PLOS ONE

RESEARCH ARTICLE

A Comparative Study of Five Association Tests Based on CpG Set for Epigenome-Wide Association Studies

Qiuyi Zhang, Yang Zhao, Ruyang Zhang, Yongyue Wei, Honggang Yi, Fang Shao, Feng Chen*

Department of Biostatistics, School of Public Health, Nanjing Medical University, Nanjing, China, 211166

* fengchen@njmu.edu.cn

Citation: Zhang Q, Zhao Y, Zhang R, Wei Y, Yi H, Shao F, et al. (2016) A Comparative Study of Five Association Tests Based on CpG Set for Epigenome-Wide Association Studies. PLoS ONE 11(6): e0156895. doi:10.1371/journal.pone.0156895

Editor: Karen Conneely, Emory University, UNITED STATES

Received: February 13, 2016

Accepted: May 20, 2016

Published: June 3, 2016

Data Availability Statement: The rheumatoid arthritis dataset is available from Gene Expression Omnibus (accession number GSE42861).

Funding: This work was supported by National Natural Science Foundation of China (No. 81530088, 81473070, 81373102, 81402763, 81402764, 81202283), Jiangsu Natural Science Foundation (No. BK20140907), Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No. 14KJA310002), the Priority Academic Program Development of Jiangsu Higher Education Institution (PAPD) and Top-notch Academic Programs Project of Jiangsu Higher Education Institutions (TAPP:

Abstract

An epigenome-wide association study (EWAS) is a large-scale study of human disease-associated epigenetic variation, specifically variation in DNA methylation. High throughput technologies enable simultaneous epigenetic profiling of DNA methylation at hundreds of thousands of CpGs across the genome. The clustering of correlated DNA methylation at CpGs is reportedly similar to that of linkage-disequilibrium (LD) correlation in genetic single nucleotide polymorphisms (SNP) variation. However, current analysis methods, such as the *t*-test and rank-sum test, may be underpowered to detect differentially methylated markers. We propose to test the association between the outcome (e.g case or control) and a set of CpG sites jointly. Here, we compared the performance of five CpG set analysis approaches: principal component analysis (PCA), supervised principal component analysis (SPCA), kernel principal component analysis (KPCA), sequence kernel association test (SKAT), and sliced inverse regression (SIR) with Hotelling's T^2 test and *t*-test using Bonferroni correction. The simulation results revealed that the first six methods can control the type I error at the significance level, while the *t*-test is conservative. SPCA and SKAT performed better than other approaches when the correlation among CpG sites was strong. For illustration, these methods were also applied to a real methylation dataset.

Introduction

DNA polymorphisms explain only a small proportion of inheritance patterns in many complex diseases [1]. Some of the missing heritability might be explained by epigenetic variation, especially DNA methylation [2]. Indeed, the DNA methylation state, rather than DNA sequence, is more determinative of gene expression levels [3]. Further, levels of DNA methylation may "record" an individual's environmental exposures, and thus methylation is a potential biomarker for disease diagnosis and risk stratification [4,5]. Because of the reversibility of DNA

PPZY2015A067). The work was also supported by the Qing-lan Project of Jiangsu Province and the Excellent Young Teacher Project of Nanjing Medical University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

methylation, it may provide a potential therapeutic target for complex diseases, especially cancer [6,7].

Global DNA methylation status can now be profiled to determine its involvement with disease, via epigenome-wide association studies (EWASs). As an example, the HumanMethylation450 array from Illumina can assess methylation levels at more than 485,000 CpG markers [8]. The estimated proportion of DNA methylation (β-value) varies between 0 (unmethylated) and 1 (completely methylated). The aim of the analysis of EWAS data from case-control studies is to detect differentially methylated positions (DMPs), namely, CpGs that show a significant change in methylation between cases and controls [9]. Among the existing methods for DMP detection, the *t*-test and rank-sum test are the most commonly used [10]. Several advanced methods, such as mixture models, logistic M values, and generalized exponential tilt model, have been proposed recently [11–13].

Liu et al. have shown that the clustering of correlated DNA methylation at CpGs is similar to that of linkage disequilibrium (LD) correlation in genetic SNP variation but for a much shorter distance—the correlation is reduced by half for CpGs within 500bp, and it is weak for CpGs within 2kb [14]. Some clustering of methylated CpGs appears to be genetically driven, thus, they call these sets of correlated CpGs "GeMes", for genetically controlled methylation clusters. Similar to LD blocks in GWASs, this type of correlated methylation structure can be a useful tool for guiding custom array design, efficient statistical approaches, and interpretation of EWASs [15].

Considering the correlation structure among CpG sites, the above methods for DMP detection, which are based on single-locus analysis, may be underpowered to detect associations. We hypothesized that an association test on a set of biologically related CpG sites may improve the power in EWAS analysis. This improvement may result from two characteristics: First, the number of tests is reduced if CpG sites are tested by set [16]. Second, a joint test can fully utilize information contained among the multiple loci.

In this study, we sought to identify joint testing methods that may offer improved power to detect associations and tested the association between disease outcome and CpG levels using several set-based methods: PCA, SPCA, KPCA, SKAT, and SIR (briefly described below). We then used simulated datasets to compare the performance of these five CpG set analysis approaches with Hotelling's T^2 test and *t*-test with Bonferroni correction. Additionally, we analyzed publicly available DNA methylation data from a rheumatoid arthritis (RA) dataset [17] for practical application of the methods.

Methods

Let *i* denote the *i*th individual. For a CpG set, we used $G_{i1},G_{i2},\ldots,G_{ip}$ to denote DNA methylation proportions at the *p* CpG sites from the *i*th individual. When the outcome variables are dichotomous (e.g., $y = 1/0$ for case or control):

$$\boldsymbol{Logit}\ \boldsymbol{P}(\boldsymbol{y}_i = \boldsymbol{1}) = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}'\,\boldsymbol{X}_i + \boldsymbol{\beta}'\,\boldsymbol{G}_i,$$

where $\boldsymbol{X_i} = (X_{i1}, X_{i2},\ldots, X_{im})$ denotes the covariates.

PCA

Principal component analysis (PCA) is a classical multivariate method for the analysis of non-independent variables. When the *p* explanatory variables are correlated, it is possible to use a few ($k<<p$) top principal components (PCs) to replace the explanatory variables in the

regression analysis [18–21]. In our analysis, we used the first k PCs instead of p CpGs to test the association with the disease outcome, in which k is the number of PCs that explain more than 80% percent of the total variation. A k-df likelihood ratio test can be used to test the significance of the CpG set.

SPCA

SPCA (supervised principal component analysis) is a supervised dimension reduction approach [22–24]. The SPCA model is:

$$\boldsymbol{Logit}\ \boldsymbol{P}(\boldsymbol{y}_i = \mathbf{1}) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \boldsymbol{PC}_1 + \boldsymbol{\varepsilon}_j.$$

Compared to traditional PCA, which uses all CpGs in a set to extract the PCs, only those CpGs with the strongest correlation with the outcome are used to perform SPCA, and PC_1 is the first principal component. After variable selection, the test statistic $\boldsymbol{T} = \hat{\boldsymbol{\beta}}_1 / \boldsymbol{s.e.}(\hat{\boldsymbol{\beta}}_1)$ is no longer approximated well by a t-distribution, so we used the distribution proposed by Chen et al. for the hypothesis testing [25].

KPCA

Kernel principal component analysis (KPCA) is a nonlinear extension of traditional PCA that has been studied intensively recently in the field of machine learning [26–29]. Given the observations, we first map the data nonlinearly into a higher-dimensional feature space F by

$$\begin{aligned} \Phi: \quad & \boldsymbol{R}^M \rightarrow \boldsymbol{F} \\ & \boldsymbol{x} \rightarrow \boldsymbol{X} \end{aligned},$$

where $\boldsymbol{\phi}$ is a nonlinear function. Then, a kernel matrix $\boldsymbol{K}$ is formed using the inner products of new feature vectors. A standard PCA is performed on the centralized $\boldsymbol{K}$, which is the estimate of the covariance matrix of the new feature vector in F_1. Such a nonlinear PCA from the original data may be constructed to a linear PCA from the kernel matrix $\boldsymbol{K}$.

Commonly used kernel functions include linear kernel, polynomial kernel, radial basis function (RBF) kernel, IBS kernel, and weighted IBS kernel [30]. In particular, KPCA with linear kernel is standard linear PCA. In this study, we chose the RBF kernel due to its flexibility in choosing the associated parameter. The parameter $\boldsymbol{\sigma}$ is set to 0.01 and the threshold is set to 80%.

SKAT

The sequence kernel association test (SKAT) is a supervised, flexible, computationally efficient regression method. It has been used to test for the association between a set of genetic variants and a continuous or dichotomous trait [31]. Considering the correlation among the CpG markers, we used SKAT to test the association between the trait and a set of CpGs.

To increase power, SKAT tests H_0 by assuming each β_j follows an arbitrary distribution with a mean of zero and a variance of $w_j\boldsymbol{\tau}$, where $\boldsymbol{\tau}$ is a variance component and w_j is a prespecified weight for variant j. H_0: $\boldsymbol{\beta} = 0$ is equivalent to testing H_0: $\boldsymbol{\tau} = 0$, which can be conveniently tested with a variance-component score test in the corresponding mixed model. The variance-component score statistic is

$$\boldsymbol{Q} = (\mathrm{y} - \hat{\boldsymbol{\mu}})' \mathrm{K} (\mathrm{y} - \hat{\boldsymbol{\mu}}),$$

where $\boldsymbol{K} = \boldsymbol{GWWG}'$, $\boldsymbol{G}$ is an $n \times p$ matrix with the (i, j)-th element being the genotype of variant j of subject i, and $\boldsymbol{W} = \mathrm{diag}(w_1, w_2, \ldots, w_p)$ contains the weights of the p variants. In this study, the matrix $\boldsymbol{G}$ is quantitative and denotes the methylation values. We set $w_j = 1$; that is, all variants are weighted equally.

SIR

Sliced inverse regression (SIR) is a novel data-analytic tool for reducing the dimension of the input variable $\boldsymbol{G}$ [32]. Instead of regressing y against $\boldsymbol{G}$ directly, SIR regresses $\boldsymbol{G}$ against y (inverse regression) by fitting $\boldsymbol{\eta}(y) = E(G|y)$.

To perform a SIR analysis, we first standardized the explanatory variable $\boldsymbol{G}$ to $Z = \sum_{GG}^{-1/2}[G - E(G)]$, where Σ_{GG} is the sample covariance matrix of G. Second, we sliced the range of the response variable y into H intervals, $I_1, \ldots, I_h$, and partitioned the whole dataset into several slices according to the y value. Let the proportion of the y_i that falls in the slice h be denoted as $\hat{p}_h = (1/n)\sum_{i=1}^{n} \delta_h(y_i)$. The value of $\boldsymbol{\delta}_h(y_i)$ is 0 or 1 depending on whether y_i falls into the hth slice or not. Third, we calculated the sample mean of $\boldsymbol{Z}$ within each slice, denoted as $\hat{m}_h = (1/n\hat{p}_h)\sum_{y_i \in I_k} z_i$. A principal component analysis was then applied to $\hat{m}_h$, extracting the most important K-dimensional affine subspace for tracking the inverse regression curve $E(G|y)$. Finally, we output SIR after retransforming these components back to the original scale.

Simulations

We performed simulations to evaluate the type I error and power of the five CpG set analysis approaches, in comparison to a t-test using Bonferroni correction and Hotelling's T^2 test. For the t-test, we extracted the minimum P-value as the whole P-value of the CpG sites (the P value of a CpG set). We generated the simulated datasets by using a disease model

$$\boldsymbol{Logit}\ \boldsymbol{P}(\boldsymbol{D}_i = \mathbf{1}) = \boldsymbol{\beta}_0 + \sum_{j-1}^{C} \boldsymbol{\beta}_j \boldsymbol{G}_{ij},$$

in which C is the number of causal CpGs. We used the program RandGen, a free program for generating random numbers, to generate the correlated CpGs [33]. Users can specify sample size, the number of variables, distributions, and correlations through the RandGen input file. If we specify the correlations between variables using the Pearson correlation parameter, then RandGen conducts a possibly time-consuming search to find the necessary copula correlation (RhoController) values to produce those desired correlations.

Simulations based on virtual datasets

Each simulated dataset contained 1,000 cases and 1,000 controls. For each individual, we first generated methylation values using RandGen. Correlation coefficients for any pairs of CpGs were set from 0.2 to 0.8 by 0.2 increments. Here, we assumed that the CpG set contained 10 CpGs. Two scenarios were simulated; they differed by whether or not the distributions of each CpG site were the same.

Scenario 1 (same distribution): In each situation, the mean of the corresponding distribution was 0.2/0.4/0.6 or 0.8.

Scenario 2 (different distributions): The means of each CpG site were 0.2, 0.2, 0.3, 0.4, 0.5, 0.5, 0.6, 0.7, 0.8, and 0.8, respectively.

The outcome for each individual was determined by the above disease model. We set $C = 0$ (no causal CpG site in the set) to evaluate type I error, which was defined as the proportion of

Table 1. Parameter settings of virtual datasets.

Simulations	Number of causal CpGs	Location of causal CpGs	Correlation coefficient (r)	Values of β_j
Scenario 1				
1.1	0	-	0.2/0.4/0.6/0.8	-
1.2	1	1	0.2/0.4/0.6/0.8	0.5/0.6/0.7/0.8/0.9/1.0
1.3	2	1 and 2	0.2/0.4/0.6/0.8	0.1/0.2/0.3/0.4/0.5
Scenario 2				
2.1	0	-	0.2/0.4/0.6/0.8	-
2.2	1	1	0.2/0.4/0.6/0.8	0.5/0.6/0.7/0.8/0.9/1.0
2.3	1	5	0.2/0.4/0.6/0.8	0.5/0.6/0.7/0.8/0.9/1.0
2.4	1	10	0.2/0.4/0.6/0.8	0.5/0.6/0.7/0.8/0.9/1.0
2.5	2	1 and 5	0.2/0.4/0.6/0.8	0.1/0.2/0.3/0.4/0.5
2.6	2	1 and 10	0.2/0.4/0.6/0.8	0.1/0.2/0.3/0.4/0.5
2.7	2	5 and 10	0.2/0.4/0.6/0.8	0.1/0.2/0.3/0.4/0.5

doi:10.1371/journal.pone.0156895.t001

"falsely" rejected H_0 in the 5,000 replications. To evaluate the power of the seven methods, we assumed $C = 1$ and $C = 2$. For each parameter setting, we generated 1,000 simulated datasets to calculate the power at the significance level of 0.05. Parameters of simulations are described in Table 1.

Simulations based on real DNA methylation datasets

We also simulated the CpG sets in a more realistic scenario by using a real DNA methylation dataset as the template. We used data from the Gene Expression Omnibus (GEO) generated from the Illumina HumanMethylation450 array data on whole blood (accession number GSE42861). This study examined methylation differences between RA patients (n = 354) and healthy controls (n = 335). We selected protein tyrosine phosphatase, receptor type, D (*PTPRD*) and mutL homolog 1 (*MLH1*) gene regions to generate the simulated methylation data. *PTPRD* is located on Chr 9 and *MLH1* is located on Chr 3. The CpG sites we chose are located within 1Kb of *PTPRD* and *MLH1* genes. Six CpGs on *PTPRD* (IlmnID: cg08719869, cg09371281, cg09781601, cg13723825, cg14080967, cg14458619) and nine on *MLH1* (IlmnID: cg02103401, cg04726821, cg04841293, cg05670953, cg10990993, cg11291081, cg18320188, cg21109167, cg24607398) were considered. Respectively, the correlation coefficient matrices of the two CpG sets are

$$R_1 = \begin{bmatrix} \mathbf{1} & & & & & \\ \mathbf{0.264} & \mathbf{1} & & & & \\ \mathbf{0.469} & \mathbf{0.257} & \mathbf{1} & & & \\ \mathbf{0.778} & \mathbf{0.224} & \mathbf{0.458} & \mathbf{1} & & \\ \mathbf{0.890} & \mathbf{0.248} & \mathbf{0.410} & \mathbf{0.819} & \mathbf{1} & \\ \mathbf{0.374} & \mathbf{0.894} & \mathbf{0.286} & \mathbf{0.315} & \mathbf{0.364} & \mathbf{1} \end{bmatrix}$$

and

$$R_2 = \begin{bmatrix} 1 & & & & & & & & \\ 0.432 & 1 & & & & & & & \\ 0.138 & 0.303 & 1 & & & & & & \\ 0.331 & 0.709 & 0.292 & 1 & & & & & \\ 0.679 & 0.576 & 0.146 & 0.440 & 1 & & & & \\ 0.272 & 0.600 & 0.268 & 0.660 & 0.308 & 1 & & & \\ 0.311 & 0.421 & 0.538 & 0.518 & 0.307 & 0.443 & 1 & & \\ 0.113 & 0.585 & 0.453 & 0.668 & 0.262 & 0.417 & 0.449 & 1 & \\ 0.610 & 0.498 & 0.065 & 0.378 & 0.710 & 0.299 & 0.210 & 0.160 & 1 \end{bmatrix}.$$

We simulated 100 individuals with half as cases and half as controls. We set $C = 0$ to evaluate type I error and $C = 1$ to evaluate power, with 5,000 datasets produced to calculate the type I error rate, and 1,000 datasets generated for the evaluation of power. The first CpG site of each gene was defined as the causal CpG. Details of parameter settings are shown in Table 2.

Applications

The seven methods under comparison were applied to an RA DNA methylation dataset (described above). Original analysis of the dataset showed that the effect of genotype on RA risk appears to be mediated by DNA methylation changes in five genes: *GSTA2*, *PBX2*, *C6orf10*, *HLA-DQB2*, and *GPSM3* [17]. We restricted our analysis to *GSTA2* and *PBX2*, which are about 13.4 kb and 5.9 kb in length and include 6 CpGs and 51CpGs, respectively. The analyses were performed using R packages (version 3.0.2). The "superpc", "kernlab", "SKAT", "dr", and "Hotelling" packages were used to perform SPCA, KPCA, SKAT, SIR, and T^2 analyses, respectively.

Results

Results of simulations based on virtual datasets

Type I error. Type I error rates of PCA, SPCA, KPCA, SKAT, SIR, T^2, and *t*-test based on 10 CpGs are presented in Table 3 and Tables A-C in S1 File. Whether or not the CpGs were from the same distribution, the five set-based methods, as well as the T^2 test, could control the type I error at the 0.05 significance level, while the *t*-test became increasingly conservative as the correlation among CpGs increased.

Power. Estimated power from scenarios 1.2 and 2.3 is presented in Fig 1. The powers of PCA, SPCA, KPCA, and SKAT increased as the correlation among CpGs increased, but there were no apparent trends for SIR, T^2, and *t*-test. When correlation was strong ($r = 0.6$, $r = 0.8$), SKAT and SPCA were more powerful than the *t*-test. However, if there was weak correlation, there were no apparent differences among the three methods. We also found that the power was lower for SIR and T^2 than for other methods, in general.

Table 2. Parameter settings based on real methylation datasets.

Simulations	Number of causal CpGs	Location of causal CpGs	Values of β_j
1	0	-	-
2	1	1	4.0/5.0

doi:10.1371/journal.pone.0156895.t002

Table 3. Empirical Type I error rates at $\alpha = 0.05$ level under different scenarios.

Same distribution (mean methylation level = 0.6)							
r	**PCA**	**SPCA**	**KPCA**	**SKAT**	**SIR**	**T^2**	***t*-test**
0.2	0.0504	0.0504	0.0546	0.0456	0.0492	0.0412	0.0486
0.4	0.0512	0.0506	0.0500	0.0490	0.0536	0.0498	0.0452
0.6	0.0472	0.0530	0.0572	0.0478	0.0438	0.0506	0.0340
0.8	0.0470	0.0514	0.0456	0.0446	0.0460	0.0468	0.0244
Different distributions							
r	**PCA**	**SPCA**	**KPCA**	**SKAT**	**SIR**	**T^2**	***t*-test**
0.2	0.0494	0.0496	0.0518	0.0560	0.0552	0.0544	0.0450
0.4	0.0424	0.0478	0.0482	0.0504	0.0486	0.0524	0.0422
0.6	0.0476	0.0464	0.0498	0.0510	0.0454	0.0512	0.0356
0.8	0.0420	0.0538	0.0506	0.0484	0.0482	0.0514	0.0184

doi:10.1371/journal.pone.0156895.t003

Results from the simulations in scenarios 1.3 and 2.5 are presented in Fig 2. Power for all seven methods increased when the correlations became stronger. When the distributions of each CpG were the same, both SKAT and SPCA were more powerful than the other methods, independent of correlation strength. In contrast, when the distributions of each CpG were different, the powers of SKAT and SPCA were higher than *t*-test when the correlation was strong.

Considering that the average number of CpGs per gene on the HM450K array is ~17, we simulated a CpG set with 20 CpGs. Figure A in S1 File shows that SKAT and SPCA remained more powerful than other methods when the correlation was strong.

Results of simulations based on a real DNA methylation dataset

Table 4 presents the results based on the *PTPRD* gene. PCA, SPCA, KPCA, SKAT, SIR, and T^2 could control type I error at the significance level of 0.05, while the *t*-test was conservative. Power results are presented in Fig 3(a). For both $\beta_1 = 4.0$ and 5.0, SKAT and SPCA were more powerful than the other five methods. Among the seven methods, the power was lowest for the *t*-test. Results based on the *MLH1* gene were similar to those from the *PTPRD* gene [Table 4 and Fig 3(b)].

Application to real data

We applied the seven methods to two CpG sets from an RA methylation dataset from the GEO data repository. The *P*-values for the CpG sets are presented in Table 5. For the first CpG set, the *P*-value of SPCA was 4.06E-05, the lowest of the seven methods. SKAT was second to SPCA with a *P*-value of 5.36E-04. *P*-values for PCA, *t*-test, KPCA, T^2, and SIR were 1.11E-03, 1.88E-03, 2.42E-03, 3.74E-03 and 5.39E-03, respectively. For the *PBX2* gene, the result also showed that SPCA had the best performance. The *t*-test was slightly superior to SKAT. All seven approaches yielded significant results at the significance level of 0.05 and were consistent with the original report of this dataset [17].

Discussion

The correlation structure of the DNA methylation data enables testing of the association between the disease outcome and a set of CpGs simultaneously. Here, we demonstrate that analyzing DNA methylation data using CpG set-based analysis for epigenome-wide association studies offers superior power over individual analysis. The set-based CpG association analysis

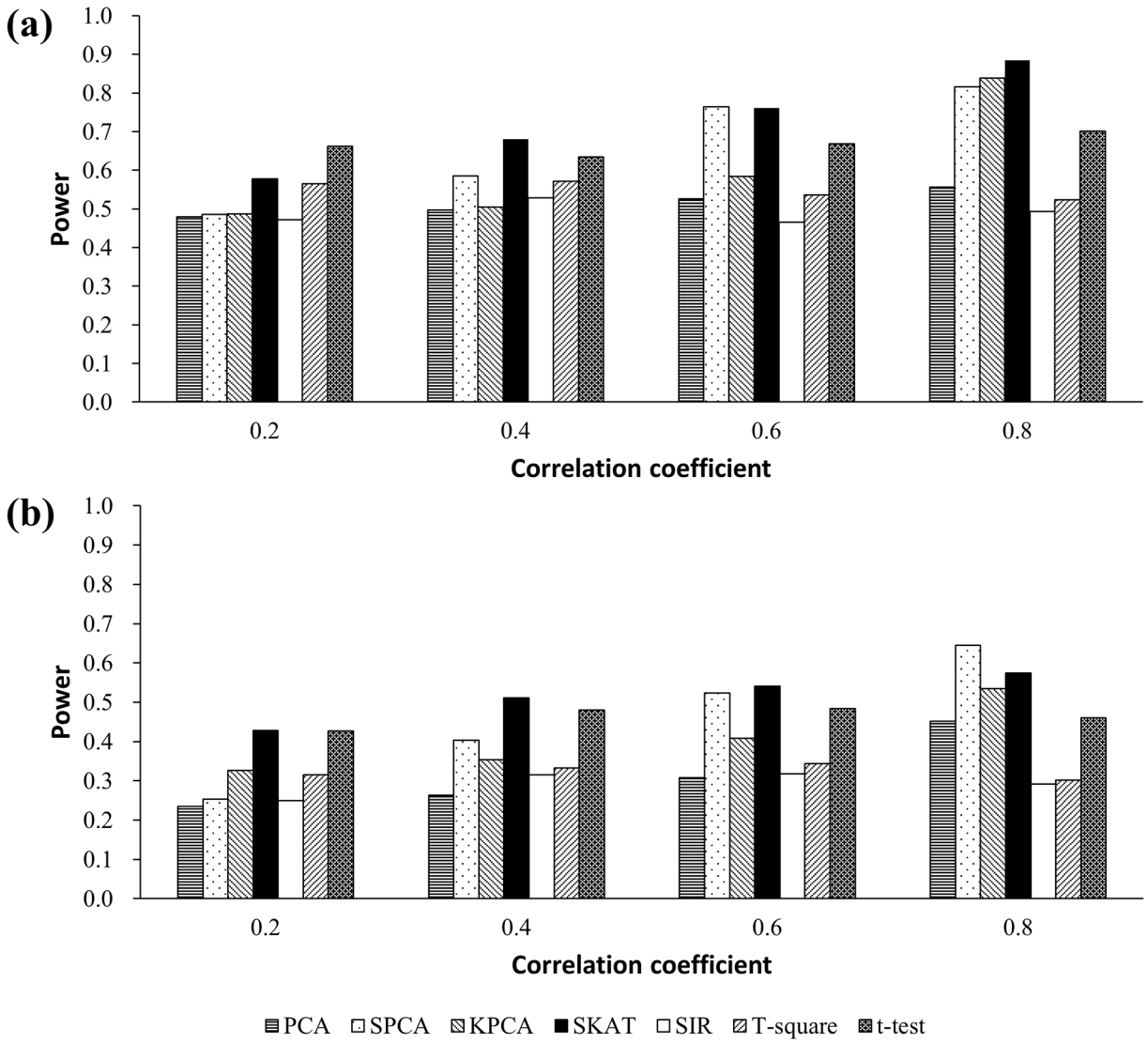

Fig 1. (a) Simulated power at single causal CpG model based on 10 CpGs from the same distribution (mean methylation level = 0.6). The regression coefficient in the disease model, $\beta_1 = 0.7$. **(b) Simulated power at single causal CpG model based on 10 CpGs from different distributions**. The 5th CpG is set as the causal CpG. The regression coefficient in the disease model, $\beta_1 = 0.7$.

doi:10.1371/journal.pone.0156895.g001

has several advantages: first, the set-based methods can "borrow" information from the correlated CpG sites; second, the set-based methods decrease the number of multiple comparisons [34–36].

In this research, we compared the performance of five CpG set analysis approaches (PCA, SPCA, KPCA, SKAT, and SIR) with Hotelling's T^2 test and *t*-test using Bonferroni correction. We found that all of these set-based methods can control the type I error at the target significance level. The *t*-test with Bonferroni correction is conservative, especially when the correlations between CpG sites are strong, and thus can be less powerful to identify the association between the outcome variable and CpG set. When the CpG sites in the set have high correlation with each other, SPCA and SKAT can combine their information and provide better-simulated power among the seven approaches. We suggest that SPCA and SKAT can be used for

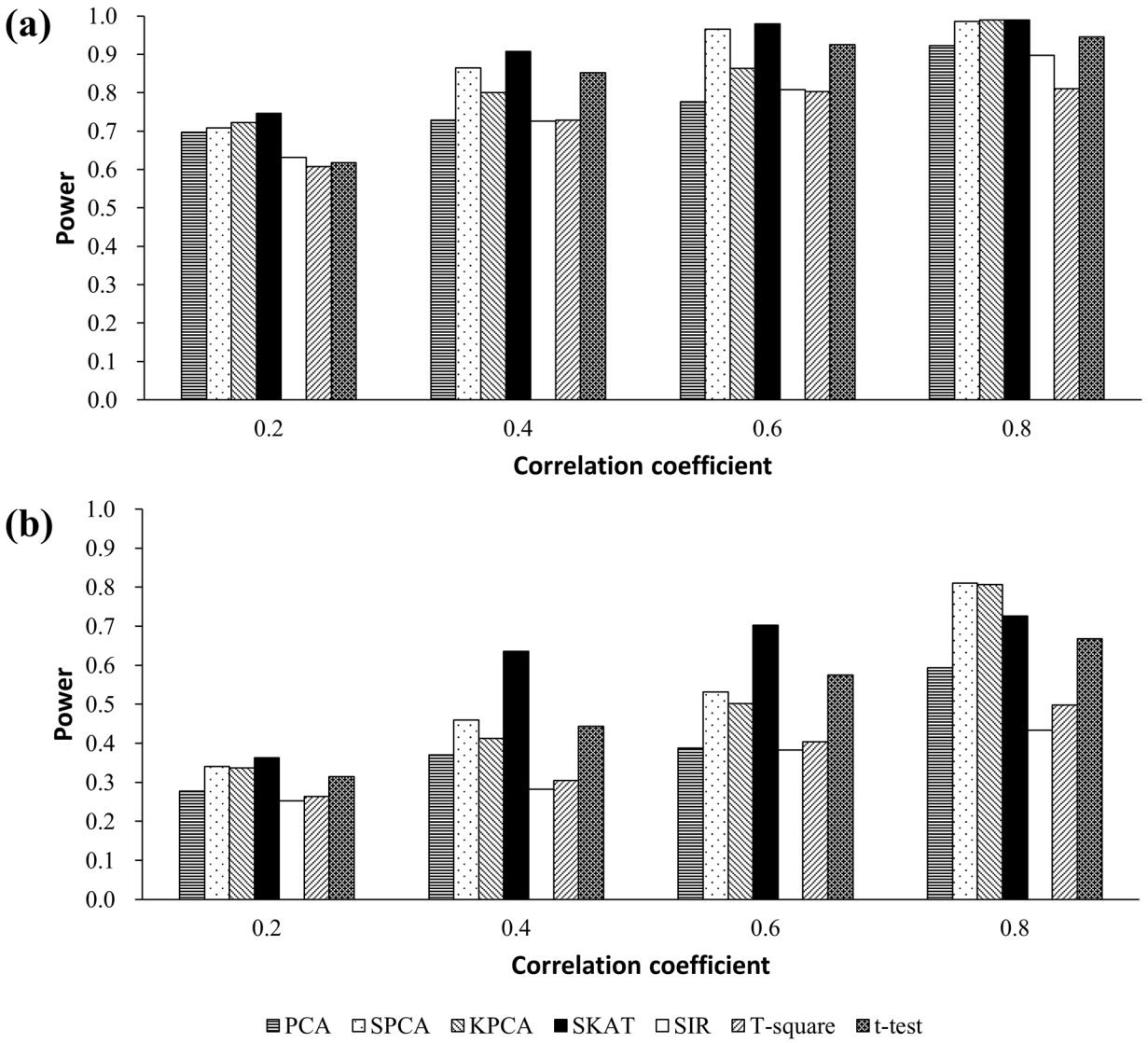

Fig 2. (a) Simulated power at two causal CpGs model based on 10 CpGs from the same distribution (mean methylation level = 0.6). The regression coefficients in the disease model, $\beta_1 = \beta_2 = 0.5$. **(b) Simulated power at two causal CpGs model based on 10 CpGs from different distributions**. 1st and 5th CpGs are set as the causal CpGs. The regression coefficients in the disease model, $\beta_1 = \beta_2 = 0.5$.

doi:10.1371/journal.pone.0156895.g002

CpG set analysis and screening the association across the entire epigenome. In the RA methylation dataset, we compared the methylation differences of two CpG sets (*GSTA2* and *PBX2*) between patients and healthy controls. All these simulated studies and applications in real data analysis suggest that set-based methods may be used in DNA methylation data analysis.

Table 4. Empirical Type I error rates based on a real methylation dataset.

Gene	PCA	SPCA	KPCA	SKAT	SIR	T^2	*t*-test
PTPRD	0.0572	0.0578	0.0554	0.0480	0.0568	0.0584	0.0372
MLH1	0.0537	0.0556	0.0483	0.0514	0.0582	0.0518	0.0422

doi:10.1371/journal.pone.0156895.t004

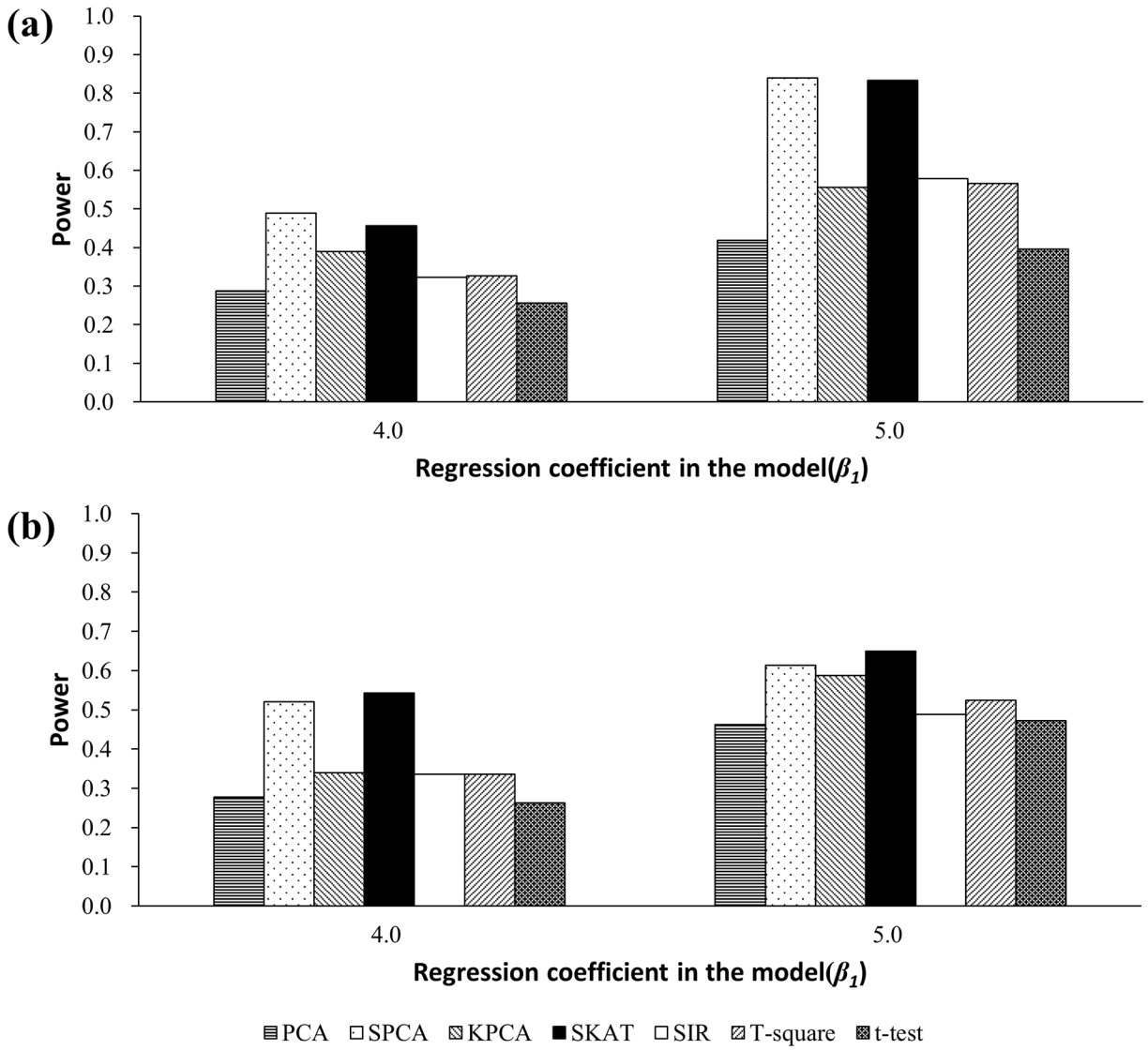

Fig 3. (a) Simulated power based on the *PTPRD* gene. (b) Simulated power based on the *MLH1* gene.

doi:10.1371/journal.pone.0156895.g003

SPCA is a supervised dimension reduction approach, which only includes the disease-relevant CpGs before the extraction of the principle components. Thus, this method performs better than PCA under different situations. As the KPCA does not rule out the irrelevant CpGs, the power of KPCA is inferior to SPCA but better than PCA in most occasions. We also find KPCA consumes more computational resources. SKAT uses variance component testing

Table 5. CpG set analysis results of DNA methylation datasets from epigenome studies.

Gene	Number of CpGs	*P*-value for the CpG set						
		PCA	SPCA	KPCA	SKAT	SIR	T^2	*t*-test
GSTA2	6	1.11E-03	4.06E-05	2.42E-03	5.36E-04	5.39E-03	3.74E-03	1.88E-03
PBX2	51	3.32E-06	1.03E-10	9.55E-08	1.57E-08	2.80E-02	1.23E-03	4.08E-09

doi:10.1371/journal.pone.0156895.t005

framework to increase test power. If the CpGs in a region are highly correlated, the reduced degree of freedom improves the statistical power. The power of SIR is almost the lowest throughout the simulations. One possible reason is that the outcome variable is binary and can be divided into only two slices. Although, in theory, Hotelling's T^2 test has the ability to summarize the information from correlated CpGs, it is less powerful than SPCA and PCA in our simulations. This may result from the assumptions of multiple normal distribution being violated for DNA methylation data. Thus, SIR and Hotelling's T^2 test are not recommended for application in CpG set analysis for EWAS studies.

In summary, we propose to use set-based method for DNA methylation data analysis and compare the performance of CpG set-based analysis for DNA methylation data. We suggest using SPCA and SKAT to improve test power. However, there remain some limitations in our study. First, the virtual simulated datasets were generated based only on multivariate beta distribution. Other distributions such as inverse logit transformation of a multivariate normal distribution should be considered. Several recent studies have noticed that measured methylation may exhibit different levels of variability in different groups, possibly due to batch effects [7]. Therefore, some new tests that capture differences in both mean and variance of methylation levels, such as semiparametric tests [13], have been proposed. We will discuss the performance of these methods in the same situation later. Second, for the methods PCA, KPCA, and SIR, further studies should be performed to identify the effect of the number of PCs on the power. Third, more complicated situations, such as the interactions between CpG sets and the methylation-mediated genetic risks in the genome-wide scan, are not covered here but will be considered in future studies.

Supporting Information

S1 File. Table A, Empirical Type I error rates at $\alpha = 0.05$ level based on 10 CpGs from the same distribution (mean level = 0.2). Table B, Empirical Type I error rates at $\alpha = 0.05$ level based on 10 CpGs from the same distribution (mean level = 0.4). Table C, Empirical Type I error rates at $\alpha = 0.05$ level based on 10 CpGs from the same distribution (mean level = 0.8). Figure A, Simulated power at single causal CpG model based on 20 CpGs from the same distribution (mean methylation level = 0.6). The regression coefficient in the disease model, $\beta_1 = 0.7$. (DOCX)

Acknowledgments

The authors thank all of the study participants for their contributions to this study. The comments and suggestions from the reviewers are also deeply appreciated.

Author Contributions

Conceived and designed the experiments: QYZ YZ FC. Performed the experiments: QYZ YZ RYZ. Analyzed the data: QYZ. Contributed reagents/materials/analysis tools: QYZ YYW HGY FS. Wrote the paper: QYZ YZ.

References

1. Manolio TA, Brooks LD, Collins FS (2008) A HapMap harvest of insights into the genetics of common disease. Journal of Clinical Investigation 118: 1590–1605. doi: 10.1172/JCI34772 PMID: 18451988
2. Maher B (2008) Personal genomes: The case of the missing heritability. Nature 456: 18–21. doi: 10.1038/456018a PMID: 18987709
3. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747–753. doi: 10.1038/nature08494 PMID: 19812666

4. Laird PW (2003) The power and the promise of DNA methylation markers. Nature Reviews Cancer 3: 253–266. PMID: 12671664
5. Bock C (2009) Epigenetic biomarker development. Epigenomics 1: 99–110. doi: 10.2217/epi.09.6 PMID: 22122639
6. Das PM, Singal R (2004) DNA methylation and cancer. Journal of Clinical Oncology 22: 4632–4642. PMID: 15542813
7. Bock C (2012) Analysing and interpreting DNA methylation data. Nature Reviews Genetics 13: 705–719. doi: 10.1038/nrg3273 PMID: 22986265
8. Laird PW (2010) Principles and challenges of genome-wide DNA methylation analysis. Nature Reviews Genetics 11: 191–203. doi: 10.1038/nrg2732 PMID: 20125086
9. Saadati M, Benner A (2014) Statistical challenges of high-dimensional methylation data. Statistics in Medicine 33: 5347–5357. doi: 10.1002/sim.6251 PMID: 25042556
10. Xu HY, Podolsky RH, Ryu DW, Wang XL, Su SY, Shi HD, et al. (2013) A method to detect differentially methylated loci with next-generation sequencing. Genetic Epidemiology 37: 377–382. doi: 10.1002/gepi.21726 PMID: 23554163
11. Wang S (2011) Method to detect differentially methylated loci with case-control designs using Illumina arrays. Genetic Epidemiology 35: 686–694. doi: 10.1002/gepi.20619 PMID: 21818777
12. Du P, Zhang XA, Huang CC, Jafari N, Kibbe WA, Hou LF, et al. (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. Bmc Bioinformatics 11.
13. Chen Y, Ning Y, Hong C, Wang S (2014) Semiparametric tests for identifying differentially methylated loci with case-control designs using Illumina arrays. Genetic Epidemiology 38: 42–50. doi: 10.1002/gepi.21774 PMID: 24301455
14. Liu Y, Li X, Aryee MJ, Ekstrom TJ, Padyukov L, Klareskog L, et al. (2014) GeMes, clusters of DNA methylation under genetic control, can inform genetic and epigenetic analysis of disease. American Journal of Human Genetics 94: 485–495. doi: 10.1016/j.ajhg.2014.02.011 PMID: 24656863
15. Yip WK, Fier H, DeMeo DL, Aryee M, Laird N, Lange C. (2014) A novel method for detecting association between DNA methylation and diseases using spatial information. Genetic Epidemiology 38: 714–721. doi: 10.1002/gepi.21851 PMID: 25250875
16. Pan W (2009) Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genetic Epidemiology 33: 497–507. doi: 10.1002/gepi.20402 PMID: 19170135
17. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, et al. (2013) Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nature Biotechnology 31: 142–147. doi: 10.1038/nbt.2487 PMID: 23334450
18. Jolliffe IT (2002) Principal component analysis. New York: Springer. xxix, p. 487.
19. Zhao Y, Chen F, Zhai R, Lin X, Diao N, Christiani DC. (2012) Association test based on SNP set: logistic kernel machine based test vs. principal component analysis. PLoS One 7: e44978. doi: 10.1371/journal.pone.0044978 PMID: 23028716
20. Cai M, Dai H, Qiu Y, Zhao Y, Zhang R, Chu M, et al. (2013) SNP set association analysis for genome-wide association studies. PLoS One 8: e62495. doi: 10.1371/journal.pone.0062495 PMID: 23658731
21. Yi H, Wo H, Zhao Y, Zhang R, Dai J, Jin G, et al. (2015) Comparison of dimension reduction-based logistic regression models for case-control genome-wide association study: principal components analysis vs. partial least squares. Journal of biomedical research 29: 298–307.
22. Bair E, Tibshirani R (2004) Semi-supervised methods to predict patient survival from gene expression data. PLoS Biology 2: 511–522.
23. Bair E, Hastie T, Paul D, Tibshirani R (2006) Prediction by supervised principal components. Journal of the American Statistical Association 101: 119–137.
24. Chen X, Wang L, Smith JD, Zhang B (2008) Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes. Bioinformatics 24: 2474–2481. doi: 10.1093/bioinformatics/btn458 PMID: 18753155
25. Chen X, Wang L, Hu B, Guo MS, Barnard J, Zhu XF. (2010) Pathway-based analysis for genome-wide association studies using supervised principal components. Genetic Epidemiology 34: 716–724. doi: 10.1002/gepi.20532 PMID: 20842628
26. Scholkopf B, Smola AJ, Muller KR (1997) Kernel principal component analysis. Artificial Neural Networks—ICANN' 97: 583–588.
27. Mika S, Schölkopf B, Smola AJ, Müller K, Scholz M, Rätsch G. Kernel PCA and De-Noising in Feature Spaces; 1998. Citeseer. pp. 7.
28. Liu ZQ, Chen DC, Bensmail H (2005) Gene expression data classification with kernel principal component analysis. Journal of Biomedicine and Biotechnology: 155–159. PMID: 16046821

29. Gao QS, He YG, Yuan ZS, Zhao JH, Zhang BB, Xue FZ. (2011) Gene- or region-based association study via kernel principal component analysis. Bmc Genetics 12.

30. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, et al. (2010) Powerful single-set analysis for case-control genome-wide association studies. American Journal of Human Genetics 86: 929–942. doi: 10.1016/j.ajhg.2010.05.002 PMID: 20560208

31. Wu MC, Lee S, Cai TX, Li Y, Boehnke M, Lin XH. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. American Journal of Human Genetics 89: 82–93. doi: 10.1016/j.ajhg.2011.05.029 PMID: 21737059

32. Li KC (1991) Sliced inverse regression for dimension reduction. Journal of the American Statistical Association 86: 316–327.

33. Miller J (2002) RandGen: A program for generating random numbers. Department of Psychology, University of Otago, New Zealand, Version 2.

34. Wang X, Qin H, Sha Q (2009) Incorporating multiple-marker information to detect risk loci for rheumatoid arthritis. BMC Proc 3 Suppl 7: S28. PMID: 20018018

35. Thomas M, De Brabanter K, De Moor B (2014) New bandwidth selection criterion for Kernel PCA: approach to dimensionality reduction and classification problems. BMC Bioinformatics 15: 137. doi: 10.1186/1471-2105-15-137 PMID: 24886083

36. Zeng P, Zhao Y, Qian C, Zhang L, Zhang R, Gou J, et al. (2015) Statistical analysis for genome-wide association study. Journal of biomedical research 29: 285–297. doi: 10.7555/JBR.29.20140007 PMID: 26243515