

Designing and efficiently analyzing two-stage genome-wide association studies

Andrew Skol

PhD Candidate

University of Michigan

Abstract:

Genome-wide association is a promising approach to identify common genetic variants that predispose to human disease. Because of the high cost of genotyping hundreds of thousands of markers on thousands of subjects, genome-wide association studies will often follow a staged design in which a proportion (π_{samples}) of the available samples are genotyped on a large number of markers in stage 1, and a proportion (π_{markers}) of these markers are later followed up by genotyping them on the remaining samples in stage 2. The standard strategy for analyzing such two-stage data is to view stage 2 as a replication study, and to focus on findings that reach statistical significance when stage 2 data are considered alone. I will demonstrate that the alternative strategy of jointly analyzing the data from both stages almost always results in increased power to detect genetic association, despite the need to use more stringent significance levels, even when effect sizes differ between the two stages.

In addition, I will explore how the optimal two-stage genome-wide association design is affected by several factors, including the genome-wide type I error rate and the difference between the per genotype cost in stage 1 and stage 2. Specifically, I will demonstrate how these factors affect the proportion of subjects allocated to each stage, the proportion of markers selected for follow-up in stage 2, and the overall cost of the genome-wide association study. The implications of misspecifying the stage 2 per genotype cost will also be addressed.