Biomedical Data Science - Final Project

Analysis of Carl Zimmer's Personal Genome

Presentations: 4/23/2025, 1:00 PM, BASS 305
Writeup Deadline: 4/30/2025, 11:59 PM

Overview

Group Assignment

- Students will work in teams and work on one of the topics of interest. A team will be made up of students enrolled in both non-programming and programming modules.

- 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.

Submission

- Each team is required to submit **three documents** as well as any supplementary information, all together in one zipped folder.
    - The first document is a final write-up of the investigation with four sections: Introduction, Methods, Results, and Discussion. The text portion of the write-up should be at least 1500 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. This document must be in PDF format.
    - The second document includes the slides of the presentation students will be delivering on their results. This document must be in PDF format (~15 slides).
    - The third 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 submission deadline is April 30, 11:59 PM (CANVAS)**.

- Please make sure to submit any supplementary files (variant file(s), code, or any other documents) as well. **Only one member of each team should make a submission**.

- Submitted final projects will be published on the class website. It will also serve as a reference for you and later students and researchers. If you have any issues or concerns regarding publishing material, please feel free to let us know.

Presentation

- Final Presentation Details:
- Date and Location : April 23, 2025, at 1:00pm at BASS 305
- Format: 3-minute recorded presentation + 2-minute Q&A
- Submission deadline: **April 22, noon**
- Required: ~4 slides + one summary slide + MPEG file

- We will invite Carl Zimmer on the presentation day and will openly discuss interesting results your team finds in his personal genome. We anticipate this would be an interesting experience for all of us.

Grading and Attendance Policies

- Final Presentation and Discussion Session Attendance Policies:
- **All team members must attend and present.**
- **Absence on April 23rd or assigned discussion sessions: Requires a Dean's excuse for undergraduates or a comparable reason for graduate students in advance** (see a nice resource for a list of DE level reasons to receive the team grade).
- **No Dean's excuse & no-show: Penalty** grade.
- **Same rules apply to discussion session presentations.**

- Final grades will be based on the content and clarity of written summary, presentation, analysis, and any submitted code.
- All team members will receive the same grade unless a contribution complaint is made. If a complaint is valid, the responsible member may receive a penalty.

Analysis Topics

- Each team will analyze **chromosome 22** and compare it to **autosomes (1-22)** or using previous analyses on the website or both.
- Ensure a **comparison summary** is included on the final slide.
- 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. Example variant file name is *gene_variants_chr2.vcf*. 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.

- **Gene Expression Analysis:** Find the expression profiles of prioritized genes using the Genotype-Tissue Expression (GTEx) database (https://gtexportal.org/home/). Compare gene expression profiles across available tissues. Evaluate how expression profiles of prioritized genes vary across tissues. Broadly discuss what these tissue-specific expression differences suggest regarding gene function, regulatory mechanisms, or potential tissue-specific roles in health and disease. Support your discussion with two or more relevant references.

- **Network Analysis:** Perform either of the following:

    - Protein-Protein Interaction Networks: Utilize databases such as STRING to identify interaction networks involving your prioritized genes. Generate and include visualizations of the networks. Explain the interactions depicted and their biological implications.

    - Pathway Analysis: Explore databases like KEGG, Reactome, or MSigDB to identify pathways significantly affected by your prioritized genes. Visualize and interpret these pathways.

    - Explain how these networks or pathways inform the biological functions of prioritized genes and discuss how genetic variants (SNP/SNVs) may influence protein function, network interactions, or pathways.

- **Protein Structure Analysis:** Identify available protein structures from the Protein Data Bank (PDB) for the gene products (proteins) of your prioritized genes. For proteins lacking experimentally-determined structures, use AlphaFold to predict the 3D structures. Visualize these structures using PyMOL or another suitable tool. Clearly highlight amino acids affected by SNP/SNVs identified in Carl's genome, especially if they reside within exonic regions. Discuss the functions of these proteins and precisely identify affected structural regions (e.g., loops, binding pockets, alpha-helices, beta-sheets). Evaluate how these variants may impact protein structure and function.

- **Text Mining Analysis:** Conduct an extensive literature review and text mining analysis using at least 20 publications retrieved from PubMed on your prioritized genes (include

PMIDs). Leverage a Large Language Model (LLM), such as GPT-4 or similar, to systematically extract frequent biological terms, key findings, and gene-disease associations. Identify correlations between specific terms within and across publications. Discuss the implications for disease based on these correlations. Compare your LLM-assisted findings with protein function annotations from comprehensive databases such as UniProt or GeneCards. Highlight any consistencies or discrepancies identified in this comparative analysis and discuss their significance.

If you need any clarification on the final project, please do not hesitate to email TAs at cbb752@gersteinlab.org.