RNA-seq hands-on tutorial using Chipster: Find differentially expressed genes between lung and lymph node using 10 samples

Eija Korpelainen, 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 compare the result lists using an interactive Venn diagram. You also filter the result list based on a given column. Finally, you annotate genes and perform pathway analysis.

1. Open new session and fill in phenodata

Select the session **course_RNAseq_lung_lymphnode_comparison_10samples**. In this session you have a count table of 10 samples (5 from lung and 5 from lymph node) and a phenodata file describing the samples.
Save your own copy of the session: go to the **Session info** section, click the **three dots** by the session name, select **Rename**, and give your session a new name.
The group column of the phenodata is pre-filled for you. Please check to ensure that the phenodata is **1** for lung samples, and **2** for lymph node samples.

2. Check the experiment level quality with PCA plot

Select the file **ngs-data-table.tsv** and run the tool **Quality control / PCA and heatmap of samples with DESeq2.**
-Do the groups separate along the first principal component (PC1)? How much variance does this PC explain?
-According to the heatmap, do there seem to be subgroups within the lung and lymph node samples which are more similar to each other?

3. Analyze differential expression with edgeR

Select the file **ngs-data-table.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 = 5**.
-Why do we use the criteria of 5 samples in filtering?
-How many differentially expressed genes do you get (click "View in full screen" to see the number of rows)?

4. Analyze differential expression with DESeq2

Select the file **ngs-data-table.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?
- Inspect **deseq2_report.pdf**. Can you see the shrinkage of fold change values in the MA plot?

5. Compare the DE gene lists found by edgeR and DESeq2 using a Venn diagram

Select files **de-list-deseq2.tsv** and **de-list-edger.tsv**. In the visualization panel Chipster proposes **Venn diagram,** click **Draw**.
-How many genes do the lists have in common?
Select the common genes: Click on that section in the graph and click **Create file**.

6. Check how many genes have a linear fold change greater than the absolute value of 4

Select the file **de-list-deseq2.tsv** and run the tool **Utilities / Filter table by column value** like this:

-**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 = outside

-How many genes have a fold change higher than 4?

7. Add gene symbols and description to a gene list

Select the file **filtered-NGS-results.tsv** and run the tool **Utilities / Annotate Ensembl identifiers**.

-If you look at the description of the genes, do you see lung-specific functions (e.g. surfactant proteins or mucins)?

8. We do not do this exercise now: Find pathways which are over-represented in the list of annotated genes

Select the file **annotated.tsv** and run the tool **RNA-seq / Hypergeometric test for ConsensusPathDB**.

Open the result file **cpdb-pathways.tsv**. Can you see pathways related to lymph node or lung function?