RNA-Seq hands-on tutorial using Chipster: Differential expression analysis when there is a strong batch effect (Parathyroid dataset)

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

In this tutorial, you will detect differentially expressed genes between cultured adenocarcinoma samples before and after treatment with DPN. In the analysis, you need to take into account that the samples are
-paired (before and after samples are from the same patient)
-from different time points (treated either 24h or 48h)
-from different patients

1. Open a session

Open the session **course_RNAseq_parathyroid**. Inspect the session description and the phenodata: This data contains 12 samples from 3 different patients, 2 different treatments (control and DPN), and 2 different time points (24h and 48h).

3. Study the effect of different factors with PCA

Select the **parathyroid_counts.tsv** and run the tool **Quality control / PCA and heatmap of samples with DESeq2** so that you set the parameters as follows:
Phenodata column for coloring samples in PCA plot = group
Phenodata column for the shape of samples in PCA plot = patient
Do the samples separate nicely according to the experimental groups? If not, why not?

4. Analyze differential expression with DESeq2

Select the file **parathyroid_counts.tsv** and run the tool **RNA-seq / Differential expression using DESeq2**.
-How many differentially expressed genes do you get?

Repeat the run so that you set **Column describing additional experimental factor = patient**.
-How many differentially expressed genes do you get? What if you set the **P-value cutoff** to **0.1**?
-How many genes are removed by the automatic independent filtering (check summary.txt)?

5. Analyze differential expression with edgeR

Let's run edgeR three times, adding one more effect (factor) on each run.
a) Select the file **parathyroid_counts.tsv** and run the tool **RNA-seq / Differential expression using edgeR for multivariate experiments**, and set the effects so that the **Main effect 1 = group**, and leave the other two effect fields EMPTY for now. Set the parameter **Analyze only genes which have counts in at least this many samples** = **3.**
b) Run as above, but set also **Main effect 2 = time**.
c) Run as above, but set also **Main effect 3 = patient**.

6. Extract differentially expressed genes from the edgeR result files

Run the tool **Utilities / Filter table by column value** for each of the **edger-glm.tsv** files setting the parameters as follows:

Column to filter by = **FDR-as.factor(group)2**
Does the first column lack a title = yes
Cut-off value = 0.05

-Which edgeR run produced most DE genes? How many are they?

7. Annotate the results

In order to see the gene symbols and descriptions of the DE genes, choose the **filtered-ngs-result.tsv** and the tool **Utilities / Annotate Ensembl IDs**.

8. Draw a heatmap of the results

Select the **annotated.tsv** from previous step and the original file **parathyroid_counts.tsv**. Select the tool **RNA-seq / Heatmap for RNA-seq results** and make sure that the input files have been correctly assigned. Select to **Represent genes with** = **gene names**.