QuantSeq hands-on tutorial using Chipster: From raw reads to counts per genes

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

In this tutorial you learn to analyze QuantSeq FWD UMI 3' RNA-seq data starting from the raw reads (FASTQ files). You will learn how to
-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 data
-align reads to genome
-check the alignment level quality
-remove PCR duplicates
-count reads per genes

We perform these preprocessing steps with a small subset of reads so that the analysis jobs run fast enough in a course setting. In the next exercise set a bigger dataset is used for differential expression analysis etc.

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.
The full dataset is available in GEO, accession GSE141072. We use here a subset of reads from two male samples:
-1 sample treated with LCWE only (sample marked with ML)
-1 sample treated with LCWE and Anakinra (sample marked with MLA)

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_2samples**. This session has two 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 both 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 (in General Statistics part you can configure columns and select length)?
-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 both 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 ML3.
-How many reads had TATA in positions 7-10 and were therefore processed and kept?

4. Check with MultiQC that TATA and UMIs were removed

Select both **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 both **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 ML3.
-What percentage of reads contained adapters or polyA?
-How many reads were removed and why?

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

Select both **trimmed** FASTQ files and run the **MultiQC tool** again.
-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)?
-What is the length of the reads now? Are all the reads the same length?

7. Check the strandedness of the reads

Select **ML3_subset_extracted_trimmed.fq.gz** and run the tool **Quality control / RNA-seq strandedness inference** so that you set **Organism = Mus_Musculus.GRCh38.95**. Open the resulting **experiment_data.txt**.
-Is the data stranded? Mark down the parameters for HISAT2 and HTSeq.

8. Align reads to reference genome using HISAT2

Select both **trimmed** FASTQ files and the tool **Alignment / HISAT2 for single end reads** and set the parameters as described below and click **Run tool for each file (2 jobs)**.
Genome = Mus_musculus.GRCm38.95
RNA-strandness = F
When the results arrive, check the **log_final.txt** of sample ML3.
-What percentage of reads mapped uniquely?
-What percentage of reads mapped to multiple loci?

Inspect the contents a BAM file: Select **ML3_subset_extracted.bam** and run the tool **Utilities / Create a preview for BAM**. Open the result file **ML3_subset_extracted.prev.sam**.
-Can you see the UMIs in read names? Do there seem to be PCR duplicates?
-Can you see short reads?

Do NOT run STAR now during the course, you can try it afterwards if you like with these parameters:
Select both **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 (2 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

9. Run this in the very end (it takes a lot of time). Check the alignment-level quality with RseQC

Select **ML3_subset_extracted.bam** and the tool **Quality control / RNA-seq quality metrics with RseQC**. In parameters set **organism = Mus_musculus.GRCm38.95** and click **Run.** You can follow the run by clicking on the **Jobs** link: select the **RseQC** run, and scroll down in the **Job** panel (bottom-right).

When the result files arrive,
-Inspect the file **ML3_subset_extracted.txt**. How many alignments does the BAM file contain? Is the tag (~read) density higher in exons than in introns?
-Inspect the result file **ML3_subset_extracted.pdf**. Is the coverage uniform along transcripts (check the first plot)?

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

Select the result file **ML3_subset_extracted_deduplicated.bam** and run the tool **Utilities / Create a preview for BAM** again.
-Are the PCR duplicates gone?

11. 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 (2 jobs)**.
-Inspect **ML3_subset_extracted_deduplicated.tsv**. Can you find genes with counts?
-Check in **htseq-count-info.txt** how many alignments were not counted for any gene and why?

12. Create a count table and description file for the experiment
Select both **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 sample and **2** for the LA (Anakinra-treated) sample.

Well done! You can now close this session and move to the next exercise session **course_QuantSeq_KawasakiDisease_6samples** where we start with a ready-made count table containing counts from 6 full samples. With that data we practice
-expression level QC
-differential expression analysis
-annotation of gene identifiers
-heatmap
-pathway analysis