Introduction to Bioinformatics (236523)

HW 3 – Spring 2021

**General instructions:**

• Deadline: 07/06/21 23:59.

• Submission in pairs only.

• The submission is via the course website.

• You should submit the markdown file (.Rmd) and its word/pdf version in a .ZIP format named according to next format: <HW#>\_<ID1>\_<ID2>.zip .

**Detailed instruction for the HW:**

The answers to the questions should be submitted as a markdown file. Please make sure that the output of the markdown file as doc or pdf (applying the Knit) option will give us the question number in bold large font. Additionally to the markdown file submit the pdf versions of two plots that are required at Question 2.4.

**Question 1:**

Search in internet and answer the following questions:

1. What is the difference between a review paper and a research paper?
2. What is the definition of journal impact factor?

Open the PubMed site and search for studies containing the word “GWAS”.

Based on your search results, choose two papers –

1. A review paper on GWAS.
2. A research paper in which a GWAS meta-analysis is performed.
3. Cite the papers – write down the name of the paper, the authors’ names, the name of the journal in which it was published and, the publication date. You can easily approach the paper citations via the page with the search results.

Enter the full-length text in the journal in which the paper was published. Choose the papers that can be accessed without a fee. Please pay attention that using the Technion’s network you’ll be able to access much more papers free of charge.

1. For each of the following papers answer:
   1. What is the value of the impact factor of the journal in which the paper was published at?
   2. How many authors’ affiliations there are?
2. Only for the research paper answer:
   1. What trait was studied in the paper?
   2. If the authors used data from databases - how many databases, the data used in the study was retrieved from? Name the databases. If the authors didn’t use any databases where the data was retrieved from?
   3. Did the authors get to a conclusion that single SNP/set of SNPs **cause** the trait that has been studied in the GWAS? If yes – bring here the citation of the paragraph that mentions this. If not – which words were used to describe the possible effect of the SNP/group of SNPs on the trait.
   4. Use the sample size that was tested in the research paper you’ve found and calculate the minimal effect size that can be found in GWAS with this number of samples at power=0.9. You can use the code for power calculations that was given at the lecture.

**Question 2:**

Download the ANGST (Anxiety Neuro Genetics Study) GWAS summary statistics results from <https://www.med.unc.edu/pgc/download-results> .

1. What is the name of the paper these results were taken from?
2. Create a Manhattan plot relying on the data from the anxiety.meta.full.cc.tbl file.

To create the plot you should read the file using the `fread` function from the data.table package. This function allows you to read large files.

As you probably noticed this file structure is different from the gData object structure that was used to create the aforementioned plot in tutorial 8.

To create the Manhattan plot from this data, without creating the gData object, you should use the `fastman` package.

The installation commands are:

install.packages("remotes")

remotes::install\_github("danioreo/fastman")

The manual for the function that you’ll use to generate the Manhattan plot is:

<https://rdrr.io/github/danioreo/fastman/man/> .

Please pay attention that building the Manhattan plot using this package requires you to provide the reference genome the genomic locations were based on. This data can be found in the research paper in the “Materials and Methods” section.

In case it’s NCBI build 37 or UCSC hg 19 you should set the `chr\_build` parameter as "GRCh37". In case it’s NCBI build 38 or UCSC hg38 you should choose the chr\_build= "GRCh38".

In your Manhattan plot remove the suggestive threshold line and leave only the genome-wide significant line.

* 1. Where are the most significant SNPs are located?
  2. Can you assume that they are in LD? Why?

2.3. Choose one of the most significant SNPs and use the dbSNP site to search on which gene this SNP it is located. Write down the name of the SNP and the name of the gene. Is this gene protein-coding or non-protein coding?

\*Hint – to create the Manhattan plot faster you can remove all rows with P <= 0.1 using one of the `fastman` parameters.

1. Create the qqplot using the appropriate function from the `fastman` package. Based on the qqplot, would you continue with the analysis or go back to perform additional QC ?
2. Export the qqplot and Manhattan plot as PDF in size of 10x10 inches from the “Plots” section in the RStudio and submit them together with the rest of the files in this HW.