QuantSeq hands-on tutorial using Chipster: From count table to differentially expressed genes

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

In this tutorial you perform differential expression analysis using QuantSeq FWD UMI 3' RNA-seq data. The data is a small subset of samples from the experiment by Porrit et al (Arteriosclerosis, Thrombosis, and Vascular Biology. 2020;40:802–818) where they studied the role of IL-1 in the Kawasaki Disease (KD). They used Lactobacillus casei cell wall extract (LCWE) to induce a mouse model of KD, and then treated some samples with the IL-1 receptor antagonist Anakinra. Your job is to **find out which genes changed their expression in response to the Anakinra treatment**.
The full dataset is available in GEO, accession GSE141072. The subset used here consists of 6 male samples:
-3 samples treated with LCWE only (samples marked with L)
-3 samples treated with LCWE and Anakinra (samples marked with LA)

We have already perfomed the following analysis steps for you, but they are still described here (steps 1-11) so that you can use this document as a guide when analyzing your own data later.
-check the read quality
-remove UMIs and TATA from reads and store UMIs in read names
-remove polyA read-through, adapters and low quality ends
-check the strandedness of the data
-align reads to genome
-remove PCR duplicates
-count reads per genes
-combine the count files in one table and describe the experimental setup using the phenodata file

You can start from step 12 and learn how to
-peform experiment level QC
-detect differentially expressed (DE) genes
-annotate DE genes
-make a heatmap
-perform pathway analysis

1. Start Chipster and open a session

Go to **https://chipster.rahtiapp.fi/** and **Log in** using your HAKA (=university) account.
Go down to **Training sessions** and select **course_QuantSeq_KawasakiDisease_6samples**. This session has six FASTQ files, one for each sample. 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.

2. Check the quality of reads with MultiQC

Select all the FASTQ files and the tool **Quality control / Read quality with MultiQC for many FASTQ files** and click **Run Tool (1 job)**. Select the resulting **multiqc_report.html** and click **Open in tab**.
-How many reads are there in each file and how long are they (configure columns in General Statistics part)?
-Are there a lot of duplicate reads?
-Is the base quality good all along the reads?
-Is the base distribution random in the beginning of the reads?
-Do the reads have adapters left?

3. Remove TATA and store UMIs in read names

Select all the FASTQ files and run the tool **Preprocessing / Extract UMIs from QuantSeq reads** so that you set **Create log file = yes**. When the result files arrive, inspect the log file of sample ML1.
-How many reads had TATA in positions 7-10 and were therefore processed and kept?

4. Check with MultiQC that TATA and UMI were removed

Select all the **extracted** FASTQ files and run the **MultiQC tool** again.
-How long are the reads now?
-Is the TATA motif gone?

5. Remove polyA read-through, adapters and low-quality ends

Select all the **extracted** FASTQ files and run the tool **Preprocessing / Trim QuantSeq reads using BBDuk** so that you set **Create log file = yes**. When the result files arrive, inspect the log file of sample ML1.
-What percentage of reads contained adapters or polyA?
-What is the length of the reads now? Are all the reads the same length?
-How many reads were removed and why?

6. Check with MultiQC that polyA read-through and adapters were removed

Select all the **trimmed** FASTQ files and run the **MultiQC tool** again.
-How long are the reads now?
-Are the adapters gone?
-Is there less A in the reads towards the ends (is the plot "Per Base Sequence Content" less green towards the end than before)?

7. Check the strandedness of the reads

Select **ML1_extracted_trimmed.fq.gz** and run the tool **Quality control / RNA-seq strandedness inference** so that you set **Organism = Homo_sapiens.GRCh38.95**. Open the resulting **experiment_data.txt**.
-Is the data stranded? Mark down the parameters for HTSeq.

8. Align reads to reference genome using STAR

Select all the **trimmed** FASTQ files and the tool **Alignment / STAR for single end reads**. Set the parameters as described below and click **Run tool for each file (6 jobs)**.
Genome = Mus_musculus.GRCm38.95
Maximum alignments per read = 20
Maximum mismatches per alignment = 999
Mismatch ratio = 0.1
Minimum intron size = 20
Maximum intron size = 1000000
Maximum gap between two mates = 1000000
When the results arrive, check the **log_final.txt** of sample ML1.
-What percentage of reads mapped uniquely?
-What percentage of reads mapped to multiple loci?

Inspect the contents a BAM file: Select **ML1_extracted.bam** and run the tool **Utilities / Create a preview for BAM**. Open the result file **ML1_extracted.prev.sam**.
-Can you see the UMIs in read names? Do there seem to be PCR duplicates?
-Can you see short reads?

9. Remove PCR duplicates

Select all the **BAM** files and run the tool **Preprocessing / Deduplicate aligned QuantSeq reads**.

You can run the tool **Utilities / Create a preview for BAM** again for **ML1_extracted_deduplicated.bam**.
-Are the PCR duplicates gone?

10. Count reads per genes using HTSeq

Select all the **deduplicated BAM** files and the tool **RNA-seq / Count aligned reads per genes with HTSeq**. Make sure that the parameters are set correctly for your data:
Reference organism = Mus_musculus.GRCm38.95
Does the BAM file contain paired-end data = no
Is the data stranded and how = **yes in HTSeq**
Run the tool by clicking **Run for each file (6 jobs)**.
-Inspect **ML1.tsv**. Can you find genes with counts?
-Check in **htseq-count-info.txt** how many alignments were not counted for any gene and why?

11. Create a count table and description file for the experiment

Select **all the tsv files** containing the read counts, and the tool **Utilities / Define NGS experiment**. Set the parameters **Does your data contain genomic coordinates** = **yes** and **Count column** = **count**. Run the tool by clicking **Run tool (one job)**.
In the resulting **phenodata.tsv** file, fill in the **group** column: enter **1** for the L samples and **2** for the LA (Anakinra-treated) samples.

12. START HERE: Check the experiment level quality with PCA

Select the file **ngs-data-table.tsv** and run the tool **Quality Control / PCA and heatmap of samples with DESeq2**.
-Do the sample groups separate along the principal component 1? How much variance does this PC explain?

13. Detect differentially expressed genes with DESeq2

Select the file **ngs-data-table.tsv** and run the tool **RNA-seq / Differential expression using DESeq2**.
-How many differentially expressed genes are detected (check the number of rows in de-list-deseq2.tsv)?

14. Retrieve the top 50 differentially expressed genes

Select the file **de-list-deseq2.tsv** and run the tool **Utilities / Modify text** by setting the parameters as follows:
Operation = Select a set of rows from the file
First row to select = 1
Last row to select = 51
Input file format = Table

15. Visualize the top 50 differentially expressed genes in a heatmap

Select the file **selected.tsv** and **ngs-data-table.tsv** and run the tool **RNA-seq / Heatmap for RNA-seq results** and make sure that the input files have been correctly assigned.
-Are these genes up- or down-regulated by the Anakinra treatment?

16. Annotate the top 50 differentially expressed genes

Select the file **selected.tsv** and run the tool **Utilities / Annotate Ensembl identifiers**.
-Do you see muscle-related genes in the result list?

17. Perform pathway analysis on the gene list

Select the file **annotated.tsv** and run the tool **RNA-seq / Hypergeometric test for ConsensusPathDB** so that you set **Organism = mouse**.
Inspect the result file **cpdb-pathways.html**. What pathway is statistically most significantly enriched in this list of genes?