INFO 510: Final problems

Malay (malay@uab.edu)

1 How to submit you answers

Unless otherwise mentioned, submit your answer as single `Rmd` file with runnable embedded code. You should also submit the generated `html` or `PDF` along with the source `.Rmd`. Scripts files should be submitted as `.R` files. Please email your answers to `cb2edu@gmail.com` with CC to `malay@uab.edu`. Write your full name and the string "Info 510 solutions to final problems" in the subject line. Do not send word or any other document types.

2 Problem 1

Swissvar is a database of human gene, their variations, and disease associations. The file can be downloaded from here: ftp://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 (A) list out the top five genes that are mutated in various human disease. (B) plot the frequency distribution of disease variants in human genome across all the genes in the file (C) 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.

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` in your current directory. Read this file using the starndard R way.

3 Problem 2

(A) Write two R scripts one for upper quantile normalization and other for DSEQ normalization of a count table file. Assume the first col to be gene name. The pograms should accept the count table name (use R `commandArgs()` function to read the file name) and should generate the normalized output file.

(B) Use one of the example count table files (you can use this file: https://github.com/cb2edu/info510/raw/2019/04-RNASeq/data/pnas_expression.txt) and generate one 2x2 multiplot figure. Each figure should show the boxplots for each sample for all the genes. 4 figures should represnt before and after upper quantile and DESEQ normalization. Label the figures properly and generate the figure as publication ready image.
