Modeling Informatively Missing Genotypes in Haplotype Analysis

by
Nianjun Liu
The University of Alabama at Birmingham
1665 University Blvd
Birmingham, AL 35294, USA
nliu@uab.edu

Abstract

It is common to have missing genotypes in practical genetic studies. The majority of the existing statistical methods, including those on haplotype analysis, assume that genotypes are missing at randomb assumption can induce both false-positive and false-negative evidence of association. To address this limitation in the current methods, we propose a general missing data model to characterize missing data patterns across a set of two or more markers simultaneously. We prove that haplotype frequencies and missing data probabilities are identifiable if and only if there is linkage disequilibrium between these markers under our general missing data model. Simulation studies on the analysis of haplotypes consisting of two markers illustrate that our proposed model can reduce the bias for haplotype frequency estimates due to incorrect assumptions on the missing data mechanism. Finally, we illustrate the utilities of our method through its application to real data.