ready-made read count table, and perform experiment level quality control. You then detect differentially expressed genes using DESeq2 and edgeR, and learn how to take confounding factors into account in differential expression analysis. Finally, you filter data based on a given column and play with different visualizations. For example, you learn how to compare gene lists using the interactive Venn diagram.

We use Drosophila data from an RNAi knock-down experiment of the pasilla splicing factor gene. The experiment is a two-group comparison with 4 untreated samples and 3 RNAi-treated samples. Unfortunately, some samples were sequenced single end and some paired end. Now it is your job to correct for this mistake in the differential expression analysis!

1. Open session

Click **Sessions** in the top panel, scroll down to **Training sessions**, and select **course_RNAseq_drosophila**. Save your own copy of the session: go to the **Session info** section, click the **three dots** by the session name, select **Save a copy**, and give your session a new name.
Inspect the session description and check the phenodata file. In particular, pay attention to the group, description and readtype columns of the phenodata.

2. Check the experiment level quality with PCA

Select the file **pasilla_counts.tsv** and the tool **Quality control / PCA and heatmap of samples with DESeq2.** Choose
Phenodata column for coloring samples in PCA plot = treatment_description
Phenodata column for the shape of samples in PCA plot = readtype_description

-Do the groups separate along the first principal component (PC1)? How much variance does this PC explain?
-How much variance is explained by PC2? Do the single end and paired end samples separate along PC2?
-According to the heatmap, are there subgroups within the treated and untreated samples which are more similar to each other?

3. Analyze differential expression with DESeq2

Select the file **pasilla_counts.tsv** and run the tool **RNA-seq / Differential expression using DESeq2**.

-How many differentially expressed genes do you get?
-Inspect **summary.txt**. How many genes had some reads mapping to them? How many of those genes had too low read counts and were hence left out of the analysis? What was the low count threshold that DESeq2 decided?
-BONUS EXERCISE: Are the final dispersion values (blue small spots) always higher than the original ones (black)?

4. Analyze differential expression with DESeq2 so that you take read type into account

Select the file **pasilla_counts.tsv** and the tool **RNA-seq / Differential expression using DESeq2**, and set the parameter **Column describing additional experimental factor = readtype.** Rename the resulting DE list to **de-list-deseq2-rt.tsv** (click on the three dots by the file name and select **Rename**).

-How many differentially expressed genes do you get now? Was it a good idea to include the readtype?
-Did DESeq2 decide to use the same low count threshold as before?

5. Compare the gene lists from exercises 3 and 4 using a Venn diagram

Select the two DE lists (the .tsv files) from exercises 3 and 4 by keeping the ctrl/cmd key down. In the visualization panel select the method **Venn diagram**.

-How many genes do the lists have in common?

6. Check how many genes have been up-regulated more than 4-fold and visualize their profiles

Select the file **de-list-deseq2-rt.tsv** and run the tool **Utilities / Filter table by column value** by setting the parameters as follows:

-**Column to filter by** = **log2FoldChange**
-**Does the first column lack a title** = **yes**
-Cutoff = 2 (remember that 2 in log2 scale means 4 in linear scale)
-Filtering criteria = larger-than

-How many genes have a fold change higher than 4? Visualize them as an interactive **expression profile**.

7. Draw a heatmap of the results

Select the **filtered-NGS-results.tsv** from the previous step and the original file **pasilla_counts.tsv**. Select the tool **RNA-seq / Heatmap for RNA-seq results** and make sure that the input files have been correctly assigned.
- View the heatmap and save it on your computer.

8. Visualize the read counts of the gene which has the smallest padj-value

Select the file **de-list-deseq2-rt.tsv** and the tool **RNA-seq / Plot normalized counts for a gene**. In the parameters, indicate the **gene name** (FBgn0039155) and set **Show names in plot = yes.**

-Do the groups differ clearly in the read counts? Is the within-group variability large?

9. BONUS EXERCISE: Analyze differential expression with edgeR

Select the file **pasilla_counts.tsv** and run the tool **RNA-seq / Differential expression using edgeR** so that you set **Filter out genes which don't have counts in at least this many samples = 3**.

-Why do we use the criteria of 3 samples in filtering?
-How many differentially expressed genes do you get?
-Look at the **MA plot**. How large of a fold change is required for a gene to be considered statistically significantly differentially expressed? Does the MA plot look different from the one made by DESeq2? Why?