CBB752 Spring 2017 Final Project

About the Course

- **Title:** Biomedical Data Science: Mining and Modeling
- **Instructor:** Mark Gerstein (http://www.gersteinlab.org)
- **TAs:** Mengting Gu, Paul Muir
- **Introduction:** Bioinformatics encompasses the analysis of gene sequences, macromolecular structures, and functional genomics data on a large scale. It represents a major practical application for modern techniques in data mining and simulation. Specific topics to be covered include sequence alignment, large-scale processing, next-generation sequencing data, comparative genomics, phylogenetics, biological database design, geometric analysis of protein structure, molecular-dynamics simulation, biological networks, normalization of microarray data, mining of functional genomics data sets, and machine-learning approaches to data integration.
- **More Information:** Check out the course website (http://cbb752b17.gersteinlab.org).

Final Project

Due: May 9th 11:59PM

Students will form teams to work on one of the following topics of interest, though we are open to other potential projects if they are clearly articulated and of comparable scope to these listed below. The submitted final projects will be published on this website. It will also serve as a reference for later students and researchers.

Topics

Part 1: Comparative analysis of personal genomes:

-1. Identifying SVs in personal genomes.

-2. Calculate how many SNPs are shared among all genomes and how many are person specific.

-3. Intersect each person's SNPs with HGMD (http://www.hgmd.cf.ac.uk/ac/index.php) and identify which are disease associated. Are the shared or specific SNP sets enriched for disease associated SNPs?

Part 2: Personal genomes and personalized medicine (CRISPR):

-1. Read this paper (http://palgrave.nature.com/nbt/journal/v34/n2/full/nbt.3437.html) and other papers about the design of guide RNAs. Discuss how individual SNPs would impact the off-target effects in the presence of the SNP.

-2. Propose a tool that identifies off target CRISPR sites given a genome and guide RNA sequence.

-3. Propose a tool that finds all PAM sites in the human reference genome as well as a personal genome and compares the similarity of the two sets.

-4. Propose a tool that determines the usefulness of CRISPR-targeted deaminases. How many disease associated SNPs could this technology edit?

-5. Propose a tool that integrates additional genomic information to better predict CRISPR activity (e.g. DNase I hypersensitivity data from ENCODE).

Part 3: Network analysis of personal genomes:

-1. Propose a tool that calculates the degree centrality and betweenness centrality of proteins containing and not containing SNPs in the personal genomes using a PPI file.

-2. One of the personal genomes has an accompanying transcriptomic time course. Identify pathways or gene networks whose expression are altered after HRV or RSV infection.

Part 4: Structure Analysis

-1. Check the protein-coding SNPs in Carl Zimmer's genome and Micheal Synder's genome (http://www.cell.com/abstract/S0092-8674%2812%2900166-3). Look for SNPs that are potentially deleterious (for example, intersecting the SNPs with a database like HGMD). What structural changes might happen due to the mutation that could cause a phenotypic change?