Final project - Introduction to Bioinformatics (236523)

Final project submission date: By August 26, 2021

Workflow submission: You must send a workflow of the project at least three weeks before your final submission. The workflow submission deadline is August 5, 2021. We ask that you submit a workflow of your project so we can help you focus your research. The description of what should be included in the workflow is detailed below.

The project should be performed in pairs.

Project goal

The **goal** of the project is to investigate **how genomic data can be useful to better understand diseases**. To do this you will choose a disease, or some aspect of the disease, such as a group of patients that receive some treatment, a complication of the disease, etc. You will then identify all the gene expression data sets that have been deposited in the public domain for this disease. You will choose one (or more) datasets and perform various analyses to extract insights from the data.

Details

Step 1: Go to arrayexpress.com and search for your disease of interest. Sort the results by number of assays. Identify a dataset that has at least a dozen of samples (preferably more) and also has a link to "Expression Atlas" or "Processed data" (we do not expect you to work with raw FASTQ files, although that we will not stop you to go beyond). Preferably, identify more than one dataset, as more data, from more sources, can improve the signal from the noise.

Explore this dataset to be sure that it will be possible to use it for further analysis. You might want to look at the metadata that is coupled with the gene expression data, which will allow to perform supervised analyses (but it might be that unsupervised analyses are the way to go with your data).

Disease ideas: You may choose any disease to work on. You can find some ideas in this link: https://www.dph.illinois.gov/topics-services/diseases-and-conditions/diseases-a-z-list

Note 1: you may choose single-cell RNA-seq data, not just bulk gene expression (RNA-seq or microarray data). We did not learn in class how to analyze microarray data, but it is not substantially different than RNA-seq. You can find here a nice tutorial for analyzing the most abundant type of microarrays: https://bioconductor.org/packages/devel/workflows/vignettes/maEndToEnd/inst/doc/MA-Workflow.html

A nice tutorial for RNA-seq can be found here:
https://www.bioconductor.org/packages/devel/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html

Note 2: If your disease of interest is a type of cancer, we are still required to use a dataset on ArrayExpress, and not use TCGA or cBioPortal as learned in class.

Step 2: Now that you are sure you have data that you will use for your research project on the disease, write a literature review (up to 5 pages, font size 12). Identify multiple relevant studies about this disease (with a major focus on studies that are based on analysis of gene expression data), learn what is already known about the disease, what is unknown, and most importantly how genomic data can help advance our understanding of the disease and may eventually improve therapeutic strategies in the clinic.

In the introduction of the manuscript that you will submit you will need to provide a summary of this literature review:

- Start with an introduction about the disease, and summarize what your learned as described above.
- Provide a summary table of all the datasets you found with links to the studies (not just in those you reviewed, but all relevant datasets you can find in ArrayExpress or GEO). Include the number of samples (number of patients separate from number of controls, if later are available), which tissue/cell type was profiled, and main findings from each dataset.

Step 3: Go back to the dataset(s) you chose in step 1. Download the data and perform all relevant analyses. This includes, but not limited to, analysis types that we learned in class such as differential expression analysis, unsupervised clustering, linear/logistic regression predictions, dimensionality reduction (PCA/tSNE), gene set enrichment analyses (and pathway analysis in general), cell type composition (xCell), and survival analysis. You can also analyze genes that have been previously implicated in the disease in functional studies or genetic studies (e.g. GWAS catalog).

Your goal in this step is to extract data-driven insights. Are your findings consistent with the findings described in other studies? In addition, perform hypothesis-driven analysis based on what you learned from your literature review – can you validate findings described in other studies?

Provide an R markdown file of your analyses and create figures from your main findings. Describe your analyses and findings in the results section of your main document.

Step 4: Write a discussion that summarize what you've learned from your research and which questions are still open. Refer to existing studies in your explanations. Describe a possible theoretic study that can help solve some of those questions – what data is needed to answer the questions.

What to submit?

You should submit a main document with following sections:

- **Title** – provide a title that gives a glimpse to what you did and what you found. Also include your names, affiliation and IDs.
- **Abstract** – one paragraph of up to 250 words. Should include a sentence or two for each of the following:
    - General sentence about the studied disease
    - The knowledge gap
    - What you did
    - What you found
    - What the findings mean
- **Introduction**
    - This is the literature review summary you performed in step 2.
    - Don't forget to include a summary table of all the datasets you found.
    - Provide references to the relevant manuscripts!
- **Results**
    - Description of your analyses and figures from step 3.
    - My tip of writing a result – each paragraph contains the following:
        1. Why you did the analysis
        2. How you did the analysis (in one or two sentences, the full description goes into the methods section)
        3. What you found
        4. What does it mean
    - Most importantly: a figure, or figure panel, that allows to easily understand all 4. Go over the presentation from the last lecture for tips on how to generate an effective figure.
    - Pay attention to put a figure number and a figure legend under each figure. Refer to the figure numbers when you describe the result that can be seen in the figure.
- **Discussion**
    - Your discussion from step 4. What did you learn? What would you do next?
- **Methods**
    - Explain what you did in the results section. Use subsections. The results section should be very succinct in describing what was done.
- **References**

  Use a citation manager to add references to your document. We suggest using Mendeley (the Technion has subscription for it) or Zotero, but you are welcome to use any citation manager.

In addition to the main document, we ask that you submit an **R markdown** file with your code and the downloaded files that were used for the data processing – the original files and

modified in case some changes were done in the files that were used in the analysis. Please add documentation so we can all understand what is going on.

What should the workflow include?

We want you to submit your study workflow, so we can help you focus your research to the right direction. The workflow does not need to include paragraph, just points. It should include the following:

1. Your chosen disease (or disease problem).
2. In points: why this disease is interesting to study and how genomics data can help.
3. The dataset(s) you plan to analyze – provide information on the platform type (e.g. RNA-seq, microarray) and number of samples.
4. List of 5 manuscripts you are planning to review in step 2.
5. Open questions you have for us so we can help you focus. This is your chance to get feedback on your study design and approach.

Good luck!