Hindawi
International Journal of Genomics
Volume 2020, Article ID 4708152, 10 pages
https://doi.org/10.1155/2020/4708152

Research Article

Statistical Method Based on Bayes-Type Empirical Score Test for Assessing Genetic Association with Multilocus Genotype Data

Yi Tian,[1] Li Ma,[2] Xiaohong Cai,[2] and Jiayan Zhu[2]

[1]*School of Mathematics and Statistics, Central China Normal University, Wuhan 430079, China*
[2]*School of Information Engineering, Hubei University of Chinese Medicine, Wuhan 430065, China*

Correspondence should be addressed to Jiayan Zhu; zhujiayan999@163.com

Received 11 December 2019; Accepted 21 April 2020; Published 7 May 2020

Academic Editor: Atsushi Kurabayashi

Simultaneous testing of multiple genetic variants for association is widely recognized as a valuable complementary approach to single-marker tests. As such, principal component regression (PCR) has been found to have competitive power. We focus on exploring a robust test for an unknown genetic mode of all SNPs, an unknown Hardy-Weinberg equilibrium (HWE) in a population, and a large number of all SNPs. First, we propose a new global test by means of the use of codominant codes for all markers and PCR. The new global test is built on an empirical Bayes-type score statistic for testing marginal associations with each single marker. The new global test gains power by robustly exploiting the Hardy-Weinberg equilibrium in the control population and effectively using linkage disequilibrium among test markers. The new global test reduces to PCR when the genotype for each marker is coded as the number of minor alleles. This connection lends insight into the power of the new global test relative to PCR and some other popular multimarker test methods. Second, we propose a robust test method based on the new global test and the ordinary PCR test built on a prospective score statistic for testing marginal associations with each single marker when the genotype for each marker is coded as the number of minor alleles by taking the minimum p value of these two tests. Finally, through extensive simulation studies and analysis of the association between pancreatic cancer and some genes of interest, we show that the proposed robust test method has desirable power and can often identify association signals that may be missed by existing methods.

1. Introduction

Association analyses that test multiple genetic markers as a set rather than individually have been appreciated for their potential power. These statistical methods largely fall into three classes: those for summarizing p values from the tests of each single marker [1–5], those that synthesize single-marker test statistics, such as Hotelling T^2 (standard Chi-squared) statistic [6–8] and the burden test [9, 10], and those based on a direct test of joint associations of multiple markers, such as variance component tests (VC) [11–13], the sequence kernel association test (SKAT) [14–18], and principal component regression (PCR) methods [19–21]. The relative performance of these methods has been comprehensively compared in previous work [22]. When the number of single-nucleotide polymorphisms (SNPs) is small, these methods have similar power; however, when the number of SNPs is large, the effects of SNPs are not constant and may have different directions, the linkage disequilibrium (LD) among multiple markers is somewhat strong, and the SNPs adopt additive genetic code. Three methods, namely, VC, SKAT, and PCR, have been found to have competitive power in this case [22, 23]. A major reason is that all 3 methods can decrease the degrees of freedom of the test to some extent [12]. In this work, we focus on exploring a robust test for unknown genetic modes of SNPs of interest, unknown Hardy-Weinberg equilibrium (HWE) in a population, and a large number of SNPs of interest.

We first propose a novel multi-SNP test under the case-control study design, which we term the principal Chi-squared test. The principal Chi-squared test applies a two-degree-of-freedom score statistic based on the empirical Bayes method for each SNP and derives a global test based on the eigenvalue decomposition of the asymptotic variance-

covariance matrix of each SNP test. The global test achieves improved power by robustly exploiting the HWE in the control population and effectively exploiting the LD among all SNPs. We denote the global test by PChiB (see Methods). In addition to competitive power, PChiB is conveniently implemented and is easily comprehensible by the nonstatistics community because of the well-known eigenvalue decomposition method. The global test is closely related to standard PCR in that it reduces to the score test of PCR when each SNP is coded as the number of minor alleles. This relation not only lends insight into its power relative to PCR but also into the connection between PCR and variance-component-based tests. We show that both classes of these methods are weighted combinations of uncorrelated Chi-squared random variables, each of which is a weighted combination of a single SNP test with weights equal to the loadings of the eigenvectors of their joint asymptotic variance-covariance matrix. This observation, while supporting documented conclusions that none of the two classes of methods is uniformly more powerful than the other [22], reveals theoretically that the LD structure among SNPs plays a critical role in the powers of these methods. When a real disease causal SNP adopts recessive and dominate codes, test PChiB can gain desirable power. When a real disease causal SNP adopts an additive code, test PChiB may have somewhat lower power. Thus, we propose a robust test by taking the minimum p value of the new global test PChiB and the ordinary prospective score test of PCR in which each SNP is coded as the number of minor alleles, regardless of the actual genetic code of each SNP. We denote the robust test by Min2.

Suppose that q diallelic SNPs in a genomic region of interest are genotyped for n_1 case samples and n_0 control samples. Let Y_i denote the binary case-control status ($Y_i = 1$: case; $Y_i = 0$: control) for sample i $(i = 1, 2, \cdots, n)$, where $n = n_1 + n_0$, the first n_1 samples are cases, and the remaining n_0 samples are controls. Denote G_{ik} as the count of the minor alleles of SNP k from sample i for $i = 1, 2, \cdots, n$, and $k = 1, 2, \cdots q$. A new global test is designed to test the null hypothesis that the genomic region spanned by q SNPs is not associated with the phenotype status of interest against the general alternative that one or more SNPs, which may or may not be genotyped, are associated with the phenotype status of interest. We fit an ordinary logistic regression model for the binary case-control status and all SNPs.

Incorporating HWE constraints into the control population based on the retrospective likelihood for testing a diallelic marker may lead to increased power under dominant and recessive genetic models compared to standard prospective likelihood-based tests [24]. To address the issue that deviation from HWE may lead to an inflated type I error rate in this test, an empirical Bayes score test, which is a data-adaptive linear combination of the prospective likelihood score test and retrospective likelihood score test under the HWE constraint, was proposed [25]. This test can maintain nominal type I error rates under deviations from HWE that are observed in real settings and largely maintains the power gain under the recessive genetic model. Here, our new global statistic uses this test principal as the building block. We expect that our method achieves considerably improved power when aggregating the small power gains at each SNP.

The rest of this paper is organized as follows. In Results, we demonstrate, through simulation studies and analysis of pancreatic cancer data [26, 27], that the proposed robust test can often have desirable power compared to some popular tests across a broad range of scenarios. In Discussion, we further discuss the merits and disadvantages of our proposed test method and note some directions for future research. In Methods, we present the new global test in detail and discuss its connections to PCR and other existing methods. We also briefly introduce the robust test by taking the minimum p value of the new global test and the score test of PCR, where each single SNP is coded as the number of minor alleles, regardless of the actual genetic code of each SNP.

2. Results

2.1. A Robust Statistical Method Based on Two Types of Principal Chi-Squared Tests. For real genotype data, we can first calculate the prospective score test, denoted by $\tilde{U}_P = (\tilde{U}_{P,1},\cdots,\tilde{U}_{P,q})$, in which all SNPs are supposed to adopt additive codes. We denote a consistently estimated covariance $\tilde{V}_P$ for $\tilde{U}_P$ and calculate the ordinary principal components regression (PCR) score statistic, which is denoted by *PChiP* (selecting the top PCs explaining 85% of genetic variability) based on the estimated covariance $\tilde{V}_P$, as in Gauderman et al. [19]. Second, we can obtain the p value of PChiP, which is denoted by $PV_{A,P}$ because PChiP follows a Chi-squared distribution asymptotically under the null hypothesis. Third, we calculate the empirical Bayes score denoted by $U_B = (U_{B,1},\cdots,U_{B,q})$ and its consistently estimated covariance, which is denoted by V_B, based on codominant codes (see Methods). Similarly, we calculate the new aforementioned principal Chi-squared statistic PChiB based on the estimated covariance V_B. Note the dimension of U_B is $2q$, and we can estimate the p value of PChiB, which is denoted by $PV_{C,B}$, because PChiB also follows a Chi-squared distribution asymptotically under the null hypothesis. Finally, we take the minimum of the two p values of PChiP and PChiB as a robust test, as follows:

$$\text{Min } 2 = \min\left(PV_{A,P}, PV_{C,B}\right). \quad (1)$$

We estimate the p value of Min2 via statistical permutation. We conduct extensive simulations to investigate the power performance of Min2.

To view the performance of Min2 comprehensively, we can compare it to 4 other tests, namely, PChiB, PChiP, SSUP (see Methods) and GOLD, where GOLD is constructed as follows. Suppose the first SNP is the real causal SNP satisfying the logistic regression model $\text{logitPr}(Y = 1) = \beta_0 + \beta_1 * G_1$, where β_0 and β_1 represent log odds ratios. Other SNPs are correlated with the first SNP with genotypes $G_2, \cdots, G_q$. The Gold method (denoted by GOLD) is an ordinary score test based on the above real statistical model. Clearly, in real data analysis scenarios, we do not know the causal SNP. GOLD only has a value in simulation studies and is not

Table 1: Size and haplotypes with frequencies for gene NAT2.

Haplotype	Frequency
443423442114244211	0.279
214242244112422433	0.246
413443444332224231	0.211
214242224112422433	0.092
214243444332222431	0.042
413243444112422233	0.025
413443444332244231	0.018
443423444332224231	0.017
214242244112224233	0.017
413423442134244211	0.011
244242244112422433	0.008
413243224112422433	0.008
413443442332422433	0.008
214242224132422433	0.008
413423422134244211	0.006
214242244132422433	0.002

practical in real data analysis. We consider 3 scenarios for analysing genotype data. First, we apply PChiB, Min2, PChiP, SSUP, and GOLD to analyse genotype data, including all SNPs. Second, we apply PChiB, Min2, PChiP, SSUP, and GOLD to analyse genotype data, excluding the first SNP, which is the causal SNP. Third, we apply PChiB, Min2, PChiP, SSUP and GOLD to analyse genotype data, including only labelled SNPs. To comprehensively assess the performance of these 5 methods, we designate every SNP among all q SNPs as the causal SNP in turn in the simulation procedure.

2.2. Simulation Procedure. We conduct extensive simulation studies to assess the relative power of Min2 by comparing its performance with that of 4 other test statistics, namely, PChiB, PChiP, SSUP, and GOLD. We consider real LD structures defined by haplotypes inferred from the International Hapmap Project CEU samples. We set the haplotype information for gene NAT2 studied by Kwee et al. [28] as the basis of our simulations. To generate multilocus genotype data based on real haplotypes, we estimated haplotypes and their frequencies in a genomic region via HaploView software [29]. The LD structures plot based on the complete set of SNPs for gene NAT2 is displayed in Supplementary Figure 1 (See Supplementary File). For gene NAT2, we select SNPs with MAFs > 0.05 and genotype rates ≥ 75%, for a total of 18 SNPs. Haplotypes based on the complete set of SNPs and their frequencies are provided in Table 1. Five SNPs, rs13277605, rs1799930, rs1208, rs1961456, and rs2410556, are tag SNPs.

To obtain the n_0 control samples, we generated multilocus genotype data, as follows. Let $\{f_H\}$ denote the set of estimated haplotype frequencies with $\sum_H f_H = 1$. Then, a pair of haplotypes for each control sample was generated under HWE, where the frequency of haplotype pairs (H, H') takes the form $\phi_{HH'} = f_H^2$ as $H = H'$ and $\phi_{HH'} = f_H f_{H'}$ as $H \neq H'$. The haplotype phase information was then deleted, and only locus-specific genotype data were retained. To generate multilocus genotype data for each case sample (total number n_1), we generated the pair of haplotypes (H, H') using the following probabilities:

$$\phi^1_{HH'} = \frac{R_{\mathrm{Aa}}^{I(HH' \text{includes} \, ''\mathrm{Aa}'')} R_{\mathrm{aa}}^{I(HH' \text{includes} \, ''\mathrm{aa}'')} \phi_{HH'}}{\sum_{H,H'} R_{\mathrm{Aa}}^{I(HH' \text{includes} \, ''\mathrm{Aa}'')} R_{\mathrm{aa}}^{I(HH' \text{includes} \, ''\mathrm{aa}'')} \phi_{HH'}}, \quad (2)$$

where R_{Aa} and R_{aa} are the odds ratios for genotypes "Aa" and "aa", 'A' is the major allele for the disease causal SNP, 'a' is the minor allele of the disease-causal SNP, and indicator functions $I(HH' \text{includes} \, ''Aa'')$ and $I(HH' \text{includes} \, ''aa'')$ refer to whether haplotype pair (H, H') has allele combinations (A,a) and (a, a), respectively, at the causal SNP.

To evaluate the impact of deviation from HWE on the power of PChiB, we additionally generated multilocus genotype data from real haplotypes based on gene NAT2, as described above, but with the frequency of haplotype pairs (H, H') equal to $\phi_{HH'} = (1 - F_{st}) f_H f_{H'} + \delta_{HH'} F_{st} f_H$. Here, $\delta_{HH'}$ is an indicator function, with $\delta_{HH'} = 1$ if $H = H'$ and $\delta_{HH'} = 0$ if $H \neq H'$, and F_{st} is the fixation parameter, which represents mild deviation from HWE, as observed in real gene association analysis studies.

We set $n_1 = 1000$ and $n_0 = 1000$ and consider two scenarios with HWE indicator $F_{st} = 0$ and $0.5 \log(2)$, as in Luo et al. [25]. Furthermore, we designate every SNP as the causal SNP in turn. When the causal SNP adopts an additive code, we obtain the genotype and case-control status based on the logistic model with causal SNP odds ratio 1 for estimating the empirical type I error rates and with causal SNP odds ratio 1.2 for estimating the empirical power. When the causal marker adopts a dominant code, we obtain the genotype and case-control status based on the logistic model with causal SNP odds ratio 1.3 for estimating the empirical power. When the causal marker adopts a recessive code, we obtain the genotype and case-control status based on the logistic model with causal SNP odds ratio 1.5 for estimating the empirical power. With the genotype and case-control status information, we calculate the p value of Min2 via 200 permutations. The empirical type I error rates and powers of the 4 tests were considered under a significance level of 0.05 by means of 500 repetitions, as Kwee et al. [28] examined the type I error and power of the semiparametric and single-tag SNP approaches assuming a nominal significance level of 0.05.

2.3. Numerical Results. To comprehensively assess the performance of Min2, we construct test statistics under 3 scenarios, namely, using all SNPs, using all SNPs except the causal SNP, and using only tag SNPs.

Because the empirical type I error rates are nearly the same when the real causal SNP adopts an additive code, dominant code, and recessive code, we present the empirical type I error rates for only the case where the real causal SNP adopts an additive code. The results based on all 18 SNPs with $F_{st} = 0$ are displayed in Figure 1, and the results based on all 18 SNPs with $F_{st} = 0.05 log(2)$ are displayed in

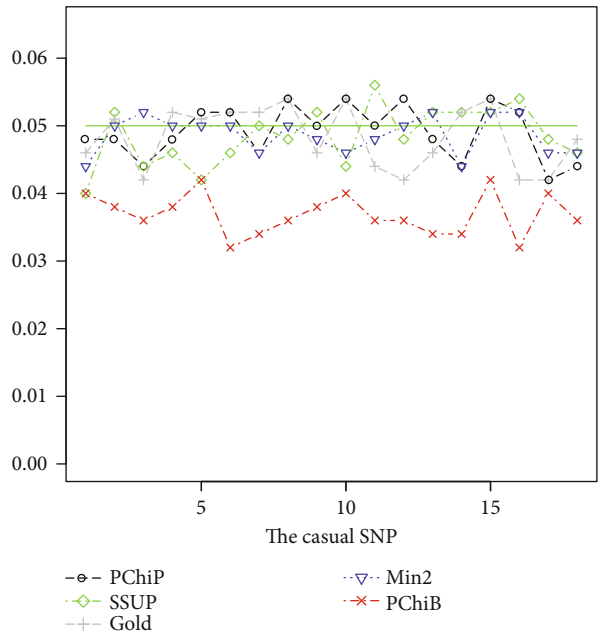

Figure 1: Empirical null hypothesis rejection rates (based on all 18 SNPs) of GOLD, PChiP, SSUP, PChiB, and Min2. Each SNP is treated as the causal locus in turn, which has an additive effect, with simulated odds ratio 1.0 and $F_{st} = 0$ based on 1000 controls, 1000 cases and 500 iterations.

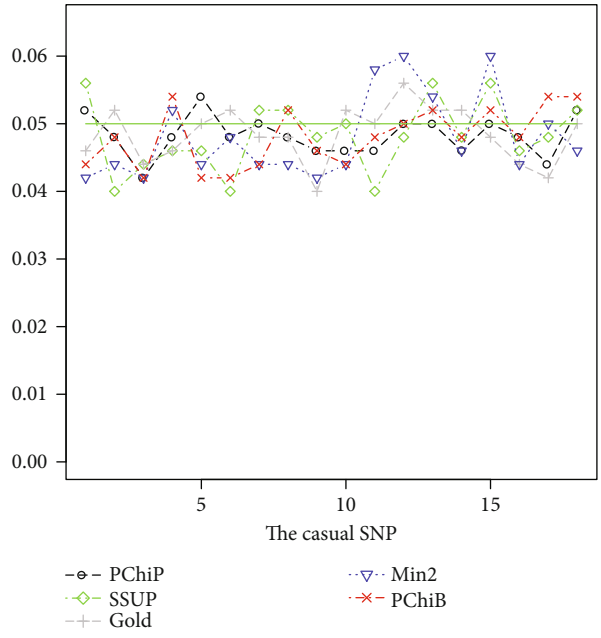

Figure 2: Empirical null hypothesis rejection rates (based on all 18 SNPs) of GOLD, PChiP, SSUP, PChiB, and Min2. Each SNP is treated as the causal locus in turn, which has an additive effect, with simulated odds ratio 1.0 and $F_{st} = 0.5 \log(2.0)$ based on 1000 controls, 1000 cases, and 500 iterations.

Figure 2. Other results based on all 17 SNPs (excluding the causal SNP) and 5 tag SNPs are displayed in Supplementary Figure 2a, Figure 2b, Figure 3a, and Figure 3b (See Supplementary File). From Figures 1 and 2, we can see that

Table 2: Empirical powers (based on all 18 SNPs) of GOLD, PChiP, SSUP, PChiB, and Min2. Each SNP is treated as the causal locus in turn, which has a recessive effect, with simulated odds ratio 1.5 and F_{st}t = 0 based on 1000 controls, 1000 cases, and 500 iterations.

Causal SNP no.	PChiP	SSUP	GOLD	Min2	PChiB
1	0.672	0.738	0.948	0.764	0.764
2	0.364	0.352	0.826	0.504	0.492
3	0.678	0.768	0.954	0.784	0.796
4	0.748	0.826	0.972	0.846	0.842
5	0.428	0.394	0.816	0.546	0.534
6	0.642	0.704	0.926	0.726	0.732
7	0.588	0.638	0.932	0.736	0.73
8	0.048	0.042	0.186	0.054	0.024
9	0.42	0.366	0.81	0.506	0.524
10	0.348	0.168	0.778	0.372	0.286
11	0.398	0.186	0.844	0.378	0.2
12	0.434	0.4	0.812	0.554	0.542
13	0.586	0.642	0.938	0.684	0.708
14	0.428	0.426	0.836	0.54	0.522
15	0.73	0.818	0.978	0.822	0.826
16	0.678	0.746	0.972	0.78	0.808
17	0.34	0.328	0.778	0.51	0.518
18	0.71	0.768	0.954	0.802	0.794

Min2 can control the type I error rate well when the HWE indicator coefficient F_{st} equals 0 or 0.5log(2.0), but PChiB has a conservative empirical type I error rate when F_{st} equals 0. We further investigate this phenomenon: when the real genetic model adopts additive code, PChiB adopts a codominant code with F_{st} equal to 0, so the correlations between every two SNPs are decreased and test PChiB may absorb a large number of degrees of freedom. For example, when considering the scenario with all 18 SNPs and designating the 1st SNP as the causal SNP, PChiP absorbs 2 degrees of freedom and PChiB absorbs 5 degrees of freedom, according to the simulation data. When the real genetic model adopts recessive and dominant codes, all 5 tests control the type I error rate well, regardless of whether F_{st} is 0 or 0.5log(2.0).

For the empirical power comparison, when the real causal SNP adopts a recessive code, we display the results based on all 18 SNPs in Tables 2 and 3 for $F_{st} = 0$ and $F_{st} = 0.5 \log(2)$. Other results based on 17 SNPs (excluding the causal SNP) and 5 tag SNPs are displayed in Supplementary Figure 4a, 4a, Figure 5a, and Figure 5b (See Supplementary File). From Table 2, Supplementary Figure 4a and Supplementary Figure 4b for $F_{st} = 0$, we can see that the GOLD test always performs best because it is an oracle test, and Min2 performs nearly as good as PChiB in all 3 scenarios. Additionally, Min2 always performs better than PChiP and SSUP, regardless of which of the 18 SNPs is the causal SNP. For example, in Table 2, the empirical powers of PChiP, SSUP, GOLD, Min2, and PChiB are 0.364, 0.352, 0.826, 0.504, and 0.492, respectively, when the 2nd SNP is the causal SNP. From Table 3, Supplementary Figure 5a,

Table 3: Empirical powers (based on all 18 SNPs) of GOLD, PChiP, SSUP, PChiB, and Min2. Each SNP is treated as the causal locus in turn, which has a recessive effect, with simulated odd ratios 1.5 and $F_{st} = 0.5 \log(2.0)$ based on 1000 controls, 1000 cases, and 500 iterations.

Causal SNP no.	PChiP	SSUP	GOLD	Min2	PChiB
1	0.755	0.795	0.97	0.84	0.875
2	0.45	0.46	0.865	0.51	0.605
3	0.765	0.835	0.965	0.855	0.885
4	0.835	0.925	0.99	0.84	0.88
5	0.555	0.58	0.85	0.605	0.67
6	0.69	0.77	0.935	0.785	0.765
7	0.65	0.715	0.965	0.74	0.79
8	0.06	0.085	0.37	0.07	0.1
9	0.515	0.48	0.905	0.65	0.755
10	0.535	0.28	0.825	0.6	0.59
11	0.58	0.33	0.88	0.625	0.665
12	0.48	0.475	0.83	0.62	0.665
13	0.695	0.765	0.94	0.735	0.79
14	0.58	0.595	0.895	0.655	0.7
15	0.79	0.875	0.98	0.84	0.88
16	0.725	0.805	0.97	0.805	0.865
17	0.52	0.495	0.83	0.61	0.65
18	0.785	0.825	0.955	0.86	0.875

and Supplementary Figure 5b for $F_{st} = 0.5\log(2)$, we can see that Min2, when using all 18 SNPs, using all 18 SNPs except for the causal SNP, and using only tag SNPs, always performs much better than PChiP and SSUP, regardless of which of the 18 SNPs is the causal SNP. For example, in Table 3, the empirical powers of PChiP, SSUP, GOLD, and Min2 are 0.755, 0.795, 0.970, 0.840, and 0.875, respectively, when the 1st SNP is the causal SNP.

When the real causal SNP adopts a dominant code, we display all the results based on all 18 SNPs in Tables 4 and 5 for $F_{st} = 0$ and $F_{st} = 0.5 \log(2)$. Other results based on 17 SNPs (excluding the causal SNP) and 5 tag SNPs are displayed in Supplementary Figure 6a, Figure 6b, Figure 7a, and Figure 7b (See Supplementary File). From these figures, we can see that Min2 performs robustly among all 5 tests over all 3 scenarios with $F_{st} = 0$ and 0.5log(2). For example, in Table 4, the empirical powers of PChiP, SSUP, GOLD, Min2, and PChiB are 0.598, 0.588, 0.846, 0.636, and 0.556, respectively, when the 9th SNP is the causal SNP, and the empirical powers of PChiP, SSUP, GOLD, Min2, and PChiB are 0.638, 0.382, 0.826, 0.628, and 0.496, respectively, when the 10th SNP is the causal SNP. In Table 5 for $F_{st} = 0.05 \log(2)$, the empirical powers of PChiP, SSUP, GOLD, Min2, and PChiB are 0.585, 0.310, 0.786, 0.545, and 0.455, respectively, when the 11th SNP is the causal SNP.

When the real causal SNP adopts an additive code, we display all results based on all 18 SNPs in Tables 6 and 7 for $F_{st} = 0$ and $F_{st} = 0.5 \log(2)$. Other results based on 17 SNPs (excluding the causal SNP) and 5 tag SNPs are displayed in Supplementary Figure 8a, Figure 8b, Figure 9a,

Table 4: Empirical powers (based on all 18 SNPs) of GOLD, PChiP, SSUP, PChiB, and Min2. Each SNP is treated as the causal locus in turn, which has a dominant effect, with simulated odds ratio 1.3 and $F_{st} = 0$ based on 1000 controls, 1000 cases, and 500 iterations.

Causal SNP no.	PChiP	SSUP	GOLD	Min2	PChiB
1	0.51	0.56	0.76	0.532	0.456
2	0.57	0.552	0.822	0.564	0.49
3	0.486	0.576	0.79	0.532	0.438
4	0.448	0.532	0.74	0.476	0.416
5	0.644	0.626	0.824	0.628	0.518
6	0.556	0.61	0.808	0.576	0.516
7	0.568	0.63	0.79	0.596	0.504
8	0.13	0.152	0.712	0.128	0.078
9	0.598	0.588	0.846	0.636	0.556
10	0.638	0.382	0.826	0.628	0.496
11	0.574	0.338	0.818	0.586	0.51
12	0.614	0.614	0.836	0.622	0.56
13	0.506	0.58	0.79	0.548	0.502
14	0.584	0.578	0.836	0.576	0.51
15	0.458	0.518	0.756	0.482	0.388
16	0.462	0.538	0.808	0.478	0.418
17	0.694	0.662	0.822	0.676	0.598
18	0.492	0.55	0.76	0.51	0.448

Table 5: Empirical powers (based on all 18 SNPs) of GOLD, PChiP, SSUP, PChiB, and Min2. Each SNP is treated as the causal locus in turn, which has a dominant effect, with simulated odd ratio 1.3 and $F_{st} = 0.5 \log(2.0)$ based on 1000 controls, 1000 cases, and 500 iterations.

Causal SNP no.	PChiP	SSUP	GOLD	Min2	PChiB
1	0.56	0.645	0.774	0.47	0.45
2	0.55	0.505	0.816	0.475	0.455
3	0.5	0.55	0.778	0.44	0.445
4	0.455	0.5	0.75	0.4	0.43
5	0.615	0.645	0.834	0.54	0.51
6	0.62	0.68	0.816	0.565	0.61
7	0.56	0.61	0.812	0.46	0.515
8	0.15	0.19	0.712	0.12	0.145
9	0.58	0.56	0.834	0.53	0.515
10	0.67	0.435	0.822	0.6	0.56
11	0.585	0.31	0.786	0.545	0.455
12	0.61	0.6	0.852	0.57	0.555
13	0.485	0.575	0.812	0.445	0.455
14	0.68	0.645	0.804	0.56	0.53
15	0.505	0.545	0.776	0.415	0.43
16	0.455	0.55	0.79	0.43	0.44
17	0.66	0.64	0.826	0.61	0.6
18	0.51	0.55	0.76	0.475	0.46

and Figure 9b (See Supplementary File). From these figures, we can see that Min2 performs robustly among all 5 tests over all 3 scenarios for $F_{st} = 0$ and 0.5log(2). Under these 3

Table 6: Empirical powers (based on all 18 SNPs) of GOLD, PChiP, SSUP, PChiB, and Min2. Each SNP is treated as the causal locus in turn, which has an additive effect, with simulated odds ratio 1.2 and $F_{st} = 0$ based on 1000 controls, 1000 cases, and 500 iterations.

Causal SNP no.	PChiP	SSUP	GOLD	Min2	PChiB
1	0.694	0.748	0.798	0.644	0.466
2	0.614	0.584	0.784	0.594	0.404
3	0.654	0.728	0.818	0.614	0.424
4	0.706	0.78	0.8	0.678	0.482
5	0.666	0.656	0.798	0.626	0.428
6	0.654	0.736	0.796	0.618	0.462
7	0.724	0.77	0.814	0.702	0.484
8	0.102	0.114	0.504	0.09	0.052
9	0.632	0.63	0.76	0.594	0.408
10	0.644	0.36	0.77	0.614	0.352
11	0.626	0.402	0.77	0.616	0.4
12	0.674	0.652	0.774	0.606	0.432
13	0.708	0.768	0.802	0.682	0.468
14	0.644	0.618	0.804	0.604	0.422
15	0.678	0.782	0.816	0.656	0.484
16	0.632	0.71	0.794	0.612	0.428
17	0.696	0.662	0.754	0.652	0.444
18	0.688	0.756	0.798	0.652	0.454

Table 7: Empirical powers (based on all 18 SNPs) of GOLD, PChiP, SSUP, PChiB, and Min2. Each SNP is treated as the causal locus in turn, which has an additive effect, with simulated odds ratio 1.2 and $F_{st} = 0.5\log(2.0)$ based on 1000 controls, 1000 cases, and 500 iterations.

Causal SNP no.	PChiP	SSUP	GOLD	Min2	PChiB
1	0.76	0.805	0.855	0.705	0.625
2	0.66	0.6	0.795	0.585	0.55
3	0.745	0.78	0.825	0.725	0.655
4	0.765	0.84	0.86	0.735	0.695
5	0.755	0.72	0.825	0.69	0.57
6	0.77	0.795	0.825	0.685	0.61
7	0.725	0.765	0.84	0.665	0.625
8	0.155	0.19	0.585	0.145	0.14
9	0.66	0.67	0.8	0.645	0.57
10	0.725	0.53	0.82	0.69	0.565
11	0.665	0.435	0.835	0.63	0.51
12	0.73	0.695	0.82	0.62	0.59
13	0.695	0.74	0.85	0.675	0.6
14	0.7	0.67	0.81	0.655	0.55
15	0.66	0.725	0.82	0.635	0.545
16	0.635	0.725	0.82	0.58	0.535
17	0.72	0.705	0.81	0.705	0.57
18	0.72	0.76	0.845	0.695	0.62

scenarios, the real genetic codes are additive, so it is not unexpected that the performance of PChiP is always a little better than that of Min2, regardless of which of the 18 SNPs is the causal SNP. Although SSUP sometimes has slightly better power than PChiP and Min2, it can sometimes have very low power. For example, in Table 6, the empirical powers of PChiP, SSUP, GOLD, Min2, and PChiB are 0.626, 0.402, 0.770, 0.616, and 0.400 when the 11th SNP is the causal SNP. In Table 7, for $F_{st} = 0.5\log(2)$, the empirical powers of PChiP, SSUP, GOLD, Min2, and PChiB are 0.660, 0.670, 0.800, 0.645, and 0.570, respectively, when the 9th SNP is the causal SNP.

2.4. The Analysis of High-Density Lipoprotein Cholesterol (HDL-C) Data from GWAS Pancreatic Cancer Data. Herein, we present an analysis of HDL-C data from GWAS pancreatic cancer data [26, 27] to illustrate our method. Plasma levels of high-density lipoprotein cholesterol are known to be heritable, but only a fraction of the heritability is explained. We developed a high-density genotyping array populated with HDL-C candidate loci selected based on the known biology of HDL metabolism, mouse genetic studies, human genetic association studies, and available GWAS data. SNP selection was based on tag SNPs but also included low-frequency nonsynonymous SNPs. We performed association analysis on the majority of reported GWAS loci (including ABCA1, CETP, GALNT2, LCAT, LIPG, LIPC, and LPL).

The data set consists of 1231 samples (case: 625 and control: 606) with 64 SNPs from the above 13 genes. Basic information about the 13 genes is presented in Supplementary Table 1 (additional file 2). We calculate the p values of 4 test methods, i.e., PChiP, SSUP, Min2, and PChiB, when analysing the data set. The numerical results are displayed in Table 8. From Table 8, we can see that the numerical results of Min2 are consistent with those of the other tests. For example, when investigating the association between HDL-C and gene GALNT2, including 2 SNPs, the p values of PChiP, SSUP, Min2, and PChiB are 0.1065, 0.1065, 0.0370, and 0.0272, respectively. For another example, when investigating the association between HDL-C and gene LPL, including 15 SNPs, the p values of PChiP, SSUP, Min2, and PChiB are 0.002, 0.00016, 0.002, and 0.0044, respectively. For the third example, when investigating the association between HDL-C and gene LIPG, including 2 SNPs, the p values of PChiP, SSUP, Min2, and PChiB are 0.0012, 0.0012, 0.0001, and 0.0002.

Table 8: p values of tests PChiP, SSUP, Min2, and PChiB when analysing 7 genes.

Gene	SNP nos.	PChiP	SSUP	Min2	PChiB
GALNT2	2	0.1065	0.1065	0.0370	0.0272
LPL	15	0.0020	0.00016	0.0020	0.0044
ABCA1	3	0.0311	0.0121	0.040	0.0782
LIPC	9	0.0069	0.0019	0.0050	0.0669
CETP	25	6.051e-13	3.278e-13	7.615e-14	1.114e-16
LCAT	2	0.9981	0.9999	0.9700	0.9297
LIPG	2	0.0012	0.0012	0.0001	0.0002

Because the number of SNPs in each gene is not very large in the real data, the real data do not provide a good

example to illustrate the merit our test. However, this limitation does not affect our purpose of deriving a robust test. Our method focuses on the robustness in the following 3 scenarios: the genetic code for all SNPs is unknown, whether the HWE is satisfied in the original population is unknown, and a large number of SNPs exists.

3. Discussion

One key factor of the improved power of kernel-machine-based tests [17] and PCR is the reduced degrees of freedom. Kernel-machine-based tests make full use of possible correlations among score statistics, which is known to be advantageous for high-dimensional data [30], and are robust to the directions of association of different SNPs. Principal component analysis is a standard method of reducing the dimensionality of a large number of variables. Despite this seemingly obvious argument, the relative merits of PCR and kernel-machine-based tests remain understudied. We provide insights into the theoretical connection between kernel-machine-based tests and the PCR method. We find that when the LD extent of each pair of SNPs is somewhat strong, principal component analysis methods may have higher power than kernel-machine-based tests. PCR often has similar or higher power than kernel-machine-based tests, where the LD pattern is an important parameter for power. We will further explore the principle of selecting the number of PCs in future work.

In this work, we consider an association test between human complex diseases and genetic SNPs based on principal component analysis (PCA) since PCA is widely used in the recent literature. PCA accounts for linear combinations among SNPs. If this linearity exists, PCA is optimal. However, when how the multiple genetic SNPs influence the risk of disease is unknown, one alternative strategy is to use haplotype analysis since haplotypes can capture the LD information between markers [31–37].

We propose a novel global test (PChiB) based on the empirical Bayes score test, which is a data-adaptive linear combination of the prospective likelihood score and the retrospective likelihood score under the HWE constraint in the control population. PChiB can maintain desirable power when the real causal SNP adopts recessive and dominant codes under the HWE constraint in the control population. A small disadvantage of PChiB is that when the genetic code of the real causal SNP is additive, PChiB does not have desirable power because of the large degrees of freedom. Thus, we propose a robust test (Min2) that maintains the power gain under deviations from HWE observed in real settings, regardless of which genetic code the real causal SNP adopts. Min2 gains power by effectively using the LD among all the tested SNPs over all scenarios. Because PChiP is based on the assumption that all SNPs adopt an additive code, while PChiB and Min2 are based on the assumption that all SNPs adopt a codominant code, PChiP has low degrees of freedom and performs best when the causal SNP adopts an additive code. PChiB and Min2 may have less power than PchiP in this scenario. When the causal SNP adopts dominant or recessive codes, Min2 has desirable power, regardless of whether HWE is satisfied in the control population. We propose to use our new test Min2 for the association analysis of multilocus genotypes and complex diseases.

We propose the robust test Min2, where the p values are obtained via permutation and compared it with PChiB (empirical score based on all SNPs adopting codominant codes), PChiP (prospective score based on all SNPs adopting additive codes), and SSUP (a VC method based on the prospective score and all SNPs adopting an additive code). The main purpose of this article is to introduce the proposed test Min2, not to compare it with other existing tests for GWAS.

Notably, it would be a good idea to extend the proposed tests to include covariate adjustments in the logistic models. The derivation will be very complex and requires additional research. We will consider this problem in our future work. In simulations, we need to set a large sample size n as the number of MAF is low, so we have not considered rare variants. We may investigate the robustness about PChiB when the number of MAF is low in our further work.

4. Methods

4.1. A New Principal Chi-Squared Test. Suppose there are n_1 case samples and n_0 control samples and denote $n = n_1 + n_0$. For the ith ($i = 1, \cdots, n$) sample and kth ($k = 1, \cdots, q$) SNP, denote G_{ik} as the additive code, namely, the numbers of minor alleles taking values 0, 1, and 2. For the ith ($i = 1, \cdots, n$) sample and kth ($k = 1, \cdots, q$) SNP, denote $m(G_{ik})$ as the codominant code, namely, $m(G_{ik}) = (m_1(G_{ik}), m_2(G_{ik})) = (I[G_{ik} = 1], I[G_{ik} = 2])$, where $I[\cdot]$ is an indicator function. Clearly, $m(0) = (0, 0)$, $m(1) = (1, 0)$, and $m(2) = (0, 1)$.

For $k = 1, \cdots, q$, denote $\widehat{f}_k$ as the estimated minor allele frequency (MAF) for the kth SNP in the pooled case-control sample and denote g_k as the number of minor allele in a genotype for the kth SNP in a population with values 0, 1, and 2. For $k = 1, \cdots, q$, denote $P_{\widehat{f}_k}(g_k)$ as the estimated genotype frequency for the kth SNP. We can then obtain $\widehat{f}_k = \sum_{i=1}^{n}\{I[G_{ik} = 1] + 2I[G_{ik} = 2]\}/(2n)$, $P_{\widehat{f}_k}(g_k = 0) = (1 - \widehat{f}_k)^2$, $P_{\widehat{f}_k}(g_k = 1) = 2\widehat{f}_k(1 - \widehat{f}_k)$, and $P_{\widehat{f}_k}(g_k = 2) = (\widehat{f}_k)^2$. For $k = 1, \cdots, q$, denote a 2-dimensional row vector by $\tau_{(k)} = (\tau_{1k}, \tau_{2k}) = E_{\mathrm{HWE},\widehat{f}_k}[m(g_k)] - \bar{m}(g_k) = (E_{\mathrm{HWE},\widehat{f}_k}[m_1(g_k)] - \bar{m}_1(g_k), E_{\mathrm{HWE},\widehat{f}_k}[m_2(g_k)] - \bar{m}_2(g_k))$, where $E_{\mathrm{HWE},\widehat{f}_k}[m(g_k)] = \sum_{g_k=0,1,2} m(g_k)P_{\widehat{f}_k}(g_k) = (\sum_{g_k=0,1,2} m_1(g_k)P_{\widehat{f}_k}(g_k), \sum_{g_k=0,1,2} m_2(g_k)P_{\widehat{f}_k}(g_k))$ is the expected value of $m(g_k)$ under HWE, and $\bar{m}(g_k) = (\bar{m}_1(g_k), \bar{m}_2(g_k))$ is the pooled sample mean of $m(g_k) = (m_1(g_k), m_2(g_k))$, namely, $\bar{m}(g_k) = \sum_{i=1}^{n} m(G_{ik})/n = (\sum_{i=1}^{n} m_1(G_{ik})/n, \sum_{i=1}^{n} m_2(G_{ik})/n)$. For $k = 1, \cdots, q$, denote $s^2_{\bar{m}_1(g_k)}$ as the pooled sample variance of $m_1(g_k)$, namely, the variance of $m_1(G_{1k}), \cdots, m_1(G_{nk})$ and denote $s^2_{\bar{m}_2(g_k)}$ as the pooled sample variance of $m_2(g_k)$, namely, the variance of $m_2(G_{1k}), \cdots, m_2(G_{nk})$. For $k = 1, \cdots, q$, denote a diagonal matrix W_k with elements equal to $(s^2_{\bar{m}_1(g_k)}/n)/((s^2_{\bar{m}_1(g_k)})/n + \tau^2_{1k})$ and $(s^2_{\bar{m}_2(g_k)}/n)/((s^2_{\bar{m}_2(g_k)})/n + \tau^2_{2k})$. Clearly, W_k is extended from the weight proposed

by Luo et al. [25] and Chatterjee et al. [38] when an additive (dominant or recessive) code is adopted. The weight matrix W_k is data adaptive. When codominant coding is adopted, by means of W_k, we propose the empirical Bayes score for the kth $(k = 1, \cdots, q)$ SNP with the following form:

$$U_{B,k} = \sum_{i=1}^{n_1} \Big\{ m(G_{ik}) - \Big[E_{\mathrm{HWE},\hat{f}_k}[m(g_k)] W_k + \bar{m}(g_k)(I_{2\times 2} - W_k) \Big] \Big\}, \tag{3}$$

where $I_{2\times 2}$ is an identity matrix with dimension 2.

Let $U_{(B)}$ denote the vector of empirical Bayes scores for all q SNPs, namely, $U_{(B)} = (U_{B,1}, U_{B,2}, \cdots, U_{B,q})$, which is of length $2q$. Denote the estimated asymptotic covariance matrix by V_B (See Supplementary File) for empirical Bayes score vector $U_{(B)}$. A common test for whether all q markers can be jointly built, similar to the Hotelling T^2 statistic, is $T^2 = U_{(B)} V_B^{-1} U_{(B)}^T$, where '$T$' indicates the transpose of a vector or matrix. Our proposed new global statistic is based on the eigenvalue decomposition of covariance matrix V_B, as follows. For $k = 1, 2, \cdots, 2q$, denote λ_k and ξ_k (a $2q \times 1$ column vector) as the eigenvalue and corresponding eigenvector of covariance matrix V_B. Let $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_{2q})$ and $\xi = (\xi_1, \xi_2, \cdots, \xi_{2q})$ denote the eigenvalues and corresponding eigenvectors of covariance matrix V_B. We then have $\xi^T V_B \xi = \mathrm{diag}\,(\lambda_1, \lambda_2, \cdots, \lambda_{2q})$, and V_B can be written as $\xi\ \mathrm{diag}\,(\lambda_1, \lambda_2, \cdots, \lambda_{2q})\xi^T$. Since the norm of the eigenvector is unity and V_B^{-1} can be written as $\xi\ \mathrm{diag}\,(\lambda_1^{-1}, \lambda_2^{-1}, \cdots, \lambda_{2q}^{-1})\xi^T$, the test statistic T^2 can be written as

$$T^2 = \left(U_{(B)}\xi\right) \mathrm{diag}\left(\lambda_1^{-1}, \lambda_2^{-1}, \cdots, \lambda_{2q}^{-1}\right)\left(\xi^T U_{(B)}^T\right) = \sum_{k=1}^{2q} \left(U_{(B)}\xi_k\right)^2 / \lambda_k. \tag{4}$$

Note that $U_{(B)}\xi_k$, is a linear combination of the score for each individual SNP $U_{B,k}$ with $\mathrm{var}\,[U_{(B)}\xi_k] = \lambda_k$ for $k = 1, 2, \cdots, 2q$. We propose to utilize the first $s(1 \le s \le 2q)$ summands in T^2 to test the null hypothesis and denote the resultant test statistic as follows:

$$\mathrm{PChiB} = \sum_{k=1}^{s} \frac{\left(U_{(B)}\xi_k\right)^2}{\lambda_k}. \tag{5}$$

Due to the orthogonality of $\xi_1, \cdots, \xi_s$, $(U_{(B)}\xi_1)^2/\lambda_1, \cdots, (U_{(B)}\xi_s)^2/\lambda_s$ are independent. Because $(U_{(B)}\xi_1)^2/\lambda_1, \cdots, (U_{(B)}\xi_s)^2/\lambda_s$ are all asymptotically normally distributed with mean 0 and variance 1 under the null hypothesis that the genomic region spanned by the q SNPs is not associated with the phenotype status of interest, PChiB is asymptotically distributed as a Chi-squared variable with s degrees of freedom under the null hypothesis.

A remaining issue is how to select the number of summands s. Note that PChiB is based on eigenvalue decomposition, similar to the standard PCR. Many criteria for selecting s have been introduced in the literature [39]. It has been shown that using the top principal components that explain 80 ~ 90% of the genetic variability is sufficient [19, 20, 23]. We select s according to the same principal, i.e., that the top s principal components can explain approximately 85% of the genetic variability. This strategy is supported by the connection between PChiB and PCR (see the next subsection). In fact, the number of principal components affects the power of the principal component test [40]. When the LD extent of each pair of SNPs is very strong, the top one principal component alone has desirable power. When the LD extent of each pair of SNPs is somewhat strong, using the top principal components that explain 80 ~ 90% of the genetic variability is a robust method.

4.2. Understanding PChiB through an Exposition of PCR. We revisit PChiB based on only the standard prospective likelihood score under additive coding for kth $(k = 1, \cdots, q)$ and establish its equivalence to PCR [19, 20]. This equivalence sheds light on the promise of increased power of PChiB since PCR has been established to be a promising method for multi-SNP association analysis. In PCR, the phenotype variable is regressed on only a few of the top principal components (PCs) that summarize approximately 80-90% of the genetic variability. The PCs represent the directions in which most of the variability in the data occurs, as identified by the eigenvalue decomposition of the variance-covariance matrix of the centred raw genotype scores. Each principal component is a linear combination of genotype scores for all SNPs, and all principal components are uncorrelated with each other.

Here, we present the standard prospective likelihood score under additive genetic coding. The collection of all q prospective score functions, denoted by $\tilde{U}_{(P)} = (\tilde{U}_{P,1}, \tilde{U}_{P,2}, \cdots, \tilde{U}_{P,q})$, is asymptotically distributed as multivariate normal with mean $(0,\cdots,0)_{q\times 1}$ and variance-covariance matrix $\tilde{V}_P$ under the null hypothesis. Let $Y = (Y_1, \cdots, Y_n)^T$, $\bar{Y} = \sum_{i=1}^n Y_i/n$. For $k = 1, 2, \cdots, q$, let $\bar{G}_k = \sum_{i=1}^n G_{ik}/n$ and $\bar{G} = (\bar{G}_1, \cdots, \bar{G}_q)$. For $i = 1, 2, \cdots, n$, let $G_{(i)} = (G_{i1}, \cdots, G_{iq})^T$. Denote G as a genotype matrix with ith row and kth column element G_{ik} for $i = 1, \cdots, n$,and $k = 1, \cdots, q$. Let $\bar{1}$ be a column vector with all elements 1 and length n. In matrix form, $\tilde{U}_{(P)}^T = \sum_{i=1}^n (Y_i - \bar{Y})G_{(i)} = G^T(Y - \bar{Y}\bar{1})$, and its covariance matrix $\tilde{V}_P = \bar{Y}(1 - \bar{Y})\sum_{i=1}^n [G_{(i)} - \bar{G}][G_{(i)} - \bar{G}]^T$. Now, let $A = [a_1, a_2, \cdots, a_q]$ be a $q \times q$ matrix whose kth column is the characteristic vector of the matrix $\tilde{V}_P (k = 1, \cdots, q)$, and let $\tilde{\lambda}_1 \ge \tilde{\lambda}_2 \ge \cdots \ge \tilde{\lambda}_q$ be its eigenvalues. Denote orthogonal transformation $\tilde{G} = GA$. The likelihood score based on a logistic regression of Y on $\tilde{G}$ is $\tilde{U}_{(P)} = \tilde{G}^T(Y - \bar{Y}\bar{1})$. The covariance matrix of $\tilde{U}_{(P)}$ is a diagonal matrix with elements $\tilde{\lambda}_1, \tilde{\lambda}_2, \cdots, \tilde{\lambda}_q$.

Suppose that we consider the first $\tilde{s}(1 \le \tilde{s} \le q)$ PCs as follows. Let $A_{\tilde{s}} = [a_1, a_2, \cdots, a_{\tilde{s}}]$ be a $q \times \tilde{s}$ matrix containing

the first $\tilde{s}$ eigenvectors, and let $\tilde{G}_{\tilde{s}} = GA_{\tilde{s}}$. The standard PCA test based on the score statistic for testing the association between Y and $\tilde{G}_{\tilde{s}}$ from the logistic regression model is exactly equal to $(Y - \bar{Y}\bar{1})^T GA_{\tilde{s}}\, \mathrm{diag}\,(\tilde{\lambda}_1^{-1}, \tilde{\lambda}_2^{-1}, \cdots, \tilde{\lambda}_{\tilde{s}}^{-1}) A_{\tilde{s}}^T G^T (Y - \bar{Y}\bar{1})$, which is denoted by *PChiP*, and is the same as our proposed method when the adopted genetic code is additive code. Denote $\tilde{T}_k = (Y - \bar{Y}\bar{1})^T G a_k a_k^T G^T (Y - \bar{Y}\bar{1})/\lambda_k$, $k = 1, 2, \cdots, q$. When adopting additive code, the standard Hotelling T^2 statistic is equal to $\sum_{k=1}^{q} \tilde{T}_k$, and the PChiP statistic reduces to $\sum_{k=1}^{\tilde{s}} \tilde{T}_k$.

The proposed statistic (in this situation, equivalent to PCR) can be shown to be closely related to a statistic called the sum of squared score test based on prospective likelihood [12], which is denoted by SSUP. SSUP is obtained as SSUP= $\tilde{U}_{(P)} \tilde{U}_{(P)}^T = \sum_{j=1}^{q} \tilde{U}_{P,k}^2$, and it can be expressed as SSUP= $\sum_{j=1}^{q} \tilde{\lambda}_k [(\tilde{U}_{(P)} a_k^T)^2 / \tilde{\lambda}_k]$. Therefore, SSUP and PChiB use different weights for the contributions of the PCs: SSUP weights all PCs by the eigenvalues, whereas PChiB assigns equal weights to the top PCs. SSUP allows PCs with small eigenvalues to make additional contributions to the test, but PChiB discards PCs with small eigenvalues to reduce the degrees of freedom. This difference has implications on their relative power, which depends critically on the structure of variance-covariance matrix and, therefore, the LD structure of the assessed genomic region.

Data Availability

Data available on request.

Conflicts of Interest

The authors declare no competing interests.

Authors' Contributions

Jiayan Zhu and Yi Tian designed the methods, wrote the main manuscript text, and conducted some of the simulations. Ma Li and Xiaohong Cai conducted some of the simulations. Jiayan Zhu contributed to the interpretation of all results. All authors reviewed the manuscript.

Acknowledgments

We thank Prof. Jinbo Chen from the Department of Biostatistics of the University of Pennsylvania and Tonja Nansel and Prof. Kai Yu from the National Institutes of Health in the USA for providing the genetic data to demonstrate the methods and for providing some meaningful comments to improve the manuscript. We thank Prof. Qizhai Li from the China Science Academy for the discussion of the manuscript. Research of Yi Tian is partially supported by the self-determined research funds of Central China Normal University(CCNU) from the colleges basic research of MOE (No. CCNU19TD009) and National Nature Science Foundation of China (No. 61877023). Research of Jiayan Zhu is partially supported by seeding project funding (No. 2019ZZX026), scientific research project funding of talent recruitment, and start up funding for scientific research of Hubei University of Chinese Medicine.

Supplementary Materials

Derivation of the asymptotic variance estimate for EB score test UB. 2. Supplementary tables and figures for Results. *(Supplementary Materials)*

References

[1] W. Chen, X. Chen, K. J. Archer et al., "A rapid association test procedure robust under different genetic models accounting for population stratification," *Human Heredity*, vol. 75, no. 1, pp. 23–33, 2013.

[2] Q. Yang, J. Zhu, and Z. Li, "Maximin efficiency robust test for multiple nuisance parameters and its statistical properties," *Acta Mathematica Scientia*, vol. 37, no. 1, pp. 223–234, 2017.

[3] K. Yu, Q. Li, A. W. Bergen et al., "Pathway analysis by adaptive combination of *P*-values," *Genetic Epidemiology*, vol. 33, no. 8, pp. 700–709, 2009.

[4] H. Zhang, J. Shi, F. Liang, W. Wheeler, R. Stolzenberg-Solomon, and K. Yu, "A fast multilocus test with adaptive SNP selection for large-scale genetic- association studies," *European Journal of Human Genetics*, vol. 22, no. 5, pp. 696–702, 2014.

[5] S. Zhang, J. Zhu, and Z. Li, "Adaptive group-combined P-values test for two-sample location problem with applications to microarray Data," *Scientific Reports*, vol. 8, no. 1, pp. 8117–8119, 2018.

[6] J. M. Chapman, J. D. Cooper, J. A. Todd, and D. G. Clayton, "Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power," *Human Heredity*, vol. 56, no. 1-3, pp. 18–31, 2003.

[7] R. Fan and M. Knapp, "Genome association studies of complex diseases by case-control designs," *American Journal of Human Genetics*, vol. 72, no. 4, pp. 850–868, 2003.

[8] M. Xiong, J. Zhao, and E. Boerwinkle, "Generalized T2 Test for Genome Association Studies," *The American Journal of Human Genetics*, vol. 70, no. 5, pp. 1257–1268, 2002.

[9] B. Li and S. M. Leal, "Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data," *American Journal of Human Genetics*, vol. 83, no. 3, pp. 311–321, 2008.

[10] Z. Z. Tang and D. Y. Lin, "MASS: meta-analysis of score statistics for sequencing studies," *Bioinformatics*, vol. 29, no. 14, pp. 1803–1805, 2013.

[11] J. J. Goeman, S. A. van de Geer, and H. C. van Houwelingen, "Testing against a high dimensional alternative," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 3, pp. 477–493, 2006.

[12] W. Pan, "Asymptotic tests of association with multiple SNPs in linkage disequilibrium," *Genetic Epidemiology*, vol. 33, no. 6, pp. 497–507, 2009.

[13] J. Y. Tzeng and D. Zhang, "Haplotype-based association analysis via variance-components score test," *American Journal of Human Genetics*, vol. 81, no. 5, pp. 927–938, 2007.

[14] I. Ionita-Laza, S. Lee, V. Makarov, J. D. Buxbaum, and X. Lin, "Sequence kernel association tests for the combined effect of rare and common variants," *American Journal of Human Genetics*, vol. 92, no. 6, pp. 841–853, 2013.

[15] S. Lee, M. J. Emond, M. J. Bamshad et al., "Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies," *American Journal of Human Genetics*, vol. 91, no. 2, pp. 224–237, 2012.

[16] S. Lee, M. C. Wu, and X. Lin, "Optimal tests for rare variant effects in sequencing association studies," *Biostatistics*, vol. 13, no. 4, pp. 762–775, 2012.

[17] M. C. Wu, P. Kraft, M. P. Epstein et al., "Powerful SNP-set analysis for case-control genome-wide association studies," *American Journal of Human Genetics*, vol. 86, no. 6, pp. 929–942, 2010.

[18] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin, "Rare-variant association testing for sequencing data with the sequence kernel association test," *American Journal of Human Genetics*, vol. 89, no. 1, pp. 82–93, 2011.

[19] W. J. Gauderman, C. Murcray, F. Gilliland, and D. V. Conti, "Testing association between disease and multiple SNPs in a candidate gene," *Genetic Epidemiology*, vol. 31, no. 5, pp. 383–395, 2007.

[20] K. Wang and D. Abbott, "A principal components regression approach to multilocus genetic association studies," *Genetic Epidemiology*, vol. 32, no. 2, pp. 108–118, 2008.

[21] F. Zhang, X. Guo, S. Wu et al., "Genome-wide pathway association studies of multiple correlated quantitative phenotypes using principle component analyses," *PLOS One*, vol. 7, no. 12, article e53320, 2012.

[22] S. Basu and W. Pan, "Comparison of statistical tests for disease association with rare variants," *Genetic Epidemiology*, vol. 35, no. 7, pp. 606–619, 2011.

[23] D. H. Ballard, J. Cho, and H. Zhao, "Comparisons of multi-marker association methods to detect association between a candidate region and disease," *Nature Genetics*, vol. 34, no. 3, pp. 201–212, 2010.

[24] J. Chen and N. Chatterjee, "Exploiting Hardy-Weinberg equilibrium for efficient screening of single SNP associations from case-control studies," *Human Heredity*, vol. 63, no. 3-4, pp. 196–204, 2007.

[25] S. Luo, B. Mukherjee, J. Chen, and N. Chatterjee, "Shrinkage estimation for robust and efficient screening of single-SNP association from case-control genome-wide association studies," *Genetic Epidemiology*, vol. 33, no. 8, pp. 740–750, 2009.

[26] L. Amundadottir, P. Kraft, R. Z. Stolzenberg-Solomon et al., "Genome-wide association study identifies variants in the *ABO* locus associated with susceptibility to pancreatic cancer," *Nature Genetics*, vol. 41, no. 9, pp. 986–990, 2009.

[27] G. M. Petersen, L. Amundadottir, C. S. Fuchs et al., "A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33," *Nature Genetics*, vol. 42, no. 3, pp. 224–228, 2010.

[28] L. C. Kwee, D. Liu, X. Lin, D. Ghosh, and M. P. Epstein, "A powerful and flexible multilocus association test for quantitative traits," *American Journal of Human Genetics*, vol. 82, no. 2, pp. 386–397, 2008.

[29] J. C. Barrett, B. Fry, J. Maller, and M. J. Daly, "Haploview: analysis and visualization of LD and haplotype maps," *Bioinformatics*, vol. 21, no. 2, pp. 263–265, 2005.

[30] S. X. Chen and Y. L. Qin, "A two-sample test for high-dimensional data with applications to gene-set testing," *Annals of Statistics*, vol. 38, no. 2, pp. 808–835, 2010.

[31] D. Fallin, A. Cohen, L. Essioux et al., "Genetic analysis of Case/Control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease," *Genome Research*, vol. 11, no. 1, pp. 143–151, 2001.

[32] J. C. Long, R. C. Williams, and M. Urbanek, "An E-M algorithm and testing strategy for multiple-locus haplotypes," *American Journal of Human Genetics*, vol. 56, no. 3, pp. 799–810, 1995.

[33] D. J. Schaid, C. M. Rowland, D. E. Tines, R. M. Jacobson, and G. A. Poland, "Score Tests for Association between Traits and Haplotypes when Linkage Phase Is Ambiguous," *American Journal of Human Genetics*, vol. 70, no. 2, pp. 425–434, 2002.

[34] Q. Sha, J. Dong, R. Jiang, and S. Zhang, "Tests of association between quantitative traits and haplotypes in a reduced-dimensional space," *Annals of Human Genetics*, vol. 69, no. 6, pp. 715–732, 2005.

[35] J. Y. Tzeng, B. Devlin, L. Wasserman, and K. Roeder, "On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit," *American Journal of Human Genetics*, vol. 72, no. 4, pp. 891–902, 2003.

[36] D. V. Zaykin, P. H. Westfall, S. S. Young, M. A. Karnoub, M. J. Wagner, and M. G. Ehm, "Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals," *Human Heredity*, vol. 53, no. 2, pp. 79–91, 2002.

[37] S. Zhang, A. J. Pakstis, K. K. Kidd, and H. Zhao, "Comparisons of Two Methods for Haplotype Reconstruction and Haplotype Frequency Estimation from PSopulation Data," *American Journal of Human Genetics*, vol. 69, no. 4, pp. 906–912, 2001.

[38] N. Chatterjee, Y. H. Chen, S. Luo, and J. Carroll, "Genome-wide pathway association studies of multiple correlated quantitative phenotypes using principle component analyses," *European Journal of Human Genetics*, vol. 24, pp. 489–502, 2009.

[39] S. Valle, W. Li, and S. J. Qin, "Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods," *Industrial and Engineering Chemistry Research*, vol. 38, no. 11, pp. 4389–4401, 1999.

[40] Z. Li, W. Zhang, D. Pan, and Q. Li, "Power calculation of multi-step combined principal components with applications to genetic association studies," *Scientific Reports*, vol. 6, no. 1, pp. 1–10, 2016.