Received: 10 April 2017 | Revised: 7 October 2017 | Accepted: 19 October 2017

DOI: 10.1002/gepi.22102

RESEARCH ARTICLE

WILEY Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

On the substructure controls in rare variant analysis: Principal components or variance components?

Yiwen Luo[1,2] | Arnab Maity[2] | Michael C. Wu[3] | Chris Smith[1] | Qing Duan[4] | Yun Li[4,5] | Jung-Ying Tzeng[1,2,6,7]

[1]Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, United States of America

[2]Department of Statistics, North Carolina State University, Raleigh, North Carolina, United States of America

[3]Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America

[4]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America

[5]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America

[6]Department of Statistics, National Cheng-Kung University, Tainan, Taiwan

[7]Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taipei, Taiwan

Correspondence
Yun Li, Department of Genetics, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America.
Email: yunli@med.unc.edu
Jung-Ying Tzeng, Bioinformatics Research Center, Department of Statistics, North Carolina State University, Raleigh, North Carolina, United States of America. Department of Statistics, National Cheng-Kung University, Tainan, Taiwan. Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taipei, Taiwan.
Email: jytzeng@ncsu.edu

Funding information
Ministry of Science and Technology, Taiwan, Grant/Award Number: 106-2811-B-002-006; National Institutes of Health, Grant/Award Numbers: P01CA142538, R01HG006292, R01HG006703, R01HL129132

Abstract

Recent studies showed that population substructure (PS) can have more complex impact on rare variant tests and that similarity-based collapsing tests (e.g., SKAT) may suffer more severely by PS than burden-based tests. In this work, we evaluate the performance of SKAT coupling with principal components (PC) or variance components (VC) based PS correction methods. We consider confounding effects caused by PS including stratified populations, admixed populations, and spatially distributed nongenetic risk; we investigate which types of variants (e.g., common, less frequent, rare, or all variants) should be used to effectively control for confounding effects. We found that (i) PC-based methods can account for confounding effects in most scenarios except for admixture, although the number of sufficient PCs depends on the PS complexity and the type of variants used. (ii) PCs based on all variants (i.e., common + less frequent + rare) tend to require equal or fewer sufficient PCs and often achieve higher power than PCs based on other variant types. (iii) VC-based methods can effectively adjust for confounding in all scenarios (even for admixture), though the type of variants should be used to construct VC may vary. (iv) VC based on all variants works consistently in all scenarios, though its power may be sometimes lower than VC based on other variant types. Given that the best-performed method and which variants to use depend on the underlying unknown confounding mechanisms, a robust strategy is to perform SKAT analyses using VC-based methods based on all variants.

KEYWORDS

population substructure, principal components analysis, rare variant association tests, variance components

1 | INTRODUCTION

With the advance of next-generation sequencing, rare variants association study has emerged as an important paradigm for mapping human complex traits. Recent studies have revealed that the impact of population substructure can be more complex on rare variants (RVs) than on common variants (CVs) (Mathieson & McVean, 2012; Moore et al., 2013; Nelson et al., 2012; O'Connor et al., 2013; The 1000 Genomes Project Consortium, 2012; Zawistowski et al., 2014) and dealing with population substructure is indispensable for RV testing (Cardon & Palmer, 2003; Kang et al., 2010). Compared to CVs, RVs have relatively short evolutionary history and are more geographically localized or even private to specific subpopulations. For example, even in Europe, there is a gradient in diversity from Southern to Northern Europe (Mathieson & McVean, 2012; Nelson et al., 2012; The 1000 Genomes Project Consortium, 2012). Many works have explicitly quantified the degree of inflation under various simulated scenarios (e.g., Mathieson & McVean, 2012; Wang et al., 2014) and in real data (e.g., Hoffman, Krause, Lehmann, & Krüger, 2014; Nelson et al., 2012; Wang et al., 2014), and showed that the degree of inflation is more severe in RVs than in CVs. Furthermore, RV studies require large sample sizes, for which multiethnic samples start to gain attractions in association studies of RVs (Haiman et al., 2013).

Recent studies reach different conclusions about if the correction methods designed for CVs can effectively account for population substructures in RV studies (e.g., Babron, de Tayrac, Rutledge, Zeggini, & Génin, 2012; Baye et al., 2011; Jiang, Epstein, & Conneely, 2013; L. Liu et al., 2013; Liu, Nicolae, & Chen, 2013; Zhang, Guan, & Pan, 2013). By focusing on the principal component (PC)-based methods (Price et al., 2006), it is noted that the correction performance depends on (i) the types of collapsing methods (Q. Liu et al., 2013; Zawistowski et al., 2014) e.g., burden-based methods (Asimit, Day-Williams, Morris, & Zeggini, 2012; Li & Leal, 2008; Madsen & Browning, 2009; Morgenthaler & Thilly, 2007; Morris & Zeggini, 2010; Price et al., 2010) or similarity-based methods such as the Sequencing Kernel Association Test (SKAT) (Wu et al., 2011) and others (Tzeng et al., 2011; Tzeng, Lu, & Hsu, 2014; Zhao, Marceau, Zhang, & Tzeng, 2015); (ii) the degree of complication of the substructures (Q. Liu et al., 2013; Mathieson & McVean, 2012); and (iii) the number of variants collapsed (Q. Liu et al., 2013; Mathieson & McVean, 2012; Zawistowski et al., 2014). In addition, it is also not clear which types of variants (e.g., RVs or CVs or both) should be used to capture the subtle substructure of RVs (Babron et al., 2012; L. Liu et al., 2013) and how many PCs should be used to sufficiently remove the confounding caused by population substructure (Babron et al., 2012; L. Liu et al., 2013; Q. Liu et al., 2013; Mathieson & McVean, 2012). It has been suggested that for simple stratifications (e.g., two continental groups), top few PCs based (e.g., 5–10) on CVs can effectively account for population substructures (L. Liu et al., 2013; Q. Liu et al., 2013; Zhang, Shen, & Pan, 2013). For regional and complex substructures, many PCs (20–100 PCs) are required to remove the inflation induced by population substructures (Mathieson & McVean, 2012).

Besides the PC-based approaches, there are also VC-based approaches for adjusting for population substructures, e.g., Kang et al., 2010; Lippert et al., 2011; Schaid, Mcdonnell, Sinnwell, & Thibodeau, 2013; Thornton & McPeek, 2010. Instead of including the PCs as covariates, VC correction methods account for the substructure effect by empirically estimating the genealogical relatedness based on whole genome sharing information. For association studies of CVs, VC correction methods are known to perform better than PC-based approaches when populations have subtle and complex substructures and be applicable in a wide range of substructures, including continental to regional stratification, admixture, and cryptic relatedness (Kang et al., 2010). Nevertheless, only a few studies discussed the use of VC correction methods in RV analysis (Babron et al., 2012; Listgarten, Lippert, & Heckerman, 2013; Mathieson & McVean, 2012) and majority of these studies are under the context of burden tests. It is suggested that VC-based correction for burden test can effectively correct for the inflations caused by complex confounding structure and give maximum power among different controlling methods (Listgarten et al., 2013).

Recent work has suggested that similarity-based RV tests (e.g., SKAT) may suffer more severe impact by population substructure than burden-based RV tests (Q. Liu et al., 2013; Zawistowski et al., 2014). However, the performance of PC vs. VC correction methods has not been comprehensively examined for similarity-based RV tests. In this work, we aim to evaluate the PC-based and VC-based correction methods for SKAT and provide practical guidelines on the population substructure control for RV testing. Specifically, we implement the VC-based correction approaches under the framework of SKAT (referred to as SKAT-VC) and compare their performance with SKAT incorporating PC covariates (referred to as SKAT-PC). We consider a variety of confounding effects due to population structure ranging from stratification, admixture, and geographically distributed nongenetic risk; we consider COalescent Simulation (COSI) simulated sequence data and real sequence data from CoLaus samples (Caucasians residents of Lausanne, Switzerland (Firmann et al., 2008)). We investigate the effectiveness of using CVs (minor allele frequency (MAF) > 5%), RVs (MAF < 1%), less frequent variants (LFVs; 1%≤MAF≤5%), and all variants (AVs; including RVs, LFVs, and CVs) in reconstructing the substructures under each scenario and correcting for the inflation. We hope that our findings can provide helpful guidelines in the practice

of substructure controls in RV association tests using similarity based tests.

2 | METHODS

Consider a study consisting of n individuals indexed by $i = 1, \ldots, n$. For individual i, let Y_i be the trait value; X_i be a vector of covariates excluding population substructures; G_i be the design vector of L single nucleotide polymorphisms (SNPs) in the gene region to be evaluated for association. Each element of G_i, denoted by $G_{i\ell}$, corresponds to the minor allele count for individual i at locus ℓ and takes values 0, 1 or 2. We also obtain genome-wide SNPs for each individual, using which we compute genetic relationship matrix (GRM) (Price et al., 2006; Yang et al., 2011) based on linkage disequilibrium (LD)-pruned RVs, LFVs, CVs, and AVs, and denote the corresponding GRMs as $K^{RV}_{GRM}, K^{LFV}_{GRM}, K^{CV}_{GRM}$ and K^{AV}_{GRM}, respectively. Specifically, given a set of whole genome SNPs, e.g., the M LD-pruned CVs, we obtain GRM by first computing the normalized genomic design matrix $Z_{n\times M}$ so that each locus has mean 0 and variance 1 for the genotypic value, and then get $K^{CV}_{GRM} = ZZ^T/M$.

In this paper, we use SKAT (Wu et al., 2011), a similarity-based test, to evaluate the association between traits and rare variants and account for population substructure by two different methods: SKAT-PC and SKAT-VC. In SKAT-PC method, we treat substructure as fixed effects and account for its effects by including the top PCs of a particular K^{υ}_{GRM}, with $\upsilon \in$ {RV, LFV, CV, AV}. In SKAT-VC method, we treat substructure as random effects and account for the substructure-induced relatedness by including a particular K^{υ}_{GRM} as the variance-covariance matrix for the substructure effects. Below we briefly describe each of the methods.

2.1 | SKAT-PC method

For continuous traits, the SKAT-PC model has the form of

$$Y_i = X_i\,\beta + X_{PC,i}\beta_{PC} + h\left(G_i\right) + \epsilon_i,$$

where $X_{PC,i}$ the $q \times 1$ vector of the top q PC scores based on a certain K^{υ}_{GRM} for subject i (and $\upsilon \in$ {RV, LFV, CV, AV}) obtained by eigensoft (Price et al., 2006); β and β_{PC} are the coefficient vectors for the covariates and population substructures, respectively; $h(G_i)$ is the genetic effect for individual i; and ϵ_i is the random error with $N(0, \sigma^2)$. As in the standard kernel machine framework, the genetic effect is modeled using random effects, i.e., $h^T \equiv [h(G_1), \cdots h(G_n)] \sim MVN(0,\ \tau_G K_G)$, where τ_G is an unknown variance component and K_G is the $n \times n$ similarity matrix based on G_i and G_j. The (i, j) entry of K_G records the similarity between subjects i and j, and is defined using a prespecified kernel function $k_G(\cdot,\cdot)$ based on the L variants in the gene, i.e., $K_G\{i, j\} = k_G(G_i,\ G_j)$. In this article, we use the weighted linear kernel function $k_G(i, j) = \sum_{\ell=1}^{L} w_\ell G_{i,\ell} G_{j,\ell}$ with the suggested weight function $\sqrt{w_\ell} = (1 - MAF_\ell)^{24}$. The association between trait and gene can be examined by testing for $H_0:\ \tau_G = 0$, and the corresponding score test statistic is $T_{PC} = \frac{1}{2\sigma^2}\, Y^T P_1 K_G P_1 Y|_{\ \sigma^2=\hat{\sigma}^2}$, where $P_1 = I - \tilde{X}(\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T$, $\tilde{X} = (X,\ X_{PC})$ and $\hat{\sigma}^2$ is the moment estimator for σ^2 under $H_0:\ \tau_G = 0$. The test statistic T_{PC} follows a weighted χ^2_1 distribution.

2.2 | SKAT-VC method

For continuous traits, the SKAT-VC model has the form of

$$Y_i = X_i\,\beta + b_i + h\left(G_i\right) + \epsilon_i,$$

where the terms are defined as in SKAT-PC model, except that b_i is the substructure random effects for subject i, and $b = [b_1, \ldots, b_n]^T \sim \text{MVN}(0,\ \tau_R K^{\upsilon}_{GRM})$ with $\upsilon \in$ {RV, LFV, CV, AV}. Compared to the SKAT-PC method, herein additional variance component of population substructure τ_R is introduced. Consequently, the score test statistic for $H_0:\ \tau_G = 0$ is $T_{VC} = \frac{1}{2}\ Y^T P_2 K_G P_2 Y|_{\ \tau_R=\hat{\tau}_R\ ,\ \sigma^2=\hat{\sigma}^2}$, where $P_2 = V_G^{-1} - V_G^{-1}X(X^T V_G^{-1} X)^{-1} X^T V_G^{-1}$; $V_G = \tau_R\ K^{\upsilon}_{GRM} + \sigma^2 I$, and $(\hat{\tau}_R, \hat{\sigma}^2)$ are the restricted maximum likelihood estimates under $H_0:\ \tau_G = 0$. The test statistic T_{VC} approximately follows a weighted χ^2_1 distribution, i.e., $\sum_j \lambda_j \chi^2_1$, where λ_j are the nonzero eigenvalues of $\frac{1}{2} V_G^{\frac{1}{2}} P_2 K_G P_2 V_G^{\frac{1}{2}}$.

2.3 | Design of comparative study

We consider two simulation studies with different scenarios of confounding effects to evaluate the performance of SKAT-PC and SKAT-VC. In Simulation I, we explore the **substructure confounders** and use simulated sequence data using software package COSI (Schneider, Roessli, & Excoffier, 2000). In Simulation II, we explore **the spatially distributed confounders** and use the sequence data obtained from CoLaus (Cohorte Lausannoise) sequencing study (Firmann et al., 2008; Song et al., 2011).We describe the specific design for each simulation study below.

2.4 | Simulation I: Exploration of substructure confounders

In Simulation I, we use COSI (Schneider et al., 2000) to generate sequence data for a 1 Mb region for 10,000 European haplotypes, 10,000 African haplotypes, and 10,000 admixture of European and African haplotypes based on the coalescent model that mimics the corresponding population history. To create a stratified population, we first focus on the African haplotypes and European haplotypes. We randomly sample 2,000 haplotypes from each haplotype population with replacement and form the genotypes of 1,000 European individuals and 1,000 African individuals. A total of 21,621

polymorphic SNPs are obtained, including 3,558 CVs, 2,555 LFVs, and 15,508 RVs. We evenly partition the 21,621 SNPs into 500 genes, with an average of 43 SNPs per gene.

To generate admixture population, we randomly sample 4,000 haplotypes with replacement from the admixture haplotypes and form 2,000 individuals with admixed European and African ancestry. And total of 16,571 polymorphic SNPs (with 3,907 CVs, 2,427 LFVs, and 10,237 RVs) are evenly partitioned into 500 genes. Next, we simulate trait values based on the ancestry background and causal rare variants. We then perform gene-based SKAT tests on the RVs of every gene. Specifically, given the COSI simulated genotypes of individual i, we simulate trait value Y_i based on their genotypes from Normal$(\mu_i, 1)$ with $\mu_i = \beta_0 + \beta_{PS} X_{PS,i} + \sum_{\ell=1}^{L} r_\ell G_{i\ell}$, where $X_{PS,i}$ is the ancestry of individual i as defined below, β_{PS} is the population substructure effect, $G_{i\ell}$ is the minor allele count of locus ℓ, and $r_\ell = \gamma \times |\log \text{MAF}_\ell|$ if SNP ℓ is causal and is 0 otherwise. We consider three scenarios: (i) stratification ($\beta_{PS} \neq 0$), where we have 1,000 individuals from European population with $X_{PS,i} = 1$, and 1,000 individuals from African population with $X_{PS,i} = 0$; (ii) admixture ($\beta_{PS} \neq 0$), where we have 2,000 individuals from the African European admixture population with $X_{PS,i}$ being the proportion of European ancestry; and (iii) no confounding from population substructure ($\beta_{PS} = 0$), i.e., phenotypes were independent of ancestry. We set $\beta_{PS} = -1$ for (i) and -20 for (ii), which lead to detectable confounding effects caused by substructure.

2.5 | Simulation II: Exploration of spatially distributed confounders

Simulation II was conducted using the genome-wide association study (GWAS), target sequencing, and grandparental origin data from the Cohorte Lausannoise (CoLaus) cohort to create confounding. These data allow us to investigate the impact of substructure based on realistic population substructures in Europe. The CoLaus GWAS study (Firmann et al., 2008) contains SNPs data from 500K Affimetrix chips for a cohort of 6,188 Caucasians residents of Lausanne, Switzerland, aged 35–75 years old. The CoLaus sequence study (Nelson et al., 2012; Song et al., 2011) contained targeted sequence data for 202 genes (11,839 loci) for 2,000 of the 6,188 subjects. We focus on the 1,769 subjects that have both GWAS and sequence SNPs, use MArkov Chain Haplotyping (MACH) to impute the missing genotypes, and obtain 442,171 loci. These SNPs include 340,973 CVs, 40,006 LFVs, and 61,192 RVs. We test the association between the RVs of 202 genes and a simulated response variable derived from origin of grandparents. The birth place data of the four grandparents of the CoLaus samples show that, among the 1,769 subjects, ≈ 50% have all four paternal and maternal grandparents born in Switzerland, about 7%, 6.6%, 5%, and 5% with all four grandparents born in Italy, Portugal, Spain, and France, respectively. The remaining subjects (≈ 500 subjects) had grandparents born in same or different countries in Europe.

When simulating confounding effects, we adopt the design of Mathieson and McVean (2012), which considers a nongenetic risk factor that follows a certain spatial distribution and the spatial distribution correlates with the population substructure. To apply this design on the CoLaus samples, we make the following modifications. First, we use the birth places of the four grandparents to define an individual's "location"; such set-up allows us to introduce a natural correlation between the population substructure and the spatial distribution of the nongenetic risk factor. Second, we set Portugal and Spain as the geographic origin of the nongenetic risk factor, and consider a discrete risk distribution and a continuous risk distribution of the nongenetic factor. In the data, there are ≈ 11.6% of CoLaus individuals with all four grandparents born in Portugal or Spain. We simulate Y_i based on the distance between the birth places of the four grandparents of individual i and the risk center (Portugal and Spain), i.e., $Y_i \sim \text{Normal}(\mu_i, 1)$ with $\mu_i = \beta_{BP} X_{BP,i} + \sum_{\ell=1}^{L} r_\ell G_{i\ell}$, where $X_{BP,i}$ indicates the birth places of the grandparents of individual i (as defined below); β_{BP} is the confounding effect; $G_{i\ell}$ and r_ℓ are the same as defined in Simulation I. In Scenario (i), we let the nongenetic risk have a discrete spatial distribution. To do so, we set $\beta_{BP} > 0$ and $X_{BP,i}$ equal to the number grandparents born in Spain or Portugal for subject i. In Scenario (ii), we let the nongenetic risk have a continuous spatial distribution. Besides letting $\beta_{BP} > 0$, we define $X_{BP,i}$ using the following procedure: First, define a risk origin at geographic coordinate (39.75, −6) (i.e., latitude 39.75 N and longitude 6 W, the average coordinate of Spain and Portugal). Next, we obtain the geographic coordinate of the birth place for grandparent k $(k = 1, \ldots, 4)$ of Subject i, denoted by (A_{ik}, O_{ik}), and calculate the distance to the high-risk origin, i.e., $d_{ki} = \sqrt{(A_{ki} - 39.75)^2 + [O_{ki} - (-6)]^2}$ for grandparent k. Finally, for Subject i, we compute $D_i = \frac{1}{4} \sum_{k=1}^{4} d_{ki}$, i.e., the average distance among the four grandparents to the risk origin and set $X_{BP,i} = (D_i - \bar{D}) / s_D$, where $\bar{D}$ and s_D are the mean and standard deviations of $D_1, \ldots, D_n$, respecitively. The resulting Y_i has a larger value if the grandparents of Subject i were born near the high-risk origin. We also consider Scenario (iii) of $\beta_{PS} = 0$, i.e., no confounding caused by the nongenetic risk factor. We set $\beta_{BP} = 5$ for (i) and 10 for (ii), which lead to detectable confounding effects.

2.6 | Evaluation of the performance of SKAT-PC and SKAT-VC

We evaluate the type I error rates and power of the gene-based RV test using SKAT-PC and SKAT-VC. The tests are performed on RVs for each of the 500 genes in Simulation I and for each of the 202 genes in Simulation II. We adjust the

population substructure using PC or VC method, based on RVs, LFVs, CVs, or AVs. An ideal method would be able to correct for inflation caused by the confounding effects and retain a high power to detect association signal of the causal rare variants.

For type I error analysis, we set $\gamma = 0$ (i.e., no effect of the causal RVs) with the effect size of causal variant ℓ as $r_\ell = \gamma \times |\log \text{MAF}_\ell|$. For a given SKAT-PC/-VC method, we collect the *P*-values from all genes and compare its distribution with the expected *P*-value distribution of no genetic effect (i.e., Uniform (0,1)) using quantile–quantile plots (QQ plots). In the QQ plot, if the dots fall along the 45 degree line, it indicates the observed distribution agrees with the expected distribution. To permit a visual inspection of major deviations, we shade the 95% confidence band of the 45 degree line as the "allowable" zone, and an empirical distribution that follows outside the shaded area would imply major deviation from the expected null distribution. We note that this confidence band only serves as a coarse criterion to detect gross deviations from the expected distribution. We repeat the process 10 times and present the QQ plots averaged over the 10 replications as did in Mathieson and McVean (2012). In addition, we also report the type I error rate at nominal level $\alpha = 0.05$ and 0.005, which is the proportion of rejection among all genes, averaged over the 10 replications. We perform statistical tests to examine if the type I error rate (denoted by π) of a method is significantly larger than the nominal level α. The rejection region for testing for $H_0 : \pi \le \alpha$ vs. $H_A : \pi > \alpha$ and adjusting for 17 strategies (i.e., to correct for confounders using 10 PCs, 50 PCs, 100 PCs, and VC combined with RVs, LFVs, CVs, and AVs as well as no correction of 0 PC) is $\hat{\pi} > \alpha + |Z_{0.05/17 \text{ tests}}| \sqrt{\frac{\alpha \times (1-\alpha)}{\#\text{ of replicates} \times \#\text{ of genes}}}$.

For power analysis, we randomly select a causal gene and set $\gamma > 0$ for the causal gene. The value of γ is determined so that the SKAT power is around 0.7 to 0.8 at $\alpha = 0.05$ when no confounders and no correction methods are used under each scenario. We conduct 200 replications and compute the power as the proportion of rejection among the 200 replications. We report power only for those methods whose type I error rate is not significantly larger than nominal level.

3 | RESULTS

3.1 | Simulation I: Exploration of substructure confounders

Under the scenario of no confounding (Table 1a; top row of Fig. 1/Supplementary Figure 1), both SKAT-PC and

TABLE 1 Estimated type I error rates $\hat{\pi}$ of SKAT-PC and SKAT-VC in Simulation I (substructure confounders) at nominal level $\alpha = 0.05$ and $\alpha = 0.005$. Bold cells indicate that the true type I error rate π of a method is not significantly greater than α (i.e., a correction method can adjust for confounding effect). The rejection region for testing for $H_0 : \pi \le \alpha$ vs. $H_A : \pi > \alpha$ and adjusting for 17 strategies (i.e., to correct for confounders using 10 PCs, 50 PCs, 100 PCs, and VC combined with RV, LFV, CV, and AV as well as no correction (i.e., 0 PC)) is $\hat{\pi} > \alpha + |Z_{0.05/17 \text{ tests}}| \sqrt{\frac{\alpha \times (1-\alpha)}{5000}}$, which is > 0.0585 for $\alpha = 0.05$ and > 0.00775 for $\alpha = 0.005$

a. No population substructure

$\alpha = 0.05$	0 PC	10 PC	50 PC	100 PC	VC	$\alpha = 0.005$	0 PC	10 PC	50 PC	100 PC	VC
RV	**0.0466**	**0.0450**	**0.0504**	**0.0500**	**0.0466**	RV	**0.0040**	**0.0044**	**0.0050**	**0.0048**	**0.0040**
LFV		**0.0464**	**0.0446**	**0.0474**	**0.0466**	LFV		**0.0038**	**0.0042**	**0.0046**	**0.0040**
CV		**0.0462**	**0.0480**	**0.0494**	**0.0466**	CV		**0.0036**	**0.0048**	**0.0052**	**0.0040**
AV		**0.0454**	**0.0440**	**0.0472**	**0.0466**	AV		**0.0034**	**0.0050**	**0.0050**	**0.0040**

b. Stratified population

$\alpha = 0.05$	0 PC	10 PC	50 PC	100 PC	VC	$\alpha = 0.005$	0 PC	10 PC	50 PC	100 PC	VC
RV	0.5020	0.2578	0.0818	0.0604	0.1490	RV	0.3054	0.0988	0.0122	**0.0064**	0.0314
LFV		0.0792	0.0658	**0.0584**	0.0680	LFV		0.0130	**0.0072**	**0.0064**	0.0092
CV		0.0656	**0.0494**	**0.0532**	**0.0512**	CV		0.0100	**0.0056**	**0.0050**	**0.0058**
AV		**0.0578**	**0.0488**	**0.0458**	**0.0394**	AV		**0.0066**	**0.0048**	**0.0056**	**0.0034**

c. Admixed population

$\alpha = 0.05$	0 PC	10 PC	50 PC	100 PC	VC	$\alpha = 0.005$	0 PC	10 PC	50 PC	100 PC	VC
RV	0.1616	0.1738	0.1670	0.1650	0.0638	RV	0.0478	0.0448	0.0442	0.0464	0.0078
LFV		0.1652	0.1660	0.1468	0.1210	LFV		0.0488	0.0420	0.0420	0.0256
CV		0.1690	0.1618	0.1754	0.1390	CV		0.0454	0.0406	0.0470	0.0304
AV		0.1720	0.1556	0.1428	**0.0484**	AV		0.0488	0.0422	0.0300	**0.0060**

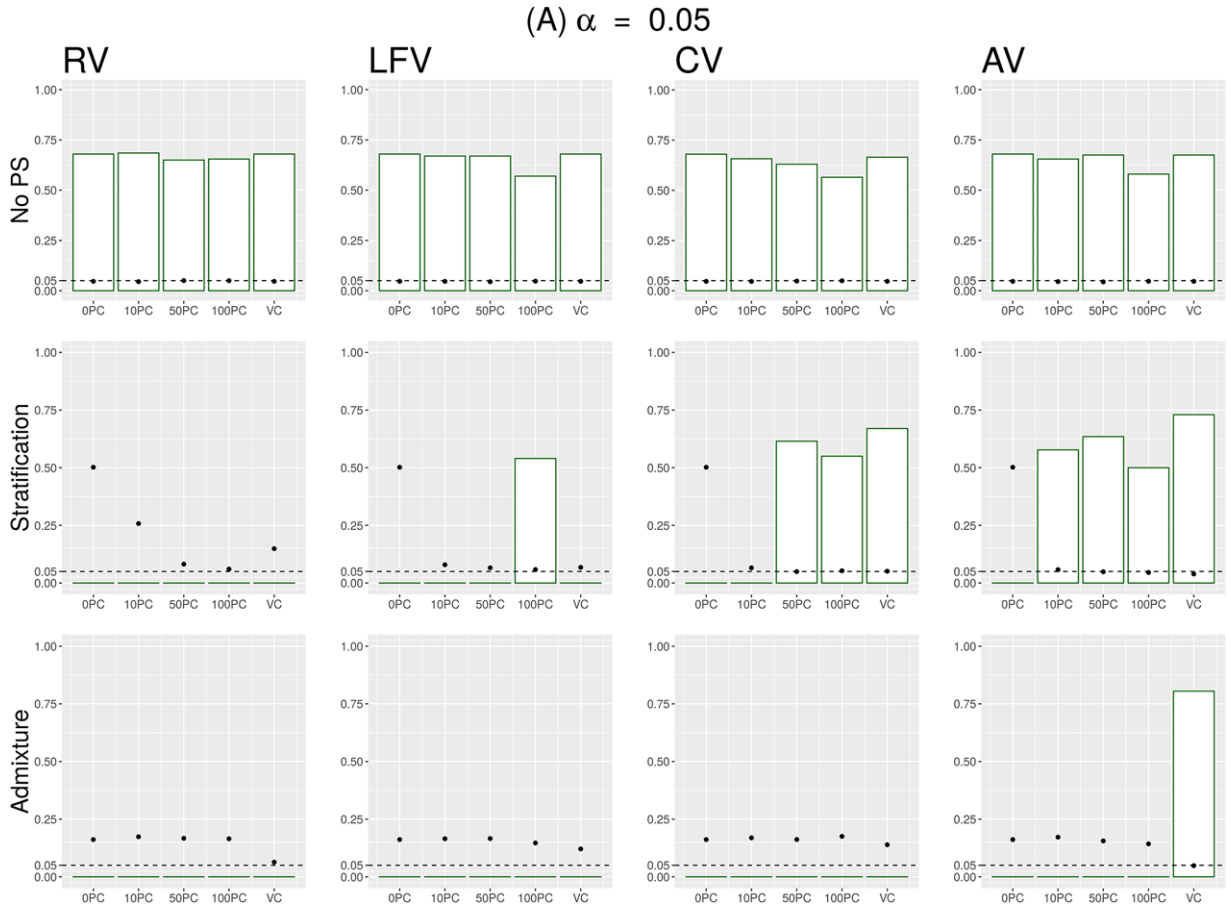

FIGURE 1 Power and type I error rates of SKAT-PC and SKAT-VC in Simulation I (substructure confounders) at nominal level $\alpha = 0.05$ (Panel A) and 0.005 (Panel B). The bars indicate power and the dots indicate type I error rates. Different rows are for different scenarios of population substructure (PS), i.e., (from top to bottom) no PS, stratification, and admixture. Different columns indicate the types of variants used to construct PC/VC, i.e., RV for rare variants, LFV for less frequent variants, CV for common variants and AV for all variants

SKAT-VC have reasonable performance, regardless of the types of variants used to construct PCs and VCs (i.e., RVs, LFVs, CVs, or AVs). The dots in the QQ plots fall within the shaded area and the type I error rates are not significantly higher than nominal level. For power (Fig. 1 top row), SKAT-VC has similar power as SKAT-0PC (i.e., no substructure adjustment) regardless which types of variants are used for obtaining VCs. Though not very obvious, the power of SKAT-PC tends to drop when more PCs are included.

Under the scenario of stratification (Table 1b; middle row of Fig. 1/Supplementary Figure 1), the ability of SKAT-PC and SKAT-VC to correct for the confounding depends on the types of variants used to estimate substructure. For SKAT-PC, the number PCs required increases when rarer variants are used, e.g., SKAT-PC based on RVs or LFVs requires 50–100 PCs and SKAT-PC based on AVs only requires 10 PCs. The power of SKAT-PC seems to be similar regardless it is based on LFVs CVs or AVs. For SKAT-VC, using AVs and CVs can correct for stratification, but using RVs and LFVs cannot. For power of those methods that can effectively correct confounding (Fig. 1 middle row), SKAT-VC-AV has the highest power when $\alpha = 0.05$; SKAT-100PC-RV and SKAT-VC-AV have the highest power when $\alpha = 0.005$.

Under the scenario of admixture (Table 1c; bottom row of Fig. 1/Supplementary Figure 1), SKAT-VC with AVs is the only method that can provide effectively control. Regardless based on which types of variants, SKAT-PC with even 100 PCs cannot correct for the substructure effect regardless which types of variants are used.

3.2 | Simulation II: Exploration of spatially distributed confounders

Under the scenario of no confounding (Table 2a; top row of Fig. 2/Supplementary Figure 2), both SKAT-PC and SKAT-VC have reasonable performance– the dots of QQ plot all fall within the shaded area and the type I error rates are not significantly higher than the nominal level. For power (Fig. 2 top row), all methods have similar values comparing to no correction (i.e., SKAT-0PC).

When the nongenetic risk has a discrete spatial distribution (Table 2b; middle row of Fig. 2/Supplementary Figure 2),

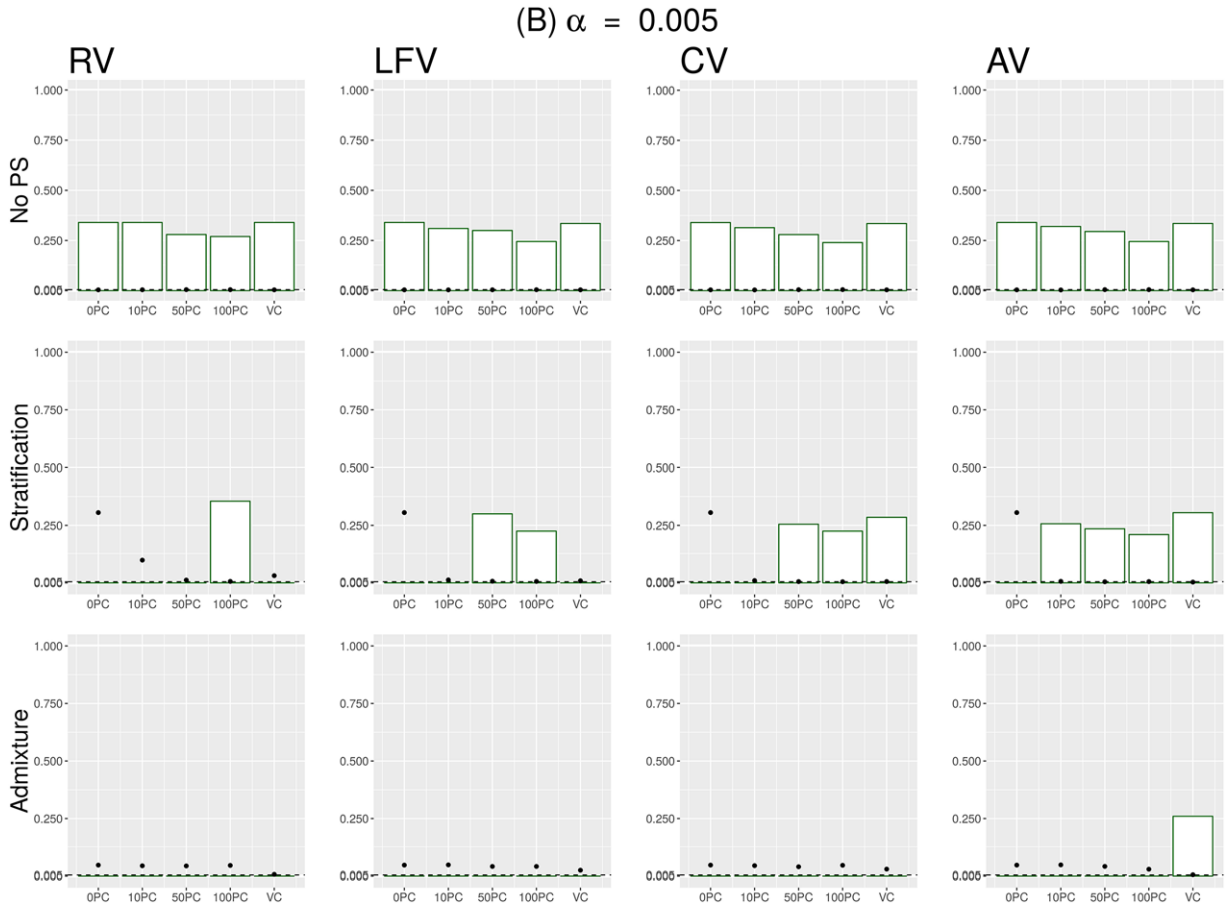

FIGURE 1 Continued

both SKAT-PC and SKAT-VC can correct for confounding effect, though (i) certain types of variants have to be used for SKAT-VC, and (ii) all types of variants works for SKAT-PC. For SKAT-PC, only 10 PCs are sufficient to capture the confounding effect regardless which types of variants are used. On the other hand, SKAT-VC works when RVs and AVs are used. For power analysis (Fig. 2 middle row), SKAT-PC-AV has the highest power, followed by SKAT-PC-RV and SKAT-VC-RV; SKAT-VC-AV tends to have lower power than SKAT-PC.

When the nongenetic risk has a continuous spatial distribution (Table 2c; bottom row of Fig. 2/Supplementary Figure 2), both SKAT-PC and SKAT-VC can correct for the confounding effect regardless which types of variants are used, though SKAT-VC-CV has some slight inflation when $\alpha = 0.05$. For SKAT-PC, 10 PCs are sufficient to correct for confounding. For power analysis (Fig. 2 bottom row), SKAT-100PC-AV, SKAT-50PC-AV, and SKAT-VC-RV have the highest power; SKAT-VC-AV again tends to have lower power than SKAT-PC.

In summary, when the confounding effects is caused by nongenetic risk whose spatial distribution is related to substructure, SKAT-PC can adjust for the confounders with just 10 PCs regardless of which types of variants used. The power of SKAT-PC based on AVs tends to be the highest. SKAT-VC can correct for the inflation except when RVs or AVs are used; SKAT-VC-RV yields similar power to SKAT-PC and SKAT-VC-AV yields lower power than SKAT-PC.

4 | DISCUSSION

Focusing on SKAT RV tests, we evaluate the performance of PC-based and VC-based methods for correcting inflation caused by confounders related to population substructure. We consider simulated and real sequence data, and confounding caused by population stratification, population admixture, and spatially distributed nongenetic factors. We find that these correcting methods developed for CV association analysis can work for RV analysis. Specifically, SKAT-PC can correct for the substructure-related confounders in all scenarios investigated in this work except for admixed populations; and SKAT-VC is capable to correct for confounding effects in all scenarios. However, which variants to use in order to reach effective correction depend on the specific scenarios, and for SKAT-PC, it would also depend on the number of PCs included. Overall speaking, for SKAT-PC, using AVs often requires smaller number of PCs and yields

TABLE 2 Estimated type I error rates $\hat{\pi}$ of SKAT-PC and SKAT-VC in Simulation I (spatially distributed confounders) at nominal level $\alpha = 0.05$ and $\alpha = 0.005$. Bold cells indicate that the true type I error rate π of a method is not significantly greater than α (i.e., a correction method can adjust for confounding effect). The rejection region for testing for $H_0 : \pi \leq \alpha$ vs. $H_A : \pi > \alpha$ and adjusting for 17 strategies (i.e., to correct for confounders using 10 PCs, 50 PCs, 100 PCs, and VC combined with RV, LFV, CV, and AV as well as no correction (i.e., 0 PC)) is $\hat{\pi} > \alpha + |Z_{0.05/17 \text{ tests}}| \sqrt{\frac{\alpha \times (1-\alpha)}{2020}}$, which is > 0.0633 for $\alpha = 0.05$ and > 0.0093 for $\alpha = 0.005$

a. No confounders

$\alpha = 0.05$	0 PC	10 PC	50 PC	100 PC	VC
RV	**0.0546**	**0.0551**	**0.0500**	**0.0469**	**0.0541**
LFV		**0.0536**	**0.0536**	**0.0505**	**0.0546**
CV		**0.0536**	**0.0546**	**0.0546**	**0.0541**
AV		**0.0561**	**0.0490**	**0.0464**	**0.0531**

$\alpha = 0.005$	0 PC	10 PC	50 PC	100 PC	VC
RV	**0.0083**	**0.0077**	**0.0061**	**0.0066**	**0.0082**
LFV		**0.0066**	**0.0071**	**0.0066**	**0.0082**
CV		**0.0071**	**0.0066**	**0.0061**	**0.0077**
AV		**0.0088**	**0.0082**	**0.0082**	**0.0082**

b. Nongenetic confounder with discrete spatial distribution

$\alpha = 0.05$	0 PC	10 PC	50 PC	100 PC	VC
RV	0.1061	**0.0597**	**0.0571**	**0.0546**	**0.0347**
LFV		**0.0464**	**0.0500**	**0.0495**	0.0725
CV		**0.0622**	**0.0628**	**0.0612**	0.0898
AV		**0.0587**	**0.0551**	**0.0475**	**0.0489**

$\alpha = 0.005$	0 PC	10 PC	50 PC	100 PC	VC
RV	0.0327	**0.0056**	**0.0041**	0.0097	**0.0026**
LFV		**0.0056**	**0.0082**	**0.0046**	0.0128
CV		**0.0051**	**0.0015**	**0.0020**	0.0163
AV		**0.0077**	**0.0071**	**0.0005**	**0.0051**

c. Nongenetic confounder with continuous spatial distribution

$\alpha = 0.05$	0 PC	10 PC	50 PC	100 PC	VC
RV	0.0908	**0.0490**	**0.0577**	**0.0367**	**0.0383**
LFV		**0.0510**	**0.0546**	**0.0526**	**0.0582**
CV		**0.0378**	**0.0418**	**0.0418**	0.0710
AV		**0.0464**	**0.0439**	**0.0301**	**0.0393**

$\alpha = 0.005$	0 PC	10 PC	50 PC	100 PC	VC
RV	0.0148	**0.0031**	**<0.0005**	0.0117	**0.0031**
LFV		**0.0056**	**0.0066**	**0.0041**	**0.0092**
CV		**0.0046**	**<0.0005**	**0.0005**	**0.0077**
AV		**0.0077**	**0.0026**	**0.0010**	**0.0036**

reasonable power (although SKAT-PC cannot adjust for admixture effects regardless which variants are used to construct PCs). For SKAT-VC, VC-AV appears to be the optimal strategy because it can adjust for confounding effects in all scenarios investigated in this work, although sometimes it may have less power compared to its PC counterpart (such as in the scenario of spatially distributed confounders in the CoLaus simulations). Given the underlying confounding sources is often unknown in a prior, SKAT-VC using AVs would serve as a more reliable strategy to adjust for substructure-related confounding effects.

In this work, the investigations of PC vs. VC correction methods in SKAT tests are conducted using quantitative traits because there are currently no methods available for SKAT-VC with binary traits. The extension of SKAT-VC from continuous traits to binary traits is computationally and numerically nontrivial–the binary SKAT-VC method requires the estimation of nuisance variance component for the confounding effects under logistic mixed models; such estimation involves optimization of a penalized likelihood of a generalized linear mixed model (Liu, Ghosh, & Lin, 2008) and/or inversion of a high-dimensional covariance matrix as well as high-dimensional integration (e.g., Zhang & Lin, 2003). Although several algorithms have been proposed in the field of SNP-set GxE kernel tests for binary traits, e.g., penalized methods of Lin, Lee, Christiani, and Lin (2013) and EM algorithms of Zhao et al. (2015), these approaches are computationally intensive and it is not clear if these methods, which deal with a low-rank similarity matrix computed from a candidate SNP set, can appropriately handle the GRM (which is computed from whole genome SNPs and has its rank equal to the sample size) and yield computationally feasible and numerically stable results for binary SKAT-VC tests. Nevertheless, given the robust performance observed in quantitative SKAT-VC tests, it is worth the effort to develop computationally efficient algorithms for binary SKAT-VC tests for future study.

In Simulation I with substructure confounders, we observe that SKAT-PC does not always provide effective control of confounding effects. Recently, Sha, Zhang, and Zhang (2016) proposes a new PC-based method, called PC-nonp, to control for population substructure. Unlike the typical PC-based correction methods that assume the top PCs have a linear effect on the traits, PC-nonp uses a nonparametric regression to model the potential nonlinear or complex effects of the PCs. Sha et al. (2016) show the effectiveness of PC-nonp in controlling for population substructure with burden RV tests, and it is of great interest to evaluate the performance of PC-nonp with

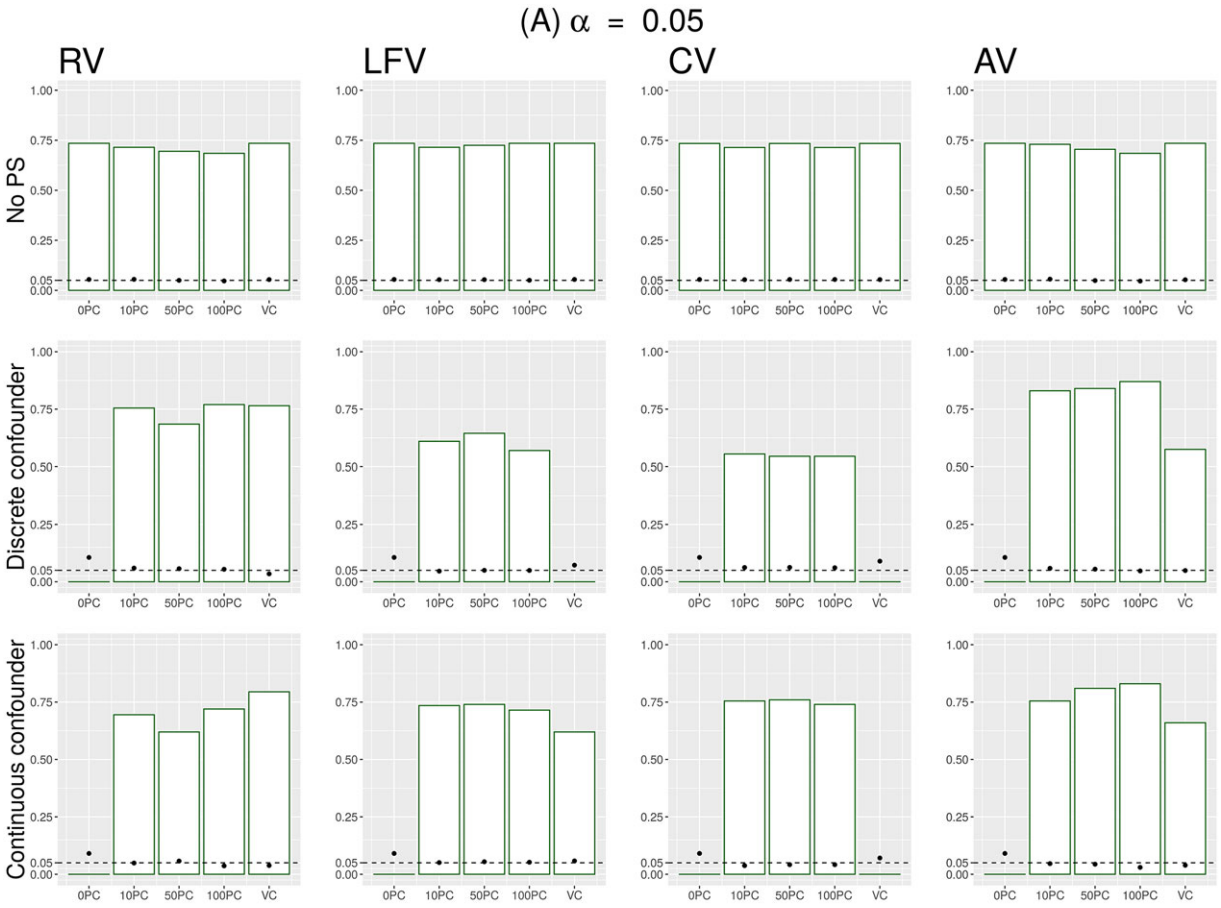

FIGURE 2 Power and type I error rates of SKAT-PC and SKAT-VC in Simulation II (spatially distributed confounders) at nominal level $\alpha = 0.05$ (Panel A) and 0.005 (Panel B). The bars indicate power and the dots indicate type I error rates. Different rows are for different type of confounders, i.e., (from top to bottom) no confounders, confounders with a discrete spatial distribution, and confounders with a continuous spatial distribution. Different columns indicate the types of variants used to construct PC/VC, i.e., RV for rare variants, LFV for less frequent variants, CV for common variants, and AV for all variants

SKAT. However, in our preliminary explorations, it seems that SKAT-PC-nonp was not able to control for confounding regardless which type of variants are used to obtain PCs (data not shown). As noted in the literature (e.g., Q. Liu et al., 2013 and Zawistowski et al., 2014), similarity-based RV tests can be more severely impacted by population substructure than burden-based RV tests, and those PS methods that can effectively account for confounding for one type of RV tests (e.g., burden-based tests) may not always work for the other types (e.g., similarity based) of RV tests. We see that the optimal usage in coupling confounder correction methods with SKAT remains open for further investigations.

4.1 | PC vs. VC

VC-based methods treat the population substructure as random effects and it is known to be able to correct inflation caused by complex confounding even when PC methods failed (Listgarten et al., 2013). In our study, we see that SKAT-VC can correct for the inflation caused by both stratification and admixture and achieve higher power than SKAT-PC. On the other hand, when the inflation is caused by spatially confined nongenetic confounders, we see that both SKAT-VC (except for SKAT-VC using CVs and LFVs) and SKAT-PC provide effective adjustment, and SKAT-PC tends to have higher power than SKAT-VC. The results agree with the observation of Zhang and Pan (2014). Focusing on CV association tests with PC and VC constructed using CVs, Zhang and Pan (2014) investigate the ability of PC and VC methods in adjusting for spatially distributed nongenetic confounders. They found that (i) PC can more effectively adjust for the confounding effects than VC because top PCs of genetic variants can represent geographic coordinates (Wang, Zöllner, Rosenberg, Weinblatt, & Shadick, 2012; Zhang & Pan, 2014); and (ii) VC based on CVs may fail to correct for the confounding effects in CV association tests. Besides reaching similar conclusions as theirs in our investigation on RV association tests, we also found that in RV tests, if AVs are used to correct for potential confounding effects, VC-AV methods can successfully adjust for spatially distributed confounders, and PC-AV methods can

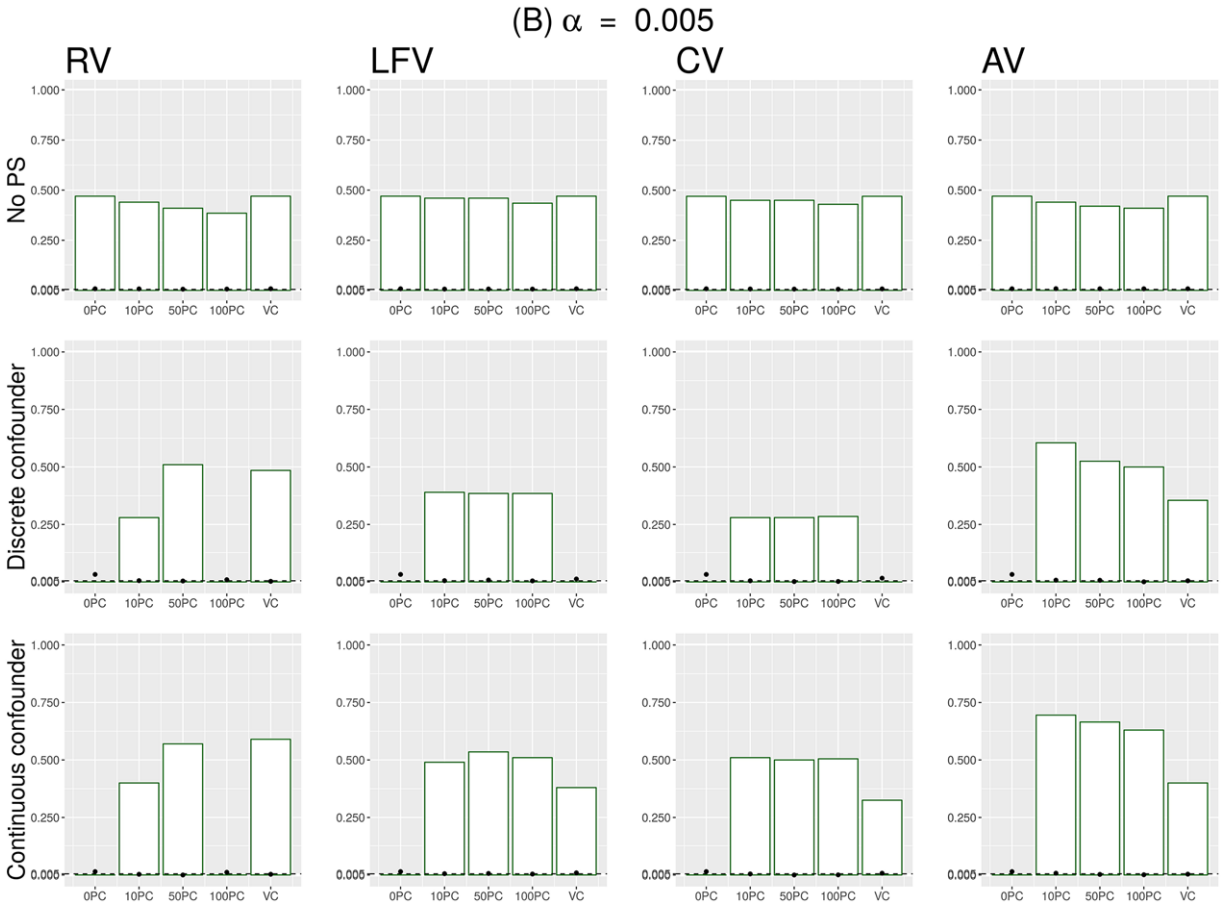

FIGURE 2 Continued

achieve higher power than PC methods using other types of variants.

4.2 | Which variants to use to obtain PC and VC?

Several works investigated the performance of PC-based correction methods under stratified populations (L. Liu et al., 2013; Liu, Nicolae, & Chen 2013; Zhang et al., 2013; Zhang et al., 2013). Although some found that using RVs to construct PCs can more effectively adjust for inflation (Q. Liu et al., 2013), most found that using CVs or AVs to construct PCs would be more effective (L. Liu et al., 2013; Zhang et al., 2013; Zhang et al., 2013). Our results in general agree with the latter, i.e., using AVs to construct PCs provided more effective adjustment (i.e., requiring fewer number of PCs) than RVs for inflation caused by stratification. We also observed that for confounding caused by spatially distributed nongenetic risk factor, PCs-based AVs would provide the most effective and efficient (i.e., yielding highest power) adjustment than PCs based on other types of variants. As pointed out by Zhang et al. (2013), although RVs were more likely to cluster in a few subpopulations, only a low proportion (≈ 25%) of RVs was population specific. In contrast, a relatively high proportion of CVs and LFVs (e.g., > 70%) were subgroup specific, which may lead to a better adjusting performance under stratified population.

For VC-based methods, VC-AV provides the most reliable performance across all scenarios, including admixed population (where all PC methods failed), stratified population (where VC-RV and VC-LFV failed), and discrete distributed nongenetic risk factor (where VC-CV and VC-LFV failed). VC-AV has the highest power among all PC and VC methods in stratified populations; yet it does not yield the highest power when accounting for spatially distributed confounders. On the other hand, VC-RV, though fails to adjust for substructure confounders, provides satisfactory adjusting performance for spatially distributed confounders and often yields high power among all PC and VC methods. This observation is not unexpected according to the results of Q. Liu et al. (2013), though their evaluation focused on SKAT-PC and we focus on SKAT-VC.

In conclusion, it has been noticed that similarity-based association test of rare variants, such as SKAT, which can accommodate a mixture of risky and protective variants in a gene, can also be more vulnerable to substructures than burden-based test (Zawistowski et al., 2014). Our work attempts to provide some guidance regarding the choice of

correction methods for SKAT RV tests. First, we would suggest using AVs to obtain PCs or VCs to achieve effective correction and reasonable power across various scenarios. Second, depends on the complexity of population substructure and the type of variants used, the minimum number of PCs required varies from 10 to 100, though 10 PCs are often sufficient for SKAT-PC if AVs are used to construct PCs in the scenarios explored in this study. Third, when population substructure becomes more complex, such as admixture or localized structure, SKAT-VC would be preferred. Overall, given the underlying confounding mechanism is not known in a priori, SKAT-VC-based AVs would be the most robust correction method.

ACKNOWLEDGMENTS

The authors deeply thank Drs. Peter Vollenweider and GerardWaeber, PIs of the CoLaus study, and Drs. Meg Ehm and Matthew Nelson, collaborators at GlaxoSmithKline, for providing the CoLaus genotype data. This work was supported in part by National Institutes of Health grants P01CA142538, R01HG006292, R01HG006703, and R01HL129132 and Ministry of Science and Technology of Taiwan grant 106-2811-B-002-006.

ORCID

Yun Li http://orcid.org/0000-0002-3467-2599
Jung-Ying Tzeng http://orcid.org/0000-0002-5505-1775

REFERENCES

Asimit, J. L., Day-Williams, A. G., Morris, A. P., & Zeggini, E. (2012). ARIEL and AMELIA: Testing for an accumulation of rare variants using next-generation sequencing data. *Human Heredity*, *73*(2), 84–94.

Babron, M.-C., de Tayrac, M., Rutledge, D. N., Zeggini, E., & Génin, E. (2012). Rare and low frequency variant stratification in the UK population: Description and impact on association tests. *PloS One*, *7*(10), e46519.

Baye, T. M., He, H., Ding, L., Kurowski, B. G., Zhang, X., & Martin, L. J. (2011). Population structure analysis using rare and common functional variants. *BMC Proceedings*, *5*(9), S8.

Cardon, L. R., & Palmer, L. J. (2003). Population stratification and spurious allelic association. *The Lancet*, *361*(9357), 598–604.

Firmann, M., Mayor, V., Vidal, P. M., Bochud, M., Pecoud, A., Hayoz, D., … Vollenweider, P. (2008). The CoLaus study: A population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. *BMC Cardiovascular Disorders*, *8*(1), 6.

Haiman, C. a., Han, Y., Feng, Y., Xia, L., Hsu, C., Sheng, X., … Stram, D. O. (2013). Genome-wide testing of putative functional exonic variants in relationship with breast and prostate cancer risk in a multiethnic population. *PLoS Genetics*, *9*(3), e1003419.

Hoffman, J. I., Krause, E. T., Lehmann, K., & Krüger, O. (2014). MC1R genotype and plumage colouration in the zebra finch (Taeniopygia guttata): Population structure generates artefactual associations. *PLoS One*, *9*(1), e86519.

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. Y., Freimer, N. B., … Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, *42*(4), 348–354.

Li, B., & Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *The American Journal of Human Genetics*, *83*(3), 311–321.

Lin, X., Lee, S., Christiani, D. C., & Lin, X. (2013). Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics*, *14*(4), 667–681.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., & Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods*, *8*(10), 833–835.

Listgarten, J., Lippert, C., & Heckerman, D. (2013). Fast-LMM-Select for confounding from spatial structure and rare variants. *Nature Genetics*, *45*(5), 470–471.

Liu, D., Ghosh, D., & Lin, X. (2008). Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics*, *24*(9), 292.

Liu, L., Sabo, A., Neale, B. M., Nagaswamy, U., Stevens, C., Lim, E., … Roeder, K. (2013). Analysis of rare, exonic variation amongst subjects with autism spectrum disorders and population controls. *PLoS Genetics*, *9*(4), e1003443.

Liu, Q., Nicolae, D. L., & Chen, L. S. (2013). Marbled inflation from population structure in gene-based association studies with rare variants. *Genetic Epidemiology*, *37*(3), 286–292.

Madsen, B. E., & Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, *5*(2), e1000384.

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

Moore, C. B., Wallace, J. R., Wolfe, D. J., Frase, A. T., Pendergrass, S. a., Weiss, K. M., & Ritchie, M. D. (2013). Low frequency variants, collapsed based on biological knowledge, uncover complexity of population stratification in 1000 genomes project data. *PLoS Genetics*, *9*(12), e1003959.

Morgenthaler, S., & Thilly, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutation Research*, *615*(1), 28–56.

Morris, A. P., & Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology*, *34*(2), 188–193.

Nelson, M. R., Wegmann, D., Ehm, M. G., Kessner, D., St. Jean, P., Verzilli, C., … Mooser, V. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, *337*(6090), 100–104.

O'Connor, T. D., Kiezun, A., Bamshad, M., Rich, S. S., Smith, J. D., Turner, E., … Akey, J. M. (2013). Fine-scale patterns of population

stratification confound rare variant association tests. *PLoS One*, *8*(7), e65834.

Price, A. L., Kryukov, G. V., de Bakker, P. I. W., Purcell, S. M., Staples, J., Wei, L.-J., & Sunyaev, S. R. (2010). Pooled association tests for rare variants in exon-resequencing studies. *American Journal of Human Genetics*, *86*(6), 832–838.

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.

Schaid, D. J., Mcdonnell, S. K., Sinnwell, J. P., & Thibodeau, S. N. (2013). Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. *Genetic Epidemiology*, *37*(5), 409–418.

Schneider, S., Roessli, D., & Excoffier, L. (2000). Arlequin: A software for population genetics data analysis. *User Manual Ver*, *2*, 2496–2497.

Sha, Q., Zhang, K., & Zhang, S. (2016). A nonparametric regression approach to control for population stratification in rare variant association studies. *Scientific Reports*, *18*(6), 37444.

Song, K., Nelson, M. R., Aponte, J., Manas, E. S., Bacanu, S.-A. A., Yuan, X., … Waterworth, D. M. (2011). Sequencing of Lp-PLA2-encoding PLA2G7 gene in 2000 Europeans reveals several rare loss-of-function mutations. *The Pharmacogenomics Journal*, *12*(5), 425–431.

The 1000 Genomes Project Consortium. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *135*(7422), 56–65.

Thornton, T., & McPeek, M. S. (2010). ROADTRIPS: Case-control association testing with partially or completely unknown population and pedigree structure. *The American Journal of Human Genetics*, *86*(2), 172–184.

Tzeng, J.-Y., Lu, W., & Hsu, F.-C. (2014). Gene-level pharmacogenetic analysis on survival outcomes using gene-trait similarity regression. *The Annals of Applied Statistics*, *8*(2), 1232–1255.

Tzeng, J.-Y., Zhang, D., Pongpanich, M., Smith, C., McCarthy, M. I., Sale, M. M., … Sullivan, P. F. (2011). Studying gene and gene-environment effects of uncommon and common variants on continuous traits: A marker-set approach using gene-trait similarity regression. *American Journal of Human Genetics*, *89*(2), 277–288.

Wang, C., Zhan, X., Bragg-Gresham, J., Kang, H. M., Stambolian, D., Chew, E. Y., … Abecasis, G. R. (2014). Ancestry estimation and control of population stratification for sequence-based association studies. *Nature Genetics*, *46*(4), 409–415.

Wang, C., Zöllner, S., Rosenberg, N. A., Weinblatt, M., & Shadick, N. (2012). A quantitative comparison of the similarity between genes and geography in worldwide human populations. *PLoS Genetics*, *8*(8), e1002886.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, *89*(1), 82–93.

Zawistowski, M., Reppell, M., Wegmann, D., St Jean, P. L., Ehm, M. G., Nelson, M. R., … Zöllner, S. (2014). Analysis of rare variant population structure in Europeans explains differential stratification of gene-based tests. *European Journal of Human Genetics*, *22*(9), 1137–1144.

Zhang, D., & Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics*, *4*(1), 57–74.

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, Y., & Pan, W. (2014). Principal Component Regression and Linear Mixed Model in Association Analysis of Structured Samples: Competitors or Complements? *Genetic Epidemiology*, *39*(3), 149–155.

Zhang, Y., Shen, X., & Pan, W. (2013). Adjusting for population stratification in a fine scale with principal components and sequencing data. *Genetic Epidemiology*, *37*(8), 787–801.

Zhao, G., Marceau, R., Zhang, D., & Tzeng, J.-Y. J.-Y. (2015). Assessing gene-environment interactions for common and rare variants with binary traits using gene-trait similarity regression. *Genetics*, *199*(3), 695–710.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Luo Y, Maity A, Wu MC, et al. On the substructure controls in rare variant analysis: Principal components or variance components?. *Genet Epidemiol*. 2018;42:276–287. https://doi.org/10.1002/gepi.22102