Regression Modeling of Allele Frequencies and Testing Hardy Weinberg Equilibrium

Daniel J. Schaid  Jason P. Sinnwell  Gregory D. Jenkins
Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, Minn., USA

Key Words
Logistic regression · Over-dispersion quasi-likelihood · Residual correlation

Abstract
Background/Aims: Tests for whether observed genotype proportions fit Hardy Weinberg Equilibrium (HWE) are widely used in population genetics analyses, as well as to evaluate quality of genotype data. To date, all methods testing for HWE require subjects to be classified into discrete categories, yet it is becoming clear that the distribution of allele frequencies tends to be smooth over geographic regions.

Methods: To evaluate the HWE assumption, we develop new approaches to model allele frequencies as functions of covariates, and use these models to test whether there is residual correlation between the two alleles of subjects; lack of residual correlation supports the null hypothesis of HWE, but conditional on how the covariates influence the allele frequencies.

Results: By simulations, we illustrate that a simple statistical test of residual correlation of alleles adequately controls the type I error rate, while maintaining power that is comparable to standard tests for HWE.

Conclusion: Our approach can be implemented in standard software, enabling more flexible and powerful ways to evaluate the association of covariates with allele frequencies and whether these associations ‘explain’ departures from HWE when the covariates are ignored, opening new strategies to evaluate the quality of genotype data generated by next-generation sequencing assays.

Introduction

Tests for whether observed genotype proportions fit Hardy Weinberg Equilibrium (HWE) have received much attention in population genetics and more recently in analysis of large-scale genotype data. Departure from HWE indicates that the two alleles within subjects are correlated, and this correlation can be parameterized in terms of a fixation index [1]. When evaluating the quality of single nucleotide polymorphisms (SNPs) measured in genome-wide association studies, it is common practice to use tests for HWE as a way to filter out poor-quality SNPs, expecting that extreme departures from HWE are due to poorly called genotypes. An example might be a SNP that has common alleles, yet no heterozygotes. In some studies, however, an unusually large number of SNPs might depart from HWE, indicated by an overall
departure of the distribution of the observed HWE statistics from the large-sample theoretical $\chi^2$ distribution, displayed as a quantile-quantile plot. When this occurs, exploratory analyses might be used to determine whether major factors, such as institutional source of DNA, or sample preparation, could explain departures from HWE. This becomes challenging when the factors are continuous variables, or when there are many potential factors, because all existing tests for HWE are based on counts of genotypes, and the genotype counts in turn are assumed to be determined by allele frequencies that are constant. If, in fact, the allele frequencies are continuous functions of explanatory factors, then testing departures from HWE becomes challenging.

We developed new approaches to model allele frequencies as functions of covariates and used these models to test whether there is residual correlation between the two alleles of subjects; lack of residual correlation supports the null hypothesis of HWE, but conditional on how the covariates influence the allele frequencies. In our developments, we used both theory and simulations to refine our approach. A significant challenge was to allow for small allele frequencies, which can invalidate large-sample statistical tests. Nonetheless, simulations and theoretical insights showed that a simple strategy can be used with widely available software.

**Methods**

**Background**

To simplify notation and the development of our models, we focus on a locus with only two alleles, which cover the majority of SNPs used in studies today. We also focus on samples with unrelated subjects. We later remark how our models could be extended to more than two alleles, and to pedigree-based analyses. But, before we describe our regression approach, we first review some basic statistical foundations for genotype probabilities when there is Hardy Weinberg Disequilibrium (HWD). Suppose that a locus has the two alleles $A$ and $a$, with allele frequencies $p$ and $q = 1 - p$. Although not necessary, one could view $A$ as the ‘major’ allele and $a$ as the ‘minor’ allele, where $q < p$. As described by Weir [2], the three genotype probabilities are determined by the allele frequencies and the disequilibrium coefficient, $D$:

- $P_{AA} = p^2 + D$
- $P_{Aa} = 2pq - 2D$
- $P_{aa} = q^2 + D$

Now, to motivate our modeling, recognize that each genotype can be represented by two allele indicators, $y_1$ and $y_2$, each of which has a value of 0 (for the A allele) or 1 (for the a allele). The joint distribution of the pair $(y_1, y_2)$ is represented by the above genotype probabilities. It is important to recognize that $(y_1, y_2)$ is unordered, because we typically do not know the parental origin of each allele, which is why the probability for the heterozygote, $P_{Aa}$, has a factor of 2. So, to describe the joint distribution for $(y_1, y_2)$, we ‘split’ $P_{Aa}$ equally, as illustrated in table 1. From the joint distribution of $(y_1, y_2)$, it is straightforward to show that $E[y_1] = E[y_2] = p$ and $Var(y_1) = Var(y_2) = pq$, as expected. More important for our discussion, $Cov(y_1, y_2) = D$. This emphasizes that departures from HWE do not influence estimates of allele frequencies, but rather influence the covariance of the two alleles within subjects. We use this perspective to formulate regression models with two aims: (1) to determine which covariates influence allele frequencies and (2) to test the null hypothesis that the residual correlation of allele indicators is zero, supporting HWE.

**Regression Models for Subject Level Covariates**

When covariates are measured at the subject level, we can model the influence of covariates on allele frequencies using standard software for generalized linear models, such as glm in the R software environment. For this approach, let $g$ denote the coded genotype for subject $i$ having values of 0, 1, 2, for the count of the number of minor alleles. A regression model for the influence of covariate vector $x_i$ on the allele frequency can be fit by converting $g_i$ to a proportion, $y_i = g_i/2$, using a weight of 2 (for the 2 alleles per subject), and using quasi-binomial to allow for estimation of the scale parameter in logistic regression; in ordinary logistic regression with binomial data, the scale parameter has value 1. By using quasi-likelihood, we allow for residual correlation between alleles, which translates to over-dispersion captured by estimating the scale parameter in the quasi-likelihood. This allows one to test whether covariates have a statistically significant association with allele frequencies, adjusting for the possibility of residual correlations. An example of the method in R is given below.

```
glm(y ~ x, weight = rep(2, length(y)), family = quasi-binomial).
```

A few remarks on alternative regression strategies are worthwhile. Simulations (not shown) illustrated that type I error rates for testing the regression coefficients when ignoring the over-dispersion can be inflated. We also evaluated the utility of multinomial regression, a possible strategy for genotype categories with three levels. Yet, we found several limitations of multinomial regression. First, for each covariate in the regression model, there are two regression coefficients, because of the three levels of $g$ (one level is treated as a baseline). This means that multinomial regression will have twice the number of covariates as logistic regression, possibly weakening power to detect the effect of covariates. Second, multinomial regression does not distinguish between the effects of covariates on allele frequencies versus their effects on

---

**Table 1. Joint distribution of allele indicators, $(y_1, y_2)$**

<table>
<thead>
<tr>
<th>Genotype</th>
<th>$y_1$</th>
<th>$y_2$</th>
<th>$P(y_1, y_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>0</td>
<td>0</td>
<td>$p^2 + D$</td>
</tr>
<tr>
<td>Aa</td>
<td>0</td>
<td>1</td>
<td>$pq - D$</td>
</tr>
<tr>
<td>aA</td>
<td>1</td>
<td>0</td>
<td>$qp - D$</td>
</tr>
<tr>
<td>aa</td>
<td>1</td>
<td>1</td>
<td>$q^2 + D$</td>
</tr>
</tbody>
</table>
the disequilibrium coefficient, $D$. To see this, consider a simple example of two subpopulations with the same allele frequencies, yet different values of the disequilibrium coefficient, $D$. Because of different values of $D$, they have different genotype frequencies. Using a covariate to indicate the subpopulations, the multinomial regression will determine the covariate to be associated with the genotype frequencies, yet logistic regression would not find the covariate to be associated with the allele frequencies. Because our objective is to include only covariates that influence allele frequencies, so that in later steps we can test whether there is residual correlation between alleles, we advocate use of $\text{glm}$ for logistic regression with quasi-binomial family option.

**Tests for Uncorrelated Residuals**
To test the null hypothesis that the residuals are uncorrelated, we considered two general approaches: (1) Fisher’s transformation for the correlation coefficient of the residuals, and (2) various $C(a)$ statistics to detect departure from binomial variance. It is important to understand that testing correlation of residuals for a subject’s alleles is equivalent to testing for over-dispersion in logistic regression, and that over-dispersion can be caused by a number of factors, including omission of covariates, miss-modeling of covariates (e.g. assuming linear influence when quadratic is needed), the wrong link function in $\text{glm}$, or the existence of large outliers. Hence, before testing for uncorrelated residuals, or binomial variance, care is needed to be sure that measured covariates are adequately modeled and large outliers are adequately handled.

**Fisher’s Transformation**
We explored the utility of Fisher’s transformation to test whether a correlation coefficient differs from zero, because, as we will show, it generalizes the usual $\chi^2$ goodness-of-fit (GOF) test for HWE. Fisher’s transformation is based on the estimated correlation of the allele indicator residuals, $r_{ij} = (y_{ij} - p_i)$, where $y_{ij}$ is the binary allele indicator for the $j$-th allele ($j = 1, 2$) of subject $i$, and $p_i$ is the fitted value from the quasi-binomial $\text{glm}$. However, caution is needed on how to compute the correlation coefficient, because heterozygotes are arbitrarily coded as having a pair of allele indicators $(1, 0)$. In truth, the order of allele indicators should be based on parental origin, and hence randomly $(1, 0)$ versus $(0, 1)$. Or, two pairs of allele indicators could be assigned to each heterozygote, with weights of 0.5 each. Using the order $(1, 0)$ for all heterozygotes falsely introduces correlation. This is because the marginal mean and variance of $r_{11}$ and $r_{12}$ wrongly differ, simply because of the arbitrary order of $(1, 0)$ for heterozygotes. To get around this, we compute the sample moment for the residual cross product by

$$m_{12} = \frac{1}{n} \sum_{i=1}^{n} r_{1i} r_{2i},$$

but compute the first and second sample moments of $r$ based on the pool of $r_{1i}$ and $r_{2i}$:

$$m_1 = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{2} r_{ij}^2$$

$$m_2 = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{2} r_{ij}$$

From the above sample moments, we compute $\text{Cov}(r_1, r_2) = m_{12} - m_1^2$ and $\text{Var}(r_1) = \text{Var}(r_2) = m_2 - m_1^2$, to then compute $\rho = \text{Cov}(r_1, r_2)/\text{Var}(r)$.

We propose using Fisher’s transformation for the correlation coefficient to test $H_0: \rho = 0$, resulting in the statistic

$$z(\rho) = \sqrt{n-3} \log\left(\frac{1 + \rho}{1 - \rho}\right)/2.$$  

Fisher showed that under the null hypothesis, and for a bivariate normal distribution, $z$ has a standard normal distribution. We recognize that the assumption of a bivariate normal distribution might seem problematic for our residuals based on logistic regression. However, without covariates, Fisher’s $z$ statistic is nearly identical to the usual $\chi^2$ GOF statistic commonly used to test HWE. To see this, suppose that $o_i$ is the observed count of the number of subjects with genotype $i$, for $i = 0, 1, 2$. The usual GOF statistic, which compares each of the three observed and HWE-expected counts, can be expressed as

$$T_{\text{GOF}} = n\left(\frac{4o_1o_2 - o_0^2}{(2o_1 + o_2)(2o_1 + o_2)}\right).$$

An alternative way to derive $T_{\text{GOF}}$ is based on binary allele indicators, $a_1$ and $a_2$. By creating two pairs for heterozygotes $[(a_1 = 1, a_2 = 0), (a_1 = 0, a_2 = 1)]$, each with weight 1/2, one can create a $2 \times 2$ table to count allele indicators, resulting in table 2. Applying Pearson’s $\chi^2$ statistic for independence to the counts in table 2 results in $T_{\text{GOF}}$. Now, for two binary variables, Pearson’s statistic is also well known to be equal to $T_{\text{GOF}} = np^2$, where $p$ is the correlation coefficient for the binary variables. This illustrates how the usual GOF test for HWE depends on the correlation of allele indicators.

Because Fisher’s $z$ has a standard normal distribution under the null hypothesis, $z^2 \sim \chi_1^2$, the same as the distribution of $T_{\text{GOF}}$. Using a second-order Taylor series expansion of $z(\rho^2)$ about $\rho = 0$ results in $z(\rho^2) = (n-3)\rho^2$, close to $T_{\text{GOF}} = np^2$. Because of these close statistical properties of the usual GOF test for HWE with Fisher’s $z$ statistic for correlation, we evaluated Fisher’s method to test for residual correlations among allele indicators after adjusting for covariates.

**C(a) Statistics**
Because Fisher’s $z$ statistic is based on a heuristic approach and does not account for estimating regression coefficients for the adjusting covariates, we evaluated alternative approaches based on more rigorous theory – $C(a)$ statistics for testing departure from binomial variation. Because each subject has two alleles, we can view each subject as representing a binomial random sample of

<table>
<thead>
<tr>
<th>$y_1$</th>
<th>$y_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$o_0$</td>
</tr>
<tr>
<td>1</td>
<td>$o_1/2$</td>
</tr>
<tr>
<td></td>
<td>$o_2$</td>
</tr>
</tbody>
</table>

**Table 2. Genotype counts based on $2 \times 2$ table of allele indicators**
size 2, with binomial probability the predicted allele frequency based on subject-specific covariates. Without covariates, implying a common binomial parameter, others have derived statistics to test departures from the binomial assumption [3, 4], and have shown that the resulting C(α) statistic is optimal against correlated binomial data. The C(α) statistic is appealing because it is simple to compute and provides intuition by comparing the empirical variance with the null binomial variance. For our situation with n subjects, each a sample of size 2 alleles, the score statistic for testing the binomial distribution against the β-binomial alternative was derived by Tarone [4], but we extended it to allow subject-specific allele frequencies (p_i). The statistic has the form T_{Tarone} = (Σ_i f_i^2)/(Σ_i b_i^2), where f_i is the contribution of the i-th subject to the score equation evaluated under the null hypothesis:

\[ f_i = \left\{ \left( g_i - 2 p_i \right)^2 + g_i (2 p_i - 1 - 2 p_i) \right\} / 2 p_i (1 - p_i) \]

It is easy to verify that under the null hypothesis of binomial data, E[f_i] = 0, and that for large samples, T_{Tarone} \sim χ^2. Through simulations, however, we found T_{Tarone} to have inflated type I error rates for p < 0.3. Because this was likely caused by the factor 2p_i(1 - p_i)^2 in the denominator of f_i, we evaluated a slightly different score:

\[ S = \sum_{i=1}^{n} (g_i - 2 p_i)^2, \]

so that S is the empirical residual variance. Noting that S can be written as a quadratic form, S = r'r, where r is the vector of residuals, we can view r as having an asymptotic multivariate normal distribution with mean zero and diagonal variance matrix (V) with v_i = 2p_i(1 - p_i) along the diagonal. This makes it easy to consider the null mean and variance of S:

\[ E_r[S] = tr(VV) = \sum_i v_i, \]
\[ Var_r(S) = 2tr(VV) = 2\sum_i v_i^2. \]

With these in hand, we computed p values by two different approaches. The first was to create a standardized statistic, T_1 = (S - E_r[S])^2/Var_r(S), and assume that T_1 has an asymptotic χ^2_i distribution. The second was to use a Satterthwaite approximation for the distribution of quadratic forms [5–7]. Here, the distribution of S was approximated by a scaled χ^2 distribution with the scale and degrees of freedom estimated by the first two moments, E_r[S] and Var_r(S) [5–7]. That is, the scale was estimated as δ = Var_r(S)/(2E_r[S]), the degrees of freedom as d = 2E_r[S]^2/Var_r(S), and p values were computed by assuming T_{scaled} = S/δ \sim χ^2_i.

By simulations, we found expression (1) to work well for the null mean, but the null variance was smaller than the estimate from expression (2). The problem is exacerbated by small p. To obtain more accurate estimates of the variance of S, we used binomial moments to determine the correct variance, which accounts for the binomial kurtosis:

\[ Var_r(S) = \sum_i (\beta_i - 1)v_i^2. \]

Here, β_i is the coefficient of kurtosis, β_i = μ_4_i/σ_4_i, where μ_k, i is the 4th central moment for a binomial sample of size 2: μ_4_i = v_i. For normally distributed data, β_i = 3, which explains the factor of 2 in expression (2), Var_r(S) = 2trVV. In our case, β_i = 1/ν_i, which varies with the fitted allele frequency, p_i. Through simulations, we found expression (3) to accurately estimate Var_r(S), but it nonetheless led to conservative statistics (both T_1 and T_{scaled}). The problem is that as p_i decreases, E_r[S] and Var_r(S) become highly correlated. To overcome this problem, we tried traditional transformations for binomial data, such as

\[ \sqrt{g_i} \]
\[ \text{arcsin}(\sqrt{g_i}) \]

but these did not lead to adequate tests. Because expression (3) created a test much more conservative than that based on expression (2), we fixed β_i to its minimum value of 2 (which occurs when p_i = 1/2), which results in the simple estimate Var_r(S) = Σ_i v_i. This estimate lead to statistics with reasonable type I error rates and power for a wide range of allele frequencies, illustrated by our simulations.

We recognize that our approach might seem to have limitations, because the above methods do not correct for having estimated the regression coefficients for the covariates. As emphasized elsewhere [8, 9], it can be important to account for estimating the regression parameters used for the fitted values, because the fitted values will be closer to the observed values, and the residuals will have smaller variances, compared to use of the true regression parameters. For multivariate normal data, this can be easily accomplished by computing P_0, the projection matrix for the fitted model, P_0 = V - VX(X'VX)^{-1}X'Y, and using it to compute the null mean and variance of S: E_r[S] = tr(P_0), and Var_r(S) = 2tr(P_0P_0). We evaluate these moments in the statistical tests, but found the statistics to be too conservative (when the factor 2 was used to compute Var(S)) or too liberal (when the factor 2 was eliminated from Var(S)). In fact, ignoring the fact that the regression coefficients were estimated, and using Var_r(S) = Σ_i v_i^2, worked well in terms of controlling the type I error rate and had reasonable power.

In summary, we chose three statistics to test whether there is residual correlation between alleles, after adjusting for covariates. For the following statistics, we used E_r[S] = Σ_i v_i, and Var_r(S) = Σ_i v_i^2:

1. Fisher’s z(ρ)
2. C(α) as χ^2_i: T_1 = (S - E_r[S])^2/Var_r(S)
3. C(α) as scaled χ^2: T_{scaled} = S/δ, with δ = Var_r(S)/(2E_r[S]) and d = 2E_r[S]^2/Var_r(S).

Simulations

For the simulations, we varied the total sample size over 50, 100, 200, and 500, and simulated scenarios without and with departures from HWE. In one scenario, we simulated from a single population with departure from HWE, to compare the power of our methods with standard tests for HWE (χ^2 GOF and exact test). The level of departure from HWE was determined by D', the fraction of the maximum value of D from table 1. This way, we could contrast the power of our methods when there is residual correlation with standard tests for HWE.

For other scenarios, we divided the total sample size into equally sized subpopulations (2 or 5). For some of these scenarios, we kept the allele frequency constant over all subpopulations, to
evaluate the type I error rates, whether or not adjustment was made for the different subpopulations. In other scenarios, we varied the allele frequencies across the subpopulations within a simulation. This allowed us to evaluate two features of the statistics: (1) without adjustment for the subpopulations (i.e. only an intercept in the regression models), the simulations give an impression of the power of our methods to detect residual correlations induced by a sample of genotypes from a stratified population; (2) with adjustment for the subpopulations, modeled as indicator covariates, the resulting simulations illustrate the type I error rates of our methods after adjusting for subpopulation covariates. This latter was used to evaluate how well the subpopulation covariates could explain away the HWD. From this perspective, correct modeling of the allele indicators should result in independent residuals, so that statistical tests on the residuals provide estimates of the type I error. If a model adequately 'explains' HWD, then the simulated type I error rate should be near the nominal. For all our simulations, we used nominal error rates of 0.01 and 0.05, and 1,000 simulations.

When varying allele frequencies across subpopulations, we fixed one subpopulation to have an allele frequency of 0.1, and then used equally sized increments to increase the allele frequency of the other subpopulations, up to a maximum specified allele frequency. For example, with 5 subpopulations, and a maximum allele frequency of 0.2, the array of allele frequencies was (0.100, 0.125, 0.150, 0.175, 0.200). The larger the maximum allele frequency, the larger the increments between the subpopulations.

To evaluate our methods for continuously varying covariates, we simulated a normally distributed covariate, $x$, and then used it to compute a subject-specific allele frequency $p$ based on the logistic model, $p = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$. The parameter $\beta_0$ was chosen based on a specified 'population' average allele frequency ($\bar{p}$), and based on the logit transformation: $\beta_0 = \log(\frac{\bar{p}}{1 - \bar{p}}) - \beta_1 \bar{x}$. The parameter $\beta_1$ ranged from 0, for type I error, up to 2. From this arrangement, two independent allele indicators were simulated for each subject.

**Results**

**One Population**

Simulation results for a single population are illustrated in figure 1 (type I error) and figure 2 (power), based on nominal type I error rates of 0.05. For these results, we illustrate our three proposed statistics, $z(\rho)$, $T_1$, and $T_{scaled}$, as well as the commonly used tests for departure from
HWE, \(T_{GOF}\) and the exact test [2]. Figure 1 shows that the type I error of \(T_{scaled}\) can be grossly inflated, while \(z(\rho)\) and \(T_1\) behave much like the usual test for HWE, \(T_{GOF}\). Note that all statistics tend to be conservative for small minor allele frequency (MAF; 0.01 in our simulations), mainly because many simulated data sets were not informative for such small MAF. Figure 2 illustrates that the power of \(z(\rho)\) and \(T_1\) is similar to that for \(T_{GOF}\) and the exact test. Power for \(T_{scaled}\) is misleading because of its inflated type I error rates. Results for a nominal type I error rate of 0.01 were similar to those in figures 1 and 2 (not shown), with the important exception that larger sample sizes were required to avoid inflated type I error rates for small MAF. The general pattern suggested that for sample sizes of at least 100, the type I error rates were close to the nominal level of 0.01 for MAF at least 0.10, and that for sample sizes of at least 500, the type I error rates were adequate for MAF at least 0.05.

**Two Populations**

Simulations for two subpopulations with common allele frequencies showed that the type I error rates of \(z(\rho)\) and \(T_1\) were near the nominal rate of 0.05, while \(T_{scaled}\) tended to have inflated type I error rates for small MAF, for both unadjusted and covariate adjusted tests (results not shown). Simulations with different allele frequencies across the subpopulations give the power of our three statistics, \(z(\rho)\), \(T_1\), and \(T_{scaled}\) to detect allelic correlations when no covariates are modeled, as illustrated in the left panels of figure 3 (note that the y-axis on the left panels ranges from 0 to 1). As for one population, the power of \(z(\rho)\) and \(T_1\) was close to that for \(T_{GOF}\) and the exact test.

The right panels of figure 3 illustrate the type I error rate of the three statistics \(z(\rho)\), \(T_1\), and \(T_{scaled}\), as tests for residual correlation after modeling the subpopulations with indicator variables. Note that the y-axis for the right panels ranges from 0 to 0.10. These results illustrate that \(z(\rho)\) tends to have grossly inflated type I error rates, \(T_{scaled}\) tends to be conservative, and \(T_1\) fluctuates as being slightly conservative in some situations and slightly anticonservative in others. The larger the sample size, the closer the error rate of \(T_1\) approximates the nominal level of 0.05. The conclusions for simulations with a nominal type I error rate of 0.01 (results not shown) were similar to those summarized in figures 2 and 3.
Regression Modeling of Allele Frequencies and Testing HWE

Five Populations
Simulation results for five subpopulations were generally consistent with those for two subpopulations. The left panels of figure 4 illustrate the power of the various tests to detect allelic correlations when no covariates are modeled, and the right panels illustrate the type I error rates for residual correlation after adjusting for the five subpopulations via indicator variables. Again, we find that \( z(\rho) \) tends to have grossly inflated type I error rates, \( T_{\text{scaled}} \) tends to be conservative, and the error rate of \( T_1 \) fluctuates about the nominal level of 0.05, with less fluctuation for larger sample sizes. Similar conclusions were made for simulations with a nominal type I error rate of 0.01 (results not shown).

Quantitative Covariate
For the normally distributed \( x \) covariate, we present in figures 5 and 6 the power to detect departures from HWE when no covariates are modeled, as well as the type I error rates to detect residual correlations when the \( x \) covari-
ate is included in the model. Figure 5 gives results for when the average allele frequency is 0.05, and figure 6 for when the average is 0.2. These figures illustrate that our three evaluated statistics, $z(\rho)$, $T_1$, and $T_{scaled}$, have power that is similar to $T_{GOF}$ and the exact test. For testing residual correlations, we find the type I error rates to be grossly inflated for $z(\rho)$, yet close to the nominal for both $T_1$ and $T_{scaled}$. Similar conclusions were made for simulations with a nominal type I error rate of 0.01 (results not shown).

**Discussion**

Because testing for departures from HWE has become widely used to evaluate the quality of SNP genotypes, and because large-scale studies are often mixtures of subpopulations with different allele frequencies, it is worthwhile to evaluate whether any observed departures from HWE can be attributed to measured covariates. Without the types of modeling we have suggested, the alternative would be to stratify a sample into mutually exclusive sub-
groups, and either perform a global stratified test for HWE [10], or perform separate tests for the different subgroups. The global stratified test achieves the same objective as our proposed test for residual correlation, yet without the flexibility of the modeling approach that allows multiple covariates, some of which could be quantitative. Performing separate tests for HWE within different subgroups is plagued by multiple testing and weak power within small subgroups. In contrast, our modeling approach provides much flexibility on ways to model the influence of covariates on allele frequencies, even geographic spatial structure [11], as well as a global test for residual correlation of alleles.

By iterating between large sample approximations of the distribution of our proposed methods and simulations, we propose the simple $T_1$ statistic as a way to test
for residual correlations between alleles, after adjusting for covariates. This statistic is intuitively appealing, because it directly compares the empirical variance with the binomial expected variance, illustrated by the numerator of $T_1$ expressed as $\sum_i (g_i^2 - 2p_i^2) - 2p_i(1 - p_i)$. The denominator, the variance, was challenging to determine, yet in the end the simple variance $\sum_i [2p_i(1 - p_i)]^2$ worked best. Through simulations, we showed that using $T_1$ to test for residual correlations resulted in type I error rates that were close to the desired nominal rate, and that $T_1$ had power to detect HWE that was similar to the commonly used tests, $T_{GOF}$ and the exact test. In contrast, our other evaluated statistics, $z(\rho)$ and $T_{scaled}$ performed poorly in terms of type I error rates.

Throughout our developments and simulations, we assumed a locus with only two alleles, which covers the majority of genetic markers used in modern large-scale genetic studies (i.e. SNPs). Yet, it would be appealing to account for more than two alleles. This could be achieved by using the multinomial distribution to model the influ-

---

Fig. 6. Quantitative covariate and average MAF = 0.20. Left panels for power to detect allelic correlations when not adjusting for a normally distributed quantitative covariate. Right panels are type I error rates for residual correlation after adjusting for the quantitative covariate in quasi-binomial logistic regression. See figure 1 for legend.
nce of covariates on allele frequencies, using a quasi-likelihood to allow for over-dispersion at this modeling step. Then, the fitted values could be used to construct a test for residual allelic correlations. Paul et al. [3] extended Tarone’s test to construct a test for departures from a multinomial distribution against a Dirichlet-multinomial alternative. Like Tarone, they assumed the multinomial probabilities were constant over all subjects. Using the approach by Paul et al. [3], assuming $m$ alleles, but extending to allow multinomial probabilities to depend on subject-specific covariates, we extended the score statistic of Paul et al. [3] to take the form

$$S = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\text{hom}_{ij} - p_{ij}^2}{p_{ij}},$$

where $\text{hom}_{ij}$ is an indicator of whether subject $i$ is homozygous for allele $j$, and $p_{ij}$ is the fitted probability of allele $j$ for subject $i$. Using Paul et al.’s [3] approach, it is easy to show that $\text{Var}(S) = n(m-1)$, and $T = S^2/\text{Var}(S)$ has an asymptotic $\chi^2$ distribution with 1 d.f. This makes it straightforward to extend our modeling approach to test for residual correlation of allele indicators when there are more than two alleles.

We have assumed that all subjects are unrelated, not allowing for pedigree data. Others have developed methods to estimate allele frequencies from pedigree data. McPeek et al. [12] considered best linear unbiased estimation of allele frequencies from pedigree data, using allele indicators and known structures of pedigrees to account for correlations of allele indicators among pedigree members, yet assumed HWE. Olson [13] also used allele indicators and pedigree data, but based methods on generalized estimating equations to estimate allele frequencies, HWD disequilibrium coefficients, and linkage disequilibrium coefficients (i.e. correlation of allele indicators between loci). A critical distinction between these two approaches is that Olson [13] used a ‘working independence’ for pedigree members, relying on robust covariance estimates, in contrast to McPeek et al. [12] who used the pedigree structure to determine kinship coefficients to build an assumed covariance structure. Building on the work of McPeek et al. [12], Bourgain et al. [14] used quasi-likelihood methods to derive tests for HWE with pedigree data. Although none of these authors allowed for covariates to model allele frequencies, their general formulations could be extended to allow for covariates. The quasi-likelihood approach is particularly appealing, because one could allow for subject-specific covariates influencing subject-specific allele frequencies, an approach that might be especially useful when evaluating residual correlations between alleles of subjects in pedigrees composed of different racial or ethnic groups. A challenging aspect of this situation is that the covariates for a subject will depend on a mixture of the covariates of its ancestors. The pedigree structure influences both the mean and the covariance matrix of the allele indicators.

Much of our discussion has focused on subject-specific covariates, such as race or ethnic group, or quantitative covariates, such as principal components derived from large-scale genotype data. One could also allow for ancestral ambiguity by using the probability of ancestral group membership, such as probabilities output from the STRUCTURE program [15], as subject-specific covariates. Furthermore, as genomic technology and research progresses, it might be informative to extend our methods to model allele-specific covariates, such as features of genomic sequence data including the number of sequence reads or the quality of the reads. To extend our approach to allele-specific covariates, one could model the effects of covariates on allele indicators ($a_{ij}$ for allele $j$ of subject $i$) by treating each subject as a cluster of size 2 alleles, and using estimating equations to model the effects of covariates on allele indicators. Then, our score statistic could be extended by changing the numerator of $T_1$ to be $\left[\Sigma_i(a_{ij} - p_{ij}) + (a_{i2} - p_{i2})\right]^2 - \left[p_{i1}(1 - p_{i1}) + p_{i2}(1 - p_{i2})\right]^2$, which still contrasts the empirical variance with the assumed binomial variance, but now allows for allele frequencies that depend on allele-specific covariates. The denominator of $T_1$ would change to $\Sigma_i[p_{i1}(1 - p_{i1}) + p_{i2}(1 - p_{i2})]^2$. This might provide new insights to measured genomic features that influence allele frequencies and add to the methods useful to evaluate the quality of called genotypes.

**Acknowledgements**

This research was supported by the US Public Health Service, National Institutes of Health, contract grant number GM065450.
References