# Report

## Robust Genomic Control for Association Studies

Gang Zheng,[1] Boris Freidlin,[2] and Joseph L. Gastwirth[3]

[1]Office of Biostatistics Research, Division of Epidemiology and Clinical Applications, National Heart, Lung and Blood Institute, and [2]Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, MD; and [3]Department of Statistics, George Washington University, Washington, DC

**Population-based case-control studies are a useful method to test for a genetic association between a trait and a marker. However, the analysis of the resulting data can be affected by population stratification or cryptic relatedness, which may inflate the variance of the usual statistics, resulting in a higher-than-nominal rate of false-positive results. One approach to preserving the nominal type I error is to apply genomic control, which adjusts the variance of the Cochran-Armitage trend test by calculating the statistic on data from null loci. This enables one to estimate any additional variance in the null distribution of statistics. When the underlying genetic model (e.g., recessive, additive, or dominant) is known, genomic control can be applied to the corresponding optimal trend tests. In practice, however, the mode of inheritance is unknown. The genotype-based $\chi^2$ test for a general association between the trait and the marker does not depend on the underlying genetic model. Since this general association test has 2 degrees of freedom (df), the existing formulas for estimating the variance factor by use of genomic control are not directly applicable. By expressing the general association test in terms of two Cochran-Armitage trend tests, one can apply genomic control to each of the two trend tests separately, thereby adjusting the $\chi^2$ statistic. The properties of this robust genomic control test with 2 df are examined by simulation. This genomic control–adjusted 2-df test has control of type I error and achieves reasonable power, relative to the optimal tests for each model.**

For mapping disease-susceptibility genes for complex human diseases, case-control studies testing linkage disequilibrium or association are useful approaches for detecting markers with small-to-moderate genetic effects on traits (Risch and Merikangas 1996; Khoury and Yang 1998). However, because of population stratification or cryptic relatedness, case-control studies may produce spurious associations. Case-control studies, on the other hand, are easier than family-based association studies to conduct, because they use population controls and do not require genetic data from family members. Statistical methods have been developed for adjusting population stratification and/or cryptic relatedness in case-control studies. One is based on inferring the number of strata in a population and estimating the probability of each sample member belonging to these strata (Pritchard and Rosenberg 1999; Pritchard et al. 2000; Satten et al. 2001; Zhu et al. 2002). Another approach is genomic control (GC) (Devlin and Roeder 1999; Bacanu et al. 2000; Devlin et al. 2001; Reich and Goldstein 2001; Zheng et al. 2005), which adjusts the variance of the Cochran-Armitage trend test by use of data from null loci. Here, we focus on developing a robust GC test.

In case-control studies, the Cochran-Armitage (CA) trend tests are preferred to the allele-based test, as they are valid when Hardy-Weinberg equilibrium (HWE) does not hold. Furthermore, the two types of tests are asymptotically equivalent under HWE (Sasieni 1997). To apply the CA trend test, increasing scores are assigned a priori to the genotypes. Thus, the trend statistic is a function of scores. The choice of scores depends on the underlying genetic model—for example, recessive, additive, or dominant (Sasieni 1997; Zheng et al. 2003)—which is a typical problem in the application of trend tests (Graubard and Korn 1987). The GC developed by Devlin and Roeder (1999) was based on the trend test with scores optimal for the additive model. Zheng et al. (2005) studied GC for recessive and dominant models. For many complex diseases, the underlying genetic models are usually unknown, and a single trend test for case-control studies may lose substantial power when the model is misspecified (Freidlin et al. 2002). Thus, an

Received September 13, 2005; accepted for publication November 23, 2005; electronically published December 22, 2005.

Address for correspondence and reprints: Dr. Gang Zheng, Office of Biostatistics Research, Division of Epidemiology and Clinical Applications, National Heart, Lung, and Blood Institute, 6701 Rockledge Drive, Bethesda, MD 20892. E-mail: zhengg@nhlbi.nih.gov

*Am. J. Hum. Genet.* 2006;78:350–356. © 2005 by The American Society of Human Genetics. All rights reserved. 0002-9297/2006/7802-0017$15.00

**Table 1**

**Genotype Distributions for Case-Control Data**

| | Genotype | | | |
|---|---|---|---|---|
| Data | *NN* | *NM* | *MM* | Total |
| Case | $r_0$ | $r_1$ | $r_2$ | $r$ |
| Control | $s_0$ | $s_1$ | $s_2$ | $s$ |
| Total | $n_0$ | $n_1$ | $n_2$ | $n$ |

efficiency-robust test (Gastwirth 1966) having fairly high power across a set of models should be useful. Here, we show that the usual $\chi^2$ test of general association (GA) between the disease and the marker is robust and can be modified to account for population stratification. This test is widely used in genetic data analysis and is also supported by many existing software packages (Weir 1996; Sham 1998; Gibson and Muse 2004).

The GC method adjusts the variance of a trend test by estimating the variance inflation caused by population stratification by use of the null loci. It is not directly applicable to the GA test statistic, which has a complicated variance-covariance matrix. A direct adjustment at the scale level may not be applicable. To circumvent this problem, we express the GA test in terms of two CA trend tests. Then, adjusting each CA trend test by the usual GC method provides the adjustment of the GA test.

Consider a genetic marker with two alleles $M$ and $N$ with frequencies $p$ and $q = 1 - p$, respectively, where $M$ is a disease-associated allele, referred to as the "risk allele." The genotype distributions of case-control data are displayed in table 1, where $(r_0,r_1,r_2)$ and $(s_0,s_1,s_2)$ are genotype counts of cases and controls. They are independent and follow multinomial distributions $(r_0,r_1,r_2) \sim \text{Mul}(r; p_0,p_1,p_2)$ and $(s_0,s_1,s_2) \sim \text{Mul}(s; q_0,q_1,q_2)$. Denote the disease prevalence in the population as $K = \Pr(\text{disease})$; the genotypes as $G_0 = NN$, $G_1 = NM$, and $G_2 = MM$; and their frequencies by $g_i = \Pr(G_i)$, $i = 0,1,2$. The penetrances are defined as the conditional probabilities of disease given each of three genotypes

$$f_i = \Pr(\text{disease} \mid G_i), i = 0,1,2 \ .$$

The genotype frequencies can be written as $p_i = \Pr(G_i \mid \text{disease}) = g_i f_i / K$ and $q_i = \Pr(G_i \mid \text{control}) = g_i(1 - f_i)/(1 - K)$ for $i = 0,1,2$ in cases and controls. Under the null hypothesis of no association, $H_0: p_i = q_i = g_i$ for $i = 0,1,2$; that is, $H_0: f_0 = f_1 = f_2 = K$. As $M$ is a risk allele, under the alternative hypothesis $H_1: f_0 \leqslant f_1 \leqslant f_2$ with at least one equality strictly holding. A genetic model is recessive, additive, or dominant when the penetrances satisfy $f_1 = (1 - \lambda)f_0 + \lambda f_2$ and $\lambda = 0$, 1/2, or 1, respectively. For local alternatives ($f_0 \approx f_1 \approx f_2$), the multiplicative model $f_1^2 = f_0 f_2$ is equivalent to the additive model. To see this, write $\gamma_1 = f_1/f_0 = 1 + \epsilon_1 > 1$ and $\gamma_2 = f_2/f_1 = 1 + \epsilon_2 > 1$, and $\epsilon_i \approx 0$ for $i = 1,2$. Then, the additive model implies that $\epsilon_2 = 2\epsilon_1$. Thus, $\gamma_1^2 = (1 + \epsilon_1)^2 \approx 1 + 2\epsilon_1 = 1 + \epsilon_2 = \gamma_2$.

When the genetic model is known, a more powerful and directed test is the CA trend test (Agresti 1990). To apply this CA trend test, increasing scores $(0,x,2)$ are assigned to three genotypes $(NN,NM,MM)$, respectively, where $0 \leqslant x \leqslant 2$. The trend test can be written (Sasieni 1997) as

$$Z(x) = \frac{n^{1/2}\left[\sum_{i=0}^{2} x_i(sr_i - rs_i)\right]}{\left\{rs\left[n\sum_{i=0}^{2} x_i^2 n_i - \left(\sum_{i=0}^{2} x_i n_i\right)^2\right]\right\}^{1/2}} \ , \qquad (1)$$

where $(x_0,x_1,x_2) = (0,x,2)$. For a given $x$, $Z(x)$ asymptotically follows a standard normal distribution under $H_0$. Thus, the null hypothesis is rejected when $|Z(x)| > z_{1-\alpha/2}$. The trend test $Z(x)$ is optimal when $x$ is properly specified a priori. For recessive, additive (multiplicative), and dominant models, the respective values of optimal $x$ are 0, 1, 2. From equation (1), it follows that the trend test $Z(x)$ is invariant to a linear transformation of $x$—that is, the scores $(0,x,2)$ and $(0,x/2,1)$ yield the same trend test. Thus, a general model can be expressed as $f_1 = (1 - \lambda)f_0 + \lambda f_2$, where $\lambda \in [0,1]$ and the optimal choice of $x$ in $Z(x)$ is $\lambda$ (Zheng et al. 2003). Unfortunately, $Z(x)$ is not robust to a misspecification of the genetic model.

The GC of Devlin and Roeder (1999) is based on $Z(1)$, the optimal test for the additive model. When the population is stratified, they considered the test statistic $Z_*^2(1) = Z^2(1)/\hat{\lambda}(1)$, which follows $\chi^2$ distribution with 1 df ($\chi_1^2$), where $\lambda(1)$ is the variance inflation factor that can be estimated using the null loci. Let the trend test $Z(1)$, calculated on $c$ null loci, be denoted as $Z_1(1), \ldots, Z_c(1)$, which are realizations of a random variable $Z_0(1)$, where the subscript 0 indicates the null loci, which are not associated with diseases and are not under linkage disequilibrium with the disease loci. Since $Z_*^2(1)$ follows $\chi_1^2$ and $\lambda(1)$ is a constant, $\lambda(1) = E[Z_0^2(1)]$. Therefore, $\lambda(1)$ can be estimated by the expected value of a random variable by using its realizations $Z_1^2(1), \ldots, Z_c^2(1)$. Devlin and Roeder (1999) studied both Bayesian and frequentist approaches for estimating $\lambda(1)$. Here, we use the latter—that is, $\hat{\lambda}(1) = \text{median}\,[Z_1^2(1), \ldots, Z_c^2(1)]/0.456$. Zheng et al. (2005) showed that the idea can be applied to the optimal tests for the recessive and dominant models. Also, we assume that the minor-allele frequencies of null loci are close to that of the marker (Reich and Goldstein 2001).

A robust test that does not depend on the underlying

**Table 2**

**Type I Error and Empirical Power Performance of Three GC Trend Tests With No Population Stratification**

| Allele Frequency and Model | $Z_*(2)$ | $Z_*(1)$ | $Z_*(0)$ | $T_2^*$ | $T_2^{**}$ |
|---|---|---|---|---|---|
| $p = .1$: | | | | | |
| Null[a] | .063 | .062 | .052 | .043 | .041 |
| Null[b] | .051 | .050 | .040 | .028 | .028 |
| DOM[c] | .795 | .772 | .079 | .681 | |
| ADD[d] | .790 | .800 | .161 | .702 | |
| REC[e] | .176 | .424 | .795 | .678 | |
| $p = .5$: | | | | | |
| Null[a] | .063 | .062 | .062 | .056 | .068 |
| Null[b] | .051 | .051 | .049 | .037 | .037 |
| DOM[f] | .811 | .622 | .158 | .703 | |
| ADD[g] | .651 | .774 | .572 | .701 | |
| REC[h] | .193 | .677 | .818 | .715 | |

NOTE.—Type I error and empirical power performance are shown for the three GC trend tests $Z_*(2)$, $Z_*(1)$, and $Z_*(0)$ under the dominant (DOM), additive (ADD), and recessive (REC) models and for the RGC $T_2^*$ and the 2-df $\chi^2$ test with direct GC adjustment, $T_2^{**}$, by use of two subpopulations of sizes $a_1 = 200$, $a_2 = 0$ for cases and $b_1 = 0$, $b_2 = 200$ for controls, with no population stratification ($F = 0$), and two-sided $\alpha = 0.05$ with 10,000 replications for power and 100,000 for type I error. In all models, the baseline penetrance $f_0 = 0.1$.

[a] With GC.
[b] Without GC.
[c] $f_1 = f_2 = 0.18$.
[d] $f_1 = 0.175$, $f_2 = 0.25$.
[e] $f_1 = 0.1$, $f_2 = 0.552$.
[f] $f_1 = f_2 = 0.187$.
[g] $f_1 = 0.15$, $f_2 = 0.2$.
[h] $f_1 = 0.1$, $f_2 = 0.175$.

genetic model is a test of the general association for the 2 × 3 table (table 1), which is given by

$$T_{\mathrm{GA}} = \frac{(r_0 - \frac{n_0 r}{n})^2}{\frac{n_0 r}{n}} + \frac{(r_1 - \frac{n_1 r}{n})^2}{\frac{n_1 r}{n}} + \frac{(r_2 - \frac{n_2 r}{n})^2}{\frac{n_2 r}{n}} + \frac{(s_0 - \frac{n_0 s}{n})^2}{\frac{n_0 s}{n}} + \frac{(s_1 - \frac{n_1 s}{n})^2}{\frac{n_1 s}{n}} + \frac{(s_2 - \frac{n_2 s}{n})^2}{\frac{n_2 s}{n}} . \quad (2)$$

Under the null hypothesis of no association between disease status and genotypes, $T_{\mathrm{GA}}$ follows asymptotically the $\chi^2$ distribution with 2 df ($\chi_2^2$). Note that GC has been applied to the test statistics that have a $\chi_1^2$ distribution, whereas $T_{\mathrm{GA}}$ has a $\chi_2^2$ distribution. Thus, direct application of GC to equation (2) is inappropriate.

However, the general association test, $T_{\mathrm{GA}}$, is asymptotically equivalent to the 2-df score test obtained from the logistic regression model. Define two indicator variables $(x_1,x_2)$ as (0,0), (0,1), and (1,1) to designate the genotypes *NN*, *NM*, and *MM*, respectively. For the *j*th individual, his genotype is denoted by two indicator variables $(x_{1j},x_{2j})$, and its status is denoted as $y_j = 1$ for case and $y_j = 0$ for control. Then, applying

$$\Pr(y_j = 1|x_{1j},x_{2j}) = \frac{\exp(\alpha + \beta_1 x_{1j} + \beta_2 x_{2j})}{1 + \exp(\alpha + \beta_1 x_{1j} + \beta_2 x_{2j})} ,$$

the likelihood function is proportional to

$$L(\alpha,\beta_1,\beta_2) = \prod_{j=1}^{n} \{[\Pr(y_j = 1|x_{1j},x_{2j})]^{y_j} \times [1 - \Pr(y_j = 1|x_{1j},x_{2j})]^{1-y_j}\} .$$

The null hypothesis of no association is $H_0$:$\beta_1 = \beta_2 = 0$. The score function evaluated under $H_0$ can be written (see appendix A) as

$$U_1 = \frac{\partial L}{\partial \beta_1}|_{H_0,\hat{\alpha}} = \frac{1}{n}(sr_2 - rs_2)$$

$$U_2 = \frac{\partial L}{\partial \beta_2}|_{H_0,\hat{\alpha}} = \frac{1}{n}[s(r_1 + r_2) - r(s_1 + s_2)] ,$$

where $\hat{\alpha}$ is the maximum-likelihood estimate of the nuisance parameter $\alpha$ under $H_0$. Denote $\mathbf{U}$ as $(U_1,U_2)$ and the observed Fisher information matrix evaluated under $H_0$ and $\alpha = \hat{\alpha}$ as $I(\alpha,\beta_1,\beta_2)$. The submatrix of $I^{-1}(\alpha,\beta_1,\beta_2)$ corresponding to $(\beta_1,\beta_2)$ is denoted by $\Sigma^{-1}$ and is a consistent estimate of the inverse of the covariance matrix of $\mathbf{U}$. Thus, under $H_0$,

$$T_2 = \mathbf{U}^T\Sigma^{-1}\mathbf{U} = \frac{1}{1 - \hat{\rho}^2}[Z^2(0) + Z^2(2) - 2\hat{\rho}Z(0)Z(2)] \quad (3)$$

has an asymptotic $\chi_2^2$ distribution, where

$$\hat{\rho} = \left(\frac{n_0 n_2}{(n_1 + n_2)(n_0 + n_1)}\right)^{1/2} \quad (4)$$

is a consistent estimator of the null correlation between $Z(0)$ and $Z(2)$ (appendix A). Note that $T_2$ is approximately $\chi_2^2$ when there is no population stratification. To adjust for possible population stratification, we can apply GC to equation (3) by replacing $Z(0)$ and $Z(2)$ by $Z_*(0)$ and $Z_*(2)$, respectively, and $\hat{\rho}$ by $\rho_*$, which is estimated using null loci. The resulting test statistic will be referred to as the "robust genomic control"

**Table 3**

**Type I Error and Empirical Power Performance of Three GC Trend Tests With Population Stratification**

| *F*, Allele Frequency, and Model | $Z_*(2)$ | $Z_*(1)$ | $Z_*(0)$ | $T_2^*$ | $T_2^{**}$ |
|---|---|---|---|---|---|
| $F = .005$: | | | | | |
| $p = .1$: | | | | | |
| Null[a] | .063 | .061 | .031 | .041 | .081 |
| Null[b] | .254 | .260 | .072 | .192 | .192 |
| DOM[c] | .807 | .769 | .087 | .683 | |
| ADD[d] | .803 | .798 | .239 | .721 | |
| REC[e] | .153 | .282 | .790 | .631 | |
| $p = .2$: | | | | | |
| Null[a] | .062 | .061 | .054 | .053 | .091 |
| Null[b] | .240 | .258 | .125 | .201 | .201 |
| DOM[f] | .823 | .739 | .156 | .703 | |
| ADD[g] | .790 | .778 | .464 | .733 | |
| REC[h] | .160 | .353 | .811 | .685 | |
| $p = .5$: | | | | | |
| Null[a] | .062 | .061 | .061 | .055 | .097 |
| Null[b] | .199 | .260 | .200 | .206 | .206 |
| DOM[i] | .795 | .481 | .140 | .654 | |
| ADD[j] | .789 | .799 | .673 | .786 | |
| REC[k] | .198 | .569 | .800 | .673 | |
| $F = .05$: | | | | | |
| $p = .2$: | | | | | |
| Null[a] | .052 | .043 | .045 | .045 | .141 |
| Null[b] | .657 | .674 | .458 | .628 | .628 |
| DOM[l] | .795 | .689 | .131 | .623 | |
| ADD[m] | .734 | .717 | .505 | .677 | |
| REC[n] | .168 | .281 | .779 | .654 | |
| $p = .5$: | | | | | |
| Null[a] | .052 | .048 | .053 | .047 | .142 |
| Null[b] | .613 | .679 | .613 | .637 | .637 |
| DOM[o] | .804 | .449 | .130 | .633 | |
| ADD[p] | .741 | .796 | .776 | .777 | |
| REC[q] | .149 | .443 | .735 | .558 | |

NOTE.—Type I error and empirical power performance of the three GC trend tests $Z_*(2)$, $Z_*(1)$, and $Z_*(0)$ under the dominant (DOM), additive (ADD), and recessive (REC) models and for the RGC $T_2^*$ and the 2-df $\chi^2$ test with direct GC adjustment, $T_2^{**}$, by use of two subpopulations of sizes $a_1 = 200$, $a_2 = 0$ for cases and $b_1 = 0$, $b_2 = 200$ for controls, with population stratification, and two-sided $\alpha = 0.05$ with the same replications as in table 2. In all models, the baseline penetrance $f_0 = 0.1$

[a] With GC.
[b] Without GC.
[c] $f_1 = f_2 = 0.264$.
[d] $f_1 = 0.254$, $f_2 = 0.408$.
[e] $f_1 = 0.1$, $f_2 = 0.717$.
[f] $f_1 = f_2 = 0.232$.
[g] $f_1 = 0.213$, $f_2 = 0.326$.
[h] $f_1 = 0.1$, $f_2 = 0.366$.
[i] $f_1 = f_2 = 0.254$.
[j] $f_1 = 0.215$, $f_2 = 0.330$.
[k] $f_1 = 0.1$, $f_2 = 0.225$.
[l] $f_1 = f_2 = 0.65$.
[m] $f_1 = 0.545$, $f_2 = 0.99$.
[n] $f_1 = 0.1$, $f_2 = 0.98$.
[o] $f_1 = f_2 = 0.715$.
[p] $f_1 = 0.52$, $f_2 = 0.94$.
[q] $f_1 = 0.1$, $f_2 = 0.539$.

(RGC) test and is denoted as $T_2^* = [Z_*^2(0) + Z_*^2(2) - 2\rho_* Z_*(0)Z_*(2)]/(1 - \rho_*^2)$, which has a $\chi^2$ distribution of $\chi_2^2$ under $H_0$ and population stratification. The fact that the RGC test is a function of the adjusted optimal test statistics for the two extreme genetic models, recessive and dominant, is not surprising, since the optimal tests for the "extreme" models are components of nearly all efficiency-robust tests (Gastwirth 1966, 1985).

To evaluate the performance of the proposed genotype-based $\chi^2$ test, we conducted simulation studies and estimated empirical power and type I error for three trend tests $Z_*(2)$, $Z_*(1)$, and $Z_*(0)$ and the RGC test $T_2^*$ under a range of underlying conditions and genetic models. For comparison, we also applied the GC adjustment to the 2-df $\chi^2$ test statistic (eq. [3]). This modified GA test is denoted as $T_2^{**}$. The SAS macro running the simulations is available on request. In the simulations, we assumed that the candidate gene and the null loci have the same minor-allele frequency. Our simulations follow an algorithm similar to that of Devlin and Roeder (1999), Bacanu et al. (2000), and Zheng et al. (2005), which assumes that each subpopulation is in HWE. We specified the minor-allele frequency $p$, the Wright's coefficient of inbreeding $F$, the penetrances $f_0$, $f_1$, and $f_2$ under various genetic models, the sample sizes of cases $a_k$ and controls $b_k$ for the $k$th subpopulation $k = 1, \ldots, m$, and the number of null loci $c$ used to estimate variance inflation factors. In step 1, the allele frequency $p_k$ was generated for the $k$th subpopulation from the beta distribution, Beta$[(1 - F)p/F, (1 - F)q/F]$, for $k = 1, \ldots, m$. In step 2, for individuals from the $k$th subpopulation, two alleles were drawn at random from the binomial distribution $(2, p_k)$ to create a genotype at the candidate allele locus. Disease status was randomly generated conditional on the number, $i$, of candidate alleles in the genotype by use of the Bernoulli distribution with parameter $f_i$. The process continued until $a_k$ cases and $b_k$ controls were obtained. In step 3, genotypes for each of $c$ null loci were generated using the same beta-binomial algorithm as above. The statistics $Z_k(j)$ $(j = 0,1,2)$ at the $k$th locus $(k = 1, \ldots, c)$ were calculated, and the variance inflation factors, $\lambda(j)$, were estimated as $\hat{\lambda}(j) = \text{median}[Z_1^2(j), \ldots, Z_c^2(j)]/0.456$. Then, the GC trend test statistics $Z_*(j)$ were obtained by $Z_*(j) = Z(j)/\hat{\lambda}^{1/2}(j)$. The RGC test $T_2^*$ was calculated using equations (3) and (4), with $\rho_*$ estimated as the average of $\hat{\rho}$ over $c$ null loci. For the 2-df $\chi^2$ test with direct GC adjustment, $T_2^{**}$, we calculated $T_{2,k}$ for the $k$th null locus $(k = 1, \ldots, c)$ and the variance inflation factors as $\hat{\lambda}(T_2) = \text{median}(T_{2,1}, \ldots, T_{2,c})/1.386$, where 1.386 is the median of the $\chi^2$ distribution $\chi_2^2$. Then, $T_2^{**} = T_2/\hat{\lambda}(T_2)$, where $\rho_*$ was used in place of $\hat{\rho}$ in equation (3).

**Table 4**

**Type I Error and Empirical Power Performance of Three GC Trend Tests with Larger Population Sizes and Population Stratification**

| $F$, Allele Frequency, and Model | $Z_*(2)$ | $Z_*(1)$ | $Z_*(0)$ | $T_2^*$ | $T_2^{**}$ |
|---|---|---|---|---|---|
| $F = 0$: | | | | | |
| $p = .1$: | | | | | |
| Null[a] | .062 | .063 | .058 | .053 | .059 |
| Null[b] | .050 | .049 | .052 | .038 | .038 |
| DOM[c] | .793 | .768 | .101 | .680 | |
| ADD[d] | .788 | .797 | .213 | .710 | |
| REC[e] | .124 | .310 | .788 | .672 | |
| $F = .005$: | | | | | |
| $p = .1$: | | | | | |
| Null[a] | .061 | .061 | .060 | .055 | .096 |
| Null[b] | .284 | .293 | .103 | .233 | .233 |
| DOM[f] | .813 | .774 | .146 | .696 | |
| ADD[g] | .788 | .781 | .404 | .731 | |
| REC[h] | .105 | .202 | .819 | .697 | |
| $F = .05$: | | | | | |
| $p = .1$: | | | | | |
| Null[a] | .044 | .038 | .060 | .055 | .141 |
| Null[b] | .691 | .696 | .336 | .652 | .652 |
| DOM[i] | .800 | .767 | .278 | .686 | |
| ADD[j] | .777 | .770 | .607 | .747 | |
| REC[k] | .115 | .189 | .784 | .680 | |

NOTE.—Type I error and empirical power performance of the three GC trend tests $Z_*(2)$, $Z_*(1)$, and $Z_*(0)$ under the dominant (DOM), additive (ADD), and recessive (REC) models and for the RGC $T_2^*$ and the 2-df $\chi^2$ test with direct GC adjustment, $T_2^{**}$, by use of two subpopulations of sizes $a_1 = 750$, $a_2 = 250$ for cases and $b_1 = 250$, $b_2 = 750$ for controls, with population stratification, and two-sided $\alpha = 0.05$ with the same replications as in table 2. In all models, the baseline penetrance $f_0 = 0.1$

[a] With GC.
[b] Without GC.
[c] $f_1 = f_2 = 0.132$.
[d] $f_1 = 0.13$, $f_2 = 0.16$.
[e] $f_1 = 0.1$, $f_2 = 0.246$.
[f] $f_1 = f_2 = 0.166$.
[g] $f_1 = 0.161$, $f_2 = 0.222$.
[h] $f_1 = 0.1$, $f_2 = 0.301$.
[i] $f_1 = f_2 = 0.374$.
[j] $f_1 = 0.355$, $f_2 = 0.61$.
[k] $f_1 = 0.1$, $f_2 = 0.630$.

Table 2 reports the type I error rates and empirical power of three trend tests and two 2-df $\chi^2$ tests after GC corrections when there is no population stratification. Only the power for $T_2^*$ is reported. When there was no population stratification, the GC-adjusted type I error rates for three trend tests and the directly GC-adjusted $\chi^2$ test $T_2^{**}$ were slightly greater than those of the corresponding unadjusted tests, because of the variation that GC method adds by estimating the variance inflation factor from $c$ null loci. For the RGC statistic $T_2^*$, the type I error rates were $\alpha < 0.05$ because of the estimation of the null correlation. For empirical power comparison, when the genetic model is unknown, a test statistic is highly efficiency robust if it has high minimum power across the genetic models—that is, if it has high power when the model is misspecified. From table 2, the RGC test $T_2^*$ was efficiency robust, relative to each of the three trend tests optimal for a specific genetic model. Across the three genetic models, $Z_*(0)$ or $Z_*(2)$ had power <20% when the dominant or recessive model is true, respectively. The trend test optimal for the additive model $Z_*(1)$ was the most efficiency robust among three trend tests. However, $T_2^*$ had greater minimum power than $Z_*(1)$. When there was population stratification (tables 3 and 4), type I error rates for all tests were inflated. When GC controls were applied, the type I error for the three trend tests and RGC was near the nominal 0.05 level. Use of direct GC adjustment of the 2-df $\chi^2$ statistic (eq. [3]), however, failed to fully adjust for population stratification. The pattern of power performance among three GC trend tests and the RGC was similar to that shown in table 2 when there is no population stratification.

For genetic case-control association studies when the genetic model is unknown and there is no population stratification, the $\chi^2$ test with 2 df testing the general association is highly efficient, relative to the optimal trend tests (Zheng et al. 2006). Moreover, this genotype-based 2-df $\chi^2$ test has been applied more often in genetic association studies than the trend tests have. When there is population stratification, the type I error rates may be inflated using either the trend tests or the 2-df $\chi^2$ test because of the inflation of variances of the test statistics. GC is a useful method for adjusting the variance of three trend tests to ensure the desired type I error rate. However, GC cannot be directly applied to the 2-df $\chi^2$ test statistic, which has the complicated variance-covariance matrix. After expressing the 2-df score test from the logistic regression model as a function of two trend tests, we apply the GC approach to the 2-df $\chi^2$ test by adjusting the variance of each trend test. Simulation results show that this 2-df $\chi^2$ test is efficiency robust across the recessive, additive, and dominant models, compared with the three GC trend tests.

## Acknowledgment

The research of J.L.G. was partially supported by grant SES-0317956 from the National Science Foundation.

## Appendix A

The log-likelihood can be written as $\log L(\alpha,\beta_1,\beta_2) = r\alpha + \beta_1 r_2 + \beta_2(r_1 + r_2) - n_0 \log[1 + \exp(\alpha)] - n_1 \log[1 + \exp(\alpha + \beta_2)] - n_2 \log[1 + \exp(\alpha + \beta_1 + \beta_2)]$. From $\partial \log L/\partial\alpha|_{H_0} = 0$, $\hat{\alpha} = \log(r/s)$ is the maximum-likelihood estimate of $\alpha$. The score functions are given by $\partial \log L/\partial\beta_1 = r_2 - n_2 \exp(\alpha + \beta_1 + \beta_2)/[1 + \exp(\alpha + \beta_1 + \beta_2)]$ and $\partial \log L/\partial\beta_2 = (r_1 + r_2) - n_1 \exp(\alpha + \beta_2)/[1 + \exp(\alpha + \beta_2)] - n_2 \exp(\alpha + \beta_1 + \beta_2)/[1 + \exp(\alpha + \beta_1 + \beta_2)]$. Hence, $U_1$ and $U_2$ are obtained by substituting $\beta_1 = \beta_2 = 0$ and $\exp(\hat{\alpha}) = r/s$.

Let $I(\alpha,\beta_1,\beta_2)$ be the observed Fisher information matrix evaluated under $H_0$ and $\alpha = \hat{\alpha}$. Then, under $H_0$ and $\alpha = \hat{\alpha}$, $-\partial^2 \log L/\partial\alpha^2 = n\phi(1-\phi)$, $-\partial^2 \log L/(\partial\alpha\partial\beta_1) = -\partial^2 \log L/\partial\beta_1^2 = -\partial^2 \log L/\partial\beta_1\beta_2 = n_2\phi(1-\phi)$, and $-\partial^2 \log L/\partial\alpha\partial\beta_2 = -\partial^2 \log L/\partial\beta_2^2 = (n_1 + n_2)\phi(1-\phi)$, where $\phi = r/n$. Thus,

$$I^{-1}(\alpha,\beta_1,\beta_2) = \frac{1}{\phi(1-\phi)} \begin{pmatrix} \frac{1}{n_0} & 0 & -\frac{1}{n_0} \\ 0 & \frac{n_1 + n_2}{n_1 n_2} & -\frac{1}{n_1} \\ -\frac{1}{n_0} & -\frac{1}{n_1} & \frac{n_0 + n_1}{n_0 n_1} \end{pmatrix},$$

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\phi(1-\phi)} \begin{pmatrix} \frac{n_1 + n_2}{n_1 n_2} & -\frac{1}{n_1} \\ -\frac{1}{n_1} & \frac{n_0 + n_1}{n_0 n_1} \end{pmatrix},$$

and

$$\boldsymbol{\Sigma} = \frac{\phi(1-\phi)}{n} \begin{pmatrix} (n_0 + n_1)n_2 & n_0 n_2 \\ n_0 n_2 & n_0(n_1 + n_2) \end{pmatrix},$$

where $\Sigma$ is a consistent estimate of covariance matrix of $(U_1,U_2)$. Note that $Z^2(0) = U_1^2/\hat{\mathrm{V}}\mathrm{ar}(U_1)$ and $Z^2(2) = U_2^2/\hat{\mathrm{V}}\mathrm{ar}(U_2)$, where $\hat{\mathrm{V}}\mathrm{ar}(U_1) = \phi(1-\phi)(n_0 + n_1)n_2/n$ and $\hat{\mathrm{V}}\mathrm{ar}(U_2) = \phi(1-\phi)n_0(n_1 + n_2)/n$. Hence, a consistent estimate of the asymptotic null correlation $\rho = \mathrm{Corr}[Z(0),Z(2)] = \mathrm{Corr}(U_1,U_2)$ is

$$\hat{\rho} = \left\{ \frac{n_0 n_2}{(n_1 + n_2)(n_0 + n_1)} \right\}^{1/2}.$$

Thus, $T_2 = \mathbf{U}^T \boldsymbol{\Sigma}^{-1} \mathbf{U}$ can be written as

$$\begin{aligned} T_2 &= \frac{1}{\phi(1-\phi)n_0 n_1 n_2} [U_1^2 n_0(n_1 + n_2) + U_2^2(n_0 + n_1)n_2 - 2U_1 U_2 n_0 n_2] \\ &= \frac{(n_0 + n_1)(n_1 + n_2)}{n_1 n} [Z^2(0) + Z^2(2) - 2Z(0)Z(2)\hat{\rho}], \end{aligned}$$

where $(n_1 n)/[(n_0 + n_1)(n_1 + n_2)] = 1 - \hat{\rho}^2$.

## References


Agresti A (1990) Categorical data analysis. John Wiley & Sons, New York

Bacanu SA, Devlin B, Roeder K (2000) The power of genomic control. Am J Hum Genet 66:1933–1944

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004

Devlin B, Roeder K, Wasserman L (2001) Genomic control, a new approach to genetic-based association studies. Theor Popul Biol 60: 155–166

Freidlin B, Zheng G, Li Z, Gastwirth JL (2002) Trend tests for case-control studies of genetic markers: power, sample size and robustness. Hum Hered 53:146–152

Gastwirth JL (1966) On robust procedures. J Am Stat Assoc 61:929–948

——— (1985) The use of maximin efficiency robust tests in combining contingency tables and survival analysis. J Am Stat Assoc 80:380–384

Gibson G, Muse SV (2004) A primer of genome science. 2nd ed. Sinauer Associations, Sunderland, MA

Graubard BI, Korn EL (1987) Choice of column scores for testing independence in ordered $2 \times K$ contingency tables. Biometrics 43: 471–476

Khoury MJ, Yang Q (1998) The future of genetic studies of complex human diseases: an epidemiologic perspective. Epidemiology 9:350–354

Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220–228

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67: 170–181

Reich DE, Goldstein DB (2001) Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol 20:4–16

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Sasieni PD (1997) From genotypes to genes: doubling the sample size. Biometrics 53:1253–1261

Satten GA, Flanders WD, Yang Q (2001) Account for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. Am J Hum Genet 68:466–477

Sham P (1998) Statistics in human genetics. Arnold Publishers, London

Weir BS (1996) Genetic data analysis II: methods for discrete population genetic data. Sinauer Associations, Sunderland, MA

Zheng G, Freidlin B, Gastwirth JL (2006) Comparison of robust tests for genetic association using case-control studies. Institute of Mathematical Statistics, Lecture Notes and Monograph Series (The 2nd special issue in honor of E. L. Lehmann) (in press)

Zheng G, Freidlin B, Li Z, Gastwirth JL (2003) Choice of scores in trend tests for case-control studies of candidate-gene associations. Biom J 45:335–348

——— (2005) Genomic control for association studies under various genetic models. Biometrics 61:186–192

Zhu X, Zhang SL, Zhao H, Cooper RS (2002) Association mapping, using a mixture model for complex traits. Genet Epidemiol 23:181–196