A Nonparametric Regression Approach to Control for Population Stratification in Rare Variant Association Studies

Qiuying Sha, Kui Zhang & Shuanglin Zhang

Recently, there is increasing interest to detect associations between rare variants and complex traits. Rare variant association studies usually need large sample sizes due to the rarity of the variants, and large sample sizes typically require combining information from different geographic locations within and across countries. Although several statistical methods have been developed to control for population stratification in common variant association studies, these methods are not necessarily controlling for population stratification in rare variant association studies. Thus, new statistical methods that can control for population stratification in rare variant association studies are needed. In this article, we propose a principal component based nonparametric regression (PC-nonp) approach to control for population stratification in rare variant association studies. Our simulations show that the proposed PC-nonp can control for population stratification well in all scenarios, while existing methods cannot control for population stratification at least in some scenarios. Simulations also show that PC-nonp’s robustness to population stratification will not reduce power. Furthermore, we illustrate our proposed method by using whole genome sequencing data from genetic analysis workshop 18 (GAW18).

Recently, there is increasing interest to detect associations between rare variants and complex traits. The variant by variant methods used to detect associations of common variants may not be optimal for detecting associations of rare variants due to allelic heterogeneity as well as the extreme rarity of individual variants1. Many statistical methods for testing the association of rare variants have been developed by using joint information of multiple variants in a genomic region. These methods can be roughly divided into three groups: burden tests, quadratic tests, and combined tests.

Burden tests1–5 collapse rare variants in a genomic region into a single burden variable and then regress the phenotype on the burden variable to test for the cumulative effects of rare variants in the region6. Burden tests implicitly assume that all rare variants are causal and directions of effects are all the same. Quadratic tests include tests with statistics of quadratic form of score vector7–9 and also adaptive weighting methods10–13. Quadratic tests are robust to directions of effects of causal variants and are less affected by neutral variants than burden tests do. If most of the rare variants are causal and directions of effects of causal variants are all the same, burden tests can outperform quadratic tests; otherwise, quadratic tests perform better. Combined tests6,14 combine information from burden tests, quadratic tests, and possibly other tests aiming to have advantages of multiple tests and to increase the robustness of tests.

All the aforementioned methods are population-based methods for unrelated individuals. It has been long recognized that, for population-based association studies, population stratification can seriously confound association results15,16. For rare variants this problem can be more serious, because the spectrum of rare variations can be very different in different populations. In common variant association studies, several methods that use a set of genomic markers genotyped in the same samples have been developed to control for population stratification. These methods include genomic control (GC) approach17–19, principal component (PC) based linear regression (PC-linear) approach20, and mixed linear model (MLM) approach21,22 among others. GC approach adjusts the ordinary chi-square test statistic $X^2$ to $X^2/\lambda$ and assumes $X^2/\lambda$ follows a chi-square distribution, where

Department of Mathematical Sciences, Michigan Technological University, Houghton, MI 49931, USA. Correspondence and requests for materials should be addressed to S.Z. (email: shuzhang@mtu.edu)
the inflation factor $\lambda$ can be estimated using genotypes at genomic markers. PC-linear approach summarizes the genetic background or ancestry information through the PCs of genotypes at genomic markers. The PCs can be further used to eliminate the effect resulting from population stratification through linear regressions. MLM approach corrects for a wide range of sample structures by explicitly accounting for pairwise relatedness between individuals.

Although several methods for controlling for population stratification have been developed for common variants, it remains unclear whether these methods are equally effective for rare variants. Because rare variants have typically arisen recently, they tend to show greater geographic clustering or more latent subpopulations than common variants that are typically older. The more geographic clusters or latent subpopulations, the more difficult it will be to control for population stratification. Mathieson and McVean29 demonstrated that rare variants can show a stratification that is systematically different from common variants. They also demonstrated that the commonly used methods such as GC, PC-linear, and MLM to control for population stratification in common variant associations are not necessarily controlling for population stratification in rare variant associations. Zhang et al.24 showed that the use of PCs calculated from common variants were effective to control for population stratification in rare variant associations. Jiang et al.30 also found that the PC based methods performed quite well while GC often yielded lower power. Note that both studies of Zhang et al.24 and Jiang et al.30 did not explicitly model the spatial structure of populations in their simulation studies. Zhang et al.24 used two continental groups from the 1000 Genomes Project with six and four subpopulation groups, respectively. Jiang et al.25 simulated data with two populations. Lissgarten et al.26 reported that FaST-LMM Select (a MLM approach) could control for population stratification when samples were from spatially structured populations. However, their approach reduced power substantially when causal rare variants are spatially clustered.

In this article, we propose a PC based nonparametric regression (PC-nonp) approach to control for population stratification in rare variant association studies. PC-nonp adjusts population effects of both trait values and genotypes at candidate loci for PCs of genotypes at genomic markers by applying nonparametric regressions. We use extensive simulation studies to evaluate the performance of the proposed method PC-nonp and compare the performance of PC-nonp with that of GC and PC-linear developed for common variants and recently proposed biased urn permutation test (BiasePerm)28 developed for rare variants. Simulation results show that PC-nonp can control for population stratification well in all scenarios while GC, PC-linear, and BiasedPerm cannot control for population stratification at least in some scenarios. Results also show that PC-nonp’s robustness to population stratification will not reduce power. Furthermore, we evaluate the performance of our approach by applying it to the whole genome sequencing data from genetic analysis workshop 18 (GAW18) and find that only PC-nonp is effective to control for population stratification.

**Method**

Consider a sample of $n$ unrelated individuals. Suppose that each individual has been genotyped at a candidate locus (single variant or multiple variants) and at $L$ genomic markers. Let $y_i$, $x_i$, and $p_i$ denote the trait value, genotypic score at the candidate locus (weighted sum of genotypic scores if there are multiple variants), and the first $k$ PCs (rescaled to the interval $[0, 1]$) of genotypes at genomic markers of the $i^{th}$ individual. The PCs of genotypes at genomic markers are good summary measures of ancestry or genetic background. PC-linear is probably the most popular method to control for population stratification. However, this method is based on linear combinations of PCs. Furthermore, recently developed BiasePerm28 is based on linear combinations of PCs on logistic scale if we use PCs as covariate vector29. The relationships between trait values and PCs can be highly nonlinear and the forms of the relationships are different in different populations. When the relationships are highly nonlinear and the forms of relationships are unknown, we should use more flexible regression methods rather than use linear regression. Nonparametric regression is a very flexible regression method and it does not require the form of regression function.

In this article, we propose a PC based nonparametric regression (PC-nonp) approach that adjusts population effects of both trait values and genotypes at candidate loci for PCs of genotypes at genomic markers by applying nonparametric regressions. That is,

$$y_i = \mu_1(p_i) + \epsilon_i \quad \text{and} \quad x_i = \mu_2(p_i) + \epsilon_i,$$

(1)

where $\mu_1(\cdot)$ and $\mu_2(\cdot)$ are regression functions with unknown forms and will be estimated using smoothing techniques. Let $y^*$ and $x^*$ be the residuals of the nonparametric regressions. We can consider $y^*$ and $x^*$ as the trait value and genotypic score at the candidate locus of the $i^{th}$ individual after adjusting for population effects. We can construct association tests based on the residuals.

Many methods have been developed to estimate the unknown regression function, including local linear method29–31, kernel smoothing method32,33, and wavelet method34,35. We propose to use kernel smoothing method. Let $K(\cdot)$ be a kernel function with mode at 0. The kernel estimators of $\mu_1(p_i)$ and $\mu_2(p)$ are given by

$$\hat{\mu}_1(p_i) = \frac{1}{n} \sum_{j=1}^{n} y_j \frac{K(\frac{p_j - p_i}{H})}{H} \quad \text{and} \quad \hat{\mu}_2(p) = \sum_{j=1}^{n} x_j \frac{K(\frac{p_j - p}{H})}{H},$$

(2)

respectively, where $p_j = (p_{j1}, \ldots, p_{jk})$ is the first $k$ PCs for the $i^{th}$ individual, $H = (h_1, \ldots, h_k)$ is the smoothing parameter, and $p_j - p = \left(\frac{p_{j1} - p_1}{h_1}, \ldots, \frac{p_{jk} - p_k}{h_k}\right)$. If we denote $W_j(p) = K(\frac{p_j - p_i}{H})\sum_{l=1}^{n} K(\frac{p_l - p_i}{H})$, then
\[ \mu = \sum_{i,j} \hat{p}_W p(W_{ij}) \] and \[ \mu = \sum_{i,j} \hat{p}_W p(W_{ij}). \] With these nonparametric estimators, the fitted values of trait and the fitted values of genotypic scores at the candidate locus are given by \[ \hat{y} = \sum_{i,j} \hat{p}_W p(W_{ij}) y_{ij} \] and \[ \hat{x} = \sum_{i,j} \hat{p}_W p(W_{ij}) x_{ij}. \] Intuitively, \[ \hat{y} = \sum_{i,j} \hat{p}_W p(W_{ij}) y_{ij} \] and \[ \hat{x} = \sum_{i,j} \hat{p}_W p(W_{ij}) x_{ij} \] are the weighted mean of trait values and weighted mean of genotypic scores of those individuals whose genetic background is similar to that of the \( i \)th individual. Thus, we can consider residuals \[ \hat{y} = \sum_{i,j} \hat{p}_W p(W_{ij}) y_{ij} \] and \[ \hat{x} = \sum_{i,j} \hat{p}_W p(W_{ij}) x_{ij} \] as the trait value and genotypic score of the \( i \)th individual after adjusting for population stratification.

In this study, we use the quartic kernel\(^3\),

\[
k(t) = \begin{cases} \frac{15}{16} (1 - t^2)^2, & |t| \leq 1, \\ 0, & |t| > 1. \end{cases}
\] (3)

For computational consideration, we assume that \( h_1 = \ldots = h_k = h \). Then,

\[
K \left( \frac{P - P_{ij}}{H} \right) = \prod_{i=1}^{k} \left( \frac{P - P_{ij}}{h} \right).
\]

To test association between trait values and genotypes based on \( y^* \) and \( x^*_i \), we can use score test with test statistic \( T_{score} = U^2/V \), where \( U = \sum_{i,j} (y^*_i - \overline{y})(x^*_i - \overline{x}) \) and \( V = \sum_{i,j} (y^*_i - \overline{y})^2 \sum_{i,j} (x^*_i - \overline{x})^2 \). The statistic \( T_{score} \) asymptotically follows a chi-square distribution with one degree of freedom (df)\(^3\). For rare variants, \( x_i \) can be a weighted combination\(^2\) or collapsing\(^1\) of genotypes at multiple variants in a genomic region. Based on the residuals of the nonparametric regression, we can construct other rare variant association tests such as CMC\(^1\), SKAT\(^9\), and TOW\(^8\). We will discuss this issue in more details later in the discussion section. In this study, we use a single-variant test in which \( x_i \) is the genotypic score of a single variant and a regional test in which \( x_i \) is the weighted combination of genotypes at the variants in a genomic region\(^2\) to evaluate the performance of our proposed method.
We have so far assumed a given smoothing parameter in the kernel estimates. It is well known that choosing a proper value for smoothing parameter \( h \) is critical to kernel estimates of regression functions\(^{35-37} \). We use a method similar to that of Zhang et al.\(^{35} \) to choose smoothing parameter \( h \). This method is based on the genotypes at a set of genomic markers. Suppose there are \( L \) genomic markers. We perform PC-nonp single-variant test for all the \( L \) genomic markers and denote \( P_1, \ldots, P_L \) as the associated \( P \)-values. If population stratification is well controlled for, \( P \)-values \( P_1, \ldots, P_L \) should follow a uniform distribution under the null hypothesis of no association. Let \( F_\lambda \) be the empirical distribution function of the \( P \)-values \( P_1, \ldots, P_L \) and \( F \) be the uniform distribution function. The Kolmogorov test statistic \( \text{Kol}(h) = \max |F_{\lambda}(x) - F(x)| \) measures how close the distribution of the \( P \)-values \( P_1, \ldots, P_L \) and the uniform distribution are. We propose to choose \( h^* \) that minimizes the Kolmogorov test statistic, i.e.,

\[
h^* = \arg \min_h \text{Kol}(h)
\]

as the value of the smoothing parameter. \( h^* \) can be obtained by a simple grid search across a range of \( h \). We divide the interval \([0, \infty]\) into subintervals \( 0 \leq h_0 < \ldots < h_{k-1} < h_k < \infty \). Then, \( h^* = \arg \min_h \text{Kol}(h) \). The computational time to find \( h^* \) increases linearly with \( S \). However, \( h^* \) needs to be calculated only once. We can use this \( h^* \) to calculate the residuals of the nonparametric regression for trait values and genotypes at each variant. Let \( k \) denote the number of PCs used. In this study, we use \( h_i = 2^{2(i-23)/30+k} \), where \( s = 1, \ldots, 30 \) and \( k = 10 \). It is worth noting that the smoothing parameter \( h \) is chosen with the \( P \)-values of a single-variant test, whichever test is actually used in testing associations.

**Software.** R code for implementing our proposed method is given at Shuanglin Zhang’s homepage http://www.math.mtu.edu/~shuzhang/software.html. The R code includes three functions: PCA, choose_OPT_SMPL and Resid_Nonp. PCA gives the first \( k \) principal components of genotypes at genomic markers. choose_OPT_SMPL chooses the optimal value of smoothing parameter. Given the value of the smoothing parameter, Resid_Nonp calculates the residuals of trait values and genotypes at a candidate region by applying nonparametric regression for PCs of genotypes at genomic markers.

**Comparison of Tests.** We compare the performance of the proposed test with that of the following four tests. (1) Uncorrected: this test is also based on the score test statistic \( T^U_{\text{score}} \), \( T^U_{\text{score}} \) is the same as \( T_{\text{score}} \), but \( T^U_{\text{score}} \) is based on the original trait values \( y_i \) and genotypic scores \( x_{ij} \) instead of based on the residuals. (2) GC**: GC divides \( T^U_{\text{score}} \) by an inflation factor \( \lambda \) and \( \lambda = \text{median}(T^U_{\text{score}}(1), \ldots, T^U_{\text{score}}(L))/0.456 \), where \( T^U_{\text{score}}(i) \) is the value of \( T^U_{\text{score}} \) when \( T^U_{\text{score}} \) is applied to the \( i \)-th genomic marker. (3) PC-linear**: in this permutation procedure, the odds of a subject being selected as a case are equal to his or her odds of disease conditional on confounder variables. In this study, PC-linear, PC-nonp, and BiasedPerm are based on the first 10 PCs of genotypes at the genomic markers.

**Simulations.** We consider two sets of simulations: populations with \( k_0 \) subpopulations and populations with spatially structured populations. In each set of simulations, we consider both qualitative and quantitative traits. To generate a qualitative disease affection status, we use a liability threshold model based on a continuous phenotype (quantitative trait). An individual is defined to be affected if the individual’s phenotype is at least one standard deviation larger than the phenotypic mean. This yields a prevalence of 16% for the simulated disease in the general population. In the following, we describe how to generate genotypes and how to generate a quantitative trait in the two sets of simulations.

**Simulation Set 1: Populations with \( k_0 \) Subpopulations.** This set of simulations is based on allele frequencies at 24,487 variants calculated from the empirical Mini-Exome genotype data provided by the genetic analysis workshop 17 (GAW17). The genotypes of GAW17 data set are extracted from the sequence alignment files provided by the 1000 Genomes Project for their pilot3 study (http://www.1000genomes.org). GAW17 data contain genotypes of 697 unrelated individuals at 24,487 variants. The distributions of MAF at rare variants (MAF < 0.01) and MAF at common variants of 24,487 variants are given in Figure S1.

To generate genotypes of individuals in a population with \( k_0 \) subpopulations, we follow Price et al.,\(^{20} \) Ikonita-Laza et al.,\(^{38} \) and Qin et al.\(^{39} \) For each variant, we randomly select a variant from 24,487 variants and take the MAF at this variant as the ancestral population allele frequency \( p \). Then, independently draw \( k_0 \) values \( p_1, \ldots, p_{k_0} \) from a beta-distribution with parameters \( (1-F_\lambda)/F_\lambda \) and \( (1-p)(1-F_\lambda)/F_\lambda \) where \( F_\lambda \) is the Wright’s measure of population subdivision\(^{40} \) in this study, \( F_\lambda = 0.01 \). For each variant, we accept \( p_1, \ldots, p_{k_0} \) as allele frequencies for the \( k_0 \) subpopulations if \( \frac{1}{k_0} \sum_{i=1}^{k_0} p_i > 0.002 \); we redraw \( p_1, \ldots, p_{k_0} \) otherwise. The MAF distributions at the rare variants (MAF < 0.01) and at the common variants for \( k_0 = 5 \) are given in Figure S1.

To evaluate type I error, we generate trait values independent of genotypes by using the model:

\[
y_{ij} = \mu_i + \epsilon_{ij},
\]

where \( y_{ij} \) denotes the trait value of the \( j \)-th individual in the \( i \)-th subpopulation, \( \epsilon_{ij} \) follows a standard normal distribution, and \( \mu_i \) is the population mean of the \( i \)-th subpopulation. In this study, if \( k_0 \leq 2 \), we set \( \mu_1 = 0 \) and \( \mu_2 = \mu \); otherwise, we set \( \mu_1 = \cdots = \mu_{k_0-2} = 0 \) and \( \mu_{k_0-1} = \mu_{k_0} = \mu \), where \( \mu = 5 \) if \( k_0 = 20 \); otherwise \( \mu = 2 \).
To evaluate power, we consider $n_T$ variants (possibly both rare and common variants) in a genomic region. We randomly choose $n_i$ from the $n_T$ variants as causal variants (in this study, $n_i = n_T / 2$). For the $i^{th}$ individual in the $j^{th}$ subpopulation, let $x_{ij}$ denote the genotypic score at the $i^{th}$ variant. For the $j^{th}$ individual in the $i^{th}$ subpopulation at the $k^{th}$ causal variant. We assume that all the $n_i$ causal variants have the same heritability such that rarer variants have larger effects. Under this assumption, the disease model is given by

$$y_{ij} = \mu_i + \sum_{l=1}^{n_T} \beta_l x_{ij} + \epsilon_{ij},$$  

where $\beta_l$ are constants and their values depend on the total heritability.

**Simulation Set 2: Spatially Structured Populations.** We generate genotypes and phenotypes under spatially structured populations using the methods similar to those of Mathieson and McVean. Briefly, the space is divided into $K_0 \times K_0$ grid squares. Then, we generate genotypes by starting with a number of individuals and their locations on the grid. We work backward in time to generate random genealogical events. Each event is either a coalescence of two lineages or a migration of a single lineage from one square to another. 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 (see Supplemental materials or Mathieson and McVean for details).

To generate quantitative traits under null hypothesis, let $\phi_i = [1, n] \rightarrow [1, K_0] \times [1, K_0]$ be a function that maps each individual to the grid square from which they originated. Then, we generate the trait value of the $i^{th}$ individual by $y_{i} = \beta_{R,0} + \epsilon_{i}$, where $\phi(i) = (l, j)$ if the $i^{th}$ individual originates from grid square $l, j$. $R_{ij}$ is the nongenetic risk in grid square $l, j$, $\epsilon_i$ is a standard normal random number, and $\beta$ is a constant. We use the following three models to determine the value of $R_{ij}$. Model 0: no population stratification in which $R_{ij} = 0$ for all $i$ and $j$. Model 1: a small and sharp spatial distribution in which $R_{ij} = 1$ if $l_0 \leq l \leq l_0 + 3$ and $j_0 \leq j \leq j_0 + 3$ for $l_0 = j_0 = 6$, or $20 - l_0 = j_0 = 6$, or $l_0 = j_0 + 14$; $R_{ij} = 0$ otherwise. Model 2: a wide and smooth spatial distribution in which $R_{ij} = 0.4e^{-((l-l_0)^2 + (j-j_0)^2)/14}$ for $l_0 = j_0 = 6$. In this study, we use the following parameters: $K_0 = 20$, $M = 0.01$, and $\beta = 2$.

Under alternative hypothesis, we assume that there are $n_T$ variants in a genomic region. We randomly choose $n_i$ from the $n_T$ variants as causal variants. For an individual, let $x_i$ denote the genotypic score at the $i^{th}$ causal variant. Under the assumption that all the $n_i$ causal variants have the same heritability, the trait value for an individual is generated by

$$y = \sum_{l=1}^{n_T} \beta_{l} x_{l} + y_0,$$

where $y_0$ is the trait value generated under null hypothesis.

**Results**

**Existence of the minimum of Kolmogorov test statistic $Kol(h)$.** We first perform simulation studies to evaluate the existence of the minimum of Kolmogorov test statistic $Kol(h)$. We generate trait values and genotypes at 10,000 variants under simulation set 1 for $k_0 = 5$ and $k_0 = 10$ and under simulation set 2 models 1 and 2. Under each of the four scenarios, we calculate $Kol(h)$ for different values of $h$. The relationships between $Kol(h)$ and $-\log(h)$ under the four scenarios are given in Fig. 2. This figure shows that the curves of $Kol(h)$ under the four scenarios are all bowed shaped and thus have minimum. The histograms of 10,000 $P$-values of the proposed test for different values of $h$ are given in Figures S2–S5 for the four scenarios, respectively. From these figures, we can see that when $h$ is large, population effects are not adjusted enough and thus the number of small $P$-values are more than expected; when $h$ is small, population effects are over adjusted and thus the number of large $P$-values are more than expected; when $h$ minimizes $Kol(h)$, the distribution of $P$-values is very close to the uniform distribution.

**Evaluate type I error rates.** We use 10,000 replicated samples to evaluate type I error rates. For BiasedPerm, we use 5,000 permutations to evaluate $P$-values. For all other tests, we use asymptotic distributions to evaluate $P$-values. For 10,000 replicated samples, the 95% confidence intervals (CIs) for type I error rates of nominal levels 0.01 and 0.001 are (0.008, 0.012) and (0.00037, 0.00163), respectively.

To evaluate type I error rates, we first want to see the performance of the asymptotic distributions we used. For this purpose, we perform simulations under null hypothesis in a homogenous population ($k_0 = 1$ in simulation set 1) and in the case of no population stratification (model 0 in simulation set 2). Type I error rates are given in Tables 1 and 2 for quantitative traits and qualitative traits, respectively. Table 1 shows that, for quantitative traits, type I error rates of all the four tests in all the scenarios are within the corresponding 95% confidence intervals, which indicates that the asymptotic distributions work very well. Table 2 shows that, for qualitative traits, most of the type I error rates are within the corresponding 95% CIs and those of the type I error rates that are not in the 95% CIs are very close to the corresponding 95% CIs, which indicates that the asymptotic distributions approximately work well.

Type I error rates under structured populations in simulation set 1 for $k_0 = 2$, 10, 20 are given in Tables 3 and 4 for quantitative traits and qualitative traits, respectively. As shown by these two tables, Uncorrected has inflated type I error rates in all the scenarios. GC cannot control for population stratification for quantitative traits when $k_0 = 10$ and 20 because most variants have very small correlation with the trait. PC-linear and BiasedPerm cannot control for population stratification when $k_0 = 20$ because the linear combinations of the first 10 PCs cannot
discriminate 20 subpopulations. Only PC-nonp can control for population stratification in all simulation scenarios. If we increase the number of PCs, PC-linear and BiasedPerm may control for population stratification when \( k_0 = 20 \). The problems to use PC-linear and BiasedPerm to control for population stratification are (1) we do not know how many PCs should be used and (2) increasing the number of PCs may decrease the power.

Type I error rates under spatially structured populations in simulation set 2 for models 1 and 2 are given in Tables 5 and 6 for quantitative traits and qualitative traits, respectively. These two tables show that Uncorrected has inflated type I error rates in all the scenarios. GC cannot control for population stratification for single variant test because most variants have very small correlation with the trait. PC-linear and BiasedPerm have inflated type I error rates under model 1 because these two methods try to correct highly nonlinear relationships on the basis of linear functions of relatedness. PC-nonp can control for population stratification well in all simulation scenarios because nonparametric regressions can adapt any function, linear or nonlinear.

Power comparison. To evaluate if PC-nonp's robustness to population stratification will reduce power, we perform simulation studies to compare power using regional tests under \( k_0 = 1 \) and \( k_0 = 10 \) in simulation set 1.

![Figure 2. The relationships between \(-\log(\text{smoothing parameter } h)\) and Kolmogorov test statistic Kol(h) in four structured populations.](image)

### Table 1. Type I error rates of four tests in a homogenous population and in the case of no population stratification for quantitative traits.

| Population Structures | Test         | Alpha | Un-corrected   | GC       | PC-linear | PC-nonp |
|-----------------------|--------------|-------|----------------|----------|-----------|---------|
| A Homogenous Population | Single Variant | 0.01  | 0.01054        | 0.01104  | 0.01096   | 0.01106  |
|                       | Regional     | 0.01  | 0.0103         | 0.01112  | 0.01078   | 0.01048  |
| No Population Stratification | Single Variant | 0.01  | 0.00086        | 0.00112  | 0.00088   | 0.00096  |
|                       | Regional     | 0.01  | 0.00942        | 0.00974  | 0.00936   | 0.0096   |
|                       |              | 0.001 | 0.00074        | 0.00078  | 0.00076   | 0.00072  |
and under models 0 and 2 in simulation set 2, in which all tests except Uncorrected can control for population stratification well. Power comparisons under $k_0 = 1$ and $k_0 = 10$ in simulation set 1 are given in Fig. 3. This figure shows that, when there is no population stratification (a homogenous population), all tests have very similar powers. When there is population stratification (a structured population with 10 subpopulations), PC-nonp and PC-linear are more powerful than Uncorrected and BiasedPerm, and GC has the lowest power. GC loses power because it has a larger inflation factor when there is population stratification. BiasedPerm essentially performs permutation within subpopulations and thus it will lose power when there are a large number of subpopulations. Uncorrected loses power because, in the structured population with 10 subpopulations, different trait value means in subpopulations weaken the association signal. PC-nonp and PC-linear do not lose power because, after adjusted for population effects, it appears that PC-nonp and PC-linear perform association tests in a homogenous population.

Table 2. Type I error rates of four tests in a homogenous population and in the case of no population stratification for qualitative traits. Note: “A homogenous population” means $k_0 = 1$ in simulation set 1; “no population stratification” means model 0 in simulation set 2.

| Population Structures | Test          | Alpha | Un-corrected | GC     | PC-linear | PC-nonp |
|-----------------------|---------------|-------|--------------|--------|-----------|---------|
| A Homogenous Population | Single Variant | 0.01  | 0.0131       | 0.0076 | 0.0129    | 0.01178 |
|                       | Regional      | 0.001| 0.00168      | 0.00112| 0.00182   | 0.00172 |
|                       |               | 0.01  | 0.00992      | 0.00958| 0.00992   | 0.0121  |
|                       |               | 0.001| 0.00124      | 0.00116| 0.00122   | 0.00162 |
| No Population Stratification | Single Variant | 0.01  | 0.01082      | 0.0055 | 0.01      | 0.0079  |
|                       | Regional      | 0.001| 0.00169      | 0.0007 | 0.00162   | 0.0014  |
|                       |               | 0.01  | 0.01113      | 0.00964| 0.01176   | 0.00926 |
|                       |               | 0.001| 0.00178      | 0.00138| 0.00166   | 0.00134 |

Table 3. Type I error rates of four tests based on simulation set 1 for quantitative traits.

| Number of Subpopulations | Test          | Alpha | Un-corrected | GC     | PC-linear | PC-nonp |
|--------------------------|---------------|-------|--------------|--------|-----------|---------|
| 2                        | Single Variant | 0.01  | 0.1516       | 0.00414| 0.01026   | 0.01038 |
|                          | Regional      | 0.001| 0.05998      | 0.00032| 0.00108   | 0.001   |
|                          |               | 0.01  | 0.258        | 0.0081 | 0.00995   | 0.00925 |
|                          |               | 0.001| 0.1489       | 0.0004 | 0.00115   | 0.00097 |
| 10                       | Single Variant | 0.01  | 0.0503       | 0.01498| 0.01096   | 0.01068 |
|                          | Regional      | 0.001| 0.01606      | 0.0026 | 0.0012    | 0.00108 |
|                          |               | 0.01  | 0.0549       | 0.0101 | 0.0107    | 0.0102  |
|                          |               | 0.001| 0.0139       | 0.00085| 0.0008    | 0.001   |
| 20                       | Single Variant | 0.01  | 0.04896      | 0.02662| 0.04234   | 0.01004 |
|                          | Regional      | 0.001| 0.0214       | 0.0102 | 0.01632   | 0.00082 |
|                          |               | 0.01  | 0.0488       | 0.01285| 0.0391    | 0.011   |
|                          |               | 0.001| 0.0137       | 0.00255| 0.0096    | 0.00075 |

Table 4. Type I error rates of five tests based on simulation set 1 for qualitative traits.
Table 5. Type I error rates of four tests based on simulation set 2 for quantitative traits.

| Model | Test       | Alpha | Un-corrected | GC   | PC-linear | PC-nonp |
|-------|------------|-------|--------------|------|-----------|---------|
| 1     | Single     | 0.01  | 0.0207       | 0.0135 | 0.0228   | 0.0132  |
|       | Variant    | 0.001 | 0.0042       | 0.0021 | 0.0048   | 0.0014  |
|       | Regional   | 0.01  | 0.1256       | 0.1213 | 0.1171   | 0.1065  |
|       |            | 0.001 | 0.052        | 0.0024 | 0.0463   | 0.0008  |
| 2     | Single     | 0.01  | 0.019        | 0.0139 | 0.0111   | 0.0115  |
|       | Variant    | 0.001 | 0.0048       | 0.0026 | 0.0011   | 0.0011  |
|       | Regional   | 0.01  | 0.0374       | 0.0128 | 0.0112   | 0.0082  |
|       |            | 0.001 | 0.0102       | 0.0009 | 0.0008   | 0.0006  |

Table 6. Type I error rates of five tests based on simulation set 2 for qualitative traits.

| Model | Test       | Alpha | Un-corrected | GC   | PC-linear | BiasedPerm | PC-nonp |
|-------|------------|-------|--------------|------|-----------|------------|---------|
| 1     | Single     | 0.01  | 0.0384       | 0.0128 | 0.0409   | 0.02564    | 0.007   |
|       | Variant    | 0.001 | 0.0132       | 0.0022 | 0.0151   | 0.00652    | 0.0005  |
|       | Regional   | 0.01  | 0.0248       | 0.0095 | 0.0288   | 0.02752    | 0.0114  |
|       |            | 0.001 | 0.0056       | 0.0009 | 0.005    | 0.00584    | 0.0008  |
| 2     | Single     | 0.01  | 0.0453       | 0.0174 | 0.0087   | 0.00514    | 0.0088  |
|       | Variant    | 0.001 | 0.0182       | 0.0048 | 0.0006   | 0.00032    | 0.001   |
|       | Regional   | 0.01  | 0.0366       | 0.0115 | 0.0121   | 0.00684    | 0.0093  |
|       |            | 0.001 | 0.0078       | 0.0007 | 0.0016   | 0.00046    | 0.0012  |

Figure 3. Power comparisons based on populations with $k_0$ subpopulations. "Homo" means that simulations are based on a homogenous population ($k_0 = 1$ in simulation set 1). "Structured" means that simulations are based on a structured population with 10 subpopulations.
Power comparisons under models 0 and 2 in simulation set 2 are given in Fig. 4. As shown by this figure, for quantitative traits, the pattern of power comparisons is very similar to that in Fig. 3. For qualitative traits, Uncorrected is the most powerful one. The pattern of power comparisons among PC-nonp, PC-linear, BiasedPerm, and GC is very similar to that in Fig. 3.

Analysis of GAW18 whole genome sequencing data set. The data set for GAW18 includes whole genome sequencing (WGS) data of 959 individuals (464 directly sequenced and the rest imputed) from 20 Mexican American pedigrees from San Antonio, Texas. There are 21–76 individuals in each pedigree. Phenotype data include sex, age, year of examination, systolic and diastolic blood pressure (SBP and DBP), use of antihypertensive medications, and tobacco smoking at up to four time points.

Since Mexican American population is admixture population, association studies based on unrelated individuals from this population may be subjected to bias due to population stratification. For our purpose, we extract 132 genetically unrelated individuals from the 20 pedigrees with phenotypes and WGS data and select SBP as the trait of interest while take sex, age, use of antihypertensive medications, and tobacco smoking as covariates. For WGS data, we only consider one chromosome (chromosome 17). Among the 132 unrelated individuals, there are 404,032 SNPs on chromosome 17. Since the sample size is small, we only consider the 41,754 uncommon SNPs with MAF between 0.02 and 0.05 instead of including rare SNPs. We randomly draw 10,000 SNPs from the 41,754 SNPs without replacement and test association between the phenotype and each of the 10,000 SNPs using each of the four tests: Uncorrected, GC, PC-linear, and PC-nonp. We repeat the drawing procedure 4 times with re-drawing 10,000 SNPs from the 41,754 SNPs. Quantile-quantile plots of the observed \(-\log_{10}(P\text{-values})\) of the four tests and expected \(-\log_{10}(P\text{-value})\) under the assumption of uniform distribution of P-values are given in Fig. 5. All quantile-quantile plots are averaged over 4 draws in order to show the average effect. Since we randomly draw 10,000 SNPs across chromosome 17, it is unlikely that there are a large number of SNPs in the 10,000 SNPs associated with SBP. Therefore, if population stratification can be well controlled for, P-values should proximately follow a uniform distribution. Figure 5 shows that only P-values of PC-nonp nearly follow a uniform distribution while for all other tests, the number of small P-values is more than expected.

Discussion

With the development of next-generation sequencing technology, there is increasing interest to detect associations between rare variants and complex traits. Many statistical methods have been developed for detecting rare variant associations. However, these methods may be subject to bias due to population stratification and, as
and and mU in TOW, denote yi
a
taxis can be written as for single-variant tests. SPTA is difficult to extend to regional rare variant association tests such as SKAT and TOW because it is designed
candidate marker. Although SPTA and PC-nonp are equivalent for single-variant tests under quantitative traits, trait values and PCs is assumed unknown, but trait values have linear relationship with genotypic scores at the
genotypes at genomic markers through a semi-parametric model, where the exact form of relationship
ination. SPTA models the relationship between trait values, genotypic scores at the candidate marker, and PCs of
population stratification in rare variant association studies. To apply PC-nonp, we first calculate PCs of genotypes at
the genomic markers. Then, we use these PCs to adjust population effects of both trait values and genotypes at a
candidate locus by applying nonparametric regressions. Our simulations show that the proposed PC-nonp can
control for population stratification well in all scenarios while existing methods cannot control for population
stratification at least in some scenarios. Simulations also show that PC-nonp's robustness to population stratification
will not reduce power. Applications to the GAW18 whole genome sequencing data set also show that our
proposed method can control for population stratification better than existing methods.

Although we describe our proposed method using a single-variant test and a weighted sum regional test, our
method can be applied to most existing rare variant association tests such as CMC, SKAT, and TOW. To apply
our method to SKAT and TOW, denote yi
a
and xim as the trait value and genotypic score at the mth variant of the ith individual. Let yim and xim denote the residuals of nonparametric regressions \( y_i = \mu(p_i) + \varepsilon_i \) and \( x_{im} = \mu_m(p_i) + \varepsilon_{im} \), where \( i = 1, \ldots, n \) and \( m = 1, \ldots, M \). Based on the residuals yim and xim, the test statistics of both SKAT and TOW can be written as

\[
T = \sum_{m=1}^{M} \frac{y_m^2}{U_m}, \quad \text{where } U_m = \sum_{i=1}^{n} (y_{im} - \bar{y}_m) (x_{im} - \bar{x}_m).
\]

In TOW, \( V_m = \sum_{i=1}^{n} (y_{im} - \bar{y}_m)^2 \) while, in SKAT, \( \frac{1}{V_m} = \text{Beta}(\text{MAF}; a_1, a_2) \), the beta distribution density function with pre-specified parameters
\( a_1 \) and \( a_2 \) evaluated at the sample MAF for the mth variant in the data. To apply our method to CMC, suppose that
M variants can be classified as \( S_r \) groups of rare variants and \( S_i \) individual variant sites. Define indicator variables
\( x_{is}(i = 1, \ldots, n; s = 1, \ldots, S_r) \) for all individuals and the \( S_r \) groups of rare variants, where \( x_{is} = 1 \) if minor alleles at any
variant in the \( s \)th group of the \( i \)th individual are present; \( x_{is} = 0 \) otherwise. Let \( S = S_r + S_i \) and define \( x_{is}(s+1) \) \( (s = 1, \ldots, S_r) \) as the genotypic score of the \( i \)th individual at the \( s \)th individual variant site. Let yim and xim denote the residuals of nonparametric regressions \( y_i = \mu(p_i) + \varepsilon_i \) and \( x_{im} = \mu_m(p_i) + \varepsilon_{im} \), where \( i = 1, \ldots, n \) and \( s = 1, \ldots, S_r \). Based on residuals yim and xim, we cannot use T
test because \( x_{im}^2 \) are not 0 and 1. We can use a score test or the improved score test66.

Zhang et al.41 proposed a semi-parametric test for association (SPTA) to control for population stratification.
SPTA models the relationship between trait values, genotypic scores at the candidate marker, and PCs of
genotypes at genomic markers through a semi-parametric model, where the exact form of relationship between
trait values and PCs is assumed unknown, but trait values have linear relationship with genotypic scores at the
candidate marker. Although SPTA and PC-nonp are equivalent for single-variant tests under quantitative traits, 
SPTA is difficult to extend to regional rare variant association tests such as SKAT and TOW because it is designed
for single-variant tests.

References
1. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence
data. Am J Hum Genet 83, 311–21 (2008).
2. Madsen, B. E. & Browning, S. R. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. PLoS Genetics
5 (2009).
3. Morgensthaler, S. & Thilly, W. G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a
cohort allelic sums test (CAST). Mutat Res 615, 28–56 (2007).
4. Price, A. L. et al. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86, 832–8 (2010).
5. Zawistowski, M. et al. Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. Am J Hum
Genet 87, 604–17 (2010).
6. Lee, S. et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-
exome sequencing studies. Am J Hum Genet 91, 224–37 (2012).
7. Neale, B. M. et al. Testing for an unusual distribution of rare variants. PLoS Genet 7, e1001322 (2011).
A Nonparametric Regression Approach to Control for Population Stratification

Sha, Q. et al. A Nonparametric Regression Approach to Control for Population Stratification in Rare Variant Association Studies. Scirp. Rep. 6:37444, doi: 10.1038/srep37444 (2016).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
