Final Project

Submission due dates

- Workflow – 23/02/2023 or earlier
- Final project – 16/03/2023
- Presentation week – 21-23/03/2023 (exact times will be notified later)

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. 10-minutes presentation

Project goal

The goal of the project is to extract meaningful insights from genomic data to better understand a certain medical condition. **Note** that by saying meaningful it doesn't mean that you are required to find anything novel or statistically significant. Meaningful could also be that you couldn't get any results given the data and methods you were using.

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 or biological question.

We encourage you to work 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 look at the last tutorial presentation and recording.

The workflow should include the following:

- Students' names and IDs.
- Your disease of choice.

  Example: Asthma

- What biological question do you want to answer?

  Example: We would like to explore how different medications affect the airway smooth muscle cells in asthma.

- How can gene expression data help?

  Example: We will analyze bulk RNA-seq expression data to run a differential expression analysis to compare samples of asthma patients that consumed different medications. In addition, we use the differentially expressed genes to identify enriched gene sets and pathways.

- What other type of analysis are you planning to use and how can it help?

  Example: We will also use GWAS to check whether the differentially expressed genes correspond to statically significant GWAS markers of asthma patients vs. controls.

- 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)
4. Is it a raw/normalized count matrix data?
5. Different groups that can be found in the data

Example:

Accession ID	***Location***	***Data type***	***Count matrix***	***Groups***
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***

- Do not use raw sequencing files!
- If you are not sure that this is a count matrix data, download the data, load it into R and make sure that you have a genes x samples/cells matrix.
- If you wish to use DESeq2, the count matrix should be in raw counts (not normalized!). Make sure that the values in the matrix are integers (they might be normalized to TMP/FPKM/RPKM).
- Send the workflow to Almog (almog.angel@campus.technion.ac.il) as a PDF file with the students' ID as 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 then you can start the next step.**

2. Bioinformatic analysis (45 points)

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

The analysis should consist of two parts: the main analysis would be by using either bulk or scRNA-Seq. The second analysis can be done by using other methods we learned in class such as GWAS, survival analysis, CRISPR screens, etc.

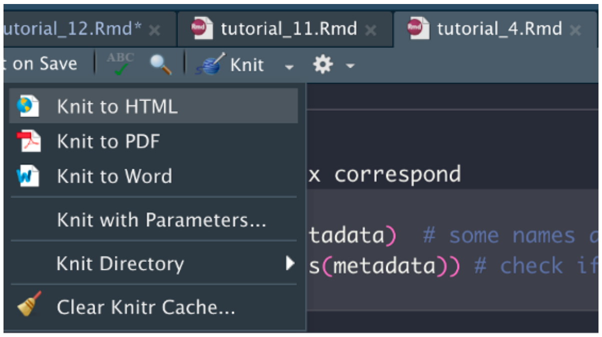
Main analysis:

Use at least one of the datasets from in step 1 and perform all relevant analyses. Your main analysis should be comprehensive as possible, covering multiple methods that we learned in class/tutorials, such as: unsupervised clustering, linear/logistic regression predictions, dimensionality reduction (PCA/tSNE/UMAP), gene set enrichment analyses (and pathway analysis in general), cell type composition analysis (xCell).

Second analysis:

The second analysis should cover one or more of the other methods we learned in class. You can find data for this analysis in the literature, TCGA, DepMap and other repositories that can be found in Google. Note that the main and second analysis should be related. 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 and CRISPR screen is straightforward. However, for other type of diseases it can be challenging. If you like to use GWAS as a second analysis and cannot find data, we recommend using GWAS summary statistics from the GWAS catalog.

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



```
x correspond

tadata)  # some names c
s(metadata)) # check if
```

We will grade this step based on:

a. How comprehensive is your analysis (the more method the better).
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 a broad statement about your topic.
2. State the problem or knowledge gap related to this topic that your study explores.
3. Describe what specific aspect of this problem you investigated.
4. Briefly explain how you went about doing this.

5. Describe the most meaningful outcome(s) of your study.
6. Close your abstract by explaining the broad implication(s) of your findings.

4. Presentation (45 points)

- The presentation will take 10 minutes + 2-3 minutes for questions.
- Both students should present equally as possible.
- **For more recommendation and instructions check Dvir's final project presentation.**

In brief, the presentation should include the following:

1. Introduction of the disease/condition you are dealing with
2. The biological question and motivation for the project
3. The data you use - short description
4. The main results and figures
5. Your conclusions

We will grade this step based on:

a. Your presentation appearance and clarity
b. Your understanding of the analysis and results
c. Using the 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! ☺