

Title: Improving statistical analysis of genome-wide association studies

Research Mentor: Professor Nathan Tintle, Mathematics Department

Project Description:

Note: Prior knowledge of genetics is NOT necessary for this project.

Genome-wide association (GWA) studies are an increasingly popular way to attempt to identify the genetic components of complex human diseases. In short, individual's genotypes (AA, AB, or BB) are measured at thousands of locations across the genome. The distribution of genotypes for people with the disease of interest is compared to the distribution of genotypes for individuals without the disease of interest using standard statistical methods (e.g. chi-squared tests; trend tests; logistic regression). Strong differences in the distributions suggest that the genomic location is associated with the disease under study. GWA studies have successfully identified genes associated with diabetes, Crohn's disease and Rheumatoid Arthritis, among many others.

This summer we will investigate one or both of the following statistical research questions for GWA studies:

1. Traditionally, genotypes have been "called" for each sample presented for analysis. Calling an individual's genotype means identifying individuals as "AA", "AB" or "BB" for each location on the genome. New technology, however, assigns posterior probabilities of genotype assignment (e.g. 92%, 7% and 1% for the AA, AB and BB genotypes, respectively). We will explore the implications of using posterior probabilities instead of called genotypes and explore different statistical methods of using posterior probabilities in analysis.
2. New technology is available to predict an individual's genotype at particular genomic locations, even when those locations are not measured directly. However, errors in these predicted genotypes can increase both the type I and type II error rates in related statistical tests of association. We will explore how genotype errors are created and, subsequently, document how much type I and type II error rates increase as a result.

Students will participate in addressing the research questions using a combination of mathematical proof, computer data simulation and real data analysis.

Background: Students should have some experience with computer programming. Some statistics and/or calculus would also be helpful. Prior knowledge of genetics is not necessary.