CENG 465
Spring 2018-2019

Due Date: May 14, 2019, 23:59

Assignment #4

Finding Differential Expressed Genes in an RNA-Seq Experiment

In this assignment, your goal is to use any existing workflow of your choice to determine the differentially expressed genes in an RNA-Seq Experiment. You will start with raw sequencing data in SRA format and your analysis will results in a list of differentially expressed genes (both upregulated and downregulated) in some condition with respect to a baseline/control condition. You will also identify whether there is a significanly enriched biological process in the list of differentiall expressed genes you find.

You will analyze the following public dataset available at NCBI GEO:

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE54846

The dataset contains six RNA-Seq samples to investigate the effects of the knockdown of the macroH2A1 in cancer cells. Luciferase knockdown, with no effect on the investigated mechanism, is used as control. In other words, the differentially expressed genes in the following three samples:

GSM1325077	macroH2A1 KD replicate 1
GSM1325078	macroH2A1 KD replicate 2
GSM1325079	macroH2A1 KD replicate 3

are sought against the following set of background control samples:

GSM1325074	Luc KD replicate 1
GSM1325075	Luc KD replicate 2
GSM1325076	Luc KD replicate 3

All of the short reads in these samples are obtained using the Illumina HiSeq 2500 technology as “unpaired” single-end reads.

You will first need to use the SRA Toolkit to convert/to download the SRA files to/as FASTQ files. See the following link for more information:

https://www.ncbi.nlm.nih.gov/books/NBK158900/

https://ncbi.github.io/sra-tools/fastq-dump.html

Then you are free to use any workflow for finding differentially expressed genes in RNA-Seq datasets. Refer to the following materials for further assistance:

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0190152

https://f1000research.com/articles/4-1070/v2

http://user.ceng.metu.edu.tr/~tcan/ceng465_s1819/Schedule/NGS_RNASeq.pdf

You may also use existing workflows on online/cloud based platforms such as the Galaxy Toolkit (https://usegalaxy.org/) or the Cancer Genomics on the Cloud (CGC) by Seven Bridges (https://cgc-accounts.sbgenomics.com/auth/register).

After you determine the list of differentially expressed genes, use a gene set enrichment tool such as DAVID (https://david.ncifcrf.gov/gene2gene.jsp) to determine the biological processes that may be involved in the studied biological condition.

Submission:

We wanted this assignment to be a Kaggle like competition and for this, you will need to register at https://darwin.2strand.com/competitions and submit your solution as a .zip file which should contain a short description of the workflow you have used (the sequence of tools and the specific parameters used in these tools), your list of differentially expressed genes (as a plain text file or as an Excel file), and a short report about the biological processes you have found enriched in the list of genes (i.e., result of the DAVID analysis). Also submit a copy of the zip file in ODTU-Class as a regular assignment submission. Follow the announcements coming from ODTU-Class for instructions on registering for the competition and submitting your results to the competiton interface.