Biomedical Data Science - Final Project

Analysis of Carl Zimmer's Personal Genome

Presentations: 4/22/2019, Monday, 1:00 PM, BASS 305
Write-up Deadline: 5/3/2019, Friday, 11:59 PM

Overview

Group Assignment

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

- If you would like to be in the same team with one of your classmates, please inform us after the discussion section on Friday 04/05, and we will do our best to accommodate your preferences.

- 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 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.
    - 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.

- **All documents should be submitted on CANVAS by May 3rd, 2019 at 11:59 PM**. Please make sure to submit any supplementary files (variant file(s), code or any other 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 presentations will take place on 4/22/2019, Monday, 1:00 PM at BASS 305**.

- Each presentation will be for **10 minutes followed by 5 minutes of Q&A**.

- 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

- Final grades will be based on the content and clarity of written summary, presentation, analysis, and any submitted code.

- Each team will get a group grade. Everyone in a team will get the same grade for the final project.

Analysis Topics

Each team will be assigned one chromosome to work on (ie., team1: chr1, team2: chr2, etc.).

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.

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