***December 2019***

# An Introduction to Deep-Sequencing Data Analysis
# Exercise 4: UTAP: User-friendly Transcriptome Analysis Pipeline

**Dena Leshkowitz and Bareket Dassa**

Please read all the instructions before you start the exercise.

**In this assignment**, you will practice how to set up a transcriptome analysis of MARS-seq data, using UTAP web-based pipeline. In the second part of the exercise you will analyze the results based on the UTAP report, and answer the questions below. Note: UTAP run may take a few hours.

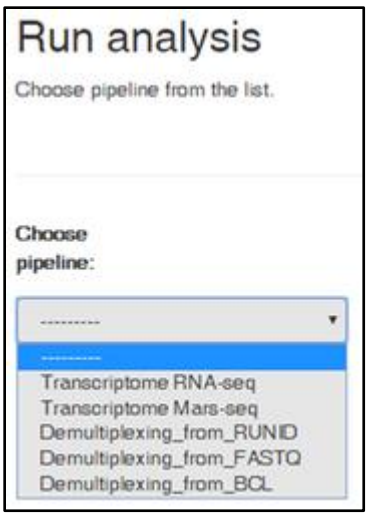
**Background:** Last year, students at the Weizmann sandbox course prepared MARS-seq libraries, for detecting changes in gene expression, upon treatment of mouse cells with Lipopolysaccharide (LPS). Eight teams of students prepared the libraries for NGS analysis. In this exercise you will analyze the differences between treated and untreated (control) cells in their experiment.

You will use the UTAP, a User-friendly Transcriptome Analysis Pipeline, which was developed at the Weizmann Bioinformatics unit. Using UTAP you will execute the full process, starting from sequences and ending with sets of differentially expressed genes summarized in a comprehensive report.

## ***Part I: Set-up a transcriptome analysis***

A. Login to “UTAP: User-friendly Transcriptome Analysis Pipeline” at:
   http://ngspipe.wexac.weizmann.ac.il:7000
   Use your class Username and password.
   (The page is accessible only inside the Weizmann network, and is compatible with Firefox and Chrome browser, but not with InternetExplorer).

B. Click "Run pipeline" from the top menu. Choose “Transcriptome MARS-seq” analysis:

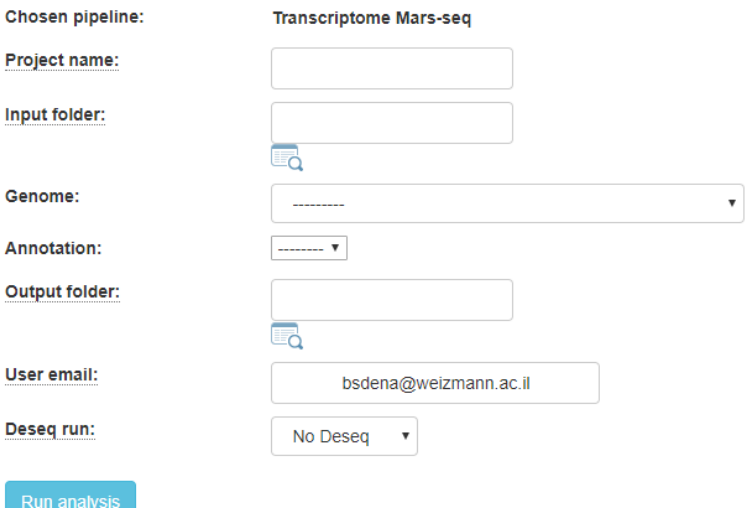



Choose a project name, for example “Class1-exercise”

C. **Select the input folder:** The raw fastq files are saved in sub-folders within the folder “190411_NB551168_0309_AHNHGTBGX9**”.** Select the entire folder as input. Once you select the input folder, a list of all samples will be displayed. They are listed as they appear in the table below (item F).

**Note-** if you do not find this input folder, please contact us.

The **output** folder will be selected automatically, **please do not change it**!

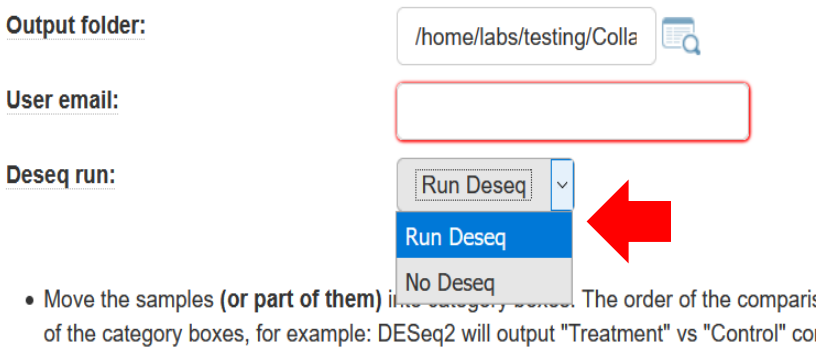



D. **Additional set ups:**
Select the **reference genome** (Mus musculus (mm10)) and **annotation** (the Refseq annotation) for the reads alignment.

E. **Change the email to your personal email.**

F. **Differential gene expression analysis:**

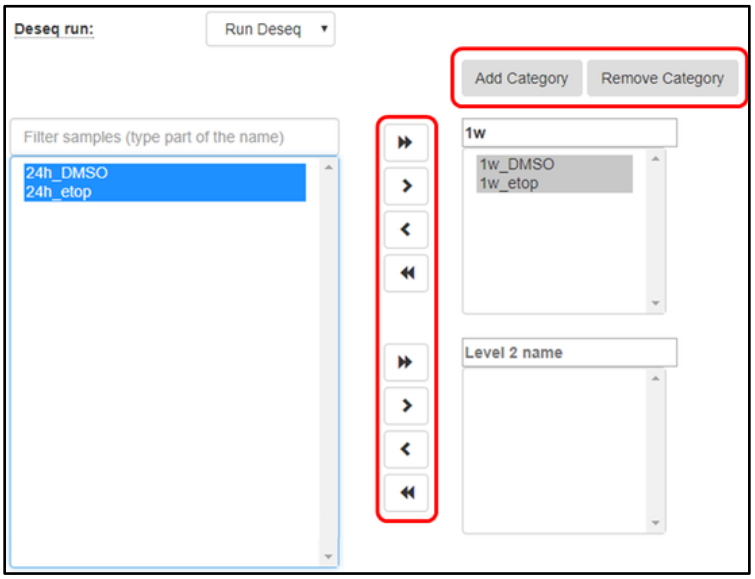
Choosing “run DESeq2” allows you to detect differentially expressed genes in your data using the DESeq2 package (DESeq2 manual):




You will then need to define the samples to categories. Use the following table to assign the samples into two treatment categories (LPS and control). Sort the samples accordingly as shown in the example figure below.

| Sample name | Treatment | Replicate |
|---|---|---|
| Team1_C1 | control | 1 |
| Team1_C2 | control | 2 |
| Team1_LPS1 | LPS | 1 |
| Team1_LPS2 | LPS | 2 |
| Team2_C1 | control | 1 |
| Team2_C2 | control | 2 |
| Team2_LPS1 | LPS | 1 |
| Team2_LPS2 | LPS | 2 |
| Team3_C1 | control | 1 |
| Team3_C2 | control | 2 |
| Team3_LPS1 | LPS | 1 |
| Team3_LPS2 | LPS | 2 |
| Team4_C1 | control | 1 |
| Team4_C2 | control | 2 |
| Team4_LPS1 | LPS | 1 |
| Team4_LPS2 | LPS | 2 |
| Team5_C1 | control | 1 |
| Team5_C2 | control | 2 |
| Team5_LPS1 | LPS | 1 |
| Team5_LPS2 | LPS | 2 |
| Team6_C1 | control | 1 |
| Team6_C2 | control | 2 |

| | | |
|---|---|---|
| Team6_LPS1 | LPS | 1 |
| Team6_LPS2 | LPS | 2 |
| Team7_C1 | control | 1 |
| Team7_C2 | control | 2 |
| Team7_LPS1 | LPS | 1 |
| Team7_LPS2 | LPS | 2 |
| Team8_C1 | control | 1 |
| Team8_C2 | control | 2 |
| Team8_LPS1 | LPS | 1 |
| Team8_LPS2 | LPS | 2 |




G. **Run the pipeline**

Finally, submit the run for analysis. Once the analysis is completed, you will be notified by email (usually lasts a few hours) and you will be able to open a report page with the results. You can check the status of the run from the “User Datasets” at the UTAP top menu.

## ***Part II: Questions***

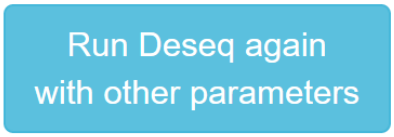
For this exercise start with creating a folder called “Exercise_4” in your home directory. Save the answers to this exercise in a file inside this folder.

1. Once the run has ended, copy and write down the link to your UTAP report.
2. Open the report and answer the following questions:
   Look at the number of reads in each sample (Fig 2). We recommend on starting the analysis with 5 million reads per sample. Name which samples do not meet the minimal number of reads?

3. From Fig 6, calculate what is the average percentage of reads that were uniquely mapped to the genome? (use the “Download figure 6 as table” link)
4. What could be the reason that not all reads are uniquely mapped to the genome?
5. Principle component analysis (PCA) is a statistical procedure used to reduce the dimensions of a data set, allowing the description of data sets and their variance with a reduced number of variables. You will learn about it in the next lecture, but in this report it is used to assess overall similarity between samples, where similar samples should be grouped together.
   Look at the PCA plot in Fig 10 (PC1 vs. PC2). Identify which are the outlier samples? It seems that the team with outlier samples switched the labels between their samples. You will need to reassign the outlier samples to their correct groups in the next question.
6. Rerun the DESeq analysis taking into consideration:
   A. The samples switching
   B. The fact that we had a batch effect, since different teams prepared the samples. Therefore, you need to consider the library of each team as a different **batch** of sample preparation.

When re-running DESeq analysis, mark the different teams as batches:

I. Choose the “Use Datasets” link at the top menu, and click on this button to set-up a different DESeq comparison.




II. Select the samples into categories boxes (take into consideration the samples that were switched), and click on "Add Batch Effect" button, then mark the samples that belongs to one batch and click on "Batch 1" button. Repeat this operation to assign all the samples into batches (of teams).

7. The rerun creates a new report. Once the rerun has ended, copy the link to the new re-run report.
8. In the new analysis report look at Table 2 ("**Table 2: Links to functional enrichments analysis**") and state:
   A. How many genes were detected as upregulated in the LPS treatment?
   B. How many genes were detected as downregulated?
9. Are the numbers in question 8 different from the results in the first UTAP run, if yes why?

# Good luck!