Published in final edited form as:
Ann Hum Genet. 2010 July ; 74(4): 351–360. doi:10.1111/j.1469-1809.2010.00588.x.

Influence of population stratification on population-based marker-disease association analysis

TENGFEI LI[1], **ZHAOHAI LI**[2], **ZHILIANG YING**[3], and **HONG ZHANG**[4,*]

[1] Department of Mathematics, Fudan University, 220 Handan Road, Shanghai 200433, P.R. China

[2] Department of Statistics, George Washington University, 2140 Pennsylvania Ave., N.W. Washington, DC 20052, USA

[3] Department of Statistics, Columbia University, 1255 Amsterdam Avenue, New York, NY 10027, USA

[4] Department of Statistics and Finance, University of Science and Technology of China, 96 Jinzhai Road, Hefei, Anhui 230026, P.R. China

Summary

Population-based genetic association analysis may suffer from the failure to control for confounders such as population stratification (PS). There has been extensive study on the influence of PS on candidate gene-disease association analysis, but much less attention has been paid to its influence on marker-disease association analysis. In this paper, we focus on the Pearson chi-square test and the trend test for marker-disease association analysis. The mean and variance of the test statistics are derived under presence of PS, so that the power and inflated type I error rate can be evaluated. It is shown that the bias and the variance distortion are not zero in the presence of both PS and penetrance heterogeneity (PH). Unlike the candidate gene-disease association analysis, when PS is present, the bias is not zero no matter whether PH is present or not. This work generalizes the results of Ewens and Spielman (1995), where only the fully recessive penetrance model is considered and only the bias is calculated. It is shown that candidate gene-disease association analysis can be treated as a special case of marker-disease association analysis. Consequently, our results extend previous study on the candidate gene-disease association analysis. A simulation study confirms the theoretical findings.

Keywords

bias; marker-disease association; penetrance heterogeneity; population stratification; variance distortion

INTRODUCTION

Population-based gene-disease association analysis is the most commonly used statistical method for detecting genetical variants underlying human diseases (Risch and Merikangas, 1996; Risch, 2000). Such an approach makes use of the case-control design, which is easy to carry out and cost-effective. However, the case-control studies often suffer from a failure to account for confounders such as population stratification (PS), resulting in spurious

[*]Corresponding author: Hong Zhang, Ph.D, Department of Statistics and Finance, University of Science and Technology of China, 96 Jinzhai Road, Hefei, Anhui 230026, P.R. China. Phone: +86-551-3600560; Fax: +86-551-3600025, zhangh@ustc.edu.cn.

associations (Knowler et al., 1988; Lander and Schork, 1994; Cardon and Palmer, 2003; Campbell et al., 2005). When parental genotypes of affected individuals are available, the transmission disequilibrium test (TDT) can be used to control for false positives due to PS. However, for diseases with a late age of onset, the parental genotypes are generally unavailable and, therefore, TDT is not applicable. There have been studies in recent years on the impacts of PS on gene-disease association analysis, particularly with respect to the bias and/or variance distortion of the test statistic (Ewens and Spielman, 1995; Gorroochurn et al., 2004; Heiman et al. 2004; Qin et al., 2006; Whittemore, 2006; Li et al., 2009; Zheng et al. 2009). Most of the existing studies focus on a candidate locus, where the null hypothesis states that the penetrance does not depend on genotype in any subpopulation. Markers are widely used in preliminary association analyses for detecting disease genes, especially in genome-wide association analyses. However, the impact of PS on marker-disease association has not been studied, with the exception of the work of Ewens and Spielman (1995), where the bias of the test statistic for marker-disease association was obtained by assuming a very special disease model, namely a fully recessive penetrance model, but the variance distortion and power function were not given.

In this paper, we extend the results of Ewens and Spielman (1995) to a more general class of models, without assuming any mode of inheritance. Besides the bias, we also derive the variance distortion and power function under both the null hypothesis and the alternative hypothesis. The null hypothesis in the marker locus case states that the linkage disequilibrium (LD) measures are zero in any subpopulation. It is shown that the bias and variance distortion under the null hypothesis are not zero in the presence of both PS and penetrance heterogeneity (PH). In addition, the bias is not zero when PS is present, even if PH is not, in contrast to the result for candidate gene-disease association analysis, where the bias is equal to zero when PH is absent. We demonstrate that candidate gene-disease association analysis can be treated as a special case of marker-disease association analysis, so that our results are extensions of the previous work on candidate gene-disease association analysis. Because the null hypothesis in the marker locus case is different from that in the candidate locus case, the existing results under the null hypothesis for a candidate locus cannot be transformed through simple reparameterization to yield our results.

Our contributions consist of the following: 1) we extend the existing results to the general case of marker-disease association analysis; 2) we find that the presence of PS can lead to bias of the marker-disease association test statistic even when the PH is absent, while in the candidate locus case the bias is always zero when PH is absent; 3) we derive the power functions for the Pearson chi-square test and the trend test so that one can study the impact of PS and PH on both the type I and type II errors of the two tests.

The rest of the paper is organized as follows. Some notation and definitions are given in the next section. The subsequent sections give the mean and variance of the Pearson chi-square test statistic and the trend test statistic and their power functions. A small-scale simulation study is conducted to verify the theoretical results. This is followed by some concluding remarks.

NOTATION

Suppose that in a case-control study n_1 cases and n_2 controls are sampled from their respective populations, where $n = n_1 + n_2$. A marker with alleles M and m is then genotyped, with the counts of genotypes and alleles given in Tables 1 and 2, respectively.

Let the proportions of cases and controls with allele M be denoted by $\hat{q}_D = (2D_2 + D_1)/(2n_1)$ and $\hat{q}_C = (2C_2 + C_1)/(2n_2)$, respectively. In addition, let $\hat{q} = (2D_2 + D_1 + 2C_2 + C_1)/(2n)$ be the

proportion of the pooled sample with allele *M*. The commonly used Pearson chi-square test statistic based on allele counts is the square of the following test statistic:

$$T=\frac{\widehat{q}_D-\widehat{q}_C}{V^{1/2}}, \tag{1}$$

where

$$V=\widehat{q}(1-\widehat{q})(\frac{1}{2n_1}+\frac{1}{2n_2}) \tag{2}$$

is an estimate of the variance of $\hat{q}_D-\hat{q}_C$. The estimate V is used when Hardy-Weinberg equilibrium (HWE) holds in the overall population. In the "variance adjustment" section, we will present a variance estimate that is valid even when HWE does not hold.

We assume that the total population consists of K subpopulations, with HWE holding within each subpopulation. Throughout this paper, we shall use S_i to denote the event that a randomly selected individual is from subpopulation i and w_i to denote the proportion of the total population that belongs to subpopulation i. Assume that only one locus, with alleles A and a, is responsible for the disease. For subpopulation i, let p_i and q_i denote the frequencies of alleles A and M, respectively. Thus, the frequencies of alleles A and M in the overall population are $p=\sum_{i=1}^{K} w_i p_i$ and $q=\sum_{i=1}^{K} w_i q_i$, respectively. Furthermore, let x_{i1}, x_{i2}, x_{i3} and x_{i4} denote the frequencies of gametes MA, Ma, mA and ma, respectively, and $\delta_i = x_{i1}x_{i4}-x_{i2}x_{i3}$ the LD measure between the marker locus and the disease locus. Finally, denote by f_{2i}, f_{1i} and f_{0i} the penetrances of genotypes AA, Aa and aa, respectively. Under the HWE, the frequencies of genotypes AA, Aa and aa at the disease locus for subpopulation i are $p_{2i}=p_i^2$, $p_{1i} = 2p_i(1-p_i)$ and $p_{0i} = (1-p_i)^2$, respectively. The null hypothesis of linkage equilibrium becomes

$$H_0:\delta_1=\ldots=\delta_K. \tag{3}$$

Under the null hypothesis H_0, all LD measures δ_i, $i = 1,\ldots,K$, are equal to 0, while under the alternative hypothesis, at least one LD measure is not equal to 0. It is clear that the null hypothesis implies that the marker is not associated with the disease.

Definition 1

PS is said to be present if the allele frequencies at the marker locus are heterogenous, i.e., the q_i vary with i.

EXPECTATION OF FREQUENCY DIFFERENCE

In this section, we calculate the expectations of $\hat{q}_D$ and $\hat{q}_C$ under both the null and the alternative hypotheses. We study the null expectation of $\hat{q}_D-\hat{q}_C$, which is termed *bias*. Hereafter, let $Y = 1$ denote the event that a randomly chosen individual is a case, and $Y = 2$ the event that a randomly chosen individual is a control.

By definition, the expectations of $\hat{q}_D$ and $\hat{q}_C$ are equal to

$$E(\widehat{q}_D)=P(MM|Y=1)+\frac{1}{2}P(Mm|Y=1) \tag{4}$$

and

$$E(\widehat{q}_C)=P(MM|Y=2)+\frac{1}{2}P(Mm|Y=2), \tag{5}$$

respectively. The disease prevalence, which we denote by B, satisfies $B=\sum_{i=1}^{K} w_i \sum_{j=0}^{2} f_{ji} p_{ji}$ by the Law of Total Probability. In Appendix I, we show that,

$$P(MM|Y=1)=\frac{1}{B}\{\sum_{i=1}^{K} w_i q_i^2 \sum_{j=0}^{2} f_{ji} p_{ji}+\sum_{i=1}^{K} w_i \delta_i^2 (f_{2i}+f_{0i}-2f_{1i})+2\sum_{i=1}^{K} w_i q_i \delta_i [\, p_i (f_{2i}-f_{1i})+(1-p_i)(f_{1i}-f_{0i})]\} \tag{6}$$

and

$$P(Mm|Y=1)=\frac{2}{B}\{\sum_{i=1}^{K} w_i q_i (1-q_i) \sum_{j=0}^{2} f_{ji} p_{ji}-\sum_{i=1}^{K} w_i \delta_i^2 (f_{2i}+f_{0i}-2f_{1i})+\sum_{i=1}^{K} w_i (1-2q_i) \delta_i [\, p_i (f_{2i}-f_{1i})+(1-p_i)(f_{1i}-f_{0i})]\}. \tag{7}$$

Similarly, we have

$$P(MM|Y=2)=\frac{1}{1-B}\{\sum_{i=1}^{K} w_i q_i^2 \sum_{j=0}^{2} (1-f_{ji}) p_{ji}-\sum_{i=1}^{K} w_i \delta_i^2 (f_{2i}+f_{0i}-2f_{1i})-2\sum_{i=1}^{K} w_i q_i \delta_i [\, p_i (f_{2i}-f_{1i})+(1-p_i)(f_{1i}-f_{0i})]\} \tag{8}$$

and

$$P(Mm|Y=2)=\frac{2}{1-B}\{\sum_{i=1}^{K} w_i q_i (1-q_i) \sum_{j=0}^{2} (1-f_{ji}) p_{ji}+\sum_{i=1}^{K} w_i \delta_i^2 (f_{2i}+f_{0i}-2f_{1i})-\sum_{i=1}^{K} w_i (1-2q_i) \delta_i [\, p_i (f_{2i}-f_{1i})+(1-p_i)(f_{1i}-f_{0i})]\}. \tag{9}$$

Substituting the above four probabilities for those in (4) and (5) gives

$$\Delta=E(\widehat{q}_D-\widehat{q}_C)=\frac{A_1+A_2}{B(1-B)}, \tag{10}$$

where

$$A_1=\sum_{j=0}^{2}[\sum_{i=1}^{K}w_i f_{ji}p_{ji}q_i - q\sum_{i=1}^{K}w_i f_{ji}p_{ji}] \tag{11}$$

and

$$A_2=\sum_{i=1}^{K}w_i\delta_i[(f_{2i}-f_{1i})p_i+(f_{1i}-f_{0i})(1-p_i)]. \tag{12}$$

Since $A_2 = 0$ under the null hypothesis, the bias is $A_1/[B(1-B)]$. Define a random variable Z with probability function $P(Z = i) = w_i$, $i = 1, \cdots, K$. Then $A_1=\mathrm{Cov}(q_Z, \sum_{j=0}^{2} f_{jZ}p_{jZ})$, where q_Z, f_{jZ} and p_{jZ} are conditional probabilities that are equal to q_i, f_{ji} and p_{ji}, respectively, conditional on $Z = i$.

The following are some scenarios that occur in practice.

Scenario 1

If PS is absent, then random variable q_Z degenerates to a constant. In this scenario, $A_1 = 0$ and the bias is zero.

Scenario 2

If PS is present but PH is absent (i.e., $f_{ji} = f_{j1}$, $j = 0,1,2$, $i = 1,\cdots,K$), then the random variables f_{jZ}, $j = 0,1,2$, degenerate to constants, but A_1 is not zero. Hence the bias is not zero in general since q_Z and p_{jZ}, $j = 0,1,2$, are not necessarily constant.

Scenario 3

If both PS and PH are present, then the bias is not zero in general.

Remark 1—When $f_{2i} = 1$, $f_{0i} = f_{1i} = 0$, $i = 1, \cdots, K$, the model degenerates to the so-called *fully recessive penetrance model* and the expectation becomes

$$\frac{\sum_{i=1}^{K}w_i q_i p_i^2 - q\sum_{i=1}^{K}w_i p_i^2+\sum_{i=1}^{K}w_i\delta_i p_i}{B(1-B)}. \tag{13}$$

The above expression is almost identical to expression (5) in Ewens and Spielman (1995).

Remark 2—If the marker locus and the disease locus coincide, so that $p_i = q_i$ and the LD measures are $\delta_i=p_i - p_i^2$, then the marker locus becomes a candidate locus. In this case, (6)–(9) become

$$P(MM|Y=1)=(\sum_{i=1}^{K} w_i f_{2i} p_{2i})/(\sum_{i=1}^{K} w_i \sum_{j=0}^{2} f_{ji} p_{ji}),$$
$$P(Mm|Y=1)=(\sum_{i=1}^{K} w_i f_{1i} p_{1i})/(\sum_{i=1}^{K} w_i \sum_{j=0}^{2} f_{ji} p_{ji}),$$
$$P(MM|Y=2)=(\sum_{i=1}^{K} w_i (1-f_{2i}) p_{2i})/(\sum_{i=1}^{K} w_i \sum_{j=0}^{2} (1-f_{ji}) p_{ji})$$

and

$$P(Mm|Y=2)=(\sum_{i=1}^{K} w_i (1-f_{1i}) p_{1i})/(\sum_{i=1}^{K} w_i \sum_{j=0}^{2} (1-f_{ji}) p_{ji}).$$

The resulting expectation corresponds to the candidate locus case studied by Li et al. (2009). In the candidate locus case, the null hypothesis is $f_{0i} = f_{1i} = f_{2i}$ for $i = 1,\ldots, K$, and the bias is equal to zero if either PS or PH is absent. In the marker locus case that we study in the current paper, however, the bias is generally not zero if PH is absent but PS is present.

VARIANCE OF THE FREQUENCY DIFFERENCE

In Appendix II, we derive the following variance formula for $\hat{q}_D$:

$$\mathrm{Var}(\widehat{q}_D)=\frac{1}{4n_1}[4P(MM|Y=1)(1-P(MM|Y=1))+P(Mm|Y=1)(1-P(Mm|Y=1))-4P(MM|Y=1)P(Mm|Y=1)]. \quad (14)$$

Under the null hypothesis H_0, the conditional probabilities $P(MM|Y = 1)$ and $P(Mm|Y = 1)$ given by (6) and (7) are equal to

$$P_{H_0}(MM|Y=1)=\frac{1}{B}\sum_{i=1}^{K} w_i q_i^2 \sum_{j=0}^{2} f_{ji} p_{ji}=\sum_{i=1}^{K} q_i^2 \alpha_i=\bar{q}_D^2+\sigma_D^2 \quad (15)$$

and

$$P_{H_0}(Mm|Y=1)=\frac{2}{B}\sum_{i=1}^{K} q_i(1-q_i)\sum_{j=0}^{2} f_{ji} p_{ji}=\sum_{i=1}^{K} 2q_i(1-q_i)\alpha_i=2\bar{q}_D(1-\bar{q}_D)-2\sigma_D^2,$$

respectively, where

$$\bar{q}_D=\sum_{i=1}^{K}\alpha_i q_i,\ \ \sigma_D^2=\sum_{i=1}^{K}\alpha_i(q_i-\bar{q}_D)^2 \quad \text{and}\ \alpha_i=\frac{w_i}{B}\sum_{j=0}^{2} f_{ji} p_{ji}.$$

It follows from (14), (15) and (16) that the null variance of $\hat{q}_D$ is

$$\mathrm{Var}_{H_0}(\widehat{q_D})=\frac{\overline{q}_D(1-\overline{q}_D)+\sigma_D^2}{2n_1}.$$

Similarly,

$$\mathrm{Var}_{H_0}(\widehat{q_C})=\frac{\overline{q}_C(1-\overline{q}_C)+\sigma_C^2}{2n_2},$$

where

$$\overline{q}_C=\sum_{i=1}^{K}\gamma_i q_i,\ \ \sigma_C^2=\sum_{i=1}^{K}\gamma_i(q_i-\overline{q}_C)^2 \quad \text{and } \gamma_i=\frac{w_i}{1-B}\sum_{j=0}^{2}(1-f_{ji})p_{ji}.$$

Under the alternative hypothesis, the variance of $\hat{q}_D$ can be expressed as

$$\begin{aligned}\mathrm{Var}(\widehat{q_D})=&\frac{\overline{q}_D(1-\overline{q}_D)+\sigma_D^2}{2n_1}\\&+\frac{1}{2n_1B}\sum_{i=1}^{K}w_i\delta_i^2(f_{2i}+f_{0i}-2f_{1i})\\&-\frac{1}{n_1B^2}\{\sum_{i=1}^{K}w_i\delta_i[p_i(f_{2i}-f_{1i})+(1-p_i)(f_{1i}-f_{0i})]\}^2\\&+\frac{1}{2n_1B}\sum_{i=1}^{K}\delta_i w_i[p_i(f_{2i}-f_{1i})+(1-p_i)(f_{1i}-f_{0i})](1\\&+2q_i-4\overline{q}_D).\end{aligned} \tag{16}$$

We refer to Appendix III for its detailed derivation. Similarly, the variance of $\hat{q}_C$ is equal to

$$\begin{aligned}\mathrm{Var}(\widehat{q_C})=&\frac{\overline{q}_C(1-\overline{q}_C)+\sigma_C^2}{2n_2}\\&-\frac{1}{2n_2(1-B)}\sum_i w_i\delta_i^2(f_{2i}+f_{0i}-2f_{1i})\\&-\frac{1}{n_2(1-B)^2}\{\sum_i w_i\delta_i[p_i(f_{2i}-f_{1i})+(1-p_i)(f_{1i}-f_{0i})]\}^2\\&-\frac{1}{2n_2(1-B)}\sum_i w_i\delta_i[p_i(f_{2i}-f_{1i})+(1-p_i)(f_{1i}-f_{0i})](1\\&+2q_i-4\overline{q}_C).\end{aligned} \tag{17}$$

By virtue of the independence between the cases and controls, the variance of $\hat{q}_D-\hat{q}_C$ is

$$\sigma^2 \equiv \mathrm{Var}(\hat{q}_D)+\mathrm{Var}(\hat{q}_C). \tag{18}$$

In particular, the null variance of $\hat{q}_D-\hat{q}_C$ is

$$\sigma_0^2 \equiv \frac{\bar{q}_D(1-\bar{q}_D)+\sigma_D^2}{2n_1}+\frac{\bar{q}_C(1-\bar{q}_C)+\sigma_C^2}{2n_2}. \tag{19}$$

Definition 2

We say *variance distortion* exists if under the null hypothesis, the variance estimator V for $\hat{q}_D-\hat{q}_C$ as given by (2), is not asymptotically equivalent to the true variance, that is, if V/σ_0^2 does not converge to 1 with probability 1 under the null hypothesis.

By the Law of Large Numbers, $\hat{q}_D \to \bar{q}_D$ and $\hat{q}_C \to \bar{q}_C$, which imply that $\hat{q}$ converges to $c_1\bar{q}_D + c_2\bar{q}_C$ with probability 1, where $c_j = \lim_{n\to\infty} n_j/n$, $j = 1,2$. It follows that under the null hypothesis V is asymptotically equivalent to

$$\bar{\sigma}^2 \equiv [(c_1\bar{q}_D+c_2\bar{q}_C)(1-c_1\bar{q}_D-c_2\bar{q}_C)](\frac{1}{2n_1}+\frac{1}{2n_2}). \tag{20}$$

Remark 3

If PS is absent, then under the null hypothesis $\bar{q}_D = \bar{q}_C = q$ and $\sigma_D^2=\sigma_C^2=0$. Hence, $\sigma_0^2=\bar{\sigma}^2$ and the variance distortion vanishes under the null hypothesis and HWE. Otherwise, the variance distortion is present in general.

VARIANCE ADJUSTMENT

In this section, we derive the power function of the test statistic T, which is given by (1). By the Central Limit Theorem, $T_A = (\hat{q}_D-\hat{q}_C-\Delta)/\sigma$ converges in distribution to the standard normal distribution, since the mean and variance of $\hat{q}_D-\hat{q}_C$ are Δ and σ^2. The two-sided T test at level of significance α is determined by rejection region $\{|T|> u_{\alpha/2}\}$, where $u_{\alpha/2}$ is the upper $\alpha/2$ -quantile of the standard normal distribution. The corresponding power function is therefore approximated by

$$1-\Phi(\frac{u_{\alpha/2}\bar{\sigma}-\Delta}{\sigma})+\Phi(\frac{-u_{\alpha/2}\bar{\sigma}-\Delta}{\sigma}), \tag{21}$$

where Φ is the standard normal distribution function and $\bar{\sigma}^2$ is defined by (20).

As we mentioned earlier, variance distortion exists in presence of PS. Therefore, it is necessary to use a consistent estimate of the variance σ^2. Notice that under the null hypothesis, σ^2 becomes $\sigma_0^2=(\bar{q}_D(1-\bar{q}_D)+\sigma_D^2)/(2n_1)+(\bar{q}_C(1-\bar{q}_C)+\sigma_C^2)/(2n_2)$. We can estimate it with

$$V^*=\frac{\widehat{q}_D(1-\widehat{q}_D)+\widehat{\sigma}_D^2}{2n_1}+\frac{\widehat{q}_C(1-\widehat{q}_C)+\widehat{\sigma}_C^2}{2n_2}, \tag{22}$$

where $\widehat{\sigma}_D^2=D_2/n_1-\widehat{q}_D^2$ and $\widehat{\sigma}_C^2=C_2/n_2-\widehat{q}_C^2$ are consistent estimates of σ_D^2 and σ_C^2, respectively. The estimator V^* was used by Li et al. (2009) for the candidate locus. In the marker locus case, we can show that V^* is asymptotically equivalent to σ^2 under both the null hypothesis and the alternative hypothesis. Actually, V^* is a special estimate of the trend test statistic that will be studied in the next section, and it will be shown that V^* is asymptotically equivalent to σ^2 even when HWE does not hold.

Now, a modification of T takes the form

$$T^*=\frac{\widehat{q}_D-\widehat{q}_C}{(V^*)^{1/2}}. \tag{23}$$

The T^* test with rejection region $\{|T^*|>u_{\alpha/2}\}$ has an approximate power function

$$1-\Phi(u_{\alpha/2}-\Delta/\sigma)+\Phi(-u_{\alpha/2}-\Delta/\sigma). \tag{24}$$

EXTENSION TO TREND TEST

The trend test statistic is defined as

$$T_x=\frac{(D_2/n_1-C_2/n_2)+x(D_1/n_1-C_1/n_2)}{V_x^{1/2}},$$

where x is a given real number between 0 and 1 and V_x is an estimator of the variance of the numerator. From (6)–(9), it follows that the expectation of $(D_2/n_1-C_2/n_2)+x(D_1/n_1-C_1/n_2)$ is

$$\begin{aligned}\Delta_x=&\frac{1}{B(1-B)}\{\sum_{i=1}^{K}w_i[q_i^2+2xq_i(1-q_i)]\sum_{j=0}^{2}f_{ji}p_{ji}\\&-\sum_{i=1}^{K}w_i\sum_{j=0}^{2}f_{ji}p_{ji}\sum_{i=1}^{K}[q_i^2\\&+2xq_i(1-q_i)]\\&+(1-2x)\sum_{i=1}^{K}w_i\delta_i^2(f_{2i}+f_{0i}-2f_{1i})\\&+2\sum_{i=1}^{K}w_i\delta_i[q_i\\&+x(1\\&-2q_i)][p_i(f_{2i}\\&-f_{1i})+(1\\&-p_i)(f_{1i}-f_{0i})].\end{aligned}$$

Under the null hypothesis, this expectation is equal to $\mathrm{Cov}(\sum_{j=0}^{2}p_{jZ}f_{jZ},q_Z^2+2xq_Z(1-q_Z))$, where Z is the random variable defined below (12). In the absence of PS, the random variable Z becomes non-random, making the null expectation 0. Otherwise, the expectation is nonzero in general. Furthermore, under the assumptions in Remark 2, the expression Δ_x reduces to that given by Zheng et al. (2009).

For notational simplicity, we use $g_{21}=P(MM|Y=1)$, $g_{11}=P(Mm|Y=1)$, $g_{22}=P(MM|Y=2)$ and $g_{12}=P(Mm|Y=2)$ for the expressions given by (6)–(9). Using the facts that (D_2, D_1, D_0) and (C_2,C_1,C_0) follow trinomial distributions, we obtain the following formula

$$\begin{aligned}\sigma_x^2=&\mathrm{Var}[(\tfrac{D_2}{n_1}-\tfrac{C_2}{n_2})+x(\tfrac{D_1}{n_1}-\tfrac{C_1}{n_2})]\\=&\tfrac{1}{n_1}[g_{21}(1-g_{21})+x^2g_{11}(1-g_{11})-2xg_{11}g_{21}]+\tfrac{1}{n_2}[g_{22}(1-g_{22})+x^2g_{12}(1-g_{12})-2xg_{12}g_{22}].\end{aligned}$$

Replacing the g_{ij} by their consistent estimators, we get the following estimate of the variance of $(D_2/n_1-C_2/n_2)+x(D_1/n_1-C_1/n_2)$:

$$V_x=\frac{1}{n_1}[\frac{D_2}{n_1}(1-\frac{D_2}{n_1})+x^2\frac{D_1}{n_1}(1-\frac{D_1}{n_1})-2x\frac{D_1}{n_1}\frac{D_2}{n_1}]+\frac{1}{n_2}[\frac{C_2}{n_2}(1-\frac{C_2}{n_2})+x^2\frac{C_1}{n_2}(1-\frac{C_1}{n_2})-2x\frac{C_1}{n_2}\frac{C_2}{n_2}]. \quad (25)$$

By the Law of Large Numbers, V_x is a consistent estimate under both the null hypothesis and the alternative hypothesis, even if HWE does not hold in any subpopulation. When $x=0.5$, we have that $V_{0.5}=V^*$ and $T_{0.5}=T^*$. This shows that V^* is a consistent estimate of the variance of $\hat{q}_D-\hat{q}_C$.

The asymptotic power function of the T_x test with rejection region $\{|T_x|>u_{\alpha/2}\}$ is

$$1-\Phi(u_{\alpha/2}-\Delta_x/\sigma_x)+\Phi(-u_{\alpha/2}-\Delta_x/\sigma_x). \quad (26)$$

A SIMULATION STUDY

To study the finite sample performance of the mentioned tests, we conducted some simulations. We studied the impact of PS on the powers and type I error rates of the T test (defined in (1)) and the T^* test (defined in (23)). In the simulations, we assumed the study population consisted of 2 subpopulations of equal sizes.

First we considered an additive mode of inheritance in each subpopulation. The underlying models were specified as follows. The allele frequencies of M at a disease locus were 0.2 and 0.2 (PS is absent) or 0.1 and 0.3 (PS is present) for the two subpopulations. The allele frequencies of M at a marker locus were 0.3 and 0.3 (PS is absent) or 0.4 and 0.2 (PS is present) for the two subpopulations. HWE was assume to hold in the 2 subpopulations at both the marker and the disease loci. The penetrances of genotypes *aa*, *Aa* and *AA* in subpopulation 1 were 0.1, 0.2 and 0.3, respectively, and they were either 0.2, 0.3, 0.4 for subpopulation 2 (PH is present) or the same as those in subpopulation 1 (PH is absent). The LD measures in the two subpopulations were the same, that is, either 0 (null hypothesis) or 0.05 (alternative hypothesis).

We randomly generated the genotypes of 1000 cases and 1000 controls. The empirical type I error rates/powers at a 0.05 level of significance were estimated based on 5,000,000 replications. The asymptotic type I error rates/powers of the T^* test were calculated using formula (24). The resulting powers are presented in Table 3.

For the T^* test, it is seen that the asymptotic type I error rates/powers and the empirical type I error rates/powers are very close to each other, with differences of no more than 0.001, showing an accurate approximation of the power function. As expected, when PS is absent, the type I errors are virtually equal to the nominal level 0.05; when PS is present, the type I error could be inflated a great deal, especially when PH is also present (0.876). The power is also influenced by the presence of PS and PH. For example, when both PS and PH are present, the power is only 0.142, compared with 0.809 for the case where neither PS nor PH is present.

The T test has type I error rates/powers close to those of the T^* test in the absence of PS, with differences of no more than 0.001. In the presence of PS, there are minor differences that vary from 0.009 to 0.017.

The above simulations assumed that HWE held in any subpopulation. Our further simulations without an assumption of HWE (results not shown) showed that the T^* test had type I error rates close to the nominal levels in the presence of PS, but the T test could distort the type I error rate, with its magnitude depending upon the strength of Hardy-Weinberg disequilibrium.

Second we considered a fully penetrance recessive model in each subpopulation, with the penetrance being 1 for *AA* and 0 for *Aa* and *aa*. The other parameters are the same as those in Table 3, except that the LD measures under the null hypothesis are 0.01. The simulation results are reported in Table 4. For this mode of inheritance, the impact of PS on the type I error rates and the powers has a trend the same as that for the additive mode of inheritance.

Third we considered a special case, where the the marker locus and the disease locus coincide, with a common allele frequency p_i for the i th subpopulation, i = 1,2, and where the LD measures are $\delta_i = p_i - p_i^2$. The other parameters are the same as those in Table 3 except that the null hypothesis (in each subpopulation the penetrances are independent of the genotypes) is different and the mode of inheritance under the alternative hypothesis is recessive. The detailed parameter settings are described and the simulation results are

presented in Table 5. As expected, the T^* test has type I error rates controlled at the nominal level when either PS or PH is absent (this is different from the marker locus case), but the type I error rate is inflated when both PS and PH are present. Furthermore, the presence of PS and/or PH also has an impact on the powers of the T^* test, with the trend similar to that for the marker locus case. There are only minor differences between the T test and the T^* test, except under the null hypothesis with the absence of PS.

DISCUSSION

Population-based marker-disease association analysis is a powerful tool but may suffer from PS. Our work provides closed forms for the expectation and variance of two commonly used test statistics, which enable us to study the type I error rate and power under various scenarios. We extend the work of Ewens and Spielman (1995) by relaxing the assumption of fully recessive penetrance and studying bias and variance distortion simultaneously. Our simulation results are in agreement with those from asymptotic approximations, confirming that the theoretical findings are correct. Both analysis and simulation results show that the presence of PS can inflate the type I error rate and decrease the power dramatically in the marker-disease association analysis. Therefore, it is necessary to modify the test statistics to accommodate PS. Methods have been proposed in the literature for correcting bias and/or variance distortion in candidate gene-disease association analysis, including genomic control (Devlin and Roeder, 1999; Devlin et al., 2001), structured association (Pritchard et al., 2000; Satten et al., 2001; Pritchard and Donnelly, 2001), the delta centralization (Gorroochurn et al., 2006). Whittermore (2006) suggested sensitivity analysis. However, the performance qualify for these methods for marker-disease association analysis is unclear and needs further investigations.

Acknowledgments

We would like to thank the Managing Editor, the Handling Editor and two reviewers for their helpful comments and suggestions leading to an improvement of the paper. We are grateful to Dr. B. J. Stone for editorial help. This research was supported in part by the National Natural Science Foundation of China 10701067 (HZ), the Outstanding Overseas Chinese Scholars Fund of Chinese Academy of Sciences (ZL), and NIH grant 5R37GM047845 (ZY).

References

Armitage P. Tests for linear trends in proportions and frequencies. Biometrics. 1955; 11:375–386.

Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN. Demonstrating stratification in a European American population. Nat Genet. 2005; 37:868–872. [PubMed: 16041375]

Cardon LR, Palmer LJ. Population stratification and spurious allelic association. Lancet. 2003; 361:598–604. [PubMed: 12598158]

Ewens WJ, Spielman RS. The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet. 1995; 57:455–464. [PubMed: 7668272]

Gorroochurn P, Heiman GA, Hodge SE, Greenberg DA. Centralizing the non-central chi-square: A new method to correct for population stratification in genetic case-control association studies. Genet Epi. 2006; 30:277–289.

Gorroochurn P, Hodge SE, Heiman G, Greenberg DA. Effect of population stratification on case-control association studies. II. False-positive rates and their limiting behavior as number of subpopulations increases. Hum Hered. 2004; 58:40–48. [PubMed: 15604563]

Heiman GA, Hodge SE, Gorroochurn P, Zhang J, Greenberg DA. Effect of population stratification on case-control association studies. Hum Hered. 2004; 58:30–39. [PubMed: 15604562]

Knowler WC, Williams RC, Pettitt DJ, Steinberg AG. Gm $^{3;5,13,14}$ and type 2 diabetes mellitus: an association in American Indians with genetic admixture. Am J Hum Genet. 1988; 43:520–526. [PubMed: 3177389]

Lander ES, Schork NJ. Genetic dissection of complex traits. Science. 1994; 265:2037–2048. [PubMed: 8091226]

Li CC. Population subdivision with respect to multiple alleles. Ann Hum Genet. 1969; 33:23–29. [PubMed: 5821316]

Li CC. Genetics of subdivided populations and its relationships with certain measures of association. Genet Epi. 1991; 8:1–11.

Li Z, Zhang H, Zheng G, Gastwirth JL, Gail MH. Excess false positive rate caused by population stratification and disease rate heterogeneity in case-control association studies. Comput Statist Data Anal. 2009; 53:1767–1781.

Qin, H.; Zhang, H.; Li, Z. The impact of population stratification on commonly used statistical procedures in population-based QTL association studies. In: Hsiung, A.; Zhang, C.; Ying, Z., editors. Random Walk, Sequential Analysis and Related Topics-A Festschrift in Honor of Yuan-Shih Chow. Singapore: World Scientific Publisher; 2006. p. 311-333.

Risch N. Searching for genetic determinants in the new millennium. Nature. 2000; 405:847–856. [PubMed: 10866211]

Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996; 273:1516–1517. [PubMed: 8801636]

Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet. 1993; 52(3):506–516. [PubMed: 8447318]

Whittemore, AS. Population structure in genetic association studies. Proceedings of the American Statistical Association, Statistics in Epidemiology Section [CD-ROM]; Alexandria, VA: ASA; 2006.

Zheng G, Li Z, Gail MH, Gastwirth JL. Impact of population substructure on trend tests for genetic case-control association studies. Biometrics. 200910.1111/j.1541–0420.2009.01264.x

APPENDIX I. Proof of (6) and (7)

By the definition of the linkage equilibrium measures δ_i, the probabilities of gametes *MA*, *Ma*, *mA* and *ma* are

$$P(MA|S_i)=p_iq_i+\delta_i \tag{27}$$

$$P(Ma|S_i)=(1-p_i)q_i-\delta_i, \tag{28}$$

$$P(mA|S_i)=(1-q_i)p_i-\delta_i \tag{29}$$

and

$$P(ma|S_i)=(1-q_i)(1-p_i)+\delta_i, \tag{30}$$

respectively. Random mating gives

$$P((MM,AA)|S_i)=[P(MA|S_i)]^2=[p_iq_i+\delta_i]^2, \tag{31}$$

$$P((MM,Aa)|S_i)=2[p_iq_i+\delta_i][(1-p_i)q_i-\delta_i], \tag{32}$$

$$P((MM,aa)|S_i)=[(1-p_i)q_i-\delta_i]^2, \tag{33}$$

$$P((Mm,AA)|S_i)=2[p_iq_i+\delta_i][p_i(1-q_i)-\delta_i], \tag{34}$$

$$P((Mm,Aa)|S_i)=2[p_iq_i+\delta_i][(1-p_i)(1-q_i)+\delta_i]+2[(1-p_i)q_i-\delta_i][p_i(1-q_i)-\delta_i], \tag{35}$$

$$P((Mm,aa)|S_i)=2[(1-p_i)q_i-\delta_i][(1-p_i)(1-q_i)+\delta_i]. \tag{36}$$

Here (*MM*, *AA*) is the joint genotype at the marker locus (*MM*) and the disease locus (*AA*), so are the other 5 pairs. It follows from (31)–(36) that

$$P(MM|Y=1)=\frac{\sum_{i=1}^{K}w_i[P((MM,AA)|S_i)f_{2i}+P((MM,Aa)|S_i)f_{1i}+P((MM,aa)|S_i)f_{0i}]}{\sum_{i=1}^{K}w_i\sum_{j=0}^{2}f_{ji}p_{ji}}$$

$$=\frac{1}{B}\{\sum_{i=1}^{K}w_iq_i^2\sum_{j=0}^{2}f_{ji}p_{ji}+\sum_{i=1}^{K}w_i\delta_i^2(f_{2i}+f_{0i}-2f_{1i})+2\sum_{i=1}^{K}w_iq_i\delta_i[p_i(f_{2i}-f_{1i})+(1-p_i)(f_{1i}-f_{0i})]\}$$

and

$$P(Mm|Y=1)=\frac{\sum_{i=1}^{K}w_i[P((Mm,AA)|S_i)f_{2i}+P((Mm,Aa)|S_i)f_{1i}+P((Mm,aa)|S_i)f_{0i}]}{\sum_{i=1}^{K}w_i\sum_{j=0}^{2}f_{ji}p_{ji}}$$

$$=\frac{2}{B}\{\sum_{i=1}^{K}q_i(1-q_i)\sum_{j=0}^{2}f_{ji}p_{ji}-\sum_{i=1}^{K}w_i\delta_i^2(f_{2i}+f_{0i}-2f_{1i})+\sum_{i=1}^{K}w_i(1-2q_i)\delta_i[p_i(f_{2i}-f_{1i})+(1-p_i)(f_{1i}-f_{0i})]\}.$$

APPENDIX II. Proof of (14)

Define two indicator functions

$$I_{MM}=\begin{cases}1, & \text{if randomly selected marker genotype of a case is } MM;\\ 0, & \text{otherwise.}\end{cases}$$
$$I_{Mm}=\begin{cases}1, & \text{if randomly selected marker genotype of a case is } Mm;\\ 0, & \text{otherwise.}\end{cases}$$

Then,

$$\widehat{q}_D=\frac{\sum(2I_{MM}+I_{Mm})}{2n_1},$$

where the summation is taken over all cases. Since

$$\begin{aligned}\mathrm{Var}(I_{MM})&=P(MM|Y=1)(1-P(MM|Y=1)),\\ \mathrm{Var}(I_{Mm})&=P(Mm|Y=1)(1-P(Mm|Y=1)),\\ \mathrm{Cov}(I_{MM},I_{Mm})&=-P(MM|Y=1)P(Mm|Y=1),\end{aligned}$$

the variance of $\hat{q}_D$ is

$$\begin{aligned}\mathrm{Var}(\widehat{q}_D)&=\mathrm{Var}\left(\frac{\sum(2I_{MM}+I_{Mm})}{2n_1}\right)\\ &=\frac{1}{(2n_1)^2}[4n_1\mathrm{Var}(I_{MM})+n_1\mathrm{Var}(I_{Mm})+4n_1\mathrm{Cov}(I_{MM},I_{Mm})]\\ &=\frac{1}{4n_1}[4P(MM|Y=1)(1-P(MM|Y=1))+P(Mm|Y=1)(1-P(Mm|Y=1))-4P(MM|Y=1)P(Mm|Y=1)].\end{aligned}$$

APPENDIX III. Proof of (16)

Define $c_1 \equiv \sum_i q_i^2\alpha_i$, $c_2 \equiv \Sigma_i 2q_i(1-q_i)\alpha_i$, $x_1 \equiv P(MM|Y=1)-c_1$, $x_2 \equiv P(MM|Y=1)-c_2$. Substituting c_1, c_2, x_1 and x_2 into (14), we have

$$\mathrm{Var}(\widehat{q}_D)=\frac{\widehat{q}_D(1-\widehat{q}_D)+\sigma_D^2}{2n_1}-\frac{(2x_1+x_2)^2}{4n_1}+\frac{4(1-2c_1-c_2)x_1+(1-2c_2-4c_1)x_2}{4n_1} \tag{37}$$

By the definitions,

$$2x_1+x_2=\frac{2}{B}\sum_i \delta_i w_i[p_i(f_{2i}-f_{1i})+(1-p_i)(f_{1i}-f_{0i})], \tag{38}$$

$$2c_1+c_2=\sum_i[2q_i^2\alpha_i+2q_i(1-q_i)\alpha_i]=2\overline{q}_D, \tag{39}$$

$$4x_1+x_2=\frac{2}{B}\sum_i \delta_i w_i[p_i(f_{2i}-f_{1i})+(1-p_i)(f_{1i}-f_{0i})](1+2q_i)+\frac{2}{B}\sum_i w_i\delta_i^2(f_{2i}+f_{0i}-2f_{1i}). \quad (40)$$

It follows from (39) and (40) that

$$\begin{aligned}
&\frac{4(1-2c_1-c_2)x_1+(1-2c_2-4c_1)x_2}{4n_1}\\
&\quad=\frac{4(1-2\bar{q}_D)x_1+(1-4\bar{q}_D)x_2}{4n_1}\\
&\quad=\frac{4x_1+x_2-4(2x_1+x_2)\bar{q}_D}{4n_1}\\
&\quad=\frac{1}{2n_1B}\sum_i \delta_i w_i[p_i(f_{2i}-f_{1i})+(1-p_i)(f_{1i}-f_{0i})](1\\
&\quad+2q_i\\
&\quad-4\bar{q}_D)\\
&\quad+\frac{1}{2n_1B}\sum_i w_i\delta_i^2(f_{2i}+f_{0i}-2f_{1i}).
\end{aligned} \quad (41)$$

Equation (16) follows from (37), (38) and (41).

Table 1

Genotype counts

	MM	***Mm***	***mm***	**Sum**
Cases	D_2	D_1	D_0	n_1
Controls	C_2	C_1	C_0	n_2
Sum	r_2	r_1	r_0	n

Table 2

Allele counts

	M	***m***	**Sum**
Cases	$2D_2+D_1$	$2D_0+D_1$	$2n_1$
Controls	$2C_2+C_1$	$2C_0+C_1$	$2n_2$
Sum	$2r_2+r_1$	$2r_0+r_1$	$2n$

Table 3

Type I error rates/powers for marker locus under additive mode of inheritance

			T test	*T** test	
Hypothesis [1]	**PS** [2]	**PH** [3]	**Empirical**	**Asymptotic**	**Empirical**
Null	Absent	Absent	0.050	0.050	0.050
Null	Absent	Present	0.050	0.050	0.050
Null	Present	Absent	0.217	0.203	0.204
Null	Present	Present	0.885	0.876	0.876
Alternative	Absent	Absent	0.808	0.809	0.809
Alternative	Absent	Present	0.604	0.604	0.605
Alternative	Present	Absent	0.402	0.385	0.386
Alternative	Present	Present	0.153	0.142	0.143

[1]"Null": $\delta_1 = \delta_2 = 0$; "Alternative": $\delta_1 = \delta_2 = 0.05$.

[2]"Absent": $p_1 = p_2 = 0.2$ and $m_1 = m_2 = 0.3$; "Present": $p_1 = 0.1$, $p_2 = 0.3$ and $m_1 = 0.4$, $m_2 = 0.2$.

[3]"Absent": the penetrances of genotypes *aa*, *Aa*, *AA* are 0.1,0.2 and 0.3, respectively, in both of the subpopulations; "Present": the penetrances are 0.1, 0.2 and 0.3 in subpopulation 1 and 0.2, 0.3 and 0.4 in subpopulation 2.

Table 4

Type I error rates/powers for marker locus under fully recessive mode of inheritance[1]

		T test	*T** test	
Hypothesis[2]	**PS**[3]	**Empirical**	**Asymptotic**	**Empirical**
Null	Absent	0.050	0.050	0.050
Null	Present	1.000	1.000	1.000
Alternative	Absent	0.941	0.941	0.941
Alternative	Present	0.838	0.828	0.828

[1] In both of the subpopulation, the penetrances of genotypes *aa*, *Aa* and *AA* are 0, 0 and 1, respectively.

[2] "Null": $\delta_1 = \delta_2 = 0$; "Alternative": $\delta_1 = \delta_2 = 0.01$.

[3] "Absent": $p_1 = p_2 = 0.2$ and $m_1 = m_2 = 0.3$; "Present": $p_1 = 0.1$, $p_2 = 0.3$ and $m_1 = 0.4$, $m_2 = 0.2$.

Table 5

Type I error rates/powers for candidate locus [1]

			T test	*T** test	
Hypothesis [2]	**PS [3]**	**PH [4]**	**Empirical**	**Asymptotic**	**Empirical**
Null	Absent	Absent	0.050	0.050	0.050
Null	Absent	Present	0.050	0.050	0.050
Null	Present	Absent	0.057	0.050	0.050
Null	Present	Present	0.850	0.838	0.838
Alternative	Absent	Absent	0.749	0.730	0.731
Alternative	Absent	Present	0.482	0.467	0.467
Alternative	Present	Absent	0.891	0.867	0.868
Alternative	Present	Present	0.108	0.089	0.090

[1]The marker locus and the disease locus coincide, and the LD measure δ_i in the *i* th subpopulation is $p_i - p_i^2$ with p_i being the allele frequency of both marker and disease loci in the *i* th subpopulation.

[2]"Null": the penetrances of the genotypes *aa*, *Aa* and *AA* are the same in each subpopulation; "Alternative": the penetrances are different for the genotypes *aa*, *Aa* and *AA* in each subpopulation.

[3]"Absent": $p_1 = p_2 = 0.2$; "Present": $p_1 = 0.3$, $p_2 = 0.1$.

[4]"Absent": under the null hypothesis, the penetrances of genotypes *aa*, *Aa* and *AA* are 0.2 in each subpopulation, and under the alternative hypothesis, the penetrances of genotypes *aa*, *Aa* and *AA* are 0.1, 0.1 and 0.2, respectively, in both of the two subpopulations; "Present": under the null hypothesis, the penetrances of genotypes *aa*, *Aa* and *AA* are 0.2 in subpopulation 1 and 0.1 in subpopulation 2, and under the alternative hypothesis, the penetrances of genotypes *aa*, *Aa* and *AA* are 0.1, 0.1 and 0.2 in subpopulation 1 and 0.2, 0.2 and 0.3 in subpopulation 2.