Final Project

Submission due dates

- Workflow – February 16th, 2025 (or earlier)
- Final project – March 27th, 2025
- Presentations – First week of next semester (to be announced)

We recommend submitting your workflow ASAP to give yourself more time for the main project

General instructions

- Submissions in pairs.
- Presentations in English or Hebrew.

Project steps

1. Workflow submission
2. Bioinformatic analysis
3. Abstract
4. 15-minutes presentation

Project goal

Using bioinformatics tools to extract meaningful insights from genomic data.

Detailed instructions

1. **Workflow submission: find datasets and phrase a biological question**

To ensure your efforts are focused in the right direction, we ask you to submit a workflow. This step is crucial so that you won't waste time analyzing unsuitable data. We encourage working on a biological question related to a disease you are personally interested in. Although there is a huge amount of publicly available datasets, it is sometimes challenging to find those that meet your needs. Therefore, we recommend first making a list of all datasets related to a specific disease you found, and only then coming up with a biological question that can be answered by analyzing one or more of those datasets.

For instructions on finding the "right" datasets watch the tutorial recordings.

The workflow should include the following:

- Students' names and IDs.
- Your disease of choice.
- Which biological question(s) do you want to answer?
- How do you plan to address those questions using the datasets below?
- Which secondary analysis are you planning to conduct?
- A table of the datasets (at least five) you collected with the following information:

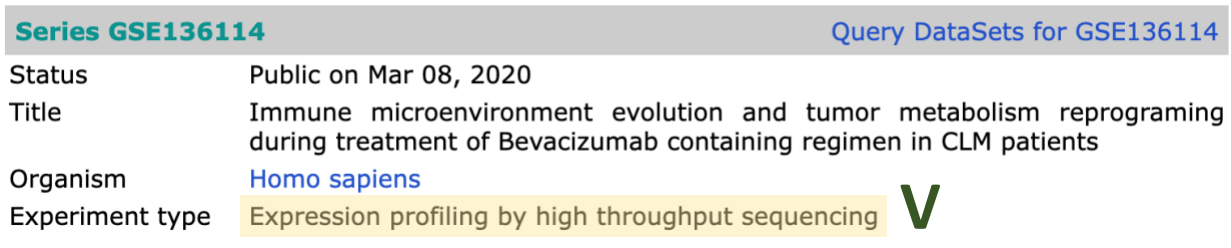
1. Dataset accession ID
2. Link for the website where the data can be found
3. Type of the data (bulk/scRNA-Seq/GWAS summary/Survival)
4. For gene expression data: Is the data in raw counts or normalized?*
5. Data description

* We prefer gene expression data in counts (not normalized) since they are suitable for differential gene expression analysis with DESeq2. However, if you wish to use normalized data, you can use a similar package called **limma.**

Example:

Accession ID	*Location*	*Data type*	*Normalization?*	*Description*
GSE58434	*https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58434*	*Bulk RNA-seq*	*FPKM normalized*	*12 samples in total, 6 of them are healthy control, and 6 are asthma patients in which 3 have taken Dexamethasone and 3 Albuterol*

- **Make sure that you download sequencing data (not data by array)!**
  For example:



Series GSE20271	Query DataSets for GSE20271
Status	Public on Sep 15, 2010
Title	Expression data from breast cancer FNA biopsies from patients
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	The behavior of breast cancers and their response to different chemotherapy

- **Do not use RAW sequencing files (FASTQ/FASTA)!**

Supplementa…	Size	Download	File type/resource
GSE20271_RAW.tar	590.3 Mb	(http)(custom)	TAR (of CEL)

- Look for "counts" data
  For example:

Supplementary file	Size	Download	File type/resource
GSE145325_rawcounts_valencia.csv.gz	2.7 Mb	(ftp)(http)	CSV

And sometimes you will have this button to access the counts data:

Download RNA-seq counts

- **If you are not sure whether the data is in counts, download, open the expression matrix, and check if the values are integers.**
- **Normalized gene expression data might be in various forms such as TPM, FPKM, RPKM, and CPM. You can still use them, but not with DESeq2.**

Supplementary file	Size	Download	File type/resource
GSE150368_Count_matrix.txt.gz	1.0 Mb	(ftp)(http)	TXT
GSE150368_RAW.tar	35.2 Mb	(http)(custom)	TAR (of TXT)
GSE150368_transcript_tpm_all_samples.txt.gz	11.4 Mb	(ftp)(http)	TXT

- **Send the workflow to Almog (almog.angel@campus.technion.ac.il) as a PDF file with the students' IDs in the file name (ID1_ID2.pdf). The title of the e-mail should be "Workflow submission ID1 ID2".**
- **Only if your workflow is approved and signed by Almog/Dvir you are allowed to proceed to the next steps.**

2. Bioinformatic analysis (45 points)

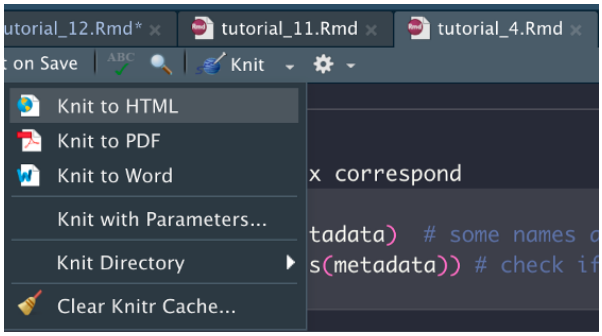
- Keep your code as clean as possible
- Use comments with proper documentation for each step in your analysis (explain what happens in this step).
- Make sure that you create beautiful plots.
- Summarize your project in the abstract.

The analysis should consist of two parts. The main analysis would be done using either bulk or scRNA-Seq. The secondary analysis can be done by using the other method (scRNA-Seq if you chose bulk as the main analysis and vice versa) or other methods we learned in class such as GWAS and survival analysis.

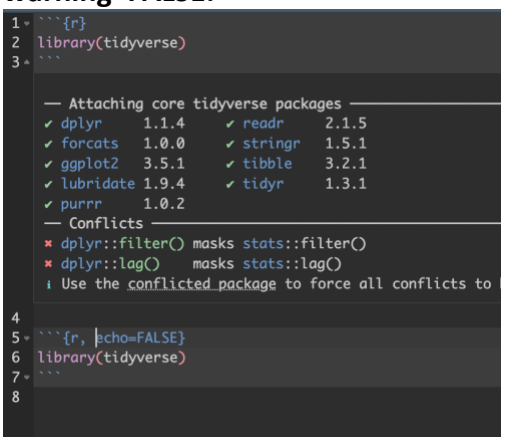
Main analysis:
Use at least one of the datasets from step 1 (more is recommended) and perform all relevant analyses. Your main analysis should be as comprehensive as possible, covering multiple methods that we learned in class/tutorials.

Secondary analysis:
The secondary analysis should address a different type of data. The main and secondary analyses should be related somehow. Namely, they should target the same biological question or at least the same disease. For some topics, such as cancer, using a second analysis such as survival analysis is straightforward. However, for other types of diseases, it can be challenging to find such data. If you would like to use GWAS as a second analysis, look for "GWAS summary statistics", as raw genotyping data of humans is usually confidential. Otherwise, you can perform the secondary analysis with gene expression data of a different type.

Document your analysis in R markdown and when you are done save it as an HTML file. To do so you need to click on "Knit to HTML":



Note: Do not present irrelevant console messages. Use echo=FALSE, message=FALSE, or warning=FALSE:



We will grade this step based on:

a. How comprehensive is your analysis.
b. The quality of your analysis (make sure you do not forget steps and use the right functions properly).
c. How well you documented the analysis.
d. The plots (figures) you generate.

3. Abstract (10 points)

- Write an abstract for your project (in English, up to 300 words).
- Include the abstract at the beginning of your R markdown file.

There are six steps to writing a standard abstract:

1. Begin with an introduction about the medical condition you chose.
2. State the problem or knowledge gap related to this disease you wish to address.
3. Write a brief description of what you have done (your analysis strategy).
4. Describe the most meaningful outcome(s) of your analysis.
5. Close your abstract by discussing the broad implication(s) of your findings.

4. Presentation (45 points)

- The presentation will take 15 minutes (including questions).
- Both students should present as equally as possible.
- **For more recommendations and instructions check Dvir's final project presentation.**

We will grade this step based on:

a. Your presentation appearance and clarity
b. Your understanding of the analysis and results
c. Using correct terminology

Instructions for submission

Submit your project as a ZIP file that contains:

- Workflow signed and approved by Almog/Dvir
- HTML R markdown file of your analysis

Good luck! ☺