## June 2017

# An Introduction to Deep-Sequencing Data Analysis

## Exercise #6: Analysing transcript assembly of RNA-Seq data

***Dena Leshkowitz***

### Introduction

In this exercise we will run and evaluate assembly of transcripts using RNA-Seq data and a model genome (mouse). We will use RNA-Seq data from the article [Using Synthetic Mouse Spike-In Transcripts to Evaluate RNA-Seq Analysis Tools.](https://www.ncbi.nlm.nih.gov/pubmed/27100792) PLoS One. 2016 21;11(4). In this study, in vitro synthesized mouse spike-in control transcripts were added to the total RNA of differentiating mouse embryonic bodies, and their expression patterns were measured. This approach enables assessing the accuracy of the assembly tools.

Answers for exercise 5 are found at- http://dors.weizmann.ac.il/course/AnswersForExercise5.docx

### Instructions

We will analyze one sample from the experiment explained above, namely sample C2 which is ES from day 0. The library was created from total RNA (including spike-in mixes) and was processed using the Illumina TruSeq Strand Specific total RNA with RiboZero Gold protocol (Illumina). The source for spike-in was in*-vitro* transcription (IVT) products of plasmid made from cDNA clones from the Fantom2 mouse cDNABook collection (RIKEN, Japan). The spike-ins were selected based on previous knowledge that the gene is not expressed in our biological system.

A list of the spike-ins and the quantities added to this sample are found in ~/course\_2017/ES\_RNA/journal.pone.0153782.s009.docx (download to open).

You will build transcripts from the bam file created with Tophat using ~50M Paired-end 100-bp reads sequences.

1. Run the Stringtie program
   1. Make directory (folder) C2-stringtie
   2. Load module stringtie
   3. Run the following command from C2-stringtie folder

*bsub -q bio-guest -R "rusage[mem=50000] span[hosts=1]" -o ./string\_job.txt -e ./string\_error.txt -J strin -n 10 "stringtie -p 10 ~/course\_2017/ES\_RNA/C2.bam -o C2\_stringtie.gtf"*

* 1. Run an additional stringtie which is guided , using an additional parameter:

-G ~/course\_2017/ES\_RNA/AK\_igenomes\_mm10\_final\_resub\_011215.gtf

Remember to change the name of the output from the previous run.

* 1. Look at the gtf file outputs using the command -more.

Gtf is a file format used to hold information about a [gene](https://en.wikipedia.org/wiki/Gene) structure (see details <http://mblab.wustl.edu/GTF22.html>)

1. Inspecting the built Cufflinks and Stringtie transcripts
   1. Open IGV viewer (see exercise 1) and
   2. Change the genome to mm10
   3. Load the two gtf files that were created by stringtie
   4. Load the three gtf files found in ~/course\_2017/ES\_RNA/

Two are cufflinks outputs that we have prepared for you, one of the outputs was built using an annotation files (C2\_cufflinks\_rabt\_transcripts.gtf). The second output was built without any guide (C2\_cufflinks\_de\_novo\_transcripts.gtf) the third file (AK\_igenomes\_mm10\_final\_resub\_011215.gtf) is refSeq annotation that includes the spike-ins transcripts.

* 1. Load ~/course\_2017/ES\_RNA/C2.bam
  2. Expand the gtf tracks by right clicking on the gtf file name in the IGV viewer and select the option “Expanded”
  3. Go to gene Pomc – this locus contains three spike-ins

|  |
| --- |
| AK030714 |
| AK017581 |
| AK017492 |

These are the only transcripts that are supposed to be assembled since the endogenous gene is not expressed in this RNA sample.

**Question: What is the accuracy of the assemblies we see here?**

The criteria to judge accuracy is the number of transcripts assembled and how well the exons agree with the spike-ins exons, take into account also agreement in exons boarders.

If time permits, look at other spike-in genes such as Acbd4 and Jph3. For more spike-in genes open: ~/course\_2017/ES\_RNA/journal.pone.0153782.s009.docx (download first).

### Good job. You completed this exercise.