Improving population-specific allele frequency estimates by adapting supplemental data: An Empirical Bayes Approach

Hua Tang
Department of Genetics, Stanford University

Abstract

Estimation of the allele frequency at genetic markers is a key ingredient in biological and biomedical research, such as studies of human genetic variation or of the genetic etiology of heritable traits. As more and more genetic data becomes publicly available, investigators face a dilemma: when should data from other studies and population subgroups be pooled with the primary data? Pooling additional samples will generally reduce the variance of the frequency estimates; however, used inappropriately, pooled estimates can be severely biased due to population stratification. Because of this potential bias, most investigators avoid pooling, even for samples with the same ethnic background and residing on the same continent. We propose an empirical Bayes approach for estimating allele frequencies of single nucleotide polymorphisms. This procedure adaptively incorporates genotypes from related samples, so that more similar samples have a greater influence on the estimates. In every example we have considered, our estimator achieves a mean squared error (MSE) that is smaller than either pooling or not, and sometimes substantially improves over both extremes. The bias introduced is small, as is shown by a simulation study that is carefully matched to a real data example. Our method is particularly useful when small groups of individuals are genotyped at a large number of markers, a situation we are likely to encounter in a genome-wide association study.