Population Genetic Analysis with PCA Informative Markers

Jamey Lewis

An Abstract of a Thesis submitted to the Graduate
Faculty of Rensselaer Polytechnic Institute
in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy
Major Subject: Computer Science
The original of the complete thesis is on file
in the Rensselaer Polytechnic Institute Library

Examining Committee:

Prof. Petros Drineas, Thesis Advisor

Prof. Christopher Bystroff, Member

Prof. Russell J. Ferland, Member

Dr. Peristera Paschou, Member

Prof. Mohammed J. Zaki, Member

Rensselaer Polytechnic Institute
Troy, New York
August 2010
(For Graduation December 2010)

Abstract

The analysis of large-scale genetic data from thousands of individual humans has revealed the fact that subtle population genetic structure can be detected at levels that were previously unimaginable. Such genetic structure reflects waves of migration and recent gene flow among different human populations. This complex structure can introduce bias in association studies that seek to identify genes affecting susceptibility to certain disorders. Using Principal Components Analysis (PCA), we analyze and reveal stratification within genetic data, and apply an algorithm to select small panels of PCA-informative markers (PCAIMs) that can reproduce said structure. Importantly, we develop a novel method that can remove redundancy from the selected SNP panels, and show that we can effectively remove correlated markers thus significantly decreasing the expensive genotyping costs in large studies.

Conducting simulated association studies, we couple our method with a PCA-based stratification correction tool and demonstrate that a small number of PCAIMs can efficiently remove false correlations with almost no loss in power. We further present an exploration of the extent and resolution to which individual ancestry can be predicted, first in Europe and then around the world. We show that it is possible to predict geographic coordinates of origin within Europe down to a few hundred kilometers from actual individual origin, using information from carefully selected panels of 500 or 1,000 SNPs. Using the Human Genome Diversity Panel as reference (51 populations - 650,000 SNPs), we show that, in most cases here, the number of SNPs needed for ancestry inference can be successfully reduced to less than 650 of the original 650,000 while retaining close to 100% accuracy. Finally, we apply the same methods to non-human data, namely the Bovine HapMap. Performing extensive cross-validation experiments, we demonstrate that 250-500 carefully selected SNPs suffice in order to achieve close to 100% accuracy in the prediction of individual breed of ancestry, when this particular set of 19 breeds is considered.

The methods we describe, in combination with the increasingly more comprehensive databases of human and animal genetic variation, open new horizons in a variety of fields, ranging from the study of human evolution and population history, to medical genetics and forensics. Applied to non-human datasets, they can be used to inform the design of studies of the genetic basis of economically important traits in cattle, as well as breeding programs and efforts to conserve biodiversity. Furthermore, the SNPs that we have identified can provide a reliable solution for the traceability of breed-specific branded products.