

Score Tests for Genome-wide Association Studies Adjusting for Population Structure and Cryptic Relationships

Astle, W.J.¹, Holmes C.C.² and Balding D.J.¹

¹ Section of Biostatistics, Department of Epidemiology and Public Health, Imperial College London.

² Department of Statistics, University of Oxford.

In the analysis of population association studies, Genomic Control^[1] (GC) adjusts the Armitage test statistic to correct the type I error for the effects of population substructure, but its power is often sub-optimal. We present score statistics which generalise GC to incorporate co-variation of kinship and phenotype, retaining control over type I error while improving power. Our model is similar to that of Yu et al.^[2] who use mixed models, but we derive an explicit statistic that generalizes the Armitage test statistic and is fast to compute. There is also some similarity between our methodology and that employed by the popular EIGENSTRAT^[3] However, while EIGENSTRAT can adjust for population stratification, it is unable to deal with strong cryptic kinship.

The problems of population structure^[4] and cryptic relatedness^[5] are essentially the same: if patterns of shared ancestry differ between cases and controls, whether distant (coancestry) or recent (cryptic relatedness), false positives can arise and power can be diminished. With large numbers of widely-spaced genetic markers, coancestry can now be measured accurately for each pair of individuals via patterns of allele-sharing. Instead of modeling subpopulations, we work instead with a coancestry coefficient for each pair of individuals in the study.

We explain the relationships between our method, GC and EIGENSTRAT. We present simulation studies and real data analyses to illustrate the power advantage of our method in a range of scenarios incorporating both substructure and cryptic relatedness.

References

- [1] Devlin B. and Roeder K., Genomic control for association studies. *Biometrics* 55(4) December 1999.
- [2] Yu J. et al., A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* 38(2) February 2006.
- [3] Price A.L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38(8) (August 2006).
- [4] Clayton D.G. et al. Population structure, differential bias and genomic control in a large-scale case-control association study. *Nature Genetics*, 37(11) November 2005.
- [5] Voight B.J. and Pritchard J.K., Confounding from cryptic relatedness in case-control association studies. *Public Library of Science Genetics*, 1(3) September 2005.