RNA-seq hands-on tutorial using Chipster: From raw reads to differentially expressed genes

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

In this tutorial you start with raw reads (FASTQ files), and learn how to check the quality and strandedness of the data, align reads to genome, and count reads per genes. Then you combine the count files for all the samples in one table, and describe your experimental setup using the phenodata file. You also learn how to check coverage uniformity, and whether novel splice junctions were found. Finally, you detect differentially expressed genes and annotate them.

The data is a small subset of paired end RNA-seq reads from human lung and lymph node samples. Note that when analyzing differential expression you should always have at least 3 biological replicates! We use this small dataset for the first steps of the analysis for the interest of time. In the second exercise set we use a ready-made count table, generated from 10 full size samples: five lung and five lymph node samples.

1. Start Chipster and open a session

Go to **https://chipster.rahtiapp.fi/** and **Log in** using your HAKA (=university) account. In **Training sessions**, select **course_RNAseq_lung_lymphnode_comparison_2samples**. This session has four FASTQ files, two 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 New Tab**.

-How many reads are there in each file and how long are they (click **Configure Columns** and select **Read Length**)?

-Is the base quality good all along the reads?

-Is the base distribution random in the beginning of the reads (you can click on the samples in the **Per Base Sequence Content plot** to see the profiles)?

-Do the reads have adapters left?

3. Assign paired FASTQ files to samples

Select all the FASTQ files. Go to the **Files** panel (bottom-right), click on the three dots, and select **Define Samples**. Check that the forward and reverse identifiers are correct and click **Find Pairs**. Check that the files are paired correctly and click **Save**.

4. Check the strandedness of the reads

Select the two FASTQ files of the **lymphnode sample** and the tool **Quality control / RNA-seq strandedness inference with RSeQC.** Go to parameters and set **Organism = Homo_sapiens.GRCh38** and click **Run tool (1 job)**. Open the resulting **experiment_data.txt**.

-Is the data stranded? Mark down the parameters for HISAT2 and HTSeq.

5. Align reads to reference genome using HISAT2

Select all the FASTQ files and the tool **Alignment / HISAT2 for paired end reads.** Go to parameters and set **genome = Homo_sapiens.GRCh38.108** and **RNA strandness = unstranded**. Click **Run for each sample (2 jobs)**. Open one of the **hisat.log** files.

-What was the overall alignment rate? What percentage of reads aligned concordantly exactly once?

Inspect the contents of a BAM file: Select **lymphnode4a_R1.bam** and run the tool **Utilities / Create a preview for BAM**. Open the result file **lymphnode4a_R1.prev.sam**.

-Is the BAM file sorted? If so, based on what?
-What was the version of HISAT2 used?
-BONUS EXERCISE: Look at the CIGAR strings and scroll down. Can you spot a read which spans an intron?

6. Perform alignment level quality check with RSeQC

NOTE: This run takes a long time, so we put it to run when we finish today and inspect the results tomorrow.
Select **lymphnode4a_R1.bam** and the tool **Quality control / RNA-seq quality metrics with RSeQC**. In parameters set **organism = Homo_sapiens.GRCh38.108** 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 **lymphnode4a_R1.txt**. How many alignments does the BAM file contain? Is the tag (~read) density higher in exons than in introns?
-Inspect the result file **lymphnode4a_R1.pdf**. Is the coverage uniform along transcripts (check the first plot)? Were novel splice junctions found (check the splice junctions plot)?

7. Count reads per genes using HTSeq

Select both **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 = **Homo_sapiens.GRCh38.108**
Does the BAM file contain paired-end data = yes
Is the data stranded and how = **no in HTSeq**
Run the tool by clicking **Run for each file (2 jobs)**.

-Inspect **lymphnode4a_R1.tsv**. Can you find genes with counts?
-Inspect **htseq-count-info.txt** for the lymph node sample. How many alignments were not counted for any gene and why?

8. 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 lung sample and **2** for the lymph node sample.

9. Detect differentially expressed genes with edgeR

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

10. Annotate the list of differentially expressed genes

Select the file **de-list-edger.tsv** and run the tool **Utilities / Annotate Ensembl identifiers**.
-If you look at the description of the genes with smallest padj-values, do you see lung-specific functions (e.g. surfactant proteins or mucins)?

11. View a summary of the analysis steps taken

Select **annotated.tsv**, click on the three dots, and select **History** in the visualization panel.

12. Share your analysis session with a colleague

Select your session, click on the three dots, and select **Share**. Create a new access rule: enter **haka/ekorpela@csc.fi** in the UserID field, set **Rights = read-only**, and click **Save**.