On a Unifying 'Reverse' Regression for Robust Association Studies and Allele Frequency Estimation with Related Individuals

Lin Zhang[1] and Lei Sun[1,2]

[1]Department of Statistical Sciences, University of Toronto, 100 St. George Street, Toronto, Ontario M5S 3G3, Canada, linzhang@utstat.toronto.edu

[2]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, 155 College Street, Toronto, Ontario M5T 3M7, Canada, sun@utstat.toronto.edu

June 4, 2019

Abstract

For genetic association studies with related individuals, standard linear mixed-effect model is the most popular approach. The model treats a complex trait (phenotype) as the response variable while a genetic variant (genotype) as a covariate. An alternative approach is to reverse the roles of phenotype and genotype. This class of tests includes quasi-likelihood based score tests. In this work, after reviewing these existing methods, we propose a general, unifying 'reverse' regression framework. We then show that the proposed method can also explicitly adjust for potential departure from Hardy–Weinberg equilibrium. Lastly, we demonstrate the additional flexibility of the proposed model on allele frequency estimation, as well as its connection with earlier work of best linear unbiased allele-frequency estimator. We conclude the paper with supporting evidence from simulation and application studies.

Keywords: Robust association tests; Quasi-likelihood score tests; Dependent samples; Hardy–Weinberg equilibrium; Allele frequency estimation.

1 Introduction

Genetic association studies aim at identifying genetic variants, Gs, that influence a heritable trait, Y, of interest. To this end, allele-based association tests or allelic tests, comparing allele frequencies between case and control groups, are locally most powerful (Sasieni, 1997) in a sample of unrelated individuals. However, traditional allelic tests (i) analyze only binary outcomes, (ii) cannot easily accommodate covariates, (iii) are limited to independent samples, and (iv) have type I error control issue if there is a departure from Hardy–Weinberg equilibrium (HWE) in the study population.

HWE states that the two alleles in a genotype are independent draws from the same distribution, or, equivalently, genotype frequencies depend solely on the allele frequencies. For a bi-allelic SNP with two possible alleles A and a, let p and $1-p$ be the respective allele frequencies. Under HWE, $p_{aa} = (1-p)^2$, $p_{Aa} = 2p(1-p)$, and $p_{AA} = p^2$, where p_{aa}, p_{Aa}, and p_{AA} are the genotype frequencies of genotypes aa, Aa and AA, respectively. To measure the departure from HWE or the amount of Hardy–Weinberg disequilibrium (HWD),

$$\delta = p_{AA} - p^2 \tag{1}$$

is a widely used quantity, and $\delta = 0$ corresponds to HWE (Weir, 1996).

Genotype-based association tests treat phenotype Y as the response variable and genotype G as an explanatory variable. Due to the regression nature of the framework, genotype-based association tests can easily handle continuous traits and incorporate covariates. It is commonly assumed that genotype-based association tests are robust to departure from HWE. With a sample of independent individuals, both theoretical and empirical results support this (Sasieni, 1997; Schaid and Jacobsen, 1999). However, in the presence of sample dependency, little has been discussed.

When individuals in a sample are genetically related with each other, linear mixed-effect models (LMM) have become the most popular approach for association testing (Eu-Ahsunthornwattana et al., 2014). The variance-covariance matrix of the phenotype is partitioned into a weighted sum

of correlation structure due to genetic relatedness and shared environmental effects, where the weight is usually referred to as 'heritability' (Visscher et al., 2008). The genetic relatedness is typically represented by a known kinship coefficient matrix, or estimated based on the available genome-wide genetic data if the pedigree information was not collected (Yang et al., 2011; Sun and Dimitromanolakis, 2012).

An alternative approach is to reverse the roles of Y and G in the regression model. O'Reilly et al. (2012) proposed MultiPhen, a method that treats the genotype G of a SNP as the response variable and phenotype values Ys of multiple traits as predictors. However, MultiPhen relies on ordinal logistic regression and analyzes only independent samples. MultiPhen does not require the assumption of HWE, but insights to its robustness to HWD was not provided (O'Reilly et al., 2012).

Thornton and McPeek (2007) extended the traditional allelic tests to study binary traits with related individuals. Their test was then generalized by Feng et al. (2011) and Feng (2014) to a quasi-likelihood score test for either binary or continuous traits. However, none of these methods can directly incorporate covariates. Jakobsdottir and McPeek (2013) later proposed a 'retrospective' approach, MASTOR, to study the association between G and one (approximately) normally distributed trait Y, while accommodating covariates in related individuals. All methods in this category, however, implicitly assumed HWE.

In this paper, we first review and provide some insights into the aforementioned genetic association tests. We then propose a robust and flexible 'reverse' regression framework that (a) unifies several existing association methods, and (b) explicitly includes a correction factor in the variance-covariance matrix to adjust for potential departure from HWE. Further, we show that the proposed 'reverse' regression framework (c) can also be used to estimate allele frequency in complex pedigrees. Interestingly, we reveal that for the simple case of no covariates and HWE, the proposed estimator is the best linear unbiased estimator of McPeek et al. (2004). We conclude the paper with supporting evidence from simulation and application studies, and some discussion points.

2 Method

2.1 The Traditional Allele-based Association Test, or Allelic Test, T_{allelic}

Consider a bi-allelic SNP with two alleles a and A. Without loss of generality, A is the minor allele with population minor allele frequency (MAF) less than 0.5. Consider a case-control study with independent observations. We use r_k, for $k \in \{0,1,2\}$, to denote the genotype counts, respectively, for genotypes aa, Aa and AA, in the case group of size r. Similarly, s_k for the control group of size s, and n_k for the combined sample of size n.

Let $\hat{p}_r = (2r_2 + r_1)/2r$, $\hat{p}_s = (2s_2 + s_1)/2s$, and $\hat{p} = (2n_2 + n_1)/2n$ be the sample allele frequencies of allele A, respectively, in the case, control and combined samples. The classical allelic association test is based on,

$$T_{\text{allelic}} = \frac{(\hat{p}_r - \hat{p}_s)^2}{(\frac{1}{2r} + \frac{1}{2s})\hat{p}(1-\hat{p})}.$$

Under the null of no association, $T_{\text{allelic}} \sim \chi^2_1$.

It has been shown that T_{allelic} is locally most powerful for a sample of unrelated individuals (Sasieni, 1997). However, the validity of T_{allelic} requires the assumption of Hardy–Weinberg equilibrium. Some remedies have been proposed. For example, Schaid and Jacobsen (1999) considered a variance adjustment and recommended $(\hat{p}_r - \hat{p}_s)^2/(\frac{1}{2r} + \frac{1}{2s})(\hat{p}(1-\hat{p}) + \hat{\delta})$ as a robust allelic test, where

$$\hat{\delta} = \hat{p}_{AA} - \hat{p}^2,\ \hat{p}_{AA} = n_2/n,\ \hat{p} = (2n_2 + n_1)/2n,$$

and $\hat{p}$ and $\hat{p}_{AA}$ are the sample frequency estimates of allele A and genotype AA, respectively. Nevertheless, existing (classical and robust) allelic tests are limited to binary Y without consideration of covariate effects.

2.2 The Traditional Phenotype-on-Genotype (Y-on-G) Association Tests, T_{indep} and T_{LMM}

2.2.1 Independent Samples, T_{indep}

Define $G = 0$, 1 and 2 for genotypes aa, Aa and AA, respectively, and let Y be a (continuous) trait of interest. With a sample of n unrelated individuals, the traditional genotype-based association test assumes that,

$$y = \alpha^* 1 + \beta^* g + \gamma^* z + \varepsilon^*, \quad \varepsilon^* \sim N(0, \sigma^{*2} I), \tag{2}$$

where $y = (y_1, y_2, \cdots, y_n)$ is a $n \times 1$ vector for the phenotypic values, $g = (g_1, g_2, \cdots, g_n)$ is a $n \times 1$ vector for the genotypes of the SNP, ε^* is the error term with variance σ^{*2}, 1 is a $n \times 1$ vector of 1s, and I is the identity matrix. For notation simplicity but without loss of generality, we assume that there is only one additional covariate, denoted by z.

Score tests are often used for genetic association analyses (Derkach et al., 2015). In this case, the score statistic of testing $H_0 : \beta^* = 0$ can be easily derived as

$$T_{\text{indep}} = n \cdot \frac{\{(g-\bar{g}1)^T(y-\bar{y}1) - \frac{(g-\bar{g}1)^T(z-\bar{z}1)(y-\bar{y}1)^T(z-\bar{z}1)}{(z-\bar{z}1)^T(z-\bar{z}1)}\}^2}{[(g-\bar{g}1)^T(g-\bar{g}1) - \frac{\{(g-\bar{g}1)^T(z-\bar{z}1)\}^2}{(z-\bar{z}1)^T(z-\bar{z}1)}][(y-\bar{y}1)^T(y-\bar{y}1) - \frac{\{(y-\bar{y}1)^T(z-\bar{z}1)\}^2}{(z-\bar{z}1)^T(z-\bar{z}1)}]}. \tag{3}$$

After some simple algebraic manipulations, one can show that

$$\frac{1}{n}(g-\bar{g}1)^T(g-\bar{g}1) = \widehat{\text{var}}(G) = 2(\hat{p}(1-\hat{p}) + \hat{p}_{AA} - \hat{p}^2) = 2(\hat{p}(1-\hat{p}) + \hat{\delta}).$$

Because $\hat{\delta} = \hat{p}_{AA} - \hat{p}^2$ measures the amount of Hardy–Weinberg disequilibrium present in the data (Weir (1996)), T_{indep} inherently adjusts for departure from HWE through $\widehat{\text{var}}(G) = 2(\hat{p}(1-\hat{p}) + \hat{\delta})$. As a result, the traditional genotype-based association test is robust to HWD in independent samples.

When Y is binary, logistic regression is commonly used. However, Chen (1983) showed that

under common regularity conditions, the score test statistic takes identical form for exponential family in independent samples with no covariates. (Derkach et al. (2015) also showed that for Y-dependent sampling, "the score statistics are identical for conditional and full likelihood approaches, and are of the same form as for ordinary random sampling.") Thus, in terms of association testing, we can conclude that genotype-based association studies of binary traits in independent samples are also robust to departure from HWE.

2.2.2 Dependent Samples, $T_{\mathbf{LMM}}$

When individuals in a sample are related to each other as in pedigree data, a common practice is to replace $\text{var}(\varepsilon^*) = \sigma^{*2}I$ in (2) with $\sigma_y^2\Sigma_y$ to reflect the sample dependence, and use the linear mixed-effect model,

$$y = \alpha^* 1 + \beta^* g + \gamma^* z + \varepsilon^*, \quad \varepsilon^* \sim N(0, \sigma_y^2\Sigma_y), \quad \Sigma_y = h^2\Sigma_\Phi + (1-h^2)I, \tag{4}$$

where h^2 is the 'heritability', and Σ_Φ is the kinship coefficient matrix. Among the total variance of Y, σ_y^2, $\sigma_a^2 = h^2\sigma_y^2$ can be interpreted as the variance of Y due to additive genetic variation, while $\sigma_e^2 = (1-h^2)\sigma_y^2$ as the variance due to environmental variation.

For the Σ_Φ matrix, $\Sigma_\Phi(i,j) = 2\phi_{i,j}$ where $\phi_{i,j}$ is the kinship coefficient between individual i and individual j. When the pedigree information is not available, it is a common practice to estimate $\Sigma_\Phi(i,j)$ with the averaged sample correlation across K autosomal SNPs (Yang et al., 2011),

$$\hat{\Sigma}_\Phi(i,j) = \frac{1}{K}\sum_{k=1}^{K} \frac{(g_{ik}-2\hat{p}_k)(g_{jk}-2\hat{p}_k)}{2\hat{p}_k(1-\hat{p}_k)}, \tag{5}$$

where g_{ik} and g_{jk} are the genotypes of SNP k for individual i and j respectively, and $\hat{p}_k$ is the estimated allele frequency.

Without the assumption of HWE, $\text{var}(G_k) = 2(p_k(1-p_k)+\delta_k)$. Thus, (5) implies $\delta_k = 0$, and the traditional kinship coefficient estimation implicitly assumes that HWE holds at each and every

SNP k, $k = 1, \cdots, K$. Consequently, T_{LMM} derived from the linear mixed model (4) can be sensitive to departure from HWE. In Section 3, we will demonstrate with a simple sib-pair design that when the true heritability is known, the empirical type I error rate of the standard linear mixed-effect model is inflated when $\delta > 0$ (or deflated if $\delta < 0$), and when heritability is estimated from the data, the test is accurate but the estimated heritability is then biased when $\delta \neq 0$.

2.3 The Proposed 'Reverse' Genotype-on-Phenotype (G-on-Y) Regression Model and Association Test, T_{reverse}

To account for potential departure from Hardy–Weinberg equilibrium yet adjusting for covariate effect and sample dependency, one useful direction is to reverse the roles of Y and G in the regression model. We define the 'reverse' regression model as,

$$g = \alpha 1 + \beta y + \gamma z + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 \Sigma_g), \quad \sigma^2 \Sigma_g = \sigma^2 \Sigma_\Phi + \Sigma_\delta, \tag{6}$$

where Σ_Φ is the kinship coefficient matrix as defined earlier, and Σ_δ is a function of δ, that explicitly models the amount of Hardy–Weinberg disequilibrium. For example, for a pair of siblings,

$$\sigma^2 \Sigma_g = \sigma^2 \Sigma_\Phi + \Sigma_\delta = \sigma^2 \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} + \begin{pmatrix} 0 & 0.5\delta \\ 0.5\delta & 0 \end{pmatrix}. \tag{7}$$

Under the null hypothesis of $H_0 : \beta = 0$, the corresponding score test statistic is

$$T_{\text{reverse}} = n \cdot \frac{(V_{gy} V_{zz} - V_{gz} V_{yz})^2}{(V_{gg} V_{zz} - V_{gz}^2)(V_{yy} V_{zz} - V_{yz}^2)}, \tag{8}$$

where

$$V_{uv} = u^T \hat{\Sigma}_g^{-1} v - u^T \hat{\Sigma}_g^{-1} 1 (1^T \hat{\Sigma}_g^{-1} 1)^{-1} 1^T \hat{\Sigma}_g^{-1} v, \text{ for } u, v \in \{g, z, y\}.$$

T_{reverse} is asymptotically χ^2_1 distributed under H_0.

The response variable G, in the proposed 'reverse' regression (6), is a discrete random variable with three possible values 0, 1, and 2. However, in light of the result of Chen (1983), we modelled G using a linear model. We show in the following that indeed, in the simple case of independent sample, HWE, or no covariates, T_{reverse} derived from (6) shares similar properties with previous methods that treat G as discrete.

2.4 Connections with T_{indep} and T_{LMM}

In the absence of sample correlation, $\Sigma_g = I$, and the proposed test statistic in (8) is then simplified to T_{indep} in (3); T_{reverse} is symmetric with respect to G and Y, provided that the same Σ is used. Thus, for independent samples, T_{reverse} adjusts for HWD through $\text{var}(G) = 2(p(1-p)+\delta)$ as in the traditional association test, and

$$T_{\text{reverse},\ \Sigma_g = I} = T_{\text{indep}}.$$

In the presence of sample correlation. it is not difficult to see that if both T_{LMM} and T_{reverse} were to use the same, known correlation matrix Σ in their respective regression models (4) and (6), then the two tests are equivalent with each other,

$$T_{\text{reverse},\ \Sigma_g = \Sigma} = T_{\text{LMM},\ \Sigma_y = \Sigma}.$$

However, the traditional linear mixed-effect model uses kinship coefficient matrix Σ_Φ to model Σ_y, while the proposed method adjusts for sample correlation through Σ_g that includes both Σ_Φ for relatedness and a correction factor δ in Σ_δ for potential HWD at the tested SNP.

2.5 Connections with Quasi-likelihood Association Tests without Covariates, $T_{\text{QL-binary}}$ and T_{GQLS}

For case-control studies with related samples, Thornton and McPeek (2007) proposed an association test that compares sample allele frequency estimates between the case and control groups, while adjusting for relatedness between individuals. The test statistic is defined as,

$$T_{\text{QL-binary}} = \frac{(\hat{p}_r - \hat{p}_c)^2}{\widehat{\text{var}}(\hat{p}_r - \hat{p}_c)}, \tag{9}$$

where

$$\hat{p}_r = \frac{1_r^T \Sigma_\Phi^{-1} x}{1_r^T \Sigma_\Phi^{-1} 1}, \qquad \hat{p}_c = \frac{1^T \Sigma_\Phi^{-1} x}{1^T \Sigma_\Phi^{-1} 1}, \qquad x = g/2,$$
$$\hat{\text{var}}(\hat{p}_r - \hat{p}_c) = \frac{1}{2}\hat{p}_c(1-\hat{p}_c)[(1_r^T \Sigma_\Phi^{-1} 1)^{-2}(1_r^T \Sigma_\Phi^{-1} 1_r) - (1^T \Sigma_\Phi^{-1} 1)^{-1}],$$

and 1_r is a $n \times 1$ vector with the ith observation to be 1 if individual i is from the case group, and 0 otherwise. Note that $\hat{p}_r$ is a sample estimate of the allele frequency in cases, while $\hat{p}_c$ is the pooled estimate using the case-control combined sample, both adjusting for sample correlation through Σ_Φ.

Feng et al. (2011) and Feng (2014) reformulated $T_{\text{QL-binary}}$ as a generalized quasi-likelihood score test that handles both continuous and binary traits but does not allow for covariates. The quasi-likelihood based regression framework assumes that,

$$E(x_i|y_i) = \mu_i, \quad \log(\frac{\mu_i}{1-\mu_i}) = \beta_0 + \beta_1 y_i, \quad \text{var}(x|y) = \Lambda^{1/2}\Sigma_\Phi\Lambda^{1/2},$$

where Λ is an $n \times n$ diagonal matrix, and $\{\text{diag}\{\Lambda\}\} = \{\mu_1(1-\mu_1), \mu_2(1-\mu_2), \cdots, \mu_n(1-\mu_n)\}$.

Under the null of no association, the framework also assumes that $E(x_i) = p_c$, and $\text{var}(x) =$

$p_c(1-p_c)\Sigma_\Phi$. Thus, the generalized quasi-score statistic of testing $H_0 : \beta_1 = 0$ is,

$$T_{\text{GQLS}} = \frac{1}{\frac{1}{2}\hat{p}_c(1-\hat{p}_c)} \cdot \frac{[(y-\bar{y}1)^T\Sigma_\Phi^{-1}(x-\hat{p}_c 1)]^2}{(y-\bar{y}1)^T\Sigma_\Phi^{-1}(y-\bar{y}1)}, \tag{10}$$

where $\bar{y} = (1^T\Sigma_\Phi^{-1}1)^{-1}1^T\Sigma_\Phi^{-1}y$, and $\hat{p}_c$ is identical to that in $T_{\text{QL-binary}}$ of (9). When Y is binary, $y = 1_r$. Substituting y with 1_r in (10), it is easy to show that

$$T_{\text{GQLS, } Y=\text{binary}} = T_{\text{QL-binary}}.$$

When individuals are independent of each other in a sample, $\Sigma_\Phi = I$, then $\hat{p}_c = \hat{p}$. Thus, $\hat{p}_c(1-\hat{p}_c)$ in the denominator of (10) assumes $\delta = 0$ (HWE) in estimating $\text{var}(G/2)$. In contrast,

$$T_{\text{reverse, } \Sigma_g=I, \text{ no } Z} = \frac{1}{1+\frac{\hat{\delta}}{\hat{p}(1-\hat{p})}} \times T_{\text{GQLS, } \Sigma_\Phi=I}. \tag{11}$$

2.6 Connections with MASTOR, an Alternative Mixed-Model Association Test with Covariates, T_{MASTOR}

For continuous traits with covariates, Jakobsdottir and McPeek (2013) proposed MASTOR, a mixed-effect model-based association test. The MASTOR test statistic of no association is defined as,

$$T_{\text{MASTOR}} = \frac{(x^T P_y^* y)^2}{\hat{\text{var}}(x^T P_y^* y|W,y)} = \frac{(x^T P_y^* y)^2}{\hat{\sigma}_x^2 y^T P_y^{*T}\Sigma_\Phi P_y y}, \tag{12}$$

where

$$P_y^* = V_{xy}^* - V_{xz}^* V_{yz}^* / V_{zz}^*, \ \hat{\sigma}_x^2 = x^T\{\Sigma_\Phi^{-1} - \Sigma_\Phi^{-1}1(1^T\Sigma_\Phi^{-1}1)^{-1}1^T\Sigma_\Phi^{-1}\}x/(n-1),$$
$$V_{uv}^* = u^T\hat{\Sigma}_y^{-1}v - u^T\hat{\Sigma}_y^{-1}1(1^T\hat{\Sigma}_y^{-1}1)^{-1}1^T\hat{\Sigma}_y^{-1}v, \text{ for } u,v \in \{x,z,y\}, \ x = g/2,$$

and Σ_y is the same as that in the linear mixed-effect model of (4), and $\hat{\Sigma}_y$ is the maximum likelihood estimate under the null of no association.

To compare T_{reverse} and T_{MASTOR}, we rewrite T_{reverse} in (8) as

$$T_{\text{reverse}} = \frac{(g^T P_g y)^2}{\hat{\sigma}_g^2 y^T P_g y}, \tag{13}$$

where $P_g = V_{gy} - V_{gz} V_{yz} / V_{zz}$, and $\hat{\sigma}_g^2 = g^T P_g g / n$.

Note that V_{uv} in (8) (or re-formulated as in (13)) has the same form as V_{uv}^* above. Thus, if we were to use $\Sigma_y = \Sigma_\Phi$ in the calculation of P_y^* for T_{MASTOR} (i.e. assume $h^2 = 1$ in $\Sigma_y = h^2 \Sigma_\Phi + (1 - h^2) I$), and not accounting for HWD in the calculation of T_{reverse} (i.e. assume $\delta = 0$ in $\sigma^2 \Sigma_\Phi + \Sigma_\delta$), then

$$T_{\text{reverse},\ \delta=0} = T_{\text{MASTOR},\ \Sigma_y = \Sigma_\Phi}.$$

However, although both T_{MASTOR} and T_{reverse} measure the correlation between G and Y while adjusting for covariate effects, the two approaches are distinct from each other. T_{reverse} in (8) is directly derived from a regression framework with G as the response variable and Y as a covariate, explicitly adjusting for HWD through $\sigma^2 \Sigma_g = \sigma^2 \Sigma_\Phi + \Sigma_\delta$ for the tested SNP using model (6). That is, V_{uv} uses Σ_g that contains both Σ_Φ and Σ_δ as illustrated in (7) for sibling pairs. In contrast, T_{MASTOR} in (12) is a hybrid score statistic, where the score function is derived from the linear mixed-effect model with Y as the response variable, while the variance of the score function is estimated from a 'retrospective' approach that models $X = G/2$ as the response variable. When estimating σ_x^2 and modelling Σ_y, T_{MASTOR} considers Σ_Φ alone to account for the genetic relationship between related individuals. Hence, T_{MASTOR} implicitly assumes HWE.

2.7 Allele Frequency Estimation with the Proposed 'Reverse' Regression

Another feature of the proposed 'reverse' regression framework is its inherent ability to estimate allele frequencies in dependent samples while adjusting for covariate effects. When no phenotype Y is included in (6), the 'reverse' regression is simplified to

$$g = \alpha 1 + \gamma z + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 \Sigma_g). \tag{14}$$

The intercept α can be interpreted as twice the population allele frequency, p, for the SNP under the study. The maximum likelihood estimator of α is

$$\hat{\alpha}_z = \frac{(z^T \hat{\Sigma}_g^{-1} z)^{-1} z^T \hat{\Sigma}_g^{-1} g - (1^T \hat{\Sigma}_g^{-1} z)^{-1} 1^T \hat{\Sigma}_g^{-1} g}{(z^T \hat{\Sigma}_g^{-1} z)^{-1} z^T \hat{\Sigma}_g^{-1} 1 - (1^T \hat{\Sigma}_g^{-1} z)^{-1} 1^T \hat{\Sigma}_g^{-1} 1}. \tag{15}$$

When no covariate is included in (14) and Σ_g is replaced with the kinship coefficient matrix Σ_Φ, $\hat{\alpha}_z$ is then reduced to

$$\hat{\alpha} = (1^T \Sigma_\Phi^{-1} 1)^{-1} 1^T \Sigma_\Phi^{-1} g.$$

Interestingly, $\hat{\alpha}/2$ is the best linear unbiased estimator of p as studied in McPeek et al. (2004). Thus, the proposed $\hat{\alpha}_z/2$ is a generalized allele frequency estimator that, in addition to accounting for sample dependency, can also adjust for covariate effects and potential departure from Hardy–Weinberg equilibrium.

3 Empirical Results

Given the analytical insights provided above, here we briefly examine the performance of the commonly used T_{LMM} based on the linear mixed-effect model of (4) and the proposed T_{reverse} based on (6) for association analyses, through application and simulation studies. We also demonstrate numerically the additional utility of T_{reverse} for allele frequency estimation.

3.1 Cystic Fibrosis Sib-pair Data

We first extracted 65 sibling pairs from a cystic fibrosis (CF) gene modifier study, as previously described in Sun et al. (2012) and Wright et al. (2011). The outcome Y of interest is the lung function measurements of the 130 CF subjects. In total, there were $570,539$ SNPs genotyped using the Illumina 610-Quad Beadchip. To stabilize the variance estimation, we also required SNPs to have minor allele frequency greater than 5%. We then applied the T_{LMM} and the proposed T_{reverse} to the remaining $505,172$ SNPs. For the implementation of T_{LMM}, we treated h^2 as unknown and estimated it based on the linear mixed-effect model of (4) as in convention.

As expected, the results of T_{LMM} and the proposed T_{reverse} are similar to each other (Figure 1), and both have good type I error control (results not shown). However, the estimated h^2, obtained using the 65 sibling data, is $\hat{h}^2 = 0.82$. This value is substantially greater than 0.5, the commonly-believed 'true' heritability of lung function in CF (Vanscoy et al., 2007). To verify if the biased heritability estimate is due to chance, we conducted a simulation study assuming that only one causal SNP, G_{causal} with MAF of 0.2, affects Y with $h^2 = 0.5$. Genotype and phenotype values for 65 sibling pairs were then simulated under the assumption of HWE. Among the $100,000$ independently generated replicates, only 4.24% of the heritability estimates was greater than $\hat{h}^2 = 0.82$, the value that was observed in the real CF data.

One possible explanation is that the causal SNP(s) may be not be in HWE, resulting in a biased estimator of heritability, $\hat{h}^2$. Following the same sib-pair design, we next used simulations to demonstrate that (a) assuming the true h^2 is known, the empirical type I error rate of linear mixed-effect model (4) inflates if $\delta > 0$, or deflates if $\delta < 0$, and (b) when h^2 is treated as a nuisance parameter, its estimate based on model (4) can be biased in the presence of Hardy–Weinberg disequilibrium.

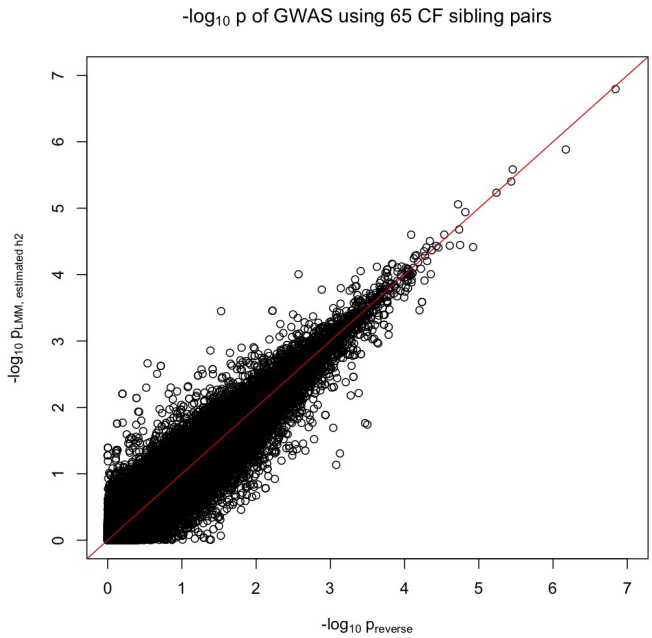

Figure 1: **CF sib-pair data application results, comparing $T_{\mathbf{LMM}}$ based on the linear mixed-effect model (4) and $T_{\mathbf{reverse}}$ based on the proposed 'reverse' regression model (6).** The association is between lung function measurements in 65 sibling paris with cystic fibrosis and $505,172$ autosomal SNPs (with minor allele frequency greater than 0.05). Implementation of T_{LMM} assumed h^2 unknown, and $\hat{h}^2$ estimated based on 65 samples was 0.82; heritability of lung function in CF is believed to be around 0.5 based on CF epidemiological studies (Vanscoy et al., 2007).

3.2 Simulated Sib-pair Data

Consider a continuous trait Y with $h^2 = 0.5$ and influenced by one causal SNP with minor allele frequency of 0.2, G_{causal}, for which the HWD factor is δ_{causal}. We conducted association testing between Y and a non-associated SNP, G_{tested} (with its own δ_{tested}) using simulated data from 65 sibling pairs. The sample size 65 was chosen to match with the number of sibling pairs in the cystic fibrosis dataset in Section 3.1.

Figure 2(a) plots the empirical type I error rates (black circles) of T_{LMM} using the true $h^2 = 0.5$, for a nominal level of 0.05, estimated from independently simulated $10,000$ replicates for each δ_{causal} value. The trend of type I error inflation is clear as δ_{causal} increases. In Figure 2(a) we set

$\delta_{\text{tested}} = 0.06$, but we note that the root cause of the type I error issue is $\delta_{\text{causal}} \neq 0$ when using the LMM (4). Figure 2(b) shows that even if G_{tested} is in HWE ($\delta_{\text{tested}} = 0$) the problem remains, albeit less severe, as long as $\delta_{\text{causal}} \neq 0$. The simulation results also confirm that the proposed T_{reverse} has the correct type I error control (red triangles in Figure 2); results are similar if G_{tested} were not simulated but based on the real genotype data observed in the 65 sibling pairs.

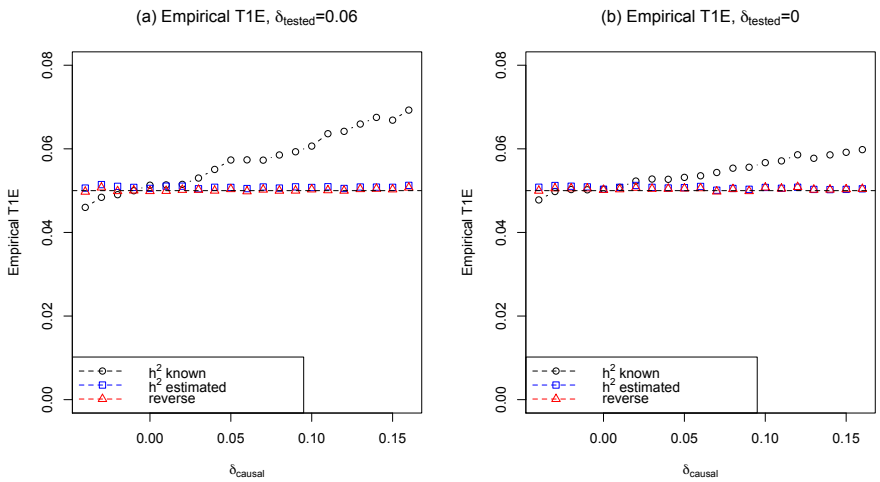

Figure 2: **Empirical type I error rate of T_{LMM} based on the linear mixed-effect model (4) and T_{reverse} based on the proposed 'reverse' regression model (6), against δ_{causal}.** (a) When G_{tested} of tested SNPs are in HWD with $\delta_{\text{tested}} = 0.06$. (b) When G_{tested} of tested SNPs are in HWE with $\delta_{\text{tested}} = 0$. The true heritability of the phenotype is $h^2 = 0.5$, the minor allele frequencies $p_{\text{causal}} = p_{\text{tested}} = 0.2$, and $10,000$ independent replicates of phenotypes and genotypes for 65 sibling pairs were simulated for each δ_{causal} value. Black circles are for T_{LMM} using the true $h^2 = 0.5$, blue circles are for T_{LMM} while estimating h^2 (results of $\hat{h}^2$ shown in Figure 3), and red triangles are for T_{reverse}.

In most practical implementations of the linear mixed-effect model (4), h^2 is treated as a nuisance parameter, and no type I error issue has been reported. Indeed, when h^2 was estimated in our simulation study, the size of T_{LMM} was correct (blue squares in Figure 2). However, in this situation, the impact of HWD is now on the estimation of h^2. Specifically, we applied the LMM (4) to the same simulated data as above but assumed h^2 to be unknown. Figure 3 clearly shows that $\hat{h}^2$ is downward biased when $\delta_{\text{causal}} < 0$, and upward biased if $\delta_{\text{causal}} > 0$. The bias can be

substantial. For example, when $\delta_{\text{causal}} = 0.10$, the estimated heritability $\hat{h}^2$ is centred at 0.78 as compared to the true value of 0.5.

In Figure 3, it is notable that when $\delta_{\text{causal}} > 0.1$, $\hat{h}^2 > 1$. Since h^2 is the proportion of variance in Y explained by additive genetic variation, $0 \leq h^2 \leq 1$ by definition. However, if $\delta_{\text{causal}} \neq 0$, $\hat{h}^2$ based on the LMM, without additional truncation, is a biased estimate of h^2 by a factor of $1 + \hat{\delta}/\hat{\sigma}^2$.

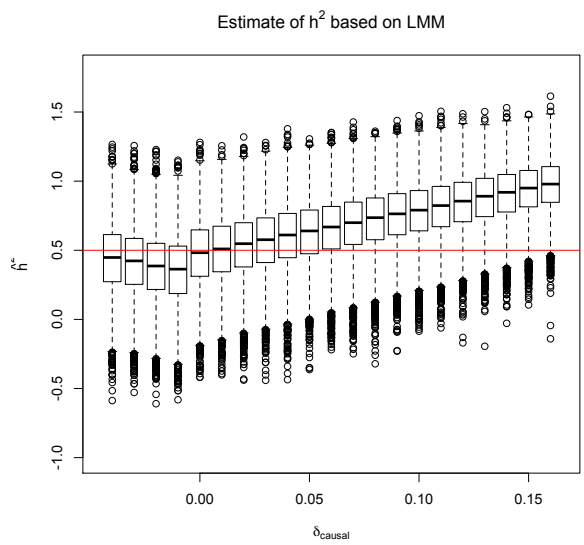

Figure 3: **Box-plots of $\hat{h}^2$, estimated from the linear mixed-effect model (4) against δ_{causal}.** The true heritability of the phenotype is $h^2 = 0.5$. The minor allele frequencies $p_{\text{causal}} = p_{\text{tested}} = 0.2$, and $10,000$ independent replicates of phenotypes and genotypes for 65 sibling pairs were simulated for each δ_{causal} value. The empirical type 1 error rates are shown in Figure 2 as blue circles.

3.3 Estimation of minor allele frequency

Here we compare the best linear unbiased estimator (BLUE) by McPeek et al. (2004) with the proposed 'reverse' regression estimator, in the presence of sample correlation and Hardy–Weinberg disequilibrium. We chose the parent-child relationship for this study. Note that even though the kinship coefficient of a parent-child pair is identical to that of a sib-pair, Σ_δ is not the same for these

two relationship types. In a parent-child pair, the parent is considered as the founder of the pedigree and the child as non-founder, while in a sib-pair both siblings are considered as non-founders. As a result,

$$\mathrm{var}(G_{\mathrm{parent}}) = \sigma^2 + \delta, \text{ vs. } \mathrm{var}(G_{\mathrm{child}}) = \mathrm{var}(G_{\mathrm{sibling}}) = \sigma^2.$$

For a parent-child pair,

$$\sigma^2 \Sigma_g = \sigma^2 \Sigma_\Phi + \Sigma_\delta = \sigma^2 \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} + \begin{pmatrix} \delta & 0.5\delta \\ 0.5\delta & 0 \end{pmatrix}. \tag{16}$$

Let $\bar{g}_{\mathrm{parent}} = \sum_{i=1}^{m} G_{i,\mathrm{parent}}/m$ and $\bar{g}_{\mathrm{child}} = \sum_{i=1}^{m} G_{i,\mathrm{child}}/m$, where m is the number of independent families, the minor allele frequency estimates are

$$\hat{p}_{\mathrm{BLUE}} = (\bar{g}_{\mathrm{parent}} + \bar{g}_{\mathrm{child}})/4, \text{ and} \tag{17}$$

$$\hat{\alpha}_{\mathrm{reverse}}/2 = \{(1-\hat{\rho})\bar{g}_{\mathrm{parent}} + (1+\hat{\rho})\bar{g}_{\mathrm{child}}\}/4,$$

where

$$\hat{\rho} = \hat{\delta}/\hat{\sigma}^2.$$

It is easy to show that $E(\hat{\alpha}_{\mathrm{reverse}}/2) = E(\hat{p}_{\mathrm{BLUE}}) = p$ regardless of the value of ρ.

In equation (17), $\hat{p}_{\mathrm{BLUE}}$ always weighs $\bar{g}_{\mathrm{parent}}$ and $\bar{g}_{\mathrm{child}}$ equally. For the proposed estimator $\hat{\alpha}_{\mathrm{reverse}}/2$, if $\delta > 0$, $\mathrm{var}(\bar{g}_{\mathrm{parent}}) > \mathrm{var}(\bar{g}_{\mathrm{child}})$. Accordingly, because $\rho > 0$, $\hat{\alpha}_{\mathrm{reverse}}/2$ would weigh $\bar{g}_{\mathrm{parent}}$ less than $\bar{g}_{\mathrm{child}}$. Conversely, if $\delta < 0$, $\mathrm{var}(\bar{g}_{\mathrm{parent}}) < \mathrm{var}(\bar{g}_{\mathrm{child}})$, and because $\rho < 0$, $\hat{\alpha}_{\mathrm{reverse}}/2$ would weigh $\bar{g}_{\mathrm{parent}}$ more than $\bar{g}_{\mathrm{child}}$. Therefore, variance of $\hat{\alpha}_{\mathrm{reverse}}/2$ is smaller than that of $\hat{p}_{\mathrm{BLUE}}$.

To confirm this analytical insights numerically, we simulated $1,000$ parent-child pairs, where the parents' genotypes were out of HWE, with $\delta_{\mathrm{parent}} = 0.06$ and $p_{\mathrm{parent}} = 0.2$. A total of $50,000$ independent replicates were simulated, and the corresponding minor allele frequency estimates

by the two methods are shown in Figure 4(a). To demonstrate the empirical standard error of the estimates, we first randomly split the $50,000$ estimates into 100 even sets, then calculated the sample standard error based on the 500 estimates from each set. The corresponding 100 standard error estimates are shown in Figure 4(b). Results clearly demonstrate the estimation efficiency of the proposed $\hat{\alpha}_{\text{reverse}}/2$ as compared with $\hat{p}_{\text{BLUE}}$ in the presence of HWD for parent-child pairs.

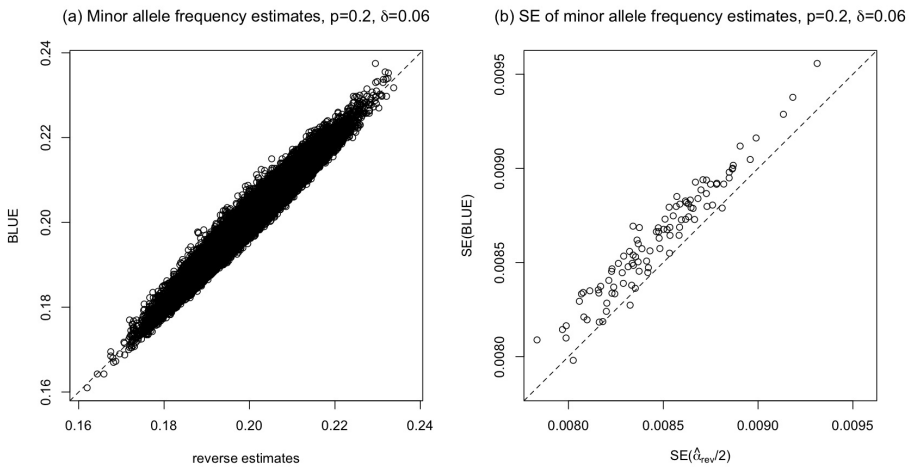

Figure 4: **Results of the minor allele frequency estimation by BLUE and the proposed estimator (17) for simulated parent-child data.** The proposed estimator is a special case of $\hat{\alpha}_z$ of (15), derived from the proposed 'reverse' regression without covariate Z. (a) Minor allele frequency estimates based on $50,000$ independent replicates. (b) Empirical standard error of the minor allele frequency estimates, based on 100 independent sets of 500 estimates. Each minor allele frequency estimation was based on $1,000$ parent-child pairs, $p_{\text{parent}} = 0.2$ and $\delta_{\text{parent}} = 0.06$ (offspring data are determined by parental data), with 500×100 independently simulated replicates.

4 Discussion

In this note, after reviewing a catalogue of genetic association tests we proposed a 'reverse' regression model, treating genotype G as the response variable and phenotype Y as a covariate. The proposed method is robust and flexible, and it can (i) analyze both binary and continuous outcomes, (ii) adjust for additional covariate effects, (iii) handle dependent samples, and (iv) incorpo-

Table 1: **Comparison of genotype-based association tests.**

Descriptions			Performance[a]				Proposed unifying framework
Test statistic	Equation	var-cov matrix[b]	(i) Diverse Y	(ii) Z adjustment	(iii) Dependency	(iv) HWD	T_{reverse} (6), based on the 'reverse' regression model (8) $g = \alpha 1 + \beta y + \gamma z + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 \Sigma_g)$, $\sigma^2 \Sigma_g = \sigma^2 \Sigma_\Phi + \Sigma_\delta$.
Traditional phenotype-on-genotype ($Y \sim G$) association tests							
T_{indep}	(2), (3)	I	✓	✓	✗	✓	$T_{\text{reverse},\, \Sigma_g = I} = T_{\text{indep}}$
T_{LMM}	(4)	Σ_y	✓	✓	✓	✗	$T_{\text{reverse},\, \Sigma_g = \Sigma} = T_{\text{LMM},\, \Sigma_y = \Sigma}$
Genotype-on-phenotype ($G \sim Y$) association tests							
MultiPhen[c]	N/A	N/A	✓	✓	✗	✓	N/A
$T_{\text{QL-binary}}$	(9)	Σ_Φ	✗	✗	✓	✗	$T_{\text{GQLS},\, Y = \text{binary}} = T_{\text{QL-binary}}$
T_{GQLS}	(10)	Σ_Φ	✓	✗	✓	✗	$T_{\text{reverse},\Sigma_g = I, \text{no } Z} = \frac{1}{1 + \frac{\hat{\delta}}{\hat{p}(1-\hat{p})}} \times T_{\text{GQLS},\Sigma_\Phi = I}$
T_{MASTOR}	(12)	Σ_Φ, Σ_y	✗	✓	✓	✗	$T_{\text{reverse},\delta = 0} = T_{\text{MASTOR},\Sigma_y = \Sigma_\Phi}$
Proposed 'reverse' regression model							
T_{reverse}	(6), (8)	Σ_Φ, Σ_δ	✓	✓	✓	✓	

[a]: (i) analyze both continuous and binary outcomes, Y;
(ii) adjust for covariate Z effects;
(iii) handle related individuals in dependent samples;
(iv) robust to Hardy–Weinberg disequilibrium (HWD).

[b]: I is the identity matrix, Σ_Φ is the kinship coefficient matrix, $\Sigma_y = h^2 \Sigma_\Phi + (1 - h^2) I$, and Σ_δ depends on the relationship type, e.g.

$$\begin{pmatrix} 0 & 0.5\delta \\ 0.5\delta & 0 \end{pmatrix}$$

for a sib-pair, where δ models potential departure from Hardy–Weinberg equilibrium (HWE).

[c]: ordinal logistic regression applied to independent samples.

rate a factor δ to adjust for potential departure from Hardy–Weinberg equilibrium. Furthermore, the proposed T_{reverse} unifies many existing genotype-based tests, including T_{LMM} based on the linear mixed-effect model, quasi-likelihood score test for binary traits $T_{\text{QL-binary}}$ (Thornton and McPeek, 2007), generalized quasi-likelihood score test T_{GQLS} (Feng et al., 2011; Feng, 2014), and T_{MASTOR} (Jakobsdottir and McPeek, 2013). Finally, the proposed regression framework can also estimate allele frequency while adjusting for covariates and departure from HWE. Table 1 provides a summary of the method comparison.

In the 'reverse' regression model, we treated the discrete genotype data G as continuous, because the method is purposed for association testing and earlier work has shown that, under some regularity conditions, the score test statistic for the exponential family is identical (Chen, 1983). Indeed, for the simple cases of independent samples, no covariates, or HWE, desirably the proposed test numerically coincides with a number of existing methods as shown in Section 2.

Departure from HWE generally falls into two categories: disequilibrium at the causal SNP(s) and disequilibrium at the tested SNP. If HWE does not hold at the causal SNP(s), models that rely on Σ_y (e.g. the linear mixed-effect model and MASTOR) may have inflated type I error rate if $\delta_{\text{causal}} > 0$, or deflated type I error rate if $\delta_{\text{causal}} < 0$, assuming the true heritability h^2 is known; if h^2 is estimated based on the data, the estimate is then biased if $\delta \neq 0$. If Hardy–Weinberg disequilibrium occurs at the tested SNP, models that assumes $\text{var}(g_{\text{tested}}) = \sigma_g^2 \Sigma_\Phi$ (e.g. quasi-likelihood score test and MASTOR) may be sensitive to $\delta_{\text{tested}} \neq 0$.

At a tested SNP, because the proposed 'reverse' framework is conditional on Y, the variance-covariance matrix only concerns G_{tested}, i.e. Σ_g. The modelling and estimation of Σ_g can account for potential departure from HWE through Σ_δ, in addition to genetic correlation as captured by the kinship coefficient matrix of Σ_Φ, resulting in a more robust association test for related individuals.

We demonstrated the type I error issue of the linear mixed-effect model, in the presence of HWD and assuming the true heritability is known, using data consists of related individuals only. In practice, this issue diminishes if the sample includes a large number of independent individuals

or the magnitude of HWD is small. Nevertheless, the analytical framework presented here can be an useful alternative that not only account for potential HWD but also unifies a number of existing tests. Even if HWD is of no concern, the proposed framework generalizes the MultiPhen approach (O'Reilly et al., 2012) to jointly analyze multiple phenotypes using related individuals. Further, we demonstrated that when h^2 is treated as unknown, its estimation can be biased and often upwardly (if $\hat{\delta} > 0$) as seen in the CF application. This insight offers a possible alternative interpretation of the 'missing heritability' issue (Maher, 2008).

In practice, SNPs out of HWE are typically not analyzed due to concerns for low genotyping quality. However, the issue identified here remains relevant: HWE quality control screening uses a p-value threshold in the range of 10^{-8} (Consortium et al., 2007), and this practice itself can be called into question because a truly associated SNP is likely to be out of HWE (Turner et al., 2011; Ryckman and Williams, 2008). The potential of using the proposed framework to jointly test δ and β to increase the power of association testing is of future research interest.

For genetic epidemiological interpretation, it may appear that the traditional Y-on-G regression approach is more intuitive. But, as pointed out by Schaid et al. (2013) in a different setting, the retrospective alternative that treats G as random while conditioning on Y "overcomes the problem of modelling the ascertainment process, which would be particularly challenging for highly enriched pedigrees." How to model gene-environment interaction using the proposal 'reverse' regression framework, however, remains an open question.

Acknowledgements

We thank Dr. Lisa J. Strug and her lab for providing the cystic fibrosis genotype data. This research is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC, RGPIN-04934 and RGPAS-522594), and the Canadian Institutes of Health Research (CIHR, MOP-310732) to LS. LZ is a trainee of the CIHR STAGE (Strategic Training in Advanced Genetic Epidemiology)

training program at the University of Toronto.

References

Chen, C.-F., 1983: Score tests for regression models. *Journal of the American Statistical Association*, **78 (381)**, 158–161.

Consortium, W. T. C. C., and Coauthors, 2007: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447 (7145)**, 661.

Derkach, A., J. F. Lawless, and L. Sun, 2015: Score tests for association under response-dependent sampling designs for expensive covariates. *Biometrika*, **102 (4)**, 988–994.

Eu-Ahsunthornwattana, J., E. N. Miller, M. Fakiola, S. M. Jeronimo, J. M. Blackwell, H. J. Cordell, W. T. C. C. C. 2, and Coauthors, 2014: Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLoS Genetics*, **10 (7)**, e1004 445.

Feng, Z., 2014: A generalized quasi-likelihood scoring approach for simultaneously testing the genetic association of multiple traits. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **63 (3)**, 483–498.

Feng, Z., W. W. Wong, X. Gao, and F. Schenkel, 2011: Generalized genetic association study with samples of related individuals. *The Annals of Applied Statistics*, 2109–2130.

Jakobsdottir, J., and M. S. McPeek, 2013: Mastor: mixed-model association mapping of quantitative traits in samples with related individuals. *The American Journal of Human Genetics*, **92 (5)**, 652–666.

Maher, B., 2008: Personal genomes: The case of the missing heritability. *Nature News*, **456 (7218)**, 18–21.

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

O'Reilly, P. F., C. J. Hoggart, Y. Pomyen, F. C. Calboli, P. Elliott, M.-R. Jarvelin, and L. J. Coin, 2012: Multiphen: joint model of multiple phenotypes can increase discovery in gwas. *PloS One*, **7 (5)**, e34 861.

Ryckman, K., and S. M. Williams, 2008: Calculation and use of the hardy-weinberg model in association studies. *Current Protocols in Human Genetics*, **57 (1)**, 1–18.

Sasieni, P. D., 1997: From genotypes to genes: doubling the sample size. *Biometrics*, 1253–1261.

Schaid, D. J., and S. J. Jacobsen, 1999: Biased tests of association: comparisons of allele frequencies when departing from hardy-weinberg proportions. *American Journal of Epidemiology*, **149 (8)**, 706–711.

Schaid, D. J., S. K. McDonnell, J. P. Sinnwell, and S. N. Thibodeau, 2013: Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. *Genetic Epidemiology*, **37 (5)**, 409–418.

Sun, L., and A. Dimitromanolakis, 2012: Identifying cryptic relationships. *Statistical Human Genetics: Methods and Protocols*, 47–57.

Sun, L., and Coauthors, 2012: Multiple apical plasma membrane constituents are associated with susceptibility to meconium ileus in individuals with cystic fibrosis. *Nature genetics*, **44 (5)**, 562.

Thornton, T., and M. S. McPeek, 2007: Case-control association testing with related individuals: a more powerful quasi-likelihood score test. *The American Journal of Human Genetics*, **81 (2)**, 321–337.

Turner, S., and Coauthors, 2011: Quality control procedures for genome-wide association studies. *Current Protocols in Human Genetics*, **68 (1)**, 1–19.

Vanscoy, L. L., and Coauthors, 2007: Heritability of lung disease severity in cystic fibrosis. *American journal of respiratory and critical care medicine*, **175 (10)**, 1036–1043.

Visscher, P. M., W. G. Hill, and N. R. Wray, 2008: Heritability in the genomics era—concepts and misconceptions. *Nature Reviews Genetics*, **9 (4)**, 255.

Weir, B. S., 1996: *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. 575.1072 W4.

Wright, F. A., and Coauthors, 2011: Genome-wide association and linkage identify modifier loci of lung disease severity in cystic fibrosis at 11p13 and 20q13. 2. *Nature genetics*, **43 (6)**, 539.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher, 2011: Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, **88 (1)**, 76–82.