Exercise 3. RNA-seq hands-on tutorial using Chipster: Parathyroid dataset

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

1. **Launch Chipster**.

In this tutorial you start with a ready-made read count table and look for differentially expressed genes. You learn how to perform statistical testing so that you take into account up to 3 experimental factors. You also learn to add gene information to tables containing Ensembl identifiers.

The data contains 12 samples, which are cultured parathyroid adenocarcinomas from 3 patients, treated with DPN or nothing, for 24 and 48 hours. This data is available in the Bioconductor package parathyroid.

1. Open a session. Select **Open Example session** and **course_RNAseq_parathyroid**. Inspect the session description and the phenodata.

2. 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**. Run the tool twice with the following parameter settings:

a) Phenodata column for coloring samples in PCA plot = **treatment**
Phenodata column for the shape of samples in PCA plot = **patient**

b) Phenodata column for coloring samples in PCA plot = **hours**
Phenodata column for the shape of samples in PCA plot = **patient**

-Which factor seems to affect the samples most? Do we need to take it into account in analysis?

3. 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? And if you set the P-value cutoff to 0.1?

-How many genes are removed by the automatic independent filtering (check summary.txt)?

4. 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. Filter out genes that are not expressed in at least **3 samples.**

b) Run as above but set also **Main effect 2 = time**.

c) Run as above but set also **Main effect 3 = patient**.

Note that the result files contain all the genes that went to the analysis, as edgeR doesn't know which comparison you are interested in (treatment, patient or time).

5. Filter the results based on the desired comparison. Select each **edger-glm.tsv** file (one at a time) and run the tool **Utilities / Filter table by column value** setting the parameters as follows:

Column to filter by = **FDR-as.factor(group)2**

Does the first column have a title = no

-Which edgeR run produced most DE genes? How many are they? And if you filter that run with FDR < 0.1?

-How would you get the number of differentially expressed genes between the different time points?

A small p-value means there is a small chance that the observed results occurred by chance. FDR (False Discovery Rate) is a modification of the p-value.

DESeq2 result table

- baseMean = mean of counts (divided by size factors) taken over all samples
- **log2FoldChange = log2 of the ratio meanB/meanA**
- lfcSE = standard error of log2 fold change
- stat = Wald statistic
- pvalue = raw p-value
- **padj = Benjamini-Hochberg adjusted p-value**

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
FBgn0026562	47282.42	-2.4	0.08	-30.26	4.159e-201	3.309e-197
FBgn0039155	924.27	-4.46	0.16	-27.04	4.476e-161	1.781e-157
FBgn0029167	4287.44	-2.21	0.08	-26.75	1.107e-157	2.937e-154
FBgn0035085	654.94	-2.5	0.11	-22.08	5.278e-108	1.050e-104
FBgn0034736	231.7	-3.29	0.18	-18.28	1.261e-74	2.006e-71
FBgn0000071	359.53	2.6	0.14	17.98	2.741e-72	3.635e-69
FBgn0034434	153.84	-3.69	0.21	-17.26	9.008e-67	1.024e-63
FBgn0039827	342.77	-3.83	0.23	-16.54	1.742e-61	1.733e-58
FBgn0029896	513.08	-2.34	0.14	-16.29	1.168e-59	1.033e-56
FBgn0052407	220.26	-2.2	0.15	-14.99	8.597e-51	6.841e-48
FBgn0037754	299.03	-2.23	0.15	-14.94	1.916e-50	1.386e-47

6. Annotate the results. In order to see the gene symbols and descriptions of the DE genes, choose the **filtered-ngs-result.tsv** (FDR < 0.1) and the tool **Utilities / Annotate Ensembl IDs**. Set **species = human**.

-Which gene has the smallest FDR for the group comparison? Note that you need to sort the result table for the column **FDR-as.factor(group)2**.