

# Biomedical Data Science - Final Project (Spring 2023)

## Analysis of Carl Zimmer's Personal Genome

Key Due Dates:

3-minute recorded presentations: April 26<sup>th</sup>, 2023, Wed, in class

15-minute presentations in discussion: April 27<sup>th</sup> and 28<sup>th</sup>, 2023 in discussion section

Written report and other deliverables: May 9<sup>th</sup>, Tues, submission should be done in Canvas

### Group Assignment

- Students will work in teams on topics of interest. Team composition will be balanced by students enrolled in non-programming and programming modules. The teams have been assigned already. Email us ASAP if there are issues with your team.
- We encourage team members to work together in a collaborative environment on both the analysis and written parts of the project. If any student feels their voice was not heard while working on the project, please reach out to the TAs as soon as possible. At the end of the submitted write-up, please include each team member's contribution.

### Documents and Deliverables

- Each team is required to submit to do **SIX documents** as well as any supplementary information, all together in **one zipped folder**:
  - The first document is a write-up of the investigation with five sections: Introduction, Methods, Results, Discussion, and Team Members' Contributions. The text portion of the write-up should be at least **1000 words** in length and should provide a background on the topic the team investigated, a description of the approaches taken and a discussion of the results with suggestions for potential future work. References and individual team member contributions don't count toward the word requirement. This document must be in PDF format.
  - The second document includes the slides of the presentation students will be delivering on their results. There should be **under 10 slides**. The **last slide** should be a **one-slide summary** of all your methods and results, which will be shared with future classes. Please follow the template at **\*\*\***.
  - The third document is a single text file with 5 elements. Each element should be on a new line, with a total of 5 lines in the text file.
    - Line 1: Project Description
    - Line 2: Section # and Group # (i.e., section 1, group 1)
    - Line 3: Chromosome (i.e., chrX)
    - Line 4: Top 10 Key Genes separated by commas (i.e. GENE1, GENE2, ...)

- Line 5: Key Variants if Applicable (separated by commas) (i.e. rs283843, rs2383828 OR chr17:3832322:C:A, chr17:28383839:T:G; etc.)
- The fourth document is a VCF file that includes a subset of the variants the team identifies in selected genes of interest. Please see the description of Part 1 later in this document for more details. The VCF should be annotated to include which genes the SNPs intersect.
- The fifth document should be all the codes for the project. This could be a single bash/python/R file, or a folder including all the codes.
- Finally, the sixth document is a 1 slide summary of your project. Please use the exact layout and template found here: [One-slide Summaries](#)
- All documents should be submitted on CANVAS by 11:59 PM, May 9th 2023. Only one member of each team should make a submission.

## Presentation

- Final presentations will take place in class on April 26th 2023, Wednesday, 1:00 PM in BASS 305. Each group will create a 3 minute recorded presentation to be played in class (i.e. mp4 file). We will make a Canvas submission specific for this 3 minute recording so you can upload it before class on April 26. We will be strict on time, so please make sure your video is 3 minutes or less. Please also email the slides for the 3 minute presentation to [cbb752@gersteinlab.org](mailto:cbb752@gersteinlab.org) before class on April 26 as well.
- In addition to the 3 minute recorded presentation, each group is responsible for a 15 minute presentation to be presented in discussion section on April 27 or 28. Here, it is expected that everyone presents and participates.
- Carl Zimmer will join us on the presentation day and we will openly discuss interesting results your team finds in his personal genome. We anticipate this would be an interesting experience for all of us.
- Note that the presentation is earlier than the writing due. You should prioritize getting the results and focus on the writing later.

## Grading

- Final grades will be based on the content and clarity of written summary, both presentations, analysis, and any submitted code.
- Generally, group members will share the same grades.

## Analysis Topics

Each team will be assigned one chromosome to work on.

Carl's germline SNPs are found [here](#) under [Germline SNP call set for subjectZ](#). Coordinates are based on the GRCh37 version of the human genome. The file is in VCF format. For more information about VCF, please see [here](#).

### Part 1: Gene Prioritization

Given the germline variant call (VCF), find 10 genes on the chromosome you are assigned with the highest mutational burden (i.e., number of mutations). List the genes and submit records of the variants you identified in the prioritized genes in a file called *gene\_variants\_chr{i}.vcf*, where *i* is the number of the chromosome your team is assigned. In your report, describe the steps you take to identify the variants in the genes of interest. Make sure to mention any database or software tool you use. If you write your own code, please make sure to include it in the final submission.

**[Extra credit]** Suggest an alternative approach (besides using the number of point mutations in each gene) to prioritize 10 genes. These can include methods that rely on genomic mutations (finding genes with more pathologically relevant mutations) or other information (scoring genes using information other than variant counts). Please submit preliminary results of your alternative approach in a supplementary PDF should you decide to work on the extra credit section.

### Part 2: In-Depth Analysis of 10 Genes

Now that you selected 10 genes from Part 1, each team will choose one of the following areas and perform in-depth analysis on the prioritized genes.

1. Gene expression analysis. Find the expression profiles of prioritized genes using data from Genotype-Tissue Expression (GTEx) data (<https://gtexportal.org/home/>). Compare gene expression profile across available tissues. How do expression profiles of the prioritized genes vary across tissues? Broadly speaking, what might differences in expression levels of the same gene across tissues suggest? Provide two or more references to support your arguments.
2. Network analysis. Either: (1) Find protein-protein interaction network(s) involving one or more of the genes you prioritized in Part 1 (example: using "Multiple proteins" option in STRING database) or (2) Find relevant pathway(s) affected by the prioritized genes (examples: use KEGG, Reactome, MSigDB, etc. as reference databases). Provide a figure or more of the network(s) you selected and justify your choice. Explain the interactions or processes taking place in the network(s) you select. What can the network(s) tell us about the functions of the prioritized genes? How might variants in these genes affect resulting protein functions?
3. Protein structure analysis. The Protein Data Bank (PDB) includes structure files of a large compendium of proteins. PDB is an evolving database, and some protein structures might not be included yet. Find the available structure files (in PDB format) of the products (i.e. proteins) of the genes you prioritized and visualize them using PyMOL or another tool of your choice. On the resulting figures, highlight the amino acid(s) affected by the SNP/SNVs you identified in Carl's genome if any of the variants lie in exonic regions.

What are the functions of these proteins? Which areas of the protein structure are affected (i.e. loop, binding pocket, alpha/beta-sheet regions)? Broadly speaking, what are the possible implications of SNP/SNVs on protein structure?

4. Text mining analysis. Perform text mining analysis on publications relating to the prioritized genes. Please use at least 20 publications returned by PubMed when searching for your genes of interest and include their PMIDs in the submission. What are the most frequent biological terms in these publications? Can you find correlations between specific terms occurring in the same paragraphs throughout the combined texts? Is there any possible implication for disease? What does this literature survey tell us about the prioritized genes? Compare your findings with the description of the gene product (i.e. protein) functions you can find in UniProt or GeneCards, which are examples of comprehensive protein annotation databases.

If you have any question regarding the final project, please contact TFs at [cbb752@gersteinlab.org](mailto:cbb752@gersteinlab.org).