

# The Effect of Missing Genotypes on Genetic Association Analysis

Nianjun Liu, Ph.D. Candidate  
Yale University

## Abstract:

It is common to have missing genotypes in practical genetic studies. In general, the exact underlying missing data mechanism is unknown to the investigators. Currently, the missing data mechanism is assumed to be missing at random (i.e. different genotypes and different alleles are missing with the same probability) in most statistical treatments, including haplotype frequency estimation and consequently in haplotype association analysis. However, very few studies have examined the magnitude of the effects when this simplifying assumption is violated. In this study, we show, by simulations, that the haplotype frequency estimates can be biased using methods assuming missing at random if genotypes/alleles are not missing at random. Accordingly, haplotype association analysis based on haplotype frequency estimates may be biased as well, inducing both false-positive and false-negative evidence of association. We propose a model to characterize missing data patterns across a set of two or more markers simultaneously. We use simulations of two SNPs (single-nucleotide polymorphisms) to illustrate that our proposed model can reduce the bias caused by incorrectly assuming missing at random, and have reliable estimates. We also prove that haplotype frequencies and missing probabilities are identifiable if and only if there is linkage disequilibrium (LD) between these markers. Finally, we illustrate the utilities of our method through its application to a real data set.

