RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Adjustment for Population Stratification via Principal Components in Association Analysis of Rare Variants

Yiwei Zhang, Weihua Guan, and Wei Pan*

Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota

Received 12 April 2012; Revised 11 September 2012; accepted revised manuscript 13 September 2012.
Published online 12 October 2012 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21691

ABSTRACT: For unrelated samples, principal component (PC) analysis has been established as a simple and effective approach to adjusting for population stratification in association analysis of common variants (CVs, with minor allele frequencies MAF > 5%). However, it is less clear how it would perform in analysis of low-frequency variants (LFVs, MAF between 1% and 5%), or of rare variants (RVs, MAF < 5%). Furthermore, with next-generation sequencing data, it is unknown whether PCs should be constructed based on CVs, LFVs, or RVs. In this study, we used the 1000 Genomes Project sequence data to explore the construction of PCs and their use in association analysis of LFVs or RVs for unrelated samples. It is shown that a few top PCs based on either CVs or LFVs could separate two continental groups, European and African samples, but those based on only RVs performed less well. When applied to several association tests in simulated data with population stratification, using PCs based on either CVs or LFVs was effective in controlling Type I error rates, while nonadjustment led to inflated Type I error rates. Perhaps the most interesting observation is that, although the PCs based on LFVs could better separate the two continental groups than those based on CVs, the use of the former could lead to overadjustment in the sense of substantial power loss in the absence of population stratification; in contrast, we did not see any problem with the use of the PCs based on CVs in all our examples.
Genet Epidemiol 37:99–109, 2013. © 2012 Wiley Periodicals, Inc.

KEY WORDS: 1000 Genomes Project; association tests; logistic regression; next-generation sequencing; SNP; SSU test

Introduction

With the availability of next-generation sequencing data, there has been increasing interest in studying associations between complex traits and low-frequency variants (LFVs, with minor allele frequency [MAF] between 1% and 5%) or rare variants (RVs, with MAF < 1%); see two recent reviews [Asimit and Zeggini, 2010; Bansal et al., 2010]. Due to the low MAFs of LFVs and RVs, statistical tests developed for common variants (CVs, with MAF > 5%) in genome-wide association studies (GWASs) may no longer be powerful. Accordingly, there have been intensive efforts in developing new statistical tests for LFVs and RVs. Basu and Pan [2011] conducted a comprehensive review and comparison of many existing association tests for LFVs and RVs with unrelated samples. Although there does not exist a uniformly most powerful test, they used simulated data to demonstrate the generally good performance of the sum of squared score (SSU) test, which has been shown [Pan, 2011] to be closely related to an empirical Bayes test for high-dimensional data [Goeman et al., 2006], kernel machine regression (KMR) [Kwee et al., 2008; Wu et al., 2010, 2011b], genomic-distance based regression (GDBR) [Wessel and Schork, 2006], and the C-alpha test [Neale et al., 2011]. A limitation of their study is the lack of use of real sequence data. Furthermore, Basu and Pan [2011] also did not consider the small sample size issue and use of covariates, which may include principal components (PCs) to adjust for population stratification. Here we use a low-coverage whole-genome sequencing dataset generated by the 1000 Genomes Project [1000 Genomes Project Consortium, 2010] to address the above issues.

Intuitively, population stratification can arise in association studies of LFVs and RVs, and some existing techniques for CVs, for example, PC analysis, might be applicable to LFVs and RVs [Lin and Tang, 2011]. However, two recent studies [Baye et al., 2011; Siu et al., 2012] achieved different conclusions on the relative effectiveness of CV- or RV-based PCs in uncovering population structures. More importantly, to our knowledge, the issue has not been experimentally demonstrated in the context of association tests. Among the many existing techniques for CVs, Wu et al. [2011a] demonstrated that adding a few top PCs as covariates in a regression analysis is a simple and effective approach to adjusting for population stratification for unrelated samples. Hence, we adopt this approach throughout. Furthermore, with the availability of sequence data, as pointed out by Price et al. [2010a], it is not completely clear whether LFVs or RVs can be used to infer genetic ancestry. If so, importantly, it is natural to ask whether using LFVs or RVs (or both LFVs/RVs and CVs) can perform better than using CVs alone in adjusting for population stratification. We show that, in agreement with Siu et al. [2012], based on the 1000 Genomes Project data for two continental groups, 174 African (AFR) and 283 European

*Correspondence to: Wei Pan, Division of Biostatistics, MMC 303, School of Public Health, University of Minnesota, Minneapolis, MN 55455-0392. E-mail: weip@biostat.umn.edu

(EUR) samples, the top PC based on a large number of LFVs could better separate the two groups than that based on CVs; however, the PCs based on either CVs, LFVs, or RVs could not separate the underlying subgroups. More interestingly and perhaps surprisingly, although using PCs based on either CVs or LFVs can effectively control inflated Type I error rates in the presence of population stratification, using PCs based on CVs maintained power while using PCs based on some randomly selected LFVs might suffer from substantial power loss in the absence of population stratification, which was likely due to the high linkage disequilibrium (LD) among the randomly selected LFVs.

Methods

Data

We downloaded a low-coverage whole genome sequencing dataset released in August 2010 on the 1000 Genomes Project web site. The dataset included 629 individuals: 174 Africans (AFR), 283 Europeans (EUR), and 194 Asians; we only used the data from the first two groups. In the first two continental groups, there were four and six subgroups, respectively (Table I). Due to the small sample size, we mainly focus on the two continental groups for association testing with chromosome 1 data, though we will also explore the use of PCA in separating the subgroups with the whole genome data. We defined RVs as single nucleotide polymorphisms (SNPs) with MAFs less than 1%, LFVs as those with MAFs between 1% and 5%, and CVs as those with MAFs greater than 5%. On chromosome 1, among the 694,231 common SNPs in both groups, there were 478,208 CVs, 146,353 LFVs, and 69,670 RVs.

For the purpose of this project, we selected a few regions of multiple LFVs or RVs associated with the continental group. As pointed out by Price et al. [2010a], because spurious associations often arise at differentiated variants whose MAFs are unusually different between different ancestral groups, it is crucial to consider these SNPs when correcting for population stratification. We used sliding windows with various sizes on chromosome 1 and tested the association between the continental group and the LFVs or RVs inside each window using a few statistical tests (discussed below). We identified three regions, termed R1 to R3, as representatives for unusually differentiated LFVs or RVs with various characteristics.

For each region, based on a statistical model and the selected LFVs or RVs from the sequence data, we generated simulated datasets with a simulated disease status for each subject. Then we tested possible association between the generated disease status and the observed SNPs in each region, based on which and the truth we assessed the performance of each test in terms of its statistical power and Type I error. Of particular interest was to investigate how the performance of a test depended on whether and how to use PCs constructed from the genome-wide sequence data.

Statistical Tests

We applied two sets of some representative statistical tests for association analysis of LFVs or RVs. The first set includes the score test, the sum of squared score (SSU) test, the weighted sum of squared score (SSUw) test, the Sum test, and the univariate minimum *P*-value (UminP) test [Pan, 2009], while the second includes the T1, T5, Fp, VT, and EREC tests [Lin and Tang, 2011]. We will first introduce the first set of the five tests. They were chosen based on the following reasons. The score test is a classical test in general statistical applications, asymptotically equivalent to the Wald test and likelihood ratio test. The UminP test is perhaps most popular in association analysis of CVs, as used in GWASs. The Sum test is a representative of the so-called pooled association tests [Han and Pan, 2010], similar to the well-known CAST [Morgenthaler and Thilly, 2007] and CMC test [Li and Leal, 2008]. The SSU test is closely related to GDBR [Wessel and Schork, 2006], KMR [Wu et al., 2010, 2011b], and C-alpha test [Neale et al., 2011]; in an extensive simulation study, Basu and Pan [2011] found that the SSU test performed similarly to the KMR and C-alpha, and was an overall winner with the highest or close to the highest power in association analysis of RVs. With either CVs or RVs, the SSUw test often performed similarly to the SSU test; however, with both RVs and CVs, the SSUw test might perform better [Basu and Pan, 2011]. All the five tests are based on the score vector of a regression model, for example, a generalized linear model (GLM), hence only a reduced model under the null hypothesis H_0 is to be fitted, leading to their being computationally faster and numerically more stable than those based on fitting a full model, for example, the Wald or likelihood ratio test. Importantly, because the tests are formulated in the general regression framework, it is easy to incorporate covariates or extend them to other more complex studies, for example, with censored event times as traits, correlated family data, or multiple traits.

For a binary trait Y_i for subject i with k SNPs $X_i = (X_{i1}, \ldots, X_{ik})'$ and covariates $Z_i = (Z_{i1}, \ldots, Z_{iJ})'$, all the five tests are based on the null model

$$LogitPr(Y_i = 1) = \beta_0 + \sum_{j=1}^{J} Z_{ij}\gamma_j,$$

which is simpler than the full model:

$$LogitPr(Y_i = 1) = \beta_0 + \sum_{j=1}^{k} X_{ij}\beta_j + \sum_{j=1}^{J} Z_{ij}\gamma_j.$$

Table I. A summary of the two continental groups

Group	EUR						AFR			
Subgroups	CEU	FIN	GBR	TSI	MXL	PUR	YRI	LWK	ASW	PUR2
No. of samples	90	36	43	92	17	5	78	67	24	5

We use the additive coding for each SNP; that is, $X_{ij} = 0$, 1, or 2 is the the number of minor alleles in SNP j. Due to the extremely low MAF, it is unlikely to have two copies of the minor allele for a given RV, and thus there is little difference between various coding schemes for RVs. In the current context, for population stratification, we include the top J PCs as covariates.

All the five tests are global tests with the null hypothesis H_0: $\beta = (\beta_1, \ldots, \beta_k)' = 0$; it is global in the sense of not identifying specific zero subcomponents of β. Given a score vector $U = (U_{.1}, \ldots, U_{.k})'$ and its covariance estimate $V = Cov(U)$, the five test statistics are, respectively:

$$Score = U'V^{-1}U,$$

$$SSU = U'U = \sum_{j=1}^{k} U_{.j}^2,$$

$$SSUw = U'diag(V)^{-1}U = \sum_{j=1}^{k} U_{.j}^2/V_{jj},$$

$$Sum = 1'U = \sum_{j=1}^{k} U_{.j},$$

$$UminP = \max_{j=1}^{k} U_{.j}^2/V_{jj},$$

where $diag(V)$ is a diagonal matrix with diagonal elements (V_{jj}'s) of V.

Under H_0, based on the asymptotic Normality of U, $U \sim N(0, V)$, the asymptotic distribution of the first four tests can be easily derived and used to obtain their P-values [Pan, 2009], while numerical integrations with a multivariate normal density can be used for the UminP test [Conneely and Boehnke, 2007]. For relatively small sample sizes, especially with RVs, the above asymptotics may not be applicable. Alternatively, as suggested by other authors [Lin and Tang, 2011; Wu et al., 2011a], we can apply the parametric bootstrap [Efron and Tibshirani, 1993] in the following steps:

(1) fit the null model;
(2) use the fitted null model to generate $Y_{i,b}^*$'s as the bth bootstrap dataset with $b = 1, \ldots, B$;
(3) calculate a test statistic T with the original data $(Y_i, X_i)'$s, and T_b with the bth bootstrap data $(Y_{i,b}^*, X_i)'$s;
(4) the P-value is $\sum_{b=1}^{B} I(|T| > |T_b|)/B$.

We used $B = 200$ throughout (and using $B = 1,000$ gave similar results in all the simulations), though in practice we might need to use a much larger B to achieve a higher level of statistical significance.

As to be shown, for the score test and to a lesser degree for the UminP test, the asymptotics might give inflated Type I error rates, while the bootstrap gave much better results; in contrast, the SSU and SSUw tests are more robust to small samples with Type I error rates always close to the nominal level in all our experiments.

For comparison, we also included the T1, T5, Fp, VT, and EREC tests [Lin and Tang, 2011], all implemented in software SCORE-Seq available at http://www.bios.unc.edu/dlin/software/SCORE-Seq/

As shown by Lin and Tang [2011], a general class of the score-based association tests can be formulated as

$$T_G = \sum_{j=1}^{k} \zeta_j U_j,$$

where ζ_j is a weight for SNP j. Different choices of the weights lead to a variety of tests:

(1) the T1 (or T5) test corresponds to $\zeta_j = 1$ if the MAF of SNP j is less than 1% (or 5%) and $\zeta_j = 0$ otherwise;
(2) in the Fp test, we have $\zeta_j = 1/\sqrt{\tilde{p}_j(1-\tilde{p}_j)}$, where $\tilde{p}_j$ is an estimate of the MAF of SNP j with pseudocounts from the pooled sample, giving higher weights to rarer SNPs [Madsen and Browning, 2009];
(3) the VT test combines multiple tests based on multiple thresholds, and for each threshold, $\zeta_j = 1$ if the MAF of SNP j is less than the threshold and $\zeta_j = 0$ otherwise [Price et al., 2010b]; it is a form of the adaptive Neyman's test [Pan and Shen, 2011];
(4) the EREC test uses $\zeta_j = \tilde{\beta}_j \pm c$ with $\tilde{\beta}_j$ as the (univariate) maximum likelihood estimate of β_j and $c = 1$ for binary traits.

Although an asymptotic null distribution is available for each of the first three tests, it is not available for the EREC test. Furthermore, the asymptotic approximations might result in inflated Type I error rates for RVs. Hence, we only show the results of the second set of the tests with their P-values calculated by the parametric bootstrap with the minimum allowable $B = 10^6$ resamples.

We also note that the score, SSU, SSUw, and Sum tests are also special cases of the general T_G test. In particular, the SSUw test uses the weight $\zeta_j = U_j/V_{jj} \approx \tilde{\beta}_j$ [Pan, 2009], suggesting its close connection to the EREC test, whose weight ζ_j can be regarded as shrinking $\tilde{\beta}_j$ toward constants c or $-c$.

Results

Data Description

As shown in Figure 1, there are clear differences between the MAF distributions of the two continental groups. In particular, the difference seems to be larger for low MAFs than for high MAFs.

We selected three regions, named R1 to R3: the first two contained 19 and 40 consecutive LFVs (and only LFVs) respectively, while the third one consisted of 40 consecutive RVs (and only RVs); we calculated the MAF of any SNP based on the pooled sample. The LFVs or RVs within each region were associated with the continental group; that is, the MAFs of the SNPs were different between the AFR and EUR groups (Figure 2). These three regions also showed different LD patterns (Figure 3): LD was weak in R1, moderate in R3, and strong in R2.

We randomly selected a large number of CVs (or LFVs) from chromosome 1 to construct PCs. As shown in

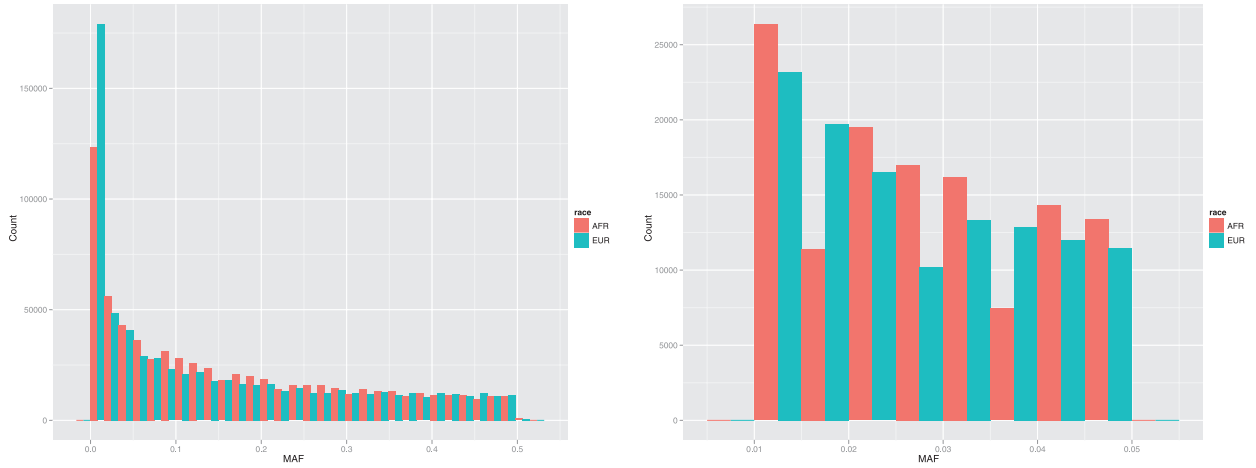


Figure 1. Distributions of MAFs for the EUR and AFR groups.

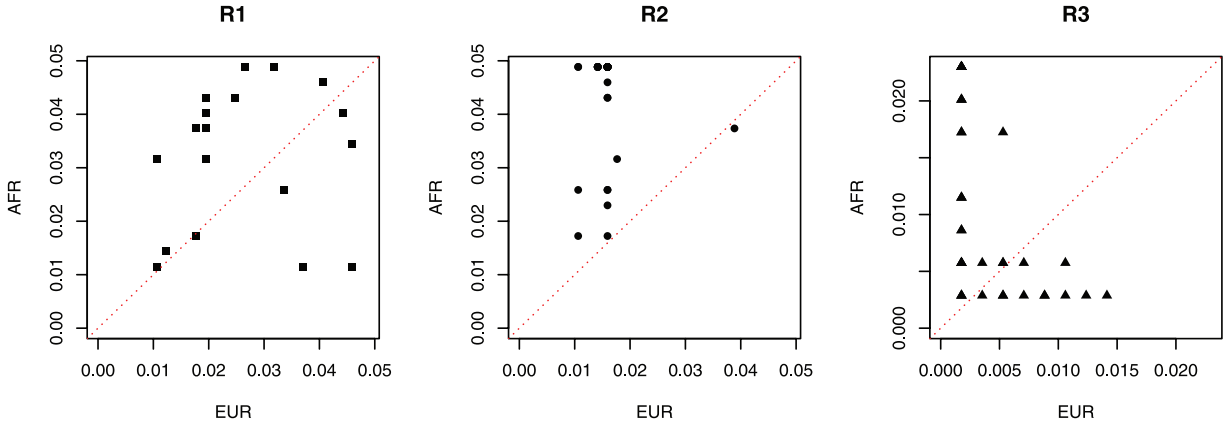


Figure 2. Comparison of the MAFs between the EUR and AFR groups for the SNPs in regions R1–R3.

Figure 4, the top PC based on CVs could largely separate the two AFR and EUR groups; however, perhaps surprisingly, the top PC based on LFVs did better in completely separating the two groups. When using some randomly selected SNPs, including CVs, LFVs, and RVs, the results were between those based on either CVs or LFVs alone (not shown). We will present and discuss results based on RVs later. Because the results with 100,000 CVs (or LFVs) (not shown) were similar, in the following, we used a few top PCs based on either 10,000 CVs or 10,000 LFVs.

Association Testing With LFVS: Type I Error

We first generated simulated data under H_0 with population stratification. Specifically, we randomly selected 90% of the EUR samples and 10% of AFR samples as cases (i.e., $Y_i = 1$), while the remaining ones as controls (i.e., $Y_i = 0$). In this way, none of the SNPs caused the "disease," and there was a clear association between the continental group and the "disease" (i.e., population stratification). We applied the five tests to the two LFV regions with 1,000 simulated datasets for each case. Because the results are similar for PCs based on either CVs or LFVs, we only show that for the former. Table II lists the Type I error rates at the nominal level $\alpha = 0.05$. It is clear that, without adjustment for population stratification, all the tests could have dramatically inflated Type I error rates (except the Sum test for R1), suggesting the necessity of adjusting for population stratification. With PCs, including with even just the single top PC (i.e., #PCs=1), the problem with inflated Type I error rates largely disappeared; there was almost no difference between using various numbers of PCs, as long as at least one PC was used.

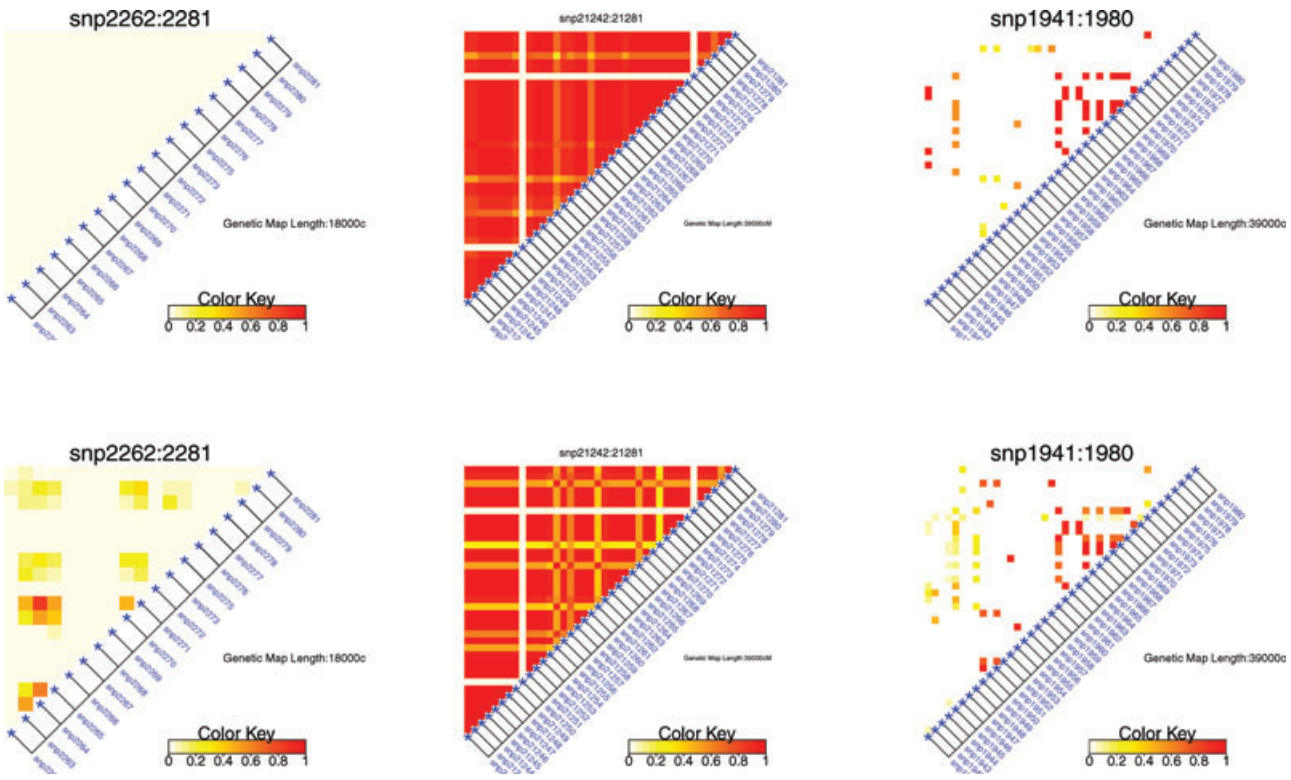
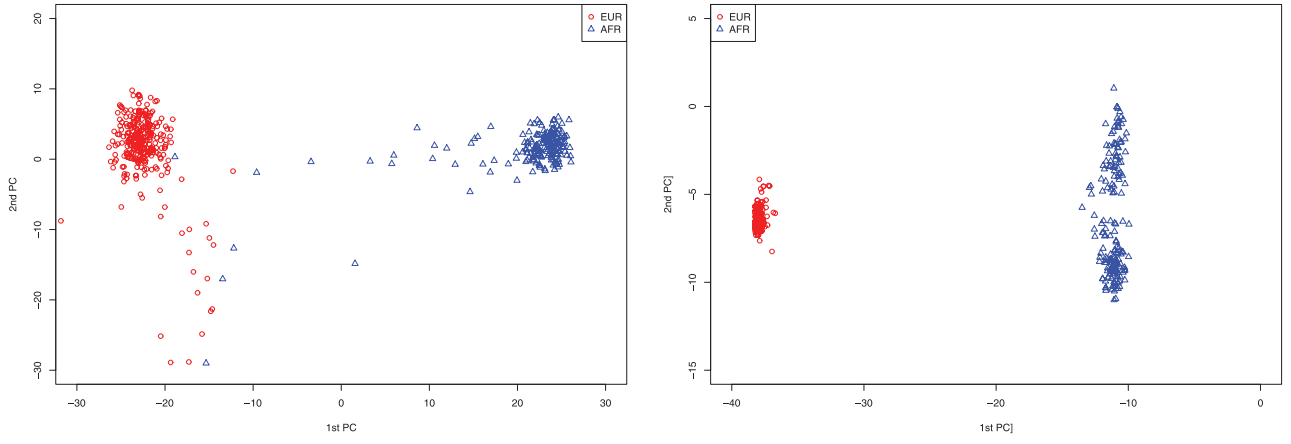


Figure 3. LD plots in r^2 for the EUR (top row) and AFR (bottom row) groups in regions R1–R3.



(a) Based on 10,000 CVs (b) Based on 10,000 LFVs

Figure 4. The top two PCs constructed with CVs or LFVs.

It is noted that the asymptotics-based score test could have severely inflated Type I error rates, even in the presence of PCs for region R2, and that the asymptotics-based UminP test could also have slightly inflated Type I error rates. The bootstrap-based tests all had their Type I error rates better controlled.

Association Testing With LFVS: Power

We generated a disease status from the following logistic regression model:

$$LogitPr(Y_i = 1) = \beta_0 + \sum_{j=1}^{k_1} X_{ij}\beta_j,$$

Table II. Type I error rates with population stratification. The PCs were constructed using 10,000 CVs

		Asymptotics				Bootstrap			
Loc	Test	#PCs=0	1	5	10	#PCs=0	1	5	10
R1	Score	0.693	0.055	0.052	0.048	0.716	0.044	0.044	0.042
	SSU	0.618	0.062	0.039	0.035	0.647	0.061	0.039	0.044
	SSUw	0.620	0.052	0.036	0.035	0.642	0.053	0.038	0.037
	Sum	0.067	0.066	0.050	0.047	0.070	0.044	0.048	0.048
	UminP	0.201	0.084	0.065	0.067	0.232	0.068	0.039	0.047
R2	Score	0.155	0.180	0.192	0.171	0.624	0.062	0.071	0.061
	SSU	0.709	0.053	0.055	0.054	0.684	0.047	0.051	0.055
	SSUw	0.700	0.052	0.055	0.054	0.669	0.049	0.052	0.055
	Sum	0.684	0.054	0.055	0.055	0.652	0.049	0.059	0.056
	UminP	0.677	0.052	0.061	0.066	0.685	0.044	0.050	0.047

Table III. Empirical power of various tests based on the parametric bootstrap for the two regions with k_1 causal SNPs. The PCs were constructed using 10,000 CVs

Loc	Test	#PCs=0	1	5	10	#PCs=0	1	5	10
		$\beta_i \sim U(-\log 3, \log 3)$				$\beta_i \sim U(0, \log 3)$			
R1 $k_1 = 8$	Score	0.489	0.489	0.489	0.496	0.838	0.836	0.824	0.826
	SSU	0.500	0.492	0.497	0.501	0.882	0.880	0.891	0.881
	SSUw	0.507	0.480	0.481	0.486	0.883	0.883	0.889	0.886
	Sum	0.240	0.230	0.234	0.230	0.860	0.860	0.856	0.852
	UminP	0.401	0.397	0.393	0.383	0.813	0.820	0.813	0.802
		$\beta_i \sim U(-\log 2, \log 2)$				$\beta_i \sim U(0, \log 1.5)$			
R1 $k_1 = 19$	Score	0.483	0.467	0.479	0.475	0.504	0.504	0.493	0.479
	SSU	0.507	0.493	0.511	0.492	0.773	0.771	0.758	0.738
	SSUw	0.479	0.479	0.483	0.477	0.769	0.764	0.765	0.756
	Sum	0.207	0.202	0.201	0.204	0.842	0.839	0.839	0.832
	UminP	0.288	0.286	0.280	0.287	0.558	0.544	0.546	0.541
		$\beta_i \sim U(-\log 3, \log 3)$				$\beta_i \sim U(0, \log 2)$			
R2 $k_1 = 4$	Score	0.256	0.240	0.251	0.244	0.707	0.702	0.699	0.687
	SSU	0.401	0.406	0.400	0.406	0.786	0.787	0.785	0.793
	SSUw	0.404	0.404	0.403	0.408	0.783	0.787	0.787	0.795
	Sum	0.405	0.406	0.406	0.409	0.784	0.785	0.789	0.793
	UminP	0.360	0.344	0.347	0.349	0.761	0.763	0.758	0.756
		$\beta_i \sim U(-\log 2, \log 2)$				$\beta_i \sim U(0, \log 1.1)$			
R2 $k_1 = 30$	Score	0.364	0.365	0.362	0.366	0.776	0.761	0.765	0.761
	SSU	0.601	0.602	0.607	0.606	0.871	0.868	0.871	0.869
	SSUw	0.601	0.603	0.607	0.610	0.873	0.867	0.872	0.869
	Sum	0.599	0.601	0.602	0.605	0.874	0.866	0.869	0.869
	UminP	0.563	0.550	0.546	0.544	0.848	0.834	0.836	0.835

where X_{ij} was the jth SNP of the ith subject (AFR or EUR), $\beta_0 = -\log 3$ was chosen to generate a background disease incidence of 25% (when all $X_{ij} = 0$), and the causal effect sizes β_j were randomly generated from a uniform distribution $U(-a, a)$ or $U(0, a)$ for a constant $a > 0$. With $U(-a, a)$, some causal effects were deleterious while others were protective against disease; with $U(0, a)$, all causal effects were in the same direction of being deleterious. We used $k_1 \leq k$: if $k_1 < k$, we randomly selected a subset of the SNPs to be causal while others were neutral or noncausal, but a test was always applied to all the k SNPs; it is important to assess a test's robustness to the number of noncausal SNPs, since in practice we expect causal SNPs to be mixed with some neighboring noncausal ones. Each subject i's genotype was input to the above model to generate his/her disease status. In such a way, we generated a dataset of 457 subjects with various numbers of cases and controls.

Because the general conclusions remained the same, we only chose a small subset of results to present in Tables III and IV. Recall that there were weak and strong LD, and a small and a large number of LFVs, in the two regions R1 and R2, respectively. The PCs were constructed based on the 10,000 randomly selected CVs. First, because all the SNPs had MAF between 1% and 5%, the T1 test was not applicable (with all weights $\zeta_j = 0$), and the T5 test (with all weights $\zeta_j = 1$) was essentially the same as the Sum test. In addition, because the causal SNPs were selected randomly and were not correlated with lower or higher MAFs, the Fp and VT tests were not expected to improve over the T5 and Sum tests. Second, because there was no population stratification, it is good to

Table IV. Empirical power of various tests based on the parametric bootstrap for region R1. The PCs were constructed using either 10,000 CVs or 10,000 LFVs

		10,000 CVs				10,000 LFVs			
Loc	Test	#PCs=0	1	5	10	#PCs=0	1	5	10
		$\beta_i \sim U(-\log 3, \log 3)$				$\beta_i \sim U(-\log 3, \log 3)$			
R1 $k_1 = 8$	T5	0.200	0.207	0.195	0.188	0.200	0.205	0.193	0.182
	Fp	0.203	0.197	0.186	0.192	0.203	0.195	0.184	0.179
	VT	0.214	0.212	0.208	0.199	0.214	0.211	0.202	0.195
	EREC	0.401	0.399	0.381	0.375	0.401	0.401	0.373	0.368
		$\beta_i \sim U(0, \log 3)$				$\beta_i \sim U(0, \log 3)$			
R1 $k_1 = 8$	T5	0.872	0.878	0.869	0.857	0.872	0.876	0.866	0.812
	Fp	0.872	0.873	0.864	0.852	0.872	0.872	0.856	0.805
	VT	0.838	0.832	0.826	0.813	0.838	0.820	0.784	0.755
	EREC	0.916	0.914	0.903	0.891	0.916	0.912	0.873	0.850
		$\beta_i \sim U(-\log 2, \log 2)$				$\beta_i \sim U(-\log 2, \log 2)$			
R1 $k_1 = 19$	T5	0.222	0.225	0.223	0.216	0.222	0.232	0.215	0.193
	Fp	0.220	0.223	0.214	0.209	0.220	0.219	0.208	0.198
	VT	0.230	0.235	0.228	0.229	0.230	0.225	0.233	0.200
	EREC	0.395	0.396	0.398	0.382	0.395	0.390	0.392	0.375
		$\beta_i \sim U(0, \log 1.5)$				$\beta_i \sim U(0, \log 1.5)$			
R1 $k_1 = 19$	T5	0.844	0.846	0.845	0.828	0.844	0.843	0.823	0.768
	Fp	0.849	0.852	0.837	0.827	0.849	0.852	0.806	0.762
	VT	0.779	0.779	0.752	0.747	0.779	0.760	0.729	0.685
	EREC	0.826	0.820	0.807	0.796	0.826	0.817	0.782	0.698

see that using or not using PCs, or using different numbers of PCs, gave similar results for all the tests. We emphasize that this is a desired property. In practice, for a given dataset, population stratification may or may not be present; to be safe in avoiding spurious associations, we might still want to apply an adjustment, for example, based on PCs. Hence, it would be desirable to have no or minimum power loss when adjusting for population stratification. Third, the identity of the most powerful tests varied with the setup. For example, in region R1,

(1) with all 19 causal SNPs being deleterious, the Sum, T5, Fp, and EREC tests performed similarly and were most powerful;
(2) with the 19 causal SNPs with opposite association directions, as expected, the Sum, T5, and Fp tests were low powered, while the SSU, SSUw, and score tests were most powerful. Although there was no uniformly most powerful test, the SSU and SSUw tests seemed to be the overall winners.

Fourth, due to the small sample size, the asymptotics-based score test might lose power as compared to the bootstrap-based score test (not shown). In contrast, other tests seemed to be more robust to small samples: their asymptotics-based version and bootstrap-based version always gave similar results (not shown).

Because the PCs based on LFVs could better separate the AFR and EUR groups, it would be interesting to see the performance of the tests with PCs constructed from LFVs. In many situations, a test with LFV-based PCs and with CV-based PCs performed similarly; however, as shown in Tables IV and V, when all the causal effects were in the same direction, it is clear that adjusting with more than one PC led to power loss, which was often substantial. For example, in the setup with 30 causal SNPs with positive effects in region R2 (Table V), the SSU and SSUw tests were most powerful; however, with 1, 5, and 10 PCs, the power of the SSU test monotonically decreased from 0.871 to 0.865, 0.803, and 0.781, respectively. This is a case called overadjustment in the sense of losing substantial power when adjusting for population stratification (or more generally, confounders). This phenomenon was not specific to the first set of the five tests shown in Table V; it also appeared for the second set (Table IV): for example, for region R1 with $k_1 = 8$ causal SNPs with the same effect direction, the power of the T5, Fp, VT, and EREC tests reduced, respectively, from 0.872, 0.872, 0.838, and 0.916 with no PC to 0.812, 0.805, 0.755, and 0.850 with 10 PCs constructed from 10,000 LFVs. It is interesting to note that, in all our examples, using only the top PC could largely control the Type I error rate while maintaining the power (with no or negligible power loss).

We explored the reason for the overadjustment. We first hypothesized that the LFV-based PCs might reflect some hidden ethnic structure. When adjusting for ancestry using either reported two continental groups or reported ethnic subgroups, the test results were similar to those with no or only one PC; in other words, there was no loss of power. Second, we regressed the binary trait on the top 10 PCs and referred to the corresponding linear combination of the top 10 PCs as a PCs-defined group score. We found that in the cases with overadjustment, the PCs-defined group score was much more significantly associated with the sum of the LFVs to be tested

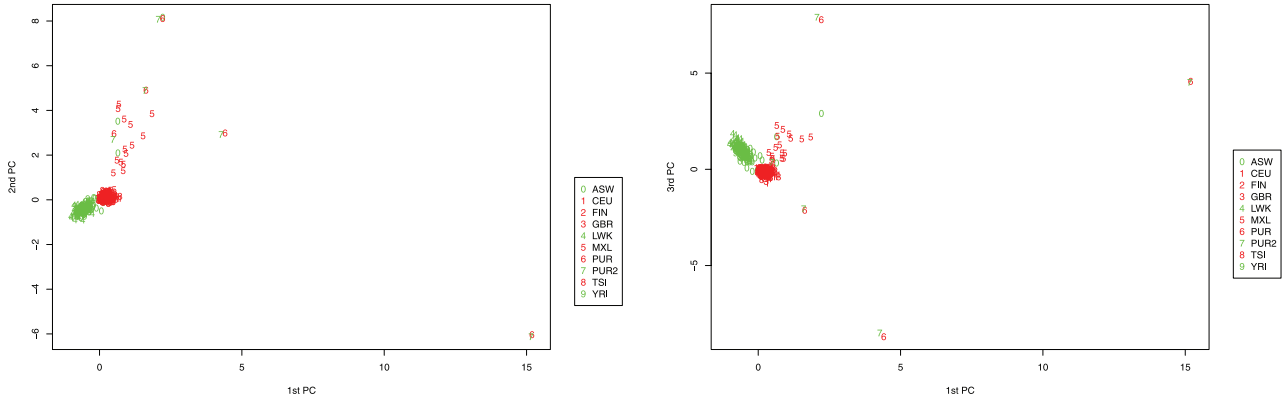
Table V. Empirical power of various tests based on the parametric bootstrap for the two regions with k_1 causal SNPs. The PCs were constructed using 10,000 LFVs

Loc	Test	#PCs=0	1	5	10	#PCs=0	1	5	10
		$\beta_i \sim U(-\log 3, \log 3)$				$\beta_i \sim U(0, \log 3)$			
R1 $k_1 = 8$	Score	0.489	0.498	0.499	0.486	0.838	0.836	0.805	0.787
	SSU	0.500	0.502	0.495	0.506	0.882	0.881	0.880	0.851
	SSUw	0.507	0.490	0.489	0.493	0.883	0.885	0.882	0.857
	Sum	0.240	0.223	0.226	0.219	0.860	0.857	0.850	0.821
	UminP	0.401	0.402	0.386	0.396	0.813	0.822	0.792	0.753
		$\beta_i \sim U(-\log 2, \log 2)$				$\beta_i \sim U(0, \log 1.5)$			
R1 $k_1 = 19$	Score	0.483	0.474	0.481	0.490	0.504	0.497	0.465	0.452
	SSU	0.507	0.501	0.507	0.503	0.773	0.765	0.724	0.663
	SSUw	0.479	0.476	0.484	0.488	0.769	0.763	0.727	0.681
	Sum	0.207	0.202	0.194	0.180	0.842	0.835	0.819	0.791
	UminP	0.288	0.286	0.273	0.283	0.558	0.547	0.498	0.441
		$\beta_i \sim U(-\log 3, \log 3)$				$\beta_i \sim U(0, \log 2)$			
R2 $k_1 = 4$	Score	0.256	0.242	0.233	0.229	0.707	0.703	0.628	0.616
	SSU	0.401	0.410	0.386	0.382	0.786	0.790	0.734	0.713
	SSUw	0.404	0.408	0.384	0.380	0.783	0.791	0.737	0.713
	Sum	0.405	0.407	0.384	0.380	0.784	0.786	0.737	0.712
	UminP	0.360	0.341	0.330	0.317	0.761	0.762	0.703	0.672
		$\beta_i \sim U(-\log 2, \log 2)$				$\beta_i \sim U(0, \log 1.1)$			
R2 $k_1 = 30$	Score	0.364	0.361	0.347	0.337	0.776	0.761	0.657	0.634
	SSU	0.601	0.604	0.585	0.582	0.871	0.865	0.803	0.781
	SSUw	0.601	0.600	0.583	0.580	0.873	0.866	0.805	0.782
	Sum	0.599	0.604	0.579	0.578	0.874	0.863	0.804	0.787
	UminP	0.563	0.546	0.511	0.503	0.848	0.833	0.770	0.737

than those in other cases. Although the LFVs were randomly selected to construct PCs, we found that surprisingly many of them were highly correlated, as shown by an exome sequence dataset from the 1000 Genomes Project [Tintle et al., 2011]. For CVs, it is highly recommended to use only nearly independent SNPs to construct PCs [Lee et al., 2011; Patterson, 2006] to avoid the resulting PCs' representing some peculiar features of the data. Hence, we first tried to remove highly correlated LFVs by using PLINK with a threshold of $r^2 \leq 0.5$ or $r^2 \leq 0.05$, respectively. Then using the top 10 PCs constructed with the remaining LFVs, we obtained the results (not shown) similar to those without adjustment (or to using CV-based PCs). In conclusion, the multiple PCs based on the original possibly highly correlated LFVs perhaps represented some unknown and possibly artificial structure in the data.

Subgroup Analysis With RVS

We used a subset of 786,487 nonmonomorphic RVs from chromosomes 2 to 22 to construct PCs. We first used PLINK



(a) PC1 vs PC2 (b) PC1 vs PC3

Figure 5. The top three PCs constructed with 10,000 RVs. The red/dark ones are the EUR samples and the green/gray ones are the AFR samples.

Table VI. The *P*-values of the Tracy-Widom (TW) test and one-way ANOVA applied to the eigenvalues or PCs constructed from 10,000 RVs

# Eigenvalue or PC	TW	2-group ANOVA	10-subgroup ANOVA
1	0	6.667×10^{-24}	2.955×10^{-25}
2	0	3.061×10^{-21}	9.428×10^{-28}
3	0	1.212×10^{-28}	6.830×10^{-27}
4	0	0.1283	0.0393
5	0	3.165×10^{-33}	8.635×10^{-34}
6	0	6.745×10^{-10}	1.593×10^{-27}
7	0	9.098×10^{-8}	5.310×10^{-18}
8	0	0.0349	1.002×10^{-25}
9	0	0.1118	1.221×10^{-55}
10	0	0.0002	3.654×10^{-9}
11	0	0.2260	1.376×10^{-5}
12	0	0.9646	0.9393
13	0	0.0013	9.934×10^{-9}
14	0	0.1094	0.3517
15	0	0.2797	0.3142
16	0	0.8881	0.1215
17	0	0.4054	0.0243
18	3.800×10^{-3}	0.9982	0.9916
19	4.230×10^{-2}	0.1583	0.7185
20	1.580×10^{-1}	0.7069	0.3861

[Purcell et al., 2007] to prune correlated SNPs with a sliding window of size 50 (shifted by 5) and a threshold of $r^2 < 0.05$; after this thinning process, we had 305,036 RVs. We then selected a random set of 10,000 RVs to construct PCs. As shown in Figure 5, the first PC could largely separate the two continental groups, but not the 10 subgroups. Several PUR and PUR2 samples appeared to be outliers, which might have unduly influenced the PCA results; on the other hand, it may be argued that the RVs could better separate the PUR and PUR2 samples from other subgroups.

We applied the Tracy-Widom (TW) test [implemented in R package `EigenCorr`, Lee et al., 2011], yielding a statistically significant *P*-value less than 0.05 for each of the top 19 eigenvalues (Table VI). We also applied one-way ANOVA to test the significance of each PC with varying mean values across the two continental groups or the 10 subgroups; in both cases, the most significant PCs were in the top 13.

A visual examination of the scatter plots of the top PCs did not reveal that the top PCs could separate the 10 subgroups. Hence, we applied finite Gaussian mixture model-based clustering (implemented in R package `mclust`) to top 10, 20, and 50 PCs. Based on the Rand index (calculated using R package `clue`), using top 20 PCs led to the highest agreement between the resulting 12 clusters and 10 true subgroups (with a Rand index value of 0.812 and an adjusted Rand index value of 0.408). As shown in Table VII, in agreement with Figure 5, although the two continental groups could be largely but not perfectly separated, the subgroups in the EUR group could not be distinguished: most of them were mixed into two clusters.

Association Testing With RVS

We also conducted a simulation study with the 40 RVs in region R3. To assess the Type I error rates, we generated a binary trait as before under population stratification; for power, we randomly selected $k_1 = 15$ RVs as causal ones with their effect sizes $\beta_j \sim U(0, \log(3))$. We used the parametric bootstrap for *P*-value calculation for each test. Note that because all the 40 SNPs had MAF less than 1%, the results for the T1 and T5 tests were exactly the same. As shown in Table VIII, under H_0, if no adjustment was made, all the tests resulted in dramatically inflated Type I errors. Because the first PC could not completely separate the two continental groups (Fig. 5), using only the top PC still yielded largely inflated Type I error rates. In contrast, using the top 10 or 20 PCs could largely remedy the problem, though there were still some slightly inflated Type I error rates for some tests, which could be due to the fact that even the top PCs could not completely separate the two continental groups (Fig. 5 and Table VII). For power, with the exception of the score, SSU and SSUw tests, all other tests seemed to have some power loss with 20 PCs.

Table VII. The numbers of samples assigned to each of the 12 clusters based on top 20 PCs constructed from 10,000 RVs

Subgroup/ Cluster	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
CEU	76	14	0	0	0	0	0	0	0	0	0	0
FIN	9	27	0	0	0	0	0	0	0	0	0	0
GBR	29	13	0	0	0	0	1	0	0	0	0	0
TSI	78	14	0	0	0	0	0	0	0	0	0	0
MXL	0	0	0	0	0	3	14	0	0	0	0	0
PUR	0	0	0	0	0	0	0	1	1	1	1	1
YRI	0	0	68	1	9	0	0	0	0	0	0	0
LWK	0	0	0	38	24	5	0	0	0	0	0	0
ASW	0	0	1	0	14	9	0	0	0	0	0	0
PUR2	0	0	0	0	0	0	0	1	1	1	1	1

Table VIII. Type I error and power for region R3. The PCs were constructed using 10,000 RVs

	Type I error				Power			
Test	#PCs=0	1	10	20	#PCs=0	1	10	20
Score	0.972	0.521	0.114	0.086	0.525	0.519	0.504	0.537
SSU	0.995	0.301	0.040	0.052	0.654	0.639	0.623	0.640
SSUw	0.992	0.542	0.056	0.070	0.659	0.652	0.634	0.652
Sum	0.995	0.818	0.076	0.050	0.671	0.664	0.628	0.630
UminP	0.900	0.561	0.108	0.088	0.492	0.476	0.427	0.434
T1, T5	0.995	0.821	0.061	0.038	0.663	0.673	0.624	0.608
Fp	0.993	0.816	0.064	0.041	0.658	0.654	0.615	0.606
VT	0.984	0.647	0.056	0.057	0.605	0.590	0.537	0.533
EREC	0.997	0.119	0.012	0.066	0.662	0.648	0.609	0.594

Discussion

We have used a low-coverage whole-genome sequencing dataset generated by the 1000 Genomes Project to empirically investigate some characteristics of LFVs or RVs that are relevant to their association analysis. For example, some might argue that, due to the low MAFs, LFVs, and RVs are expected to be independent; we have demonstrated that the neighboring LFVs or RVs in a region may be in either low,

moderate, or high LD, suggesting that future studies on the performance of any association test should consider varying LD as a factor. Furthermore, as a useful complement to the extensive simulation studies of Basu and Pan [2011], we have used real sequence data to demonstrate the power properties of the various tests with or without PCs, though it was not the main aim of the current study. In particular, it is confirmed that the Sum test, a representative of simple pooled association tests [Dering et al., 2011], is not powerful in the presence of different association directions or of many non-causal SNPs; in contrast, the SSU and SSUw tests are much more powerful in these situations. It is also shown that the asymptotics of the Sum, SSU, and SSUw tests seemed to work well with a reasonable sample size for LFVs, much more robust than the score test. Of course, with small samples sizes or RVs with extremely low MAFs, one has to be cautious in using asymptotics. As shown here and in other places [Tang and Lin, 2011; Wu et al., 2011b], the parametric bootstrap is a useful alternative. Given the generally good performance of the SSU and SSUw tests, we would recommend their use in practice; if the applicability of the asymptotics is of concern, a two-step procedure can be taken: one could first use the asymptotics-based SSU or SSUw test to quickly scan the genome, then apply the more computing-intensive bootstrap-based SSU or SSUw test to the more significant regions identified in the first step.

Perhaps the most interesting finding of this study is that, in accordance with Siu et al. [2012] but differing from Baye et al. [2011], PCs constructed with LFVs could potentially separate different continental or ethnic groups better than those with CVs, though either can be used to adjust for population stratification effectively. We note that Siu et al. [2012] used a similar whole genome sequence dataset as ours while Baye et al. [2011] used a smaller subset of the exome sequence dataset with much fewer LFVs or RVs. In addition, differing from Mathieson et al. [2012], we focused on two relatively well-separated populations, that is, AFR and EUR samples; further studies are warranted for other more challenging cases. In all our numerical examples, in contrast to that using PCs based on CVs led to no or little power loss in the absence of population stratification, surprisingly using multiple PCs based on LFVs might result in *overadjustment* in the sense of substantial power loss. It is also interesting to note that, in all our examples, using only the top PC based on LFVs could largely control the Type I error rate while maintaining the power (with no or minimum power loss). The overadjustment with multiple PCs based on LFVs in our experiments was likely due to the use of many LFVs in high LD; once we used LFVs not in high LD, the problem largely disappeared. This is in agreement with two known results: first, it is highly recommended to use only almost independent CVs to construct PCs [Lee et al., 2011; Patterson et al., 2006]; second, for unknown reasons, there seems to exist long-range correlations among LFVs or RVs in real sequence data [Tintle et al., 2011]. Hence, one has to be careful in selecting LFVs or RVs to construct PCs; in particular, a random subset of far-away LFVs or RVs may not be sufficient. Furthermore, our preliminary analysis also shows that PCA of RVs with MAFs < 1% might not be effective in separating subpopulations. One possible reason is the sensitivity of PCA to outliers, which are present with some diverse subpopulations and largely varying numbers of subpopulation samples; it would be interesting to apply other more robust methods [e.g., Lee et al., 2011]. We also emphasize that our conclusions are based on the use of a low-coverage whole-genome sequencing dataset, which may be different from high-coverage sequencing data; for example, high-coverage sequencing tends to uncover more RVs [Tennessen et al., 2012]. Importantly, we only considered using CVs, LFVs, or RVs, but not their combined use; it remains to be investigated how to select and combine CVs, LFVs, and RVs to best capture population structures. Finally, because our current study focuses on unrelated samples and the PC-based adjustment for population stratification, it would be interesting to investigate the same issues with other adjustment methods [e.g., Engelhardt and Stephens, 2010; Guan et al., 2009; Lee et al., 2010; Pritchard et al., 2000; Zhu et al., 2002, Zhu et al., 2008] or for family studies [Feng et al., 2011; Zhu et al., 2010].

Acknowledgments

We thank the reviewers for many helpful and constructive comments and suggestions. Y.Z. and W.P. were supported by NIH grants R21DK089351, R01HL65462, R01HL105397, and R01GM081535.

Web Resources

R code will be posted on our web site at http://www.biostat.umn.edu/∼weip/prog.html.

References

1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467: 1061–1073.

Asimit J, Zeggini E. 2010. Rare variant association analysis methods for complex traits. *Annu Rev Genet* 44:293–308.

Bansal V, Libiger O, Torkamani A, Schork NJ. 2010. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* 11:773–785.

Basu S, Pan W. 2011. Comparison of statistical tests for association with rare variants. *Genet Epidemiol* 35:606–619.

Baye TM, He H, Ding L, Kurowski BG, Zhang X, Martin LJ. 2011. Population structure analysis using rare and common functional variants. *BMC Proc* 5(Suppl 9):S8.

Conneely KN, Boehnke M. 2007. So many correlated tests, so little time! Rapid adjustment of p values for multiple correlated tests. *Am J Hum Genet* 81: 1158–1168.

Dering C, Hemmelmann C, Pugh E, Ziegler A. 2011. Statistical analysis of rare sequence variants: an overview of collapsing methods. *Genet Epidemiol* 35(S1):S12–S17.

Efron B, Tibshirani R. 1993. *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman and Hall/CRC.

Engelhardt BE, Stephens M. 2010. Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS Genet* 6(9):e1001117.

Feng T, Zhu X. 2010. Genome-wide searching of rare genetic variants in WTCCC data. *Hum Genet* 128:269–280.

Goeman JJ, van de Geer S, van Houwelingen HC. 2006. Testing against a high dimensional alternative. *J R Stat Soc* B 68:477–493.

Guan W, Liang L, Boehnke M, Abecasis GR. 2009. Genotype-based matching to correct for population stratification in large-scale case-control genetic association studies. *Genet Epidemiol* 33:508–517.

Han F, Pan W. 2010. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered* 70:42–54.

Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. 2008. A powerful and flexible multilocus association test for quantitative traits. *Am J Hum Genet* 82:386–397.

Lee AB, Luca D, Klei L, Devlin B, Roeder K. 2010. Discovering genetic ancestry using spectral graph theory. *Genet Epidemiol* 34:51–59.

Lee S, Wright FA, Zou F. 2011. Control of population stratification by correlation-selected principal components. *Biometrics* 67:967–974.

Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 83:311–321.

Li Y, Byrnes AE, Li M. 2010. To identify associations with rare variants, Just WHaIT: weighted haplotype and imputation-based tests. *Am J Hum Genet* 87:728–735.

Lin DY, Tang ZZ. 2011. A general framework for detecting disease associations with rare variants in sequencing studies. *Am J Hum Genet* 89:354–367.

Madsen BE, Browning SR. 2009. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 5(2):e1000384.

Mathieson I, McVean G. 2012. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* 44:243–246.

Morgenthaler S, Thilly WG. 2007. 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.

Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Ogho-Melander M, Katherisan S, Purcell SM, Roeder K, Daly MJ. 2010. Testing for an unusual distribution of rare variants. *PLoS Genet* 7(3):e1001322.

Pan W. 2009. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet Epidemiol* 33:497–507.

Pan W. 2011. Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. *Genet Epidemiol* 35:211–216.

Pan W, Shen X. 2011. Adaptive tests for association analysis of rare variants. *Genet Epidemiol* 35:381–388.

Patterson N, Price A, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet* 2(12):e190.

Price AL, Zaitlen NA, Reich D, Patterson N. 2010a. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* 11:459–463.

Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei L-J, Sunyaev SR. 2010b. Pooled association tests for rare variants in exon-resequenced studies. *Am J Hum Genet* 86:832–838.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945–959.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575.

Siu H, Jin L, Xiong M. 2012. Manifold learning for human population structure studies. *PLoS One* 7(1):e29901.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, Leal SM, Gabriel S, Rieder MJ, Abecasis G, Altshuler D, Nickerson DA, Boerwinkle E, Sunyaev S, Bustamante CD, Bamshad MJ, Akey JM, NHLBI Exome Sequencing Project. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337(6090):64–69

Tintle N, Aschard H, Hu I, Nock N, Wang H, Pugh E. 2011. Inflated type I error rates when using aggregation methods to analyze rare variants in the 1000 Genomes Project exon sequencing data in unrelated individuals: summary results from Group 7 at Genetic Analysis Workshop 17. *Genet Epidemiol* 35(S1):S56–S60.

Wessel J, Schork NJ. 2006. Generalized genomic distance-based regression methodology for multilocus association analysis. *Am J Hum Genet* 79:792–806.

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. 2010. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet* 86:929–942.

Wu C, DeWan A, Hoh J, Wang Z. 2011a. A comparison of association methods correcting for population stratification in case-control studies. *Annals Hum Genet* 75:418–427.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011b. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 89:82–93.

Zhu X, Zhang S, Zhao H, Cooper RS. 2002. Association mapping, using a mixture model for complex traits. *Genet Epidemiol* 23:181–196.

Zhu X, Li S, Cooper RS, Elston RC. 2008. A unified association analysis approach for family and unrelated samples. *Am J Hum Genet* 82:352–365.

Zhu X, Feng T, Li Y, Lu Q, Elston RC. 2010. Detecting rare variants for complex traits using family and unrelated data. *Genet Epidemiol* 34:171–187.