A new association test based on disease allele selection for case–control genome-wide association studies

Zhongxue Chen

Abstract

Background: Current robust association tests for case–control genome-wide association study (GWAS) data are mainly based on the assumption of some specific genetic models. Due to the richness of the genetic models, this assumption may not be appropriate. Therefore, robust but powerful association approaches are desirable.

Results: In this paper, we propose a new approach to testing for the association between the genotype and phenotype for case–control GWAS. This method assumes a generalized genetic model and is based on the selected disease allele to obtain a p-value from the more powerful one-sided test. Through a comprehensive simulation study we assess the performance of the new test by comparing it with existing methods. Some real data applications are also used to illustrate the use of the proposed test.

Conclusions: Based on the simulation results and real data application, the proposed test is powerful and robust.

Keywords: Generalized genetic model, Robust test, Single-nucleotide polymorphism

Background

In a case–control genome-wide association study (GWAS), to detect the associated single-nucleotide polymorphisms (SNPs), we need to conduct a test for each individual SNP data, which are summarized as a 2-by-3 table. Although Pearson's chi-square test can be used, it is usually less powerful than the Cochran-Armitage trend test (CATT) if the genetic model is known [1-3]. However, if the genetic models are unknown or various, the optimal scores in the CATTs are difficult or unable to find. If we use a CATT with fixed scores for all SNPs, we may lose power for some situations [4-13]. To circumvent this disadvantage, in the literature, many robust association tests have been proposed [7,12,14-22]. Those tests do not rely solely on one specific genetic model; rather they consider several possible genetic models simultaneously. In addition, many of them are based on the assumption that the underlying genetic model is one of the following three: additive, recessive, and dominant. For example, the maxmin efficiency robust test (MERT) by Gastwirth [23,24], and the maximum of the three optimal CATTs under recessive, additive, and dominant models (MAX3) have been studied [6]. Zheng and Ng proposed a two-phase procedure called genetic model selection (GMS) method [12] which selects a genetic model from the three models in its first stage. On the contrary, Joo et al. proposed a test which eliminates genetic models [25]. Due to the environmental interaction, there are unlimited genetic models besides the three ideal models (recessive, additive, and dominant). Chen and Ng proposed a robust association test based on the generalized genetic model (GGM) [19], which includes the recessive, additive, and dominant models as special cases. Their approach obtains a p-value from a one-sided test for each of the two possible disease alleles. With the uncertainty of the disease allele, the overall p-value is then approximated from the two dependent tests.

In this paper, we propose a new robust association test which utilizes the GGM and obtains a one-sided p-value based on the selected disease allele. The performance of the new test is compared with existing methods in terms of controlling type I error rate and detecting power. Some real data applications are also used to demonstrate the use of the new test.
Methods

GGM and existing methods

The data of a diallelic SNP with alleles A and a in a case control GWAS are summarized in Table 1, where \( r_1, r_2, \) and \( r_3 \) are the frequencies of genotypes AA, Aa, and aa, respectively, for the \( r \) cases; and \( s_1, s_2, \) and \( s_3 \) are the frequencies of genotypes AA, Aa, and aa, respectively, for the \( s \) controls.

Among the three genotypes AA, Aa, and aa, the relative risks of genotypes Aa and aa to AA are defined as:

\[
\begin{align*}
\lambda_1 &= \frac{\Pr(\text{case}|AA)}{\Pr(\text{case}|AA)} \\
\lambda_2 &= \frac{\Pr(\text{case}|aa)}{\Pr(\text{case}|AA)}
\end{align*}
\]

Under the null hypothesis that there is no association between the genotype and phenotype, \( \lambda_1 = \lambda_2 = 1 \). Regarding the alternative hypothesis, we assume the underlying genetic model is a GGM, which is also called order-restricted relative risks model [19]. For the case where \( a \) is the disease allele, GGM assumes \( \lambda_1 \geq 1 \) and \( \lambda_2 \geq \lambda_1 \) with at least one of the inequalities is strictly greater than. It is easy to see that the aforementioned ideal models, recessive (\( \lambda_1 = 1, \lambda_2 > \lambda_1 \)), additive (\( \lambda_1 = (1 + \lambda_2)/2 \)), and dominant (\( \lambda_1 = \lambda_2 > 1 \)), are all special cases of the generalized model. If \( A \), rather \( a \), is the disease allele, GGM assumes \( 1 \geq \lambda_1 \) and \( 1 \geq \lambda_2 \) with at least one of the inequalities is strict.

Suppose the frequencies for AA, Aa, and aa are \( p_1, p_2, p_3 \) for cases and \( q_1, q_2, \) and \( q_3 \) for controls, respectively. Under the null hypothesis that there is no association between the disease and the genotype, it is easy to show \( p_1 = q_1, p_2 = q_2, \) and \( p_3 = q_3 \). In this paper, we assume the cases and controls follow trinomial distributions TN \((r, p_1, p_2, p_3)\) and TN \((s, q_1, q_2, q_3)\), respectively.

The test statistics for some well-known existing methods are summarized as follows:

The CATT test statistic is [8]:

\[
Z_c = \frac{n^{1/2} \sum_{i=0}^2 x_i(s_i-r_i)}{\left(\frac{\sum_{i=0}^2 x_i^2 n_i - \left(\sum_{i=0}^2 x_i n_i\right)^2}{n} \right)^{1/2}}, \text{ where } \left(x_0, x_1, x_2\right) = (0, 1, 1).
\]

They are the scores assigned to the three columns.

| Genotype | AA | Aa | aa | Total |
|----------|----|----|----|-------|
| Case     | \( r_1 \) | \( r_2 \) | \( r_3 \) | \( r \) |
| Control  | \( s_1 \) | \( s_2 \) | \( s_3 \) | \( s \) |
| Total    | \( n_1 \) | \( n_2 \) | \( n_3 \) | \( n \) |

The statistic for MAX3 is [6]:

\[
MAX3 = \text{MAX}\{\left|\frac{Z_0}{Z_1/2}\right|, |Z_1|\}.
\]

The statistic for GMS is [12,13]:

\[
GMS = Z_0 I(\frac{Z_1}{2} > 0) I(\frac{Z_{HWDTT}}{c} > c)
+ Z_1 I(\frac{Z_1}{2} > 0) I(\frac{Z_{HWDTT}}{c} < c)
- Z_0 I(\frac{Z_1}{2} \leq 0) I(\frac{Z_{HWDTT}}{c} > c)
- Z_0 I(\frac{Z_1}{2} \leq 0) I(\frac{Z_{HWDTT}}{c} < c)
\]

where \( I \) is the indicator function, and the Hardy-Weinberg disequilibrium trend test (HWDTT) statistic is given by [15]:

\[
Z_{HWDTT} = \frac{(r/n)^{1/2}((\Delta s - \Delta a)/(\sqrt{r_s n - n_s})/\sqrt{r_a n - n_a}]}{1/(r_s n - n_s)}.
\]

The statistic for MERT is [24]:

\[
\text{MERT} = \frac{Z_0 + Z_1}{\left\{2(1 + \hat{\rho}_{01})\right\}^{1/2}}, \text{ where } \hat{\rho}_{01} = \left(\frac{n_0 n_2}{n_0 + n_1}\right)^{1/2}/\left(\frac{n_0 + n_1}{n_1 + n_2}\right)^{1/2}.
\]

The proposed test

Let \( T = \begin{bmatrix} T_1 \\ T_2 \\ T_3 \end{bmatrix} = \begin{bmatrix} r_2 s_1 - r_1 s_2 \\ r_3 s_2 - r_1 s_3 \\ r_3 s_1 - r_1 s_3 \end{bmatrix}, \) it can be shown that under the null hypothesis of no association, the mean of the vector \( E[T] = 0 \), and variance-covariance matrix is.

\[
\Sigma_T = \begin{bmatrix}
\text{MERT} = \frac{Z_0 + Z_1}{\left\{2(1 + \hat{\rho}_{01})\right\}^{1/2}}, \text{ where } \hat{\rho}_{01} = \left(\frac{n_0 n_2}{n_0 + n_1}\right)^{1/2}/\left(\frac{n_0 + n_1}{n_1 + n_2}\right)^{1/2}.
\]

We define the following “standardized” statistics:

\[
Z = \begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \end{bmatrix} = \begin{bmatrix}
(r_2 s_1 - r_1 s_2) / \sqrt{r_s n_2 (n + (2-n) n_s)/n^2}
(r_3 s_2 - r_1 s_3) / \sqrt{r_s n_3 (n + (2-n) n_s)/n^2}
(r_3 s_1 - r_1 s_3) / \sqrt{r_s n_3 (n + (2-n) n_s)/n^2}
\end{bmatrix}
\]

It is easy to prove the following result.
Theorem 1. Asymptotically, under the null hypothesis the statistics in (2) follow a multivariate normal distribution, $Z \sim MVN(0, \Sigma_Z)$, where the variance-covariance matrix is

$$
\Sigma_Z = \begin{bmatrix}
1 & \sqrt{\frac{p_3 p_2}{(1-p_2)(1-p_3)}} \\ 0 & 1 \\
\sqrt{\frac{p_3 p_2}{(1-p_2)(1-p_3)}} & \sqrt{\frac{p_1 p_0}{(1-p_2)(1-p_3)}} \\
\end{bmatrix}.
$$

From theorem 1, $Z_1, Z_2$ and $Z_3$ are linear dependent; it is not difficult to show that asymptotically $Z_3 = aZ_1 + bZ_2$, where $a = \sqrt{\frac{p_3 p_1}{(1-p_2)(1-p_3)}}$ and $b = \sqrt{\frac{p_1 p_0}{(1-p_2)(1-p_3)}}$. It can also be shown that under the GGM, if $a$ (or $A$) is the disease allele, the expectations of $Z_i$ ($i = 1,2,3$) in (2) are all greater (or less) than 0. In addition, under the GGM, if $z_1$ is close to 0, the genetic model is close to the recessive model. On the other hand, the genetic model is unlikely to be the recessive model if $z_1$ is far from 0.

We will select the disease allele based on the sign of $Z_3$ and combine the two one-sided $p$-values obtained based on the selected disease allele. Following the idea of combining $p$-values from independent studies using robust test [26], we define the test statistic as follows:

$$
C = \begin{cases} 
F^{-1}(\Phi(Z_1)) + F^{-1}(\Phi(Z_2)) & \text{if } Z_2 \geq 0 \\
F^{-1}(\Phi(-Z_1)) + F^{-1}(\Phi(-Z_2)) & \text{if } Z_2 < 0
\end{cases},
$$

where $\Phi(\cdot)$ is the cumulative density function (CDF) of the standard normal distribution (N(0,1)) and $F^{-1}(\cdot)$ is the inverse of the CDF of the chi-square distribution with one degree of freedom.

Usually it is difficult to directly calculate the p-value (or critical value) for statistic $C$ in (3). However, the p-value can be easily estimated using resampling method. Specifically, we first simulate two independent samples, $z_1$ and $z_2$, both from N(0,1). Then we calculate $z_3 = a z_1 + b z_2$, where $a = \sqrt{\frac{p_3 p_1}{(1-p_2)(1-p_3)}}$ and $b = \sqrt{\frac{p_1 p_0}{(1-p_2)(1-p_3)}}$ are the estimates for $a$ and $b$, and $p_1 = \frac{n_1}{n}$ are the estimates for $p_i$ ($i = 1,2,3$) from the data. Next we calculate

$$
g = \begin{cases} 
F^{-1}(\Phi(z_1)) + F^{-1}(\Phi(z_2)) & \text{if } z_3 \geq 0 \\
F^{-1}(\Phi(-z_1)) + F^{-1}(\Phi(-z_2)) & \text{if } z_3 < 0
\end{cases},
$$

We repeat the above steps $K$ times and get $K$ values for $g$. The $p$-value can then be estimated as (number of $g's$ that are greater than or equal to $c)/K$, where $c$ is the observed statistic calculated from data using (3).

### Table 1

| $\lambda_1 \lambda_2$ | (1,1) | (1,1.4) RM* | (1,1.1, 1.4) AM* | (1.2, 1.4) AM* | (1.3, 1.4) | (1.4, 1.4) DM* | (1.5, 1.4) ODM* | (0.9, 1.4) UDM* |
|----------------------|-------|-------------|-----------------|----------------|-------------|----------------|----------------|----------------|
| ChiSQ               | 0.053 | 0.927       | 0.759           | 0.528          | 0.395       | 0.428          | 0.524          | 0.992          |
| MAX3                | 0.051 | 0.941       | 0.794           | 0.568          | 0.438       | 0.445          | 0.484          | 0.994          |
| GMS                 | 0.05  | 0.933       | 0.772           | 0.577          | 0.431       | 0.432          | 0.459          | 0.993          |
| CATT                | 0.05  | 0.934       | 0.807           | 0.644          | 0.427       | 0.25           | 0.136          | 0.986          |
| MERT                | 0.051 | 0.879       | 0.744           | 0.642          | 0.479       | 0.377          | 0.261          | 0.938          |
| GGM                 | 0.053 | 0.944       | 0.793           | 0.598          | 0.447       | 0.432          | 0.412          | 0.995          |
| New                 | 0.049 | 0.927       | 0.761           | 0.574          | 0.455       | 0.46           | 0.512          | 0.982          |

*RM: Recessive Model; AM: Additive Model; DM: Dominant Model; ODM: Over-dominant Model; UDM: Under-dominant Model.
### Results and discussion

**Simulation study**

In this section, we will assess the performance of the proposed test by comparing it with existing methods in terms of controlling type I error rate and detecting power through a comprehensive simulation study. In the simulation study we assume that cases and controls follow trinomial distributions with probabilities \( p = (p_1, p_2, p_3) \) and \( q = (q_1, q_2, q_3) \), respectively. It can be shown that, for given \( q_i/s \), and relative risks \( \lambda_1, \lambda_2 \), the values of the corresponding \( p_i/s \) can be obtained as follows \([17-21]\):

\[
p_1 = \frac{q_1}{q_1 + q_2 + q_3}, \quad p_2 = \frac{\lambda_1 q_2}{q_1 + \lambda_1 q_2 + \lambda_2 q_3}, \quad p_3 = \frac{\lambda_2 q_3}{q_1 + \lambda_1 q_2 + \lambda_2 q_3}.
\]

We first assume Hardy–Weinberg equilibrium (HWE) holds for controls, and the minor allele frequencies (MAF) are 0.3 and 0.5. The disease allele is either the minor or the major allele. The numbers of cases \((r)\) and controls \((s)\) are both set to be 1000. Different pairs of \( \lambda_1 \) and \( \lambda_2 \) are used in the simulation to compare the performance of our proposed method with those of GMS, MERT, MAX3, Pearson’s chi-square test, and CATT with \( x = 0.5 \), which is the commonly used test under the assumption of additive genetic model. More specifically, we fix \( \lambda_2 \) to be 1.4 and let \( \lambda_1 \) vary from 1.0 to 1.4 with increment 0.1. Therefore, the three special genetic models are included in the simulation study. To assess the robustness of the proposed test, we also simulate data from genetic models other than the GGM: over-dominant model \((\lambda_1 = 1.5, \lambda_2 = 1.4)\) and under-dominant model \((\lambda_1 = 0.9, \lambda_2 = 1.4)\). The significance level is set to be 0.05, and the type I error rate and power are estimated by the proportions of rejections from 1000 replicates. To estimate the p-value for the proposed test, we resample 10,000 times \((i.e., K = 10,000)\). The p-values from MAX3, GMS, and MERT are obtained by using the R package “Rassoc” \([13]\).

When the null hypothesis is true \((\lambda_1 = \lambda_2 = 1)\), all methods have rejection proportions close to the preset significance level 0.05 \((\text{For power values, see Additional file 1: Power plots of Figures S1-S5 for Tables 2-6}); this indicates that they can control type I error rate. Table 2 reports the empirical powers of each method when HWE holds for controls with MAF equals to 0.3, and the minor allele is the risk allele. The chi-square test is usually less powerful than the proposed test except for the condition where the genetic model is over-dominant, under which the proposed test is more powerful than other methods. The CATT and MERT perform relatively poorly when the genetic models are far from the additive model. MAX3 and GMS perform similarly and have comparable power values with the proposed test; while the new test performs better when the genetic model is dominant or over-

### Table 4 Empirical type I error rates and powers for each method from 1000 replicates at significance level 0.05 when the sample sizes are 1000 for cases and controls and HWE holds for controls MAF equals 0.5

| \( \lambda_1 \), \( \lambda_2 \) | \( (1.1) \), \((1.4) \) RM* | \((1.1), (1.4) \) AM* | \((1.2), (1.4) \) AM* | \((1.3), (1.4) \) DM* | \((1.4), (1.4) \) DM* | \((1.5), (1.4) \) ODM* | \((0.9), (1.4) \) UDM* |
|---|---|---|---|---|---|---|---|
| ChiSQ | 0.046 | 0.03 | 0.743 | 0.539 | 0.426 | 0.481 | 0.565 | 0.995 |
| MAX3 | 0.04 | 0.044 | 0.771 | 0.601 | 0.473 | 0.486 | 0.507 | 0.995 |
| GMS | 0.04 | 0.029 | 0.756 | 0.609 | 0.471 | 0.49 | 0.517 | 0.992 |
| CATT | 0.043 | 0.036 | 0.807 | 0.659 | 0.459 | 0.314 | 0.194 | 0.974 |
| MERT | 0.044 | 0.088 | 0.763 | 0.656 | 0.517 | 0.409 | 0.302 | 0.933 |
| GGM | 0.043 | 0.092 | 0.786 | 0.635 | 0.484 | 0.46 | 0.437 | 0.995 |
| New | 0.049 | 0.092 | 0.755 | 0.616 | 0.485 | 0.514 | 0.556 | 0.988 |

*RM: Recessive Model; AM: Additive Model; DM: Dominant Model; ODM: Over-dominant Model; UDM: Under-dominant Model.

### Table 5 Empirical type I error rates and powers for each method from 1000 replicates at significance level 0.05 when the sample sizes are 1000 for cases and controls and the genotype frequencies for controls are \((0.1, 0.36, 0.54)\) with minor allele is the disease allele

| \( \lambda_1 \), \( \lambda_2 \) | \( (1.1) \), \((1.4) \) RM* | \((1.1), (1.4) \) AM* | \((1.2), (1.4) \) AM* | \((1.3), (1.4) \) DM* | \((1.4), (1.4) \) DM* | \((1.5), (1.4) \) ODM* | \((0.9), (1.4) \) UDM* |
|---|---|---|---|---|---|---|---|
| ChiSQ | 0.046 | 0.03 | 0.743 | 0.539 | 0.426 | 0.481 | 0.565 | 0.995 |
| MAX3 | 0.04 | 0.044 | 0.771 | 0.601 | 0.473 | 0.486 | 0.507 | 0.995 |
| GMS | 0.04 | 0.029 | 0.756 | 0.609 | 0.471 | 0.49 | 0.517 | 0.992 |
| CATT | 0.043 | 0.036 | 0.807 | 0.659 | 0.459 | 0.314 | 0.194 | 0.974 |
| MERT | 0.044 | 0.088 | 0.763 | 0.656 | 0.517 | 0.409 | 0.302 | 0.933 |
| GGM | 0.043 | 0.092 | 0.786 | 0.635 | 0.484 | 0.46 | 0.437 | 0.995 |
| New | 0.049 | 0.092 | 0.755 | 0.616 | 0.485 | 0.514 | 0.556 | 0.988 |

*RM: Recessive Model; AM: Additive Model; DM: Dominant Model; ODM: Over-dominant Model; UDM: Under-dominant Model.
dominant. Table 3 lists the power values when HWE holds for controls with MAF equals to 0.3, and the major allele is the risk allele. Once again, the CATT and MERT lose power dramatically compared to other methods when the genetic model is recessive or under-dominant. The proposed test is more powerful than the MAX3 and GMS when the genetic models are close to recessive model. The chi-square test is less powerful than the proposed test when the genetic models are between recessive and dominant.

Table 4 gives the empirical powers when HWE holds for controls with MAF equals to 0.5. As seen in Tables 2 and 3, the CATT and MERT are less powerful when the genetic model is dominant or over-dominant. On the contrary, chi-square test is reasonably powerful under those situations. The proposed test has the largest powers when the genetic model is close to the dominant one (i.e., $\lambda_1 = 1.3, 1.4, 1.5$). For other situations, the new test has comparable powers with those from MAX3 and GMS.

Table 5 and 6 report the empirical powers when HWE does not hold with the genotypic frequencies for controls are $(0.1, 0.36, 0.54)$, and the disease risks are minor allele and major allele, respectively. Once again we can see that CATT and MERT may lose power dramatically under some conditions (e.g., over-dominant in Table 5 and under-dominant in Table 6). The proposed test has comparable power values with those from the MAX3 and GMS; under some situations, it is more powerful (e.g., over-dominant model in Table 5).

The method based on GGM by Chen and Ng (GGM in Tables 2–6) has similar performance as the proposed test. However, if the disease allele is the minor allele, the new test is more powerful than the GGM method when the genetic model is dominant or over-dominant (see Tables 2, 4 and 5). In addition, unlike the proposed test, the GGM method doesn’t report the disease allele.

Real data application
To illustrate the use of the proposed test, we apply it to some real data. The SNP rs181489 has been shown to be associated with Parkinson disease [27]. Table 7 lists the genotypic counts of cases and controls of this SNP from fifteen sites [27]. We apply the proposed test and others to the data of each site to obtain $p$-values. Table 8 reports the results along with the three statistics, $z_1$, $z_2$ and $z_3$. Out of the 15 populations, 14 have positive values for $z_3$, indicating the disease allele is A rather than G for this SNP. Recall the statistic $z_1$ compares genotype GG to GA between controls and cases. A small $z_1$ indicates the relative risk $\lambda_1$ is close to 1, and therefore the genetic model is close to the recessive one. On the other hand, if $z_1$ is large, the underlying genetic model is unlikely to be a recessive model. For population D, the estimated $z_1$ is $-0.090$, the genetic model is close to a recessive one, under which the

Table 6 Empirical type I error rates and powers for each method from 1000 replicates at significance level 0.05 when the sample sizes are 1000 for cases and controls and the genotype frequencies for controls are $(0.1,0.36,0.54)$ with major allele is the disease allele

| $\lambda_1 \lambda_2$ | (1,1) | (1.1,1.4) RM* | (1.1,1.4) AM* | (1.3,1.4) | (1.4,1.4) DM* | (1.5,1.4) ODM* | (0.9,1.4) UDM* |
|-----------------------|-------|---------------|---------------|------------|---------------|---------------|---------------|
| ChiSQ                | 0.042 | 0.55          | 0.553         | 0.673      | 0.812         | 0.927         | 0.981         | 0.749         |
| MAX3                 | 0.044 | 0.568         | 0.6            | 0.718      | 0.843         | 0.939         | 0.987         | 0.712         |
| GMS                  | 0.041 | 0.585         | 0.61           | 0.718      | 0.82          | 0.922         | 0.98          | 0.727         |
| CATT                 | 0.041 | 0.431         | 0.607          | 0.765      | 0.854         | 0.918         | 0.962         | 0.248         |
| MERT                 | 0.042 | 0.512         | 0.635          | 0.755      | 0.808         | 0.866         | 0.896         | 0.385         |
| GGM                  | 0.036 | 0.55          | 0.611          | 0.743      | 0.848         | 0.942         | 0.986         | 0.66          |
| New                  | 0.043 | 0.552         | 0.597          | 0.734      | 0.849         | 0.941         | 0.981         | 0.654         |

*RM: Recessive Model; AM: Additive Model; DM: Dominant Model; ODM: Over-dominant Model; UDM: Under-dominant Model.

Table 7 Genotypic count data for rs181489 from different populations (data obtained from [27])

| Population | Case | Control | Total |
|------------|------|---------|-------|
|            | GG   | GA      | AA    | GG   | GA      | AA    |
| A: Australia | 400  | 402      | 99    | 320  | 307      | 58    | 1586  |
| B: France   | 244  | 245      | 57    | 5    | 61       | 11    | 623   |
| C: Germany  | 86   | 119      | 19    | 16   | 18       | 7     | 265   |
| D: Germany  | 222  | 176      | 85    | 133  | 107      | 25    | 748   |
| E: Germany  | 144  | 149      | 39    | 169  | 140      | 29    | 670   |
| F: Greece   | 44   | 67       | 17    | 44   | 37       | 10    | 219   |
| G: Greece   | 119  | 126      | 47    | 130  | 123      | 18    | 563   |
| H: Ireland  | 140  | 147      | 58    | 229  | 157      | 38    | 769   |
| I: Italy    | 78   | 86       | 21    | 87   | 71       | 9     | 352   |
| J: Italy    | 73   | 88       | 28    | 44   | 43       | 8     | 284   |
| K: Italy    | 33   | 47       | 10    | 41   | 21       | 9     | 161   |
| L: Norway   | 290  | 233      | 80    | 240  | 228      | 56    | 1127  |
| M: Poland   | 158  | 144      | 47    | 171  | 135      | 30    | 685   |
| N: Sweden   | 50   | 30       | 10    | 91   | 68       | 17    | 266   |
| O: USA      | 156  | 170      | 50    | 191  | 137      | 32    | 736   |

Table 8 Genotypic count data for rs181489 from different populations (data obtained from [27])
The test statistic in (3) is defined based on the idea of combining p-values from independent studies using chi-square distribution with 1 degree of freedom [26]. Although there are many other approaches available in the literature [26,28-31], it remains a research topic to choose the best one if there is any. However, it should be noticed that for case control GWAS, a robust method, such as the proposed test, is desirable due to the various underlying genetic models. In addition, when we combine the p-values from the 15 independent studies using the chi-square distribution with 1 degree of freedom [26], the overall p-value is $2.6 \times 10^{-11}$.

Through simulation studies and real data applications, we have shown that the proposed test is robust and powerful. In addition, the three statistics, $z_1$, $z_2$, and $z_3$, may also provide useful information about the disease allele and the genetic model.

### Conclusions

Although CATT has been widely used in case–control GWAS with the assumption that the underlying genetic model is additive, its performance may be very poor if the true genetic model is not additive. Therefore, robust but powerful association tests are more appropriate when detecting the associated SNPs. Many existing association tests make the assumption that the genetic model is one of the three special genetic models (recessive, additive, and dominant), which may be a too strong assumption in practice. In this paper, we propose a robust association test without making strong assumption about the genetic model. Our simulation results show that even the assumption of GGM is violated (e.g., over- and under-dominant models), the proposed test still has reasonable power; indicating it is a robust test. In terms of computational cost, the proposed test is reasonably fast. For instance, it took my desktop about 70 seconds to get the results in Table 8 for the real data application. Our simulation study also confirmed that the proposed test can control type I error rate with smaller cutoff p-value, e.g., $10^{-4}$ and $10^{-5}$ (see Additional file 2: Table S1-S2).

### Additional files

**Additional file 1:** Power plots of Figures S1-S5 for Tables 2-6.

**Additional file 2:** Table S1. Empirical type I error rate ($\times 10^{-4}$) for each method from $10^5$ replicates at significance level $10^{-4}$ with the sample sizes 1000 for cases and controls and given genotype frequencies for controls. **Table S2.** Empirical type I error rate ($\times 10^{-5}$) for each method from $10^5$ replicates at significance level $10^{-5}$ with the sample sizes 1000 for cases and controls and given genotype frequencies for controls.

### Abbreviations

GWAS: Genome-wide association study; SNP: Single-nucleotide polymorphism; CATT: Cochran-Armitage trend test; MERT: Maxmin efficiency robust test; MAX3: The maximum of the three optimal; CATTs: Under recessive, additive, and dominant models; GMS: Genetic model selection; GGM: Generalized genetic model; CDF: Cumulative density function; HWDTT: Hardy-Weinberg disequilibrium trend test; HWE: Hardy-Weinberg equilibrium.
Competing interests
The author declares that he has no competing interests.

Acknowledgement
The author would like to thank the support from several faculty research funds awarded to the author by the Indiana University School of Public Health-Bloomington. The author is also grateful to the three anonymous referees for their constructive comments which highly improved the presentation of the manuscript.

Received: 5 February 2014 Accepted: 6 May 2014
Published: 12 May 2014

Reference
1. Cochran W: Some methods for strengthening the common chi-square tests. Biometrics 1954, 10(4):417–451.
2. Armitage P: Tests for linear trends in proportions and frequencies. Biometrics 1955, 11(3):375–386.
3. Zheng G, Freidlin B, Gastwirth JL: Comparison of robust tests for genetic association using case-control studies. IMS Lect Notes-Monogr Ser 2006, 49:253–265. (Optimality: The Second Erich L. Lehmann Symposium).
4. Chen Z, Zheng G: Exact robust tests for detecting candidate-gene association in case-control trio design. J Data Sci 2005, 3:19–33.
5. Freidlin B, Podgor MJ, Gastwirth JL: Efficiency robust tests for survival or ordered categorical data. Biometrics 1999, 55(3):883–886.
6. Freidlin B, Zheng G, Li Z, Gastwirth JL: Trend tests for case-control studies of genetic markers: power, sample size and robustness. Hum Hered 2002, 53(3):146–152.
7. Gonzalez JR, Carasco JL, Dudbridge F, Armgad L, Estivill X, Moreno V: Maximizing association statistics over genetic models. Genet Epidemiol 2008, 32(2):246–254.
8. Sasienski PD: From genotypes to genes: doubling the sample size. Biometrics 1997, 53(4):1253–1261.
9. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pihlstrom AV, Penttila M, Posner B, Balding DJ, Meyre D, Polychronakos C, Froguel P: A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 2007, 445(7130):881–885.
10. Slager SL, Schaid DJ: Case-control studies of genetic markers: power and sample size approximations for Armitage's test for trend. Hum Hered 2001, 52(3):149–153.
11. Zheng G, Freidlin B, Li Z, JL G: Choice of scores in trend tests for case-control studies of candidate gene associations. Biom J 2003, 45:335–348.
12. Zheng G, Ng HKT: Genetic model selection in two-phase analysis for case-control association studies. Biostatistics 2008, 9(3):391–399.
13. Zang Y, Fung WK, Zheng G: Simple algorithms to calculate the asymptotic null distributions of robust tests in case-control genetic association studies in R. J Stat Softw 2010, 33(8):1–24.
14. Kwak M, Joo J, Zheng G: A robust test for two-stage design in genome-wide association studies. Biometrics 2009, 65(4):1288–1295.
15. Song K, Elston RC: A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine-mapping in case-control studies. Stat Med 2006, 25(11):105–126.
16. Wang K, Sheffield VC: A constrained-likelihood approach to marker-trait association studies. Am J Hum Genet 2005, 77(5):768–780.
17. Chen Z, Huang H, Ng HKT: Testing for association in case-control genome-wide association studies with shared controls. Stat Methods Med Res 2013, Published online before print February 1, 2013, doi:10.1177/0962280213470461.
18. Chen Z: Association tests through combining p-values for case-control genome-wide association studies. Stat Probability Lett 2013, 83(8):1854–1862.
19. Chen Z, Ng HKT: A robust method for testing association in genome-wide association studies. Hum Hered 2012, 73(1):26–34.
20. Chen Z, Huang H, Ng HKT: Design and analysis of multiple diseases genome-wide association studies without controls. GENE 2012, 510(1):87–92.
21. Chen Z: A new association test based on Chi-square partition for case-control GWA studies. Genet Epidemiol 2011, 35(7):658–663.
22. Chen Z, Huang H, Ng HKT: An improved robust association test for GWAS with multiple diseases. Stat Probability Lett 2014, 91:153–161.
23. Gastwirth JL: On robust procedures. J Am Stat Assoc 1966, 61:929–948.
24. Gastwirth JL: The use of maximin efficiency robust tests in combining contingency tables and survival analysis. J Am Stat Assoc 1985, 80:380–384.
25. Joo J, Kwak M, Zheng G: Improving power for testing genetic association in case-control studies by reducing the alternative space. Biometrics 2010, 66(1):266–276.
26. Chen Z, Nadarah SJ: On the optimally weighted z-test for combining probabilities from independent studies. Comput Stat Data Anal 2014, 70:287–294.
27. Elbaz A, Ross OA, Ioannidis J, Soto-Otrolaza AJ, Moisan F, Aaby J, Anna N, Bozi M, Brighina L, Chartier-Harlin MC, Destee A, Ferrarese C, Ferraris A, Gibson JM, Gispert S, Hadjigeorgiou GM, Jasinski-Mya B, Klein C, Krüger R, Lambert JC, Lohmann K, van de Loo S, Loriot MA, Lynch T, Mellick GD, Mutez E, Nilsson C, Opala G, Puschmann A, Quattrone A, et al: Independent and joint effects of the MAPT and SNCA genes in Parkinson disease. Ann Neurol 2011, 69(5):778–792.
28. Chen Z, Nadarah S: Comments on 'Choosing an optimal method to combine p-values' by Sungho Won, Nathan Morris, Qing Lu and Robert C. Elston, Statistics in Medicine 2009; 28: 1537–1553. Stat Med 2011, 30(24):2959–2961.
29. Chen Z: Is the weighted z-test the best method for combining probabilities from independent tests? J Evol Biol 2011, 24(4):926–930.
30. Loughin TM: A systematic comparison of methods for combining p-values from independent tests. Comput Stat Data Anal 2004, 47(3):467–485.
31. Fisher RA: Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd, 1932.

doi:10.1186/1471-2164-15-358
Cite this article as: Chen: A new association test based on disease allele selection for case-control-genome-wide association studies. BMC Genomics 2014, 15:358.