RNA-seq hands-on tutorial using Chipster: lymph node and lung comparison

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

In this tutorial you will compare gene expression in human lung and lymph node samples. To make things faster, the data is given in two separate sessions:
-In the first session you have two paired-end samples in fastq format. In order to make the analysis faster during the course, only a small subset of the reads (200 000) is used. You will preprocess and align the reads, make a count table and perform differential expression analysis.
-In the second session you will work with a ready-made count table, generated from 10 full size samples. You will perform experiment level quality control, differential expression analysis, filtering and annotation.

1. From raw reads to differentially expressed genes using the session with 2 samples

Open the session **course_RNAseq_lung_lymphnode_comparison_2samples**. It has 4 fastq files: 2 for each sample. Note that normally fastq files are zipped and Chipster can use them like that.
Based on what you have learned, analyse the data and answer the questions below. Remember to save your session from time to time.

-What can you tell about the samples based on their name?
-How long are the reads? How many read pairs are there for each sample?
-How is the base quality, do you need to trim the reads?
-Is the data stranded? What is the inner distance of the paired reads?
-What is the alignment rate? How many reads have multiple alignments?
-Can you find any genes with large counts? How many alignments were not counted for any gene?
-Are there genes which are differentially expressed between these two samples? How many? Which gene has the highest positive fold change? How does it look like in the genome browser? Save your session.

2. Differential expression analysis using a count table containing 10 samples

Open the session **course_RNAseq_lung_lymphnode_comparison_10samples**. In this session, you have a count table of 10 samples and a phenodata file describing the samples. Based on what you have learned, analyse the data and answer the questions below.

-Do the experimental groups separate based on the PCA plot?
-How many differentially expressed genes do you get with DESeq2? How many of them have positive fold change? How many genes had low counts?
-How many differentially expressed genes you get with edgeR? Are many of those found also by DESeq2?
-How many genes have changed their expression more than 4-fold based on the DESeq2 result?
-If you look at the description of the genes with smallest padj-values, do you see lung-specific functions (e.g. surfactant proteins and mucins)?