CB2-101 2023 Assignment

Malay (mbasu@kumc.edu)

How to submit you answers

You should fork the assignment directory on github. You should clone the forked repository to your local computer, using standard `git` commands.

Once cloned, you should create separate directory for each problem in this assignment. Once you have completed the assignments don't forget to `add`, `commit` and `push` to the repository. After you pushed your answer to github, you will generate a "pull request".

You many modify the answers as many times as you want before the deadline. But please don't forget to `add`, `commit`, and `push` to the repository each time. Also, don't forget to generate a pull request each time.

Unless otherwise mentioned, submit your answer as single `Rmd` or Jupyter notebook file with runnable embedded code. You should also submit the generated `html` or `PDF` along with the source `.Rmd` or Jupyter notebook. Scripts files should be submitted as `.R` or `.sh` files.

Note:

1. You'll be judged by the intent and your sincerity, but not on correctness of the answer. As the course master, my decision about you sincerity of attempt is final.
2. You must answer all the problems in this assignment. If you leave one out, you fail the assignment.
3. If you feel that you want more discussions on specific topic then contact me. I will happy to go through a refresher.

All the best and happy computing!

1 Problem

Write a `bash` script that takes a fasta file as an input and print out the average protein length in the file.

Note: You can use *E. coli* MG1655 proteome file to test your script. The file can be downloaded from here: https://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Bacteria/Escherichia_coli_K_12_substr__MG1655_uid57779/NC_000913.faa.

Note:

1. You must use only `bash` commands. No other programming language is allowed.
2. You may need the following commands in `bash` to complete this task. `wget, zcat, wc, tr, bc, and grep`. You are not restricted to any of these commands. You can use any or all or any other bash commands in your script or command line.

2 Problem

Swissvar is a database of human gene, their variations, and disease associations. The file can be downloaded from here: https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/variants/humsavar.txt. The first column of this file contains the gene name and the rest of the columns contains the other information. Using this file

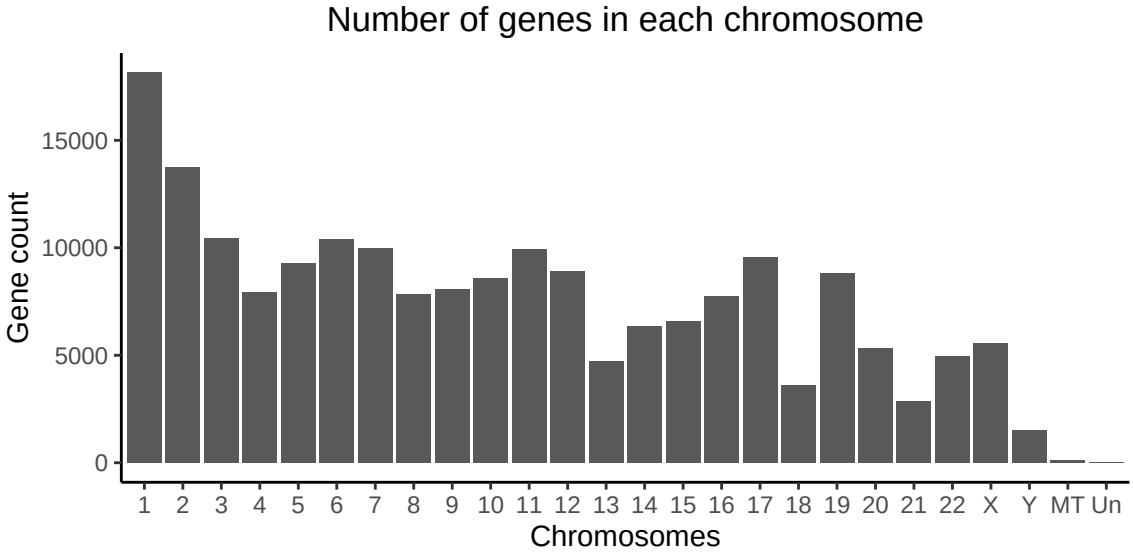
1. list out the top five genes that are mutated in various human disease.
2. plot the frequency distribution of disease variants in human genome across all the genes in the file.
3. calculate the average number disease causing mutations across all genes in human genome and mark this number on the previous plot as veritcal red line.
4. The 4th column of this file contains the amino acid affected by the the mutation like this: `p.Gly477Arg`. The `p` indicates it is protein sequence. Then the 3 letter code of the aa affected then the position in number and then three letter code for the aa that the position changed to. You should write a regular expression to extract the affected aa. Plot a graph showing the fraction of mutations affecting each 20 amino acid on the x-axis. Which amino acid has the highest probablity of getting mutated?

Hint: Remember to skip the information lines in the file and also note that `type of variant` column contains both disease causing and non-disease causing variants.

Note: Try to parse this file yourself. If you cannot do it, run the script "create_data_file.R" to create the data file `humsavar.tsv.gz` in data directory. A ready made file for use is also present in the data directory. Read this file using the standard R way.

3 Problem

1. Use `R` and `ggplot2` package to draw a a plot of number of genes per chromosome in human genomes. This task requires the data file `Homo_sapiens.gene.info.gz`. You need to use columns 3 and 7 indicating `Symbol` and `chromosome` respectively. You script should create a plot exactly as shown below. Save the plot to PDF file.



2. The longer chromosome might have higher frequency of genes most probably by chance. We will test this hypothesis. You can find the length information of each chromosome here: https://www.ncbi.nlm .nih.gov/grc/human/data. Is there any correlation between number of genes and chromosome length? Also, plot the regression data with the trendline. Calculate the R^2 and other statistics to determine whether the fit is significant. From the regression equation estimate the number of genes expected for each chromosome. Then evaluate whether any chromosome has higher or lower concentration of genes. **Hint**: Ignore `MT` and `UN`. You need to find 95% confidence interval of the regression. Read more about it here: https://rpubs.com/aaronsc32/regression-confidence-prediction-intervals

3. In earlier problem we calculated the frequency of disease variant in each gene in human genome. Can you evaluate whether any human chromosome has more concentration of disease variants?

Data: `Homo_sapiens.gene_info.gz` . This is a tab-delimited text file that contains information about

all the genes in the human genome. If you are interested in more about this file format check here: https://ftp.ncbi.nih.gov/gene/DATA/README.

Note:: The figure should exactly look like the above figure. There are some data in the `chromosome` column that are ambiguous and looks like this: `10|19|3`. You need to discard all row where the chromosome value contains a `|`.

4 Problem

Use Fermi estimation (Lecture 1) to estimate a quantity starting from very little knowledge. The more creative you are in creating the problem, the more kudos you will get. Describe the question you are trying to answer and how did you derive the answer.
