## June 2018

# An Introduction to Deep-Sequencing Data Analysis

## Final Exercise

***Ester Feldmesser and Dena Leshkowitz***

### Introduction

The data for this exercise was downloaded from GEO (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE77105>). This is the summary of the experiment as it appears on the website: “MicroRNAs (miRNAs) are important regulators of cell fate decisions in immune responses. They act by coordinate repression of multiple target genes, a property that we exploited to uncover regulatory networks that govern T helper-2 (Th2) cells. A functional screen of individual miRNAs in primary T cells uncovered multiple miRNAs that inhibited Th2 cell differentiation. Among these were miR-24 and miR-27, which each functioned independently to limit interleukin-4 (IL-4) production”.

In this experiment, gene expression analysis of miRNA-deficient mouse CD4+ T cells transfected with miRNA mimics was compared to gene expression of cells transfected with miR-23, miR-24 or miR-27.

The experiment was performed in mice cultured cells in quadruplicates, for the exercise we choose three of the treatments in duplicates.

### Assignment Instructions

Read all the instructions before you start the exercise.

1. You need to submit a report by email (Dena.Leshkowitz@weizmann.ac.il and Ester.Feldmesser@weizmann.ac.il) until July 15th. The report should include the specified below.
2. The fastq files are located in your home directory in ~/utap\_data/Mouse\_input. There is one folder for each sample and inside each folder there is a fastq file. There is also a file called mouse\_pheno\_data.txt under the course2018 directory that relates between the fastq name and the biological sample information.
3. Create in your home directory a folder called Final\_exercise. Perform quality control for all the samples using FastQC as was done in exercise2. Make a new folder for the fastqc output under the final exercise folder. Look at the results and write in the **report** a short summary regarding the sequence quality.
4. Run the UTAP pipeline: <http://ngspipe.wexac.weizmann.ac.il:7000> for the mouse samples.

Use the Mus musculus (mm10) genome and as annotation mm10 (Refseq). Do not change the default for the output folder. The protocol used for preparing these sample libraries was un-stranded. Run Deseq through the pipeline. To define the attributes use the mouse\_pheno\_data.txt to see which sample belongs to which treatment. **Include in the report a link to the pipeline output.**

1. Next step will be to perform K-means clustering on the differentially expressed genes.

You can continue from here in two ways.

1. The first way starts from using an output derived from the UTAP pipeline:
	1. Open the UTAP results report in a browser, go to “Links to Results” (on the left side of report) and download the file that includes “Quantification data including: raw counts, normalized counts and rld (log normalized counts) and pairwise deseq2 statistics” by clicking on the word “here” (red arrow in the figure below).



* 1. Open the file in Excel. Make sure that the symbol names are defined in the import as text.



* 1. We will select genes for clustering. We are interested in the comparison of each of the miRNAs to the control (miRNA mimics). You need to filter the genes that are differentially expressed in any of the two relevant comparisons and prepare a text file with these gene names and their rld values. The easiest way to filter is by writing a formula in a new column, in row2.

Find the columns containing the pass information for each contrast and use them in a query such as:

=OR(AD2="yes",AJ2="yes"). Copy the formula to all the other cells (genes) in the column. Apply a filter to this column and choose the genes that correspond to TRUE.

Copy to a new file the filtered gene names and the rld values.

Remove “.rld” from column names (we do not want long names). Save the new file as text.

* 1. Upload the txt file that you created to the final exercise folder.
	2. Copy the Rmd script to the Final\_exercise with the command:

*cp ~/course\_2018/Arabidopsis\_input/Arabidopsis\_script.Rmd ~/*Final\_exercise

* 1. Open the WEXAC Rstudio through the web. Open the R script that you just copied.
	2. Save it with a new name in the Final\_exercise directory.

Now you are ready to start changing it.

* 1. Clean the Global Environment window at the upper right panel by clicking on the broom.
	2. You need to run the script starting from the “Clustering of differentially expressed genes” section. First, you need to do several changes.
		1. Read the text file with the differentially expressed genes using the read.table command (*data <- read.table("file.txt")*). Change the command that appears in line 242 by this one. The table you just read in is a matrix equivalent to the one called data in this section.
		2. After the scaling, you need to add column names (sample names), change the command in line 252 by *colnames(data.scaled)<-colnames(data)*.
	3. You do not need to order the samples, they are already ordered.
	4. Perform clustering using K-means in a similar way to the used in the clustering exercise (exercise 5) and create a heatmap. Try at least 2 and 3 clusters. Include the path to the plots and tables in the **report**.
	5. Note that you have less samples than in the Arabidopsis exercise, therefore you will need to change the column numbers in the K-means clustering procedure, in the commands: order and pheatmap. Take into account that you have six samples so that in m.kmeans object the cluster number is in column 7.
	6. When the script runs correctly, you can delete the irrelevant lines and create an html report. This html is part of the final exercise (we will look for it in the Final\_exercise folder).
1. The second way starts from the count files produced by STAR in a similar way to exercise five. Your ReadsPerGene.out.tab files are located in the pipeline output directory (inside utap\_data) under the 3\_mapping directory. Be sure that these are the files you are using in the Rmd script. The mouse phenodata file is located in course\_2018. You need to copy it to the exercise folder and read it into samples (line 46 in Arabidopsis\_script.Rmd).

See directions e-m in the previous paragraph, except for i that is not relevant.

1. Include the path to the clustering html file in your final **report**. Describe which number of clusters you chose.
2. Let us assume that you are interested in the specific effect of miRNA24. Choose the cluster that best reveals this effect and perform a functional analysis in David, InterMine for mouse (MouseMine) or GeneAnalytics (GeneCards suite). In David, use as background the default mouse background.
3. In additions, perform an analysis in GSEA, using the log2fold changes between miRNA24 and the control. To convert mouse symbols into human symbols as required GSEA use the table in http://dors2.weizmann.ac.il/course/course2018/ Mouse2human\_id.xlsx. The table was created from a table downloaded from the Mouse Genome Informatics (MGI) website (<http://www.informatics.jax.org/downloads/reports/HMD_HumanPhenotype.rpt>). Convert the mouse gene symbols to human using the Excel function vlookup. A detailed explanation of the function can be found in <https://www.timeatlas.com/vlookup-tutorial/>. Save the converted file as txt and after saving it change the suffix to rnk. The suffix of the text file name to upload to GSEA should be “.rnk”. Perform the analysis against the Hallmark gene sets.
4. Pay attention where the GSEA results folder is saved. Upload this folder to the Final\_exercise directory or zip it and send us by email with the final exercise report.
5. Summarize in the **report** what is the main specific biological effect/s of miRNA24 transfection.