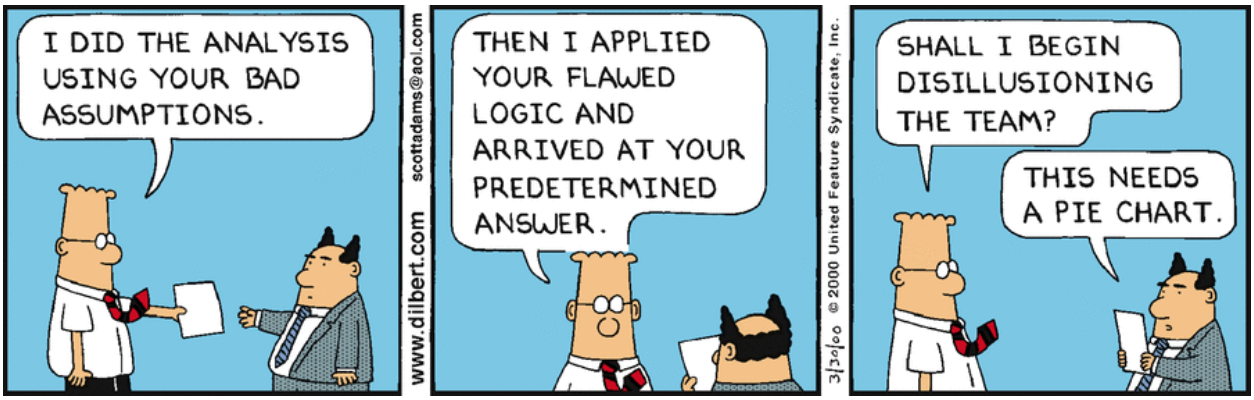
Exam homework for BOHTA 2023

Use this setup chunk to load the libraries you use and define a work directory. I have written in the libraries I used to solve all exercises.



(Remove inspirational picture from your hand-in. It is also fine to remove introduction texts)

Name: Abdullah Faqih Al Mubarok

KU id: vpx267

Instructions

For deadlines etc, see Digital Exam

You have to supply both the answer (numbers, tables, plots, discussion or combinations thereof) as well as the R or Unix code you used to make the plots/results. This should be done using this R markdown template: **We want the R markdown file and the resulting html file (not a PDF!), BUT NOTHING ELSE (no extra tables or similar)**.

The html file is what you will present at the oral exam. At the oral exam, there may be questions on the homework and/or anything else from the course.

In the home work part, all aids are allowed except

i) working with others since it is an individual exam. To be clear, no copying of code or anything else between students or from other sources than the course material (e.g. slides) are allowed. Reading articles etc is of course fine, but then cite them and never copy ad verbatim.

ii) use of chat bots or anything similar.

If we find something like that you will be reported.

In the oral exam part, no aids are allowed except your html file - which we will have ready for you when you arrive. This means no notes or similar.

Note that:

Some questions may request you to use R or UNIX commands /options we have not covered explicitly in the course: this is part of the challenge

Much like in an academic paper, the analysis and results should be presented on a level of detail that someone else could replicate the analysis. If not, we will deduct points.

Use tidyverse for all analyses and plotting, unless we explicitly used something else in class for the relevant plot or analysis (e.g. plotting trees, expression analysis on matrices, heat maps, etc)

Do not use additional libraries than those we used in class (my solution for the home work only uses tidyverse, deSeq2 and ggrepel)

For statistical tests, you have to:

Motivate the choice of test

State exactly what the null and alternative hypothesis are (depends on test!) and the P-value threshold

Comment the outcome: do you reject the null hypothesis or not, and what does this mean for the actual question we wanted to answer (interpretation)?

Question 1- What transcription factors bind muscle- and liver-specific enhancers?

All data for this question is in the data_q1 folder, except enhancer regions which are defined by the slidebase tool - see below

Enhancers are regulatory regions far away from genes, but regulate genes through 3d looping to the gene promoter. Such enhancers are often tissue-specific, and are bound by specific transcription factors that are used in the specific tissue. Earlier work has established that liver-specific regulation is mediated by HNF1, HNF3, HNF4, HNF6, and CCAAT/enhancer binding protein (C/EBP) transcription factors while skeletal muscle specific regulation is mediated by MyoD and Myf- factors, MEF1, SREBP and TEF1.

In this question, we will explore whether we can confirm these earlier observations using two genomics data sets: the FANTOM5 enhancer database (predicted enhancer regions across nearly all tissues and cells -see https://www.nature.com/articles/nature12787) and ENCODE ChIP-seq peaks from the UCSC browser.

To find liver- and muscle-specific enhancers, we will use the SlideBase resource (https://slidebase.binf.ku.dk/) : this is a selection tool where one can select enhancer regions based on their activity in a given tissue. First, view the videos on the SlideBase web site to learn how it works. In the tutorial, enhancer selection is used as an example.

1: Use SlideBase to define two bed files corresponding to liver- and skeletal muscle- specific enhancers. Use the 'Organ expression', not the 'Cell expression' sliders, and obtain two bed files:

1: Enhancer regions where >=80% of expression comes from liver

2: Enhancer regions where >=80% of expression comes from skeletal muscle

How many enhancers are selected in each? How much do they overlap?

Solution goes here:

2: The chip_tfbs.bed file in the data_q1 folder is taken from the UCSC browser: it is the ChIP-seq peaks for a large number of cells from the ENCODE project. Each row is one ChIP peak where the name corresponds to the ChIPed transcription factor. Find a way to overlap the enhancers we defined above with these ChIP results.

Overlap each enhancer set with the ChIP-seq regions. What are the 10 transcription factors whose ChIP-sites overlap the most with liver specific enhancers? What are the top 10 for muscle specific enhancers? Show the results in two tables using R. Do the transcription factor names match the previously established liver- and muscle-specific transcription factors that were discussed in the introduction?

Solution goes here:

3: To make a proper comparison of counts of transcription factors in each enhancer set, make an xy plot where y is the number of overlaps in liver-specific enhancers, x is he number of overlaps in muscle-specific enhancers, and each dot is one transcription factor. Make sure that

you show if points are overlapping each other in a good way. The dots should also show the name of the factor, at least for the most interesting dots - geom_text_repel() is recommended from the ggrepel package.

Tip: pivot_wider() may be useful.

Comment on your plot: what is shared and what are those transcription factors: what factors are mostly binding muscle- or liver-specific enhancers only?

Solution goes here:

Q2: Is DNA accessibility associated to cancer sub types?

All data for this question is in the data_q2 folder

In the data_q2 folder, there is a large file called atac_brca.tsv. This shows ATAC-seq peak intensities from samples from 74 breast cancer patients. Specifically:

Each row shows one patient (sample).

Column 1 shows patient ID

Column 2 shows a PAM50 type - a classification for cancer sub types - in this data, there are six PAM50 types

Columns 3...N shows all ATAC-seq peaks and their intensity

The ATAC-seq peak intensity is measuring how accessible that part of the DNA is. The reason this experiment was made was that there was a hypothesis that DNA accessibility can tell us something about breast cancer and in particular breast cancer sub type.

Let's find out if this is true.

1: Make a PCA of the ATAC-seq data where dots indicate patients. You should scale and center the data, and show percent explained variance on axes. Do note this is a big file: it make take a bit more time to analyze than in class exercises. Comment on your plot

Solution goes here:

2: The PAM50 classification has six classes. We would hope those classes correspond to clusters in the data, and in particular one the PCA that we made. Use k means where k =6 on the atac-seq matrix. Then:

i) Visualize both PAM50 classification and what k means cluster each patient belong to in the PCA plot (with % variance indicated on axes) and

ii) Make a table that shows the overlaps between k means and PAM50 classifications (e.g. how many patients are in k means group 1 and PAM50 classification LumA, etc etc)

In the ideal scenario, these k means and PAM50 classifications would agree almost 100%. Do they? Discuss and interpret the outcome.

Solution goes here:

Question 3: Helping Novo Nordisk

All data for this question is in the data_q3 folder.

You have been hired by the Danish pharmaceutical giant Novo Nordisk to analyze an RNA-Seq study they have recently conducted. The study involves treatment of pancreatic islet cells with new experimental drugs for treatment of type 2 diabetes. Novo Nordisk wants to investigate how the drugs affects cellular mRNA levels in general, and whether the expression of key groups of genes are affected.

As the patent for the new experimental drug is still pending, Novo Nordisk has censored the names of genes. The experiment uses four treatments, A,B,C,D - we are not really told what these are, except that C and D are two different types of controls, but there is a suspicion these two controls are so similar that should be merged into one group. This is one of your jobs to find out.

You have been supplied with 4 files in the question 3 data folder:

- `studyDesign.tsv`: File describing treatment of the 30 samples included in the study.
- `countMatrix.tsv`: Number of RNA-Seq reads mapping to each of the genes.
- `normalizedMatrix.tsv`: Normalized expression to each of the genes.
- `diabetesGene.tsv`: Collection of IDs of genes known to be involved in type 2 diabetes.

1: Read all data sets into R, and make sure the top three files have matching numbers and names of both samples and genes

Solution goes here

2: Use a PCA to see how well separated the groups are, if there are any outliers and whether it makes sense to make C and D into a single control group. You should center but not scale the data before PCA. As above, indicate % variance on axes.

Solution goes here

3: Discuss the PCA and make the appropriate changes - remove potential outliers if any, decide if C+D samples should be merged into one control group and if so, do that without editing the files. If you decide not to merge, select one of C or D to be the control group that we will use below. Regardless, we will compare A and B to this control group below and we will call it 'control'. Make a new PCA plot after making the changes and comment on that. As above, indicate % variance on axes.

Solution goes here

4: Use DESeq2 to obtain differentially expressed (DE) genes comparing i) A vs control and ii) B vs control. Use default parameters, except use a logFC threshold of 0.25 and an adjusted P-value threshold of 0.05. How many genes are up- and down-regulated in respective comparison?

Solution goes here

5: To check if our data is fine, make the following 'diagnosis' plots and comment on them for each comparison

For each of these questions, make a single plot with one facet for each comparison

5.1: MA plots. Discuss whether the MA-plots indicate appropriate normalization (max 70 words discussion)

Solution goes here

5.2: P value distributions.

6.3 Volcano plots:

Solution goes here

7: Is any of treatment preferentially acting on diabetes-related genes?

-Novo Nordisk claims that treatments A and B affects up-regulated genes related to diabetes. Your task is to investigate whether this is true. They have supplied you with a long list of genes that are diabetes-related - diabetesGenes.tsv

Principally there are two ways of doing this, by answering two questions:

7.1: Are significantly up-regulated genes (FDR <0.05 and log2FC >0) in a given comparison (defined as above) significantly more likely to be diabetes related compared non-differentially expressed genes?

7.2: For each comparison, do diabetes-related genes have a significantly different change in expression vs control than other genes?

Answer both of these questions using the appropriate visualizations and statistical tests. If you use 2x2 contingency tables, show the tables. Comment your results and compare the answers of the two questions - if they are the same, why is that? If not, why is that?

Solution goes here