

Rare variant analysis in large-scale association and sequencing studies

Eleftheria Zeggini¹

Abstract

Recent advances in whole-genome genotyping technologies, the availability of large, well-defined sample sets, and a better understanding of common human sequence variation, coupled with the development of appropriate quality control and analysis pipelines, have led to the identification of many novel common genetic determinants of complex traits.

However, despite these successes, much of the genetic component of these traits remains unaccounted for. One largely unexplored paradigm which may contribute to this missing heritability is a model of multiple rare causal variants, each of modest effect and residing within the same functional unit, for example, a gene. Joint analysis of rare variants, searching for accumulations of minor alleles in individuals, for a dichotomous or quantitative trait, may thus provide signals of association with complex phenotypes that could not have been identified through traditional association analysis of single nucleotide polymorphisms (SNPs). However, statistical methods to perform such joint analyses of rare variants have not yet been fully developed or evaluated.

We have implemented rare variant analysis methods in user-friendly software and have extended simple approaches to collapsing rare allele tables by incorporating variant-specific quality scores (for example arising from next generation sequencing studies in which different positions have been covered at different depths) and genotype-specific probabilities (for example arising from 1000 genomes project-imputed data). We have carried out simulations to evaluate these methods and find increases in power to detect association under varying allelic architectures and parameters. We have additionally extended rare variant analysis methods to be robust to different directions of effect and to the presence of correlation structure across SNPs within the same locus. We make recommendations for the analysis of rare variants in large-scale association and next generation sequencing studies.

¹ Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1HH, United Kingdom