BMC Genetics

BioMed Central

Proceedings

Open Access

# Robust trend tests for genetic association in case-control studies using family data

Xin Tian*, Jungnam Joo, Gang Zheng and Jing-Ping Lin

Address: Office of Biostatistics Research, National Heart, Lung and Blood Institute, 6701 Rockledge Dr., Bethesda, Maryland 20892, USA

Email: Xin Tian* - tianx@nhlbi.nih.gov; Jungnam Joo - jooj@nhlbi.nih.gov; Gang Zheng - zhengg@nhlbi.nih.gov; Jing-Ping Lin - linj@nhlbi.nih.gov

* Corresponding author

*from* Genetic Analysis Workshop 14: Microsatellite and single-nucleotide polymorphism
Noordwijkerhout, The Netherlands, 7-10 September 2004

Published: 30 December 2005

*BMC Genetics* 2005, **6**(Suppl 1):S107 doi:10.1186/1471-2156-6-S1-S107

## Abstract

We studied a trend test for genetic association between disease and the number of risk alleles using case-control data. When the data are sampled from families, this trend test can be adjusted to take into account the correlations among family members in complex pedigrees. However, the test depends on the scores based on the underlying genetic model and thus it may have substantial loss of power when the model is misspecified. Since the mode of inheritance will be unknown for complex diseases, we have developed two robust trend tests for case-control studies using family data. These robust tests have relatively good power for a class of possible genetic models. The trend tests and robust trend tests were applied to a dataset of Genetic Analysis Workshop 14 from the Collaborative Study on the Genetics of Alcoholism.

## Background

Testing for linkage disequilibrium or association provides a useful alternative to testing linkage for complex traits with relatively small genetic effects [1]. Among the tests for association between a candidate-gene and a disease within a case-control design, the Cochran-Armitage (CA) trend test [2,3] is preferable to the allele-based test and the Pearson's chi-squared test [4-6]. In such studies, cases and controls are usually independent random samples. Genotypes on each individual at markers in or near candidate genes are observed. For a marker with two alleles, the CA trend test can be used to test a linear trend between the disease and the number of the high-risk alleles at this marker.

Recently, there has been an increasing interest in statistical methods that evaluate association between genetic markers and disease status using family-based data [7,8]. This would allow data available from linkage studies to be efficiently used to test for association. Unlike the traditional case-control studies in which all individuals are unrelated, cases and controls drawn from family data are often correlated because these individuals are often biologically related. Consequently, the frequencies of the high-risk alleles at a marker locus will be increased among related individuals. This may affect the false positive rate (type I error) for the association test, compared to case-control design based on independent samples. Hence, any test of genetic association must account for the correlations among family members. Slager and Schaid [7] extended the original CA trend test to case-control studies with family data, in which they modeled the correlations among related cases or controls as functions of the probability of their marker alleles shared identically by descent (IBD). This method can be applied to complex family structures and it obtains different correlations for different types of relative pairs. Thus, it is more flexible than the method assuming a common correlation for each pair of relatives within a family. With this correlation adjusted, the resulting trend test in Slager and Schaid [7] is similar to the orig-

**Table 1: The data in a case-control study**

| Status | *NN* | *NM* | *MM* | Total |
|---|---|---|---|---|
| Case | $r_0$ | $r_1$ | $r_2$ | $R$ |
| Control | $s_0$ | $s_1$ | $s_2$ | $S$ |
| Total | $n_0$ | $n_1$ | $n_2$ | $n$ |

inal one but it uses appropriate variance formulation. Note that this trend test uses different scores depending on assumptions of the underlying genetic models. In practice, because the genetic model is unknown for most, if not all, complex diseases, applying a trend test with one set of scores would result in loss of power if the genetic model is misspecified. Therefore, more robust tests have been proposed to protect against model uncertainty [9,10].

In this paper we study the two robust trend tests, the maximum test (MAX) and maximin efficiency robust test (MERT), in case-control design applied to family data. These two robust tests account for the correlated individuals and do not rely on the assumption of any particular genetic model. The performance of the robust trend tests and the extended CA trend test is compared by a simulation study. These tests are illustrated using a Genetic Analysis Workshop 14 dataset from the Collaborative Study on the Genetics of Alcoholism (COGA).

## Methods

### *The trend tests*

Consider data for a case-control study of genetic association as in Table 1. Assume a marker with two alleles: *N* and *M*, where *N* is a normal allele and *M* is an allele with high risk. Denote genotypes as $g_0 = NN$, $g_1 = NM$, and $g_2 = MM$. Let the genotype frequencies for cases and controls to be $p_j$ and $q_j$, $j = 0, 1, 2$, respectively, and $\sum_{j=0}^{2} p_j = \sum_{j=0}^{2} q_j = 1$. Hence, the null hypothesis of no association is to test $p_j = q_j$ for each $j$.

Given the data, the CA trend test for association [4] between a disease and the marker is written as $Z_x = U(\boldsymbol{x})/(\mathrm{Var}[U(\mathbf{x})])^{1/2}$, where $U(\boldsymbol{x}) = \mathrm{n}^{-1}\sum_{\mathrm{j}=0}^{2} \mathrm{x_j}(Sr_j - Rs_j)$, and $\boldsymbol{x} = (x_0, x_1, x_2)'$ is a set of increasing scores (weights) assigned to the three genotypes $(g_0, g_1, g_2)$ *a priori* based on the underlying genetic model. Note that $(x_0, x_1, x_2)'$ can be reparameterized as $(0, x, 1)'$ with $0 \le x \le 1$. If cases and controls are from independent random samples, the counts $(r_0, r_1, r_2)$ and $(s_0, s_1, s_2)$ in Table 1 follow multinomial distributions $mul(R; p_0, p_1, p_2)$ and $mul(S, q_0, q_1, q_2)$, respectively. Under the null hypothesis, it can be shown that $\mathrm{Var}[U(\mathbf{x})] = n^{-1}RS\left[\sum_{j=0}^{2} x_j^2 P_j - \left(\sum_{j=0}^{2} x_j p_j\right)^2\right]$, and $Z_x$ asymptotically follows a standard normal distribution $N(0, 1)$.

The null hypothesis $H_0$ is rejected in favor of the alternative that *M* is the high risk allele associated with disease when $Z_x > z_{1-\alpha}$, where $z_{1-\alpha}$ is the upper $100(1 - \alpha)^{\text{th}}$ percentile of $N(0, 1)$. When it is not certain which allele is high-risk, $H_0$ is rejected when $|\mathrm{Z_x}| > \mathrm{z}_{1-\alpha/2}$.

However, since for case-control studies drawn from family data, cases and controls within the same family may be biologically related, Slager and Schaid [7] proposed the following method for estimating the variance to account for correlations among related cases or controls. Let $\boldsymbol{\gamma}_i = (\gamma_{i0}, \gamma_{i1}, \gamma_{i2})'$ be the genotype indicator vector for the $i^{\text{th}}$ case, where $\gamma_{ij} = 1$ for the $i^{\text{th}}$ case with genotype $\mathrm{g}_j$ and $\gamma_{ij} = 0$ otherwise, $i = 1, \ldots, R$. Similarly, we use $\boldsymbol{z}_j$ for controls. Then $\mathbf{r} = (r_0, r_1, r_2)' = \sum_{i=1}^{R} \boldsymbol{\gamma}_i$, and $\mathbf{s} = (s_0, s_1, s_2)' = \sum_{j=1}^{S} \mathbf{z}_j$. Furthermore, $\boldsymbol{\gamma}_i$ and $\boldsymbol{z}_j$ follow the multinomial distributions $mul(1; p_0, p_1, p_2)$ and $mul(1; q_0, q_1, q_2)$, respectively. Let $\phi = R/n$. The test statistic $U(\mathbf{x})$ can also be written as $U(\mathbf{x}) = \mathbf{x}'[(1 - \phi)\,\mathbf{r} - \phi\mathbf{s}]$. Then,

$$Var[U(\mathbf{x})] = \mathbf{x}'\{\mathbf{Var}[(1-\varphi)\,\mathbf{r} - \varphi\mathbf{s}]\}\mathbf{x}$$
$$= \mathbf{x}'\{(1-\varphi)^2 Var(\sum_{\mathrm{i}} \mathbf{y}_{\mathrm{i}}) + \varphi^2 Var(\sum_{\mathrm{j}} \mathbf{z}_j) - 2\varphi(1-\varphi)Cov(\sum_{i} \mathbf{y}_{i}, \sum_{j} \mathbf{z}_j)\}\mathbf{x},$$

where the variances and covariances can be calculated based on the multinomial distributions and IBD-sharing probabilities for pairs of the related individuals [7],

### *Robust trend tests when the genetic model is unknown*

Because for most complex diseases the underlying genetic model is unknown, we consider two robust trend tests [9,10], the MERT and the MAX in the case-control study,

**Table 2: Empirical powers of trend tests and robust trend tests**

| Model | $(RR_1, RR_2)$ | $Z_{(x = 0)}$ | $Z_{(x = 1/2)}$ | $Z_{(x = 1)}$ | MERT | MAX |
|---|---|---|---|---|---|---|
| Null | (1,1) | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 |
| Recessive | (1,2.6) | 0.80 | 0.62 | 0.26 | 0.70 | 0.76 |
| Additive | (1.2,2.4) | 0.44 | 0.80 | 0.73 | 0.77 | 0.75 |
| Dominant | (1.9,1.9) | 0.18 | 0.72 | 0.80 | 0.64 | 0.73 |

where the *cases* and controls may be related. Note that for the special case in which cases and controls are independent random samples, the tests have been studied by Friedlin et al. [10].

Suppose we have a family of trend test statistics $Z_i$ corresponding to different genetic models. The first robust test, MERT, can be written as a linear combination of the two test statistics with minimum correlation $\rho_0$. Denoting these two tests as $\{Z_s, Z_t\}$, then MERT is written as $Z_{\text{MERT}}=(Z_s + Z_t)/\{2(1 + \rho_0)\}^{1/2}$, which asymptotically follows a standard normal distribution. The second robust trend test, MAX, can be defined as $Z_{\text{MAX}}=\max(Z_s, Z_{\text{MERT}}, Z_t)$ for a one-sided test, and $Z_{\text{MAX}} = \max(|Z_s|, |Z_{\text{MERT}}|, |Z_t|)$ for a two-sided alternative, where $Z_{\text{MERT}}$ is chosen as the "middle" test because it has equal correlations with $Z_s$ and $Z_t$. MAX is more powerful than MERT when $\rho_0$ is small, and the two tests have similar power when the minimum correlation is relatively large (e.g., $\rho_0 \geq 0.75$) [11].

For case-control studies drawn from family data, we can derive the correlations for the trend tests defined in the previous section. Let the variance-covariance matrix $\sum = Var[(1-\varphi)\mathbf{r} - \varphi\mathbf{s}]$. Then the correlation between any two test statistics can be obtained

$$Corr\left[Z_{\mathbf{x}_0}, Z_{\mathbf{x}_1}\right] = \frac{Cov[\mathrm{U}(\mathbf{x}_0), \mathrm{U}(\mathbf{x}_1)]}{(Var[\mathrm{U}(\mathbf{x}_0)])^{1/2}(Var[\mathrm{U}(\mathbf{x}_1)])^{1/2}} = \frac{\mathbf{x}_0' \sum \mathbf{x}_1}{(\mathbf{x}_0' \sum \mathbf{x}_0)^{1/2}(\mathbf{x}_1' \sum \mathbf{x}_1)^{1/2}},$$

where $\mathbf{x}_0$ and $\mathbf{x}_1$ are two sets of scores used for two different genetic models.

To test for association between a marker and disease status, the optimal scores for the recessive, additive, and dominant models are $x = 0$, 1/2, and 1 in $\mathbf{x} = (0, x, 1)'$ [12]. Based on the prior scientific knowledge, other possible choices of genetic models can also be assumed, which leads to different trend tests. The correlation of any two tests can then be calculated to determine the pair of tests with minimum correlation, so the MERT test can be performed. To apply the MAX test, the critical value and the *p*-value are obtained from simulation.

### *The trend tests with multiple alleles*

The above trend tests $Z_x$ can be extended to test the association with a multiallelic marker in a case-control study [7]. For a marker with $K$ different alleles, there are $m = K(K + 1)/2$ possible genotypes and we can obtain a case-control table with $r_i$ and $s_i$, $i = 1, ..., m$, similar to Table 1. The trend test statistic can be written as a $(K\text{-}1) \times 1$ vector, $U = U(\mathbf{X}) = \mathbf{X}'[(1\text{-}\phi)\mathbf{r} - \phi\mathbf{s}]$, where $\mathbf{X}$ is a $m \times (K - 1)$ matrix with the $j^{\text{th}}$ column, $\mathbf{x}_j$, as a score vector for the m genotypes corresponding to the $j^{\text{th}}$ allele, and $\text{Var}(U) = \mathbf{X}'\sum\mathbf{X}$ can be obtained similarly as in the previous section to adjust for correlations among family members. To test the association with this marker, Slager and Schaid [7] proposed to use the statistic $U'[\text{Var}(U)]^{-1}U$ as it asymptotically follows a chi-squared distribution with $(K - 1)$ degrees of freedom.

Here, we can apply MERT and MAX as alternatives to this chi-squared test. Corresponding to the $j^{\text{th}}$ allele, the $j^{\text{th}}$ element *of* $U$ is $U_j = \mathbf{x}'_j[(1\text{-}\phi)\mathbf{r}\text{-}\phi\mathbf{s}]$, and we have $\sigma_j^2 = \text{Var}(U_j) = \mathbf{x}'_j\sum\mathbf{x}_j$ and $\text{Cov}(U_i, U_j) = \mathbf{x}_i'\sum\mathbf{x}_j, i, j = 1,\ldots,(K-1)$. Then the trend test for each allele, $Z_j = \mathrm{U}_j/\sigma_j$, $j = 1,..., (K - 1)$, and the correlation for any two tests can be obtained. Hence, for the family of trend tests, MERT and MAX can be used to test for association with a multi-allelic marker.

## Results

### *A simulation study*

To illustrate the robustness of the statistics, MERT, and MAX, and to compare their performance with individual trend tests for given models, we simulated the case-control datasets and computed the empirical powers for all the tests under three genetic models: the recessive, additive and dominant models.

The simulations were based on the assumptions that the disease prevalence $K = 0.1$ and the allele frequency p = 0.3 with 20,000 replications. To facilitate the calculation, each case-control dataset included 160 cases generated as 80 sib-pairs drawn from 80 different families, and 160 controls as unrelated random samples. It can be shown that the probabilities of 0, 1, 2, alleles shared IBD are 1/4, 1/2, and 1/4 for the sib-pairs when parents' genotype information was unknown. Assuming these IBD probabilities, the variance of the trend test was adjusted for the cor-

**Table 3: A case-control dataset from the COGA study**

| Status | *NN* | *NM* | *MM* | Total |
|---|---|---|---|---|
| Case | 153 | 278 | 178 | 609 |
| Control | 69 | 141 | 51 | 261 |
| Total | 222 | 419 | 229 | 870 |

relations among related cases. Let the genotype relative risks $RR_1 = f_1/f_0$ and $RR_2 = f_2/f_0$, where $f_0$, $f_1$, and $f_2$ are penetrances for genotypes $g_0$, $g_1$, and $g_2$. Thus, equivalently, the null hypothesis $H_0$ can be written as $RR_1 = RR_2 = 1$. The alternative hypothesis can be specified by varying $RR_1$ and $RR_2$.

Table 2 displays the empirical powers of the trend tests and the robust tests, MERT and MAX. The relative risks $RR_1$ and $RR_2$ were chosen so that a particular trend test had about 80% power for each given model. When the true underlying model was recessive inheritance and the corresponding optimal test $Z_{(x = 0)}$ had power of 80%, the tests $Z_{(x = 1/2)}$ and $Z_{(x = 1)}$ only had power of 62% and 26%, respectively. However, the test $Z_{(x = 0)}$ was underpowered when the true model was dominant or additive. Compared to these trend tests, the MERT and MAX tests had relatively good powers for all the three models.

### *Application*

The COGA data consist of 1,614 individuals from 143 families, with alcoholism diagnosis, microsatellite, and single-nucleotide polymorphism (SNP) marker information. The preliminary genome scan by linkage analysis using the microsatellite data suggested that *ADH3* of chromosome 4 may be an alcoholism susceptibility gene. Without adjusting for family structure, a logistic regression with backward selection of SNPs from the Illumina dataset near the *ADH* genes indicated that SNP marker rs1037475 was a significant predictor. Here we applied the association tests to case-control data using the ALDX1 diagnosis of "affected" and "purely unaffected" status to define case status and genotypes for this SNP marker. Table 3 presents the data including cases from 143 families and controls from 111 families.

Results of trend tests for the data in Table 3 with or without adjusting for the family-based correlations are shown in Figure 1. For individuals from the same family, their shared alleles IBD probabilities were calculated using software GENEHUNTER [13], and the correlations and the adjusted variances of the test statistics were obtained. We then applied the two-sided trend tests under recessive, additive, and dominant models, corresponding to the scores $x = 0$, $1/2$, and 1. The tests showed significant association under both the recessive and additive model assumption ($Z_{(x = 0)} = 2.89$, $p = 0.004$; $Z_{(x = 1/2)} = 2.02$, $p = 0.043$), but it failed to show any significant result assuming a dominant model ($Z_{(x = 1)} = 0.40$, $p = 0.69$). Note that after adjusting for the correlations among family members, standard errors were larger, resulting in smaller test statistics $Z_x$ and thus larger $p$-values compared to the tests without adjusting for the correlations (see Figure 1).

Figure 1 also shows the trend test results depend on the scores $x = (0, x, 1)$ for the underlying genetic models. The trend tests $Z_x$ with $0 \le x \le 1$ correspond to different models, where the statistics $Z_x$ above the horizontal dotted line are significant. Due to the uncertainty about the mode of inheritance, different conclusions could be reached and using any single trend test may result in significant loss of power when the model is misspecified. Therefore, we also applied the two robust tests to these data. Given the tests for the recessive, additive, and dominant models, the pairwise correlations were calculated as $\mathrm{Corr}(Z_{(x = 0)}, Z_{(x = 1)}) = 0.334$, $\mathrm{Corr}(Z_{(x = 0)}, Z_{(x = 1/2)}) = 0.818$, and $\mathrm{Corr}(Z_{(x = 1/2)}, Z_{(x = 1)}) = 0.813$. Then we obtained $Z_{\mathrm{MERT}} = (2.89 + 0.40)/\{2(1 + 0.334)\}^{1/2} = 2.01$ with $p$-value = 0.044. By simulations with 1,000,000 replications, the empirical $p$-value for $Z_{\mathrm{MAX}} = 2.89$ was $p = 0.009$. In this example, because the correlation between the test statistics under the recessive and dominant models is small, MAX appears to be more powerful than MERT to detect associations between disease status and a marker. Both robust trend tests showed significant association between this SNP marker and alcoholism.

## Conclusion

In this paper, we applied the trend tests of genetic association to case-control studies drawn from the COGA families. Although the significant results under the recessive, additive, and dominant models were similar for this example, the tests ignoring the correlations among family members would have yielded large false-positive rates and moreover, unadjusted tests would not be valid.

We have also studied two robust trend tests, MERT and MAX, for case-control studies with family data. When the genetic model is unknown, these robust tests based on a family of possible genetic models tend to be more con-

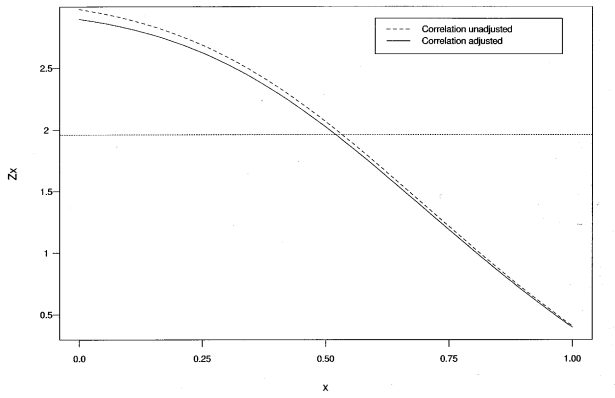


**Figure 1**
Plot of the trend tests $Z_x$ versus scores *x*.

servative against model misspecification. Although we have focused on the examples and models for genetic association, these results hold generally for trend tests of association with correlated cases or controls when the exposure variables have some natural ordering.

## Abbreviations

CA: Cochran-Armitage

COGA: Collaborative Study on the Genetics of Alcoholism

IBD: Identical by descent

MAX: Maximum test

MERT: Maximin efficiency robust test

SNP: Single-nucleotide polymorphism

## Authors' contributions

XT involved in the design of the study and statistical analysis, and drafted the manuscript. JJ, GZ, and JPL participated in its design and performed the statistical analysis. All authors read and approved the final manuscript.

## References


1. Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996, 273:1516-1517.
2. Armitage P: **Tests for linear trends in proportions and frequencies.** *Biometrics* 1955, **11:**375-386.
3. Cochran WG: **Some methods for strengthening the common chi-squared tests.** *Biometrics* 1954, **10:**417-451.
4. Sasieni PD: **From genotypes to genes: doubling the sample size.** *Biometrics* 1997, **53:**1253-1261.
5. 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:**149-153.
6. Czika W, Weir BS: **Properties of the multiallelic trend test.** *Biometrics* 2004, **60:**69-74.
7. Slager SL, Schaid DJ: **Evaluation of candidate genes in case-control studies: a statistical method to account for related subjects.** *Am J Hum Genet* 2001, **68:**1457-1462.
8. Rabinowitz D, Laird NM: **A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information.** *Hum Hered* 2000, **50:**211-223.
9. Gastwirth JL: **The use of maximin efficiency robust tests in combining contingency tables and survival analysis.** *J Am Stat Assoc* 1985, **80:**380-384.
10. 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:**146-152.
11. Freidlin H, Podgor MJ, Gastwirth JL: **Efficiency robust tests for survival or ordered categorical data.** *Biometrics* 1999, **55:**883-886.
12. Zheng G, Freidlin B, Gastwirth JL: **Choice of scores in trend tests for case-control studies of candidate-gene associations.** *Biometrical J* 2003, **45:**335-348.
13. Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES: **Parametric and nonparametric linkage analysis: a unified multipoint approach.** *Am J Hum Genet* 1996, **58:**1347-1363.