Emily Brahma and Nirmal Krishnan
Dr. Alexis Battle
Computational Genomics: Data Analysis
11 April 2017

Final Project Proposal

Problem:

An estimated 80 percent of men who reach the age of 80 have prostate cancer (PCa) cells. This may sound alarming, but disease progression is generally slow with low mortality. In many men with low-grade prostate cancers (LGPCa), the best treatment is no treatment at all, as the risk of complications from surgery or radiation outweighs the generally minor symptoms of the disease. However, for a small subset of men diagnosed with high-grade prostate cancer (HGPCa), the disease can be extremely worrisome in terms of symptoms, disease progression, and mortality. For these kinds of cancers, the most important part of a good prognosis is an early diagnosis. With an early diagnosis, a procedure like a prostatectomy, in which the full-prostate is removed, generally results in long-term disease free survival. Therefore, it is critical that these cancers are caught early so that patients can take advantage of these surgical option, rather than radiation therapy (which typically has worse complications rates and poorer long-term disease free survival).

In this paper, we are proposing a genome-wide associate study (GWAS) to determine single-nucleotide polymorphisms (SNPs) that are significantly associated with HGPCa. After finding these SNPs significantly associated with HGPCa, we will build a classification engine using those SNPS to predict patients who are significantly at risk of developing HGPCa. What makes our study unique, is that this is the first GWAS study that focuses on predicting risk of severe PCa.

Data:

The data we are using is publicly available through the National Cancer Institute GDC Data Portal. We plan on pulling 500 patients with PCa that have SNP profiles and clinical pathology reports available. The clinical pathology reports are critical for our analysis because they contain the Gleason scores for each patient. A Gleason score is a rating from 2 to 10 based on indicators in prostate tissue samples of the likelihood of disease metastasis. Patients with Gleason scores of 8-10 are considered to have HGPCa.

Approach:

We will conduct a GWAS on every SNP in our dataset by individually testing each SNP for association with HGPCa. After determining which SNPs are significantly associated with the disease using the Bonferroni correction, we will extract these SNPs to build a feature set. Next, we will train a neural network on this feature set to classify patients at risk of developing HGPCa. In order to avoid overfitting, we will be using a cross validation approach, in which we will separate the data into train, validation, and test sets. The training data will be used to build our model, the validation data will be used to fine tune our model's parameters, and the test set will be used to evaluate accuracy.

Tests and Metrics:

We will be using Hinge Loss in evaluating our model, given by:

$$loss = \sum_{1}^{n} max(0, 1 - y \times f(x))$$

In terms of our GWAS, we will be using a standard α of 0.05 (using Bonferroni correction).