G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2016;14(4):187-195
https://doi.org/10.5808/GI.2016.14.4.187

ORIGINAL ARTICLE

Multiple Group Testing Procedures for Analysis of High-Dimensional Genomic Data

Hyoseok Ko, Kipoong Kim, Hokeun Sun*

Department of Statistics, Pusan National University, Busan 46241, Korea

In genetic association studies with high-dimensional genomic data, multiple group testing procedures are often required in order to identify disease/trait-related genes or genetic regions, where multiple genetic sites or variants are located within the same gene or genetic region. However, statistical testing procedures based on an individual test suffer from multiple testing issues such as the control of family-wise error rate and dependent tests. Moreover, detecting only a few of genes associated with a phenotype outcome among tens of thousands of genes is of main interest in genetic association studies. In this reason regularization procedures, where a phenotype outcome regresses on all genomic markers and then regression coefficients are estimated based on a penalized likelihood, have been considered as a good alternative approach to analysis of high-dimensional genomic data. But, selection performance of regularization procedures has been rarely compared with that of statistical group testing procedures. In this article, we performed extensive simulation studies where commonly used group testing procedures such as principal component analysis, Hotelling's T^2 test, and permutation test are compared with group lasso (least absolute selection and shrinkage operator) in terms of true positive selection. Also, we applied all methods considered in simulation studies to identify genes associated with ovarian cancer from over 20,000 genetic sites generated from Illumina Infinium HumanMethylation27K Beadchip. We found a big discrepancy of selected genes between multiple group testing procedures and group lasso.

Keywords: genetic association studies, genetic selection, genetic testing, principal component analysis

Introduction

In human genetic association studies with high-dimensional genomic data, a multiple group testing procedure is often required to identify genes or genetic regions that are associated with a disease or a trait since a gene or a genetic region usually contains multiple genetic sites or variants. For instance, single nucleotide polymorphism data, DNA methylation data and sequencing data consist of tens of thousands of genes where each gene has multiple genetic sites. In order to identify genes or genetic regions associated with a phenotype outcome, we need to conduct an individual group test for each gene. However, an individual test for high-dimensional genomic data suffers from multiple testing issues such as the control of family-wise error rate (FWER) or dependent tests. In this reason, Bonferroni adjustment or false discovery rate (FDR) control methods [1-3] should be performed after computing the p-value of a multiple group test for individual genes or genetic regions.

Alternatively, regularization procedures using a penalized likelihood can be applied for analysis of high-dimensional genomic data. Basically, regularization procedures perform variable selection based on a parametric regression with a penalty function, where a phenotype outcome regresses on all of genetic sites. As a tuning parameter for sparsity is decreasing, the most outcome-related genetic sites can be sequentially selected. One of the most popular regularization procedures for high-dimensional genomic data is lasso (least absolute shrinkage and selection operator) [4-6]. For group selection such as a gene or a genetic region, group lasso can be applied to high-dimensional genomic data that has a group or a cluster structure [7, 8]. Since regularization procedures do not test but select a gene or a genetic region associated with a phenotype outcome, the control of FWER or FDR is not required. However, the selection of the optimal

Received July 27, 2016; Revised October 20, 2016; Accepted October 26, 2016

***Corresponding author:** Tel: +82-51-510-2257, Fax: +82-51-581-1459, E-mail: hsun@pusan.ac.kr

tuning parameter is crucial to determine the number of the outcome-related genes or genetic regions.

The main goal of both the individual group test and the group selection procedure is to identify disease/trait-related genes or genetic regions in analysis of high-dimensional genomic data. Although these two different statistical methods have the same goal, there have been rarely statistical literatures which compare either test performance or selection performance of two statistical methods since hypothesis testing and variable selection have been considered as completely different approaches in statistics. In genetic association studies, however, the testing procedure essentially determines whether each gene or each genetic region is significantly associated with a phenotype outcome or not, while variable selection makes a conclusion which genes or genetic regions are associated with a phenotype outcome. Therefore, we can directly compare the true positives of the group test procedure and the group selection procedure when they identify the same number of disease/trait-related genes or genetic regions.

In this article, we conduct extensive simulation studies in order to compare the performance of both group testing procedures and group selection procedure in terms of true positives. That is, the total number of correctly identified genes or genetic regions is compared when the same number of genes or genetic regions is detected. The simulation studies focus on a case-control association study with high-dimensional genomic data which has a group or a cluster structure. For group testing procedures, we consider commonly used three methods such as principal component analysis (PCA), Hotelling's T^2 test, and permutation test. For group selection procedure, we employ group lasso. These four statistical methods are also applied to real high-dimensional DNA methylation data where DNA methylation beta values of CpG sites from approximately 12,000 genes between ovarian cancer cases and healthy controls were generated from Illumina Infinium HumanMethylation27K Beadchip (Illumina Inc., San Diego, CA, USA).

Methods

Principal component analysis

PCA is one of the most common statistical approaches to data dimension reduction [9]. It basically transforms multiple variables to have orthogonality so the first principal component can be expressed as a weighted linear combination of variables. The weights of the first component are computed such that the component can account for the greatest possible variance of multiple variables. In case-control association studies of genomic data with a group structure, we can apply PCA for each gene or genetic region, where data information of multiple genetic sites within the same gene or genetic region can be reduced to a single numerical vector corresponding to the first principal component. Then, a phenotypic association of the principal component can be tested based on the independent two sample T-test. Statistical approaches based on PCA have been widely applied to analysis of high-dimensional genomic data [10-12].

Hotelling's T^2 test

Hotelling's T^2 test is one of the representative multivariate tests used when significant differences between the mean vectors of two multivariate data sets are tested. It is known as a powerful multivariate test as long as data satisfies necessary assumptions such as random samples, multivariate normality and equivalent variance and covariance matrices between two groups. In genetic association studies with microarray data, the Hotelling's T^2 test is often applied to find differentially expressed genes [13, 14]. We also applied the Hotelling's T^2 test for each gene or genetic region to test a significant difference between cases and controls. We employed an R package 'Hotelling' for simulation studies and real data analysis.

Permutation test

Permutation test is a nonparametric statistical test used when an underlying distribution of genetic data is not need to be assumed. If the derivation of a theoretical distribution of a test statistic is challenging, permutation test can be employed to compute an empirical p-value of the test statistic. For genomic data with a group structure, we first compute an individual p-value of a phenotypic association test for each genetic site. We then combine K p-values such that $-\sum_{i=1}^{K} \log p_i$, where p_i is the p-value of the the i-th genetic site, and K is the total number of genetic sites in the gene or genetic region. Next, we repeatedly permute case-control labels and compute $-\sum_{i=1}^{K} \log p_i$ for each permutation set. After we obtain the empirical distribution of $-\sum_{i=1}^{K} \log p_i$, we can easily calculate the empirical p-value from the original case and control set. This permutation based test for genomic data with a group structure has been demonstrated to be very efficient and powerful when we need to aggregate the information of multiple genetic sites [15].

Group lasso

Statistical association tests above should be conducted to each gene or each genetic region one at a time, so the individual tests cannot consider genetic correlations among genes. But, genes linked with each other on genetic pathways or genes that have an interaction effects on a phenotype outcome can have a functional relationship with each other. Since their correlations could be important information for genetic association studies, a statistical model that can includes all of genetic information is often preferred. Regularization procedures can be conducted to entire genomes in a regression framework, where a phenotype outcome is regressed on all of genetic data. The solution of regression coefficients can be achieved, constraining the parameter space of regression coefficients. This constraint actually enables to obtain the coefficient solution even if the number of genes is much greater than a sample size. Depending on the parameter space constraint, a type of regularization procedure is determined. Group lasso regularizes regression coefficients such that the sum of L_2 norm of the coefficients for each group is less than an arbitrary value [7]. The arbitrary value is corresponding to a tuning parameter value for sparsity. For a fixed tuning parameter λ, the estimated regression coefficients $\beta = (\beta_1, \beta_2, \cdots, \beta_m)^T$ of group lasso maximizes

$$Q_\lambda(\beta) = -l(\beta) + \lambda \sum_{k=1}^{m} \sqrt{m_k} \, \| \beta_k \|_2,$$

where $l(\beta)$ is a logistic likelihood, and m_k is the total number of genetic sites of the k-th gene, i.e., $\beta_k = (\beta_{k1}, \beta_{k2}, \cdots, \beta_{km_k})^T$. The L_2 norm of β_k is defined as $\| \beta_k \|_2 = \sqrt{\sum_{j=1}^{m_k} \beta_{kj}^2}$.

In genomic data analysis with a group structure, group lasso sequentially selects the most outcome-related gene or genetic region. The selection results rely on the tuning parameter value λ since we have different coefficient estimates for a value of λ. As λ decreases, the number of selected genes is gradually increased. In general, we select the k-th gene if the estimated regression coefficients β_k are nonzero. In group lasso, the estimated regression coefficients of all genetic sites belong to selected genes or genetic regions are nonzero while all genetic sites of unselected genes or genetic regions have the exactly zero regression coefficients. Therefore, group selection can be performed based on the solution of regression coefficients for a fixed value of λ. Note that the numerical values of estimated coefficients β_k are not of interest since group lasso performs selection but not prediction. That is, we see if $\beta_k = 0$ or not.

In our simulation studies and real data analysis, we first started with a relatively large λ value which is large enough that the solution to all regression coefficients can be exactly zero. In this case, no genes are selected. We then gradually decreased the value of λ until a single gene has nonzero regression coefficients of its genetic sites. This gene is considered as the top rank gene. In the same way, we can find the second ranked gene as λ continues to be decreasing. Eventually, we can obtain a list of top ranked genes. Since we compared a particular number of top selected genes with the same number of top significant genes computed by multiple testing procedures, we didn't need to find the optimal λ value in the simulation studies and real data analysis. Group lasso has been widely applied for analysis of high-dimensional genomic data [8, 16]. We used an R package 'gglasso' for simulation studies and real data analysis.

Results

Simulation studies

In this simulation study, we compared true positive rates of three group testing procedures and one selection procedure such as PCA, Hotelling's T^2 test, permutation test, and group lasso selection procedure when the same number of genes is identified by four statistical methods. We conducted two different simulation studies, where the first simulation assumed that all genes are independent with each other and the second simulation assumed that genes are correlated with each other according to a given network information.

In the first simulation studies, we assumed that a single gene consists of 5 genetic sites. Numerical data of the five genetic sites within the same gene were generated from a multivariate normal distribution of $N(\mu, \Sigma)$, where a mean vector μ has a different value between cases and controls if the gene is causal, but μ has the same value between cases and controls if the gene is noncausal. The covariance matrix Σ represents a correlation pattern of the five genetic sites within the same gene and we assumed an AR(1) correlation matrix, i.e., $\Sigma = \{\sigma_{ij}\}_{1 \le i,j \le 5} = \rho^{|i-j|}$, where ρ is a correlation coefficient fixed as ρ=0.3,0.5, or 0.7 in the simulation. We considered 1,000 genes so we have a total of 5,000 genetic sites in the simulation data, where 100 cases and 100 controls were generated. Only 20 genes out of 1,000 genes are assumed to be causal. Note that the simulated 1,000 genes are independent with each other.

Three group testing procedures were applied to individual genes and the p-values of testing the mean difference between cases and controls were computed for 1,000 genes. The p-values of each testing procedure were then listed from the smallest to the largest. Finally, top 20, 30, and 40 genes were selected for each testing procedure based on the 20, 30, and 40 smallest p-values, respectively. In contrast, group lasso procedure sequentially selects the most outcome-related genes as a tuning parameter for sparsity is decreasing. Since we can easily control the tuning parameter value, we were able to select the top 20, 30, and 40 genes. True positives rate of the selected genes from the four statistical methods were

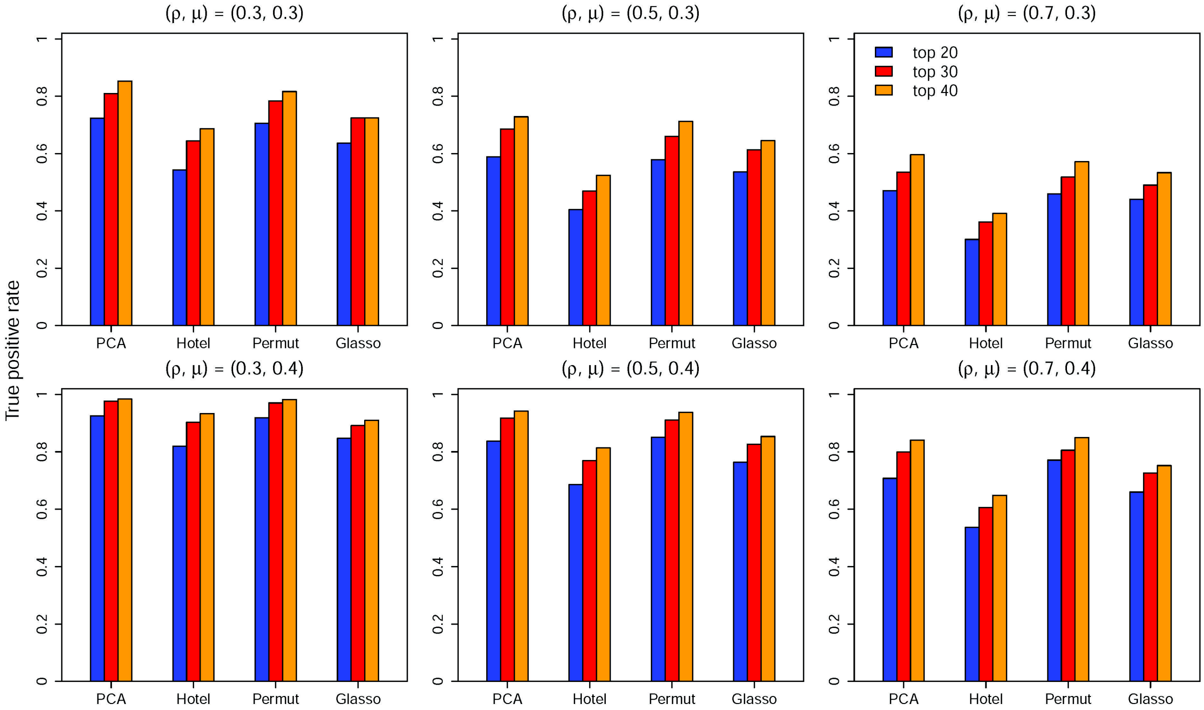

Fig. 1. Averaged true positive rates of top 20, 30, and 40 genes detected by principal component analysis (PCA), Hotelling's T^2 test (Hotel), permutation test (Permut), and group lasso are displayed along with a different correlation coefficient ρ of 0.3, 0.5, and 0.7, and a different mean difference μ of 0.3 and 0.4 between cases and controls.

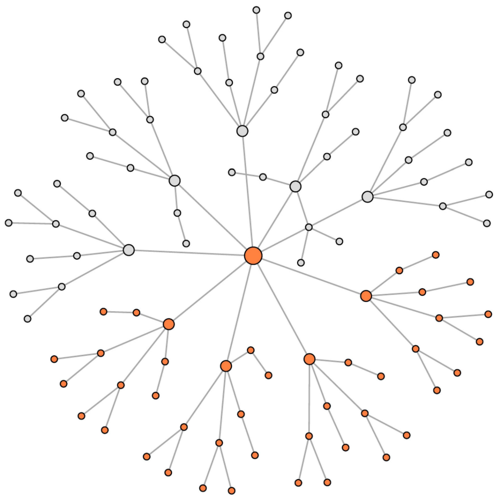

Fig. 2. An example of a simulated network graph with 100 genes used in the second simulation study is present. The colored 45 genes are assumed to be causal genes.

computed along with two different values of μ=0.3 and 0.4. That is, only 20 causal genes were assumed to have a mean difference by either 0.3 or 0.4 between cases and controls. True positive rates stand for the proportion of correctly identified genes among the 20 causal genes. The simulation was repeated 100 times and averaged true positive rates of four statistical methods over 100 simulation replications were summarized in Fig. 1.

In Fig. 1, it appears that all of true positive rates are overall increased as the mean difference is increasing and the correlation coefficient is decreasing. In high-dimensional data analysis, it is often observed that detection power is decreased due to highly correlated variables. We can see the similar result of decreased true positive rates for highly correlated genetic sites. When we compared four statistical methods, both PCA and permutation test seem to have the largest true positive rates in all simulation settings, while Hotelling's T^2 test show the lowest true positive rates in all settings. The true positive rates of group lasso procedure seem to be higher than those of Hotelling's T^2 test, but slightly lower than those of both PCA and permutation test. Consequently, we can conclude that group lasso procedure shows similar selection performance as the group hypothesis testing procedures in the first simulation study.

In the second simulation study, we generated 1,000 genes where each 100 genes are truly correlated with each other according to the simulated genetic network in Fig. 2. So, we have 10 network groups each of which consists of 100 genes. Similar to the first simulation study, genetic data was generated from N(μ,Σ), where the covariance matrix Σ is an inverse of a precision matrix Ω In a Gaussian graphical model [17], nonzero entries of the precision matrix correspond to links between two genes of a network graph.

Therefore, we could obtain the precision matrix Ω according to the given network in Fig. 2 in the same way described [12, 18]. Our simulation data consists of 100 cases and 100 controls over 1,000 genes, where only 45 genes within the same network have a different mean μ between cases and controls. Let us denote the *j*-th gene by χ_j. Since our simulation study should be conducted for genetic sites with a group structure, we additionally generated 10 genetic sites for each gene. If the first *N* genetic sites among 10 sites are causal and the other *10 - N* genetic sites are noncasual, the *N* sites of the the *j*-th gene were generated such that $\chi_{jk} = \chi_j + N(0, \sigma^2)$, $k = 1, 2, ..., N$ and $\chi_{jk} = N(0, \sigma^2)$ for $k = N + 1, ..., 10$. Finally, we have a total of 10,000 genetic sites with 200 samples. In the simulation, we discarded all simulated genes χ_j for $j = 1, 2, ..., 1{,}000$. Instead, we used only 10,000 genetic sites where each 10 sites consist of one gene. In this simulation setting, 10 genetic sites within the same gene are not only correlated with each other, but they are also correlated with the other 10 genetic sites within the different genes that are linked with each other according to the genetic network.

We also applied four statistical methods used in the first simulation study. We fixed the number of causal genetic sites per gene as *N* = 2, 5, or 10. Since only 45 genes are causal, the number of causal sites is 90, 225, or 450 among 10,000 sites, respectively. The standard deviation σ to control a noise level was set to be 1.5, 2, or 2.5. Higher standard deviation is likely to produce stronger noises, so true positive rates are expected to be decreased. Simulation was repeated 100 times and averaged true positive rates of top 10 genes to top 100 genes selected by four methods are summarized in Fig. 3.

In Fig. 3, PCA overall shows the best selection performance except when the noise level is either moderate or strong, and the number of causal genetic sites is small, i.e., *N* = 2 and σ = 2 or 2.5. As the number of causal sites in a gene is decreasing, the true positive rates of three group testing procedures seem to be decreasing together. However, we can see that the true positive rates of group lasso are almost the same regardless of the number of causal sites in a gene. As we mentioned earlier, group lasso enforces the regression coefficients of genetic sites in the selected genes to be nonzero even if only a few genetic sites are causal and majority is noncausal. In this reason, the selection performance of group lasso does not affected by the number of causal and noncausal sites in the same gene. When all of genetic sites in the same gene are causal (*N* = 10), both PCA and permutation test overwhelm the other two methods. Since computation of the test statistics of two methods is based on individual genetic sites, they should be statistically powerful when all of sites in the same gene are causal. In the second simulation, the Hotelling's T^2 test shows the worst

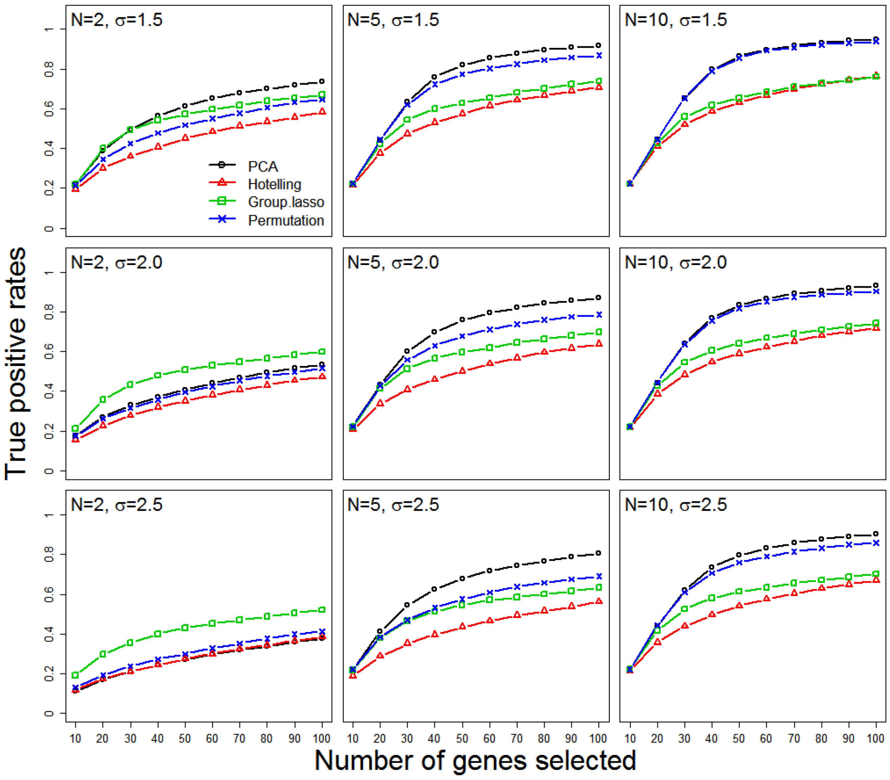

Fig. 3. Averaged true positive rates of principal component analysis (PCA), Hotelling's T^2 test, permutation test, and group lasso are displayed along with a different number of selected genes when the number of causal genetic sites (*N*) among 10 sites in a gene is 2, 5, or 10, and the standard deviation of an error terms (σ) to control a noise level is 1.5, 2, or 2.5.

selection of true positives in all simulation settings. This is due to relatively high correlation among genetic sites in the same genes. We have already seen that the true positive rates of Hotelling's T^2 test were drastically decreased as the correlation was increasing in the first simulation study.

Analysis of ovarian cancer DNA methylation data

Next, we applied four statistical methods to real ovarian cancer DNA methylation data. Ovarian cancer DNA methylation data generated from Illumina Infinium HumanMethylation27K Beadchip has been already applied to identify CpG sites and genes associated with ovarian cancer from some different studies [19-21]. The methylation data set consists of 20,461 CpG sites from 12,770 genes with 152 controls and 119 cases. Many genes have either one or two CpG sites and some genes have up to 9 CpG sites. Since our four statistical methods can be applied to genomic data with a group structure, and our main goal of this study is to compare group testing procedures with group selection procedure, we excluded genes that have only one CpG site in the analysis. So, we ended up with 14,627 CpG sites from 6,936 genes which have at least two CpG sites.

In the exactly same way used in the simulation studies, we identified top 20 genes for each method. Table 1 shows the top 20 genes and their p-values computed by PCA and Hotelling's T^2 test. Also, 20 selected genes by group lasso procedure are included in the table. For the permutation test, we permuted the data over 1,000,000 times, but we found that the empirical p-values of 246 genes are still less than 10^{-7}. Due to time limit, we cannot reduce down the number of genes in the top list of permutation test anymore. Therefore, we had to finalize with top 246 genes detected by permutation test. For each statistical method, Fig. 4 summarizes the number of overlapped genes among top 20 genes by the other three methods.

In Fig. 4, it appears that 15 genes in the top 20 list of PCA is also in the top 20 list of Hotelling's T^2 test, while only 2 genes in the top 20 list of PCA is in the top 20 list of group lasso. Similarly, we can see that only 3 genes in the top 20 list of Hotelling's T^2 test is in the top 20 list of group lasso. Top 246 genes detected by permutation test include all of the 20 genes detected by both PCA and Hotelling's T^2 test, but only 7 genes among the 246 genes are overlapped with top 20 genes selected by group lasso. This result indicates that

Table 1. Top 20 gene lists and their p-values detected by PCA and Hotelling's T^2 test and top 20 gene list selected by group lasso

Top	PCA		Hotelling's T^2		Group lasso
	Gene	p-value	Gene	p-value	Gene
1	*KCNE1*	7.70×10^{-17}	*AIF1*	$<10^{-18}$	*BRD4*
2	*LMO2*	2.78×10^{-16}	*BRD4*	$<10^{-18}$	*C20orf65*
3	*RNASE3*	4.44×10^{-16}	*FCGR3B*	$<10^{-18}$	*C21orf56*
4	*GPR97*	5.89×10^{-16}	*KCNE1*	$<10^{-18}$	*CCL26*
5	*NFE2*	1.39×10^{-15}	*FYB*	1.11×10^{-16}	*CDH3*
6	*ENTPD1*	4.50×10^{-15}	*NFE2*	2.22×10^{-16}	*CXorf36*
7	*CSF3R*	5.79×10^{-15}	*CTSG*	4.44×10^{-16}	*GPR97*
8	*POR*	6.21×10^{-15}	*PNPLA2*	5.55×10^{-16}	*H2BFS*
9	*FYB*	7.59×10^{-15}	*LMO2*	6.66×10^{-16}	*HKR1*
10	*PPP2R4*	1.64×10^{-14}	*GPR97*	8.88×10^{-16}	*HLA-DQB2*
11	*PNPLA2*	3.78×10^{-14}	*ELOVL3*	3.77×10^{-15}	*KCNE1*
12	*IL27*	4.52×10^{-14}	*RNASE3*	3.77×10^{-15}	*LAX1*
13	*CSTA*	5.30×10^{-14}	*NR1I2*	4.44×10^{-15}	*LY9*
14	*CIAS1*	1.32×10^{-13}	*CSF3R*	5.33×10^{-15}	*NALP2*
15	*ELOVL3*	2.79×10^{-13}	*POR*	8.55×10^{-15}	*NYD-SP18*
16	*CD22*	2.85×10^{-13}	*TRPM2*	1.27×10^{-14}	*OLFML2A*
17	*MPO*	2.95×10^{-13}	*PPP2R4*	1.37×10^{-14}	*PTPN20B*
18	*CTSG*	5.58×10^{-13}	*CIAS1*	1.83×10^{-14}	*SLC9A2*
19	*C10orf27*	7.96×10^{-13}	*CSTA*	2.86×10^{-14}	*TM4SF1*
20	*GPR109A*	9.45×10^{-13}	*ENTPD1*	2.89×10^{-14}	*ZNF681*

PCA, principal component analysis.

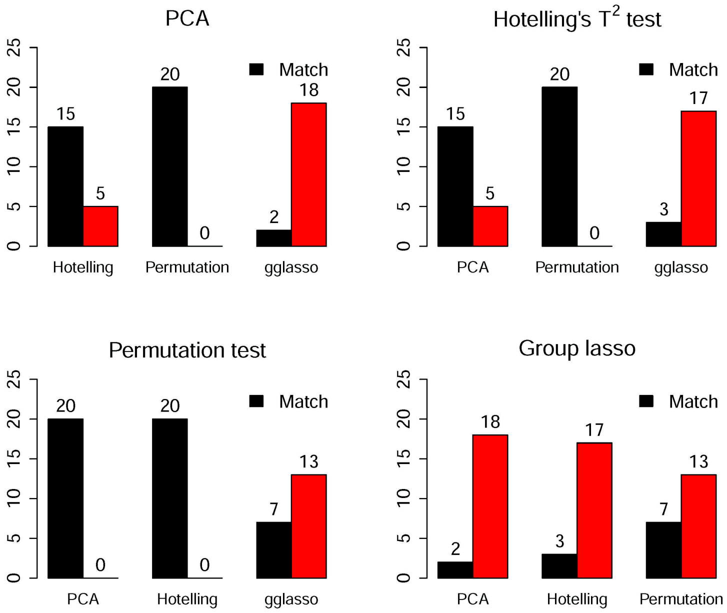

Fig. 4. The number of overlapped genes and non-overlapped genes in the top 20 lists of principal component analysis (PCA), Hotelling's T^2 test, permutation test, and group lasso are displayed in analysis of ovarian cancer DNA methylation data.

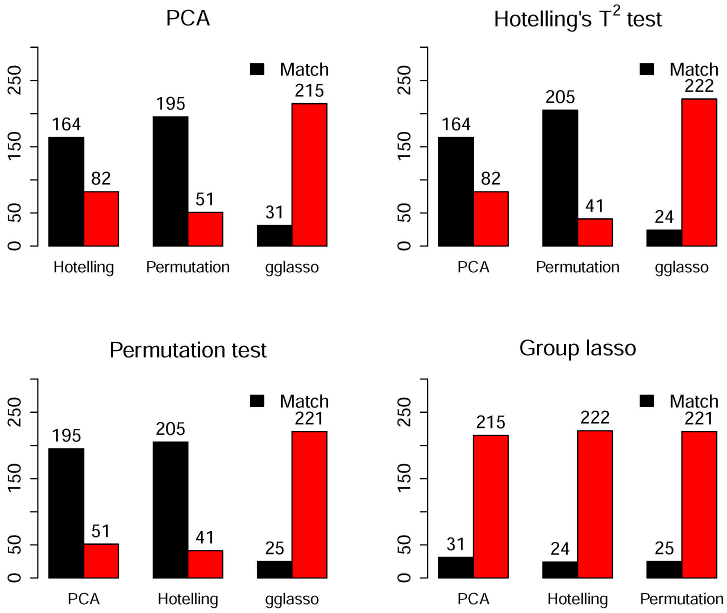

Fig. 5. The number of overlapped genes and non-overlapped genes in the top 246 lists of principal component analysis (PCA), Hotelling's T^2 test, permutation test, and group lasso are displayed in analysis of ovarian cancer DNA methylation data.

genes selected by group lasso are quite different from genes detected by three group testing procedures.

Next, we identified top 246 genes by each of four different statistical methods since the minimum number of genes detected by permutation test is 246. Fig. 5 summarizes the number of overlapped genes among 246 genes by the other three methods. It seems that three group testing procedures of PCA, Hotelling's T^2 test and permutation test have from 164 (66.67%) to 205 (83.73%) overlapped genes with each other. In contrast, group lasso selection procedure have only 24 (9.76%) to 31 (12.60%) overlapped genes with three group testing procedures. In this result, we can conclude that most of genes selected by group lasso are very different from genes detected by multiple group testing procedures in high-dimensional DNA methlyation data analysis.

In the first simulation study three group testing procedures and group lasso selection procedure show very similar selection performance. However, selection performance of four methods was quite different from each other in the second simulation study. We demonstrated that true positive selection could vary on the number of causal sites and the noise level. In analysis of ovarian cancer DNA methylation data, we found that group lasso identified quite different genes, compared with three group testing procedures. In general, regularization procedures like group lasso are known as a good alternative method to testing procedures when we identify disease or trait associated genes with high-dimensional genomic data, where the number of genes far exceeds the number of samples. However, our investigation found that many genes selected by group lasso are rarely overlapped with genes detected by three group testing procedures, which detected many overlapped genes. The similar situation was observed when the number of causal sites is small and the noise level is relatively large in our second simulation study. In that case, true positive rates of group lasso are much higher than those of three group testing procedures, where three testing procedures shows the almost same selection performance. In real DNA methylation data multiple genes are usually highly correlated with each other. Moreover, the number of causal sites could be small and the noise level could be large. But, further biological investigation with genes selected by group lasso should be conducted to figure out the main reason of this big discrepancy between group lasso and group tests in analysis of DNA methylation data.

Discussion

In this article, we compared group testing procedures with group lasso selection procedure when high-dimensional genomic data with a group structure are used for case-control genetic association studies. In statistics, hypothesis testing and variable selection are regarded as totally different statistical methods since their objectives are different from each other. Therefore, the comparison of these two statistical approaches has not been studied much. However, in genetic association studies with high-dimensional genomic data, both testing and selection procedures have the same goal which is to identify genes of genetic regions that are associated with either a disease or a trait. Particularly, many types of high-dimensional genomic data consist of multiple groups where each gene or each genetic region contains some genetic sites or variants. So, our research focused on the comparison of group testing procedures and group lasso selection procedure.

In simulation studies, we found that the selection performance of group lasso and group testing procedures could be similar or very different from each other. It depends on data structure such as correlation strength and patterns among genes, the number of causal sites in a gene and noise levels. In real data analysis, it was surprising that group lasso identified many different genes that are not detected by group testing procedures in ovarian cancer association studies of DNA methylation data. Although group lasso is known as one of the most commonly used selection methods in statistics when the number of variables is much greater than a sample size, it shows unexpected selection results in ovarian cancer data analysis. In contrast, multiple group testing procedures including PCA, Hotelling's T^2 test and permutation test identified almost the same genes associated with ovarian cancer. Since multiple group testing procedures are still the most popular method for medical doctors and geneticists to apply for high-dimensional genomic data with a group structure, we might need to further investigate the validity of group lasso selection procedure in genetic association studies. In future study, our investigation will focus on finding additional reasons that we had many genes detected by group test procedures but missed by group lasso in ovarian cancer data.

Acknowledgments

We thank the reviewers for very helpful and valuable comments. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2014R1A1A2054178).

References

1. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Methodol* 1995;57:289-300.
2. Storey JD. A direct approach to false discovery rates. *J R Stat Soc Series B Stat Methodol* 2002;64:479-498.
3. Storey JD, Tibshirani R. Statistical significance for genome-wide studies. *Proc Natl Acad Sci U S A* 2003;100:9440-9445.
4. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Methodol* 1996;58:267-288.
5. Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* 2009;25:714-721.
6. Alexander DH, Lange K. Stability selection for genome-wide association. *Genet Epidemiol* 2011;35:722-728.
7. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *J R Stat Soc Series B Stat Methodol* 2006;68:49-67.
8. Ma S, Song X, Huang J. Supervised group Lasso with applications to microarray data analysis. *BMC Bioinformatics* 2007;8:60.
9. Jolliffe IT. *Springer Series in Statistics. Principal Component Analysis.* New York: Springer-Verlag, 2002.
10. Chen M, Cho J, Zhao H. Incorporating biological pathways via a Markov random field model in genome-wide association studies. *PLoS Genet* 2011;7:e1001353.
11. Lee S, Epstein MP, Duncan R, Lin X. Sparse principal component analysis for identifying ancestry-informative markers in genome-wide association studies. *Genet Epidemiol* 2012;36:293-302.

12. Sun H, Wang S. Network-based regularization for matched case-control analysis of high-dimensional DNA methylation data. *Stat Med* 2013;32:2127-2139.
13. Lu Y, Liu PY, Xiao P, Deng HW. Hotelling's T^2 multivariate profiling for detecting differential expression in microarrays. *Bioinformatics* 2005;21:3105-3113.
14. Kong SW, Pu WT, Park PJ. A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics* 2006;22:2373-2380.
15. Cheung YH, Wang G, Leal SM, Wang S. A fast and noise-resilient approach to detect rare-variant associations with deep sequencing data for complex disorders. *Genet Epidemiol* 2012;36:675-685.
16. Park H, Niida A, Miyano S, Imoto S. Sparse overlapping group lasso for integrative multi-omics analysis. *J Comput Biol* 2015; 22:73-84.
17. Whittaker J. *Graphical Models in Applied Multivariate Statistics*. New York: John Wiley & Sons, 1990.
18. Peng J, Wang P, Zhou N, Zhu J. Partial correlation estimation by joint sparse regression models. *J Am Stat Assoc* 2009;104:735-746.
19. Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, Shen H, *et al*. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res* 2010;20:440-446.
20. Sun H, Wang S. Penalized logistic regression for high-dimensional DNA methylation data with case-control studies. *Bioinformatics* 2012;28:1368-1375.
21. Chen Y, Ning Y, Hong C, Wang S. Semiparametric tests for identifying differentially methylated loci with case-control designs using Illumina arrays. *Genet Epidemiol* 2014;38:42-50.