Use of Empirical Kinship Matrices in Whole Genome Case-Control Studies of Disease in Stratified Populations

Daniel O Stram and Cyril Rakovski

Division of Biostatistics
University of Southern California
Email: stram@usc.edu

Concern about the effects of hidden population stratification remains an issue of interest in the design and analysis of studies of the genetics of complex disease traits. Two recent approaches, the mixed model of Yu et al (Nat Genet 2005) and principal components, c.f. Price et al (Nat Genet 2006), both involve the estimation and use of a large scale empirical kinship matrix to correct for the effects of hidden population structure. We note that, under a now standard model, (Balding Nichols, Genetica 1995) for hidden structure, a given marker *S*, with genotype vector $\mathbf{G}_S$, will have a covariance matrix in a stratified population of form Cov($\mathbf{G}_S$)=$2p_S(1-p_S)\boldsymbol{\kappa}$ where p_S is the minor allele frequency of marker *S* in an ancestral population and $\boldsymbol{\kappa}$ is a correlation matrix having off diagonal elements equal to either zero or $2F_j/(1+F_j)$ where F_j is Wright's genetic distance between the *j*th strata and the ancestral population.

Based on this simple observation, we advocate the adoption, for the use in case control studies in more complex stratified populations, of a model for the covariance matrix of any given marker as Cov($\mathbf{G}_S$)=V_S **K** in which **K** is a unknown but constant correlation matrix and V_S is the variance of marker *S* in the stratified population. With this model case control differences in allele frequency due to a major gene effect of marker *S* on case control status can be estimated as $\hat{\beta}=\left(\mathbf{C}'\mathbf{K}^{-1}\mathbf{C}\right)^{-1}\mathbf{C}'\mathbf{K}^{-1}(\mathbf{G}_S-\mu_S)$ where **C** is a vector of zeros and ones denoting case control status and μ_S is the mean of marker *S* in the stratified population. An adjusted Armitage test for the significance of observed case-control differences in allele frequency is then equal to $\hat{\beta}^2/\mathrm{Var}(\hat{\beta})$ with $\mathrm{Var}(\hat{\beta})=V_S\left(\mathbf{C}'\mathbf{K}^{-1}\mathbf{C}\right)^{-1}$. We discuss estimation of **K** using large scale SNP data, and compare the computational requirements of this method to that of either the mixed model or the principal components method. In addition we show by simulation that this adjusted Armitage test has good statistical properties settings in which cases and controls are closely related but where pedigree structure is unknown, and where publicly available marker data (e.g. from CGEMS or other studies releasing large scale genotype data) is used as a convenience control sample for studies in which only disease cases are genotyped. We find that this test compares very favorably with other methods both computationally and in terms of power and type I error rate and thus is practical for whole genome scans.