Title: Using Imputation as a Method of Improving Genetic Data Analysis
Author: Kimberly Diaz Perez
Faculty Sponsor: Dr. Jessica Turner

Background.
Genome-wide association studies (GWAS) are an important technique for identifying genetic traits, or genotypes, of complex human disorders. Genotypic imputations are now widely used to enhance the accuracy and variation of a GWAS dataset. Non-imputed data can often lack significant genetic information about single nucleotide polymorphisms (SNPs), which are variations in DNA nucleotides in different individuals. Therefore, the genetic dataset is compared to a reference panel, such as 1000 Genomes Project, to increase the imputation accuracy, which maximizes the number of single nucleotide polymorphisms that are observed. We apply imputation techniques to a genome-wide scan dataset to evaluate the improvements.

Methods.
Imputing a dataset involves three distinct stages such as phasing, haploid imputation, and the post-processing of the imputation. Phasing is used to estimate the number of haplotypes in the sample. Haploid imputation uses the haplotypes estimated from phasing to compare it to the reference panel. Lastly, the dataset needs to be processed to ensure the quality of the imputed SNPs is reliable. We obtained genome-wide sequence data from 211 Caucasian, Indian, and Hispanic subjects, including approximately 1,939,155 SNPs from an Illumina Omni-5 beadchip; we phased the data using the software MaCH and imputed the data using MaCH-minimac to associate it with the reference panel to increase the coverage of SNPs. Finally, the imputed data was processed to eliminate low-quality SNPs, such as SNPs with a low-confidence imputation, low minor allele frequency or high missing genotype rate. We also tested confidence values of 0.9 to 0.3 to determine the effect on the final imputed genome-wide sample.

Results.
Following all steps, the number of SNPs increased from 1.9 million to 6.3 million SNPs with high confidence (0.9). The number of SNPs in the imputed dataset increased dramatically with lower confidence in the imputations.

Conclusion.
In summary, the results demonstrate the importance of genotypic imputations in data analysis due to the expansion and association of our data with other GWAS genotyping platforms previously employed. Decreasing the confidence in the imputations leads to more SNPs for analysis but possibly increases the noise in the sample.

Keywords: Imputation, Single Nucleotide Polymorphisms, Human Genetics, Genotype, Bioinformatics, Genome-Wide Association Studies