BMC Genetics

METHODOLOGY ARTICLE Open Access

Robust tests for matched case-control genetic association studies

Yong Zang, Wing Kam Fung*

Abstract

Background: The Cochran-Armitage trend test (CATT) is powerful in detecting association between a susceptible marker and a disease. This test, however, may suffer from a substantial loss of power when the underlying genetic model is unknown and incorrectly specified. Thus, it is useful to derive tests obtaining the plausible power against all common genetic models. For this purpose, the genetic model selection (GMS) and genetic model exclusion (GME) methods were proposed recently. Simulation results showed that GMS and GME can obtain the plausible power against three common genetic models while the overall type I error is well controlled.

Results: Although GMS and GME are powerful statistically, they could be seriously affected by known confounding factors such as gender, age and race. Therefore, in this paper, via comparing the difference of Hardy-Weinberg disequilibrium coefficients between the cases and the controls within each sub-population, we propose the stratified genetic model selection (SGMS) and exclusion (SGME) methods which could eliminate the effect of confounding factors by adopting a matching framework. Our goal in this paper is to investigate the robustness of the proposed statistics and compare them with other commonly used efficiency robust tests such as MAX3 and χ^2 with 2 degrees of freedom (df) test in matched case-control association designs through simulation studies.

Conclusion: Simulation results showed that if the mean genetic effect of the heterozygous genotype is between those of the two homozygous genotypes, then the proposed tests and MAX3 are preferred. Otherwise, χ^2 with 2 df test may be used. To illustrate the robust procedures, the proposed tests are applied to a real matched pair case-control etiologic study of sarcoidosis.

Background

The population-based case-control association study is a powerful approach in detecting the association between a candidate marker and a disease. Compared with the family-based association study which recruits samples from family members, the case-control study is more cost effective because cases and controls are unrelated hence easy to recruit from population. To test the genetic association using the case-control design, the genotypic data for a bi-allelic marker are usually described by a 2 × 3 table where rows represent the disease status and columns represent the genotypic counts. Hence, to test for genetic association is equivalent to test for association between the rows and the columns. Generally, the Pearson's χ^2 with 2 df test can be used to detect such an association. Besides, if a linear trend among the rows can be assumed, a more powerful test which utilizes the score test for a logistic regression can be obtained. This score test is known as the Cochran-Armitage trend test (CATT) [1-3].

To apply the CATT, increasing scores are specified a priori for the underlying genetic model. A genetic model refers to the model of inheritance, which defines some relationship of the risks of having the disease given different genotypes. The common genetic models include, but not limit to, recessive (REC), additive (ADD) and dominant (DOM) models. If the underlying genetic model is known, the asymptotically optimal CATT can be used. Otherwise, the CATT is not robust when the scores are misspecified [4]. Unfortunately, the underlying genetic model is usually unknown in practice and an incorrect choice of the genetic model may result in a substantial loss of power for the CATT. Thus, a robust method which does not assume a prior knowledge of the underlying genetic model is often useful.

* Correspondence: wingfung@hku.hk
Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, China

BioMed Central

© 2010 Zang and Fung; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Table 1 Type I error rates of GMS and GME based on 10,000 replicates without confounding (Scenario 1) and in the presence of confounding factors (Scenarios 2-8), with the significance level 0.05 using r_l cases and s_l controls; p_l is the risk allele frequency and k_l is the prevalence, l = 1,2

Scenario	r_1	r_2	s_1	s_2	p_1	p_2	k_1	k_2	GMS	GME
1	250	250	250	250	0.3	0.3	0.05	0.05	0.0510	0.0502
2	250	250	250	250	0.05	0.5	0.01	0.1	0.0199	0.0141
3	250	250	250	250	0.1	0.5	0.01	0.1	0.0190	0.0167
4	250	250	250	250	0.2	0.4	0.03	0.07	0.0391	0.0384
5	300	200	200	300	0.2	0.4	0.03	0.07	0.3923	0.4403
6	325	175	175	325	0.2	0.4	0.03	0.07	0.7337	0.7880
7	350	150	150	350	0.2	0.4	0.03	0.07	0.9077	0.9567
8	375	125	125	375	0.2	0.4	0.03	0.07	0.9625	0.9954

Methods robust for a variety of underlying model of inheritance have recently become an important area of research. The Pearson's χ^2 test with 2 df does not assume any structure of a genetic model so it is a robust test against the genetic model. Moreover, the maximin efficiency robust test (MERT) and the MAX method using the maximum of the CATTs optimal for REC, ADD and DOM respectively were extensively studied [5-10]. Recently, Zheng and Ng [11] proposed the genetic model selection (GMS) method to test for genetic association. Different from other robust tests, the GMS approach is a two-phase analysis which uses the Hardy-Weinberg disequilibrium trend test (HWDTT) [12] to choose the most suitable genetic model in the first phase followed by the CATT optimal for the selected genetic model to detect the association in the second phase. Since the same data were used twice in the analysis, the nominal type I error for the second test needs to be adjusted so that the GMS can obtain a correct size. This GMS has an assumption that the marker allele associated with the disease allele is known. Such assumption can be difficult to justify, for instance, in many complex diseases. Thus, to remove this restriction, Joo et al. [13] proposed to use the CATT optimal for the ADD model to detect the risk allele. After the risk allele is determined, the GMS corresponding to the detected risk allele is then carried out. As a result, using their GMS, people do not need to assume that the risk allele is known. Besides the modified GMS, Joo et al. [13] developed another two-phase test called the genetic model exclusion (GME) which excludes the most unlikely genetic model rather than selecting the most likely one. They also showed that when the genetic relative risks (GRRs) are small, the GME is more efficiency robust than the GMS. Besides the frequentist analysis, a Bayesian hierarchical model which regards the genetic model parameter as a fixed effect has been proposed by Minelli et al. [14]. If expert opinion or external evidence is available, an informative prior distribution of the genetic model parameter could be adopted; otherwise, a vague prior distribution should be used to avoid the undue influence on the posterior distribution.

Although the population based case-control study is powerful and feasible to implement, spurious association may arise due to known confounding factors such as gender, age and race. Intuitively, the GMS and GME do not work in the presence of confounding factors. One of the reasons is that when the samples are divided into several sub-populations via the confounding factors, the Hardy-Weinberg equilibrium (HWE) assumption needed in the first phase of the GMS and GME does not hold any more. Besides, the CATTs used in the second phase of the GMS and GME do not control the size well due to the confounding factors.

Typically, when the confounding factors can be observed, they could be treated as the covariates of interest and incorporated in the logistic regression. However, further calculation to adjust for the covariates may complicate the trend test. Alternatively, the matching strategy is frequently used as a much simpler way to control potential confounding factors in epidemiological studies. Specifically, a single case is matched with a certain number of controls based on the confounding factors constructing for each matched set. Then, a conditional logistic regression analysis is normally used to fit the matched data. Recently, an increasing number of matching studies are conducted by either adopting the matched design [15-18] or developing statistical procedures [19-24] for matched genetic association studies.

Similar to the unmatched case-control association study, when the underlying genetic model is unknown, the robustness of the statistics for the matched case-control design is also worth studying. Zheng and Tian [21] proposed the MAX3 test based on the matching trend test (MTT) derived from a conditional logistic regression. However, to our knowledge, this is the only paper discussing the robust tests in matched case-control design, compared to the large amount of literature in the unmatched design. Thus, in this paper, we start by developing the stratified genetic model selection (SGMS) and exclusion (SGME) methods for matched case-control association, then we study the robustness of the test statistics. The performance of the robust tests and MTTs is compared by simulation for a wide range of scenarios. Finally, the tests are applied to a real matched pair case-control etiologic study of sarcoidosis.

Methods

Genetic model selection and exclusion

When the genetic model is unknown, the T_{HWDTT} test proposed by Song and Elston [12] can be used to detect

the latent genetic model. Zheng and Ng [11] demonstrated that under Hardy-Weinberg equilibrium (HWE) and when the allele investigated is the risk allele (denoted as D), $T_{\mathrm{HWDTT}} > 0$ under the REC model and $T_{\mathrm{HWDTT}} < 0$ under the DOM model. Denote T_0, $T_{0.5}$ and T_1 as the CATTs optimal for REC, ADD and DOM respectively, Zheng and Ng [11] proposed to use T_0 if $T_{\mathrm{HWDTT}} > c$; T_1 if $T_{\mathrm{HWDTT}} < -c$ and $T_{0.5}$ otherwise to test for genetic association, where c is a pre-specified threshold.

Note that for the original GMS mentioned above, the risk allele is assumed to be known. However, if the risk allele cannot be correctly specified, such GMS may have some problems. Specifically, consider a bi-allelic marker with alleles D and d and assume D is the risk allele. So if D is really the risk allele, T_0, $T_{0.5}$ and T_1 are optimal for the REC, ADD and DOM models respectively. On the other hand, if d is the true risk allele, then $-T_1$, $-T_{0.5}$ and $-T_0$ are optimal for the REC, ADD and DOM models respectively. Joo et al. [13] proposed to use $T_{0.5}$ to decide which one is the risk allele followed by the corresponding GMS which depends on the determined risk allele. Joo et al. [13] suggested the modified GMS which can be written as

$$\begin{aligned} T_{\mathrm{GMS}} = {} & T_0 I(T_{0.5} > 0) I(T_{\mathrm{HWDTT}} > c) + T_{0.5} I(T_{0.5} > 0) I(|T_{\mathrm{HWDTT}}| \le c) \\ & + T_1 I(T_{0.5} > 0) I(T_{\mathrm{HWDTT}} < -c) - T_1 I(T_{0.5} \le 0) I(T_{\mathrm{HWDTT}} > c) \\ & - T_{0.5} I(T_{0.5} \le 0) I(|T_{\mathrm{HWDTT}}| \le c) - T_0 I(T_{0.5} \le 0) I(T_{\mathrm{HWDTT}} < -c), \end{aligned} \quad (1)$$

where I(.) is an indicator function.

When the GRRs are small, Joo et al. [13] found that the probability of selecting the true genetic model by using T_{HWDTT} becomes small. On the other hand, the probability of correctly excluding the most unlikely genetic model remains high against the GRRs. Furthermore, when the most unlikely genetic model is excluded, the simple MERT [5] can be carried out to build a robust test against the remaining models. These facts inspired Joo et al. [13] to develop the genetic model exclusion (GME) approach. Specifically, first denote $T_0^* = (T_0 + T_{0.5}) / (\sqrt{2(1 + \hat{\rho}_{0,0.5})}), T_{0.5}^* = T_{0.5}$ and $T_1^* = (T_1 + T_{0.5})/(\sqrt{2(1 + \hat{\rho}_{1,0.5})})$ where $\hat{\rho}_{x_1,x_2}$ is an estimate of the correlation between T_{x_1} and T_{x_2} under the null hypothesis of no association, then one can obtain the GME statistic from the GMS test by replacing T_0, $T_{0.5}$ and T_1 in (1) by T_0^*, $T_{0.5}^*$ and T_1^* respectively. Since the GMS and GME are two stage tests and the same data set is used twice, the critical values of the tests in the second stage need to be adjusted to control the overall type I error rates; see Zheng and Ng [11] and Joo et al. [13].

Although GMS and GME are efficiency robust tests, they could be seriously affected by confounding factors. In the presence of sub-populations, GMS and GME may not keep the correct size. Therefore, to overcome this limitation, we propose the stratified genetic model selection (SGMS) and exclusion (SGME) approaches in the following.

Notation

Consider a bi-allelic marker with alleles d and D and assume D is the risk allele. Denote the three genotypes of this marker as $G_0 = dd$, $G_1 = Dd$ and $G_2 = DD$. Suppose that the confounding factors define L strata, denoted by C_l, $l = 1,..., L$. In the lth stratum, r_l cases are drawn from the population and m controls are matched to each case. Thus, the total number of controls in the lth stratum is $s_l = mr_l$ for $l = 1,..., L$. The genotype counts for (G_0, G_1, G_2) in cases and controls in the lth stratum are denoted by (r_{0l}, r_{1l}, r_{2l}) and (s_{0l}, s_{1l}, s_{2l}), respectively. Hence, $r_l = \sum_{i=0}^{2} r_{il}$ and $s_l = \sum_{i=0}^{2} s_{il}$. The total number of cases is $r = \sum_l r_l$ and the total number of controls is $s = mr$. The total sample size is then $n = (m + 1)r$.

In the lth stratum ($l = 1,..., L$), denote the penetrance by $f_{il} = Pr(\mathrm{case}|G_i, C_l)$ for $i = 0, 1, 2$, the disease prevalence by $k_l = Pr(\mathrm{case}|C_l) = \sum_i f_{il} Pr(G_i|C_l)$, and the genotype frequencies in cases and controls by $p_{il} = Pr(G_i|\mathrm{case}, C_l) = f_{il} Pr(G_i|C_l)/k_{il}$ and $q_{il} = Pr(G_i|\mathrm{control}, C_l) = (1 - f_{il}) Pr(G_i|C_l) = (1 - k_{il})$, respectively. Define GRRs in the lth stratum as $\lambda_{1l} = f_{1l}/f_{0l}$ and $\lambda_{2l} = f_{2l}/f_{0l}$ ($f_{0l} > 0$). A genetic model is REC, ADD and DOM if $\lambda_{1l} = 1$, $\lambda_{1l} = (\lambda_{2l} + 1)/2$ and $\lambda_{1l} = \lambda_{2l}$, respectively. We assume that HWE holds in each stratum. Thus, $Pr(G_0|C_l) = q_l^2$, $Pr(G_1|C_l) = 2p_l q_l$ and $Pr(G_2|C_l) = p_l^2$ where p_l is the allele frequency of A in the lth stratum and $q_l = 1 - p_l$.

Stratified genetic model selection and exclusion

Let X_{1lj} and X_{2ljk} denote the genotypic scores for the jth case and the kth control matched with the jth case in the lth stratum, $j = 1,..., r_l$, $k = 1,..., m$ and $l = 1,..., L$. Each score takes one of the three possible values: 0, x or 1 for the genotypes G_0, G_1 or G_2 respectively, where x is 0, 0.5 or 1 for the REC, ADD or DOM model. Following Day and Byar [25] and Zheng and Tian [21], the likelihood function conditional on the outcomes of cases and matched controls for the candidate marker can be written as

$$\begin{aligned} L(\beta \mid G, C_1, \ldots, C_L) &= \prod_{l=1}^{L} \prod_{j=1}^{r_l} \frac{\exp(\alpha_l + \beta X_{1lj})}{\exp(\alpha_l + \beta X_{1lj}) + \sum_{k=1}^{m} \exp(\alpha_l + \beta X_{2ljk})} \\ &= \prod_{l=1}^{L} \prod_{j=1}^{r_l} \frac{\exp(\beta X_{1lj})}{\exp(\beta X_{1lj}) + \sum_{k=1}^{m} \exp(\beta X_{2ljk})}. \end{aligned} \quad (2)$$

The null hypothesis of no association $H_0 : \beta = 0$ can be tested by the score statistic given by $Z_{\mathrm{MTT}}(x) = U(x) / \{\widehat{\mathrm{Var}}_{H_0}(U(x))\}^{1/2} = (\partial \log L / \partial \beta)_{H_0} / \{-(\partial^2 \log L / \partial \beta^2)_{H_0}\}^{1/2}$. Using the matched case-control data, the closed form of the matching trend test (MTT) can be written as [21]

$$Z_{\mathrm{MTT}}(x) = \frac{U(x)}{[V(x)]^{1/2}} = \frac{\sum_{l=1}^{L}\sum_{j=1}^{r_l}(mX_{1lj} - \sum_{k=1}^{m}X_{2ljk})}{\left[\sum_{l=1}^{L}\sum_{j=1}^{r_l}\left\{(m+1)(X_{1lj}^2 + \sum_{k=1}^{m}X_{2ljk}^2) - (X_{1lj} + \sum_{k=1}^{m}X_{2ljk})^2\right\}\right]^{1/2}}. \quad (3)$$

Obviously, $\sum_{j=1}^{r_l} X_{1lj} = r_{2l} + xr_{1l}$ and $\sum_{j=1}^{r_l}\sum_{k=1}^{m} X_{2ljk} = s_{2l} + xs_{1l}$. $Z_{\mathrm{MTT}}(x)$, follows $N(0,1)$ under the null hypothesis of no association.

Suppose a family of scientifically plausible models is defined. Similar to the CATTs, corresponding to each model, an asymptotically optimal normally distributed MTT can be obtained. For example, $Z_{\mathrm{MTT}}(0)$, $Z_{\mathrm{MTT}}(0.5)$ and $Z_{\mathrm{MTT}}(1)$ are optimal for the REC, ADD and DOM models respectively. When the genetic model is uncertain, a pre-specified test from this family is not fully efficient, hence, MTTs are not suggested to be directly used when the underlying genetic model is unknown. This underlying genetic model, however, can be ascertained using the Hardy-Weinberg Disequilibrium (HWD) coefficient which is de-noted as $\Delta = Pr(DD) - [Pr(DD)+Pr(Dd)/2]^2$. In the unmatched study, denote the HWD coefficients in the case group and the control group as $\Delta_p = Pr(DD|case) - [Pr(DD|case)+Pr(Dd|case)/2]^2$ and $\Delta_q = Pr(DD|control) - [Pr(DD|control) + Pr(Dd|control)/2]^2$, Zheng and Ng [11] obtained that $\Delta_p - \Delta_q > 0$ under REC and $\Delta_p - \Delta_q < 0$ under DOM. Using the matched design described above, we denote Δ_{pl} and Δ_{ql} as the HWD coefficients in the case group and the control group of the lth sub-population respectively, $l = 1,..., L$. Similar to the unmatched counterpart, we have $\Delta_{pl} - \Delta_{ql} > 0$ for each l, $l = 1,..., L$ thus $\sum_{l=1}^{L}(\Delta_{pl} - \Delta_{ql}) > 0$ under REC and $\Delta_{pl} - \Delta_{ql} < 0$ for each l, $l = 1,..., L$ thus $\sum_{l=1}^{L}(\Delta_{pl} - \Delta_{ql}) < 0$ under DOM.

Denote $\hat{\Delta}_l = \hat{\Delta}_{pl} - \hat{\Delta}_{ql} = [\hat{p}_{2l} - (\hat{p}_{2l} + \frac{1}{2}\hat{p}_{1l})^2] - [\hat{q}_{2l} - (\hat{q}_{2l} + \frac{1}{2}\hat{q}_{1l})^2]$ where $\hat{p}_{il} = r_{il} / r_l$ and $\hat{q}_{il} = s_{il} / (mr_l)$ for $i = 0, 1, 2$ and $l = 1,..., L$. Under the null hypothesis of no association and assume HWE in each stratum, using simple algebra we can obtain $\hat{\Lambda}_l = \widehat{Var}_{H_0}(\hat{\Delta}_l) = (m+1)(1-\hat{p}_l)^2\hat{p}_l^2 / (r_l m)$ where $\hat{p}_l = [2(r_{2l} + s_{2l}) + (r_{1l} + s_{1l})] / [2(m+1)r_l]$. Thus, using the same motivation as the Cochran-Mantel-Haenszel (CMH) statistic [1,25,26] we can construct the stratified model reduction test (SMRT):

$$Z_{\mathrm{SMRT}} = \frac{\hat{\Delta}}{\sqrt{\hat{\Lambda}}} = \frac{\sum_{l=1}^{L}\hat{\Delta}_l}{\sqrt{\sum_{l=1}^{L}\hat{\Lambda}_l}} = \frac{\sum_{l=1}^{L}\{[\hat{p}_{2l} - (\hat{p}_{2l} + \frac{1}{2}\hat{p}_{1l})^2] - [\hat{q}_{2l} - (\hat{q}_{2l} + \frac{1}{2}\hat{q}_{1l})^2]\}}{\sqrt{\sum_{l=1}^{L}(m+1)(1-\hat{p}_l)^2\hat{p}_l{}^2/(r_l m)}}. \quad (4)$$

Notice that the denominator of Z_{SMRT} is estimated under the null hypothesis thus Z_{SMRT} is a score test [27]. We may also use the Wald test or likelihood ratio test. However, if we adopt the Wald test, the statistic becomes much more complex; if we adopt the likelihood ratio test, the statistic cannot be expressed explicitly thus it is hard to derive the correlations between the two stage tests and calculate the p-value of the overall test. For these reasons, the score test is adopted. Under the null hypothesis, Z_{SMRT} asymptotically follows a standard normal distribution $N(0, 1)$. Z_{SMRT} tends to be large if the true genetic model is REC and tends to be small if the true genetic model is DOM. Hence, with a pre-specified threshold $c > 0$ (set to be $\Phi^{-1}(0.95)$), we can classify the underlying genetic model as REC if $Z_{\mathrm{SMRT}} > c$, DOM if $Z_{\mathrm{SMRT}} < -c$ and ADD otherwise. So when the underlying genetic model is decided, $Z_{\mathrm{MTT}}(x)$ optimal for the corresponding genetic model can be used to test for association. Notice that in the above discussion, we assume that D is the risk allele. If D is the risk allele as we assume, $Z_{\mathrm{MTT}}(0)$ and $Z_{\mathrm{MTT}}(1)$ are optimal for the REC and DOM models respectively. On the other hand, if d is the risk allele, then $Z_{\mathrm{MTT}}(0)$ and $Z_{\mathrm{MTT}}(1)$ are optimal for the DOM and REC models respectively. Besides, the expected values of $Z_{\mathrm{MTT}}(0)$ and $Z_{\mathrm{MTT}}(1)$ are negative in this case. Similar to Joo et al. [13], we use $Z_{\mathrm{MTT}}(0.5)$ to determine the risk allele. That is, if $Z_{\mathrm{MTT}}(0.5) > 0$, $Z_{\mathrm{MTT}}(0)$, $Z_{\mathrm{MTT}}(0.5)$, $Z_{\mathrm{MTT}}(1)$ are optimal for the REC, ADD and DOM models; if $Z_{\mathrm{MTT}}(0.5) \le 0$, $-Z_{\mathrm{MTT}}(1)$, $-Z_{\mathrm{MTT}}(0.5)$, $-Z_{\mathrm{MTT}}(0)$ are optimal for the REC, ADD and DOM models. Hence, the stratified genetic model selection (SGMS) test is proposed as

$$Z_{\mathrm{SGMS}} = \begin{cases} Z_{\mathrm{MTT}}(0)I(Z_{\mathrm{MTT}}(0.5) > 0) - Z_{\mathrm{MTT}}(1)I(Z_{\mathrm{MTT}}(0.5) \le 0) & \text{if } Z_{\mathrm{SMRT}} > c \\ \mathrm{sign}(Z_{\mathrm{MTT}}(0.5))Z_{\mathrm{MTT}}(0.5) & \text{if } |Z_{\mathrm{SMRT}}| \le c \\ Z_{\mathrm{MTT}}(1)I(Z_{\mathrm{MTT}}(0.5) > 0) - Z_{\mathrm{MTT}}(0)I(Z_{\mathrm{MTT}}(0.5) \le 0) & \text{if } Z_{\mathrm{SMRT}} < -c. \end{cases} \quad (5)$$

Under the null hypothesis of no association, we show that $(Z_{\mathrm{MTT}}(0.5), Z_{\mathrm{SMRT}}, Z_{\mathrm{MTT}}(x))$ asymptotically follows a multivariate normal distribution $N(0, \Sigma_x)$ where

$$\Sigma_x = \begin{pmatrix} 1 & 0 & \rho_{x,0.5} \\ 0 & 1 & \rho_x \\ \rho_{x,0.5} & \rho_x & 1 \end{pmatrix},$$

$x = 0, 1$. In addition, $Z_{\mathrm{MTT}}(0.5)$ and Z_{SMRT} are asymptotically independent. Detailed proof and the forms of ρ_x and $\rho_{x,0.5}$ as well as their consistent estimates are derived in the **Appendix**. Define $\boldsymbol{\phi}_x(z_1, z_2, z_3)$ as the density function of $N(0, \boldsymbol{\Sigma}_x)$ and $\phi(z)$ as the density function of the standard normal distribution. Let $t > 0$ be the observed value of Z_{SGMS} and the corresponding p-value is obtained as

$$\begin{aligned} PV_s &= Pr(Z_{\mathrm{SGMS}} > t) \\ &= 2\left\{\sum_{x=0,1}\int_t^{+\infty}\int_c^{+\infty}\int_0^{+\infty}\phi_x(z_1, z_2, z_3)dz_1dz_2dz_3\right\} \\ &\quad +2\left\{\int_{-c}^{c}\phi(z)dz\int_t^{+\infty}\phi(z)dz\right\}. \end{aligned} \quad (6)$$

With a pre-specified significance level ζ, we declare a significant association if $PV_s < \zeta$.

Table 2 Type I error rates of $Z_{MTT(0)}$, $Z_{MTT(0.5)}$, $Z_{MTT(1)}$, Z_{SGMS}, Z_{SGME}, Z_{MAX3} and Z_{χ^2} with 2 df based on 10,000 replicates in the presence of confounding factors with the significance level α using *R* cases and *S* controls

Scenario	α	Z_{MTT} (0)	Z_{MTT} (0.5)	Z_{MTT} (1)	Z_{SGMS}	Z_{SGME}	Z_{MAX3}	Z_{χ^2}
A	0.05	0.0527	0.0498	0.0518	0.0501	0.0490	0.0527	0.0531
B		0.0487	0.0503	0.0493	0.0494	0.0502	0.0515	0.0481
C		0.0510	0.0512	0.0509	0.0506	0.0510	0.0513	0.0490
D		0.0526	0.0524	0.0537	0.0529	0.0528	0.0516	0.0484
E		0.0519	0.0512	0.0534	0.0536	0.0526	0.0501	0.0507
F		0.0485	0.0486	0.0479	0.0488	0.0467	0.0501	0.0481
G		0.0493	0.0497	0.0490	0.0492	0.0498	0.0457	0.0491
H		0.0522	0.0493	0.0480	0.0522	0.0521	0.0525	0.0522
A	0.01	0.0092	0.0081	0.0100	0.0083	0.0081	0.0075	0.0084
B		0.0096	0.0091	0.0106	0.0106	0.0103	0.0096	0.0101
C		0.0093	0.0101	0.0109	0.0108	0.0104	0.0101	0.0109
D		0.0121	0.0101	0.0098	0.0093	0.0093	0.0095	0.0105
E		0.0098	0.0094	0.0101	0.0093	0.0092	0.0083	0.0114
F		0.0109	0.0095	0.0106	0.0109	0.0102	0.0102	0.0109
G		0.0100	0.0094	0.0105	0.0114	0.0103	0.0111	0.0093
H		0.0087	0.0111	0.0104	0.0099	0.0103	0.0117	0.0107

A : *R* = (150, 150, 200), *S* = (300, 300, 400), *P* = (0.1, 0.3, 0.5), *K* = (0.01, 0.05, 0.02).
B : *R* = (100, 300, 100), *S* = (200, 600, 200), *P* = (0.1, 0.3, 0.5), *K* = (0.01, 0.05, 0.02).
C : *R* = (300, 300, 400), *S* = (300, 300, 400), *P* = (0.1, 0.3, 0.5), *K* = (0.01, 0.05, 0.02).
D : *R* = (200, 600, 200), *S* = (200, 600, 200), *P* = (0.1, 0.3, 0.5), *K* = (0.01, 0.05, 0.02).
E : *R* = (250, 250), *S* = (500, 500), *P* = (0.2, 0.4), *K* = (0.01, 0.02).
F : *R* = (150, 350), *S* = (300, 700), *P* = (0.2, 0.4), *K* = (0.01, 0.02).
G : *R* = (500, 500), *S* = (500, 500), *P* = (0.2, 0.4), *K* = (0.01, 0.02).
H : *R* = (300, 700), *S* = (300, 700), *P* = (0.2, 0.4), *K* = (0.01, 0.02).

Although Z_{SMRT} can be used to determine the underlying genetic model, the probability of selecting the correct genetic model is low when the GRRs are small or moderate. On the other hand, the probability of correctly excluding the most unlikely genetic model remains high when GRRs are very small. That is, when $Z_{SMRT} > c$, the underlying genetic model is likely to be either REC or ADD rather than just REC and excluding the DOM model is more reasonable than just selecting the REC model. Similarly, when $Z_{SMRT} < -c$, excluding REC is more reasonable than just selecting the DOM model. Therefore, when the GRRs are low, the strategy of excluding the most unlikely genetic model is more preferred to that of selecting the most suitable genetic model.

Similar to Joo et al. [13], we define the MERT-type statistic named the matching averaged test (MAT) as

$$Z_{MAT}(0) = (Z_{MTT}(0) + Z_{MTT}(0.5))/\sqrt{2(1+\hat{\rho}_{0,05})} \quad \text{and}$$

$$Z_{MAT}(1) = (Z_{MTT}(1) + Z_{MTT}(0.5))/\sqrt{2(1+\hat{\rho}_{1,0.5})}\,.$$ The definition of MATs indicate that $Z_{MAT}(0)$ is optimal for either REC or ADD and $Z_{MAT}(1)$ is optimal for either DOM or ADD. Besides, $Z_{MTT}(0.5)$ is still optimal for just ADD. Utilizing the stratified genetic model exclusion (SGME) strategy, we use $Z_{MAT}(0)$ to test for association if $Z_{SMRT} > c$ thus DOM is excluded; use $Z_{MAT}(1)$ if $Z_{SMRT} < -c$ thus REC is excluded and $Z_{MTT}(0.5)$ otherwise. In addition, similar to SGMS, $Z_{MTT}(0.5)$ is used at the beginning to determine the risk allele. Hence, the statistic for the SGME approach can be written as

$$Z_{SGME} = \begin{cases} Z_{MAT}(0)I(Z_{MTT}(0.5) > 0) - Z_{MAT}(1)I(Z_{MTT}(0.5) \le 0) & \text{if } Z_{SMRT} > c \\ \text{sign}(Z_{MTT}(0.5))Z_{MTT}(0.5) & \text{if } |Z_{SMRT}| \le c \\ Z_{MAT}(1)I(Z_{MTT}(0.5) > 0) - Z_{MAT}(0)I(Z_{MTT}(0.5) \le 0) & \text{if } Z_{SMRT} < -c \end{cases} \quad (7)$$

Under the null hypothesis of no association, we obtain that $(Z_{MTT}(0.5), Z_{SMRT}, Z_{MAT}(x))$ asymptotically follows a multivariate normal distribution $N(0, \Sigma_x^*)$ where

$$\Sigma_x^* = \begin{pmatrix} 1 & 0 & \sqrt{\frac{(1+\rho_{x,0.5})}{2}} \\ 0 & 1 & \frac{\rho_x}{\sqrt{2(1+\rho_{x,0.5})}} \\ \sqrt{\frac{(1+\rho_{x,0.5})}{2}} & \frac{\rho_x}{\sqrt{2(1+\rho_{x,0.5})}} & 1 \end{pmatrix},$$

$x = 0, 1$. Define $\phi_x^*(z_1, z_2, z_3)$ as the density function of $N(0, \Sigma_x^*)$, similar to the test of Z_{SGMS}, the p-value of Z_{SGME} can be derived as

$$\begin{aligned} PV_e &= Pr(Z_{SGME} > t) \\ &= 2\left\{ \sum_{x=0,1} \int_t^{+\infty} \int_c^{+\infty} \int_0^{+\infty} \phi_x^*(z_1, z_2, z_3) dz_1 dz_2 dz_3 \right\} \\ &\quad +2\left\{ \int_{-c}^{c} \phi(z)dz \int_t^{+\infty} \phi(z)dz \right\}. \end{aligned} \quad (8)$$

We declare a significant association if $PV_e < \zeta$ where ζ is the pre-specified significance level.

Other robust procedures

In equation (2), we use one indicator to code three genotypes. One the other hand, if we define two dummy variables $((X_{1lj1}, X_{1lj2})$ for the cases and (X_{2ljk1}, X_{2ljk2}) for the controls) taking values (0,0), (0,1) and (1,1) to code three genotypes G_0, G_1 and G_2, the conditional likelihood function becomes [10]

$$\begin{aligned} &L(\beta_1, \beta 2 \,|\, G, C_1, ..., C_L) \\ &= \prod_{l=1}^{L} \prod_{j=1}^{r_l} \frac{\exp(\beta_1 X_{1lj1} + \beta_2 X_{1lj2})}{\exp(\beta_1 X_{1lj1} + \beta_2 X_{1lj2}) + \sum_{k=1}^{m} \exp(\beta_1 X_{2ljk1} + \beta_2 X_{2ljk2})}. \end{aligned} \quad (9)$$

The score test derived from equation (9), denoted by Z_{χ^2} with 2 df, has an asymptotic χ^2 distribution with 2

Table 3 Type I error rates of $Z_{MTT(0)}$, $Z_{MTT(0.5)}$, $Z_{MTT(1)}$, Z_{SGMS}, Z_{SGME}, Z_{MAX3} and Z_{χ^2} with 2 df for small sample size

Scenario	α	Z_{MTT} (0)	Z_{MTT} (0.5)	Z_{MTT} (1)	Z_{SGMS}	Z_{SGME}	Z_{MAX3}	Z_{χ^2}
*A**	0.05	0.0474	0.0501	0.0487	0.0493	0.0495	0.0452	0.0502
*B**		0.0531	0.0479	0.0497	0.0480	0.0485	0.0462	0.0503
*C**		0.0569	0.0526	0.0489	0.0507	0.0526	0.0516	0.0488
*D**		0.0470	0.0480	0.0516	0.0482	0.0484	0.0488	0.0503
*E**		0.0492	0.0498	0.0489	0.0518	0.0511	0.0535	0.0498
*F**		0.0519	0.0503	0.0514	0.0535	0.0537	0.0486	0.0489
*G**		0.0484	0.0505	0.0526	0.0483	0.0466	0.0551	0.0502
*H**		0.0504	0.0453	0.0451	0.0451	0.0456	0.0484	0.0504
*A**	0.01	0.0076	0.0083	0.0092	0.0102	0.0089	0.0108	0.0126
*B**		0.0075	0.0091	0.0097	0.0078	0.0088	0.0086	0.0081
*C**		0.0078	0.0089	0.0096	0.0082	0.0084	0.0126	0.0092
*D**		0.0080	0.0095	0.0116	0.0093	0.0092	0.0111	0.0099
*E**		0.0072	0.0093	0.0098	0.0099	0.0092	0.0091	0.0120
*F**		0.0087	0.0081	0.0089	0.0085	0.0081	0.0091	0.0116
*G**		0.0077	0.0120	0.0120	0.0113	0.0125	0.0081	0.0102
*H**		0.0079	0.0087	0.0088	0.0073	0.0081	0.0098	0.0095

The results are simulated based on 10,000 replicates in the presence of confounding factors with the significance level α using R cases and S controls.
*A**: R = (15, 15, 20), S = (30, 30, 40), P = (0.1, 0.3, 0.5), K = (0.01, 0.05, 0.02).
*B**: R = (10, 30, 10), S = (20, 60, 20), P = (0.1, 0.3, 0.5), K = (0.01, 0.05, 0.02).
*C**: R = (30, 30, 40), S = (30, 30, 40), P = (0.1, 0.3, 0.5), K = (0.01, 0.05, 0.02).
*D**: R = (20, 60, 20), S = (20, 60, 20), P = (0.1, 0.3, 0.5), K = (0.01, 0.05, 0.02).
*E**: R = (25, 25), S = (50, 50), P = (0.2, 0.4), K = (0.01, 0.02).
*F**: R = (15, 35), S = (30, 70), P = (0.2, 0.4), K = (0.01, 0.02).
*G**: R = (50, 50), S = (50, 50), P = (0.2, 0.4), K = (0.01, 0.02).
*H**: R = (30, 70), S = (30, 70), P = (0.2, 0.4), K = (0.01, 0.02).

df under $H_0 : \beta_1 = \beta_2 = 0$. Note that Z_{χ^2} with 2 df does not rely on any information of the underlying genetic model so it is a robust test against model of inheritance.

Another robust test is the MAX3 which was also proposed as an efficiency robust test for unmatched genetic association studies [7,28]. Analogy to the unmatched counterpart, Zheng and Tian [21] proposed the MAX3 statistic for matched case-control association study which is defined as

$$Z_{\mathrm{MAX3}} = max(|Z_{\mathrm{MTT}}(0)|, |Z_{\mathrm{MTT}}(0.5)|, |Z_{\mathrm{MTT}}(1)|).$$

Compared with the optimal MTTs and χ^2 test with 2 df, MAX3 has the largest minimum power across the three genetic models [21,29]. As mentioned in Zheng and Tian [21] and Joo et al. [10], the null distribution of Z_{MAX3} can be approximated by Monte-Carlo simulation. In addition, the p-value of Z_{MAX3} can also be obtained according to the asymptotic formula given by Zang et al. [29].

Results

Simulation

To check whether GMS and GME can keep the correct size in the presence of confounding factors, we carried out simulation studies to examine the performance of GMS and GME in the presence of sub-populations. The nominal level was set at 0.05. We assumed that due to confounding factors, each of the case and control populations was divided into two sub-populations with equal probability. The simulation results are summarized in Table 1. We find that when there are no confounding factors (Scenario *1*), GMS and GME can control the size well. On the other hand, in the presence of confounding factors and adopt the matched design to each of the sub-populations, GMS and GME are found to be conservative (Scenarios *2*, *3* and *4*). Furthermore, without the matched design, the type I error rates of GMS and GME are seriously inflated (Scenarios *5* to *8*). The simulation results show that in the presence of sub-populations, GMS and GME cannot keep the correct size whether or not the matched design is utilized.

To check if the ability of Z_{SMRT} to select the correct genetic model is low when GRRs are small, we conducted a simulation to compare the selection procedure with the exclusion procedure. Considered 300 cases with 600 matched controls, the samples were divided into 3 sub-populations with proportions being 0.3, 0.3 and 0.4 respectively. Set the MAFs and the penetrance in the three strata as (p_1, p_2, p_3) = (0.1, 0.3, 0.5) and (f_{01}, f_{02}, f_{03}) = (0.01, 0.05, 0.02). The threshold c was fixed as $\Phi^{-1}(0.95)$ and let GRR2 = λ_{2l} increase from 1.1 to 2.0 with increments of 0.1, l = 1,..., L.

The results are summarized in Figure 1, with circles representing the probabilities of selecting the correct genetic models and triangles representing the probabilities of correctly excluding the most unlikely genetic models. From Figure 1 we can find that under REC and DOM the triangles are always higher than 90%, whereas the circles can be less than 20% when the GRR2 is small. However, when ADD is the true genetic model, the circles coincide with the triangles. This is because under ADD, both REC and DOM are the most unlikely models thus the selection procedure is just the same as the exclusion procedure.

Next, we performed simulations with no disease association and under various genetic models to evaluate the performance of the proposed robust methods. Moreover, we also considered the MTTs optimal for the REC, ADD and DOM models, i.e. $Z_{\mathrm{MTT}}(0)$, $Z_{\mathrm{MTT}}(0.5)$ and $Z_{\mathrm{MTT}}(1)$ respectively. Let R, S, F_i, K and P denoted the vectors of r_l, s_l, f_{il}, k_l and p_l respectively across sub-populations, l = 1,..., L, i = 0, 1, 2. Each sub-population was in HWE. We first examined the type I error rates

Table 4 Empirical powers of $Z_{MTT(0)}$, $Z_{MTT(0.5)}$, $Z_{MTT(1)}$, Z_{SGMS}, Z_{SGME}, Z_{MAX3} and Z_{χ^2} with 2 df based on 10,000 replicates

Scenario	Model	$Z_{MTT}(0)$	$Z_{MTT}(0.5)$	$Z_{MTT}(1)$	Z_{SGMS}	Z_{SGME}	Z_{MAX3}	Z_{χ^2}	ρ^*
A	REC	0.8059	0.5590	0.1369	0.6981	0.6760	**0.7340**	0.7154	0.3267
	ADD	0.4890	0.7998	0.7126	0.7594	**0.7896**	0.7629	0.7188	0.7623
	DOM	0.1237	0.6818	0.8040	0.7142	0.7155	**0.7300**	0.7158	0.2639
B	REC	0.8073	0.5367	0.1356	0.6725	0.6497	**0.7229**	0.7147	0.3423
	ADD	0.4637	0.7977	0.7258	0.7646	**0.7908**	0.7502	0.7140	0.7445
	DOM	0.1287	0.7011	0.8054	0.7168	0.7259	**0.7383**	0.7199	0.2756
C	REC	0.8057	0.5503	0.1330	0.6970	0.6691	**0.7244**	0.7153	0.3038
	ADD	0.4896	0.8052	0.7139	0.7654	**0.7952**	0.7648	0.7094	0.7295
	DOM	0.1193	0.6877	0.8062	0.7153	0.7210	**0.7445**	0.7112	0.2710
D	REC	0.7978	0.5235	0.1400	0.6655	0.6433	**0.7177**	0.7144	0.3124
	ADD	0.4639	0.8045	0.7308	0.7654	**0.7934**	0.7499	0.7090	0.7037
	DOM	0.1204	0.7024	0.8071	0.7144	0.7225	**0.7250**	0.7033	0.2792
E	REC	0.7974	0.5276	0.1453	0.7166	0.6782	**0.7288**	0.7218	0.3492
	ADD	0.4667	0.8014	0.7294	0.7554	**0.7884**	0.7517	0.7131	0.7396
	DOM	0.1261	0.6989	0.8027	0.7135	**0.7234**	0.7161	0.7077	0.2795
F	REC	0.8056	0.5547	0.1535	0.6991	0.6742	**0.7317**	0.7173	0.3561
	ADD	0.5014	0.8068	0.7195	0.7650	**0.7955**	0.7524	0.7112	0.7661
	DOM	0.1389	0.6933	0.8003	0.7167	0.7231	**0.7291**	0.7141	0.2887
G	REC	0.8045	0.5241	0.1499	0.7233	0.6786	**0.7316**	0.7082	0.3172
	ADD	0.4562	0.8020	0.7313	0.7522	**0.7883**	0.7560	0.7068	0.6970
	DOM	0.1247	0.7091	0.8004	0.7167	0.7297	**0.7468**	0.7099	0.2825
H	REC	0.7967	0.5481	0.1467	0.6950	0.6651	**0.7203**	0.7136	0.3279
	ADD	0.4885	0.8017	0.7154	0.7610	**0.7911**	0.7529	0.7009	0.7272
	DOM	0.1254	0.6944	0.8055	0.7183	0.7248	**0.7366**	0.7091	0.2945

The settings are the same as those in Table 2 except that the GRRs are determined so that the optimal MTT has the maximum power of about 80%. The significance level is 0.05. ρ^* is the minimum correlation of the optimal tests.

of the mentioned tests under the null hypothesis of no association with nominal levels taken as 0.05 and 0.01 respectively. The results are summarized in Table 2.

We considered eight separate scenarios (A to H) with different numbers of cases, controls, risk allele frequencies and disease prevalences. For example, in scenario A, 150, 150 and 200 cases from 3 different sub-populations comprised the whole case group and each case was matched with 2 controls within the same sub-population. The risk allele frequencies of the 3 sub-populations were 0.1, 0.3 and 0.5 respectively and the disease prevalences equalled to 0.01, 0.05 and 0.02. Table 2 shows that the type I error rates of all the mentioned tests are close to the nominal levels and so the robust tests and MTTs can control the sizes well. Besides, although we assume that HWE holds in each sub-population, a moderate departure from HWE has little impact to the sizes of SGMS and SGME (results skipped for brevity).

We also conducted simulation to investigate the performance of the proposed tests for small sized samples, where the number of cases is at most 100. The results are summarized in Table 3. The settings were the same as those in Table 2 except that the sample sizes in Table 3 were only 10% of those in Table 2. The results show that the proposed tests can keep the size reasonably well even for small sample case.

The powers of the MTTs and robust tests were compared under three genetic models (REC, ADD and DOM). The settings were the same as those in Table 2 except that the nominal level was set to be 0.05 and the GRR was determined so that the optimal MTT has the maximum power of about 80%. The results are summarized in Table 4. In each row, the power of the robust test which performs best among the four robust tests considered in Table 4 is bold-faced.

From Table 4 we notice that although the MTTs can obtain the highest power if the genetic models are correctly specified, the minimum powers of $Z_{\mathrm{MTT}}(0)$ and $Z_{\mathrm{MTT}}(1)$ are below 20% and the minimum powers of $Z_{\mathrm{MTT}}(0.5)$ are between 50% to 60%. On the other hand, the minimum powers of the robust tests are about 65% across all genetic models. Table 4 clearly shows the advantage of the robust tests that, when the genetic model is unknown, the robust tests are more preferred than the MTTs. Besides, from Table 4 we can conclude that if only the REC, ADD and DOM models are

Table 5 The pair-matched case-control study of ACCESS

		Controls			
Caucasian		'11'	'13'	'33'	Total
Cases	'11'	0	0	1	1
	'13'	0	9	36	45
	'33'	2	29	201	232
	Total	2	38	238	278
		Controls			
Female/African-American		'11'	'13'	'33'	Total
Cases	'11'	1	11	8	20
	'13'	8	26	40	74
	'33'	4	34	24	62
	Total	13	71	72	156
		Controls			
Male/African-American		'11'	'13'	'33'	Total
Cases	'11'	1	2	5	8
	'13'	1	14	17	32
	'33'	1	11	11	23
	Total	3	27	33	63
		Controls			
Combined		'11'	'13'	'33'	Total
Cases	'11'	2	13	14	29
	'13'	9	49	93	151
	'33'	7	74	236	317
	Total	18	136	343	497

considered, Z_{SGME} and Z_{MAX3} perform better than the other two robust tests and Z_{MAX3} always dominate Z_{χ^2} with 2 df under such situations. Table 4 also reports ρ^*, which is defined as the minimum correlation of the optimal tests [5]. For example, when REC is the true model, then ρ^* = min(corr($Z_{MTT}(0)$, $Z_{MTT}(0.5)$), corr ($Z_{MTT}(0)$, $Z_{MTT}(1)$)). ρ^* is considered here as a guideline for choosing efficiency robust tests between Z_{SGME} and Z_{MAX3}. From Table 4 we find that when ρ^* is small (around 0.3), Z_{MAX3} performs better or at least as powerful as Z_{SGME}. However, when ρ^* is large (around 0.7), Z_{SGME} is a better choice. Notice that this finding is similar to the property of the efficiency robust procedures in survival data analysis studied by Freidlin et al. [30] who also suggest to use the MAX-type statistic if ρ^* is less than 0.6 or 0.7.

We further compared Z_{SGMS}, Z_{SGME}, Z_{MAX3} and Z_{χ^2} with 2 df under different genetic models. The parameter settings were the same as those of scenario *A* in Table 2 except that λ_2 increased from 1.1 to 2.0 with increments of 0.1 and $\lambda_{1l} = 1 + x(\lambda_{2l} - 1)$. The results are summarized in Figures 2, 3, 4 and 5 with titles *a*, *b*, *c*, *d*, *e*, *f*, *g* representing x = 0.25, 0, 0.25, 0.5, 0.75, 1, 1.25 respectively.

Notice that x = 0, 0.5, 1 (figures b, d, f) correspond to the REC, ADD and DOM models respectively. Under these three commonly used genetic models, Z_{SGMS}, Z_{SGME} and Z_{MAX3} have comparable powers although Z_{MAX3} may be slightly more powerful than the other two tests under the REC and DOM models, and Z_{SGME} may dominate Z_{MAX3} and Z_{SGMS} under the ADD model. Z_{χ^2} with 2 df has the least power among all the tests considered here. x = 0.25 (figure c) indicates a genetic model between REC and ADD and x = 0.75 (figure e) corresponds to a genetic model between ADD and DOM. The performance of the robust tests under such two genetic models is similar to that under ADD. Z_{SGME} is slightly more powerful than Z_{SGMS} and Z_{MAX3}, and Z_{χ^2} with 2 df still obtains the least power.

x = - 0.25 and 1.25 indicate two less plausible models, the under-recessive model (figure a) and over-dominant model (figure g). Under the under-recessive model where $f_{1l} < f_{0l}$, Z_{χ^2} with 2 df is the most powerful test followed by Z_{SGMS} and Z_{MAX3}. Z_{SGME} performs the worst in such a situation. Under the over-dominant model where $f_{1l} > f_{2l}$, all the robust tests perform very similarly.

To summarize, if the mean genetic effect of the heterozygous genotype is between those of the two homozygous genotypes, then we suggest Z_{MAX3}, Z_{SGMS} and Z_{SGME}. On the other hand, if the genetic effects are not ranked in accordance with the genotypes, then Z_{χ^2} with 2 df is preferred. This is reasonable because Z_{χ^2} with 2 df does not take the order of the genetic effects into consideration so it should perform well if the genetic effects are not ranked in accordance with the genotypes.

Notice that in our simulation we consider the common disease common variant (CDCV) which is currently the most popular theory underlying complex disease etiology. However, if the common disease rare variant (CDRV) assumption holds which implies that the disease etiology is caused collectively by multiple rare variants with moderate to high penetrances, the proposed tests perform conservatively and underpowered for detecting association [31]. In this case, the combined multivariate and collapsing (CMC) method proposed by Li and Leal [31] may be used to increase the power of the proposed tests.

An application

We applied MTTs and the robust tests to a matched pair case-control etiologic study of sarcoidosis (ACCESS) [15]. In this study, a total of 497 matched pairs of case-control sets samples based on their age (within 5 years), race (Caucasian and African-American) and gender were recruited to test for association between immunoglobulin gene polymorphism and

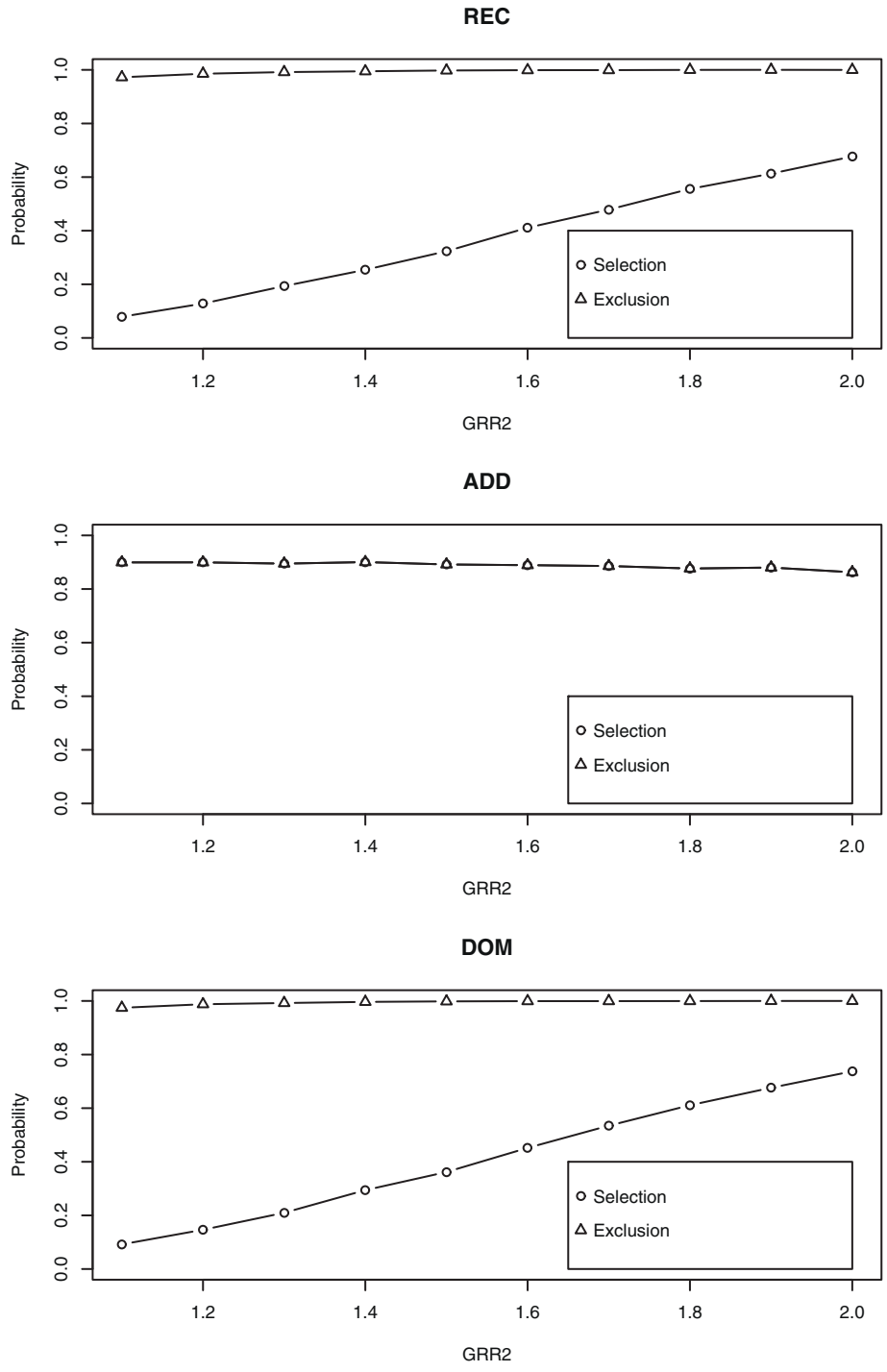

Figure 1 The probabilities of correctly selecting the genetic models and of correctly excluding the most unlikely genetic models based on 10,000 replicates.

sarcoidosis. A subset containing 219 African-American matched pairs was used by Zheng and Tian [21]. We consider the KM(1,3) polymorphism as the candidate marker. After estimating the risk allele frequencies in controls of the matched sets defined by the two confounding factors (gender and race), we find three sub-populations namely Caucasian, Female African-American and Male African-American. The details of the matched data and sub-structure information are summarized in Table 5.

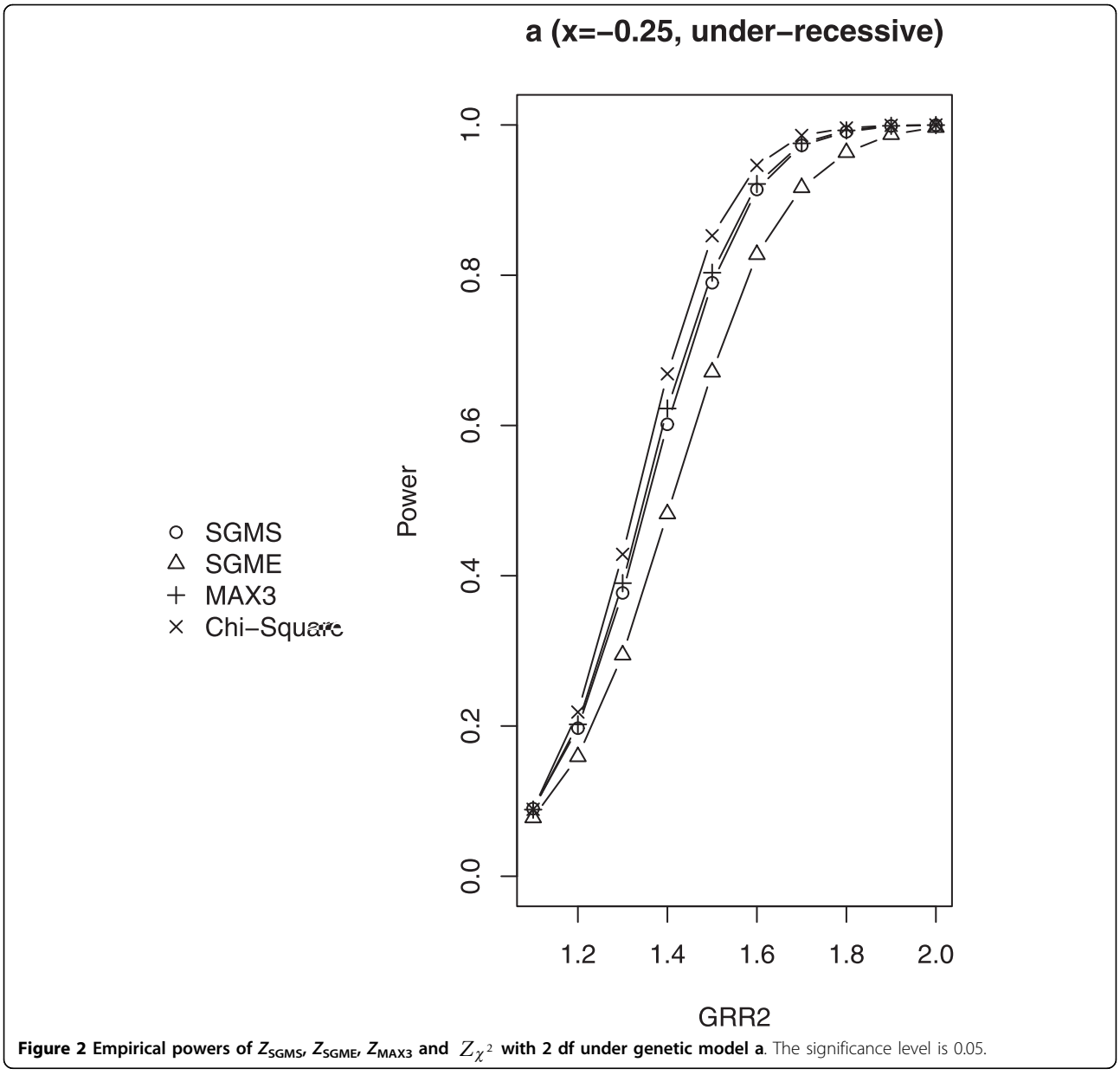

Figure 2 Empirical powers of Z_{SGMS}, Z_{SGME}, Z_{MAX3} and Z_{χ^2} with 2 df under genetic model a. The significance level is 0.05.

First we applied the MTTs optimal for the REC, ADD and DOM models to the data set and obtained the p-values being 0.058, 0.025 and 0.093 for $Z_{\mathrm{MTT}(0)}$, $Z_{\mathrm{MTT}(0.5)}$ and $Z_{\mathrm{MTT}(1)}$ respectively. Thus, whether or not there is a significant association is unclear under a nominal level 0.05 because different genetic models give different answers.

Then we applied Z_{χ^2} with 2 df and Z_{MAX3} to the data set and obtained the p-values as 0.076 and 0.056, which were also hard to provide a more conclusive finding under a significance level of 0.05. Note that the p-value of Z_{MAX3} was calculated according to the asymptotic formula obtained by Zang et al. [29]. Thereafter we applied Z_{SGMS} and Z_{SGME} to the same data. We obtained Z_{SMRT} = 0.124, which falls in the interval [-1.645,1.645] and strongly suggested an ADD model. Thus, for SGMS we select ADD and for SGME we exclude REC and DOM. Using formulas (6) and (8) we obtained the p-values as 0.0398 for Z_{SGMS} and 0.0310 for Z_{SGME}, both suggesting a marginally significant association. According to our simulation, Z_{SGME} is the most powerful robust test under the ADD model. We also obtained the minimum correlation of the optimal tests ρ^* = 0.603, which indicates that Z_{SGME} is a better choice than Z_{MAX3} according to our previous discussion for Table 4. Obviously, our results are consistent with the findings in that discussion. To sum up, we

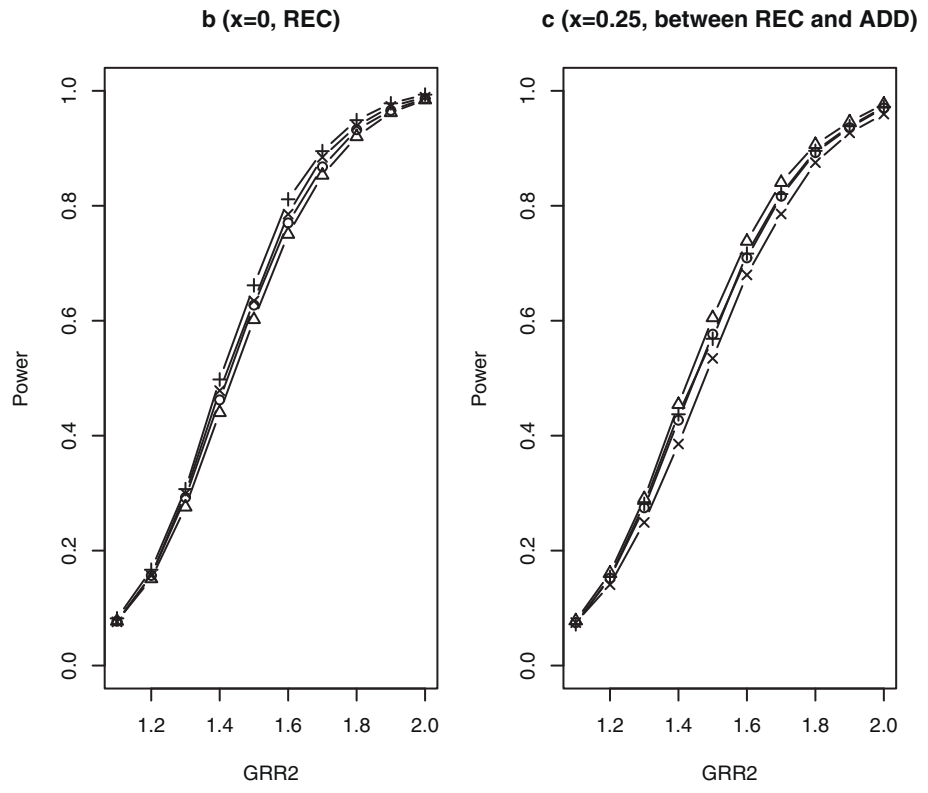

Figure 3 Empirical powers of Z_{SGMS}, Z_{SGME}, Z_{MAX3} and Z_{χ^2} with 2 df under genetic models b and c. The significance level is 0.05.

observe that there is some association between the candidate marker and sarcoidosis.

Discussion

In this paper, we extended the GMS [11] and GME [13] methods to the matched case-control association study and proposed the SGMS and SGME methods so that they can be used when there are confounding factors in the recruited samples. We showed that the p-values of both tests can be determined analytically based on the asymptotic tri-variate normal distributions. Besides, we also reviewed some other robust tests in matched case-control association study such as the MAX3 test and the χ^2 with 2 df test. Simulations were carried out to examine the robustness of all these tests. The tests were also used to analyze a real pair matched data set of sarcoidosis. Simulation results indicate that when the genetic model is unknown, a mis-specification of the genetic model may result in a substantial loss of power for the MTTs. In this situation, robust tests are preferred. Further comparisons among the robust tests were also conducted. According to our simulation, when the genetic effects are ordered in accordance with their genotypes, MAX3, SGMS and SGME are preferred. On the other hand, if the less plausible genetic models such as the over-dominant and under-recessive models cannot be excluded, then χ^2 with 2 df test is a good choice.

We adopted the matching framework in the stage of recruiting samples so our study is a pre-matched case-control association study. In practice, even in the unmatched case-control design matching is still an important tool to eliminate the effect of latent confounding factors such as the population stratification and cryptical relatedness. For example, Guan et al. [24] recently proposed a matched design in an unmatched case-control study. They post-matched individuals by their genotypes followed by a conditional matching analysis to correct for population stratification in genome-wide association studies. In fact, after applying their method or the principal components method [32] and its extension [33] to classify the latent population structure, all the robust tests discussed in this paper can be used as robust approaches as well as correcting the latent population stratification in the unmatched case-control or genome-wide association studies. The regression approach is also suggested in the literature to adjust for confounding factors other than markers. However, if the whole population has many subpopulation due to confounding, the performance of the regression method could be affected because too many

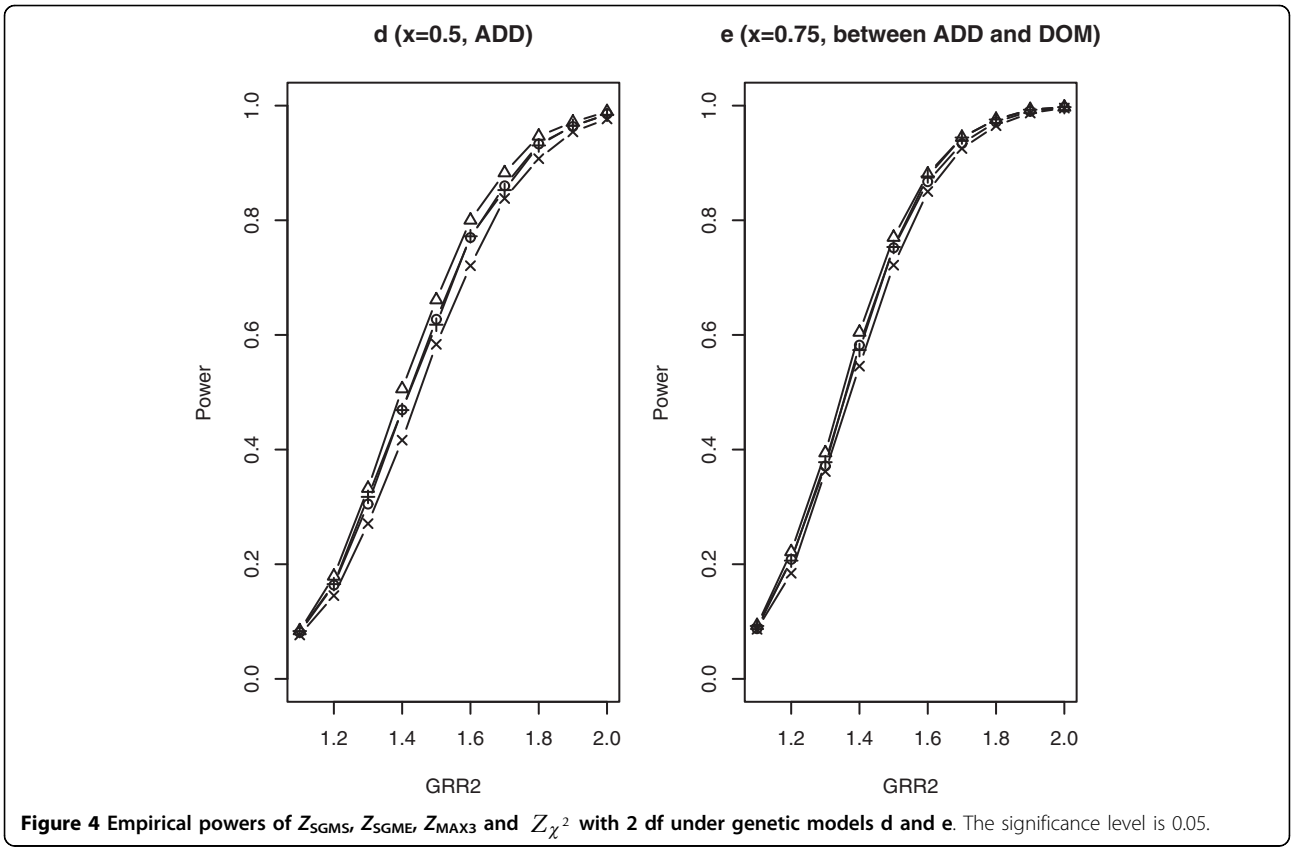

Figure 4 Empirical powers of Z_{SGMS}, Z_{SGME}, Z_{MAX3} and Z_{χ^2} with 2 df under genetic models d and e. The significance level is 0.05.

nuisance parameters need to be estimated. Furthermore, how to derive the variance-covariance matrices of the distribution of the robust tests in this case is still uncertain. Further research in this area is needed.

Conclusion

Simulation results and real data analysis show that SGMS and SGME can keep a correct Type I error rate for stratified data while have good efficiency robustness against genetic model uncertainty. Besides, the proposed formulas in this paper can easily be used to calculate the corresponding p-values. Thus, SGMS and SGME are useful for genetic data analysis of matched case-control design.

Appendix

First we derive the correlation ρ_x between Z_{SMRT} and $Z_{\mathrm{MTT}}(x)$. Define $U(x)$ as the numerator of Z_{MTT}, under the null hypothesis,

$$cov_{H_0}\left(\sum_{l=1}^{L}\hat{\Delta}_l, U(x)\right) = \sum_{l=1}^{L} cov_{H_0}(\hat{\Delta}_l, U(x))$$

Following Zheng and Ng [11], we can obtain that

$$cov_{H_0}(\hat{\Delta}, U(0)) = (m+1)\sum_{l=1}^{L}[p_l^2(1-p_l)^2] + O(1/r)$$

$$cov_{H_0}(\hat{\Delta}, U(0.5)) = O(1/r)$$

$$cov_{H_0}(\hat{\Delta}, U(1)) = -(m+1)\sum_{l=1}^{L}[p_l^2(1-p_l)^2] + O(1/r).$$

When $n \to \infty$, $V(x) \to m(m+1)\sum_{l=1}^{L} r_l(x^2 p_{1l} + p_{2l} - x^2 p_{1l}^2 - p_{2l}^2 - 2x p_{1l} p_{2l})$ and $\hat{\Lambda} \to \sum_{l=1}^{L} \frac{m+1}{r_l m}(1-p_l)^2 p_l^2$. Hence, we have

$$\rho_0 = cov_{H_0}(Z_{\mathrm{SMRT}}, Z_{\mathrm{MTT}}(0)) = \frac{\sum_{l=1}^{L}[p_l^2(1-p_l)^2]}{\sqrt{[\sum_{l=1}^{L}\frac{1}{r_l}p_l^2(1-p_l)^2][\sum_{l=1}^{L} r_l p_l^2(1-p_l^2)]}}$$

$$\rho_{0.5} = cov_{H_0}(Z_{\mathrm{SMRT}}, Z_{\mathrm{MTT}}(0.5)) = 0$$

$$\rho_1 = cov_{H_0}(Z_{\mathrm{SMRT}}, Z_{\mathrm{MTT}}(1)) = -\frac{\sum_{l=1}^{L}[p_l^2(1-p_l)^2]}{\sqrt{[\sum_{l=1}^{L}\frac{1}{r_l}p_l^2(1-p_l)^2][\sum_{l=1}^{L} r_l(2p_l - p_l^2)(1-p_l)^2]}}.$$

Substitute $\hat{p}_l = [2(r_{2l} + s_{2l}) + (r_{1l} + s_{1l})]/[2(m+1)r_l]$

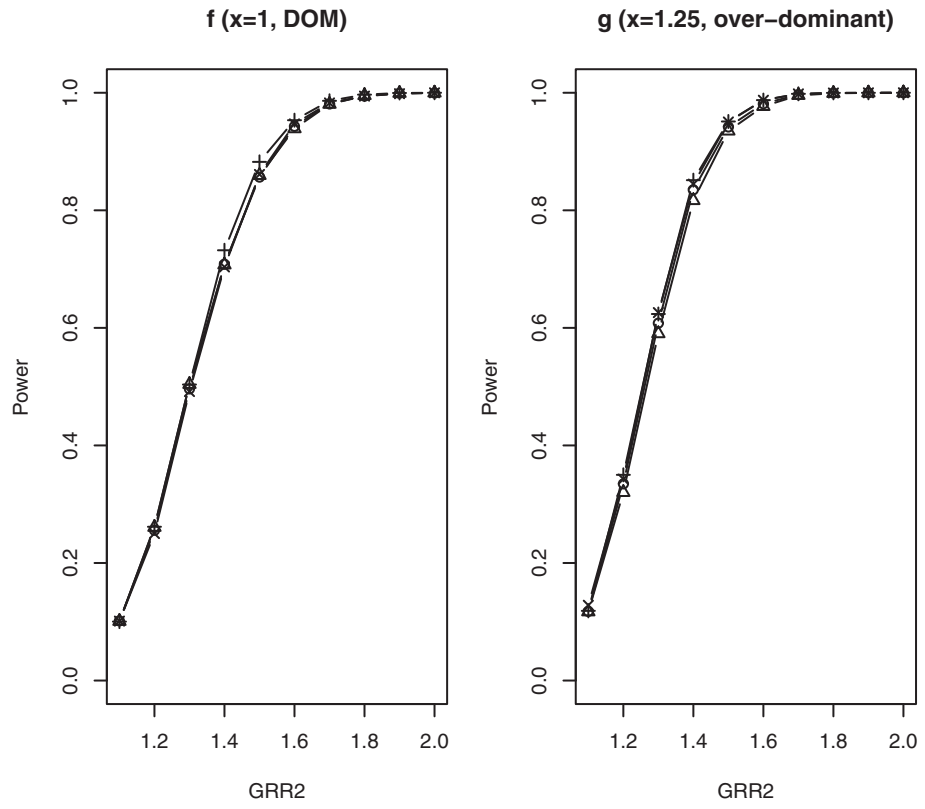

Figure 5 Empirical powers of Z_{SGMS}, Z_{SGME}, Z_{MAX3} and Z_{χ^2} with 2 df under genetic models f and g. The significance level is 0.05.

for p_l, we obtain the estimate $\hat{\rho}_x$ for ρ_x (x = 0, 0.5, 1).

Next we report the correlation $\rho_{x,0.5}$ between $Z_{\mathrm{MTT}}(x)$ and $Z_{\mathrm{MTT}}(0.5)$ (x = 0, 1). Under the null hypothesis,

$$
\begin{aligned}
&covH_0(U(0),U(0.5))\\
&=\sum_{l=1}^{L}\sum_{j=1}^{r_l} cov_{H_0}\left((mX_{1lj|x=0}-\sum_{k=1}^{m}X_{2ljk|x=0}),\ (mX_{1lj|x=0.5}-\sum_{k=1}^{m}X_{2ljk|x=0.5})\right)\\
&=\sum_{l=1}^{L}\sum_{j=1}^{r_l}\left[m^2 cov_{H_0}(X_{1lj|x=0},X_{1lj|x=0.5})+\sum_{k=1}^{m}cov_{H_0}(X_{2ljk|x=0},X_{2ljk|x=0.5})\right]\\
&=(m^2+m)\sum_{l=1}^{L}r_l(p_{2l}-\frac{1}{2}p_{1l}p_{2l}-p_{2l}^2)
\end{aligned}
$$

$$
\begin{aligned}
&covH_0(U(1),U(0.5))\\
&=\sum_{l=1}^{L}\sum_{j=1}^{r_l} cov_{H_0}\left((mX_{1lj|x=0.5}-\sum_{k=1}^{m}X_{2ljk|x=0.5}),\ (mX_{1lj|x=1}-\sum_{k=1}^{m}X_{2ljk|x=1})\right)\\
&=\sum_{l=1}^{L}\sum_{j=1}^{r_l}\left\{m^2 cov_{H_0}(X_{1lj|x=0.5},X_{1lj|x=1})+\sum_{k=1}^{m}cov_{H_0}(X_{2ljk|x=0.5},X_{2ljk|x=1})\right\}\\
&=(m^2+m)\sum_{l=1}^{L}r_l(\frac{1}{2}p_{1l}+p_{2l}-\frac{1}{2}p_{1l}^2-\frac{3}{2}p_{1l}p_{2l}-p_{2l}^2).
\end{aligned}
$$

Since $V(x)\to m(m+1)\sum_{l=1}^{L}r_l(x^2p_{1l}+p_{2l}-x^2p_{1l}^2-p_{2l}^2-2xp_{1l}p_{2l})$, after simple algebra, we have,

$$
\rho_{0,0.5}=\frac{\sum_{l=1}^{L}r_l(p_{2l}-\frac{1}{2}p_{1l}p_{2l}-p_{2l}^2)}{\sqrt{\left[\sum_{l=1}^{L}r_l(p_{2l}-p_{2l}^2)\right]\left[\sum_{l=1}^{L}r_l(\frac{1}{4}p_{1l}+p_{2l}-\frac{1}{4}p_{1l}^2-p_{2l}^2-p_{1l}p_{2l})\right]}}
$$

$$
\rho_{1,0.5}=\frac{\sum_{l=1}^{L}r_l\left(\frac{1}{2}p_{1l}+p_{2l}-\frac{1}{2}p_{1l}^2-\frac{3}{2}p_{1l}p_{2l}-p_{2l}^2\right)}{\sqrt{\left[\sum_{l=1}^{L}r_l(p_{1l}+p_{2l}-p_{1l}^2-p_{2l}^2-2p_{1l}p_{2l})\right]\left[\sum_{l=1}^{L}r_l\left(\frac{1}{4}p_{1l}+p_{2l}-\frac{1}{4}p_{1l}^2-p_{2l}^2-p_{1l}p_{2l}\right)\right]}}
$$

Substitute $\hat{p}_{il}=(r_{il}+s_{il})/[(m+1)r_l]$ for p_{il} (i = 0,1,2), we obtain the estimate $\hat{p}_{x,0.5}$ for $\rho_{x,0.5}$ (x = 0,1).

Acknowledgements

The research of Y. Zang was partially supported by the China Natural Science Foundation grant 10701067 and the research of W. K. Fung was partially supported by the HKU Research Output Prize Funding.

Authors' contributions

ZY carried out the project and wrote the draft of the manuscript. FWK proposed the idea and revised the manuscript. Both authors read and approved the manuscript.

Received: 30 March 2010 Accepted: 12 October 2010
Published: 12 October 2010

References

1. Cochran WG: **Some methods for strengthening the common chi-square test.** *Biometrics* 1954, **10**:417-451.
2. Armitage P: **Test for linear trends in proportions and frequencies.** *Biometrics* 1955, **11**:375-386.
3. Sasieni PD: **From genotypes to genes: doubling the sample size.** *Biometrics* 1997, **53**:1253-1261.
4. Zheng G, Freidlin B, Li Z, Gastwirth JL: **Choice of scores in trend tests for case-control studies of candidate-gene associations.** *Biometrical Journal* 2003, **45**:335-348.
5. Gastwirth JL: **On robust procedures.** *Journal of the American Statistical Association* 1966, **61**:929-948.
6. Gastwirth JL: **The use of maximin efficiency robust tests in combining contingency tables and survival analysis.** *Journal of the American Statistical Association* 1985, **80**:380-384.
7. Freidlin B, Zheng G, Li Z, Gastwirth JL: **Trend tests for case-control studies of genetic markers: power, sample size and robustness.** *Human Heredity* 2002, **53**:146-152.
8. Gonzalez JR, Carrasco JL, Dubridge F, Armengol L, Estivill X, Moreno V: **Maximizing association statistics over genetic models.** *Genetic Epidemiology* 2008, **32**:246-254.
9. Li Q, Zheng G, Li Z, Yu K: **Efficient approximation of p-value of the maximum of correlated tests, with applications to genome-wide association studies.** *Annals of Human Genetics* 2008, **27**:397-406.
10. Joo J, Kwak M, Chen Z, Zheng G: **Tutorial in biostatistics: Efficiency robust statistics for genetic linkage and association studies under genetic model uncertainty.** *Statistics in Medicine* 2010, **29**:158-180.
11. Zheng G, Ng HKT: **Genetic model selection in two-phase analysis for case-control association studies.** *Biostatistics* 2008, **9**:391-399.
12. Song K, Elston RC: **A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine-mapping in case-control studies.** *Statistics in Medicine* 2006, **25**:105-126.
13. Joo J, Kwak M, Zheng G: **Improving power for testing genetic association in case-control studies by reducing alternative space.** *Biometrics* 2010, **66**:266-276.
14. Minelli C, Thompson JR, Abrams KR, Lambert PC: **Bayesian implementation of a genetic model-free approach to the meta-analysis of genetic association studies.** *Statistics in Medicine* 2005, **24**:3845-3861.
15. ACCESS Research Group: **Design of a case control etiologic study of sarcoidosis (ACCESS).** *Journal of Clinical Epidemiology* 1999, **52**:1173-1186.
16. Manusirivithaya S, Siriaunkgul S, Khunamornpong S, Sripramote M, Sampatanukul P, Tangjitgamol S, Srisomboon J: **Association between Bcl-2 expression and tumor recurrence in cervical cancer: A matched case-control study.** *Gynecologic Oncology* 2006, **102**:263-269.
17. Suzuki H, Li YN, Dong XQ, Hassan MM, Abbruzzese JL, Li DH: **Effect of insulin-like growth factor gene polymorphisms alone or in interaction with diabetes on the risk of pancreatic cancer.** *Cancer Epidemiology Biomarker and Prevention* 2008, **17**:3467-3473.
18. Yin JY, Vogel U, Ma Y, Qi R, Wang HW: **Association of DNA repair gene XRCC1 and lung cancer susceptibility among nonsmoking Chinese women.** *Cancer Genetics and Cytogenetics* 2009, **188**:26-31.
19. Lee WC: **Case-control association studies with matching and genomic controlling.** *Genetic Epidemiology* 2004, **27**:1-13.
20. Kraft P, Cox DG, Paynter RA, hunter D, De Vivio I: **Accounting for haplotype uncertainty in matched association studies: A comparison of simple and flexible techniques.** *Genetic Epidemiology* 2005, **28**:261-272.
21. Zheng G, Tian X: **Robust trend tests for genetic association using matched case-control design.** *Statistics in Medicine* 2006, **25**:3160-3173.
22. Zhang H, Zhang H, Li Z, Zheng G: **Statistical methods for haplotype-based matched case-control association studies.** *Genetic Epidemiology* 2007, **31**:316-326.
23. Chen J, Rodriguez C: **Conditional likelihood methods for haplotype-based association analysis using matched case-control data.** *Biometrics* 2007, **63**:1099-1107.
24. Guan W, Liang L, Boehnke M, Abecasis GR: **Genotype-based matching to correct for population stratification in large-scale case-control genetic association studies.** *Genetic Epidemiology* 2009, **33**:508-517.
25. Day NE, Byar DP: **Testing hypotheses in case-control studies-equivalence of Mantel-Haenszel statistics and logit score tests.** *Biometrics* 1979, **35**:623-630.
26. Mantel N, Haenszel W: **Statistical aspects of the analysis of data from retrospective studies of disease.** *Journal of the National Cancer Institute* 1959, **22**:719-748.
27. Agresti A: *Categorical data analysis* John Wiley & Sons, Inc, second 2002.
28. Zheng G, Chen Z: **Comparison of maximum statistics for hypothesis testing when nuisance parameter is present only under the alternative.** *Biometrics* 2005, **61**:254-258.
29. Zang Y, Fung WK, Zheng G: **Asymptotic powers for matched trend tests and robust matched trend tests in case-control genetic association studies.** *Computational Statistics and Data Analysis* 2010, **54**:65-77.
30. Freidlin B, Podgor MJ, Gastwirth JL: **Efficiency robust tests for survival or order categorical data.** *Biometrics* 1999, **55**:883-886.
31. Li B, Leal SM: **Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data.** *American Journal of human genetics* 2008, **83**:311-321.
32. Price AL, Patterson NJ, Plenge RM, Reich D: **Principle components analysis corrects for stratification in genome-wide association studies.** *Nature Genetics* 2006, **33**:904-909.
33. Li Q, Wacholder S, Hunter DJ, Hoover RC, Chanock S, Thomas G, Yu K: **Genetic background comparison using distance-based regression with application in population stratification evaluation and adjustment.** *Genetic Epidemiology* 2009, **33**:432-441.

doi:10.1186/1471-2156-11-91
Cite this article as: Zang and Fung: **Robust tests for matched case-control genetic association studies.** *BMC Genetics* 2010 **11**:91.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central