RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Efficient Generalized Least Squares Method for Mixed Population and Family-based Samples in Genome-wide Association Studies

Jia Li,[1]* James Yang,[2] Albert M. Levin,[1] Courtney G. Montgomery,[3] Indrani Datta,[1] Sheri Trudeau,[1] Indra Adrianto,[3] Paul McKeigue,[4] Michael C. Iannuzzi,[5] and Benjamin A. Rybicki[1]

[1]*Department of Public Health Sciences, Henry Ford Health System, Detroit, Michigan, United States of America;* [2]*School of Nursing, University of Michigan, Ann Arbor, Michigan, United States of America;* [3]*Arthritis and Clinical Immunology Research Program, Oklahoma Medical Research Foundation, Oklahoma City, Oklahoma, United States of America;* [4]*Public Health Sciences Section, University of Edinburgh Medical School, Edinburgh, Scotland;* [5]*Department of Medicine, Upstate Medical University, Syracuse, New York, United States of America*

Received 24 June 2013; Revised 26 March 2014; accepted revised manuscript 26 March 2014.
Published online 20 May 2014 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21811

ABSTRACT: **Genome-wide association studies (GWAS) that draw samples from multiple studies with a mixture of relationship structures are becoming more common. Analytical methods exist for using mixed-sample data, but few methods have been proposed for the analysis of genotype-by-environment (G×E) interactions. Using GWAS data from a study of sarcoidosis susceptibility genes in related and unrelated African Americans, we explored the current analytic options for genotype association testing in studies using both unrelated and family-based designs. We propose a novel method—generalized least squares (GLX)—to estimate both SNP and G×E interaction effects for categorical environmental covariates and compared this method to generalized estimating equations (GEE), logistic regression, the Cochran–Armitage trend test, and the W_{QLS} and M_{QLS} methods. We used simulation to demonstrate that the GLX method reduces type I error under a variety of pedigree structures. We also demonstrate its superior power to detect SNP effects while offering computational advantages and comparable power to detect G×E interactions versus GEE. Using this method, we found two novel SNPs that demonstrate a significant genome-wide interaction with insecticide exposure—rs10499003 and rs7745248, located in the intronic and 3' UTR regions of the *FUT9* gene on chromosome 6q16.1.**
Genet Epidemiol 38:430–438, 2014. © 2014 Wiley Periodicals, Inc.

KEY WORDS: GWAS; G×E; gene-by-environment; generalized least squares; mixed samples; sarcoidosis

Introduction

Most genome-wide association studies (GWAS) of chronic diseases have used case-control samples of unrelated individuals—however, family-based designs can also be useful to test for both linkage and association. "Mixed" samples result from combining data from both case-control and family-based sampling methods, or where family sampling is incomplete. Such samples present analytic challenges due to correlation between individuals. There are three commonly used approaches that model the association between genotypes and disease status for mixed samples: the Efficient Mixed-Model Association eXpedited (EMMAX) [Kang et al., 2010] method; W_{QLS} [Bourgain et al., 2003] and M_{QLS} [Thornton et al., 2007]; and generalized estimating equations (GEE), recommended by Gray-McGuire et al. [2009] as well as Chen and Yang [2010] in their GWAF R package.

The EMMAX approach models phenotypes using a linear-mixed model with fixed and random effects; fixed effects include the candidate single nucleotide polymorphism (SNP) and covariates such as gender and age, while random effects are based on a phenotypic covariance matrix. Linear-mixed models assume the phenotype is a continuous variable. Although Kang et al. [2010] suggest that the same model can be used for dichotomized variables, the resulting estimates of the linear coefficients may not provide meaningful interpretation, as the estimated proportion difference is used less commonly than odds ratios or relative risks for the effect size estimation. The same problem exists for W_{QLS} and M_{QLS}, in which the test statistics are constructed by quadratic forms with the genotype data treated as a linear outcome. In these methods, different choices of symmetric weight matrices in the quadratic form are used to accommodate different population and pedigree structures; in particular, the W_{QLS} statistic [Bourgain et al., 2003]—which uses the Kinship matrix calculated from pedigree data—was proposed for related individuals without additional population structure. Neither EMMAX or W_{QLS}/M_{QLS} can be easily extended to genotype-by-environment (G×E) analyses; specifically, the multiplicative interaction term cannot be directly estimated because both methods treat the categorical data (case-control status or genotype data) as a continuous outcome. Conversely, the GEE model in the GWAF package uses an independent

Supporting Information is available in the online issue at wileyonlinelibrary.com.

*Correspondence to: Jia Li, One Ford Place, 3E, Detroit, MI 48202, USA. E-mail: jiajiaysc@gmail.com

working correlation structure, with each family being a cluster in the robust variance estimate to test association between the phenotype of interest and each SNP. This method offers flexibility by modeling binary outcome with different link functions (e.g., identity or logit link) and the G×E effect can be tested by including an interaction term. However, the use of the independent working correlation and the computation burden for GWAS data makes the GEE approach less efficient.

We propose an extension of the GSK method originally developed by Grizzle et al. [1969], which has been used for categorical data analysis in traditional observational studies. It assumes that the hypotheses of interest can be expressed in terms of an underlying $(S \times R)$ contingency table, with S representing the cross-classification of a limited number of discrete covariates (e.g., case/control status), and R identifying the number of multinomial response profiles (e.g., genotypes). This approach retains flexibility to model marginal proportions, marginal logits, mean scores, and cumulative logits with increased power and computation efficiency versus competing methods.

For the analysis of GWAS data in mixed samples, our approach—the extended generalized least squares (GLX)—extends the GSK approach by incorporating kinship into the covariance matrix, as well as proposed different response functions to estimate additive, dominant, and recessive effects and G×E interaction effects. We outline the proposed approach and detail methods for genotype and G×E testing. We also present simulation results comparing the GLX method with the Cochran–Armitage trend test, ordinary logistic regression, EMMAX, W_{QLS}, M_{QLS}, and GEE (as implemented in GWAF). Finally, the proposed method is applied to GWAS data from a study of sarcoidosis susceptibility genes in African Americans.

Methods

Extended Generalized Least-Squares (GLX)

We start with the notation of GLX under the setting for individual SNP analysis. Let N subjects be categorized into three possible genotype categories ($R = 3$) for a SNP (i.e., AA, Aa, aa). Individuals with similar covariate values are grouped into stratum s, $s = 1, \ldots, S$. Let n_{sr}, $r = 1, 2, 3$ represents number of subjects within stratum s and genotype r, and n_s stands for total count of subjects within strata s.

Following the definitions of Grizzle et al. [1969], let the expected cell probabilities be π_{sr} and the observed cell probabilities be $p_{sr} = n_{sr}/n_{s.}$, $r = 1, 2, 3$. Define $P_s^t = [p_{s1}\ \ p_{s2}\ \ p_{s3}]$ as a vector of observed probabilities within stratum s, and $P^t = [\,P_1' \ldots P_s'\,]$ as the long vector across strata; similarly, define $\pi_s^t = [\pi_{s1}\ \ \pi_{s2}\ \ \pi_{s3}]$ as a vector of expected probabilities in stratum s, and $\pi^t = [\,\pi_1' \ldots \pi_s'\,]$ as the vector across strata. Without loss of generality, assume that a response function (e.g., $F(\pi)$) of the marginal probabilities is linearly related to the covariates X and parameter β (i.e., $F(\pi)_{u\times 1} = X_{u\times v}\beta_{v\times 1}$), where $\boldsymbol{X}$ is a design matrix of rank $v(\leq u)$ and u is associated with the choice of response function as illustrated in the following sections. The covariance matrix of response function F can be estimated using the delta method: $\hat{V}(F) = \dot{H}[\hat{V}(P)]\dot{H}'$, where $\dot{H}$ is dF^t/dP, and $\hat{V}(P)$ is the estimated covariance matrix of observed probabilities. The estimation of $\hat{V}(P)$ is discussed in detail in the following section. Therefore β is consistently estimated by $\hat{\beta} = (X'[\hat{V}(F)]^{-1}X)^{-1}X'[\hat{V}(F)]^{-1}\hat{F}$, using the inverse of $\hat{V}(F)$ as the weight matrix and $\hat{F}$ as the response vector. The covariance matrix of $\hat{\beta}$ is $V(\hat{\beta}) = (X'[\hat{V}(F)]^{-1}X)^{-1}$.

$F(\pi)$ includes a wide range of possible functions; the most commonly used can often be expressed in two families: (i) linear functions $F(\pi) = A \times \pi$; and (ii) log-linear functions $F(\pi) = K \times \log(A \times \pi)$, where A and K are matrices of arbitrary constants that formulate a specific response function. For details, refer to Grizzle et al. [1969]. Examples for GWAS are illustrated in the following sections.

Estimating the Covariance- $\hat{V}(P)$

In the above model, in order to estimate the weight matrix $\hat{V}(F)$, we must estimate $\hat{V}(P)$. In the case of SNP data, P^t can be expressed as $C \times Y$ where $\boldsymbol{C}$ is a block diagonal matrix having $C_s = 1/n_{s.} \times 1_{n_{s.}} \otimes I_3$ on the diagonal, $\boldsymbol{I_n}$ is the identity matrix of size n, 1_n denotes a vector of size n with all entries one and $\otimes$ is the Kronecker product. $Y = 1_{n_{s.}} \otimes Y_i$, in which $Y_i = [\,Y_{i1}\ Y_{i2}\ Y_{i3}\,]$ is the 3×1 vector of indicator variables for genotypes for each subject $\boldsymbol{i}$. $Y_{ig} = 1$ if $Y_i = g$, and $\boldsymbol{Y_{ig}} = \mathbf{0}$ otherwise, $g = 1, 2, 3$. Note that the covariance of Y between a pair of individuals is

$$\Sigma = \mathrm{cov}(Y_i, Y_j) = \begin{bmatrix} f^4k_{0ij} + f^3k_{1ij} + f^2k_{2ij} - f^4 & 2f^3qk_{0ij} + f^2qk_{1ij} - 2f^3q & f^2q^2k_{0ij} - f^2q^2 \\ . & 4f^2q^2\pi_{0ij} + fqk_{1ij} + 2fqk_{2ij} - 4f^2q^2 & 2fq^3k_{0ij} + fq^2k_{1ij} - 2fq^3 \\ . & . & q^4k_{0ij} + q^3k_{1ij} + q^2k_{2ij} - q^4 \end{bmatrix}, \quad (1)$$

where f is the minor allele frequency, $q = 1 - f$ and k_{mij} is the probability that two individuals i and j share m alleles identity by descent (IBD) under a given relationship, $m = 0, 1, 2$. Thus, $\hat{V}(P)$ can be estimated by $C\hat{\Sigma}C'$. The "theoretical relationship IBD" statistics k_{mij} can be inferred using known pedigree structures. When errors of pedigrees exist, the degree of relationship can still be robustly estimated using the genome-wide genotype data, which is known as the "empirical relationship IBD."

In this paper, we adopted the Kinship-based inference for genome-wide association studies (KING) method proposed by Manichaikul et al. [2010] to estimate kinship coefficient and IBD statistics in the real data analysis. The allele frequency $\boldsymbol{f}$ can be estimated by: (1) the sample frequency $\bar{Y}/2$; or (2) the best linear unbiased estimator (BLUE), given by $\hat{f} = (1_n^t\Phi^{-1}1_n)^{-1}1_n^t\Phi^{-1}(\bar{Y}/2)$, where Φ is the kinship matrix and as suggested by McPeek et al. [2004]. However, we have not

found significant differences between the two estimates in the simulations.

Association Testing Between SNP and a Binary Outcome

To test for association between a single SNP and a binary outcome (e.g., case-control), we developed a general framework of the GLX method with several options for response function $F(\pi)$ that provides estimates based on proportions. One is the linear response function $F(\pi_s) = 0 \times \pi_{s1} + 1 \times \pi_{s2} + 2 \times \pi_{s3}$, $s = 1$ (case) or 2 (control). We can use the following design matrix such that $F(\pi) = X\beta$: $X = \left[\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right]$, $\beta = [\beta_1, \beta_2]^t$. $\hat{V}(F)$ is estimated from $C\hat{\Sigma}C^t$ and $\hat{\boldsymbol{\Sigma}}$ is estimated from equation (1). This parameterized model allows for the estimation of genotypic means while accounting for the dependence between subjects. After estimation, the value of $\hat{\beta}_1$ typically represents average effect of risk allele of two groups, and $\hat{\beta}_2$ represents the differentially effect of risk allele between two groups. The association test can be constructed by the Wald test for $\hat{\beta}_2$. The Wald statistic is computed as a ratio of $\hat{\beta}_2$ over its standard error $\sqrt{V(\hat{\beta})}$. We can assess the level of statistical significance using the normal approximation of the Wald statistic. Note that under this model, $\hat{V}(F)$ is reduced to $J\Phi J^t$, where $\boldsymbol{J}$ is a block diagonal matrix having $1/n_{s.} \times 1_{n_s}$ on the diagonal and Φ is the kinship matrix.

Another option is the log-linear response function. Specifically, one can choose the adjacent logit link function for log-additive model, $F(\pi_s)^t = [\log \frac{\pi_{s2}}{\pi_{s1}} \ \log \frac{\pi_{s3}}{\pi_{s2}}]$, $s = 1$ or 2, if the effect was expected to be additive on the log odds scale. In matrix form, $F(\pi)$ is constructed as $F(\pi) = K \times \log(A \times \pi)$ with $K = \left[\begin{smallmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{smallmatrix}\right] \otimes I_2$ and $\boldsymbol{A}$ is an identity matrix with dimension of 6. We note that dominant or recessive effects can similarly be tested for by modifying K appropriately. Recall that $F(\pi) = X\beta$, $\beta = [\beta_1, \beta_2, \beta_3]^t$. In this situation, the design matrix X is chosen as:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

β_1 and β_2 are the intercepts for each response function across groups. The effect of each copy of the second allele is $\hat{\beta}_3$ on the log odds scale, corresponding to an odds ratio of $\exp(\hat{\beta}_3)$. Similarly a Wald statistic $z = \hat{\beta}_3/\sqrt{\text{var}(\hat{\beta}_3)}$ can be constructed for testing the null hypothesis of $\beta_3 = 0$ as described above.

More generally, a test of the hypothesis $H_0 : L\beta = 0$ is produced by conventional methods of weighted multiple regression, where $\boldsymbol{L}$ is a matrix of full rank $\boldsymbol{d}$. Given the model, the test is produced by $\hat{\beta}^t L^t [L(X^t[\hat{V}(F)]^{-1}X)^{-1}L^t]^{-1}L\hat{\beta}$ that is asymptotically a χ^2 distribution with $\boldsymbol{d}$ degrees of freedom under H_0[Grizzle et al., 1969].

The choice between linear and log-linear response functions depends on data assumptions underlying the statistical analysis. Linear response functions using simple proportions assume that the theoretical distance between proportions is equal. Log-linear response functions, on the other hand, stretch out the distance between proportions. Therefore, the results obtained based on log-linear functions interpret differently from the results obtained from the linear model.

Interaction Between SNP and Environmental Factor on a Binary Outcome

The proposed approach can be extended to test the interaction for categorical outcome and environmental factors (G×E). Consider, an environmental risk factor ($\boldsymbol{E}$), a "high-risk genotype" ($\boldsymbol{G}$), and a disease of interest ($\boldsymbol{D}$). In general, statistical gene–environmental interaction is defined as departure from additive or greater-than-multiplicative joint effects of gene/environmental effects. Statistical interactions are scale dependent; choice of measurement scale will affect the assessment of G×E interaction. Ottman [1996] presents a variety of definitions for gene–environment interactions using relative risks and odds ratios under additive and multiplicative scales; for illustrative purposes, we focus on the interaction based on odds ratios.

In order to apply the GLX test for the G×E multiplicative effect, we first conduct the SNP association test within each environmental factor stratum. Let $\hat{\beta}_t$ and $\text{var}(\hat{\beta}_t)$ represent the coefficient estimates and corresponding variance for the association test between G and D within stratum t. The adjacent logit link function, as described above, is used here. The G×E effect then can be tested using Cochran's Q test [Cochran, 1954]. The test statistic consists of a weighted sum of squared deviations around the mean of the effect. Specifically for our G×E interaction testing, $\boldsymbol{Q}$ is defined as $\sum w_t(\hat{\beta}_t - \bar{\beta})^2$, where $\bar{\beta} = \sum w_t\hat{\beta}_t / \sum w_t$ is the weighted mean of the log odds ratio. Here we choose the weight $w_t = 1/\text{var}(\hat{\beta}_t)$, that is, the inverse of the estimated variance of log odds ratio in each strata. Under the null hypothesis of no interaction, $\boldsymbol{Q}$ follows a $\boldsymbol{\chi}^2$distribution with $\boldsymbol{T} - \boldsymbol{1}$ degrees of freedom, where $\boldsymbol{T}$ is the number of strata in $\boldsymbol{E}$.

Simulation Studies

We examined the type I error and power of the proposed estimators in a variety of simulated pedigrees based on three real datasets: the Ancestry Mapping of African genes of Sarcoidosis Susceptibility study (AMASS) [Rybicki et al., 2011]; the multi-ethnic study of atherosclerosis (MESA) [Bild et al., 2002]; and the Framingham Heart Study [Govindaraju et al., 2008].

AMASS study data were compiled from three previously conducted studies: (1) a multisite case-control study; (2) a multisite affected sib-pair study; and (3) a single institution family-based study. Of 2,494 genotyped specimens, 1,877 had both genotype and environmental data. A total of 1,283 specimens were collected from 475 pedigrees with 277 sibships ranging in size from 2 to 6. The remaining 594 specimens were from the case-control study.

MESA was a study of characteristics of subclinical cardiovascular disease. One of its ancillary studies—the MESA

Table 1. Simulation configurations for each data type

Type	Total *N*	*N* families	*N* unrelated
AMASS	1,877	475	594
MESA	3,735	702	0
Framingham	6,870	765	435

Family Study—applied genetic analysis and genotyping methodologies to delineate the genetic determinants of early atherosclerosis. MESA pedigrees were more complex than those in AMASS. Among 3,735 specimens, there were approximately 702 families with family size varying from 3 to 16; however, a majority of the MESA families consisted of only parents and offspring.

The Framingham Heart Study included 6,870 individuals in families of up to three generations, including over 900 pedigrees and 230 singletons. The pedigree sizes varied from 2 to 296, more complex than the AMASS and MESA studies. Summary information for each study sample is provided in Table 1 and ordered by complexity of pedigree structures. We randomly selected 500 mixed families and singletons from the MESA and Framingham studies for the simulation.

Using Merlin [Abecasis et al., 2002] and pedigree structures from the real examples, we simulated 10,000 datasets of genotype "G" data through gene dropping, with a minor allele frequency of 0.2. The binary environmental factor "E" was generated assuming that individuals within the same family have the same environmental exposure—that is, it is more correlated than exposure for individuals in different families. The quantitative traits of all individuals in the pedigrees were generated according to the linear-mixed model described below. The phenotype of individual *j* in family *i* was generated by:

$$Y_{ij} = \beta_0 + \beta_1 G + \beta_2 E + \beta_3 G \times E + r_i + e_{ij},$$

where β_0 is the population mean, β_1 is the additive effect for a SNP, β_2 is the environmental main effect and β_3 is the interaction effect between G and E. To allow for correlation between phenotypes within each family, we assumed a random family effect $\boldsymbol{r_i}$ that follows a multivariate normal distribution $N(0, \sigma_f^2 2\Phi_i)$, where the elements of $2\Phi_i$ denote the kinship coefficient between individuals in family *i*. The residual $\boldsymbol{e_{ij}}$ is normally distributed with mean 0 and variance σ_e^2. We considered $\sigma_f^2 = 3$ and $\sigma_e^2 = 1$, which corresponds to a heritability value of 0.75. Dichotomous traits were then generated from the Bernoulli distribution with probability $\boldsymbol{p_{ij}}$, where $\boldsymbol{p_{ij}}$ is calculated from the inverse logit function, $\text{logit}^{-1}(Y_{ij})$.

We set β_0 to –3, representing a baseline disease risk of approximately 5%, independently of *G* and *E*. Type I error was estimated for testing $H_0 : \beta_1 = 0$ and $H_0 : \beta_3 = 0$, corresponding to no genetic effect and no G×E interaction, respectively. For power analysis, we considered three types of values for β_1, β_2, and β_3. First, we set them to 0.405, 0, and 0, respectively, yielding an OR of approximately 1.5 for the gene main effect. Second, we set β_1, β_2, and β_3 to be 0.405, 0.405, and 0.69, respectively, yielding ORs of approximately 1.5 for the main effect of gene *G*, approximately 1.5 for the main effect of the environmental factor, and approximately 2 for the G×E interaction effect. This model represents the synergistic effect between G and E. Third, we set β_1, β_2, and β_3 to be 0, 0, and 0.69, respectively, yielding G×E interaction effect without marginal effects.

GLX was compared to the Cochran Armitage trend test, GEE with identity link, EMMAX, W_{QLS}, and M_{QLS} (for the linear mean score model), ordinary logistic regression, and GEE (for the log-additive effect model). The kinship coefficient and IBD estimates were calculated from the known pedigree structures. In the simulation, the allele frequency $\boldsymbol{p}$ was estimated by the sample frequency $\bar{Y}/2$, as we did not see a significant difference in the results using the sample frequency or BLUE for allele frequency estimation.

An Example From a Study of Sarcoidosis

To examine the performance of our proposed method in a real dataset, we applied GLX to the AMASS GWAS samples, consisting of 1073 African-American sarcoidosis cases and 804 controls drawn from unrelated case-control and family samples. Genotyping was performed on the Illumina Human Omi1-Quad at the Oklahoma Medical Research Foundation (Oklahoma City, OK) for 1.1 M SNPs across the genome. The details of genotyping and quality control process have been described in Adrianto et al. [2012].

After quality control processes, the final set comprised 887,296 autosomal SNPs. The SNP-based pairwise kinship coefficients and identity-by-descent coefficients were estimated using KING [Manichaikul et al., 2010]. Before applying the kinship estimates to the GLX method, a pair of individuals with kinship coefficient (ϕ) less than $\mathbf{1/2^{9/2}}$ was considered to be unrelated; the corresponding probability of zero IBD-sharing (k_0) was set at 1, and probability of one IBD-sharing (k_1) was set at 0. To test if the individual SNPs had significant effect on the risk of sarcoidosis, we applied the GLX method with both linear response function and adjacent logit response function.

Rossman et al. [2008] found that carrying *HLA-DRB1**1101 and exposure to workplace insecticides was associated with increased risk of developing sarcoidosis ($P < 0.10$), suggesting a potential G×E interaction. To illustrate the proposed G×E test from the GLX method, we focus here on evaluating the G×E interaction effects of SNP and insecticide exposure on sarcoidosis risk.

Simulation Results

Figure 1A and B present the estimated type I error rates and power for a variety of SNP association tests. The GLX method with either linear response or log-additive response functions controls for Type I error rates when the nominal rate of statistical significance is $P = 0.01$. Ordinary logistic regression performs similarly to the trend test, and both are anticonservative under certain scenarios. This inflation of type I error is expected from a method that ignores relatedness within

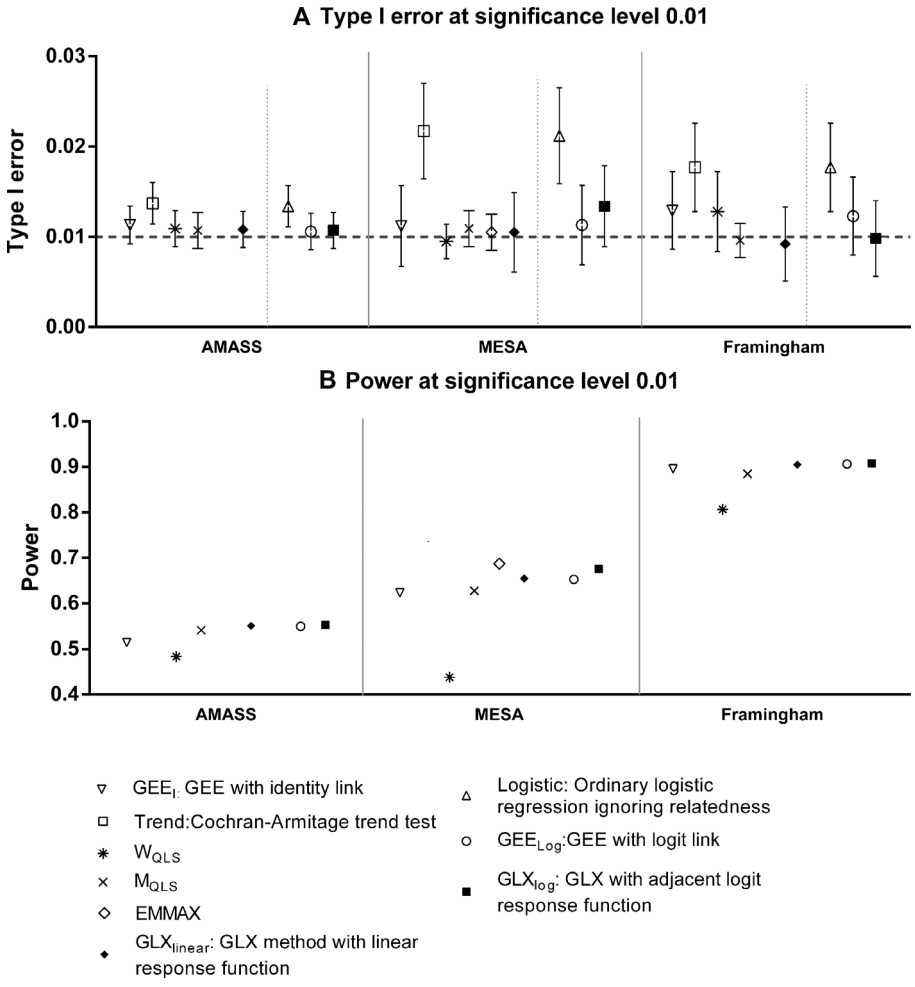

Figure 1. Comparison of different methods for testing SNP association with risk of disease when minor allele frequency is 0.2. Nominal type I error rate was set at 0.01. Point estimates and 95% confidence intervals of type I error and power were presented.

the mixed sample, consistent with the simulation results for the trend test reported by Manichaikul et al. [2012] and Feng et al. [2011]. When the related portion of the total samples is small (as with AMASS), the type I error of a simple logistic regression or trend test may not be greatly affected. But Type I error is more severely inflated when the sample consists of more complex pedigrees—for instance, using the MESA study sample, the type I error rates under logistic regression or the trend test are double those of the nominal level. Due to above observations, power is not reported for ordinary logistic regression and the trend test in Figure 1B. GEE models that use robust variance estimation with an independent working covariance control type I error for all three scenarios. The same conclusion is made for the W_{QLS} and M_{QLS} tests.

However, GLX outperforms them in terms of power. The power of GLX with an adjacent logit response function is 68% ± 1% (MESA) and 91% ± 0.3% (Framingham) for each study sample configuration. The power of M_{QLS} with logit link is 63% ± 1% (MESA) and 88% ± 0.3% (Framingham) sample configurations. We observe larger differences in power between GLX, GEE, and W_{QLS} when using the linear function: GLX has 65% ± 1% power compared to 62% ± 1% (GEE with identity link), 44% ± 1% (W_{QLS}), and 63% ± 1% (M_{QLS}) in the MESA study; similar trends are seen using the AMASS and Framingham study sample configurations. Although GLX methods did not systematically show greater power than EMMAX, both methods are comparable in most of the scenarios.

Figure 2A, B, and C present the results of estimated type I error rates and power for SNP-by-environment (SNP×E) interaction tests for different methods. As we observe for the SNP association testing, type I error is reasonably controlled by both GEE and GLX for all sample configurations (Fig. 2A). Logistic regression again demonstrates inflated type I error. Type I errors rates are around 2% for both MESA and Framingham pedigrees when the nominal significance was set at 1%. For these two sample configurations, GEE and GLX achieve nearly equivalent power (Fig. 2B). Both methods have power of 47% ± 1% (MESA) and 71% ± 0.6% (Framingham) in these study sample configurations. The performance of these three tests under the interaction-only model leads to the same conclusions (Fig. 2C).

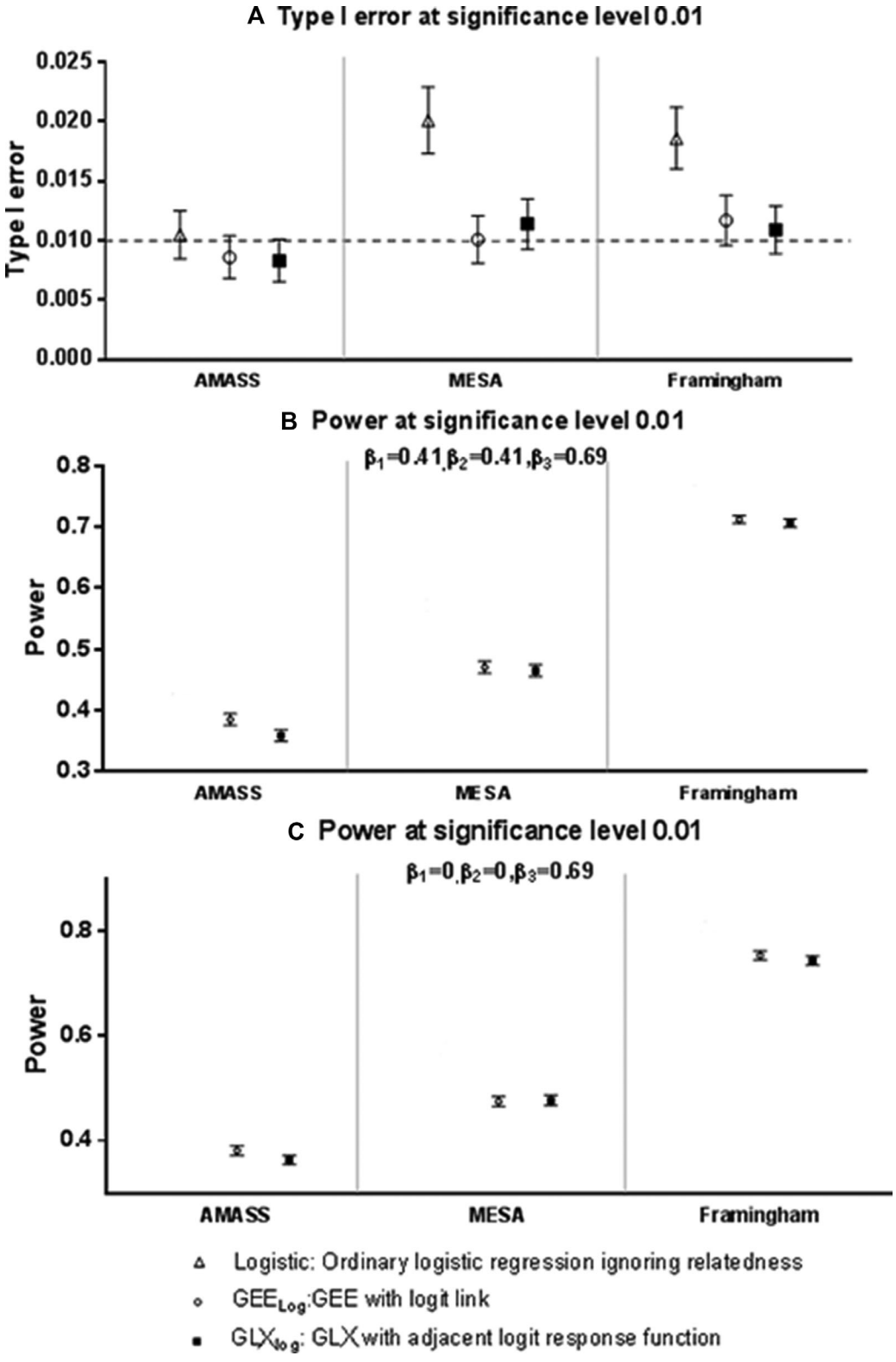

Figure 2. Comparison of different methods for testing SNP-by-environment multiplicative interaction with risk of disease when minor allele frequency is 0.2. Nominal type I error rate was set at 0.01. Point estimates and 95% confidence intervals of type I error and power were presented.

Table 2. AMASS data results: SNPs associated with risk of Sarcoidosis with *P*-values passed genome-wide significance level

CHR	SNP	GLX_{linear} (GI = 1.10)	GLX_{log} (GI = 1.05)	OR	95% CI	EMMAX (GI = 1.01)	MQLS (GI = 1.15)
6	rs6931646	2.1×10^{-7}	2.1×10^{-7}	0.70	(0.61, 0.80)	1.1×10^{-7}	1.1×10^{-6}
6	rs2239804	1.7×10^{-7}	1.8×10^{-7}	0.69	(0.61, 0.79)	9.1×10^{-8}	9.7×10^{-7}
6	rs2239803	1.1×10^{-7}	1.2×10^{-7}	0.69	(0.60, 0.79)	5.4×10^{-8}	6.0×10^{-7}
6	rs6911419	1.3×10^{-7}	1.4×10^{-7}	0.69	(0.61, 0.79)	6.7×10^{-8}	6.7×10^{-7}
6	rs9268658	1.6×10^{-7}	1.7×10^{-7}	0.69	(0.61, 0.79)	8.2×10^{-8}	4.5×10^{-7}

GLX_{linear}, *P*-value for GLX method with linear response function; GLX_{log}, *P*-value for GLX method with log-additive response function; OR, odds ratio; 95% CI, 95% confidence interval for OR; GI, genomic inflation factor.

Table 3. AMASS data results: SNP by insecticide exposure interaction with *P*-values < 1E^{-}5

SNP	CHR	Position	Alleles	MAF	Gene	Insecticide exposure OR (95% CI)	No insecticide exposure OR (95% CI)	*P*-value (GI = 0.94)
rs6720972	2	98039926	G/A	0.07		0.46 (0.43, 0.50)	1.60 (1.48, 1.73)	5.35×10^{-6}
rs13417566	2	111309077	A/C	0.35	ACOXL	0.76 (0.75, 0.78)	1.45 (1.42, 1.47)	7.85×10^{-6}
rs10499003*	6	96611940	C/A	0.07	FUT9	2.06 (1.88, 2.25)	0.40 (0.37, 0.43)	1.47×10^{-8}
rs7745248	6	96766835	A/G	0.08	FUT9	1.92 (1.76, 2.09)	0.38 (0.35, 0.41)	1.82×10^{-8}
rs11156352	6	96794243	G/A	0.09		1.66 (1.55, 1.78)	0.50 (0.47, 0.53)	2.69×10^{-6}
rs2341786	6	159848314	A/G	0.10		1.75 (1.65, 1.86)	0.55 (0.52, 0.58)	1.56×10^{-6}

P-value: *P*-values for gene-by-insecticide-exposure interaction were obtained from GLX method with log-additive response function; OR, odds ratio; 95% CI, 95% confidence interval for OR; GI, genomic inflation factor.

Exemplar Data Set

As a complement to the simulation studies, we analyzed GWAS data from the AMASS study. The single-marker association analysis using the GLX method with linear and logit response function showed modest genomic inflation factors (1.10 and 1.05). We further corrected the *P*-values based on the genomic inflation factors for each method. The six most significant SNPs are in the Major Histocompatibility Complex (chromosome 6p21); method-specific results are displayed in Table 2. The EMMAX analysis identified one SNP (rs2239803) approaching the genome-wide significance level ($P < 5 \times 10^{-8}$, using Bonferroni correction). As expected, M_{QLS} failed to control for population stratification. Odds ratios and 95% confidence intervals for the six SNPs were estimated using the GLX method with log-additive response function. The *P*-values for testing the odd ratios are slightly higher than the *P*-values using the linear response function. Table 3 lists the six SNPs that reach a suggestive significance level (*P*-values < 1×10^{-5}) for a SNP-by-insecticide-exposure interaction. Two of these passed the genome-wide significance level: rs10499003 and rs7745248, both located in the *FUT9* gene on chromosome 6. The other two SNPs from chromosome 6 are neighboring SNPs in high linkage disequilibrium (LD) with rs10499003 ($r^2 > 0.6$). The estimated odds ratio of sarcoidosis risk for each additional copy of allele C at rs10499003 is 2.06 (95% confidence interval (CI): 1.88–2.25) if exposed to insecticide. The risk decreases (OR = 0.40 (95% CI: 0.37–0.43) for those who were not exposed to insecticide.

Discussion

Large association studies that collect both unrelated case-control and family data have been conducted in several diseases [Bild et al., 2002; Edenberg et al., 2005; Govindaraju et al., 2008; Rybicki et al., 2011]. Motivated by a GWAS of sarcoidosis susceptibility genes in African Americans, we proposed an extended Generalized Least Squares approach to test for the association between disease and genes, as well as gene-by-environment interactions, when data comes from mixed samples of family-based and population studies. This method allows us to account for correlations among family members by using kinship estimated from GWAS data or known pedigrees. This approach is flexible in the sense that many response functions under varying model assumptions can be used, and corresponding effect size and confidence intervals can be estimated.

Insecticide exposure has been previously shown to be associated with sarcoidosis on both marginal and G×E interaction levels [Newman et al., 2004; Rossman et al., 2008]. Using the GLX method, we found significant interactions between insecticide exposure and the *FUT9* SNPs rs10499003 (intronic) and rs7745248 (3′UTR). The protein encoded by this gene belongs to the glycosyltransferase family [Brito et al., 2008]; recent human cell-line work suggests that the FUT9 protein has an important role in the biosynthesis of human E-selectin ligands [Buffone et al., 2013]. Sarcoidosis patients are known to have elevated levels of circulating E-selectin in peripheral blood [Berlin et al., 1998; Hamblin et al., 1994]. Furthermore, recent work in mice has shown that E-selectin knockouts have a more severe grade of granuloma

formation in lungs upon exposure to *P. acnes* versus wild-type mice [Kamata et al., 2013]; in humans, a common E-selectin polymorphism is associated with significantly reduced risk of developing sarcoidosis in patients with erythema nodosum [Amoli et al., 2004]. Although it is too preliminary to implicate *FUT9* in sarcoidosis pathogenesis, this finding demonstrates the potential utility of the GLX method to uncover provocative G×E interactions worthy of further study.

GLX has also shown advantages over other methods in simulations. Traditional methods that assume independent sampling are anticonservative if applied to mixed sample data that is dominated by complex family structures, such as those found in the MESA and Framingham studies. Another option for mixed sample data is GEE, which provides robust inferences for most of the scenarios in our simulation; however, its computation time is relatively long. As a result, some researchers run the association test using EMMAX or M_{QLS} first, then use a GEE model with logit link function on the regions highlighted by EMMAX or M_{QLS} to obtain the estimates of odds ratios and corresponding standard errors. In contrast, the GLX method is computationally appealing for its simpler, noniterative procedure. To scan 10,000 SNPs for a study of sample size 4,000 using a 2.67 GHz Intel® Xeon® CPU running Linux, it took 15.4 min for GEE and just 1.3 min for GLX versus. In this case, GLX is 12 times faster than GEE.

The current version of GLX was written in the R programming language; GEE in the GWAF package was written in C language and called by R. We are currently developing a version of GLX implemented in C—we expect this to be even more computationally efficient. More importantly, our simulations show that GLX has superior power to GEE under linear and log-additive models for SNP association tests for a number of scenarios. The final advantage to using GLX rather than EMMAX or W_{QLS}/M_{QLS} is that GLX offers many different forms of response functions that can be specified by the user. Unlike EMMAX and W_{QLS}/M_{QLS}, which use linear models on categorical data, the interpretation of the model coefficient of the GLX method may be more meaningful than these two methods and may be easily extended to test for G×E interactions.

There are some limitations to this method. First, the behavior of tests in small samples is unknown. Occasional small cell counts may require adjustment of the data so that the weight matrix is not singular. However, this problem is not unique to our GLX test; in actual practice, SNP genotypes can be collapsed into dominant or recessive coding. Second, the GLX method is primarily developed for categorical data analysis. Continuous covariates may be used by considering them as categorical variables based on their unique values. However, computational difficulties may arise if a continuous covariate has a large number of unique values; in this case, we can still use this method by discretizing the variables. Third, while the proposed GLX method is an efficient method for accurately estimating both main and marginal effects from family-based data, effect estimates may not be generalizable to the general population if ascertainment bias exists. Correcting for ascertainment bias was beyond the scope of this study, but this issue has been extensively addressed by others [Epstein et al., 2002; Noh et al., 2005; Schaid et al., 2010]. And finally, population stratification was not observed in the exemplar dataset, nor was the GLX method assessed in the presence of population heterogeneity with simulated data. However, our method can be extended to account for population stratification by modifying the covariance matrix, similar to the approach used in ROADTRIPS [Thornton and McPeek, 2010].

In summary, we propose a novel generalized least squares (GLX) method to estimate both SNP and G×E interaction effects in mixed samples. Our simulation results demonstrate that this method improves upon existing methods used to analyze these types of data, both in terms of type I error and power under a variety of pedigree structures. Given the computational efficiency of the GLX method and its ability to be easily extended to test for G×E interactions, it should be very attractive for analysis of genome-wide marker datasets of mixed samples.

Acknowledgments

This work was supported by National Institutes of Health grant numbers R56-AI072727, R01-HL092576 (BAR); R01-HL54306, U01-HL060263 (MCI); 1RC2HL101499, R01HL113326 (CGM); P20GM103456 (IA). MESA data were accessed through dbGaP (study accession phs000209). The MESA Family Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support is provided by grants and contracts R01HL071051, R01HL071205, R01HL071250, R01HL071251, R01HL071258, R01HL 071259, UL1-RR-025005; by the National Center for Research Resources, Grant UL1RR033176; and the National Center for Advancing Translational Sciences, Grant UL1TR000124.

The authors have no conflict of interest to declare.

References

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. 2002. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 30(1):97–101.

Adrianto I, Lin CP, Hale JJ, Levin AM, Datta I, Parker R, Adler A, Kelly JA, Kaufman KM, Lessard CJ and others. 2012. Genome-wide association study of African and European Americans implicates multiple shared and ethnic specific loci in sarcoidosis susceptibility. *PLoS One* 7:e43907.

Amoli MM, Llorca J, Gomez-Gigirey A, Garcia-Porrua C, Lueiro M, El-Magadmi M, Fernandez ML, Ollier WE, Gonzalez-Gay MA. 2004. E-selectin polymorphism in erythema nodosum secondary to sarcoidosis. *Clin Exp Rheumatol* 22(2):230–232.

Berlin M, Lundahl J, Sköld CM, Grunewald J, Eklund A. 1998. The lymphocytic alveolitis in sarcoidosis is associated with increased amounts of soluble and cell-bound adhesion molecules in bronchoalveolar lavage fluid and serum. *J Intern Med* 244(4):333–340.

Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux AV, Folsom AR, Greenland P, Jacob DR Jr, Kronmal R, Liu K and others. 2002. Multi-ethnic study of atherosclerosis: objectives and design. *Am J Epidemiol* 156(9):871–881.

Bourgain C, Hoffjan S, Nicolae R, Newman D, Steiner L, Walker K, Reynolds R, Ober C, McPeek MS. 2003. Novel case-control test in a founder population identifies P-selectin as an atopy-susceptibility locus. *Am J Hum Genet* 73(3):612–626.

Brito C, Kandzia S, Graça T, Conradt HS, Costa J. 2008. Human fucosyltransferase IX: specificity towards N-linked glycoproteins and relevance of the cytoplasmic domain in intra-Golgi localization. *Biochimie* 90(9):1279–1290.

Buffone A Jr, Mondal N, Gupta R, McHugh KP, Lau JT, Neelamegham S. 2013. Silencing α1,3-fucosyltransferases in human leukocytes reveals a role for FUT9 enzyme during E-selectin-mediated cell adhesion. *J Biol Chem* 288(3):1620–1633.

Chen MH, Yang Q. 2010. GWAF: an R package for genome-wide association analyses with family data. *Bioinformatics* 26(4):580–581.

Cochran WG. 1954. The combination of estimates from different experiments. *Biometrics* 10:101–129.

Edenberg HJ, Bierut LJ, Boyce P, Cao M, Caulwey S, Chiles R, Doheny KF, Hansen M, Hinriches T, Jones K and others. 2005. Description of the data from the Collabroative Study on the Genetics of Alcoholism (COGA) and single-nucleotide polymorphism genotyping for Genetic Analysis Workshop 14. *BMC Genet* 6(Suppl 1):S2.

Epstein MP, Lin X, Boehnke M. 2002. Ascertainment-adjusted parameter estimates revisited. *Am J Hum Genet* 70(4):886–895.

Feng Zeny, Wong WL, Gao X, Schenkel F. 2011. Generalized genetic association study with samples of related individuals. *Ann Appl Stat* 5(3):2109–2130.

Govindaraju DR, Cupples LA, Kannel WB, O′Donnell CJ, Atwood LD, D′Agostino RB Sr, Fox CS, Larson M, Levy D, Murabito J and others. 2008. Genetics of the Framingham Heart Study population. *Adv Genet* 62:33–65.

Gray-McGuire C, Bochud M, Goodloe R, Elston RC. 2009. Genetic association tests: a method for the joint analysis of family and case-control data. *Hum Genomics* 4(1):2–20.

Grizzle JE, Starmer CF, Koch GG. 1969. Analysis of categorical data by linear models. *Biometrics* 25:489–504.

Hamblin AS, Shakoor Z, Kapahi P, Haskard D. 1994. Circulating adhesion molecules in sarcoidosis. *Clin Exp Immunol* 96(2):335–338.

Kamata M, Tada Y, Mitsui A, Shibata S, Miyagaki T, Asano Y, Sugaya M, Kadono T, Sato S. 2013. ICAM-1 deficiency exacerbates sarcoid-like granulomatosis induced by propionibacterium acnes through impaired IL-10 production by regulatory T cells. *Am J Pathol* 183(6):1731–1739.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* 42(4):348–354.

Katayama Y, Hidalgo A, Chang J, Peired A, Frenette PS. 2005. CD44 is a physiological E-selectin ligand on neutrophils. *J Exp Med* 201(8):1183–1189.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. 2010. Robust relationship inference in genome-wide association studies. *Bioinformatics* 26(22):2867–2873.

Manichaikul A, Chen WM, Williams K, Wong Q, Sale MM, Pankow JS, Tsai MY, Rotter JI, Rich SS, Mychaleckyj JC. 2012. Analysis of family- and population-based samples in cohort genome-wide association studies. *Hum Genet* 131(2):275–287.

McPeek MS, Wu X, Ober C. 2004. Best linear unbiased allele-frequency estimation in complex pedigrees. *Biometrics* 60(2):359–367.

Newman LS, Rose CS, Bresnitz EA, Rossman MD, Barnard J, Frederick M, Terrin ML, Weinberger SE, Moller DR, ACCESS Research Group and others. 2004. A case control etiologic study of sarcoidosis: environmental and occupational risk factors. *Am J Respir Crit Care Med* 170(12):1324–1330.

Noh M, Lee Y, Pawitan Y. 2005. Robust ascertainment-adjusted parameter estimation. *Genet Epidemiol* 29(1):68–75.

Ottman R. 1996. Gene–environment interaction: definitions and study designs. *Prev Med* 25(6)764–770.

Rossman MD, Thompson B, Frederick M, Iannuzzi MC, Rybicki BA, Pander JP, Newman LS, Rose C, Magira E, Monos D and others. 2008. HLA and environmental interactions in sarcoidosis. *Sarcoidosis Vasc. Diffuse Lung Dis* 25(2):125–132.

Rybicki BA, Levin AM, McKeigue P, Datta I, Gray-McGuire C, Colombo M, Reich D, Burke RR, Iannuzzi MC. 2011. A genome-wide admixture scan for ancestry-linked genes predisposing to sarcoidosis in African-Americans. *Genes Immun* 12:67–77.

Schaid DJ, McDonnell SK, Riska SM, Carlson EE, Thibodeau SN. 2010. Estimation of genotype relative risks from pedigree data by retrospective likelihoods. *Genet Epidemiol* 34(4):287–298.

Thornton T, McPeek MS. 2010. ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. *Am J Hum Genet* 86(2):172–184.

Thornton T, McPeek MS. 2007. Case-control association testing with related individuals: a more powerful quasi-likelihood score test. *Am J Hum Genet* 81(2):321–337.