Fast and Accurate False Positive Control in Genome-wide Association Studies

Yu Zhang
Penn State University
422A Thomas
University Park
yulili2000@gmail.com

Abstract

Genome-wide disease association studies commonly involve simultaneous testing of millions of single nucleotide polymorphisms (SNP). The SNP-based association tests are often highly correlated due to linkage disequilibrium (LD) among the SNPs. Simple Bonferroni corrections for multiple comparisons are often too conservative. Permutation tests, which are often used in practice, are on the other hand too computationally expensive for genome-wide studies. We present a new method to compute genome-wide significance of SNP associations, which is computationally extremely fast and at the same time attains a high accuracy. We demonstrate analytically that the accuracy and the computation time of our method are almost independent of the sample size, the SNP size, and the p-values being adjusted. When applied to genome-wide SNP datasets, we observed highly variable conservativeness of multiple comparisons evaluated from different genomic regions. The conservativeness of p-values is significantly correlated with SNP LD and SNP density. The method can further be used to control false discovery rate, and we discuss a novel approach using SNP-specific thresholds to detect genome-wide significant associations utilizing prior knowledge of the spatial distribution of disease susceptible loci.