DOI: 10.1111/j.1541-0420.2010.01454.x

Unbiased and Locally Efficient Estimation of Genetic Effect on Quantitative Trait in the Presence of Population Admixture

Yuanjia Wang,[1,*] Qiong Yang,[2] and Daniel Rabinowitz[3]

[1]Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 W168th Street, New York, New York 10032, U.S.A.

[2]Department of Biostatistics, Boston University, Boston, Massachusetts 02118, U.S.A.

[3]Department of Statistics, Columbia University, New York, New York 10027, U.S.A.

[*]*email:* yw2016@columbia.edu

Summary. Population admixture can be a confounding factor in genetic association studies. Family-based methods (Rabinowitz and Larid, 2000, *Human Heredity* **50,** 211–223) have been proposed in both testing and estimation settings to adjust for this confounding, especially in case-only association studies. The family-based methods rely on conditioning on the observed parental genotypes or on the minimal sufficient statistic for the genetic model under the null hypothesis. In some cases, these methods do not capture all the available information due to the conditioning strategy being too stringent. General efficient methods to adjust for population admixture that use all the available information have been proposed (Rabinowitz, 2002, *Journal of the American Statistical Association* **92,** 742–758). However these approaches may not be easy to implement in some situations. A previously developed easy-to-compute approach adjusts for admixture by adding supplemental covariates to linear models (Yang et al., 2000, *Human Heredity* **50,** 227–233). Here is shown that this augmenting linear model with appropriate covariates strategy can be combined with the general efficient methods in Rabinowitz (2002) to provide computationally tractable and locally efficient adjustment. After deriving the optimal covariates, the adjusted analysis can be carried out using standard statistical software packages such as SAS or R. The proposed methods enjoy a local efficiency in a neighborhood of the true model. The simulation studies show that nontrivial efficiency gains can be obtained by using information not accessible to the methods that rely on conditioning on the minimal sufficient statistics. The approaches are illustrated through an analysis of the influence of apolipoprotein E (APOE) genotype on plasma low-density lipoprotein (LDL) concentration in children.

Key words: Family-based study; Genetic association study; Population stratification.

1. Introduction

Association studies are used to locate genetic locus influencing a trait of interest by evaluating association between allelic variability at a candidate locus and a trait. These association-based methods are especially useful for evaluating genetic factors with small-to-moderate effects (Risch and Merikangas, 1996). However, it is well known that population admixture can be a confounding factor for association-based methods (see for example, Elston, 1998). When the study sample consists of subjects drawn from subpopulations with different allele frequencies and trait distributions, spurious association may be detected even when the trait and the gene are not biologically linked.

Family-based association studies have been proposed to adjust for population admixture. The transmission disequilibrium test (TDT) and its extensions (Falk and Rubinstein, 1987; Terwilliger and Ott, 1992; Spielman, McGinnis, and Ewens, 1993; Lazzeroni and Lange, 1998; Spielman and Ewens, 1998) examine the transmission of parental alleles to the offsprings in a family given the parental genotypes. By conditioning on the parental genotypes, the bias due to population admixture is avoided. However when the parental genotypes are not all observed, these tests cannot be applied. Applying TDT tests to restricted data sets of families with complete parental genotypes can result in bias (Curtis and Sham, 1995). Rabinowitz and Larid (2000) proposed a general approach to adjust for population admixture by comparing the test statistics to their conditional distributions given the minimal sufficient statistics for the genetic model under the null hypothesis that does not require complete parental genotype information.

When the association between a trait and a locus has been established by testing, it may be desirable to estimate the form and strength of the trait-genotype association and to evaluate the interaction between genotypes and other environmental factors. Yang et al. (2000) proposed an approach to estimate candidate gene effect in a linear model that is not affected by spurious association. In this approach, the paradigm of conditioning on minimal sufficient statistics is achieved through augmenting the standard regression models with appropriate additional covariates. The approach is computationally convenient: after obtaining additional covariates, the analysis can be carried out by standard statistical packages such as SAS or R.

The approaches proposed in Rabinowitz and Larid (2000) for testing and in Yang et al. (2000) for estimation do not capture all of the available information (the minimal sufficient statistic is not always complete). Rabinowitz (2002) proposed a general framework to develop efficient test statistic that exploits all of the available information that is not potentially confounded by population stratification. Whittemore (2004) proposed a general framework for efficient estimation functions that protects against population stratification. Allen, Satten, and Tsiatis (2005) developed locally efficient estimation of haplotype-disease association in case-parent trio designs that is robust to confounding.

Here is shown that the approach of eliminating bias by adding supplemental covariates to a linear model in Yang et al. (2000) can be combined with the method of deriving efficient estimating equations in Whittemore (2004) or Rabinowitz (2002) to obtain a computationally tractable estimation approach that is efficient but not confounded by population admixture. The optimal supplemental covariates are obtained through matrix algebra calculations. In the cases where the family-specific effects are absent, the additional covariates have closed-form expressions. The proposed methods enjoy a local efficiency in a neighborhood of the true model. We use simulation studies to investigate unbiasedness and efficiency of the methods under conditions including violation of assumptions and departure from the true model. The simulation results show that nontrivial efficiency gains can be obtained by using information not accessible to methods that rely on conditioning on the minimal sufficient statistics. The approaches are illustrated through an analysis of the influence of apolipoprotein E (APOE) genotype on plasma low-density lipoprotein (LDL) concentration in children.

2. Methods

In this section, linear models for quantitative traits are introduced and optimal additional covariates to be included in the models are derived. Two models are considered: one does not involve family-specific terms and the other includes random family-specific effects.

2.1 *Model Without Family-Specific Effects*

Let Y_{ij} denote a quantitative trait of the jth individual in the ith family, and let G_{ij} denote the genotype of the same individual at a candidate locus. A simple linear model relating Y_{ij} to G_{ij} is

$$Y_{ij} = X(G_{ij})\beta + \varepsilon_{ij}, \tag{1}$$

where $X(G_{ij})$ is a coding for the genotype, and β is the effect of the genotype on the trait. For example, for a recessive trait, $X(G_{ij})$ can take value one for subjects carrying two copies of the disease allele and value zero for subjects carrying zero or one copy of the disease allele. The ε_{ij} are residual effects other than the genotypes under examination, which may be environmental factors that are independent of the genotypes or genetic factors at unlinked loci. When there is no population admixture, ε_{ij} are independent of G_{ij} and the usual least square estimate of β is unbiased. However, when there is population admixture, the membership of subpopulation is part of the residual effects ε_{ij}. Since the subpopulation membership influences the genotype distribution in the subpopulation, ε_{ij} and G_{ij} are correlated so that the ordinary least squares estimate of β is biased.

To motivate the derivation of efficient estimator of β when the genotypes and the residual effects are not independent, it is useful to review the method proposed in Yang et al. (2000). The validity of Yang et al. (2000) relies on the fact that even though G_{ij} and ε_{ij} in model (1) may be marginally correlated, they are conditionally independent, given the minimal sufficient statistic for the genetic model under the null hypothesis. In Yang et al. (2000), after adding the conditional expectation of $X(G_{ij})$ given the minimal sufficient statistics as additional covariates to the model, the least squares estimate for β is unbiased even when population admixture is present. In the current work, the form of the optimal additional covariates that leads to efficient estimator of β is unknown, and is derived from the constrained optimization problem.

Similar to Yang et al. (2000), the key assumption underlying the proposed approaches is that although the genotype-related covariates and the residuals are correlated due to population admixture, they are conditionally independent given the founder genotypes. As noted in Yang et al. (2000), this assumption corresponds to the transmission of parental alleles to the offspring generation being independent to any other factors that influence the trait, given the parental genotypes. While this assumption is surely an approximation of the biological truth, it is the basis for other methods adjusting for population admixture in family-based association studies (for example TDT and FBAT).

Let U_{ij} denote the unknown additional covariate for the jth individual from the ith family, let n denote the total number of families, let n_i denote the number of subjects in the ith family, and let N denote the total number of subjects. A linear model relating a trait to a genotype with additional covariates is

$$Y_{ij} = X(G_{ij})\beta + U_{ij}\gamma + \varepsilon_{ij}.$$

Without loss of generality, the above model can be written as

$$Y_{ij} = (X(G_{ij}) - U_{ij})\beta + U_{ij}\tilde{\gamma} + \varepsilon_{ij}, \tag{2}$$

where $\tilde{\gamma} = -\beta + \gamma$. In standard multiple regression analysis, the least squares estimates of a subset of regression coefficients can be acquired by first regressing the covariates of these coefficients on the covariates of the remaining coefficients, taking the residuals, and then regressing the response variable on the residuals. We apply this observation to model (2), where we are interested in estimating β. When the covariates $X(G_{ij}) - U_{ij}$ are uncorrelated with U_{ij} (in terms of expectation), the residuals of regressing $X(G_{ij}) - U_{ij}$ on U_{ij} are $X(G_{ij}) - U_{ij}$ themselves. Therefore the least squares estimate of β can then be written as

$$\widehat{\beta} = [(X-U)^T(X-U)]^{-1}(X-U)^T Y, \tag{3}$$

where the unknown covariates U satisfy

$$E(X-U)^T U = 0. \tag{4}$$

Here, $X_i = (X(G_{i1}), \ldots, X(G_{in_i}))^T, U_i = (U_{i1}, \ldots, U_{in_i})^T$, $X = (X_1, \ldots, X_n)^T, U = (U_1, \ldots, U_n)^T$, and $Y = (Y_1, \ldots, Y_n)^T$.

Let $G_i^\star$ denote the genotypes of founders in the ith family. Note that $G_i^\star$ may not be completely observed. An important assumption ensuring the validity of the proposed methods analogous to that in Yang et al. (2000) is that even though G_{ij} and ε_{ij} may be marginally dependent, they are conditionally independent given $G_i^\star$. It is shown in the Appendix that under this assumption, the expectation of $\widehat{\beta}$ in (3) is

$$\beta + E\Bigg\{ [(X-U)^T(X-U)]^{-1} \sum_i E\big[(X_i - U_i)^T \mathbf{1}_{n_i} \,|\, G_i^\star\big] E\big[\varepsilon_i \,|\, G_i^\star\big] \Bigg\}, \tag{5}$$

and the variance of $\widehat{\beta}$ is

$$\sigma^2 E[(X-U)^T(X-U)^{-1}], \tag{6}$$

where $\mathbf{1}_{n_i}$ is the $n_i \times 1$ vector of one. From (5), the condition for obtaining unbiased estimate of β is then

$$E\big[(X_i - U_i)^T \mathbf{1}_{n_i} \,|\, G_i^\star\big] = 0, \qquad i = 1, \ldots, n. \tag{7}$$

It follows that the optimal unbiased estimator for β can be obtained by minimizing (6) under the constraints (4) and (7).

It is convenient to introduce some notations to describe the solution to the above constrained maximization problem. Let $\vartheta_i^\star$ denote all possible founder genotypes compatible with that observed in the ith family, and let ϑ_i denote all possible combinations of offspring genotypes in the ith family. Let c_i denote the dimension of ϑ_i and let d_i denote the dimension of $\vartheta_i^\star$. Let W_i denote the $c_i \times c_i$ diagonal matrix with the diagonal entries given by the (possibly misspecified) probabilities of founder genotypes, $P(g)$. In reality, W_i are usually computed from observed founder genotypes. Let Z_i denote the $c_i \times d_i$ matrix with the $(g, g^\star)$ entry given by the conditional probability of an offspring genotype given the founder genotypes, $P(g \,|\, g^\star)$. Let X_i denote the $n_i \times c_i$ matrix with rows index individuals in a family and columns index components in ϑ_i; that is, the kth row in the matrix is $(X_{ik}(g_1), \ldots, X_{ik}(g_{c_i}))$. Let V_i denote the $n_i \times c_i$ matrix with the (m, g)th component being $U_{im}(g)$. The corresponding matrices W_i, Z_i, X_i, and V_i for an example pedigree are given in the Appendix. With these notations, it is shown in the Appendix that the solution to the constrained optimization problem of minimizing (6) subject to (4) and (7) is

$$V_i = X_i Z_i \big(Z_i^T W_i^{-1} Z_i\big)^{-1} Z_i^T W_i^{-1}. \tag{8}$$

The additional covariates U_i can be picked from the rows of V_i that correspond to the observed genotypes in members of the ith family. A simple example illustrating the computations is presented in the Appendix.

We use family-specific residual sums to estimate the variance of $\hat{\beta}$. By the conditions (4) and (7), the estimating equation for $\hat{\beta}$,

$$\sum_i (X_i - U_i)^T [Y_i - (X_i - U_i)\beta],$$

has expectation zero. Since the family-specific residual terms are independent, the variance of the solution to this estimating equation is (Cox and Hinkley, 1979)

$$\frac{\sum_i [(X_i - U_i)^T (Y_i - (X_i - U_i)\hat{\beta})]^2}{\left[\sum_i (X_i - U_i)^T (X_i - U_i)\right]^2}. \tag{9}$$

One advantage of the proposed estimator is that it is unbiased even when the founder genotype distributions W_i are misspecified. To see this, note that the condition for obtaining unbiasedness (7) holds regardless of whether the marginal probabilities W_i are correctly specified. To be specific, denote the misspecified W_i as W_i^*. As derived in the Appendix, the solutions for the corresponding U_i^* are the components of

$$X_i Z_i \big(Z_i^T W_i^{*-1} Z_i\big)^{-1} Z_i^T W_i^{*-1}.$$

By the introduced notations, taking the conditional expectation of a random variable amounts to multiplying it by components of Z_i. Therefore, from

$$\begin{aligned} E\big[\big(X_i - U_i^*\big)^T \mathbf{1}_{n_i} \,|\, G_i^\star\big] &= \mathbf{1}_{n_i}^T Z_i^T X_i^T \mathbf{1}_{n_i} - \mathbf{1}_{n_i}^T Z_i^T W_i^{*-1} \\ &\quad \times Z_i \big(Z_i^T W_i^{*-1} Z_i\big)^{-1} Z_i^T X_i^T \mathbf{1}_{n_i} \\ &= \mathbf{1}_{n_i}^T Z_i^T X_i^T \mathbf{1}_{n_i} - \mathbf{1}_{n_i}^T Z_i^T X_i^T \mathbf{1}_{n_i} = 0, \end{aligned}$$

it follows that the unbiasedness condition (7) is satisfied with misspecified W_i (denoted as W_i^*) and misspecified U_i (denoted as U_i^*). Although in this case the adjusted estimator remains unbiased, it is not efficient. We study the magnitude of efficiency loss due to misspecification of W by simulations in Section 3.

2.2 *Including Family-Specific Effects*

In some situations, there may exist family-specific effects influencing a trait. The linear model in this case can be expressed as

$$Y_{ij} = X(G_{ij})\beta + \alpha_i + \varepsilon_{ij}, \tag{10}$$

where α_i is a random family-specific factor. Due to population stratification, α_i may not be independent of G_{ij}, but may be independent of the subject-specific residual effects ε_{ij}. An assumption ensuring validity of the proposed methods in this case is that even though G_{ij} and α_i may be marginally correlated, they are conditionally independent given $G_i^\star$.

When there are family-specific effects, to obtain the optimal estimator, weighted least squares

$$\widehat{\beta}' = [(X-U)^T \Sigma^{-1} (X-U)]^{-1} (X-U)^T \Sigma^{-1} Y \tag{11}$$

should be used, where Σ is the conditional covariance matrix of Y given $G^\star$. From the Appendix, the expectation of $\widehat{\beta}'$ is

$$\begin{aligned} \beta &+ E\Bigg\{ \big[(X-U)^T \Sigma^{-1} (X-U)\big]^{-1} \\ &\qquad \times \sum_i E\big[(X_i - U_i)^T \Sigma_i^{-1} \mathbf{1}_{n_i} \,|\, G_i^\star\big] E\big[\alpha_i \,|\, G_i^\star\big] \Bigg\} \\ &+ E\Bigg\{ \big[(X-U)^T \Sigma^{-1} (X-U)\big]^{-1} \\ &\qquad \times \sum_i E\big[(X_i - U_i)^T \Sigma_i^{-1} \mathbf{1}_{n_i} \,|\, G_i^\star\big] E\big[\varepsilon_{ij} \,|\, G_i^\star\big] \Bigg\}, \end{aligned} \tag{12}$$

where Σ_i is the conditional covariance of the observations from the ith family given the founder genotypes, $\mathrm{cov}(Y_i Y_i^T \,|\, G_i^\star)$. The unbiasedness condition analogous to (7) should be modified as

$$E\left[(X_i - U_i))^T \Sigma_i^{-1} \mathbf{1}_{n_i} \mid G_i^\star\right] = 0, \qquad i = 1, \ldots, n. \tag{13}$$

The constraint analogous to (4) should be modified as

$$E(X - U)^T \Sigma^{-1} U = 0. \tag{14}$$

In addition, we show in the Appendix that the variance to be minimized is

$$E(X - U)^T \Sigma^{-1} (X - U). \tag{15}$$

Now define

$$X_i' = \Sigma_i^{-\frac{1}{2}} X_i, U_i' = \Sigma_i^{-\frac{1}{2}} U_i, Y_i' = \Sigma_i^{-\frac{1}{2}} Y_i, \text{and } \mathbf{1}_{n_i}' = \Sigma_i^{-\frac{1}{2}} \mathbf{1}_{n_i}.$$

With these notations, the constraint (13) becomes (7), the constraint (14) becomes (4), and the minimization term (6) becomes (15) with X_i, U_i, and $\mathbf{1}_{n_i}$ replaced as X_i', U_i', and $\mathbf{1}_{n_i}'$. Consequently, carrying out the same constrained optimization procedure for model (1) using the newly defined variables leads to the solution

$$V_i' = X_i' Z_i \left(Z_i^T W_i^{-1} Z_i\right)^{-1} Z_i^T W_i^{-1}. \tag{16}$$

In practice, we can fit a linear mixed effects model that has a random family-specific effect and include the original supplemental covariates without considering the family-specific effects (the $\hat{\beta}$ obtained from fitting a linear mixed effects model is estimated by weighted least squares).

It is worth mentioning that population admixture behaves as a source of family-specific effects (Fulker et al., 1999; Abecasis, Cardon, and Cookson, 2000). The weighted least squares approach therefore provides efficiency gain over the ordinary least squares even when there are no additional family-specific effects, α_i. We compare the two approaches by simulations. For the weighted least squares, the variance can be estimated by

$$\frac{\displaystyle\sum_i [(X_i - U_i)^T \hat{\Sigma}_i^{-1} (Y_i - (X_i - U_i)\hat{\beta})]^2}{\left[\displaystyle\sum_i (X_i - U_i)^T \hat{\Sigma}_i^{-1} (X_i - U_i)\right]^2}. \tag{17}$$

3. Simulation Studies

In this section, we use extensive simulation studies to evaluate properties of the proposed methods. We examine the effect of misspecifying marginal probabilities in W and the influence of the family-specific effects. We also compare ordinary least squares with weighted least squares, and the proposed methods with Yang et al. (2000).

We generated 100 nuclear families each with two children. To simulate population stratification, we drew parental genotypes from a 50:50 admixture of two populations. Parents in the same family were drawn from the same population. The disease allele frequency in each subpopulation was 0.1 and 0.3, so that the marginal disease allele frequency in the whole population was 0.2. The parental genotypes within each subpopulation were simulated based on the Hardy–Weinberg equilibrium. The offspring genotypes were simulated based on Mendelian transmission probabilities. We assumed a dominant effect of the disease allele. We simulated a linear model with different intercept for each population. To investigate the impact of varying severity of population admixture, we simulated several combinations of the intercepts. The intercept in the first population was 5, while in the second population was 10 or 20. The genetic effect was chosen to be 10, 20, 50, or 100. In the models with random family-specific effects, the variances of these effects were 5, 15, or 25. There were 1000 replications in each set of the simulations. We simulated residuals from a normal distribution with mean zero and standard deviation five.

Table 1
Type I error rates of various methods

α level	Unadjusted OLS	Adjusted OLS	Unadjusted WLS	Adjusted WLS	Adjusted Yang*
			$\mu_1 = 5, \mu_2 = 10$		
0.01	0.208	0.007	0.138	0.008	0.007
0.05	0.432	0.041	0.343	0.05	0.043
0.1	0.57	0.091	0.477	0.099	0.09
			$\mu_1 = 5, \mu_2 = 20$		
0.01	0.83	0.007	0.201	0.012	0.007
0.05	0.952	0.043	0.425	0.053	0.046
0.1	0.975	0.103	0.571	0.105	0.106

*Adjusted by method in Yang et al. (2000).

We first examine performance of the proposed methods under the null hypothesis ($\beta = 0$). Table 1 shows the type I error rates of test statistics computed using various methods. When the population admixture is moderate, the type I error rates of the unadjusted ordinary least squares and unadjusted weighted least squares are clearly much higher than the nominal level. For example, for $\alpha = 0.05$, the type I error is 0.43 for the former and 0.34 for the latter. In contrast, the proposed adjusted ordinary and weighted least squares methods have maintained the desirable error rates. We also investigate the method in Yang et al. (2000) and find its type I error rate to be close to the nominal level. When the population admixture is more severe, the type I error rates of the two unadjusted analyses are substantially higher than the nominal level while the proposed approach and Yang et al. (2000) have maintained the correct *α-level.*

Next we examine performance of various methods under the alternative hypothesis ($\beta \neq 0$). The first set of simulations corresponds to model (1) where there are no family-specific effects. Table 2 summarizes results for the ordinary least squares and the weighted least squares method. We first examined the properties of the proposed estimator when there was no genetic effect, that is, $\beta = 0$. Under this null model, the unadjusted estimator reported a large spurious genetic effect, that is, the mean $\hat{\beta} = 1.54$ for the ordinary least squares and the mean $\hat{\beta} = 1.39$ for the weighted least squares analysis. In contrast, the adjusted estimators were very close to zero: the mean $\hat{\beta}$ was 0.02 (average empirical SE = 1.38) for ordinary least squares and the mean $\hat{\beta} = 0.02$ (average empirical SE = 1.37) for weighted least squares. Next we examined the estimator where the true genetic effect was greater than zero ($\beta = 10$, 20, 50 or 100). It can be seen that the unadjusted estimates had large bias while the adjusted estimates had negligible bias: the bias of the former ranged from 1.48

Table 2
Estimates with correctly specified W: no family-specific effects

	$\mu_1 = 5, \mu_2 = 10$				$\mu_1 = 5, \mu_2 = 20$			
True β	Unadjusted mean $\hat{\beta}$	Adjusted mean $\hat{\beta}$	Empirical S.E.	Estimated S.E.	Unadjusted mean	Adjusted mean $\hat{\beta}$	Empirical S.E.	Estimated S.E.
			Ordinary least squares estimates (OLS)					
0	1.54	0.02	1.38	1.31	4.48	−0.02	1.89	1.88
10	11.48	9.98	1.32	1.31	14.53	9.98	1.90	1.87
20	21.53	20.01	1.35	1.31	24.48	19.93	1.88	1.87
50	51.51	50.00	1.33	1.31	54.50	49.99	1.89	1.89
100	101.51	100.05	1.34	1.31	104.43	100.00	1.89	1.87
			Weighted least squares estimates (WLS)					
0	1.39	0.02	1.37	1.32	2.27	0.02	1.45	1.47
10	11.32	9.98	1.32	1.32	12.27	10.02	1.49	1.47
20	21.31	20.01	1.33	1.32	22.26	19.95	1.48	1.47
50	51.35	49.99	1.33	1.32	52.30	50.02	1.48	1.46
100	101.38	100.05	1.32	1.32	102.22	100.05	1.47	1.46
			Comparing to Yang et al. (2000)					
True β	Yang mean $\hat{\beta}$	Empirical S.E.	Eff. gain OLS*	Eff. gain WLS†	Yang mean $\hat{\beta}$	Empirical S.E.	Eff. gain OLS*	Eff. gain WLS†
10	10.32	2.20	40%	40%	10.88	2.91	35%	49%
20	20.33	2.25	40%	41%	20.85	2.88	35%	49%
50	50.3	2.25	41%	41%	50.78	2.93	35%	49%
100	100.2	2.20	39%	40%	100.84	2.83	33%	48%

*[SE(Yang) − SE(OLS)]/SE(Yang); †[SE(Yang) − SE(WLS)]/SE(Yang).

to 4.53, while for the latter it ranged from zero to 0.07. The magnitude of the bias for the unadjusted methods increased with the severity of population stratification. When the population admixture was moderate ($\mu_1 = 5, \mu_2 = 10$), the bias of the unadjusted least squares estimator was around 1.5. When the population admixture was more severe ($\mu_1 = 5, \mu_2 = 20$), the bias increased to around 4.5. The bias for the unadjusted weighted least squares estimator in each scenario of the population admixture was 1.3 and 2.3, respectively. The magnitude of the bias was similar across all values of the genetic effect for both estimators.

Note that when the population admixture was moderate, the standard errors of the ordinary and weighted least squares estimators were similar. However, when the population admixture was more substantial ($\mu_1 = 5, \mu_2 = 20$), the weighted least squares method was more efficient even when there were no family-specific effects. This is because population admixture acts as a source of family-specific effects (Fulker et al., 1999; Abecasis et al., 2000) in which case the weighted least squares method is more efficient. The efficiency gains of the weighted least squares increased with the severity of admixture. When the difference between the intercepts of the two populations was 15, the reduction of the empirical standard error of the estimator was up to 24%. The estimated standard errors were close to the empirical ones.

We compare the efficiency of the proposed methods with Yang et al. (2000), that is, adding conditional expectations of the genotypes given the minimal sufficient statistics of the null model as additional covariates in the linear model. We see from the bottom panel of Table 2 that the efficiency gains of the proposed methods ranged from 33% to 49%, which were nontrivial. Note that the efficiency gains of ordinary least squares versus weighted least squares were similar when the admixture was moderate (the left panel in Table 2). When the admixture was more severe (the right panel in Table 2), the efficiency gains increased from about 35% in ordinary least squares to about 49% in weighted least squares.

The second set of simulations corresponds to model (10) where there are random family-specific effects. The variance of these effects was 15. Again we examined the estimator both under a null model ($\beta = 0$) and under several alternative models ($\beta > 0$). The same phenomenon of the unadjusted estimators reporting spurious genetic effect when the true β was zero while the adjusted estimators were very close to zero was also observed for this set of simulations. From Table 3, we also see that both the ordinary least squares and the weighted least squares methods provided unbiased estimates. As expected, the weighted least squares estimates were more efficient. When the admixture was moderate, using weighted least squares instead of ordinary least squares reduced the empirical standard error by up to 10%. When the population admixture was more severe, the corresponding reduction was up to 29%. We noticed larger efficiency gains of the proposed methods over Yang et al. (2000) in this set of simulations (the bottom panel of Table 3). The efficiency gains ranged from 40% to 60%, which are again substantial.

In the third set of simulations, we investigate the unbiasedness of $\hat{\beta}$ when W_i is misspecified. We analyzed simulated data with correctly specified allele frequency (0.2), moderately misspecified frequency (0.4), and substantially misspecified frequency (0.9). Tables 4 and 5 summarize results under different severity of admixture for models (1) and (10). We see that

Table 3
Estimates with correctly specified W: with family-specific effects, $var(\alpha_i) = 15$

	$\mu_1 = 5, \mu_2 = 10$				$\mu_1 = 5, \mu_2 = 20$			
True β	Unadjusted mean $\hat{\beta}$	Adjusted mean $\hat{\beta}$	Empirical S.E.	Estimated S.E.	Unadjusted mean	Adjusted mean $\hat{\beta}$	Empirical S.E.	Estimated S.E.
			Ordinary least squares estimates (OLS)					
0	1.53	−0.05	1.57	1.51	4.51	0.02	2.02	2.01
10	11.54	10.03	1.53	1.53	14.53	9.92	2.04	2.01
20	21.44	19.95	1.51	1.52	24.45	20.02	2.04	2.02
50	51.50	49.99	1.51	1.52	54.50	49.92	2.05	2.02
100	101.49	99.91	1.54	1.52	104.49	99.98	2.09	2.02
			Weighted least squares estimates (WLS)					
0	1.07	−0.04	1.42	1.40	1.96	−0.03	1.42	1.48
10	11.05	10.04	1.41	1.41	11.96	9.96	1.53	1.49
20	21.07	20.0	1.36	1.40	22.04	20.02	1.45	1.48
50	51.07	50.0	1.36	1.40	51.96	49.95	1.50	1.48
100	101.1	99.9	1.40	1.41	102.02	100.08	1.50	1.48
			Comparing to Yang et al. (2000)					
True β	Yang mean $\hat{\beta}$	Empirical S.E.	Eff. gain OLS*	Eff. gain WLS†	Yang mean $\hat{\beta}$	Empirical S.E.	Eff. gain OLS*	Eff. gain WLS†
10	10.50	3.01	49%	53%	10.81	3.59	43%	57%
20	20.23	2.99	49%	55%	20.78	3.65	44%	60%
50	50.24	3.02	50%	55%	50.80	3.61	43%	58%
100	100.23	3.09	50%	55%	100.80	3.49	40%	57%

*[SE(Yang) − SE(OLS)]/SE(Yang); †[SE(Yang) − SE(WLS)]/SE(Yang).

Table 4
Unbiasedness under misspecification of W: no family-specific effects, $\mu_1 = 5, \mu_2 = 20$

	$p = 0.2$			$p = 0.4$			$p = 0.9$		
True β	Adjusted mean $\hat{\beta}$	Empirical S.E.	Estimated S.E.	Adjusted mean $\hat{\beta}$	Empirical S.E.	Estimated S.E.	Adjusted mean $\hat{\beta}$	Empirical S.E.	Estimated S.E.
				Ordinary least squares estimates					
10	9.94	1.89	1.88	9.73	1.95	1.86	9.80	1.93	1.83
20	19.97	1.94	1.87	19.68	1.90	1.87	19.81	1.94	1.85
50	49.90	1.92	1.87	49.75	1.94	1.88	49.74	1.91	1.84
100	99.99	1.91	1.87	99.70	1.91	1.87	99.70	1.9	1.82
				Weighted least squares estimates					
10	9.94	1.51	1.47	9.93	1.52	1.48	9.94	1.55	1.51
20	20.00	1.49	1.47	49.90	1.51	1.48	49.88	1.53	1.51
50	49.98	1.48	1.47	49.90	1.51	1.48	49.9	1.56	1.5
100	99.93	1.47	1.47	99.90	1.51	1.48	99.90	1.47	1.51

the unbiasedness holds with misspecified W. The mean bias of both methods ranged from 0.01 to 0.3. There appears to be a small sample bias for the proposed methods when the allele frequency was severely misspecified as 0.4 and 0.9. However, the bias went away when we increased the sample size to 200 families. Specifically, for ordinary least squares, the mean bias decreased from approximately 0.2 (when $p = 0.4$) to 0.04 and from approximately 0.3 (when $p = 0.9$) to 0.05. For weighted least squares, the mean bias decreased from approximately 0.1 (when $p = 0.4$) to 0.02 and from approximately 0.1 (when $p = 0.9$) to 0.03.

In the fourth set of simulations, we investigate efficiency loss due to misspecification of W. In Tables 6 and 7, we present the empirical standard errors of the point estimates with different scenarios of misspecification. For model (1), where there are no family-specific effects, when the allele frequency was moderately misspecified ($p = 0.4$), the efficiency loss ranged from 1% to 5%, which was moderate. The efficiency loss for ordinary least squares and weighted least squares was similar. When the allele frequency was severely misspecified ($p = 0.9$), the efficiency loss ranged from 7% to 13 %. For model (10), where there are family-specific effects, the efficiency loss ranged from 0% to 6%. The magnitudes of the loss were comparable for the moderately and severely misspecified allele frequency.

In the fifth set of simulations, we examine the efficiency loss due to adjusting for population admixture when it is in fact absent. We considered both the cases when the family-specific

Table 5

Unbiasedness under misspecification of W: with family-specific effects, $\mu_1 = 5, \mu_2 = 20, var(\alpha_i) = 25$

	$p = 0.2$			$p = 0.4$			$p = 0.9$		
True β	Adjusted mean $\hat{\beta}$	Empirical S.E.	Estimated S.E.	Adjusted mean $\hat{\beta}$	Empirical S.E.	Estimated S.E.	Adjusted mean $\hat{\beta}$	Empirical S.E.	Estimated S.E.
Ordinary least squares estimates									
10	9.95	2.20	2.10	9.80	2.10	2.10	9.70	2.10	2.00
20	19.98	2.15	2.11	19.80	2.20	2.10	19.70	2.05	2.00
50	50.00	2.18	2.11	49.80	2.15	2.10	49.80	2.14	2.02
100	99.95	2.20	2.10	99.70	2.18	2.10	99.70	2.10	2.00
Weighted least squares estimates									
10	10.05	1.56	1.50	9.97	1.52	1.51	9.99	1.52	1.51
20	19.99	1.50	1.48	20.01	1.49	1.50	19.90	1.49	1.52
50	50.02	1.53	1.49	50.02	1.50	1.49	49.90	1.50	1.52
100	99.95	1.52	1.49	99.90	1.53	1..50	99.90	1.48	1.51

Table 6

Efficiency loss due to misspecification of W: no family-specific effects, $\mu_1 = \mu_2 = 5$

	$p = 0.2$		$p = 0.4$			$p = 0.9$		
True β	Adjusted mean $\hat{\beta}$	Empirical S.E.	Adjusted mean $\hat{\beta}$	Empirical S.E.	Efficiency loss	Adjusted mean $\hat{\beta}$	Empirical S.E.	Efficiency loss
Ordinary least squares estimates								
10	10.01	1.25	10.04	1.30	4%	10.05	1.34	7%
20	20.01	1.26	19.97	1.27	1%	19.97	1.35	8%
50	50.01	1.27	49.99	1.30	2%	49.99	1.36	7%
100	99.99	1.27	100.03	1.28	1%	100.02	1.40	10%
Weighted least squares estimates								
10	9.90	1.25	10.04	1.31	5%	10.05	1.35	8%
20	20.01	1.24	19.95	1.27	2%	19.99	1.35	8%
50	50.02	1.24	49.99	1.30	5%	49.99	1.37	10%
100	99.99	1.25	100.02	1.29	3%	100.00	1.41	13%

Table 7

Efficiency loss due to misspecification of W: with family-specific effects, $\mu_1 = \mu_2 = 5, var(\alpha_i) = 25$

	$p = 0.2$		$p = 0.4$			$p = 0.9$		
True β	Adjusted mean $\hat{\beta}$	Empirical S.E.	Adjusted mean $\hat{\beta}$	Empirical S.E.	Efficiency loss	Adjusted mean $\hat{\beta}$	Empirical S.E.	Efficiency loss
Ordinary least squares estimates								
10	10.02	1.61	9.99	1.67	4%	10.00	1.64	2%
20	19.96	1.59	19.99	1.66	4%	20.05	1.69	6%
50	50.03	1.63	49.99	1.66	2%	49.94	1.67	2%
100	99.99	1.60	99.95	1.60	0%	100.00	1.64	2%
Weighted least squares estimates								
10	10.03	1.42	9.99	1.46	3%	10.00	1.49	5%
20	19.98	1.43	20.02	1.49	4%	20.05	1.45	1%
50	50.00	1.43	50.01	1.50	5%	49.96	1.51	6%
100	99.98	1.43	99.95	1.44	1%	100.00	1.50	5%

effects were present and absent. Tables 8 and 9 summarize results under these two scenarios. When there was no admixture and the family-specific effects were absent, the efficiency loss of the adjusted analysis compared to the unadjusted ranged from 34% to 38%. The magnitude of loss was similar for ordinary and weighted least squares. When there were family-specific effects, the efficiency loss was around 28%. The efficiency loss of ordinary and weighted least squares was also similar.

In the sixth set of simulations, we investigate the influence of departure from constant family-specific effects model. In Table 10, we compare the empirical standard errors of $\hat{\beta}$

Table 8

Efficiency loss due to adjusting for admixture when no admixture is present: no family-specific effects, $\mu_1 = \mu_2 = 5$

	Unadjusted		Adjusted		
True	Mean	Emp.	Mean	Emp.	Efficiency
β	$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.	loss*
	Ordinary least squares estimates				
10	10.00	0.73	9.99	1.11	34%
20	19.95	0.79	19.95	1.19	34%
50	49.99	0.77	50.02	1.23	37%
100	100.01	0.71	100.03	1.15	38%
	Weighted least squares estimates				
10	10.00	0.73	9.99	1.11	34%
20	19.95	0.79	19.95	1.19	34%
50	49.99	0.78	50.02	1.23	37%
100	100.01	0.71	100.03	1.15	38%

* [SE(Adjusted) − SE(Unadjusted)]/SE(Adjusted).

Table 9

Efficiency loss due to adjusting for admixture when no admixture is present: with family-specific effects, $\mu_1 = \mu_2 = 5, var(\alpha_i) = 15$

	Unadjusted		Adjusted		
True	Mean	Emp.	Mean	Emp.	Efficiency
β	$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.	loss*
	Ordinary least squares estimates				
10	9.99	1.01	9.97	1.41	28%
20	20.01	1.02	19.98	1.42	28%
50	50.02	1.01	50.02	1.39	27%
100	99.98	1.02	99.98	1.41	28%
	Weighted least squares estimates				
10	9.99	0.97	9.97	1.34	28%
20	19.99	0.97	19.98	1.32	27%
50	50.01	0.95	50.02	1.31	27%
100	99.99	0.95	99.98	1.32	28%

*[SE(Adjusted) − SE(Unadjusted)]/SE(Adjusted).

under varying values of the variance of the family-specific effects. The variance of α_i in each model was 0, 15, or 25. For the ordinary least squares method, the standard error of $\hat{\beta}$ increased with increasing variance of the family-specific effects, and the loss of efficiency ranged from 8% to 17%. For the weighted least squares method, the standard errors of the estimators were similar regardless of the value of the family-specific variance (efficiency loss up to 3%). In other words, the family-specific effects have little influence on the efficiency of the estimates obtained by weighted least squares.

4. Data Analysis

In this section, the proposed approaches are applied to an analysis of the influence of APOE genotype on plasma LDL concentrations in young children. There are three common alleles at the APOE locus ($\varepsilon2, \varepsilon3, \varepsilon4$). The apo $\varepsilon3$ is the most prevalent allele in the general population, with a frequency of 75% to 80%. The frequency of apo $\varepsilon4$ allele varies with ethnicity (Howard, Gidding, and Liu, 1998). Previous studies of adults have shown that apo $\varepsilon4$ allele is associated with higher LDL cholesterol levels compared to $\varepsilon3$ (Davignon, Gregg, and Sing, 1988), while the role of $\varepsilon2$ is more complicated.

The effect of APOE gene was found to be larger in younger people (Hixson, 1991). Children included in this data analysis were recruited through the Columbia University BioMarkers Study, a cross-sectional study of children and their parents conducted from 1994 to 1998 (Shea et al., 1999; Isasi et al., 2000). Families were recruited from lists of cardiac patients generated through the Presbyterian Hospital Clinical Information System, private cardiology practices, lipid clinics, pediatric practices at Columbia-Presbyterian Medical Center, and fliers posted within the medical center. Families with at least one healthy child, 4 to 25 years of age, were eligible for participation. Healthy was defined as not having any chronic medical condition under treatment by a pediatrician, other than high blood pressure or high lipids (referral criteria to the Children's Cardiovascular Health Center). Some subjects were recruited through family members other than the children. Around 75% of the children were Hispanic and the remaining 25% were non-Hispanic White.

There were 621 children recruited for the study, among whom 55 did not have data on the APOE genotype. Among

Table 10

Effect of departure from constant family-specific effects model ($\mu_1 = 5, \mu_2 = 20$)

	Var(α_i) = 0		Var(α_i) = 15			Var(α_i) = 25		
True	Adjusted	Empirical	Adjusted	Empirical	Increase	Adjusted	Empirical	Increase
β	mean $\hat{\beta}$	S.E.	mean $\hat{\beta}$	S.E.	of S.E.	mean $\hat{\beta}$	S.E.	of S.E.
			Ordinary least squares estimates					
10	9.98	1.90	9.87	2.13	12%	9.95	2.22	17%
20	19.93	1.88	19.83	2.03	8%	19.98	2.15	14%
50	49.99	1.89	49.90	2.10	11%	50.00	2.18	15%
100	100.00	1.89	99.97	2.11	12%	99.95	2.20	16%
			Weighted least squares estimates					
10	10.02	1.49	9.93	1.51	1%	10.05	1.49	0%
20	19.95	1.48	19.87	1.51	2%	19.99	1.50	1%
50	50.02	1.48	49.93	1.50	1%	50.02	1.49	1%
100	100.05	1.47	99.99	1.48	1%	99.95	1.52	3%

the children with genotype data, 10 were excluded because of Mendelian genotyping errors. There were 13 children without LDL concentration data. These children contributed to the computation of the additional covariates, but were not included in the subsequent regression analysis relating LDL to genotype because of missing LDL levels. The mean LDL of the 534 children was 103.8 and the standard deviation was 43.5. The frequencies of children with APOE genotypes $\varepsilon 2\varepsilon 2, \varepsilon 3\varepsilon 2, \varepsilon 3\varepsilon 3, \varepsilon 4\varepsilon 2, \varepsilon 4\varepsilon 3$, and $\varepsilon 4\varepsilon 4$ were 9.1%, 7.9%, 61.8%, 1.8%, 25.6%, and 2.0%. The mean LDL concentrations for children in each genotype group were 49.8, 84.9, 106.1, 107.2, 105.4, and 105.9.

Among the 547 children from 322 families with genotype data, 153 in 78 families had complete parental genotypes, 389 in 241 families had parental genotype available in one of the parents, and 5 of them in 3 families had no genotype information on any of the parents.

Three sets of analyses were presented. First, the unadjusted analysis was carried out. Then the analysis was repeated with two approaches to adjust for admixture: the first was the methods proposed in Yang et al. (2000), which computed the additional covariates as conditional expectation of the founder genotypes given the minimal sufficient statistics of the genetic model under the null; the second was the methods proposed here, which computed the additional covariate by (8) or (16). Results from the three analyses were compared.

In each set of the analyses, models with and without a family-specific random effect were fit to the data. The ordinary least squares estimates for model (1) were obtained by fitting a simple linear model, while weighted least squares estimates for model (10) were obtained by fitting a mixed effects model with random family-specific effects. Standard errors were computed as in (9) or (17). The APOE genotype was coded as the number of each of the three APOE alleles carried by a subject. In this data analysis example, Y_{ij} is the LDL concentration for the *j*th child in the *i*th family, $X(G_{ij}) = (X^{\varepsilon 2}(G_{ij}), X^{\varepsilon 3}(G_{ij}), X^{\varepsilon 4}(G_{ij}))^T$ are the numbers of $\varepsilon 2, \varepsilon 3$, and $\varepsilon 4$ allele carried by the child, $\beta = (\beta_2, \beta_3, \beta_4)^T$ are the effect of each of the three alleles, and α_i is the family-specific random effect.

Results of the unadjusted analysis are summarized in Table 11. The significant contrasts were $\beta_3 - \beta_2$ and $\beta_4 - \beta_2$. The estimated differences were 19.9 (SE: 5.3) and 20.7 (SE: 6.1). The interpretation for this analysis was that children carrying the apo $\varepsilon 2$ allele had a significantly lower LDL concentration than children with the $\varepsilon 3$ or $\varepsilon 4$ allele.

For the first adjusted analysis, the FBAT (Rabinowitz and Larid, 2000; Horvath, Xu, and Laird, 2001) was used to compute the conditional expectation of $X(G_i)$ given the minimal sufficient statistics of parental genotypes. Then a linear model was fit using these conditional expectations as additional covariates as in Yang et al. (2000). The results are summarized in Table 12. In this analysis, the parameter estimates for β were not identifiable, but the contrasts remained identifiable. The significant contrasts were still $\beta_3 - \beta_2$ and $\beta_4 - \beta_2$ as in the unadjusted analysis. The values of the contrasts were 24.5 (SE: 11.1) and 31.7 (SE: 11.0) for ordinary least squares, and 22.0 (SE: 9.8) and 30.1 (SE: 10.1) for weighted least squares. Note that when using the weighted least squares method, the effect for $\beta_3 - \beta_2$ changed from 18.8 (unadjusted) to 22.0 (adjusted), and the effect for $\beta_4 - \beta_2$ changed from 22.7 (unadjusted) to 30.1 (adjusted). Similar magnitude of increase was observed for the ordinary least squares estimates.

Table 11

Real data example: the unadjusted analysis

Parameter	Estimate	Standard error	*p* value
Ordinary least squares estimates			
$\varepsilon 2$	33.0	5.2	<0.001
$\varepsilon 3$	52.9	1.2	<0.001
$\varepsilon 4$	53.7	3.2	<0.001
$\varepsilon 3 - \varepsilon 2$	19.9	5.4	<0.001
$\varepsilon 4 - \varepsilon 3$	0.8	3.7	0.83
$\varepsilon 4 - \varepsilon 2$	20.7	6.2	<0.001
Weighted least squares estimates			
$\varepsilon 2$	33.9	6.3	<0.001
$\varepsilon 3$	52.7	1.3	<0.001
$\varepsilon 4$	56.6	2.8	<0.001
var(α_i^2)	721.5	145.4	<0.001
var(σ_i^2)	1159.7	115.1	<0.001
$\varepsilon 3 - \varepsilon 2$	18.8	6.5	<0.001
$\varepsilon 4 - \varepsilon 3$	3.9	3.4	0.24
$\varepsilon 4 - \varepsilon 2$	22.7	6.7	<0.001

Table 12

Real data example: adjusting by Yang et al. (2000)

Parameter	Estimate	Standard error	*p* value
Ordinary least squares estimates			
$\varepsilon 3 - \varepsilon 2$	24.5	11.1	0.03
$\varepsilon 4 - \varepsilon 3$	7.2	4.7	0.12
$\varepsilon 4 - \varepsilon 2$	31.7	11.0	0.004
Weighted least squares estimates			
var(α_i^2)	718.8	146.1	<0.001
var(σ_i^2)	1165.0	116.0	<0.001
$\varepsilon 3 - \varepsilon 2$	22.0	9.8	0.03
$\varepsilon 4 - \varepsilon 3$	8.2	4.3	0.06
$\varepsilon 4 - \varepsilon 2$	30.1	10.1	0.003

For the second adjusted analysis using the methods developed here, the matrices W_i and Z_i are required. The genotype frequencies in W_i were computed using observed founder genotypes. We have shown that misspecification of W_i does not affect the unbiasedness of β. To illustrate the computation of Z_i and the additional covariates, the calculation was carried out for an example pedigree with two children and one observed heterozygous parent in the Appendix.

The results of these analyses were summarized in Table 13. The significant estimated contrasts were $\beta_3 - \beta_2$ and $\beta_4 - \beta_2$, with values 18.6 (SE: 6.9) and 19.9 (SE: 6.9) for ordinary least squares, and 17.6 (SE: 7.3) and 22.0 (7.3) for weighted least squares. The changes of the contrasts in the proposed adjustment were smaller compared to the Yang adjustment, and the standard errors were also smaller. These comparisons suggest that applying Yang et al. (2000) may have overcorrected for population admixture. Furthermore, the larger standard errors for the contrasts in the Yang analysis compared to the proposed reflected the loss of information by conditioning on the minimal sufficient statistics in the Yang analysis. We can

Table 13
Real data example: adjusting by the proposed method

Parameter	Estimate	Standard error	p value
	Ordinary least squares estimates		
$\varepsilon 3 - \varepsilon 2$	18.6	6.9	0.007
$\varepsilon 4 - \varepsilon 3$	1.3	3.6	0.72
$\varepsilon 4 - \varepsilon 2$	19.9	6.9	0.004
	Weighted least squares estimates		
$\text{var}(\alpha_i^2)$	734.7	146.6	<0.001
$\text{var}(\sigma_i^2)$	1157.2	114.6	<0.001
$\varepsilon 3 - \varepsilon 2$	17.6	7.3	0.02
$\varepsilon 4 - \varepsilon 3$	4.4	3.4	0.20
$\varepsilon 4 - \varepsilon 2$	22.0	7.3	0.003

also see that the weighted least squares estimates had smaller standard error than the ordinary least squares estimates.

5. Discussion

Here a locally efficient approach to adjusting for population admixture when estimating genetic effect on a quantitative trait is proposed. The main step is to augment a linear regression model with supplemental covariates that provides unbiased minimal variance estimator for the genetic parameter of interest. The form of the additional covariates is similar in spirit to that proposed in Rabinowitz (2002) and Whittemore (2004). The models in (1) and (10) can be extended to include environmental factors.

In the testing context, it was observed by Whittemore and Halpern (2003) that both Rabinowitz and Larid (2000) and Rabinowitz (2002) can be formulated as solutions to a constrained optimization problem: the coefficient of variation of the test statistic is to be maximized under some constraint. In Rabinowitz and Larid (2000), the constraint is that the conditional expectation of the test statistic given the minimal sufficient statistic of the genetic model under the null hypothesis is zero; while in Rabinowitz (2002) the constraint is that the conditional expectation of the test statistic given the founder genotypes is zero. The column space of the constraints in the former contains the corresponding column space of the constraints of the latter: the vectors in the latter space can be expressed as a linear combination of vectors in the former space. This observation implies that the constraints in Rabinowitz and Larid (2000) are too restrictive and there is potential loss of information incurred by conditioning on a larger space. For example, families where all children have the same genotype do not contribute to the analysis. In contrast, the methods in Rabinowitz (2002) capture all the available information.

In the estimation context, the analogous comparison of efficiency is between Yang et al. (2000), which was based on Rabinowitz and Larid (2000), and the proposed approach, which has a similar form to Rabinowitz (2002) and Whittemore (2004). Yang et al. (2000) corresponds to projecting covariates involving genotypes subject to population admixture onto the space of the minimal sufficient statistics, while the proposed approach corresponds to projecting the genotype-related covariates onto an appropriate smaller space. The larger the space of projection, the more information is lost. Our simulation results suggest nontrivial efficiency gains of the proposed methods over Yang et al. (2000). It can also be seen from the real data analysis example that the standard errors of the proposed methods were smaller, which suggests overadjustment in Yang et al. (2000) because conditioning on the minimal sufficient statistic may be too restrictive.

It is shown in the Appendix that the proposed approach is optimal when there is no family-specific effect or when the family-specific effect is a constant. When such effect is not a constant, the proposed approach is locally optimal because the family-specific effect is approximately a constant (the variance of such effect is zero) when considered locally. We used simulation studies to investigate the efficiency of the estimators when the family-specific effect is not a constant. The efficiency of the ordinary least squares method was reduced to up to 17% (Table 10), but the efficiency of the weighted least squares method was not greatly influenced by the departure from the constant family-specific effect model (efficiency loss up to 3%, see Table 10).

The added covariates U_i involve marginal probabilities W_i, which are estimated (possibly incorrectly) from the data. The estimators depending on these estimated values would normally introduce extra variability. However, since the additional covariates $X_i - U_i$ can be viewed as residuals from the projection of X_i onto the space spanned by W_i, they are orthogonal to W_i. By the orthogonality, there is no additional variability introduced by estimating W_i.

Since population admixture acts as a source of family-specific effects, weighted least squares is more efficient than ordinary least squares even when there are no additional family-specific effects. For the real data analysis, weighted least squares should be used.

The proposed methods are designed for the single-locus model or the multilocus model without interaction. When there are multiple loci predisposing a disease and there is no interaction between the loci, we compute a set of optimal covariates for each locus using founder genotypes at this locus. All the optimal covariates will then be included in the linear model analyses. When there is interaction between the loci, the current approaches need to be modified to a haplotype-based method to account for this effect because haplotype association analysis may be more powerful than genotype association analysis (Morris and Kaplan, 2002). However, one complication faced by a haplotype analysis is that the phase of a haplotype is usually not observed. In a haplotype analysis, when the phase is known, one uses functions of haplotypes as predictors. When the phase is unknown, one uses the conditional distribution of the haplotypes given the genotypes to compute the conditional expectation of phase-unknown haplotype scores. The conditional distribution depends on the marginal distribution of haplotypes, which may be subject to population stratification or may be estimated using only approximated assumptions (e.g. Hardy–Weinberg equilibrium). To extend the methods developed here in this context, the covariates X_i in the equation (8) should be replaced by the estimated conditional expectation of the functions of haplotypes. The W_i should be replaced by the marginal distribution used in the calculations of X_i, and Z_i should be replaced by the matrix of conditional probability of all possible haplotypes given the parental haplotypes. Further research along this direction is underway.

The proposed methods are easy to implement and the computational cost is low. The extra computation involved other than fitting a linear model is to compute the additional optimal covariates. The form of these covariates (see (8)) suggests that they are constructed using marginal distribution of founder genotypes and conditional distribution of the offspring genotypes and involve some matrix algebra. These computations do not entail iterations and can be completed in seconds. In our simulations, it took half a minute to compute the optimal covariates for 1000 repetitions on a Dell Workstation with 2.00 GHz CPU. A link to the code to compute the optimal covariates can be found at `www.columbia.edu/∼yw2016`.

Here the methods are developed in the context of random sampling. When subjects with extreme values of a quantitative trait are oversampled or when certain outcomes are oversampled, these methods are generally biased. In such settings, the more general estimating equation conditioning on the outcomes proposed in Whittemore (2004) may be applicable.

Acknowledgements

Yuanjia Wang's research is supported by NIH AG031113-01A2. We thank Dr Steve Shea for providing the LDL data used to illustrate the proposed methods.

References

Abecasis, G. R., Cardon, L. R., and Cookson, W. O. C. (2000). A general test of association for quantitative traits in nuclear families. *American Journal of Human Genetics* **66,** 279–292.

Allen, A. S., Satten, G. A., and Tsiatis, A. A. (2005). Locally-efficient robust estimation of haplotype-disease association in family-based studies. *Biometrika* **92,** 559–571.

Cox, D. R. and Hinkley, C. V. (1979). *Theoretical Statistics*. London: Chapman & Hall.

Curtis, D. and Sham, P. C. (1995). A note on the application of the transmission disequilibrium test when a parent is missing. *American Journal of Human Genetics* **56,** 811–812.

Davignon, J., Gregg, R. E., and Sing, C. F. (1988). Apolipoprotein E polymorphism and atherosclerosis. *Arteriosclerosis* **8,** 1–21.

Elston, R. C. (1998). Linkage and association. *Genet Epidemiology* **15,** 565–576.

Falk, C. T. and Rubinstein, P. (1987). Haplotype relative risks: An easy reliable way to construct a proper control sample for risk calculations. *Annals of Human Genetics* **51,** 227–233.

Fulker, D. W., Cherny, S. S., Sham, P. C., and Hewitt, J. K. (1999). Combined linkage and association sib-pair analysis for quantitative traits. *American Journal of Human Genetics* **64,** 259–267.

Hixson, J. E. (1991). Apolipoprotein E polymorphisms affect atherosclerosis in young males: Pathobiological Determinants of Atherosclerosis in Youth (PDAY) research group. *Arteriosclerosis Thrombosis* **11,** 237–244.

Horvath, S., Xu, X., and Laird, N. (2001). The family based association test method: Strategies for studying general genotype-phenotype associations. *European Journal of Human Genetics* **9,** 301–306.

Howard, B. V., Gidding, S. S., and Liu, K. (1998). Association of apolipoprotein E phenotype with plasma lipoproteins in African-American and white young adults. *American Journal of Epidemiology* **148,** 859–868.

Isasi, C. R., Shea, S., Deckelbaum, R. J., Couch, S. C., Starc, T. J., Otvos, J. D., and Berglund, L. (2000). Apolipoprotein $\varepsilon 2$ allele is associated with an anti-atherogenic lipoprotein profile in children: The Columbia University Biomarker Study. *Pediatrics* **106,** 568–575.

Lazzeroni, L. C. and Lange, K. (1998). A conditional interference framework for extending the transmission/disequilibrium test. *Human Heredity* **48,** 67–81.

Morris, R. W. and Kaplan, N. L. (2002). On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genetic Epidemiology* **23,** 221–233.

Rabinowitz, D. (2002). Adjusting for population heterogeneity and misspecified haplotype frequencies when testing nonparametric null hypotheses in statistical genetics. *Journal of the American Statistical Association* **92,** 742–758.

Rabinowitz, D. and Larid, N. (2000). A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Human Heredity* **50,** 211–223.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273,** 1516–1517.

Shea, S., Isasi, C. R., Couch, S., Starc, T. J., Tracy, R. P., Deckelbaum, R., Talmud, P., Berglund, L., and Humphries, S. E. (1999). Relations of plasma fibrinogen level in children to measures of obesity, the (G-455->A) mutation in the beta-fibrinogen promoter gene, and family history of ischemic heart disease: The Columbia University BioMarkers Study. *American Journal of Epidemiology* **150,** 737–746.

Spielman, R. S. and Ewens, W. J. (1998). A sib-ship test for linkage in the presence of association: The sib transmission/disequilibrium test. *American Journal of Human Genetics* **62,** 450–458.

Spielman, R. S., McGinnis, R. E., and Ewens, W. J. (1993). Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* **52,** 506–516.

Terwilliger, J. D. and Ott, J. (1992). A haplotype-based "haplotype relative risk" approach to detecting allelic associations. *Human Heredity* **42,** 337–346.

Whittemore, A. (2004). Estimating genetic association parameter from family data. *Biometrika* **91,** 219–225.

Whittemore, A. and Halpern, J. (2003). Genetic association tests for family data with missing parental genotypes: A comparison. *Genetic Epidemiology* **25,** 80–91.

Yang, Q., Rabinowitz, D., Isasi, C., and Shea, S. (2000). Adjusting for confounding due to population admixture when estimating the effect of candidate genes on quantitative traits. *Human Heredity* **50,** 227–233.

Received October 2009. *Revised March* 2010.
Accepted April 2010.

Appendix

In this section, the expectation and the variance of $\widehat{\beta}$ were computed, the solution to the constrained optimization problem was derived, and the efficient additional covariates for an example pedigree was computed.

From the expression $\widehat{\beta} = [(X-U)^T(X-U)]^{-1}(X-U)^TY$ and the model (2), the expectation of $\widehat{\beta}$ is

$$\begin{aligned} E\widehat{\beta} &= E\{[(X-U)^T(X-U)]^{-1}(X-U)^TY\} \\ &= \beta + E\{[(X-U)^T(X-U)]^{-1}(X-U)^TU\tilde{\gamma}\} \\ &\quad + E\{[(X-U)^T(X-U)]^{-1}(X-U)^T\varepsilon\} \end{aligned}$$

$$= \beta + E\left\{ [(X-U)^T(X-U)]^{-1} \times \sum_i E\left[(X_i - U_i)^T \mathbf{1}_{n_i} \,|\, G_i^\star\right] E\left[\varepsilon_{ij} \,|\, G_i^\star\right] \right\},$$

where $\mathbf{1}_{n_i}$ denotes the $n_i \times 1$ vector of 1. Here the third equality follows from the constraint (4) and the fact that given the founder genotypes, $X(G)$ is conditionally independent of ε. It follows that the condition to ensure the unbiasedness of $\widehat{\beta}$ is that

$$E\left[(X_i - U_i)^T \mathbf{1}_{n_i} \,|\, G_i^\star\right] = 0, i = 1, \ldots, n.$$

Now turn to the computation of the variance. We have the expression

$$\mathrm{var}(\widehat{\beta}) = \mathrm{var}(E(\widehat{\beta} \,|\, G^\star)) + E(\mathrm{var}(\widehat{\beta} \,|\, G^\star)).$$

Under the conditions (4) and (7), we have $E(\widehat{\beta} \,|\, G^\star) = \beta$. Therefore the first term on the right-hand side of the expression is zero. Let A_i denote the vector $X(G_i) - U_i$, and let A denote the matrix $(A_1^T, \ldots, A_n^T)^T$. The second term can be calculated as

$$\begin{aligned} \mathrm{var}(\widehat{\beta} \,|\, G^\star) &= E\left\{ [A^T A]^{-1} \sum_i A_i^T \varepsilon_i \varepsilon_i^T A_i [A^T A]^{-1} \,|\, G^\star \right\} \\ &= \sum_i E\left(\varepsilon_{ij}^2 \,|\, G_i^\star\right) E\{[A^T A]^{-1} A_i^T A_i [A^T A]^{-1} \,|\, G^\star\} \\ &= \sigma^2 E\{[A^T A]^{-1} \,|\, G^\star\}. \end{aligned}$$

Here σ^2 is the variance of the residuals. The second equality follows from the conditional independence of ε_{ij} and G_{ij} given $G_i^\star$. Taking the expectation we obtain

$$\mathrm{var}(\hat{\beta}) = \sigma^2 E[A^T A]^{-1}. \tag{A.1}$$

Similarly, when there are family-specific effects, expectation of the weighted least squares $\widehat{\beta}'$ as in (11) is

$$\begin{aligned} E\widehat{\beta}' &= E\{[(X-U)^T \Sigma^{-1} (X-U)]^{-1} (X-U)^T \Sigma^{-1} Y\} \\ &= E\{[(X-U)^T \Sigma^{-1} (X-U)]^{-1} (X-U)^T \\ &\quad \times \Sigma^{-1}[(X-U)\beta + U\tilde{\gamma} + \alpha + \varepsilon]\} \\ &= \beta + E\left\{ [(X-U)^T \Sigma^{-1} (X-U)]^{-1} \right. \\ &\quad \left. \times \sum_i E\left[(X_i - U_i)^T \Sigma_i^{-1} \mathbf{1}_{n_i} \,|\, G_i^\star\right] E\left[\alpha_i \,|\, G_i^\star\right] \right\} \\ &\quad + E\left\{ [(X-U)^T \Sigma^{-1} (X-U)]^{-1} \right. \\ &\quad \left. \times \sum_i E\left[(X_i - U_i)^T \Sigma_i^{-1} \mathbf{1}_{n_i} \,|\, G_i^\star\right] E\left[\varepsilon_{ij} \,|\, G_i^\star\right] \right\}. \end{aligned}$$

The conditional variance is

$$\begin{aligned} \mathrm{var}(\widehat{\beta}' \,|\, G^\star) &= E\left\{ [A^T \Sigma^{-1} A]^{-1} \sum_i A_i^T \Sigma_i^{-1} \left(\alpha_i^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T + \varepsilon_i \varepsilon_i^T\right) \right. \\ &\quad \left. \times \Sigma_i^{-1} A_i [A^T \Sigma^{-1} A]^{-1} \,|\, G^\star \right\} \\ &= E\left\{ [A^T \Sigma^{-1} A]^{-1} \sum_i A_i^T \Sigma_i^{-1} E\left[\alpha_i^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \right.\right. \\ &\quad \left.\left. + \varepsilon_i \varepsilon_i^T \,|\, G_i^*\right] \Sigma_i^{-1} A_i [A^T \Sigma^{-1} A]^{-1} \,|\, G^\star \right\} \\ &= \sum_i E\{[A^T \Sigma^{-1} A]^{-1} A_i^T \Sigma_i^{-1} A_i [A^T \Sigma^{-1} A]^{-1} \,|\, G^\star\} \\ &= E\{[A^T \Sigma^{-1} A]^{-1} \,|\, G^\star\}. \end{aligned}$$

Here the second equality follows from the conditional independence of α_i and G_{ij} given G_i^*, and the third equality follows from $\Sigma_i = \mathrm{var}(Y_i Y_i^T \,|\, G_i^*)$. Taking expectation, we have

$$\mathrm{var}(\hat{\beta}') = E[A^T \Sigma^{-1} A]^{-1}. \tag{A.2}$$

Therefore minimizing the variance of weighted least squares when there are family-specific effects amounts to minimizing $(X-U)^T \Sigma^{-1} (X-U)$.

Since by a linear transformation solving U for the weighted least squares estimator can be converted to solving U for ordinary least squares, we solve the constrained optimization problem for the latter. To minimize the variance (A.1) subject to constraints (4) and (7), we introduce Lagrange equations. Recall that $\vartheta_i^\star$ denotes all possible founder genotypes compatible with that observed in the ith family, and ϑ_i denotes all possible combination of offspring genotypes in the ith family. The object function is

$$\begin{aligned} &\sum_{i=1}^{n} \sum_{g \in \vartheta_i} \sum_{j=1}^{n_i} (X_{ij}(g) - U_{ij}(g))^2 P(g) \\ &\quad - \sum_{i=1}^{n} \sum_{g \in \vartheta_i} \sum_{g^\star \in \vartheta_i^\star} \sum_{j=1}^{n_i} \lambda_{i,g^\star} (X_{ij}(g) - U_{ij}(g)) P(g \,|\, g^\star) \\ &\quad - \eta \sum_{i=1}^{n} \sum_{g \in \vartheta_i} \sum_{j=1}^{n_i} (X_{ij}(g) - U_{ij}(g)) U_{ij}(g) P(g). \end{aligned}$$

Here $\lambda_{i,g^\star}$ and η are Lagrange multipliers. The term $U_{ij}(g)$ is the additional covariate to add when the observed genotypes in the offspring is g.

It is convenient to write the objective function in a matrix form and do the calculation in matrix algebra. Recall the notations for W_i, X_i, and V_i defined in Section 2, and let λ_i denote the $d_i \times 1$ vector of $\lambda_{i,g^\star}$. The objective function can be written as

$$\begin{aligned} &\sum_i \mathbf{1}_{n_i}^T (X_i - V_i) W_i (X_i - V_i)^T \mathbf{1}_{n_i} - \sum_i \lambda_i^T Z_i^T (X_i - V_i)^T \mathbf{1}_{n_i} \\ &\quad - \eta \sum_i \mathbf{1}_{n_i}^T (X_i - V_i) W_i V_i^T \mathbf{1}_{n_i}, \end{aligned}$$

and the Lagrange equations can be written as

$$2W_i(X_i - V_i)^T\mathbf{1}_{n_i} - Z_i\lambda_i - \eta W_i(X_i - 2V_i)^T\mathbf{1}_{n_i} = 0 \tag{A.3}$$

$$Z_i^T(X_i - V_i)^T\mathbf{1}_{n_i} = 0 \tag{A.4}$$

$$\sum_i \mathbf{1}_{n_i}^T(X_i - V_i)W_iV_i^T\mathbf{1}_{n_i} = 0. \tag{A.5}$$

Multiplying both sides of (A.3) on the left by $Z_i^TW_i^{-1}$ and using the condition (A.4) results in

$$\lambda_i = \eta\left(Z_i^TW_i^{-1}Z_i\right)^{-1}Z_i^TX_i^T\mathbf{1}_{n_i}. \tag{A.6}$$

Plug (A.6) into (A.3) to get

$$(2-\eta)W_i(X_i - V_i)^T\mathbf{1}_{n_i} - \eta Z_i\left(Z_i^TW_i^{-1}Z_i\right)^{-1}Z_i^TX_i^T\mathbf{1}_{n_i} + \eta W_iV_i^T\mathbf{1}_{n_i} = 0.$$

Solve this equation to arrive at

$$V_i^T\mathbf{1}_{n_i} = W_i^{-1}Z_i\left(Z_i^TW_i^{-1}Z_i\right)^{-1}Z_i^TX_i^T\mathbf{1}_{n_i}, \tag{A.7}$$

and $\eta = 2$. Here V_i is identifiable up to the sum of its components. Adding and subtracting a constant from the elements of V_i does not change the sum of all the elements. Nevertheless, the expectation of the estimator is the same. We simply pick

$$V_i = X_iZ_i(Z_i^TW_i^{-1}Z_i)^{-1}Z_i^TW_i^{-1}.$$

The desirable additional covariates U_i can be picked from the row of V_i that corresponds to the observed genotypes in family members of the ith family. The marginal probabilities W_i can be computed as $P(g) = \sum_{g^*\in\vartheta_i(g^{**})} P(g \mid g^*)P(g^* \mid g^{**})$. Here g^{**} index the observed founder genotypes in ϑ_i. Note that the solution under the constraints (A.3) and (A.4) satisfies (A.5) automatically.

Finally the computation for an illustrative example pedigree with two children is presented. Suppose that the example pedigree has two children with genotypes (DD, Dd), one parent with observed genotype Dd, and the other parent with no genotype information. The parental genotypes compatible with the observed genotypes are $\vartheta_i^\star = \{(Dd, DD), (Dd, Dd), (DD, dd)\}$. The nine possible genotype configurations for the children are listed as the rows in Table A1. Here $c_i = 9$, and $d_i = 3$. The entries of matrix Z_i are presented in Table A1.

Table A1
The entries of Z_i for the example pedigree

	Parental genotypes		
Offspring genotypes	(Dd, DD)	(Dd, Dd)	(Dd, dd)
(DD, DD)	1/4	1/16	0
(DD, Dd)	1/4	1/8	0
(DD, dd)	0	1/16	0
(Dd, DD)	1/4	1/8	0
(Dd, Dd)	1/4	1/4	1/4
(Dd, dd)	0	1/8	1/4
(dd, DD)	0	1/16	0
(dd, Dd)	0	1/8	1/4
(dd, dd)	0	1/16	1/4

Table A2
The matrix X_i and V_i for the example pedigree

Offspring genotypes	X_i^T	V_i^T		
		$p = 0.1$	$p = 0.2$	$p = 0.4$
(DD, DD)	(2, 2)	(2.04, 2.04)	(1.98, 1.98)	(1.86, 1.86)
(DD, Dd)	(2, 1)	(1.68, 1.68)	(1.66, 1.66)	(1.62, 1.62)
(DD, dd)	(2, 0)	(1.24, 1.24)	(1.18, 1.18)	(1.06, 1.06)
(Dd, DD)	(1, 2)	(1.68, 1.68)	(1.66, 1.66)	(1.62, 1.62)
(Dd, Dd)	(1, 1)	(0.6, 0.6)	(0.7, 0.7)	(0.9, 0.9)
(Dd, dd)	(1, 0)	(0.48, 0.48)	(0.46, 0.46)	(0.42, 0.42)
(dd, DD)	(0, 2)	(1.24, 1.24)	(1.18, 1.18)	(1.06, 1.06)
(dd, Dd)	(0, 1)	(0.48, 0.48)	(0.46, 0.46)	(0.42, 0.42)
(dd, dd)	(0, 0)	(0.44, 0.44)	(0.38, 0.38)	(0.26, 0.26)

Code X_i as the number of D alleles, then from the equation (A.7), the matrix V_i can be calculated. The results under different assumptions of the allele frequency are recorded in Table A2. It can be seen that there is no big difference in V_i when we change the allele frequency.

The observed genotypes for children in the example pedigree is (DD, Dd), which correspond to the second entry in Table A2. Therefore the additional covariate U_i for this family when the allele frequency is 0.1 is (1.68, 1.68). When the allele frequency is 0.2 or 0.4, the additional covariates for this family are (1.66, 1.66) and (1.62, 1.62), respectively. These covariates are not substantially affected by the specification of allele frequency.