Received: 24 June 2021 | Revised: 20 May 2022 | Accepted: 25 May 2022

DOI: 10.1002/gepi.22493

RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

WILEY

Control for population stratification in genetic association studies based on GWAS summary statistics

Shijia Yan | **Qiuying Sha** | **Shuanglin Zhang**

Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan, USA

Correspondence
Shuanglin Zhang, Department of Mathematical Sciences, Michigan Technological University, 1400 Townsend Dr, Houghton, MI 49931, USA.
Email: shuzhang@mtu.edu

Abstract

Over the past years, genome-wide association studies (GWAS) have generated a wealth of new information. Summary data from many GWAS are now publicly available, promoting the development of many statistical methods for association studies based on GWAS summary statistics, which avoids the increasing challenges associated with individual-level genotype and phenotype data sharing. However, for population-based association studies such as GWAS, it has been long recognized that population stratification can seriously confound association results. For large GWAS, it is very likely that there exist population stratification and cryptic relatedness, which will result in inflated Type I error in association testing. Although many methods have been developed to control for population stratification, only two of these approaches can be used to control population stratification without individual-level data: one is based on genomic control (GC) and the other one is based on linkage disequilibrium score regression (LDSC). However, the performance of these two approaches is currently unknown. In this study, we use extensive simulation studies including populations with subpopulations, spatially structured populations, and populations with cryptic relatedness to compare the performance of these two approaches to control for population stratification using only GWAS summary statistics without individual-level data. Data sets from the genetic analysis workshop 19 and UK Biobank are also used to evaluate these two approaches. We demonstrate that the intercept of LDSC can be used as a more accurate correction factor than GC. The results from this study will provide very useful information for researchers using GWAS summary statistics while trying to control for population stratification.

KEYWORDS

association study, GWAS summary statistics, LD score regression, population stratification

1 | INTRODUCTION

Over the past 16 years, genome-wide association studies (GWAS) have generated a wealth of new information. Summary data from many GWAS are now publicly available, promoting the development of many statistical methods for association studies based on GWAS summary statistics, which avoids the increasing challenges associated with individual-level genotype and phenotype data sharing. However, for population-based association studies, it has been long recognized that population stratification can seriously confound association results

(Knowler et al., 1988; Lander & Schork, 1994). For large GWAS, it is very likely that there exist population stratification and cryptic relatedness, which will result in inflated Type I error in association testing. To correct the inflation, many methods that use a set of genomic markers genotyped in the same samples have been developed to control for population stratification. These methods include the genomic control (GC) approach (Devlin & Roeder, 1999; Devlin et al., 2001; Reich & Goldstein, 2001), linkage disequilibrium score regression (LDSC) (Bulik-Sullivan et al., 2015), principal component (PC)-based approaches (Chen et al., 2003; Price et al., 2006; S. Zhang et al., 2003), approaches by dividing the underlying population into several homogeneous subpopulations and then constructing test statistics based on homogeneous subpopulations (Pritchard, Stephens, & Donnelly, 2000; Pritchard, Stephens, Rosenberg, et al., 2000; S. Zhang & Zhao, 2001), mixed linear model approaches (Kang et al., 2010; Z. Zhang et al., 2010), and approaches for rare variants association studies (Jiang et al., 2013; Sha et al., 2016; Y. Zhang et al., 2013).

Although many methods have been developed to control for population stratification, only two of these approaches can be used to control for population stratification without individual-level data: one is based on GC (Devlin & Roeder, 1999) and the other one is based on LDSC (Bulik-Sullivan et al., 2015). GC assumes that only a small fraction of single-nucleotide polymorphisms (SNPs) are associated with a trait, and no association exists for other SNPs (Yang et al., 2011). Under this assumption, GC corrects the inflation of test statistics by dividing a correction factor. However, GC fails to distinguish polygenicity (i.e., many small genetic effects) from confounding bias based on that assumption and polygenicity also contributes to the inflation of test statistics (Bulik-Sullivan et al., 2015). Furthermore, with the evidence that the GC correction factor increases as a sample size increases in the presence of polygenicity, GC is too conservative and suffers a loss of power for large samples (Devlin & Roeder, 1999; Yang et al., 2011).

The other approach to control for population stratification without individual-level data is based on LDSC (Bulik-Sullivan et al., 2015). As discussed by Bulik-Sullivan et al., the intercept of LDSC provides a more robust quantification of inflation. In LDSC, the LD score is constructed to measure the degree of trait-associated genetic variation tagged by an SNP. For a given trait, SNPs with higher LD scores are more likely to tag causal SNPs and thus have more inflated corresponding test statistics. Besides, the inflation from cryptic relatedness or population stratification is not correlated with LD scores (Bulik-Sullivan et al., 2015; Devlin & Roeder, 1999; Voight & Pritchard, 2005; Yang et al., 2011). Based on these two key evidences, LDSC associates χ^2 statistics with LD scores by a simple linear regression model and the inflation caused by confounding is distributed to the intercept. Therefore, LDSC can distinguish polygenicity from confounding.

In fact, under the null hypothesis of GWAS, the theoretical expectation of a χ^2 test statistic is one for each SNP. Then, if there is an inflation due to confounding or other artifacts, the intercept of LDSC can measure the contribution. In spite of this, as mentioned in Bulik-Sullivan et al. (2015), if there were a positive correlation between LD score and Wright F_{st} (Bhatia et al., 2013), the intercept of LDSC would underestimate the contribution of population stratification to the inflation in χ^2 statistics.

Although the intercept of LDSC plays a key role in estimating the inflation, how to use the intercept to control for population stratification and the performance of using the intercept as a correction factor are unknown. In this paper, we consider two approaches to control for population stratification using the intercept of LDSC: (1) every χ^2 statistic will be corrected by minus the intercept (LD-M); (2) similar to GC, let the intercept be a correction factor, then every χ^2 statistic will be corrected by dividing the correction factor (LD-D). We use extensive simulations, including (1) populations with k_0 subpopulations; (2) spatially structured populations; (3) populations with cryptic relatedness, to compare LD-M and LD-D with GC and evaluate the performance of these methods. We also apply LD-M, LD-D, and GC to data sets from the genetic analysis workshop 19 (GAW19) and UK Biobank for further evaluations.

2 | METHODS

If an individual-level genotype and phenotype data set is available for a GWAS, we can use a score test statistic to test the association between a trait and an SNP. Suppose that there are a total of n individuals and J SNPs in a GWAS. For the jth SNP, we let y_i and x_i denote the trait and genotype for the ith individual, where $i = 1, ..., n$. The score test statistic is given by $T^U_{\text{score}}(j) = U^2/V$, where $U = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})$ and $V = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\sum_{i=1}^{n}(x_i - \bar{x})^2$. Under the null hypothesis that the jth SNP is not associated with the trait, the test statistic $T^U_{\text{score}}(j)$ asymptotically follows a χ^2 distribution with one degree of freedom (df) (Sha et al., 2011). If individual-level genotype and phenotype data are not available and only GWAS summary statistics are available, we let Z_j be the Z-score for the jth SNP, then $T^U_{\text{score}}(j) = Z_j^2$ for $j = 1, 2, ..., J$.

In GC (Devlin & Roeder, 1999), the inflation of the score test statistic $T^U_{\text{score}}(j)$ for the jth SNP is corrected by dividing the correction factor λ, where $\lambda = median\left(T^U_{\text{score}}(1), ..., T^U_{\text{score}}(J)\right)/0.456$, the ratio of the median of the observed test statistics and the median of the $\chi^2_{df=1}$ distribution.

In LDSC (Bulik-Sullivan et al., 2015), we first calculate the LD score for the jth SNP by $l_j = \sum_k r^2_{jk}$, where $j = 1, ..., J$, $k = 1, ..., J$, and $r^2_{jk} = R^2_{jk} - \frac{1 - R^2_{jk}}{n-2}$ is the squared correlation between SNP j and SNP k, and R^2_{jk} is the squared Pearson correlation between SNP j and SNP k. Then, we obtain the score test statistic T^U_{score} for each SNP from a GWAS either using individual-level genotype and phenotype data or using summary statistics. At last, by LDSC, $E\left[T^U_{\text{score}}(j)\middle|l_j\right] = c + 1 + \beta l_j$, we can estimate the intercept $c + 1$.

Under different scenarios, we compare the following four test statistics to control for population stratification: (1) GC: $T^U_{\text{score}}/\lambda$; (2) LD-D: $T^U_{\text{score}}/\left(c+1\right)$, the score test statistic divided by the intercept in LDSC; (3) LD-M: $T^U_{\text{score}} - \left(c+1\right)$, the score test statistic minus the intercept in LDSC; and (4) uncorrected: T^U_{score}.

3 | RESULTS

3.1 | Simulation studies

To compare the performance of the above four methods, we consider scenarios that confounding bias is due to population stratification and cryptic relatedness. We use the simulation procedures similar to the simulations in Devlin and Roeder (1999) and Sha et al. (2016). For scenarios where confounding bias is due to population stratification, we consider both qualitative and quantitative traits. To generate qualitative traits, we use a liability threshold model with a 30% prevalence for the simulated disease status and define cases and controls based on the generated quantitative traits (case:control ≈ 3:7). We consider three sets of simulations: (1) populations with k_0 subpopulations; (2) populations with spatially structured populations; and (3) populations with cryptic relatedness.

3.1.1 | Simulation Set 1: Populations with k_0 subpopulations

In this simulation, we use the minor allele frequencies (MAFs) of 24,487 SNPs from the GAW17. In GAW17, there are 697 unrelated individuals. We follow the procedures of Price et al. (2006) and Sha et al. (2016) to generate genotypes of individuals in a population with k_0 subpopulations. For each SNP, we randomly choose a MAF from 24,487 SNPs in GAW17 as the ancestral population allele frequency p. Then, we independently draw k_0 values $p_1, ..., p_{k_0}$ from the β-distribution with parameters $p(1 - F_{\text{st}})/F_{\text{st}}$ and $(1-p)(1-F_{\text{st}})/F_{st}$, where F_{st} is the Wright measure of population subdivision (Balding & Nichols, 1995) (in this study, $F_{\text{st}} = 0.01$). We accept $p_1, ..., p_{k_0}$ as allele frequencies for the k_0 subpopulations if $\frac{1}{k_0}\sum_{i=1}^{k_0} p_i \geq 0.002$; otherwise, we redraw $p_1, ..., p_{k_0}$.

To generate quantitative traits, we use the model $y_{ik} = \mu_k + \beta x_{ik} + \varepsilon_{ik}$, where $i = 1, ..., n_k$, $k = 1, ..., k_0$, the number of individuals $n = n_1 + \cdots + n_{k_0}$, y_{ik} and x_{ik} are the trait and genotype of the ith individual in the kth subpopulation, and $\varepsilon_{ik} \sim N(0,1)$. Under the null hypothesis, we set $\beta = 0$. In this study, we consider $n = 1000$, $n_k = n/k_0 (k = 1, ..., k_0)$, $k_0 = 1, 5, 10,$ and 20, $\mu_1 = 0$, and $\mu_2 = \mu_3 = \cdots = \mu_{k_0} = (k_0 - 1)\mu$. We set $\mu = 0.1$ for evaluating powers and $\mu = 0.3$ for evaluating Type I error rates.

3.1.2 | Simulation Set 2: Population with spatially structured populations

To generate spatially structured populations, we follow the simulation procedures in Mathieson and McVean (2012) and Sha et al. (2016). We first divide the space into $K_0 \times K_0$ grid squares. Then, we start with the number of individuals and their locations on the grid. Based on random genealogical events, including the coalescence of two lineages and a migration of a single lineage from one square to another, we generate genotypes backwards in time. The relative rates of coalescence and migration depend on the population-scaled migration rate M and the number and distribution of lineages on the grid (Sha et al., 2016).

For n individuals, let $\phi(i) = (k, h)$ if the ith individual originates from grid square k, h, where $k = 1, ..., K_0$ and $h = 1, ..., K_0$. Denote the nongenetic risk in grid square k, h by $R_{k,h}$. Then, under the null hypothesis, the trait of the ith individual is generated by $y_i = \alpha R_{\phi(i)} + \varepsilon_i$, where $\varepsilon_i \sim N(0,1)$ and α is a constant. Under the alternative hypothesis, the trait for an individual is generated by $y = \beta x + y_0$, where y_0 is the trait generated under the null hypothesis.

In this study, we use $n = 800$, $K_0 = 20$, the population-scaled migration rate $M = 0.01$, and $\alpha = 2$. We generate genotypes for $J = 1000$ SNPs and there is one causal SNP in the simulation of power. Spatially structured populations can be analogized under three scenarios for different values of $R_{k,h}$. Scenario 1: There is

no population stratification with $R_{k,h} = 0$ for all k and h. Scenario 2: There is a small and sharp spatial distribution with $R_{k,h} = 1$ for $k = 6, 7, 14, 15$ and $h = 6, 7, 14, 15$, and $R_{k,h} = 0$ for other k, h. Scenario 3: There is a wide and smooth spatial distribution in which $R_{k,h} = 0.4e^{-((k-k_0)^2+(h-h_0)^2)/18}$ and $k_0 = h_0 = 6$.

3.1.3 | Simulation Set 3: Population with cryptic relatedness

To generate a population with cryptic relatedness, we follow the simulation procedures in Devlin and Roeder (1999). In this simulation, we only consider a balanced case–control study and assume that cases and controls both have a fixed allelic correlation F_1 and F_2, respectively.

To generate genotypes for the evaluation of Type I error rates, we first draw a value of p from a β-distribution with parameters $\alpha = \beta = (1 - F_{st})/2F_{st}$, where $F_{st} = F_1$ for cases and $F_{st} = F_2$ for controls. Then, we generate a binomial sample of two alleles with parameter p to form the genotype. Under the null hypothesis, we assume that $F_1 = F_2 = F_{st}$, and we let $F_{st} = 0.00001$, 0.0001, 0.001, and 0.01 in our simulation studies. Considering the influence of the large sample size, we generate genotypes at $J = 1000$ SNPs for $n = 1000$, 5000, and 10,000.

In the simulation for power comparison, cases are likely to be related compared with controls in a randomly mating population as they share a genetic disorder, so we let $F_1 = 0.00001, 0.001$, and 0.005 and $F_2 = 0.00001$ and 0.001. Under the alternative hypothesis, we generate a binomial sample of two alleles with parameter $p = \gamma/(1+\gamma)$ to form the genotype at the causal SNP for each of the cases; for noncausal SNPs in cases, we use p drawn from a β-distribution with parameters $\alpha = \beta = (1 - F_1)/2F_1$. For controls, we use p drawn from a β-distribution with parameters $\alpha = \beta = (1 - F_2)/2F_2$, and genotypes are generated under the null hypothesis.

We generate genotypes at $J = 1000$ SNPs for $n = 1000$ individuals. We only consider one causal SNP in the simulation of power. We use $\gamma = 1.25$, 1.35, 1.45, and 1.55 to generate the causal SNP in cases.

3.2 | Simulation results

3.2.1 | Type I error rates

To evaluate Type I error of the four methods, we consider different types of traits, different sample sizes, and different models. For each simulation set, we generate 1000 replicated samples. Each individual contains genotypes at 1000 SNPs and a phenotype. We consider 1000 replicated samples and 1000 SNPs as $1000 \times 1000 = 10^6$ replicated samples. For 10^6 replicated samples, the 95% confidence interval (CI) of Type I error rates divided by nominal level 0.05 is (0.9915, 1.009). The Type I error rates beyond the corresponding upper bound of 95% CIs are boldfaced in Tables 1–3.

For simulation set 1, we consider a population with $k_0 = 1$, 5, 10, and 20 subpopulations. The Type I error rates divided by the nominal level 0.05 of each method are summarized in Table 1. For a quantitative trait, we can find that when there is no subpopulation ($k_0 = 1$), all methods can control Type I error rates. When subpopulations exist, the uncorrected test has inflated Type I error rates, LD-M can control Type I error rate for $k_0 = 5$, but fail to control Type I error for more subpopulations; LD-D and GC both have correct Type I error rates. For a qualitative trait, we can find similar results, and the Type I error rate of the uncorrected test is inflated even for a homogeneous population ($k_0 = 1$).

For simulation set 2, we consider three scenarios of spatially structured populations. The results of Type I error rates are summarized in Table 2. For a quantitative trait, we find that when there is no population stratification (Scenario 1), all methods have correct Type I error rates; but in the cases of population stratification (Scenarios 2 and 3), only LD-D can control Type I error rates. For a qualitative trait, only the uncorrected test cannot control Type I error rates under any of the three scenarios.

For simulation set 3, we consider a balanced case-control study with different sample sizes. Type I error

TABLE 1 Type I error rates are divided by the nominal level 0.05 of uncorrected test, GC, LD-D, and LD-M for simulation set 1

Trait	k_0	Uncorrected	GC	LD-D	LD-M
Quantitative	1	1.002	1.01	0.9936	0.5525
	5	**2.294**	0.9814	0.9856	0.9736
	10	**2.935**	0.9488	0.9789	**1.536**
	20	**2.767**	0.9487	0.9836	**1.459**
Qualitative	1	**1.059**	0.7442	0.9307	0.4628
	5	**1.834**	0.7739	0.8995	0.8844
	10	**2.314**	0.7144	0.892	**1.138**
	20	**2.334**	0.7261	0.8911	**1.162**

Note: Type I error rates in boldface indicate the values beyond the upper bound of the 95% CIs.

Abbreviations: CI, confidence interval; GC, genomic control; LD-D, every χ^2 statistic will be corrected by dividing the intercept; LD-M, every χ^2 statistic will be corrected by minus the intercept.

Trait	Scenario	Uncorrected	GC	LD-D	LD-M
Quantitative	1	1.004	0.9998	0.9956	0.5516
	2	**2.475**	**1.404**	1.008	**1.406**
	3	**1.974**	**1.156**	1.002	**1.113**
Qualitative	1	**1.217**	0.5119	0.9529	0.3286
	2	**2.056**	0.6668	0.8263	0.8617
	3	**1.829**	0.66	0.8014	0.6798

Note: Type I error rates in boldface indicate the values beyond the upper bounds of the 95% CIs.

Abbreviations: CI, confidence interval; GC, genomic control; LD-D, every χ^2 statistic will be corrected by dividing the intercept; LD-M, every χ^2 statistic will be corrected by minus the intercept.

TABLE 2 Type I error rates divided by the nominal level 0.05 of uncorrected test, GC, LD-D, and LD-M for simulation set 2

TABLE 3 Type I error rates divided by the nominal level 0.05 of uncorrected test, GC, LD-D, and LD-M for simulation set 3

n	F_{st}	Uncorrected	GC	LD-D	LD-M
1000	0.00001	**1.022**	1.007	0.9992	0.566
1000	0.0001	**1.232**	**1.01**	0.9976	0.6812
1000	0.001	**3.314**	0.997	0.9967	**1.74**
1000	0.01	**11.14**	0.9491	0.9846	**4.952**
5000	0.00001	**1.118**	1.003	0.9965	0.6152
5000	0.0001	**2.198**	1.008	0.9952	**1.182**
5000	0.001	**8.479**	1.002	0.998	**4.012**
5000	0.01	**15.7**	0.9416	0.9765	**6.068**
10,000	0.00001	**1.231**	1.006	0.9948	0.6772
10,000	0.0001	**3.318**	**1.02**	0.9999	**1.755**
10,000	0.001	**11.08**	1.009	0.9944	**4.918**
10,000	0.01	**16.93**	0.9523	0.9793	**6.235**

Note: Type I error rates in boldface indicate the values beyond the upper bounds of the 95% CIs.

Abbreviations: CI, confidence interval; GC, genomic control; LD-D, every χ^2 statistic will be corrected by dividing the intercept; LD-M, every χ^2 statistic will be corrected by minus the intercept.

rates of each method are summarized in Table 3. We find that (1) LD-D always has correct Type I error rates; (2) GC has inflated Type I error rates when $F_{st} = 0.0001$; (3) LD-M only can control Type I error rate when F_{st} is small; and (4) uncorrected test always has inflated Type 1 error rates.

3.2.2 | Powers

To evaluate the power of the four methods, we consider 1000 replicated samples. Each sample contains 1000 SNPs and a trait. We use a significant level of 0.05 in the power comparison.

For simulation set 1 with a population including $k_0 = 1$, 5, 10, and 20 subpopulations, the powers of each method are summarized in Figures 1 and 2 for a quantitative trait and a qualitative trait, respectively. For a quantitative trait, we can find that (1) when there is only one subpopulation ($k_0 = 1$), LD-D, GC, and uncorrected test have comparable powers, but LD-M has lower power; and (2) when there is more than one subpopulation, the uncorrected test has the highest power, but its Type I error rates are inflated; LD-M is more powerful than GC and LD-D, but the Type I error rates of LD-M are also inflated; LD-D is more powerful than GC, although both of them can control Type I error rates. For a qualitative trait, we can find similar results.

For simulation set 2, there are three scenarios of spatially structured populations. From Figure 3, we find that for a quantitative trait (1) when there is no population stratification (Scenario 1), all of these four methods, GC, LD-D, LD-M, and uncorrected test, have comparable powers; and (2) for spatially structured populations with a small and sharp spatial distribution or with a wide and smooth spatial distribution (Scenarios 2 and 3), LD-D has the smallest power, but only LD-D can control Type I error rates in these cases. For the powers of a quantitative trait shown in Figure 4, the uncorrected test and LD-D have higher powers when there is no population stratification (Scenario 1); however, the Type I error rate of the uncorrected test is inflated; LD-M has the lowest power. For the two scenarios of spatially structured populations (Scenarios 2 and 3), although the uncorrected test has the highest power, it cannot control Type I error rate; among the three methods, LD-M, LD-D, and GC can control Type I error rates, and LD-M and LD-D have higher power than GC.

For simulation set 3, we compare the powers of each method with the sample size $n = 1000$. From Figure 5, we can find that (1) the powers of GC and LD-D are comparable; (2) LD-M has higher powers than LD-D and GC, but it only can control Type I error rate when F_{st} is

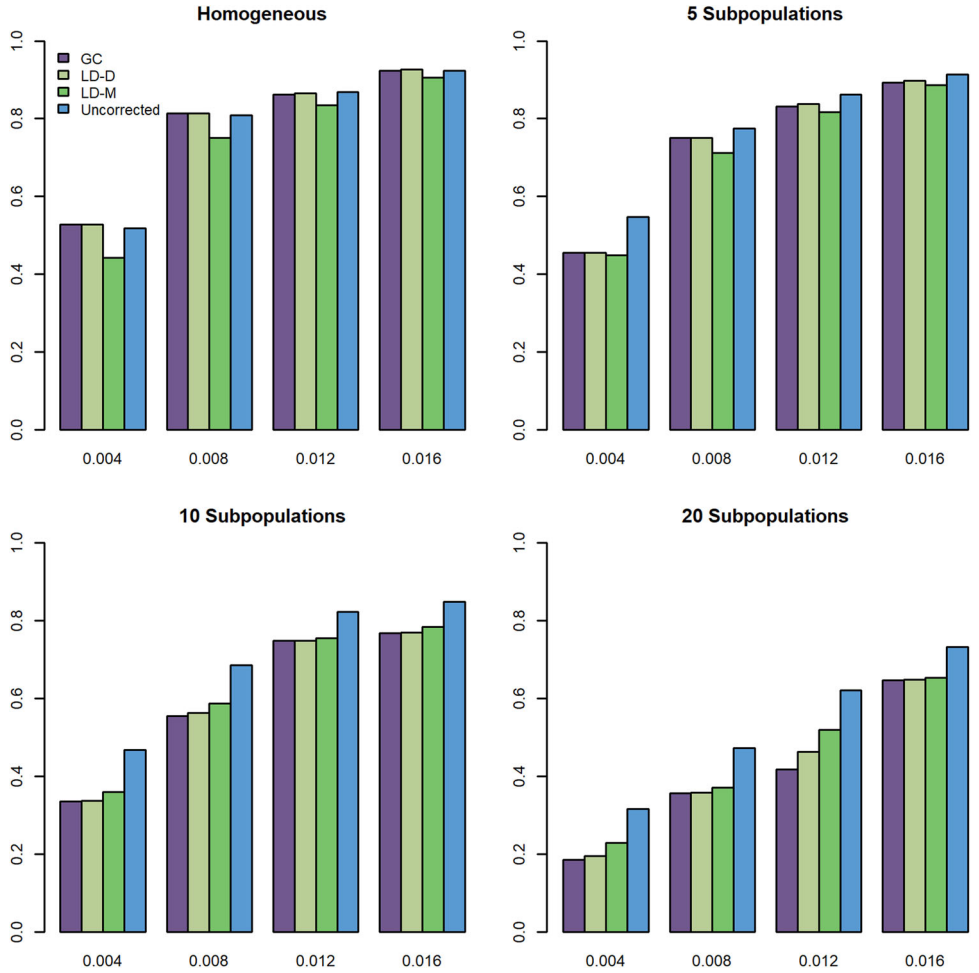

FIGURE 1 Power comparisons of GC, LD-D, LD-M, and uncorrected test in the simulation set 1 for a quantitative trait with *Heritability* = 0.004, 0.008, 0.012, and 0.016. GC, genomic control; LD-D, every χ^2 statistic will be corrected by dividing the intercept; LD-M, every χ^2 statistic will be corrected by minus the intercept.

small; and (3) the power of the uncorrected test is always the highest, but its Type I error rates are inflated.

In summary, (1) LD-D can control Type I error rates for all simulation scenarios and it is also more powerful than GC; (2) GC cannot control Type I error rates under some simulation scenarios; (3) LD-M cannot control Type I error rates for more simulation scenarios than GC; and (4) uncorrected test cannot control Type I error rates for all simulation scenarios that have population stratifications or cryptic relatedness, although it has the largest power.

3.3 | Real data analysis

3.3.1 | Application to GAW19

The first data set we use to conduct our analyses is a combination of true genotypes and simulated hypertension phenotypes across 849 Mexican-American individuals who are part of 20 separate pedigrees and provided as part of the GAW19. This data set is based on the family-based design with related individuals. In this data set, there are two related phenotypes, systolic blood pressure and diastolic blood pressure (DBP) at three time points, with 200 replicates. We consider the average of DBP at three time points as the phenotype of interest in our analyses (Zhu et al., 2016).

To evaluate the performance of these four methods, uncorrected, LD-M, LD-D, and GC, under the scenario of cryptic relatedness in real data sets, we evaluate type I error rates of these methods based on four SNP sets obtained from the data set of GAW19. We first randomly select 1000, 5000, 10,000, and 15,000 SNPs on chromosome 15 that are far away from the simulated functional loci for DBP as four SNP sets. For each of the SNP sets (1000, 5000, 10,000, and 15,000), we apply the four methods to each of the 200 phenotypes and each SNP in

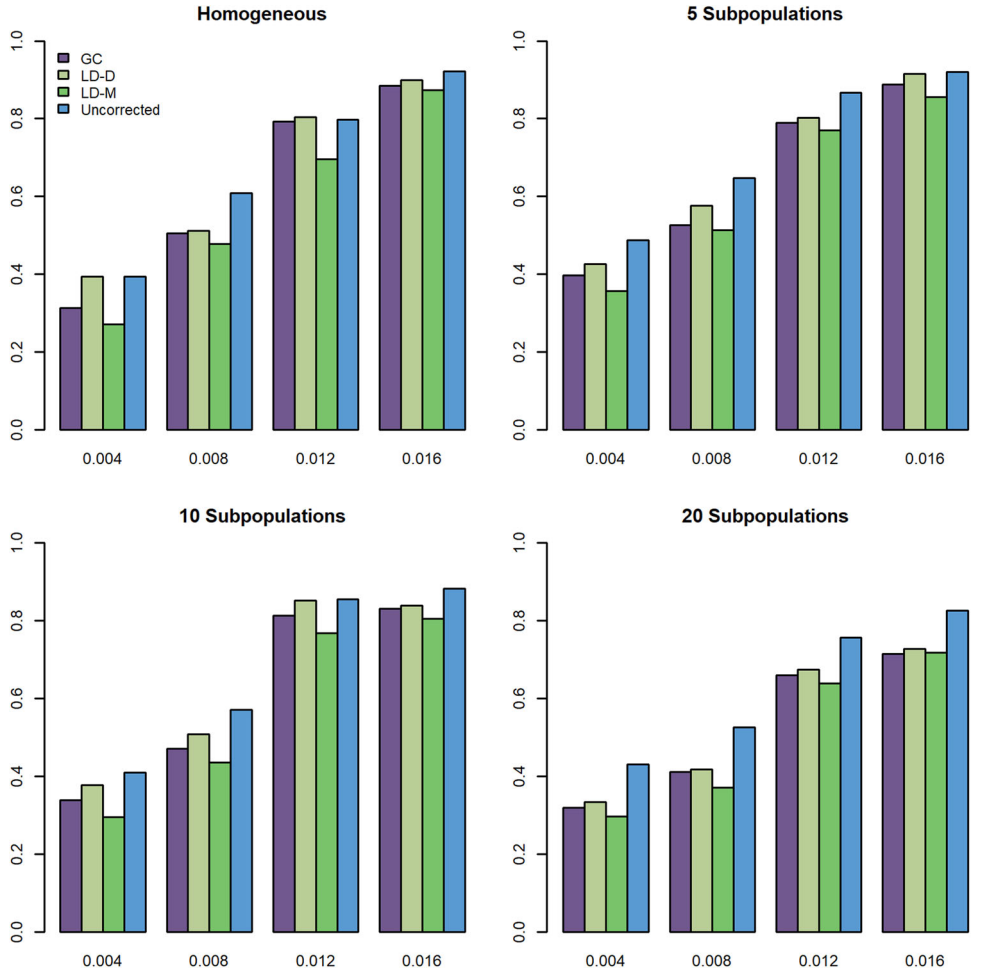

FIGURE 2 Power comparisons of GC, LD-D, LD-M, and uncorrected test in the simulation set 1 for a qualitative trait with *Heritability* = 0.004, 0.008, 0.012, and 0.016. GC, genomic control; LD-D, every χ^2 statistic will be corrected by dividing the intercept; LD-M, every χ^2 statistic will be corrected by minus the intercept.

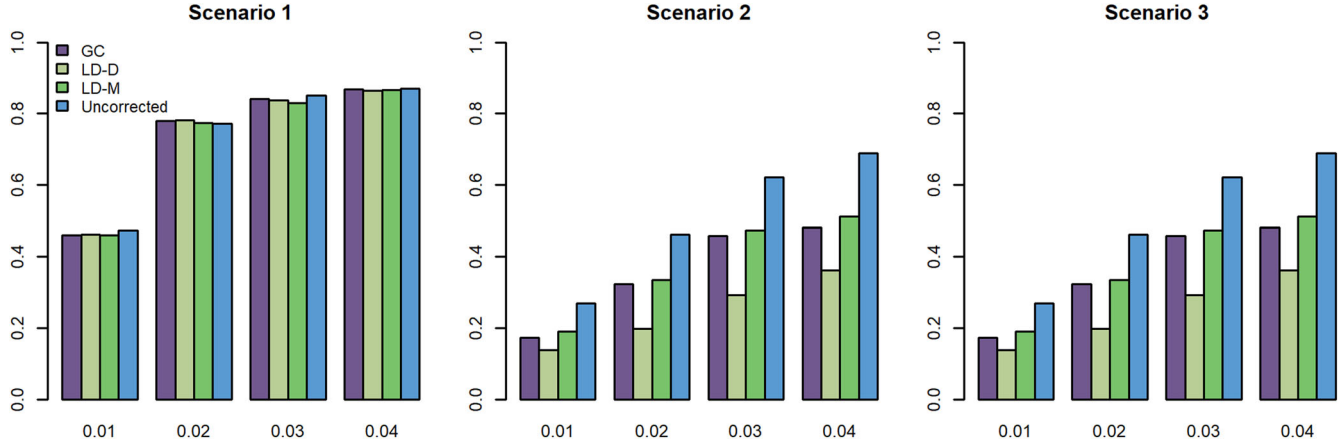

FIGURE 3 Power comparisons of GC, LD-D, LD-M, and uncorrected test in the simulation set 2 for a quantitative trait with *Heritability* = 0.01, 0.02, 0.03, and 0.04. GC, genomic control; LD-D, every χ^2 statistic will be corrected by dividing the intercept; LD-M, every χ^2 statistic will be corrected by minus the intercept.

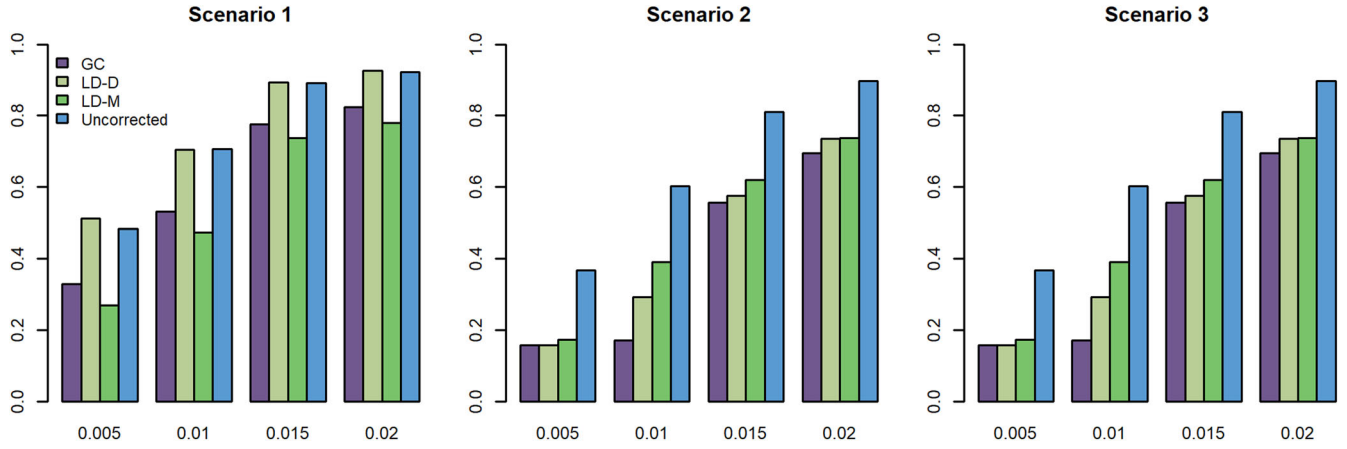

FIGURE 4 Power comparisons of GC, LD-D, LD-M, and uncorrected test in the simulation set 2 for a qualitative trait with *Heritability* = 0.005, 0.01, 0.015, and 0.02. GC, genomic control; LD-D, every χ^2 statistic will be corrected by dividing the intercept; LD-M, every χ^2 statistic will be corrected by minus the intercept.

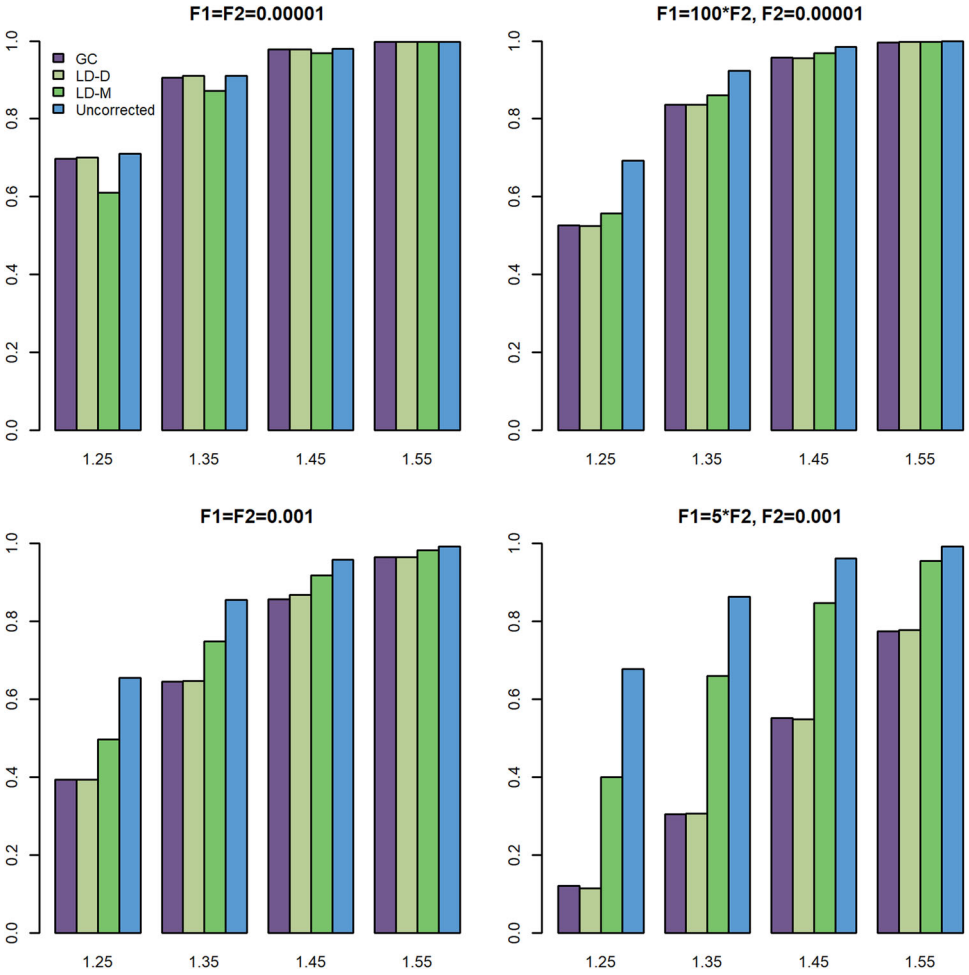

FIGURE 5 Power comparisons of GC, LD-D, LD-M, and uncorrected test in the simulation set 3 for a balanced case–control study with $\gamma = 1.25,\ 1.35,\ 1.45,$ and 1.55 and $n = 1000$. GC, genomic control; LD-D, every χ^2 statistic will be corrected by dividing the intercept; LD-M, every χ^2 statistic will be corrected by minus the intercept.

TABLE 4 Type I error rates divided by the nominal level 0.05 of uncorrected, GC, LD-D, and LD-M for the data set from GAW19

# of SNPs	Uncorrected	GC	LD-D	LD-M
1000	**2.142**	**1.095**	1.016	**1.175**
5000	**2.053**	**1.075**	0.992	**1.119**
10,000	**2.05**	**1.069**	0.9902	**1.118**
15,000	**2.056**	**1.039**	0.989	**1.114**

Note: Type I error rates in boldface indicate the values beyond the corresponding upper bound of the 95% CI divided by the nominal level 0.05.

Abbreviations: # of SNPs, number of SNPs sampled from noncausal SNPs; CI, confidence interval; GC, genomic control; LD-D, every χ^2 statistic will be corrected by dividing the intercept; LD-M, every χ^2 statistic will be corrected by minus the intercept.

an SNP set. For each SNP set, we consider 200 replicated phenotypes and SNPs in the SNP set as replicated samples to calculate the Type I error rate of each method. Table 4 summarizes the Type I error rates divided by the nominal level 0.05 of each method for each SNP set. From this table, we can see that uncorrected has inflated Type I error rates; the two methods to adjust for population stratification, LD-M and GC, still have inflated Type I error rates, and only LD-D can control Type I error rates.

3.3.2 | Application to UK Biobank GWAS of BMI

We also assess the performance of uncorrected, GC, LD-M, and LD-D using the UK Biobank data. Previous studies have highlighted that population structure within the United Kingdom is rather limited, but it occurs at a fine scale (e.g., birth location) on North–South and East–West clines (The Wellcome Trust Case Control Consortium, 2007; O'Dushlaine et al., 2010). Cook et al. (2020) have demonstrated that there is substantial inflation in GWAS with birth location and body mass index (BMI) is genetically correlated with birth location. At the same time, the fine-scale population structure in the UK Biobank GWAS of BMI cannot be fully accounted for by adjusting PCs (Cook et al., 2020).

To compare the performance of uncorrected, GC, LD-M, and LD-D for unrelated samples with fine-scale population structure, we utilize published association summary statistics for BMI available from Neale Lab. Full details of the quality control, phenotype derivation, and association analyses can be found at: https://www.nealelab.is/uk-biobank. The results of the GWAS are available for 359,983 unrelated individuals with European ancestry. There are 13,362,638 SNPs without missing information on chromosomes 1–22. With GWAS significance threshold

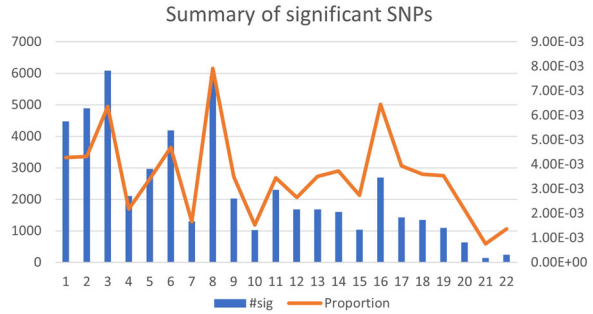

FIGURE 6 Summary of the significant SNPs associated with BMI in the UK Biobank. The *x*-axis represents the chromosome; the left *y*-axis is the number of significant SNPs; the right *y*-axis is the proportion of significant SNPs for each chromosome. BMI, body mass index; SNP, single-nucleotide polymorphism.

TABLE 5 Type I error rates divided by the nominal level 0.05 of uncorrected test, GC, LD-D, and LD-M for noncausal SNPs in UK Biobank

# of SNPs	Uncorrected	GC	LD-D	LD-M
1000	**1.6679**	**0.0735**	0.9919	**0.8577**
2000	**1.6806**	**0.0712**	0.9968	**0.8631**
3000	**1.6820**	**0.0730**	0.9974	**0.8633**

Note: Type I error rates in boldface indicate the values beyond the corresponding 95% CI divided by the nominal level 0.05.

Abbreviations: # of SNPs, number of SNPs sampled from noncausal SNPs; CI, confidence interval; GC, genomic control; LD-D, every χ^2 statistic will be corrected by dividing the intercept; LD-M, every χ^2 statistic will be corrected by minus the intercept.

5×10^{-8}, there are a total of 50,839 significant SNPs that are associated with BMI, which are summarized in Figure 6. Using GWAS summary statistics, we calculate the χ^2 statistics based on the *Z*-scores. GC inflation factor can be obtained based on all χ^2 statistics, which is 2.17. We also used LD scores computed from 1000 Genomes Project of the European sample, which is available from https://data.broadinstitute.org/alkesgroup/LDSCORE. The LDSC intercept is obtained based on 1,285,620 SNPs shared by both data sets, which is 1.25.

Based on Figure 6, we can see that chromosome 21 has the fewest significant SNPs and the smallest proportion of significant variants; therefore, we consider SNPs in chromosome 21 with $p > 0.005$ as noncausal SNPs. There are a total 172,534 noncausal SNPs. We randomly select 1000, 2000, and 3000 noncausal SNPs with 1000 replicates, and then we evaluate the Type I error rates of the four methods at a nominal significance level 0.05. The Type I error rates of the four methods are summarized in Table 5. From this table, we can see that only LD-D can control Type I error rates under all scenarios; uncorrected has inflated Type I error rates; GC and LD-M can control Type

I error rates, but are conservative; and GC is much more conservative compared with LD-M.

4 | DISCUSSION

For the development of statistical methods for association studies based on large sample GWAS summary statistics, it is not always feasible to do comprehensive quality control and correct for confounding biases without individual-level genotype and phenotype data. Typically, GC is a practicable method despite its limitations. On the other hand, the studies of LDSC have shown that the intercept of LDSC could provide a more robust inflation estimation than GCs. To explore how to use the intercept of LDSC to control for population stratification in genetic association studies, two methods, LD-D and LD-M, were investigated in this paper. We used three different simulation sets and applications to real data sets from GAW19 and UK Biobank to evaluate the performance of LD-D and LD-M with GC. In conclusion, LD-D has correct Type I error rates in all simulation scenarios and in the applications to the real data sets, which is proved to be a more reliable and accurate correction method than GC. LD-M and GC cannot control Type I error rates in some scenarios.

In the previous LDSC paper (Bulik-Sullivan et al., 2015), a potential limitation of LDSC was shown when variance explained per SNP may be correlated with LD score for some phenotypes, which result in the underestimation of confounding contributions estimated by the intercept of LDSC. Particularly, LD-M can control Type I error rates in the situation of spatially structured populations for unbalanced case–control studies indicating that the underestimation is due to the correlation between LD score and Wright F_{st} (Bulik-Sullivan et al., 2015). However, Lee et al. (2018) showed that dividing the intercept of LDSC is still a viable means of correcting confounding even in those cases, which is in accordance with the performance of LD-D. In conclusion, LD-D is a more reliable and stronger tool for controlling confounding bias in association studies than GC and LD-M.

AUTHOR CONTRIBUTIONS

Shuanglin Zhang and Qiuying Sha designed the research: Shijia Yan and Shuanglin Zhang performed the statistical analysis: and Shijia Yan, Qiuying Sha, and Shuanglin Zhang wrote the manuscript.

ACKNOWLEDGMENTS

The Genetic Analysis Workshops are supported by the National Institutes of Health (NIH) Grant R01 GM031575 from the National Institute of General Medical Sciences. The preparation of GAW 17 data was supported, in part, by NIH R01 MH059490 and used data from the 1000 Genomes Project (www.1000genomes.org). The GAW19 whole-genome sequence data were provided by the T2D-GENES (Type 2 Diabetes Genetic Exploration by Next-generation sequencing in Ethnic Samples) Consortium, which is supported by NIH Grants U01 DK085524, U01 DK085584, U01 DK085501, U01 DK085526, and U01 DK085545. The other genetic and phenotypic data for GAW19 were provided by the San Antonio Family Heart Study and San Antonio Family Diabetes/Gallbladder Study, which are supported by NIH Grants P01 HL045222, R01 DK047482, and R01 DK053889.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Qiuying Sha http://orcid.org/0000-0002-9342-3269
Shuanglin Zhang http://orcid.org/0000-0002-9478-1199

REFERENCES

Balding, D. J., & Nichols, R. A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, *96*(1–2), 3–12.

Bhatia, G., Patterson, N., Sankararaman, S., & Price, A. L. (2013). Estimating and interpreting FST: The impact of rare variants. *Genome Research*, *23*(9), 1514–1521.

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Price, A. L., & Neale, B. M. (2015). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, *47*(3), 291–295.

Chen, H. S., Zhu, X., Zhao, H., & Zhang, S. (2003). Qualitative semi-parametric test for genetic associations in case–control designs under structured populations. *Annals of Human Genetics*, *67*(3), 250–264.

Cook, J. P., Mahajan, A., & Morris, A. P. (2020). Fine-scale population structure in the UK Biobank: Implications for genome-wide association studies. *Human Molecular Genetics*, *29*(16), 2803–2811.

Devlin, B., & Roeder, K. (1999). Genomic control for association studies. *Biometrics*, *55*(4), 997–1004.

Devlin, B., Roeder, K., & Wasserman, L. (2001). Genomic control, a new approach to genetic-based association studies. *Theoretical Population Biology*, *60*(3), 155–166.

Jiang, Y., Epstein, M. P., & Conneely, K. N. (2013). Assessing the impact of population stratification on association studies of rare variation. *Human Heredity*, *76*(1), 28–35.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-y, Freimer, N. B., & Sabatti, C. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, *42*(4), 348–354.

Knowler, W. C., Williams, R. C., Pettitt, D. J., & Steinberg, A. G. (1988). Gm3;5,13,14 and type 2 diabetes mellitus: An association in American Indians with genetic admixture. *American Journal of Human Genetics*, *43*, 520–526.

Lander, E. S., & Schork, N. J. (1994). Genetic dissection of complex traits. *Science*, *265*, 2037–2048.

Lee, J. J., McGue, M., Iacono, W. G., & Chow, C. C. (2018). The accuracy of LD score regression as an estimator of confounding and genetic correlations in genome-wide association studies. *Genetic Epidemiology*, *42*(8), 783–795.

Mathieson, I., & McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics*, *44*(3), 243–246.

O'Dushlaine, C. T., Morris, D., Moskvina, V., Kirov, G., International Schizophrenia Consortium, Gill, M., Corvin, A., & Cavalleri, G. L. (2010). Population structure and genome-wide patterns of variation in Ireland and Britain. *European Journal of Human Genetics*, *18*(11), 1248–1254.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959.

Pritchard, J. K., Stephens, M., Rosenberg, N. A., & Donnelly, P. (2000). Association mapping in structured populations. *The American Journal of Human Genetics*, *67*(1), 170–181.

Reich, D. E., & Goldstein, D. B. (2001). Detecting association in a case–control study while correcting for population stratification. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, *20*(1), 4–16.

Sha, Q., Zhang, Z., & Zhang, S. (2011). An improved score test for genetic association studies. *Genetic Epidemiology*, *35*(5), 350–359.

Sha, Q., Zhang, K., & Zhang, S. (2016). A nonparametric regression approach to control for population stratification in rare variant association studies. *Scientific Reports*, *6*, 37444.

The Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*(7145), 661–678.

Voight, B. F., & Pritchard, J. K. (2005). Confounding from cryptic relatedness in case–control association studies. *PLoS Genetics*, *1*(3), e32.

Yang, J., Weedon, M. N., Purcell, S., Lettre, G., Estrada, K., Willer, C. J., Smith, A. V., O'Connell, J. R., Mangino, M., Mägi, R., Madden, P. A., Heath, A. C., Nyholt, D. R., Martin, N. G., Montgomery, G. W., Frayling, T. M., Hirschhorn, J. N., McCarthy, M. I., Goddard, M. E., & Visscher, P. M., GIANT Consortium. (2011). Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics*, *19*(7), 807–812.

Zhang, S., & Zhao, H. (2001). Quantitative similarity-based association tests using population samples. *The American Journal of Human Genetics*, *69*(3), 601–614.

Zhang, S., Zhu, X., & Zhao, H. (2003). On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, *24*(1), 44–56.

Zhang, Y., Guan, W., & Pan, W. (2013). Adjustment for population stratification via principal components in association analysis of rare variants. *Genetic Epidemiology*, *37*(1), 99–109.

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Arnett, D. K., Ordovas, J. M., & Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, *42*(4), 355–360.

Zhu, H., Wang, Z., Wang, X., & Sha, Q. (2016). A novel statistical method for rare-variant association studies in general pedigrees. *BMC Proceedings*, *10*(Suppl 7), 22.

How to cite this article: Yan, S., Sha, Q., & Zhang, S. (2022). Control for population stratification in genetic association studies based on GWAS summary statistics. *Genetic Epidemiology*, 46, 604–614. https://doi.org/10.1002/gepi.22493