Title:

Association Methods for Biobank Studies: Scalable Gene-Environment Interaction Tests and Copy Number Variant Association Tests

Abstract:

Biobank comprises rich genetic and non-genetic information of large samples, and facilitates gene-environment interaction (GxE) studies and genetic association studies beyond single nucleotide polymorphisms (SNPs). Here we present two of our recent work related to biobank studies. In the first work, we introduce SEAGLE (i.e., scalable exact algorithm for large-scale set-based G×E test) to permit GxE test on biobank-scale data for the GxE variance component (VC) test, a widely used strategy to boost overall G×E signals from a genomic region (e.g., gene). SEAGLE employs modern matrix computations to calculate the test statistic and p-value of the GxE VC test in a computationally efficient fashion, without imposing additional assumptions or relying on approximations. SEAGLE can accommodate sample sizes in the order of 10^5, is implementable on standard laptops, and does not require specialized computing equipment. The second work focuses on CNV association analysis. Unlike SNPs, CNV association assessment requires special attentions because CNVs (1) can vary in dosage and length; (2) have no natural definition of a “locus” unit due to breakpoint non-alignment; (3) can have heterogeneous etiological effects depending on the regions disrupted. To address these issues, we introduce CONCUR (i.e., copy number profile curve-based association test) that treats CNVs of each individual as curve data over genomic locations and assesses association using a kernel machine framework. CONCUR evaluates CNV-phenotype associations by comparing individuals' copy number profiles across the genomic regions using the proposed "common area under the curve (cAUC) kernel", captures the effects of CNV dosage and length, and accommodates between- and within-position etiological heterogeneity without the need to define artificial CNV loci as required in current kernel methods. We illustrate the discussed work using data analyses on Taiwan Biobank data.