A Generalized Robust Allele-based Genetic Association Test

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

March 12, 2020

Abstract

The allele-based association test or the allelic test, comparing allele frequency difference between case and control groups, is locally most powerful. However, the classical allelic test is limited in applications because it is sensitive to the Hardy–Weinberg equilibrium (HWE) assumption, not applicable to continuous traits, and not easy to account for covariate effects or sample correlation. To develop a generalized robust allelic test, we propose a unifying regression model with individual allele as the response variable. We show that the score test statistic derived from this novel regression framework contains a correction factor that explicitly adjusts for the departure from HWE and encompasses the classical allelic test as a special case. When the trait of interest is continuous, the corresponding allelic test evaluates a weighted difference between individual-level allele frequency estimate and sample estimate where the weight is proportional to an individual's trait value, and the test remains valid under Y-dependent sampling. Finally, the proposed method allows for joint allele-based association analyses of multiple (continuous or binary) phenotypes, in the presence of covariates, sample correlation and population heterogeneity. To support our analytical findings, we provide empirical evidence from both simulation and application studies.

Keywords: Allele-based association analysis; Correlation; Hardy–Weinberg equilibrium; Multiple phenotypes; Multiple populations; Relatedness; Robustness.

1 Introduction

A key component of current large-scale genetic studies of complex human traits is association analysis. An association study aims to identify genetic markers that influence a heritable trait or phenotype of interest, while accounting for environmental effects. To formulate the problem more precisely, assume that single nucleotide polymorphisms (SNPs) are the genetic markers available. For each bi-allelic SNP, let a and A be the two possible alleles, and as in convention let A denote the minor allele with population frequency $p \leq 0.5$. The SNP genotype G for an individual is a paired (but unordered) alleles, taking the form of aa, Aa or AA. For a case-control association study of a binary trait (Table 1), intuitively one can compare the estimates of allele frequency of A between the case and control groups. Indeed, the resulting allelic test is locally most powerful, but the validity of the test hinges on the assumption of Hardy-Weinberg equilibrium (HWE) (Sasieni, 1997). Counting each genotype AA contributing two *independent* copies of allele A, the allelic test 'doubles' the sample size but implicitly assumes HWE (Sasieni, 1997). That is, the genotype frequencies depend only on the allele frequencies as, $p_{aa} = (1-p)^2$, $p_{Aa} = 2p(1-p)$ and $p_{AA} = p^2$.

Table 1: **Notations for genotype and allele counts for a case-control study**. The HLA-DQ3 example is from Sasieni (1997), studying women with cervical intraepithelial neoplasia 3.

	Genotype Counts				Allele Counts		
	aa	Aa	AA	Total	a	A	Total
Case	r_0	r_1	r_2	r	$2r_0 + r_1$	$r_1 + 2r_2$	$2r$
Control	s_0	s_1	s_2	s	$2s_0 + s_1$	$s_1 + 2s_2$	$2s$
	n_{aa}	n_{Aa}	n_{AA}		n_a	n_A	
Total	n_0	n_1	n_2	n	$2n_0 + n_1$	$n_1 + 2n_2$	$2n$
The HLA-DQ3 example from Sasieni (1997)							
Case	40	45	28	113	125	101	226
Control	273	100	43	416	646	186	832
Total	313	145	71	529	771	287	1058

For a population to be in HWE, several assumptions must be (approximately) true including random mating, infinite population size, and no inbreeding, mutation, migration, or selection (Hardy et al., 1908; Weinberg, 1908). To evaluate the HWE assumption using an independent sample as in Table 1, one typically applies the Pearson goodness-of-fit χ^2 test, $\sum_{i=1}^{3} \frac{(O_i-E_i)^2}{E_i} = \frac{(n_0-n(1-p)^2)^2}{n(1-p)^2} + \frac{(n_1-n2p(1-p))^2}{n2p(1-p)} + \frac{(n_2-np^2)^2}{np^2} \sim \chi^2_2$. In practice, allele frequency p is often unknown and commonly replaced by the sample estimate resulting in loss of degrees of freedom (d.f.). The resulting Pearson-based HWE test thus has the following form,

$$T_{\text{HWE, Pearson}} = \frac{(n_0 - n(1-\hat{p})^2)^2}{n(1-\hat{p})^2} + \frac{(n_1 - n2\hat{p}(1-\hat{p}))^2}{n2\hat{p}(1-\hat{p})} + \frac{(n_2 - n\hat{p}^2)^2}{n\hat{p}^2} \sim \chi^2_1, \tag{1}$$

where $\hat{p} = (n_1 + 2n_2)/2n$. Using the HLA-DQ3 data in Table 1 as an illustration, among a total of 529 individuals 313, 145 and 71 have genotypes, respectively, *aa*, *Aa* and *AA*. Direct application of $T_{\text{HWE, Pearson}}$ yields a test statistic of 49.7623 and a p-value of 1.74×10^{-12}, suggesting that the population is not in HWE.

In the presence of Hardy-Weinberg disequilibrium (HWD), the size of the classical allelic test is not controlled at the nominal level (Sasieni, 1997). Efforts have been made to alleviate this problem, mainly along the line of improving variance estimate of the original test statistic (Schaid and Jacobsen, 1999). However, this improvement does not resolve several important issues present in more complex data, including how to analyze continuous traits, how to include covariates, and how to cope with related individuals from families or pedigree data.

Consequently, most if not all current genetic association studies rely on genotype-based regression models, where the response variable is phenotype Y and the predictors include genotype G and other covariates. For the three genotype groups, *aa*, *Aa* and *AA*, the coding is commonly additive as 0, 1 and 2 (Hill et al., 2008). Note that although the genotype *AA* is also given a value of two here, the genotype-based approach is robust to HWD. This is because the $Y - G$ regression is performed conditional on genotype G, and the value two here merely specifies that the effect of

$G = AA$ on Y is twice that of $G = Aa$ on Y (i.e. additively). Nevertheless, it is a bit mysterious how exactly a genotype-based test statistic accounts for HWD. Further, the actual data collection typically starts with sampling individuals based on Y, which can be a random or Y-dependent sampling (Derkach et al., 2015). It then genotypes the sampled individuals to obtain G. Thus, it can be argued that the $G-Y$ regression is a more fitting statistical framework. This 'reverse' regression approach can also readily analyze multiple phenotypes simultaneously, which was the motivation behind the development of MultiPhen (O'Reilly et al., 2012). To deal with the three genotype groups, O'Reilly et al. (2012) used an ordinal logistic regression and stated that the proposed likelihood ratio test does not assume HWE. However, the statistical insight is lacking and analyzing pedigree data remains a challenge.

This work generalizes the locally most powerful allele-based association test to more complex settings by developing a novel *allele-based* 'reverse' regression framework. In what follows, Section 2 first revisits the classical allelic test, providing insight about the need for a more flexible formulation of the allelic test. Section 3 then develops the new allele-based 'reverse' regression framework by first appropriately partitioning the two alleles of a genotype then specifying the individual allele as the response variable. In addition to the parameter that captures the phenotype-genotype association, the proposed regression framework includes a new parameter that models the dependency between the two alleles of a genotype, explicitly accounting for potential departure from HWE. This section also provides examples that highlight the unifying feature of the proposed framework for both association analysis and HWE testing itself. Section 4 considers more complex settings including related individuals from pedigree data, genetic markers with more than two alleles, and multiple phenotypes and populations. Given the theoretical results presented, simulation experiments in Section 5 are relatively brief with additional empirical evidence from two applications. Section 6 concludes with remarks and discussion.

2 The classical allelic test revisited

For a given SNP and a binary phenotype of interest, let p_r denote the population frequency of allele A for the cases and p_s for the controls. A test of no association between the SNP and the disease status is to test the null hypothesis that H_0: $p_r = p_s$. The classical allelic test is a direct application of the standard test that compares two proportions using a pooled sample estimate of the variance,

$$T_{\text{allelic}} = \frac{(\hat{p}_r - \hat{p}_s)^2}{(\frac{1}{2r} + \frac{1}{2s})\hat{p}(1-\hat{p})} \overset{\text{HWE}}{\sim} \chi_1^2, \tag{2}$$

where, using the notations in Table 1, $\hat{p}_r = (2r_2 + r_1)/2r = r_A/2r$, $\hat{p}_s = (2s_2 + s_1)/2s = s_A/2s$ and $\hat{p} = (2n_2 + n_1)/2n = n_A/2n$ are the sample estimates of allele frequency, respectively, in the case, control and combined groups.

The validity of T_{allelic} however requires the Hardy–Weinberg equilibrium assumption, because only under HWE $n_A \sim \text{Binomial}(2n, p)$, and

$$\widehat{\text{var}}(\hat{p}_r - \hat{p}_s) \overset{\text{HWE}}{=} (\frac{1}{2r} + \frac{1}{2s})\hat{p}(1-\hat{p}).$$

Using the HLA-DQ3 data in Table 1 as an example, the HWE test in Section 1 has shown that the assumption of HWE is violated. Thus, a direct application of the allelic association test in this case ($T_{\text{allelic}} = 44.847$ corresponding to a p-value of 2.13×10^{-11}) is not appropriate.

Indeed, Sasieni (1997) has pointed out that T_{allelic} is valid and locally most powerful if and only if the HWE assumption holds and the genetic effect is additive (Web Appendix A). It is now well known that T_{allelic} can have inflated type 1 error rate. However, we emphasize that this is true only if there is an excess of homozygotes AA, i.e. $\delta > 0$, where

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

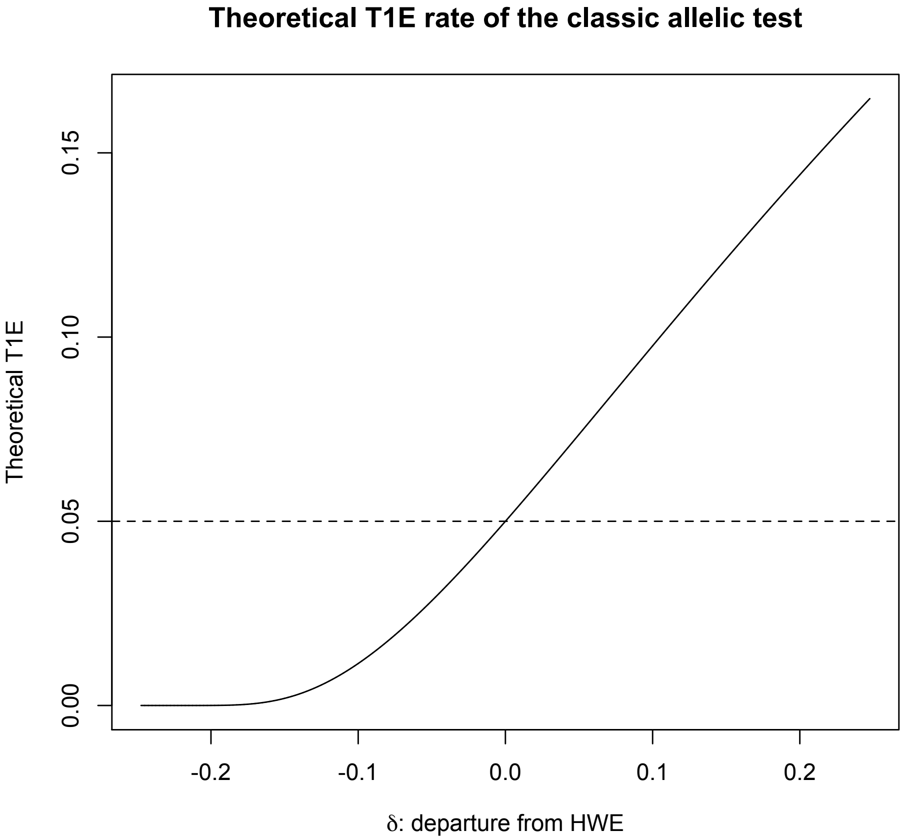

Figure 1: **The theoretical type 1 error rate of the classical allelic test, T_{allelic}, at the nominal level of $\alpha = 0.05$, with respect to departure from HWE, δ.** $\delta = p_{AA} - p^2$ is the classical measure of departure from HWE (Weir, 1996), where p is the frequency of the minor allele A and $-p^2 \leq \delta \leq p(1-p)$. When $p = 0.5$, $-0.25 \leq \delta \leq 0.25$.

100 is the most commonly used measure of Hardy–Weinberg disequilibrium (Weir, 1996). If $\delta < 0$,
101 T_{allelic} is conservative as shown in Figure 1.

To robustify T_{allelic} against HWD, Schaid and Jacobsen (1999) proposed a variance adjustment by directly modeling the genotype counts using a multinomial distribution. For the case group, $(r_0, r_1, r_2) \sim \text{Multinomial}\{r, (p_{aa}, p_{Aa}, p_{AA})\}$ under the null hypothesis of no association, and $\widehat{\text{var}}(\hat{p}_r) = \widehat{\text{var}}((2r_2 + r_1)/2r) = (\hat{p}(1-\hat{p}) + (\hat{p}_{AA} - \hat{p}^2))/2r = (\hat{p}(1-\hat{p}) + \hat{\delta}))/2r$, similarly for the control group replacing r with s. Hence,

$$\widehat{\text{var}}(\hat{p}_r - \hat{p}_s) = (\frac{1}{2r} + \frac{1}{2s})(\hat{p}(1-\hat{p}) + \hat{\delta}),$$

and the resulting test statistic is robust against HWD,

$$T_{\text{allelic, Schaid}} = \frac{(\hat{p}_r - \hat{p}_s)^2}{(\frac{1}{2r} + \frac{1}{2s})(\hat{p}(1-\hat{p}) + \hat{\delta})} \sim \chi^2_1. \tag{3}$$

The revised variance estimate has a correction term, $\hat{\delta} = (\hat{p}_{AA} - \hat{p}^2)$, which is the sample estimate of δ (Weir, 1996). Later in Section 3, we will provide analytical insight about how δ is related to $T_{\text{HWE, Pearson}}$ in (1). For now, it is clear that the denominator of T_{allelic} can be smaller or larger than that of $T_{\text{allelic, Schaid}}$, resulting in inflated (when $\hat{\delta} > 0$) or deflated (when $\hat{\delta} < 0$) type 1 error rate. In the HLA-DQ3 example, $\hat{\delta} = 0.061$. Thus, the classical allelic test will be too optimistic with $\{T_{\text{allelic}} = 44.8470\} > \{T_{\text{allelic, Schaid}} = 34.3207\}$.

This robust-variance approach is effective but limited to the simplest setting of case-control studies using independent observations with no covariates. In the presence of sample correlation, direct modifications of the $\hat{\delta}$ term, or more generally the analytical expression of $T_{\text{allelic, Schaid}}$, can be difficult. For example, it is not clear if r and s should be simply replaced by the effective numbers of sample size of the case and control groups, provided we know how to estimate them. It is also not clear how to use this comparing-two-proportions analytical framework to adjust for covariate effects or analyze other types of phenotype data, whereas many complex traits are continuous. Thus, an alternative formulation of allele-based association test is needed.

3 A Generalized Robust Allele-based (RA) Association Test

3.1 Decoupling the two alleles in a genotype

Consider a SNP with genotype $G \in \{aa, Aa, AA\}$ and for the moment assume that there are n independent observations, $G_i, i = 1, \ldots, n$. The partition of the homozygous genotypes aa and AA is straightforward, but the partition of the heterozygous genotype Aa requires additional considerations because of the unknown ordering of the two alleles (i.e. Aa and aA equally likely). We partition each G_i as follows,

$$(G_{i1}, G_{i2}) = \begin{cases} (0,0) & \text{if the genotype is } aa \\ (0,1) & \text{if the genotype is } Aa \text{ and } c_i = 0 \\ (1,0) & \text{if the genotype is } Aa \text{ and } c_i = 1 \\ (1,1) & \text{if the genotype is } AA \end{cases} \tag{4}$$

where $c_i \overset{iid}{\sim} \text{Bernoulli}(1/2)$ if $G_i = Aa$ for $i = 1, \ldots, n$.

Previous work attempted to split the n_{Aa} observations equally; exactly half of the n_{Aa} observations have $(G^*_{i1}, G^*_{i2}) = (0,1)$ and the other half have $(G^*_{i1}, G^*_{i2}) = (1,0)$ (Schaid et al., 2012; Bourgain et al., 2003). That is, $\sum_i G^*_{i1} \equiv \sum_i G^*_{i2} \equiv n_{AA} + n_{Aa}/2$. However, this even-split approach reduces the variation inherent in a randomly selected allele. One can show that $\text{var}(\sum_i G^*_{i1}) = n(p_{AA} + p_{Aa}/4 - (p_{AA} + p_{Aa}/2)^2)$ while $\text{var}(\sum_i G_{i1}) = \text{var}(\sum_i G^*_{i1}) + np_{Aa}/4$; the use of a fair coin in our proposed approach ensures that $\sum_i G_{i1} \sim \text{Binomial}(n, p_{AA} + p_{Aa}/2)$ and similarly for $\sum_i G_{i2}$ (Web Appendix B). As we will see in the following sections, this subtle difference in how we decouple the two alleles in a genotype, as compared with previous work, leads to correct inference for both association and HWE analyses.

3.2 Reformulating the test of HWE as an allele-based regression

A critical component of developing a robust allelic association test is the modelling of the Hardy–Weinberg equilibrium assumption. HWE assumes that the two alleles in a genotype are independent of each other. Thus, given the introduction of the two allele-based binary variables, G_{i1} and G_{i2} in (4), a natural approach is to use the following logistic regression,

$$logit(E(G_{i1})) = log(\frac{p_i}{1-p_i}) = \alpha + \beta G_{i2},$$

and reformulate testing of HWE as testing of the regression coefficient β. Indeed, we can show that the corresponding score test of $H_0 : \beta = 0$ closely approximates $T_{\text{HWE, Pearson}}$, the Pearson χ^2 test derived from the genotype count data (Web Appendix C).

Since our primary interest is testing (not estimation), we can also implement a Gaussian model,

$$G_{i1} = \alpha + \beta G_{i2} + \varepsilon_i, \quad \text{where } \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2). \tag{5}$$

The score test derived from this Gaussian model is in fact identical to that from the logistic model. More generally, Chen (1983) has shown that, under some regularity conditions, the score test statistics for regression models from the exponential family have identical form.

One can also show that (linearly) regressing G_{i2} on G_{i1} leads to the same conclusion. However, the differential treatment and interpretation of G_{i1} and G_{i2} is not ideal. Further, the regression framework (5) uses n alleles as the response whereas there are $2n$ alleles given a sample of n genotypes. Thus, we consider an alternative regression formulation that ‘doubles’ the sample size, with both alleles as the response.

In the revised regression, instead of using the location parameter β to represent the dependence between the two alleles, we re-parameterize it as the correlation parameter ρ in the covariance matrix to capture HWD. This model reformulation is particularly beneficial for methodology development in Section 3.3 where the regression coefficient is reserved for the primary goal of association testing. The proposed RA regression for HWE testing is

$$\begin{pmatrix} G_{i1} \\ G_{i2} \end{pmatrix} = \alpha \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \end{pmatrix}, \quad \text{where} \quad \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \end{pmatrix} \overset{iid}{\sim} N(0, \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}). \tag{6}$$

The score test statistic of testing $H_0 : \rho = 0$ is

$$T_{\text{HWE, RA}} = \frac{(\bar{g}_{12} - \bar{g}^2)^2}{\frac{1}{n}\bar{g}^2(1-\bar{g})^2} = \frac{(\hat{p}_{AA} - \hat{p}^2)^2}{\frac{1}{n}\hat{p}^2(1-\hat{p})^2} = \frac{\hat{\delta}^2}{\frac{1}{n}\hat{p}^2(1-\hat{p})^2} \sim \chi^2_1, \tag{7}$$

where $\bar{g}_{12} = \sum_i g_{i1}g_{i2}/n = n_{AA}/n = \hat{p}_{AA}$ and $\bar{g} = (\sum_i(g_{i1} + g_{i2}))/2n = (2n_{AA} + n_{Aa})/2n = \hat{p}$. Note that $\hat{\rho} = (\hat{p}_{AA} - \hat{p}^2)/(\hat{p}(1-\hat{p})) = \hat{\delta}/(\hat{p}(1-\hat{p}))$, which is a scaled estimate of HWD.

We first note that the newly developed HWE test statistic is, attractively, proportional to $\hat{\delta} = \hat{p}_{AA} - \hat{p}^2$. Interestingly, after some algebraic manipulations we can show that $T_{\text{HWE, RA}}$ in (7) is *identical* to $T_{\text{HWE, Pearson}}$ in (1),

$$\begin{aligned} T_{\text{HWE, Pearson}} &= \frac{(n_0 - n(1-\hat{p})^2)^2}{n(1-\hat{p})^2} + \frac{(n_1 - n2\hat{p}(1-\hat{p}))^2}{n2\hat{p}(1-\hat{p})} + \frac{(n_2 - n\hat{p}^2)^2}{n\hat{p}^2} \\ &= \frac{(n_2 - n\hat{p}^2)^2}{n}(\frac{1}{(1-\hat{p})^2} + \frac{2}{\hat{p}(1-\hat{p})} + \frac{1}{(\hat{p})^2}) \\ &= \frac{(\hat{p}_{AA} - \hat{p}^2)^2}{\frac{1}{n}\hat{p}^2(1-\hat{p})^2} = \frac{\hat{\delta}^2}{\frac{1}{n}\hat{p}^2(1-\hat{p})^2} \sim \chi^2_1. \end{aligned} \tag{8}$$

Remark 1. For a sample of unrelated individuals, the score test of $H_0 : \rho = 0$ based on the Gaussian regression model of (6) is identical to the classical Pearson's χ^2 test of HWE in (1) (or re-expressed in (8)) based on genotype count data, $T_{\text{HWE, RA}} = T_{\text{HWE, Pearson}}$.

This equivalence, however, is under the simplest scenario of an independent sample. For more complex data, several authors have proposed different HWE testing strategies, each addressing a specific challenge (Troendle and Yu, 1994; Bourgain et al., 2004; Lauretto et al., 2009). For example, Troendle and Yu (1994) developed a method that tests HWE across strata, while Bourgain et al. (2004) proposed a quasi-likelihood method that tests HWE in related individuals. In Section 4 we will show how the proposed regression framework (6) can be extended to derive a generalized HWE test suitable for complex data. For the moment, we still consider an independent sample but turn our attention to association analysis.

3.3 The generalized robust allele-based (RA) association test via regression

As before, we start with an independent sample of size n. For a given bi-allelic SNP, we continue to use the previous notations for the two allele-based random variables, G_{i1} and G_{i2}, $i = 1, \ldots, n$, as constructed in (4). We now also consider Y, a (categorical or continuous) phenotype of interest, and Z, an environmental factor or other covariates available; Z can be multi-dimensional but denoted as one random variable for notation simplicity but without loss of generality. The proposed RA regression for association analysis is as follows,

$$\begin{pmatrix} G_{i1} \\ G_{i2} \end{pmatrix} = (\alpha + \beta Y_i + \gamma Z_i) \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \end{pmatrix}, \quad \text{where} \quad \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \end{pmatrix} \overset{iid}{\sim} N(0, \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}). \tag{9}$$

Based on the above model, it is clear that testing $H_0 : \beta = 0$ is evaluating the relationship between the SNP and phenotype of interest while adjusting for covariate effects. The corresponding score test is

$$T_{\text{RA}} = \frac{\{\sum_{i=1}^{n} \sum_{j=1}^{2} (g_{ij} - \hat{p} - \hat{\gamma}(z_i - \bar{z}))y_i\}^2}{2(1 - \hat{\rho}_{Y,Z}^2) \sum_i (y_i - \bar{y})^2 (\hat{p}(1 - \hat{p}) + \hat{\delta})} \sim \chi_1^2, \tag{10}$$

where $\hat{p}$ and $\hat{\delta}$ are defined as before, $\bar{y}$ and $\bar{z}$ are the sample means, and

$$\hat{\alpha} = \hat{p} - \hat{\gamma}\bar{z}, \qquad \hat{\gamma} = \frac{\sum_i (g_{i1} + g_{i2}) z_i - \hat{p}\bar{z}}{\sum_i (z_i - \bar{z})^2}, \quad \text{and} \quad \hat{\rho}_{Y,Z} = \frac{\sum_i y_i z_i - \bar{y}\bar{z}}{\sqrt{\sum_i (y_i - \bar{y})^2} \sqrt{\sum_i (z_i - \bar{z})^2}}.$$

The proposed T_{RA} unifies previous methods. For example, if Y is binary and $\gamma = 0$ as in a case-control study without covariates, T_{RA} in (10) is simplified to

$$T_{\text{RA, binary, }\gamma=0} = \frac{(\hat{p}_r - \hat{p}_s)^2}{(\frac{1}{2r}+\frac{1}{2s})(\hat{p}(1-\hat{p})+\hat{\delta})}. \tag{11}$$

If we further assume HWE (i.e. let $\rho = 0$), the corresponding score test is reduced to

$$T_{\text{RA, binary, }\gamma=0,\ \rho=0} = \frac{(\hat{p}_r - \hat{p}_s)^2}{(\frac{1}{2r}+\frac{1}{2s})\hat{p}(1-\hat{p})}. \tag{12}$$

Remark 2. Under the HWE assumption and for a case-control study using an independent sample without covariates, the score test of $H_0 : \beta = 0$ based on the proposed RA regression model (9) is identical to the classical allelic test in (2), $T_{\text{RA, binary, }\gamma=0,\ \rho=0} = T_{\text{allelic}}$. In the presence of HWD, the corresponding score test has an additional correction factor $\hat{\delta} = \hat{p}_{AA} - \hat{p}^2$ for the variance estimate as compared to T_{allelic}, and $T_{\text{RA, binary, }\gamma=0} = T_{\text{allelic, Schaid}}$.

The proposed RA testing framework also generalizes. For example, T_{RA} accounts for covariate effects. T_{RA} also analyzes any phenotypes, binary or continuous, by generalizing the concept of comparing two proportions between two groups ($H_0 : p_r = p_s$) to testing regression coefficient ($H_0 : \beta = 0$). To provide additional analytical insight, consider a continuous trait and constrain the full model (9) to be without covariates. In that case, the corresponding score test statistic has the expression of

$$T_{\text{RA, }\gamma=0} = \frac{\{\sum_i((g_{i1}+g_{i2})/2-\hat{p})y_i\}^2}{\frac{1}{2}\sum_i(y_i-\bar{y})^2(\hat{p}(1-\hat{p})+\hat{\delta})}. \tag{13}$$

Thus, the generalized RA test evaluates a weighted difference between individual-level allele frequency estimate, $(g_{i1}+g_{i2})/2$, and the whole sample estimate, $\hat{p} = \sum_i(g_{i1}+g_{i2})/2n$, where the weight is an individual's trait value, y_i.

Remark 3. The proposed robust allele-based regression (9) delivers a more flexible allelic test, T_{RA} in (10), that analyzes both categorical and continuous phenotypes while accounting for covariate effects. Because the regression model is conditional on Y, the phenotype data can be subjected to Y-dependent sampling.

In hindsight, results so far may not be surprising. However, the advantages of developing the proposed RA regression framework become evident when extending allele-based association methods to more complex data such as pedigree data and data with population heterogeneity, which we investigate in the next section.

4 Complex data

4.1 Multiple populations

The classical allelic test is limited to a sample of individuals from the same population, but population heterogeneity is often present in large-scale datasets (Diaz-Papkovich et al., 2019). Intuitively, one may use a weighted average of the test statistics obtained from the individual populations. However, it is not clear how to derive the optimal weight, and it is also difficult to extend such an approach to non-discrete populations as in principal component analyses (PCA) (Reich et al., 2008).

The proposed RA regression model of (9) can naturally adjust for population effects by including population indicators, or the top principal components inferred from PCA, as part of the covariates. Here we emphasize that the potential population effects could include both difference in allele frequency and difference in Hardy–Weinberg disequilibrium between populations. The RA framework, desirably, not only models allele frequency heterogeneity through the regression coefficient γ but also accounts for HWD heterogeneity through ρ in the covariance matrix.

Without loss of generality, it is instructive to consider the simple case of a case-control study with two populations but without additional covariates. Let $Z_i = 0$ for population I and $Z_i = 1$ for population II, the corresponding RA regression model is

$$\begin{pmatrix} G_{i1} \\ G_{i2} \end{pmatrix} = (\alpha + \beta Y_i + \gamma Z_i) \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \end{pmatrix}, \quad \text{where} \quad \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \end{pmatrix} \sim N(0, \sigma_i^2 \begin{pmatrix} 1 & \rho_i \\ \rho_i & 1 \end{pmatrix}), \tag{14}$$

$\rho_i = \rho^I$ and $\sigma_i^2 = (\sigma^I)^2$ if $Z_i = 0$; $\rho_i = \rho^{II}$ and $\sigma_i^2 = (\sigma^{II})^2$ if $Z_i = 1$. Using superscripts ^{I} and ^{II} for all the other notations introduced so far, the generalized RA test of $H_0 : \beta = 0$ while accounting for population heterogeneity has the following expression,

$$T_{\text{RA, binary, 2 pop}} = \frac{\{\frac{2r^I s^I}{n^I}(\hat{p}_r^I - \hat{p}_s^I) + \frac{2r^{II}s^{II}}{n^{II}}(\hat{p}_r^{II} - \hat{p}_s^{II})\}^2}{2(\frac{r^I s^I}{n^I} + \frac{r^{II}s^{II}}{n^{II}})\{\frac{n^I}{n^I+n^{II}}(\hat{p}^I(1-\hat{p}^I) + \hat{\delta}^I) + \frac{n^{II}}{n^I+n^{II}}(\hat{p}^{II}(1-\hat{p}^{II}) + \hat{\delta}^{II})\}} \sim \chi_1^2, \tag{15}$$

where $\hat{\delta}^I = \hat{p}_{AA}^I - (\hat{p}^I)^2$ and $\hat{\delta}^{II} = \hat{p}_{AA}^{II} - (\hat{p}^{II})^2$ capture any population-specific HWD.

Finally, if evaluating HWE across multiple populations is the primary objective, we can achieve this by testing $H_0 : \rho^I = \rho^{II} = 0$ and show that the corresponding score test statistic has the following form, $T_{\text{HWE, RA, 2 pop}} = T_{\text{HWE, RA, pop I}} + T_{\text{HWE, RA, pop II}} \sim \chi_2^2$, where the expressions for $T_{\text{HWE, RA, pop I}}$ and $T_{\text{HWE, RA, pop II}}$ are given in (7). We note again the unifying feature of the proposed RA framework. For example, the test of Troendle and Yu (1994) developed specifically for testing HWE across strata has identical form as $T_{\text{HWE, RA, 2 pop}}$.

4.2 Multiple alleles

In the previous sections, we have assumed that the genetic marker under study is a bi-allelic SNP with two alleles and three unordered genotypes, the most commonly encountered genetic variation. Other types of data such as copy number of variations (CNVs) can be of interest (Jakobsson et al., 2008), but the corresponding allele-based association test has not been developed. Here we demonstrate how the RA model of (9) can be extended to derive a generalized allelic association test for multi-allelic markers, with adjustments for covariate effects and Hardy–Weinberg disequilibrium.

For a genetic marker with K different alleles, the total number of possible unordered genotypes is $K(K+1)/2$, among which $K(K-1)/2$ are heterozygotes and K are homozygotes. As in the bi-allelic marker case, a critical step in the RA methodology development is the partition of a genotype, particularly a heterozygote. Extending the partition method for a bi-allelic marker in Section 3.1, we now introduce two indicator vectors, g_{i1} and g_{i2}, where $g_{i1} = (G_{i1}^1, G_{i1}^2, \cdots, G_{i1}^{K-1})'$ and $g_{i2} = (G_{i2}^1, G_{i2}^2, \cdots, G_{i2}^{K-1})'$. $G_{i1}^l = 1$ if the first allele is l and $G_{i2}^l = 1$ if the second allele is l, for $l < K$; allele K is chosen to be the baseline without loss of generality. The partition of a homozygote $G_i = (l,l)$ is straightforward. For a heterozygote $G_i = (m,l)$, the ordering of the two alleles depends on the outcome of a Bernoulli trial, $c_i \overset{iid}{\sim} \text{Bernoulli}(1/2)$, as in the bi-allelic case of (4).

Table 2: **Allele partition of the six unordered genotypes for a genetic marker with three alleles, A, B and C.** For individual i, $g_{i1} = (G_{i1}^A, G_{i1}^B)'$ and $g_{i2} = (G_{i2}^A, G_{i2}^B)'$, denoting the allele status for the first and second allele of genotype G_i, respectively. For each heterozygous genotype, i.e. $G_i = AB$, AC or BC, the ordering of the two alleles depends on the outcome of $c_i \overset{iid}{\sim} \text{Bernoulli}(1/2)$.

Unordered	G_{i1}		G_{i2}	
Genotype, G_i	G_{i1}^A	G_{i1}^B	G_{i2}^A	G_{i2}^B
AA	1	0	1	0
AB	c_i	$1-c_i$	$1-c_i$	c_i
AC	c_i	0	$1-c_i$	0
BB	0	1	0	1
BC	0	c_i	0	$1-c_i$
CC	0	0	0	0

As an illustration, Table 2 details the allele partition of a tri-allelic marker with three possible alleles, A, B and C. The corresponding RA regression model is

$$\begin{pmatrix} G_{i1}^A \\ G_{i1}^B \\ G_{i2}^A \\ G_{i2}^B \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_1 \\ \beta_2 \end{pmatrix} Y_i + \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_1 \\ \gamma_2 \end{pmatrix} Z_i + \varepsilon_i, \quad \text{where } \varepsilon_i \overset{iid}{\sim} N(0, \begin{pmatrix} \sigma_1^2 & \delta_1 & \delta_2 & \delta_3 \\ \delta_1 & \sigma_2^2 & \delta_3 & \delta_4 \\ \delta_2 & \delta_3 & \sigma_1^2 & \delta_1 \\ \delta_3 & \delta_4 & \delta_1 & \sigma_2^2 \end{pmatrix}), \quad (16)$$

and under the null of no association, $\delta_1 = -p_A p_B$, $\delta_2 = p_{AA} - p_A^2$, $\delta_3 = \frac{1}{2} p_{AB} - p_A p_B$, and $\delta_4 = p_{BB} - p_B^2$. Testing the association between a tri-allelic marker and a phenotype trait Y is then equivalent to testing H_0: $\beta_1 = \beta_2 = 0$, and the resulting score test statistic is χ_2^2 distributed under H_0.

Here we note that for a mutli-allelic marker with K alleles, a *genotype*-based association test inherently has $(K(K+1)/2-1)$ d.f. Appropriate genotype coding can reduce the d.f. by restricting the relationships between the effects of the $K(K+1)/2$ genotypes on the phenotype, but the most parsimonious yet interpretable model is not well understood (Wang, 2011). In contrast, the proposed RA framework is allele-based with $(K-1)$ d.f., modelling the effect of each allele with the chosen baseline allele. The RA model can also be used to derive regression-based test of HWE for multi-allelic markers (Web Appendix D).

4.3 Multiple phenotypes

In settings where we are interested in testing the association between a genotype and multiple J phenotypes simultaneously, we can simply include multiple $Y_j 1$ vectors in the RA model of (9), or (16) for a multi-allelic marker, each representing one phenotype, and then test $H_0 : \beta_j = 0, \forall j \in \{1, 2, \ldots, J\}$. The corresponding score test statistic will be χ_J^2 distributed under the null. Here we re-iterate that the proposed 'reverse' regression is *allele-based*, conceptually distinct from *genotype-based* MultiPhen (O'Reilly et al., 2012) that uses an ordinal logistic regression for an independent sample.

4.4 Related individuals

We now consider a sample of n correlated individuals with known or accurately estimated pedigree structure (Dimitromanolakis et al., 2019). For notation simplicity but without loss of generality, we present the RA model for analyzing a bi-allelic marker and one phenotype of interest. Let g be a $2n \times 1$ vector of allele indicators for the n genotypes available, where $g = (g_1', g_2', \dots, g_n')'$ and $g_i = (G_{i1}, G_{i2})'$ for $i \in \{1, \dots, n\}$, following the allele-partition step as outlined in Section 3.1, and let $y = (y_1', y_2', \dots, y_n')'$, $y_i = (Y_i, Y_i)'$, $z = (z_1', z_2', \dots, z_n')'$, and $z_i = (Z_i, Z_i)'$. The generalized RA regression model for a dependent sample is,

$$g = \alpha 1 + \beta y + \gamma z + \varepsilon, \quad \text{where } \varepsilon \sim N(0, \sigma^2 \Sigma), \tag{17}$$

1 is a $2n \times 1$ vector of 1s, and Σ is a $2n \times 2n$ matrix that captures the genetic correlation between individuals as well as departure from Hardy-Weinberg equilibrium in founders. Founders are individuals that only have direct descendants or no related individuals included in the sample, and their offspring genotypes are in HWE assuming random mating (Web Appendix E).

The specification of Σ is non-trivial, where for any two individuals i and j, $\Sigma_{2(i-1)+l,\ 2(j-1)+l'}$, not only measures the genetic correlation between individual i's lth allele and individual j's l'th allele, l and $l' \in \{1,2\}$, but also accounts for potential HWD. We note that if $i = j$ and $l = l'$, $\Sigma_{2(i-1)+l,\ 2(j-1)+l'} = 1$. If $i = j$ and $l \neq l'$, $\Sigma_{2(i-1)+l,\ 2(j-1)+l'} = 0$ for a non-founder and $= \rho$ for a founder, where ρ models HWD. Finally, if $i \neq j$, $\Sigma_{2(i-1)+l,\ 2(j-1)+l'} = \phi_{i,j}(1+\rho)$, where $\phi_{i,j}$ is the kinship coefficient between the two individuals (Web Appendix F).

As an illustration, let us consider a sample of f independent sib-pairs. With a slight abuse of notations, let $\{G_{j11}, G_{j12}, G_{j21}, G_{j22}\}$ denote the the four alleles of the jth sib-pair, $j = 1, \dots, f$, where $\{G_{j11}, G_{j12}\}$ are for sibling 1 and $\{G_{j21}, G_{j22}\}$ are for sibling 2. In this case, Σ is a block diagonal matrix with

$$\Sigma_j = \begin{pmatrix} 1 & 0 & \phi(1+\rho) & \phi(1+\rho) \\ 0 & 1 & \phi(1+\rho) & \phi(1+\rho) \\ \phi(1+\rho) & \phi(1+\rho) & 1 & 0 \\ \phi(1+\rho) & \phi(1+\rho) & 0 & 1 \end{pmatrix},$$

where $\phi = 0.25$ is the kinship coefficient for a sib-pair. If we assume that there are no covariates, the score statistic of testing $H_0 : \beta = 0$ is

$$T_{\text{RA, sib-pair, } \gamma=0} = \frac{[\frac{1}{1-4\phi^2(1+\hat{\rho})^2}\{\sum_{j=1}^{f}\sum_{k=1}^{2}\sum_{l=1}^{2} y_{jk}(g_{jkl}-\bar{g}) - 2\phi(1+\hat{\rho})\sum_{j=1}^{f}\sum_{l=1}^{2}(y_{j1}(g_{j2l}-\bar{g}) + y_{j2}(g_{j1l}-\bar{g}))\}]^2}{2\bar{g}(1-\bar{g})\sum_{j=1}^{f}\{(y_{j1}-\bar{y})^2 + (y_{j2}-\bar{y})^2 - 4\phi(1+\hat{\rho})(y_{j1}-\bar{y})(y_{j2}-\bar{y})\}}, \tag{18}$$

where y_{j1} and y_{j2} are the phenotype values of the jth sib-pair, $\bar{y} = \sum_{j=1}^{f}\sum_{k=1}^{2} y_{jk}/2f$, $\bar{g} = \sum_{j=1}^{f}\sum_{k=1}^{2}\sum_{l=1}^{2} g_{jkl}/4f$, and $\hat{\rho} = \sum_{j=1}^{f}\sum_{l=1}^{2}\{(g_{j11}-\bar{g})(g_{j2l}-\bar{g}) + (g_{j12}-\bar{g})(g_{j2l}-\bar{g})\}/(\phi\bar{g}(1-\bar{g})) - 1$.

For further illustration, consider a sib-pair case-control study with all sib-pairs concordant in phenotype (i.e. r pairs of cases and s pairs of controls). In that case, (18) is reduced to

$$T_{\text{RA, sib-pair, binary-concordant, } \gamma=0} = \frac{(\bar{g}_r - \bar{g}_s)^2}{(\frac{1}{4r}+\frac{1}{4s})(1+2\phi(1+\hat{\rho}))\bar{g}(1-\bar{g})}, \tag{19}$$

where $\bar{g}_r = \sum_{j=1}^{f}\sum_{k=1}^{2}\sum_{l=1}^{2} y_{jk}g_{jkl}/4r$, $\bar{g}_s = \sum_{j=1}^{f}\sum_{k=1}^{2}\sum_{l=1}^{2}(1-y_{jk})g_{jkl}/4s$, and $\bar{g}$ and $\hat{\rho}$ are as defined above. It is compelling that the form of (19) is similar to that of the classic allelic test in (2). However, the denominator of (19) explicitly adjusts for the inherent genetic correlation between the sibling alleles through ϕ, as well as any potential HWD through $\hat{\rho}$.

Remark 4. The proposed robust allele-based regression (9) can be naturally generalized to analyze multiple populations and phenotypes. The RA model (9) can be further generalized to model (16) to analyze genetic markers with more than two alleles, and to model (17) to analyze pedigree data. With a sample of related individuals, the Σ matrix decomposes into two parts that explicitly model the genetic correlation between individuals and the departure from HWE in the founder generation.

5 Empirical evidence

5.1 Simulation studies

To numerically demonstrate the robustness of T_{RA} to HWD as compared with T_{allelic}, we simulated a case-control study with an independent sample of 1,000 cases and 1,000 controls. The minor allele frequency was $p = 0.2$ or 0.5 for the minor allele A. The amount of HWD as measured by $\delta = p_{AA} - p^2$ ranged from the minimum of $-p^2$ to the maximum of $p(1-p)$. Then $p_{AA} = \delta + p^2$ and $p_{Aa} = 2(p - p_{AA})$, and $(n_{aa}, n_{Aa}, n_{AA}) \sim \text{Multinomial}\{n, (1 - p_{Aa} - p_{AA}, p_{Aa}, p_{AA})\}$. For power evaluation at $\alpha = 0.05$, we assumed an additive model with disease prevalence $K = 0.1$ and penetrance $P(Y = 1|G = aa) = f_0 = 0.09$; $P(Y = 1|G = AA) = f_2 = (K - f_2 p)/(1-p)$ and $P(Y = 1|G = Aa) = f_1 = (f_0 + f_2)/2$. The empirical type 1 error results in Figure 2(a) and 2(b) confirm the theoretical results in Figure 1: T_{allelic} is not robust against HWD while the proposed T_{RA} is accurate across the whole range of HWD values. Further, the empirical power results in Figures 2(c) and 2(d) highlight the fact that the classical allelic test could have reduced power when the number of homozygotes AA is fewer than what is expected under the HWE assumption (i.e. $\delta < 0$), which is not well acknowledged in the existing literature.

5.2 Application 1 - revisit the study of Wittke-Thompson et al. (2005)

For the purpose of studying Hardy–Weinberg disequilibrium in case-control studies, Wittke-Thompson et al. (2005) identified 60 SNPs from 41 case-control association studies. Focusing on association

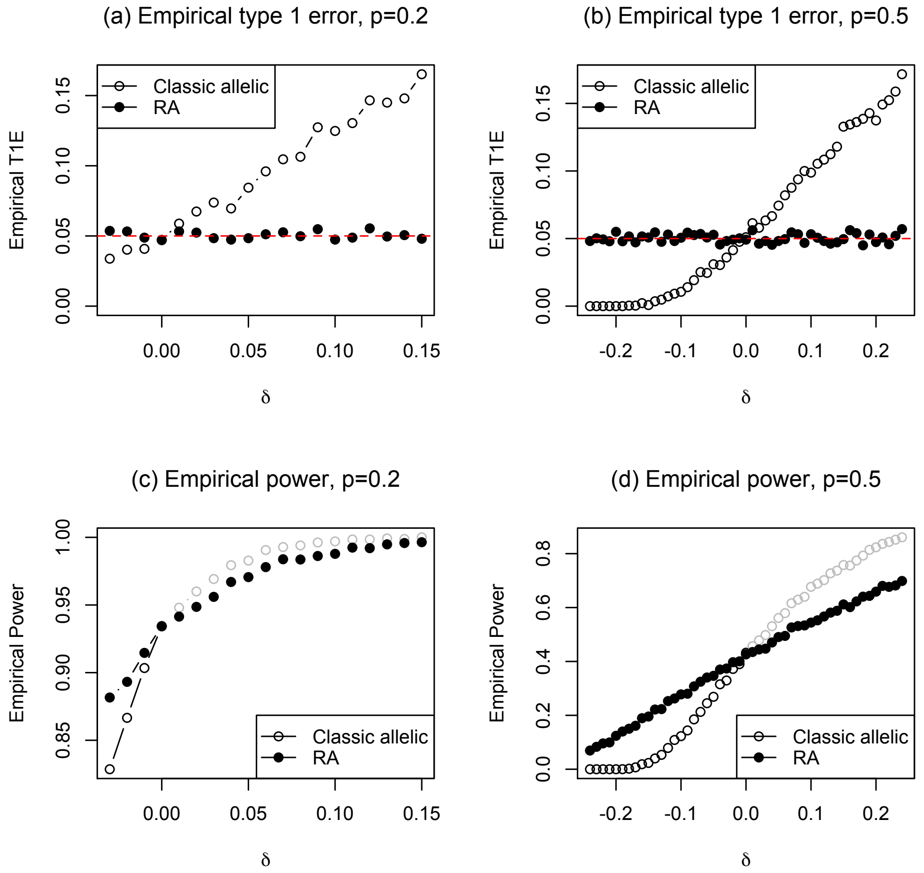

Figure 2: **Empirical type 1 error rate and power of the classical allelic association test and the proposed robust allelic (RA) test at the nominal level of $\alpha = 0.05$.** Note that when $\delta > 0$, the classical allelic test has inflated type 1 error rate as shown in (a) and (b), so the corresponding power in (c) and (d) is not meaningful and shown in a lighter shade. Also note that the HWD measure δ is bounded by the minor allele frequency p, $-p^2 \leq \delta \leq p(1-p)$.

analyses of these 60 bi-allelic markers, we compared T_{allelic} with the proposed T_{RA} while considering HWD at each SNP. Figure 3 contrasts $-\log_{10}(p\text{-values})$ of the two methods, stratified by if there was an excess ($\hat{\delta} > 0$; unfilled triangles) or lack ($\hat{\delta} < 0$; filled triangles) of the homozygotes AA with A being the minor allele.

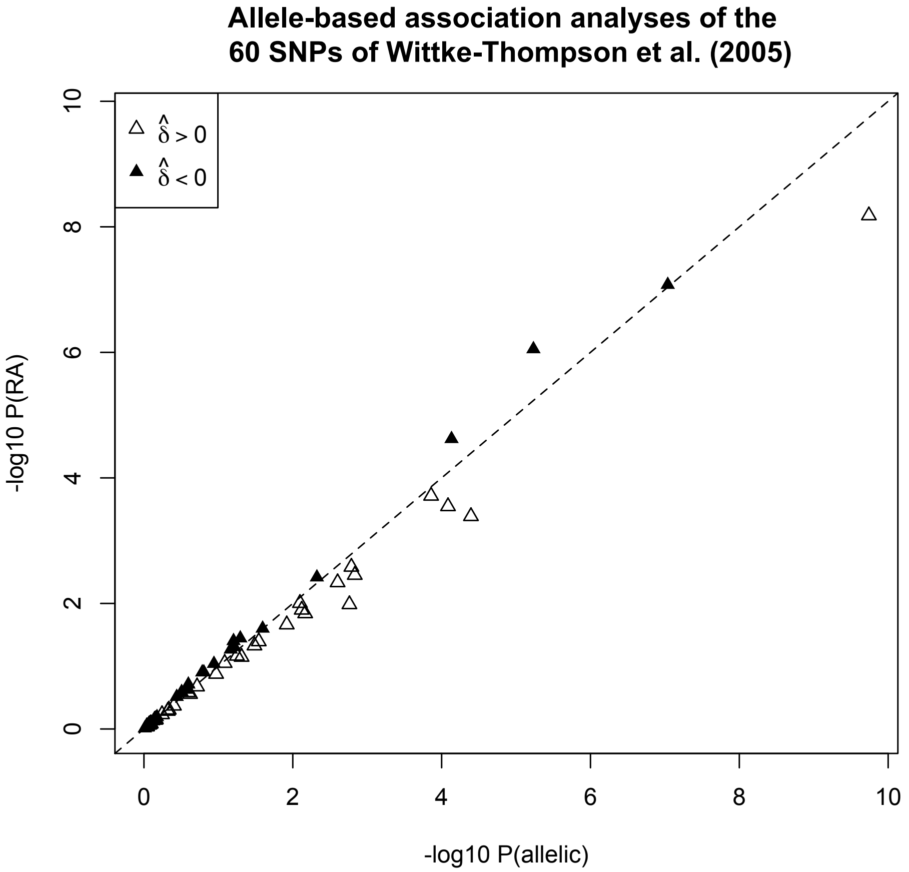

Figure 3: **Results of application 1.** Allele-based association tests of the 60 SNPs identified in Wittke-Thompson et al. (2005), contrasting the proposed RA method, T_{RA} in (11), with the classical allelic test, T_{allelic} in (2). Unfilled triangles are for SNPs with $\hat{\delta} > 0$ (T_{allelic} having inflated type 1 error), and filled triangles are for SNPs with $\hat{\delta} < 0$ (T_{allelic} having deflated type 1 error); see Figure 1 for theoretical results and Figure 2 for simulation results regarding type 1 error control of the two methods.

As anticipated based on the theoretical results in Figure 1 and simulation results in Figure 2, for SNPs with $\hat{\delta} > 0$, T_{allelic} can appear to be more powerful than the proposed T_{RA}. For example, for the most significant SNP, p-value$_{\text{allelic}} = 1.82 \times 10^{-10}$ and p-value$_{\text{RA}} = 6.60 \times 10^{-9}$. However, $\hat{\delta} = 0.052 > 0$ with p-value$_{\text{HWE}} = 3.09 \times 10^{-4}$. Thus, the result of T_{allelic} is not accurate for this SNP. In contrast, for the third most significant SNP, $\hat{\delta} = -0.031 < 0$ and p-value$_{\text{HWE}} = 0.040$. In that case, T_{allelic} is conservative while the proposed T_{RA} is not only robust but also more powerful, where p-value$_{\text{allelic}} = 5.84 \times 10^{-6}$ and p-value$_{\text{RA}} = 8.86 \times 10^{-7}$.

5.3 Application 2 - a cystic fibrosis (CF) gene modifier study

To demonstrate the generalizability of the proposed RA framework, we applied T_{RA} to jointly analyze two phenotypes using a sample of related individuals from the Canadian cystic fibrosis (CF) gene modifier study (Sun et al., 2012; Corvol et al., 2015). The two phenotypes of interest are lung function (a quantitative trait (Taylor et al., 2011)) and meconium ileus (MI, a binary trait (Gong et al., 2019)). Among the sample of $2,540$ CF subjects, $2,420$ are singletons and 60 independent sib-pairs. For completeness, we first analyzed each phenotype individually using the proposed *allele-based* RA framework, and we compared the results with the traditional *genotype-based* method via (generalized) linear mixed models (LMM or GLMM). We then analyzed both phenotypes jointly using T_{RA}.

Figures 4(a) and 4(b) show that results of genotype-based and allele-based methods are largely consistent; see Section 6 for additional discussion. Interestingly, for the most significant SNP associated with MI in Figure 4(b), p-value of T_{RA} is 2.62×10^{-6}, slightly smaller than 7.80×10^{-6} of the genotype-based GLMM method. In addition, the proposed T_{RA} method can jointly analyzed both phenotypes and appears to identify SNPs that have p-values several orders of magnitude smaller than that from studying one phenotype at a time, as shown in Figures 4(c) and 4(d). However, these results do not reach genome-wide significance and establishing true association requires additional analyses.

Table 3 summarizes the association results for previously reported and replicated SNPs associated with CF lung function (Corvol et al., 2015) and MI association (Sun et al., 2012). Note that

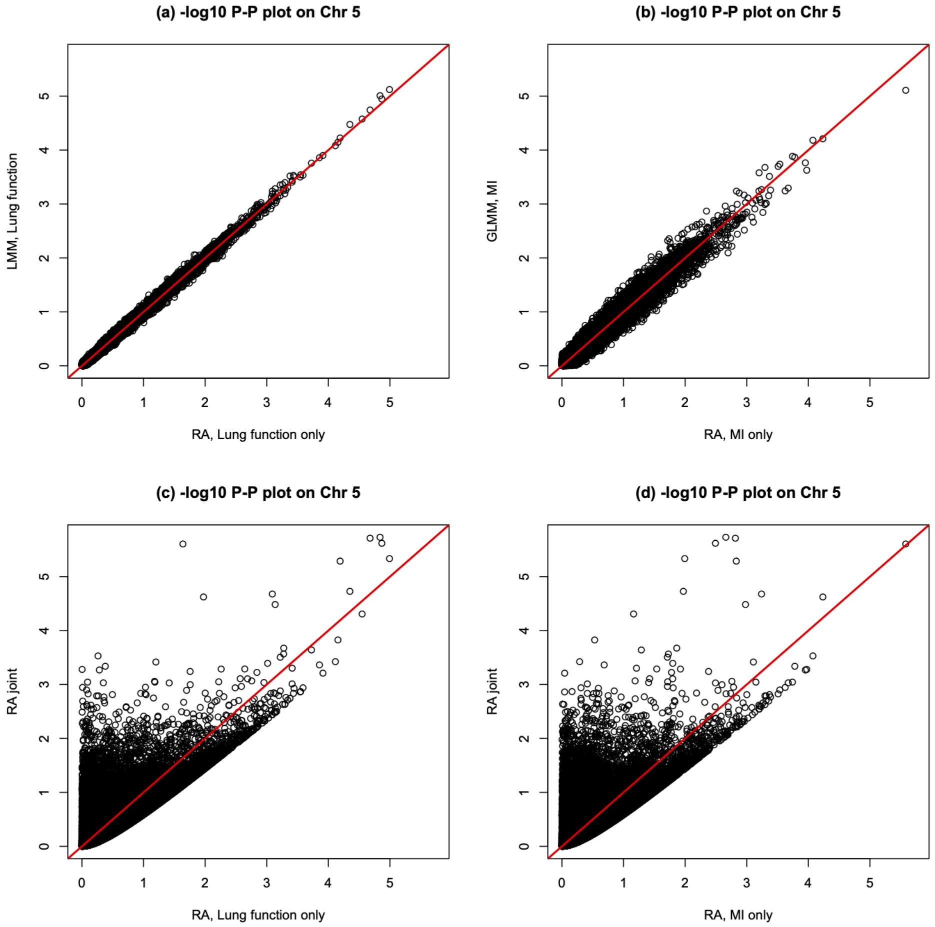

Figure 4: **Results of application 2 - Chromosome 5-wide.** Genetic association studies of lung function and meconium ileus of $34,378$ bi-allelic markers on chromosome 5, using a sample of $2,540$ individuals with cystic fibrosis of which $2,420$ are singletons and 120 are from 60 sib-pairs. LMM and GLMM are *genotype-based* association analyses based on, respectively, linear mixed model for a continuous trait (i.e. lung) and generalized LMM for a binary trait (i.e. MI), and RA is the proposed allele-based association method that can also jointly analyze multiple traits using a sample of related individuals. Genome-wide results are shown in Web Figure 1.

the p-values in Table 3 differ from those in Sun et al. (2012) and Corvol et al. (2015), because the analyses here only included the Canadian sample and individuals with both phenotypes measured. For all the SNPs in Table 3, the proposed RA test yields slightly larger $-\log_{10}(p\text{-values})$ than LMM or GLMM, suggesting that the allele-based method has the potential to be more powerful than the traditional genotype-based approach. The joint RA analysis of the two phenotypes did not

Table 3: **Results of application 2 - Previously reported SNPs.** Other details see legend to Figure 4.

Top CF lung function associated SNPs from Corvol et al. (2015)				
Chr	SNP	Lung function only		MI and lung function jointly
		$-\log_{10} p_{\text{LMM}}$	$-\log_{10} p_{\text{RA}}$	$-\log_{10} p_{\text{RA, joint}}$
3	rs2246901	3.21	**3.25**	2.63
5	rs3749615	3.25	**3.27**	3.57
6	rs2395185	6.65	**6.77**	6.08
11	rs10466455	5.84	**5.86**	4.84
Top CF meconium ileus associated SNPs from Sun et al. (2012)				
Chr	SNP	MI only		MI and lung function jointly
		$-\log_{10} p_{\text{GLMM}}$	$-\log_{10} p_{\text{RA}}$	$-\log_{10} p_{\text{RA, joint}}$
1	rs4077468	5.34	**5.47**	4.87
1	rs7512462	4.54	**4.82**	4.56
1	rs7419153	3.68	**4.07**	3.35
1	rs12047830	3.10	**3.20**	2.63

lead to more significant results; this is not surprising because these SNPs were selected based on the single-phenotype analyses.

6 Discussion

The classical allele-based association test, examining the difference in allele frequency of a bi-allelic genetic marker between cases and controls, is intuitive and locally most powerful. As pointed out by Sasieni (1997), for a sample of n individuals the allelic test ‘doubles’ the sample size by considering $2n$ alleles instead of n genotypes. However, the work of Sasieni (1997) also highlighted the sensitivity of the allelic test to the assumption of Hardy–Weinberg equilibrium. The subsequent development of Schaid and Jacobsen (1999) based on improving variance estimate is effective, but its application is restricted to case-control studies using independent samples and without covariates.

Here we developed a novel, robust allele-based (RA) regression framework that regresses the individual alleles on the phenotype of interest and covariates if available, generalizing the concept of comparing allele frequencies for more complex data. Utilizing the earlier work by Chen (1983), the proposed regression relies on the Gaussian model of (9) that (i) leads to a valid allelic association test through testing the regression coefficient β, (ii) analyzes either a binary or a continuous phenotype, or both, where the phenotype data can be subjected to Y-dependent sampling, (iii) adjusts for covariate effects, including population heterogeneity, through additional regression coefficient γ, (iv) accounts for sample correlation through kinship coefficient ϕ in the covariance matrix Σ, and (v) explicitly models potential departure from HWE through ρ in Σ; see *Remark 3*. Appealingly, the generalized allelic association test also unifies previous methods; see *Remark 2*.

The pivotal stage of this work is designing the two allele-based random variables, G_{i1} and G_{i2}, and leveraging the regression framework in new settings. The idea of reformulating an existing test statistic as a regression to facilitate method extension is not new. In their Reader Reaction to the generalized non-parametric Kurskal-Wallis test of Acar and Sun (2013) for handling group uncertainty, Wu and Guan (2015) presented "*a rank linear regression model and derived the proposed GKW statistic as a score test statistic*". More recently, Soave and Sun (2017) showed that by first reformulating the original Levene's test, testing for variance heterogeneity between k groups in an independent sample without group uncertainty, as a two-stage regression, the extension to more complex data is more straightforward.

In our study, the correct representation of G_{i1} and G_{i2} is critical. In Section 3.1, we have argued that splitting the n_{Aa} heterozygotes into exact halves (G_{i1}^* and G_{i2}^*) reduces the variation inherent in a randomly selected allele. Looking at it from a different angle, assume that there are only two individuals with Aa. In that case, if G_{11}^* is one for individual 1 then G_{21}^* must be zero for individual 2, introducing additional dependence between alleles beyond the underlying kinship relationship and HWD. In contrast, if G_{11} is one then G_{21} is yet to be independently determined by the outcome of tossing a fair coin as defined in (4).

The concept of 'reverse' regression has also been explored before, focusing on regressing *genotype* on phenotype, notably by O'Reilly et al. (2012) for joint analyses of multiple phenotypes. The corresponding MultiPhen method uses an ordinal logistic regression for the three genotype groups and then applies a likelihood ratio test. Although MultiPhen does not require the assumption of HWE, its application is limited to independent samples and bi-allelic markers.

Another stream of genotype-based 'reverse' or retrospective approach started with the quasi-likelihood method of Thornton and McPeek (2007) for case-control association testing with related individuals. The method first defines $X_i = G_i/2 \in \{0, 1/2, 1\}$, then links the mean of X_i with Y_i via a logit transformation and uses the kinship coefficient matrix as the covariate matrix of X_i, and finally obtains a quasi-likelihood score test. Subsequently, Feng (2014) and Feng et al. (2011) extended the method of Thornton and McPeek (2007) to a quasi-likelihood regression model that can incorporate multiple phenotypes. We note that although $X_i = G_i/2$ was interpreted as the allele frequency per individual i by the previous work, the quasi-likelihood score test is fundamentally a genotype-based association method. Further, the use of the kinship matrix alone as the covariance matrix requires the assumption of HWE. Recently, we showed that genotype-based 'reverse' regression can be specified in a robust fashion that guards against HWD in related individuals (Zhang and Sun, 2019).

Most existing family-based association studies rely on the $Y-G$ prospective regression framework via LMM or GLMM (Eu-Ahsunthornwattana et al., 2014). For the application study in Section 5.3, we applied both the proposed RA method and LMM (for the continuous CF lung function) and GLMM (for the binary meconium ileus status). Although there are differences in the (single-phenotype) analyses (Figures 4(a) and 4(b)), results are remarkably consistent. Interestingly, in the simplest case of an independent sample with no covariates, we can show analytically that the corresponding RA test statistic has identical form as that derived from genotype-based prospective regression model, as well as that from the non-parametric trend test (Web Appendix G). The similarity with the existing methods indirectly confirms the validity of the proposed approach but does not take away the contributions of this work. In particular, unlike LMM and GLMM, the proposed 'reverse' regression can analyze more than one phenotype at a time as shown in Figures 4(c) and 4(d).

One of the challenges related to the proposed framework is the interpretation of parameter estimate for β even though its corresponding hypothesis testing is valid. Thus, we emphasize that the method developed here is tailored for variant detection, providing a statistically efficient and computationally fast way for genome-wide association scans. Another difficulty present in any 'reverse' regression approach is the modelling and interpretation of gene-gene or gene-environment interactions. It is also not clear how to perform allelic association test for X-chromosomal variants; see Chen et al. (2018) for genotype-based association methods. However, the proposed framework is flexible and promising in a number of other ways.

For example, the inclusion of parameter ρ in the RA model (9) is advantageous for both method comparison and further development. In the absence of Y and Z and sample correlation, the score test derived from the reduced model is equivalent to the traditional Pearson χ^2 test of HWE using a sample of independent genotype observations; see *Remark 1*. For more complex data, instead of developing individual remedies addressing specific challenges, the proposed method provides a principled approach for extensions. For example, we have shown in Section 4.1 that by introducing a population indicator we can derive a HWE test across populations. Similarly, testing $H_0 : \delta_2 = \delta_3 = \delta_4 = 0$ using model (16) in Section 4.2 leads to a HWE test for tri-allelic markers. Finally, using the generalized RA model (17) in Section 4.4, we can develop a score test of HWE that naturally accounts for sample correlation present in pedigree data.

In terms of association testing, the value of introducing ρ in the regression model is two fold. First, if there is a strong prior evidence for HWE, we can restrict ρ to be zero and establish a locally most powerful score test. Second, for the special case of a case-control study, Song and Elston (2006) and Wang and Shete (2010) have argued that departure from HWE in the case group provides additional association evidence. However, their methods are ad-hoc. For example, the method of Song and Elston (2006) first conducts genotype-based association test and Pearson χ^2 test of HWE separately, then aggregates the two (dependent) tests by a weighted sum, and finally evaluates the statistical significance via simulations. The proposed RA regression framework offers a conceivable approach to directly incorporate group-specific ρ into association inference, which we will explore as future work.

7 Acknowledgment

The authors thank Dr. Lisa J. Strug and her lab for providing the cystic fibrosis application 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

Acar, E. F., and L. Sun, 2013: A generalized kruskal–wallis test incorporating group uncertainty with application to genetic association studies. *Biometrics*, **69 (2)**, 427–435.

Bourgain, C., M. Abney, D. Schneider, C. Ober, and M. S. McPeek, 2004: Testing for hardy-weinberg equilibrium in samples with related individuals. *Genetics*, **168 (4)**, 2349–2361.

Bourgain, C., and Coauthors, 2003: Novel case-control test in a founder population identifies p-selectin as an atopy-susceptibility locus. *The American Journal of Human Genetics*, **73 (3)**, 612–626.

Chen, B., R. V. Craiu, L. J. Strug, and L. Sun, 2018: The x factor: A robust and powerful approach to x-chromosome-inclusive whole-genome association studies. *arXiv preprint arXiv:1811.00964*.

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

Corvol, H., and Coauthors, 2015: Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis. *Nature communications*, **6**, 8382.

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.

Diaz-Papkovich, A., L. Anderson-Trocmé, and S. Gravel, 2019: Umap reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS genetics*, **15 (11)**.

Dimitromanolakis, A., A. D. Paterson, and L. Sun, 2019: Fast and accurate shared segment detection and relatedness estimation in un-phased genetic data via truffle. *The American Journal of Human Genetics*, **105 (1)**, 78–88.

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.

Gong, J., and Coauthors, 2019: Genetic association and transcriptome integration identify contributing genes and tissues at cystic fibrosis modifier loci. *PLoS genetics*, **15 (2)**, e1008 007.

Hardy, G. H., and Coauthors, 1908: Mendelian proportions in a mixed population. *Science*, **28 (706)**, 49–50.

Hill, W. G., M. E. Goddard, and P. M. Visscher, 2008: Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics*, **4 (2)**, e1000 008.

Jakobsson, M., and Coauthors, 2008: Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, **451 (7181)**, 998–1003.

Lauretto, M. S., F. Nakano, S. R. Faria Jr, C. A. Pereira, and J. M. Stern, 2009: A straightforward multiallelic significance test for the hardy-weinberg equilibrium law. *Genetics and Molecular Biology*, **32 (3)**, 619–625.

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.

Reich, D., A. L. Price, and N. Patterson, 2008: Principal component analysis of genetic data. *Nature genetics*, **40 (5)**, 491.

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., J. P. Sinnwell, and G. D. Jenkins, 2012: Regression modeling of allele frequencies and testing hardy weinberg equilibrium. *Human heredity*, **74 (2)**, 71–82.

Soave, D., and L. Sun, 2017: A generalized levene's scale test for variance heterogeneity in the presence of sample correlation and group uncertainty. *Biometrics*, **73 (3)**, 960–971.

Song, K., and R. C. Elston, 2006: A powerful method of combining measures of association and hardy–weinberg disequilibrium for fine-mapping in case-control studies. *Statistics in Medicine*, **25 (1)**, 105–126.

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.

Taylor, C., and Coauthors, 2011: A novel lung disease phenotype adjusted for mortality attrition for cystic fibrosis genetic modifier studies. *Pediatric pulmonology*, **46 (9)**, 857–869.

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.

Troendle, J., and K. Yu, 1994: A note on testing the hardy-weinberg law across strata. *Annals of human genetics*, **58 (4)**, 397–402.

Wang, J., and S. Shete, 2010: Using both cases and controls for testing hardy-weinberg proportions in a genetic association study. *Human Heredity*, **69 (3)**, 212–218.

Wang, T., 2011: On coding genotypes for genetic markers with multiple alleles in genetic association study of quantitative traits. *BMC genetics*, **12 (1)**, 82.

Weinberg, W., 1908: On the demonstration of heredity in man. *(1963) Papers on human genetics*.

Weir, B., 1996: Genetic analysis ii. *Sinauer: Sunderland, MA*.

Wittke-Thompson, J. K., A. Pluzhnikov, and N. J. Cox, 2005: Rational inferences about departures from hardy-weinberg equilibrium. *The American Journal of Human Genetics*, **76 (6)**, 967–986.

Wu, B., and W. Guan, 2015: Reader reaction on the generalized kruskal–wallis test for genetic association studies incorporating group uncertainty. *Biometrics*, **71 (2)**, 556–557.

Zhang, L., and L. Sun, 2019: On a unifying "reverse" regression for robust association studies and allele frequency estimation with related individuals. *bioRxiv*, 470328.