Submitted to the Annals of Applied Statistics
arXiv: math.PR/0000000

GENERALIZED GENETIC ASSOCIATION STUDY WITH SAMPLES OF RELATED INDIVIDUALS

By Zeny Feng[*,§], William Wong[†], Xin Gao[‡,§] and Flavio Schenkel[*,§]

University of Guelph[*], University of Toronto[†], and York University[‡]

Genetic association study is an essential step to discover genetic factors that are associated with a complex trait of interest. In this paper, we present a novel generalized quasi-likelihood score (GQLS) test that is suitable for a study with either a quantitative trait or a binary trait. We use a logistic regression model to link the phenotypic value of the trait to the distribution of allelic frequencies. In our model, the allele frequencies are treated as a response and the trait is treated as a covariate that allows us to leave the distribution of the trait values unspecified. Simulation studies indicate that our method is generally more powerful in comparison with family-based association test (FBAT) and controls the type I error at the desired levels. We apply our method to analyze data on Holstein cattle for an estimated breeding value phenotype, and to analyze data from the Collaborative Study of the Genetics of Alcoholism for alcohol dependence. The results show a good portion of significant SNPs and regions consistent with previous reports in the literature, and also reveal new significant SNPs and regions that are associated with the complex trait of interest.

1. Introduction. Recent biological technology allows researchers to perform genome-wide association studies using dense panel of SNPs at an affordable cost. Association studies have been widely used to identify genome regions that are associated with a complex trait of interest. Current methods in genetic association studies can be roughly categorized into two approaches: (1) studies on samples of unrelated subjects; (2) studies on samples of related subjects, from nuclear families, extended families, or from isolated/founder populations which often include inbred individuals that are related through multiple lines of descent.

The classical population-based association test in a case-control study design is the simplest approach where unrelated affected (cases) and unaffected (controls) individuals are typed. However, for a rare disease, it is

[§]Research grant: Natural Sciences and Engineering Research Council of Canada (NSERC) individual discovery grant.

Keywords and phrases: Genetic association test; kinship-inbreeding coefficient; logistic regression; quasi-likelihood

difficult to recruit independent cases in the general population, and more importantly, the naive analysis of data from a general population recruitment design may lead to false positive signals due to confounding effects caused by the population structure. Many researchers (Ewan and Spielman, 1995; Khoury and Yang, 1998; Lander and Schork, 1994) have reported and discussed aspects of this problem. For example, the confounding effect of ethnicity is well known as the population stratification effect in the genetics literature. For an association test with a quantitative trait, a simple linear regression model is often used. As noted, the association tests of quantitative traits via population-based approaches are also subject to the same problem of confounding by the population stratifications.

The family-based association study design using the family based association test (FBAT) analysis method has become popular as this strategy is robust to the population heterogeneity (Horvath et al, 2001; Laird et al., 2000). In FBAT analysis, a statistic U is computed on the basis of the linear combinations of offsprings' genotype and phenotype expression functions. The mean and the variance of U under the null hypothesis of no association is calculated conditional on the parental genotype. Thus, FBAT methods typically require the typing of family members, such as parents or siblings (for inferring a missing parental genotype) of each affected subject to make use of such a subject in the test. This becomes a limitation of the method. For example, for a late onset disease, it is difficult and sometimes impossible to collect the information of the family members of an affected subject. On the other hand, FBAT typically requires heterozygous parents to compute the null distribution of the test statistic. Moreover, when dealing with a large pedigree, FBAT breaks down the pedigree to small nuclear families, such that the relationship among remotely related individuals are ignored. Similarly, FBAT does not take into account for the relationship across related families in the analysis. For these reasons, a family-based approach is generally less powerful in comparison with population-based approaches (Bourgain et al., 2003; Risch and Teng, 1998; Thornton and McPeek, 2007).

Slager and Schaid (2001) have proposed a method that was based on the Armitage trend test with the inclusion of a variance that accounts for the relationships among individuals from an outbred population. However, this method cannot handle large, complex, inbred pedigrees. A different approach, a pedigree disequilibrium test, proposed by Martin et al. (2001) can be employed to handle large pedigree association analysis. A founder/isolated population-based study design has been suggested (Lander and Schork, 1994; Wright et al., 1999) for association mapping. This study design efficiently controls the confounding effect due to population structure

and has been useful for complex trait mapping. Recently, Bourgain et al. (2003) proposed a case-control association test where subjects are sampled from a founder population with known genealogy. They adapted the idea of a population-based association test to test whether the allele frequencies of a specified allele are equal between the case group and control group, taking into account the correlations among subjects and the inbreeding configuration within subjects. This method can be used to analyze data from a large inbred pedigree and is also suitable for data from multiple pedigrees with careful control of ethnic homogeneity (Thornton and McPeek, 2007). The test is based on a quasi-likelihood scoring (QLS) approach and has been shown to be more powerful than the traditional transmission/disequilibrium test (TDT) when samples are from homogeneous populations. However, these approaches are limited to binary traits.

Following the line of quasi-likelihood approach proposed by Bourgain et al. (2003) and Thornton and McPeek (2007) to handle the correlation structure among related subjects, we propose a generalized linear model framework to accommodate other types of traits. We use a logistic regression model to link the trait to the distribution of allelic frequencies. In our model, the observed trait of each individual is treated as a covariate. The proportion of a specified allele in the genotype is the response. In conventional models, the phenotypic trait is treated as the response and the distribution of the trait values needed to be specified. For example, the normality assumption is often required for a quantitative trait. In our method, the trait is treated as explanatory variable, which allows us to leave the distribution unspecified. On the other hand, treating the allele frequencies of the marker as the response, we have the exact covariance structure for the responses with the provision of the pedigree structure or the documented genealogy. Under this innovative modeling, we derive the test statistic (W_G) and show that W_G asymptotically follows a χ^2_{k-1} distribution, where k is the number of alleles of the marker. Our proposed GQLS test generalizes the existing approaches in three aspects: 1) the GQLS method can establish associations between marker's allele frequencies and all types of traits; 2) it uses a general link function to connect the mean value of the allele frequency with the traits; 3) our GQLS method can be extended to solve the problem when a sample is collected from multiple sub-populations. In this article, we focus on the logistic link, but the extension of our test to other link functions, e.g., the probit function, would be straightforward.

This paper is motivated by the challenges of analyzing data on Holstein cattle in North America. The aim of this study is to identify SNPs or genome regions that are associated with the estimated breeding values (EBVs) of a

proven bull. The EBV of a bull predicts its genetic merit. For example, the milk yield EBV of a bull predicts the milk yield of its female descendants. Conducting an association study in this data set is challenging. First, dams are not typed, and sires are typed only if they appear as proven bulls in the data set. Thus, FBAT is not applicable to analyze this data set. Second, most of the bulls, sires and dams, are inbred. They are descendants from a single complex pedigree and the relationships among them are known but complicated. The conventional population-based association test does not account for this complex relationship among subjects. Ignoring the correlation structure among subjects would lead to an inflated positive result. This will be shown by simulation studies in the paper. Third, the case-control founder-population-based approach proposed by Bourgain et al. (2003) is limited to binary traits where most of the EBVs are quantitative. Thus, the challenge of analyzing this data set becomes a motivation for the development of our method.

We perform simulation studies on collections of pedigrees of various sizes and on single complex pedigrees with different sizes to validate our method. We compare the empirical performance of our method with others. In application, we also apply our method to the Collaborative Study of the Genetics of Alcoholism (COGA) data provided by Genetic Analysis Workshop (GAW) 14 (Bailey-Wilson et al., 2005; Edenberg et al., 2005) to demonstrate the application in the binary trait and multiple small families study design.

The paper is organized as follows. Section 2 presents the proposed generalized quasi-likelihood association test. Section 3 presents the details of simulation studies to assess the validity and the power of the proposed test compared with other methods. In Section 4, applications to real data are provided to illustrate the practical application of the proposed method. Discussions are provided in Section 5.

2. Methods.

2.1. *Association test with a biallelic marker.* Suppose that in a genetic study we have a sample of n subjects that is from a single isolated/founder population or a single pedigree. Subjects may be arbitrarily related with a known relationship. It is assumed that the inbreeding configuration for each subject is also known. Let $\boldsymbol{X} = (X_1, ...X_n)'$ with X_i being the phenotypic observation of the ith subject. The X_i can be binary with $X_i = 1$ or 0 coding for "affected" or "unaffected" respectively, or can be continuous for a quantitative trait. Given a biallelic marker of interest, alleles are labelled by "0" and "1". Let $\mathbf{Y} = (Y_1, ..., Y_n)'$ with $Y_i = \frac{1}{2} \times$ (the number of allele 1 in subject i) being the proportion of the allele 1 in the observed genotype of subject i,

and $Y_i = 0, \frac{1}{2}$, or 1. Let $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)' = \mathrm{E}(\mathbf{Y}|\boldsymbol{X})$ that $0 < \mu_i < 1$. We propose a logistic regression model to link the expected allele frequency $\boldsymbol{\mu}$ of the marker with the trait $\boldsymbol{X}$. We let

$$\mu_i = E(Y_i|X_i) = \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}. \tag{2.1}$$

To test the association between the marker and the trait, we test

$$H_0 : \beta_1 = 0 \qquad \text{against} \qquad H_a : \beta_1 \neq 0.$$

Our model provides a natural constraint that $0 < \mu_i < 1$ for all $i = 1, ..., n$. Under the null hypothesis, we have $\mu_i = \mu = \frac{e^{\beta_0}}{1+e^{\beta_0}}$ for all $i = 1, ..., n$. The mean vector of $\mathbf{Y}$ no longer depends on X_i and becomes $\boldsymbol{\mu} = \mathrm{E}(\mathbf{Y}) = \mu\mathbf{1}$ where $\mathbf{1}$ is a n-vector of 1's. It can be shown that, under H_0, the covariance matrix of $\mathbf{Y}$ is given by $\boldsymbol{\Sigma_0} = \frac{1}{2}\mu(1-\mu)\boldsymbol{\rho}$, and

$$\boldsymbol{\rho} = \begin{pmatrix} 1+\phi_1 & 2\phi_{12} & \cdots & 2\phi_{1n} \\ 2\phi_{12} & 1+\phi_2 & \cdots & 2\phi_{2n} \\ \vdots & \cdots & \ddots & \vdots \\ 2\phi_{1n} & 2\phi_{2n} & \cdots & 1+\phi_n \end{pmatrix}, \tag{2.2}$$

where ϕ_i is the inbreeding coefficient of individual i and ϕ_{ij} is the kinship coefficient between individual i and individual j. See Appendix A in the supplementary material for the justification. The covariance matrix $\boldsymbol{\Sigma_0}$ will be invertible if $\mu \neq 1$ or 0, and $\boldsymbol{\rho}$ is invertible provided that the monozygous twins (twins that are genetically identical as they originate from a single fertilized egg) are merged and represented by one single individual. This can be done using the multiple outputation procedure (Follmann et al, 2003). The quasi-likelihood score function is in the form of

$$S(\boldsymbol{\beta}) = (S_{\beta_0}(\boldsymbol{\beta}), S_{\beta_1}(\boldsymbol{\beta}))' = D'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \boldsymbol{\mu}), \tag{2.3}$$

where D is a $n \times 2$ derivative matrix in the form of

$$D = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} = \left(\frac{\partial \boldsymbol{\mu}}{\partial \beta_0}, \frac{\partial \boldsymbol{\mu}}{\partial \beta_1}\right), \tag{2.4}$$

and Σ is the covariance matrix of $\mathbf{Y}$. Under the null hypothesis, we have $\boldsymbol{\mu} = \mu\mathbf{1}$ and the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$. The solution to the equation of the quasi-likelihood score function $S_{\beta_0}(\beta_0, 0) = 0$ gives an estimate of μ as

$$\hat{\mu} = (\mathbf{1}'\boldsymbol{\rho}^{-1}\mathbf{1})^{-1}\mathbf{1}'\boldsymbol{\rho}^{-1}\mathbf{Y}, \tag{2.5}$$

and therefore gives the estimate of β_0 as $\hat{\beta}_0 = \log \frac{\hat{\mu}}{1-\hat{\mu}}$ under the null hypothesis. See Appendix B in the supplementary material for the derivation.

When $\beta_1 \neq 0$, the marker is associated with the trait and the expected value of Y_i given the X_i is given by Equation (2.1). For a binary trait, the two-sample model of Bourgain et al. (2003) in the form of

$$\mu_i = \begin{cases} p + r, & \text{if } i \text{ is affected, with } 0 < p + r < 1, \\ p, & \text{if } i \text{ is unaffected, with } 0 < p < 1, \end{cases}$$

becomes a special case of our model that $p = \frac{e^{\beta_0}}{1+e^{\beta_0}}$ and $r = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} - \frac{e^{\beta_0}}{1+e^{\beta_0}}$. We propose a generalized quasi-likelihood scoring statistic to test the association between the marker and the trait. Under the null hypothesis that $\beta_1 = 0$,

$$\mathrm{E}[S_{\beta_1}(\beta_0, \beta_1 = 0)] = \mathrm{E}\left[\frac{\partial \boldsymbol{\mu}}{\partial \beta_1} \boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \boldsymbol{\mu})\right] = 0.$$

As described by Cox and Hinkley (1974), the quasi-score statistic is given by

$$W = S_{\beta_1}(\hat{\beta}_0, 0)' \mathrm{var}_0^{-1}(S_{\beta_1}(\hat{\beta}_0, 0)) S_{\beta_1}(\hat{\beta}_0, 0), \tag{2.6}$$

where $\hat{\beta}_0$ is the quasi-likelihood estimate of β_0 and $\mathrm{var}_0^{-1}(S_{\beta_1}(\hat{\beta}_0, 0))$ is the $(2, 2)$th entry of the inverse of the information matrix $\mathbf{I}(\boldsymbol{\beta})$ that are computed under the null hypothesis that $\beta_1 = 0$. As demonstrated by Heyde (1997), under the null hypothesis, W follows a χ^2 distribution with 1 degree of freedom asymptotically. In our case, we obtain an explicit expression for our generalized quasi-likelihood scoring statistic in the form of

$$\begin{aligned} W_{\mathrm{G}} \quad = \quad & \tfrac{2}{\hat{\mu}(1-\hat{\mu})}[\boldsymbol{X}'\boldsymbol{\rho}^{-1}(\mathbf{Y} - \hat{\mu}\mathbf{1})]' \\ & [\boldsymbol{X}'\boldsymbol{\rho}^{-1}\boldsymbol{X} - (\boldsymbol{X}'\boldsymbol{\rho}^{-1}\mathbf{1})(\mathbf{1}'\boldsymbol{\rho}^{-1}\mathbf{1})^{-1}(\mathbf{1}'\boldsymbol{\rho}^{-1}\boldsymbol{X})]^{-1} \\ & [\boldsymbol{X}'\boldsymbol{\rho}^{-1}(\mathbf{Y} - \hat{\mu}\mathbf{1})], \end{aligned} \tag{2.7}$$

where $\hat{\mu}$ is given by Equation (2.5). See Appendix B in supplementary material for the derivation. Note that, in Equation (2.7), we do not need $\hat{\beta}_0$ to compute the W_G statistic. W_G is expressed in a general form for both the quantitative and binary traits. When the trait is binary, the quasi-likelihood scoring statistic proposed by Bourgain et al. (2003) becomes a special case of our W_G that they are the same. Under the null hypothesis, W_G follows a χ_1^2 distribution asymptotically.

Following the same line as in Bourgain et al. (2003), we generalize the W_G statistic to accommodate F independent families in an outbred population. Among n subjects, let n_f be the number of subjects that are from family

f and let $\mathbf{Y}_f = (Y_{1f}, \ldots, Y_{n_f f})'$ be the vector of Y's for subjects that are from family f, $f = 1, ..., F$. Then, we have $n = n_1 + ... + n_F$. Let $\mathbf{\Sigma}_f$ and $\boldsymbol{\rho}_f$ be the covariance and correlation matrix of Y's for those subjects that are from the fth family. If all the individuals in the sample are outbred, the diagonal entries of matrix $\boldsymbol{\rho}_f$ are equal to 1 for all $f = 1, ..., F$. The overall covariance matrix under the null hypothesis is a block diagonal matrix that consists of $\mathbf{\Sigma}_1, ..., \mathbf{\Sigma}_F$. We derive that explicit form for the quasi-likelihood estimate of μ under the null hypothesis as

$$\hat{\mu} = (\sum_{f=1}^{F} \mathbf{1}_f' \boldsymbol{\rho}_f^{-1} \mathbf{1}_f)^{-1} (\sum_{f=1}^{F} \mathbf{1}_f' \boldsymbol{\rho}_f^{-1} \mathbf{Y}_f), \tag{2.8}$$

where $\mathbf{1}_f$ is the n_f-vector of 1's. We derive an explicit form that

$$W_G = \frac{2}{\hat{\mu}(1-\hat{\mu})} A' B^{-1} A \tag{2.9}$$

where

$$A = \sum_{f=1}^{F} [\boldsymbol{X}_f' \boldsymbol{\rho}_f^{-1} (\mathbf{Y}_f - \hat{\mu} \mathbf{1}_f)],$$

$$B = \sum_{f=1}^{F} \boldsymbol{X}_f' \boldsymbol{\rho}_f^{-1} \boldsymbol{X}_f - (\sum_{f=1}^{F} \boldsymbol{X}_f' \boldsymbol{\rho}_f^{-1} \mathbf{1}_f)^2 (\sum_{f=1}^{F} \mathbf{1}_f' \boldsymbol{\rho}_f^{-1} \mathbf{1}_f)^{-1},$$

ans $\boldsymbol{X}_f$ is the n_f-vector of the traits of the individuals from the fth family.

2.2. *Association test with a multiallelic marker.* Now, suppose the marker under investigation has k different alleles and there are n individuals being sampled from a single pedigree. Let $\mathbf{Y} = (\mathbf{Y}_1', ..., \mathbf{Y}_{k-1}')'$ be an $n(k-1)$-vector with $\mathbf{Y}_j = (Y_{j1}, ..., Y_{jn})'$ being an $n-$vector that $Y_{ji} = \frac{1}{2} \times$(the number of allele j in individual i). Similarly to the biallelic case, we let $\boldsymbol{\mu} = \mathrm{E}(\mathbf{Y}|\boldsymbol{X}) = (\boldsymbol{\mu}_1', ..., \boldsymbol{\mu}_{k-1}')'$ with $\boldsymbol{\mu}_j = (\mu_{j1}, ..., \mu_{jn})'$ and

$$\mu_{ji} = \frac{e^{\beta_{0j} + \beta_{1j} X_i}}{1 + \sum_{j=1}^{k-1} e^{\beta_{0j} + \beta_{1j} X_i}}.$$

Each random vector $2 \times (Y_{1i}, ..., Y_{k-1,i})'$ follows a multinomial$(2, (\mu_{1i}, ..., \mu_{k-1,i})')$ distribution with $0 < \mu_{ji} < 1$ and $\sum_{j=1}^{k} \mu_{ji} = 1$ for all $i = 1, ..., n$. Under the null hypothesis that the marker is not associated with the trait, all β_{1j}'s are 0. Thus, we perform a simultaneous hypothesis test that

$$H_0 : \beta_{11} = ... = \beta_{1,k-1} = 0 \quad \text{vs} \quad H_a : \text{at least one } \beta_{1j} \neq 0, j = 1, ..., k-1.$$

Here, we generalize the notation of vector $\boldsymbol{\beta}$ as in the biallelic case that $\boldsymbol{\beta} = (\boldsymbol{\beta}_0', \boldsymbol{\beta}_1')'$ with $\boldsymbol{\beta}_0 = (\beta_{01}, ..., \beta_{0,k-1})'$ and $\boldsymbol{\beta}_1 = (\beta_{11}, ..., \beta_{1,k-1})'$. Under the null hypothesis that $\boldsymbol{\beta}_1 = \mathbf{0}$, we have $\mu_{ji} = \mu_j$ for all i and rewrite the mean vector $\boldsymbol{\mu} = (\mu_1 \mathbf{1}', ..., \mu_{k-1}\mathbf{1}')'$ where $\mathbf{1}$ is an n-vector of 1's. Under the null hypothesis, the covariance matrix of $\mathbf{Y}$ is given by $\boldsymbol{\Sigma} = \mathbf{F} \otimes \boldsymbol{\rho}$ (the Kronecker product of matrices $\mathbf{F}$ and $\boldsymbol{\rho}$) where $\mathbf{F}$ is a $(k-1)\times(k-1)$ matrix, which is the same as in Bourgain et al. (2003). Here, let $\boldsymbol{\mu}^* = (\mu_1, ..., \mu_{k-1})$ be the $(k-1)$-vector such that $\boldsymbol{\mu} = \boldsymbol{\mu}^* \otimes \mathbf{1}$ under the null hypothesis. We show that, under the null hypothesis, the quasi-likelihood estimate of $\boldsymbol{\mu}^*$ is given by

$$\hat{\boldsymbol{\mu}^*} = (\hat{\mu}_1, ..., \hat{\mu}_{k-1})' = (\mathbf{1}'\boldsymbol{\rho}^{-1}\mathbf{1})^{-1}(\mathbf{I}_{k-1} \otimes (\mathbf{1}'\boldsymbol{\rho}^{-1}))\mathbf{Y} \tag{2.10}$$

where $\mathbf{I}_{k-1}$ is a $(k-1) \times (k-1)$ identity matrix. Thus, $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}^*} \otimes \mathbf{1}$. We obtain an explicit form of the generalized quasi-likelihood scoring statistic as

$$W_G = C \cdot (\mathbf{Y} - \hat{\boldsymbol{\mu}})'(\hat{\mathbf{F}}^{-1} \otimes (\boldsymbol{\rho}^{-1}\boldsymbol{X}\boldsymbol{X}'\boldsymbol{\rho}^{-1}))(\mathbf{Y} - \hat{\boldsymbol{\mu}}). \tag{2.11}$$

where $C = [\boldsymbol{X}'\boldsymbol{\rho}^{-1}\boldsymbol{X} - \boldsymbol{X}'\boldsymbol{\rho}^{-1}\mathbf{1}(\mathbf{1}'\boldsymbol{\rho}^{-1}\mathbf{1})^{-1}(\mathbf{1}'\boldsymbol{\rho}^{-1}\boldsymbol{X})]^{-1}$ is a constant depending on the trait vector $\boldsymbol{X}$ and the correlation matrix $\boldsymbol{\rho}$, and $\hat{\mathbf{F}}$ is computed by using the $\hat{\boldsymbol{\mu}^*}$. See Appendix C in the supplementary material for derivations of $\hat{\boldsymbol{\mu}^*}$ and W_G in the multiallelic case. Under the null hypothesis, W_G follows a χ^2 distribution with $k-1$ degrees of freedom asymptotically. Alternatively, we can express the statistic in the form

$$W_G = C \sum_{j=1}^{k-1}\sum_{l=1}^{k-1}(\hat{\mathbf{F}}^{-1})_{jl}(\mathbf{Y}_j - \hat{\mu}_j\mathbf{1})'\boldsymbol{\rho}^{-1}\boldsymbol{X}\boldsymbol{X}'\boldsymbol{\rho}^{-1}(\mathbf{Y}_l - \hat{\mu}_l\mathbf{1}). \tag{2.12}$$

In the biallelic case that $k = 2$, we have $\mathbf{F} = \frac{1}{2}\mu(1-\mu)$ and $\boldsymbol{\Sigma} = \frac{1}{2}\mu(1-\mu)\boldsymbol{\rho}$, $\hat{\boldsymbol{\mu}^*}$ and W_G reduce to those that are derived under the biallelic case. When the n individuals in the sample comprise subjects that are from F independent families, we retain the notation of $\boldsymbol{X}_f, \mathbf{1}_f$ and $\boldsymbol{\rho}_f$ as in the biallelic case. Let $\mathbf{Y}_f = (\mathbf{Y}_{1f}', ..., \mathbf{Y}_{k-1,f}')'$ and $\mathbf{Y}_{jf} = (Y_{j1}, ..., Y_{jn_f})'$. The statistic W_G is given by

$$W_G = C \cdot \sum_{j=1}^{k-1}\sum_{l=1}^{k-1}(\hat{\mathbf{F}}^{-1})_{jl}\left\{\sum_{f=1}^{F}(\mathbf{Y}_{jf} - \hat{\mu}_j\mathbf{1}_f)'\boldsymbol{\rho}_f^{-1}\boldsymbol{X}_f\sum_{f=1}^{F}(\mathbf{Y}_{lf} - \hat{\mu}_l\mathbf{1}_f)'\boldsymbol{\rho}_f^{-1}\boldsymbol{X}_f\right\}, \tag{2.13}$$

where $C = \left\{\sum_{f=1}^{F}\boldsymbol{X}_f'\boldsymbol{\rho}_f^{-1}\boldsymbol{X}_f - (\sum_{f=1}^{F}\boldsymbol{X}_f'\boldsymbol{\rho}_f^{-1}\mathbf{1}_f)^2(\sum_{f=1}^{F}\mathbf{1}_f'\boldsymbol{\rho}_f^{-1}\mathbf{1}_f)^{-1}\right\}^{-1}$. Under the null hypothesis, W_G follows a χ^2_{k-1} distribution asymptotically.

2.3. *Data collected from multiple sub-populations.* In this paper, we extend our GQLS method to a solution that overcome the problem of population stratification. Suppose a sample is collected from S different sub-populations , denoted by $pop_1, ..., pop_S$. For illustration, let the marker of interest be bi-allelic (e.g., a SNP). For each sub-population, pop_s, we compute a GQLS test statistic, $W_G^{(s)}$. We know that the $W_G^{(s)}$ follows χ_1^2 distribution asymptotically. In statistical theory, the sum of S independent χ^2 random variables follows a χ^2 distribution with the degrees of freedom being the sum of the S degrees of freedom. Thus, a new overall statistic, which is the sum over all sub-populations, having the form as

$$W_{all} = W_G^{(1)} + W_G^{(2)} + ... + W_G^{(S)}$$

follows a χ_S^2 distribution asymptotically under the null hypothesis.

It is well-known that FBAT is robust to the analysis of family data collected from different populations. We will compare the performance of our overall test method with FBAT in the population stratification problem via simulation studies. We will also apply this overall test method to the COGA data set. See Section 3.3 and Section 4.2 for details.

3. Simulation study. We conduct simulation studies to validate the χ^2 distribution approximation to the distribution of the W_G statistic and to compare the power achieved by our approach with the power achieved by the FBAT. We consider three different study designs. First, we simulate single large complex pedigrees. Second, we simulate multiple small families. Third, for each study design, we combine samples simulated under settings to mimic a sample collected from different sub-populations to investigate the robustness of our extended method using the W_{all} statistic. Since SNPs are popular for genetic association studies and SNPs are typically biallelic, we simulate biallelic markers for demonstration. We use the software KinInbcoef (Bourgain, 2003) to compute the kinship-inbreeding coefficient correlation matrix $\boldsymbol{\rho}$. We will describe the simulation procedures and summarize the results for each design in the following three subsections.

3.1. *Single large pedigree study design.* In this study design, a family is grown starting from a single individual. Each single individual is assigned a spouse with probability 0.8 or remains single with probability 0.2. For each couple, we generate the number of offspring according to a Poisson distribution with mean 3. Any pedigree that stops growing before the completion of six generations by natural degeneration, or stops before reaching to a desired family size, is disregarded. A new pedigree is grown until we obtain one single pedigree that consists of six generations and has a desirable

number of family members in the last three generations. In our simulation study, we generate three large single outbred pedigrees that have sizes of 136, 273, and 557, respectively. Family members of the top three generations are removed to mimic the practical situations (especially, in human data) in which clinical information and DNA samples are most likely not available for more than three generations back. The genealogy of the entire pedigree remains for calculating the correlation matrix $\boldsymbol{\rho}$. Removing the family members from the top three generations, the pedigree sizes reduce to 124, 251, and 526, respectively. For each founder (an individual with parents' genetic information unknown), the marker genotype is simulated by random mating. The genotypes of descents are generated according to the Mendelian law of segregation.

To assess the type I error rate, for each individual, traits are generated genetically according to a SNP with the minor allele frequency (MAF) of the SNP being set to 0.3. Denote the genotype of the SNP by G that $G = 0, 1$, or 2 for having 0, 1, or 2 allele 1 in the genotype. We simulated the quantitative trait, X, from $N(-1 + G, \sigma^2)$ with $\sigma = 1.2$. The binary trait was simulated from Bernoulli(p_G) with $p_0 = 0.1, p_1 = 0.3, p_2 = 0.4$. Then, a SNP that is unlinked to the causal SNP is generated. The minor allele frequency of the SNP is set to 0.3 and 0.1. For each combination of settings, we generate 1000 replicates. For each simulated data set, we compute the W_G statistic for the unlinked SNP, and take the rejection threshold to be the $(1 - \alpha)$th quantile of the χ^2_1 distribution. We run FBAT on each simulated data set. In FBAT, default options are chosen in most of the cases except that the "minsize" (the minimum number of informative families) is set to 4. To illustrate the preservation of the type I error by considering the correlation among related subjects, we perform the standard Armitage trend test (Armitage, 1955) that assumes independent subjects in the sample. The Armitage trend test was implemented using the "independence_test" function in the R package "coin" (R Development Core Team, 2009). This function also allows testing on the quantitative trait. We consider $\alpha = 0.05$ and 0.01. In Table 1, we summarize the empirical rejection rates at each significance level for each combination of settings. The simulation results indicate that the χ^2_1 distribution approximates the distribution of the W_G statistic well. The inflation of the null empirical rejection rate using the trend test is obvious (indicated in bolded numbers) in the single large pedigree study.

To compare the power with the FBAT method, we simulate the quantitative trait and the binary trait conditioning on the genotype of each individual. The minor allele frequency of the association marker is set to 0.3

and 0.1. Three different genetic models are considered for both the quantitative and binary trait. The quantitative trait X is generated according to an additive model: $X_i = a + bG_i + \varepsilon_i$, where

$$G_i = \begin{cases} -1 & \text{for the homozygous genotype that } Y_i = 0, \\ 0 & \text{for the heterozygous genotype that } Y_i = 1/2, \\ 1 & \text{for the homozygous genotype that } Y_i = 1. \end{cases}$$

The random environmental errors ε_i, are generated from $N(0, \sigma^2)$. Without loss of generality, we set the intercept $a = 0$. We specify three different association models: (1) $b = 0.5, \sigma = 1.2$; (2) $b = 1, \sigma = 1.5$; and (3) $b = 1, \sigma = 1.2$. The coefficient b quantifies the effect of the marker. The different values of σ^2 pose different levels of difficulty for the detection of genetic association. These three models are denoted by qt1, qt2, and qt3 respectively in the tables that summarize the results of power assessments.

For the binary trait, we generate the affection status of individuals according to three disease models. In model 1, we consider a recessive epistasis disease controlled by two SNPs that are unlinked to each other. Individuals having two copies of allele 1 at both SNPs have a penetrance (defined as $f = \mathrm{P}(\text{affected}|\text{genotype})$) of $f_1 = 0.5$. Individuals having two copies of allele 1 at one SNP but not at the other SNP have a penetrance of $f_2 = 0.4$. Individuals with fewer than two copies of allele 1 at both SNPs have a penetrance of $f_3 = 0.1$. In model 2, we consider a dominant epistasis disease controlled by two SNPs that are unlinked to each other. Individuals with at least one copy of allele 1 at both SNPs have a penetrance of $f_1 = 0.5$. All other individuals have a penetrance of $f_2 = 0.1$. In model 3, we consider a single disease locus model with $f_1 = 0.5$ if an individual has two allele 1's at the SNP, $f_2 = 0.3$ if an individual has one allele 1 at the SNP, and $f_3 = 0.1$ otherwise. These three models are denoted by bt1, bt2 and bt3 respectively in the tables that summarize the results of power assessments.

For each combination of settings, we generate 1000 replicates. For each simulated data set, we compute the W_G and obtain the p-value by the χ^2_1 approximation. We run FBAT on each simulated data set. The proportions of p-values $\leq \alpha$ are reported in Table 2. Simulation results show that our method outperforms the FBAT for a higher detection power in all scenarios. Results are particularly striking for the binary trait with small sample size.

We extend our simulation studies to a single pedigree that consists of nine generations. Genotypes and clinical information of family members in

TABLE 1
Type I error assessment-single large pedigree study design (6 generations)

			Sample size								
			124			251			526		
MAF	α[1]	trait	GQLS	FBAT	Trend	GQLS	FBAT	Trend	GQLS	FBAT	Trend
0.3	0.05	bt[2]	0.049	0.051	**0.086**	0.045	0.049	**0.103**	0.049	0.051	0.069
		qt[3]	0.048	0.046	**0.152**	0.057	0.048	**0.11**	0.053	0.051	**0.107**
	0.01	bt	0.006	0.007	**0.019**	0.009	0.009	**0.033**	0.011	0.011	0.015
		qt	0.015	0.011	**0.069**	0.009	0.008	**0.03**	0.013	0.01	**0.034**
0.1	0.05	bt	0.052	0.076	0.048	0.047	0.038	0.053	0.048	0.048	0.065
		qt	0.045	0.057	**0.086**	0.054	0.051	0.059	0.054	0.044	0.062
	0.01	bt	0.013	0.004	0.009	0.009	0.006	0.007	0.013	0.012	**0.018**
		qt	0.014	0.008	0.016	0.012	0.012	0.017	0.011	0.009	0.008

[1]Monte Carlo standard deviation = 0.0069 or 0.0031 for $\alpha = 0.05$ or 0.01 respectively.
[2]bt: binary trait. [3]qt: quantitative trait.

the top six generations are removed. The genealogy of the entire pedigree remains for calculating the correlation matrix $\boldsymbol{\rho}$. We generate two single large pedigrees having sizes of 704 and 875, respectively. After removing the family members in the top six generations, there are 615 and 795 individuals remaining. Similarly, we set the MAF of 0.3 and 0.1. The results of type I error and power assessments are summarized in the Table 1 and 2 in the supplementary material. The simulation results are consistent to the results of the studies with six generations. The empirical type I error rates obtained by our method and the FBAT are close to each of the nominal significance level. The trend test generally inflates the empirical rejection rate under the null hypothesis (indicated in bolded numbers). Our method is generally more powerful than the FBAT.

3.2. *Multiple families study design.* In this study, families are grown following the similar procedure as for the single large family study design except that families will grow for a maximum of three generations. The simulated sample comprises families and independent individuals. Family sizes range from 1 to 23 with an average size of 6.3. As in the single large pedigree study design, the genotype of founders are generated by random mating and the genotype of non-founders are generated according to the Mendelian law of segregation. We let the sample size (number of subjects) be 100, 200 and 500, respectively. To assess the type I error rate, we generate a quantitative trait and a binary trait for each individual as described in the single large family study design. Then, a SNP that is unlinked to the causal SNP is generated. The minor allele frequency of the SNPs is set to 0.3 and 0.1. For each combination of settings, we generate 1000 replicates. In Table 3, we

TABLE 2
Power comparison - single large family study design (6 generations)

			Sample size					
			124		251		526	
MAF	trait	α	GQLS	FBAT	GQLS	FBAT	GQLS	FBAT
0.3	bt1	0.05	0.228	0.186	0.368	0.277	0.603	0.466
		0.01	0.106	0.06	0.186	0.106	0.386	0.257
	bt2	0.05	0.421	0.264	0.663	0.451	0.928	0.764
		0.01	0.218	0.089	0.444	0.237	0.791	0.548
	bt3	0.05	0.610	0.421	0.91	0.705	0.997	0.96
		0.01	0.385	0.181	0.758	0.439	0.868	0.851
0.1	bt 1	0.05	0.098	0.063	0.103	0.076	0.135	0.099
		0.01	0.03	0.01	0.03	0.016	0.053	0.027
	bt2	0.05	0.164	0.116	0.211	0.132	0.294	0.187
		0.01	0.063	0.029	0.088	0.036	0.145	0.77
	bt3	0.05	0.452	0.33	0.624	0.424	0.911	0.721
		0.01	0.263	0.119	0.442	0.167	0.804	0.463
0.3	qt1	0.05	0.468	0.416	0.781	0.734	0.977	0.96
		0.01	0.242	0.19	0.572	0.491	0.905	0.869
	qt2	0.05	0.831	0.791	0.991	0.970	1	1
		0.01	0.656	0.558	0.946	0.897	0.999	0.999
	qt3	0.05	0.943	0.909	0.999	0.994	1	1
		0.01	0.836	0.758	0.995	0.981	1	1
0.1	qt1	0.05	0.261	0.228	0.401	0.364	0.707	0.650
		0.01	0.126	0.074	0.214	0.173	0.483	0.403
	qt2	0.05	0.511	0.451	0.730	0.674	0.968	0.944
		0.01	0.326	0.23	0.54	0.428	0.916	0.842
	qt3	0.05	0.658	0.602	0.868	0.82	0.994	0.979
		0.01	0.469	0.349	0.736	0.642	0.978	0.922

summarize the null empirical rejection rates. The results indicate that the χ_1^2 distribution approximates the distribution of the W_G statistic well. The inflation of the null empirical rejection rate using the trend test is observed. For power comparisons, we simulate the quantitative traits and binary traits according to the six models that have been described in the previous section. The MAF of the association marker is also set to 0.3 and 0.1. The powers achieved by our method and the FBAT under each combination of settings are summarized in Table 4. Simulation results show that our method consistently outperforms FBAT for all scenarios.

3.3. *Data with sub-populations.* In this section, we consider the situation that a sample contains individuals from different populations. Similarly to the previous section, we consider biallelic markers. For illustration, we con-

TABLE 3
Type I error assessment-multiple families study design

			Sample size								
			100			200			500		
MAF	α	trait	GQLS	FBAT	Trend	GQLS	FBAT	Trend	GQLS	FBAT	Trend
0.3	0.05	bt	0.055	0.037	**0.1**	0.048	0.053	**0.09**	0.056	0.051	0.057
		qt	0.056	0.058	0.0654	0.052	0.049	0.07	0.055	0.054	0.064
	0.01	bt	0.012	0.005	**0.025**	0.012	0.010	0.018	0.012	0.013	0.012
		qt	0.013	0.010	0.009	0.013	0.008	**0.020**	0.011	0.011	0.015
0.1	0.05	bt	0.054	0.037	0.059	0.050	0.045	0.0068	0.05	0.05	0.065
		qt	0.048	0.043	**0.082**	0.047	0.048	**0.088**	0.043	0.055	0.070
	0.01	bt	0.015	0.006	0.011	0.01	0.009	0.013	0.007	0.006	0.013
		qt	0.013	0.007	**0.022**	0.007	0.006	**0.031**	0.007	0.006	0.013

sider a sample collected from two sub-populations only. In fact, for each of the previous study designs: the single large pedigree and the multiple small pedigrees, we combine two simulated data set with different MAF to make up a sample that consists individuals from two different populations. For example, in the single large pedigree study design, we combined the two simulated samples from two sub-populations with MAF being set to 0.1 and 0.3, and with different combinations of sample sizes for each sub-population. For each combined sample, the W_{all} is the sum of the two W_G statistics from two sub-samples. The p-values is obtained by the χ^2_2 distribution. The type I error rate and the power are compared between our method and FBAT.

In the supplementary material, Table 3, we summarize the results of type I error rates assessment by combining two single large pedigrees. Similarly, in the supplementary material, Table 4, we summarize the results of type I error assessment by combining the two simulated samples of multiple small pedigrees. Overall, the empirical type I error rates obtained by our method using the W_{all} test statistics and the empirical type I error rates obtained by FBAT are close to each of the nominal significance levels. However, FBAT is slightly less stable. For example, in Table 3, the empirical type I error rate is 0.005 at 0.01 significance level for a quantitative trait when combining the sample size of 124 from population 1 and sample size of 526 from the population 2. Table 4, the empirical error rate is 0.033 at at 0.05 significance level for a binary trait when combining the sample sizes of 100 from both population 1 and population 2. Both the 95% confidence intervals constructed based on these two empirical type I error rate do not cover the true values of $\alpha = 0.01$ and 0.05.

In the supplementary material, Table 5 and 6, we summarize the results of power assessment. The simulation results indicated that the performance of our method and FBAT are comparable that one shows some advantages

TABLE 4
Power comparison - multiple small pedigree study design

			Sample size					
			124		251		526	
MAF	trait	α	GQLS	FBAT	GQLS	FBAT	GQLS	FBAT
0.3	bt 1	0.05	0.22	0.17	0.302	0.171	0.610	0.388
		0.01	0.093	0.04	0.135	0.07	0.402	0.215
	bt2	0.05	0.354	0.178	0.561	0.317	0.93	0.675
		0.01	0.170	0.044	0.339	0.11	0.818	0.42
	bt3	0.05	0.639	0.328	0.829	0.514	0.996	0.917
		0.01	0.38	0.13	0.643	0.255	0.982	0.78
0.1	bt1	0.05	0.095	0.063	0.095	0.082	0.118	0.073
		0.01	0.029	0.008	0.041	0.01	0.032	0.011
	bt2	0.05	0.132	0.078	0.183	0.107	0.322	0.16
		0.01	0.046	0.011	0.078	0.024	0.148	0.06
	bt3	0.05	0.118	0.073	0.322	0.16	0.942	0.634
		0.01	0.032	0.011	0.148	0.06	0.843	0.361
0.3	qt1	0.05	0.433	0.309	0.709	0.546	0.98	0.928
		0.01	0.217	0.114	0.478	0.302	0.921	0.777
	qt2	0.05	0.795	0.624	0.969	0.849	1	1
		0.01	0.58	0.369	0.913	0.704	1	1
	qt3	0.05	0.934	0.792	0.999	0.976	1	1
		0.01	0.817	0.552	0.99	0.691	1	1
0.1	qt1	0.05	0.215	0.153	0.351	0.264	0.746	0.597
		0.01	0.079	0.046	0.157	0.091	0.520	0.331
	qt2	0.05	0.456	0.279	0.734	0.502	0.986	0.726
		0.01	0.228	0.09	0.515	0.251	0.943	0.796
	qt3	0.05	0.647	0.391	0.895	0.625	0.99	0.977
		0.01	0.419	0.149	0.745	0.4	0.992	0.916

over the other under some suituations, and *vice versa*.

4. Real data analysis.

4.1. *Application to Holstein cattle data.* The data set contains 821 progeny-tested proven bulls born between 1965 and 2001. Each bull was genotyped using the Affymetrix MegAllele GeneChip Bovine mapping 10K SNP array (Affymetrix Inc., 2005). Among 821 bulls, some bulls also appear as the sires of other bulls. The relationships among bulls and their sires and dams are complicated. All the 821 bulls sampled have genetically contributed to the current Canadian cow population. Most of the animals in the population have non-zero inbreeding coefficient. A genealogy of the population tracing back 25 generations with the oldest animal born in 1909, was used to compute the kinship-inbreeding coefficient with the software CFC (Sargolzaei et al., 2006). Out of 9,919 genotyped SNPs, only 8,624 SNPs have known location on the 29 Bos Taurus autosome chromosomes (BTA). SNPs with

more than 20% of missing values or MAF of less than 5% were excluded from the study. A total of 7103 SNPs were analyzed. The experimental design is mainly a grand-daughter design that the milk productivities of daughters and granddaughters of a bull are used to estimate the breeding value of the bull. The phenotypes used in the analysis were trait EBVs released in November 2008 and provided by the Canadian Dairy Network (CDN, Guelph, Canada). For illustration, we only present results of the association tests with milk yield EBV.

In Table 5, we report the top 81 most significant SNPs that have p-value$\leq$ 0.001 that can be grouped into 36 regions (SNPs at a close inter-distance, less than 1cM, define a region) on 16 BTAs. Out of 36 significant SNPs or regions, 16 significant SNPs or regions have been found in agreement with the quantitative traits loci or associated SNPs reported in the literature. In BTA14, 22 SNPs concentrated in 0-27cM have strong association with milk yield and their p-values range from 6.45×10^{-10} to 0.001. At the telomere of BTA14, Daicylglycerol acyl transferase 1 (*DGAT1*) at 0cM has been considered to be a quantitative trait nucleotide with a major effect on milk yield (Bennewitz et al., 2003; Boichard et al., 2003; Grisart et al., 2004). A SNP at 0.27cM has a strong association signal. Twelve SNPs in the region of 3.38-8.47cM are consistent with 3 SNPs at 4cM, 5cM, and 6cM that have been reported significantly associated with milk yield by Daetwyler et al. (2007) and Bennewitz et al. (2003). A SNP at 11.2cM also confirms the association with milk yield reported by Daetwyler et al. (2007). The most significant SNP is found at 94cM on BTA5 and confirms a QTL at the same location reported by Viitala et al. (2003). A significant SNP at 98cM also confirms a QTL at the same location reported by Viitala et al. (2003). Note that, after adjusting for Bonferroni's correction at 5% significance level (or, at 7.13×10^{-6} individual significant level), 11 regions remain significant. However, for many complex traits that are controlled by several genes, each individual gene may only have a small effect. When thousands of SNPs are tested, using the Bonferronis correction may result in low power of the study. Therefore, when we interpret the Bonferroni result, we need to be careful that some signals disappear after the adjustment may be due to the conservativeness of the Bonferroni's correction.

4.2. *Application to COGA data.* The Collaborative Study on the Genetics of Alcoholism (COGA) dataset was provided by the Genetic Analysis Workshop 14 (GAW14). The dataset included 1614 individuals from 143 families. Among 1614 individuals, 1,351 individuals were genotyped for a

panel of 11,555 SNPs from Affymetrix. A set of alcoholism phenotypes and covariates were provided. We use the ALDX1 as the phenotype. Individuals who are coded as "affected" in the ALDX1 variable are considered as affected individuals. Unaffected individuals are those coded as "pure" unaffected in the ALDX1. Individuals with other codings are considered to have unknown phenotypes. In this study, we compare our method with FBAT under three scenarios. In scenario 1, we consider a large sample from a single population. We only include individuals who are coded as "white, non-Hispanic". There are 119 such families consisting of 1074 individuals. In scenario 2, we consider a small sample from a single population. We only include individuals who are coded as "white, Hispanic". There are 11 such families consisting of 78 individuals. In scenario 3, we combine the two samples from the the two populations of "white, Hispanic" and "white, non-Hispanic". In our studies, we use the software KinInbcoef to compute the kinship coefficient for correlation matrix $\boldsymbol{\rho}$. We only analyze SNPs that are on autosomes. In total, there are 10532 SNPs on autosomes.

The results based on our method are summarized in the supplementary material, Table 7 . In total, there are 22 SNPs found to be significant (p-values$<$ 0.001) in the "white, hispanic" sample, 19 SNPs are found to be significant based on the "white, non-Hispanic" sample, and 24 SNPs are found to be significant based on the pooled samples of "white". There are 19 SNPs that are significant in both the pooled sample and the "white, hispanic" or in both the pooled sample and the "white, non-Hispanic" sample. On chromosome 2, SNP tsc0052826 is significant in both the "white hispanic" sample and the pooled sample, it is 0.344cM from a marker that had been reported for a significant linkage with alcohol dependence (Hill et al., 2004; Valdes, McWeeney, and Thomson, 1999). On chromosome 6, SNP tsc1395926 is significant in both the "white hispanic" sample and the pooled sample. It is very close to two loci (less than 1Mb) that had been found to link to the alcoholism (Hill et al., 2004; Ma et al., 2005). On Chromosome 7, SNPs tsc0333356 is significant in both the "white hispanic" sample and the pooled sample, it is 1.47cM away from a marker that had been reported to significantly link to ALDX1 by Zhu et al. (2005) and is 0.811cM from a marker shown significant linkage to alcohol dependence by Hill et al. (2004). The most significant SNP is SNP tsc0059716 on chromosome 13 (p-value=4×10^{-6}), which is about 2.4cM away from a SNP that had been reported to significantly associate with ALDX1 (Zhu et al., 2005). In total, there are 12 SNPs are found to be very close to regions or SNPs that had been reported to link or associate with alcohol dependence or alcoholism related traits in the literature. After adjusting for Bonferroni's correction

at 5% significance level (or, at 4.75×10^{-6} individual significant level), four SNPsSNPs (tsc0587314 on chromosome 3, tsc0506913 on chromosome 5, tsc0630829 on chromosome 7, and tsc0059716 on chromosome 13) remain significant.

The results based on FBAT are summarized in Table 8 in the supplementary material. In total, there are 43 SNPs are found to be significant (p-value$<$ 0.001) in the pooled sample, 29 SNPs are significant in the "white, non-hispanic" sample, and only one SNP is significant in the "white, hispanic" sample. Among these significant SNPs, SNP tsc0056748 on chromosome 13 is significant in more that one sample (the pooled sample and the "white, non-Hispanic" sample). There are 17 significant SNPs in the pooled sample that had been reported significantly associated with the ALDX1 by Zhu et al. (2005). Note that, the results in Zhu et al. (2005) are based on the same pooled sample of "white, hispanic" and "white, non-hispanic", same definition of "affected" individual, and are analyzed by the FBAT as well. The only difference is the definition of "unaffected" individual that we only use "pure-affected" individuals while Zhu et al. (2005) use "pure-unaffected" and "never drank". Therefore, there would be more significant SNPs confirmed by Zhu et al. (2005). In addition, SNP tsc0046578 on chromosome 1 is 1.37cM away from a SNP that had been reported to significantly link to alcohol dependence by Prescott et al. (2006). SNP tsc0697701 on chromosome 8 is 0.7Mb away from a SNP that significantly link to the alcoholism by Hill et al. (2004). SNP tsc0896393 on chromosome 12 is 1.5Mb away from a SNP that significantly link to ALDX1 reported by Ma et al. (2005). After adjusting for the bonferroni correction, three SNPs (tsc0515272 on chromosome 3, tsc0029429 on chromosome 9, and tsc 1750530 on chromosome 16) remain significant.

5. Discussion. In this article, we adopt the framework of the generalized linear model and assume that the expected marker allelic frequency is connected to the linear predictor based on the trait of interest through an arbitrary specified link function. Although we focus on the logistic link, which is the canonical link for a binomial random variable, models utilizing other link functions can be built with minor modifications of the approach herein. The population-based association study is still a popular study design for common traits. To prevent spurious association due to a confounding population structure, association studies should be performed within a relative homogeneous population. Such a population-based association study is a special case of our method in which the $\boldsymbol{\rho}$ matrix will be an identity matrix for independent subjects. For stratified population, Lander and Schork

(2006) suggested using "internal controls" to balance the ethnicity between the cases and controls in the sample in order to eliminate the confounding effects. Our proposed generalized association method uses all available family members to provide natural "internal controls". Conneally (2003) pointed out that for any choice of study design, whether based on families or population-based, a large sample size is needed to detect an associated gene with only a partial effect on the trait. The quasi-likelihood scoring method fully utilizes the correlation information among the sampled individuals. It accommodates various data types for genetic association studies including the conventional population-based association studies, and those using founder/isolated populations with documented genealogy, or multiple complex pedigrees. Thus, this method essentially increases the sample size and becomes more powerful. On the other hand, when data set contains samples from multiple sub-populations, we propose a solution that combines the W_G statistics from each sub-population to construct a new test statistic W_{all}. The W_{all} statistic is founded to follow a χ^2 distribution asymptotically with the degrees of freedom depends on the number of sub-populations and the number of alleles of the marker being tested. Simulation results confirm that the χ^2 distribution approximates the distribution of W_{all} well. Simulation results also shown that our method has comparable power to the FBAT. However, our approach is limited to known subpopulations. If unknown subpopulations exist, it is possible to extend our approach to a mixture population with more population parameters to be estimated.

It is known that pedigrees error can easily arise in the study of large pedigree and even in the study of small pedigrees. Our GQLS method cannot handle this error directly. However, many methods and software are available to detect such error under different study designs (PREST by McPeek and Sun (2000); RELATIVE by Göring and Ott (1997); RELPAIR by Epstein, Duren, and Boehnke (2000)). When the pedigree errors are found, involved individuals could be either removed from the study accordingly, or, the relationship, i.e., the kinship and inbreeding coefficients, among involved individuals can be inferred through the genome scan (if genome data are available) as a substitute in the $\boldsymbol{\rho}$ matrix. However, the approximation of the χ^2 distribution to the resulting W_G statistic is needed to be further investigated.

Acknowledgements. The authors thank Professor Mary Thompson (Department of Statistics and Actuarial Science, University of Waterloo) for a critical reading of the original version of this paper. The COGA Data were provided by the Collaborative Study on the Genetics of Alcoholism

(U10AA008401), GAW grant, R01 GM031575.

Software. The R code for computing the GQLS test statistic is available at http://www.uoguelph.ca/∼zfeng/software/.

References.

Affymetrix MeAllele GeneChip Bovine 10K SNP array. Affymetrix Inc., South San Francisco, CA, USA Online. Available at http://www.affymetrix.com/support/technical/datasheets/bovine10k_snp_datasheet.pdf Accessed on December, 2009.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, **11**(3),375-386

Bailey-Wilson, J. E., Almasy, L., Andrade, M., Bailey, J., Bickeböller, H., Cordell, H. J., Daw, E. W., Goldin, L., Goode, E. L., Gray-McGuire, C., Hening, W., Jarvik, G., Maher, B. S., Mendell, N., Paterson, A. D., Rice, J., Satten, G., Suarez, B., Vieland, V., Wilcox, M., Zhang, H., Ziegler, A., and MacCluer, J. W. (2005). Genetic analysis workshop 14: microsattellite and single-nucleotide polymorphism marker loci for genome-wide scans. *BMC Genetics* **6**(Suppl I), S1

Bennewitz, J., Reinsch, N., Grohs, C., Levziel, H., Malafosse, A., Thomsen, H., Xu, N., Looft, C., Khn, C., Brockmann, G. A., Schwerin, M., Weimann, C., Hiendleder, S., Erhardt, G., Medjugorac, I., Russ, I., Frster, M., Brenig, B., Reinhardt, F., Reents, R., Averdunk, G., Blmel, J., Boichard, D., and Kalm, E. (2003). Combined analysis of data from two granddaughter designs: A simple strategy for QTL confirmation and increasing experimental power in dairy cattle. *Genetics Selection Evolution* **35**,319-338

Boichard, D., Grohs, C., Bourgeois, F., Cerqueira, F., Faugeras, R., Neau, A., Rupp, R., Amigues, Y., Boscher, M. Y., and Levziel, H. (2003). Detection of genes influencing economic traits in three French dairy cattle breeds. *Genetics Selection Evolution* **35**, 77-101

Bourgain, C. (2003). KinInbcoef: Calculation of kinship and inbreeding coefficients. Available at http://www.stat.uchicago.edu/ mcpeek/software/KinInbcoef/index.html. Accessed on December, 2009.

Bourgain, C., Hoffjan, S., Nicolae, R., Newman, D., Steiner, L., Walker, K., Reynolds, R., Ober, C., and McPeek, M. S.(2003). Novel case-control test in a founder population identifies P-seletin as an Antopy-susceptibility locus. *American Journal of Human Genetics* **73**,612–626.

Conneally, P. M. (2003). 2002 ASHG presidential address: the complexity of complex diseases. *American Journal of Human Genetics* **72**,228-232

Cox, D. R. and Hinkley, D. V. (1974). Theoretical statistics. *Chapman and Hall,* London

Daetwyler, H. D., Schenkel, F. S., Sargolzaei, M., and Robinson, J. A. B. (2007). A Genome Scan to Detect Quantitative Trait Loci for Economically Important Traits in Holstein Cattle Using Two Methods and a Dense Single Nucleotide Polymorphism Map. *Journal of Dairy Science* **91**,3225-3236

Dick, D. M. and Bierut, L. J. (2006) The genetics of alcohol dependence. *Current Psychiatry Reports* **8**,151-157.

Edenberg, H. J., Bierut, L. J., Boyce, P., Cao, M., Cawley, S., Chiles R., Doheny, K. F., Hansen, M., Hinrichs, T., Jones, K., Kennedy, G. C., Liu, G., Marcus, G., McBride, C., Murray, S. S., Oliphant, O., Pettengill, J., Porjesc, B., Pugh, E. W., Rice, J. P., Rubano, T., Shannon, S., Steeke, R., Tischfield, J. A., Tsai, Y. Y., Zhang, C., and Begleiter, H.

(2005). Description of the data from the Collaborative Study on the Genetics of Alcoholism (COGA) and single-nucleotide polymorphism genotyping for Genetic Analysis Workshop 14. *BMC Genetics*, **6**(Suppl I), S2

Epstein, M.P., Duren, W.L., and Boehnke, M. (2000). Improved inference of relationships for pairs of individuals. *American Journal of Human Genetics* **67**,1219-1231.

Ewans, W. J. and Spielman, R. S. (2003). The transmission/disequilibrium test: history, subdivision, and admixture. *American Journal of Human Genetics*, **57**,455-464

Follmann, D., Proschan, M., Leifer, E. (2003). Multiple outputation: inference for complex clustered data by averaging analyses from independent data. *Biometrics*, **59**, 420-429

Göring, H. H. and Ott, J. (1997). Relationship estimation in affected sib pair analysis of late-onset diseases. *European Journal of Human Geneics* **5**, 69-77.

Grisart, B., Farnir, F., Karim, L., Cambisano, N., Kim, J., Kvasz, A., Mni, M., Simon, P., Frre, J. M., Coppieters, W., and Georges, M. (2004). Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(8), 2398-2403

Heyde, C. (1997). Quasi-likelihood and its application: a general approach to optimal parameter estimation. *Spring,* New York

Heyen, D. W., Weller, J. I., Ron, M., Band, M., Beever, J. E., Feldmesser, E., Da, Y., Wiggans, G. R., VanRaden, P. M., and Lewin. H. A. (1999). A genome scan for QTL influencing milk production and health traits in dairy cattle. *Physiological Genomics* **1**,165-175

Hill, S. Y., Shen, S., Zezza, N., Hoffman, E. K., Perlin, M., and Allan, W. (2004). A genome wide search for alcoholism susceptibility genes. *American Journal of Medical Genetics Part B (Neuropsychiatric Genetics)*, **128B**, 102-113

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

Khoury, M. J. and Yang, Q. (1998). The future of genetic studies of complex human diseases: an epidemiologic perspective. *Epidemiology* **9**,350-354.

Laird, N. M., Horvath, S., Xu, X. (2000). Implementing a unified approach to family-based tests of association. *Genetics Epidemiology* **19**(Suppl 1), S36-S42

Lander, E. S. and Schork, N. J. (1994). Genetic dissection of complex traits: guideline for interpreting and reporting linkage results. *Nature Genetics* **11**,2037-2048.

Lander, E. S. and Schork N. J. (2006). Genetic dissection of complex traits. *The Journal of Lifelong Learning in Psychiatry* **4**,442-458

Ma, Q., Yu, Y., Meng, Y., Farrell, J., Farrer, L. A., and Wilcox, M. A. (2005). Genome-wide linkage analysis for a alcohol dependence: a comparison between single-nucleotide polymorphism and microsatellite marker assays. *BMC Genetics*, **6**, (suppl 1):S8

Martin, E. R., Bass, M. P., Kaplan, N. L. (2001). Correcting for a potential bias in the pedigree disequilibrium test. *American Journal of Human Genetics* **68**, 1065-1067

McPeek, M. S. and Sun, L. (2000). Statistical tests for detection of misspecified relationships by use of genome-screen data. *American Journal of Human Genetics* **66**, 1076-1094.

Prescott, C. A., Sullivan, P. F., Kuo, P. H., Webb, B. T., Vittum, J., Patterson, D. G., Thiselton, D. L., Myer, J. M., Devitt, M., Halberstadt, L. J., Robinson, V. P., Neale, M. C., van den Oord, E. J., Walsh, D., Riley, B. P., and Kendler, K. S. (2006). Genomewide linkage study in the Irish affected sib pair study of alcohol dependence: evidence for a susceptibility region for symptoms of alcohol dependence on chromosome 4. *Molecular Psychiatry*, **11**, 603-611

R Development Core Team (2009). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria. URL: http://www.R-project.org

Risch, N. and Teng, J. (1998). The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Research* **8**,1273-1288

Sargolzaei, M., Iwaisaki, H., and Colleau, J. J. (2006) CFC: A tool for monitoring genetic diversity. *8th World Congress of Genetics Applied to Livestock Production, CD-ROM Communication 27-28*. Belo Horizonte, Brazil, Aug. 13-18, 2006

Slager, L. and Schaid, D. (2001). Evaluation of candidate genes in case-control studies: a statistical method to account for related subjects. *American Journal of Human Genetics* **68**, 1457-1462

Spielman, R., McGinnis, R., and Ewens, W. (1993). Transmission test for linkage disequilibrium and association. *American Journal of Human Genetics* **52**,506-516

Thornton, T. and McPeek, M. S. (2007). Case-control association testing with related individuals: a more powerful quasi-likelihood score test. *American Journal of Human Genetics* **81**,321-337

Valdes, A. M., McWeeney, S. K., amd Thomson, G. (1999). Evidence for linkage and association to alcohol dependence on chromosome 19. *Genetics Epidemiology*, **17**(suppl 1): S367-372

Viitala S. (2008). Identification of genes controlling milk production in dairy cattle. PhD thesis, MTT Agrifood Research Finland, University of Turku, Finland.

Viitala, S. M., Schulman, N. F., Koning, D. J., Elo, K., Kinos, R., Virta, A., Virta, J., Mki-Tanila, A., and Vilkki, J. H. (2003). Quantitative Trait Loci Affecting Milk Production Traits in Finnish Ayrshire Dairy Cattle *Journal of Dairy Science* **86**,1828-1836

Wright, A. F., Carothers, A. D., and Pirastu, M. (1999). Population choices in mapping for complex diseases. *Nature Genetics* **23**, 387-404.

Zhu, X., Cooper, R., Kan, D., Cao, G., and Wu, X. (2005). A genome-wide linkage and association study using COGA data. *BMC Genetics*, **6**, (Suppl 1):S128

Dept of Mathematics and Statistics,
University of Guelph,
Guelph, Ontario N1G2W1, Canada
E-mail: zfeng@uoguelph.ca

Toronto Health Economics and
Technology Assessment Collaborative
University of Toronto,
Toronto, ON M5S3M2, Canada
E-mail: wwl.wong@utoronto.ca

Dept of Matematics and Statistics,
York University,
North York, Ontario M3J1P3, Canada
E-mail: xingao@mathstat.yorku.ca

Dept of Animal and Poultry Science,
University of Guelph,
Ontario N1G2W1, Canada
E-mail: schenkel@uguelph.ca

TABLE 5
Most significant loci (p-value ≤0.001) found for milk yield trait

BTA	No. of SNPs	Location (cM)[1]	p-value [2]
1	1	47.90 [5]	2.18×10^{-5}
4	1	20.05 [5]	0.000105
	2	56.65, 59.81 [5]	0.000664
	1	101.74	0.000126
5	1	1.03 [5]	0.00086
	1	8.32	2.91×10^{-5}
	3	29.59-34.46	7.5×10^{-5}
	6	45.51-50.53	4.65×10^{-6}*
	1	69.89	8.8×10^{-6}
	8	73.49-77.77	8.44×10^{-7}*
	12	90.76-101.06 [6,8,9]	3.14×10^{-11}*
	1	114.90 [3]	0.000125
6	1	47.66 [7]	0.000355
7	1	75.07 [5]	0.001
8	1	41.75	0.000215
	1	55	0.000126
11	1	113.46	0.000853
12	1	61.77 [5]	5.81×10^{-7}*
14	1	0.27 [3,4,5,6]	2.06×10^{-6}*
	12	3.38-8.47 [3,5]	3.86×10^{-8}*
	3	11.2 [4,5]	4.94×10^{-6}*
	4	21.50 [4]	6.45×10^{-10}*
	2	26.69	0.000691
15	1	21	0.000291
16	1	31.66	3.39×10^{-8}
	1	54	0.000364
	1	62	0.000946
	1	90.54 [4]	3.85×10^{-5}
17	1	16	0.000595
	1	72	0.00034
	1	78.58	0.000868
18	1	15.78 [4]	1.75×10^{-6}*
23	2	9.36	4.44×10^{-6}*
26	2	44, 45	0.000341
	1	53 [3,4]	0.00061
27	1	57	0.000962

[1]Chromosomal region that the SNPs span on
[2]Minimum p-value if there are more that one SNPs in the region
[3]In agreement with Bennewitz et al. (2003) [4]In agreement with Boichard et al. (2003)
[5]In agreement with Daetwyler et al. (2007) [6]In agreement with Grisart et al. (2004)
[7]In agreement with Heyen et al. (1999) [8]In agreement with Viitala et al. (2003)
[9]In agreement with Viitala (2008)
*Significant at 5% Bonferroni's correction (at 7.13×10^{-6} individual significance level)