Testing for genetic association in the presence of population stratification in genome-wide association studies

Running title: Testing for genetic association

KAI WANG

Department of Biostatistics

College of Public Health

The University of Iowa

Iowa City, Iowa

Corresponding author:

Kai Wang, PhD

Department of Biostatistics, C227 GH

College of Public Health

University of Iowa

Iowa City, IA 52242

e-mail: kai-wang@uiowa.edu

phone:(319) 384-5175

fax: (319) 384-5018

Genome-wide case-control association study is gaining popularity, thanks to the rapid development of modern genotyping technology. In such studies, population stratification is a potential concern especially when the number of study subjects is large as it can lead to seriously inflated false positive rates. Current methods addressing this issue are still not completely immune to excess false positives. A simple method that corrects for population stratification is proposed. This method modifies a test statistic such as the Armitage trend test by using an additive constant that measures the variation of the effect size confounded by population stratification across genomic control markers. As a result, the original statistic is deflated by a multiplying factor that is specific to the marker being tested for association. This deflating multiplying factor is guaranteed to be larger than 1. These properties are in contrast to the conventional genomic control method where the original statistic is deflated by a common factor regardless of the marker being tested and the deflation factor may turn out to be less than 1. The new method is introduced first for regular case-control design and then for other situations such as quantitative traits and the presence of covariates. Extensive simulation study indicates that this new method provides an appealing alternative for genetic association analysis in the presence of population stratification.

Keywords: genetic association, population stratification, genomic control, variance inflation factor, SNP

INTRODUCTION

With the rapid development of large-scale high-throughput genotyping technology, genome-wide association study is becoming more and more popular in the mapping of genetic variants underlying complex human disorders. However, there are some concerns regarding regular association methods [McCarthy et al., 2008; Lunetta, 2008]. A major one is that, in the presence of population stratification, they tend to generate excess false positives than expected. Population stratification may be a phenomenon more prevalent than it appears to be [Epstein et al., 2007]. This issue has been discussed extensively in the literature and many methods have been proposed to address it [Horvath & Laird, 1998; Devlin & Roeder, 1999; Bacanu et al., 2000; Pritchard & Rosenberg, 1999; Pritchard et al., 2000; Price et al., 2006; Kimmel et al., 2007; Zhu et al., 2008].

A popular method for handling population stratification is genomic control (GC) [Devlin & Roeder, 1999; Bacanu et al., 2000; Devlin et al., 2001; Bacanu et al., 2002]. This method modifies a test statistic, for instance, the Armitage test for trend [Armitage, 1955], by a multiplying factor. This factor is estimated using markers that are believed to be not in association with the phenotype. The structured population method tries to first infer the subpopulation structure underlying the sample. Subsequent analyses are conducted within each subpopulation and the results are summarized [Pritchard & Rosenberg, 1999; Pritchard et al., 2000; Pritchard & Donnelly, 2001]. The third method is to create surrogates for population stratification using markers that are not in association with the phenotype. For example, the principal component method [Price et al., 2006] uses the first few principal components based on the correlation matrix of marker genotypes as covariates in subsequent regression analysis. Following the same idea, one can use the first few components from a partial least-squares regression [Epstein et al., 2007].

There are some drawbacks to these methods. The genomic control intends to deflate the test statistic in order to reduce the false positive rates due to population stratification. However, the multiplying correction factor may not be always less than 1 [Bacanu et al., 2002]. In addition, the same multiplying factor is used regardless of the testing position. The result from structured association method depends on the assumed number of subpopulations. It is also computation intensive. Finally, despite its increasingly popularity, the principal

component method is inherently unable to completely account for the impact of population stratification (appendix A). It has been documented that the principal component method can fail to control false positive rate [Kimmel et al., 2007; Epstein et al., 2007] although it is of less concern when the number of markers used to extract principal components is large [Lee et al., 2008]. So does the partial least-squares regression method [Lee et al., 2008].

In this study, I propose a new method for association study that corrects for the effect of population stratification. This method is similar to GC in the sense that it also results in a multiplying factor to a statistic such as Armitage test for trend. However, contrary to the GC method, this multiplying factor is no longer constant. Instead, it changes as the marker being tested at changes. In addition, this factor is guaranteed to deflate the test statistic.

In what follows, I first describe the proposed method for association study of case-control data. This method is then generalized to situations of continuous phenotype with and without covariates. The performance of this method is evaluated through extensive simulation studies.

METHOD

In a case-control study, the genotype counts in cases and controls at a biallelic marker can be summarized in a 2×3 table. The notations for various genotype counts are presented in table 1. Association to this marker can be tested using a two-degrees of freedom Pearson chi-square test on the 2×3 table. An alternative popular method is the Armitage test for trend [Armitage, 1955] that possesses one degree of freedom. The Armitage trend test assumes that the impact of the disease allele on odds ratio is multiplicative. Due to its less degree of freedom, the Armitage test for trend could be more powerful than the Pearson chi-square test. Using notations in table 1, the Armitage test for trend can be expressed as [Sasieni, 1997]

$$X_G^2 = \frac{N\{N(r_1 + 2r_2) - R(n_1 + 2n_2)\}^2}{R(N-R)\{N(n_1 + 4n_2) - (n_1 + 2n_2)^2\}},$$

which equals

$$X_G^2 = \frac{NT^2}{R(N-R)\{N(n_1 + 4n_2) - (n_1 + 2n_2)^2\}}$$

with $T = (r_1 + 2r_2)S - (s_1 + 2s_2)R$. Note that $T \propto (r_1 + 2r_2)/2R - (s_1 + 2s_2)/2S$. So T measures the difference in allele, say, A frequencies between cases and controls.

It is well known that, in the presence of population stratification, regular methods for testing association, such as the Pearson chi-square method and the Armitage test for trend, can generate excess false positives compared to the nominal significance levels. As summarized in the Introduction, numerous methods have been proposed to tackle this issue. The method proposed here may be regarded as one similar to GC but its motivation and approach are quite different.

The idea behind GC is simple. It can be illustrated using the statistic X_G^2. Since X_G^2 leads to excessive false positive rate, the true variance of T is inflated but is not accounted for by X_G^2. The variance inflation factor (VIF), denoted by λ, is estimated through genomic control markers as follows. By definition, these GC markers are not associated with the affection status. The X_G^2 statistics computed at these markers are expected to be random samples from the chi-square distribution with one degree of freedom. However, their actual distribution is not because of population stratification. One modification is to align the median of these statistics with that of the chi-square distribution with one degree of freedom, which is about 0.456. A robust estimate of λ is [Devlin et al., 2001] $\hat{\lambda}$ = the median of values of statistic X_G^2 at GC markers divided by 0.456. At the testing location, the statistic $X_G^2/\hat{\lambda}$, instead of X_G^2, is used and is compared to the chi-square distribution with one degree of freedom. As a VIF, λ is expected to be larger than one. Although this is often the case, it is possible that $\hat{\lambda}$ is less than 1 [Devlin & Roeder, 1999, page 999, last line of left column]. The chance is higher when the extent of population stratification is less. In the extreme, the probability that $\hat{\lambda}$ is less than 1 will be 0.5 in the absence of population stratification. It has been suggested that $\hat{\lambda}$ is set to 1 whenever its afore-mentioned estimate of VIF is less than 1 [Bacanu et al., 2000]. Note that in GC, the same VIF estimate is applied regardless of testing locations.

In the Armitage trend test, the variance of T, denoted by σ_T^2, is computed as

$$\sigma_T^2 = R(1 - R/N)[N(n_1 + 4n_2) - (n_1 + 2n_2)^2].$$

This variance is correct in the absence of population stratification. When there is population stratification, it is smaller than the true variance of T [Devlin & Roeder, 1999] which explains why the test statistic is inflated. In the context of genetic association testing, this variance

is conditional in the sense that the variance of T is for the marker being tested. On the other hand, the variance of T for a randomly picked marker is

$$\begin{aligned} Var(T) &= E[Var(T|\text{marker } l)] + Var[E(T|\text{marker } l)] \\ &= E[Var(T|\text{marker } l)] + \Lambda, \end{aligned}$$

where $\Lambda = Var[E(T|\text{marker } l)]$. The variance $Var(T)$ is unconditional since the marker is randomly chosen. The quantity Λ is the variance of conditional mean $E(T|\text{marker } l)]$ across all the markers. It can be estimated by the sample variance of $\{T_l\}$, where T_l is the quantity T at marker l. In the next section, robust estimation of Λ will be discussed. An estimate of Λ is denoted by $\hat{\Lambda}$.

The above consideration motivates the following modification of the Armitage trend test:

$$X^2 := \frac{T^2}{\sigma_T^2 + \hat{\Lambda}}.$$

That is, σ_T^2 in the Armitage trend test is substituted by $\sigma_T^2 + \hat{\Lambda}$. Compared to the Armitage trend test statistic $X_G^2 = T^2/\sigma_T^2$, there is

$$\begin{aligned} X^2 &= X_G^2 \cdot \frac{\sigma_T^2}{\sigma_T^2 + \hat{\Lambda}} \\ &= \frac{X_G^2}{1 + \hat{\Lambda}/\sigma_T^2}. \end{aligned}$$

So the proposed method deflates X_G^2 with a factor $1 + \hat{\Lambda}/\sigma_T^2$. This factor is always larger than one. Instead of being a constant, its size depends on σ_T^2: the larger the value of σ_T^2, the smaller its value. Since σ_T^2 equals $R(1 - R/N)[n_1(N - n_1) + 4n_0n_2]$, given n_1, σ_T^2 is larger for n_0 and n_2 that are closer to each other.

ROBUST ESTIMATION OF Λ

As mentioned in the previous section, a natural estimate of $\Lambda = Var[E(T|p_l)]$ is the sample variance of $\{T_l\}$ across all genome control markers. However, this estimate is affected by sampling error. Consider an extreme case where the allele frequencies are the same across all genome control markers such that $E(T|\text{location } l)$ is a constant. The value of Λ is 0 but the sample variance of $\{T_l\}$ is not unless T_ls are the same at all genome control markers.

Another disadvantage of using sample variance is that it is sensitive to extreme values in $\{T_l\}$. Presence of extreme values would seriously bias Λ estimate upward and jeopardize the power of the proposed statistic X^2. For these reasons, a robust estimate of Λ is necessary. Of particular interest are those that are smaller than the sample variance of $\{T_l\}$.

Two candidate estimates considered here are the α-trimmed variance and the α-winsorized variance of $\{T_l\}$ [Huber, 1981]. For the α-trimmed variance, the top $100\alpha\%$ and the lower $100\alpha\%$ values of $\{T_l\}$ are trimmed and the α-trimmed variance is set to equal the sample variance of the remaining values. That is,

$$\hat{\Lambda}_{\text{trim},\alpha} := \text{sample variance of } T_{([\alpha L]+1)}, \ldots, T_{(L-[\alpha L])},$$

where $T_{(l)}$ is the lth ordered statistic of $\{T_l\}$, and the notation $[x]$ means the integer part of x. Here I did not use the version of α-trimmed variance that uses a normalizing factor in order to avoid an estimate possibly larger than the usual sample variance.

For the α-winsorized variance, all the T_l values in the upper $100\alpha\%$ tail are replaced by the next largest one and all the values in the lower $100\alpha\%$ tail are replaced by the next smallest one. Specifically, the α-winsorized variance is defined as $\hat{\Lambda}_{\text{win},\alpha}$ = sample variance of $T_{(k+1)}$, $\ldots, T_{(k+1)}, T_{(k+2)}, \ldots, T_{(L-k-1)}, T_{(L-k)}, \ldots, T_{(L-k)}$, where $k = [\alpha L]$. Here the value $T_{(k+1)}$ is repeated $k+1$ times, so is the value $T_{(L-k)}$. In formula, it can be expressed as

$$\hat{\Lambda}_{\text{win},\alpha} = (L-1)^{-1}\left[(k+1)(T_{(k+1)} - \bar{T})^2 + \sum_{l=k+2}^{L-k-1}(T_{(l)} - \bar{T})^2 + (k+1)(T_{(L-k)} - \bar{T})^2\right],$$

where $\bar{T}$ is the winsorized mean which equals

$$\bar{T} = L^{-1}\left[(k+1)T_{(k+1)} + \sum_{l=k+2}^{L-k-1} T_{(l)} + (k+1)T_{(L-k)}\right].$$

It is not clear as to whether there exists a generally applicable level α of trimming or winsorization. It seems to depend on the extent of changes in the difference in genotype counts between cases and controls along the genome. It is also not clear which method is theoretically better. In the simulation to be reported later, impact of different level of α and the two robust variance estimation methods are investigated.

QUANTITATIVE TRAITS AND MORE GENERAL MODELS

Let y denote a quantitative phenotype. For the ease of exposition, assume that a simple regression analysis over genotype score is justified. Let G denote the genotype dose at a biallelic marker being tested for association. That is, $G = 0, 1$ or 2 if there are 0, 1 or 2 copies of a designated reference allele. Denote the regression model by

$$E(y) = a + bG. \tag{1}$$

In a regular association analysis, one would expect b to be 0 in the absence of association. Association can be tested by using the regular t test:

$$t = \frac{\hat{b}}{s_b},$$

where $\hat{b}$ is the least-squares estimate of b and s_b is the standard deviation of $\hat{b}$. However, in the presence of population association, this t test is liberal [Bacanu et al., 2000], similar to the situation of case-control data.

To correct for population stratification, one can compute $\hat{b}$ at GC markers and obtain the sample variance of these $\hat{b}$s. Denote this sample variance by $\hat{\Lambda}$. At the marker being tested for association, the modified t statistic would be

$$\frac{\hat{b}}{(s_b^2 + \hat{\Lambda})^{1/2}}.$$

For moderate or large sample size, refer this statistic to the standard normal distribution. The robust variance estimation methods discussed in the previous section apply to $\hat{\Lambda}$.

Covariates can be easily incorporated into regression model (1) as well.

SIMULATION

The data are simulated in almost the same way as in [Devlin & Roeder, 1999], which is often used in simulation studies [Price et al., 2006; Devlin et al., 2001]. Specifically, an ancestral allele frequency p is randomly sampled from interval [0.1, 0.9]. Denote the Wright's coefficient of inbreeding F_{st} in cases by F_1 and in controls by F_2. The allele frequency p_1 in cases is sampled from a beta distribution with parameters $p(1-F_1)/F_1$ and $(1-p)(1-F_1)/F_1$ and the allele frequency p_2 in controls is sampled from a beta distribution with parameters

$p(1-F_2)/F_2$ and $(1-p)(1-F_2)/F_2$. Instead of assuming HWE [Price et al., 2006; Devlin & Roeder, 1999; Devlin et al., 2001], the genotype frequencies in the case population are taken to be $F_1(1-p_1)+(1-F_1)(1-p_1)^2$, $2(1-F_1)p_1(1-p_1)$, $F_1p_1+(1-F_1)p_1^2$ for having 0, 1 and 2 copies of the disease allele. Genotype frequencies in the control population are determined in the same manner for the values of F_2 and p_2. Particularly, HWE is not assumed. This process is repeated for every SNP except the disease SNP.

The genotypes of the disease SNP in the control population are simulated in the same manner as described in the previous paragraph. It is possible that the controls carry the disease allele. In the cases, the disease genotype is simulated in the following manner. Let γ be the relative risk of carrying one copy of the disease allele to carrying 0 copy. Under a multiplicative model, the relative risk for carrying two copies of the disease allele to carrying 0 copy of the disease allele would be γ^2. The frequencies of the genotypes at the disease locus in cases are taken to be proportional to $F_1(1-p_1)+(1-F_1)(1-p_1)^2$, $2(1-F_1)p_1(1-p_1)\gamma$, and $F_1p_1+(1-F_1)p_1^2\gamma^2$, respectively, for carrying 0, 1 and 2 copies of the disease allele.

In the simulation study, the number of GC SNPs is taken to be 1000. The number of cases is set to 100 and so is the number of controls. Simulations are conducted with varying inbreeding coefficient F_{st}s (F_1 in cases and F_2 in controls) and genotype relative risk γ. For each parameter configuration, 1000 genomic control SNPs are simulated first from which the VIF for the genomic control method, the variance estimate $\hat{\Lambda}$ for the proposed method, and the principal components for the principal components (PC) method are computed. Another 1000 SNPs are then generated under the specified parameter configuration for which the following statistics are computed: $X^2_{\text{trim},\alpha=0.01}$, $X^2_{\text{trim},\alpha=0.05}$, $X^2_{\text{trim},\alpha=0.1}$, $X^2_{\text{win},\alpha=0.01}$, $X^2_{\text{win},\alpha=0.05}$, $X^2_{\text{win},\alpha=0.1}$, the GC method, the Armitage test for trend, and the PC method as implemented in computer program EIGENSTRAT [Price et al., 2006]. The first two principal components are used as covariates in the PC method. Type I error rates are reported in table 2 and the power is reported in table 3.

It is obvious from table 2 that the Armitage trend test has inflated type I error rate in all situations considered. So is the PC method, although not as severe. Statistics $X^2_{\text{win},0.1}$ and $X^2_{\text{trim},0.1}$ are also liberal for large values of F_1 and F_2. For $\alpha = 0.01$ or 0.05, statistics $X^2_{\text{win},\alpha}$ and $X^2_{\text{trim},\alpha}$ seem to have size close to the nominal level. So is the GC method. More

detailed comparison of these methods under the null hypothesis will be provided later.

After looking at the type I error rates of these statistics, it is of particular interest to compare their power. It can be seen from table 3 that increasing Wright's coefficient of inbreeding tends to decrease their power. However, the pattern of power is different for these statistics. The GC method and the PC method are more powerful than $X^2_{\text{win},\alpha}, X^2_{\text{trim},\alpha}, \alpha = 0.01, 0.05, 0.1$, when F_1 and F_2 are small. But it is less powerful when F_1 and F_2 are in the upper range of their values considered here.

To have a more detailed comparison of the GC method, the PC method, and the proposed method, the distribution of each method under the null hypothesis of no association is compared to the chi-square distribution with 1 degree of freedom via quantile versus quantile plot (Q-Q plot). The data are generated in exactly the same manner as in the study of type I error rate except that there are 500 cases and 500 controls. This plot is produced for different values of F_1 and F_2 and are shown in figures 1, 2, and 3. In all these figures, it is not surprising that the Q-Q plot for Armitage trend test is very far from the 45 degree line $y = x$. The PC method performs better but still demonstrate inflated type I error rate. The GC method tends to be conservative in the tail of its distribution, the part arguably most relevant to testing for association. The Q-Q plot curves seem to be bending downwards and away from the 45 degree line in the tail, indicating that the assumption the VIF is constant across the genome is questionable. This phenomenon tends to be more severe for larger F_1 and F_2 values. One the other hand, plots for the proposed method seem to be on a straight line, regardless of the level of winsorization or trimming and the values of F_1 and F_2 but their slopes are dependent on the level of winsorization or trimming used in the robust variance estimation. Overall, 0.01-winsorized variance and the 0.01-trimmed variance appear to outperform variances at other winsorized or trimmed levels. In addition, the plot is less sensitive to the level of winsorization than to the level of trimming, suggesting that the α-winsorized variance is preferred to the α-trimmed variance.

DISCUSSION

Current genetic association methods adjusting for the effect of population stratification are not very successful in controlling excessive type I error rates. The method proposed in

this report adopts a different approach. It modifies the variance of the estimated effect size by adding to it a constant that measures the variation of the expected effect size across the genome. Unlike the GC method, such a modification of a test statistic results in a VIF that is no longer constant across the genome. Since the proposed method and the GC method both adjust a regular test statistic by applying a VIF, the following discussion is focused on comparison between these two methods.

The VIF in the GC method depends on Wright's coefficient of inbreeding F_{st}. It was argued that it is roughly a constant across the genome as long as F_{st} is so [Devlin & Roeder, 1999]. However, this assumption is not true at least in theory. For the situation where the number of cases and the number of controls are the same, it has been shown that the VIF depends not only on F_{st} within cases and within controls but also on allele frequencies as well as the composition of sub-populations in cases and in controls [Devlin & Roeder, 1999]. The impact of allele frequencies and population structure on the VIF is larger when F_{st} in cases or in controls gets larger. This explains why in my simulation study the GC method performs worse as F_{st} gets larger. Another problem with the GC method is that the estimate of the VIF can be less than 1, especially when the number of GC loci is not large or the degree of inbreeding is small in both cases and controls [Bacanu et al., 2000; Pritchard & Donnelly, 2001]. For instance, in the absence of population structure, the distribution of the Armitage test for trend should follow a chi-square distribution with degree of freedom 1. By definition, 50% of the test statistic at GC markers will be less than the median of the chi-square distribution with degree of freedom 1. That is, the chance for the VIF estimate to be less than 1 is 50%. Whenever the VIF is less than 1, the GC method will not provide any improvement over the original statistic. A remedy is to set the VIF estimate to 1 whenever its initial estimate is less than 1 [Bacanu et al., 2000]. Simulation studies [Bacanu et al., 2000; Pritchard & Donnelly, 2001] suggest that the such a practice generates roughly correct type I error rates but it is not clear theoretically why this is the case.

In contrast, the proposed method always deflates the original test statistic since it adds a non-negative term to its denominator. The extent of modification depends on the variance of the estimated size of the genetic effect – the larger the variance, the smaller the extent of modification.

The proposed method is motivated by a consideration of the unconditional variance of estimated genetic effect size. Instead of focusing on its variance at a particular locus, this method considers the variance at an arbitrary locus. There are two advantages of this approach. First, it provides a natural way of modifying the variance of the estimated genetic effect as delineated in this report. The second advantage that has not been discussed yet is it implicitly treats the mean genetic effect size across the genome is 0, an assumption less stringent than assuming the locus-wise mean genetic effect is 0. Because of population stratification, the locus-wise mean genetic effect size is unlikely to be 0. It is argued for the GC method that the locus-wise mean genetic effect size is 0 and is thus ignored. The unconditional variance of the estimated genetic effect size has also been discussed in the context of GC [Devlin et al., 2001] but it is used to justify the GC method.

Ideally, the additive constant Λ in the proposed method reflects the variation in effect size purely due to population stratification. If it is estimated by its sample counterpart, the estimate will also include sampling variation. The α-trimmed variance and the α-winsorized variance are used to not only remove outliers but also reduce sampling variation. In the simulation study, the 0.01-winsorized variance seems to perform best.

In summary, a new approach to correcting for population stratification in genetic association study is proposed. It appears to be advantageous over the GC method. This method is very easy to implement. A code for the R statistical computing environment [R Development Core Team, 2008] is provided in Appendix B.

ACKNOWLEDGEMENT

I thank an anonymous reviewer and the Editor-in-Chief Dr. Nancy Cox for their useful comments and insights.

REFERENCES

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, *11*, 375–386.

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

Bacanu, S., Devlin, B., & Roeder, K. (2002). Association studies for quantitative traits in structured populations. *Genet Epidemiol*, *22*, 78–93.

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

Devlin, B., Roeder, K., & Wasserman, L. (2001). Genomic control, a new approach to genetic-based association studies. *Theoretical Population Biology*, *60*, 155–166.

Epstein, M. P., Allen, A. S., & Satten, G. A. (2007). A simple and improved correction for population stratification in case-control studies. *Am J Hum Genet*, *80*, 921–930.

Horvath, S. & Laird, N. M. (1998). A discordant-sibship test for disequilibrium and linkage: no need for parental data. *Am J Hum Genet*, *63*, 1886–1897.

Huber, P. (1981). *Robust Statistics*. John Wiley & Sons.

Kimmel, G., Jordan, M. I., Halperin, E., Shamir, R., & Karp, R. M. (2007). A randomization test for controlling population stratication in whole-genome association studies. *Am J Hum Genet*, *81*, 895–905.

Lee, S., Sullivan, P. F., Zou, F., & Wright, F. A. (2008). Comment on a simple and improved correction for populatoin stratication. *Am J Hum Genet*, *82*, 524–526.

Lunetta, K. L. (2008). Genetic association studies. *Circulation*, *118*, 96–101.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-side association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews: Genetics*, *9*, 356–369.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratication in genome-wide association studies. *Nature Genetics*, *38*, 904–909.

Pritchard, J. K. & Donnelly, P. (2001). Case-control studies of association in structured or admixed populatoins. *Theoretical Population Biology*, *60*, 227–237.

Pritchard, J. K. & Rosenberg, N. A. (1999). Use of unlinked genetic markers to detect population stratication in association studies. *Am J Hum Genet*, *65*, 220–228.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*, 945–959.

Pritchard, J. K., Stephens, M., Rosenberg, N. A., & Donnelly, P. (2000). Association mapping in structured populations. *Am J Hum Genet*, *67*, 170–181.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Sasieni, P. D. (1997). From genotypes to genes: Doubling the sample size. *Biometrics*, *53*, 1253–1261.

Zhu, X., Li, S., Cooper, R. S., & Elston, R. C. (2008). A unified association analysis approach for family and unrelated samples correcting for stratication. *Am J Hum Genet*, *82*, 352–365.

APPENDIX A INHERENT INABILITY FOR THE PRINCIPAL COMPONENT METHOD TO COMPLETELY ACCOUNT FOR THE IMPACT OF POPULATION STRATIFICATION ON ASSOCIATION STUDY

The essence of the principal component method is to use the top principal components extracted from the correlation or the covariance matrix of genotype scores as surrogates of the underlying population structure. Let S denote population stratum. Let g_1 and g_2 be genotype scores at two markers that are not associated with phenotype y. That is, they are not correlated with y within each subpopulation. g_1 and g_2 are centered around their individual mean. The covariance σ_{yg_1} between y and g_1 is

$$\begin{aligned}\sigma_{yg_1} &= Cov(y, g_1) \\ &= Cov[E(y|S), E(g_1|S)] + E[Cov(y, g_1|S)] \\ &= Cov[E(y|S), E(g_1|S)].\end{aligned}$$

Similarly, the covariance σ_{yg_2} between g_2 and y is $\sigma_{yg_2} = Cov[E(y|S), E(g_2|S)]$. If these two markers are used for a principal components method, only the largest principal component will be of interest. Denote it by P. That is, $P = a_1g_1 + a_2g_2$ with $(a_1, a_2)^t$ the largest eigenvector of the covariance matrix of g_1 and g_2. Here it is assumed that the principal component is covariance matrix based.

In the principal components method, it is expected that, if P is used as a covariate, the partial correlation between the phenotype y and any unassociated marker is 0. If this conjecture is true, it must be the case that the partial correlation between y and g_2 is 0. However, this is generally not the case. What follows explains why.

This partial correlation can be expressed

$$Corr(y, g_2|P) = \frac{r_{yg_2} - r_{yP}r_{g_2P}}{[(1 - r_{yP}^2)(1 - r_{g_2P}^2)]^{1/2}},$$

where $r_{x,y}$ denotes the correlation coefficient between variables x and y. The numerator of this fraction is proportional to

$$\sigma_{yg_2}Var(P) - \sigma_{yP}\sigma_{g_2P},$$

which equals

$$\begin{aligned} & \sigma_{yg_2}(a_1^2\sigma_{g_1}^2 + 2a_1a_2\sigma_{g_1g_2} + a_2^2\sigma_{g_2}^2) - (a_1\sigma_{yg_1} + a_2\sigma_{yg_2})(a_1\sigma_{g_1g_2} + a_2\sigma_{g_2}^2) \\ = \; & a_1[\sigma_{yg_2}(a_1\sigma_{g_1}^2 + a_2\sigma_{g_1g_2}) - \sigma_{yg_1}(a_1\sigma_{g_1g_2} + a_2\sigma_{g_2}^2)] \\ = \; & \lambda a_1(a_1\sigma_{yg_2} - a_2\sigma_{yg_1}), \end{aligned}$$

where λ is the eigenvalue associated with $(a_1, a_2)^t$. For it to be 0, there must be $a_1\sigma_{yg_2} - a_2\sigma_{yg_1} = 0$. This relationship seldom holds as a_1 and a_2 depends only on the covariance between markers g_1 and g_2 while σ_{yg_1} and σ_{yg_2} are covariances that depend in addition on phenotype y.

As a numerical example, consider a cohort study with two subpopulations, each subpopulation has 0.5 probability. The probability of being affected is 0.1 in population 1 and 0.2 in population 2. At SNP 1, the A allele frequency is 0.2 in population 1 and 0.4 in population 2. At SNP 2, the A allele frequency is 0.1 in population 1 and 0.4 in population 2. Both SNPs are in HWE within each population and are in linkage equilibrium with each other. None of them is associated with the phenotype y in each population. Each SNP is scored by the number of A alleles. It turns out that the variance matrix for the vector of genotype scores (g_1, g_2) is

$$\begin{pmatrix} 0.44 & 0.06 \\ 0.06 & 0.42 \end{pmatrix},$$

for which the largest eigenvector is $(a_1, a_2)^t = (-0.76302, -0.6463749)^t$. In addition, $\sigma_{yg_1} = 0.01$ and $\sigma_{yg_2} = 0.015$. It is easy to verify that $a_1\sigma_{yg_2} - a_2\sigma_{yg_1} = -0.01790905 \neq 0$.

APPENDIX B R CODE FOR THE PROPOSED METHOD

```
Var.Est = function(null.cases, null.ctrls, alpha, method = "win")
# null.cases, null.ctrls:
#         n x 3 matrices of genotype counts at n genomic control SNPs.
#         Column 1, 2 and 3 are counts of subjects that have 0, 1 and 2
#               copies of, say, A allele.
# alpha: level for alpha-trimmed variance or alpha-winsorized variance
# method: method used in the robust variance estimation.
#               method = "win" for the alpha-winsorized method and
#               method = "trim" for the alpha-trimmed method.
#
{
    RR = apply(null.cases, 1, sum)
    NN = RR + apply(null.ctrls, 1, sum)
    T = (NN*null.cases - RR*(null.ctrls+null.cases)) %*% c(0,1,2)

    y = sort(as.vector(T))
    L = length(y)
    k = floor(L*a)
    if (var.method == "win")
        Lambda = var(c(rep(y[k+1], k+1), y[(k+2):(L-k-1)], rep(y[L-k], k+1)))
    if (var.method == "trim")
        Lambda = var(y[(k+1):(L-k)])

    Lambda
}

new.stat = function(rs, ss, Lambda)
# rs: a vector of length 3. It contains genotype counts of 0, 1 and 2 copies
```

```
#            of, say, A allele, respectively, in cases.
# ss: the same as rs but is for controls
#
# Lambda: an estimate of Lambda. Obtained from function Var.Est(...).
#                See the text for further explanation.
{
    r1 = rs[2]   # number of 1's in cases
    r2 = rs[3]   # number of 2's in cases
    R = sum(rs)
    n0 = rs[1]+ss[1]
    n1 = rs[2]+ss[2]
    n2 = rs[3]+ss[3]
    N = n0+n1+n2

    sigma2T = R*(1-R/N)*(N*(n1+4*n2)-(n1+2*n2)^2)
    (N*(r1+2*r2)-R*(n1+2*n2))^2/(sigma2T + Lambda)
}
```

Table 1: Genotype counts for a case-conrol study

	# of A allele			
	0	1	2	Total
Case	r_0	r_1	r_2	R
Control	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

Table 2: Type I rates of test statistics $X^2_{\text{win},\alpha}, X^2_{\text{trim},\alpha}, \alpha = 0.01, 0.05, 0.1$, the genomic control (GC) method, the Armitage trend test, and the principal control (PC) method. The Wright's coefficient of inbreeding is denoted by F_1 for the case population and by F_2 for the control population. The critical value is obtained from a chi-square distribution with 1 degree of freedom with significance level 0.01.

F_1	F_2	$X^2_{\text{win},0.01}$	$X^2_{\text{win},0.05}$	$X^2_{\text{win},0.1}$	$X^2_{\text{trim},0.01}$	$X^2_{\text{trim},0.05}$	$X^2_{\text{trim},0.1}$	GC	Trend Test	PC
0.001	0.001	0.000	0.000	0.001	0.000	0.001	0.001	0.003	0.011	0.017
0.001	0.003	0.000	0.001	0.001	0.000	0.002	0.004	0.010	0.021	0.036
0.001	0.010	0.002	0.002	0.002	0.002	0.002	0.008	0.004	0.069	0.014
0.001	0.030	0.002	0.002	0.006	0.002	0.007	0.021	0.000	0.190	0.014
0.003	0.001	0.001	0.003	0.004	0.003	0.005	0.009	0.011	0.030	0.032
0.003	0.003	0.003	0.004	0.004	0.003	0.004	0.009	0.009	0.042	0.031
0.003	0.010	0.001	0.003	0.010	0.003	0.011	0.026	0.006	0.104	0.032
0.003	0.030	0.002	0.004	0.015	0.004	0.019	0.037	0.004	0.236	0.019
0.010	0.001	0.001	0.003	0.005	0.002	0.007	0.013	0.012	0.075	0.027
0.010	0.003	0.002	0.002	0.004	0.002	0.004	0.014	0.007	0.097	0.027
0.010	0.010	0.003	0.007	0.009	0.006	0.010	0.023	0.008	0.138	0.019
0.010	0.030	0.008	0.011	0.022	0.010	0.025	0.057	0.007	0.281	0.019
0.030	0.001	0.003	0.009	0.012	0.009	0.017	0.028	0.006	0.196	0.017
0.030	0.003	0.003	0.005	0.012	0.003	0.014	0.031	0.003	0.191	0.021
0.030	0.010	0.003	0.006	0.011	0.005	0.011	0.034	0.003	0.234	0.017
0.030	0.030	0.008	0.011	0.021	0.009	0.031	0.057	0.003	0.341	0.023

Table 3: Power of test statistics $X^2_{\text{win},\alpha}, X^2_{\text{trim},\alpha}, \alpha = 0.01, 0.05, 0.1$, the genomic control (GC) method, the Armitage trend test, and the principal component (PC) method. The Wright's coefficient of inbreeding is denoted by F_1 for the case population and by F_2 for the control population. The genotype relative risk for the multiplicative model is denoted by γ. The critical value is obtained from a chi-square distribution with 1 degree of freedom with significance level 0.01.

F_1	F_2	γ	$X^2_{\text{win},0.01}$	$X^2_{\text{win},0.05}$	$X^2_{\text{win},0.1}$	$X^2_{\text{trim},0.01}$	$X^2_{\text{trim},0.05}$	$X^2_{\text{trim},0.1}$	GC	Trend Test	PC
0.001	0.001	1.75	0.129	0.150	0.183	0.147	0.197	0.252	0.311	0.431	0.435
		2.75	0.668	0.704	0.748	0.687	0.758	0.805	0.899	0.910	0.910
0.003	0.001	1.75	0.134	0.164	0.191	0.159	0.206	0.266	0.335	0.467	0.463
		2.75	0.573	0.621	0.663	0.603	0.684	0.738	0.791	0.886	0.858
0.003	0.003	1.75	0.092	0.121	0.161	0.113	0.170	0.240	0.212	0.461	0.303
		2.75	0.527	0.574	0.629	0.560	0.653	0.725	0.777	0.887	0.698
0.010	0.001	1.75	0.075	0.103	0.141	0.087	0.149	0.212	0.170	0.447	0.127
		2.75	0.494	0.542	0.574	0.521	0.595	0.664	0.660	0.858	0.514
0.010	0.003	1.75	0.079	0.096	0.127	0.087	0.151	0.198	0.120	0.459	0.101
		2.75	0.406	0.457	0.529	0.440	0.550	0.631	0.548	0.849	0.305
0.010	0.010	1.75	0.067	0.097	0.125	0.081	0.137	0.192	0.108	0.476	0.066
		2.75	0.325	0.404	0.450	0.380	0.469	0.549	0.363	0.838	0.167
0.030	0.001	1.75	0.056	0.086	0.103	0.073	0.120	0.170	0.066	0.501	0.050
		2.75	0.282	0.340	0.398	0.314	0.410	0.504	0.321	0.815	0.127
0.030	0.003	1.75	0.050	0.075	0.097	0.066	0.108	0.157	0.058	0.466	0.049
		2.75	0.227	0.279	0.345	0.262	0.365	0.460	0.254	0.812	0.114
0.030	0.030	1.75	0.047	0.069	0.098	0.057	0.114	0.166	0.034	0.511	0.036
		2.75	0.134	0.185	0.253	0.162	0.270	0.371	0.124	0.731	0.067

Figure 1: Q-Q plot of the distributions of test statistics when there is no association. From left to right, top to bottom: $X^2_{\text{win},0.01}, X^2_{\text{win},0.05}, X^2_{\text{win},0.1}, X^2_{\text{trim},0.01}, X^2_{\text{trim},0.05}, X^2_{\text{trim},0.1}$, the genomic control method, the Armitage trend test, and the principal component method. The Wright's coefficient of inbreeding is 0.01 for cases and 0.001 for controls. The reference line is $y = x$, the 45 degree line.

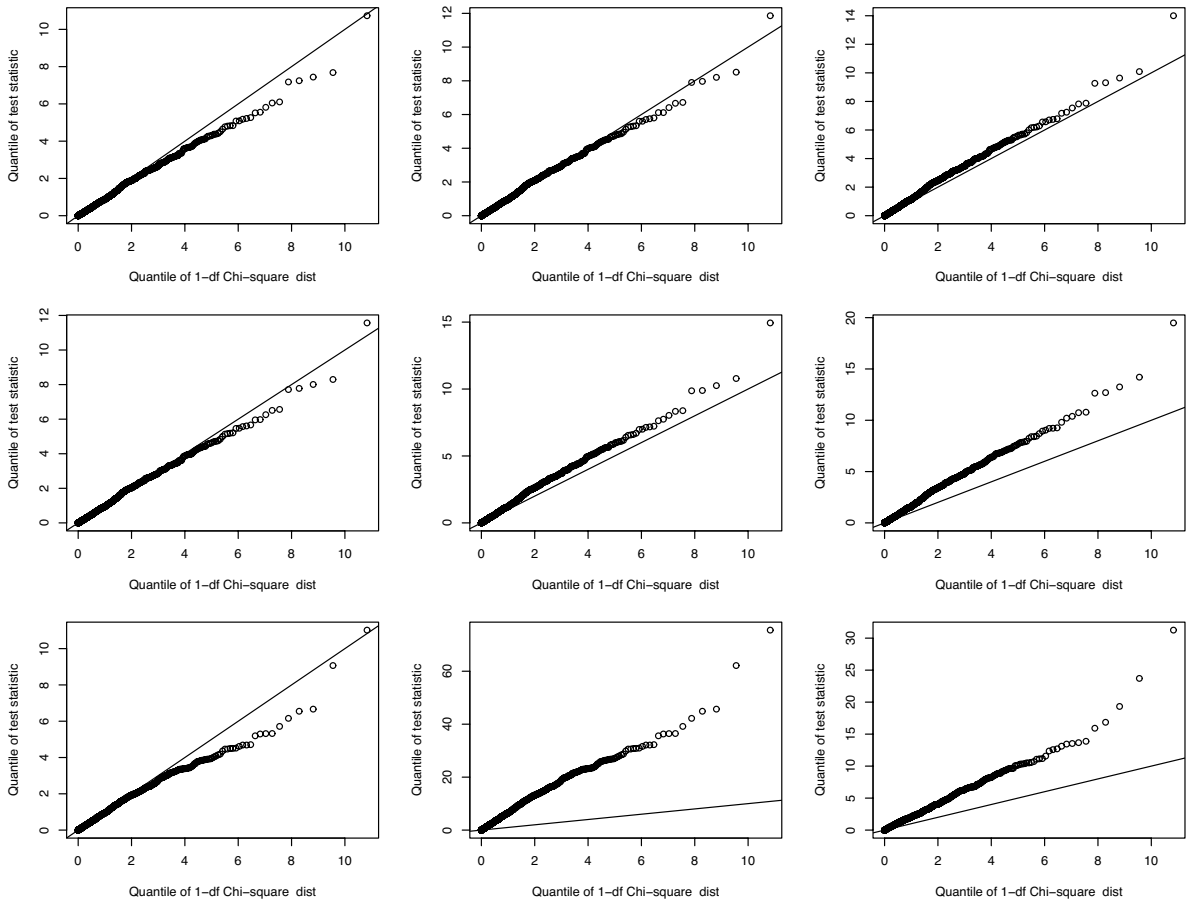

Figure 2: Q-Q plot of the distributions of test statistics when there is no association. From left to right, top to bottom: $X^2_{\text{win},0.01}, X^2_{\text{win},0.05}, X^2_{\text{win},0.1}, X^2_{\text{trim},0.01}, X^2_{\text{trim},0.05}, X^2_{\text{trim},0.1}$, the genomic control method, the Armitage trend test, and the principal component method. The Wright's coefficient of inbreeding is 0.03 for cases and 0.003 for controls. The reference line is $y = x$, the 45 degree line.

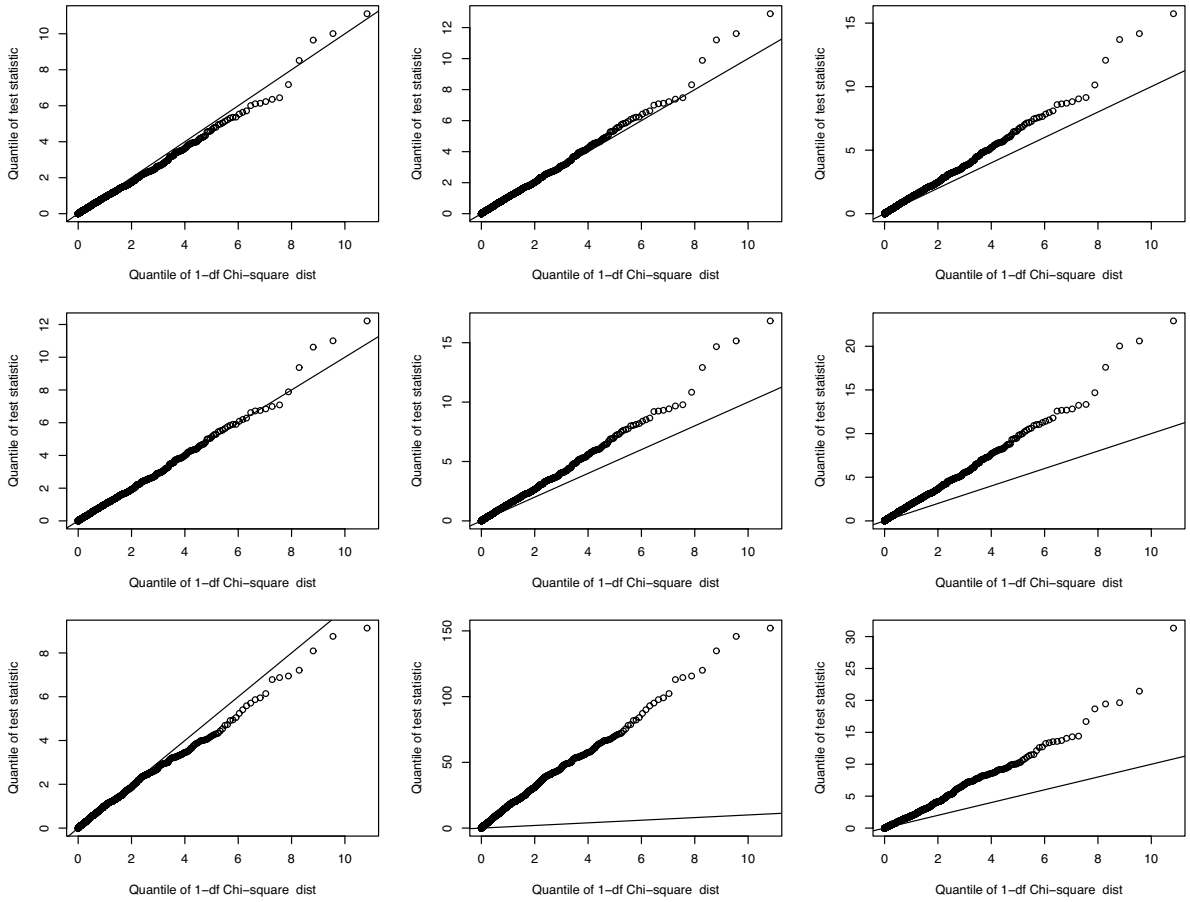

Figure 3: Q-Q plot of the distributions of test statistics when there is no association. From left to right, top to bottom: $X^2_{\text{win},0.01}, X^2_{\text{win},0.05}, X^2_{\text{win},0.1}, X^2_{\text{trim},0.01}, X^2_{\text{trim},0.05}, X^2_{\text{trim},0.1}$, the genomic control method, the Armitage trend test, and the principal component method. The Wright's coefficient of inbreeding is 0.1 for cases and 0.03 for controls. The reference line is $y = x$, the 45 degree line.